Konstanty Andrzejczak (276044), Arkadiusz Urbaniak (276034)

Raport
Analiza oraz porównanie kwot transferów i wartości rynkowych zawodników z wykorzystaniem metod statystyki opisowej

1. Wstęp

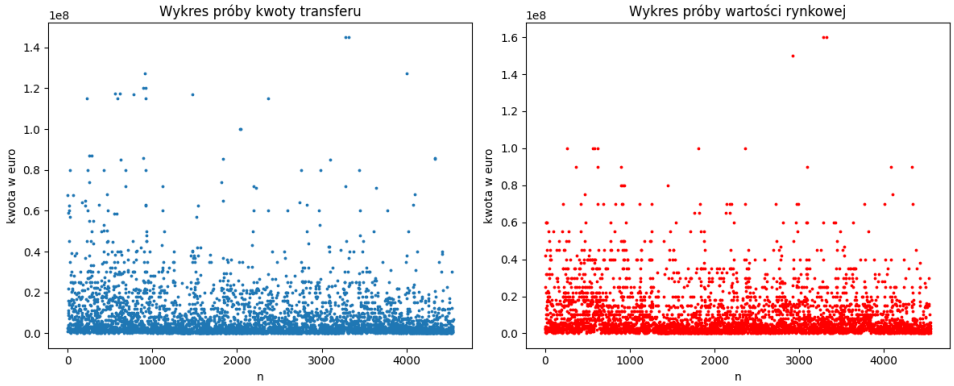
1.1. Cel

Celem raportu jest zbadanie oraz porównanie, przy wykorzystaniu metod statystyki opisowej, dwóch zbiorów danych dotyczących kwot transferów i wartości rynkowych zawodników.

1.2. Dane

Baza danych użyta do analizy została pozyskana z serwisu GitHub `https://github.com/d2ski/football-transfers-data.git`, bazując na danych udostępnionych na stronie Transfermarkt.com. Transfermarkt to platforma internetowa specjalizująca się w danych dotyczących piłki nożnej, obejmująca informacje o zawodnikach, klubach, transferach oraz wartościach rynkowych zawodników. Wartości rynkowe zawodników na Transfermarkcie są obliczane na podstawie różnych czynników, takich jak wiek, umiejętności, forma, statystyki, potencjalny wpływ na drużynę oraz ostatnie transfery podobnych zawodników. Natomiast kwoty transferów opierają się głównie na danych z oficjalnych ogłoszeń transferowych klubów, doniesień prasowych, informacji od agentów zawodników oraz własnej analizie rynku transferowego w celu określenia wartości. Baza danych zawiera informacje o wszystkich transferach zawodników w głównych ligach europejskich (angielskiej, włoskiej, hiszpańskiej, niemieckiej, francuskiej, portugalskiej, holenderskiej) w latach 2009-2021.

Analiza w tym raporcie skoncentrowana została na latach 2018-2021 i obejmuje jedynie transfery pieniężne (nie zawiera transferów za darmo, wymian i wypożyczeń). Zestawienie obejmuje zarówno transfery zawodników opuszczających wspomniane ligi, jak i pozyskanych do nich nowych piłkarzy. Próba składa się z 4553 obserwacji, a wszystkie wymienione kwoty są podane w euro.



Rysunek 1. Wizualizacja danych

2. Miary statystyczne

W tej części zostały umieszczone podstawowe statystyki. W poniższych sekcjach badane próby będą oznaczane odpowiednio:

X - kwota transferu,

Y - wartość rynkowa

2.1. Miary położenia

2.1.1. Średnia arytmetyczna

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

gdzie:

- X_i ozancza daną obserwacje z próby,
- n oznacza rozmiar próby.

