Predicting the Most Suitable Place for a Restaurant in Paris

Shubham

December17, 2020

1. Introduction

1.1 Background

Economics revolves around two major factors: Demand and Supply. The success of any financial establishment like hotel, restaurant and supermarket depend largely on demand and supply of a particular region. Demand factors include population and per capita income while supply factors include number of suppliers for that particular service. Restaurants are a part of consumer foodservice industry. In 2017, only in Europe, the market value for consumer foodservice reached 508 billion dollars, where France accounts for 61 billion dollars' market. These figures may look captivating at first, but it comes at a cost. It takes a long and tiring analysis to select a right place for something. All the demand and supply factors should be analyzed very carefully, failing of which may lead to severe consequences like revenue loss or in some cases complete shutdown. From all the influencing factors, population on demand side and number of suppliers on supply side becomes very important for consideration.

1.2 Problem

Data that might contribute to determine the most suitable type and place for a new restaurant are number of people living, per capita income, number and type of existing restaurants, distance from nearest attractions and connectivity to various places. This project aims to predict the suitable type and place for a new restaurant in Paris region, using type and number of existing restaurants and distance from most famous tourist attractions of Paris.

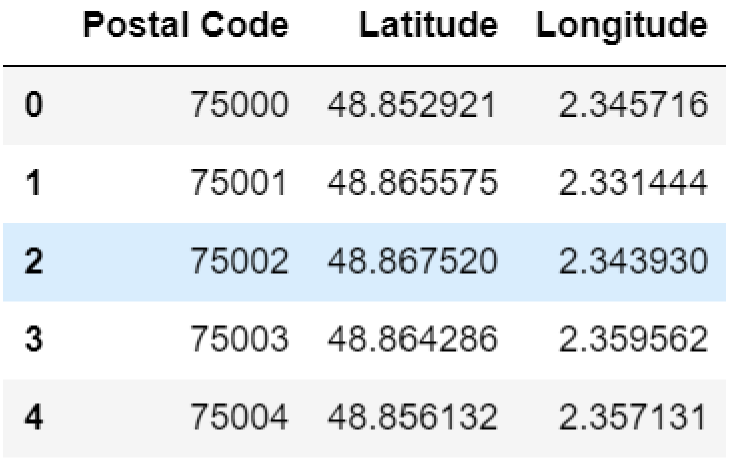
1.3 Interest

Individuals who are interested in opening new restaurant in Paris regions, are the primary users.

2. Data acquisition and cleaning

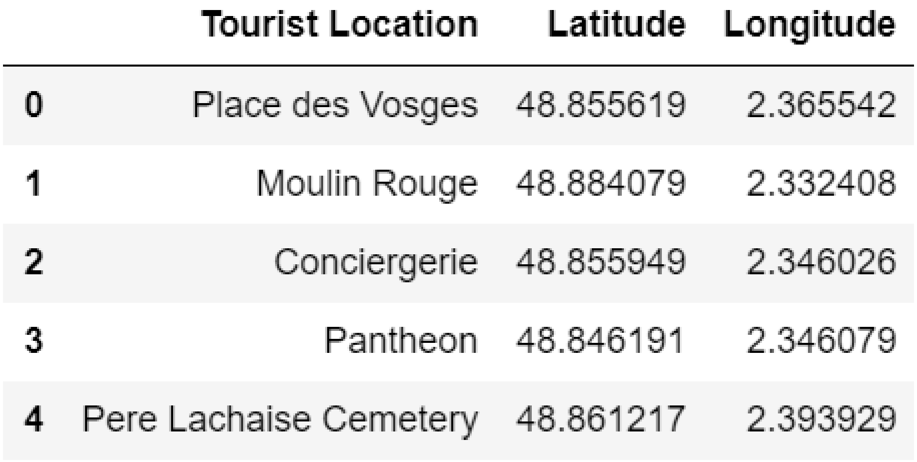
2.1 Data Sources

First, all the postal codes for Paris were scraped from here (Figure 1) and major tourist attractions of Paris were scraped from here (Figure 2). Further, latitude and longitude for each postal code and tourist attraction were obtained from python's geopy library using Nominatim class. Next, Foursquare API was used for exploring each postal code.



