RUTGERS

Rutgers Business School
Newark and New Brunswick

DATA ANALYSIS AND VISUALIZATION FINAL PROJECT-

'CRIME ANALYSIS IN BOSTON'

PROFESSOR- Prof. Michail Xyntarakis

PROJECT SUBMITTED BY-

SIDDHANT POKLE (sp1931)

SIDDHARTHA KANTIPUDI (sk2396)

We have selected a dataset which evaluates the crimes done in Boston over the years. The dataset selected will help us to analyse various array of crimes which occurred in Boston from 2015 to 2018. This visualization will help us considerably to plot down:

- Types of crimes most common in Boston
- Frequency of crimes as per days/weeks/years
- Region wise crime analysis

Data sets explored:

For the project, we considered quite a few datasets and shortlisted two of them.

- Dataset from Analyze Boston - Boston Police Department
- Dataset from Massachusetts Crime Statistics

Dataset selected for analysis:

Dataset from Analyze Boston - Boston Police Department

We are selecting this dataset. This dataset fits in perfectly with the objectives of this visualization as it contains records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Crime incident reports are provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond.

The visualization aims to create an interactive dashboard for the BPD officers to track and see the trends of crimes in Boston over the last 4 years. This will help them to analyse, plan and patrol the regions as per the data at hand.

Dataset from Massachusetts Crime Statistics

This dataset consists of data based on the jurisdiction, year, and theme of the crime. The dataset is concrete with its variables, but it is divided as per the themes like violent crimes, DUI/Drugs, Hate crime, arrestees, etc.

This is not in line with our objective as we must considered maximum types of crimes divided region wise. Hence, we ignore this dataset

Variables of Interest & Data cleaning:

These are the variables from our dataset which we are going to consider modelling the visualizations. We have classified the variables and cleaned them.

Variable_name	Variable_type	Comments on variable_type	Description
INCIDENT_NUMBER	Numeric	Looks good	Unique id as per crime
OFFENSE_CODE	Numeric	Looks good	Unique code for types of crime
OFFENSE_CODE_GROUP	Text	Looks good	Descriptions of offenses
OFFENSE_DESCRIPTION	Text	Looks good	Subset of offenses
DISTRICT	Text	Looks good (NULL values are handled)	District codes mentioned
REPORTING_AREA	Numeric	Looks good	Reporting area codes
SHOOTING	Text	This filed will be removed due to many NULL values	N/A
OCCURRED_ON_DATE	Text	Looks good	Time of crime (MM/DD/YYYY) timestamp
YEAR	Numeric	Looks good	Crime Year
MONTH	Numeric	Looks good	Crime Month
DAY_OF_WEEK	Text	Looks good	Crime Day
HOUR	Numeric	Looks good	Crime hour
UCR_PART	Text	Looks good	Uniform Crime reporting as per parts
STREET	Text	Looks good (NULL values are handled)	Location of Street where crime occurred
Lat	Text	Looks good (NULL values are handled)	Latitude value
Long	Float	Looks good (NULL values are handled)	Longitude value
Location	Float	Looks good	Lat_Long pair

EDA

Data size: The data consists of 3,19,074 rows & 17 columns

Number of Crime incidents (2015-2018) : 3,19,074

Types of crimes listed: 576

Number of districts in Boston: 13

Period of Data: 2015 to 2018 (monthly)

Python Code for Data cleaning:

jupyter DAV_Project Last Checkpoint: 5 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Help Trusted Python 3

Run Code

In [2]:
```
# Importing libraries
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
from pylab import rcParams
from pandas.api.types import CategoricalDtype
import warnings
import os
import folium
from folium.plugins import HeatMap
import pandas_profiling
import plotly.express as px
import plotly as py
import plotly.graph_objs as go
import plotly.graph_objs as go
import plotly.tools as tls
import tkinter
import matplotlib
matplotlib.use('TkAgg')
```

