Analysis of Restaurants and their Inspection grade in New York

I. Ibrahim Rinub Babu
x19207387
M.Sc in Data Analytics
National College of Ireland

II. Iswarya Yogeashwaran
x20155034
M.Sc in Data Analytics
National College of Ireland

III. Pranavi Pusapati
x20155301
M.Sc in Data Analytics
National College of Ireland

Abstract—In New York, fifty percentage of Funds are spent on meals made outside the house. The inspections of restaurants become a major thing of public health rules to protect people from the outbreak of food-borne sickness. Food grading entails inspecting, assessing, and filtering different foods based on quality, freshness, regulatory requirements, and selling price and restaurant features. The main objective of this report is to visualize the food quality and other features of the restaurants in New York that influence the inspection grade given by the food inspector. This report includes interpretation and visualization of various restaurants in New York, for analysis. This project involves the extraction of unstructured data through API, importing to and accessing from the database, and pre-processing into structured data and then interpretation and visualization.

Keywords- Inspection grade, Restaurant, Food, New York

I.INTRODUCTION

According to the survey of Baruch College at the City University of New York in 2012, the restaurant grading is being accepted by 91 percentage of New Yorkers, 81 percentage people decide their dining decisions by Letter grading and A-grade restaurants are being chosen confidently by 76 percentage people. The process of the project is extracting unstructured data through API and loading the raw data into a non-relational database. Then, the raw data is accessed from a non-relational database using python, which is then pre-processed by cleaning, transforming, imputation, and converting the raw data into a structured form by transforming it and pushing that into a relational database. The stored data in a relational database is then retrieved, for interpreting and visualizing to know the insights of the data, which helps for the analysis purpose. The Non-relational database used in this project is MongoDB Atlas on Amazon Web Service client, whereas PostgreSQL is used as a Relational database in this project, which is deployed on AWS RDS (Amazon Webservice-Relational Database Service). This project analysis is being conducted to know the following research questions:

1. Which seating arrangement outside the restaurant influences the grades?
2. Does the type of Alcohol permit influence the grades?
3. Did the restaurant's special offer such as kid meal, shareable facility, regional food, and less nutritional content influence the food inspection grading of the restaurants?

II. LITERATURE REVIEW

This research paper [4] is about a recommendation system that is used in the tripadvisor.com search data. The recommendation of restaurants to the users is depending upon the user's concern that applied to multiple fields such as the highest rating restaurants, places, other user's comments, and so on. In this recommendation system, the user chooses the features of the restaurant and it recommends the highest rating of the restaurant with the help of the user's comments. The user can select the basic hotel amenities by using a restaurant recommendation system, and the perfectly matched hotels will then be inhibited depending on this factor. This research concludes the recommendation system considers the hotel industry domain wherein the reviews of the restaurant's user's comment sentiments concern the hotel characteristics which gives the examination of user's feedback. The results show the recommendation system's higher accuracy. The author [3] here discusses Salmonella infection rates in the United States remained constant. Restaurants are the common reason for the outbreak of Salmonella. The author examined and proved the effect of before and after the restaurant inspection grade correlation on foodborne illness. The results of the restaurant inspection grade are in the form of Letters. After the implementation of grading letters, the rate of salmonella infection decreased. The author collected yearly laboratory-confirmed case counts from the NYC DOHMH and the New York State Department of Health for this study from 1994 to 2015. They computed the percentage growth from one year to the next. They used t-tests to compare mean rates of Salmonella infection in NYC and the rest of the state before and after the implementation of a point scoring system in 2005 and grade cards in 2010. (NYS). The autocorrelation and partial autocorrelation functions revealed that the outcome was not autocorrelated. They calculated incidence rate ratios (IRR) comparing Salmonella infections in New York City to those in New York State before and after the implementation of the point rating system and the uploading of letter grades. During the period 1994–2015, the annualized rate of Salmonella infections dropped significantly in both NYC and NYS. This author [10] proposes three enhancements to the standard UCF technique. The precision of the UCF methods was significantly lower since this customer's choice for a restaurant was adversely affected by many factors. Eventually, overall personal details of approved online users are used to calculate the similarity of user features. The results clearly show that the ACF-modified algorithm improves the accuracy of similarity computation, providing customers with a high precision restaurant recommendation.

This paper [13] presents a methodology for assessing the quality of hotel services in China. Initially, a questionnaire based on the HSQ-CS Model is created. Furthermore, AHP is used to determine the weight of each variable in the list of questions. A series of practical methods are used in data analysis with polling data to evaluate service quality enhancing the customer (CS). First, hotel service quality is assessed using the Customer Satisfaction Degree. Second, using discriminant analysis, correlation analysis, and other techniques, some informative outcomes were obtained.

