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

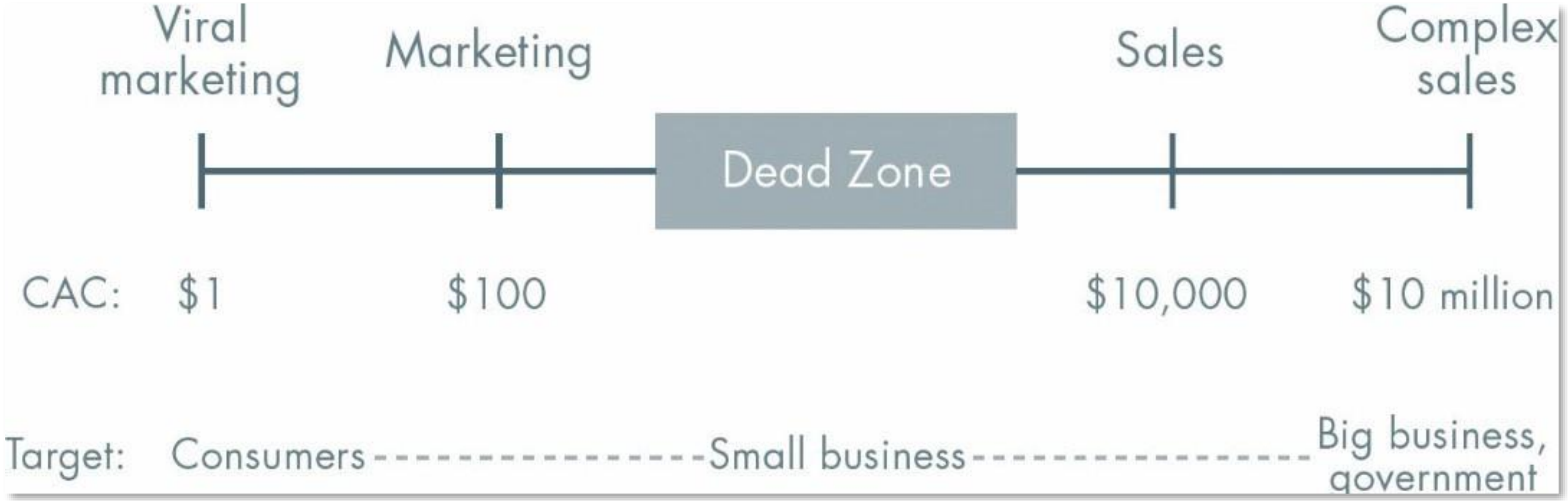
I. Introduction.

Without doubt choosing the right location for a new business is one of the most crucial steps in the life of a project.

This work have as a goal determine which factor must be taken in consideration when choosing a location for a new restaurant.

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.



