# Title

New York City Restaurant Inspection Score Prediction

# Introduction & Problem Description

The Health Department of New York City conducts unannounced inspections of restaurants at least once a year. Inspectors check for compliance in food handling, food temperature, personal hygiene and vermin control. Each violation of a regulation gets a certain number of points. At the end of the inspection, the inspector totals the points, and this number is the restaurant's inspection score—the lower the score, the better the Grade.

Due to many factors, some restaurants inspection data may be missing or for new restaurants that are not yet inspected; there would be no scores present.

This project used Foursquare API to explore neighborhoods in New York City to get the most common restaurant categories in a particular neighborhood. Then various other dataset will be used along with restaurant inspection data to build a model to predict scores for a restaurant based which does not have a pre-inspected score.

Target Audience:

This model would be useful for a restaurant that has not been inspected to have an idea of possible inspection score and take careful measures to improve its score during first inspection from NYC Health Department.

# Data Description

Below are the data set used for this project along with the source of availability.

## DOHMH New York City Restaurant Inspection Results

The dataset contains every sustained or not yet adjudicated violation citation from every full or special program inspection conducted up to three years prior to the most recent inspection for restaurants and college cafeterias in an active status on the RECORD DATE (date of the data pull). When an inspection results in more than one violation, values for associated fields are repeated for each additional violation record. Establishments are uniquely identified by their CAMIS (record ID) number. Keep in mind that thousands of restaurants start business and go out of business every year; only restaurants in an active status are included in the dataset.

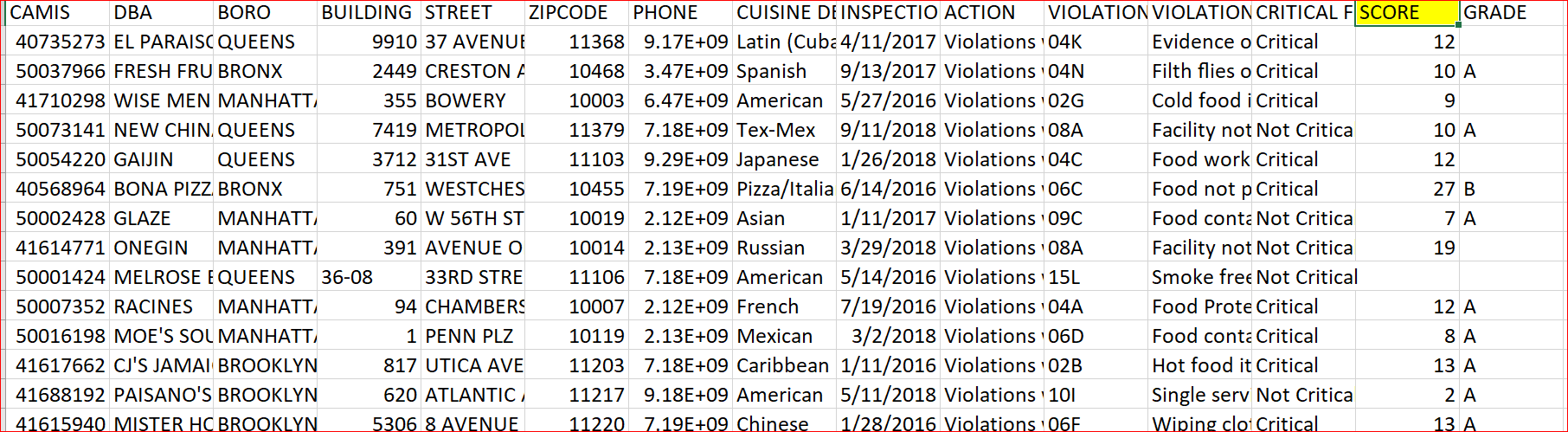
Records are also included for each restaurant that has applied for a permit but has not yet been inspected and for inspections resulting in no violations

Website : <https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j>

Data is downloaded from this site as renamed to resturant\_inspect.csv for use in the project.

A subset of the data used to consider volume constraints.

