DATA ANALYSIS – Mini project

# Team

Vijji, Mathew & Komala

# Objective

To do a mini project to on Data Analysis & Visualization using python/Jupyter notebook and related libraries.

Different topic of interest was discussed. The three topics shortlisted were, Diet plans, Crime data & Finance/Credit card related.

Github project was created, added other team members as contributors.

# Process

Data Collection

Multiple data sets were searched through with few constraints,

* + Datasets with at least above 10k
  + Datasets with meaningful values & attributes
  + Datasets with at least more than 10 attributes
  + Minimal invalid/empty values

Data Cleanup

Tasks related to data cleanup was discussed and agreed. Tasks & issues were created in Github and issues were assigned to each team members

* Unwanted columns dropped
* Columns renamed as needed
* Delete rows with empty values
* Combining similar values

# Classification

Analyze 10 years of DC crime data, detect patterns, and classification of the sample data was done with the following questions in mind

* What method & type of Crime across years in DC area
* During which time of the day crime happens
* Which district &quad has the highest level of Crime & what type of Crime
* Which weapon is more used Gun or Knife
* Homicides using Gun across DC

Some of the groupings done,

Dataset is grouped/classified into rate of crime in different Quad, districts, time of a day

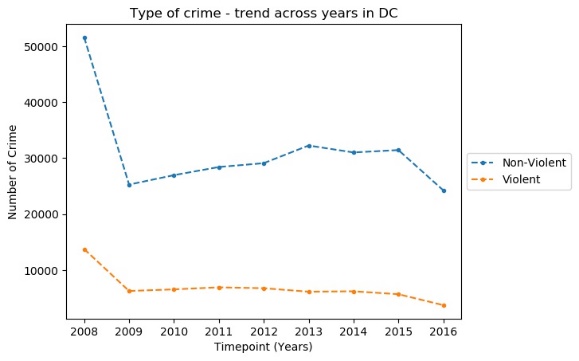
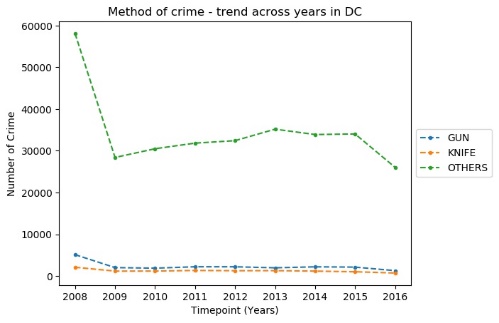
Number of crimes were ground for Method or Type of crime

Crime locations for homicides with Gun

# Patterns

Crime pattern – Type & Method:

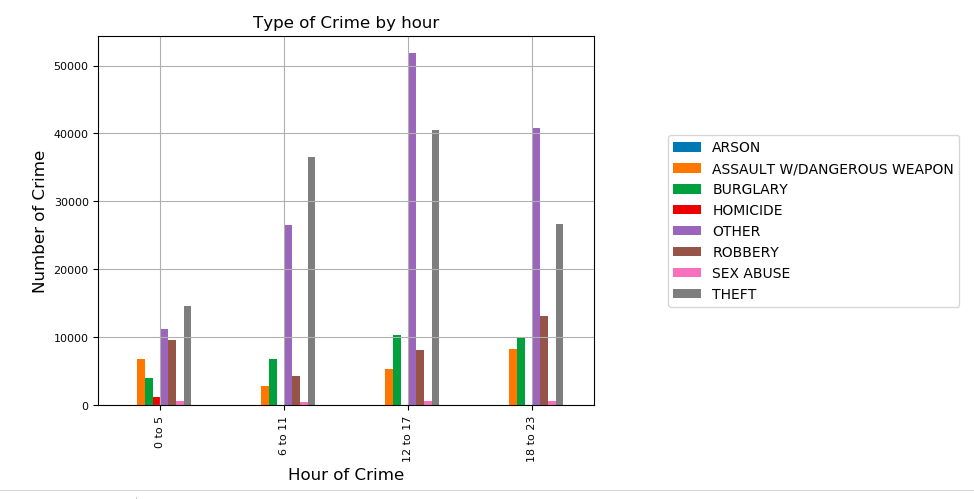
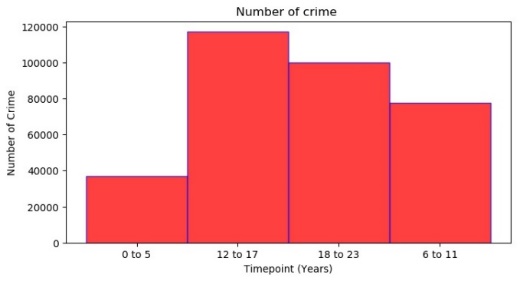
* Relatively less violent crime
* Very high crime rate before year 2009
* Number of crimes stayed almost same for 7 years
* Reducing trend of over all and violent crime starting year 2016