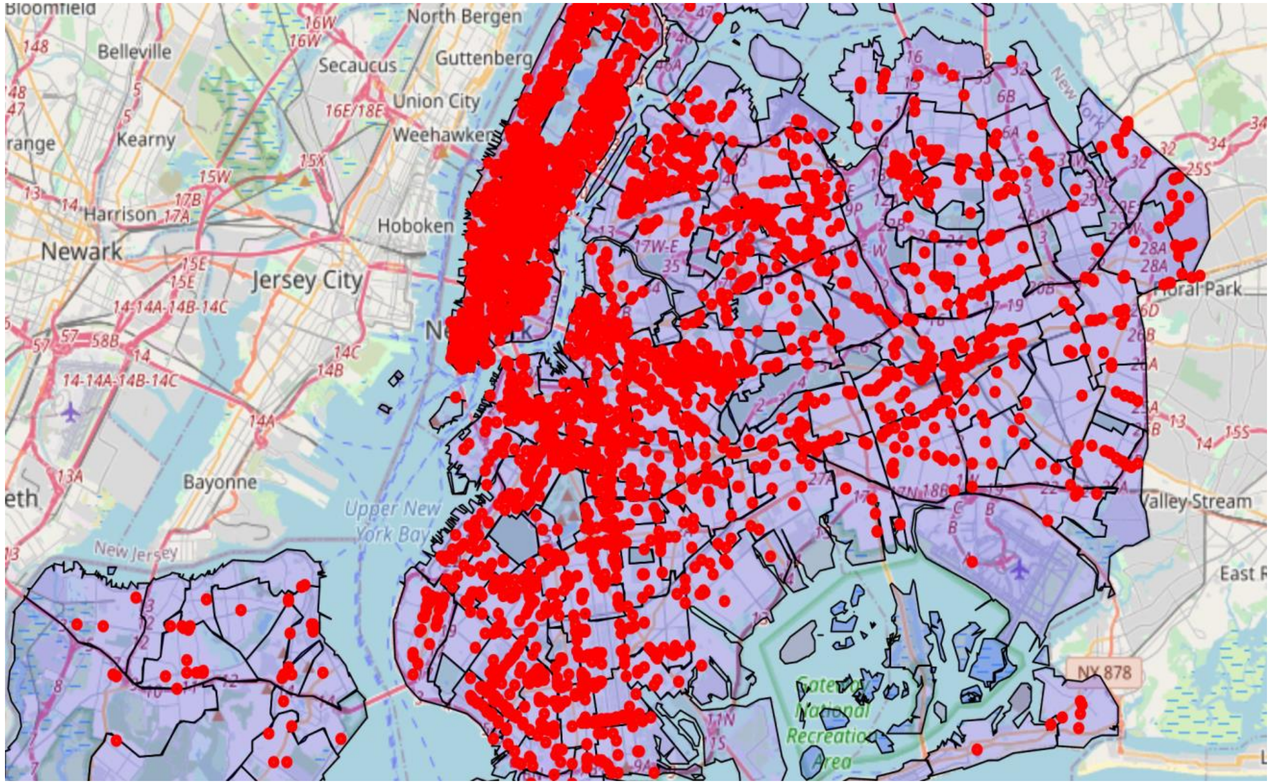
II. Data.

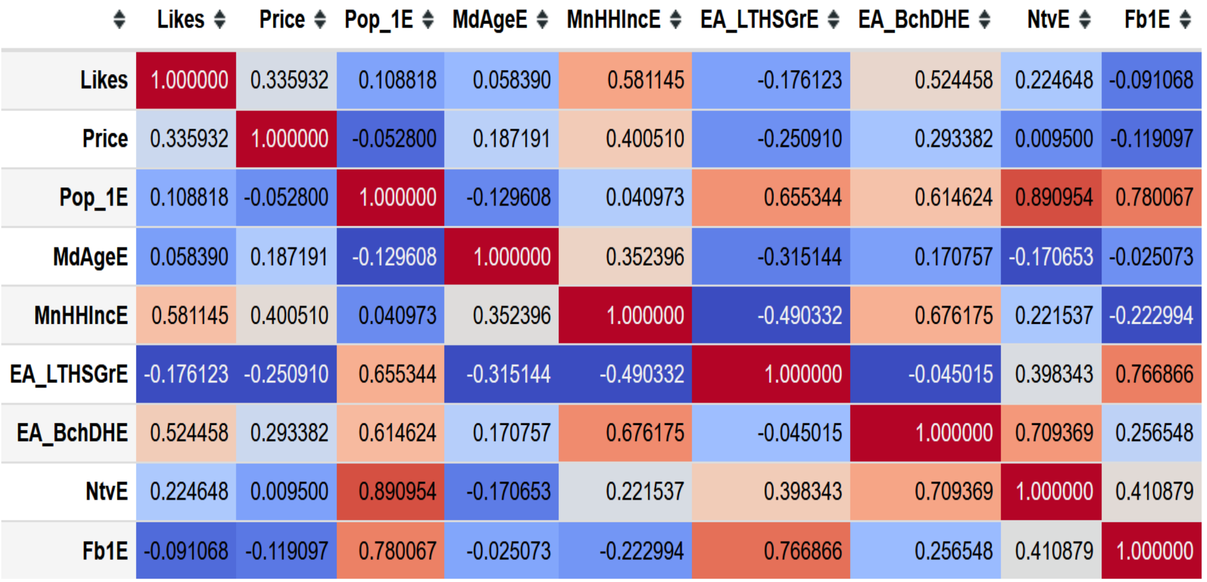
With the objective of predict how the location influence in the success or failure of a small/mid restaurant in the neighborhoods of New York City we collected data from different sources to complete this project.

- Foursquare API.
- Open City NY
- Census2020 website.

