LA Crime Report Analysis

Oluwatomi Hassan

2023-10-22

Introduction

The raw dataset shows Los Angeles California crime incidents in from 2020 till 2023. This data is compiled from crime reports on paper before being entered into the database with variables in such as report date, occur date, occur time, area of crime, type of crime, and victim info(age,sex,and descent). The aim of this study is to answer the following questions: Does the premise contribute to the odds of the victim gender? Are there any seasonality in the numbers of crime reported in a given year? Does the count of reported crime depend on the neighborhood area?. Time series analysis was performed to explore any seasonality and trend in the number of crime reported in LA within those two years. To determine the gender(male) of the victim depends on the premise, logistic regression is performed on the data. Poisson regression is used to examine if the differences in the number of reported crime in depends on the L.A neighborhood.

Models and Methods

The raw data was cleaned to group the premise description in the crime reports into five premise groupings of street/sidewalk, parking lots, house, apartment and stores. Additionally, I extracted crime reported by male and female victims to examine association between gender and premise. The raw data was aggregated to the number of count of crime reported daily from 2020 till 2022. Exploratory data analysis is performed of the variable of interest is performed aand statistical analysis is conducted to determine if the findings are significant.

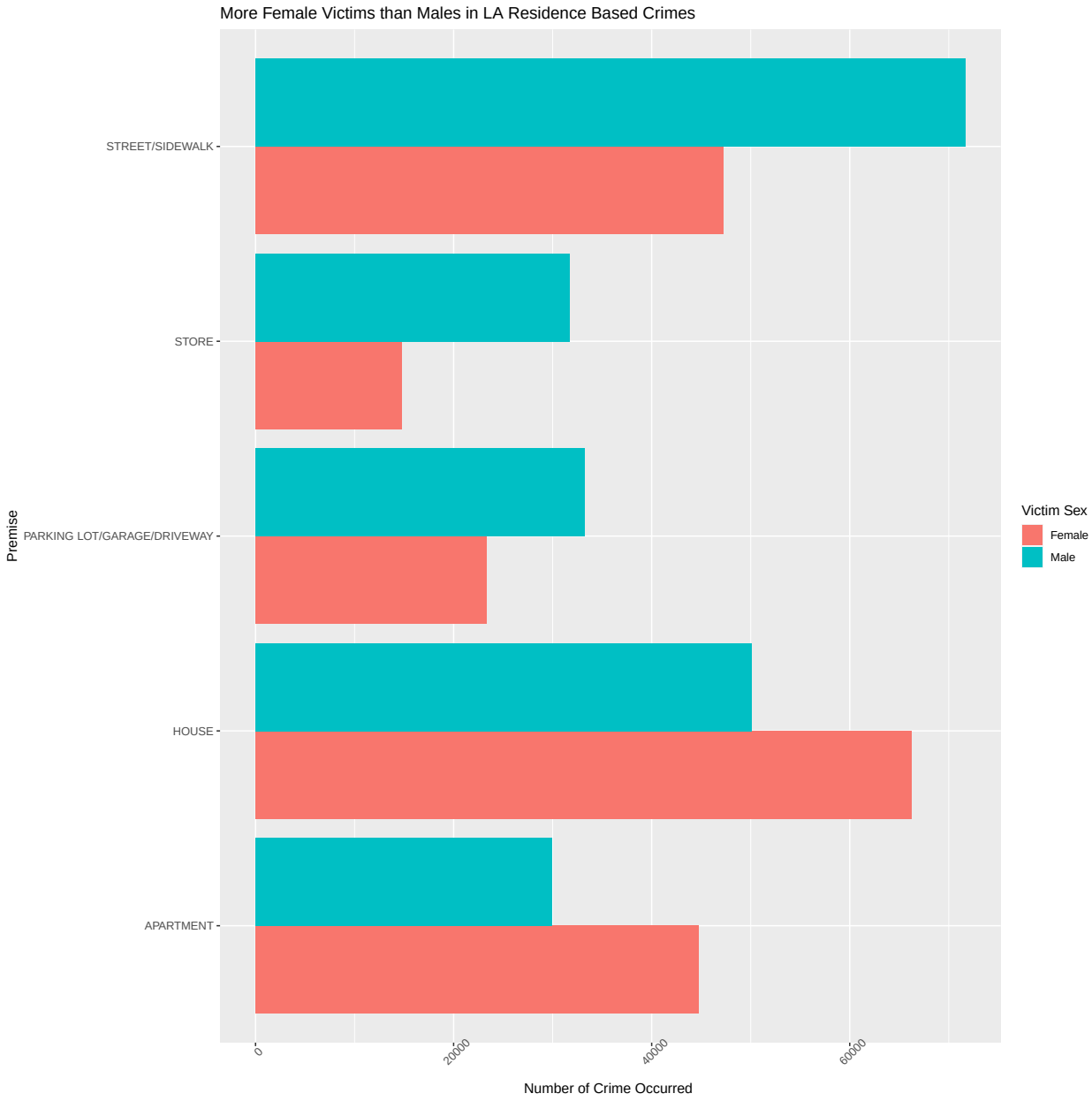
The glm() function in *R* caret package is used to determine any association between crime reported in various premise groupings and the gender(male or female) of the victim. A logistic regression is modeled with gender as the response variable and categorical variable premise as the explanatory variable. The model to fit is $logit(p) = B_1Apartment - B_2House + B_3Parking + B_4Store + B_5Street$. To evaluate the performance of the model, the AIC and drop in deviance test of the full model and null model is compared. To determine if there are differences of in the number of crime reported in each LA neighborhood. Poisson regression is performed and drop in deviance test is used to compare the full model with the neighborhood to the null model. The full model is as described $log(\lambda_i) = \beta_0 + \beta_1 Neighborhood_{1i} + \cdots + \beta_k Neighborhood_{ki}$ where k is the 1- number of neighborhoods and i is the neighborhood. To examine the trend and seasonality of the daily crime reported, the R function decompose() was use to decompose a time series into trend, seasonality and irregular or random components using moving averages. The random plot of the time series plot is accessed to evaluate the trend and seasonality observed.

