Auto Theft Analysis

Cruel Cookies - Lixiang Zhu, Park Yongjun, James Lee, Omar Rojas

1. Background & Motivation

Our project was motivated by the high rates of auto theft in Seattle. Seattle ranked 8th for auto theft for all metro areas in the US. Due to the high rates, in depth research on auto theft rates would surely lead to interesting findings. To do this kind of analysis would require a large amount of reliable data, and luckily this is provided publicly by the city of Seattle. We were motivated by the idea of an application where users could input their address, and an algorithm would produce a risk score for parking their car in that area. Hopefully incorporating many variables, such as car model and time of year. It would provide a recommendation of whether it was safe to park or not, or suggest alternate locations. Such an application could potentially reduce the rate of car thefts in Seattle, effectively saving individuals the thousands of dollars of losses and emotional distress that come with auto theft.

Research Questions

Our main mission that we wanted to accomplish from this research was to find out:

- Where and when in Seattle are cars more likely to be stolen?

In order to arrive at a conclusion for the main question, we first needed general understanding of how the incidents are distributed location-wise and time-wise. Therefore our next research question was:

- Is all crime distributed across Seattle census tracts?

In more detail, we wanted to know if the the characteristics of specific neighborhoods play a significant role in the auto theft rate, so we wanted to ask:

- how does the frequency of auto theft incidents in Seattle correlate to other factors in the census tracts specifically white/non-white ratio and population?

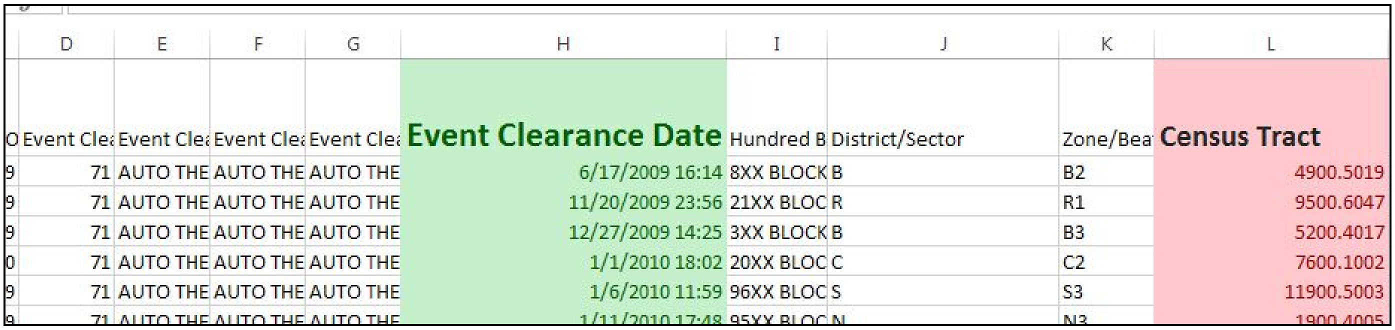
Then, after we find a correlation or do not find a correlation among these data sets, we can proceed to ask:

- Is there a predictive model for auto theft in Seattle?

2. Descriptive Data Analysis

2.1 What is our Data?

We obtained the data for rate of auto theft by census tract from data.seattle.gov. The data was formatted as a CSV file with columns for census tract, type of crime, and yearly rate. The data was available for years 2007- 2010. We used data for every year in our visualization.



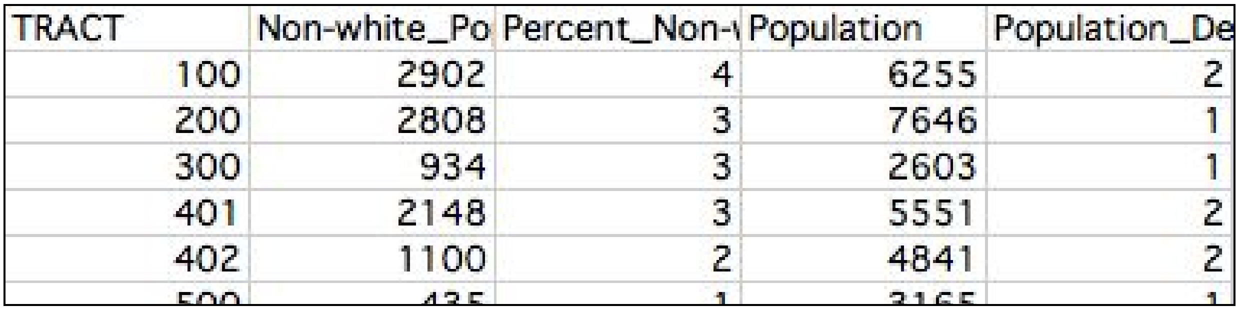
	D	E	F	G	H	I	J	K	L
O	Event Clea	Event Clea	Event Clea	Event Clea	Event Clearance Date	Hundred B	District/Sector	Zone/Bea	Census Tract
9	71	AUTO THE	AUTO THE	AUTO THE	6/17/2009 16:14	8XX BLOCK	B	B2	4900.5019
9	71	AUTO THE	AUTO THE	AUTO THE	11/20/2009 23:56	21XX BLOC	R	R1	9500.6047
9	71	AUTO THE	AUTO THE	AUTO THE	12/27/2009 14:25	3XX BLOCK	B	B3	5200.4017
0	71	AUTO THE	AUTO THE	AUTO THE	1/1/2010 18:02	20XX BLOC	C	C2	7600.1002
9	71	AUTO THE	AUTO THE	AUTO THE	1/6/2010 11:59	96XX BLOC	S	S3	11900.5003
9	71	AUTO THE	AUTO THE	AUTO THE	1/11/2010 17:48	95XX BLOC	N	N3	1900.4005

DATA 1 - Auto theft data by census tract

Census tracts are how the US government divides cities for census purposes. We are basing our visualizations on census tracts due to their inclusion in the data. Seattle has total of 265 census tracts of various sizes.