III. TOOLS

A. Python

Python is an exquisite and flexible language with a rich modular and code library ecosystem. Python Work for ETL will commence with knowledge of appropriate frameworks and library systems such as workflow management tools, data access and extraction libraries, and fully functional ETL toolkits.

B. Mongo DB Atlas

In this project, we hosted MongoDB databases on AWS cloud clients and automate time-consuming administration tasks such as upgrades and backups. MongoDB Atlas is a highly scalable cloud database service built by the authorized MongoDB. Despite the MongoDB edition, Atlas uses the default Transport Layer Security (TLS) protocol version 1.2 to secure Atlas clusters.

Cloud client: MongoDB Atlas AWS cloud client

C. PostgreSQL

Amazon RDS facilitates PostgreSQL deployment in cloud installation, operation, and scaling. With Amazon RDS, cost-effective and redesignable hardware capability for scalable PostgreSQL deployments can be done within a few minutes. Installing and updating PostgreSQL software, storage management, replication for high accessibility and read-output, as well as disaster recovery backup, are all complex and time-consuming administrative tasks of Amazon RDS. Python is used throughout the project. This is because it is one of the best and efficient languages for data analysis.

Cloud client: PostgreSQL in AWS RDS instance

IV. METHODOLOGY

A. Description of Data

The dataset used in this project analysis is extracted through API by using SODA (Socrata Open Data API) programmatic access in the NYC Open Data website, which is approachable by the public.

DATASET-1: The Open restaurant application

This dataset contains open(street) restaurant applications in New York, which has a plan for seating arrangement in front of their business on the sidewalk or roadway. It also includes the details of alcohol availability in the restaurants and types of permission received for alcohol supply. This dataset includes 12,165 rows and 35 columns.

DATASET-2: Inspection Grade

This dataset contains the Inspection grade of the restaurants in New York. This dataset includes 87,883 rows and 9 columns which have latitude and longitude columns also for locating the place exactly.

DATASET-3: Menu information

This dataset contains the restaurant menu information of various states in New York. It provides information about the extra availabilities such as kid meals, regional food, and allowing shareability. This dataset includes 65,220 rows and 49 columns. The three datasets are in unstructured form, which is converted into structured form. Before that, multiple steps are followed for analysis to get insights about the structured dataset. The steps in this analysis are all explained below in Figure-1.

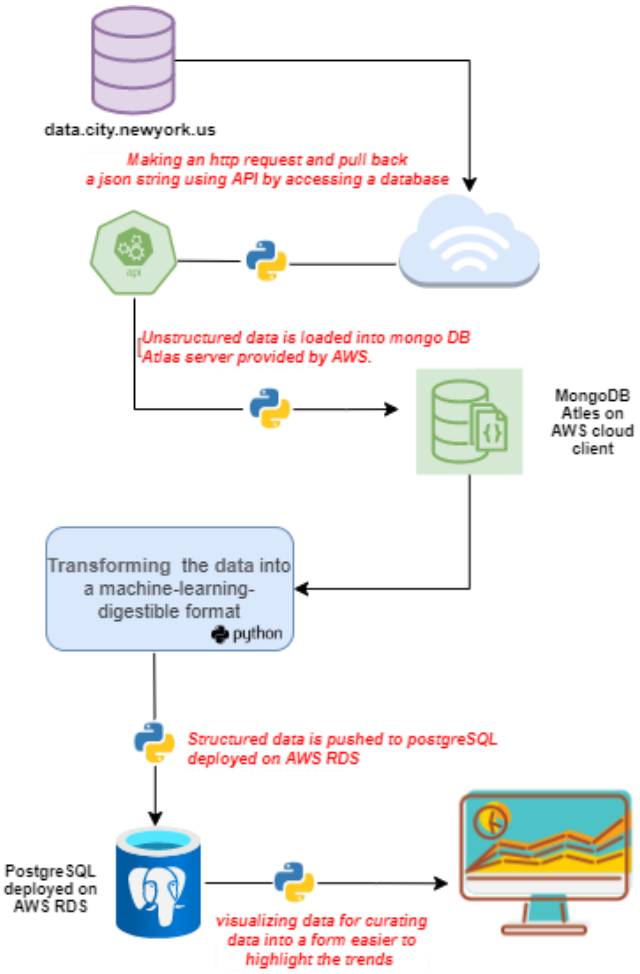

Fig. 1. PROJECT MANAGEMENT WORKFLOW

B. Data Extraction:

The three datasets are extracted from NYC Open data repository data.cityofnewyork.us in JSON format using SOCRATA API, which is also called SODA API. The Soapy package is programmatically used to extract datasets from the website using API keys. The extracted datasets are in JSON format.

