Coursera Capstone Project for IBM Data Science Specialization - Week 2

By: Arnav Ashank

1. Introduction

1.1. Background

New York City is the most populous city in U.S. and is home to many immigrant population in New York after. Furthermore, it is the largest metropolitan area in the world with 18 million people as of 2010 with an estimated population of 18,897,109 residents.

Being a metropolitan city, New York City is also home to many restaurants which serves wide variety of cuisines. Owing to significant number of Indian expatriate population, New York City and its nearby Suburbs have handful of Indian restaurant.

If someone from India visits New York City for the first time, it will be useful if he/she have some prior information about the Indian Restaurants in New York City and how good they are. Moreover, prior information on location of other restaurants and their violation history will help in coming to an informed decision.

So, as a part of this project using the New York City Inspection Data and FourSquare API Indian restaurants in New York City will be listed, visualized and rated.

1.2. Problem Description

By utilizing the New York City restaurants inspection data, Indian Restaurants in New York City and their risk category will be analysed. Secondly, a classsifier model will be built to predict the risk categories of restaurants. Furthermore, using the foursquare API the ratings of Indian Restaurants in New York City will be obtained.

1.3. Target Audience

- People looking to open new restaurants
- Restaurants
- Travellers who love Indian food

2. Data

For this project I will use the following data :

1. New York City restaurants inspection data from 2016-2019
    - Data source: https://data.cityofnewyork.us/api/views/43nn-pn8j/rows.csv?accessType=DOWNLOAD
    - Description : This data set contains 386000 rows and 26 columns with Restaurant Name, Street Name, violation descriptions along with their latitude and longitude. This data will be downloaded and used.

	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION	...	RECORD DATE	INSPECTION TYPE
0	50054214	GUAC	Manhattan	179	AVENUE B	10009.0	2122544822	Mexican	12/19/2018	Establishment Closed by DOHMH. Violations wer...	...	05/16/2020	Cycle Inspection / Re-inspection
1	40401912	O'NIEALS	Manhattan	174	GRAND STREET	10013.0	2129419119	American	09/20/2017	Violations were cited in the following area(s).	...	05/16/2020	Smoke-Free Air Act / Initial Inspection
2	40861946	SUPPER	Manhattan	156	EAST 2 STREET	10009.0	2124777600	Italian	03/27/2018	Violations were cited in the following area(s).	...	05/16/2020	Cycle Inspection / Initial Inspection
3	50078428	CHAMPION COFFEE	Manhattan	319	EAST 14 STREET	10003.0	9172615949	Café/Coffee/Tea	12/27/2019	Violations were cited in the following area(s).	...	05/16/2020	Cycle Inspection / Initial Inspection
4	50057673	YAMA RAMEN	Manhattan	60	WEST 48 STREET	10036.0	2128326688	Japanese	11/08/2018	Violations were cited in the following area(s).	...	05/16/2020	Cycle Inspection / Initial Inspection

Fig.1. Snapshot of the New York City Restaurants Inspection data loaded into a data frame

This dataset contains most of the information that will be needed for the project such as location information, street name, etc., However, this dataset contains mixed datatypes in all 26 columns and needs extensive cleaning before it can be used for the project.

```
CAMIS                    int64
DBA                     object
BORO                    object
BUILDING                object
STREET                  object
ZIPCODE                float64
PHONE                   object
CUISINE DESCRIPTION     object
INSPECTION DATE         object
ACTION                  object
VIOLATION CODE          object
VIOLATION DESCRIPTION   object
CRITICAL FLAG           object
SCORE                  float64
GRADE                   object
GRADE DATE              object
RECORD DATE             object
INSPECTION TYPE         object
Latitude               float64
Longitude              float64
Community Board        float64
Council District       float64
Census Tract           float64
BIN                    float64
BBL                    float64
NTA                     object
```

Fig.2. A snapshot of datatypes of the columns in the dataframe.

2. Ratings of Indian restaurants for selected locality in New York City
   - Data source : FourSquare API
   - Description : By using this api we will get all the ratings for Indian restaurants in selected neighbourhood

2.1. Data Pre-processing

The loaded data frame is further subjected to processing before it has been utilized. So in this regard, a new data frame with only columns of interest were created. The rows that were having any null values have been dropped.

	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION	...	RECORD DATE	INSPECTION TYPE	La
0	50054214	GUAC	Manhattan	179	AVENUE B	10009.0	2122544822	Mexican	12/19/2018	Establishment Closed by DOHMH. Violations wer...	...	05/16/2020	Cycle Inspection / Re-inspection	40.7
1	40401912	O'NIEALS	Manhattan	174	GRAND STREET	10013.0	2129419119	American	09/20/2017	Violations were cited in the following area(s).	...	05/16/2020	Smoke-Free Air Act / Initial Inspection	40.7
2	40861946	SUPPER	Manhattan	156	EAST 2 STREET	10009.0	2124777600	Italian	03/27/2018	Violations were cited in the following area(s).	...	05/16/2020	Cycle Inspection / Initial Inspection	40.7
3	50078428	CHAMPION COFFEE	Manhattan	319	EAST 14 STREET	10003.0	9172615949	Café/Coffee/Tea	12/27/2019	Violations were cited in the following area(s).	...	05/16/2020	Cycle Inspection / Initial Inspection	40.7
4	50057673	YAMA RAMEN	Manhattan	60	WEST 48 STREET	10036.0	2128326688	Japanese	11/08/2018	Violations were cited in the following area(s).	...	05/16/2020	Cycle Inspection / Initial Inspection	40.7

5 rows × 26 columns

Fig. 3.A snapshot of the modified data frame.

