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?

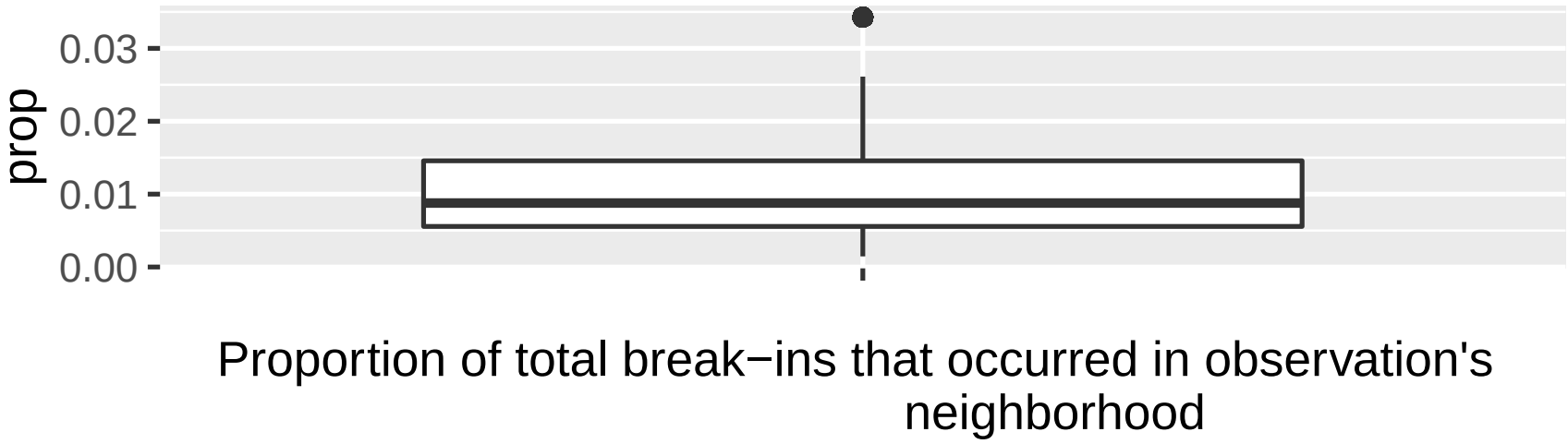
Objectives Continued

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
```

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 and Linear Regression

Subquestion 2:
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.
-

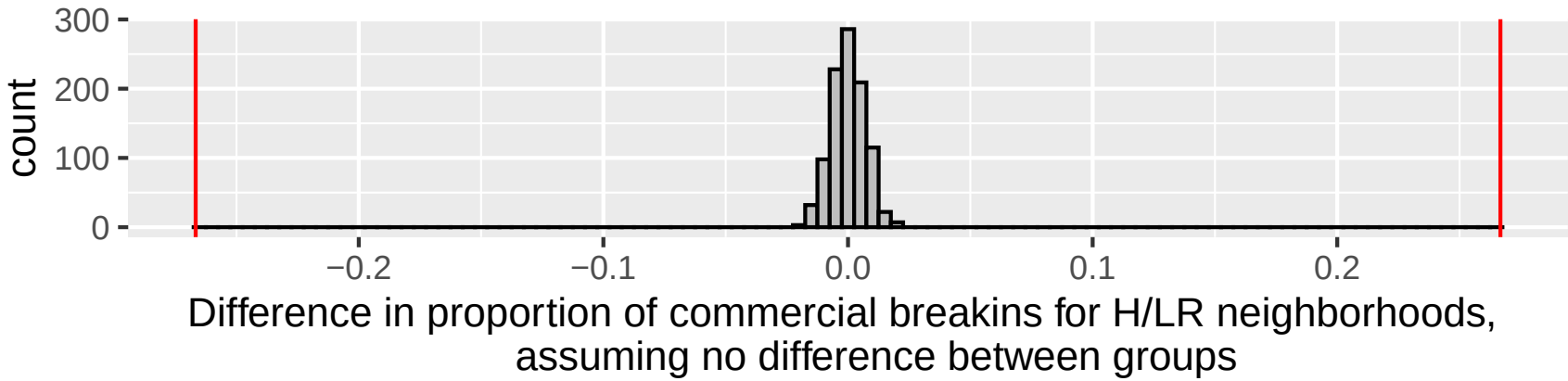
Results–Subquestion 1: Hypothesis Test

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. 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
```



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.

