ANALYSIS OF SEATTLE CRIME DATA

Brandon Mendoza & Melanie Ma

2023-02-19

Introduction

The dataset used in this study is the "Seattle Crime Data: 2008-Present" dataset provided by the Seattle government. The dataset was selected because it is relevant to the study of the general well-being and safety of Seattle's citizens. The dataset comprises a sample of Seattle's entire crime history data, with a focus on recent data (since 2008). The statistical analysis and visualizations were conducted using R and Tableau. It is important to note that the dataset contains missing values across several columns, which were accounted for and removed accordingly during the analysis.

Research Questions

Upon closer inspection of the dataset, the aim is to perform more statistical analysis on two research questions:

1) What is the overall trend of total crime in Seattle, and how did the pandemic affect crime?

This question is of interest as it seeks to investigate the validity of claims that Seattle has become more dangerous. Raw data will be analyzed to determine if this is the case.

2) In order to maximie safety and societal well-being, which precincts are in need of staff (namely, police reports) the most, and for which times of day?

In 2022, the Seattle Police Department staff count reached a new low for the first time in 30 years. This question aims to better understand how SPD can improve community safety by analyzing staffing needs of each precinct and determining specific time intervals where more staff is required. By exploring which precincts are most in need of staff during specific time intervals, resources can be better allocated, and community safety prioritized.

Statistical Analysis

1) What is the overall trend of total crime in Seattle across time, and how did the pandemic affect crime?

```
# Loading the data
spd <- read.csv("spd.csv")

spd$Offense.Start.DateTime <- as.POSIXct(spd$Offense.Start.DateTime,
format="%m/%d/%Y %I:%M:%S %p")

spd$Date.Of.Crime <- as.Date(spd$Offense.Start.DateTime)

recentSpd <- spd %>%
  filter(year(spd$Offense.Start.DateTime) >= 2008 & year(spd$Offense.Start.DateTime) <= 2023)
```

To investigate this inquiry, a time series analysis will be conducted on the dataset to discern the overall pattern of criminal incidents in Seattle.

