City of Stars, City of Crimes:

Data Driven Insights into

LA’s Crime Report

Prepared by

Haiying Lin, Nathaniel Zhu, Youyi Fu, Maohua Wu

Team 11 BizStat Innovators

MGT 256 Final Project Paper

Professor—Adem Orsdemir

University of California Riverside

**1. Introduction**

Crime data analysis holds an important role in understanding and analyzing criminal activities in urban areas. Throughout this research paper, we will comprehensively study the crime data specifically in the Los Angeles area, with surrounding smaller urban cities. Los Angeles is known for its diverse demographics and complex urban environments, emerging as the ideal focus for our research topic - predicting crime. The motivations behind this topic are: firstly, to leverage statistical methods in deciphering patterns and trends using the LAPD crime data from 2020 to 2023, a four-year period, and secondly, to explore the potential of predicting victim sex by implementing predictive modeling.

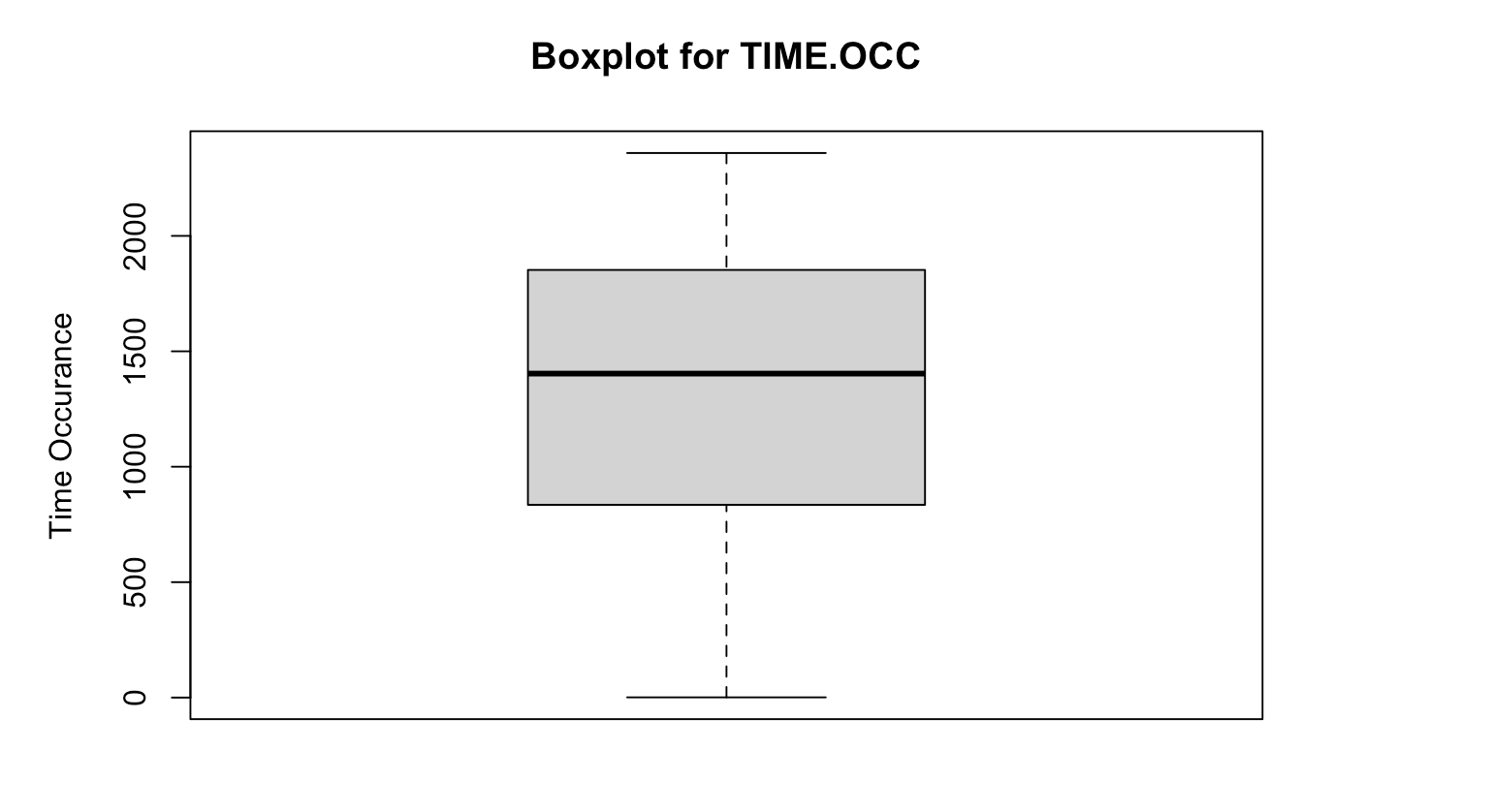
Los Angeles, with its large metropolitan expanse, presents a unique opportunity for crime analysis. The city’s crime rates not only impact local communities but also serve as indicators of urban criminal trends at a national level. Additionally, as residents of the Los Angeles metropolitan area, this study sparked our interest delving into the secrets and patterns of crime activities. Therefore, This study is significant for policymakers, law enforcement agencies, and urban city planners seeking to enhance the living conditions of the neighborhood and resource allocation. By predicting victim sex, it can also help law enforcement and community organizations develop useful targeted prevention strategies and policies.

Throughout this paper, we will organize our discussion into four main sections: “Descriptive Analysis”, “Logistic Regression”, “KNN”, and “Conclusions”. The Descriptive Analysis section will illustrate the current crime landscape of LA through various graphical representations, starting our research with a better understanding of the background of LA crime data. These representations include distribution of victim sex, crime codes, time trends during the year, and seasonality of crime. Next, the two Predictive Modeling sections will detail our predictive modeling process with logistic regression and KNN, including model selection, validation, and interpretation of results. These methods are chosen for their accuracy and efficiency in predicting categorical and numerical outcomes, especially useful in the context of crime prediction. Finally, the Conclusion section will synthesize our findings, offering recommendations we gathered based on our research and challenges and limitations we encountered during the data gathering, research, and analysis.

