‘Varaipadam’ - Exploratory Analysis of Geolocational Data

Aditya Prakash Singh , 19016, Department of Electrical Engineering & Computer Science

Ayush Yadav , 19070, Department of Data Science and Engineering

Indian Institute of Science Education and Research, Bhopal

1. Project Overview

1.1. Objective

This project involves using K-Means Clustering to find the best accommodation for students in Bhopal city by classifying accommodation (Hotels) for incoming students based on their preferences on amenities, budget, and proximity to the location.

1.2 Project Context

In the fast-moving, effort-intense environment that the average person inhabits, It’s a frequent occurrence that one is too tired to fix oneself a home-cooked meal. And of course, even if one gets home-cooked meals every day, it is not unusual to want to go out for a good meal every once in a while for social/recreational purposes. Either way, it’s a commonly understood idea that regardless of where one lives, the food one eats is an essential aspect of the lifestyle one leads. Imagine a scenario where a person has newly moved into a new location. They already have certain preferences, certain tastes. It would save both the student and the food providers a lot of hassle if the student lived close to their preferred outlets. Convenience means better sales and saves time for the customer.

Food delivery apps aside, managers of restaurant chains and hotel managers can also leverage this information. For example, suppose a restaurant manager already knows the demographic of his current customers. In that case, they’d ideally want to open at a location where this demographic is at its highest concentration, ensuring short commute times to the location and more customers served. If potential hotel locations are being evaluated, a site that caters to various tastes would be ideal since one would want every guest to have something to their liking.

2. Approach

The High-Level approach of this project is as follows:

- **Data Collection:** Fetching datasets from relevant locations/sources.
- **Data Cleaning/Preparation:** Using Pandas to clean the datasets and prepare them for analysis and visualization.
- **Visualization:** Using Matplotlib, Seaborn, Pandas to visualize the data using different graphs and plots. E.g., boxplots, etc.
- **REST API:** Using Google Places API to fetch geolocational data.

- **ML Algorithm:** Applying the K-Means Clustering algorithm to cluster the locations using the ScikitLearn module.
- **Map Plotting:** Present the findings on Map using Folium and Seaborn.

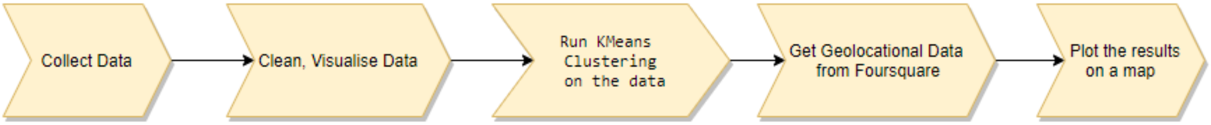

Collect Data
Clean, Visualise Data
Run KMeans Clustering on the data
Get Geolocational Data from Foursquare
Plot the results on a map

Fig. 1. The approach towards the objective.

3. Applications

The K-Means clustering algorithm is used in a variety of applications in real life, like:

- Determining the academic performance of students by grouping them by their learning rate.
- Diagnostic systems (grouping system faults under various reasons).
- Grouping search results from search engines.
- Wireless sensor networks (Mapping networks)

The Google Places API can be used for various purposes, like:

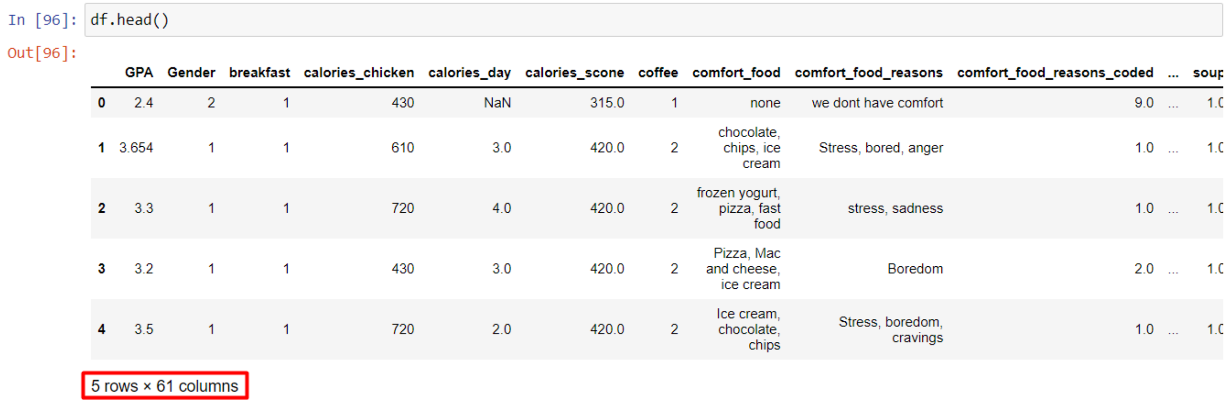
- Building a restaurant review app like Swiggy Zomato etc.
- Supporting a ride sharing service like Uber Pool.