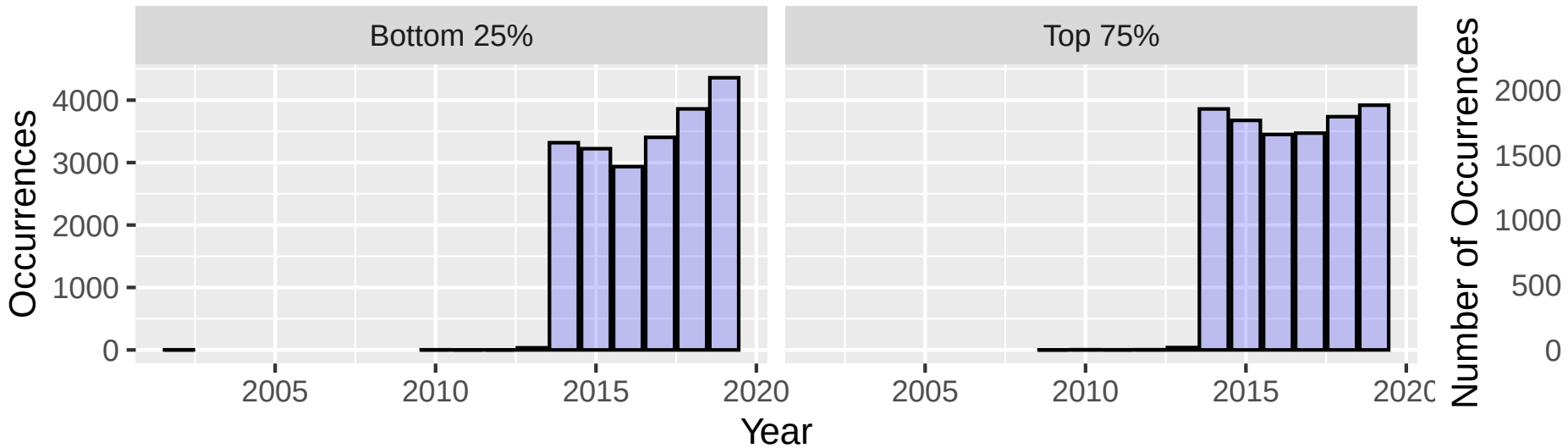
Results–Subquestion 1: Linear Regression RMSE

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

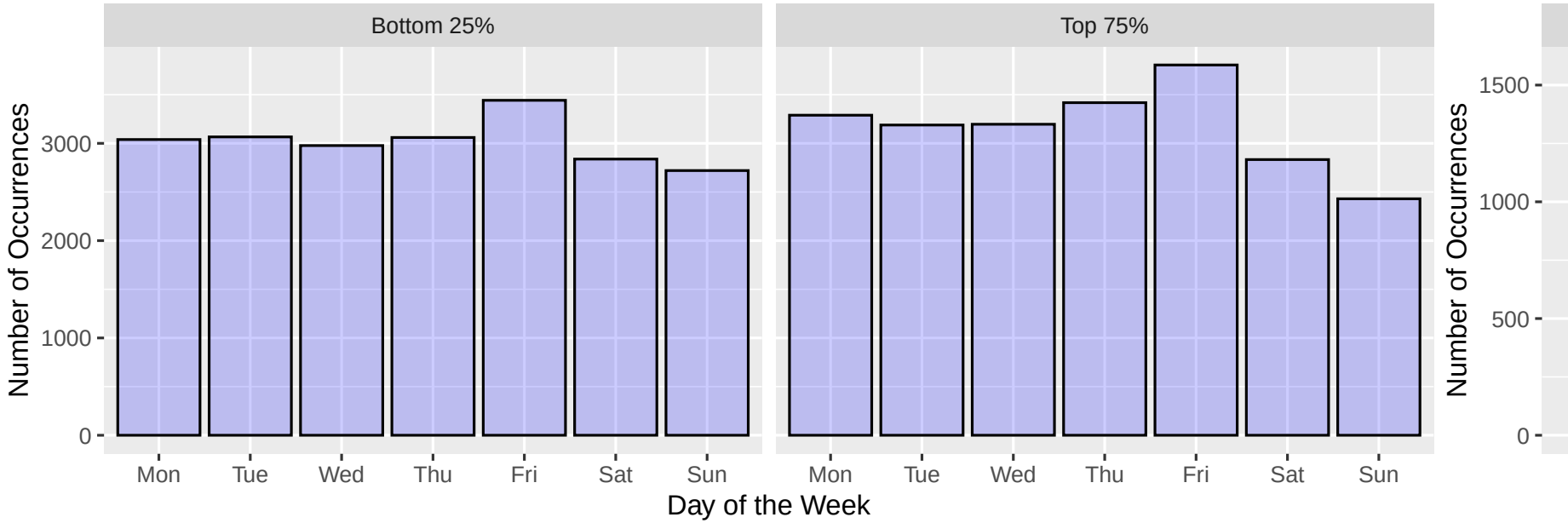
The Root Mean Squared Error measures prediction error for predictions from a linear regression model. Splitting the data up into testing (20%) and training (80%), we see that the RSME for both is around 0.006. Therefore, on average, predicted **vulnerabilities** (or proportions) are about 0.006 away 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.

