End-to End Data Analysis

Turning Data into Insights

Zhoubin Maneshi

07.03.2025

Objectives

5 Key Steps to Analyze Data

To analyze data effectively, start by understanding the dataset—loading it, checking its structure, and identifying missing values. Next, clean and preprocess the data by handling missing values, fixing data types, removing outliers, and standardizing where necessary. Once the data is ready, explore it using visualizations, correlation analysis, and pattern detection to uncover trends. Then, apply statistical methods or build predictive models, validating their performance to ensure accuracy. Finally, present insights through charts, tables, and reports, making the findings clear and actionable for decision-making.

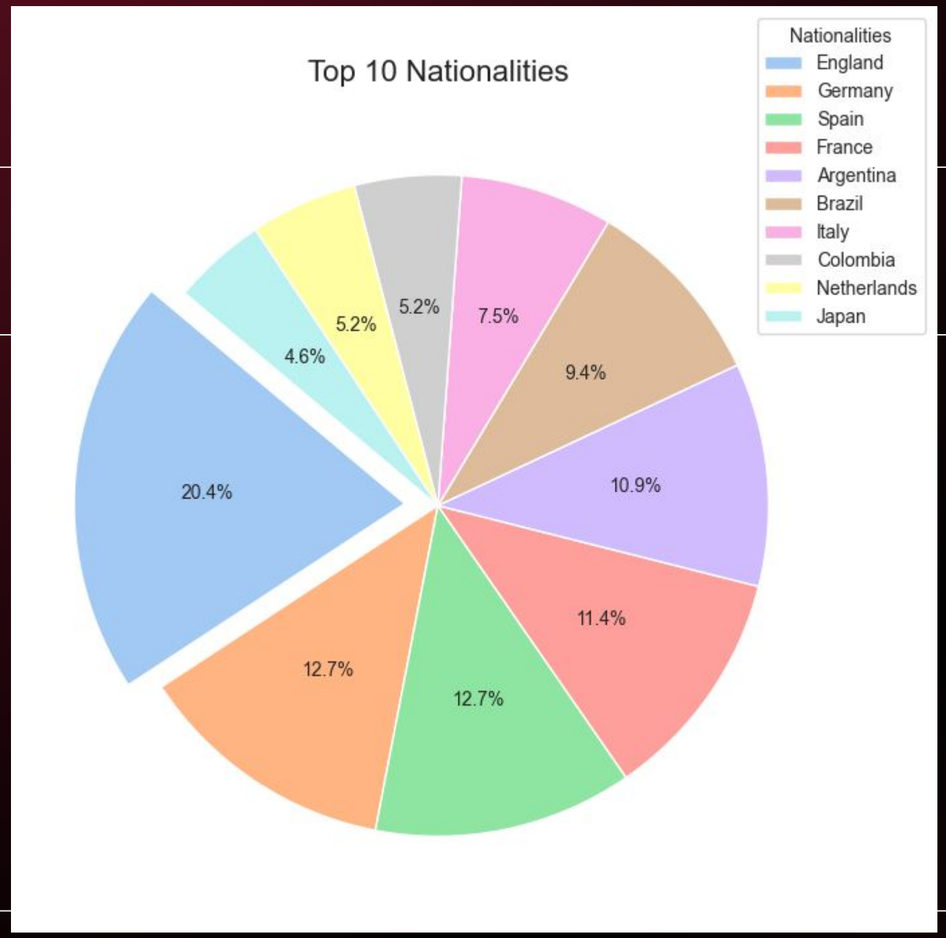
1. UNDERSTAND DATA
2. CLEAN & PREPROCESS
3. ANALYZE DATA
4. VISUALIZE FINDINGS
5. INTERPRET RESULTS

