CHICAGO CRIME DATA ANALYSIS

Advance in Data Sciences and Architecture
INFO 7390 - SPRING 2017

PROFESSOR:
SRIKANTH KRISHNAMURTHY

TEAM MEMBERS (Team 9):
AASHRI TANDON
PRAGATI SHAW
SARTHAK AGARWAL

Table of Contents

Project Overview

Implemented time series analysis to predict the crime count for a given month and year. The predictions are done for the districts of Chicago.

Performed geospatial analysis and statistical analysis to study the trend of crime across districts of Chicago.

Hosted Application

The application is hosted on Azure and can be accessed using the following URL http://138.91.152.159:8080/ADS_FinalProject/

Infrastructure

Python – Data processing

R - Machine Learning (Time Series analysis)

Docker – For easy distribution and submission.

Java – Web application.

Microsoft Azure – Machine learning Rest API

ArcGIS pro – Geospatial Analysis

D3 – Statistical Analysis

FME workbench – Creating shape files

AZURE ML account login information:

username: team9ads@outlook.com

password: Northeastern

Data wrangling

Data set : Chicago crime data

Dataset Description:

Source: City of Chicago link

Data Size: 1.4 gigabytes

Columns: Dataset has 22 columns.

Columns
ID
Case Number
Date
Block
IUCR
Primary Type
Description
Location Description
Arrest
Domestic
Beat
District
Ward
Community Area
FBI Code
X Coordinate
Y Coordinate
Year
Updated On
Latitude
Longitude
Location

(A) Data Download and pre-processing

Dockerizing

docker pull aashritandon/chicagocrime

Running the docker image:
docker run aashritandon/chicagocrime python DataDownload.py
accessKey=AWS_ACCESS_KEY_HERE
secretKey=AWS_SECRET_ACCESS_KEY_HERE

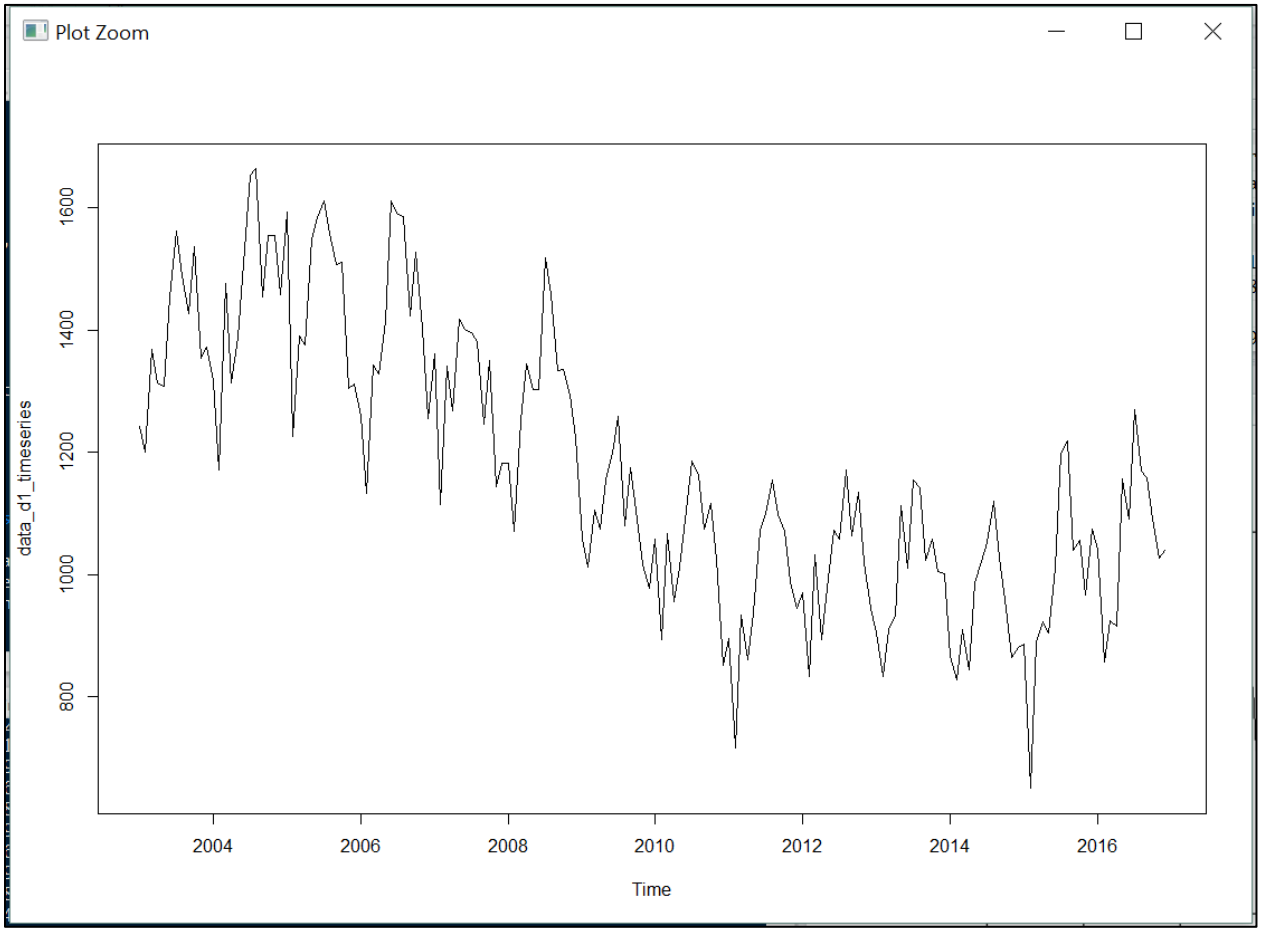
Time Series Analysis

Our goal is to analyze the time series data of Chicago dataset where we observe the monthly data from 2003 to 20016 and crime rate for each district of Chicago and then make future predictions on the crime rate for 2017, 2018, 2019.

