CSE 564 Visualization Project Proposal

VISUAL ANALYTICS OF CRIME WITHIN CITY

Viraj Kamat (SBU-ID112818603) Saketh Chintapalli (SBU-ID112686022)

Objective

We have datasets from multiple locations pertaining to crime, these datasets contain information on the crime, the location of the crime, the type of crime and the time. Given this information we wish to draw a correlation between crimes occurring in an area to the external factors which would help explain how crimes are spread out in an area.

We wish to use Visual Data analytics to chart our observations; implemented with the help of software libraries such as d3.js Javascript module and Pandas Python library.

Motivation & Intuition of our project

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 cities such as Chicago and D.C Metro area but we wish to use crime datasets from other cities as well

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.

Dataset

We sourced our dataset from Kaggle which included crime statistics for the D.C Metro area in Washington and crime statistics for New York. As for external data, we incorporated data for unemployment rates from the official websites of the Department of Labor and the Department of employment services for the City of New York and D.C Metro Area respectively.

An initial dataset that we are using 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 data appears to be fairly clean, with some values missing in some cells. We would be taking measures to impute missing values.

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

We have other fields as well that would contribute to evaluating crimes in a region, however for now we have an intuition that these fields would contribute largely to our analysis.

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

The dataset procured is clean and well documented so as to make it effective for us to perform quick analyses across the various data attributes. Note that we have selected the important variables only in this snippet as we have more than 30 variables in total.

Exploratory Analysis

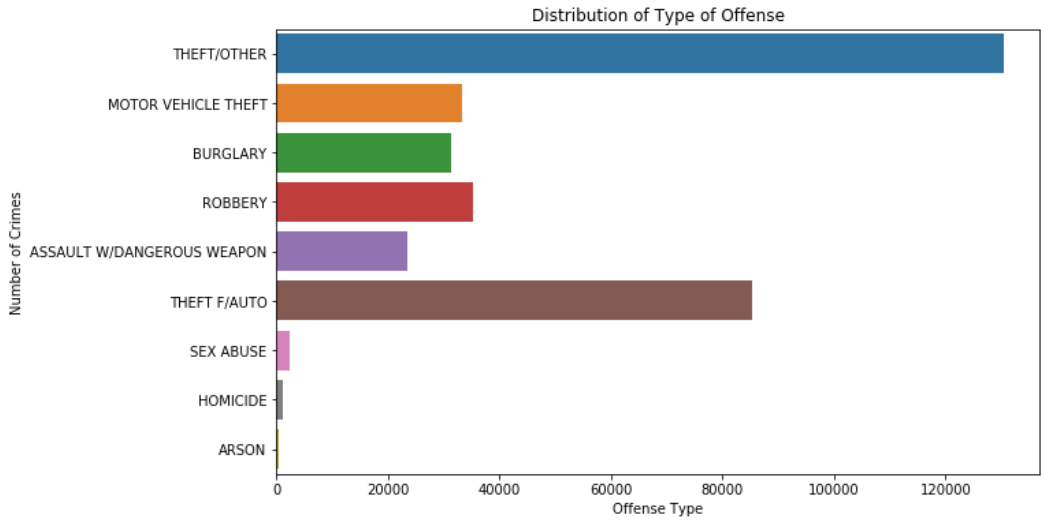


Fig 1: Bar-chart of crime based on type of crime

As we can see, theft crimes were the most frequently occuring in the D.C. area for the given time frame we chose to analyze. In addition, burglary and robbery were also significant in number during the same time frame. We then went on to see which hour of day reported the most criminal activity in D.C.