	Postal Code	Latitude	Longitude
0	75000	48.852921	2.345716
1	75001	48.865575	2.331444
2	75002	48.867520	2.343930
3	75003	48.864286	2.359562
4	75004	48.856132	2.357131

Figure 1. List of postal codes of Paris



	Tourist Location	Latitude	Longitude
0	Place des Vosges	48.855619	2.365542
1	Moulin Rouge	48.884079	2.332408
2	Conciergerie	48.855949	2.346026
3	Pantheon	48.846191	2.346079
4	Pere Lachaise Cemetery	48.861217	2.393929

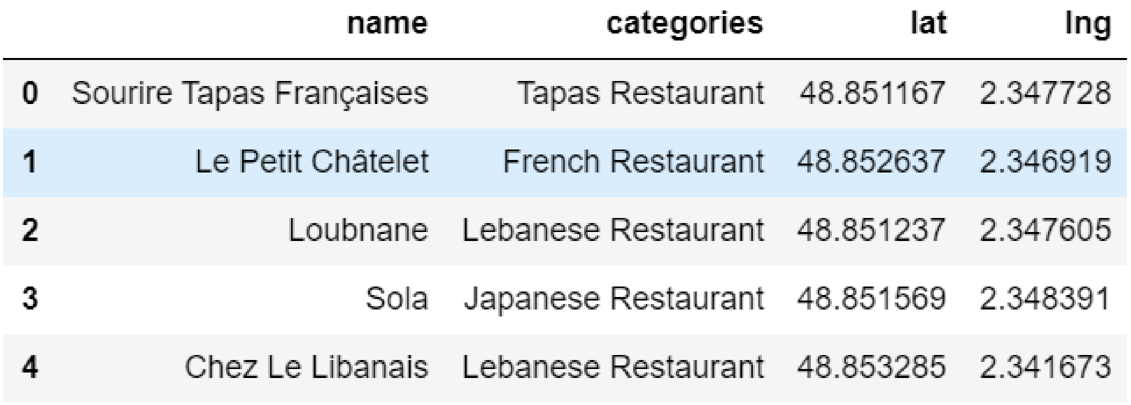
Figure 2. List of top 25 tourist attractions in Paris

2.2 Data Cleaning

All irrelevant data were removed from websites during scarping. Entries whose coordinates could not be obtained by Nominatim, were manually corrected. Foursquare explore API's JSON data was filtered for only food category results. Further, it was filtered by only selecting food outlets, containing restaurant in their name.

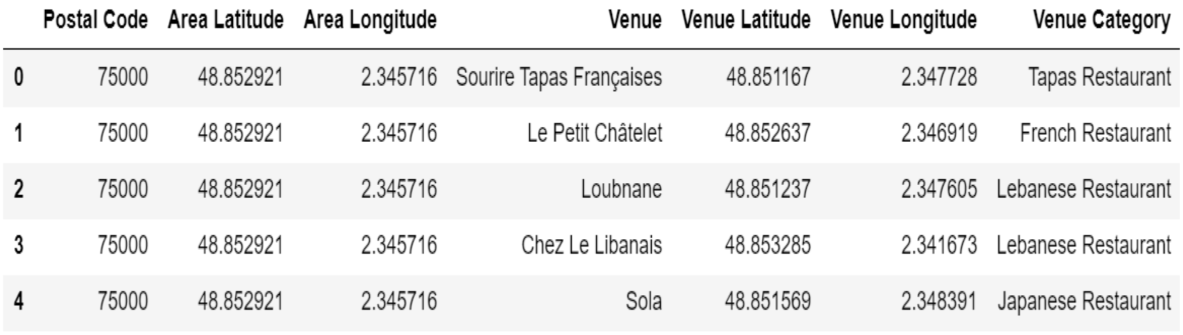
2.3 Feature Selection

Original data frame, obtained from foursquare API using food category, consisted of 2028 samples and many features. Upon examining the features, it could be concluded that many features were redundant for our analysis. Thus, they were removed and most prominent 4 features were selected. Further, it was filtered for restaurants and revamped to 1465 samples and 4 features (Figure 3). Features selected for a venue were - name, latitude, longitude and category. Next, 3 additional features were added namely postal codes and corresponding latitudes and longitudes. So, final data frame used for analysis contained 1465 samples and 7 features (Figure 4).