Glimpse of the data:



The Score attribute from the data set will be used in the prediction model for restaurants on a particular zip that do not have inspection score yet

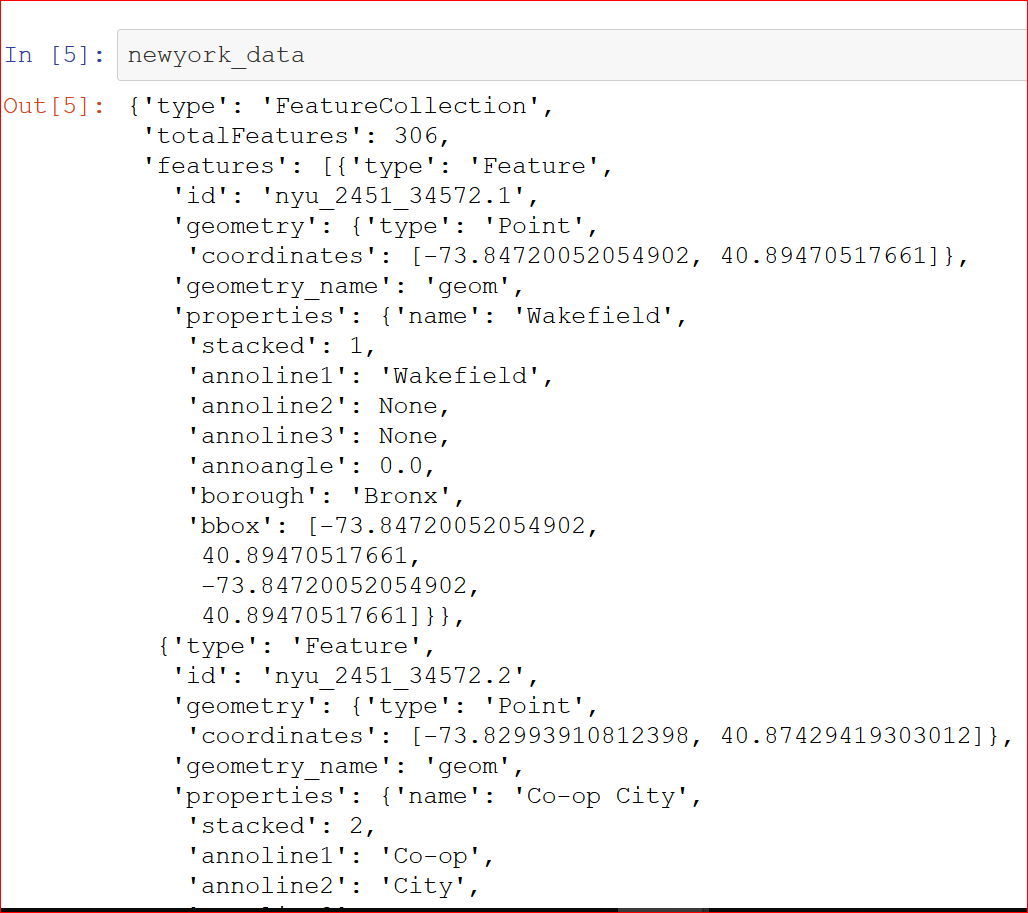
## NYC boroughs and the neighborhoods data

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Luckily, this dataset exists for free on the web. Feel free to try to find this dataset on your own, but here is the link to the dataset: <https://geo.nyu.edu/catalog/nyu_2451_34572>

For convenience, the course admins. downloaded the files and placed it on the server, so we can simply run a wget command and access the data as used in other labs. The data is available in json format.

Glimpse of the data



## Foursquare API data

Foursquare API to explore neighborhoods in New York City. You will use the **explore** function to get the most common venue/restaurant categories in each neighborhood.

