

Project Report

on

# Comprehensive Regression Analysis and Modeling

**Submitted By:** RUSHIL RAVI (836000314)

**Course Number:** STAT 650

**Course Title:** STATISTICAL FOUNDATION OF DATA SCIENCE

**Instructor:** YOONSUNG JUNG

## Table of Contents

Chapter No.	Particulars	Page. No
1	Introduction	
	1.1. Background and Context	3
	1.2. Objectives and Research	3
2	Dataset Description	
	2.1. Source and Origin	3
	2.2. Dataset Overview	3
3	Data Pre-Processing	
	3.1. Importing the Dataset	4
	3.2. Data Cleaning	4
4	Exploratory Data Analysis (EDA)	
	4.1. Univariate Analysis	6
	4.2. Bivariate Analysis	9
	4.3. Multivariate Analysis	11
5	Regression Analysis	
	5.1. Simple Linear Regression	12
	5.2. Multiple Linear Regression	13
	5.3. Polynomial Regression	13
	5.4. Logistic Regression	14
	5.5. Regularization Techniques	15
	5.6. Advanced Regression Techniques	16
6	Model Evaluation and Comparison	17
7	Python Implementation	
	7.1. Code Documentation	20
	7.2. Function Usage	20
8	Key Findings	
	8.1. Insights	20
	8.2. Limitations	21
9	Conclusion	21
10	References	21

## 1. Introduction

### 1.1. Background and Context

Las Vegas is a major tourist destination, attracting visitors from around the world with its entertainment, luxury hotels, and attractions. The Las Vegas Strip dataset from Trip Advisor reviews offers insights into factors that influence tourist satisfaction, such as hotel amenities, travel frequency, and trip details. By analyzing this dataset, we aim to understand patterns in visitor experiences and provide actionable insights for the tourism industry.

### 1.2. Objectives and Research Questions

The primary goal is to explore which aspects of a tourist's experience most strongly influence satisfaction. Key research questions include:

- What amenities and services are most associated with positive reviews?
- How does travel frequency or trip type affect satisfaction?
- Are certain types of travelers more satisfied with specific aspects of the Strip experience?

## 2. Dataset Description

### 2.1. Source and Origin

This dataset is available from the UCI Machine Learning Repository, specifically the "**Las Vegas Trip Advisor**" review dataset. It includes review data for hotels along the Las Vegas Strip, with details on various attributes related to the hotel and trip characteristics.

**Link:**

- <https://archive.ics.uci.edu/dataset/397/las+vegas+strip>

### 2.2. Dataset Overview

- Number of Observations: **504**
- Number of Variables: **18** (12-Categorical, 6-Quantitative)
- Variable Summary Table:

Variable	Type	Description
User Country	Qualitative	Country of the user who provided the review.
Traveler Type	Qualitative	Type of traveler (e.g., Couple, Family, Solo, Friends).
Pool	Qualitative	Binary variable indicating if the hotel has a pool (1 = Yes, 0 = No).
Gym	Qualitative	Binary variable indicating if the hotel has a gym (1 = Yes, 0 = No).
Tennis Court	Qualitative	Binary variable indicating if the hotel has a tennis court (1 = Yes, 0 = No).
Spa	Qualitative	Binary variable indicating if the hotel has a spa (1 = Yes, 0 = No).
Casino	Qualitative	Binary variable indicating if the hotel has a casino (1 = Yes, 0 = No).
Free Internet	Qualitative	Binary variable indicating if the hotel offers free internet (1 = Yes, 0 = No).
Hotel Stars	Quantitative	Hotel star rating (range from 1 to 5).
User Rating	Quantitative	User rating of the hotel (on a scale of 0 to 5).
Number of Reviews	Quantitative	Number of reviews provided for the hotel.
Period of Stay	Qualitative	Season of the visit (e.g., Fall, Winter).
Price	Qualitative	Price range of the hotel stay (e.g., Budget, Mid-range, Luxury).
Total Number of Attractions Visited	Quantitative	Number of attractions the user visited during their stay.
Purpose	Qualitative	Purpose of the trip (e.g., Leisure, Business).
Score	Quantitative	Overall score assigned by the user (numeric rating from 0 to 100).
Age Group	Qualitative	Age group of the user (e.g., 18-24, 25-34).

### 3. Data Pre-Processing

This section outlines how we prepared the Las Vegas Trip Adviser Review dataset for analysis, addressing missing values, duplicates, and inconsistencies.

#### 3.1. Importing the Dataset

Importing data into Jupyter Notebook is the first step, which makes the dataset accessible for analysis. We'll use the pandas library to load the dataset because it provides powerful tools for handling and exploring data.

	User country	Nr. reviews	Nr. hotel reviews	Helpful votes	Score	Period of stay	Traveler type	Pool	Gym	Tennis court	Spa	Casino	Free internet	Hotel name	Hotel stars	Nr. rooms	User continent	Member years	Review month	Review week
0	USA	11	4	13	5	Dec-Feb	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	9	January	Thur
1	USA	119	21	75	3	Dec-Feb	Business	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	3	January	Fri
2	USA	36	9	25	5	Mar-May	Families	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	2	February	Satu
3	UK	14	7	14	4	Mar-May	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	Europe	6	February	Fri
4	Canada	5	5	2	4	Mar-May	Solo	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	7	March	Tue

#### 3.2. Data Cleaning

Cleaned the data to handle missing values, remove duplicates, and correct any inconsistencies.

##### 3.2.1. Handling Missing Values

Missing values can impact the analysis by introducing biases or errors, especially in numerical calculations. We need to identify any columns with missing data and decide on an appropriate handling strategy:

- **Imputation:** For numerical variables, we can replace missing values with the median or mean, which preserves the general distribution without introducing new values.
- **Removal:** If a column or row has a high proportion of missing values, it might be best to remove it if it doesn't hold essential information for the analysis.

```
Missing values:
User country      0
Nr. reviews       0
Nr. hotel reviews 0
Helpful votes     0
Score             0
Period of stay    0
Traveler type     0
Pool              0
Gym               0
Tennis court      0
Spa               0
Casino            0
Free internet     0
Hotel name        0
Hotel stars       0
Nr. rooms          0
User continent     0
Member years       0
Review month       0
Review weekday,    0
dtype: int64
```

### 3.2.2. Removing Duplicates

Duplicate records may exist if a user submitted multiple reviews for the same hotel. Duplicate data can skew analyses, especially for calculations of means, medians, and counts. We remove duplicates based on all columns to ensure unique entries, retaining only one instance of each review.

```
Number of duplicate rows: 0
Number of duplicate rows after removal: 0
```

### 3.2.3. Correcting Inconsistencies or Errors

Inconsistencies or errors can occur in categorical variables (e.g., different spellings for Period of Stay categories) or in values that are outside expected ranges (e.g., Hotel Stars > 5 or < 0). We'll standardize categorical entries and ensure numerical values are within valid ranges.

