To Eat or not to Eat: An inspection of restaurants in NYC

Introduction

The goal of our analysis is to figure out which variables in our dataset have the most influence on the inspection score awarded to each data observation, which is each individual restaurant in the city of New York.

If we were to dine at one restaurant in NYC every day, it would take us 68 years to have dined in each and every one of them. When we consider the limited amount of resources the government has for the inspection of restaurants to maintain a good standard of health and hygiene we see that there is a big discrepancy. Essentially, there are 24,294 unique restaurants in NYC and there are limited resources that can be used towards there regulation. Our analysis of the data will allow us to make recommendations to improve the allocation of the government's resources which will work towards not only maintain the health and hygiene of the city but also work towards better usage of the budget.

We start parsing through the data with an exploratory data analysis (EDA) approach. First, we remove the absolute duplicates, values that are the same on all levels of the data. Then, we proceed to conduct a multiple univariate, and bivariate analysis to get a better understanding of our data set to see how the data is distributed and, in this exploration, we also look for outliers, missing values, and unexpected values. Once we cleaned the data we created a linear multiple regression model using the variables. Furthermore, we create more models that include more of the independent variables to find the best representation of the dataset.

We did find one study on LinkedIn performed by a masters group from Carnegie Melon University that had used the same dataset. This team used supervised and unsupervised learning to make recommendations on identifying violation trends, predicting violations, and reducing violations. Their approach was similar to ours in so far as to use one statistical method, in this case supervised and unsupervised learning, to identify certain opportunities and make some recommendations.

Computational Setup/Steps:

Firstly, we imported the libraries that were required to perform our analysis and these libraries ranged from libraries such as numpy to libraries such as sklearn.preprocessing. We then went on to import the dataset from an online location and loading it into a pandas data frame. We then start an initial cleaning of the data frame by removing the absolute duplicates and before we get into the exploration of the dataset we preprocess the data, where we parse through the data to look for missing values, not acceptable values, and data points.

Once the data is clean and appropriate for the analysis we proceed with multiple univariate analysis to the observe the patterns for certain variables. These analyses give us information on how the data points are distributed amongst geographical variables such as 'Boroughs' (column name: boro), how the data is distributed amongst the descriptive variables such as cuisine, information on the violation codes breached, the spread of the inspection scores awarded. This information was garnered through descriptive functions inherent in the pandas library, donut charts, and boxplots. Using tools such as bar graphs in the bivariate analysis we get information on how the grades are classified and how the violations are distributed geographically. These graphs are present in the appendix of this document.

Once we have observed the data we get an idea of which variables might be influential in affecting the inspection score. On that basis, using our discretion for the independent variables, we created a regression model which had a certain R-square value. We then went on to add more independent variables that could possibly have an effect on the inspection score to observe what change it would have to the R-square value as a higher value would indicate that the new model would be a better representation of the dataset. From the observations of our regression model we are able to achieve the goal of our analysis which was to figure out which variables have the most impact on the inspection score awarded to restaurants.

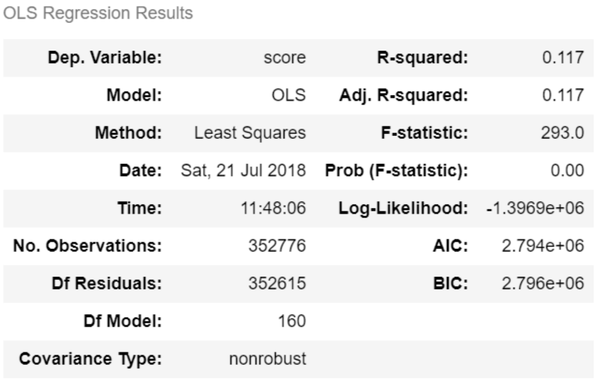
To extend our analysis we also performed an unsupervised Kmeans clustering analysis such that the model would calculate the accuracy of the grade distribution of our dataset. In this process we took five clusters (as there are five grade types) to cluster the categorical data that affects the awarded grade. This model predicted the accuracy of the grades awarded in our dataset with approximately thirty six percent accuracy.

