A SYSTEMATIC ANALYTICAL STUDY OF RESTAURANTS IN NEW YORK CITY

Masters in Science in Data Analytics (MSCDAD_B)
Database and Analytics Programming

RITIK VERMA
Database and Analytics Programming
National College of Ireland
x20165374@student.ncirl.ie

SILAMBARASAN PRABAKARAN
Database and Analytics Programming

SOURAV RAMALINGAN
Database and Analytics Programming
National College of Ireland
x20199911@student.ncirl.ie

KISHORE LAKSHMANAN
Database and Analytics Programming
National College of Ireland
x20253583@student.ncirl.ie

I. Abstract

The complicated architecture of restaurants, with several operations and a wide range of permissions and approvals, necessitates a comprehensive examination. In this study, we fetched four distinct unstructured datasets from State of New York | Open Data | State of New York (ny.gov) and NYC Open Data - (cityofnewyork.us), in various formats and successfully inserted them into Python to operate on a non-relational database, MongoDB on Amazon Web Service client, and relational database, PostgreSQL is used which is deployed on AWS RDS (Amazon Webservice Relational Database Service). We saved the datasets in MongoDB in XML and JSON forms and returned them to Python as structured datasets. We later pre-processed the retrieved data and put it in Postgre SQL. We pulled the data from PostgreSQL using select query and returned it to Python as a merged dataset containing all of the columns from the four distinct datasets. Later, we utilized numerous Python commands to visualize the data and explain its findings.

***Index Terms*—NYC, Restaurants, Liquor grant, SODA API, MongoDB, Python, Postgre SQL.**

II. Introduction

A. Motivation

Some of the world's most recognized restaurants may be found in New York. The greatest restaurants in New York City reflect a diverse range of cuisines and styles, with new establishments opening and generating attention on a regular basis. The most renowned restaurants in New York City are classics, places that everyone — natives and visitors alike — should eat at least once. These eateries include steakhouses, delis, and pizza shops, all of which are traditional New York

choose to follow the pattern and seek the issue that is impacting the same. We discovered that the grades are impacted by the Sidewalk size and the seating arrangements made by the restaurant employees. Surprisingly, the authorities permit roadside seating restaurants belonging to a certain county based on the grades they earned during the inspection.

B. Objective

The primary goal of our model is to determine which counties have restaurants with higher inspection gradings in terms of seating arrangements and adjoined sidewalk dimensions, using databases such as MongoDB and PostgreSQL, and to visualize the extracted data for interpretation Python libraries are being utilized. Open Restaurant Applications, Food Safety Inspections – Current Ratings, Liquor Authority Current List of Active Licenses, and Restaurants (rolled up) are the four datasets we used.

III. Related Work

Fakhri [1] has implemented a suggested system for the restaurant, which provides the highest rated option based on input from other users. They used a collaborative filtering methodology to accomplish this solution. The disadvantage is that the user may provide inaccurate feedback based on their one-time visit. Harpanahalli [2] suggested a digital payment technique in restaurants based on RFID tags in his paper. This can be difficult to implement, time-consuming, and not always as precise or dependable as barcode scanners.

Qadami [3] has leveraged cloud technology to create a shared restaurant model, such as sharing hardware and software to connect the menu, recipes, and chef prep, among other things, of two or more restaurants. This may be beneficial as well as unpleasant because they are revealing the privacy of their own business. Based on the stimulus organism response theory, Rajput [4] discovered consumer intention and satisfaction. This was accomplished by a poll of a group of consumers. Jadhav [5] and his team built digital restaurants and went through the kitchen, cashier everywhere utilizing smartphones, and customers may even pay cash from them with many alternatives. This has a disadvantage since if there are any problems or in emergency situations, we need another option.