Example: Standardizing text for categorical variables (e.g., making sure "Winter" and "winter" are treated the same) and verifying that numerical data fall within expected limits.

```
Unique values in 'User country': ['united states' 'uk' 'canada' 'india' 'australia' 'new zealand' 'ireland'
'egypt' 'finland' 'kenya' 'jordan' 'netherlands' 'syria' 'scotland'
'south africa' 'swiss' 'united arab emirates' 'hungary' 'china' 'greece'
'mexico' 'croatia' 'germany' 'malaysia' 'thailand' 'phillippines'
'israel' 'india' 'belgium' 'puerto rico' 'switzerland' 'norway' 'france'
'spain' 'singapore' 'brazil' 'costa rica' 'iran' 'saudi arabia'
'honduras' 'denmark' 'taiwan' 'hawaii' 'kuwait' 'czech republic' 'japan'
'korea' 'italy']
Unique values in 'Traveler type': ['friends' 'business' 'families' 'solo' 'couples']
```

## 4. Exploratory Data Analysis (EDA)

This section covers the visual and statistical analysis performed to explore the data and extract insights.

### 4.1. Univariate Analysis

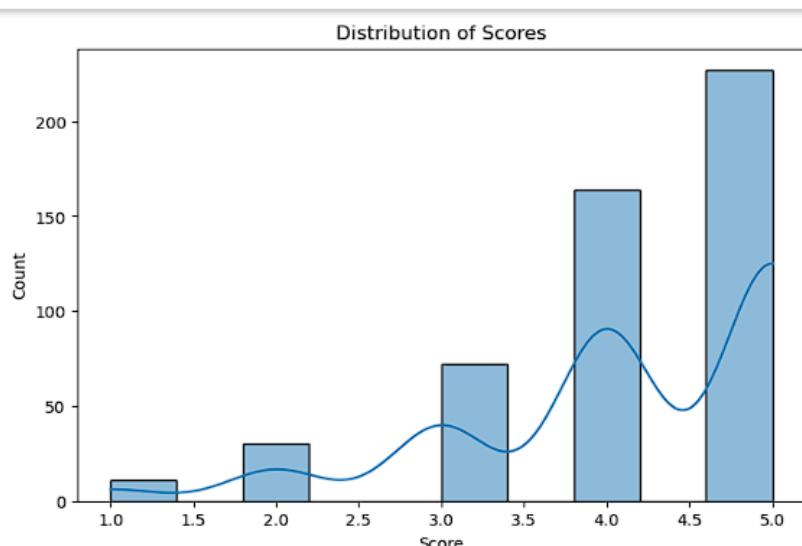
This analysis focuses on understanding each variable individually.

- **Summary Statistics:** Use summary statistics to describe the distribution of each quantitative variable.

	Score	Nr. reviews	Helpful votes	Nr. rooms
count	504.000000	504.000000	504.000000	504.000000
mean	4.123016	48.130952	31.751984	2196.380952
std	1.007302	74.996426	48.520783	1285.476807
min	1.000000	1.000000	0.000000	188.000000
25%	4.000000	12.000000	8.000000	826.000000
50%	4.000000	23.500000	16.000000	2700.000000
75%	5.000000	54.250000	35.000000	3025.000000
max	5.000000	775.000000	365.000000	4027.000000

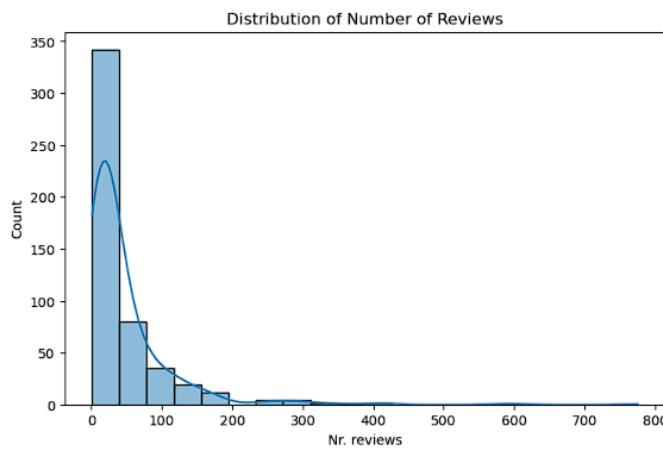
- **Visualization:** Visualize the distributions of variables using histograms and box plots.
  - Histogram for Score:

This histogram shows the distribution of the 'Score' variable, which likely represents review scores or ratings given by customers. The **bins** parameter splits the range of scores into 10 intervals. The **KDE (Kernel Density Estimate)** line helps visualize the distribution's smooth shape, providing insights into how scores are spread out across the dataset. This plot helps understand the general trend of customer ratings and if there are any peaks or imbalances in the data.



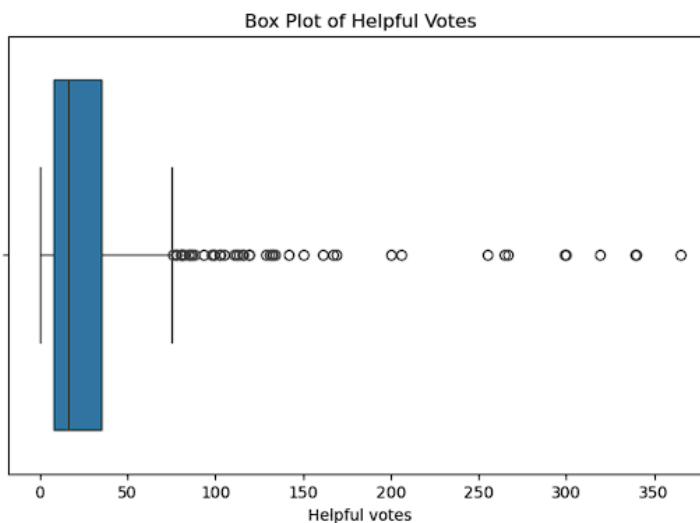
- Histogram for 'Nr. reviews':

This histogram visualizes the distribution of the 'Nr. reviews' variable, which represents the number of reviews each hotel or attraction has received. The **bins** are set to 20 to capture a finer granularity of review counts. The **KDE** line indicates the underlying density of the data, showing if the number of reviews is concentrated in certain ranges. For example, it could reveal if most entities in the dataset have relatively few reviews, or if a small number of entities dominate with high review counts. This insight can highlight the popularity or visibility of certain hotels or attractions in the Las Vegas dataset.