C. Unstructured Data storage:

The unstructured data in JSON format is loaded into Mongo dB Atlas server (non-relational database) provided by Amazon Web Service (AWS). The Pymongo package is used to access Mongo DB. In Mongo dB Atlas, a new database is created to store all three datasets. These data can be obtained from the database and programmatically loading the dataset into a data frame.

D. Data Transformation:

The data transformation process is done by converting into a separate data frame by using the panda's package. The seaborn package is harnessed heatmap to visualize the null values in the raw datasets. The data in the data frame are transformed into structured form by filling or removing missing values, dropping the unnecessary columns, renaming the columns. The missing values are not removed in every case, because it would lead to the loss of other features. The missing numerical values are filled by using the KNN imputation algorithm. KNN (K- nearest neighbor imputation) uses the frequent value among the nearest neighbor in the column. Then, the missing values of categorical values are filled by using frequent categorical imputation. This imputation fills the NaN values with the most frequently occurred category in the column. Besides, mean imputation is utilized to fill the missing values in numerical columns in dataset 3. Moreover, the Pandas Series package is a one-dimensional labeled array. This package is applied in the transformation process for splitting a column, which is a dictionary and choosing only two key-value pairs for further process (that two values are only needed for analysis). The selected key-value pair in the array is converted into a separate data frame for adding that column into the main data frame, which is used to do the entire process.

DATASET-1:

In dataset 1, the data is first extracted from the non-relational database and the heatmap is composed to determine the missing values in the dataset. After identifying the missing values the unwanted columns are removed from the data frame. Some columns are renamed with appropriate names. All the continuous variables in the data frame are loaded to the new data frame and the correlation between them is identified. All the missing values in the data frame are imputed using the k-Nearest Neighbor Imputation method. After KNN imputation is accomplished for all the continuous values in the data frame. The output will be in the form of an array. Then the array of all the continuous variables are converted into a data frame. Missing values in the categorical variables are filled with the most frequent values. Both the continuous variable data frame and categorical variable data frame are conceded to gather to form the final transformed dataset. Final data is pushed to PostgreSQL with the AWS RDS cloud provider.

DATASET-2:

In dataset 2, the data is first extracted from the non-relational database and the heatmap is composed to determine the missing values. All the unwanted columns are removed. In the geocoded column, the latitude and longitude are stored together as a dictionary. It is separated into two different columns by using the pandas series module. It helped to split strings around the given separator or delimiter. All the continuous and categorical variables are loaded separately into two different data frames. Continuous variables are imputed using KNN imputation and the categorical variables are imputed using the most frequent values. The two data frames after imputation conceded together finally and pushed to RDS.

DATASET-3:

In dataset 3, all the unwanted variables are removed. The missing values are dropped using the thresh function. The thresh value is assigned as 11. After removing some null values using thresh function, the remaining null values in the continuous variables are filled by KNN imputation and some values are imputed by the random values of mean, mode, and median. The categorical variables are imputed by the most frequent values. Some categorical variables consist of values 0's and 1's. These values are replaced with the 'Yes' and 'No' values. And to make the inner joint with dataset-2 to achieve resultant data 2, the unique values of the 'restaurant name ' column are listed. Since there were minor changes in the strings of restaurant names between the two datasets. The names of some restaurant are corrected programmatically and the data is pushed to RDS

RESULTANT DATASET 1 :

After pulling two datasets parallelly from the database on AWS RDS it is loaded separately in two data frames. All the variables which are not essential for merging and achieving the final interpretation are removed from both the data frame. Some of the columns are renamed to performed the inner joint of two datasets. From the total datasets, multiple samples are taken by performing the inner joint of different variables. For each sample, the duplicate values are removed from each sample taken. All the samples are conceded together the resultant dataset is created.

RESULTANT DATASET 2 :

After pulling the dataset 2 and 3 from the AWS RDS PostgreSQL, It is loaded separately in two data frame. Some column names are changed in both the data frames to similar names to perform inner joint. The inner joint is performed using restaurant names of both the restaurant. The inner joint is performed by using the merge function. The merged data frame is examined for repeated values. All the duplicated values are dropped from the merged data frame. Some columns were converted to float values since it was in object type and all the columns are rearranged. An additional column called 'hotel Id' is generated to perform visualization effectively.