	name	categories	lat	lng
0	Sourire Tapas Françaises	Tapas Restaurant	48.851167	2.347728
1	Le Petit Châtelet	French Restaurant	48.852637	2.346919
2	Loubnane	Lebanese Restaurant	48.851237	2.347605
3	Sola	Japanese Restaurant	48.851569	2.348391
4	Chez Le Libanais	Lebanese Restaurant	48.853285	2.341673

Figure 3. List of restaurants received from Foursquare API



	Postal Code	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	75000	48.852921	2.345716	Sourire Tapas Françaises	48.851167	2.347728	Tapas Restaurant
1	75000	48.852921	2.345716	Le Petit Châtelet	48.852637	2.346919	French Restaurant
2	75000	48.852921	2.345716	Loubnane	48.851237	2.347605	Lebanese Restaurant
3	75000	48.852921	2.345716	Chez Le Libanais	48.853285	2.341673	Lebanese Restaurant
4	75000	48.852921	2.345716	Sola	48.851569	2.348391	Japanese Restaurant

Figure 4. Final list containing postal code and restaurant venues

3. Exploratory Data Analysis

3.1 Calculation of target variable

It’s evident, “Divide and Conquer” is the most effective approach in computer science. Thus, one postal code region was randomly selected and complete analysis was performed on that. A bar chart was plotted for various categories of restaurant present in this randomly selected region (Figure 5). It was concluded, what kind of restaurants are most common and how well it is situated between top 25 tourist attractions. Further, the analysis was scaled up for all postal codes and corresponding results were presented in both tabular and map format.

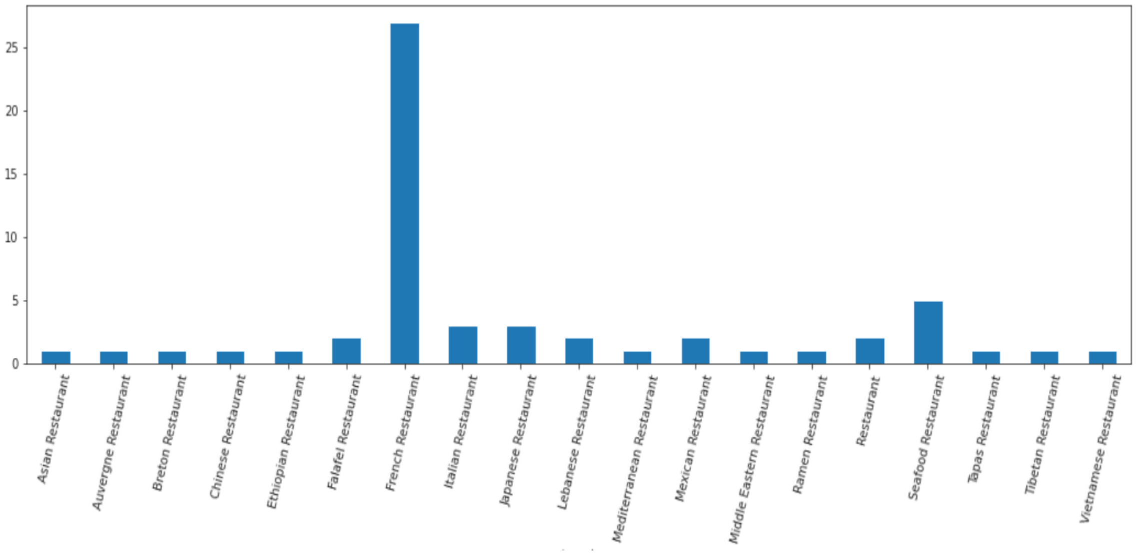


Figure 5. Bar chart showing number of different category of restaurants

3.2 Map for various postal code

To better understand the locations, a map was plotted for the coordinates of all the postal code regions of Paris (Figure6).

