City of Seattle crime trend analysis

University of Washington, Data 557
Winter 2019
Tara Wilson, Ben Brodeur Mathieu, Lauren Heintz, Will Wright

I. Abstract

An increased understanding of the distribution of crime in Seattle provides a platform to help prevent offenses, keeping residents and their belongings secure. Crimes within the Seattle city boundaries recorded by the Seattle Police Department between 2009 and 2018, were analyzed to communicate the influence of certain factors on crime frequency. Demographic data was brought in to allow linking of reported incidents with details about their location. To understand the cyclic variation of crime occurrences, an ANOVA test of independence was conducted. It was found that crime is related closely with hour and time of day, but is essentially constant between different months and seasons. A linear regression was conducted to model the influence of median income on crime count by neighborhood. This revealed that median income is not statistically tied to crime occurrence when controlled for additional demographic factors. A chi-square test showed that crime types and Seattle city regions interact, with certain crime types occurring more frequently in certain regions. Overall, it was seen that time and location influence the occurrence and types of crime more heavily than other factors such as time of year and median neighborhood income.

II. Introduction

The first question addressed in this analysis is whether or not there are differences in the counts of crime for different hours, times of day, months and seasons. Frequency counts of crime were examined over these various time classifications to see which seemed to be associated with higher occurrences of reported crime.

Next, demographic data was brought in to address the relationship between the median income of Seattle neighborhoods and crime volume. This investigation explored the connection between these factors while adjusting for a neighborhood's population, gender and age.

Finally, the distribution of crime types for different regions of Seattle was explored. This question was answered by conducting significance tests under the assumption that specified types of crime were equally likely to happen in all regions of the city.

Altogether, these separate investigations helped provide a more complete picture about when and where crime waslikely to happen as well as which types were more common than others in an area.

III. Dataset description

The chosen dataset is updated daily and lists crimes reported to the Seattle Police Department. It is hosted on the City of Seattle website (1). Each row contains information about the crime report record including time, location and offense categorization. This data is used for strategic planning, accountability, and performance management (2). The categorizations are set up to simulate the standard reported by the Federal Bureau of Investigation. Only records from years 2009-2018 were analyzed.

For the first question, the selected features were the date and time the crime occured, and collated totals derived from this data. The time listed for each crime was reported in two ways: A 'Reported Time' and an 'Occurred Time'. Since the questions for this analysis pertained to when crime is likely to occur, 'Occurred Time' was selected. It is crucial to consider that the 'Occured Time' for a reported incident may be inaccurate in some cases, as some crimes may occur in the absence of an observer and a best-guess estimate may have been recorded.

For the second question, external information about Seattle neighborhoods was brought in to complement the original dataset. Data was sourced from City-Data (3) and included demographic information such as population size, median income, median age, and gender proportions. These were added to the model to support the income predictor. It was assumed variables such as age would have a confounding relationship with median income as people tend to have higher incomes further along in their career, for example. It should be noted that the City-Data dataset provided data for 109 Seattle neighborhoods whereas the City of Seattle Police Department crime data was assigned to 58 neighborhoods. Some neighborhoods in the City-Data listing had different names, or were separated into multiple sub-neighborhoods where the Seattle PD data only had one. Local knowledge, as well as neighborhood definitions from Google Maps were used in order to consolidate this data in a systematic way.

To assess the association between crime types and Seattle regions, five specific crimes were chosen from the provided list of 'Primary Offenses'. These crimes were selected based on two criteria: A large enough sample size and interest and relevance to the University of Washington student population. The five 'Primary Offenses' chosen were: car prowl, bicycle theft, motor vehicle theft, street robbery and residential burglary. To elaborate on some of these in legal terms: 'Theft' is defined as taking another's property without intent to return, 'Car prowl' indicates there was theft of property inside a vehicle, 'Robbery' differs from theft in it includes the use of force, intimidation, or threat and 'Residential burglary' is defined as entering a home/residence with the intention of theft.

As it is impossible to generate and randomize crime, the study is considered observational. According to the data's source, records represent independent events however there may be some dependence between records. For example, one person committing multiple crimes in one evening would be represented as multiple independent records but would have an inherent and unmarked dependency. For these reasons, there may exist slight dependence between crimes, but they will be treated as independent as there is no way to assess which may have underlying commonalities.

VI. Statistical methods

Question 1 - Crime volume and time

The aim of this question is to evaluate the association between crime volume and time specifically within days and years. To address this, some aggregation of the data was required. This analysis took the 'Occurred Time' and assigned each crime an 'Hour of Day' indicator where 00:00-01:00 is Hour 0 and so on for all 24 hours of the day. Since looking at each hour of the day may be too narrow in order to grasp general trends in crime throughout the day, another factor called 'Time of Day' which breaks down the 24 hours in four sections was added. 'Time of Day' is defined as Morning (6am-12pm), Afternoon (12pm-6pm), Evening (6pm-12am), and Night (12am-6am). Within the provided date, crimes were assigned an 'Occurred Month' where 1 is January and so on for all 12 months of the year. Looking at each month of the year may also be too narrow to grasp trends in crime throughout the year, which are impacted by factors such as weather, and hours of daylight. Consequently, 'Season' was added as a factor. To simplify the analysis, Winter was defined as the months of December, January, and February, Spring as March, April, and May, Summer as June, July, and August and Fall as September, October, and November. The crime totals were then collated for each hour and time of day and for each month and season for all crimes occurring in the years 2009 to 2018.

