Wielowymiarowa analiza danych na przykładzie statystyk klubów piłkarskich

Projekt edukacyjny
Uniwersytet Ekonomiczny w Katowicach
Informatyka, st. 2, sem. 1
Analiza i modelowanie danych
Tomasz Chmiel, Artur Mendela

Streszczenie:

Projekt zawiera analizy statystyczne dotyczące klubów piłkarskich. Wykorzystana została baza danych FIFA, skupiająca się na statystykach z 2022 roku. Do każdej analizy wyodrębniono różne części bazy danych, tak, aby można było przeprowadzić czytelne badania, spełniające oczekiwania twórców projektu. Przeprowadzono łącznie pięć analiz - porządkowanie liniowe, analizę skupień, analizę korespondencji, dwuczynnikową analizę wariancji oraz analizę regresji. Każda z analiz miała za zadanie wysnuć odpowiednie wnioski, które zostały opisane w poszczególnych etapach dokumentacji. Pod uwagę wzięto głównie kluby z najwyższej ligi angielskiej Premier League, jednak w niektórych sytuacjach, na potrzebę projektu, dołączano do nich kluby z innych lig.

Cel projektu:

Celem projektu jest zbadanie zależności pomiędzy statystykami klubów piłkarskich w 2022r. Głównym założeniem jest zdobycie wiedzy o tych liczbach, które odgrywają największą rolę dla klubów. Każda analiza ma swój własny cel, który opisany jest na początku dokumentacji konkretnej analizy.

Postawiono pięć hipotez:

1. „Klub z najlepszymi ogólnymi statystykami w angielskiej Premier League, powinien charakteryzować się najbardziej korzystnymi predyspozycjami do skutecznej gry ofensywnej w lidze.”

Wykorzystana metoda:
Porządkowanie liniowe

2. „W wyniku analizy skupień wszystkich zespołów angielskie Premier League, można oczekiwać wyraźnego podziału pomiędzy czołowymi drużynami, które bezpośrednio walczą o mistrzostwo kraju a pozostałymi, co będzie odzwierciedleniem różnic w ich stylu gry, umiejętnościach i osiągnięciach.”

Wykorzystana metoda:
Analiza skupień

3. „Istnieje bliska relacja między ilością zdobytych tytułów mistrzowskich a popularnością klubów w angielskiej Premier League”

Wykorzystana metoda:
Analiza korespondencji

4. “Wiek i szybkość piłkarzy ma wpływ na potencjał ataku danej drużyny.”

Wykorzystana metoda:
Analiza wariancji dwuczynnikowa

5. “Potencjał ataku danej drużyny ma wpływ na jej końcową lokatę w tabeli.”

Wykorzystana metoda:
Analiza regresji

Porządkowanie liniowe

Hipoteza: „Klub z najlepszymi ogólnymi statystykami w angielskiej Premier League, powinien charakteryzować się najbardziej korzystnymi predyspozycjami do skutecznej gry ofensywnej w lidze.”

Cel: Z wykorzystaniem porządkowania liniowego ustalić, czy w sezonie 2022 ogólne statystyki drużyn w angielskiej Premier League miały zasadniczy wpływ na ich potencjał ataku. Dzięki takiemu badaniu, można wysunąć wniosek, czy jakość ataku jest w pewnym stopniu uzależniona od innych statystyk.

Wykorzystano porządkowanie liniowe, które wydaje się odpowiednie ze względu na swoje kryteria - w tym określanie wag, stymulantów, destymulantów i nominantów.

Na początku utworzono ranking klubów wykorzystując metodę wzorcową porządkowania liniowego w której każda cecha będzie mieć identyczną wagę. Następnie stworzono ranking nazwany ‘Superofensywa’. Owy ranking wszystkim cechom statystycznym, które sprzyjają grze ofensywnej ustawi większe wagi a cechę agresja zamieni z destymulanty w stymulantę, gdyż agresja jest pożądana gdy zespołowi bardziej zależy na zdobywaniu bramek niż nie traceniu goli oraz swoich zawodników. Wykorzystując wspomnianą metodę porządkowania liniowego najpierw należało określić charakter zmiennych przez wyznaczenie:

- Stymulant - gdy wyższe wartości danej cechy statystycznej są pożądane. Oznacza to, że im wyższa wartość, tym bardziej pozytywny wpływ ma na analizę.

- Destymulant - gdy niższe wartości danej zmiennej statystycznej są preferowane. To oznacza, że im niższa wartość, tym bardziej korzystne są wyniki analizy.

- Nominant - dana cecha statystyczna jest neutralna, czyli pożądana jest wartość z określonego zakresu, ale brak skrajnych preferencji co do jej konkretnego poziomu. W przypadku nominantów wartość z danego zakresu nie wpływa bezpośrednio na wynik analizy, ale jest istotna dla ogólnej interpretacji danych.

W analizie skupiono się na ośmiu najlepszych drużynach w angielskiej Premier League. Są to kolejno: Arsenal, Chelsea, Leicester City, Liverpool, Manchester City, Manchester United, Tottenham Hotspur oraz West Ham United. Wybór właśnie tych topowych zespołów pozwolił na szczegółową analizę ich statystyk, umiejętności i stylu gry.

1. **Dane i określenie preferencji**

Wedle określonych preferencji, odnośnie danych statystycznych każdej drużyny zostały przyjęte stymulanty i destymulanty. Żadna z cech nie musiała zostać przekształcona z zmiennej jakościowej na zmienną ilościową. Wartości wszystkich cech z wyjątkiem wieku, są uśrednionymi umiejętnościami każdej drużyny w danym aspekcie, opracowanymi na podstawie statystyk meczowych z ubiegłych sezonów oraz potencjału kadrowego. Cechy zostały przedstawione poniżej.

Wiek:

Interpretacja: Średni wiek zawodników każdej drużyny. Preferencje mogą skierować się w kierunku młodszych zawodników, gdyż często charakteryzują się większym potencjałem rozwojowym i szybkością.

Drybling:

Interpretacja: Zmienna ilościowa mierząca umiejętność zręcznego manewrowania piłką w trakcie gry. Wartości tej cechy są pożądane, gdyż wyższy poziom dryblingu może oznaczać zdolność do skutecznego omijania przeciwników.

Podania:

Interpretacja: Miara precyzji i skuteczności podań zawodników. Wartości tej cechy są preferowane, ponieważ precyzyjne podania są kluczowe dla efektywnej gry zespołowej.

Wykończenie:

Interpretacja: Ocena umiejętności zdobywania bramek przez zawodników. Wyższe wartości są pożądane, sugerując efektywność w zakresie strzelania goli.

Rzuty wolne:

Interpretacja: Mierzy umiejętność skutecznego wykonywania rzutów wolnych. Wyższe wartości są pożądane, ponieważ mogą przekładać się na zdolność zdobywania goli z tego typu sytuacji.

Główkowanie:

Interpretacja: Ocena umiejętności zdobywania goli głową oraz gry defensywnej z wykorzystaniem głowy. Wyższe wartości są preferowane, szczególnie dla zawodników grających na pozycjach ofensywnych.

Kontrola piłki:

Interpretacja: Zmienna ilościowa mierząca zdolność do skutecznego panowania nad piłką. Wyższe wartości są pożądane, umożliwiając utrzymanie piłki w posiadaniu i efektywną grę.

Agresja:

Interpretacja: Mierzy skłonność zawodników do agresywnego podejścia w grze. Niższe wartości są preferowane, aby unikać zbędnych fauli i kar.

Przyśpieszenie:

Interpretacja: Zmienna ilościowa oceniająca szybkość rozpędzania się zawodników. Wyższe wartości są preferowane, zwłaszcza dla graczy ofensywnych, aby mogli szybko reagować i przyspieszać w trakcie akcji.

W poniższej tabeli przedstawiono przyjęte stymulanty i destymulanty.

Rysunek 1: Preferencje cech statystycznych

Cechy statystyczne	Preferencje
Wiek	Destymulanta
Drybling	Stymulanta
Podania	Stymulanta
Wykończenie	Stymulanta
Rzuty wolne	Stymulanta
Główkowanie	Stymulanta
Kontrola piłki	Stymulanta
Agresja	Destymulanta
Przyśpieszenie	Stymulanta

Źródło: Opracowanie własne

2. Średnia i odchylenie standardowe po przekształceniu wartości na stymulanty

Rysunek 2: Średnia i odchylenie standardowe

	Wiek	Drybling	Podania	Wykończenie	Rzuty wolne	Główkowanie	Kontrola piłki	Agresja	Przyśpieszenie
Średnia	1.5237	83.95313	82.625	81.8	78.9166667	79.5	83.65	2.6875	83.5625
Odchylenie standardowe	1.2737	2.135944	2.245969	2.5	2.34372685	2.66974737	2.391652149	1.5971	2.343541657

Źródło: Obliczenia własne

a) Wiek:

Średnia wieku drużyn w lidze wynosi 1.52. Taka wartość powstała gdyż cecha będąca destymulantą została przekształcona na stymulantę. Odchylenie standardowe 1.27 sugeruje pewne zróżnicowanie wieku wśród czołowych drużyn, choć ogólnie jest to niewielka zmienność.

b) Drybling:

Średnia umiejętności dryblingu wynosi 83.95, co potwierdza ogólnie wysoki poziom tej zdolności wśród najlepszych drużyn. Niewielkie odchylenie standardowe (2.14) wskazuje na równomierną jakość tej umiejętności w zespołach.

c) Podania:

Średnia efektywności w podaniach wynosi 82.63, co wskazuje na solidną umiejętność gry zespołowej w najlepszych drużynach. Odchylenie standardowe (2.25) sugeruje pewne zróżnicowanie jakości podań w ligowej rywalizacji.

d) Wykończenie:

Średnia umiejętności wykończenia na poziomie 81.8 sugeruje solidne zdolności strzeleckie wśród najlepszych drużyn. Odchylenie standardowe (2.5) wskazuje na pewne zróżnicowanie skuteczności wśród zawodników w czołowych zespołach ligi.

e) Rzuty wolne:

Średnia skuteczność w rzutach wolnych wynosi 78.92, co wskazuje na pewną stabilność w tej dziedzinie wśród topowych drużyn. Odchylenie standardowe (2.34) sugeruje pewne zróżnicowanie umiejętności w tej specyficznej sytuacji gry.

f) Główkowanie:

Średnia umiejętność główkowania na poziomie 79.5 wskazuje na równomierną zdolność do zdobywania goli głową wśród najlepszych drużyn. Wyższe odchylenie standardowe (2.67) może sugerować większe zróżnicowanie tej umiejętności w czołowych zespołach ligi.

g) Kontrola piłki:

Średnia kontrola piłki wynosząca 83.65 wskazuje na wysoki poziom tej umiejętności w najlepszych drużynach ligi. Odchylenie standardowe (2.39) sugeruje pewne zróżnicowanie w jakości kontrolowania piłki w czołowych zespołach.

h) Agresja:

Średnia skłonność do agresywnej gry na poziomie 2.69 może wpływać na dynamikę zespołów w czołówce. Niższe odchylenie standardowe (1.60) wskazuje na pewną jednorodność w tym aspekcie w ligowej rywalizacji.

i) Przyśpieszenie:

Średnia szybkość rozpędzania się zawodników wynosi 83.56, co sugeruje ogólnie wysoką dynamikę w grze czołowych drużyn. Odchylenie standardowe (2.34) wskazuje na pewne zróżnicowanie w tej umiejętności wśród najlepszych drużyn.

3. Wzorzec i antywzorzec

Po wykonaniu standaryzacji zmiennych wyznaczono wzorzec Z+, czyli maksymalną wartość zmiennej po standaryzacji oraz antywzorzec Z-, czyli minimalną wartość zmiennej po standaryzacji.

Rysunek 3: Wzorzec Z+ i antywzorzec Z-

	Wiek	Drybling	Podania	Wykończenie	Rzuty wolne	Główkowanie	Kontrola piłki	Agresja	Przyśpieszenie
Z+	1.9442	1.60204	1.5027	1.2	2.026686684	1.257475856	1.69339007	1.3697	0.986754382
Z-	-1.196	-1.49963	-1.347	-1.84	-1.52890399	-1.68555274	-1.191644864	-1.683	-2.160192026

Źródło: Obliczenia własne

Natomiast w przypadku rankingu ogólnego wszystkie cechy mają te samą wagę równą 11,(1).

4. Ranking klubów na podstawie cech statystycznych z identycznymi wagami.

Rysunek 4: Ogólny ranking klubów

	Klub	di (SMR)
1	Manchester City	0.77630229
2	Liverpool	0.70342005
3	Manchester United	0.623918765
4	Chelsea	0.524295201
5	Tottenham Hotspur	0.470398558
6	Arsenal	0.42203192
7	Leicester City	0.307673563
8	West Ham United	0.21453554

Źródło: Obliczenia własne

Z przeprowadzonej analizy rankingu, opartej na metodzie wzorcowej porządkowania liniowego, wynika, że przy jednolitych wagach dla wszystkich kryteriów, najbardziej korzystnym wyborem okazuje się być klub Manchester City, zajmujący pierwsze miejsce z wynikiem di (SMR) wynoszącym około 0,776. Na drugim miejscu znajduje się Liverpool z rezultatem di (SMR) około 0,703, natomiast trzecie miejsce zajmuje Manchester United z wynikiem di (SMR) około 0,624.

