An analysis of Break and Enters in Toronto

Mohsin Reza, Ryan Wong, Joyce Huang, Justin Orial

Group 106-1 | STA130 Final Project

March 30, 2020

Introduction

- We are going to analyze data provided by the Toronto Police Service for 43,302 Break and Enters that were reported in Toronto from 2014 to 2019
- The data contains various facts related to each Break and Enter including the type of premise, neighbourhood, the year, month, day of week and time of day the crime occured in etc.
- We hope to find relevant trends in the data that will assist the Toronto Police Service in preventing Break and Enters in the future.

Objectives

The specific questions about the data that we want to answer are as follows:

1. Do certain police divisions in Toronto have a larger number/proportion of Break and Enters?
2. Do certain hours of the day have a significantly larger number/proportion of Break and Enters?
3. Is there a relationship between the day of the week a Break and Enter occurred and the hour of the day a Break and Enter occurred?
4. Do certain premise types (apartments, houses etc.) have a larger number/proportion of Break and Enters?
5. Do certain premise types have more Break and Enters on certain hours of the day/days of the week?

Data Summary - Part 1

We did all the following to "clean" the data:

- Removed all variables except *premisetype*, *occurenceyear*, *occurencemonth*, *occurencedayofweek*, *occurencehour*, and *Division*.
- We only kept these variables as we felt that those were the only variables that would help us answer our research questions.
- We created a new variable, which we named *is_midnight*. We assigned this variable a value of "midnight" if the hour of occurrence of the Break and Enter was 0 and "not midnight" otherwise

Data Summary - Part 2

- We created a new variable, which we named *day_number*. The value of *day_number* was assigned based on the value of *occurencedayofweek* - the table below shows the value *day_number* was assigned based on the value of *occurencedayofweek*:

Value of *occurencedayofweek*	Value of *day_number*
Monday	1
Tuesday	2
Wednesday	3
Thursday	4
Friday	5
Saturday	6
Sunday	7

