Projekt część I - raport

Maciej Durkalec
Informatyka Stosowana
MSiD Lab grupa 4
280582@student.pwr.edu.pl

I. Wstęp

CEL PROJEKTU

Zgodnie z założeniami pierwszej części projektu, przeprowadziłem kompleksową analizę zbioru danych FIFA 22, skupiając się na:

- Poznaniu narzędzi Python do analizy danych
- Identyfikacji kluczowych zależności między cechami zawodników
- Przygotowaniu wizualizacji wspierających wnioskowanie

WYKORZYSTANE NARZĘDZIA

Środowisko programistyczne:

- Środowisko wirtualne Conda (fifa_env)
- Plik environment.yml z wymaganymi bibliotekami

Biblioteki Python:

- pandas – manipulacja danymi i obliczenia statystyczne
- seaborn & matplotlib – zaawansowane wizualizacje
- numpy – obliczenia numeryczne

Metody analityczne:

- Statystyki opisowe (średnia, mediana, percentyle)
- Wizualizacje rozkładów (histogramy, boxploty, violinploty)
- Analiza korelacji (heatmapa, regresja liniowa)

II. Charakterystyka zbioru danych

Oryginalny zbiór danych FIFA 22

- Rozmiar: 19,000+ wierszy (zawodników)
- Kolumny: 110 cech opisujących zawodników
- Zakres danych:
  - Informacje biometryczne (wiek, wzrost, waga)
  - Statystyki sportowe (oceny, pozycje, umiejętności)
  - Dane finansowe (wartość rynkowa, zarobki)
  - Kontekst klubowy i narodowy

Wybrane cechy do analizy

Ze względu na złożoność zbioru, skupiłem się na 11 kluczowych zmiennych:

Kategoria	Wybrane cechy	Typ danych
Oceny	overall, potential	numeryczne
Biometria	age, height_cm, weight_kg	numeryczne
Finanse	value_eur, wage_eur	numeryczne
Kontekst	nationality_name, club_name	kategorialne
Charakterystyka	preferred_foot, player_positions	kategorialne

Uzasadnienie wyboru:

- Zmienne numeryczne umożliwiły analizę rozkładów i korelacji
- Zmienne kategorialne pokazały zależności grupowe (np. preferowana noga vs wartość rynkowa)
- Ograniczenie liczby cech zwiększyło czytelność analizy

Przetwarzanie danych:

- Filtracja: Usunięto wiersze z brakującymi wartościami w kluczowych kolumnach
- Transformacje - Wydobycie głównej pozycji z player_positions (np. "CAM" z "CAM,RW")
- Normalizacja: Standaryzacja jednostek (wzrost w cm, waga w kg, zarobki w EUR/tydzień)

Kategoria	Wybrane cechy	Typ danych
Oceny	overall, potential	numeryczne
Biometria	age, height_cm, weight_kg	numeryczne
Finanse	value_eur, wage_eur	numeryczne
Kontekst	nationality_name, club_name	kategorialne
Charakterystyka	preferred_foot, player_positions	kategorialne

Analiza Statystyk Wstępnych Danych

1. Statystyki Numeryczne

Cecha	Średnia	Mediana	Min	Max	Odchylenie standardowe	5%	95%	Brakujące wartości
overall	65.77	66	47	93	6.88	54	77	0
potential	71.08	71	49	95	6.09	62	82	0
age	25.21	25	16	54	4.75	18	34	0
height_cm	181.30	181	155	206	6.86	170	193	0
weight_kg	74.94	75	49	110	7.07	64	87	0
value_eur	2,850,452	975,000	9,000	194,000,000	7,613,700	180,000	11,500,000	74
wage_eur	9,018	3,000	500	350,000	19,470	500	37,150	61

Kluczowe obserwacje:

- Rozkład ocen (overall): Większość graczy ma ocenę między 54 a 77 (5-95 percentyl).
- Wartość rynkowa (value_eur): Skrajnie prawostronny rozkład – mediana (975k EUR) jest znacznie niższa niż średnia (2.85M EUR), co wskazuje na nieliniową zależność.
- Wiek (age): 95% graczy ma mniej niż 34 lata, ale rekordzista ma 54 lata (K. Miura z Japoni, natomiast drugi najstarszy zawodnik ma jedynie 43 lata).
- Pensje (wage_eur): Ogromne dysproporcje – 5% najgorzej opłacanych zarabia ≤500 EUR/tydzień, a 5% najlepszych >37k EUR/tydzień.

2. Statystyki Kategorialne

Narodowości (nationality_name)

- 163 unikalne kraje, ale silna koncentracja:
- 9% zawodników pochodzi z Anglii (najliczniejsza grupa)
- Prawie 1/3 wszystkich zawodników pochodzi z zaledwie 5 krajów.

Kluby (club_name)

- 701 unikalnych klubów, brak wyraźnego lidera:
- Najczęstsze: Arsenal, Everton, Betis, Borussia Mönchengladbach (po 0.17% każdy)
- 61 brakujących wartości (0.3% danych)
- Wnioski: Rozproszenie zawodników – brak klubu z >0.2% reprezentacją w zbiorze.

