CSE-564: Visualization

FINAL PROJECT REPORT

Analysis and Visualization of Crime Data in United States

ANKIT AGRAHARI (110946823)
TARUN LOHANI (110921666)

PROJECT PROPOSAL

Problem Statement:

The newly elected government wants to eradicate various crimes from US has instructed the federal and state police to take preventive measures. The police department want to appoint their force based on the number of criminal cases as well as based on the severity of the crime. They want to see the recent trends and patterns through visualization in the crime across the country with detailed comparision among states and cities.

Objectives:

We will develop an Interactive web visualization which will be helpful for the government to gain insight about the trends; increase/decrease in number of crimes in US during recent years. We will visualize the regions in United States where the crime rates are high along with the type of crime so that the appropriate police staff could be appointed accordingly. We will also get a picture of the ratio of solved cases with respect to total number of cases, effect of population on crime rates and severity and try to get interesting correlations in any.

We intent to evaluate the following:

1) Figure out the regions with low, medium and high crime rates densities.
2) Filter different regions according to the type of crime.
3) Draw different plots for different regions with the frequency of different kind of crimes occurred in that region.
4) Draw plots for recent years of crime data to see the increase or decrease in the crime rates for each state.
5) Seek some correlation among different type of crimes in all the states and effect of population and other factors on crime rates and severity.

Literature Review:

The reported U.S. violent crime rate includes:

- Murder
- Rapes and sexual assault
- Robbery
- Aggravated assault

Crime rates are necessarily altered by averaging neighborhood higher or lower local rates over a larger population which includes the entire city. Having small pockets of dense crime may lower a city's average crime rate. US violent crime rates includes:

- Homicide
- Gun Violence
- Property Crime
- Crime Against Children

Crime rates vary in the United States depending on the type of community, within metropolitan statistical areas both violent and property crime rates are higher than the national average; in cities located outside metropolitan areas, violent crime was lower than the national average, while property crime was higher for rural areas, both property and violent crime rates were lower than the national average.

Dataset:

Dataset	**Timeline/Year**	**No. of Entries**	**Major Columns/Features**	**Data Source**
Fbi gov crime data	2013	10000 x 14	Population, Violent Crime, Murder and nonnegligent manslaughter, Rapes and sexual assault, Robbery, Aggravated assault etc	https://ucr.fbi.gov/
Crime Rates	1975-2015	3000 x 16	Year, State, Population, Homicide, Rapes , Assaults, Robberies, Crime per capita etc.	https://www.kaggle.com/marshallproject/crime-rates
Effect of Population against crimes	2012	300 x 15	Population, Violent_crime_total, Murder_and_Manslaughter, Forcible_rape, Robbery, Aggravated_assault, Property_crime_total, Burglary, Larceny_theft etc.	https://www.kaggle.com/mascotinme/population-against-crime

Homicide Reports	1980-2015	638455 x 24	City, States, Year. Month, Incident etc.	https://www.kaggle.com/murderaccountability/homicide-reports

Analysis:

We have a number of datasets having information of crime data. We have a dataset which contains type of crime for all the cities in a period of one year. Another dataset has information about crime rate per capita which gives us the effect of population on crime rate and severity. Other datasets have records of various crimes of past 35 years and ratio of solved cases with respect to the total number of cases. To get a meaningful visualization from these data, first we will perform following operations with the data:

- Club different datasets having different information to form one data set where every information related to crime (i.e years, crime rates, types of crimes, crime per capita etc.) can be found and processed.
- Transform the data and find correlation among the rows and columns.
- Clean the data to remove inconsistency and redundant data.

Once having both the city wise and state wise crime data, we will start our analysis with

- Putting different weights to different crime types which will denote the severity of that crime.
- Figuring out the correlation between different type of crimes, areas and population.
- Compare the crimes among cities and states by using different visualization techniques.

Goals\Deliverables:

Our final goal will be to deliver the following -

- Show the low, medium and high crime density areas on the US map.
- Provide two modes of analysis - city-wise and state-wise.
- Provide user interaction so that they can set the severity of different type of crimes.
- Provide user interaction to increase and decrease a particular type of crime to filter the areas on the map.
- Draw plots and graphs to compare and correlate among interesting features of the crime data.

Methodology:

- We will use python for data cleaning and processing.
- Server: We will run server on localhost through python Flask framework.
- Client: Our Client side implementation will be on javascript, html and d3.js libraries.
- We will use different python libraries for our data analysis and other plottings.

References:

Literature Reviews:

- https://en.wikipedia.org/wiki/Crime_in_the_United_States
- https://en.wikipedia.org/wiki/Crime_in_the_United_States#International_comparison
- http://chicago.cbslocal.com/2015/10/22/violent-crime-statistics-for-every-city-in-america/

Data Set:

- https://crime-data-explorer.fr.cloud.gov/
- https://ucr.fbi.gov/crime-in-the-u.s
- https://www.kaggle.com/datasets

PROJECT PROGRESS REPORT

We tried to develop an Interactive web visualization which will be helpful for the government to gain insight about the trends; increase/decrease in number of crimes in US during recent years. We have done visualization of the regions in United States where the crime rates are high along with the type of crime so that the appropriate police staff could be appointed accordingly.

