Analiza danych NBA

Kacper Skwarek 275992, Paweł Wojarnik 276027

2024-12-05

Spis treści

Opis danych

Cel analizy

Celem analizy jest zbadanie zależności pomiędzy miejscem zawodnika w drafcie NBA, drużyną, do której został wybrany, a jego osiągnięciami w karierze zawodowej. Analiza pozwoli lepiej zrozumieć, jakie cechy draftowanych graczy mogą być predyktorami sukcesu w NBA, a także wskazać, czy wyższe pozycje w drafcie korelują z wyższymi osiągnięciami w karierze.

Źródło danych

Dane pochodzą z publicznego zestawu zawierającego informacje o graczach draftowanych do NBA od roku 1989 do 2021, znajdującego się na stronie https://www.basketball-reference.com/draft/.
Zestaw zawiera dane zebrane z publicznie dostępnych źródeł statystycznych, takich jak strony drużyn NBA, bazy danych draftów oraz archiwa ligowe. Szczegóły na temat licencji nie zostały podane, ale dane są wykorzystywane w celach edukacyjnych i analitycznych.

Link do zestawu danych:

https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021?resource=download

Opis zmiennych

Poniżej przedstawiono kluczowe zmienne w zbiorze danych:

1. **Rok draftu (`year`)**: Rok, w którym zawodnik został wybrany w drafcie NBA (wartości od 1989 do 2021).

2. **Miejsce w drafcie (`rank`)**: Pozycja zawodnika w drafcie w danym roku (wartości od 1 do 60).

3. **Drużyna (`team`)**: Drużyna NBA, która wybrała zawodnika.

4. **Zawodnik (`player`)**: Imię i nazwisko draftowanego zawodnika.

5. **Uczelnia (`college`)**: Uczelnia, na której zawodnik grał przed przystąpieniem do draftu (jeśli dotyczy).

6. **Lata aktywności (`years_active`)**: Liczba lat spędzonych w lidze NBA (jednostka: lata, wartości od 0 do 22).

7. **Rozegrane mecze (`games`)**: Liczba rozegranych meczów w karierze zawodnika (jednostka: mecze, wartości od 0 do 1541).

8. **Średnie punkty na mecz (`points_per_game`)**: Średnia liczba punktów zdobytych na mecz przez zawodnika (wartości od 0 do 27,2).

9. **Win shares (`win_shares`)**: Win Shares to kompleksowa metryka używana do oceny ogólnego wpływu koszykarza na sukces jego drużyny. Łączy różne aspekty gry zawodnika w jedną statystykę, która szacuje liczbę zwycięstw, do których zawodnik się przyczynił. (wartości od -1,7 do 249,5).

10. **Box plus-minus (`box_plus_minus`)**: Zaawansowana miara wydajności gracza w stosunku do średniego gracza NBA (wartości od -52 do 51,1).

Tabela 1: Przykładowe dane draftowanych zawodników NBA

Rok draftu	Miejsce w drafcie	Drużyna	Zawodnik	Uczelnia	Lata aktywności	Rozegrane mecze	Średnie punkty na mecz	Win shares	Box plus-minus
1989	1	SAC	Pervis Ellison	Louisville	11	474	9.5	21.8	-0.5

Statystyki opisowe

- **Lata aktywności (`years_active`)**: Średnia: 5,5 lat, mediana: 4 lata, maksymalna wartość: 22 lata.

- **Rozegrane mecze (`games`)**: Średnia: 302 mecze, mediana: 163 mecze, maksymalna wartość: 1541 meczów.

- **Średnie punkty na mecz (`points_per_game`)**: Średnia: 7,3 punktów, mediana: 6,2 punktów, maksymalna wartość: 27,2 punktów.

- **Win shares (`win_shares`)**: Średnia: 17,9, mediana: 5,3, maksymalna wartość: 249,5.

Pytania badawcze

1. Czy miejsce w drafcie (np. top 10) koreluje z wynikami zawodnika w karierze (np. liczba rozegranych meczów, średnie punkty, win shares)?

