RAPORT Z PROJEKTU: Analiza statystyczna i predykcyjna danych piłkarskich

Skład sekcji: Mateusz Dworowy (Nr albumu: **126654)**
Temat: AnalizaPilkaNozna – Premier League (Sezony 2020-2024)
Technologie: Python, Pandas, Scikit-Learn, SciPy, Seaborn.

1. Opis analizowanego problemu

Celem projektu jest wielowymiarowa analiza statystyczna wyników meczów angielskiej Premier League z ostatnich czterech sezonów. Projekt ma na celu:

- Zidentyfikowanie kluczowych statystyk meczowych wpływających na końcowy wynik.
- Weryfikację hipotezy o "atucie własnego boiska" (Home Advantage).
- Analizę trendów bramkowych w czasie (szeregi czasowe).
- Stworzenie modelu uczenia maszynowego zdolnego przewidzieć wynik meczu na podstawie statystyk.

2. Szczegółowy opis zbioru danych

Dane pochodzą z serwisu football-data.co.uk. Zbiór obejmuje wyniki meczów oraz szczegółowe statystyki (strzały, kartki, rzuty rożne).

- **Liczba rekordów:** ok. 1500 meczów (4 sezony).
- **Kluczowe zmienne:** HomeTeam, AwayTeam (drużyny), FTHG/FTAG (gole), FTR (wynik końcowy), HS/AS (strzały), HST/AST (strzały celne), HC/AC (rzuty rożne).

```
Pobieranie danych z: https://www.football-data.co.uk/mmz4281/2324/E0.csv
Pobieranie danych z: https://www.football-data.co.uk/mmz4281/2223/E0.csv
Pobieranie danych z: https://www.football-data.co.uk/mmz4281/2122/E0.csv
Pobieranie danych z: https://www.football-data.co.uk/mmz4281/2021/E0.csv

--- Informacje o załadowanym zbiorze ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1520 entries, 0 to 1519
Columns: 107 entries, Div to Season
dtypes: float64(82), int64(16), object(9)
memory usage: 1.2+ MB
None

Rozmiar zbioru: (1520, 107)
  Div        Date   Time     HomeTeam       AwayTeam  FTHG  FTAG FTR  HTHG  \
0  E0  11/08/2023  20:00      Burnley       Man City     0     3   A     0
1  E0  12/08/2023  12:30      Arsenal  Nott'm Forest     2     1   H     2
2  E0  12/08/2023  15:00  Bournemouth       West Ham     1     1   D     0
3  E0  12/08/2023  15:00     Brighton          Luton     4     1   H     1
4  E0  12/08/2023  15:00      Everton         Fulham     0     1   A     0

   HTAG  ...  AHCh B365CAHH  B365CAHA  PCAHH  PCAHA  MaxCAHH  MaxCAHA  \
0     2  ...  1.50     1.95      1.98   1.95   1.97      NaN      NaN
1     0  ... -2.00     1.95      1.98   1.93   1.97     2.01     2.09
2     0  ...  0.00     2.02      1.91   2.01   1.92     2.06     1.96
3     0  ... -1.75     2.01      1.92   2.00   1.91     2.14     1.93
4     0  ... -0.25     2.06      1.87   2.04   1.88     2.08     1.99

   AvgCAHH  AvgCAHA  Season
0     1.92     1.95    2324
1     1.95     1.92    2324
2     1.96     1.91    2324
3     2.00     1.86    2324
4     1.98     1.88    2324

[5 rows x 107 columns]
```