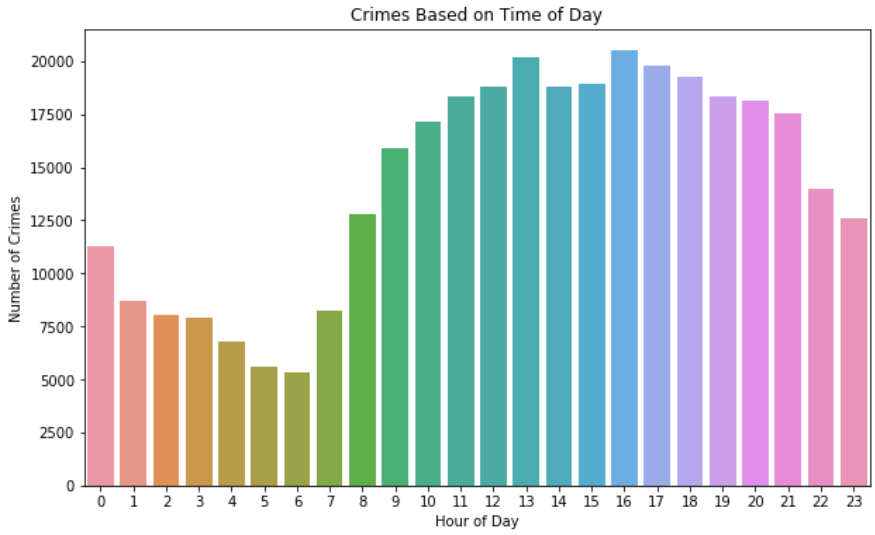


Fig 2 : Bar-chart of crime against hour of the day

Based on the bar plot above, we can see that D.C. reported a peak in its crimes usually around rush hour (4-7 PM).

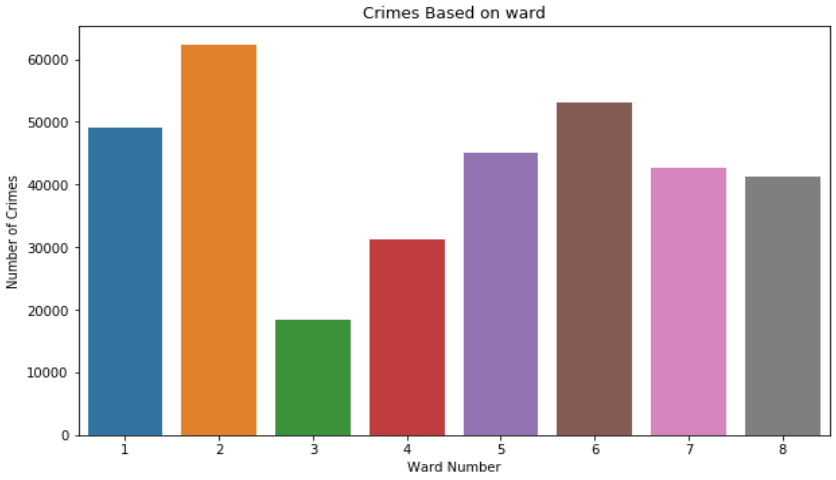


Fig 3: Bar-chart of crime based on ward

The regions in the D.C. area are divided into “Wards” with there being 8 wards in the whole area. The bar plot above shows the specific wards and their corresponding crime rates between 2008 and 2017. We went on to see that Ward 2 reported the most crimes and Ward 3 reported the least crimes in the area.

Furthermore, we decided to explore a possible correlation between unemployment rate in the wards with the crime rates. We decided to use Ward 8 for our initial analysis as it had reported the highest unemployment rates amongst all the wards. The line plots below show us the breakdown of unemployment rate and crime rate over the timeline for Ward 8.