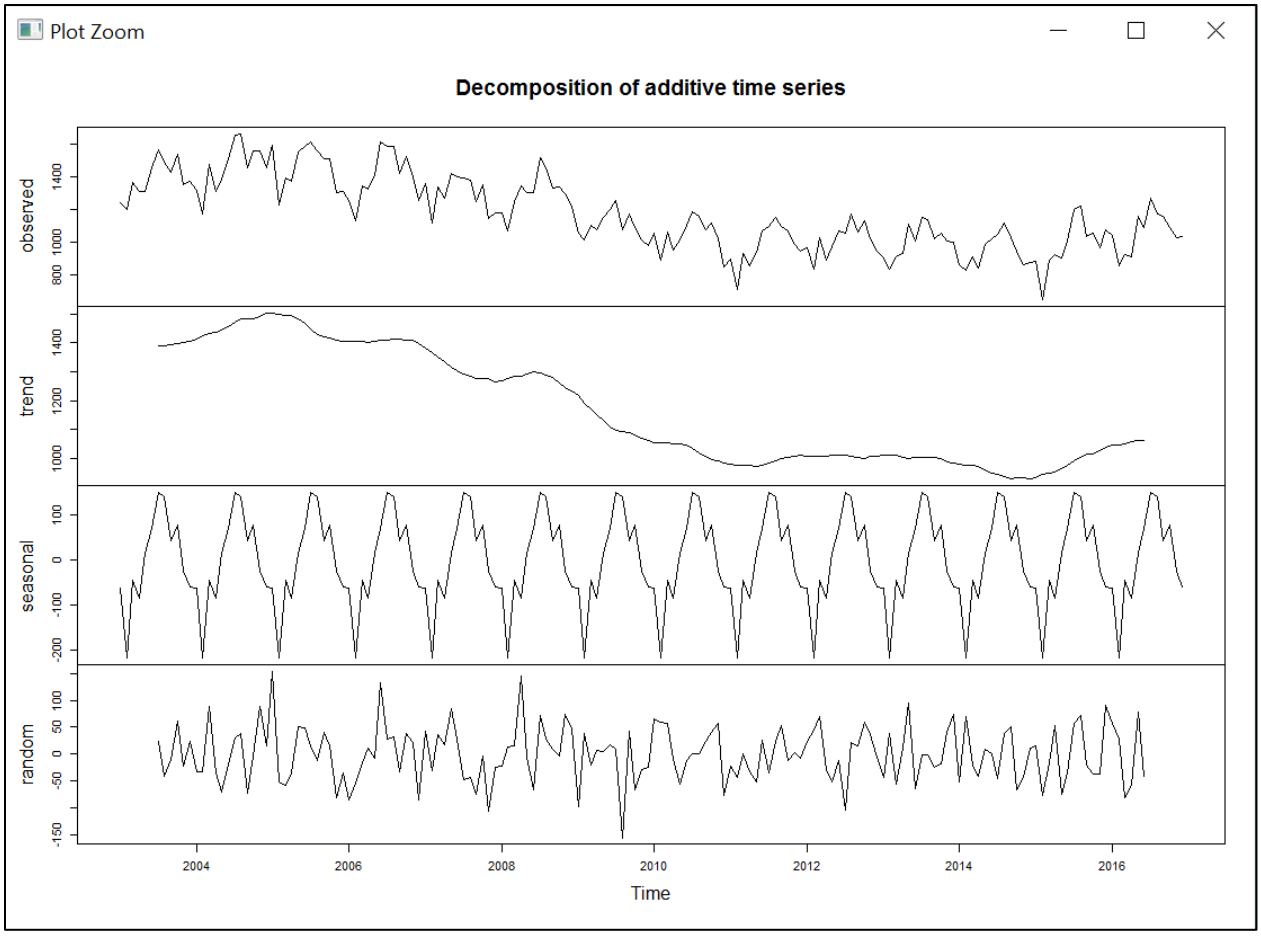
Here, we will explain the analysis done for District 1, it is same for all the districts.

Our first step was to manipulate the dataset into time series format.

The time series for crime rate looks like this:



Then we decomposed the series to analyze if the series has seasonal, trend component or both.



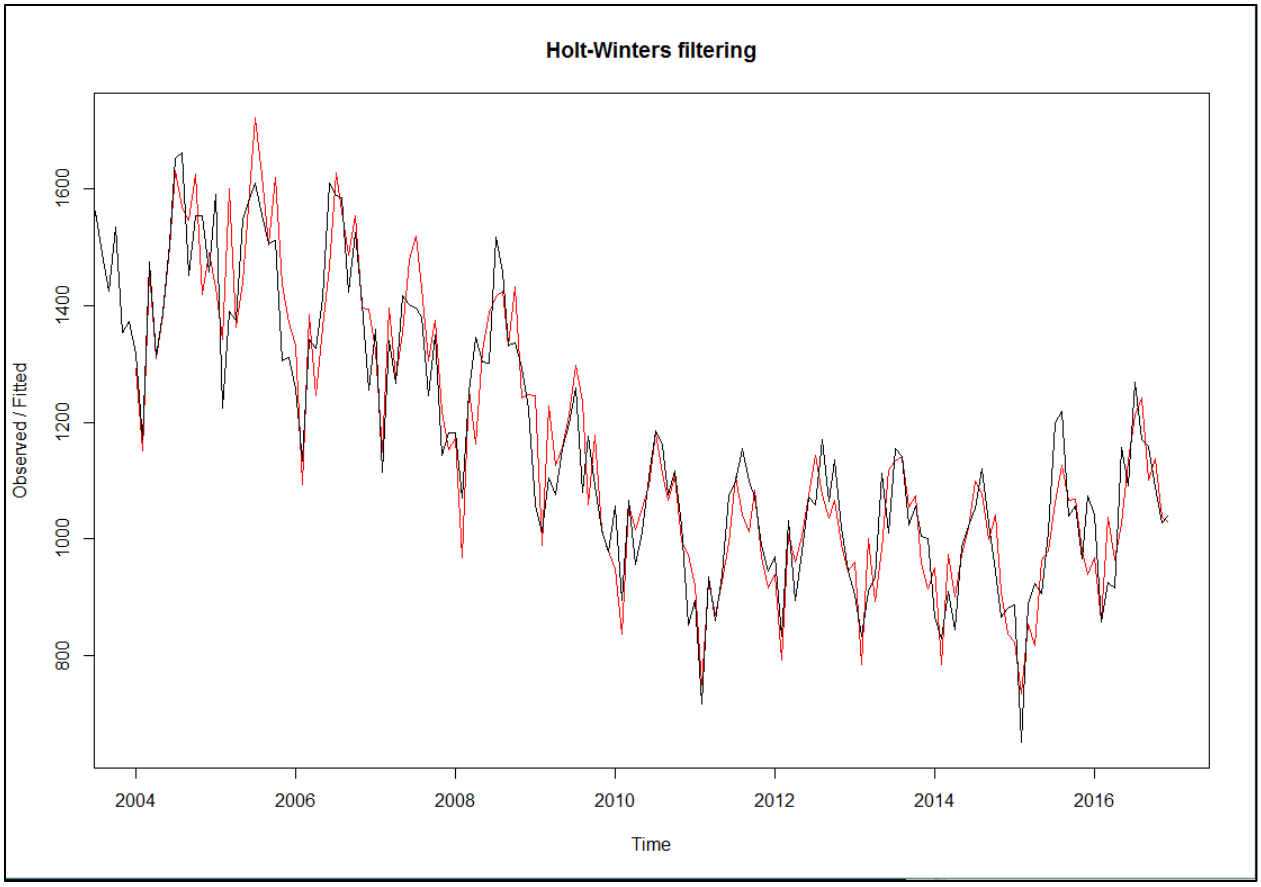
We can clearly see that there is a trend where the series maxima is at ~1500 and then it lowers down to ~1000 then oscillates up and down a few times.

The series has a noticeable seasonal component as well.

Since the series has increasing and decreasing trend and seasonality, we used **Holt-Winters exponential smoothing** to make short-term forecasts.

