**PROJECT REPORT**

RESTAURANT SITE SELECTION IN CALIFORNIA

Preeti Sethi

September 27, 2019

# 

INDEX

[**1.BUSINESS PROBLEM**](#_o0mo8mlrtuu1) **3**

[1.1 INTRODUCTION](#_4x0wn34pxfnb) 3

[1.2 AIM](#_md6x6f38hv17) 3

[1.3 TARGETED AUDIENCE](#_a38yhdwuy65a) 3

[**2. DATA**](#_62i2peawykvg) **4**

[2.1 DATA REQUIREMENT](#_2bnv2qfe1jyt) 4

[2.2 DATA SOURCES](#_l88qv7zfrm5r) 4

[2.2.1 United States cities database](#_ve3mqq6rcped) 4

[2.2.2 The Foursquare’s Places API](#_7ma2rvcc165y) 5

[2.3 DATA CLEANING](#_so42cib1khjm) 5

[**3. METHODOLOGY**](#_o3fehfjycjmj) **6**

[3.1 DATA COLLECTION AND WRANGLING](#_1hbave2z34zu) 6

[3.2 DATA ANALYSIS](#_xdizqfhlhauo) 7

[3.2.1 Analyzing most preferred food places for all the cities](#_x77ej9wgakvg) 7

[3.2.2 Analyzing the best cities to open a particular food place](#_pc15ozv6qazz) 8

[3.3 MACHINE LEARNING](#_ri4fe53m9wd4) 8

[**4. RESULTS**](#_m9ebfknbctva) **9**

[4.1 RESULT SET I](#_4ekoeefqu7jk) 9

[4.2 RESULT SET II](#_m4g3cam6la5f) 10

[4.3 RESULT SET III](#_rsl9q2o4wiry) 10

[**5. DISCUSSION**](#_detjg33lcg7o) **11**

[**6. CONCLUSION**](#_mgdgqrxg38gw) **11**

# 

# 

# 1. BUSINESS PROBLEM

## 1.1 INTRODUCTION

One of the most important factors contributing to the relative success of a Restaurant is the popularity of its cuisine among the occupants of the neighborhood. For instance, it is most suitable to open a Chinese restaurant in an area where people generally visit and like the already existing Chinese restaurants. This is the key idea behind the implementation of this project.

## 1.2 AIM

The main motive of this project is to help a Restaurant owner identify the most suitable site for a restaurant of their cuisine in a particular area, taken as California in this project. Hence, the target audience for this project consists of people looking forward to set up their restaurants in California, United States. The results of this project will reveal what are the most liked cuisines among people living in towns and cities of California. Through these results, a new Restaurant owner can identify which city or town will be most suitable for the type of restaurant they are willing to open in California.

## 1.3 TARGETED AUDIENCE

This project will interest the prospective investors willing to set up their restaurants in California. Having knowledge about the popular cuisines of an area beforehand, they can start their restaurant in an area where it is more likely to be visited and liked by the occupants of that area. Initial setup in an appropriate location will help them boost their businesses and earn greater profits. Undeniably then this project will help solve a baffling issue of site selection for a new business and hence will interest stakeholders.

# 2. DATA

To implement the above idea, it is important to have relevant data at hand.

## 2.1 DATA REQUIREMENT

The data required includes:

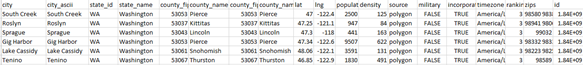
* List of cities and towns in California, US.
* Latitudes and longitudes of California and towns/cities within it.
* Other data related to population of a city/town and their cuisine interests.
* Details of Restaurants and their popularity in an area within these cities/towns.

## 2.2 DATA SOURCES

The above data is collected through two sources:

### 2.2.1 United States cities database

This dataset is taken from<https://simplemaps.com/data/us-cities>. The data in this database is built from the ground up using authoritative sources such as the U.S. Geological Survey and U.S. Census Bureau. The image below shows a sample preview of this database.