III. Methodology.

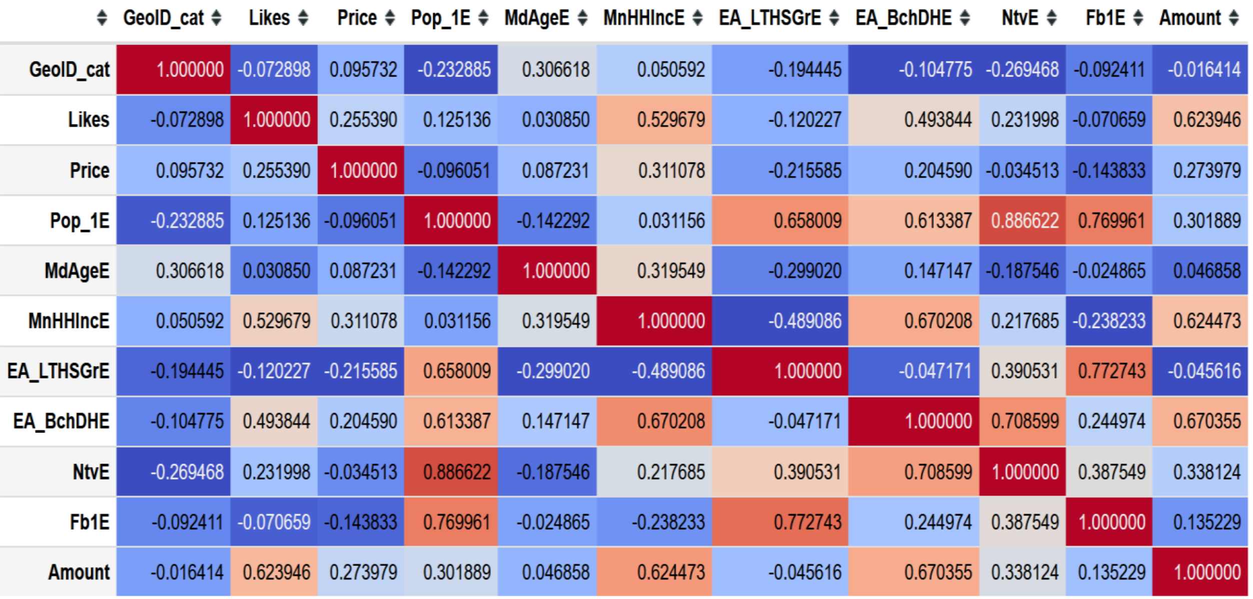
Our goal in this project was to find some relationships between the location of a restaurant and it success. With this in mind, we proceeded to analyze data from the neighborhoods of the city where every restaurant works.





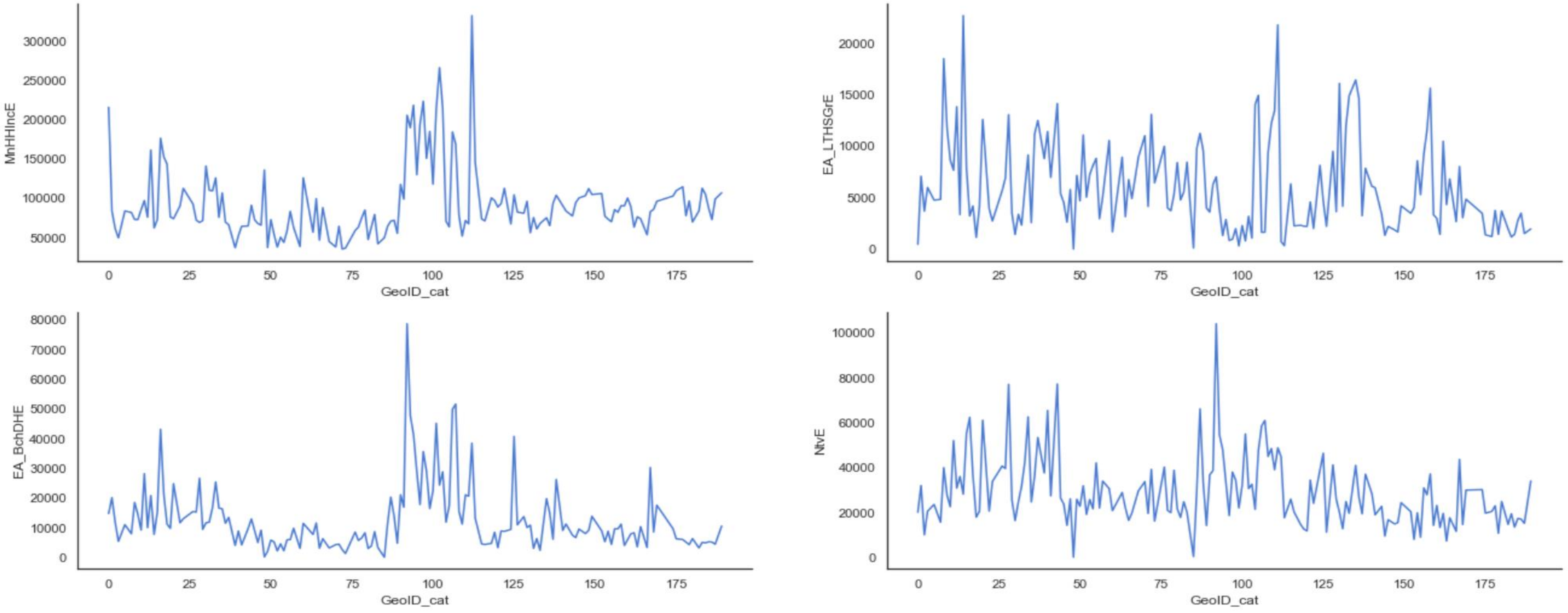
	Likes	Price	Pop_1E	MdAgeE	MnHHIncE	EA_LTHSGrE	EA_BchDHE	NtvE	Fb1E
Likes	1.000000	0.335932	0.108818	0.058390	0.581145	-0.176123	0.524458	0.224648	-0.091068
Price	0.335932	1.000000	-0.052800	0.187191	0.400510	-0.250910	0.293382	0.009500	-0.119097
Pop_1E	0.108818	-0.052800	1.000000	-0.129608	0.040973	0.655344	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

