

FOSTERING UNDERSTANDING OF NATIONAL PARK VISITATION TRENDS  
THROUGH QUANTITATIVE METHODS AND VISUALIZATION

by

Adrienne Camille Torielli  
A Thesis  
Submitted to the  
Graduate Faculty  
of  
George Mason University  
in Partial Fulfillment of  
The Requirements for the Degree  
of  
Master of Science  
Geoinformatics and Geospatial Intelligence

Committee:

_____	Dr. Dieter Pfoser, Thesis Director
_____	Dr. Arie Croitoru, Committee Member
_____	Dr. Andreas Züfle, Committee Member
_____	Dr. Anthony Stefanidis, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Fall Semester 2017 George Mason University Fairfax, VA

Fostering Understanding of National Park Visitation Trends through Quantitative  
Methods and Visualization

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of  
Science at George Mason University

by

Adrienne Camille Torielli  
Master of Arts  
American Military University, 2015  
Bachelor of Science  
United States Air Force Academy, 2011

Director: Dieter Pfoser, Professor  
Department of Geography and Geoinformation Science

Fall Semester 2017  
George Mason University  
Fairfax, VA

Copyright 2017 Adrienne Camille Torielli  
All Rights Reserved

## **DEDICATION**

This is dedicated to my kind and patient husband, for making me laugh through all challenges, and my parents, for their constant support and understanding.

## **ACKNOWLEDGEMENTS**

I would like to thank the many friends, relatives, and supporters who allowed me the time to pursue this program. My husband, Bryan, assisted me in many computer-related endeavors. The instruction of Drs. Pfoser, Croitoru, and Züfle was invaluable in developing this research.

## TABLE OF CONTENTS

	Page
List of Tables .....	vii
List of Figures .....	viii
List of Abbreviations .....	x
Abstract .....	xi
Chapter One .....	1
Introduction .....	1
Purpose / Motivation .....	2
Relation to Previous Research.....	3
Chapter Two.....	5
Literature Review .....	5
National Parks Reports & Research .....	5
NPS Reports .....	6
National Parks Research .....	12
Parks & Tourism .....	14
Other National Park Systems.....	14
Tourism.....	15
Methods .....	16
Knowledge Discovery in Databases .....	16
KDD & Visualization .....	19
Research Question.....	21
Chapter Three.....	23
Methodology .....	23
Data .....	25
Sources.....	26
Cleaning.....	33
Integration.....	38

Selection .....	40
Analysis .....	42
Transformation .....	43
Data Mining .....	45
Interpretation .....	46
Knowledge .....	46
Chapter Four .....	48
Results & Interpretation .....	48
NPS Visitation .....	49
External Variables .....	68
Chapter Five.....	79
Key Findings .....	79
Significant Results .....	80
Discussion .....	81
Limitations .....	85
Chapter Six.....	87
Conclusion.....	87
Future Research.....	88
References .....	89

## LIST OF TABLES

Table	Page
Table 1 Annual Visitation Summary Report (NPS 2016) .....	3
Table 2 Excerpt of Park Metadata Table .....	35
Table 3 External Variables.....	36
Table 4 Data after Preprocessing .....	37
Table 5 Total Visits By Type 1979-2016 .....	44
Table 6 Parks by Visit Type.....	50
Table 7 Percent Increase since 1979 .....	51
Table 8 Least-visited Parks .....	55
Table 9 Park Visitation (other than Recreation) .....	59



## LIST OF FIGURES

Figure	Page
Figure 1 Percentage of Recreation Visits by Type of Unit (NPS 2016) .....	8
Figure 2 Excerpt of Parks with Fees (NPS "Plan Your Visit," 2016).....	9
Figure 3 Recreation Visits based on proximity to MSA.....	10
Figure 4 Percentage Change of Recreation Visits by State or Territory, 2015 to 2016 ...	11
Figure 5 KDD Process (Fayyad, Piatetsky-Shapiro & Smyth1996).....	18
Figure 6 Modified KDD Process .....	23
Figure 7 NPS Boundary Layer in CARTO .....	26
Figure 8 Center points of all NPS System Units.....	27
Figure 9 NPS IRMA Query Builder .....	28
Figure 10 EIA Annual Average Gas Price not adjusted for inflation .....	30
Figure 11 NCEI Average Annual Temperature .....	32
Figure 12 Database Design showing relationships between tables .....	39
Figure 13 Nearest Neighbor Join .....	41
Figure 14 Excerpt of Pivot Table for Visits by Year and Unit Code.....	44
Figure 15 All Visits by Year .....	48
Figure 16 Visits by Park Type 1979 - 2016.....	52
Figure 17 Parks by Number of Visits, color denotes park fee .....	53
Figure 18 Annual Difference by Visit Type .....	56
Figure 19 Visits Averaged by Number of System Units .....	58
Figure 20 All Other Visit Types .....	59
Figure 21 Parks with Net Decrease.....	60
Figure 22 Closest Large City to Each Park (color).....	61
Figure 23 Parks Clustered by Number of Visits and Distance to Large City .....	63
Figure 24 Clusters by Visit Type .....	65
Figure 25 Cluster Description (not including recreation visits).....	66
Figure 26 Parks Grouped by Similar Number of Visits.....	67
Figure 27 Visits Normalized by Annual Population, Visits, and Annual Population .....	69
Figure 28 Relationship between Gas Price and Visits .....	70
Figure 29 Relationship between Unemployment and Visits.....	70
Figure 30 Relationship between Average Temperature and Visits .....	71
Figure 31 Visits under “normal” conditions .....	73
Figure 32 Visits under “abnormal” conditions .....	73
Figure 33 Regression Model Summary (Year) .....	75
Figure 34 Regression Model Summary (Pop) .....	75
Figure 35 Regression Model Summary (number of parks) .....	76

Figure 36 Regression using US-only visits..... 77

Figure 37 Capture of Interactive Web Map ..... 79

## LIST OF ABBREVIATIONS

Bureau of Labor and Statistics .....	BLS
Consumer Price Index .....	CPI
Department of the Interior .....	DOI
Energy Information Administration .....	EIA
Geographic Information System .....	GIS
Geographic JavaScript Object Notation.....	GeoJSON
Integrated Resource Management Portal .....	IRMA
Knowledge Discovery in Databases .....	KDD
Metropolitan Statistical Area .....	MSA
National Centers for Environmental Information .....	NCEI
National Oceanic and Atmospheric Administration .....	NOAA
National Park Foundation .....	NPF
National Park Service .....	NPS
Structured Query Language .....	SQL
United States Geological Survey .....	USGS

## **ABSTRACT**

### **FOSTERING UNDERSTANDING OF NATIONAL PARK VISITATION TRENDS THROUGH QUANTITATIVE METHODS AND VISUALIZATION**

Adrienne Camille Torielli, M.S.

George Mason University, 2017

Thesis Director: Dr. Dieter Pfoser

The United States Department of the Interior recognizes and protects over 400 entities, designated as National Park System Units, for public use and enjoyment. Celebrations for the National Park Service (NPS) centennial in 2016 highlighted the growth of the park system, particularly the recent uptick in visitors beginning in 2014. For the first time in 2016, overall visits reached over 500 million. Concerns over the ability to maintain the parks amid their resurgent popularity and financial constraints highlight the challenge of administering the areas comprising the parks. Historical visitation data is available for public use through the National Park Service dating back to 1979. This research sought to analyze the public use data set to determine if the current park popularity is part of an existing trend and discern the reasons for any changes. Annual summaries of visitation by NPS group certain trends by administrative region, as parks are located throughout the United States and its

territories. However, visual representations of the park system are limited to charts depicting annual changes and aggregations of data by state. Previous reports indicate park visitors often do not come from communities that most closely and geographically surround the parks themselves. In addition, the parks are considered a tourist destination, influenced by variables impacting leisure expenditures, like economic conditions. Conducting exploratory visual analysis to identify spatial and temporal attributes of visitation allowed for an assessment of the entire historical data set, while other data science techniques provided methods for testing variables outside of park data. A web-based map highlighting the significant findings will serve as a reference for users interested in visualizing the information and exploring additional trends.

## **CHAPTER ONE**

### **Introduction**

The National Parks Service (NPS), overseen by the United States Department of the Interior, provides access to federally-designated locations throughout the U.S. and its territories for visitors, campers, and researchers. Although there are fifty-nine National Parks, there are four-hundred seventeen total sites comprising twenty different park categories (NPS FAQ). These categories include national battlefields, historic sites, memorials, monuments, preserves, and seashores, among others (NPF Blog 2016). For three consecutive years beginning in 2014, the overall number of recreation visitors to the National Park system broke the previous record. This came after several years of a steady decline (Hetter 2015, Flowers 2016, NPS FAQ 2017). While the Park Service addresses some possible reasons for the increase in their annual summary report, it is not a primary objective to understand why visitation changes each year. Both the NPS and the National Park Foundation (NPF), which is charged with fundraising and outreach for the parks, support diversity initiatives meant to increase visitors. Of the few existing reports, both NPS and NPF acknowledge the parks largely attract a demographic of older, white Americans. Ultimately, examining a variety of possible impacts to park visitation will assist NPS in determining the cycle of visitors and possibly reveal some spatial correlation to the park visitation increases.

## **Purpose / Motivation**

The increasing popularity of the National Parks provides both challenges and opportunities to visitors and the Park Service. NPS welcomes new and returning visitors, but is also hoping to increase visitation among a younger and more diverse demographic. In addition, more visitors may stretch limited funding and resources. Therefore, it is important to look beyond the national trend to determine the spatial characteristics of park visitation. In their annual report, NPS alluded to weather, the improving economy, gas prices, and greater publicity as possible reasons for the number of visitors. However, it is not known whether, or to what extent, these variables impacted visitation. It is possible that park services may suffer due to an overwhelming number of visitors, thereby threatening the future viability of services offered. Understanding these relationships to park visitation may allow NPS to anticipate and prepare for park visitation fluctuations and better anticipate visitor travel motivations. This research serves as a macro-level study of park visitation to provide greater access to almost four decades worth of information in a manner not yet attempted. Visualizing information, particularly spatial information, provides functionality to NPS data that does not currently exist on such a wide scale. Existing NPS reports focus only on annual comparisons and presenting data, without necessarily delving into explaining the possible causes or reasons for year-to-year change. Literature on the KDD process, particularly geospatial datasets, provides a framework for further research without explicitly indicating the exact or best steps necessary. Therefore, the methodology for assessing the visitation and its possible external factors will fit within the general KDD process, while the exact methodology remained unique.

## Relation to Previous Research

Few research papers examined the park system including all park and visitation types. System-wide research tended to assess a subset of park types or identify trends over a shorter time-period. None of these tested the creation web-based map or test variables concerning visitation. NPS publishes reports detailing statistical changes in visitors for all parks compared to the previous year. An example of a typical summary report, shown in Table 1, illustrates the type of information readily available for public query (NPS IRMA Summary Report 2016).

<b>Table 1 Annual Visitation Summary Report (NPS 2016)</b>	
Category Summary	
<b>Recreation Visits</b>	330,971,689
<b>Recreation Visitor Hours</b>	1,427,664,670
<b>Non-Recreation Visits</b>	172,285,627
<b>Non-Recreation Visitor Hours</b>	92,895,795
<b>Concessioner Lodging Overnights</b>	3,272,026
<b>Concessioner Camping Overnights</b>	1,294,573
<b>Tent Camper Overnights</b>	3,858,162
<b>Recreation Vehicle (RV) Overnights</b>	2,543,221
<b>Backcountry Overnights</b>	2,154,698
<b>Miscellaneous Overnights (Groups and Aboard Boats)</b>	2,156,818
<b>Non-Recreation Overnights</b>	150,982
<b>Total Overnight Stays</b>	15,430,454

NPS attempts to forecast upcoming visitation for the next two years based on past visitor statistics. These forecasts only consider recreation visits for each park, with no assessment to determine the accuracy of those forecasts. Other existing studies focus on



comparison to previous year, on certain types of visits, and generalize the impact of external variables on visitation without direct comparison. Some research regarding national parks was not conducted concerning NPS, as a host of other countries also have protected lands equivalent to the major U.S. system units categorized as “National Parks.” These studies were included, even though they were not specific to American parks, because researchers there identified similar gaps and challenges not previously addressed. A wide variety of journal publications, textbooks, and other primary sources were reviewed, as the topics researched required an interdisciplinary study.

However, previous geoscience, spatial analysis and even social media studies have been done in single parks. Much of the previous research investigated for this effort was to discover techniques for using web mapping services and conducting spatial analysis. As Haklay, Singleton, and Parker indicated, web-based GIS does not offer additional functionality compared to standalone systems. Rather, it is different due to its ability to reach a broad audience and design an interface that is intuitive for users (Haklay, Singleton, & Parker 2008). Therefore, searching for web mapping resources was limited to only those sources that offered insight into how to design and run efficient queries. Much of this was not in the form of academic research papers, however, because many of these theories are standardized and no specialized approach was necessary.

## **CHAPTER TWO**

### **Literature Review**

The National Park Service provides access to a wealth of primary sources for use by the public, from raw data to annual reports. In addition, research about within this, and other, national park systems, was reviewed to gather a broad range of perspectives about parks and visitation. This included literature from other disciplines, including research on travel and tourism, who work closely with national parks to promote visitation (Blaszak 2006). This industry is designed to predict visitation based on numerous factors, and was instrumental in defining certain variables. Because this research efforts required knowledge outside of the National Parks, literature regarding knowledge discovery in databases (KDD), data visualization, and other techniques was reviewed to understand various processes and methods.

### **National Parks Reports & Research**

While the National Parks host researchers of all disciplines, little research exists about the 417 system units, or even 59 Parks, on a national scale. The Park Service itself releases reports on activities, and supports research within its lands, while the National Park Foundation offers grants to projects of interest. This at least provided several primary sources to gather detailed information. One of the only discussions of trends in the NPS system was the transcript of a statement provided by the Directors of two different Park regions. They provided a statement to Congress in 2006, which mentioned

a few of the relationships between the variables examined in this research (Blaszak 2006). This statement was on a relatively obscure website, hosted by the office representing DOI in Congress and furthering legislation at achieving its goals. Fluctuating visitor numbers across the entire park system for specific years were attributed to extreme weather events, changes in American travel patterns, openings and closings of certain parks, economic conditions, and demographics. One or more of these reasons contributed to the overall stagnation of visits for the previous ten-year-period, even including the record-high year of 1999 (Blaszak 2006).

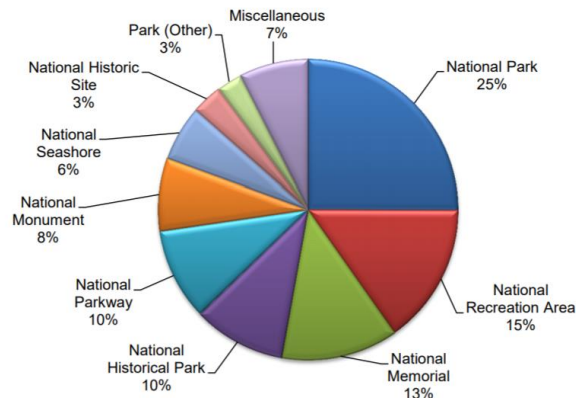
In this statement, NPS also explained visitor trends using similar external variables to this research. Unfortunately, there was no supporting documentation or additional references to discern the data used or how NPS reached those conclusions. They were presented as fact, like how research from one university discovered meaningful connection between rising real disposable income and decreasing park visitation. Separately, this statement highlighted that higher unemployment led to a decline in visitation. Therefore, while other reports may reference this information or accept it as truth, there was still additional value in using unemployment in this research, especially as combined with other variables. Searches for references included to develop this report were unable to return any of the original research.

### **NPS Reports**

The National Park Service publishes statistical abstract reports on an annual basis dating back to 1965. These primarily report the number of all visitors to each park, but also include the breakdown of visitors by visit type. Some reports also include

comparison to the previous year, such as the difference and percentage difference of types of visits at each park. Their abstracts also adjust park visitation statistics to normalize them as much as possible between years when there are special events that might skew that data. However, there was no definitive information on whether these changes were reflected in the raw visitor use statistics available for download. For example, the report in 2016 detailed errors in visit estimates or changes in the process for counting visits for twenty-nine parks during 2015. It calculated adjusted visit values for these parks to compare them. But previous reports did not detail whether these changes were retroactively applied to previous visitor statistics, or if that was necessary, as there were no summaries encompassing multiple years.

In the same way that park visitation experiences increased visitation based on the season, a historical approach to visitation revealed some interesting characteristics. This was displayed in some instances as a chart or on a map of the contiguous United States, or in other cases as a comparison of the average of certain visit types over a multi-year period. One type of representation, used in the annual statistical abstract report, included a pie chart comparing percentages of recreation visits based on the visit type, indicated in Figure 1 (Ziegler 2016). However, there is not a consistent statistical study across this entire period.



**Figure 1 Percentage of Recreation Visits by Type of Unit (NPS 2016)**

Discovering simple information often required multiple searches. The information was not concise, nor was one master document discovered that listed every detail about each park. For example, the Visitor Use Statistics page has many datasets formatted and ready to download, while they also offer a query to add additional information (NPS IRMA 2017). However, it did not include information such as the year the park began operation, or the entrance fee to that park. There was no current single listing of parks charging entrance fees, although the information about the types of fees and rates were available on each specific park website. The closest thing to a list of parks charging entrance fees was a post on the NPS website highlighting the days in which fees were waived at those parks, shown in Figure 2. There may be parks charging fees not accounted for on that list. Some parks charge more than an entrance fee based on activities and amenities available, but this information was even more obscure.

## Free Entrance Days - Participating Parks |

Parks listed on this page waive their entrance fees on nationally designated [fee-free dates](#), hours and event schedules.

By Name | [By State](#)

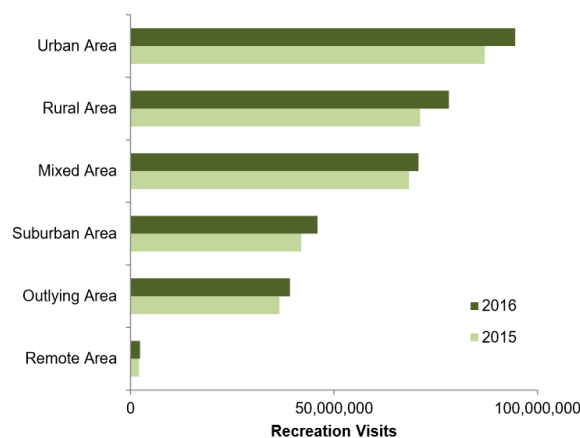
[Acadia National Park](#), Maine  
[Adams National Historical Park](#), Massachusetts  
[Antietam National Battlefield](#), Maryland  
[Arches National Park](#), Utah  
[Assateague Island National Seashore](#), Maryland-Virginia  
[Aztec Ruins National Monument](#), New Mexico  
[Badlands National Park](#), South Dakota  
[Bandelier National Monument](#), New Mexico  
[Big Bend National Park](#), Texas  
[Bighorn Canyon National Recreation Area](#), Montana-Wyoming

**Figure 2 Excerpt of Parks with Fees (NPS "Plan Your Visit," 2016)**

Another challenge acknowledged widely throughout NPS involves capturing visitor information. Because parks have multiple entrance points, are in urban or public spaces with no entrances and are open to visitors that may be visiting for a day, camping, or even kayaking through a space, NPS has devised a system to standardize counting of persons within the park. However, they do not regularly ask for demographic information, track the visits by pass holder type, or ask for even the state of residence from visitors. This type of data would possibly answer questions about park use, but a system has not been designed yet to capture it. However, the NPS published the “National Park Service Comprehensive Survey of the American Public” in 2011 after a study between 2008 and 2009 (Taylor, Grandjean & Gramann 2011). This study highlighted an overwhelming majority of visitors were white and not Hispanic. High travel costs and long travel times were two of the top three reasons cited for not visiting a park site more often (11). The study was conducted in both Spanish and English, however, but with only approximately four thousand responses to the phone survey.

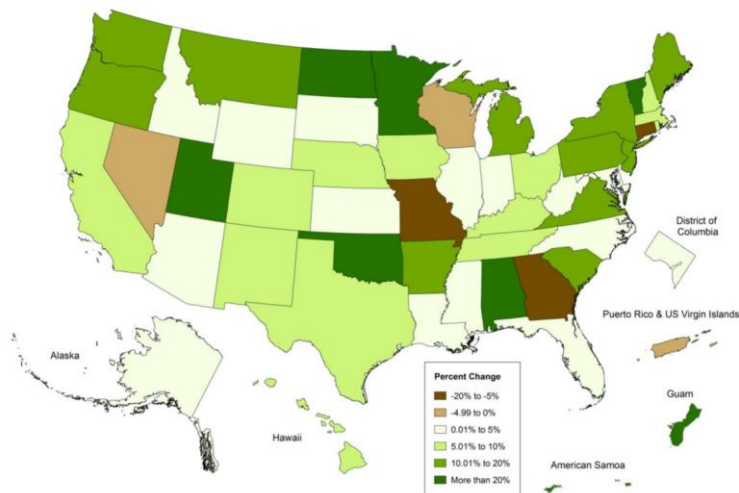
Additional languages and methods of conducting interviews may yield different results in future surveys.

