

## **Hurtownie danych – Projekt**

Proces tworzenia hurtowni danych powinien być poprzedzony zrozumieniem „potrzeb biznesu” oraz rzeczywistości (dziedziny problemowej) reprezentowanej przez dostępne zasoby danych. Realizacja poniższego zadania ma uzmysłwić występujące problemy w określonym (wybranym) wycinku rzeczywistości, a następnie umożliwić zidentyfikowanie (określenie) potrzeb, celu i możliwości analiz biznesowych, by wspierać procesy decyzyjne (podejmowanie właściwych decyzji biznesowych).

Projekt końcowy powinien zawierać przynajmniej jedną kostkę Analysis Services, dotyczącą danych wybranych i przetworzonych przez studenta przy użyciu Integration Services. Utworzona kostka powinna:

- zawierać przynajmniej 5 wymiarów, w tym co najmniej dwa o strukturze hierarchicznej (np. czas, miejsce, itp)
- posiadać co najmniej 3 miary, w tym min. jedną nieaddytywną
- odpowiadająca jej tabela faktów powinna posiadać co najmniej 10000 rekordów.

### **Projekt – etap I (15.05./18.05.)**

#### **Propozycja tematu**

Proszę przygotować zakres realizacji projektu zgodnie z poniższą specyfikacją oraz przedyskutować propozycję projektu z osobą prowadzącą zajęcia. Poczynione uzgodnienia zarejestrować w formie wniosków.

#### **Zakres opracowania projektu HD**

- Tytuł projektu

Analiza wyników piłkarzy w poszczególnych meczach w trakcie konkursów europejskich

- Charakterystyka dziedziny problemowej, krótki opis obszaru analizy, problemy i potrzeby

Piłka nożna budzi ogromne emocje wśród szerokiej publiczności. Ta popularność przekłada się na wielkie pieniądze, którymi obracają sponsorzy, kluby, piłkarze, fani oraz zakłady bukmacherskie i ich gracze. Wszystkim im zależy na wiedzy o tym jak radzą sobie na boisku poszczególni piłkarze i drużyny, aby móc przewidywać pewne aspekty rozgrywek i możliwe wyniki.

Centrum analizy będą dane o wystąpieniach piłkarzy w meczach w trakcie rozgrywek europejskich. Część danych sięga aż do sezonu dotyczącego 2011 roku. Jednak główne informacje dotyczące występów poszczególnych piłkarzy dotyczą okresu od 2014-07-01 do 2023-04-17. Analizie poddane zostaną informacje dotyczące występów piłkarzy w meczach, występów klubów w meczach, całych meczów, piłkarzy i ich zarobków w czasie, klubów oraz konkursów europejskich.

Sport za kurtyną to już nie tylko ciężkie treningi zawodników, ale praca setek ludzi, w tym analityków. Wyciąganie informacji z danych jest kluczowe w budowaniu dobrze radzących sobie drużyn, przewidywaniu możliwych wyników oraz problemów i ich rozwiązań.

Kluby potrzebują informacji, dzięki którym będą mogły wybierać piłkarzy do swojego składu (lub decydować, którego piłkarza należałoby sprzedać lub posadzić na ławce). Problemem jest wyciągnięcie informacji z poprzednich występów gracza i jego warunków fizycznych oraz osadzenie jego wpływu na wyniki całej drużyny.

Sponsorzy klubów oraz gracze i zakłady bukmacherskie potrzebują informacji o całościowych wynikach klubów. Problemem jest analiza przeszłych występów drużyn w zależności od zarządu, przeciwnika czy graczy, tak aby móc przewidywać potencjalne wyniki całego zespołu.

- Cel przedsięwzięcia – oczekiwania

Stworzenie zestawień, które wykorzystają zgromadzone dane i pozwolą na kompleksową analizę osiągnięć zawodników ze względu na ich przeszłe występy w meczach i warunki fizyczne. Stworzona hurtownia powinna zapewnić strukturę, która pozwoli na powtórzenie takich analiz w przyszłości.

- **Zakres analizy – badane aspekty (min. 10 wielowymiarowych zestawień, które zostaną utworzone po wdrożeniu kostki)**

1. Jak wiek i pozycja graczy wpływają na czas spędzany na boisku? (Średnia liczba minut spędzonych na boisku w zależności od grupy wiekowej i pozycji).
2. Jak sezon i wiek graczy wpływają na czas spędzany na boisku? (Średnia liczba minut spędzonych na boisku w zależności od grupy wiekowej i sezonu).
3. Jak pozycja i wiek graczy wpływają na procentowy udział w golach drużyny? (Średni procentowy udział w golach drużyny w zależności od pozycji i grupy wiekowej).
4. Jak poziom rozgrywek i liczba kibiców wpływają na liczbę goli strzelanych przez graczy? (Średnia liczba goli w zależności od grupy wielkości widowni i poziomu rozgrywek).
5. Jak poziom rozgrywek i liczba kibiców wpływają na liczbę asyst graczy? (Średnia liczba asyst w zależności od grupy wielkości widowni i poziomu rozgrywek).
6. Jak doping wpływa na liczbę goli strzelanych przez graczy?
  - (Średnia liczba goli w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).
  - (Maksymalna liczba goli w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).
7. Jak doping wpływa na kulturę gry?
  - (Średnia liczba żółtych kartek w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).
  - (Średnia liczba czerwonych kartek w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).
8. Jak typ konkursu w zależności od doświadczenia życiowego (wieku) gracza wpływa na kulturę gry?
  - (Średnia liczba żółtych kartek w zależności od typu konkursu i grupy wiekowej graczy).
  - (Średnia liczba czerwonych kartek w zależności od typu konkursu i grupy wiekowej graczy).

Analogiczne zestawienie można zrealizować rozważając zamiast typu konkursu poziom rozgrywek (np. finały), co będzie wymagało utworzenia grup.

9. Jak warunki fizyczne wpływają na wyniki graczy? (Liczba zwyciężonych meczy ze względu na stopę dominującą i wzrost).
10. Jak warunki fizyczne wpływają na wyniki graczy? (Suma zdobytych goli ze względu na stopę dominującą i wzrost).
11. Jak wielkość populacji kraju, w którym rozgrywany jest mecz wpływa na agresję graczy (wyrażoną poprzez żółty lub czerwone kartki)?

- (Średnia liczba żółtych kartek w zależności od grupy wielkości populacji i grupy wiekowej graczy).
- (Średnia liczba czerwonych kartek w zależności od grupy wielkości populacji i grupy wiekowej graczy).
- Źródła danych (lokalizacja, format, dostępność)

#### Wstępna analiza źródeł danych

Lp.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1.	appearances	csv	1166215	89,267	Wystąpienie piłkarza podczas jednego meczu. Rekord zawiera informacje o jego występie i wynikach.
2.	club_games	csv	123096	7,579	Wystąpienie klubu podczas danego meczu. Rekord zawiera informacje o wynikach rywalizacji jednej drużyny.
3.	clubs	csv	411	0,068	Zawiera informacje o klubach biorących udział w rozgrywkach.
4.	competitions	csv	43	0,008	Zawiera informacje o konkursach/ligach, w ramach których rozgrywane są mecze.
5.	games	csv	61548	15,434	Zawiera informacje o jednym meczu na poziomie wyników obydwóch drużyn i informacji specyficznych dla meczu.
6.	player_valuations	csv	421564	22,769	Zawiera informacje o wartości pieniężnej graczy w zależności od czasu.
7.	players	csv	28503	9,217	Zawiera informacje o piłkarzach biorących udział w rozgrywkach.
8.	stadiums	csv	2024	0,146	Dodatkowe źródło danych. Zawiera informacje o stadionach, na których rozgrywane są mecze.

- Profilowanie danych

#### Analiza danych

Plik: appearances				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	appearance_id	Int	Długość 7 – 15 znaków	Klucz główny (jedyne klucze kandydujący), konkatenacja game_id i player_id
2.	game_id	Int	-	Klucz obcy, połączenie z games, gra w której występował piłkarz
3.	player_id	Int	-	Klucz obcy, połączenie z players, piłkarz, którego dotyczy wystąpienie w meczu
4.	player_club_id	Int	-	Klucz obcy, połączenie z clubs, klub do którego należał piłkarz w trakcie rozgrywania spotkania
5.	player_current_club_id	Int	-	Klucz obcy, połączenie z clubs, klub do którego aktualnie należy piłkarz

6.	date	Date	od 2014-07-01 do 2023-04-17 Występują wartości NULL (-1)	Data rozgrywania meczu, w danych źródłowych zamiast wartości NULL została wprowadzona wartość -1. Przy imporcie danych przekształcana przez system na datę 1899-12-30, konieczne będzie przekształcenie danych. Dana zaczerpnięta z games (połączenie przez game_id).
7.	player_name	String	Długość 0 – 31 znaków	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Dana zaczerpnięta z players (połączenie przez player_id).
8.	competition_id	String	Długość 2 – 4 znaki	Klucz obcy, połączenie z competitions. Dana zaczerpnięta z games (połączenie przez game_id).
9.	yellow_cards	Int	Min 0 Max 2 Średnia 0,15 Brak wartości NULL	Liczba żółtych kartek otrzymanych przez gracza w trakcie meczu. 85,52% wartości to 0. Ważne przy analizie kultury gry gracza, zwłaszcza, że dane są kompletne. Wartości mają na tyle mały rozstrzał, że nie będzie konieczności tworzenia grup.
10.	red_cards	Int	Min 0 Max 1 Średnia 0,0036 Brak wartości NULL	Liczba czerwonych kartek otrzymanych przez gracza w trakcie meczu. Aż 99,64% wartości to 0. W sumie zebrano informacje o przyznanych 4218 czerwonych kartek. Ważne przy analizie kultury gry gracza, zwłaszcza, że dane są kompletne. Wartości mają na tyle mały rozstrzał, że nie będzie konieczności tworzenia grup.
11.	goals	Int	Min 0 Max 6 Średnia 0,097 Brak wartości NULL	Liczba goli strzelonych przez gracza w trakcie meczu. Aż 91,36% wartości to 0. Im wyższa liczba goli tym mniejszy jej procentowy udział występowania. Ważne przy analizie wyników gry gracza, zwłaszcza, że dane są kompletne. Wartości mają na tyle mały rozstrzał, że nie będzie konieczności tworzenia grup.
12.	assists	Int	Min 0 Max 6	Liczba asyst przy strzeleniu gola gracza w trakcie meczu. Aż

			Średnia 0,074 Brak wartości NULL	93,17% wartości to 0. Ważne przy analizie wyników gry gracza, zwłaszcza, że dane są kompletne. Wartości mają na tyle mały rozstrzał, że nie będzie konieczności tworzenia grup.
13.	minutes_played	Int	Min 0 Max 120 Średnia 69,77 Brak wartości NULL	Liczba minut spędzonych na boisku przez gracza w trakcie meczu. 54,65% wartości to 90, . kolejne jest 45 z udziałem 3,70%. W sumie jest 121 różnych wartości. Przy niektórych analizach konieczne będzie stworzenie grup długości przebywania na boisku. Ważne przy analizie wpływu na rozgrywkę i doświadczenia gracza, zwłaszcza, że dane są kompletne.

Plik: club_games				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	club_id	Int	-	Klucz obcy, połączenie z clubs, może stworzyć klucz złożony (z atrybutem game_id). Zespół, który rozgrywa mecz.
2.	game_id	Int	-	Klucz obcy, połączenie z games, może stworzyć klucz złożony (z atrybutem club_id). Dla każdego game_id występują dwa rekordy (z perspektywy każdego z dwóch zespołów biorących udział w jednym meczu). Informacja o tym, dla którego meczu dane są informacje o występie zespołu.
3.	own_goals	Int	Min 0 Max 16 Średnia 1,45 Brak wartości NULL	Suma goli strzelonych przez piłkarzy danego zespołu podczas danego meczu. Przydatne przy analizie występu całego zespołu, ale jest to wartość wyliczona (dane są kompletne, bo są wyliczane z kompletnych danych).
4.	own_position	Int	Min 0 Max 21 Występują wartości NULL (-1)	29,33% wartości to NULL, reszta wartości ma równomierny rozkład oscylujący w okolicach 2-4%. Konieczne będzie obrobienie wartości -1 symbolizujących brak danych.
5.	own_manager_name	String	Długość 0 – 33 znaki	Imię managera zespołu, którego wystąpienia dotyczy rekord. Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że

				kodowanie jest odpowiednie przy zestawieniach. Ważne przy analizie wpływu managera na występy drużyny.
6.	opponent_id	Int	-	Klucz obcy, połączenie z clubs. Zespół przeciwko któremu rozgrywany jest mecz.
7.	opponent_goals	Int	Min 0 Max 16 Średnia 1,45 Brak wartości NULL	Suma goli strzelonych przez piłkarzy przeciwnego zespołu podczas danego meczu. Ważne przy analizie występu całego zespołu, ale jest to wartość wyliczona (dane są kompletne, bo są wyliczane z kompletnych danych).
8.	opponent_position	Int	Min 0 Max 21 Występują wartości NULL (-1)	29,33% wartości to NULLe, reszta wartości ma równomierny rozkład oscylujący w okolicach 2-4%. Konieczne będzie obrobienie wartości -1 symbolizujących brak danych.
9.	opponent_manager_name	String	Długość 0 – 33 znaki	Imię managera zespołu, przeciwko któremu gra aktualnie rozpatrywana drużyna. Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Ważne przy analizie wpływu managera na występy drużyn.
10.	hosting	String	Długość 4 znaki 2 wartości {‘Home’, ‘Away’} Brak wartości NULL	Każdy mecz rozgrywany jest przez dwie drużyny na stadionie jednej z nich. Zawiera informacje o tym, czy mecz był grany na stadionie danej drużyny. Występują tylko dwie możliwe wartości z udziałem wystąpień 50%. Ważne przy analizie wpływu dopingu na występy gracza.
11.	is_win	Int	Mix 0 Max 1 Średnia 0,39 Brak wartości NULL	Należy zwrócić uwagę, że wartość ‘1’ oznacza zwycięstwo, a wartość ‘0’ oznacza zarówno porażkę jak i remis. Może być to problematyczne jeżeli chciałoby się rozróżniać remisy i porażki. Z tego wynika średnia wartość i rozkład występowania (60,81% wartości 0 i 39, 19% wartości 1). Ważne przy analizie wyników graczy, ale może być wyliczane na podstawie liczby goli drużyn biorących udział w meczu.

Plik: clubs				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	club_id	Int	-	Klucz główny
2.	club_code	String	Długość 4 – 31 znaków	Kod jednoznacznie określa klub. Użycie niepraktyczne, w zestawieniach zdecydowanie łatwiej

			Wartości unikatowe	będzie prowadzić analizy, gdy kluby będą reprezentowane przez nazwy
3.	name	String	Długość 4 – 31 znaków Wartości unikatowe	Nazwa jednoznacznie określa klub. Ważne w analizach dotyczących konkretnych klubów.
4.	domestic_competition_id	String	Długość 2 – 4 znaki Brak wartości NULL	Klucz obcy, połączenie z competitions. Mówi o konkursie, w którym uczestniczy klub. Ważne przy analizach dla konkretnych konkursów.
5.	total_market_value	Float	Max 1,1 Min 585 Średnia 34,13 Występują wartości NULL	Suma wartości rynkowych wszystkich członków klubu, 69,58% wartości to NULL. Lepiej skorzystać z wartości dla danych graczy, zwłaszcza, że te są zmienne w czasie. W celu analizy należałoby stworzyć podział na grupy.
6.	squad_size	Int	Min 0 Max 40 Średnia 24,91	Liczba członków zespołu, może być przydatne w analizach (np. czas gry graczy podczas meczy może zależeć od liczby zawodników zespołu).
7.	average_age	Float	Min 20 Max 29.6 Średnia 25,57 Występują wartości NULL	Jedynie 7,06% wartości to NULL. Atrybut zmienny w czasie, ale ważny podczas analiz (np. gracz może sobie lepiej radzić w zespole doświadczonych graczy).
8.	foreigners_number	Int	Min 0 Max 29 Średnia 11,27 Brak wartości NULL	Liczba piłkarzy z zagranicy w zespole. Ważne w analizie (może mieć wpływ na wyniki zespołu), zwłaszcza, że dane są kompletne. Duży przedział wartości, trzeba rozważyć wprowadzenie grup.
9.	foreigners_percentage	Float	Min 3.3 Max 100 Występują wartości NULL	Procent piłkarzy z zagranicy w zespole. Ważne w analizie (może mieć wpływ na wyniki zespołu), zwłaszcza że dane są dość kompletne. Jedynie 9,73% to wartości NULL.
10.	national_team_players	Int	Min 0 Max 23 Średnia 4,92 Brak wartości NULL	Liczba piłkarzy będących w reprezentacji kraju w zespole. Ważne w analizie (może mieć wpływ na wyniki zespołu), zwłaszcza że dane są kompletne. Duży przedział wartości, trzeba rozważyć wprowadzenie grup.
11.	stadium_name	String	Długość 4 – 54 znaki	Nazwa stadionu klubu, nie zauważono wartości analitycznej.
12.	stadium_seats	Int	Min 1600 Max 99354 Średnia 24935,81	Liczba miejsc na stadionie. Ważne w analizie, zwłaszcza, że dane są kompletne. Duży przedział wartości, konieczne wprowadzenie grup.

			Brak wartości NULL	
13.	net_transfer_record	String	Od 3 do 11 znaków Występują wartości NULL	Rekordowy balans zysków i kosztów poniesionych w wyniku transferu zawodników. Aż 77,85% wartości to NULL, co sprawia, że nie jest to informacja przydatna w analizie.
14.	coach_name	String	Długość 7 – 25 znaków Występują wartości NULL	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Informacja zmienna czasie, w tej postaci nieprzydatna do analizy. 58,63% wartości to NULL. Gdyby coache byli przypisani do poszczególnych meczy teamów byłaby to cenna informacja do analizy.
15.	url	String	Długość 59 – 87 znaków	Adres strony internetowej klubu. Nieprzydatny w analizie.

Plik: competitions				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	competition_id	String	Długość 2 – 4 znaki	Klucz główny
2.	competition_code	String	Długość 6 – 43 znaków	Kod zawodów, wartości niemalże unikatowe (jeden duplikat). Użycie niepraktyczne, w zestawieniach zdecydowanie łatwiej będzie prowadzić analizy, gdy konkursy będą reprezentowane przez nazwy
3.	name	String	Długość 6 – 43 znaków	Nazwa zawodów. Wartości niemalże unikatowe (jeden duplikat), niemalże jednoznacznie określające konkurs. Ważne w analizach dotyczących konkretnych konkursów.
4.	type	String	Długość 5 – 17 znaków, 4 wartości {‘domestic_cup’, ‘domestic_league’, ‘international_cup’, ‘other’} Brak wartości NULL	Typ konkursu. Ważne przy analizie ze względu na typ konkursu, zwłaszcza, że dane są kompletne. Uogólnia informację o podtypie, możliwość utworzenia hierarchii.
5.	sub_type	String	Długość 10 – 40 znaków, 11 wartości, Brak wartości NULL	Podtyp konkursu. Ważne przy analizie ze względu na podtypkonkursu, zwłaszcza, że dane są kompletne. Uzupełnia informację o typie, możliwość utworzenia hierarchii.
6.	country_id	Int	14 unikatowych wartości,	16% wartości to NULL. Atrybut nie odnosi się do rekordów żadnej innej tabeli, aby był użyteczny konieczne



			Występują wartości NULL (-1)	byłoby wydzielenie tabeli państw. Konieczne będzie obrobienie wartości -1 symbolizujących brak danych.
7.	country_name	String	Długość 0 – 11 znaków, 14 unikatowych wartości, Występują wartości NULL	16% wartości to NULL, atrybut ważny w zestawieniach ze względu na kraj konkursu.
8.	country_latitude	Float	Min 38.95976 Max 64.68631	Szerokość geograficzna, ważne przy tworzeniu zestawień opartych o mapy. Można rozważyć grupowanie wartości ze względu na wiele unikatowych wartości.
9.	country_longitude	Float	Min -8.135352 Max 97.74531	Długość geograficzna, ważne przy tworzeniu zestawień opartych o mapy. Można rozważyć grupowanie wartości ze względu na wiele unikatowych wartości.
10.	domestic_league_code	String	Długość 0 – 4 znaki, 14 unikatowych wartości	16% to wartości NULL, może być przydatne w analizie, rozróżnia różne ligi w danym kraju.
11.	confederation	String	Długość 6 znaków, 1 wartość {‘europa’ } Brak wartości NULL	Jednostka, pod którą podlega konkurs. Atrybut w żaden sposób nie różnicuje rekordów, więc nie jest przydatny w analizie
12.	url	String	Długość 64 – 107 znaki	Adres strony internetowej konkursu. Nieprzydatny w analizie.

Plik: games				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	game_id	Int	-	Klucz główny
2.	competition_id	String	Długość 2 –4 znaki, Brak wartości NULL	Klucz obcy, połączenie z competitions
3.	competition_type	String	Długość 5 – 17 znaków, 4 wartości {‘domestic_cup’, ‘domestic_league’, ‘international_cup’, ‘other’} Brak wartości NULL	Informacja zaczerpnięta z competitions. Typ konkursu. Ważne przy analizie ze względu na typ konkursu, zwłaszcza, że dane są kompletne.
4.	season	Int	od 2011 do 2022,	Konieczne będzie obrobienie wartości -1 symbolizujących brak danych. Dana istotna w tworzeniu wymiaru czasu.

			Występują wartości NULL (-1)	
5.	round	String	Długość 1 – 28 znaków, 115 unikatowych wartości, Brak wartości NULL	Runda czy etap rozgrywania zawodów. Ze względu na dużo unikatowych wartości potrzebne byłoby utworzenie grup lub analiza jedynie konkretnych rund.
6.	date	Date Time	od 2012-07-03 do 2023-05-06, Nieliczne wartości NULL	Nieliczne puste wartości, w hurtowni bardziej się przyda ta wartość w tabeli faktów
7.	home_club_id	Int	-	Klucz obcy, połączenie z clubs
8.	away_club_id	Int	-	Klucz obcy, połączenie z clubs
9.	home_club_goals	Int	Min 0 Max 15 Średnia 1,6, Brak wartości NULL	Informacja zaczerpnięta z club_games (a tam jest to suma liczby goli piłkarzy z klubu z pliku appearances)
10.	away_club_goals	Int	Min 0 Max 16 Średnia 1,3, Brak wartości NULL	Informacja zaczerpnięta z club_games (a tam jest to suma liczby goli piłkarzy z klubu z pliku appearances)
11.	aggregate	String	Brak wartości NULL	Informacja o tym czy jest częścią większego agregatu czy pojedynczym meczem. Nie zauważono przydatności w zestawieniach. Wymagana zmiana formatu, bo atrybut odczytywany jest jako data (choć nią nie jest).
12.	home_club_position	Int	Min 1 Max 21, Występują wartości NULL (-1)	Konieczne będzie obrobienie wartości -1 symbolizujących brak danych. Przydatny w analizach ze względu na pozycję klubu w rankingu.
13.	away_club_position	Int	Min 1 Max 21, Występują wartości NULL (-1)	Konieczne będzie obrobienie wartości -1 symbolizujących brak danych. Przydatny w analizach ze względu na pozycję klubu w rankingu.
14.	club_home_name	String	Długość 4 – 31 znaków, Występują wartości NULL	17% wartości NULL, nazwa klubu pochodząca z clubs, istotne w zestawieniach ze względu na club
15.	club_away_name	String	Długość 4 – 31 znaków, Występują wartości NULL	15% wartości NULL, nazwa klubu pochodząca z clubs, istotne w zestawieniach ze względu na club
16.	home_club_manager_name	String	Długość 4 – 29 znaków,	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne

			Nieliczne wartości NULL	upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Około 1% wartości to NULL, ważne przy analizie wpływu managera na wyniki.
17.	away_club_manager_name	String	Długość 3 – 33 znaków, Nieliczne wartości NULL	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Około 1% wartości to NULL, ważne przy analizie wpływu managera na wyniki.
18.	stadium	String	Długość 4 – 55 znaków	2129 unikatowych wartości, istotne przy analizie wyników na danym stadionie.
19.	attendance	Float	Min 0 Max 99354, Średnia 15356, Brak wartości NULL	Liczba widzów gry. Bardzo dużo unikatowych wartości, do analizy należałoby wykonać grupowanie. Przydatne w analizie wpływu publiki na występ graczy i drużyn.
20.	referee	String	Długość 7 – 42 znaków, Nieliczne wartości NULL	Mnie niż 1% wartości NULL. Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Ważne w analizie ze względu na wpływ sędziego na wyniki meczu i przyznawane kartki.
21.	url	String	Długość 71 – 115 znaków	Adres strony internetowej meczu. Nieprzydatny w analizie.

Plik: player_valuations				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	date	Date	od 2003-12-09 do 2023-12-19, Brak wartości NULL	Ważne w analizie ze względu na czas danej wartości gracza, informacje z długiego okresu
2.	datetime	Date	od 2003-12-09 do 2023-12-19, Brak wartości NULL	Ważne w analizie ze względu na czas danej wartości gracza, informacje z długiego okresu
3.	dateweek	Date	od 2003-12-08 do 2023-12-18, Brak wartości NULL	Ważne w analizie ze względu na czas danej wartości gracza, informacje z długiego okresu
4.	player_id	Int	-	Klucz obcy, połączenie z players
5.	current_club_id	Int	-	Klucz obcy, połączenie z clubs

6.	market_value_in_eur	Int	Min 10000 Max 200000000 Średnia 2359188, Brak wartości NULL	Ważne w analizie, kluczowa informacja z pliku, kompletna dana (choć niekoniecznie player_valuations musi zawierać pełnię danych dotyczących każdego gracza)
7.	player_club_domestic_competition_id	String	14 unikatowych wartości	Klucz obcy, połączenie z competitions. Ważne w analizie konkursu, w którym gracz występował przy danej pensji.

