**NYC Restaurant Analysis**

**Final Report**

**Hasmitha Ganesh Babu**

**Submitted to: Professor Jennifer Shin**

**Date: 10th May 2022**

**Table of Content**

[**Background**](#_5pqbxdhufuot) **3**

[**Objective**](#_a4fz3fgtz8i1) **3**

[**Tools and Technology Used**](#_smfscnub2nd6) **4**

[**Data Science Processes Used**](#_9c7k2nudepq3) **4**

[**Exploratory Data Analysis**](#_z3e410twrz48) **6**

[**Further Analysis**](#_sn8kh7t9u00d) **9**

[**Conclusion**](#_sx6z6yqbutnt) **10**

[**References**](#_bh1wo7teyjs5) **11**

##### 

##### Background

New York City has one of the best food scenes in the world, with a diverse selection of cuisine and restaurant types. According to the New York Restaurant Industry Report, it has roughly about 23,650 restaurants listed in 2019. Because of that, OpenTable shares that it would take 22.7 years for someone to eat at every restaurant without going to the same place twice. With so many options to choose from it makes it very difficult to choose the restaurants. For this project, we decide to choose 2 main factors to determine the success of a restaurant that would highly impact our decision in choosing a restaurant to dine in. The first factor is the inspection score in the range of highest A, B, and lowest C, which reflects the restaurant’s cleanliness and sanitary level. The second factor is the customer ratings, which reflect customer satisfaction in the range of 1 to 5 stars.

This analysis would be useful for consumers looking to enjoy the restaurants in New York City, as well as, for new restaurant business owners who are looking for more insights into which locations and cuisines for high potential success in terms of inspection scores and customer ratings.

##### **Objective**

The goal of this dataset is to determine the grade associated with an establishment (restaurant) inspection, by filtering the zip code of the restaurant location and the cuisine. Another objective is to determine the highest user ratings for each borough and cuisine. The last objective is to determine the relationship between the inspection score and user ratings.

The intended audience is an ordinary New Yorker who is attempting to decide on an appropriate restaurant as well as any new restaurant business owners who are looking for more insights into which locations and cuisines for high potential success in terms of inspection scores and customer ratings. Following our study, they will have all the information and tools they need to make the best choice.

##### Tools and Technology Used

1. Python was used to clean and merge the two datasets.
2. Pandas and NumPy were used to create visualizations as a part of the supplementary analysis.
3. Excel was used to open and extract relevant data points within the dataset.
4. Tableau was used for data visualization.

##### Data Science Processes Used

We used the OSEMiN (Obtain, Scrub, Explore, Model, Interpret) process for this project.

1. **Obtaining the Data**

1. The Restaurant Inspection Data

* The inspection score is reflected by the inspection grade (A, B, and C) obtained from the DOHMH Restaurant Inspection Score on NYC Open Data.
* Provides information about restaurant inspections for all NYC restaurants.
* URL:<https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j>

2. Yelp Data

* The customer ratings (1 to 5 stars) are obtained from Yelp Dataset using Yelp API.
* This dataset contains information on the restaurant businesses like name, address, phone number, and the star ratings as well as reviews within the five boroughs.
* URL: <https://www.yelp.com/dataset>

1. **Data Transformation**

After obtaining the datasets from the two data sources, transformations needed to be done in order to merge the two datasets together.

There were six different steps that were performed to clean, transform the data and prepare it for analysis. They include:

1. Remove duplicate data - certain restaurants were listed in the dataset numerous times. Duplicative rows were deleted because the criteria were to have one entry per restaurant.
2. Remove unnecessary columns - We removed any columns that were not going to be important to our analysis (such as restaurant addresses, longitude, latitude...)
3. Feature Engineering - A few columns of data in the Yelp dataset (categories, coordinates, and location) were contained in a dictionary format. So, dummy variables were created to isolate these values and make them easier to deal with.

In addition, the frequency of each cuisine type was used to determine which cuisines are mainstream (featured at least 150 times in the dataset) and which are rare.

The Yelp dataset's PHONE column was modified to match the format of the Inspection dataset's display phone column so that this field could later be used to integrate and merge the datasets.

