Hoang Anh (Benjamin) Nguyen

Data Preprocessing: Statistics and Exploratory Data Visualization

1. Basic statistics:

1.1. 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
```

1.2. 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
```

2. Visualizations

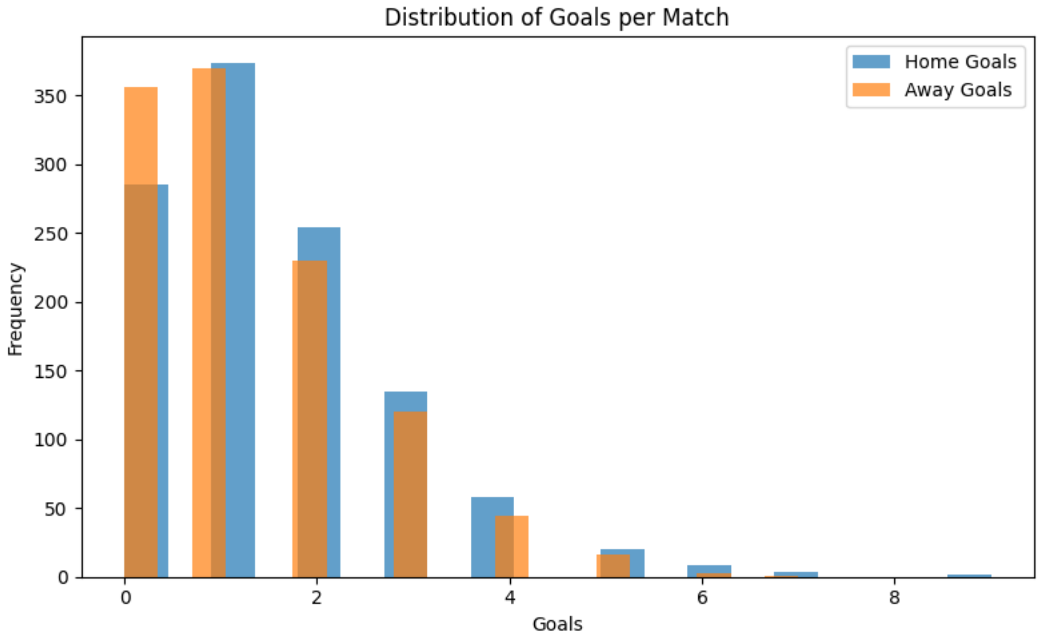


Figure 1. Histogram for numerical data: goals per match

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.

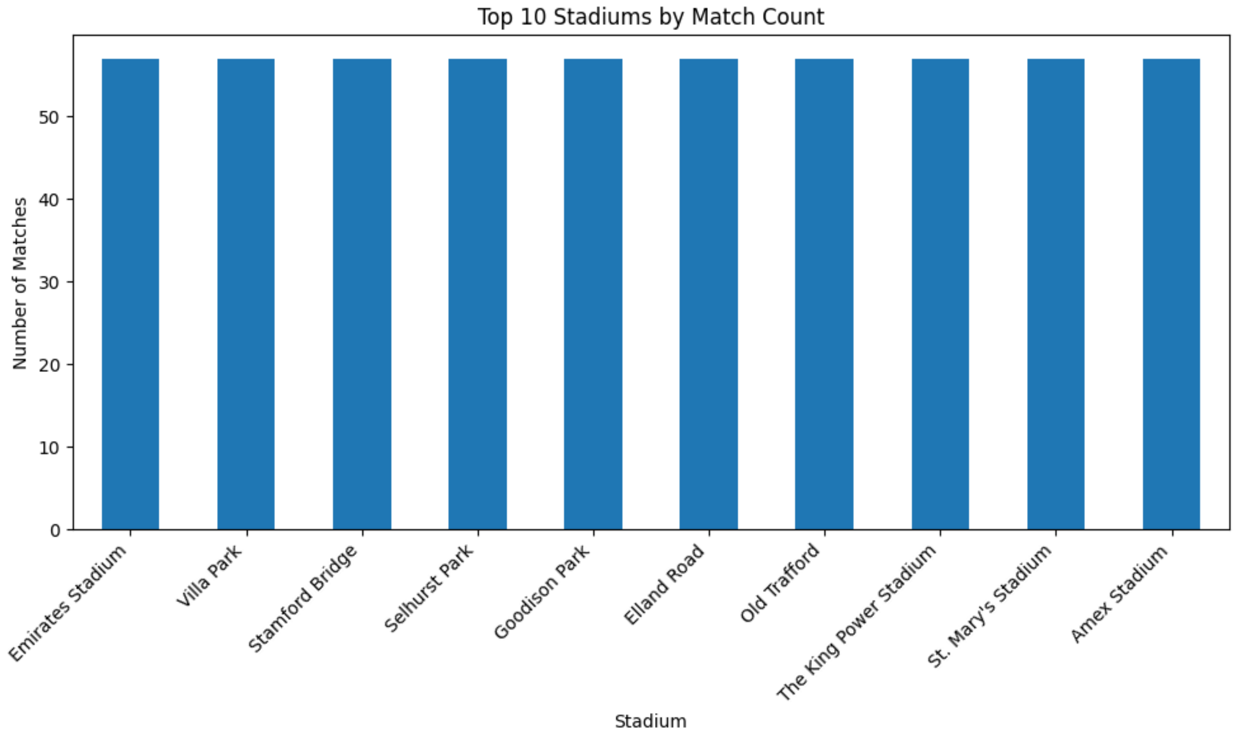


Figure 2. Bar chart for top 10 stadiums by match count

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.

