Comprehensive Exploratory Data Analysis Report

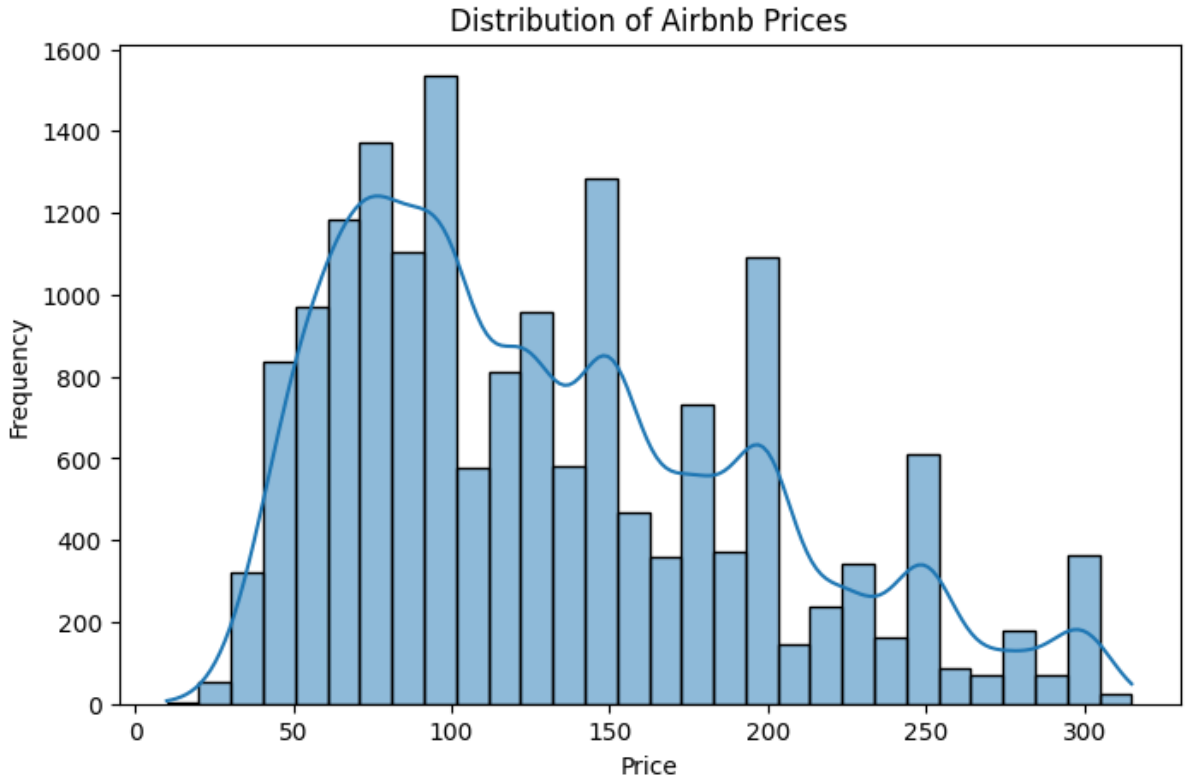
This report summarizes the key findings, patterns, and anomalies observed from the Exploratory Data Analysis (EDA) of two distinct datasets: NYC Airbnb Listings and World Cup Match Results.

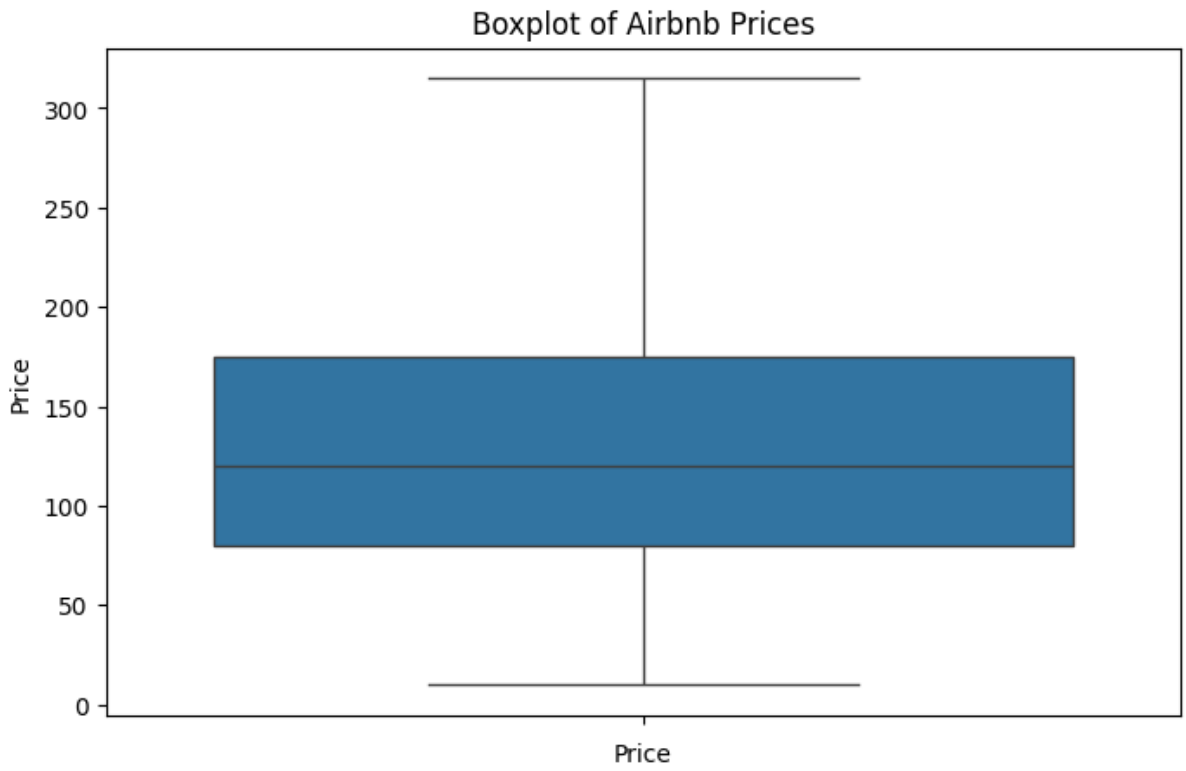
I. Visual Analysis Report: NYC Airbnb Listings

The analysis of the Airbnb dataset provided deep insights into how structural and geographical factors influence listing prices and performance.

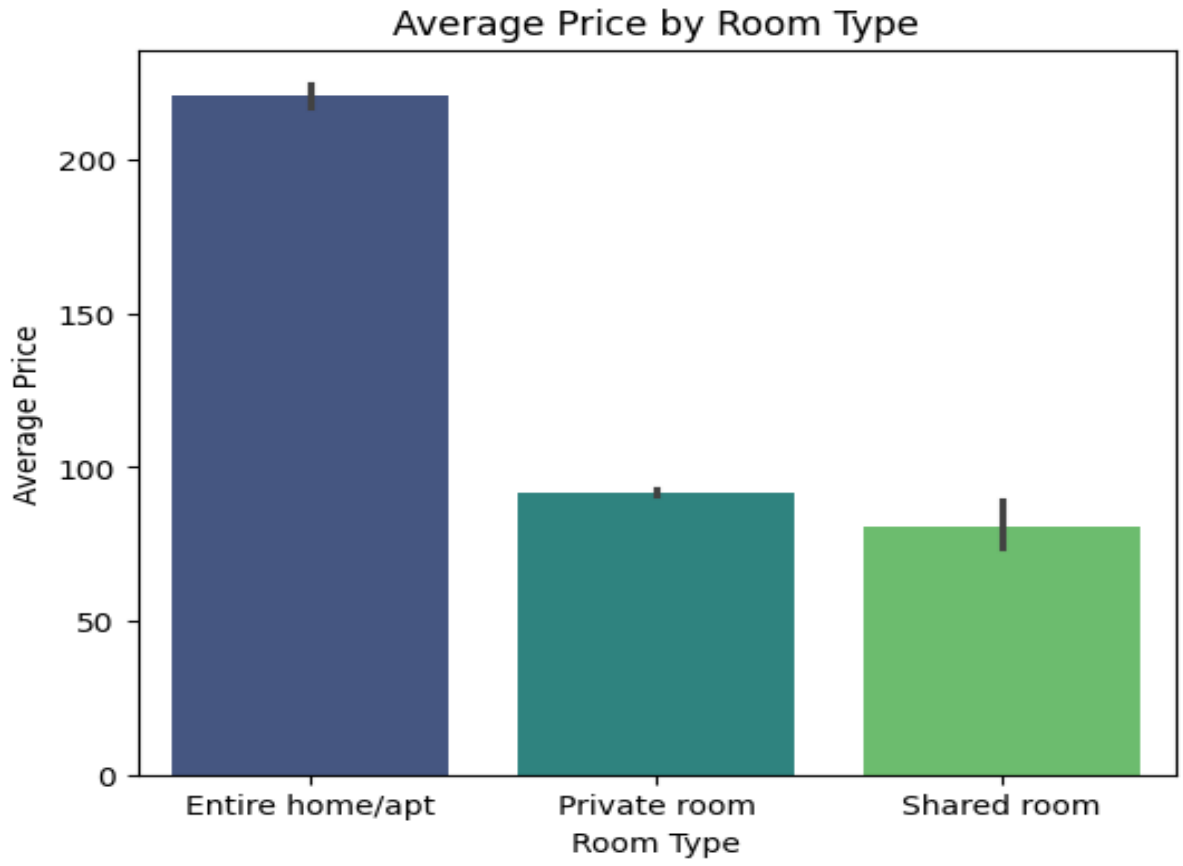
1. Price Distribution (Histogram and KDE Plot)

The distribution of the raw Price variable is heavily right-skewed, concentrating the majority of listings below $200. The use of a log-transformed price (Log Price) was crucial to normalize the data, confirming that a small number of ultra-high-priced luxury listings significantly inflate the average price across the dataset.