Crime pattern – Time of crime:

* High crime rate is observed during afternoon time, but they are mostly non-violent
* Theft is common and most happening crime throughout the day
* Violent crimes like Homicide happens past midnight but relatively insignificant during other time of the day
* Relatively minimal sexual abuse

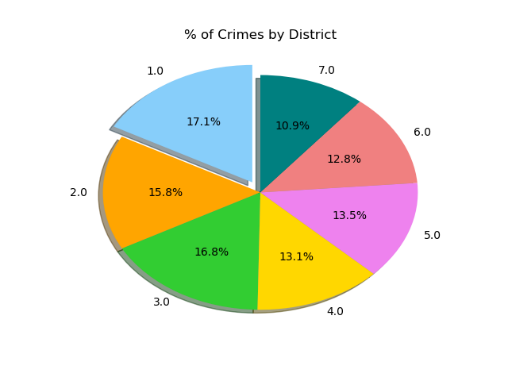
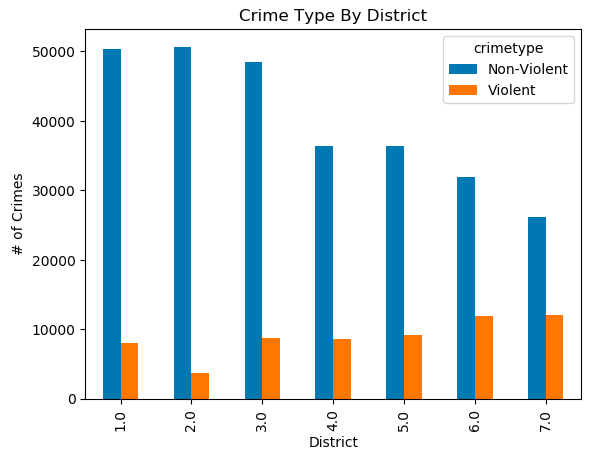
Visuals:

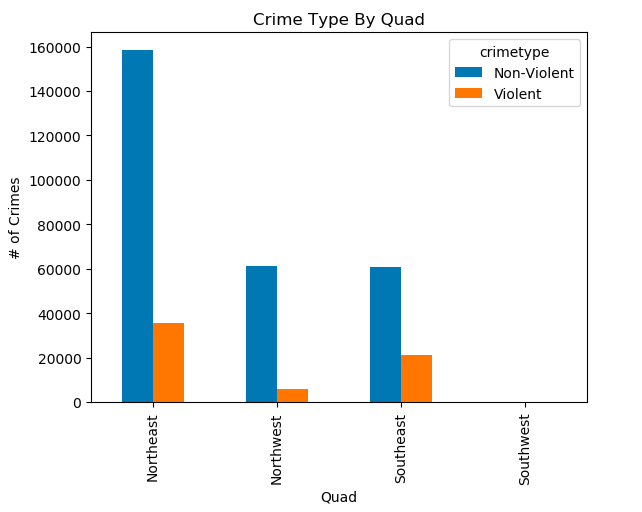
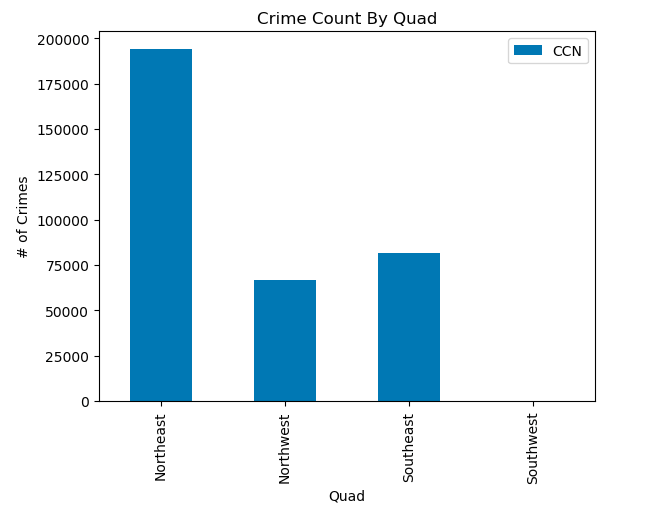