Wyniki:

$$\overline{X} \approx 7\,785\,190.51[€]$$

$$\overline{Y} \approx 9\,067\,133.60[€]$$

Średnia arytmetyczna kwoty transferu jest niższa od średniej arytmetycznej wartości rynkowych. Wynika z tego, że dla większości piłkarzy cena za transfery wyniosła mniej niż ich szacowana wartość rynkowa. W obu przypadkach średnia arytmetyczna jest znacznie bardziej zbliżona do minimum, co sugeruje, że większość obserwacji znajduje się poniżej podanej wartości. Z danych wynika także, że przeciętny transfer wynosi około 7 800 000 euro. Dzięki tej informacji klub może szacować swoje wydatki na dany rok. Średnia arytmetyczna jest precyzyjniejsza do tego rodzaju szacunków, ponieważ zespół składa się zarówno z zawodników o większych jak i o mniejszych zdolnościach piłkarskich (większej, bądź mniejszej wartości rynkowej).

2.1.2. Średnia harmoniczna

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \tag{2}$$

Wyniki:

$$H_X \approx 330\,220.75[€]$$

$$H_Y \approx 1\,422\,451.77[€]$$

W przypadku średniej harmonicznej można zauważyć niższe wyniki. Jest to spowodowane wrażliwością tej statystyki na wartości bliższe zeru. Obie próby muszą zawierać znaczną ilość takich wartości. Warto zauważyć, że pośród obu prób ten efekt jest o wiele bardziej widoczny dla próby X, co sugeruje, że zawiera ona więcej małych liczb w porównaniu do próby Y.

2.1.3. Średnia geometryczna

$$G = \sqrt[n]{\prod_{i=1}^{n} X_i} \tag{3}$$

Wyniki:

$$G_X \approx 2\,797\,934.98[€]$$

$$G_Y \approx 3\,914\,842.54[€]$$

Średnia geometryczna jest bardziej odporna na wartości odstające, przez co jej wynik jest zbliżony do mediany($X_{med} = 3\,000\,000[€], Y_{med} = 4\,000\,000[€]$). W kontekście analizowanych danych, wynik ten wskazuje na to, że średnia arytmetyczna, została znacząco zawyżona, przez odstające wartości.

2.1.4. Średnia ucinana

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} X_i \tag{4}$$

gdzie:

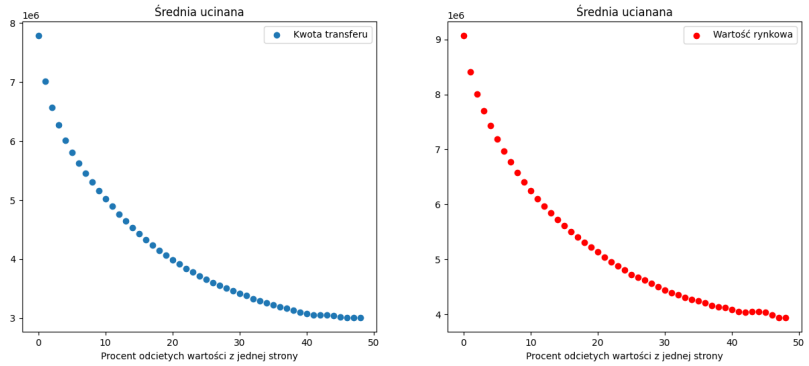
- k oznacza liczbę skrajnych wartości, które ucinamy

Dla przykładu przeprowadziliśmy obliczenia przy ucięciu 25% skrajnych wartości z obu stron. Dla tak zdefinowanego k wyniki prezentują sią następująco:

$$\overline{X}_{0.25} \approx 3\,652\,555.99[€]$$

$$\overline{Y}_{0.25} \approx 4\,725\,516.03[€]$$

Wykresy:



Rysunek 2. Średnia ucinana

Średnia ucinana pokazuje, jak zmienia się wartość średniej wraz z wyrzucaniem (ucinaniem) obustronnie kolejnych odstających wartości. Wraz ze wzrostem ilości odrzucanych obserwacji, wartość średniej dąży do mediany. Natomiast dla parametru k=0 średnia ucinana przyjmuje wartość średniej arytmetycznej. Dla naszych danych możemy zaobserwować, że średnia ta na początku zmniejsza się dużo bardziej. Wskazuje to na to, iż najbardziej odstające wartości wywoływały największe odchylenia średniej. Natomiast z powodu, że średnia ucinana maleje wraz ze wzrostem k, wywnioskować można, że najbardziej odstające wartości były większe od mediany.