To determine if there were any statistically significant associations between intervals within a day ('Time of Day' and hour) or during the year ('Season' and month) on the total crime counts over the years (2009-2018), ANOVA tests of significance were conducted. These units of factors were tested as well under an independent hypothesis model. There were four independent null hypotheses which were addressed by four ANOVA tests:

1. There is no difference in volume of crime based on hour of the day.
2. There is no difference in volume of crime based on time of the day.
3. There is no difference in volume of crime based on month of the year.
4. There is no difference in volume of crime based on season of the year.

The assumptions required for the ANOVA test are independence, equality of variance, and normality or large size of the sample data. The statement of independence for the test is addressed in Section III. The equality of variance assumption was verified through analysis of residuals, specifically using box and whisker plots. As seen in Figures 1 through 4, the residuals for crime volume have roughly equivalent spread regardless of the time aggregation, with several noted outliers specifically for Hour 0 and the early hours of the day (from 3 to 7 a.m.). The assumption was denoted as met.

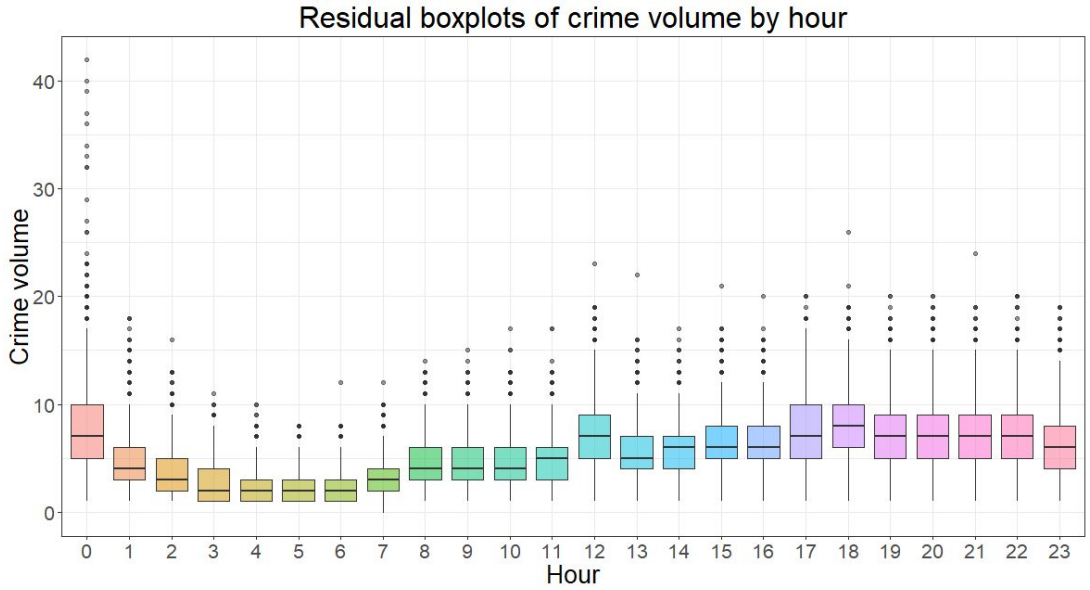

Figure 1. Residuals of crime volume aggregated by hour of day.

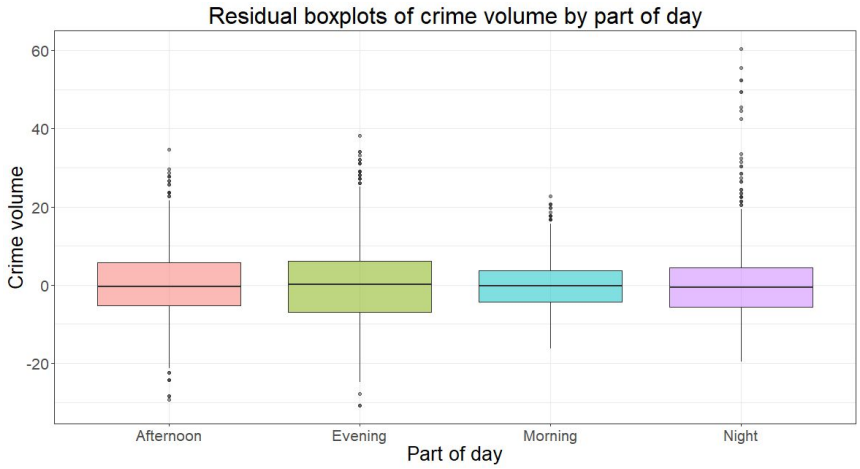

Figure 2. Residuals of crime volume aggregated by time of day.

