PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

TITLE	**DOHMH New York City Restaurant EDA Project**
NAME	**Pavithra M**
COURSE	**DADS - Offline**
BATCH	**July - 2025**

1.INTRODUCTION

The NYC dataset contains detailed information about food establishments, their inspections, and geographic attributes. It includes variables such as inspection **SCORE**, **GRADE**, establishment identifiers (CAMIS, BIN, BBL), and location details (**ZIPCODE, Latitude, Longitude, Street, Community Board, Council District, Census Tract, NTA**). This dataset enables analysis of food safety compliance across different neighborhoods, identification of geographic patterns, and exploration of governance-related impacts on inspection outcomes.

Project Overview:

- Analyze food inspection scores and grades.
- Study the impact of location factors on results.
- Find geographic and trend patterns.
- Clean data for accurate analysis.
- Provide insights to support better decisions.

2. AIM OF THE PROJECT

- To study NYC food inspection scores and grades.
- To understand the effect of location factors on outcomes.
- To clean and transform data to enhance quality and accuracy for analysis.
- To apply statistical hypothesis testing to validate observed patterns and relationships.
- To identify geographic and trend patterns through EDA.
- To handle data issues for reliable analysis.
- To provide insights for better food safety decisions.

3. PROBLEM STATEMENT

Food safety is a critical public health concern in New York City, where thousands of establishments are inspected regularly.The project aims to analyze NYC restaurant inspection data to identify key factors affecting health inspection scores. By finding trends, outliers, and common issues, it seeks to understand why some restaurants perform poorly. The ultimate goal is to provide insights that help improve food safety and maintain high hygiene standards across restaurants.

4. PROJECT WORKFLOW

1. Data Understanding

- Collect and explore the NYC food inspection dataset.
- Identify key variables (SCORE, GRADE, location features, etc.).
- Understand data structure, types, and relevance.

2. Data Cleaning & Preprocessing

- Handle missing values, invalid entries, and duplicates.
- Treat outliers using IQR or statistical methods.
- Standardize formats (ZIP codes, phone numbers, etc.).
- Encode categorical variables if needed.

3. Statistical Tests

- Performed Independent T-Statistic test
- Performed Chi Square Test

4. Exploratory Data Analysis (EDA)

- **Univariate Analysis**: Distribution of SCORE and BORO.
- **Bivariate Analysis**: Relationship between SCORE by GRADE & SCORE vs INSPECTION DATE.
- **Multivariate Analysis**: Barplots and Correlation heatmaps.
- Detect multicollinearity among features.

5. Insights & Interpretation

- Identified geographic trends.
- Highlight weak, moderate, and strong correlations.
- Detect multicollinearity and recommend feature selection.
- Provided data-driven insights for food safety compliance.

6. Visualization

- Created histograms,Countplots,Violin plots, barplots, Line plots, and boxplots.
- Developed correlation heatmaps.
- Summarized findings through clear visual storytelling.

7. Conclusion & Recommendations

- Summarize key findings (influence of geographic patterns).
- Suggest improvements for inspection monitoring.
- Provide recommendations for policy and decision-making.

5. Data Understanding

This step involves exploring the NYC food inspection dataset to get familiar with its structure, variables, and contents. Key aspects include identifying important features (like SCORE, GRADE, location details), checking data types, and summarizing basic statistics. The goal is to understand what the data represents and assess its quality before further analysis.

Dataset Overview

- **Rows : 290022**
- **Columns : 27**

Key Variables

1. Inspection Outcome Variables:

- SCORE
- GRADE

2. Location Variables:

- Community Board

- Council District
- Latitude & Longitude
- ZIPCODE
- Street, Census Tract, BIN, BBL, NTA

3. Establishment Identifiers & Contact:

- CAMIS
- PHONE

6. Data Cleaning

Data cleaning ensures the dataset is accurate, consistent, and ready for analysis.