The shape of the modified data frame is now reduced to 188969 rows and 26 columns from the original size of 386000 rows and 26 columns.

The risk categories are explicitly define in the inspection data. Hence, a new column named risk is created in the above data frame. by binning the inspection scores into following categories using the guide https://www1.nyc.gov/site/doh/business/food-operators/letter-grading-for-restaurants.page

Three categories were created for ease of classification:

a. Low Risk - Inspection Score between -2 to 13
b. Medium Risk - Inspection Score between 13 to 27
c. High Risk - Inspection Score between 27 to 100

In [13]:
```
# Create risk category bins
bins = [-2, 13, 27, 100]
labels =["Low Risk","Medium Risk","High Risk"]
nyc_mod_df['Risk'] = pd.cut(nyc_mod_df['SCORE'], bins,labels=labels)
nyc_mod_df.head()
```

Out[13]:

	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION	...	INSPECTION TYPE
5	41582828	CIBO EXPRESS GOURMET MARKET/SALOTTO	Queens	0	LA GUARDIA AIRPORT	11369	3478675394	Sandwiches/Salads/Mixed Buffet	02/06/2020	Violations were cited in the following area(s).	...	Cycle Inspection / Initial Inspection
8	50068643	RISBO	Brooklyn	701	FLATBUSH AVENUE	11225	3472259115	American	03/21/2019	Violations were cited in the following area(s).	...	Cycle Inspection / Initial Inspection
9	41272855	BOBO RESTAURANT	Manhattan	181	WEST 10 STREET	10014	2124882626	French	07/18/2019	Violations were cited in the following area(s).	...	Cycle Inspection / Re-inspection
14	40394701	THE END ZONE BAR	Queens	14944	14 AVENUE	11357	7187469654	American	01/29/2019	Violations were cited in the following area(s).	...	Cycle Inspection / Re-inspection
17	50064444	818 FAST FOOD	Brooklyn	818	NOSTRAND AVENUE	11216	3474353450	Chicken	06/07/2017	Violations were cited in the following area(s).	...	Pre-permit (Operational) / Initial Inspection

Fig.4. Data frame after creating risk category column

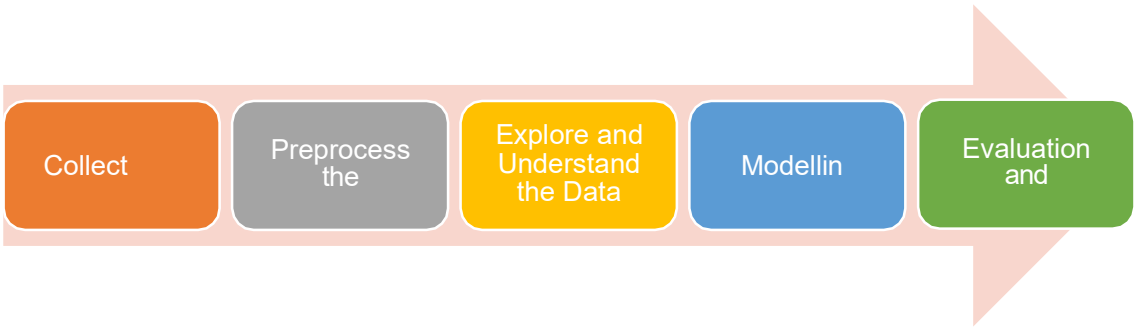
Now after further cleaning and some type conversion, the final dataframe named nyc_mod_df looks like this.

```
: nyc_mod_df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 188969 entries, 5 to 388684
Data columns (total 29 columns):
CAMIS                    188969 non-null int64
DBA                      188969 non-null object
BORO                     188969 non-null object
BUILDING                 188969 non-null object
STREET                   188969 non-null object
ZIPCODE                  188969 non-null int32
PHONE                    188969 non-null object
CUISINE DESCRIPTION      188969 non-null object
INSPECTION DATE          188969 non-null datetime64[ns]
ACTION                   188969 non-null object
VIOLATION CODE           188969 non-null object
VIOLATION DESCRIPTION    188969 non-null object
CRITICAL FLAG            188969 non-null object
SCORE                    188969 non-null float64
GRADE                    188969 non-null object
GRADE DATE               188969 non-null object
RECORD DATE              188969 non-null object
INSPECTION TYPE          188969 non-null object
Latitude                 188969 non-null float64
Longitude                188969 non-null float64
Community Board          188969 non-null float64
Council District         188969 non-null float64
Census Tract             188969 non-null float64
BIN                      188969 non-null float64
BBL                      188969 non-null float64
NTA                      188969 non-null object
Risk                     188969 non-null category
Inspection Year          188969 non-null int64
Day of Week              188969 non-null int64
```

Fig.5. Information about Processed dataframe

3.Methodoloy



3.1.1. Overall performance of Restaurants based on risk category for the period 2016-2020

As we have binned the restaurants that were inspected during the period of 2016-2020 into three categories based on their inspection scores it can be easily visualized the percentage of restaurants in each risk categories.