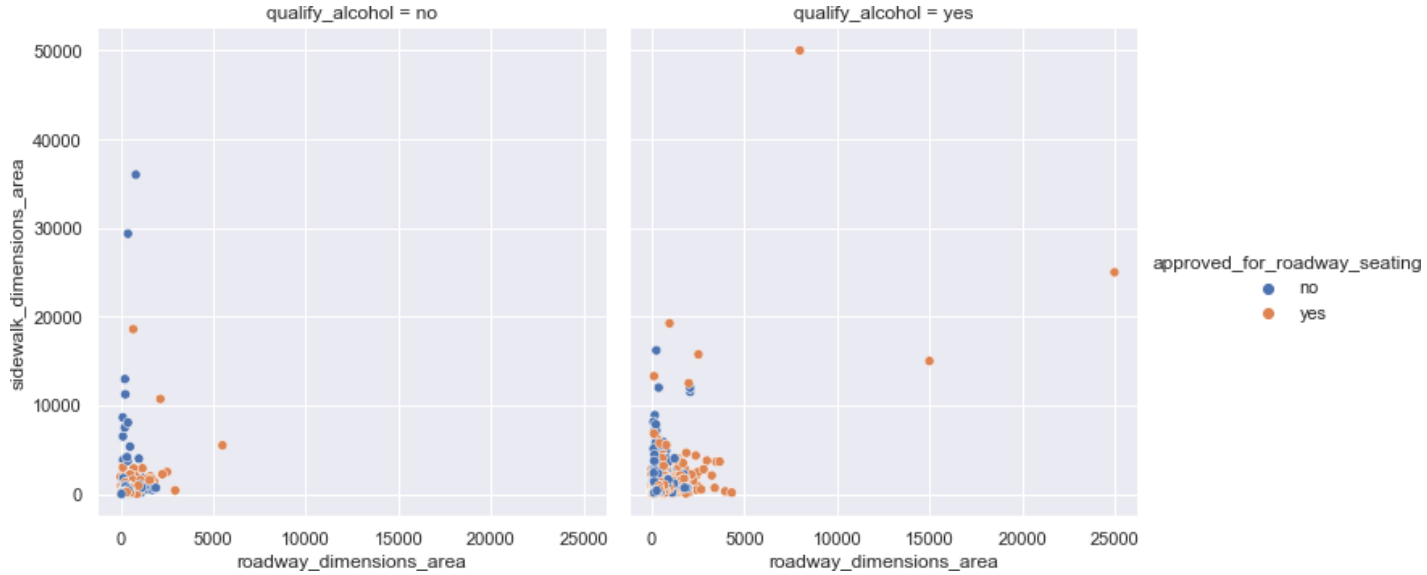

Fig. 2. Distribution of alcohol availability and the area of road-walk accepted restaurants

E. Structured Data Storage

The Structured Data is saved in .csv format is loaded into Amazon RDS (Relational Database Service) of Postgre SQL(Relational database). The connection is established with the help of the psycopg2 package.

F. Data Concatenation

The dataset1 (The street restaurant) and dataset2 (Inspectiongrade for the restaurants) are pulled from Postgre SQL. Three samplings are taken from merging dataset1 and dataset 2. The first sampling is taken inner joint by having common columns of zip-code and latitude. The second sampling is taken inner joint for the column legal business name, which is the common column for both the dataset. The third sampling is taken inner joint for latitude, which is the common column for both the dataset. The duplicate values are checked after taking every sampling and removed them. The excess unwanted columns are removed by using the drop function. The three samplings stored in three different variables are concatenated into a single variable to analyze and to get useful insights. The dataset2(Inspection grade for the restaurants) and dataset3(The menu information) are pulled from the database Postgre SQL. The unique values are checked in both dataset 3 and dataset2. The trade name column in dataset 2 is renamed into the restaurant. The first sampling is taken inner joint by having a common column of the restaurant. Then, unique values are checked in the first sampling. The unwanted columns are removed and again checked duplicate values and removed. Since unique values are low in number, the duplicate values are being examined and removed. The average is taken for all the food categories ' calories, total fat, saturated fat, sodium, carbohydrates, protein, dietary fiber for each restaurant, and it is grouped by restaurant column.

G. Data Visualization

The data is pulled from PostgreSQL and saved in Pandas Dataframe for visualization. The visualization is done with the help of the Plot Panda Package and seaborn Package.

The Initial Visualization is done for all three datasets, whichare given below.

DATASET-1

In dataset1, the analysis has taken place using the seaborn package, replot is plotted with x-axis as roadway dimensions area and y-axis as sidewalk dimensions area in the figure-2. This distribution of replotting shows that the restaurants with alcohol availability have larger roadway dimensional areas, which is approved for roadway seating for their business. In the figure-3, it describes that the alcohol available restaurants, which is approved for sidewalk seating have larger sideway dimensional area.