5. Ranking “Superofensywa”

Aby wyłonić najlepiej grający ofensywnie zespół należy wszystkim cechą statystycznym, które sprzyjają grze ofensywnej ustawić większe wagi a cechę agresja z destymulanty zmienić w stymulantę, gdyż agresja sprzyja atakowaniu i jest pożądana gdy zespołowi bardziej zależy na zdobywaniu bramek niż nie traceniu goli oraz swoich zawodników.

Rysunek 5: Wagi

	Wiek	Drybling	Podania	Wykończenie	Rzuty wolne	Główkowanie	Kontrola	Agresja	Przyśpieszenie
Wagi	20%	5%	5%	5%	15%	5%	5%	15%	25%

Źródło: Obliczenia własne

Zmiana wag w kontekście analizy cech piłkarskich sugeruje, że teraz większy nacisk kładziony jest na aspekty ofensywne gry. Zmiany, które zostały wprowadzone względem analizy ogólnej, prezentują się następująco:

Waga Wieku (20%):

Zwiększenie wagi dla wieku sugeruje, że skład z młodszymi zawodnikami jest teraz bardziej ceniony. Młodszy skład może oznaczać większą dynamikę, szybkość oraz ambicję. To podejście sprzyja bardziej energicznej i ofensywnej grze.

Waga Rzutów Wolnych (15%):

Rzuty wolne stanowią okazję do zdobycia bramki. Zwiększenie wagi tej cechy sugeruje, że drużyny mogą bardziej eksponować zawodników o umiejętnościach w tej dziedzinie, co może przyczynić się do zwiększenia liczby bramek zdobywanych z tego stałego fragmentu gry.

Waga Agresji (15%):

Zwiększenie wagi agresji wskazuje na skoncentrowanie się na bardziej ofensywnym podejściu do gry. Agresywna gra może przyczynić się do szybszego odzyskiwania piłki w środku pola, co pozwala drużynie na częstsze przewagi liczbowe i szybkie ataki. Wzrost wagi na agresję sprzyja zatem intensywnej, pressującej grze ofensywnej.

Waga Przyśpieszenia (25%):

Przyśpieszenie jest kluczowym elementem szybkich ataków. Przy zwiększonej wadze tej cechy, drużyny preferują dynamiczne akcje, co może być szczególnie skuteczne w przełamywaniu obrony przeciwnika.

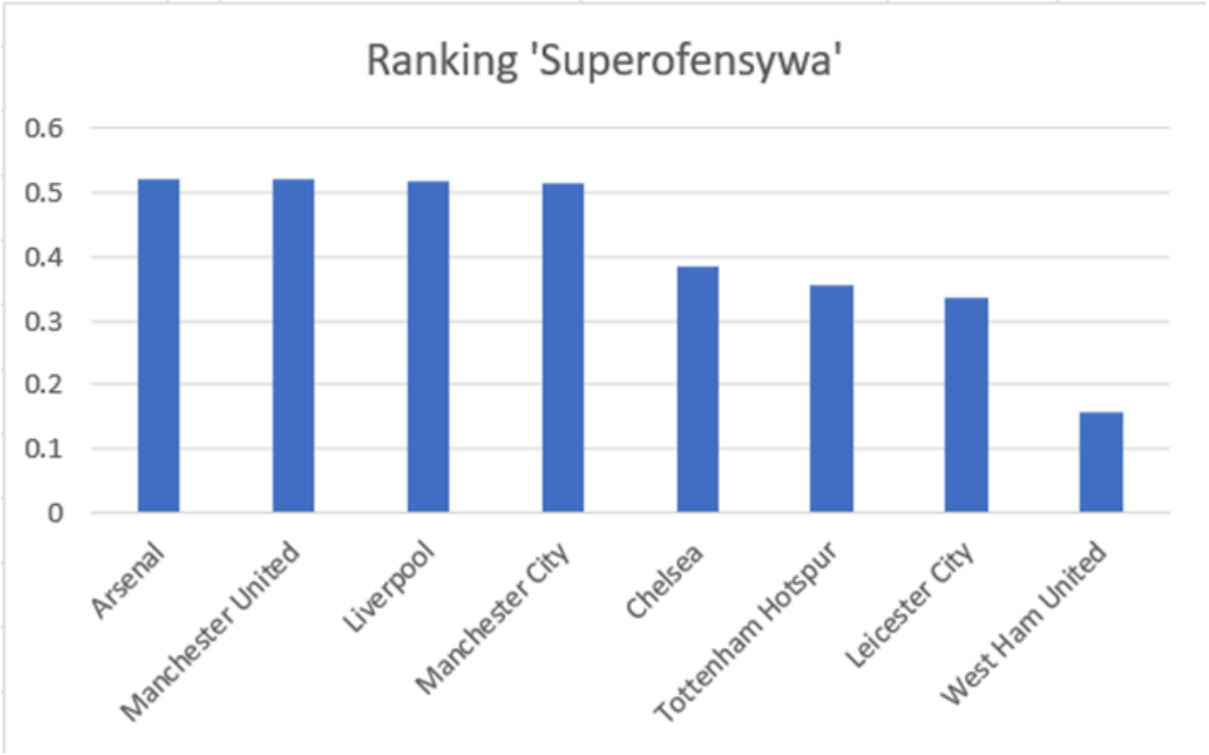
Rysunek 6: Wyniki analizy

	Klub	di (SMR)
1	Arsenal	0.522197756
2	Manchester United	0.520293054
3	Liverpool	0.516786044
4	Manchester City	0.513432412
5	Chelsea	0.383440821
6	Tottenham Hotspur	0.356153322
7	Leicester City	0.334934622
8	West Ham United	0.156079732

Źródło: Opracowanie własne

Z wyników analizy, opartej na metodzie wzorcowej porządkowania liniowego z zastosowaniem zróżnicowanych wag promujących grę ofensywną, wynika, że klub Arsenal zajmuje pierwsze miejsce z wynikiem di (SMR) wynoszącym 0,522. Następnie, na drugim miejscu plasuje się Manchester United z rezultatem di (SMR) równym 0,520, a trzecie miejsce zajmuje Liverpool z wynikiem di (SMR) na poziomie 0,517. Ranking kontynuują kolejno Manchester City (0,513), Chelsea (0,383), Tottenham Hotspur (0,356), Leicester City (0,335), oraz West Ham United (0,156). Analiza uwzględniała różne aspekty gry, ze szczególnym naciskiem na te sprzyjające grze ofensywnej. Kluby, które osiągnęły wyższe pozycje w rankingu, wykazały się lepszymi wynikami pod względem założonych kryteriów, co sugeruje, że ich styl gry sprzyja efektywnej grze ofensywnej.

Rysunek 7: Ranking “Superofensywa”



Źródło: Opracowanie własne

6. Podsumowanie

Analiza ogólna wskazała, że klub Manchester City zajmuje pierwsze miejsce w rankingu ogólnym, co sugeruje, że ma największe predyspozycje do gry ofensywnej. Jednak, po dostosowaniu wag cech w analizie ‘Superofensywa’, Arsenal zdobył pierwsze miejsce, a Manchester City dopiero czwarte, co oznacza, że w kontekście specjalnie dostosowanych kryteriów, to Arsenal wykazuje największe predyspozycje do skutecznej gry ofensywnej. Ostatecznie, hipoteza, zakładająca, że klub z najlepszymi ogólnymi statystykami w angielskiej Premier League, powinien charakteryzować się najbardziej korzystnymi predyspozycjami do skutecznej gry ofensywnej w lidze, nie została potwierdzona.

Analiza skupień

Hipoteza: „W wyniku analizy skupień wszystkich zespołów angielskiej Premier League, można oczekiwać wyraźnego podziału pomiędzy czołowymi drużynami, które bezpośrednio walczą o mistrzostwo kraju a pozostałymi, co będzie odzwierciedleniem różnic w ich stylu gry, umiejętnościach i osiągnięciach.”

Cel: W kontekście analizy skupień, celem jest zrozumienie struktury i podobieństw pomiędzy zespołami Premier League na podstawie różnych cech, takich jak wiek, drybling, kontrola piłki czy przyśpieszenie. Przyjęta hipoteza zakłada, że w wyniku analizy skupień, grupy zespołów wytworzą się w sposób, który jednoznacznie wyodrębni czołowe drużyny z największym potencjałem od reszty.

Wykorzystano analizę skupień, ponieważ najlepiej spełnia oczekiwania dotyczące podziału danych na dwie zasadnicze grupy oraz odzwierciedlenia różnić dla kilku statystyk.

1. **Dane**

Przeprowadzono kompleksowy proces agregacji i kompletowania danych statystycznych dla wszystkich klubów grających w angielskiej Premier League w sezonie 2022/2023. Do tego celu zastosowano język programowania Python, który ze względu na swoją wszechstronność i bogatą gamę narzędzi analizy danych, stanowi niezastąpione wsparcie we wszelkich zadaniach badawczych. Python umożliwił efektywną filtrację, analizę oraz prezentację danych, co pozwoliło uzyskać kompleksowy obraz statystyk dla każdego klubu, zapisany w czytelnej formie w pliku Excel.

Rysunek 8: Kompletowanie danych

```python
import pandas as pd
from tabulate import tabulate

file_path = 'C:\\Users\\tomek\\Downloads\\archive\\teams-stats.ds.csv'

df = pd.read_csv(file_path)

desired_clubs = ['AFC Bournemouth', 'Arsenal', 'Aston Villa', 'Brentford', 'Brighton & Hove Albion',
                 'Chelsea', 'Crystal Palace', 'Everton', 'Fulham', 'Leeds United', 'Leicester City',
                 'Liverpool', 'Manchester City', 'Manchester United', 'Newcastle United', 'Nottingham Forest',
                 'Southampton', 'Tottenham Hotspur', 'West Ham United', 'Wolverhampton Wanderers']

english_clubs_2022 = df[(df['Year'] == 2022) & (df['Club'].isin(desired_clubs))]

selected_columns = ['Club', 'Age', 'Dribbling', 'Crossing', 'Finishing', 'FKAccuracy', 'HeadingAccuracy',
                    'BallControl', 'Aggression', 'Acceleration']

excel_file_path = 'C:\\Users\\tomek\\Downloads\\archive\\PLclubs.xlsx'
english_clubs_2022[selected_columns].to_excel(excel_file_path, index=False)

print(tabulate(english_clubs_2022[selected_columns], headers='keys', tablefmt='pretty'))
print(f"\nDataFrame saved to Excel file: {excel_file_path}")
```

Źródło: Opracowanie własne

Tak prezentują się dane po zaimplementowaniu ich w programie Statistica przed przeprowadzeniem procesu standaryzacji.

