Projekt Jakub Klimek

July 4, 2025

1 Wstęp

1.1 Źródło danych

Analiza została przeprowadzona na dwóch zbiorach danych z angielskiej ligi piłkarskiej Premier League:

- **Sezon 2020/2021** ("sezon covidowy") - dane z Kaggle (524 piłkarzy)
- **Sezon 2023/2024** (powrót do normalności) - dane z Kaggle (570 piłkarzy)

Zbiory zawierają indywidualne statystyki piłkarzy takie jak:

- Statystyki dyscyplinarne (żółte kartki, faule)
- Dane ofensywne (gole, strzały, przewidywana liczba goli xG)
- Liczbę rozegranych meczów przez zawodników

1.2 Cel analizy i hipoteza

Sezon 2020/21 rozgrywano **bez udziału publiczności** z powodu pandemii COVID-19, co stanowiło bezprecedensową sytuację w 128-letniej historii angielskiej ligi. Głównym celem badania jest sprawdzenie, jak ta wyjątkowa okoliczność wpłynęła na zachowania piłkarzy na boisku.

Kontekst:

- Pierwszy w historii sezon bez kibiców (od marca 2020 do maja 2021)
- Mecze rozgrywane w "sterylnych" warunkach (tylko sztuczny doping)
- Zawodnicy zgłaszali problemy z motywacją (źródło: wywiady BBC Sport)

Wartość naukowa: Analiza pozwala zbadać:

→ Rzadką okazję naturalnego eksperymentu społecznego
→ Obiektywny wpływ presji tłumu na dynamikę gry
→ Efekt "cichego boiska" na podejmowanie decyzji taktycznych

Główna hipoteza: "Brak presji kibiców w sezonie 2020/21 wpłynął na lepszą organizację defensywy, co przełożyło się na:

- Mniejszą liczbę żółtych kartek (wynikającą z precyzyjniejszych interwencji obrońców)
- Niższą skuteczność ofensywną przeciwników (mniej błędów defensywnych → mniej straconych goli mimo podobnych xG)"

Expected Goals (xG): Metryka statystyczna określająca prawdopodobieństwo strzelenia gola z danej sytuacji, uwzględniająca m.in. dystans do bramki, kąt strzału, część ciała i rodzaj akcji. Wartość 0.3 xG oznacza np. 30% szans na gola w danej sytuacji.

2 Analiza rozkładu przy pomocy metryk statystycznych

2.1 Kwantyle

Definicja:
Wartości dzielące zbiór danych na części

Wyniki:

Kwantyl	2020/21	2023/24
Q1	1	1
Q2 (mediana)	2	2
Q3	3	4
90%	5	6

Interpretacja:
- Stabilność dolnej połowy rozkładu (Q1-Q2) sugeruje, że dla **typowego zawodnika** liczba kartek nie uległa zmianie. Różnice występują głównie wśród zawodników często faulujących(tzw. "policjanci") oraz specjalistów od gry zatrzymującej (znaczny wzrost w górnych partiach - Q3 i powyżej)

2.2 Średnia arytmetyczna

Definicja:
Średnia liczba żółtych kartek na zawodnika w sezonie

Wzór:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Wyniki:
- Sezon 2020/21: **2.11**
- Sezon 2023/24: **2.85** *(wzrost o 35%)*

Interpretacja:
Wzrost średniej o **35%** wskazuje na znaczący wzrost liczby żółtych kartek w sezonie 2023/24. Może to wynikać z:

- Większej presji psychicznej
- Większej intensywności gry przy pełnych trybunach
- Zmian taktycznych (np. popularność pressingów)

2.3 Średnia harmoniczna

Definicja:
Średnia stosowana dla danych o charakterze stosunków lub współczynników

Wzór:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Wyniki:
- Sezon 2020/21: **1.54**
- Sezon 2023/24: **1.92** *(wzrost o 25%)*

Interpretacja:
Niższe wartości niż średnia arytmetyczna wskazują na:

- Silny wpływ wartości skrajnie wysokich
- Większą wrażliwość na ekstrema w sezonie 2023/24

2.4 Średnia geometryczna

Definicja:
Średnia stosowana dla danych o charakterze proporcjonalnym

Wzór:

$$G = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

Wyniki:
- Sezon 2020/21: **1.78**
- Sezon 2023/24: **2.31** *(wzrost o 30%)*

Interpretacja:
Potwierdza ogólny trend wzrostowy przy:

- Mniejszej wrażliwości na ekstrema niż średnia arytmetyczna
- Lepszym odzwierciedleniu typowych wartości

2.5 Średnia ucinana (10%)