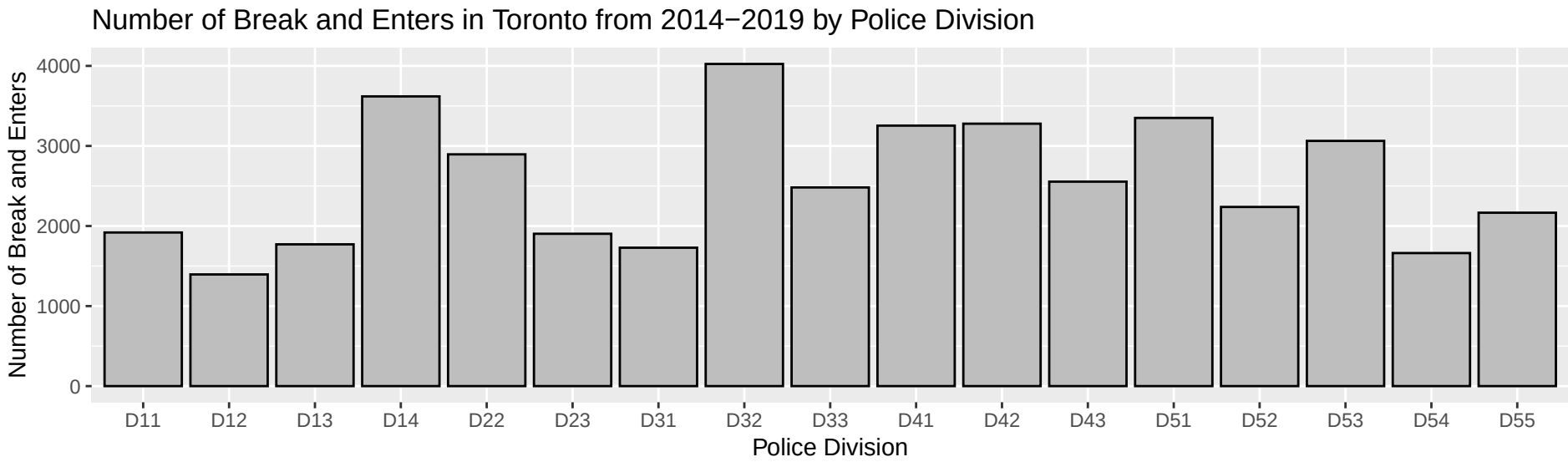
Statistical Methods - Part 1

- To investigate if certain police divisions in Toronto had a larger number/proportion of Break and Enters, we created a bar graph showing the number of break and enters per division.
- We also created a summary table displaying the minimum, maximum, median, mean, and standard deviation of the break and enters across all divisions.
- To investigate if certain premise types and hours of the day had a larger number/proportion of Break and Enters, we again created bar plots, histograms and summary tables.
- Based on our results, we saw that there was a significantly larger number of break and enters at midnight. Hence, to determine if there was really a difference between the amount of Break and Enters at midnight compared to at other times, we conducted a randomization/hypothesis test between the two groups

Statistical Methods - Part 2

- To determine if there is a relationship between the day of the week a Break and Enter occurred and the hour of the day a Break and Enter occurred, we created a scatterplot and tried to fit a linear regression model using the day of the week (*day_number*) as the explanatory variable and the hour of the day (occurrencehour) as the numerical response.
- Finally, to determine if certain premise types have more Break and Enters on certain hours of the day/days of the week, we created a classification tree using the hour of the day (*occurrencehour*) and day of the week (*occurrencedayofweek*) as predictors and the type of premise as the response.

Results - Part 1: B&Es and Police Divisions

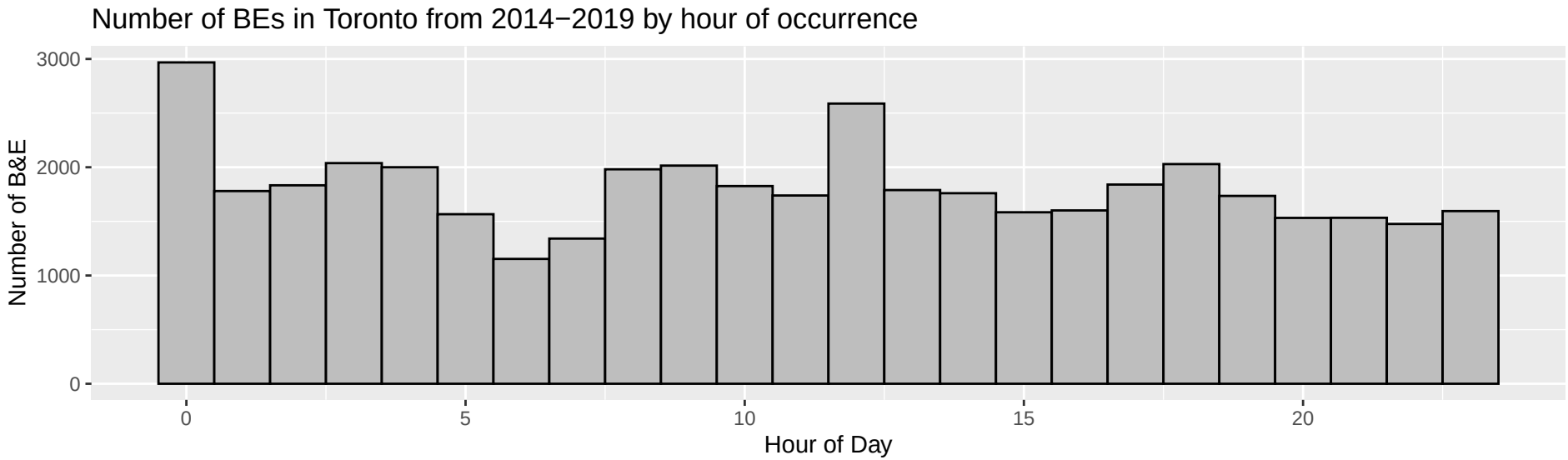



```
## # A tibble: 1 x 5
##     Min   Max   Med  Mean Standard_Deviation
##   <int> <int> <int> <dbl>              <dbl>
## 1  1395  4024  2482 2547.               784.
```

What does the graph and summary table tell us?