2. Które drużyny historycznie wybierały najbardziej produktywnych zawodników?

3. Czy uczelnie, takie jak Duke czy Kentucky, są związane z lepszymi wynikami zawodników w NBA?

Wczytanie danych

Nadanie odpowiednich nazw zmiennych

Dane zostały wczytane i zmieniono nazwy zmiennych na bardziej opisowe, aby ułatwić ich interpretację. Na przykład:

- `three_point_percentage` zamiast `3_point_percentage` (prostsze odwołanie),
- `rank` zamiast `draftRank` (bardziej intuicyjne).

Pełna lista zmienionych nazw znajduje się w dokumentacji.

Typy kolumn

W celu poprawienia analizy zadbano o odpowiednie typy kolumn:

- Zmienne kategoryczne, takie jak `team` (drużyna) oraz `college` (uczelnia), zostały skonwertowane do typu `category`, co zmniejsza użycie pamięci oraz poprawia wydajność.
- Pozostałe zmienne, takie jak `games` (liczba rozegranych meczów) czy `win_shares` (miara sukcesu zawodnika), są traktowane jako zmienne numeryczne.

Czyszczenie wartości

W przypadku braków danych zastosowano następujące podejście:

1. **Zmienna `years_active` (liczba lat aktywności)**: Uzupełniono brakujące wartości wartością `0`, co odpowiada zawodnikom, którzy nie rozegrali żadnego meczu w NBA.

2. **Statystyki indywidualne** (`games`, `points`, `minutes_played` itp.): Braki uzupełniono wartością `0`, co oznacza brak aktywności w lidze.

3. **Dane o uczelniach (`college`)**: Pozostawiono brakujące wartości jako `NA`, ponieważ te dane nie wpływają na analizy statystyczne wydajności zawodników.

Obsługa braków danych

- Zmienna `college` posiada 337 brakujących wartości (ok. 17,5%), co oznacza, że nie wszystkie dane o edukacji zawodników są dostępne. Nie wpłynie to jednak na ogólną analizę.
- Dane zaawansowane, takie jak `win_shares_per_48_minutes`, mają niewielką liczbę braków (do 1,5%). Te braki zostały zignorowane, ponieważ ich skala jest niewielka w stosunku do całości zbioru danych.

Dyskusja

Braki danych w zestawie wynikają prawdopodobnie z tego, że niektórzy gracze:

1. Zostali wybrani w drafcie, ale nigdy nie grali w lidze NBA.
2. Nie mieli pełnych statystyk zebranych w systemach statystycznych NBA.
3. Nie ukończyli uczelni amerykańskich, co utrudnia przypisanie wartości dla `college`.

Po przetworzeniu danych są one gotowe do dalszej analizy.