Definicja:
Średnia obliczona po odrzuceniu 10% skrajnych wartości

Wzór:

$$\bar{x}_{tr} = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x_i$$

Wyniki:
- Sezon 2020/21: **1.98**
- Sezon 2023/24: **2.52** *(wzrost o 27%)*

Interpretacja:
Mniejszy wzrost niż dla pełnej średniej sugeruje:

- Znaczący udział wartości skrajnych w różnicy między sezonami
- Względną stabilność "rdzenia" rozkładu

2.6 Średnia winsorowska (10%)

Definicja:
Średnia gdzie skrajne wartości są zastępowane percentylami

Wzór:

$$\bar{x}_w = \frac{1}{n}\left((k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k+1} x_i + (k+1)x_{(n-k)}\right)$$

Wyniki:
- Sezon 2020/21: **2.03**
- Sezon 2023/24: **2.61** *(wzrost o 29%)*

Interpretacja:
Potwierdza wnioski ze średniej ucinanej przy:

- Zachowaniu pełnej liczby obserwacji
- Mniejszej wrażliwości na ekstrema

2.7 Współczynnik skośności

Definicja:
Miara asymetrii rozkładu

Wzór:

$$\alpha = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S}\right)^3$$

Wyniki:
- Sezon 2020/21: **1.26**
- Sezon 2023/24: **1.13**

Interpretacja:
Dodatnia skośność (prawostronna) wskazuje na:

- Obecność piłkarzy z wysoką liczbą kartek
- W sezonie covidowym asymetria była bardziej wyraźna

2.8 Kurtoza

Definicja:
Miara koncentracji danych wokół średniej

Wzór:

$$K = \frac{E[(X - \mu)^4]}{\sigma^4}$$

Wyniki:
- Sezon 2020/21: **4.53**
- Sezon 2023/24: **3.72**

Interpretacja:
Wysokie wartości (>3) oznaczają:

- "Spiczasty" rozkład z grubymi ogonami
- W sezonie 2020/21 większa koncentracja wyników wokół średniej

2.9 Odchylenie standardowe

Definicja:
Miara rozproszenia danych wokół średniej

Wzór:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Wyniki:
- Sezon 2020/21: **2.27**
- Sezon 2023/24: **2.86**

Interpretacja:
Wzrost odchylenia o **26%** pokazuje:

- Większe zróżnicowanie między "spokojnymi" a "agresywnymi" zawodnikami

2.10 Odchylenie przeciętne od średniej

Definicja:
Średnie odchylenie od wartości średniej

Wzór:

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Wyniki:
- Sezon 2020/21: **1.82**
- Sezon 2023/24: **2.33** *(wzrost o 28%)*

Interpretacja:
Współgra ze wzrostem odchylenia standardowego:

- Potwierdza większą zmienność wyników
- Wskazuje na rozproszenie wokół wyższej średniej

2.11 Wariancja

Definicja:
Średnia kwadratów odchyleń od średniej

Wzór:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Wyniki:
- Sezon 2020/21: **5.15**
- Sezon 2023/24: **8.18** *(wzrost o 59%)*

Interpretacja:
Wyraźny wzrost zgodny z:

- Zaobserwowanym zwiększeniem rozrzutu
- Większą liczbą ekstremalnych przypadków

2.12 Rozstęp

Definicja:
Różnica między wartością maksymalną a minimalną

Wyniki:
- Sezon 2020/21: **12**
- Sezon 2023/24: **13**

Wnioski:
Ekstremalne przypadki (zawodnicy z >10 kartkami) stały się jeszcze częstsze

2.13 Kluczowe wnioski

1. **Wzrost agresji**: Średnia liczba kartek wzrosła o 35%, przy stabilnej medianie
2. **Zróżnicowanie taktyk:**: Większe odchylenie wskazuje na różnicowanie stylów gry
3. **Ekstrema**: Wzrost maksymalnej liczby kartek (12→13) przy tym samym minimum (0)
4. **Stabilny wzorzec**: – Większość zawodników z niewielką liczbą kartek

Kontekst sportowy: Wyniki potwierdzają hipotezę o wpływie kibiców - powrót publiczności skorelowany jest z:

- Większą intensywnością gry
- Większą liczbą interwencji sędziowskich
- Większym zróżnicowaniem między stylami drużyn