The closest proxy to comparing demographics or population to visitor statistics was analysis based on the extent of which a park's boundaries were contained within a given Metropolitan Statistical Area (MSA). Using the park boundary may be better than other methods for approximating demographics, but this does not account for travel over roads or the placement of park entrances. One challenge to using this, and other data from the U.S. census, is the frequency in which the populations are estimated and the MSAs are updated. Dramatic population shifts in some cities or regions may result in a park changing from one MSA to another between years, although the details of the MSAs and the parks within them was not listed within the NPS reports. The 2010 MSAs, defined by the US. Census Bureau, were categorized as shown in Figure 3, to compare visitation to the population density within or surrounding a park (Ziegler 2016).



**Figure 3 Recreation Visits based on proximity to MSA**

No comprehensive study on visitor statistics dating back to 1979 was conducted using spatial information. However, the National Park Service published a special report on activities related to the NPS Centennial (Ziegler 2016). Many news articles reference this report and the annual visitation statistics, with indications of record-breaking years sometimes acknowledged when the mid-year report by the NPS is released. In any case, most new articles report the basic facts regarding the number of recreation visits to the park without detailing the different types of visitors, how the numbers are estimated, and often only focus on the system units called “National Parks.” Very few attempts exist to visually display this information, although NPS plotted the number of visitors in the line graph over the past five years. Visual representations included a display of parks based on the percentage change of visits. The visit information was displayed based on the park location by state, shown in Figure 4 (Ziegler 2016).



**Figure 4 Percentage Change of Recreation Visits by State or Territory, 2015 to 2016**



## **National Parks Research**

One of the few studies about the Park System was in terms of its financial value, modelling the amount park visitors would be willing to pay for certain activities based on NPS survey data (Neher, Duffield & Patterson 2013). On average, it estimated the park visitors would be willing to pay an average of \$102 across all parks, greater than the current annual pass price of \$80. Even more interesting, however, was a study conducted at Yellowstone National Park. Benson, Watson, Taylor, Cook, and Hollenhorst grouped visitors depending on the activities they participated in during their time at the park, then estimated how much that visit was worth to the visitor. In a sense, this study assigned value to visits based on activity instead of requesting that information. The visitors received a benefit between approximately one-hundred to seven-hundred dollars, well above the annual pass price (Benson et al. 2013). However, based on the results from Taylor et al.'s report, cost is still a significant barrier for many people. Weber and Sultana approached the challenge of demographics differently than the NPS reports, focusing on the geographic distribution of parks to determine its effect on visits for non-white visitors (Weber & Sultana 2013). They studied a smaller subset of park units and determined that non-white visitors tended to visit parks closer to them (2013). Similar to the construct used in this research, Schuett, Le, and Hollenhorst attempted to understand visitation trends based on a combination of variables. Their study focused on the composition of groups of visitors, and included an analysis of visitor surveys and how far each group had traveled to visit each park type, resulting in a better understanding of how different types of parks attract different types of people, based on group size and whether the visitor was new to the parks or a repeat visitor (Schuett, Le & Hollenhorst 2010, 206-208).

Another more fascinating study regarding the entire park system measured the correlation between monthly visitation from NPS and the monthly mean air temperature from the Climatic Research Unit (Fisichelli, Schuurman, Monahan & Ziesler 2015). This research grouped 340 of the system units to determine which exhibited strong relationships between weather and visitation. Park visitation increased with temperature until around 77 degrees Fahrenheit, along with other trends based on park location. This information was utilized with other temperature data for this study. One limit to this effort was that the researchers did not include a listing of the parks they analyzed by name, only providing an output map, requiring generalizations to be made if the concepts were to be included. Another limitation was from the temperature data, which was only available at a resolution of 0.5 decimal degrees. This covers approximately between 1250 and 1850 square miles on the surface of the Earth at most latitudes where parks are located. While some parks indeed cover large swathes of space, there is a risk of generalization in highly concentrated park areas. Given the dearth of other park-wide efforts, and the difficulty in assessing temperature across the parks over such a long span of time, this study was instrumental in establishing a relationship between temperature and park visitation at over 80 percent of parks. Therefore, further analysis will seek to use this general relationship when assessing an effect of temperature, but not make it the primary focus.

Dye and Shaw studied the ability to create essentially a menu for users to define the type and difficulty of a trail, along with other activities within Great Smoky Mountain National Park (2005). This type of function could easily be transitioned to a web map,

since this research was conducted in 2005 and relied on Visual Basic computer interface. The NPF website allows for searching parks based on location, but not searching activities and trails within parks. Although older, this research indicates the long history of attempting to modernize NPS and make information more readily available.

### **Parks & Tourism**

It was imperative to gather information from journals and publications dedicated to topics ranging from computer science, geography, and politics to fully grasp the current state of research for this effort. Because park visitation was considered among other leisure activities and tourist travel, research conducted by, or assessing the impact to, the tourism industry provided insight into the external factors influencing visitation. Assumptions made by NPS in its press releases and statistical reporting regarding the impact of temperatures, the overall economy, and population were not explicitly outlined, but rather presented as common knowledge. This type of assumption, such as lower gas prices leading to increased vehicle travel and visitation to tourist destinations, is often repeated in news articles and other studies without acknowledgement of its veracity or the strength of its correlation to different types of tourism.

### **Other National Park Systems**

A study on a Swedish park, by Fredman, Friberg, and Emmelin evaluated the change in visitation to a park after it was officially designated a “National Park,” (Fredman, Friberg & Emmelin 2007). The concept of designation as an influence factor was not explicitly studied in this research, but provided an explanation for assessing parks based on their type, particularly after the researchers noticed a forty-percent

increase in visitation in the year after the change was made (87-89). Although Rodger, Taplin, and Moore conducted their study on a remote Australian national park, the methods used highlight the unique challenges experienced in park systems around the world. Rather than relying on observed information, they attempted to prove the same causal relationship by manipulating customer satisfaction in an experiment, but discovered diverging results. Instead of focusing on customer satisfaction, Cessford and Muhar discussed management techniques for observing visitor behavior and estimating visitor numbers (Cessford & Muhar 2003). The purpose was two-fold: to learn how groups behave to protect the resources, and cost-effectively track visitors to improve estimates without accounting for every individual (2003).

### **Tourism**

Eagles provided a list of research areas requiring additional study to best benefit and maintain parks (Eagles 2013). It included improvements to two overarching categories, visitors and tourism, to best plan and manage the park system's resources (544). Each of the ten primary research areas was meant to integrate and influence the other, highlighting the relationship between parks and tourism. Bonn, Line, and Cho modeled the effect of lower gas prices on tourism and tourism-related activities in a unique study. In it, the researchers focused on individuals traveling by car to vacation destinations in Florida (Bonn, Line & Cho 2016). The study cited previous research, highlighting there is not much focus on whether the hospitality industry suffers, and to what extent, when gas prices are high enough to impact the lodging industry. There is not much information about NPS regarding how many travel to reach a specific park, but

because there are lodging alternatives and the existence of parks within one day's drive of many urban areas, the effect of gas prices on other travel expenses provides another lens with which to consider this as a variable. Bonn, Line, and Cho discovered that a decrease in gas prices resulted in an increase in spending to attractions and events, as well as the desire to return to that destination (2016).

## **Methods**

Collecting and interpreting large amounts of data from various sources requires its own criteria and an understanding of various techniques, processes, and tools. With increasing interest in uncovering the hidden meaning in large datasets, or even the possibility that more data would lead to better conclusions, a review of prevailing Knowledge Discovery in Databases, or KDD, was required. Studying the process of KDD, to include the special situation of analyzing geospatial data, did not reveal one specific and clear solution to best understanding this dataset. In fact, it became apparent that developing one's own specifics was necessary, while still employing a set of standards to create a repeatable set of steps to further continue this research. A combination of two KDD processes provided the general framework used to assess the data, along with additional details involving visualization.

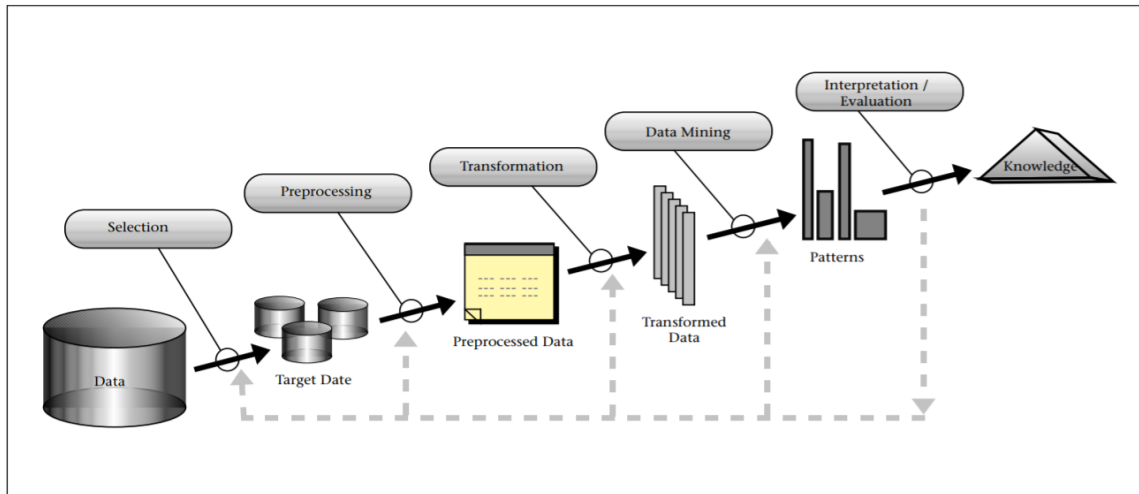
## **Knowledge Discovery in Databases**

One of the seminal resources on KDD by Fayyad, Piatetsky-Shapiro, and Smyth was published in 1996, long before the current obsession with data mining. The authors describe the relationship between both KDD and data mining, as well as the process for conducting KDD (Fayyad, Piatetsky-Shapiro & Smyth 1996). They argue that the

purpose of KDD is to make data relevant, so it can be summarized, reveal context, or establish meaning to better understand what similar information means in the future. Some of the definition for data mining continues to be debated, such as how to define whether the patterns or results meet the criteria of being novel or useful. While this opens fascinating possibilities, depending on how these terms are interpreted, the main purpose of their research was to illustrate KDD as a series of steps, and clarify its relationship to data mining. For this, they condensed the outline of KDD from previous research, while still recognizing the importance of iterating as necessary (Brachman and Anand, as cited by Fayyad, Piatetsky-Shapiro & Smyth 1996, 42).

Establishing the five steps for KDD was the primary lasting impact of this article. These steps are: selection, preprocessing, and transformation of the data, which wrangle data into a more useful format, followed by data mining and evaluation for analysis and creation of knowledge. The authors introduced data mining tasks such as linear regression, clustering, and outlier detection, but within the context of finding patterns in databases. In addition, they highlighted the necessity of tailoring the techniques to the specific problem, rather than establishing a hierarchy of data mining techniques. Fayyad, Piatetsky-Shapiro, and Smyth instead outlined the type and purposes of different techniques, indicating certain conditions for using each. Understanding that not all techniques or methods will be useful in every context, keeping the KDD steps and data mining overview vague, the provided KDD process is deliberately open to interpretation and flexible for a given context. Much research has cited this article, as it was introduced a method for KDD which could be applied to large datasets using computational

methods. These were increasingly available during that time. Their KDD Process has intermediate steps, but ultimately comprises five main functions, as seen in Figure 5.



**Figure 5 KDD Process (Fayyad, Piatetsky-Shapiro & Smyth1996)**

Han, Kamber, and Pei posited a more detailed KDD process, added steps for combining data sources and expanding the data interpretation process. As a textbook, the authors have more space to explain and expand on the ideas of earlier KDD research (Han, Kamber & Pei 2011). This changed the order of some of the initial steps and added additional details to acknowledge the presence of a database for managing information. For example, where earlier efforts adopted data selection as the first step in the KDD process, Han, Kamber, and Pei defined selection differently. Instead of dealing with the data source, the second set of research began their process with the assumption that the data sources already exist, and the first step of KDD involves cleaning the data. They also add a step for data integration for combining data sources. Although not explicitly stated

in their process, the use of a database for storing data tables was inferred. From there, it would be possible to conclude that data integration also involves the creation of the schema in a database or the establishment of relations between tables. This becomes clearer as they define the data selection step to involve retrieving the necessary data for a specific type of analysis from a database. From there, the process mirrors that of other KDD outlines. The authors do not argue that this is the best or only way to conduct KDD, instead attempting to offer another framework. However, the KDD process detailed by Han, Kamber and Pei provided the better framework for this research.

### **KDD & Visualization**

One aspect of KDD extends into another discipline, that of visualization. As mentioned in previous KDD research, visualizing results, whether they have a spatial, temporal, or other component, can be valuable to identifying patterns. Rather than delve into the arguments of whether the human brain or computers can better identify patterns or outliers, as there may be arguments on either side depending on the exact circumstances, the more important takeaway was highlighted by Gahegan, Wachowicz, Harrower, and Rhyne (Gahegan, Wachowicz, Harrower & Rhyne 2001). The authors summarized previous research efforts and the use of visualization in other disciplines, using both machines or manual human interpretation. Their article acknowledged the challenges of combining KDD and visualization, particularly with geographic data, but also that both serve to achieve goals related to exploring data and conducting analysis (Gahegan et al. 2001, 30). They also included greater detail connecting KDD to the overall scientific method, exploring models of reasoning based on their own and others'



research (Leedy 1993 and Baker 1999, as cited in Gahegan et al. 2001). This differed from the other studies of the KDD process, in illustrating how different models of reasoning like induction and deduction benefit from exploratory visual analysis, particularly when it comes to spatial patterns. Other challenges faced by early geovisualization proponents still exist, such as the challenge of incorporating data from non-spatial sources (MacEachren, Gahegan, Pike, Brewer, Cai & Lengerich 2004). Newer software provides improved functionality for displaying combined information, but usually by limiting the type of spatial analysis that can be performed.

At the time of their publication, the tools for databases and computational models were not as advanced, but it did not limit their ability to advocate for greater interaction between KDD and visualization. More recent research acknowledges certain advances, but also indicated that greater accessibility to data representing both space and time still requires an improved method, specifically for performing KDD (Gahegan & Kraak 2001, 4-7). Many of the concepts for geovisualization and KDD were reviewed by Mennis and Guo, who also indicated how the early stages of KDD were particularly important using spatial data (Mennis & Guo 2009). The human brain can indeed identify patterns, and in some cases, it is more important for the outcome to be seen by a given audience or open for interaction, than it is to determine which method is better. As this research attempts to process data on a much larger scale than previously attempted, creating visualization and geovisualization of preprocessed data.

## **Research Question**

News articles and the Park Service itself highlight the number of visitors, most only concerning the past few years at a time. In addition, none calculated whether these numbers are part of a longer trend based on the statistics dating back to 1979. The purpose of investigating park trends was to determine whether there was a historical precedent to the most recent influx of visitors and highlight visitation changes over the years. This research was foremost concerned with visualizing the trends in park visitation using the annual public-use statistics from 1979 to 2016. In addition, using the KDD process, various data transformations were performed, along with data mining in the form of clustering and regression. Comparisons of park visitation based on location is limited to states and regions, or percentage of the park boundary located within areas of differing population. Visualizations of the visitor statistics history and an interactive web-based map of visitation provided a method for exploratory visual analysis and a method for others to view and interact with the data. Currently, visual representations included a display of parks based on the percentage change of visits, with the visit information displayed based on the park location by state. Another primary goal was to determine if the increase in visitation has been equally distributed amongst the system units and visit types. Parks that show a decrease in visitors will be highlighted and further explored.

In addition, there has been no system-wide research attempting to identify the strength of correlation external variables to the number of visitors. Few visual representations of visitor statistics exist, predominantly displayed in charts comparing year-over-year visitor differences. Given the previous research into the tourism industry, a few hypotheses were tested. This research also attempted to determine the strength to

which favorable economic conditions, population, and anomalous weather correlate to the number of recreation visits to the National Parks within the continental United States using regression. For this research, “good” economic conditions were described with the following parameters: no park entrance fee, lower than average annual gas price, and lower than average unemployment. It is also hypothesized that the presence of entrance fees will not be related to visitation, as NPS offers a variety of methods for people of all incomes to visit parks throughout each year. Both low gas prices and low unemployment were hypothesized to correlate to higher park attendance. Although not exactly tested the same way, the effects of weather and population density on park attendance have already been measured. Therefore, only a very high-level overview of the relationship between these two indicators, over time, was tested. Given the current research outcomes, temperatures closer to normal levels and closer distances to large cities are expected to correlate with higher visitor levels. Regardless of the correlation to visits, the outcome should be visually and spatially represented to encourage further research.

## CHAPTER THREE

### Methodology

The Knowledge Discovery in Databases process outlined by Han et. al provided a framework for this research. Although it has more steps than the process illustrated by Fayyad, Piatetsky-Shapiro, and Smith, these seven functions were appropriate as well. Integrating visualization within KDD, as highlighted by Gahegan et al., provided an additional avenue for investigating visitation. Rather than listing each step on its own, related processes were grouped together, resulting in the process shown in Figure 6.

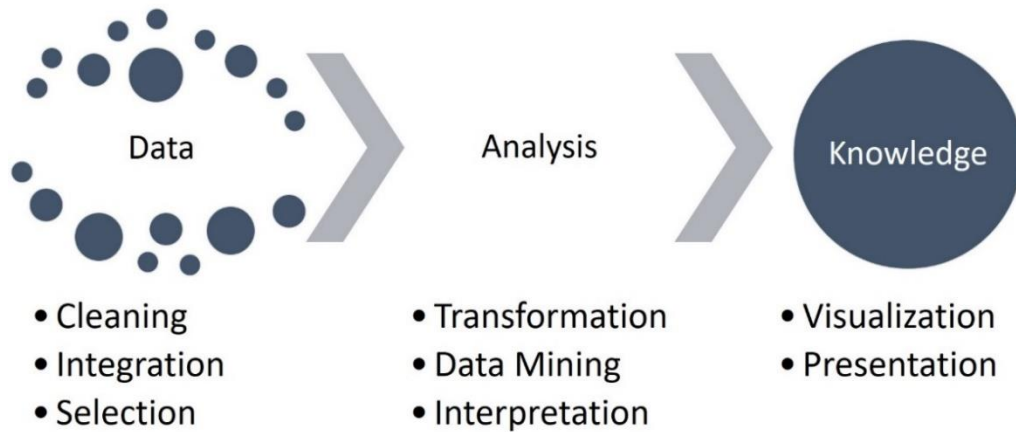


Figure 6 Modified KDD Process

Three major categories were established, with their relevant KDD processes as a subset of that section. This allowed for a more simplified methodology, particularly because iteration occurred within each major section: data, analysis, and knowledge. General best practices in KDD and visualization, allowed for exploration of NPS visitation on a greater scale than previously attempted across all park system units.

The data section of the KDD process involved finding primary sources, processing or cleaning that data, creating a database for storage, integrating multiple sources together as required, and the function of retrieving or selecting the data from storage. Data was stored as tables in a database managed through CARTO, which could be joined based on location or relationship. This resulted in a database to manage all the NPS visitation data, NPS park boundary polygons, NPS park centroids, and all external variables. The relationships between the park visitation and location data, were joined based on the unique park unit code.

The second section, analysis, of the KDD process included any effort to understand the prepared data. In this case, the transformation step carried more importance than in Fayyad et al.'s initial model, as it included all efforts to summarize and aggregate the data. This step resulted in a basic understanding of the visitor statistics, such as calculating sums and averages of information across different dimensions, such as by park, year, or visit type. This step was also important to understanding the external variables, resulting in information detailing average, or normal, conditions to determine which values existed outside of that. Another facet of the analysis section involved data mining using a variety of tools. The current NPS polygon, point layer, and annual

visitation statistics are managed through CARTO, an online database and visualization service with some spatial and Structured Query Language (SQL) functions. In addition, both Excel and Tableau, a data visualization software for relational databases, were used to discover patterns within the visitation statistics not yet explored. Applying existing algorithms, conducting outlier detection through exploratory visualization, and computing linear regression of external variables provided a wealth of results. This section details this process through the analysis section, while actual outcome of this data transformation, mining, and interpretation will be reported later.

The final section in the KDD process included the presentation of the previous steps. The results were interpreted in order to find the information that was nontrivial and novel. Finally, the knowledge gleaned from this research comprised two parts: visualization and presentation. This visualization differs from the exploratory visual analysis conducted in the analysis section of the KDD process, as it exists to convey the significant findings. This will be discussed in the conclusion. This study was designed to utilize only open source and publicly available information, rather than using internal NPS datasets or previously assessed information. In addition, a variety of free and licensed for analysis were also necessary for making the process and results widely accessible.