*Fig 2.1: Sample preview of the US cities dataset*

The fields of interest include:

* *state\_name* containing the name of the state or territory that contains the city/town. This will be filtered to ‘California’ in this project.
* *city* containing the name of the city/town. These will be considered individually and analyzed for their occupants’ cuisine interests.
* *lat* containing the latitude of the city/town.
* *lng* containing the longitude of the city/town.
* *incorporated* containing TRUE if the place is a city/town. FALSE if the place is just a commonly known name for a populated area. This will be filtered for all the TRUE values in our case.
* *population* containing an estimate of the city’s urban population. This will be used to quantify the interest against the number of customers in an area.
* *density* containing the estimated population per square kilometer. This will be used to divide the *population* and get an estimated area of the city(in km).
* *ranking* containing an integer from 1-5 that captures the importance of a city (1 is most important, 5 least important). The cities with upper rankings will be filtered for further analysis.

### 2.2.2 The Foursquare Places API

These API return responses from a global database of rich venue data and user content. This will be particularly used to get details about the restaurants and their popularity in an area. The above results can be acquired through following calls to these APIs:

* Call to search for venue near given location with the category Food

A typical response to this call will return all the restaurants (naturally belonging to food category) with the details about their other categories. This will give an idea about their cuisine.

* Call to search for users who liked a particular venue

This will respond with information about the users who have liked the above found venues. A greater number of likes will indicate more popularity of the venue in general.

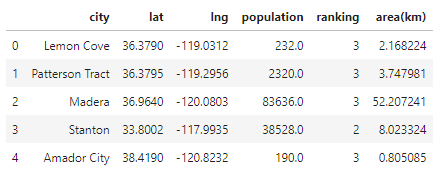
Therefore, the towns/cities in California with their coordinates provided by the first dataset will be searched for nearby restaurants and their popularity through Places API responses.

## 2.3 DATA CLEANING

This section introduces what all changes were made to data before analyzing it including its cleaning, preprocessing and formatting.

* Since the *us\_cities* dataset contains data of all the states of US, our first step should be to filter the *state\_name* field to California. This will ensure that we focus only on the data related to California which is our area of interest in this project.
* The next step is to remove all the fields like Timezone, zip, id etc. which are not of significance to our analysis.
* The next step is to check if the listed city/town are actual cities or just a name given to popular areas. Since we want to focus on cities and towns of California, we will ensure this by filtering the *incorporated* field to True. After the appropriate filter, this column shall also be removed from the dataset.
* The last step is to calculate the *area (in km)* of the city by dividing the *population* field by *density* field. This will be followed by deletion of *density* field from the dataset.

Figure 2.2 below contains the snippet of how our final dataset will look like.



*Fig 2.2: Snippet of final dataset after data cleaning*

# 3. METHODOLOGY

This section will include everything that was done from data wrangling to statistical testing and application of machine learning in order to analyse the above data.

## 3.1 DATA COLLECTION AND WRANGLING

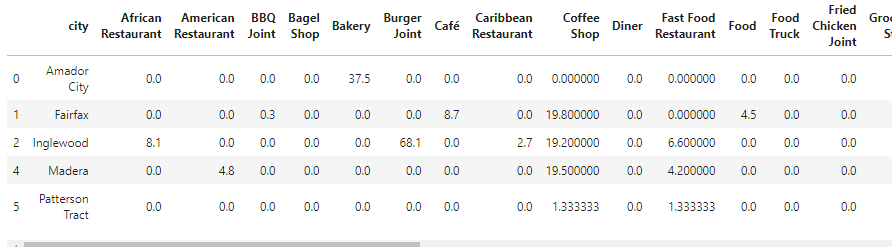
The data from *figure 2.2* was further used to do our exploratory analysis.

