Analiza zmian statystyk ligi NBA od początku lat 90-tych oraz ich porównanie

Kacper Balbus

15 czerwca 2023

Spis treści

1 Wstęp **3**

2 Słownik **3**

3 Zbiór danych i jego przetwarzanie **3**
3.1 Zbiór danych 3
3.2 Przetwarzanie wstępne 4
3.3 Przykładowe dane 4

4 Ekspolracja danych **5**
4.1 Zmiany statystyk na przestrzeni sezonów 5
4.2 Zależności między statystykami 15
4.3 Analiza zależności między statystykami 16

5 Modelowanie **32**
5.1 Przygotowanie danych 32
5.2 Dobór modelu 32
5.3 Eksperymenty na modelu 35

6 Wnioski **37**

1 Wstęp

Liga NBA (National Basketball Association) jest najbardziej rozpoznawalną i oglądaną ligą koszykówki na całym świecie. Od początku lat 90-tych do dziś, przeszła ona wiele zmian, zarówno pod względem stylu gry, jak i statystyk osiąganych przez zawodników. Coraz częściej pojawiają się opinie sugerujące, że liga zmierza w złym kierunku, jej poziom się obniżył, a jej "złotymi czasami"były właśnie lata 90-te. Zestawienie i analiza wszytskich danych, powinna w widoczny sposób przedstawić nam, gdzie najbardziej widać różnicę teraźniejszości względem lat 90-tych oraz czy rzeczywiście, w tamtych czasach, na boiskach NBA panował wyższy poziom. Celem projektu jest zbadanie, jak zmieniały się poszczególne statystyki od sezonu 1990-91 do sezonu 2022-23, znalezienie korelacji pomiędzy poszczególnymi statystykami, stworzenie modelu, który umożliwi określenie czy drużyna osiągająca konkretne statystyki, w podanym sezonie zasadniczym, byłaby w stanie dostać się do fazy play-off oraz przy jego pomocy sprawdzenie, jak poradziłyby sobie drużyny, gdyby osiągnęły konkretne statystyki w podanym sezonie. Naszym głównym narzędziem do odczytywania danych oraz pracy na nich, będzie Python.

2 Słownik

- FG (Field goals) - liczba udanych rzutów.
- 3P (3 pointers) - liczba udanych rzutów za 3 punkty.
- FT (Free throws) - liczba udanych rzutów osobistych.
- TRB (Total rebounds) - liczba udanych zbiórek w ataku i obronie.
- AST (Assists) - liczba asyst.
- STL (Steals) - liczba odbiorów.
- BLK (Blocks) - liczba bloków.
- TOV (Turnovers) - liczba strat.
- PF (Personal fouls) - liczba popełnionych fauli.
- PTS (Points) - liczba punktów.

3 Zbiór danych i jego przetwarzanie

3.1 Zbiór danych

Do rozwiązania problemu skorzystałem ze zbiorów danych dostępnych na stronie internetowej basketball-reference.com:

1. Zbiór "NBA League Averages - Per Game"zawiera dane średnich statystyk całej ligi, w jednym meczu, w każdym sezonie od 1990-91. Posłuży on do obserwacji zmian poszczególnych statystyk na przestrzeni lat oraz znalezienia korelacji pomiędzy nimi.

2. Zbiory "Per Game Stats"(jeden zbiór dotyczy jednego sezonu) zawierają średnie statystyki każdej drużyny, z sezonu zasadniczego, na jeden mecz oraz informację, czy dana drużyna dostała się do fazy play-off. Te zbiory zostaną przez nas wykorzystane do opracowania modelu i wykonania na nim eksperymentów.

3.2 Przetwarzanie wstępne

1. Ze zbioru "NBA League Averages - Per Game", korzystając z programu Excel, usunięto niepotrzebne kolmuny danych. Po zmianach zawiera on kolumnę określającą sezon, z którego pochodzi wiersz oraz kolumny zawierające średnie statystyki wymienione w słowniku.

2. Zbiory "Per Game Stats"zostały połączone w jeden zbiór, zawierajacy średnie statystyki drużyn z każdego sezonu od 1990-91. Korzystając z programu Excel dodane zostały dwie dodatkowe kolumny:

   - Season - określająca z jakiego sezonu pochodzi konkretny wiersz.
   - Playoffs - określająca, binarnie ("YES"lub "NO"), czy dana drużyna dostała się do fazy play-off. Wcześniej określane to było w kolumnie Team i symbolizowane było gwiazdką występującą po nazwie zaspołu.

   Następnie usunięto niepotrzebne kolumny danych, aby pozostały jedynie wcześniej stworzone oraz zawierające średnie statystyki wymienione w słowniku.

3.3 Przykładowe dane

1. "NBA League Averages - Per Game"

Season	**FG**	**3P**	**FT**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**
2022-23	42.0	12.3	18.4	43.4	25.3	7.3	4.7	14.1	20.0	114.7
2021-22	40.6	12.4	16.9	44.5	24.6	7.6	4.7	13.8	19.6	110.6
2020-21	41.2	12.7	17.0	44.3	24.8	7.6	4.9	13.8	19.3	112.1

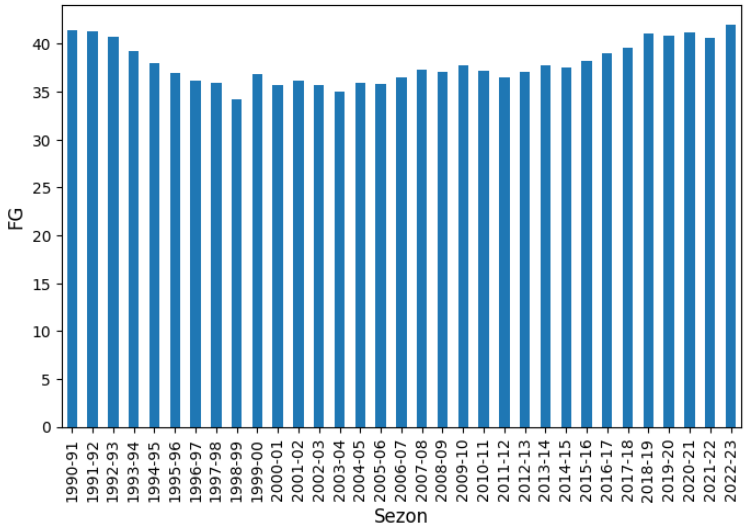
2. "Per Game Stats"

Season	**FG**	**3P**	**FT**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**	**Playoffs**
2022-23	43.6	13.8	19.8	42.5	27.3	7.0	3.4	13.5	19.7	120.7	YES
2022-23	42.0	13.6	18.7	41.5	27.0	7.7	5.8	14.9	21.2	116.3	NO
2002-03	32.8	2.8	15.8	42.4	21.2	8.7	5.1	18.5	25.1	84.2	NO

4 Ekspolracja danych

4.1 Zmiany statystyk na przestrzeni sezonów

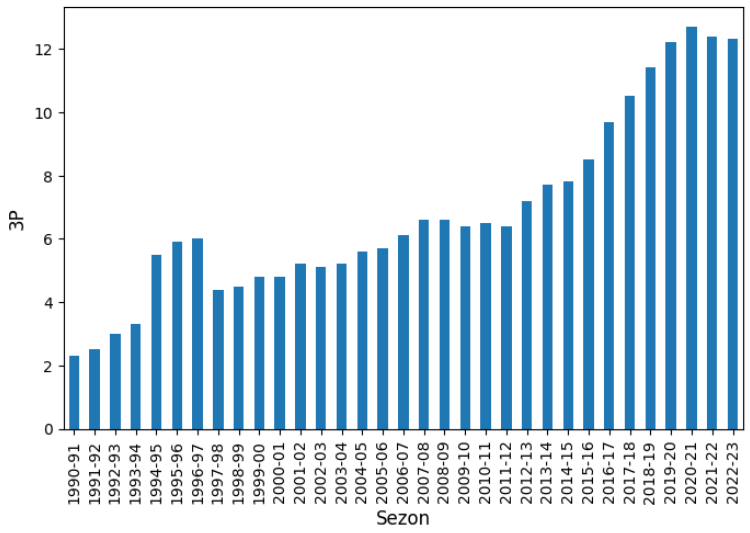
1. Liczba trafionych rzutów:



Rysunek 1: Średnia liczba trafionych rzutów na mecz na przestrzeni sezonów

Statystyka w latach 90-tych spadła o 5 i utrzymywała się w okolicy 35 trafionych rzutów na mecz. Od sezonu 2006-07 statystyka znów zaczęła się zwiększać i w obecnym sezonie znów wyniosła więcej niż 40.