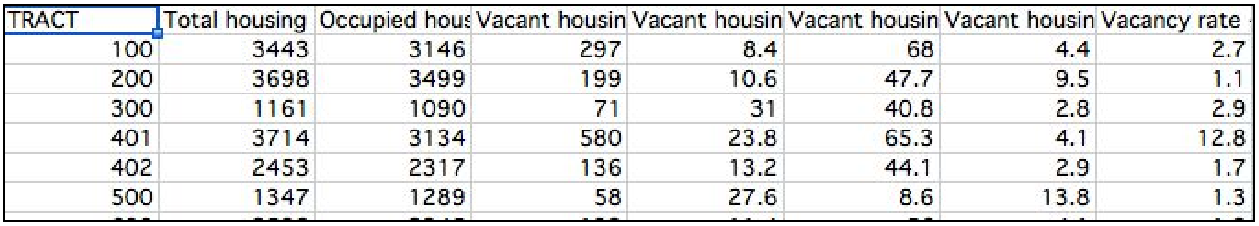
Census tracts are a good choice for data analysis because they are created with respect to their characteristics. According to the US Census Bureau, they are "Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants" (United States Census Bureau).

In order to find correlation between variables, we also found population data and housing data of Seattle by census tract. The data is from the United States Census Bureau:



TRACT	Non-white_Po	Percent_Non-\	Population	Population_De
100	2902	4	6255	2
200	2808	3	7646	1
300	934	3	2603	1
401	2148	3	5551	2
402	1100	2	4841	2
500	435	1	3165	1

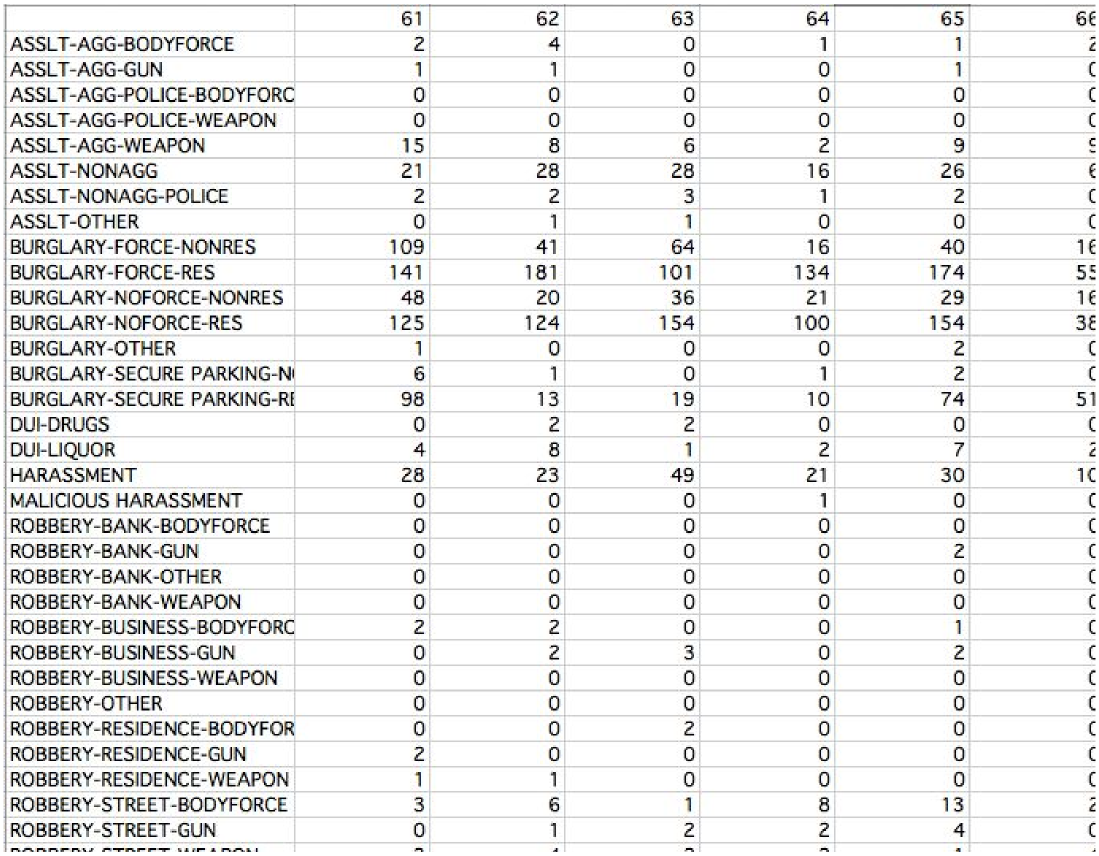
DATA 2 - Population data by census tract



TRACT	Total housing	Occupied hous	Vacant housin	Vacant housin	Vacant housin	Vacant housin	Vacancy rate -
100	3443	3146	297	8.4	68	4.4	2.7
200	3698	3499	199	10.6	47.7	9.5	1.1
300	1161	1090	71	31	40.8	2.8	2.9
401	3714	3134	580	23.8	65.3	4.1	12.8
402	2453	2317	136	13.2	44.1	2.9	1.7
500	1347	1289	58	27.6	8.6	13.8	1.3

DATA 3 - Housing data by census tract

We used this data to find the factors that may be correlated with auto thefts. They have the population of non-whites and the number of vacant housing units for each census tract in the Seattle downtown area. We hypothesized that these variables may be correlated with auto theft, so we created a multivariable linear regression based on this data.