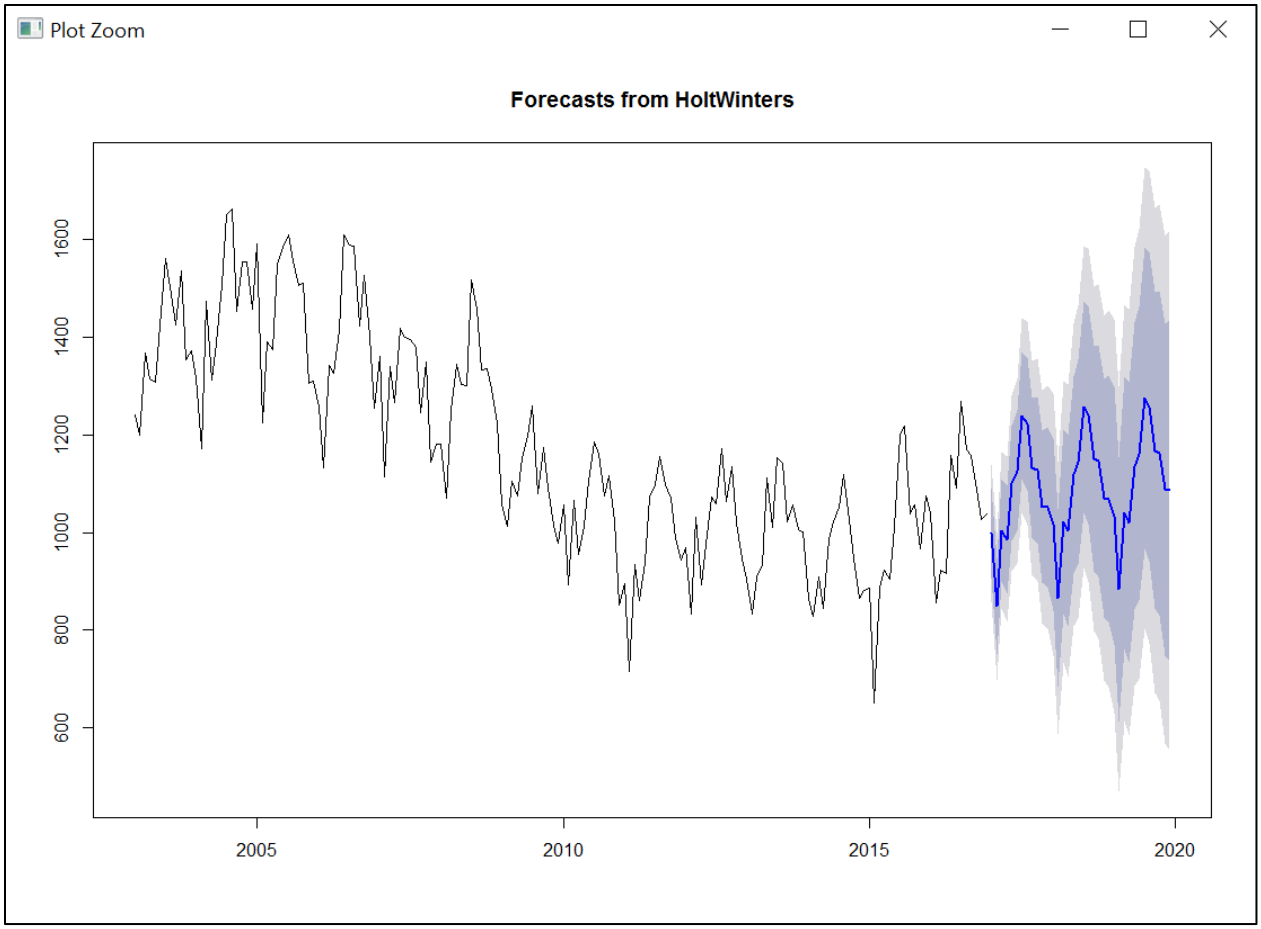
Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.

In the below plot you can see the trained model. The original series is shown in black, the forecasted values are shown in red.



For visualizing the future 3 years of forecast which is not included in the original time series, we use the “forecast.HoltWinters()” function in the “forecast” package.

The forecasts are shown as a blue line, and the dark grey and light grey shaded areas show 80% and 95% prediction intervals, respectively.


Plot Zoom
Forecasts from HoltWinters
1600
1400
1200
1000
800
600
2005
2010
2015
2020

Geospatial Analysis

Using ArcGIS PRO Desktop

We performed geospatial analysis to study the trend of crime of Chicago.

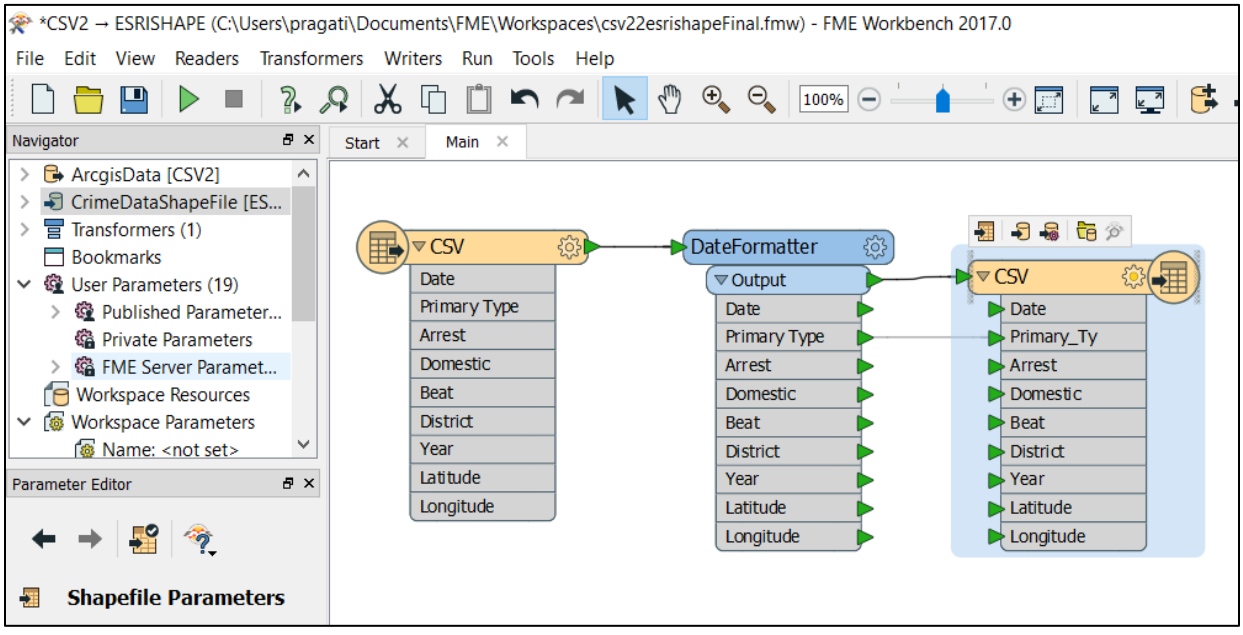
Steps:

1. Manually clustered the data set based on districts
2. Summarized the data based on different crime types.
3. Created categorical data
4. Created a new feature - yearly total crime of a district

5. Exported data to a csv

Year	District	narcotics_	theft_cour	criminal_d	burgalary_	assault_co	robbery_c	others_co	total_crim	Latitude	Longitude
2001	1	1	45	0	0	1	1	90	138	41.98741	-87.6101
2001	2	0	60	4	1	2	1	104	172	41.97608	-87.5814
2001	3	0	63	2	0	1		100	166	41.79325	-87.5573
2001	4	2	66	0	5	2	4	118	197	41.76044	-87.5271
2001	5	0	53	2	3	0	0	76	134	41.72173	-87.5934
2001	6	1	61	0	1	1	7	116	187	41.75847	-87.5981

