Projekt zaliczeniowy

Zaawansowane programowanie w języku Python
Rok akademicki 2024/2025

Autorzy:
Mateusz Strojek

[Analiza najlepszych klubów piłkarskich – projekt funkcyjny]

Projekt skupia się na wstępnej analizie danych dotyczących klubów piłkarskich oraz na analizie skupień. Dane są scrapowane ze strony https://www.whoscored.com/Statistics. Pobierane są poszczególne zmienne: drużyna, liga, liczba goli, strzały na mecz, żółte kartki, średnie posiadanie piłki na mecz, średni odsetek dokładnych podań, liczba wygranych powietrznych pojedynków na mecz oraz średnia ocena za mecz. Strona zawiera 4 zakładki; na pierwszej pokazane są podstawowe informacje o ramce danych, na drugiej pokazany jest wykres pudełkowy, wykres punktowy oraz wykres typu 'countplot' wraz z korelogramem. Na trzeciej zakładce pokazana jest analiza k-średnich zawierająca wartości indeksu sylwetkowego w zależności od liczby skupień, wykres punktowy w zależności od poszczególnych zmiennych oraz statystyki opisowe każdej grupy. Na ostatniej zakładce zostało przeprowadzone porządkowanie liniowe metodą Hellwiga.

Podział ról

Mateusz Strojek: Całość.

Opis funkcjonalności

Projekt składa się z pięciu plików:
Strona.py – plik zawierający funkcje czyszczące ramkę danych, układ strony, takie jak:

1. transform_variable_to_categorical(column, data) - Zmienia typ danej kolumny na kategorie.
2. change_premier_league_to_pl(column, data) - Zmienia nazwę "Premier League" na "PL" w kolumnie "Tournament".
3. plot_scatter_plot(df, x_axis, y_axis, hue=False) - Tworzy wykres punktowy między dwiema wybranymi zmiennymi.
4. show_dtypes(data) - Wyświetla typy danych w ramce danych.
5. show_stats(data) - Wyświetla podstawowe statystyki opisowe dla danych.

6. silhouette_print(data) - Oblicza i wyświetla wartość indeksu sylwetkowego dla różnych liczb klastrów.
7. plot_correlation_matrix(df) - Tworzy macierz korelacji dla danych numerycznych i wyświetla ją w formie korelogramu.
8. plot_boxplot(variable_box, df, hue=True) - Tworzy wykres pudełkowy dla wskazanej zmiennej, z możliwością dodania podziału według ligi.
9. plot_countplot(df) - Tworzy wykres słupkowy przedstawiający liczbę drużyn w różnych ligach.
10. plot_kmeans_clusters(data, num_clusters, x_axis, y_axis) - Przeprowadza analizę K-means na danych i rysuje wykres dla dwóch zmiennych.
11. generate_h1(text) - Generuje nagłówek H1 w aplikacji Streamlit.
12. generate_h2(text) - Generuje nagłówek H2 w aplikacji Streamlit.
13. generate_h3_centered(text) - Generuje nagłówek H3 w aplikacji Streamlit.
14. generate_head_table(df:pd.DataFrame, slides=0) - Wyświetla pierwsze wiersze tabeli z danymi.
15. clean_data(df_raw) - Czyści dane, zmieniając wartości w kolumnie "Tournament" i usuwając duplikaty.
16. get_data() - Ładuje dane, z możliwością skorzystania z istniejącego pliku CSV lub skrapowania nowych danych.
17. run_tab1(df) – wygenerowanie pierwszej zakładki „Overview”
18. run_tab2(df) – wygenerowanie drugiej zakładki „Plots”
19. run_tab3(df) – wygenerowanie trzeciej zakładki „K-means analysis”
20. run_tab4(df) – wygenerowanie czwartej zakładki „Linear ordering”
21. change_variables(variable_types, df) - zamiana wartości zmiennych w zależności czy zmienna jest stymulanta czy destymulantą
22. stimulant(column) – funkcja do stymulanty
23. destimulant(column) – funkcja do destymulanty
24. distances(dataframe, maxima, weights=None) - Funkcja licząca dystans pomiędzy wzorcem a aktualną wartością
25. di_count(distance_frame) – Funkcja licząca mi wartość d1 do wyniku indeksu
26. linear_ordering_Hellwig(df_original, df, weights, variable_types) - Funkcja do porządkowania liniowego
27. run_app() – generowanie architektury strony z wykresami, itd.
28. main() - Główna funkcja aplikacji Streamlit, która generuje interaktywne wykresy i statystyki.