Plik: player				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	player_id	Int	-	Klucz główny
2.	name	String	Długość 2 – 32 znaków, Brak wartości NULL	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. Konkatenacja imienia i nazwiska, ważne w zestawieniach ze względu na gracza (czytelność zestawienia).
3.	current_club_id	Int	-	Klucz obcy, połączenie z clubs, aktualny klub gracza, zmienne z upływem czasu. Mało przydatne w zestawieniach, bardziej czytelne będzie wykorzystanie nazwy klubu.
4.	current_club_name	String	Długość 4 – 31 znaków, Brak wartości NULL	Klucz obcy, połączenie z clubs, aktualny klub gracza, zmienne z upływem czasu. Ważne w zestawieniach.
5.	country_of_citizenship	String	Długość 4 – 24 znaków, 178 unikatowych wartości, Tylko jedna wartość NULL	Kraj obywatelstwa, ważne w analizach, zwłaszcza, że dane są niemalże kompletne, ale mogą ulec zmianie.
6.	country_of_birth	String	Długość 3 – 30 znaków, 181 unikatowych wartości, Występują wartości NULL	Kraj urodzenia, ważne w analizach pochodzenia graczy i nie mogą ulec zmianie. Wartości NULL stanowią tylko 6,6%. Może utworzyć hierarchę z miastem urodzenia, jednak przy tej liczbie unikatowych wartości warto rozważyć wprowadzenie grup.
7.	city_of_birth	String	Długość 5 – 52 znaków, 8028 unikatowych wartości, Występują wartości NULL	Miasto urodzenia, dane nie mogą ulec zmianie. Wartości NULL stanowią tylko 6,5%. Może utworzyć hierarchę z krajem urodzenia, jednak przy tak dużej liczbie unikatowych wartości konieczne będzie wprowadzenie grup (albo właśnie

				rozpatrywanie po kraju wyższym w hierarchii).
8.	date_of_birth	Date	od 31-07-1938 do 18-01-2007, Nieliczne wartości NULL	Jedynie 0,12% wartości NULL, bardzo ważne w analizach. W przypadku analizy meczy ważne będzie wyliczanie wieku w danym momencie.
9.	position	String	Długość 6 – 10, znaków, 4 wartości {'Attack', 'Defender', 'MidField', 'Goalkeeper'}, Brak wartości NULL	Ważne w analizie ze względu na konkretną pozycję, zwłaszcza, że dane są kompletne.
10.	sub_position	String	Długość 9 – 18 znaków, 12 różnych wartości, Występują wartości NULL	12% wartości NULL. Możliwe utworzenie hierarchii z position, ważne w analizach.
11.	foot	String	Długość 4 – 5 znaków, 3 wartości {'Both', 'Right', 'Left'}, Występują wartości NULL	7,6% wartości NULL, ważne w analizach.
12.	height_in_cm	Int	Min 159 Max 206, Występują nieliczne wartości NULL (0)	Konieczne będzie obrobienie wartości 0 symbolizujących brak danych. Jest to jednak jedynie około 6% wartości. Ważne przy analizach.
13.	market_value_in_eur	Int	Min 10000 Max 180000000 Średnia 2099269, Występują wartości NULL	31% wartości to NULL, jest to aktualna wartość rynkowa gracza, co może nie mieć przełożenia na jego wartość z konkretnych występów w meczach. Dana zmienna w czasie. W tej postaci raczej nie warta analizy.
14.	highest_market_value_in_eur	Int	Min 10000 Max 200000000 Średnia 3571998, Występują wartości NULL	3,7% wartości to NULL, jest to najwyższa wartość rynkowa gracza do tej pory, co może nie mieć przełożenia na jego wartość z konkretnych występów w meczach. Dana może ulec zmianie w czasie. W tej postaci może być analizowana, ale raczej nie jako priorytet.
15.	agent_name	String	Długość 2 – 51 znaków, Występują wartości NULL	39% wartości to NULL, jest to aktualny agent gracza, co może nie mieć przełożenia na jego konkretne wystąpienia w meczach z przeszłości, bo dana jest zmienna w

				czasie. W tej postaci może być analizowana, ale raczej nie jako priorytet. Trzeba pamiętać o brakach informacji o zmianie w czasie.
16.	contract_expiration_date	Date	od 2023-01-01 do 2032-06-30, Występują wartości NULL	39% wartości to NULL, dana jest zmienna w czasie analizy
17.	current_club_domestic_competition_id	String	Długość 2 – 4 znaki, Brak wartości NULL	Zmienna wartość zależna od obecnego klubu, niepotrzebna w analizie
18.	first_name	String	Długość 2 – 18 znaków, Występują wartości NULL	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. W zestawieniach ważniejszy jest atrybut name (konkatenacja imienia i nazwiska).
19.	last_name	String	Długość 2 – 22 znaków, Brak wartości NULL	Pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Konieczne upewnienie się, że kodowanie jest odpowiednie przy zestawieniach. W zestawieniach ważniejszy jest atrybut name (konkatenacja imienia i nazwiska).
20.	player_code	String	Długość 4 – 35 znaków, Tylko jedna wartość NULL	W zestawieniach ważniejszy jest atrybut name (konkatenacja imienia i nazwiska).
21.	image_url	String	Długość 71 – 91 znaków, Brak wartości NULL	Adres zdjęcia gracza. Nieprzydatny w analizie.
22.	last_season	Int	od 2012 do 2022, Brak wartości NULL	Ważny w analizie
23.	url	String	Długość 55 – 88 znaków, Brak wartości NULL	Adres strony internetowej o gracz. Nieprzydatny w analizie.

Plik: stadiums			
Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
Confederation	String	Długość 3 – 8 znaków, Brak wartości NULL	Konfederacja, informacja nieprzydatna do analizy kiedy rozpatrujemy tylko mecze europejskie
Stadium	String	Długość 4 – 40 znaków, Brak wartości NULL	Pozwoli połączyć dane z bazowym źródłem i rozszerzyć je o informacje o mieście. Nie wszystkie nazwy są w pełni zgodne z bazowym źródłem

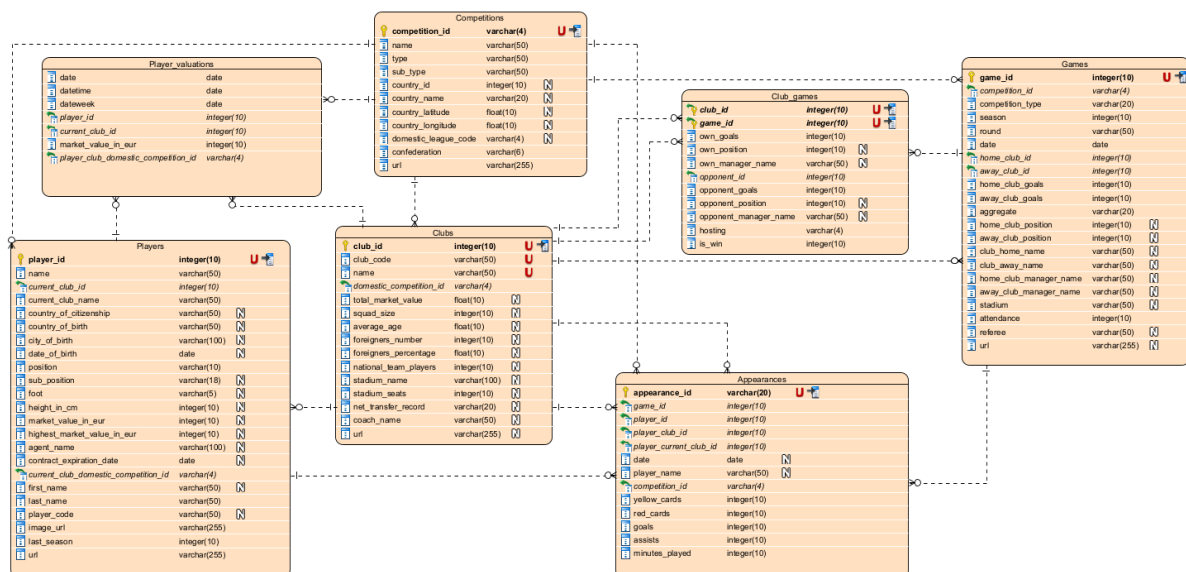
			danych, konieczne będzie wykorzystanie fuzzy lookup.
City	String	Długość 3 – 25 znaków, Brak wartości NULL	Miasto, w którym znajduje się stadion. Kluczowy powód dołączania zbioru danych. Z założenia miał stanowić bazę pod dołączenie informacji o pogodzie, jednak z pomysłu zrezygnowano co zostało opisane we wnioskach.
HomeTeams	String	Długość 1 – 51 znaków, Brak wartości NULL	Informacja o klubach, które są rezydentami stadionu. Nieprzydatne w analizie.
Capacity	String	Min 244 Max 153000 Średnia 22905	Liczba osób, które mogą zostać pomieszczone na stadionie. Priorytetem będzie informacja ze zbioru bazowego, która jest pełna.
Country	String	Długość 4 – 24 znaków, Brak wartości NULL	Kraj, w którym znajduje się stadion. Może stanowić ciekawe miejsce do analizy, chociaż jest to
IOC	String	Długość 5 – 52 znaków, Brak wartości NULL	Nieprzydatne w analizie.
Population	int	Min 32194 Max 1403500365 Średnia 99649306, Brak wartości NULL	Populacja państwa, informacje ważna w analizie, ale należy ją podzielić na grupy ze względu na bardzo dużą liczbę możliwych wartości.

#### Ocena przydatności danych w pliku do tworzenia hurtowni danych

Lp.	Plik	Ocena jakości danych
1.	appearances	Zawiera informacje o pojedynczych wydarzeniach (wystęпах gracza w meczu) znakowanych czasowo. Bardzo dobry kandydat na tabelę faktów, gdyż przechowuje jednostkowe niezagregowane informacje, które warto poddawać analizie. Zawiera informacje, które mogą mieć znaczenie strategiczne (np. liczba goli czy asyst gracza w meczu).
2.	club_games	Zawiera informacje o pojedynczych meczach z perspektywy całej drużyny (czyli po 2 rekordy na każdy mecz, dla perspektywy każdej z drużyn). Zawiera ważne dane do analizy z perspektywy całego meczu. Zawiera wiele unikatowych w skali bazy informacji (np. o drużynie właściciela spotkania i przyjezdnej), ale ma też wygodnie wyliczone informacje, które trzeba byłoby zbierać na podstawie appearances (sumaryczna liczba goli i wynik spotkania). Zawiera ważną informację o klubie aktualną w czasie, managera.
3.	clubs	Zawiera informacje o klubach. Zawiera wiele informacji, które ulegają zmianie z biegiem czasu (np. dotyczące składu zespołu). Należy dokonywać bardzo ostrożnej analizy ze względu na te atrybuty, chociaż są one istotne.

4.	competitions	Istotny dla zestawień dotyczących konkursów. Wykorzystanie zależy od kierunku obranej analizy. Może stanowić bazę pod stworzenie regionów i nanoszenie danych na mapę podczas zestawień.
5.	games	Zawiera informacje o pojedynczych meczach łącząc perspektywę oby drużyn (1 rekord na mecz). Zawiera ważne dane do analizy z perspektywy całego meczu. Zawiera wiele unikatowych w skali bazy informacji (np. o stadionie, na którym odbywało się spotkanie i widowni czy poziomie rozgrywek), ale ma też informacje powtórzone z club_games. Ważne na wielu poziomach analizy (np. ze względu na wpływ poziomu rozgrywek na formę zawodnika czy znaczenie dopingu w jego wynikach).
6.	player_valuations	Istotny plik do zestawień dotyczących płac zawodników. Dane są aktualne dla danego przedziału czasowego, należy je podpiąć pod konkretne mecze, aby móc np. analizować wpływ wartości na zawodnika na jego osiągi w spotkaniach.
7.	players	Bardzo istotny plik do zestawień. Nie wszystkie dane mogą zostać w pełni wykorzystane, gdyż nie są aktualne na przestrzeni poszczególnych występów graczy w meczach. Należy uważać na atrybuty brane pod uwagę i skupić się na tych, które faktycznie powinny zostać wykorzystane.
8.	stadiums	Zewnętrzne źródło danych, które umożliwia stworzenie wymiaru miejsca meczu.

- Definicja typów encji/klas (wraz z własnościami) oraz związków pomiędzy nimi dla bazy źródłowej.



- Propozycja faktów, wymiarów, hierarchii, miar (w tym nieaddytywnych)

#### Fakty:

- wystąpienia piłkarzy w meczach

#### Miary:

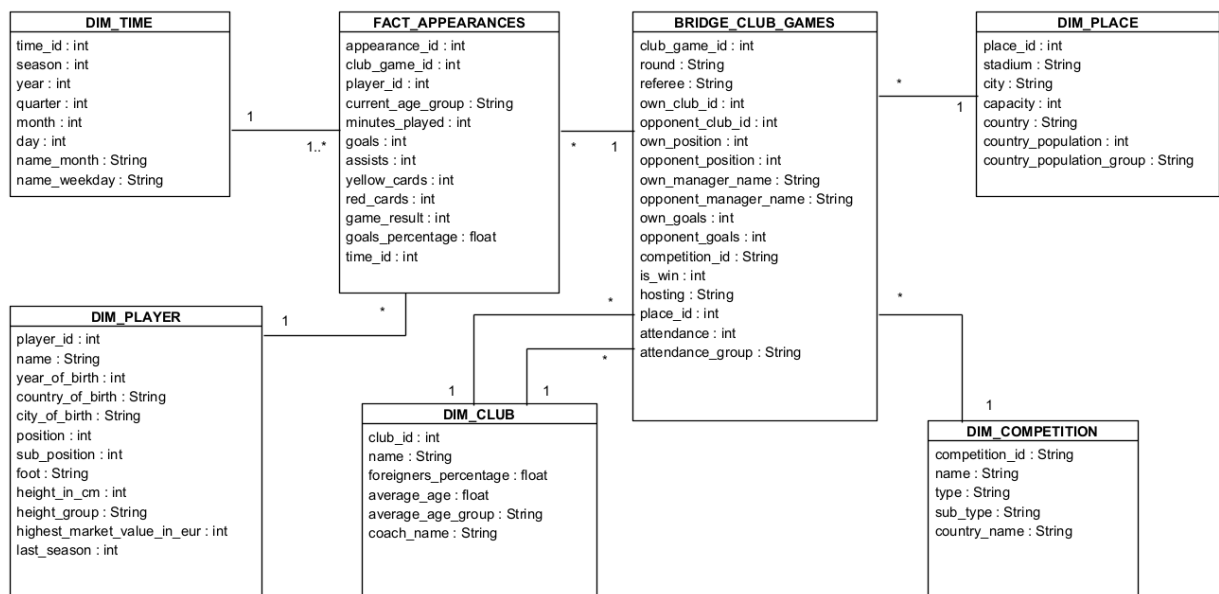
- Średnia liczba zdobytych goli
- Średnia liczba zdobytych asyst



- Suma goli
- Suma asyst
- Średnia liczba minut na boisku
- Maksymalna liczba goli
- Suma żółtych kartek
- Suma czerwonych kartek
- Liczba rozegranych meczy
- Liczba wygranych meczy
- Maksymalny procentowy udział w liczbie goli drużyny w meczu
- Liczba przegranych meczy

### Wymiary:

- Czas (hierarchia: rok, sezon, kwartał, miesiąc, dzień)
  - Piłkarz (hierarchia: pozycja, podpozycja, hierarchia: kraj urodzenia, miasto urodzenia)
  - Miejsce rozgrywania meczu (hierarchia: kraj, miasto)
  - Konkurs
  - Klub
- Diagram klas – model danych hurtowni utworzony na podstawie danych zgromadzonych w plikach.



### Wnioski:

Wybrane dane są już częściowo zdenormalizowane. Następuje przez to w plikach źródłowych pewna redundancja danych. Przy tworzeniu hurtowni należy wziąć to pod uwagę i odpowiednio dobrać model.

Część danych w plikach źródłowych nie jest aktualna dla wszystkich swoich powiązań. Może to prowadzić do nieprawdziwych wniosków podczas analizy, np. podczas starszych meczy piłkarza analizowana będzie jego aktualna płaca. Należy unikać analizy ze względu na te atrybuty albo upewnić

się, że były one aktualne podczas wydarzenia, co i tak sprowadzałoby się do posiadania rozszerzonych danych.

W wielu miejscach w danych, zwłaszcza w imionach i nazwiskach piłkarzy oraz innych osób zaangażowanych w piłkę nożną, pojawiają się znaki, które nie są rozpoznawane przez niektóre typy kodowania. Przy tworzeniu zestawień konieczne upewnienie się, że kodowanie jest poprawione i dane są czytelne. Jeżeli ta kwestia zostanie w przyszłości pominięta, dane uzyskane w zestawieniach mogą być pozbawione wartości analitycznej ze względu na brak czytelności.

W plikach źródłowych zawarte jest dość mało aspektów rozgrywki, przez co pole analizy jest dość ograniczone. Poza informacją o asystach, golach, czasie gry na boisku i przyznanych kartkach brakuje chociażby informacji o obronionych golach w przypadku piłkarzy, nieudanych podejściach na bramkę i wielu innych czynnikach, które mają wpływ na jakość wystąpienia gracza. Można by także pokusić się o bardziej złożone aspekty, jak liczba podań do gracza albo czas posiadania piłki drużyn lub zawodników.

Niektóre atrybuty w bazie mogą być trudne w analizie, ze względu na nietypowe wartości. Np. miasta urodzenia niektórych graczy zawierają w sobie po przecinku prowincję/region/stan, co może być mylące przy łączeniu np. z zewnętrznym zbiorem danych.

Przy wyborze atrybutów do analizy istotne jest wzięcie pod uwagę jak wiele wśród nich to wartości NULL i rozpoznanie przyczyny występowania tych wartości. Jeżeli są to informacje niepełne, których nie zebrano, analizy mogą być przekłamanie ze względu na wybiórcze dobranie danych. Jeżeli jednak te dane po prostu nie dotyczą niektórych rekordów, analiza byłaby już możliwa.

Hurtownie przechowują ogromne ilości danych, dlatego kluczowe jest profilowanie i poznanie rozmiaru przechowywanych atrybutów. Pozwoli to na nałożenie ograniczeń na rozmiar, które przełożą się na mniejsze wykorzystanie pamięci.

Bardzo ciężkie okazało się dobranie dodatkowych danych. Rozważano początkowo dodanie informacji o pogodzie. Z tego względu specjalnie szukano zbiorów danych o położeniu stadionów. Zbiór bezpośrednio zawierający informacje o położeniu geograficznym stadionów był jednak mocno okrojony. Później znaleziono zbiór zawierający dane o stadionach przedstawiony w tym raporcie jako zewnętrzne źródło danych. Nawet jednak z takimi informacjami pozyskanie danych o pogodzie stanowiło ogromne wyzwanie ze względu na konieczność wykorzystania API. Wszystkie znalezione propozycje miały jednak duże ograniczenia na liczbę możliwych bezpłatnych zapytań, które nie pozwoliłyby zebrać informacji dla wszystkich wydarzeń z tabeli faktów.

Istotne jest również grupowanie atrybutów, które przyjmują wiele unikatowych wartości. Bez tego zestawienia będą zbyt nieczytelne, a przez to nieprzydatne do analizy.

## Projekt – etap II (29.05./1.06.)

### Proces ETL

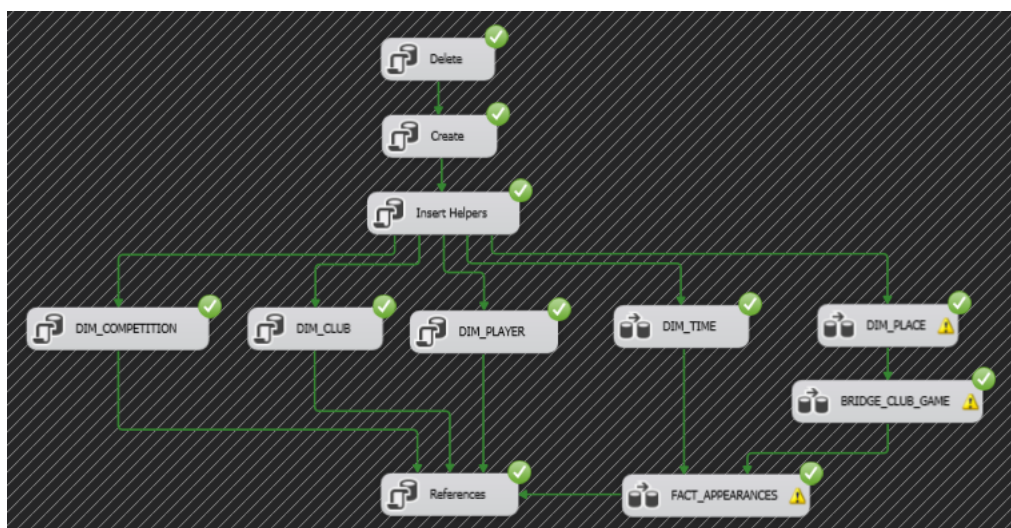
1. Utworzone tabele w poprzednim punkcie wypełnić danymi zgodnie z ustalonymi założeniami projektowymi wykorzystując zapytania SQL lub inne narzędzia dostępne w Integration Services.

Przy ocenie będą brane następujące elementy pakietu(ów):

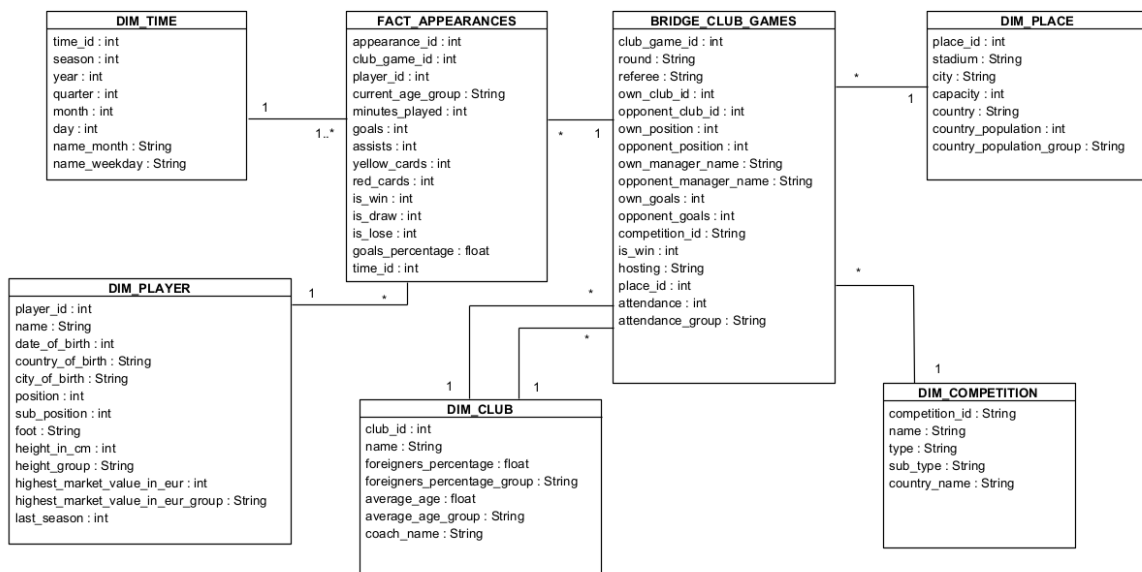
- właściwa struktura procesu ETL (odpowiednie rozbięcie procesu ETL na zadania/pakiety, dobrze dobrane nazwy poszczególnych zadań, wprowadzona automatyzacja, obsługa błędów, itp.)
- stabilność i prawidłowe, bezbłędne wykonanie
- złożoność przeprowadzonych operacji. Przykładowo, jeżeli dane źródłowe już są w pełni zdenormalizowane proszę nie spodziewać się maksymalnej liczby punktów za ten element
- dokumentacja powinna zawierać krótki opis dotyczący każdego zadania, które pozwoli zorientować się, jaki jest jego cel (np. zadanie Z kopiuje dane z tabeli X i Y do tabeli T dokonując denormalizacji) oraz mapę logiczną procesu ETL.

W tworzeniu procesu ETL wykorzystywane były dane z dwóch źródeł, Football Data from Transfermarkt oraz Stadiums. Początkowo występowały problemy z kodowaniem znaków dla plików źródłowych. Wszystkie pliki ze źródła Football Data from Transfermarkt zapisane były w formacie UTF8. Domyślnie SQL Server tworzy bazy o kodowaniu SQL\_Latin1\_General\_CP1\_CI\_AS, dla którego niemożliwe było prawidłowe odczytanie części znaków nawet pomimo ustawienia odpowiedniego kodowania dla tabel i ich kolumn. Konieczne było utworzenie nowej bazy o kodowaniu Latin1\_General\_100\_CS\_AS\_KS\_WS\_SC\_UTF8. Dzięki temu możliwe było prawidłowe odczytywanie znaków o kodowaniu UTF8. Plik Stadiums już przy pobraniu ze strony źródłowej miał nieprawidłowo ustawione kodowanie i wiele znaków było nieprawidłowo rozpoznawane. Konieczne było ręczne zmienienie kodowania i zapisanie pliku ze zmienionym rozszerzeniem.

Dla każdego zadania dodana została obsługa błędów. W przypadku błędu pojawia się okienko z informacją o error code oraz error description.



Utworzony proces ETL



Uzupełniony diagram klas – model danych hurtowni

### Zadanie Delete:

Usunięcie tabel DIM\_COMPETITION, DIM\_CLUB, DIM\_PLAYER, DIM\_TIME, DIM\_PLACE, BRIDGE\_CLUB\_GAME oraz tabel pomocniczych Months i Weekdays o ile takie istnieją w bazie.

### Zadanie Create:

Utworzenie tabel DIM\_COMPETITION, DIM\_CLUB, DIM\_PLAYER, DIM\_TIME, DIM\_PLACE, BRIDGE\_CLUB\_GAME oraz tabel pomocniczych Months i Weekdays.

### Zadanie Insert Helpers:

Wypełnienie danymi tabel pomocniczych Months i Weekdays. Months wypełniana jest numerami miesięcy i odpowiadającymi im nazwami słownymi, Weekdays wypełniana jest numerami dni tygodnia i odpowiadającymi im nazwami słownymi.

### Zadanie DIM\_COMPETITION:

SQL Task, wypełnia danymi tabelę DIM\_COMPETITION. Wszystkie informacje pobiera z tabeli competitions ze źródła 1 i w żaden sposób ich nie przekształca.

### Zadanie DIM\_CLUB:

SQL Task, wypełnia danymi tabelę DIM\_CLUBS. Wszystkie informacje pobiera z tabeli clubs ze źródła 1. Dla atrybutu average\_age dla na pewno nieprawidłowej wartości średniego wieku wynoszącej 0 przypisuje NULL. Wprowadza dwie nowe kolumny, foreigners\_percentage\_group oraz average\_age\_group.

foreigners\_percentage\_group rozróżnia 4 grupy udziału obcokrajowców w składzie drużyny: 0 – 25%, 25 – 50%, 50 – 75%, 75 – 100% oraz wartość NULL w przypadku, gdy nie zdefiniowano atrybutu foreigners\_percentage.

average\_age\_group rozróżnia 5 grup wiekowych 16 – 20, 20 – 23, 23 – 27, 27 – 30 oraz 30+. W przypadku, gdy nie zdefiniowano średniego wieku lub średni wiek to nieprawidłowa wartość 0 przypisywana jest wartość NULL.

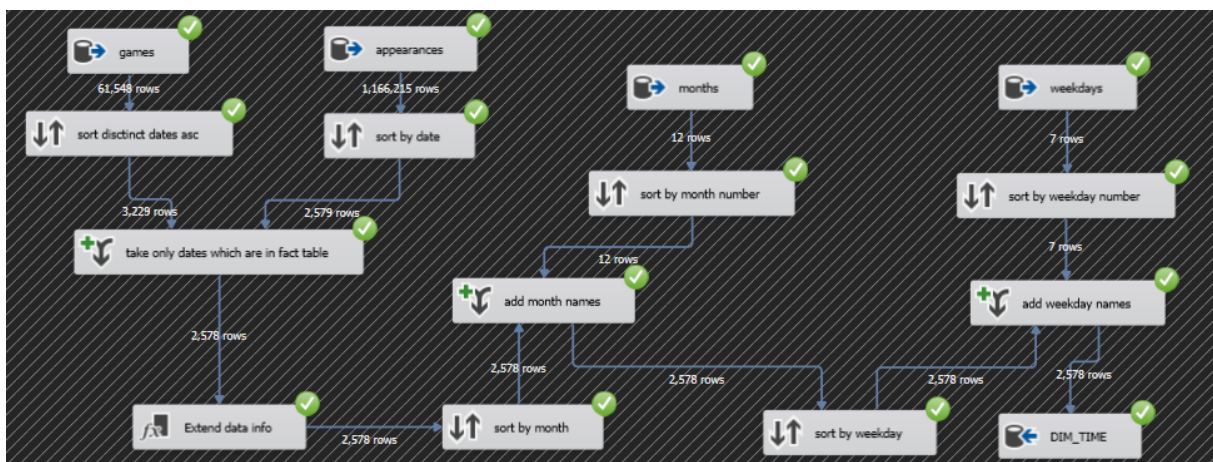
### Zadanie DIM\_PLAYER:

SQL Task, wypełnia danymi tabelę DIM\_PLAYERS. Wszystkie informacje pobiera z tabeli players ze źródła 1. Dla atrybutów height\_in\_cm i highest\_market\_value\_in\_eur dla na pewno nieprawidłowych wartości wynoszących 0 przypisuje NULL. Wprowadza dwie nowe kolumny, height\_group oraz highest\_market\_value\_in\_eur\_group.

height\_group rozróżnia 7 grup wzrostu: 150 – 160 cm, 160 – 170 cm, 170 – 180 cm, 180 – 190 cm, 190 – 200 cm, 200+ cm oraz 'Inne'. W przypadku, gdy nie zdefiniowano wzrostu lub wzrost to nieprawidłowa wartość 0 przypisywana jest wartość NULL.