In [3]:
```
# Reading the csv files
crime_df = pd.read_csv("C:\\Users\\siddh\\OneDrive\\Documents\\Rutgers_Mita\\Data Analysis & Visualization\\Project\\crime.csv",
offense_df = pd.read_csv("C:\\Users\\siddh\\OneDrive\\Documents\\Rutgers_Mita\\Data Analysis & Visualization\\Project\\offense_c
```

In [4]:
```
crime_df.dtypes
```

Out[4]:
```
INCIDENT_NUMBER         object
OFFENSE_CODE             int64
OFFENSE_CODE_GROUP      object
OFFENSE_DESCRIPTION     object
DISTRICT                object
REPORTING_AREA          object
SHOOTING                object
OCCURRED_ON_DATE        object
YEAR                     int64
MONTH                    int64
DAY_OF_WEEK             object
HOUR                     int64
UCR_PART                object
STREET                  object
Lat                    float64
Long                   float64
Location                object
dtype: object
```

In [5]:
```
offense_df.dtypes
```

Out[5]:
```
CODE     int64
NAME    object
dtype: object
```

In [6]:
```
crime_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319073 entries, 0 to 319072
Data columns (total 17 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   INCIDENT_NUMBER      319073 non-null  object 
 1   OFFENSE_CODE         319073 non-null  int64  
 2   OFFENSE_CODE_GROUP   319073 non-null  object 
 3   OFFENSE_DESCRIPTION  319073 non-null  object 
 4   DISTRICT             317308 non-null  object 
 5   REPORTING_AREA       319073 non-null  object 
 6   SHOOTING             1019 non-null    object 
 7   OCCURRED_ON_DATE     319073 non-null  object 
 8   YEAR                 319073 non-null  int64  
 9   MONTH                319073 non-null  int64  
 10  DAY_OF_WEEK          319073 non-null  object 
 11  HOUR                 319073 non-null  int64  
 12  UCR_PART             318983 non-null  object 
 13  STREET               308202 non-null  object 
 14  Lat                  299074 non-null  float64
 15  Long                 299074 non-null  float64
 16  Location             319073 non-null  object 
dtypes: float64(2), int64(4), object(11)
memory usage: 41.4+ MB
```

In [7]:
```
offense_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 576 entries, 0 to 575
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   CODE    576 non-null    int64 
 1   NAME    576 non-null    object
dtypes: int64(1), object(1)
memory usage: 9.1+ KB
```

In [8]:
```
crime_df.describe().T
```

Out[8]:

	count	mean	std	min	25%	50%	75%	max
OFFENSE_CODE	319073.0	2317.546956	1185.285543	111.000000	1001.000000	2907.000000	3201.000000	3831.000000
YEAR	319073.0	2016.560586	0.996344	2015.000000	2016.000000	2017.000000	2017.000000	2018.000000
MONTH	319073.0	6.609719	3.273691	1.000000	4.000000	7.000000	9.000000	12.000000
HOUR	319073.0	13.118205	6.294205	0.000000	9.000000	14.000000	18.000000	23.000000
Lat	299074.0	42.214381	2.159766	-1.000000	42.297442	42.325538	42.348624	42.395042
Long	299074.0	-70.908272	3.493618	-71.178674	-71.097135	-71.077524	-71.062467	-1.000000

In [9]:
```
offense_df.describe().T
```

Out[9]:

	count	mean	std	min	25%	50%	75%	max
CODE	576.0	1727.970486	1163.050098	111.0	542.0	1768.0	2900.25	3831.0

In [10]:
```
crime_df.head()
```

Out[10]:

T	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR	UCR_PART	STREET	Lat	Long	Location
4	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	347	NaN	2018-08-21 00:00:00	2018	8	Tuesday	0	Part Two	HECLA ST	42.306821	-71.060300	(42.30682138, -71.06030035)
4	151	NaN	2018-09-03 19:27:00	2018	9	Monday	19	Part Three	CAZENOVE ST	42.346589	-71.072429	(42.34658879, -71.07242943)
4	272	NaN	2018-09-03 21:16:00	2018	9	Monday	21	Part Three	NEWCOMB ST	42.334182	-71.078664	(42.33418175, -71.07866441)
3	421	NaN	2018-09-03 21:05:00	2018	9	Monday	21	Part Three	DELHI ST	42.275365	-71.090361	(42.27536542, -71.09036101)