Rysunek 9: Dane do analizy skupień

Kluby	x1	x2	x3	x4	x5	x6	x7	x8	x9
AFC Bournemouth	76	73.4	71.4	73.3333333	72.5714286	74.9	78.5	81.5	3.3
Arsenal	83.125	80.6	77.2	79.3333333	75	81.8	81.875	85.375	3.6
Aston Villa	80.75	80.2	78.8	82.6666667	78.5714286	80.7	82.625	79.5	3.6
Brentford	76.375	73.4	73.4	74	74.2857143	75.4	80.75	83.5	3.3
Brighton & Hove Albion	76.75	79	76.8	76.3333333	72.7142857	75.7	76.875	84.875	3.3
Chelsea	84.75	84.8	81.6	79.3333333	78	84.3	83.125	82.375	3.9
Crystal Palace	79.875	75	76.2	76.3333333	75.4285714	78.6	78.625	80.5	3.7
Everton	79.5	76.2	77	75.6666667	81	79.2	83	78.625	3.5
Fulham	77.25	75.8	75.8	81.6666667	75.8571429	76.6	82.375	83.75	3.3
Leeds United	78.5	74.2	75.8	69	74.4285714	77.3	77.5	84.5	3.3
Leicester City	80.75	80.4	81.4	77	75.8571429	81.4	82.25	82.125	3.5
Liverpool	86.375	85	83.6	79	80.8571429	86.6	84.25	85.5	4.2
Manchester City	87.375	86	84.8	80.3333333	81.1428571	87.7	84.125	83.625	4.1
Manchester United	83.875	82.6	82.6	83.6666667	81.2857143	84.6	86.25	85.875	4.3
Newcastle United	81.375	79.4	77.4	76.3333333	80	79.7	79.875	82.5	3.6
Nottingham Forest	78.625	75.2	76.2	69	77	77	83.625	81.75	3.6
Southampton	77.875	78	74.6	80.3333333	72.1428571	76.8	79.125	81.625	3.4
Tottenham Hotspur	84.125	82	84.4	75.3333333	81	82	85.25	85.125	3.9
West Ham United	81.25	79.6	78.8	77.3333333	82.8571429	80.8	81.375	78.5	3.8
Wolverhampton Wanderers	83.75	79.6	80.2	77.3333333	76.7142857	81.5	80.375	88.25	3.6

Źródło: Opracowanie własne

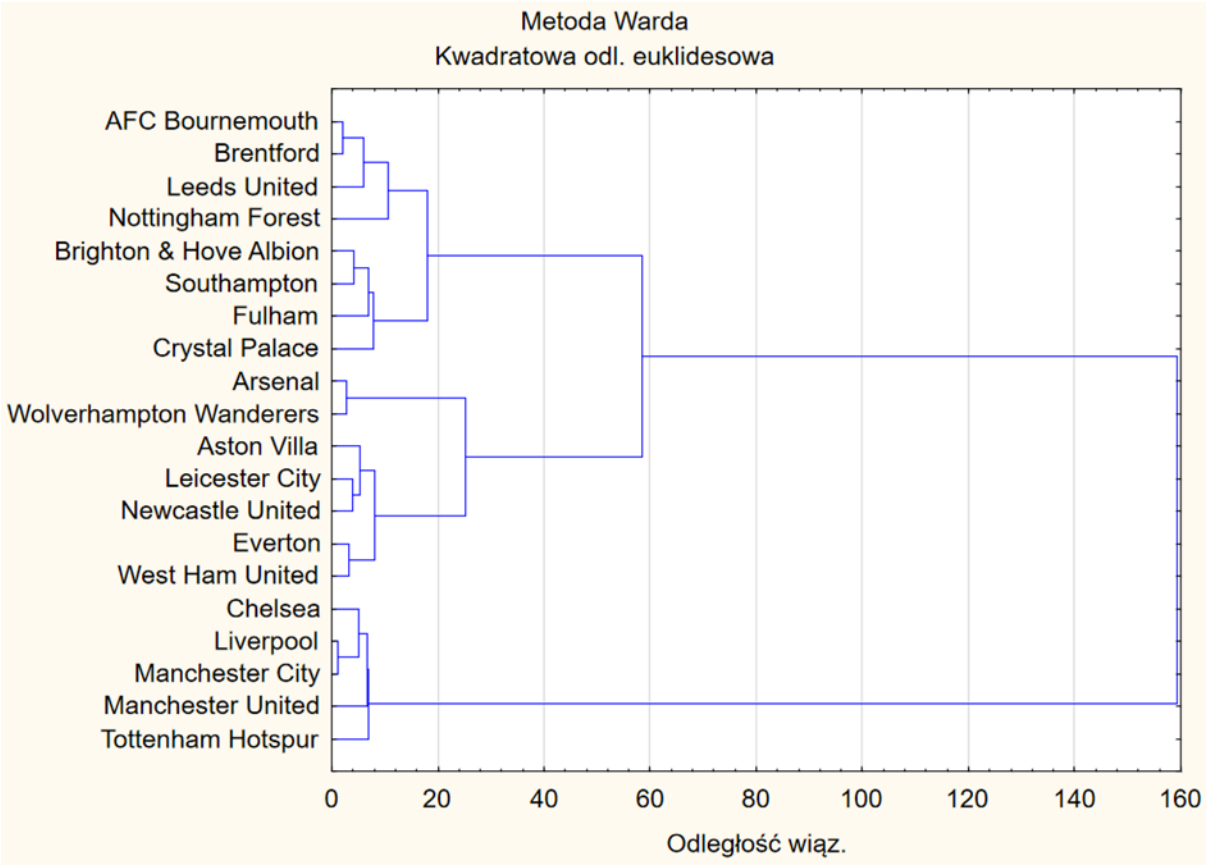
2. Wybór parametrów w aglomeracyjnej analizie skupień

Dane wykorzystane do analizy skupień muszą być standaryzowane, gdyż nie mogą posiadać jednostek, jeśli mają być porównywalne. Jako metodę grupowania wybrano Aglomerację. W sekcji grupuj wybrano przypadki, ponieważ celem jest grupowanie klubów. Jako metodę aglomeracji wybrano metodę Warda, która sprawia, że skupienia cechują się wysokim stopniem wewnętrznej spójności i podobieństwa jak również jednocześnie dużym zróżnicowaniem pomiędzy sobą.

3. Diagram drzewa

W kontekście przeprowadzonej aglomeracji, udało się wygenerować diagram drzewa, będący reprezentacją hierarchicznej struktury między klubami. Diagram ten ilustruje połączenia między klubami, ułatwiając analizę ich grupowania. Na wstępie każdy klub jest traktowany jako odrębne skupienie. W kolejnych krokach aglomeracji skupienia te łączą się w struktury wyższego rzędu, a proces ten kontynuuje się, aż do uzyskania jednego, zintegrowanego klastra zawierającego wszystkie kluby.

Rysunek 10: Diagram drzewa



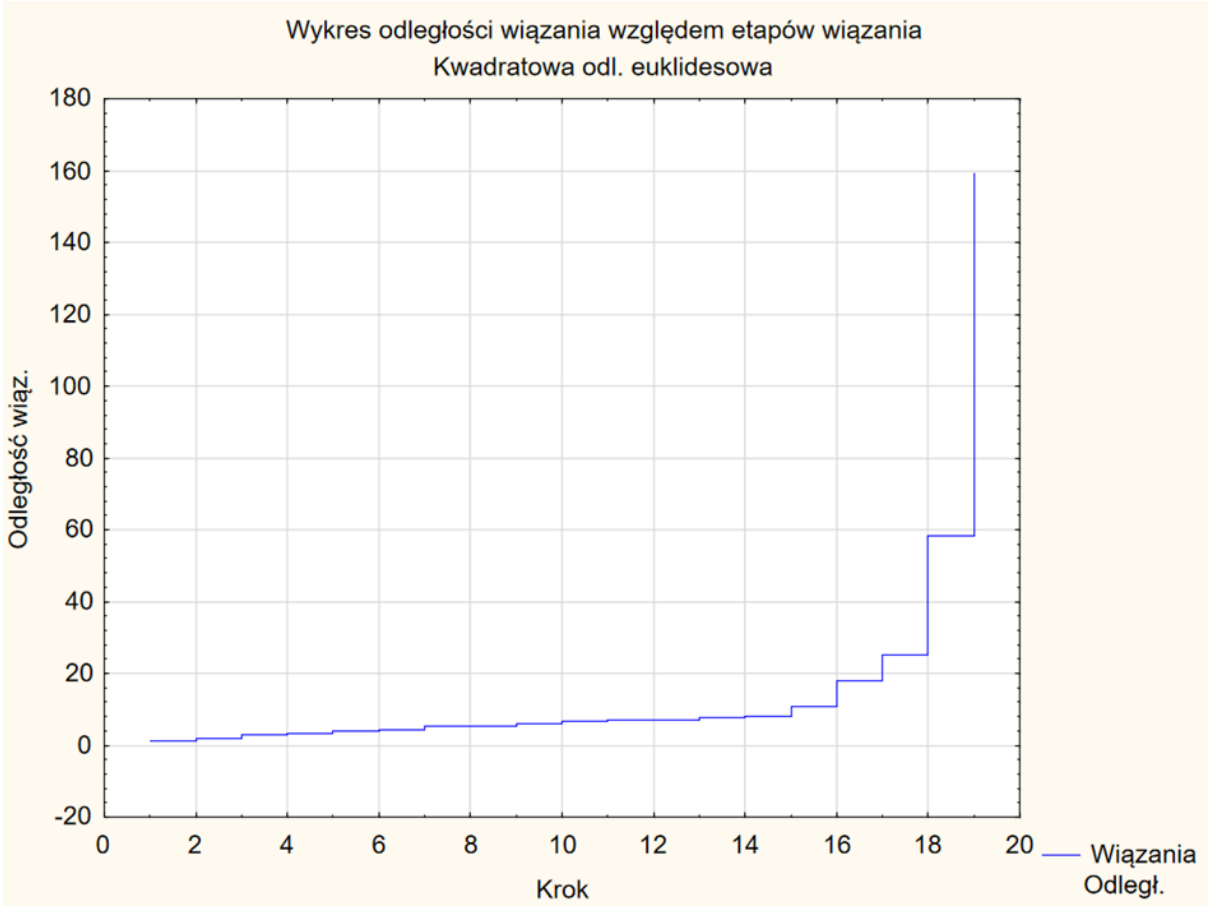
Źródło: Opracowanie własne

Diagram drzewa jest narzędziem wizualnym umożliwiającym precyzyjną analizę procesu aglomeracji. Pozwala on na identyfikację klastrów i obserwację ich ewolucji w czasie. Struktura hierarchiczna drzewa odzwierciedla, jak pojedyncze jednostki (kluby) konsolidują się w większe zespoły, aż do uzyskania jednego, kompleksowego klastra obejmującego wszystkie kluby.

4. Decyzja o przycięciu drzewa

Wykorzystując wykres przebiegu aglomeracji ustalono punkt przecięcia drzewa. Dokonano tego na podstawie analizy linii pionowych tworzących wykres.

Rysunek 11: Wykres przebiegu aglomeracji



Źródło: Opracowanie własne

Zidentyfikowano moment, w którym dochodzi do znacznego wzrostu odległości między połączonymi skupieniami, sygnalizując naturalny punkt podziału między klubami. Przeprowadzając analizę rysunku, taki moment występuje w przypadku dziewiętnastego kroku aglomeracji.

Na osi Y przypisana mu jest odległość między 59 a 160. Wybierając punkt, w którym odległość ta mieszcząca się w tym zakresie, możemy zauważyć, że dla tej konkretnej odległości powstają dwa odrębne skupienia.

5. Przynależność do skupień

Do powstałych skupień należą następujące kluby:

- skupienie 1: Chelsea, Liverpool, Manchester United, Manchester City, Tottenham Hotspur

- skupienie 2: AFC Bournemouth, Arsenal, Aston Villa, Brentford, Brighton & Hove Albion, Crystal Palace, Everton, Fulham, Leeds United, Leicester City, Newcastle United, Nottingham Forest, Southampton, West Ham United, Wolverhampton Wanderers

Rysunek 12: Przynależność do skupień

	Przynależność do skupień (analizaskupien) Odległość łączenia = 140 Metoda Warda Kwadratowa odl. euklidesowa
	Przynależność do skupień
Chelsea	1
Liverpool	1
Manchester City	1
Manchester United	1
Tottenham Hotspur	1
AFC Bournemouth	2
Arsenal	2
Aston Villa	2
Brentford	2
Brighton & Hove Albion	2
Crystal Palace	2
Everton	2
Fulham	2
Leeds United	2
Leicester City	2
Newcastle United	2
Nottingham Forest	2
Southampton	2
West Ham United	2
Wolverhampton Wanderers	2

Źródło: Opracowanie własne

Zgodnie z ustalonymi kryteriami, każde z utworzonych skupień obejmuje kluby uczestniczące w angielskiej Premier League, które wykazują wyraźne podobieństwo między sobą, charakteryzując się ograniczonym zróżnicowaniem. Nie bez znaczenia jest fakt, że skupienia te prezentują znaczące różnice między sobą, co potwierdza istnienie wyraźnych kontrastów pomiędzy dwoma grupami klubów.

6. Podsumowanie

Wyniki analizy skupień mają na celu zrozumienie struktury i podobieństw pomiędzy zespołami na podstawie różnych cech statystycznych, takich jak wiek, drybling, kontrola piłki czy choćby przyśpieszenie. Przeprowadzono kompleksowy proces agregacji i kompletowania danych statystycznych dla wszystkich klubów w angielskiej Premier League w sezonie 2022/2023. Dane zostały standaryzowane, a analiza skupień została przeprowadzona przy użyciu aglomeracji z metodą Warda, celem wyodrębnienia klarownych grup zespołów. Wygenerowano diagram drzewa, prezentujący hierarchiczną strukturę między klubami. Drzewo ilustruje połączenia między zespołami, umożliwiając analizę ich grupowania. Zidentyfikowano punkt przecięcia drzewa, sygnalizujący naturalny podział między klubami. Przyjęta decyzja o przycięciu drzewa pozwoliła na identyfikację dwóch skupień zróżnicowanych pod względem charakterystyk. W rezultacie analizy skupień, utworzono dwa skupienia klubów. Skupienie 1 obejmuje najlepsze pięć drużyn ligi, które corocznie plasują się w czubie tabeli. Natomiast skupienie 2 zawiera pozostałe zespoły, znacznie słabsze od tych ze skupienia 1. Potwierdzono zatem, że w wyniku analizy skupień wszystkich zespołów angielskie Premier League, można oczekiwać wyraźnego podziału pomiędzy czołowymi drużynami, które bezpośrednio walczą o mistrzostwo kraju a pozostałymi, co będzie odzwierciedleniem różnic w ich stylu gry, umiejętnościach i osiągnięciach.