For clarification purposes, the summary table shows the mean, median, minimum, maximum and standard deviation of the number of Break and Enters aross all divisions. We can see from the bar graph and summary table that the number of break and enters per division varies widely, from a minimum of 1395 break and enters to a maximum of 4024. In addition, the median and mean of the data, 2482 and 2547.176 respectively, differ slightly, with the mean being 65.176 break and enters greater than the median. We can now look at a map of Toronto to find out which part of the city contains the police divisions with the most number of Break and Enters - this will be presented in the conclusions.

Results - Part 2 - Break & Enters and Days of the Week

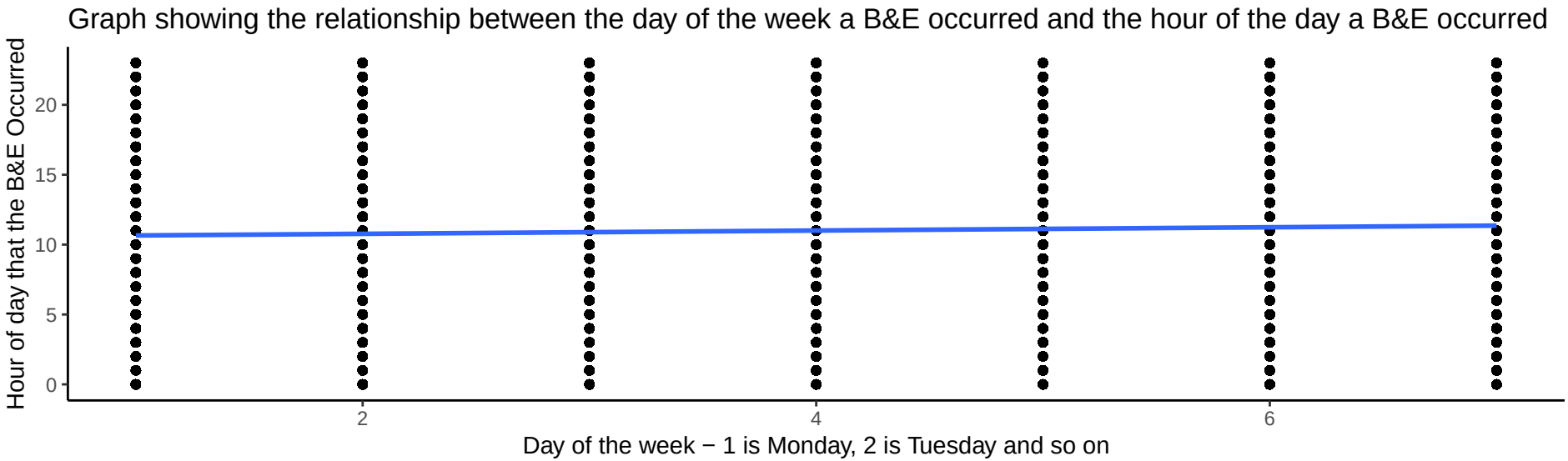



```
## # A tibble: 1 x 3
##   mean_hour median_hour sd_hour
##       <dbl>       <dbl>   <dbl>
## 1      11.0          11    6.97
```

What does the graph and summary table tell us?

Possible outliers at 6 am in the morning. slightly right-skewed (due to mode at midnight), sort of bimodal with peaks at 12 am (midnight) and 12 pm (noon). The center is the median hour of 11 am and the spread is a standard deviation of 6.966893, which means the values are quite spread out from the mean. Notice that the number of B&Es is significantly higher at 12am. We conducted a hypothesis test between the two groups hour of occurrence at 12am and hour of occurrence not at 12am to determine if there was actually a difference between their number of B&Es. Our null hypothesis was there is no difference between the number of B&Es in the two groups described above and our alternate hypothesis is there is a difference between in the number of B&Es between the two groups described above. After conducting the test, we deduced the following:
If there is no difference between the proportion of occurrences at midnight and the proportion of occurrences at every other hour of the day, the chance of observing a difference in proportions as large or larger than 0.02804369 (test statistic) is VERY close to 0 (p-value). In other words, these data provide VERY strong evidence against our null hypothesis that the proportion of occurrences is not different at midnight and every other hour of the day.

