**IBM DATA SCIENCE CAPSTONE PROJECT**

**INTRODUCTION**

The capstone project aims to give a taste of what data scientists do in real life when working with data. The project requires the knowledge of location data and different location data providers, such as Foursquare. One also needs the understanding of how to make RESTful API calls to the Foursquare API to retrieve data about venues in different neighborhoods around the world. It also encourages one to be creative in situations where data is not readily available by scraping web data and parsing HTML code. Python and its pandas library are utilized to manipulate data, which help to explore and analyze data. Folium library is also used to great maps of geospatial data and to communicate results and findings.

**BUSINESS PROBLEM**

An entrepreneur would like to open a café that specializes in tea, cocoa and cheesecakes in Dubai. Being a city with diverse population and a tourist hub, Dubai is considered great choice for café and restaurant owners to begin or expand their business. The entrepreneur would like for the café to be affordable and easily accessible to the public.

Taking into account the price level and easy access to public, the café needs to be located within a community in Dubai that has a good amount of footfall of the general public as well as the tourists.

Being a resident of Dubai for 20 years now, the intention behind is to derive optimal communities within Dubai for the location of the restaurant using unsupervised machine learning in addition to application of knowledge of the environment of the city.

Although this business problem is specific for a particular café owner, this model can also be extrapolated to the audience of any potential entrepreneur looking to open a new restaurant or café.

**OBJECTIVE**

The primary objective of this project was to use K-Means Clustering to identify the optimal list of communities in Dubai that would be a great fit for our entrepreneur to open their café.

**DATA**

To build this model, 3 different data sources will be used:

1. List of Communities in Dubai

<https://en.wikipedia.org/wiki/List_of_communities_in_Dubai>

This data will be retrieved from the URL using Web Scraping. The *pandas* package on Python will be used to retrieve this data.

1. Geospatial data of the Communities in Dubai from the above list

The latitude and longitude of the communities in Dubai will be retrieved by using the *geocoder* package on Python.

This data will then be merged with the data obtained from Wikipedia to create the base data.

**BLOCKER:**

The *geocoder* package did not retrieve the right coordinates for the various communities in Dubai. This was identified when the communities were superimposed over Dubai’s map using Folium.

Therefore, the coordinates for the communities were retrieved manually from Google Maps and was consolidated in a CSV file. This was then loaded using the *pandas* package read\_csv command. The link to the CSV file can be found below:

<https://github.com/varsha30051996/datascience-personal/blob/master/Dubai.csv>

1. Top Venues per Community

The top venues per community will be retrieved by using Foursquare through an API by using the data collected in points 1&2 as base data.

**METHODOLOGY**

1. **DATA PREPARATION AND EXPLORATION**

***Step 1: Webscraping of Wikipedia page with the list of communities in Dubai***

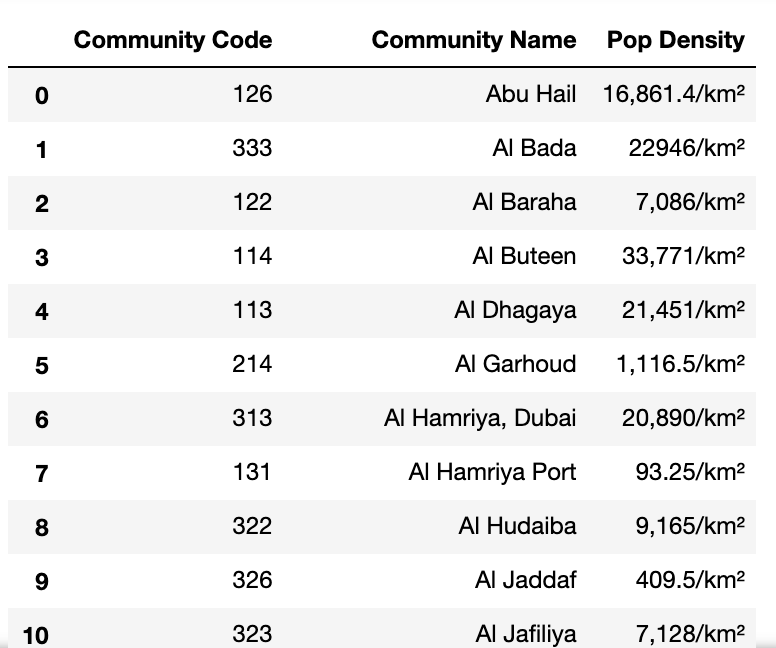
To get the first dataset containing the list of communities of Dubai, the Wikipedia was scraped using the **pandas** package *read\_html* command.