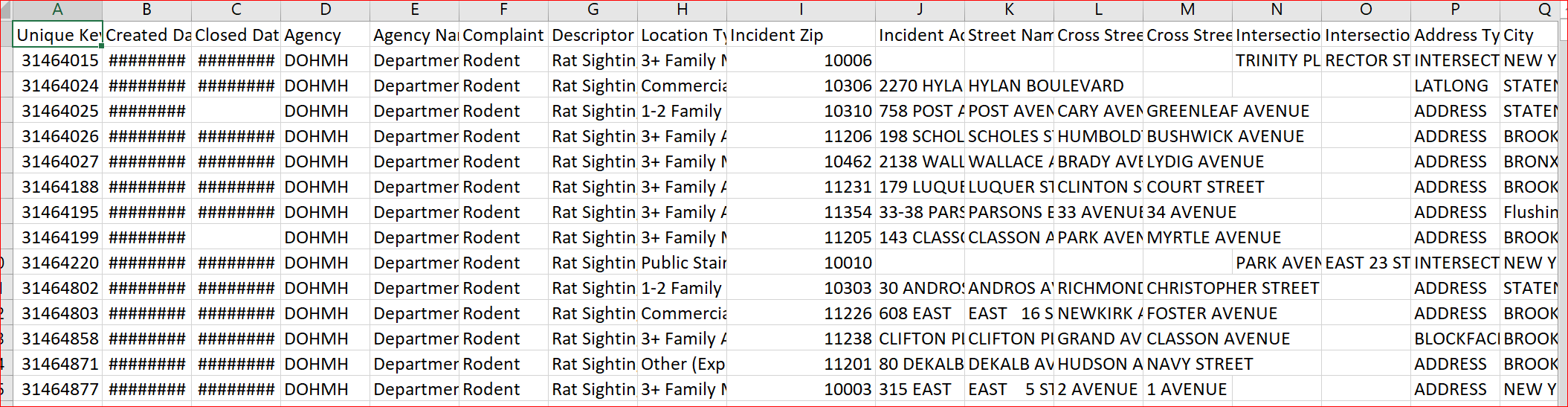
This project used Foursquare API to explore neighborhoods in New York City to get the most common restaurant categories in a particular neighborhood

## NYC Rat Inspection Data

The [Rat Information Portal (RIP)](http://maps.nyc.gov/doitt/nycitymap/template/?applicationName=DOH_RIP) is a web-based mapping application where users can view rat inspection data. Findings from the Health Department's inspections are searchable by address, or by borough, block and lot (BBL).

Information about the most recent inspections, compliance, baiting’s, and cleanups on any given property are available by city , Incident zip etc.

Glimpse of the data:



In this project, it will be analyzed if there is any correlation with rat sighting incident in a zip with the restaurant inspection score. The data is downloaded from the portal and stores as Rat\_Sightings.csv in the project

## NYC Rolling Sales Data

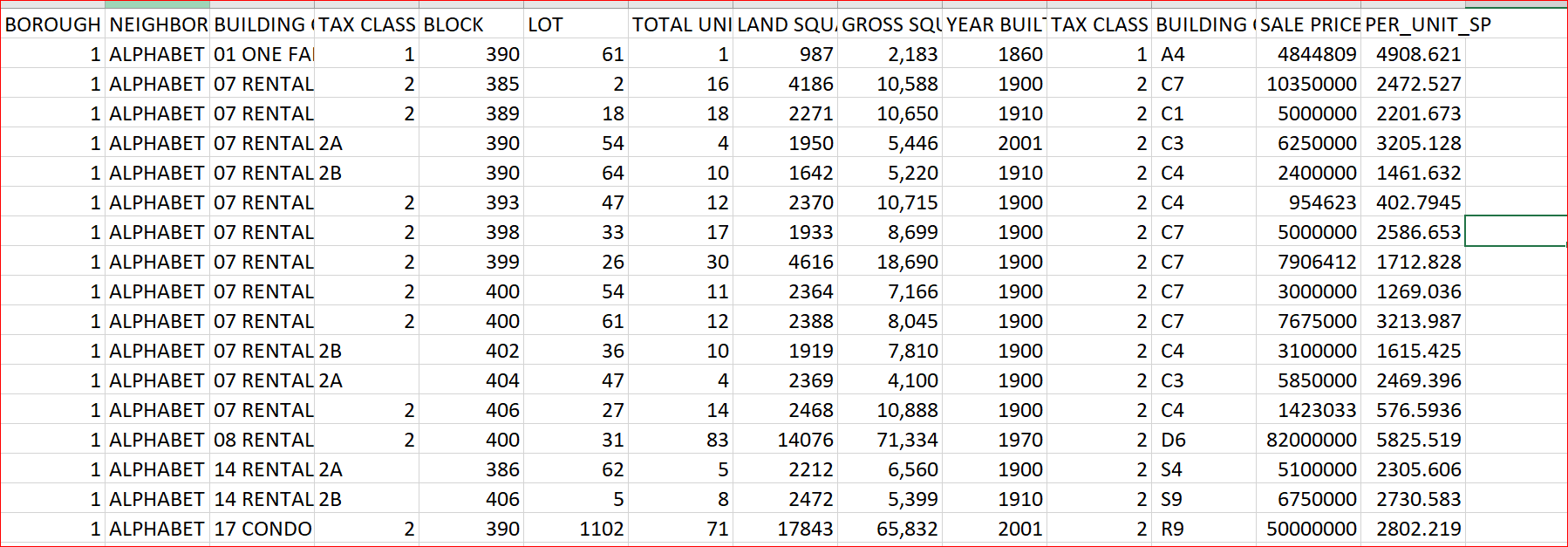
The Department of Finance’s Rolling Sales files lists properties that sold in the last twelve-month period in New York City for all tax classes. These files include:

* the neighborhood;
* building type;
* square footage;
* other data.

Site : <https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>

The per unit sales prices of the data for zipcodes will be used to find any correlation with the restaurant inspection score.

Glimpse of the Data:



# Methodology

INDEPENDENT/DEPENDENT VARIABLES:

For the analysis, the dependent variable is Restaurant Inspection Score and everything else is considered to be an independent variable.

It is worth noting that lesser score is an indication of good inspection report and vice versa.

Since all variables utilized contain continuous numerical data and we want to predict the Inspection Score given certain input parameters, it would make most sense to apply Regression Analysis using Python.

### Part A : Explore neighborhoods in New York City to get the most common restaurant categories in a particular neighborhood

1. In this project, we will start of NYC boroughs and the neighborhoods data.

#### The json data will be transformed into a pandas data frame & goopy library will be used to get the latitude and longitude values of New York City

1. Next, we are going to start utilizing the Foursquare API to explore the venues in Manhattan neighborhoods using pre-registered Foursquare client ID

#### Once Manhattan data is available, the top 100 venues that are in ‘Marble Hill’ within a radius of 500 meters are collected

### Part B: Data Wrangling & Grouping Necessary Data Sets

1. All the other data sets are collected and data wrangling approaches are applied to clean null values , group data and keep only necessary attribute columns
2. Data sets are joined compared based on zip codes

### Part C: Exploratory Data Analysis

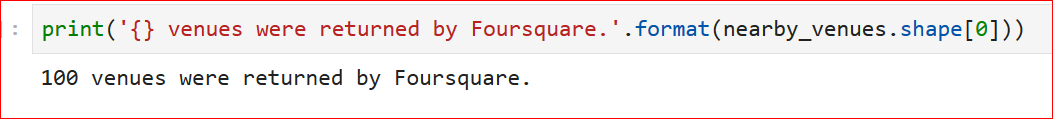
1. Here different numerical variables like Unit Sales price, Rat sight counts are used to explore the main characteristics which have the most impact on Inspection score for different locations based on zip code
2. Scatter plots with fitted lines using packages "Matplotlib" and "Seaborn" are used to visualize the relationship between Inspection score with different continuous numeric variables
3. Then **Pearson Correlation Coefficient & P value method** is usedto measure the linear dependence between Inspection Score variable and other dependent variables.
4. All the above steps give a better idea of what our data looks like and which variables are important to take into account when predicting the Inspection Score.

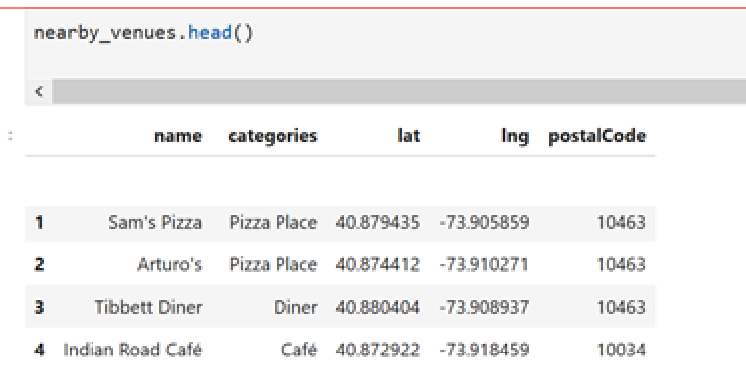
### Part D : Model Development

1. In this section, we will develop predictive Linear Regression model that will predict the Inspection Score of a restaurant in a particular zip code that do not have an inspection score yet.
2. Using simple linear regression model , we will create a linear function with the predictor variable and the "Inspection Score" as the response variable.