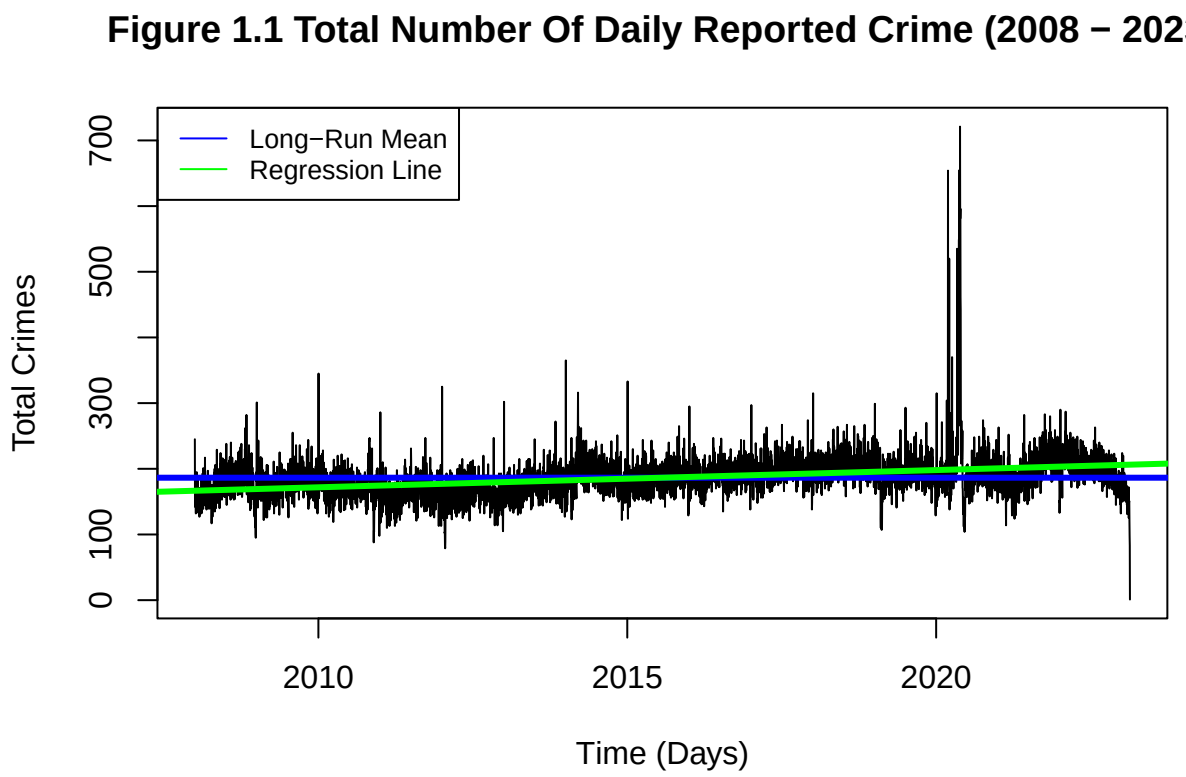
```
# Group data from 2008 to 2023 by date of crime committed
crime_counts <- recentSpd %>%
  group_by(Date.Of.Crime) %>%
  summarize(Total.Offenses = n()) %>%
  arrange(Date.Of.Crime)

# Create time series data with a daily frequency
crime_ts_data <- ts(crime_counts$Total.Offenses, start = c(year(crime_counts$Date.Of.Crime)[1],
              month(crime_counts$Date.Of.Crime)[1]), frequency = 365)

# Linear Regression Model for Time Series
model <- lm(crime_ts_data ~ time(crime_ts_data))
summary(model)
```

```
##
## Call:
## lm(formula = crime_ts_data ~ time(crime_ts_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205.30  -17.20   -1.84   14.48  521.91
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -5.101e+03  1.988e+02  -25.66   <2e-16 ***
## time(crime_ts_data)  2.623e+00  9.862e-02   26.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.04 on 5524 degrees of freedom
## Multiple R-squared:  0.1135, Adjusted R-squared:  0.1134
## F-statistic: 707.6 on 1 and 5524 DF,  p-value: < 2.2e-16
```

```r
# Construct a time series graph to model total number of daily reported crime
crime_ts_plot <- plot(crime_ts_data, xlab = "Time (Days)", ylab = "Total Crimes",
     main = "Figure 1.1 Total Number Of Daily Reported Crime (2008 - 2023)")
abline(h = mean(crime_ts_data), col="blue", lwd = 3)
abline(model, col="green", lwd=3)
legend("topleft", legend=c("Long-Run Mean", "Regression Line"),
       col=c("blue", "green"), lty=1, cex=0.8)
```



Upon examination of Figure 1.1, it can be observed that there has been a general increase in the number of daily reported crimes in Seattle over a period of fifteen years (2008 - 2023). Since the p-value of the linear regression model is significantly lower than the alpha 0.05, there is sufficient evidence to believe that the number of daily reported crimes over the last fifteen years has been changing over time. It is worth noting that the data shows a significant spike in the middle of 2020, which could potentially skew the overall results. In order to further investigate this anomaly, a time series analysis will be conducted on crime data recorded since the beginning of the pandemic (March 11, 2020).

```r
#Filter the data to be after the pandemic
postCovidData <- spd %>%
  filter(spd$Date.Of.Crime >= as.POSIXct("2020-03-11"))
covid_crime <- postCovidData %>%
  group_by(Date.Of.Crime) %>%
  summarize(Total.Offenses = n()) %>%
  arrange(Date.Of.Crime)

# Create a time series data for Covid crimes with daily frequency
covid_crime_ts <- ts(covid_crime$Total.Offenses, start = c(year(covid_crime$Date.Of.Crime)[1],
```