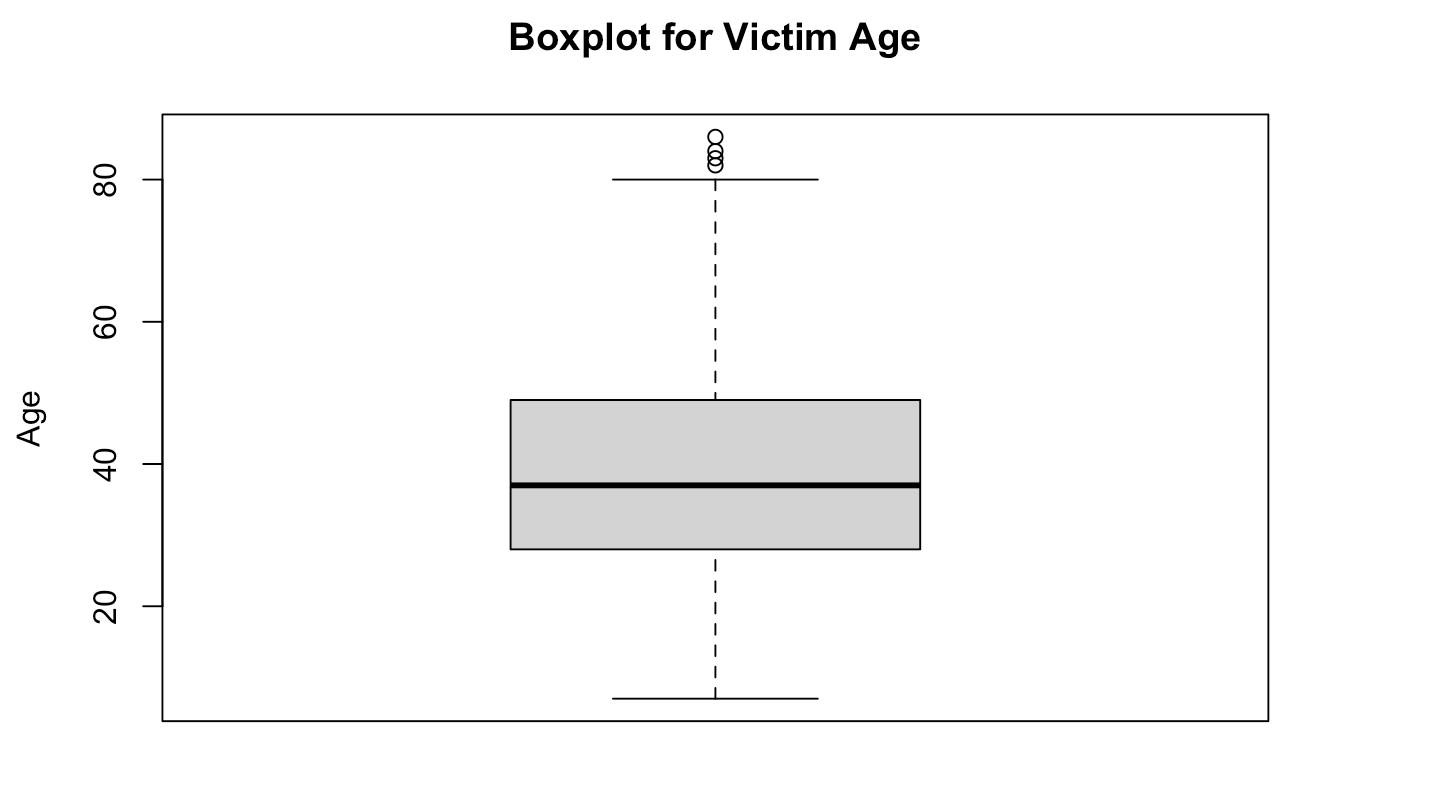
By the end of this paper, we aim to have a detailed statistical analysis of LA’s crime data, showcasing the effectiveness of descriptive and predictive modeling in understanding and forecasting crime patterns. Ultimately, this study serves as a practical application of classroom materials and knowledge, preparing us for future professional circumstances. We look forward to presenting our findings in the subsequent sections.