Lastly, we have provided the reader with a set of heat maps and interactive plots to allow the reader to better visualize the data. Moreover, a codebook is also provided to aid in the understanding of the dataset.

The K-means clustering is the slowest part of the code as it takes 40.7s +/- 989 ms per loop(mean +/- standard deviation of 7 runs, 1 loop each). If we double the data the clustering will take about 500.14 s+/- 876 ms per loop.

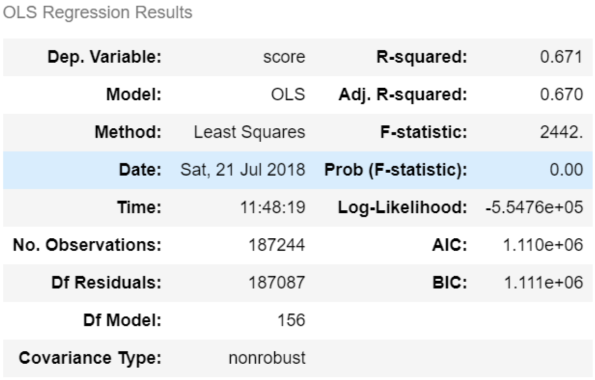
What were some computational challenges and how did you go about solving them?
One of the computational challenge we came across was the plotting of folium map that shows distribution of restaurants across various zip codes. As we had a lot of restaurants to plot, which was creating computational challenges, we decided to have a single latitude and longitude plot for each zip code that made the plot possible.

Results


OLS Regression Results

Dep. Variable:	score	R-squared:	0.117
Model:	OLS	Adj. R-squared:	0.117
Method:	Least Squares	F-statistic:	293.0
Date:	Sat, 21 Jul 2018	Prob (F-statistic):	0.00
Time:	11:48:06	Log-Likelihood:	-1.3969e+06
No. Observations:	352776	AIC:	2.794e+06
Df Residuals:	352615	BIC:	2.796e+06
Df Model:	160		
Covariance Type:	nonrobust		

Fig. Reg.1


OLS Regression Results

Dep. Variable:	score	R-squared:	0.671
Model:	OLS	Adj. R-squared:	0.670
Method:	Least Squares	F-statistic:	2442.
Date:	Sat, 21 Jul 2018	Prob (F-statistic):	0.00
Time:	11:48:19	Log-Likelihood:	-5.5476e+05
No. Observations:	187244	AIC:	1.110e+06
Df Residuals:	187087	BIC:	1.111e+06
Df Model:	156		
Covariance Type:	nonrobust		

Fig. Reg.2

Based on the correlation table in our jupyter notebook, for our initial regression model (results in Fig.Reg.1) we have taken into consideration the boroughs, critical flags, cuisine description, and violation codes, which resulted in the model having a low R-square value indicating that the model is not a very good representation of the data. So, we changed the variables in the equation using the p-values from the first regression and we built a second regression (results in Fig.Reg.2) which has a better R-square value indicating that this second model better represents our data.

The appendix also provides a boxplot indicating the distribution of inspection scores for restaurant chain and independent restaurants, and it is evident that the restaurant chains have a lower inspection score indicating that as a collection of restaurants they are more hygienic than the independent restaurants.

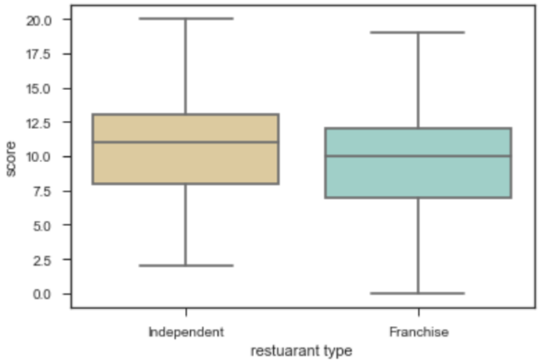
Conclusion:

From the above results it is clear that the variables that have a high influence on the model are the grade, borough, cuisine description and violation code.