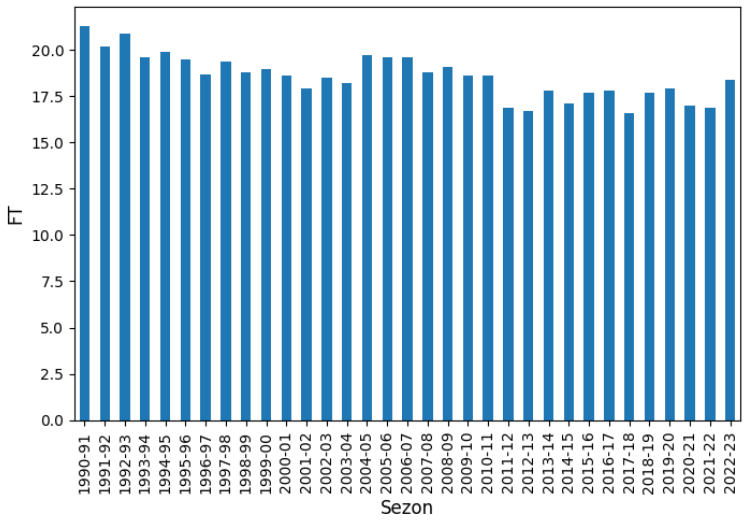
2. Liczba trafionych rzutów za 3 punkty:



Rysunek 2: Średnia liczba trafionych rzutów za 3 punkty na mecz na przestrzeni sezonów

Od sezonu 1990-91 wzrosła ona sześciokrotnie osiągając w ostatnich 4 sezonach okolice 12 na mecz. Najbardziej gwałtowny wzrost wystąpił od sezonu 2011-12 i wyniósł 6. Warto również zwrócić uwagę na sezony od 1994-95 do 1996-97, które widocznie wyróżniają się w porównaniu z innymi sezonami w latach 90-tych, mając prawie dwukrotnie większą średnią punktów. Było to skutkiem tymczasowgo skróceniem, w tych sezonach, odległości linii rzutu za 3 punkty.

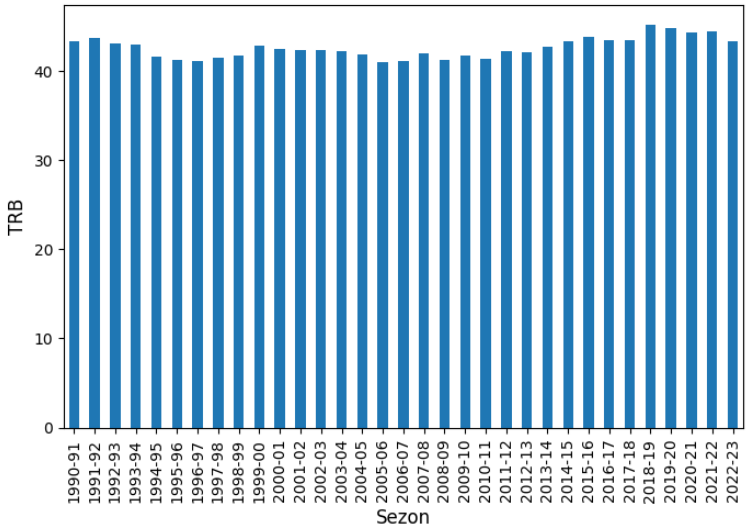
3. Liczba trafionych rzutów osobistych:



Rysunek 3: Średnia liczba trafionych osobistych na mecz na przestrzeni sezonów

Średnia liczba trafianych rzutów osobistych spadła z około 21 w sezonie 1990-91 do 17.5 w sezonie 2003-04, następnie w sezonach od 2004-05 do 2010-11 ponownie wynosiła około 19. Od sezonu 2011-12 wynosiła najmniej, bo utrzymywała się w okolicy 17.5 na mecz.

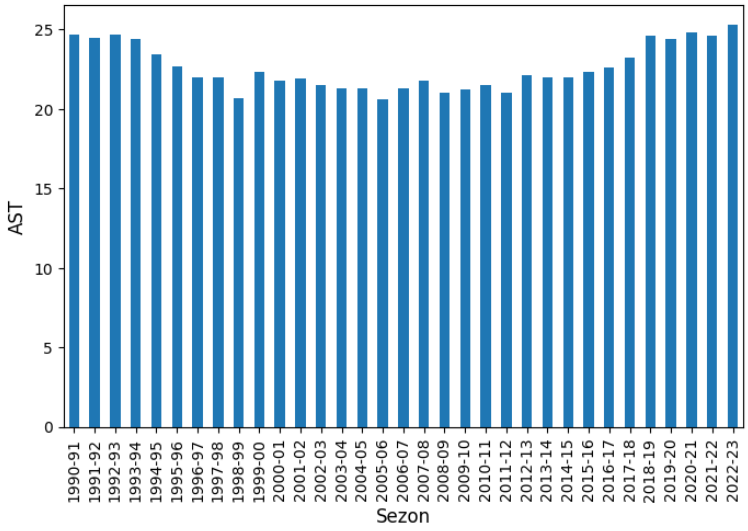
4. Liczba zbiórek:



Rysunek 4: Średnia liczba zbiórek na mecz na przestrzeni sezonów

Liczba zbiórek zanotowała nieznaczny spadek o 2 na początku lat 90-tych, finalnie wracając w sezonie 2013-14 do początkowego poziomu 43 zbiórek na mecz.

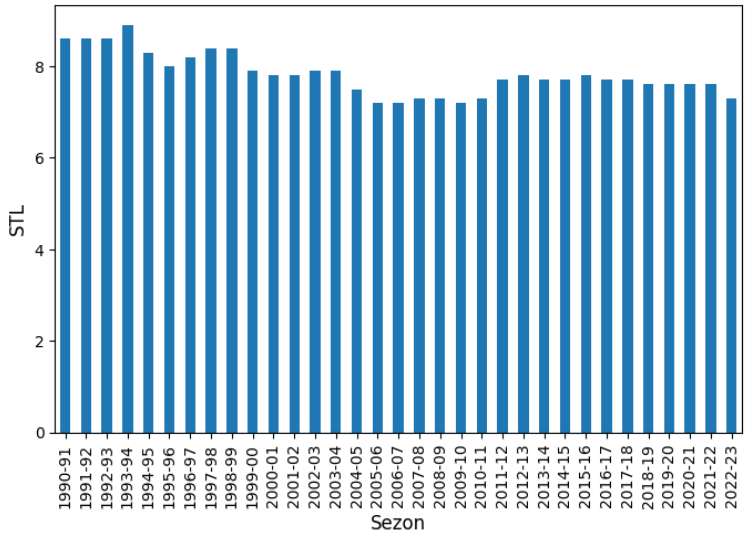
5. Liczba asyst:



Rysunek 5: Średnia liczba asyst na mecz na przestrzeni sezonów

Liczba asyst zanotowała w latach 90-tych znaczny spadek, bo aż o 5, następnie utrzymywała się w okolicach 20, aby ponownie od sezonu 2012-13 wzrosnąć do 25, czyli nieco powyżej liczby z sezonu 1990-91.

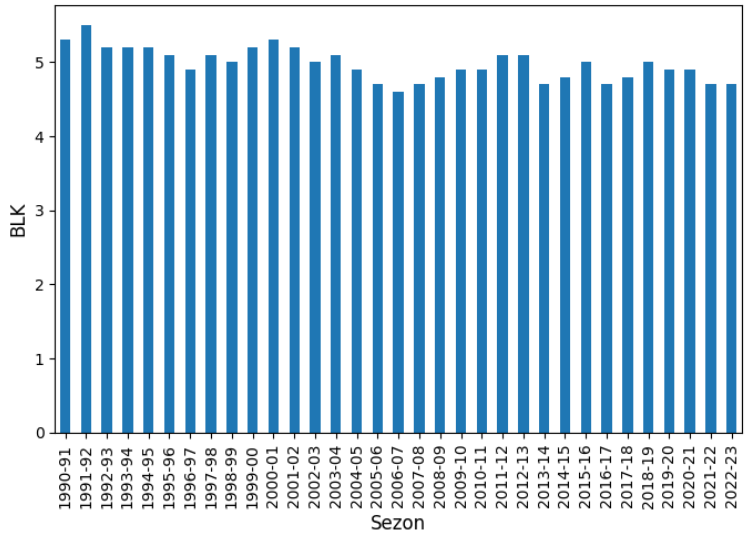
6. Liczba odbiorów:



Rysunek 6: Średnia liczba odbiorów na mecz na przestrzeni sezonów

Liczba odbiorów sukcesywnie zmiejszała się od połowy lat 2000-nych z 8.5 do 7.5. Od sezonu 2011-12 liczba ta minimalnie wzrosła do 8 i utrzymywała się na podobnym poziomie.

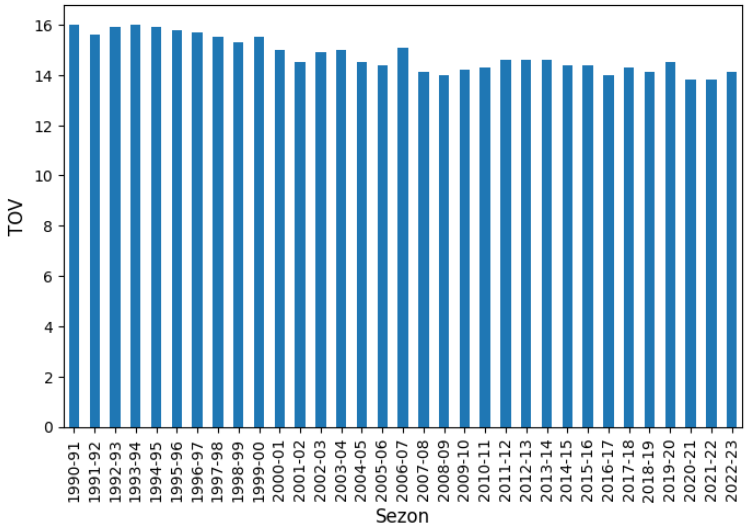
7. Liczba bloków:



Rysunek 7: Średnia liczba bloków na mecz na przestrzeni sezonów

Liczba bloków nie zanotowała znaczących zmian, notując wzrosty i spadki na przestrzeni analizowanych seoznów, ostatecznie zmniejszając się o 0.5 w porównaniu do sezonów na początku lat 90-tych.