**2. Descriptive Analysis:**

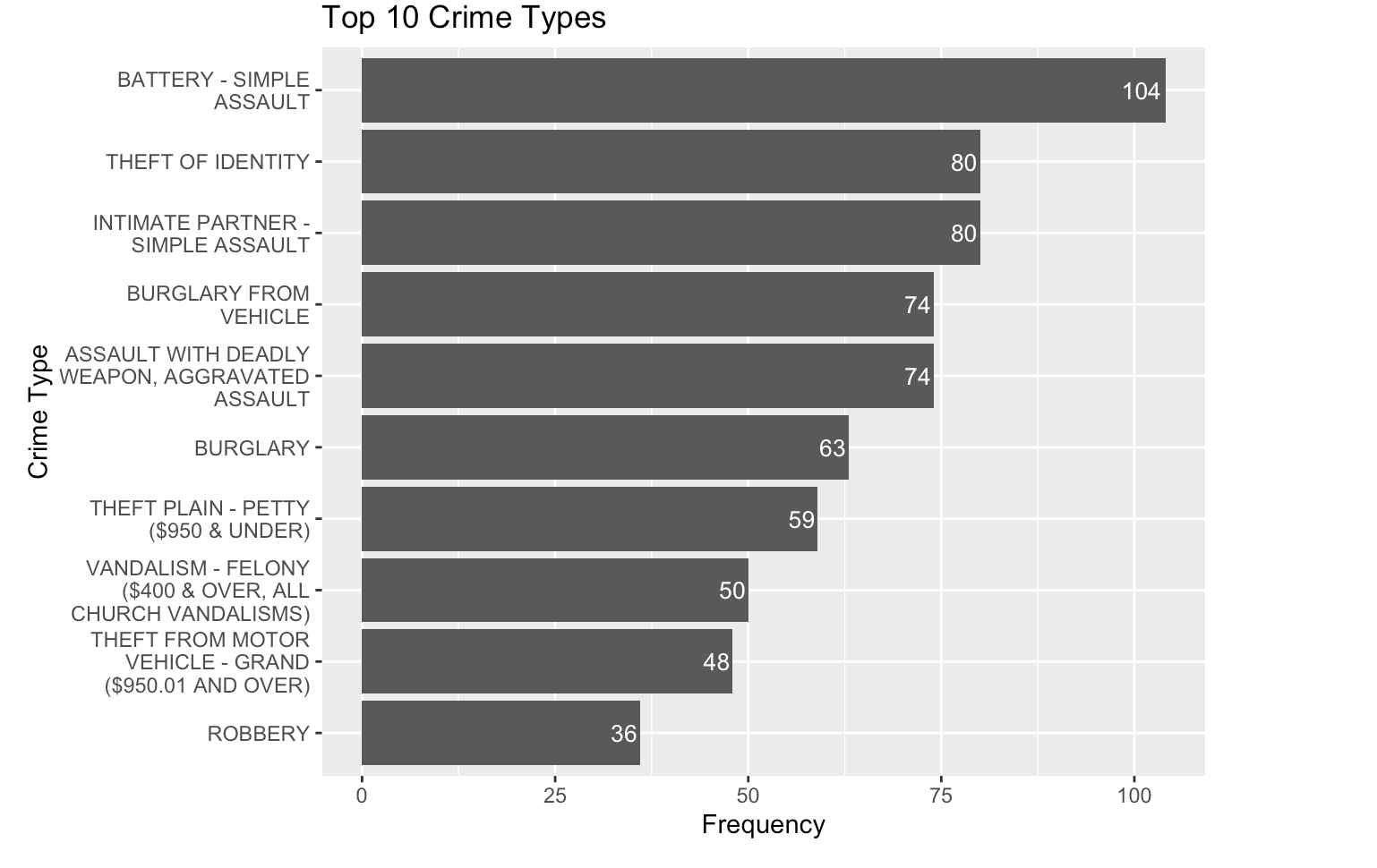
We gathered our data from Data.gov: *Crime Data from 2020 to Present*. This data includes a total of 870,000 raw data without proper cleaning. Most of the columns inside are not properly labeled and used solely by LAPD, which means those columns are dropped for our future studies. First, since we are predicting the victim sex, we are going to drop blank information or “X” labeled data in Vict.Sex, this is because some of the crimes are non victim related, like vandalism, or victim refuse or include unisex. We also filter out the number “0” in Vict.Age, the reason is because it is either non victim related crime or the victim refuses to include age. After we have cleaned the data and dropped unuseful columns, we randomly select 1000 observations with a set seed of 123. Below graphs are the result from our sample data:

**2.1 Figure 1: Boxplot for TIME.OCC**

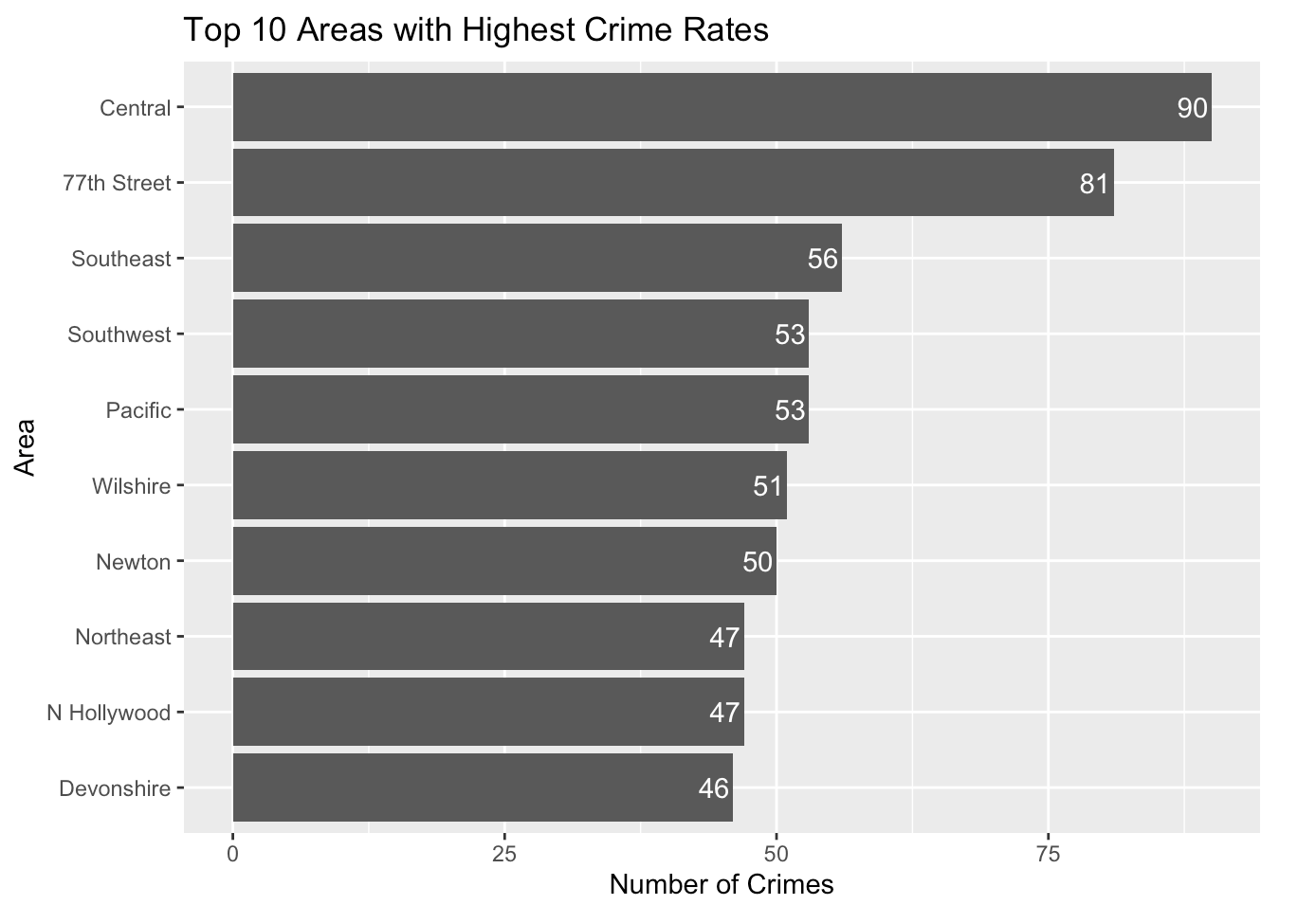
From observing the time occurrence, we created the boxplot for TIME.OCC, one of our independent variables. As we can see from the graph above, it shows there are no outliers, which means there are cases going on throughout the day. We can conclude that crimes happen more at daylight time, and the boxplot tells us that 50% of data is observed from 9am to 7pm. Thus, we believe that in Los Angeles, the crime rate is higher in the daylight period rather than during the night.