2.1.5. Średnia Winsorowska

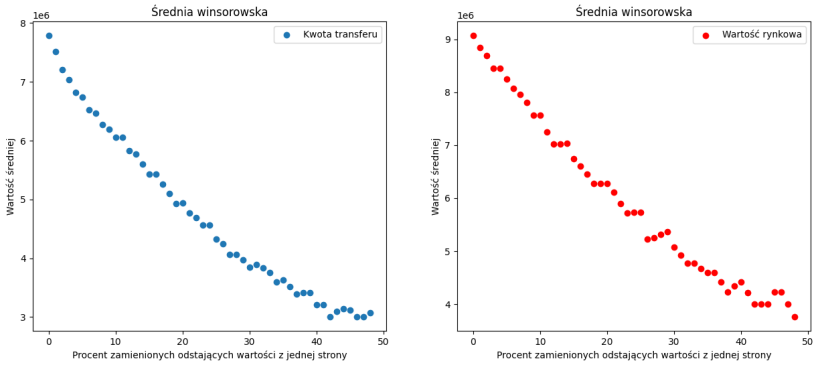
$$\frac{1}{n}\left[(k+1)X_{(k+1)} + \sum_{i=k+2}^{n-k-1} X_{(i)} + (k+1)X_{(n-k)}\right] \tag{5}$$

Jako parametr k ponownie wybraliśmy 25% skrajnych wartości z obu stron. Wyniki wyglądają następująco:

$$\overline{X}_{win[0.25]} \approx 4\,325\,833.88[€]$$

$$\overline{Y}_{win[0.25]} \approx 5\,737\,090.75[€]$$

Wykresy:



Rysunek 3. Średnia Winsorowska

W przypadku średniej winsorowskiej wnioski są bardzo podobne, co do średniej ucinanej. Obie średnie różnią się tym, że w przypadku średniej winsorowskiej, zamiast ucinać skrajnych wartości, zamienia się je na wartości maksymalne i minimalne pozostałego zbioru.

2.1.6. Pierwszy kwartyl (Q1)

$$Q_1 = \begin{cases} X_{((n+1)/4)} & \text{, gdy n jest nieparzyste} \\ \frac{1}{2}(X_{(n/4)} + X_{(n/4-1)}) & \text{, gdy n jest parzyste} \end{cases} \tag{6}$$

Wyniki:

$$Q_{1_X} = 1\,000\,000[€]$$

$$Q_{1_Y} = 1\,500\,000[€]$$

Pierwszy kwartyl mówi nam, że cena transferu najtańszych 25% piłkarzy wyniosła poniżej lub równo 1 milion euro, a ich wartość rynkowa wynosiła mniej lub równo 1,5 miliona euro.

2.1.7. Mediana, drugi kwartyl (Q2)

$$X_{med} = \begin{cases} X_{((n+1)/2)} & \text{, gdy n jest nieparzyste} \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2-1)}) & \text{, gdy n jest parzyste} \end{cases} \tag{7}$$

Wyniki:

$$X_{med} = 3\,000\,000[€]$$

$$Y_{med} = 4\,000\,000[€]$$

W obu przypadkach mediana jest ponad 2 razy mniejsza niż średnia arytmetyczna co może świadczyć o tym, że średnia jest znacząco zawyżona przez mniejszy odsetek, bardzo drogich piłkarzy. Z wartości mediany, wiadomo także, że cena 50 % zawodników jest niższa lub równa 3 000 000 euro w przypadku kwoty transferu i 4 000 000 euro w przypadku wartości rynkowej. Na podstawie tych danych możemy stwierdzić, że przeciętny zawodnik jest wyceniany na większą kwotę niż cena, za jaką został sprzedany/kupiony.