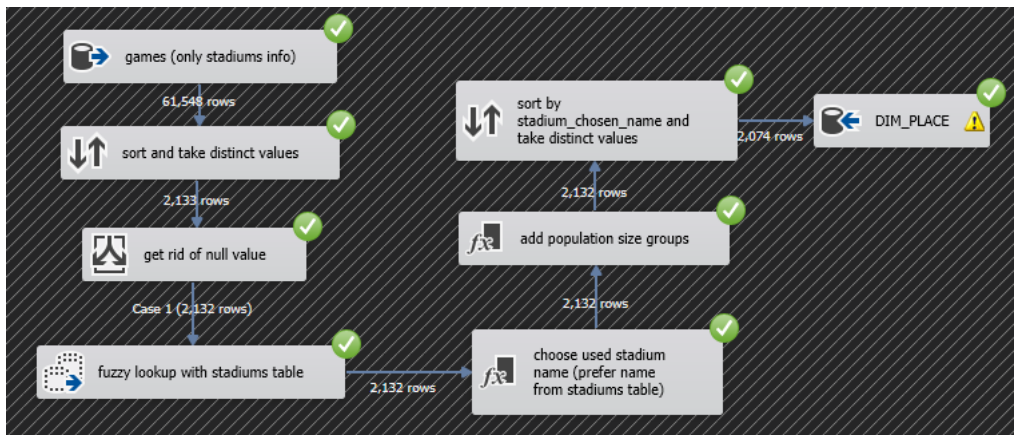
highest\_market\_value\_in\_eur\_group rozróżnia 7 grup wartości 0 – 1mln, 1 – 5 mln, 5 – 10 mln, 10– 20 mln, 20 – 50 mln, 50+ mln. W przypadku, gdy nie zdefiniowano wartości lub wartość to nieprawidłowa wartość 0 przypisywana jest wartość NULL.

### Zadanie DIM\_TIME:



Konieczne jest pobieranie daty z games. Sezon rozgrywek nie jest możliwy do wyliczenia, a zdefiniowano go w tabeli games, dlatego połączono tabele games i appearances. Konieczność połączenia wynika również z faktu, że nie wszystkie rekordy appearances mają poprawnie zdefiniowane daty (daty z roku 1899). Występują także niepoprawne rekordy dotyczące występów piłkarzy z przyszłości, które mają już przypisane wyniki meczu. Konieczna jest zatem korekta poprzez połączenie z istniejącymi meczami. Następnie na podstawie daty wyliczane są dodatkowe informacje dotyczące wymiaru czasu. Obliczany jest klucz time\_id, którego 4 pierwsze cyfry reprezentują rok, 2 kolejne cyfry reprezentują miesiąc, a dwie ostatnie cyfry reprezentują dzień miesiąca. Wyodrębniane są rok, miesiąc, dzień oraz dzień tygodnia. Następnie na podstawie tych danych następuje połączenie z tabelami pomocniczymi, które zapewniają nazwę miesiąca i dnia tygodnia słownie.

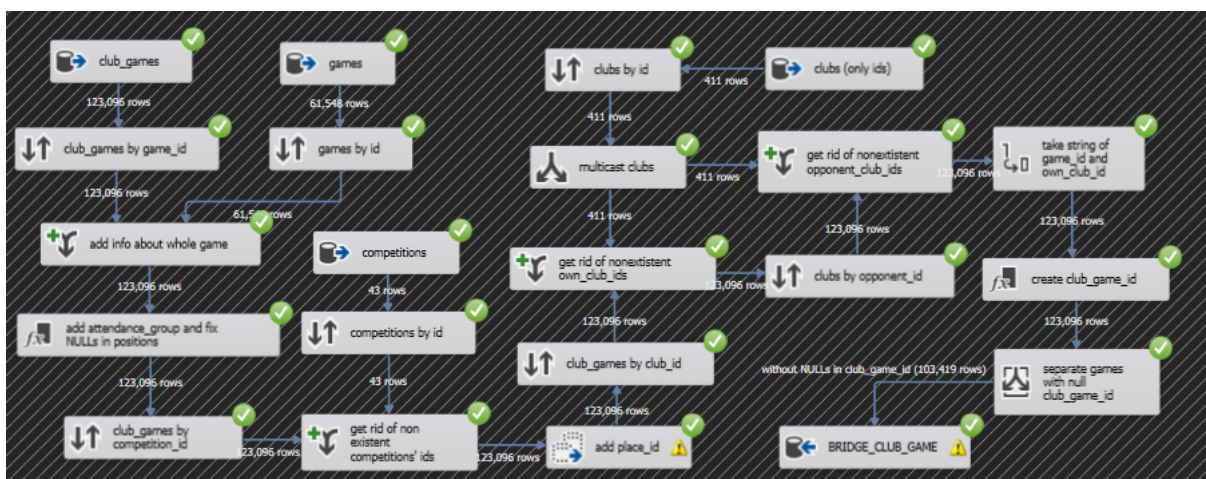
### Zadanie DIM\_PLACE:



Najpierw zaciągane są dane z tabeli games. Pobierany jest jedynie atrybut przechowujący nazwę stadionu. Następnie wyodrębniane są jedynie unikatowe wartości oraz usuwana wartość NULL. Za pomocą fuzzy lookup dopasowywane są nazwy stadionów z pliku stadiums z podobieństwem 0,60. Na tej podstawie dobierane są również pozostałe informacje dotyczące miejsca, takie jak miasto, pojemność stadionu, kraj i populacja kraju. Dla stadionów, dla których dopasowano nazwę z tabeli stadiums preferowana jest nazwa z tabeli stadiums, co pozwoli na eliminację błędnej pisowni. Dla stadionów, dla którym nie udało się dopasować nazwy wybierana jest nazwa z tabeli games. Na koniec dodawana jest kolumna country\_population\_group i usuwane są duplikaty powstałe w wyniku poprawiania pisowni. Ponieważ w tabeli źródłowej nie było id, jest ono nadawane poprzez autonumerowanie.

country\_population\_group rozróżnia 6 grup wielkości populacji państwa: 0 – 1 mln, 1 – 5 mln, 5 – 10 mln, 10 – 20 mln, 20 – 50 mln, 50+ mln. W przypadku, gdy nie zdefiniowano populacji kraju przypisywana jest wartość NULL.

### Zadanie BRIDGE\_CLUB\_GAME:



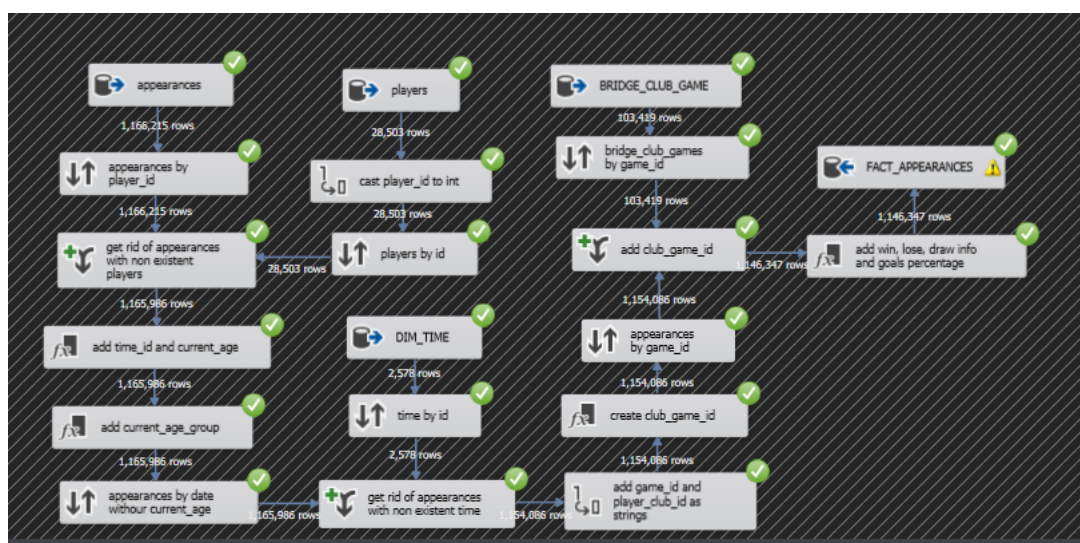
Zadanie BRIDGE\_CLUB\_GAME włącza do club\_games informacje o sędzi, stadionie i wielkości publiki w trakcie meczu. Następnie dodany jest atrybut attendance\_group. Nieprawidłowe wartości (-1) w pozycji w rankingu klubu rozgrywającego i oponenta zamieniane są na NULLe. Następnie weryfikowane są przyporządkowane competition\_id. Następowały niespójności w danych i istniały



połączenia do konkursów, dla których nie istniały odpowiadające rekordy w competitions. Ich identyfikatory zamieniono na NULLe. Następnie za pomocą fuzzy lookup i tabeli DIM\_PLACE dopasowano place\_id stadionu, na którym rozgrywano mecz. Kolejnym krokiem było poprawienie przypisanych id klubów rozgrywającego i oponenta. Miała miejsce sytuacja analogiczna jak z konkursami i trzeba było zamienić odwołania do nieistniejących rekordów z clubs na NULLe. Na koniec stworzone zostało club\_game\_id, które składało się z konkatencasji game\_id + „\_” + club\_id castowanych na stringi.

attendance\_group rozróżnia 6 grup wielkości widowni: bez widowni , 0 - 10 tys., 10 - 30 tys, 30 - 50 tys., 50 - 70 tys.

### Zadanie FACT\_APPEARANCES:



Zdecydowano się na usuwanie wybrakowanych, anomalnych danych, które mogą zaburzać analizę, a są na pewno nieprawidłowe. Zdecydowano się na rezygnację z rekordów bez przypisanego gracza bo nie można analizować wystąpienia gracza, który nie istnieje. Następnie dla każdego wystąpienia wyliczono wiek aktualny podczas rozgrywania danego meczu. Dodano atrybut current\_age\_group. Następnie zweryfikowano połączenie z time\_id, usuwając wystąpienia graczy, które na pewno są nieprawidłowe, np. mające miejsce w przyszłości. Na koniec utworzono zostało club\_game\_id, które składało się z konkatencasji game\_id + „\_” + current\_club\_id castowanych na stringi. Dzięki temu zweryfikowane zostało połączenie z meczami z perspektywy drużyn, co pozwoliło na usunięcie nieistniejących powiązań. Dodano atrybuty pod utworzenie miar – informacje o wygranej, remisie bądź przegranej oraz procentowy udział gracza w bramkach drużyny w trakcie meczu.

current\_age\_group rozróżnia 6 grup wiekowych: 16 – 20 lat, 20 – 25 lat, 25 – 30 lat, 30 – 35 lat, 35 – 40 lat, 40+ lat. W przypadku, gdy nie zdefiniowano wieku gracza przypisywana jest wartość NULL.

### Zadanie References:

Nakłada więzy integralności na tabele. Nakłada ograniczenie klucza głównego na competition\_id w DIM\_COMPETITION, time\_id w DIM\_TIME, club\_id w DIM\_CLUB, player\_id w DIM\_PLAYER, place\_id w DIM\_PLACE, club\_game\_id w BRIDGE\_CLUB\_GAME oraz appearance\_id w FACT\_APPEARANCES.

Dla BRIDGE\_CLUB\_GAME nakłada ograniczenie klucza obcego na competition\_id jako powiązanie z DIM\_COMPETITION, place\_id jako powiązanie z DIM\_PLACE, own\_club\_id jako powiązanie z DIM\_CLUB oraz opponent\_club\_id jako powiązanie z DIM\_CLUB.

Dla FACT\_APPEARANCES nakłada ograniczenie klucza obcego na club\_game\_id jako powiązanie z BRIDGE\_CLUB\_GAME, player\_id jako powiązanie z DIM\_PLAYER oraz time\_id jako powiązanie z DIM\_TIME.



Cel					Źródło			Przekształcenie
Nazwa tabeli	Nazwa kolumny	Typ danych	Typ tabeli	Baza	Nazwa tabeli	Kolumna	Typ danych	
FACT_APPEARANCES	appearance_id	NVARCHAR(20)	FAKT	1	appearances	appearance_id	NVARCHAR	Brak
FACT_APPEARANCES	club_game_id	NVARCHAR(20)	FAKT	1	appearances	game_id	INT	Castowanie game_id i player_club_id na typ DT_STR, a następnie utworzenie konkatenacji game_id + „_” + player_club_id. Połączenie INNER JOIN utworzonej konkatenacji z tabelą BRIDGE_CLUB_GAME na atrybucie club_game_id, w celu otrzymania jedynie wystąpień o przypisanym meczu.
				1	appearances	player_club_id	INT	
				3	BRIDGE_CLUB_GAME	club_game_id	NVARCHAR(20)	
FACT_APPEARANCES	player_id	INT	FAKT	1	appearances	player_id	INT	Połączenie INNER JOIN eliminujące wystąpienia o player_id, które nie odnosi się do żadnego piłkarza w players.
				1	players	player_id	INT	
FACT_APPEARANCES	current_age_group	NVARCHAR(20)	FAKT	1	appearances	date	DATE	Obliczenie wieku piłkarza (w latach) w momencie rozgrywania danego meczu. Następnie podział uzyskanych wyników na 6 przedziałów.
				1	players	date_of_birth	DATE	
FACT_APPEARANCES	minutes_played	INT	FAKT	1	appearances	minutes_played	INT	Brak
FACT_APPEARANCES	goals	INT	FAKT	1	games	date	INT	Brak
FACT_APPEARANCES	yellow_cards	INT	FAKT	1	appearances	yellow_cards	INT	Brak

FACT_APPEARANCES	red_cards	INT	FAKT	1	appearances	red_cards	INT	Brak
FACT_APPEARANCES	is_win	INT	FAKT	1	appearances	is_win	INT	Połączenie INNER JOIN eliminujące wystąpienia bez odpowiadającego meczu. Gdy own_goals jest większe od opponent goals przydzielona zostaje wartość 1, w przeciwnym przypadku przydzielana jest wartość 0.
				1	appearances	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	own_goals	INT	
				3	BRIDGE_CLUB_GAME	opponent_goals	INT	
FACT_APPEARANCES	is_draw	INT	FAKT	1	appearances	is_win	INT	Połączenie INNER JOIN eliminujące wystąpienia bez odpowiadającego meczu. Gdy own_goals jest równe opponent goals przydzielona zostaje wartość 1, w przeciwnym przypadku przydzielana jest wartość 0.
				1	appearances	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	own_goals	INT	
				3	BRIDGE_CLUB_GAME	opponent_goals	INT	
FACT_APPEARANCES	is_lose	INT	FAKT	1	appearances	is_win	INT	Połączenie INNER JOIN eliminujące wystąpienia bez odpowiadającego meczu. Gdy own_goals jest mniejsze od opponent goals przydzielona zostaje wartość 1, w przeciwnym przypadku przydzielana jest wartość 0.
				1	appearances	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	own_goals	INT	
				3	BRIDGE_CLUB_GAME	opponent_goals	INT	
FACT_APPEARANCES	goals_percentage	INT	FAKT	1	appearances	goals	INT	Połączenie INNER JOIN eliminujące wystąpienia

				1	appearances	club_game_id	INT	bez odpowiadającego meczu. Wyliczenie procentowego udziału piłkarza w golach drużyny w trakcie meczu stworzone na podstawie wyliczenia: w przypadku gdy drużyna nie strzeliła goli przydziel 0, gdy drużyna zdobyła jakieś bramki castuj goals oraz home_club_goals na typ float i oblicz (goals/home_club_goals * 100).
				3	BRIDGE_CLUB_GAME	club_game_id	INT	
				3	BRIDGE_CLUB_GAME	own_goals	INT	
FACT_APPEARANCES	time_id	INT	FAKT	1	appearances	date	DATE	Data przedstawiona w postaci liczby, której 4 pierwsze cyfry reprezentują rok, 2 kolejne cyfry reprezentują miesiąc, a 2 ostatnie cyfry odpowiadają dniowi miesiąca. Złączenie INNER JOIN z DIM_TIME w celu wyeliminowania wystąpień bez podanej daty i takich, w których występują niespójności pomiędzy datą meczu a wystąpienia.
				3	DIM_TIME	time_id	INT	
DIM_PLACE	place_id	INT	WYMIAR	-	-	-	-	Wprowadzono klucz sztuczny oparty o autonumerowanie ze

								względ na brak klucza w pliku źródłowym
DIM_PLACE	stadium	NVARCHAR(100)	WYMIAR	2	stadiums	stadium	NVARCHAR	<p>Najpierw wzięto informacje jedynie o stadionach z tabeli games. Zgrupowano je po nazwie stadionu i przy użyciu bloczka „Conditional Split” odrzucono wartość NULL. W tak przygotowanych danych występowało wiele literówek, te same stadiony występowały po kilka razy zapisane inaczej. Wykorzystano więc fuzzy lookup na nazwach stadionów z pliku stadiums (z podobieństwem 0,60). Dla stadionów, dla których nie udało się znaleźć przyporządkowania w stadiums zostawiono samą nazwę stadionu z games, nie rozszerzając informacji o miejscu. Dla stadionów ze znalezionym przyporządkowaniem zmieniono nazwę na tę ze stadiums. Ze względu na liczne literówki w games doprowadziło to do</p>
				1	games	stadium	NVARCHAR	

								powstania duplikatów, które zostały usunięte poprzez wzięcie unikatowych wartości.
DIM_PLACE	city	NVARCHAR(40)	WYMIAR	2	stadiums	city	NVARCHAR	Brak
DIM_PLACE	capacity	INT	WYMIAR	2	stadiums	capacity	INT	Brak
DIM_PLACE	country	NVARCHAR(40)	WYMIAR	2	stadiums	country	NVARCHAR	Brak
DIM_PLACE	country_population	INT	WYMIAR	2	stadiums	country_population	INT	Brak
DIM_PLACE	country_population_group	NVARCHAR(20)	WYMIAR	2	stadiums	country_population_group	NVARCHAR	Przydzielenie jednej z 5 grup wielkości populacji lub NULLa, gdy brak informacji o wielkości populacji kraju
DIM_CLUB	club_id	INT	WYMIAR	1	clubs	club_id	INT	Brak
DIM_CLUB	name	NVARCHAR(50)	WYMIAR	1	clubs	name	NVARCHAR	Brak
DIM_CLUB	foreigners_percentage	FLOAT	WYMIAR	1	clubs	foreigners_percentage	FLOAT	Brak
DIM_CLUB	foreigners_percentage_group	NVARCHAR(20)	WYMIAR	1	clubs	foreigners_percentage	FLOAT	Przydzielenie jednej z 4 grup udziału obcokrajowców w składzie drużyny lub wartości NULL, gdy tej informacji dla klubu brakuje.
DIM_CLUB	average_age	FLOAT	WYMIAR	1	clubs	average_age	FLOAT	Dla nieprawidłowej wartości average_age wynoszącej 0 przypisywany jest NULL.
DIM_CLUB	average_age_group	NVARCHAR(20)	WYMIAR	1	clubs	average_age	FLOAT	Przydzielenie jednej z 6 grup średniego wieku piłkarzy w składzie drużyny lub wartości NULL, gdy tej informacji dla klubu brakuje.
BRIDGE_CLUB_GAME	club_game_id	NVARCHAR(20)	BRIDGE	1	club_games	game_id	INT	W przypadku, gdy połączeniach albo game_id
				1	club_games	club_id	INT	

								albo club_id jest NULLeM, club_game_id również zostanie NULLeM. Dla reszty przypadków game_id i club_id zostaną castowane na typ DT_STR, a następnie utworzona zostanie konkatencja game_id + „_” + club_id. Wszystkie rekordy o club_game_id będącym NULLeM zostaną odseparowane z użyciem blozku „Conditional split” i niezapisane w tabeli.
BRIDGE_CLUB_GAME	round	NVARCHAR(30)	BRIDGE	1	games	round	NVARCHAR	Runda turnieju uzyskana na podstawie połączenia club_games z games.
				1	club_games	game_id	INT	
				1	games	game_id	INT	
BRIDGE_CLUB_GAME	referee	NVARCHAR(50)	BRIDGE	1	games	referee	NVARCHAR	Sędzia meczu uzyskany na podstawie połączenia club_games z games.
				1	club_games	game_id	INT	
				1	games	game_id	INT	
BRIDGE_CLUB_GAME	own_club_id	INT	BRIDGE	1	club_games	own_club_id	INT	Połączenie LEFT JOIN z tabelą clubs w celu zamienia id klubów w club_games nieistniejących w clubs na wartości NULL.
				1	clubs	club_id	INT	
BRIDGE_CLUB_GAME	opponent_club_id	INT	BRIDGE	1	club_games	opponent_club_id	INT	Połączenie LEFT JOIN z tabelą clubs w celu zamienia id klubów w club_games nieistniejących w clubs na wartości NULL.
				1	clubs	club_id	INT	

BRIDGE_CLUB_GAME	own_position	INT	BRIDGE	1	club_games	own_position	INT	Brak
BRIDGE_CLUB_GAME	opponent_position	INT	BRIDGE	1	club_games	opponent_position	INT	Brak
BRIDGE_CLUB_GAME	own_manager_name	NVARCHAR(50)	BRIDGE	1	club_games	own_manager_name	NVARCHAR	Brak
BRIDGE_CLUB_GAME	opponent_manager_name	NVARCHAR(50)	BRIDGE	1	club_games	opponent_manager_name	NVARCHAR	Brak
BRIDGE_CLUB_GAME	own_goals	INT	BRIDGE	1	club_games	own_goals	INT	Brak
BRIDGE_CLUB_GAME	opponent_goals	INT	BRIDGE	1	clubs	opponent_goals	INT	Brak
BRIDGE_CLUB_GAME	competition_id	NVARCHAR(4)	BRIDGE	1	games	competition_id	NVARCHAR	Identyfikator konkursu uzyskany na podstawie połączenia club_games z games.
				1	club_games	game_id	INT	
				1	games	game_id	INT	
BRIDGE_CLUB_GAME	is_win	INT	BRIDGE	1	club_games	is_win	INT	Brak
BRIDGE_CLUB_GAME	hosting	NVARCHAR(4)	BRIDGE	1	club_games	hosting	NVARCHAR	Brak
BRIDGE_CLUB_GAME	place_id	INT	BRIDGE	1	club_games	game_id	INT	Najpierw pozyskano nazwę stadionu, na którym rozgrywano mecz poprzez połączenie club_games z games. Następnie w DIM_PLACE z wykorzystaniem fuzzy_lookup wyszukano nazwy stadionów z podobieństwem 0,60. Dzięki temu dla każdego meczu, któremu przypisany był stadion w games uzyskano odpowiednie połączenie poprzez zaciągnięcie identyfikatora miejsca (place_id). W przypadku braku przypisania
				1	games	game_id	INT	
				1	games	stadium	NVARCHAR	
				3	DIM_PLACE	stadium	NVARCHAR	

								wstawiona została wartość NULL.
BRIDGE_CLUB_GAME	attendance	INT	BRIDGE	1	games	attendance	INT	Liczba osób oglądających mecz na stadionie uzyskana na podstawie połączenia club_games z games.
				1	club_games	game_id	INT	
				1	games	game_id	INT	
BRIDGE_CLUB_GAME	attendance_group	NVARCHAR(20)	BRIDGE	1	games	attendance	INT	Przydzielenie liczbie osób oglądających mecz na stadionie pozyskanej z połączenia club_games z games jednej z 6 grup opisujących wielkość publiki.
				1	club_games	game_id	INT	
				1	games	game_id	INT	
DIM_COMPETITION	competition_id	NVARCHAR(4)	WYMIAR	1	competitions	competition_id	NVARCHAR	Brak
DIM_COMPETITION	name	NVARCHAR(50)	WYMIAR	1	competitions	name	NVARCHAR	Brak
DIM_COMPETITION	type	NVARCHAR(17)	WYMIAR	1	competitions	type	NVARCHAR	Brak
DIM_COMPETITION	sub_type	NVARCHAR(40)	WYMIAR	1	competitions	sub_type	NVARCHAR	Brak
DIM_COMPETITION	country_name	NVARCHAR(11)	WYMIAR	1	competitions	country	NVARCHAR	Brak
DIM_TIME	time_id	INT	WYMIAR	1	games	date	DATE	Przekształcenie daty na liczbę, gdzie pierwsze 4 cyfry to rok, następne 2 to miesiąc i ostatnie 2 to dzień
DIM_TIME	season	INT	WYMIAR	1	games	season	INT	Brak
DIM_TIME	year	INT	WYMIAR	1	games	date	DATE	Rok wyizolowany z daty meczu
DIM_TIME	quarter	INT	WYMIAR	1	games	date	DATE	Kwartał wyizolowany z daty meczu
DIM_TIME	month	INT	WYMIAR	1	games	date	DATE	Miesiąc wyizolowany z daty meczu
DIM_TIME	day	INT	WYMIAR	1	games	date	DATE	Dzień miesiąca wyizolowany z daty meczu



DIM_TIME	month_name	NVARCHAR	WYMIAR	1	games	date	DATE	Nazwa miesiąca przydzielona na podstawie połączenia wyliczonego z daty meczu miesiąca z nazwą z tabeli pomocniczej. Tabela pomocnicza zawiera informacje o nazwach miesięcy wyrażonych liczbowo.
DIM_TIME	day_of_week	NVARCHAR	WYMIAR	1	games	date	DATE	Nazwa dnia tygodnia przydzielona na podstawie połączenia wyliczonego z daty meczu dnia tygodnia z nazwą z tabeli pomocniczej. Tabela pomocnicza zawiera informacje o nazwach dni tygodnia wyrażonych liczbowo.
DIM_PLAYER	player_id	INT	WYMIAR	1	players	player_id	INT	Brak
DIM_PLAYER	name	NVARCHAR(40)	WYMIAR	1	players	name	NVARCHAR	Brak
DIM_PLAYER	date_of_birth	DATE	WYMIAR	1	players	date_of_birth	DATE	Brak
DIM_PLAYER	country_of_birth	NVARCHAR(40)	WYMIAR	1	players	country_of_birth	NVARCHAR	Brak
DIM_PLAYER	city_of_birth	NVARCHAR(60)	WYMIAR	1	players	city_of_birth	DATE	Brak
DIM_PLAYER	position	NVARCHAR(10)	WYMIAR	1	players	position	NVARCHAR	Brak
DIM_PLAYER	sub_position	NVARCHAR(18)	WYMIAR	1	players	sub_position	NVARCHAR	Brak
DIM_PLAYER	foot	NVARCHAR(5)	WYMIAR	1	players	foot	NVARCHAR	Brak
DIM_PLAYER	height_in_cm	INT	WYMIAR	1	players	height_in_cm	INT	Dla nieprawidłowej wartości 0 przypisywany jest NULL
DIM_PLAYER	height_group	NVARCHAR(20)	WYMIAR	1	players	height_in_cm	INT	Przydzielenie jednej z 7 grup wzrostu lub wartości NULL w przypadku

								braku informacji o wzroście.
DIM_PLAYER	highest_market_value_in_eur	FLOAT	WYMIAR	1	players	market_value	INT	Dla nieprawidłowej wartości 0 przypisywany jest NULL
DIM_PLAYER	highest_market_value_in_eur_group	VARCHAR(15)	WYMIAR	1	players	market_value	INT	Przydzielenie jednej z 7 grup wartości piłkarza lub NULLa w przypadku braku informacji o wartości.
DIM_PLAYER	last_season	INT	WYMIAR	1	players	market_value	INT	Brak

#### LEGENDA:

- 1 – numer odpowiadający bazie Football Data from Transfermarkt
- 2 – numer odpowiadający bazie Stadiums
- 3 – numer odpowiadający bazie Soccer

## Wnioski:

Bardzo istotne jest prawidłowe profilowanie danych. Dzięki temu przy uzupełnianiu tabel wymiarów i tabeli faktów możliwe było dobranie rozmiarów pól pozwalających na zoptymalizowanie zajmowanej pamięci. Ponadto poprzez rozpoznanie nieprawidłowych wartości atrybutów możliwe było przeprowadzenie transformacji na danych, które powinny zostać NULLami. Było to szczególnie istotne, gdyż wartości takie jak wzrost wynoszący 0 cm czy -1 zamieniające się w datę z 1899 roku mogłyby zaburzyć przeprowadzane analizy.