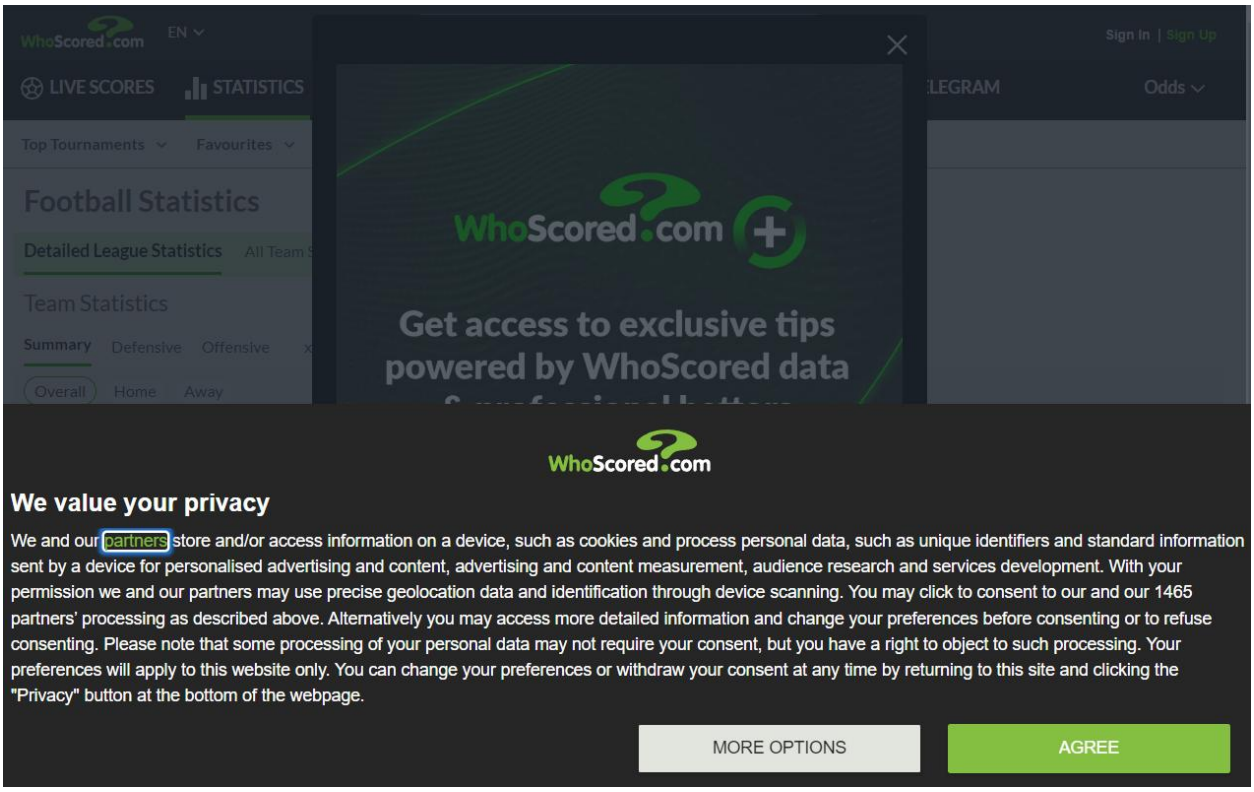
ScrapingData.py – plik zawierający funkcje odpowiedzialne za scrapowanie danych, takie jak:

1. setup_driver(url) - Inicjalizuje przeglądarkę Chrome i otwiera stronę internetową podaną w argumencie url.
2. extract_all_data(driver) - Wyciąga dane ze wszystkich kolumn tabeli oraz z kolumny 'Discipline'.
3. extract_cards(discipline) - Wyciąga informacje o żółtych i czerwonych kartkach z HTML w kolumnie 'Discipline'.
4. clean_data(dataframe) - Czyści dane, usuwając zbędne kolumny oraz numerację drużyn.
5. scrape_statistics(file_name) - Główna funkcja do scrapowania danych statystycznych z zewnętrznej strony i zapisania ich do pliku CSV.