## **Data**

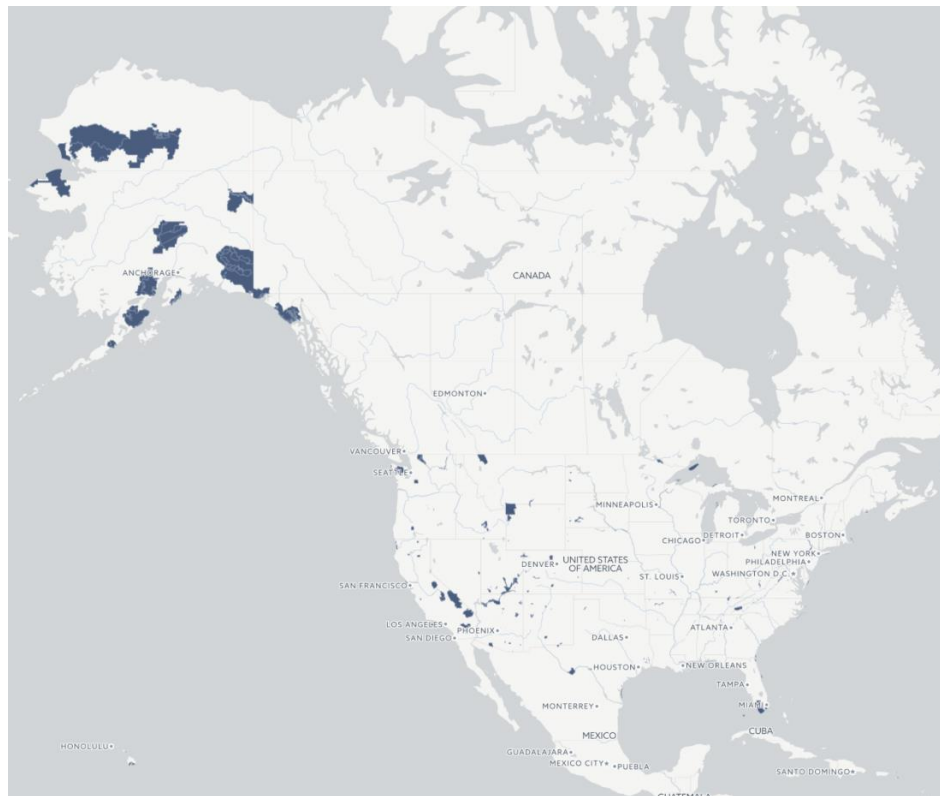
The National Park Service hosts a wealth of data on its Integrated Resource Management Applications (IRMA) portal. NPS also hosts the NPS Visitor Use Statistics Portal through IRMA, which allows for searching annual reports, as well as full or user-

defined datasets. Other primary government sources were queried or referenced to find additional datasets.

## Sources

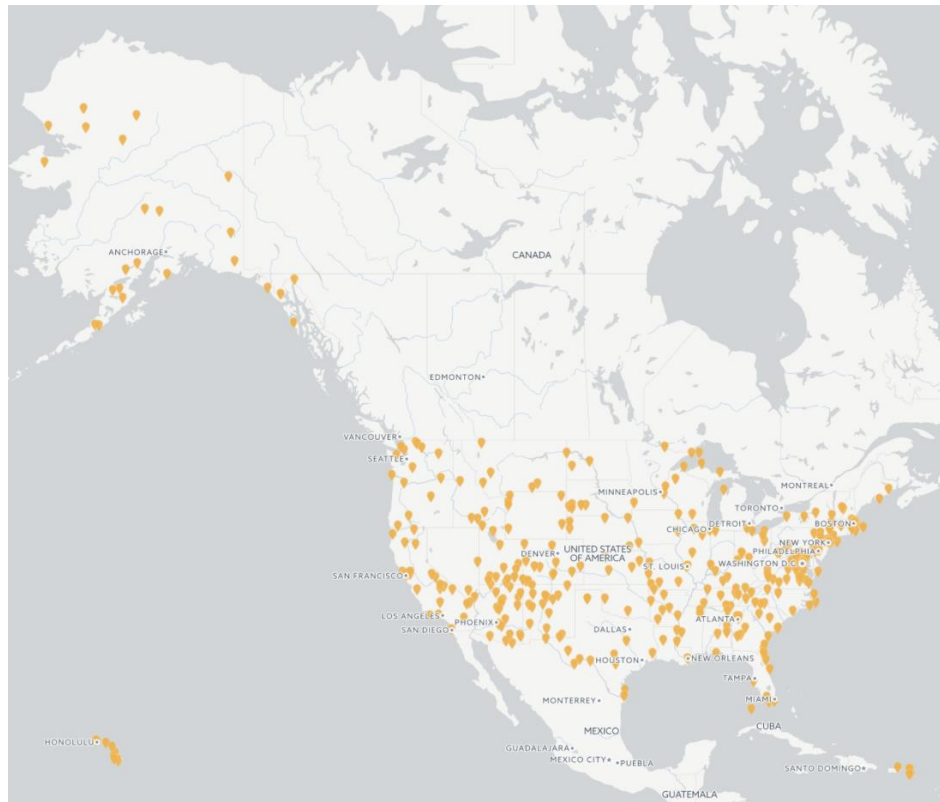
### *National Park Service Data*

Spatial datasets were available in multiple places, but the most updated source was the NPS Data Store. This service provided access to the layers created by NPS and included a polygon layer of the park and system unit boundaries. The NPS boundary layer contains polygons outlining the extent of each national park. This information was displayed in CARTO, an online source for web-based GIS, as shown in Figure 7.



**Figure 7 NPS Boundary Layer in CARTO**

There is a separate point layer indicating the centroids of each of those polygons, which was on CARTO or through the Data.gov website, a repository for federally-collected data open for public use (USG 2017). Calculating the centroids of the polygon layer within CARTO yielded the same results, so the centroid layer available on CARTO was used. The centroid layer was included, as seen in Figure 8, because smaller parks were easy to overlook using the NPS boundaries.



**Figure 8 Center points of all NPS System Units**



The annual visitor data from all park system units from 1979-2016 was downloaded. Data prior to 1979 was considered “historical” data and was available to query, but was not used because the categories for visits were not well-defined prior to this point. This data not only includes the total number of recreation visitors, but also defines over a dozen other visitor categories including those visiting the parks for overnight stays, camping, and research (NPS IRMA 2017). These were downloaded to include the park name and four-letter unit code. The four-letter code was used to join this table with the annual visitation, and again with the polygon and point layers. An excerpt of the data, shown in Figure 9, highlights the various menus and user-defined input options.

Select Year(s)	<input type="text" value="2016, 2015, 2014, 2013, 2012, 2011"/>	Select Month(s)	<input type="text" value="January, February, March, April, May"/>	<input type="button" value="View Report"/>
Select Region(s)	<input type="text" value="Alaska Region, Intermountain Region"/>	Select State(s)	<input type="text" value="AK, AL, AR, AS, AZ, CA, CO, CT, DE"/>	
Select Park Type(s)	<input type="text" value="International Historic Site, National Historic Site"/>	Select Park(s)	<input type="text" value="Abraham Lincoln Birthplace NHP, Adams National Historical Park"/>	
Select Field Name(s)	<input type="text" value="Recreation Visits, NonRecreation Visits"/>	Select Additional Field(s)	<input type="text" value="Unit Code"/>	
Annual Summary Only	<input checked="" type="radio"/> True <input type="radio"/> False			

### Figure 9 NPS IRMA Query Builder

Other NPS data proved more difficult to obtain. For example, park entrance fee information was only available on the website for each park, or from a blog post by NPS from 2011. This data was viewed in the context of an economic factor possibly affecting

visitation. No information was available detailing how many of those visitors counted in the annual report used the various passes available. However, the presence of a park entrance fee on its own was considered as a possible variable to determine whether it impacted visitation in any way. One attempt to include this information merely listed whether the parks had a fee or not. NPS listed the parks allowing entrance fees to be waived on certain days throughout the year. This was used to derive a listing of parks that normally charge fees.

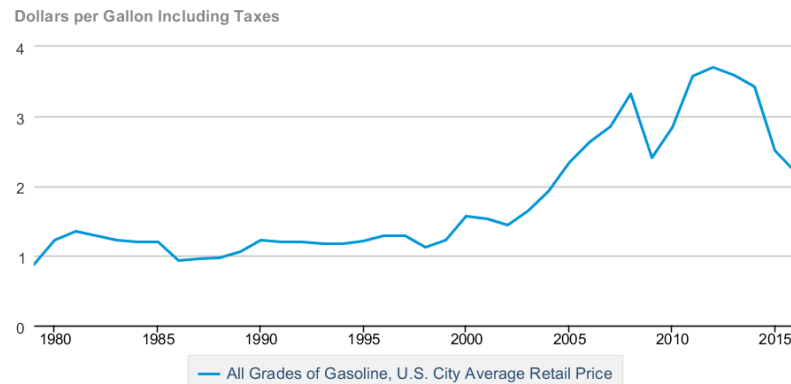
#### *External Variables*


Other variables outside of NPS-created data required additional collection and processing, to test how park visitation was affected by the types of external factors that typically also impact other tourist attractions and leisure travel. Although some previous reports and research established correlations between one variable and visitation, none tested a combination of external factors. Therefore, the most important test of external variables was not so much attempting to determine whether park visitation increased or decreased due to the change in one factor, but rather how a combination of factors impacted the parks annually. The exact type of data used in previous studies was either unavailable for the U.S. annually throughout the research period, or was not specified. Therefore, there were some challenges to collecting this information across the country, and some assumptions made after additional research.

The dataset used to estimate the average gas price nationwide every year since 1979 was based on information collected by the Department of Energy (DOE). DOE's Office of Energy Efficiency and Renewable Energy listed the average annual gas prices collected from an older version of the EIA website (DOE Energy.gov 2017). This data

was published in March of 2016 but only contained information through 2015. The U.S. Energy Information Association (EIA) provided an interactive query for downloading different types of gas price information. table provided a field containing the average retail gas price for all grades of leaded and unleaded gas based on the cities reported. These cities were dispersed throughout the country, and since it provided the only consistent measurement for the entire time, this data was used. The current retail gas price was adjusted using the U.S. Department of Commerce’s calculations for Consumer Price Index (CPI) as part of the data cleaning and processing step, based on the data illustrated in Figure 10.

**Table 9.4 Retail Motor Gasoline and On-Highway Diesel Fuel Prices**



 Source: U.S. Energy Information Administration

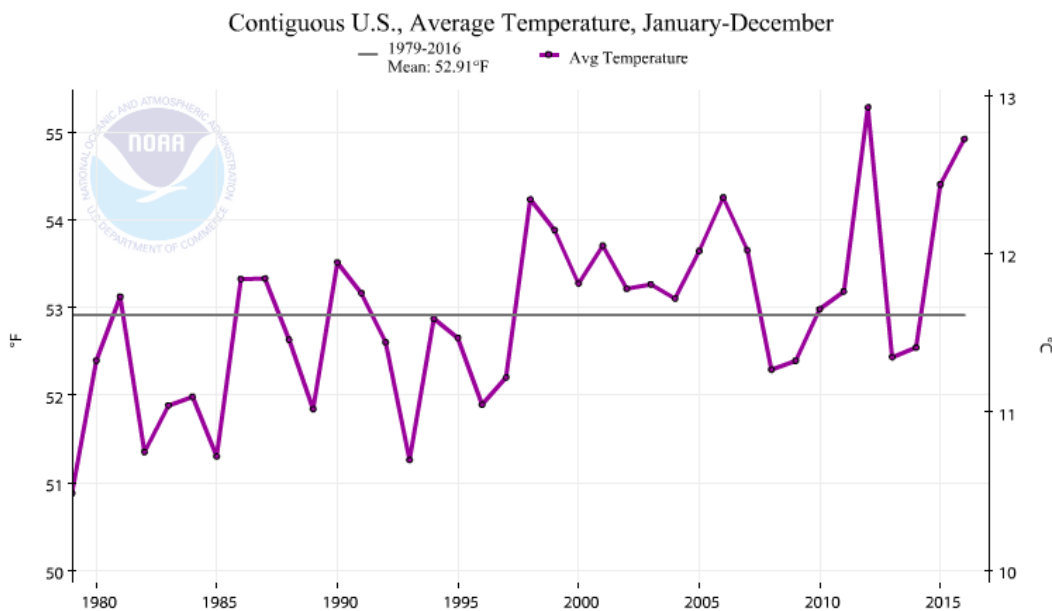
**Figure 10 EIA Annual Average Gas Price not adjusted for inflation**

The final economic indicator was the overall unemployment rate, retrieved from the U.S. Department of Labor’s Bureau of Labor and Statistics (BLS). Datasets were

available to query for several considerations, such as the types of labor, age groups, or race. The overall “Labor Force Statistics from the Current Population Survey” series provided the unemployment rate on a monthly basis, which would have to be processed before use. Two data types exist, one that reports the unemployment rate with a seasonal adjustment, and one without (BLS 2001). This adjustment attempts to limit the effect of a predictable seasonal employment trends, and are most commonly referenced (BLS 2001). Although an option exists to add a field for the average annual unemployment rate to the seasonally adjusted unemployment series selected, the functions was broken. The largest difference between the sets was less than 8/100, or 0.08 percent, therefore the seasonally adjusted data was collected.

Given Fisichelli et al.’s research into the effects of climate on visitation, a national-level assessment of visitation based on weather would not provide as much detail. In many cases, historical weather data was either too detailed, as much of the state and national averages are calculated based on inputs from thousands of individual weather stations, or already processed and analyzed. In the latter case, products and visualizations were available, but the underlying data was not accessible. Climatology for the entire United States between 1895 to 2010 containing average monthly temperature data for over 100 years across the country was considered, but there was no way to create an output shapefile from the data. The National Centers for Environmental Information (NCEI), part of the National Oceanic and Atmospheric Administration (NOAA), displays an interactive “Climate at a Glance Map” of statewide average temperature in a specific month and year, compared to the mean temperature calculated from 1901 to 2000

(NOAA 2017). Again, this data was too detailed and could not be summarized on a national scale. The NCEI also allows web users to compare two maps of the US on their “U.S. Climate at a Glance” page (NOAA NCEI 2017). There was a table of data including the same information on an annual basis, including the temperature anomalies on average for the entire year, also available a chart illustrated in Figure 11.



**Figure 11 NCEI Average Annual Temperature**

The Park Service compares visitation based on a park’s proximity to or within areas of different population density. This suffers from the same challenge as the rest of their statistical data, in that comparisons are only made based on the previous year’s information. Instead of using the MSA designation, this research instead compared visitation based on a different measure of population based on a point layer of U.S. cities.

The dataset, created by the U.S. Geological Survey (USGS) in 2014, contained the population of cities based on the 2010 decennial census, collected by the U.S. Census Bureau. Although not accounting for population changes between the 1970 census and today, nor available on an annual basis, these cities provided a minimum capacity for estimating proximity between parks and people. Most reports do not include whether park visitation has changed at a similar rate to the size of national population. Due to the limitations of NPS visitor surveys, understanding the nationalities of the people visiting the parks is limited. This provided a frame of reference for a change in population, although it was not intended to provide any correlation to visits. The overall U.S. population dataset, estimated based on mid-year population from the World Bank, was hosted through the Economic Research division of the U.S. Federal Reserve Bank of St. Louis (World Bank 2017).

### **Cleaning**

The cleaning and preprocessing of the downloaded tables was mostly performed in Excel, although CARTO or other methods would have also been acceptable. There were 12,628 rows of records detailing visits to parks system units operating between 1979 and 2016. With nine visit types, plus the unit code and date attributes, the raw visit records totaled 138,908. Because the NPS IRMA Portal allowed for creating custom queries, there was no superfluous or empty fields in that data. However, the NPS Park Boundary and Park centroids GeoJSON layers contained additional information irrelevant to this research. For example, the boundary layer contained two fields that only contained null values, a field with notes detailing changes made to the polygons, and

dates that the fields were created and updated. In the centroids layer, there were also two fields that contained entirely null values, as well as a field for an ID that was duplicative considering CARTO creates its own ID field. These fields were removed from the dataset.

Both CARTO and Tableau allow for queries and displays based on temporal attributes of the data. However, the “year” field in the NPS and external data was often formatted as a string, text, or number. For the best functionality, these fields were changed to a date field type. In Excel, this required an arbitrary month and day field to be concatenated with the existing year field in a new column. Each year was formatted as a date using the first of January as the month and day to simplify the process.

Between the three main NPS data tables, each followed a different naming convention for the park names. In the polygon and point layer, the first part of the full park name was separated from the park type, separating “San Juan Island” as the name from “National Historical Park” as the type. In the park centroids layer, the same park also contains a field for the full unit name of “San Juan Island National Historical Park.” The visitor statistics from IRMA have the same park name, under the field “park,” listed as “San Juan Island NHP.” Rather than have duplicative information, the visitor statistics information was processed to remove lengthy full park names. Therefore, the unit code was a more reliable method of joining tables for park data, while the full name was joined to each spreadsheet before any visualizations were made for consistency.

Processing the park fees list required importing the text data into a table, deleting the state name, and joining the unit names to the park names. The new table was created

using the park names, codes, regions, and states from the visitor statistics and was called “Park Metadata.” From there, the parks with fees were imported into a new table using a rule. If the park code existed, then the attribute “Fee” in the metadata table was a Boolean value of yes, or true. Otherwise, the value was no, or False. In addition, the values for park type, region, and state were removed from the visitor statistics table and added to the Metadata table. The newly created table, shown in Table 2, was used so only the unit code and visitor statistics remained in their original table.

**Table 2 Excerpt of Park Metadata Table**

Park	UnitCode	ParkType	Region	State	Fee
Abraham Lincoln Birthplace NHP	ABLI	National Historical Park	Southeast	KY	N
Acadia NP	ACAD	National Park	Northeast	ME	Y
Adams NHP	ADAM	National Historical Park	Northeast	MA	Y
African Burial Ground NM	AFBG	National Monument	Northeast	NY	N
Agate Fossil Beds NM	AGFO	National Monument	Midwest	NE	N
Alibates Flint Quarries NM	ALFL	National Monument	Intermountain	TX	N
Allegheny Portage Railroad NHS	ALPO	National Historic Site	Northeast	PA	N
Amistad NRA	AMIS	National Recreation Area	Intermountain	TX	N
Andersonville NHS	ANDE	National Historic Site	Southeast	GA	N
Andrew Johnson NHS	ANJO	National Historic Site	Southeast	TN	N

The EIA retail annual gas price, averaged across all grades and types of gasoline except for diesel, was initially reported in original U.S. dollars. Standardizing the dataset required a step for calculating the price in 2016 USD. The Bureau of Labor and Statistics (BLS), hosts a tool on their website for converting prices based on the Consumer Price Index (CPI). The calculator has the option of comparing data on a monthly scale (BLS 2017). Given that most NPS park visits occur in the summer, the price adjustment for both input and output was compared using the month of June.



The unemployment rate information was processed to calculate the average annual unemployment rate, based on each month's individual rate. Then, the monthly rates were removed from the table. The temperature data was similar, requiring only the necessary values were removed. The average annual temperature was kept, to be used in the regression model and for visualization, as it was easier to understand. The field detailing the extent of the temperature anomaly already indicated the amount of abnormality in the weather attribute, which proved useful in a binary type of comparison. This field ensured that no additional calculations were necessary, as it already indicated the amount above or below "normal," and was easier to use in a different type of analysis. The final table of external variables, illustrated in Table 3, indicates the values to be tested based on the strength of their correlation to park visitation.