1. Working with categorical data - In order to execute the analysis, numerical data is more efficient to work with. So, dummy variables were utilized to convert the cuisine type, transactions, GRADE, neighborhood, critical flag, count range, and price value columns into numerical data columns.
2. Handling missing data - Any missing price values were set to zero to indicate that the price level is unknown. The columns ACTION, CRITICAL FLAG, INSPECTION TYPE, GRADE, SCORE, and GRADE DATE were altered so that any null values were replaced with values based on the dataset. The remaining rows with null values were eliminated.
3. Dealing with outliers - Restaurants with a significant number of reviews (more than 1,500) were changed to 1500 and those with a high number of inspections (more than 70) were changed to 70 to eliminate outliers.

##### Exploratory Data Analysis

For this analysis, the two factors: inspection grades and Yelp ratings are examined for different types of cuisine and boroughs throughout the city. Some of the questions that we would like to explore include:

1. Which cuisines have the highest user ratings and practice better sanitary, as reflected by higher inspection grades?
2. Which borough has the highest user ratings and practices better sanitary, as reflected by higher inspection grades?
3. What is the relationship between inspection scores and user ratings?

To answer the first question, we would like to explore which different types of cuisine receive the highest ratings and inspection grades, as shown in charts 1 and chart 2.

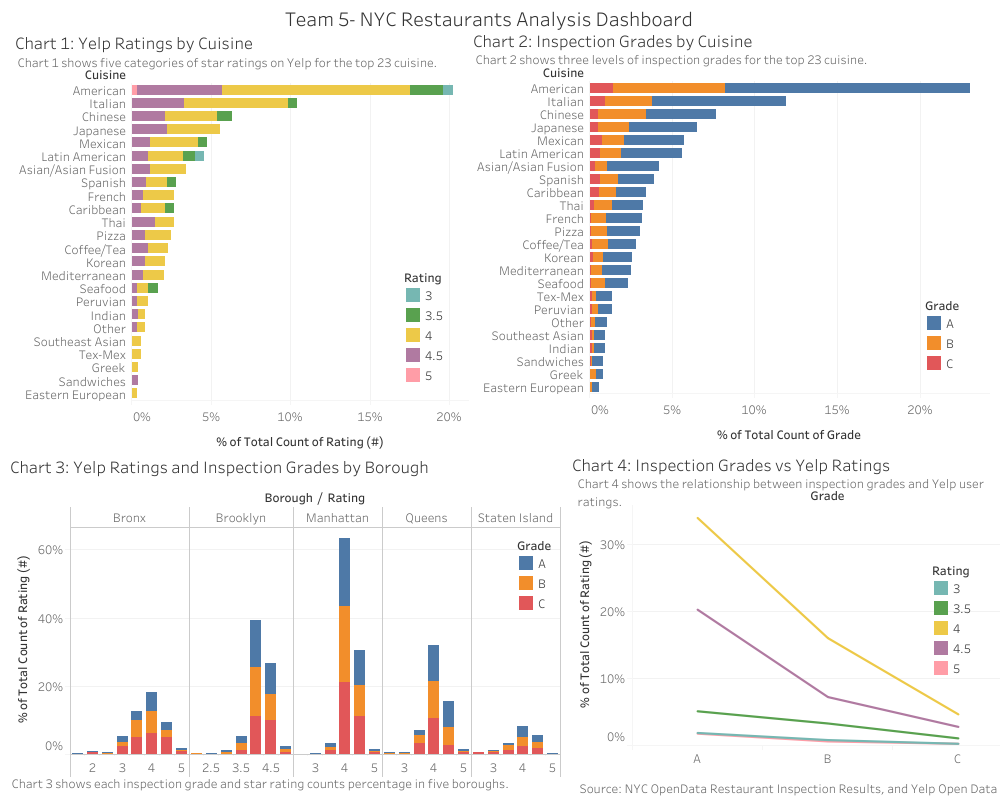


Chart 1 is a bar chart showing the percent of restaurants for certain cuisines that received a rating of 3 stars and above. Because there are many cuisines listed, we decided to show the top 23 that have the highest star rating percentage. From the chart, we can see the overall trend is that 4 stars dominate across cuisines, followed by 4.5 stars. Cuisine wise, American has the highest percentage of all-star ratings, followed by Italian cuisine, Chinese and Japanese. Interestingly enough, Chinese cuisine has more 3.5-star ratings than Italian, and Japanese has no star ratings below 4 stars.