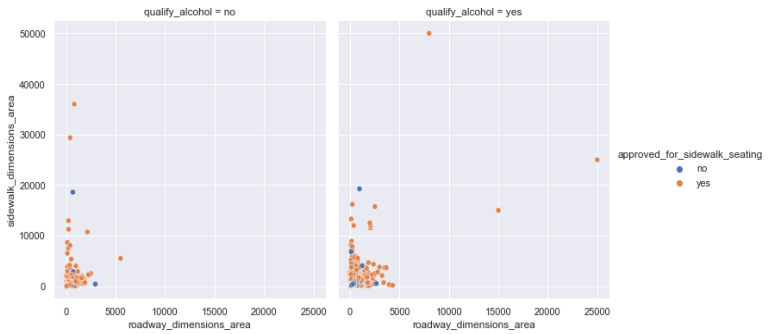

Fig. 3. Distribution of alcohol availability and the area of side-walk accepted Restaurants

DATASET-2

In dataset 2, the pie chart as a circle represents the percentage of inspection grade over all the counties in New York, which can be seen in the figure-4. A-graded restaurants are about 22 percentage all over New York, where B and C grades are around 20.3 percentage and 57.7 percentage.

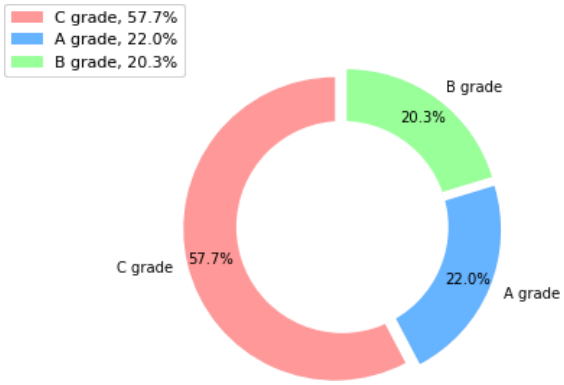

Fig. 4. Pie-chart of inspection grade over NewYork in Percentage

In the figure-5, inspection grades of the restaurant's percentage over all the counties are analyzed in a pie chart. The C-graded restaurants are more in Kings, Bronx counties.

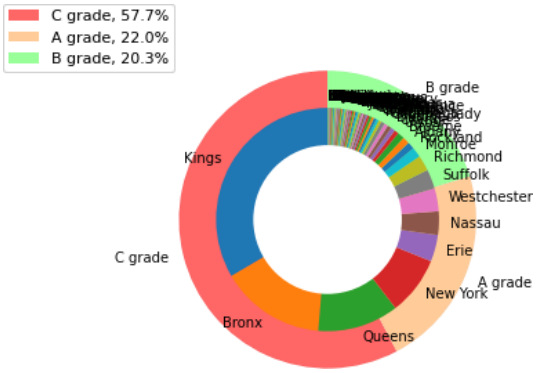

Fig. 5. Pie-chart of inspection grade over all the counties in Percentage

H. MERGED RESULTANT DATASETS

A. RESULTANT DATASET-1
(STREET RESTAURANT AND INSPECTION GRADE)

In this analysis, the figure-6 shows the sidewalk dimensional area of the restaurants over all the counties according to their inspection grade in cat-plot. Queens county and Kings county have a higher number of A-graded restaurants, which depends on their sidewalk dimensional area. It explains that the sidewalk dimensional area is lesser for A-graded restaurants.

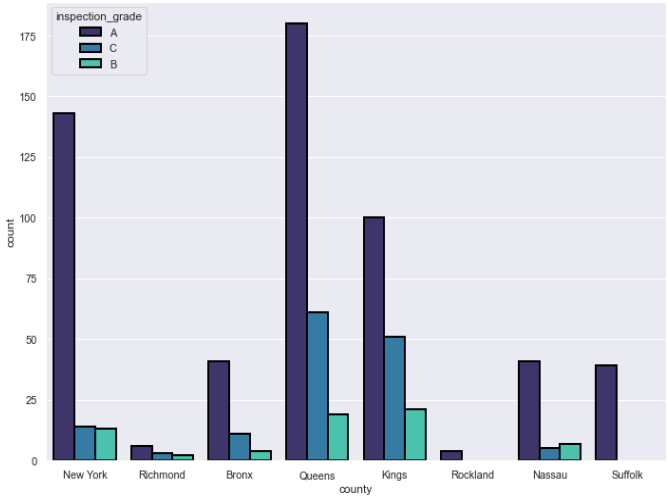

Fig. 8. count-plot for finding inspection grade count all over New York

In the figure-8, the count-plot is implemented to find the inspection grade count over all the counties of New York. Queens is the county, which has the highest number of A-graded restaurants in New York, while Suffolk and Rockland are the two counties that have only A-graded restaurants and no other graded. The count-plot shows that the inspection grade does not base on the alcohol permit in the figure-9. Both A and C-graded restaurants have a higher number of alcohol permit restaurants. This figure-10 shows the influence of alcohol permit license providers both the highest A graded and C gradedrestaurants are OP type alcohol permit.