Patterns across districts & quads:

* District 6 & 7 have lesser number of crimes, but most of them is Violent crimes
* Northeast part of DC is the high crime and unsafe Quad
* SouthWest has very minimal crimes and the most safe area in entire DC
* NorthWest, though has significant number of crime, it is mostly non violent

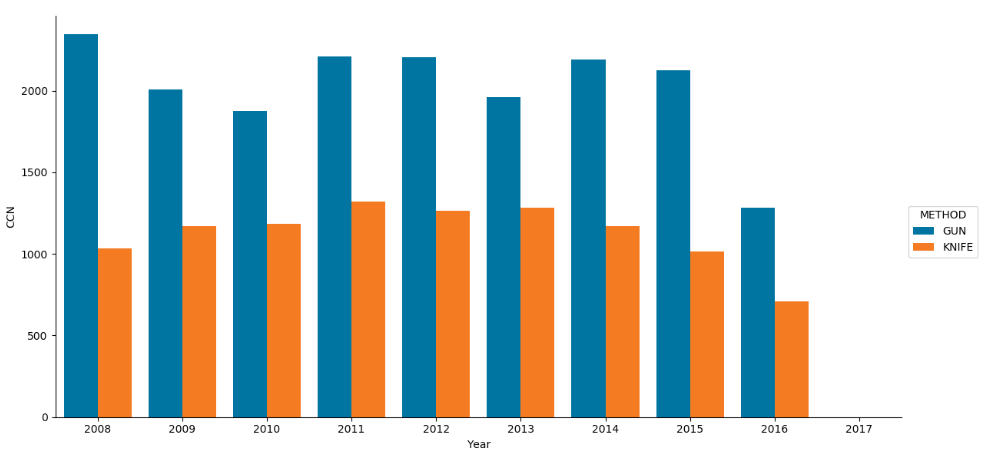
Visuals:

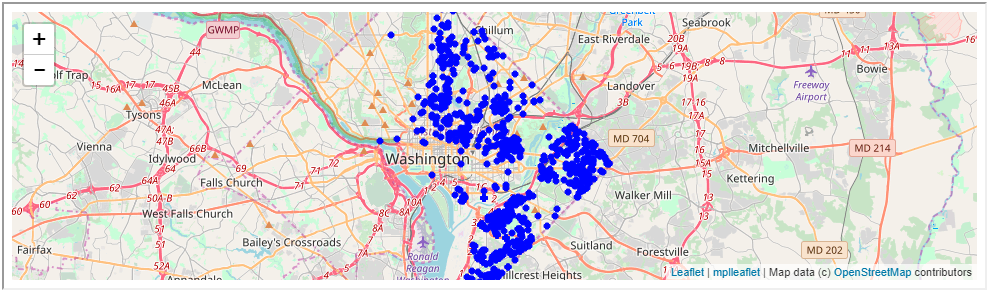
Gun Vs Knife:

* On an average, across years, Gun is the most used weapon in all crimes



Gun Homicide

* Northeast Quad has the most crimes (homicides) with Gun
* Very few crime in Southwest Quad
* Even in high crime Quads, crime happens mostly on specific areas



# Summary & Conclusion:

* High rate of theft/burglary/robbery
* Northeast is the most unsafe Quad in DC
* Southwest is the most safe Quad in DC
* High usage of Gun then knife or any other weapon
* Overall crime is reducing significantly in the past 10 years

Learnings:

* Set coding standards – variable names, where to store output file everything should have been agreed early. It would have helped and saved few hours of work
* Split the work equally among team members
* Prepare who should present which part, before presentation