Finding the best option of an apartment in Bucharest

Dolganiuc Ana

06.05.2021

1. **Introduction**

1.1 Background

Bucharest is a big city with plenty of neighborhoods and options for renting or buying real estate. For a businessman,or a simple person who wants to buy an apartment the variety of options can be overwhelming. I'm part of the DataScientists group of a Real Estate agency and I was requested to find the best apartment in Bucharest, Romania and to fit the criteria set by the client.
His main requests are:

- to have 1 or 2 rooms
- to be in the range of the average price
- to be near the central zone
- to be near the subway
- to be in the area with most venues
- to have the price between 55.000 and 90.000

1.2 Problem

What is the best option of an apartment for a couple in Bucharest in 2021?

2. Data acquisition and cleaning

2.1 Data sources

I used the dataset about the Real Estate found on kaggle , and the Foursquare API

- *The Foursquare location data*- I will use the Foursquare location data to fill all the client's requested criteria about how the zone should be like
- *Real Estate csv*- I will use a Real Estate csv found on kaggle with the prices and the locations of the apartments(from april 2021), which will give me the opportunity to search for an apartment that will suit my client the best
  The columns of the csv are: location_area, price, rooms_count and subway_distance

2.2 Data cleaning

After downloading the csv into the notebook, I started filtering the data. First of all I've dropped all the unnecessary columns, changed the row values with more

significant ones, displayed only the apartments (because the data set contained other types of real estate), and replaced some location names that didn't match with the geocoder library. In addition, I dropped some columns that contained missing values which were insignificant in the whole process. Ultimately, I checked the data type of every column and the number of rows and columns necessary to continue the process.

2.3 Feature selection

After cleaning the data frame, I began the process of selecting the apartments by the criteria set by the client.

Firstly, I looked for the areas closer to the subway, with a distance with less than 900 meters. Then I selected the apartments that had 1 or 2 rooms

Secondly, I found the latitude and longitude of every neighborhood using the geopy and geocode libraries. After that I used the folium library to locate the central areas, and found out that there were 13 areas of this kind.

After Data cleaning and Feature selection my Data Frame looked like this:

	location	location_area	subway_dist	price	real_estate_type	rooms_count	latitude	longitude
5	Bucuresti, zona Mosilor	Mosilor	776	79999	apartment	2	47.230328	28.856537
77	Bucuresti, zona Cismigiu	Cismigiu	485	80000	apartment	1	44.437424	26.087402
478	Bucuresti, zona Mosilor	Mosilor	895	85996	apartment	2	47.230328	28.856537
825	Bucuresti, Sector 3, zona Calea Calarasilor	Calea Calarasilor	873	81000	apartment	1	45.254743	27.957209
858	Bucuresti, zona Universitate	Universitate	473	77000	apartment	2	44.435563	26.102468

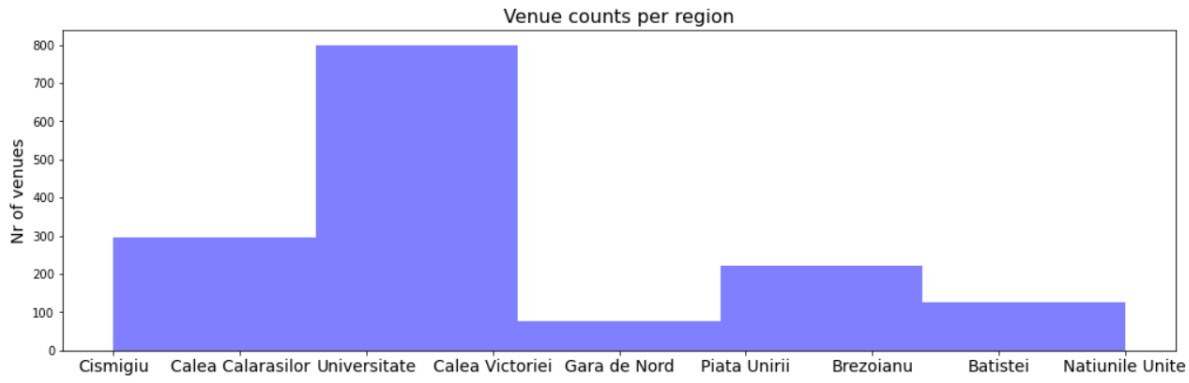
3.Exploratory Data Analysis

3.1 Displaying the target variables

At this stage I used the Foursquare API to identify the venues within 500 m from the neighborhood and set a limit of not more than 100 venues per borough.

