Case of Study:

Location, Location and Location: A Machine Learning approach to choose location for a new business in NYC.

Author: **P.A. Spring**

Applied Data Science Capstone

by IBM

I. Introduction.

Without doubt choosing the right location for a new business is one of the most crucial steps in the life of a project. Determine which factor must be taken in consideration is core to solve this problem. Our case of study is centered in one of the business with the highest rate of failures within a year of operation, the small and mid-restaurants, this enterprise have percentages of failure of around the 26% within a year. Therefore, for restaurants owners and entrepreneurs is decisive to identify how different variables are related with the success or failure of his business.

The goal of this work is to develop a machine learning model with the objective of predict and describe how the right location influence the success or failure of a restaurant. Specifically, we are going to analyze the restaurants distribution around the neighborhoods of New York City using foursquare, google and opencity data. The target audience for this study will be restaurants owners, entrepreneurs and researchers.

II. Data.

With the objective of predict how the location influence in the success or failure of a small/mid restaurant in the neighborhoods of New York City we are going to collect data from different sources. In the following lines is a description of all the necessary data for our project.

Restaurant descriptions: it is necessary to create a database with restaurants around the city of NY, with information about geo-location, name, address, type of restaurant, average rating, customers review and level of price. For this we will need to get the data from the foursquare api and complementary information from the google maps api.

Geo-location data from the neighborhoods of the city: Every neighborhood of the city has relevant and useful information for our case of study. For this we need the planning data of the city, in a geojson file with the geometry of every neighborhoods, with the specific name, codes and boroughs. This data will be extracted from the NYC data site.

Information about point of interest around the selected restaurants: In the NYC data site we can find different kind of locations of interest around the city, how college, schools, touristic points and metro stations. This kind of places and them closeness with the restaurants could be in some way related with the success or failure of them, because of that we will try to include them in the analysis.

Description of the groups of people living in the neighborhoods of the city: Demographic data is another important source of information for this study, some researches point at some specific demography characteristics around the success of a small restaurant, because of this we will use the demographic information of the neighborhoods of the city as: average age, income, education level. All this information is available in the census2020 website.

Finally, the next table shows how the data was acquired and transformed in the data acquisition step of this work.

Step	Objective	Input	Output
2.1	Dataset creation with the neiborhoods of the city and geodata	Neighborhood Tabulation Areas (NTA).geojson	shape_neighborhoods_df.csv
2.2	Adding a center point to the neighborhood data	shape_neighborhoods_df.csv	shape_neighborhoods_df.csv
2.3	Creation of a list of point inside every neighborhood to improve the venues searching	shape_neighborhoods_df.csv	shapes_k3.csv, shapes_k5.csv, shapes_k10.csv
2.4	Search of restaurants and food places around the points created in the step 2.3	shapes_k3.csv	foursquare_venues.csv
2.5	Classification of the foursquare venues in his correspondent neighborhood	foursquare_venues.csv, nta.geojson	foursquare_with_neighborhood.csv
2.6	Selection of type restaurants for the study	denom.csv, foursquare_with_neighborhood.csv	small_restaurants_foursquare.csv
2.7	Filling restaurants with premium information from foursquare API	small_restaurants_foursquare.csv	restaurants_only.csv
2.8	Addition of demographic information to restaurants depending of the neighborhood to which belongs	restaurants_only.csv, nyc_demo_df.csv	Model_dataset.csv

The following image shows how the dataset was transformed in every step described in the last table.

	Neighborhood	Borough	BoroCode	NTACode	Polygon_Coordinates
0	Borough Park	Brooklyn	3	BK88	[[[[-73.97604935657381, 40.631275905646774], [...
1	Murray Hill	Queens	4	QN51	[[[[-73.80379022888098, 40.775610112295055], [...
2	East Elmhurst	Queens	4	QN27	[[[[-73.8610972440186, 40.7636644770877], [-73...
3	Hollis	Queens	4	QN07	[[[[-73.75725671509139, 40.71813860166257], [-...
4	Manhattanville	Manhattan	1	MN06	[[[[-73.94607828674226, 40.82126321606191], [-...