Analiza danych

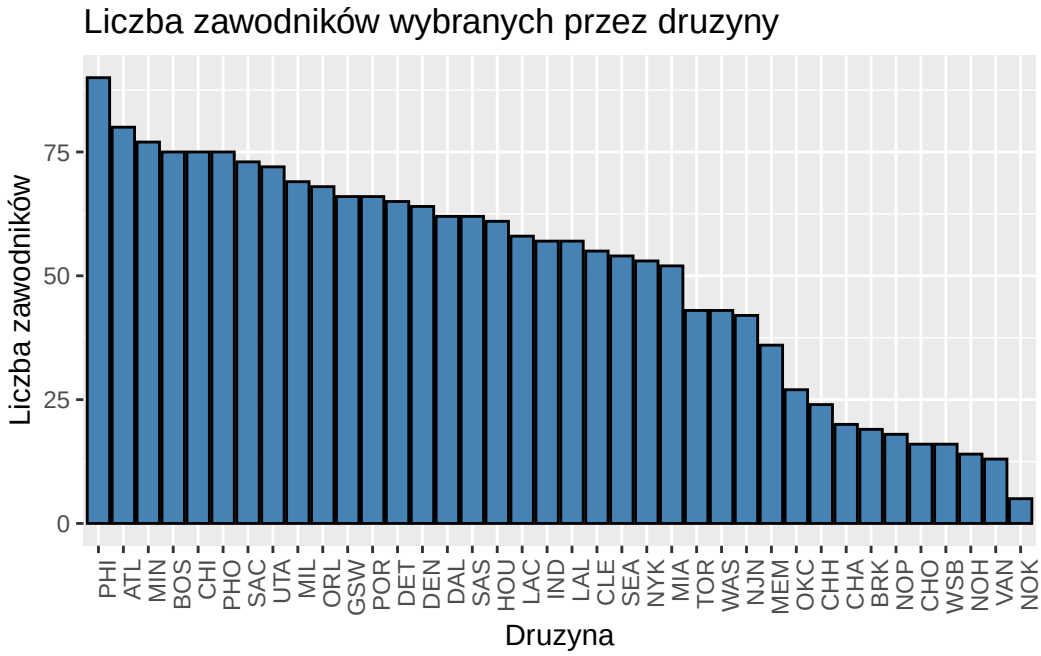
Wstęp

Celem tej sekcji jest analiza zależności między miejscem zawodnika w drafcie NBA, drużyną, do której został wybrany, a jego osiągnięciami w karierze. Skupimy się na odpowiadaniu na pytania badawcze, analizując zarówno cechy kategoryczne, jak i ciągłe.

Analiza cech kategorycznych

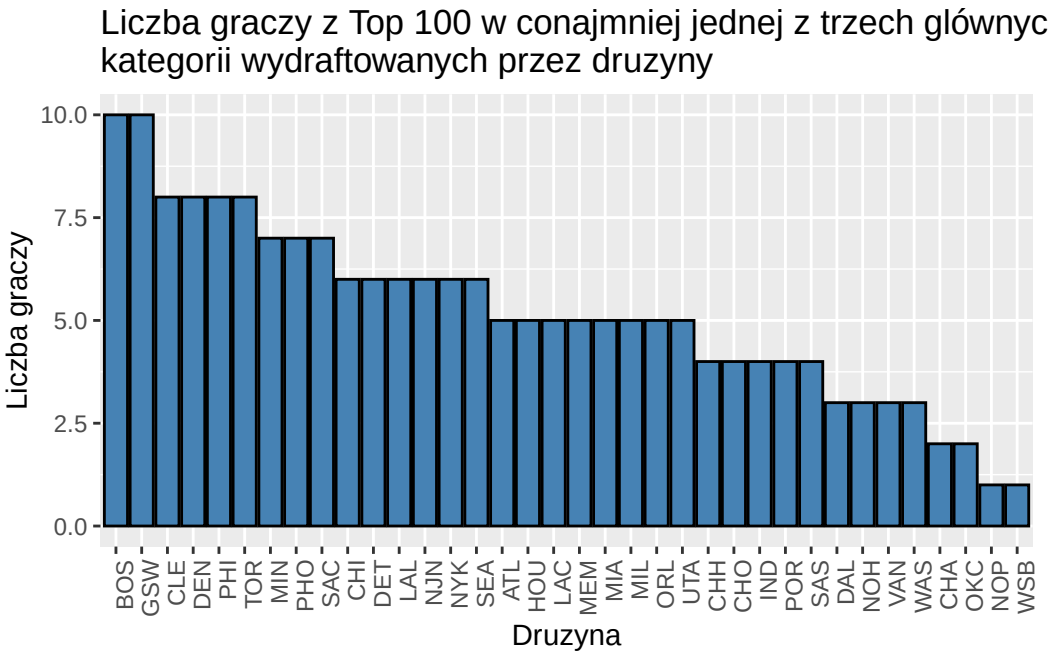
Liczba zawodników w zależności od drużyny

Zbadano, które drużyny NBA wybierały największą liczbę zawodników w analizowanym okresie.



Wnioski: Najwięcej zawodników zostało wybranych przez drużyny, takie jak PHI (Philadelphia 76ers) i ATL (Atlanta Hawks). Jest to związane z ich częstym udziałem w draftach, oraz strategi budowania drużyn od podstaw.