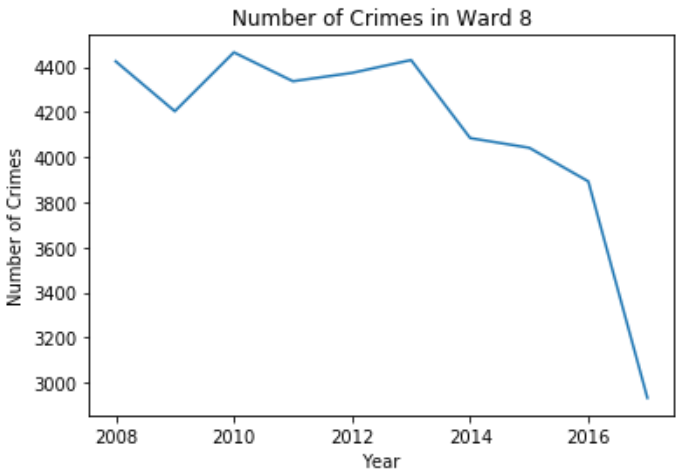


Fig 4: Line-chart of crime based on year

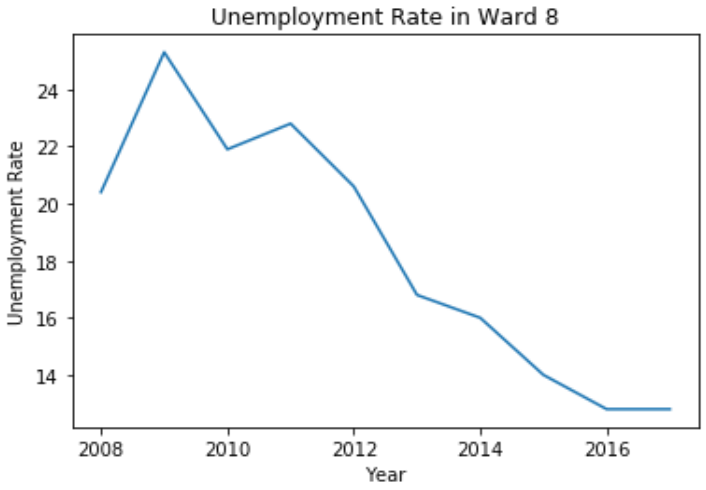
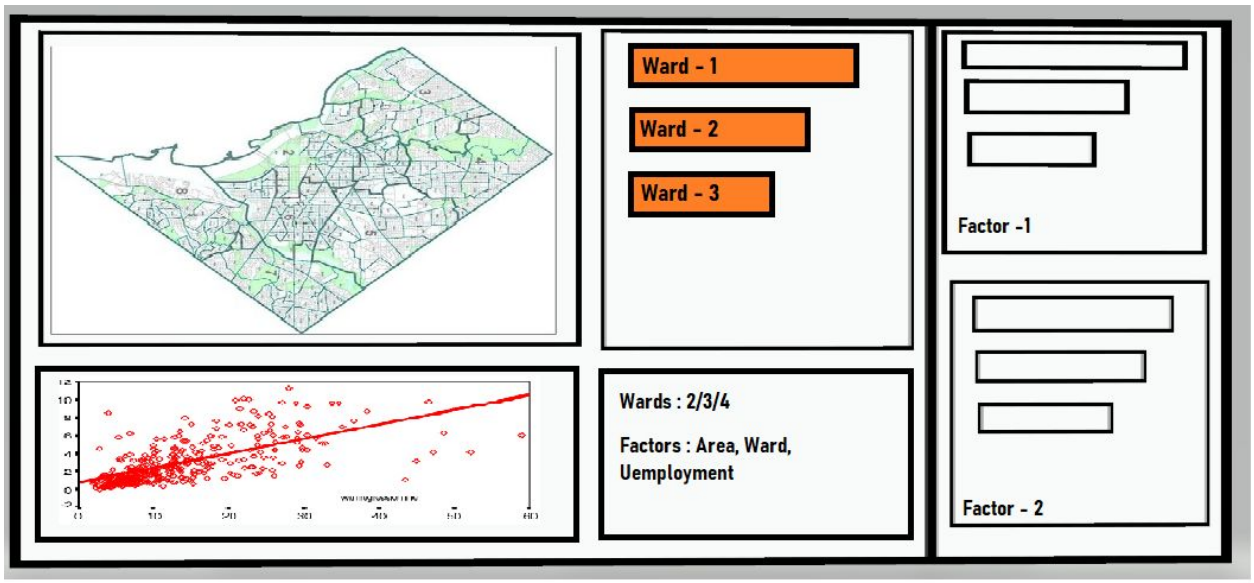


Fig 5: Line-chart of unemployment by year

We can notice a similar trend in the decrease of crime rate along with the unemployment rate from the line plots above.

