USING CITY ATTRIBUTES TO PREDICT HUMAN TRAFFICKING BUSINESS TYPES

by

Lydia Shirey
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Systems Engineering

Committee:

| | |
|---|---|
| ___________________________________ | Dr. Edward Huang, Thesis Director |
| ___________________________________ | Dr. Jessica Lin, Committee Member |
| ___________________________________ | Dr. Ran Ji, Committee Member |
| ___________________________________ | Dr. John Shortle, Department Chair |
| ___________________________________ | Dr. Kenneth S. Ball, Dean, College of Engineering and Computing |
| Date:______________________________ | Fall Semester 2021<br>George Mason University<br>Fairfax, VA |

Using City Attributes to Predict Human Trafficking Business Types

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

by

Lydia Shirey
Bachelor of Science
George Mason University, 2016

Director: Edward Huang, Associate Professor
Systems Engineering and Operations Research Department

Fall Semester 2021
George Mason University
Fairfax, VA

Copyright 2021 Lydia Shirey
All Rights Reserved

## ACKNOWLEDGEMENTS

I would like to thank my family, my friends, and my colleagues who have supported me through this process. I would especially like to thank: PhD student Julia Cheng for being an invaluable resource for data analytics, answering all my questions, and guiding me through unfamiliar processes; my advisor, Dr. Huang, for keeping me on track and assisting me through this whole process, helping me narrow my research down into a sensible scope, and providing excellent constructive feedback; and my review committee, Dr. Jessica Lin and Dr. Ran Ji, for their support and review of this thesis. Finally, composer and conductor Howard Shore receives an honorable mention for providing unparalleled musical scores for my research and writing ambiance.

Soli Deo gloria.

**TABLE OF CONTENTS**

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF EQUATIONS

# LIST OF ABBREVIATIONS

Area Under the Curve ........................................................................................ AUC
Domestic Minor Sex Trafficking ................................................................................ DMST
Federal Bureau of Investigation ........................................................................................ FBI
Homeland Security Investigations ........................................................................................ HIS
Intimate Partner Violence ........................................................................................ IPV
Receiver Operating Characteristic ........................................................................................ ROC
Synthetic Minority Oversampling Technique ........................................................ SMOTE

## ABSTRACT

USING CITY ATTRIBUTES TO PREDICT HUMAN TRAFFICKING BUSINESS TYPES

Lydia Shirey, M.S.

George Mason University, 2021

Thesis Director: Edward Huang

Human traffickers employ a variety of business models with which to traffic victims and profit from their victimization. Awareness of human trafficking, particularly sex trafficking, and how to recognize it are key facilitators in reducing the problem and preventing further victimization. Determining predictors of sex trafficking business models will contribute to building awareness, and this analysis seeks to establish if sex trafficking business models and city characteristics are correlated. Logistic regression, multinomial logistic regression, and random forest models were trained on data collected from the court documents of federally prosecuted human trafficking cases in the State of California as well as city attributes from Census data – e.g. population density, median home value, education statistics, etc. – to determine their correlation. Results indicate that certain city characteristics are correlated with sex trafficking business models, which can

be used to provide indications of what type of sex trafficking business models will be problematic in a given city based on these key characteristics.

# CHAPTER ONE: INTRODUCTION

## Section One: Background

Human trafficking is a broad term that encompasses several different, though interrelated, crimes. Labor trafficking is the exploitation of persons, adult or child, to perform work against their will or under coercion; this can include domestic servitude, service industry labor, or manual labor. Similarly, sex trafficking is the exploitation of persons, adult or child, to work in the commercial sex industry against their will or under coercion. Traffickers profit financially off of both types of trafficking, but the direct means of profit may vary [1]. Human trafficking contains other types of trafficking in addition to sex and labor, but the majority of trafficking incidences are related to sex trafficking and labor trafficking [2].

Sex trafficking is one of the most common type of human trafficking on a global level, and labor trafficking is becoming increasingly more common in recent years. Additionally, trends in the trafficking victims are changing accordingly, with the proportion of adult women decreasing in recent years while the proportion of male victims and adolescent girls increasing, although adult women are still the most trafficked out of all gender and age categories. In the United States, prosecution of sex trafficking has also increased over time, reflecting the implementation of policies to combat the problem domestically [2].

For the purposes of this analysis, the focus is on sex trafficking as opposed to labor trafficking, and specifically as relates to the United States. Sex trafficking can be accomplished on a small scale, such as an individual pimp selling one or more individuals for commercial sex, or on a large scale, such as human trafficking networks that span multiple states or countries and involve hundreds of victims. The nature of the trafficking model utilized to traffic or sell the victims of trafficking can have an impact on the best method to be used for combatting trafficking [1].

The scale of an operation will, by necessity, change the approach taken by law enforcement, where an individual pimp could be combatted by local law enforce while a national trafficking ring may require the assistance of the Federal Bureau of Investigation (FBI) or Homeland Security Investigations (HIS), in addition to state and local law enforcement, due to limited jurisdictions. With this in mind, a better understanding of how various sex trafficking businesses are organized and which are the key indicators of trafficking could help to reduce the problem [1].

## Section Two: Sex Trafficking Business Models

Shelley [3] gives very detailed descriptions of various business models that can be found internationally and domestically. The focus of her work is centered around the disparities in how human trafficking looks in different regions, seeking to describe why and how these disparities may have arisen as well as to characterize the unique aspects of each regional model. Proposed regional human trafficking models are the Chinese Trade and Development Model, the Post-Soviet Natural Resource Model, the Balkan Crime

Group Model, the United States Pimp Model, the U.S.-Mexican Supermarket Model, and the Nigerian and West African Trafficking Model.

The Chinese Trade and Development Model Shelley proposes is rooted in organized crime and human smuggling, or the movement of persons desiring to migrate illegally. Cultural and societal practices in Asia may also make the use of prostitutes a more common issue than other regions, and thus brothels and escort services are key aspects of the Chinese model. Trafficking networks are both local and international and often operate with little legal interference locally [3].

The Post-Soviet Natural Resource Model is less polished and more disjointed than the Chinese model, but is characterized by both the domestic utilization of prostitution and the export of prostitutes to be exploited internationally. Further insidious practices involve child trafficking for the production of pornography and the adept and prolific use of the internet to traffic victims. The Balkan Crime Group Model, in contrast, is dominated by localized organized crime although it too has a base of operations predominantly in Europe. The Balkan model utilizes a gang-type approach and keeps expenses low by exploiting women as streetwalkers [3].

The United States Pimp Model is marked by the prevalence of small-scale sex trafficking businesses, whereby pimps capitalize on the vulnerable members of society that are easy to coerce into prostitution. The internet is widely used to advertise victims and hotels or transient dwellings are often utilized as well. Similar to the Balkan Crime Group Model, the U.S.-Mexican Supermarket Model is characterized by violent gangs operating large trafficking rings. A primary distinction is that the U.S.-Mexican

Supermarket Model emphasizes the quantity of customers rather than an exclusivity of services, the latter of which is often more lucrative. The trafficked victims are exploited locally in various lower-level business models, but the gang managing the trafficking ring ensures its presence is known through the use of violence and forcible control [3].

The final model proposed by Shelley is the Nigerian and West African Trafficking Model. Organized crime is present in this model as well, utilizing coercion to maintain control of the victims that have been trafficking to be prostitutes in Europe and other regions. These networks draw victims from all over West Africa in addition to Nigeria, and the corruption in Nigeria encourages the trade [3].

Many of the regional trafficking models proposed by Shelley have overlapping themes and commonalities in the use of lower-level operational business models, such as brothels, streetwalking, and prostitutes often managed by pimps. Even the models characterized by the presence of organized crime and gangs still utilized these operational models at a lower level to facilitate the transactions. These operational models align generally with the public perception of sex trafficking operations, oftentimes being the primary physical indication or “storefront” encountered by the public while the higher-level regional model being utilized may be too removed to be easily recognized by the public or law enforcement. With that in mind, this analysis focuses on operational business models utilized by sex traffickers as a way to define localized manifestations of the regional business models.

Relying on characteristics of the various models presented by Shelley as well as analyzing the different business models present in United States sex trafficking court cases [4], a brief description of operational business models is incorporated in the following sections. The proposed operational business model breakdown is by no means meant to represent an exhaustive list of the different types of business models used by sex traffickers, but rather is meant to categorize and distinguish between the most common types of operational business models encountered in the court documentation and the regional business models referenced in this analysis.

**The Pimp Model**

The Pimp Model generally follows a template of one trafficker facilitating the sale of one or more victims for the purposes of commercial sex. Pimps do occasionally trade or sell their victims to other pimps, based on information from United States court documentation, and some may operate in business together, but the general pattern is for pimps to operate alone. Hotels and other transient occupancy residences are utilized to foil detection by law enforcement, to facilitate easily moving the operation to a different location, and to keep expenses relatively low [4].

The victims are often controlled by violence and are coerced into submission. It is not uncommon for pimps to initiate romantic relationships with their victims prior to forcing them into sex work, and victims may also be enticed into the industry by pimps exaggerating their profits and promising a glamorous lifestyle. In reality, pimps take most or all of the victim's earnings, leaving them financially dependent on their pimps and unable to leave [4].

### The Streetwalker Model

The Streetwalker Model is inherently similar to the pimp model, primarily distinguished by the method of sale. Whereas in the Pimp Model, the pimp himself will often seek out customers through online advertisements or interpersonal networks, streetwalkers are left to solicit customers on their own. This model also utilizes exclusively face-to-face advertisement rather than internet advertisements. Otherwise, there are similarities in the use of hotels or transient residences and oftentimes there is a pimp-like figure that manages the sex worker [4].

### The Brothel Model

The Brothel Model is distinguished by the use of a private residence or business out of which is run a sex trafficking operation. Rather than a transient location more common in the Pimp and Streetwalker Models, these operations are restricted to one or more long-term locations. Common businesses used as fronts for sex work are massage parlors, although a wide variety of other business models were referenced in the court documentation. Despite generally being more profitable than the other business models, the risk of detection is greater with the Brothel Model since the operation is not able to be relocated easily and is harder to conceal from law enforcement and the general public [4].

With the Brothel Model, the victims are usually coerced into the trade by their traffickers and not uncommonly have been trafficked in from foreign countries. The traffickers force them into sex work to pay off their inflated debts from being smuggled into the country, and threats of personal or familial violence effectively keep the victims submissive. Most international victims were promised other forms of work more consistent with labor trafficking, such as restaurant or farm work, and were informed of the reality of their situation only upon arrival [4].

**The Gang Model**

The Gang Model differs slightly from the previous three in that it operates on a higher and more organized level. This model is most often characterized by a large trafficking network, organized by one or more gangs, for the purposes of trafficking a wide variety of individuals. Based on information gathered from federally prosecuted human trafficking court cases, these operations often span multiple states and regularly transport victims from location to location to avoid detection and increase profits. In order to facilitate the sale of victims, the Gang Model utilizes all three of the aforementioned models, but the overarching organization distinguishes this model [4].

## Section Three: Motivation

Despite having a large and well-documented problem with sex trafficking, the greater United States public is not well aware of the scale of the problem and the degree to which it has spread domestically [3]. Messaging to improve the public's awareness has increased in the recent past, for example the notable work done to combat trafficking at large sporting events or the Polaris project, but an increased public awareness of the issue and warning signs could help to combat trafficking across the nation.

Since all sex trafficking utilizes some kind of victim-trafficker dynamic, approaches to prevent sex trafficking may focus on identifying common victim characteristics in an effort to understand those most vulnerable to being trafficking, or they may focus on how to find or identify traffickers. The assertion could be made that combatting sex trafficking by eliminating businesses is a much more efficient approach than focusing on individual victims. Such an approach may focus on the best methods for identifying the perpetrator or may seek to identify what type of business model is likely to be used by the perpetrator. In either case, focusing on the traffickers rather than the trafficked provides a method to combat the localized or regional source of the issue rather than release a single individual and leave a trafficker free to find additional victims.

Such an approach would likely be more valuable to focus law enforcement efforts since the identification of individual sex trafficking victims may be accomplished by social or healthcare workers. Providing local law enforcement with methods to determine what type of business models may be common in a particular locality could assist in narrowing the scope of their work to better facilitate the detection of trafficking. Furthermore, it may enable the general public to be more informed of the type of trafficking that may be common in their localized area.

## Section Four: Research Question

With all of this in mind, the goal of this research is to determine if sex trafficking business models and regional attributes, such as tax rates, population statistics, and other common city characteristics, are correlated. If such a correlation exists, characterizing the relationship between business models and regional characteristics could prove to be a valuable tool to inform or assist law enforcement. For the purposes of assisting law enforcement, it was considered likely that a finer granularity of analysis, such as at the zip code or the city level, would be more beneficial than a characterization at the county or a higher regional level due to the limits of local jurisdictions; the city level proved the most feasible option based on the limited location information recorded in the court cases. This correlation may then be used to predict the type of sex trafficking operational business models – pimps, streetwalkers, brothels, or gangs – that may be most common in a particular city based on general socioeconomic characteristics of that city. The following analysis seeks to determine if there is a correlation between the types of operational business models and the characteristics of the cities in which they were discovered using data from actual sex trafficking court cases prosecuted in the United States [4] and multiple machine learning algorithms.

# CHAPTER TWO: LITERATURE REVIEW

There exists extensive literature on the topic of sex trafficking as well as utilizing regional characteristics to predict crime. The literature reviewed in the following sections represents the most relevant material related to this research question, but only represents a small subset of the total volume of research into related topics.

## Section One: Characterizing Victims

Much of the research relating to human trafficking centers around characterizing actual and potential victims through the use of surveying healthcare or social workers as well as analyzing past victims to establish key patterns. A quantity of literature has focused on surveys or questionnaires of social and healthcare workers as they are often some of the first caregivers that would have the opportunity of detecting sex trafficking in their patients or clients. Much emphasis is placed on the importance of training healthcare and social workers on how recognize sex trafficking indicators in clients and how to assist in making positive determinations [5], [6]. Cole and Sprang [5] conducted a survey to analyze healthcare workers' awareness of child sex trafficking as well as to collect data on victim risk factors from those cases confirmed to be child sex trafficking.

One goal of Cole and Sprang [5] was to gather more data on communities largely neglected in previous studies, particularly micropolitan and rural communities, as well as collect further data on metropolitan communities. Key results negated the common perception that sex trafficking victims are not United States citizens and affirmed that a significant proportion of traffickers are a familial relative of the victim.

Bloom [7], though primarily focused on Intimate Partner Violence (IPV), also analyzed the role of healthcare workers in identifying human trafficking victims and what indicators may contribute to detection. In addition to healthcare and social workers, Konstantopoulos et al. [6], interviewed a wide range of professions that could be, or are, involved with sex trafficking victims, such as law enforcement officers, government workers, philanthropists, and those in the legal profession to gather information on human trafficking victims.

Possibly due to this broadened survey population, the analysis of Konstantopoulos et al. [6] provided more information regarding the current problems faced by those combatting human trafficking. Interviewees indicated they believed: that reported sex trafficking numbers underestimated the true extent of the issue; that child sexual and physical abuse, poverty, lack of education, and similar factors were predictors of who is likely to be victimized by sex trafficking; and that a lack of engagement of health care systems in responding to sex trafficking has perpetuated the problem.

Further studies focused on analyzing populations of the already-confirmed victims and sought to determine predictive risk factors based on empirical data. Fedina et al. [8] sought to analyze domestic child sex trafficking in the United States, both instances of victims trafficked from inside the country as well as those brought into the country from other nations. Researchers surveyed youth sex trafficking victims and the multivariate analysis of survey results indicated two key victim characteristics as being strongly correlated with child sex trafficking: victims were runaways prior to becoming victims and are racial minorities.

Other attributes considered by Fedina et al. [8], such as physical or sexual abuse at a young age or close association with sex workers, were not determined to be significant indicators of children being trafficked for sex, contrary to the impression of the interviewees in the Konstantopoulos et al. [6] analysis. This underscores the importance of incorporating both qualitative and quantitative analysis in such studies to potentially alleviate cognitive bias, although further analysis is needed to determine which case is likely to be more accurate. Cockbain et al. [9] emphasize the importance, from the viewpoint of combatting sex trafficking, of focusing on empirical evidence rather than popular assumptions about crime, as evidenced by traffickers being observed to operate opportunistically rather than in a highly-organized or sophisticated manner, contrary to the common perception.