	61	62	63	64	65	66
ASSLT-AGG-BODYFORCE	2	4	0	1	1	2
ASSLT-AGG-GUN	1	1	0	0	1	0
ASSLT-AGG-POLICE-BODYFORC	0	0	0	0	0	0
ASSLT-AGG-POLICE-WEAPON	0	0	0	0	0	0
ASSLT-AGG-WEAPON	15	8	6	2	9	9
ASSLT-NONAGG	21	28	28	16	26	6
ASSLT-NONAGG-POLICE	2	2	3	1	2	0
ASSLT-OTHER	0	1	1	0	0	0
BURGLARY-FORCE-NONRES	109	41	64	16	40	16
BURGLARY-FORCE-RES	141	181	101	134	174	55
BURGLARY-NOFORCE-NONRES	48	20	36	21	29	16
BURGLARY-NOFORCE-RES	125	124	154	100	154	38
BURGLARY-OTHER	1	0	0	0	2	0
BURGLARY-SECURE PARKING-N	6	1	0	1	2	0
BURGLARY-SECURE PARKING-R	98	13	19	10	74	51
DUI-DRUGS	0	2	2	0	0	0
DUI-LIQUOR	4	8	1	2	7	2
HARASSMENT	28	23	49	21	30	10
MALICIOUS HARASSMENT	0	0	0	1	0	0
ROBBERY-BANK-BODYFORCE	0	0	0	0	0	0
ROBBERY-BANK-GUN	0	0	0	0	2	0
ROBBERY-BANK-OTHER	0	0	0	0	0	0
ROBBERY-BANK-WEAPON	0	0	0	0	0	0
ROBBERY-BUSINESS-BODYFORC	2	2	0	0	1	0
ROBBERY-BUSINESS-GUN	0	2	3	0	2	0
ROBBERY-BUSINESS-WEAPON	0	0	0	0	0	0
ROBBERY-OTHER	0	0	0	0	0	0
ROBBERY-RESIDENCE-BODYFOR	0	0	2	0	0	0
ROBBERY-RESIDENCE-GUN	2	0	0	0	0	0
ROBBERY-RESIDENCE-WEAPON	1	1	0	0	0	0
ROBBERY-STREET-BODYFORCE	3	6	1	8	13	2
ROBBERY-STREET-GUN	0	1	2	2	4	0

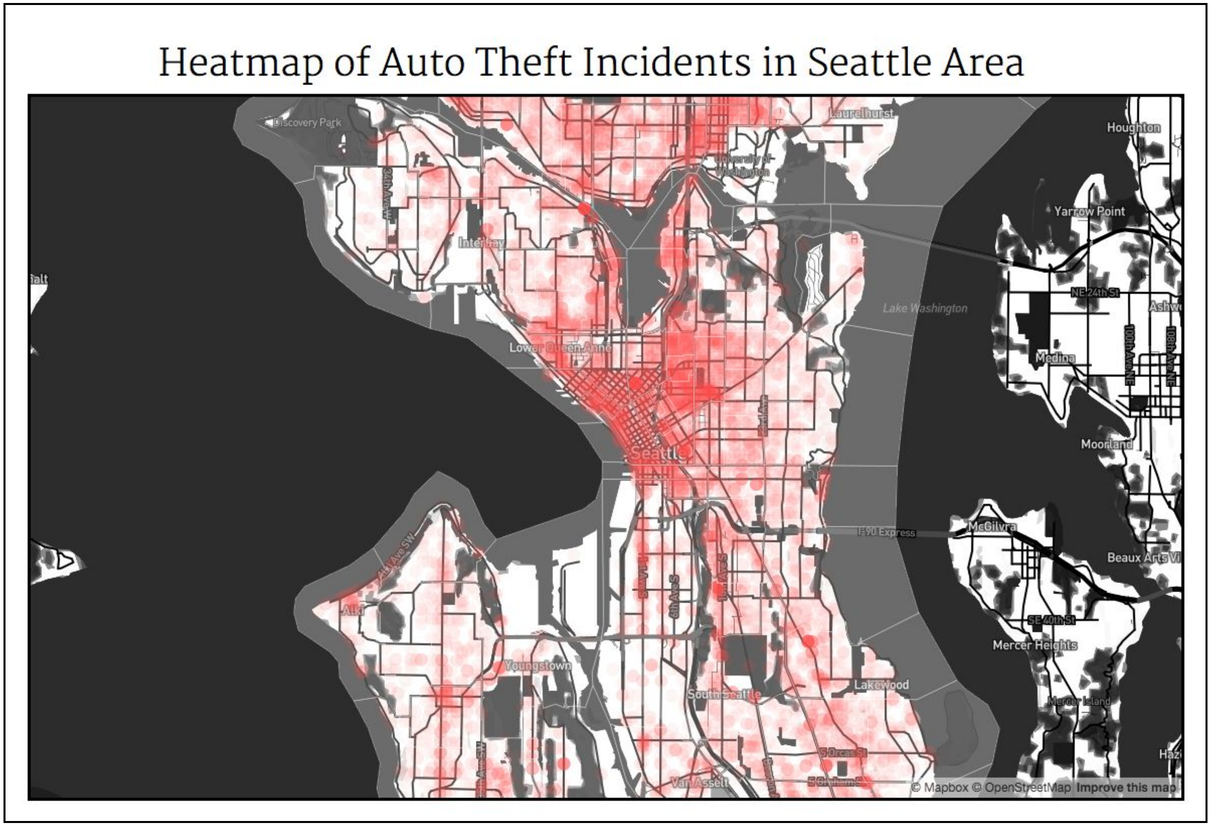
DATA 4 - Overall crime data by census tract

This data keeps track of the frequency of the different crimes in each census tract. This data table was only available for the most populated downtown areas of Seattle (census tracts 61-91). The data consists of records spanning back to 1996, allowing us to find more detailed insights about these specific areas.