DataSets: Dataset we worked primarily during our analysis and visualization for first phase are:

- https://www.kaggle.com/murderaccountability/homicide-reports : This dataset contains the year wise (1990- 2014) data of all Homicide crimes of different states. The Data contains all the details of Victim and Perpetrator. Some of the Important columns of the dataset are

➔ State and Year of crime
➔ Victim Age, Race and Sex
➔ Perpetrator Age, Race and Sex
➔ Relationship of Perpetrator and victim
➔ Weapons Used

- https://ucr.fbi.gov/: This dataset contains the information of all the crime occurred in year 2015 in different Counties. It contains crime rate per capita of all counties in US and the count of different crimes held in different counties. Some of the Important columns of the dataset are

➔ County_Name
➔ Crime Rate per 100000 population and population
➔ Count of Different Crimes (Murder, Rape, Robbery, Burglary, Larceny, Arson, Population)

Methodology: Methodology used for our data analysis and visualization are:

- **MongoDb**: We have used MongoDb for storing and processing our data. We have stored data after cleanup and data gets loaded at run time from mongo.
- **Python:** Python have been used for background data analysis and cleanup
- **Python Flask Framework:** We have used python flask for running server on localhost:5050

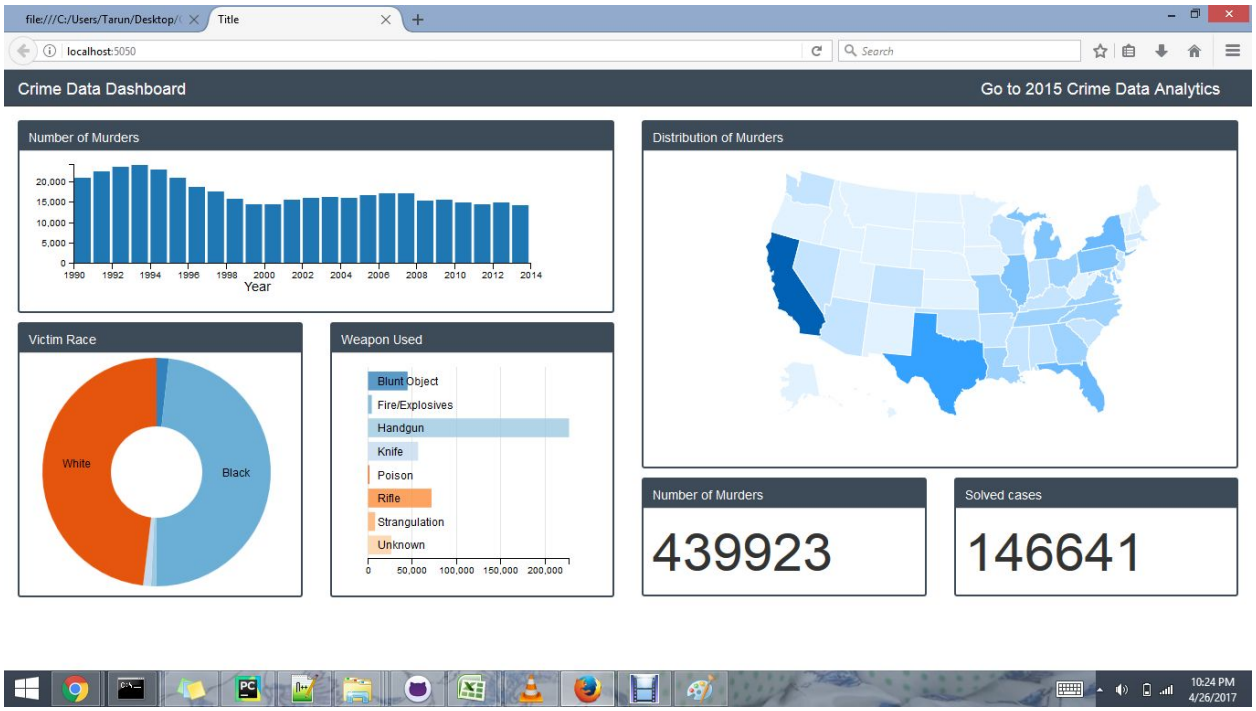
- **D3.js, Javascript, html/css:** D3.js library and javascript has been used for front end visualization and styling.
- **Crossfilter.js, dc.js and queue.js** has been used to filter the data and design the dashboard.

Crime Data Analysis: Dashboards provide a central location for users to access, interact and analyze up-to-date information so they can make smarter, data-driven decisions. During first phase of our crime data analysis, we mostly focussed on visualizing crime data on dashboard which helps engaging the viewer within few seconds and provide clear understanding of the visualization.

Task Completed in First Phase of our analysis and visualization -

1. **Dashboard:** Designed dashboard using Homicide dataset which will be capable of user interaction. Used 6 different charts which will help to analyse and visualize crime in different U.S states.
2. **DataMap:** Visualized the crime data on US map for all the counties using fbi dataset of year 2015.

1. Dashboard:



Our dashboard has six different charts which are explained below -

1. **Number of murders(Histogram)** - This chart shows the number of murders reported from year 1990 to 2014. The chart is rendered using the dc.barchart. User can select only a few bars and other charts will show the stats of only selected years.

2. **Distribution of Murders(DataMap)** - The U.S datamap shows the distribution of the murders in different states. We have used color distribution for our analysis where light blue corresponds to lesser number of murders and dark blue is for higher number of murders. User can select a particular state to visualize the stats of that state. Also, hovering on the state shows the state code and number of murders in that state for the particular configuration selected by the user at that time. The map is rendered using the dc.geoChoroplethChart.

