Daniela Cristina Bula Gonçalves Balaniuc

Solving Crimes in the City of Boston

São Paulo
2021

Daniela Cristina Bula Gonçalves Balaniuc

Solving Crimes in the City of Boston

Course
completion work for
obtaining the
degree in
MBA
in Data Science
presented to
the Anhembi University
Morumbi.

São Paulo
2021

Summary: A Plan of action was made it with the analises and filters of the data of the crimes at the Boston city, for the police to have a plan and strategies of action and schedule turns of police to work. Were used analise, filters, data mining, pre-processing and grafics for the visualization of the data. In 3 fases of process were processed the predictive model and the descriptive model of grafics of the visualization of the data.

Key-words – Analise. Data Mining. Pre-Processing of Data, Visualization.

Resumo: Um plano de ação foi realizado de acordo com as análises e filtragem dos dados de crimes da cidade de Boston para que a policia tivesse um plano e estratégias de ação e turnos do trabalho policial. Foram utilizadas análises, filtragem, mineração de dados, pré-processamentos e gráficos para a visualização dos dados. Em 3 fases de processo foram processados nos modelos preditivos e modelos descritivos os gráficos para a visualização dos dados.

Key-words – Análise. mineração de dados. Pré-processamento de dados. Visualização.

I. Introduction

Action plan based exclusively on statistical/analytical foundations and accompanied by an implementation in Python as described in the data set. Actions used for pre-processing and included in the study; crime data was imported to MS Access to make a pre-analysis and visualization of information. Because the Boston City Police want to optimize the use of human resources, and reinforce new strategies based on data intelligence, a strategic plan of action for the police was carried out. According to Gordon S. Linoff and Michael J. A. Berry (2011, p. 398) - Data Mining Techniques - Third Edtion. " Optimization is about finding the extreme value -

maximum or minimum depending on the problem. More precisely, optimization is about finding the conditions that leads to those extreme values".
O plano de ação indicando quais os principais turnos e horários em que a corporação deve se atentar para ocorrência de crimes.

II. Materials and Methods

A cleanup and filtering was carried out because it needed to determine the types of crimes and when they happened. Including the days when attention should have distinct behaviors that favor the protection of the community are: Monday, Tuesday, Wednesday, Thursday and Friday, being on Friday the largest event of occurrence. In the days that have occurred these crimes are necessary to have more police working.
Identifying strategy that is able to identify new crimes that are high risk (1). The most appropriate technique for this type of approach is descriptive models. The technique in Python as the graph of bars showing the data of high-risk crimes.

A. Techniques used

The statistical tools and techniques were able to generate with the analyses are:
Predictive Models: It is a regression technique using data mining that allows the development of models that act in the optimization of the way we interpret an existing data set to identify the data to a future behavior, the Linear Regression technique is an alternative, since it allows predicting the result of a given unknown pending variable through values allocated in independent variables. Simple Linear Regression to predict results in continuous variables, as a plan of action indicating with which the main shifts and times in which the corporation should pay attention to the occurrence of crimes.
Descriptive Models: Data mining and machine learning algorithms perform two functions. One of them is to predict how a new data point will fit into historical patterns. For example, an algorithm can use data history that medical professionals have about certain treatments and diseases to determine which treatment would be most effective for a new patient. The other function is to

discover patterns in unlabeled data. For example, a company wants to analyze all its data about customer interactions to determine the different types of customers the company has and how to approach them effectively. In this example, the data is "unlabeled". The company does not know how to define its customer groups, but wants to examine all of its customers to see if certain patterns can be detected and then label the groups based on those patterns. This data mining and machine learning function often relies on unsupervised machine learning techniques.

B. Strategies for Validation

Identifying strategy that is able to identify new crimes that are high risk (1).
The most appropriate technique for this type of approach is descriptive models. The technique in Python as the graph Bars the data of high-risk crimes.

C. Premises

In the city of Boston and regions, the validation criteria used to demonstrate the success of its solution were used. This technique of Bar Graph and Clusters were used to show the data of high-risk crimes. The pre-processing actions that have been included are data filtering, data analysis, visualization and localization with k-average comparison.

D. Desired Results

Identify the times that deserve more attention from the police team, the days when the crimes happened the most. The days when attention should have different behaviors that favor the protection of the community. The days that occurred these crimes are necessary to have more police working. The new crimes that high risk (1). Identification of possible criminal organizations affecting district B2.

