Visual Analytics of D.C. Crime

Saketh Chintapalli
112686022

Viraj Kamat
112818603

Introduction:

Crimes are a very serious matter that afflicts society on a day to day basis. Police and other law enforcement agencies have today turned towards science to better explain crime and possibly prevent it, and hence a lot of data is collected on crime from different cities across the USA and we wish to use these datasets along with Visual Data Analytics to help explain criminal activities in an area.

Glancing through our dataset we observed that crimes in areas of a city were indeed associated with external factors. Some of these factors include time of day and location of a crime, but other factors such as unemployment and overall quality of life in a region definitely contributed to a crime. We have these datasets from the D.C. metropolitan area.

Using Visual Data Analytics we wish to highlight how external factors would better explain the crime in an area. Using scatterplots, bar-charts and histogram we wish to highlight those factors that would indicate a strong correlation between a crime in an area and other external factors such as time-of-day, nature of crime, unemployment in a region, etc. The key aspect here is to leverage visual charts to amplify these contributing features.

Data:

The dataset we used to do our analyses is the crime data reported in the Washington D.C. area ranging from 2008 to 2017. The dataset contains fields regarding the various elements to the crimes committed, like the type of offense, date/time, location of incident, type of crime(violent/non-violent), etc.

We explored approximately 10,000 data points to get an initial look at the data, we used the pandas profiling module and the output is as follows :

Dataset statistics

Number of variables	33
Number of observations	10000
Missing cells	557
Missing cells (%)	0.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	11.8 MiB
Average record size in memory	1.2 KiB

The following are the important fields D.C Metro Area crime stats dataset :
Offense/method - Type of offense committed
Time - Timing of the offense
Block/District/Ward - Area of the offense

A snippet of the dataset for the D.C. area can be seen below:

REPORT_DAT	SHIFT	OFFENSE	METHOD	BLOCK	DISTRICT
9/17/2012 4:40:00 PM	EVENING	THEFT/OTHER	OTHERS	1300 - 1399 BLOCK OF 49TH STREET NE	6.0
8/6/2013 7:38:00 AM	DAY	THEFT F/AUTO	OTHERS	800 - 899 BLOCK OF 20TH STREET NE	5.0
5/26/2014 10:05:00 AM	DAY	ROBBERY	OTHERS	5000 - 5069 BLOCK OF BENNING ROAD SE	6.0
7/16/2015 12:16:00 AM	MIDNIGHT	ROBBERY	GUN	2700 - 2799 BLOCK OF 13TH STREET NE	5.0
8/18/2017 10:04:22 PM	EVENING	THEFT F/AUTO	OTHERS	1 - 60 BLOCK OF K STREET NW	1.0
9/15/2014 2:53:00 PM	DAY	THEFT F/AUTO	OTHERS	364 - 399 BLOCK OF MASSACHUSETTS AVENUE NW	1.0
8/7/2016 1:31:40 PM	DAY	THEFT F/AUTO	OTHERS	5100 - 5199 BLOCK OF JAY STREET NE	6.0
9/12/2009 1:50:00 PM	DAY	THEFT/OTHER	OTHERS	100 - 199 BLOCK OF DIVISION AVENUE NE	6.0
1/29/2009 9:15:00 PM	EVENING	ASSAULT W/DANGEROUS WEAPON	KNIFE	1730 - 1797 BLOCK OF LANIER PLACE NW	3.0
6/6/2011 3:45:00 PM	EVENING	BURGLARY	OTHERS	400 - 499 BLOCK OF JEFFERSON STREET NW	4.0

Data Preprocessing:

Data preprocessing was done with the Python pandas module and was used extensively when serving the data from our backend server to the front end web application to be rendered by d3js.

Our initial step was to clean the dataset and remove any outliers and imputing any missing values. To impute missing values in our dataset we randomly sampled values from the column of the dataset to fill in those that were absent. Next we went on to remove a few columns in our dataset that would not yield any important results in our analysis such as *start_date*, *end_date*, *census_tract* and *psa* .

Since the dataset was very large we decided to perform our data analysis for a single year, which in our case was 2017. To include some other exogenous factors that would help in our analysis we included an external dataset of unemployment rate in a particular ward and using the Pandas dataframe merge function we were able to merge the DC Metro area crime stats dataset with the unemployment rate dataset on the ward column. This would help us then build correlation between the unemployment rate and crime in a particular ward.

Lastly since we had categorical variables, and we needed to perform a scatterplot of the top two PCA's which requires only numerical values we used the Pandas factorize function to convert categorical variables such as shift of day, crime type, offense type to numerical variables where needed.

This dataset was stored in a global variable in our code that could be accessed by other server functions that would process data required for frontend rendering for visual analysis.

Methods:

For our analyses, we decided to create an interactive dashboard with a quad partitioned layout to be able to make apt inferences. In particular, we decided to analyze the latest available crime data for the D.C. area which is 2017. We wished to pinpoint high-risk areas in D.C. with respect to various types of crimes and other factors such as time of day and the particular methods of the crime. In addition, we wished to highlight a possible correlation between certain types of crimes and the unemployment rates of certain wards in the D.C. area. Washington, D.C. is geographically divided into 8 wards, and we have statistics regarding the unemployment rate in each ward.