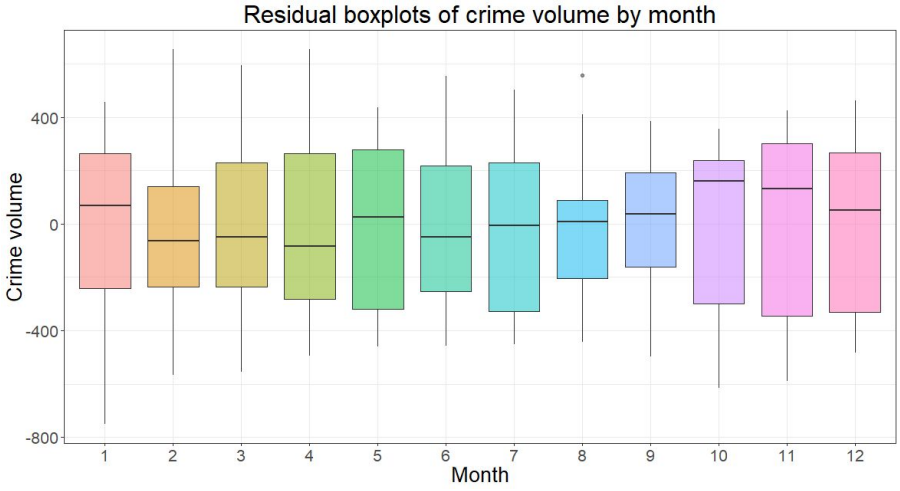

Figure 3. Residuals of crime volume aggregated by month.

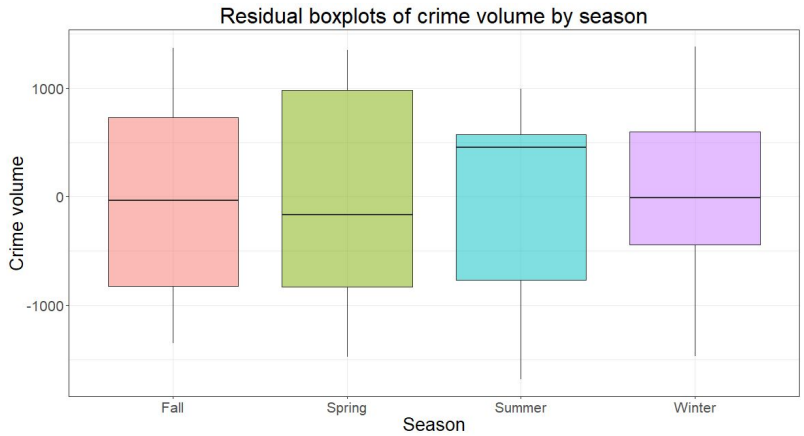

Figure 4. Residuals of crime volume aggregated by season.

	Min Sample Size	Median Sample Size	Mean Sample Size	Max Sample Size
Sample Size of Month x Hour Groups	438	1,714	1,611	2,644
Sample Size of Season x Time of Day Groups	18,989	28,552	29,001	41,061

Table 1: Sample size distributions by groups

There was a minimum of 438 entries per group with an average around 1,611 when using Month and hour for grouping. There was a minimum of almost 19,000 and an average of 29,000 when grouping by Season and time of day. These sample size were large enough to meet the assumptions requirements and provided even more power for the test.

After analyzing the p-values from the results of the ANOVA test, a determination was made on whether or not the null hypotheses were rejected.

Question 2 - Crime rate and median income

A linear regression was conducted to examine a potential association between crime rate and median income for Seattle neighborhoods. Demographic data brought in for this analysis included neighborhood median income, population, the ratio of male to female residents, the average age of male residents and the average age of female residents (3). The demographic data complimenting the original crime dataset was only available for 2016. To prevent the addition of confounding factors through assumptions about the trends of demographic data over the years, crime data was also limited to events occurring in 2016.

As mentioned in Section III, the data from the two sources were matched using neighborhood names. While most samples had direct alignment between the two datasets, some samples were either too specific or not specific enough in one dataset or the other. In these cases, data were matched as closely as possible, with North Ballard and South Ballard being summed to represent all of Ballard, and other similar adjustments. This is recognized as a limitation of this analysis, however it was decided that the matches were very accurate overall and should have limited effect on the interpretation of the results.

The response variable, crime count, is a sum of all types of crimes aggregated for each neighborhood in the year 2016. Median income was measured in dollars, rounded to the nearest whole number, and represents the median household income for the neighborhood measured in 2016. Population is the number of listed inhabitants for the specified neighborhood. The ratio of male to female inhabitants was derived from the proportional population counts for each gender and becomes a singular identifier for the split of men and women in an area. The median age for men and women was measured in years and reported to the nearest tenth of a year.

