## Definition

### Project Overview

The current project is about how to determine the level of clients’ satisfaction and what parameters affects it to a greater extent. The restaurant business is taken as an example. The data for the project is taken from Kaggle [“Restaurant Data with Consumer Ratings”](https://www.kaggle.com/uciml/restaurant-data-with-consumer-ratings) (<https://www.kaggle.com/uciml/restaurant-data-with-consumer-ratings>).

The input data has details about 138 users which rated 130 restaurants. The details of restaurants are provided as well.

Several models have been built to predict what characteristics of restaurants are dominant in the assessment and in what category - nice, good or satisfied - a restaurant falls into.

It might help a restaurant’s owner to correct deficiencies and improve strengths of business. The restaurant business has been taken as it’s always on demand. In addition, this project can be used as basis for the analysis of other services with different variables.

### 

### Problem Statement

The purpose of the project is to build the model which will define whether a client likes a restaurant or not. It will be based on the restaurant’s characteristics. The first step is to clean and analyze the given data. Some details are redundant and cannot help to build a model. The lack of details for some restaurants will be taken into consideration as well.

During the first step, I explored the data and tried to find some relations between restaurants and users data.

This step helped me to understand how the data should be encoded and what algorithms are to be used to make a prediction when dealing with categorical input and output. I used the mix of different algorithms to get the result. For more accurate predictions, I divided restaurant into few clusters and selected features when runninga model. Clustering and features selection were conducted independently.

The algorithms for clustering and features selection:

* K-means
* Backward elimination
* Chi2

The algorithms for a prediction:

* Random Forest Classifier
* Multiclass classifier from Linear Learner from Amazon SageMarker
* Balance Classifier

The data was splitted into the training and test sets to see how good or bad the model is.

## Analysis

### Data Exploration and Visualization

Those two segments were combined as closely intertwined with each other. I took few examples of graphs from Jupyter Notebook to display the data.

The data for this project was taken from ‘Restaurant Data with Consumer Ratings’. There are 9 .csv files which can be divided into three groups: 1) restaurants related data ; 2) users related data; 3) ratings given by users.

Files related to restaurants:

1. geoplaces.csv stores the general data about restaurants. There are 20 self explanatory columns in the file.
   1. placeID;
   2. latitude;
   3. longitude;
   4. the\_geom\_meter;
   5. name;address;
   6. city;
   7. state;
   8. country;
   9. fax;
   10. zip;
   11. alcohol;
   12. smoking\_area;
   13. dress\_code;
   14. accessibility;
   15. price;
   16. url;
   17. Rambience;
   18. franchise;
   19. area;
   20. other\_services
2. chefmozaccepts.csv stores payment options of restaurants.
3. chefmozcuisine.csv stores restaurants’ cuisines.
4. chefmozhours.csv stores restaurants’ working hours.
5. chefmozparking.csv stores parking options of restaurants.

Files related to users:

1. userprofile.csv stores the general data about users. There are 19 self explanatory columns in the file.
   1. userID;
   2. latitude;
   3. longitude;
   4. smoker;
   5. drink\_level;
   6. dress\_preference;
   7. ambience;
   8. transport;
   9. marital\_status;
   10. hijos (meaning children from Spanish);
   11. birth\_year;
   12. interest;
   13. personality;
   14. religion;
   15. activity;
   16. color;
   17. weight;
   18. budget;
   19. height
2. usercuisine.csv stores restaurants’ cuisines which users evaluated.
3. userpayment.csv stores payment options of users.

All files related to restaurants and users have the common value in datasets: placeID for restaurants and userID for users. Those values will be used as a key value when merging the data. The last file rating\_final.csv stores placeID, userID, and three types of ratings given by users: rating, food\_rating, service\_rating.

All files are saved in the folder raw\_data and then will be used when exploring and visualising the data.

The first step is to upload the data into Jupyter Notebook and then save all files in the correct format as the initial files have as mixed types of delimiters.