4. Data Preparation and Cleaning

- The data used for this project is the '*food_coded.csv*' data, which was downloaded from here.
- This dataset was explored using the pandas library and over 70 parameters (columns) were found. All of them were not useful.
- The more useful and relevant features required for the Project's objective were sorted out and extracted into a pandas dataframe.
- Also, the rows (entries) with missing values were either removed or the missing values were replaced by some other values to complete the data.
- This process of Extracting the features, (and dealing with different kinds of values as well as NaN values) is known as Data Cleaning.

The outcome of this task is as follows:

Before Cleaning:



```
In [96]: df.head()
```

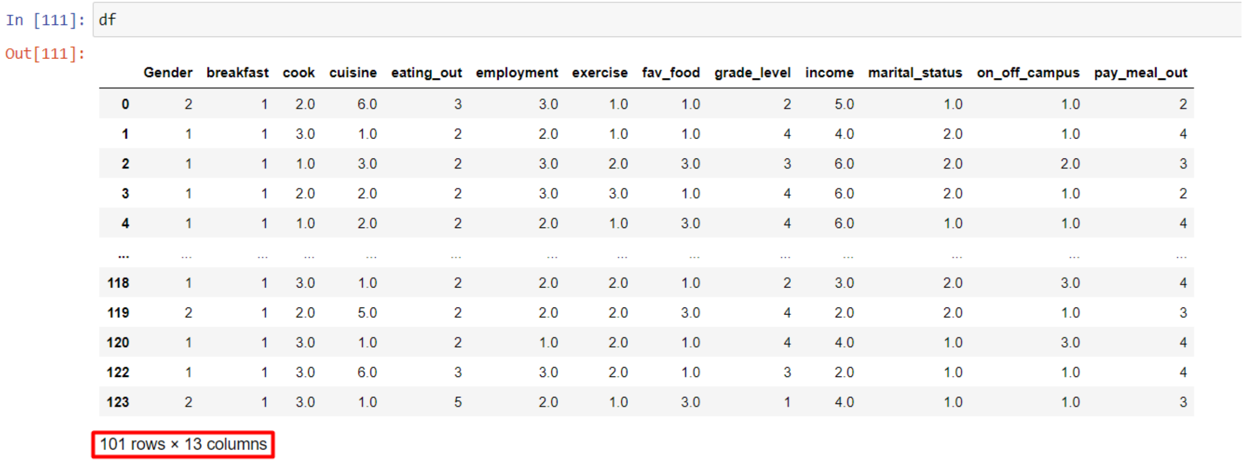
Out[96]:

	GPA	Gender	breakfast	calories_chicken	calories_day	calories_scone	coffee	comfort_food	comfort_food_reasons	comfort_food_reasons_coded	...	soup
0	2.4	2	1	430	NaN	315.0	1	none	we dont have comfort	9.0	...	1.0
1	3.654	1	1	610	3.0	420.0	2	chocolate, chips, ice cream	Stress, bored, anger	1.0	...	1.0
2	3.3	1	1	720	4.0	420.0	2	frozen yogurt, pizza, fast food	stress, sadness	1.0	...	1.0
3	3.2	1	1	430	3.0	420.0	2	Pizza, Mac and cheese, ice cream	Boredom	2.0	...	1.0
4	3.5	1	1	720	2.0	420.0	2	Ice cream, chocolate, chips	Stress, boredom, cravings	1.0	...	1.0

5 rows × 61 columns

Fig. 2. First 5 rows of the complete raw data.