**Table 3 External Variables**

<b>Year</b>	<b>AvgTemp</b>	<b>TempAnomaly</b>	<b>Unemployment</b>	<b>GasPrice</b>	<b>ParkNum</b>	<b>Population</b>
<b>1979</b>	50.88	-1.14	5.85	2.93	269	225055000
<b>1980</b>	52.39	0.37	7.18	3.56	273	227225000
<b>1981</b>	53.12	1.1	7.62	3.6	278	229466000
<b>1982</b>	51.35	-0.67	9.71	3.18	289	231664000
<b>1983</b>	51.88	-0.14	9.6	2.98	293	233792000
<b>1984</b>	51.98	-0.04	7.51	2.79	297	235825000
<b>1985</b>	51.3	-0.72	7.19	2.69	304	237924000
<b>1986</b>	53.32	1.3	7	2.05	305	240133000
<b>1987</b>	53.33	1.31	6.18	2.04	306	242289000
<b>1988</b>	52.63	0.61	5.49	1.97	310	244499000
<b>1989</b>	51.84	-0.18	5.26	2.06	315	246819000
<b>1990</b>	53.51	1.49	5.62	2.26	317	249623000
<b>1991</b>	53.16	1.14	6.85	2.13	320	252981000
<b>1992</b>	52.6	0.58	7.49	2.05	325	256514000
<b>1993</b>	51.26	-0.76	6.91	1.96	328	259919000
<b>1994</b>	52.87	0.85	6.1	1.91	328	263126000
<b>1995</b>	52.65	0.63	5.59	1.91	329	266278000
<b>1996</b>	51.89	-0.13	5.41	1.98	330	269394000
<b>1997</b>	52.2	0.18	4.94	1.94	337	272657000

1998	54.23	2.21	4.5	1.66	343	275854000
1999	53.88	1.86	4.22	1.77	342	279040000
2000	53.27	1.25	3.97	2.19	345	282162411
2001	53.7	1.68	4.74	2.07	346	284968955
2002	53.21	1.19	5.78	1.93	350	287625193
2003	53.26	1.24	5.99	2.15	354	290107933
2004	53.1	1.08	5.54	2.44	357	292805298
2005	53.64	1.62	5.08	2.9	357	295516599
2006	54.25	2.23	4.61	3.14	360	298379912
2007	53.65	1.63	4.62	3.3	361	301231207
2008	52.29	0.27	5.8	3.65	361	304093966
2009	52.39	0.37	9.28	2.68	361	306771529
2010	52.98	0.96	9.61	3.14	364	309348193
2011	53.18	1.16	8.93	3.82	368	311663358
2012	55.28	3.26	8.08	3.89	368	313998379
2013	52.43	0.41	7.37	3.7	369	316204908
2014	52.54	0.52	6.17	3.47	369	318563456
2015	54.4	2.38	5.26	2.54	369	320896618
2016	54.91	2.89	4.85	2.20	369	323127513

After the processing, five main tables remained, a combination of downloaded, cleaned, and transferred attributes within each. Table 4 below indicates the fields in each table, with an asterisk indicating the attributes calculated or created, such as a properly formatted “date” field.

**Table 4 Data after Preprocessing**

Table Name	Fields (Attributes)	Data Type
<b>NPS Boundary</b>	the_geom	Geometry - multipolygon
	the_geom_webmercator	Geometry
	unit_code	Text / String
<b>NPS Centroids</b>	the_geom	Geometry- point
	the_geom_webmercator	Geometry
	unit_code	Text / String
<b>Visits</b>	Unit Code	Text / String
	Year	Number
	Date*	Date
	Recreation Visits	Number
	Non-Recreation Visits	Number
	Concessioner Lodging	Number

	Concessioner Camping	Number
	Tent Campers	Number
	RV Campers	Number
	Backcountry Campers	Number
	Non-Recreation Overnight Stays	Number
	Misc. Overnight Stays	Number
<b>Park Metadata</b>	Park (park name)	Text / String
	Unit Code	Text / String
	Park Type	Text / String
	Region	Text / String
	State	Text / String
	Park Fee	Boolean
<b>External Variables</b>	Year	Number
	Average Temperature	Number
	Temperature Anomaly	Number
	Unemployment Rate	Number
	Gas Price	Number
	Number of Parks	Number
	Population	Number

## Integration

There were a few considerations prior to integrating the data. Although an external database management system was considered, CARTO was used because it offered additional visualization functionality. However, it caps cached data at 250MB was a limitation in some queries. Unlike a traditional database management system (DBMS), CARTO does not require a schema to be built. However, to best understand the interaction between datasets and the capacity for them to be joined, a relational database was sketched, illustrated in Figure 12.

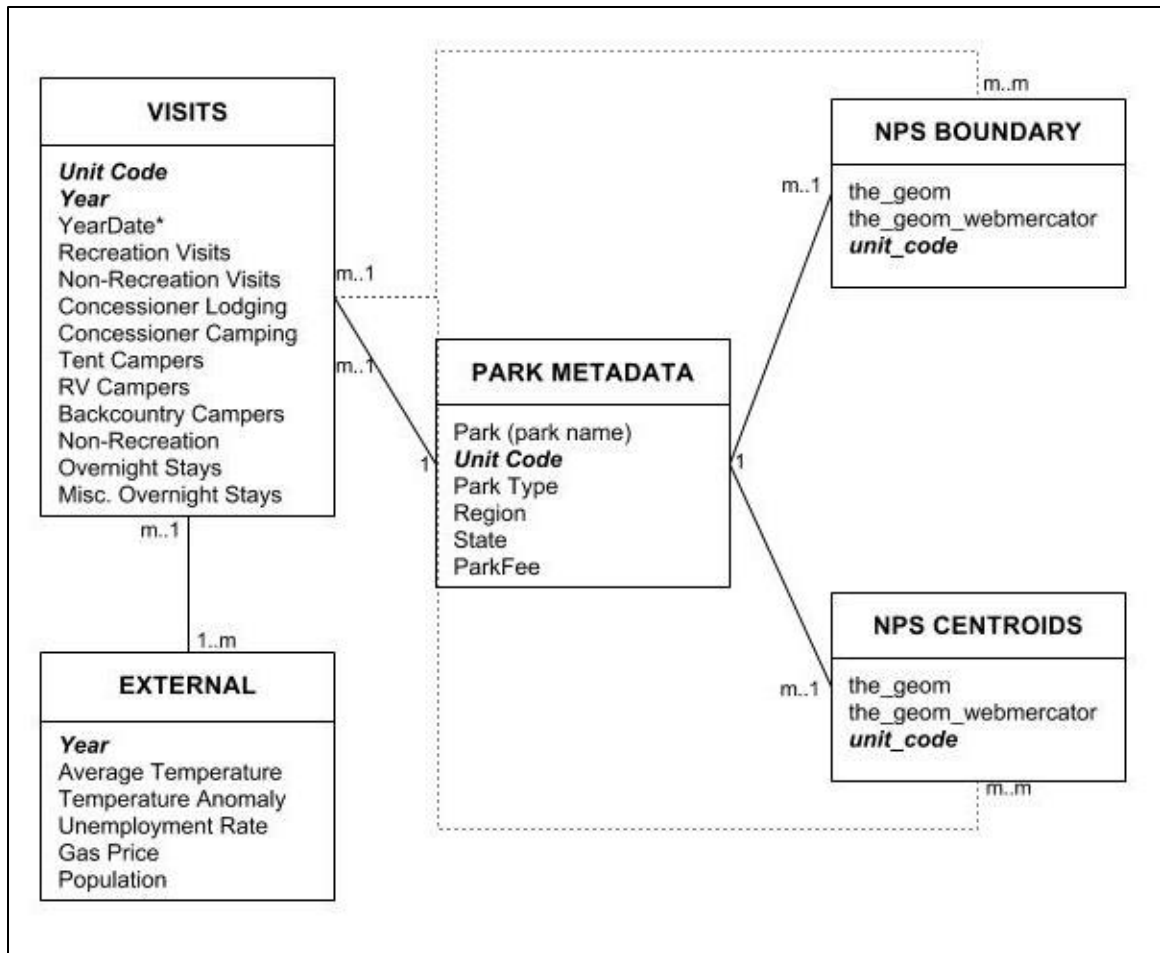


Figure 12 Database Design showing relationships between tables

Ensuring each of the NPS data layers was formatted to contain the park unit code allowed for the ability to create connections between information. Although not used in most of the calculation and analysis, a master file was created to include all the park visitor statistics, linking the geospatial point and polygon data with the visitor statistics for every park over the 38-year period. In order to efficiently process the basic information, it was decided that using only the visitor statistics would suffice for most queries. Then, the point and polygon information could be joined and query through

CARTO or Tableau for other analysis. In addition, the joined tables, containing only the relevant information, could be exported again for visualization and to identify outliers. In addition, the joined tables, containing only the relevant information, could be exported again for visualization and to identify outliers.

### **Selection**

Depending on the purpose, most of the park visitor statistics were selected by either year, unit code, or visit type, while additional information, either from the spatial datasets or the metadata, was linked as needed to prevent duplication of attributes. Selecting data for analysis differed based on the purpose of the analysis and the tool used. In CARTO, data was selected based on the required attributes. The most frequently joined tables were the visitor statistics to the point and polygon layer. In addition to CARTO, Tableau provided other tools for visualizing spatial and temporal aspects of the visitor statistics. Selecting data in Tableau required importing the necessary tables, then joining them together using a built-in function, rather than conducting a SQL query. For example, choosing fields to view or access from one or more tables was as simple as dragging the desired attributes into a workspace and manipulating the display. With the updated datetime functions in the table, selecting and analyzing different statistics and locations over the 38-year period was much easier.

The dataset of U.S. cities was processed early to limit its size when querying. Because this study was more concerned with displaying cities with a large population, rather than conducting analysis based on the actual population size, only the relevant information was kept. For this research, a large city was defined as anything with a

population of over five hundred thousand people as of the 2010 Census. This value was queried in CARTO, and the results were exported to a new table. The original table was removed to save space. To measure the distance between each park and the closest large city, the information from the two tables had to be joined. The “cross lateral join” function and spatial operator “<->” to compare distance between the centroids of each park in the “park\_boundary\_centroids” table and the “large\_cities” table. This functioned as a nearest neighbor search between the centers of the bounding boxes of the index automatically created by CARTO (Boundless Geo 2011). The distance was output into a new field for distance, as indicated in Figure 13.

```
SELECT bc.*, large_cities.city_name, large_cities.pop_2010, large_cities.cartodb_id as cities_cartodb_id,
ST_Distance(geography(large_cities.the_geom), geography(bc.the_geom)) AS distance
FROM nps_boundary_centroids bc
CROSS JOIN LATERAL
  (SELECT cartodb_id, the_geom, city_name, pop_2010
   FROM large_cities
   ORDER BY bc.the_geom_webmercator <-> the_geom_webmercator
   LIMIT 1) AS large_cities
```

**Figure 13 Nearest Neighbor Join**

Only one park was linked to incorrect nearest city, the War in the Pacific Memorial in Guam. The distances between this park and both San Diego and San Francisco are within one mile. The way the tables were joined calculated the centroid distance using a bounding box index. Either of these two challenges, the two similar values or the bounding box index, contributed to this error. Rather than manually update the table, the unit code was removed from spatial visualizations.

## **Analysis**

The analysis component of the KDD process created for this research contains three main steps: data transformation, data mining, and interpretation. As most authors agree when discussing KDD, the process requires iteration. One of the methods used to interpret transformed and mined data was by exploratory visual analysis. Therefore, the interpretation of results occurred in this research throughout this process, as certain patterns were detected after an initial review. This spurred additional data integration and selection. A few assumptions were made after an initial review of the data after its cleaning and processing.

Because there was no accessible and complete source for park opening dates, the visitor statistics information contains some flaws. For parks created or opening after 1979, there were null values in reporting for those years in sum of all visits and recreation visitors. Averaging the number of visitors across parks throughout the reporting period required that only the years without null values were included. For example, if a park was not open, the number of visits in each visit type field were initially blank when the data was downloaded. In Tableau and CARTO, those values were automatically changed to null. If a park had a certain number of visits during one year, then none the following year, then that zero value was still included. This change in visitors was typically either due to a temporary park closure, the transfer of a park outside of NPS management, or the designating a new unit code.

Regardless of the specific date and month a park opened, the first-year reporting visitor statistics was analyzed as a full year of visitation instead of attempting to prorate the number of visitors based on the months a park was open. The gas prices and

temperature data were national averages extrapolated from specific cities or states. Unemployment and annual population were based on measures of U.S. citizens, even though there are parks in U.S. territories outside of the contiguous United States and international tourist visitors to the parks. While all of those have some impact on park visitation, this research sought to better illuminate historical trends and spatial characteristics of the data.

### **Transformation**

The bulk of data transformation was conducted in Excel and Tableau, as these tools offered a simple way to add fields to each table with newly-created fields based on the aggregation of existing attributes. Excel allowed for quick calculations to be made for important data summaries and to understand the basic features of the visit statistics, such as sums and averages over time. Tableau offered the ability to easily visualize the spatial component of the data after calculating or displaying data after being imported from Excel. Using the “Visits” table, the transformation of visitor statistics included a calculation to create a sum of all visitors, as they were initially dispersed based on the type of visit. In addition, the difference of visitors at each park each was calculated, along with the percentage difference. The percentage difference in visitors was calculated in many of the statistical abstracts, but not tracked and reported for the previous timeframe. Other transformations included averaging the number of visitors by visit type over the course of 38 years. These calculations were displayed in a pivot table, with each one field containing all park unit codes and each year as its own field, depicted in Figure 14.



UnitCode	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
ABLI	-32157	13171	66462	-62581	1897	6882	15715	12205	-4501	21499	-2154
ACAD	107657	140916	366922	531666	-435741	27434	181707	370208	224130	937461	-3091882
ADAM	-9548	2081	-1615	722	-2764	1157	-2862	7084	-7675	6997	-4797
AFBG	0	0	0	0	0	0	0	0	0	0	0
AGFO	9664	1200	-1872	4392	-1264	37	-1650	-660	229	2237	3023
ALFL	289	173	14	-479	332	-51	2	292	-354	1702	-424
ALPO	15756	-8392	-5580	-4829	4803	9688	-534	25110	29954	9796	-1535
AMIS	-192705	-98005	294692	-38411	-79709	-29845	64150	29146	58150	39327	-99824
ANDE	-24423	-13003	-9644	6868	-9892	17411	11358	316	-202	6156	-4662
ANIA	0	0	0	0	0	0	0	0	0	2235	-16
ANJO	17917	8509	3478	41832	-34521	-16282	3993	4495	-529	-5216	-333
ANTI	32845	115792	-57168	22469	94358	57208	122642	-95500	-475984	11219	72008
APCO	10393	26565	-5086	5872	71750	-40414	46171	12291	-23382	64747	25507

Figure 14 Excerpt of Pivot Table for Visits by Year and Unit Code

Based on previous reports, certain visit types displayed similar trends. Rather than compare each visit type independently, certain groups were transformed by aggregating similar visit types. Creating pivot tables in Excel and visualizations in Tableau from the aggregated values provided insight into the dataset, as well as illustrating certain trends that required additional analysis. Measuring the variance and standard deviation in the external variables provided a baseline for what values could be considered “normal,” which allowed for exploration of the visitation based on outlying years. The overall of visits by visit type was summarized, resulting in totals for all visit types for the time period, displayed in Table 5.

Table 5 Total Visits By Type 1979-2016

Recreation Visits	Non-Recreation Visits	Concessioner Lodging	Concessioner Camping	Tent Campers
10,324,521,705	4,765,451,150	135,027,904	37,828,463	136,039,174
RV Campers	Backcountry Campers	Non-Recreation Overnight Stays	Misc. Overnight Stays	All Visits
115,428,819	74,556,117	13,703,609	91,703,783	15,694,260,724

## **Data Mining**

After transforming the data and ensuring its compatibility with the analysis tools, the next step in the KDD process was data mining. The purpose was to discover previously unknown patterns using some of the common data mining techniques, such as regression and clustering. Other tools exist for machine learning and computation data mining, but these techniques were not applied for this research. Rather than use a computational solution, outliers were detected based on visualization of the data. This section chronicles those main analysis techniques.

### *Regression*

One of the most common methods for data mining is regression. Using the open source software, R, linear regression was calculated to assess the strength of the linear relationship between the external variables and park visitation. If a variable shows a strong correlation to park visitation, it could be further investigated, and possibly later used to provide a better forecast of visitation. In addition, the linear regression of the park visitation based on the year was also conducted to determine the temporal aspect to visitation. These trends attempted to answer whether park visitation was generally increasing or decreasing each year.

### *Clustering*

Another data mining technique, clustering, provided a method for grouping similar objects based on a specific attribute. Rather than only seeing spatial when visualizing results, clustering data grouped parks based on similarities on number of park visitors or their distance from a city. This technique highlighted other similarities between parks, particularly those without any distinguishing features, as well as a generalized approach to assessing almost four hundred system units. Tableau leverages

min-max normalized k-means clustering that groups data based on variance. It automatically calculates the number of clusters, although this can be changed if necessary. This provided a method for finding similarities in the data based on components that were not spatial in nature.

### **Interpretation**

The tools for interpretation and visualization included Excel, CARTO, and Tableau. The iterative nature of KDD was reinforced during this step, as visualization efforts were assessed to determine possible outliers and trends in the data. Charts and plots depicting transformed data was reviewed to visualize overall trends and perform exploratory visual analysis. Detecting outliers was easier using a visual representation. By conducting clustering and regression, measurements of groups and correlation revealed certain characteristics. The results from the analysis methods, such as evidence of correlation or measures of percentage change, comprised the results from this research. Explaining the cause of the results, however, was part of the final KDD step.

### **Knowledge**

Presenting new knowledge learned from the data, the final step of the KDD process, leverages previous research and newly-discovered patterns to discern possible reasons for the correlations and trends discovered. This was important for identifying potential causes for trends, to determine whether the hypotheses were proven true, and provide a new beginning for future research. This research focused on attempting to explain the most recent record-breaking visitation in the context of historical visitation on a national scale. Using data visualization tools to reveal patterns beyond year-to-year

comparisons and highlight visitation changes since 1979 indicated limitations in NPS statistical reporting. To better plan for the Parks in the future, adopting some of these techniques and leveraging the existing wealth of data may provide an additional resource.

## CHAPTER FOUR

### Results & Interpretation

With the KDD process as a guide, interpreting data after cleaning and analysis yielded a wealth of results. Beginning at the data transformation step, visual analysis was used to identify certain trends or patterns. Results of visualization required an iterative approach before accepting or recognizing a specific result. No previous reporting established whether the most recent increase in park visitation was like any historical fluctuations in visitation. Plotting all park visitation based on the year from 1979 through 2016 illustrated a steadily increasing number of visitors each year, with few exceptions. Additional iterations of calculation, visual exploration, and data mining were conducted to discern additional patterns, beyond the overall trend illustrated in Figure 15.

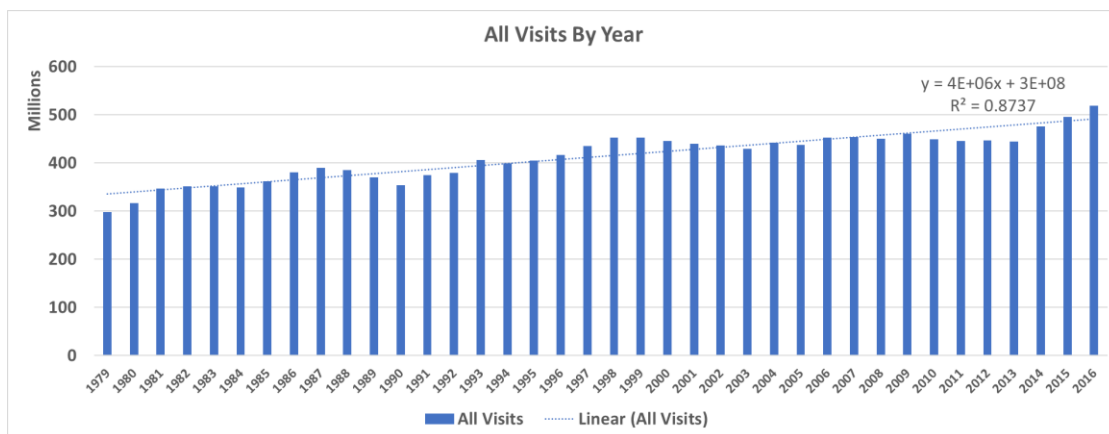


Figure 15 All Visits by Year

## **NPS Visitation**

Summarizing the visit types across both unit code and year, then comparing them to the summation provided by NPS illustrated that calculations in both Excel and Tableau were accurate. For example, park visitation overall showed a steady increase over time. NPS reports 417 system units, but grouping parks by unit code returned only 372 unique values. This was the same value in their annual summary data report, and NPS acknowledges that they do not collect, or report, visits from certain parks that are jointly administered. In addition, some unit codes are used for parks that have multiple locations and unit types. For example, a park can be both a “National Park” and a “National Park and Preserve.” This was one drawback from using only unit codes to compare data, instead of the full park name, especially when comparing visitation by unit type.

Querying the database resulted in some basic measures of the park system. Of these 12,628 records for each system unit and year, there were only 35 occurrences when no recreation visits were reported for a given park and year. From 1979 through 2016, there were over fifteen billion visits of all types to all parks still reporting visitation. Two-thirds of these were recreation visits, displayed in charts to identify any obvious patterns. In 1979 there were 269 reporting system units by unit code, which had grown to 369 by 2016. These parks reported at least one year with a recreation visit value greater than zero. Other visit types proved to not only be significantly less popular, but were also unavailable at many locations. For example, only 30 parks have concessioner camping facilities, while 96 parks reported RV campers, illustrated in Table 6.

**Table 6 Parks by Visit Type**

<b>Visit Type</b>	<b>Number of Parks</b>	<b>Count of Records</b>
<b>Recreation Visits</b>	372	12539
<b>Non-Recreation Visits</b>	235	7020
<b>Concessioner Lodging</b>	51	1666
<b>Concessioner Camping</b>	30	668
<b>Tent Campers</b>	107	3531
<b>RV Campers</b>	95	3127
<b>Backcountry Campers</b>	118	3734
<b>Non-Recreation Overnight Stays</b>	64	750
<b>Misc. Overnight Stays</b>	139	3251

Two park system units had stopped reporting to NPS prior to 2016: The John F. Kennedy Center for Performing Arts and the Oklahoma City National Memorial. Multiple parks are administered under unit codes NACA, NCPC, and NCPE as part of the National Capital Parks region. Previous designations included National Capital Park Area, National Capital Parks - Central, and just National Capital Parks, with unit codes NACA and NCPC. The NACA unit code was used until 1996, when both NCPC and NCPE started being calculated separately. However, the combined values of those two codes was equivalent to the previous NACA designation. Neither NCPC or NCPE were used in the GIS data from the park service, however, which proved an added complication. Therefore, spatial representations of the data for visits to unit codes NCPC and NCPE starting in 1997 were assigned to the NACA code.

