Data Busters: DPD Data Project

Daniel Peck
CSE 5331

Table of Contents

- **Executive Understanding**

Crime exists. We need to look at data to find patterns in this crime. With these patterns, action can be taken.

The data that I am evaluating is the Dallas Police Crime data for 2014/2015. This is data that has been collected for the past year and can be used to make actionable steps for the Dallas PD to take and become more effective at crime stopping. Data Mining can take data and create these steps intelligently, and effectively.

I am analyzing this data using R data analyzing language. I am also using a few different libraries to compute the data and find patterns. I am using Plotly for the graphs and Lubridate to help manage time frames for the project.

- **Business Understanding**

The challenge for this project was to take the Dallas PD's data for 2014/2015 and use data analytics to find innovative/intelligent patterns that the DPD can use to enhance their crime prevention and peace keeping abilities. The DPD needs to find better patterns and correlations for crime, so that it can use that to affect its own patterns for patrol, profiling, amongst other actions. The information could also be used to add more effectiveness to different political campaigns and other demographically focused data.

The measurement of this project would most likely require a medium length term of a changed protocol by the police force, and then doing the same tests again after that period of time to understand if the changes affects were successful. The different data that would be required to measure effectiveness would include all of the previous data rerecorded, and then

The change in protocol would likely not require any substantial funds, so the ability to put into practice wouldn't be financially difficult; that being said, the cost of having misinformation and changing protocol unsuccessfully has very real consequences for real people.

- **Data Understanding**

The data for this project gave us a great deal of variables to consider, however, only those relevant to the correlations found will be included.

Variable Name	Type of Variable	Variable Description
UCROffDesc	Nominal	The type of crime committed: i.e. Murder, Assault, Theft, etc. There are 33 types of crime included in this data set
Zip Code	Nominal	Zip code of reported crime There were 106 included zip codes in the data set
Lat/Lon	Interval	The geographic coordinates of reported crime These
Call Received	Ratio	The time at which the 911 call was received
Call Dispatched	Ratio	The time at which the 911 call was used to dispatch help/police
Status	Nominal	The status of the case, ranging from closed by arrest to open case
Response Time	Interval	The difference between the receiving and dispatching of each call to police This had to be cleaned more specifically, as some of the data reported that there was a negative response time, though probably

		from entered data incorrectly
Reported Date	Ratio	The date that the call was reported. This data was also a bit faulty have 2002 and 2037 entries, again most likely from incorrectly entered data.
CompRace	Nominal	The race of the complainant, which had to be cleaned from having many unknowns, blanks, or "TEST"s
CompSex	Nominal	The gender of the complainant This was a binary of M or F, with a few unknowns

- **Data Quality**

The data quality of the data was actually quite bad in parts. The worst aspect was the amount of duplicate reports there were. There might be an understanding of why these reports are duplicated, be it that the department requires two different reports on each incident, or something of that nature. That being said, since each of the duplicates was exactly the same as the original, with a few having very small differences, it seemed the best way to clean this data was to remove the duplicates with a remove function, which naturally cut the amount of data down by near half, from 236,000 to about 119,000 tuples, using this line of code:

```
data_clean <- subset(data_clean, !duplicated(data_clean[,1]))
```

Outside of that occurrence, the data itself had some holes or peculiarities such as the dispatch time being before the initial reported time, and when doing analysis on those stats it gave funky results that had outliers lying in negative space, so I ran some code to remove all sub-zero differences between dispatch and report times.

Other ways that the data had to be cleaned was with small codes removing non-logical data, such as when periods of times for sequential events was negative. The data for Response Time had a few negative outliers that had to be removed for quality.

- **Simple Statistics**

For some of the most used data of mine, these are some simple statistics for each, with non numeric data having different levels of occurrence:

UCROffDesc	Mode (top 3): Theft, Criminal Mischief/Vandalism, Burglary
Zip Code	Mode (top 3): 75217, 75243, 75228
Status	Mode (top 3): Suspended, Clear by Arrest, Open
Response Time	Min: 0 1st Quarter: 198 Median: 780 Mean: 4860 3rd Quarter: 2997 Max: 125300
Reported Date	Min: 2002-07-27 12:00:00 1st Quarter: 2014-10-12 12:00:00 Median: 2015-02-21 12:00:00 Mean: 2015-02-20 20:14:33 3rd Quarter: 2015-07-03 12:00:00 Max: 2039-10-04 12:00:00
CompRace	Mode: B, W, L
CompSex	Mode: Male, then Female

- **Single Variable Analysis**

These are some minor analyses on the most prominent variables. Most of which are generally histograms detailing frequency, as the most effective way to represent this data is to show where crimes are committed the most, or what type. The single variables alone are often Nominal, and therefore aren't very useful on their own, except to determine their frequency.