Griffith [6] shows how information technology contributes to the food inspection link, which connects inspection to design and construction. It is employed in the healthcare system. Oria [7] advocated improving the inspection system's efficiency. They used a risk-based method to examine via the FDA. Both models merely provide an overview of information technology. Kaskela [8] took two approaches to the food safety inspection: compliance and noncompliance. Then they stated that unannounced inspections result in high accuracy in their project. The disadvantage is that it is only done for item-specific examination. The report [9] discusses food safety problems in restaurants and school foodservice foundations. They demonstrated statistical methodologies such as mean, MSE, and other aspects in this article. The document [10] depicts the inspection capabilities of the Philippine government's food service. They also used a risk-based strategy for food inspection. Treno [11] has deduced the future likely path of the alcohol sector and other drug legislation these days. It's similar to a standard alcohol literature review. From 1940 through 2013, they published the drug policy. The article [12] depicts a health research board study that provides an overview of alcohol usage and damage in Ireland. This mostly demonstrates the involvement of young people in alcohol drinking. LeClercq's [13] idea was that raising alcohol costs would lower consumption. This paradigm is being compared to viewpoint analysis. This method has become more complicated as they divide the technique into five distinct groups. Wolf [14] was discussing the association between social host policies and juvenile alcohol use according to age. To overcome this problem, they used a negative binomial and logistic regression model. Zahoor [15] built a machine learning method to do emotional analysis and categorization of restaurant evaluations. This is a useful model that accepts either a 1 or a 0 input. In the work [16], demand forecasting in restaurants is demonstrated using machine learning and statistical methodologies. To generate the result in this work, they employ a variety of ML algorithms in conjunction with statistical approaches. Priya [17] used a machine learning technique and regression models to estimate restaurant ratings. To make an accurate prediction, they must have access to all previous data. Holmberg [18] uses machine learning methods to anticipate restaurant revenues. There are several eateries in various cities. Out of all the procedures, predicting the data became critical.

IV. Methodology

1. Data Set description

TO analyze the data 4 data sets have been chosen, 3 related to restaurants and 1 being the current list of active Liquor licenses.

-Open Restaurant Applications- *Open Restaurant Applications* is an unstructured dataset with the JSON extension that contains applications from food service establishments seeking permission to re-open under Phase Two of the State's New York Forward Plan and place outdoor seating in front of their business on the sidewalk and/or roadway. It is made up of 13,057 rows and 36 columns. It has columns for Restaurant name, Borough (county), Postcode, sidewalk measurements, and so on. The dataset is available at Open Restaurant Applications | NYC Open Data (cityofnewyork.us)

- Food Safety Inspections – Current Ratings-
Food Safety Inspections – Current Ratings is an unstructured dataset with the JSON extension that describes the inspection grade, inspection date, owner name, and a variety of other factors connected to food safety inspections. There are 3,83,233 rows and 26 columns in the data set. The data set can be found at Food Safety Inspections – Current Ratings | State of New York (ny.gov).

- Liquor Authority Current List of Active Licenses-
Liquor Authority Current List of Active Licenses is an unstructured dataset that was downloaded in a JSON file format. It specifies the serial number of the license, County, Certificate number, Premise addresses, and numerous other variables related to the Liquor Authority's current list of active liquor licenses. There are 50,121 rows and 36 columns in the data set. The data set may be found at Liquor Authority Current List of Active Licenses | State of New York (ny.gov).

- Restaurants (rolled up)-
Restaurants (rolled up) is an unstructured dataset that was retrieved in a JSON extension, from Restaurants (rolled up) | NYC Open Data (cityofnewyork.us), which entitles restaurant details. Having ZIPCODE, DBA(Restaurant name), BORO(County) and several other variables relating to the various restaurants in NYC. The data set is made up of 29,962 rows and 6 columns.

2. Process flow

In this project, a certain technique is being used. It starts with importing JSL File Format datasets into Python for analysis, then the Unstructured data sets are imported to MONGO DB, making them Structural in data frames, and combining all the datasets. It is retrieved back into Python for visualization once it has been organized and merged.

At various stages, the ETL process is utilized to extract, convert, and load data. The relevant data is downloaded in JSON file format from several sources using SODA API and imported into Python in the first stage. Later, it is translated into dictionary format to add uniformity to the dataset, which is then placed into MongoDB database collections.

What occurs with MongoDB is that the MongoDB Atlas on Amazon Web Service client is one of the most popular DBMS designed specifically for document-oriented datastores. Despite the fact that it stores data in JSON format, MongoDB saves JSON files in a binary encoded format, which is the binary java – script object notation (BSON), for efficiency and a high-performance index. It also produces a worldwide unique identifier for objects called ObjectID, which is a 12-byte BSON primary key that aids in the uniqueness and integrity of the file in a database.