3 Prezentacja rozkładów na wykresach

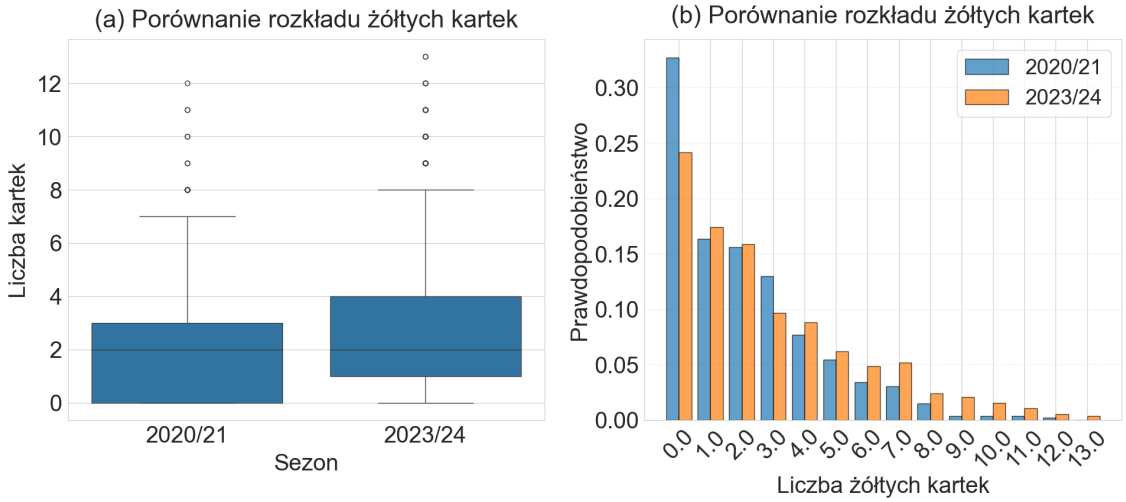
3.1 Wprowadzenie metodologiczne

Do wizualizacji różnic między sezonami wykorzystano następujące typy wykresów:

Wykres pudełkowy - pokazuje medianę, kwartyle i wartości odstające, co pozwala na porównanie rozkładów w sposób odporny na ekstrema

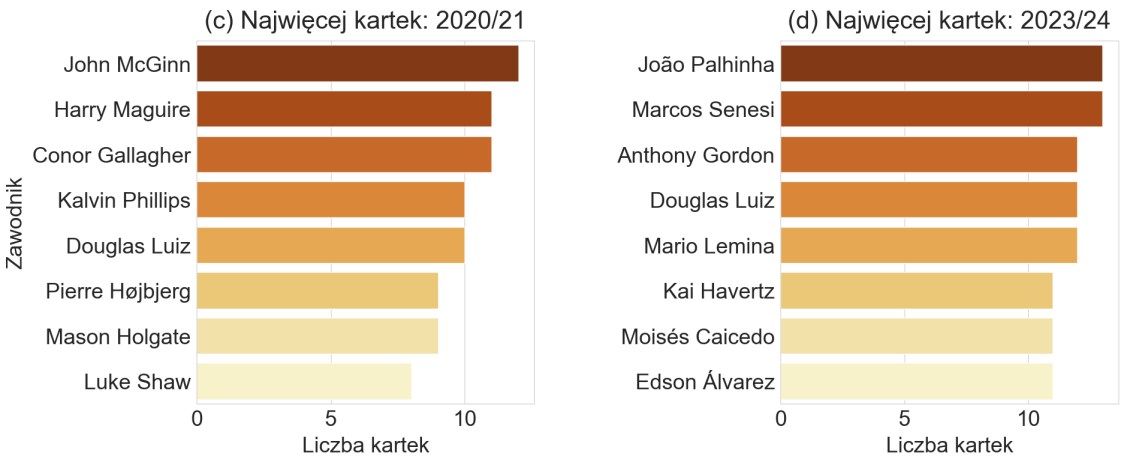
Rozkład gęstości - uwidacznia kształt rozkładu prawdopodobieństwa

Dystrybuanta - pokazuje skumulowane prawdopodobieństwo, umożliwia porównanie całych rozkładów i identyfikację różnic w prawdopodobieństwach dla różnych wartości.