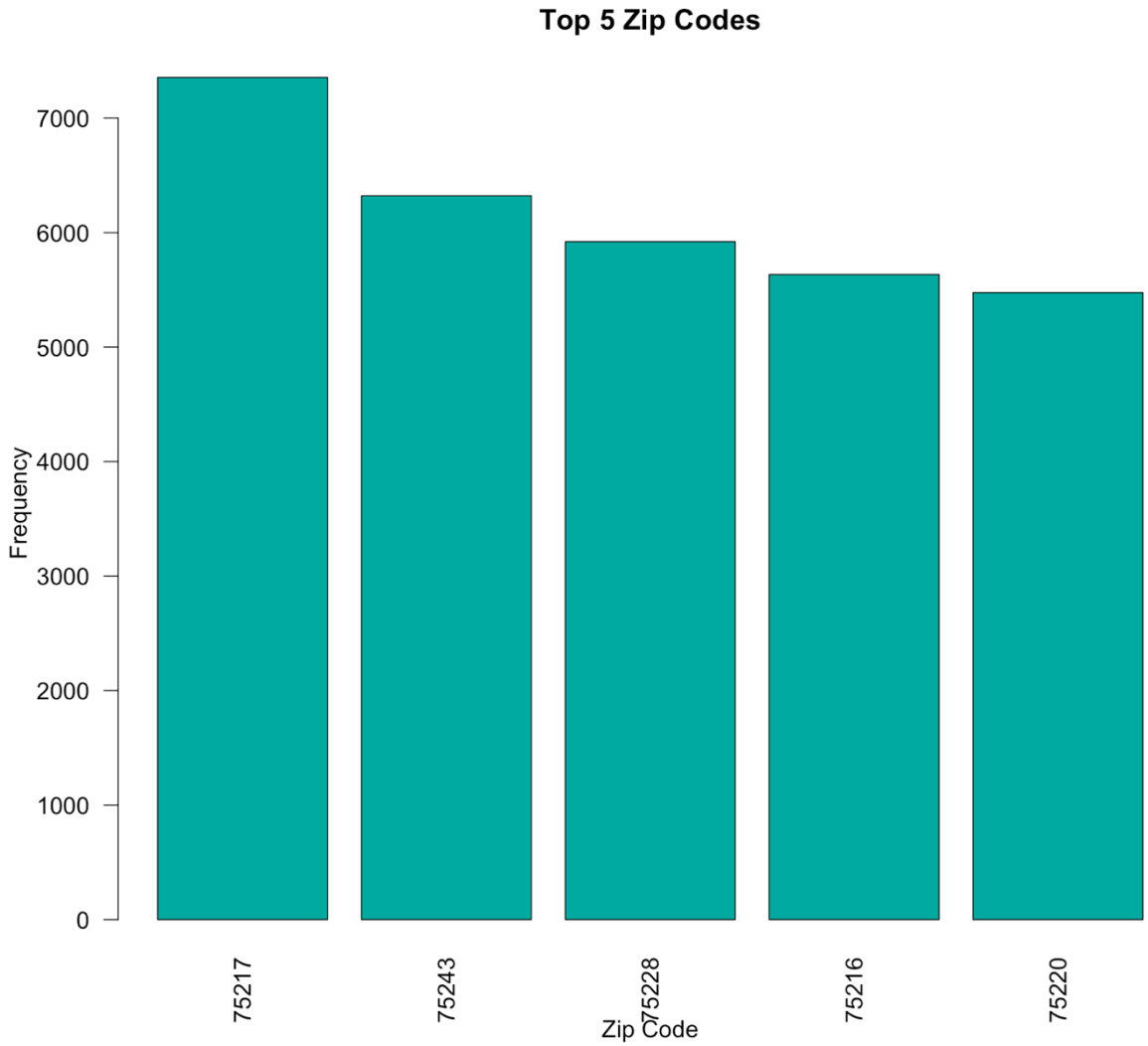

Figure 1

This is the frequency of the 5 most common zip codes. This is better seen in the geolocation map, but perhaps just to briefly mention here that about 10 zip codes or so make up a large majority of the crimes committed.