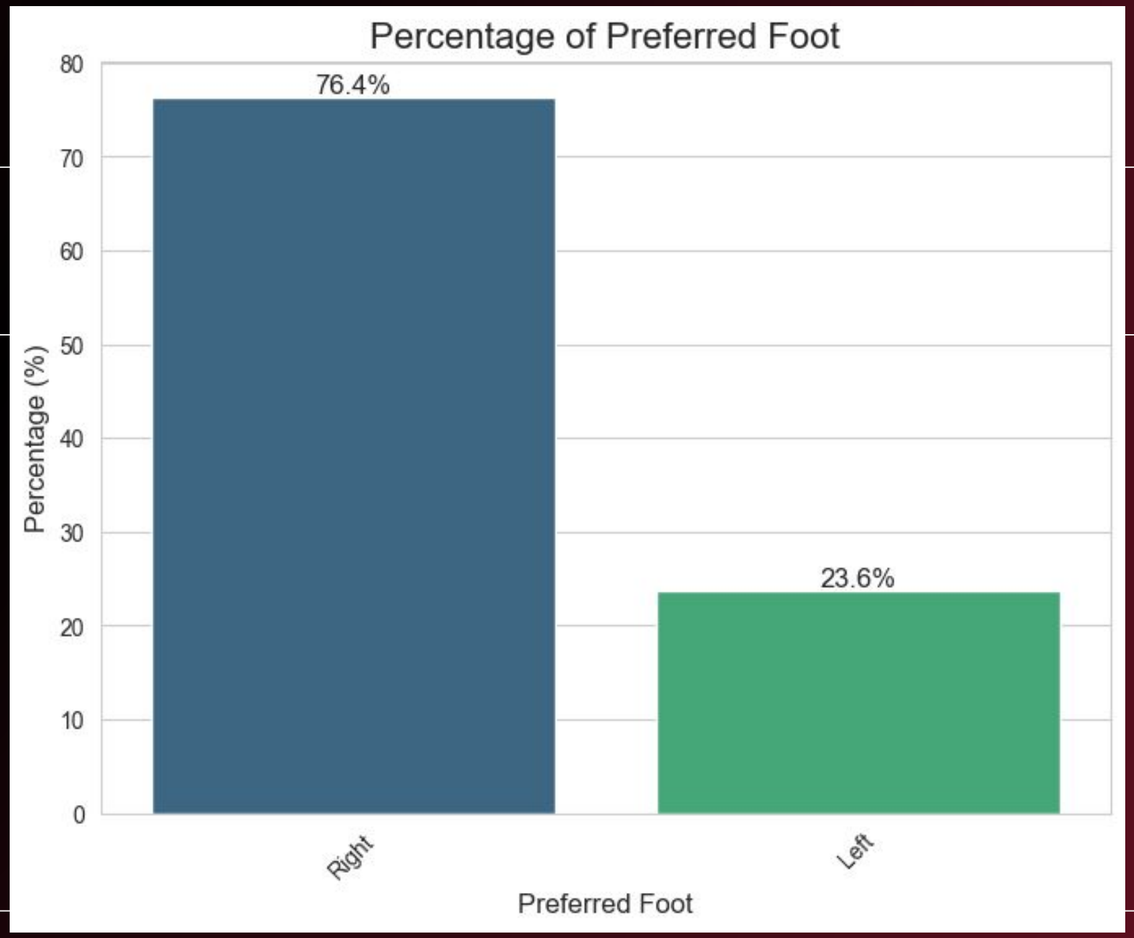
FIFA GAME DATASET

Making Sense of Raw Data

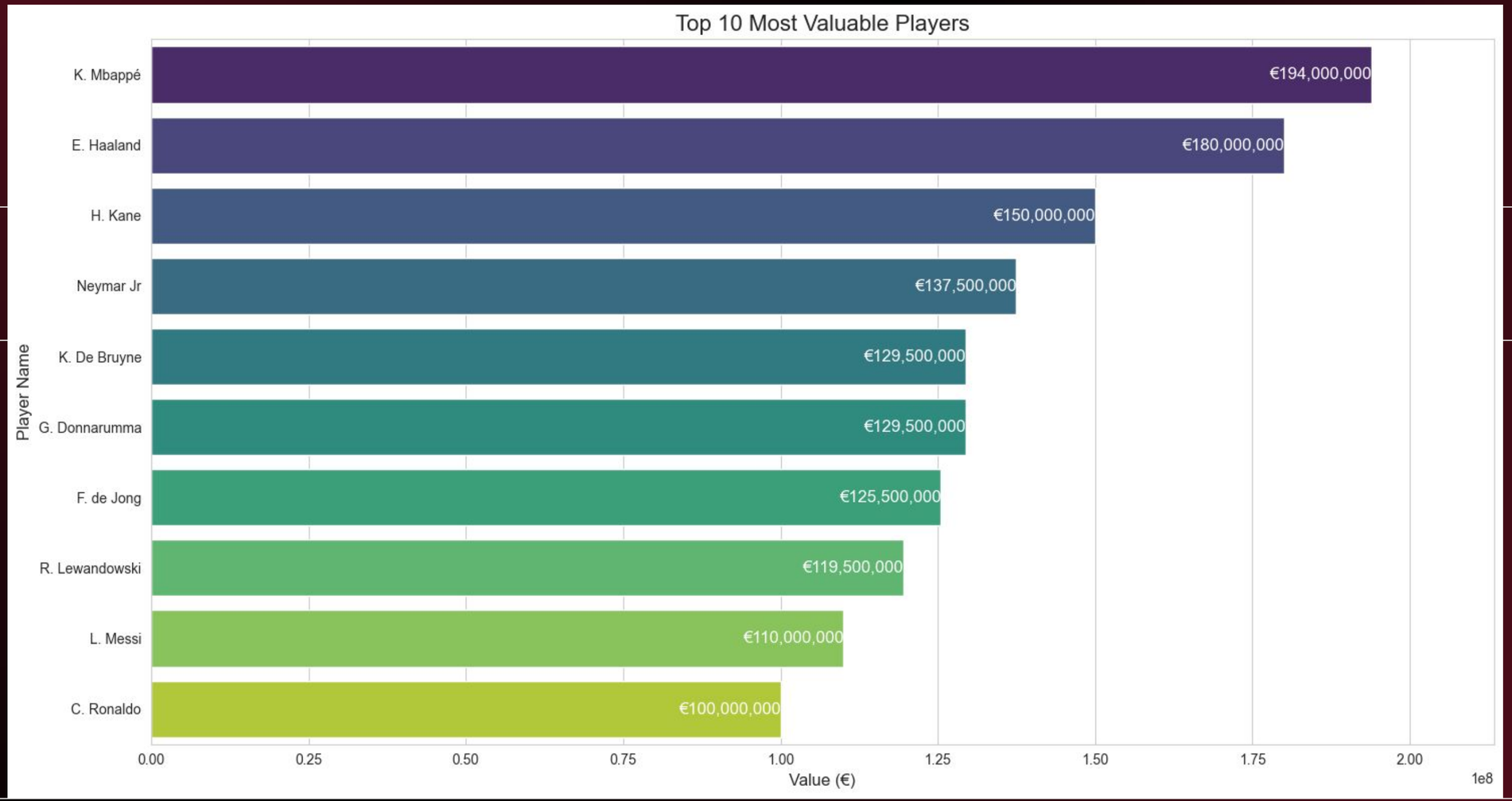
Seven separate raw FIFA datasets from 2017 to 2023 were sourced from Kaggle, available at this link. This is a rich dataset with 66 columns, each assigning a numerical or categorical value to a player. The preprocessing began with a quick data survey using standard pandas methods to understand the structure. The datasets were then concatenated into a single DataFrame, which was subsequently cleaned by removing empty spaces, dropping duplicates, filling missing values, and deleting special characters from numerical columns. All columns with more than 30% missing values were rejected. Finally, the processed data was saved as a single CSV file named **"cleaned_data"** for further analysis.