Figure 1: Creation of a dataframe with the name and geometry of the neighborhoods of NYC.

	Neighborhood	Borough	BoroCode	NTACode	Polygon_Coordinates	Centroid
0	Borough Park	Brooklyn	3	BK88	((-73.97604935657381, 40.631275905646774), (-7...	(-73.98866123064063, 40.630949655424594)
1	Murray Hill	Queens	4	QN51	((-73.80379022888098, 40.775610112295055), (-7...	(-73.80954590118573, 40.768351587973775)
2	East Elmhurst	Queens	4	QN27	((-73.8610972440186, 40.7636644770877), (-73.8...	(-73.86839559825498, 40.7633522011835)
3	Hollis	Queens	4	QN07	((-73.75725671509139, 40.71813860166257), (-73...	(-73.76113705097167, 40.71063933392971)
4	Manhattanville	Manhattan	1	MN06	((-73.94607828674226, 40.82126321606191), (-73...	(-73.95378199219353, 40.817975566129746)

Figure 2: Addition to the neighborhood of the coordinates of their centers.

	Neighborhood	Borough	BoroCode	NTACode	Polygon_Coordinates	Centroid	Points
0	Borough Park	Brooklyn	3	BK88	((-73.9760507905698, 40.6312841471042), (-73.9...	(-73.98866266865282, 40.630957896376444)	[POINT (-73.98464694508898 40.6251869300012), ...
1	Murray Hill	Queens	4	QN51	((-73.8037916164002, 40.7756183880718), (-73.8...	(-73.80954729011228, 40.76835986212814)	[POINT (-73.81153220836251 40.77802647165134),...
2	East Elmhurst	Queens	4	QN27	((-73.8610986495631, 40.7636727481715), (-73.8...	(-73.86839700569737, 40.76336047230126)	[POINT (-73.87526718508542 40.7670774233572 3),...
3	Hollis	Queens	4	QN07	((-73.7572580842358, 40.7181468677945), (-73.7...	(-73.76113842090989, 40.71064759845558)	[POINT (-73.76944885731125 40.7137679873343), ...
4	Homecrest	Brooklyn	3	BK25	((-73.9585942121111, 40.6104112689022), (-73.9...	(-73.96433509678462, 40.59996252149724)	[POINT (-73.96158399922773 40.59534622246872),...

Figure 3: Creation of random points inside the geometry of the neighborhoods.

	id	name	categories	referralId	hasPerk	location.address	location.lat	location.lng
0	4f324d4719836c91c7ca4a1e	Ambrosia Italian and Albanian Restaurant	[{'id': '4d4b7105d754a06374d81259', 'name': 'F...	v-1592953214	False	90 Church Ave	40.642365	-73.979874
1	4f325b4819836c91c7cfc0a7	Red's Italian Cuisine	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	v-1592953214	False	3714 13th Ave	40.641598	-73.985633
0	585d400ecf4451174c84de78	Brooklyn Italians Soccer Club	[{'id': '52e81612bcbc57f1066b7a2e', 'name': 'S...	v-1592953214	False	5725 18th Ave	40.623125	-73.985773
0	4f325b4819836c91c7cfc0a7	Red's Italian Cuisine	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	v-1592953215	False	3714 13th Ave	40.641598	-73.985633
0	585d400ecf4451174c84de78	Brooklyn Italians Soccer Club	[{'id': '52e81612bcbc57f1066b7a2e', 'name': 'S...	v-1592953215	False	5725 18th Ave	40.623125	-73.985773

Figure 4: Restaurants found around the city using the foursquare app.