In [11]:
```
crime_df.shape
```

Out[11]:
```
(319073, 17)
```

In [12]:
```
crime_df.head()
```

Out[12]:

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:00:00
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:00:00
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:27:00
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:16:00
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:05:00

In [17]:
```
crime_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 319073 entries, 0 to 319072
Data columns (total 22 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   INCIDENT_NUMBER  319073 non-null  object
 1   OFFENSE_CODE     319073 non-null  int64
 2   Group            319073 non-null  category
 3   Description      319073 non-null  category
 4   District         317308 non-null  category
 5   REPORTING_AREA   319073 non-null  object
 6   Shooting         1019 non-null    object
 7   Date             319073 non-null  datetime64[ns]
 8   Year             319073 non-null  int64
 9   Month            319073 non-null  int64
 10  Day              319073 non-null  category
 11  Hour             319073 non-null  int64
 12  UCR_PART         318983 non-null  category
 13  Street           308202 non-null  object
 14  Lat              299074 non-null  float64
 15  Long             299074 non-null  float64
 16  Location         319073 non-null  object
 17  dayofweek        319073 non-null  category
 18  quarter          319073 non-null  category
 19  dayofyear        319073 non-null  category
 20  dayofmonth       319073 non-null  category
 21  weekofyear       319073 non-null  int64
dtypes: category(9), datetime64[ns](1), float64(2), int64(5), object(5)
memory usage: 37.5+ MB
```

In [18]:
```
crime_df.head()
```

Out[18]:

	INCIDENT_NUMBER	OFFENSE_CODE	Group	Description	District	REPORTING_AREA	Shooting	Date	Year	Month	...	UCR_PART	Street
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:00:00	2018	9	...	Part One	LINCOLN ST
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:00:00	2018	8	...	Part Two	HECLA ST
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:27:00	2018	9	...	Part Three	CAZENOVE ST
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:16:00	2018	9	...	Part Three	NEWCOMB ST
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:05:00	2018	9	...	Part Three	DELHI ST

5 rows × 22 columns

In [19]:
```
crime_df.isnull().sum()
```

Out[19]:
```
INCIDENT_NUMBER      0
OFFENSE_CODE         0
Group                0
Description          0
District          1765
REPORTING_AREA       0
Shooting        318054
Date                 0
Year                 0
Month                0
Day                  0
Hour                 0
UCR_PART            90
Street           10871
Lat              19999
Long             19999
Location             0
dayofweek            0
quarter              0
dayofyear            0
dayofmonth           0
weekofyear           0
dtype: int64
```

In [20]:
```
crime_df = crime_df.drop(columns='Shooting')
crime_df = crime_df.dropna(axis=0)
print(crime_df.isnull().sum(),"\nShape:",crime_df.shape)
```

```
INCIDENT_NUMBER    0
OFFENSE_CODE       0
Group              0
Description        0
District           0
REPORTING_AREA     0
Date               0
Year               0
Month              0
Day                0
Hour               0
UCR_PART           0
Street             0
Lat                0
Long               0
Location           0
dayofweek          0
quarter            0
dayofyear          0
dayofmonth         0
weekofyear         0
dtype: int64
Shape: (296573, 21)
```

In [21]:
```
crime_df.describe()
```

Out[21]:

	OFFENSE_CODE	Year	Month	Hour	Lat	Long	weekofyear
count	296573.000000	296573.000000	296573.000000	296573.000000	296573.000000	296573.000000	296573.000000
mean	2293.856737	2016.550954	6.611856	13.124576	42.300079	-71.046943	27.070617
std	1182.909512	1.001118	3.279314	6.278383	0.981040	1.586477	14.359695
min	111.000000	2015.000000	1.000000	0.000000	-1.000000	-71.178674	1.000000
25%	802.000000	2016.000000	4.000000	9.000000	42.297521	-71.097223	15.000000
50%	2907.000000	2017.000000	7.000000	14.000000	42.325574	-71.077562	28.000000
75%	3201.000000	2017.000000	9.000000	18.000000	42.348624	-71.062529	39.000000
max	3831.000000	2018.000000	12.000000	23.000000	42.395042	-1.000000	53.000000