6. Converted csv data to a shape file using FME Workbench



7. Installed R-ArcGIS bridge to connect R-studio to ArcGIS Desktop

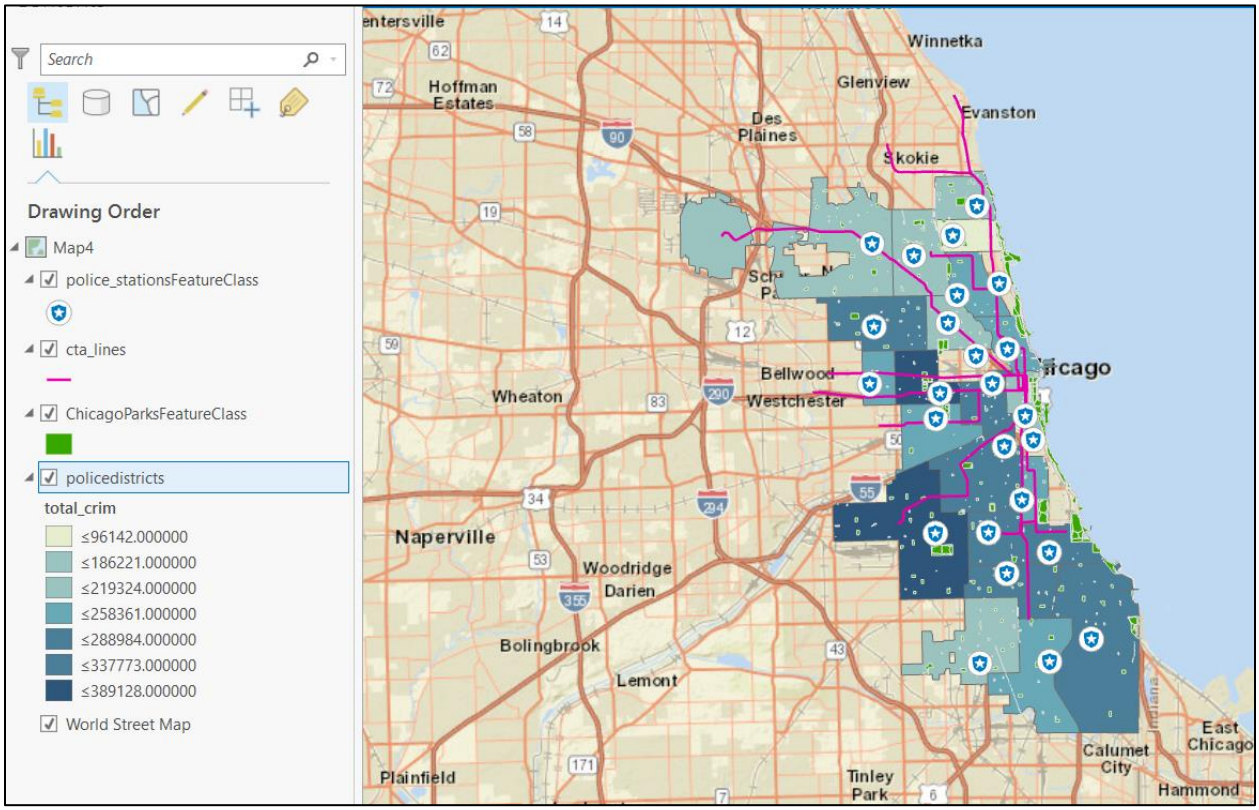
```
library(arcgisbinding)
arc.check_product()

enrich_df <- arc.open(path = 'D:/Primary2012/CSV.shp')
enrich_select_df <- arc.select(object = enrich_df, fields = c('FID', 'Shape','Year','District','narcotics_','theft_coun','
library(sp)
enrich_spdf <- arc.data2sp(enrich_select_df)
head(enrich_spdf@data)
arcgis_df <- arc.sp2data(enrich_spdf)
arc.write('C:/Users/pragati/Documents/ArcGIS/Projects/CrimeDataAnalysis/CrimeDataAnalysis.gdb/Primary2012', arcgis_df)
```

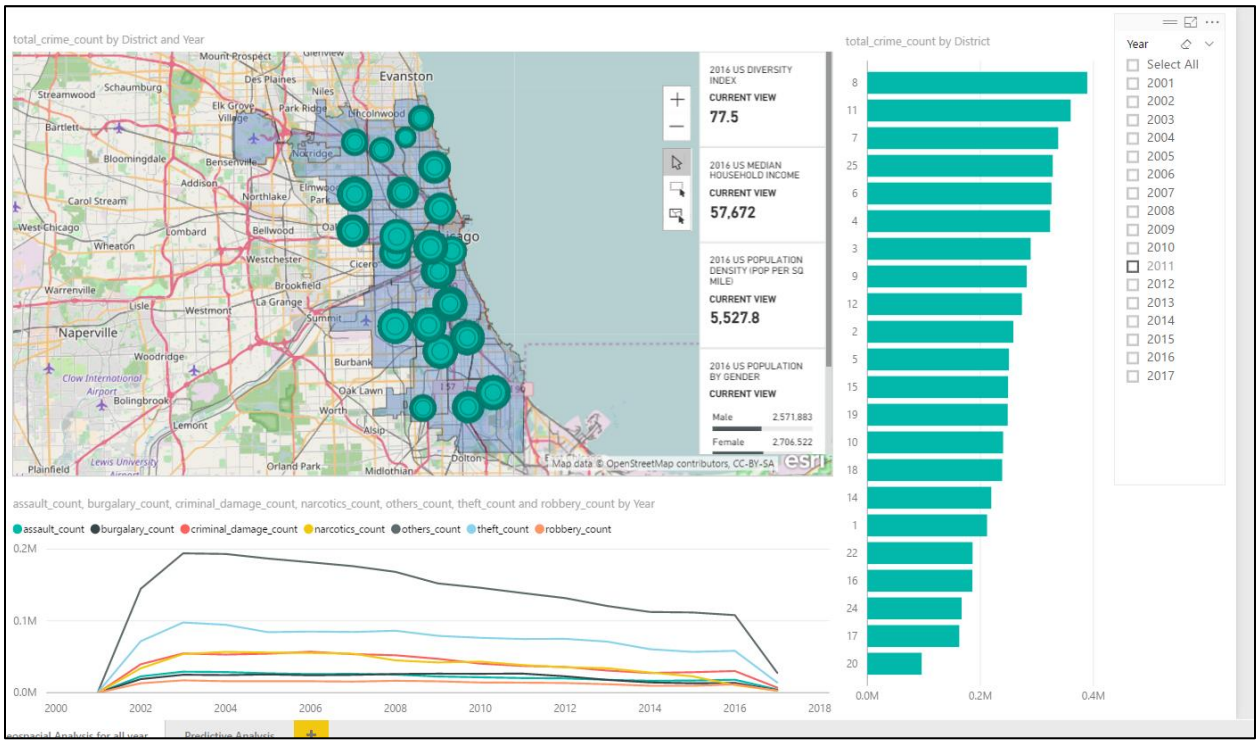
8. Created features classes to create feature layers in ArcMap