From empirical evidence, Twis et al. [10] utilized a multiple case study approach to investigate the relationships between human trafficking victims and their romantic partners, specifically seeking to provide an understanding of domestic minor sex trafficking (DMST) victims. Themes of the case studies include gang involvement by the trafficker/romantic partner, seemingly innocuous romantic relationships that turn into victimization, victims having familial involvement with the sex industry or familial sexual violence, and victims being under the control of the trafficker by means of socio-economic dependencies and/or abuse.

Twis [11] also sought to characterize the relationships between traffickers and their victims using multinomial logistic regression and an Odds Ratio model. Results indicated that characteristics of the victim were correlated with a higher risk of being trafficked by a family member, such as involvement of the victims in welfare and juvenile justice systems.

Panlilioa et al. [12] also analyzed child victim characteristics in detail, concluding that key victim characteristics are that they are or were runaways, drug users, and/or have experienced serious injuries in the recent past before being victimized by traffickers. Of particular interest to the researchers was to narrow the scope of common victim indicators to better assist law enforcement efforts and simplify characterization.

Reid et al. [13] utilized logistic regression to investigate factors that predispose children to fall victim to human trafficking, providing insight into how these indicators differ between boys and girls. Results reinforced the perceptions of those surveyed in Konstantopoulos et al. [6] and differed from Fedina et al. [8] in that neglect and abuse were found to be correlated with incidences of child trafficking. Despite being just a subset of the total research conducted on characterizing human trafficking victims, the breadth and disparity in the approaches and results of the various researchers indicates how widely-varying and difficult it can be to characterize human trafficking. However, the research into characterizing victims is quite extensive, and represents a wealth of information for assisting caregivers and law enforcement in detecting those being victimized by human traffickers.

## Section Two: Characterizing Offenders

With the primary focus of the literature seeking to characterize the victims of human trafficking, only a small grouping of sources were found that sought to characterize the perpetrators involved. Contesting the narrative that sex trafficking in the United Kingdom is a highly organized phenomenon, Cockbain and Wortley [9] characterized traffickers as having varied criminal experience, easy access to their victims, and provided some insight into how they targeted more vulnerable individuals to traffic.

Walters and Davis [14] establish some of the primary characteristics of the traffickers common at the United States and Mexico border, describing their strategies for recruiting unsuspecting victims and their primary motivations to conduct sex trafficking. Hughes [15] not only presents the importance of focusing on the perpetrators, both the consumers of commercial sex and the traffickers, but also stresses the reasons why the benefits of victim-focused approaches to characterize and combat sex trafficking are limited. The customers of the illicit sex industry are allowed to remain vague and uncensured while the victims of trafficking are stigmatized for their involvement without regard for the force, fraud, and coercion involved in their recruitment. Hughes proposes that focusing on the perpetrators is a much more constructive method of combatting sex trafficking as it focuses on the source of the problem.

Carpinteri et al. [16] further expounds upon offender characteristics, both traffickers and customers of the trafficked, utilizing logistic regression to analyze predictors of various sex trafficking offenders based on social and personal characteristics. The authors note the lack of research on the topic of characterizing perpetrators and posit this to be a beneficial topic of further research. These few sources were the primary research that could be found on the topic of characterizing sex trafficking offenders, in stark contrast to the abundance of research that has been conducted in order to characterize human trafficking victims. This dearth of information may simply be caused by the general difficulty of conducting interviews and collecting information from the traffickers themselves. It is easier for researchers or caregivers to gain the trust of the victims than for researchers or law enforcement to draw reliable information from the traffickers, but these difficulties contribute to a general focus on the victims rather than the victimizers, and thus represent a large gap in the field of research.

## Section Three: Spatial Analyses

Another body of research applicable to this analysis is composed of studies utilizing data analytics and machine learning techniques to characterize and predict crime trends in time and space. There is a large quantity of research which focuses on crime prediction and the following subset was referenced as being most related to the topic of this analysis. Using Bayesian analysis, Mahfoud et al. [17] sought to characterize spatio-temporal trends in residential burglary in the city of Amsterdam. The resulting analysis was proposed to be utilized for setting the scheduling routines of local police, and their conclusions were that recency and proximity of crimes were characteristics not conducive to crime forecasting. Also utilizing Bayesian networks, Kiss et al. [18] analyzed Nepalese migration with causal variables and determined that the destination country of the individuals was a key factor in whether or not they were victims of labor trafficking.

Pal and Mondal [19] focused their research on changes in spatio-temporal trends in human trafficking in the country of India. The authors analyzed the rate of sex trafficking across the country and discovered that, while overall human trafficking numbers have been declining in the country as a whole, it has become a more widespread issue as it traveled to states that previously had lower rates.

Helderop et al. [20] analyzed hotel reviews written by guests at the hotels to determine spatial hotspots where prostitution was occurring. Information was restricted to the city of Phoenix in Arizona, and natural language processing was utilized to gather data from the online reviews. Hotspots were then determined using clustering based on the severity of prostitution, and they concluded there is a negative correlation between quantity of prostitution activity and the average cost of a night at the hotel. A random forest model was also employed to predict in which severity cluster a hotel might belong, and it was determined that the cost of the hotel was not a key predictor, but rather location of the hotel and vectors from the natural language processing. Key location characteristics for indicating higher prostitution activity were proximity to major highways and the presence of industrial and commercial centers.

Xia et al. [21] analyzed drug trends in regions compared to location, social, and demographic information to determine correlations. A random forest model was trained on drug-related death locations rather than police information to prevent incorporating any bias from law enforcement, and the analysis was conducted spatiotemporally in order to characterize trends over time. The authors noted that incorporating temporal information into the model improved its performance compared to exclusively a spatial analysis because of the ability to distinguish trends over time rather than average the drug-related incidents over time. This research was similar to that conducted by Barnum et al. [22], which utilized risk terrain modeling on drug-dealing locations and sought to find common characteristics in the locations; they found that accessibility to the buyers as well as security from legal and physical interventions were key to determining drug-dealing hotspots.

Fry [23] utilized logistic and stepwise regression to determine key factors which could be used to predict human trafficking. It was determined that a country's corruption index and the percent of the population under the age of 14 were statistically significant predictors for the country of origin in human trafficking. Building upon previous studies, this analysis also utilized destination countries and determined that population and the Human Development Index – a national rating referencing life expectancy, literacy rates, and gross domestic product – were key predictors for determining the destination for the trafficking victims.

Focusing on crime hotspot mapping, Chainey et al. [24] analyzed the predictive capabilities of various hotspot mapping techniques and sought to determine the most effective tool for predicting crime hotspots based on previous hotspots and the type of crime. Results concluded that spatial prediction abilities did vary across the different types of crime being considered, and that kernel density estimation was generally the best performing method tested by the researchers to predict crime hotspots.

Of particular interest due to its focus on sex trafficking specifically, Mletzko et al. [25] studied indicators of sex trafficking hotspots using kernel density estimation and spatial clustering techniques to analyze sex trafficking distribution as well as binomial regression models to measure hotspots in relation to spatial characteristics. Results indicated that proximity to highways, hotels and other transient occupancy facilities, sexually oriented businesses, and sociodemographic attributes such as residential instability were correlated with sex trafficking hotspots. Their focus was also specifically on urban settings, not suburban or rural regions.

After reviewing this literature on detecting crime hotspots and researching predictive crime analytics, it can be seen that there are many varying procedures that can be conducted to characterize crime in various localities. Although there can be large disparities between the key predictors and aspects of different types of crime, the general procedures followed by these researchers ought to be adaptable to a wide variety of criminal disciplines since they rely on common data analytics concepts and are not narrowly focused on a single discipline. Therefore these various approaches informed the methodology utilized herein, despite the fact that few of the referenced literature focused on sex trafficking as the focal point of their research.

## Section Four: Characterizing Trafficking Types and Business Models

As a final primary applicable category in the literature, there is some research that has been conducted to characterize trafficking types or business models. Cockbain and Bowers [26] utilized multinomial logistic regression to predict the type of human trafficking involved based on individual cases in the United Kingdom. They analyzed the key predictors from the model results in terms of predicting sex trafficking, domestic labor trafficking, and other labor trafficking. Age, gender, and nationality among other predictors were analyzed, and gender was an especially important predictor across the three trafficking types.

Shelley [3] and her regional model characterizations were described in Chapter One of this analysis, but her focus is on large-scale regional models of human trafficking that transcend national borders. Shelley's six observed regional models describe all aspects of human trafficking, but also specifically mention lower-level operational business models from which the four focused on by this research – pimps, streetwalkers, brothels, and gangs – were summarized.

Oblinger [27] further analyzes the lower-level and more operationally-focused business models specifically used by sex traffickers. Describing sex workers in general, Oblinger states that most utilize a pimp for protection, are increasingly using internet ads to streamline their business, and some, in a more traditional model, utilize streetwalking. Larger organized rings move victims from location to location on a prearranged circuit, under the management of a pimp. The generic term, sexually oriented businesses, encompasses escort services, strip clubs, and massage parlors, and these businesses seek to disguise or conflate the sale of sexual services as a legitimate business and thus are generally able to thrive. Notably absent is the mention of brothels, which would conceivably also fall under the category of sexually oriented businesses and is one of the more common business models found in the United States sex trafficking court cases [4].

Stearns [28] analyzes the relationship between gangs and human trafficking, including sex trafficking, and notes the recent trend of normally rival gangs operating together to form human trafficking enterprises. Stearns characterizes human trafficking in terms of the gangs as being low risk and high reward compared to their other criminal activities, and thus provides insight into why gangs have entered the world of human trafficking as well as why rival gangs would join forces. These final analyses were the primary sources found that provide descriptive explanations of the common business models employed by sex traffickers and were referenced when deciding the four categories of business models used in this analysis.

From this review of the available literature, it can be seen that many studies focus on characterizing the victims of human trafficking, either using surveys of care workers or by analyzing data gathered from victims. Significantly less research has been conducted to characterize the perpetrators of sex trafficking, both the buyers or consumers of commercial sex acts and the traffickers themselves, and represents an opportunity for future analysis as it appears to be a beneficial approach to the topic.

Many of the studies employing the technique of analyzing spatial crime trends focused on drug dealing or burglary and only a few considered human trafficking characterization. Perhaps this is due to a perceived simplicity of the question; it is not difficult to conjecture that areas of larger populations would likely have a worse problem with trafficking, localized poverty tends to foster exploitation of the vulnerable, and access to transit infrastructure (airports, highways, shipping ports) enables traffickers to transport their victims more easily and thereby avoid detection. However, the ability to characterize spatial characteristics that could be used to predict types of sex trafficking businesses ought to prove at least reasonably useful to trafficking reduction efforts.

With all of this in mind, there appears to be a gap in the literature when considering predictive analysis of sex trafficking business models, and the proposal herein is to seek to determine the relationship between sex trafficking business models and predictors related to cities that have federally prosecuted sex trafficking court cases. A description of the data, methodology, and findings is enclosed in the following sections.

# CHAPTER THREE: METHODOLOGY

This chapter contains an in-depth explanation of the data used in this analysis – sources, characteristics, and processing – as well as the machine learning methods utilized and the general procedure followed to get the results.

## Section One: Human Trafficking Data

Data for this analysis was gathered from human trafficking court cases. The documentation reviewed to gather the data was pulled from the Human Trafficking Data website [4], which contains a repository of the court documents and occasional articles related to the various federally-prosecuted human trafficking court cases. Cases in the database are organized by state (this repository is restricted to the United States) and concern both labor trafficking and sex trafficking. In addition to court documents and articles, general case information, such as the identity of the perpetrator, the number of victims, a short description of the case details, and a few other data points, were gathered by the investigators that established the database.

The procedure employed by the owners of the database to collect cases included searching legal databases for human trafficking court cases between the year 2000 and 2015. The court documents associated with the human trafficking court cases were then obtained by the researchers in order to populate the repository, and general information about each of the cases was collected from the court records and media reports and provided in a short summary. All of the human trafficking cases pertaining to a particular state can then be referenced by the users as needed.

Since the database provided limited details outside of the court documentation, more in-depth information regarding the specific location of the trafficking, the nature of the trafficking, and other specifics from these cases were gleaned from the documents themselves and recorded. A separate database was populated with this information on a case-by-case basis, specifically data regarding: use of hotels, use of online ads, use of violence, length and type of sentencing, business type used for trafficking, recruitment location, solicitation or incident location, etc. Some attributes were binary encoded – such as a 0 if no hotels were mentioned in the documentation or a 1 if hotels were mentioned – while others were stored as text or numeric values – such as the type of sentencing the perpetrator received or the months of the sentence.

After processing the cases and assembling the database of case details summarized from the court documentation, data specific to this research question was consolidated into a separate repository and superfluous data for this analysis – such as sentencing information and use of violence – was removed. The other information was gathered for other analyses. Additionally, since not every case included essential details for this analysis, it was necessary to further discard some cases when extracting the subset of applicable data. Specifically, if the documentation did not include references to where the incidents occurred, the case was not included into the subset.

Due to the scale of the dataset and the widely geographically-distributed nature of the data, this analysis was paired down to focus solely on one state rather than the entirety of the United States. This allowed for more easily refining the parameters of the study and helped to alleviate the uncertainty of geographically-related factors which could be problematic if all states were included in the same model. For example, of the three states with definitively enough data to train a model – California, Florida, and Texas – there were enough disparities between their population demographics, population densities, political leanings, etc., that combining them into one predictive model would likely have interfered with accuracy. It is quite likely that the predictors of human trafficking common to one state may not be common to another, which provides an excellent opportunity for further research. Therefore, the scope of this analysis was narrowed to focus on the state with the most cases and the most widely varying city demographics across the dataset, the State of California, with the assumption that this methodology could be adjusted and adapted for any other state with high enough case numbers to train a model.

The final dataset used in this analysis, based on the Human Trafficking Data [4] and derived from the documents provided therein, was thus focused on data from one individual state. It is important to note that many of the court cases listed multiple incident locations. Some cases had recorded incidents in over 18 different states and had as many as 10 different city or county locations within the State of California alone. This meant that a single case might account for multiple different entries in the dataset, although no duplicate case-location-business model incidents were included. For example, if a case recorded incidents at two separate hotels within the same city and they used the same type of trafficking business model, only one incident was recorded in this dataset so as not to be duplicative. However, all business types mentioned in the court documentation associated with each incident were recorded, so a case will have one or more businesses associated with it.

When conducting pre-processing on the data, it could be seen that certain types of business models had as few as a single case associated with it. Therefore, the cases were binned according to the type of operational business model used by the traffickers. It is also important to note that not all cases specifically mentioned the type of business model involved. Therefore, the Pimp Model was used as a default if the type of business model was not specified since the general model of at least one individual controlling the movements and actions of a victim was common to all cases not specifying a business model. This means there is a higher quantity of pimp cases compared to the other business types. The other business models listed in the court documents included: brothels, gangs, massage parlors, labor camps, escort services, streetwalking, and the specific mention of pimps. These various business models were then binned into subcategories so as to simplify the model.

Table 1 shows the original breakdown of the raw data, before any binning of the business models occurred. It can be seen that binning the business models was necessary due to some categories having extremely low incident rates, such as the massage parlors, escort services, and the single labor camp incident. Every business model mentioned in a given court case was listed in the database, which is why there are sometimes disparate business models associated together in the table. Also note that the final sums from this table will not match those used in the final analysis; some cases had to be removed later on in the data processing for not having the necessary cross-referenced attributes.

**Table 1 - Business Models from the Raw Data**

| Business Types Mentioned in Court Documentation | Total |
|---|---|
| Brothel | 6 |
| Brothel; Massage Parlor | 1 |
| Brothel; Streetwalking | 6 |
| Escort Service | 3 |
| Pimp/prostitution | 58 |
| Pimp/prostitution; Brothel; Labor Camp | 1 |
| Pimp/prostitution; Brothels; Massage Parlors | 1 |
| Pimp/prostitution; Gangs | 23 |
| Pimp/prostitution; Streetwalking | 18 |
| Totals | 117 |

The cases which were marked brothels in this dataset either specifically mentioned brothels in the documentation or they contained references to a residence or place of business where a prostitution enterprise was based rather than a transient location typical of the Pimp Model. For the California dataset, all of these businesses specifically referenced were massage parlors, while the non-business-related brothels were repurposed residences, either houses or apartments. Due to the lack of specifics in the court documentation and news articles related to each of the cases, it is unclear whether the residences were purchased specifically to be used as brothels or whether they were previously owned by members of the trafficking network and had been donated or commandeered to be used as a brothel.