Merging Datasets

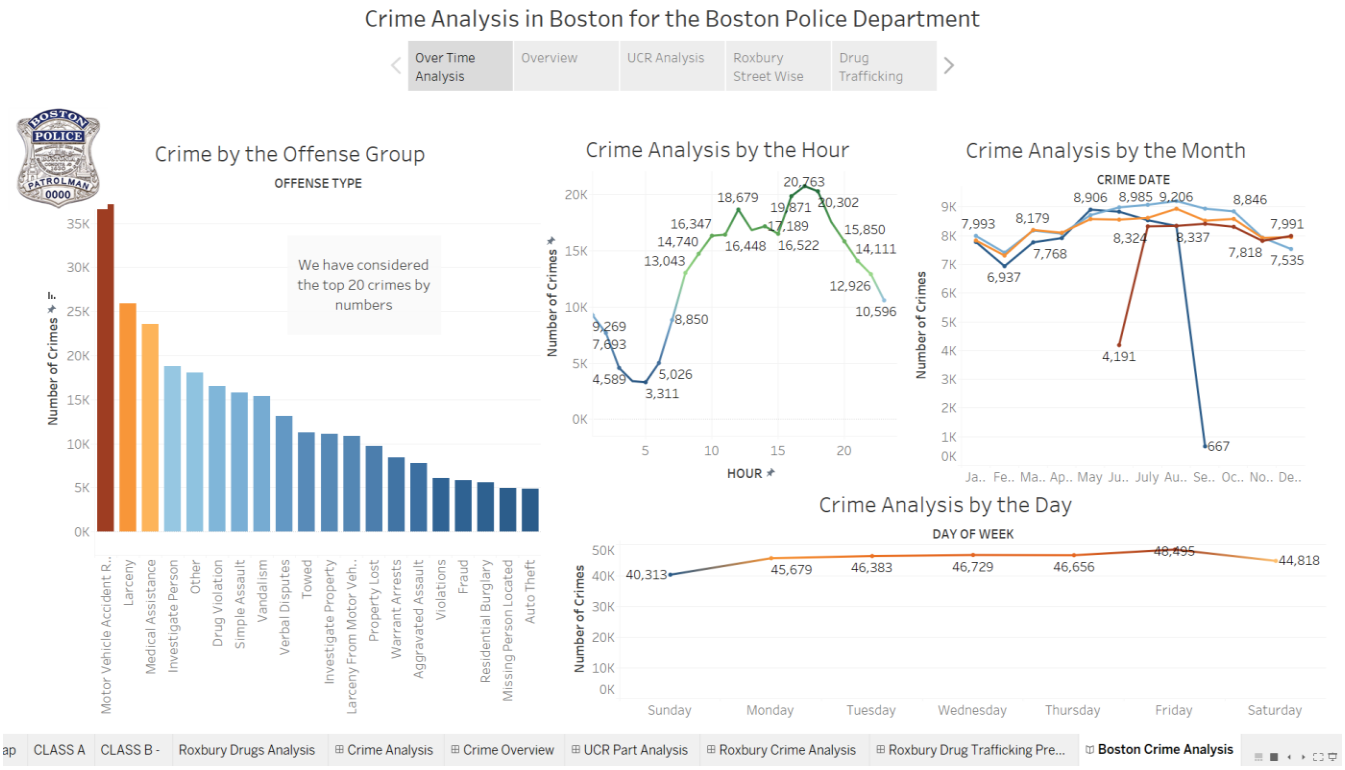
We have merged two datasets from the source Analyze Boston , crime.csv & offence_codes.csv.

This dataset merging provides district wise , section wise crime classification which helps to give additional insights.