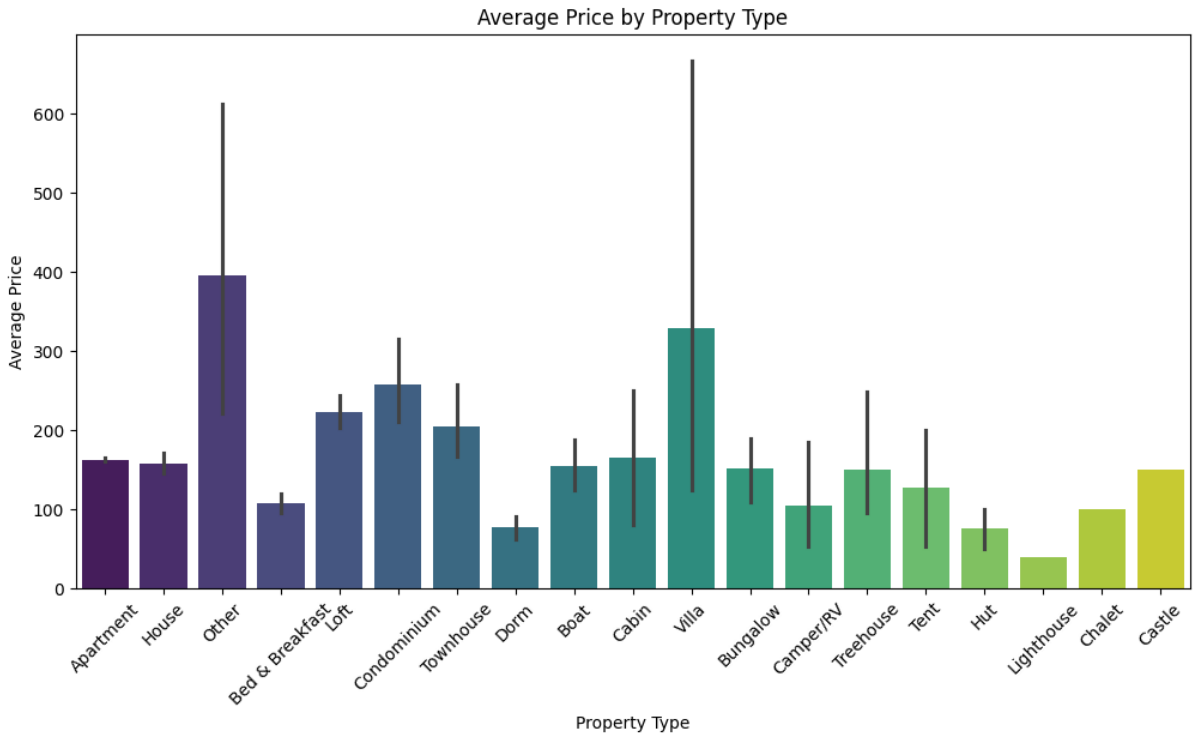
2. Room Type vs. Price (Bar Plot)



This bar plot confirms that **Room Type significantly affects price**. **Entire home/apt** listings command the highest average price, followed by **Private rooms**, and then **shared rooms** at the lowest end. This establishes room type as the primary structural factor influencing cost.

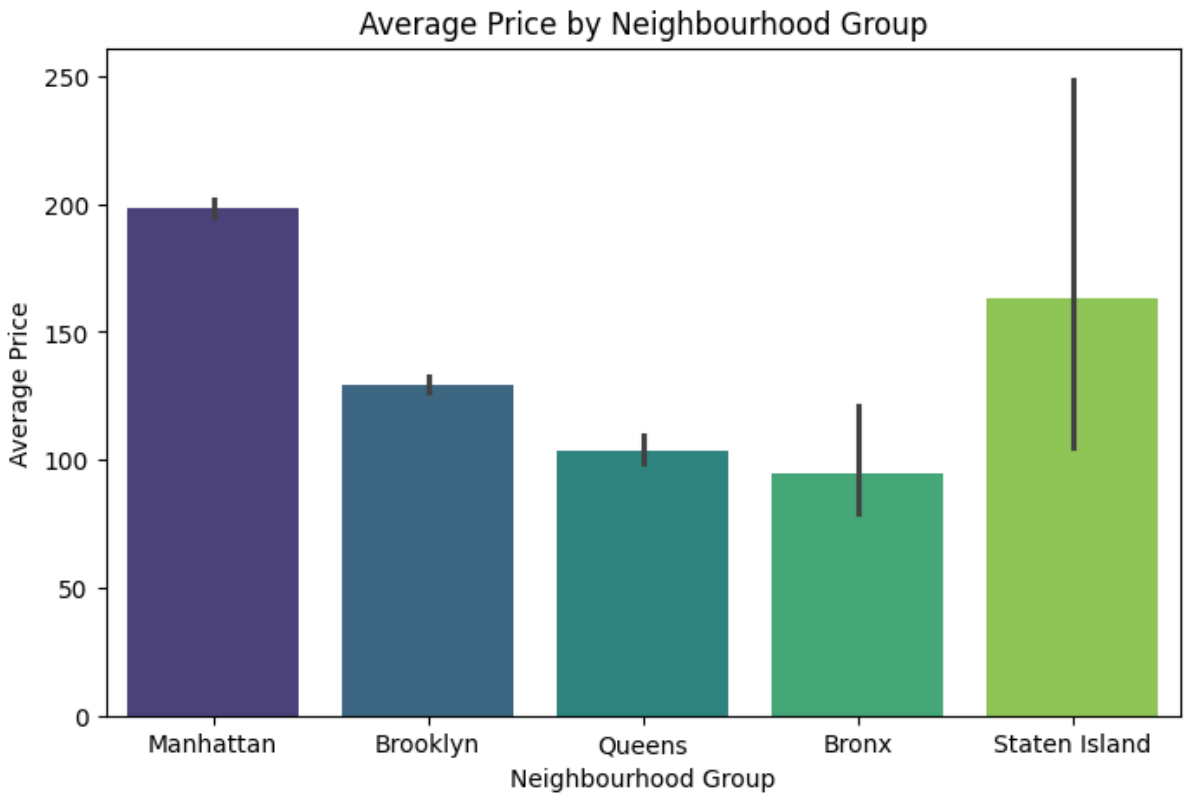
3. Price Differ By Property Type (Bar Plot)

The visualization of **Property Type** reveals a wide range of average pricing, with specialized categories like **Villa** and **Loft** generally showing higher prices. The high variability suggests that beyond the room type, the actual structure type (e.g., apartment vs. townhouse) plays a significant role in price setting.



4. Neighbourhood Group vs. Average Price (Bar Plot)

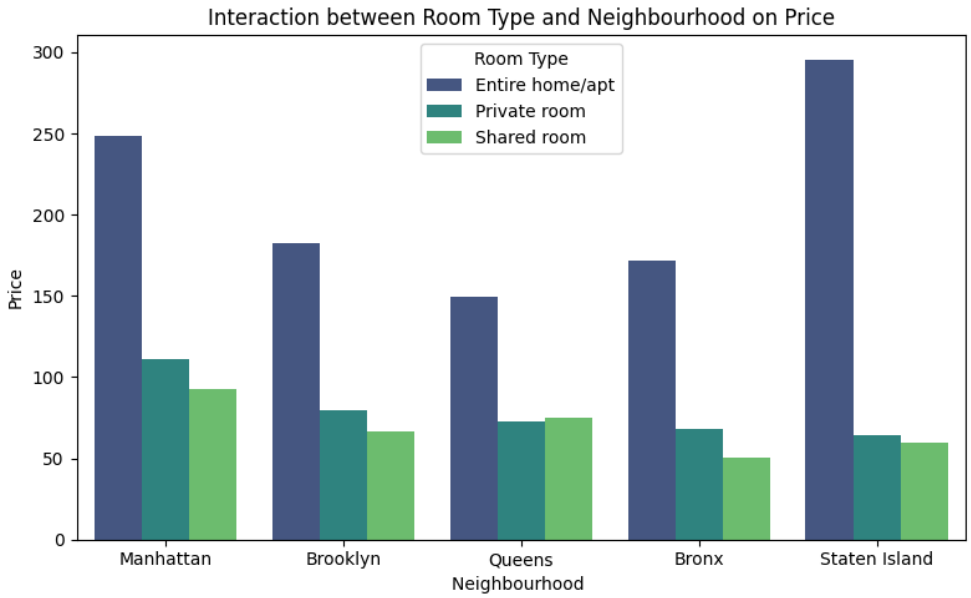
This plot directly addresses the impact of geography, showing that **Neighbourhood group influences price due to location-based demand**. As expected, **Manhattan** listings possess the highest average price, followed by **Brooklyn**, confirming the price premium associated with prime tourist and business locations.