The largest year-over-year increase in visitation occurred in from 1980 to 1981, with an increase in visits of 9.49 percent over the previous year. Between 1989 and 1990, NPS recorded a four-and-a-half percent decrease in visitors. The worst year-to-year

decline was still only almost half of the amount of increase. Averaging the number of visitors across parks throughout the reporting period required that only the years without null values were being included. For example, if a park was not open, the number of visits in each visit type field were initially blank when the data was downloaded. If a park had a certain number of visits for one year, then none the following year, then that zero value was still included. Aggregating visit types and comparing the change in visitors through the extent of the historical data, shown in Table 7. The percentage increase of visits to parks by visit type from 1979 to 2016 mirrored some of the changes in NPS accounting and reporting. For example, as the number of parks increased, so did the number of visits. In addition, NPS has changed and updated methods for estimating visitation, particularly what counts as a non-recreation visit, that differed from data collected in 1979.

**Table 7 Percent Increase since 1979**

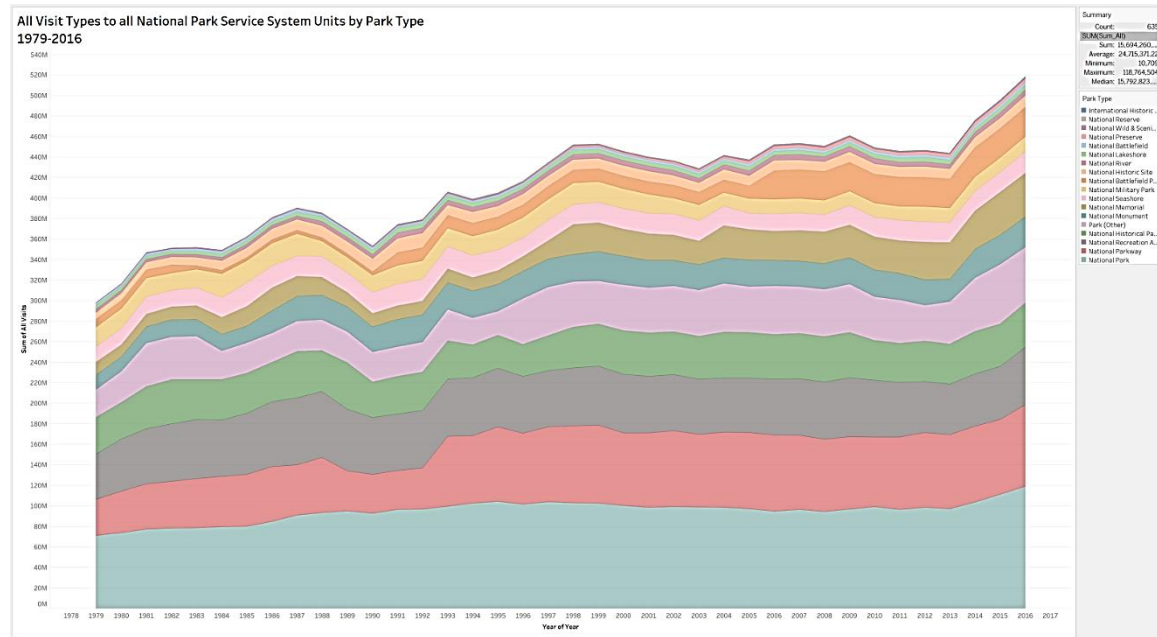
<b>Recreation Visits</b>	<b>Non-Recreation Visits</b>	<b>Concessioner Lodging</b>	<b>Concessioner Camping</b>	<b>Tent Campers</b>
261%	324%	204%	254%	213%
<b>RV Campers</b>	<b>Backcountry Campers</b>	<b>Non-Recreation Overnight Stays</b>	<b>Misc. Overnight Stays</b>	<b>All Visit Types</b>
157%	190%	127%	287%	274%

### *Exploratory Visual Analysis*

Attempting to discern patterns and other information by reading all the park number values in table was tedious, but were much more obvious when presented on a chart. Visualizing the information highlighted areas that required further study, or revealed some sort of error. For example, a gap in visits for a system unit was discovered



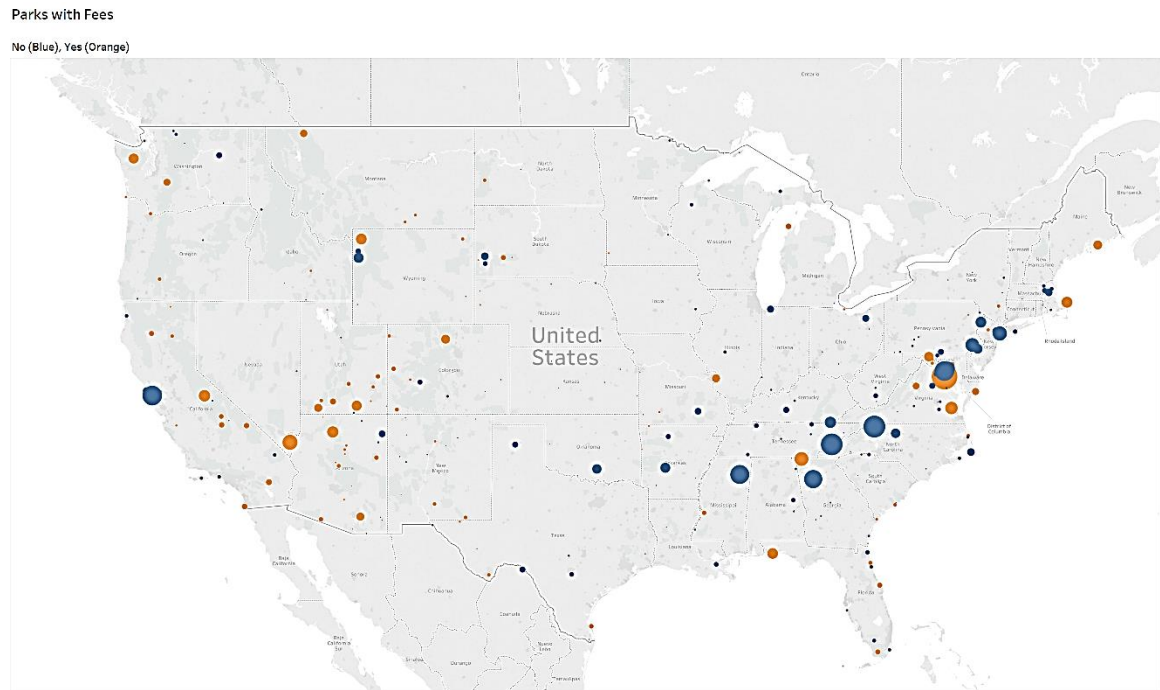
through the visual analysis of park visitation. The sum of annual visitation revealed certain patterns, illustrated in Figure 16, such as changing visits by park type.



**Figure 16 Visits by Park Type 1979 - 2016**

Displaying parks on a map with markers sized to various attributes, such as the number of visitors or the distance to a city provided a way to identifying patterns spatially. Rather than just sorting through tables and querying for information about park fees, for example, the information could be presented in a more user-friendly method. In addition, one of the assessments and investigations lacking in NPS reporting is the limited ability to view how certain characteristics of parks are distributed through the entire park system. Instead of using just the region or state as a method for aggregating information, highlighting the parks that normally charge fees by color, and sizing the

marker quantify the number visits to each, a trend emerges. As illustrated in Figure 17, from this display, the parks charging fees generally have fewer visitors.



**Figure 17 Parks by Number of Visits, color denotes park fee**

Ranking the parks provided another method for comparing the change over time, without quantifying the amount of difference between the parks. It was much easier to determine if any parks were consistently visited than comparing raw numbers, or even differences between years. System units were ranked for their overall visitor attendance each year. The only system unit to be among the top five most-visited units since 1979 was Blue Ridge Parkway (BLRI). There were ten distinct parks within the top five ranking, but counting the highest number of years resulted in the following parks: Golden

Gate National Recreation Area (GOGA), Great Smoky Mountains National Park (GRSM), and National Capital Parks – East (NCPE). However, comparing the average ranking revealed a different list, with the George Washington Memorial Parkway (GWMP) and Kennesaw Mountain National Battlefield Park (KEMO) consistently ranked in the first two spots, respectively, since they reached a high enough level of visitors. Previous visualizations revealed a large spike in visits counted at the GWMP in 1993, due to a change in park visitor collection and estimation. Removing the parkways resulted in a new list of high-ranking parks. The most interesting result, however, was that Chickamauga and Chattanooga National Military Park (CHCH) was in the top five from 1979 to 1999, apart from 1991. In that year, KEMO joined the group and has remained there consistently. Identifying these parks on map indicated these parks were within one hundred miles from one another, but the KEMO was much closer to the city of Atlanta. While it cannot be proven that one park contributed to a decline in visitation at another park, this process highlighted these unique scenarios for future exploration.

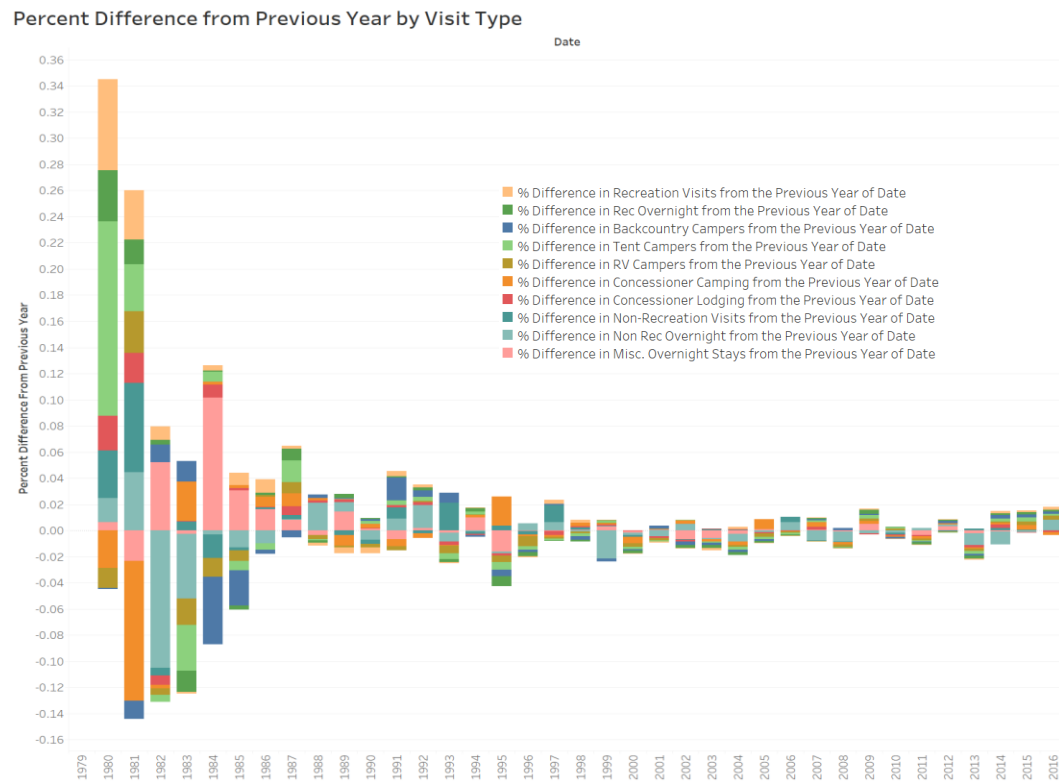
Assessing the least-visited parks was more challenging, because these parks experience changes on a difference scale than the larger system units. Unlike the most-visited parks, there were no parks with the lowest ranking of visitors since 1979. Of the parks whose rankings decreased, most experienced a decline for one to five years before climbing in visitation. The two most recent parks, tied for lowest visitation in 2016, had previously not even been in the lowest ten percent. These were the Clara Barton (CLBA) and the Pennsylvania Avenue National Historic Site (PAAV). While CLBA has been closed for renovations, there was no indication explaining why the visitation decreased at

PAAV, since it is part of the larger Washington D.C. National Mall area and has no defined boundaries. The best way to estimate the overall least-visited parks in the historical data was to calculate their individual ranking of visitors each year, then find the median of those values. Using this method, the top five parks had a median ranking between two and six, illustrated in Table 8.

Table 8 Least-visited Parks	
PARK	MEDIAN
ANIAKCHAK NM & PRES	2.00
PORT CHICAGO NAVAL MAGAZINE NM	4.00
ALIBATES FLINT QUARRIES NM	4.00
EUGENE O'NEILL NHS	5.00
SALT RIVER BAY NHP & ECOLOGICAL PRES	6.00

Another method to understand park visitation over time was to create a measure for assessing consistency. To do so, the amount of change each year had to be considered. Parks could then be compared by how much, or little, their level of visitors changed from year to year. The smallest and largest park visitation values for each park were identified, and the span between them became the range of possible visits. Large positive numbers highlighted parks that experienced major increases, whereas a negative range illustrated parks with decreasing attendance. A symmetrical range indicated parks with relatively consistent visitor patterns. The easier way to display this was by comparing parks to the percent increase or decrease in visitation from their previous

records, then depicting these changes over time illustrated an interesting temporal pattern. While actual visitation numbers for each park type may vary widely, the amount of that change is becoming more consistent each year, as illustrated in Figure 18.

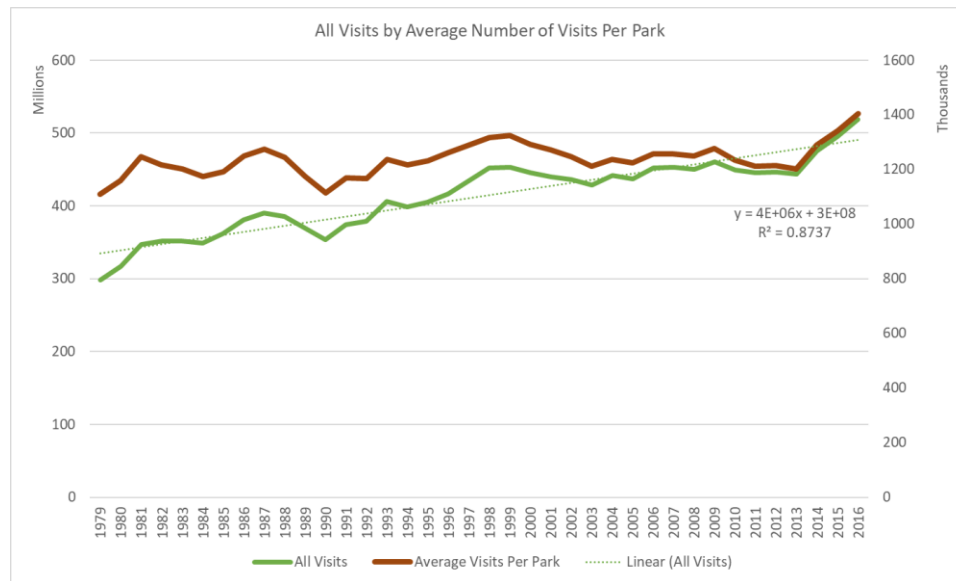


**Figure 18 Annual Difference by Visit Type**

Fewer than fifty system units exhibited large percent differences between 2016 and their first reporting year. This measure was useful for eliminating the overall park visitation number itself, which skewed most visual representations to accommodate high-visitation parks. For example, system units like George Washington Memorial Parkway, Golden Gate National Recreation Area and Great Smoky Mountains National Park, with

annual visitation in the tens or hundreds of millions, overwhelmed the densely displayed smaller system units. These and other parks were relatively unlikely to have large changes in visitation between years. Highlighting parks with the greatest percent visitor increase was one way to discern whether these parks contributed to the overall park trend.

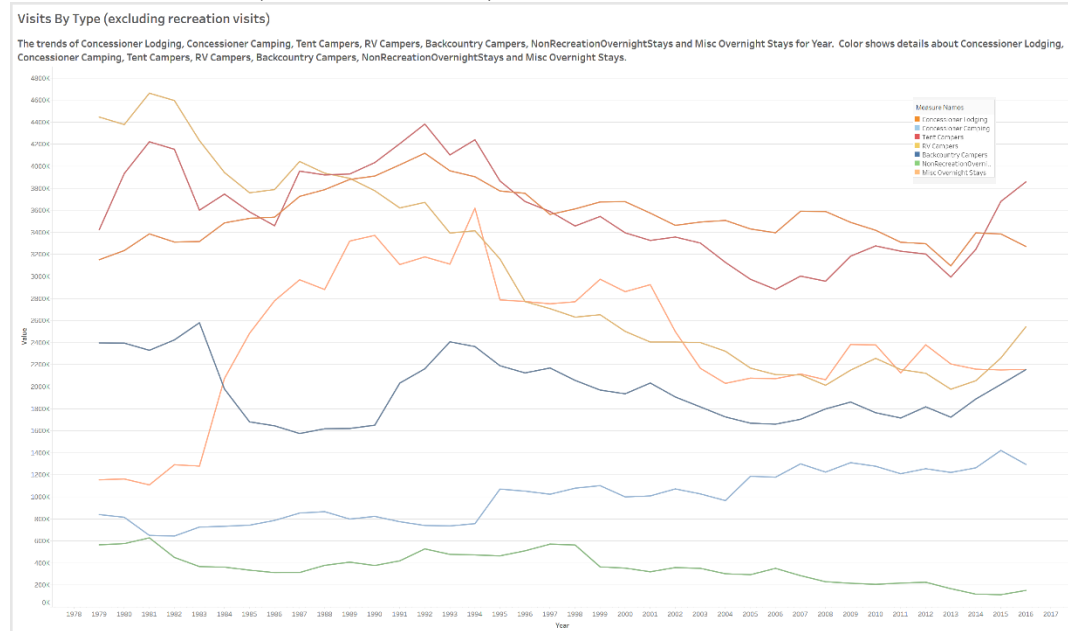
Investigating the change in visitors required searching for specific park information, although it was typically either due to a temporary park closure or the re-designation of a unit code. This illustrated the iterative nature of KDD, in that creating a visualization allowed for human recognition of patterns, errors, or other aspects of the dataset which could then be calculated or reviewed to glean insight. The average number of visits per reporting system unit each year was calculated to determine whether the increased number of parks was related to the overall number of visitors. The average number of visits per reporting unit increased by three hundred thousand, 1.1 million average visits per system unit to 1.4 million, from 1979 to 2016. This chart, in Figure 19, illustrated that an increase in system units alone does not explain the overall number of visitors.



**Figure 19 Visits Averaged by Number of System Units**

Because recreation visits comprise the bulk of visitation, however, other trends are overlooked. NPS does track other park types, but most of their and news agency reporting only highlights overall and recreation visits. Most other park visits trends exhibited an overall decline, apart from miscellaneous overnight stays and concessioner lodging. Recent reports highlighted the sharp increase in tent and backcountry camper visits, but neither of those surpassed their previous record high levels. These trends, illustrated in Figure 20, illustrated much more fluctuation than the overall recreation and non-recreation visits.

**Table 9 Park Visitation (other than Recreation)**



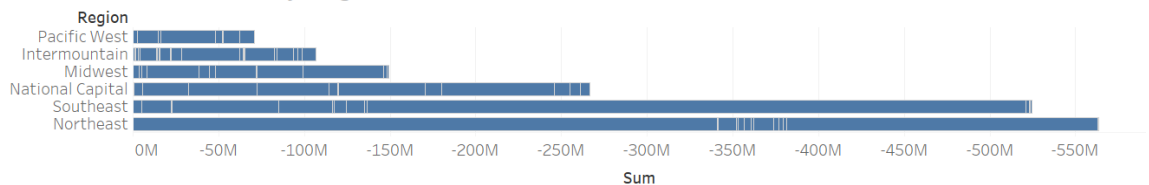
**Figure 20 All Other Visit Types**

Given the recent popularity of the parks, one main research goal was to determine if any parks experienced a decrease in visitors over their history. The difference between the sum of all visit types was calculated for each year and park. Only forty-two parks displayed a net decrease in number of visitors based on the annual differences. This list did not include those park units that no longer report their statistics, as discussed above. The minimum and maximum values were then compared to create a range of visits. This measure provided an estimation to determine whether most park visitation stays the same each year, and to discover whether parks with a wider range of visitation were related to the most-visited parks. Only sixteen parks experienced a net decrease on the order of twenty-thousand visitors or more. The park losing the most visitors since 1979 was Chickamauga and Chattanooga National Military Park (CHCH) in Georgia, whose total



loss was 380 million. Parks with the greatest decreases were in the Northeast and Southeast regions. This result was unexpected, as there tends to be a higher concentration of parks and large cities in those regions. The region with the smallest net loss of visitors was the Pacific region, illustrated in Figure 21.