	Likes	Price	Pop_1E	MdAgeE	MnHHIncE	EA_LTHSGrE	EA_BchDHE	NtvE	Fb1E	categories_cat	GeoID_cat
0	0.0	2.0	35510	34.3	76511.0	4140	9845	20759	14751	3	18
1	4.0	2.0	102494	24.7	64990.0	14159	9810	77270	25224	8	38
2	5.0	1.0	35510	34.3	76511.0	4140	9845	20759	14751	9	18
3	0.0	2.0	90847	38.3	73256.0	18526	18582	40121	50726	6	7
4	7.0	1.0	90847	38.3	73256.0	18526	18582	40121	50726	3	7

Figure 5: Dataset filtered and prepared to be used in the different machine learning algorithms.

Where:

- Categories represent the type of restaurant.
- Geoid represent the neighborhood of the restaurant.
- Ntv represent the estimation of natives living in that neighborhood.
- Fb represent the estimation of foreign people living in that neighborhood.

- Likes represents the popularity of the restaurant.
- Price represent the price range of the restaurant, number between 1-4.
- Pop_1E represent the estimated population by the restaurant.
- MdAgeE represent the average age of the population round the restaurant.
- MnHH represent the mean income by family.

III. Methodology.

Our goal in this project is to find some relationships between the location of a restaurant and their success. With this in mind, we proceeded to analyze complementary data from the neighborhood where every restaurant works. This section of our study point to find how the price and the popularity of this kind of business is influenced by location and demography. With these two things as goals we are going to train a series of machine learning algorithms that answer our questions.

3.1 Restaurants distribution around the city.

Our dataset was constructed with the goal of recollect the most amount of detailed data related with small restaurants in the neighborhoods of NYC. Using the foursquare API we could retrieved information about 15000 business, but almost the 90% of them without all the required information. Finally, just 1500 observations were suitable for us. In the next image is shown how the restaurants are scattered around the of the city.

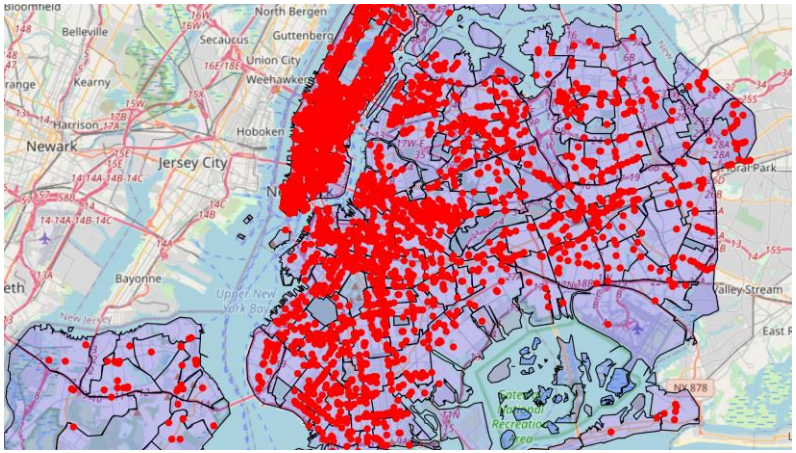

Figure 6: Restaurants around NYC.

3.2 Statistical analysis and initial visualization.

The main problem after the data selection was the low number of restaurants with price and rating in our dataset, just the 10% of all the observations recollected had have enough information to our purpose. This problem was related with the data available in foursquare and the lack of actualized information in their database, also because of the nature of our objectives we needed to make premium calls to the API, which make the process slow.

3.2.1 Price and popularity analysis.

Some studies showed that the location of a restaurant is a capstone for their future success, but how the location influence the popularity of a restaurant and how an owner project the revenues related with the location of the restaurant is interesting to insight.

The next two table shows the correlation between the features of our dataset. This analysis helped us to choose the principal variables for our models. In the table 1 we could see that the price is mostly correlated with demography of the neighborhood, being the neighborhoods with professionals highly educated (EA_BChDHE) and with great income (MnHHIncE) the areas with most restaurants, in this case Manhattan.