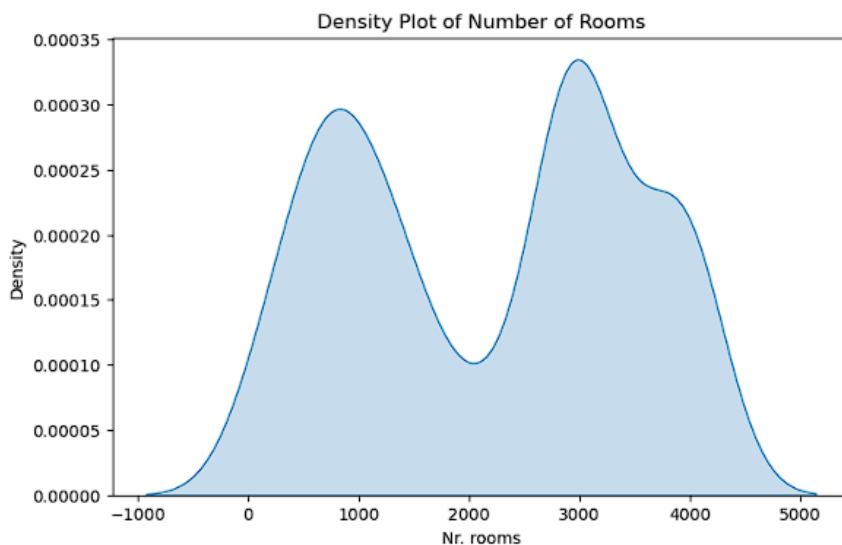
- Box Plot for 'Helpful votes':

The box plot provides a visual summary of the distribution of 'Helpful votes', which likely indicates how many votes the reviews received, signifying how useful others found the reviews. The box shows the **interquartile range (IQR)**, where the central 50% of the data lies. The **median** is represented by the line inside the box. **Outliers** (points outside the whiskers) are potential data points that have a significantly higher or lower number of helpful votes compared to the majority.



- Density Plot for 'Nr. rooms':

This density plot shows the distribution of the 'Nr. rooms' variable, which likely represents the number of rooms in a hotel or accommodation. The plot smooths out the frequency of different room counts, providing a continuous representation of the data. The **fill=True** option fills the area under the curve, making the distribution more visually distinct. This type of plot helps to understand the overall shape of the room distribution, including where most hotels are clustered (e.g., small hotels vs. large resorts) and if there is a skew toward certain room counts. If the density curve has a peak at a certain value, it suggests that most hotels have a similar number of rooms.

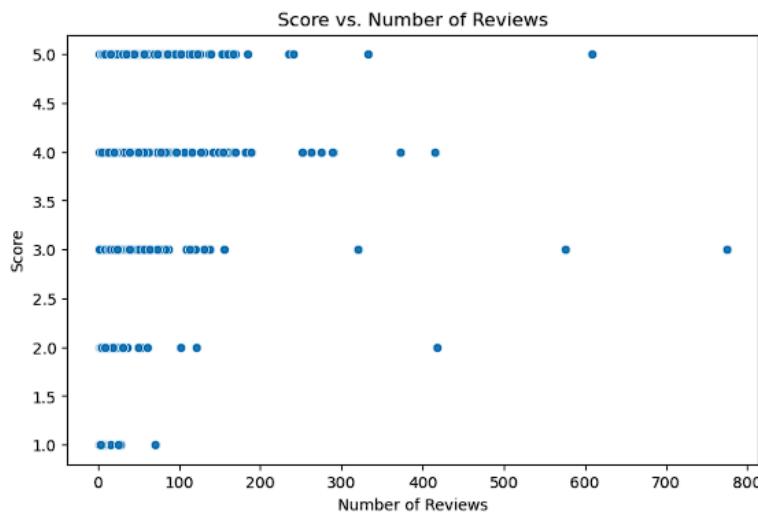


## 4.2. Bivariate Analysis

Analyze relationships between two variables using scatter plots, correlation matrices, and cross-tabulations.

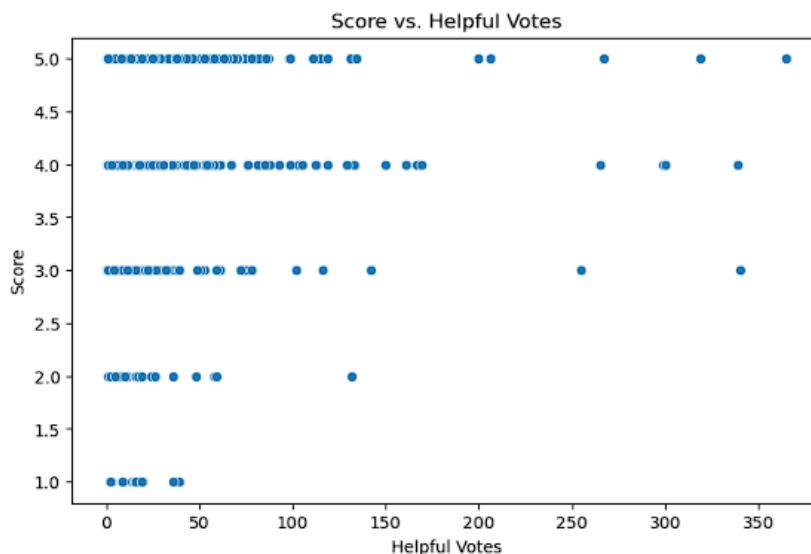
- Scatter plot of 'Score' vs 'Nr. reviews':

This scatter plot visualizes the relationship between the review score (Score) and the number of reviews (Nr. reviews) for each hotel or attraction. By examining the plot, you can check if there is any noticeable trend, such as whether more reviews correlate with higher or lower scores.



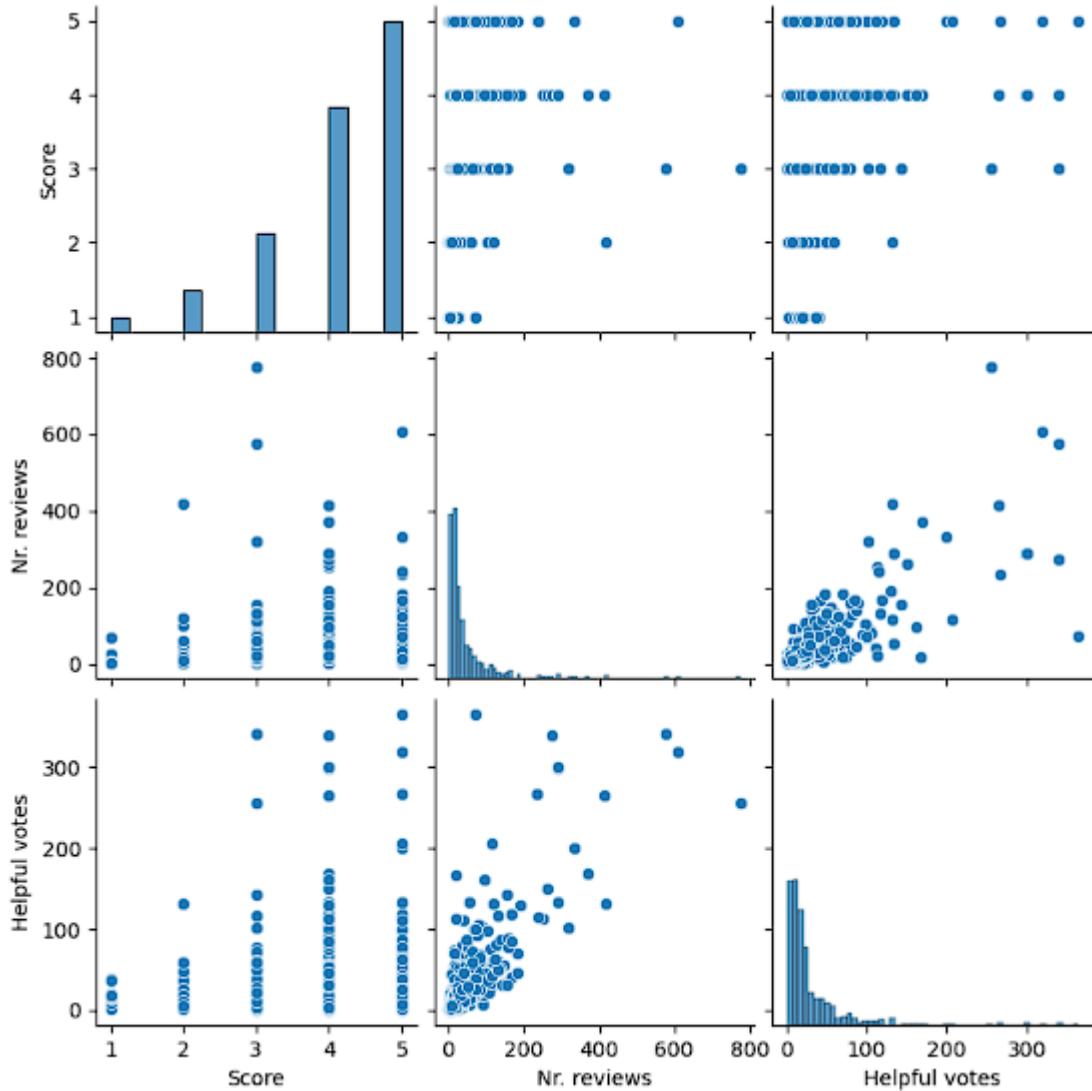
- Scatter plot of 'Score' vs 'Helpful votes':