5. Interaction: Room Type and Neighbourhood on Price (Grouped Bar Plot)

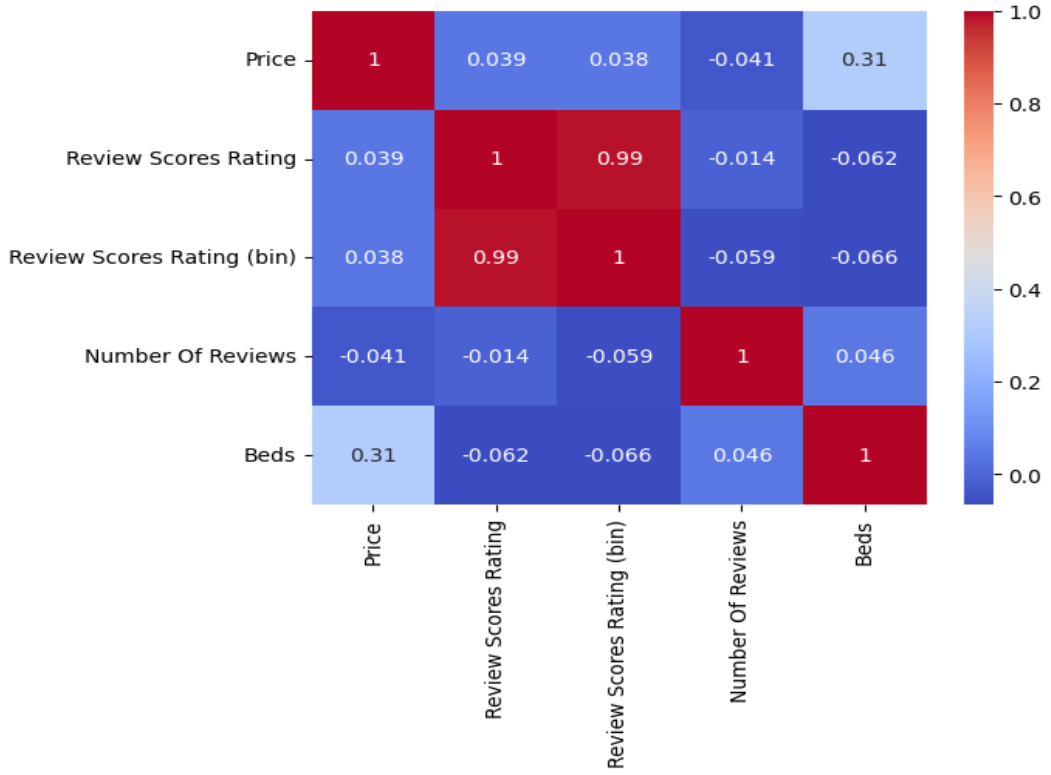
This crucial visualization confirms that **Room type and neighbourhood jointly influence the average price**. The price hierarchy remains consistent across all neighbourhoods,

But the price difference between Entire Home/apt in **Manhattan** versus the **Bronx** is enormous, indicating a compounding effect of location and exclusivity.



6. Correlation HeatMap

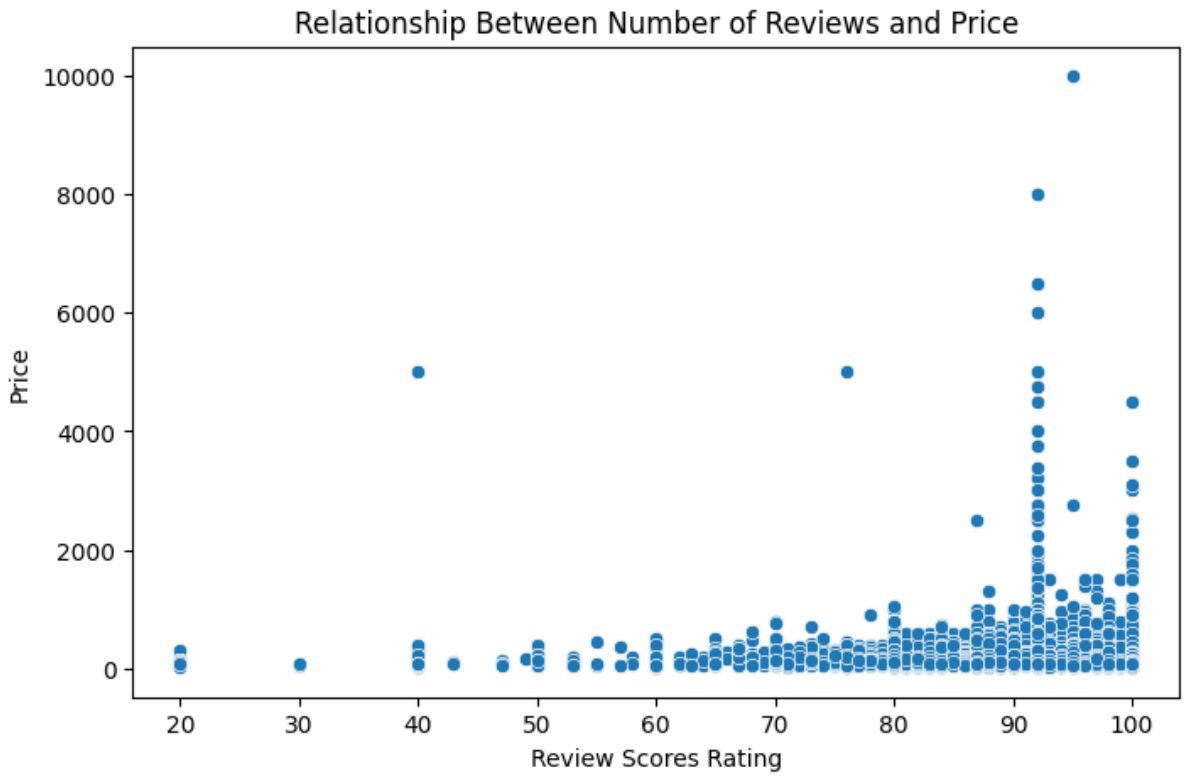
The heat map of key numeric variables shows relatively weak correlations. The strongest positive correlation exists between **Price** and **Beds**, which is intuitive. A notable weak correlation between **Number of Reviews** and **Price** suggests that high price does not deter reviews, nor does a low price guarantee high volume.



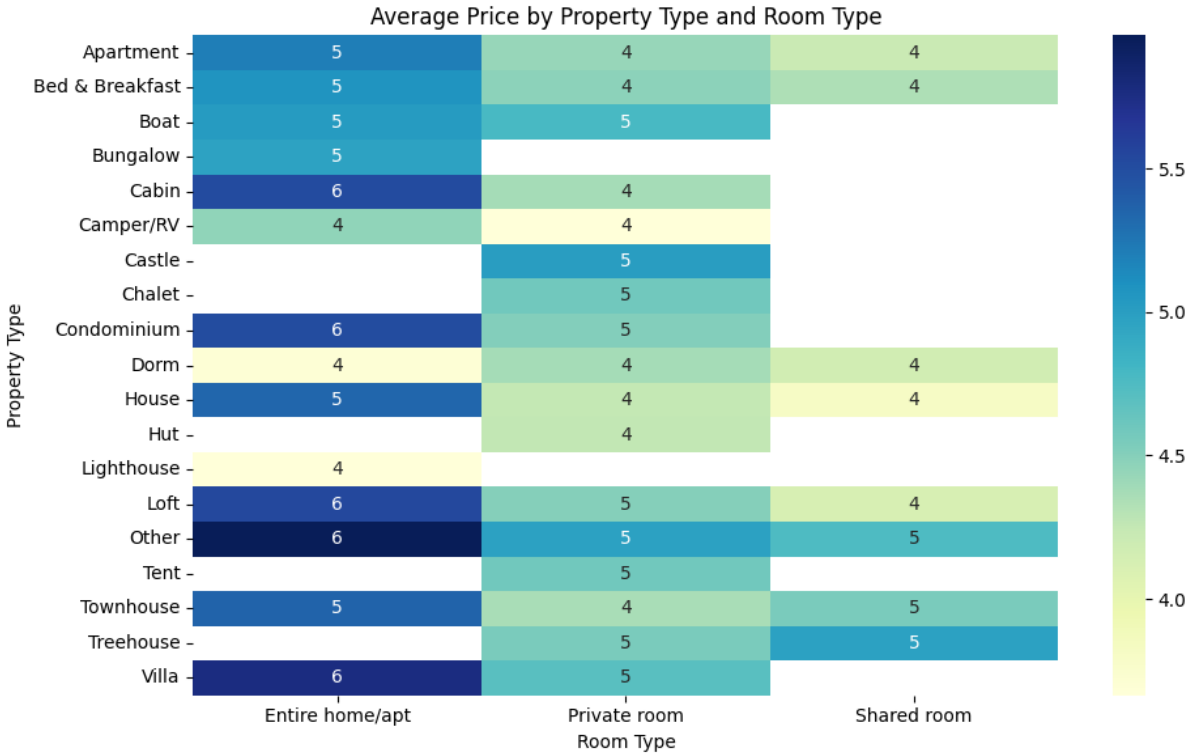
7.