	Likes	Price	Pop_1E	MdAgeE	MnHHIncE	EA_LTHSGrE	EA_BchDHE	NtvE	Fb1E
Likes	1.000000	0.335932	0.108818	0.058390	0.581145	-0.176123	0.524458	0.224648	-0.091068
Price	0.335932	1.000000	-0.052800	0.187191	0.400510	-0.250910	0.293382	0.009500	-0.119097
Pop_1E	0.108818	-0.052800	1.000000	-0.129608	0.040973	0.655344	0.614624	0.890954	0.780067
MdAgeE	0.058390	0.187191	-0.129608	1.000000	0.352396	-0.315144	0.170757	-0.170653	-0.025073
MnHHIncE	0.581145	0.400510	0.040973	0.352396	1.000000	-0.490332	0.676175	0.221537	-0.222994
EA_LTHSGrE	-0.176123	-0.250910	0.655344	-0.315144	-0.490332	1.000000	-0.045015	0.398343	0.766866
EA_BchDHE	0.524458	0.293382	0.614624	0.170757	0.676175	-0.045015	1.000000	0.709369	0.256548
NtvE	0.224648	0.009500	0.890954	-0.170653	0.221537	0.398343	0.709369	1.000000	0.410879
Fb1E	-0.091068	-0.119097	0.780067	-0.025073	-0.222994	0.766866	0.256548	0.410879	1.000000

Table 1: Correlation table for the price analysis of our dataset.

The second correlation table was created with the mean values of the selected features in our dataset, this because the scope of the model is to predict the popularity of a restaurant related with their location, and this is a continues variable, different was the construction of the table 1, because the scope of that data is to predict a price range (1,2,3,4) this means a categorical variable.

As we could noticed the popularity of a restaurant, represented by likes in the social media, is mostly related with the income, education and the people who belong to that area (NtvE).

	GeoID_cat	Likes	Price	Pop_1E	MdAgeE	MnHHIncE	EA_LTHSGrE	EA_BchDHE	NtvE	Fb1E	Amount
GeoID_cat	1.000000	-0.072898	0.095732	-0.232885	0.306618	0.050592	-0.194445	-0.104775	-0.269468	-0.092411	-0.016414
Likes	-0.072898	1.000000	0.255390	0.125136	0.030850	0.529679	-0.120227	0.493844	0.231998	-0.070659	0.623946
Price	0.095732	0.255390	1.000000	-0.096051	0.087231	0.311078	-0.215585	0.204590	-0.034513	-0.143833	0.273979
Pop_1E	-0.232885	0.125136	-0.096051	1.000000	-0.142292	0.031156	0.658009	0.613387	0.886622	0.769961	0.301889
MdAgeE	0.306618	0.030850	0.087231	-0.142292	1.000000	0.319549	-0.299020	0.147147	-0.187546	-0.024865	0.046858
MnHHIncE	0.050592	0.529679	0.311078	0.031156	0.319549	1.000000	-0.489086	0.670208	0.217685	-0.238233	0.624473
EA_LTHSGrE	-0.194445	-0.120227	-0.215585	0.658009	-0.299020	-0.489086	1.000000	-0.047171	0.390531	0.772743	-0.045616
EA_BchDHE	-0.104775	0.493844	0.204590	0.613387	0.147147	0.670208	-0.047171	1.000000	0.708599	0.244974	0.670355
NtvE	-0.269468	0.231998	-0.034513	0.886622	-0.187546	0.217685	0.390531	0.708599	1.000000	0.387549	0.338124
Fb1E	-0.092411	-0.070659	-0.143833	0.769961	-0.024865	-0.238233	0.772743	0.244974	0.387549	1.000000	0.135229
Amount	-0.016414	0.623946	0.273979	0.301889	0.046858	0.624473	-0.045616	0.670355	0.338124	0.135229	1.000000

Table 2: Correlation table from the popularity dataset.

