Visual Analysis Report

Datasets Used

- Airbnb Dataset
- Football World Cup Dataset

Objective

This project performs a visual exploratory data analysis (EDA) on two datasets: Airbnb listings and Football World Cup matches, to uncover patterns and relationships among their attributes. For each dataset, I will clean and explore variables (Univariate, bivariate, and multivariate), form testable hypotheses based on the EDA, and defend those hypotheses with focused visualizations. I will also compute the data-ink ratio for the plots and comment on graphical integrity and visualization principles. Finally, I will build a network diagram showing all match pairings from the 2014 World Cup to visualize team connections and tournament structure.

Airbnb Dataset

Exploratory Data Analysis (EDA):

1. Data Preparation

To begin the analysis, I first explored the dataset using *info(), describe()*, and by examining individual columns to understand the data types, missing values, and distributions.

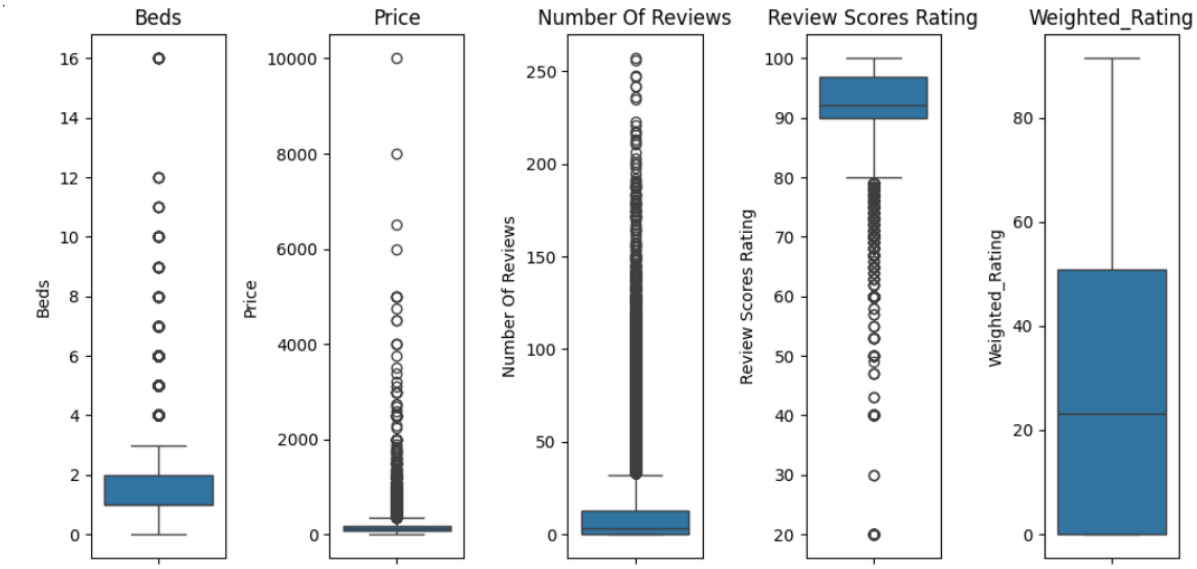
Cleaning and Preprocessing:

- **Missing Values:**
  - ✓ The columns *'Host Since'* and *'Property Type'* had very few missing values, so those records were dropped.
  - ✓ The *'Zipcode'* column was not essential for this analysis, so missing values were filled with `0`.
  - ✓ The *'Beds'* column was filled with the median value since it best handles outliers and skewed data.
  - ✓ For *'Review Scores Rating'* and *'Review Scores Rating (bin)'*, missing values were filled based on similar listings (by *property type*, *room type*, or *neighborhood*) to maintain contextual accuracy.
  - ✓ After filling, around 25 records still had null review scores with no clear pattern, so these were dropped as random data gaps.
- **Duplicates:**
  - ✓ Only 17 duplicate rows were found and removed.
- **Data Type Correction:**
  - ✓ The *'Beds'* column was converted from float to integer for better clarity.
- **Feature Engineering:**

- ✓ To make review scores more reliable, I created a new column called *Weighted_Rating*. It adjusts the ratings by giving more weight to listings with a larger number of reviews, making highly rated listings with very few reviews less misleading.

2. Univariate Analysis

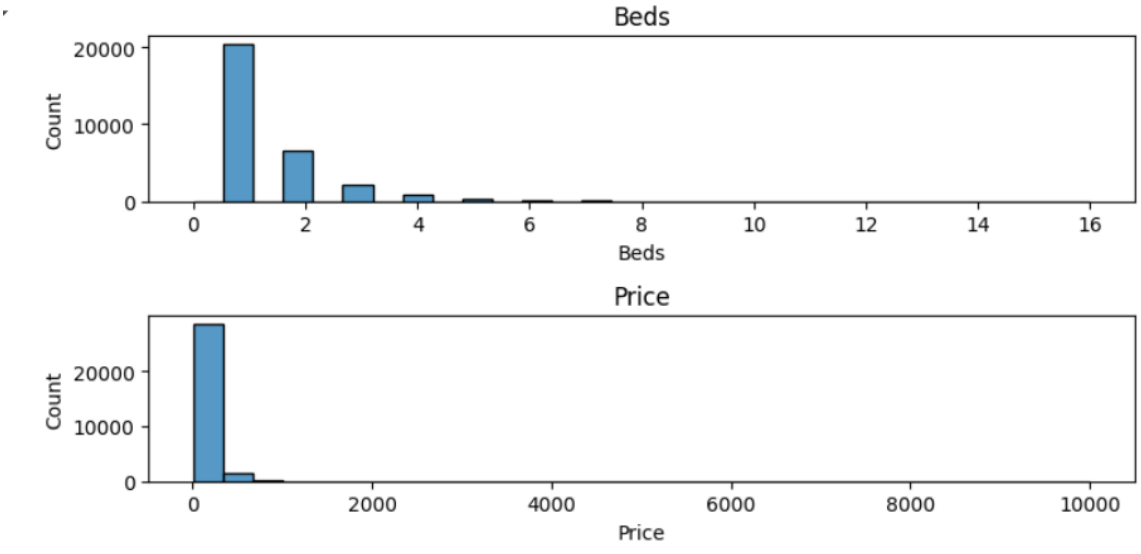
Box plots:

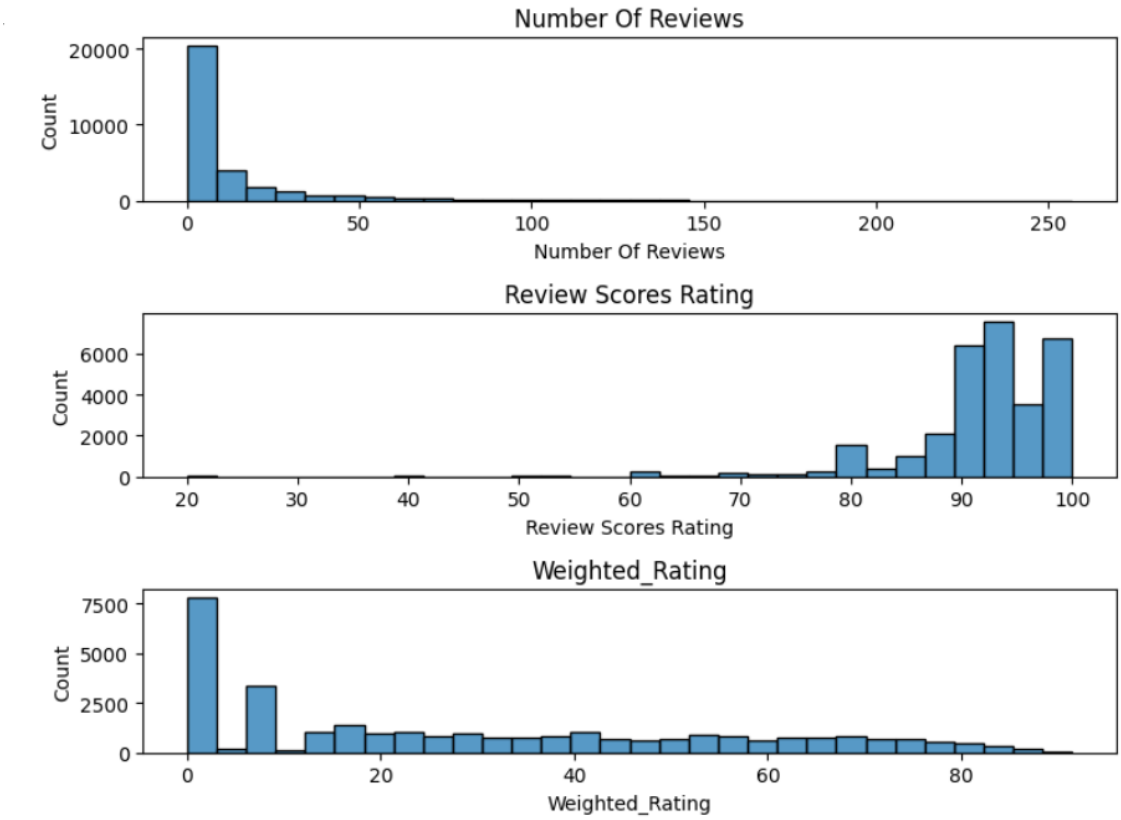


Description and Interpretation:

These outliers seem to reflect real-world differences (like larger properties or luxury listings), so I chose not to remove them

Histograms:



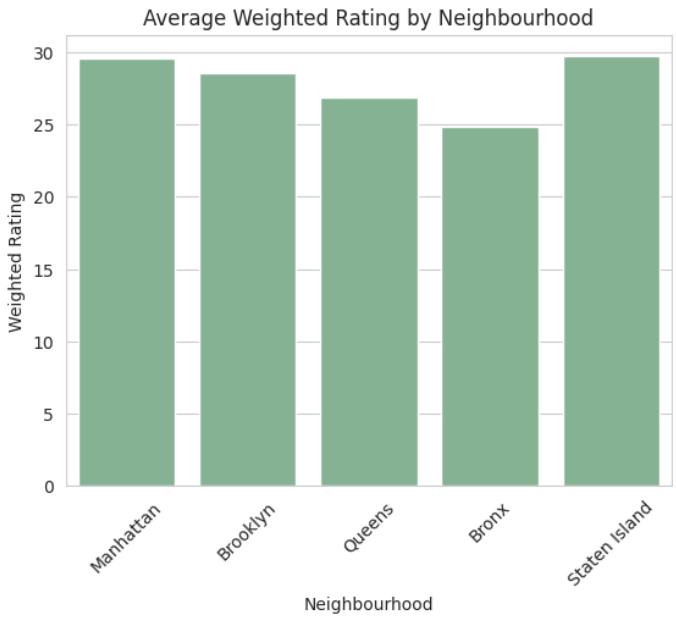


Description and Interpretation:

We can clearly see that the data distribution is skewed for most attributes

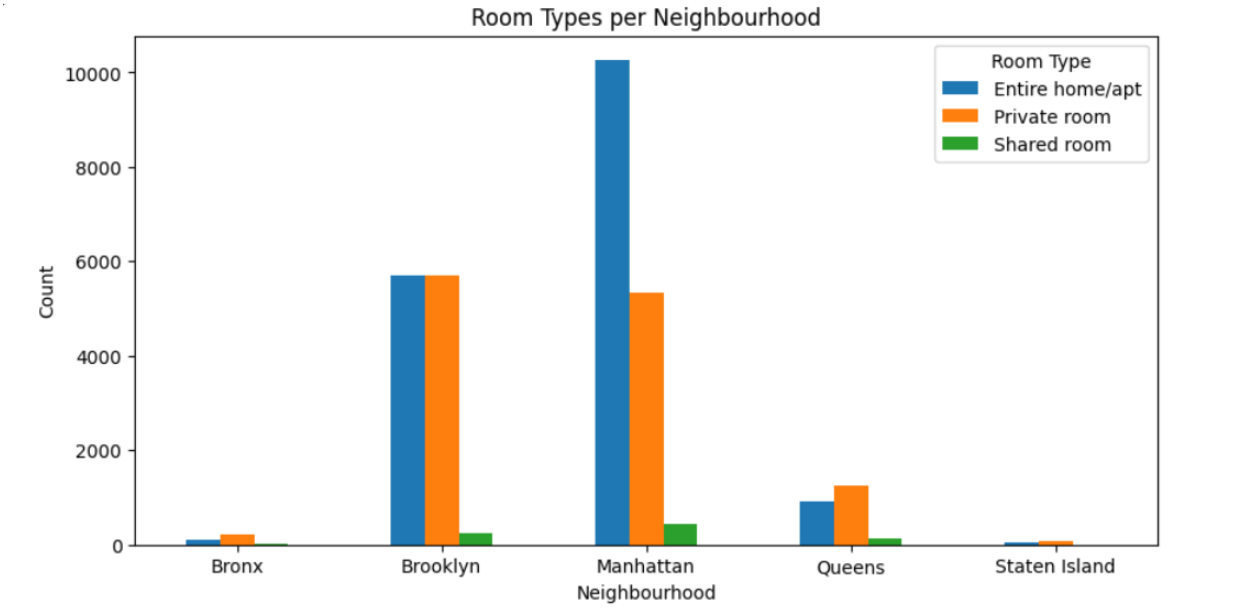
3. Bivariate Analysis

Bar plots:



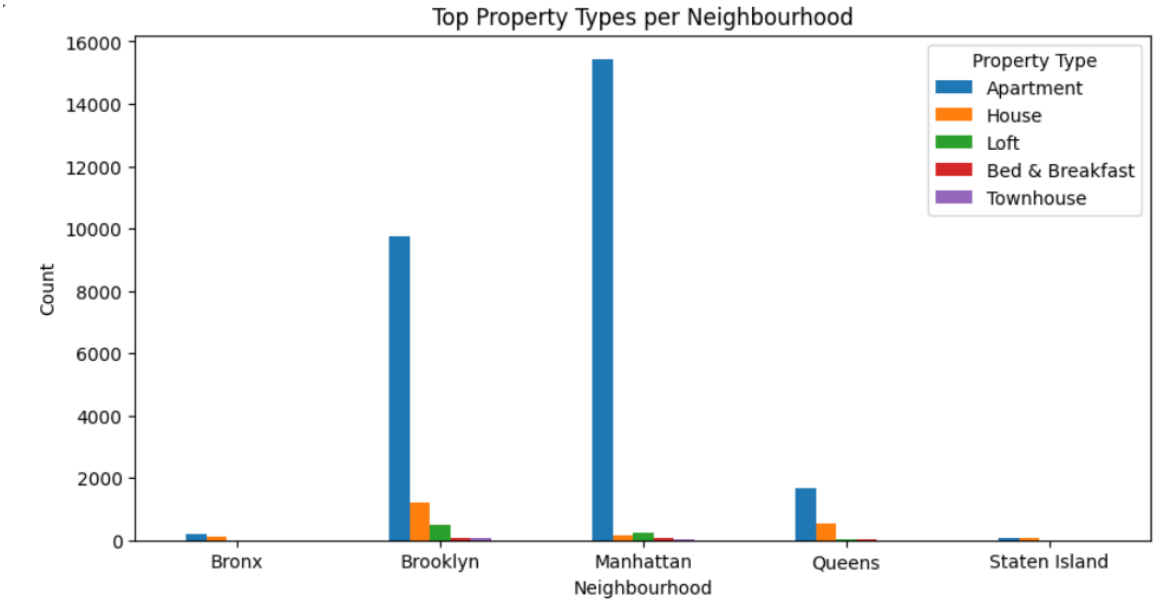
Description and Interpretation:

The bar plot shows the average weighted rating for each neighborhood from the dataset. We can see all neighborhoods have almost similar performance.



Description and Interpretation:

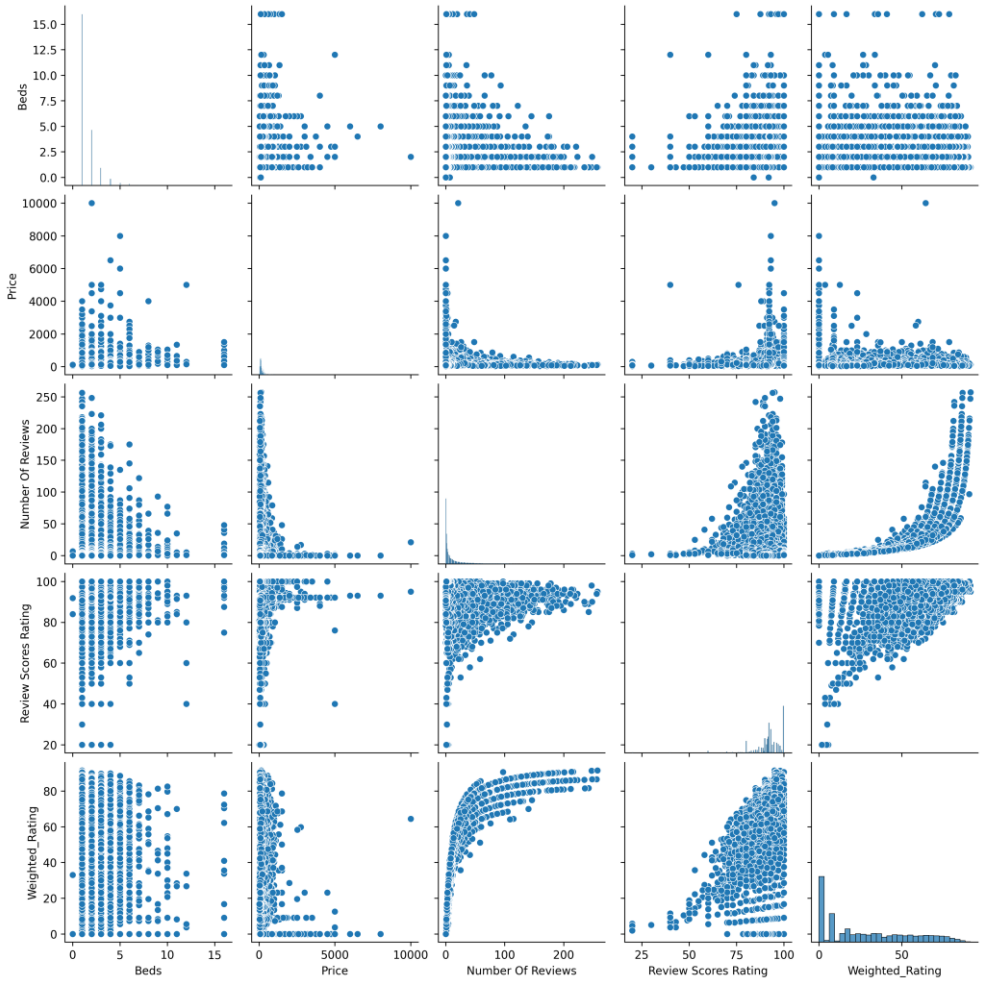
The grouped bar plot above shows the distribution of room types per neighborhood. We see that Manhattan has a large no of entire home/apt listings and how other neighborhoods' listings are performing. We can also see that shared rooms are very less.