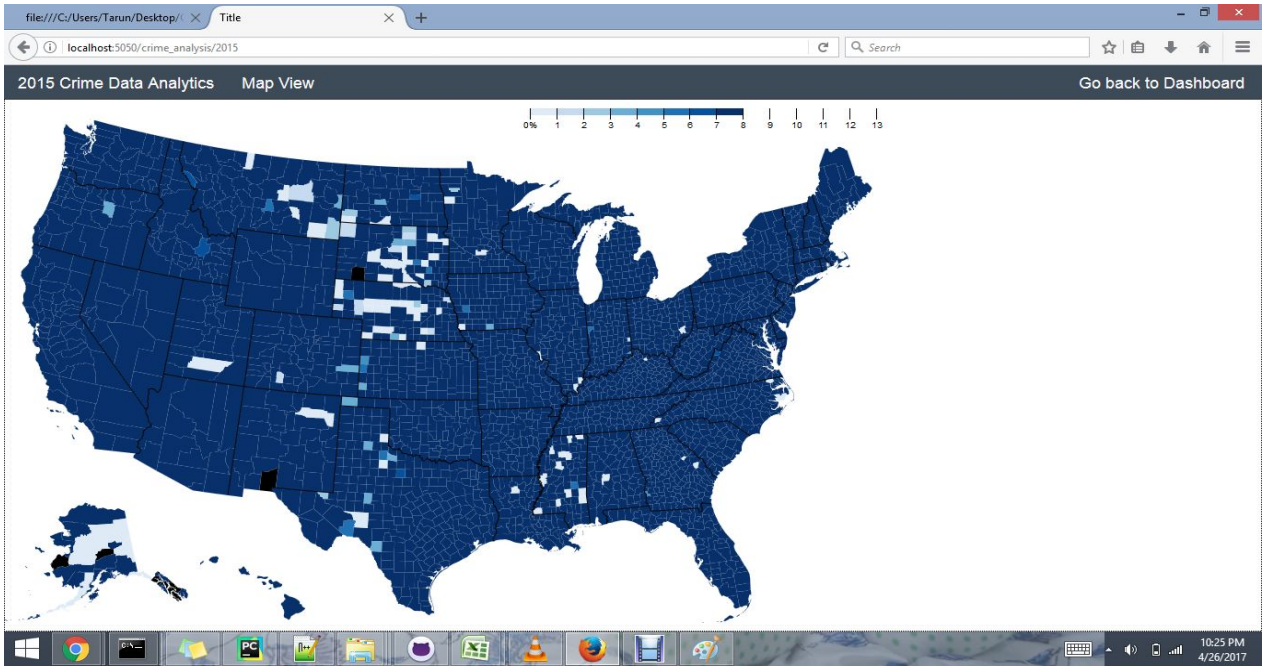
3. **Victim Race(Pie Chart)** - This donut chart shows the percentage of various type of victims involved in the murders. The chart is rendered using the dc.pieChart library. It also supports user interaction and the user can select the slice to see the stats of victims of a particular race.

4. **Weapon Used(Histogram)** - This chart shows the various types of weapons used in the murder with the number of murders on the x-axis. The chart has been rendered using the dc.rowChart library. By selecting a bar, user can see the stats of victims murdered by the selected weapon.

5. **Total Cases**- This area displays the murder count for the selected configuration. Its value gets updated accordingly as the user make selections. It is rendered using the dc.numberDisplay library.

6. **Solved cases** - This area similar to the above, displays the count of solved cases for the selected configuration. Its value gets updated accordingly as the user make selections. It is rendered using the dc.numberDisplay library.

2. DataMap:



In this data map we tried to show the distribution of crime at county level. On hover over a county, the count of Murders, Rapes, Robbery, Assault, Burglary, Larceny, Theft and Arson along with total number of criminal incident and total population is displayed.

Next Steps:

Our further analysis on crime data will be in categorized in different stages:

1. We will enhance our dashboard to get a more clear and engaging visualization.

2. We will extend the data map view for states as well. We also intend to provide sliders on the right side of the map where the users can increase or decrease the weights of different type of crimes. Also we intend to provide sider chart capable of comparing the crimes among different states/counties.

3. We will enhance our crime data analysis for different states and counties of US through bubble chart. Compared with many charts and graphs, bubble charts have lots of advantages. They can legibly show data values that differ by a ratio of 100,000, and can display hundreds of individual values at once. We will do our analysis as per below marks.

➔ **Analysis of crimes for all states:** We will show the population , crime and crime per capita of all states in US through Bubble chart. The size of the Bubble will vary based on the number of crimes and population per capita of all states. We will mostly focus our analysis on violent crimes of all states i.e. Murder, Rape, Robbery and Assault

➔ **Analysis of crimes for all counties of a states:** We will extend our analysis for all counties of all states. Our data visualization will contain the information of number of crimes, population and crime per population of any county. The size of the Bubble will vary based on the number of crimes and population per capita of all counties. Here also our analysis will mostly focus on violent crimes of all counties i.e. Murder, Rape, Robbery and Assault

➔ **Implement Search Bar:** Our next task will be to implement search bar in our html which enable user to search for any state or county and all the crime information will get populated.

References:

- https://github.com/dc-js/dc.js/tree/master
- https://crime-data-explorer.fr.cloud.gov/
- https://ucr.fbi.gov/crime-in-the-u.s

FINAL PROJECT REPORT

We developed Interactive web visualizations which will be helpful for the government to gain insight about the trends; compare crime rates between different states and counties. increase/decrease in number of crimes in US during recent years. We found the correlation and top attributes between different type of crimes. We did our analysis by visualizing data in several ways which help us to analyze data and predict results more accurately. We have done visualization of the regions in United States where the crime rates are high along with the type of crime so that the appropriate police staff could be appointed accordingly.