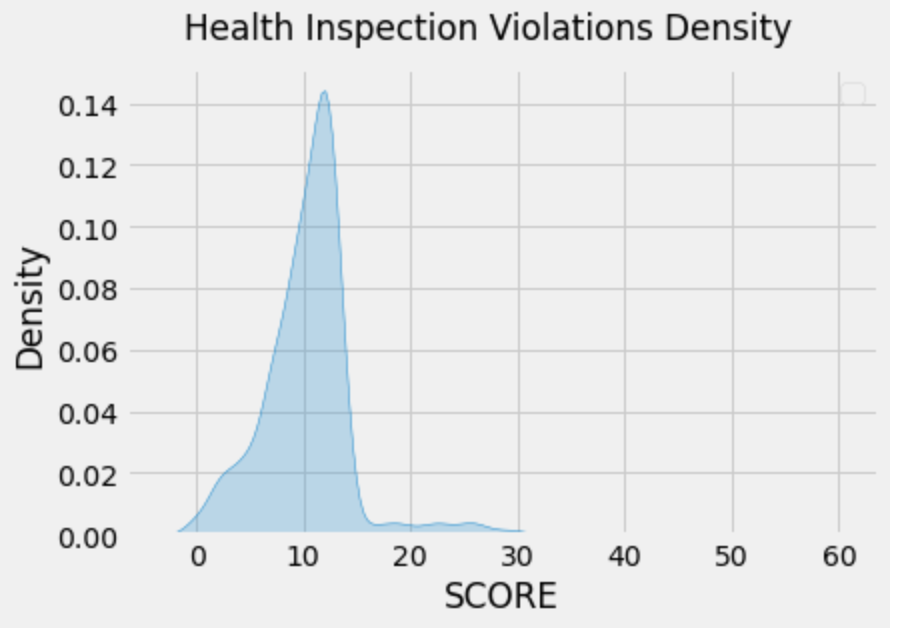
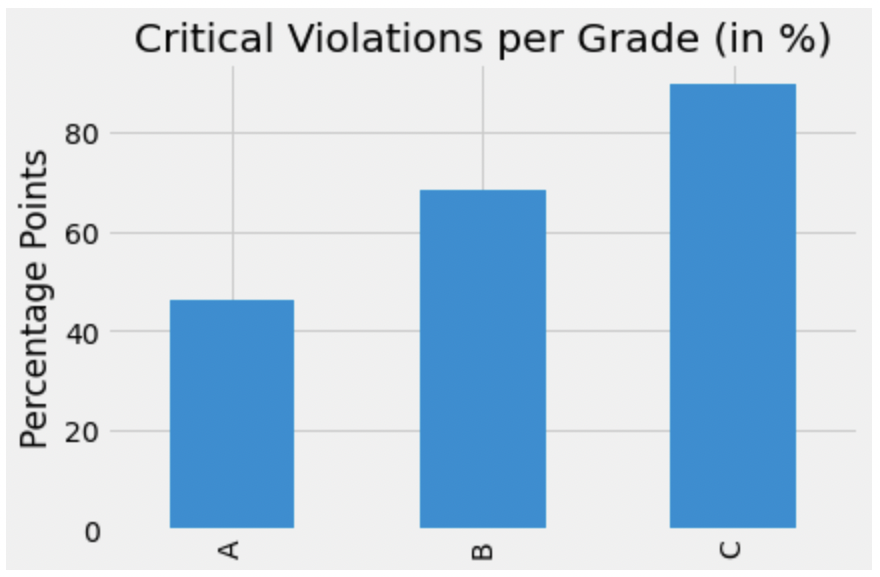
Similarly, chart 2 is a bar chart showing the percent of restaurants for certain cuisines that received inspection grades of A, B, or C. The overall trend is grade A dominates across the restaurant, which means that most of the restaurants in NYC across cuisines satisfy the sanitary inspection level. American cuisine has the highest number of A grades compared to others. This could mean that American restaurants tend to focus more on hygiene and cleanliness.