Univariate: Pie Chart of Top 10 Nationalities
Bar chart of Preferred Foot Percentage

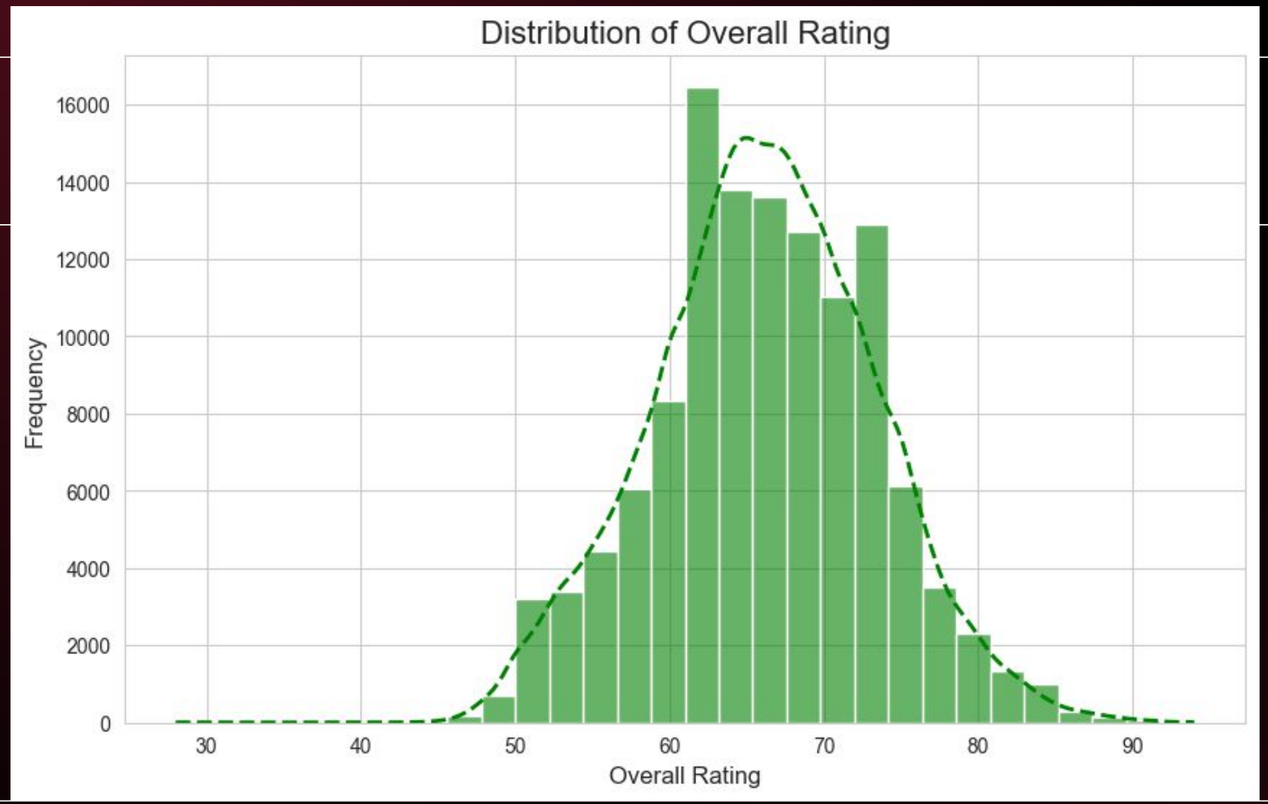


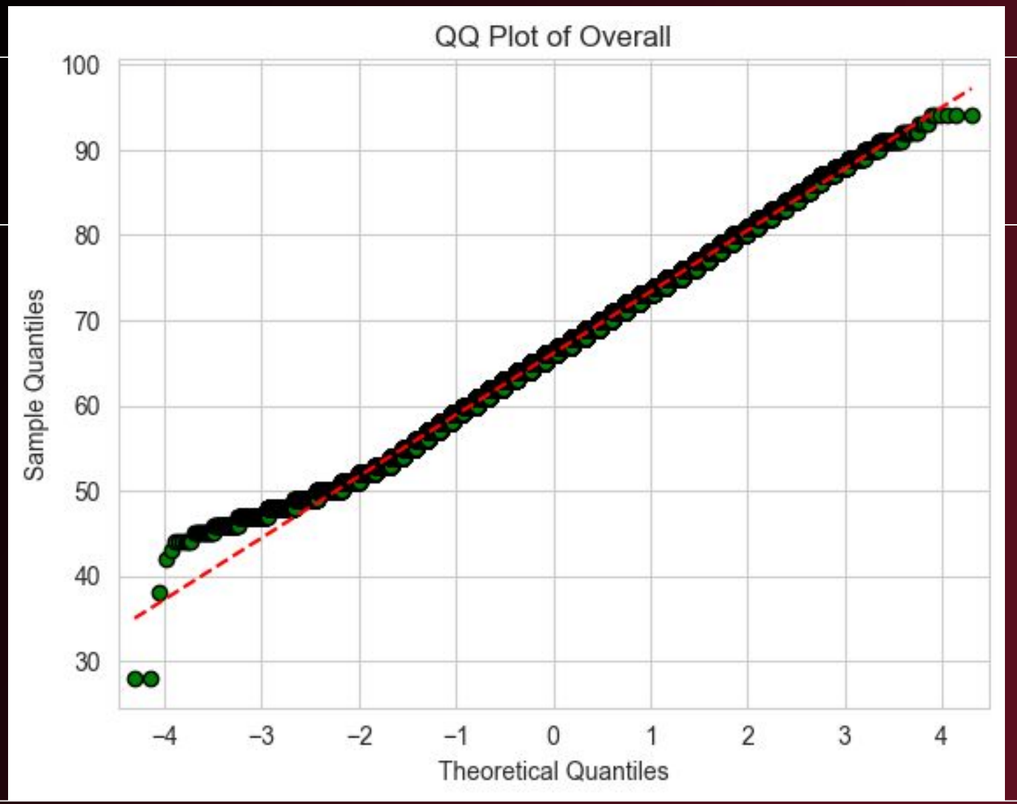