Results–Subquestion 2: Frequency of Occurrence Year



Results–Subquestion 2:



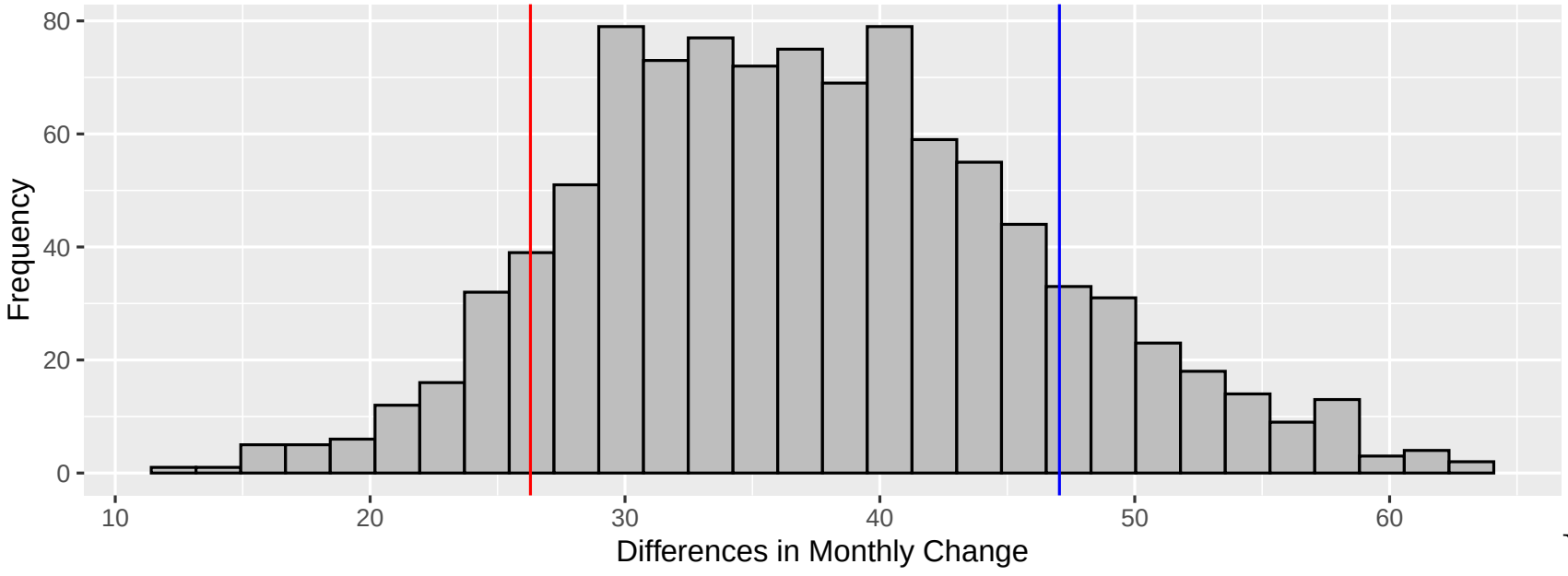
Subquestion 2–Are these trends any different?