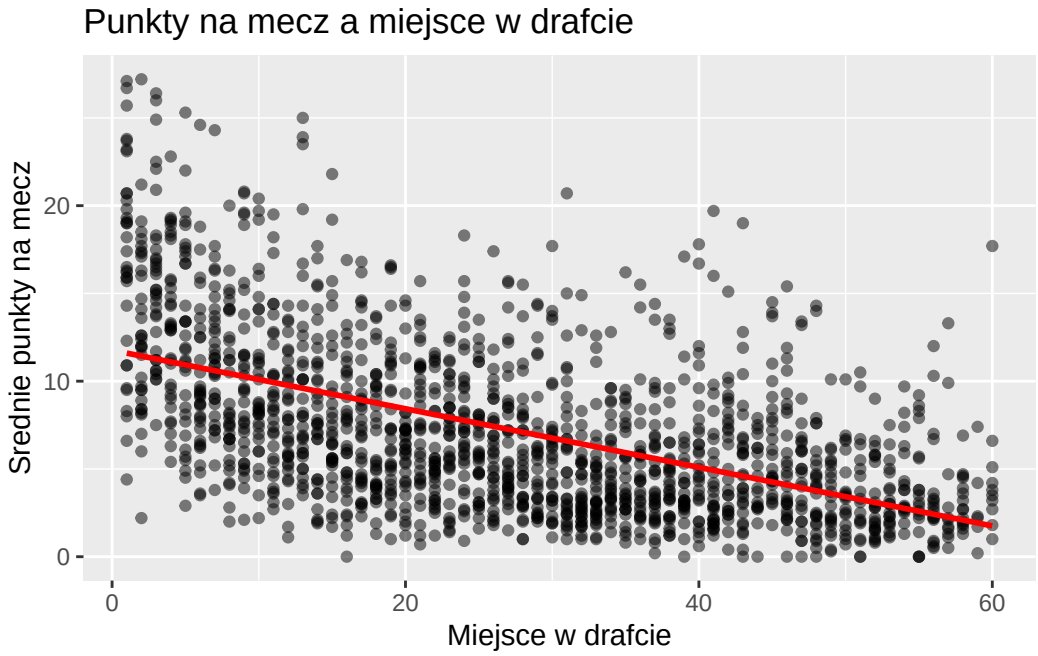
Liczba graczy z Top 100 w conajmniej jednej ze statystyk: punkty na mecz, win shares oraxz box plus-minus wydraftowanych przez drużyny