**2.2 Figure 2: Boxplot for victim age**

In this boxplot, the interquartile range upper limit is 80 and the lower limit is 3. The 75th quartile is around 50, which means 75% of victims are younger than 50. And there are some obvious outliers on the top of this boxplot that shows some of these victims are older than 80.

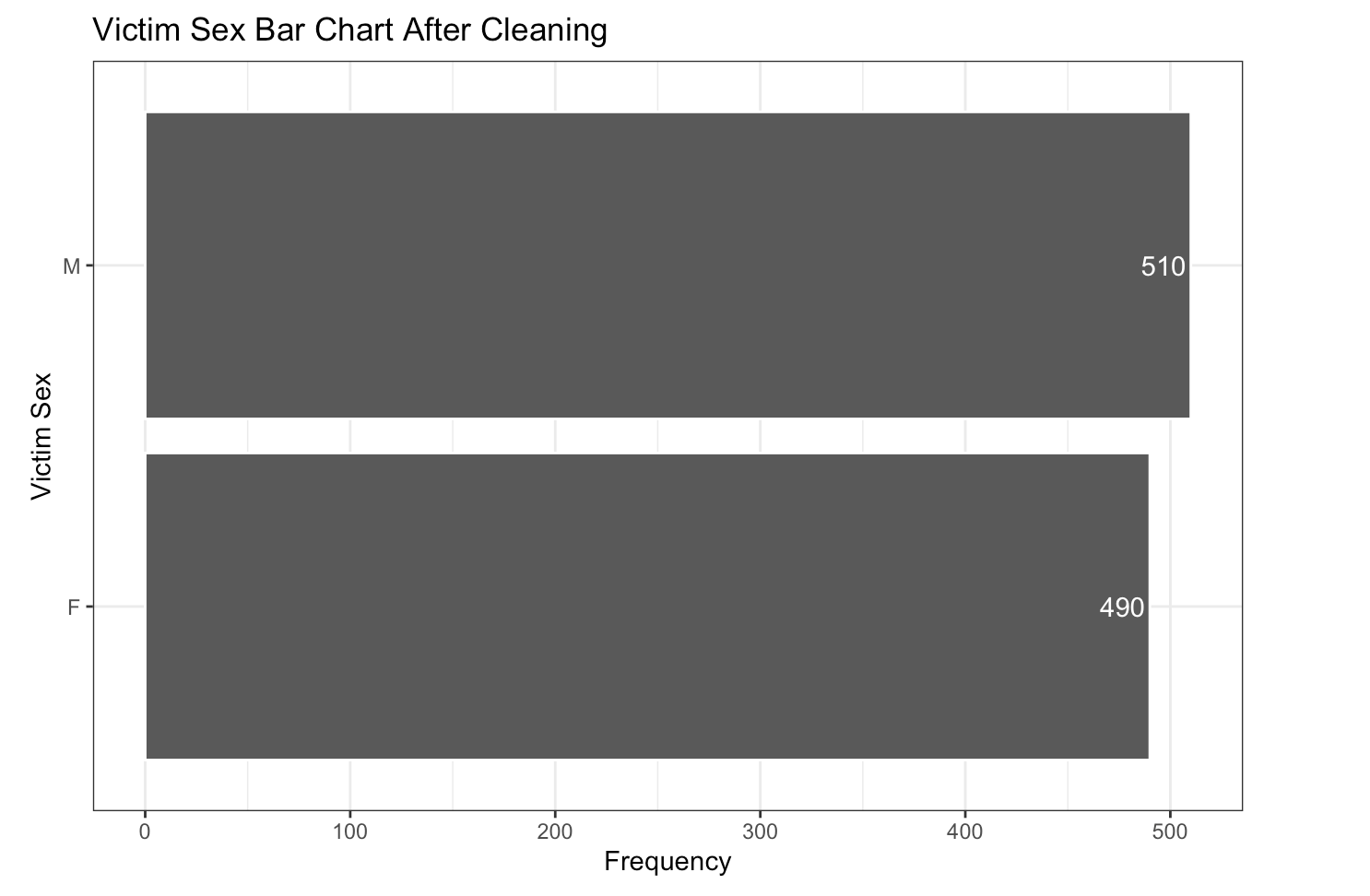
**2.3 Figure 3: Top 10 crime types**

In this barplot, we analyze the top 10 crime types from total crimes, the y axis is the crime type and the x axis represent their frequency. The most frequent crime is battery - simple assault. Based on this descriptive analysis, city planners and governors can create specific community outreach or support groups to the victims to reduce this type of crime from happening. Following the barplot, there are a lot of cases related to cars, such as burglary from vehicle and theft from vehicle, which could advise people not to leave the stuff in the cars when they leave their cars, especially in the Los Angeles area.