Zmienna	Unikalne klasy	Brakujące wartości	Najczęstsza klasa	Udział najczęstszej klasy	Top 5 klas (udział %)
nationality_name	163	0	England	9.0%	England (8.9%), Germany (6.3%), Spain (5.6%), France (5.1%), Argentina (5.0%)
club_name	701	61	Arsenal	0.17%	Real Betis (0.17%), Everton (0.17%), Borussia M'gladbach (0.17%), Arsenal (0.17%), Celta Vigo (0.17%)
preferred_foot	2	0	Right	76.2%	Right (76.3%), Left (23.7%)
player_positions	674	0	CB	12.6%	CB (12.6%), GK (11.1%), ST (9.2%), CDM/CM (5.0%), CM (3.8%)
short_name	18,145	0	J. Rodríguez	0.07%	J. Rodríguez (0.07%), J. Hernández (0.05%), J. Brown (0.04%), Paulinho (0.04%), L. Rodríguez (0.04%)

Preferowana noga (preferred_foot)

- 76% prawonożnych, 24% lewonożnych
- Wnioski: Lewonożni są 3x rzadsi, co może tłumaczyć ich wyższą wartość rynkową (potwierdzone w późniejszej analizie boxplotów). Oraz potwierdzona teoria o wyższym potencjale sportowym osób leworęcznych/lewonożnych (ok. 12% ludzi świata jest lewonożna, a w sporcie ten odsetek jest znacząco wyższy)

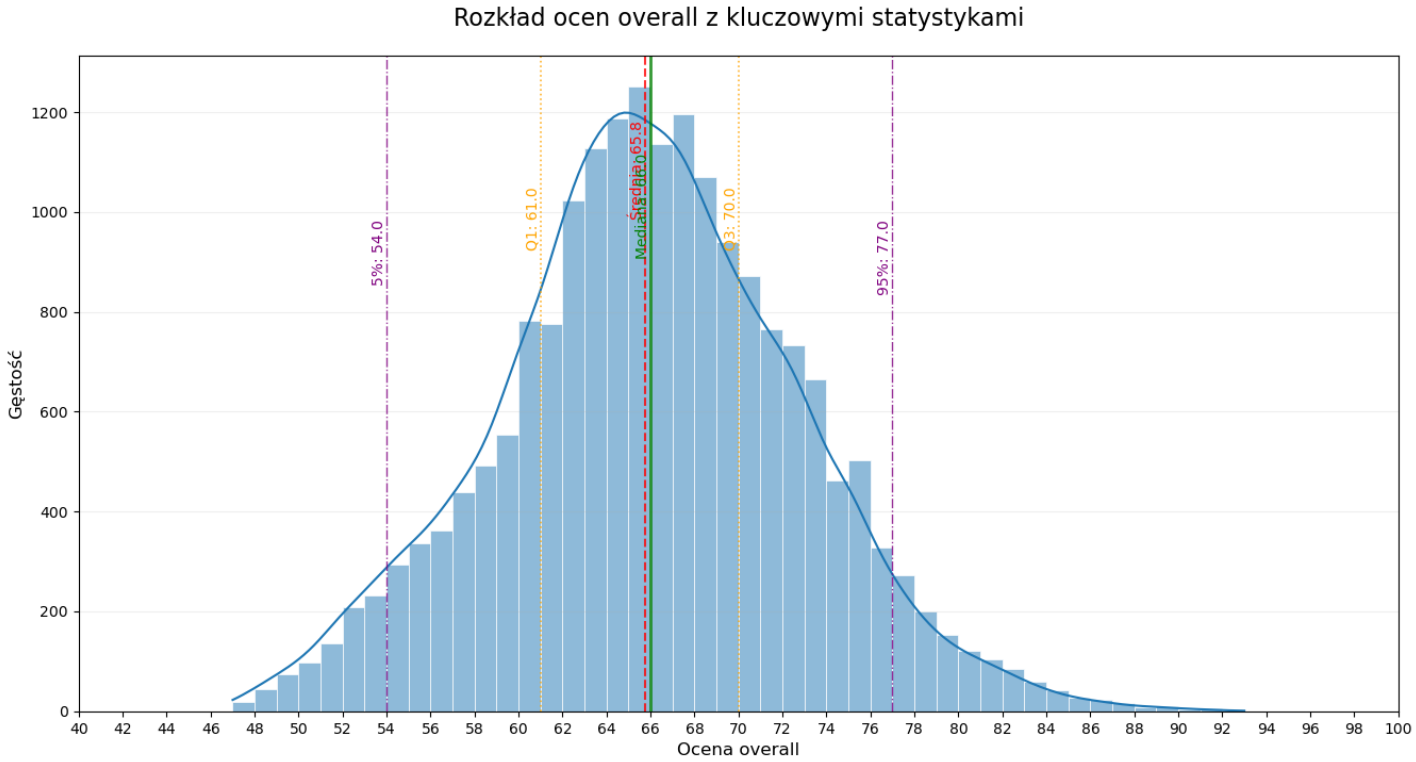
Pozycje (player_positions)

- 674 unikalne kombinacje, ale dominują proste role:
- 12.6% środkowi obrońcy (CB)
- 11.1% bramkarze (GK)
- 9.2% napastnicy (ST)

Analiza wykresu: Rozkład ocen overall

Opis techniczny

- Oś X: Ocena overall (skala 40-100)
- Oś Y: Liczba zawodników (gęstość rozkładu)
- Dodatkowe elementy:
- Czerwona przerywana linia: średnia (70.0)
- Zielona ciągła linia: mediana (77.0)
- Pomarańczowe linie: kwartyle (Q1 i Q3)
- Fioletowe linie: percentyle 5% i 95%