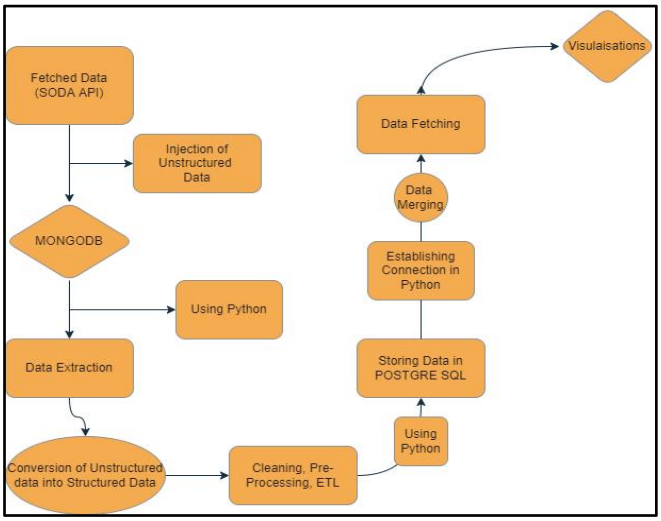

Fig. 2.1 FLOW CHART

GridFS is a function that allows MongoDB to store and retrieve binary temporary files. It stores files in two collections: 'files' and 'chunks.' GridFS separates files into chunks and stores each as a distinct document in the chunks collection, while the file collection maintains metadata.

The data was extracted from MongoDB again in the following stage, and all datasets were cleaned.

After cleaning and filtering the data, it is imported into PostgreSQL. Finally, the data is taken from PostgreSQL and different libraries and functions are applied to the data to conduct visualization in order to gain meaningful insights.

3. Extract

SODA API is used to retrieve data from various datasets available at Data.gov.us. The datasets in.json format were first transformed to dict format before being exported to MongoDB via a connection established between the Python server and the MongoDB server. The JSON files are stored in MongoDB using the MongoClient and Pymongo Library.

4. Pre-processing

- **Identifying Missing or Null values**: Missing values are recognized and handled by either removing the associated rows or replacing them with a sequence of the mean/median of their respective columns. The Null values were identified using merge data 1342.isnull()sum() function.
- **Removing Special characters and Null values**: The Data frame has several rows and columns that include special characters and Null values, which may create errors during execution or skew the results. To eliminate special characters and Null values, the *encode()* and *merge_data_1342.dropna()* functions were used, respectively.
- **KNN Imputation:** Missing values of an attribute are imputed using the supplied number of attributes that are most comparable to the missing attribute's values. imputer = *KNNImputer(n_neighbors=2)* and *imputer.fit_transform*, are being used to impute.

5. Load

PostgreSQL is used to analyze the cleaned and filtered data frame. *Psycopg2*, a well-known library function in the PostgreSQL adapter for Python, was used to load the data into the PostgreSQL database.

6. Data Retrieve from POSTGRE SQL

The compiled data is acquired from four tables using the aliases 1, 2, 3, and 4, using the "pd.read_sql" function, with the assistance of the shared primary key ID, and placed in a single data frame in Python, ready for display.

7. Visualizations

Following data extraction from the PostgreSQL database, the aggregated data was placed in a single data frame to allow for analysis and interpretation of the results.

The Graphs, Catplots and Countplots shown below were used to visualize the results of the complied data after extraction.

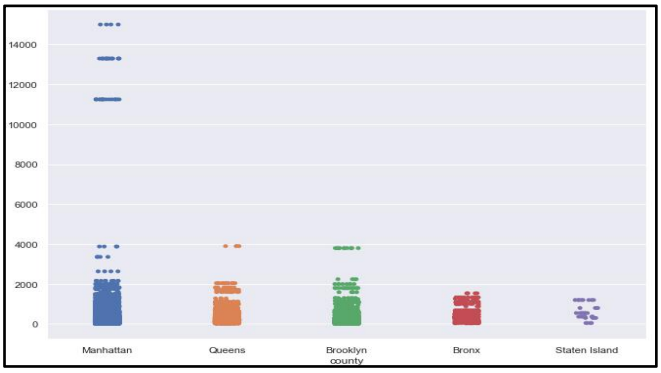

Fig. 7.1 Sidewalk_dimensions_area x Counties

Because Manhattan's sidewalks are fairly large, more restaurants receive the inspection grade 'A.' The preceding picture (Fig. 7.1) indicates which restaurants in the different Counties have a larger Sidewalk dimension area and so receive inspection results, 'A', 'B', and 'C'.