After Cleaning:



```
In [111]: df
```

Out[111]:

	Gender	breakfast	cook	cuisine	eating_out	employment	exercise	fav_food	grade_level	income	marital_status	on_off_campus	pay_meal_out
0	2	1	2.0	6.0	3	3.0	1.0	1.0	2	5.0	1.0	1.0	2
1	1	1	3.0	1.0	2	2.0	1.0	1.0	4	4.0	2.0	1.0	4
2	1	1	1.0	3.0	2	3.0	2.0	3.0	3	6.0	2.0	2.0	3
3	1	1	2.0	2.0	2	3.0	3.0	1.0	4	6.0	2.0	1.0	2
4	1	1	1.0	2.0	2	2.0	1.0	3.0	4	6.0	1.0	1.0	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...
118	1	1	3.0	1.0	2	2.0	2.0	1.0	2	3.0	2.0	3.0	4
119	2	1	2.0	5.0	2	2.0	2.0	3.0	4	2.0	2.0	1.0	3
120	1	1	3.0	1.0	2	1.0	2.0	1.0	4	4.0	1.0	3.0	4
122	1	1	3.0	6.0	3	3.0	2.0	1.0	3	2.0	1.0	1.0	4
123	2	1	3.0	1.0	5	2.0	1.0	3.0	1	4.0	1.0	1.0	3

101 rows × 13 columns

Fig. 3. Cleaned data.

5. Data Exploration and Visualisation

Now that we have our data, we need to understand it. A good way to do this is by visualising the data via graphs. Graphs help us quickly get a sense of the data, and are a much more user-friendly way of understanding data as compared to reading thousands of rows of data.

A good graph to look at distributed groups is a Boxplot. It can tell us at glance where the population is concentrated, and how the outliers compare to the average object in the group.

Data analysis and visualization is done by understanding the correlation between the different parameters (columns) of the dataset and this is done by plotting different graphs and charts.

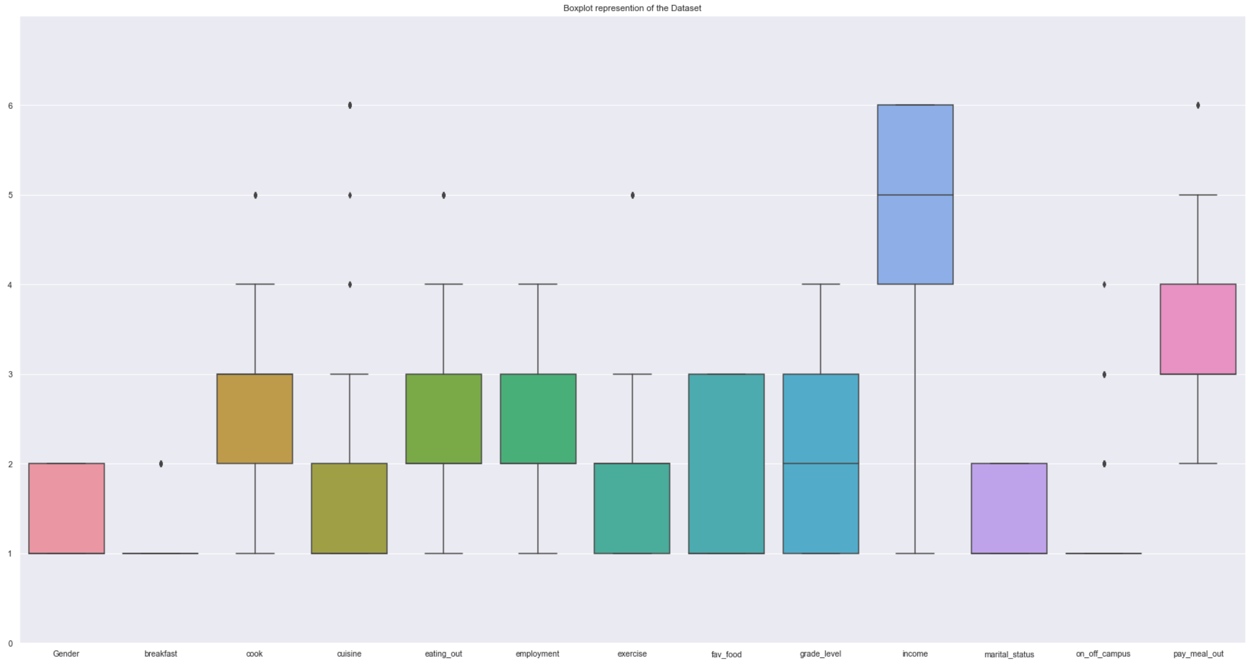
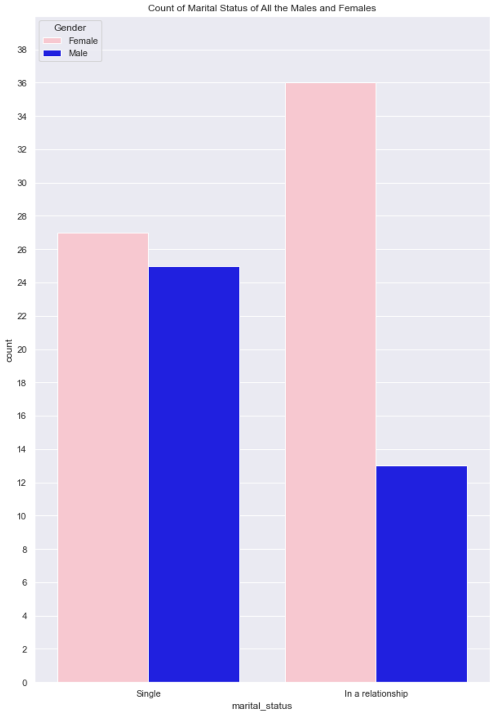