Relationship between Review Scores Rating and Price (Scatter Plot)

This scatter plot highlights the **limited relationship between review scores and price**. High prices are found across the full spectrum of review scores, and similarly, low prices existacross all scores. The lack of a clear linear trend implies price is driven more by intrinsic value (location, size) than perceived quality (review score).



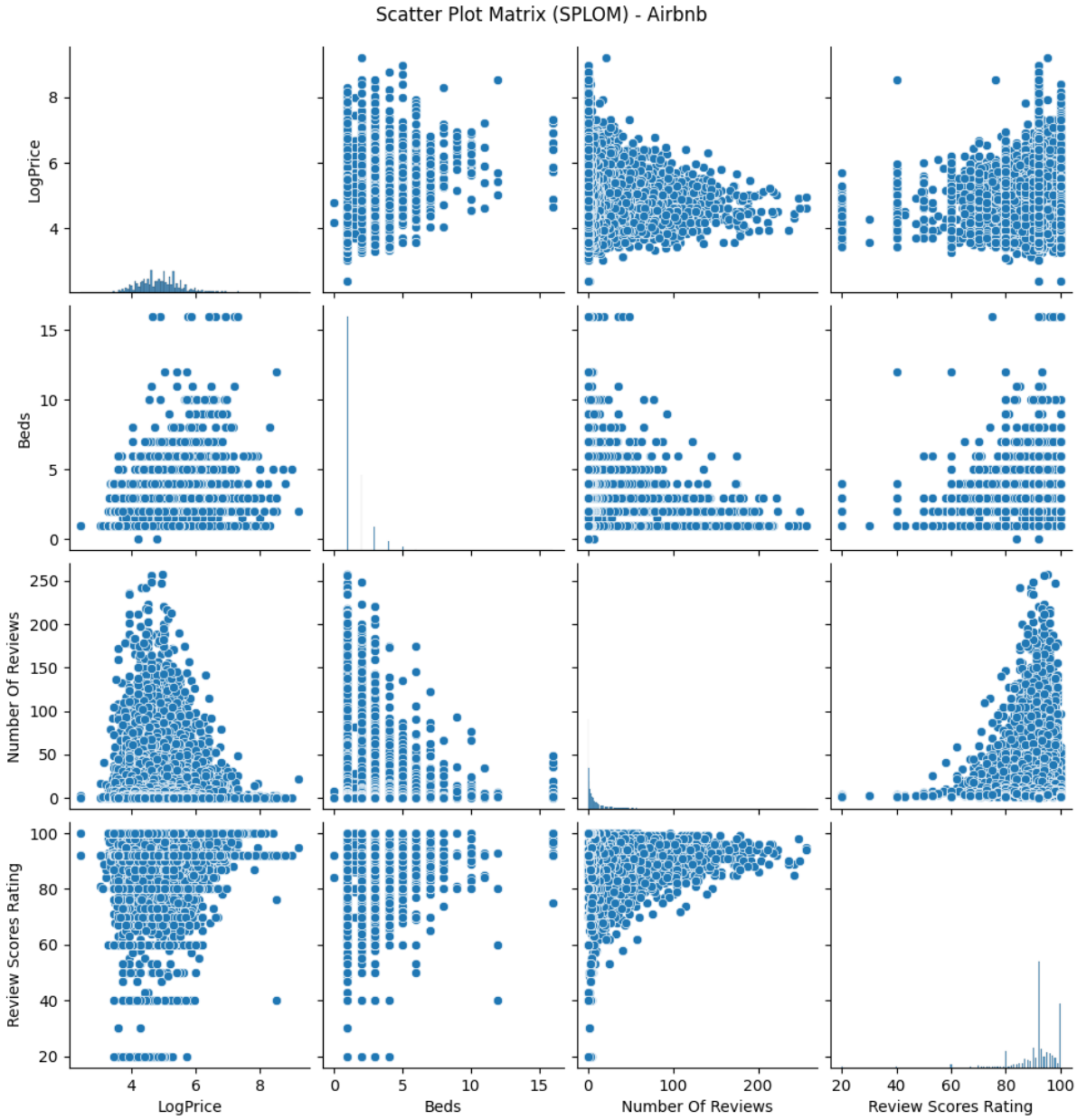
8. Joint Influence: Property Type and Room Type (Heat map)



This heat map visualizes how **Property Type and Room Type jointly influence Price**. By comparing the average log price, it shows that "Entire Home/Apt" in a high-end property type (like a Serviced Apartment or Loft) results in the highest price segment, demonstrating complex pricing interactions.

9. Scatter Plot Matrix (SPLOM)

The SPLOM provides a holistic view of relationships between numeric columns (Log Price, Beds, Number Of Reviews, Review Scores Rating). This matrix reconfirms the weak linear relationships found in the heat map, emphasizing that the complex pricing and performance of Airbnb listings are best understood through categorical variables rather than simple numeric correlations.

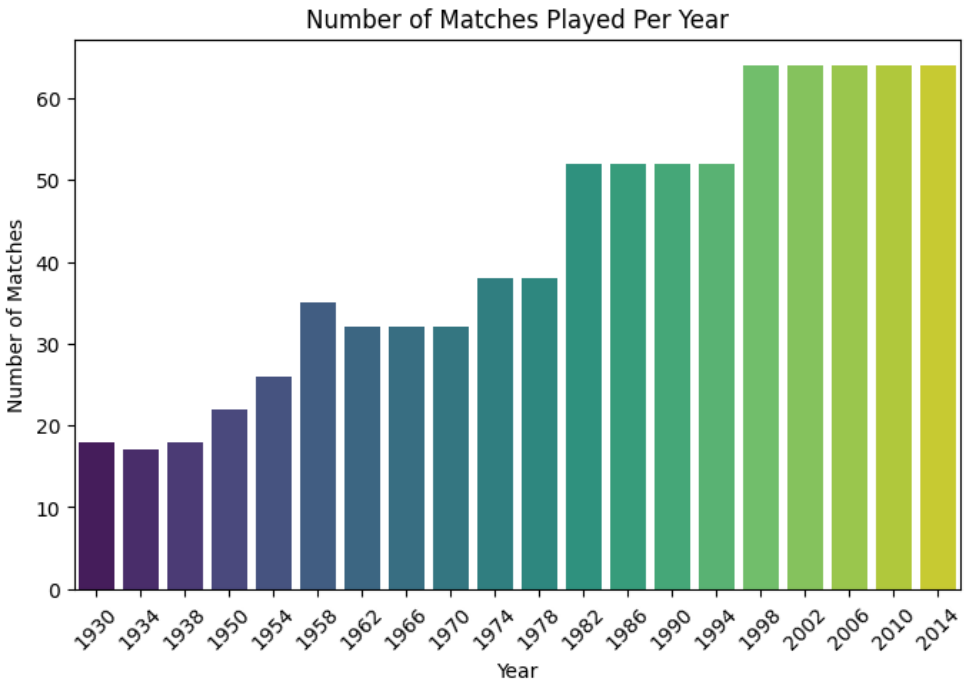


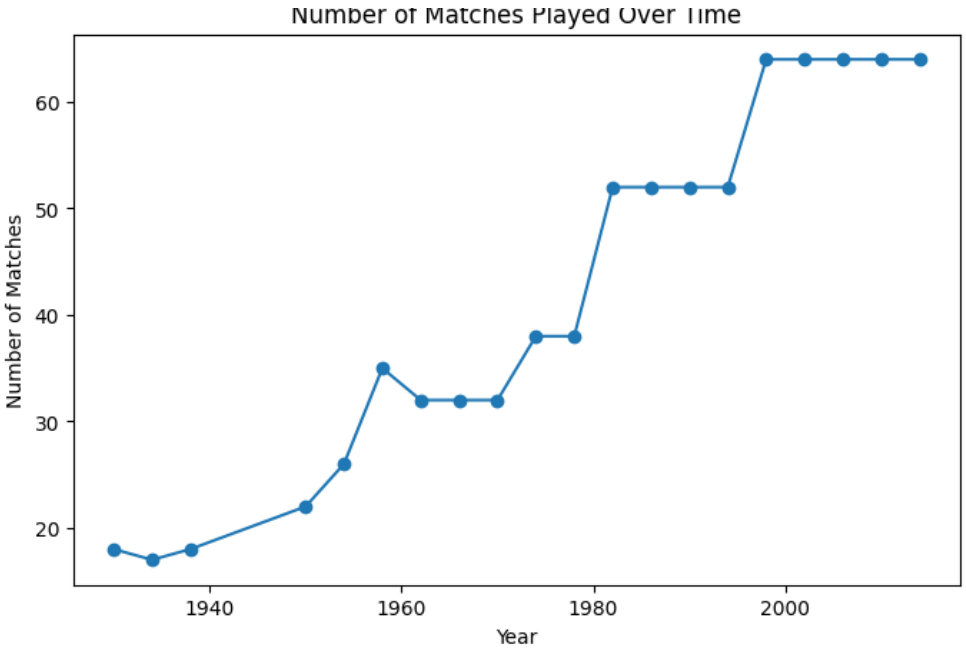
II. Visual Analysis Report: Football World Cup Results

The analysis of the World Cup data focused on tournament trends, historical goal metrics, and the influence of the 'Home' team designation.