II. Obit Results

The days when attention should have distinct behaviors that favor the protection of the community are: Monday, Tuesday, Wednesday, Thursday and Friday, being on Friday the largest event of occurrence.
The days that these crimes took place require more police working.
The data visualization approach for your results to be relevant to a Boston police analysis are: Aggravated Assault, Ballistics, Robbery, Simple Assault, Violations.
Data analysis and identification of possible criminal organizations affecting district B2.

This table shows the groups with the highest number of crimes: Total 1244 and some of the crimes with the lowest number of occurrences.

Grupos com Mais Número de Incidentes	
Tipo de crime	**Total**
Motor Vehicle Accident Response	1244
Medical Assistance	743
Verbal Disputes	654
Other	631
Investigate Person	528
Simple Assault	525
Drug Violation	501
Larceny	488
Vandalism	433
Investigate Property	333
Aggravated Assault	317
Property Lost	300

Grupos com Mais Número de Incidentes	
Tipo de crime	**Total**
Larceny from Motor Vehicle	277
Violations	266
Towed	251
Warrant Arrests	202
Missing Person Located	199
Auto Theft	173
Fraud	164
Residential Burglary	161
Harassment	151
Missing Person Reported	145
Robbery	126
Liquor Violation	106

II. Analysis of Results

1. I analyzed the crimes that happened in the city of Boston all crimes by zones, neighborhoods and streets showing the total number in the year 2018.
2. Using grouping (clustering), which can be considered one of the most important techniques of unsupervised machine learning. Its implementation consists in identifying structures or patterns in a set of unlabeled data. Some authors define this approach as a process of organizing objects into groups that share, in some way, similar characteristics. Thus, groups (or clusters) are nothing more than collections of objects that are "similar" to each

other, but that "differ" from objects belonging to other groups.

3. 1. A cleaning and filtering was carried out because it needed to determine the types of crimes and when they happened.

4. 2. The days when attention should have distinct behaviors that favor the protection of the community are: Monday, Tuesday, Wednesday, Thursday and Friday, being on Friday the largest event of occurrence.

5. 3. The days that these crimes occurred need to have more police working.

6. Identifying strategy that is able to identify new crimes that are high risk (1).

7.

8. 1. The most appropriate technique for this type of approach is descriptive models.

9. 2. 2. the Python technique as shown in the Bars chart shows the data of high-risk crimes.

10. 3. The validation criteria that were used to demonstrate the success of your solution.

12. 4. This bar Graph technique was Chosen because it shows data from high-risk crimes.

13.

14.

15. The pre-processing actions that have been included are data filtering

16. The data visualization approach for your results to be relevant to a Boston police analysis are: Aggravated Assault, Ballistics, Robbery, Simple Assalt, Violations.

Data analysis and identification of possible criminal organizations affecting district B2.

1. Grouping technique making it possible to form groups representing crimes

2. The cluster technique shall be applied, k-media based on the locations and quantity of crimes. According to Gordon S. Linoff and Michael J. A. Berry 2011, p. 468, 490 - Data Mining Techniques - Third Edition. "The goal of Algorithm clustering is to find the k point that makes a good central cluster. The central cluster defines the