**Net Decrease in Visits by Region**

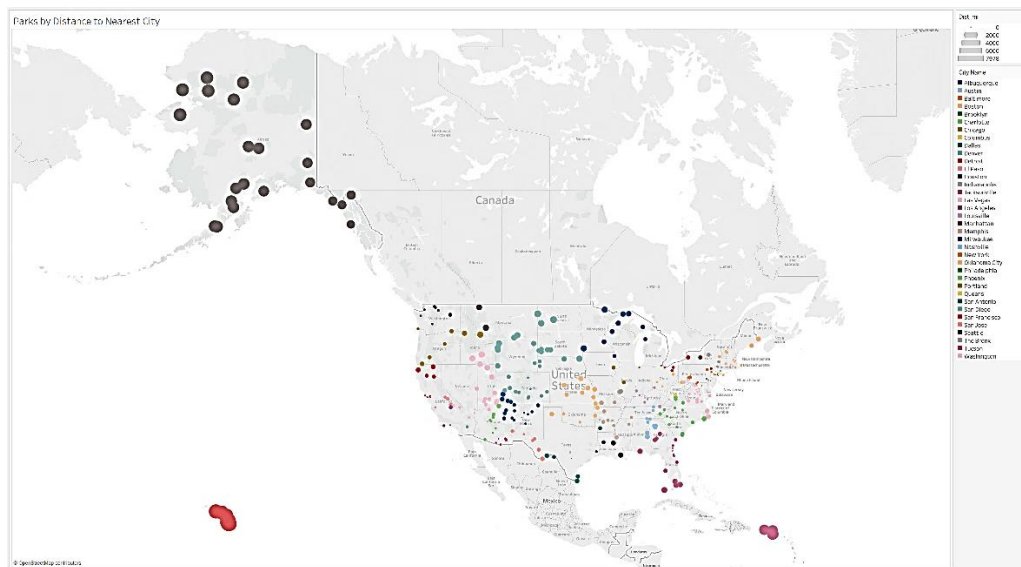


**Figure 21 Parks with Net Decrease**

Visualizing the parks based on their distance to the nearest large city highlighted some limitations. This information was skewed because only U.S. cities were used as a comparison. Lacking an understanding of visitor demographics, a widely known issue within NPS, also made the measure somewhat arbitrary. The average distance of all parks to a large city was 272 miles, while the median distance was even closer at 136 miles. Depending on availability of transit and traffic conditions, this meant that more than half of parks were less than an estimated three hours by car of a large city.

Grouping the system units by the name of the city itself presented some challenges, which in one case included parks in a range between less than half of a mile to six thousand miles. For example, all parks in the Alaska region were grouped together, with some other parks, because their closest large city was Seattle. Within that subset was

one of the least-visited parks, Aniakchak National Memorial and Preservation, along with Denali National Park, which is much more popular. Parks outside of the contiguous United States also tended to skew the distances. However, most parks were in much closer proximity to a large city. Another method for assessing these groupings was required to better understand the impact of distance of visitation.



**Figure 22 Closest Large City to Each Park (color)**

### *Clustering*

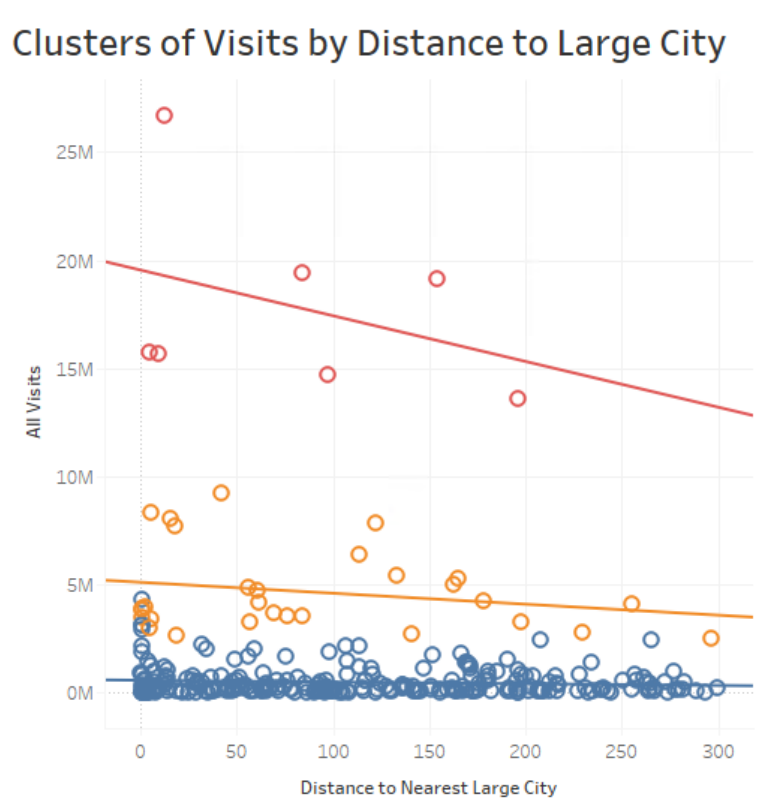
Rather than solely grouping visitor statistics based on their unit name or region, clustering parks based on original or calculated attributes allowed the ability to determine other associations between system units. It was also useful for aggregating many smaller system units and their attributes, as these were often harder to identify in other plots and even displayed within a park boundary. Clustering the parks based on their total number

of historical visitors, however, provided one way to examine parks that were much more alike in visitation. This helped determine groupings of parks by their popularity, which might not be obvious when presented in a different format.

This first attempt at clustering result in one large cluster comprised the parks within 850 miles of a large city, while another consisted of the parks much further away. The smallest cluster, within only seven members, contained the parks located very closely to a large city, but with very low annual average visitation. If like the other system units in the other clusters, these parks should have exhibited much higher visitation. These units presented an opportunity to investigate why distance to a populated does not always equate to more visitors, and possibly identify ways to improve upon this. It could also highlight other factors to be included in a later model for predicting park visitation.

Clustering was also done to aggregate parks based on their distance to the largest city and the annual visitation. This first set was calculated based on parks within the contiguous United States. Grouping all parks resulted in one cluster of system units within one thousand miles of a large city, while the second cluster contained a distance range between that and almost five thousand miles. The first cluster contained almost all the parks, 338, compared to only 25 parks in the second cluster. The between-group sum of squares, the space separating clusters, was 3.27. The clusters were not very cohesive, with a within-group sum of squares value of 4.76. The total variance, using the total sum of squares of 8.03, was 0.41, which was likely because only two clusters were calculated. The average distance to a large city for units within the first cluster was just under 147

miles. By contrast, the second cluster's average distance was over 1914 miles from a large city. This explained the high F-statistics value for distance, because it was a much more important factor in defining clusters than park visitation. Clustering parks based on the number of visits, for the 301 parks within three hundred miles of a large city revealed that visits decrease based on longer distance, Figure 23.



**Figure 23 Parks Clustered by Number of Visits and Distance to Large City**

Another set of clusters was calculated to determine the natural grouping of the cities themselves. For each city, the average distance to all parks for which it was the

nearest neighbor was calculated, as well as the number of visits associated with those parks. This returned four clusters, with a between-group Sum of Squares value of 2.69. The first cluster contained cities whose nearest parks had small average distances but low visitation. The second was comprised of higher visitation, but even closer average distance of approximately 126 miles. The average distance for the remaining two parks was widely divergent, with the third cluster containing parks almost one thousand miles away, while the fourth contained parks within less than five miles.

Clustering parks based on all the visit types individual aggregations per system unit highlighted something interesting as well. Only ten parks comprised the second of two clusters. Most of the visit types were insignificant in the calculation of the clusters. Based on the number of parks reporting these statistics, it seemed most likely that concessioner lodging and RV campers would have more of an impact on the cluster composition. When including recreation visits, the clusters included some parks that were not among the most-visited, likely because of the influence of other visit types. The ten parks in the cluster were all still located in the contiguous United States, but included some units such as Yellowstone, Yosemite and Olympic National Parks. These parks, while not having the highest visitation, do report a wider range of visit type. Both RV and backcountry campers had a large f-statistics, which mirrored the intuition regarding the impact of these less-popular visit types. Three visit types, non-recreation visits and non-recreation overnight stays, and miscellaneous overnight stays, were the only ones without a p-value of 0. However, the non-recreation visits were the only category not considered significant for creating these clusters, shown in Figure 24.

Analysis of Variance:

Variable	F-statistic	p-value	Model		Error	
			Sum of Squares	DF	Sum of Squares	DF
Sum of Rv Campers	162.6	0.0	1.692	1	3.767	362
Sum of Backcountry Campers	161.0	0.0	0.9447	1	2.124	362
Sum of Concessionerlodging	153.7	0.0	0.8308	1	1.957	362
Sum of Recreation Visits	132.3	0.0	1.316	1	3.6	362
Sum of Concessionercamping	131.3	0.0	0.604	1	1.665	362
Sum of Tent Campers	122.9	0.0	0.5868	1	1.728	362
Sum of Misc Overnight Stays	46.82	3.341e-11	0.1328	1	1.027	362
Sum of Non Recreation Overnight Stays	36.3	4.166e-09	0.1045	1	1.042	362
Sum of Non Recreation Visits	5.581	0.01868	0.03937	1	2.554	362

**Figure 24 Clusters by Visit Type**

Removing recreation visits as a variable for another set of clusters improved the proportion of variance within clusters, although only six parks comprised the second of two clusters. However, these six parks were also in the original second cluster as well, so there was not much change. Backcountry campers and concessioner lodging had the largest F-statistic measures in the model, and p-values of 0.0, indicating that these variables had a significant impact on the nature of the clusters. Removing the recreation visits did change the system units contained within the clusters, primarily ignoring the influence of non-recreation visits again, illustrated in Figure 25.

Analysis of Variance:

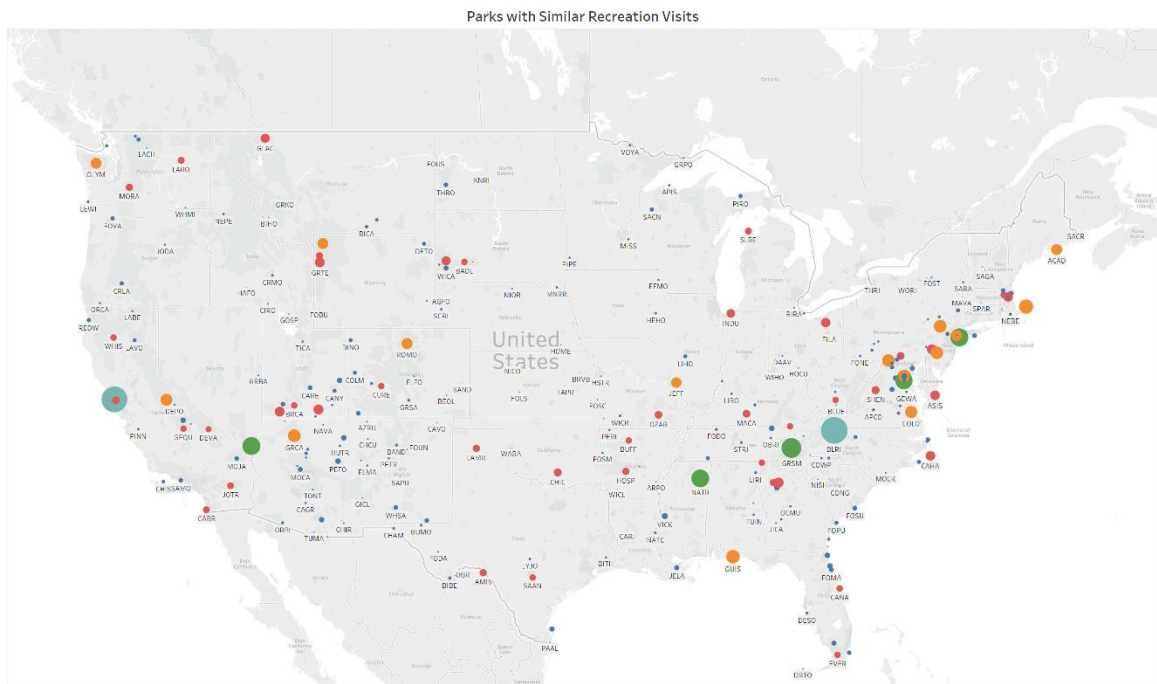
Variable	F-statistic	p-value	Model		Error	
			Sum of Squares	DF	Sum of Squares	DF
Sum of Backcountry Campers	214.0	0.0	1.256	1	2.124	362
Sum of Concessionerlodging	190.2	0.0	1.028	1	1.957	362
Sum of Rv Campers	184.1	0.0	1.915	1	3.767	362
Sum of Concessionercamping	146.5	0.0	0.6739	1	1.665	362
Sum of Tent Campers	139.9	0.0	0.6679	1	1.728	362
Sum of Misc Overnight Stays	75.71	1.11e-16	0.2147	1	1.027	362
Sum of Non Recreation Overnight Stays	60.19	8.86e-14	0.1733	1	1.042	362
Sum of Non Recreation Visits	4.948	0.02674	0.0349	1	2.554	362

**Figure 25 Cluster Description (not including recreation visits)**

Further exploration revealed that only by comparing similar categories would the clusters calculate different relationships. For example, comparing both types of overnight stays resulted in four clusters, due to the variance in number of overnight stays between parks. However, three clusters accounted for a total of seven parks. Experimenting with these clusters at least illustrated one method to identify parks with similar visits by their less-popular visit types. Most times, however, they displayed as only two clusters.

Running the algorithm on only recreation visits, as every unit code reported this type of visit, resulted in a set of five clusters depending of the number of visits. Mapping these clusters revealed that not all parks with similar parks were located in the expected areas. In some areas, parks located closer together tend to have similar trends in visitation, particularly in the National Capital region due to the density of parks. However, these clusters highlighted some of the smaller parks in that area as being comparable to dozens of parks in the southwest United States. In the third cluster, colored in red on the map, were parks that are typically in the middle of the visit rankings. Three parks in Hawaii and one in the U.S. Virgin Islands belonged to this cluster, whereas all the parks in Alaska have fewer visits in cluster two. Finding similarities in parks from disparate

regions may not solve any problems for park visitors or NPS, but indicates connections across the dataset that would not have been obvious. Ultimately, analyzing and displaying visitor information and park type spatially was better achieved in other ways.



**Figure 26 Parks Grouped by Similar Number of Visits**

### *Geovisualization*

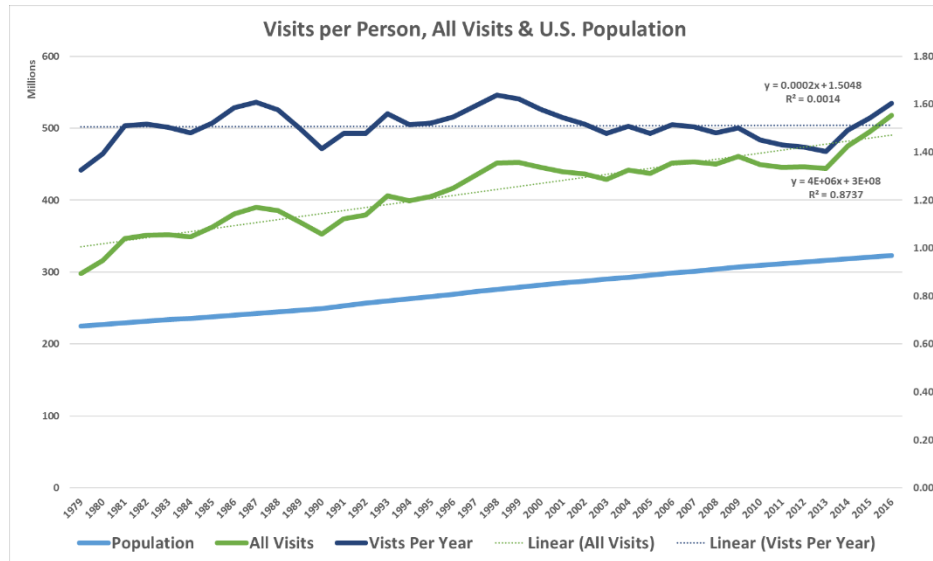
Visually analyzing the visitor statistics and clusters, even when integrated with spatial points, did not provide the same level of functionality as creating a web-based map to display results. Because CARTO functioned as the database for this research, accessing the necessary information required only a few queries. The nearest city to each park was previously calculated and added as a new attribute to the table detailing visits.



The layer for large cities was added to a new map and styled to highlight its population. Then, the layer containing park centroids was added and queried to join it the table containing park visits. The points were styled on a scale to approximate the distance from the large city. The final layer contained the polygon boundaries, which was styled as a choropleth map with colors signifying the number of visits in 2016. A copy of that layer with the visitation in 1979 was also added as a comparison. Additional styling and editing added functionality like information displays when clicking on the layers and an animation showing the parks based on the year they first reported visitors.

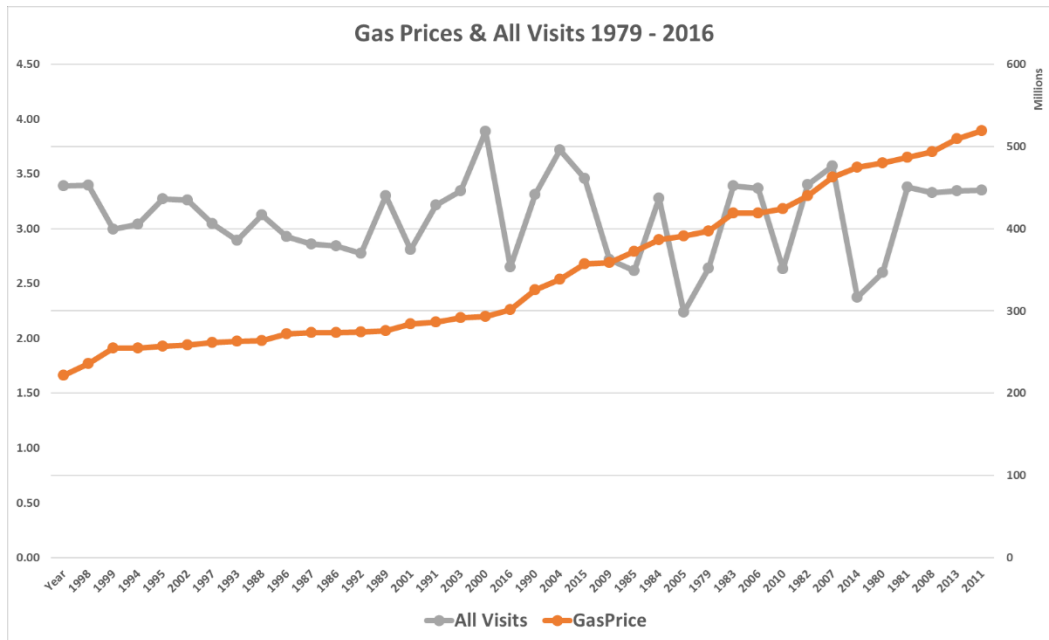
### **External Variables**

Two important metrics were lacking in the NPS reporting, a comparison of visitation to the overall population and to the number of system units. The positive trend of park visitation appears much less drastic, however, when plotted as a function in which the sum of all visitation is normalized based on the annual population measure. This measure highlighted that the most visits per population occurred in 1998 and 1999, for ratios of 1.64 and 1.62 visits per person. Although 2016 broke visitation records by sheer numbers, there were fewer visits per person in the U.S. population than the previous high. Another peak park year was 1987 at 1.61 visits per person. Each year was buffered but at least one year with a high rate of visitation. Using this measure, the park visitation increase is much subtler when plotted in Figure 27. Because the population increases very consistently each year, the number of visits to population displayed the same general shape. However, the amount of an increase in visitation between 1979 and 2016 was much less drastic, with large changes peaking in 1981, 1987, 1993, 1998, and 2016.