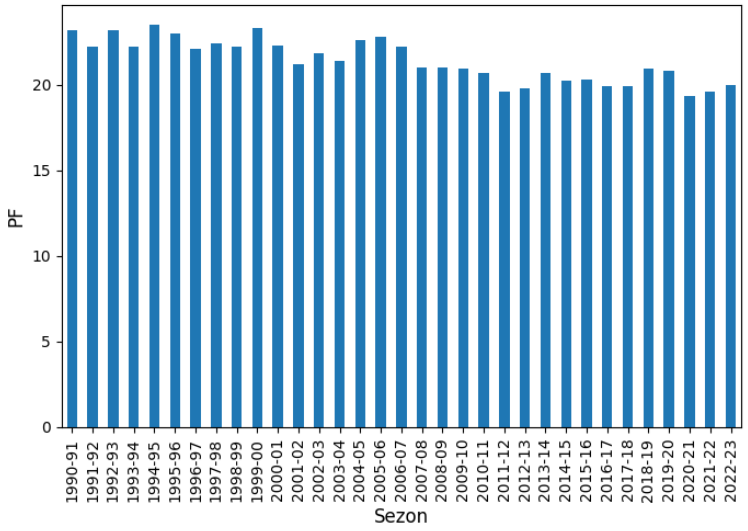
8. Liczba strat:



Rysunek 8: Średnia liczba strat na mecz na przestrzeni sezonów

Liczba strat konsekwentnie zmniejszała się z seoznu na sezon, finalnie notując spadek o 2 w porównaniu do wartości początkowej.

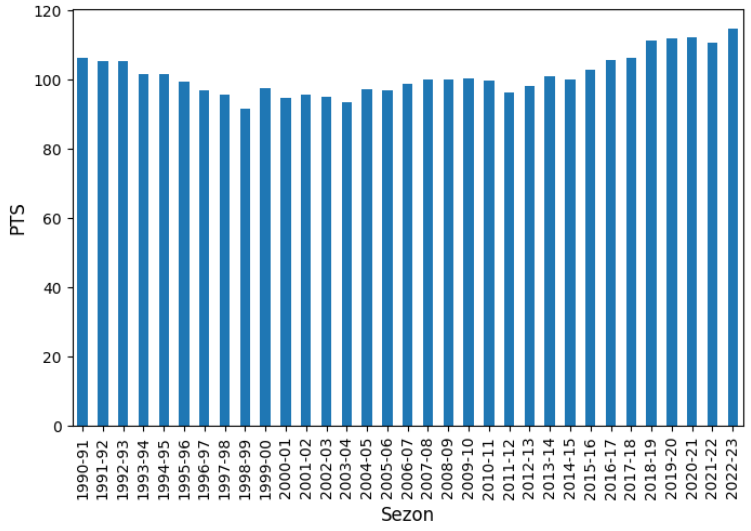
9. Liczba popełnionych fauli:



Rysunek 9: Średnia liczba popełnionych fauli na mecz na przestrzeni sezonów

Statystyka popełnionych fauli wzrastała i zmniejszała się z sezonu na sezon, finalnie notując spadek wynoszący około 3 na mecz.

10. Liczba punktów:



Rysunek 10: Średnia liczba punktów na mecz na przestrzeni sezonów

Liczba punktów spadała gwałtownie z każdym sezonem od 1990-91 do 1998-99, zmniejszając się, aż o 15. Potem jednak trend się odwrócił i trwa aż do dziś, skutkując dotychczasowym wzrostem z 91.5 na 114.7.

11. Podsumowanie

Wykresy przygotowane zostały korzystając z biblioteki matplotlib.

Warto zauważyć, że statystyki zmieniały się na 3 sposoby:

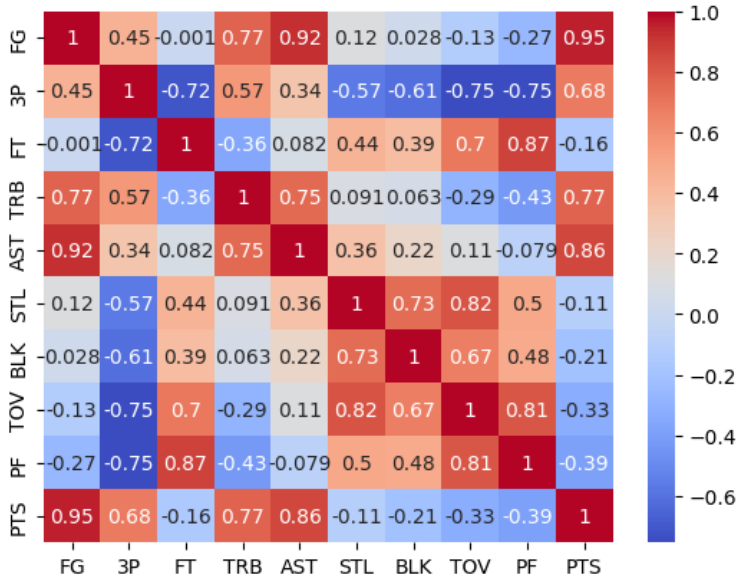
- Spadek (FT, STL, BLK, TOV, PF)
- Spadek, a następnie wzrost (FG, TRB, AST, PTS)
- Wzrost (3P)

Jedyną statystyką, która konsekwentnie zwiększała się jest liczba udanych rzutów za 3 punkty. Jest to również statystyka notująca największą różnicę w porównaniu do sezonu 1990-91, ponieważ jej aktualna wartość jest prawie 6-krotnie większa. Najwięcej statystyk zaliczyło spadek (5). Jednak każda ze zmniejszających się statystyk zmniejszyła się tylko nieznacznie. Cechą wspólną statystyk notujących spadek, a następnie wzrost jest fakt, że spadek występował w każdym przypadku na przestrzeni lat 90-tych, następnie utrzymując się na podobnym poziomie do późnych lat 2000-nych lub początku lat 2010-tych, aby ponownie wzrosnąć do wartości osiąganych na początku lat 90-tych, a nawet je przekraczając.

4.2 Zależności między statystykami

Mapa korelacji stworzona została, korzystając z biblioteki seaborn.

Jako wartości mocno skorelowane zostały uznane wszystkie związki, posiadające współczynnik korelacji wynoszący powyżej 0.65 lub poniżej -0.65.



Rysunek 11: Mapa korelacji średnich statystyk

Powyższa mapa korelacji wykazuje silny wpływ wielu statystyk na siebie. Najwięcej mocnych zależności, wykazuje średnia liczba punktów na mecz oraz liczba udanych rzutów za 3 punkty (4). Liczba punktów jest statystyką rozstrzygająca wynik meczu, co powoduje, że wpływa na nią największa ilość innych statystyk. Można spostrzec, że poza statystykami ewidentnie ofensywnymi, duży wpływ ma na nią również liczba zbiórek, czyli statystyka, bardziej defensywna niż ofensywna. Duża ilość silnych korelacji liczby rzutów za 3 punkty, najprawdopodobniej związana jest z faktem, że jest to statystyka, która zmieniała się najbardziej. Była to również jedyna statystyka posiadająca więcej niż 1 mocną korelację ujemną, co najprawdopodobniej spowodowane jest jej konsekwentnym wzrostem, przy spadku aż 5 statystyk. Najmocniej korelują ze sobą dwie grupy statystyk: FG, AST, TRB, PTS oraz STL, TOV, BLK.

4.3 Analiza zależności między statystykami

Każdy z dopasowanych trendów został obliczony korzystając z metody najmniejszych kwadratów i wykorzystując algorytm analityczny:

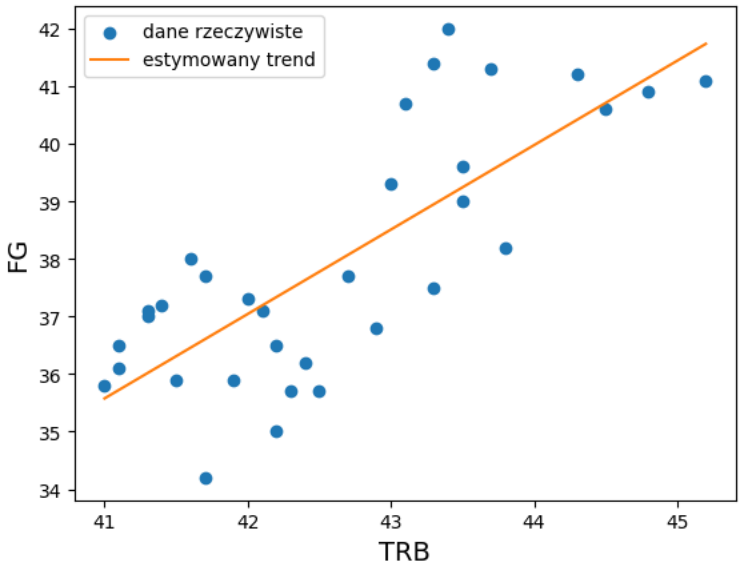
$$\hat{\theta} = (XX^T)^{-1}XY^T \tag{1}$$

Przy obliczeniach i wizualizacji wykresów wykorzystano biblioteki pandas, numpy oraz matplotlib.