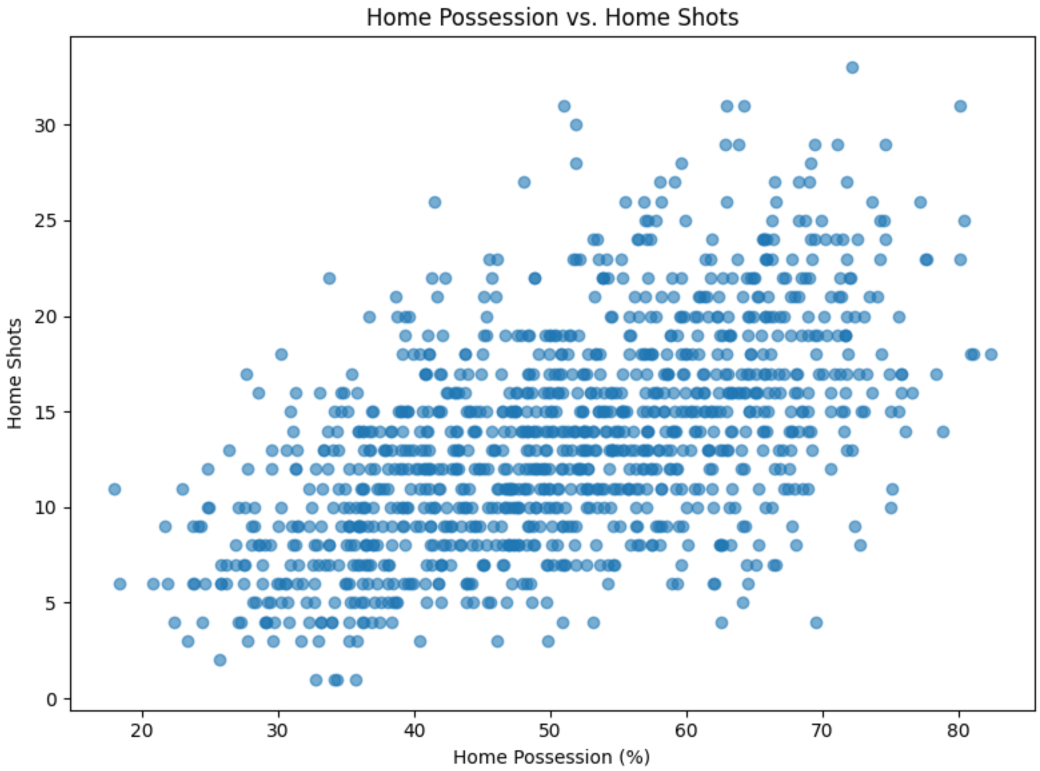
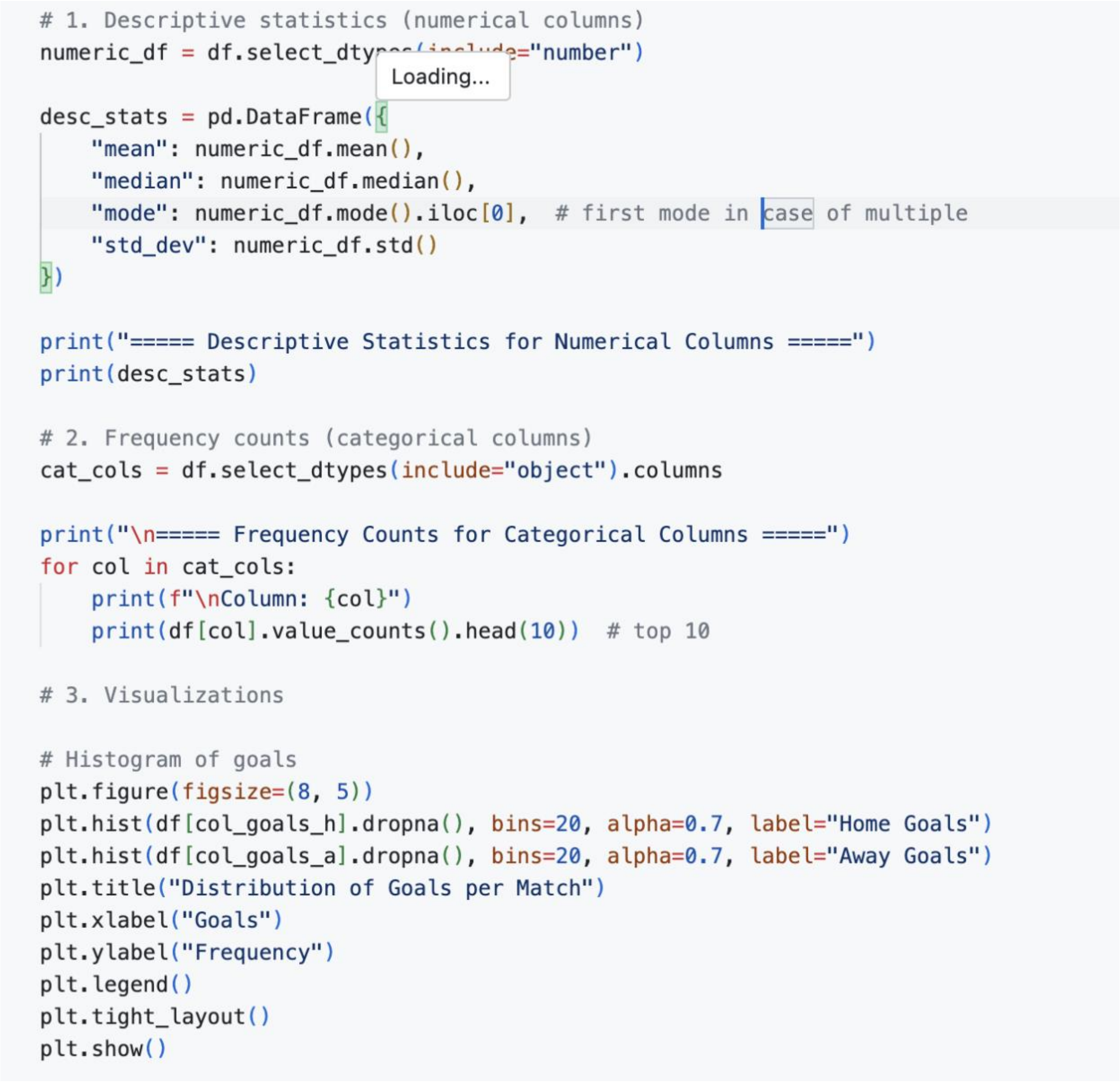


