# FINAL PROJECT OVERVIEW – 4
# AIT-664 (003)

# BOSTON CRIME
# GROUP 4

**HARITHA TUMMANAPALLY (G01197018)**
**SAISUDHA SURYANARAYANAN (G01228903)**
**SANJANA SRAVYA NAGULAPATI (G01245458)**

**INTRODUCTION:**

Boston is one of the most beautiful and highly populated cities in the United States of America. Safety is something every family looks for when relocating from one city to another. It is all about reducing or preventing crime which will help build strong and vibrant communities. So, how safe is Boston when compared to other cities around the USA? In order to find that, we have taken the crime dataset provided by Boston Police Department (BPD).

They have taken an initiative to improve the city by publishing the data to common people. Similar to the case in any significant city, crimes can be an issue. With the whole tourist destinations, exhibitions, and landmarks, it's ideal to not stroll around with a camera. One should be very careful while they are alone while walking on the streets when it's darker outside. Crimes are most likely to happen when it is dark because of obvious reasons.

**PROBLEM STATEMENT:**

Our main analysis aim is to see if Bostin is a safer place compared to other cities in the USA by analyzing the crime percentage in Boston through investigating different cities in Massachusetts. our analysis can help the Boston police to reduce the crimes that are frequently occurring by exploring some key questions that we have come up with which are related to Affected districts, districts with more crime, the time at which most crimes occur and unemployment rate which can also be helpful to see the most affordable districts depending upon the unemployment rate. To achieve solutions to these analyses it completely depends on the data, whether the data has null values or not. So first these values must be removed if there are any. This also improves the dataset so that it can be more helpful and useful for future use.

**DATASET DESCRIPTION:**

This data set is about the crime rate in Boston and it's been taken from Kaggle. It contains following attributes:

| **ATTRIBUTES** | **DESCRIPTION** |
|---|---|
| INCIDENT NUMBER | Number to identify the incidents |
| OFFENSE_CODE | It is the number that implies the crime is belonging to a particular type |
| OFFENSE_CODE_GROUP | Type of crimes |
| OFFENSE_DESCRIPTION | Elaborates the crime |
| DISTRICT | The district the crime has occurred in and the dataset contains 12 districts |

| | |
|---|---|
| REPORTING_AREA | The area in which crime has occurred |
| SHOOTING | Whether shooting has happened or not |
| OCCURED_ON_DATE | This column contains the date and time with seconds of the crime occurred |
| YEAR | Year of crime |
| MONTH | Month of crime |
| DAY_OF_THE_WEEK | Day on which crime occurred |
| UCR_PART | UCR- Uniform Crime Reporting Program which divides the crime into 3 parts |
| STREET | Name of the street |
| HOUR | Hour using 24hrs clock |
| LONGITUDE | Shows the longitude |
| LATITUDE | Shows the latitude |
| LOCATION | Using both longitude and latitude identifies the exact location of the crime occurred |

The file is 78.5 Mb the data set contains 17 columns with more than 75000 rows. It has main details regarding the location. One major thing to notice is that the districts given in codes which will be hard to analyse. So, the first step will be to recode the district codes to district names. The dataset is been divided into two types: Categorical and Numerical

Categorical - street, ucr_part, day_of_week, offense_description, offense_code_group, incident_number, district

Numerical- offense_code, reporting_area, shooting, year, month, hour, latitude, longitude, location, occurred_on_date.

**INITIAL HYPOTHESIS:**

When compared to other cities in Massachusetts and all across the US, you will see that Boston is more secure than 8% of cities in Massachusetts. Furthermore, Boston is more secure than 15% of cities in the whole US.

To accomplish these outcomes, cities in Massachusetts and the US were investigated to see the number of cities that can have a lower crime percentage in Boston. By studying this, they were able to verify that Boston is below normal in safety when compared and different urban communities in the United states. Overall, the crime rate in Boston is greater than the US average crime rate. These crimes mostly are caused because of poverty, drugs, theft, unemployment, etc.

After analyzing the data, we have come up with a few key questions which have to be answered:

***1) What are the 5 highly affected districts?***

***2) What is the average Boston Unemployment rate?***

***3) Do crimes occur in any particular months?***

***4) When does more crimes tend to occur?***

***5) Does the rent matter in accordance with the unemployment rate? Which district is affordable?***

***6) Explain the trends and pattern using Year, Month, Day, Hour and District***

**FOCUS AREA:**

Some of the main things to concentrate are:

1) Making it easy to analyze, we are going to change the district codes to district names.
2) We will identify the total number of crimes in each district which will help us find the highest and lowest crime districts.
3) We will be able to plot the districts with high and low frequency of incidents.
4) We will be able to plot the districts from 2015 to 2020 with the crimes.

So, the variables that we will be mainly focusing on for the initial analysis will be DISTRICTS, Total number of crimes occurred and OFFENSE_CODE_GROUP. For plotting districts with year, the variables are YEAR and DISTRICTS.