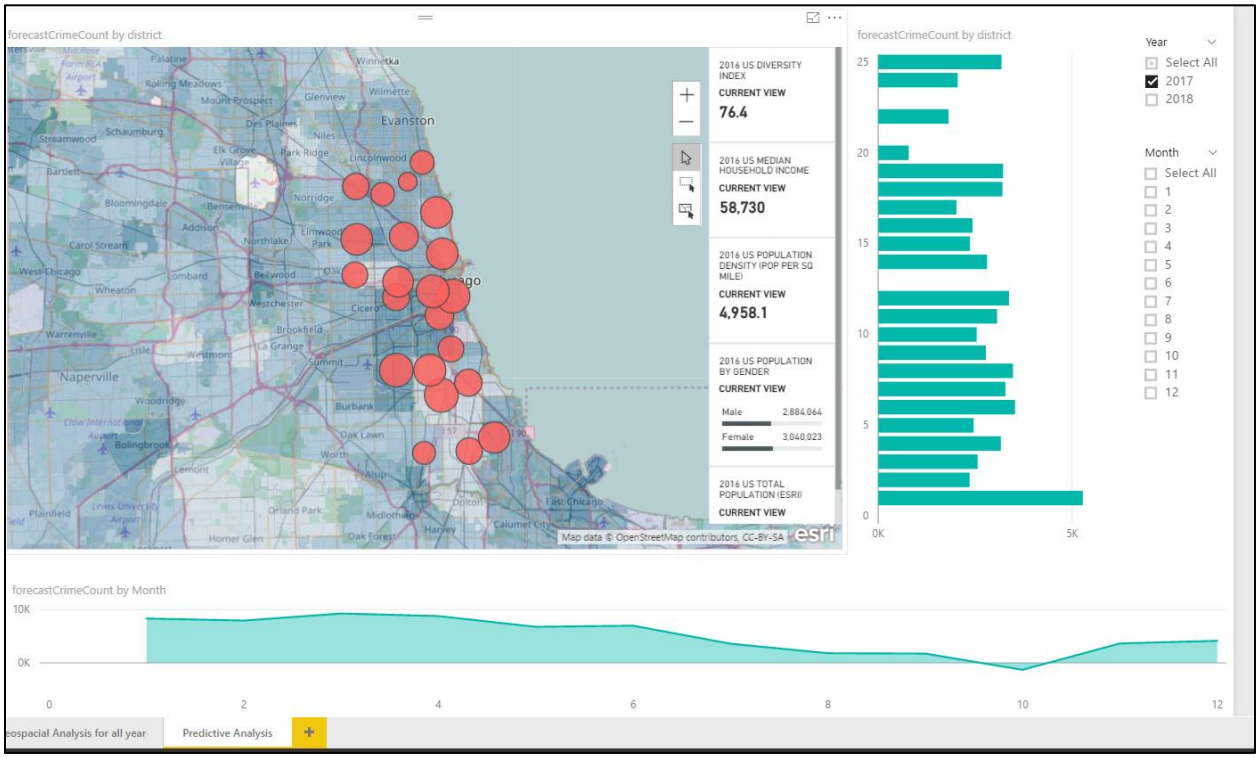
The below ArcMap shows the heat Map of Chicago district based on total crime count for all years(2001-2017).

The Chicago train lines, parks and recreation areas have also been plotted to analyze the crime trend.



Using ArcGIS + PowerBI Desktop





Analysis

- Top 2 districts with total crime has been district number 8 and 11 for more than a decade.
- From 2011 there has been a decrease in crime rate of district 25
- For the period of 2001 to present, it has been observed that theft and burglary crime type has been the most.
- For district number 8 and 11 crime type assault has been the most.

Statistical Analysis

Till date, we have been using data visualization tools like Tableau and PowerBI for analysis the trends in our data. For Chicago Crime Data analysis, we have leveraged the power of D3 JavaScript library. D3 has provided us with the capability to generate bar charts, line charts, pie charts dynamically. It can fetch the data from csv/tsv/excel files and create the chart accordingly.

Steps to create a chart using D3

- Append a svg component to the division where the chart needs to be displayed and specify the width and height.

```
var svg = this.parentElement.append("svg")
.attr("width",width + margin.left + margin.right)
.attr("height",height + margin.top + margin.bottom)
.append("g")
.attr("transform","translate(" + margin.left + "," + margin.top + ")");
```

- Add data to the chart

```
var g = svg.selectAll(".bars")
.data(data)
.enter().append("g")

g.append("rect")
.attr("class", "bar1").style("fill",
"orange")
.attr("x", function(d) {
  return x(d.Year) + 10; // center it
})
.attr("width", x.rangeBand() - 20) // make it slimmer
.attr("y", function(d) {
  return y(d.IDCount);
})
.attr("height", function(d) {
  return height - y(d.IDCount);
});
```

- Add axis to the chart

```
svg.append("g").attr("class", "x axis")
.attr("transform","translate(0," + height + ")").call(xAxis);

svg.append("g").attr("class", "y axis").call(yAxis).append("text")
        .attr("transform", "rotate(-90)").attr("y", 6).attr("dy",
                ".71em").style("text-anchor", "end").text("Values");
```

- For updating the chart, we need to do the following step

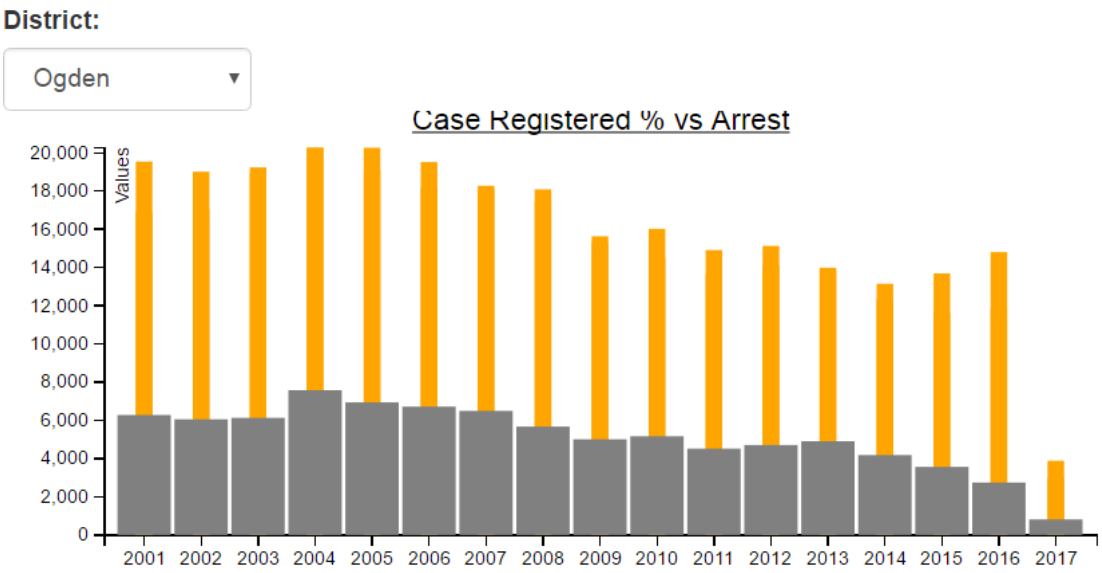
```
var bars = svg.selectAll("rect").data(data, function(d) { return d.Year; });

bars.exit().transition().duration(5).attr("x", -x.rangeBand()).remove();
```

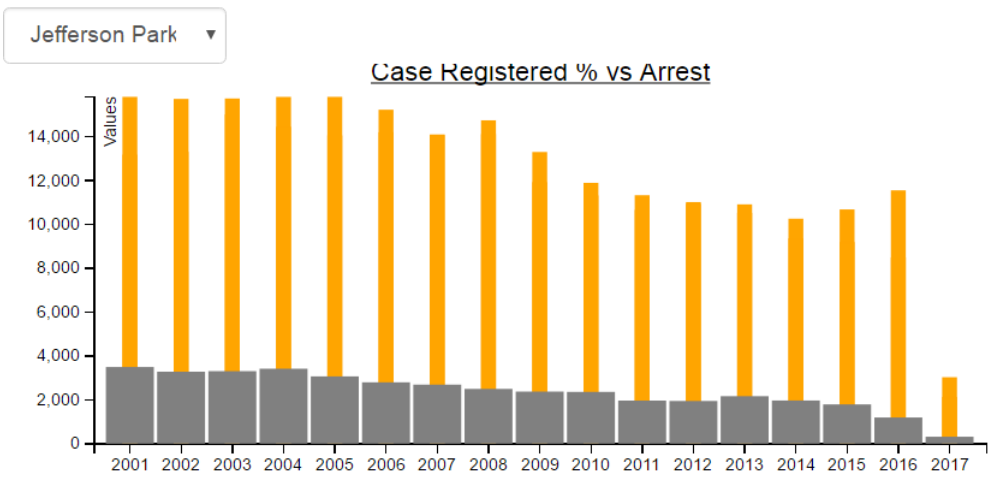
Case Registered v/s Arrest Bar Chart

We have generated a bar chart that displays the comparison between the number of cases registered in a year and the number of arrests. The graph gets dynamically updated for selected district.