Univariate: Value Bar Chart for Top 10 Players



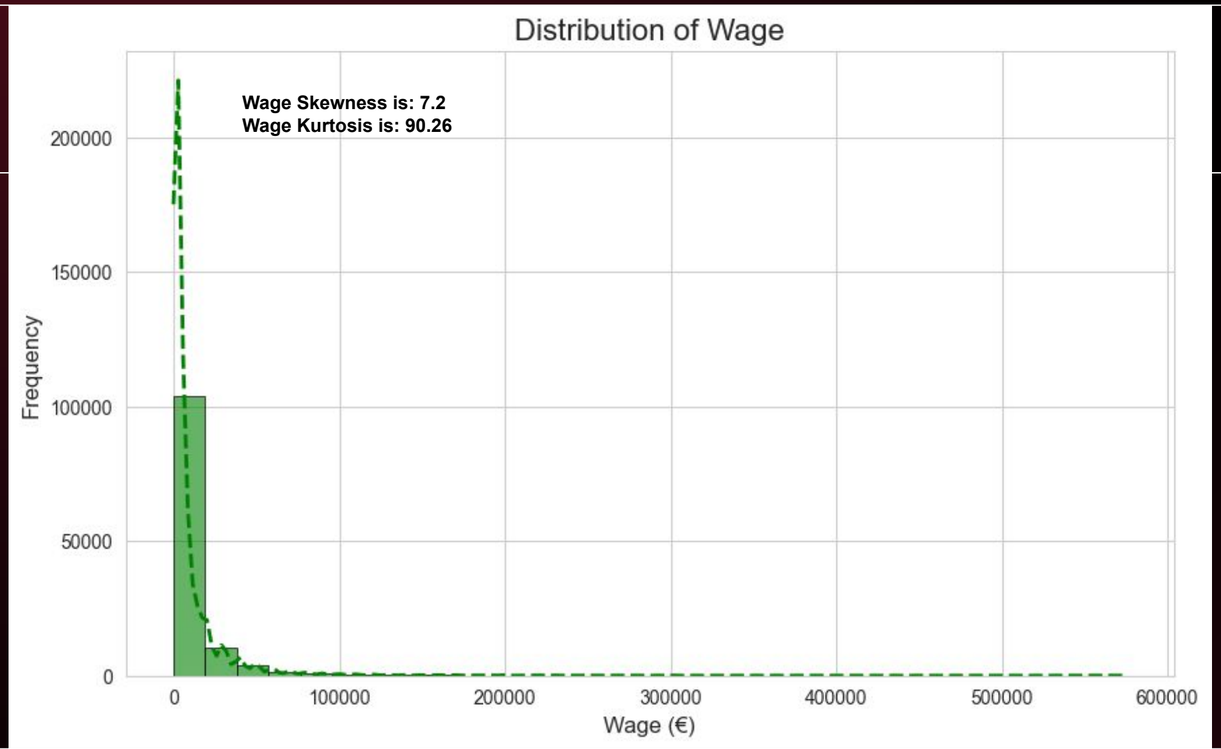
Univariate: Distribution of Overall Rating and its QQ Plot