# Results

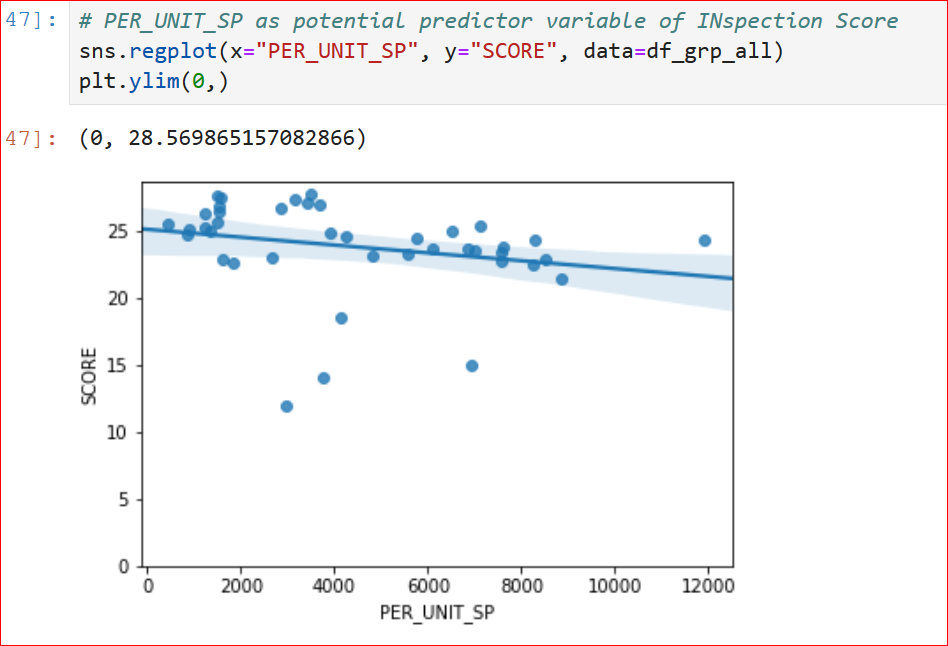
#### The top 100 venues that are in Marble Hill within a radius of 500 meters





Exploring the relation between Inspection Score with mean Per Unit Sale Price for same zip codes

Visualization via Scatter Plots with Fitted Lines

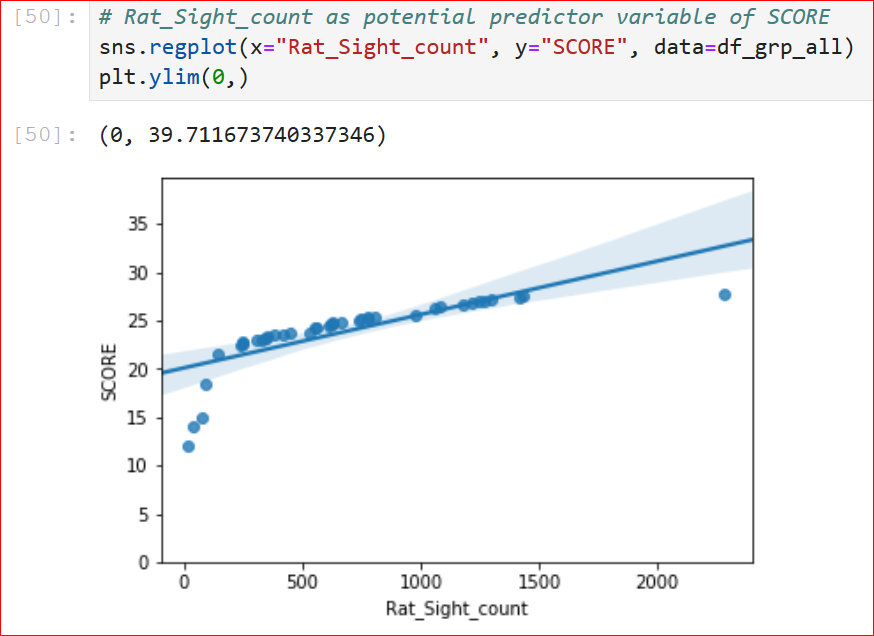


Correlation Analysis

The Pearson Correlation Coefficient is -0.24234552621138997 with a P-value of P = 0.12685080654974348