Fig. 4. A Boxplot of the cleaned data.



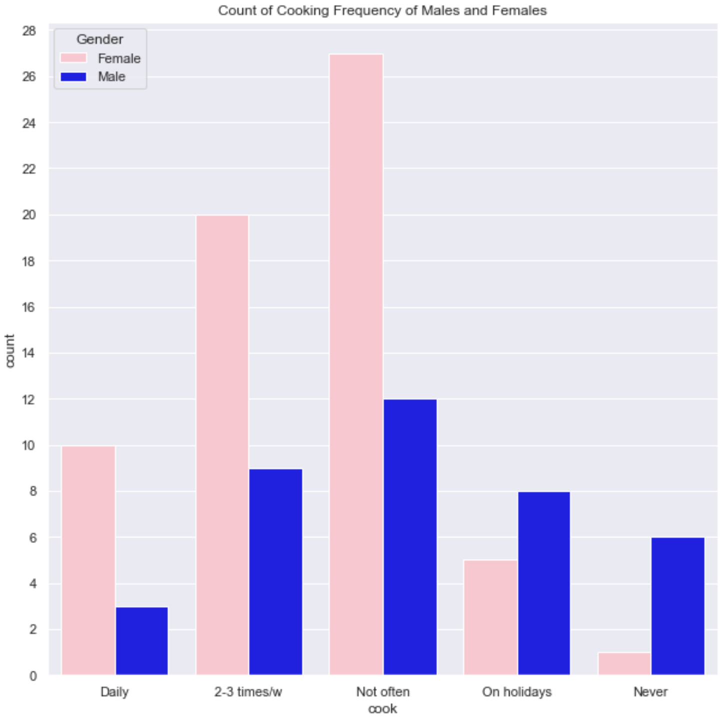


Fig. 5. Gender and Marital Status Countplot

Fig. 6. Gender and Cooking Frequency Countplot

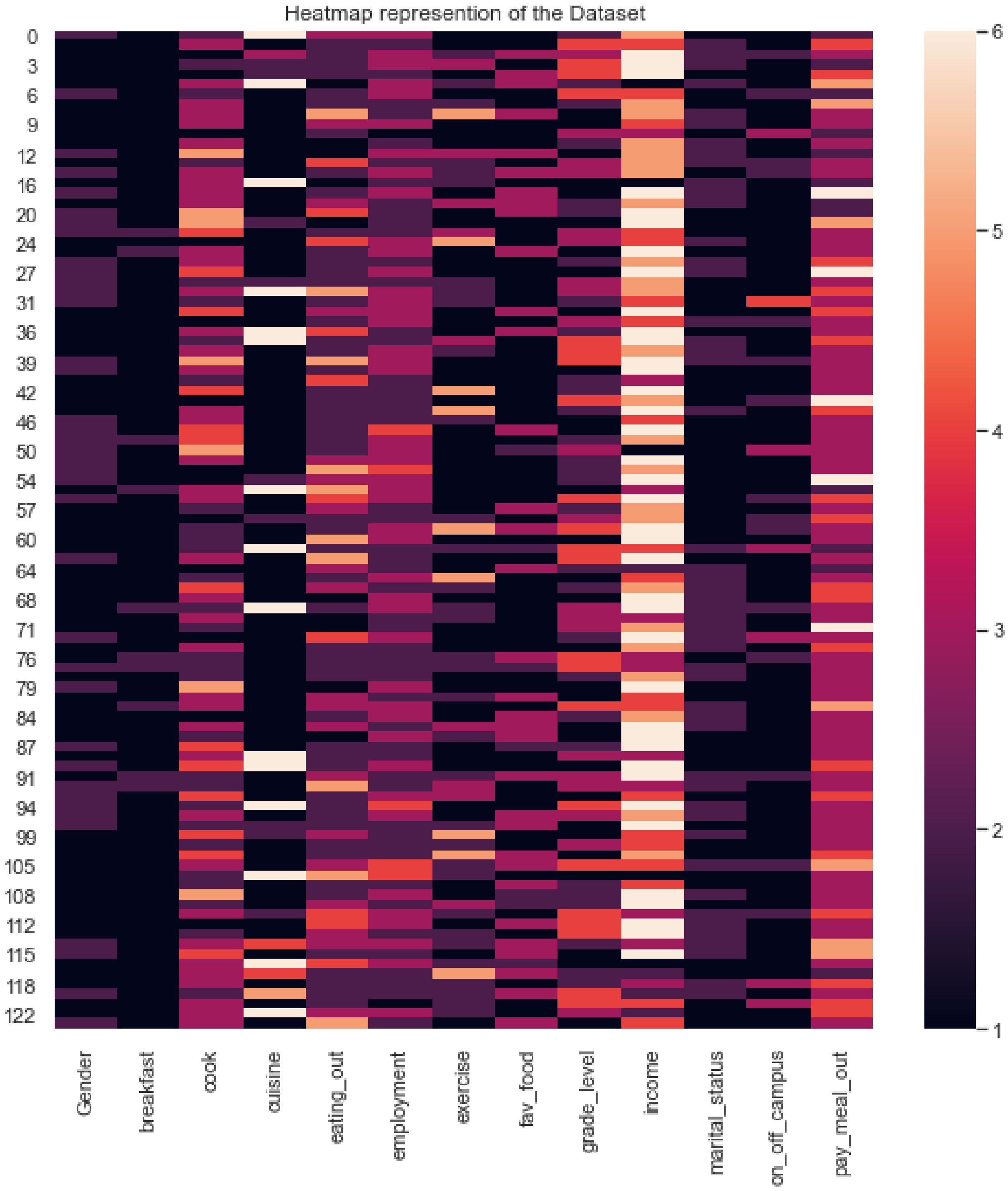


Fig. 7. Heatmap of the completely cleaned data.

6. Run KMeans Clustering on the Data

What is K-Means Clustering? K-Means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

We will run the K-Means algorithm and figure out the best value for K, which will be used later. Applying the K-Means algorithm on the dataset will help us organise the population into groups. Further, we will apply the algorithm again on a different dataset.

In our project, we have used the Elbow Method to figure out the best optimum value for K, which is where the clusters are clearly demarcated on particular attributes. E.g., income (in our case).

