Toronto Police Break and Enter Data Analysis

Studying Crime Trends in Neighborhoods with high B&E rates

Adele, Wayne, Alexandra, Andrei & Cindy(Shih-Ting),
TUT0201, Group 4

March 30, 2020

Introduction

The offence of break and enter encompasses situations where individuals tresspass or attempt to trespass on private enclosed property.

Though break and enters are not the most violent crimes, the preventative measures for such offenses can be exhaustive and costly–alarm systems, surveillance cameras, security teams, etc. In order to ensure adequate protection of vulnerable spaces and appropriate usage of resources, it is important to understand the spatial and temporal patterns of break and enters.

The data used to analyze these patterns comes from the Toronto Police Service, and includes reported break and enters from 2014-2019.

Objectives

Our main question which we seek to answer is: **"In neighborhoods with high B&E rates, what are the commonalities between the crimes?"** We then divided this question into a series of smaller subquestions, focused on specific variables:

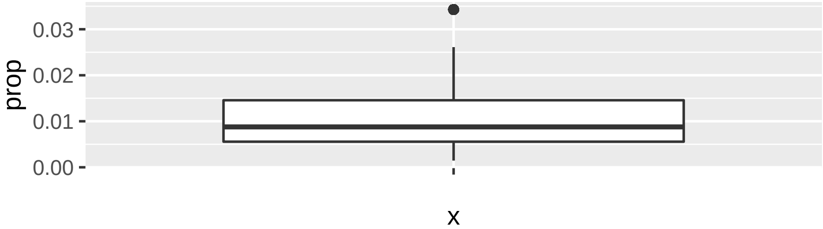
- Are particular premise types more susceptible to B&E's in high-risk compared to low-risk neighborhoods?
- Is there a particular day of the week or hour of the day that inhabitants should be more wary of B&E's?
- Is there a certain kind of B&E that occurs more often in these high-rate neighborhoods?

Using the data given to us by the TPS, we want to understand commonalities between offenses in the neighborhoods that experience the highest proportion of break-and-enters. The purpose of doing so is to help the TPS understand the characteristics of B&E offences in vulnerable neighborhoods in order to allocate officers and other resources efficiently to reduce crime across the city. Once these strategies are implemented, it is likely that the city will change and different neighborhoods will become hotbeds for B&E's. Hopefully, once a strategy is developed for high-risk neighborhoods, it can be transplanted and altered slightly to work elsewhere.

We hypothesize that premise type, as well as temporal trends such as time of day/day of the week, influences vulnerability in high-risk neighborhoods. Furthermore, we expect that these neighborhoods might experience a higher density of B&E with intent.

Data Summary

We first created a new variable for each offence based off of the neighborhood it occurred in, called *prop*, which represented the proportion of all break and enters that occurred in that neighborhood. We then created a boxplot and a corresponding summary table to help us understand how the proportions were distributed, where *prop* represents, for each observation, the total proportion of reported break-ins that occurred in that observation's neighborhood.



```
##            Q1         med         Q3        max
## 1 0.005565563 0.008775576 0.01457208 0.03427093
```

We were interested in the extremes of this data–in other words, neighborhoods that were high-risk and neighborhoods that were low-risk. This allows us look at data that is "extreme"–those with proportions <= the first quartile, and those >= the third (aka, the bottom 25% and the top 25% neighborhoods).

Statistical Methods

Subquestion 1:

- Hypothesis Test for Difference in Proportions of commercial break-ins between high/low risk neighborhoods: randomly shuffled values in original data without replacement, created a distribution of the differences under the assumption that there is no difference in proportion of commercial breakins in high- and low-risk neighborhoods.
- Linear Regression using whether a premise was residential as a predictor

Subquestion 2:

- Hypothesis Test for Difference in Proportions of Occurrence Hour and Occurrence Month

Subquestion 3:

- Bar Graph: By examining the data specifically for high-risk neighborhoods, we identify which offence type appears the most and hence occurs the most often.
- Hypothesis Test for Difference in Proportions: We randomly shuffled the values in the original data without replacement and got their distribution to see if the observed difference in proportion of B&E offences is “normal” under the assumption that there’s no proportion difference between high-risk and low-risk neighborhoods.