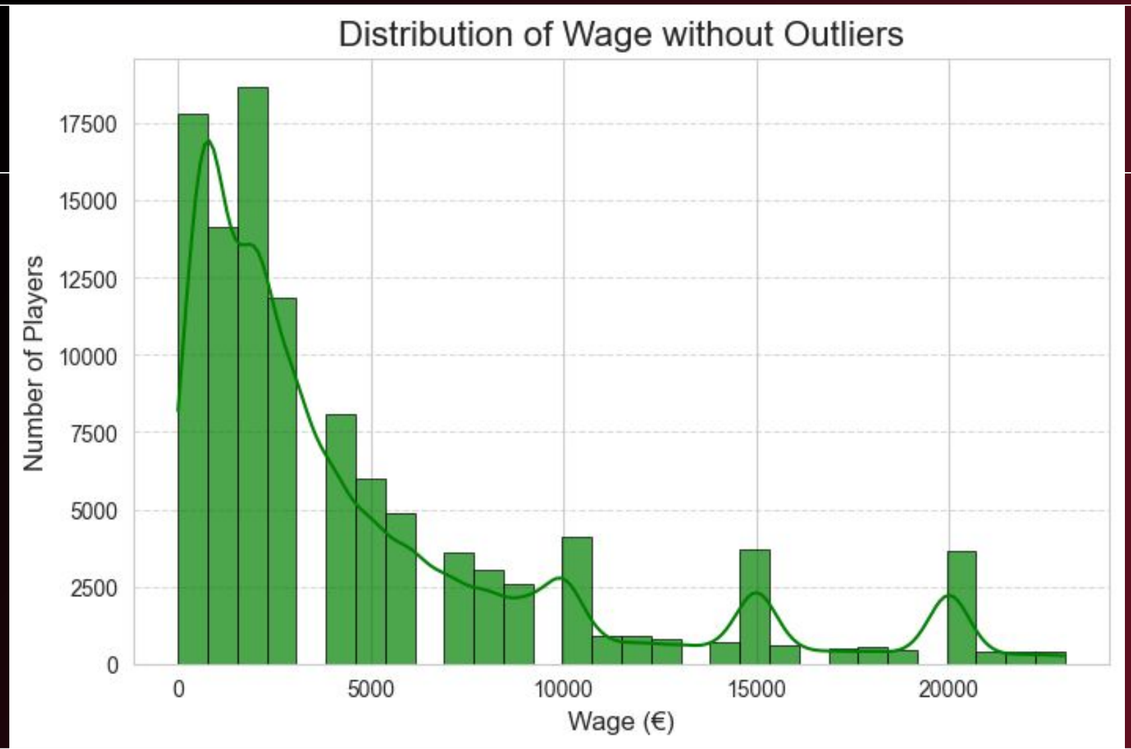




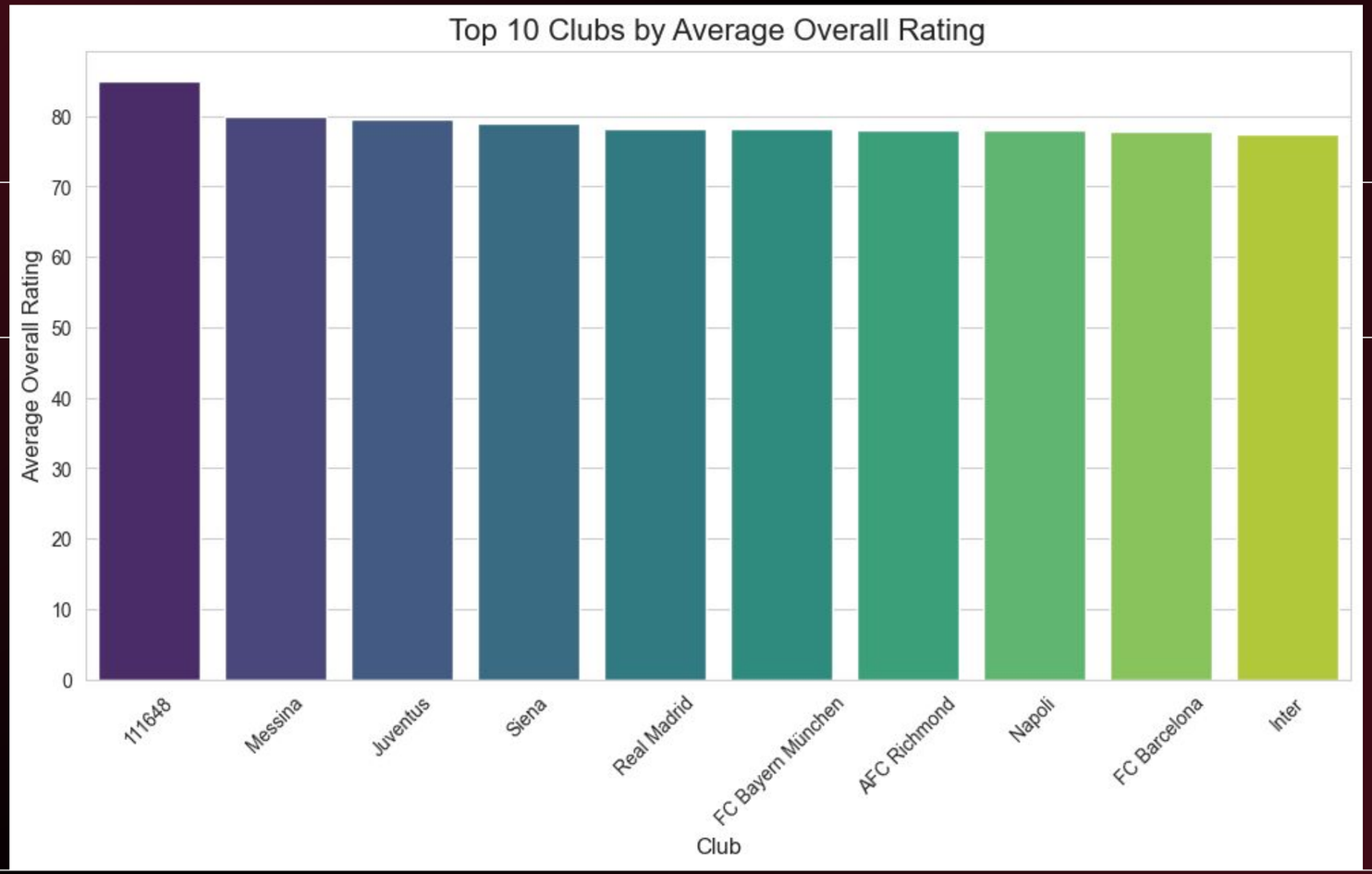
Univariate: Distribution of Wage (number of players vs. Wage)