Visualization Analysis

Fig 1:

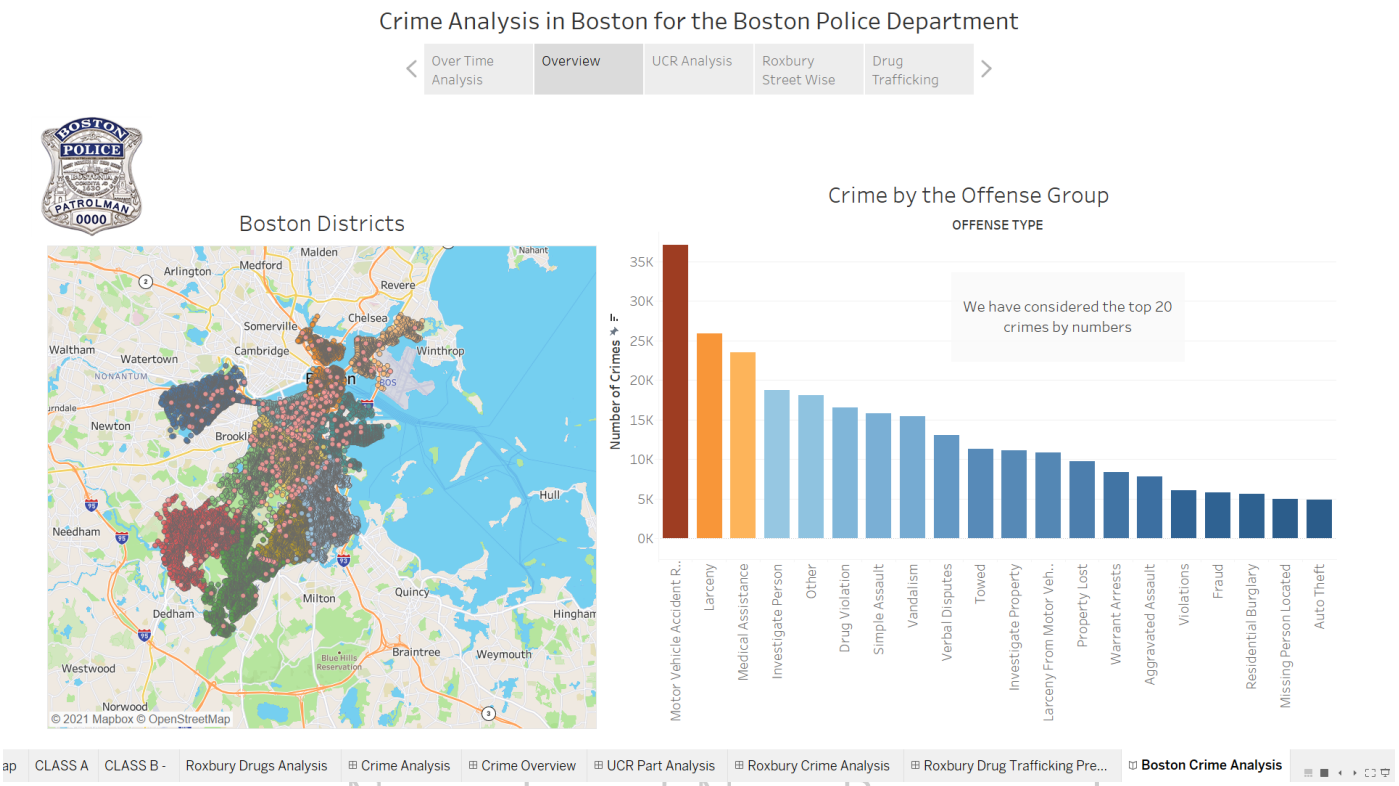
Types of crime in terms of hours, days and months



The above is an interactive visualization between the type of crime and its analysis in monthly, hourly, and weekly occurrences. We can infer the trends in the number of crimes occurred hourly, weekly, and monthly to understand the frequency of crimes and what time and place they take place. It also provides a monthly representation of number of crimes committed and identify those months portray high criminal activity.

