IBM – Coursera

Data Science Professional Certificate

Finale Capstone Project Report

“Best lands to buy in Casablanca “

Marouane BELMALLEM

September 2020

A. Introduction

This report is a part of Coursera's Data Science Professional Certificate provided by IBM. The certificate includes 9 courses such as Data Science Methodology, Databases and SQL, Data Analysis, Data Visualization, and Machine Learning. The requirement for the final report is to use Foursquare API to explore or compare neighborhoods or cities of our choice. The learner is given full creative freedom to decide what problem they will focus on and what methodologies will be used to solve that problem.

A.1. Business Problem

For this project, I was inspired to delve into land values to look for a place in Casablanca, the biggest city in Morocco and my hometown. Therefore, I will be using methods such as K-means and linear regression to create predictive analytics of land values in Casablanca. Real estate values are determined by many factors and different buyers have different priorities. Some factors that many people consider when buying lands are location, size, usable space and neighborhood comps.

In this context, the investor needs to know accurate data in which the decision he makes is based on the assertiveness of the treatment of this data. It is for these reasons it will try to give a compass on investment in real estate based on the distribution of the lands categories and their prices.

B. Data acquisition and cleaning

B.1. Data Sources

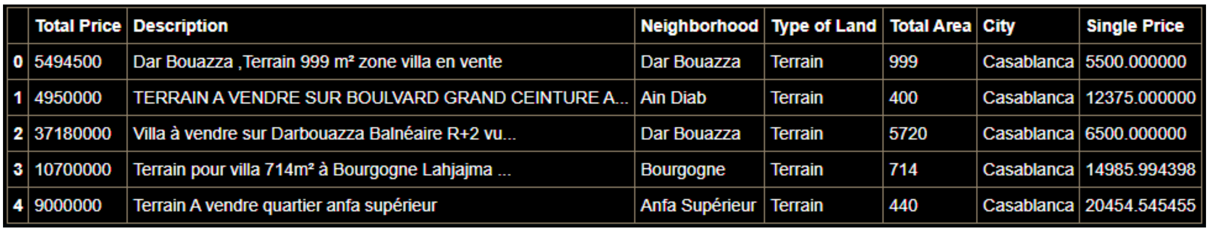
It will use 4 main sources to obtain information that will allow its manipulation and subsequent analysis, which are:

- **Google Maps API**: It Allow to make requests about the coordinates, places, and specific routes (Google, s.f.).
- **Sarouty:** Leading real estate classifieds platform in Morocco.
- **Colaboratory:** Plateform of developing the notebook.

B.2. Data Cleaning

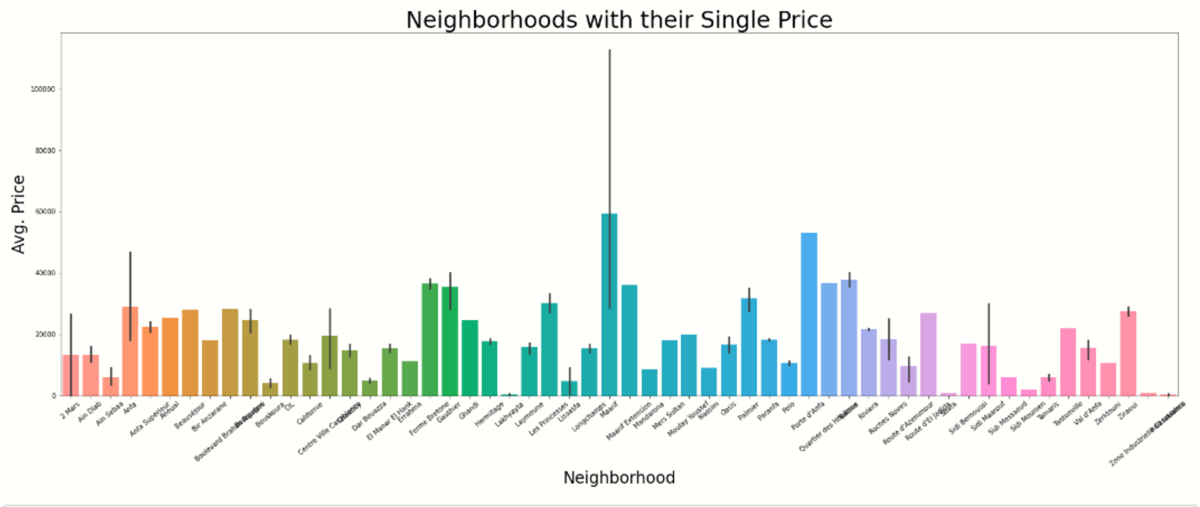
Many families consider the ***quality of local schools***, ***employment opportunities***, ***proximity to shopping*** etc. before buying a land or home. However, this remains valid in general just in case of families, so for the investors the story changes.