The assumptions for linear regression are independence of the observations, linearity, constant variance of the residuals, and normality or large sample size. As mentioned earlier, the crime data was treated as if from an observational study and thus was regarded as independent observations. The linearity assumption

was verified through a plot of residuals against the model's fitted values. The output is roughly linear with some values trending towards negative residuals for the median fitted values. The constant variance assumption was checked with a scale-location plot (Figure 5). The fitted line displayed a clear positive slope so it can be concluded that the variance is non-constant. To account for this assumption not being met, robust standard errors were used for the model analysis. Finally, the normality assumption was validated through both a q-q plot and a histogram. The tails of the data were heavier than expected for a normal distribution, but this criteria was roughly met, and was supplemented by the fairly large sample size of 58 neighborhoods.

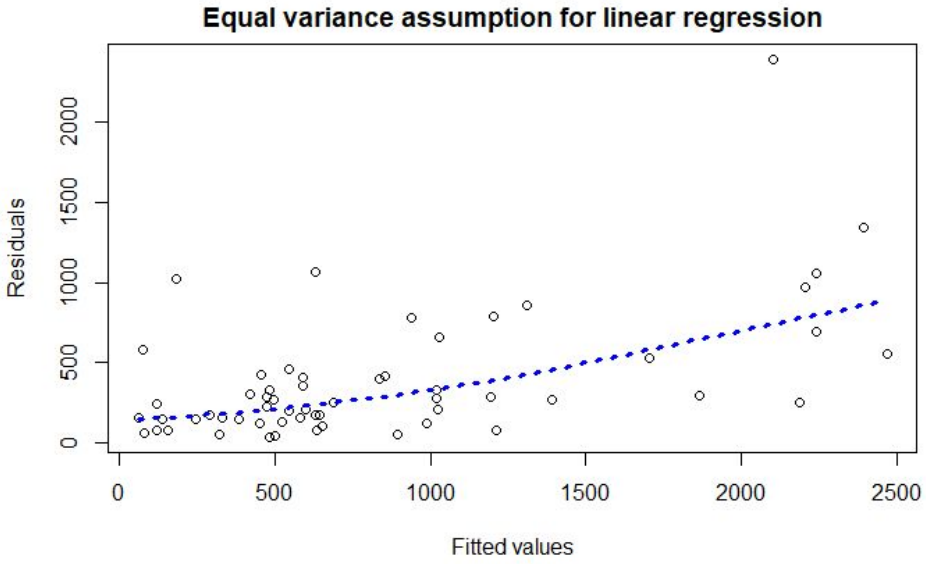

Figure 5. Residual plot for linear regression model.

Question 3 - Crime types and Seattle regions

To assess independence between regions of Seattle and crime types, a chi-square test of independence was conducted. The results form the standard chi-square test were compared with an exact (Fisher) test. The assumptions of the chi-square test are that each sample were mutually independent and that the cell counts and sample sizes overall were large enough to approximate the sample distribution with a chi-square.

To ensure the independence assumption was maintained for location and to alleviate confounding interactions between neighboring regions, the Seattle city neighborhoods contained in the dataset were classified into four geographic locations (Map 1). The quadrants were split by Interstate-5 and Interstate-90. It is acknowledged that in the context of this observational study, neighborhoods near boundaries might influence each other.

Map 1. Classifications for Seattle neighborhoods.

The independence of this data is difficult to assess due to the observational nature of this study, but the assumption was considered met for the analysis of this question as each crime is assigned to a single crime category and region. It is important to acknowledge that the dataset provided no way to ensure independence between crimes as discussed in Section III. This will be revisited in the result discussion. The size of the datasets, 500,000 samples and 450 minimal cell value, ensured that the distribution can be approximated by a chi-square due to the central limit theorem.

V. Results

Question 1 - Crime volume and time

Before analyzing the ANOVA results, it is important to analyze the bubble plot below (Figure 6), which provided a descriptive overview of the data results. The darker the shade of red fill and the larger the circle diameter indicated that more crimes were committed at the associated time. It is clearly displayed that crime occurs more frequently during the hours most people are awake. More specifically, there were clear peaks around noon, 5 p.m., and around midnight. To further analyze, the x and y axis were broken up into separate plots in the next two figures.