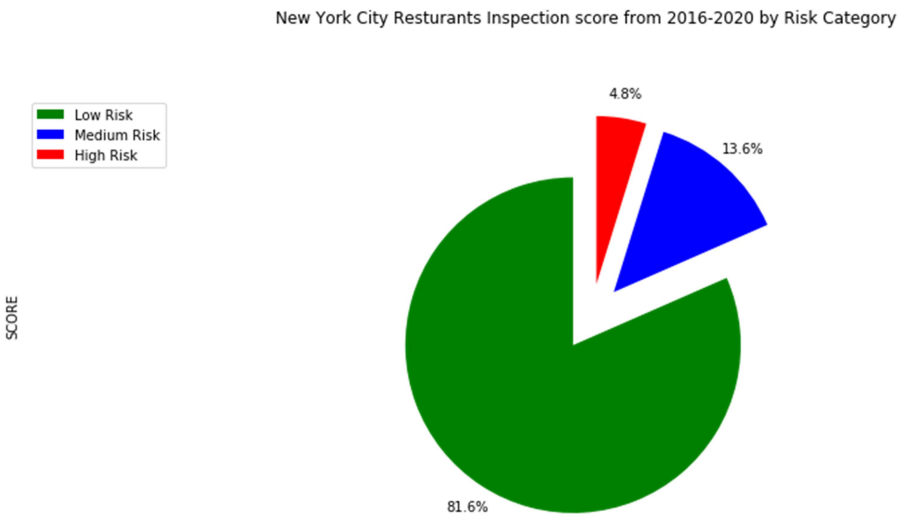


Fig.6. New York Restaurants Inspection Score from 2016-2020 by Risk Category

It can be seen from Fig.6. that almost 82% of restaurants that were inspected from 2016- 2020 have been placed under low risk category. This shows either the New York City officials are very lenient or the restaurants maintain a very high standard.

The count of restaurants in each category is given as a bar plot (Fig.7). It is inferred from the same that 154136 restaurants were classified as low risk, 25762 as medium risk and only 9071 were classified as high risk.

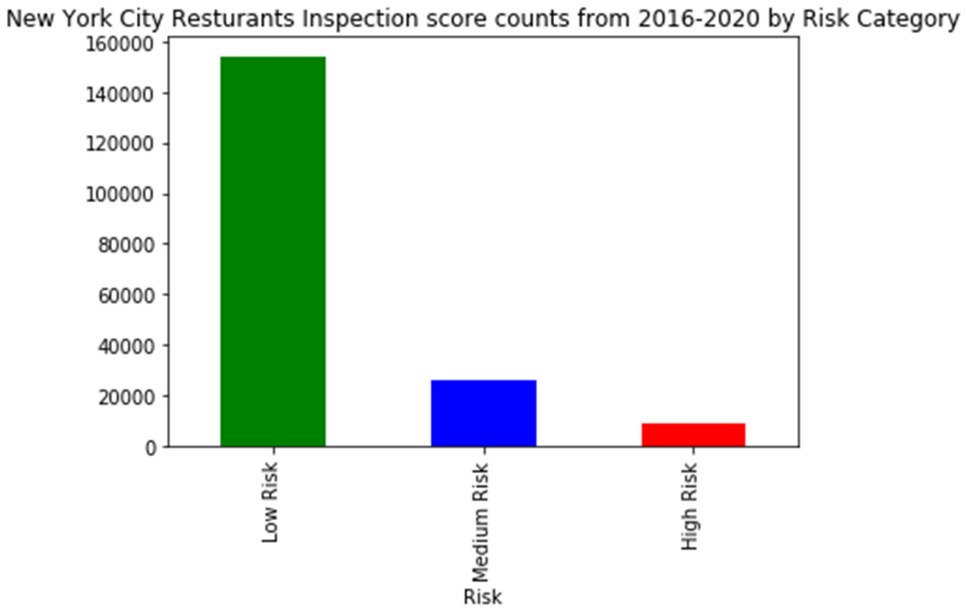


Fig.7. New York Restaurants Inspection Score counts from 2016-2020 by Risk Category

Since the overall statistics of restaurants that were placed in each risk category in the period 2016-2020 did convey some meaning, it will be more insightful if we gain some information on the percentage of restaurants that were places in each risk category for individual years. The percentage of restaurants in each risk category for individual years is shown in Fig.8.
It is understood from the same that almost 80% of the restaurants were placed in low risk category in each year and the restaurants have improved their standards since 2014 as the percentage of restaurants in high risk category have decreased over the years.