Description and Interpretation:

The grouped bar plot above shows the distribution of the top 5 property types per neighborhood. We see that all of them have a large no of apartment listings compared to other property types. Maybe people prefer to rent out apartments.

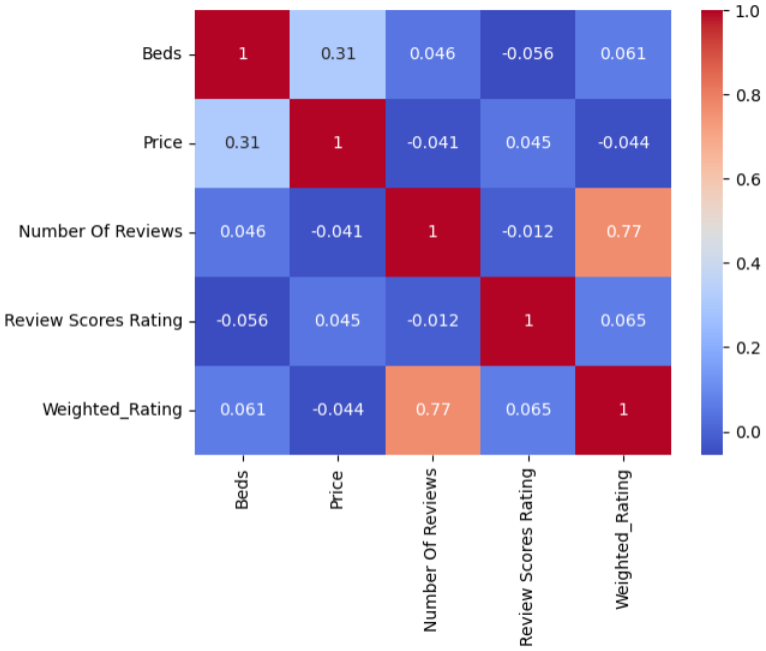
Pair plot:



Description and Interpretation:

The pair plot effectively shows the relationships between all numerical variables without having to plot every single scatter plot. Here we can also see that none of them has a strong relationship with the other. Some can be seen to have weak non-linear relationships, but they aren't strong enough that we can deduce a pattern. This is also confirmed by the correlation matrix plotted below.

Correlation Matrix:

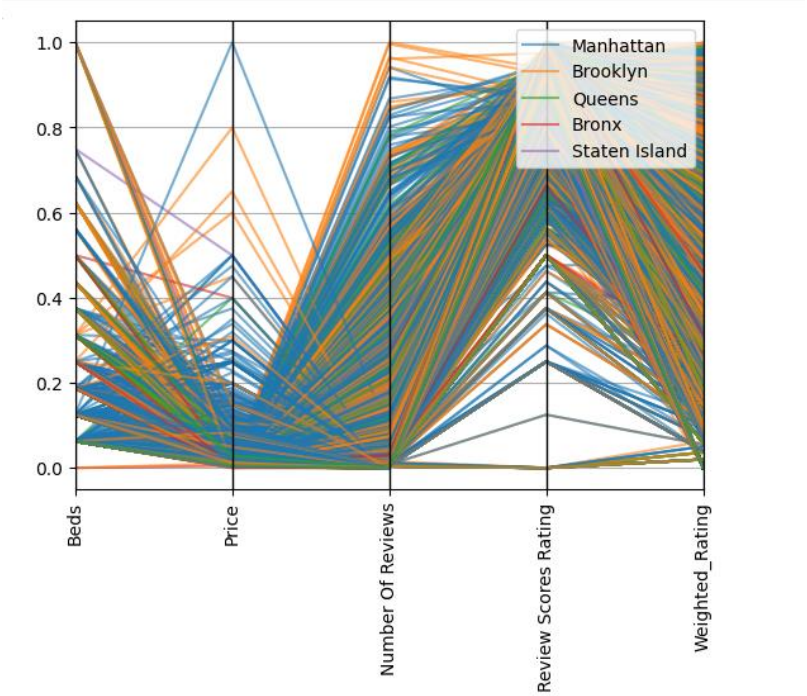


Description and Interpretation:

As discussed in the pair plot, here we can also see that none of them has a strong relationship with the other

4. Multivariate Analysis

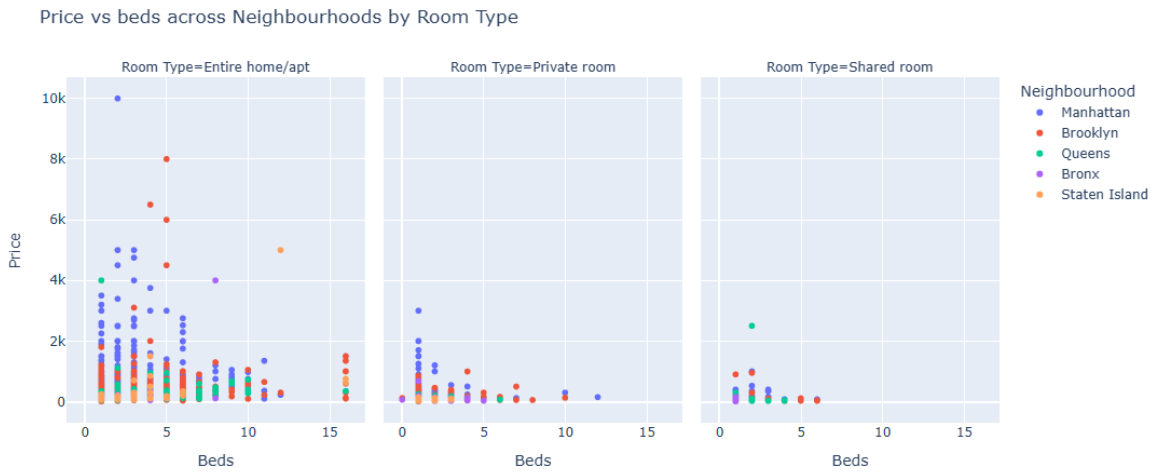
Parallel Coordinates plot:



Description and Interpretation:

The parallel coordinates plot shows how, for each record, the attributes are connected. Here we see an interesting pattern of how listings with a large no of beds also tend to have lower prices, and how the number of reviews and review score are interacting with the weighted rating.

Trellis Plot:



Description and Interpretation:

I also decided to plot a trellis plot where we get to see separately for each room type how the beds and prices are related for each neighborhood. We can also see over here how, as the no of beds increases, they still remain in the same price lane as the smaller bed listings do.