Na kodowanie należy zwracać uwagę już przy tworzeniu hurtowni danych. Źle ustawione kodowanie wpłynie na niepoprawne wyświetlanie znaków, co pozbawi zestawienia czytelności. Nie zawsze możliwe jest ustawienie poprawnego kodowania poszczególnych tabel i kolumn, gdy źle kodowana jest cała baza, co sprawi, że potrzebne będzie stworzenie bazy i importowanie danych od nowa. Kontrolę kodowania należy przeprowadzić już na plikach źródłowych, gdyż same importowane dane mogą być źródłem problemów z kodowaniem.

Czasami dla prawidłowego przekształcenia danych konieczne jest wykorzystanie elementów fuzzy. O ile fuzzy lookup przebiega dość sprawnie, to już fuzzy grouping trwa bardzo długo. Dla tworzonej hurtowni fuzzy grouping wykonywany na 60k rekordów trwał ponad 30 minut, dlatego zdecydowano się na rezygnację z jego użycia i zrealizowanie zadań z wykorzystaniem fuzzy lookupów. Elementy fuzzy pozwalają na pracę z danymi, w których występują błędy w pisowni. Dzięki zastosowaniu fuzzy lookup dla fajnych dotyczących stadionów udało się dopasować dane z dwóch różnych źródeł i ujednolicić pisownię. Dzięki zastosowaniu fuzzy lookup udało się zwiększyć liczbę dopasowanych stadionów z 327 do 642.

Przy tworzeniu zadań data flow konieczne jest pamiętanie o przekazywaniu jedynie faktycznie wykorzystywanych atrybutów. Niepotrzebne przekazywanie danych może wpłynąć na wydajność zadań. Ostrzeżenia o niewykorzystywanych danych można zobaczyć w konsoli i odpowiednio zoptymalizować przebieg data flow.

## Projekt – etap III (12.06./15.06.)

### Kostka:

1. Przygotować projekt kostki, edytować wymiary, dodać miary kalkulowane. Przygotować zestawienia z punktu 4. pierwszego etapu oraz pokazać inne ciekawe zależności w analizowanych danych (analiza w głąb, a nie tylko tabele przestawne).

Przy ocenie będą brane następujące elementy kostki:

- prawidłowa struktura kostki – model kostki powinien analitykowi na intuicyjne i łatwe korzystanie z danych
- miary kalkulowane
- dokumentacja, która powinna zawierać krótki opis wszystkich wymiarów, wszystkich ich atrybutów oraz wszystkich miar

### Dokumentacja kostki

Wymiar DIM_CLUB	
Wymiar zawierający informacje o klubach piłkarskich.	
Atrybut	Opis
club_id	Klucz główny
name	Nazwa klubu
coach_name	Konkatenacja imienia i nazwiska coacha, który prowadzi klub
foreigners_percentage_group	Grupa procentowego udziału obcokrajowców w składzie drużyny klubu
average_age_group	Grupa średniego wieku piłkarzy w drużynie klubu

Wymiar DIM_PLAYER	
Wymiar zawierający informacje o piłkarzu.	
Atrybut	Opis
player_id	Klucz główny
name	Konkatenacja imienia i nazwiska piłkarza
date_of_birth	Data urodzenia piłkarza
country_of_birth	Kraj urodzenia piłkarza
position	Pozycja gry piłkarza
sub_position	Podpozycja gry piłkarza
foot	Stopa preferowana przez piłkarza
height_group	Grupa wzrostu, do której należy piłkarz
highest_market_value_in_eur_group	Grupa wartości w euro, do której należy piłkarz
last_season	Ostatni sezon, w którym rozgrywał piłkarz

Wymiar DIM_PLACE	
Wymiar opisujący stadion, na którym rozgrywane są mecze piłkarskie.	
Atrybut	Opis
place_id	Klucz główny
stadium	Nazwa stadionu (na którym mogą być rozgrywane mecze)
city	Nazwa miasta, w którym znajduje się stadion
capacity	Pojemność stadionu
country	Nazwa kraju, w którym znajduje się stadion
country_population_group	Grupa wielkości populacji kraju, w którym znajduje się stadion

Wymiar DIM_TIME	
Wymiar opisujący czas rozgrywania meczu piłkarskiego.	
Atrybut	Opis
time_id	Klucz główny, złożony w formacie RRRRMMDD, gdzie RRRR to rok, MM to miesiąc i DD to dzień
season	Sezon piłkarski, który zawiera dany termin
year	Rok
quarter	Kwartał
month	Miesiąc
day	Dzień
month_name	Nazwa miesiąca słownie
day_of_week	Nazwa dnia tygodnia

Wymiar DIM_COMPETITION	
Konkurs, w ramach którego rozgrywane są mecze.	
Atrybut	Opis
competition_id	Klucz główny
name	Nazwa konkursu
type	Typ konkursu
sub_type	Podtyp konkursu
country_name	Nazwa kraju, który jest organizatorem konkursu

Bridge BRIDGE_CLUB_GAME	
Zawiera informacje o meczu z perspektywy jednego klubu. Dla jednego meczu przypada perspektywa dwóch drużyn.	
Atrybut	Opis
club_game_id	Klucz główny
round	Etap konkursu, w ramach którego rozgrywany jest mecz
referee	Sędzia, który sędziował podczas meczu
own_club_id	Klucz obcy klubu, z którego perspektywy przedstawiane są informacje o meczu
opponent_club_id	Klucz obcy klubu, który jest stroną przeciwną w przedstawianym meczu
own_position	Pozycja w tabeli rankingowej klubu, z którego perspektywy przedstawiany jest mecz
opponent_position	Pozycja w tabeli rankingowej klubu, który jest stroną przeciwną w przedstawianym meczu
own_manager_name	Konkatenacja imienia i nazwiska menagara klubu, z którego perspektywy przedstawiany jest mecz
opponent_manager_name	Konkatenacja imienia i nazwiska menagara klubu, który jest stroną przeciwną w przedstawianym meczu
own_goals	Liczba goli strzelonych podczas meczu przez klub, z którego perspektywy przedstawiany jest mecz
opponent_goals	Liczba goli strzelonych podczas meczu przez klub, który jest stroną przeciwną w przedstawianym meczu
competition_id	Klucz obcy konkursu, w ramach którego rozgrywany jest mecz
is_win	Wartość 1 w przypadku wygranej, wartość 0 w przypadku remisu lub przegranej
hosting	Informacja o tym, czy klub, z którego perspektywy przedstawiany jest mecz, był właścicielem spotkania
place_id	Klucz obcy miejsca, w którym rozgrywany był mecz
attendance_group	Grupa wielkości publiczności, która pojawiła się na meczu

Wymiar FACT_APPEARANCES	
Wymiar złożony z atrybuty wizyolowanego z tabeli faktów FACT_APPEARANCES. Wymiar opisujący grupę wiekową piłkarza w momencie rozgrywania meczu.	
Atrybut	Opis
appearance_id	Klucz główny wystąpienia w meczu
current_age_group	Grupa wiekowa piłkarza w momencie rozgrywania meczu

Utworzone hierarchie

DIM\_PLAYER:

- position -> sub\_position

DIM\_PLACE:

- country -> city -> stadium

DIM\_COMPETITION:

- type -> sub\_type

DIM\_TIME:

- year -> quarter -> month -> day

Miary dla tabeli faktów FACT_APPEARANCES	
Tabela faktów FACT_APPEARANCES zawiera informacje o pojedynczych wystąpieniach piłkarzy w meczach.	
Miara	Opis
minutes_played (sum)	Suma minut spędzonych na boisku
goals (sum)	Suma zdobytych goli
Maximum goals	Maksymalna liczba zdobytych goli
assits (sum)	Suma zdobytych asyst
yellow_cards (sum)	Suma zdobytych żółtych kartek
red_cards (sum)	Suma zdobytych czerwonych kartek
is_win (sum)	Liczba meczy, które zostały wygrane (wyliczana na podstawie sumy)
is_draw (sum)	Liczba meczy, które zostały zremisowane (wyliczana na podstawie sumy)
is_lose (sum)	Liczba meczy, które zostały wygrane (wyliczana na podstawie sumy)
FACT_APPEARANCES Count	Zliczenie wystąpień piłkarza w meczach
player_id Distinct Count	Zliczenie unikatowych piłkarzy
average_goal_percentage (average)	Miara kalkulowana, średni procent udziału procentowego piłkarza w golach drużyny. Najpierw wyliczony został procentowy udział piłkarza w golach w pojedynczym meczu. Następnie wyliczona została średnia.
average_goals (average)	Miara kalkulowana, średnia liczba goli. Wyliczana jako podzielenie sumy liczby zdobytych goli przez liczbę wystąpień w meczach.
average_assists (average)	Miara kalkulowana, średnia liczba asyst. Wyliczana jako podzielenie sumy liczby zdobytych asyst przez liczbę wystąpień w meczach.
average_yellow_cards (average)	Miara kalkulowana, średnia liczba otrzymanych żółtych kartek. Wyliczana jako podzielenie sumy

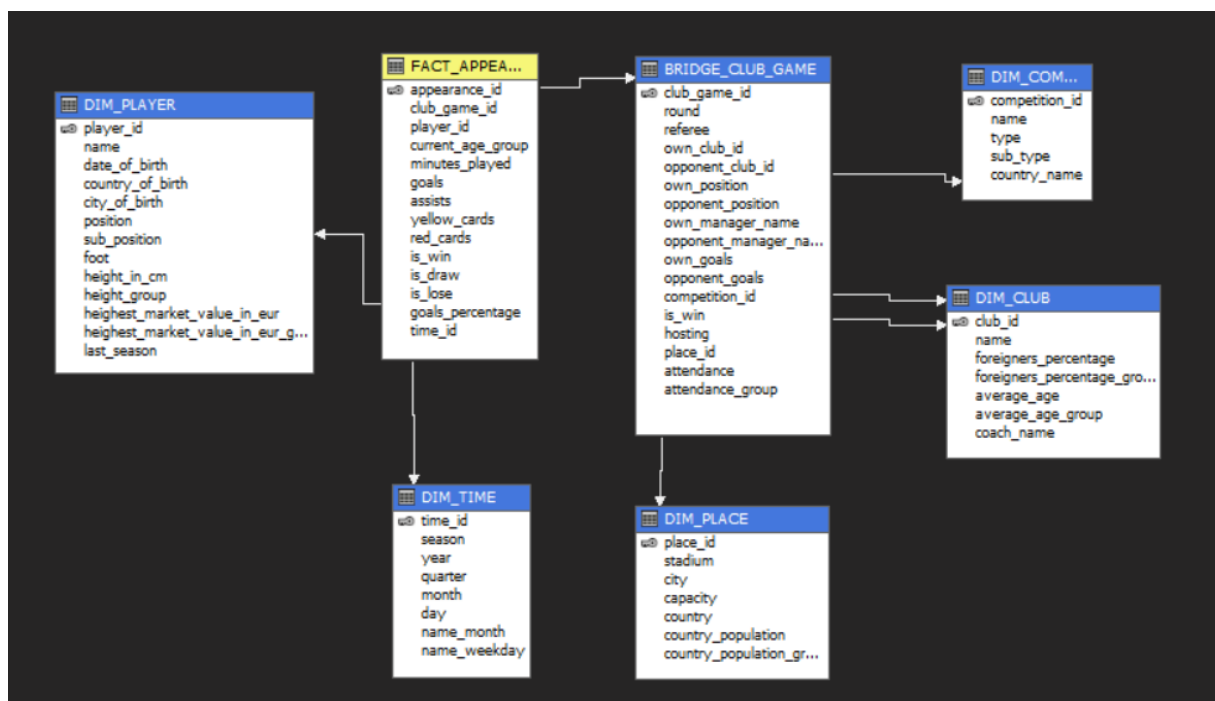
	liczby otrzymanych żółtych kartek przez liczbę występów w meczach.
average_red_cards (average)	Miara kalkulowana, średnia liczba otrzymanych czerwonych kartek. Wyliczana jako podzielenie sumy liczby otrzymanych czerwonych kartek przez liczbę występów w meczach.
average_minutes_played (average)	Miara kalkulowana, średnia liczba minut spędzonych na boisku. Wyliczana jako podzielenie sumy liczby minut spędzonych na boisku przez liczbę występów w meczach.
Average_won_games (average)	Miara kalkulowana, średnia liczba wygranych meczy. Wyliczana jako podzielenie liczby wygranych meczy przez liczbę występów w meczach.

KPI dla tabeli faktów FACT_APPEARANCES	
Tabela faktów FACT_APPEARANCES zawiera informacje o pojedynczych wystąpieniach piłkarzy w meczach.	
KPI	Opis
GoalsKPI	KPI istotne przy ocenie formy piłkarzy. Bierze pod uwagę średnią liczbę zdobytych goli i sprawdza, czy został ustalony cel (0.2). Jeżeli cel został przekroczony przyznawana jest wartość 1, jeżeli cel został równo osiągnięty przydzielana jest wartość 0, a jeżeli nie osiągnięto celu przydzielana jest wartość -1. Pomaga w ocenie formy piłkarza.
AssistsKPI	KPI istotne przy ocenie formy piłkarzy. Bierze pod uwagę średnią liczbę zdobytych asyst i sprawdza, czy został ustalony cel (0.2). Jeżeli cel został przekroczony przyznawana jest wartość 1, jeżeli cel został równo osiągnięty przydzielana jest wartość 0, a jeżeli nie osiągnięto celu przydzielana jest wartość -1.
AggresionKPI	KPI istotne przy ocenie kultury gry piłkarzy. Bierze pod uwagę średnią liczbę otrzymanych czerwonych kartek i sprawdza, czy nie został przekroczony cel (0.005). Jeżeli cel został przekroczony przyznawana jest wartość -1, jeżeli cel został równo osiągnięty przydzielana jest wartość 0, a jeżeli nie przekroczono celu przydzielana jest wartość 1. Pomaga wybrać graczy, którzy grają kulturalnie.
GeneralAggresionKPI	KPI istotne przy ocenie kultury gry piłkarzy. Bierze pod uwagę średnią liczbę żółtych kartek oraz średnią liczbę czerwonych kartek wymnożoną przez liczbę dni w roku (365) i sprawdza, czy nie został przekroczony cel (0.05). Jeżeli cel został przekroczony przyznawana jest wartość -1, jeżeli cel został równo osiągnięty przydzielana jest wartość 0, a jeżeli nie



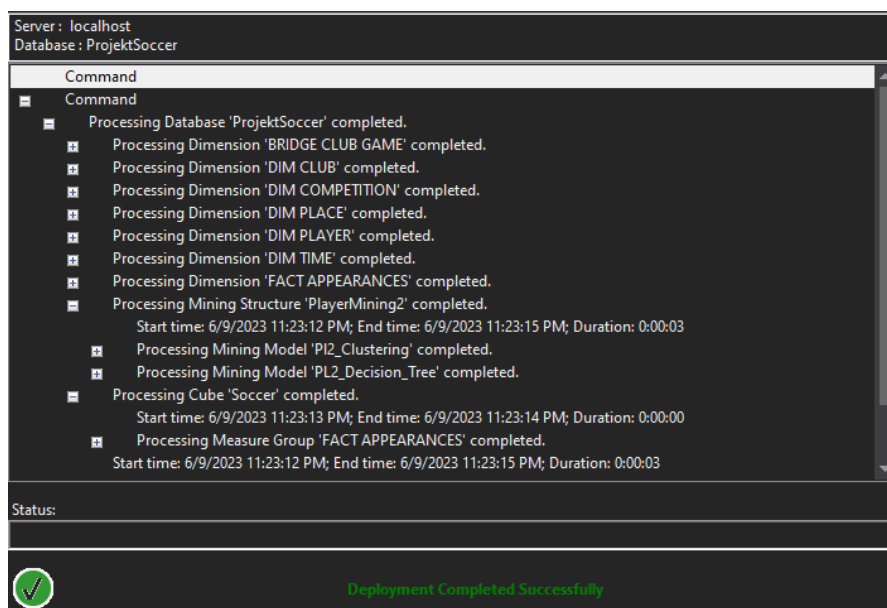
	<p>przekroczono celu przydzielana jest wartość 1. Pomaga wybrać graczy, którzy grają kulturalnie. Zdecydowano się na wymnożenie czerwonych kartek przez liczbę dni w roku, tak aby podkreślić ich ogromne znaczenie (zejście piłkarza z boiska) i piętnować je bardziej niż kartki żółte. Gdyby chciano całkowicie wyeliminować czerwone kartki można by stworzyć KPI biorące pod uwagę liczbę otrzymanych czerwonych kartek. Takie wymnożenie jednak sprawia, że całkowicie odrzuceni nie zostaną gracze, którzy rozegrali dużo meczy i czerwone kartki były jedynie pojedynczymi epizodami.</p>
--	---

### Struktura utworzonej kostki

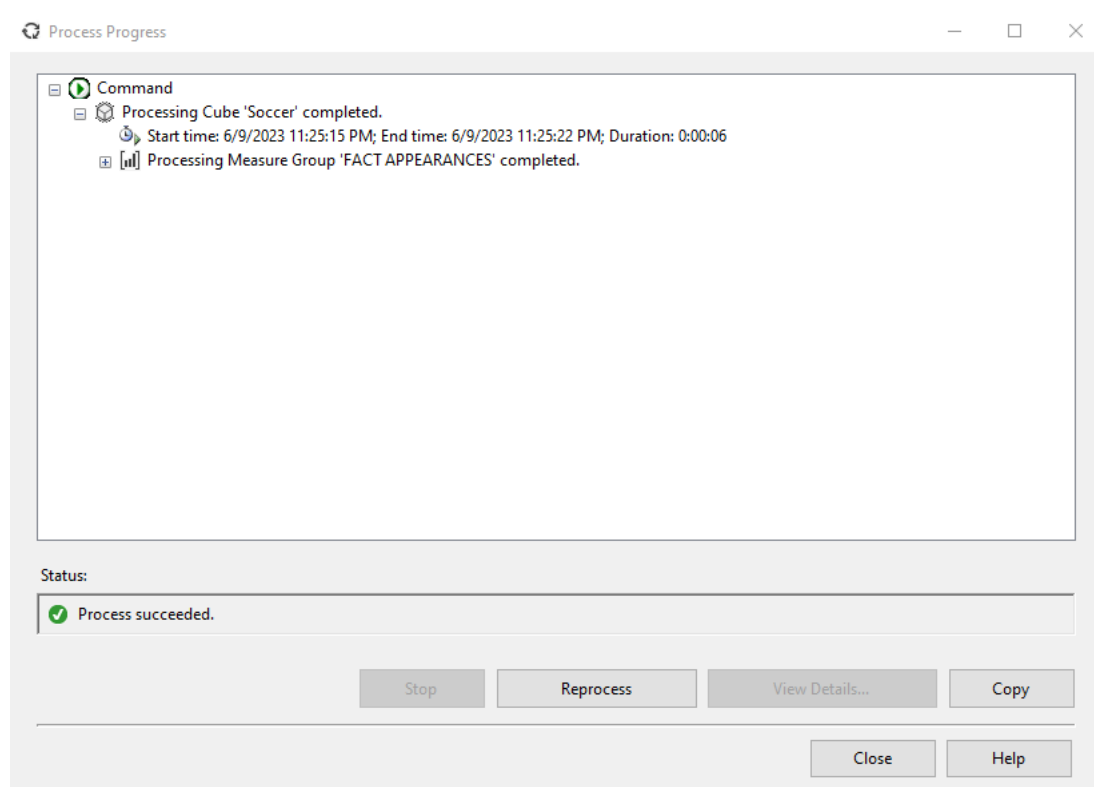


W utworzonych wymiarach kostki zdecydowano się na zachowanie atrybutów z hurtowni, które będą istotne z punktu zestawień, co opisano w dokumentacji kostki w postaci tabel. W hurtowni zdecydowano się na przechowywanie podstaw wartości wyliczanych (np. oprócz attendance\_group zachowano również atrybut attendance), aby w razie zmiany w sposobie wyliczania przedziałów można było zmienić to dla danych historycznych. W kostce ze względu na zestawienia potrzebne były jednak już zgrupowane wartości.

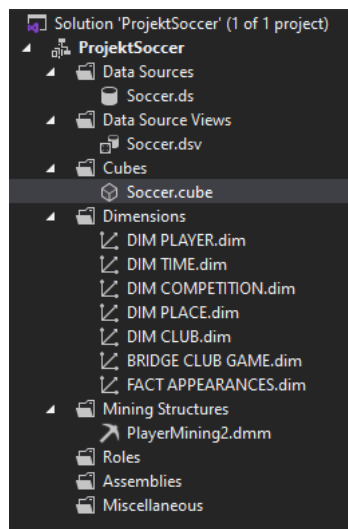
## Zakończony powodzeniem deployment kostki



## Zakończone powodzeniem przeprocesowanie kostki



## Widok utworzonych wymiarów i kostki



## Utworzone partycje

Item	Partition Name	Source	Estimated Rows	Storage Mode	Aggregation Design
1	FACT APPEARANCES	SELECT [Rosinska].[FACT_APPEARANCES].[appearance_id],[Ro...	0	MOLAP	
2	FACT APPEARANCES 2014-2015	SELECT [Rosinska].[FACT_APPEARANCES].[appearance_id],[Ro...	201115	MOLAP	AggregationDesign
3	FACT APPEARANCES 2016-2017	SELECT [Rosinska].[FACT_APPEARANCES].[appearance_id],[Ro...	263652	MOLAP	AggregationDesign 1
4	FACT APPEARANCES 2018-2019	SELECT [Rosinska].[FACT_APPEARANCES].[appearance_id],[Ro...	248783	MOLAP	AggregationDesign 2
5	FACT APPEARANCES 2020-2021	SELECT [Rosinska].[FACT_APPEARANCES].[appearance_id],[Ro...	256628	MOLAP	AggregationDesign 3
6	FACT APPEARANCES 2022-2023	SELECT [Rosinska].[FACT_APPEARANCES].[appearance_id],[Ro...	176169	MOLAP	AggregationDesign 4

Zdecydowano się na utworzenie 5 partycji, gdzie każda z nich obejmuje dwa lata rozgrywek piłkarskich. Partycje utworzone zostały z wykorzystaniem SQLa. Tworzenie nowych tabel przyspieszyłoby zapytania, ale schemat tabel źródłowych kostki bardzo straciłby na czytelności. Liczba rozegranych meczy jest dość zbliżona w każdym okresie dwuletnim, co przemawiało za takim podziałem na partycje.