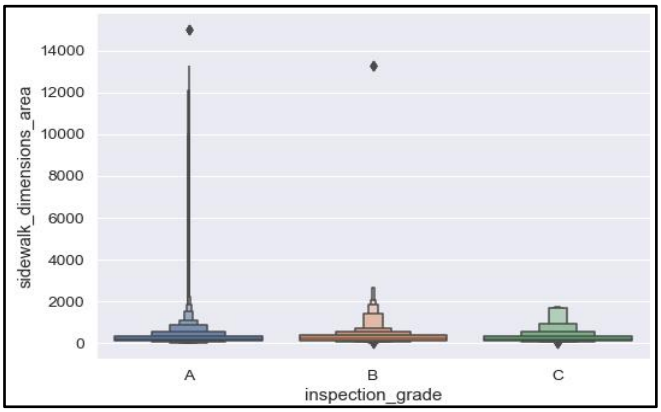

Fig. 7.2 Sidewalk_dimensions_area x Inspection_grades

The restaurants that obtained Grade 'A' during the inspection cover an area of more than 12000m^2. Outdoor seating may be simply placed in front of the individual businesses on sidewalks. Restaurants that do not receive grade 'A' have an area allotment of less than 4000m^2.

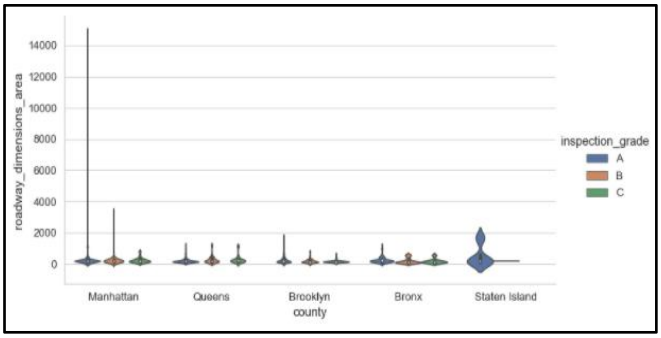

Fig. 7.3 Roadside_dimensions_area x Counties

The 'A' rated restaurants in Manhattan City have a roadside area of more than 14000m^2, whereas the 'B' and 'C' rated restaurants don't even have 4000m^2 for themselves. The other counites are being isolated in the area allotment race. Apart from Manhattan, none of the other counties, exceed the 2000m2 level, Except for a few 'A' graded enterprises on Staten Island.

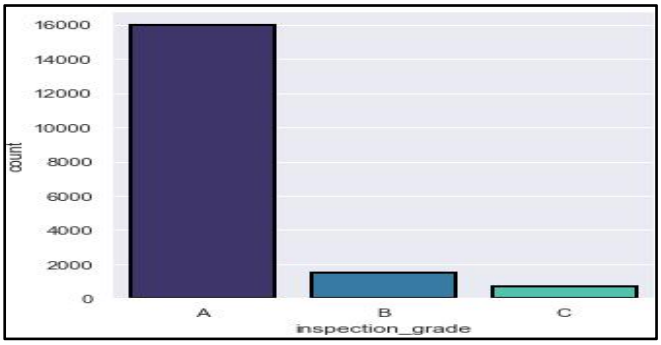

Fig. 7.4 Count of Inspection_grade

More than 16000 full-service restaurants earned 'A' grade inspection ratings across all counties. On the other side, there are about 2000 'B' and 'C' graded restaurants.

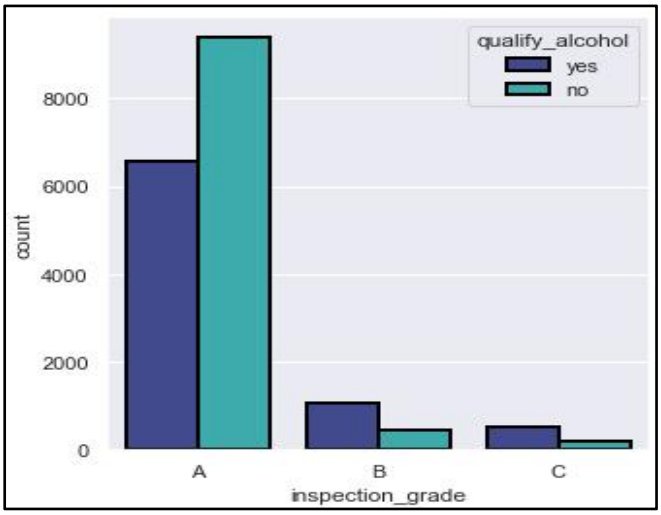

Fig. 7.5 Inspection_grade count for Alcohol qualification

More than 6000 restaurant chains did not meet the requirements for serving alcohol on their premises. Despite the fact, they received an 'A' on their inspections. 'B' and 'C' rated restaurants are still below the 2000 standard.