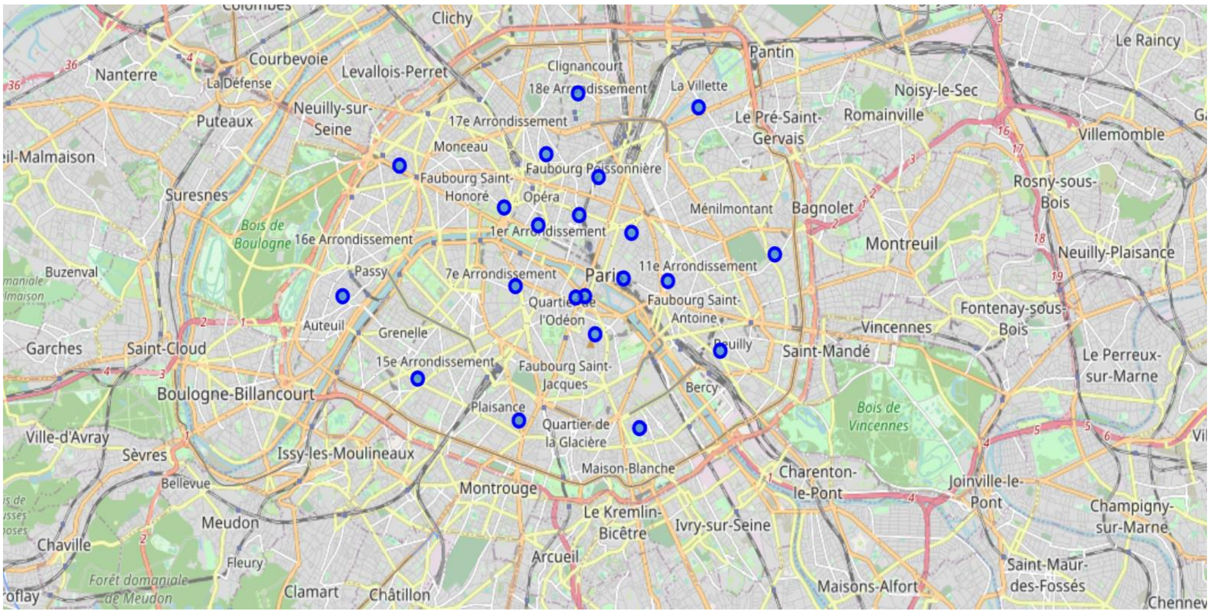
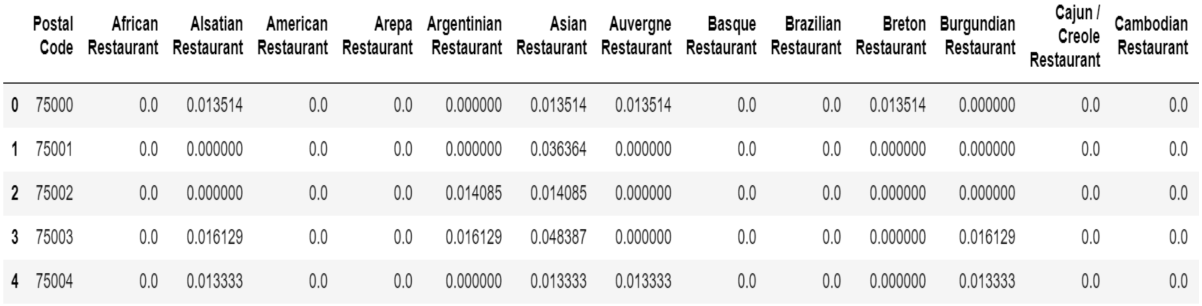


Figure 6. Map showing various postal codes in Paris region

3.3 Grouping results by postal code and category

The final data frame after all the cleaning consisted of 1465 samples and 7 features. For better understanding, it was grouped by postal code and category. Further, mean was calculated based on each category (Figure 7).