For District Ogden:



For District Jefferson Park



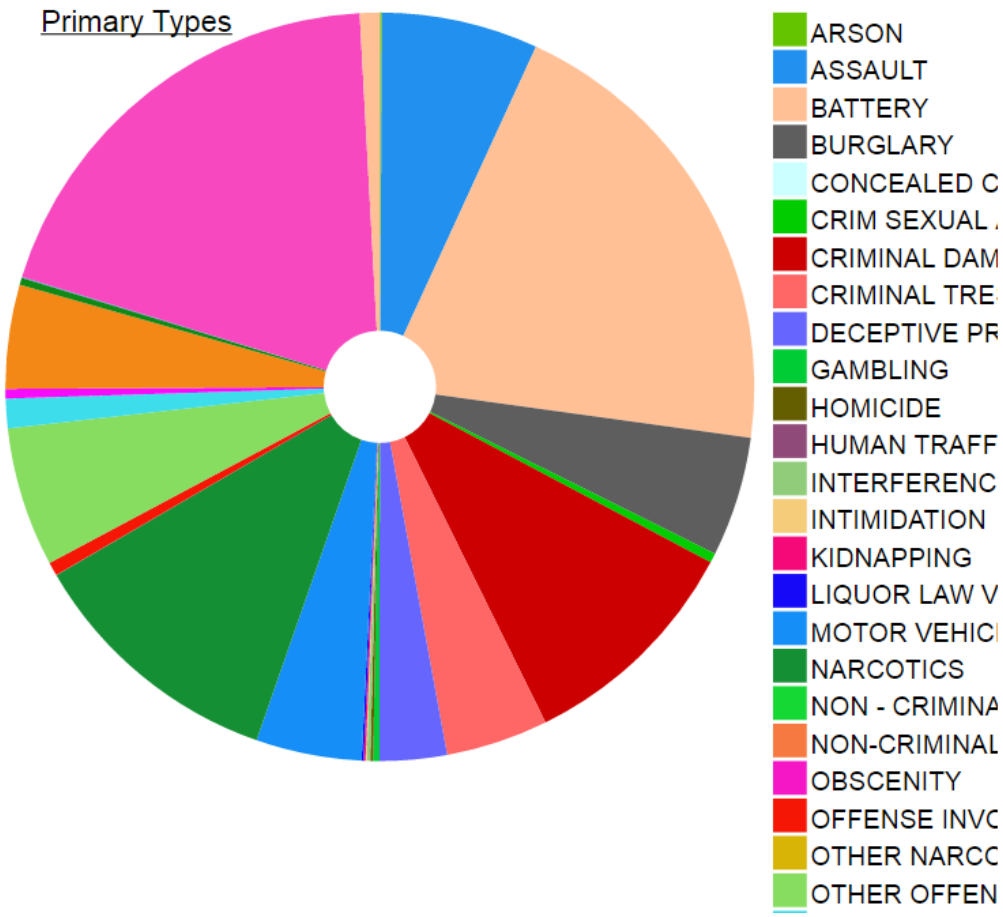
Sample Data file: TotalVsArrest.csv

```
District,Year,IDCount,ArrestCount
1.0,2001,14599,4364
1.0,2002,14742,4757
1.0,2003,16628,6179
1.0,2004,17515,6488
1.0,2005,17517,7184
1.0,2006,16868,6401
1.0,2007,15618,5315
1.0,2008,15635,3571
1.0,2009,13212,4326
1.0,2010,12536,3054
1.0,2011,11807,3529
```

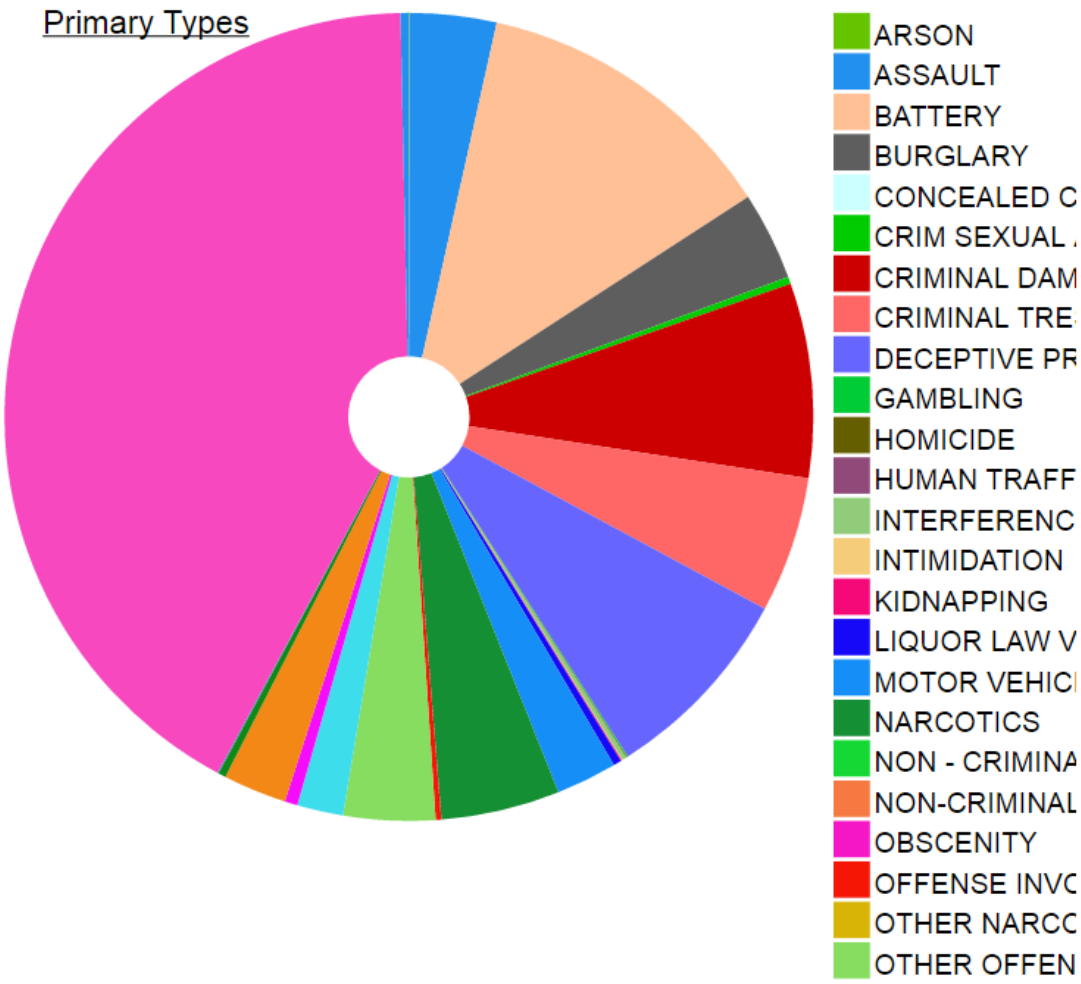
Primary Types Pie Chart

We have generated a pie chart that displays the count of crimes based on primary crime types that happened in a district. The chart dynamically gets updated when a new district is selected.