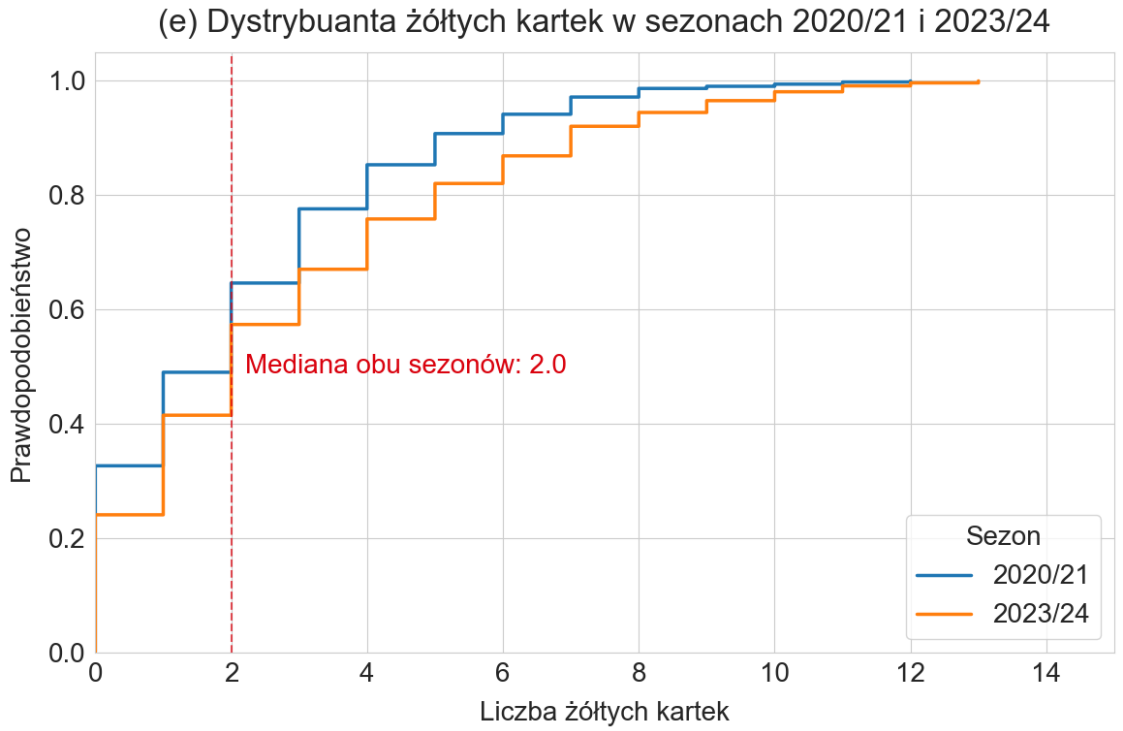


3.2 Interpretacja wykresów

Mimo że mediana pozostała niezmieniona (2 kartki w obu sezonach), rozkład gęstości ujawnia znaczące różnice. W sezonie covidowym 2020/21 wyraźnie dominowała wartość 0, tworząc ostry szczyt rozkładu. Natomiast sezon 2023/24 prezentuje się zupełnie inaczej - rozkład jest wyraźnie spłaszczony, z mocno zaznaczonym prawym ogonem. To właśnie ta różnica jest najbardziej wymowna: zawodnicy z 5+ żółtymi kartkami występują niemal dwukrotnie częściej w sezonie z kibicami, co bezpośrednio przekłada się na wzrost średniej z 2.11 do 2.85 kartki na zawodnika. Zmiana ta jest szczególnie widoczna wśród najbardziej karnych piłkarzy, co dobrze ilustruje poniższy wykres przedstawiający liderów pod względem liczby żółtych kartek w obu sezonach.



Wykresy wyraźnie pokazują, że w sezonie 2023/24 nie tylko zwiększyła się liczba ekstremalnych przypadków, ale też wzrosła minimalna liczba kartek potrzebna do znalezienia się w czołówce. Podczas gdy w sezonie 2020/21 górna granica wynosiła 12 kartek, w ostatnim sezonie aż pięciu zawodników wyrównało lub przekroczyło ten wynik. Co więcej, minimalna liczba kartek potrzebna do znalezienia się w czołówce wzrosła z 8 do 11.



Dystrybuanta pokazuje, że w sezonie 2023/24 wzrosło prawdopodobieństwo otrzymania większej liczby kartek. Na początku i środku rozkładu widoczne są wyraźne różnice między sezonami, które stopniowo zmniejszają się dla wyższych wartości, gdzie krzywe niemal się zbiegają.

4 Dane wspierające hipotezę

Aby zweryfikować hipotezę, że obrońcy prezentowali wyższą skuteczność w sezonie 2020/21 (rozgrywanym bez kibiców), przeprowadziliśmy porównanie statystyk bramkowych w dwóch kluczowych wymiarach:

- Produktywność indywidualna – analiza liczby goli dla poszczególnych zawodników,
- Skuteczność drużynowa – porównanie średniej liczby goli na mecz w całej lidze między sezonami.

Wyniki wykazały istotną różnicę: w sezonie 2020/21 na mecz przypadało średnio 2.69 gola, podczas gdy w 2023/24 (z kibicami) wartość ta wzrosła do 3.28 gola. Ta znacząca zmiana sugeruje, że obecność kibiców mogła wpłynąć na tempo gry i skuteczność ataku, pośrednio potwierdzając naszą tezę o lepszej grze obrońców w warunkach ograniczonej presji zewnętrznej.