In section of our study the point was to find how the price and the popularity of a restaurant is influenced by the location and demography.

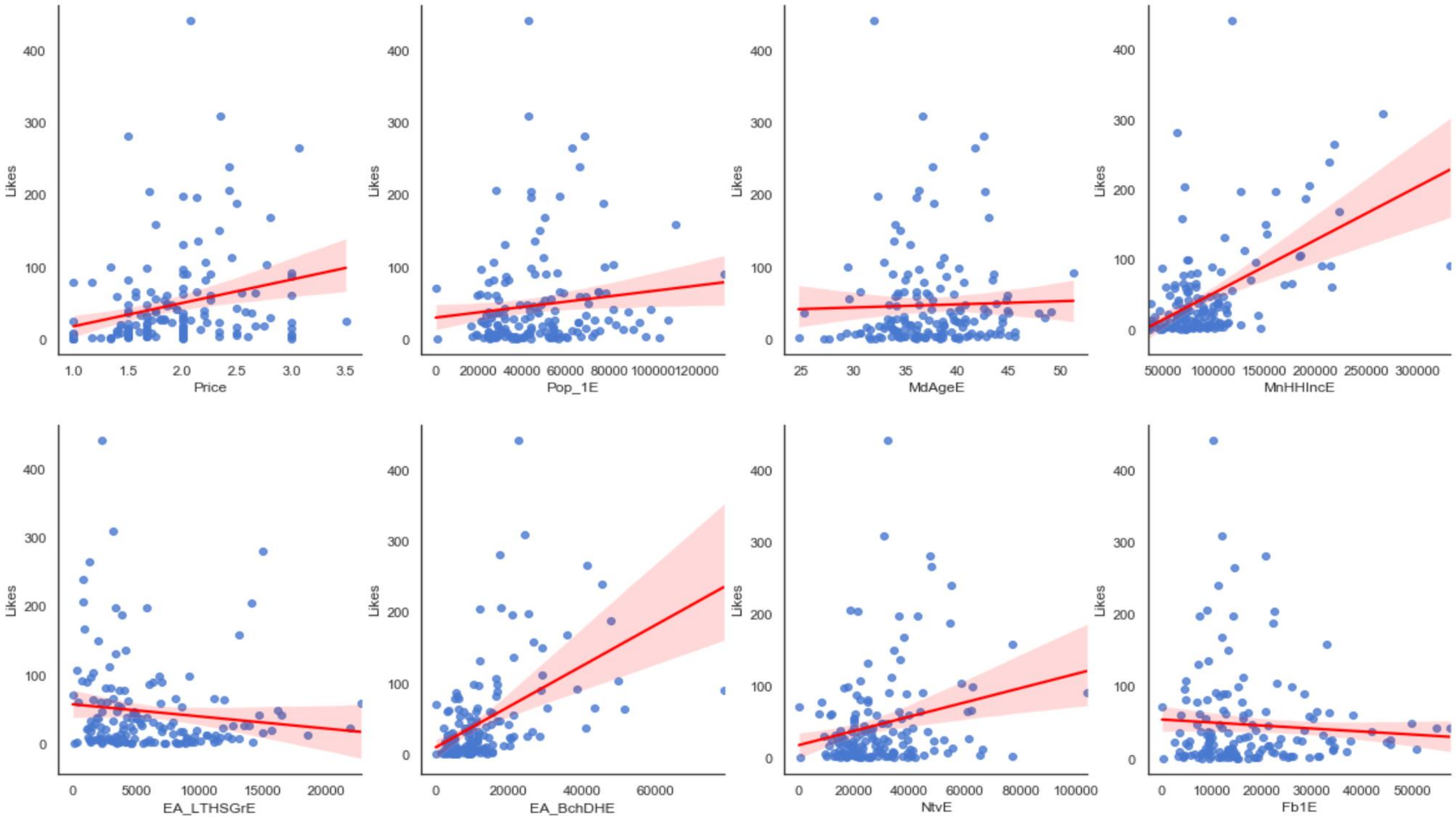


	GeoID_cat	Likes	Price	Pop_1E	MdAgeE	MnHHIncE	EA_LTHSGrE	EA_BchDHE	NtvE	Fb1E	Amount
GeoID_cat	1.000000	-0.072898	0.095732	-0.232885	0.306618	0.050592	-0.194445	-0.104775	-0.269468	-0.092411	-0.016414
Likes	-0.072898	1.000000	0.255390	0.125136	0.030850	0.529679	-0.120227	0.493844	0.231998	-0.070659	0.623946
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