Are discreptencies in between trends in our two groups of neighbourhoods significant? Do our two groups have the same trends, or do the neighbourhoods with more frequent B&Es show different time trends?

Hypothesis Test:
H_0: There is no difference in the patterns describing when B&Es occur in our worst-affected neighbourhoods.
H_a: The patterns describing when B&Es occur are different in our worst-affected neighbourhoods.

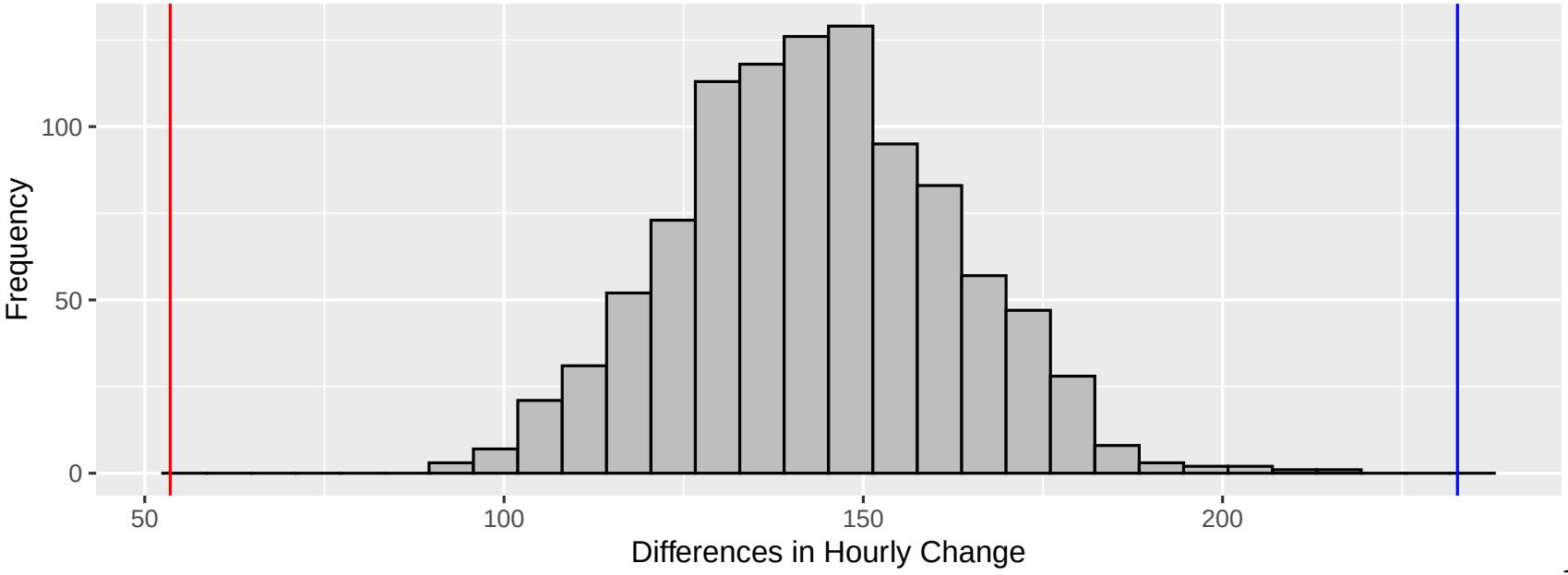
For the sake of simplicity, we will only analyze occurrence month and hour of day.

