Data Analysis and Visualization

Report :

Taiba Tariq

23i-2618

DS-A

Airbnb listings

Introduction

This project focuses on exploring and analyzing an Airbnb dataset to uncover patterns and insights related to hosts, properties, and guest experiences. The dataset contains various attributes, including host details, property characteristics, pricing information, and review scores, offering a comprehensive view of the Airbnb market dynamics.

The primary objective of this project is to perform exploratory data analysis (EDA) to understand the distribution, relationships, and trends within the data. Key variables such as *Host Id*, *Neighborhood*, *Property Type*, *Room Type*, *Price*, and *Review Scores Rating* serve as the foundation for examining how different factors influence listing performance and guest satisfaction.

Ultimately, this project contributes to a deeper understanding of Airbnb's operational and customer landscape, helping to highlight patterns that can assist hosts, customers, and platform managers in making more data-driven decisions.

Objectives

1. **Comprehensive Visual EDA:**
   Conduct detailed visual analysis to explore relationships among attributes such as price, review scores, property type, and room type. Formulate and defend hypotheses using visualizations.
2. **Graphical Integrity and Data-Ink Ratio:**
   Calculate and evaluate the Data-Ink Ratio for visualizations while assessing them against principles of graphical integrity and effective design to ensure accurate and meaningful data representation.
3. **Data Preparation and Analysis Techniques:**
   - Clean and preprocess the Airbnb dataset to handle missing or inconsistent data.
   - Perform **Univariate Analysis** to study individual feature distributions.
   - Conduct **Bivariate Analysis** to examine pairwise relationships and correlations.
   - Apply **Multivariate Analysis** techniques such as Scatter Plot Matrices, Parallel Coordinates, and Heatmaps to explore multidimensional patterns and interactions.

Here is a detailed explanation of the Data Preprocessing section for your logistic regression project:

Data Preprocessing

1. Handling Missing Values

Objective: To ensure that the dataset does not contain any NaN values or any outliers which can disrupt

model training.

Steps Taken:

● Used the .isnull().sum() function to check for null values in each column.

● NO missing values were found in our dataset but still applied:

○ For numerical columns: Missing values to be replaced using the mean or median of the column.(although none found)

○ For categorical columns: Mode could be used to fill missing values.

● In our case, it appears that there were no significant missing values, so no imputation was necessary.

2. Cleaning column names:

Spaces were removed from column names

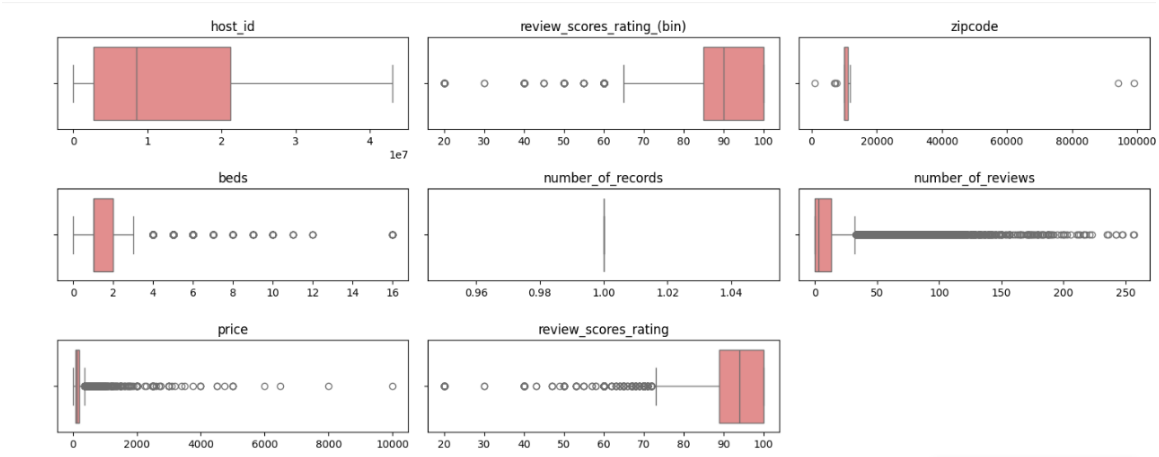
3. Outlier Detection:

Outliers are extreme values in a dataset that can negatively affect the performance of logistic regression and other machine learning models. They can bias the model.

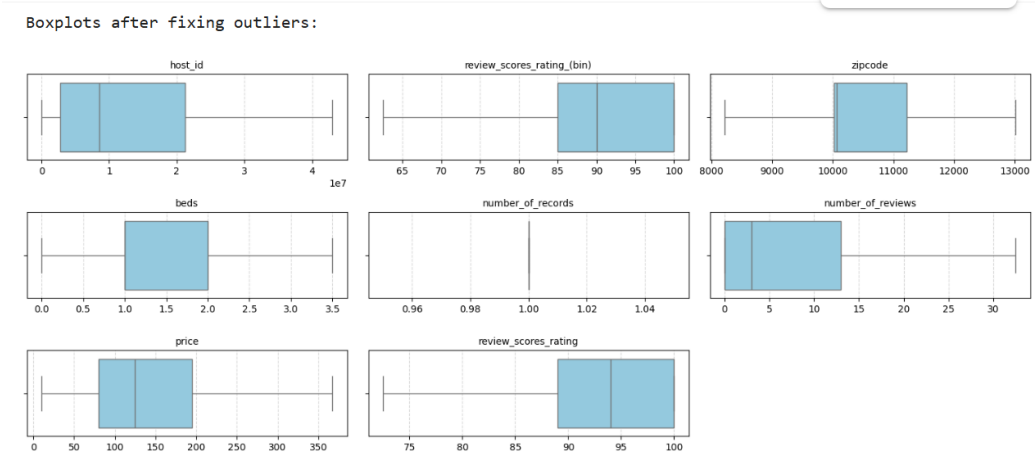
Outliers were detected using the Interquartile Range (IQR) method, a common statistical technique:

We Removed Outliers to make data set cleaner and reduce the noise Ensuring cleaner data for model training.