Also, there are some missing values in the dataset, and they are as dangerous as inaccurate data and might lead to inconsistent results in the analysis of data, hence we have to clean the dataset first.
We collected our dataset from all the references provided below and we are hoping it is detailed as it gives us the information like at what exact times and days the crime was committed, locations of the crime committed as it is one of the crucial aspects of analysis in the project.

**DATA PREPARATION:**

One of the main aims of data preparation is to make sure the precision and consistency of raw data being prepared for data processing and analysis, so that the outcomes of BI and analytics applications are valid. Data is commonly generated which includes the missing values, errors, etc. Furthermore, different data sets also have various formats that need to be accommodated. A major part of the data processing process is the correction of data errors, inspection of data quality and the joining of data sets.

So as to guarantee that they have the data that experts or business clients are looking for, data planning regularly requires identifying appropriate data to be used in analytics applications. The data can also be improved and optimized to make it more useful.

The data set contains more than 47k rows and 14 columns. While cleaning the data out of all types of crimes there were 5 crimes that had missing value, they are shooting, UCR part, street, latitude, longitude. These columns have been dropped which led to the dataset that had zero null values. New columns have been added to the dataset to enhance the analysis.

Here are the new variables present in the data:

**POPULATION IN DISTRICT:** population in districts of Boston.

**UNEMPLOYMENT RATE:** unemployment rate in each district.

**AVERAGE RENT PER NIGHT:** average rent in each district of Boston.

**TRANSPORT 1:** the first mode of transport a person could use in accordance with his/her earning i.e., car if he earns more and walk if his earnings are less

**TRANSPORT 2:** the second mode of transport i.e., bus/train

**TRANSPORT 3:** the third mode of transport i.e., bus/train

**TRANSPORT 4:** the least mode of transport i.e., walk/car depending on individual earnings

**TOOLS USED:**

For this analysis, the tool used are PYTHON and TABLEAU.

Python is used for analyzing the dataset, finding the missing values, dropping the columns with missing values, it shows the statistical details. Python has been used for visualizing as well.

For this analysis, Tableau helped compare in a single visualization the trends & patterns on how Year, Month, Day, Hour and District are related.

**DATA CLEANING:**

After processing the data, the dataset has no errors which emphasizes that there are no missing values. But, while cleaning the data, date and time column is replaced by 2 different columns respectively. Since the data is clean there is no need to replace the values or change the null values. Here is the summary of the data after cleaning the data. There are no null values in the dataset.

```
In [300]: #concise summary of the dataframe
          df.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 477612 entries, 0 to 477611
          Data columns (total 27 columns):
          INCIDENT_NUMBER                          477612 non-null object
          OFFENSE_CODE                             477612 non-null int64
          OFFENSE_CODE_GROUP                       477612 non-null object
          OFFENSE_DESCRIPTION                      477612 non-null object
          DISTRICT                                 477612 non-null object
          DISTRICT_NAME                            477612 non-null object
          REPORTING_AREA                           477612 non-null object
          SHOOTING                                 52519 non-null object
          OCCURRED_ON_DATE                         477612 non-null datetime64[ns]
          Date                                     477612 non-null datetime64[ns]
          Time                                     477612 non-null object
          YEAR                                     477612 non-null int64
          MONTH                                    477612 non-null int64
          DAY_OF_WEEK                              477612 non-null object
          HOUR                                     477612 non-null int64
          UCR_PART                                 426729 non-null object
          STREET                                   460683 non-null object
          Lat                                      449404 non-null float64
          Long                                     449404 non-null float64
          Location                                 477612 non-null object
          Population_in_District_Demographics      477612 non-null int64
          Unemployment_Rate_in_District            477612 non-null float64
          Average_rent_per_night                   477612 non-null int64
          Transport _1                             477612 non-null object
```

Out of all types of columns there were 5 columns that had missing values, they are shooting, ucr part, street, latitude, longitude [out 311].

```
Out[311]: INCIDENT_NUMBER                          0
          OFFENSE_CODE                             0
          OFFENSE_CODE_GROUP                       0
          OFFENSE_DESCRIPTION                      0
          DISTRICT                                 0
          DISTRICT_NAME                            0
          REPORTING_AREA                           0
          SHOOTING                            425093
          OCCURRED_ON_DATE                         0
          Date                                     0
          Time                                     0
          YEAR                                     0
          MONTH                                    0
          DAY_OF_WEEK                              0
          HOUR                                     0
          UCR_PART                             50883
          STREET                               16929
          Lat                                  28208
          Long                                 28208
          Location                                 0
          Population_in_District_Demographics      0
          Unemployment_Rate_in_District            0
          Average_rent_per_night                   0
          Transport _1                             0
          Transport _2                             0
          Transport _3                             0
          Transport _4                             0
          dtype: int64
```