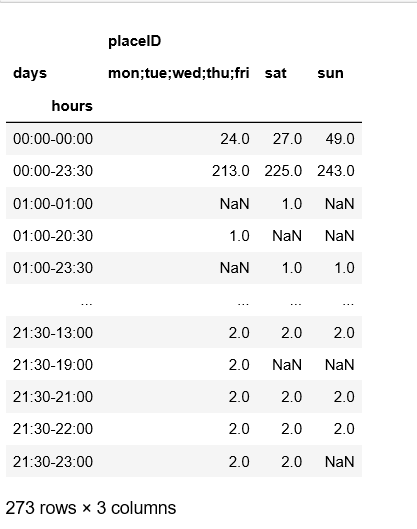
Before analyzing the data, all values in the datasets were lowercased in case same values are written in different ways, which was the case for restaurants’ payment methods. Then, all columns’ values were checked and some values were rearranged to decrease the number of options in one column.

Some examples of values’ updates.

1. Different ways of writing. Values ‘slp’, ‘san luis potosi’, ‘s.l.p.’, ‘san luis potos’ from the column ‘state’ all refer to one state ‘San Luis Potose’. They were replaced by 'san luis potosi'.
2. Similar meanings. There are 5 different options in smoking\_area: ‘none’, ‘not permitted’, ‘only at bar’, ‘section’, ‘permitted’. First two values were combined into ‘not permitted’; the remaining three values were combined into ‘permitted’.

All datasets were checked and the required changes were done for all values. In case some options were reassigned or merged into one group, it was done basing on a subjective opinion of a developer and her logical intuition.

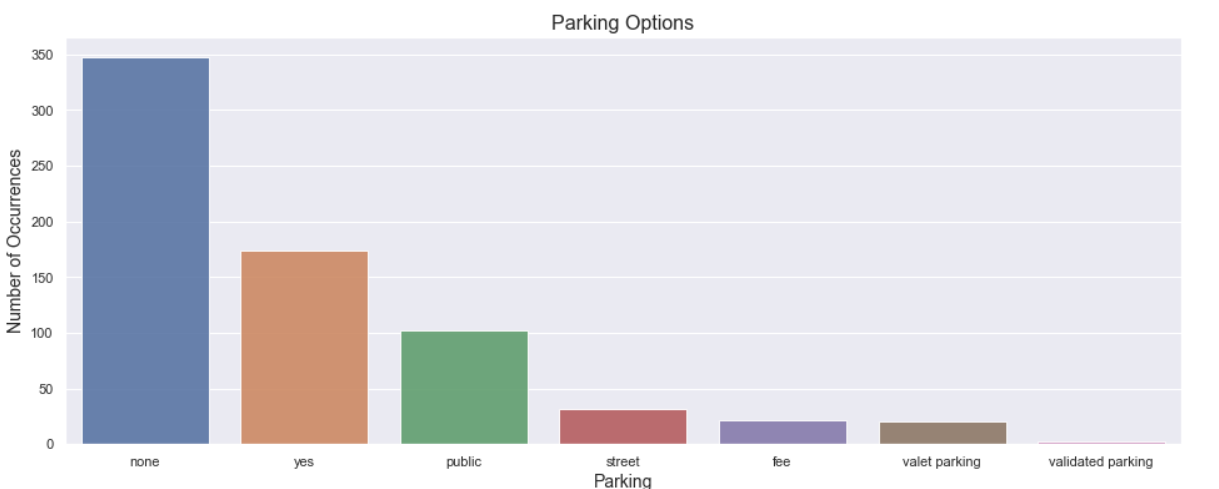
During the analysis, columns which values are irrelevant, e.g. zip/url which are mostly empty, or difficult to evaluate, e.g. latitude/longitude, were dropped. Moreover, the dataset of restaurants working hours was not considered during the further analysis. There are 273 different shifts over three types of working days. Many working hours are intersected with each other. As there are many different options of working hours of restaurants, it's been decided to disregard this dataset. If a client evaluated a restaurant, it assumes he attended it and the working shift is not an issue for the client. In addition,the data does not state what hours are more profitable for a restaurant and it's not clear how it can be evaluated.



Basing on the restaurants and users data, it’s displayed how restaurants are divided into different groups per each category and what group of users is dominant in users’ categories.

