Criminal Minds

Steve White, Monica Wisdom, Minh Nguyen, Hank Flury

Introduction

As students at the University of Washington, we have access to safety alerts and notifications of criminal incidents through the University of Washington Police Department. However, the frequency of these notifications can make it easy to fall under the perception that there is a disproportionate amount of crime happening near us. Our data analysis seeks to determine the validity of this notion and investigate the landscape of crime in the University District in relation to Seattle proper.

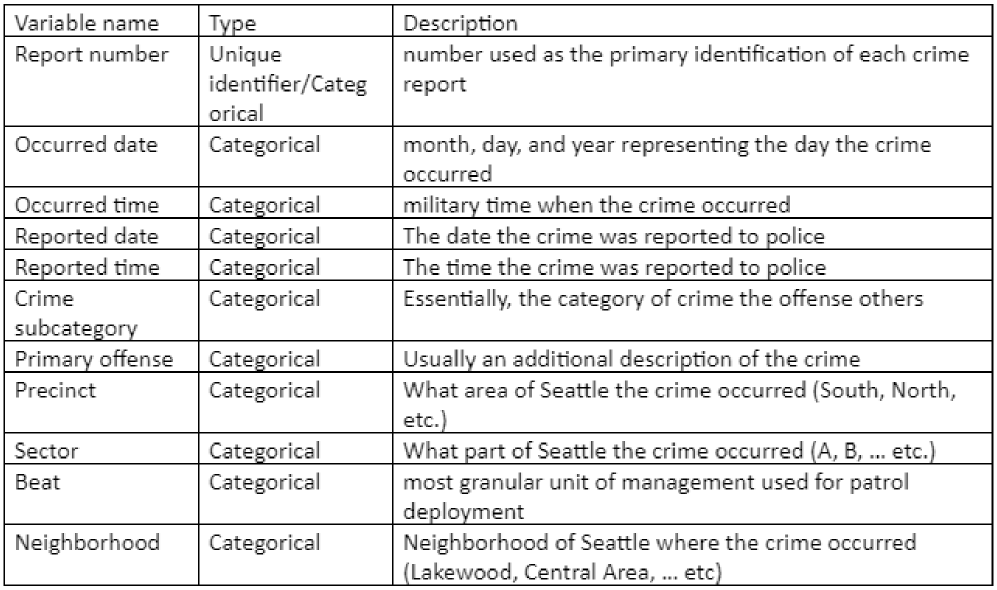
For the purpose of this project, we used two questions to guide our data analysis. The University District is the focus of our questions because we all spend a great deal of time here studying, working, and living. First, does the proportions of different crimes in the University District differ from the whole of Seattle? Second, in the University District, does the average number of days between the date a crime occurred and the date that crime was reported differ from the whole of Seattle, particularly in the cases of rape, other sexual offenses, and nonviolent family offenses?

These questions are meant to put life into perspective and reveal information about the neighborhood many of us reside in. Often, it's easier to rely on internal realizations of life and what we occur first hand. By adopting a more external perspective, we take into account the impact of society and groups of individuals to influence how we view the world.

Another reason, we hoped if these questions will allow us to see if it's better to adopt a reactive or proactive approach to policing. Then, this will provide a reason to pursue the implementation of more proactive policies surrounding rape, other sexual offenses, and nonviolent family offenses.

Methods

In order to better understand our methods, it is important one is familiar with the variables we worked with. Here is a chart detailing such variables:



Variable name	Type	Description
Report number	Unique identifier/Categorical	number used as the primary identification of each crime report
Occurred date	Categorical	month, day, and year representing the day the crime occurred
Occurred time	Categorical	military time when the crime occurred
Reported date	Categorical	The date the crime was reported to police
Reported time	Categorical	The time the crime was reported to police
Crime subcategory	Categorical	Essentially, the category of crime the offense others
Primary offense	Categorical	Usually an additional description of the crime
Precinct	Categorical	What area of Seattle the crime occurred (South, North, etc.)
Sector	Categorical	What part of Seattle the crime occurred (A, B, … etc.)
Beat	Categorical	most granular unit of management used for patrol deployment
Neighborhood	Categorical	Neighborhood of Seattle where the crime occurred (Lakewood, Central Area, … etc)

The main resampling method we utilized was the empirical bootstrap. We faced some limitations in our analysis due to all of our data being categorical. However, we could have used one-hot encoding, but at the time, the idea did not come to us. This severely limited our scope.