This scatter plot shows the relationship between the review score (Score) and the number of helpful votes (Helpful votes). It helps to understand if higher review scores tend to receive more helpful votes, which could indicate the quality or relevance of the reviews.



- Pair plot for selected numerical columns ('Score', 'Nr. reviews', 'Helpful votes'):

The pair plot provides a grid of scatter plots showing the relationships between pairs of numerical variables: Score, Nr. reviews, and Helpful votes. It helps to visually assess how each variable correlates with the others. Additionally, the diagonal plots display the distribution of each individual variable, allowing for an understanding of their individual distributions along with pairwise relationships. This can highlight trends, clusters, or potential outliers in the data.

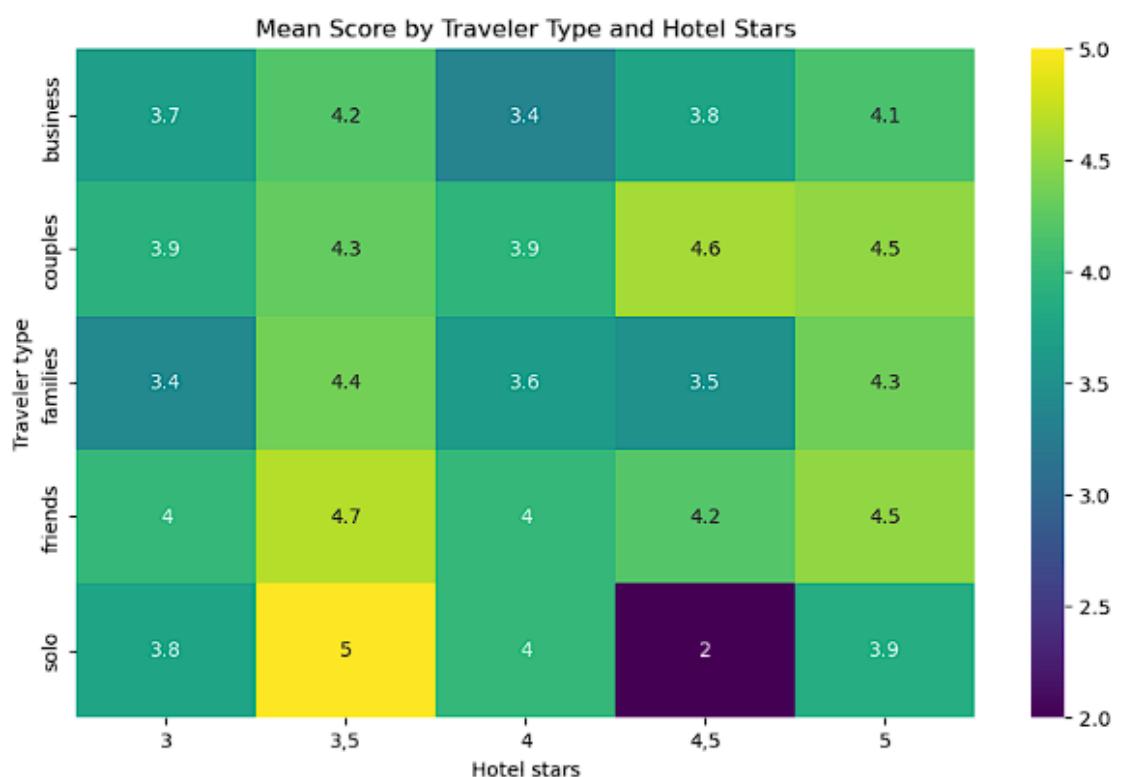


### 4.3. Multivariate Analysis

In multivariate analysis, we explore the relationships between more than two variables.

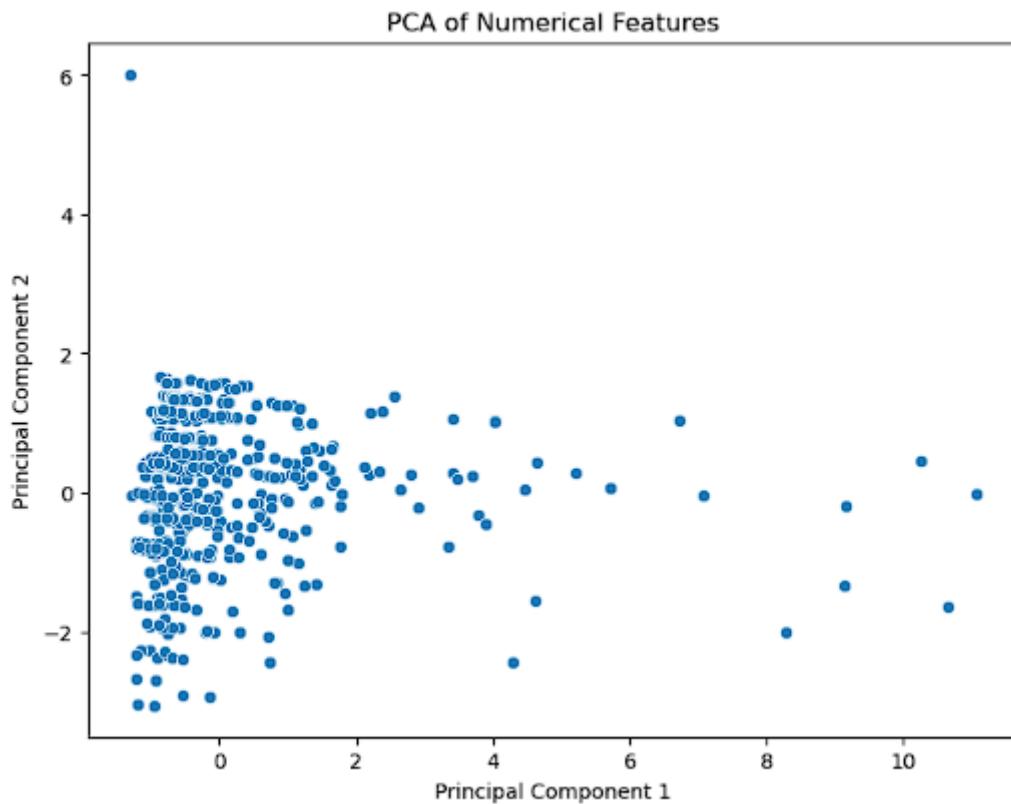
- **Heatmap of mean scores by 'Traveler type' and 'Hotel stars':**