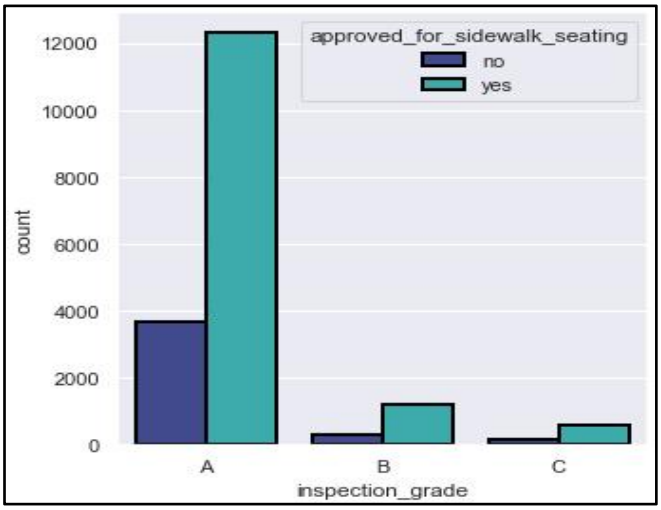

Fig. 7.6 Side_walk Seating approval on the basis of Inspection results

There were around 4000 eateries in New York City that were not allowed for sidewalk dining. Despite receiving an 'A' grade. Fortunately, several restaurants with grades other than A were still gaining permission from the authorities to have sidewalk seating arrangements.

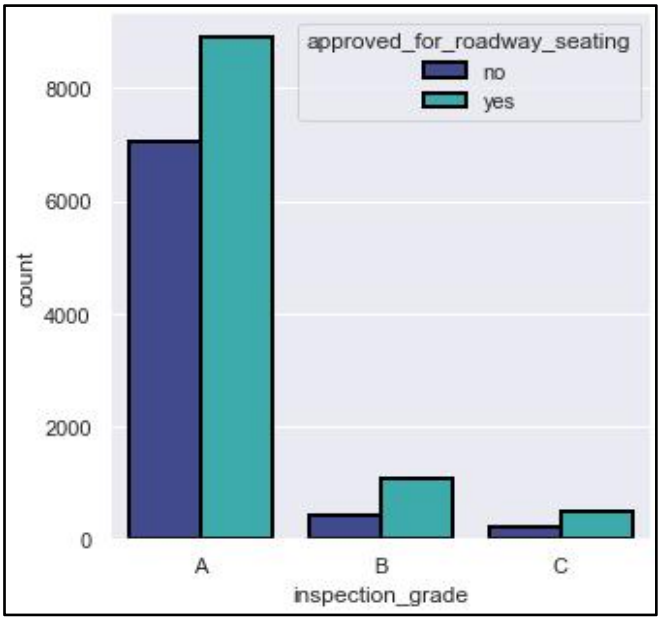

Fig. 7.7 Road_way seating approvals on basis of Inspection results

More than 8000 'A' rated cafés were given permission to place seating configurations beside the road. Almost 7000 eateries, on the other hand, were not granted the go-ahead to place the seating arrangement, regardless of the gradings they obtained during inspection.

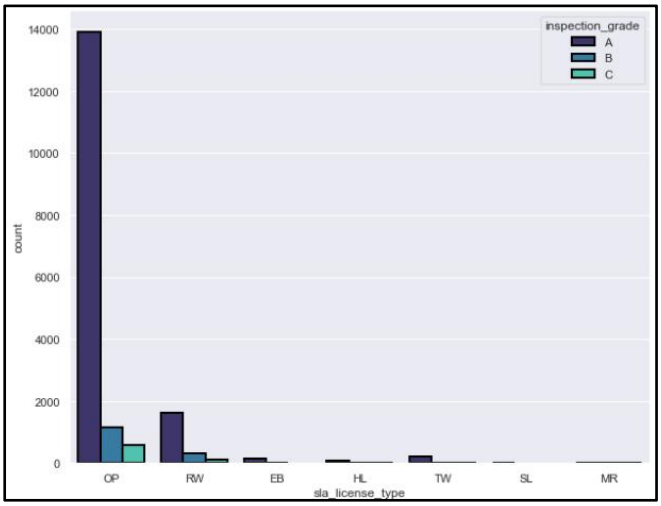

Fig. 7.8 Count of SLA_license_type basis of Inspection_grade

When their inspection findings were an 'A,' around 14000 cafés acquired an 'OP' SLA-license-type. Apart from A, the grades achieved by the restaurants did not help them gain a position when applying for a SLA license.