Subquestion 2–Occurrences by Month



With our test stat of 47.04 units of discreptency, we get a p-value of 0.236, suggesting no evidence against the null hypothesis. In other words, **monthly trends likely remain consistent in the most vulnerable neighbourhoods.**

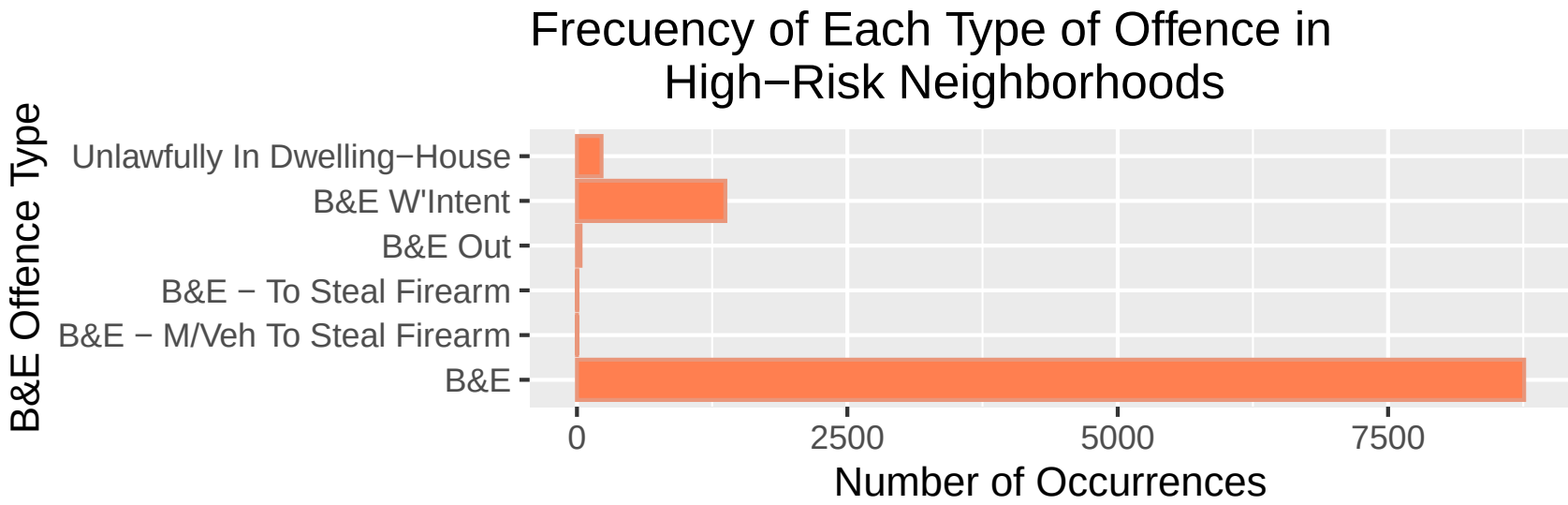
Subquestion 2–Occurrences by Hour of Day



Again with our test stat of 232.69 units of discreptency, we get a p-value of 0, suggesting strong evidence against the null hypothesis. 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.