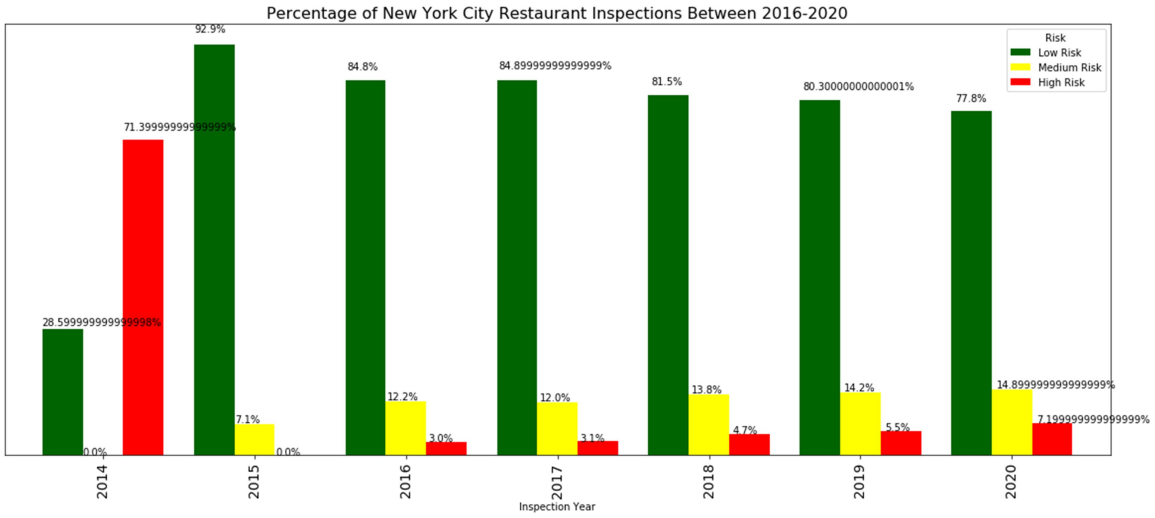


Fig.8. Percentage of New York Restaurants Inspections from 2016-202

The inspection score distribution of the restaurants for each year is shown in Fig.9. It was inferred from the same that most of the restaurants scored below 20 during the inspections in each year. Also the no inspected restaurants in 2014 and 2015 are very low.

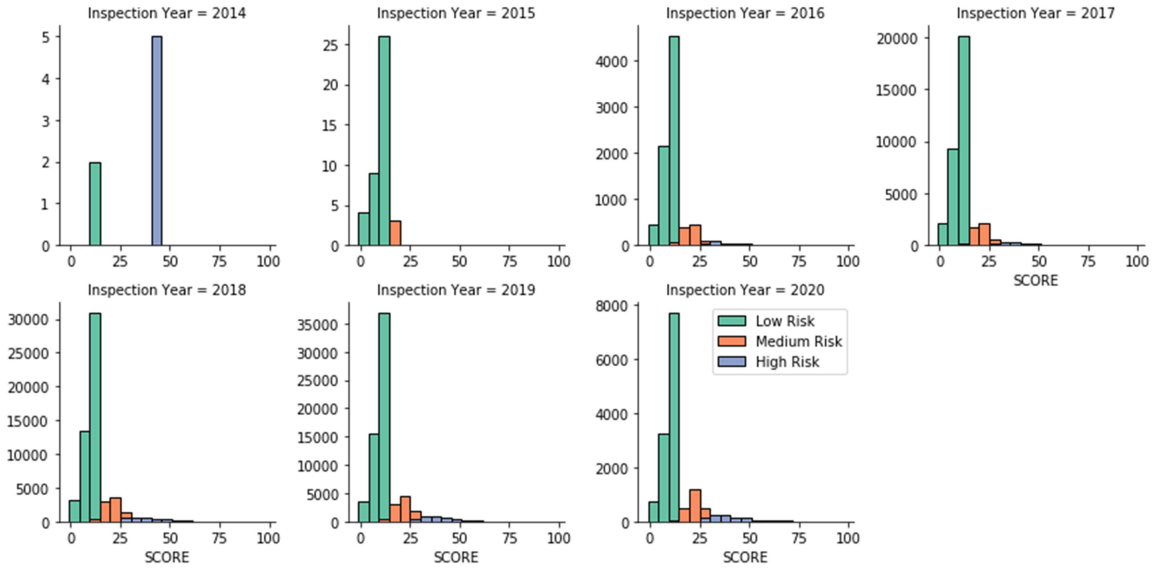


Fig.9. New York Restaurants inspection score distribution from 2016-2020

The restaurants will be better prepared if they know in prior on which day, they can expect an inspection. The day in which the restaurants were inspected in each year from 2014-2020 is depicted in Fig.10.

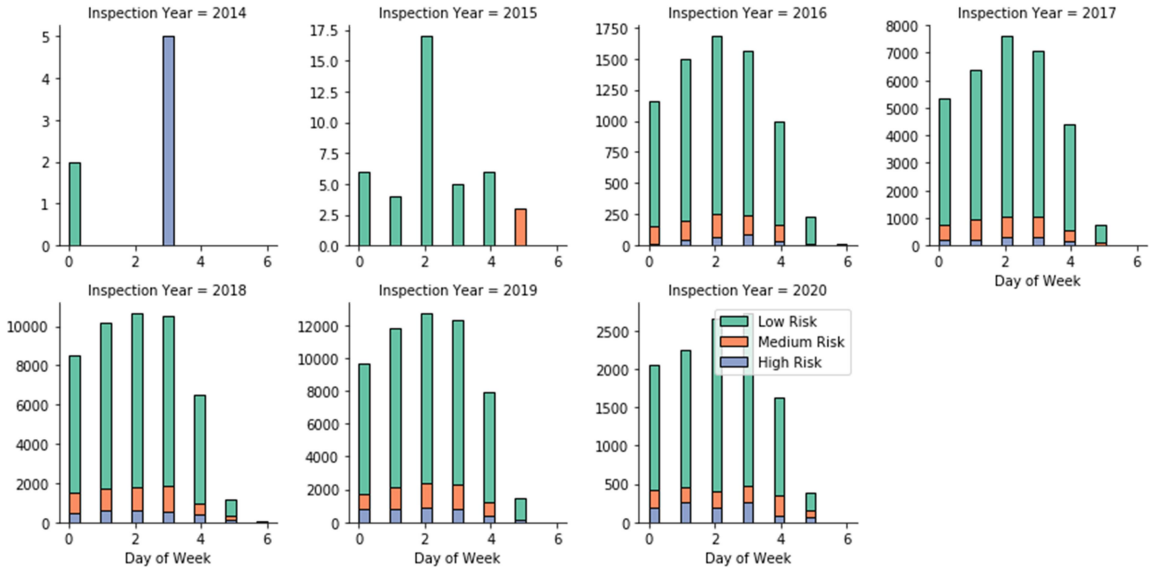


Fig.10. New York Restaurants inspection conducted days from 2014-2020

It is inferred from Fig.10. that most of the restaurants were inspected during the beginning of week.