Analiza korespondencji

Hipoteza: „Istnieje bliska relacja między ilością zdobytych tytułów mistrzowskich a popularnością klubów w angielskiej Premier League.”

Cel: W kontekście hipotezy dotyczącej angielskiej Premier League, analiza korespondencji pozwoli na zrozumienie relacji między ilością zdobytych tytułów mistrzowskich przez kluby, a ich popularnością wśród kibiców.

Wybrano analizę korespondencji, ponieważ naliza korespondencji jest metodą statystyczną wykorzystywaną w badaniach nad relacjami między kategorialnymi zmiennymi, a to spełnia oczekiwania hipotezy.

Można analizą zbadać, czy istnieje statystycznie istotny trend w kierunku większej popularności klubów, które osiągnęły więcej tytułów mistrzowskich. Jeśli hipoteza zostanie potwierdzona, będzie to oznaczać, że zdobycie tytułów mistrzowskich wpływa pozytywnie na poziom popularności klubu. Dzięki osiąganiu sukcesów, kluby mogą zyskiwać większą uwagę kibiców, medialną ekspozycję i tym samym zwiększać swoją popularność wśród społeczności piłkarskiej.

1. **Podział danych**

Tabela z danymi, które zostały wykorzystane do analizy zawiera informacje dotyczące wszystkich klubów grających w angielskiej Premier League w sezonie 2022/2023. Kluby zostały sklasyfikowane na trzy kategorie związane z ilością zdobytych tytułów mistrzowskich oraz trzy kategorie dotyczące popularności. Oto ogólny podział:

Ilość Tytułów Mistrzowskich:

- Brak mistrzostw
- Od 1 do 7 mistrzostw
- Powyżej 7 mistrzostw

Popularność:

- Mało popularny
- Popularny
- Bardzo popularny

Popularność klubów została poddana klasyfikacji na podstawie subiektywnego osądu, który uwzględniał kilka czynników. Klasyfikacja "Mało popularny,"

"Popularny," i "Bardzo popularny" opierała się na ogólnym spojrzeniu na aspekty takie jak liczba kibiców, obecność w mediach społecznościowych, a także ogólna widoczność klubu w przestrzeni publicznej. Wartości te nie zostały jednak szczegółowo liczbowo zdefiniowane, a ich ustalenie opierało się na subiektywnych ocenach ogólnej popularności poszczególnych klubów.

Rysunek 13: Dane do analizy korespondencji

Klub	Ilość tytułów mistrzowskich	Popularność
AFC Bournemouth	brak mistrzostw	mało popularny
Arsenal	od 1 do 7 mistrzostw	bardzo popularny
Aston Villa	od 1 do 7 mistrzostw	popularny
Brentford	brak mistrzostw	mało popularny
Brighton & Hove Albion	brak mistrzostw	mało popularny
Chelsea	od 1 do 7 mistrzostw	bardzo popularny
Crystal Palace	brak mistrzostw	mało popularny
Everton	powyżej 7 mistrzostw	popularny
Fulham	brak mistrzostw	mało popularny
Leeds United	od 1 do 7 mistrzostw	mało popularny
Leicester City	od 1 do 7 mistrzostw	popularny
Liverpool	powyżej 7 mistrzostw	bardzo popularny
Manchester City	powyżej 7 mistrzostw	bardzo popularny
Manchester United	powyżej 7 mistrzostw	bardzo popularny
Newcastle United	od 1 do 7 mistrzostw	popularny
Nottingham Forest	od 1 do 7 mistrzostw	mało popularny
Southampton	brak mistrzostw	mało popularny
Tottenham Hotspur	od 1 do 7 mistrzostw	popularny
West Ham United	brak mistrzostw	popularny
Wolverhampton Wanderers	od 1 do 7 mistrzostw	popularny

Źródło: Opracowanie własne

2. Współrzędne wierszy i kolumn, oraz ich wkład do bezwładności

Powstały dwie tabele, które dostarczają sporo cennych informacji. Na początku zinterpretowana zostanie tabela z współrzędnymi wierszy. Kolumna „Masa” zawiera częstości względne pokazujące wkład poszczególnych kategorii wierszowych w tworzenie wiersza. Można powiedzieć, że częstości względne pokazują, jak jednostka masy jest rozłożona na poszczególne komórki. Największy udział w tworzeniu masy wiersza ma kategoria „od 1 do 7 mistrzostw” – 45%, na drugim miejscu plasuje się „brak mistrzostwa” – 35 %, a na końcu „powyżej 7 mistrzostw”. Kolumna „Jakość” zawiera informacje dotyczące jakości reprezentacji wiersza przez punkt w wybranym układzie współrzędnych. W przeprowadzanej analizie jakość jest maksymalna i wynosi 1. Kolejna kolumna dotyczy względnej bezwładności.

Bezwładność względna wyraża udział danego punktu w bezwładności ogólnej mierzony niezależnie od liczby wymiarów wybranych przez badacza. Im wartość jest bliższa 0, tym mniejsza jest różnica pomiędzy profilem a profilem przeciętnym. Najbardziej zbliżona do profilu przeciętnego jest kategoria „od 1 do 7 mistrzostw", natomiast najmniej zbliżona jest kategoria „brak mistrzostw".

Rysunek 14: Współrzędne wierszy i wkład do bezwładności

Nazwa wiersza	Wiersz Liczba	Współrz. Wymiar1	Współrz. Wymiar2	Masa	Jakość	względna bezwład.	bezwład. Wymiar1	Cos^2 Wymiar1	bezwład. Wymiar2	Cos^2 Wymiar2
brak mistrzostw	1	-0.922574	0.209525	0.350000	1.000000	0.453844	0.533141	0.950951	0.116859	0.049049
od 1 do 7 mistrzostw	2	0.234687	-0.384375	0.450000	1.000000	0.132227	0.044357	0.271558	0.505643	0.728442
powyżej 7 mistrzostw	3	1.086459	0.498175	0.200000	1.000000	0.413929	0.422502	0.826275	0.377498	0.173725

Źródło: Opracowanie własne

Wartości w kolumnie Masa dla tabeli z współrzędnymi kolumn prezentują się następująco:

- Kategoria "bardzo Popularny" ma największy udział w tworzeniu masy kolumny – aż 40%.
- Kategoria "mało Popularny" ma udział w masie na poziomie 35%.
- Kategoria "popularny" przyczynia się do masy wiersza także w stopniu 35%.

We wszystkich przypadkach jakość reprezentacji osiąga maksymalną wartość 1. Kolumna "Względna Bezwładność" wyrażająca udział danego punktu w bezwładności ogólnej mierzony niezależnie od liczby wymiarów wybranych przez badacza.

- Kategoria "mało popularny" jest najmniej zbliżona do profilu przeciętnego, co oznacza większą różnicę pomiędzy nią a przeciętnym profilem.
- Kategoria "bardzo popularny" jest mniej zbliżona do profilu przeciętnego porównaniu do kategorii "popularny".
- Kategoria "popularny" jest najbardziej zbliżona do profilu przeciętnego, co wskazuje na mniejszą różnicę pomiędzy nią a przeciętnym profilem.

Rysunek 12: Współrzędne kolumn i wkład do bezwładności

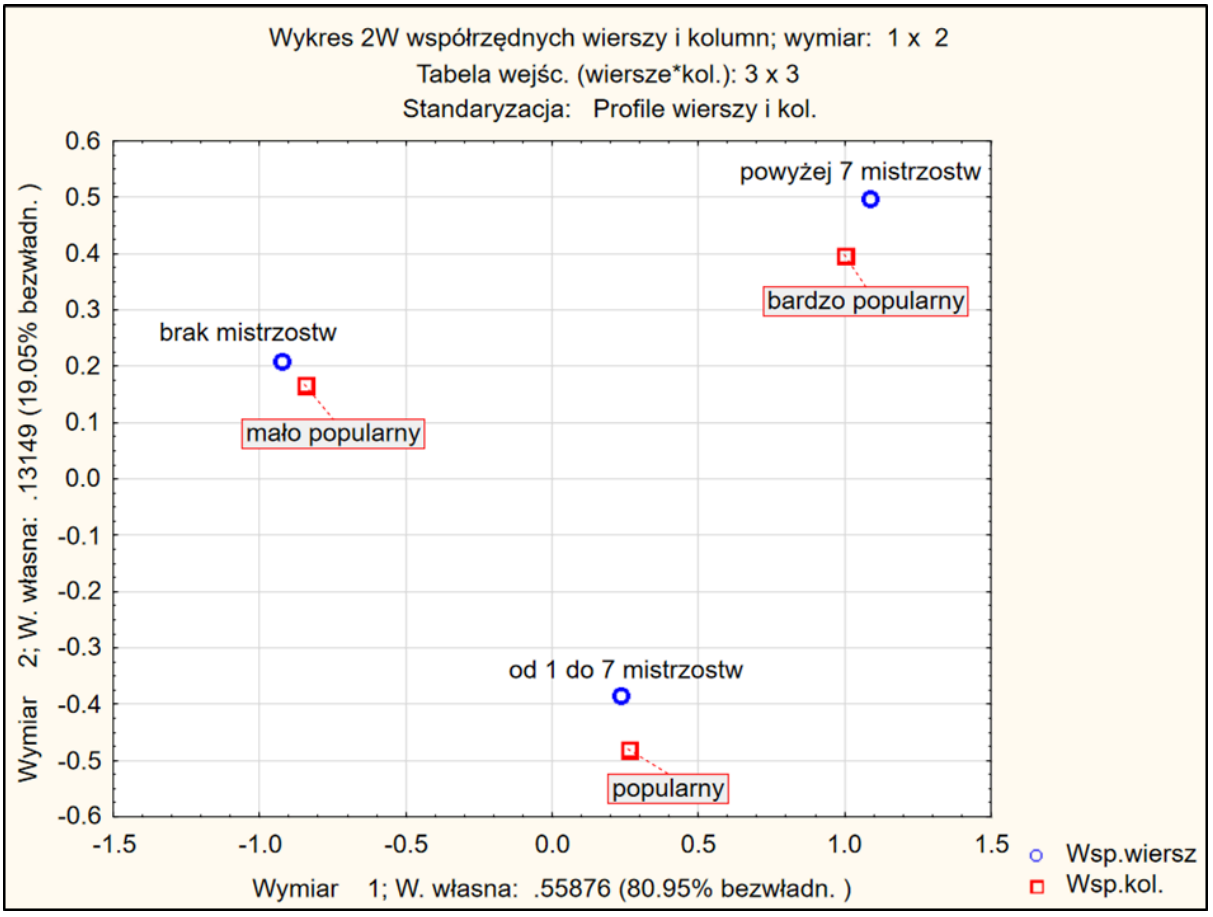
Kolumna wiersza	Kolumna Liczba	Współrz. Wymiar1	Współrz. Wymiar2	Masa	Jakość	względna bezwład.	bezwład. Wymiar1	Cos^2 Wymiar1	bezwład. Wymiar2	Cos^2 Wymiar2
mało popularny	1	-0.847163	0.168363	0.400000	1.000000	0.432326	0.513767	0.962004	0.086233	0.037996
bardzo popularny	2	0.997652	0.400306	0.250000	1.000000	0.418528	0.445318	0.861326	0.304682	0.138674
popularny	3	0.255578	-0.478348	0.350000	1.000000	0.149146	0.040915	0.222074	0.609085	0.777926

Źródło: Opracowanie własne

3. Wykres współrzędnych

Stworzono wspólny wykres dla profili wierszowych i kolumnowych. Nowo powstały graficzny reprezentant danych umożliwia jednoczesną wizualizację i porównanie struktury kategorii zarówno w kontekście wierszy, jak i kolumn.

Rysunek 15: Wykres współrzędnych wierszy i kolumn



Źródło: Opracowanie własne

Z informacji zawartych na wykresie wynika, że kluby, które zdobyły w swej historii powyżej 7 mistrzostw kraju cieszą się największą popularnością wśród kibiców. Natomiast brak choćby jednego mistrzostwa oznacza małą popularność zespołu.

4. Podsumowanie

Analiza korespondencji w kontekście relacji między ilością zdobytych tytułów mistrzowskich ligi angielskiej przez dany klub, jego popularnością powiodła się. Hipoteza zakładała bliską zależność między wspomnianymi zmiennymi, a rezultaty analizy korespondencji potwierdzają, że istnieje silna korelacja między sukcesami sportowymi klubów, a ich pozycją w oczach kibiców. Kluby, które zdobyły więcej niż 7 mistrzostw, cieszą się największą popularnością.