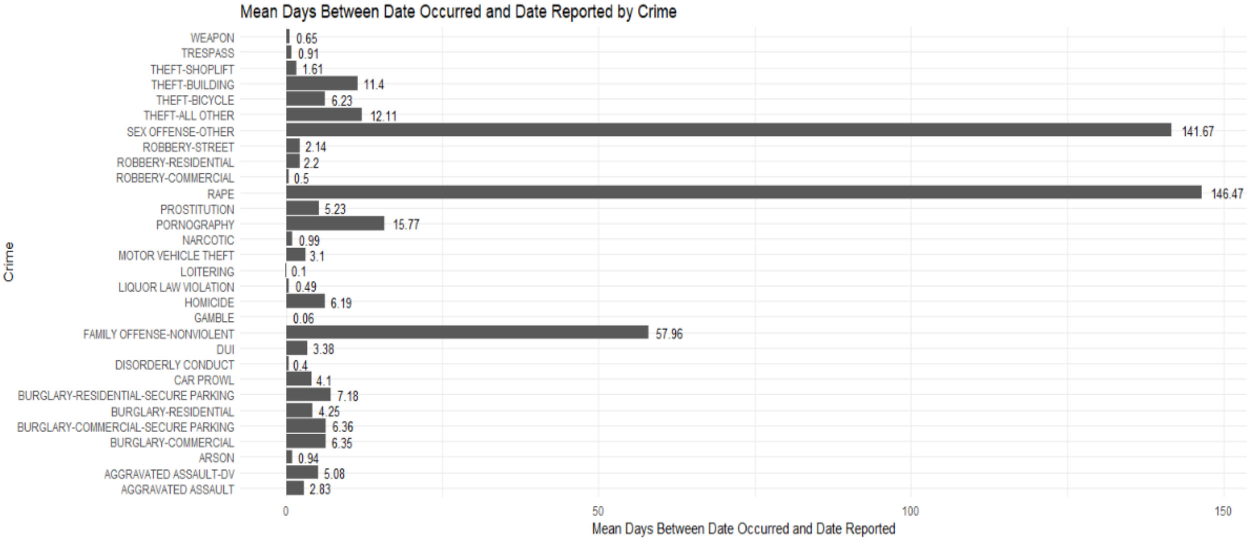
The dataset we explored is publicly provided by the Seattle Police Department. This data consists of crimes reported by a community member or witnessed by a police officer. These crimes have been reported on the basis of the hierarchical rule, where only the most serious crime is reported from a single offense. To ensure consistency, we used data from offenses that occurred after December 31st, 2007 and before March 21, 2019. This decision was made after data exploration revealed data prior to January 1st, 2008 was sparse, supposedly 40% of the data was missing. However, out of the 522,000 rows, only around 10,000 rows of the data was reported on January 1, 1975 through December 31, 2007, accounting for only a small portion.

Beyond that, we chose to convert any crimes with blank labels as "UNKNOWN" for the sake of convenience and readability. Furthermore, we created a new column denoted "Time Difference" which is the difference between the date when the crime occurred and when the crime was reported (Reported - Occurred).

To address our first question, we tested whether that the proportion of crime in the U-District is equivalent to the proportion of crime in Seattle proper for each type of crime in our dataset. To do this we made 31 null hypotheses, one for each type of crime we tested. We took 10,000 empirical bootstraps of the crimes in Seattle, sampled with a sample size equal to the amount of crime in the U district. This gave us 10,000 samples of total Seattle crime with the sample size of the U district. Next, we found quantiles for it that gave us a 90% CI. If the true University proportion lied within that 90% confidence interval, then we had no strong evidence to reject that the crime proportions between all of Seattle and U-District differs, meaning that type of crime in Seattle happens at the same rate as that type of crime in the U district. However, we decided to use the Bonferroni correction because we tested roughly 31 hypotheses with this bootstrap. We chose this method to ensure that we had significant enough evidence to reject any given hypothesis. The bonferroni correction turned our significance level into $\alpha = \frac{0.05}{31} = 0.00161229$.