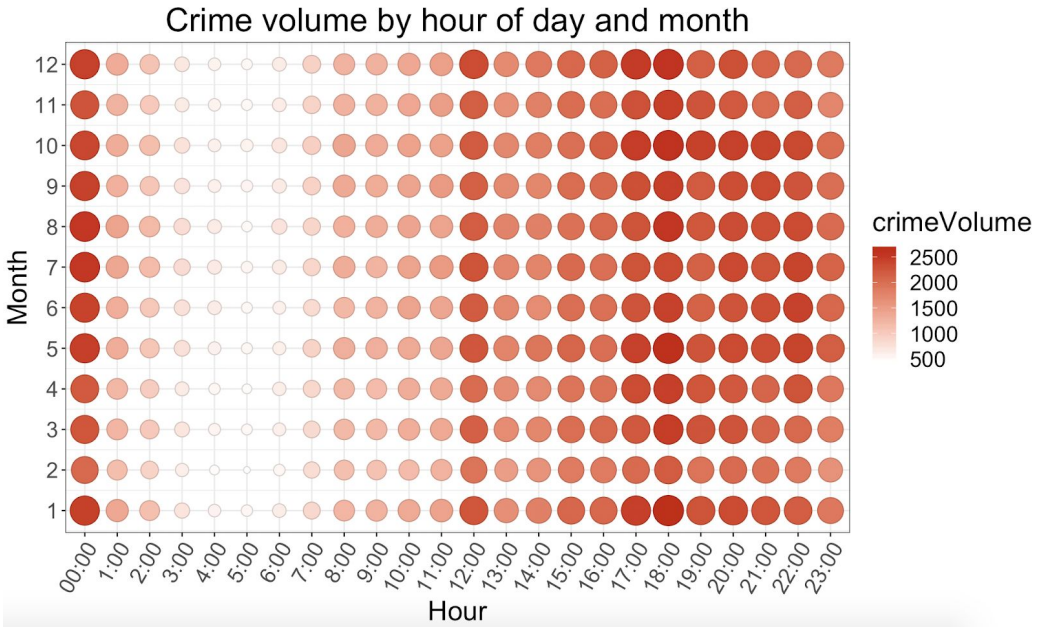

Figure 6. Crime volume by hour of day and month

Figure 7 shows the relationship between crime volume and hour of the day, averaged for year 2009-2018. Each month is shown by a different line color. It is clear that all months followed this same hourly trend. Note that lengths of the different months were not standardized by the number of days, which might explain the slightly lower crime volume for the month of February.

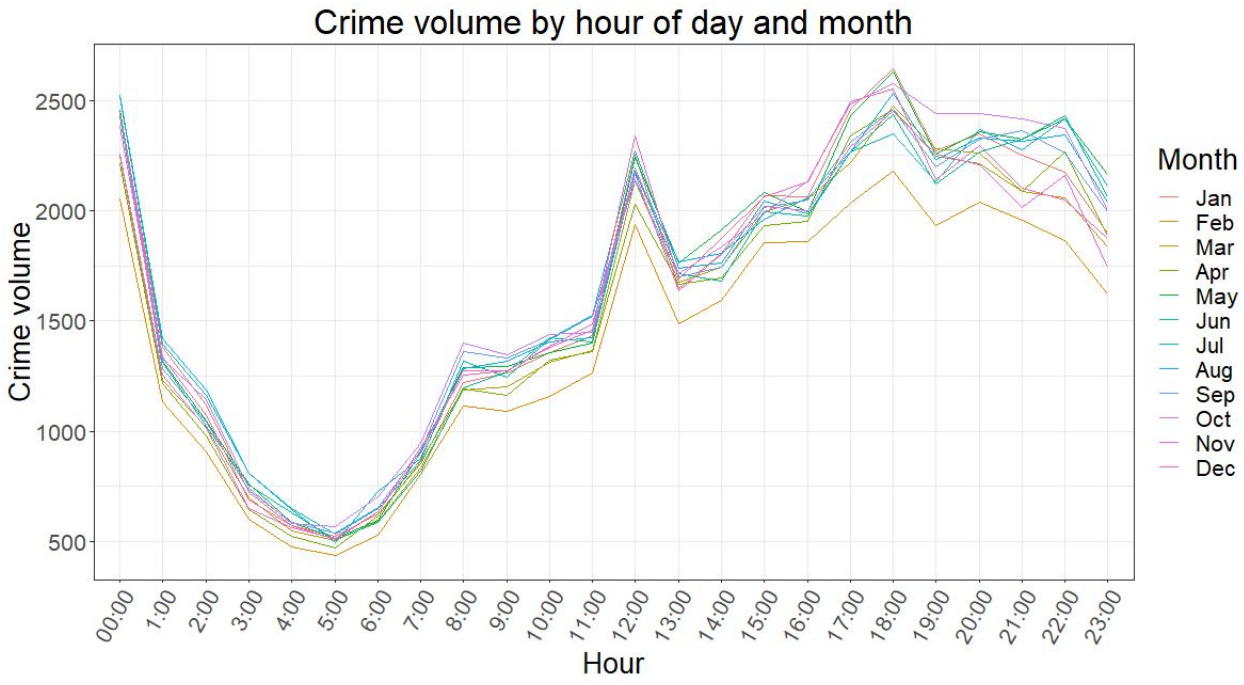

Figure 7. Average crime volume per hour for 2009-2018, grouped by month

Figure 8 shows the relationship between volume of crime and month, averaged over year 2009-2018. It is apparent that the total crime per month remained relatively constant. Again data was not standardized based on the month's length which might explain the perceived variations.