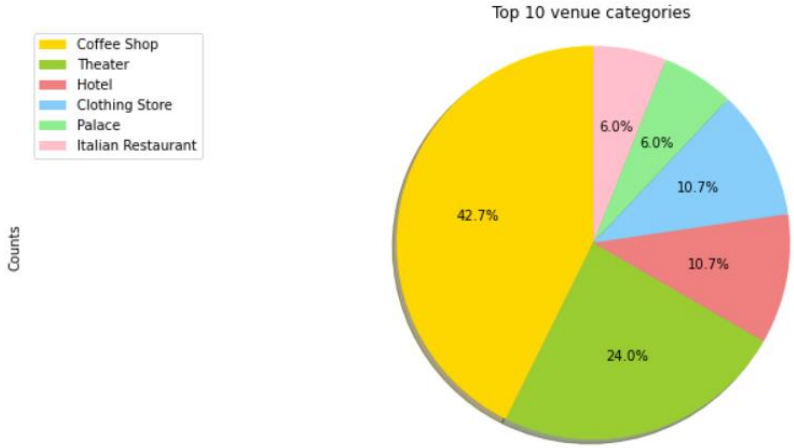
3.2 Relationship between venues and neighborhoods

Here I identified which neighborhood had the biggest number of venues and found out that the most popular neighborhood was "Universitate" with around 800 venues



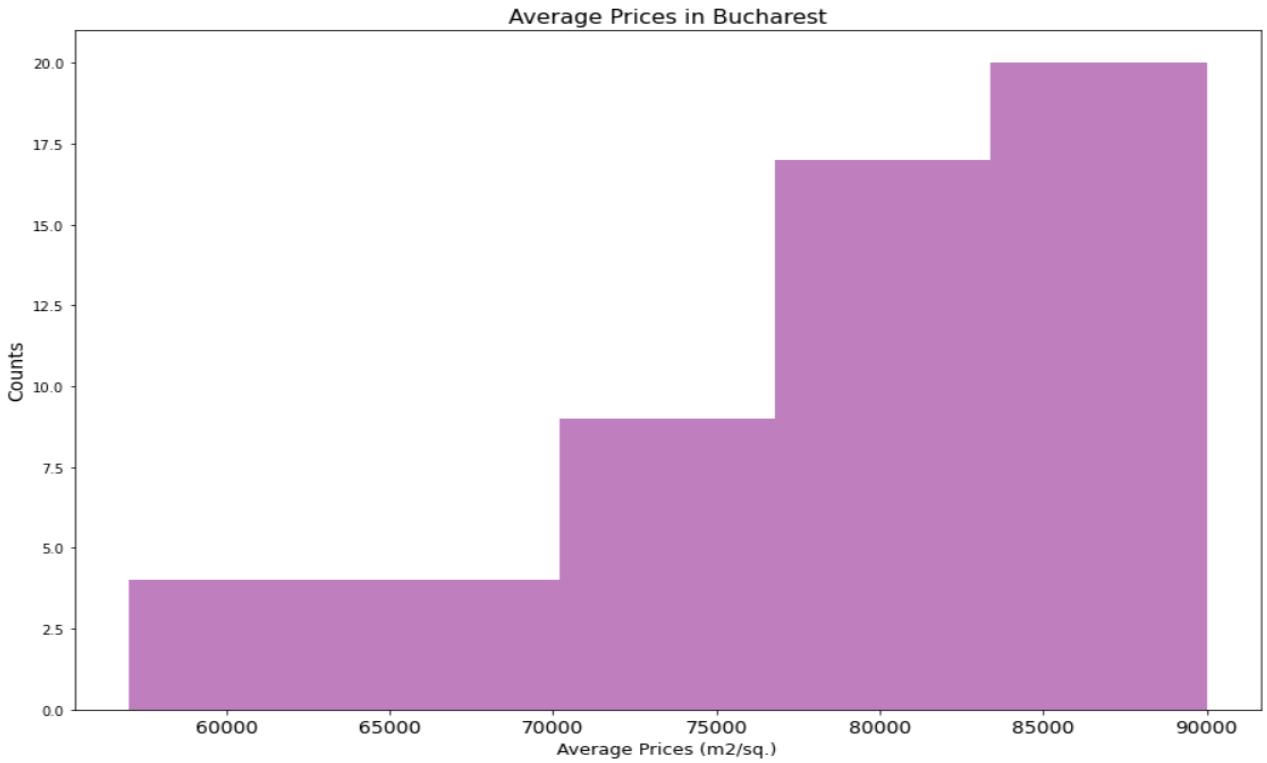
3.3 Relationship between venue category and neighborhood

I decided to see top 10 most popular venues in our target area- Universitate, and as displayed on the pie chart, the top most important venue categories are : Coffee Shops(42,7%), Theaters, Hotels, Clothing Stores, Palaces and Italian Restaurants