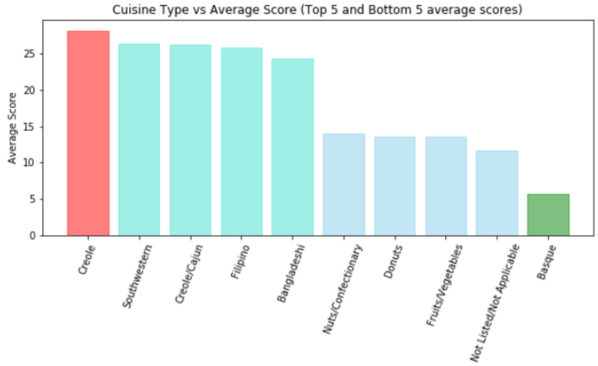
An interesting thing to investigate further would be to see if any particular set of violation codes always occur together, and if so, do they have any pattern of occurrence. This analysis could provide valuable insight into which particular violation problems could be targeted to improve the inspection scores of particular areas.

Appendix

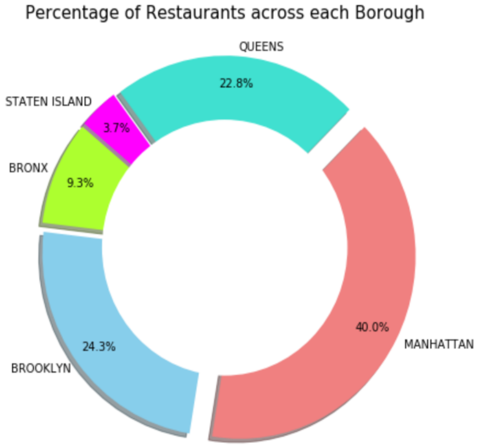
Univariate and Bivariate Plots



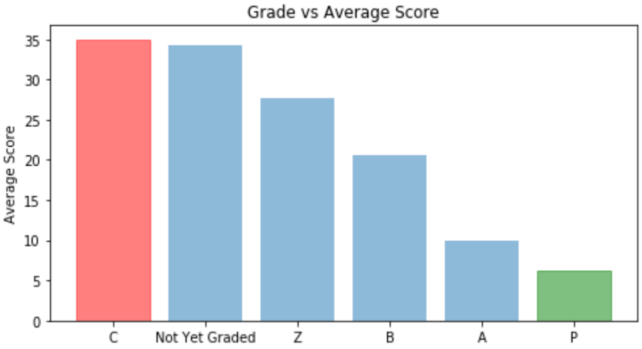
Shows the distribution scores across restaurant types



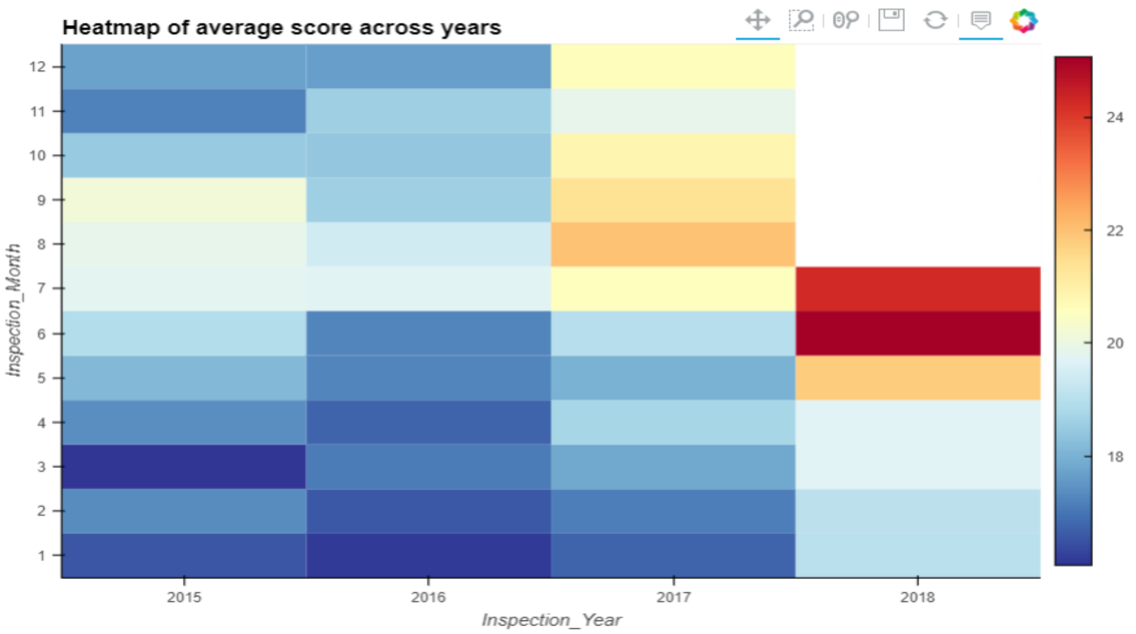
Displays top 5 and bottom 5 cuisine types based on average score



