

DATA ANALYSIS OF CRIME IN BOSTON

21CSS101J – PROGRAMMING FOR PROBLEM SOLVING

Mini Project Report

Submitted by

C DHIYA [Reg. No.: RA2211026010221]

B.Tech. CSE - AI & ML

DAKSHYA T [Reg. No.: RA2211026010250]

B.Tech. CSE - AI & ML



**SCHOOL OF COMPUTING
COLLEGE OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

(Under Section 3 of UGC Act, 1956)

S.R.M. NAGAR, KATTANKULATHUR – 603 203

KANCHEEPURAM DISTRICT

December 2022

TABLE OF CONTENTS

Chapter No.	Title	Page No.
1	Problem Statement	3
2	Methodology / Procedure	4
3	Coding (C or Python)	5
4	Results	7
5	Conclusion	13

PROBLEM STATEMENT

Data Analysis of any dataset using NumPy, Pandas and Matplotlib.

METHODOLOGY/PROCEDURE

We decided to analyse the datasheet of crimes in Boston for the mini project.

We can understand which are the most violent neighbourhoods in Boston, the most common crimes in the area and also what time we can walk safely in the beautiful tourist spots of the city.

Here we can extract some relevant information, such as the most violent neighbourhoods, the most common crimes, the year with the highest number of crimes, the most frequent time, among other things.

CODING IN PYTHON

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import folium
from folium.plugins import HeatMap
%matplotlib inline

data = pd.read_csv('/content/crime.csv', encoding='latin-1')
data.head()

data.isnull().sum()

max_street_crime = data['STREET'].value_counts().index[0]
max_year_crime = data['YEAR'].value_counts().index[0]
max_hour_crime = data['HOUR'].value_counts().index[0]
max_month_crime = data['MONTH'].value_counts().index[0]
max_day_crime = data['DAY_OF_WEEK'].value_counts().index[0]

month = ['January', 'February', 'March', 'April', 'May', 'June', 'July',
         'August', 'September', 'October', 'November', 'December']

print('Street with higher occurrence of crimes:', max_street_crime)
print('Year with highest crime occurrence:', max_year_crime)
print('Hour with highest crime occurrence:', max_hour_crime)
print('Month with highest crime occurrence:', month[max_month_crime-1],
      max_month_crime)
print('Day with highest crime occurrence:', max_day_crime)

plt.subplots(figsize=(15, 6))
sns.countplot('YEAR', data=data, palette='RdYlGn_r', edgecolor=sns.color_p
alette('dark', 7))
plt.xticks(rotation=90)
plt.title('Number Of Crimes Each Year')
plt.show()
plt.subplots(figsize=(15, 6))
sns.countplot('HOUR', data=data, palette='RdYlGn_r', edgecolor=sns.color_p
alette('dark', 7))
plt.xticks(rotation=90)
plt.title('Number Of Crimes Each Hour')
plt.show()

plt.subplots(figsize=(15, 6))
```

```

sns.countplot('OFFENSE_CODE_GROUP',data=data,palette='RdYlGn_r',edgecolor=sns.color_palette('dark',7))
plt.xticks(rotation=90)
plt.title('Types of serious crimes')
plt.show()

```

```

plt.subplots(figsize=(15,6))
sns.countplot('DISTRICT',data=data,palette='RdYlGn',edgecolor=sns.color_palette('Paired',20),order=data['DISTRICT'].value_counts().index)
plt.xticks(rotation=90)
plt.title('Number Of Crimes Activities By District')
plt.show()

```

```

pd.crosstab(data.DISTRICT,data.OFFENSE_CODE_GROUP).plot.barh(stacked=True,width=1,color=sns.color_palette('RdYlGn',5))
fig=plt.gcf()
fig.set_size_inches(12,8)
plt.show()

```

```

pd.crosstab(data.YEAR,data.OFFENSE_CODE_GROUP).plot.barh(stacked=True,width=1,color=sns.color_palette('RdYlGn',5))
fig=plt.gcf()
fig.set_size_inches(12,8)
plt.show()

```