2.2 What did we do with the data?

The Heat Map

To answer our first research question, we decided to plot every data entry in our Data 1 (Auto theft data by census tract) to create a heat map of the Seattle area, where auto theft has occurred most frequently. Below is the result we have created through the open source mapping tools Leaflet.js and Mapbox Studio:



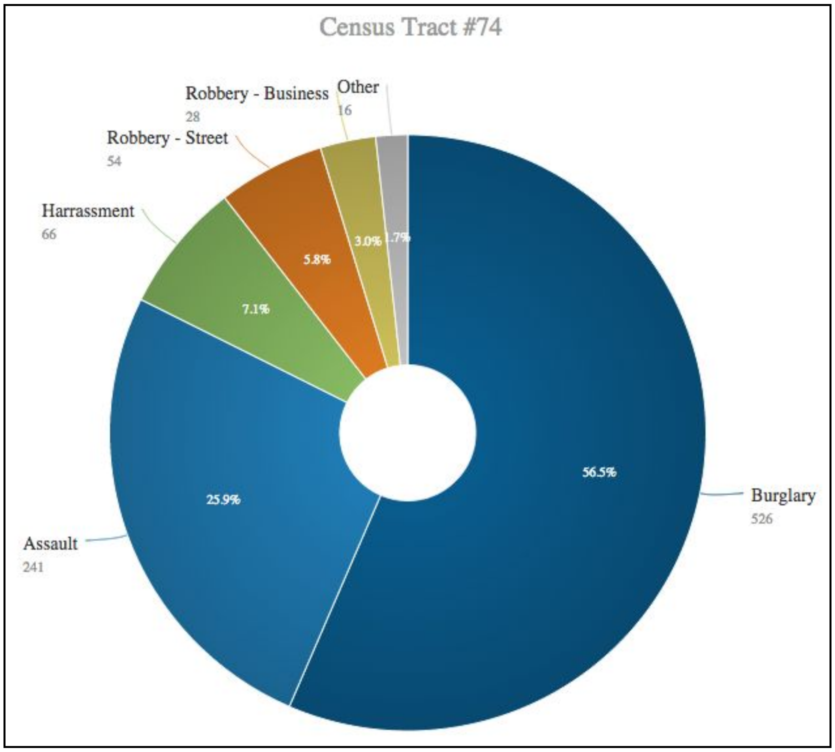
VIZ 1 - Heatmap of Auto Theft Incidents

The heat map was also provided with the geojson data of the census tracts in Seattle. The geo geojson data included the boundaries of different census tracts of Seattle, allowing us to have a better visual perception of which census tracts are more concentrated with auto theft incidents.The thin white borders layer of the map is representative of the boundary data.

VIZ 1 shows that auto theft is concentrated in the most populated areas of Seattle such as Pike Place (Census Tract 74) and Southern Downtown (Census Tract 91). However, there were some outliers such as South Seattle, where some secluded industrial areas became hotspots for for auto theft. However, simply knowing where these incidents were happening was not enough information for us to conclude how or why incidents are happening frequently in these areas.

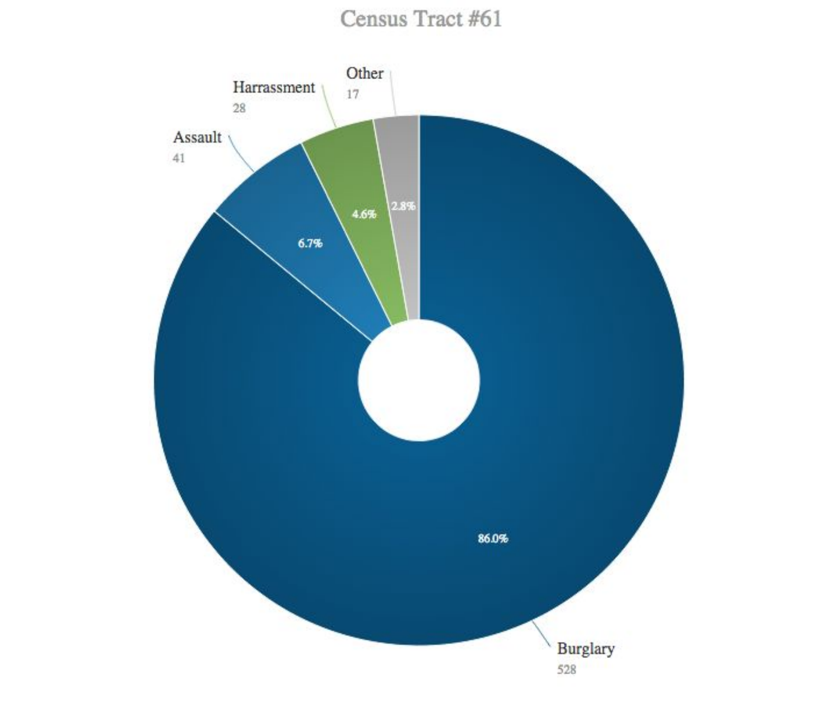
<u>Pie Chart</u>

For further analysis, we decided to take a look at overall crime data (DATA 4) among these central downtown areas (census tracts 61-91). We organized similar types of crime into same category, such as grouping Assault w/ body force and Assault w/ gun into “assault.” However, we decided to separate the two different types of robberies (business and street) since they are completely different in characteristics of what type of crimes they are. Mugging people in the back alleyways in the streets would be an example of street robbery, whereas running into banks and businesses and threatening people on the spot with large guns would be considered business robberies. We chose this approach to better view the ratio of different crimes occurring in Seattle rather than detailed characteristics of each type of crime. These different categories of crime represented in pie charts as shown below:



VIZ 2.1 - Pie Chart of Different Crimes for Census Tract #74 (Central Downtown, Pike Place)

This pie chart shows one of the census tracts in the central downtown area where auto theft was most prevalent. These census tracts in particular have higher rates of assault and robberies than other areas with fewer auto theft incidents.