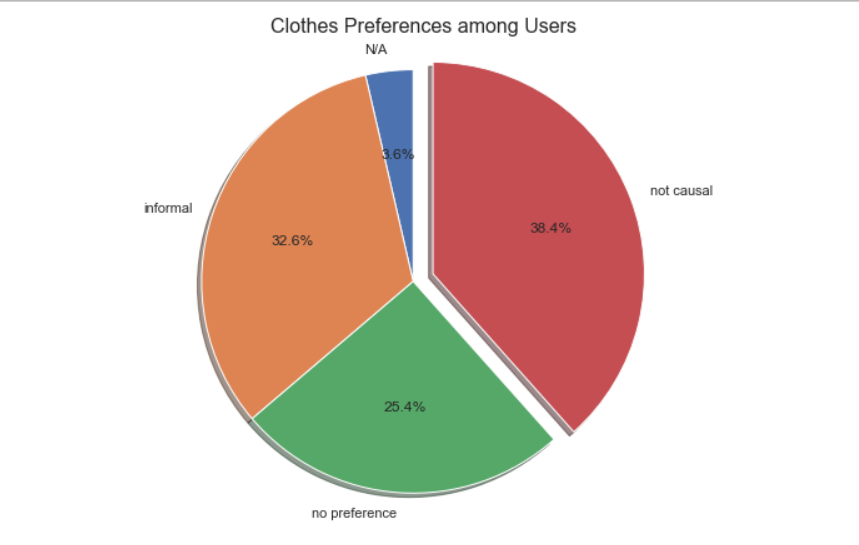
The restaurants’ graphs display how many restaurants occur in each group. It can be used to check restaurants with the most frequent features and align your own strategy when running the business.

Restaurants graph’s example: Parking Options of 130 restaurants.

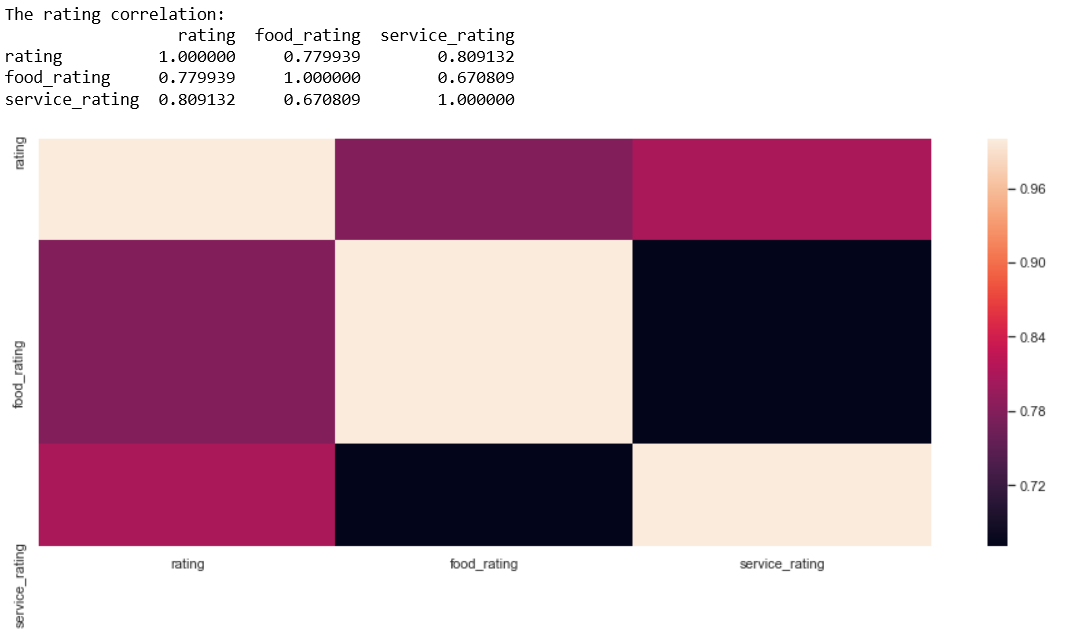


The users’ graphs display users preferences and how many users occur in a group. A restaurant’s owner might change some restaurant’s features if the majority of clients fall into a specific category and the owner wants to target this group of clients.

Users graph’s example: Clothes Preferences among 138 clients.



During the data analysis, the correlation between three types of ratings from the rating\_final.csv was checked.



Lastly, the average rating per four different categories (franchise, area, rambience, other\_services) from r\_general as those categories were selected as features which affect most the rating of restaurants. Its selection will be explained in details in the section 'methodology'.

Example: Rating for Other Services Types Options