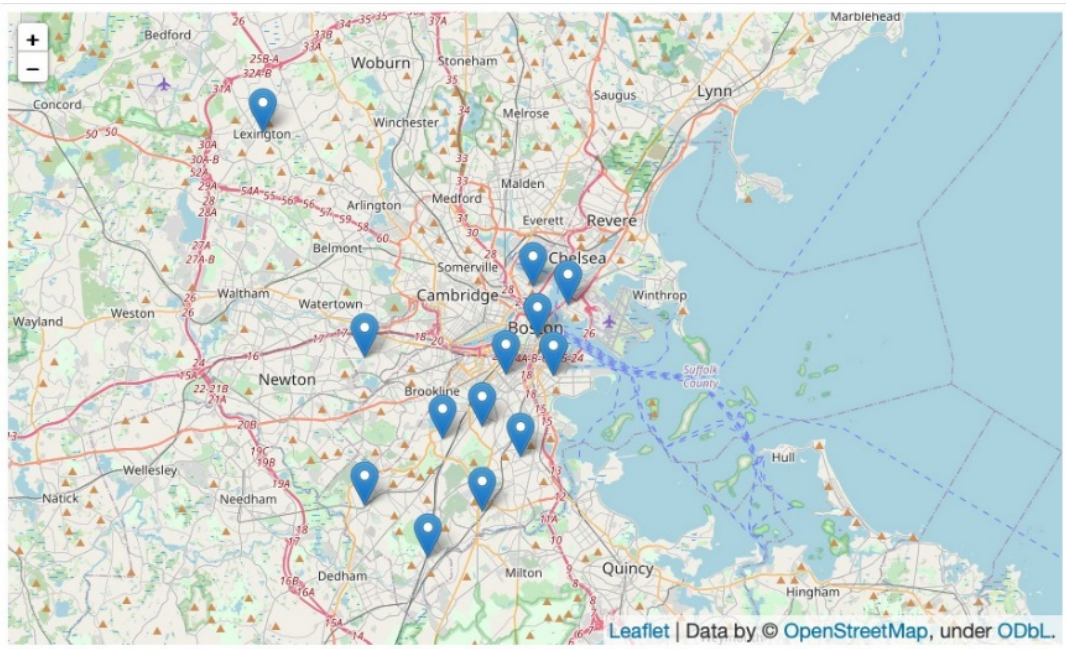
The columns with missing values are dropped which leads to the result [0ut 313].

```
Out[313]: INCIDENT_NUMBER                          0
          OFFENSE_CODE                             0
          OFFENSE_CODE_GROUP                       0
          OFFENSE_DESCRIPTION                      0
          DISTRICT                                 0
          DISTRICT_NAME                            0
          REPORTING_AREA                           0
          OCCURRED_ON_DATE                         0
          Date                                     0
          Time                                     0
          MONTH                                    0
          DAY_OF_WEEK                              0
          HOUR                                     0
          Location                                 0
          Population_in_District_Demographics      0
          Unemployment_Rate_in_District            0
          Average_rent_per_night                   0
          Transport _1                             0
          Transport _2                             0
          Transport _3                             0
          Transport _4                             0
          dtype: int64
```