Following columns included outliers:



Columns after removing outliers:

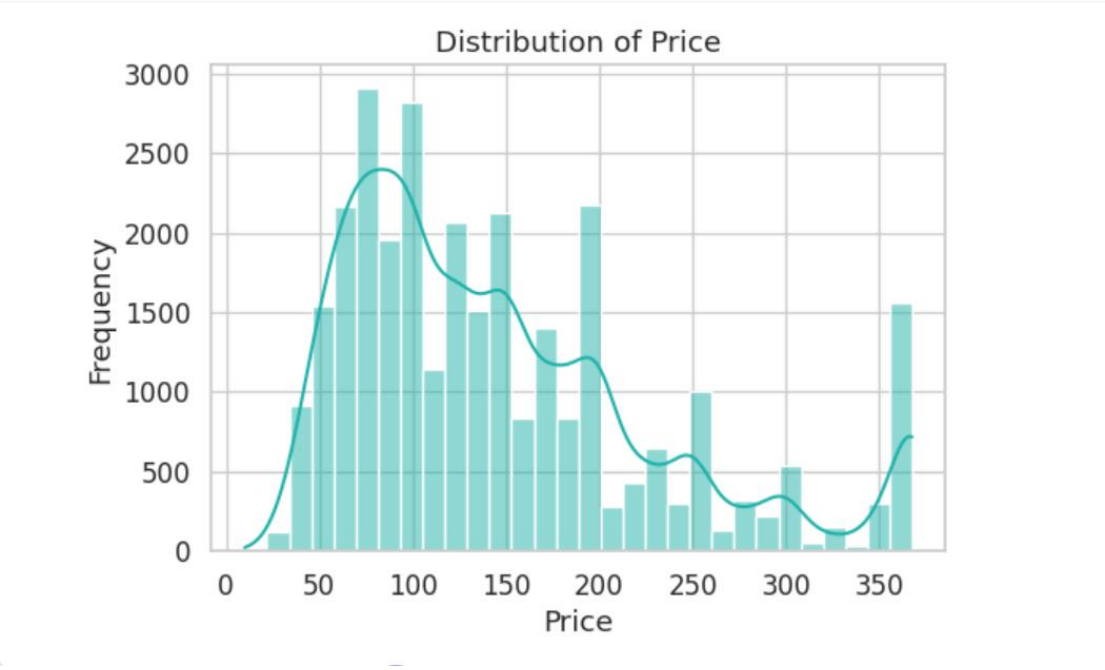


Exploratory Data Analysis:

Univariate Analysis

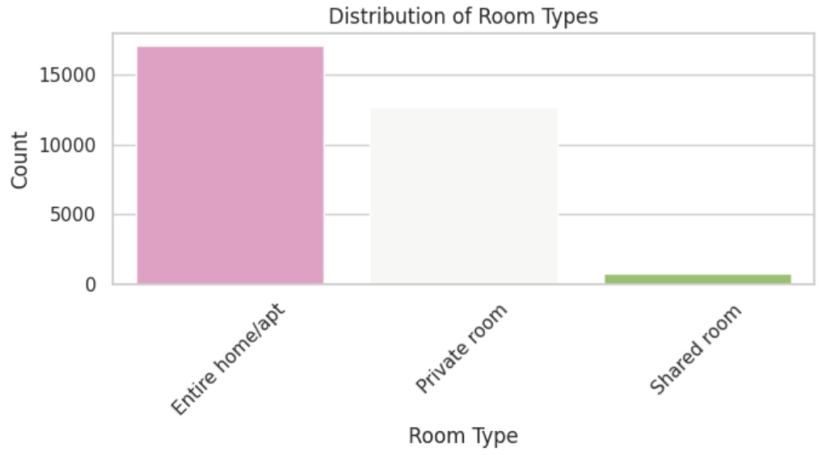
Distribution of Price:

The distribution of *Price* among Airbnb listings to understand the spread of property prices. The graph shows that most listings fall within the lower price range, indicating a right-skewed distribution. A few properties with significantly higher prices create long tails on the higher end, suggesting the presence of outliers or premium listings that influence the overall pricing trend**.**



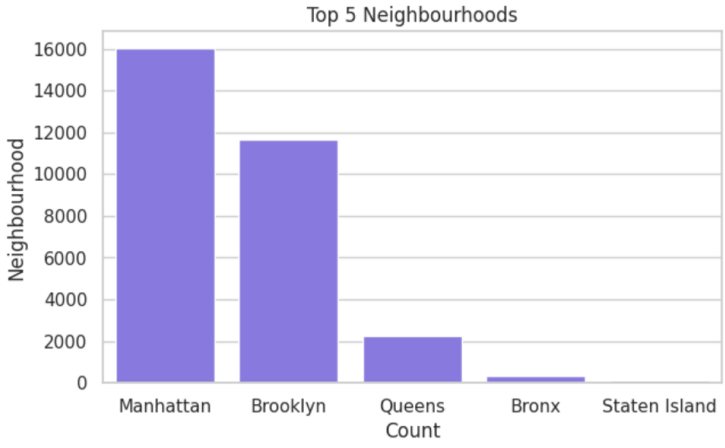
Distribution of Room Types:

This plot shows the kind of accommodation for example, *Entire home/apartment*, *Private room*, *Shared room*, or *Hotel room*. It helps in understanding how different room types affect pricing, availability, and guest preferences.