In sum, the most needed data for every piece of land that'll be extracted from the web are its ***size***, ***price*** and ***neighborhood***.



	Total Price	Description	Neighborhood	Type of Land	Total Area	City	Single Price
0	5494500	Dar Bouazza ,Terrain 999 m² zone villa en vente	Dar Bouazza	Terrain	999	Casablanca	5500.000000
1	4950000	TERRAIN A VENDRE SUR BOULVARD GRAND CEINTURE A...	Ain Diab	Terrain	400	Casablanca	12375.000000
2	37180000	Villa à vendre sur Darbouazza Balnéaire R+2 vu...	Dar Bouazza	Terrain	5720	Casablanca	6500.000000
3	10700000	Terrain pour villa 714m² à Bourgogne Lahjajma ...	Bourgogne	Terrain	714	Casablanca	14985.994398
4	9000000	Terrain A vendre quartier anfa supérieur	Anfa Supérieur	Terrain	440	Casablanca	20454.545455

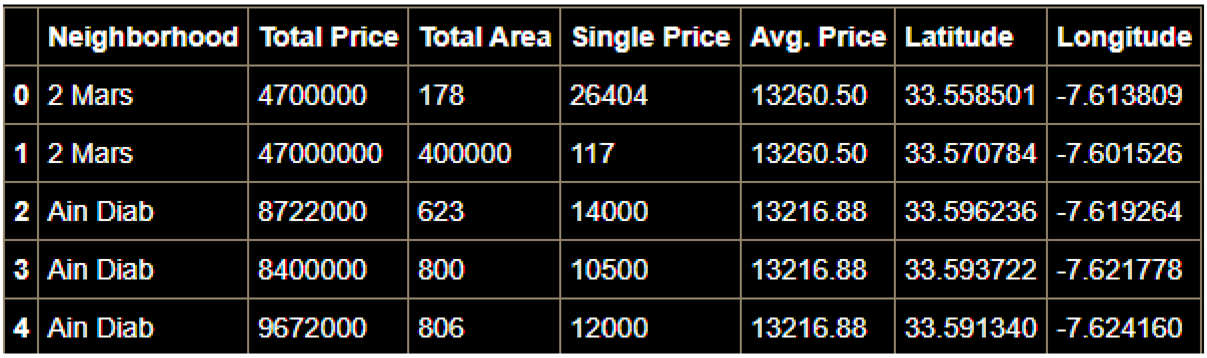
A graph was made comparing neighborhoods in Casablanca by their lands' prices and types:



In this way it can be ensured that the investment will not have a negative benefit when choosing the type of land that is required to invest, because the real estate sector around it, is profitable.