Kluczowe obserwacje

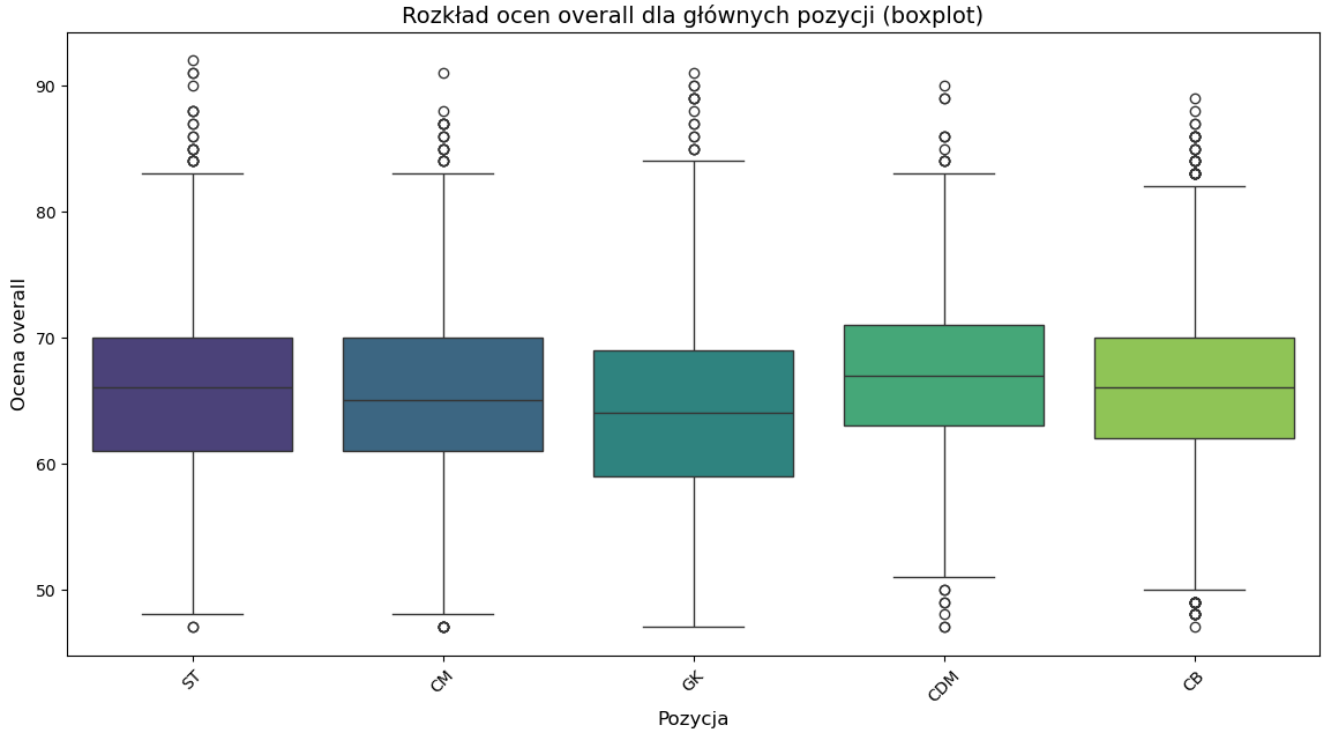
- Rozkład normalny lekko prawostronnie skośny. Lewa strona rozkładu (niższe oceny) jest bardziej stroma, prawa (wyższe oceny) - bardziej płaska i rozciągnięta
- 50% zawodników mieści się w wąskim przedziale 61-70 mimo, że overall przyjmuje wartości od 47 do 93
- Już overall 77 stanowi 95 centyl, a mimo tego najlepsi gracze osiągają oceny 90+ (np. Messi, Ronaldo, Lewandowski) co oznacza że różnica pomiędzy czołówką, a absolutną elitą jest niesamowicie duża, co w późniejszej analizie będzie zauważalne i wykresy będą pokazywać wyraźne wychylenia dla danych elitarnych zawodników.

Analiza boxplotu: Oceny overall wg pozycji

Boxplot wizualizuje rozkład danych poprzez **medianę** (linia środkowa), **kwartyle** (pudełko), typowy zakres wartości (wąsy) oraz wartości odstające (kropki), pozwalając na szybkie porównanie grup.

Obserwacje

- Najwyższą medianę overall mają zawodnicy na pozycji CDM (Środkowy pomocnik defensywny), a najniższą na pozycji GK (bramkarz)
- Bramkarze mają najszersze wąsy na boxplotach, co oznacza większe zróżnicowanie ich ocen overall, podczas gdy rozkład ocen CDM i CB jest najbardziej skupiony – ich wąsy są najwęższe.
- Na każdej pozycji widać ekstremalne wartości – wybitnych zawodników