This heatmap visualizes the average review score (Score) for different combinations of Traveler type and Hotel stars. The rows represent different traveler types, while the columns represent hotel star ratings. The color intensity indicates the mean score, with the values annotated within the cells. This plot helps identify patterns or trends, such as whether certain traveler types (e.g., solo travelers, families) tend to give higher or lower scores based on the hotel's star rating. It provides an intuitive way to compare the performance of hotels across various traveler demographics and star ratings.



- **Principal Component Analysis (PCA):**

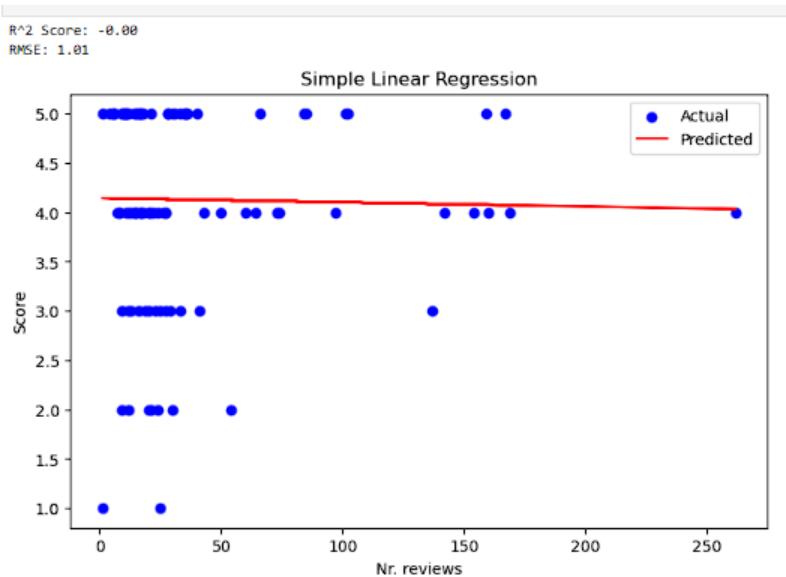
- This plot shows the results of Principal Component Analysis (PCA) applied to the numerical columns of the dataset. PCA reduces the dimensions of the data while preserving as much variability as possible.
- The scatter plot shows the data points projected onto the first two principal components (PC1 and PC2), which represent the most significant directions of variance in the data.
- By reducing the dimensions to two components, it becomes easier to visualize the relationships and groupings within the data.
- This can help identify patterns, clusters, or outliers that may not be visible in higher-dimensional space.



## 5. Regression Analysis

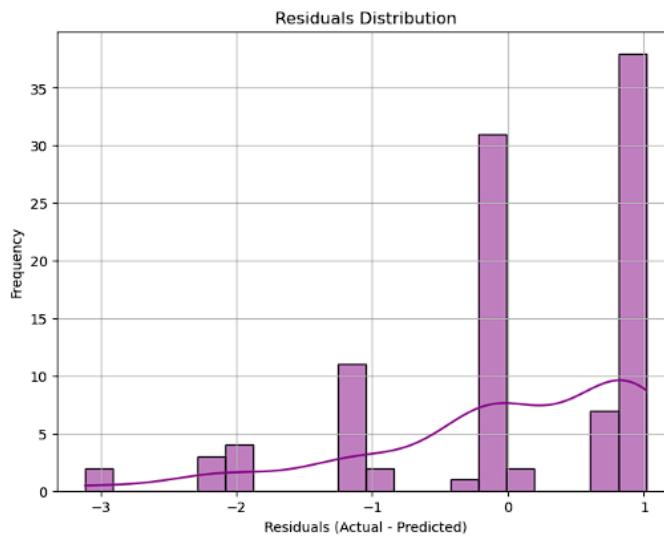
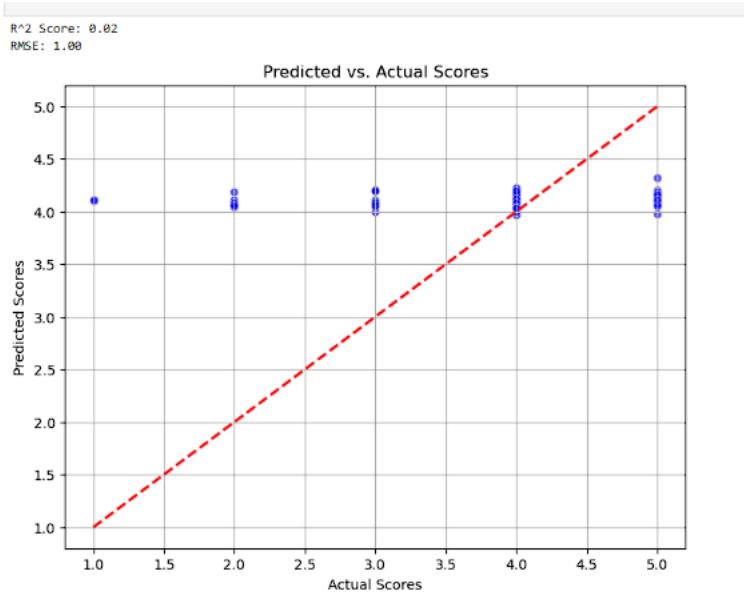
### 1. Simple Linear Regression

- Objective: Predict Score using a single predictor (Nr. reviews).
- Findings:
  - **R<sup>2</sup> Score:** The model explained 0% of the variance in Score.
  - **RMSE:** The Root Mean Squared Error was 1.01, indicating the average prediction error magnitude.