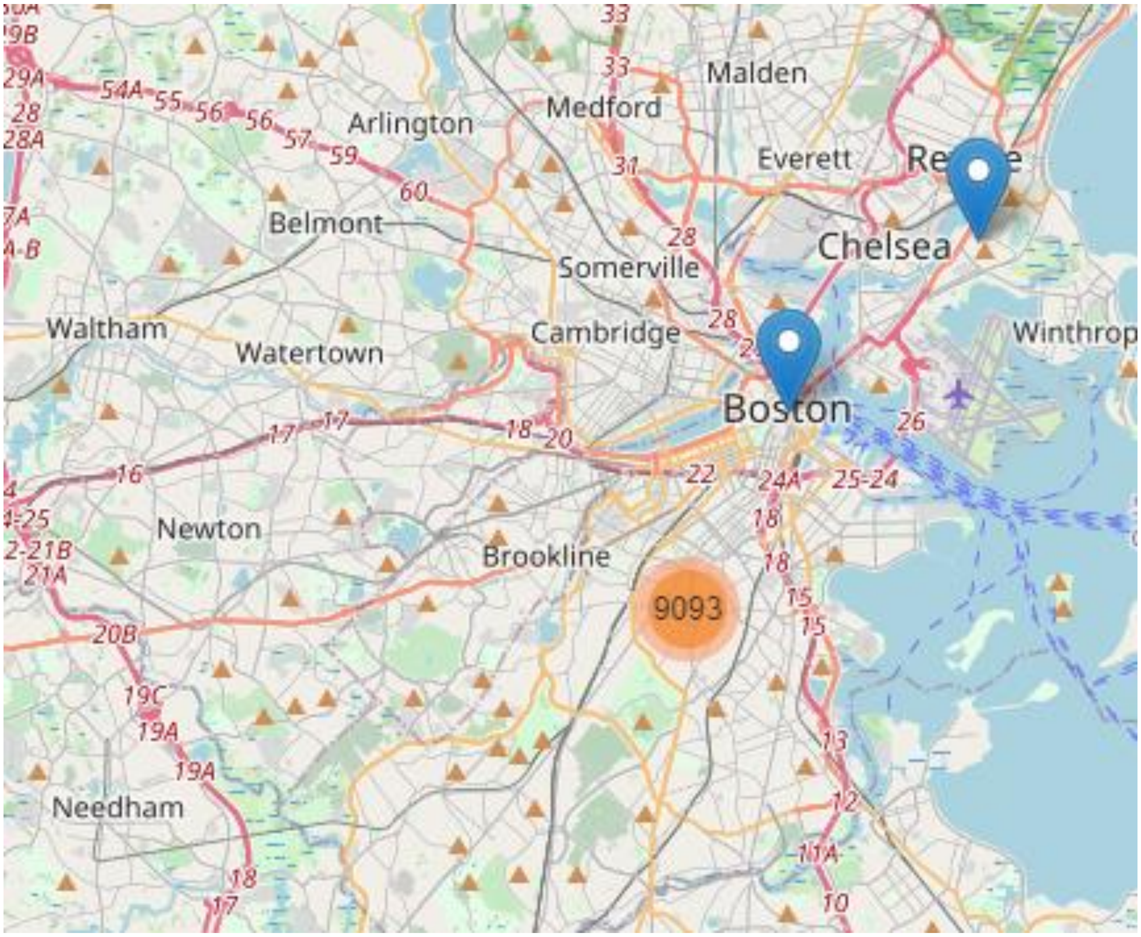
clusters: Each point is subjugated to the cluster identified by the nearest central cluster. The goal is to minimize the distance between the cluster members and a central point. The cluster has a "Y" shaped demarcation boundary with a cluster on the left, one on the right, and one on top. Drawing this marking between clusters is very useful to show the process geometrically".

3. Imported the database
4. Processing of the database with filtering of types of crime
5. Visually representing results with the matplotlib library
6. Presenting an analysis of the results of the most frequent crimes in distrtio B2
7. The types of information and insights that can be delivered as a result to the police team are the most frequent crimes as shown in the cluster and k-average graphs