Wspólny wykres dla profili wierszowych i kolumnowych potwierdza te powiązania, umożliwiając jednoczesną wizualizację struktury kategorii. Małe odległości na wykresie między punktami współrzędnych wierszy i współrzędnych kolumn, wskazują na silną relację między ilością tytułów mistrzowskich a popularnością, co raz jeszcze potwierdza sformułowaną hipotezę.

Analiza wariancji dwuczynnikowa

Hipoteza: “Wiek i szybkość piłkarzy ma wpływ na potencjał ataku danej drużyny.”

Cel: Z wykorzystaniem analizy wariancji dwuczynnikowej ustalić, czy w sezonie 2022 wiek i szybkość piłkarzy miały znaczenie na jakość ataku danej drużyny. Dzięki takiemu badaniu, potencjalnie będzie można wysunąć wniosek, czy starsi i szybszy zawodnicy, jak powszechnie się uważa, są skuteczniejsi w ataku.

Wykorzystano wariancję dwuczynnikową, ponieważ badamy zależność pomiędzy dwiema zmiennamy jakościowymi - wiek, szybkość - a zmienną zależną - potencjał ataku.

1. Dane

Do analizy wariancji dwuczynnikowej wykorzystano tabelę składającą się z 15 wierszy (rekordów) i 8 kolumn (zmiennych). Pod uwagę wzięto najlepsze kluby Premier League oraz kilka dodatkowych klubów z różnych lig w celu uzupełnienia statystyk o bardziej skrajnych wynikach. W danej analizie skupiamy się na pięciu zmiennych:

Age, Age category, SprintSpeed, SprintSpeed category oraz AttOverall.

Age, czyli średni wiek zawodników danej drużyny, mieści się w przedziale 24-29 i na tej podstawie zostały utworzone kategorie wiekowe w kolumnie Age category. SprintSpeed, czyli średnia szybkość zawodników danej drużyny, została sklasyfikowana do kategorii Wolny (Slow, <83.00) oraz Szybki (Fast, >83.01).

Rysunek 16: Dane do analizy wariancji dwuczynnikowej

	1 Club	2 Age	3 Age category	4 Dribbling	5 SprintSpeed	6 SprintSpeed category	7 Strength	8 AttOverall
1	Arsenal	24.09	<24.5	83.13	82.00	Slow	80.13	82.00
2	Aston Villa	25.42	24.5-25.5	80.75	81.63	Slow	81.25	79.00
3	Chelsea	28.09	>27.6	84.75	85.25	Fast	84.00	86.67
4	Crystal Palace	26.64	26.6-27.5	79.88	80.25	Slow	81.00	78.33
5	Leicester City	26.73	26.6-27.5	80.75	81.38	Slow	78.50	79.33
6	Liverpool	26.91	26.6-27.5	86.38	86.75	Fast	84.25	87.33
7	Manchester City	27.18	26.6-27.5	87.38	82.75	Slow	83.50	87.00
8	Manchester United	25.36	24.5-25.5	83.88	86.50	Fast	83.88	86.67
9	Newcastle United	26.80	26.6-27.5	81.38	82.75	Slow	82.00	78.33
10	Tottenham Hotspur	26.08	25.6-26.5	84.13	84.75	Fast	82.63	87.00
11	West Ham United	28.09	>27.6	81.25	80.38	Slow	83.75	81.00
12	AS Monaco	24.36	<24.5	80.75	85.38	Fast	83.88	81.67
13	PSV	24.00	<24.5	79.38	85.25	Fast	81.88	79.00
14	Burnley	26.09	25.6-26.5	75.75	80.50	Slow	76.00	75.33
15	Norwich City	26.18	25.6-26.5	72.88	81.50	Slow	80.25	71.67

Źródło: Opracowanie własne

2. Jednowymiarowe testy istotności dla Ogólnej Oceny Ataku

Rysunek 17: Jednowymiarowe testy istotności dla Ogólnej Oceny Ataku

Efekt	Jednowymiarowe testy istotnosci dla AttOverall (teams-stats_2factor Parametryzacja z sigma-ograniczeniami Dekompozycja efektywnych hipotez SS	Stopnie swobody	MS	F	p
Wyraz wolny	82350,40	1	82350,40	6534,476	0,000000
Age category	29,49	4	7,37	0,585	0,688272
SprintSpeed category	122,20	1	122,20	9,697	0,026432
Age category*SprintSpeed category	78,05	4	19,51	1,548	0,317835
Blad	63,01	5	12,60		

Źródło: Opracowanie własne

Powyższa tabela przedstawia kilka testów istotności dla Ogólnej Oceny Ataku dla zmiennych niezależnych “Age category” i “SprintSpeed category” oraz interakcji pomiędzy obiema zmiennymi. Szczególnie zwracamy uwagę na kolumnę ‘p’, która mówi, że statystycznie istotna jest jedynie szybkość piłkarzy.

Zupełnie nieistotny wydaje się wiek piłkarzy dla powyższych danych. Biorąc wszystko pod uwagę, naturalnym jest fakt, że interakcja pomiędzy wiekiem i szybkością będzie z założenia niezbyt znacząca. W tabeli widać przy interakcji wartość 0.31, co potwierdza fakt niewielkiej istotności.

3. Średnie nieważone

Rysunek 18: Średnie nieważone, tabela

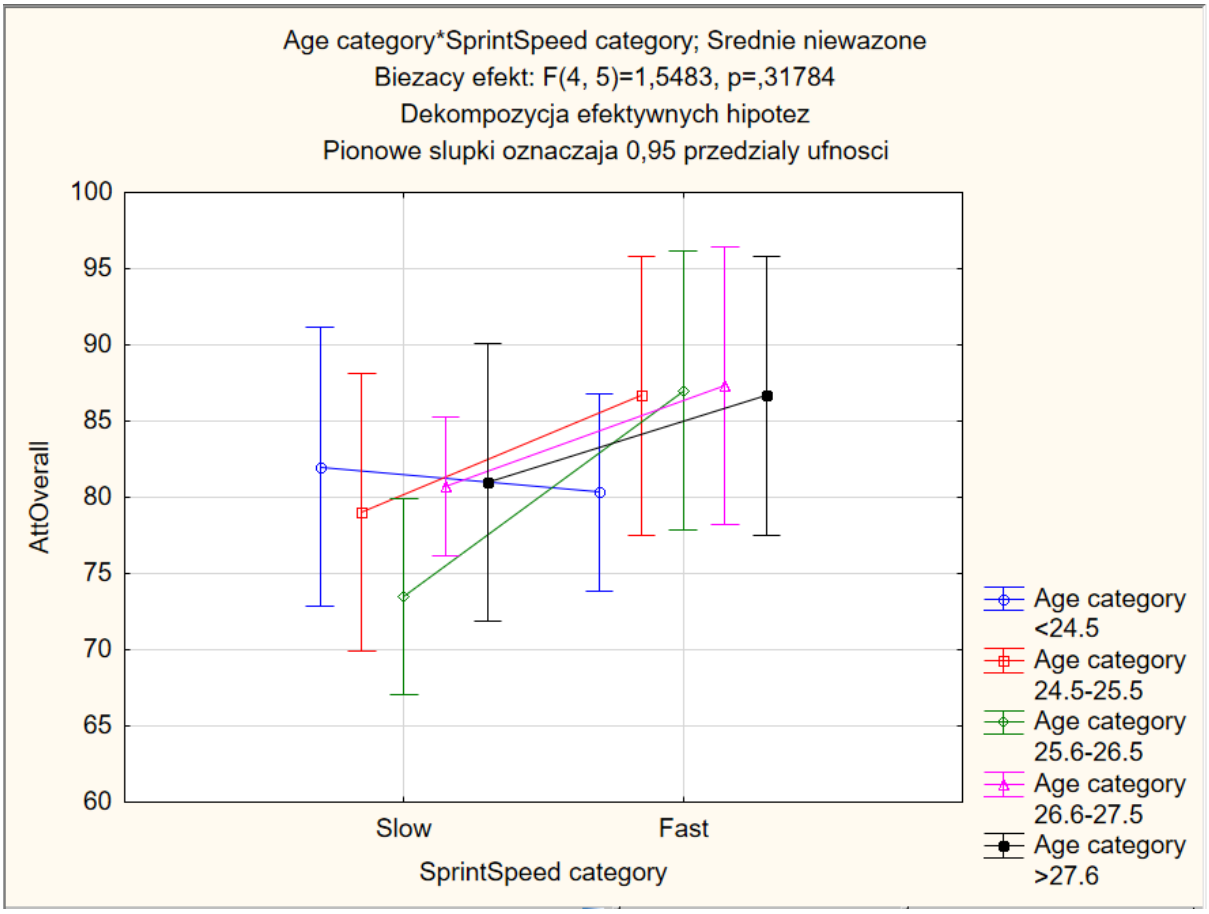
Nr podkl.	Age category*SprintSpeed category; Srednie niewazone (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw) Biezacy efekt: F(4, 5)=1,5483, p=,31784 Dekompozycja efektywnych hipotez Age category	SprintSpeed category	AttOverall Srednie	AttOverall Bl. Std.	AttOverall -95,00%	AttOverall +95,00%	N
1	<24.5	Slow	82,00000	3,549993	72,87445	91,12555	1
2	<24.5	Fast	80,33500	2,510224	73,88226	86,78774	2
3	24.5-25.5	Slow	79,00000	3,549993	69,87445	88,12555	1
4	24.5-25.5	Fast	86,66667	3,549993	77,54112	95,79221	1
5	25.6-26.5	Slow	73,50000	2,510224	67,04726	79,95274	2
6	25.6-26.5	Fast	87,00000	3,549993	77,87445	96,12555	1
7	26.6-27.5	Slow	80,75000	1,774996	76,18723	85,31277	4
8	26.6-27.5	Fast	87,33333	3,549993	78,20779	96,45888	1
9	>27.6	Slow	81,00000	3,549993	71,87445	90,12555	1
10	>27.6	Fast	86,66667	3,549993	77,54112	95,79221	1

Źródło: Opracowanie własne

Powyższa tabela przedstawia średnie wyniki zmiennej zależnej we wszystkich wariantach zmiennych niezależnych. Skupiamy swoją uwagę na różnicach w kolumnie “AttOverall Średnie”. Można zauważyć, że wyniki różnią się od siebie znacząco, np. porównując wiersze nr 5 i 6, czyli wartości 73,5 oraz 87. Taka obserwacja pozwala wysunąć wniosek, że zachodzą zależności, które warto zbadać.

a) Wykres interakcji (x - kategoria szybkości biegu)

Rysunek 19: Wykres interakcji (x - kategoria szybkości biegu)

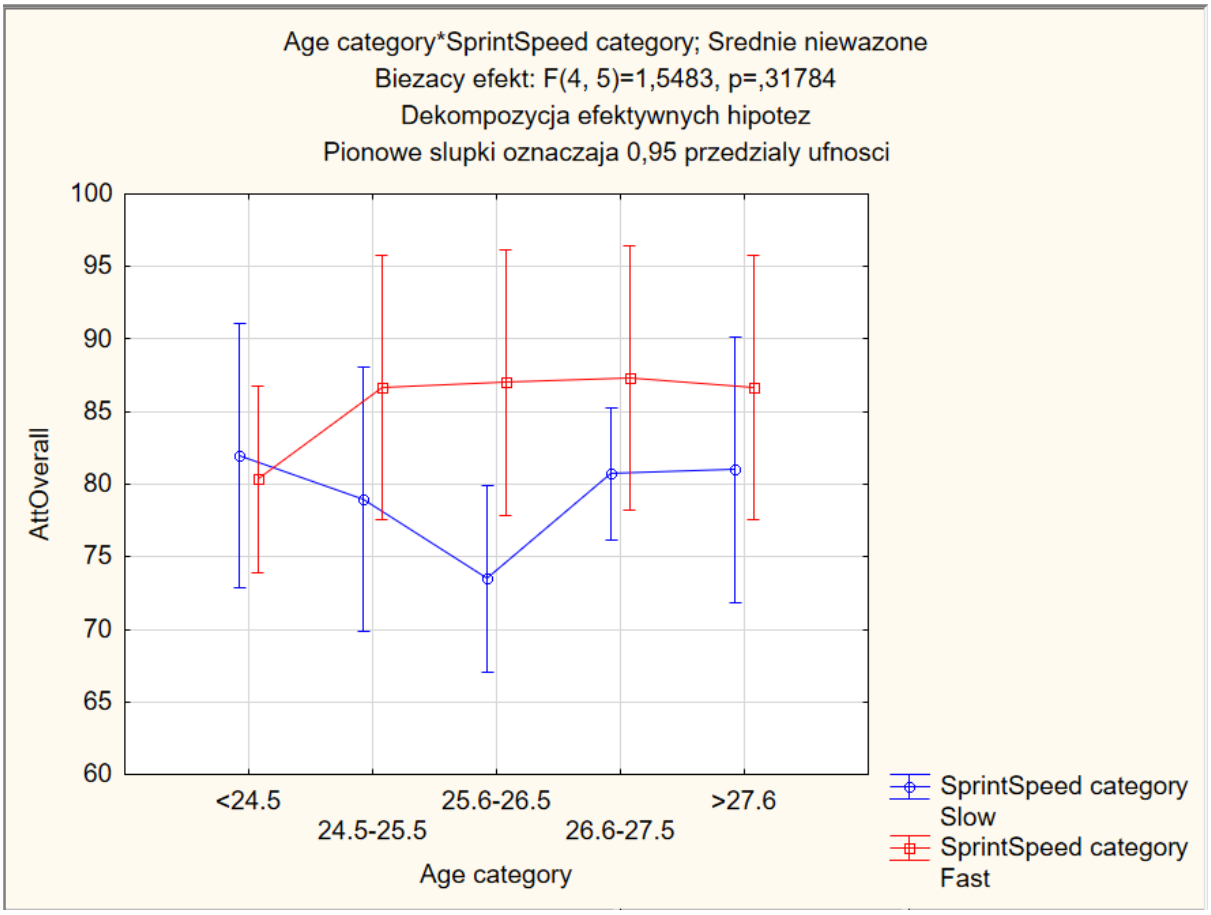


Źródło: Opracowanie własne

Z powyższego wykresu interakcji można wysunąć kilka wniosków. Rozbieżność danych jest zauważalna, przez co trudno stwierdzić jednoznacznie konkretne zależności. Zaobserwować można, że zawodnicy cechujący się wysoką szybkością biegu osiągają lepszy potencjał ataku niezależnie od wieku. Jedynym przypadkiem, który neguje tę tezę są młodzi zawodnicy, z kategorii poniżej 24.5. W ich przypadku wolniejsi zawodnicy wypadają lepiej w kontekście “AttOverall”.