Hypothesis:

Statement

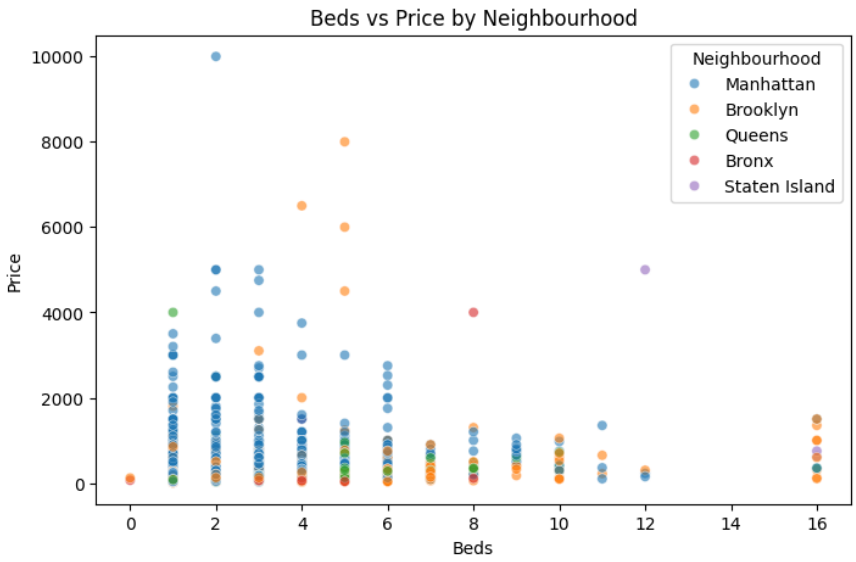
In the Airbnb dataset, listings with more beds tend to have lower to moderate prices

Reasoning:
From the parallel coordinates and the trellis plot, you can see many lines where beds increase, but price decreases or are affordable to those with lower beds.

Another Observation:
A few listings with fewer beds but high prices are likely luxury or central-location properties, which supports that price is driven more by location and quality than room size. Hence, they might have higher Ratings?

Testing

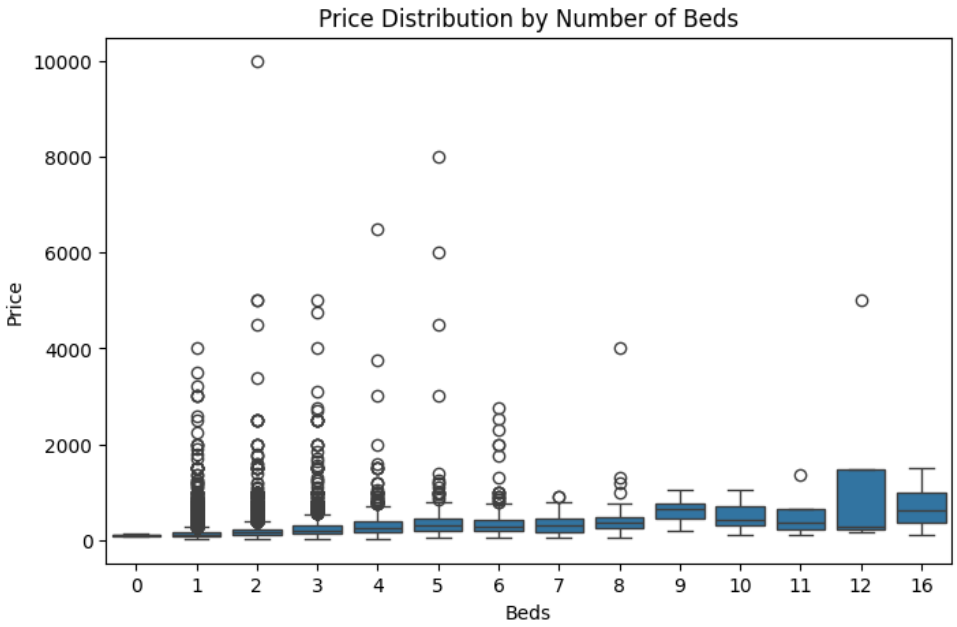
Scatter plot:



Description and Interpretation:

The scatter plot is similar to the trellis plot we showed earlier. This is showing that majority cases are following this low price and high bed rule but it's not strict because there are some exceptional luxury cases existing in the dataset.

Box plot:



Description and Interpretation:

The median line stays roughly the same across 1–7 beds. Some boxes have many high outliers, meaning a few expensive listings exist, but most prices are low to moderate. This again means: adding more beds doesn't really increase price overall.

Insights and Conclusion

The hypothesis is partly supported. While the data shows that prices don't rise consistently with more beds, it doesn't strongly prove that more beds lead to lower prices either. This suggests that factors like location, room type, or property type have a greater impact on pricing than the number of beds.

Recommendations for Further Analysis

- Include more attributes like *amenities*, *host response time*, or *availability* to better explain price differences.
- Perform geospatial analysis using latitude/longitude to visualize price trends across neighborhoods.
- Explore sentiment analysis on review text to connect guest opinions with ratings and prices.
- Build a predictive model (e.g., regression) to estimate listing prices based on location, room type, and reviews.

Football World Cup

Exploratory Data Analysis (EDA):

1. Data Preparation

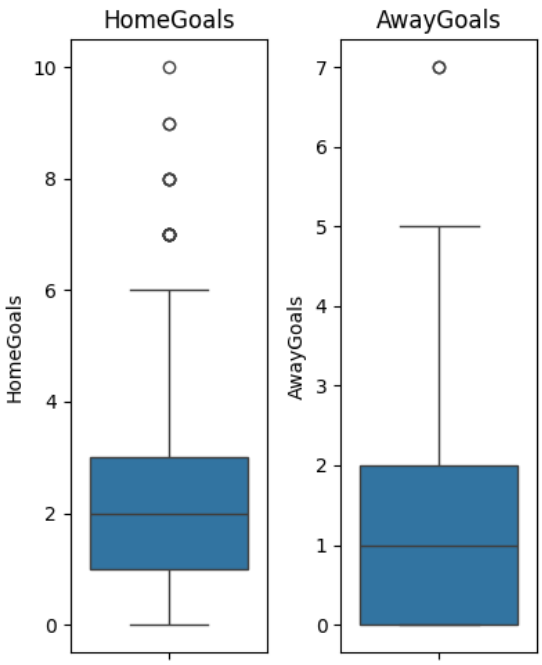
I began by exploring the dataset using *info(), describe()*, and by inspecting individual columns to understand the structure, variable types, and overall data consistency.