The cases in the Gangs Model were distinguished by utilizing a trafficking ring organized by one or more gangs. This is separate from individual gang members engaging in trafficking, which would have been represented in one of the other business model categories. Since the geographic range of this analysis was limited to a single state, it was not considered beneficial to focus on the nature of the gangs themselves and thus they were all combined into one generic gang category. A more in-depth analysis of gang involvement in human trafficking rings, such as the characteristics of gangs involved in trafficking, would provide significant opportunity for further research. It was also observed that in multiple instances, gangs joined forces to organize a human trafficking ring even in situations when the gangs might have otherwise or normally been rivals.

It is worth noting that many gangs utilized other human trafficking business models, such as pimps or brothels, to conduct their business at a lower level while operating as a trafficking ring. As mentioned in Chapter One, the argument could be made that these cases involving gangs should have been combined with the other business models due to the overlap. However, since the gang-related cases, at least in the State of California, primarily utilized large human trafficking networks to organize their operations and only secondarily operated with the typical small-scale business models, significant value was perceived in maintaining a separate business type for the gang-related cases. This allowed for a broader view of sex trafficking as a whole which included larger enterprises rather than a narrowed scope for this analysis, though with the limitation that there were not many discrete cases in the category.

The Streetwalking Model was usually a more difficult category to characterize. Some of the cases specifically mentioned the victims streetwalking, but many others only mentioned that the victims were found on the street by police or seen on the street by family members. However, in the concerned cases, the presence of the victim on the street was used by law enforcement as an indication that the victim was being trafficking. Taking that into account, as well as the fact that the cases did not specifically mention a situation which would be better characterized by another business model, the Streetwalking Model seemed most applicable. Although this model does have a heavy overlap with the pimp category, the fact that the process for recovering streetwalkers is different from that of recovering the victims of pimps meant that there was value perceived in categorizing the cases into a separate business model rather than combining the Streetwalking and Pimp Models.

The final category, the Pimp Model, included cases where a specific pimp was mentioned and was also used as a default category when only a primary trafficker was referenced and no specific mention of the business model was made in the documentation. Oftentimes the cases mentioned transient occupancy locations, such as hotel rooms, which were reserved by the pimp and utilized for commercial sex acts. Another common characteristic of these cases was the mention of only one perpetrator in the documentation whereas it was not uncommon to have more than one perpetrator involved in the gangs and brothel models (though it is worth reiterating that a pimp business model does not necessitate a single individual acting as a pimp, multiple individuals may be involved).

A small subset of the California cases included business models that were rarely utilized, specifically a labor camp and several escort services. The three incidences of escort services were included in the pimp category since the services were being organized by an individual acting like a pimp and since these businesses usually utilize hotels rather than established residences or businesses as in the Brothel Model. The labor camp incident was included under the brothel category due to the case also mentioning an apartment being used by the trafficking victims. Once all of the cases were processed into the categories for the business models, the final breakdown of the business models is shown in Table 2.

**Table 2 - Final Business Model Breakdown**

| Businesses | Total |
|---|---|
| Brothels | 14 |
| Gangs | 21 |
| Pimps | 61 |
| Streetwalking | 18 |
| **Totals** | **114** |

## Section Two: Control Cities

For the purposes of this analysis, additional data points were required in order to train a model to predict human trafficking based on city characteristics. Once the original data from the human trafficking court case repository had been processed and cleaned, further data was then gathered for each of the cities referenced in the dataset which would then become predictors for the predictive model. However, one additional set of data was necessary before gathering predictors.

In addition to the main human trafficking locations, it was necessary to also include cities that did not have federally-prosecuted human trafficking cases associated with them to ensure the dataset was not biased. For the cities that were chosen to be "controls," or cities without human trafficking cases, this does not mean that human trafficking may not an issue in that city, but rather the assumption was made for the purposes of this analysis that the city did not have human trafficking if there was no associated court case in the Human Trafficking Data repository. Therefore, an equal number of cities with and without human trafficking cases were pulled into the dataset.

These control cities without human trafficking cases were chosen randomly. This was accomplished using a list of all the cities in the State of California. Those cities specifically mentioned in the human trafficking court documentation were removed from the list of California cities and then an equal number of cities were chosen at random from the remainder. A random selection of control cities was determined to be preferable than other methods considered as it would better enable the model to find key distinguishing characteristics for the cities with human trafficking since a random sample of control cities would presumably contain a wider variety of characteristics. After collecting all the information from the court documentation as well as incorporating the control cities into the dataset, the final breakdown of the human trafficking business models can be seen in Table 3.

**Table 3 - Final Dataset Breakdown**

| Businesses | Total |
|---|---|
| Brothels | 14 |
| Gangs | 21 |
| Pimps | 61 |
| Streetwalking | 18 |
| No Human Trafficking | 39 |
| **Totals** | **153** |

## Section Three: City Attribute Data

Once the full list of cities was obtained, the next step was to gather key attributes across the cities to be used in the classification model. A primary difficulty was finding attributes which were available for all of the chosen cities; a number of attributes had to be discarded for not having enough data. Another difficulty was the need for the data to be organized by city rather than by county or zip code. This data proved to be more difficult to find, but the value of city-level data was decided to be greater than that of county or zip code levels of data due to the widely varying size of counties and zip codes across the state.

The State of California maintains a repository of some budgetary and population data which was utilized for this analysis [29]. Data was taken from the state's repository website, which contains datasets organized by city, county, year, etc. and includes a variety of tax information, income and expenditure data, population statistics, and other characteristics for around the time corresponding with the court case timeframe, the years 2000 to 2015.

After reviewing the datasets available, the key attributes which were pulled as predictors were generally related to the following categories: city budgets spent on infrastructure and general upkeep (city maintenance costs, street light costs, etc.), city expenditures on public safety (police, fire and rescue, etc.), city income related to tourism or travel (airport income, taxes on transitory living such as hotels or short-term home rentals), tax information (income taxes, sales taxes, etc.), and city expenditures on public or community offerings (culture and leisure, community centers, libraries, etc.).

This data was also cross-referenced with census data [30] regarding population and population density, education statistics (percent of population that graduated high school and percent of population that achieved a bachelor's degree or higher), racial and age-related population demographics, housing information (median rent in the city, median home value, homeownership statistics, etc.), and income statistics (poverty rate and median household income). All of these attributes were included in the dataset as predictors for the model.

## Section Four: Data Cleaning

After assembling all of the data, both the court case information as well as the city attribute data, it was necessary to clean and refine the dataset prior to working with the models. The first issue encountered was the quantity of missing information from the data gathered from the State of California's repository. Multiple attributes had to be discarded due to having too many missing data points. Those discarded included aircraft movements, airport revenue and expenditures, total transit costs, transient occupancy information, public safety data related to fire and rescue and emergency medical services, and library expenditures. Once these data points were removed, the number of null values in the dataset was significantly reduced, thus improving the likely outcomes of the models. Figure 1 shows the proportion of missing values in a subset of the original predictors; additional predictors were not included in the image as they had no missing values across the list of cities.

Another issue with the dataset was skewing of the predictors. The data was preprocessed in order to normalize the predictors. The predictors were unskewed by taking the logarithm of each value plus one (since the logarithm of zero is undefined). All values were then scaled to between zero and one in order to assist the multinomial logistic regression model. This reduced the skewing of the data attributes using minimal processing and further refined the dataset in preparation for training and testing the machine learning models.

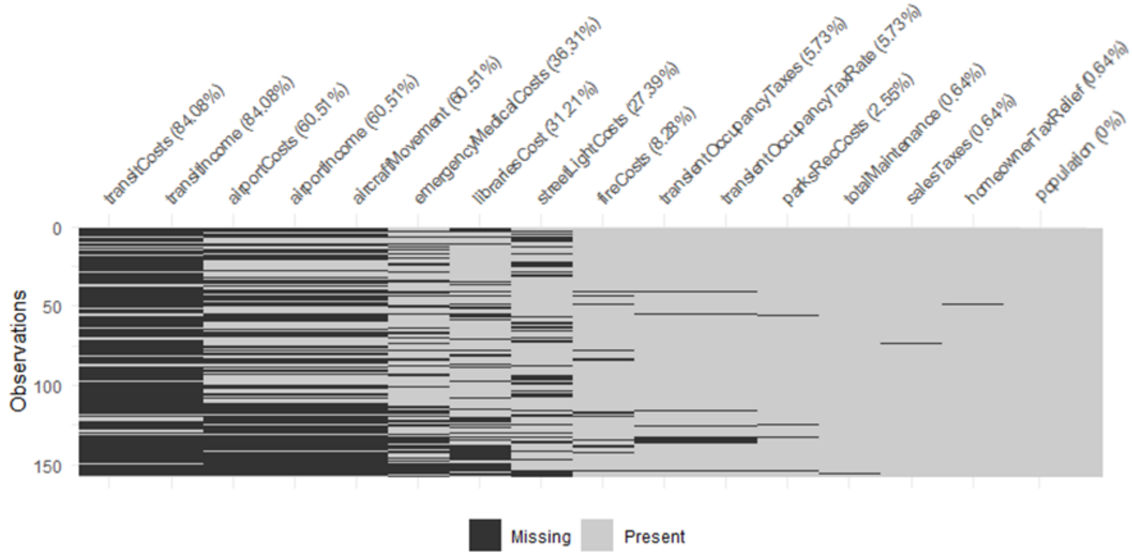


**Figure 1 - Missingness Plot**

The multicollinearity between all of the predictors was then analyzed using correlation plots and Pearson correlation values. A number of predictors were significantly correlated, as can be seen in Figure 2, where the total correlation plot was cropped to show the most correlated predictors; these predictors thus had to be removed from the dataset so as to avoid interference with the predictive models. Figure 3 shows the final correlation of the remaining predictors in the dataset. The final dataset did still have a few moderately correlated predictors, but the most severe correlation values remained below a threshold of ±0.9 and were not considered to be correlated enough to interfere with the model predictions. Therefore, no additional predictors were removed from the dataset due to multicollinearity.

| | population | totalMaintenance | totalExpenditures | propertyTaxes | salesTaxes | transientOccupancyTaxes | totalTaxes | homeownerTaxRelief | policeCosts | fireCosts | streetLightCosts | publicSafetyCosts | parksRecCosts |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| totalMaintenance | 0.99 | | | | | | | | | | | | |
| totalExpenditures | 1 | 0.98 | | | | | | | | | | | |
| propertyTaxes | 1 | 0.98 | 1 | | | | | | | | | | |
| salesTaxes | 0.99 | 0.99 | 0.99 | 0.99 | | | | | | | | | |
| transientOccupancyTaxes | 0.96 | 0.93 | 0.97 | 0.96 | 0.95 | | | | | | | | |
| totalTaxes | 1 | 0.98 | 1 | 1 | 0.99 | 0.97 | | | | | | | |
| homeownerTaxRelief | 1 | 0.99 | 0.99 | 1 | 0.99 | 0.96 | 1 | | | | | | |
| policeCosts | 1 | 0.98 | 1 | 1 | 0.99 | 0.97 | 1 | 1 | | | | | |
| fireCosts | 1 | 0.99 | 0.99 | 0.99 | 1 | 0.96 | 1 | 1 | 0.99 | | | | |
| streetLightCosts | 0.99 | 0.98 | 0.99 | 0.99 | 0.98 | 0.96 | 0.99 | 0.99 | 0.99 | 0.99 | | | |
| publicSafetyCosts | 1 | 0.98 | 1 | 1 | 0.99 | 0.97 | 1 | 1 | 1 | 1 | 0.99 | | |
| parksRecCosts | 1 | 0.98 | 1 | 1 | 0.99 | 0.97 | 1 | 1 | 1 | 0.99 | 0.99 | 1 | |
| cultureCosts | 1 | 0.98 | 1 | 1 | 0.99 | 0.96 | 1 | 1 | 1 | 1 | 0.99 | 1 | 1 |

**Figure 2 - Correlation Plot of Highly Correlated Predictors**

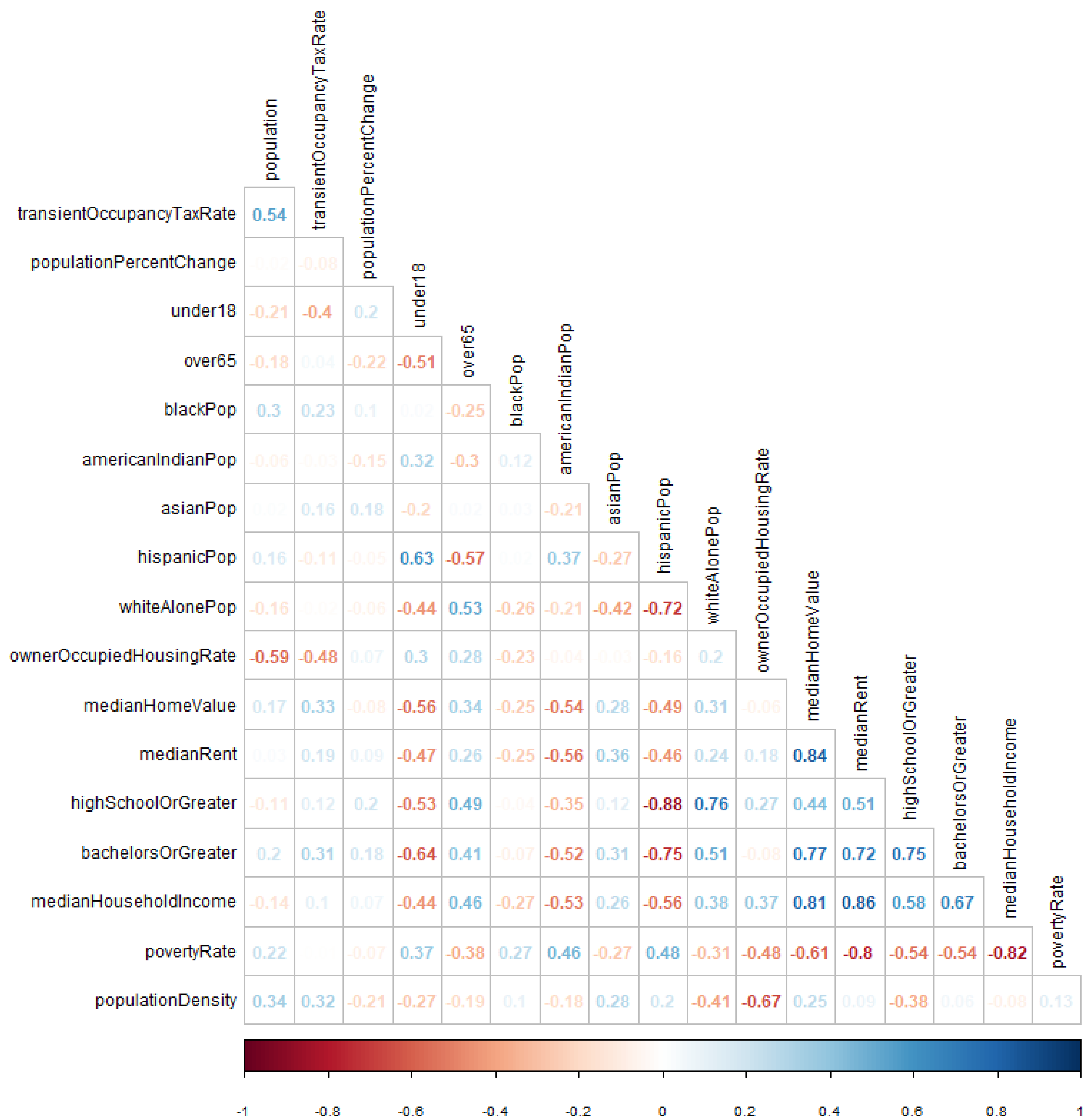

**Figure 3 - Correlation Plot of Final Dataset Without Highly Correlated Predictors**

Table 4 provides a full list of the attributes used as inputs to the models and includes a description of each attribute in the final dataset. The source of the attribute is also listed, either the State of California's data repository [29] or the Census Bureau data [30].