DataSets: Apart from the dataset in which we worked till progress report, we refined our existing data to fit our requirements according to the state and county views. The datasets in which we worked for our entire project are:

- https://www.kaggle.com/mascotinme/population-against-crime : This dataset contains information about the crime rate against per capita of the population. We tried to analyze our data for state and counties based on per capita, which will give more meaningful insight than count. Important columns in which we worked in this dataset are:

➔ Number of Crimes
➔ Year, State
➔ Population
➔ Homicide, Rapes , Assaults, Robberies

- https://www.kaggle.com/mascotinme/population-against-crime : This crime dataset is submitted to UCR by varios US county police departments. This dataset contains the effect of population on crimes and how crime rate increases/decreases with increase/decrease in population. The dataset contains:

➔ Population
➔ Violent Crime Total Murder and Manslaughter
➔ Forcible Rape, Robbery Aggravated Assault

- https://www.kaggle.com/murderaccountability/homicide-reports : This dataset contains the year wise (1980- 2014) data of all Homicide crimes of different states. The Data contains all the details of Victim and Perpetrator. Some of the Important columns of the dataset are

➔ State and Year of crime
➔ Victim Age, Race and Sex

➔ Perpetrator Age, Race and Sex
➔ Relationship of Perpetrator and victim
➔ Weapons Used

- https://ucr.fbi.gov/: This dataset contains the information of all the crime occurred in year 2015 in different Counties. It contains crime rate per capita of all counties in US and the count of different crimes held in different counties. Some of the Important columns of the dataset are

➔ County_Name
➔ Crime Rate per 100000 population and population
➔ Count of Different Crimes (Murder, Rape, Robbery, Burglary, Larceny, Arson, Population)

Methodology:

We followed the same methodology as in progress report for our data analysis and visualization:

- **MongoDb**: We have used MongoDb for storing and processing our data. We have stored data after cleanup and data gets loaded at run time from mongo.
- **Python:** Python have been used for background data analysis and cleanup
- **Python Flask Framework:** We have used python flask for running server on localhost:5050
- **D3.js, Javascript, html/css:** D3.js library and javascript has been used for front end visualization and styling.
- **Crossfilter.js, dc.js and queue.js** has been used to filter the data and design the dashboard.

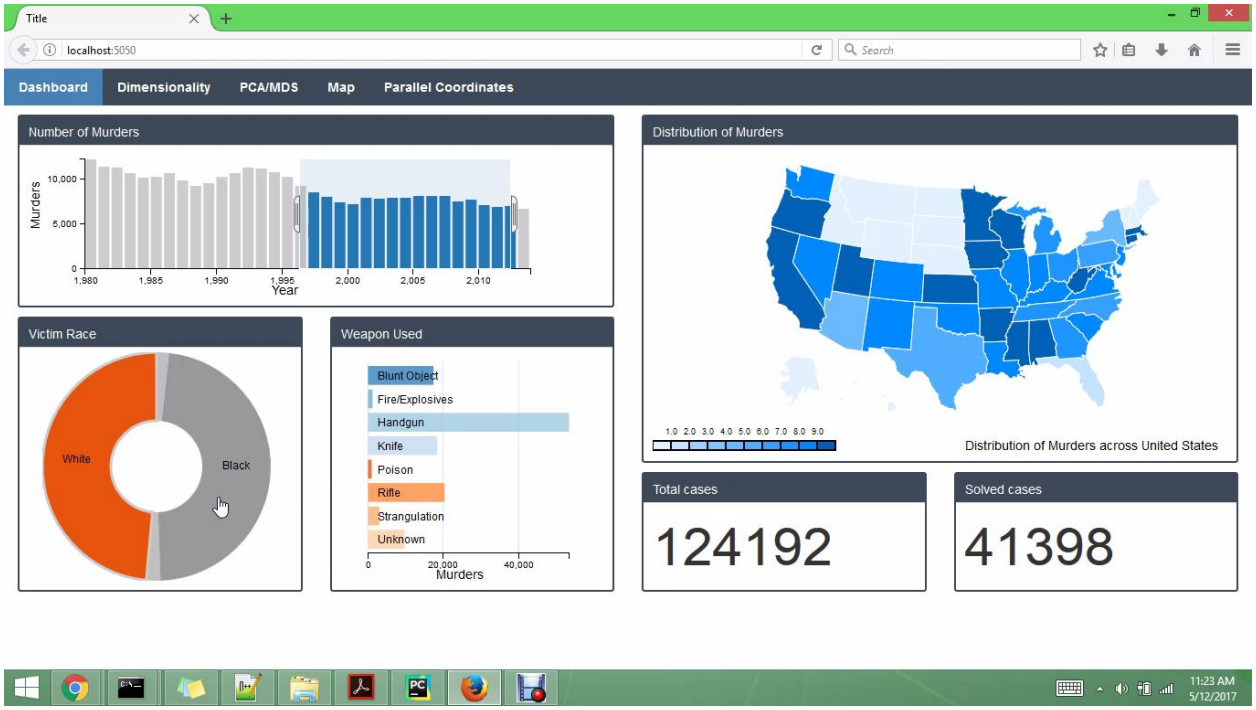
Crime Data Analysis:

We did our crime data analysis in different steps and analyzing different data through different visualization techniques.