b) Wykres interakcji (x - kategoria wieku)

Rysunek 20: Wykres interakcji (x - kategoria wieku)



Źródło: Opracowanie własne

Z Rysunku 20 można wysunąć podobne wnioski, jak z Rysunku 19, aczkolwiek dochodzi kilka nowych obserwacji. Widać pewną tendencję spadkową dla środkowej kategorii wieku w przypadku wolniejszych zawodników. Można stwierdzić, że zawodnicy cechujący się wolniejszym biegiem, osiągają najmniejszy potencjał ataku będąc w “średnim” wieku, a największy na początku kariery lub już po zdobyciu kilkuletniego doświadczenia.

4. Tabela wartości z Histogramów

Rysunek 21: Tabela danych z Histogramów

Efekt	K-S (D)	Wartość p-Lillieforsa	S-W	Wartość p
Wiek <24.5	0,349	<0,15	0,8314	0,1919
Wiek 25.6-26.5	0,2973	<1	0,9166	0,4404
Wiek 26.6-27.5	0,3206	<0,1	0,7456	0,0271
Szybkość Slow	0,2053	<1	0,9615	0,8134
Szybkość Fast	0,3768	<0,01	0,7623	0,0262
Slow * 26.6-27.5	0,3823	0,05	0,7106	0,0153

Źródło: Opracowanie własne

Powyższa tabela przedstawia wartości z 6 histogramów. Tylko dla podanych efektów z kolumny “Efekt” zaobserwowano poprawne statystyki, dlatego reszta nieistotny wyników nie została umieszczona w tabeli. Można zauważyć, że największą istotność, kolejny raz posługując się wartością “p”, mają: kategoria wiekowa 26.6-27.5, kategoria szybkości biegu “Fast” oraz interakcja “Slow * 26.6-27.5.

5. Testy jednorodności wariancji

Testy jednorodności wariancji to statystyczne metody używane do oceny, czy różne grupy w próbie badawczej mają równe wariancje. Są one kluczowe w analizie statystycznej, ponieważ wiele testów, takich jak analiza wariancji (ANOVA), wymaga, aby wariancje w grupach były mniej więcej równe, aby wyniki były wiarygodne. Istnieje kilka różnych testów jednorodności wariancji, w tym Test Levene'a, Test Hartleya F-max, Test Cochrana C, i Test Bartletta, które różnią się metodologią i wrażliwością na różne założenia dotyczące danych, takie jak rozkład normalny.

a) Efekt: Kategoria wiekowa

Rysunek 22: Test Levene’a jednorodności wariancji z efektem: Kategoria wiekowa

Test Levene'a jednorodnosci wariancji (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw)
Efekt: "Age category"
Stopnie swobody dla kazdego F : 4, 10

	MS Efekt	MS Blad	F	p
AttOverall	8,997641	2,503782	3,593620	0,045896

Źródło: Opracowanie własne

Rysunek 23: Pozostałe testy jednorodności wariancji z efektem: Kategoria wiekowa

Testy jednorodnosci wariancji (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw)
Efekt: "Age category"

	Hartleya F-maks	Cochrana C	Bartlett Chi-kw.	a df	p
AttOverall	23,68507	0,477975	3,184570	4	0,527427

Źródło: Opracowanie własne

b) Efekt: Kategoria szybkości biegu

Rysunek 24: Test Levene'a jednorodności wariancji z efektem: Kategoria szybkości biegu

Test Levene'a jednorodnosci wariancji (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw)
Efekt: "SprintSpeed category"
Stopnie swobody dla kazdego F : 1, 13

	MS Efekt	MS Blad	F	p
AttOverall	0,013390	6,272429	0,002135	0,963850

Źródło: Opracowanie własne

Rysunek 25: Pozostałe test jednorodności wariancji z efektem: Kategoria szybkości biegu

Testy jednorodnosci wariancji (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw)
Efekt: "SprintSpeed category"

	Hartleya F-maks	Cochrana C	Bartlett Chi-kw.	a df	p
AttOverall	1,471401	0,595371	0,204643	1	0,650999

Źródło: Opracowanie własne

c) Efekt: Interakcja Kategorii wiekowej i szybkości biegu

Rysunek 26: Test Levene'a jednorodności wariancji z efektem: Interakcja Kategorii wiekowej i szybkości biegu

Test Levene'a jednorodnosci wariancji (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw)
Efekt: "Age category"*"SprintSpeed category"
Stopnie swobody dla kazdego F : 9, 5

	MS Efekt	MS Blad	F	p
AttOverall	2,854091	2,737500	1,042590	0,510214

Źródło: Opracowanie własne

Rysunek 27: Pozostałe testy jednorodności wariancji z efektem: Interakcja Kategorii wiekowej i szybkości biegu

Testy jednorodnosci wariancji (teams-stats_2factor_analysis w Analiza_dwuczynnikowa.stw)
Efekt: "Age category"*"SprintSpeed category"

	Hartleya F-maks	Cochrana C	Bartlett Chi-kw.	a df	p
AttOverall	4,932972	0,631459	0,660846	2	0,718620

Źródło: Opracowanie własne

Test Levene'a jest używany do oceny, czy grupy mają równe wariancje, co jest ważnym założeniem w wielu testach statystycznych, takich jak analiza wariancji (ANOVA). Kluczowe elementy tego testu to:

MS Efekt (Mean Square Between): Jest to średnia kwadratów różnic między średnimi grup a ogólną średnią. Wskaźnik ten mierzy zmienność między grupami.
MS Błąd (Mean Square Within): Odzwierciedla średnią kwadratów różnic wewnątrz grup. Mierzy zmienność wewnątrz grup.
F: Jest to statystyka testowa obliczana jako stosunek MS Efekt do MS Błąd. Wysoka wartość F wskazuje na to, że wariancje między grupami są znacząco różne.
P (p-value): Wartość p wskazuje, jak prawdopodobne jest uzyskanie obserwowanych wyników, jeśli założymy, że hipoteza zerowa (jednorodność wariancji) jest prawdziwa. Jeśli wartość p jest mniejsza niż wybrany poziom istotności (zazwyczaj 0.05), odrzucamy hipotezę zerową o równości wariancji.

Testy dotyczące drugich tabel:

Test Hartleya F-max: Używany do oceny, czy kilka próbek pochodzi z populacji o równych wariancjach. W teście tym oblicza się stosunek największej próbkowej wariancji do najmniejszej. Jest szczególnie wrażliwy na różnice w wariancjach, gdy liczba próbek i ich rozmiar są równe.

Test Cochrana C: Podobnie jak test Hartleya, koncentruje się na stosunku największej wariancji do sumy wszystkich wariancji. Jest użyteczny, gdy rozmiary próbek są równe, ale może być mniej wiarygodny, gdy rozmiary próbek są różne.

Test Bartletta (Chi-kwadrat): Testuje hipotezę o równości wariancji w kilku próbkach, używając statystyki chi-kwadrat. Jest bardziej wrażliwy na odstępstwa od normalności niż test Levene'a.

df (stopnie swobody): W testach takich jak test Bartletta, stopnie swobody związane są z liczbą grup minus jeden. Są one ważne dla określenia krytycznych wartości w rozkładzie chi-kwadrat.
p (p-value): Podobnie jak w teście Levene'a, wartość p wskazuje na prawdopodobieństwo odrzucenia hipotezy zerowej o równości wariancji, gdy w rzeczywistości jest ona prawdziwa.

W każdym z tych testów, kluczowe jest zrozumienie, że odrzucenie hipotezy zerowej wskazuje na to, że wariancje w grupach są statystycznie różne, co może wpływać na ważność wyników innych testów statystycznych, takich jak ANOVA, które zakładają jednorodność wariancji.

Powyższe dane z Rysunków 22-27 ze wszystkich podpunktów wskazują, na brak istotności kategori wiekowej oraz wysoką istotność szybkości biegu w kontekście potencjału ataku.

6. Testy NIR:

Testy NIR (Najbliższego Istotnego Różnicowania), znane również jako testy post-hoc lub testy porównań wielokrotnych, są stosowane po przeprowadzeniu analizy wariancji (ANOVA). Głównym celem tych testów jest ustalenie, które konkretne grupy różnią się istotnie od siebie pod względem statystycznym, po tym jak ANOVA wykazała istnienie ogólnych różnic między grupami.

a) **Kategoria wieku**

Rysunek 28: Test NIR dla kategorii wieku

	Test NIR; zmienna AttOverall (teams-stats_2factor_analysis w Analiza_dwuczynni Prawdopodobienstwa dla testów post-hoc Blad: MS miedzygrupowe = 12,602, df = 5,0000					
Nr podkl.	**Age category**	{1} 80,890	{2} 82,833	{3} 78,000	{4} 82,067	{5} 83,833
1	<24.5		0,574865	0,364516	0,668947	0,405397
2	24.5-25.5	0,574865		0,196043	0,806598	0,789467
3	25.6-26.5	0,364516	0,196043		0,177531	0,131752
4	26.6-27.5	0,668947	0,806598	0,177531		0,577864
5	>27.6	0,405397	0,789467	0,131752	0,577864	

Źródło: Opracowanie własne

b) Kategoria szybkości biegu

Rysunek 29: Test NIR dla kategorii szybkości biegu

Nr podkl.	Test NIR; zmienna AttOverall (teams-stats_2factor_analys Prawdopodobienstwa dla testów post-hoc Blad: MS miedzygrupowe = 12,602, df = 5,0000 SprintSpeed category	{1} 79,111	{2} 84,723
1	Slow		0,030124
2	Fast	0,030124	

Źródło: Opracowanie własne

c) Interakcja kategorii wieku i szybkości biegu

Rysunek 30: Test NIR dla interakcji kategorii wieku i szybkości biegu

Nr	Age category	SprintSpeed category	{1} 82,000	{2} 80,335	{3} 79,000	{4} 86,667	{5} 73,500	{6} 87,000	{7} 80,750	{8} 87,333	{9} 81,000	{10} 86,667
1	<24.5	Slow		0.717497	0.576168	0.395277	0.107977	0.365011	0.765526	0.336682	0.849964	0.395277
2	<24.5	Fast	0.717497		0.771182	0.205091	0.112156	0.185865	0.897888	0.168397	0.884419	0.205091
3	24.5-25.5	Slow	0.576168	0.771182		0.187270	0.261623	0.171932	0.677686	0.157833	0.706799	0.187270
4	24.5-25.5	Fast	0.395277	0.205091	0.187270		0.029137	0.949636	0.196229	0.899538	0.310244	1.000000
5	25.6-26.5	Slow	0.107977	0.112156	0.261623	0.029137		0.026702	0.064896	0.024493	0.145124	0.029137
6	25.6-26.5	Fast	0.365011	0.185865	0.171932	0.949636	0.026702		0.176142	0.949636	0.285632	0.949636
7	26.6-27.5	Slow	0.765526	0.897888	0.677686	0.196229	0.064896	0.176142		0.158076	0.952217	0.196229
8	26.6-27.5	Fast	0.336682	0.168397	0.157833	0.899538	0.024493	0.949636	0.158076		0.262777	0.899538
9	>27.6	Slow	0.849964	0.884419	0.706799	0.310244	0.145124	0.285632	0.952217	0.262777		0.310244
10	>27.6	Fast	0.395277	0.205091	0.187270	1.000000	0.029137	0.949636	0.196229	0.899538	0.310244	

Źródło: Opracowanie własne

Powyższe tabele wskazują na pary statystycznie istotne. W przypadku kategorii wieku, wszystkie pary są mało lub wcale znaczące. Dla szybkości biegu, można przyjąć, że wszystkie są bardzo istotne, a dla interakcji istnieje tylko kilka par, które w tabeli są w kolorze czerwonym.