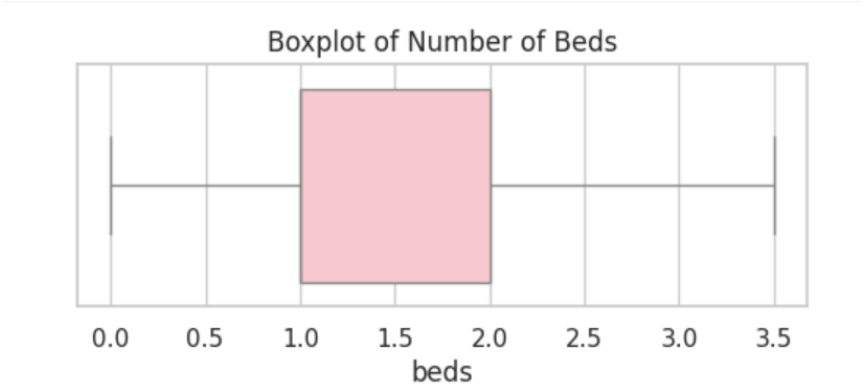
Neighborhood:

Represents the area where the listing is located, useful for identifying location-based trends in pricing and popularity



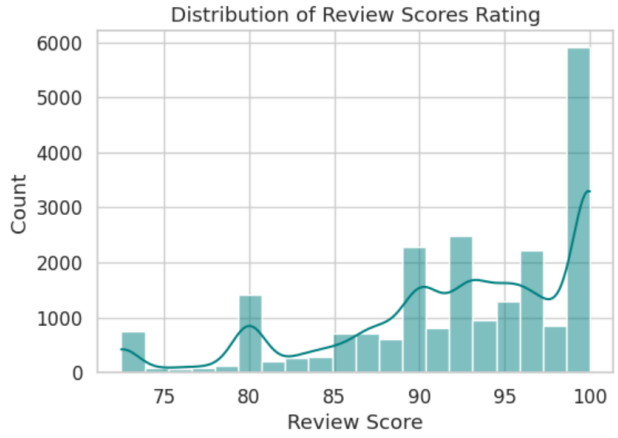
Distribution on the basis of Number of Beds:

Indicates how many beds are available in a listing. It helps assess the accommodation's capacity and analyze its impact on pricing, guest preferences, and overall demand**.**



Review Score Rating:

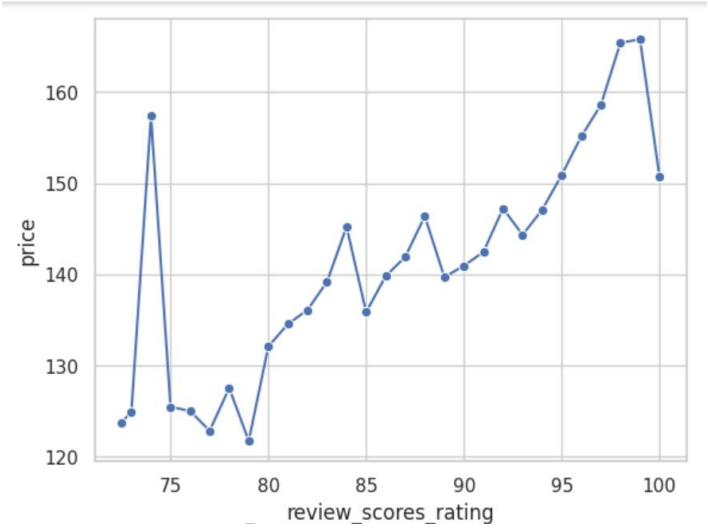
Represents the average rating given by guests based on their stay experience. It reflects overall guest satisfaction and can be used to analyze the relationship between service quality, price, and booking trends.



Bi-variate Analysis

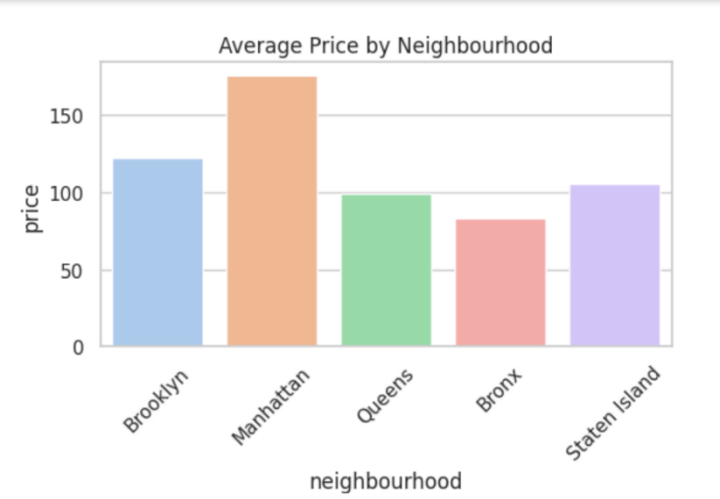
Price vs Review Score Rating:

This line plot visualizes the relationship between listing price and review score rating. It helps identify whether higher-rated listings tend to have higher or lower prices, revealing patterns between guest satisfaction and pricing strategies across different listings.



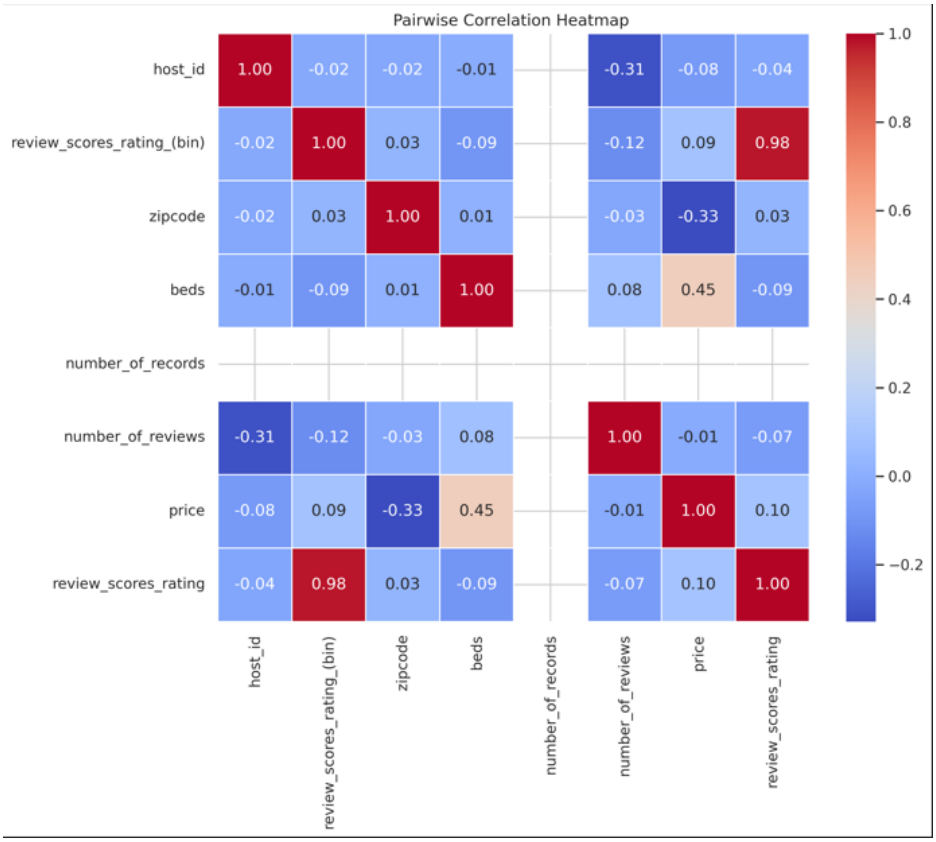
Average Price by Neighborhood:

This visualization shows the average listing price across different neighborhoods. It helps identify which areas are more expensive or affordable, providing insights into how location influences pricing patterns within the city.



Pairwise Correlation Heatmap:

This heatmap displays the correlation between all numerical variables in the dataset. It helps identify strong positive or negative relationships, revealing which factors (such as price, number of beds, or review scores) are most closely related to each other.

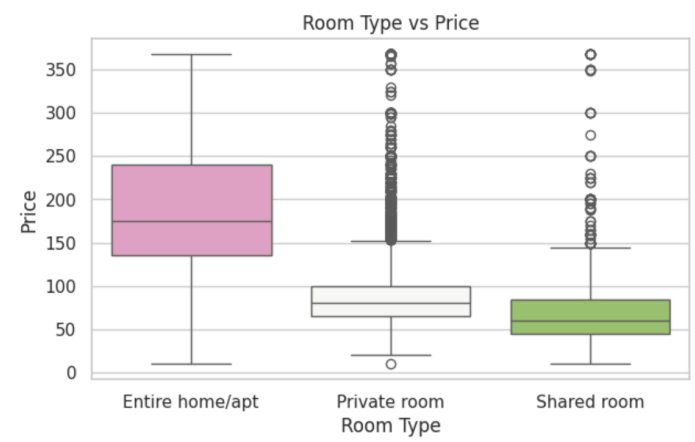


Price vs Review Score Rating (LOWESS Correlation):

This plot uses a LOWESS (Locally Weighted Scatterplot Smoothing) curve to show the trend between price and review score rating. It helps visualize whether higher-rated listings generally have higher or lower prices, highlighting any non-linear relationships between guest satisfaction and pricing**.**

Room Type vs Price:

This visualization compares the average price across different room types, such as entire homes, private rooms, and shared rooms. It helps identify how accommodation type influences pricing and highlights which room categories are typically more expensive or budget-friendly.

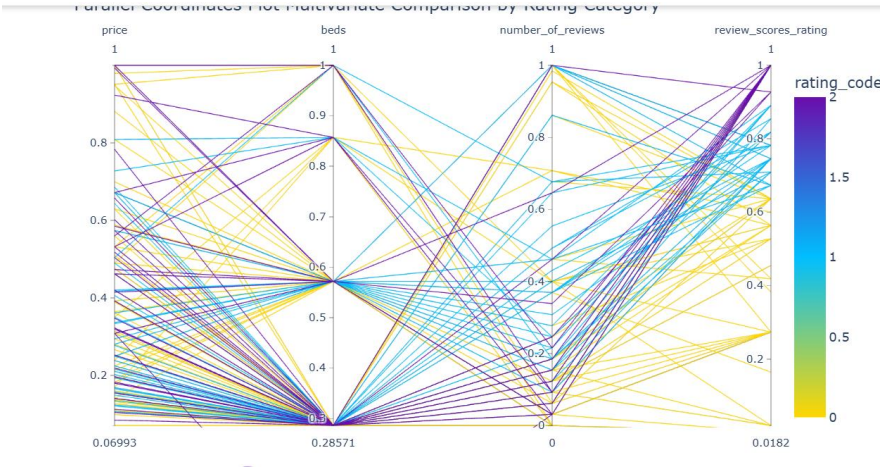


Multi-variate Analysis

Parallel Coordinates Plot :

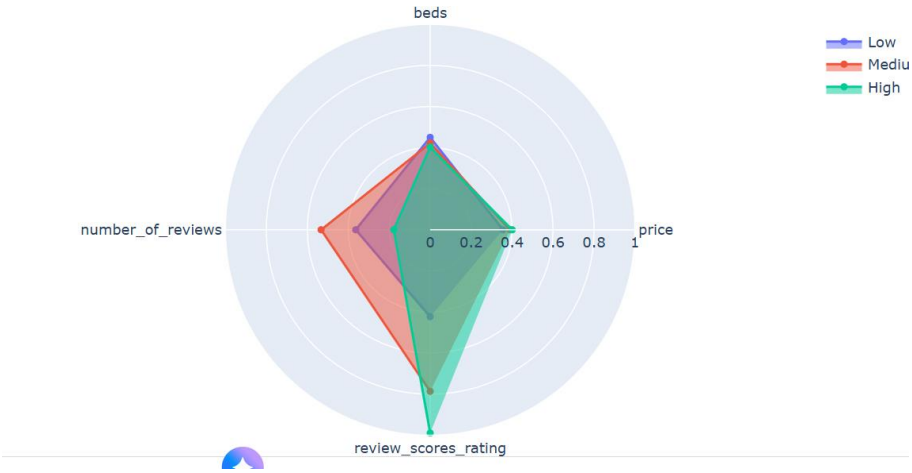
(Multivariate Comparison by Rating Category)

This plot visualizes multiple variables simultaneously, grouped by rating categories. It helps identify patterns and differences among listings with varying review scores, revealing how factors like price, number of beds, and location interact across different levels of guest satisfaction.