1. Historical Growth (Bar and Line Plots)

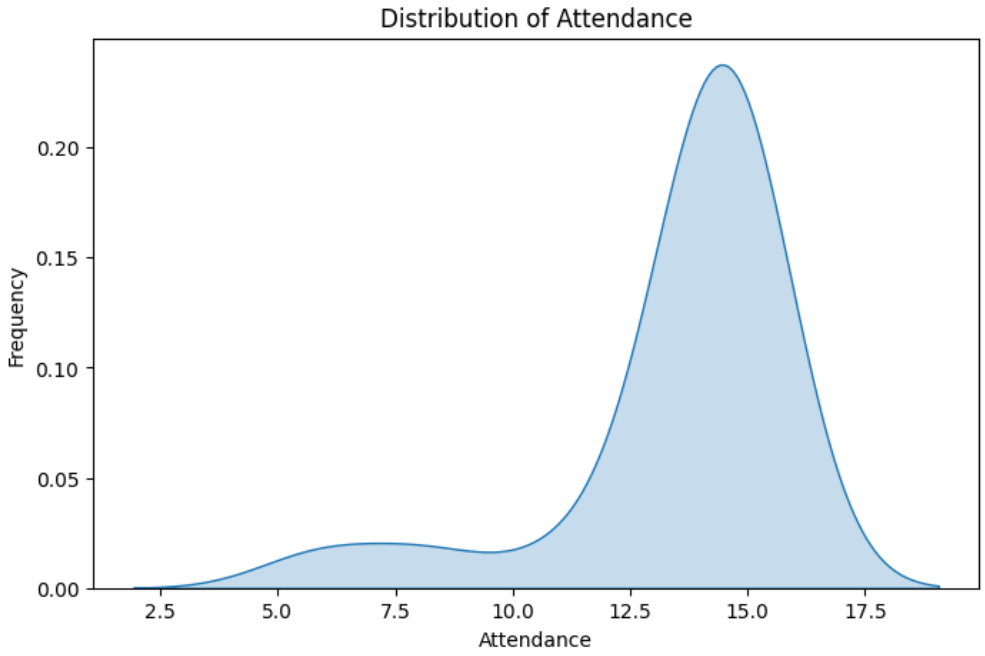
The line and bar plots of **Matches Played** over time show clear step-changes in the tournament structure, confirming that the **number of matches played has increased over time**. This increase is strongly correlated with the expansion of qualified teams and the tournament format changes.

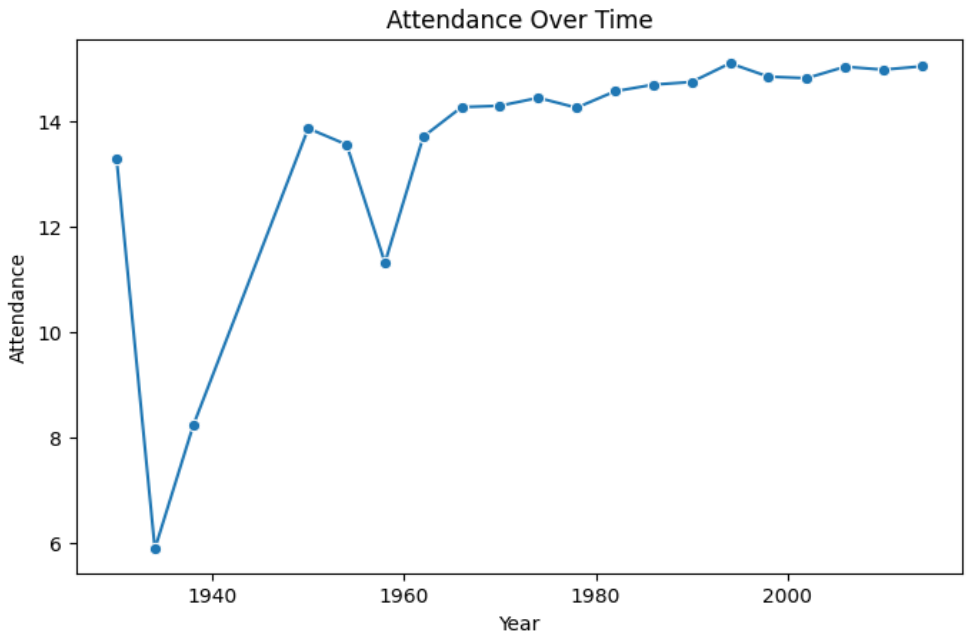


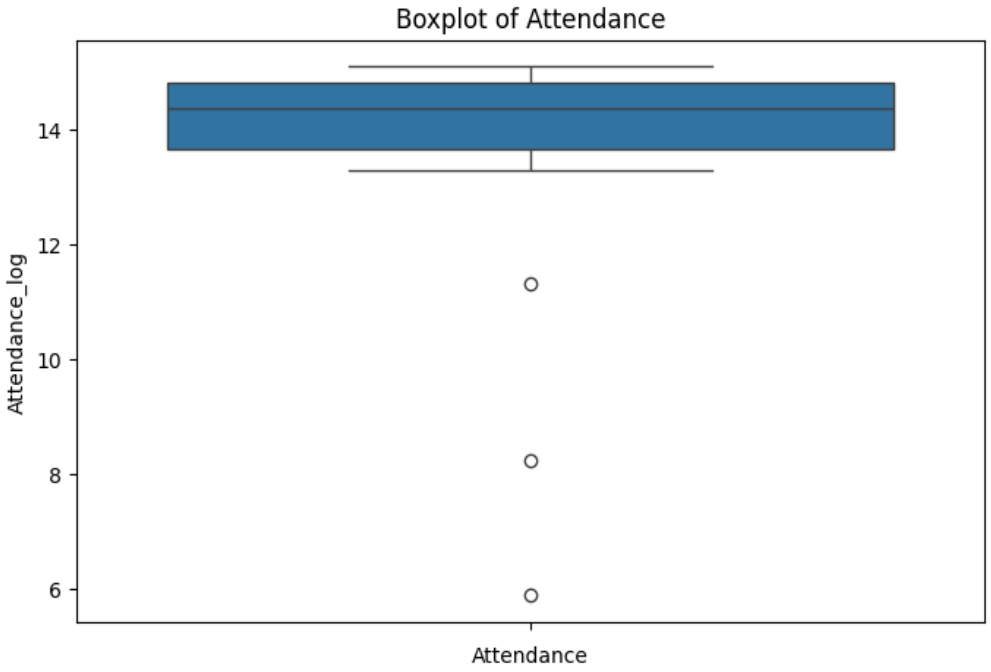


2. Distribution of Attendance (KDE, Box Plot, and Line Plot)

The log-transformed distribution shows that **attendance is left-skewed**, meaning most tournaments have high attendance, with lower attendance generally occurring only in the early years. The line plot over time confirms an upward trend, indicating that tournament size and global interest have steadily increased over the years.