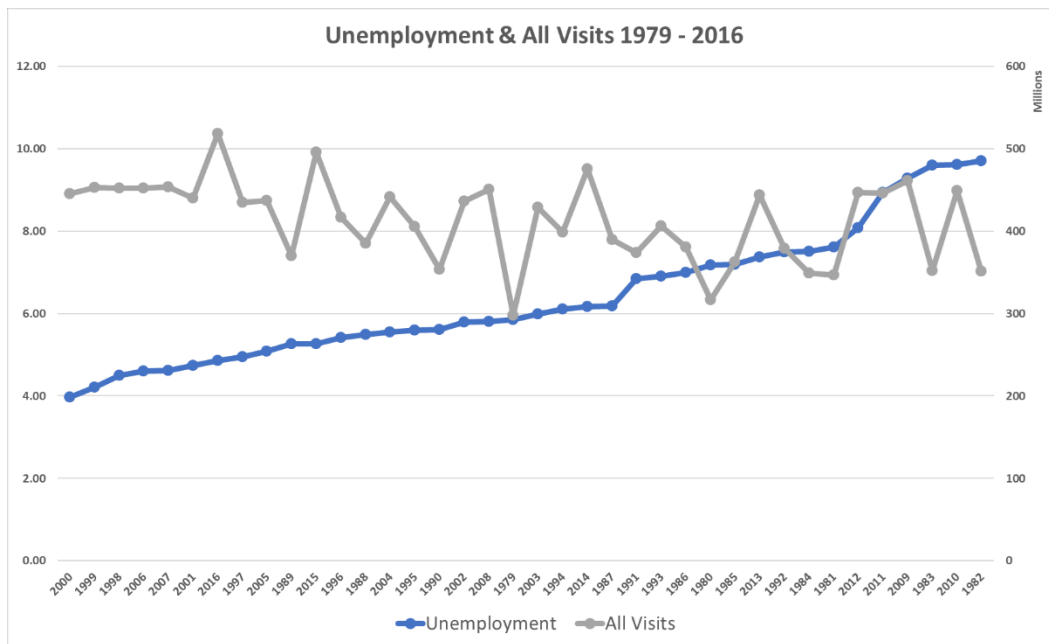


**Figure 27 Visits Normalized by Annual Population, Visits, and Annual Population**

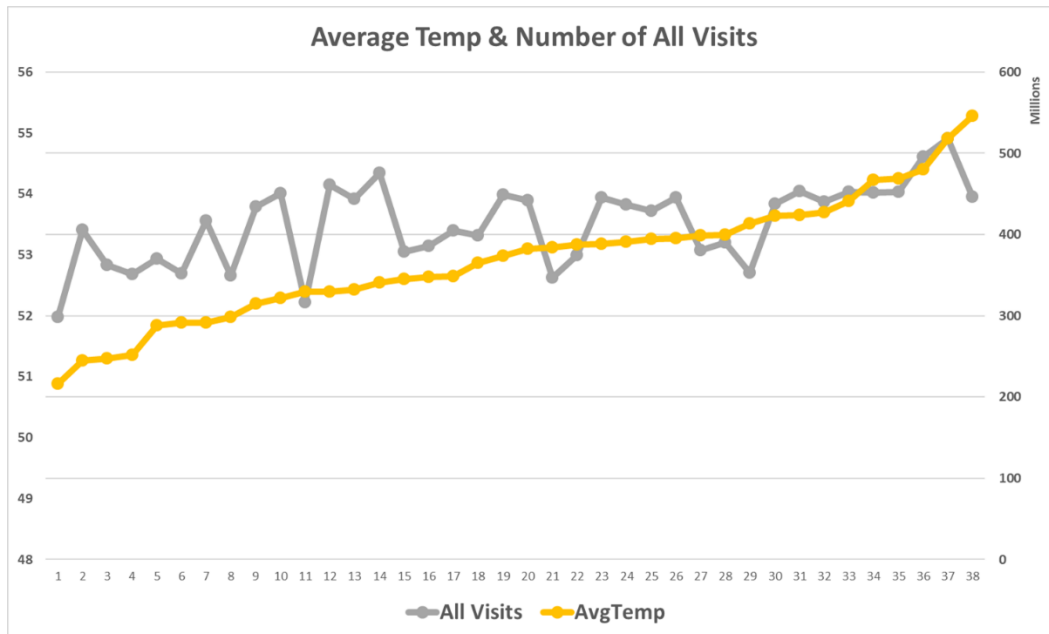
In addition to number of parks and population, tourism-related variables were tested as well. The effect of each individual variables was assumed to be directly or inversely related to visitation. This assumption allowed for the creation of various linear models accounting for a combination of those individual variables. While the relationship between the year, or even U.S. population, and park visitation was modeled as linear, actual visitation was much more cyclical. The following figures illustrate the level of park visitation based on the variable of an external variable such as gas prices, Figure 28, unemployment rate, Figure 29, and average temperature, Figure 30.



**Figure 28 Relationship between Gas Price and Visits**



**Figure 29 Relationship between Unemployment and Visits**

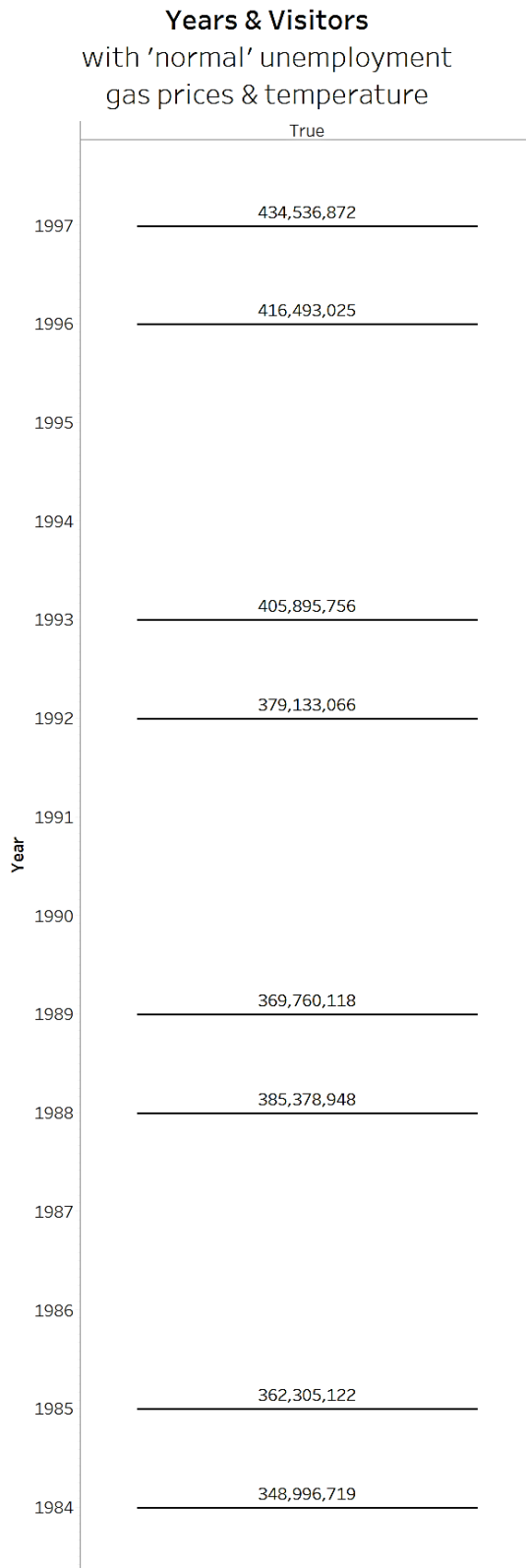


**Figure 30 Relationship between Average Temperature and Visits**

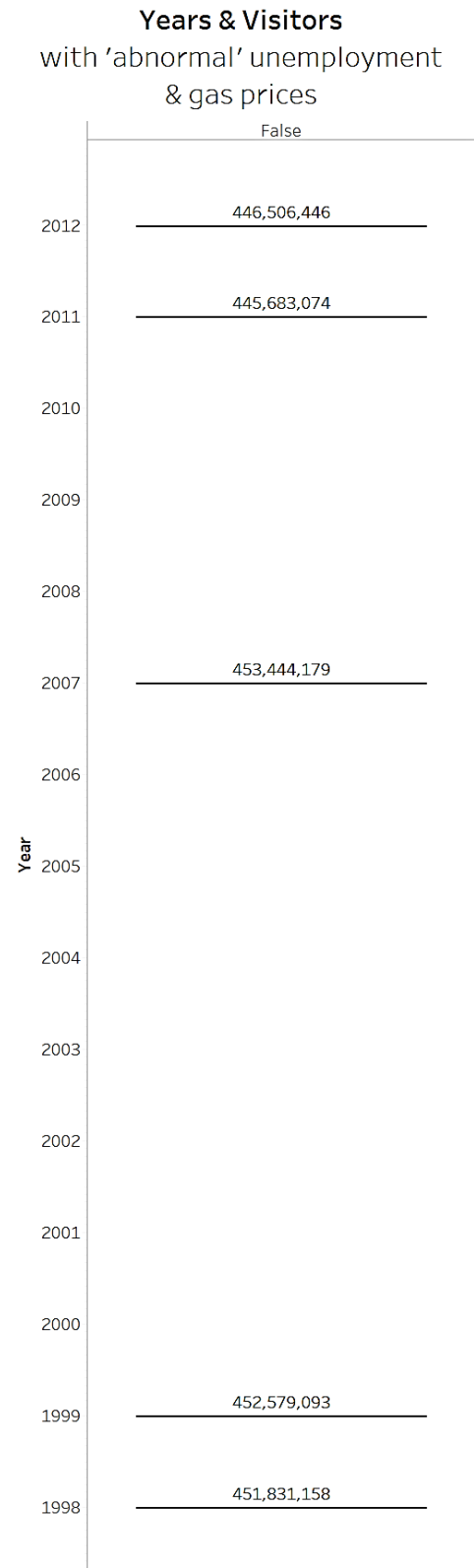
In an effort to simplify the external variables provide another method for testing desirable conditions for park visitation, each variable was categorized as either normal or abnormal as a new field for visualization. For example, average gas prices exhibited a wide range since 1979. It was useful to have a measure for determining which prices were within a one standard deviation of that population, while highlighting those values on either the very low or very high end of the data set. If a certain year had a normal gas price, then it would not be expected to have either very high or very low visitation. The expected result would be to see if any of these “normal” years did, in fact, record an unusual number of visits. The temperature anomaly field was already a measure of the level of abnormality, so initially any non-zero value was considered “abnormal.” However, each year displayed some degree of anomaly, even a small one. Therefore, the

average temperature value was used in the model. The standard deviations each field was calculated. Then a rule was created to label each year as either normal, which was displayed as “TRUE,” or not, displayed as “FALSE.” Therefore, for the same time period as visitation, the unemployment rate and the price of gas could be assessed on a simple metric, rather than based on value itself. Using a chart to visualize each of the external variables individually compared to park visitation at a minimum provided some additional context.

Locating all the years where the unemployment rate, gas prices and temperature were within a ‘normal’ range, then displaying the output illustrated the relative infrequency of those conditions. Eight years were within normal, only occurring sometime between 1984 and 1997. The visitation range in that time went from approximately 349 million to 435 million, shown in Figure 31. Comparing this to years in which these values all fell outside of their normal range returned only two results in 1998 and 2012. However, the number of visitors was five million higher in 1998. Both years recorded over 446 million visits. Eliminating the criteria for temperature in this scenario returned three additional years where the unemployment and gas prices were outside of their normal range, illustrated in Figure 32. However, while gas prices were higher normal, unemployment was technically abnormally low. This illustrated the limitation of assessing visitation annually using a proxy for normality instead of actual values. However, it was interesting to determine which years, within this time frame, were collocated with variables that could be considered normal.



**Figure 31** Visits under “normal” conditions



**Figure 32** Visits under “abnormal” conditions

### *Regression*

Multivariate linear regression of the economic factors, weather, and population resulted in a measure for assessing their effect of the sum of all visitor types. This provided a graphic display of the relationship, while regression returned the coefficient of determination, or r-squared, and calculated probability, p-value. Comparing all variables to the sum of visitors, as opposed to comparing them individually, provided a holistic overview of multiple factors influencing park visitation based on the tourism industry. Testing a model with only temperature, unemployment and gas price indicated the need for a temporal variable. Without year, population, or even park number in the equation, the model was a poor fit for estimating visits in a given year. Only average temperature was considered significant, but with an f-statistic value of 7.23, this model only account for 34 percent of the overall variance.

When using regression to construct a basic linear model of external variables effects on visitation, the first calculations included the visitor statistics from all parks. In certain situations, it was advantageous to only use the data from parks that had been in continuous operation since 1979. The second models included only parks open for the entire time period. While the relationship between the year, or even U.S. population, and park visitation was modeled as linear, actual visitation was much more cyclical. Using either the year or the population to predict future visitation would ideally yield similar results, although the year is a constant and the population would have to be estimated. Calculating the regression for each of these separately against park visitation resulted in the same coefficient of determination, or r-squared value. Essentially, 87 percent of the variation in park visitation from the past 38 years was associated with either year or

population. However, neither alone would be useful in predicting an increase or decrease in visitation, as the input values for this type of function would most likely continue to increase. The difficulty in obtaining demographic information about its visitors, and the fact that the park system units are open to international visitors, also indicate that population may not be the most reliable metric.

This research focused on other variables impacting visitation, adding them to a linear model to determine their association with the number of all visits. Both first models included the average temperature, unemployment, and gas price. One tested the impact of population and the other tested the year, finding them to be considered the most significant source of variation. This was not surprising based on their correlation without additional variables, but the overall adjust r-squared value also increased in these models, to over 89 percent. The models, residuals, and other values for models including year and population are shown in Figures 33 and 34, respectively.

```
> model_var_year <- lm(mydata$SumAllVisits~mydata$Year+mydata$AvgTemp +
> summary(model_var_year)

Call:
lm(formula = mydata$SumAllVisits ~ mydata$Year + mydata$AvgTemp +
    mydata$Unemployment + mydata$GasPrice)

Residuals:
    Min       1Q   Median       3Q      Max
-36102566  -8872299   316748  12698884  24055619

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.587e+09  5.128e+08 -14.796 4.00e-16 ***
mydata$Year    3.924e+06  2.933e+05  13.375 7.05e-15 ***
mydata$AvgTemp  3.722e+06  3.447e+06   1.080  0.2881
mydata$Unemployment -4.362e+06  1.895e+06  -2.302  0.0278 *
mydata$GasPrice -2.540e+06  4.189e+06  -0.606  0.5485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16320000 on 33 degrees of freedom
Multiple R-squared:  0.9049,    Adjusted R-squared:  0.8934
F-statistic: 78.54 on 4 and 33 DF,  p-value: < 2.2e-16
```

**Figure 33 Regression Model Summary (Year)**

```
> model_var_pop <- lm(mydata$SumAllVisits~mydata$AvgTemp+mydata$Unemployme
> summary(model_var_pop)

Call:
lm(formula = mydata$SumAllVisits ~ mydata$AvgTemp + mydata$Unemployment +
    mydata$GasPrice + mydata$Population)

Residuals:
    Min       1Q   Median       3Q      Max
-39072380  -7640849  -247266  12096781  23966620

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.559e+08  1.719e+08  -0.907  0.3712
mydata$AvgTemp  4.213e+06  3.440e+06   1.225  0.2293
mydata$Unemployment -4.179e+06  1.901e+06  -2.198  0.0351 *
mydata$GasPrice  -4.088e+06  4.211e+06  -0.971  0.3388
mydata$Population  1.400e+00  1.051e-01  13.323 7.87e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16380000 on 33 degrees of freedom
Multiple R-squared:  0.9043,    Adjusted R-squared:  0.8927
F-statistic: 77.96 on 4 and 33 DF,  p-value: 2.435e-16
```

**Figure 34 Regression Model Summary (Pop)**



The differentiating factor was the p-value in the model incorporating the year field, which was just smaller than the model including population. In addition, the difference in the minimum and maximum residuals between the two models were close, although neither was symmetrical. The model including year also had a slightly higher f-statistics. Both models also indicated that, of the other variables, only unemployment rate was significant. Another similar measure, the number of parks each year, was also tested. It returned a higher adjusted r-squared value, f-statistics, and residual error. However, its residuals were widely disparate, as illustrated in Figure 35.

```
> model_var_all <- lm(mydata$SumAllVisits~mydata$AvgTemp+mydata$Unemployme
> summary(model_var_all)
```

Call:  
lm(formula = mydata\$SumAllVisits ~ mydata\$AvgTemp + mydata\$Unemployment +  
mydata\$GasPrice + mydata\$ParkNum)

Residuals:

Min	1Q	Median	3Q	Max
-15644220	-7817173	-1613579	4443469	40747844

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-283759373	122697502	-2.313	0.0271 *
mydata\$AvgTemp	3569556	2508255	1.423	0.1641
mydata\$Unemployment	214514	1409312	0.152	0.8799
mydata\$GasPrice	-5138366	3087452	-1.664	0.1055
mydata\$ParkNum	1559608	82233	18.966	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11990000 on 33 degrees of freedom  
Multiple R-squared: 0.9487, Adjusted R-squared: 0.9425  
F-statistic: 152.6 on 4 and 33 DF, p-value: < 2.2e-16

**Figure 35 Regression Model Summary (number of parks)**

Another set of models was tested after excluding visitation from states outside of the contiguous United States. The data for gas prices and weather was only available for a

smaller subset of the system units, it was prudent to remain consistent. The residuals minimum and maximum were much more symmetrical in this model, although the f-statistic was lower, illustrated in Figure 36. However, when accounting for only US-based parks, the unemployment rate became a more significant indicator, even as the fitness of the model decreased from 0.89 to 0.87.

```
> model_US <- lm(mydata$SumUS~mydata$AvgTemp+mydata$Unemployment+myd
> summary(model_US)

Call:
lm(formula = mydata$SumUS ~ mydata$AvgTemp + mydata$Unemployment +
    mydata$GasPrice + mydata$Year)

Residuals:
    Min       1Q   Median       3Q      Max
-38674511 -9775767  1475024  12301536  32658575

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.192e+09  5.407e+08  -13.301 8.24e-15 ***
mydata$AvgTemp    2.531e+06  3.635e+06   0.696  0.4911
mydata$Unemployment -5.479e+06  1.998e+06  -2.741  0.0098 **
mydata$GasPrice   -4.637e+06  4.418e+06  -1.050  0.3016
mydata$Year       3.758e+06  3.093e+05  12.148 1.00e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17210000 on 33 degrees of freedom
Multiple R-squared:  0.8874,    Adjusted R-squared:  0.8737
F-statistic:    65 on 4 and 33 DF,  p-value: 3.523e-15
```

**Figure 36 Regression using US-only visits**

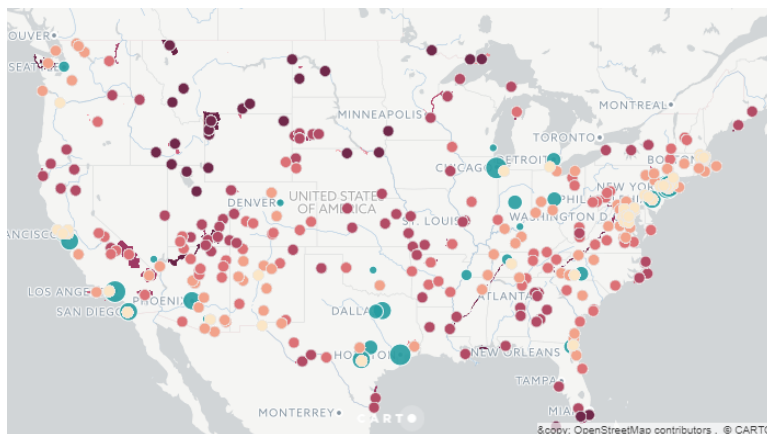
Comparing, sorting, calculating, and viewing various metrics of park visitation by park type, region, and state illustrated national-level trends. These types of comparisons highlighted the long-term growth in park visitors. However, analyzing all system units using historical data revealed some additional benefits. For example, clustering parks

based on a combination of visitor types revealed some interesting connections normally overlooked in favor of the sum of visits. Testing linear models to estimate park visitation based on variables typically affecting tourist travel, along with variables associated with the parks themselves, illustrated which ones were significant.

## CHAPTER FIVE

### Key Findings

The final step of the KDD process requires the presentation of new knowledge and significant results. Static visual displays, descriptions of regression, and discussion of hypotheses provide part of that solution. When handling spatial information, however, an interactive display provides additional benefits. A web map was published through CARTO highlighting the key findings from this research. This map utilized the existing database for selecting and integrating NPS spatial data and visitation. The fields created for this research, such as the results of the nearest-neighbor query to locate the largest city for each park, are represented as well. The final map presents the primary research objectives, highlighting both spatial and temporal characteristics, illustrated in Figure 37.



**Figure 37** Capture of Interactive Web Map

## **Significant Results**

The research illustrated the national-level trends in park visitation, highlighted unique system units, modeled variables affecting visitation, and evaluated the results of each hypothesis, to fulfill the research goals and characterize historical National Park visitation. One of the significant findings of this research, which might spur NPS to change their method for visitor forecasts, concerned the overall trend in park visitation. Ultimately no positive or negative visitor trend lasted for a period of more than five years. If the overall number of visitors was either increasing or decreasing, the trend did not continue indefinitely, but ended after only five years. Visitation fluctuates on a cycle of approximately five years when comparing raw numbers, while still increasing overall. The Park Service uses only the previous five years' visitor statistics to forecast future visitation, which does not account for the historical data.

Using the normalized value of visits per person, calculated by dividing the annual sum of visits over its corresponding population, another trend emerged. The highest rate of visits per person occurred in 1998, 1999, and 1987. The precursor to those high numbers was an increase in rate to approximately 1.58 visits per person, which last occurred in 2015. While this was an imperfect measure, as visitors from other countries use the parks, it at least provided a way to see if the park visitation increased at a similar rate to the number of people who might use them. In this case, characterizing record-breaking visitation by the rate highlights more significant changes than the visitor numbers alone.