1. Wpływ liczby zbiórek na liczbę udanych rzutów:

   Trend:

$$y = 1.466x - 24.542 \tag{2}$$



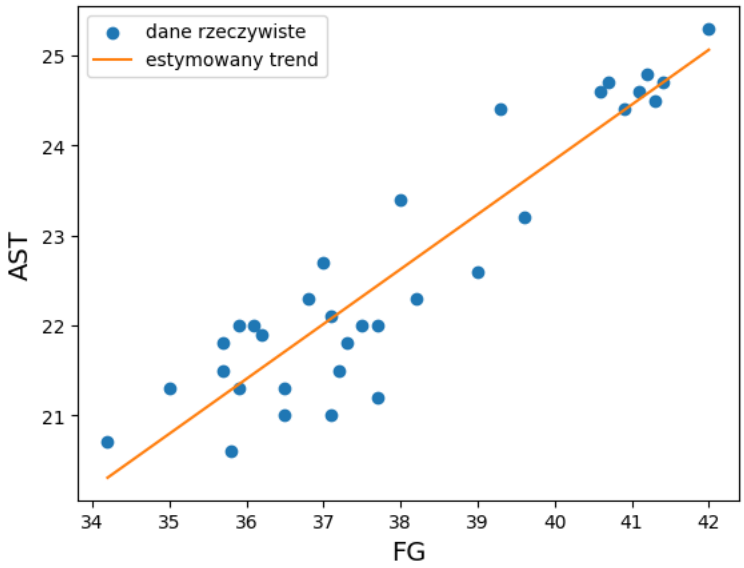
Rysunek 12: Zależność liczby udanych rzutów od liczby zbiórek

Liczba zbiórek liniowo wpływa na liczbę udanych rzutów. Im więcej piłek uda się zebrać drużynie, tym więcej może oddać rzutów, co prowadzi do większej ilości rzutów udanych.

2. Wpływ liczby udanych rzutów na liczbę asyst:

   Trend:

$$y = 0.61x - 0.56 \tag{3}$$



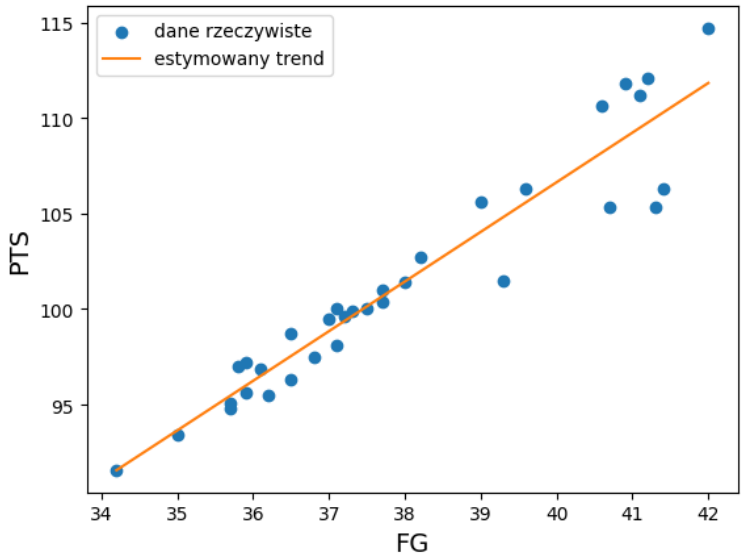
Rysunek 13: Zależność liczby asyst od liczby udanych rzutów

Większa liczba udanych rzutów bezpośrednio powoduje, że większa liczba rzutów po podaniach, również jest udana, co prowadzi do większej liczby asyst.

3. Wpływ liczby udanych rzutów na liczbę punktów:

   Trend:

$$y = 2.597x + 2.761 \tag{4}$$



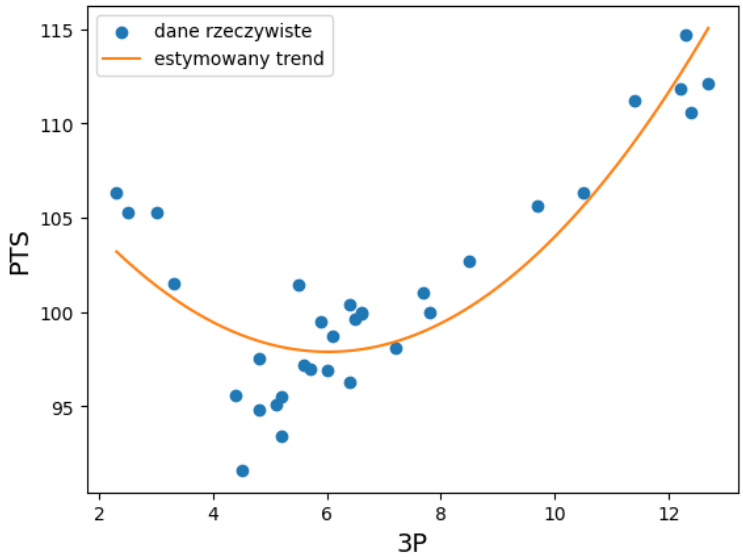
Rysunek 14: Zależność liczby punktów od liczby udanych rzutów

Jest to najmocniejsza korelacja statystyk. Liczba udanych rzutów bezpośrednio wpływa na liczbę punktów, gdyż liczba punktów jest na jej podstawie obliczana.

4. Wpływ liczby udanych rzutów za 3 punkty na liczbę punktów:

   Trend:

$$y = 0.384x^2 - 4.628x + 111.803 \tag{5}$$



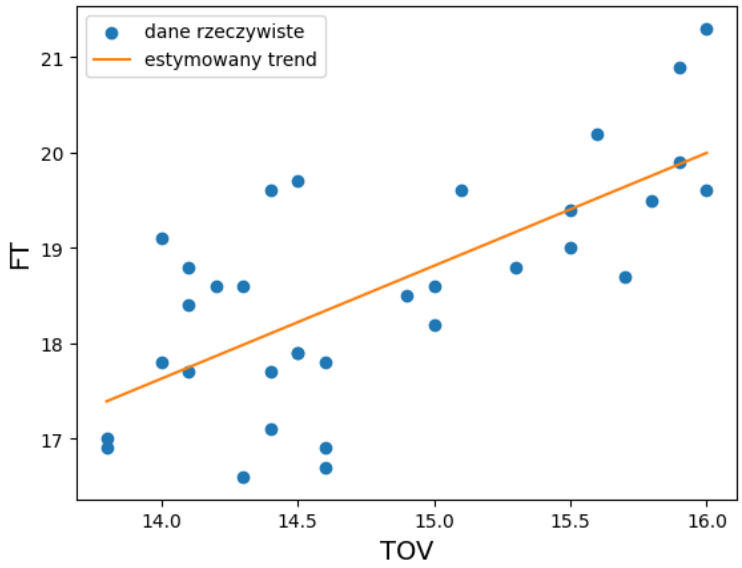
Rysunek 15: Zależność liczby punktów od liczby udanych rzutów za 3 punkty

Najlepiej dopasowanym trendem do tej korelacji jest trend kwadratowy. Liczba punktów maleje wraz ze wzrostem liczby udanych rzutów za 3 punkty na przedziale od 0 do 5.5, natomiast od tego momentu już tylko wzrasta. Oznacza to, że gdy gracze oddają mało udanych rzutów za 3 punkty, zdobywają lepszy wynik, rzucając w inny sposób, natomiast gdy liczba ich udanych rzutów za 3 punkty zaczyna przekraczać 8, znacząco zwiększa to ich sumaryczna liczba punktów na mecz. Przewidywana liczba punktów na mecz, przy braku udanych rzutów za 3 punkty (112), przekraczana jest przy zdobyciu 10 "trójek". Najgorsza liczba punktów osiągana była przy średniej wynoszącej 6 udanych rzutów za 3 punkty.