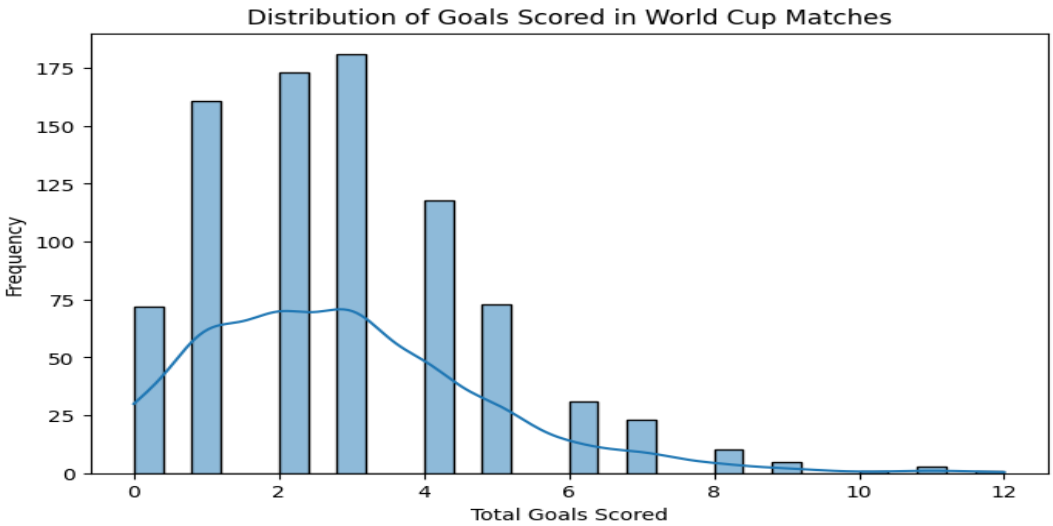


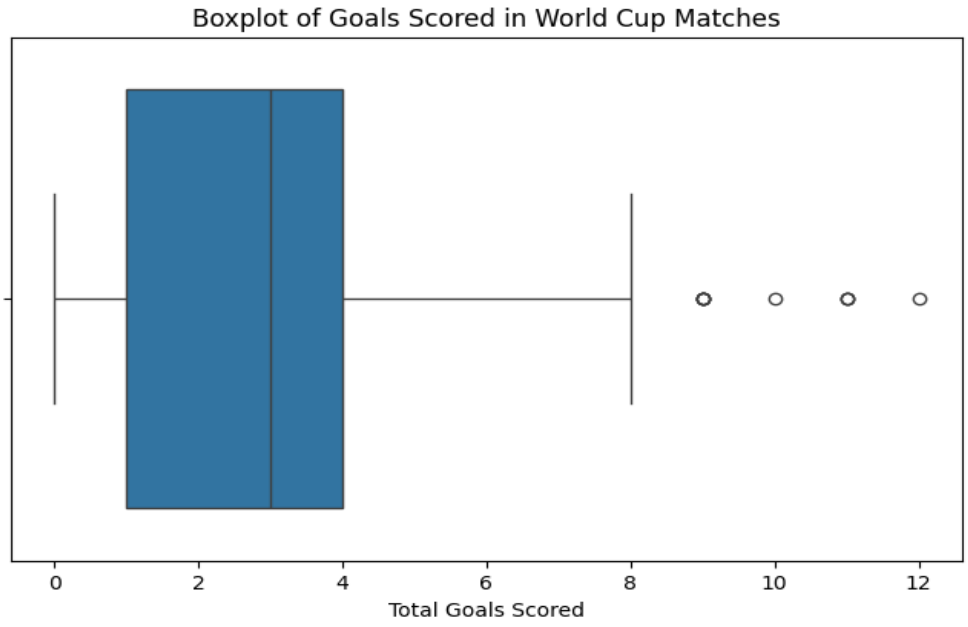




3. Distribution of Goals Scored (Histogram and Box Plot)

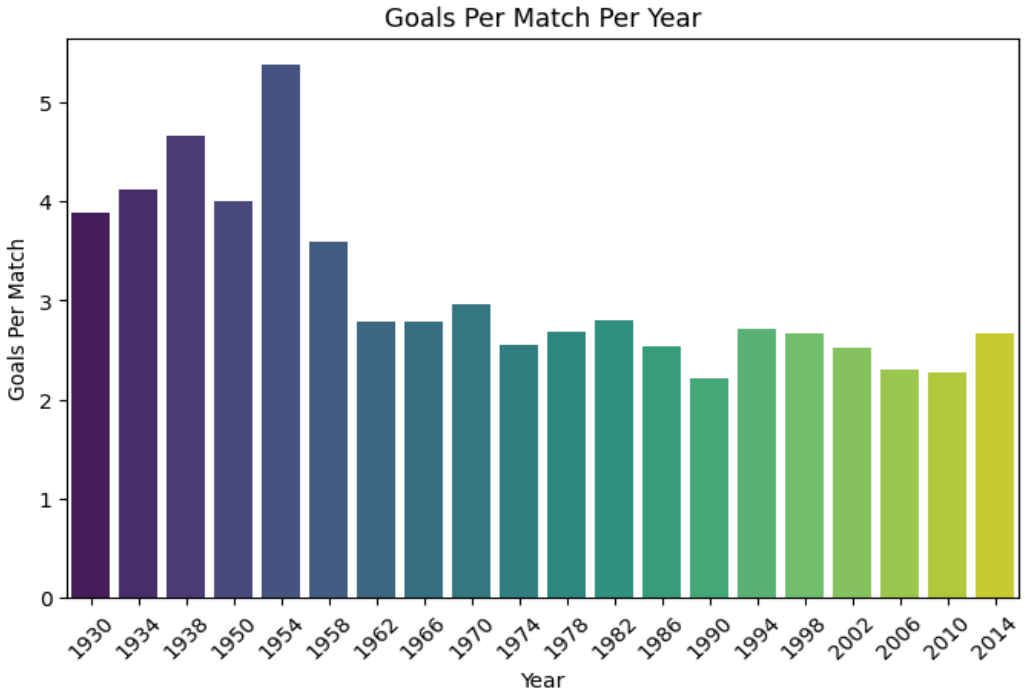
The histogram of goals scored per match is **right-skewed**, confirming that **most matches have low scores** (0-3 goals), and matches with very high scores are rare outliers. This skewness suggests that most games are tightly contested, supporting the notion of competitive balance.





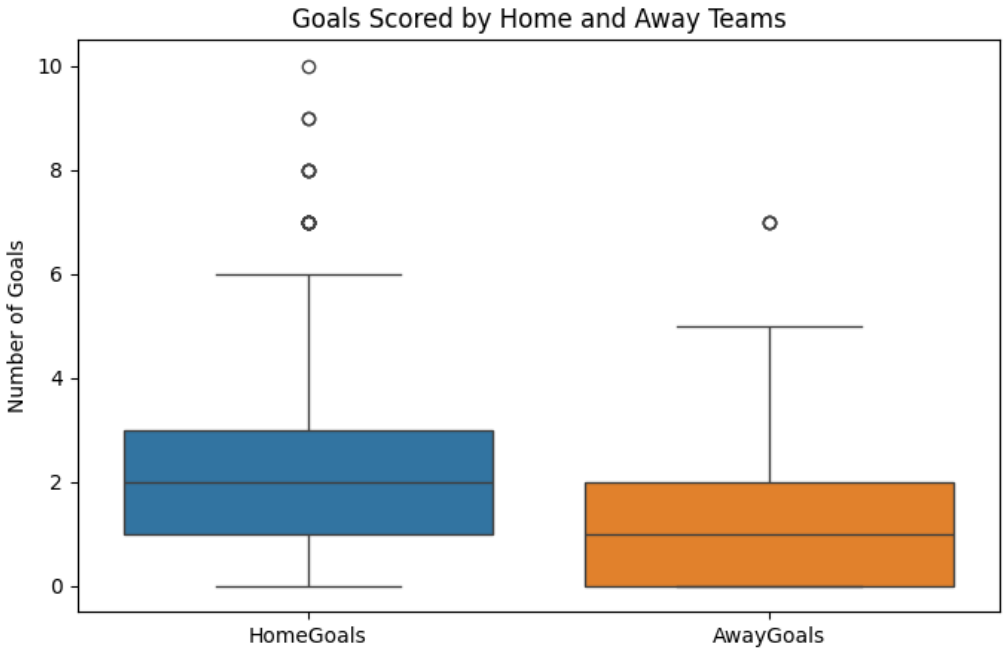
4. Goals Per Match Trend (Line and Bar Plots)

