Hoang Anh (Benjamin) Nguyen

**Data Preprocessing: Statistics and Exploratory Data Visualization**

1. **Basic statistics:**
   1. **Descriptive Statistics for Numerical Columns**

mean median mode std\_dev

attendance 26584.311404 29288.50 0.0 22827.967261

Home Team 11.450000 11.00 1.0 6.815784

Goals Home 1.502632 1.00 1.0 1.359450

Away Team 11.450000 11.00 1.0 6.815784

Away Goals 1.290351 1.00 1.0 1.233457

home\_possessions 50.816754 50.85 35.2 12.896181

away\_possessions 49.205965 49.20 64.8 12.899495

home\_shots 13.558772 13.00 15.0 5.615658

away\_shots 11.474561 11.00 10.0 5.048515

home\_on 4.715789 5.00 5.0 2.564688

away\_on 4.039474 4.00 3.0 2.359525

home\_off 5.069298 5.00 5.0 2.621838

away\_off 4.232456 4.00 4.0 2.448306

home\_blocked 3.776316 3.00 3.0 2.562984

away\_blocked 3.203509 3.00 1.0 2.266833

home\_pass 79.707368 81.10 82.7 7.442593

away\_pass 78.974298 80.40 79.1 7.329814

home\_chances 1.525439 1.00 1.0 1.389242

away\_chances 1.321930 1.00 0.0 1.328043

home\_corners 5.579825 5.00 5.0 3.055779

away\_corners 4.647368 4.00 3.0 2.767909

home\_offside 1.700877 1.00 1.0 1.433636

away\_offside 1.689474 1.00 1.0 1.496286

home\_tackles 58.001842 58.30 50.0 13.002387

away\_tackles 56.899561 57.10 50.0 12.765163

home\_duels 50.753772 50.00 50.0 10.757576

away\_duels 49.342632 50.00 50.0 10.777668

home\_saves 2.728070 2.00 2.0 1.856035

away\_saves 3.178070 3.00 2.0 1.978562

home\_fouls 10.620175 10.00 9.0 3.415676

away\_fouls 10.567544 10.00 10.0 3.560070

home\_yellow 1.587719 1.00 1.0 1.216014

away\_yellow 1.722807 2.00 1.0 1.279406

home\_red 0.051754 0.00 0.0 0.233210

away\_red 0.053509 0.00 0.0 0.236554

* 1. **Frequency Counts for Categorical Columns**

Column: date

date

28th May 2023 10

22nd May 2022 10

23/05/2021 10

29th October 2022 8

12th November 2022 8

8th April 2023 8

18th February 2023 8

11th September 2021 8

19th February 2022 8

20th November 2021 8

Name: count, dtype: int64

Column: clock

clock

3:00pm 291

8:00pm 161

2:00pm 136

5:30pm 101

4:30pm 90

12:30pm 60

7:45pm 45

8:15pm 45

6:00pm 43

7:30pm 41

Name: count, dtype: int64

Column: stadium

stadium

Emirates Stadium 57

Villa Park 57

Stamford Bridge 57

Selhurst Park 57

Goodison Park 57

Elland Road 57

Old Trafford 57

The King Power Stadium 57

St. Mary's Stadium 57

Amex Stadium 57

Name: count, dtype: int64

Column: class

class

h 494

a 390

d 256

Name: count, dtype: int64

Column: links

links

https://www.skysports.com/football/aston-villa-vs-brighton-and-hove-albion/stats/446398 2

https://www.skysports.com/football/fulham-vs-arsenal/stats/428839 1

https://www.skysports.com/football/arsenal-vs-wolverhampton-wanderers/465005 1

https://www.skysports.com/football/newcastle-united-vs-brighton-and-hove-albion/stats/428854 1

https://www.skysports.com/football/southampton-vs-tottenham-hotspur/stats/428855 1

https://www.skysports.com/football/arsenal-vs-west-ham-united/stats/428847 1

https://www.skysports.com/football/manchester-united-vs-crystal-palace/stats/428853 1

https://www.skysports.com/football/leeds-united-vs-fulham/stats/428851 1

https://www.skysports.com/football/everton-vs-west-bromwich-albion/stats/428850 1

https://www.skysports.com/football/manchester-city-vs-aston-villa/stats/428841 1

Name: count, dtype: int64

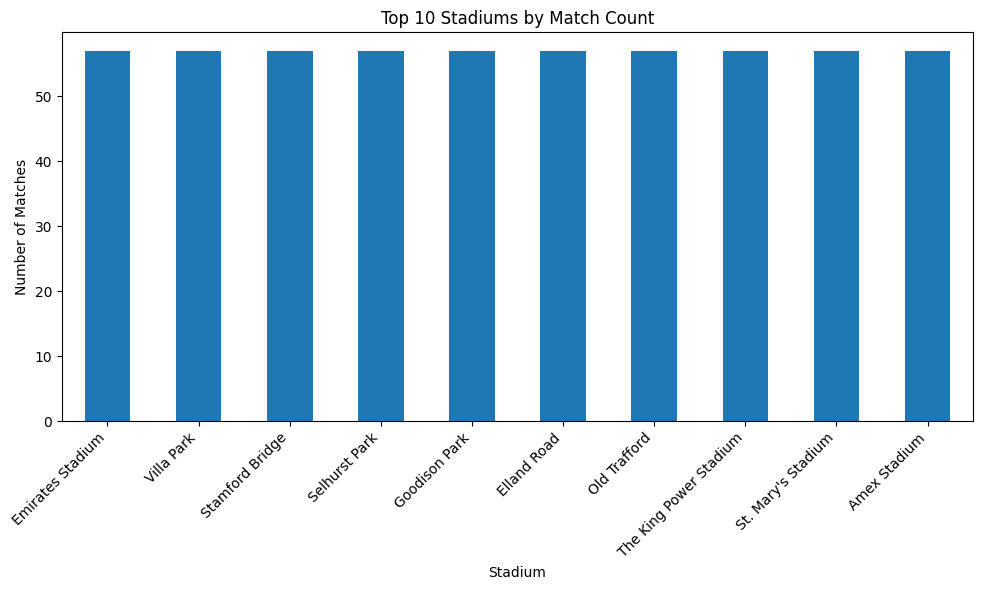
1. **Visualizations**

**A graph of goals and goals

AI-generated content may be incorrect.**

**Figure 1. Histogram for numerical data: goals per match**

Interpretation: the histogram of goals scored reveals that most matches are low scoring, with one or two goals being the most common outcomes. Home teams generally score more frequently than away teams, which is consistent with the well-known home advantage in soccer. High-scoring games with five or more goals are rare, indicating that such results are outliers rather than the norm.



**Figure 2. Bar chart for top 10 stadiums by match count**

Interpretation: the bar chart of the top 10 stadiums by match count confirms that matches are evenly distributed across the main home grounds of league teams. Each stadium has a similar number of hosted matches if the dataset represents a complete season, since every team plays an equal number of home games. Any differences in counts may be explained by relegated or promoted teams, or occasional use of neutral venues.

A diagram of a graph

AI-generated content may be incorrect.

**Figure 3. Scatter plot to explore relationships between two numerical variables: home possession verse shots**

Interpretation: the scatter plot of home possession versus home shots explores whether having more of the ball leads to more scoring opportunities. While there is generally a positive relationship, teams with higher possession often generate more shots, the scatter also shows exceptions. Some matches feature teams with high possession but relatively few shots.

1. **Screenshot of the code**

**A screenshot of a computer program

AI-generated content may be incorrect.**

****