Key steps include:

Dropping Features:

- Dropped Features that having too many missing values
- Dropped Features that are not useful for analysis

Handling Missing Values:

- Identified missing or null entries in critical columns (e.g., SCORE, GRADE..).
- Filled missing values using appropriate methods:

1. **Mean** for numerical data
2. **Not Available** for categorical data
3. **Mode** for GRADE
4. Date features with **default placeholder date** (not real inspection data) means the actual date was missing or never recorded.

Removing Duplicates:

- Checked for duplicate rows or repeated establishments.
- Removed duplicates to prevent skewed analysis or bias.

Handling Outliers:

- Detected outliers in numeric column like SCORE by using IQR Method.
- Repeated 3 more times to reduce outliers in SCORE for better analysis.

Correcting Invalid Data:

- Standardized formats for categorical variables (GRADE should be A/B/C).
- Changed data types for INSPECTION DATE & GRADE DATE to **datetime** for meaningful analysis.

Outlier Treatment:

- Used Boxplot for Identify Outliers in numeric column (SCORE).
- Used **Interquartile (IQR)** Method to remove outliers.

After cleaning my dataset :

- **Rows - 290016**
- **Columns - 10**

7. Feature Engineering

Feature engineering enhances the dataset's usefulness by creating meaningful, analysis-ready variables that improve insights, trends detection, and predictive modeling performance.

Reducing Multicollinearity

- Identify highly correlated features.
- Dropped features to prevent redundancy in modeling.

8.Filtering Data For analysis

Filtering is the process of selecting a subset of the dataset that is relevant, clean, and suitable for a specific analysis or visualization. This helps improve accuracy and focus on meaningful insights.

Key Steps:

Remove Unnecessary Columns:

- Dropped columns that are not required for analysis (e.g., DBA) to simplify the dataset.

Select Relevant Rows:

- Filtered rows based on conditions, e.g., valid SCORE values, specific grades (A, B, C).

Handle Outliers:

- Excluded extreme SCORE values to prevent skewed results.

9.Statistical analysis and Testing

Descriptive Analysis:

Descriptive statistics summarize and describe the main features of a dataset. They include metrics like mean, median, standard deviation, minimum, maximum, and quartiles, which give a snapshot of the distribution and spread of values.

- Described SCORE.

Statistical Analysis:

Hypothesis Testing

Hypothesis testing is a statistical method used to validate whether an observed pattern in the data is due to chance or represents a meaningful difference/relationship. It involves setting up a Null Hypothesis (Ho), which assumes no effect or difference, and an Alternative Hypothesis (H1), which assumes there is a significant effect or difference. The decision to reject or accept Ho is based on a p-value compared to a significance level (alpha = 0.05 in this case).

Tests Performed

1.Independent Sample t-test

Comparing SCORE between GRADE A and GRADE B Restaurants :

Null Hypothesis Ho : No Significant difference between GRADE A and GRADE B SCORES

Alternate Hypothesis Ha : Significant difference between GRADE A and GRADE B SCORES

Results : The p_value is less than 0.05.Hence,the SCORES of GRADE A and GRADE B have significant differences.

2.Chi Square Test

Taking two Features ACTION and GRADE to check they are independent or not :

Null Hypothesis Ho : ACTION and GRADE are independent

Alternate Hypothesis Ha : ACTION and GRADE are dependent

Results : The p_value is 0.0 which is less than 0.05.HenceACTION and GRADE are dependent.

10.Exploratory Data Analysis - Univariate Analysis

1.Numerical Univariate Analysis - SCORE

- **Histplot** for SCORE Distribution to show how restaurant inspection scores vary, highlighting both average performers and outliers.

Insights :

- Most of the scores falls between 10 and 30.
- The most significant peak is around the score of 23, means most restaurants have the common score.
- A few restaurants have very high scores, meaning more violations indicating poor hygiene or safety standards.
- The uneven score pattern suggests that some restaurants consistently perform better than others in inspections.

2.Categorical Univariate Analysis - BORO

- **Countplot** for analzing Geographical trends by counting how many inspections occurred in each borough (BORO).