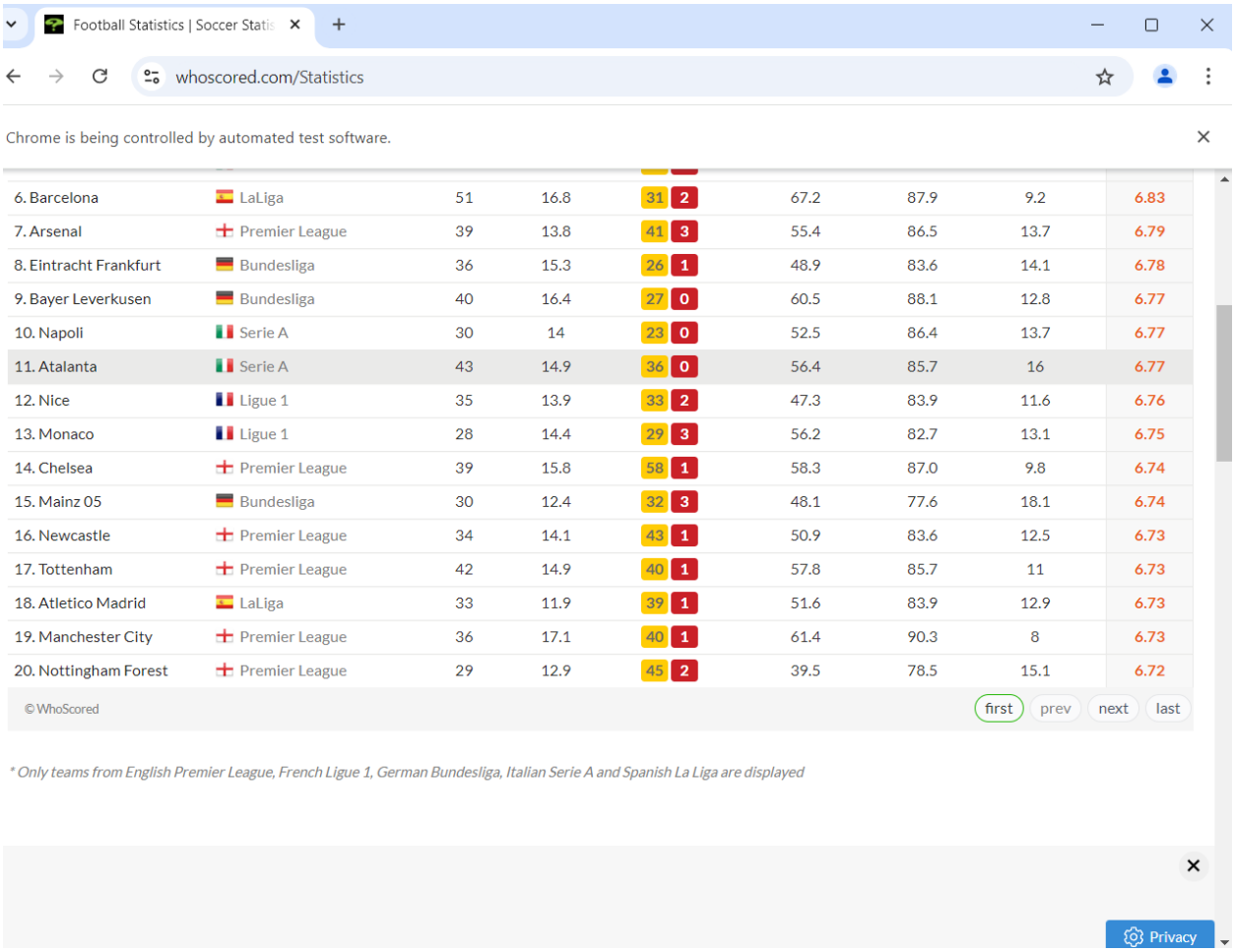
Style.py – plik zawierający style css.
Stare_kluby.csv – plik zawierające „stare dane” – plik jest potrzebny jedynie do demonstracji aktualizacji danych po zescrapowaniu.
Clubs.csv – plik z danymi zescrapowanymi ze strony.