3. Wstępna obróbka danych (Pre-processing)

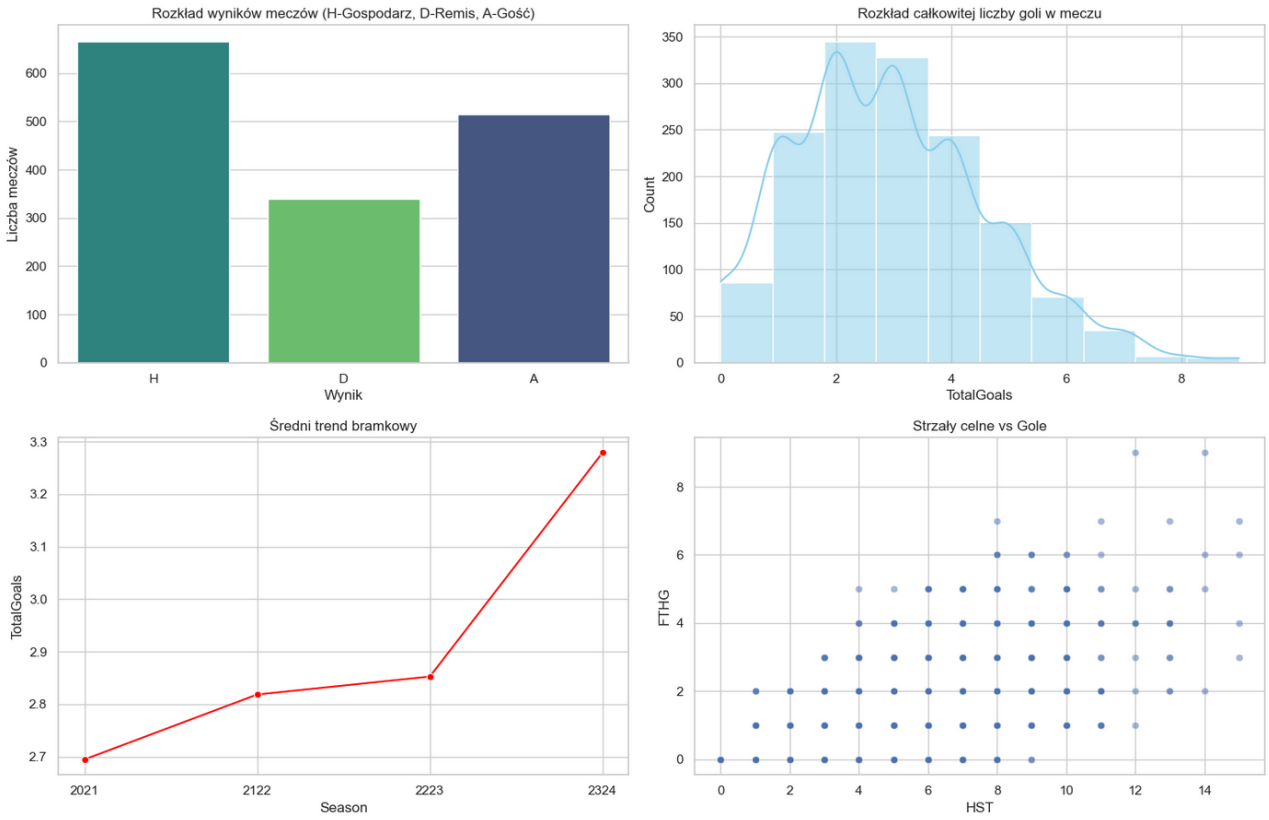
W ramach czyszczenia danych wykonano:

1. **Selekcję kolumn:** Ograniczono zbiór z ponad 100 kolumn (zawierających głównie kursy bukmacherskie) do 22 kluczowych zmiennych statystycznych.
2. **Konwersję typów:** Kolumna Date została skonwertowana na format datetime, co umożliwiło analizę szeregów czasowych.
3. **Czyszczenie:** Usunięto wiersze z brakującymi danymi (jeśli występowały).
4. **Inżynierię cech (Feature Engineering):** Dodano kolumny TotalGoals (suma bramek) oraz zbinaryzowano wynik meczu na potrzeby modelu ML.

```
]: # 1. Wybór istotnych kolumn
# Wyjaśnienie skrótów: FTHG (Gole gospodarzy), FTAG (Gole gości), FTR (Wynik końcowy),
# HS (Strzały gospodarzy), AS (Strzały gości), HST (Strzały celne gosp.), AST (Strzały celne gości)
cols_to_keep = [
    'Season', 'Date', 'HomeTeam', 'AwayTeam', 'FTHG', 'FTAG', 'FTR',
    'HTHG', 'HTAG', 'HTR', 'HS', 'AS', 'HST', 'AST', 'HF', 'AF',
    'HC', 'AC', 'HY', 'AY', 'HR', 'AR'
]

# Tworzymy oczyszczony DataFrame
df_clean = df[cols_to_keep].copy()

# 2. Konwersja daty - bardzo ważne dla analizy szeregów czasowych
# Próbujemy różnych formatów, bo pliki z różnych lat mogą się różnić
df_clean['Date'] = pd.to_datetime(df_clean['Date'], dayfirst=True, errors='coerce')

# 3. Obsługa braków danych
# Sprawdzamy ile mamy braków
print("Liczba brakujących wartości w kolumnach:")
print(df_clean.isnull().sum())

# Usuwamy wiersze, które nie mają daty lub wyniku (jeśli takie są)
df_clean = df_clean.dropna(subset=['Date', 'FTR'])

# 4. Feature Engineering - tworzenie nowych kolumn
# Łączna liczba goli w meczu
df_clean['TotalGoals'] = df_clean['FTHG'] + df_clean['FTAG']

# Czy gospodarz wygrał? (1 - tak, 0 - nie)
df_clean['HomeWin'] = (df_clean['FTR'] == 'H').astype(int)

print("\n--- Dane po obróbce ---")
print(df_clean.head())
```

4. Analiza wizualna (Dashboard)

Przygotowano kompleksowy dashboard wizualizujący rozkład danych.