- **Dashboards :** Dashboard provide a central location for users to access, interact and analyze up-to-date information so they can make smarter, data-driven decisions. During first phase of our crime data analysis, we mostly focussed on visualizing crime data on dashboard which helps engaging the viewer within few seconds and provide clear understanding of the visualization.

  Designed dashboard using Homicide dataset which will be capable of user interaction. Used 6 different charts which will help to analyse and visualize crime in different U.S states.

Our dashboard has six different charts which are shown and explained explained below -



1. **Number of murders(Histogram)** - This chart shows the number of murders reported from year 1990 to 2014. The chart is rendered using the dc.barchart and brushing is enabled. User can select only a few bars and other charts will show the stats of only selected years.

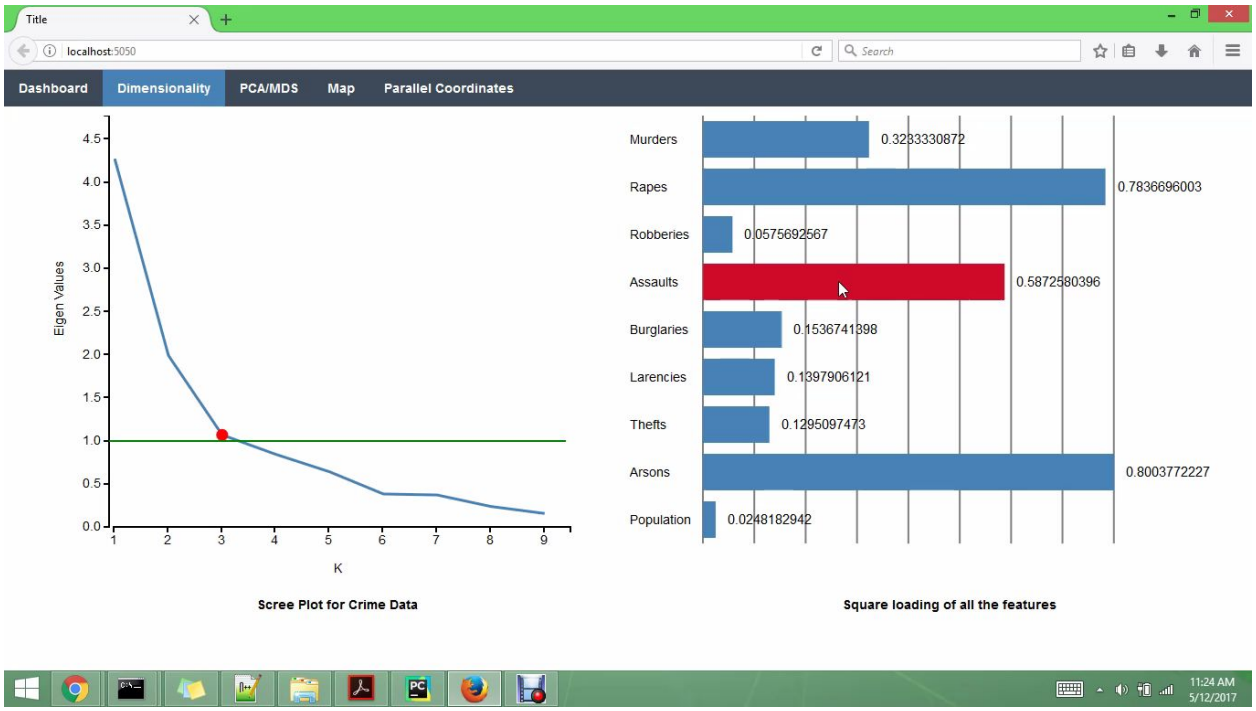
2. **Distribution of Murders(DataMap)** - The U.S datamap shows the distribution of the murders in different states. We have used color distribution for our analysis where light blue corresponds to lesser number of murders and dark blue is for higher number of murders. User can select a particular state to visualize the stats of that state. Also, hovering on the state shows the state code and number of murders in that state for the particular configuration selected by the user at that time. The map is rendered using the dc.geoChoroplethChart.

3. **Victim Race(Pie Chart)** - This donut chart shows the percentage of various type of victims involved in the murders. The chart is rendered using the dc.pieChart library. It also supports user interaction and the user can select the slice to see the stats of victims of a particular race.

4. **Weapon Used(Histogram)** - This chart shows the various types of weapons used in the murder with the number of murders on the x-axis. The chart has been rendered using the dc.rowChart library. By selecting a bar, user can see the stats of victims murdered by the selected weapon.