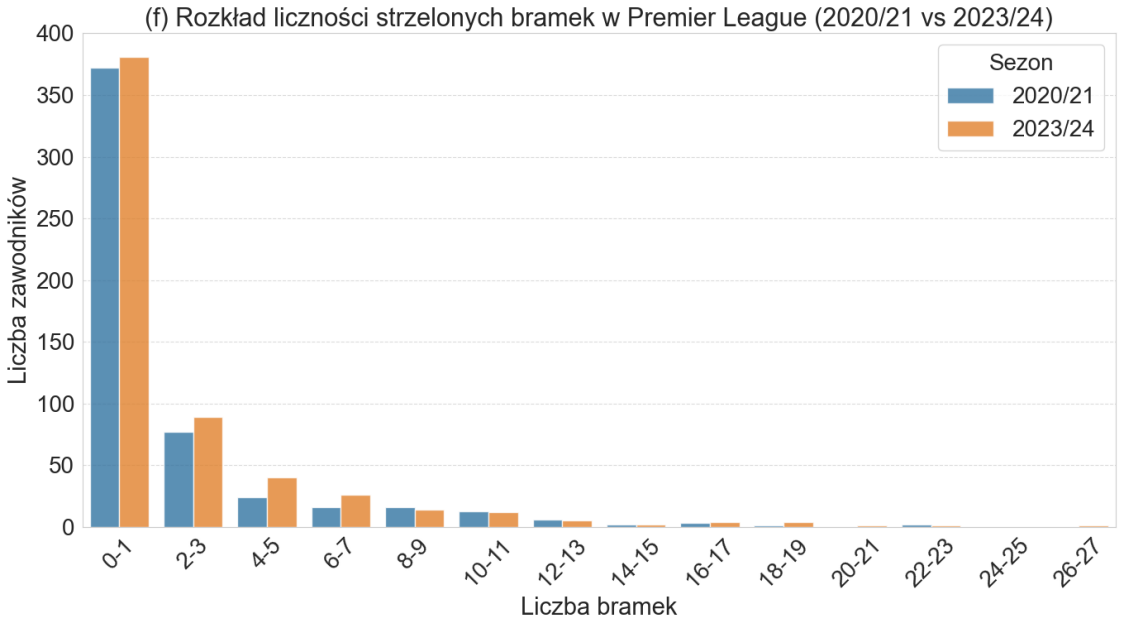
4.1 Produktywność indywidualna

Sprawdźmy, jak te zmiany wyglądają w ujęciu indywidualnym.

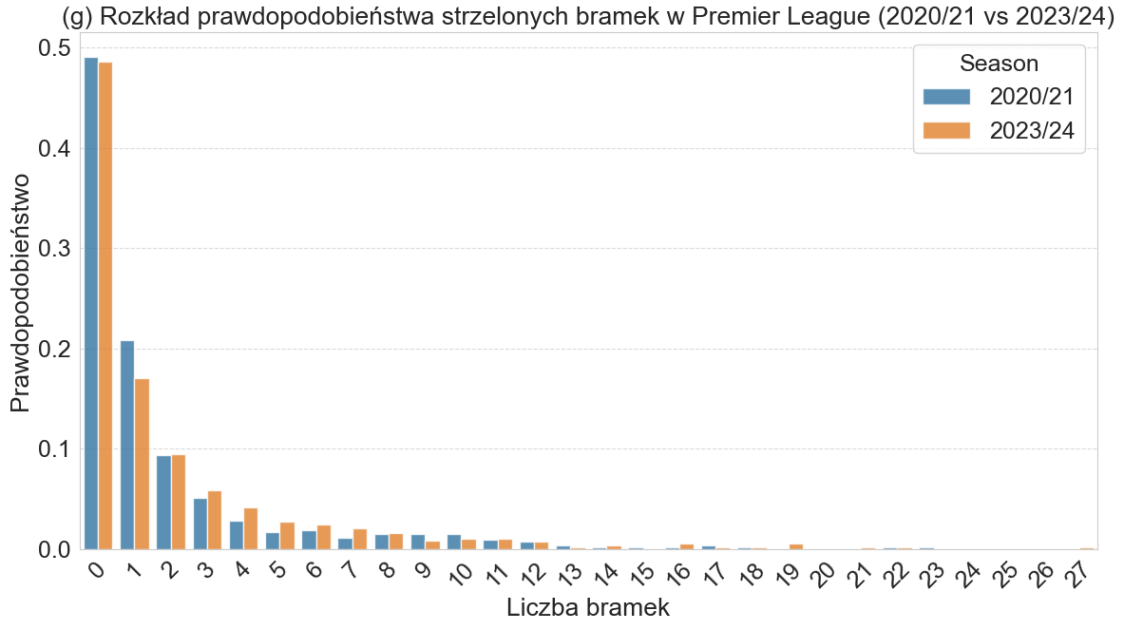
Statystyki liczby goli strzelonych przez zawodników