## Utworzone miary kalkulowane

Script Organizer

Command

1 CALCULATE

2 [Average goal percentage]

3 [Average goals]

4 [Average assists]

5 [Average minutes played]

6 [Average yellow cards]

7 [Average red cards]

Calculation Tools

Metadata

Functions

Templates

Search Model

Measure Group:

<All>

Soccer

Measures

BRIDGE CLUB GAME

DIM PLAYER

DIM TIME

FACT APPEARANCES

Name:

[Average goals]

Parent Properties

Parent hierarchy: Measures

Parent member:

Change

Expression

[Measures].[Goals] / [Measures].[FACT APPEARANCES Count]

No issues found

Ln: 1 Ch: 57 SPC CRLF

Additional Properties

Format string:

Visible: True

Non-empty behavior:

Associated measure group: (Undefined)

Display folder:

Color Expressions

Font Expressions

## Utworzono KPI

**KPI Organizer**

- GoalsKPI
- AssistsKPI
- AgresionKPI
- GeneralAgresionKPI

**Calculation Tools**

- Metadata
- Functions
- Templates

**Measure Group:** <All>

**Measures:**

- Soccer
- Measures
- BRIDGE CLUB GAME
- DIM PLAYER
- DIM TIME
- FACT APPEARANCES

**KPI Configuration**

**Name:** GeneralAgresionKPI

**Associated measure group:** <All>

**Value Expression:** [Measures].[Average yellow cards] + [Measures].[Red Cards]

**Goal Expression:** 0.05

**Status:**

**Status indicator:**

**Status expression:**

```

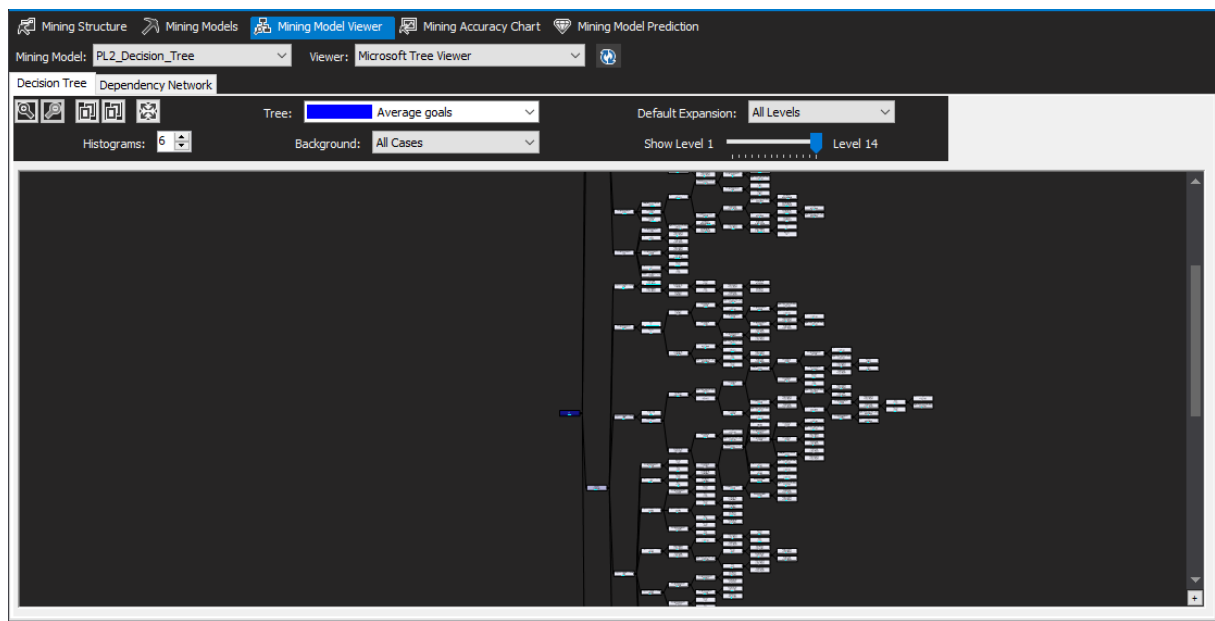
CASE
WHEN kpivalue('GeneralAgresionKPI') < kpigoal('GeneralAgresionKPI') then 1
WHEN kpivalue('GeneralAgresionKPI') > kpigoal('GeneralAgresionKPI') then -1
ELSE 0

```

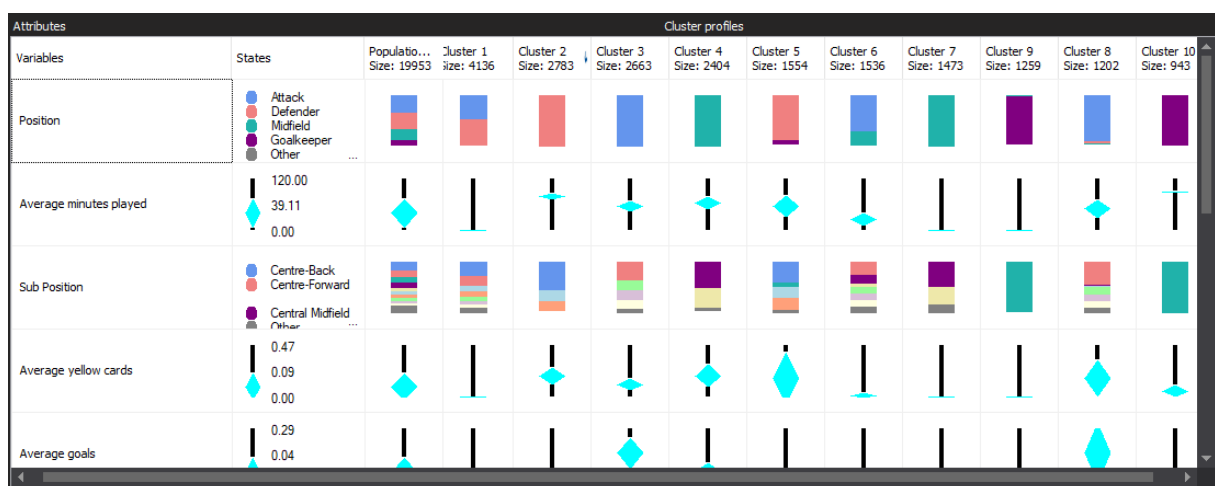
## Utworzony mining model

Structure	PL2_Decision_Tree	PL2_Clustering
	Microsoft_Decision_Trees	Microsoft_Clustering
Average assists	Input	Input
Average goals	Predict	Predict
Average minutes played	Input	Input
Average red cards	Input	Input
Average yellow cards	Input	Input
Foot	Input	Input
Heighest Market Value In E...	Input	Input
Height Group	Input	Input
Player Id	Key	Key
Position	Input	Input
Sub Position	Input	Input

## Widok fragmentu powstałego drzewa decyzyjnego



## Widok fragmentu profili utworzonych klastrow



## Wnioski:

Przy tworzeniu tabeli faktów istotne było pamiętać, by nie umieszczać w niej atrybutów liczbowych, które nie mogą zostać miarami. Przy tworzeniu kostki atrybuty liczbowe automatycznie przekształcane są w miary będące sumami. Dlatego konieczne było umieszczenie w FACT\_APPEARANCES aktualnej grupy wiekowej gracza `current_age_group` w postaci stringa przedstawiającego przedział. Gdyby umieszczony został bezpośrednio wiek jako liczba, możliwe byłoby stworzenie przez pomyłkę miary np. sumującej wiek graczy, co nie powinno mieć miejsca.

Przy tworzeniu miar należy zawsze zastanowić się nad sensem utworzonej miary. Nie można dopuścić do sytuacji, gdy miarą są sumy wartości nieaddytywnych (np. wartości procentowych).

KPI są bardzo przydatne przy ocenie sytuacji. Mogą zostać wykorzystane do oceny spełnienia pewnych warunków istotnych z punktu widzenia biznesu. W przypadku wartości zmiennych w czasie istotne są również trendy, które można wyznaczyć z pomocą KPI.

Przy tworzeniu partycji należy przemyśleć jak wyglądają dane i jaki podział tabeli faktów byłby sensowny. W przypadku rozpatrywanej tabeli faktów FACT\_APPEARANCES zdecydowano się na podział faktów w bloki dwuletnie. Można by również rozważyć partycje zawierające dane z jednego roku, jednak zdecydowano się na mniejszą fragmentaryzację danych.

Utworzono mining structure zawierające dwa modele. Wybrano model drzewa decyzyjnego oraz clustering. Na podstawie danych o graczach przewidywano jego średnie gole. Drzewo decyzyjne wykazało, że największy wpływ na liczbę goli gracza ma oczywiście pozycja, na której rozgrywa. Kolejnym najbardziej z rozpatrywanych czynników była średnia liczba minut spędzonych na boisku. Clustering pozwolił na przegląd danych. Utworzono 10 klastrów, gdzie najważniejszym kryterium podziału była pozycja piłkarza. Tak przygotowane dane zostaną punktem wyjściowym analizy.

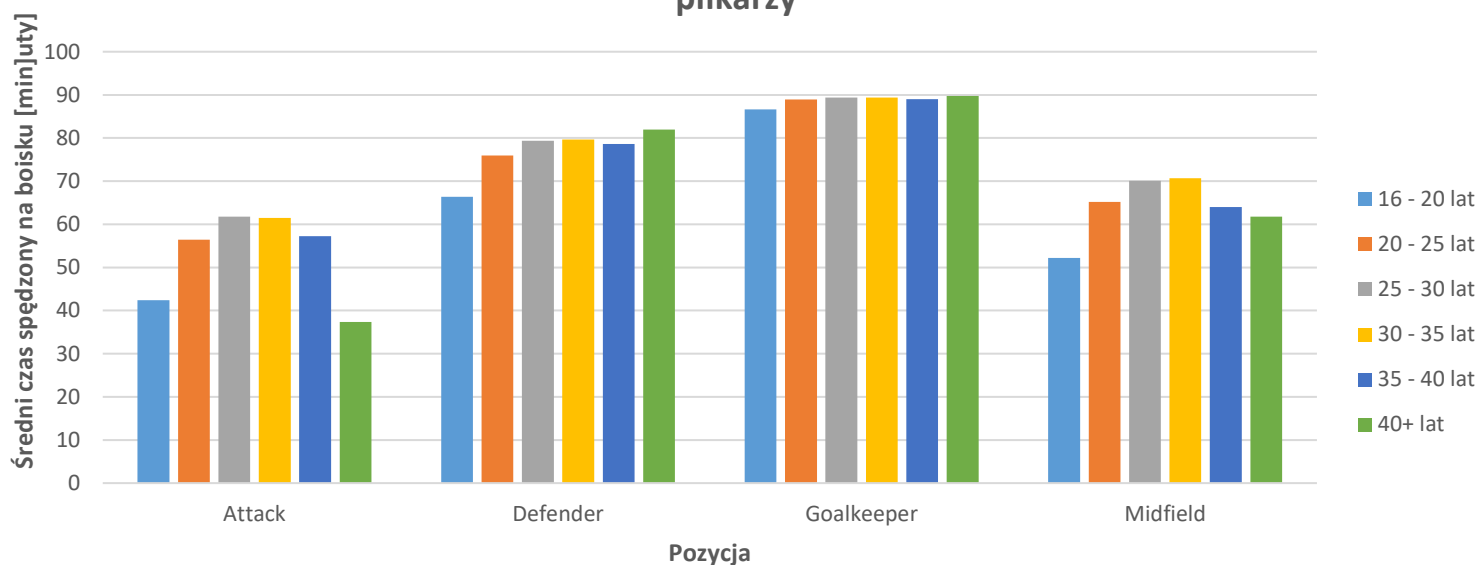
## Projekt – etap IV (19.06./22.06.)

### Prezentacja

Prezentacja powinna zawierać 4-8 slajdów (trwać ok. 8 minut) i wyjaśniać jakie dane są przedmiotem analizy. Prezentacja powinna być zakończona, krótką demonstracją, która pokaże najciekawsze związki między danymi znajdującymi się w kostce.

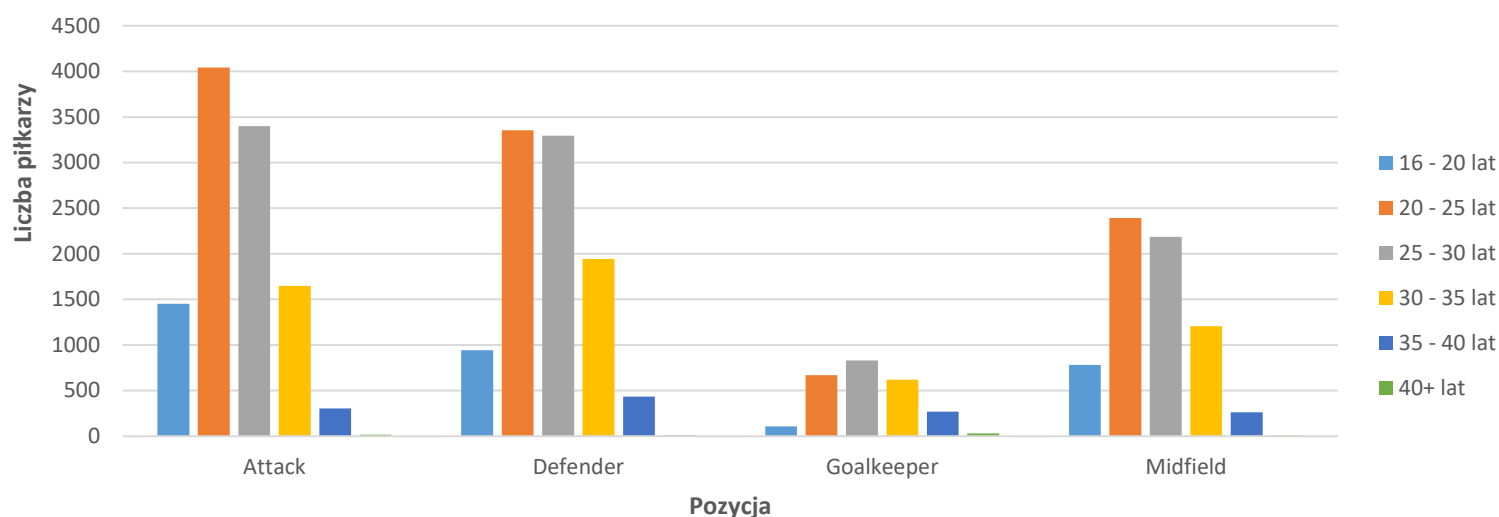
1. Jak wiek i pozycja graczy wpływają na czas spędzany na boisku? (Średnia liczba minut spędzonych na boisku w zależności od grupy wiekowej i pozycji).

Średni czas spędzany na boisku w zależności od grupy wiekowej i pozycji piłkarzy

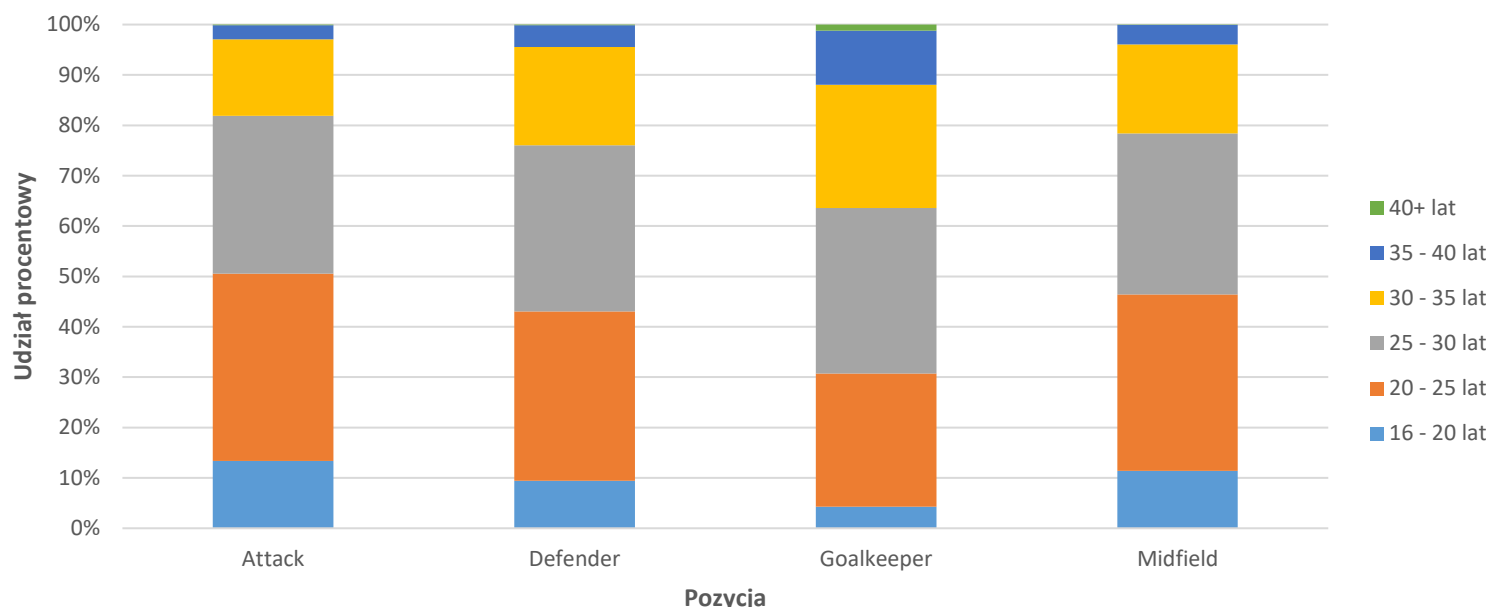


Player Id Distinct Count		Column Labels						Grand Total
Row Labels		16 - 20 lat	20 - 25 lat	25 - 30 lat	30 - 35 lat	35 - 40 lat	40+ lat	
Attack		1452	4041	3400	1648	303	16	6822
Defender		943	3355	3294	1945	434	11	6105
Goalkeeper		109	670	832	620	271	31	1565
Midfield		781	2393	2185	1208	264	5	4213
Grand Total		3285	10459	9711	5421	1272	63	18705

## Liczba piłkarzy w zależności od pozycji i grupy wiekowej



## Procentowy udział piłkarzy o danej grupie wiekowej wśród rozgrywających na danej pozycji



W celu rzetelnego przeanalizowania wpływu grupy wiekowej i pozycji piłkarzy na średni czas spędzany na boisku stworzono również zestawienia prezentujące liczebność poszczególnych grup.

Z zestawień wynika, że najliczniejszą grupą są piłkarze grający na pozycji atakującej, później obrońcy, pomocnicy i na końcu bramkarze. Widać dość duże wahania w udziale poszczególnych grup wiekowych w rozgrywających na danej pozycji. Atakujący są zdecydowanie grupą, w której widać zdecydowaną przewagę udziału młodszych piłkarzy. W ich przypadku grupa wiekowa 20 – 25 lat stanowi aż niemalże 40% wszystkich piłkarzy. Następnie widać spory udział grupy 25-30 lat oraz sporo już mniejszy udział grupy 30-35 lat. Udział piłkarzy starszych jest już zdecydowanie marginalny.

Podobnie wygląda sytuacja obrońców i pomocników. W ich przypadku nie ma już aż tak dużej przewagi młodszych zawodników i grupy wiekowe 20 – 25 lat i 25 – 30 lat są niemalże równoliczne. Udział piłkarzy w wieku 35 – 40 lat jest odrobinę mniejszy niż 5%, ale i tak jest to mocniejszy wpływ tej grupy wiekowej niż zaobserwowano u atakujących.



Najbardziej odmienny rozkład występuje u bramkarzy. Tylko u nich widać wyraźną przewagę grupy 25 – 30 lat nad grupą 20 – 25 lat. Co więcej, udział grupy wiekowej 30 – 35 lat jest praktycznie taki sam jak udział grupy wiekowej 20 – 25 lat. Również grupa 35 – 40 lat ma nieporównywalnie większy udział wśród bramkarzy względem innych pozycji sięga aż 10%. Pomimo tego, że bramkarze są najmniej liczną grupą zaobserwowano u nich największą liczbę zawodników po 40 roku życia ze wszystkich pozycji.

Najmniejsze wahania w średnim czasie spędzonym na boisku w zależności od wieku zaobserwowano dla bramkarzy. Dla wszystkich grup wiekowych średnia wynosiła niemalże 90 minut, czyli cały czas trwania meczu. W ich przypadku grupa wiekowa wydawała się nie mieć aż takiego znaczenia. Odrobinę mniejszą wartość średnią osiągnęła grupa wiekowa 16 – 20 lat, wciąż była ona jednak wyższa niż 85 minut. Można zatem przypuszczać, że bramkarz to bardzo odmienna pozycja od reszty. Nie tylko gromadzi głównie starszych zawodników, ale również spędzają oni właściwie cały czas trwania meczu na boisku. Ma to zapewne związek ze specyfiką pozycji, która wymaga mniej biegania i intensywnego wysiłku fizycznego rozłożonego na cały czas trwania gry. Bramkarze muszą za to dobrze rozumieć grę, mieć dobre wyczucie i dobrą kondycję, tak żeby jednak pozwalała im ona na gwałtowne uaktywnienie się w potrzebnych momentach. Z tego względu warto wybierać na tę pozycję bardziej doświadczonych graczy. Starsi zawodnicy i tak spędzą statystycznie tyle samo, a nawet odrobinę więcej, czasu na boisku co młodszy, a będą mieli potrzebne wyczucie i zrozumienie rozgrywki. W przypadku bramkarzy wybór młodszych zawodników nie wydaje się mieć poparcia danymi z zestawień jeżeli chodzi o czas gry. Ryzykowne jest jednak sięganie po zawodników powyżej 40 roku życia, gdyż pomimo większego udziału niż na innych pozycjach, dane zgromadzone na podstawie 31 bramkarzy powyżej 40 roku życia wydają się niewystarczające do wyciągnięcia pełnych przekonań wniosków.

Największe wahania występują dla atakujących. Jest to również pozycja, która ma najmniejsze średnie czasy przebywania na boisku. Dla żadnej z grup wiekowych średnia nie przekracza 65 minut. Jest to wymagająca fizycznie pozycja, ale wbrew pozorom to wcale nie najmłodsze grupy wiekowe spędzają średnio najwięcej czasu na boisku. W grupie wiekowej 16 – 20 lat średni czas nie przekracza nawet 45 minut, czyli połowy meczu. Grupa 20 -25 lat oscyluje około średnio 55 minut, a peak następuje dla grupy 25 – 30 lat z trochę ponad 60 minutami. Dla grupy 30 – 35 lat następuje delikatny spadek, jednak nawet starsi piłkarze z grupy 35 – 40 lat z kolejnym spadkiem wciąż spędzają nie mniej czasu na boisku niż grupy poniżej 25 roku życia. Zatem również dla atakujących wcale wybór najmłodszego zawodnika nie wydaje się być najkorzystniejszy ze względu na czas spędzany na boisku. Ze względu na specyfikę pozycji i tak z danych wynika, że niezależnie od grupy wiekowej miejsce mają zmiany na boisku. Można zatem rekrutować starszych, bardziej doświadczonych piłkarzy, bo i tak w przypadku braków kondycyjnych związanych z wiekiem nastąpi zmiana. Grupa 40+ jest tak mało liczna, że trudno o rzetelną analizę. Jednak przy 16 badanych piłkarzach widać było gwałtowny spadek w czasie na boisku, który ledwo przekraczał średnie 35 minut. Może to być zatem znak, że ryzykowne jest zatrudnianie atakujących zbliżających się do 40 roku życia.

Dość podobnie wygląda rozkład średniego czasu spędzanego na boisku dla pomocników. W ich przypadku wspinanie po średnim czasie na boisku rozpoczyna grupa 16 – 20 lat z lekko ponad 50 minutami. Następnie grupa 20 – 25 lat skacze do 65 minut, a peak następuje dla grup 30 – 35 lat i 25 – 30 lat z wynikiem rzędu 70 minut. Następnie ma miejsce przełamanie i grupa 35- 40 lat spędza nieznacznie mniej czasu na boisku niż piłkarze w wieku 20 – 25 lat. Również dla pomocników muszą występować regularne zmiany, na co wskazuje średnia liczba minut. Musi to być jednak albo mniej angażująca fizycznie albo mniej warta zmian pozycja niż atakujący, bo pomimo wzrostu średniego czasu na boisku o około 10 minut zależności między grupami wiekowymi zmieniły się jedynie nieznacznie.

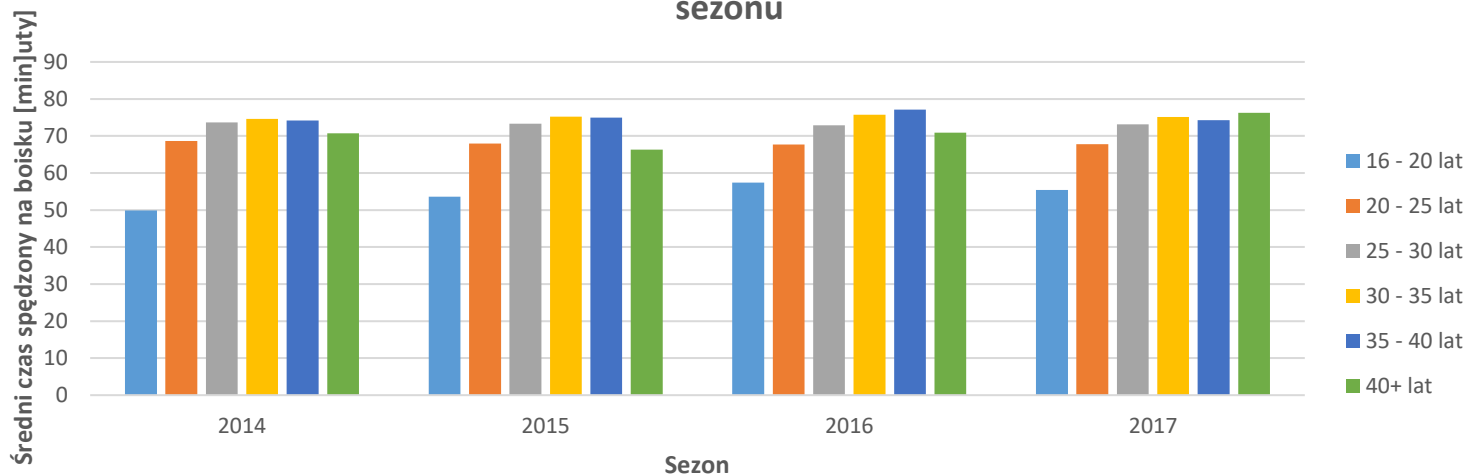
W przypadku obrońców średni czas spędzany na boisku rozpoczyna 65 minut dla grupy 16 – 20 lat. Mocny wzrost ma miejsce dla 20 – 25 lat z 75 minutami. Dla starszych grup wiekowych następują kolejne delikatne wzrosty nieprzekraczające 80 minut. Grupa wiekowa 40+ jest bardzo mało liczna (11 piłkarzy) i nie można brać jej wyniku ponad 80 minut za miarodajną wskazówkę przy wyborze piłkarzy

do składu. Zmiany tutaj muszą występować rzadziej niż wśród atakujących i pomocników, ale nie ma to również odzwierciedlenia w średnim czasie spędzonym na boisku przez starszych piłkarzy.

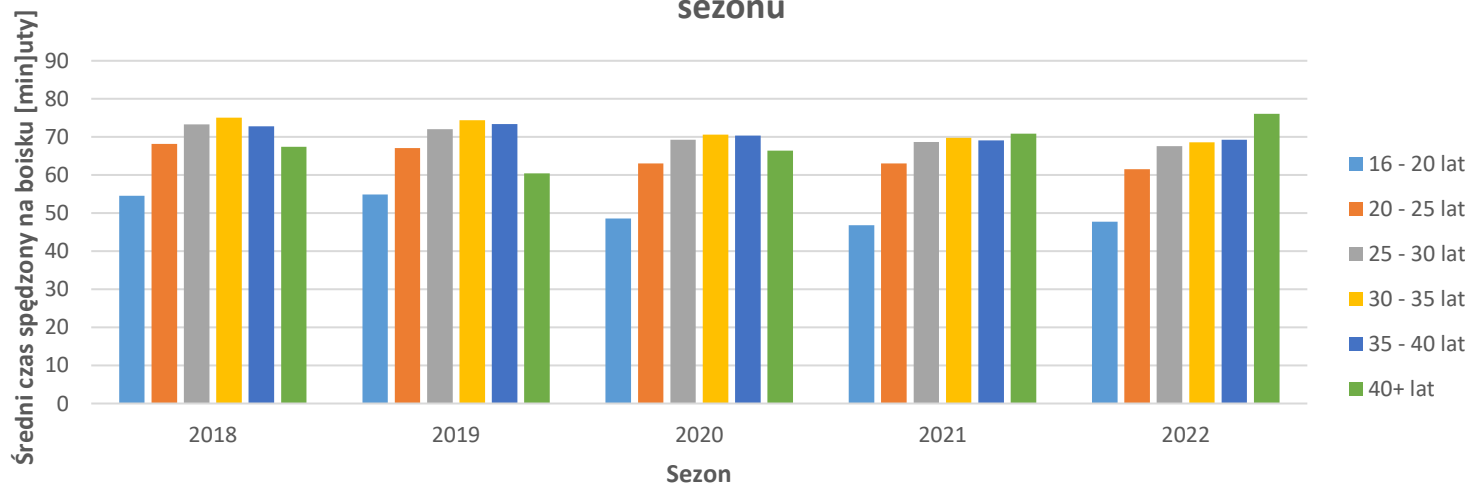
Patrząc ogólnie, zdecydowanie najliczniejszą grupą (poza bramkarzami) są zawodnicy w wieku 20 – 25 lat. Wbrew temu jednak na żadnej z pozycji nie spędzają średnio najwięcej czasu. Może to mieć związek z brakiem doświadczenia i np. koniecznością zmiany zawodnika w nowych sytuacjach, którym nie jest w stanie sprostać. Na pewno jednak analizowane zestawienia nie pozwalają na stwierdzenie, jakoby starsi piłkarze spędzali jednoznacznie mniej czasu na boisku, a wręcz przeciwnie. Może to wynikać zarówno z tego, że pozycja nie wymaga aż tak ogromnego wysiłku fizycznego przez cały czas gry, ale również ze względu na następujące zmiany w składzie w trakcie meczu. Widać jednak bardzo wyraźne spadki liczebności grup starszych. Poza bramkarzami mowa tu o nieporównywalnie mniejszej reprezentacji grup starszych. Można zatem wnioskować, że piłkarze, którzy kontynuują swoją karierę pomimo wieku są selekcyonowani naturalnie przez swoją kondycję, co zaburza wyniki. Karierę kontynuują najzdolniejsi i najwydolniejsi piłkarze, co naturalnie sprawia, że są bardziej pożądanymi podczas wyboru składu na mecz. Na pewno przełamanie następuje zbliżając się do 40 roku życia. Reprezentacja tych grup jest tak mała, że ryzykowne wydaje się powoływanie piłkarzy w tym wieku do składu, gdy dostępni są młodsi, dobrze rokujący zawodnicy.