5. Wpływ liczby strat na liczbę udanych rzutów osobistych:

   Trend:

$$y = 1.184x - 1.05 \tag{6}$$

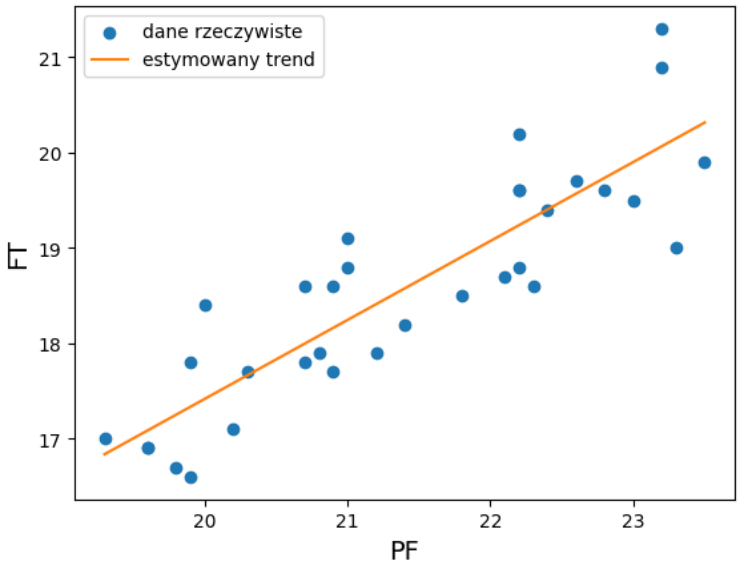


Rysunek 16: Zależność liczby udanych rzutów osobistych od liczby strat

Liczba udanych rzutów osobistych idzie liniowo w parze ze stratami. Straty często wynikają z prób zakłóceń przez drużynę przeciwną rzutu, co w wielu przypadkach skutkuje rzutami osobistymi.

6. Wpływ liczby popełnionych fauli na liczbę udanych rzutów osobistych:
   Trend:

$$y = 0.828x + 0.856 \tag{7}$$



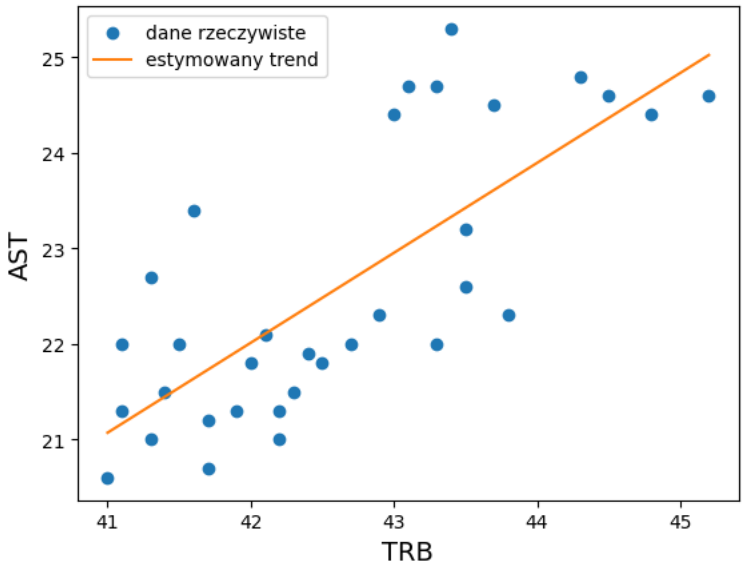
Rysunek 17: Zależność liczby udanych rzutów osobistych od liczby popełnionych fauli

Większa tendencja do popełniania fauli, skutkuje większą liczbą udanych rzutów osobistych. Rzuty osobiste bezpośrednio przyznawane są po nieprzepisowej próbie zakłócenia rzutu, co jest faulem. Obie statystyki naturalnie z siebie wynikają.

7. Wpływ liczby zbiórek na liczbę asyst:

   Trend:

$$y = 0.941x - 17.516 \tag{8}$$



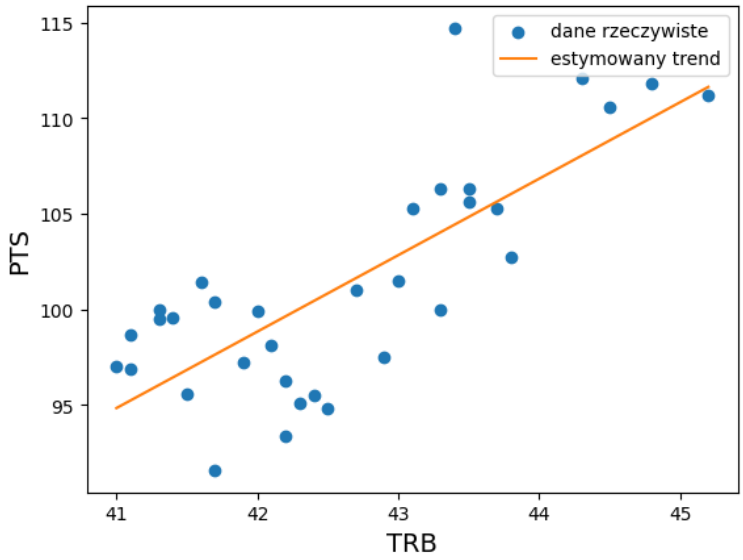
Rysunek 18: Zależność liczby asyst od liczby zbiórek

Większa liczba zbiórek prowadzi do większej ilości asyst, co powoduje występowanie trendu liniowego. Zbiórki bezpośrednio wpływają na możliwość dokonania podań, prowadzących do zdobycia punktów.

8. Wpływ liczby zbiórek na liczbę punktów:

   Trend:

$$y = 3.995x - 68.947 \tag{9}$$



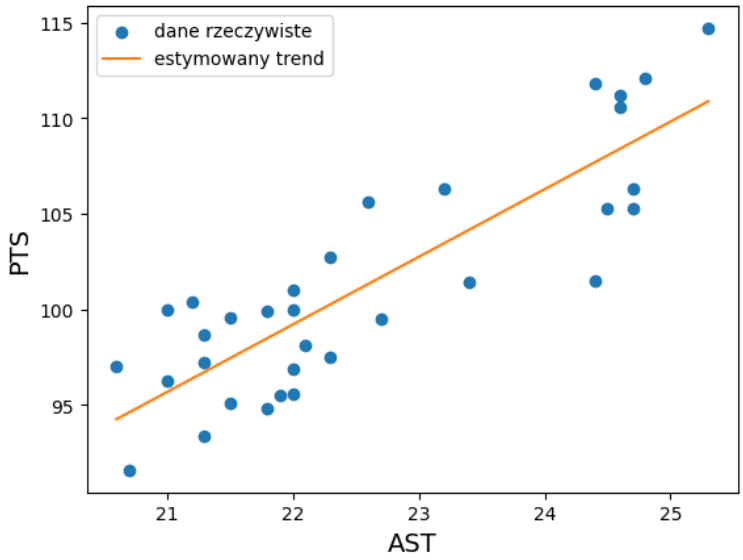
Rysunek 19: Zależność liczby punktów od liczby zbiórek

Większa liczba zbiórek, prowadzi liniowo, do większej liczby punktów. Podobnie jak w przypadku większej liczby udanych rzutów, zbiórki prowadzą do większej ilości okazji rzutowych, co przekłada się na liczbę punktów.

9. Wpływ liczby asyst na liczbę punktów:

   Trend:

$$y = 3.534x + 21.469 \tag{10}$$



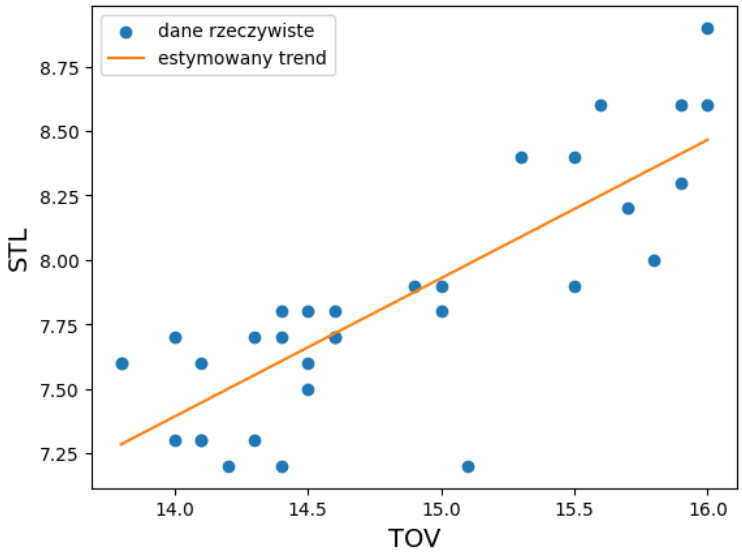
Rysunek 20: Zależność liczby punktów od liczby asyst

Asysty zaliczane są jedynie w przypadku, gdy zawodnik po podaniu zdobędzie punkt. Jest to więc naturalna zależność liniowa.