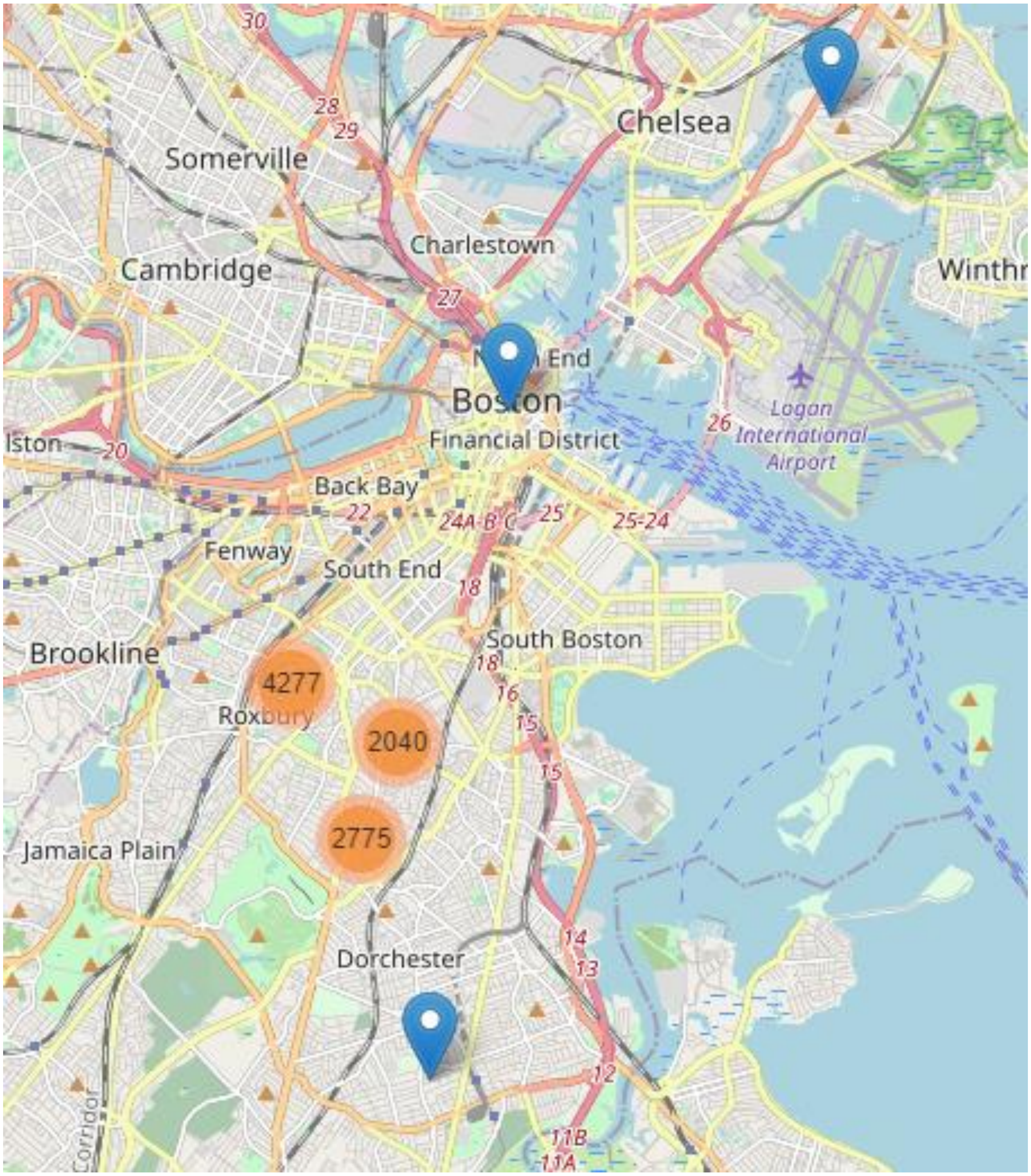
In this analysis as the data shown in the Clusters Graphs - k-means identifies possible criminal organizations affecting the district B2.

Localização e quantidade dos agrupamentos com maior número de incidentes: Total 1244 utilizando a Biblioteca Folium para gerar mapas geolicalização.