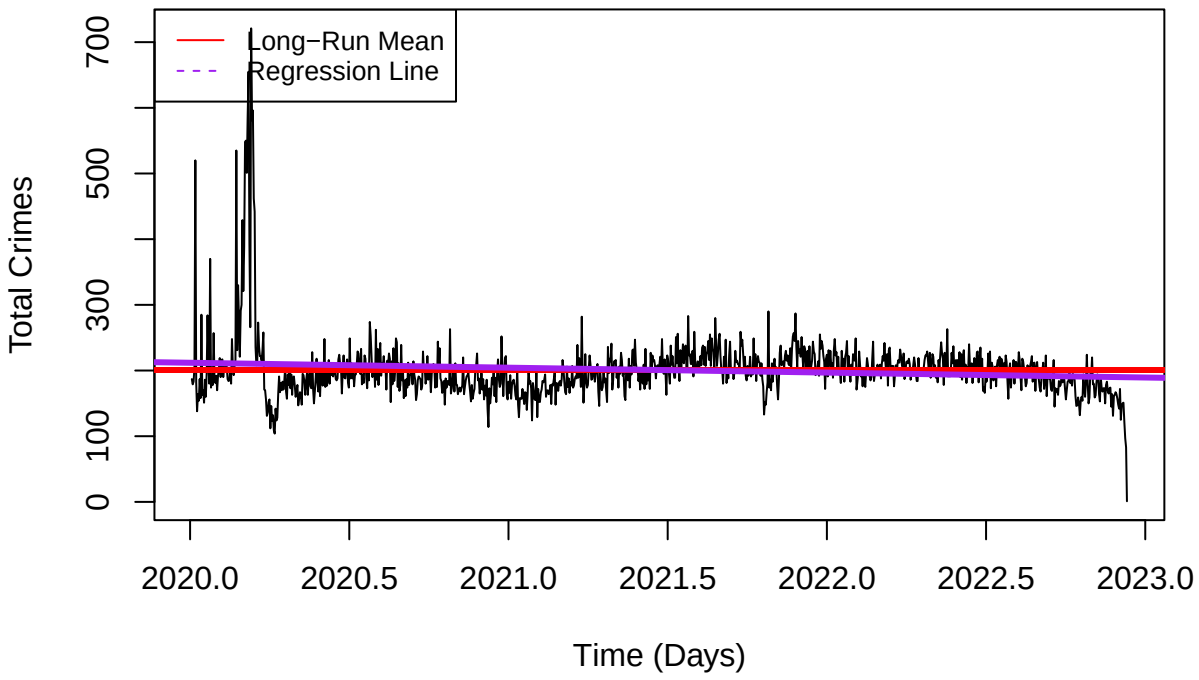
```
                  month(covid_crime$Date.Of.Crime)[1]), frequency = 365)

covidModel <- lm(covid_crime_ts ~ time(covid_crime_ts))
summary(covidModel)
```

```
##
## Call:
## lm(formula = covid_crime_ts ~ time(covid_crime_ts))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -188.83  -23.86   -3.14   15.07  510.59
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          15326.326   3604.992   4.251 2.31e-05 ***
## time(covid_crime_ts)    -7.482      1.783  -4.196 2.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.57 on 1071 degrees of freedom
## Multiple R-squared:  0.01617,    Adjusted R-squared:  0.01525
## F-statistic:  17.6 on 1 and 1071 DF,  p-value: 2.945e-05
```

```
# Construct a time series graph to model total number of daily reported crime after the pandemic
covid_crime_plot <- plot(covid_crime_ts, xlab = "Time (Days)", ylab = "Total Crimes",
     main = "Figure 1.2 Total Number Of Daily Reported Crime (2020 - 2023)")
abline(h = mean(covid_crime_ts), col="red", lwd = 3)
abline(lm(covid_crime_ts ~ time(covid_crime_ts)), col="purple", lwd=3)
legend("topleft", legend=c("Long-Run Mean", "Regression Line"),
       col=c("red", "purple"), lty=1:2, cex=0.8)
```

Figure 1.2 Total Number Of Daily Reported Crime (2020 − 2023)