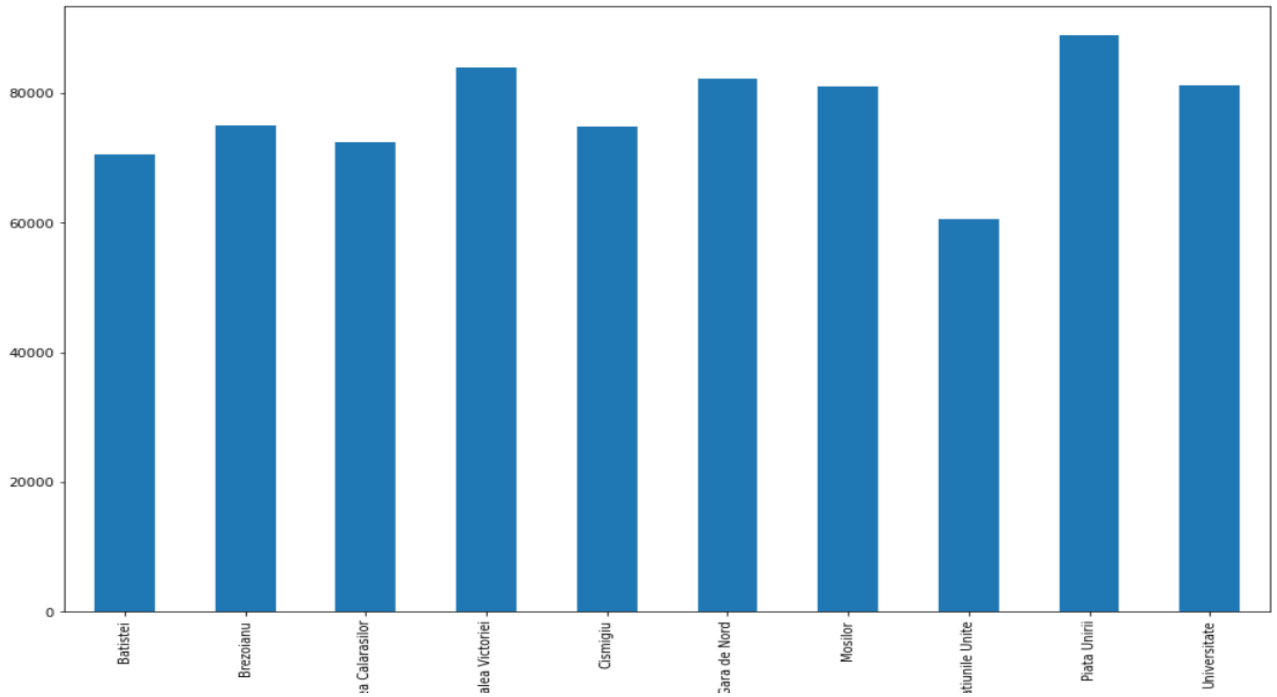


4.4 Relationship between price and neighborhood

First of all, I set the price to be in the specific range between 55.000 and 90.000 ron per m^2 , after that I represented the average prices in all the neighborhoods from Bucharest with a histogram graph