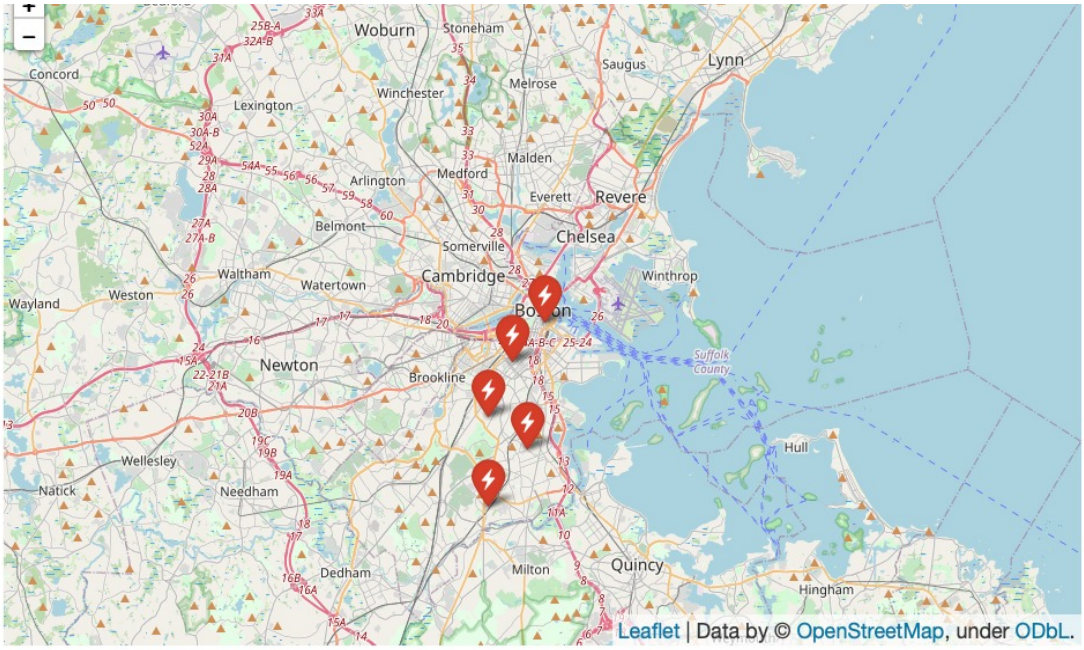
**DATA VISUALIZATION:**

Data visualization is a graphical representation of data where data is placed in a visual context. So that outliers can be detected from patterns, trends and correlations. Some of the visual plots which can be used for this dataset are bar graphs, box plots, confusion matrix, decision matrix and a scatterplot. By performing these visualizations, some facts about the dataset can be noticed.

Firstly, the district names are grouped with their location. The map below depicts the districts in Boston and their crime rates.




As you can see, the 5 highest crime count in Boston are plotted and highlighted them with a flash symbol just to highlight them and make them more noticeable.




**Correlation analysis:**

Before visualizing data via heat map and scatter matrix, finding correlation is the first step for the data.

**Heat map:**

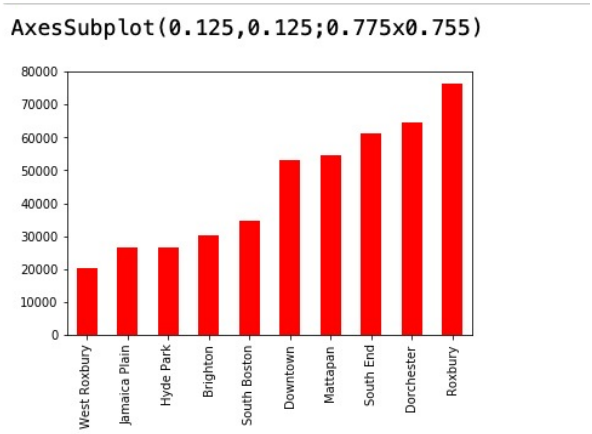
Correlation helps us to measure the relationship strength among variables ranging from -1 to 1. In the below map we have six variables –

OFFENSE_CODE, MONTH, HOUR, Population_in_District_Demographics, Average_rent_per_night , Unemployment_Rate_in_District.

**Bar plot analysis:**

**Question 1: What are the 5 highly affected districts?**

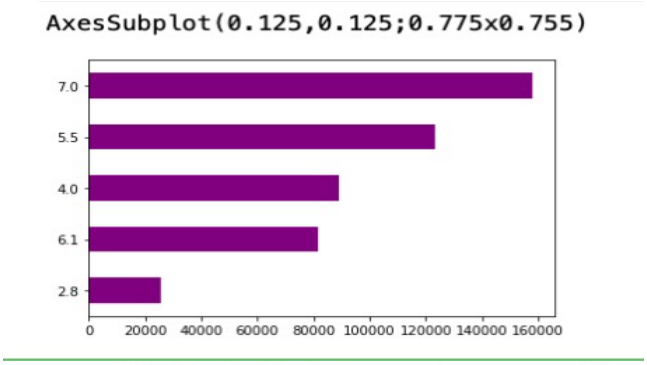
The crime count of each district has been taken and plotted in a bar graph. This bar graph depicts the crime rate in different districts in Boston. From the below image, it is obvious that Roxbury, Dorchester, South End, Mattapan and Downtown are the districts where the crime rate is the highest whereas West Roxbury, Jamaica Plain, Brighton, South Boston and Hyde park are the lowest which answers the above question as well.




**Finding the overall unemployment rate in Boston:**

**Question 2: What is the average Boston Unemployment rate?**

The unemployment rate is Boston of all the different districts are plotted in the graph below, the highest recorded is 7.0 and the lowest rate is 2.8.




**Comparing the Transport with Unemployment Rate:**

In the project of Boston Crime, there are four transport columns and unemployment rate with mode of transports utilized by public are compared. Here, Transport_1 and Transport_4 which are most used and least used respectively in the particular districts are considered. Highest Unemployment Rate is in Hyde park and Jamaica Plain districts where mode of transport is 'walk'.

| DISTRICT_NAME | Unemployment_Rate_in_District | Transport _1 | Transport _4 |
|---|---|---|---|
| Brighton | 6.1 | Walk | Car |
| Charlestown | 6.1 | Walk | Car |
| Dorchester | 5.5 | Car | Walk |
| Downtown | 5.5 | Car | Walk |
| East Boston | 5.5 | Car | Walk |
| Hyde Park | 7.0 | Walk | Car |
| Jamaica Plain | 7.0 | Walk | Car |
| Lexington | 5.5 | Car | Walk |
| Mattapan | 6.1 | Walk | Car |
| Roxbury | 5.5 | Car | Walk |
| South Boston | 5.5 | Car | Walk |
| South End | 5.5 | Car | Walk |
| West Roxbury | 5.5 | Car | Walk |

