Patterns of Auto Theft Occurrences in City of Toronto*

Yiyi Ren 1005979103

4/27/2022

Abstract

Occurrences of auto thefts is an important indicator to show whether a city is safe. Summarizing the pattern, analyzing the results of previous auto theft cases, would be able to help preventing the future occurrences. By analyzing the data, also with the adopted linear regression model, it is able to conclude that neighborhoods which have the largest areas tend to have higher occurrences of auto thefts, especially during nights.

Introduction

Auto theft is one of the Major Crime Indicators. The number of auto theft occurrences of neighborhood indicate the safety level of them. Losing of car can be very hard to accept, harmful for one's emotion, even will have consequences including inconvenience and feelings of personal violation. The prevention of auto theft occurrences is crucial for every family and individuals. The city of Toronto is one of the most important and busiest cities in Canada. Despite the importance of preventing auto thefts, there is not very much precious work done to explore this potential connection. This project aims to summarize the pattern of auto theft happened in Toronto from 2014 to 2018, finding the pattern of the occurrences of auto thefts. Hence, provide valid suggestions of how to efficiently prevent auto thefts to Toronto Police. The rest of the paper was organized into three major sections. In the Data section, I would first explain the source of the basic data and the methodology I used to process the data. Data definitions and characteristic are explained and included in this major part. In the Results section, I plotted three different figures. Figure 1 shows the distribution of occurrences in terms of time and premise type. Figure 2 is an visualization of the frequency of occurrences, putting in Toronto's map. In figure 3, I adopted a linear regression model for the relationship between area of neighborhoods and occurrences. It was rational to conclude that

Data

The datasets that were used in this project are all provided and downloaded from Open Data Toronto. The City of Toronto's Open Data Portal is an open source delivery tool which contains a various number of datasets feathering the city of Toronto. The dataset was processed and analyzed in R (R Core Team 2020) and I analyzed all these using R package including: opendatatoronto(Gelfand 2020), dplyr(Wickham et al. 2021), tidyverse(Wickham et al. 2019), patchwork(Pedersen 2020) and ggplot2(Wickham 2016). There are a variety of variables included in the dataset of Major Crime Indicators includes crime categories of Assault, Break and Enter, Auto Theft, Robbery and Theft Over. According to the selected subject, only all the variables about Auto Theft were selected.

*Code and data are available at https://github.com/Yiyi0423/sta-304.git

Data Source and Methodology

Methodology The two datasets of 'auto_theft' and 'neighbourhood' are joined together by the shared variable "Hood_ID." Informative and useful variables are selected among all variables. A new variable 'area' is mutated by pop_2016 divided by pop_density_per_square_km. Another new variable 'is_night' would be TRUE is 'occurrencehour' is less than 6 or greater than or equal to 18, otherwise FALSE. 'occurrencedayofweekn' is weekday converted to number. Overall, there were 30 variables and 18,175 observations. Specially, the variables were:

1. Index_ and event_unique_id: The unique identifier of the report and the event.
2. occurrencedate, occurrenceyear, occurrencemonth, occurrenceday, occurrencedayofyear, occurrencedayofweek and occurrencehour: Time information of the occurrence, including date, year, month day, day of year, day of week and hour.
3. reporteddate, reportedyear, reportedmonth, reportedday, reporteddayofyear and reporteddayofweek: Reported time information of the occurrence, including date, year, month, day, day of year, and day of week.
4. premisetype: Premise where occurrence took place.
5. offence: Offence related to the occurrence.
6. MCI: Type of Major Crime Indicators, all are auto theft.
7. Division: Division where event occurred.
8. Hood_ID and Neighbourhood: Neighbourhood Name and Identificator.
9. Long and Lat: Longitude and Latitude of point extracted after offsetting X and & Coordinates to nearest intersection node.

