PROJECT DOCUMENTATION

EXPLORATORY DATA ANALYSIS USING PYTHON

TITLE	Exploring Patterns in Restaurant Inspection Data from NYC
NAME	Navaneetha Krishnan A
COURSE	DA/DS Offline
BATCH	July 2025

TABLE OF CONTENT

1. INTRODUCTION

The NYC Restaurant Inspection dataset contains detailed records of inspections for restaurants throughout New York City, including restaurant name DBA, borough, inspection dates, scores, letter grades, and other related metadata. This project focuses on cleaning and preprocessing the dataset, creating derived features, and conducting exploratory analysis to uncover patterns in inspection scores and grades across boroughs and over time. The analysis emphasizes identifying frequent violations, temporal trends in inspection outcomes, geographic differences across boroughs and neighborhoods, and the relationship between numeric scores and assigned grades. The insights gained from this study aim to assist public-health officials, restaurant managers, and consumers in understanding compliance trends and targeting areas for improvement.

2. AIM OF THE PROJECT

The primary aim of this analysis is to examine NYC restaurant inspection data in order to understand the distribution and underlying drivers of inspection scores and grades. The specific objectives include: 1 cleaning and standardizing the dataset, 2 creating relevant derived features such as numeric grade mappings and inspection frequency metrics, 3 conducting univariate, bivariate, and multivariate exploratory analyses to identify trends and anomalies, and 4 performing statistical tests to assess whether observed differences across boroughs or over time are significant. The ultimate goal is to generate actionable insights and recommendations that can assist inspection offices and restaurant stakeholders in improving food-safety compliance.

3. PROBLEM STATEMENT

Food-safety inspections play a critical role in protecting public health and maintaining consumer trust. The central question addressed in this analysis is how inspection data can be leveraged to identify restaurants, neighborhoods, or time periods with systematically poor compliance, and to prioritize targeted interventions. Key stakeholders—including public health departments, restaurant owners, and policy-makers—need insight into where violations cluster, whether certain boroughs consistently show higher scores (indicating worse performance), and whether observed trends over time reflect genuine improvements or data artifacts. Repeated violations by the same establishment may indicate systemic issues that require focused training or stricter enforcement. Additionally, missing or placeholder values (such as 1/1/1900) and duplicate violation records complicate the analysis, making rigorous data cleaning essential for credible results. Addressing these challenges enables more effective allocation of inspection resources, informed restaurant training programs, and better public communication about food-safety compliance.

4. PROJECT WORKFLOW

- **Data cleaning & preprocessing**:
  - Standardized column names (lowercase, underscores).
  - Dropped irrelevant or highly sparse columns (location identifiers, geographic columns, long text fields).
  - Handled duplicates, converted date columns to datetime, and replaced placeholder/missing values.
- **Feature engineering**:
  - Created grade_num mapping (A=1, B=2, C=3, N=4).
  - Derived inspection frequency metrics.
  - Imputed missing score values with the dataset median.
- **Exploratory analysis**:
  - Univariate plots: histograms and boxplots.
  - Bivariate comparisons: boxplots, barplots, lineplots.
  - Heatmap pivot of mean scores by grade and borough.
  - **Statistical testing**: Chi-square test for independence, t-tests, and one-way ANOVA to evaluate borough differences.
  - **Synthesis & reporting**: Collected insights and produced actionable recommendations.

5. Data Understanding

- **Dataset overview**:
  - The NYC Restaurant Inspection dataset contains 290,022 rows and 27 columns, including restaurant name, borough, inspection dates, scores, and letter grades.
- **Key columns used in analysis**:
  - Restaurant_Name – name of the restaurant
  - Restaurant_location – borough of the restaurant
  - inspection_date – date of inspection
  - inspection_type – type of inspection
  - score – numeric inspection score

- grade – letter grade (A, B, C, N)
- grade_date – date the grade was assigned

- **Initial exploration findings**:

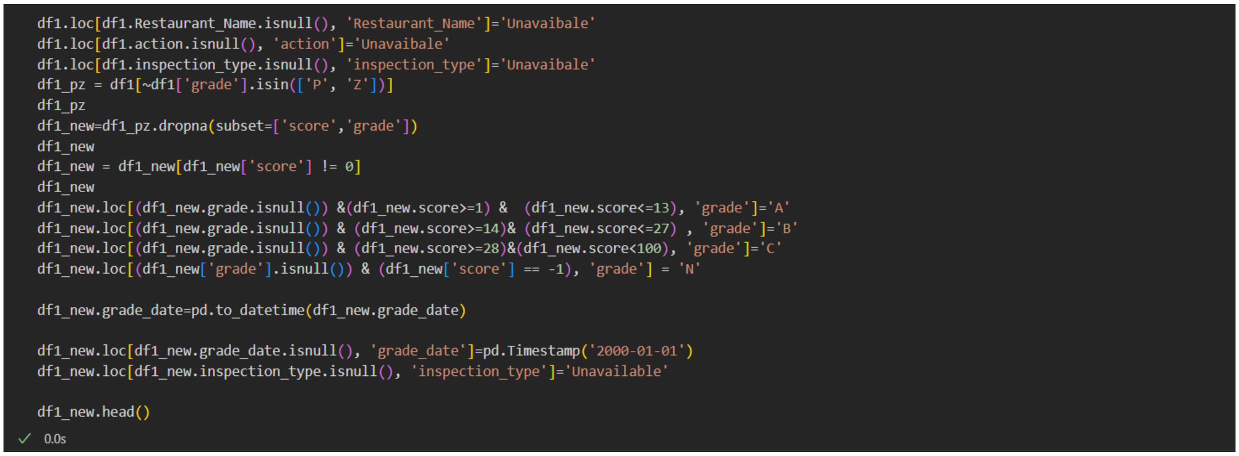
  - Some missing values, duplicate rows, and placeholder dates (e.g., 1/1/1900) were detected.
  - Multiple inspections per restaurant led to repeated records.