The next figures (7,8,9) represent a visualization of the correlation between the variables of the dataset with our scopes. The box plot shows some week dependency between price ranges of the restaurants in our dataset and our independent variables. In the figure 8 is clear that the exist a distribution of the variables around the neighborhoods of the city, all the graphs shown a distinctive peak around the neighborhood 100 (Lenox Hill Manhattan).

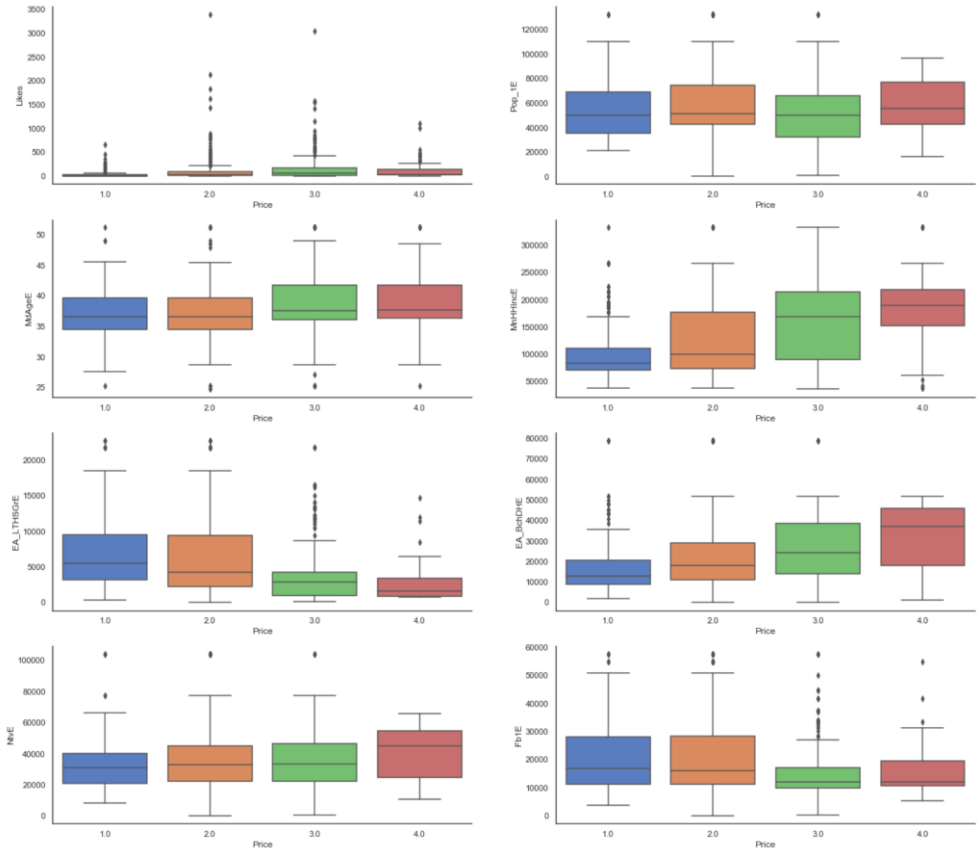

Figure 7: Box plot with the distribution of the selected features from the dataset and the price range.