**Table 4 - Final Dataset Attributes, Without Outcome Columns**

| Attribute | Description |
|---|---|
| ID | Unique identifier for each data entry |
| City | City where the commerce occurred, not necessarily the location where the victim was recruited or discovered |
| transientOccupancyTaxRate | Total transient occupancy tax rate from FY17, based on the CA repository |
| populationPercentChange | The percent change in population between the 2000 and the 2010 Census Data |
| under18 | The percent of the population that is under 18 years of age, based on the 2010 Census Data |
| over65 | The percent of the population that is over 65 years of age, based on the 2010 Census Data |
| blackPop | The percent of the population that is Black, based on the 2010 Census |
| americanIndianPop | The percent of the population that is American Indian, based on the 2010 Census Data |
| asianPop | The percent of the population that is Asian, based on the 2010 Census |
| hispanicPop | The percent of the population that is Hispanic, based on the 2010 Census Data |
| whiteAlonePop | The percent of the population that is White, not including the Hispanic percentage, based on the 2010 Census Data |
| ownerOccupiedHousingRate | The percentage of the population living in houses that they owned, based on the 2010 Census Data |
| medianHomeValue | The median home value in the city, based on the 2010 Census Data |
| medianRent | The median monthly rent, based on the 2010 Census Data |
| highSchoolOrGreater | The percentage of the population that achieved a high school diploma or greater, based on the 2010 Census Data |
| bachelorsOrGreater | The percentage of the population that achieved a bachelor's degree or greater, based on the 2010 Census Data |
| medianHouseholdIncome | The median household income, based on the 2010 Census Data |
| povertyRate | The estimated poverty rate, based on the 2010 Census Data |
| populationDensity | The population density, based on the 2010 Census Data |

To establish the final datasets to be used with the predictive models, the human trafficking business models were used as the outcomes. Several different datasets were established in order to analyze the behavior of the predictive models. One data table was made with a single text-based outcome column containing a name for each of the business models (gangs, brothels, pimps, streetwalking, and no human trafficking), which was used to determine each predictive model's overall ability to classify the test data into each of the five categories, although this dataset required algorithms that could analyze multi-class data.

The predictive models were also exercised using dummied outcomes, where the presence of each model was coded as 1 and the absence of the model was coded as 0 for each of the different business models. For example, if the model was analyzing the presence of gangs, the outcome column contained a 1 for each case involving gangs and a 0 for all the other cases, whether they belonged to a different business model or the control group. This resulted in five subsets of data as the control group was also dummied against the combined presence of the human trafficking cases. A final group of datasets was developed to analyze how the predictive models behaved when only comparing a single business model to the control group containing no sex trafficking. This resulted in four additional subsets of data, one for each of the sex trafficking business models containing cases specific to that business model as well as the no trafficking data. These four datasets were significantly smaller than the previous ones mentioned, but the two categories of data subsets allowed for a comparison of how the predictive models distinguished between the different business models when combined in the same dummied dataset versus how the predictive models behaved without the noise of the other business models.

After establishing the datasets, the final step for preprocessing of the data was to balance the datasets between the different sex trafficking business models. As can be seen from Table 2, the quantity of cases in each category vary widely, and thus the datasets for each of the sex trafficking business models were heavily imbalanced. This imbalance of data affects the outcomes of the predictive models, so Synthetic Minority Oversampling Technique (SMOTE) was utilized to create more balanced dummied datasets across the different business models. SMOTE is a method to synthetically oversample the minority case in a dataset to even out the dataset and improve true positive outcomes.

Results from using SMOTE on the dataset are captured in Table 5, where the Full Dataset columns depict the ratios of the business model cases on the right against the remainder of the dataset on the left (all other business models and the no trafficking data), before and after utilizing SMOTE. The No Trafficking : Business Model columns depict the ratio of the control group on the left against each of the four business models on the right before and after utilizing SMOTE. It can be seen that the synthetic oversampling technique did not perform well on the smaller datasets, specifically the Gangs and Streetwalking Models in the rightmost columns, but since the ratios were not too disparate for these two models this was not considered to be impactful.

**Table 5 - SMOTE Outcomes**

| Business | Full Dataset | | No Trafficking : Business Model | |
|---|---|---|---|---|
| | Before | After | Before | After |
| Brothels | 110:12 | 110:100 | 28:12 | 28:23 |
| Gangs | 104:18 | 104:95 | 28:18 | 28:18 |
| Pimps | 38:84 | 70:84 | 28:84 | 78:84 |
| Streetwalking | 101:21 | 101:97 | 28:21 | 28:21 |
| No Human Trafficking | 94:28 | 94:76 | | |

## Section Five: Predictive Models

Three predictive models were chosen for this analysis, a logistic regression model, a multinomial logistic regression model, and a random forest model. The logistic regression model is a parametric classification model, providing binary categorical outcomes based on the input data and it fits to a sigmoidal curve; this model was chosen to determine linearity. Multinomial logistic regression is the multi-class expansion of logistic regression, utilizing maximum likelihood estimation to categorize inputs; the primary reason this model was chosen is that it is not limited to a binomial dataset, and thus should be better able to characterize the five different business models. A random forest utilizes a host of decision trees to survey and analyze an ensemble of classifications from the various trees; the model was utilized to narrow down the list of predictors and find the key attributes which can be correlated with sex trafficking business models.

The dataset was split into training and testing sets, with a randomly chosen 80 percent going to training the models and 20 percent to testing. Randomly splitting the dataset was accomplished using randomized data partitions. Leave One Out Cross-

Validation was also utilized with the random forest model to help compensate for the small size of the dataset and to better stabilize the model outcomes across the various trials. As previously mentioned, SMOTE was also used on the dummied datasets to better balance the minority cases to improve model performance.

After running the various models on the full set of predictors, the key attributes reported by the random forest model were analyzed to find a subset of the predictors which could be entered back into all of the models. This was done to remove the irrelevant predictors and to narrow down the list of the key predictors which could ultimately be used for further analysis. Several iterations of this process were conducted, and each time the lowest-rated attributes were removed, the workspace was cleared to reset the models, and the refined datasets were run through the model again. In addition to analyzing model outcomes when removing the lowest-rated attributes, the models were also analyzed when the highest-rated attribute was removed in order to determine the secondary predictors.

To analyze and compare the results of the various models, several different metrics were calculated. The area under the curve (AUC) was calculated as a performance metric (specifically the AUC of the receiver operating characteristic curve) as well as the overall accuracy of the models on the testing datasets. Sensitivity (or recall), precision, and specificity were also calculated as metrics to determine the overall performance of the models. Receiver operating characteristic (ROC) curves themselves were not directly utilized due to the use of a multinomial model as well as the fact that the small size of the dataset would affect the granularity of the curve.

The three different models were run on the full datasets as well as the dummied datasets so as to compare how well the models performed under all conditions. The primary metrics used for comparison were sensitivity, precision, and AUC, which a focus on highlighting true positive identification over true negative. Secondary metrics for comparison were predictive accuracy and specificity.

**Equation 1 - Sensitivity (Recall)**

$$Sensititivy\ (recall) = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

**Equation 2 - Precision**

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

**Equation 3 - Specificity**

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

As previously mentioned, tuning the set of predictors using the random forest model was accomplished by iteratively running the random forest and multinomial logistic regression models on the datasets, recording the key predictors as reported by the random forest, and then restricted the database to remove the lowest valued predictors. This process was repeated numerous times and the results were compared. Five primary trials with this technique adequately summarize the behaviors observed, and thus the results will focus on these five iterations.

The first of the five iterations was conducted on the datasets with all predictors present, and this was performed with both the combined dataset with all sex trafficking business models in one outcome column as well as the dummied datasets. Eight of the most important predictors from each of these first trials were then made into subsets of input data which became the inputs for the second iteration of the models. A further three of the most important predictors from each of these secondary trials were then used on a third iteration and the resulting metrics were recorded. Other variations of this procedure did not significantly impact the outcomes of the model, so these are the three being considered. Since the total population predictor scored quite highly across all of the models, it was removed for iteration four in order to analyze the model performances in its absence. Iteration five, similarly, removed all racial demographics and the total population.

# CHAPTER FOUR: RESULTS

The logistic regression model was run first in order to determine the linearity of the dataset. From the model accuracy results when using the dummied datasets with all business models present, the accuracies were substantially lower than the other predictive models and were thus not analyzed further in the different iterations to record the other comparison metrics. However, the logistic regression model performed better on the dummied datasets that only compared a single business model to the control group, with accuracies comparable to the other predictive models; therefore, all five of the comparison metrics – AUC, Specificity, Precision, Sensitivity, and Predictive Accuracy – were collected to analyze against the results from the other predictive models. Since the logistic regression model can only utilize binomial data, only the dummied datasets were used with this model.

Both the random forest and the multinomial logistic regression models performed relatively poorly when trained on the multinomial dataset with the text-based outcome column containing all of the business models. The All Business Models columns in Table 6 show the average accuracy of the predictive models for the five different dummied datasets containing all business models, as well as the dataset with all sex trafficking business models combined into one outcome column in the All Businesses row; the four different datasets with the single business model and the control group are shown in the Single Business Model & No Trafficking columns.

The results of the Single Business Model & No Trafficking columns appear to show generally higher predictive accuracies than the All Business Models columns. However, the accuracies for the Brothels and Streetwalking business models were 100% in four out of the five iterations, which, given the small size of the dataset and the fact that the corresponding AUCs were also 1, there is reason to suspect the models may be overfit. When working with smaller datasets such as this, overfitting can become a serious issue. One way to detect overfitting is when a model has repeatedly very high accuracy results (at or near 100% predictive accuracy), indicating the model has learned the sample dataset too specifically and will not perform well on other similar data. Despite the possible overfitting, the results of the models are still useful. For the other two business models, Pimps and Gangs, the predictive accuracies are reasonable and not excessively high, so the apprehension of overfitting is low. Results from the Pimps model had better accuracies than that of the All Business Model iterations, but in all other cases there was either the suspicion of overfitting or the All Business Model averages were higher.

For additional results from the different iterations, the following tables show the metric comparisons for the three predictive models across all of the datasets. Table 7 shows the mean, maximum, minimum, and standard deviation of the predictive accuracy across the five iterations of running the models. Table 8 shows the Area Under the Curve for the five iterations. Tables 9, 10, and 11 show the Specificity, Precision, and

Sensitivity across the five iterations, as calculated from the confusion matrices of the prediction against the test data using Equations 1, 2, and 3.

**Table 6 - Average Predictive Accuracy of All Models**

| | All Business Models | | | Single Business Model & No Trafficking | | |
|---|---|---|---|---|---|---|
| | LR Accuracy | MLR Accuracy | RF Accuracy | LR Accuracy | MLR Accuracy | RF Accuracy |
| Brothels | 41% | 86% | 89% | 98% | 98% | 100% |
| Pimps | 60% | 74% | 75% | 79% | 81% | 87% |
| Gangs | 40% | 85% | 86% | 77% | 80% | 71% |
| Streetwalking | 73% | 82% | 79% | 88% | 96% | 98% |
| No Trafficking | 39% | 87% | 87% | | | |
| **All Businesses** | | **54%** | **57%** | | | |

**Table 7 - Accuracy Across Five Iterations**

| | | All Business Models | | | | Single Business Model & No Trafficking | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Max | Min | Mean | Standard Deviation | Max | Min | Mean | Standard Deviation |
| Multinomial Logistic Regression | Brothels | 93% | 75% | 87% | 6% | 100% | 89% | 98% | 4% |
| | Pimps | 79% | 71% | 74% | 3% | 85% | 77% | 81% | 4% |
| | Gangs | 93% | 75% | 85% | 6% | 92% | 64% | 80% | 11% |
| | Streetwalking | 89% | 68% | 84% | 8% | 100% | 90% | 96% | 5% |
| | No Trafficking | 93% | 75% | 87% | 7% | | | | |
| | **All Businesses** | **61%** | **54%** | **57%** | **3%** | | | | |
| Random Forest | Brothels | 93% | 89% | 90% | 2% | 100% | 100% | 100% | 0% |
| | Pimps | 82% | 68% | 76% | 5% | 88% | 81% | 87% | 3% |
| | Gangs | 86% | 86% | 86% | 0% | 82% | 64% | 71% | 9% |
| | Streetwalking | 79% | 79% | 79% | 0% | 100% | 90% | 98% | 4% |
| | No Trafficking | 89% | 82% | 88% | 3% | | | | |
| | **All Businesses** | **61%** | **54%** | **57%** | **2%** | | | | |
| Logistic Regression | Brothels | | | | | 100% | 89% | 98% | 4% |
| | Pimps | | | | | 85% | 73% | 79% | 4% |
| | Gangs | | | | | 91% | 55% | 77% | 13% |
| | Streetwalking | | | | | 90% | 80% | 88% | 4% |

**Table 8 - AUC Across Five Iterations**

| | | All Business Models | | | | Single Business Model & No Trafficking | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Max | Min | Mean | Standard Deviation | Max | Min | Mean | Standard Deviation |
| Multinomial Logistic Regression | Brothels | 0.94 | 0.75 | 0.84 | 0.07 | 1.00 | 0.93 | 0.99 | 0.03 |
| | Pimps | 0.82 | 0.73 | 0.77 | 0.03 | 0.99 | 0.86 | 0.93 | 0.04 |
| | Gangs | 0.77 | 0.51 | 0.66 | 0.09 | 0.96 | 0.64 | 0.80 | 0.11 |
| | Streetwalking | 0.65 | 0.52 | 0.60 | 0.05 | 1.00 | 0.83 | 0.96 | 0.07 |
| | No Trafficking | 0.97 | 0.92 | 0.96 | 0.02 | | | | |
| | **All Businesses** | **0.81** | **0.72** | **0.75** | **0.03** | | | | |
| Random Forest | Brothels | 0.96 | 0.94 | 0.96 | 0.01 | 1.00 | 1.00 | 1.00 | 0.00 |
| | Pimps | 0.89 | 0.80 | 0.85 | 0.03 | 0.98 | 0.90 | 0.95 | 0.03 |
| | Gangs | 0.87 | 0.64 | 0.71 | 0.08 | 0.96 | 0.75 | 0.86 | 0.07 |
| | Streetwalking | 0.81 | 0.73 | 0.76 | 0.03 | 1.00 | 1.00 | 1.00 | 0.00 |
| | No Trafficking | 0.99 | 0.95 | 0.96 | 0.02 | | | | |
| | **All Businesses** | **0.75** | **0.59** | **0.70** | **0.06** | | | | |
| Logistic Regression | Brothels | | | | | 1.00 | 0.86 | 0.97 | 0.06 |
| | Pimps | | | | | 0.98 | 0.85 | 0.92 | 0.05 |
| | Gangs | | | | | 0.96 | 0.57 | 0.79 | 0.14 |
| | Streetwalking | | | | | 1.00 | 0.95 | 0.99 | 0.02 |

**Table 9 - Specificity Across Five Iterations**

| | | All Business Models | | | | Single Business Model & No Trafficking | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Max | Min | Mean | Standard Deviation | Max | Min | Mean | Standard Deviation |
| Multinomial Logistic Regression | Brothels | 1.00 | 0.93 | 0.97 | 0.04 | 1.00 | 0.88 | 0.98 | 0.05 |
| | Pimps | 0.62 | 0.53 | 0.57 | 0.03 | 0.64 | 0.55 | 0.59 | 0.04 |
| | Gangs | 0.18 | 0.08 | 0.11 | 0.04 | 0.33 | 0.13 | 0.22 | 0.09 |
| | Streetwalking | 0.90 | 0.88 | 0.89 | 0.01 | 1.00 | 0.88 | 0.95 | 0.06 |
| | No Trafficking | 1.00 | 0.94 | 0.98 | 0.03 | | | | |
| Random Forest | Brothels | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | Pimps | 0.67 | 0.50 | 0.60 | 0.06 | 0.70 | 0.58 | 0.68 | 0.05 |
| | Gangs | 0.86 | 0.86 | 0.86 | 0.00 | 0.78 | 0.64 | 0.71 | 0.06 |
| | Streetwalking | 0.88 | 0.88 | 0.88 | 0.00 | 1.00 | 0.88 | 0.95 | 0.06 |
| | No Trafficking | 1.00 | 0.90 | 0.98 | 0.04 | | | | |
| Logistic Regression | Brothels | | | | | 0.50 | 0.50 | 0.50 | 0.00 |
| | Pimps | | | | | 0.64 | 0.50 | 0.57 | 0.05 |
| | Gangs | | | | | 0.88 | 0.60 | 0.75 | 0.11 |
| | Streetwalking | | | | | 0.88 | 0.78 | 0.86 | 0.04 |