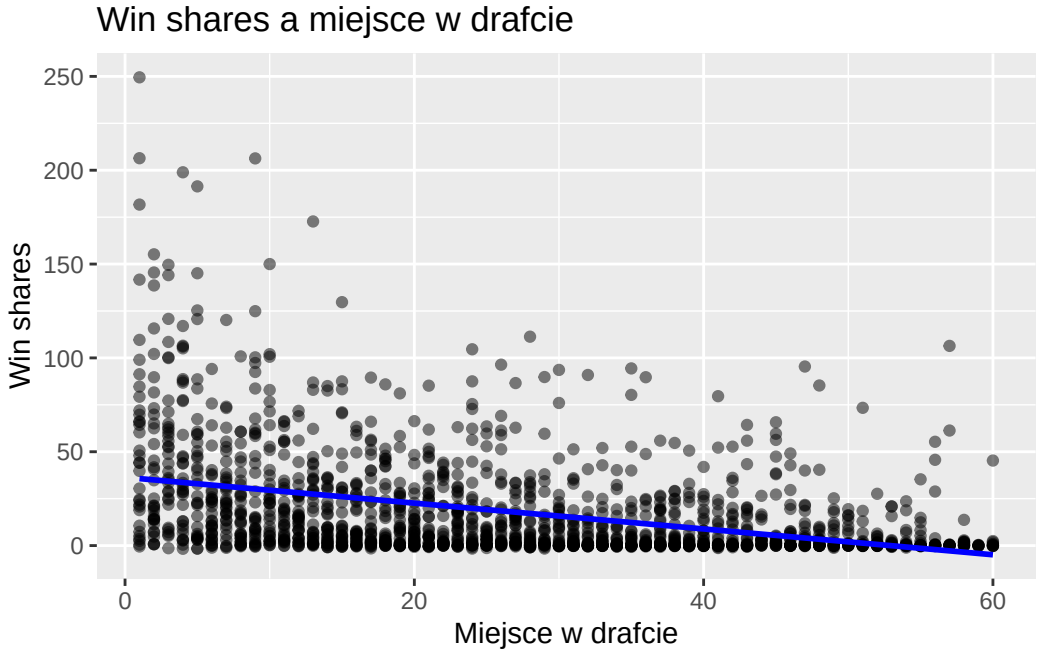


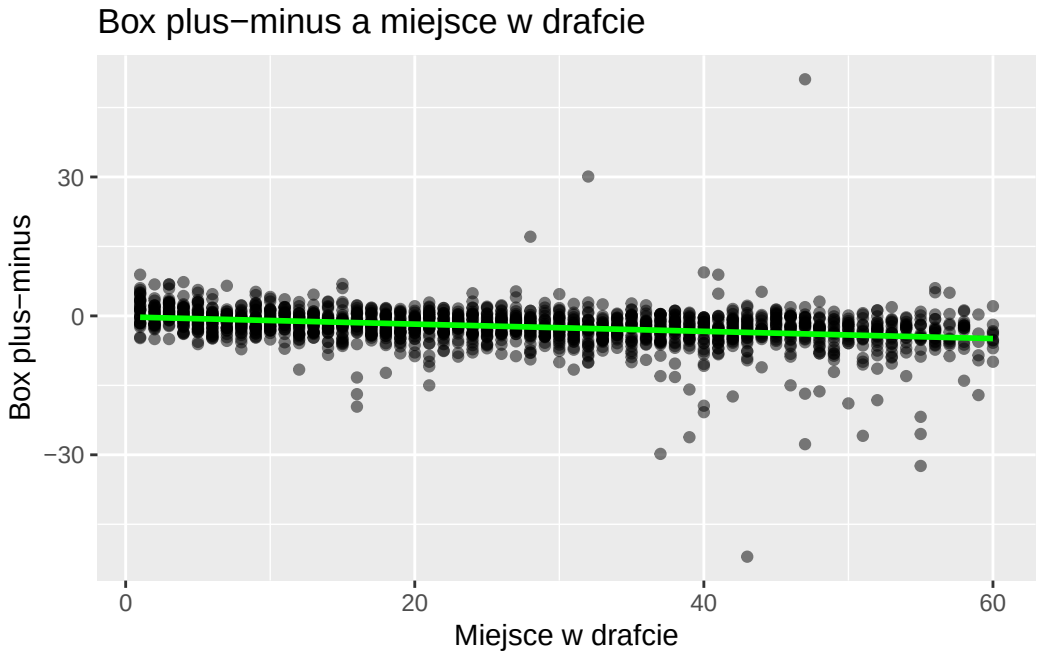
Wnioski: Największą skuteczność w wybieraniu graczy w drafcie miają drużyny takie jak BOS(Boston Celtics) czy GSW(Golden State Warriors). Można zaobserwować, że drużyny z największą ilością picków z top 100, plasują się również w czołówce łącznej ilości wyborów w drafcie, np. Boston Celtics - pierwsze miejsce w ilości zawodników z top 100 oraz 4 w łącznej iości picków. Inne czynniki wpływające na wyniki widoczne na wykresie to jakie picki miały dostępne drużyny (można mieć wile niskich picków co z reguły skutkuje gorszej jakości graczami) oraz skuteczność skautów poszczególnych drużyn. ## Analiza cech ciągłych

Średnie punkty na mecz a miejsce w drafcie

Zbadano, czy wyższe miejsce w drafcie (np. top 10) koreluje z wyższymi osiągnięciami w karierze.


Punkty na mecz a miejsce w drafcie
Srednie punkty na mecz
0
10
20
0
20
40
60
Miejsce w drafcie


Win shares a miejsce w drafcie
Win shares
0
50
100
150
200
250
0
20
40
60
Miejsce w drafcie



Wnioski: Wyższe miejsca w drafcie (np. top 10) mają wyraźną tendencję do wyższych średnich punktów na mecz oraz lepszych statystyk win shares i box plus-minus. Jednak istnieją wyjątki - niektóre niskie pozycje osiągają bardzo dobre wyniki.