**Comparing Unemployment Rate with Average Rent:**

The below output is the result of comparing individual District's Unemployment rate with the Rents per night in that district. The highest rents per night are predicted at East Boston, Roxbury and West Roxbury. Least rent average is recorded at Mattapan, Brighton and Charlestown.

| DISTRICT_NAME | Unemployment_Rate_in_District | Average_rent_per_night |
|---|---|---|
| Brighton | 6.1 | 37 |
| Charlestown | 6.1 | 37 |
| Dorchester | 5.5 | 60 |
| Downtown | 5.5 | 60 |
| East Boston | 5.5 | 72 |
| Hyde Park | 7.0 | 66 |
| Jamaica Plain | 7.0 | 66 |
| Lexington | 5.5 | 60 |
| Mattapan | 6.1 | 37 |
| Roxbury | 5.5 | 72 |
| South Boston | 5.5 | 60 |
| South End | 5.5 | 60 |
| West Roxbury | 5.5 | 72 |

**Line Graphs – Plotting trends using MONTHS:**

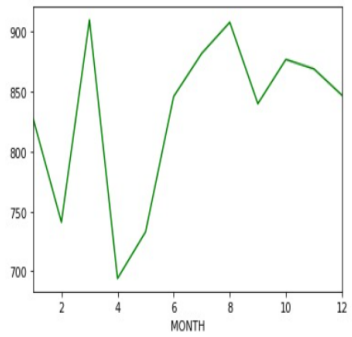
**Question 3:** **Do crimes occur in any particular months?**

**Colour code –**

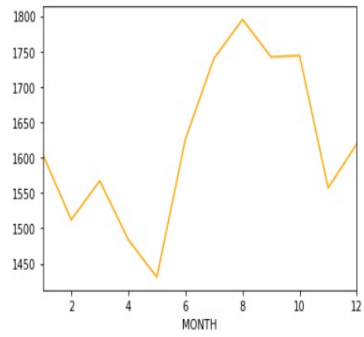
**GREEN – Charlestown; YELLOW – East Boston; RED – Jamaica Plain**

In the below graphs, 3 districts are compared. Initially, separate data frames for – Jamaica Plain, East Boston and Charlestown are created. The results show that in all three districts the crime rate is high around 8 months. When looked closer at the graphs Jamaica Plain and East Boston have almost a similar line graph but analysis of month column alone will be difficult to conclude. Hence hour, unemployment and average rents are compared below.

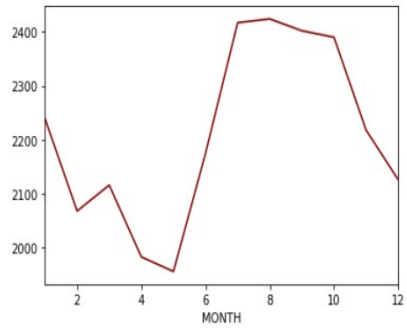
<matplotlib.axes._subplots.AxesSubplot at 0x7fd6863b7e48>




<matplotlib.axes._subplots.AxesSubplot at 0x7fd675d91b70>




<matplotlib.axes._subplots.AxesSubplot at 0x7fd6897ed0b8>



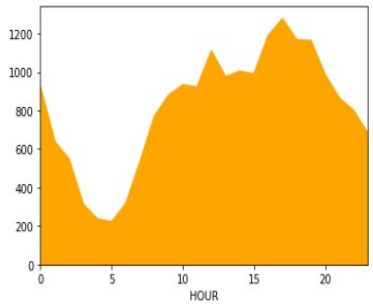

**Comparing Crime in Districts on Hour Basis:**

**Question 4:** **When does more crimes tend to occur?**

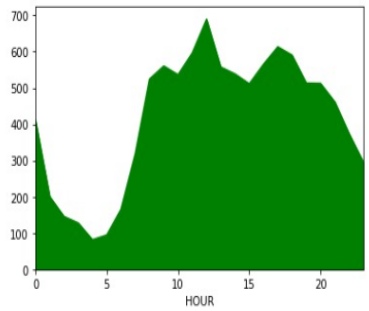
The below graphs are area graphs with the same districts i.e., Jamaica Plain, East Boston and Charlestown with the same colour coding. The below visuals are based on the rate of crimes happening at 'HOUR' of the day. From the results, the maximum amount of the crimes is happening at the night, at that time there might be less people wandering at that hour and also,

as the unemployment rate is high in Jamaica Plain people walk at that time and people whose mode of transport is walking might be more prone to crime.