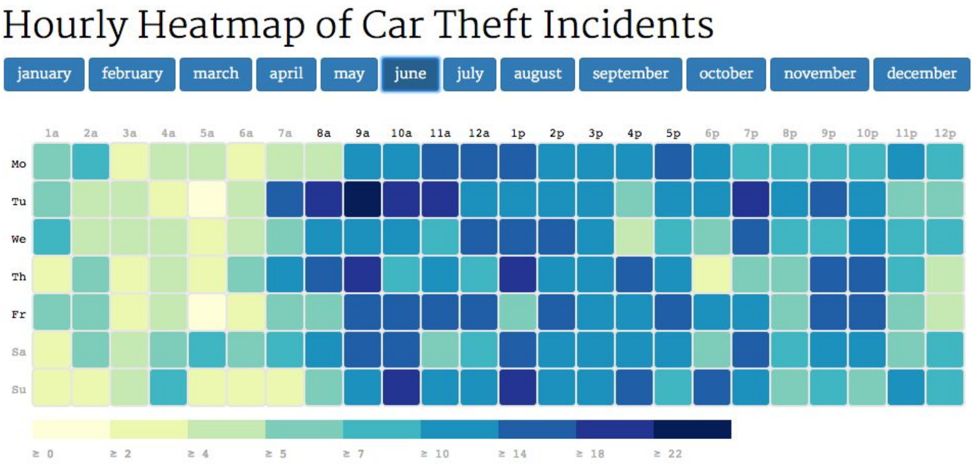
VIZ 2.2 - Pie Chart of Different Crimes for Census Tract #61 (Madison Park)

This pie chart illustrates the different types of crimes that were occurring in areas with fewer auto theft incidents. We observed that these census tracts not only had less crime in general, but also had an overwhelming ratio of burglary over any other type of crime.

Judging from the two different types of pie charts, we could analyze and create a new hypothesis that Seattle areas with higher levels of assault and robbery are likely to have more auto theft incidents than other areas that have simply higher rates of burglary. Even though auto theft is categorized as burglary, the census tracts with low level of auto theft still had higher levels of burglary because they tended to be residential areas with more house break-ins. In these areas, people usually have garages and safer places to park their car, which would explain the less frequent auto theft incidents in these census tracts. We would have liked to have separated the auto theft data and the house break-ins or other types of burglary data, but this was not possible due to the lack of consistency among the provided resources.

The Timemap

To find out hourly frequency of auto thefts, we visualized auto theft data (DATA 1) and plotted frequency of incidents using the D3 visualization tool. The timemap represents frequency of auto thefts per hour and is divided by month and weekday.



X: Time
Y: Day of Week

Darker - more frequent
Lighter - less frequent

VIZ 3 - Pie Chart of auto theft by census tract

We expected that auto thefts would happen mostly at night, but this turned out to not be the case. According to the heatmap, most auto thefts occur during daytime. Attempting to find out why the frequency is higher from 9 am to 5 pm, we examined our data more closely and discovered that timestamps recorded on our data are the reported timestamp and not actual time of the theft occurrence. Because most people are asleep during nighttime and report auto thefts after they find their cars gone in the morning, the timemap visualization displays higher frequency in the span of 7-10 AM. If we had timestamp of actual auto theft crime, we expect that the distribution would be more evenly distributed along the timeline. However, we can still see that there is more auto theft incidents distributed among the span of day time during 11 AM - 5 PM compared to evening time during 6 PM to 12 PM. This allowed us to come to a conclusion that auto theft is more frequent during daytime compared to evening time.

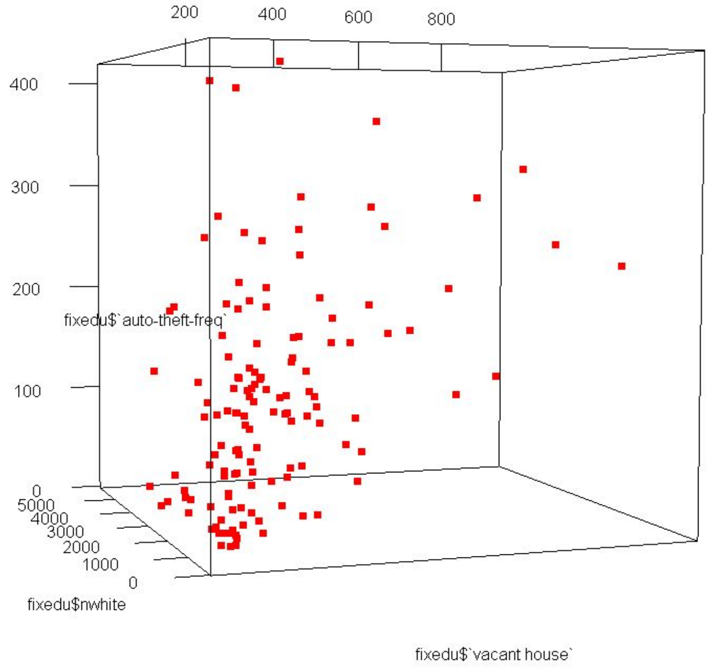
3. Analysis and model

3.1 Attempted method of analysis and result

The first analysis we did is a multivariable linear regression of the relationship between car theft, non-white population, and the number of vacant housing units among census tracts. Using population data (DATA 2) and housing data (DATA 3), we created a

3-dimensional scatter plot and calculated multivariable linear regression among the variables of each census tract.

3-D scatter plot



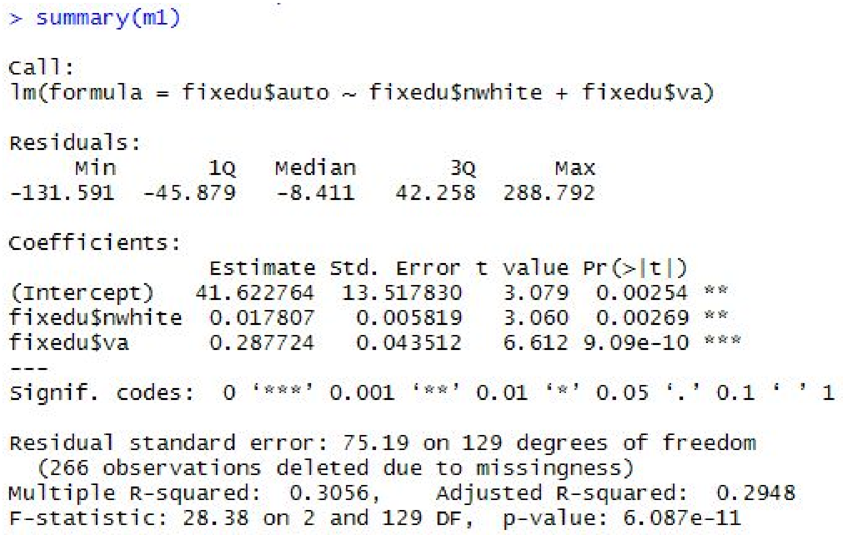
X: non-white people population (each census)
Y: the auto-theft frequency (each census)
Z: vacant housing unit number (each census)