2.1.8. Trzeci kwartyl (Q3)

$$Q_3 = \begin{cases} X_{(3(n+1)/4)} & \text{, gdy n jest nieparzyste} \\ \frac{1}{2}(X_{(3n/4)} + X_{(3n/4-1)}) & \text{, gdy n jest parzyste} \end{cases} \tag{8}$$

Wyniki

$$Q_{3_X} = 9\,000\,000[€]$$

$$Q_{3_Y} = 12\,000\,000[€]$$

Trzeci kwartyl mówi, że cena transferu 25% najdroższych piłkarzy wyniosła powyżej lub równo 9 milionów euro, a ich wartość rynkowa powyżej lub równo 12 milionów euro (cena za transfer $\frac{3}{4}$ piłkarzy poniżej lub równo 9 mln, wartość rynkowa 75% poniżej lub równa 12 mln) Jest to wynik, który nie odstaje bardzo od średniej. Pokazuje to, że cena minimum $\frac{3}{4}$ wszystkich piłkarzy znajduje się względnie blisko w okolicach średniej.

2.2. Miary rozproszenia

2.2.1. IQR (rozstęp międzykwartylowy)

$$IQR = Q_3 - Q_1 \tag{9}$$

Wyniki:

$$IQR_X = 8\,000\,000[€]$$
$$IQR_Y \approx 10\,500\,000[€]$$

Rozstęp międzykwartylowy mówi nam, że różnica między cenami 75% najtańszych piłkarzy (Q1) a 25% najdroższych piłkarzy (Q3) wynosi odpowiednio 8 i 10,5 mln. euro. (zakres, w którym znajduje się środkowa połowa (50%) danych). Te rezultaty są relatywnie wąskie porównując z obszarem max-min. Oznacza to, że dane środkowe są relatywnie blisko siebie, a wyniki zawyża 25% najdroższych piłkarzy.

2.2.2. Rozstęp

$$R = X_{(n)} - X_{(1)} \tag{10}$$

Dla naszych danych

$$X_{\min} = 1000[€] \qquad X_{\max} = 145\,000\,000[€]$$
$$Y_{\min} = 25000[€] \qquad Y_{\max} = 160\,000\,000[€]$$

Wyniki:

$$R_X = 144\,999\,000[€]$$
$$R_Y = 159\,975\,000[€]$$

Rozstęp z próby wskazuje na to, że ceny zawodników przyjmują szeroki zakres.

2.2.3. Wariancja z próby

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{11}$$

Wyniki

$$s_X^2 \approx 168\,721\,451\,989\,415.5[€]$$
$$s_Y^2 \approx 170\,997\,793\,536\,096[€]$$

2.2.4. Odchylenie standardowe z próby

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{12}$$

Wyniki:

$$s_X \approx 12\,990\,709.51[€]$$
$$s_Y \approx 13\,078\,049.38[€]$$

Odchylenie standardowe mówi o tym, że ceny piłkarzy i ich wartość rynkowa rozproszone są od wartości średniej arytmetycznej o około 13 mln. Klub dopłacając kwotę mniej więcej 13 mln. euro statystycznie byłby w stanie zakupić większość zawodników.

2.2.5. Odchylenie przeciętne od wartości średniej

$$d = \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}| \tag{13}$$

Wyniki:

$$d_X \approx 7\,839\,767.11[€]$$

$$d_Y \approx 8\,596\,914.72[€]$$

Przeciętnie kwota transferu zawodnika różni się od średniej arytmetycznej o 7 839 767,11 euro a wartość rynkowa o 8 596 914,72 euro. Wskazuje to, iż większość zawodników powinna mieścić się w takich widełkach cenowych.