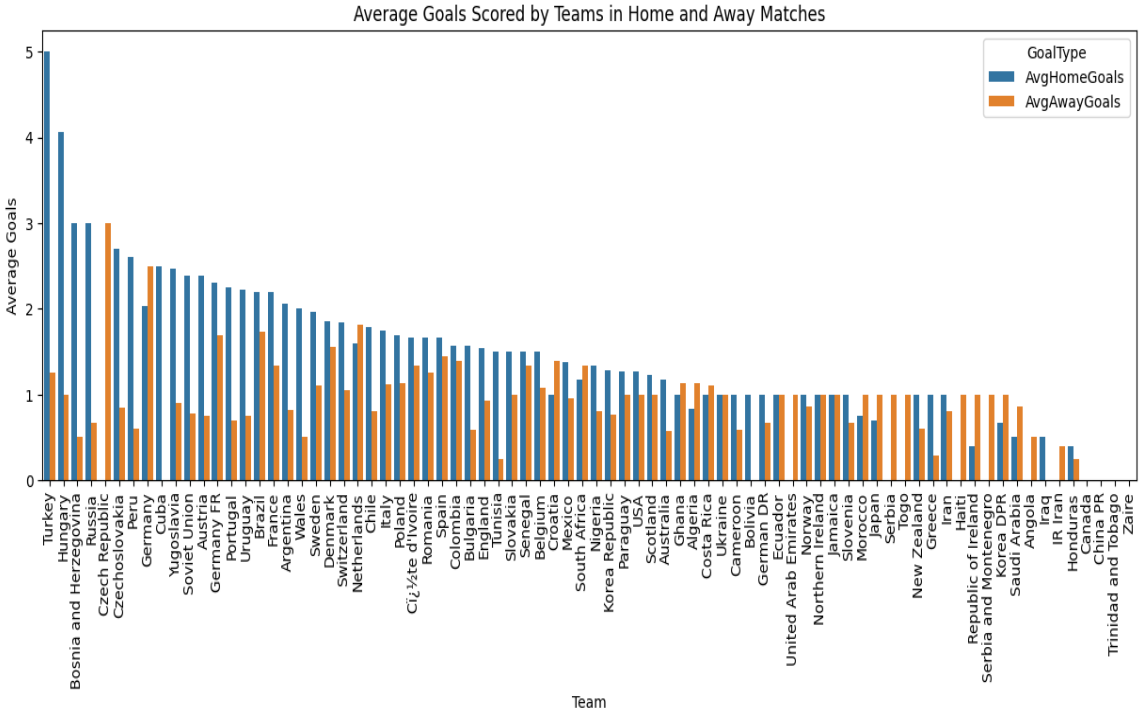
The line plot of **Goals Per Match Over Time** clearly shows a significant decline in the average goals per game since the mid-20th century. This supports the observation that goals per match have decreased over time, likely **due to more tactical and defensive play** as global football standards have risen.



5. Home vs. Away Goals (Box Plot and Bar Plot)

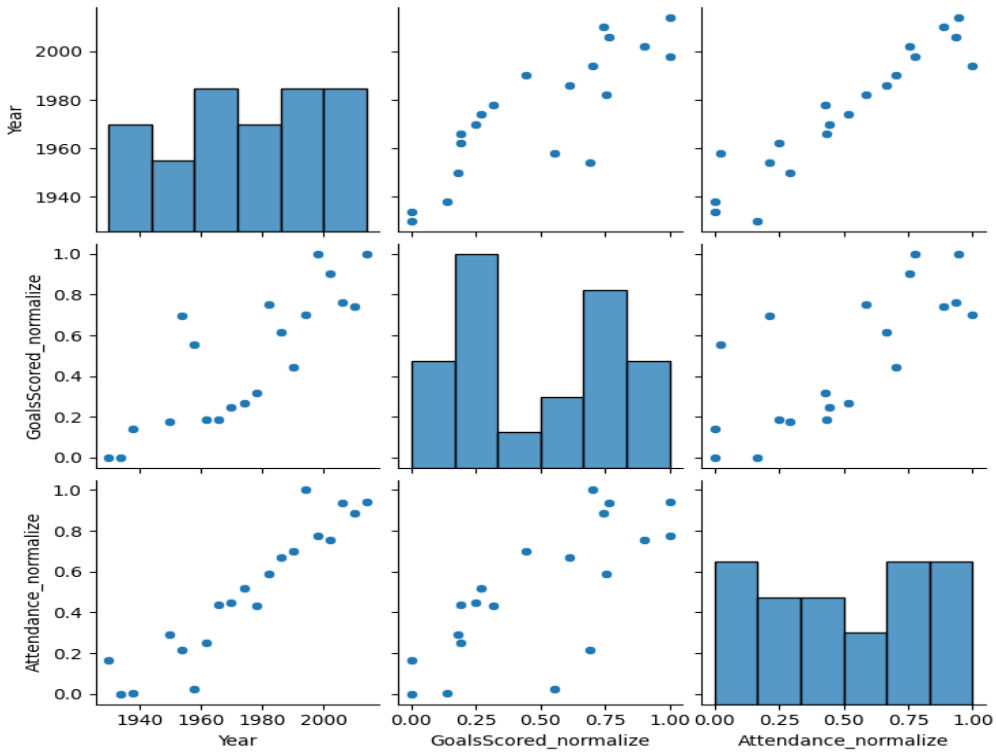
The box plot confirms that **Home teams tend to score more goals than away teams**. The average goals bar plot, which groups by team, further illustrates this by showing that most teams have a higher average goal tally when designated as the 'Home Team' in a match, highlighting a clear structural advantage.