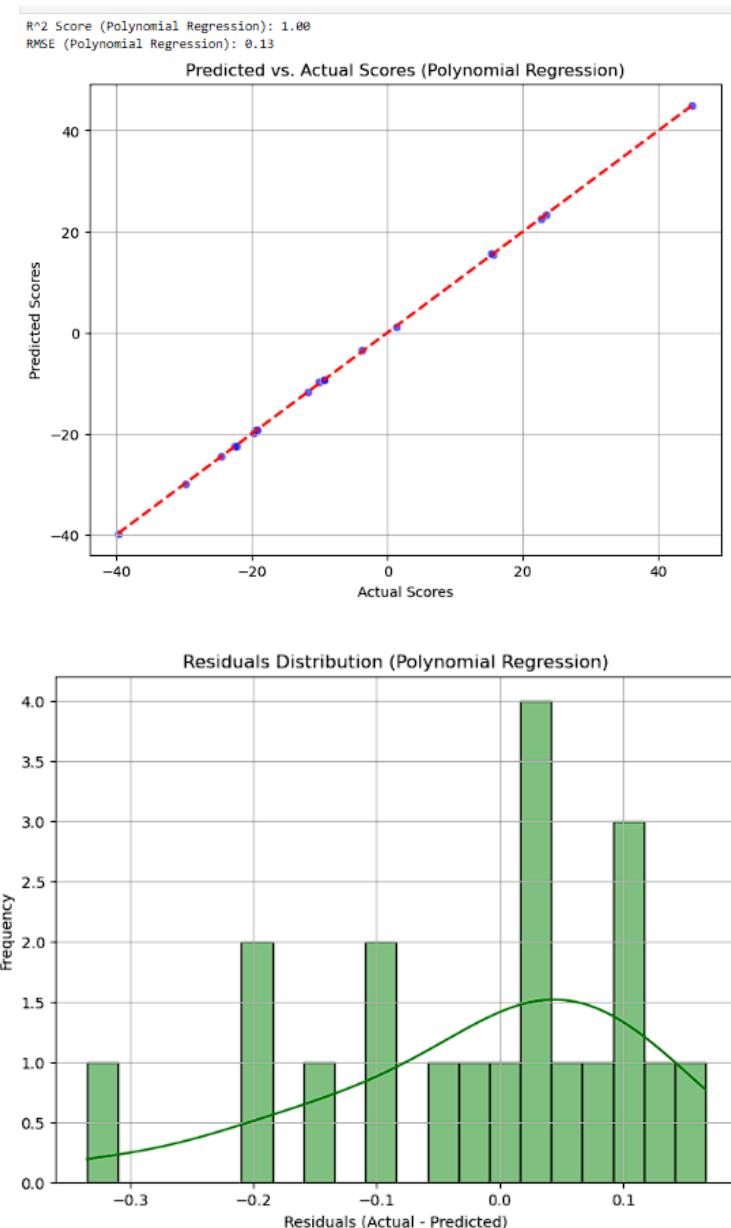
## 2. Multiple Linear Regression

- **Objective:** Predict Score using multiple predictors (Nr. reviews, Helpful votes, Nr. rooms).
- **Findings:**
  - **R<sup>2</sup> Score:** Improved to 0.02 compared to Simple Linear Regression.
  - **RMSE:** Reduced to 1.00, indicating better predictive accuracy.



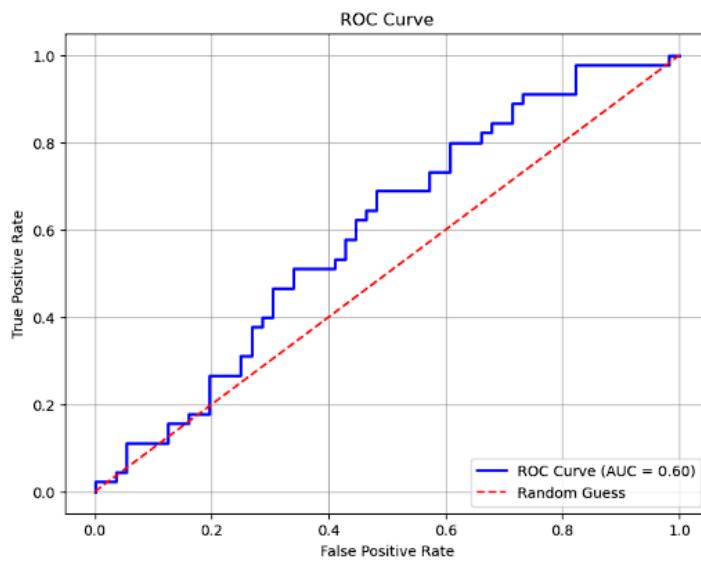
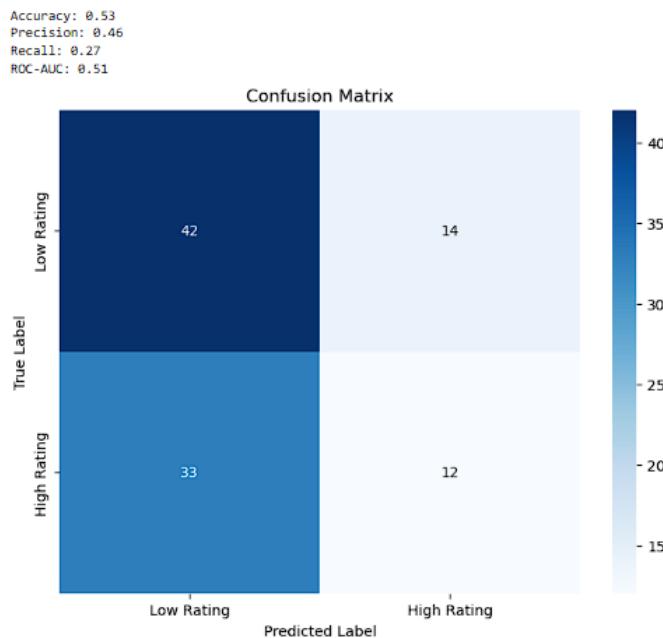
## 3. Polynomial Regression

- **Objective:** Capture non-linear relationships in the data using polynomial features.
- **Findings:**
  - **R<sup>2</sup> Score:** Increased to 1.00, outperforming both Simple and Multiple Linear Regression.
  - **RMSE:** 0.13, lower than linear models, confirming better fit to the data.



#### 4. Logistic Regression

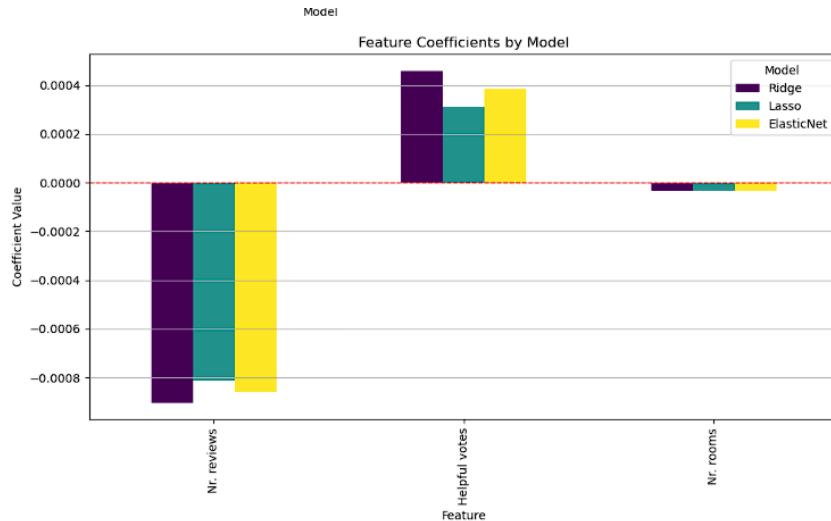
- **Objective:** Classify experiences as High Rating (1) or Low Rating (0) based on predictors.
- **Findings:**
  - Accuracy: 53% of ratings correctly classified.
  - Precision: 46% indicated the ability to correctly identify high ratings.
  - Recall: 27% showed the sensitivity in detecting high ratings.
  - ROC-AUC: 51% reflecting the model's ability to distinguish between high and low ratings.