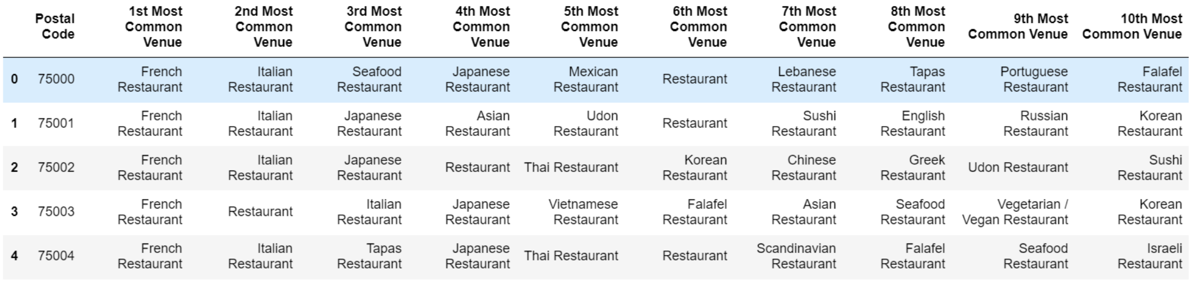


	Postal Code	African Restaurant	Alsatian Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Auvergne Restaurant	Basque Restaurant	Brazilian Restaurant	Breton Restaurant	Burgundian Restaurant	Cajun / Creole Restaurant	Cambodian Restaurant
0	75000	0.0	0.013514	0.0	0.0	0.000000	0.013514	0.013514	0.0	0.0	0.013514	0.000000	0.0	0.0
1	75001	0.0	0.000000	0.0	0.0	0.000000	0.036364	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0
2	75002	0.0	0.000000	0.0	0.0	0.014085	0.014085	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0
3	75003	0.0	0.016129	0.0	0.0	0.016129	0.048387	0.000000	0.0	0.0	0.000000	0.016129	0.0	0.0
4	75004	0.0	0.013333	0.0	0.0	0.000000	0.013333	0.013333	0.0	0.0	0.000000	0.013333	0.0	0.0

Figure 7. List showing each postal region based on mean of restaurant category

3.4 Getting top 10 restaurant categories for a particular region

Top 10 restaurant categories for a region was calculated from mean table designed in previous step (Figure 8).

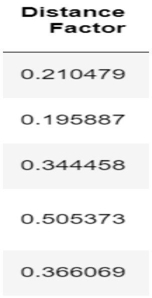


	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	75000	French Restaurant	Italian Restaurant	Seafood Restaurant	Japanese Restaurant	Mexican Restaurant	Restaurant	Lebanese Restaurant	Tapas Restaurant	Portuguese Restaurant	Falafel Restaurant
1	75001	French Restaurant	Italian Restaurant	Japanese Restaurant	Asian Restaurant	Udon Restaurant	Restaurant	Sushi Restaurant	English Restaurant	Russian Restaurant	Korean Restaurant
2	75002	French Restaurant	Italian Restaurant	Japanese Restaurant	Restaurant	Thai Restaurant	Korean Restaurant	Chinese Restaurant	Greek Restaurant	Udon Restaurant	Sushi Restaurant
3	75003	French Restaurant	Restaurant	Italian Restaurant	Japanese Restaurant	Vietnamese Restaurant	Falafel Restaurant	Asian Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant
4	75004	French Restaurant	Italian Restaurant	Tapas Restaurant	Japanese Restaurant	Thai Restaurant	Restaurant	Scandinavian Restaurant	Falafel Restaurant	Seafood Restaurant	Israeli Restaurant

Figure 8. List showing top 10 restaurant categories of a particular region

3.5 Calculation of distance for a region from top 25 tourist attractions

Linear difference between a region’s and a tourist attraction place’s latitude and longitude was calculated. But since some values were negative, it didn’t make much sense to keep them as it is. So, it was squared and further adjusted for ranks on the basis of popularity of tourist places. Further, a suitable scaling factor was multiplied to finally arrive at “Distance Factor” (Figure 9).



Distance Factor
0.210479
0.195887
0.344458
0.505373
0.366069

Figure 9. List showing Distance Factor of various regions

3.6 Combined map for top 25 tourist attractions and all postal regions

A combined map for top 25 tourist attractions (shown using star) and all postal regions (shown using circles) was prepared for better visualization (Figure 10).