* For every city listed, a Foursquare API call was used to get response of all the venues belonging to food category within its area as given in the dataset.
* The categories of every venue obtained were appropriately put in another dataframe.
* These categories constitute the most important element of our dataset. These categories are further one-hot encoded to give a clearer picture of these venues in different cities.



*Fig 3.1: Snippet of dataset after one-hot encoding of category field*

* Now through another call to Foursquare API,the number of likes for each place are obtained and multiplied with the above values to take into account the likeability of these places in a city.
* The ranking of each city indicating its significance is also multiplied to above values to make these values more realistic and reliable.
* Next this data is grouped by cities using mean values of all categories in these cities.
* Finally, the dataset with 0 likeability or 0 ranking making all values 0 in a row are removed since such places should not be considered for setting up a new food venue.



*Fig 3.2: Snippet of dataset with final values used for further analysis*

The above dataset with 31 categories of food venues are analysed for each city in California.

## 3.2 DATA ANALYSIS

The dataset above is analysed in two ways to obtain the required results and give the answer to the question of best site for opening up a food venue in California.

### 3.2.1 **Analyzing the most preferred food places for all the cities**

This analysis will take into consideration each city and tell which are the most liked or preferred places in these cities. This will be useful in a scenario when a person, living in a California city, is reluctant to relocate. Within the same city, the analysis will give the best food places to a stakeholder willing to open up a new food business.

To analyze the same, the following steps were taken:

* A new dataframe containing 6 columns is created.
  + The first column will contain all the cities from above dataset.
  + The columns from 2nd to 6th will contain the most preferred venues with the value in 1st column as most preferred.
* The values of categories are sorted in descending order for each city.
* The top 5 results are carefully used as most preferred food locations of each city.

### 3.2.2 **Analyzing the best cities to open a particular food place**

This analysis will take into consideration every food category of venues present in cities of California. It will then show which cities should be preferred for opening up venues of each category. This will be useful in a scenario when a person, say is an expert chef of American food, and is willing to open up an American restaurant only. The analysis below will indicate him which are the best cities within California to open up an American Restaurant.

To achieve the same, the following steps were taken:

* A new dataframe is constructed with 5 columns where each column will indicate 5 most preferred cities for a food location with first column showing the 1st most preferred city and likewise.
* The index of this dataframe will be set as the category for which the other 5 columns will contain the values. The categories will comprise of all the categories one-hot encoded in the above dataframe.
* The values of most preferred cities for each food category are calculated by sorting the dataset of fig 2.2 in descending order for each category one-by-one. For every sorting, the top 5 values against that category are inserted into the dataframe.

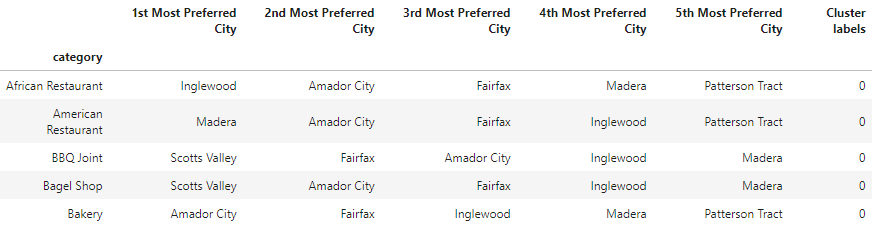
## 3.3 MACHINE LEARNING

The above analysis is further explored by applying Machine Learning algorithms. In this project, k-means clustering is used in order to cluster all the food categories together which are equally preferred in a city.

This is done through following steps:

* A new dataframe is created by transposing the dataset in fig 3.2
* K-means clustering is done on this dataframe for number of clusters, k value, as 5.
* The cluster labels obtained are then merged with the original dataset.

The figure below shows this merged data.



*Fig 3.3: Dataset merged with cluster labels*

Finally, the categories above are grouped by the cluster labels to give the categories which are collectively and similarly preferred by people.

# 4. RESULTS

The figures below will provide us with the results of our analysis described above.

