Toronto Police Break and Enter Data Analysis

The subtitle of my project

Adele, Wayne, Alexandra, Andrei & Cindy, 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?

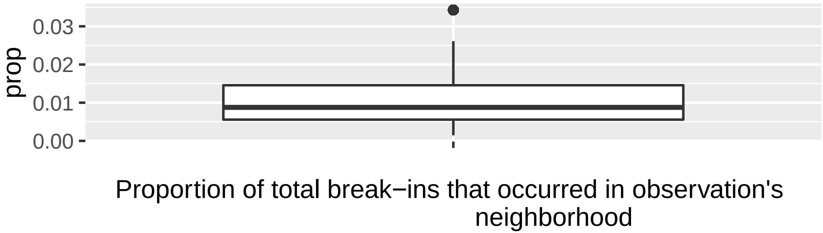
Objectives Continued

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.



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

Data Summary Continued

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

Describe here what you have done to the data without presenting any results (output). If you want to indicate variables by symbols or variable names, define them here. Subquestion 1: Hypothesis Test for Difference in Proportions and Linear Regression

Results–Subquestion 1:

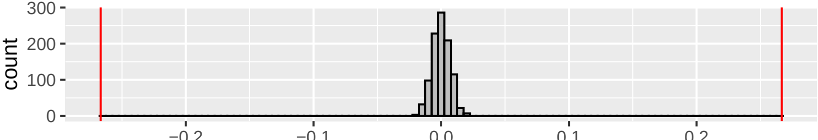
Upon initial observation, we noticed that the proportion of commercial break-ins in high-risk neighborhoods appeared much larger than commercial break-ins in low-risk neighborhoods. This was tested using a difference of proportions hypothesis test, using the following hypotheses:

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. 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.

```
## # 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
```



Results–Subquestion 1 Continued:

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.

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 Continued:

```
## # 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. Therefore, on average, predicted **vulnerabilities** (or proportions) are about 0.006 away from the true value in this data set restricted to high-risk neighborhoods.

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.

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.

Acknowledgements (optional)

If you received any help from someone other than your team members you can acknowledge them. For example:
The authors would like to thank "TA name" for helpful suggestions and comments that improved the presentation of this poster.