Scrapowanie danych – ważne informacje

- Przed scrapowaniem należy szybko zaakceptować warunki strony i szybko usunąć niepotrzebne reklamy sprzed widoku, aby nie zasłaniały pierwszej tabeli z góry.



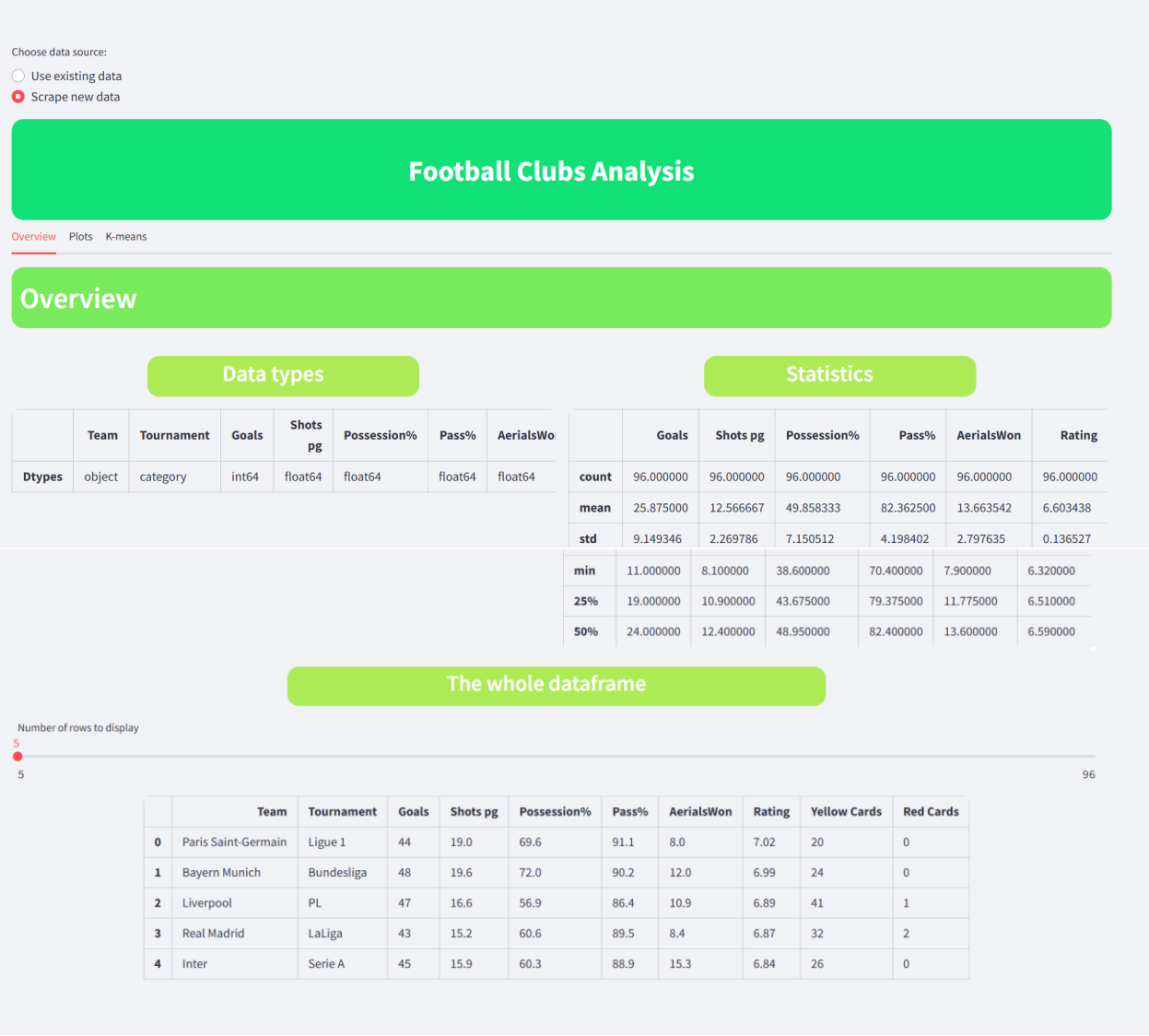
Następnie zescrolować stronę do poziomu, aż przycisk „next” będzie widoczny, najlepiej mniej więcej do tego momentu:



Gdy nie będzie chciało przejść do następnej strony tabelki, należy poruszyć zescrolować lekko w górę/dół.