Result:

Fig.1a: Distribution of occurrences in terms of month

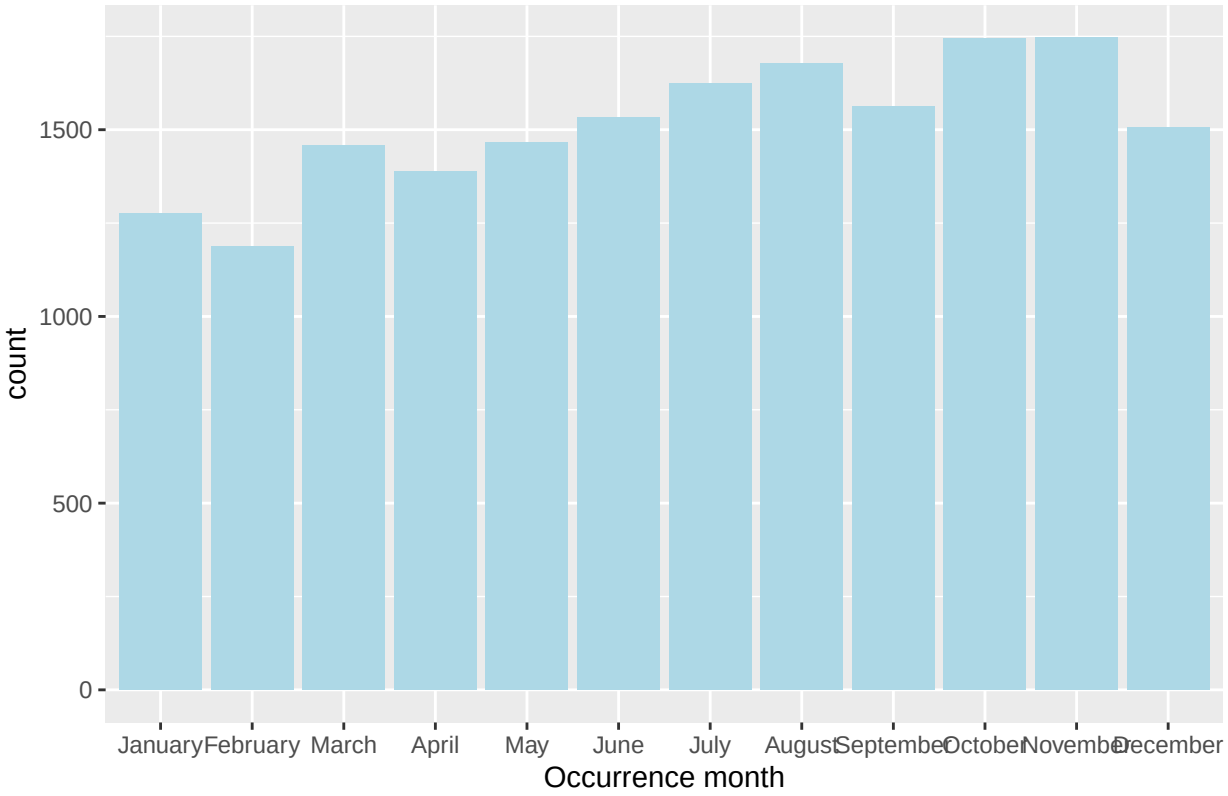


Fig.1b: Distribution