**Table 10 - Precision Across Five Iterations**

| | | All Business Models | | | | Single Business Model & No Trafficking | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Max | Min | Mean | Standard Deviation | Max | Min | Mean | Standard Deviation |
| Multinomial Logistic Regression | Brothels | 0.40 | 0.00 | 0.20 | 0.18 | 1.00 | 1.00 | 1.00 | 0.00 |
| | Pimps | 0.93 | 0.83 | 0.89 | 0.04 | 1.00 | 0.93 | 0.97 | 0.03 |
| | Gangs | 1.00 | 0.00 | 0.41 | 0.39 | 1.00 | 0.50 | 0.90 | 0.20 |
| | Streetwalking | 0.13 | 0.00 | 0.04 | 0.05 | 1.00 | 1.00 | 1.00 | 0.00 |
| | No Trafficking | 0.78 | 0.50 | 0.69 | 0.11 | | | | |
| Random Forest | Brothels | 0.50 | 0.40 | 0.42 | 0.04 | 1.00 | 1.00 | 1.00 | 0.00 |
| | Pimps | 0.94 | 0.81 | 0.90 | 0.05 | 1.00 | 1.00 | 1.00 | 0.00 |
| | Gangs | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.50 | 0.45 |
| | Streetwalking | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | No Trafficking | 0.70 | 0.63 | 0.69 | 0.03 | | | | |
| Logistic Regression | Brothels | | | | | 1.00 | 1.00 | 1.00 | 0.00 |
| | Pimps | | | | | 1.00 | 0.93 | 0.97 | 0.03 |
| | Gangs | | | | | 1.00 | 0.00 | 0.60 | 0.49 |
| | Streetwalking | | | | | 1.00 | 1.00 | 1.00 | 0.00 |

**Table 11 - Sensitivity Across Five Iterations**

| | | All Business Models | | | | Single Business Model & No Trafficking | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Max | Min | Mean | Standard Deviation | Max | Min | Mean | Standard Deviation |
| Multinomial Logistic Regression | Brothels | 1.00 | 0.00 | 0.60 | 0.49 | 1.00 | 0.50 | 0.90 | 0.20 |
| | Pimps | 0.79 | 0.63 | 0.72 | 0.05 | 0.79 | 0.74 | 0.76 | 0.03 |
| | Gangs | 0.50 | 0.00 | 0.30 | 0.24 | 0.75 | 0.25 | 0.50 | 0.22 |
| | Streetwalking | 0.33 | 0.00 | 0.13 | 0.16 | 1.00 | 0.67 | 0.87 | 0.16 |
| | No Trafficking | 1.00 | 0.86 | 0.94 | 0.07 | | | | |
| Random Forest | Brothels | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | Pimps | 0.79 | 0.68 | 0.74 | 0.05 | 0.84 | 0.74 | 0.82 | 0.04 |
| | Gangs | 0.00 | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 0.30 | 0.24 |
| | Streetwalking | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.67 | 0.93 | 0.13 |
| | No Trafficking | 1.00 | 0.71 | 0.94 | 0.11 | | | | |
| Logistic Regression | Brothels | | | | | 1.00 | 0.50 | 0.90 | 0.20 |
| | Pimps | | | | | 0.79 | 0.63 | 0.73 | 0.06 |
| | Gangs | | | | | 0.75 | 0.00 | 0.35 | 0.34 |
| | Streetwalking | | | | | 0.67 | 0.33 | 0.60 | 0.13 |

Since it can be difficult to analyze results when looking at averages across the five iterations, Tables 12 and 13 provide additional comparisons pulled from the same results data. Table 12 is a comparison of which model performed better with each iteration and each business model, though exclusively with the All Business Model datasets since a direct comparison is more challenging when considering all three predictive models. If the outcomes were equivalent between the random forest and the multinomial logistic regression models, there is an equal sign in the cell. Otherwise, the better of the two models is shown with RF for random forest and MLR for multinomial logistic regression. These outcomes were compared across all five metrics, and the most important metrics are shaded in grey. The Overall column shows the general trend in the outcomes, looking to the AUC, Specificity, and Precision for an initial assessment of the better predictive model and considering the Predictive Accuracy and Sensitivity for secondary assessments.

In general, the multinomial logistic regression model appears to better characterize the datasets, although there is some variation across the different sex trafficking business models and there is no definitive majority across the results. The random forest generally performed better for predicting the Brothel and Pimp business models, results were mixed for the gangs although the random forest was marginally better, and for the Streetwalking, No Trafficking, and All Business models, the multinomial logistic regression model performed better.

**Table 12 - RF Versus MLR Results for the All Business Models Datasets**

| | | Accuracy | AUC | Specificity | Precision | Sensitivity | Overall |
|---|---|---|---|---|---|---|---|
| Brothels | 1 | = | RF | = | = | = | = |
| | 2 | = | RF | RF | RF | RF | RF |
| | 3 | = | RF | RF | RF | RF | RF |
| | 4 | = | RF | = | = | = | = |
| | 5 | RF | RF | = | RF | = | RF |
| Pimps | 1 | RF | RF | RF | RF | RF | RF |
| | 2 | RF | RF | RF | RF | RF | RF |
| | 3 | RF | RF | RF | RF | = | RF |
| | 4 | = | RF | = | = | = | = |
| | 5 | MLR | MLR | MLR | MLR | MLR | MLR |
| Gangs | 1 | RF | RF | RF | MLR | MLR | RF |
| | 2 | RF | RF | RF | = | = | RF |
| | 3 | = | RF | RF | = | = | RF |
| | 4 | MLR | MLR | RF | MLR | MLR | MLR |
| | 5 | MLR | MLR | RF | MLR | MLR | MLR |
| Streetwalking | 1 | RF | RF | MLR | MLR | MLR | MLR |
| | 2 | MLR | RF | MLR | MLR | MLR | MLR |
| | 3 | MLR | RF | MLR | = | = | MLR |
| | 4 | MLR | RF | MLR | = | = | MLR |
| | 5 | MLR | RF | MLR | = | = | MLR |
| No Trafficking | 1 | RF | = | = | RF | = | RF |
| | 2 | MLR | RF | = | MLR | = | MLR |
| | 3 | = | MLR | RF | MLR | RF | MLR |
| | 4 | MLR | MLR | = | MLR | = | MLR |
| | 5 | RF | RF | MLR | RF | MLR | RF |
| All Businesses | 1 | RF | MLR | | | | MLR |
| | 2 | MLR | MLR | | | | MLR |
| | 3 | = | MLR | | | | MLR |
| | 4 | RF | MLR | | | | MLR |
| | 5 | = | MLR | | | | MLR |

Another item of interest that can be seen from these results is that, for the Pimps and Gangs models, the multinomial logistic regression model performed better when the population data was removed from the datasets in iterations four and five. This does not appear to be a trend across the entire dataset, but it is important to note since it is an obvious way to see that the different predictive models may perform better on different subsets of the data. These analyses were not utilized for the Single Business Model & No Trafficking datasets since all three of the predictive models were utilized, and since finding the best predictive model is not a primary focus of this analysis.

From Table 13, it can be seen that the random forest model generally performed best on the first iteration, when the full dataset was used. Note Specificity, Precision, and Sensitivity were not analyzed for the All Businesses model due to the limitations of generating confusion matrices when the data is multinomial. Thus the primary metric for the All Businesses model is the AUC and the secondary metric is the predictive Accuracy. The multinomial logistic regression model performed best on the first and fifth iterations, where the full dataset and the subset with no population metrics were used. These results indicate that tuning the predictors did not substantially improve the results of either model, but that the most effective subset of data was generally the one without any population demographics.

**Table 13 – Comparison of the Best Iteration**

| | | All Business Models | | | | | | Single Business Model & No Trafficking | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | AUC | Specificity | Precision | Sensitivity | Mode | Accuracy | AUC | Specificity | Precision | Sensitivity | Mode |
| Multinomial Logistic Regression | Brothels | 3 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Pimps | 5 | 5 | 5 | 5 | 3 | 5 | 2 | 3 | 2 | 2 | 2 | 2 |
| | Gangs | 5 | 5 | 2 | 5 | 1 | 5 | 4 | 3 | 2 | 1 | 1 | 1 |
| | Streetwalking | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | No Trafficking | 2 | 1 | 1 | 2 | 1 | 1 | | | | | | |
| | All Businesses | 2 | 1 | | | | | | | | | | |
| Random Forest | Brothels | 3 | 2 | 1 | 3 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Pimps | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| | Gangs | 1 | 3 | 1 | 1 | 1 | 1 | 3 | 2 | 3 | 3 | 3 | 3 |
| | Streetwalking | 1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | No Trafficking | 1 | 2 | 1 | 1 | 1 | 1 | | | | | | |
| | All Businesses | 3 | 1 | | | | | | | | | | |
| Logistic Regression | Brothels | | | | | | | 1 | 1 | 1 | 1 | 1 | 1 |
| | Pimps | | | | | | | 5 | 3 | 5 | 2 | 1 | 5 |
| | Gangs | | | | | | | 4 | 3 | 4 | 1 | 4 | 4 |
| | Streetwalking | | | | | | | 1 | 1 | 1 | 1 | 1 | 1 |

To cross-correlate the results from Table 13 with the key predictors, the best predictors resulting from the iteration listed in the Mode column were extracted. This means that for each business model (Brothels, Pimps, Gangs, Streetwalking, and No Trafficking) and for each predictive model (logistic regression, random forest and multinomial logistic regression), the most valuable predictors based on results from the random forest model were taken from the iterations listed in the Mode columns in Table 13. However, there were commonalities between the different iterations, and, if a predictor was among the best five for several different iterations, the predictor was also included in the subset of best predictors. Table 14 summarizes the best predictors and depicts which business models utilized them. This table also merges the results from the All Business Models iterations with the Single Business Model & No Trafficking iterations since separating the two was no longer necessary and there was significant overlap in the best predictors chosen despite the different methodologies.

**Table 14 - Best Performing Predictors**

| | Brothels | Pimps | Streetwalking | None | Gangs |
|---|---|---|---|---|---|
| americanIndianPop | ✓ | | | | |
| asianPop | | | ✓ | | ✓ |
| bachelorsOrGreater | ✓ | | | | |
| blackPop | | | ✓ | | |
| highSchoolOrGreater | ✓ | | | | |
| hispanicPop | ✓ | | ✓ | | |
| medianHomeValue | ✓ | | | | ✓ |
| medianHouseholdIncome | ✓ | | | | |
| medianRent | | | | | ✓ |
| over65 | | ✓ | | | |
| ownerOccupiedHousingRate | | | | ✓ | |
| population | | ✓ | ✓ | ✓ | ✓ |
| populationDensity | ✓ | ✓ | | ✓ | |
| populationPercentChange | | ✓ | | | ✓ |
| transientOccupancyTaxRate | | ✓ | ✓ | | |
| whiteAlonePop | ✓ | | | | ✓ |

A comparison of these predictors for the presence and absence of the various business models was then necessary to determine what kind of relationship there is between the business model and the data in the predictor. Each of the predictors in Table 14 was split into two sets based on the presence or absence of the concerned business model. In other words, the Asian Population predictor was split into two groups for the Streetwalking Model – the data for the streetwalking cases in one group and the No Trafficking data in the other group – as well as two groups for the Gangs Model. For the ownerOccupiedHousingRate, the Population, and the populationDensity predictors, the two groups were composed of all of the trafficking business models in one group and the No Trafficking data in the other group since the predictors were key for the No Trafficking dataset or for the majority of the different business models. The Student's T-Test was then employed on each of the split datasets to determine if the differences between the two groups was statistically significant. The resulting p-values for each predictor are listed in Table 15.

With the results from Table 15 in consideration, all of the predictors with p-values greater than 0.05 are considered to be not statistically significant. Therefore, the statistically significant predictors are: the percent of the population with a bachelor's degree or greater, the percent of the population with a high school diploma or greater, the median home value, the median household income, the percent of the population that is white, the percent of the population over the age of 65, the transient occupancy tax rate, the percent of the population that is black, the percent of houses that are occupied by their owners, the population, and the population density, all showed in bold in Table 15. This is not to say that the other predictors are not of value, but their value cannot be quantified as being statistically significant for this analysis.

**Table 15 - P-Values for Each Key Predictor**

| | Predictor | p-value |
|---|---|---|
| Brothels | americanIndianPop | 0.3573961 |
| | **bachelorsOrGreater** | **0.03148731** |
| | **highSchoolOrGreater** | **0.01922695** |
| | hispanicPop | 0.1133503 |
| | **medianHomeValue** | **0.0138436** |
| | **medianHouseholdIncome** | **3.34563E-05** |
| | **whiteAlonePop** | **0.000663655** |
| Pimps | **over65** | **0.001726693** |
| | populationPercentChange | 0.1260379 |
| | **transientOccupancyTaxRate** | **1.58217E-05** |
| Gangs | asianPop | 0.8642899 |
| | medianHomeValue | 0.2072762 |
| | medianRent | 0.6913434 |
| | populationPercentChange | 0.399926 |
| | whiteAlonePop | 0.1074769 |
| Streetwalking | asianPop | 0.07807507 |
| | **blackPop** | **7.05765E-05** |
| | hispanicPop | 0.3101495 |
| | **transientOccupancyTaxRate** | **8.63525E-06** |
| No Trafficking | **ownerOccupiedHousingRate** | **9.153E-09** |
| All Businesses | **population** | **6.523E-13** |
| | **populationDensity** | **0.00001143** |

Table 16 summarizes the maximum, minimum, mean, and standard deviation between the predictors, split by the presence and absence of the concerned sex trafficking business model referenced from Table 14. In addition, Figures 4-15 are histograms comparing the presence and absence of the aforementioned business model or models. As previously mentioned, for the ownerOccupiedHousingRate, the Population, and the populationDensity predictors, the two groups present in the histograms were all of the trafficking business models in one group and the No Trafficking data in the other group. For the rest of the histograms, only the business model mentioned is present in the dataset and is compared to the No Trafficking data.

**Table 16 - Statistical Differences Between the Presence and Absence of the Trafficking Business Models**

| | | Brothels | | | | | Pimps | | Streetwalking | | No Trafficking | All Businesses | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | bachelorsOrGreater | highSchoolOrGreater | medianHomeValue | medianHouseholdIncome | whiteAlonePop | over65 | transientOccupancyTaxRate | blackPop | transientOccupancyTaxRate | ownerOccupiedHousingRate | population | populationDensity |
| Trafficking | Max | 37.4% | 92.5% | $636,900 | $84,557 | 70.6% | 22.1% | $15.00 | 23.8% | $15.00 | 72.3% | 4040079 | 18226 |
| | Min | 2.6% | 36.4% | $164,900 | $34,426 | 1.7% | 7.5% | $7.50 | 2.1% | $8.00 | 19.1% | 4216 | 2444 |
| | Mean | 27.0% | 76.7% | $458,433 | $60,172 | 26.0% | 12.7% | $11.50 | 8.4% | $12.31 | 47.3% | 1049083 | 6195 |
| | Standard Deviation | 9.4% | 12.2% | $164,901 | $10,163 | 17.2% | 2.2% | $1.86 | 4.9% | $1.94 | 8.1% | 1345800 | 3347 |
| No Trafficking | Max | 82.6% | 100.0% | $2,000,000 | $223,217 | 89.1% | 48.7% | $14.00 | 17.5% | $14.00 | 96.9% | 208297 | 19291 |
| | Min | 6.4% | 47.3% | $183,700 | $42,447 | 1.5% | 7.3% | $4.00 | 0.0% | $4.00 | 27.2% | 1243 | 822 |
| | Mean | 36.5% | 86.1% | $700,330 | $95,647 | 48.5% | 16.7% | $9.72 | 3.1% | $9.72 | 62.4% | 37510 | 3600 |
| | Standard Deviation | 22.6% | 12.1% | $546,644 | $47,408 | 25.1% | 7.6% | $1.95 | 3.9% | $1.95 | 13.5% | 41285 | 2984 |

Figure 4 shows a histogram of the percent of the population with a bachelor's degree or higher, comparing No Trafficking with the Brothels Model. The disparity between the two groups is quite evident from the histogram, with the typical percentages being quite lower in the presence of brothels and higher in the cities with no human trafficking. The mean percentage for cities with no trafficking was 35% higher than that of the cities with brothels present, and the standard deviation was nearly 2.5 times larger, indicating a much narrower range in the percentage for cities with brothels, predominantly between 15% and 40%.