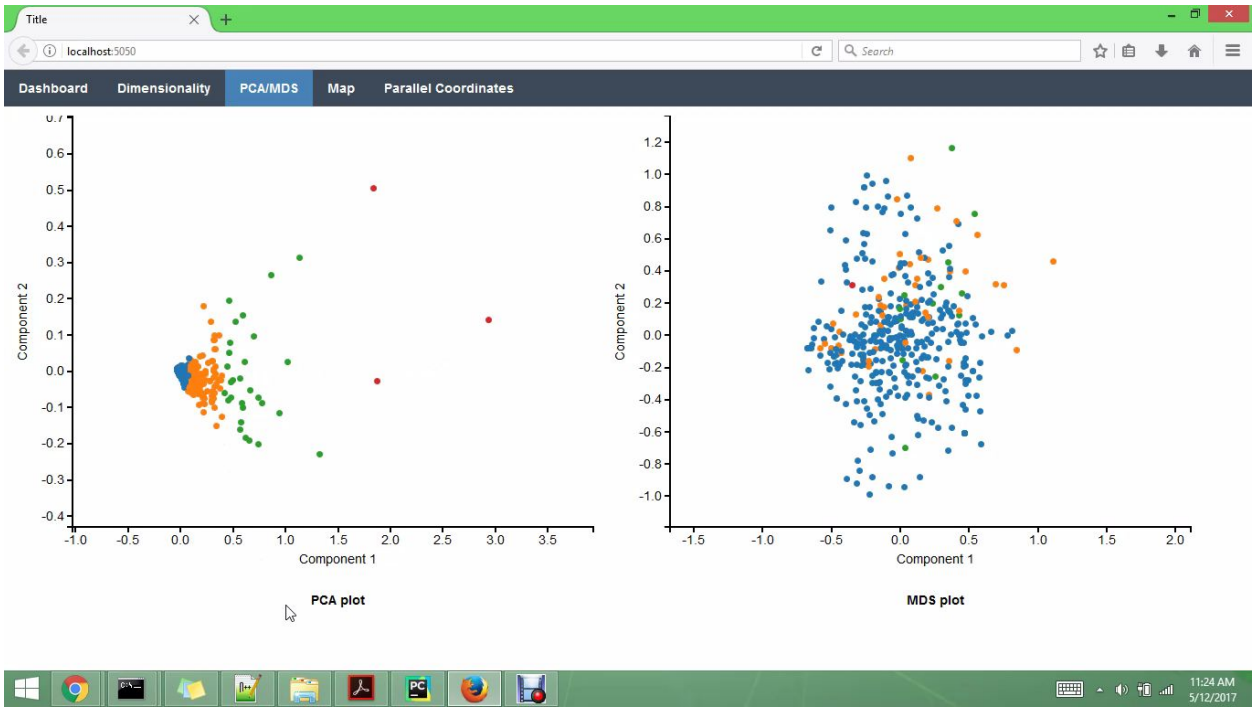
5. **Total Cases**- This area displays the murder count for the selected configuration. Its value gets updated accordingly as the user make selections. It is rendered using the dc.numberDisplay library.

6. **Solved cases** - This area similar to the above, displays the count of solved cases for the selected configuration. Its value gets updated accordingly as the user make selections. It is rendered using the dc.numberDisplay library.



- **Dimensionality:**
    1. Found the intrinsic dimensionality of the dataset used. We plotted the scree plot i.e. eigen values for each dimensions and found that 3 features had eigen values more than 1, which are considered as the the primary components of the data and comprise of more than 98% of the total variation in the data.
    2. To get the top 3 features with maximum intrinsic dimensionality, we plotted the square Loading for each feature and found that Arsons, Assaults and Rapes are the highest three components in the intrinsic dimensionality.

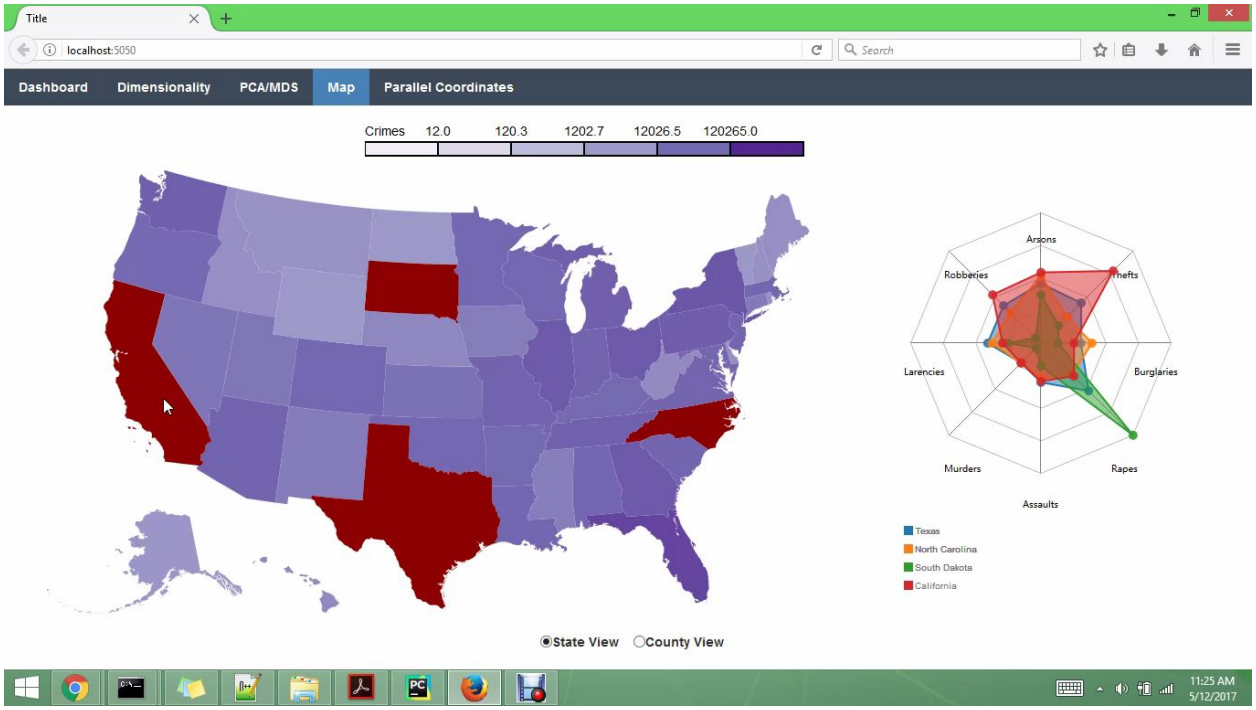
- **PCA/MDS**
    1. **PCA** - Principle component analysis - We plotted the top two principal components with highest variation on the 2D x-y plane. We found the k means clustering for the data and got elbow values as 3. We then clustered the data into 3 components. The clustered data are highly correlated between each other showed the relation.
    2. **MDS Correlation** - We used the MDS technique with the correlation to plot the MDS plot which has its centre at (0, 0) and shows the variation along the top two MDS components. It shows the correlaiton between the multidimensional data and its plot.