```

G1=data[data['MONTH'].isin(data['MONTH'].value_counts()[1:11].index)]
pd.crosstab(G1.YEAR,G1.DISTRICT).plot(color=sns.color_palette('dark',6))
fig=plt.gcf()
fig.set_size_inches(18,6)
plt.show()

```

RESULTS

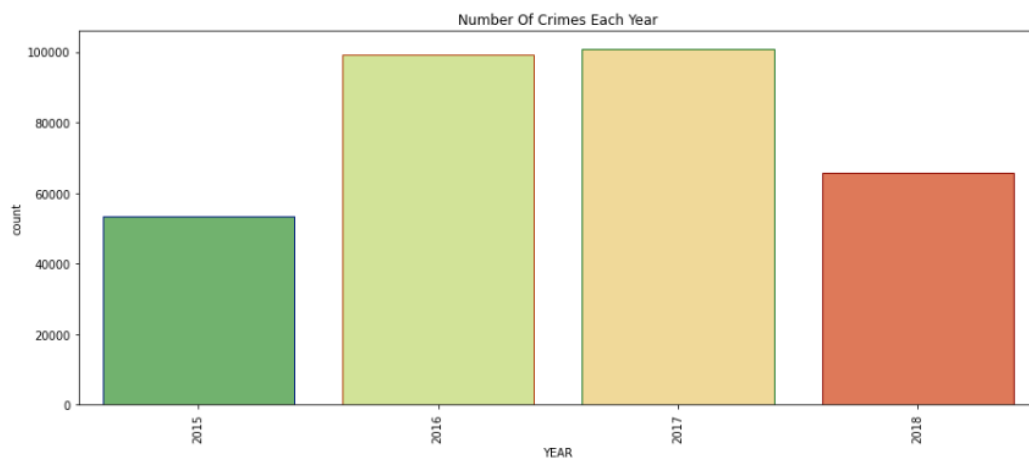
	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR	UCR_PART	STREET	Lat	Long	Location
0	1182070945	619	Larceny	LARCENY-ALL OTHERS	D14	808	NaN	2018-09-02 13:00:00	2018	9	Sunday	13	Part One	LINCOLN ST	42.357791	-71.139371	(42.35779134, -71.13937053)
1	1182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:00:00	2018	8	Tuesday	0	Part Two	HECLA ST	42.306821	-71.060300	(42.30682138, -71.06030035)
2	1182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:27:00	2018	9	Monday	19	Part Three	CAZENOVE ST	42.346589	-71.072429	(42.34658879, -71.07242943)
3	1182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:16:00	2018	9	Monday	21	Part Three	NEWCOMB ST	42.334182	-71.078664	(42.33418175, -71.07866441)
4	1182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:05:00	2018	9	Monday	21	Part Three	DELHI ST	42.275365	-71.090361	(42.27536542, -71.09036101)

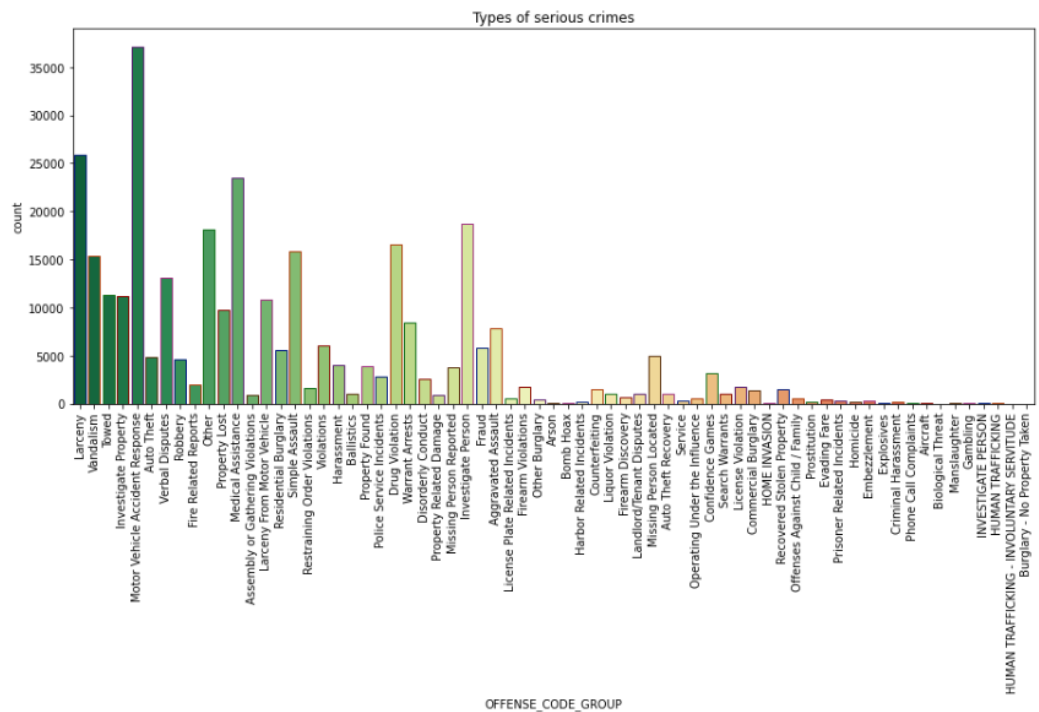
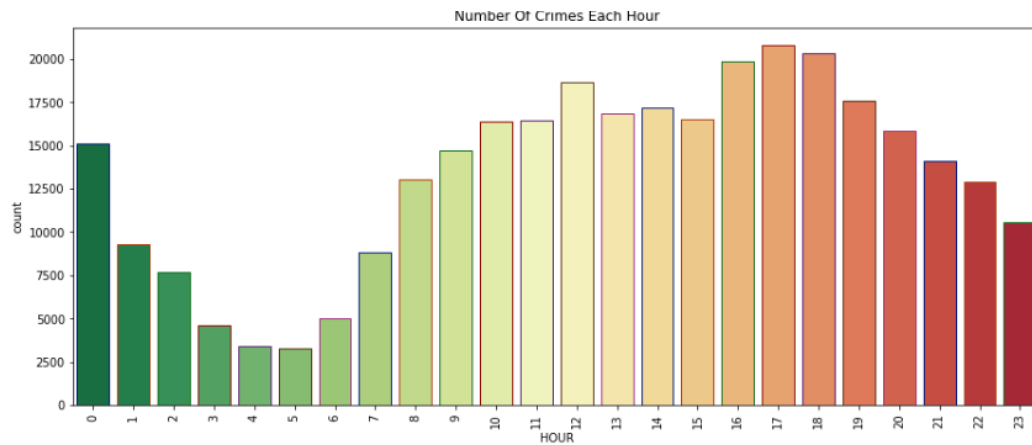
```