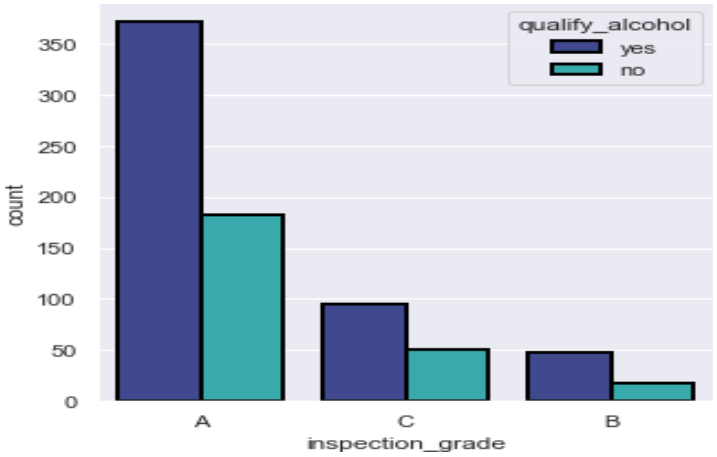

Fig. 9. Alcohol Permitted Restaurants and their grades

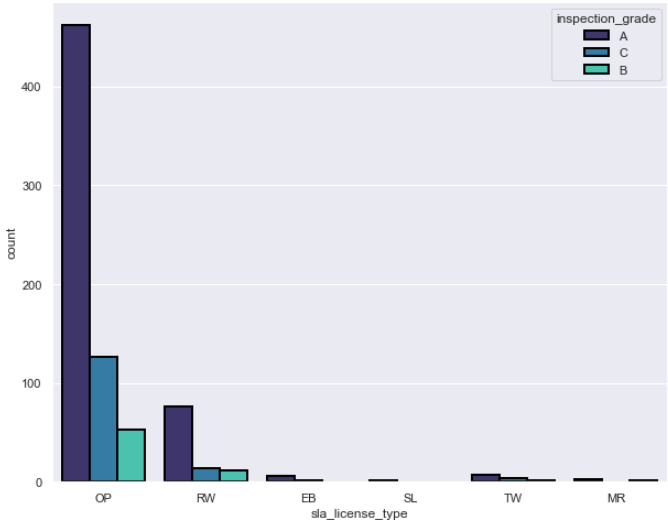

Fig. 10. Types of alcohol permit license providers and their grades in restaurants.

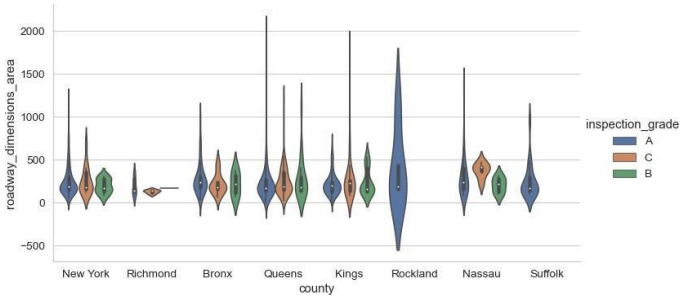

Fig. 11. Size of the road dimensions over all-county differentiated by grades

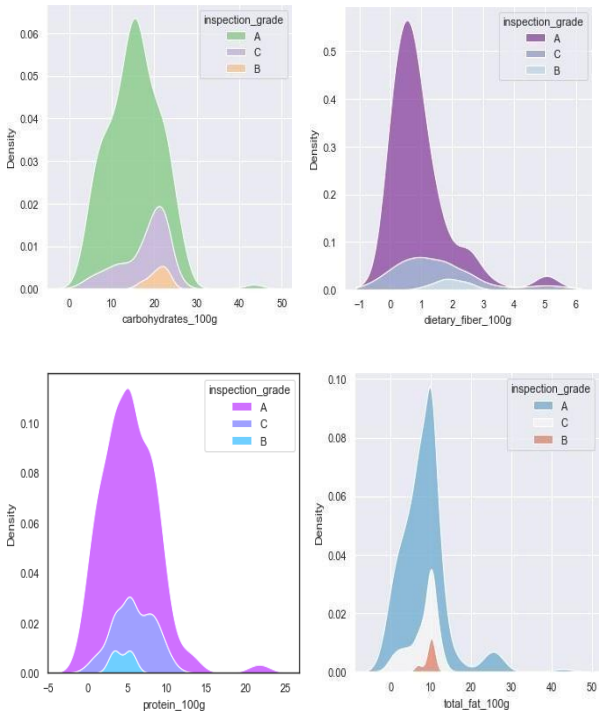

Fig. 12. Distribution of nutrients and their influence over restaurant grades

The road dimension over the different counties is explained and their influence over grades is explained clearly over figure-11.