of occurrences in terms of days in a week

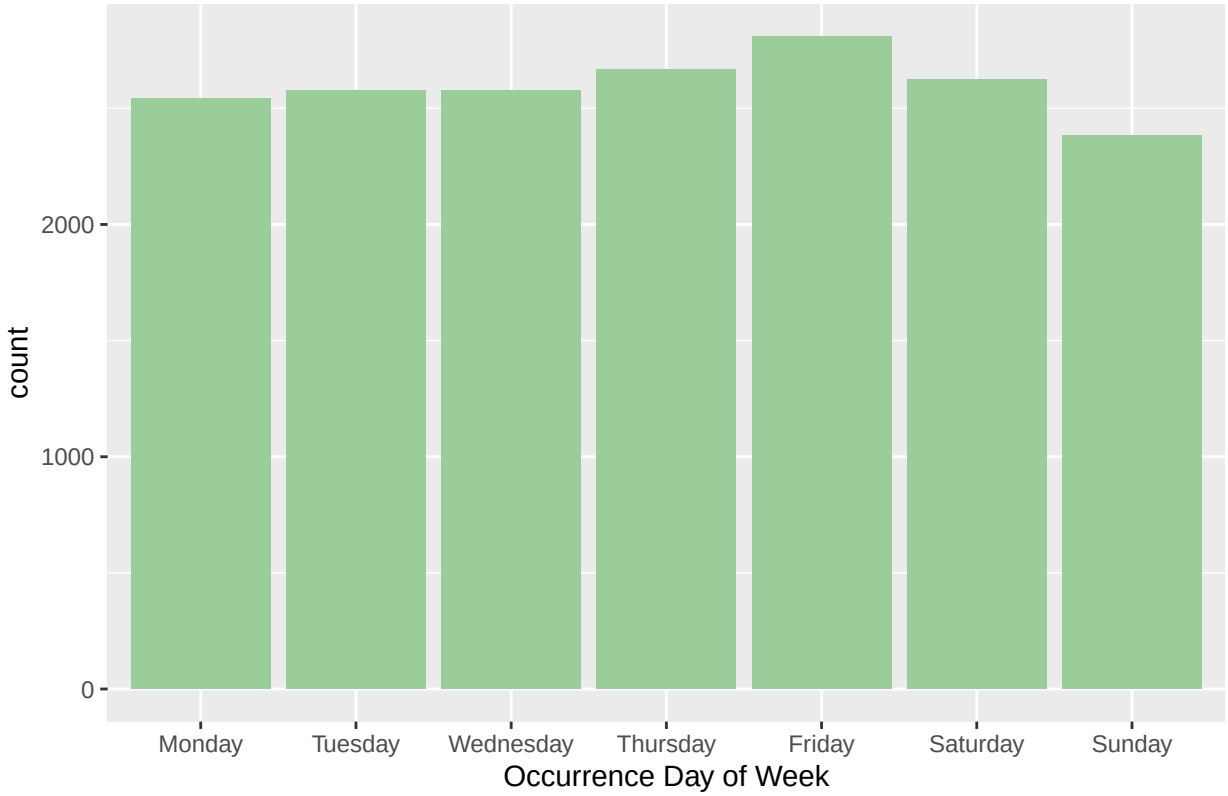
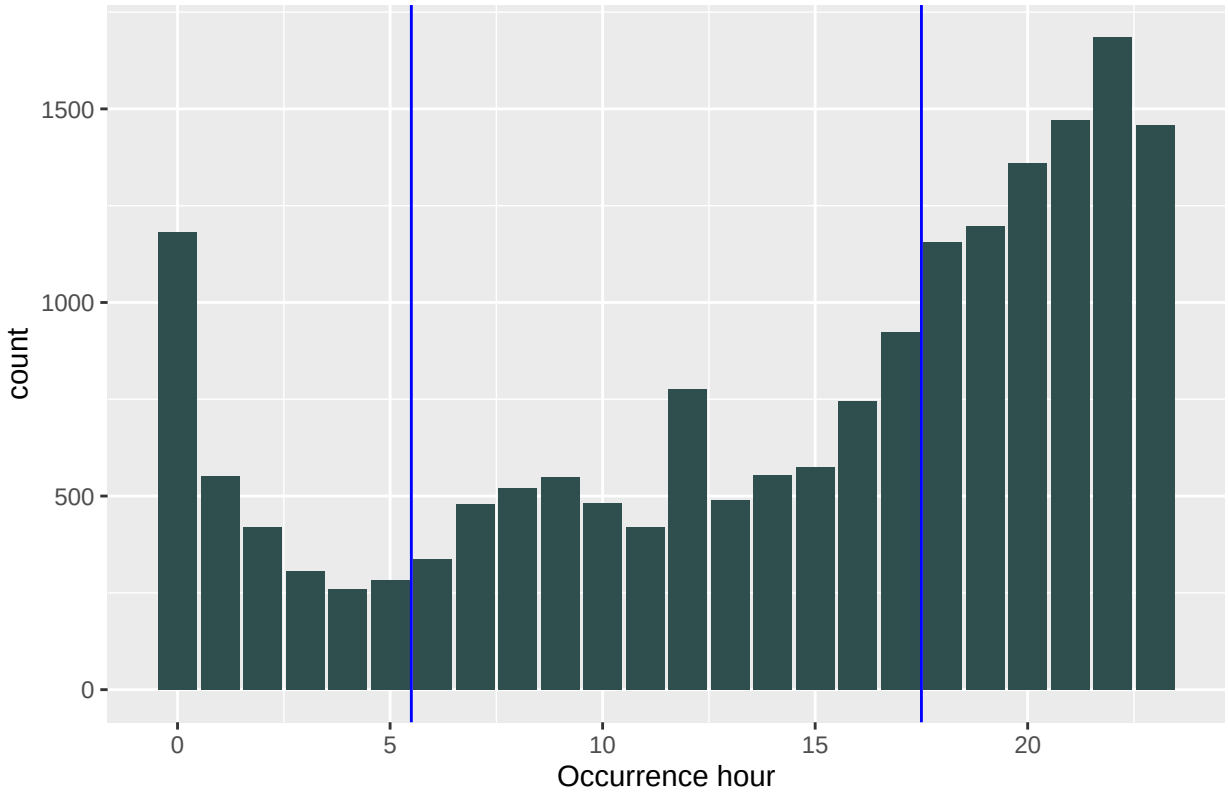


