PES UNIVERSITY, Bangalore
(Established under Karnataka Act
No. 16 of 2013)

UE18CS203

B. Tech , Sem III
Session : Aug-Dec, 2019

UE18CS203 – INTRODUCTION TO DATA SCIENCE

REPORT ON

EXPLORATORY ANALYSIS ON PORTLAND OREGON'S CRIME DATA

ABOUT THE DATASET:

The contents of this data set comes from public data available on the city of Portland website. It describes the crimes reported in the city of Portland Oregon between 1972 and October of 2018. Each individual crime reported lists the location, time and date of the incident as well as the neighborhood in which the event has occurred.

DATA SET SIZE: 202552 rows, 17 columns

ABSTRACT:

The purpose of this analysis is to understand the dataset, perform exploratory data analysis on it, plot visualizations and derive insights from it. On this dataset, we have performed analysis using plots and graphs. Address, case number report month year and few other columns have been neglected because they were way too difficult to analyze due to the lack of consistency in datapoints.

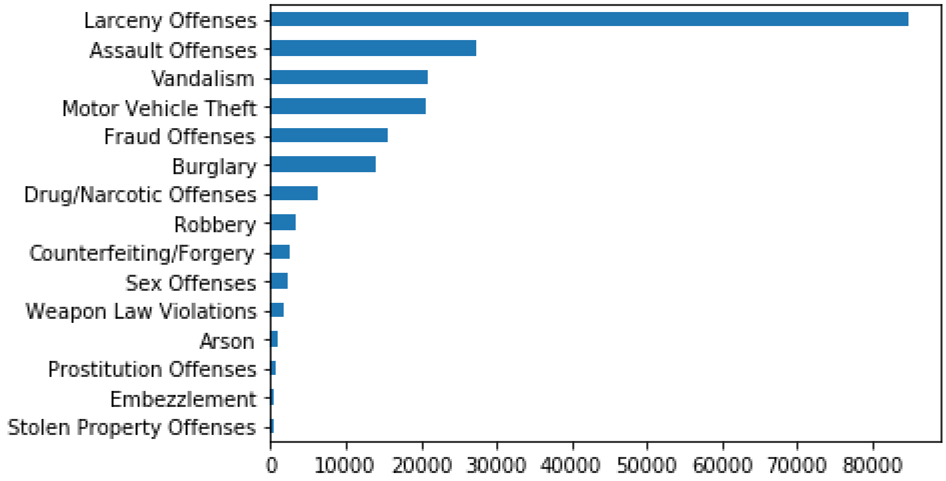
EXPLORATORY ANALYSIS:

1.data cleaning: This dataset had many incorrect, incomplete, inaccurate, and irrelevant parts of data. This has been cleaned, replaced or modified accordingly.
Address and neighborhood columns had 10% and 3% of missing data respectively. OpenDataLat and OpenDataLon values had 11% of missing values. These rows have been skipped since they are categorical values, they cannot be replaced by mean or previous values.Date and time columns have been formatted to their default format that is mm/dd/yyyy and HH:MM.

2.Visualization: Individual columns of the dataset are represented with certain graphs or plots to show the data variation in it. For example,

- ‘Crime Against’ column is visualized using Pie chart.
- Similarly OpenDataX and OpenDataY are the columns which contain discrete numerical values. Their visualization is done with histograms, outliers are estimated using boxplot, and correlation is found using scatterplot.

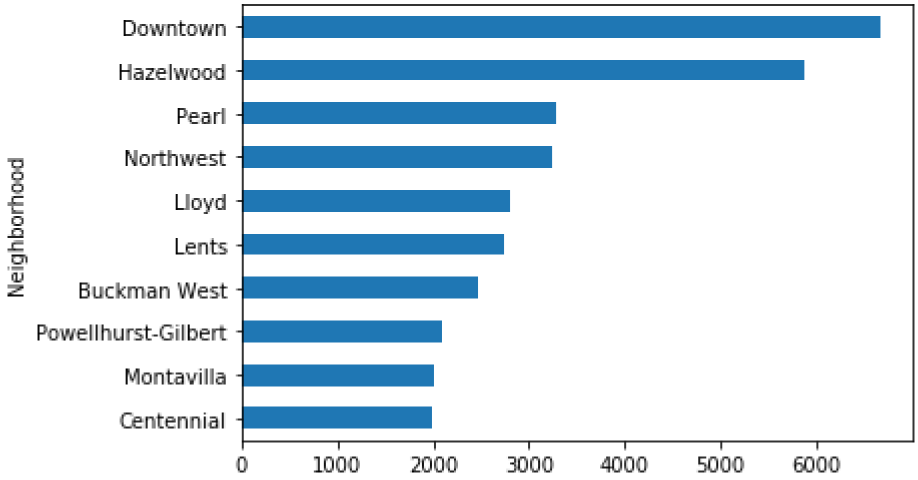
- Report date and occur date has been extracted from respective columns, and their correlation is found using scatterplot.
- Analyzing variation in the occurrence of different types of crime using bar chart, by plotting it against its count.



From the above graph, we can conclude that larceny offenses happen to occur the most in the city.