In this case, we chose to resample because it allows us to construct confidence intervals. Had we not been interested in confidence intervals we could have used a Chi-Square test Goodness of Fit.

To address our second question, we first conducted basic exploratory analysis of the problem by finding the average difference between the date a crime occurred and the date the crime was reported (denoted time difference) for each different crime type within all of Seattle proper.



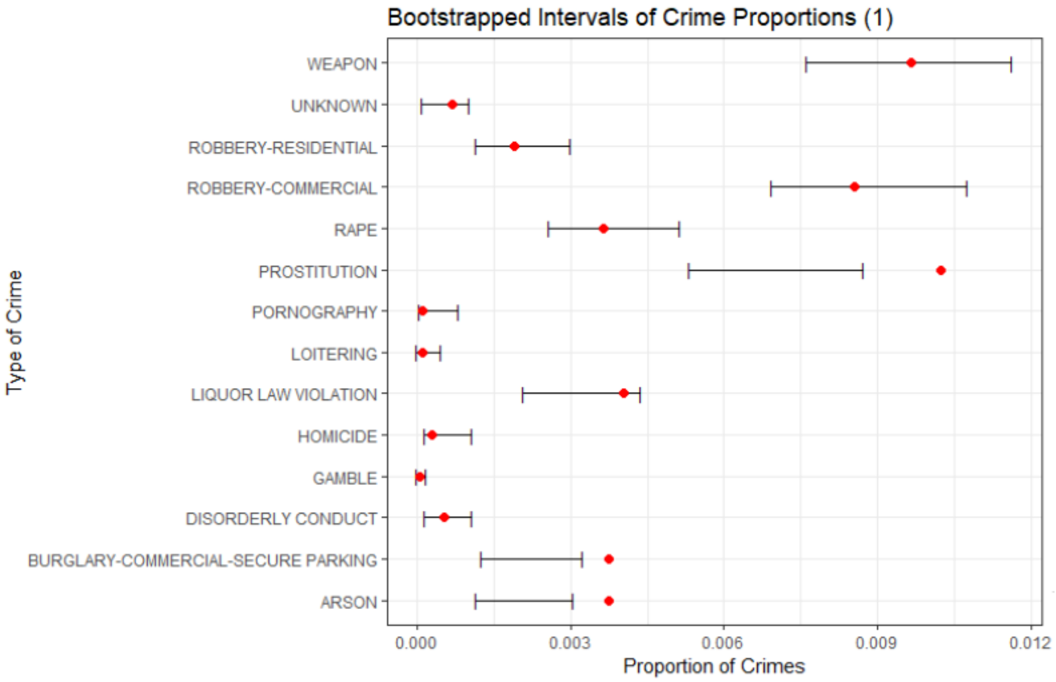
We found that rape, other sexual offenses, and nonviolent family offenses had extremely high average time difference values compared to the other crimes. After seeing this, we decided to proceed in our investigation of whether U-District's average time difference differs from Seattle proper's average time difference only for these crimes.