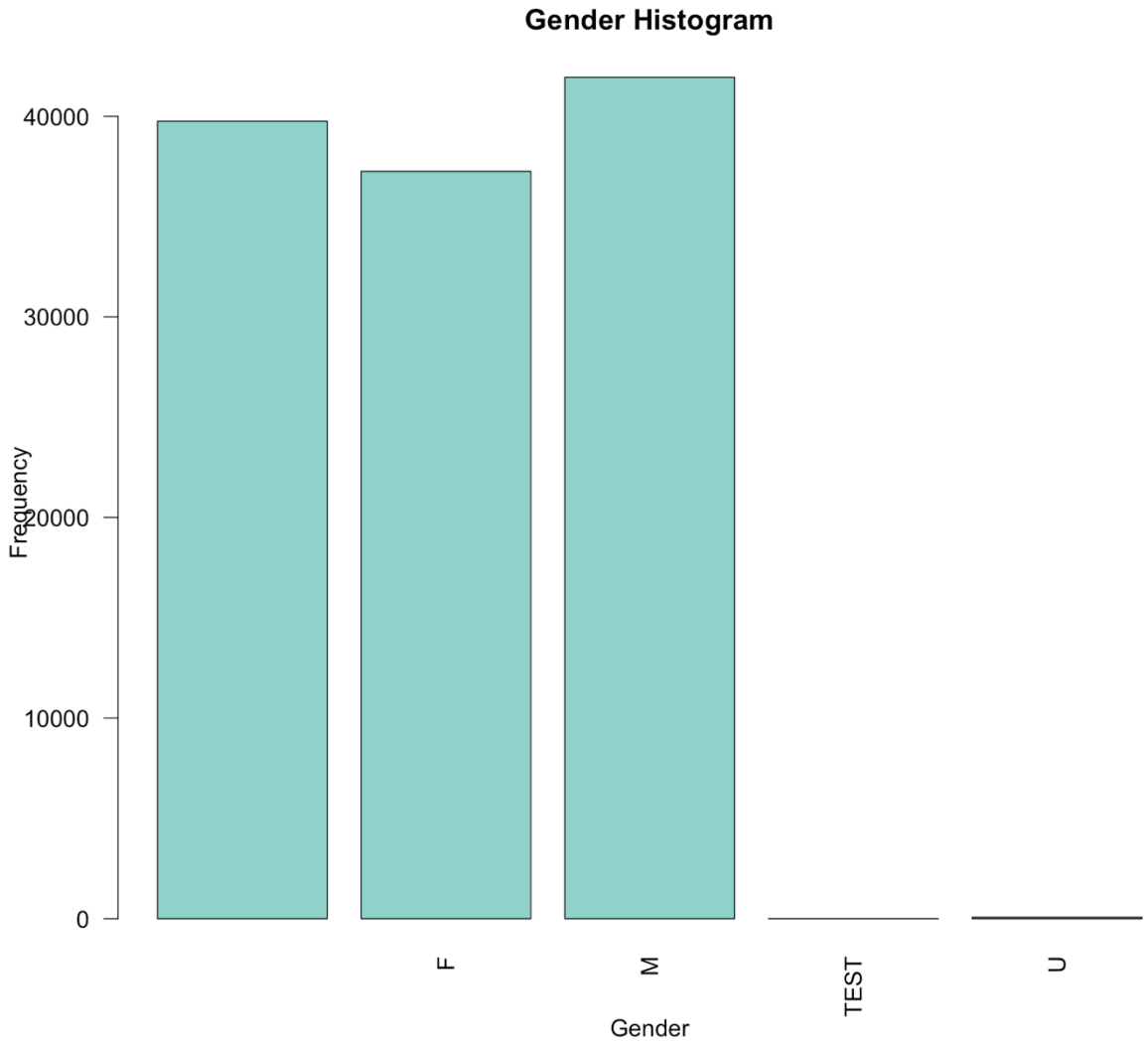

Figure 2

This Graph is rather simplistic as well. It simply shows that most crimes are called in my males and females, with empty rows also being a significant portion. The gender of the caller isn't largely important but can be a factor that would contain useful information. As might be obvious the distribution follows woman almost equal to men, with another 3rd being carried by incorrectly entered or forgotten data.

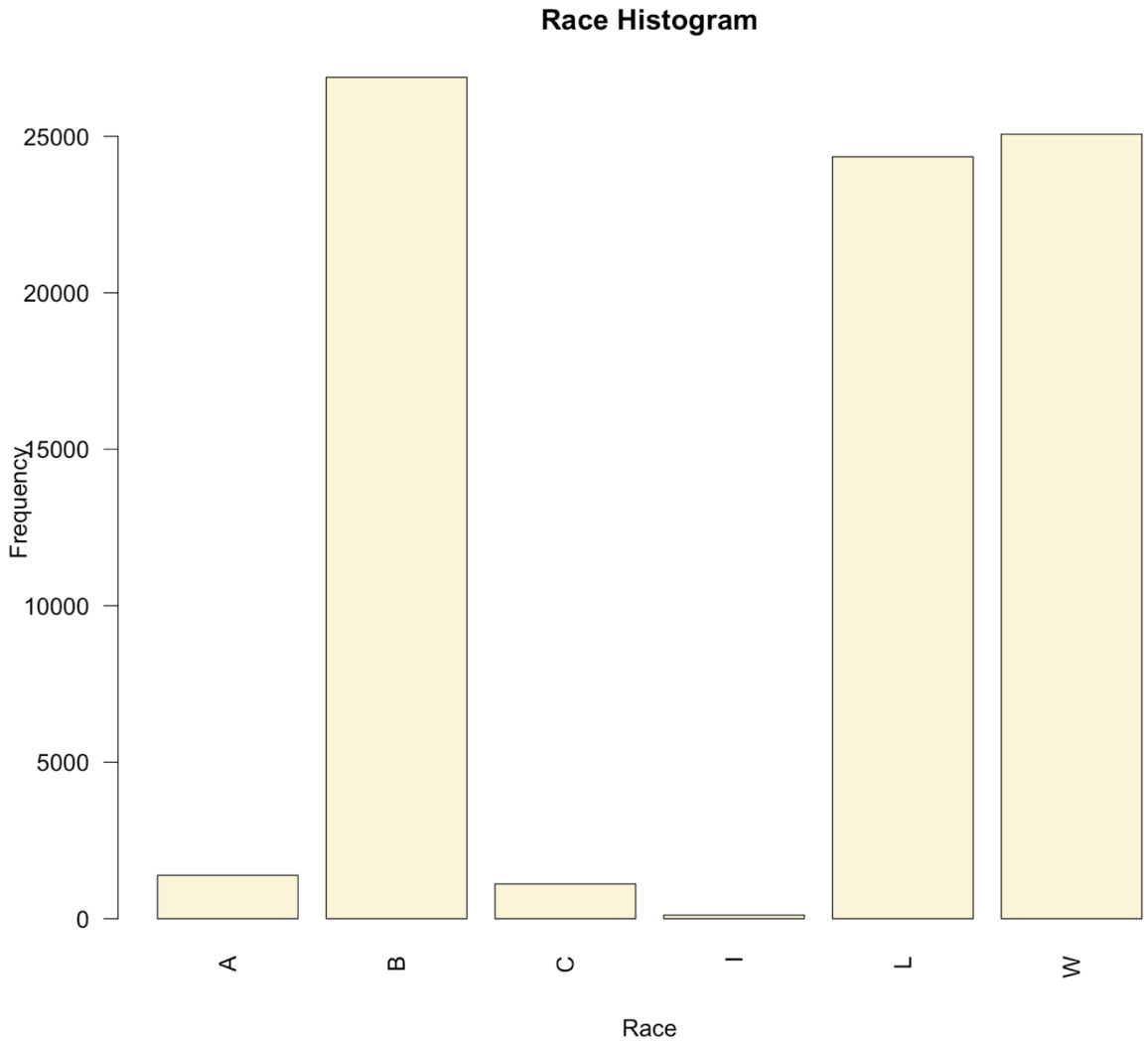

Figure 3

this histogram shows the different races of most complainants. With Texas's distribution of Race, this distribution is also not very surprising, though the extremely sharp decrease from Blacks, Whites, and Latinos to Asians and below could be significant.