Shows %restaurants across each borrow with Manhattan having highest number of restaurants

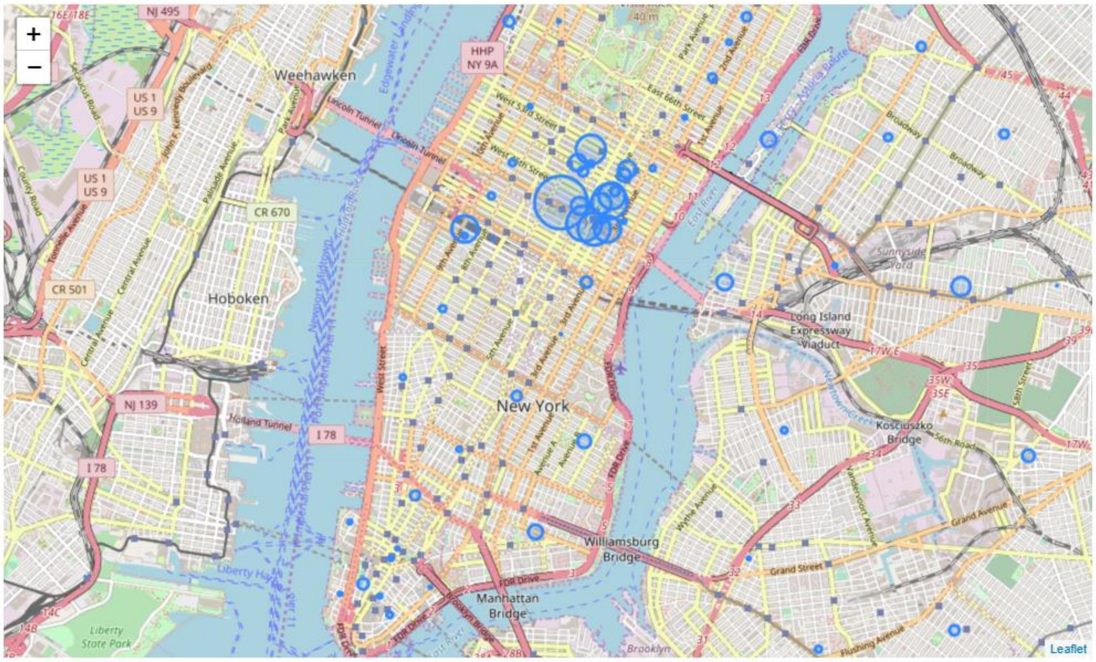


Displays the average scores across different grades with the C having the highest average score marked in red



Shows the average score distribution of the relevant year based on data 2015 to 2018

Data Visualization


+
−
NJ 495
Weehawken
HHP
NY 9A
US 1
US 9
CR 670
Hoboken
CR 501
NJ 139
Holland Tunnel
I 78
Lincoln Tunnel
West 53rd Street
West 49th Street
East 66th Street
New York
Williamsburg Bridge
Manhattan Bridge
FDR Drive
Long Island Expressway Viaduct
Kosciuszko Bridge
Sunnyside Yard
Grand Street
Grand Avenue
Flushing Avenue
Liberty State Park
Brooklyn
Leaflet