Our visualization show us the correlation between the variables of the dataset. In this figure is clear that exist neighborhoods with more restaurants in the city, Here all the graphs shows a distinctive peak around the neighborhood 100 (Lenox Hill Manhattan), one of the neighborhoods with more professionals and the greatest annual income.



This graph show of how some variables strongly correlated with the price and popularity of the restaurants.



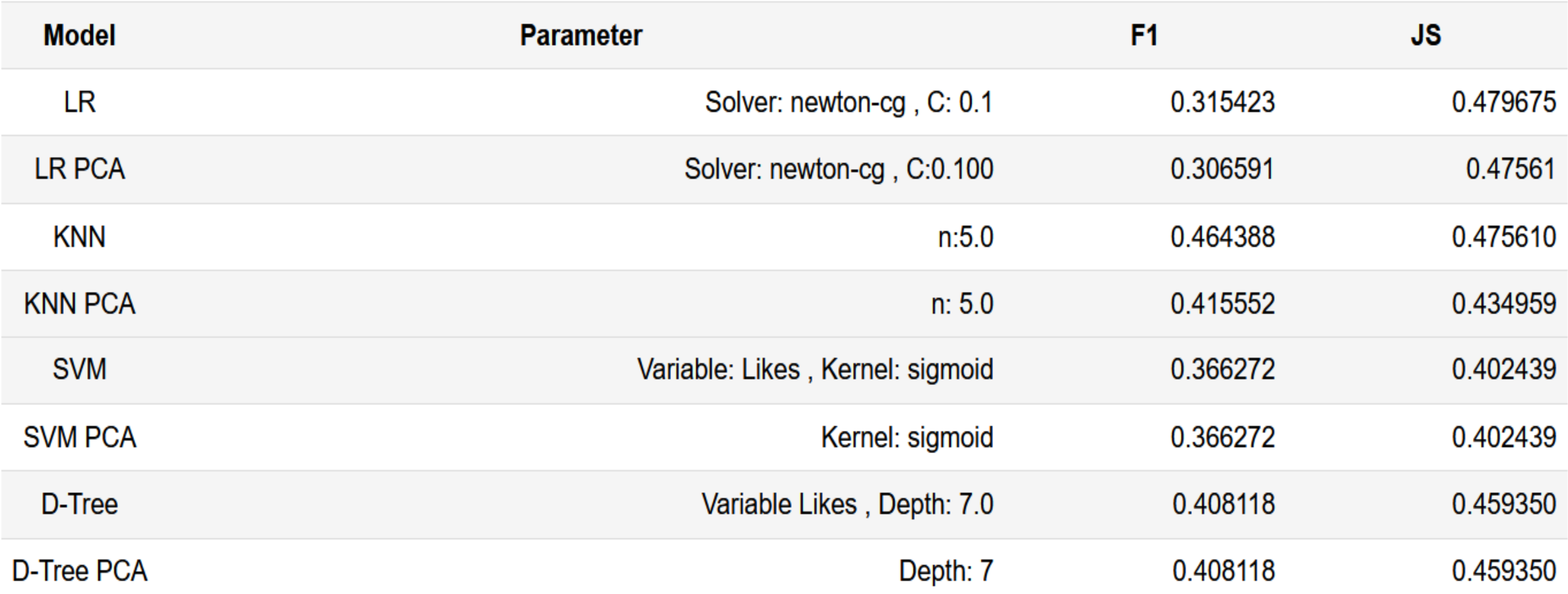
Machine learning model selection.