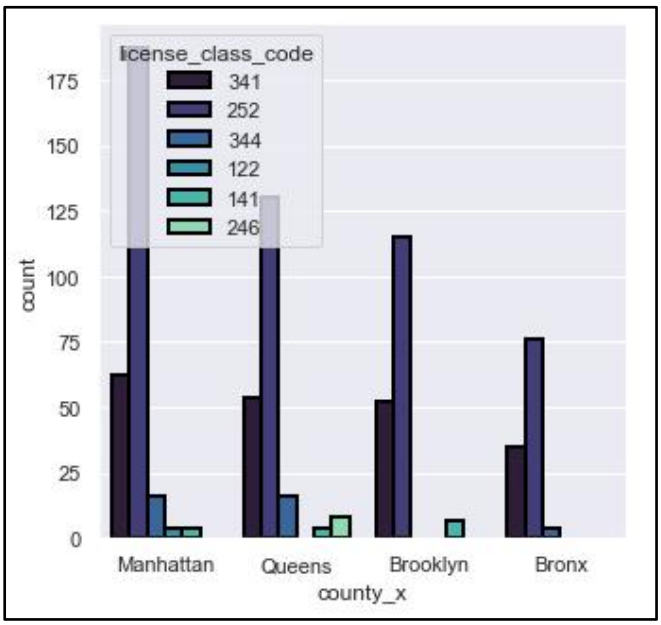

Fig. 7.9 Count of Counties on basis of License_classes

No matter which county they were from, license class 252 was the most popular license class received by restaurants. Whereas restaurants in Queens County did not acquire license 122, restaurants in Brooklyn did not receive licenses 344, 122, or 246.

V. Results and Evaluation

The visual outcome below demonstrates that our goal of locating a county with restaurants that received higher gradings throughout the inspection in terms of seating arrangements, roadside sitting arrangements, and alcohol serving qualification was attained.

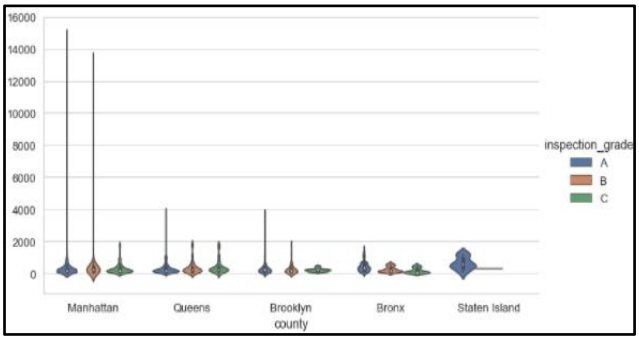

Fig. 1 Results

Manhattan, being the county with the most 'A' graded restaurants (almost 16000), has earned the status of finest county servicing the customers with the maximum potential while complying by the legal constraints and consumer demands. In comparison, just 4000 eateries in Queens and Brooklyn County obtained a 'A' rating during their inspection. On the other hand, the Bronx and Staten Island did not even reach the 2000 restaurant mark.

VI. Conclusion and Future Work

As a result, a systematic analytical research of restaurants was constructed by assembling four separate unstructured datasets of,. JSON is the file extension. The MongoDB database was used effectively to convert non-relational databases to relational databases, and complicated data was stored and retrieved. Various pre-processing methods were used to sanitize the data before it was successfully entered into PostgreSQL, where it was consolidated and extracted by queries as a single compliant data frame using the inner-join function. Machine learning analyses were used to create several visuals based on the dataset. The models were accurate in depicting Manhattan as the finest county. As part of future development, full-text queries in the MongoDB database may be created to search for specific data and extract it from the collection database. On the final dataset, many Classification models may be implemented to assess and forecast the values for future phases.