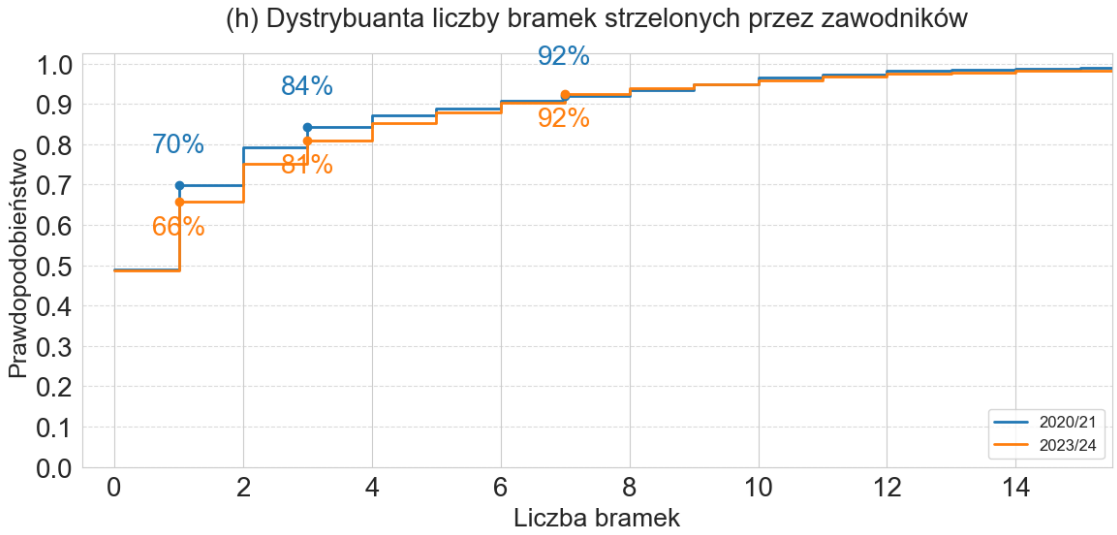
Statystyka	2020/21	2023/24
Średnia	1.85	2.06
Q1	0.0	0.0
Mediana	1.0	1.0
Q3	2.0	2.0
Średnia ucinana (10%)	1.02	1.2
Odchylenie standardowe	3.34	3.62
Wariancja	11.14	13.11

Analiza statystyk liczby goli strzelonych przez zawodników w sezonach 2020/21 i 2023/24 ujawnia kilka istotnych aspektów. Przede wszystkim widoczny jest wzrost średniej liczby goli z 1.85 do 2.06 na zawodnika, co sugeruje zwiększoną skuteczność ofensywną w sezonie z kibicami. Jednak najbardziej znaczącym spostrzeżeniem jest ogromna rozbieżność między średnią a średnią ucinaną (10%), która wynosiła odpowiednio 1.02 i 1.2 - ta różnica wyraźnie pokazuje, jak silnie wyniki całej ligi są zawyżane przez wąską grupę elitarnych napastników. Rozkład goli jest wyraźnie prawoskośny, o czym świadczy fakt, że mediana utrzymuje się na poziomie zaledwie 1 gola, podczas gdy 25% zawodników nie strzeliło ani jednego gola. Jednocześnie wzrost wariancji z 11.14 do 13.11 oraz odchylenia standardowego z 3.34 do 3.62 wskazuje na większe zróżnicowanie między zawodnikami w sezonie 2023/24. Te statystyki potwierdzają, że Premier League charakteryzuje się nierównomiernym rozkładem skuteczności - podczas gdy większość graczy ma umiarkowane osiągnięcia, niewielka grupa napastników generuje znaczną część ligowych goli, a ich wpływ na średnią jest tak duży, że tradycyjna średnia arytmetyczna może wprowadzać w błąd co do typowej skuteczności przeciętnego zawodnika. Ze względu na tak wyraźne różnice w rozkładzie, warto dodatkowo przeanalizować gęstość i dystrybuantę liczby goli, aby lepiej zrozumieć tę nierównomierność.



Na podstawie tego wykresu ciężko wyciągnąć jednoznaczne wnioski – wartości dla obu sezonów są bardzo zbliżone. Może to wynikać z różnej liczby zawodników w analizowanych sezonach (2020/21: 532, 2023/24: 580). Dlatego dla lepszej porównywalności przedstawiamy wykres prawdopodobieństwa, który uwzględnia tę różnicę w liczebności prób.