B. RESULTANT DATASET-2 (INSPECTION GRADE AND MENU INFORMATION)

The analysis has been made between the hotel grades of every restaurant and the nutrients such as carbohydrates 100g, dietary fiber 100g, calories, total fat 100g, protein 100g in the figure-11. This clearly describes the distribution density of the nutrition and their influence on the restaurant grades. All the diagrams make us conclude that the A graded hotel shows the vast distribution of nutrition when compared to hotels that are graded B and C.

The distribution of carbohydrates 100g represents that the inspection grade of A given to restaurants is between the average range of 0 to 25, according to the hotel id. The inspection grades B and C are also in the range of 20 to 25. This means the inspection grade is correlated with the lesser carbohydrate foods of the restaurants in the figure-13. The relplot is utilized for this analysis, with the help of the seaborn package.

The correlation is examined between Carbohydrates 100g and Dietary fiber 100g. It gives an r value equal to 0.645978 and a p-value equal to 0.000000, which means the correlation is strong. This analysis shows the carbohydrate and dietary fiber are linearly correlated to each other in every restaurant in New York in the figure-14.

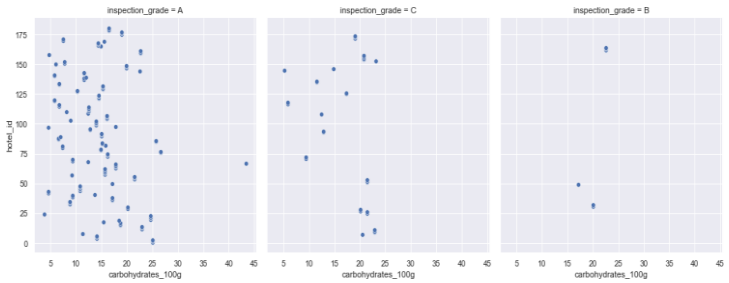

Fig. 13. Distribution of nutrients and their influence over restaurant grades

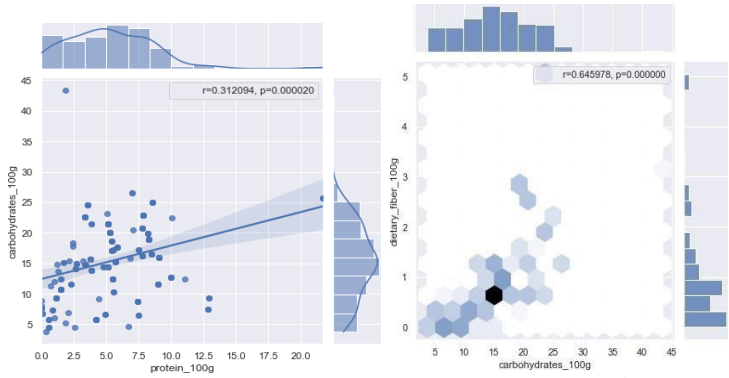

Fig. 14. Correlation between carbohydrates, proteins, and dietary fibers

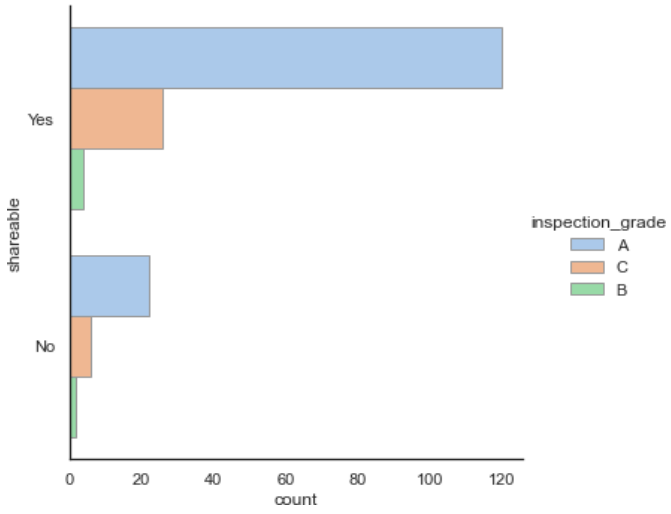

Fig. 15. Distribution of Shareable option in the restaurant

V. FUTURE WORKS

Further, We are going to develop a secured database by using Amazon Elastic Computed Cloud (EC2) platform. The AWS RDS is launched into the virtual private cloud which enables us to choose our IP address range, create subnets, and configure routing and access control lists. This helps to secure the datasets in the database. The interpretation of grades is done by fitting the dataset to certain machine learning algorithms and developed into an application to help all the restaurants to examine their grades before attending the inspection.