Further studies can be performed to see if male or females are victims of different types of crimes such as assault or theft and if these crime rate changes with premise and location.

Results

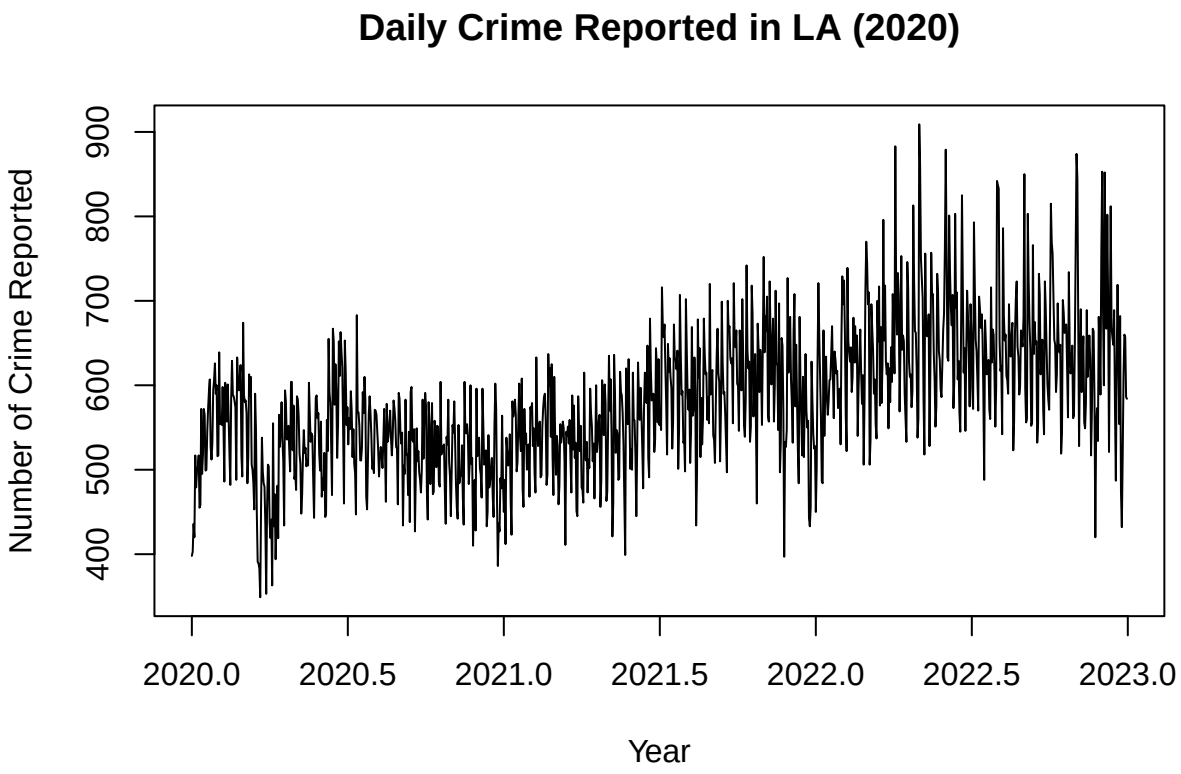
Logistic Regression - Probability of Gender based on Premise