Results - Part 4: Relationship between day of week and hour of occurrence




```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 10.5365764 0.07517159 140.167003 0.000000e+00
## day_number   0.1173345 0.01718776   6.826629 8.807613e-12
```

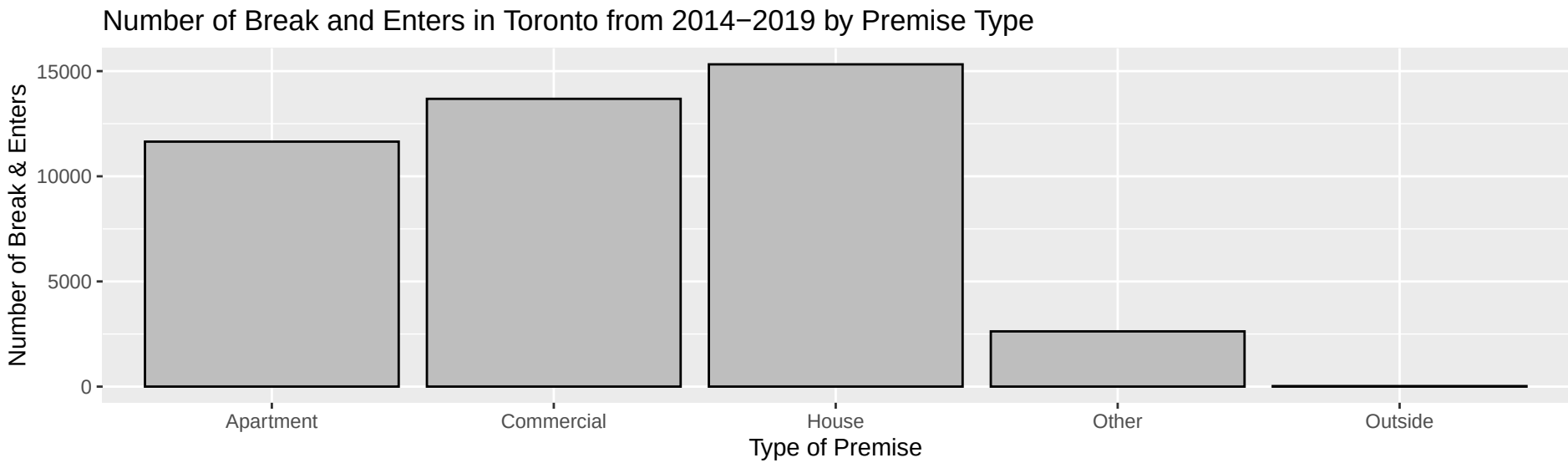
What does the plot and linear regression model tell us?

The plot tells us that there is no association, or at best, an extremely weak association between the day of the week and the hour of occurrence of the Break and Enter. This is backed up by our value for the coefficient of correlation which is:

$$r = 0.03278905$$

However, one very suprising result we saw was that even though there is seemingly no relationship between the day of the week and the hour of occurrence based on the graph, we obtained a p-value of $8.81 \cdot 10^{-12}$ (this can be seen in the table above). This p-value, since it is extremely small, tells us to reject the notion that there is no linear relationship between the day of the week and the hour of occurrence, even though there seems to be no relationship at all between the variables.