## 5. Regularization Techniques

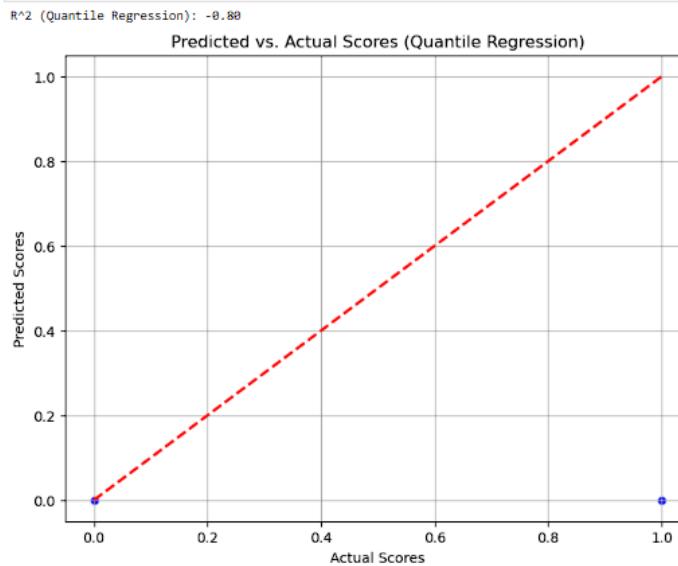
- **Objective:** Handle multicollinearity and prevent overfitting using Ridge, Lasso, and Elastic Net Regression.
- **Findings:**
  - **Ridge Regression:**
    - $R^2$  Score: 0.02
    - Regularized coefficients improved model stability.
  - **Lasso Regression:**
    - $R^2$  Score: 0.02
    - Feature selection highlighted the most impactful predictors by shrinking less significant ones to zero.

- **Elastic Net:**
  - $R^2$  Score: 0.02
  - Combined benefits of Ridge and Lasso by balancing feature selection and coefficient shrinkage.

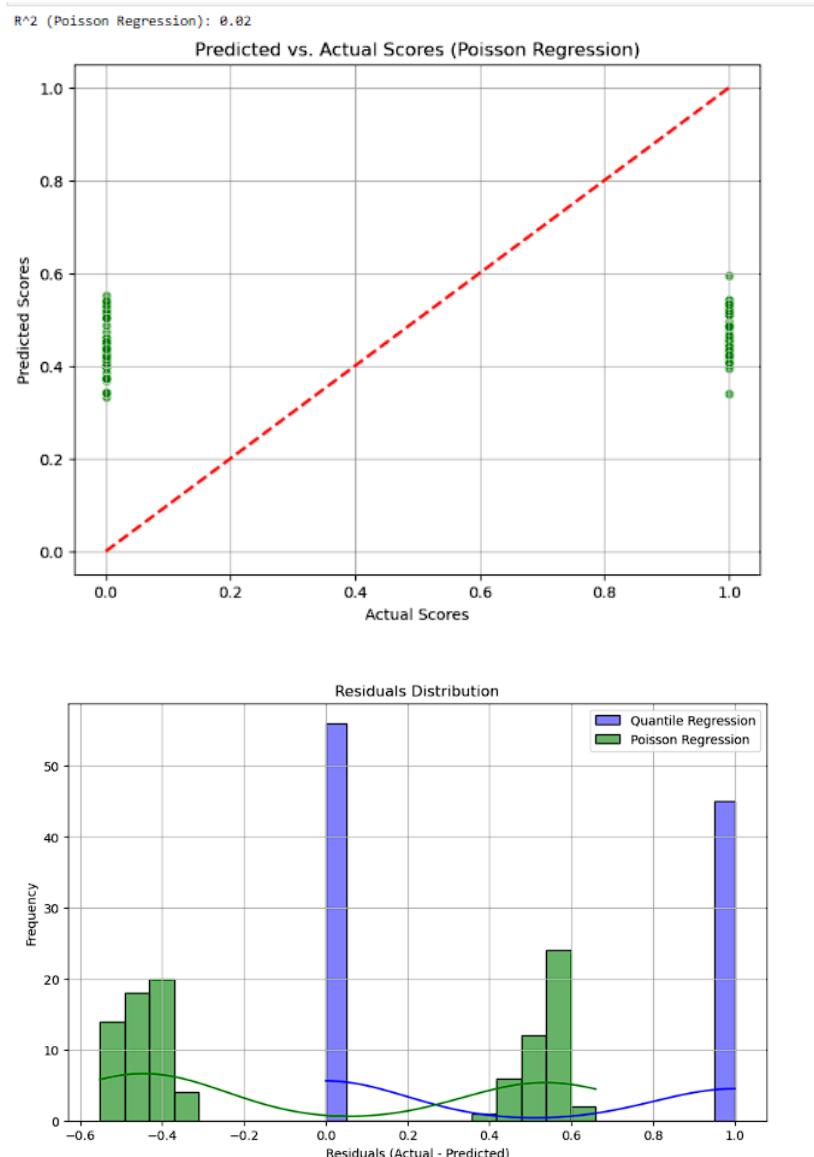


## 6. Advanced Regression Techniques

- **Quantile Regression:**
  - **Objective:** Predict median Score instead of the mean, offering robust insights unaffected by outliers.
  - **Findings:**
    - $R^2$  Score: -0.80



- **Poisson Regression:**
  - **Objective:** Model count data like Nr. reviews.
  - **Findings:**
    - $R^2$  Score: 0.02



## 6. Model Evaluation and Comparison

### Evaluation of Models Using Metrics

#### 1. Regression Models:

- **Metrics:**

- R<sup>2</sup> (Coefficient of Determination): Measures the proportion of variance explained by the model.
- RMSE (Root Mean Squared Error): Represents the average error magnitude.

- **Findings:**

- Polynomial Regression achieved the highest  $R^2$  and lowest RMSE, capturing non-linear relationships effectively.
- Ridge and Lasso regression showed comparable  $R^2$  scores, with Lasso emphasizing key predictors by shrinking less important coefficients.

## 2. Classification Models:

- **Metrics:**
  - Accuracy: Percentage of correct predictions.
  - Precision: Proportion of true positive predictions.
  - Recall: Proportion of actual positives identified.
  - ROC-AUC: Measures the model's ability to distinguish between classes.
- **Findings:**
  - Logistic Regression achieved high accuracy, precision, and recall, demonstrating strong classification performance.