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.

Results section

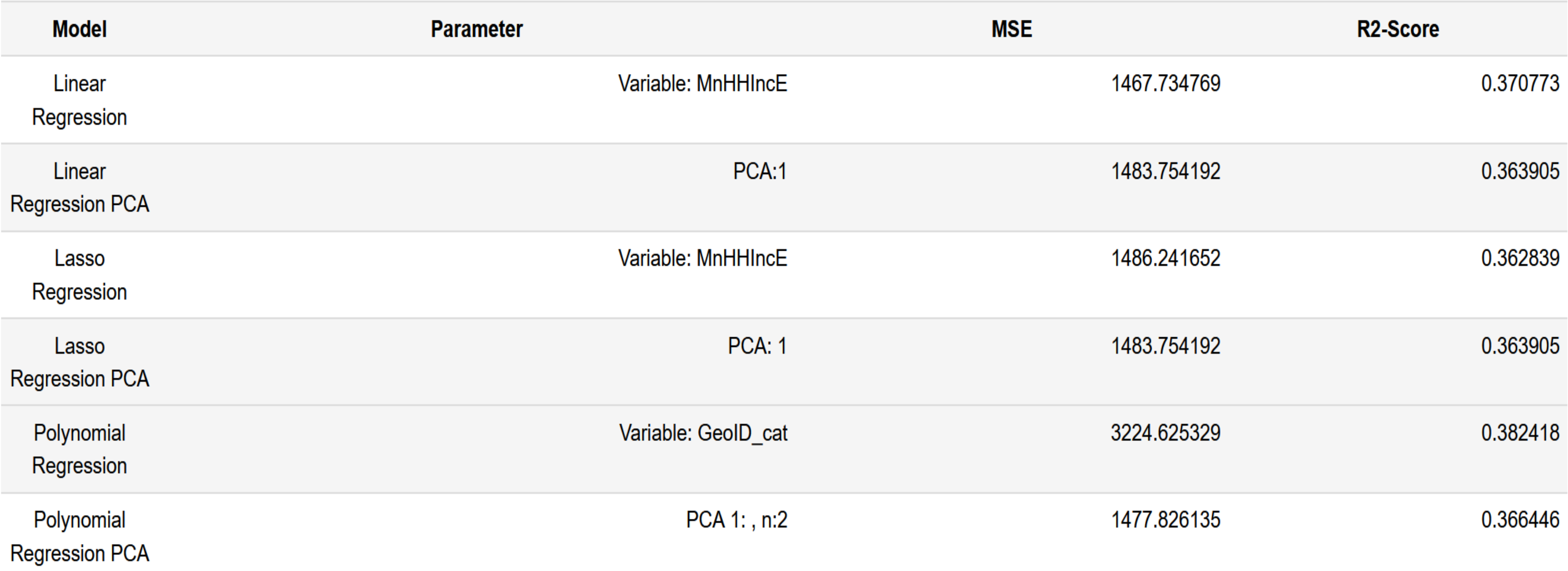
For the price prediction algorithm, the most accurate model was the K- Nearest Neighborhood with n=5, with an F1 Score of 0.46 and a Jaccard Similarity Score of 0.47. This result , should be improvable with more observations, this because our final dataset was reduced to 1500 observation, this is just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



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

Results section

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

Conclusion.

- 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.