For District Wentworth:



For District Near North

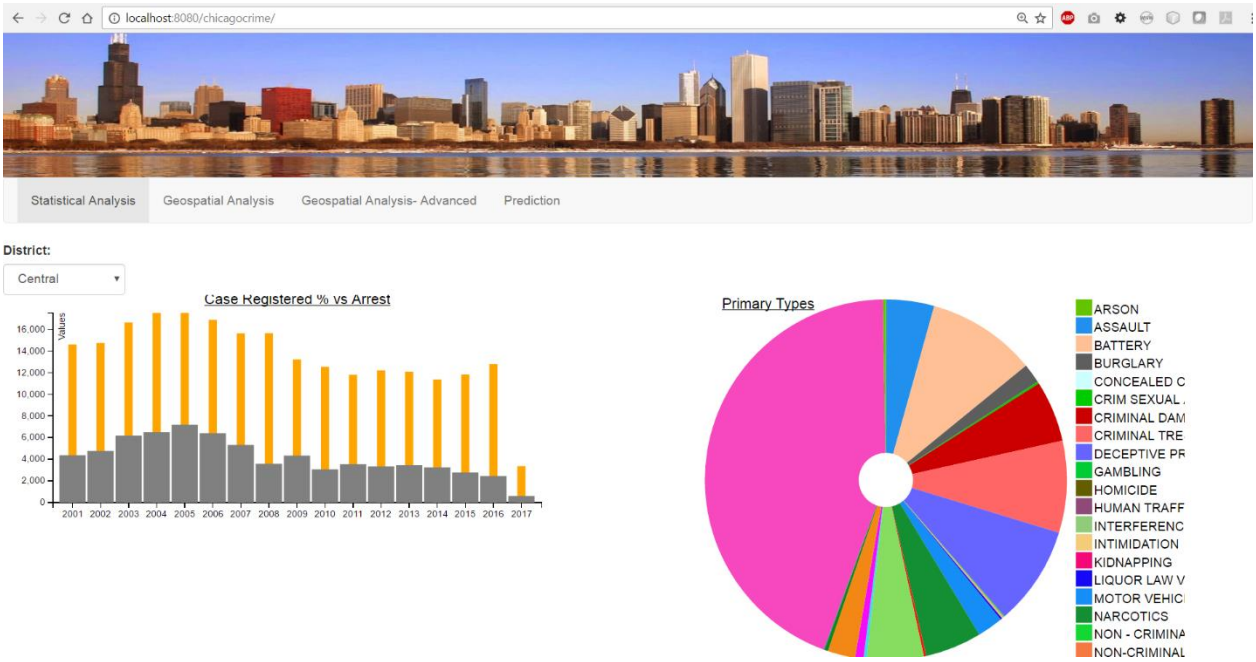


Sample Data File: PrimaryType.csv

```
District,PrimaryType,IDcount
1.0,ARSON,67
1.0,ASSAULT,9857
1.0,BATTERY,22530
1.0,BURGLARY,4012
1.0,CONCEALED CARRY LICENSE VIOLATION,2
1.0,CRIM SEXUAL ASSAULT,452
1.0,CRIMINAL DAMAGE,12606
1.0,CRIMINAL TRESPASS,18925
1.0,DECEPTIVE PRACTICE,20462
1.0,GAMBLING,88
1.0,HOMICIDE,81
1.0,HUMAN TRAFFICKING,3
1.0,INTERFERENCE WITH PUBLIC OFFICER,281
1.0,INTIMIDATION,185
1.0,KIDNAPPING,71
```