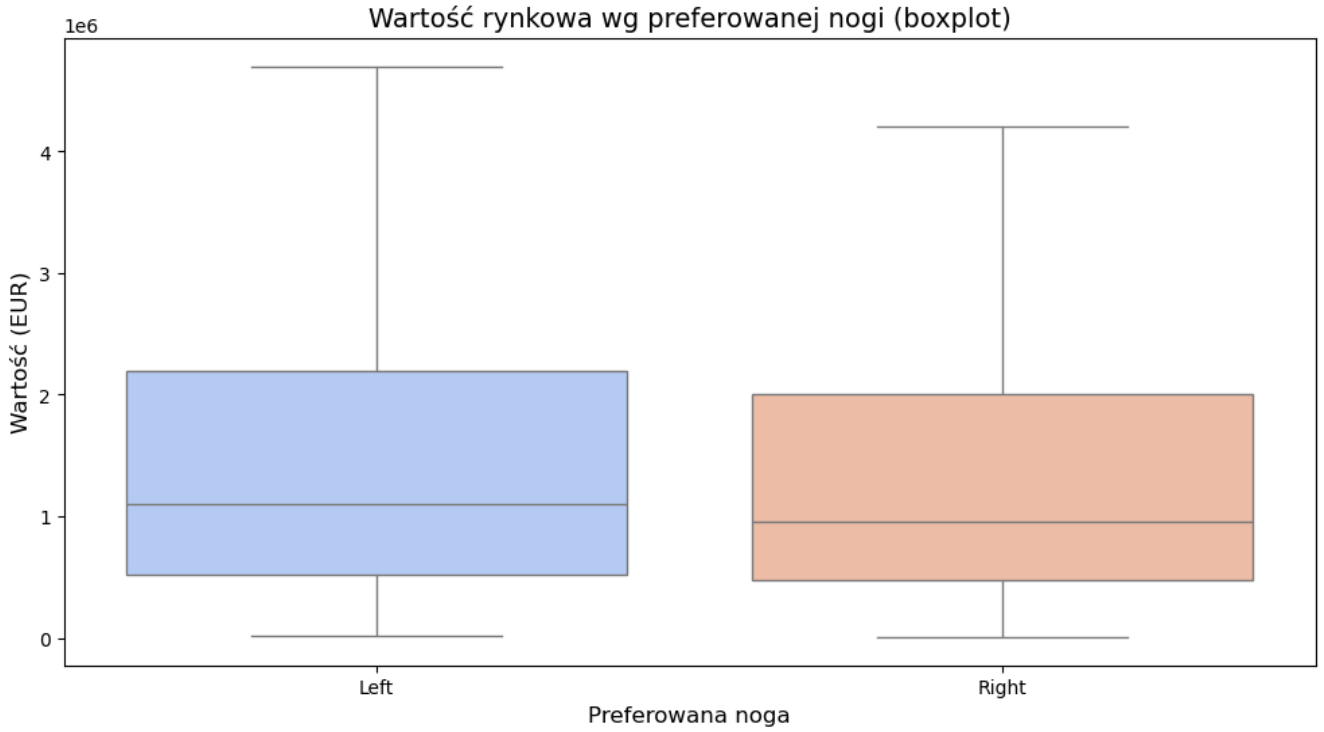


Analiza boxplotu: Wartość rynkowa vs preferowana noga

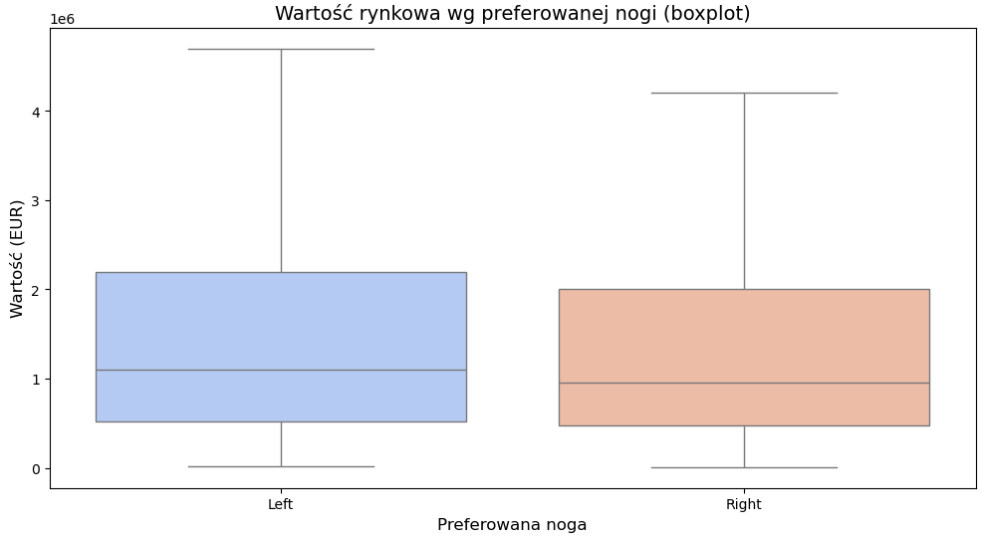
Outliery zostały ukryte (showfliers=False), ponieważ gdyby próbować je pokazać wykres stałby się nieczytelny ze względu na znacznie wyższą wartości topowych zawodników.