2. Jak sezon i wiek graczy wpływają na czas spędzany na boisku? (Średnia liczba minut spędzonych na boisku w zależności od grupy wiekowej i sezonu).

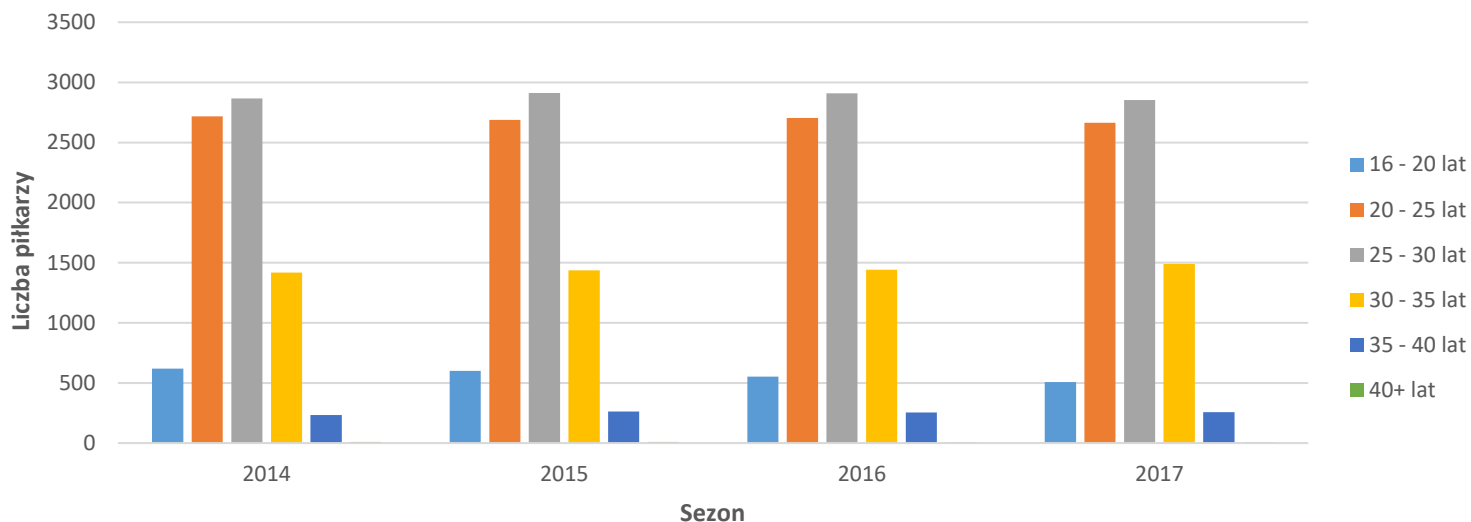
**Średni czas spędzany na boisku w zależności od grupy wiekowej piłkarzy i sezonu**



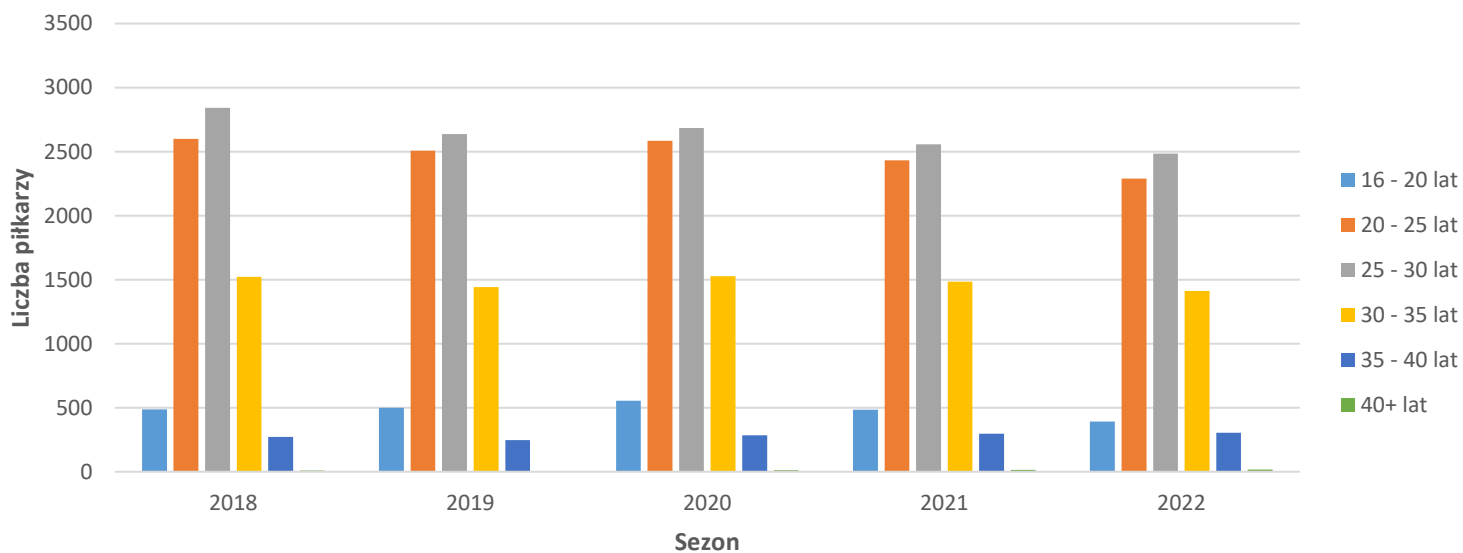
**Średni czas spędzany na boisku w zależności od grupy wiekowej piłkarzy i sezonu**



### Liczba piłkarzy w zależności od sezonu i grupy wiekowej



### Liczba piłkarzy w zależności od sezonu i grupy wiekowej



Player Id Distinct Count		Column Labels						Grand Total
Row Labels		16 - 20 lat	20 - 25 lat	25 - 30 lat	30 - 35 lat	35 - 40 lat	40+ lat	
2014		618	2717	2866	1419	233	10	6836
2015		601	2688	2911	1436	262	10	6843
2016		552	2703	2909	1442	254	8	6853
2017		506	2663	2852	1491	257	8	6798
Grand Total		1726	5928	5645	2999	601	27	12239

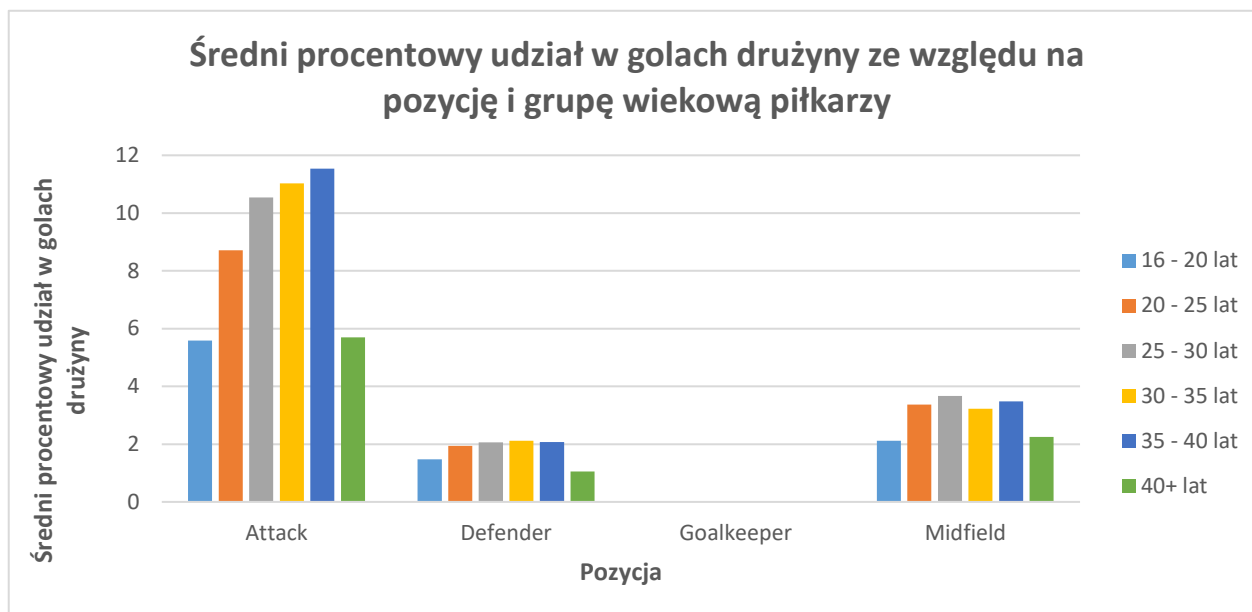
Player Id Distinct Count	Column Labels						
		20 - 25	25 - 30	30 - 35	35 - 40	40+	
Row Labels	16 - 20 lat	lat	lat	lat	lat	lat	Grand Total
2018	488	2599	2841	1521	271	10	6749
2019	499	2507	2638	1443	246	5	6417
2020	555	2584	2684	1526	285	12	6605
2021	485	2432	2557	1485	297	14	6224
2022	393	2290	2485	1413	304	18	5927
Grand Total	1737	6385	6193	3527	826	39	12902

Na przestrzeni sezonów można zaobserwować spadek liczby graczy, zwłaszcza z najbardziej wybijających się grup wiekowych. Dla sezonów do 2018 liczba piłkarzy w wieku 25 – 30 lat sięgała niemalże 3 tysięcy. Następnie zaczęła opadać i już w sezonie 2022 spadła poniżej 2500 piłkarzy. Ogólnie, dla każdej z grup wiekowych liczba piłkarzy spadła na przestrzeni lat. Wyjątkiem są grupy 35-40 lat i 40+ lat, jednak są one tak mało liczne, że mowa tutaj o trendzie, który dotyczy mimo wszystko znikomej liczby piłkarzy. Może to wskazywać na to, że albo część lig stała się mniej opłacalna i została zamknięta przez co rozgrywano mniej meczy lub też drużyny postawiły na pewniejszych zawodników, rzadziej dokonując zmian w składach, przez co część piłkarzy straciła pracę w rozgrywkach europejskich.

Na przestrzeni sezonów zaobserwowano także zmiany związane z czasem spędzonym na boisku. Dla żadnego z sezonów wartości maksymalne nie przekroczyły 80 minut. Jednak dla grup wiekowych 20 – 25, 30 – 35 i 35 – 40 lat na przestrzeni sezonu 2020 średni czas spędzany na boisku spadł poniżej kolejnego progu i nie wynosił już nawet 70 minut. Temu trendowi nie poddawała się bardzo mało liczna grupa 40+, którą ciężko poddawać obiektywnej analizie. Największe spadki miały miejsce dla grupy 20 – 25 lat, gdzie ze średnich około 70 minut wartości poszybowwały do średnich 60 minut.

Można zatem zauważyć, że pomimo spadku liczebności piłkarzy spadły średnie czasy spędzane na boisku. Można wykluczyć zatem teorię, że piłkarze byli rzadziej zmieniani, co przełożyło się na spadek ich liczebności. Być może przez pandemię nie wszystkie drużyny zdołały się utrzymać i albo całych zespołów jest mniej, albo mają mniejsze zmiany w całościowym składzie zespołu, ale za to większe zmiany w trakcie rozgrywek na boisku. Ponadto widać, że z biegiem czasu zaczęto przesuwac ciężar z młodych zawodników w wieku 20 – 25 lat. Można zatem wnioskować, że zmieniona została taktyka. Postawiono również na trochę starszych graczy, jednocześnie bardziej rozkładając czas spędzany na boisku pomiędzy różnymi zawodnikami drużyny. Warto zatem przy tworzeniu składu zbadać te trendy i rozważyć wcielanie do drużyn zawodników starszych (choć wcale nie powiedziane, że o gorszej kondycji), jednocześnie rozkładając ciężar meczu na więcej niż jednego zawodnika.

3. Jak pozycja i wiek graczy wpływają na procentowy udział w golach drużyny? (Średni procentowy udział w golach drużyny w zależności od pozycji i grupy wiekowej).

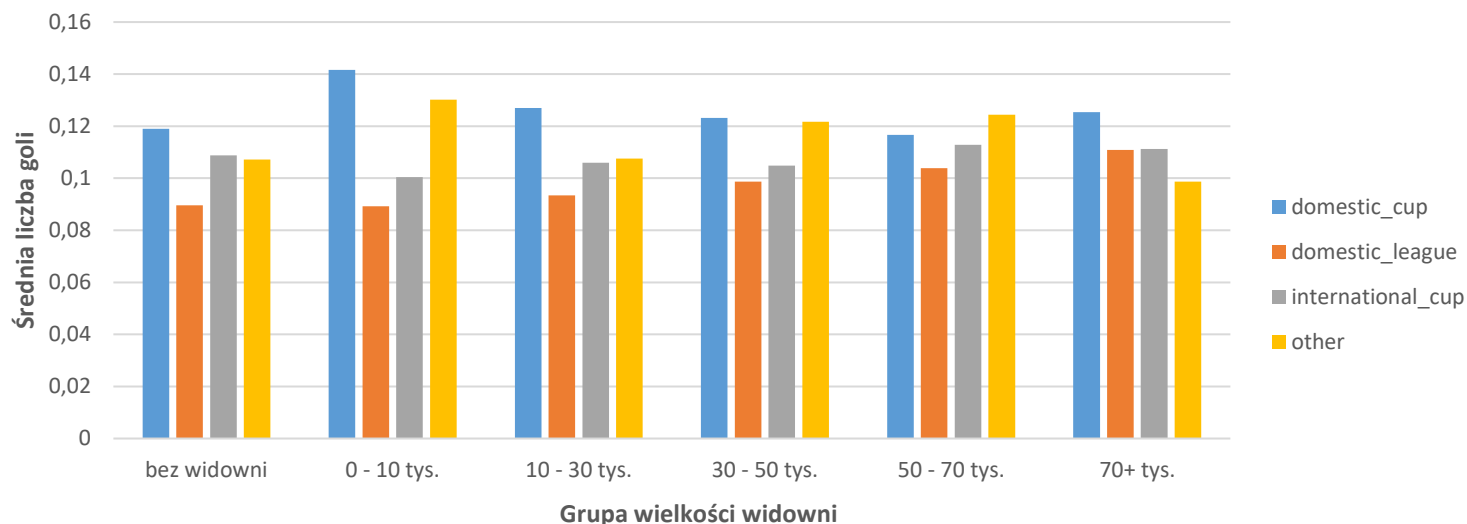


Gole strzelane przez bramkarzy są tak znikome dla wyników drużyn, że ich wykres niczego nie przedstawia. Później od końca najważniejsi w kwestii udziału w golach są obrońcy, pomocnicy i atakujący, co zgadza się z zasadami podziału pozycji pomiędzy piłkarzy. Zarówno dla obrońców jak i dla pomocników nie zauważono związku pomiędzy wiekiem, a udziałem w golach. Można zatem wnioskować, że albo umiejętności tych zawodników nie mają aż tak związku z wiekiem, albo gole strzelane przez te pozycje są po prostu strzelane okazyjnie i zależą bardziej od sytuacji na boisku, niż od piłkarzy.

Dla atakujących sytuacja rysuje się bardzo jasno. Przede wszystkim są oni głównym źródłem cennych dla zwycięstwa drużyny goli. Widać bardzo dużą zależność pomiędzy wiekiem, a procentowym udziałem w golach drużyny. Zawodnicy starsi są albo bardziej doświadczeni, przez co łatwiej im o szukanie sytuacji i oddawanie celnych strzałów, albo po prostu jeżeli wciąż zostają w składach drużyn (po tym jak się prezentowali za młodu), są już sprawdzeni jako dobrzy rozgrywający. Zestawiając te dane z wnioskami związanymi ze średnim czasem spędzonym na boisku, warto przy tworzeniu drużyn wcielać do niej starszych zawodników. Oczywiście aby mieć starszych zawodników najpierw trzeba inwestować w najmłodszych i odpowiednio dawać im się rozwijać na boisku, jednak zarówno pod względem czasu gry jak i wyników starsi doświadczeni zawodnicy wydają się być bardzo dobrym wyborem. Jest to szczególnie istotne przy trendach, które badano w związku z czasem gry na przestrzeni sezonów. Wybór starszych zawodników bardzo dobrze łączy się ze zmianami podczas meczy, które rozkładają ciężar fizyczny na więcej niż jednego zawodnika.

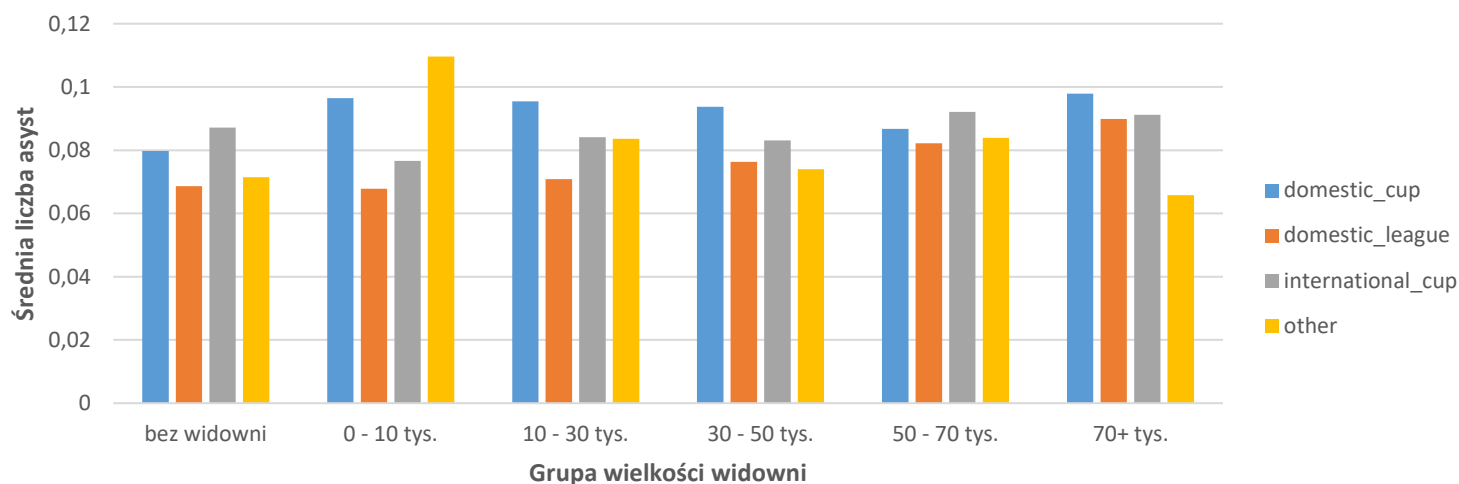
4. Jak typ konkursu i liczba kibiców wpływają na liczbę goli strzelanych przez graczy? (Średnia liczba goli w zależności od grupy wielkości widowni i typu konkursu).

Średnia liczba goli ze względu typ konkursu i grupę wielkości widowni



5. Jak poziom rozgrywek i liczba kibiców wpływają na liczbę asyst graczy? (Średnia liczba asyst w zależności od grupy wielkości widowni i poziomu rozgrywek).

Średnia liczba asyst ze względu typ konkursu i grupę wielkości widowni



W przypadku „domestic league” widać wyraźny wzrost średniej liczby goli i asyst wraz ze wzrostem wielkości widowni obecnej na meczu. W przypadku tego typu konkursu można przypuszczać, że wielkości grono kibicującego ma duży wpływ na wyniki piłkarzy. Informacje te można wykorzystać np. w przypadku zakładów bukmacherskich, obstawiając wyższe wyniki w przypadku zawodów z większą widownią.

W przypadku „domestic cup” ta zależność jest już bardziej rozmyta i ciężko wykazać zależność pomiędzy wielkością widowni a wynikami drużyn. Występują tutaj także największe średnie liczby goli, co co potwierdza, że są to zawody o innym charakterze. W „domestic cup” mogą brać udział również drużyny, które nie są w pełni profesjonalne. Może to być powodem zawyżonych i chwiejnych statystyk goli i asyst, gdyż profesjonalne drużyny uzyskują lepsze wyniki grając na drużyny niższej klasy, niż uzyskałyby grając na drużynę o podobnym charakterze. Wybijają się średnia liczba goli dla małych spotkań o widowni z przedziału 0 – 10 tys., co tym bardziej wskazuje na zawyżenie wyników przez potyczki drużyn niższej klasy.

Dla „international cup” widać delikatne tendencje wzrostowe statyk wraz ze wzrostem wielkości widowni. Są to konkursy o ogólnie bardzo dużej stawce, prawdopodobnie również powszechnie oglądane w telewizji. Wahania ze względu na wielkości widowni następują jednak są naprawdę delikatne, co ma zapewne związek z bardzo wysokim poziomem rozgrywek i trudnością rywalizacji.

Wpływ widowni widać dla rozgrywek, które rozgrywano zupełnie bez niej. W danych znajdują się informacje o meczach rozgrywanych w trakcie pandemii gdzie żadna publika nie była obecna na stadionie. Dla „domestic league” i „domestic cup” uzyskiwano bez widowni zdecydowanie gorsze wyniki. Dla konkursów international cup i other nie zauważono aż takich spadków w przypadku braku oglądających na żywo, ale może to mieć związek powszechnym oglądaniem przez widzów.

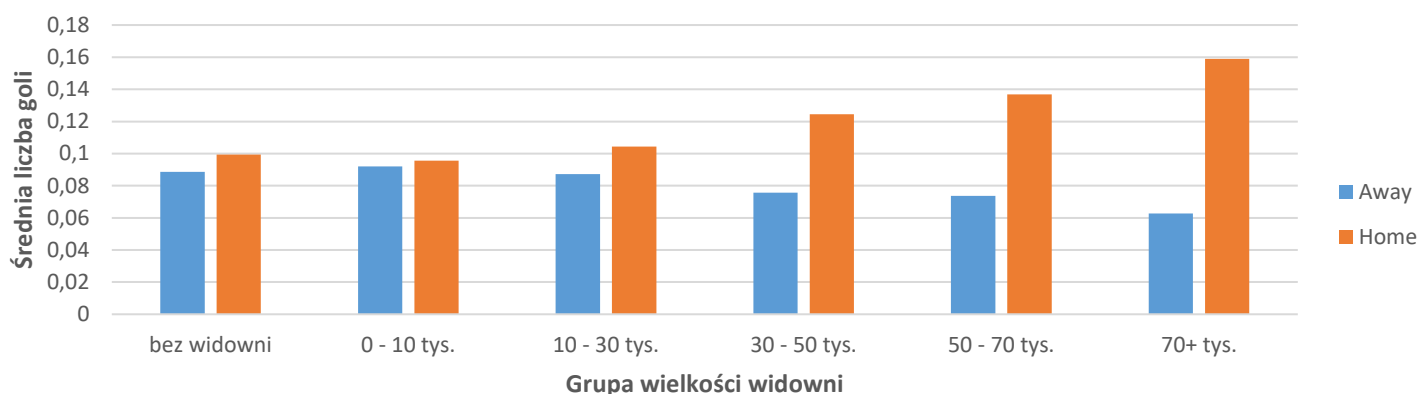
Można zauważyć, że dla większych widowni jest więcej asyst w przypadku konkursów „domestic league” i „international cup”. Wyjątkiem jest typ konkursu „other”. Dla innych przypadków jednak można wnioskować, że gole wymagają średnio zaangażowania większej liczby piłkarzy niż jedynie strzelającego gola. Warto zatem przy szykowaniu się na spotkanie omówić odpowiednią taktykę. Przygotować się na to, że przy większej publice atakujący będą potrzebowali większego wsparcia, żeby strzelić gola.

W tych wynikach należy wziąć pod uwagę, że średnia jest rozpatrywana ze względu na obydwie strony spotkania. Wyniki może zaburzać fakt, że nie ma rozróżnienia na drużynę, której publika sprzyja i tą przyjeżdżną, która nie otrzymuje wsparcia. Zostanie to przeanalizowane w kolejnym zestawieniu.

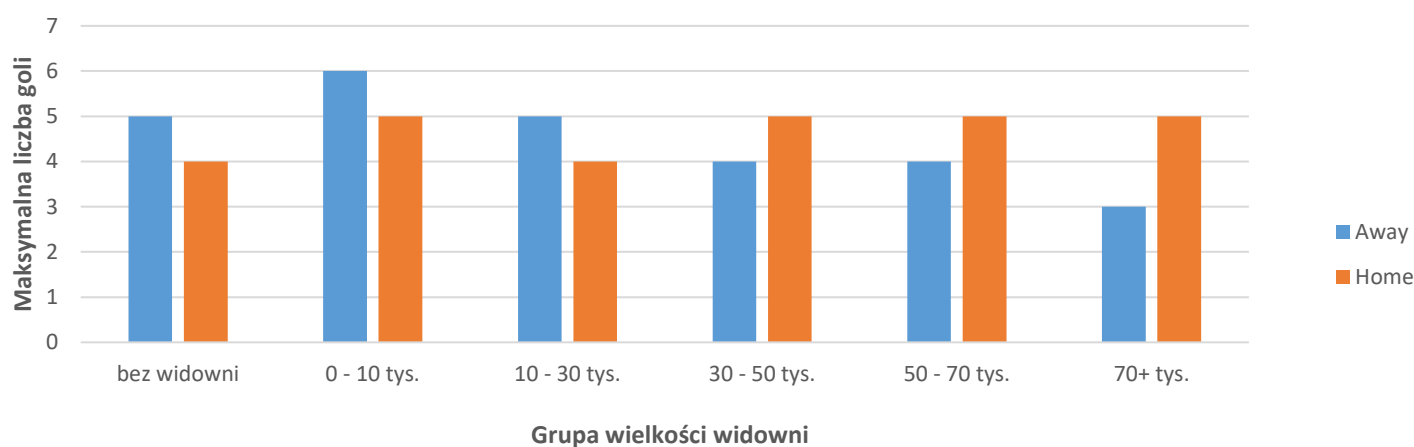
6. Jak doping wpływa na liczbę goli strzelanych przez graczy?

- (Średnia liczba goli w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).
- (Maksymalna liczba goli w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).

**Średnia liczba goli ze względu na grupę wielkości widowni i to czy drużyna jest gospodarzem**



**Maksymalna liczba goli ze względu na grupę wielkości widowni i to czy drużyna jest gospodarzem**



Przy powyższym zestawieniu widać już bardzo wyraźnie wpływ widowni na średnią i maksymalną liczbę goli. Średnia liczba goli niezależnie od liczebności widowni większa jest dla gospodarzy. Różnica pomiędzy średnimi golami gospodarzy i przyjezdnych rośnie jednak wraz ze wzrostem publiki. Dla publiki rzędu 70 tys. jest to ponad dwukrotna różnica. Zależność widać również dla maksymalnej liczby goli. W przypadku braku widowni i widowni małych (do 30 tysięcy), widać przewagę wyniku przyjezdnych. Jednak wraz ze wzrostem publiki na żywo ten wynik zmienia się na korzyść gospodarzy, a dla spotkań z największą publiką maksymalny wynik przyjezdnych jest już najmniejszy ze wszystkich zanotowanych.