Insights :

- Manhattan and Brooklyn have the highest number of inspections, showing they have the most restaurants.
- Staten Island has the fewest inspections, meaning fewer food outlets operate there.
- The large difference between boroughs suggests uneven restaurant distribution across NYC.

11.Bivariate Analysis

1.Bivariate analysis - SCORE by GRADE

- Used **Violin plot** of SCORE by GRADE shows patterns in inspection scores & reveals trends across grades.

Insights :

- Scores vary by grade, this shows different performance levels.
- Grades C and Z have higher scores than others.
- Grade P has the lowest scores, which indicating weaker performance.
- This helps to identify restaurants needing attention.

2.Bivariate analysis - SCORE vs INSPECTION DATE

- **Lineplot** for analyzing Time-Based Trend by using two variables SCORE & INSPECTION DATE by how the average score changes over time.

Insights :

- The Inspection scores improved over the years, showing better hygiene awareness.
- Some years show score drops, possibly due to stricter inspection rules or new policies.
- The overall upward trend indicates positive progress in food safety standards across NYC restaurants.

12.Multivariate Analysis

1.Multivariate analysis - SCORE by GRADE & ACTION

- **Barplot** to analyze patterns in inspection scores across grades, identifying areas for improvement & trends in restaurant performance (SCORE by GRADE & ACTION).

Insights :

- The majority of restaurants receiving "Grade A" had no recorded violations during inspection.
- Establishments that were closed or re-closed by the DOHMH generally received lower health grades like 'N' or 'C'.

2.Correlation Heatmap - SCORE vs CRITICAL FLAG

- Used **Correlation Heatmap**
- Converted CRITICAL FLAG to numeric to analyze the violations based on the inspection SCORE.

Insights :

- Both SCORE and CRITICL_FLAG show a perfect positive self-correlation of 1.0, means both variables move together.
- A very weak positive correlation(0.13) exists between two variables.
- Which indicating that higher scores are only minimally realated to having a critical violation flag.
- The low correlation suggests that critical violations are not the primary key of restaurant's final health inspection score.

13. Overall Insights from analysis

Most NYC restaurants maintain good hygiene, achieving grade A.

A few restaurants consistently score low, indicating poor compliance.

Inspection scores vary by borough, with Manhattan and Brooklyn having the most inspections.

Scores have generally improved over the years, showing better food safety awareness.

Data highlights areas and restaurants needing targeted improvement for public health safety.

Outlier analysis shows a small number of restaurants significantly affecting overall score trends.

Certain time periods show dips in scores, possibly due to seasonal or policy changes.

Restaurants with repeated violations need focused interventions to raise standards.

High-score restaurants can serve as benchmarks for best practices across the city.

Visualization of grades vs scores reveals that improving B and C grade restaurants can significantly boost overall compliance.

14. Recommendations

Focus on improving hygiene in restaurants with low scores or grades B/C.

Conduct regular and stricter inspections in high-risk areas.

Provide staff training on food safety and sanitation practices.

Recognize and reward restaurants that consistently maintain high standards.

Use data trends to monitor performance and prevent future violations.

Implement corrective action plans for restaurants with repeated violations.

Promote awareness campaigns for food safety among new or small restaurants.

Encourage sharing of best practices from top-performing restaurants.

Monitor seasonal or policy-related drops in scores to address root causes.

Leverage predictive analysis to identify potential high-risk restaurants before inspections.

15. Conclusion

The project provided valuable insights through data analysis and visualization. By performing statistical analysis, key patterns in restaurant performance were revealed, helping identify areas needing improvement in food safety and hygiene. Through EDA, including univariate, bivariate, and multivariate analyses, important trends and relationships were discovered within the dataset. These findings support better decision-making and highlight areas for improvement or further exploration. Overall, the analysis provides insights to enhance restaurant quality and public health safety. Additionally, it helps prioritize inspections and training for low-performing restaurants and promotes the adoption of best practices from top-performing establishments.