COURSERA CAPSTONE

IBM Applied Data Science Capstone

CLUSTER BOSTON CITY WITH CRIME DATA

By Manasa Devi Chakka

1 Introduction

1.1 Background

Boston's colleges and universities exert a significant impact on the regional economy. Boston attracts more than 350,000 international students from all over the world. The area's schools are major employers and attract industries to the city and surrounding regions. The city is home to a number of technology companies and is a hub of biotechnology, with Milken Institute rating Boston as the top most cluster for biotechnology in the country. The Boston public schools enrol 57,000 students attending 145 schools. There are private, parochial and charter schools as well and approximately 3,300 minority students attend participating suburban schools through the Metropolitan Education Opportunity Council. This accounts for a requirement of safer locality for all these students, employers and employees but unfortunately likewise with all major cities crime rate is high in Boston.

1.2 Business Problem

With the help of crime data digitalized by Boston City Police Department I would like to develop a real time solution which divides Boston city into different clusters based on crime rate in that area. This helps people immigrating to Boston City to find a safe home.

1.3 Target Audience

Any person who is relocating to Boston will find this useful. Especially employers or employees relocating with their families consisting of children. They would want to provide their children safe environment.

2 Data Acquisition and Data Cleaning

2.1 Data Sources

The crime data of Boston city can be collected from Kaggle. This is a dataset containing 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. This dataset has more than 2 lakh rows and 17 columns. Dataset link is https://www.kaggle.com/ankkur13/boston-crime-data

We also require venue data around the area of occurrence of crime. This helps in effective clustering of city and also provides end user with information about what kind of area it is , what venues are nearer to the frequent crime scenes so that user can avoid such kind of venues in any area. This venue data can be obtained from Foursquare API. We can make 2000 regular calls and 2 premium calls per day to Foursquare API using personal developer account which is free.

2.2 Data Cleaning

The crime data obtained from Kaggle may possess nan values in some columns. For effective data analysis and modelling techniques we need to minimize these NaN values. We can either remove the rows consisting NaN values which is useful if no. of NaN values are less or replace NaN value with zero or if a column contains a lot of NaN values we can remove that column entirely. In our dataset SHOOTING column has a lot of NaN values. So we can remove this column entirely from the data frame.

The data obtained by using Foursquare API can be imported as a JSON file. When we observe it we may find that this file contains lot of information that is not required for further processing. The details of venue that we require mainly are venue name and its latitude, longitude, category. So we extract only these details from JSON file and store these in a data frame with its respective street and street's location details.

2.3 Feature Selection

After data cleaning, the crime data frame still contains many rows which include data and time of occurrence of crime that is not relevant for analysis. It also contains many columns that include serial Id of crime, area to report etc. which are irrelevant. The columns that contribute for further process are street, offense code group, latitude and longitude. Here street column is required because it is the splitting criteria of city. Latitude and longitude are required to be specified in the url to make a call for Foursquare API . Offense code group feature is required to train the model. In venue data frame venue category is used in training the model

3 Exploratory Data Analysis

3.1 Dimensionality reduction

Dimensionality reduction is an important technique we have to perform on our data. The reason for this is our dataset is large. There are 4684 streets in our dataset. For many of the streets , the crime recorded is in single digit and they are also minor crimes. So we can remove all these rows of streets. Using value_counts method in Pandas library we can count no. of crime entries each street has and sorting it in descending order we obtain the top crime rated areas. For making feasible no. of API calls we are going to work with top 30 crime rated areas. This technique reduces the rows in the dataset from around 2,60,000 rows to around 1 lakh rows.

3.2 One Hot Encoding

The variables we want to use in fitting the model are categorical variables. But the k-means algorithm we want to use accepts only numeric data. In order to overcome this issue we use one-hot encoding. One hot encoding is applied on a single or list of columns. When we apply this on a column for each distinct value in it a separate column is created and its relation with row is specified with 1 or 0. If it's value is 1 then initially the row contains this value and vice versa. In python we can apply one hot encoding directly to a column using get dummies