Wage Skewness: 7.2
Wage Kurtosis: 90.26

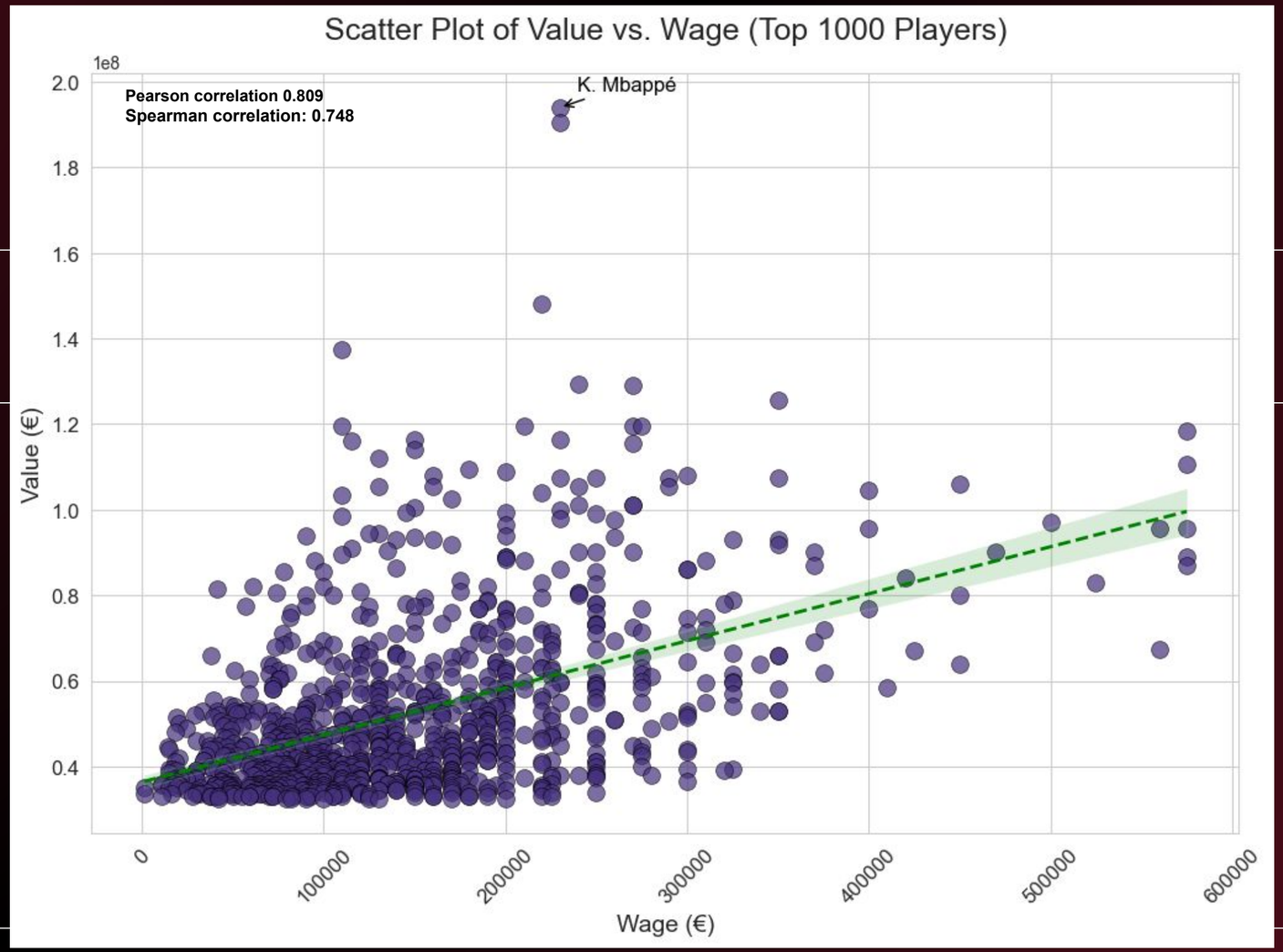




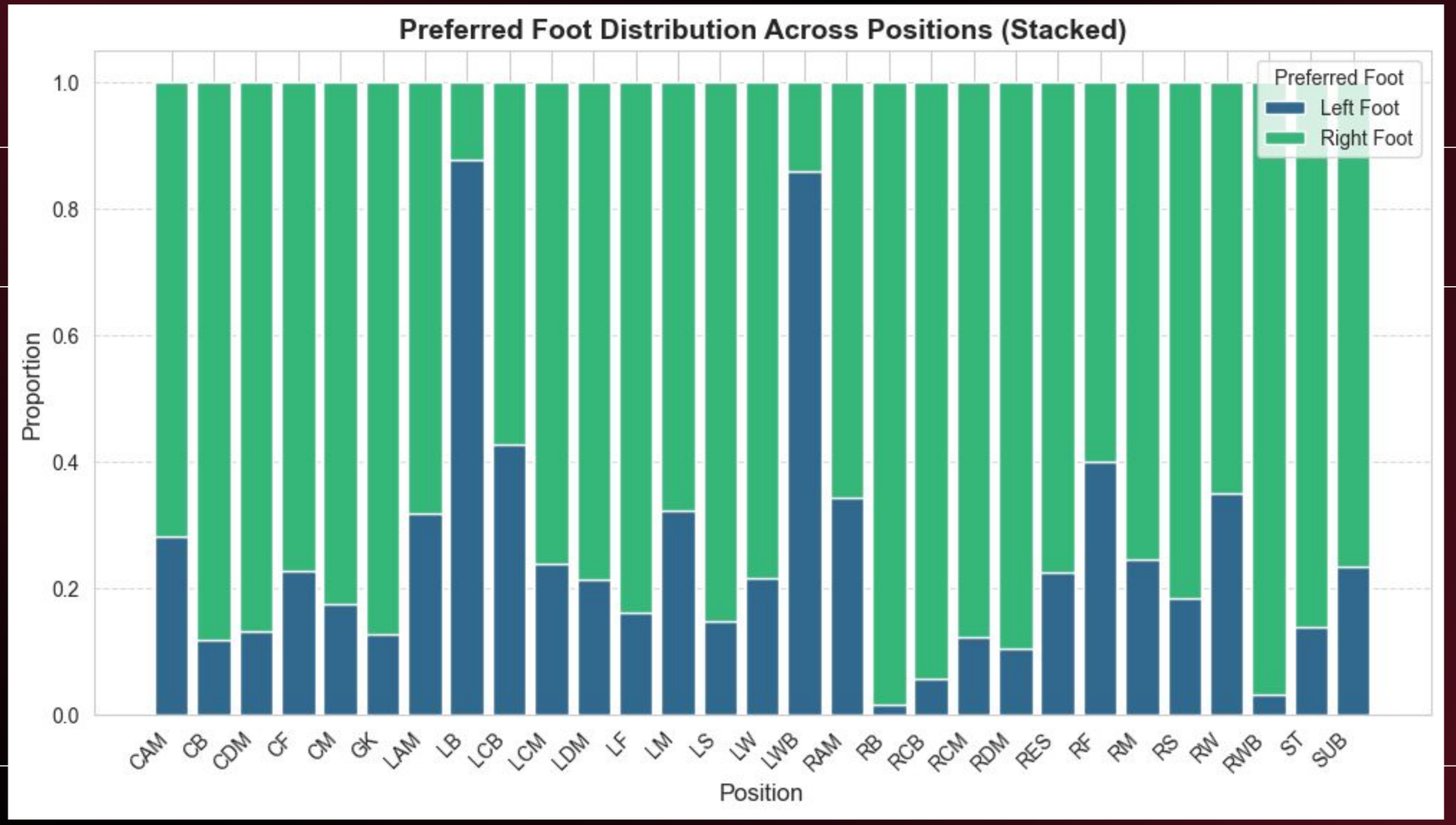
Bivariate: Average Overall Rating of top 10 Clubs



Bivariate: Scatter plot of Value vs. Wage



Bivariate: Stacked bar chart (how left/right-footed players are distributed in different positions)



Bivariate

Hypothesis from the observation

- Left-footed Players: Positions like LB and LW tend to have a higher proportion of left-footed players. This is common in football, as left-footed players are often preferred for these positions due to their ability to deliver crosses and cut inside effectively.

- Right-footed Players: Positions like CM, RB, and ST are dominated by right-footed players. This is expected, as right-footed players are more common in general, and these positions often require strong passing and shooting with the dominant foot.

- Balanced Positions: Positions like RW and CB have a more balanced distribution, indicating that both left-footed and right-footed players can excel in these roles.