V. CONCLUSION

From the research conducted on the three datasets in this project, it is evident that restaurants with the highest nutrient content such as carbohydrate, protein, dietary_fiber, calories, etc., alcohol permission, and adequate road dimensional area are inspected and graded as A-graded restaurants. It is found from the study that Queens county has the highest road dimensional area all over the counties in New york. Further analysis, the OP type alcohol permit influences the A-grade of all the restaurants of New York. Then, the final analysis on the extra offering of kid meal, shareable facility, and less calorie content of food categories influences inspection result on A- grade, which is abruptly proved.

PROJECT RELATED LINKS

GITHUB LINK:

https://github.com/Rinub/DAP_project.git

Note: Run the MASTER_DO_NOTEBOOK.ipynb file to run all the files parallelly and see all the results and visualization. This handle separate files of an overall task

REFERENCES

[1] Barsky, J.D. and Labagh, R., ”A strategy for customer satisfaction”, The Cornell Hotel and Restaurant Administration Quarterly, October, pp. 32-40, 1992.

[2] Ditdit Nugeraha Utama, Luqman Isyraqi Lazuardi, Heresy Ayu Qadrya, Bella Marisela Caroline, Tris Renanda, Atthiya Prima Sari, “Worth Eat: an Intelligent Application for Restaurant Recommendation based on Customer Preference” IEEE, Fifth International Conference on Information and Communication Technology, 2017.

[3] Firestone, M. J., & Hedberg, C. W. (2018). Restaurant Inspection Letter Grades and Salmonella Infections, New York, New York, USA. Emerging Infectious Diseases, 24(12), 2164–2168. doi:10.3201/eid2412.180544.

[4] F.M Takbir Hossain, Ismail Hossain, Samia Nawshin, “Machine Learning-Based Class Level Prediction of Restaurant Reviews”, IEEE, pp. 420-423, 2017.

[5] Gomathi, R. M., Ajitha, P., Krishna, G. H. S., & Pranay, I. H. (2019). Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities. 2019 International Conference on Computational Intelligence in Data Science (ICCIDS). doi:10.1109/iccids.2019.8862048.

[6] Ginger Zhe Jin, Phillip Leslie.The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards. The Quarterly Journal of Economics, Volume 118, Issue 2, May 2003, Pages 409–451.

[7] Khushbu Jalan, Kiran Gawande, “Context-Aware Hotel Recommendation System based on Hybrid Approach to Mitigate ColdStart-Problem”, IEEE, Communication, Data Analytics and Soft Computing, pp. 2364- 2369, 2017.

[8] Lam, T., Mok, C. and Wong, L., “Customer satisfaction v. customer retention’”, Asian Hotel and Catering Times, August, pp. 34-6, 1996.

[9] Lewis R.C., “The measurement of gaps in the quality of hotel services”, International Journal of Hospitality Management, Vol. 6 Issue. 2, pp. 83-88, 1987.

[10] Ling Li, Ya, Zhou, Han Xiong, Cailin Hu, Xiafei Wei,

[11] 3+ “Collaborative Filtering based on User Attributes and User Ratings for Restaurant Recommendation”, IEEE, pp. 2592- 2596, 2017.

[12] Marit G. Gundersen, Morten Heide, and ULF H. Olsson, “Hotel Guest Satisfaction among Business Travelers”, Cornell Hotel and Restaurant Administration Quarterly, Aprial, pp. 72-81, 1996.

[13] Nelson Tsang and Hailin Qu, “Service quality in China’s hotel industry: a perspective from tourists and hotel managers”, International Journal of Contemporary Hospitality Management, Vol. 12, No. 5, pp. 316-326, 2000.

[14] Shi, J.,& Su, Q. (2007). Evaluation of Hotel Service Quality Based on Customer Satisfaction. 2007 International Conference on Service Systems and Service Management. doi:10.1109/icsssm.2007.4280099

[15] Zhu Hang and Wang Chunxiao, “Analysis on Key Variables of Hotel Service Quality Management”, Systems Engineering-Theory Methodology Applications, Vol. 8, No.1, pp. 60-66, 1999.