2.2.6. Współczynnik zmienności

$$V = \frac{s}{\overline{X}} \cdot 100\% \tag{14}$$

Wyniki:

$$V_X \approx 166,85\%$$

$$V_Y \approx 144,24\%$$

W obu analizowanych zestawach danych wartość współczynnika zmienności jest bardzo duża. Skutkuje to tym, że odchylenie standardowe jest odpowiednio około 1,66 i 1,44 razy większe od średniej arytmetycznej. W przypadku kwoty transferu współczynnik jest wyższy. Duży współczynnik zmienności wskazuje na to, że dane cechują się znacznym zróżnicowaniem, co implikuje, że ilość odstających od średniej piłkarzy jest znacząca dla wyników pozostałych miar statystycznych.

2.3. Miary asymetrii

2.3.1. Współczynnik skośności

$$\alpha = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s}\right)^3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)^{\frac{3}{2}}} \tag{15}$$

Wyniki

$$\alpha_X \approx 4,25$$

$$\alpha_Y \approx 3,43$$

Współczynnik skośności jest dodatni, zatem wykres danych jest prawostronnie skośny. Ze względu na dużą wartość wyniku możemy stwierdzić sporą asymetrię. Z tego powodu wykres cen transferów jest bardziej prawostronnie skośny od wykresu wartości rynkowych. Na podstawie tego współczynnika stwierdzić można również, że większość cen piłkarzy jest niższa od wartości oczekiwanej.

2.4. Miary spłaszczenia

2.4.1. Kurtoza

$$K = \frac{n-1}{(n-2)(n-3)}((n+1)K_a - 3(n-1)) + 3 \tag{16}$$

gdzie:

- $K_a = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2}$

  Wyniki:

$$K_X \approx 29,28$$

$$K_Y \approx 22,27$$

Dodatnia kurtoza oznacza, że liczba wyników odstających jest większa niż w przypadku rozkładu normalnego, a także że histogram jest bardziej "szczytowy". Ponownie, takie dane sugerują, że piłkarze o odstających cenach stanowią istotną grupę w analizie. Wyniki kurtozy wskazują także na to, że oba rozkłady charakteryzują się silną ciężkoogonowością.

2.5. Miary korelacji

2.5.1. Współczynnik korelacji Pearsona

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{17}$$

Wyniki:

$$\rho_{X,Y} \approx 0,72$$

Współczynnik wyszedł większy od zera, więc zestawione dane są ze sobą dodatnio skorelowane, co sugeruje, że wraz ze wzrostem kwoty transferu rośnie także wartość rynkowa. Duża wartość współczynnika wskazuje na silne powiązania między danymi.

2.5.2. Współczynnik korelacji Spearmana

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{18}$$

gdzie:

- $d_i = R_{x_i} - R_{y_i}$,
- R oznacza rangę

Wyniki:

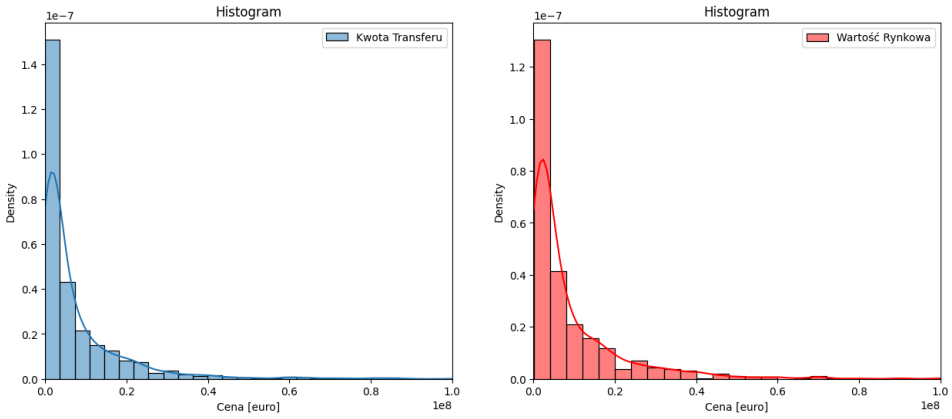
$$r_{S_{X,Y}} \approx 0,68$$

Współczynnik wyszedł większy od zera, więc zestawione dane są ze sobą dodatnio skorelowane, co oznacza, że wraz ze wzrostem kwoty transferu rośnie także wartość rynkowa. Współczynnik korelacji Spearmana jest bardziej

odporny na odstające wartości w porównaniu do współczynnika korelacji Pearsona

3. Wykresy

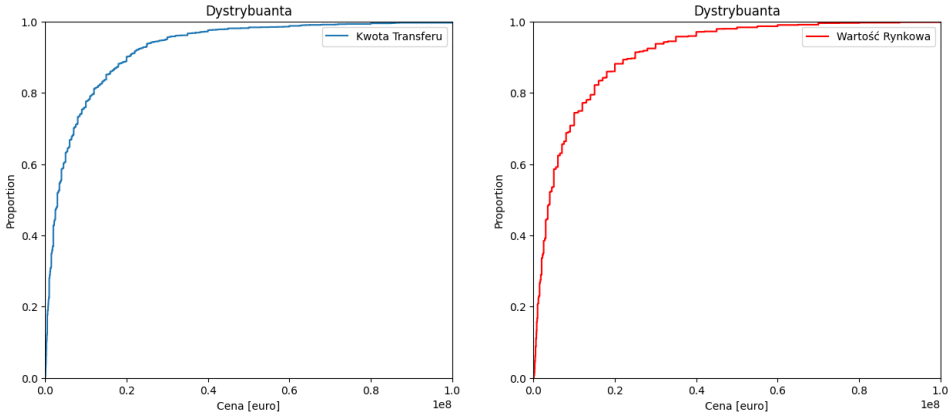
3.1. Gęstość empiryczna



Rysunek 4. Zestawienie gęstości

Analizując oba wykresy gęstości, można zauważyć strome krzywe, co wskazuje na małe rozproszenie większości obserwacji. Dodatkowo, warto zauważyć, że wysokość słupka histogramu jest większa dla kwoty transferu. Na podstawie tego można wywnioskować, że szansa na wybranie piłkarza należącego do przedziału najtańszych zawodników jest w tym przypadku większa.

3.2. Dystrybuanta empiryczna

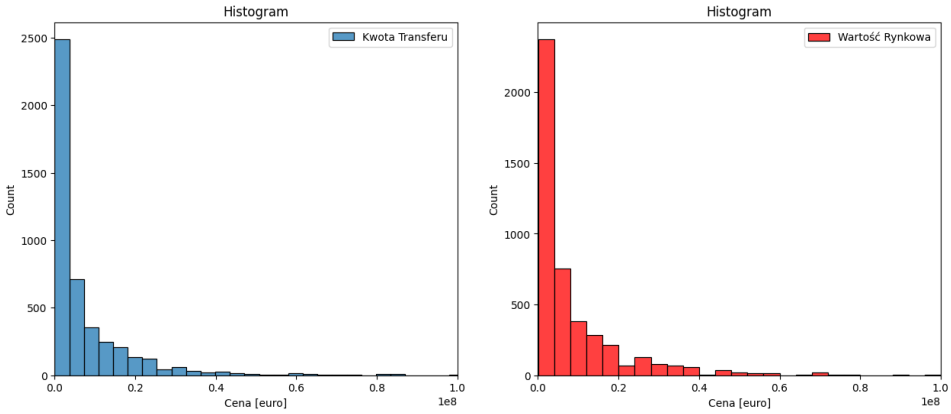


Rysunek 5. Zestawienie dystrybuant

Dystrybuanta kwoty transferu jest bardziej gładka, co świadczyć może o tym, że kwoty transferów rzadziej przyjmują te same wartości, niż ma to miejsce w przypadku wartości rynkowych, gdzie to wartość piłkarza może być zaokrąglana do tych samych kwot. Dystrybuanta w przypadku kwoty transferu na początku rośnie szybciej od dystrybuanty wartości rynkowej,