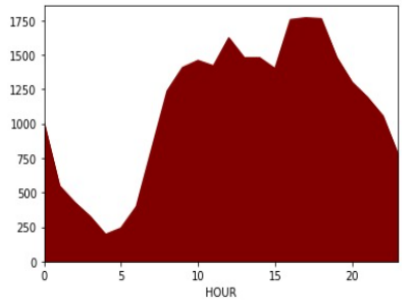
<matplotlib.axes._subplots.AxesSubplot at 0x7fd688716550>




<matplotlib.axes._subplots.AxesSubplot at 0x7fd6891cd518>



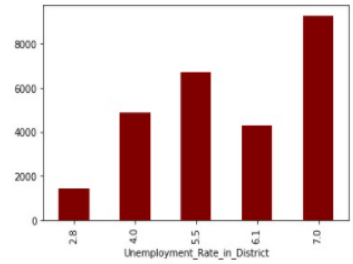

<matplotlib.axes._subplots.AxesSubplot at 0x7fd768bcc128>



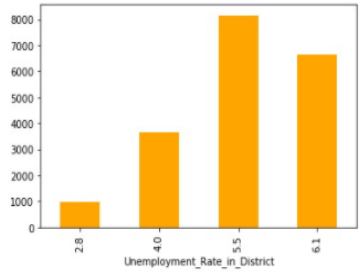

**Plotting trend using Unemployment:**

The below visualizations are histography's of Jamaica Plain, East Boston and Charlestown with the same colour coding. The results show that Jamaica Plain have the highest unemployment rate when compared to other two districts.

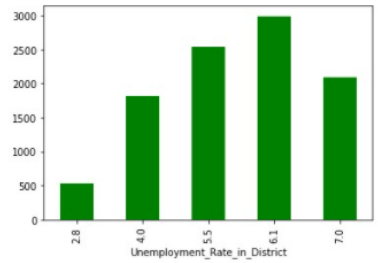
<matplotlib.axes._subplots.AxesSubplot at 0x7fd676cf77f0>




<matplotlib.axes._subplots.AxesSubplot at 0x7fd688d19470>



```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd7c4476588>
```

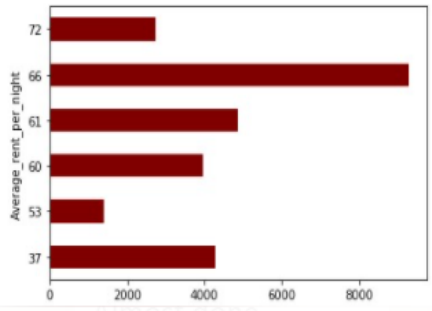



**Plotting trends using Average Rent:**

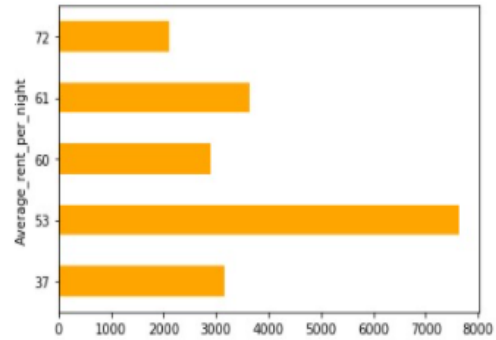
**Question 5: Does the rent matter in accordance with the unemployment rate? Which district is affordable?**

The below graphs are based on 3 different districts – Jamaica Plain (red), Charlestown (yellow) and East Boston (green). Plotting the crimes based on 'Average rent per night'. The average high rent is around 72 dollars and least is 37 dollars. In Jamaica Plain, the average rent per night is $66. In Charlestown, the average rent is $55 and $37 in East Boston which concludes that East since the unemployment rate is 6.1. So, people during unemployment will not be able to afford more for housing rents. Therefore, East Boston is affordable.