Exploring the relation between Inspection Score with rat Sight counts for same zip codes

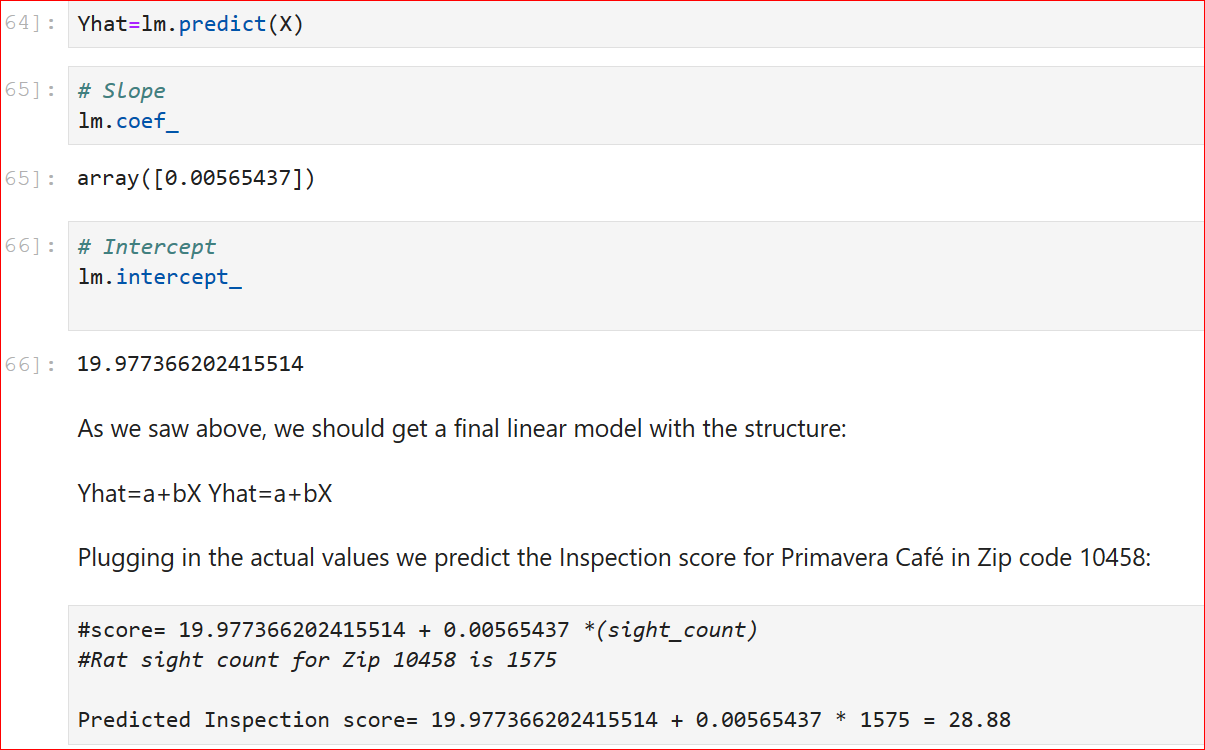
Visualization via Scatter Plots with Fitted Lines



Correlation Analysis

The Pearson Correlation Coefficient is 0.7657046069878928 with a P-value of P = 5.4659593536957175e-09

Predictive Model results



# Discussion

This project used Foursquare API to explore neighborhoods in New York City to get the most common 100 restaurant categories in Marble Hill within a radius of 500 meters.