Widok strony i jej funkcjonalność

Na pierwszej zakładce, jak wspomniałem wcześniej widnieją informacje o podstawowych informacjach ramki danych:



	Team	Tournament	Goals	Shots pg	Possession%	Pass%	AerialsWo
Dtypes	object	category	int64	float64	float64	float64	float64

	Goals	Shots pg	Possession%	Pass%	AerialsWon	Rating
count	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000
mean	25.875000	12.566667	49.858333	82.362500	13.663542	6.603438
std	9.149346	2.269786	7.150512	4.198402	2.797635	0.136527
min	11.000000	8.100000	38.600000	70.400000	7.900000	6.320000
25%	19.000000	10.900000	43.675000	79.375000	11.775000	6.510000
50%	24.000000	12.400000	48.950000	82.400000	13.600000	6.590000

	Team	Tournament	Goals	Shots pg	Possession%	Pass%	AerialsWon	Rating	Yellow Cards	Red Cards
0	Paris Saint-Germain	Ligue 1	44	19.0	69.6	91.1	8.0	7.02	20	0
1	Bayern Munich	Bundesliga	48	19.6	72.0	90.2	12.0	6.99	24	0
2	Liverpool	PL	47	16.6	56.9	86.4	10.9	6.89	41	1
3	Real Madrid	LaLiga	43	15.2	60.6	89.5	8.4	6.87	32	2
4	Inter	Serie A	45	15.9	60.3	88.9	15.3	6.84	26	0

Na tej zakładce użytkownik może rozszerzyć liczbę wierszów do pokazania. Na drugiej zakładce natomiast jest już o wiele więcej funkcjonalności. Wygląda ona tak:

Użytkownik może wybrać jeden spośród trzech wykresów (countplot, boxplot i scatterplot), wybrać zmienne dla poszczególnych osi. Jest również możliwość na np. narysowanie wykresów pudełkowych z podziałem na poszczególną ligę.
Trzecia zakładka przedstawia analizę skupień, użytkownik może wybrać liczbę skupień na podstawie np. wykresu sylwetkowego oraz zobaczyć, jak kształtują się grupy na wykresie punktowym. Dodatkowo, na samym spodzie jest tabelka ze średnimi dla poszczególnej grupy.