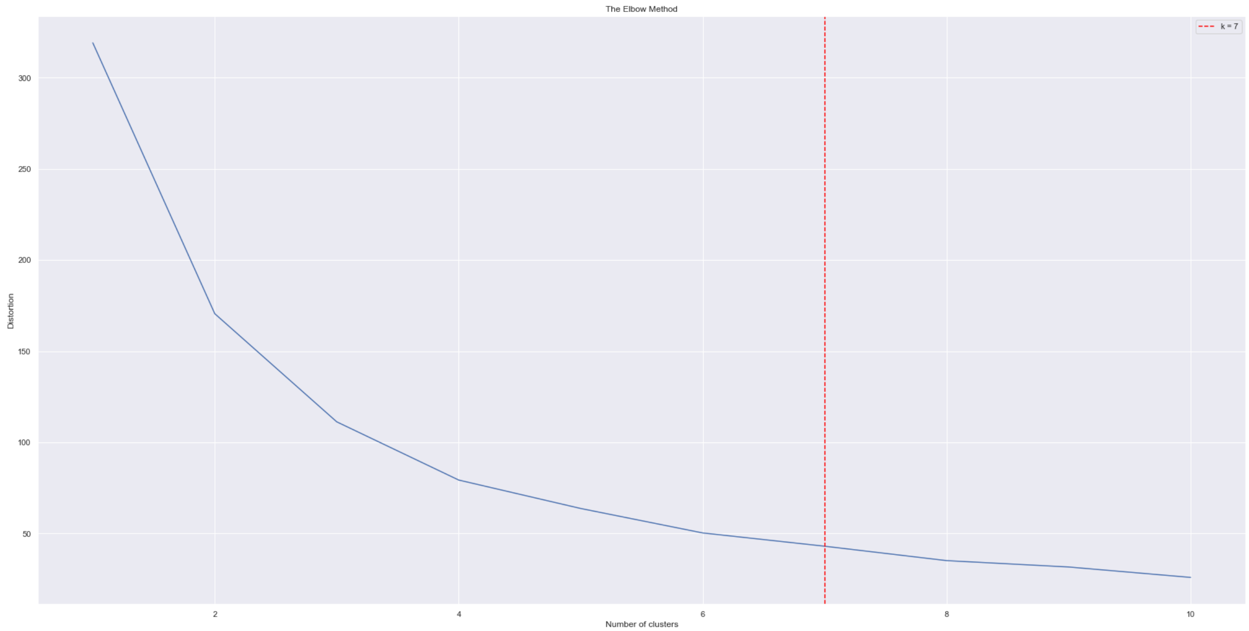


Fig. 8. The Elbow Method graph for finding the best K value.

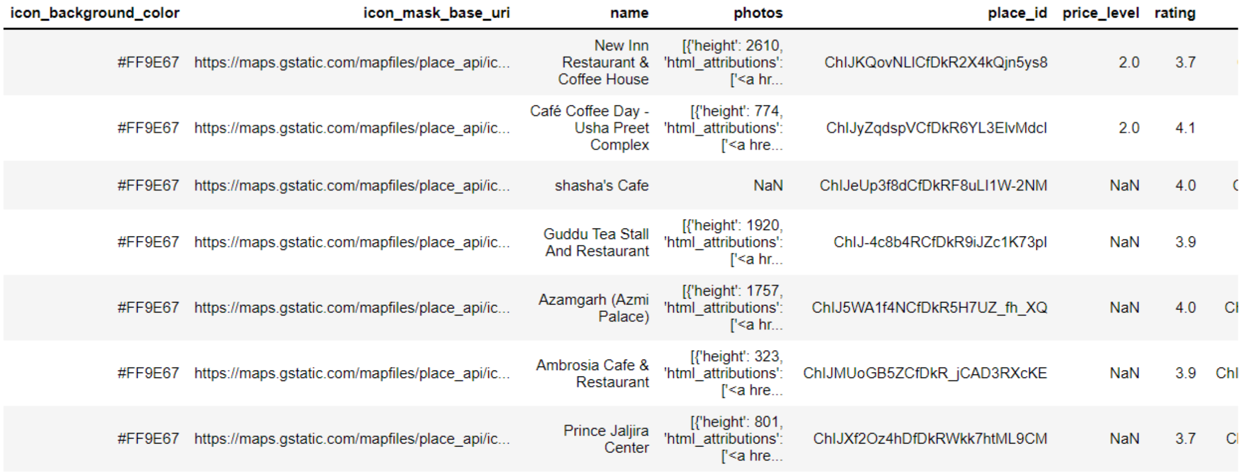
7. Get Geolocational Data from Google Places API

A working API key from Google is required to continue to this part of the project. API credentials were set up and the API was used for the retrieval of the geolocational data.