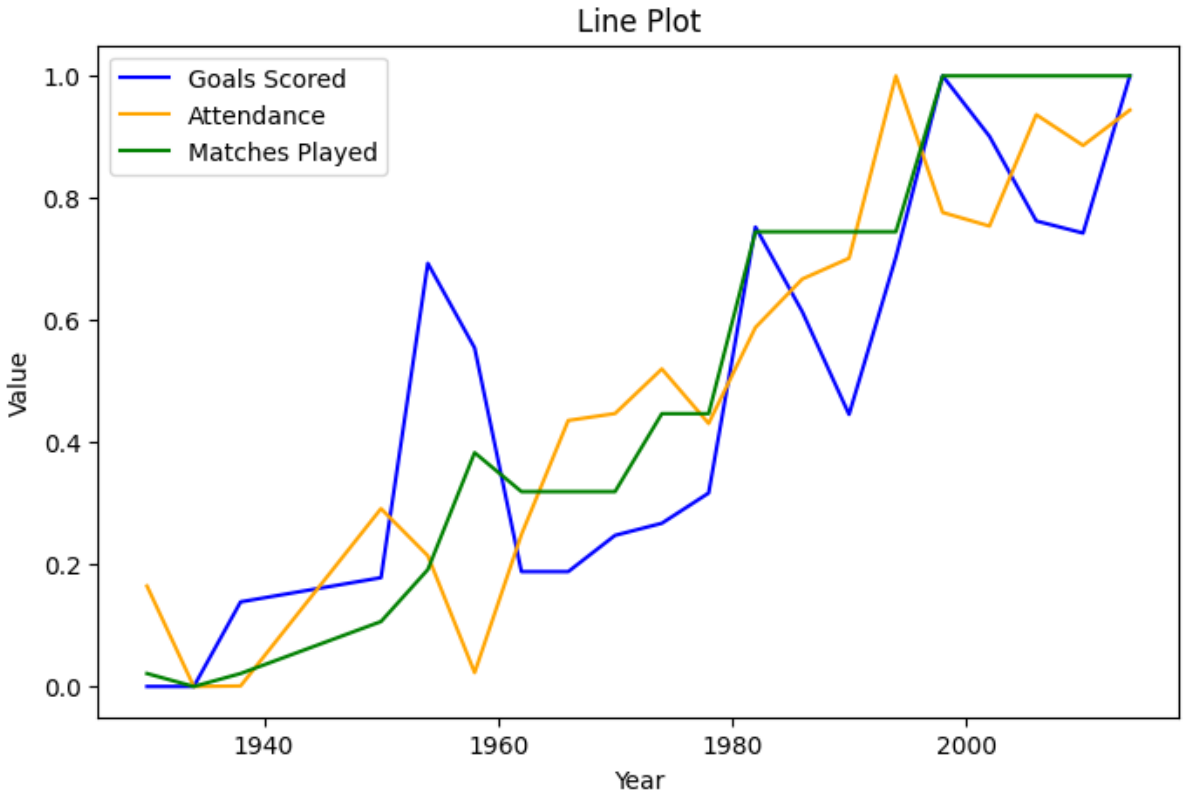


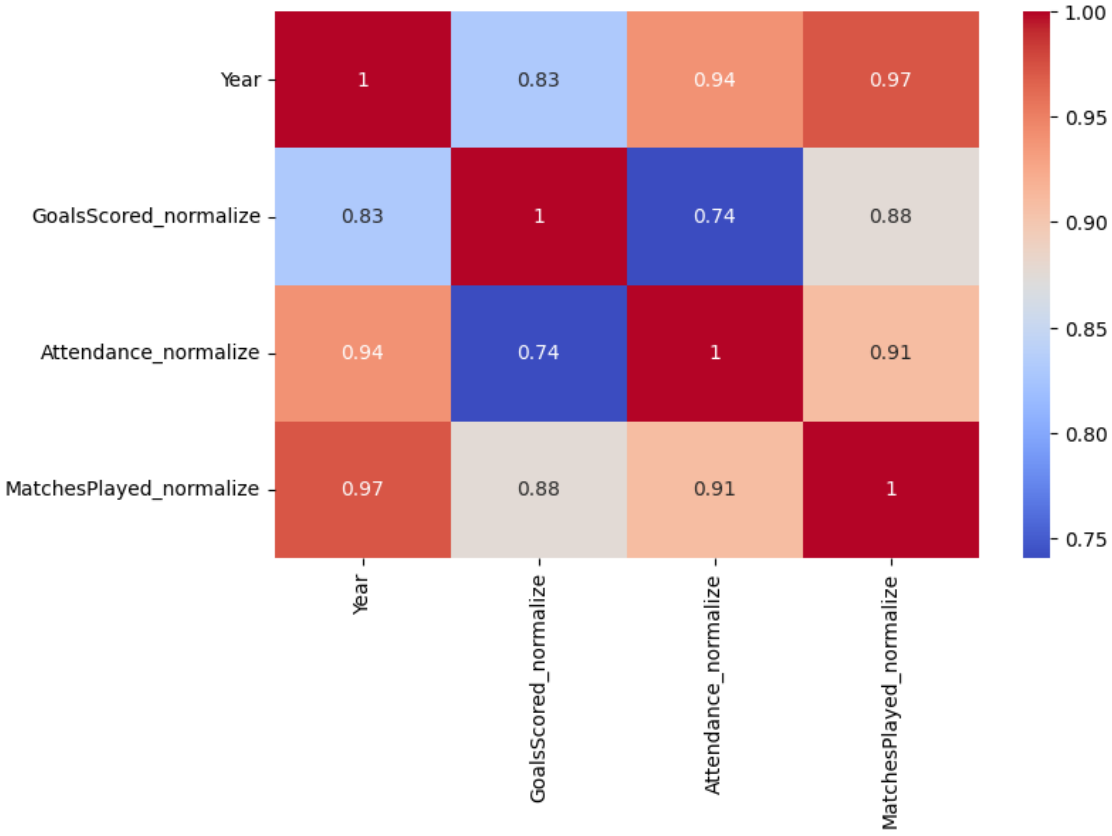


6. Interrelation of Tournament Stats (Line Plot, SPLOM, and Heatmap)

The normalized line plot shows that **Tournament year, goals scored, and attendance are highly interrelated**. As the tournament year increases, all normalized metrics (**Attendance**, **Goals Scored**, and **Matches Played**) show a general upward trend, confirming the tournament's overall growth in scale and reach. The heatmap further confirms the strong positive correlations between these scaled metrics.

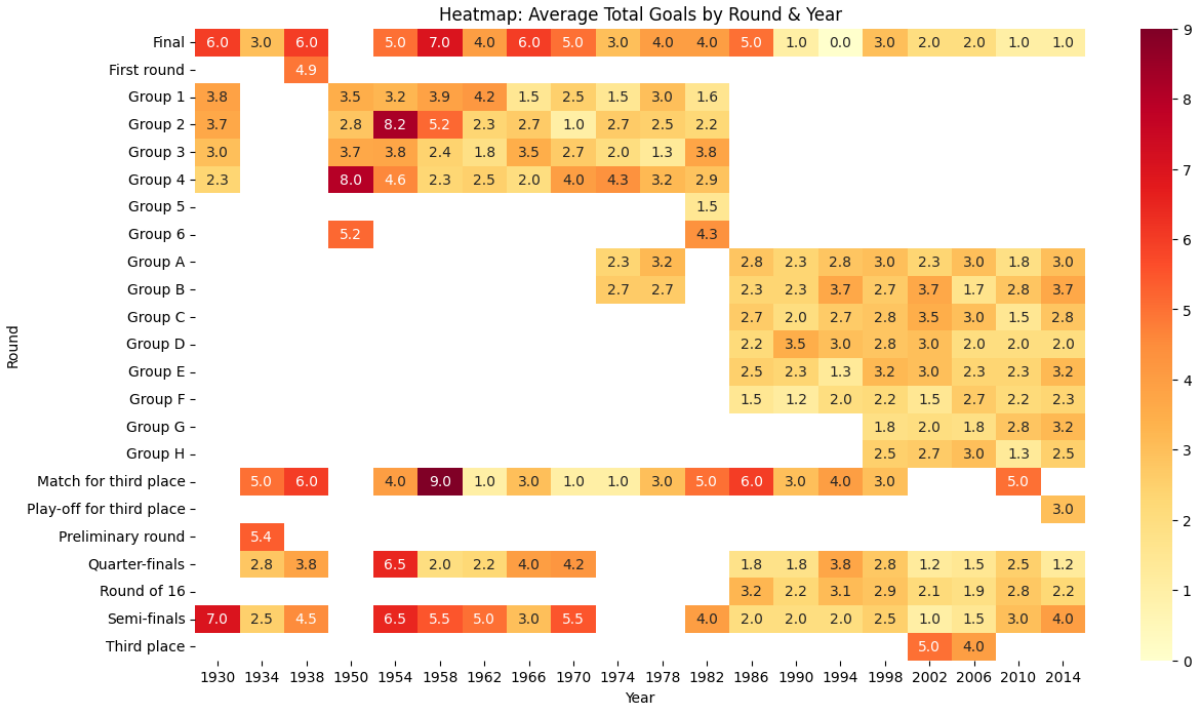


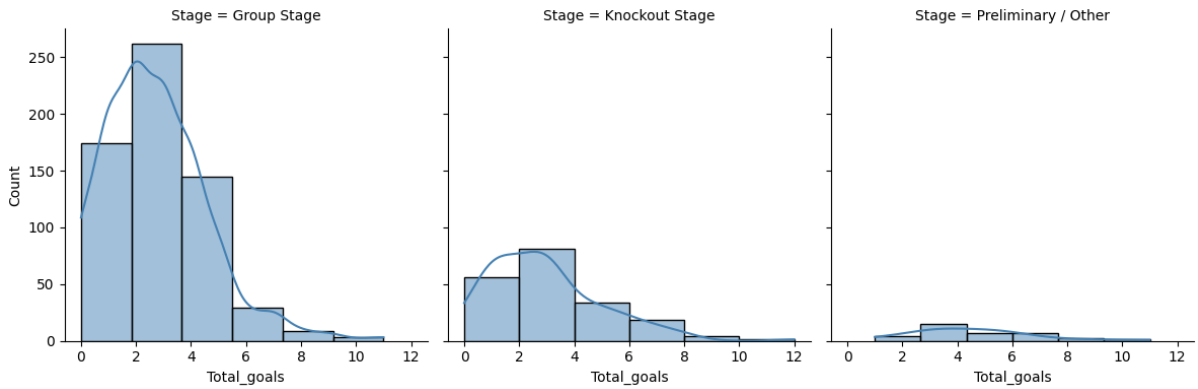

Line Plot
Goals Scored
Attendance
Matches Played
Value
Year
1.0
0.8
0.6
0.4
0.2
0.0
1940
1960
1980
2000


Year
GoalsScored_normalize
Attendance_normalize
MatchesPlayed_normalize
1
0.83
0.94
0.97
0.83
1
0.74
0.88
0.94
0.74
1
0.91
0.97
0.88
0.91
1
1.00
0.95
0.90
0.85
0.80
0.75

7. Goals by Stage and Year (Heatmap and Trellis Plot)

The heatmap showing **Average Total Goals by Round & Year** reveals that later rounds (**Knockout Stage**) do not consistently have lower goal averages than the **Group Stage**. The Trellis plot of total goals by stage (Group Stage vs. Knockout Stage) shows that while the most extreme high-scoring outliers occur in the Group Stage, the distribution of total goals per match is similar, implying scoring activity isn't drastically reduced by the higher stakes of the Knockout Stage.





III. Insights and Conclusions

Summary of Key Insights

- **Airbnb Pricing is Structural:** The NYC Airbnb market is governed by a clear hierarchy where Room Type is the primary driver of price, followed by Neighbourhood Group.
- Quality is Inelastic**:** The minimal correlation between Review Scores Rating and Price implies that demand for core features (location, size) is high and relatively inelastic to perceived quality.
- **World Cup Growth in Scale, Decline in Pace:** The World Cup has experienced massive growth in Attendance and Matches Played over time, but the excitement, measured by Goals Per Match, has steadily declined.
- **Home Team Advantage:** There is a quantifiable home team advantage in terms of average goals scored, suggesting a psychological or logistical bias within the match data.
- **Initial Hypothesis:** The general hypothesis that Manhattan listings are the most valuable and that World Cup metrics show consistent growth is strongly supported by the visual analysis, particularly in terms of scale (attendance, matches) but refuted in terms of scoring pace (goals per match).

Unexpected Findings and Implications

The most unexpected finding is the clear long-term decline in Goals Per Match**.**

Implication (Football): Despite increases in participation and spectacle, the game itself has become more defensive and tactical. This could be relevant for analyzing future tournament formats and rules designed to promote offensive play.

Implication (Airbnb): The lack of correlation between Price and Review Scores suggests hosts can potentially reduce spending on minor amenities and focus capital on location, capacity, or property type upgrades.

IV. Recommendations for Further Analysis

To deepen the understanding of both datasets, the following areas and variables should be explored:

1. **Airbnb: Host Strategy and Availability:**
    - Analyze the correlation between listing price, host cancellation rate, and minimum nights required to identify dynamic pricing strategies.
    - Use geospatial clustering to formally define sub-markets within the major boroughs beyond the simple "Neighbourhood Group" aggregation.
2. **Football World Cup: Deeper Performance Metrics:**
    - **Goal Difference (Dominance Metric):** Calculate and visualize the average goal difference per match for the top 10 nations to better quantify long-term dominance.
    - **Penalty Shootout Impact:** Analyze the frequency and success rate of penalty shootouts across different years/rounds, as this is a key factor in knockout stage outcomes.