After that different numerical variables like Unit Sales price, Rat sight counts are used to explore the main characteristics which have the most impact on Inspection score for different locations based on zip code based on scatter plot visualization & **Pearson Correlation Coefficient & P value method.**

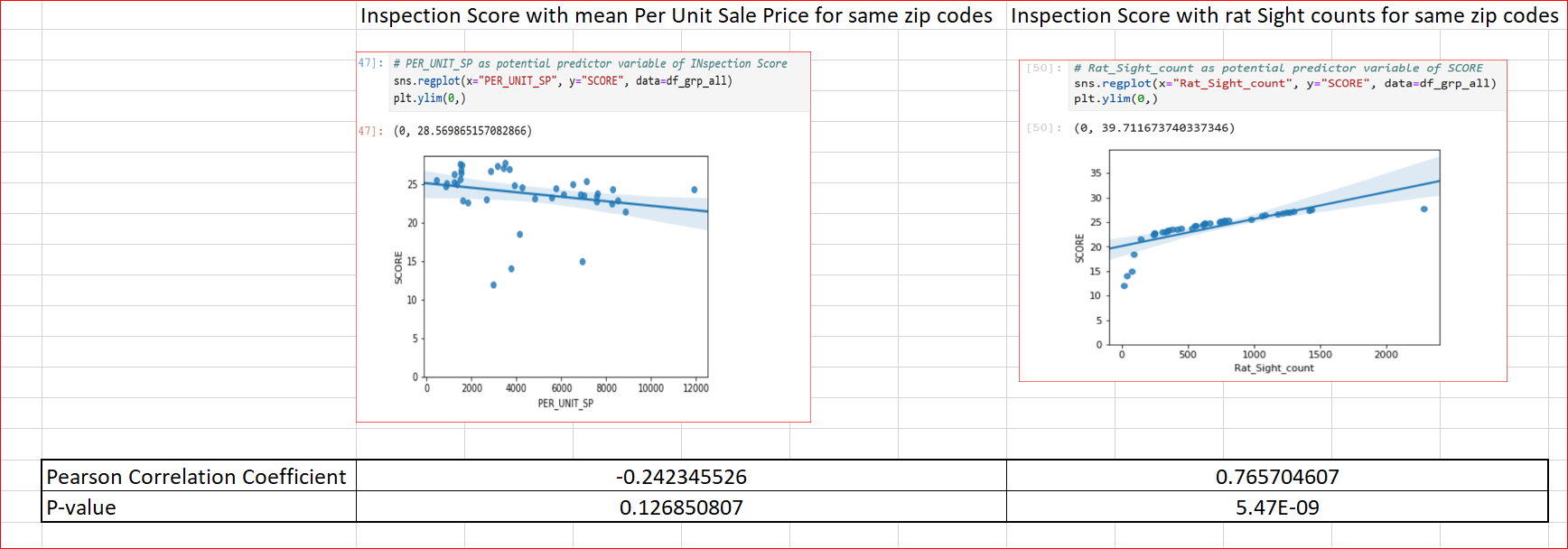
The Pearson Correlation measures the linear dependence between two variables, X and Y. The resulting coefficient is a value between -1 and 1 inclusive, where:

* **1**: total positive linear correlation,
* **0**: no linear correlation, the two variables most likely do not affect each other
* **-1**: total negative linear correlation.

The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when the p-value is:

* < 0.001 we say there is strong evidence that the correlation is significant,
* < 0.05; there is moderate evidence that the correlation is significant,
* < 0.1; there is weak evidence that the correlation is significant, and
* is > 0.1; there is no evidence that the correlation is significant.



Based on the visualization & correlation analysis , it is evident that Per unit Sales price has little or no relation for Inspection score values for same zip codes.

On the other hand , we can identify a strong positive correlation of Inspection score with the rat sight counts for same zip codes. As the rat sight count increases, the inspection score also increases.

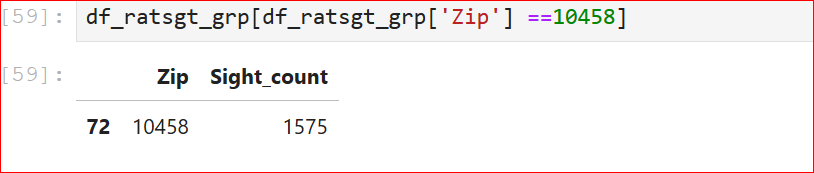
As we know that the lower the inspection score, the better is the grade for restaurant. So if an area has more right sight incidents, we can relate that the restaurants in that area will have a probability to get bad inspection score ( i.e. higher score) from authorities.