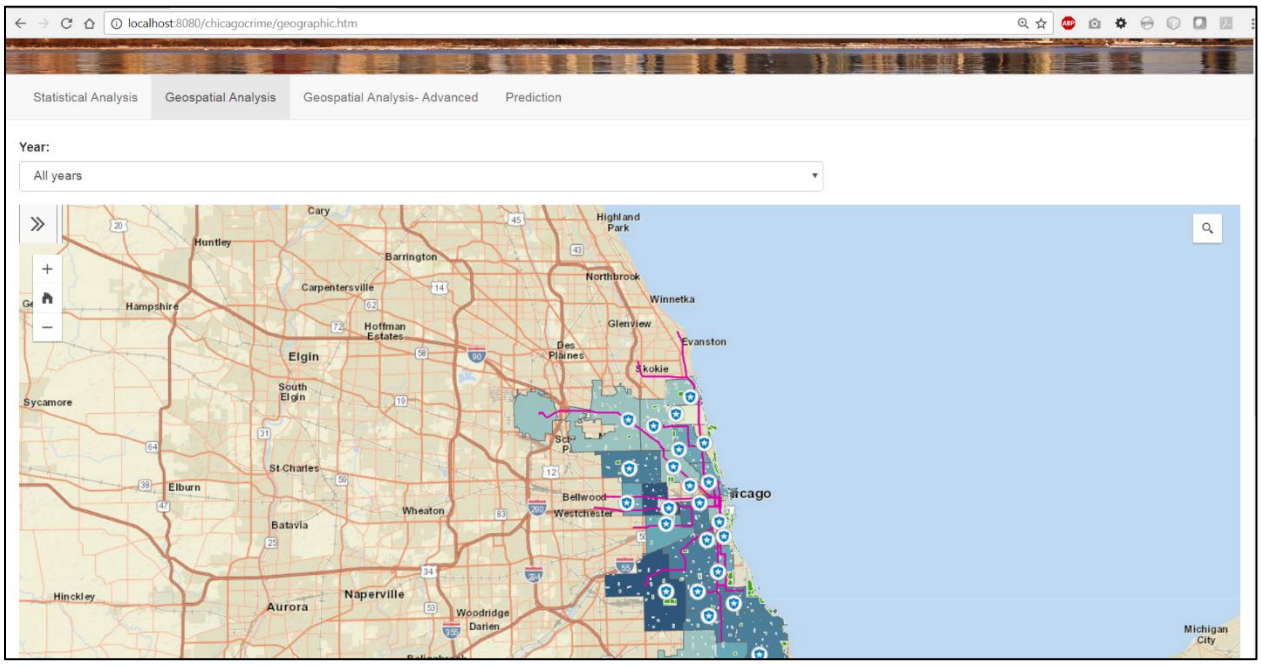
End User Web Application

Statistical Analysis

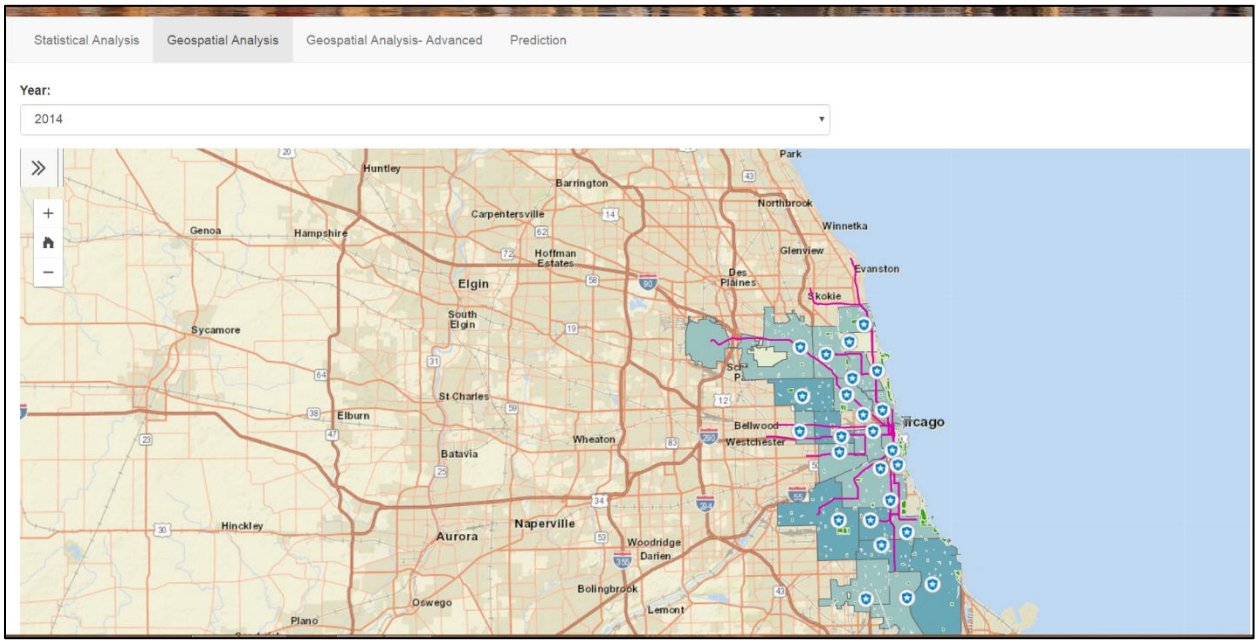


Geospatial Analysis

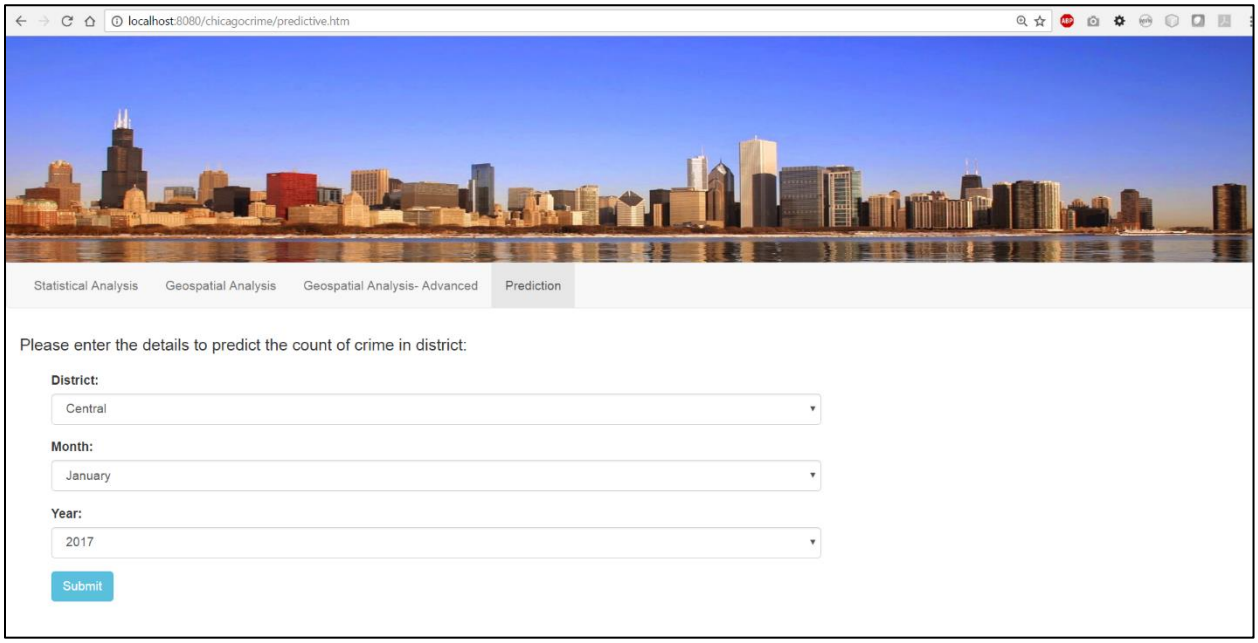
For All years

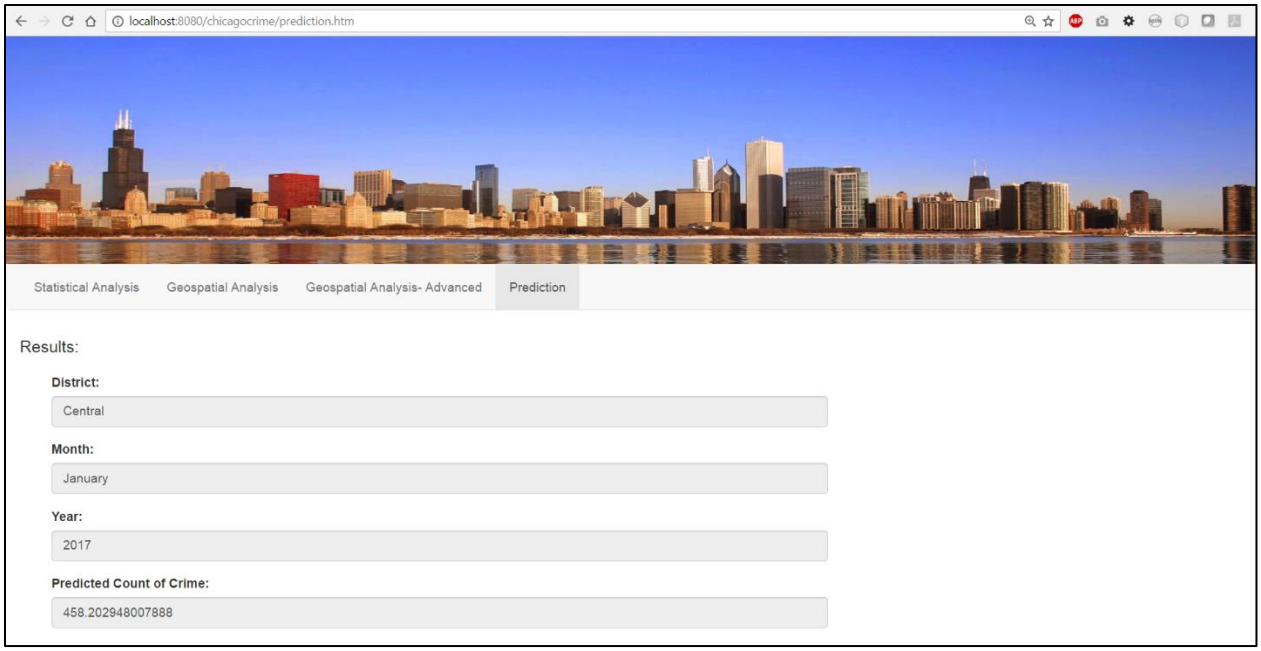


For Year 2014



Predictive Analysis





Hosted Application

The application is hosted on Azure and can be accessed using the following URL http://138.91.152.159:8080/ADS_FinalProject/

Contributions