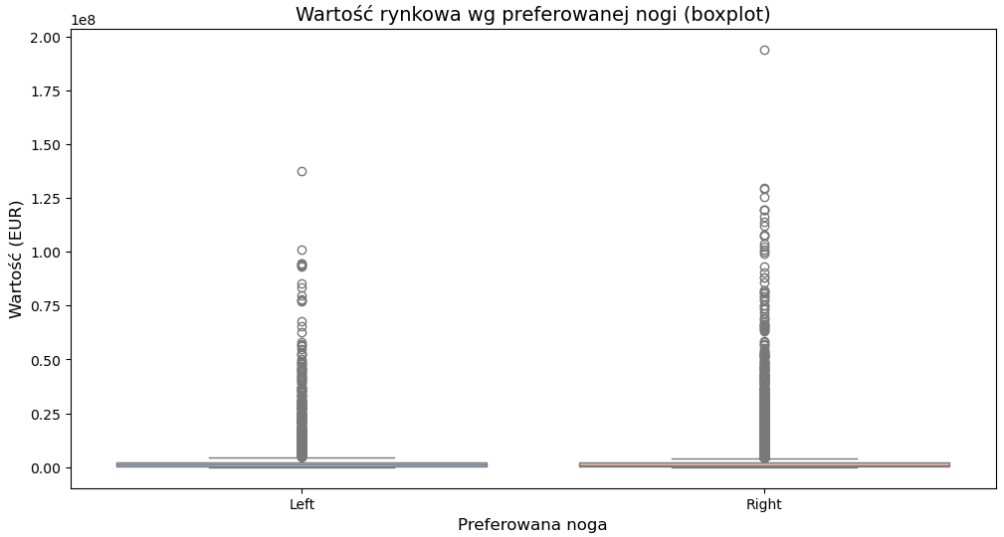
Obserwacje:

- Lewonożni (Left) mają wyższą medianę wartości rynkowej niż prawonożni (Right).
- Mediana jest bliżej Q1 niż Q3 w obu przypadkach co pokazuje, że w przypadku wartości mamy do czynienia z rozkładem prawostronnie skośnym
- Z racji, że zawodników lewonożnych jest znacznie mniej (24%) ich wartość jest znacznie wyższa przez niższą podaż



Wartość rynkowa vs preferowana noga (showfliers=false) i (showfliers=true)





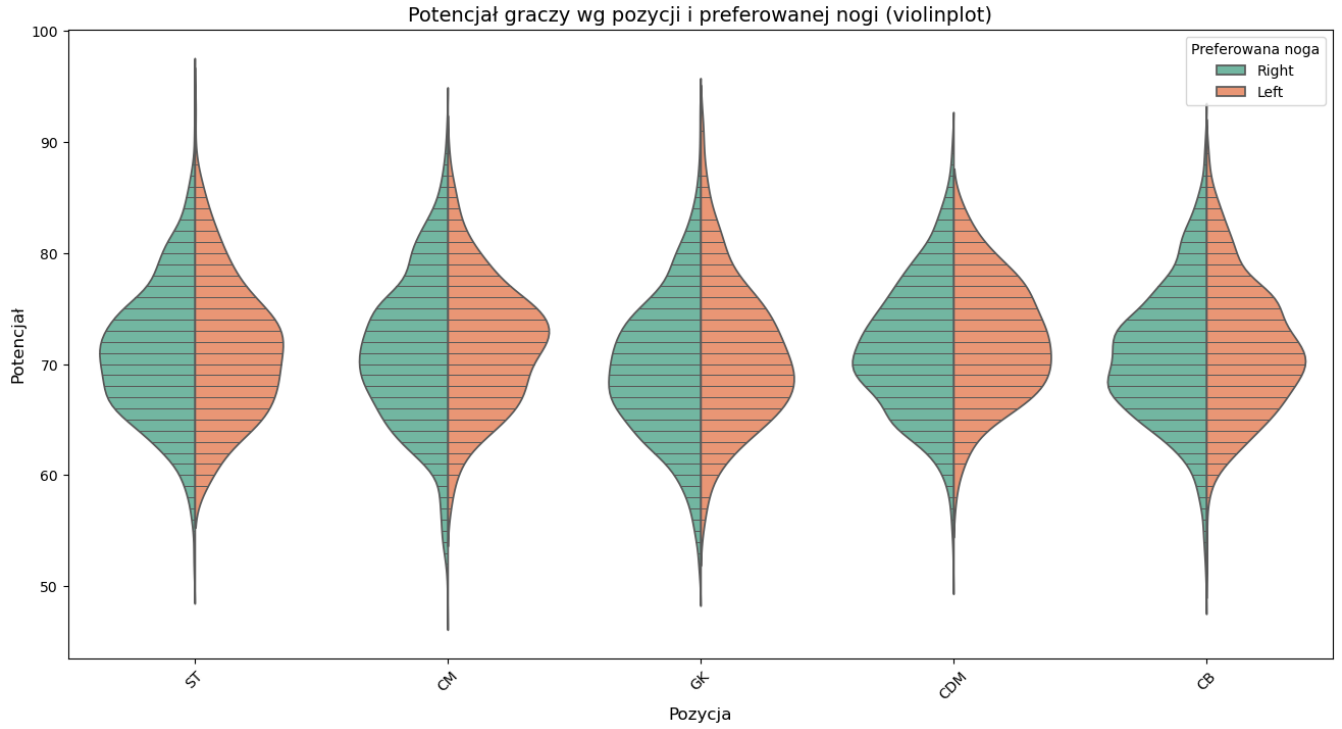
Wykres z showfliers=True jest nieczytelny, bo wartości rynkowe sięgają nawet 194 mln EUR (K. Mbappé), podczas gdy typowy zakres (IQR) mieści się w przedziale 0.5-2 mln EUR – boxy stają się wąską linią przy osi, a cały wykres zdominowany jest przez kropki outlierów.