Now let’s take any zip from the top 100 restaurant venues (from our dataset of Marble Hill within a radius of 500 meters) and identify a Zip which does not have inspection conducted yet.

As an example , Primavera Café in Zip code 10458 is taken an example.



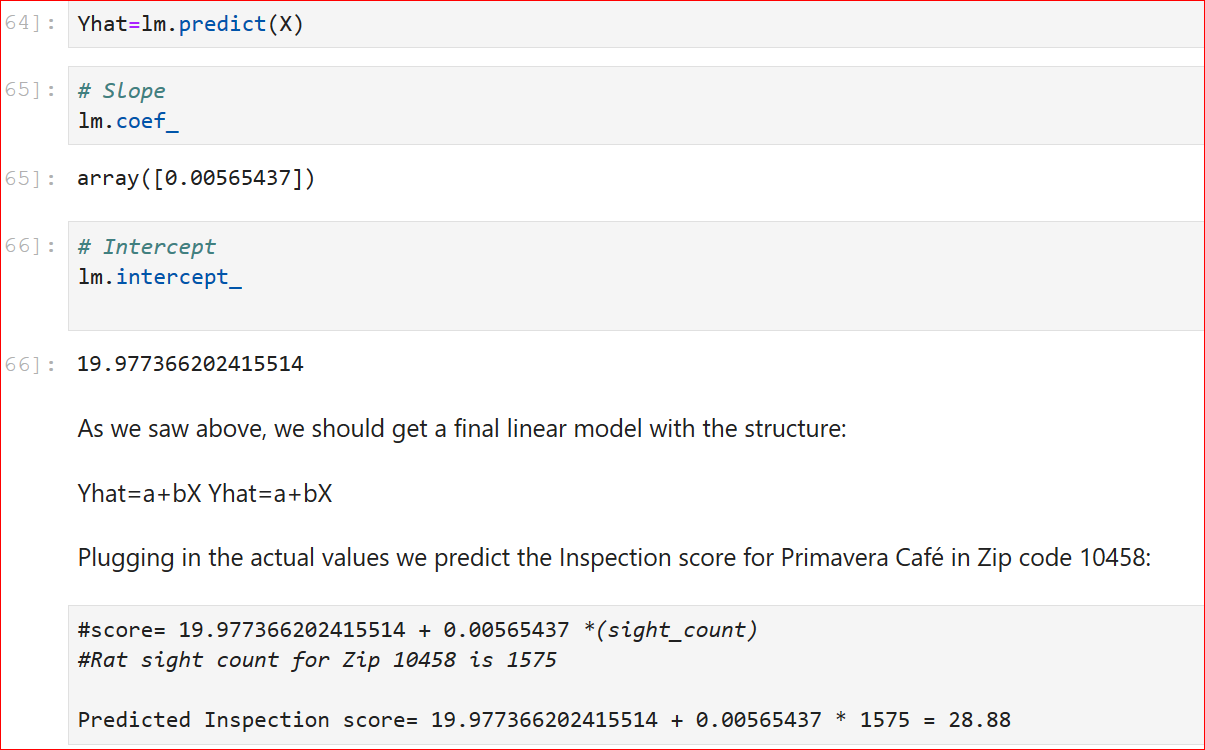
Although the Zip code 10458 does not have an entry in Inspection data , we are able to find the right sight count in the other data set.



Next we develop a predictive Linear Regression model that will predict the Inspection Score of a restaurant in a particular zip code that do not have an inspection score yet.

Using simple linear regression model , we created a linear function with the predictor variable and the "Inspection Score" as the response variable.

This can be used to predict the Inspection score for the Primavera Café in Zip code 10458 which is 28.88 as shown below.



# Conclusion

As part of this project, I was able to analyze different neighborhoods in NYC and identify top 100 venues in Marble Hill Neighborhood using Foursquare API.

Using visualization plots , exploratory & predictive analytics, I devised regression model to predict the Restaurant Inspection Scores and concluded that Rat Sighting Incidents have a positive correlation with the score.

For further evaluation , more models can be created using other combination of variables & can be split & trained accordingly. With Regression & Residual plot analysis with Multiple Regression Techniques a more accurate prediction model can be built.