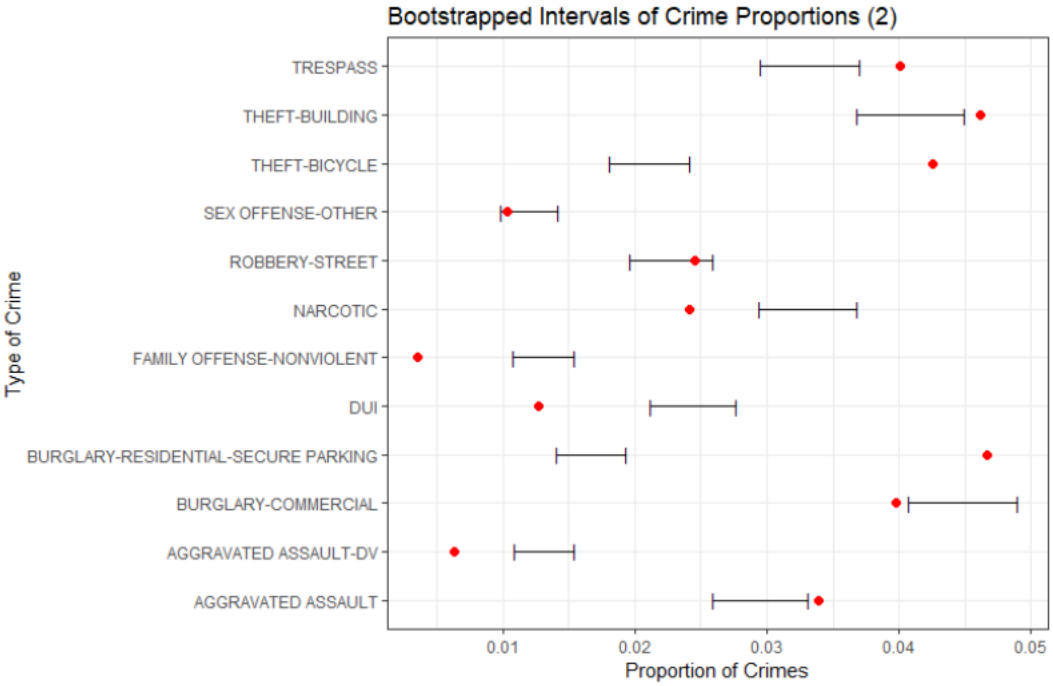
For each of these three crimes, we took 10,000 empirical bootstraps of cases of the crime from the Seattle data sampled with a sample size equal to the number of cases of the crime in the U district. This gave us 10,000 samples of cases of the crime from Seattle with the sample size equal to the number of cases of the crime in U district. For each of our 10,000 samples of cases of the crime, we found the average time difference. Next, we found quantiles for the average time difference that gave us a 90% CI. If the true University average time difference for the crime lies within that 90% confidence interval, then we had no strong evidence to reject the notion that the average time difference for the crime in Seattle does not differ from the average time difference for the crime in the U district. However, we similarly decided to use the Bonferroni correction for 3 different hypotheses. The bonferroni correction turned our significance level into $\alpha = \frac{0.05}{3} = 0.017$.

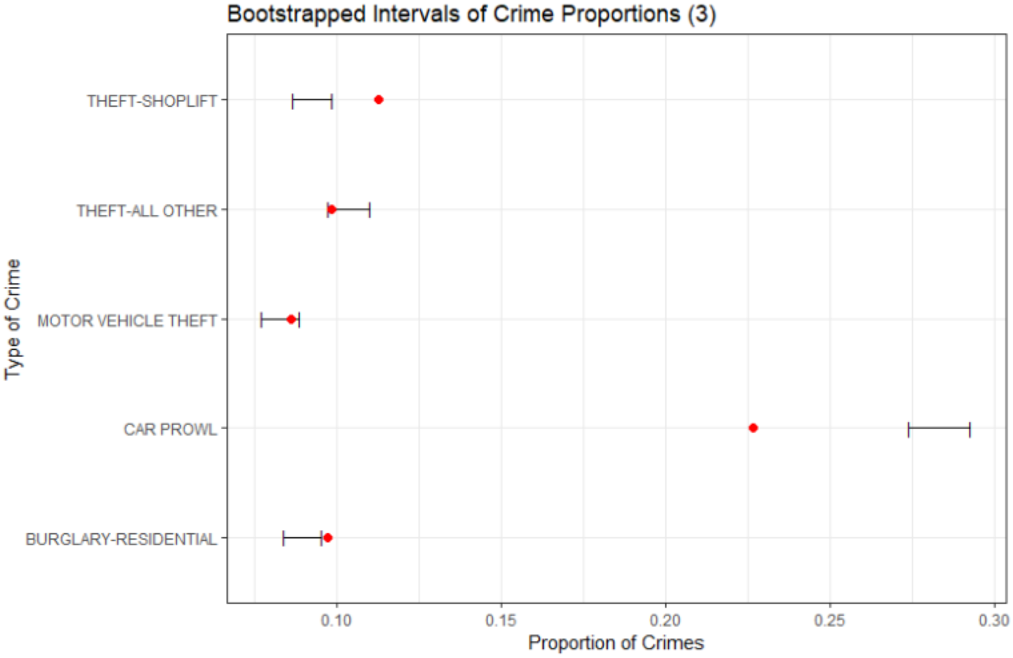
In both questions, we chose to resample because it allows us to construct confidence intervals. Had we not been interested in confidence intervals we could have used a Chi-Square test Goodness of Fit.