Dot - census tract

VIZ 4 - Multivariable regression

For this we used the R packages lubridate and scatterplot3d, which are used to handle input data and draw interactive 3D plots. By analyzing the scatter plot we can infer that there is some relationship, but to be precise, we used a method within R called the

“Fitting Linear Models” lm(), and the summary is as following:



```
> summary(m1)

Call:
lm(formula = fixedu$auto ~ fixedu$nwhite + fixedu$va)

Residuals:
     Min       1Q   Median       3Q      Max 
-131.591  -45.879   -8.411   42.258  288.792 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   41.622764  13.517830   3.079  0.00254 ** 
fixedu$nwhite  0.017807   0.005819   3.060  0.00269 ** 
fixedu$va      0.287724   0.043512   6.612 9.09e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.19 on 129 degrees of freedom
  (266 observations deleted due to missingness)
Multiple R-squared:  0.3056,	Adjusted R-squared:  0.2948 
F-statistic: 28.38 on 2 and 129 DF,  p-value: 6.087e-11
```

However, the result here shows only a very weak and noisy correlation among auto-theft frequency, non-white population, and vacant housing units. Although we were able to discover that there is some correlation among the variables, the correlation coefficient was **0.305** and therefore would not satisfy our initial purpose for a predictive model.

3.2 final model and interpretation

3.2.1 method

We used the R packages “lubridate,” "forecast," and"dplyr" to handle data.frame and forecasting.

3.2.2 analysis

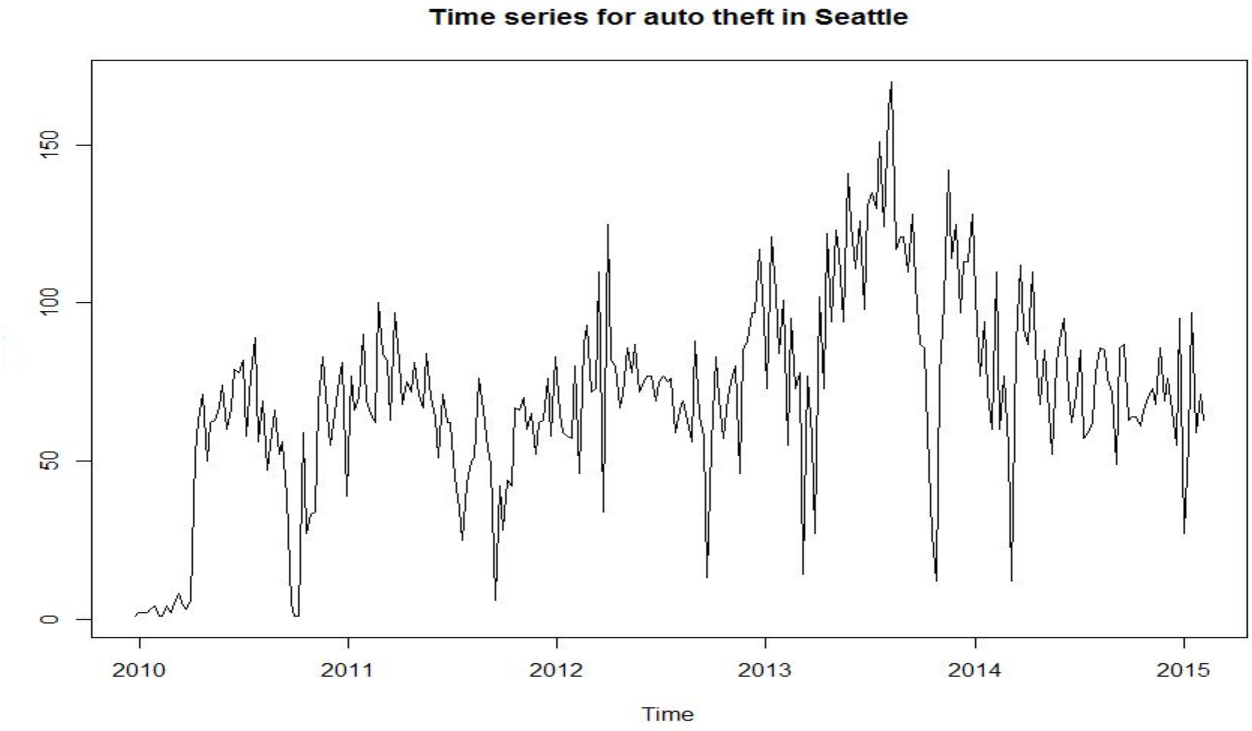
Because the regression predictive model did not satisfy our purpose, we used the time series analysis model.

Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series *forecasting*** is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time” (Imdadullah).