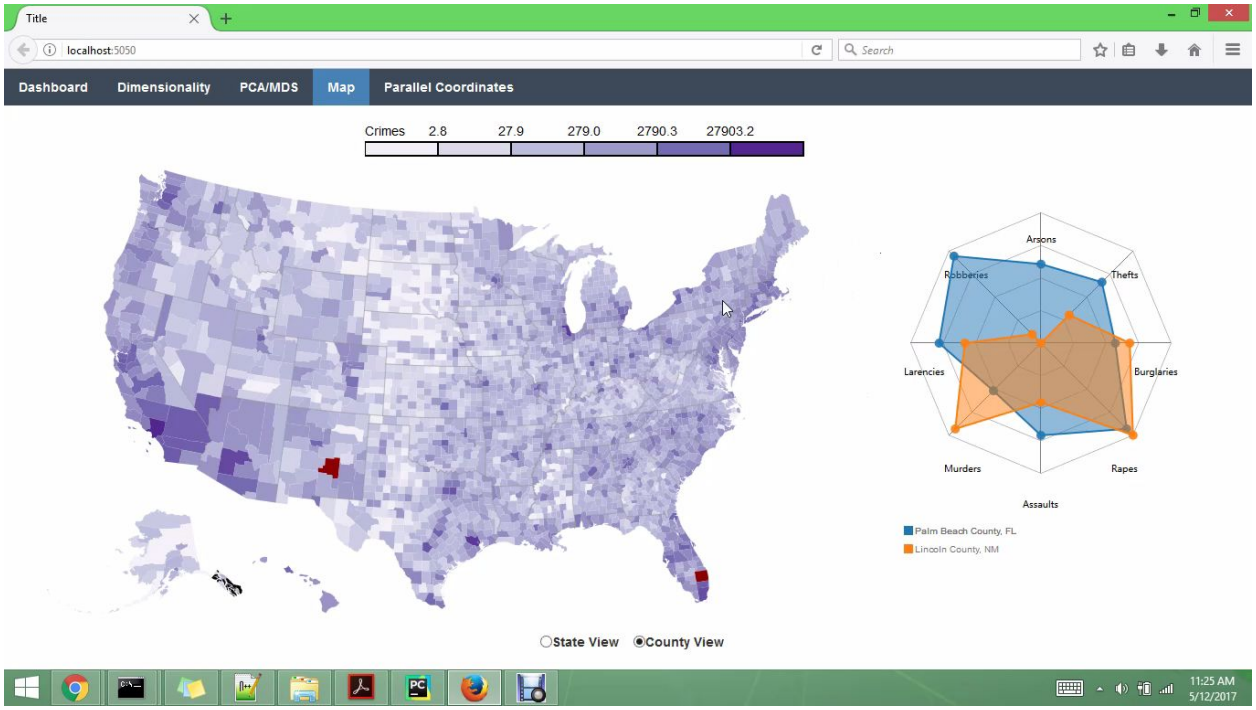


- **Data Map :** Visualized the crime data on US map for all the counties using fbi dataset of year 2015.
    1. State View - The state view allows the user to visualize the crime rate in each of the state with a color coding. User can hover on the state to check the total crimes reported in that state in that year We implemented a radar chart which provides the capability of comparing the different type of crimes among states

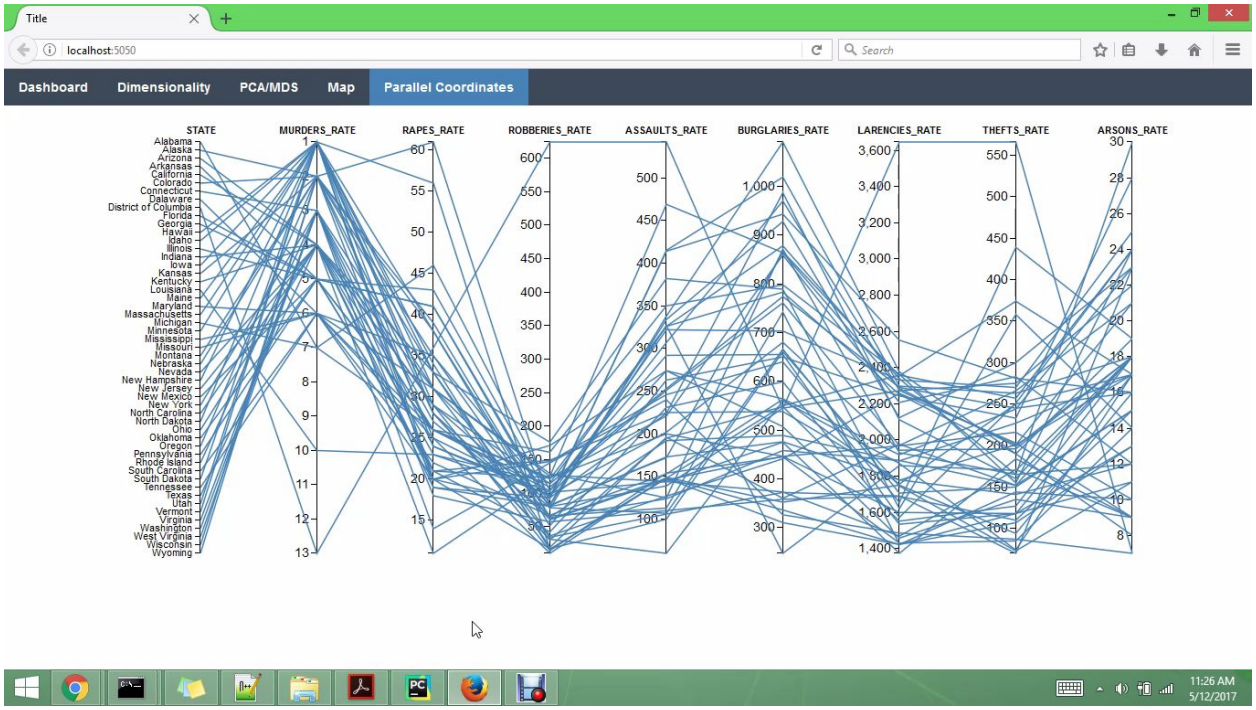
    2. County View - This is similar to the state view, but the granularity has been changed to the county level and in this view, user can compare among different counties.