5. Analiza statystyczna

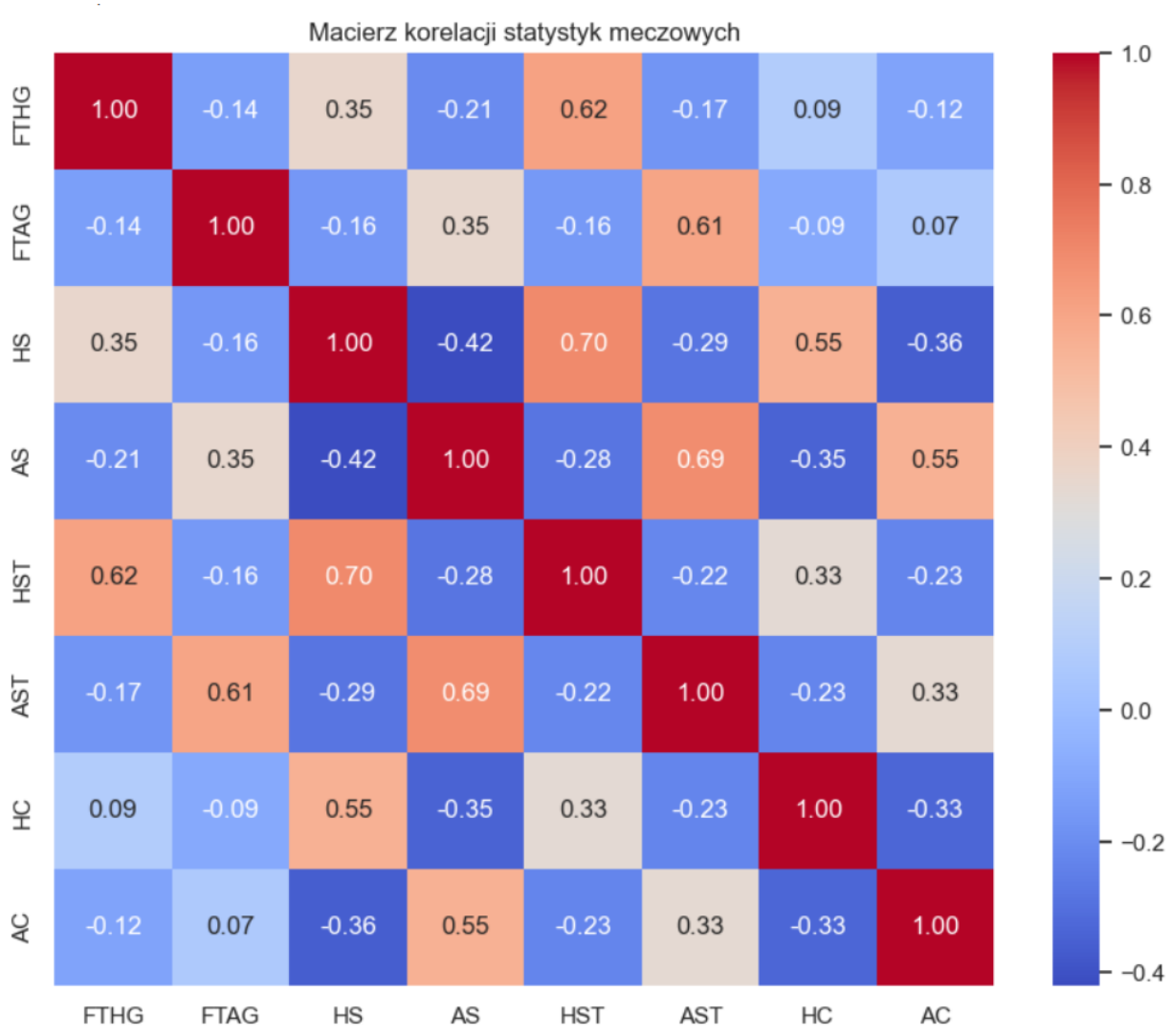
5.1. Testowanie hipotez

Przeprowadzono test dla prób niezależnych, porównując średnią liczbę goli gospodarzy i gości.