Fig.1c: Distribution of occurrences in terms of hours in a day



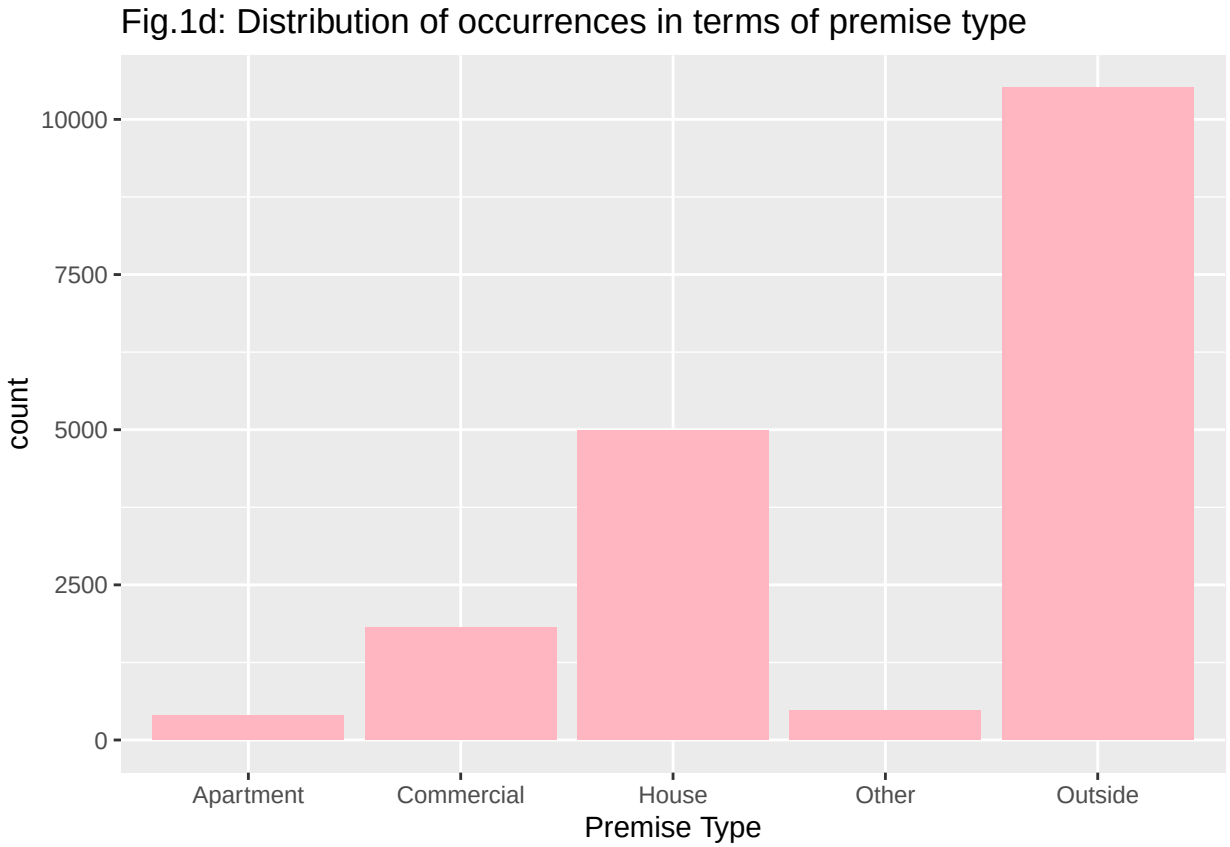


Figure 1 shows the distribution of occurrences by different time variables, including month, week, and hours in a day, also Premise type. Panel a of Figure 1 shows the distribution of occurrences in terms of months, the bar plot illustrates that we January and February had the least number of occurrences, October and November had the greatest number of cases. However, panel b suggests that each day of week had almost equal number of occurrences, with a slightly higher number of cases happened on Friday and a low peak on Sunday. Panel c illustrates the distribution of occurrences in terms of hours in a day. The two v-lines separate occurrences of day (06:00-18:00) and night. From the shape of the graph, there are apparently more occurrences of auto theft cases during the nights. Only around 37.666% of auto thefts cases happened during the daytime. Panel d of figure 1 shows the distribution of occurrence in terms of premise types. There exist obvious differences between the bars. We can see there are a significantly more auto theft cases happened in outside. Also, a considerably auto thefts happened in houses.