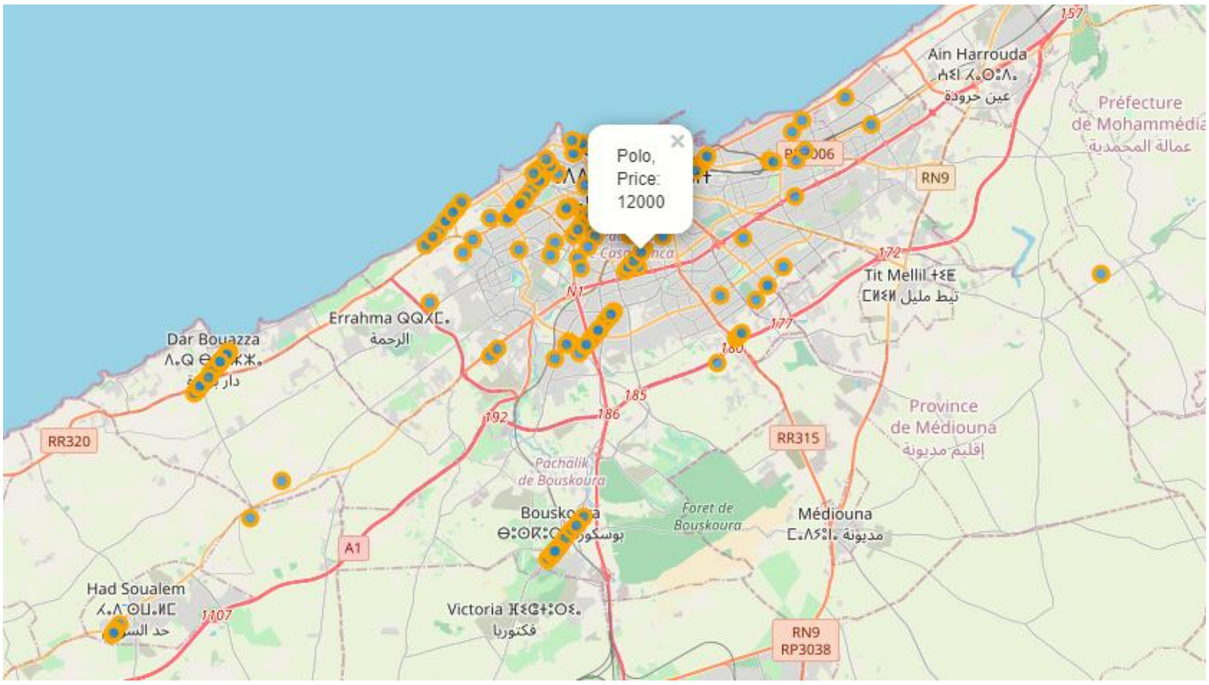
B.3. Exploratory Data Analysis

With the use of the Geopy python library, the coordinates of the different districts of Casablanca boroughs are found, that is, the districts are geocoded through their formatted address to give as a result their respective latitude and longitude coordinates and then append it with the table created earlier.

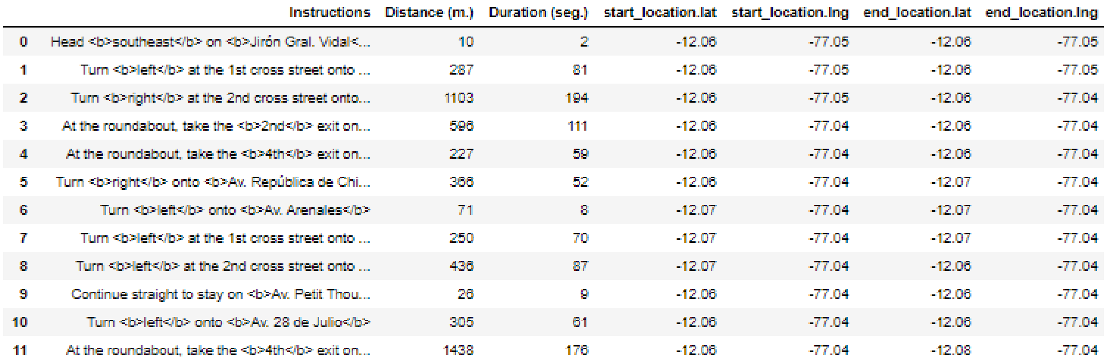


	Neighborhood	Total Price	Total Area	Single Price	Avg. Price	Latitude	Longitude
0	2 Mars	4700000	178	26404	13260.50	33.558501	-7.613809
1	2 Mars	47000000	400000	117	13260.50	33.570784	-7.601526
2	Ain Diab	8722000	623	14000	13216.88	33.596236	-7.619264
3	Ain Diab	8400000	800	10500	13216.88	33.593722	-7.621778
4	Ain Diab	9672000	806	12000	13216.88	33.591340	-7.624160

Once the data are all gotten and well treated and cleaned, it comes time of Data Visualization by plotting data in an interactive Map using Folium library, so we can see the distribution of our data in an engaging way.