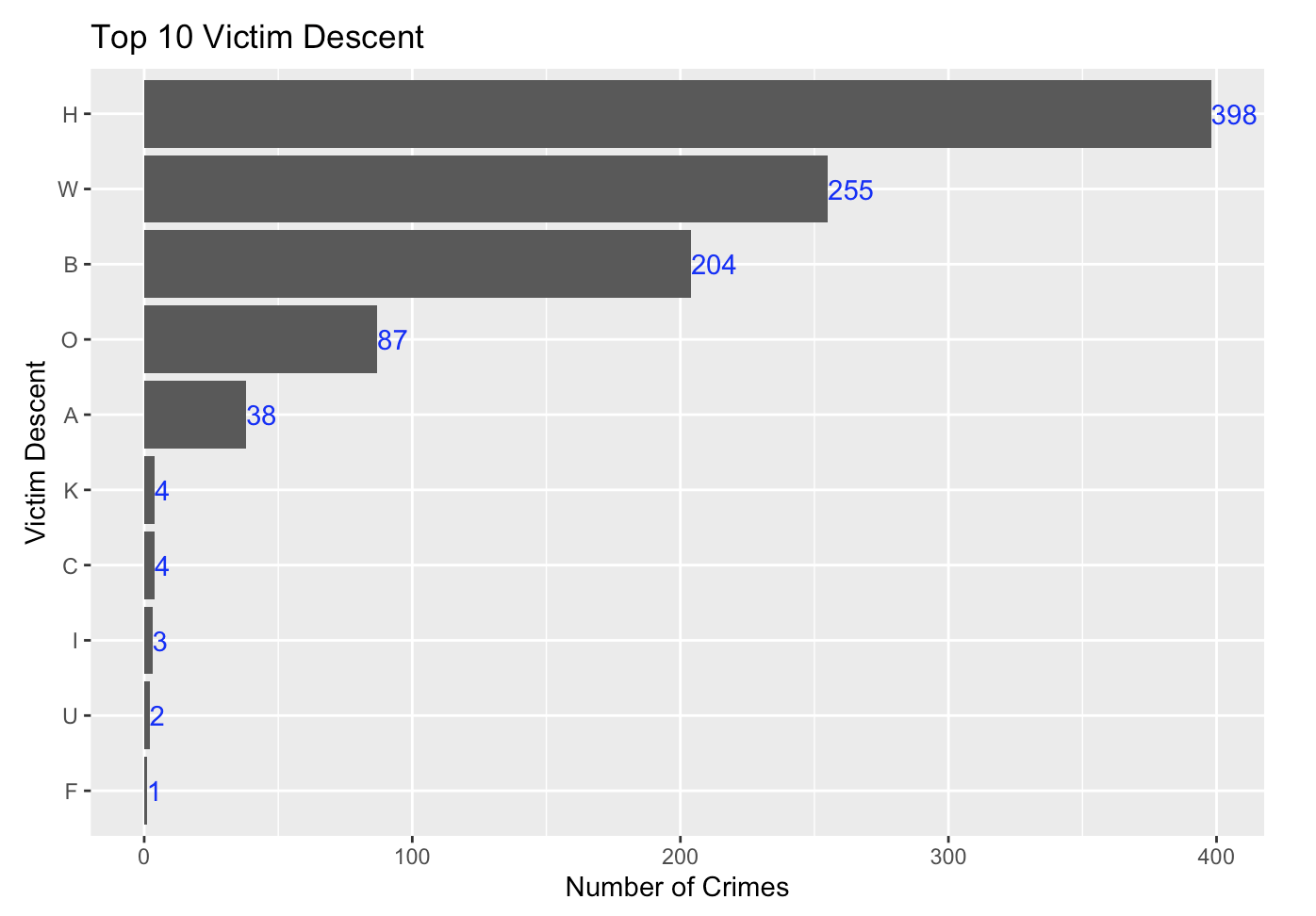
**2.4 Figure 4: Top 10 Areas with Highest Crime Rates**

In this graphical analysis, the horizontal bar chart shows the comparative crime data for these areas, providing a clear visual representation of the relative safety and criminal activity within each location from our sample data.

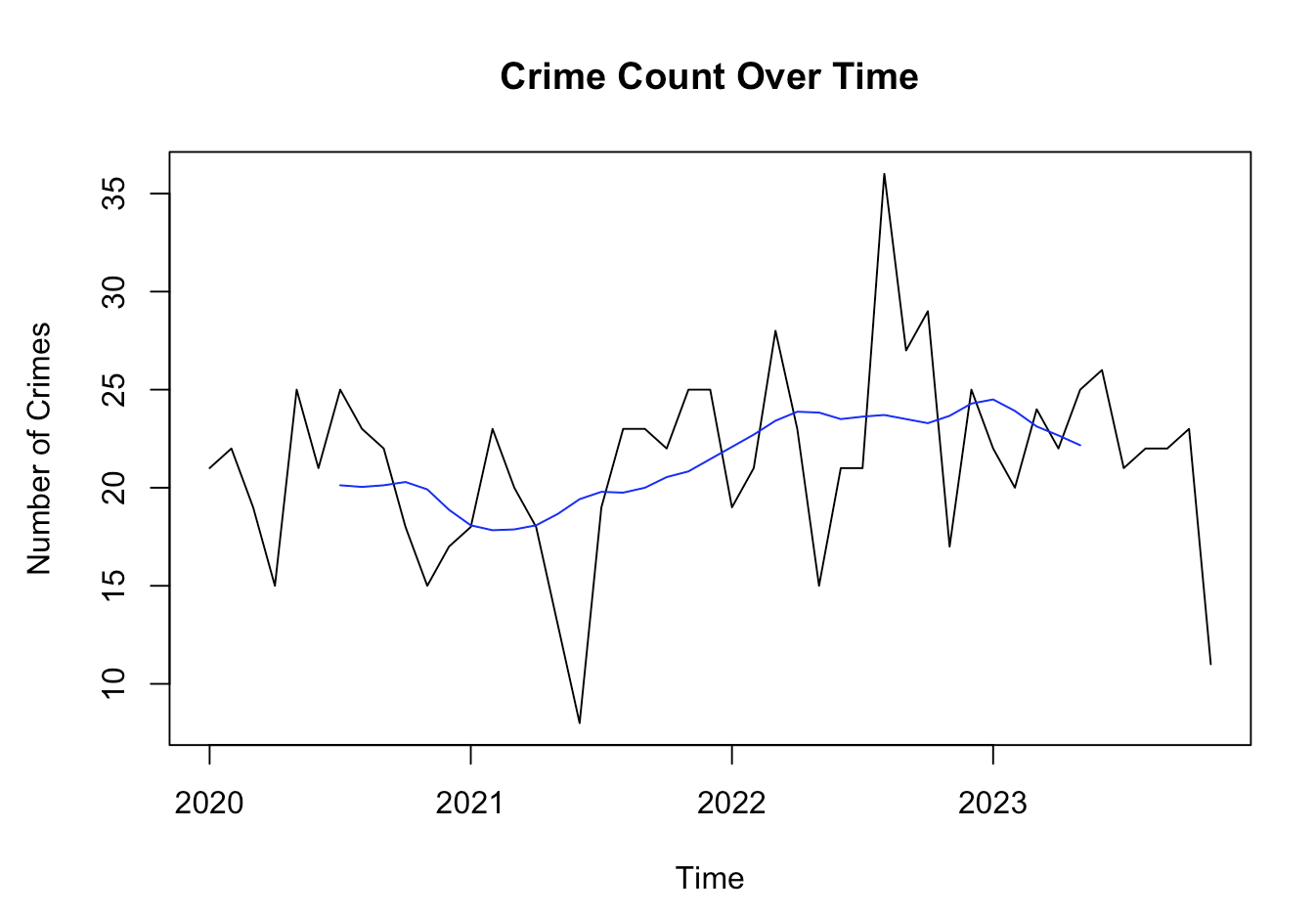
As we can see from the graph, LA Central is the most affected area, with a total of 90 reported crimes out of 1000 samples we randomly selected. Secondly, 77th Street has 81 reported crimes in total. These data bars and areas are notably higher than the rest, representing an immediate attention and concern of the potential underlying factors contributing to these high rates. This concentration may be influenced by many factors, such as population density, economic disparity, or even varying levels of police presence and resources allocated to this area.