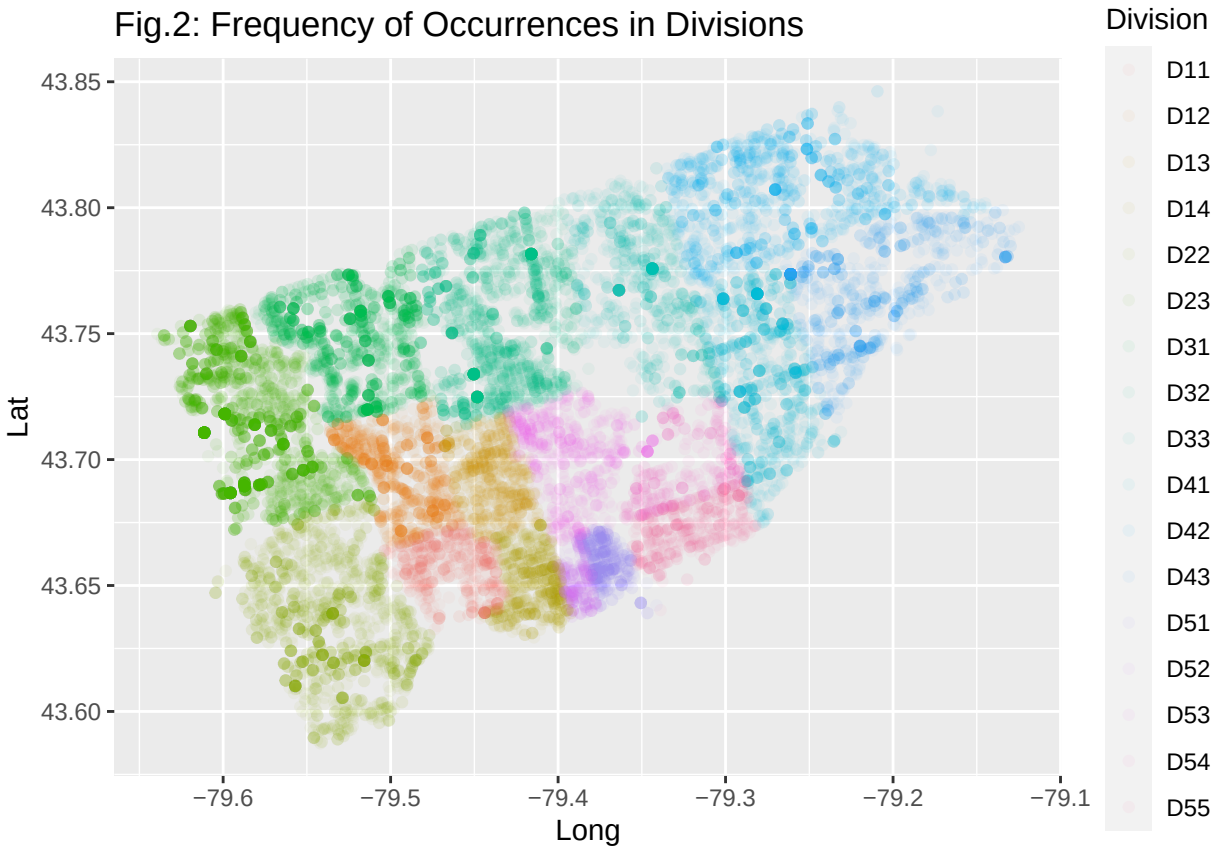
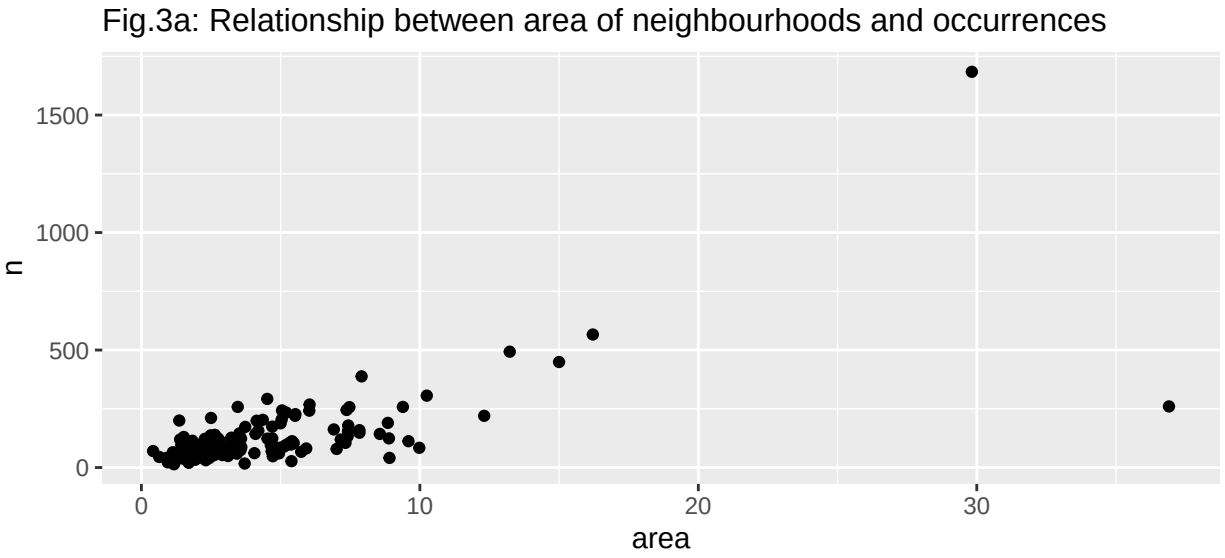