- **Exploratory steps**:

  - Dataset examined using .info(), .describe(), and .head() to understand data types, counts, and basic statistics.
  - This step helped identify relevant columns, cleaning needs, and features suitable for creating new metrics.

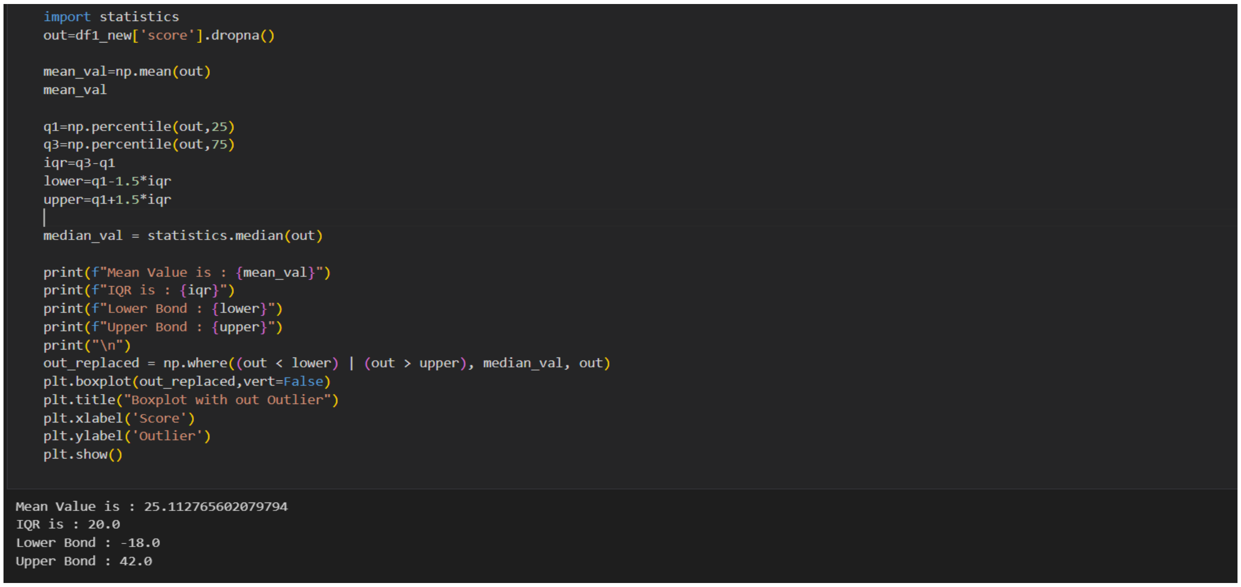
6. DATA CLEANING

- **Missing values handled**:
  - Restaurant_Name and inspection_type missing entries were filled with 'Unavailable' to retain rows.
  - score missing values were imputed with the median, which is robust to skew and outliers. Rows with score == 0 were removed, as these often represent placeholders.
  - grade_date nulls were replaced with a sentinel date (2000-01-01) to preserve temporal type while marking missingness explicitly.
- **Grade filtering and numeric mapping**:
  - Grades 'P' and 'Z' were filtered out to focus on A/B/C/N grades.
  - Letter grades were mapped to numeric values (grade_num: A=1, B=2, C=3, N=4) for numeric summaries and correlation analysis.
- **Inconsistent values handled**:
  - Standardized column names.
  - Borough code "0" entries were either removed or reclassified.