3.2. Overall performance of Indian Restaurants based on risk category for the period 2014- 2020

As this project focuses on Indian restaurants in New York city, the whole data frame have to be trimmed to have only Indian restaurants. A simple google search revealed that most of the Indian restaurants in New York have the keyword such as India etc., in their name. Hence the data frame containing only Indian restaurants were created and named as Indian_df.

The inspection score received by the Indian restaurants in the same period are given in Fig.11. It is inferred from the same that during the year 2017 few restaurants were placed in High risk category.

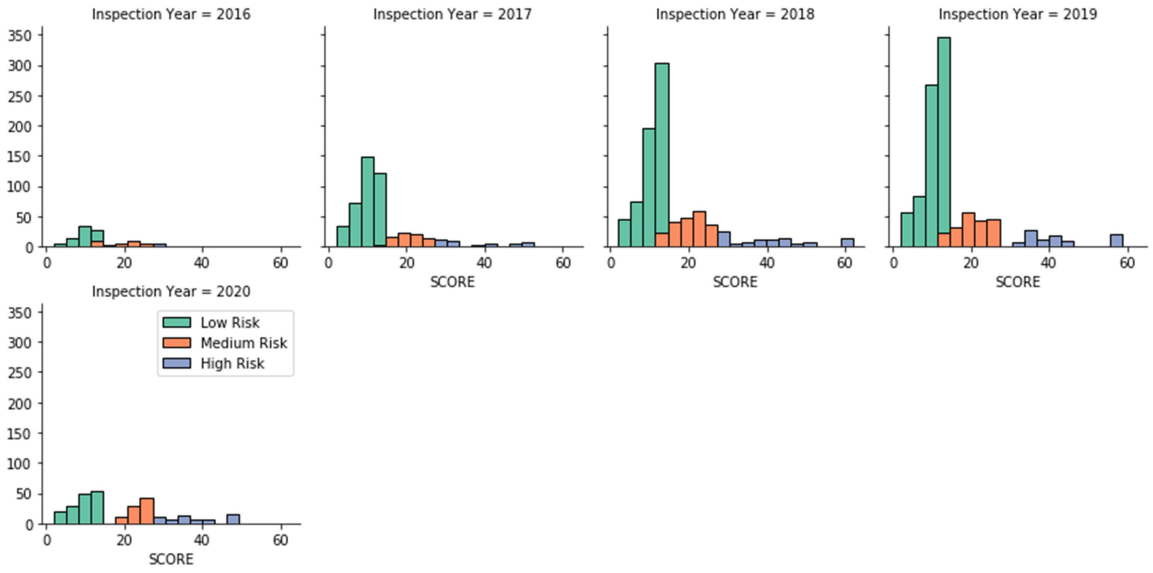


Fig.11. Indian Restaurants inspection score distribution from 2016-2020

The overall percentage of Indian restaurants in each category for the period 2016-2020 is shown in Fig. 12.

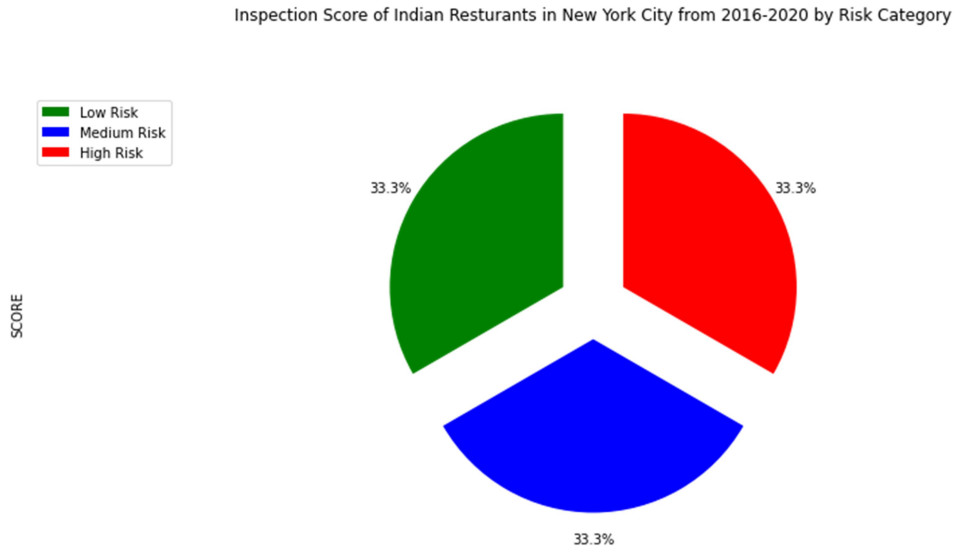


Fig.12. Indian Restaurants Inspection Score from 2016-2020 by Risk Category

The overall percentage of Indian restaurants that were placed in Low risk category is ~33.3%. This in comparison with the overall percentage of 82 is fairly low. Hence, the Indian restaurants must improve their service quality to match up to the overall performance.