Within the syntax of Directions API of Google Maps, the coordinates of the waypoints where you want the route to pass are required, which is why the districts, except those of origin and destination, will be found with the Geocoding API which is the similar to Geopy library. The result is stored in a JSON file which will be normalized and formatted to analyze it in a panda's data frame for better manipulation.

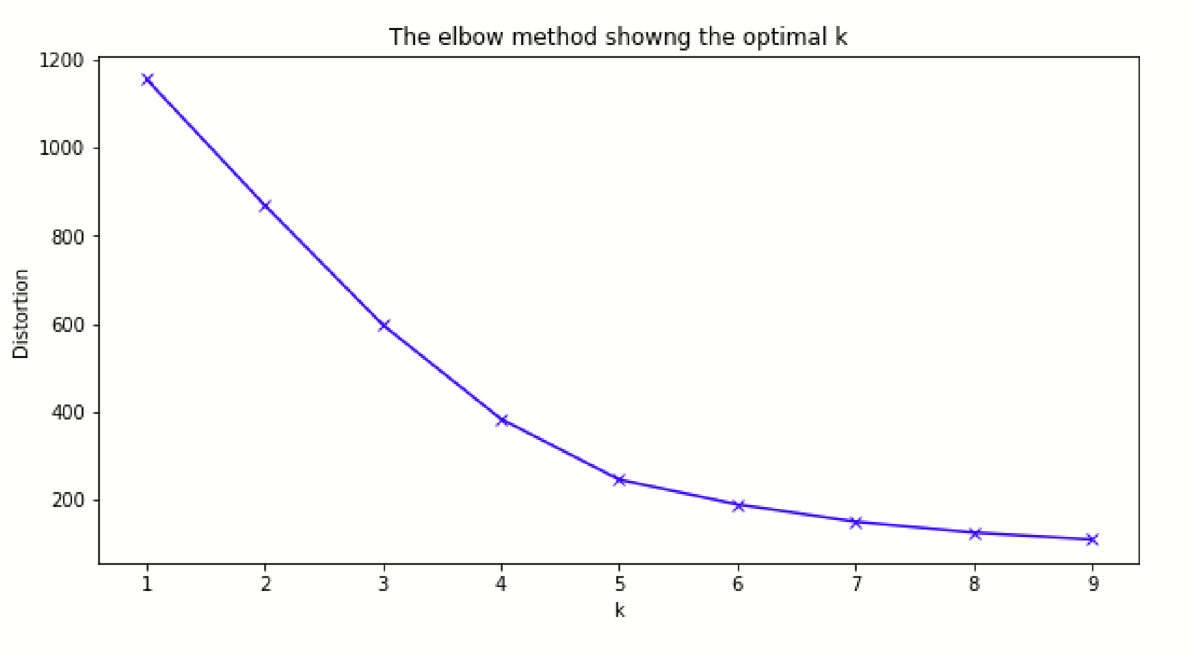


	Instructions	Distance (m.)	Duration (seg.)	start_location.lat	start_location.lng	end_location.lat	end_location.lng
0	Head <b>southeast</b> on <b>Jirón Gral. Vidal<...	10	2	-12.06	-77.05	-12.06	-77.05
1	Turn <b>left</b> at the 1st cross street onto ...	287	81	-12.06	-77.05	-12.06	-77.05
2	Turn <b>right</b> at the 2nd cross street onto...	1103	194	-12.06	-77.05	-12.06	-77.04
3	At the roundabout, take the <b>2nd</b> exit on...	596	111	-12.06	-77.04	-12.06	-77.04
4	At the roundabout, take the <b>4th</b> exit on...	227	59	-12.06	-77.04	-12.06	-77.04
5	Turn <b>right</b> onto <b>Av. República de Chi...	366	52	-12.06	-77.04	-12.07	-77.04
6	Turn <b>left</b> onto <b>Av. Arenales</b>	71	8	-12.07	-77.04	-12.07	-77.04
7	Turn <b>left</b> at the 1st cross street onto ...	250	70	-12.07	-77.04	-12.07	-77.04
8	Turn <b>left</b> at the 2nd cross street onto ...	436	87	-12.07	-77.04	-12.06	-77.04
9	Continue straight to stay on <b>Av. Petit Thou...	26	9	-12.06	-77.04	-12.06	-77.04
10	Turn <b>left</b> onto <b>Av. 28 de Julio</b>	305	61	-12.06	-77.04	-12.06	-77.04
11	At the roundabout, take the <b>4th</b> exit on...	1438	176	-12.06	-77.04	-12.08	-77.04