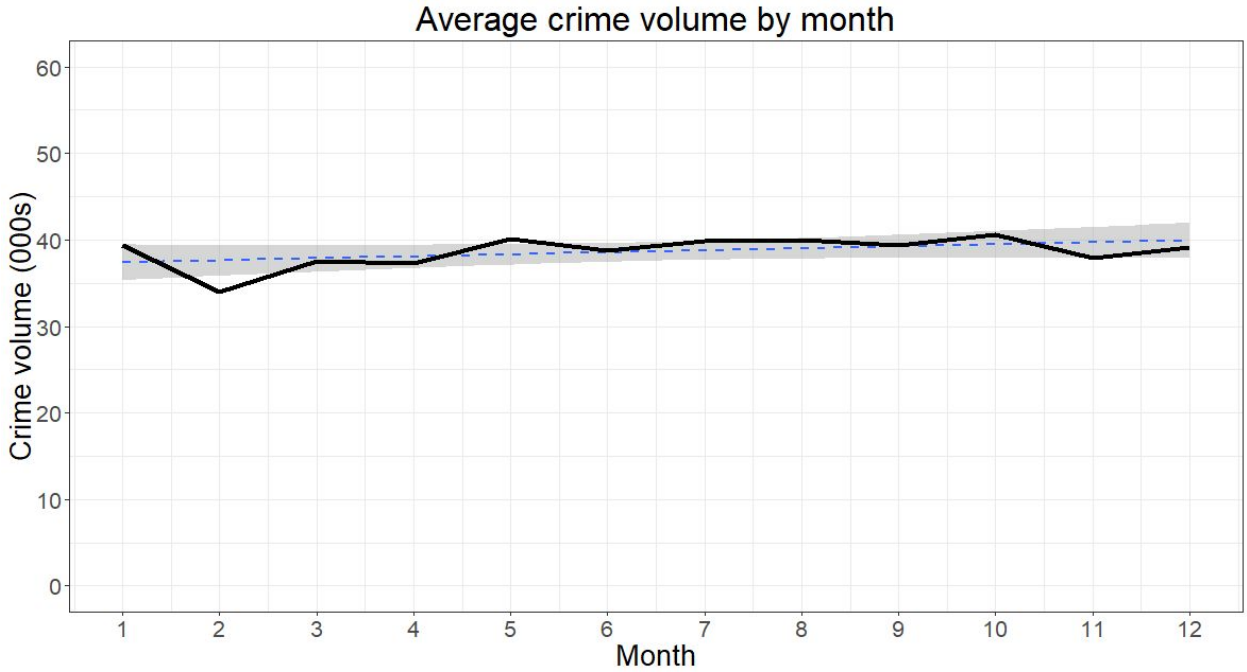

Figure 8. Average crime volume per month for 2009-2018

The ANOVA tests provided highly significant p-values ($p < 0.001$) for hour of day and time of day while p-values for season and month were not significant ($p \sim 0.98$). This corroborates the observations of Figure 7 and 8.

Question 2 - Crime rate and median income

	Estimate	**Standard error**	**Robust standard error**	**Robust z-value**
Intercept	1160	896	677	1.72
Median income	.000985	.00368	.00408	0.24
Population	.0456	.00642	.00117	3.91
Ratio male to female	602	305	287	2.10
Average age (male)	-128	45.6	36.6	-3.51
Average age (female)	83.3	43.5	31.6	2.64

Table 2: Output from the least-squares linear regression on crime count.

The multiple R-squared value for the linear regression was 0.5773. The adjusted R-squared value was 0.5367. The F-statistic was 14.21 on 5 and 52 degrees of freedom.

The linear regression produced the following equation for calculating crime count:

$$crimeCount = 1,162 + .001 * medianIncome + .046 * population + 602 * ratioMaleToFemale - 128 * medianAgeMale + 83.3 * medianAgeFemale$$

Question 3 - Crime types and Seattle regions

Both the standard chi-square test and the exact Fisher test gave highly significant p-values (< 0.001). The underlying data represented as a frequency table and a bubble plot can be found below.