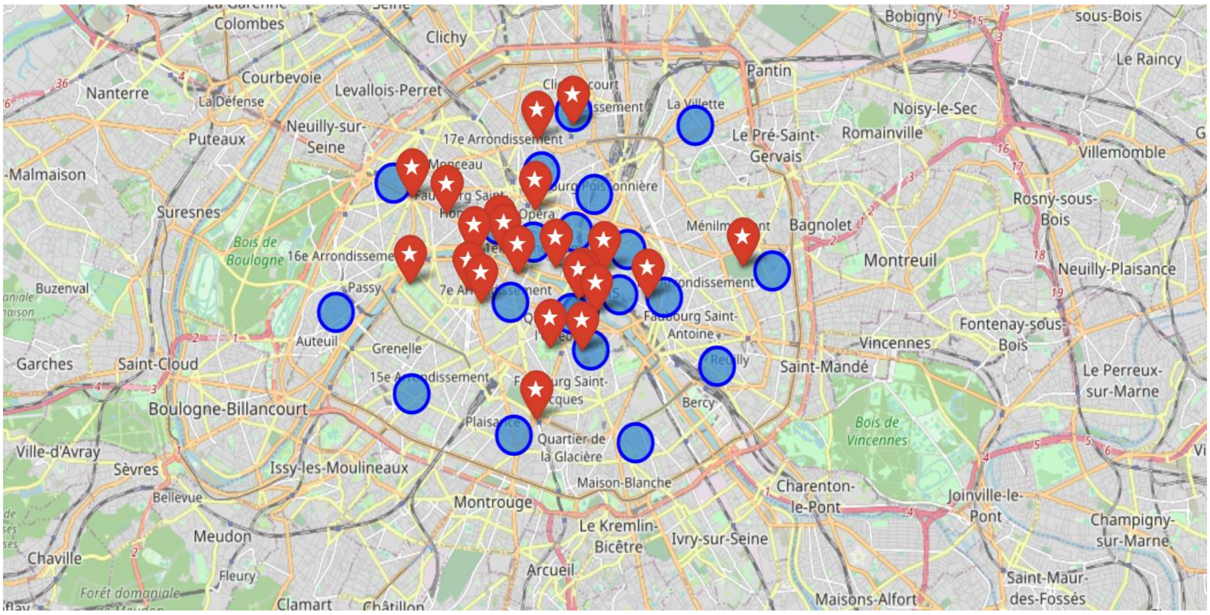


Figure 10. Map showing top 25 tourist attractions and postal regions

4. Predictive Modelling

Since, we need to separate the regions based on similarity of neighbourhood venues, clustering seems to be the most suitable approach. Out of all clustering algorithms, *k-means clustering* is highly useful in these conditions. Target of this project was to predict similar neighbourhoods based on neighbouring type of restaurants, so that an owner can select most suitable area based on type of restaurant that he wants to open.

4.1 Clustering Model

K-Means Clustering was performed for various values of k and corresponding Inertias and Distortion were calculated. Elbow method was used and based on (Figure 11) and (Figure 12), k=3 was selected.