10. Wpływ liczby strat na liczbę odbiorów:

    Trend:

$$y = 0.537x - 0.128 \tag{11}$$



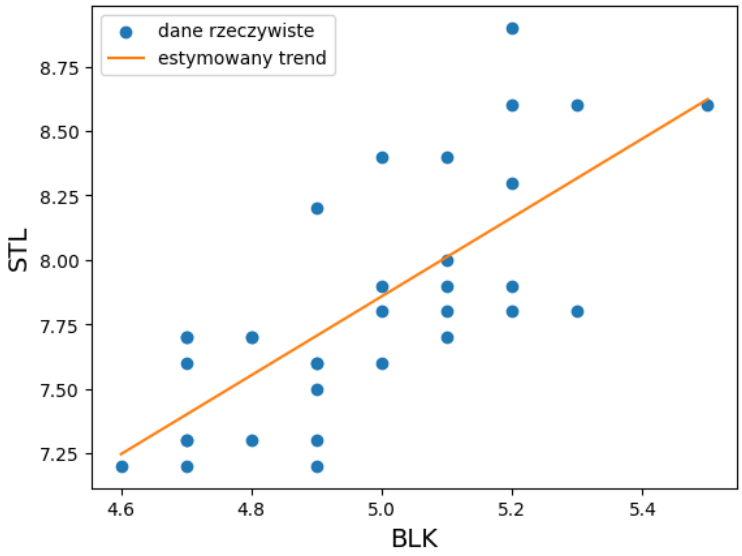
Rysunek 21: Zależność liczby odbiorów od liczby strat

Liczba strat obliczana jest na podstawie między innymi liczby odebranych piłek, przez drużynę przeciwną. W związku z tym, wymienione statystyki, są zależne od siebie liniowo.

11. Wpływ liczby bloków na liczbę odbiorów:

    Trend:

$$y = 1.53x + 0.209 \tag{12}$$



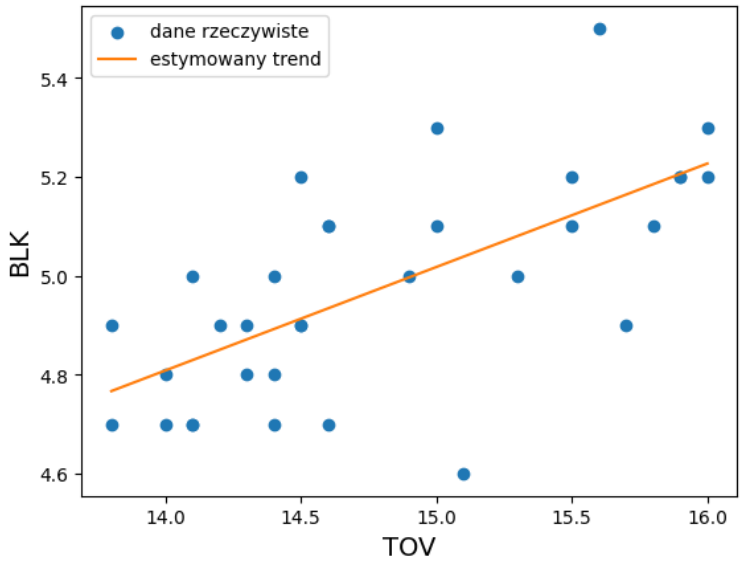
Rysunek 22: Zależność liczby odbiorów od liczby bloków

Liniowy wpływ wzrostu liczby bloków, na wzrost liczby odbiorów ukazuje, że statystyki defensywne, nawet pomimo braku bezpośredniego wpływu na siebie, są ze sobą mocno skorelowane.

12. Wpływ liczby strat na liczbę bloków:

    Trend:

$$y = 0.209x + 1.881 \tag{13}$$



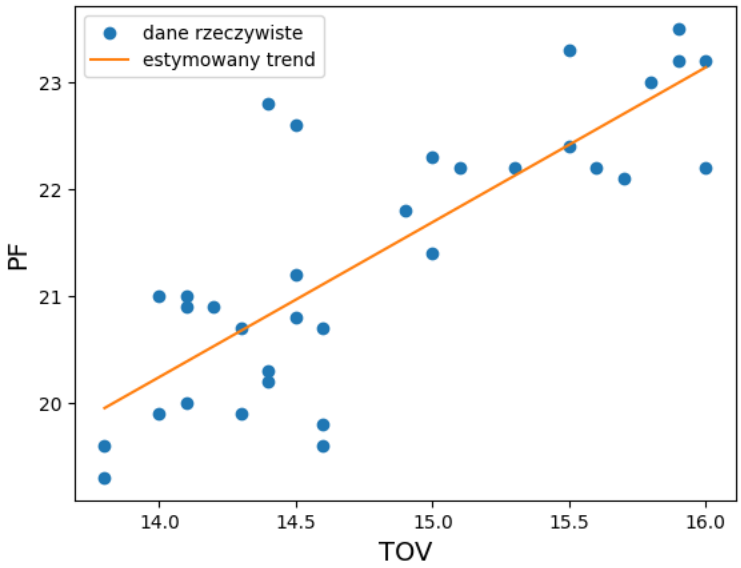
Rysunek 23: Zależność liczby bloków od liczby strat

Zwiększająca się liczba strat, daje drużynie przeciwnej więcej okazji do kontrataków, co prowadzi do zwiększenia się liczby bloków, które osiągnięte mogą zostać jedynie w defensywie.

13. Wpływ liczby strat na liczbę popełnionych fauli:

    Trend:

$$y = 1.449x - 0.043 \tag{14}$$

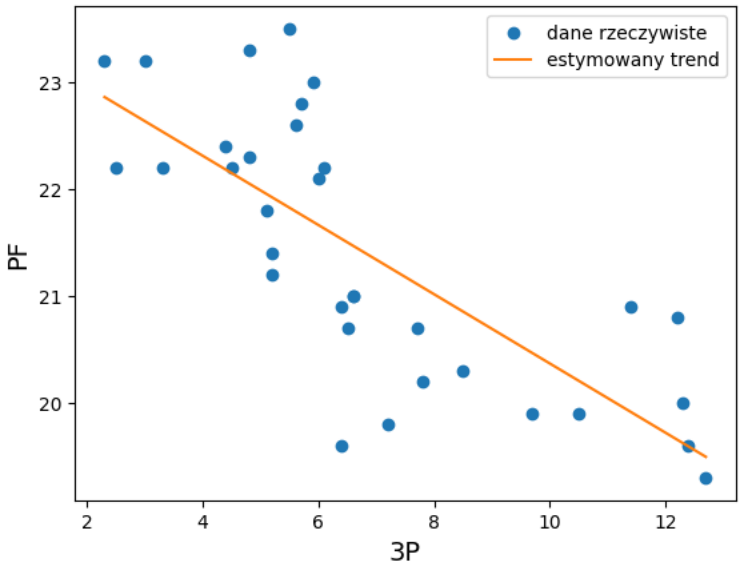


Rysunek 24: Zależność liczby popełnionych fauli od liczby strat

Wraz ze wzrostem liczby strat liniowo wzrasta liczba popełnionych fauli. Gracze mają tendencję do częstszego przekraczania przepisów, przy większej ilości prób wywołania straty u drużyny przeciwnej.

14. Wpływ liczby udanych rzutów za 3 punkty na liczbę popełnionych fauli:
    Trend:

$$y = -0.324x + 23.604 \tag{15}$$



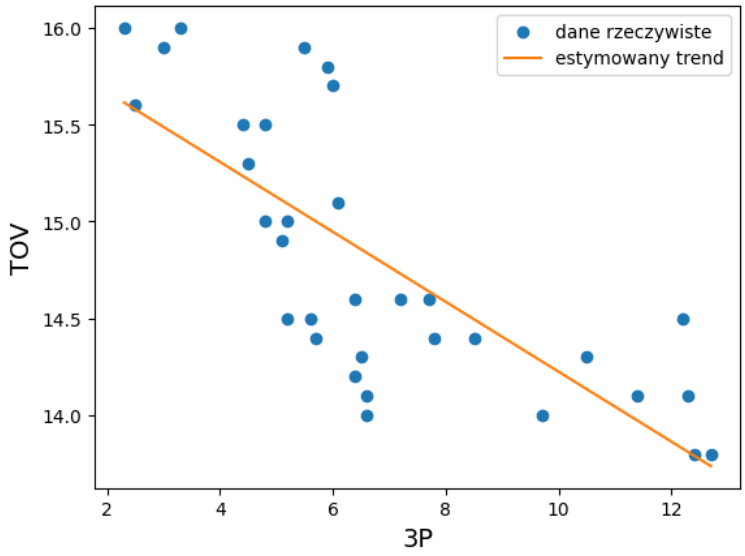
Rysunek 25: Zależność liczby popełnionych fauli od liczby udanych rzutów za 3 punkty

Wraz ze wzrostem liczby udanych rzutów za 3 punkty, liniowo maleje liczba popełnianych fauli. Jest to bezpośrednim wynikiem mniejszej potrzeby kontaktowej gry blisko kosza.