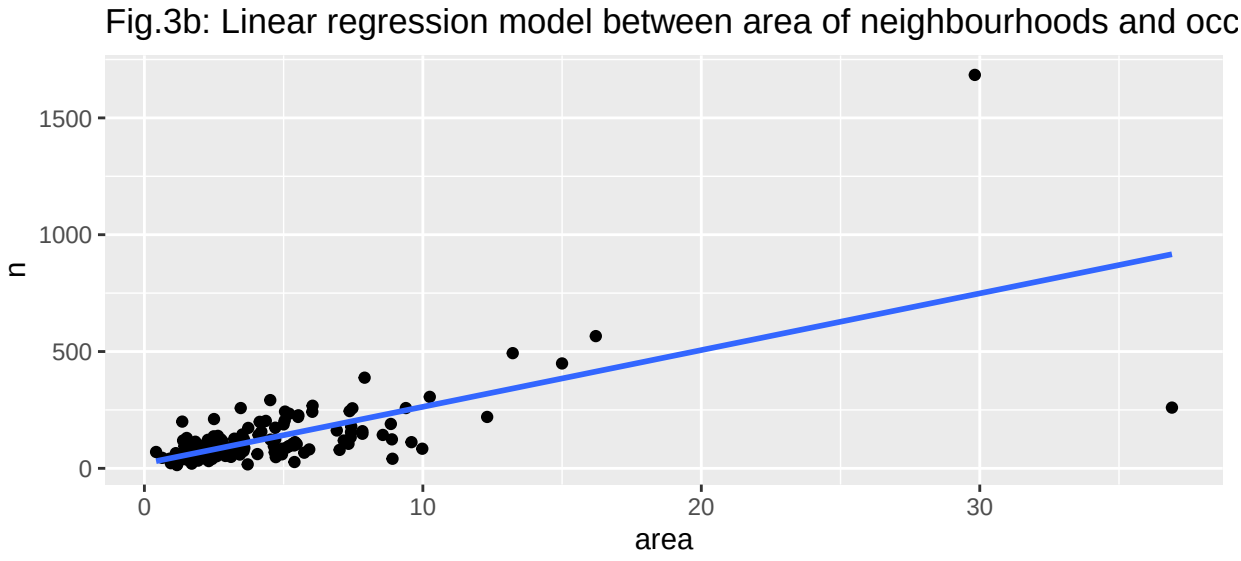