Cleaning and Preprocessing:

- **Missing Values:**
  - ✓ The dataset contained no missing values, so no imputation was required.
- **Duplicates:**
  - ✓ 16 duplicate rows were identified and removed to avoid repetition in match analysis.
- **Data Type Correction:**
  - ✓ All columns had appropriate data types; therefore, no conversions were necessary.
- **Feature Engineering:**
  - ✓ To simplify and improve the clarity of analysis, I created a *simplified round category* by grouping match stages into broader types. This made it easier to analyze trends and visualize match patterns across different stages of the tournament.

2. Univariate Analysis

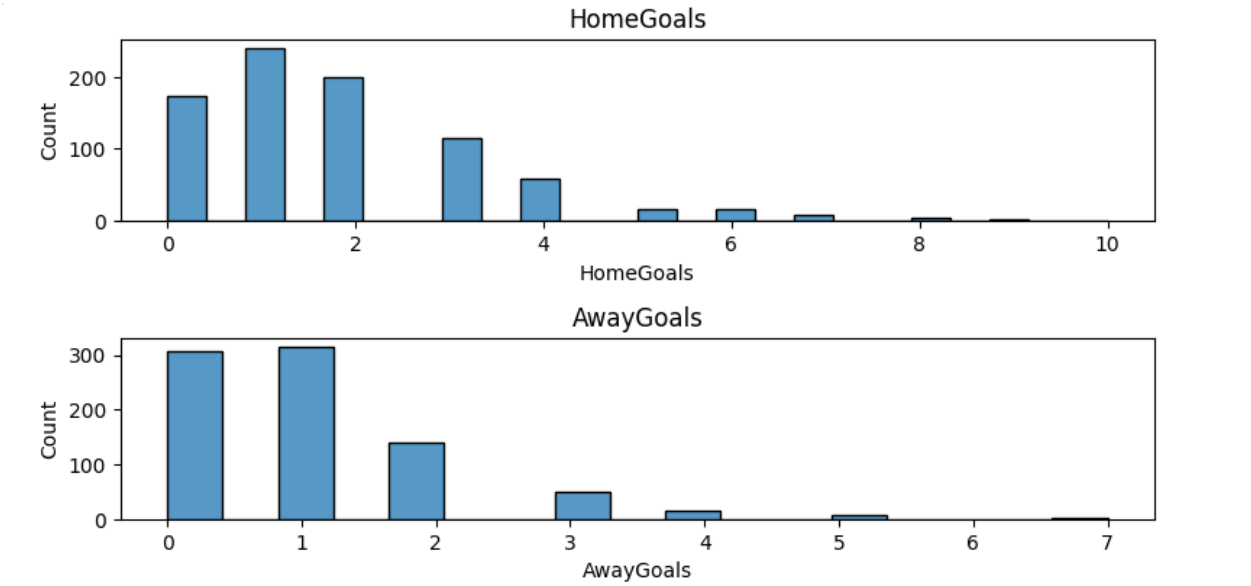
Box plots:



Description and Interpretation:

Here we can see that there are a few outlier cases existing in the Home Goals attribute showing how the home team managed to score huge achievements on some matches.

Histograms:

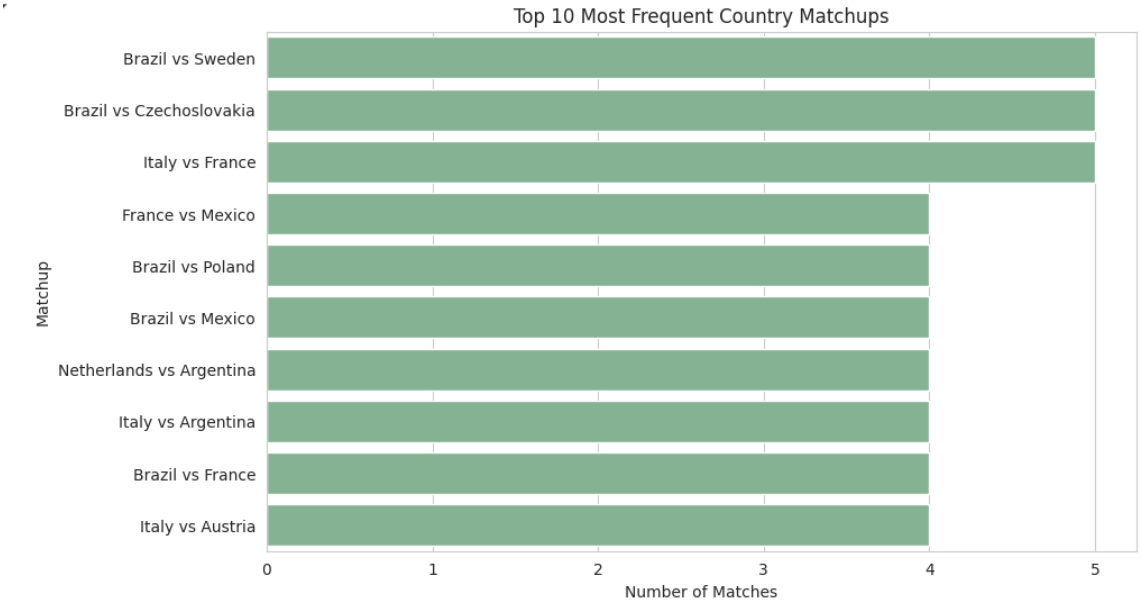


Description and Interpretation:

Again as seen in the Box plot we can see out data is right skewed.

3. Bivariate Analysis

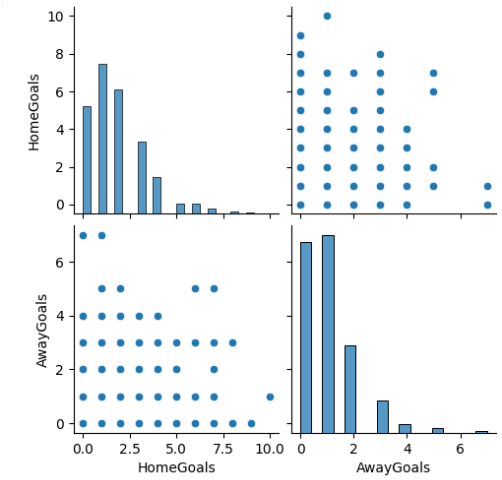
Bar plot:



Description and Interpretation:

This bar plot was plotted to analyze the frequency of specific pairing matches that occurred in the world cups. We can see that Brazil must have a strong team for it to make it to the finals with multiple pairings again and again.