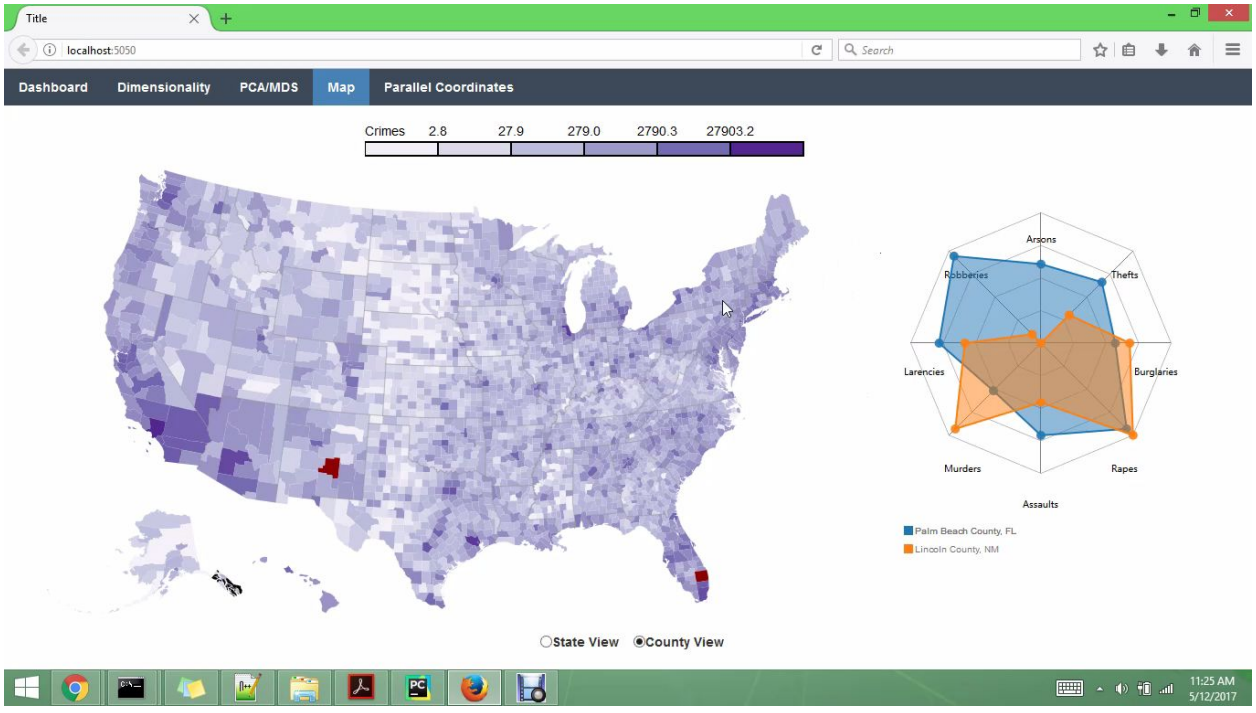
- **Radar Chart :** Radar chart has been implemented in both state and county views. It gives a good approach to compare the different type of crime rates among states or counties (depending on the state or county view). There are 8 different axis in our radar chart showing Arsons, Robberies Larencies, Murders, Assaults, Rapes, Burglaries and Thefts





- **Parallel Coordinates :** Parallel coordinates have been plotted by using each axis in the data set and the goal was to find the correlation among the different attributes in the data set. We found some interesting facts in the parallel plots.





Observations: After analyzing our data we found certain observations

1. Murder rates are higher in states and counties having higher population.
2. Rapes, Arsons and assaults rates are found highest in US as per our analysis
3. The correlation between the murder rates and rapes are highest.
4. Two states that are california and florida are found state to be highest crime rate. As florida is found to be highest for murders and Assaults rates are found highest in california.
5. Minnesota state which is having low population are found to be very less in crimes

References:

- https://github.com/dc-js/dc.js/tree/master
- https://crime-data-explorer.fr.cloud.gov/
- https://ucr.fbi.gov/crime-in-the-u.s