Bivariate

Statistical Tests

- Chi-squared test yields P-value: 0.999998
- Cramer's V yields: 0.006983

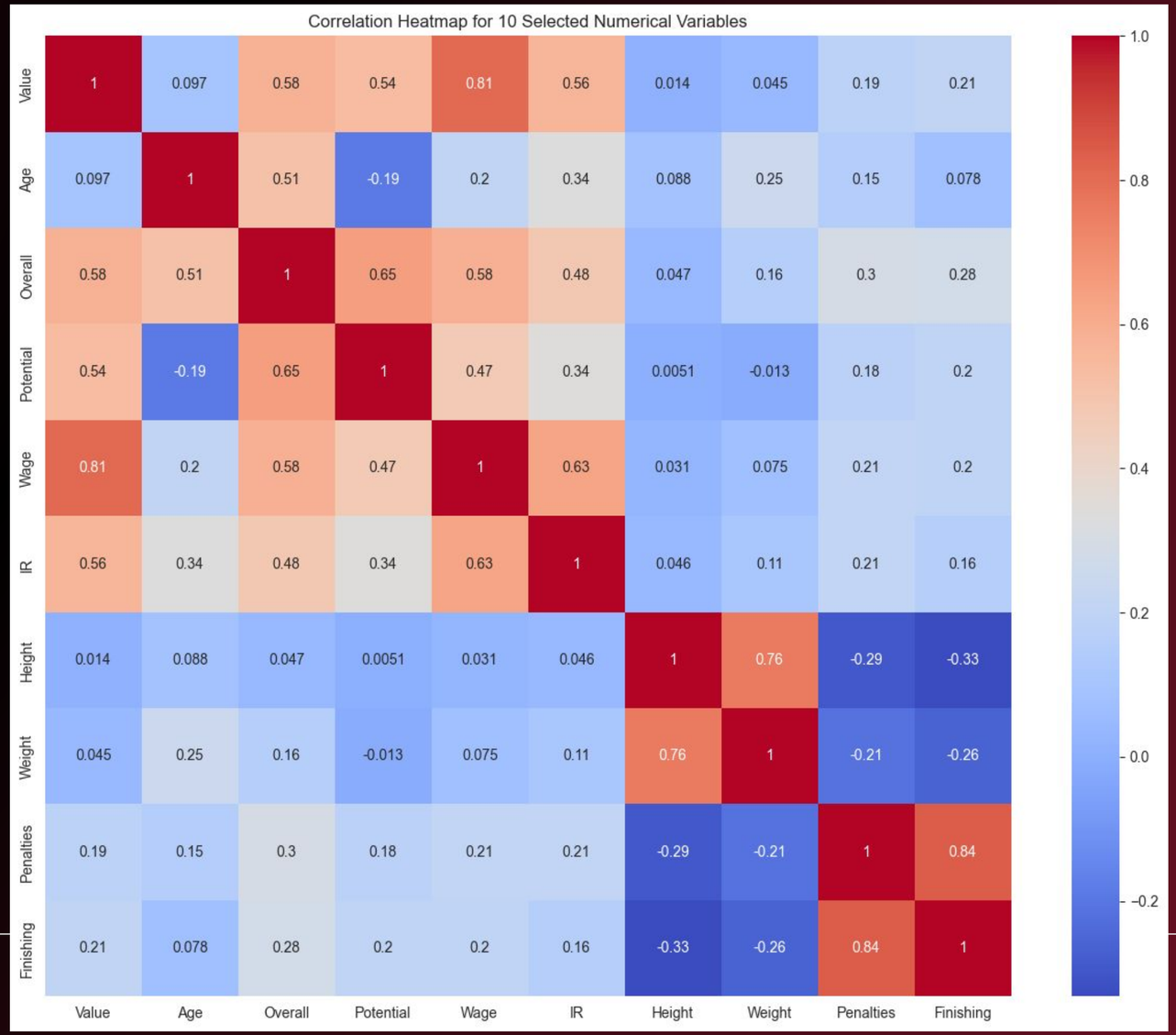
Conclusion: No significant association between ‘Preferred Foot’ and ‘Position’ is found. Therefore, the earlier observation is likely due to chance.

Bivariate: Correlation heatmap for 10 numeric values

A counter-intuitive observation!

Wage Skewness: 7.2
Wage Kurtosis: 90.26

My initial belief was that ‘finishing’ (scoring a goal) is strongly correlated with ‘Value’ (estimated monetary worth of a footballer). However, the relatively weak coefficient of 0.21 disproves this belief :|



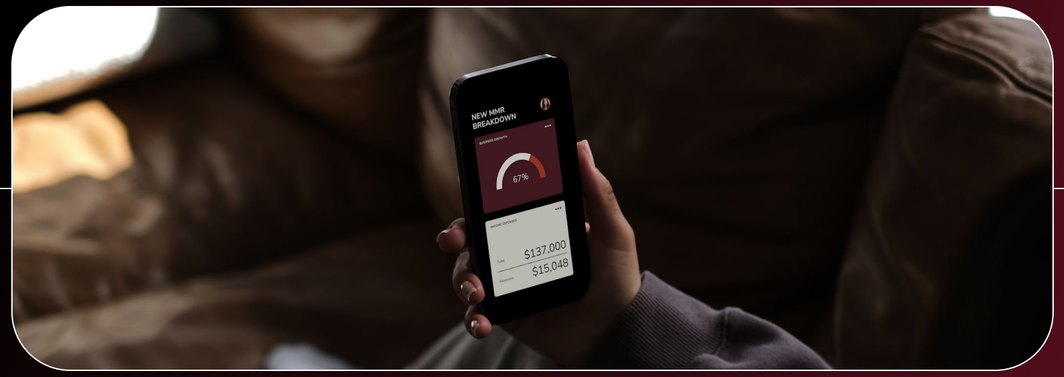
Tools for Data Analysis

Essential Software and Methods

Python data analysis and visualization ecosystem:

- pandas
- numpy
- matplotlib
- seaborn
- statsmodel
- scipy

Challenges in Data Analysis



Overcoming Common Issues

- An initial understanding and assessment of the data and observing the relations between significant parts of the data
- Converting ‘object’ type columns to raw ‘numeric’ ones using lambda functions
- Manual annotation adjustments on some plots when automated feature failed

Conclusion and Insights

- The high skewness indicates an uneven distribution of data; i.e., most data are distributed towards the lower extreme, while a small fraction of them are towards the higher extreme.
- The extreme kurtosis suggests that the wage distribution is not only highly concentrated around a central value, but also contains extreme outliers.
- While a strong correlation suggests that two variables are moving together, it does not establish a direct cause and effect relationship.
- Well-known statistical tests (Chi-square and Cramér's V) may result in a weak statistical relation between two categorical variables, contradicting the hypothesis and suggesting that the variables could be related by chance.
- Statistical tests, such as correlation coefficients and those stated above, challenge conventional beliefs. In the case of FIFA Dataset, factors like market demand, versatility, and overall skillset often play bigger roles in determining the worth of a player.

Thank you!

Questions ?!