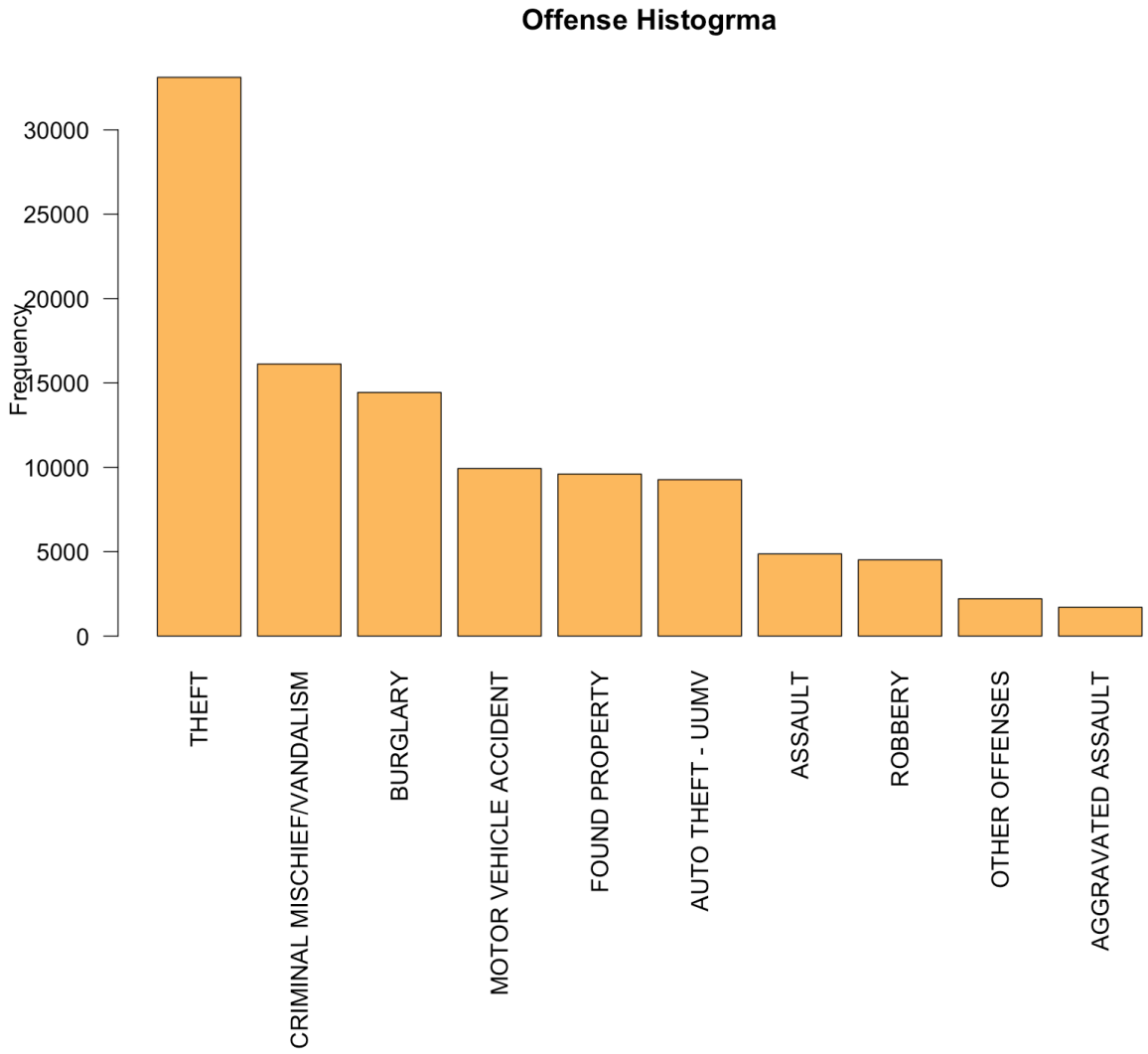

Figure 4

this graph shows the distribution of crime for the top 10 crimes in Dallas. Theft holds more than a quarter of all crime, with other non violent crimes making up more than half. There are some largely committed violent crimes such as assault and robbery, but these are significantly lower than the other less violent crimes.

- **Multi-Variable Relations**

After exploring the implicit data in the single variables, there were multi variable relations that were important to examine. These relationships are often more complex and can reveal interesting and observable patterns in data.