Analiza violinplotu: Potencjał graczy wg pozycji i preferowanej nogi

Violinplot to połączenie boxplotu i wykresu gęstości - kształt "skrzypiec" pokazuje rozkład danych (szersze miejsca = więcej zawodników), a wewnętrzne linie (inner='stick') oznaczają konkretne wartości.

Obserwacje

- Bramkarze (GK) mają najbardziej symetryczny rozkład co oznacza, że ich preferowana noga nie wpływa na potencjał.
- Wśród innych roli da się zauważyć nieznacznie wyższy potencjał zawodników lewonożnych, szczególnie na pozycji środkowego pomocnika (CM)
- Przy górnych wartościach wyraźnie dominuje kolor pomarańczowy (lewonożni), a przy dolnych - zielony, co pokazuje, że:
    - najwyższy potencjał mają głównie lewonożni,
    - zawodnicy lewonożni rzadko otrzymują niskie oceny potencjału.

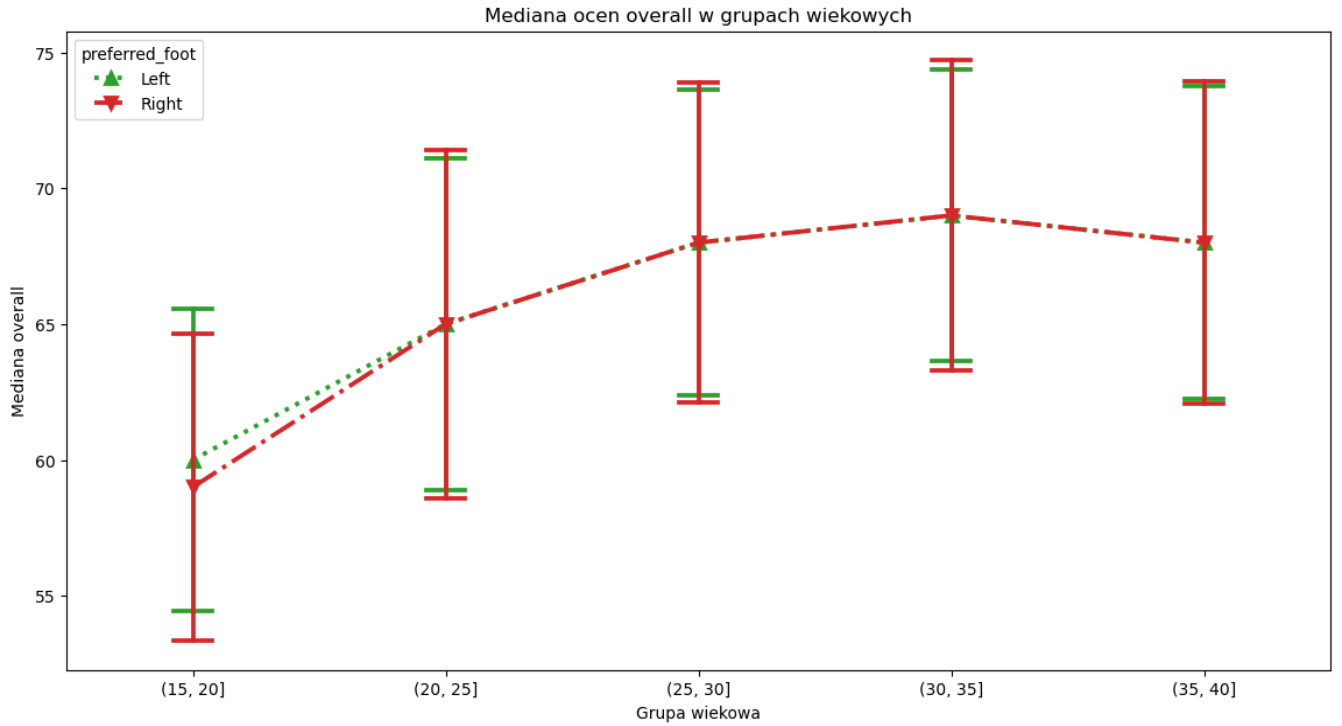


Analiza wykresu punktowego z przedziałami błędów: Mediana ocen overall w grupach wiekowych

Wykres punktowy z przedziałami błędów pokazuje medianę ocen overall dla różnych grup wiekowych, z podziałem na preferowaną nogę. Słupki błędów (error bars) reprezentują odchylenie standardowe.