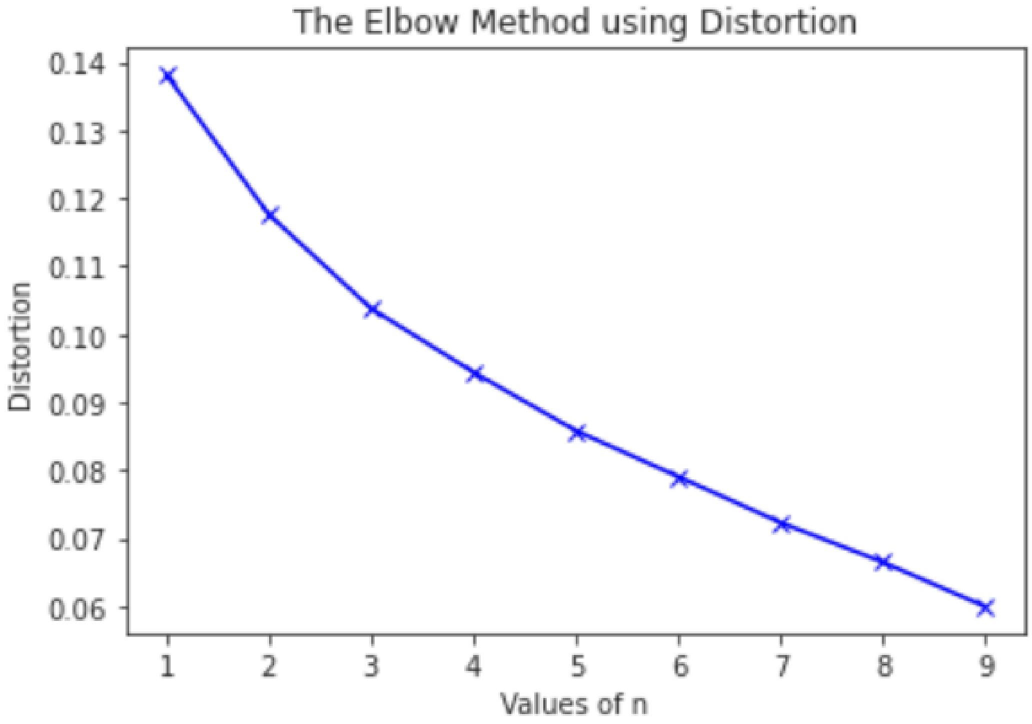


Figure 11. Graph for distortions vs n in K-Means

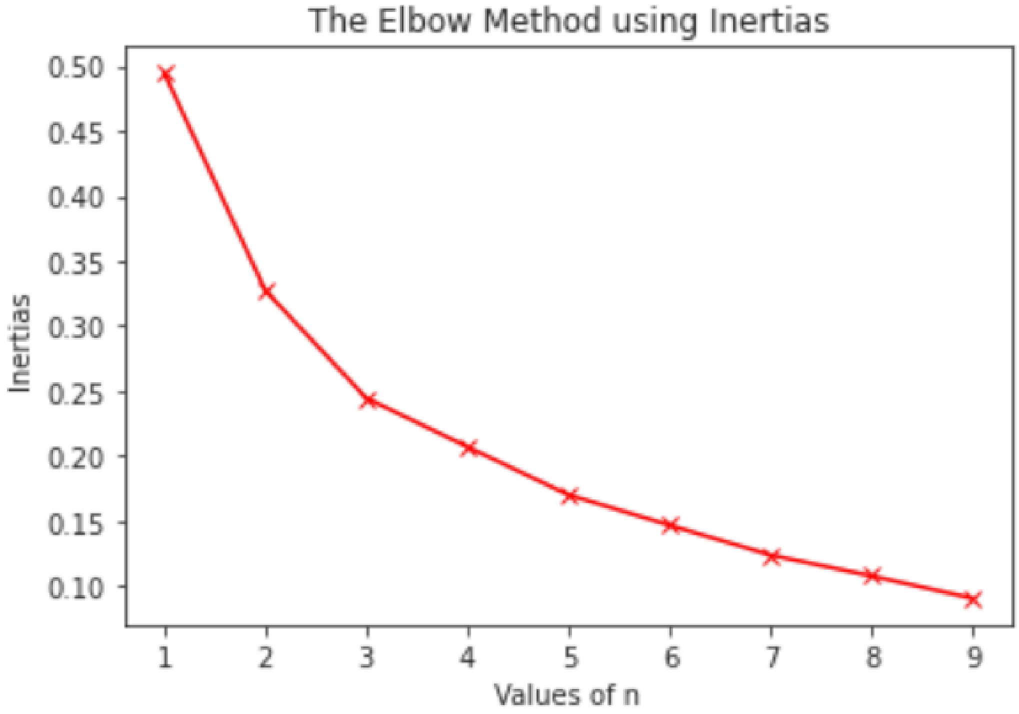


Figure 12. Graph for inertias vs n in K-Means

4.1.1 Applying K-Means Clustering

When K-Means was applied, it led to formation of highly skewed groups of 11, 1 and 9. The group containing 1 restaurant was due to an unusual fact of highly dominating Vietnamese restaurant in 750013 region. This is just an exceptional case and thus cannot be treated as noise.