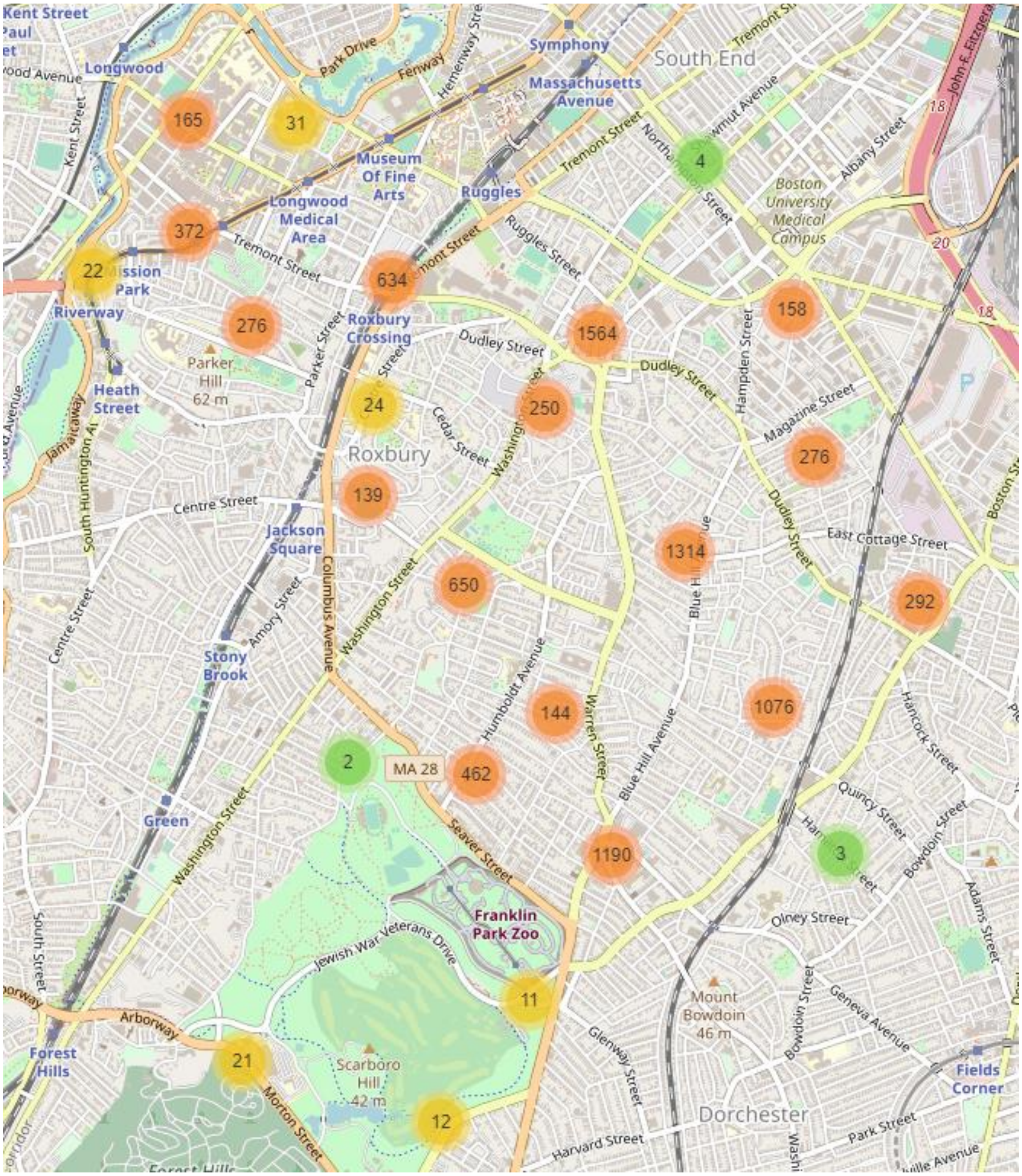
This map shows the total number of geolocation crimes in the Boston area with a total of 9,093 crimes in the vicinity of the South End in Boston and more in Roxbury.