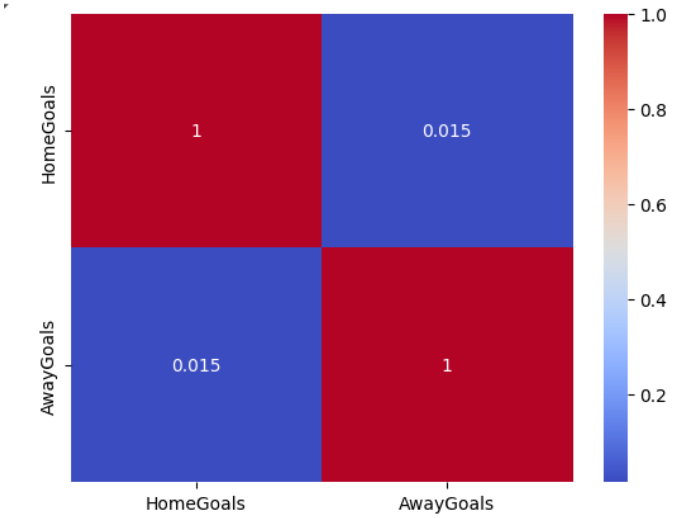
Pair plot:



Description and Interpretation:

The pair plot shows that almost no relation exists between the home goals and away goals.

Correlation Matrix:

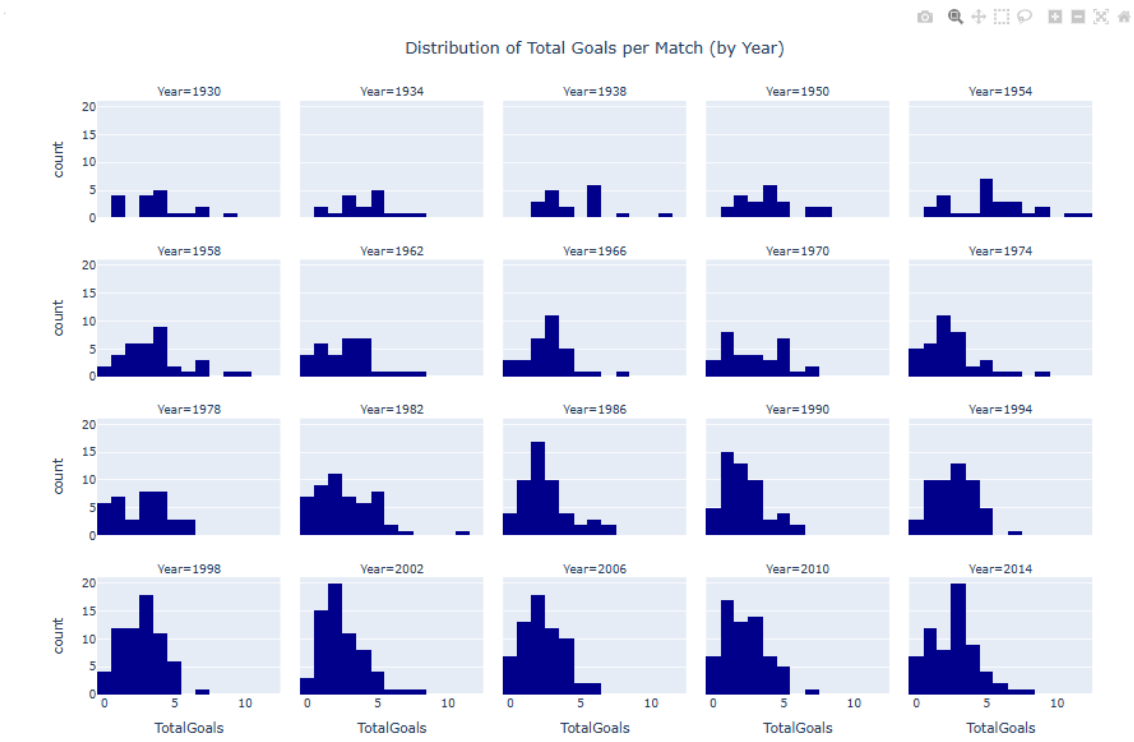


Description and Interpretation:

We can see the same results here as seen in the pair plot.

4. Multivariate Analysis

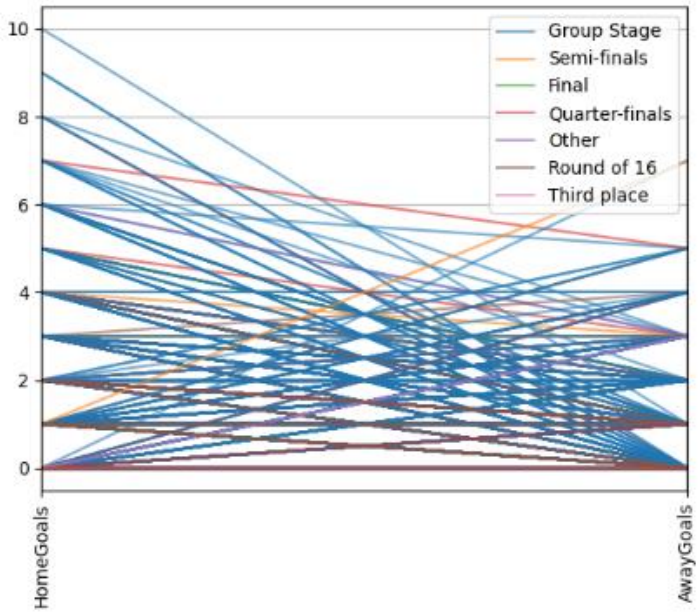
Trellis plot:



Description and Interpretation:

We decided on making a trellis plot for analyzing the total number of goals recorded during each match for all the years of world cup. And we can see that over the years the matches became more competitive.

Parallel coordinates plot:



Description and Interpretation:

This parallel coordinate plot shows how there isn't a specific pattern between these two attributes for any match, but we do see that mostly home goals tend to be larger but there are exceptional cases existing where away goals were larger too.