We can get all different kinds of locations like restaurants, gyms, hotels, etc using search queries. An API response usually looks like this:

```
{'business_status': 'OPERATIONAL',
          'geometry': {'location': {'lat': 23.236551, 'lng': 77.4001953},
                    'viewport': {'northeast': {'lat': 23.2379460302915,
                                        'lng': 77.4015009802915},
                              'southwest': {'lat': 23.2352480697085,
                                        'lng': 77.3988030197085}}},
          'icon': 'https://maps.gstatic.com/mapfiles/place_api/icons/v1/png_71/restaurant-71.png',
          'icon_background_color': '#FF9E67',
          'icon_mask_base_uri': 'https://maps.gstatic.com/mapfiles/place_api/icons/v2/restaurant_pinlet',
          'name': 'New Inn Restaurant & Coffee House',
          'opening_hours': {'open_now': False},
          'photos': [{'height': 2610,
                   'html_attributions': ['<a '
                                  'href="https://maps.google.com/maps/contrib/108022963919361589326">Dhirendra '
                                  'D</a>'],
                                                                                'photo_reference':
'Aap_uEBencBGaxWrBuW25_PNbpehSIfaXaSNRN6q32xdHcvSj82AKxOSbcpVLmJpHeQjLXpqk5YtNmRJ4zvnm1ev-SMbiPfRk_
1uzsMrAHB9NdD1txzgnbuxxMTmFUEpSKFrn_Emah8utMPwH9-2mtrchC6FBPmGKIcg8TuYtSsOQOpS9uY',
                   'width': 4640}],
          'place_id': 'ChIJKQovNLlCfDkR2X4kQjn5ys8',
          'plus_code': {'compound_code': '6CP2+J3 Bhopal, Madhya Pradesh, '
                                'India',
                    'global_code': '7JMV6CP2+J3'},
          'price_level': 2,
          'rating': 3.7,
          'reference': 'ChIJKQovNLlCfDkR2X4kQjn5ys8',
          'scope': 'GOOGLE',
          'types': ['cafe',
                 'restaurant',
                 'food',
                 'point_of_interest',
                 'store',
                 'establishment'],
          'user_ratings_total': 837,
          'vicinity': '28, Bhadbhada Road, New Market, TT Nagar, Bhopal'}
```

Now, we will clean up the data in the same way as before - drop the irrelevant values, handle the NaN values and summarise the results into a dataframe.



icon_background_color	icon_mask_base_uri	name	photos	place_id	price_level	rating	
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	New Inn Restaurant & Coffee House	[{'height': 2610, 'html_attributions': ['<a hr...	ChIJKQovNLICfDkR2X4kQjn5ys8	2.0	3.7	
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	Café Coffee Day - Usha Preet Complex	[{'height': 774, 'html_attributions': ['<a hre...	ChIJyZqdspVCfDkR6YL3ElvMdcl	2.0	4.1	
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	shasha's Cafe	NaN	ChIJeUp3f8dCfDkRF8uLI1W-2NM	NaN	4.0	C
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	Guddu Tea Stall And Restaurant	[{'height': 1920, 'html_attributions': ['<a hr...	ChIJ-4c8b4RCfDkR9iJZc1K73pl	NaN	3.9	
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	Azamgarh (Azmi Palace)	[{'height': 1757, 'html_attributions': ['<a hr...	ChIJ5WA1f4NCfDkR5H7UZ_fh_XQ	NaN	4.0	Cl
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	Ambrosia Cafe & Restaurant	[{'height': 323, 'html_attributions': ['<a hre...	ChIJMUoGB5ZCfDkR_jCAD3RXcKE	NaN	3.9	ChI
#FF9E67	https://maps.gstatic.com/mapfiles/place_api/ic...	Prince Jaljira Center	[{'height': 801, 'html_attributions': ['<a hre...	ChIJXf2Oz4hDfDkRWkk7htML9CM	NaN	3.7	Cl

Fig. 9. API response into a dataframe.

8. Plotting the clustered locations on a Map

Now we will run the K-Means clustering on the API data and plot the results on a map.

Note that here we are applying K-Means on the dataset of the locations which we chose, which will help us find the best location for each population group that we found in the last K-Means clustering.

Now, we run K-Means clustering on the dataset we prepared in the previous milestone, with the optimal K value we found.

Now that we have the results, it's time to visualise them. Using Folium, we will plot our results on a world map, centered on the location we chose i.e. Bhopal.

We define a proper colour scheme so the locations are easily differentiated by the cluster number.