Figure 2 is a scatterplot created by the latitude and longitude of point extracted after offsetting X and Coordinates to nearest intersection node. The plot therefore shows the map of city of Toronto. The different colors represent different divisions. The darker the points are, the more occurrences of auto-thefts happened in this specific location. In each division, there are several dark points, which means the auto thefts cases are separated in the city of Toronto. Each division has

Linear Regression Model





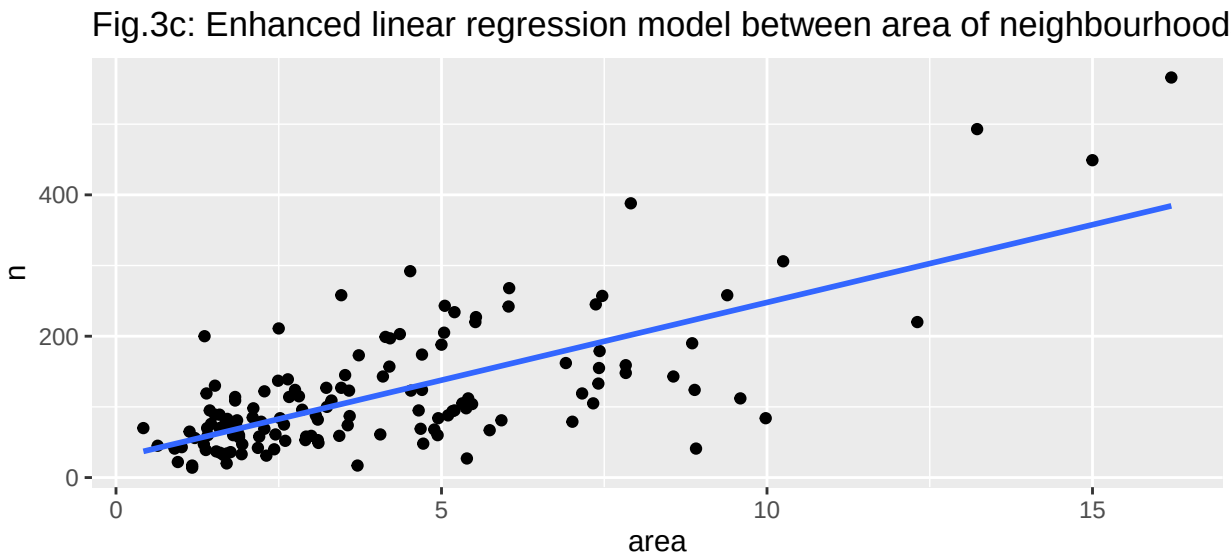


Figure 3 panel a shows the relationship between area of neighborhood and occurrences of auto theft. The variable area, in square kilometers, was calculated by population divided by population density per square kilometer. We then adopted a linear regression model for this relationship as shown in panel b. Outliers including extreme high number of occurrence and very large area, but small number of occurrences were removed. Panel c visualizes the regression model which reduced the outliers. By calculation, the correlation between area and number of occurrences is about 0.7, which indicates the relationship between the area of each neighborhood and occurrences of auto theft is linear, strong and positive.

To check how well this fitted regression line preform as a predictive model, we first randomly picked 80% of the observations in the dataset neighborhood as training dataset. Then made the rest of the observations become testing dataset. After that, training dataset was used to fit the predictive model. Made the predictions for the testing dataset and compare the predictions to the true responses. The RMSE for the predictions in the testing dataset is 2.11139. The RMSE for the predictions in the training dataset is 2.030258. The difference of the RMSEs of the training and testing datasets is not too big, thus, the predictive model is generalizing well.

Discussion

Discussion of Research Findings

There is positively strong linear relationship between number of auto thefts and area of neighborhood. A greater number of auto thefts occurred at nights than days, around 37%-38% of cased happened during days. Therefore, we suggest the Toronto Police to send extra police to those neighborhoods which have the largest areas (Hood_1, Hood_131, Hood_14), especially during nights. The results and findings of the current research can help people selecting household address and neighborhoods in terms of safety, using auto theft occurrences as an example. Moreover, the insurance company can use the results to adopt different vehicle insurance fare.

Weakness

The data set is incomplete due to privacy considerations, which could result to inaccuracy of occurrences count. To diminishing the effect of the pandemic, only the data before COVID-19 was selected to be included in the dataset. General pattern should not be affected, however, the prediction of future a few years would be not as accurate as if the recent two years' data was included. For similar reasons, the resulting prediction may not have long-lasting significance.

Reference

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal.* https://CRAN.R-project.org/package=opendatatoronto.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots.* https://CRAN.R-project.org/package=patchwork.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.