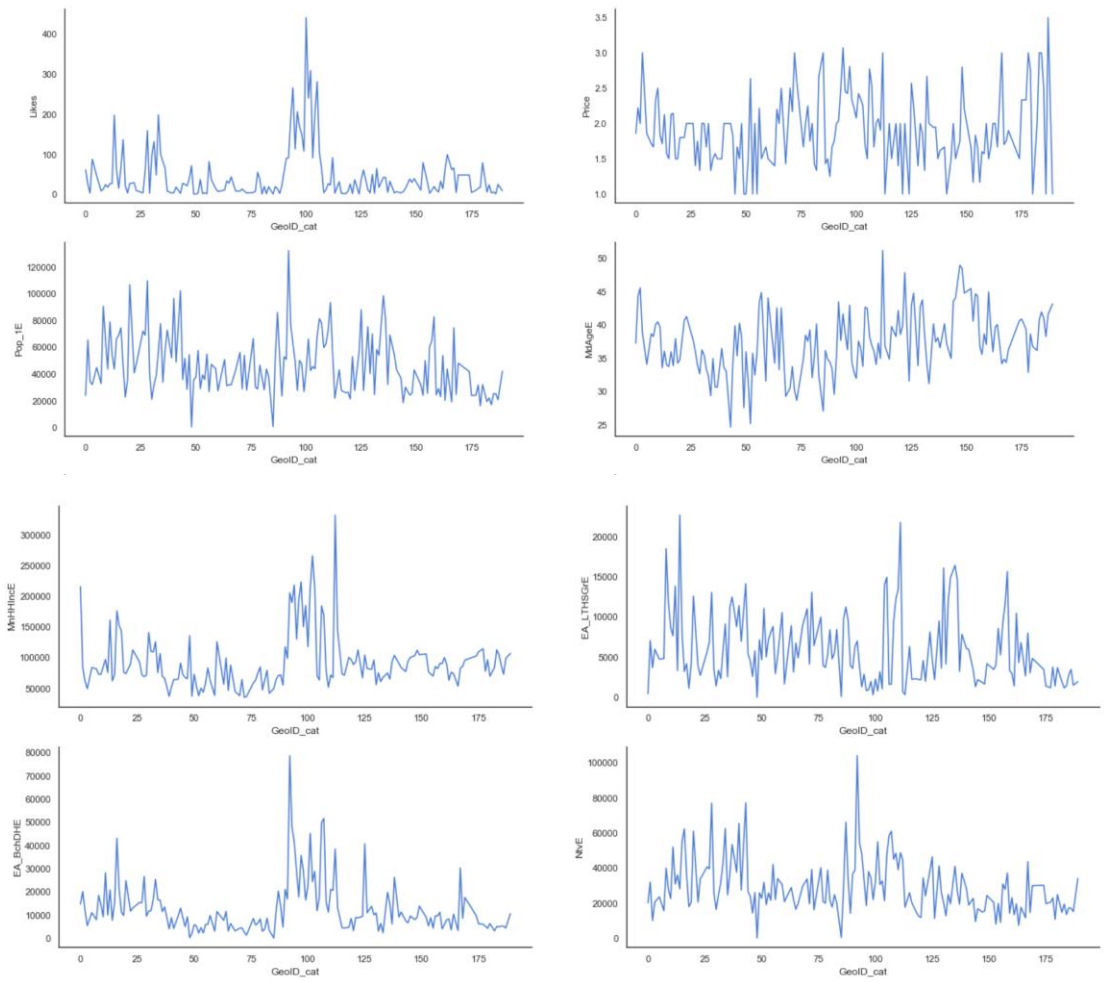

Figure 8: Restaurants distribution around the neighborhoods of the city.

In the figure 9 helped us to probe visually the information obtained in the correlation table for the popularity.