	NE	NW	SE	SW	Row totals
Burglary	25,274	26,089	10,247	9,778	**71,388**
Robbery	3,059	3,827	2,186	1,262	**10,334**
Bicycle theft	4,077	4,942	450	846	**10,315**
Car prowl	36,327	57,434	10,601	12,794	**117,156**
Vehicle theft	12,584	13,483	6,328	6,248	**38,643**
Column totals	**81,321**	**105,775**	**29,812**	**30,928**	**247,836**

Table 3. Counts of crimes by crime type and Seattle region

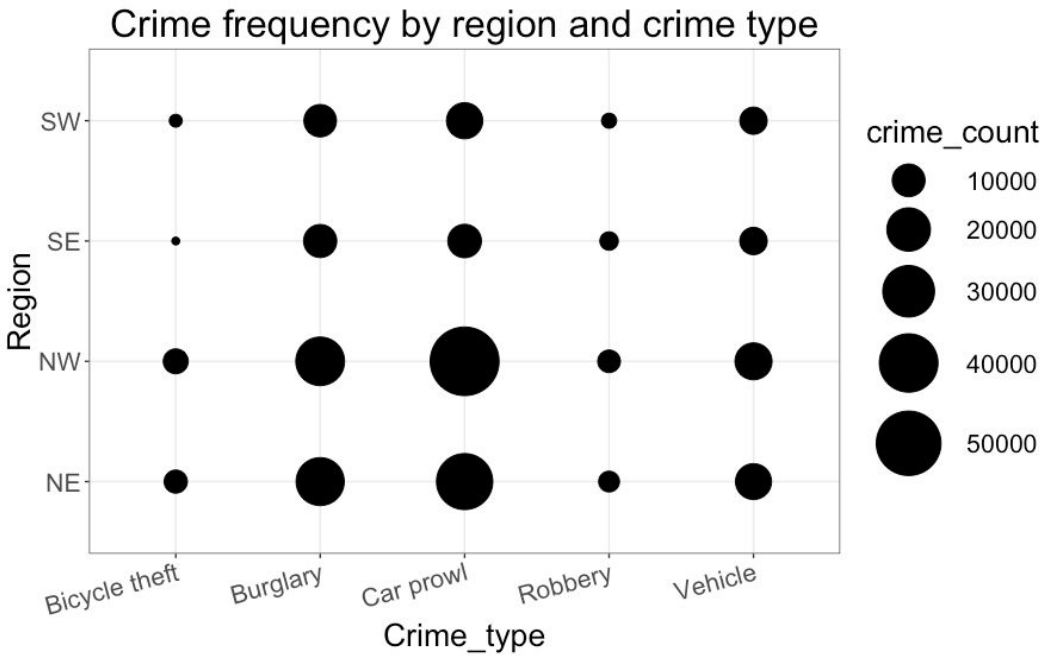

Figure 8. Crime frequency by Seattle region and crime type

Overall, there seems to be more crime in the northern regions when compared to southern regions and more in western regions when compared to eastern regions. The distribution of types of crimes vary by region (comparison of vertical lines) whereas the crime frequencies by region (horizontal lines) seem to be fairly constant.

VI. Discussion

Question 1 - Crime volume and time

Figures 6, 7, and 8 provided a descriptive overview of the data. In Figure 7, there appeared to be a drop off of crime between 2 and 6 a.m. as well as an increase around 5-6 p.m. continuing until midnight. In the same figure, each month is shown by a different colored line to emphasize the small variance in crime volume between months. Looking at each individual month seemed to indicate that they each follow the same general trend. This is reinforced by Figure 8, where there appeared to be no positive or negative relationship between any month and the volume of crime in that month. Crime appeared to be quite evenly dispersed between months. As mentioned previously, the data was not normalized by the amount of days in each month, which explained the small variations seen from month to month. Notably, Figure 6 and Figure 7 which showed February lagging slightly behind the other months in overall volume of crime was consistent with this lack of normalization with February being the shortest month by a few days.

Four separate ANOVA tests were conducted for each factor. Both hour of day and time of day were found to have extremely small p-values, both less than 0.001. This meant, for both time of day and hour of day, that there was sufficient evidence to reject the null hypothesis that 'there is no difference in the volume of crime by hour of day or time of day'. For season and month of year, there was not sufficient evidence to reject the null hypothesis ($p \sim 0.98$).

It was a limitation of this analysis that multiple hypotheses were tested for this specific question of time. As a result, the potential Type I Error rate, or rate of rejecting a true null hypothesis, of the question increased for all factors. However, even using the Bonferroni correction factor, which quite conservatively reduced the p-value necessary for a statistically significant finding, the results for time of day and hour of day were still significant and rejected the null hypothesis. Since all of the assumptions for the ANOVA model were still met, the tests and all the reported results were considered valid.

As a result, the null hypothesis that there is no relation between hour and time of day to crime volume was rejected. It is important to note that although there was an association, this result does not imply causation. Observations could have been skewed by a multitude of factors. For example, most people are asleep from 2-6 a.m., so the labelling of 'Occured time' may simply not be accurate for these hours as the crime occurrence time would be estimated when people awoke.

Question 2 - Crime rate and median income

The linear regression output indicated that the difference in mean crime rate per dollar increase in median income was only 0.001, after adjusting for the potential confounding effects of population size, the ratio of male to female residents in the neighborhood, the median age of males in the neighborhood and the median age of females in the neighborhood. This translated to an increase of 1 crime for a difference of 1,000 dollars in median income when comparing neighborhoods with identical population size, gender ratio, and median ages for males and females. The null hypothesis that median income for a neighborhood has no effect on crime count was not rejected in this case (robust $p > 0.05$).