Results - Part 5: B&Es and Premise Types



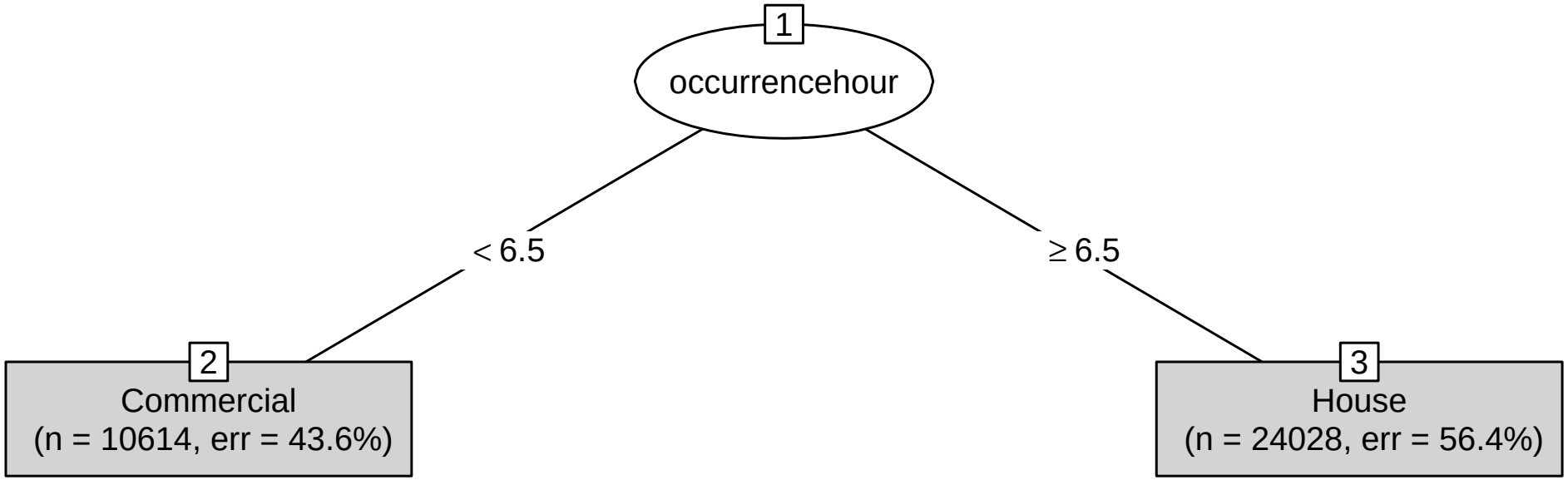

```
## # A tibble: 5 x 3
##   premisetype Number_of_BEs Proportion_of_BEs
##   <chr>                <int>             <dbl>
## 1 Apartment            11648            0.269
## 2 Commercial           13682            0.316
## 3 House                15322            0.354
## 4 Other                 2626            0.0606
## 5 Outside                 24            0.000554
```

What does the graph and summary table tells us?

We can see that houses had the highest number and proportion of Break and Enters in Toronto from 2014-2019 with 15322 Break and Enters, which made up around 35% of the total Break and Enters. This was closely followed by Commerical Premises (businesses), which has 13682 or 31.6% of Break and Enters. Next was apartments, with 11648 or 26.9% of Break and Enters. These three premises made up the vast majority of the Break and Enters, as other/outside premises only had a combined total of 2650 Break and Enters, which is approximately 6% of the total Break and Enters.

Results - Part 6 - Predicting the type of premise using occurrence hour and day of week

First, we randomly split up 80% of the data into "training" data and 20% into "testing" data. Then, using the training data, we created the classification tree which used the hour of the day (*occurrencehour*) and day of the week (*occurrencedayofweek*) as predictors and the type of premise as the response.



After creating the tree, we tested, using the testing data, how accurately our tree could predict the type of premise based on the hour of occurrence. This gave us the following confusion matrix:

	Apartment	Commercial	House	Other	Outside
Predict Commercial	529	1552	466	175	2
Predict House	1792	1262	2557	324	1

What does our classification tree tell us, and is it a "good" tree?

Based on our classification tree, we can see that a B&E that occurs before 6.5 hours (before 6:30am) is classified as a commercial premise and a B&E that occurs after 6.5 hours is classified as a house. However, our classification tree is not a good/accurate predictor of the premise type. Based on our confusion matrix, we compute the accuracy of our classification tree as 47.4%. Hence, most of the time, our classification tree will be wrong! Therefore, we cannot really say if certain hours of the day/days of the week are linked with Break and Enters at certain premise types

Conclusions and Recommendations

Based on our results, our main conclusions and recommendations are as follows:

- After considering a map of the police divisions in Toronto, we notice that many of the divisions with a large number of B&Es are centered around the southern waterfront area, which lies in the Central Field Command of the Toronto Police Service, and the northern area of the city around Steeles Ave, in the Area Field Command. Thus, we recommend that Toronto Police Service should allocate more officers to the aforementioned areas to shorten the response times to reach attempted break and enters. Moreover, residents living in these areas and businesses operating in these areas should also be told to be extra cautious.

Conclusions and Recommendations - Part 2

- A large number of Break and Enters occur at midnight and late during the night, as well as in houses, apartments, and commercial premises, so more officers should be allocated in areas with these premises and during the times mentioned above.
- No relationship was found between hour of day, day of week and premise type