Word Clouds (also known as wordle, word collage or tag cloud) are visual representations of words that give greater prominence to words that appear more frequently. This type of visualization highlights the most common words and present the data in a way that everyone can understand. The violation description of Indian restaurants from 2016-2020 in the form of a word cloud is shown in Fig.13

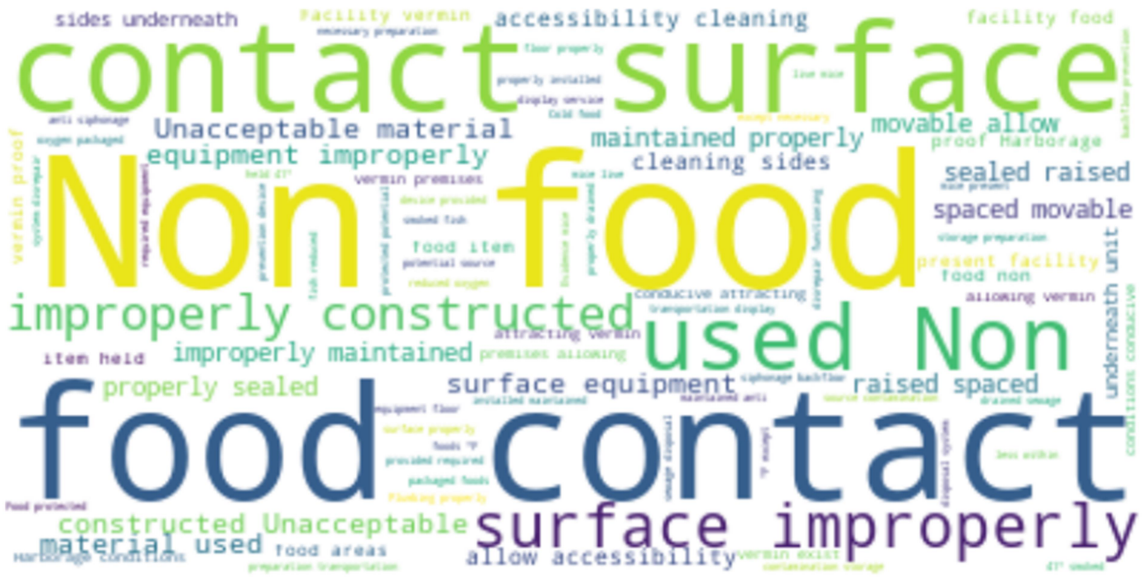
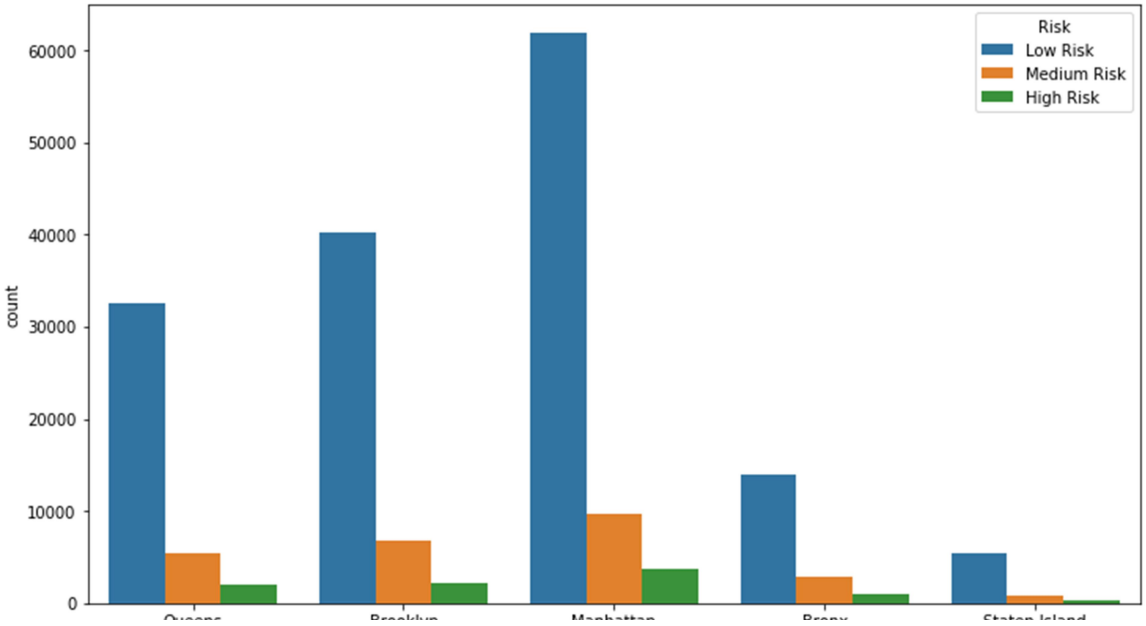


Fig.13. Word cloud of violation descriptions of Indian restaurants from 2016-2020

The word cloud shows that most violation descriptions have the words food, temp, food contact, improper surface appears more frequently. This may be due to improper hygienic conditions where food is kept.

The count of restaurants in each Borough of New York in each risk category is shown in the bar plot fig 14. Manhattan has the highest number of restaurants on low risk category which shows that the restaurants in Manhattan have high standards. Also Brooklyn has the highest restaurants in High risk category they need to improve their standards.

Fig.14. Count of restaurants in different borough depending upon risk category from 2016-2020



4. Predictive Modelling

There are two types of models, regression and classification, that can be used to predict the restaurant performance. In the present context, only classification models will be used as they predict the probabilities of the risk categories that the restaurants will be placed.

4.1. Classification models

Since the original data frame contains 188969 rows and 26 columns, it is not suitable for predictive modelling. After necessary label encoding and dropping unnecessary columns, the feature correlation between each column was analyzed. The corresponding pearson correlation plot is shown in Fig.15.