Można zatem stwierdzić, że w przypadku zakładów bukmacherskich bezpieczniej jest stawiać na gospodarzy. Trenerzy i piłkarze powinni być szczególnie dobrze przygotowani na spotkania, gdy

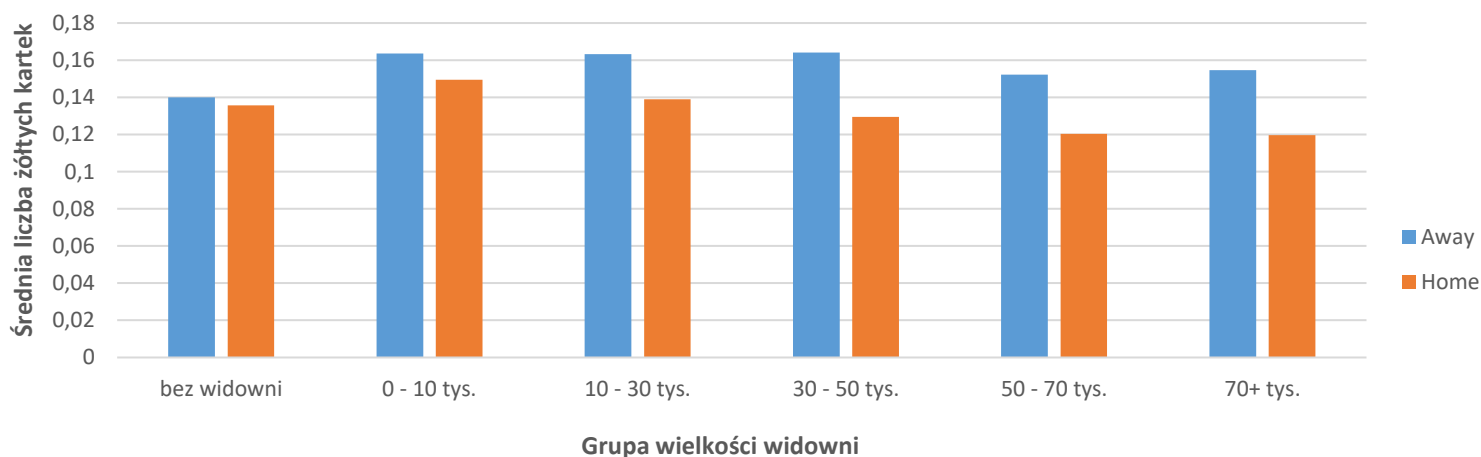


drużyna nie jest gospodarzem. Warto jest rozważyć wprowadzenie różnego rodzaju spotkań motywacyjnych i współpracy z coachami, mentorami czy psychologami sportowymi, którzy pomogą piłkarzom w rozwoju osobistym i pokonaniu bariery mentalnej związanej z publiką i dopingowaniem przeciwników. Warto też w przypadku wielu różnego rodzaju spotkań szczególnie przygotować się na mecze na terenie rywala, tak żeby piłkarze czuli się pewnie pomimo niesprzyjających warunków.

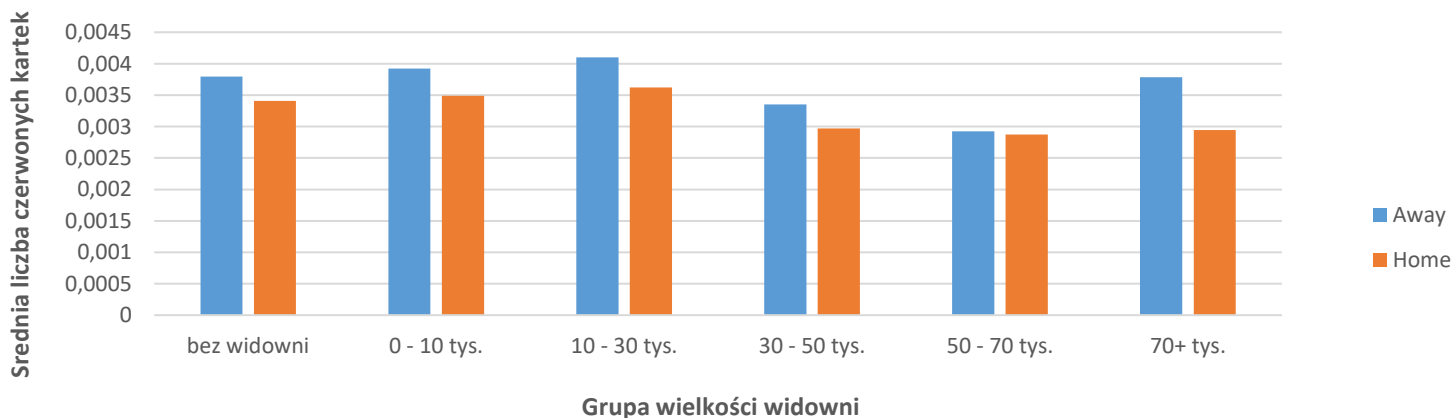
#### 7. Jak doping wpływa na kulturę gry?

- (Średnia liczba żółtych kartek w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).
- (Średnia liczba czerwonych kartek w zależności od grupy wielkości widowni na stadionie i tego czy drużyna jest gospodarzem).

**Średnia liczba żółtych kartek ze względu na grupę wielkości widowni i to czy drużyna jest gospodarzem**



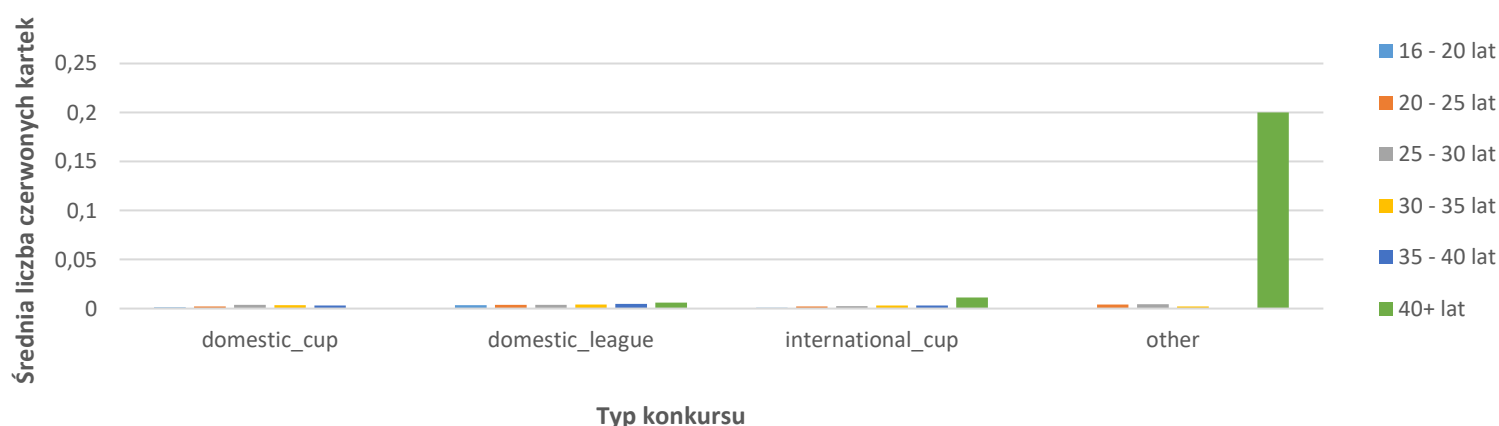
**Średnia liczba czerwonych kartek ze względu na grupę wielkości widowni i to czy drużyna jest gospodarzem**



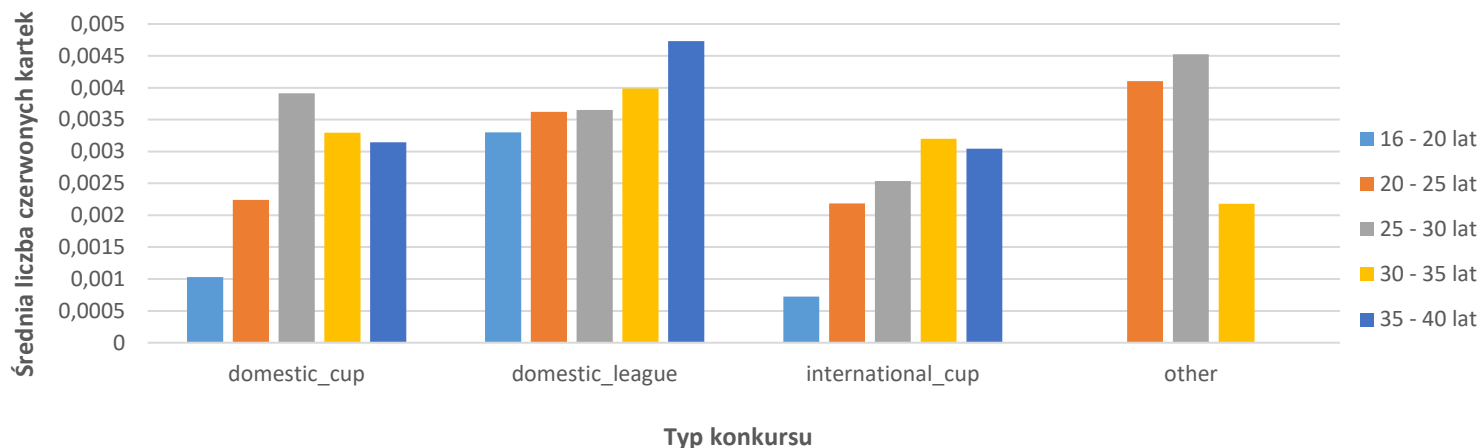
Na tej podstawie można wnioskować, że wielkość widowni wpływa na kulturę gry piłkarzy. Piłkarze grają bardziej kulturalnie w przypadku meczy odbywających się na ich terenie, a zatem z prawdopodobnie sprzyjającą im widownią. Jednak ogólnie im większa widownia, tym średnio wyższa kultura gry. Może to mieć również związek z tym, że większą publikę mają zazwyczaj mecze o większą stawkę, więc ryzykowanie problemów przez złe zachowanie jest wykluczone. Jest to jednak dość zadziwiające, biorąc pod uwagę, że takie mecze wiążą się również z większymi emocjami, zwłaszcza gdy dopingowani są przeciwnicy.

8. Jak typ konkursu w zależności od doświadczenia życiowego (wieku) gracza wpływa na kulturę gry?

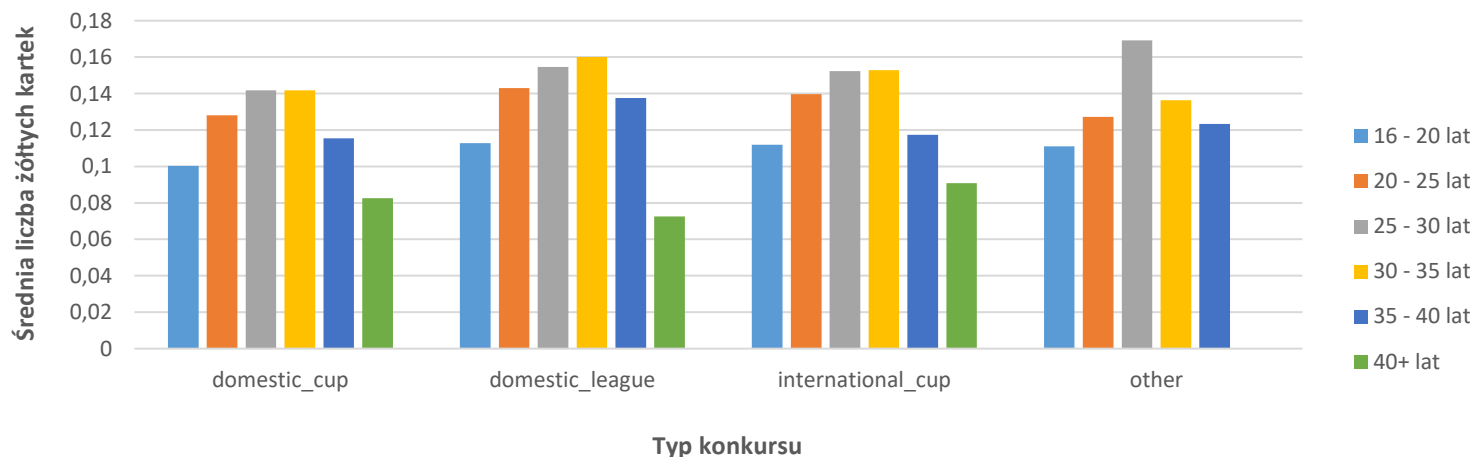
- Średnia liczba czerwonych kartek ze względu na typ konkursu i grupę wiekową  
piłkarzy**



### Średnia liczba czerwonych kartek ze względu na typ konkursu i grupę wiekową piłkarzy



### Średnia liczba żółtych kartek ze względu na typ konkursu i grupę wiekową piłkarzy



Zdecydowano się usunąć z zestawienia o średniej liczbie czerwonych kartek dane dotyczące grupy wiekowej 40+ ze względu na zaburzenie czytelności wykresu. Jest to również bardzo mało liczna grupa piłkarzy, o której informacje ciężko analizować ze względu na znikomą ilość informacji. Przedstawiono obie wersje zestawienia.

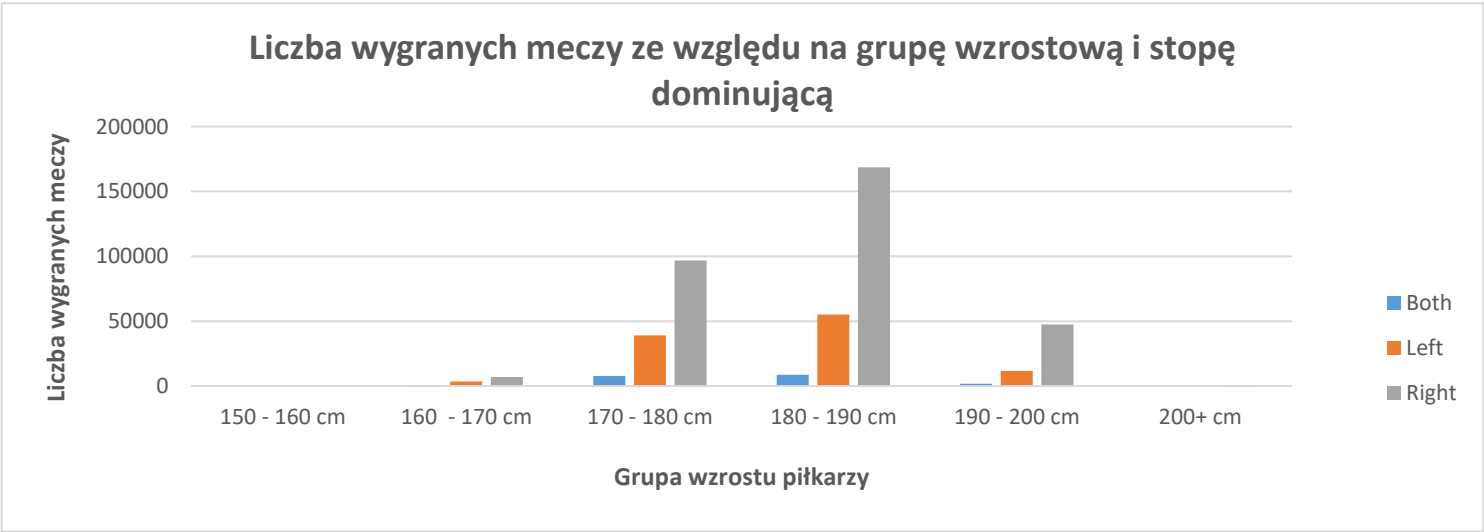
W przypadku czerwonych kartek widać tendencję, że dla różnych typów konkursów z pojedynczymi wyjątkami średnio więcej czerwonych kartek otrzymują starsi piłkarze. Podobnie wygląda sytuacja średniej żółtych kartek. Wartości rosną przez grupy 16 – 20 lat, 20 – 25 lat, 25 – 30 lat i przy grupie 30 – 35 lat utrzymują się na tym samym poziomie lub zaczynają z powrotem spadać. Można podejrzewać, że ma to związek z chociażby doświadczeniem piłkarzy. Bardziej doświadczeni piłkarze odważają się na bardziej ryzykowne zagrania, które mogą skończyć się przyznaniem kartki.

Nie zauważono zależności pomiędzy liczbą przyznawanych żółtych kartek, a typem konkursu. Inaczej jednak w przypadku kartek czerwonych. Zawody międzynarodowe „international cup” charakteryzują się mniejszą średnią liczbą czerwonych kartek w każdej grupie wiekowej. „Domestic cup” wypada pod tym względem bardziej agresywnie, ale głównie w grupie wiekowej 25 – 30 lat. Średnio najwięcej czerwonych kartek przyznaje się w „domestic league”.

Na tej podstawie można typować wyniki w zakładach bukmacherskich, dostosowując przewidywania zależnie od konkursu i grupy wiekowej. Można prowadzić również szkolenia piłkarzom związane z opracowywaniem taktyki, tak aby unikać czerwonych kartek niezależnie od typu konkursu, w którym bierze się udział.

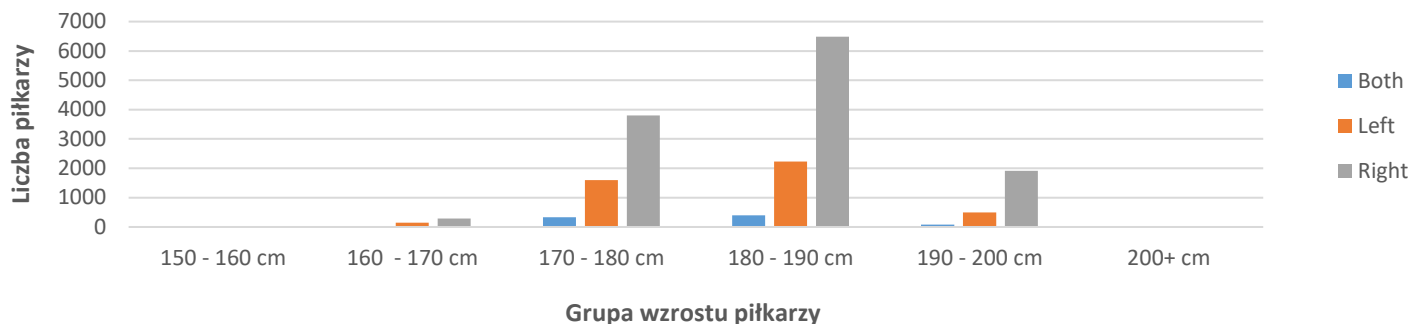
9. Jak warunki fizyczne wpływają na wyniki graczy? (Liczba zwyciężonych meczy ze względu na stopę dominującą i wzrost).

Is Win Row Labels	Column Labels			
	Both	Left	Right	Grand Total
150 - 160 cm			14	14
160 - 170 cm	598	3484	7023	11105
170 - 180 cm	7795	39147	96880	143822
180 - 190 cm	8812	55101	168585	232498
190 - 200 cm	1742	11779	47450	60971
200+ cm		406	622	1028
Grand Total	18947	109917	320574	449438

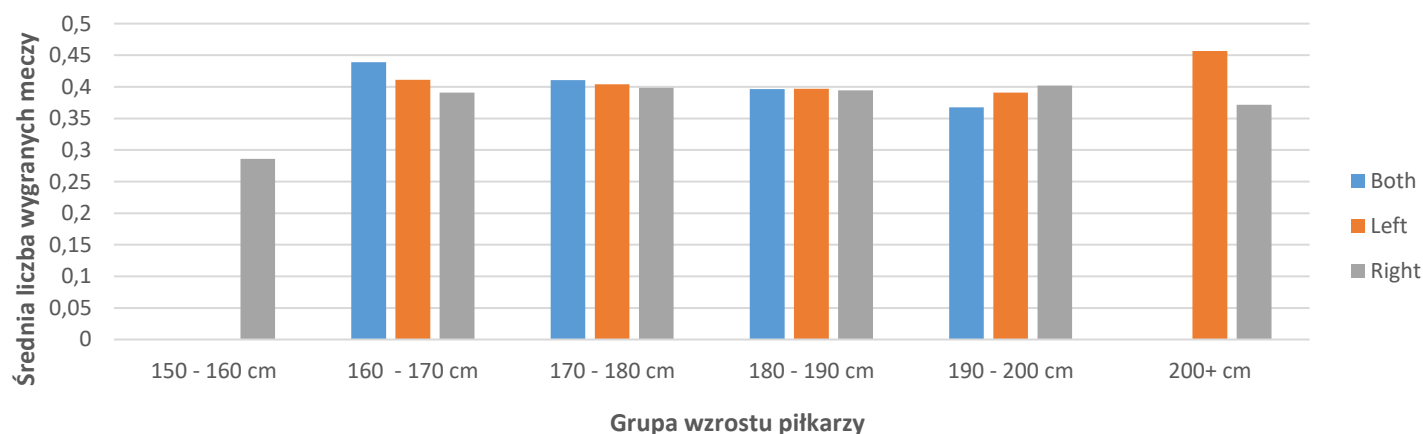


Player Id Distinct Count Row Labels	Column Labels			
	Both	Left	Right	Grand Total
150 - 160 cm			1	1
160 - 170 cm	20	147	293	460
170 - 180 cm	327	1600	3804	5731
180 - 190 cm	395	2234	6487	9116
190 - 200 cm	84	497	1916	2497
200+ cm		12	32	44
Grand Total	826	4490	12533	17849

### Liczba piłkarzy ze względu na grupę wzrostową i stopę dominującą



### Średnia liczba wygranych meczy ze względu na grupę wzrostową i stopę dominującą



Grupy wzrostowe 150 – 160 cm i 200+ cm są bardzo mało liczne, co trzeba wziąć pod uwagę. Wyniki z nimi związane mogą nie tyle odnosić się do specyfiki graczy o takich warunkach fizycznych, ale do konkretnych jednostek.

Najwięcej piłkarzy jest prawonożnych. Zdecydowanie mniej jest piłkarzy o dominującej stopie lewej, a najrzadziej piłkarze mają obie stopy dominujące. Ze względu na takie dysproporcje zdecydowano się skupić przede wszystkim na ocenie zestawienia dotyczącego średniej liczby wygranych meczy. Suma wygranych meczy dla piłkarzy o danych warunkach fizycznych jest dość proporcjonalna do ich warunków. Sam fakt nadreprezentacji pewnych grup świadczy również o profilu graczy, którzy są najczęściej wybierani. Są to jednak dane mieszczące się w okolicach średnich. Zdecydowanie preferowani są wyżsi piłkarze, o wzroście z grupy 180 – 190 cm.

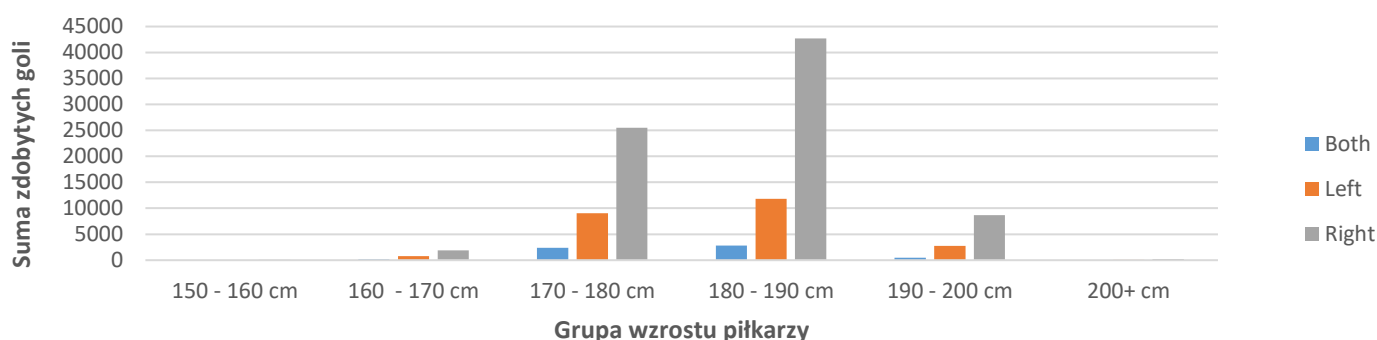
Zauważono bardzo mały wpływ wzrostu na samą średnią liczbę zwycięstw. Minimalną przewagę widać dla piłkarzy obu i lewnonożnych względem piłkarzy prawonożnych, którychz kolei jest najwięcej.

Z zestawień można wyciągnąć wnioski, że piłkarzami zostają osoby o dość konkretnych warunkach fizycznych. Zarówno osoby nieprzeciętnie niskie jak i wysokie są niedoreprezentowane wśród piłkarzy (w przeciwieństwo do np. takiej koszykówki). Najwięcej piłkarzy wpada w przedział 180 – 190 cm wzrostu, co wypada powyżej europejskiej średniej wzrostu wynoszącej dla mężczyzn 177 cm. Jednak

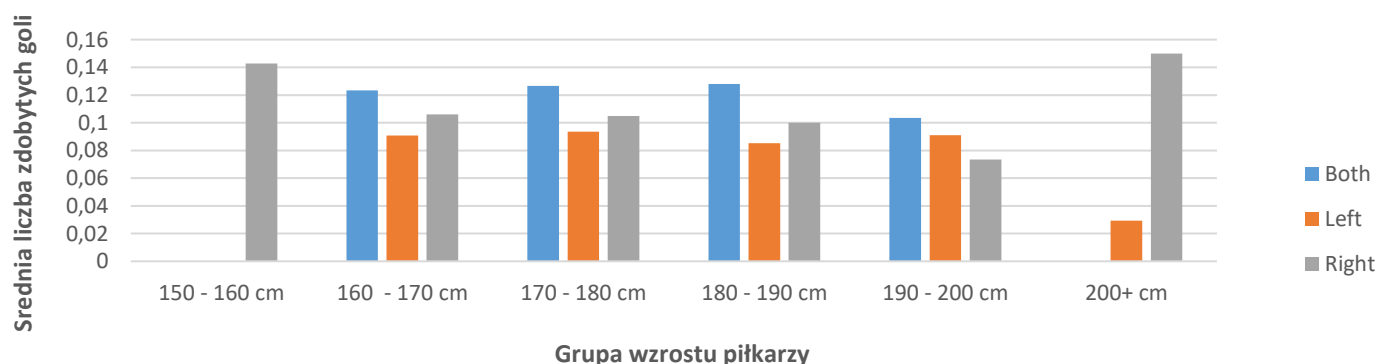
pomimo wyraźnych preferencji w samym zostaniu piłkarzem, nie widać dużych różnic w osiągniętych średnich wygranych. Można zatem przypuszczać, że niezależnie od badanych warunków fizycznych piłkarze muszą spełniać pewne inne kryteria, które determinują o ich sukcesie (poza samym zwerbowaniem). Osoby, których warunki fizyczne nie wpisują się w główny profil piłkarza muszą nadrabiać te aspekty innymi umiejętnościami lub cechami.