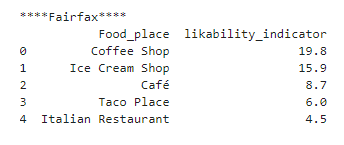
## 4.1 RESULT SET I

These results are obtained by analysis 1 where the most preferred venues for each city are obtained.



*Fig 4.1: Snippet of dataset to analyze the most preferred food places for all cities*

The above results are also visualized as below to includeother details related to likeability of each and every venue.

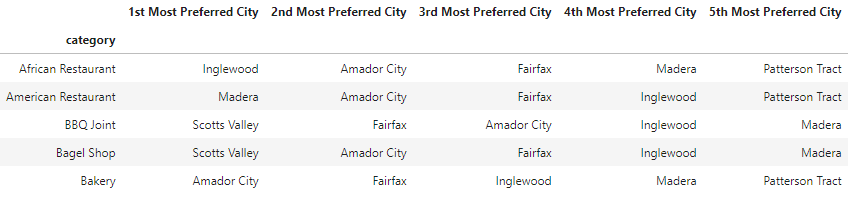


*Fig 4.2: Snippet showing likability indicator of every food place in a city(Fairfax)*

Hence, it can be estimated that a Coffee Shop in the city of Fairfax will be preferred by people over an Italian Restaurant in the same city.

## 4.2 RESULT SET II

These results are obtained by analysis II where each food venue is analysed for the top cities where they are most preferred.

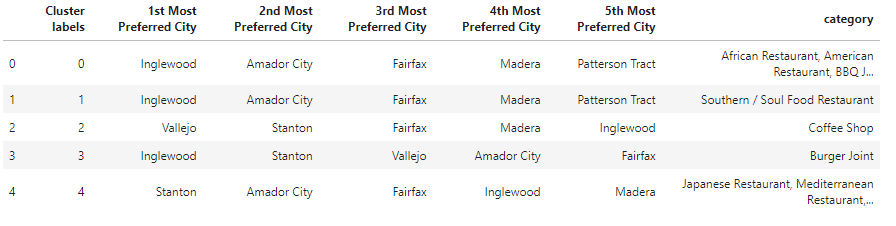


*Fig 4.3: Snippet showing results of analyzing best cities to open a particular food place.*

Hence it can be estimated than an African Restaurant will have its best business in the city of Inglewood of California.

## 4.3 RESULT SET III

These results are obtained after the application of k-means Clustering on our dataset.



*Fig 4.4: Snippet of dataset showing results after Clustering on the basis on categories*

Hence it can estimated that people who prefer African Restaurant in Inglewood, generally prefer to go to American Restaurants as well.

# 5. DISCUSSION

* The results from set I can be used to interpret that a person willing to open up a new food venue in Fairfax city of California only, should consider opening up a Coffee Shop which is most liked by people there. Otherwise he can choose to open up a restaurant having menus of most liked places there. This would mean opening up a restaurant in Fairfax having coffee, ice-cream, Tacos and Italian among other items on its menu.
* The results from set II indicate that the person willing to open up only an American Restaurant should prefer to open it up Madera city of California since it is most liked by people there. Also this city is comparatively significant and popular which will ensure higher footfall for any food places opened up here.
* The results from set III shows that all categories in a particular cluster are equally liked by people in the cities mentioned for that cluster. So these facts can be used to combine and create a menu for the new food place. For eg, a person opening up a Restaurant in Inglewood can keep its menu a mixture of African and American dishes since these are the most liked restaurants in this city of California.

# 6. CONCLUSION

This analysis will therefore be very helpful for any person willing to open up a food venue in California. The results above will help him in choosing the best city within California for their new food venue. Otherwise, they can also see it the other way round and get the best category of food venue for their city of interest. Lastly, they can also realize the people’s preference and create a menu for their new venture likewise. This analysis will help him get maximum profits resulting from increased sales due to its high demand and preference in the recommended city.