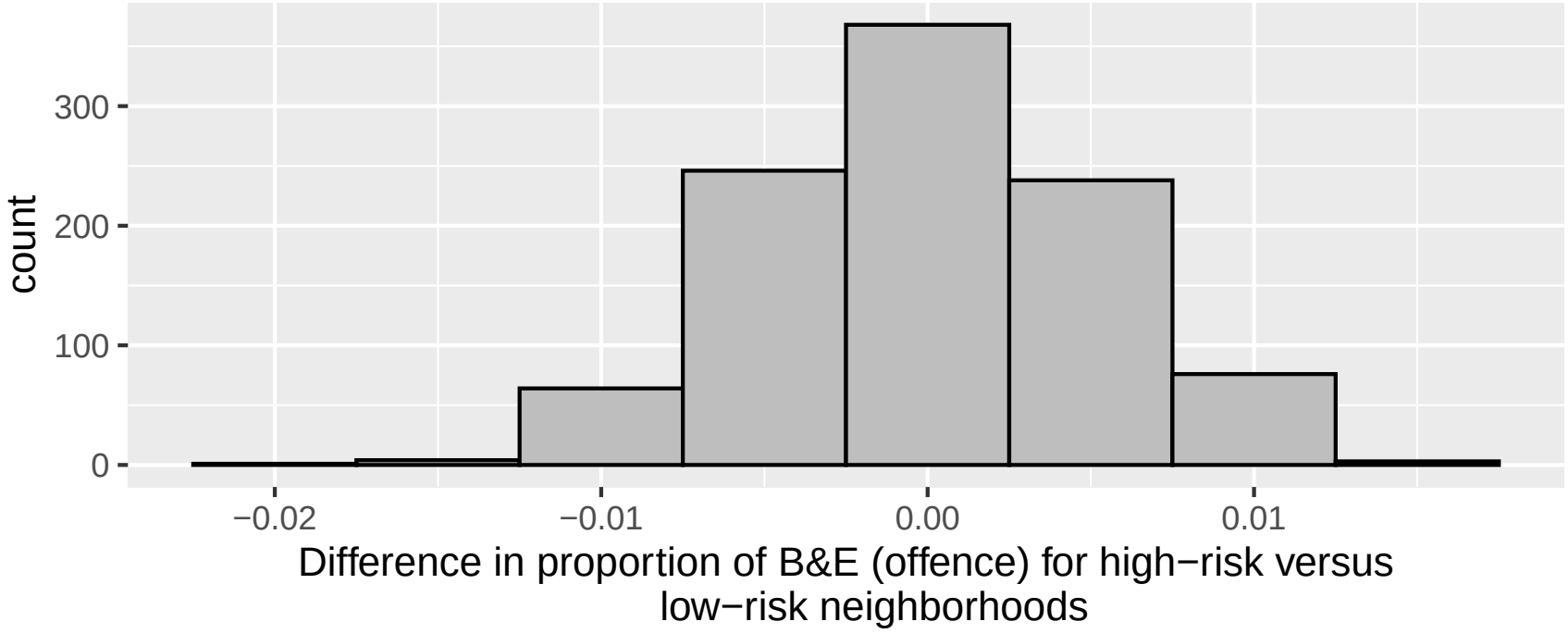


Results–Subquestion 3: Hypothesis Test

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.