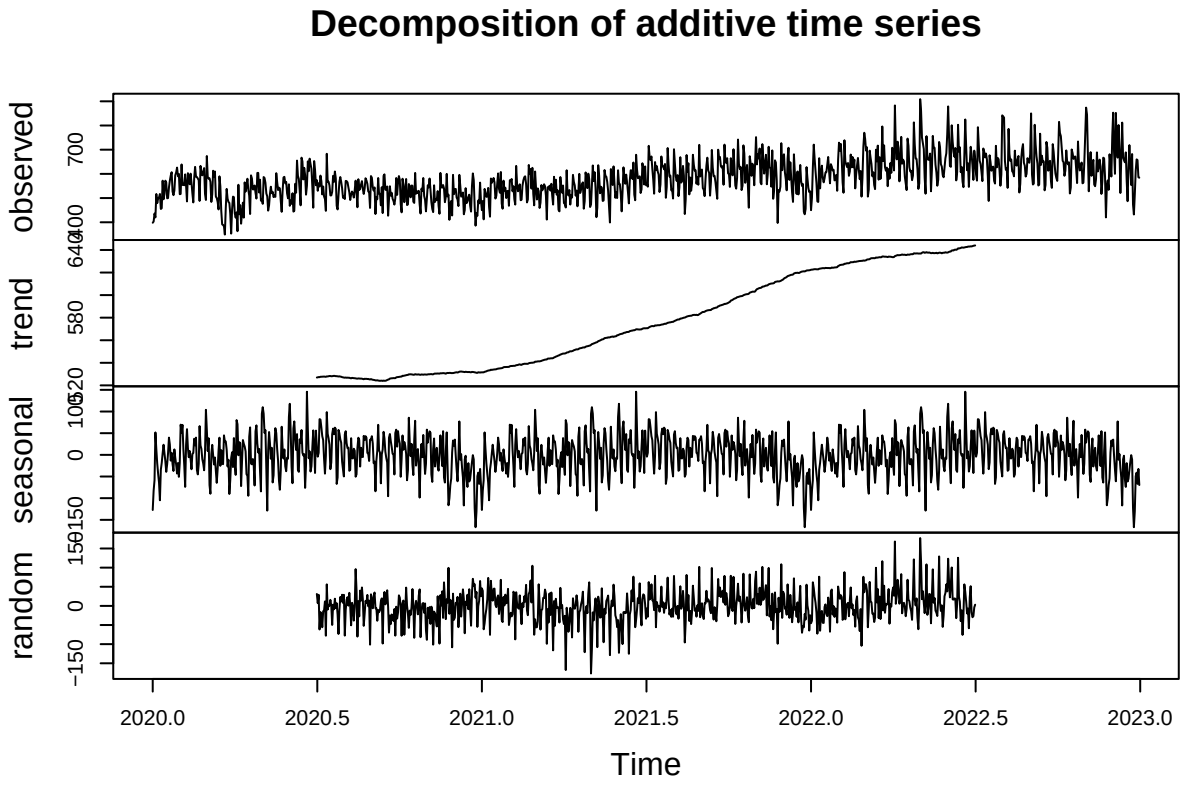
In Los Angeles(LA), exploratory data analysis suggests that there are more female victims in residence-based crimes than male victims. For crimes committed outside(parking lot or street/sidewalks) or in a store, there are more male victims in Los Angeles. Logistic regression is performed to find association between premise and the odds of the victim gender. The glm function in R caret package is used to perform a logistic regression with gender as the response variable and the five premise groupings as the explanatory variable. The model is fit without a response variable to facilitate with interpretation. The full logistic model is $logit(p) = -0.40Apartment - 0.28House + 0.35Parking + 0.76Store + 0.42Street$. The AIC for the model with Premise is smaller compared to the null model, suggesting that the full model is a more appropriate fit. There is convincing evidence in favor of the full model suggesting that the odds of a male victim differs for different premise(Drop in Deviance test, p-value= <2.2e-16). Although the dispersion parameter is estimated to be 1, a dispersion value of 1.34 suggests evidence of overdispersion not accounted for in the model. With a small value of <2e-16, the model coefficient suggests that the odds of a victim being male is about 0.67 and 0.76 for crimes that occur in an apartment and house, respectively. The odds of the victim being male for crimes committed on the street and parking lot is 1.52 and 1.42, respectively(p-value ~ 2e-16). When the premise is a store, the odds of the victim being male is about 2.14(p-value ~ 2e-16).