Fig 2 :

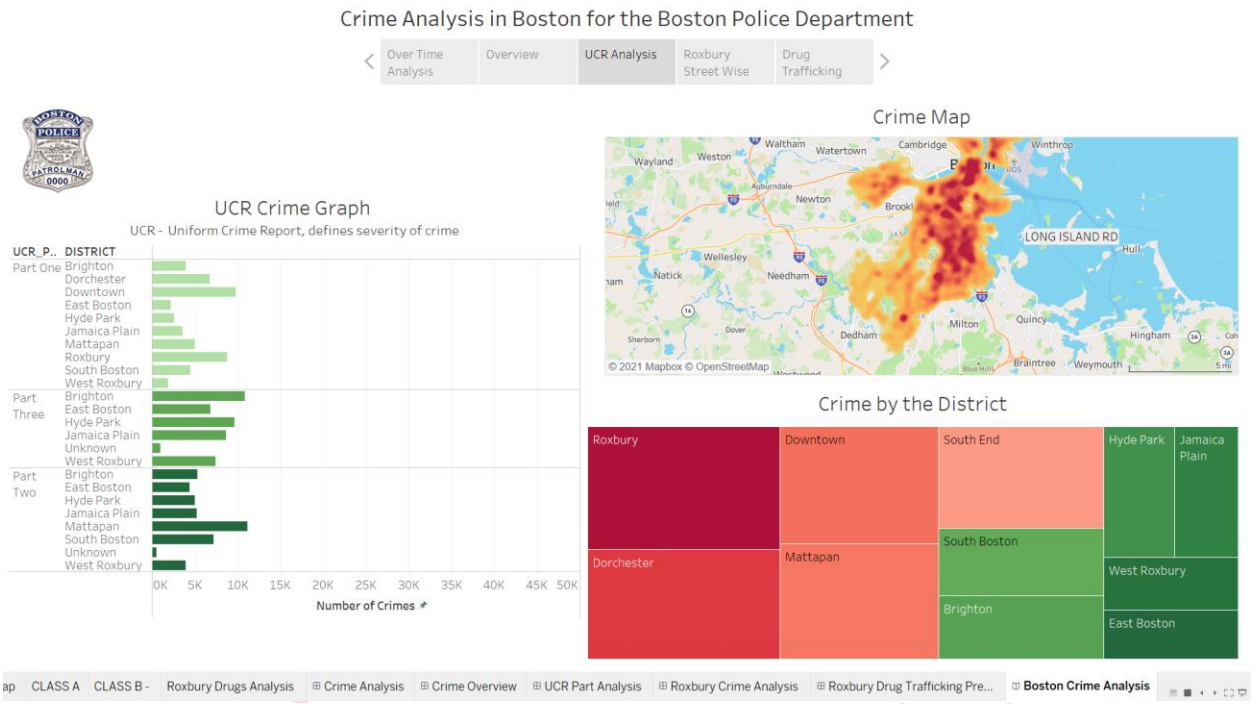
Overview of crime of



The above visualization depicts the overview of the types of crimes commited which are plotted on a map with the boston districts. You can on a macro level could see the distribution of the crime types across the districts represented by their individual colors. We can indeed infer certain types of crimes occur in specific geographical boundaries.

Fig 3:

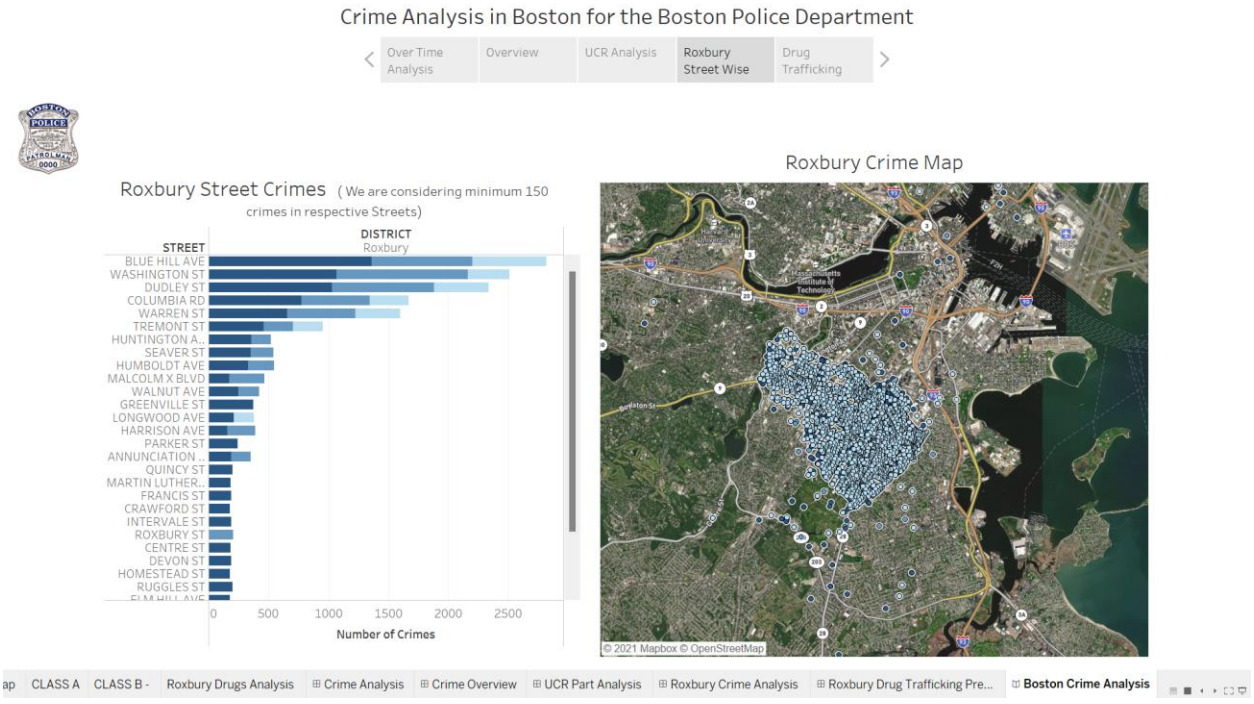
District wise heat map plotting for number of crimes



The above visualization is a heatmap of UCR crime graph (uniform crime report) which defines the severity of the crime. The heat map shows us the district wise segregation and shows those districts with highest number of crimes with respect to the severity of it. The tile sizes correspond to the number of crimes. The map itself shows a district and streetwise representation of the heatmap once zoomed in.