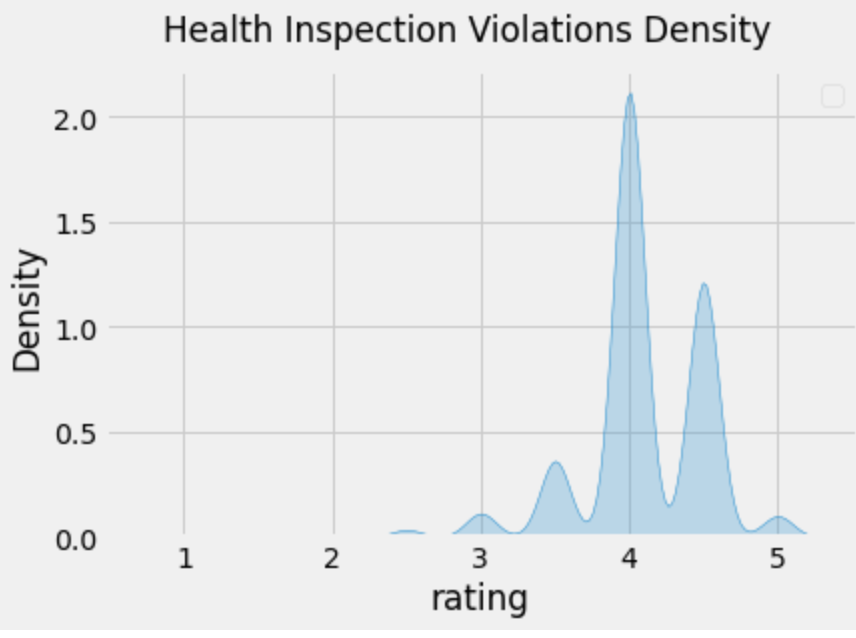
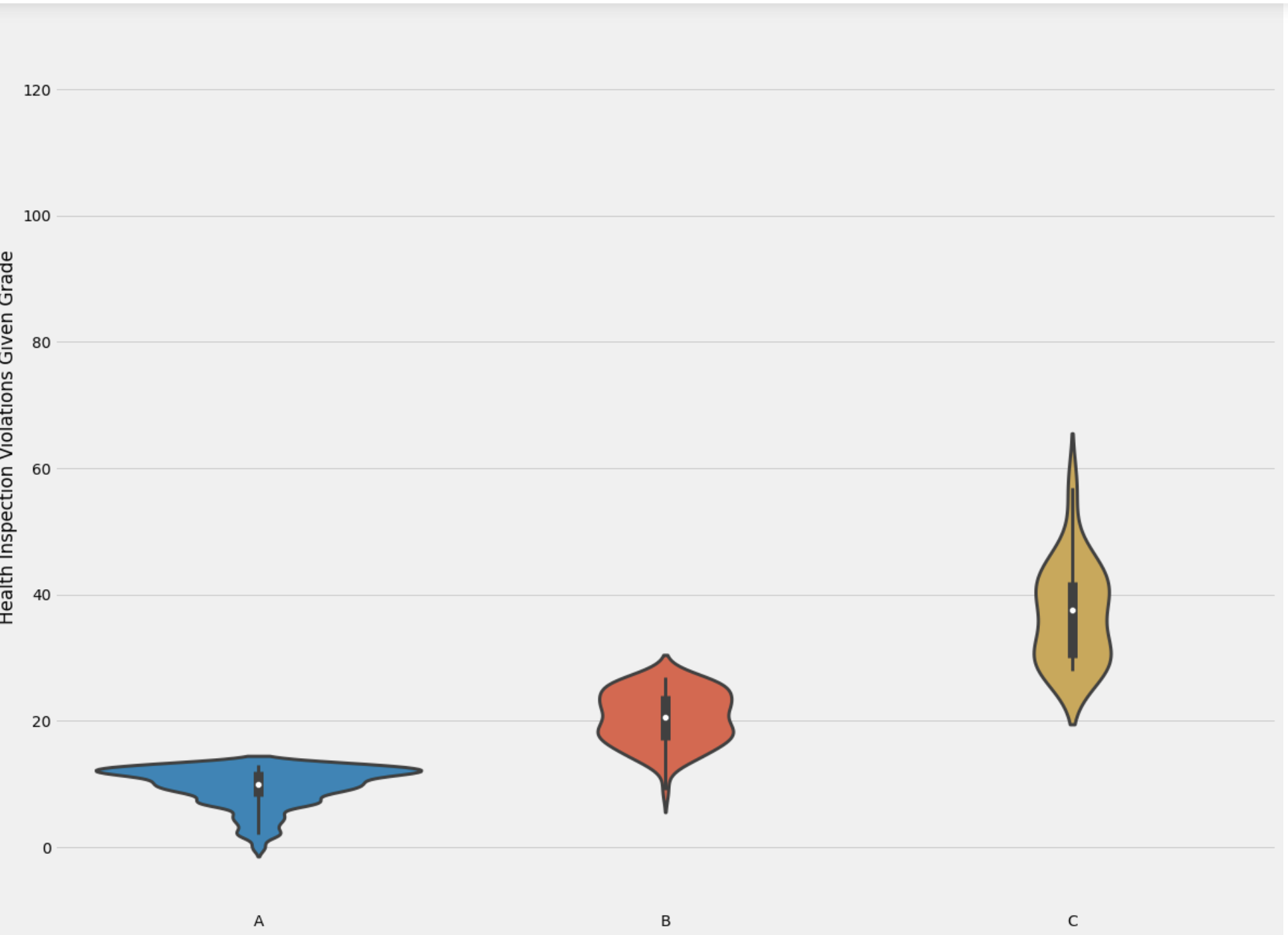
Combining charts 1 and 2, we can say that American, Italian, Chinese, and Japanese are the top 4 cuisines that receive the highest star ratings and highest inspection grades. This could mean that the number of restaurants serving these cuisines is way more than other types of restaurants serving other cuisines in the city.