15. Wpływ liczby udanych rzutów za 3 punkty na liczbę strat:

    Trend:

$$y = -0.18x + 16.029 \tag{16}$$



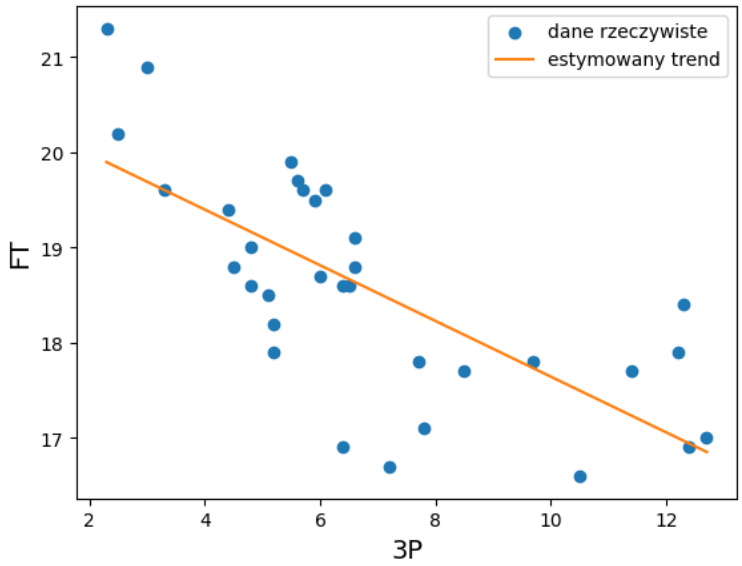
Rysunek 26: Zależność liczby strat od liczby udanych rzutów za 3 punkty

Większa liczba udanych rzutów za 3 punkty prowadzi do mniejszej liczby strat. Wynika to najprawdopodobniej z mniejszej potrzeby dryblingu w kierunku kosza, podczas którego łatwiej stracić piłkę niż przy rzucie z większej odległości. Szczególnie znaczące może się to okazać przy akcjach ofensywnych, gdzie rzut oddawany jest tuż przed końcem czasu na zegarze 24 sekund.

16. Wpływ liczby udanych rzutów za 3 punkty na liczbę udanych rzutów osobistych:

    Trend:

$$y = -0.293x + 20.569 \tag{17}$$



Rysunek 27: Zależność liczby udanych rzutów osobistych od liczby udanych rzutów za 3 punkty

    Wraz ze wzrostem liczby udanych rzutów za 3 punkty, liniowo maleje liczba udanych rzutów osobistych. Rzuty osobiste wynikają bezpośrednio z nieprzepisowego zatrzymania zawodnika podczas rzutu, co zdarza się znacznie częściej przy rzutach za 2, niż za 3 punkty.

17. Podsumowanie:

    Stopień dopasowanego trendu wybierany był na podstawie obserwacji. Dosyć istotnym problemem, napotkanym przy próbach estymacji trendu było to, że danych było stosunkowo mało oraz znajdowały się na one naprawdę małym wycinku, co w większości prowadziło do zauważalnie sporych przekłamań w wyrazie wolny wielomianów lub powodowało gwałtowne dążenie funkcji do nieskończoności.

    Prawie wszystkie silne korelacje okazywały się liniowe. Spora liczba statystyk bezpośrednio z siebie wynika, jak np. PTS i FG. Silne korelacje dodatnie wykazywały ze sobą również statystyki, które na przestrzeni lat zmieniły się w ten sam sposób np. BLK i STL, lub w przypadku silnych korelacji ujemnych w sposób przeciwny. Jedynym wyjątkiem od tej reguły

jest kwadratowa zależność liczby punktów od liczby udanych rzutów za 3 punkty. W tym przypadku liczba udanych rzutów za 3, z każdym sezonem, zaliczała wzrost, a liczba punktów wpierw malała, a następnie rosła.

5 Modelowanie

5.1 Przygotowanie danych

Zbiór "Per Game Stats"podzielony został na dane treningowe i dane testowe, w sposób losowy, w stosunku 80/20. Następnie kolumny tego zbioru zostały podzielone na cechy numeryczne (FG, 3P, FT, TRB, AST, STL, BLK, TOV, PF, PTS) i kategoryczne (Season). W przypadku braku występowania któreś z danych w wierszu, jeżeli była ona wartością numeryczną, wstawiana była wartość średnia danej kolumny. W przypadku braku wartości kategorycznej wstawiana była najczęściej występująca kategoria. Do przekształcania cech numerycznych użyty został również scaler, który pomógł wyrównać skale tych cech. Z kolei przy przekształcaniu cech kategoryczny zastosowano encoder, który utworzył dodatkowe kolumny z każdą unikalną wartością cechy kategorycznej. Modele otrzymały dane treningowe z 421 wierszami należącymi do klasy dostającej się do fazy play-off i 351 do klasy nie grającej w daleszej części sezonu. W danych testowych znalazło się 111 wierszy ze statystykami pozwalającymi na awans do fazy play-off i 83 wierszy ze statystykami nie dającymi gry w fazie play-off. Większa ilość wierszy klasy dostającej się do fazy play-off, wynika z faktu, iż w każdym sezonie do 2020-21, do fazy tej awansowało 16 drużyn, a analizowane sezony posiadały od 27 do 30 drużyn. W sezonach 2019-20, 2020-21, 2021-22 oraz 2022-23 wprowadzono turniej play-in, do którego awans traktowany jest w badaniu, jak awans do play-off, co zwiększyło liczbę awansujących drużyn z 16 do 18 w sezonie 2019-20 i do 20 w pozostałych sezonach. Do stworzenia modeli wykorzystano bibliotekę sklearn.

5.2 Dobór modelu

1. **Model KNN**

   (a) Przygotowanie modelu

   Do utworzenia modelu zastosowana została metryka cosinusowa, a ilość sąsiadów została ustawiona na 14.

   (b) Test modelu

	YES (actual)	**NO (actual)**
YES (predicted)	84	24
NO (predicted)	27	59

	precision	**recall**	**f1-score**
NO	0.69	0.71	0.70
YES	0.78	0.76	0.77
macro avg	0.73	0.73	0.73
weighted avg	0.74	0.74	0.74

Dokładność modelu: 0.74

(c) Analiza modelu

Łącznie 143 wiersze zostały przewidzianych prawidłowo i 51 błędnie. Model dał dosyć wysoką dokładność, ponieważ jest w stanie przewidzieć bliko 3/4 wierszy ze statystykami. Lepiej poradził sobie z przewidywaniem statystyk pozwalającymi na grę w fazie play-off, co wynika z precyzji predykcji klasy "YES"większej o 0.09. Jednak różnica nie była już tak znacząca w przypadku recall, gdzie zmniejszyła się ona do 0.05.

2. **Model SVC**

(a) Przygotowanie modelu

Do utworzenia modelu zastosowane zostało jądro rbf.

(b) Test modelu

	YES (actual)	**NO (actual)**
YES (predicted)	95	31
NO (predicted)	16	52

	precision	**recall**	**f1-score**
NO	0.76	0.63	0.69
YES	0.75	0.86	0.80
macro avg	0.76	0.74	0.75
weighted avg	0.76	0.76	0.75

Dokładność modelu: 0.76

(c) Analiza modelu

Łącznie 147 wierszy zostało przewidzianych prawidłowo i 47 błędnie. Model dał nam wysoką dokładność, przewidując ponad 3/4 wierszy prawidłowo. Sklasyfikował zdecydowanie większą ilość wierszy jako posiadające odpowiednie statystyki, aby dostać się do fazy play-off. Stosunek przewidzianych wystąpień klasy "YES"do "NO"wyniósł 126/68, przy rzeczywistym 111/83. Poskutkowało to wysokim recallem klasy, dostającej się do fazy play-off (0.83) i ewidentnie niższym klasy kończącej rozgrywki (0.63). Precyzja przewidywania obu cech różniła się nieznacznie.

3. **Model DecisionTree**

   (a) Przygotowanie modelu
   Zastosowano domyślne parametry modelu.

   (b) Test modelu

	YES (actual)	**NO (actual)**
YES (predicted)	79	27
NO (predicted)	32	56

	precision	**recall**	**f1-score**
NO	0.65	0.66	0.65
YES	0.74	0.73	0.74
macro avg	0.70	0.70	0.70
weighted avg	0.70	0.70	0.70

   Dokładność modelu: 0.70

   (c) Ważność atrybutów

Cecha	**Ważność**
TOV	0.1659
FT	0.1312
AST	0.1106
TRB	0.0998
PF	0.0992
STL	0.091
FG	0.0795
BLK	0.076
3P	0.0478
Season 2006-07	0.0158
Season 2013-14	0.0141
PTS	0.0114
Season 1998-99	0.011
Season 2015-16	0.0098
Season 1995-96	0.0089
Season 2022-23	0.0049
Season 2017-18	0.0048
Season 2000-01	0.0044
Season 1999-00	0.0041
Season 2018-19	0.0038
Season 2014-15	0.0035
Season 1994-95	0.0014
Season 2003-04	0.0009

(d) Analiza modelu Łącznie 147 wierszy zostało przewidzianych prawidłowo i 47 błędnie. Model dał nam dokładność wynoszącą 7/10. Ponownie, model był znacznie bardziej precyzyjny w przypadku przewidywania zespołów, które dostaną się do fazy play-off. Zarówno precyzja i recall różniły się o około 0.09, z przewagą klasy "YES". Najważniejszą cechą, przy przewidywaniu, okazała się liczba strat, liczba udanych rzutów osobistych oraz liczba asyst, a najmniej istotne okazały się liczba rzutów za 3 punkty, liczba punktów oraz informacja o sezonie. Część sezonów zostały określone przez model jako nieistotne, jednak nie została odnaleziona żadna zależność określająca, dlaczego model podjął taką decyzję.

4. Podsumowanie

   Wybory parametrów, przedstawianych modeli, dokonywane były poprzez przeprowadzanie testów na ich różnych kombinacjach i wyborze jednej prowadzącej do największej dokładności.