3.2.2.1

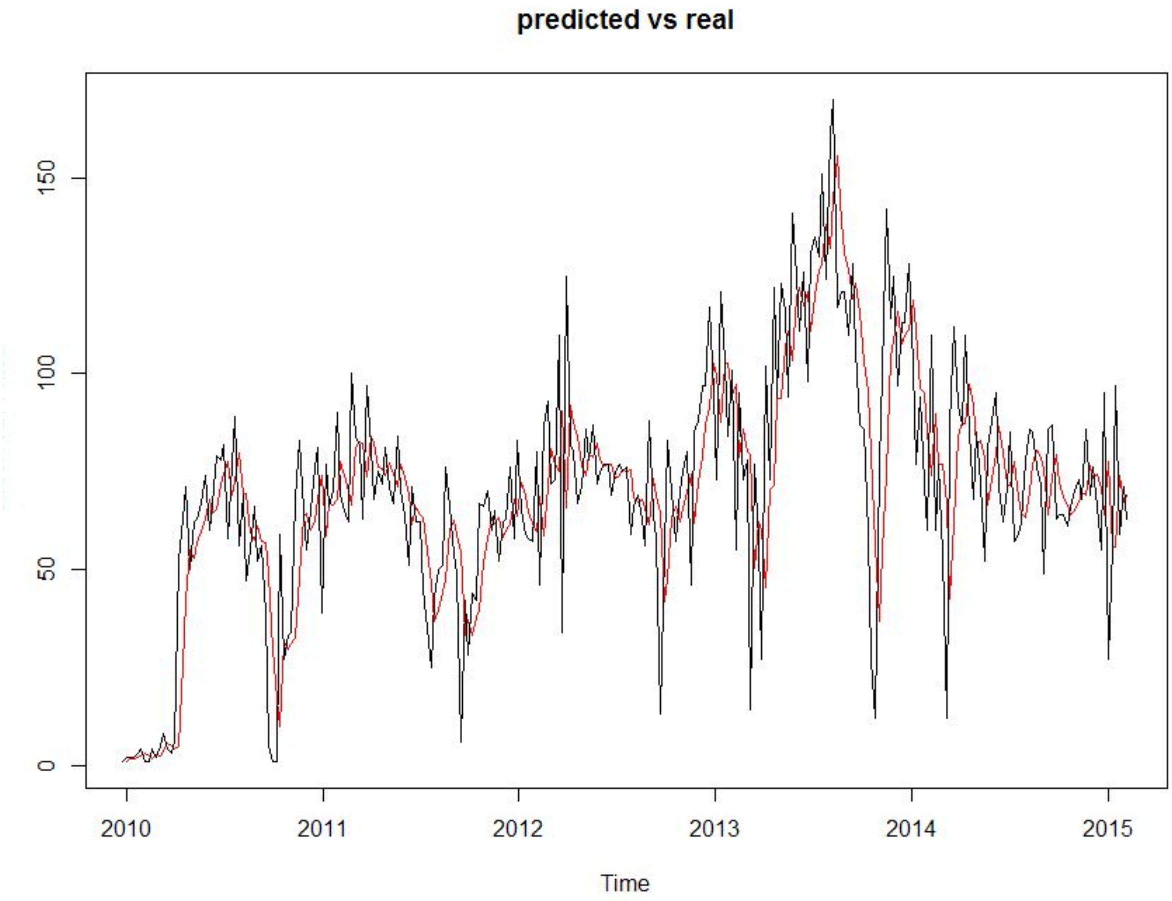
First we mutated the auto theft rate into frequency for Seattle on a weekly basis.
Our original data looks like the following:

year---#of week--- total-auto-theft
2010-----29th week 52
2010-----30th week 63
2010-----31th week 71



3.2.2.2

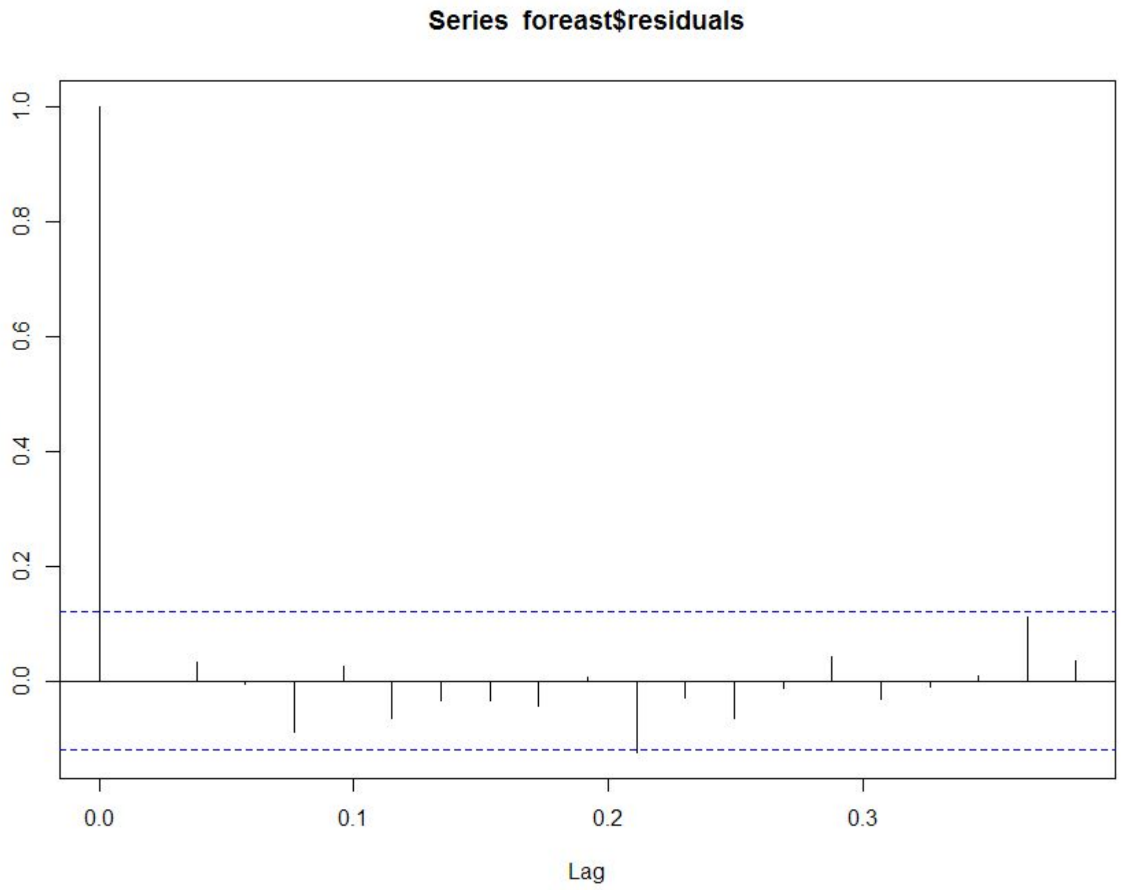
The function we used for forecasting is holt-winter. The Holt-Winters seasonal method comprises the forecast equation and three exponential smoothing equations.
This is our result :
seasonal=Holt Winters(Time1,beta = F,gamma=F,seasonal = 'additive')