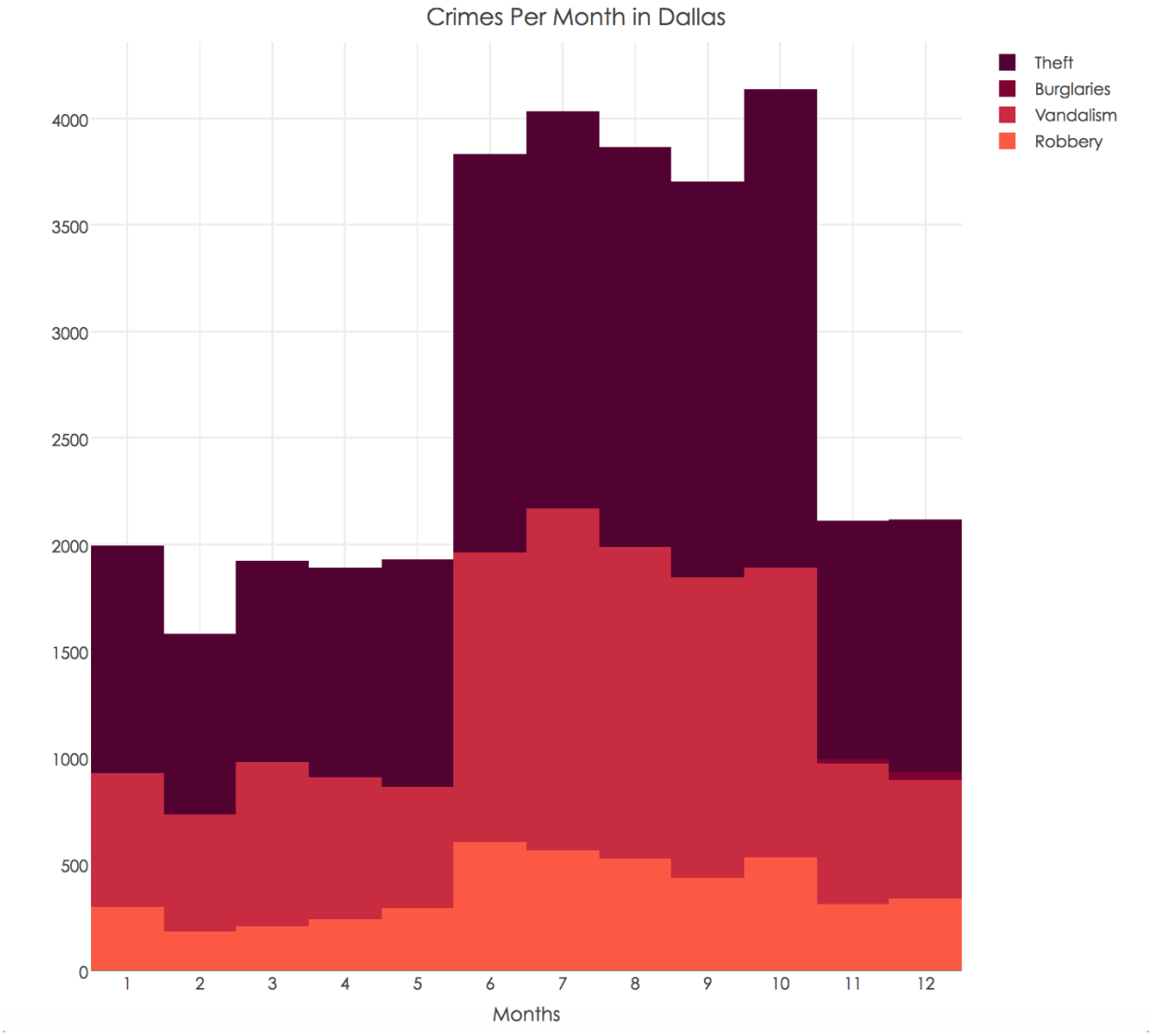

Figure 5

The above graph, figure 5, shows four different forms of crime in Dallas, for each month. This graph also shows another important detail that was interesting; From June until October, there is a very noticeable spike in criminal activity. This is across the 4 crimes observed, and shows that the crime increases uniformly. This information can extrapolate to different things, however without more data, we can only safely say there is a spike. Hypotheses might lead to a corollary with school being in session.