The below dataframe was created containing the list:



In the above data, there are lot of unnecessary columns as well as NaN values in Population Density which is an important feature to consider while choosing the optimum list of communities. Therefore, data cleaning was done on this base dataset to remove any Community with NaN Population Density. Also, Downtown Dubai’s community number which was missing was added manually since it is an important community in Dubai.

After all the necessary cleaning, the final base data was obtained as follows:



Out of 130 communities in the initial file, 91 of them had all their population density populated. Therefore, we restrict our analysis to these 91 areas.

**Assumption:**

Of the 130 areas, only 91 of them had all their fields populated. Therefore, we discard the other 29 from consideration. Since most of these were industrial areas or outskirts, its safe to discard them.

With the base data ready, the next step was to identify the geographical coordinates for these communities.

***Step 2: To get the coordinates of the 91 communities***

To get the coordinates, at first the ***geocoder*** package on Python was used. However, when these coordinates were retrieved from the geocoder package were superimposed over Dubai’s coordinates to create a map using Folium, it was seen that there was a complete mismatch and the package did not retrieve the right coordinates as seen below:



Therefore, the coordinates were filled in manually by obtaining the latitude and longitude via Google Maps search and a CSV file containing the coordinates was created.

Once these new coordinates were superimposed over Dubai’s map using Folium, the following was obtained:



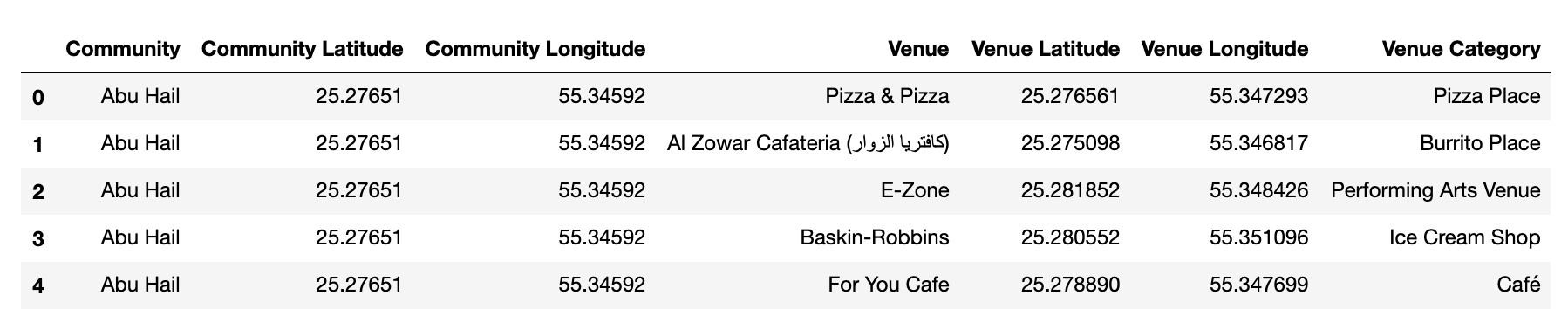
Therefore, the coordinates obtained were correct.

These coordinates were then merged to the base file to create the final input file required.