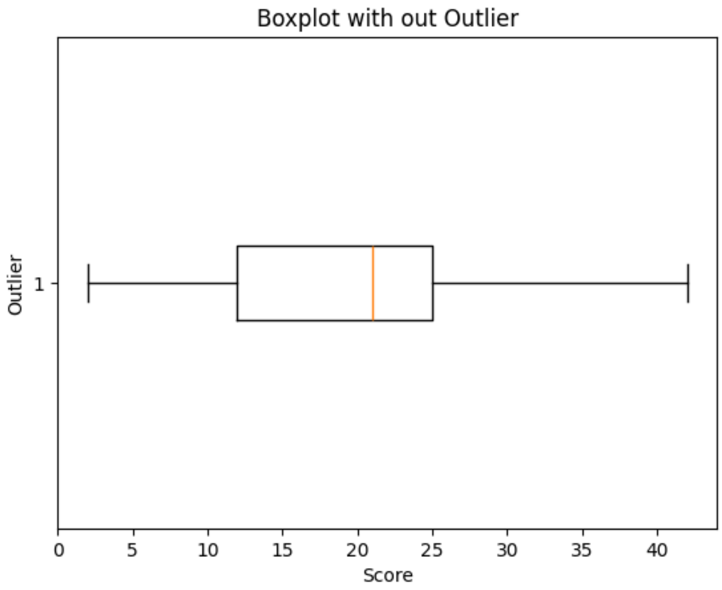
```
df1.loc[df1.Restaurant_Name.isnull(), 'Restaurant_Name']='Unavaibale'
df1.loc[df1.action.isnull(), 'action']='Unavaibale'
df1.loc[df1.inspection_type.isnull(), 'inspection_type']='Unavaibale'
df1_pz = df1[~df1['grade'].isin(['P', 'Z'])]
df1_pz
df1_new=df1_pz.dropna(subset=['score','grade'])
df1_new
df1_new = df1_new[df1_new['score'] != 0]
df1_new
df1_new.loc[(df1_new.grade.isnull()) &(df1_new.score>=1) &  (df1_new.score<=13), 'grade']='A'
df1_new.loc[(df1_new.grade.isnull()) & (df1_new.score>=14)& (df1_new.score<=27) , 'grade']='B'
df1_new.loc[(df1_new.grade.isnull()) & (df1_new.score>=28)&(df1_new.score<100), 'grade']='C'
df1_new.loc[(df1_new['grade'].isnull()) & (df1_new['score'] == -1), 'grade'] = 'N'

df1_new.grade_date=pd.to_datetime(df1_new.grade_date)

df1_new.loc[df1_new.grade_date.isnull(), 'grade_date']=pd.Timestamp('2000-01-01')
df1_new.loc[df1_new.inspection_type.isnull(), 'inspection_type']='Unavailable'

df1_new.head()
✓ 0.0s
```

- **Outlier treatment**:
    - Examined score using boxplots and the IQR method.
    - Computed Q1, Q3, and IQR; defined bounds as Q1 − 1.5×IQR and Q3 + 1.5×IQR.
    - Extreme outliers beyond these bounds were replaced with the median to reduce influence on mean-based statistics while preserving most genuine variation.



```
import statistics
out=df1_new['score'].dropna()

mean_val=np.mean(out)
mean_val

q1=np.percentile(out,25)
q3=np.percentile(out,75)
iqr=q3-q1
lower=q1-1.5*iqr
upper=q1+1.5*iqr

median_val = statistics.median(out)

print(f"Mean Value is : {mean_val}")
print(f"IQR is : {iqr}")
print(f"Lower Bond : {lower}")
print(f"Upper Bond : {upper}")
print("\n")
out_replaced = np.where((out < lower) | (out > upper), median_val, out)
plt.boxplot(out_replaced,vert=False)
plt.title("Boxplot with out Outlier")
plt.xlabel('Score')
plt.ylabel('Outlier')
plt.show()
```

```
Mean Value is : 25.112765602079794
IQR is : 20.0
Lower Bond : -18.0
Upper Bond : 42.0
```



7. OBTAINING DERIVED METRICES

- **Grade numeric mapping grade_num** – Letter grades were mapped to numbers A=1, B=2, C=3, N=4 to allow numeric summaries and correlation checks.
- **Inspection frequency / counts** – Number of inspections per restaurant was calculated to identify repeated inspections or recurring violations.
- **Temporal features inspection_year / inspection_month**– Extracted from inspection_date to study trends over time, including seasonality and compliance changes.
- **Time since last grade** – Computed for restaurants with multiple inspections to analyze intervals of improvement or regression.
- **Mean score by borough / grade** – Aggregated metrics using pivot tables to compare central tendency across boroughs and grades, supporting visualizations, statistical tests, and heatmaps.
-

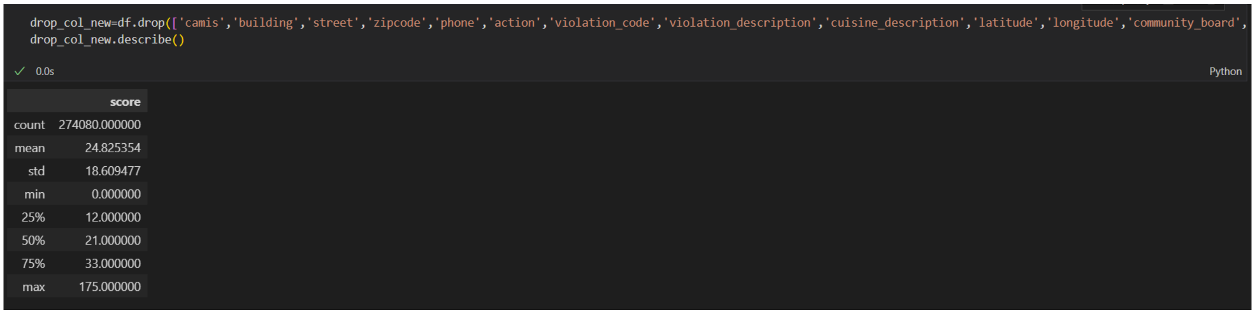
8. FILTERING DATA FOR ANALYSIS

Analytical filtering was applied to refine the dataset and ensure reliable results. Records with non-final grade values such as “P” and “Z” were removed to keep the focus on standard letter grades (A, B, C, and N). Rows with invalid entries, such as Restaurant_location == "0" or placeholder scores of zero, were also excluded. Date columns were converted and, where necessary, filtered to align with the relevant analysis period. In some cases, subsets of the data were created—for instance, isolating Manhattan and Brooklyn restaurants to compare their inspection scores and test specific hypotheses. These filtering steps aimed to preserve as much meaningful information as possible while eliminating invalid placeholders and irrelevant metadata.