**2.5 Figure 5: Victim sex bar chart after cleaning**

For the victim sex of 1000 objects we randomly choose from the original data, we created a bar chart to compare the victim sex. The y axis is victim sex and the x axis is the frequency of the crimes. We can see that the ratio of male to female victims is almost 50% from each other, which suggests that women victims are not more or less than men.

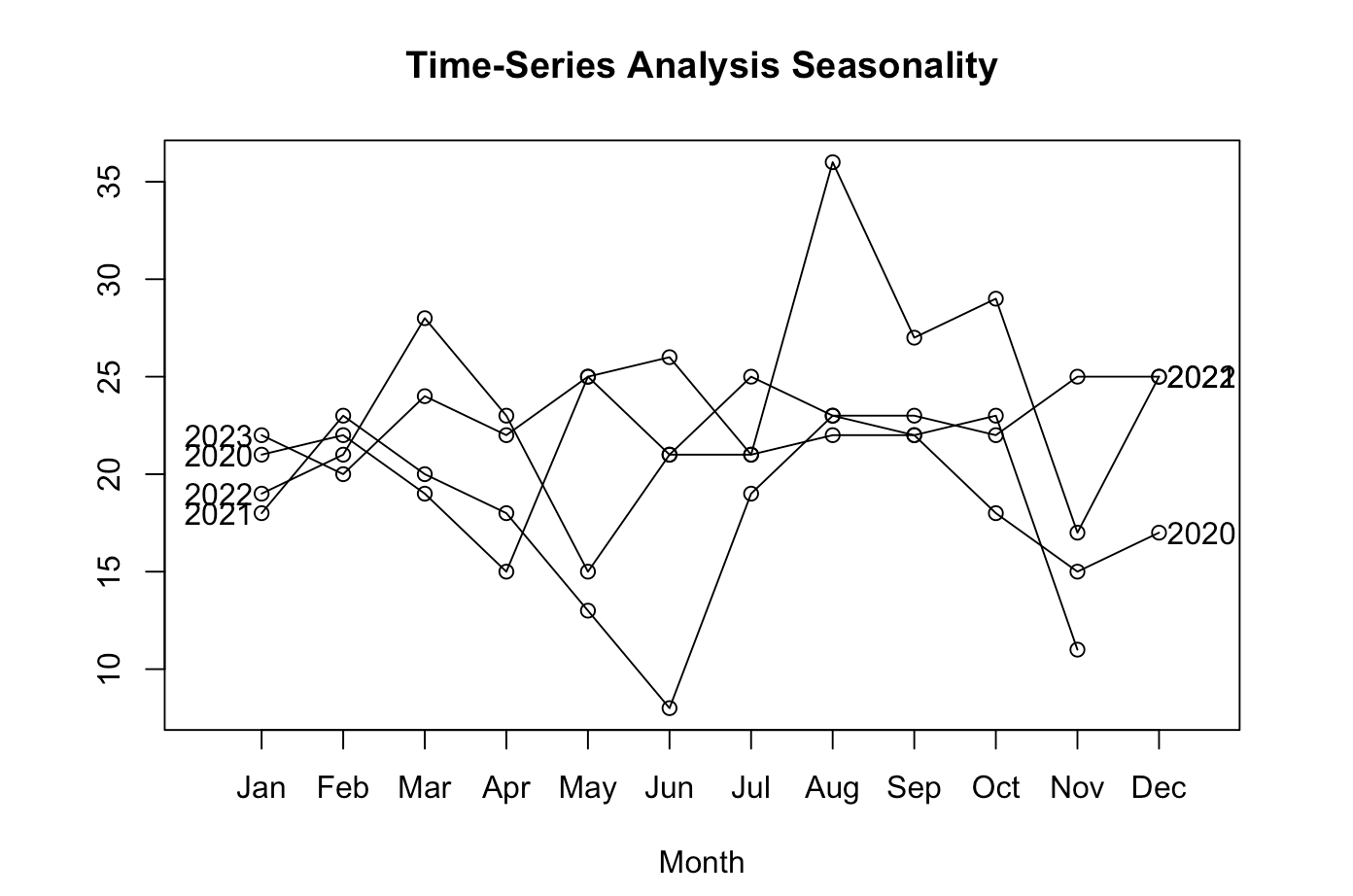
**2.6 Figure 6: Top 10 Victim Descent**

In this bar chart, we can see the categories of victims of crimes by their descent, highlighting the top three descents reported in the Los Angeles crime dataset: Hispanic, White, and Black. As we can analyze from this graph, there are a total of 398 reported Hispanic Victims in this sample dataset. This number significantly exceeds other victim descent. It could represent the various socio-economic factors, demographic representation in the population, or reporting tendencies. As the majority in this sample dataset, it is paramount to understand the context behind this graph and form a comprehensive decision. The White and Black victims are 255 and 204, respectively. While it is considerably lower than the Hispanic descent, it still represents a substantial portion of the crime victims in this dataset. The reasons behind this distribution include many factors such as residential patterns, social circumstances, or even income distribution, which requires future investigation for the city planners to draw meaningful conclusions.