Podsumowanie:

Wiek nie ma przełożenia na potencjał ataku zawodników. Zdarzają się przesłanki, które mogłyby wskazywać, że młodsi lub najstarsi zawodnicy sprawdzają się lepiej niż zawodnicy w średnim wieku, jednak nie jest to zależność stała. Największy wpływ na potencjał ataku spośród analizowanych danych ma szybkość biegu.

Analiza regresji

Hipoteza: Potencjał ataku danej drużyny ma wpływ na jej końcową lokatę w tabeli.

Cel: Z wykorzystaniem analizy regresji sprawdzić, jak siła ataku danej drużyny odnosi się do jej ogólnej skuteczności w lidze. Dzięki temu będzie można wysunąć odpowiedni wniosek, czy drużyna mająca silniejszych napastników, będzie lepsza od drugiej oraz przewidzieć jej miejsce w tabeli.

Wykorzystano analizę regresji, ponieważ badamy zależność jedną zmienną niezależną, a drugą zależną, co spełnia kryteria metody.

1. **Dane**

Do analizy regresji wykorzystano tabelę składającą się z 11 wierszy (rekordów) i 7 kolumn (zmiennych). Pod uwagę wzięto najlepsze kluby Premier League, zajmujące lokaty od 1 do 11 miejsca w 2022 roku. W danej analizie skupiamy się na dwóch zmiennych:

AttOverall, czyli ogólna ocena ataku danej drużyny
FinalPlace, czyli końcowa pozycja w tabeli danej drużyny

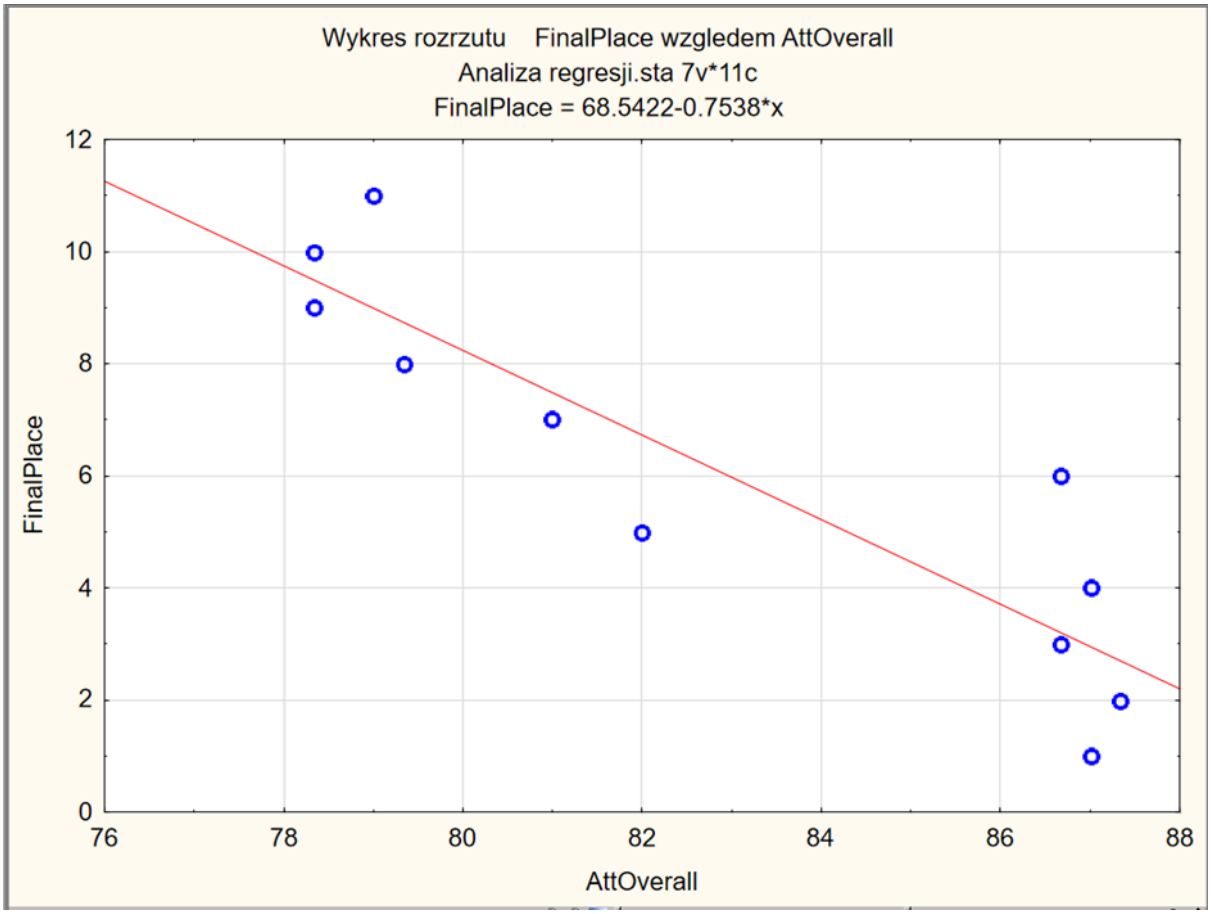
Rysunek 31: Dane do analizy regresji

	1 Club	2 Age	3 SprintSpeed	4 DefOverall	5 MidOverall	**6** **AttOverall**	7 FinalPlace
1	Arsenal	24.09	82.00	81.60	77.40	82.00	5
2	Aston Villa	25.42	81.63	82.20	77.67	79.00	11
3	Chelsea	28.09	85.25	85.40	78.40	86.67	3
4	Crystal Palace	26.64	80.25	77.00	66.20	78.33	10
5	Leicester City	26.73	81.38	81.60	71.00	79.33	8
6	Liverpool	26.91	86.75	86.20	84.00	87.33	2
7	Manchester City	27.18	82.75	87.80	80.67	87.00	1
8	Manchester United	25.36	86.50	84.40	75.00	86.67	6
9	Newcastle United	26.80	82.75	80.60	77.20	78.33	9
10	Tottenham Hotspur	26.08	84.75	82.00	73.80	87.00	4
11	West Ham United	28.09	80.38	81.20	75.75	81.00	7

Źródło: Opracowanie własne

2. Wykres rozrzutu

Rysunek 32: Wykres rozrzutu



Źródło: Opracowanie własne

Na wykresie można zaobserwować zależność liniową. Istnieje trend ujemny, który wskazuje, że drużyny z większym potencjałem ataku mają większą szansę zająć lepsze miejsce w tabeli od innych drużyn.

3. Korelacje

Rysunek 33: Tabela korelacji dla analizy regresji

Zmienna	Korelacje (Analiza regresji.sta) Oznaczone wsp. korelacji sa istotne z p < .05000 N=11 (Braki danych usuwano przypadkami) **Srednia**	Odch.std	AttOverall	FinalPlace
AttOverall	82.96970	3.945334	1.000000	-0.896688
FinalPlace	6.00000	3.316625	-0.896688	1.000000

Źródło: Opracowanie własne

Tabela przedstawia wartości średnie oraz odchylenia standardowe dla obu zmiennych - potencjał ataku oraz końcowa lokata. Druga część tabeli przedstawia macierz korelacji. Tabela wskazuje, że zależność jest całkiem silna i wraz ze wzrostem potencjału ataku drużyny, końcowa lokata jest niższa (lepsza).

4. Regresja wieloraka

a) Wyniki regresji

Podsumowanie statystyczne:

Rysunek 34: Statystyki, podsumowanie wyników regresji

statystyka	Stat.podsum.; Zmn. zal.:FinalPlace (Analiza regresji.sta) Wartosc
R wielorakie	0.896687972
Wielorakie R2	0.804049319
Skorygowane R2	0.782277021
F(1,9)	36.9299245
p	0.000184323595
Blad std. estymacji	1.5475635

Źródło: Opracowanie własne

Powyższa tabela przedstawia sześć statystyk podsumowujących, wskazujących na różne zależności.

R wielorakie, czyli współczynnik korelacji wielorakiej, przyjmuje wartości zawsze od 0 do 1. Mówi, czy zmienne niezależne są związane ze zmienną zależną. Zależność według danych z tabeli jest silna.

Wielorakie R2 wskazuje, że potencjał ataku tłumaczy końcową lokatę w tabeli w 80%.

Skorygowane R2 to dodatkowa korekta ze względu na liczbę zmiennych.

Statystyka F mówi, czy model jest statystycznie istotny.

Wartość p jest bardzo niska, czyli model jest statystycznie istotny.

Błąd standardowy estymacji wskazuje, że model może mylić się o 1.5 lokaty.

Rysunek 35: Podsumowanie regresji zmiennej zależnej AttOverall

N=11	Podsumowanie regresji zmiennej zaleznej: FinalPlace (Analiza regresji.sta) R= .89668797 R^2= .80404932 Popraw. R2= .78227702 F(1,9)=36.930 p<.00018 Blad std. estymacji: 1.5476					
	b*	Bl. std. z b*	b	Bl. std. z b	t(9)	p
W. wolny			68.54225	10.30220	6.65316	0.000093
AttOverall	-0.896688	0.147554	-0.75380	0.12404	-6.07700	0.000184

Źródło: Opracowanie własne

W powyższej tabeli najważniejsza jest wartość “b*”, czyli współczynnik standaryzowany, który mówi, że zmienna ‘AttOverall’ jest statystycznie istotna dla badanego modelu danych, wpływając na niego ujemnie.

Analiza wariancji:

Rysunek 36: Tabela analizy wariancji

Efekt	Analiza wariancji ; DV: FinalPlace (Analiza regresji.sta)				
	Suma kwadrat.	df	Srednia kwadrat.	F	p
Regres.	88.4454	1	88.44543	36.92992	0.000184
Reszta	21.5546	9	2.39495		
Razem	110.0000				

Źródło: Opracowanie własne

b) Analiza reszt regresji

Wartości przewidywane i reszty:

Rysunek 37: Tabela wartości przewidywanych i reszty

	Wartosci przewidywane i reszty FinalPlace								
	Obserw. Wartosc	Przewidyw. Wartosc	**Reszta**	Standard. Przewid.	Standard. Reszta	Bl. std. W.przew.	Mahaln. Odlegl.	Usuniete Reszta	Cooka Odlegl.
1	5.000000	6.730954	-1.730954	0.245783	-1.118502	0.481862	0.060409	-1.916786	0.074365
2	11.000000	8.992343	2.007657	1.006175	1.297302	0.678370	1.012389	2.485179	0.247756
3	3.000000	3.213238	-0.213238	-0.937049	-0.137790	0.654228	0.878060	-0.259640	0.002515
4	10.000000	9.494873	0.505127	1.175151	0.326401	0.740582	1.380980	0.655164	0.020522
5	8.000000	8.741077	-0.741077	0.921687	-0.478867	0.648981	0.849507	-0.899213	0.029687
6	2.000000	2.710707	-0.710707	-1.106025	-0.459243	0.714630	1.223291	-0.903332	0.036327
7	1.000000	2.961973	-1.961973	-1.021537	-1.267782	0.683846	1.043537	-2.438029	0.242309
8	6.000000	3.213238	2.786762	-0.937049	1.800741	0.654228	0.878060	3.393173	0.429582
9	9.000000	9.494873	-0.494873	1.175151	-0.319776	0.740582	1.380980	-0.641865	0.019697
10	4.000000	2.961973	1.038027	-1.021537	0.670749	0.683846	1.043537	1.289896	0.067827
11	7.000000	7.484750	-0.484750	0.499247	-0.313235	0.526704	0.249248	-0.548257	0.007269

Źródło: Opracowanie własne

Obserwowane wartości, czyli prawdziwe wartości w modelu.

Przewidywane wartości wskazują na przewidywane wartości naszego modelu.

Reszty, czyli różnice pomiędzy obserwacjami prawdziwymi, a obserwacjami teoretycznymi (przewidywanymi).

Standaryzowane przewidywania, czyli odchylenia standardowa dla wartości przewidywanych.