Hypothesis:

Statement

Over the years, World Cup matches have become more competitive, i.e, the total goals become fewer and fewer (more frequent at 1,2, and 3)

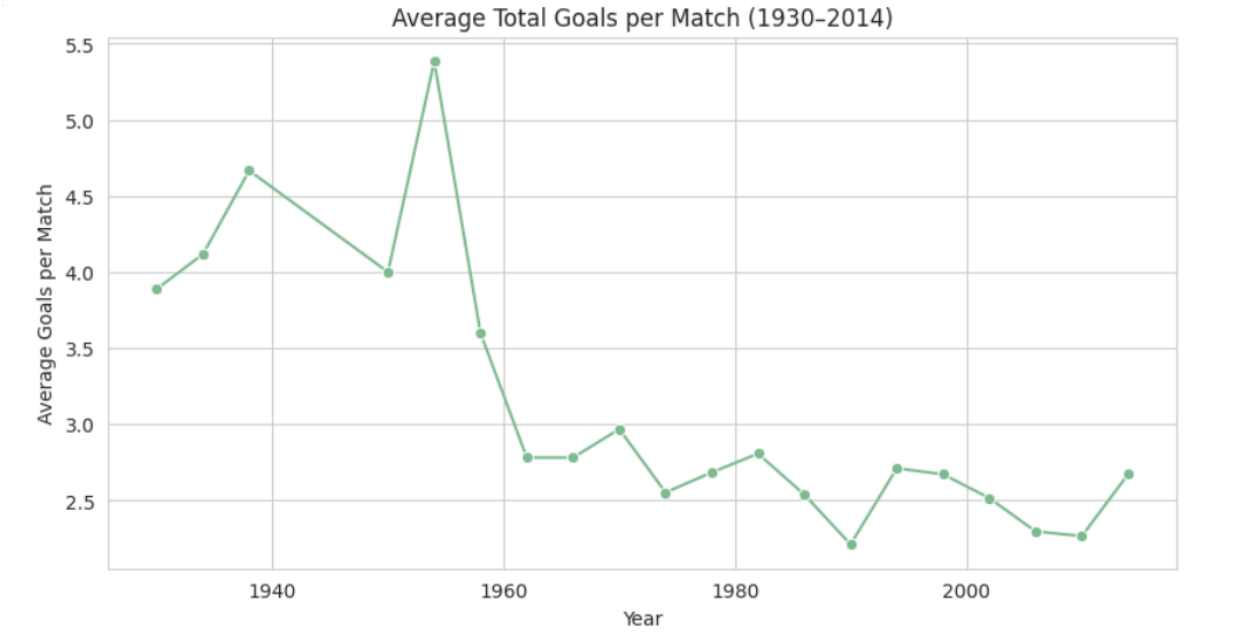
Reasoning:
From the trellis plot, later tournaments (2000s–2010s) show wider goal distributions i.e, more matches with 3 or fewer total goals, compared to older tournaments where most matches had somewhat equal distribution for all total goals. This could suggest that defensive strategies have evolved, hence making the goal scoring tougher

Another Observation:
Certain countries, such as Brazil, have consistently reached advanced World Cup stages, leading to repeated matchups with top-performing teams. From the "Top 10 Most Frequent Country Matchups" chart, Brazil

appears in most recurring match pairs, e.g., vs Sweden, vs Mexico, vs France, suggesting that Brazil frequently advances to stages where it meets other strong teams.

Testing

Line plot:



Description and Interpretation:

Hence, we can see that the average total scores drop as years progress, explaining that, indeed, scoring of goals became tougher and games became more competitive

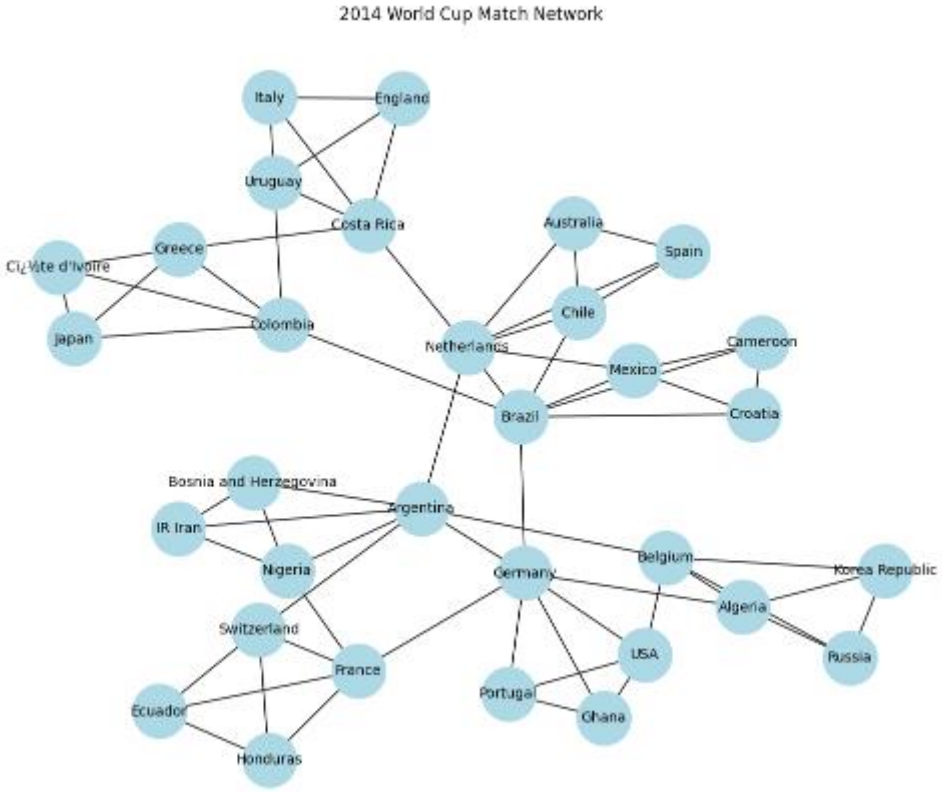
Insights and Conclusion

From the analysis, it can be concluded that football matches have become more competitive and defensively balanced over time**.** Overall, the hypothesis is supported by visual evidence, confirming that goal scoring has become tougher in recent years.

Recommendations for Further Analysis

- Combine this dataset with player or team statistics (e.g., possession, shots, rankings) to explain performance trends.
- Analyze goal timing (first vs. second half) to see how match intensity changes over time.
- Study home vs. away team advantage or referee effects on match outcomes.
- Extend the dataset beyond 2014 to check if the competitive trend continues in later tournaments.

Network Analysis (Football 2014 World Cup)



Description and Interpretation:

The network graph visualizes all matches played during the 2014 FIFA World Cup.
Each node (circle) represents a country/team, and each edge (line) connecting two nodes represents a match played between those teams.