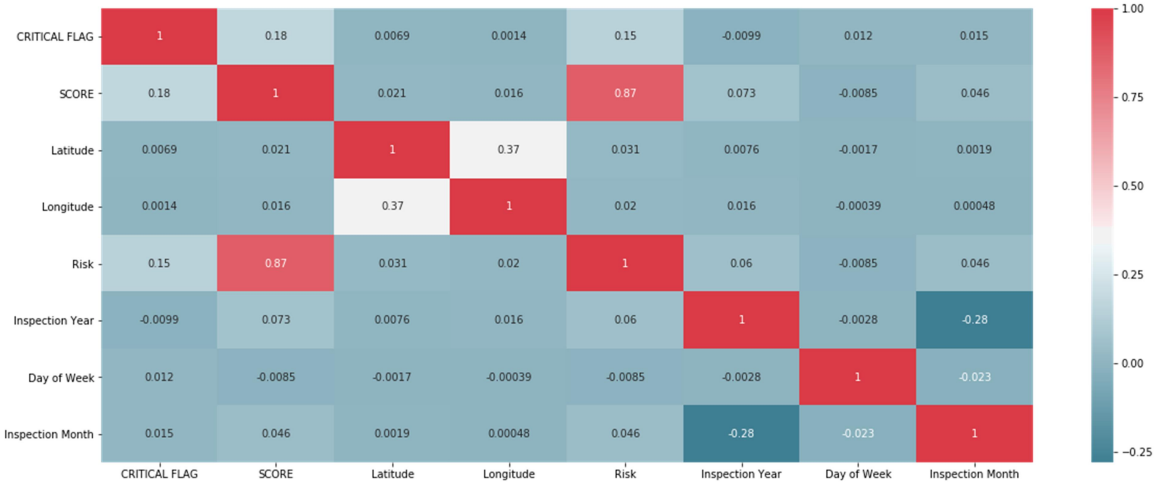


Fig.15. Person correlation plot for the features that will be used in the modelling.

It is inferred from Fig.15 that the risk is highly correlated with score.

Now the dataframe is ready to apply the machine learning algorithm. First, the Decision tree classifier was employed with minimal hyperparameters to predict the risk category and the decision tree classifier employed is shown in Fig. 16.

```
[ ]  rest_tree = DecisionTreeClassifier(criterion="entropy", max_depth = 10, max_leaf_nodes=5)
     rest_tree # it shows the default parameters
```

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=10,
                       max_features=None, max_leaf_nodes=5,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
```

Fig.16. Default parameters employed in the Decision Tree classifier model.

Then, a random forest model is employed to further improve the model and the default parameters employed are shown in Fig.17.

```
[ ]  # Build a random forest
     rf_tree = RandomForestClassifier(n_estimators = 1000, random_state = 42)
```

Fig.17. Default parameters employed in the Random Forest classifier model.

Usually, the decision tree classifier is prone to overfitting. Hence, the random forest model is employed here.

In order to evaluate the models employed, the accuracy scores were estimated. As inferred from the Table1. Both models have similar high accuracy scores and may be sufficient to predict the given data.

Table.1. Accuracy scores from the classification models.

Machine Learning Algorithm	Accuracy Score
Decision Tree Classifier	1.0
Random Forest Classifier	1.0

5. Visualization using Folium

In order, to visualize the Indian restaurants in New York city, folium package was used. For this purpose, Indian_df was used and the Indian restaurants are shown as markers in the Fig. 18.

Fig.18. Location of Indian restaurants in New York.

6. Ratings of Indian Restaurants using FourSquare API

In order to evaluate the ratings of Indian restaurants in New York City, FourSquare API was employed. For this purpose, the locations of Indian restaurants from Indian_df itself was employed to get the nearby venues using FourSquare API by employing getNearbyVenues() function. And from this the Indian restaurants were filtered to get their ID to get their ratings using venue_ratings() and get_venue_details() function.

```
[ ]  def get_Venue_Details(venue_id):
       ratings_list = []