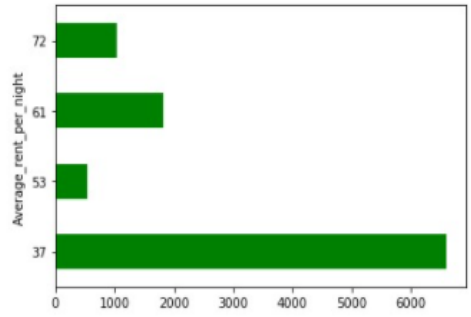
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd676bf6278>
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd68634ac88>
```

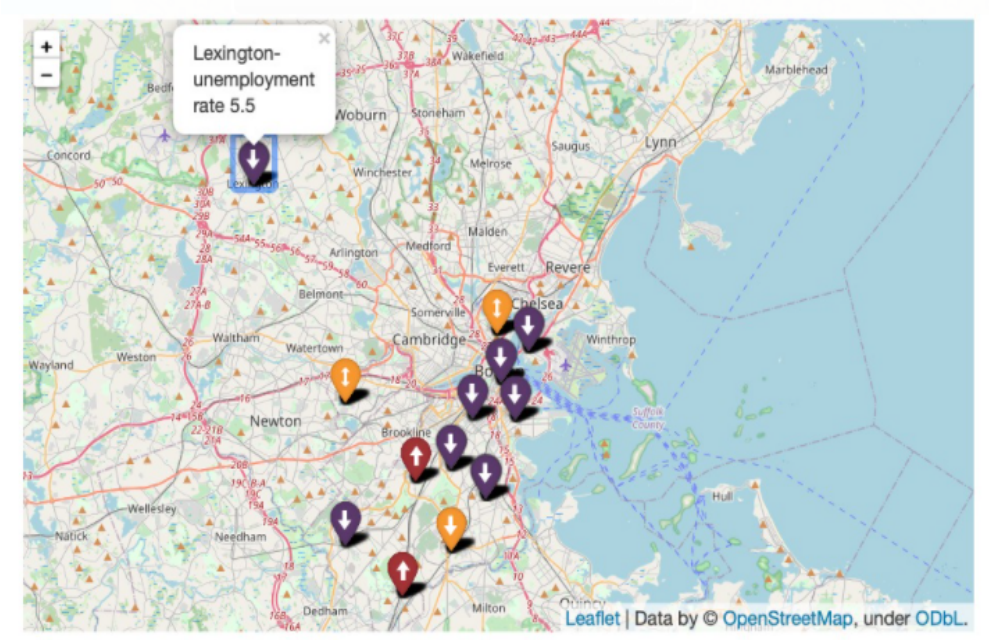


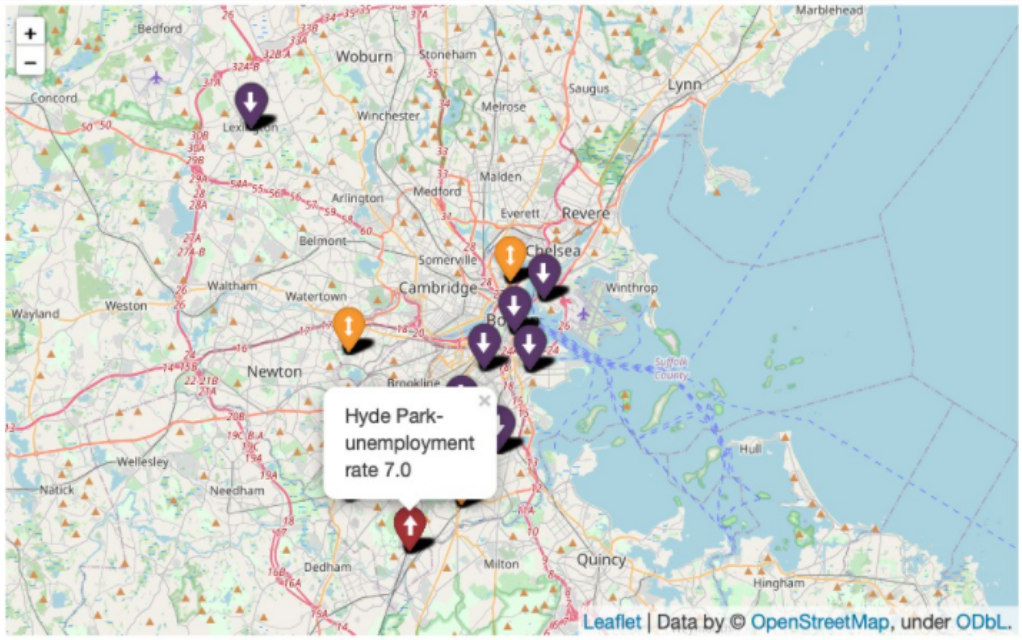
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd7c4dae1d0>
```




**Maps on Unemployment Rate:**

To make the visualization more interesting plotting districts based on the unemployment rate with 3 colour pins. The colors are Purple for 5.5 , orange 6.1 and Red 7.0 respectively. Below are few outputs. In the results **Red, Yellow, and Purple** represents the **high, moderate and less** employment rates respectively in each districts.