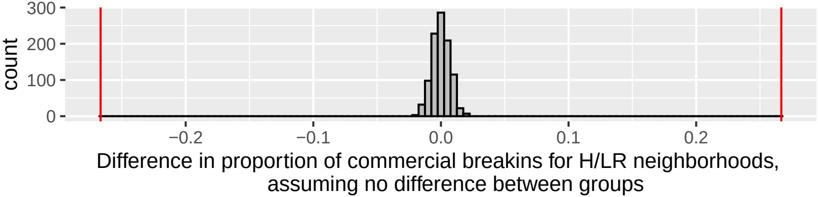
Results–Subquestion 1: Hypothesis Test

Upon initial observation, we noticed that the proportion of commercial break-ins in high-risk neighborhoods appeared much larger than commercial break-ins in low-risk neighborhoods, with an observed difference of 0.267.

```
## # A tibble: 2 x 4
##   level     n n_commercial prop_commercial
##   <chr> <int>        <int>           <dbl>
## 1 high  10394         5258           0.506
## 2 low   11418         2731           0.239
```

```
## [1] -0.266685
```

This was tested using a difference of proportions hypothesis test, using the following hypotheses: $H_0 : p_H = p_L$ & $H_A : p_H \neq p_L$, where p_H, p_L are the proportion of commercial break-ins that happened in high- and low-risk neighborhoods, respectively.



Given the low p value of $p < 0.0001$, we can conclude that there is strong evidence against the null that states that $p_H = p_L$. Therefore, we have reason to believe that the proportion of commercial breakins in these high-rate neighborhoods is different than low-rate neighborhoods.

Results–Subquestion 1: Linear Regression

This same question can be analyzed through a different lense: linear regression. For this method, a new variable **residential** was created that took the value *"Yes"* if the premise was a house or apartment, and *"No"* if it was anything else.

The equation we are testing is $prop_i = \beta_0 + \beta_1 I(premise\ is\ residential) + \epsilon_i$. In other words, we are seeing if whether a property is residential makes it more vulnerable to break-ins (**vulnerability** approximated by **prop**) We are interested in testing $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

```
##                    Estimate   Std. Error    t value    Pr(>|t|)
## (Intercept)     0.0216597997 7.996666e-05 270.860367 0.000000000
## residentialYes -0.0003972273 1.231303e-04  -3.226072 0.001258903
```

Based on the fitted regression model, the p-value corresponding to this hypothesis test is very small (the estimated p-value from R is 0.001258903), so we have very strong evidence to reject the null hypothesis and conclude that there is an association between a property being residential and proportion of neighborhood breakins, and the negative coefficient for β_1 suggests that if a property **is** residential, the **vulnerability** of that property will be lower than if it isn't residential.

Results–Subquestion 1: Linear Regression RMSE

```
## # A tibble: 1 x 3
##   RMSE_testdata RMSE_traindata ratio_of_RMSEs
##           <dbl>          <dbl>          <dbl>
## 1       0.00624        0.00619          0.991
```

The Root Mean Squared Error measures prediction error for predictions from a linear regression model. Splitting the data up into testing (20%) and training (80%), we see that the RSME for both is around 0.006. In other words, the prediction error for this model is 0.006 (referencing proportion estimates).

Subquestion 2: Yearly and Monthly

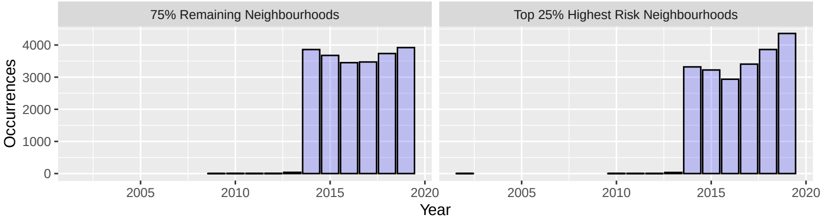
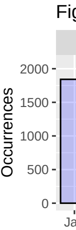
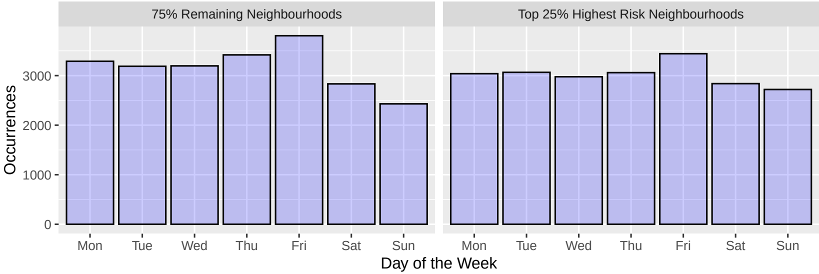