Glyph Chart — Multivariate Comparison by Rating Category:

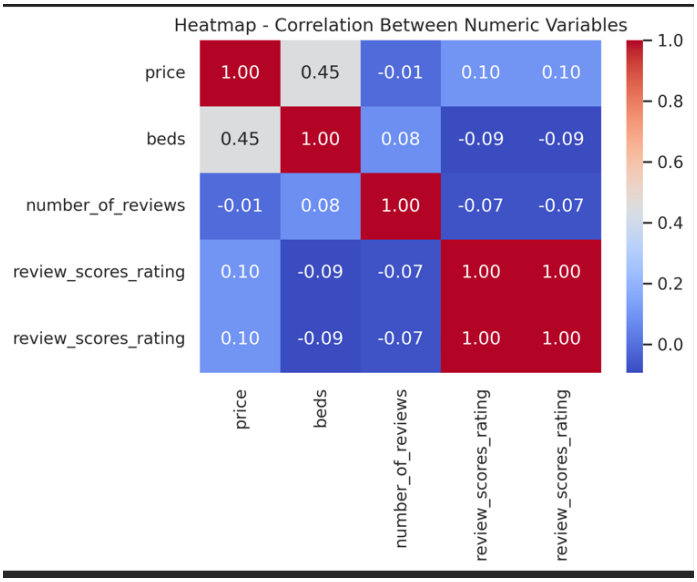
This chart represents multiple variables for each listing using visual glyphs (shapes or icons). It allows quick comparison of different features such as price, number of beds, and room type—across rating categories, helping to identify distinctive patterns among high- and low-rated listings.



Heatmap

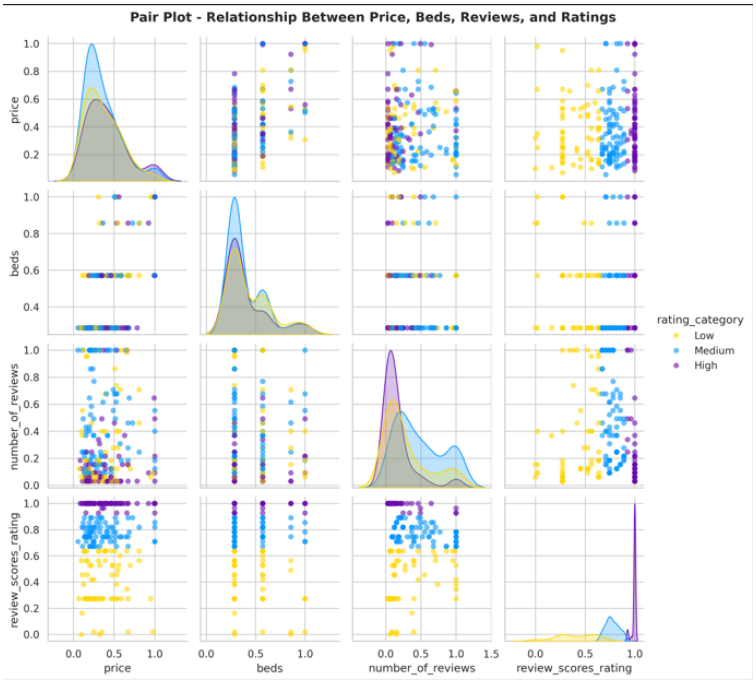
Correlation Between Numeric Variables:

This heatmap illustrates the strength and direction of relationships among all numerical variables in the dataset. It helps identify which factors, such as price, number of beds, or review scores, are closely associated, aiding in understanding underlying patterns and dependencies within the data.



Pair Plot — Relationship Between Price, Beds, Reviews, and Ratings:

This pair plot visualizes pairwise relationships among key variables such as price, number of beds, number of reviews, and review scores. It helps detect correlations, trends, and potential clusters, offering deeper insights into how these factors interact with one another**.**



Foot Ball Matches

Introduction

This report presents an exploratory data analysis of the "Football World Cup Matches (1930–2014)" dataset. The objective is to uncover patterns and insights from historical World Cup match records. The analysis begins with data analysis, mostly the provided data set was clean. It then progresses through **univariate analysis** to understand individual variable distributions, **bivariate analysis** to explore relationships between pairs of variables, and **multivariate analysis** to examine complex interactions across multiple attributes. Visualizations such as histograms, scatter plots, correlation matrices, heatmaps, and parallel coordinate plots are used throughout to facilitate meaningful interpretations of the dataset.

1. Handling Missing Values

Objective: To ensure that the dataset does not contain any NaN values or any outliers which can disrupt

model training.

Steps Taken:

● Used the .isnull().sum() function to check for null values in each column.

● NO missing values were found in our dataset but still applied:

○ For numerical columns: Missing values to be replaced using the mean or median of the column.(although none found)

- For categorical columns: Mode could be used to fill missing values.

● In our case, it appears that there were no significant missing values, so no imputation was necessary.

2. Cleaning column names:

Spaces were removed from column names

---

4. **Outlier Detection**:

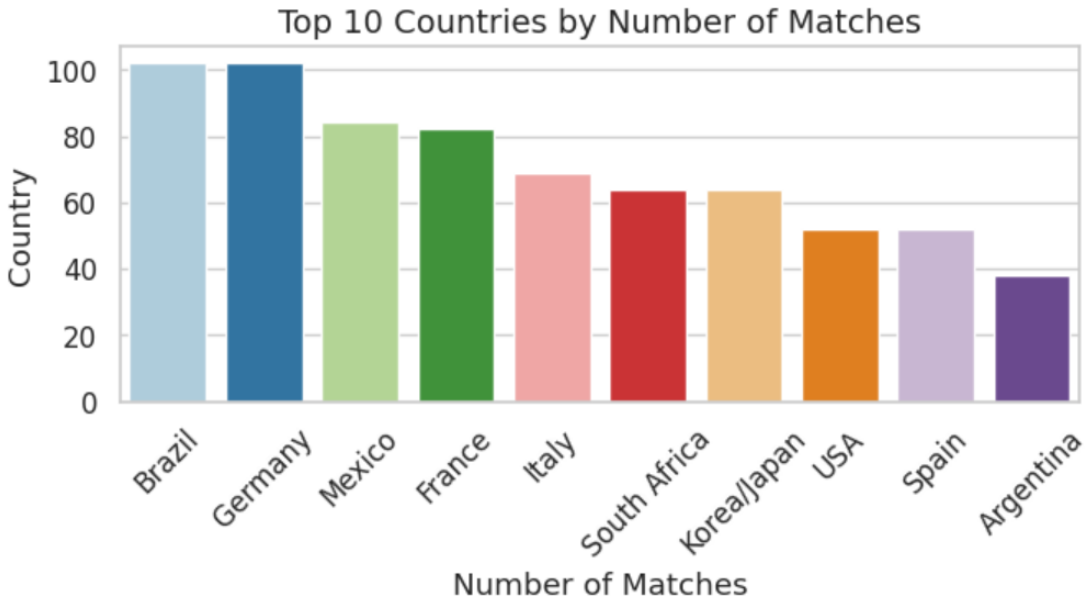
No outliers were found in our case

Exploratory Data Analysis

Univariate Analysis

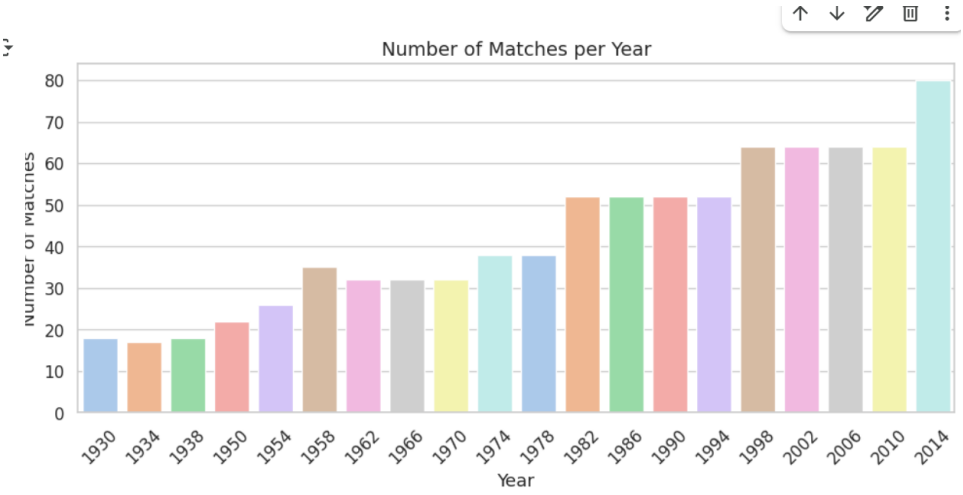
Top Countries by Number of Matches:

This univariate analysis shows the ten countries that have hosted the most football matches. It helps identify the most active or popular host nations in the dataset and provides insights into historical match distribution.



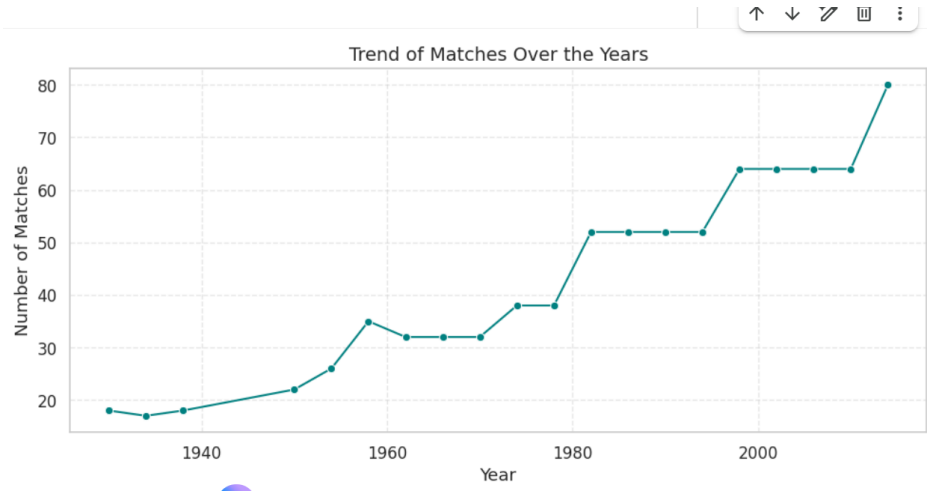
Number of Matches per Year:

This visualization shows the total number of football matches played each year. It helps identify trends over time, such as periods with more or fewer matches, and provides insights into the historical growth or fluctuations in international football activity.



Trend of Matches Over the Years (Line Plot):

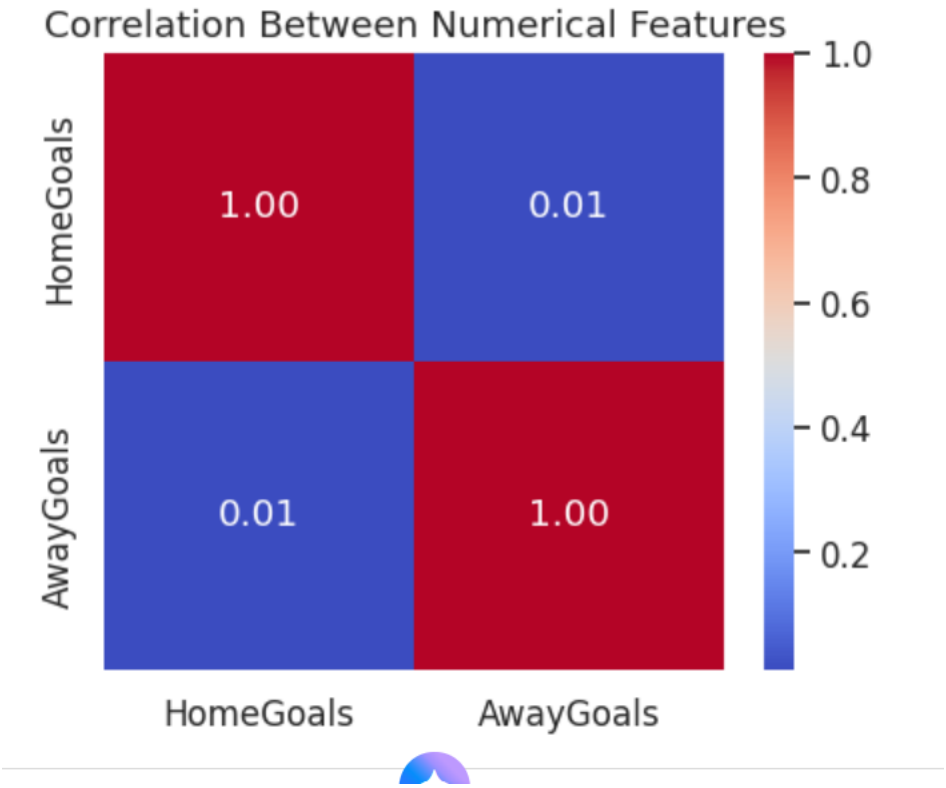
This line plot illustrates how the number of football matches has changed over time. It helps identify long-term trends, peaks, or declines in matc frequency, providing a clear view of the evolution of international football over the years