Using a difference measurement for assessing a unit's proximity to a populated area was an extension of the NPS research. Instead of categorizing parks by population

density, the nearest neighbor largest city was measured. The recorded distance was useful in ranking and plotting parks and evaluating clusters of parks. However, no direct correlation between distance and visitor was discovered. Comparing the number of park visitors to the number of park system units was also a new approach for this research. Like the estimate of park visits by population, as a variable in a linear regression model, the number of parks correlated to overall visits. Low unemployment was minimally significant, but gas prices and weather were not correlated with change in visits.

## **Discussion**

The maps, plots, and numbers only provided part of the solution in the pursuit of finding answers to the research questions. As expected, iteration was a key skill within the KDD framework for assessing the results of each calculation and query. Reconciling the output of a calculation often required multiple steps to review the outliers and posit possible solutions and reasoning for a certain result. For example, investigating the reason for abrupt visitation changes when calculating the difference over time revealed the difficulty of determining which parks were operating in which years. Locating accurate information on park opening dates was a seemingly simple task that did not have a reliable source for information to references. The only comprehensive listing was on Wikipedia which contained a table of the dates the parks were authorized. In some instances, this did not coincide with the park existing or being open to visitors. In other situations, parks were renamed, or their system unit category was reassigned. In addition, some of the oldest parks were established prior to the existence of the National Park

Service itself (NPS History). Finding details for long-term closures of parks, for work such as maintenance, can be found on the park websites themselves, not on a list of park status. For the purposes of this research, the year the park was considered “open,” was the year that it first reported visitor statistics. Parks reporting no visitors after opening were generally assumed to be undergoing a temporary closure unless information detailing a permanent change was found. However, exceptions were made, such as the case with the multiple listings for parks in the National Capital region. Even then, some parks displayed substantial differences in year-to-year visitation with no method for determining the cause.

Applying k-means clustering to data revealed unique connections between parks based on their visit types. While it typically resulted in two clusters, the clusters of recreation visits highlighted some of the parks that are normally overlooked. Instead of focusing only on highest and lowest values for visits, their spatial location, or their proximity to a city, the clusters grouped together parks that had similar trends in visitation. This provided a way to compare parks within their clusters to identify other possible similarities. The relationship between parks visitation and cities was not directly proportional, and identifying those parks near cities and other highly-visited parks that broke from this trend highlighted an opportunity to increase visitation.

The Park Service acknowledged that it does consider the number of system units and the population in its statement to Congress in 2006, as well as external factors (Blaszak 2006). However, it did not indicate whether these factors, when combined, would reveal similar results. No further reference information was available to determine

if both weather and economic variables were tested together, whether the research was conducted nationally, or the years of data used for comparison. Given the relationships between the external variables of temperature, unemployment, gas price, and population to the number of visitors, various linear models were tested to assess their utility. For example, it was hypothesized that low unemployment and low gas prices would correlate to higher visitation. However, it was determined that the number of parks and the size of the population were much better methods for estimating the number of visitors in this dataset if they were considered alone. Gas prices, temperature, and even park fees were not important to visitation, as initially hypothesized, since they were no longer considered individually. The population accounts for much of the variation in visits. Determining causation, however, requires much further study. Are there more people available to visit parks? Or does an increase number in parks, especially depending on their location, spur interest and visitation? In addition, even though visitation did not seem to be impacted by the prescence park fees, NPS offers different types of passes for frequent visitors. In late 2017, NPS increased the price of its senior lifetime pass from \$20 to \$80 (NPS “Plan Your Visit “2017).

Modelling the variables affecting visitation for one park, with a greater understanding of smaller-scale influences such as demographics, local weather, and economic factors may reveal better methods for creating new models. Although this model was rudimentary, as it does not account for some of the factors affecting visitation acknowledged by NPS. However, without better demographic information collected from park visitors, efforts to model attendance will suffer. For example, rural parks may be



more susceptible to changing visitation based on economic factors because of their distance from population centers. While some trends like lower visitation based on proximity to metropolitan areas were recorded, they have not yet been measured to determine a correlation. There may be different factors affecting different types of visitation as well. For example, backcountry camping requires specialized equipment and gear not necessary for a visit to a park in a city bypassed by a commuter. The economic conditions influencing visits for trip to a more remote park may go beyond gas prices, and include factors such as vehicle ownership and ability to afford a park pass or concessioner lodging, as discussed in demographic studies by Benson et al., Schuett, Le and Hollenhorst, and Weber & Sultana.

Even though visitor data is reported by NPS annually, NPS acknowledges most visits occur in the summer months (Blaszak 2006). Therefore, understanding the summer weather could prove to be more important. Similarly, even parks within the same region display a wide variety of characteristics. Warmer weather may be a boon to parks in the Northeast and Alaska regions, as visitors extend the season beyond the typical summer months, while the opposite could affect parks in the southern states. NPS acknowledged the impact of a changing climate, particularly by assessing visits on a monthly level. Individual researchers also study these in specific parks, but collecting local weather data and assessing it for anomalies would be difficult to coordinate across all parks.

The National Park Service routinely explains trends based on external factors, which mirrors the proven relationships between these factors and their impact of tourism and travel. And while recreation accounts for two-thirds of all visits, the number of other

visits types, particularly non-recreation visits, have limit the ability to apply tourism-related strategies to assessing park visitation. The factors and motivations that influence a traveler's decision to attend a tourist attraction like Disney World, may not be the same as another traveler's decision to eat lunch in a park near his / her workplace. NPS conducts visitor surveys, but the information primarily serves to rate the parks on variety of factors, from amenities to ease of use. Since NPS estimates the park visitation, rather than counting every individual, the ability to survey individuals who do not stop for a typical recreation may impact the understanding of visitor motivations. In addition, the centennial celebrations for NPS, as well as a social media campaign launched in late 2015, were unique factors not considered in this research (NPF 2016). Given previous research indicating the connection between tourism, which included advertising and outreach, this timing could prove to account for the difference.

### **Limitations**

No data source is perfect, and the ones used in this research needed cleaning and preprocessing like any other data source. Although monthly visitor statistic data was available, the preponderance of other information was only available on an annual basis. Using only annual park statistics prevents analysis of seasonal park visitation trends. NPS has previously established the seasonal trends in park visitation, particularly how the bulk of almost all visits occurs over the summer months. Using only point data for a population measures does limit some of the functionality, as computations of Euclidean distance do not account for distance traveled on roads or across borders. All data, apart from city population, was selected for its ability to be sorted by year, starting in 1979. In

addition, most of the external variables were averaged across the whole country, or at least the contiguous 48 states. This meant evaluating the extent to which data such as gas prices and temperature affected visitation was not applicable to all parks. Extrapolating trends on a national level may not provide a high degree of granularity, particularly for exploring visitation. Although the National Park Service provides a simple method to query data, cleaning and interpreting some of its idiosyncrasies proved challenging. While the annual abstracts and other reporting provided a resource for assessing outliers discovered in the data, such as missing unit codes, not all the information was readily attainable. In addition, even updated NPS GIS datasets were inconsistent. The GIS data did not have polygons for some parks with equivalent unit codes when more than one park was administered at a time. Because the trends were assessed over an on annual and national level, however, a loss of granularity was an acceptable solution.

Leveraging trial, free, or open-source tools limited some of the analysis capability of other systems, as well as imposed data storage limits. Therefore, outside of the NPS boundary layer, all other data was non-spatial tables or point data. Future research using these tools may be affected by changes and updates to underlying software and algorithms. For example, if Tableau changes their clustering algorithm, or if CARTO limits the number of queries performed for free, future analysis using these exact methods would become a challenge. Another limitation is the repeatability of geovisualization and other interactive visualizations as a method for detecting patterns and outliers. Everyone may view the same information differently, and certain patterns may not be distinguishable by all.

## CHAPTER SIX

### Conclusion

The interactive and visualization aspects of this research sought to distill a four decades worth of statistics into an easily understandable format and illustrate possible patterns within NPS visitation not yet uncovered. Framing this research within the KDD process revealed certain characteristics and techniques to better manage and interpret data. With the NPS Centennial in 2016, record-breaking visitor numbers for three consecutive years, and current discussions over improving park services while balancing budget requirements, understanding historical NPS visitation using free and publicly-available information proved appropriate. This research aimed to extract as much value out of the raw data as possible into an intuitive format, while providing answers to research questions generated after studying gaps in existing park research. The foremost effort was to put park visitation into context of something other than overall, system-wide visits. Then it was to determine if the visitation in the past few years was significantly greater than previous years, or whether it was part of a longer temporal pattern. Visualizing data beyond state and region, to the specific center of each park boundary, allowed for visual analysis of spatial patterns. Finally, displaying the information in a user-friendly context was paramount to presenting results and spurring interest in future research. This process provides a foundation for further exploration and research.

## **Future Research**

Predicting visitation serves a more important purpose than just an ability to test models. As part of the U.S. Federal Government, the National Park Service must budget to maintain the parks, improve services, and conduct its mission of conservation. NPS could leverage their wealth of historical data to assess the needs of each park and visitor type by furthering a national-level study of its parks. Research to improve the ability to detect trends at each park, for all types of visits, could result in more precise budgeting. This research illustrated many challenges in understanding park visitation, but the process of evaluating a long-term trend lends itself to future studies in forecasting. Assessing the forecasts developed by NPS with actual visitation, especially when considering all visit types, comprises enough material for its own study, even outside of prediction.

For the purposes of this research, a metadata table was created to store basic park information that was important but unnecessary for many queries. A similar table, available publicly, could provide the additional details needed for solving some of those issues. This table could include information about the specific park operating dates, types of fees and passes, and the types of activities or facilities within the park. Displaying this information on an interactive map would offer relevant information quickly, without the need to link between multiple web pages. In addition, the extent to which tourism-related variables impact park visitation requires further exploration of its effects across different geographic locations. Understanding visitor demographics, particularly where visitors come from, would provide fascinating insight into who visits distance and less-popular parks. This research worked to uncover possible causes and correlations for these changes, but did not attempt to predict future visitation.

## REFERENCES

- Benson, C., Watson, P., Taylor, G., Cook, P. and Hollenhorst, S., 2013. Who visits a national park and what do they get out of it?: A joint visitor cluster analysis and travel cost model for Yellowstone National Park. *Environmental management*, 52(4), pp.917-928.
- Blaszak, Marcia. "NPS Visitation Trends: Visitation Trends in the National Park System." Statement before the Subcommittee on National Parks, House Committee on Resources, Washington DC, 6 April 2006. Office of Congressional and Legislative Affairs. <https://www.doi.gov/ocl/nps-visitation-trends>.
- Bonn, M.A., Line, N.D. and Cho, M., 2017. low gasoline prices: The effects upon auto visitor spending, numbers of activities, satisfaction, and return intention. *Journal of Travel Research*, 56(2), pp.263-278.
- Boundless Geo. 2011. "How Do I Find The N Nearest Things To This Point In PostGis." September 28, 2011. <https://boundlessgeo.com/2011/09/indexed-nearest-neighbour-search-in-postgis/>.
- Brachman, R.J. and Anand, T., 1996. The process of knowledge discovery in databases. *Advances in knowledge discovery and data mining*, pp.37-57.
- Cessford, G. and Muhar, A., 2003. Monitoring options for visitor numbers in national parks and natural areas. *Journal for nature conservation*, 11(4), pp.240-250.
- Eagles, Paul FJ. "Research priorities in park tourism." *Journal of Sustainable Tourism* 22, no. 4 (2014): 528-549.
- Elwood, Sarah. 2009. "Geographic Information Science: new geovisualization technologies - emerging questions and linkages with GIScience research." *Progress In Human Geography* 33, no. 2: 256-263. Academic Search Complete, EBSCOhost (accessed August 23, 2017).
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3), p.37.
- Fisichelli, N.A., Schuurman, G.W., Monahan, W.B. and Ziesler, P.S., 2015. Protected area tourism in a changing climate: Will visitation at US national parks warm up or overheat?. *PloS one*, 10(6), p.e0128226.
- Flowers, Andrews. 2016. The National Parks Have Never Been More Popular. May 25. Retrieved June 23, 2017, from <https://fivethirtyeight.com/features/the-national-parks-have-never-been-more-popular/>
- Fredman, Peter, Lisa Hörnsten Friberg, and Lars Emmelin. 2007. Increased visitation from national park designation. *Current issues in tourism*, 10(1), pp.87-95. <http://www.sciencedirect.com/mutex.gmu.edu/science/article/pii/S0378475405000613>

- Gahegan, M., Wachowicz, M., Harrower, M. and Rhyne, T.M., 2001. The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Science*, 28(1), pp.29-44.
- Baker, V.R. 1999. Geosemiosis. *GSA Bulletin* (May 1999) 111(5): 633-45.
- Leedy, Paul D. *Practical research: Planning and design*. Macmillan, 1993.
- Haklay, M., Singleton, A. and Parker, C., 2008. Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass*, 2(6), pp.2011-2039.
- Hetter, Katia. 2015. Record Millions Flocked to National Parks Last Year - CNN. February 17. Retrieved 10 June 2017 from <http://www.cnn.com/travel/article/feat-most-visited-national-parks-sites-2014/index.html>.
- MacEachren, A.M., Gahegan, M., Pike, W., Brewer, I., Cai, G., Lengerich, E. and Hardisty, F., 2004. Geovisualization for knowledge construction and decision support. *IEEE computer graphics and applications*, 24(1), pp.13-17.
- MacEachren, A.M. and Kraak, M.J., 2001. Research challenges in geovisualization. *Cartography and geographic information science*, 28(1), pp.3-12.
- Mennis, J. and Guo, D., 2009. Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, 33(6), pp.403-408.
- NOAA National Centers for Environmental Information. 2017. Climate at a Glance: U.S. Time Series, Average Temperature. November 2017, retrieved on August 5, 2017 from <http://www.ncdc.noaa.gov/cag/>.
- Neher, Christopher, John Duffield, and David Patterson. 2013. "Valuation of National Park System Visitation: The Efficient Use of Count Data Models, Meta-Analysis, and Secondary Visitor Survey Data." *Environmental Management* 52 (3): 683–98. doi:[10.1007/s00267-013-0080-2](https://doi.org/10.1007/s00267-013-0080-2).
- Reynolds, C. 2017. National parks saw a record-setting number of visitors last year. Were they too much of a good thing? Los Angeles Times. March 1. Retrieved 10 June 2017 from <http://www.latimes.com/travel/la-tr-nps-crowds-20170228-story.html>.
- Rodger, K., Taplin, R.H. and Moore, S.A., 2015. Using a randomised experiment to test the causal effect of service quality on visitor satisfaction and loyalty in a remote national park. *Tourism Management*, 50, pp.172-183. Saxton, Gregory. "Analyzing Big Data with Python PANDAS." Social Metrics Blog. <http://social-metrics.org/analyzing-big-data-with-python-pandas/>.
- Schuett, Michael A., Lena Le, and Steven J. Hollenhorst. 2010. Who visits the US National Parks? An analysis of park visitors and visitation: 1990–2008. *World Leisure Journal*, 52(3), pp.200-210.
- Taylor, Patricia A., Burke D. Grandjean, and James H. Gramann. 2011. National Park Service comprehensive survey of the American public, 2008–2009: Racial and ethnic diversity of National Park System visitors and non-visitors. Natural Resource Report NPS/NRSS/SSD/NRR—2011432. National Park Service, Fort Collins, Colorado. Retrieved 6 June 2017 from

- [https://www.nature.nps.gov/socialscience/docs/CompSurvey2008\\_2009RaceEthnicity.pdf](https://www.nature.nps.gov/socialscience/docs/CompSurvey2008_2009RaceEthnicity.pdf).
- Torielli, AC. 2017. National Park Visitation 1979 – 2016. CARTO.  
<https://g01025023.carto.com/builder/0093f4cb-081a-45cd-b31d-8fc6f2687c53/embed>.
- U.S. Department of Commerce. Bureau of Economic Analysis. *National Income and Product Accounts, Table 1.1.9. Implicit Price Deflators for Gross Domestic Product*. February 2016.
- U.S. Department of Labor, Bureau of Labor Statistics. 2001. *Labor Force Statistics from the Current Population Survey*. “What is seasonal adjustment?” October. Retrieved 5 September 2017 from <https://www.bls.gov/cps/seasfaq.htm>.
- U.S. Department of Labor, Bureau of Labor Statistics. 2017. *Databases, Tables & Calculators by Subject*. Retrieved 5 September 2017 from <https://www.bls.gov/cps/seasfaq.htm>.
- U.S. Department of Energy. 2017. United States - Maps - U.S. Energy Information Administration (EIA). Independent Statistics & Analysis - U.S. Energy Information Administration. Retrieved 23 July 2017 from <https://www.eia.gov/state/maps.php>.
- U.S. Department of Energy. 2017. “Glossary - petroleum administration for defense district.” Independent Statistics & Analysis - U.S. Energy Information Administration. Retrieved 23 July 2017  
<https://www.eia.gov/tools/glossary/index.php?id=petroleum%20administration%20for%20defense%20district>.
- U.S. Department of the Interior. 6 July 2017. National Park Service. “Annual Visitation Highlights.” Retrieved 9 June 2017 from <https://www.nps.gov/aboutus/faqs.htm>.
- U.S. Department of the Interior. National Park Service. “Frequently Asked Questions.” Retrieved 9 June 2017 from <https://www.nps.gov/aboutus/faqs.htm>.
- U.S. Department of Interior. 30 Mar 2017. National Park Service Media Release. NPS and NPF Launch #FindYourPark. <https://www.nps.gov/subjects/centennial/nps-and-npf-launch-hashtag-findyourpark.htm>.
- U.S. Department of the Interior. National Park Service Centennial. 2016. “National Park Service Centennial Final Report - Realizing the Vision for the Second Century.” December. Retrieved 10 June 2017 from <https://www.nps.gov/subjects/centennial/upload/Centennial-Final-Report-December-2016.pdf>.
- U.S. Department of the Interior. National Park Service. Statistical Office, Denver Service Center. 1979. “National Park Statistical Abstract 1979.” Retrieved 12 September 2017 from <https://irma.nps.gov/Stats/Reports/AbstractsAndForecasts>.
- U.S. Department of the Interior. National Resource Stewardship and Science. 2017. “Statistical Abstract: 2016.” March. Retrieved 15 September 2017 from.
- U.S. Department of the Interior. National Park Service Data Store. Integrated Resource Management Applications. “Administrative Boundaries of National Park System Units 3/31/2017.” Code: 2225713. Retrieved 9 June 2017 from



- <https://irma.nps.gov/DataStore/Search/Advanced> and <https://g01025023.carto.com/dashboard/datasets/library>.
- U.S. Department of the Interior. National Park Service Data Store. Integrated Resource Management Applications. “Administrative Boundaries of National Park System Units 9/30/2017.” Code: 2225713. Retrieved 1 November 2017 from <https://irma.nps.gov/DataStore/Search/Advanced> and <https://g01025023.carto.com/dashboard/datasets/library>.
- U.S. Department of the Interior. National Park Service Data Store. Integrated Resource Management Applications. “Administrative Boundaries Centroids of National Park System Units 3/31/2017.” Code: 2225714. Retrieved 10 June 2017 from <https://irma.nps.gov/DataStore/Search/Advanced>.
- U.S. Department of the Interior. National Park Service Data Store. Integrated Resource Management Applications. National Park Service Annual Statistics. Retrieved 10 June 2017 from <https://irma.nps.gov/DataStore/Search/Advanced>.
- Weber, J. and Sultana, S., 2013. Why do so few minority people visit National Parks? Visitation and the accessibility of “America's Best Idea”. *Annals of the Association of American Geographers*, 103(3), pp.437-464.
- World Bank. FRED, Federal Reserve Bank of St. Louis. *Population, Total for United States [POPTOTUSA647NWDB]*. 15 September 2017. Retrieved 19 September 2017 from; <https://fred.stlouisfed.org/series/POPTOTUSA647NWDB>.
- Ziesler, P. S. 2017. Statistical abstract: 2016. Natural Resource Data Series NPS/NRSS/EQD/NRDS—2017/1091. National Park Service, Fort Collins, Colorado. Retrieved 15 September 2017 from <https://irma.nps.gov/Stats/Reports/AbstractsAndForecasts>.
- Ziesler, P. S. 2016. Statistical Abstract: 2015. Natural Resource Data Series NPS/NRSS/EQD/NRDS—2016/1009. National Park Service, Fort Collins, Colorado. Retrieved 15 September 2017 from <https://irma.nps.gov/Stats/Reports/AbstractsAndForecasts>.

## **BIOGRAPHY**

Adrienne Camille Torielli received her Bachelor of Science from the United States Air Force Academy in 2011. She was commissioned as a Second Lieutenant in the U.S. Air Force and is currently serving on Active Duty. She received her Master of Arts in Intelligence Studies from American Military University in 2015.