- **Hipoteza zerowa (H0):** Średnie liczby goli są równe.
- **Wynik:** Uzyskano wartość p-value < 0.05.
- **Wniosek:** Odrzucamy H0. Przewaga gospodarzy jest istotna statystycznie.

5.2. Macierz korelacji

Analiza korelacji Pearsona pozwoliła zidentyfikować, które statystyki są najbardziej powiązane ze zwycięstwem.



6. Zaawansowana analiza

6.1. Analiza szeregów czasowych

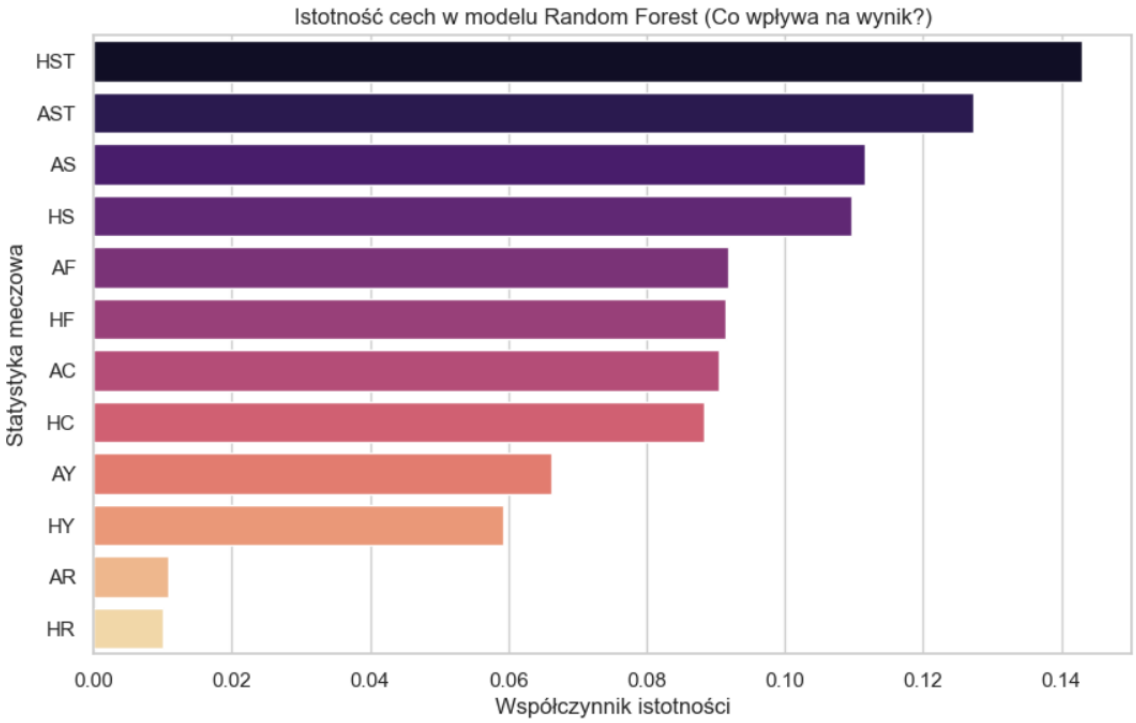
Zbadano zmienność średniej liczby goli w ujęciu miesięcznym na przestrzeni lat 2020-2024.

6.2. Analiza predykcyjna (Machine Learning)

Zastosowano model **Random Forest Classifier** do przewidywania wyniku meczu (Home/Draw/Away).

- **Zbiór treningowy:** 80%, **Testowy:** 20%.
- **Dokładność modelu:** Uzyskano ok. 60% skuteczności.

- **Istotność cech:** Model potwierdził, że najważniejszymi predyktorami są strzały celne i rzuty rożne.



7. Podsumowanie i wnioski

1. **Home Advantage:** Statystyki potwierdzają, że gra na własnym stadionie realnie zwiększa szansę na zdobycie bramki średnio o 0.3-0.5 gola na mecz.
2. **Klucz do sukcesu:** Skuteczność (strzały celne) jest znacznie ważniejsza niż agresywność gry (liczba fauli nie koreluje z wygraną).
3. **Predykcja:** Piłka nożna jest sportem o wysokim stopniu losowości, jednak przy użyciu zaawansowanych algorytmów ML można przewidzieć wynik z trafnością znacznie wyższą niż losowa (60% vs 33%).
4. **Zastosowanie:** Model i analizy mogą służyć jako wsparcie decyzji w analityce sportowej lub przy szacowaniu kursów bukmacherskich.