Figure 2a − B&Es by Year

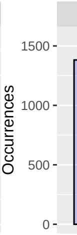


Subquestion 2: Daily and Hourly

Figure 2c − B&Es by Day of Week



Fig



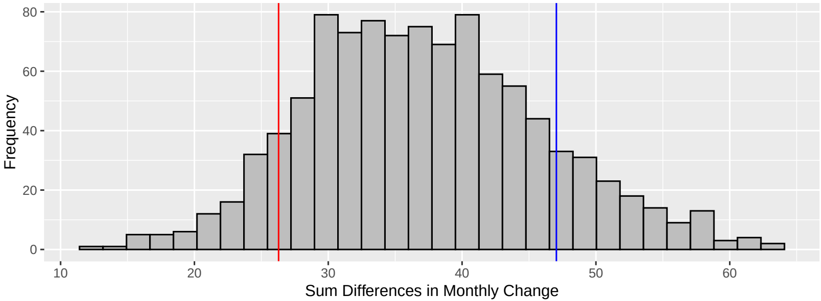
Are these trends any different?

Do the highest-risk neighbourhoods have significantly different time trends?

H_0: There is no difference in the patterns describing when B&Es occur in our most vulnerable neighbourhoods.
H_a: The patterns describing when B&Es occur are different in our most vulnerable neighbourhoods.

As our test stat, we calculate the percentage change in B&Es by time period for both categories of neighbourhoods, and then find the sum of the differences in percentage change between the two. Then we can simulate to see if these values are unusual.

Occurrences by Month



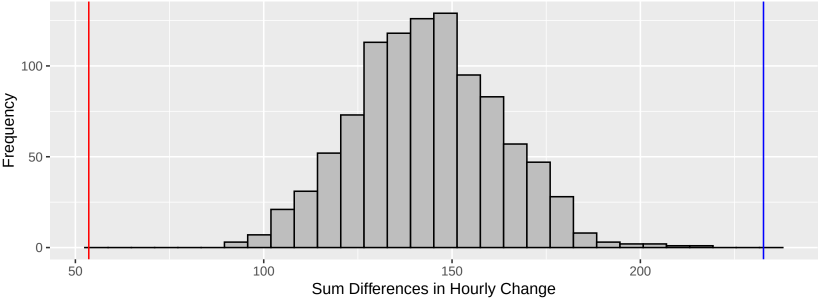
Looking back at Figure 2b, we can visually see that the trend between our two groups are very similar. Are those slight deviations in trend significant?

We get our test stat of 47.04 units of discreptency by finding the month-on-month percentage change for both groups of neighbourhoods, and getting the sum of their differences.

With this test stat, we get a p-value of 0.236, suggesting no evidence against the null hypothesis. This means 23.6% of simulated results showed a discreptency more extreme, which suggests that our test case is probably not unusual.

In other words, **monthly trends likely remain consistent in the most vulnerable neighbourhoods.**

Occurrences by Hour of Day



Similarly we run the same test on B&Es by hour of day (Figure 2d). Are the slight variation in trend between our two groups significant?

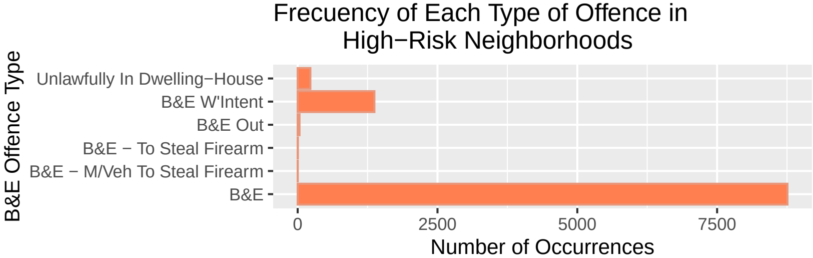
We get our test stat of 232.69 units of discreptency by finding the hour-on-hour percentage change for both groups of neighbourhoods, and getting the sum of their differences.

Running a hypothesis test, we get a p-value of 0, suggesting strong evidence against the null hypothesis. We can interpret this by saying that there are almost never more extreme cases, and thus this result must be significantly abnormal.

In other words, **hourly trends in the most vulnerable neighbourhoods are very likely different in a significant way.**

Results–Subquestion 3:

From the bar graph below, we can clearly identify that **B&E offences** occur the **most often** in high-risk neighborhoods.