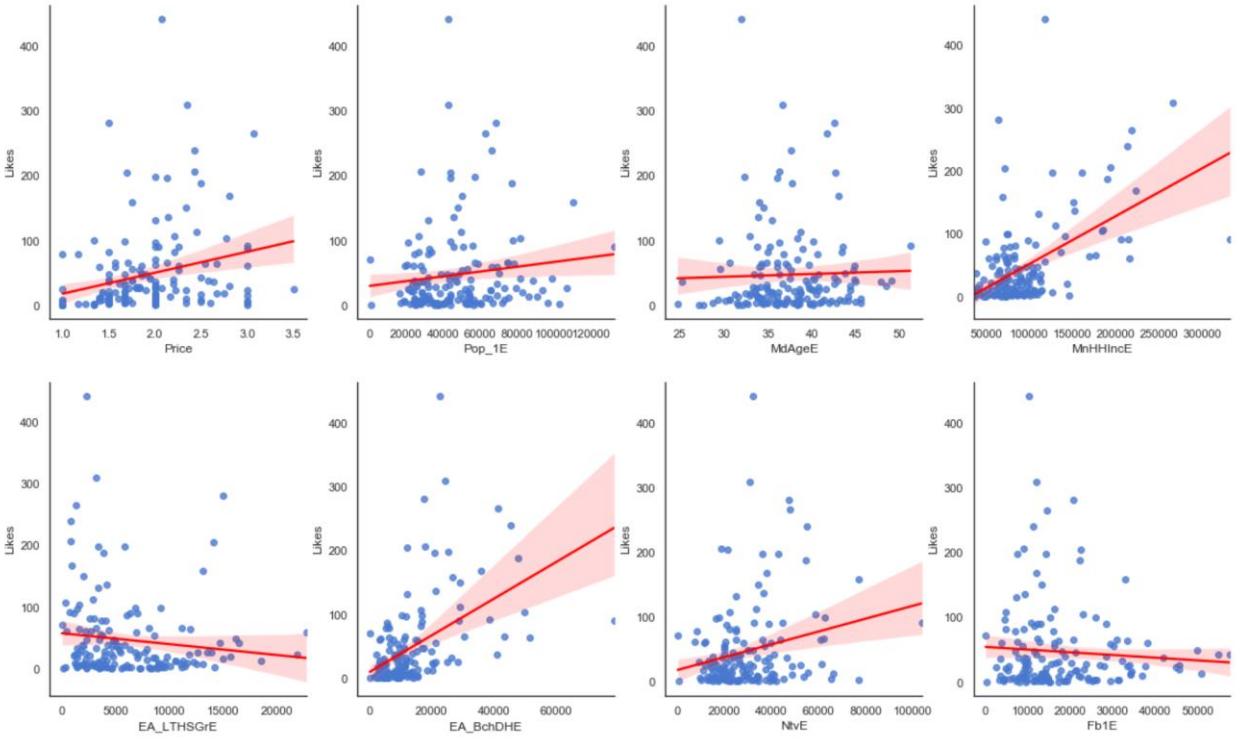

Figure 9: Correlation plots from the price dataset.

3.3 Principal Component Analysis (PCA)

The model process is divided in two steps, the first is model training between our scope (Price, Popularity) and every feature separately, the second step was grouped the most correlated features using the Principal Component Analysis (PCA).

The next image shows one example of the PCA applied to our dataset, were all the independent variables were reduced to two components.

	principal_component_1	principal_component_2
0	-1.710126	-0.027783
1	1.660479	-2.568295
2	-1.703244	-0.130704
3	1.611356	-3.288608
4	1.610354	-3.227835
...	...	...
1565	-0.247799	-1.424713
1566	-2.244598	0.106736
1567	0.827701	-0.755889
1568	0.291163	0.924632
1569	1.049629	-2.771063

Figure 10: PCA applied to the dataset.

For our training process, we decided to filter the dataset with the three features with the greatest correlation, and then applied the PCA.

3.3 Machine learning model selection.

In this step we proceed to probe different machine learning algorithms with our data, in the next table we can see all the models tested.

Objective	Machine Learning Algorithms
Range of price estimation	Logistic Regression, KNN, Decision Tree, Support Vector Machine.
Popularity estimation	Linear Regression, Polynomic Regression, Lasso Regression.

Table 3: Testing Models.

The most suitable algorithm was selected in base of the Jaccard and F1 Scores. It is important to note that our dataset with 1500 observation were divided and normalized before the training step.

III. Results section.

In this section the accuracy of the machine learning models is shown.