9. STATISTICAL ANALYSIS

- The descriptive analysis focused on both numerical and categorical variables. For inspection scores, measures of central tendency and dispersion were calculated, including count, mean, median, standard deviation, and quartiles. For categorical variables such as grade and restaurant location, frequency counts were used to summarize distributions. Visualizations like histograms and boxplots provided additional insights into skewness, spread, and potential outliers in the data.
- To move beyond description, several hypothesis tests were applied to assess whether observed differences were statistically significant. A chi-square test of independence examined whether grade distributions varied across boroughs. When p-values were below 0.05, it indicated that borough and grade were significantly associated. A one-sample t-test was also used to compare sample means against the population mean, helping to illustrate sampling behavior and validate smaller subsets.
- An independent two-sample t-test compared mean scores between boroughs, such as Manhattan and Brooklyn, to determine whether their average inspection scores differed significantly. Assumptions of normality and equal variance were checked, with Welch’s t-test applied when variances were unequal. Finally, a one-way ANOVA tested whether mean scores differed across all

five boroughs simultaneously. When results were significant, post-hoc tests such as Tukey's HSD were conducted to identify which specific boroughs differed from one another.

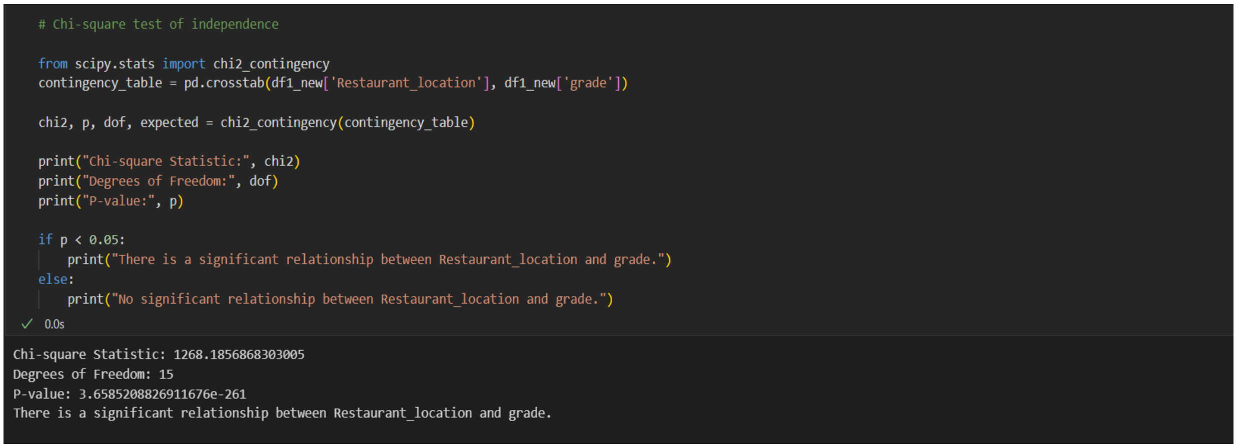


```python
drop_col_new=df.drop(['camis','building','street','zipcode','phone','action','violation_code','violation_description','cuisine_description','latitude','longitude','community_board',
drop_col_new.describe()
```

✓ 0.0s Python

	score
count	274080.000000
mean	24.825354
std	18.609477
min	0.000000
25%	12.000000
50%	21.000000
75%	33.000000
max	175.000000

Chi-square test of independence

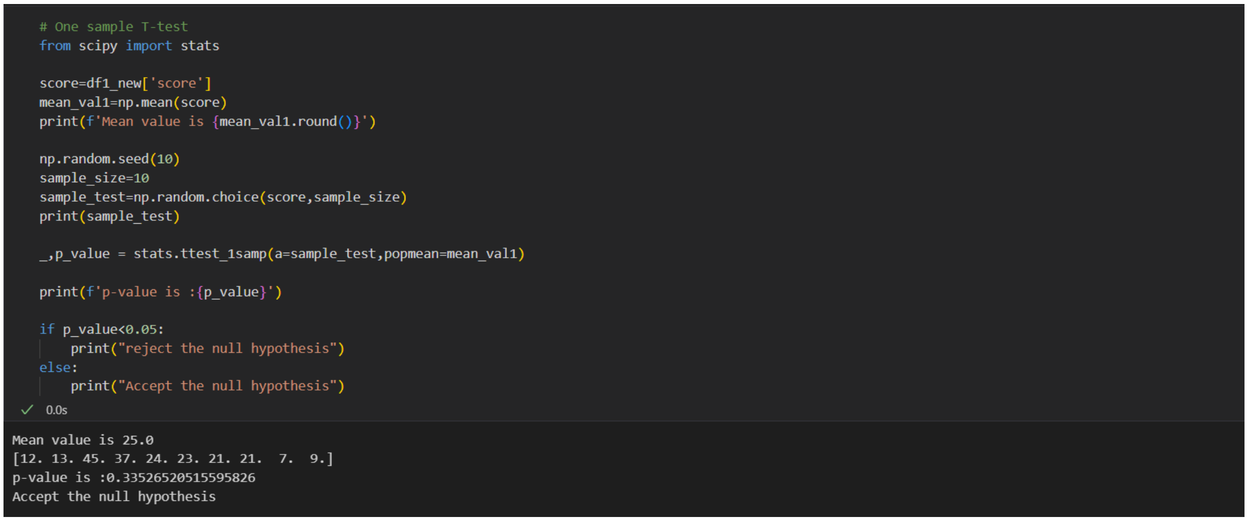


```python
# Chi-square test of independence

from scipy.stats import chi2_contingency
contingency_table = pd.crosstab(df1_new['Restaurant_location'], df1_new['grade'])

chi2, p, dof, expected = chi2_contingency(contingency_table)

print("Chi-square Statistic:", chi2)
print("Degrees of Freedom:", dof)
print("P-value:", p)

if p < 0.05:
    print("There is a significant relationship between Restaurant_location and grade.")
else:
    print("No significant relationship between Restaurant_location and grade.")
```