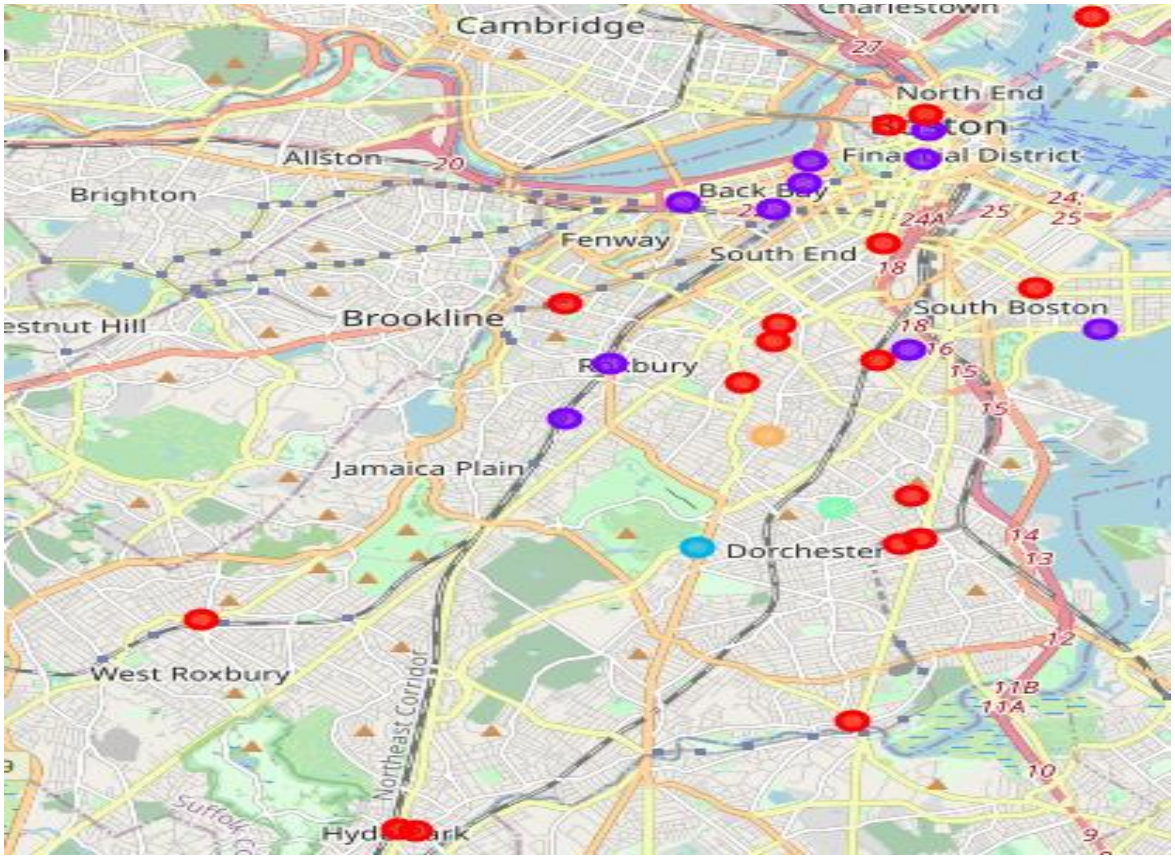
method in Pandas package. We apply one hot encoding on offense code group in crime data and on venue category in venue data. We then merge these columns obtained for fitting the model.

4 Modelling

The machine learning algorithm we employ to provide solution to our business problem is k-means clustering technique. It is an unsupervised machine learning technique where all the columns are treated as dimensions and each row is plotted as a point. We select the number of clusters the data to be divided into and when we start fitting the model with our data then it randomly allocates centroids and all the data is divided into clusters by calculating distance from all the centroids and assigning it to the cluster whose distance to centroid is small. In our model we divide our data into 5 clusters based on offense code group of crime data and venue category of venue data.

5 Results

After modelling the cluster labelled 0 contains data of seventeen streets, cluster labelled 1 has data of ten streets, cluster labelled 2 has data of one street, cluster labelled 3 has data of one street and cluster labelled 4 has data of one street. Five clusters consisting a total of more than one lakh rows.



6 Discussion

When we examine the clusters it can be observed that the cluster labelled 0 has data of seventeen streets, cluster labelled 1 has data of ten streets, cluster labelled 2 has data of one street, cluster labelled 3 has data of one street and cluster labelled 4 has data of one street. We can also observe that the most common type of crime in all the areas is motor vehicle accident. Other observations are, most common crimes in cluster 0 are larceny and drug violation. Common crimes in cluster 1 are vandalism and assault.

7 Conclusion

In this study I have used crime data digitalized by Boston police department that is available in Kaggle for processing and venue data obtained from Foursquare API in developing a solution to help relocating people find a safe home. Top 30 crime rated areas are extracted from crime data. Using features offense code group from crime data and venue category from venue data K-means clustering algorithm is fitted for these 30 areas. They are divided into 5 different clusters based on the fitted data. The crime data depicts the type of crimes that are common in those areas and venue data depicts venues where crimes occur frequently so end user can avoid such venues anywhere in the city.