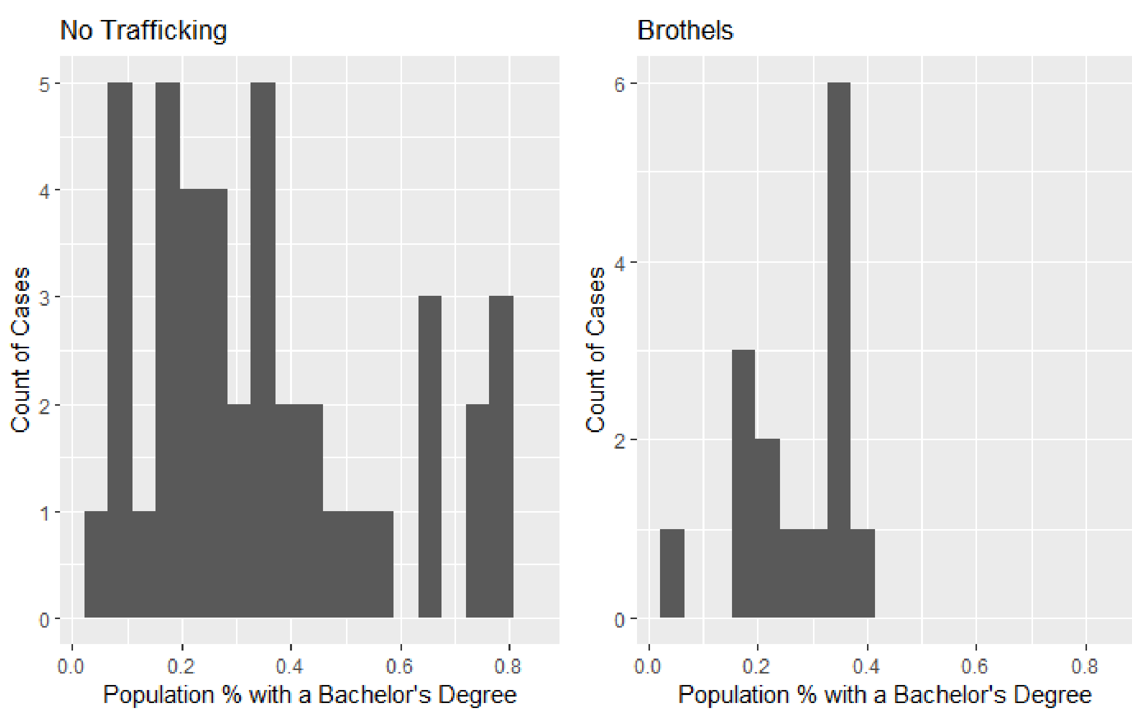


**Figure 4 - Histogram of the Percent of the Population with a Bachelor's Degree Against Brothels**

Figure 5 shows a histogram of the percent of the population with a high school diploma or higher education level, comparing No Trafficking with the Brothels Model. Results are similar to the comparison of the bachelor's degree, where the cities with brothels have distinctly lower percentages of the population with high school diplomas and the variation in range for the cities with brothels is generally lower except the few outlier cases with extremely low rates. The average percentage for cities with no trafficking was 12% higher than that of the cities with brothels present, so not quite as high a disparity as the bachelor's degrees, but still apparent.

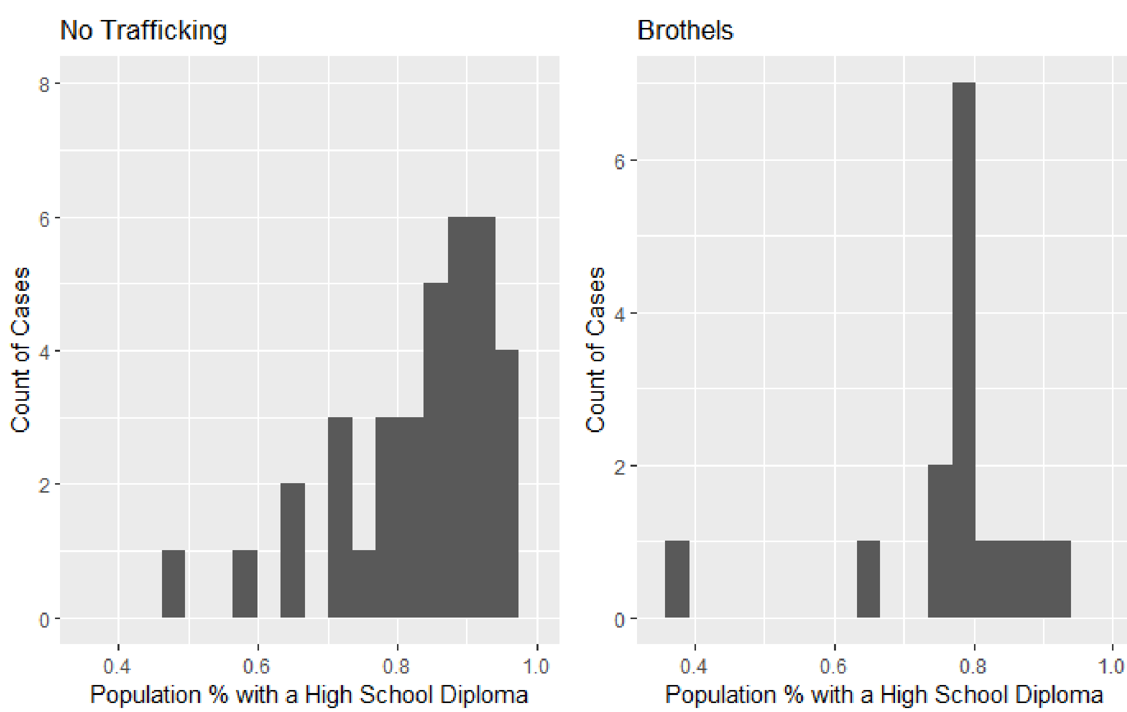


**Figure 5 - Histogram of the Percent of the Population with a High School Diploma Against Brothels**

The median home value of cities with No Trafficking against the Brothels Model is shown in the Figure 6 histogram, continuing the trend from the previous histograms where the presence of brothels correlates with lower median home values. In cities with no trafficking, the median home values averaged more than 50% higher, indicating a clear differentiation that is also apparent in the histogram. The lower home values are fairly evenly matched between the two groups, but the histogram representing the absence of human trafficking has a tail that extends much further into the higher values.

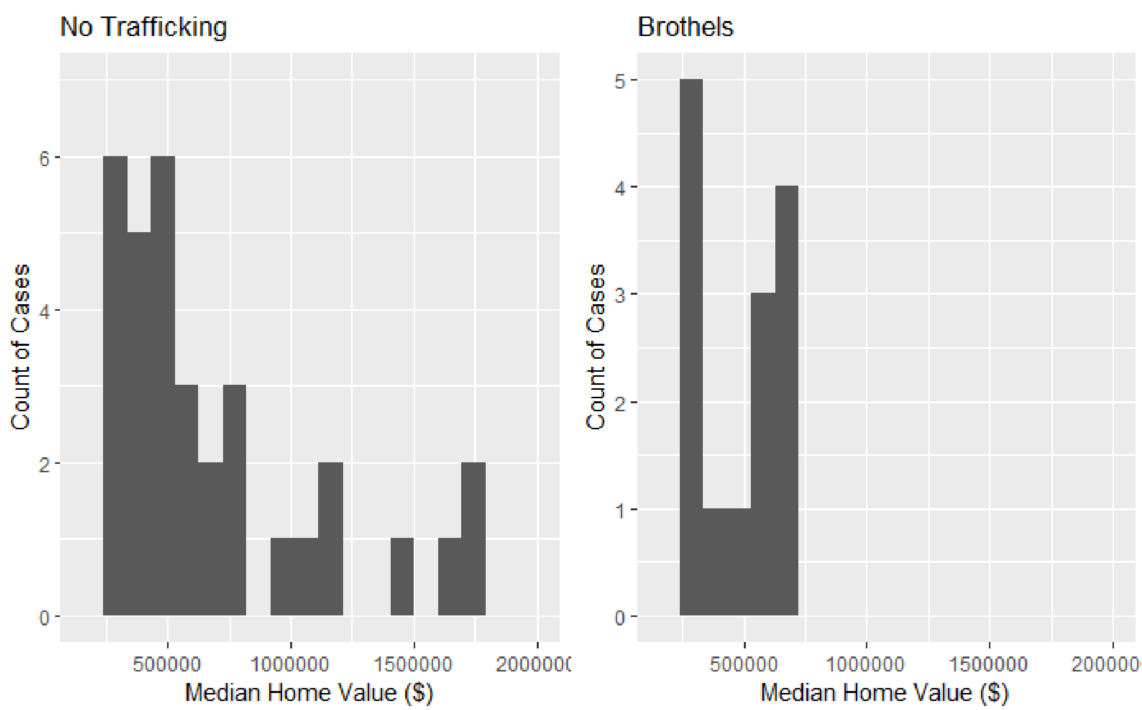


**Figure 6 - Histogram of the Median Home Value Against Brothels**

The median household income in cities with No Trafficking compared to the cities with the Brothel Model is shown in Figure 7, where a large disparity can be discerned from the histograms. In cities with no trafficking, the median home values averaged almost 60% higher than in cities with brothels, and again the tail of the histogram extends much further into the higher household income levels. The range of household incomes in the cities with brothels was also significantly smaller, where the cities without trafficking had both a number of lower household incomes than the Brothels Model as well as a substantial reach into significantly higher income levels.

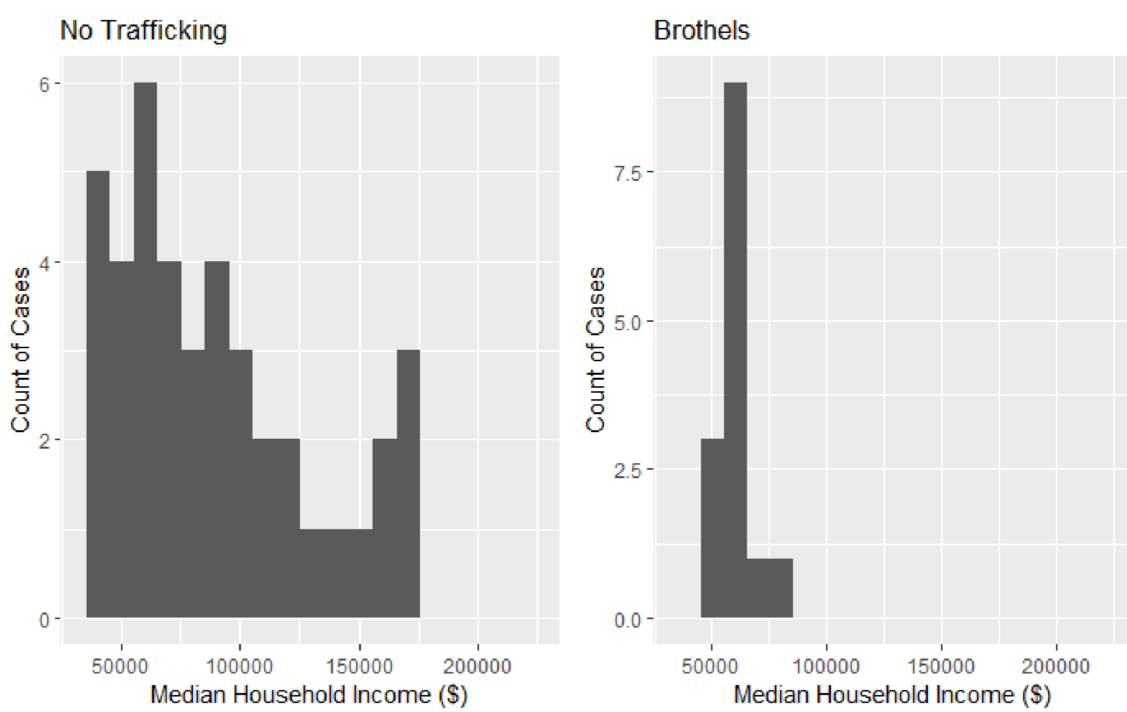


**Figure 7 – Histogram of the Median Household Income Against Brothels**

The final key predictor specific to the Brothels Model was the percentage of the population that was White, not including the Hispanic population, and Figure 8 shows this comparison against cities without human trafficking. The mean percentage of the White population in cities with no trafficking was also significantly higher than in cities with brothels, averaging more than 85% higher. There was more variation in both groups compared to the previous key brothel predictors, but it is evident from the Brothels histogram that lower percentages are correlated with the presence of brothels.

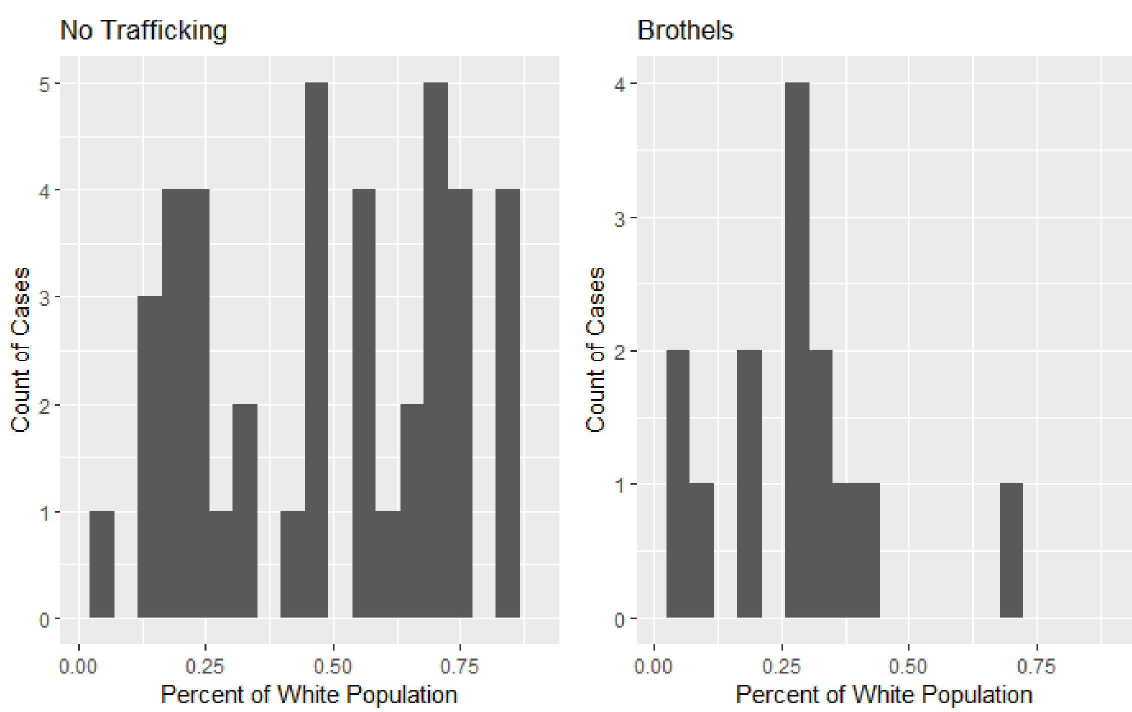


**Figure 8 - Histogram of the Percent of the Population that is White Against Brothels**

Considering the key predictors for the Pimp Business Model, Figure 9 shows a comparison of the percentage of the population over the age of 65 between the No Trafficking cities and the cities with pimps. The mean percentage for the cities with no trafficking was more than 30% higher than the cities with pimps, and it can be seen from the histograms that, although the two groups overlap significantly in the lower percentages, the No Trafficking histogram has a longer tail extending into the higher percentages. The delineation between the two groups is not as distinct as with some of the previous predictors, but the disparity is still evident when analyzing the data.

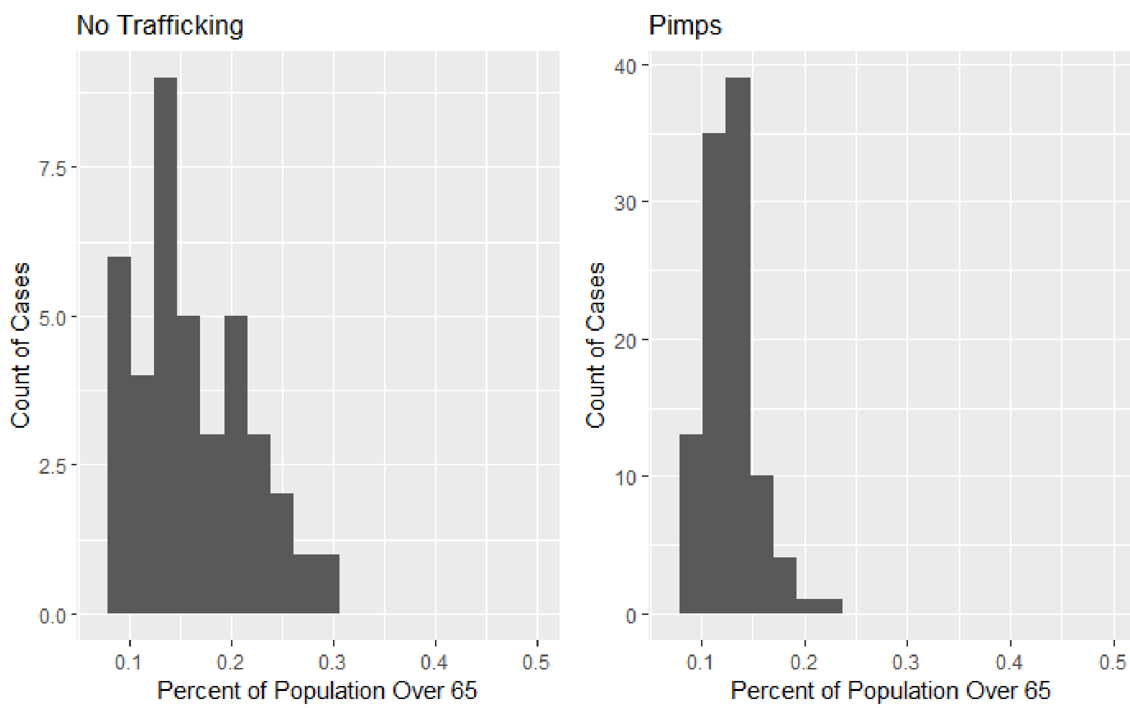


**Figure 9 - Histogram of the Percent of the Population Over the Age of 65 Against Pimps**

The other key predictor for the Pimp Business Model, the transient occupancy tax rate, is depicted in Figure 10. This predictor was also key for the Streetwalking Business Model, but the results were compared in separate histograms. Figure 10 shows that there is a significant skew between the two groups, where the rate at which hotels and other transient occupancy businesses are taxed is higher for those cities with pimps compared to cities with no human trafficking. The tax rate for the cities with pimps was almost 20% higher than the No Trafficking cities, and, although the No Trafficking cities had more variation in the tax rate, the majority of cases fell in the center of the range while the cases with pimps were predominantly on the righthand side of the histogram.