co oznacza, że ilość piłkarzy sprzedanych za niższą kwotę jest większa. Potwierdzają to wyniki średnich i mediany.

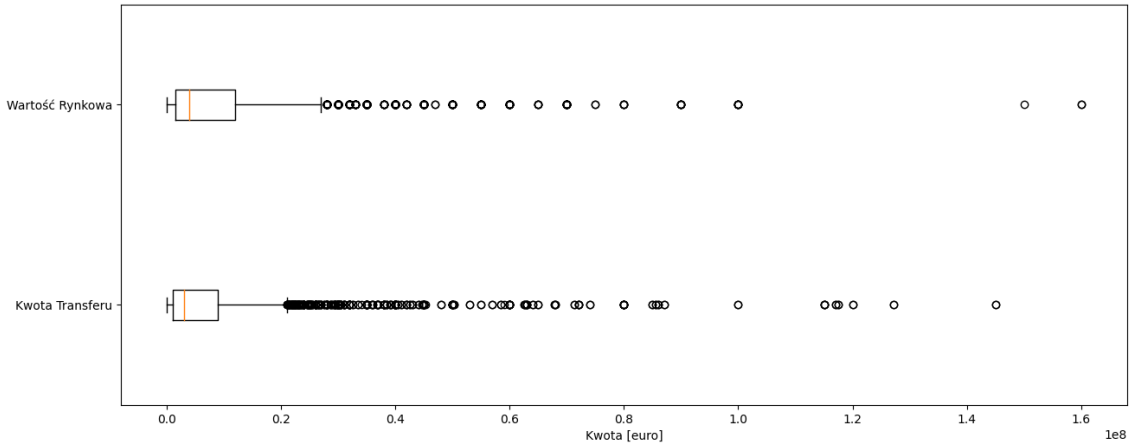
3.3. Histogram liczebności



Rysunek 6. Histogramy liczebności

Analizując histogramy, zauważalne jest, że oba są jednomodalne. Jednak moda histogramu zawiera więcej obserwacji w przypadku kwoty transferu w porównaniu z wartością rynkową. Dodatkowo, oba wykresy wykazują prawostronną skośność. Ten fakt wynika z przewagi niskich kwot transferów bądź wartości rynkowych pośród wszystkich obserwacji.

3.4. Wykres pudełkowy



Rysunek 7. Wykresy pudełkowe

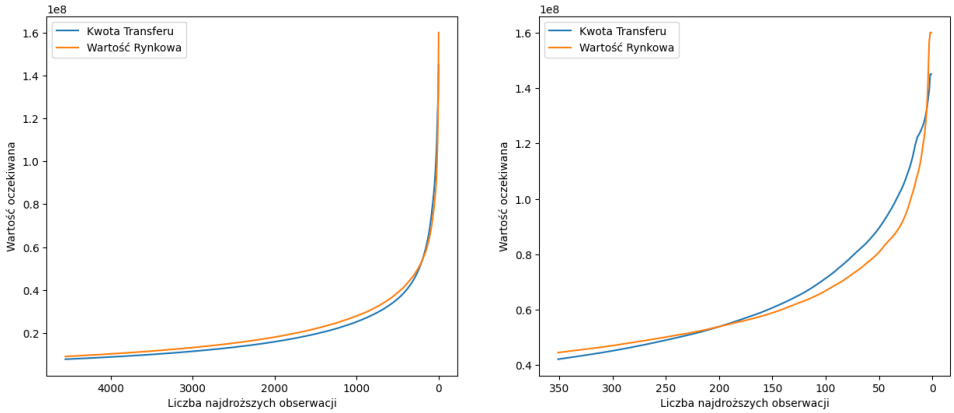
Z wykresów pudełkowych można zobaczyć, że mediany zarówno "kwot transferów", jak i "wartości rynkowych" utrzymują się na stosunkowo zbliżonym poziomie (ich różnica wynosi 1 000 000 €). Natomiast rozstęp międzykwartylowy jest już znacząco większy dla wartości rynkowej, co wskazuje na większe rozproszenie danych. Dodatkowo, można zauważyć brak występowania na obu wykresach pudełkowych niskich odstających wartości. Kwoty transferu mają więcej dodatnich odstających wartości, lecz nie tak znacząco,