Based on the information presented in Figure 1.2, it appears that the total number of daily reported crimes has remained relatively unchanged, as indicated by the relatively flat regression line. In order to gain a more comprehensive understanding of the sudden increase in reported crime, a table consisting of crime types and their absolute frequencies during May 2020 will be examined.

```
may2020 <- spd %>%
  filter(month(spd$Date.Of.Crime) == 5)

mayCovidOffense <- may2020 %>%
  group_by(Offense) %>%
  summarise(TotalOffenses = n())

mayCovidOffenseGroup <- may2020 %>%
  group_by(Offense.Parent.Group) %>%
  summarise(TotalOffenses = n()) %>%
  arrange(desc(TotalOffenses))

mayCovidOffenseGroup
```

```
## # A tibble: 31 x 2
##    Offense.Parent.Group                  TotalOffenses
##    <chr>                                         <int>
##  1 LARCENY-THEFT                                 32730
##  2 ASSAULT OFFENSES                              13074
##  3 FRAUD OFFENSES                                10265
##  4 BURGLARY/BREAKING&ENTERING                     9819
```

```
##  5 DESTRUCTION/DAMAGE/VANDALISM OF PROPERTY          8357
##  6 MOTOR VEHICLE THEFT                              5015
##  7 TRESPASS OF REAL PROPERTY                        2648
##  8 DRUG/NARCOTIC OFFENSES                           2209
##  9 ROBBERY                                          1978
## 10 DRIVING UNDER THE INFLUENCE                      1511
## # ... with 21 more rows
```

As depicted in the above table, the surge in the total number of daily reported crimes is attributed to an increase in Larceny - Theft, Fraud, and Assault offenses. This occurrence is rationalized within the context of the surge in Asian hate crimes and theft, which could be attributed to the unemployment rate increase among Seattle citizens, leading to a rise in criminal activity and credit card fraud regarding online shopping.

Analysis

Assuming the null hypothesis to be valid (lack of statistically significant correlation between time and total daily reported crime in Seattle), the probability of observing the trend illustrated in Figure 1.1 is 2.2 x 10^-16. Hence, the null hypothesis is rejected, and there is convincing evidence supporting the presence of a correlation between time and total daily reported crimes in Seattle.

2) In order to maximize safety and societal well-being, which precincts are in need of staff (namely, police reports) the most, and for which times of day?

```
# Formatting the datetime and create a new TimeGroup column
postCovidData$Report.DateTime <- as.POSIXct(postCovidData$Report.DateTime,
                                            format="%m/%d/%Y %I:%M:%S %p")

timeOfReport <- as.numeric(format(postCovidData$Report.DateTime, "%H%M%S"))
postCovidData$TimeGroup <- cut(timeOfReport, breaks =
                           c(0, 30000, 60000, 90000, 120000, 150000, 180000, 210000, 240000),
                     labels = c("12 AM to 3 AM", "3 AM to 6 AM", "6 AM to 9 AM",
                                "9 AM to 12 PM", "12 PM to 3 PM", "3 PM to 6 PM",
                                "6 PM to 9 PM", "9 PM to 12 AM"),
                     include.lowest = TRUE)
```

In order to investigate this inquiry, an analysis of recent data (after the pandemic) should be used since it has been concluded in the previous question that there is a disparity between pre-pandemic and post-pandemic data. As such, the focus is on identifying the precincts that require additional staffing and determining the most vulnerable periods of the day.

The comparison will encompass the total number of crimes committed, specifically assessing whether these crimes are categorized as violent, and the corresponding time period of occurrence.

```
# Assess the number of violent crimes in Seattle
violentData <- postCovidData %>%
  filter(Precinct != "UNKNOWN" & Precinct != "" & Precinct != "<Null>" & Precinct != "OOJ") %>%
  mutate(Violent.Crime = Offense %in% c("Robbery", "Rape", "Murder & Nonnegligent Manslaughter",
                                        "Aggravated Assault")) %>%
  group_by(Precinct, Violent.Crime) %>%
  summarise(TotalOffenses = n()) %>%
  pivot_wider(names_from = Precinct, values_from = TotalOffenses)
```

```
## `summarise()` has grouped output by 'Precinct'. You can override using the
## `.groups` argument.
```

```
violentData
```

```
## # A tibble: 2 x 6
##   Violent.Crime     E     N     S    SW     W
##   <lgl>         <int> <int> <int> <int> <int>
## 1 FALSE         32952 68995 27601 20653 48766
## 2 TRUE           2594  3949  2613  1391  4385
```

```
chisq.test(violentData)
```

```
## Warning in chisq.test(violentData): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  violentData
## X-squared = 571.16, df = 5, p-value < 2.2e-16
```

The objective is to evaluate if there is a discrepancy in the number of violent crimes committed across different precincts in Seattle. Based on the definition provided by Seattle.gov, violent crimes include robbery, rape, homicide, and aggravated assault. By utilizing Pearson's chi-squared test of independence, the observed data has a probability of 2.2*10^-16 under the null hypothesis that there is no noteworthy variance in violent crimes per precinct. Consequently, there is compelling evidence to suggest that certain precincts experience a higher incidence of violent crimes, indicating a requirement for more experienced personnel.