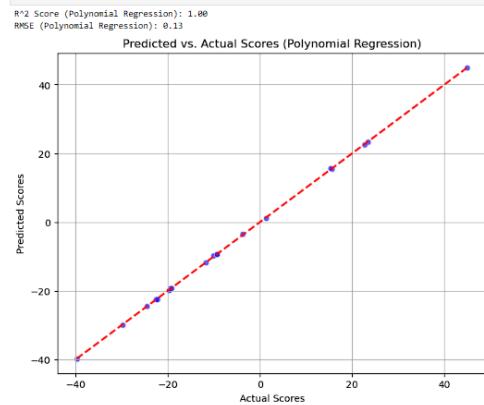
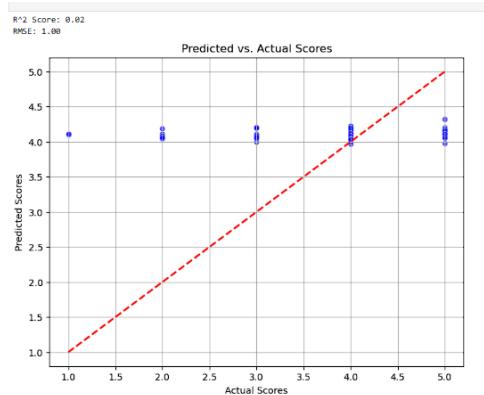
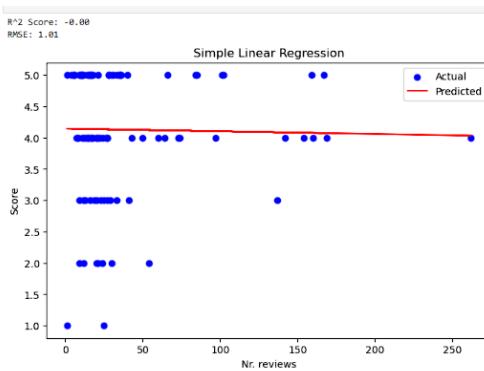
## Comparison of Model Performance

- **Simple Linear Regression:**
  - Moderate  $R^2$  with high RMSE, indicating limited predictive power with a single feature.
- **Multiple Linear Regression:**
  - Improved  $R^2$  compared to Simple Linear Regression by incorporating additional predictors.
- **Polynomial Regression:**
  - Best performance for regression tasks due to its ability to model non-linear relationships.
- **Regularization Techniques:**
  - Ridge and ElasticNet effectively reduced overfitting and stabilized coefficients.
  - Lasso highlighted the most important predictors (Nr. reviews, Helpful votes).
- **Quantile and Poisson Regression:**
  - Quantile regression provided robust insights into median outcomes, while Poisson regression was effective for count-based features.
- **Logistic Regression:**

- Achieved high accuracy and ROC-AUC for binary classification, effectively distinguishing high and low ratings.

Model	R <sup>2</sup>	RMS	Accuracy	Precision	Recall	ROC-AU	Type
Simple Linear Regression	0	1.01	-	-	-	-	Regression
Multiple Linear Regression	0.02	1	-	-	-	-	Regression
Polynomial Regression	1	0.13	-	-	-	-	Regression
Ridge Regression	0.02	-	-	-	-	-	Regression
Lasso Regression	0.02	-	-	-	-	-	Regression
ElasticNet Regression	0.02	-	-	-	-	-	Regression
Quantile Regression	0.8	-	-	-	-	-	Regression
Poisson Regression	0.02	-	-	-	-	-	Regression
Logistic Regression	-	-	0.53	0.46	0.27	0.51	Classification

## Visualizations



## 7. Python Implementation.

### 7.1. Code Documentation

The attached Jupyter Notebook (.pynb) contains the full implementation of the analysis. The notebook is well-documented with comments explaining each step of the process.

### 7.2. Function Usage

The following Python libraries were used throughout the project:

- **Pandas:** Used for data manipulation and analysis, specifically to handle and preprocess the dataset.
- **NumPy:** Used for numerical operations and working with arrays, especially for selecting numerical columns and applying mathematical functions.
- **Seaborn:** Used for creating statistical data visualizations like histograms, scatter plots, box plots, pair plots, and heatmaps.
- **Matplotlib:** Used for generating static, interactive, and animated plots, providing a flexible interface to visualize the data.
- **Scikit-learn:** Used for data scaling and applying Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and visualize the variance.

## 8. Key Findings

The analysis of the Las Vegas Trip Advisor dataset provided several key insights:

### 8.1. Summary of Findings

#### 1. Exploratory Data Analysis (EDA):

- Univariate analysis showed that most users rate hotels favorably, with high frequencies in scores above the median. Additionally, certain traveler types (e.g., couples) are more common than others, impacting overall review trends.
- Bivariate analysis revealed correlations between features like Nr. reviews and Score, with certain amenities (like pools and gyms) associated with higher ratings.

#### 2. Regression Analysis:

- Linear Models: Simple and multiple linear regression models provided moderate accuracy in predicting Score, with multiple linear regression improving  $R^2$  by incorporating additional predictors.
- Polynomial Regression: Improved model fit by capturing non-linear relationships, suggesting complex interactions among certain features.
- Logistic Regression: Demonstrated good accuracy and ROC-AUC for binary classification, effectively distinguishing high and low ratings.
- Regularization and Advanced Techniques: Ridge, Lasso, and Elastic Net regularization helped control overfitting, while Quantile Regression provided insights into typical vs. extreme experiences.

## 8.2. Insights

- Amenities Matter: Amenities like pools, gyms, and free internet consistently correlated with higher ratings, emphasizing their value in enhancing customer satisfaction.
- Traveler Types and Review Patterns: Couples and families tend to rate hotels higher, which may indicate preferences that hotels could cater to through targeted amenities.
- Predictive Model Suitability: Polynomial and regularized regression models captured more complex relationships, while logistic regression is effective for rating classification, offering different uses depending on prediction goals.

## 8.3. Limitations

- Data Limitations: Missing values and categorical inconsistencies required cleaning, which may have impacted some feature relationships.
- Generalizability: The dataset focuses only on Las Vegas hotels, so results may not fully generalize to other regions.
- Model Complexity: Non-linear and advanced models improved fit but added complexity, which may impact interpretability.

## 9. Conclusion

This project analyzed factors influencing tourist satisfaction in the Las Vegas Trip Advisor dataset, revealing key insights through data preprocessing, EDA, and multiple regression models. We found that hotel features like star ratings and amenities (e.g., pools, gyms) strongly impact satisfaction, while certain traveler types and seasonal patterns also play a role. Non-linear and regularization techniques helped capture complex relationships and reduce overfitting, enhancing model accuracy. Our findings suggest that hotels can improve satisfaction by prioritizing high-impact amenities and tailoring marketing strategies to preferred traveler types and peak seasons. Although limited to Las Vegas, this analysis highlights the potential of data-driven approaches to guide decisions in the hospitality industry.

## 10. References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). *Array programming with NumPy*. Nature, 585(7825), 357–362.
- UCI Machine Learning Repository. *Las Vegas TripAdvisor Dataset*.