jak jest to widoczne na wykresie, ponieważ wiele z wysokich wartości rynkowych się pokrywa, co ma odzwierciedlenie na dystrybuancie. Obecność wielu odstających wartości na obu wykresach potwierdza wcześniejsze wnioski o występowaniu dużej liczby zawodników, przez których ogólne wyniki są zawyżane.

4. Podsumowanie

Analizując wszystkie obserwacje dotyczące kwoty transferu oraz wartości rynkowej, można wywnioskować, że obie grupy danych są ze sobą znacząco powiązane. Wraz ze wzrostem wartości rynkowej rośnie kwota transferu. Spoglądając na miary spłaszczenia oraz miary asymetrii, obserwujemy, że największą grupę zawodników stanowią niedrodzy piłkarze, przy czym należy pamiętać, że odstający od średniej zawodnicy również stanowią znaczącą ilość. Zauważyć można również, że wszystkie obliczone średnie oraz kwartyle są większe w przypadku wartości rynkowej. Świadczy to o tym, że piłkarze sprzedawani są poniżej ich faktycznej, wycenionej wartości.

Jednakże, mimo że wartość oczekiwana, a także mediana kwoty transferu jest niższa od wartości rynkowej, nie oznacza to, że taka tendencja jest prawdziwa na przedziale wszystkich obserwacji. Analizując wykresy pudełkowe, możemy zauważyć, że w "Kwocie Transferu" znajduje się o więcej wartości odstających w porównaniu do "Wartości rynkowej". Na podstawie tego możemy wywnioskować, że wśród najdroższych zawodników tendencja jest odwrotna.



Rysunek 8. (Po lewej) Wartość oczekiwana dla zmieniającej się liczby największych danych, (po prawej) przybliżenie wykresu od 350 największych do 0

Po analizie danych odstających stwierdzić można, że tendencja średniej arytmetycznej większej w przypadku wartości rynkowej odwraca się dla około 200 najdroższych piłkarzy, tam średnia arytmetyczna jest większa w przypadku kwoty transferu i różnica ta zwiększa się wraz ze zmniejszeniem próbki. Kwota transferu takich piłkarzy często przewyższa ich wartość rynkową. Obserwacje te mają pokrycie w rzeczywistości, ponieważ klubą opłaca się zapłacić więcej za zawodnika, niż wskazuje na to wartość rynkowa, która szacowana jest tylko pod względem poziomu piłkarskiego, ponieważ znany zawodnik posiada także wartość medialną. Także najlepszych zawodników jest o wiele mniej, przez co kluby same napędzają na nich popyt, przez co ich cena znacząco wzrasta względem wartości rynkowej. Cena transferów

zawodników przeciętnych jest statystycznie niższa od ich wartości rynkowej, ponieważ kluby mają duży wybór wśród takich zawodników, przez co klub może zaoferować niższą kwotę za zawodnika ponieważ wie, że w razie nieudanych negocjacji ma do wyboru jeszcze wielu zawodników reprezentujących podobny poziom gry.