5. Conclusions

If we take closer look at the groups formed by K-Means Clustering, it can be concluded that:

1. First group (Red outlined circles) contains regions that are in vicinity of top 25 tourist attractions. Thus, a lot of tourists can be expected in these regions. Tourism is very prominent part of Paris, as more than 18 million tourists visited in 2018.
   Further, French restaurants dominate these regions as expected, but next impactful types are Italian and Southeast Asian - Japanese, Seafood and Sushi restaurants. Thus, it can be concluded that these places cater many Southeast Asian tourists round the year.
2. Second group (Green outlined circles) contains just 1 region, which presents an exception as discussed above. It caters significantly huge number of Vietnamese and Thai tourists.
3. Third group (Purple outlined circles) contains regions that are not very close to tourist attractions. Thus, a bit less tourists can be expected in this region.
   Further, French and Italian restaurants dominate these regions as well. But coming down the lane, it can be seen, it caters variety of tourists such as Mexican, Lebanese and Southeast Asian.

Results are presented in the form of a map and suitable markers are added for better visualization (Figure 13).

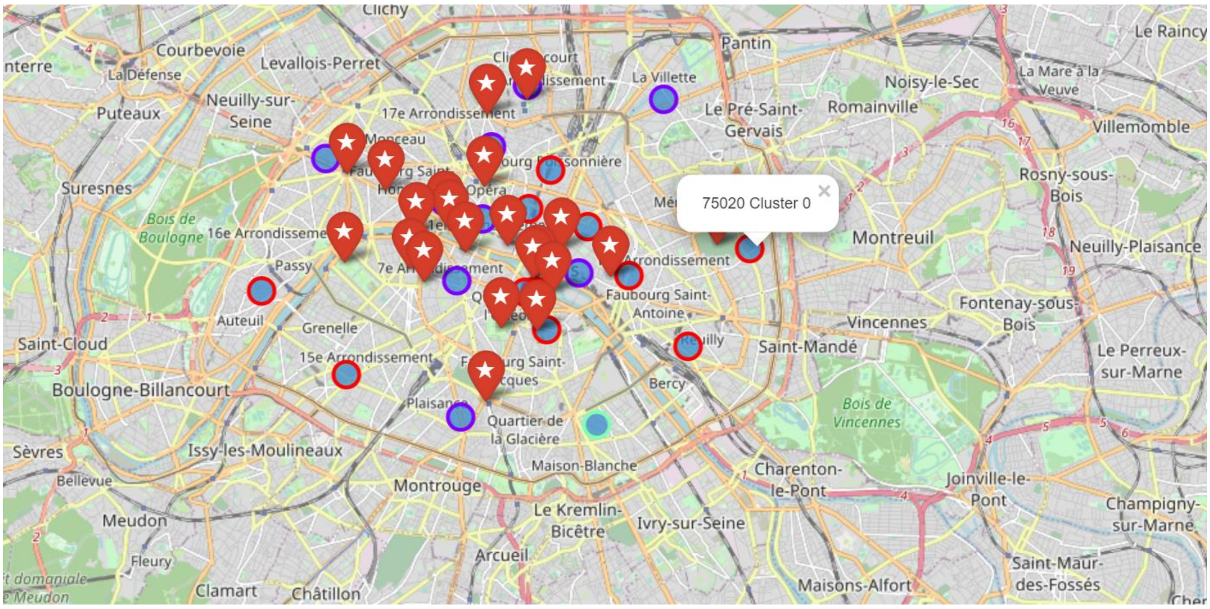


Figure 13. Map showing various clusters in different colours

6. Future Directions

There can be many factors that can be inculcated to further improve the accuracy of our restaurant predictor.

1. Per capita income can be included to show the spending capacity of people living in that area.
2. Good transportation facility means easier accessibility. Thus, it can play a major role.
3. Local taxes may significantly affect the place of interest. So, it can be included as well.

Further, market parameters are unpredictable but they affect every investment in one way or the other. Pandemic like COVID-19 can also never be predicted. So, the model will always have some factor of risk.