       # Create the API request URL
       url = 'https://api.foursquare.com/v2/venues/{}?&client_id={}&client_secret={}&v={}'.format(
                 venue_id,
                 CLIENT_ID,
                 CLIENT_SECRET,
                 VERSION)
       # make the GET request
       results = requests.get(url).json()
       print(results)
       return(results)
```

Fig.19. get_Venue_Details() function.

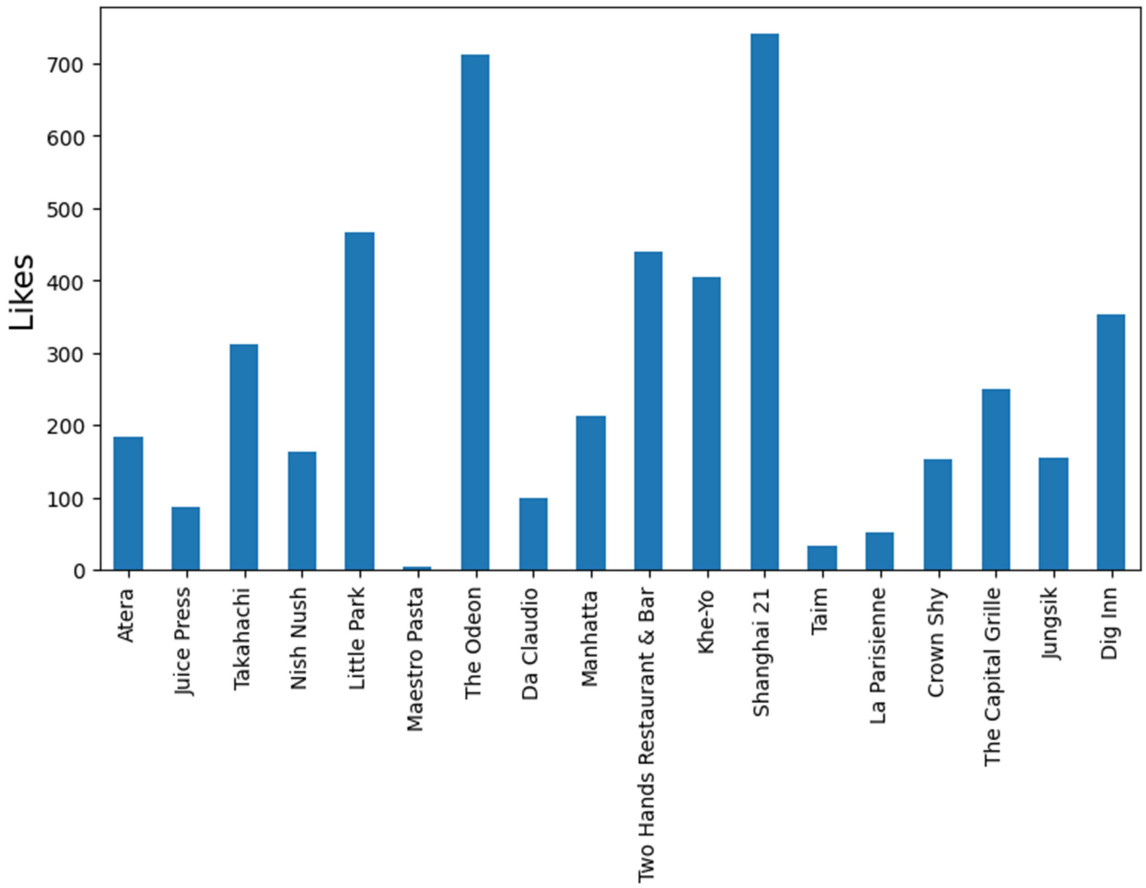
```
[ ]  ratings_list=[]
     def venue_ratings():
       for item in id_list:
         rating_details=get_Venue_Details(item)
         venue_data=rating_details['response']
         try:
             venue_id=venue_data['venue']['id']
             venue_name=venue_data['venue']['name']
             venue_likes=venue_data['venue']['likes']['count']
             venue_rating=venue_data['venue']['rating']
             venue_tips=venue_data['venue']['tips']['count']
             ratings_list.append([venue_id,venue_name,venue_likes,venue_rating,venue_tips])
         except KeyError:
           pass
         column_names=['Venue ID','Venue Name','Venue Likes','Venue Rating','Venue Tips']
         df = pd.DataFrame(ratings_list,columns=column_names)
       return(df)
```

Fig.19. venue_ratings() function

Since, the original Indian_df contains restaurants data from 2016-2019. They have duplicate entries. So using the unique IDs of the restaurants, the top five restaurants are sorted on the basis of the number of likes and other ranking values they received.

	Venue ID	Venue Name	Venue Likes	Venue Rating	Venue Tips
0	4f627061e4b05c1d57815977	Atera	183	8.9	44
1	54148bc6498ea7bb8c05b70a	Juice Press	86	9.2	12
2	4a8f2f39f964a520471420e3	Takahachi	311	8.9	80
3	564cb952498e133963c04186	Nish Nush	164	8.5	40
4	545c0436498e798e22ce4b2a	Little Park	467	8.4	108
5	5d4861420372ce0007e23375	Maestro Pasta	5	8.2	4

It can be seen from Fig. 20 that Shanghai 21 is the best Indian restaurant to dine in New York if you like Indian food. The number of likes received by the Indian restaurants in New York are shown as bar plot.


Likes
0
100
200
300
400
500
600
700
Atera
Juice Press
Takahachi
Nish Nush
Little Park
Maestro Pasta
The Odeon
Da Claudio
Manhatta
Two Hands Restaurant & Bar
Khe-Yo
Shanghai 21
Taim
La Parisienne
Crown Shy
The Capital Grille
Jungsik
Dig Inn

7. Conclusions

This project successfully completes my IBM Data Science Professional Certification Training. I am quite new to the data science and I had a steep learning curve during the course. I have really enjoyed doing all the lab exercise and the courses were really informative. The following are the conclusions that I derive from this project:

1. New York has numerous Indian Restaurants out of which we have analyzed a few of them. Thus opening a restaurant in New York will be facing a tough competition in the beginning.
2. Of the Indian restaurants that are currently present in New York which we analyzed roughly 80% are placed in low risk category based on the inspection data from 2016-2020.
3. A decision tree classifier model is built for classifying the restaurants into various risk categories and the model performs well for the given data set. This will help the restaurants in predicting their risk category.
4. The Indian restaurants in the New York were visulaized using the folium map rendering library.
5. Using FourSquare API, the venue details for the Indian restaurants were analyzed and found that among all the restaurants in Shanghai 21 is the best place to dine.

8. Limitations and Future Work

1. The restaurants are ranked solely on the data provided by FourSquare API. If data on other demographics are available this can be improved
2. The accuracy of location data depends on New York Inspection Data and FourSquare API. Hence, need to be analyzed further as there are some ambiguous entries.
3. The machine learning model will be further improved as the model developed may be prone to over-fitting