Fig 4

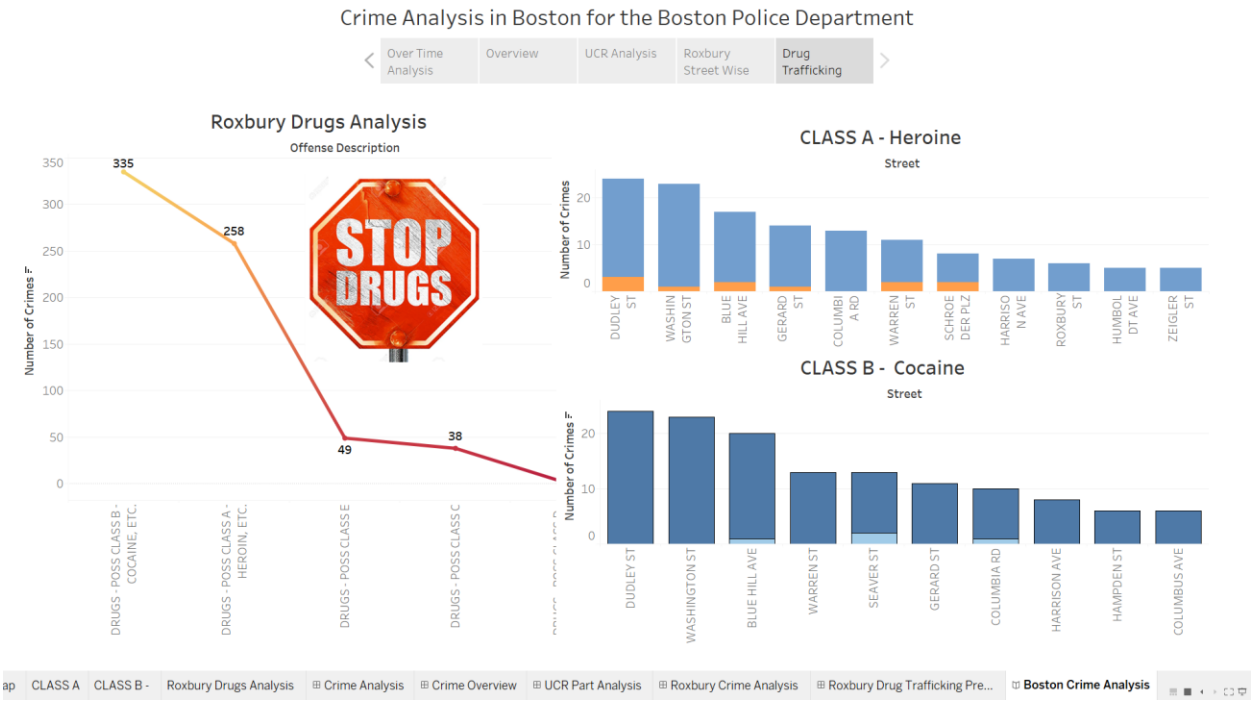
Roxbury Street wise crime analysis



We take a deeper look into the district Roxbury which was identified as the most vulnerable district with respect to the crimes committed. We have a stack bar chart with respect to the UCR bifurcation of severity of crimes. These stacked bar charts are plotted with respect to the street in that district and the corresponding plots on the map.

Fig 5

Drug Trafficking in Roxbury



The above visualization is with regards to the drug related crimes. We can infer that class A drugs have more crimes committed compared to other classes. The stacked bar charts also show street wise drug related crimes of heroine and cocaine with respect to trafficking and otherwise.

Conclusion & recommendations

A city's crime rate has always been a factor that ultimately affects home values, as many potential homebuyers aren't interested in living someplace where they may fall victim to a crime. Boston was ranked 54th nationally in the US and came in the bottom half of the list for home & community safety.

As per the FBI crime data, Boston is a very complicated community where crimes occur and aren't going down

As per our visualizations, we can observe that how maximum number of crimes occur during the peak hours of the day and mostly on Fridays. Month of August has seen the greatest number of crimes in Boston which can help the Boston Police department (BPD) to plan better as per the segregated data of crimes occurring in districts. The Street wise division of crime occurrence will be very handy for BPD patrolling schedules, active responsiveness teams and lowering of crimes in the cities.

UCR analysis also gives us the bifurcation of which crimes to prioritize and how region wise crime numbers enhance the safety protocols in districts. As per the UCR (uniform crime report) graphs and analysis we can observe that Roxbury is the district with highest number of crimes and is on a constant rise yearly.

To deep dive into the RCA (root cause analysis) of why Roxbury has the greatest number of crimes in Boston, we decided to analyse the data much further.

We could conclude from our visualizations and crisp analysis that Drug trafficking in Roxbury can one of the very strong reasons that the crimes are on a rise. We were able to see how drug trafficking may have an impact on suicides and accidents in the district and why it is crucial to stop extreme drug intake.

We are putting forward a recommendation to the BPD that they should take a district wide initiative in Boston to make people (and specially the youth) aware that how drugs are having a devastating impact on the lives of people. Drug sales and tracking down the production of high intensity drugs should be a priority.

If the BPD can control down the drug intake and quick tracking of crime occurring trends mentioned above, as per our analysis the crime rates will go down and Boston will march ahead to become one of the safest cities in the US

To conclude,

Data visualization technology can bring benefits to almost everyone within the police force – and benefits extend into the wider community. The more police can see the big picture, the higher the likelihood of crime getting solved quickly.

Thank you!