✓ 0.0s

```
Chi-square Statistic: 1268.1856868303005
Degrees of Freedom: 15
P-value: 3.6585208826911676e-261
There is a significant relationship between Restaurant_location and grade.
```

One sample T-test

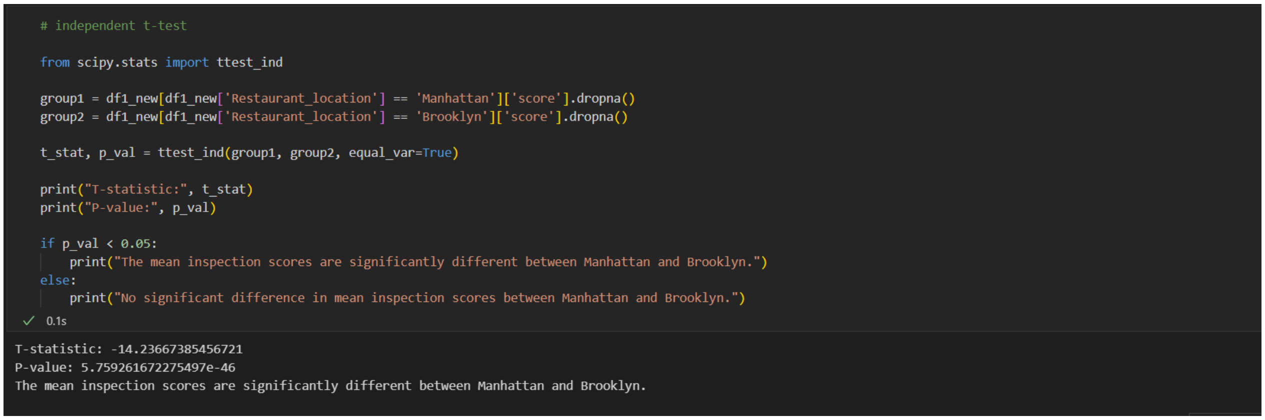


```python
# One sample T-test
from scipy import stats

score=df1_new['score']
mean_val1=np.mean(score)
print(f'Mean value is {mean_val1.round()}')

np.random.seed(10)
sample_size=10
sample_test=np.random.choice(score,sample_size)
print(sample_test)

_,p_value = stats.ttest_1samp(a=sample_test,popmean=mean_val1)

print(f'p-value is :{p_value}')

if p_value<0.05:
    print("reject the null hypothesis")
else:
    print("Accept the null hypothesis")
```

✓ 0.0s

```
Mean value is 25.0
[12. 13. 45. 37. 24. 23. 21. 21.  7.  9.]
p-value is :0.33526520515595826
Accept the null hypothesis
```

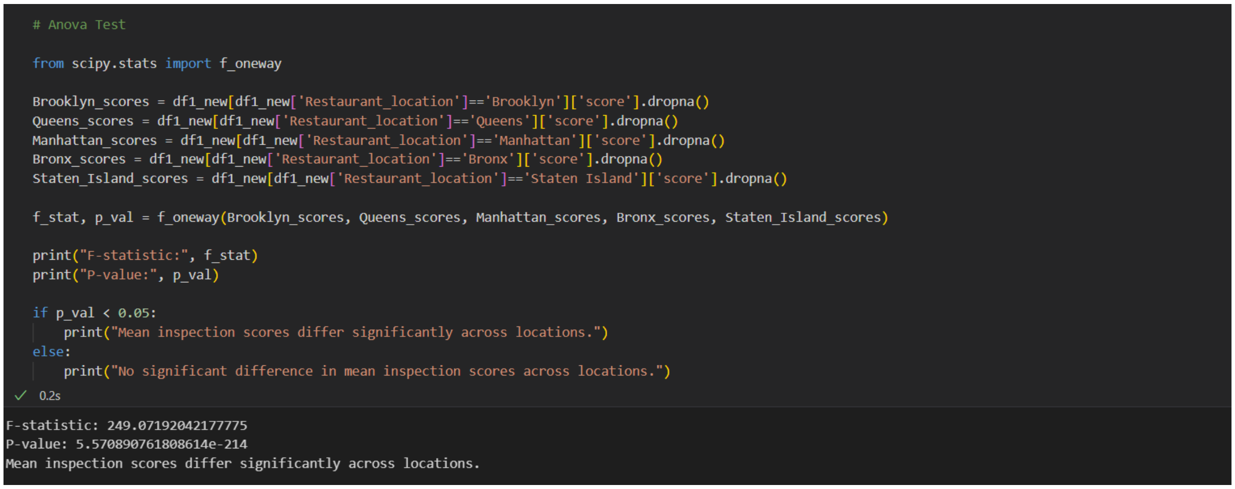
Independent t-test



```
# independent t-test

from scipy.stats import ttest_ind

group1 = df1_new[df1_new['Restaurant_location'] == 'Manhattan']['score'].dropna()
group2 = df1_new[df1_new['Restaurant_location'] == 'Brooklyn']['score'].dropna()

t_stat, p_val = ttest_ind(group1, group2, equal_var=True)

print("T-statistic:", t_stat)
print("P-value:", p_val)

if p_val < 0.05:
    print("The mean inspection scores are significantly different between Manhattan and Brooklyn.")
else:
    print("No significant difference in mean inspection scores between Manhattan and Brooklyn.")
```

```
T-statistic: -14.23667385456721
P-value: 5.759261672275497e-46
The mean inspection scores are significantly different between Manhattan and Brooklyn.
```