Bi-variate Analysis

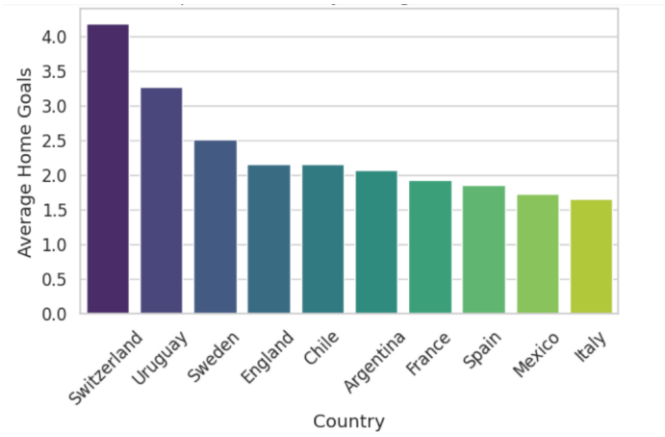
Correlation Between Numerical Features:

This analysis examines the relationships between pairs of numerical variables, such as goals scored, match counts, and other statistics. It helps identify positive or negative correlations, revealing which features tend to increase or decrease together in football match data.



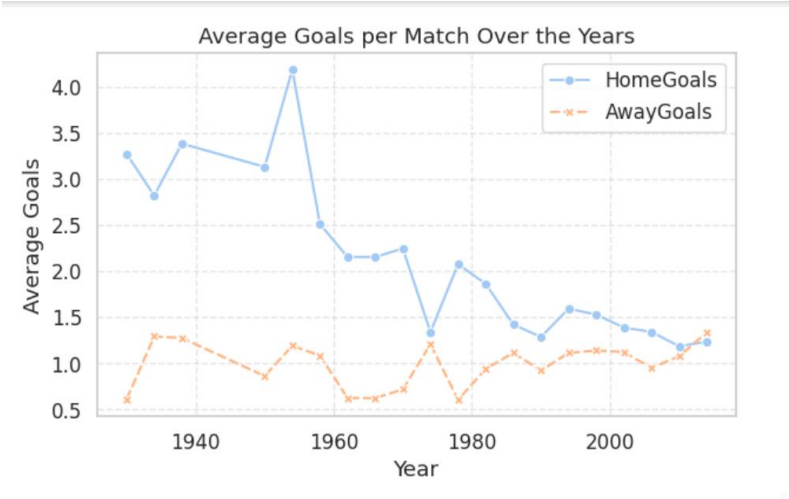
Top 10 Countries by Average Home Goals :

This visualization shows the ten countries with the highest average number of goals scored by home teams. It helps compare offensive performance across nations and provides insights into which countries tend to have stronger home-game scoring records.



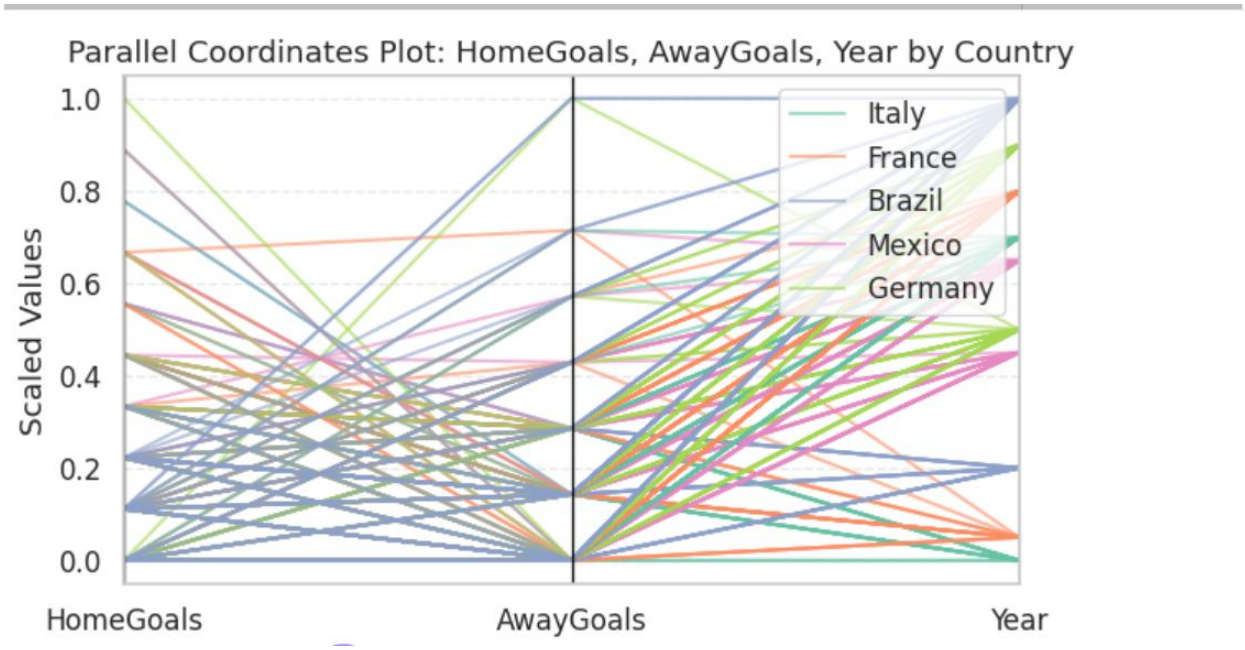
Average Goals per Match Over the Years:

This visualization shows how the average number of goals scored per match has changed over time. It helps identify trends in offensive performance, such as periods of high or low scoring, and provides insights into the evolution of gameplay in international football

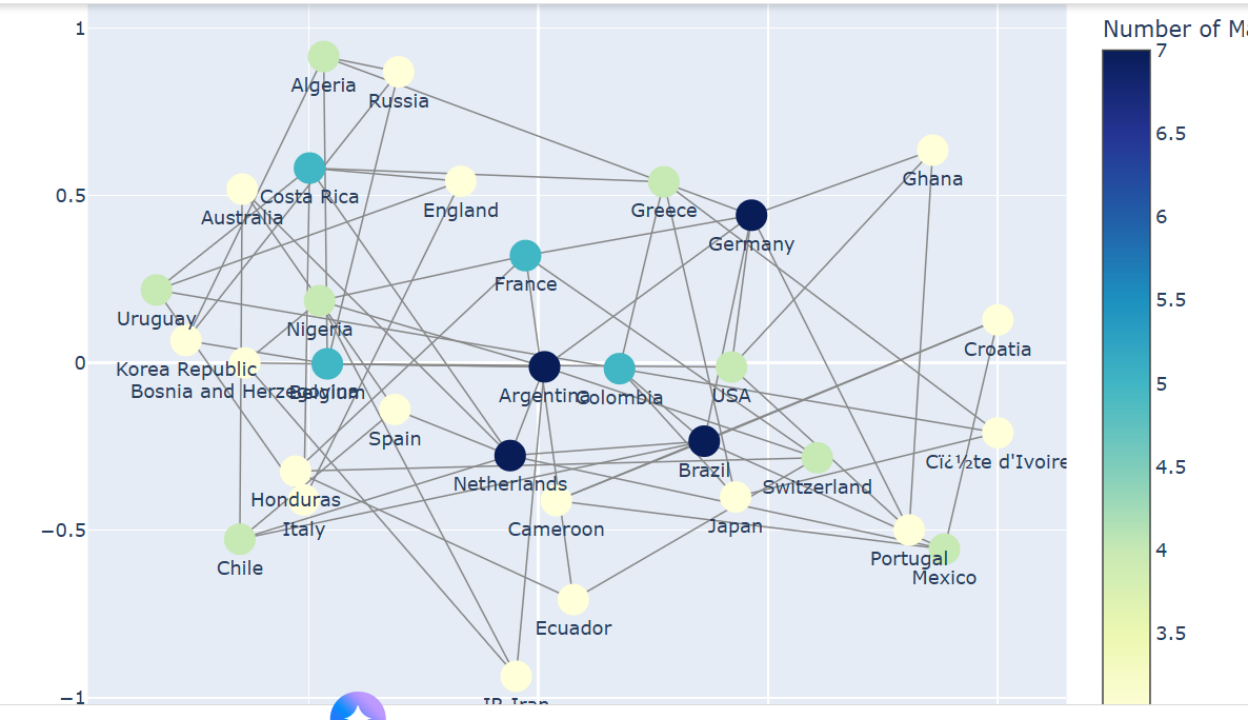


Multi-Variate Analysis

Parallel Coordinates Plot — Home Goals, Away Goals, and Year by Country:
This plot visualizes multiple variables simultaneously home goals, away goals, and match year grouped by country. It helps identify patterns and trends in scoring over time for different nations, highlighting differences in offensive and defensive performance across countries.



FOOT BALL NETWORK



Data INK Ratio

```
================================================================================
SUPPLEMENTARY ANALYSIS: LIE FACTOR IN 3D VISUALIZATION
================================================================================

Edward Tufte's Lie Factor Principle:
------------------------------------
Lie Factor = (Size of effect shown in graphic) / (Size of effect in data)

Ideal Lie Factor = 1.0 (perfect representation)
Acceptable range: 0.95 - 1.05
Problematic: < 0.95 or > 1.05

In 3D Bar Charts:
-----------------
• Bars in the foreground appear larger due to perspective
• Bars in the background appear smaller due to distance
• 3D depth adds volume that doesn't represent data
• Angle of view distorts height perception
```

For Visualization 2 (Website Traffic 3D Chart):

Without exact measurements from the image, we can estimate based on typical 3D chart distortions:

- Front bars may appear 20-30% larger than actual value
- Back bars may appear 15-25% smaller than actual value
- 3D volume effect adds ~40-60% visual mass not representing data

Estimated Lie Factor Range: 1.15 - 1.40 (PROBLEMATIC)

This means the visual representation exaggerates differences by 15-40%, making comparisons inaccurate and misleading.

CONCLUSION: The 3D effect not only reduces the Data-Ink Ratio to 0.20 but also introduces a Lie Factor > 1.15, violating graphical integrity.

```
file,total_pixels,ink_pixels,data_ink_pixels,data_ink_ratio
/content/obj_2-Sales (1).png,640926,162402,161315,0.9933
/content/obj_2-Energy (1).png,640926,56559,55960,0.9894
/content/obj_2-Energy (1).png,640926,56559,55960,0.9894
```


```
file,total_pixels,ink_pixels,data_ink_pixels,data_ink_ratio
/content/obj_2-Sales (1).png,640926,162402,161315,0.9933
/content/obj_2-Energy (1).png,640926,56559,55960,0.9894
/content/obj_2-Energy (1).png,640926,56559,55960,0.9894
```