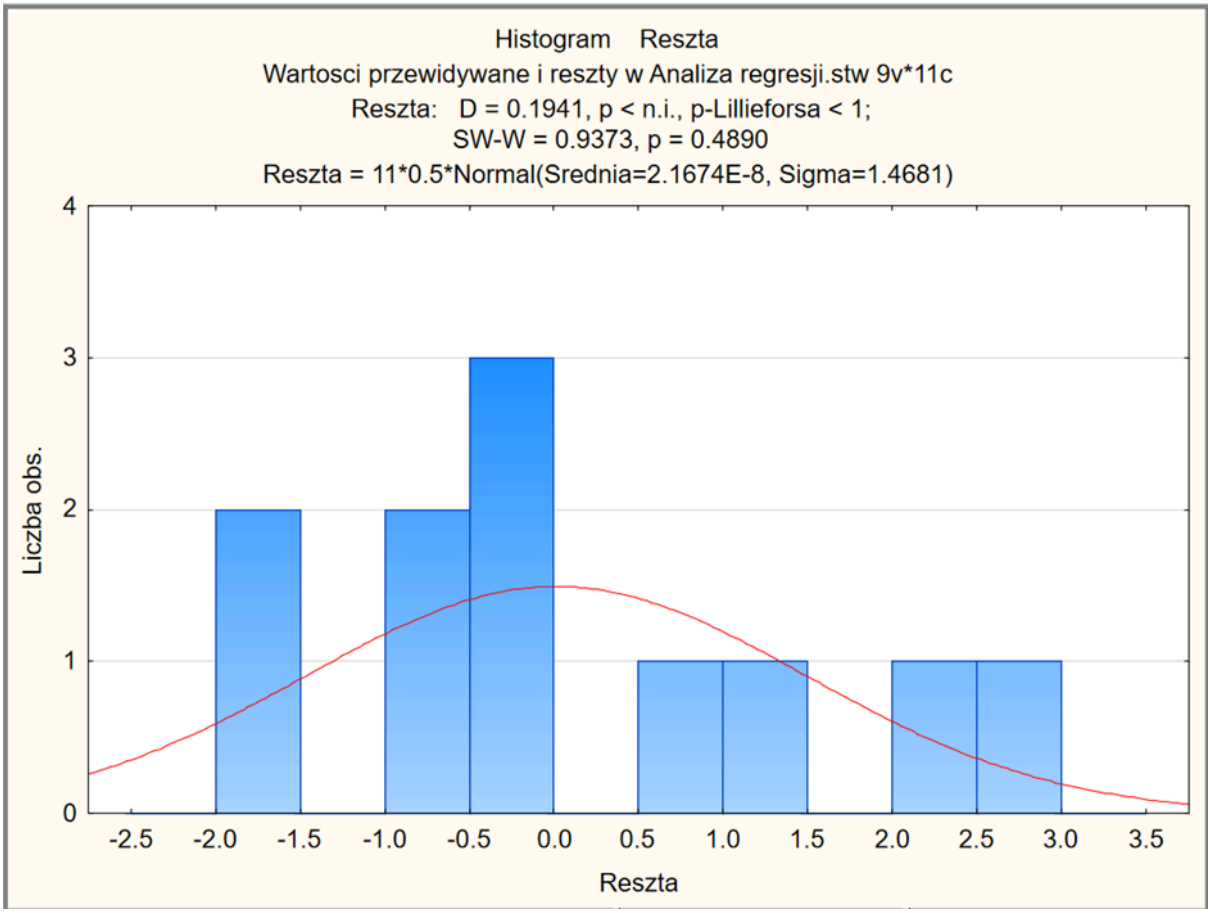
Standaryzowane reszty, czyli odchylenia standardowe dla reszt.

Błędy standaryzowane dla wartości przewidywanych.

Odległości Mahaln. oraz Cooka, które mówią czy w zbiorze pojawiają się obserwacje odstające.

Histogram Reszty:

Rysunek 38: Histogram dla zmiennej Reszta



Źródło: Opracowanie własne

Wartość D w Teście Kolmogorova-Smirnova (z poprawką Lillieforsa): Test Kolmogorova-Smirnova z poprawką Lillieforsa jest często używany do oceny normalności rozkładu. Wartość D w tym teście to maksymalna różnica między empiryczną funkcją dystrybucji próbki a funkcją dystrybucji normalnej. Im mniejsza wartość D, tym bardziej rozkład próbki przypomina rozkład normalny.

Wartość p (p-value): Wartość p wskazuje, na jakim poziomie istotności można odrzucić hipotezę zerową o normalności rozkładu. Jeśli wartość p jest większa niż wybrany poziom istotności (zazwyczaj 0.05), nie odrzucamy hipotezy o normalności rozkładu.

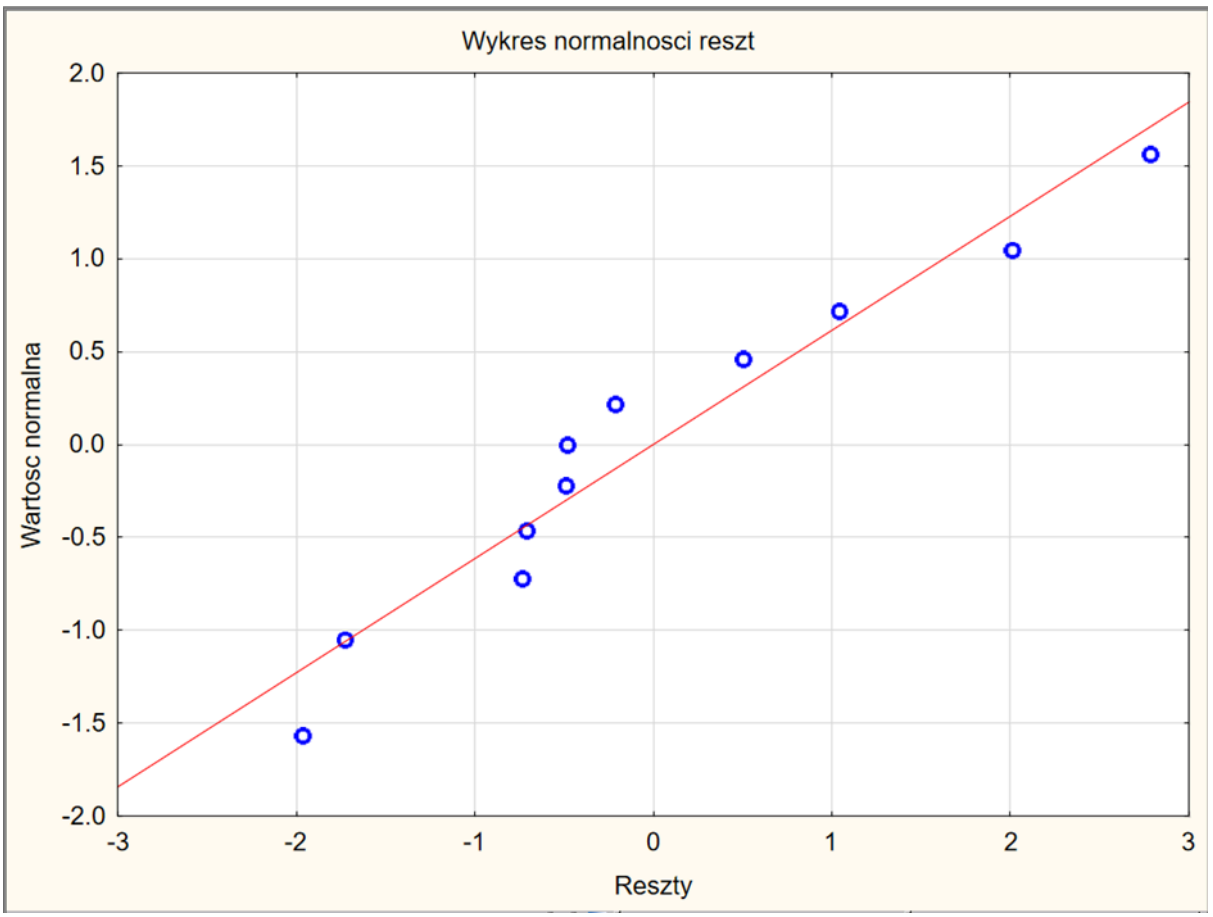
p-Lillieforsa: Jest to wartość p dla testu Lillieforsa, specyficzna wersja testu Kolmogorova-Smirnova, dostosowana do testowania normalności. Zasada interpretacji jest podobna jak wyżej - większa wartość p wskazuje na większe prawdopodobieństwo, że rozkład jest normalny.

SW-W (Test Shapiro-Wilka): Jest to inny popularny test sprawdzający normalność rozkładu. Wartość testowa w teście Shapiro-Wilka mierzy, jak bardzo próbka odchyla się od rozkładu normalnego. Podobnie jak w przypadku innych testów, większa wartość p sugeruje, że rozkład jest bliższy normalności.

Zaobserwowane wyżej wartości mogą wskazywać, że rozkład nie jest normalny.

Wykres normalności reszt:

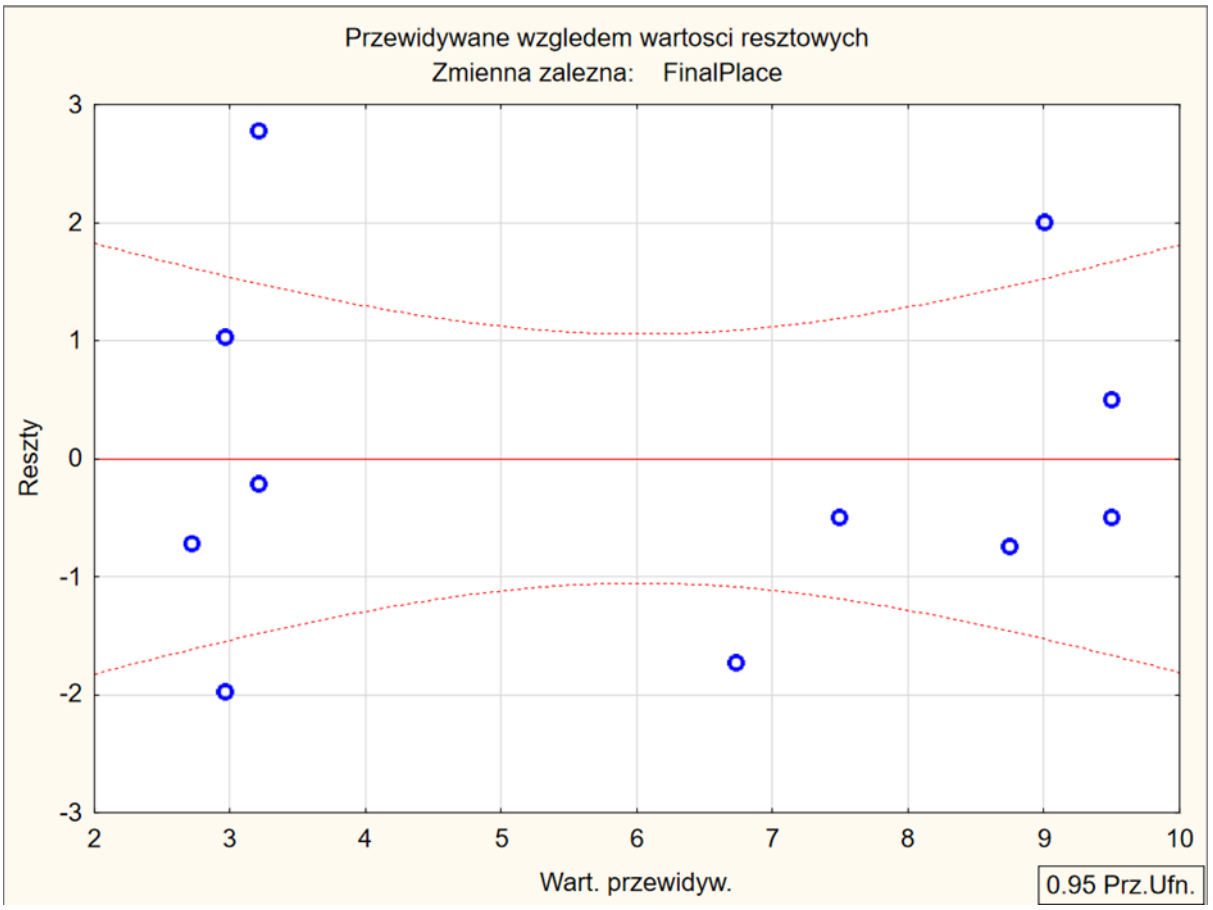
Rysunek 39: Wykres normalności reszt



Źródło: Opracowanie własne

Wykres przewidywania względem wartości resztowych:

Rysunek 40: Wykres przewidywań względem wartości resztowych



Źródło: Opracowanie własne

Na podstawie powyższego wykresu z 95% przedziałem ufności można stwierdzić, że wariacja jest stała, ponieważ większość punktów znajduje się wewnątrz przedziału.

Podsumowanie:

Na podstawie powyższej analizy regresji, przedstawionych danych oraz wykresów, można stwierdzić, że występuje stała zależność pomiędzy potencjałem ataku a końcową pozycją drużyny w tabeli.

Można przyjąć zatem, że drużynom prawdopodobnie bardziej opłaca się inwestować w aspekty ofensywne drużyny, niż defensywne, choć do potwierdzenia takiego wniosku konieczna byłaby analiza krokowa.