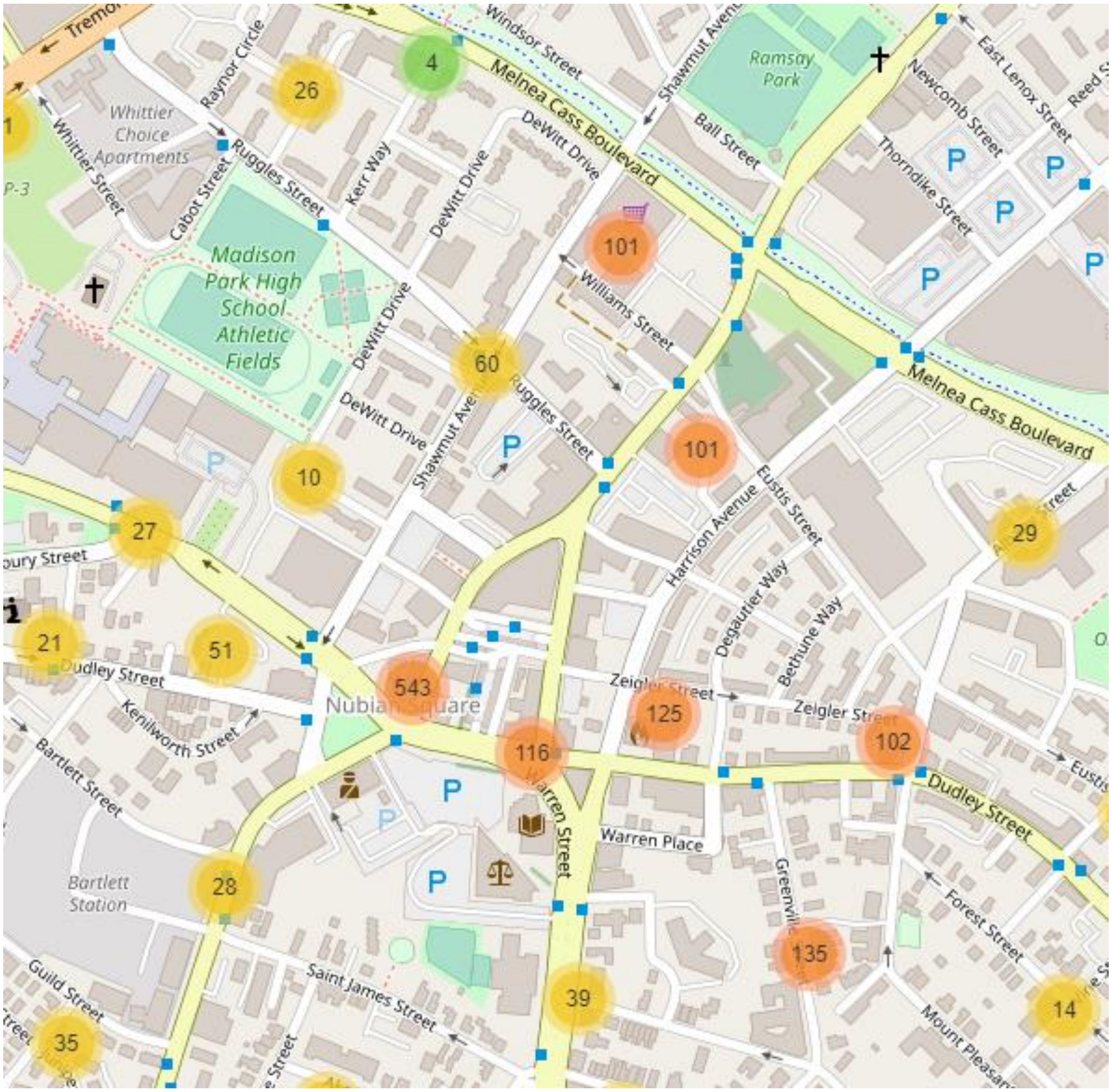
This Map shows the amount of geolocation crimes in 3 zones and neighborhoods in the Boston area in Roxbury and South Boston with 4277 crimes being this the k-compared with the other regions that have fewer crimes occurred and South with 2040 crimes and in the region near Dorchester 2275 crimes.



This map shows the amount of crimes by geolocation in zones and neighborhoods in the Boston area. In Dudley Street with 1564 crimes being this the k-average comparing with the other regions that have less crimes occurred. Also with high risk of crimes Blue Hill Avenue shows 1314 crimes, and Warren Street with Blue Hill Avenue 1190 crimes.



In this map shows the amount of crimes by geolocation in zones of neighborhoods and streets, being able to see 3 streets with the highest numbers of crimes and with minimum distance from other groups, which are 543 crimes in the vicinity of Nubian Square and that is the k-average compared to the other regions that have fewer crimes occurred, as well as 116 crimes on Warren Street and 135 crimes on Greenville Street, 125 on Zeigler Street, 101 on Eustis Street.



III. Conclusions

The objective was achieved with the action plan based exclusively on statistical/analytical foundations and accompanied by an implementation in Python as described in the data set. Actions used for pre-processing and included in the study; crime data was imported to MS Access to make a pre-analysis and visualization of information.

Because the Boston City Police wanted to optimize the use of human resources, and reinforce new strategies based on data intelligence, a strategic plan of action for the police was carried out.

The action plan indicating the main shifts and times in which the corporation should pay attention to the occurrence of crimes.

Just as in Data Mining what matters is the data, it addresses important issues about the analysis of different types of data, how this data is collected and how it can be analyzed. The quantitative results were counted, measured and expressed using the numbers, the qualitative results were worked on dates categorized based on the characteristics analyzed and used of the data, as well as the predictive model and descriptive model. It can also be used in the future techniques in Hierarchical Grouping, Logistic Regression, Classification, Decision Tree and DBSCAN.

Bibliographic references:

1. Gordon S. Linoff and Michael J. A. Berry (2011, p. 398, 468, 490) - Data Mining Techniques - Third Edtion.