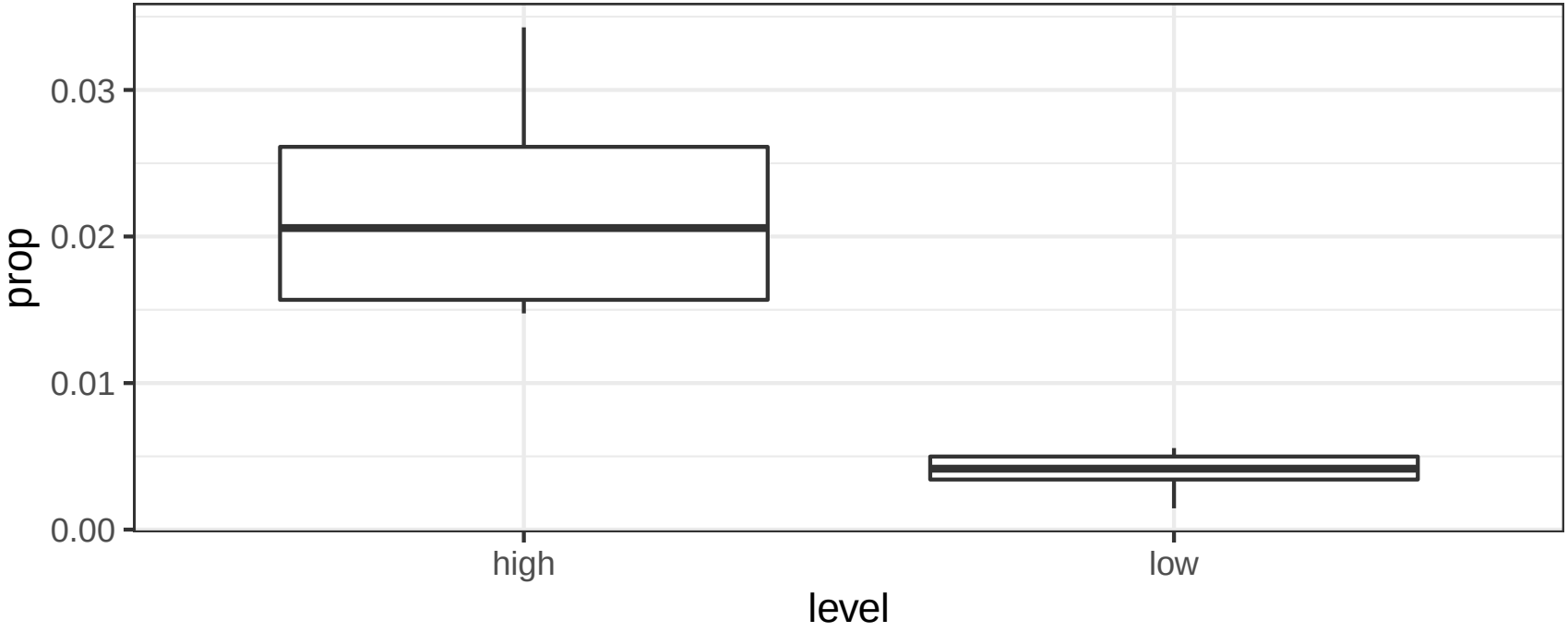


Thus, our observed value used for the test is very unusual compared to the simulated ones.

- 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 Continued:

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.

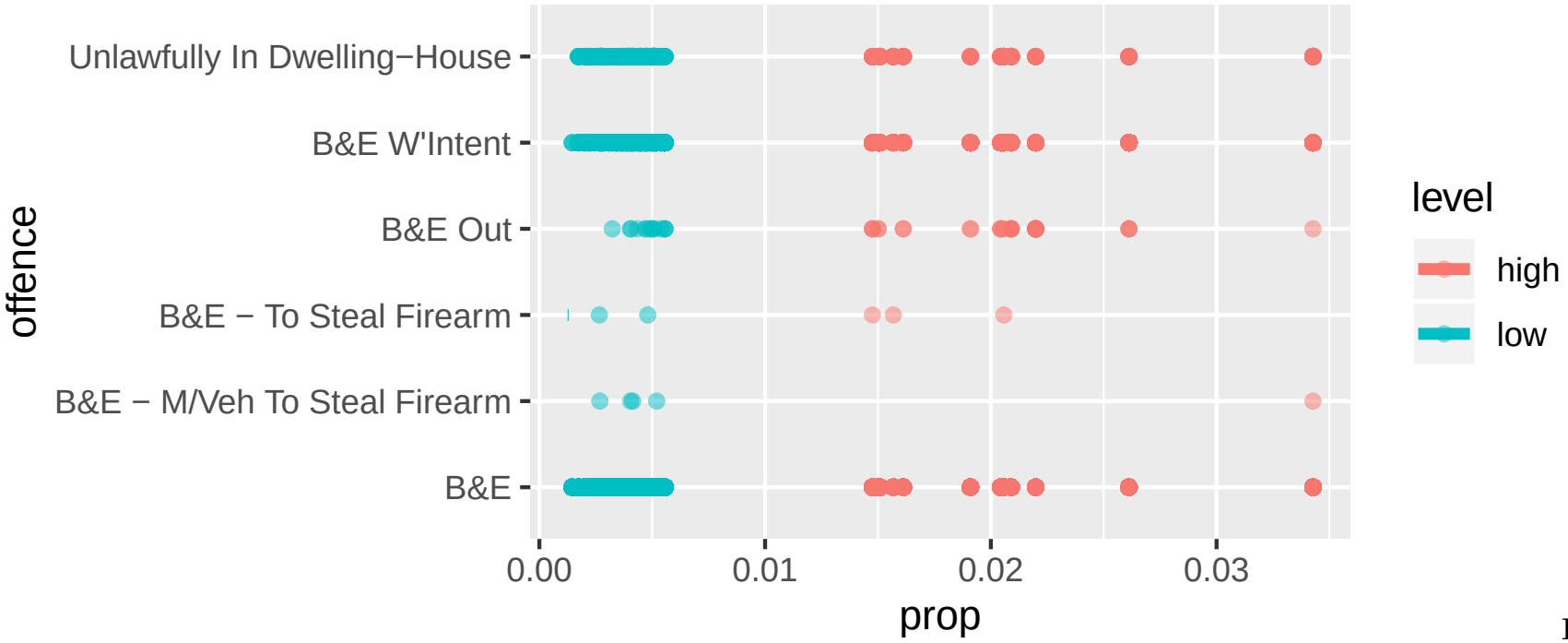
Results–Subquestion 3 Continued:

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 Continued:



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.

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.