Results

The following plots represent the results of our first question:







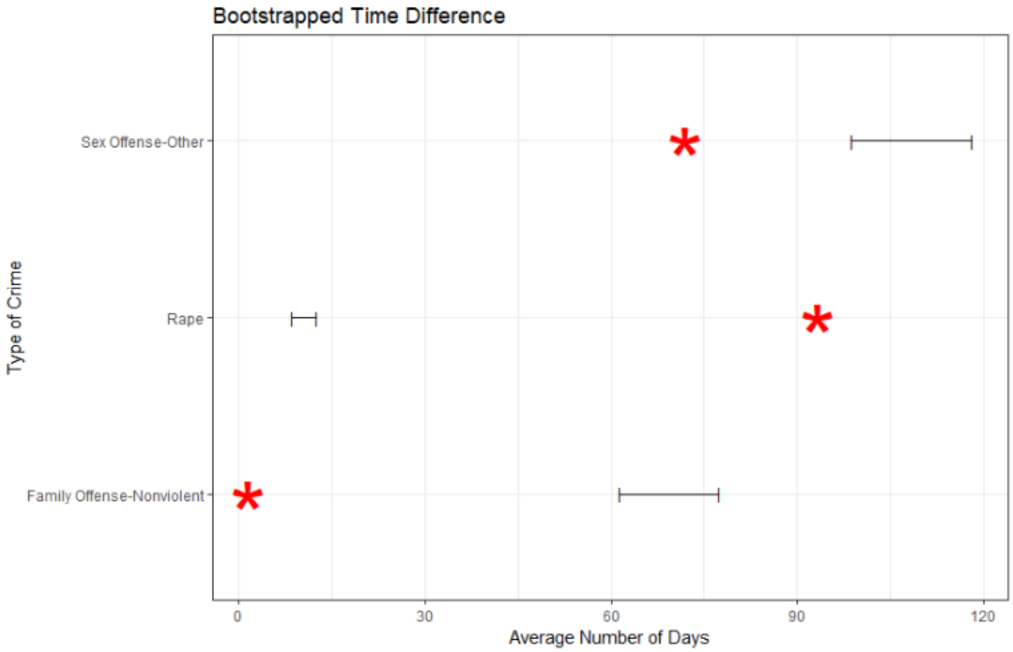
On these plots, the intervals represent our bootstrapped intervals, including the Bonferroni Correction, and the red dot represents the true value for the crimes in the U-District.

We found that 16 out of the 31 crime proportions were significant. In the U-District, the crimes that were significantly higher were: Arson, Prostitution, Trespass, Theft-Building, Theft-Bicycle, Burglary-Residential-Secure Parking, Assault, Theft-Shoplift, Burglary-Residential, Burglary-Commercial-Secure Parking, Aggravated Assault.

The crimes that were significantly lower in the U-District were: Narcotic, Family Offense-Nonviolent, DUI, Burglary-Commercial, Aggravated Assault-DV.

Overall, we have significant evidence to reject that the proportion of these specific crimes in the U-District is similar to the proportions of all of Seattle.

For the second question, we found that for all three of these crimes, the average time difference in U-District significantly differed from the average time difference in all of Seattle.



The average time difference was higher in U-District compared to Seattle proper for rape, while it was lower in U-District for other sexual offense and nonviolent family offenses.

Discussion

Our results indicate that we have support for the notion that the University District is unique compared to the whole of Seattle. We could possibly attribute this to the huge population of college aged students within the district, increasing the population and affecting the group dynamics that occur. For example, a likely reason that “Aggravated Assault-DV” (Domestic Violence) is lower is because a lower percentage of “nuclear” homes. Due to the nature of the data, we only had access to crimes committed and some broad details combined with limited background and domain research, we can not extrapolate much from our results. In the future, we can combine this data with possible individual case details and psychological research to acquire a better understanding of our results to venture into why these crimes occurred and how they happened. Ultimately, we need to supplement our analysis with a stronger understanding of the social and political landscape affecting crime in our city, especially within the University District.