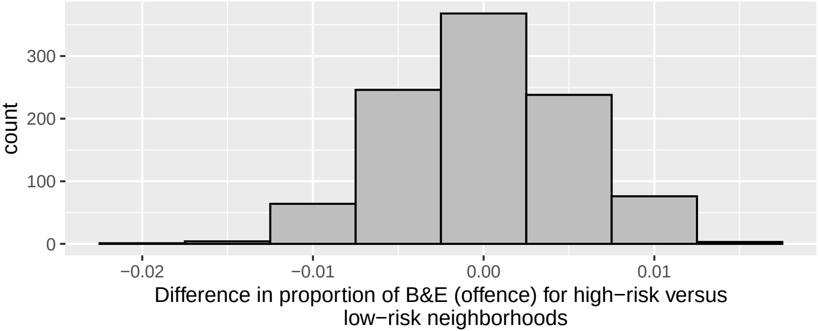


Results–Subquestion 3: Hypothesis Test

In order to determine if this specific offence type occurs more often in high-risk than low-risk neighborhoods, we conduct a *hypothesis test*:
H_0: $P_{\text{diff}} = 0$
H_a: $P_{\text{diff}} \neq 0$
where P_{diff} represents the difference between the proportion of B&E offences in high-risk neighborhoods and that of low-risk neighborhoods.

Results–Subquestion 3: Hypothesis Test

Given this sampling distribution, we compare them with the test statistic, which is the proportion of B&E offences over the total number of offences as calculated from our sample data. We want to see if the test statistic is unusual under the null.

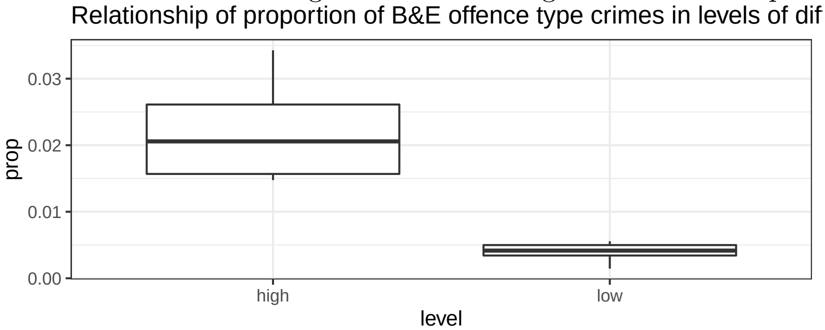


Thus, our observed value used for the test is very unusual compared to the simulated ones.

- Since the p-value is less than 0.0001, it means that there is a **low possibility** that there is data **as unusual as our observed value (test statistic)** under the null hypothesis.
- Therefore, we have **strong** evidence against the null that there is no difference in the proportion of B&E offences between high-risk and low-risk neighborhoods.

Results–Subquestion 3 - Boxplot:

Next we will show how the proportion of type Break and Enters differs between high and low rate neighbourhoods on a plot:



It is evident that the proportions of break and enters in high-rate neighbourhoods have a much larger range than in low-rate neighbourhoods of this particular type of offence.

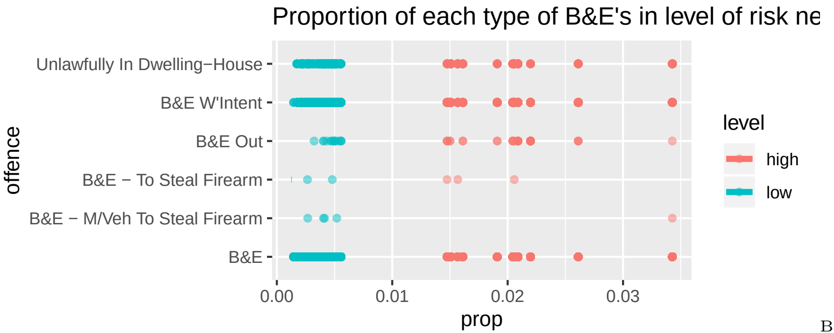
Results–Subquestion 3 multilinear regression results:

Further, we will use another method - a multilinear regresssion model to predict the proportion of Break and Enters using the offence types and level as predictors. The regression equation will be: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i}$

where our response variable is y for the ith observation, x_{1i} and x_{2i} are the independent variables, βB_0 is the intercept parameter and βB_1 is the slope parameter. We are interested to test the H_0: $P_{\text{diff}} = 0$ H_a: P_{diff} != 0 where H_0 represents the null hypothesis, H_a the alternative hypothesis, and P_{diff} is difference between the proportion of a certain offence type in B&E rated neighborhoods.

```
##                                          Estimate   Std. Error      t value
## (Intercept)                          2.148629e-02 4.443316e-05  483.5642741
## offenceB&E - M/Veh To Steal Firearm  2.502184e-03 1.943369e-03    1.2875492
## offenceB&E - To Steal Firearm       -2.829543e-03 1.943312e-03   -1.4560414
## offenceB&E Out                      -2.860470e-04 6.346983e-04   -0.4506818
## offenceB&E W'Intent                  6.586622e-05 8.692308e-05    0.7577530
## offenceUnlawfully In Dwelling-House -5.688430e-05 1.774210e-04   -0.3206176
## levellow                            -1.740534e-02 5.895986e-05 -295.2065785
##                                      Pr(>|t|)
## (Intercept)                         0.0000000
## offenceB&E - M/Veh To Steal Firearm 0.1979166
## offenceB&E - To Steal Firearm       0.1453956
## offenceB&E Out                      0.6522234
## offenceB&E W'Intent                 0.4486071
## offenceUnlawfully In Dwelling-House 0.7485033
## levellow                            0.0000000
```

Results–Subquestion 3 multilinear regression visual:



Based on the fitted regression model, the corresponding p-value test is very small (P = 0.0), so we have very strong evidence to reject the null hypothesis that there is no difference between proportions of B&E offence types. So we can conclude that there is a difference of proportion between the level and B&E offence types.

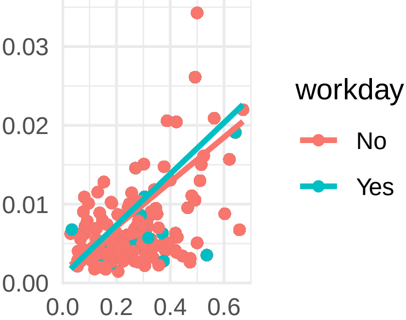
Putting It All Together

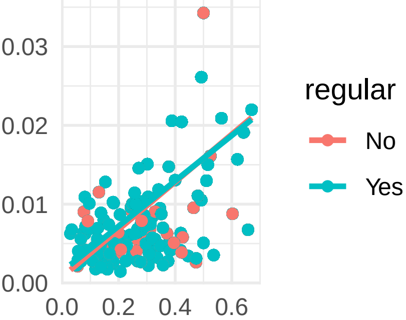
What does all this information mean? What is the intersection between time, premise type, and type of offence? This can be examined through multiple linear regression. Two regression equations were used. For each, new variables were created. In both, a variable **prop_premisetype** was created that, for each observation, corresponded to the proportion of commercial offences over total neighborhood offences. In the first model, a variable **regular** took the value "Yes" if offence was B&E, and "No" otherwise. In the second model, a variable **workday** took the value "Yes" if the offence occurred during the typical workday, adapted for commute times (7am-7pm), and "No" otherwise.

```
##                                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)                 0.0014366938 0.0001708945  8.406907 4.622437e-17
## prop_premisetype            0.0284231557 0.0003970780 71.580781 0.000000e+00
## workdayYes                 -0.0006088492 0.0004914275 -1.238940 2.153889e-01
## prop_premisetype:workdayYes  0.0040950614 0.0011251814  3.639468 2.742086e-04

## [1] 0.3095764

##                                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)                 0.0007565288 0.0004116634  1.837736  6.612293e-02
## prop_premisetype            0.0301807358 0.0009623229 31.362379 1.460799e-208
## regularYes                  0.0007004001 0.0004470409  1.566747  1.171969e-01
## prop_premisetype:regularYes -0.0014244675 0.0010435058 -1.365079  1.722508e-01

## [1] 0.3071695
```

Putting It All Together: Graphs

Each of these graphs plots *Local Commercial Rate* against *Overall Neighborhood Occurrence Rate*.





For each graph, the correlation coefficient, **r**, is about 0.31 and 0.307, respectively, indicating a weak, positive correlation. In other words, this means that roughly 31% and 30.7% of the variation in the **Overall Neighborhood Occurrence Rate** is explained by our linear regression. The for the first model, it appears as though all terms are significant except for the binary variable **workday**, yet the interaction of **workday** with **prop__premisetype** is significant. For the second model, neither **regular** nor the interaction term are significant, potentially related to this model's lower **r** value.

Conclusion

Give your main conclusions here. Follow the order of questions you presented.

Here you can also mention any additional considerations, concerns, or issues you might have. For example, if the results were unexpected, you can discuss this and perhaps offer possible explanations.