Coursera Capstone Project – SF restaurants

1. Introduction/Business Problem

A client wants to open a restaurant within the area of San Francisco.
The client is new to the city and is not bound to a specific kind of cuisine.
He needs to know:

1. what kind of cuisine would attract a lot of customers
2. which area is suitable for that kind of cuisine
3. a possible additional competitor analysis

Data science and Python will be used to weigh between the location and cuisine of restaurants, using data gathered from the website foursquare.com.

2. Data

Three kinds of data are necessary to answer the business problem:

1. The location of a restaurant (geo location, district)
2. The type of restaurant (e.g. Chinese or Mexican)
3. The business metrics of the restaurant (likes)

In the Case of the Foursquare API, the relevant data will be extracted from the *'search'* and *'likes'* endpoint:

1. The geo data from the previous Coursera assignments will mark the city border.
   Within the border arbitrary exploration points will be defined by Longitude and latitude.
2. The *'categoryId'* tag for food (*'4d4b7105d754a06374d81259'*) will be used to extract the id, category, location and name of a restaurant around an exploration point.
3. The *'categoryId'* from 2. will be used for an additional API query.
   The result contains the number of likes per restaurant and a list of a few users who liked the restaurant.

The number of premium calls for the *'details'* endpoints is very limited.
The rating of the restaurant, which is part of the response of calling the *'details'* endpoint would be a better option, but again, premium calls are limited.

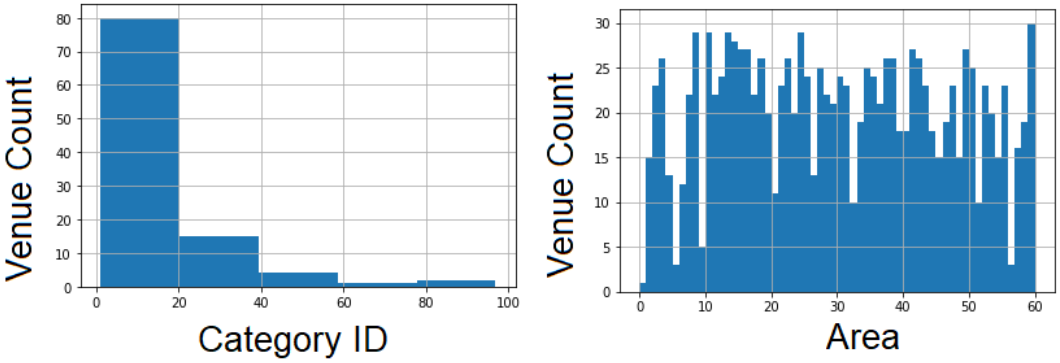
Therefore, I will use the like count as a metric for the business performance of a restaurant, which allows to collect more than 50 data points for the large area of SF.
I will assume, that many likes correspond to many and happy customers and therefore large revenue for the restaurant's owners.

3. Methodology

- The district of SF was split into multiple areas, which are labeled by integer numbers.
- The API's *'search'* endpoint is used to retrieve data for a set number of restaurants within a given area. The restaurants are selected arbitrarily, in contrast to the recommendation based and therefore preselected results of the API's *'explore'* endpoint.

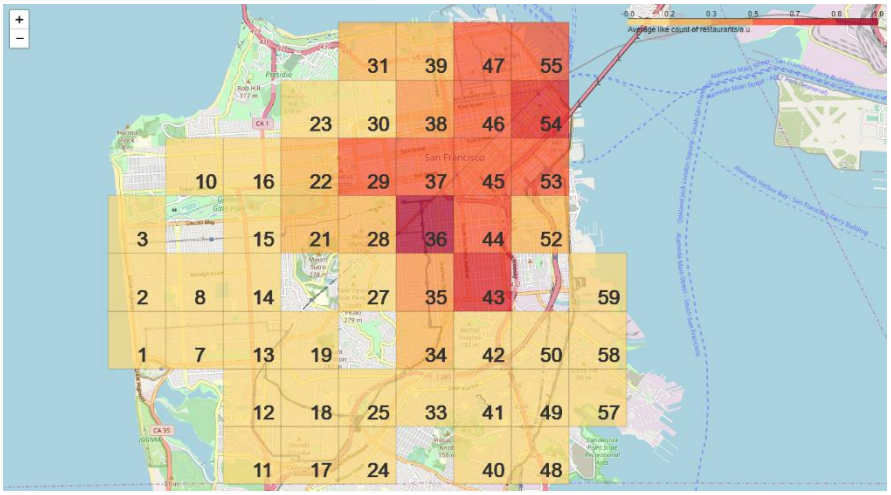
- The number of areas is related to the spatial resolution of the search results and is chosen freely to be at 60.
- The restaurant labels from the API's *'categoryId'* data contained labels, that were not useful for the customer, such as *'Office'* or *'Toys & Games'*. Search results containing those labels were dropped.