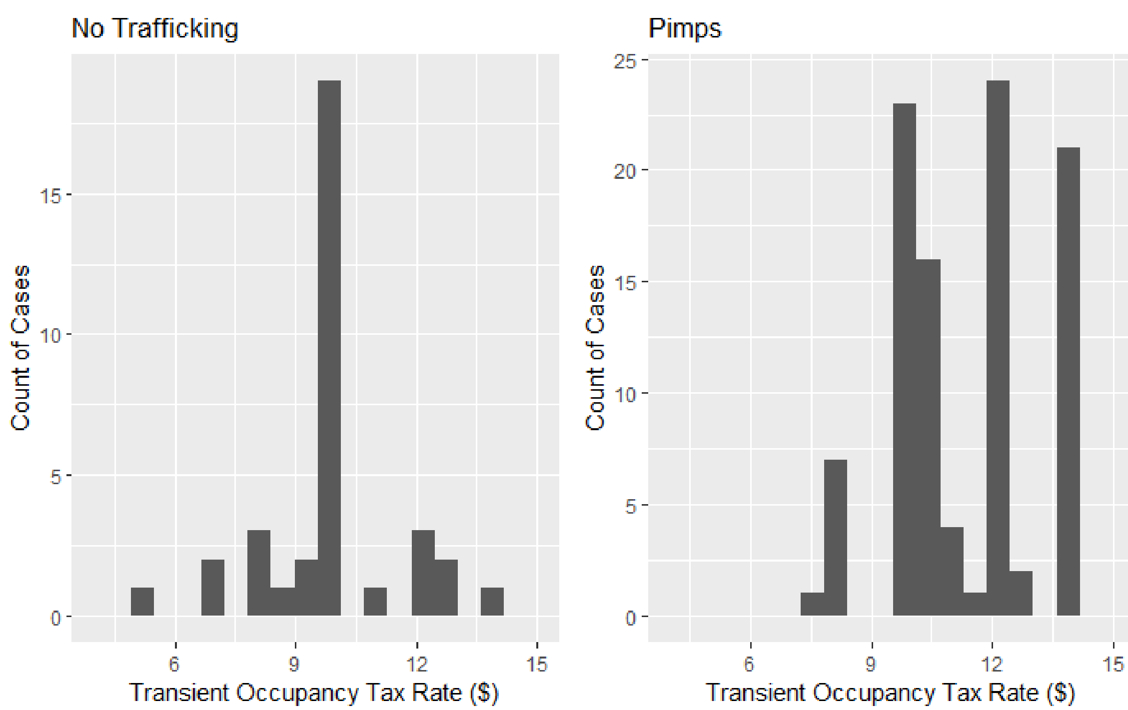


**Figure 10 - Histogram of the Transient Occupancy Tax Rate Against Pimps**

Regarding the Streetwalking Business Model, the percentage of the population that is Black is shown in Figure 11 against the No Trafficking cases. The average percentage in cities with streetwalkers was more than 2.5 times higher than in the No Trafficking cities, however it is important to note the scale of the histograms. The population percentages were overall fairly low, averaging 3% for No Trafficking and 8% for Streetwalking, and thus a magnitude-related comparison may not be the best characterization for understanding the correlation. The percent of the population that is Black is predominantly under 15% in the streetwalking cases and is predominantly below 5% in the No Trafficking cities, which may be a more useful comparison.

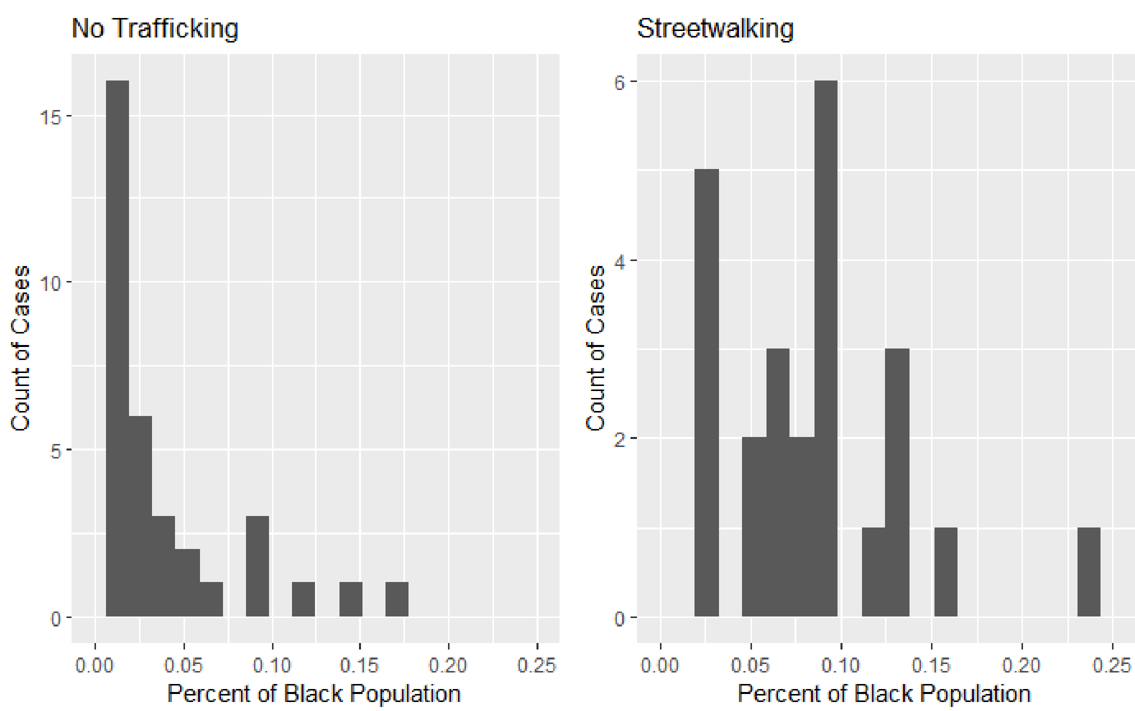


**Figure 11 - Histogram of the Percent of the Population that is Black Against Streetwalking**

As with the Pimp Business Model, the transient occupancy tax rate was also a key predictor for the Streetwalking Business Model, and the results shown in Figure 12 are similar to those shown in Figure 10. The No Trafficking histogram is naturally the same as for Figure 10, and the Streetwalking histogram is very closely related but skews higher than the Pimps histogram. The disparity in the tax rate for the cities with streetwalking was greater than with the Pimp Model, more than 25% higher than the No Trafficking cities. The range of values for the Streetwalking group is also smaller than in the Pimps group, indicating a stronger correlation between higher transient occupancy tax rates and the presence of streetwalkers in a city.

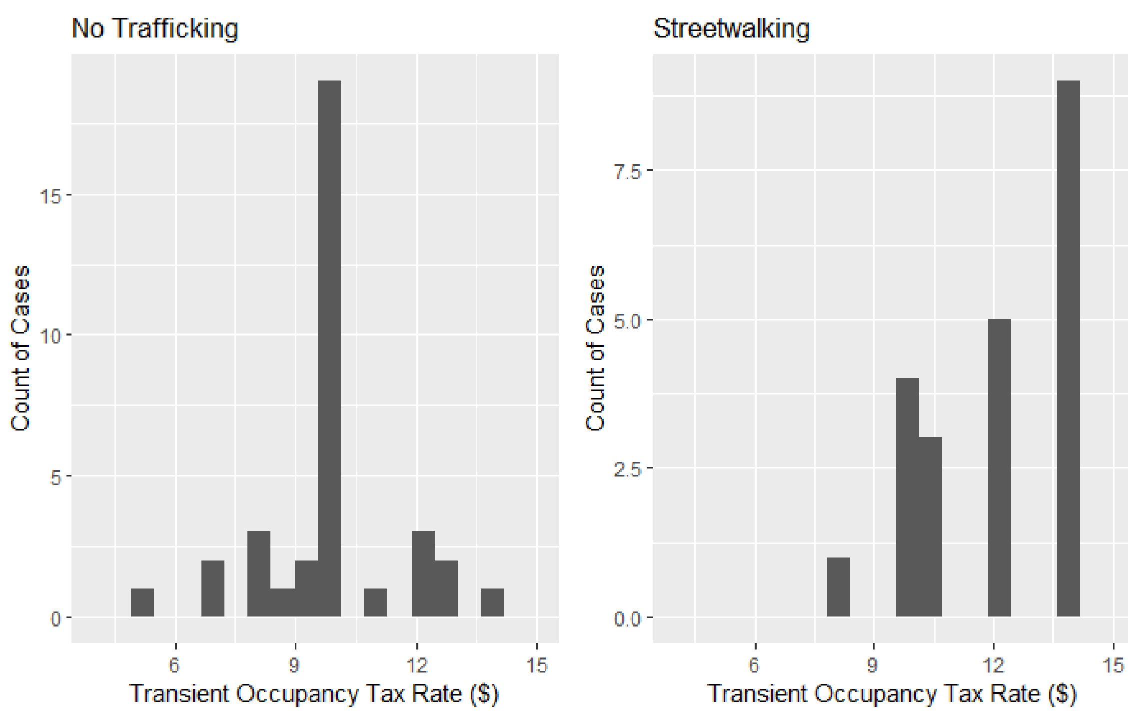


**Figure 12 - Histogram of the Transient Occupancy Tax Rate Against Streetwalking**

The percentage of homes occupied by the owners rather than by tenants was a key predictor of the No Trafficking cities, and it can be seen from Figure 13 that the percent of homeowners occupying their own homes is noticeably greater for the cases where there was no human trafficking, although the disparity in the size of the dataset can be seen via the y-axis and could influence the model outcomes. All human trafficking business models were included in the rightmost histogram since this was a predictor of the No Trafficking model. The average percentage in the cities without human trafficking was more than 30% higher than in the cities with human trafficking.

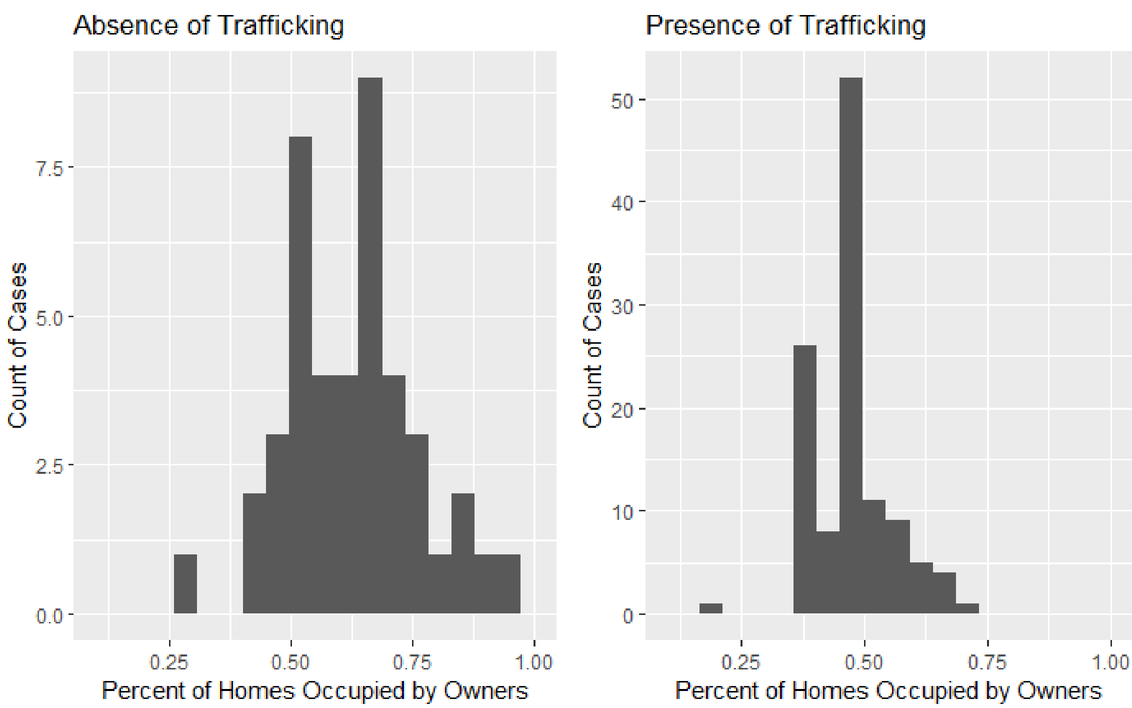


**Figure 13 - Histogram of the Percent of Homes Occupied by Their Owners Against No Trafficking**

The overall city population was a key predictor for four out of the five models (including the No Trafficking Model), the only exception being the Brothel Model which scored population quite low in the rankings. Because of how many of the models had population as a key predictor, Figure 14 compares the presence and the absence of human trafficking. The disparity in the populations can clearly be seen from the histogram, and the average population of cities with sex trafficking was nearly 28 times larger than the average of those cities with no human trafficking.

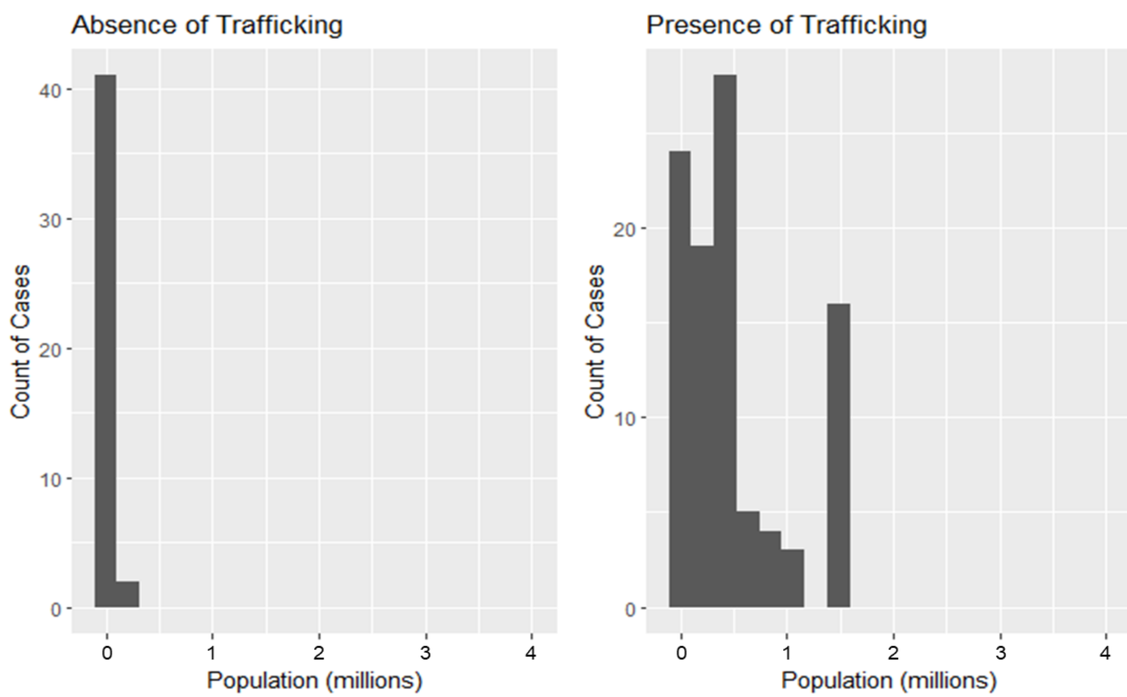


**Figure 14 - Histogram of the Population Against No Trafficking**

Finally, the population density results mirror those of the total population, where the cities with higher human trafficking problems generally have higher population densities. Population density was a key predictor for the No Trafficking, Brothels, and Pimp Business Models, and the histograms in Figure 15 were grouped in the same way as in Figure 14 to compare the overall presence and absence of human trafficking. The average population density of cities with human trafficking cases was more than 70% higher than in the cities without human trafficking, and the tail of the histogram is significantly higher in volume than in the no trafficking cities as well.