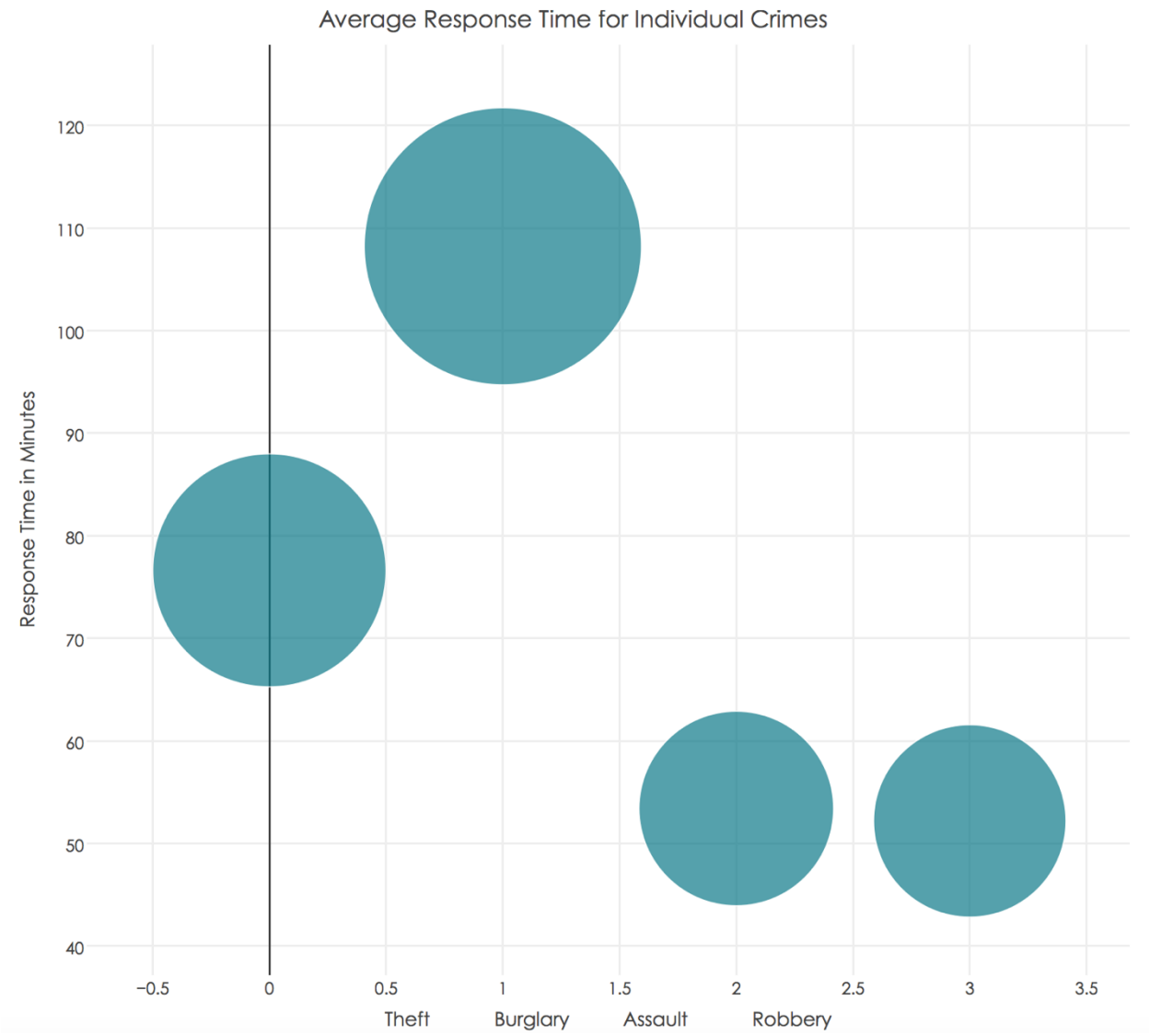

Figure 6

The above graph, figure 6, shows the various response time to different crimes on average. The size of the ball represents the standard deviation of the same crime. The response time for burglary in particular is very high, which is contrasted with both theft and robbery. Maybe the longer response times have to do with the lack of urgency that responding to burglary involves, as opposed to the immediate nature of assault and robbery.

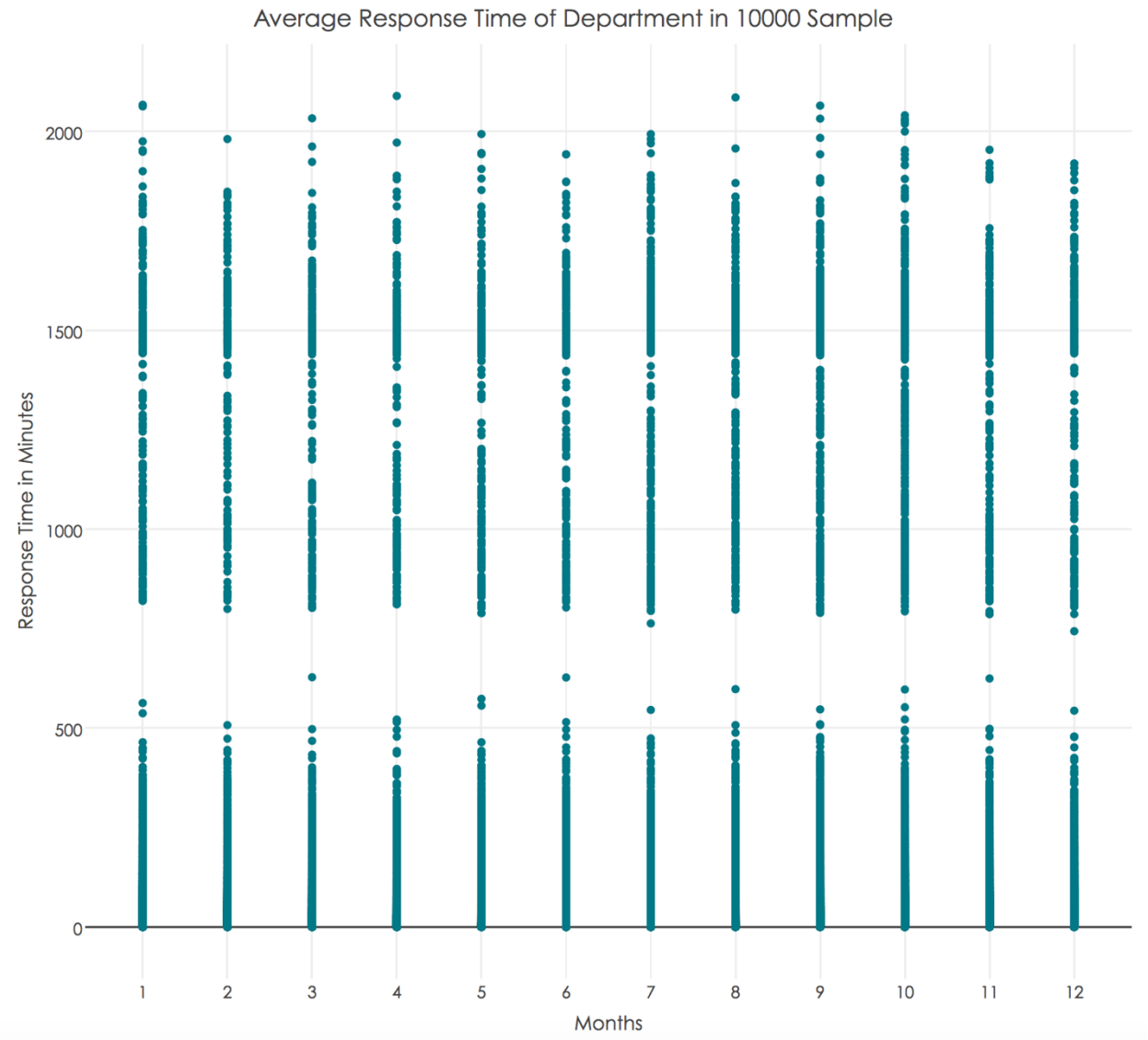

Figure 7

Figure 7 shows the response time for each month. While at first this graph shows no good information for its intended purpose, it does have a strange statistical blank in the middle, which could have to do with the police departments response ability. If they cant respond immediately it could mean that they will put if off for a regular amount of time until they can address it further. The months of the year, however, don't appear to add much value to the graph.