Time series Analysis - Daily Crime reported 2020-2022

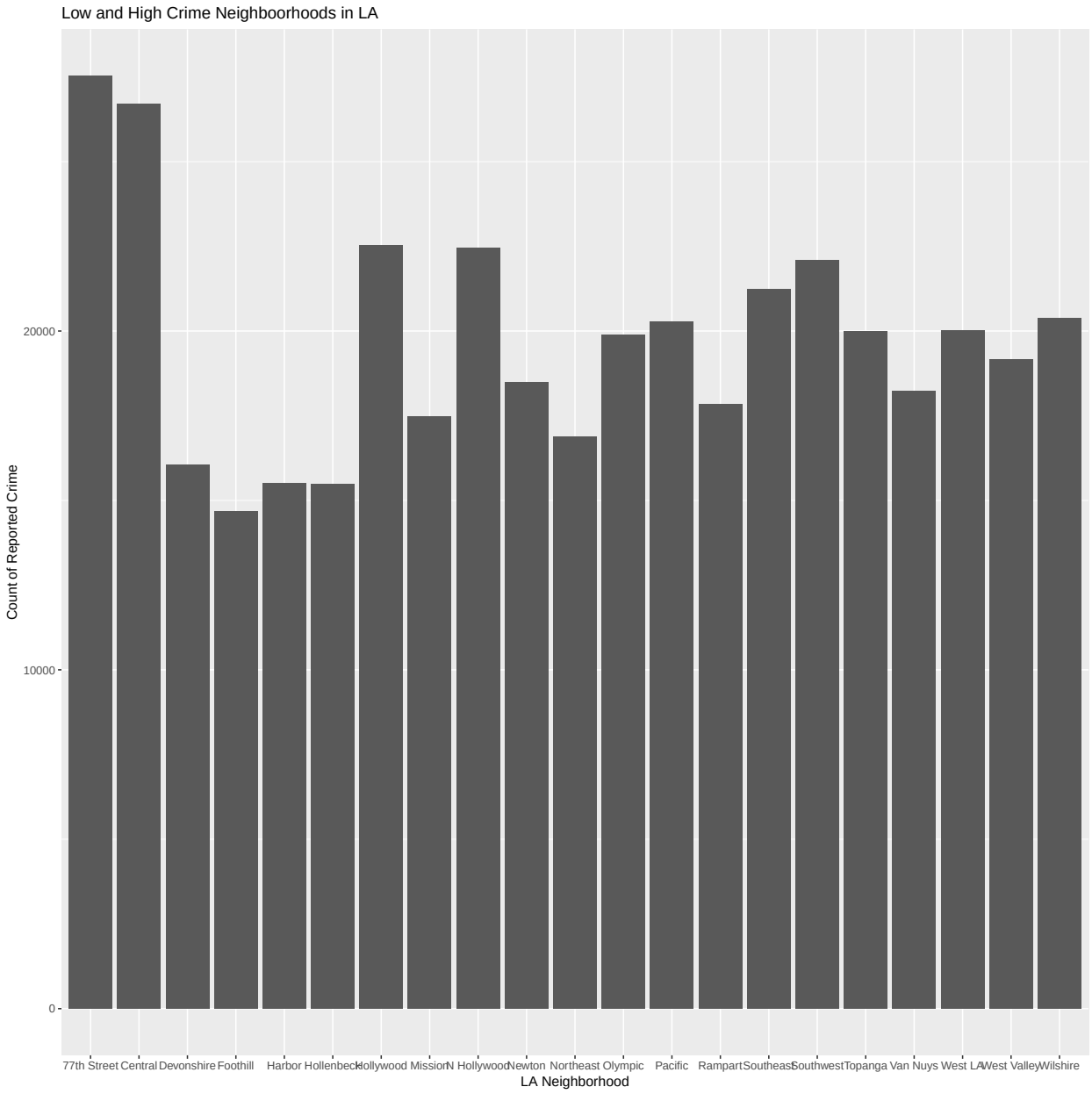
We see a gradual rise in the amount of crimes report grow from April till June and July which decreases gradually in August to december when the daily number of crime reported is plot from 2020-22. The decomposed time series plot suggest that there is a spike in crime reported in the summer months and decline in crime reported in the winter months. Analysis of the decomposed trend plot suggests that there is an increase in trend of daily crime reported with more crimes being reported in recent years compared to previous years. The decomposed random plot suggests that there is no observable pattern or seasonality in the data after the trend and seasonality is removed from the time series of daily crime reported.