3.1 Price prediction model.

Model	Parameter	F1	JS
LR	Solver: newton-cg , C: 0.1	0.315423	0.479675
LR PCA	Solver: newton-cg , C:0.100	0.306591	0.47561
KNN	n:5.0	0.464388	0.475610
KNN PCA	n: 5.0	0.415552	0.434959
SVM	Variable: Likes , Kernel: sigmoid	0.366272	0.402439
SVM PCA	Kernel: sigmoid	0.366272	0.402439
D-Tree	Variable Likes , Depth: 7.0	0.408118	0.459350
D-Tree PCA	Depth: 7	0.408118	0.459350

Table 4: Table of model evaluation from the price dataset.

For the price prediction algorithm, the most accurate model was the K- Nearest Neighborhood with n=5, with an F1 Score of 0.46 and a Jaccard Similarity Score of 0.47. This result , should be improvable with more observations, this because our final dataset was reduced to 1500 observation, this is just almost 10 restaurants in every neighborhood, but it was impossible to get more information with the limitations of the foursquare API.

3.2 Popularity prediction model.

The next table shows the results on the modeling for the popularity of a restaurant in relation with its location. Again, the accuracy of this models is not remarkable, part of the problems continues being the lack of information and the few amount of restaurants with price and popularity in the foursquare API.

From the table we can choose the polynomial regression of second degree how or better model, with a R2 Score of 0.38.

Model	Parameter	MSE	R2-Score
Linear Regression	Variable: MnHHIncE	1467.734769	0.370773
Linear Regression PCA	PCA:1	1483.754192	0.363905
Lasso Regression	Variable: MnHHIncE	1486.241652	0.362839
Lasso Regression PCA	PCA: 1	1483.754192	0.363905
Polynomial Regression	Variable: GeoID_cat	3224.625329	0.382418
Polynomial Regression PCA	PCA 1: , n:2	1477.826135	0.366446

Table 5: Model selection to restaurants popularity.

IV. Discussion section.

When we started this project, the number of venues retrieved by the foursquare API was around 15000 restaurants in the city, after the filter and selection of data, just 1500 restaurants were maintained in our dataset for training, considering the limitations of the free developer account for foursquare it is difficult to obtain more observations without incur in an excessive spend of money, which is out of the scope in this project. One of the possible solutions to the lack of accuracy in our models would be the addition of information from another sources, like google maps API. Even though, the final result for this study, with two models with improvement chances it is remarkable.

The model with the best accuracy in both indicators is the Decision Tree Algorithm, but the difference with the other is not significantly better. We can associate this low result in our price prediction with the low amount of restaurant with all the required information, with approximately 150 neighborhoods in the city and just with 1500 observation , an amount of 10 restaurants by neighborhood is not enough information to obtain a better tuning of our models. Also, could be interesting to add information of competitors and poi around the selected restaurants, this information should improve this result.

As we could see in the develop of this project, there is strong relationship between demography and the success of a restaurant. It is a trade when the location chosen to open a new business is made in places with high concentration of wealth, even though, these types of location are more expensive. This study probe that the exposition , popularity, and prices are strongly related to the location.

V. Conclusion.

This project had as goal looking for relations between the location chose for a restaurant with its success. With ratios of 26% of failure, the restaurant business is one of the most fragile on the first year of working. The are different causes for this phenomenon, one of them is the poor chose of the physical space to work.

Given the conditions on this project, it was insufficient the amount of observation acquired from the foursquare database, from the initial 15000 restaurants retrieved with the API calls, just 1500 were suitable in for our objectives, create two predictive models for the popularity and price range of a restaurant in NYC.

Our first objective, the price estimator based in the location of a restaurant the best model were the Decision Tree model, but even being our best try it result inaccurate, this could be explained by the low amount of observation in every neighborhood of the city, around 10 restaurants by area. For sure improving the collection process will expand the results of this model.

As we could see in the develop of this project, there is strong relationship between demography and the success of a restaurant, it is a trade but being in area with high concentration of wealth is a good decision, even when these types of location are more expensive. This study probe that popularity, and prices are strongly related to the location.