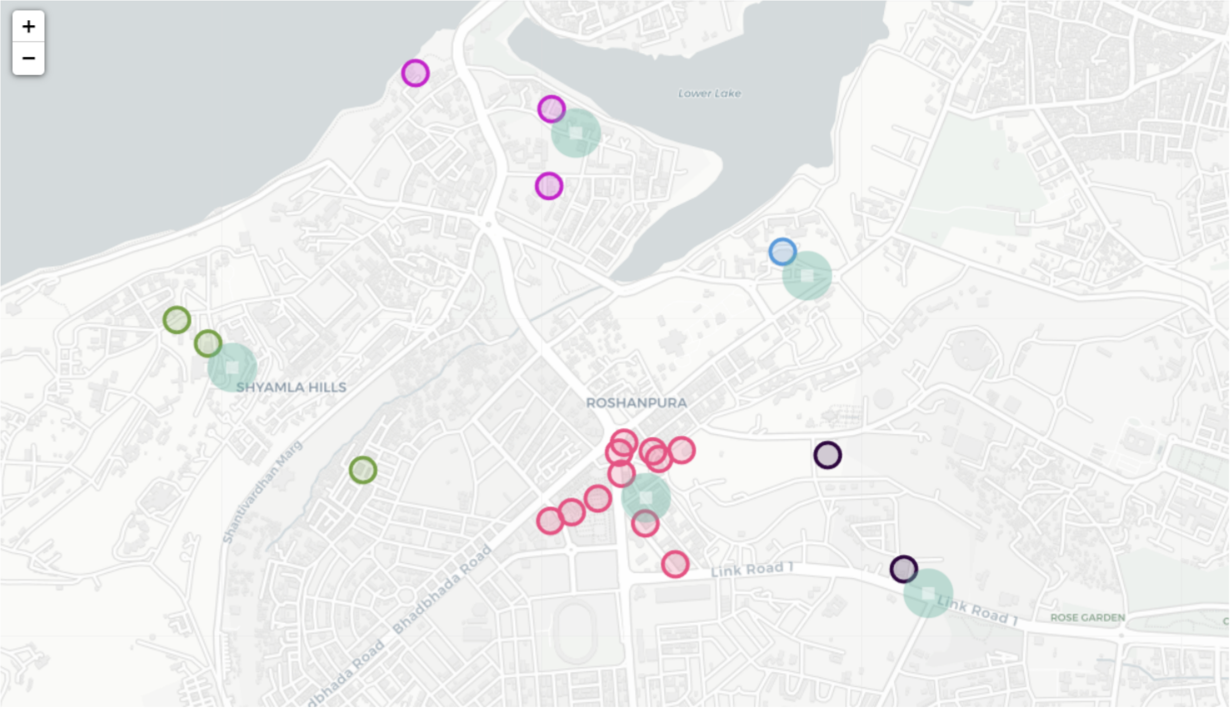


Fig. 10. The resultant map plot for Hotels dataframe.

9. Conclusion

From this project, we can conclude that the K-Means clustering algorithm is very useful for grouping data and if used with the correct dataset, it can give us very promising results and predictions.

10. Acknowledgements

Finally, we would like to thank Dr. Parthiban Srinivasan for providing us this opportunity to work on this project which helped us learn many new things relating to Python and the K-Means ML algorithm.

11. References

[1] Idea: Crio Projects - Exploratory Analysis of Geolocational Data
[2] Dataset: Food choices
[3] References (Python):
    [3.1] Pandas: pandas.read_csv — pandas 1.3.4 documentation
    [3.2] REST API: API reference — seaborn 0.11.2 documentation
    [3.3] Data Cleaning: Guide To Data Cleaning: Definition, Components & How To

[3.4] K-Means Clustering:

[3.4.1] sklearn.cluster.KMeans — documentation

[3.4.2] K-means clustering - sklearn & Python | by Dhiraj K | Heartbeat

[3.5] Folium:

[3.5.1] Quickstart — Folium 0.12.1 documentation

[3.5.2] Python Visualisation — Folium 0.12.1 documentation

[4] References (Statistics):

[4.1] Box Plots: Understanding Boxplots | by Michael Galarnyk

[4.2] K-Means Clustering: k-means clustering

[4.3] Elbow Method: Elbow method (clustering)