Figure 3. Scatter plot to explore relationships between two numerical variables: home possession verse shots

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.

3. Screenshot of the code



```
# 1. Descriptive statistics (numerical columns)
numeric_df = df.select_dtypes(include="number")

desc_stats = pd.DataFrame({
    "mean": numeric_df.mean(),
    "median": numeric_df.median(),
    "mode": numeric_df.mode().iloc[0],  # first mode in case of multiple
    "std_dev": numeric_df.std()
})

print("===== Descriptive Statistics for Numerical Columns =====")
print(desc_stats)

# 2. Frequency counts (categorical columns)
cat_cols = df.select_dtypes(include="object").columns

print("\n===== Frequency Counts for Categorical Columns =====")
for col in cat_cols:
    print(f"\nColumn: {col}")
    print(df[col].value_counts().head(10))  # top 10

# 3. Visualizations

# Histogram of goals
plt.figure(figsize=(8, 5))
plt.hist(df[col_goals_h].dropna(), bins=20, alpha=0.7, label="Home Goals")
plt.hist(df[col_goals_a].dropna(), bins=20, alpha=0.7, label="Away Goals")
plt.title("Distribution of Goals per Match")
plt.xlabel("Goals")
plt.ylabel("Frequency")
plt.legend()
plt.tight_layout()
plt.show()
```

```
# 3. Visualizations

# Histogram of goals
plt.figure(figsize=(8, 5))
plt.hist(df[col_goals_h].dropna(), bins=20, alpha=0.7, label="Home Goals")
plt.hist(df[col_goals_a].dropna(), bins=20, alpha=0.7, label="Away Goals")
plt.title("Distribution of Goals per Match")
plt.xlabel("Goals")
plt.ylabel("Frequency")
plt.legend()
plt.tight_layout()
plt.show()


# Bar chart of top stadiums
if col_stadium in df.columns:
    plt.figure(figsize=(10, 6))
    df[col_stadium].value_counts().head(10).plot(kind="bar")
    plt.title("Top 10 Stadiums by Match Count")
    plt.xlabel("Stadium")
    plt.ylabel("Number of Matches")
    plt.xticks(rotation=45, ha="right")
    plt.tight_layout()
    plt.show()

# Scatter plot: home possession vs. home shots
if col_poss_h in df.columns and col_shots_h in df.columns:
    plt.figure(figsize=(8, 6))
    plt.scatter(df[col_poss_h], df[col_shots_h], alpha=0.6)
    plt.title("Home Possession vs. Home Shots")
    plt.xlabel("Home Possession (%)")
    plt.ylabel("Home Shots")
    plt.tight_layout()
    plt.show()
```