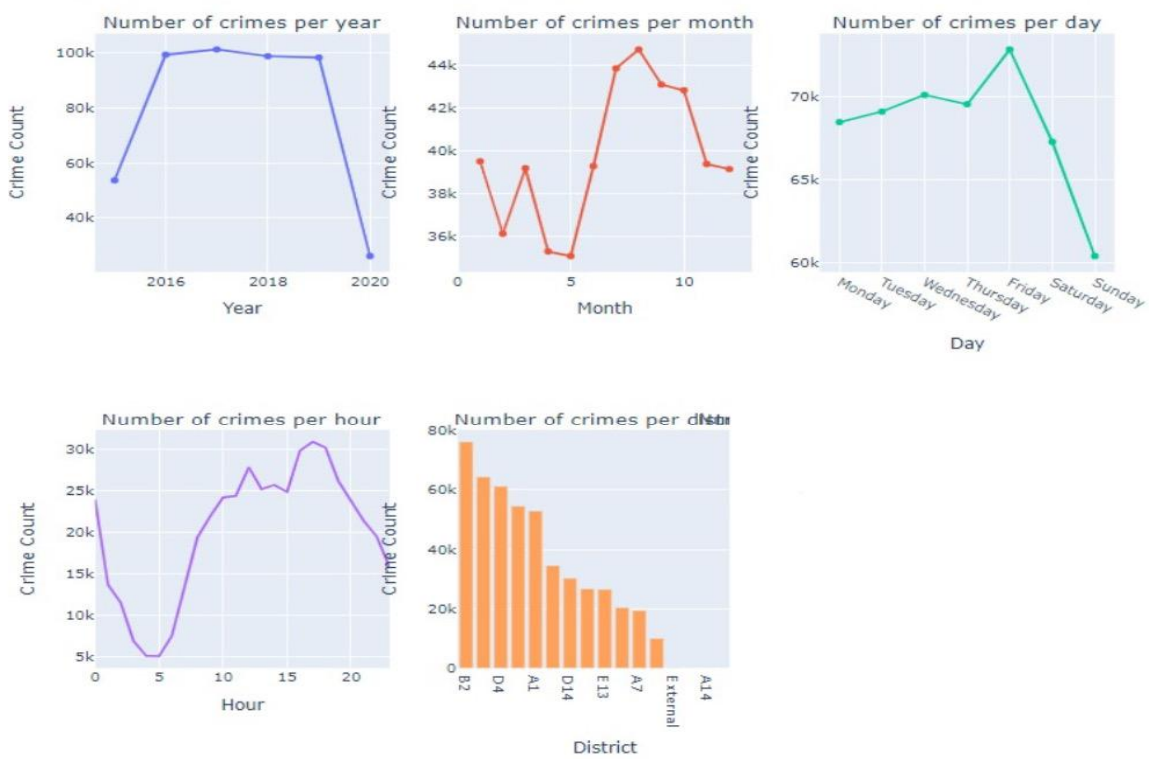



Maps for unemployment

**Total Crimes Between 2015-2020:**

**Question 6: Explain the trends and pattern using Year, Month, Day, Hour and District.**




This visualization depicts the frequency of the crime count every year, month, days, hour and district respectively.

It can be seen from the graph that in the year 2015 the crime rate was between 50k-60k, but from 2016-2019 it has increased and has remained more or less in the same range throughout these four years. Whereas, in the year 2020 the crime count has decreased to a great extent assuming because of the Pandemic.

Coming to the monthly analysis, the graph depicts how the crime count has been contrasting from month to month. July, august and September seems to be the months with highest crimes.

Now taking a look at the daily analysis per day it can be seen that crimes happen during weekdays more than weekends i.e., Friday is the day with highest crimes, whereas Sunday is the lowest. In the next graph, crimes tend to happen more during nights than during days.

**CONCLUSION:**

After performing in-depth analysis on the Boston Crime Data Set, the trends and relations are identified using Unemployment Rate, Average Rent, Population and the Occurrences of Crimes during Months, Hours, Day or Night.

Mentioning the main takeaways from the analysis below:

- People who walk during the nights especially returning back from work are suspectable to more crimes.
- The highly risky districts are Downtown, Dorchester, Roxbury, South End and Mattapan.
- Most of the people in Boston are unemployed leading to more crimes.
- To note, Charlestown is affordable to live.
- In the year 2020, the crime rates have decreased because of the Pandemic and national lockdowns.

**REFERENCES:**

*Boston crime data analysis*. (n.d.). Tableau Software. Retrieved October 11, 2020, from https://public.tableau.com/views/Bostoncrimedataanalysis/Story1?%3Aembed=y&%3AshowVizHome=no&%3Adisplay_count=y&%3Adisplay_static_image=y&%3AbootstrapWhenNotified=true&%3Alanguage=en&:embed=y&:showVizHome=n&:apiID=host0#navType=0&navSrc=Parse

*Crime Incident Reports (August 2015 - To Date) (Source: New System) - Analyze Boston*. (n.d.). Retrieved October 11, 2020, from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

*Crimes in Boston*. (n.d.). Retrieved October 11, 2020, from https://kaggle.com/AnalyzeBoston/crimes-in-boston

Geospatial Analysis using Folium in Python | Work with Location Data. (2020, June 10). *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/06/guide-geospatial-analysis-folium-python/

Holtz, Y. (2017, October 5). #312 Add markers on folium map. *The Python Graph Gallery*. https://python-graph-gallery.com/312-add-markers-on-folium-map

2019. Boston Crime Rate: Is Boston a Safe City? https://www.covesmart.com/blog/boston-crime-rate-is-boston-a-safe-city.

FBI. 2020. Boston, MA Crime. September. https://www.areavibes.com/boston-ma/crime/.

Jena, Shubhangi. 2019. Tableau — A Beginners Guide. July 27. https://towardsdatascience.com/tableau-c9d6962991ca.