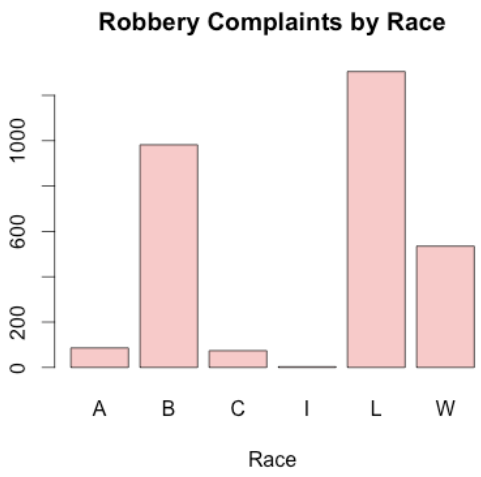

Figure 8

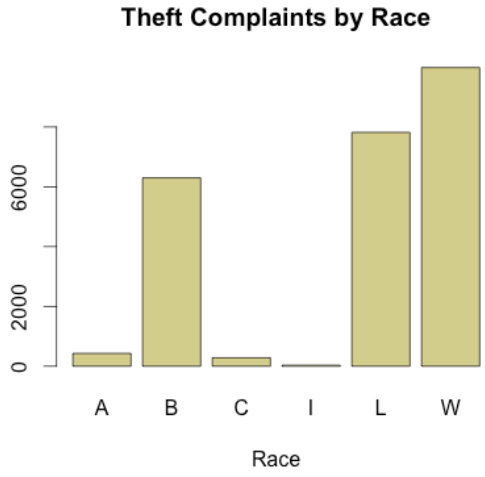

Figure 9

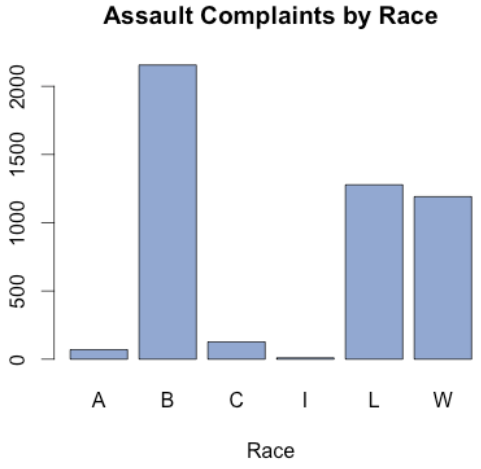

Figure 10

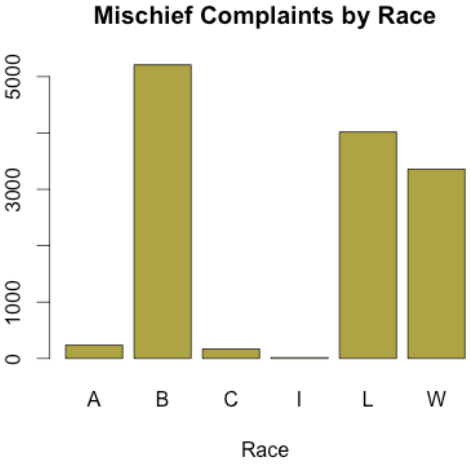

Figure 11

These graphs show the number of complaints made by each race for some particular crimes. This could mean a few things depending on the location as well. For example, if Robbery is higher amongst the Latino Population as show in figure 8, then the police department could look at methods on how to prevent robbery in those areas. These could mean spikes of crimes in certain racial populations, however it is unknown whether or not that these are valid or just statistically insignificant.