K-means analysis

Number of clusters

3

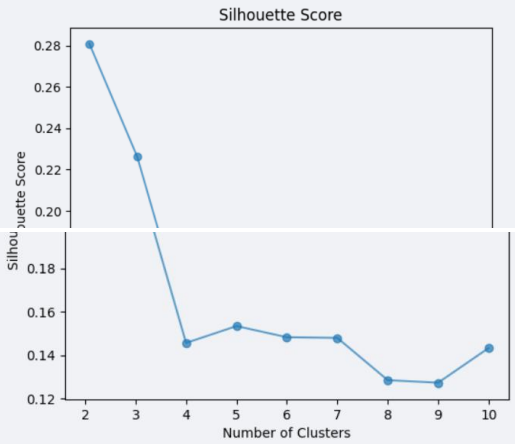
2 10

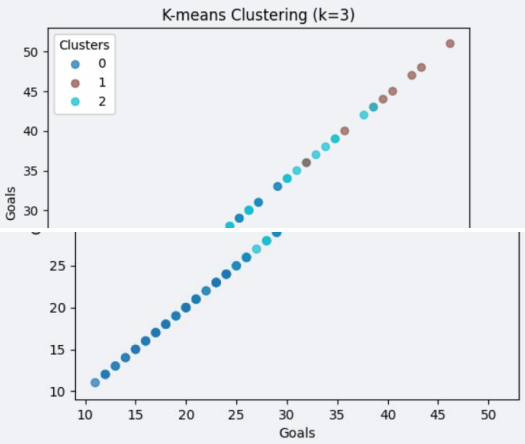
X-axis

Goals

Y-axis

Goals

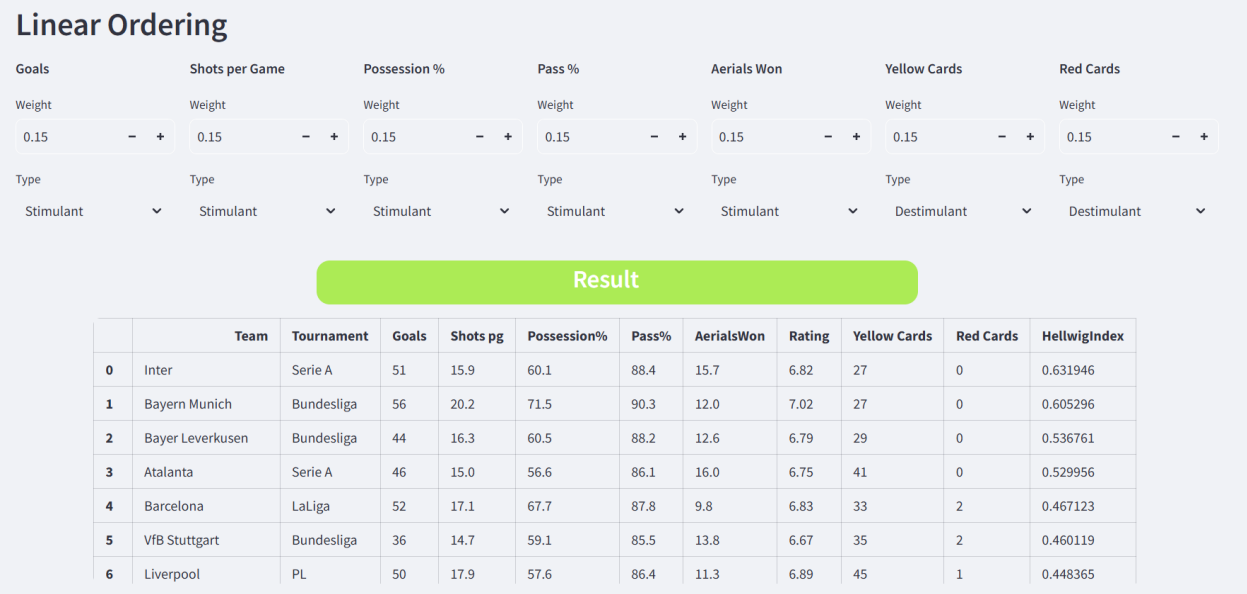




Means for a specific cluster

	Goals	Shots pg	Possession%	Pass%	AerialsWon	Rating	Yellow Cards	Red Cards
Cluster								
0	19.755102	11.148980	44.761224	79.297959	15.075510	6.513673	39.775510	1.897959
1	44.250000	17.075000	63.562500	89.050000	10.575000	6.867500	30.125000	0.750000
2	29.794872	13.423077	53.451282	84.841026	12.523077	6.662051	37.076923	1.717949

Na ostatniej zakładce występuje porządkowanie liniowe metodą Hellwiga. Można wybrać, czy zmienna jest stymulantą, czy tez nie oraz wybrać wagę danej zmiennej:



	Team	Tournament	Goals	Shots pg	Possession%	Pass%	AerialsWon	Rating	Yellow Cards	Red Cards	HellwigIndex
0	Inter	Serie A	51	15.9	60.1	88.4	15.7	6.82	27	0	0.631946
1	Bayern Munich	Bundesliga	56	20.2	71.5	90.3	12.0	7.02	27	0	0.605296
2	Bayer Leverkusen	Bundesliga	44	16.3	60.5	88.2	12.6	6.79	29	0	0.536761
3	Atalanta	Serie A	46	15.0	56.6	86.1	16.0	6.75	41	0	0.529956
4	Barcelona	LaLiga	52	17.1	67.7	87.8	9.8	6.83	33	2	0.467123
5	VfB Stuttgart	Bundesliga	36	14.7	59.1	85.5	13.8	6.67	35	2	0.460119
6	Liverpool	PL	50	17.9	57.6	86.4	11.3	6.89	45	1	0.448365

Możliwa rozbudowa projektu:

Informatyk może poszerzyć statystyczną analizę danych o np. metodę k-medoid, przeprowadzić analizę wrażliwości obydwu metod, przeprowadzić analizę głównych składowych (PCA), jak i analizę czynnikową (FA). Ponadto, można sprowadzić informację o piłkarzach i zrobić analizę w tym obszarze.