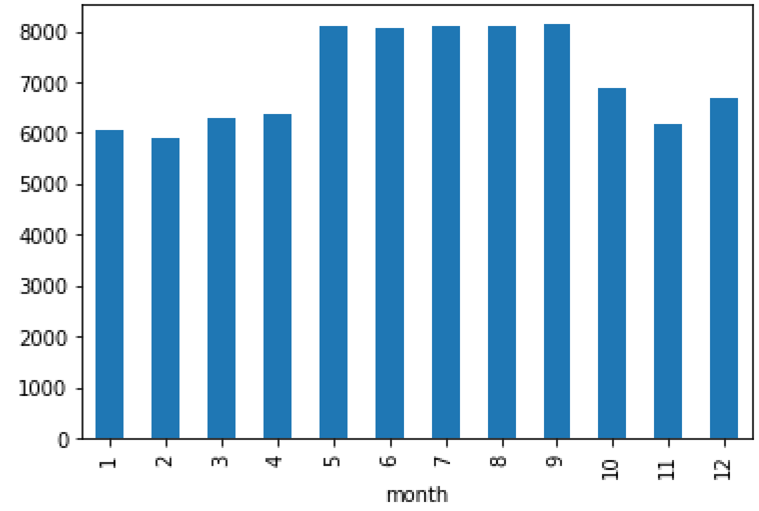
- Analyzing specific offense category:

  As we know, larceny is the maximum occurring offense. We need to locate the neighborhood where most of this offense takes place.



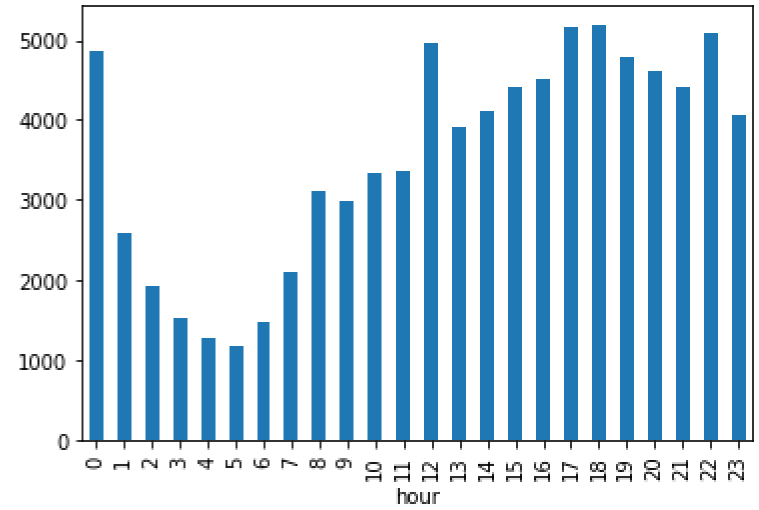
From the above graph, count of larceny offense taking place in particular neighborhood is obtained. And Downtown turns out to be the most dangerous place, with highest number of larceny offense taking place in the city.

- We further continue in finding out the month which is more dangerous, by considering specific offense category.



Still considering larceny offense as the parameter, month when plotted against count of larceny offenses happening in a year, Gives out the conclusion as January, February and November as least dangerous month. Whereas May to September is observed to be more dangerous.

- Moving ahead with time of crime occurrence, we need to find whether it is day or night time that contains more larceny offenses.

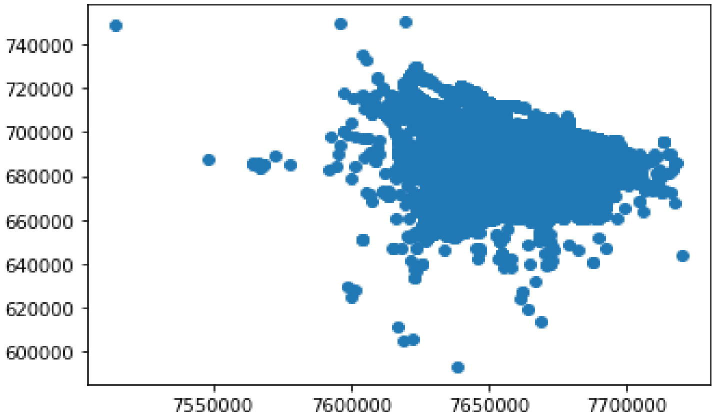


When we plot 24 hours which is being extracted from occur time column from the dataset against larceny crime count, it concludes that 4AM to 6AM is the safest whereas 4PM to 6PM is the time when most of the larceny offenses occur. Hence turns out be the most common time for crime occurance.

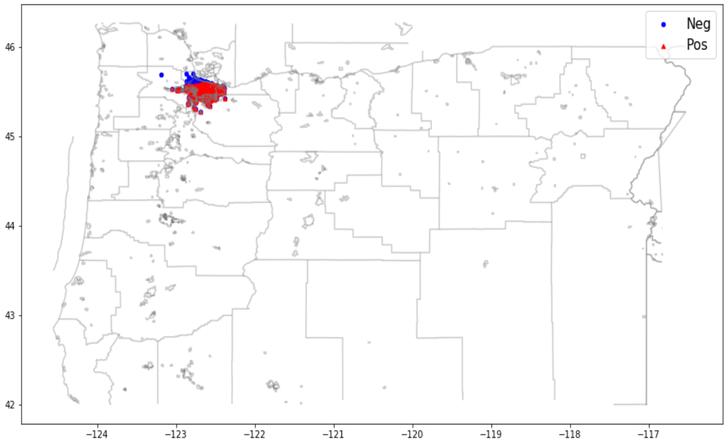
- There are few more graphs which explain which month of the year has highest occurrence of larceny offense considering 24 hours of each day.

- Correlation is found for OpenDataX and OpenDataY using pearson's correlation coefficient, which comes out to be -0.2914 that says there is a negative relationship between two variables.

  Scatterplot given below explains the case in a pictorial way.



- At the end, latitude and longitude given are plotted or marked on a map which is in .shp format.



CONCLUSION:

The exploratory analysis and visualizations what we have carried out on our dataset has helped us to understand how dataset contents are related to each other. With every plot, we came to know about relationship between respective data values. The notorious neighborhood, common time for crime occurrence, most happening offense in the city, month and year which has more crime records,, etc. many such inferences has been drawn from the given dataset.