   Różnice między dokładnością poszczególnych modeli nie okazały się duże. Najdokładniejszy okazał się SVC z dokładnością 0.76, a najmniej dokładny DecisionTree z dokładnością 0.7. Każdy z modeli radził sobie znacznie lepiej z przewidywaniem klasy cech, które dostaną się do fazy play-off, co najprawdopodobniej wynika z minimalnie większej ilości wierszy tej klasy w danych treningowych. Bazując na wynikach dokładności zastosowanych modeli, można uznać, że na podstawie wybranych w tym eksperymencie statystyk oraz wiadomości o sezonie z którego pochodzą, jesteśmy w stanie stworzyć model, który z zadowalającą dokładnością przewiduje, czy drużyna dostałaby się do fazy play-off w podanym sezonie. Wartym spostrzerzenia faktem jest mała istotność cechy związanej z konkretnym sezonem w modelu DecisionTree, a duża związana ze statystykami numerycznymi, zarówno ofensywnymi, jak i defensywnymi.

5.3 Eksperymenty na modelu

1. Opis eksperymentu

   Korzystając z modelu DecisionTree, przeprowadzony został eksperyment, pozwalający ocenić, czy aktualne drużyny NBA prezentują poziom wyższy, czy niższy, niż drużyny z sezonów w poprzednich dekadach. Wybór modelu wynika z największej dokładności, jego predykcji, dotyczącej drużyn z obecnego sezonu w obecnym sezonie, z pośród wszystkich modeli (pozostałe modele przewidywały 24 drużyny z tego sezony jako drużyny, które dostały się do fazy play-off, czyli aż o 4 za dużo) oraz z powodu jawnie znanych i przeanalizowanych wcześniej ważności poszczególnych cech wykorzystywanych przez ten model. W pierwszej fazie eksperymentu ze zbioru danych "Per Game Stats", wycięte zostały wiersze dotyczące drużyn jedynie z tego sezonu. Następnie zmieniając ich wartość w kolumnie Season, na wybrany sezon z przeszłości, policzone zostało, ile z nich, mo-

del sklasyfikował jako drużyny, które dostałyby się do fazy play-off, w podanym przeszłym sezonie. W drugim kroku, role zostały odwrócone i sprawdzone zostało jak drużyny z przeszłości, poradziłyby sobie w obecnym sezonie.

2. Eksperyment

   Wyniki drużyn z aktualnego sezonu zasadniczego w poprzednich sezonach.

Sezon	**YES**	**NO**
2022-23	20	10
2017-18	22	8
2012-13	22	8
2007-08	22	8
2002-03	22	8
1997-98	22	8
1992-93	22	8

   Wyniki drużyn z poprzednich sezonów w aktualnym sezonie zasadniczym.

Sezon	**YES**	**NO**
2022-23	20	10
2017-18	16	14
2012-13	13	17
2007-08	17	13
2002-03	16	13
1997-98	15	14
1992-93	15	12

3. Obserwacje

   Drużyny z aktualnego sezonu poradziły sobie zdecydowanie lepiej, od drużyn z przeszłości. W każdym sezonie, model przewidział, że 22 z nich dostałyby się do fazy play-off, co odpowiada jedynie 8 drużyną, którym by się to nie udało. Warto zwrócić uwagę, że w pozostałych badanych sezonach, limit drużyn awansujących do fazy play-off wynosił 16, w związku z czym, obecne drużyny przekroczyły go w każdym sezonie, aż o 6. W drugiej fazie eksperymentu, zaobserwowano, że drużyny z przeszłości miałyby spore problemy znajdując się w obecnym sezonie. Żadnej z przeszłych ekip nie udało się przebić liczby dwudziestu przechodzących dalej drużyn. Najlepszy wynik osiągnęły drużyny z sezonu 2007-08 (17), a najgorszy 2012-13 (13). Drużyny z lat 90-tych dwukrotnie osiągały przewidywania 15 drużyn dostających się do fazy play-off.

6 Wnioski

Udało się zrealizować wszystkie cele projektu.

Od początku lat 90-tych, poszczególne statystyki przechodziły różne zmiany. Porównując je bezpośrednio, zauważyć można spadek statystyk defensywnych, takich jak odbiory i bloki. Jednak wraz z nimi spadła ilość popełnianych fauli, która z pewnością wpływa na większe bezpieczeństwo zawodników, co jest niezwykle kluczowe w grze, tak kontaktowej, jak koszykówka. Drużyny osiągają również średnio o 2 straty mniej. Statystyki, które na przestrzeni analizowanych sezonów zmalały i znów wzrosły, również malały najczęściej i najgwałtowniej w latach 90-tych. Oznacza to łącznie 9 na 10 stopniowo zmniejszających się statystyk, co sugeruje, że tamto dzisięciolecie było momentem spadku poziomu NBA. Sytuacja odwraca się dopiero w latach 2010-tych, gdzie statystyki takie jak FG, TRB, AST i PTS znów zaczęły rosnąć. Zmiany wspomnianych statystyk, poza PTS i 3P, jednak nie okazały się duże, co pozwala nam stwierdzić, że mimo postępujących zmian, same fundamenty ligi pozostają niezmienne. Z całej tej analizy wyłamuję się liczba rzutów za 3 punkty, która przez cały analizowany okres, zwiększała się będąc jedyną znacznie zmieniająca sie statystyką.

Badanie korelacji między statystykami potwierdziło wiele zależności oraz zaprezentowało, że statystyki prawie w każdym przypadku wpływają na siebie liniowo, co wynika z bezpośredniego ich wpływu na siebie lub ich podobnie przebiegających zmian na przestrzeni lat. Jedyną statystyką, której wzrost powodował spadek innych statystyk były udane rzuty za 3 punkty, jednak zmiany z jej zwiększaniem się skorelowane nie były znaczne. Kwadratowa zależność liczby punktów na mecz od liczby udanych rzutów za 3 punkty oraz trwający aktualnie wzrost obu tych statystyk, z każdym sezonem, pozwala stwierdzić, że zawodnicy obecnie znacznie bardziej skupiają trening wokół tej umiejętności, co bezpośrednio wpływa na poprawę ofensywy drużyny. Biorąc pod uwagę fakt, iż aktualnie średnia liczba punktów jest prawie o 10 większa niż na początku lat 90-tych i prawie o 20 większa niż pod ich końcówkę, można to określić jako pozytywną zmianę.

Stworzone modele posiadały dokładność w okolicach 0.7 do 0.76, co uznawane jest za wynik satysfakcjonujący. Kluczowa przy analizie zmian jest tabela ważności cech w modelu DecisionTree. Brak lub mała istotność pochodzenia wiersza z poszczególnego sezonu potwierdza, że na przestrzeni sezonów 1990-91 - 2022-23 wartości cech numerycznych nie zmieniły się na tyle drastycznie, aby kluczową informacje stanowił sezon, z którego pochodzą dane. Ku zaskoczeniu, kluczowa przy przewidywaniu klasy modelu okazała się statystyka strat. Wpływ na to może mieć, fakt iż drużyny prowadzące w meczu, szczególnie w końcówce, mają tendencję by rzadziej tracić piłkę i częściej oddawać rzuty osobiste po faulach drużyny przeciwnej. Koniec, końców to procent zwycięstw w sezonie daje awans, a nie liczba punktów.

Wyniki eksperymentu prowadzą do bezpośredniego porównania formy obecnych drużyn, z tymi z przeszłości. Drużyny teraźniejsze poradziły sobie zdecydowanie lepiej w przeszłości niż drużyny przeszłe w teraźniejszości. W latach 90-tych, 22 drużyny awansowałyby do fazy play-off, czyli o dwie więcej niż w

przypadku ich gry w sezonie z którego pochodzą. Natomiast dla drużyn z początku i końca lat 90-tych przewidywane jest jedynie 15 miejsc w play-offach 2022-23. Pokazuje to, że aktualne drużyny bezpośrednio radzą sobie lepiej od drużyn z lat 90-tych.

Reasumując, wynik eksperymentu oraz analiza danych wskazuje na wzrost poziomu ligi NBA względem lat 90-tych. Co więcej, poziom ten wykazuje tendencje do ciągłego wzrostu. Statystykami mocno wyróżniającymi się, w porównaniu do sezonów sprzed 30 lat, są jedynie udane rzuty za 3 punkty oraz liczba punktów. Pomimo zmian stylu gry, 8 z 10 analizowanych statystyk nie zmieniły się znacznie. Dodatkowo są to statystyki, uznawane przez model, jako najbardziej kluczowe w predykcji awansu drużyny do fazy play-off, co tym bardziej przeczy teorii mówiącej, że aktualna koszykówka na stadionach NBA niczym nie przypomina lat 90-tych. Zmiany są nieodłączną częścią każdego sportu, a liga NBA nie jest tu wyjątkiem. Jednak w tym przypadku ewolucja nie jest gwałtowna i prowadzi do zwiększenia poziomu rozgrywek, co powinno zostać częściej zauważane i respektowane.

Literatura

[1] NBA official website. Available: `https://www.nba.com/`

[2] Basketball-Reference. Available: `https://www.basketball-reference.com/`