To answer the second question, we would like to explore which borough receives the highest ratings and inspection grades, as shown in chart 3. Chart 3 is also a bar chart that highlights the two factors within the five boroughs. The plot indicates that most restaurants do achieve the top rating of 4 stars and an inspection grade of B across the five boroughs, with the top 3 being Manhattan, Brooklyn, and Queens. Manhattan has the highest number of restaurants with ratings of four stars while Staten Island has the lowest number of restaurants with high ratings. This could also mean there are a lot more restaurants in Manhattan than in Staten Island.

Lastly, to answer the last question, we investigate the relationship between inspection scores and user ratings, as shown in chart 4. We can see the trendline is clear as the inspection grade increases from C to A, and the user rating increases as well. There is a correlation between the two factors. This means that a restaurant with high sanitary and cleanliness tends to have higher ratings on Yelp as well.

##### **Further Analysis**



1. Critical Violations per grade chart shows that more than 45% of A-rated restaurants have one or more critical violations.
2. Health inspection violation density charts with scores show that almost all of the restaurants were scored below 30 and ratings show that almost all of the restaurants were rated above 3.
3. The last violin chart shows that most A-rated restaurants have scored less than 20, B-rated restaurants have scored around 20 and 30 and C-rated restaurants have scored more than 20.

Hence, we cannot blindly trust just one of the parameters but we have to look into multiple fields before making the decision.

##### **Conclusion**

By Cuisine, Americans, Italian, and Japanese are among the top 3 highest user ratings and Grade B Inspection Grade. By Borough, Manhattan, Brooklyn, and Queens are top 3 for Grade B and 4 stars. And Cuisine and Borough, for locations of the restaurants, have an effect on the inspection grade and user ratings. Inspection grade and user ratings are highly correlated to each other.

Our next steps could be looking into performing hypothesis testing with machine learning methods and exploring other factors such as review comments, and violation types with further analysis using natural language processing.

##### 

##### References

1. Lin, Hubert, and Ekim Sarinoglu. “| notebook.community.” *| notebook.community*, https://notebook.community/NYUDataBootcamp/Projects/UG\_F16/Lin-Sarinoglu-NYC%20Inspection%20Analysis. Accessed 9 May 2022.
2. Rspiro9. “GitHub - Rspiro9/NYC-Restaurant-Yelp-and-Inspection-Analysis: Used Hypothesis Testing to Gather Analytical Insights about Manhattan Yelp and Inspection Grade Data. Additionally, Used Machine Learning Classification Techniques Such as Logistic Regression, Decision Tree, Random Forest, XGBoost, and Adaboost, to Predict Yelp Ratings for Manhattan Restaurants.” *GitHub*, github.com/rspiro9/NYC-Restaurant-Yelp-and-Inspection-Analysis#nyc-restaurant-yelp-and-inspection-analysis. Accessed 10 May 2022.