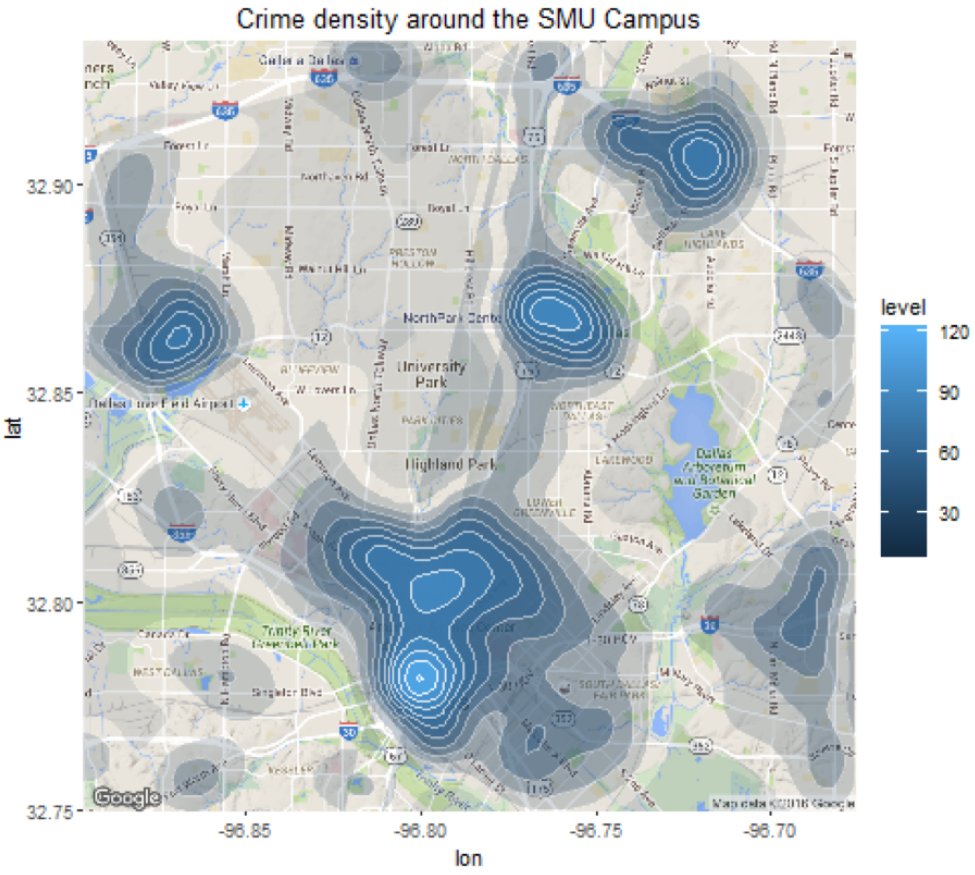

Figure 12

The above map, figure 12, is a density map of the crime in Dallas. This map shows that the crime is generally centered around more low income neighborhoods, and is rather low for the higher end neighborhoods. This could be that the police force have a bias to be patrolling the nicer neighborhoods, and aren't being as present in the lower economic neighborhoods. More information on the Police department's patrol patterns could be useful in determining if they are present enough in high density crime areas.

The following graphs, however, show a more uniform dispersal of burglary, theft, and assault around SMU campus, sans the Highland Park area, which has a different police force. This could mean that the higher density areas have more diverse crimes occurring which add to their density patterns.

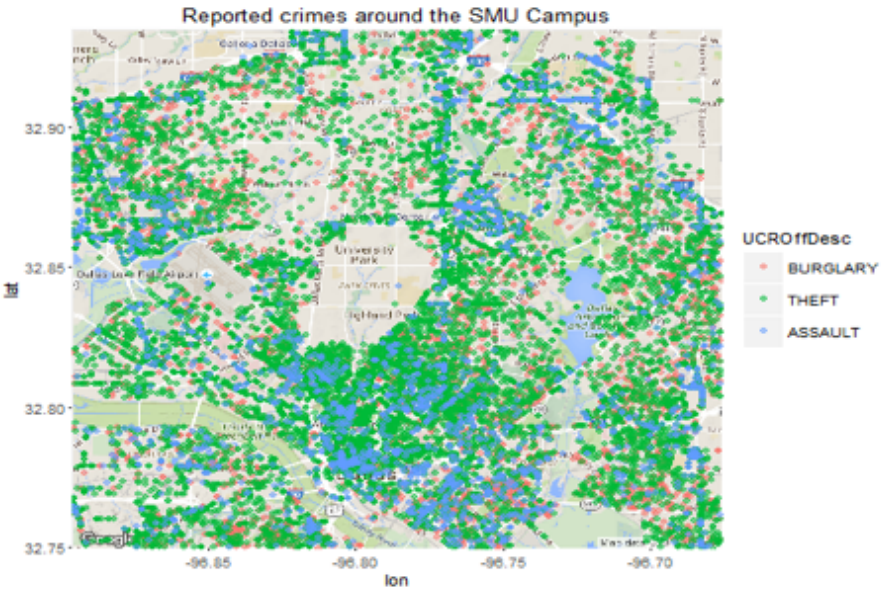

Figure 13

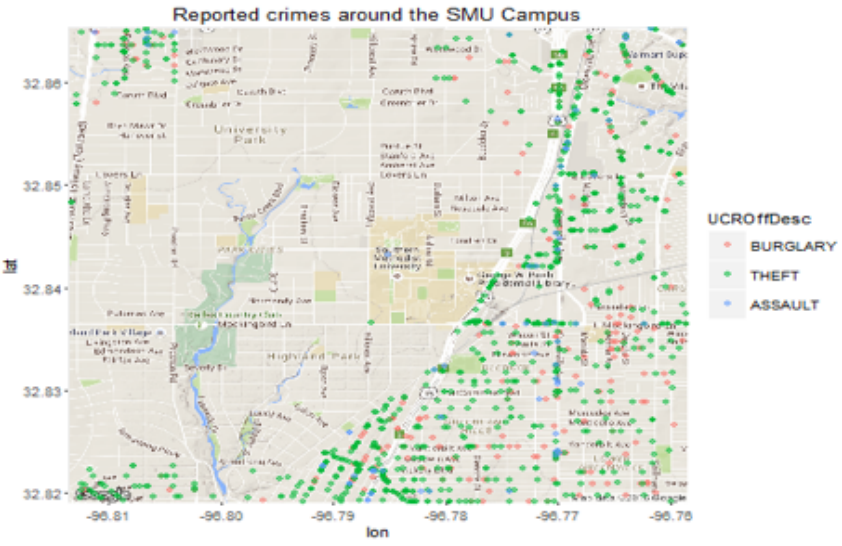

Figure 14

- **Conclusions**

This data has left a few key findings that should be noted. First it is important to note that the crime increases rather significantly in the summer and early fall months. I think that to continue this project that is one of the key findings I would want to focus on. Further data would be needed to test hypotheses, and it would be helpful to see the ages of the offenders, to observe if it has a correlation with school or other summer activities.

The other data that I would focus on in the future would be the crime density map, and the fact that low income neighborhoods have as much crime. While more data would be preferred to see if the police are effectively patrolling/crime-stopping in the area, the actions that can be made from this data are the same, in that the police should be putting more people into those areas to prevent further crime.