Anova Test



```
# Anova Test

from scipy.stats import f_oneway

Brooklyn_scores = df1_new[df1_new['Restaurant_location']=='Brooklyn']['score'].dropna()
Queens_scores = df1_new[df1_new['Restaurant_location']=='Queens']['score'].dropna()
Manhattan_scores = df1_new[df1_new['Restaurant_location']=='Manhattan']['score'].dropna()
Bronx_scores = df1_new[df1_new['Restaurant_location']=='Bronx']['score'].dropna()
Staten_Island_scores = df1_new[df1_new['Restaurant_location']=='Staten Island']['score'].dropna()

f_stat, p_val = f_oneway(Brooklyn_scores, Queens_scores, Manhattan_scores, Bronx_scores, Staten_Island_scores)

print("F-statistic:", f_stat)
print("P-value:", p_val)

if p_val < 0.05:
    print("Mean inspection scores differ significantly across locations.")
else:
    print("No significant difference in mean inspection scores across locations.")
```

```
F-statistic: 249.07192042177775
P-value: 5.570890761808614e-214
Mean inspection scores differ significantly across locations.
```

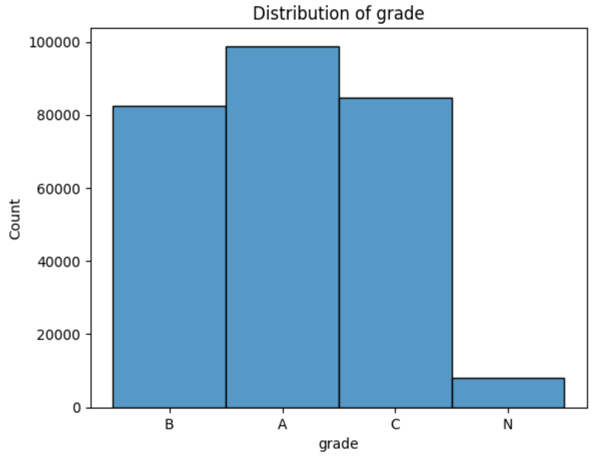
10. OVERALL INSIGHTS FROM ANALYSIS

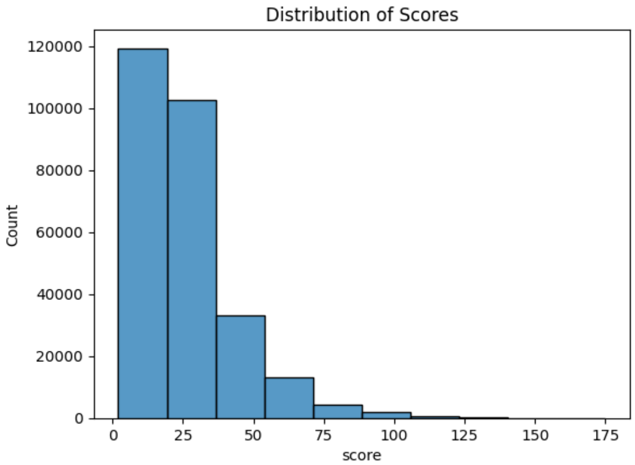
The distribution of scores follows a right-skewed pattern, with most restaurants receiving relatively low scores and fewer cases of very high violation scores. Histograms and kernel density plots highlighted this trend, showing a central tendency clustered around the lower range, while boxplots revealed several outliers beyond the interquartile range. To ensure robust statistical summaries, these outliers were addressed using median replacement.