The West and North precincts (represented by "W" and "N", respectively) exhibit a significant number of violent crimes. Consequently, these precincts are ideal candidates for analysis to determine the required staffing levels for optimal safety. Specifically, the analysis will involve assessing the total crimes occurring at different times of the day in these precincts to determine when additional staffing is required.

```
# Creating a comparison table across the N and W precincts and time intervals in the day
compare <- postCovidData %>%
  filter(Precinct == "W" | Precinct == "N") %>%
  group_by(Precinct, TimeGroup) %>%
  summarise(TotalOffenses = n()) %>%
  arrange(desc(TotalOffenses))
```

```
## `summarise()` has grouped output by 'Precinct'. You can override using the
## `.groups` argument.
```

```
compare
```

```
## # A tibble: 16 x 3
## # Groups:   Precinct [2]
##    Precinct TimeGroup     TotalOffenses
##    <chr>    <fct>                 <int>
##  1 N        12 PM to 3 PM         17202
##  2 N        9 AM to 12 PM         13851
##  3 N        6 AM to 9 AM          11935
##  4 W        12 PM to 3 PM         11070
##  5 N        3 PM to 6 PM          10304
##  6 W        9 AM to 12 PM          8992
##  7 W        6 AM to 9 AM           7946
##  8 W        3 PM to 6 PM           7714
##  9 N        6 PM to 9 PM           6663
## 10 N        9 PM to 12 AM          5745
## 11 W        6 PM to 9 PM           5624
## 12 W        9 PM to 12 AM          5123
## 13 N        12 AM to 3 AM          4117
## 14 W        12 AM to 3 AM          3746
## 15 N        3 AM to 6 AM           3127
## 16 W        3 AM to 6 AM           2936
```

Based on the table, the North and West precincts deal with the most reported crimes around the 9 AM to 3 PM interval, suggesting that staffing is needed there the most. Therefore, amongst the 5 precincts in Seattle, the North and West precincts should be prioritized with the most staffing with a heavy emphasis on the time interval between 9 AM to 3 PM for shifts.

Conclusion

The analysis of the Seattle Crime Data: 2008-Present dataset has provided insights into two key research questions. Firstly, the analysis showed that there is a statistically significant correlation between time and total daily reported crimes in Seattle, indicating that Seattle has become more dangerous over time. Specifically, the surge in the total number of daily reported crimes was attributed to an increase in Larceny-Theft, Fraud, and Assault offenses, which is potentially due to an increase in Asian hate crimes and theft, leading to a rise in criminal activity and credit card fraud regarding online shopping.

Secondly, the analysis revealed which precincts require additional staffing to maximize safety and societal well-being, and during which times of day, specifically the high volume of reported crimes to the North and West precincts, and their need for the most staff in the general 9 AM to 3 PM time interval.

By analyzing crime trends with respect to time and staffing needs of each precinct and determining specific time intervals where more staff is required, resources can be better allocated, and community safety prioritized. Overall, the results of this study have implications for Seattle policymakers to enhance community safety and well-being.

Suggestions for Future Research

Although the study investigates crime and staffing trends in Seattle, it does not offer a thorough contextual analysis of the underlying factors contributing to these trends. Despite identifying a correlation between time and reported crimes, the study does not delve into potential explanations for this correlation. In the future, the team aims to further investigate relevant data and analyze potential causes for the general trend of rising reported crimes in Seattle.

It is noteworthy that this study draws on a solitary dataset of Seattle crime data, raising concerns regarding the representativeness of crime patterns in other cities or regions. Furthermore, the dataset contains missing and erroneous values that were subsequently removed, potentially impacting the accuracy of the analysis. For instance, a considerable number of precincts were labeled as various forms of "N/A" such as "" and "UNKNOWN." As such, it is suggested that a more robust approach to data collection and validation be implemented for this dataset.