The power for this linear model was over 99%, indicating that the probability that the linear model would correctly reject an alternative hypothesis of median income being associated with crime volume was incredibly high with effect size 1.16 and significance level 0.05 (4). Due to this high power, there is confidence in the conclusion of the test to not reject the null hypothesis. Furthermore, a Poisson regression using the robust standard errors was also conducted and produced the same conclusions as the linear model with robust standard errors.

The adjusted R-squared value for the linear model was 0.5367. Therefore, just over 50% of the variability in the crime count model was explained by the included parameters and other factors are likely contributing to the control of crime count for various neighborhoods.

The linear regression model produced significant (robust $p < 0.001$) results for neighborhood population, ratio of male to female residents, average age of males and average age of females. A deeper analysis into the association of these variables on crime rate should be conducted as this model indicates they may have influence on the frequency of crime. The significance of population size and crime rate was in accordance with a recent investigation done by Seattle Post-Intelligencer using the same dataset (7). It was uncovered in this analysis that crime frequency has been rising along with the increasing population of Seattle.

When controlling for variations in population size, gender ratio and median age, the association between median income and crime count was minimal. This indicates that the driving factor of volume of crime in a neighborhood is something not determined by the median income model. Other studies have concluded that there is high correlation between differences in income and occurrence of crime (6). As a result, it is likely that the model used for this analysis does not accurately represent the economic disparity of victims and criminals. Median income may not be the best metric to encompass the wealth of an area. Accounting for the spread of income for an area, as was done in another study (6), may be more indicative of the actual demographic. Additionally, median income is a complex metric within itself: higher income may be correlated with more material possessions and opportunities for crime, but may also be associated with more crime prevention devices such as home alarm systems. Additional analysis should be conducted to uncover the best metric to represent the financial level within an area when evaluating the amount of crime experienced there.

Question 3 - Crime types and Seattle regions

Both the standard and exact chi-square tests conducted indicate, with a p-value < 0.001, that there was an association between the distribution of types of crime by region of Seattle. These results were bolstered by the fact that all assumptions were met for both tests.

The frequencies shown in Figure 8 indicated that the crime types were more likely to occur in northern areas of Seattle as opposed to those in the south part of the city. Crime also appeared slightly more likely to occur in western neighborhoods of the city as opposed to eastern areas. It was also observed that there more variations in the relative distribution of types of crimes by region (comparison of vertical lines) than the relative crime frequencies per region (comparison of horizontal lines).

The sectors of the city were divided by physical barriers so that roughly the same amount of neighborhoods (±3 neighborhoods) fall in each quadrant, however this did not ensure that equal populations fell in all four sectors. It has been seen that larger population sizes are tightly coupled with higher crime rates (7) so this may be skewing the unstandardized counts. As an example, according to our dataset, car prowl specifically seemed more likely to occur in the Northwest region. Sources such as the Seattle Times (8), show that the northwest neighborhood also has the highest amount of cars. If our study controlled for additional factors such as the number of cars per region, it is possible that the differences between region would not have been significant.

Additional investigation and data gathering are required to address discrepancies in crime distributions between regions. With enough control for confounding variables across crimes types and regions, it would then be possible to make claims such as “It is more likely for a car to be stolen from in the northwest region of Seattle.” In contrast, the current dataset can make statements of the form: “The distributions of the five crime types are significantly different in the northwest region when compared to the other regions.”

VII. References

1. City of Seattle, Seattle Police Department crime data
   https://data.seattle.gov/Public-Safety/Crime-Data/4fs7-3vj5

2. City of Seattle, Seattle Police Department API overview
   https://dev.socrata.com/foundry/data.seattle.gov/4fs7-3vj5

3. City-Data.com, Neighborhood maps
   http://www.city-data.com/nbmaps/index.html

4. R documentation, Getting started with the pwr package
   https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

5. Biochemia Medica, The Chi-square test of independence
   https://www.biochemia-medica.com/en/journal/23/2/10.11613/BM.2013.018

6. The Economist, The stark relationship between income inequality and crime
   https://www.economist.com/graphic-detail/2018/06/07/the-stark-relationship-between-income-inequality-and-crime

7. Seattle PI, 2018 crime by neighborhood: Seattle crime numbers rise steadily with population
   https://www.seattlepi.com/local/crime/article/2018-crime-by-neighborhood-Seattle-crime-numbers-13671220.php#photo-10479681

8. The Seattle Times, Housing cars or housing people? Debate rages as number of cars in Seattle hits new high
   https://www.seattletimes.com/seattle-news/data/housing-cars-or-housing-people-debate-rages-as-number-of-cars-in-seattle-hits-new-high/