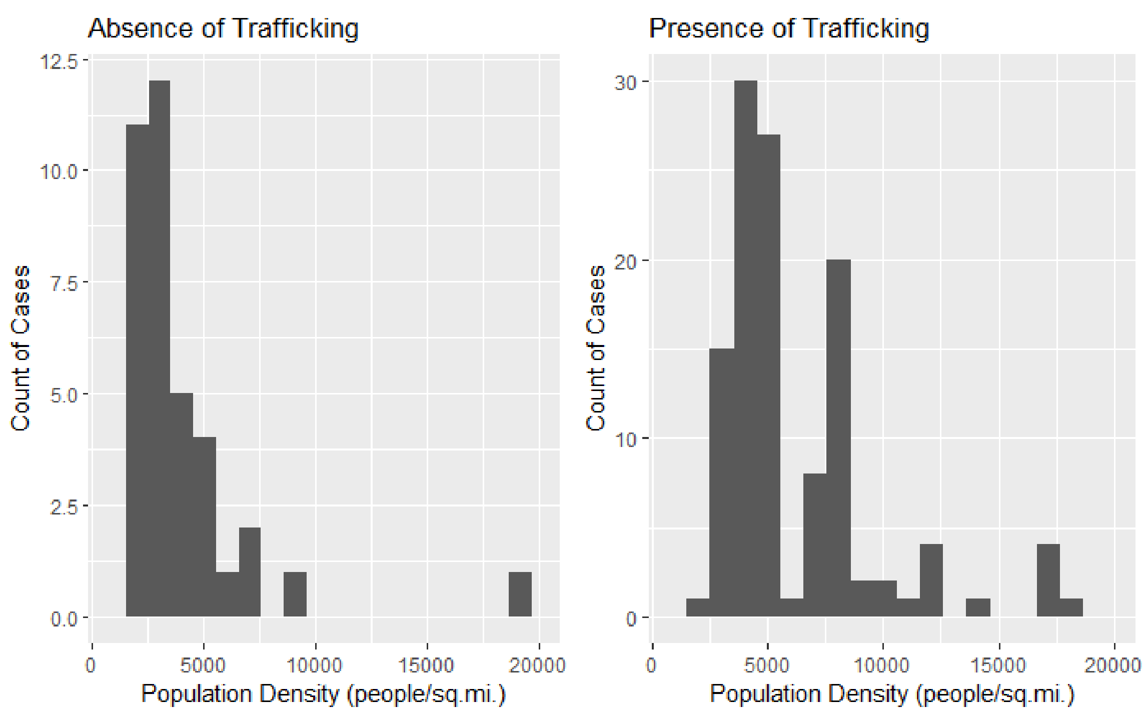


**Figure 15 - Histogram of the Population Density Against No Trafficking**

# CHAPTER FIVE: DISCUSSION

From the results presented, it can be seen that some general trends have emerged indicating there can be some correlations discerned between city characteristics and the type of sex trafficking business models present in that city. The results analyzed the performance of the predictive machine learning models – logistic regression, random forest, and multinomial logistic regression – as well as the key predictors ranked by the random forest models to determine the overall indicators of the types of business models present. The findings for each of the different business models is below.

Education statistics were key predictors for the Brothels Business Model, where lower education rates generally correlated with an increased probability of a city having brothels. The education disparity was more predominantly seen in the percent of the population that had a bachelor's degree or higher, which was 35% greater in the cities without brothels compared to the percent of the population possessing a high school diploma only being 12% higher in the cities without brothels. Based on these results, it is fairly clear that education levels are inversely correlated with the presence of brothels, although further analysis would be needed to determine how dependent the two factors are on each other.

The median home value and the median household income results also appear to be inversely correlated the presence of brothels, where lower home values and household incomes correlated with the presence of brothels in the city. At a surface level, this is not an especially surprising fact since the presence of brothels could be hypothesized to more likely occur and maintain continuity in a lower income area, but the difference in both home values and income levels is perhaps more noticeable than one would suppose. As previously mentioned, the cities without brothels had approximately 60% higher median incomes than the cities with brothels and the median home value was more than 50% higher in the cities without brothels. This clear disparity provides a strong data point in trying to predict which cities will have a greater presence of brothels for sex trafficking.

The percent of the city population that was White was the final key predictor for the Brothels Business Model and was also inversely correlated with the presence of brothels, with the average White population being 85% higher in the cities without brothels. These results then indicate that cities with lower education levels, lower household incomes and home values, and lower White population percentages are correlated with the presence of brothels in a city. This is perhaps not an unsurprising outcome considering crime in general is more prolific when education and income levels are lower, and both education and income levels are systemically lower in areas with higher non-White demographics which make up minority populations when considering the state and nation as a whole. Based on this information, it is possible that working to increase education and income levels in such areas would help to reduce the problem of sex trafficking in such areas, which is also not inconsistent with the existing literature on the topic.

Considering the Pimp Business Model, it was not especially surprising that cities with pimps generally had a lower age demographic than those without pimps. Sex trafficking in general is going to be easier to conduct in cities with younger populations since it is easier to traffic the young and vulnerable, and older populations tend to leave larger cities as they age. Furthermore, pimps are generally in the younger age ranges [4], late teens to early forties, and thus are more likely to be prolific in cities with lower percentages of those over 65 years of age.

For both the Pimp and Streetwalking Business Models, the transient occupancy tax rate was higher for cities with the presence of pimps and streetwalkers. The mean transient occupancy tax rate for the presence of pimps and streetwalkers was nearly 20% and over 25% higher, respectively, when compared to the cities with no trafficking. Initially, these results were not as would have been expected since presumably lower tax rates would facilitate the easier use of hotels by traffickers. However, after reviewing the raw data, generally the larger cities had higher transient occupancy taxes, and thus the population factor was a more significant consideration to the traffickers than higher prices for hotel rooms, likely because they were still able to maximize profits compared to utilizing smaller cities with less demand.

The other key predictor for the Streetwalking Business Model was the percent of the population in a city that was Black. As previously mentioned, the disparity between the Streetwalking cities and those without human trafficking was noticeable, but the percentages for both groupings remained quite low. These results may be somewhat related to the presence of brothels being inversely correlated with the percent of the population that is White, but further analysis would be necessary to determine if there was a correlation between the two business models. From these outcomes, all that can be reliably concluded is that marginally higher percentages of Black populations in a city correlated with the cities that had streetwalkers.

Considering the results from the percentage of homes occupied by the owners rather than by tenants, the outcomes indicated that a higher proportion of homeowners living in their own homes correlated with less likelihood of sex trafficking in the city. The percent of homes occupied by their owners was more than 30% higher in cities without human trafficking. This also correlates with a higher income area, where the general populace is able to afford their own house rather than have to rent a house or an apartment from a landlord. This also makes sense from the perspective of brothels being easier to hide in rental housing, such as apartments, since it makes it more convenient and less troublesome to move the business to another location if discovered.

The final key predictors were the total population and population density data, which showed a clear relationship between higher populations and the presence of sex trafficking. Larger city populations would mean less likelihood of traffickers being caught for human trafficking as well as providing more customers and victims, and higher population density means there is less distance between the traffickers, the victims, and their prospective clients. The mean population density was more than 70% higher for cities with the presence of trafficking compared to the control cities, and the mean city population of the control cities was dwarfed by that of the cities with sex trafficking, the latter being nearly 28 times larger than the former. It is important to note that the mean populations are quite skewed due to the outliers like Los Angeles, San Diego, and San Jose, with populations well above the median. However, even taking those outliers into account, the population and population densities are still quite disparate between the two groupings.

## CHAPTER SIX: CONCLUSION

From the results section, it can be seen that some general trends have emerged indicating there can be some correlation between city characteristics and the type of sex trafficking business models found. These trends include: lower education levels, incomes, home values, and White population percentages all correlated with the presence of brothels; lower percentages of those over the age of 65 correlated with the Pimps Business Model; higher transient occupancy tax rates correlated with both the Pimps and the Streetwalking Business Models; marginally higher Black population percentages correlated with the Streetwalking Business Model; higher percentages of homes being occupied by their owners correlated with the No Trafficking data; and higher populations and population densities were inversely correlated with the No Trafficking data.

It is interesting to note that none of the key predictors were statistically correlated with the Gangs Business Model based on the T-Test results, although the median home value and the White population percent were common to both the Brothels and Gangs Business Models before running the T-Test. This may require further research as it would likely be informative to analyze if different gangs utilized differing methods of human trafficking to meet their ends, and perhaps the reason none of these predictors were statistically significant may be due to combining all of the gangs into one category rather than analyzing the differences between the gangs. Therefore, a more in-depth analysis on gang involvement in human trafficking rings may also be of use and would provide significant opportunity for further research.

As mentioned previously, it is likely that the predictors of human trafficking common to one state or region may not be common to another due to regional and geographic discrepancies. These key predictors and their correlations to the sex trafficking business models may be specific to the State of California. Further analysis would be necessary to determine how different the regional predictors might be and whether the nation as a whole could be analyzed using similar methods or if it would be necessary to utilize a state-by-state methodology.

Finally, this analysis was prevented from considering a temporal aspect due to the small size of the dataset, but that would also be a further opportunity for research. Specifically, analyzing how these correlations between the type of sex trafficking business model and the city characteristics may change over time could provide insight into outcomes for constantly evolving cities. However, additional data would be necessary to complete this analysis.

## REFERENCES


[1] K. Feehs and A. Currier, "2019 Federal Human Trafficking Report", Human Trafficking Institute, 2020.

[2] United Nations Office on Drugs and Crime, "Global Report on Trafficking in Persons 2020", United Nations publication, 2020.

[3] L. Shelley, *Human trafficking*. Cambridge: Cambridge University Press, 2012.

[4] V. Bouché, “Human Trafficking Data”, Humantraffickingdata.org, 2021. [Online]. Available: http://www.humantraffickingdata.org/.

[5] J. Cole and G. Sprang, "Sex trafficking of minors in metropolitan, micropolitan, and rural communities", *Child Abuse & Neglect*, vol. 40, pp. 113-123, 2015.

[6] W. Macias Konstantopoulos et al., "An International Comparative Public Health Analysis of Sex Trafficking of Women and Girls in Eight Cities: Achieving a More Effective Health Sector Response", *Journal of Urban Health*, vol. 90, no. 6, pp. 1194-1204, 2013.

[7] B. Bloom, "Just Ask: Healthcare Provider Screening for Intimate Partner Violence and Human Trafficking Victims in Southern California", Master of Science in Public Health, University of California Los Angeles, 2014.

[8] L. Fedina, C. Williamson and T. Perdue, "Risk Factors for Domestic Child Sex Trafficking in the United States", *Journal of Interpersonal Violence*, vol. 34, no. 13, pp. 2653-2673, 2016.

[9] E. Cockbain and R. Wortley, "Everyday atrocities: does internal (domestic) sex trafficking of British children satisfy the expectations of opportunity theories of crime?", *Crime Science*, vol. 4, no. 1, 2015.

[10] M. Twis, L. Gillespie and D. Greenwood, "An Analysis of Romantic Partnership Dynamics in Domestic Minor Sex Trafficking Case Files", *Journal of Interpersonal Violence*, 2020.

[11] M. Twis, "Predicting Different Types of Victim-Trafficker Relationships: A

Multinomial Logistic Regression Analysis", *Journal of Human Trafficking*, vol. 6, no. 4, pp. 450-466, 2019.

[12] C. Panlilio, S. Miyamoto, S. Font and H. Schreier, "Assessing risk of commercial sexual exploitation among children involved in the child welfare system", *Child Abuse & Neglect*, vol. 87, pp. 88-99, 2019.

[13] J. Reid, M. Baglivio, A. Piquero, M. Greenwald and N. Epps, "Human Trafficking of Minors and Childhood Adversity in Florida", *American Journal of Public Health*, vol. 107, no. 2, pp. 306-311, 2017.

[14] J. Walters and P. Davis, "Human Trafficking, Sex Tourism, and Child Exploitation on the Southern Border," *Journal of Applied Research on Children: Informing Policy for Children at Risk*, vol. 2, no 6, 2011.

[15] D. Hughes, "Combating Sex Trafficking: A Perpetrator-Focused Approach", *University of St. Thomas Law Journal*, vol 6, 2008.

[16] A. Carpinteri, B. Bang, K. Klimley, R. Black and V. Van Hasselt, "Commercial Sexual Exploitation of Children: an Assessment of Offender Characteristics", *Journal of Police and Criminal Psychology*, vol. 33, no. 2, pp. 150-157, 2017.

[17] M. Mahfoud, W. Bernasco, S. Bhulai and R. van der Mei, "Forecasting Spatio-Temporal Variation in Residential Burglary with the Integrated Laplace Approximation Framework: Effects of Crime Generators, Street Networks, and Prior Crimes", *Journal of Quantitative Criminology*, 2020.

[18] L. Kiss, D. Fotheringhame, J. Mak, A. McAlpine and C. Zimmerman, "The use of Bayesian networks for realist evaluation of complex interventions: evidence for prevention of human trafficking", *Journal of Computational Social Science*, vol. 4, no. 1, pp. 25-48, 2020.

[19] B. Pal and T. Mondal, "Women's trafficking in twenty-first century India: A quest for regional variations", *Asian Journal of Women's Studies*, vol. 27, no. 1, pp. 130-151, 2021.

[20] E. Helderop, J. Huff, F. Morstatter, A. Grubesic and D. Wallace, "Hidden in Plain Sight: A Machine Learning Approach for Detecting Prostitution Activity in Phoenix, Arizona", *Applied Spatial Analysis and Policy*, vol. 12, no. 4, pp. 941-963, 2018.

[21] Z. Xia, K. Stewart and J. Fan, "Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a

major U.S. metropolitan area", *Computers, Environment and Urban Systems*, vol. 87, p. 101599, 2021.

[22] J. Barnum, W. Campbell, S. Trocchio, J. Caplan and L. Kennedy, "Examining the Environmental Characteristics of Drug Dealing Locations", *Crime & Delinquency*, vol. 63, no. 13, pp. 1731-1756, 2016.

[23] L. Fry, "Continuities in the Prediction of Human Trafficking: A Research Note", *International Journal of Comparative and Applied Criminal Justice*, vol. 34, no. 1, pp. 201-212, 2010.

[24] S. Chainey, L. Tompson and S. Uhlig, "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime", *Security Journal*, vol. 21, no. 1-2, pp. 4-28, 2008.

[25] D. Mletzko, L. Summers and A. Arnio, "Spatial patterns of urban sex trafficking", *Journal of Criminal Justice*, vol. 58, pp. 87-96, 2018.

[26] E. Cockbain and K. Bowers, "Human trafficking for sex, labour and domestic servitude: how do key trafficking types compare and what are their predictors?", *Crime, Law and Social Change*, vol. 72, no. 1, pp. 9-34, 2019.

[27] D. Oblinger, "Sex Trafficking in Sexually Oriented Businesses", in *Combating Human Trafficking: A Multidisciplinary Approach*, M. Palmiotto, Ed. CRC Press, 2015.

[28] J. Stearns, “Street Gangs and Human Trafficking", in *Combating Human Trafficking: A Multidisciplinary Approach*, M. Palmiotto, Ed. CRC Press, 2015.

[29] "Welcome - California Open Data", Data.ca.gov, 2021. [Online]. Available: https://data.ca.gov/.

[30] US Census Bureau, "Census.gov", Census.gov, 2021. [Online]. Available: https://www.census.gov/en.html.

## BIOGRAPHY

Lydia Shirey graduated from high school (home school) in Fairfax, Virginia, in 2012. She received her Bachelor of Science in Biomedical Engineering from George Mason University in 2016. She is currently employed as a systems engineer doing government contracting work and, with the completion of this thesis, will graduate with her Master of Science in Systems Engineering with a Data Analytics concentration from George Mason University in 2021.