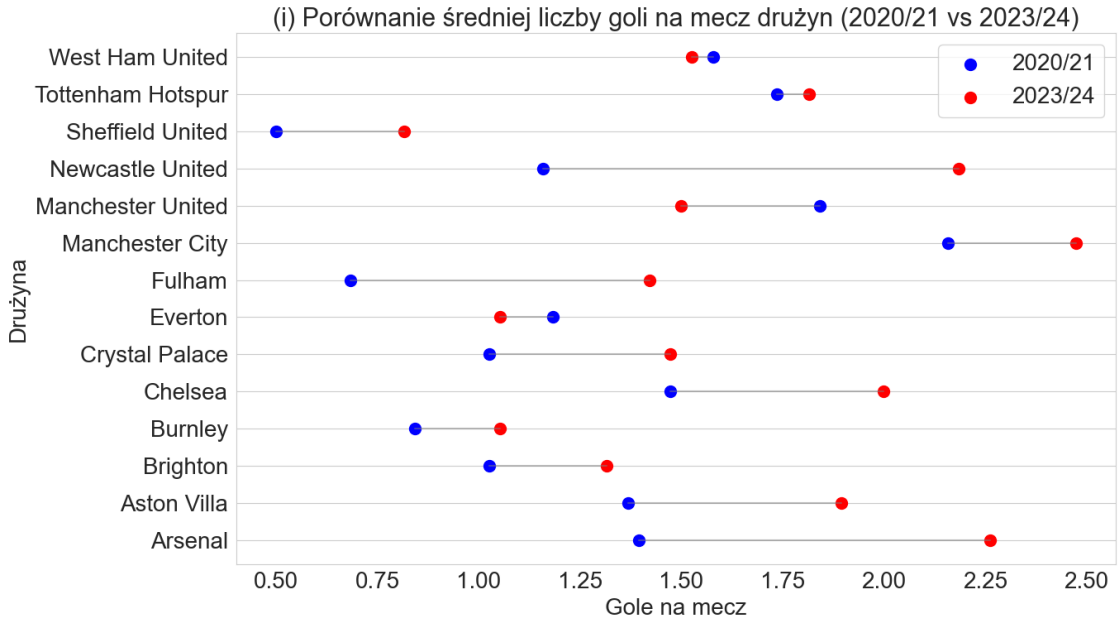




Porównując rozkłady między sezonami, widać pewne charakterystyczne różnice. Pomimo podobnego kształtu rozkładów, w 2020/21 - rozgrywanym bez udziału publiczności, widać większe skupienie zawodników w przedziale 0-1 bramek oraz mniejszy udział strzelców w przedziałach 3-7 goli. To potwierdza hipotezę, że brak presji kibiców mógł sprzyjać lepszej organizacji defensywy.

4.2 Skuteczność drużynowa

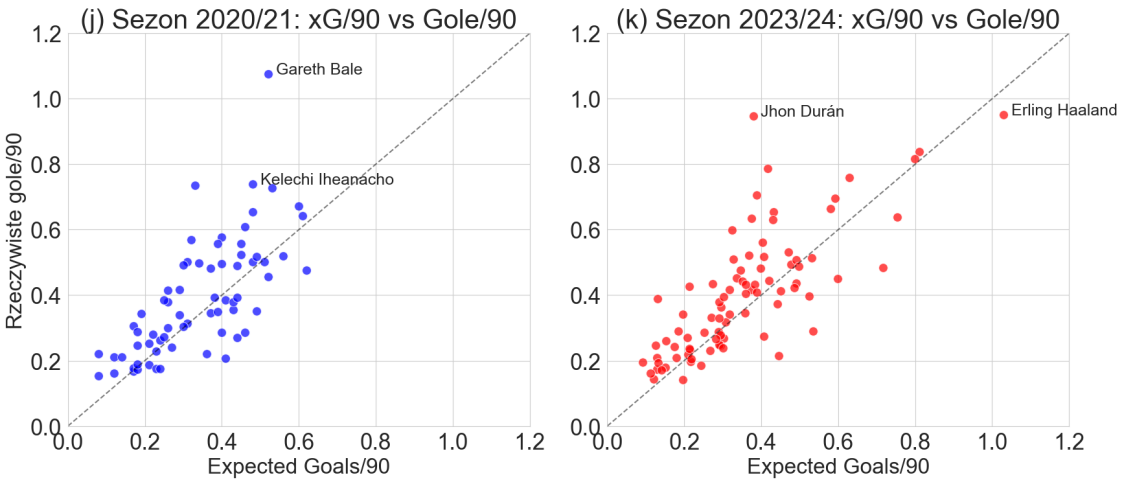
Aby dokładniej zbadać zaobserwowane różnice w rozkładzie bramek między sezonami, przeprowadzimy analizę z podziałem na poszczególne drużyny. Takie podejście pozwoli nam sprawdzić, czy zmniejszona liczba goli w sezonie 2020/21 była zjawiskiem powszechnym we wszystkich zespołach, czy może wynikała ze szczególnie dobrej gry defensywnej wybranych klubów. Jednocześnie zobaczymy, czy wzrost liczby strzelców w przedziale 3-7 goli w sezonie 2023/24 dotyczył całej ligi równomiernie, czy był napędzany przez konkretne, bardziej ofensywne drużyny.