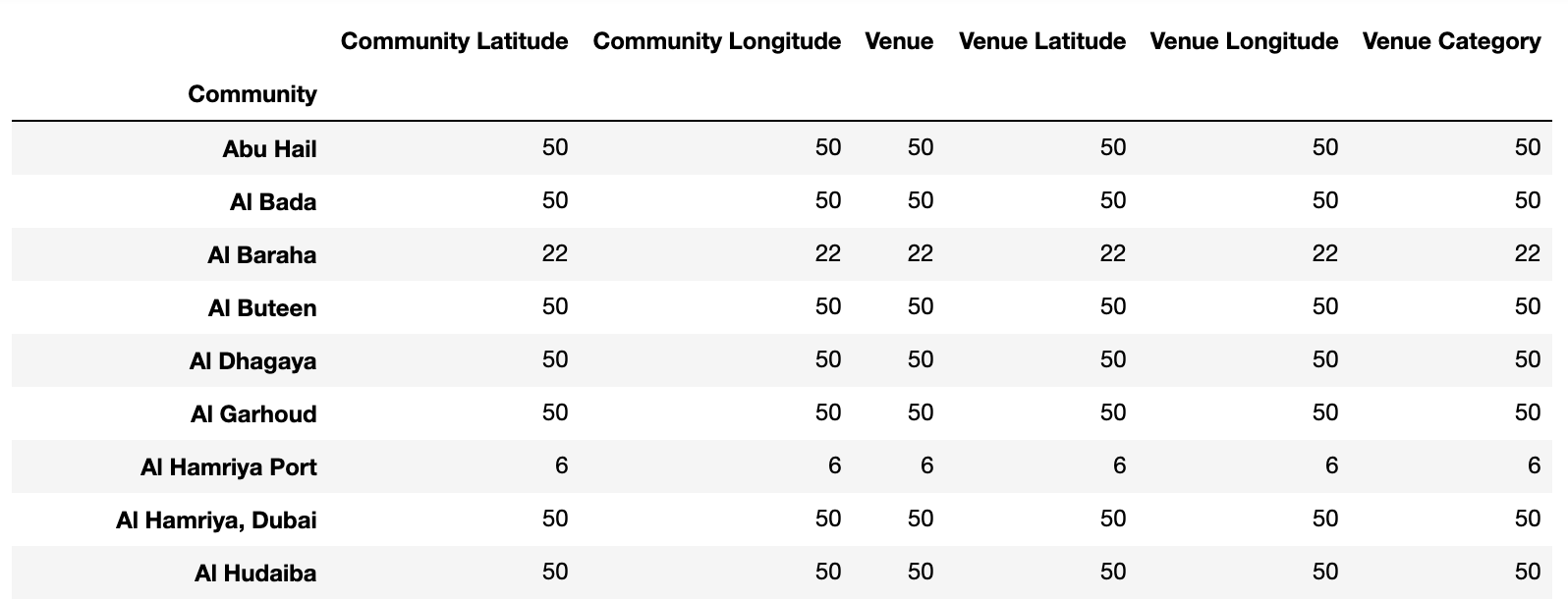
With this, the data preparation and exploration was completed.

1. **SEGMENTATION USING FOURSQUARE API**

In this step, the communities were explored further. Venues were collected for each community using the FourSquare API. Venues within a radius of 1000 m from the community coordinates were considered. The venue’s coordinates along with their category was collected and arranged into the following dataframe:



These venues were then grouped per community to understand the number of venues present in each community:



There were totally **258** unique venue categories found across the communities in Dubai.

For analyzing the communities, the focus is on venue categories to understand which venue category is the most commonly visited in each community so as to optimally place the café.

Therefore, one-hot encoding was performed to generate dummy variables for venue categories to be used for machine learning.

Communities were grouped to get the frequency of occurrence of various venue categories and the top 10 venue categories for each community was obtained as per the frequency of occurrence as below:

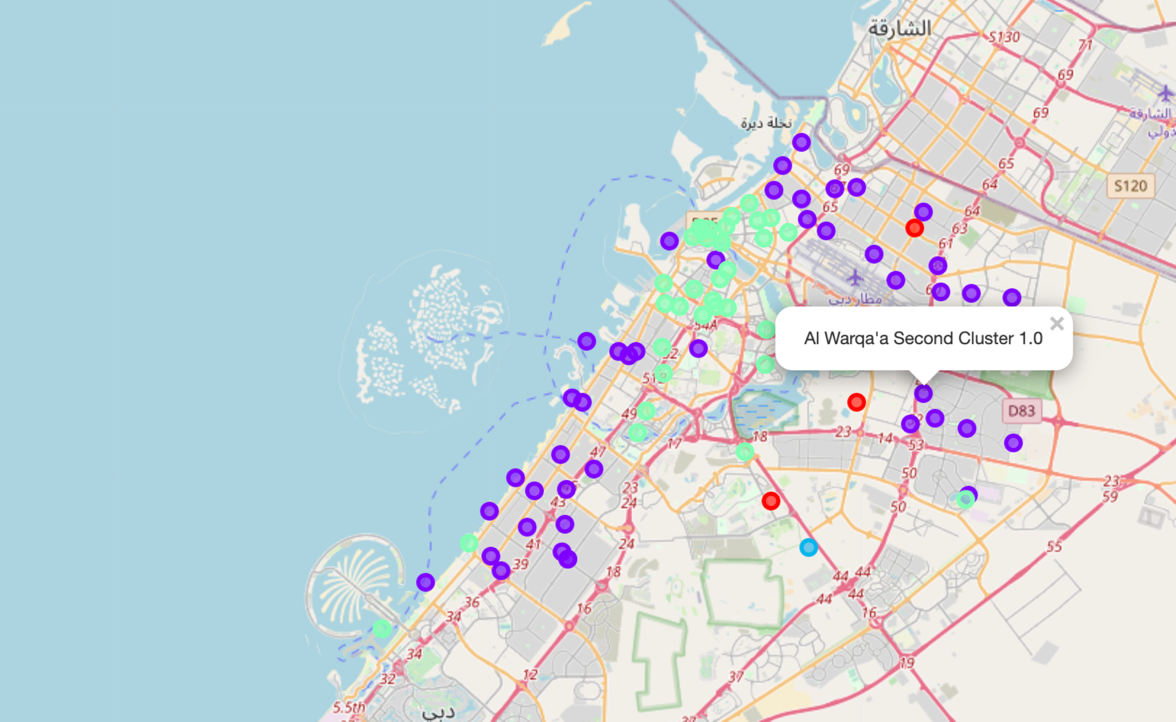
****

1. **CLUSTERING**

K-Means Clustering is performed to cluster the various communities. By the method of trial and error, the communities were clustered into 5 different clusters.



These clusters were then visualized on a map of Dubai using Folium library.



***LIMITATIONS FOR CLUSTERING:***

1. *The analysis was only performed on 91 out of the 130 communities in Dubai due to missing data.*
2. *The analysis was performed on a community level in Dubai.*
3. *While collecting the venues visited from FourSquare, a 1000 meter radius restriction from the community coordinates was applied and the number of collected venues was restricted to 50 per community.*

**RESULTS**

From the K-Means Clustering performed, it was seen that the Clusters 3 and 1 were the most popular.

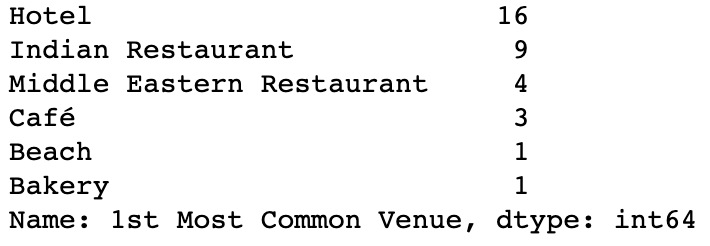
These clusters were then further analyzed to understand the 1st common venue for the various communities present within the cluster.

**Cluster 1:**



For Cluster 1, it was seen that 11 communities had coffee shop as their most common venue and 9 of them as Café.

**Cluster 3:**

****

Cluster 3 seemed to have a lot of visits to hotels and Indian Restaurants and did not seem like a good fit for a café.

**DISCUSSION**

***From the results, it is clear that Cluster 1 hosts the optimal list of communities wherein the Café can be opened.***

Any area within Cluster 1 could be chosen as a potential area to open the café.

However, further analysis was done by arranging the communities within the Cluster by population density.



When looking at the top 5 communities within Cluster 1 by population density, it seen that Al Bada and Mamzar have their top common venue as a café.

**So, these 2 could be the optimal communities to open a café.**

**CONCLUSION**

In this project, 12 communities were identified within Dubai which have most common visits to a café. Of these 12 communities, 2 of them had a high population density where we can expect high footfall – Al Bada and Mamzar. These 2 communities were chosen as optimal communities to open the café.

Although this analysis was restricted to finding the optimal location for a café. This can be utilized by any entrepreneur to understand where to open their business such as a restaurant, bakery, gym, etc.

**APPENDIX**

The Python file on which the project was done can be found on the below link:

<https://github.com/varsha30051996/datascience-personal/blob/master/Final%20Capstone%20Project.ipynb>