Obserwacje

- Mimo iż wcześniej można było zauważyć, że zarówno zarobki jak i potencjał jest wyższy u graczy lewonożnych, to w przypadku oceny overall niemal nie ma różnicy pomiędzy preferowaną nogą.
- Najwyższe oceny są w grupie 30-35 lat
- Odchylenie standardowe graczy lewonożnych jest mniejsze przez co może wynikać z ich specjalistycznej roli w drużynie.
- Jedyna różnica pomiędzy prawą, a lewą nogą występuje w grupie 15-20 lat

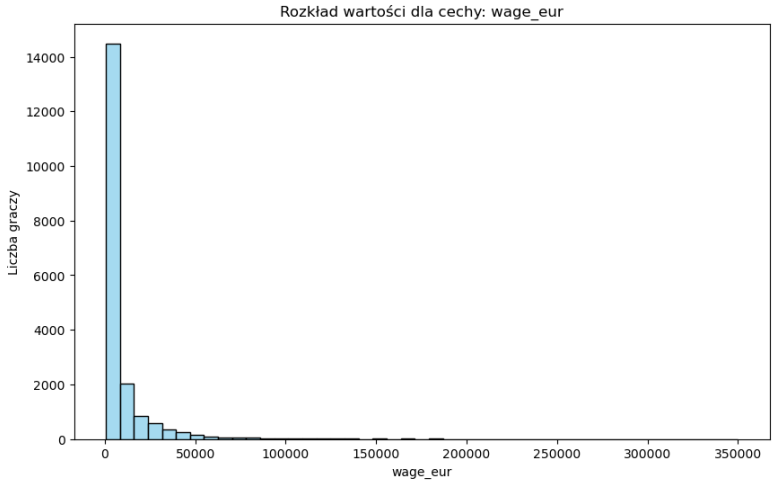


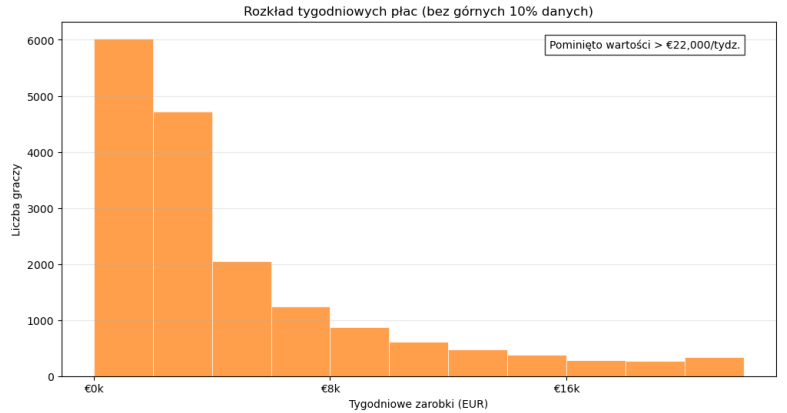
Analiza rozkładu zarobków i wartości rynkowej zawodników

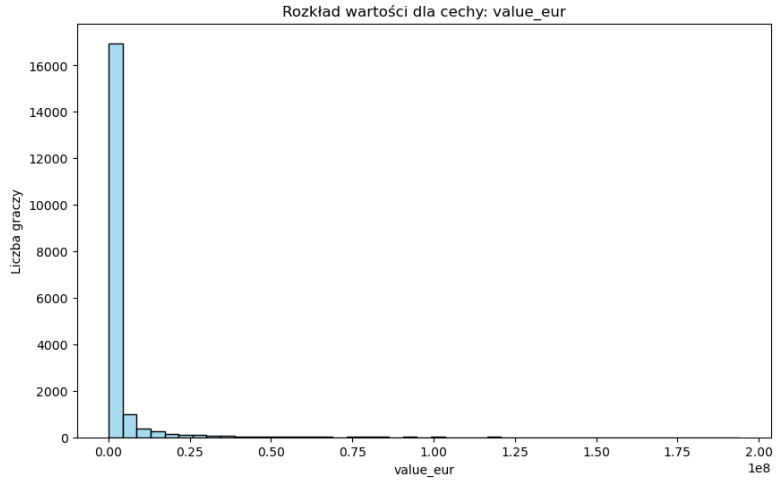
Transformacja danych
W obu przypadkach czołowi zawodnicy są warci i zarabiają znacznie więcej niż większość piłkarzy zatem by poprawić czytelność wykresu usunąłem wartości powyżej 90 centyla.

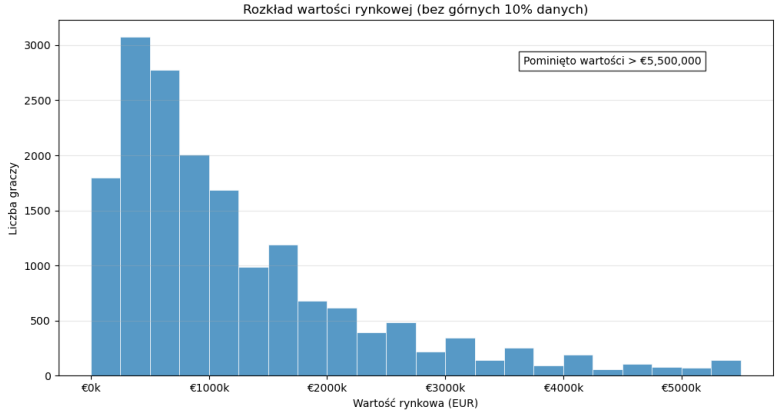
Charakterystyka rozkładów
Wspólne cechy dla zarobków i wartości rynkowej:

- Szybki spadek – większość zawodników skupiona w dolnych przedziałach wartości
- Długi prawy ogon – rozkład przypomina hiperbolę (1/x)
- ~80% graczy mieści się w pierwszych 20% skali wartości (zasada Pareto)