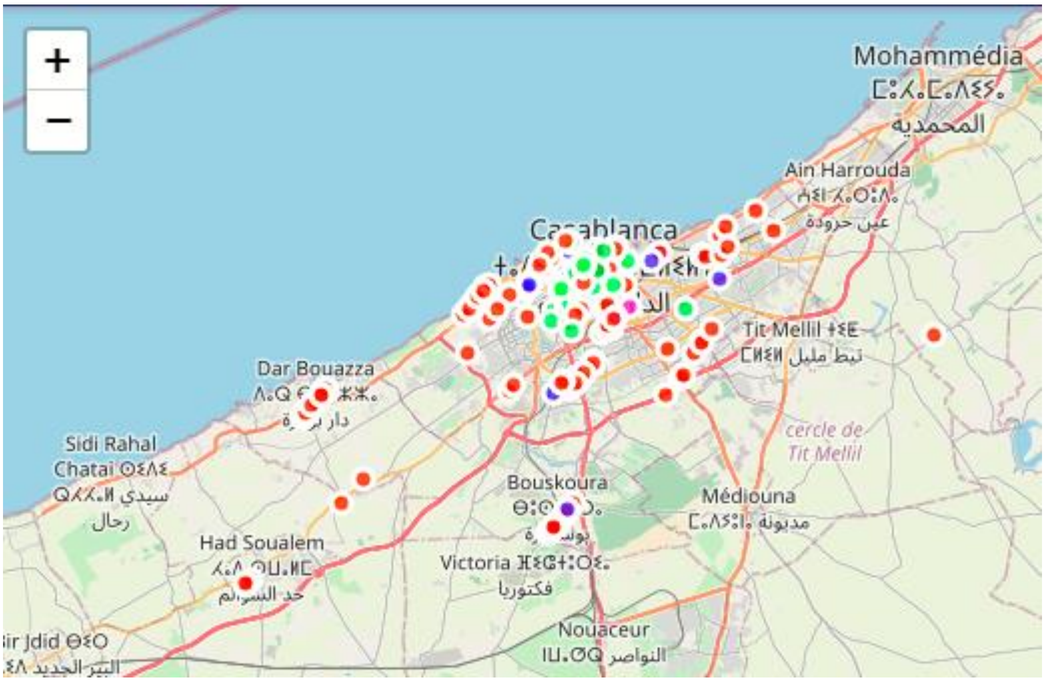
C. AI: Clustered Data

The data was well acquired, treated, analysed and visualized. Now we need a machine learning algorithm to give us deeper insights about it, so we choosed clustering as a way to do the job, using K-Mean algorithm with the Sci-Kit library.

Below is a graph showing the different Ks and their distortion ***'The elbow method'*** that gives the optimal k number to use clusters number in our algorithm.



Finally the only thing missing is the elaboration of the visualization on a map by folium where the instructions of the streets, avenues and highways were placed where mobility must pass to travel through clusters throughout Casablanca city and make the most of the different real estate options that Casa. can offer



D. Results and Discussion

Thanks to Colaboratory plateform this project was possible. After verifying the summary of the internal analysis, we made the web scraping to search of data from the web where they would give us information about the different districts within Casablanca city.

E. Conclusion

In this article conluded that the compass or north of a profitable real estate ads published on web, Google API for the proposed route and web pages for the primary collection of data concerning the districts of Casa.

F. Bibliography

Sarouty ads. Retrieved from https://www.sarouty.ma/fr/recherche?c=1&l=35&ob=pa&t=5

Google. (n.d.). *Google Colaboratory*. Retrieved from https://colab.research.google.com/