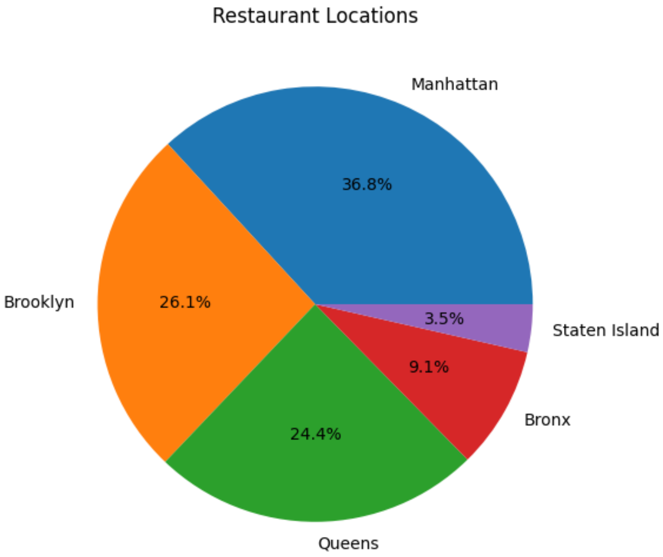
Grade distribution analysis showed that grades A, B, and C were present, with A being the most common after data cleaning. In some cases, a grade of “N” appeared, indicating restaurants that had not yet received a letter grade. The proportions of each grade were clearly illustrated using bar charts, helping to show the dominance of A-grade inspections.

Location-based distributions provided additional insights. Borough-level counts demonstrated how inspections were spread across Manhattan, Brooklyn, Queens, the Bronx, and Staten Island. Visualizations such as pie charts or bar plots highlighted each borough’s share of inspections. Further investigation into restaurant-level counts showed that many establishments underwent multiple inspections, with some consistently flagged for repeated violations. These univariate analyses set the foundation for deeper bivariate and multivariate explorations later in the study.