**2.7 Figure 7: Crime Count Over Time**

The time series chart exhibits the fluctuation in the number of crimes over a span of four years. Throughout these four years, we can see that in 2020, there is not too much fluctuation, indicating that the crimes happening during the year are with lower variance. However, in 2021, we can see that there is a significant drop. The causation was probably due to the COVID-19, since the city was on lockdown and people were staying at home. However, in the following year 2022, there is a significant increase in the crime rate, mainly due to the reopening of the city and people returning to the streets. By 2023, the graph indicates a downward trend, with the number of crimes decreasing noticeably. This could be a result of effective law enforcement policies, community intervention programs, or other socio-economic factors that may have influenced the crime rate.

The blue line represents a moving average, showing slight upward trends in the middle of 2022, suggesting an increase in average crime rates during this period.

**2.9 Figure 9: Time-Series Analysis Seasonality**

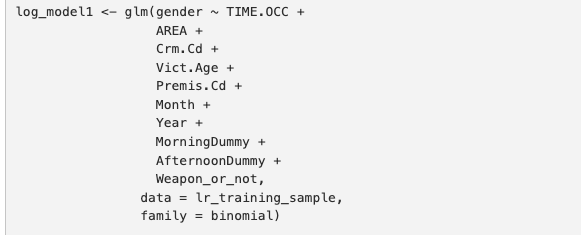
This chart shows the number of crimes each month over four years: 2020 to 2023. The lines for each year go up and down, showing how crime changes with the seasons. As we can see from the graph, there is a noticeable peak during the summer months, especially in July and August. This pattern repeats every year, which suggests that crime trends increase in the warmer months. By December, the number of crimes drops, which is consistent across all years, showing that there are fewer crimes in the winter months. With the analysis from this chart, LA city planners and governors are able to suggest new programs, introduce new strategies, or create community reach to reduce the crimes happening during the summer period.

**3 Logistic Regression**

Logistic regression is a statistical method used for binary classification, which means predicting the probability of an observation belonging to one of two classes. Despite its name, logistic regression is a classification algorithm, not a regression algorithm. It is widely used in various fields, including statistics, machine learning, and epidemiology.

**3.1 Applying Algorithm**

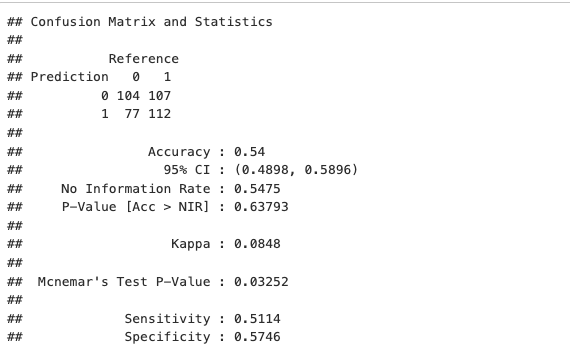
We want to predict victim sex, so we use gender as a dependent variable. Then, we choose the independent variables that most likely affect gender to form the model.



After that, we do sampling, to divide our data into two parts: training data with 60% from the sample and validation data with the rest of 40% from the sample data. Now we can use our model and two datasets to make predictions. In the next section, I will be addressing the performance of our logistic regression model.

**3.2 Assess the performance of Logistic Regression model**

To assess the performance of our model, we need to form a confusion matrix. We choose the cutoff probability of 0.49, because the sample probability of the victim being female is 49%. And here is the result we get.



The result indicates that the model has a 54% chance to predict sex of the victim successfully. The probability is not too high, and there could be several reasons that cause the result, such as the sex distribution is rather random. Thus, there is no effective way to predict it. or the independent variables we use are not highly correlated to gender.

**3.3 Discussion of the Logistic Regression model**

There are several ways that may improve our model.

1. Collect more data: The data we used might be biased. If we can have more comprehensive data, we can use them to build a stronger model.

2. Try more combinations of independent variables: Different combinations of independent variables can result in different models. The model may yield higher accuracy when more combinations are tested. However, based on our research on the data, fewer independent variables reduce the accuracy of successfully predicting the victim sex. Hence, we would use the whole numerical data points to generate our logistic regression model.

**4 K-Nearest Neighbor**

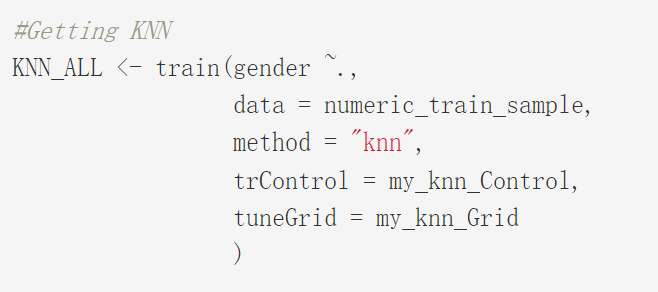
K-Nearest Neighbor Classifier is a supervised machine learning algorithm useful for classification problems. It works by finding the distances between a query and all the examples in the data, selecting the specified examples that are closest to the query, and then votes for the most frequent label. To put it in simple words, the model structure is decided by the data. It's pretty useful because in reality, most of the data does not follow the typical theoretical norms made. Hence, we decided to use the K-Nearest Neighbor Algorithm in our study.

**4.1 Applying Algorithm**

The first step is to convert the dependent categorical data (victim sex) into a dummy variable (Female=0 and Male=1). Then, we convert other categorical independent variables into dummy variables. We drop all the non-relevant categorical variables and create a new sheet of numerical data (dataset: numeric\_only\_data) using set seed of 666. It assists to create a model in a direct way.

The train and test data were split using the same technique that we mentioned in the previous logistic regression section, and then implemented a 10 fold cross validation technique.

Using the train function to implement the KNN method with numeric train sample data . Through the highest accuracy index with 10 different KNN models, finding that K=6 is the optimal model and storing the result in the object (knn\_all). We select all the variables in the numerical sample data because we have tried if just to select some of them, the accuracy would be lower than 55%.