-This project is available at: https://github.com/Simbu1212/DAP-Final-Project

VII. Bibliography

[1] A. A. Fakhri, Z. K. A. Baizal, and E. B. Setiawan, "Restaurant Recommender System Using User-Based Collaborative Filtering Approach: A Case Study at Bandung Raya Region," Journal of Physics: Conference Series, vol. 1192, p. 012023, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012023.

[2] Harpanahalli, K. Bhingradia, P. Jain, and J. Koti, "Smart Restaurant System using RFID Technology," IEEE Xplore, Mar. 01, 2020. https://ieeexplore.ieee.org/document/9076398.

[3] S. F. H. Al Qadami, "Research and Development of Shared Restaurant Platform Based on Cloud Computing," American Journal of Industrial and Business Management, vol. 08, no. 12, pp. 2321–2333, 2018, doi: 10.4236/ajibm.2018.812155.

[4] A. Rajput and R. Z. Gahfoor, "Satisfaction and revisit intentions at fast food restaurants," Future Business Journal, vol. 6, no. 1, Jun. 2020, doi: 10.1186/s43093-020-00021-0.

[5] P. Jadhav, P. Teli, S. Korade, and V. Chavan, "Implementing Digital Restaurants and Inter-Restaurant Navigation Using Smart Phones," International Journal of Computer Science and Mobile Computing, vol. 4, no. 2, pp. 319–324, 2015

[6] C. J. Griffith, "Are we making the most of food safety inspections?," British Food Journal, vol. 107, no. 3, pp. 132–139, Mar. 2005, doi: 10.1108/00070700510586452.

[7] N. R. C. (US) C. on the R. of F. and D. A. R. in E. S. Food, R. B. Wallace, and M. Oria, Enhancing the Efficiency of Inspections. National Academies Press (US), 2010.

[8] J. Kaskela, R. Sund, and J. Lundén, "Efficacy of disclosed food safety inspections in restaurants," Food Control, p. 107775, Nov. 2020, doi: 10.1016/j.foodcont.2020.107775.

[9] https://www.foodprotection.org/files/food-protection-trends/Jan-Feb-14-kwon.pdf

[10] "MDPI - Publisher of Open Access Journals," Mdpi.com, 2017. https://www.mdpi.com.

[11] A. J. Treno, M. Marzell, P. J. Gruenewald, and H. Holder, "A review of alcohol and other drug control policy research," Journal of studies on alcohol and drugs. Supplement, vol. 75 Suppl 17, no. Suppl 17, pp. 98–107, 2014.

[12] "New HRB overview presents latest research on alcohol consumption, harm and policy in Ireland," www.hrb.ie. https://www.hrb.ie/news/press-releases/single-press-release/article/new-hrb-overview-presents-latest-research-on-alcohol-consumption-harm-and-policy-in-ireland/.

[13] J. LeClercq, S. Bernard, F. Mucciaccio, and M. B. Esser, "Prospective Analysis of Minimum Pricing Policies to Reduce Excessive Alcohol Use and Related Harms in U.S. States," Journal of Studies on Alcohol and Drugs, vol. 82, no. 6, pp. 710–719, Nov. 2021, doi: 10.15288/jsad.2021.82.710.

[14] J. P. Wolf, S. Islam, G. García-Ramírez, M. J. Paschall, and S. Lipperman-Kreda, "Relationships Between Social Host Policies, Youth Drinking Contexts, and Age," Journal of Studies on Alcohol and Drugs, vol. 82, no. 6, pp. 730–739, Nov. 2021, doi: 10.15288/jsad.2021.82.730.

[15] K. Zahoor, N. Z. Bawany, and S. Hamid, "Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning," 2020 21st International Arab Conference on Information Technology (ACIT), Nov. 2020, doi: 10.1109/acit50332.2020.9300098.

[16] “ScienceDirect.com | Science, health and medical journals, full text articles and books.,” www.sciencedirect.com.
[17] J. Priya, “Predicting Restaurant Rating using Machine Learning and comparison of Regression Models,” IEEE Xplore, Feb. 01, 2020.
[18] M. Holmberg and P. Halldén, “Examensarbete 30 hp Maj 2018 Machine Learning for Restaurant Sales Forecast.” Accessed: Dec. 31, 2021. [Online]. Available: https://uu.diva-portal.org/smash/get/diva2:1216397/FULLTEXT01.pdf.