Layout & Approach

We wish to present our visualization in the form of a dashboard. The diagram below will give a brief overview of the layout of the dashboard



In the diagram above on the left pane we have the bar-chart that displays the number of crimes in each region along with a map showing crimes in a region. In the bottom left pane of the dashboard we have the scatterplots, namely of the variables that contribute to crime to amplify those features that contribute to a crime, such as type of offense, hour of day, and unemployment rate in the region.

We would have two scatter plots, namely MDS and PCA, the MDS that shows the distribution of crime across a region visually and the PCA plot that helps provide correlation between important factors (attributes) that contribute to a crime.

Note that the above dashboard will be interactive. On clicking any region in the bar-chart in the right pane would result in filtering the scatterplots, MDS/PCA plots to show data of that region only. Similarly in the right pane we would have the capability to filter out/ narrow down variables whose bar-chart scatter-plot is being displayed. We would also have the capability to select multiple variables at a time to further refine the plots in the left pane.

All the data will be rendered with the help of d3js, REST API calls would be made to the backend which would fetch data. Majority of processing of data however would happen in a backend-end server implemented with the help of Python's flask module which would return JSON based data to the front end to be rendered.

Challenges

Since we have a lot of data on our hands and a lot of plots to be developed to amplify features that could explain crime in a given region, we are bound to come across some challenges. Some of them are listed below :

1. We need to clean the datasets where variables are empty or introduce mean/average values in their place.

2. We need multiple datasets, at the moment we have crime rates and unemployment rates in a region, we would also like to combine other values such as economic conditions, living conditions, etc to better show the correlation between crime and other exogenous factors.

3. The backend servers need to work efficiently to provide requested data to the frontend i.e efficient code must be written in the Python backend servers that will server data requests with minimal time delay given that a lot of data is being processed.

4. What other interesting graphs can we introduce? The generic scatterplots/bar-charts are fine but we also need to research different approaches that eases visual analysis of the data.

5. Datasets - We have data for New York & the DC Metro area, though we also wish to add more cities to our visual analysis.

Next Steps

1. We will start to lay out all of your visual analyses onto the dashboard to create a comprehensive view of the crimes occurring in the specific wards for the D.C. area.

2. We also plan on obtaining more data on other potential factors which may have affected the crime rate to be higher or lower in certain areas as opposed to others. These may be, but are not limited to living conditions, education amongst the population, poverty rate, etc.

3. Once we obtain this data, we plan on fusing it with our original dataset on crimes to notice any significant correlations with the criminal activity amongst the wards. In addition, we plan to use MDS and PCA to see which factors affect different categories of crime the most and which components can be used to explain the variance in the dataset.

4. We plan to use brushing with the dataset specifically to analyze the subcategories of wards in D.C. in addition to the subcategories of types of offenses reported in these wards. Another idea to use brushing is to visualize the different hours in a day, to see what crimes were most frequently observed during specific times of the day. One hypothesis that could be confirmed with this visual analysis is to see whether or not assault crimes were more frequently observed during the odd hours of the day(maybe 12-5am).

5. Overall, we hope to gather interesting insights about the criminal activity in D.C. and possibly try and relate these trends with those of a nearby city such as NYC. If time permits, we would like to include another major city to see if these trends match up to those found in the D.C. area.

References

[1] https://www.kaggle.com/rblcoder/mental-health-happiness-economics-human-freedom/data

[2] https://www.kaggle.com/vinchinzu/dc-metro-crime-data

[3] https://does.dc.gov/page/unemployment-data-dc-wards

[4] https://www.citylab.com/life/2013/09/puzzling-relationship-between-crime-and-economy/6982

[5] https://www.kaggle.com/adamschroeder/crimes-new-york-city

[6] https://labor.ny.gov/stats/lslaus.shtm