Daily Crime Reported in LA (2020)
Number of Crime Reported
400
500
600
700
800
900
2020.0
2020.5
2021.0
2021.5
2022.0
2022.5
2023.0
Year



Poisson Regression - Difference in Crime reported by Neighborhood

The full model is There is convincing evidence in favor of the full model suggesting that there is a difference in the number of crime reported neighborhoods in LA (Drop in Deviance test, p-value= <2.2e-16).



Conclusions

Using logistic regression, we can conclude that the odds of the victim being male is increased for crime committed in the store, parking lot and street/sidewalk. With evidence of overdispersion not accounted by the full model of the logistic regression, other models that should be explored. A time series analysis of the aggregated count of the daily crime reported suggests that there is seasonality in the number of daily reported crime over this two-year period with the highest recorded number of crimes in the summer and lowest in the winter. Further analysis can be conducted to examine this seasonality over a longer time period i.e 5-years. The poisson regression model fit to the data suggests that there is a difference in the number of crime reported in each LA neighborhood between 2020-22. Additional analysis can be performed to determine the extent of the differences observed.

Appendix

```
# read the raw dataset
la_crime <- read.csv("./Crime_Data_from_2020_to_Present.csv",header=TRUE)
head(la_crime)

# data cleaning

## creating premise groupings, crime type and extracting 2020-2022 data

la_crime_df_2020 <- la_crime %>% mutate(Date.Rptd = mdy_hms(Date.Rptd),DATE.OCC = mdy_hms(DATE.OCC), Pr

## extracting data with only premise of interest
la_crime_df <- la_crime_df_2020[la_crime_df_2020$Premise %in% c("APARTMENT","STORE","HOUSE","PARKING LO

## extracting data with only male or female victims
la_crime_df <- la_crime_df[la_crime_df$Vict.Sex%in% c("M","F"),]

## encode victim sex to binary for logistic regression
la_crime_df <- la_crime_df %>% mutate(gender = ifelse(Vict.Sex == 'M', 1, 0))


## Data on daily crime reported from Jan 1, 2020 - Dec 31, 2022
daily_la_crime_df <- la_crime_df_2020 %>% group_by(Date.Rptd) %>% summarize(crime_count = n())
daily_la_crime <- daily_la_crime_df[,2]
## convert dataframe to time series
daily_la_crime <- ts(daily_la_crime, start = c(2020,1), frequency = 365)

# Exploratory data analysis

## EDA - Crime reported by Premise and Gender (logistic regression)
par(mfrow = c(4, 6))
ggplot(data= la_crime_df,aes(x=Premise, fill=Vict.Sex)) +  geom_bar(position = "dodge") + labs(title =
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5)) + coord_flip()

## Time series of daily crime reported between 2020-22
plot(daily_la_crime, main = "Daily Crime Reported in LA (2020-22)")


## Poisson Regression - Differences in crime report count in Neighborhood

ggplot(data= la_crime_df,aes(x=AREA.NAME))+  geom_bar()  + xlab("LA Neighborhood") + ylab("Count of Rep

# Statistical Analysis

## logistic regresion on gender with Premise
premise_gender_logit <- glm(gender ~ Premise - 1 ,data = la_crime_df, family = "binomial")
summary(premise_gender_logit)
exp(cbind(OR = coef(premise_gender_logit), confint(premise_gender_logit)))

## Decomposed time series analysis
la_crime_dec <- decompose(daily_la_crime)
plot(la_crime_dec)
```

```r
## Poisson model
summary(pois1 <- glm(crime_reported ~ AREA.NAME, family="poisson", data=la_area_df))
summary(pois_null <- glm(crime_reported ~ 1, family="poisson", data=la_area_df))

# Evaluation

## logistic regression
(disp <- premise_gender_logit$deviance/premise_gender_logit$df.residual)

### Drop in Deviance test for null and full model
anova(null_gender_logit, premise_gender_logit,test="Chisq")

## Time series residuals

plot(la_crime_dec$random)
```