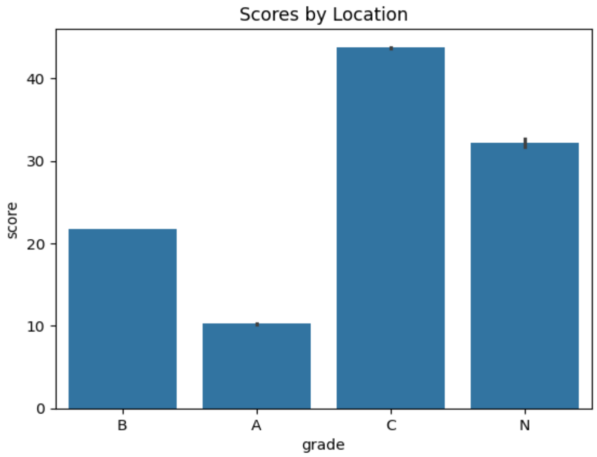
Univariate Analysis

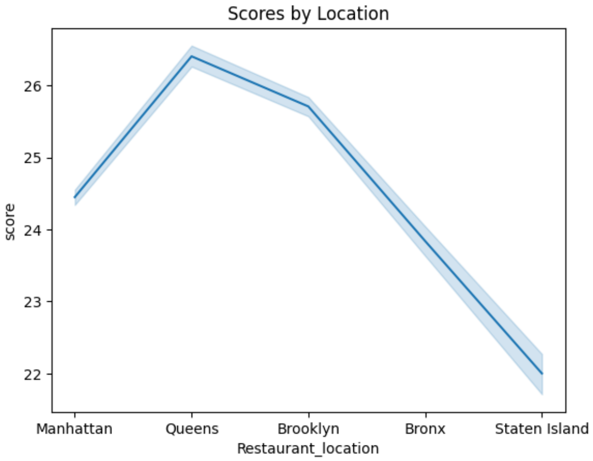


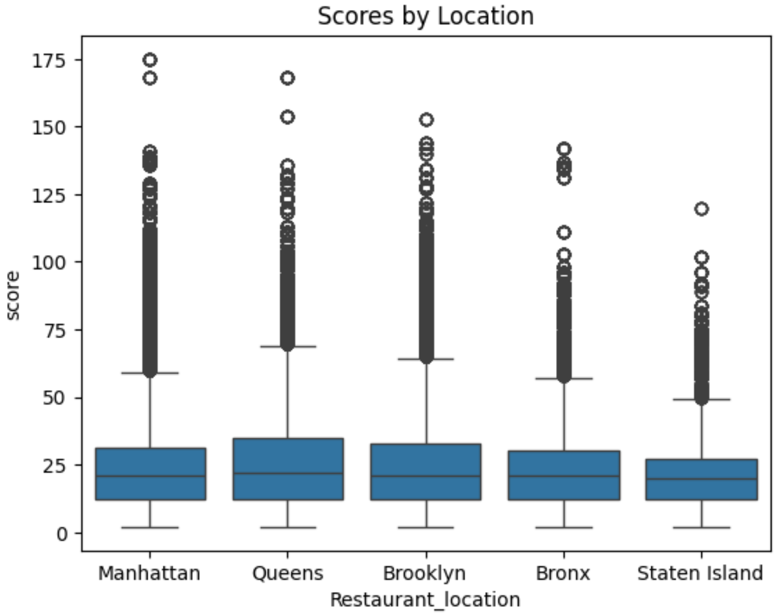




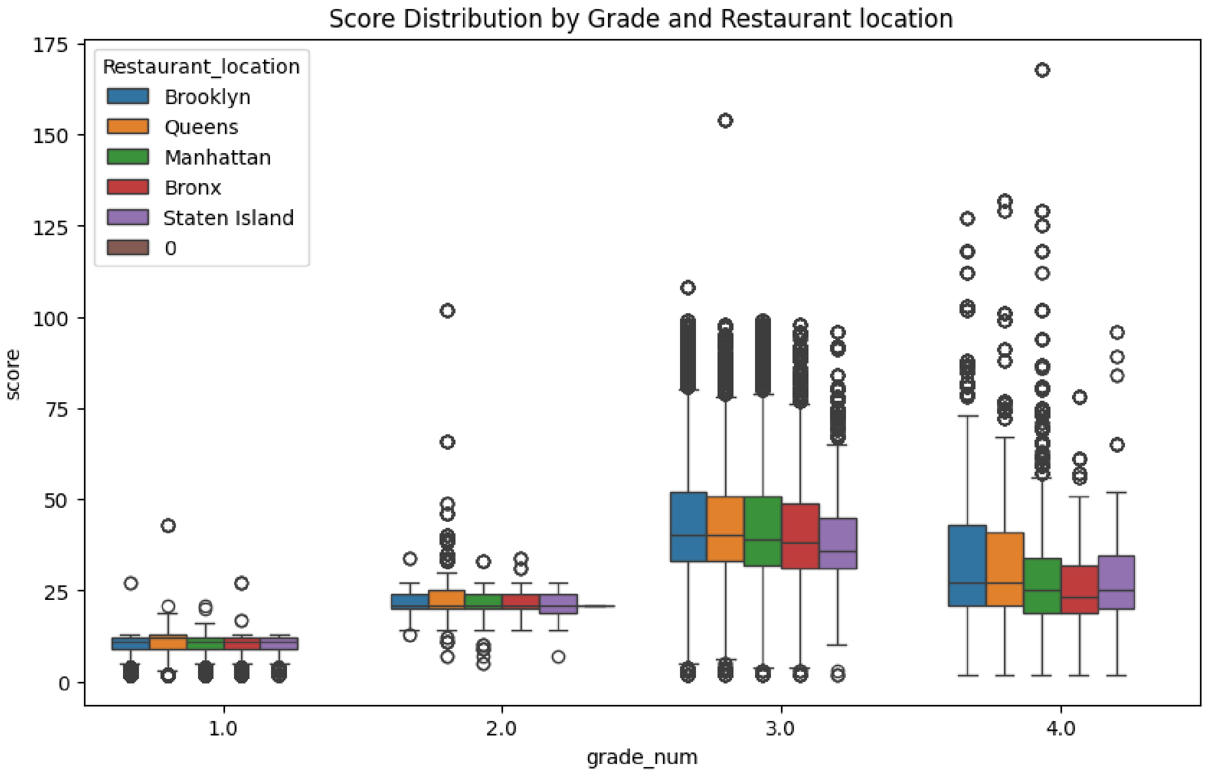
11. Bivariate Analysis







12. Multivariate Analysis



13. Overall Insights from Analysis

- The analysis confirmed that numeric scores and letter grades align as expected. A clear monotonic relationship was observed, where restaurants with an A grade consistently had lower median scores, while B and C grades showed progressively higher medians. This not only validates the grading logic but also supports the use of numeric grade values as a reliable proxy for quality.
- Differences were also evident across boroughs. Visualizations and group comparisons revealed variations in mean and median scores, with some boroughs displaying higher central tendencies or greater variability. Statistical tests such as ANOVA and chi-square confirmed that, in cases where p-values fell below 0.05, borough membership significantly influenced score and grade distributions. These disparities likely reflect differences in restaurant types, inspection frequency, or enforcement practices across boroughs.
- Another key finding was the prevalence of repeated inspections and violations. Many restaurants appeared multiple times in the dataset, often with recurring issues. Tracking inspection frequency and violation counts helps highlight chronic offenders and restaurants that may require additional training or closer monitoring.

- Data quality issues were also identified and addressed. Placeholder dates, zero scores, and invalid entries such as “Restaurant_location = 0” required domain-aware cleaning. Median imputation for missing values and sentinel replacements for missing dates helped preserve the dataset while making potential biases clear.
- Outliers presented another challenge, as unusually high violation scores tended to skew summary statistics. To counter this, both mean and median values were reported, and median imputation was applied to maintain balanced insights.
- Temporal patterns were observed as well, though they require deeper exploration. Early line plots suggested score fluctuations over time, and more advanced time-series techniques like seasonal decomposition and year-over-year analysis could reveal policy effects or seasonal influences on inspection outcomes.
- Finally, the findings support several actionable recommendations. Inspections should be prioritized in neighborhoods with higher average scores, targeted training should be developed for restaurants with repeated violations, and standardized data entry practices should be enforced to reduce placeholder and zero entries. With these improvements, the cleaned dataset provides a strong foundation for public health interventions and more strategic allocation of inspection resources.

14. CONCLUSION

This study analyzed NYC restaurant inspection data by first cleaning the dataset and then exploring patterns within it. The results highlighted a clear connection between inspection scores and letter grades, noticeable differences across boroughs, and repeated violations among some restaurants. To maintain accuracy, data quality issues such as placeholder dates, zero values, and duplicate entries were carefully handled using conservative imputation and sentinel values. Statistical tests confirmed the significance of these differences, pointing to the need for targeted measures like staff training and stricter follow-up inspections in areas with lower compliance. Looking ahead, a deeper time-series analysis, causal modeling of policy impacts, and the inclusion of neighborhood socioeconomic data could provide even stronger insights into the root causes of non-compliance.