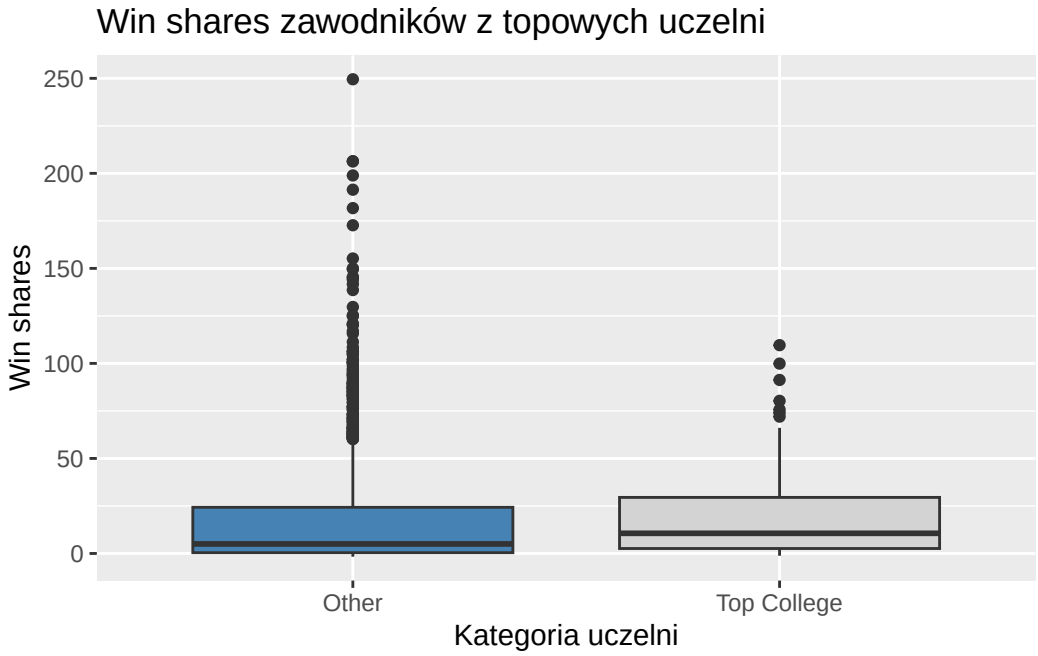
Tabela prentująca średnie punkty, liczbę rozegranych meczów, win shares oraz box plus-minus pomiędzy zawodnikami wybranymi w top 10 a resztą draftowanych.

Tabela 2: Porównanie statystyk zawodników z Top 10 draftu i reszty.

Kategoria	Średnie punkty	Średnia liczba gier	Średnie win shares	Średnie Box plus-minus
Poniżej 10	6.035474	241.6055	12.36811	-2.8117339
Top 10	12.308182	594.6879	40.21303	-0.2821212

Statystyki graczy z najlepszych uczelni

Porównano średnie “win shares” zawodników z uczelni takich jak Duke i Kentucky z resztą.



Wnioski: Zawodnicy z uczelni takich jak Duke i Kentucky osiągają wyższe mediany win shares niż zawodnicy z innych uczelni.

Podsumowanie

1. Miejsce w drafcie a sukces w karierze: Wyższe miejsce w drafcie koreluje z lepszymi wynikami, takimi jak średnia punktów na mecz.
2. Drużyny a sukces zawodników: Drużyny takie jak PHI i BOS częściej wybierają zawodników, ale ich produktywność jest różna.
3. Uczelnie a sukces: Uczelnie Duke i Kentucky mają istotny wpływ na sukces zawodników w NBA, co pokazują statystyki win shares.

Analiza potwierdziła hipotezy o roli miejsca w drafcie oraz znaczeniu wybranych drużyn i uczelni w osiągnięciach zawodników. Osiągnęliśmy cel analizy, zidentyfikowaliśmy istotne zależności oraz wskazaliśmy najważniejsze wzorce w danych.