10. Jak warunki fizyczne wpływają na wyniki graczy? (Suma zdobytych goli ze względu na stopę dominującą i wzrost).

**Suma zdobytych goli ze względu na grupę wzrostową i stopę dominującą**



**Średnia liczba zdobytych goli ze względu na grupę wzrostową i stopę dominującą**



Average goals	Column Labels			
Row Labels	Both	Left	Right	Grand Total
150 - 160 cm			0,142857143	0,142857143
160 - 170 cm	0,12325752	0,090791027	0,105951784	0,102180327
170 - 180 cm	0,126506976	0,093503653	0,104822889	0,102914801
180 - 190 cm	0,127968694	0,085191266	0,099926327	0,097509432
190 - 200 cm	0,103608356	0,090972406	0,073424885	0,077819918
200+ cm		0,029246344	0,149940263	0,108076473
<b>Grand Total</b>	<b>0,124807201</b>	<b>0,088742458</b>	<b>0,097769658</b>	<b>0,096703891</b>

Jak w przypadku samych zwycięstw nie dało się zauważyć zależności wychodzących poza samo zostawanie piłkarzem, tak w przypadku średniej liczby goli widać, że zdecydowanie najlepiej radzą sobie piłkarze obunożni. Poza grupą wzrostową 190 – 200 cm piłkarze prawonożni wypadali jednak lepiej od lewonożnych. Poza skrajnymi grupami, które raczej przez samą selekcję odpadają z rozgrywek piłki nożnej, nie zauważono szczególnego wpływu wzrostu na uzyskiwaną średnią liczbę goli.

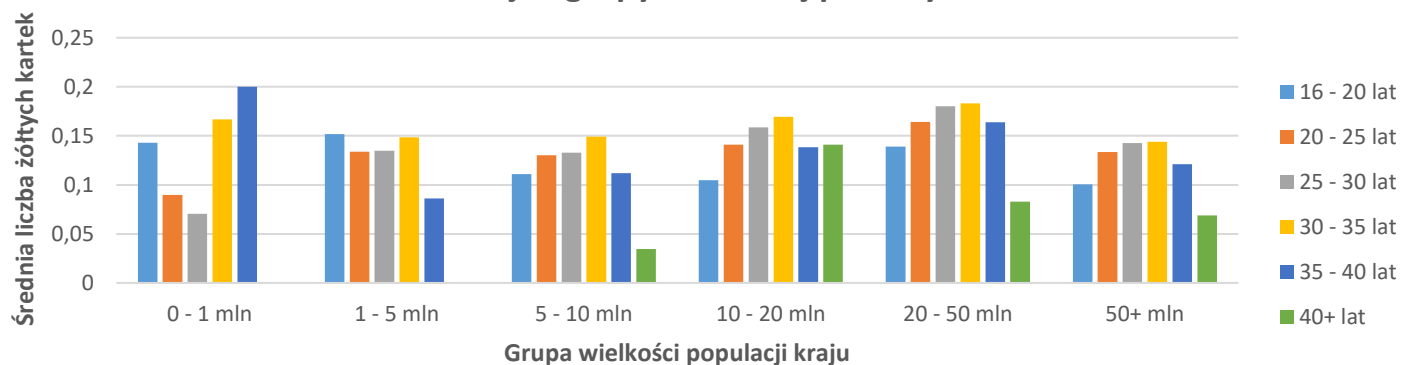
Mniej skuteczni w oddawaniu strzałów byli jednak wyżsi zawodnicy. Grupa wzrostowa 190 – 200 cm całościowo radziła sobie gorzej od innych z zestawienia. Przy tak ograniczonych danych ciężko poszukiwać przyczyn. Przydałoby się znaleźć dodatkowe dane, które pozwoliłyby prowadzić analizę zwinności piłkarzy w zestawieniu ze wzrostem. Być może też piłkarze o tym wzroście po prostu rzadziej grają na pozycji atakującej.

Należy zatem przy wyborze piłkarzy na pozycje związane z oddawaniem strzałów na bramkę w pierwszej kolejności rozpatrywać piłkarzy obunożnych. Są oni skuteczniejsi zarówno od piłkarzy lewo jak i prawonożnych. W przypadku piłkarzy lewo i prawonożnych należy się skłaniać ku zawodnikom prawonożnych, choć jest to raczej kwestia zależna od danej jednostki. Gorsze wyniki osób lewonożnych mogą być związane z np. nieodpowiednim treningiem, gdy domyślnie większość zawodników ma dominującą stopę prawą lub trudnością w zgraniu zawodników lewo i prawonożnych. Są to jednak kwestie wychodzące poza tę analizę, a bezpiecznie można jedynie stwierdzić, że preferowani powinni być strzelcy obunożni.

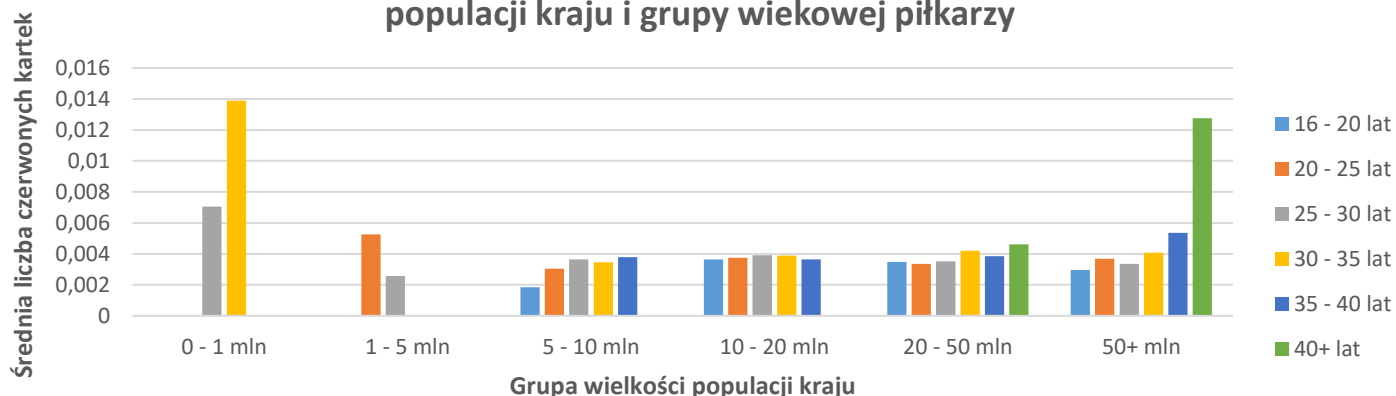
11. Jak wielkość populacji kraju, w którym rozgrywany jest mecz wpływa na agresję graczy (wyrażoną poprzez żółty lub czerwone kartki)?

- (Średnia liczba żółtych kartek w zależności od grupy wielkości populacji i grupy wiekowej graczy).
- (Średnia liczba czerwonych kartek w zależności od grupy wielkości populacji i grupy wiekowej graczy).

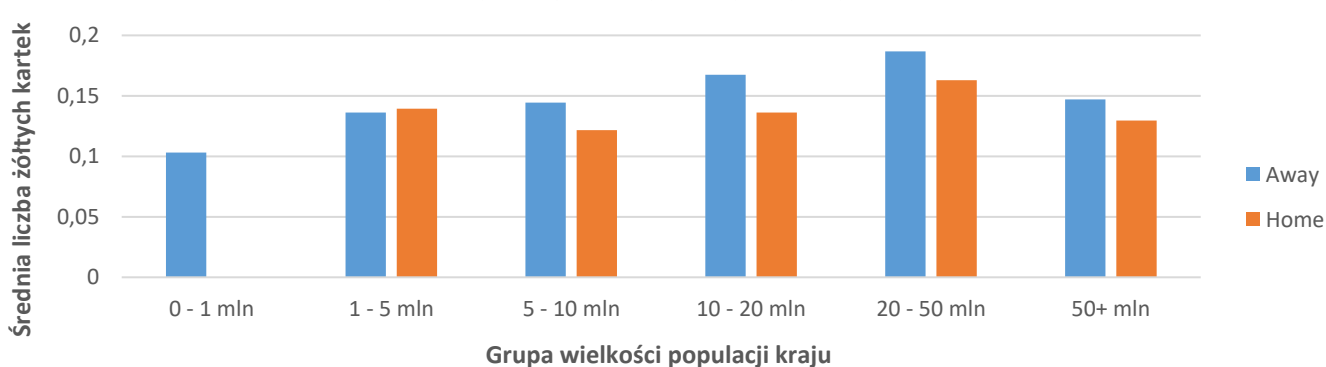
**Średnia liczba żółtych kartek w zależności od grupy wielkości populacji kraju i grupy wiekowej piłkarzy**



**Średnia liczba czerwonych kartek w zależności od grupy wielkości populacji kraju i grupy wiekowej piłkarzy**

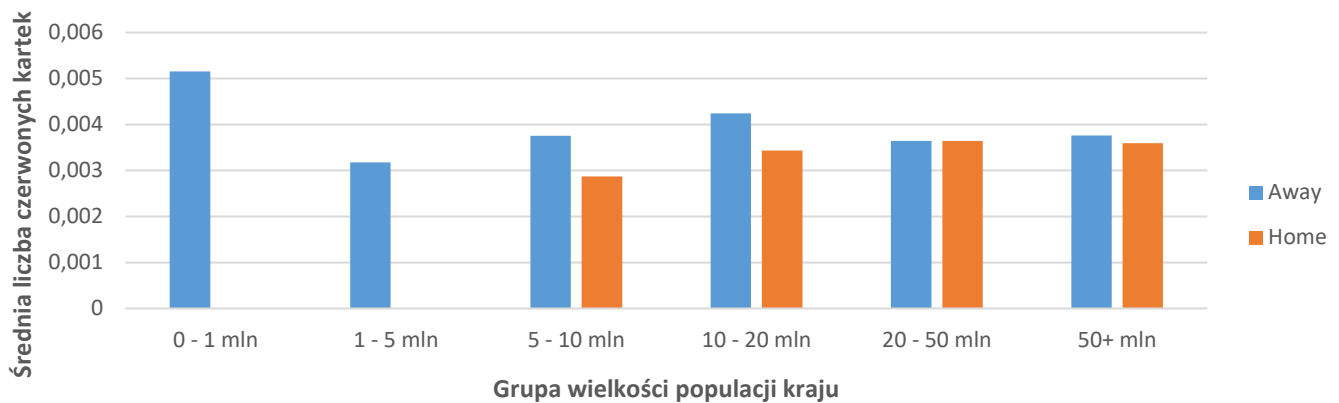


**Średnia liczba żółtych kartek w zależności od grupy wielkości populacji kraju i tego czy drużyna była gospodarzem**





### Średnia liczba czerwonych kartek w zależności od grupy wielkości populacji kraju i tego czy drużyna była gospodarzem



Zauważono, że piłkarze wraz z wiekiem doceniają agresywne, bardziej ryzykowne zagrania. Potwierdza to obserwacje z punktów 7 i 8. Być może piłkarze w tym wieku mają już bardziej ustabilizowaną pozycję w branży, dlatego nie boją się podejmować ryzyka. Sytuacja jest zaburzona dla małych państw do 5 mln mieszkańców. Dane stamtąd są dość mocno wybrakowane i odstają od wyników uzyskiwanych dla innych grup wielkości państw. Co ciekawe, wraz ze wzrostem wielkości państwa dla żółtych kartek wzrasta średnia agresja dla każdej z grup wiekowych. Spada ona dopiero dla dużych państw powyżej 50+ mln mieszkańców. Zdecydowanie inaczej wygląda sytuacja ze średnią liczbą przyznawanych czerwonych kartek. Każdy piłkarz stara się ich unikać bez względu na grupę wiekową i grupę wielkości państwa.

Doping i widoczność to nie jedynie publika na trybunach – piłkarze poruszają się pomiędzy stadionami, widzą fanów gromadzących się po drodze, a nawet na lotniskach. Same mecze transmitowane są w telewizji.

Podobnie dla zestawień związanych ze średnią liczbą przyznawanych kartek ze względu na hosting i grupę wielkości populacji kraju, dane z małych państw są wybrakowane, przez co odrzucone do analizy. Zgodnie z wcześniejszymi wnioskami, również to zestawienie ujawnia, że średnie liczby zarówno żółtych jak i czerwonych kartek są większe dla drużyn przyjezdnych niż gospodarzy. Wartości te narastają analogicznie dla kartek żółtych i czerwonych aż do progu wielkości państwa równego 20 – 50 mln mieszkańców. Dla państw największych następuje spadek zarówno dla gospodarzy jak i przyjezdnych w kartkach żółtych. Dla kartek czerwonych spadek ten odnotuje się już od 20 – 50 mln mieszkańców i wartości utrzymują się dla państw największych.

Piłkarze powinni mieć specjalistę, który będzie poruszał z nimi tematy związane z całą otoczką meczu, taką jak właśnie kibicami w miejscach publicznych czy streamowaniem spotkań w telewizji. Z zestawienia nie można jednoznacznie stwierdzić tendencji, można jednak spekulować, że narastająca presja może wzmacniać agresję zagrań, a w pewnym momencie, poprzez efekt wystawienia na świeczniku, ją również hamować.

## Data Mining

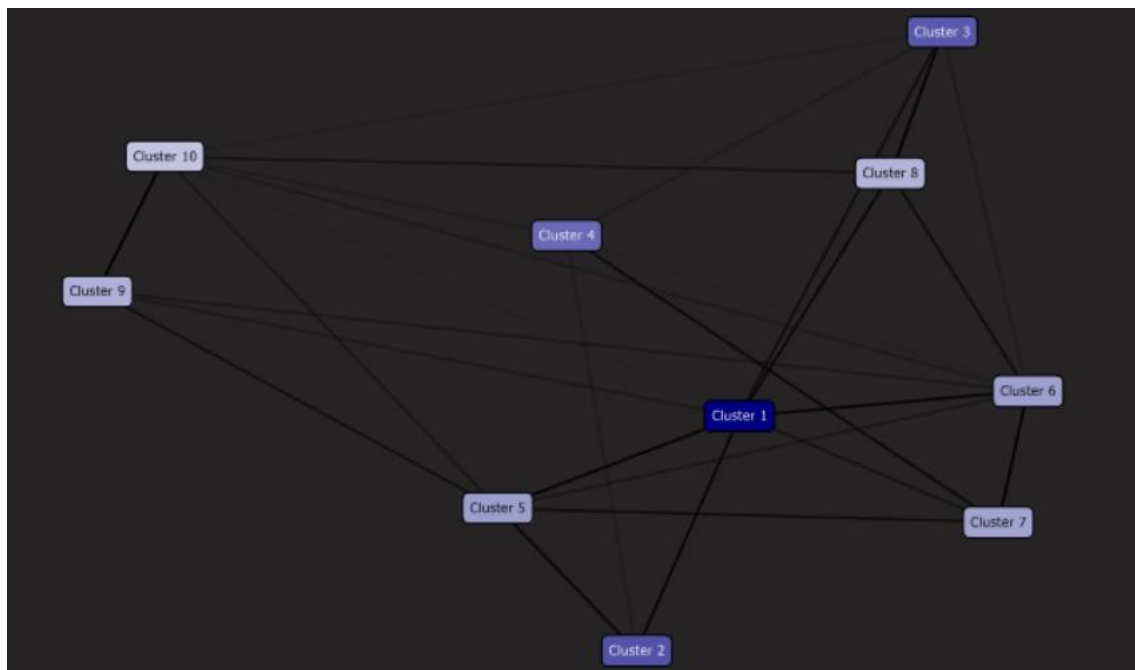
Utworzone modele:

Structure ↑	PL2_Decision_Tree	PL2_Clustering
	Microsoft_Decision_Trees	Microsoft_Clustering
Average assists	Input	Input
Average goals	Predict	Predict
Average minutes played	Input	Input
Average red cards	Input	Input
Average yellow cards	Input	Input
Foot	Input	Input
Highest Market Value In E...	Input	Input
Height Group	Input	Input
Player Id	Key	Key
Position	Input	Input
Sub Position	Input	Input





















Algorytm drzewo decyzyjne, wpływ atrybutów na wartość average goals w kolejności od najmniej znaczącego do najbardziej znaczącego (analiza dla wymiaru piłkarza):




















- Average yellow cards
- Average red cards
- Highest market value in euro group
- Foot
- Sub position
- Height group
- Average assists
- Average minutes played
- Position

Drzewo decyzyjne wykazało, że największy wpływ na średnie gole ma pozycja piłkarza i średnia liczba minut spędzanych na boisku. Zgadza się to ze wcześniejszymi analizami, gdzie czas spędzony na boisku wzrastał wraz z wiekiem (w szczególności dla napastników), ale również wtedy rosła średnia liczba zdobytych goli. Co ciekawe na liczbę goli ma mieć wpływ grupa wzrostowa, a dopiero później stopa. Z wykonanych zestawień wynikała odwrotna zależność, gdzie grupa wzrostowa była czynnikiem selekcyjnym piłkarzy, ale nie mającym aż tak dużego wpływu na ich wyniki. Ciężko określić czy algorytm nie wziął zbyt bardzo pod uwagę wartości z grup skrajnych, które są mocno niedoreprezentowane i ciężko na ich podstawie wnioskować o ogóle piłkarzy o takim profilu.

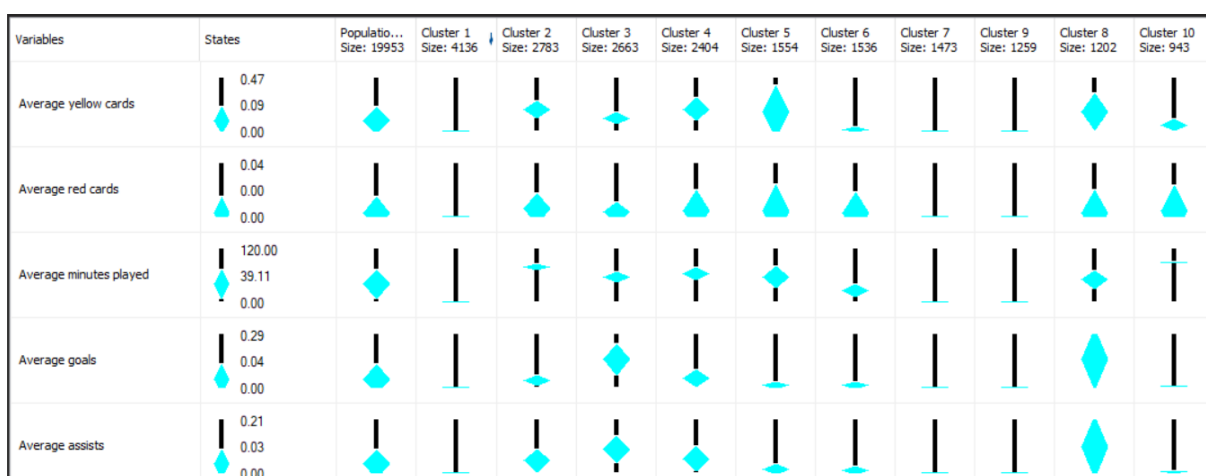
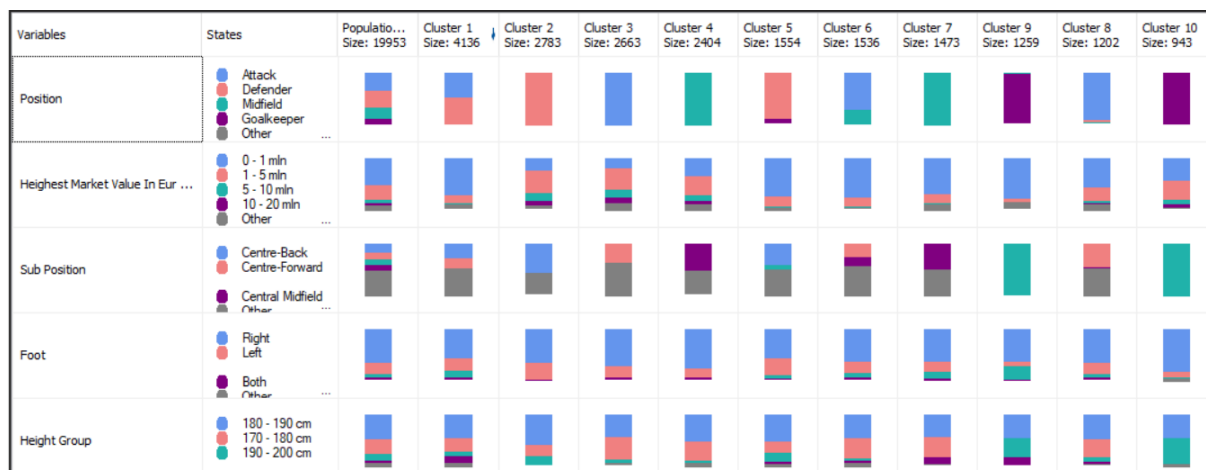


Variables	Values	Probability
Foot	Right	<div></div>
Heighest Market Value In Eur Group	0 - 1 mln	<div></div>
Height Group	180 - 190 cm	<div></div>
Position	Attack	<div></div>
Position	Defender	<div></div>
Height Group	170 - 180 cm	<div></div>
Heighest Market Value In Eur Group	1 - 5 mln	<div></div>
Average assists	0.0 - 0.1	<div></div>
Average yellow cards	0.1 - 0.2	<div></div>
Average minutes played	39.1 - 62.1	<div></div>
Average minutes played	16.1 - 39.1	<div></div>
Average yellow cards	0.0 - 0.1	<div></div>
Average red cards	0.0	<div></div>
Average goals	0.0 - 0.1	<div></div>
Average goals	0.1 - 0.3	<div></div>
Average yellow cards	0.2 - 0.5	<div></div>
Average red cards	0.0	<div></div>
Average assists	0.1 - 0.2	<div></div>
Average minutes played	62.1 - 120.0	<div></div>
Foot	Left	<div></div>

Variables	Values	Probability
Average minutes played	62.1 - 120.0	
Foot	Left	
Position	Midfield	
Average assists	0.0	
Average goals	0.0	
Sub Position	Centre-Back	
Sub Position	Centre-Forward	
Height Group	190 - 200 cm	
Average minutes played	0.0 - 16.1	
Sub Position	missing	
Position	Goalkeeper	
Sub Position	Central Midfield	
Sub Position	Defensive Midfield	
Heighest Market Value In Eur Group	5 - 10 mln	
Foot	missing	
Sub Position	Right-Back	
Average red cards	0.0	
Sub Position	Left-Back	
Sub Position	Attacking Midfield	
Sub Position	Left Winger	

Characteristics for Population (All)		
Variables	Values	Probability
Heighest Market Value In Eur Group	5 - 10 mln	
Foot	missing	
Sub Position	Right-Back	
Average red cards	0.0	
Sub Position	Left-Back	
Sub Position	Attacking Midfield	
Sub Position	Left Winger	
Height Group	missing	
Sub Position	Right Winger	
Foot	Both	
Heighest Market Value In Eur Group	10 - 20 mln	
Heighest Market Value In Eur Group	missing	
Heighest Market Value In Eur Group	20 - 50 mln	
Height Group	160 - 170 cm	
Sub Position	Left Midfield	
Sub Position	Right Midfield	
Average yellow cards	0.0	
Heighest Market Value In Eur Group	50 + mln	
Sub Position	Second Striker	

Utworzone zostało 10 klastrów. Ogólna populacja piłkarzy prezentuje się w następujący sposób, co pokrywa się z informacjami uzyskanymi w wyniku tworzenia zestawień.



Utworzone zostało 10 klastrów. Podzielone zostały one przede wszystkim ze względu na pozycję piłkarza.

Klasy obrazują jak rozkładają się profile fizyczne piłkarzy ze względu na pozycję. Przyporządkowują również wyniki i osiągi na boisku. Widać, że mała liczba piłkarzy o najwyższym wzroście przypada głównie na pozycji bramkarza. Również bramkarzy cechuje najmniejsza różnorodność stóp. Nie dość, że większość z nich jest po prostu prawonóżna, są też tam największe braki w danych wiązanych z określeniem stopy dominującej piłkarza.

Z danych widać również, na jakich pozycjach otrzymywane najczęściej są żółte i czerwone kartki. Średnio najwięcej żółtych kartek otrzymują obrońcy i pomocnicy. Atakujących cechuje duży rozrzut jeżeli chodzi o agresję powiązaną z żółtymi kartkami. Z kolei bramkarze praktycznie nie otrzymują żadnych żółtych kartek, co prawdopodobnie ma związek z naturą ich pozycji i rozgrywką odizolowaną od reszty graczy. Z kolei w przypadku średnich wartości liczby przyznawanych czerwonych kartek zacierają się różnice pomiędzy pozycjami.

Najwyższe średnie wartości związane ze zdobywaniem goli i asyst przypadają atakującym. Najniższe wartości osiągane są oczywiście dla bramkarzy.

## KPI – agresja piłkarzy

GeneralAgresionKPI Status	Column Labels	2020	2018	2022	2014	2019	2015	2021	2016	2013	2017	Grand Total
Row Labels												
A.J. Soares		●	●	●	●	●	●	●	◆	●	●	◆
Aaron Appindangoyé		◆	●	◆	◆	◆	●	◆	●	●	●	◆
Aaron Bastiaans		●	●	●	●	●	●	●	●	●	●	●
Aaron Boupendza		◆	●	●	●	●	●	●	●	●	●	◆
Aaron Chapman		●	●	●	●	●	●	●	●	●	●	●
Aaron Comrie		●	◆	●	●	●	●	●	●	●	◆	◆
Aaron Connolly		●	●	●	●	●	●	●	●	●	●	●
Aaron Cresswell		◆	▲	◆	◆	◆	●	◆	◆	●	◆	◆
Aaron Dhondt		●	●	●	●	●	●	●	●	●	●	●
Aaron Doran		●	●	●	◆	●	●	●	◆	●	●	◆
Aarón Escandell		◆	●	●	●	◆	●	◆	●	●	●	◆
Aaron Gielen		●	●	●	●	●	●	●	●	●	●	●
Aaron Herzog		●	●	●	●	●	●	●	●	●	●	●
Aaron Hickey		◆	●	◆	●	◆	●	◆	●	●	●	◆
Aaron Hughes		●	●	●	●	●	●	●	●	●	●	●
Aaron Hunt		●	●	●	●	●	◆	●	◆	●	●	◆
Aaron Kamardin		●	●	●	●	●	●	●	●	●	●	●
Aaron Kuhl		●	●	●	●	●	●	●	●	●	●	●

Utworzone KPI jest głównie skierowane dla bukmacherów i selekcjonerów składów. Pozwala bukmacherom obserwować zawodników o zwiększonej agresji, np. na przestrzeni sezonów. Umożliwia to ocenę jak skłonny do agresywnych zagrań jest piłkarz, dzięki czemu można ustalać odpowiednie wartości w zakładach. Zestawiając to z rezultatami analizy agresji piłkarzy ze względu na hosting i ich wiek, można podejmować decyzje bukmacherskie kierowane odkrytymi tendencjami.

Jest ono również przydatne dla selekcjonerów drużyn, którzy chcą dbać o kulturę rozgrywki i nie chcą dopuszczać do sytuacji, w których konieczne jest zejście zawodnika z boiska. Badanie agresji zawodników na przestrzeni sezonów może pomóc odrzucić piłkarzy, którzy nie spełniają podstawowych założeń przy tworzeniu składu.