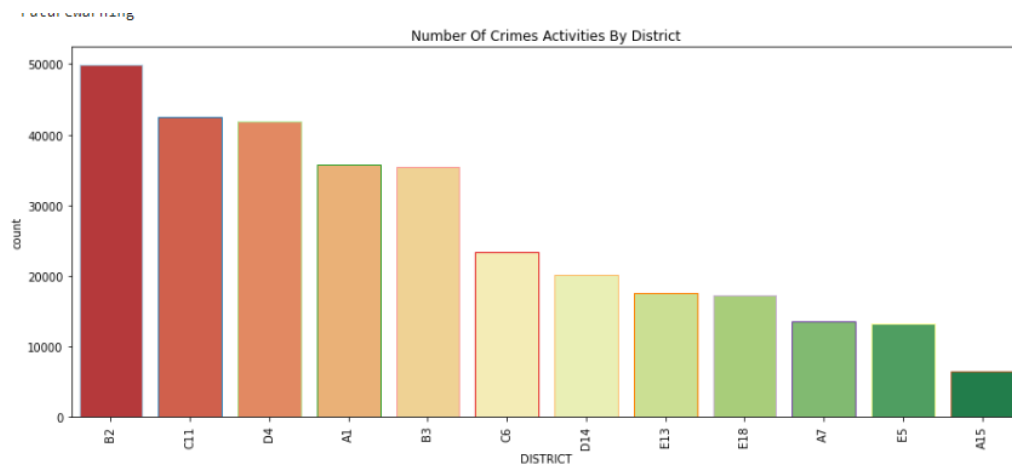
INCIDENT_NUMBER      0
OFFENSE_CODE          0
OFFENSE_CODE_GROUP    0
OFFENSE_DESCRIPTION    0
DISTRICT              1765
REPORTING_AREA        0
SHOOTING              318054
OCCURRED_ON_DATE      0
YEAR                  0
MONTH                 0
DAY_OF_WEEK           0
HOUR                  0
UCR_PART              90
STREET                10871
Lat                   19999
Long                  19999
Location              0
dtype: int64

```

- ✕ Street with higher occurrence of crimes: WASHINGTON ST
- Year with highest crime occurrence: 2017
- Hour with highest crime occurrence: 17
- Month with highest crime occurrence: August 8
- Day with highest crime occurrence: Friday

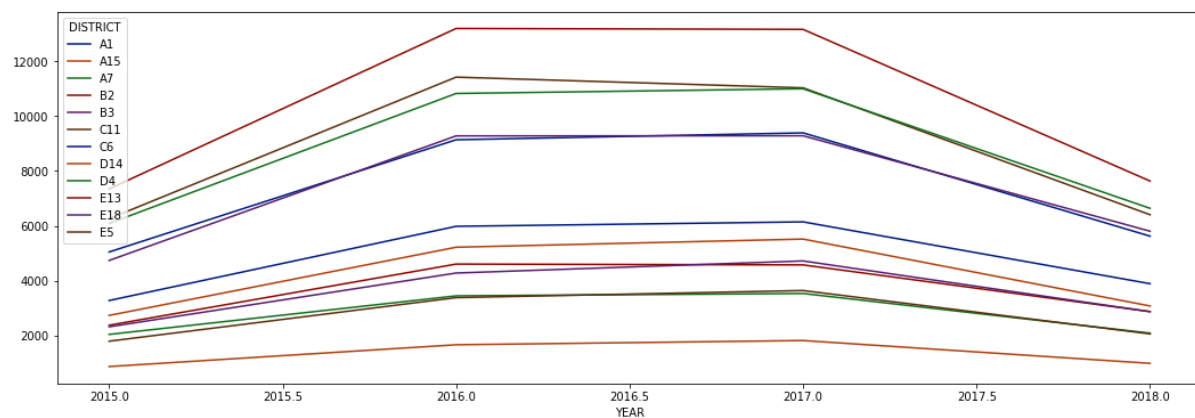












CONCLUSION

EDA is the most important part of any analysis. We will get to know many things about our data. We have tried to show most of the python functions used for exploring the data with visualizations.

From the graph we can see that 2016 and 2017 had the highest incidence of crime in the region, the occurrence of crimes begins to rise from 7am onwards peaking between 5 and 6pm and that the neighbourhood with the highest incidents of crime is B12.

We were able to successfully analyse the data.