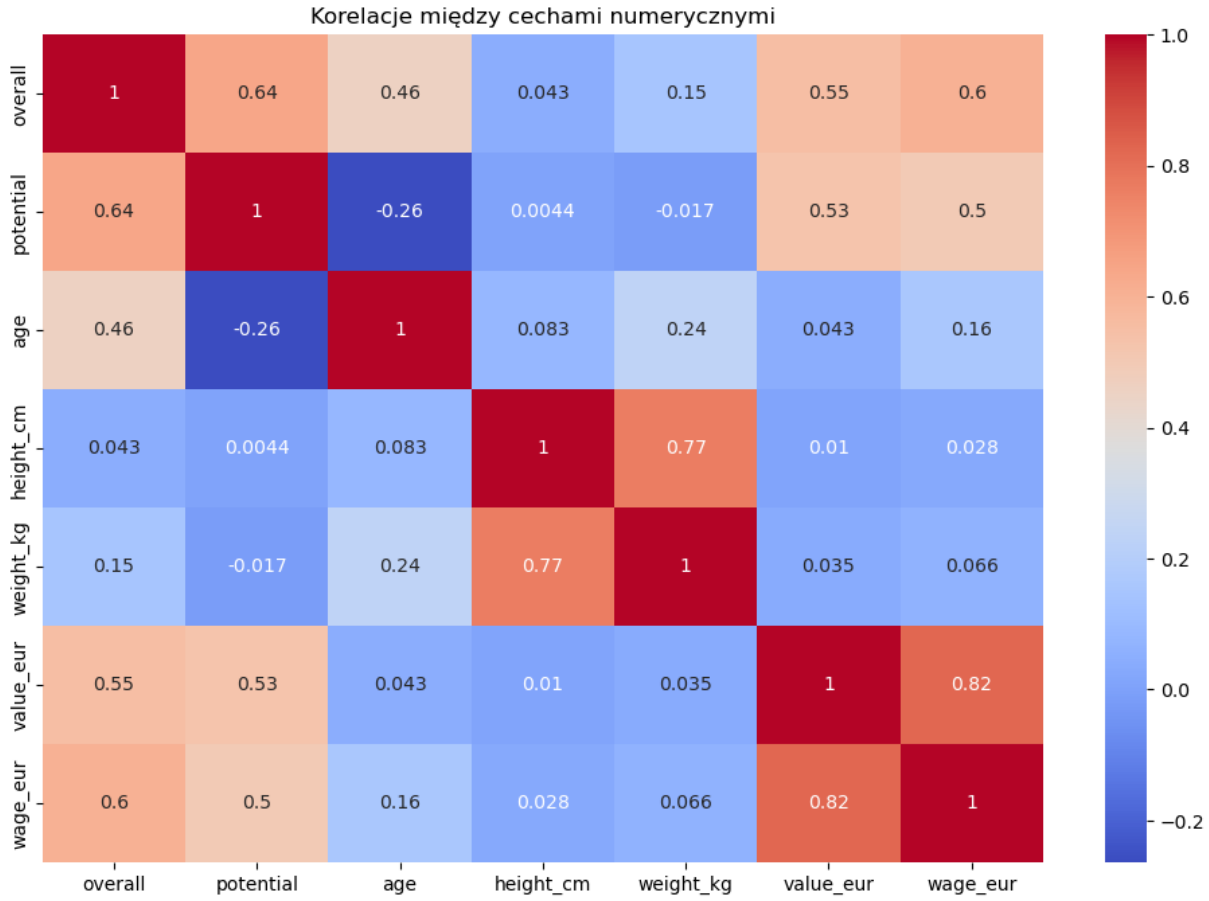


Analiza heatmapy korelacji

1. Najsilniejsze zależności
- Ocena overall & wartość rynkowa (0.6) – im lepszy zawodnik, tym droższy (logiczne)
- Zarobki & wartość (0.82) – więcej warci zawodnicy również więcej zarabiają
- Wzrost & waga (0.77) – wyżsi gracze są cięższi (oczywista fizyka)
- Ocena overall & potencjał (0.64) – obecna forma wpływa na przewidywania

2. Ciekawe braki korelacji
- Wiek & potencjał (-0.26) – młodzi mają tylko nieznacznie wyższy potencjał
- Wzrost & overall (0.043) – wzrost praktycznie nie wpływa na ocenę

3. Niespodzianki
- Wiek & wartość (0.16) – starsi gracze są nieznacznie drożsi
- Potencjał & zarobki (0.5) < Potencjał & overall (0.6) – kluby płacą za raczej za obecną formę niż za potencjał
- Wiek & wartość (0.043) – wiek nie ma żadnego wpływu na wartość



Analiza wykresu regresji liniowej: Relacja wartość-zarobki z podziałem na preferowaną nogę

Wykres przedstawia zależność między dwiema zmiennymi numerycznymi (wartość rynkowa a zarobki) za pomocą dopasowanej linii prostej.
Dodatkowo:

- Punkty: Pokazują rzeczywiste dane zawodników
- Linie regresji: Niebieska (lewonożni) i pomarańczowa (prawonożni) – wskazują ogólny trend
- Cień wokół linii: Przedział ufności (gdzie spodziewamy się większości danych)

Obserwacje

- Wbrew intuicji i wcześniejszym analizom, wykres pokazuje, że lewonożni zawodnicy z tą samą wartością rynkową zarabiają średnio nieco mniej niż prawonożni. Linia regresji dla lewonożnych (niebieska) przebiega niżej niż dla prawonożnych (pomarańczowa).