3.2.3 goodness of fit and improvement

3.2.3-1 acf

Here we used two functions provided in R to test the effectiveness our time series model:
estimates of the autocovariance or autocorrelation function



Acf is an R function to test any correlation in residual for a time series model.
For a good fitted time series model, it will show randomness for a residual most evenly distributed along the x-axis. Those residual would be mostly within a boundary line, similar to a confidence interval.

Here the forecasting residual was mostly within the boundary (blue dashed line.) therefore we can say that the coefficient and residual was significantly different from zero. From this we can see how well this model is fitted.

3.2.3-2 Box-Ljung test

Box-Ljung test a type of statistical test for whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag.(an element of a time series to produce the previous element.) it tests the "overall" randomness .The Ljung test will run a hypothesis testing where the alternative hypothesis is

H_a: The data is not independently distributed.

Output of the Ljung box test for our time series model:

```
Box.test(foreast$residuals,lag=20,type="Ljung-Box")
data:  foreast$residuals
X-squared = 15.429, df = 20, p-value = 0.7513
```

The p-value is 0.7513, which means in most of the case (ie confidence interval of 90% or 95%). we will able to fail reject the null hypothesis, therefore prove that the data we have is independently distributed.

Conclusions/Interpretations

The most unexpected finding we had was the visualization for auto theft by time of day. At the beginning of the project we expected there would be a significantly higher rate of thefts during night and early morning due to decreased visibility and increased concern of theft from car owners as a result. The visualization shows the frequency of car theft peaks around the hours between 9AM and 5PM for each month. This leads us to conclude that parking during the daytime is not significantly safer than parking during the nighttime. The visualization seems to suggest that parking at nighttime is actually safer, although we cannot conclude that from the visualization alone.

Across all months, there were high numbers of auto thefts during the mornings around 6-9 AM. We interpreted this to be because of auto thefts that occurred during the late night and early morning when most people are asleep. Car owners would not notice their cars were missing until they woke up the next morning, which would explain the high rates of auto thefts from 6-9 AM. This brings up an important point to consider when making our interpretations, the times we have on the data for auto theft may not necessarily reflect reality. We only have access to the reported time of theft, which would be when the victims notified the police. This could possibly be several hours after the actual theft, and we need to be mindful of this when making our interpretations.

We did not find any factors that make parking significantly more or less dangerous like we were expecting. Although auto theft peaked around 9 AM – 5PM, the peak was not significant enough to suggest people not park their cars during these times. There was little difference in rate of car thefts by month as well as day of week, making it unlikely that the rate of auto thefts changes during the year in a predictable pattern.

As expected, auto thefts were concentrated on more populated areas such as Downtown Seattle, Central Seattle, and University District. In contrast, less populated areas of Seattle such as Lakewood and Youngstown had very low rate of car thefts. We interpret this to be because burglars will naturally focus on more populated areas where it is easier to commit crimes. However this is simply an assumption, we cannot conclude anything from the heatmap alone. More research would need to be done to find what causes auto theft rates to be higher in some areas of seattle.

Our findings suggest that auto theft rates are not significantly influenced by factors we analyzed, that there are other factors we did not account for are more significant The Seattle police website supports this, to prevent auto theft they suggest practices such as parking in attended lots and never leaving a spare key in the car. Habits such as these would be difficult to acquire data for, unless a survey was conducted for this purpose. Seattle's high auto theft rate may not be from a high crime rate but possibly from the unavailability of attended lots or unsafe practices from car owners, although more research would be needed to make this claim.

Future Work

We have yet to answer our original research question of “How can we know how safe it is to park our cars at this location?” To answer this we would need to create an algorithm that incorporates the factors we have data for such as time, location, and date, and quantifies the level of risk into a result. From the resulting number we would have to interpret how close it is to average risk, and if we should recommend the user park elsewhere.

To make the algorithm as accurate as possible, we would need to acquire data for more factors. Considering that our visualizations did not find any factors that made parking significantly safer or more dangerous, an algorithm based on the data we have would not give users an accurate idea of how safe their cars are. More factors such as the car model would increase the accuracy of the algorithm, unfortunately not all related data is publicly available or exists at all. Also, even though it isn’t realistic, having accurate data on how many cars are ‘not’ stolen verses how many cars are stolen per census tract would be able to provide us with an extremely accurate model.

To make the algorithm useful and publicly available, we would need to create a website or mobile phone application where users can input their address to get a risk assessment for the safety of parking in that area. It could help users make safer parking decisions by suggesting safer places to park nearby and suggesting good practices.

To serve more people, we could analyze data from beyond the Seattle area. This would increase the usefulness of our application but could prove to be impractical due to the availability of auto theft data for other areas.

References

American Community Survey. United States Census Bureau.
http://www.census.gov/acs/www/data_documentation/custom_tabulation_request_form/geo_def.php

Basic Statistics and Data Analysis. Imdadullah, Muhammad. December 27 2013.
http://itfeature.com/time-series-analysis-and-forecasting/time-series-analysis-forecasting

Seattle City Database
http://data.seattle.gov/