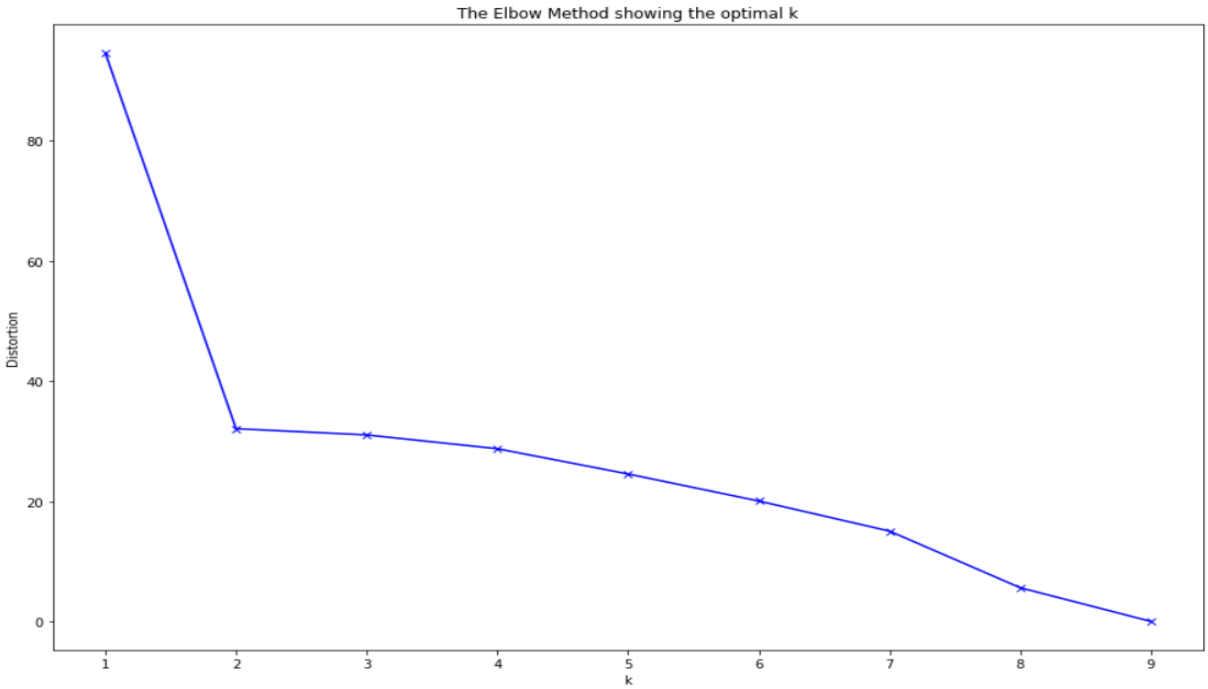


After that I visualized the price range for our specific area with a bar chart and found out that the average prices in our selected area is around 80.000 ron per m^2



4. Clustering

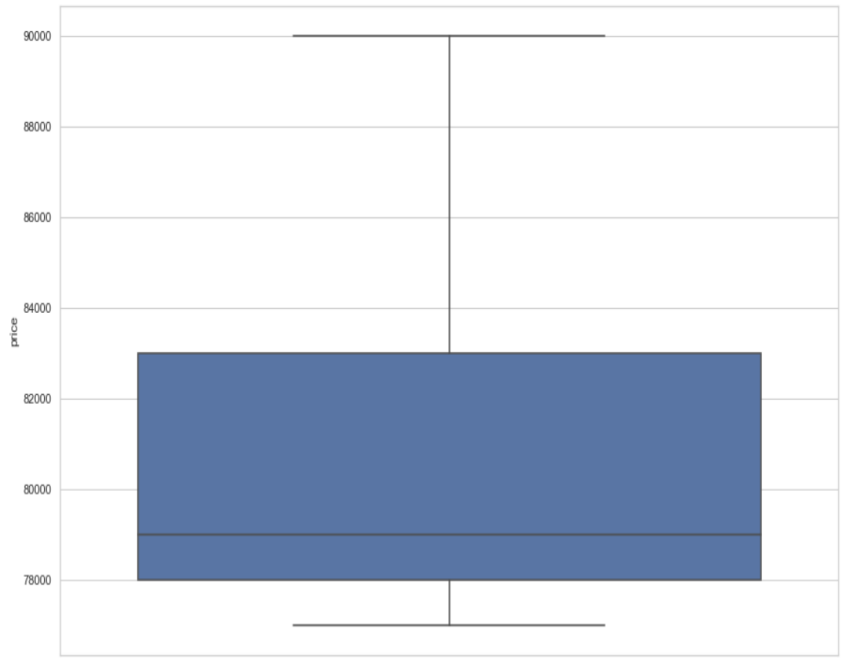
In this situation the best predictive model would be a clustering one, more precisely a K means clustering model. It is essential to use it in order to see the top 10 most popular venues in our area. For k means I used k=2 which was the best option gave by the Elbow method function.



To cluster the selected zone, I found the top 10 most popular venues in the Universitate location area and merged that with my initial data frame to get the locations for the folium map

5. Finding the best price

After all the analysis I visualized the filtered data frame with the prices and found out that the lowest price after all the set criteria would be 77.000 and the biggest would be 90.000 for a 2 rooms apartment in the Universitate area



6. Conclusions

As seen in the analysis there are a variety of apartments on sale in Bucharest(9974 more exactly). With the great multitude of neighbourhoods, it takes time to analyze which variant would be the best. Moreover the analysis depends on the customer's needs, whatever he is looking for a family-friendly zone or for a fun zone for the youth.

I used the K-means algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 2. For more detailed and accurate guidance, the data set can be expanded and the features can be changed based on the custormer's need.

I also performed data analysis through this information and ploted/clustered the most important insights of this project

I ended the study by determining the best option of a neighborhood according to price, number of rooms, distance from subways and top venues in the zone
As a result, this kind of data analysis would fit any type of customer: from business people who are looking for offices to families that decided to buy real estate

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.