- After dropping rows with useless labels, the remaining venues were counted by area. The formulated statistic contained the mean of the counted restaurants per area (μ) and the corresponding standard deviation (σ).
- Areas that had less than $(\mu - \sigma)$ venue counts were dropped from the data frame and therefore the map. Area with too few restaurants were deemed not representative of the diversity of restaurants within a given area. Additionally, some areas contain landmarks such as bigger parks, which naturally reduce the available set of data.



- A search using the API's *'likes'* endpoint was performed for all venues in the remaining 50 areas.

- The average like count of all restaurants within each area was plotted using a relative color scale. Each box represents an area, defined by the integer (called *'area'* in the data set*),* which are both shown in the map below:



- The map shows a clear segmentation into areas with a low like average (bright yellow) and high like average.
- Clear segmentation allows to sort areas into
    - **areas with few likes:** bright yellow, low business, labeled as *'0'* or

    - **areas with high like averages:** darker yellow shades into the red, a lot of business, labeled as *'1'*
- Machine learning (using logistic regression) was used to predict the binary (*'0'/'1'*) label of restaurants as a function of their geo location and their type
- Statistical testing was performed and resulted in a Jaccard similarity score of 0.78 and the below shown classification report:

	precision	recall	f1-score	support
0.0	0.82	0.93	0.87	183
1.0	0.38	0.17	0.24	46
micro avg	0.78	0.78	0.78	229
macro avg	0.60	0.55	0.55	229
weighted avg	0.73	0.78	0.74	229

4. Results

- A comparison for the most and least liked cuisines for each area was gathered from the data:

		Least liked
Most liked	**area**	
African	50	Bakery,Cajun / Creole,Desserts,Food,Restaurant...
American	3	BBQ,Bakery,Food,Food Truck,Japanese,Sandwiches
Asian	57	Bakery,Deli / Bodega,Food,Gourmet,Restaurant,S...
Bakery	36	Salad
	2	American,BBQ,Diner,Food,Food Court,Food Truck,...
	30	BBQ,Restaurant,Snacks
Breakfast	48	Bakery,Chinese,Food,Restaurant,Seafood,Tacos,V...
	52	Burgers

- The customer can't freely choose a location, he is bound to the restaurant real estate market. The developed logistic regression model can be used to predict the performance at a specific geo location/ real address, not just within the borders of the defined areas.
- It is also possible to predict the likes within areas that have been dropped from the dataset, due to invalid or too few data points.
- The results of the statistical testing of the model seem to be good enough to make rough estimations for the customer

5. Discussion

- The main task of this repost was to choose parameters for the opening of a new restaurant in SF. Depending on the customer's focus It would be possible to pursue multiple strategies. Two of them would be:

    a. Open a restaurant in an area with less business (bright yellow), focus on quality and benefit from less competition
    b. Open a restaurant in an area with a lot of business and choose a cuisine that has many likes in that area
- Using the generated tables in the notebook, we can state which kind of restaurant is most suited for a given area and business strategy.

- Future evaluation using the customer data could be done to get a user profile of typical customers for a given region or restaurant type, since this data was also extracted during the '*likes'* endpoint call.

- As data science is a highly iterative process, it is likely that the machine learning and predicting approaches could be refined with a smaller set of preselected restaurant types and more data points.
- For the future is would be interesting to compare the found results with additional data sets. An interesting question would be for example, if areas that have many likes for bakeries are also areas with many offices, where workers might grab their breakfast near their offices.

6. Conclusion

- For this assignment location data from the Foursquare API was gathered and was analyzed using multiple data science related packages in Python
- The number of data points for this analysis was limited. Despite that, it possible to get a good overview of the business metrics for specific restaurant types in the area of San Francisco
- Machine learning was a helpful asset to transfer the insight of the data into a model. This model can predict the business performance of a restaurant based on its location and cuisine.