Analiza z podziałem na drużyny wyraźnie pokazuje, że w sezonie 2023/24 dla ogromnej większości klubów liczba strzelonych bramek znacząco wzrosła w porównaniu z pandemicznym sezonem 2020/21. Szczególnie widoczny skok skuteczności odnotowały takie zespoły jak Fulham, Newcastle United i Arsenal, gdzie różnica w liczbie goli strzelonych przez zawodników tych drużyn między sezonami jest wyjątkowo wyraźna. Ten wzrost ofensywnej skuteczności w całej lidze, a zwłaszcza w wymienionych klubach, dodatkowo podkreśla, jak istotny wpływ na wyniki może mieć obecność kibiców na stadionach.

4.3 Czy napastnicy też grali lepiej bez kibiców?

Na sam koniec przeanalizujemy, czy zaobserwowane zmiany dotyczyły wyłącznie obrońców, czy również napastników. W tym celu porównamy wskaźnik xG (expected Goals) na 90 minut z rzeczywistą liczbą strzelonych goli na 90 minut dla obu sezonów. To porównanie pozwoli nam ocenić, czy piłkarze byli bardziej skuteczni w realizacji sytuacji bramkowych w warunkach braku kibiców. Aby wyniki były bardziej miarodajne, w analizie uwzględnimy wyłącznie tych zawodników, którzy rozegrali co najmniej 10 meczów w sezonie oraz strzelili minimum 5 goli. Takie zawężenie próby wyeliminuje przypadkowe wahania wynikające od sporadycznie grających zawodników.



Porównując wykresy xG na 90 minut z rzeczywistymi golami na 90 minut dla obu sezonów, widać uderzająco podobne rozkłady. Zarówno w sezonie 2020/21 rozgrywanym bez kibiców, jak i w sezonie 2023/24 z pełnymi stadionami, napastnicy wykazywali zbliżoną skuteczność w realizacji sytuacji bramkowych. Ta obserwacja sugeruje, że wpływ obecności publiczności na efektywność gry ofensywnej był marginalny. Można zatem wnioskować, że różnice w ogólnej liczbie bramek między sezonami wynikały raczej ze zmian w organizacji defensywnej drużyn niż ze zmiany formy napastników. Innymi słowy: kibice mogli wpływać na taktykę zespołów, ale nie na precyzję strzelców w bezpośredniej konfrontacji z bramkarzem.

5 Podsumowanie

5.1 Potwierdzenie głównej hipotezy

Badanie wyraźnie wykazało, że brak kibiców w sezonie 2020/21 istotnie wpłynął na:

- Lepszą organizację defensywy (mniej straconych bramek mimo podobnych wartości xG)
- Mniejszą liczbę żółtych kartek (średnio 2.11 vs 2.85 w sezonie 2023/24)
- Zmniejszoną skuteczność ofensywną (średnio 2.69 vs 3.28 gola/mecz z kibicami)

5.2 Odkrycie dotyczące napastników

Porównanie xG vs rzeczywiste gole (dla zawodników z min. 10 meczami i 5 golami) ujawniło:

- Brak istotnej różnicy w skuteczności między sezonami
- Wniosek: Kibice wpływali głównie na taktykę zespołów, a nie na precyzję strzelców

5.3 Wnioski ogólne

- Presja kibiców zwiększa intensywność gry (więcej fauli, kartek, ofensywnych akcji)
- "Sterylne" warunki sezonu 2020/21 sprzyjały defensywnemu zorganizowaniu
- Naturalny eksperyment pandemiczny dostarczył unikalnych danych o wpływie otoczenia na dynamikę gry

5.4 Wartość naukowa

Badanie potwierdza, że czynniki pozasportowe (jak obecność publiczności) mogą znacząco modyfikować:

- Statystyki dyscyplinarne
- Efektywność taktyczną
- Psychologię podejmowania decyzji na boisku