An example snippet of our dashboard can be seen below:

As we can see from the dashboard, the top left represents a heatmap of all the criminal activity occurring in specific locations in the D.C. metro area. On the top right, we have bar charts with respect to the various factors that affected criminal activity. The user has the option to choose between these specific factors which we deemed to be the most influential for crimes: Offense, Method, Crime-type, Quad, Shift. On the bottom left, we have a bar chart with respect to the distribution of crimes amongst the 8 wards in D.C. Our idea was to understand which wards are more prone to certain crimes based on filtering the factors on the top right. On the bottom right, we can see a scatter plot of the top two principal components from this dataset. The colors of the points in the scatter plot are also with respect to

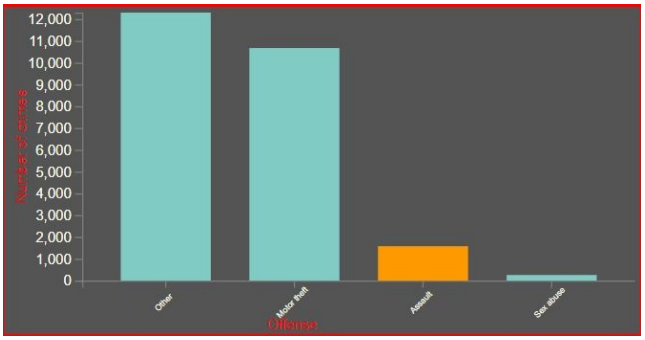
the wards which can be matched from the bar chart. For functionality in the dashboard, we are able to select certain sub-categories within the factors, and the rest of the plots will be updated accordingly. Based on the selected filters, we can see the rest of the charts being updated accordingly. As such, we will be able to identify various correlations between sub-categories of factors and criminal activity based on ward/geographic area.

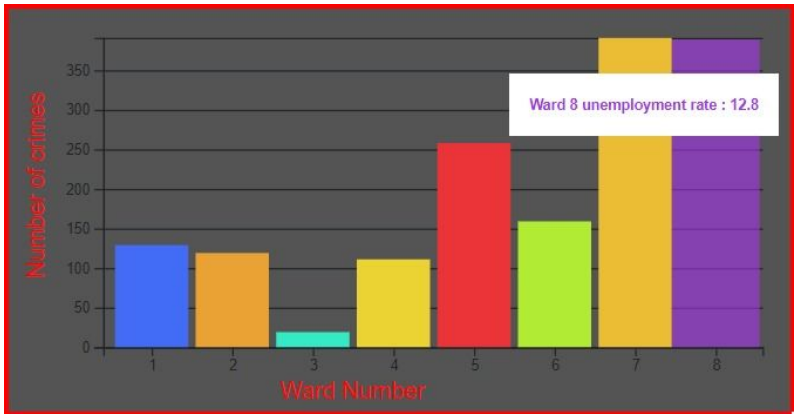
Results:

In order to draw conclusions from our plots we can select different factors that contribute to crime in our bar chart of factors contributing to crime, then subfactors and notice how the bar chart of Crimes per ward along with the heatmap of crimes in the Washington metro D.C Area updates.

For the Assault type of Crime factor, we wanted to see if there were certain wards which reported more assaults than others and the relation with respect to unemployment rate within the wards. You can see our observations below:

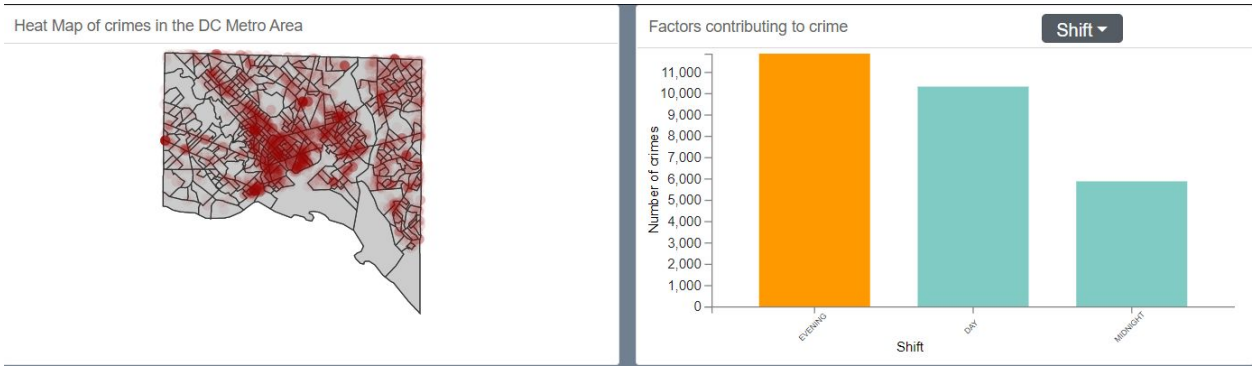
As you can see, the assault type crime has been selected and is highlighted in orange. Accordingly, the map on the left gets updated to reveal the density of assault crimes throughout the D.C. area.





The crimes-by-ward bar chart above was also updated according to assault crimes. We can now see the distribution of assault crimes amongst the different wards in D.C. Right away, we can observe that ward 8 had the highest number of assault crimes in D.C. In addition, it was particularly interesting to note that this ward also had the highest unemployment rate amongst all the words. This led to your inference of a possible positive correlation between unemployment rate and assault crimes in Washington, D.C.

Another interesting observation that was made is that most crimes occurred in the evening at the heart of the DC Metro Area. We can conclude this by looking at the following visualizations of our graph



As you can see from the heatmap of crimes in the D.C Metro area, most crimes occurred in the evening also corroborated by the bar chart and at the center of the D.C Metro area. One reason this could be the case is that either people return back to their homes from their jobs or there is a lot of movement of people that causes more interaction and hence more crime.

Conclusion :

Using the d3js javascript library to render in a browser the contents of a dataset containing the crimes occurring in the D.C area along with information of the unemployment rate in each ward we could make interesting visual analysis of crimes in the D.C Metro Area. We explored different factors that contribute to crime and how we can visually plot this on the geographical map of the D.C Metro Area to better aid in our analysis.