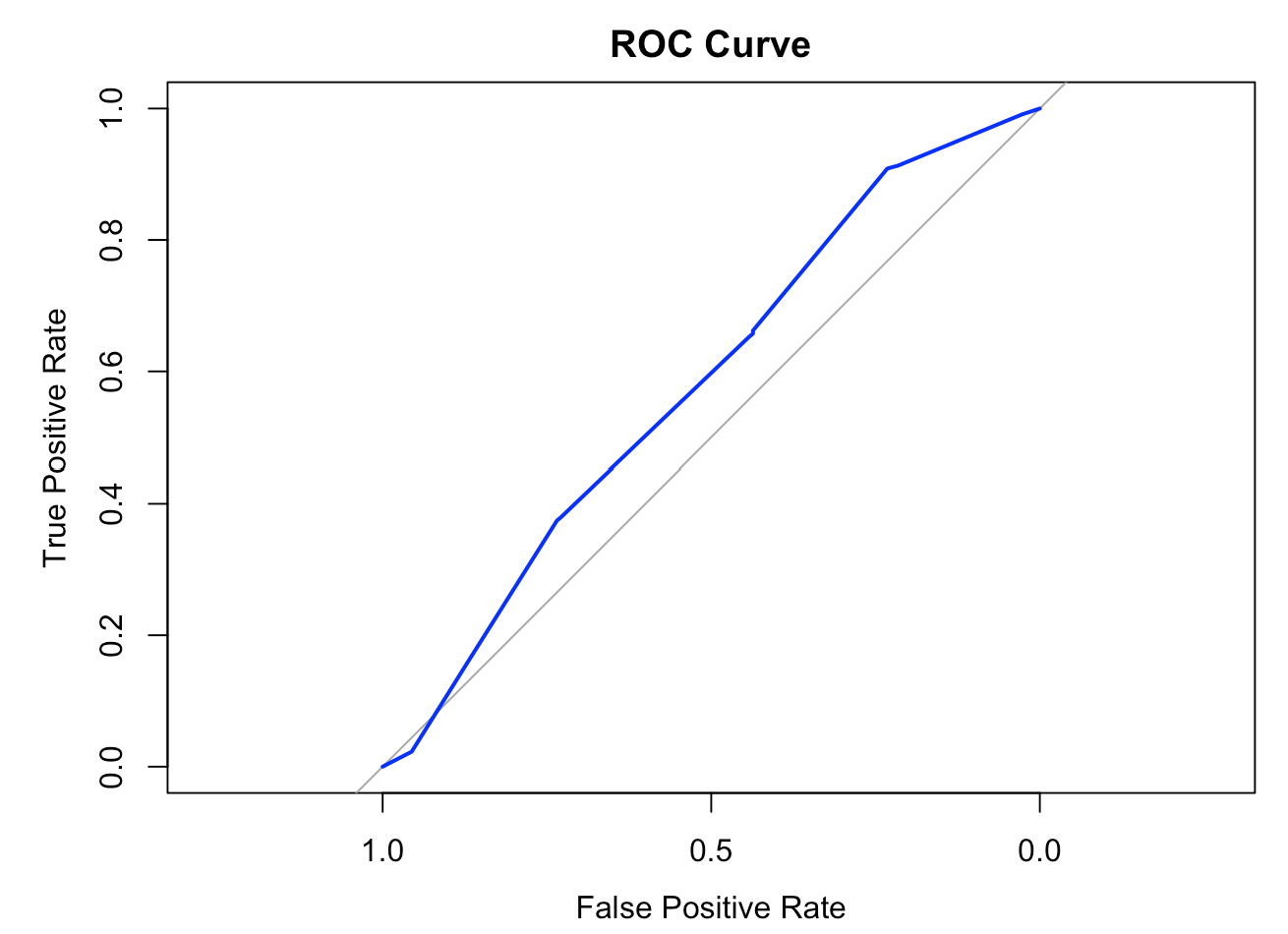


**4.2 Assess the performance of KNN model**

To evaluate the predicted performance of KNN model, using the prediction function with numeric test data. And then use a confusion matrix function to result in accuracy, sensitivity, and specificity. The results are shown below:

Accuracy : 0.55 Sensitivity : 0.5023 Specificity : 0.6077

The accuracy and specificity index is not too high. This means that the results predicted by the KNN model are only 55% accurate, and the probability of making Type I and Type II errors are 50% and 40%, respectively. The reason why the accuracy of the KNN model prediction is not high may be because the sample data is too small, and the expected outcome of females and male are always equal to 0 or 1.



We use the plot function to create the cumulative lift chart. Plotting the lift curve for the KNN model, and drawing a gray diagonal line to indicate the baseline model.The lit curve from the KNN model lies above the gray line, indicating that the KNN model performs considerably better than the baseline model. The result of the graph in the area under the curve is consistent with the confusion matrix result.

**5 Summary**

This research work offers a way to foresee and predict crimes and frauds within Los Angeles city. It focuses on having a crime prediction tool that can be helpful to law enforcement. This paper is aimed at increasing the prediction accuracy from victim gender. Along the way, many patterns of criminal activities in various areas which will be helpful for criminal investigation were known. This pattern has much greater importance than we realize. The logistic regression model and KNN system helps law enforcement agencies for improved crime analysis. By traversing through the crime dataset, we have to find out different reasons that lead to crime.

Since this paper is bearing in mind only some limited factors, such as the small sample size, limited independent variables, and low accuracy predicted result. This dataset contains too many categorical variables and few specific numerical data that can help contribute to the accuracy of our model. For getting more accurate results in prediction, we have to put more data observations and use more accurate independent variables in the training models. In the future, governments and LAPD could use well developed classification algorithms that can help to detect criminal activities more efficiently, and we can also expect to see lower crime rates using such prediction techniques.

This project enables us to utilize classroom materials alongside with the real-world implementation. The research and learning is invaluable, we expect to learn more from related studies in the future.

**Reference**

City of Los Angeles. “Crime Data from 2020 to Present.” Data Catalog, City of Los Angeles. data.lacity.org, 9 Dec. 2023, <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>