Wykorzystanie metod statystyki opisowej do analizowania danych rzeczywistych

Miłosz Gajowczyk

Wrocław 21.05.2019

Analizowane dane

Tematem raportu będzie analiza danych pochodzących ze strony: https://www.kaggle.com/drgilermo/nba-players-stats dotyczących graczy. Dane dotyczą zawodników, którzy uczestniczyli w rozgrywkach począwszy od lat 50, aż do dziś.

W swojej pracy skupie się głównie na analizie sylwetek graczy i relacji jakie zachodzą między budową ciała oraz pozycją na jakiej gra zawodnik. Sprawdzę także co można wyciągnąć z tych danych po ich połączeniu – stworzeniu współczynnika wagi do wzrostu. Poruszony zostanie także temat wyboru najlepszych przedziałów do histogramu, czy też wyciągnięcia przydatnych informacji z histogramu, który niekoniecznie nadaje się do prezentacji naszych badań (może mieć jednak duże znaczenie w trakcie analizy).

Wśród danych można znaleźć puste obserwacje, jednak już na początku analizowania scaliłem wszystkie interesujące mnie zmienne, które znajdowały się w dwóch różnych tabelach do nowej tabeli, którą wykorzystywałem w dalszych obliczeniach. W rezultacie kolumny z pustymi polami zostały zlikwidowane, a ja uzyskałem zbiór wszystkich potrzebnych danych w ilości 3814 obserwacji. Zrobiłem to za pomocą kolumny, która zawierała imię i nazwisko gracza oraz znajdowała się w obu tabelach. Zmienne użyte w dalszej analizie to:

- Wzrost – zakres od 160 do 231 cm,
- Waga – zakres od 60 do 163 kg,
- Pozycja, na której grał zawodnik – 8 pozycji.

Podstawowe statystyki

Analizę wybranych danych rozpocząłem od policzenia wartości oczekiwanej (μ) korzystając z następującego wzoru:

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

n – ilość pomiarów (ilość graczy o określonym wzroście lub wadze)
x_i – pomiar i-tego zawodnika.

Po wprowadzeniu wszelkich danych do wzoru w pakiecie matematycznym R Studio otrzymałem:

$$\mu_h \approx 198{,}69$$
$$\mu_w \approx 94{,}78$$

μ_h - wartość oczekiwana dla wzrostu zawodników w centymetrach,
μ_w – wartość oczekiwana dla wagi zawodników w kilogramach.

Taki sam wynik otrzymałem korzystając z wbudowanej w języku R funkcji „mean”, przeznaczonej do wyznaczania wartości oczekiwanej.

Następnym krokiem było policzenie odchylenia standardowego (σ) w celu sprawdzenia jak daleko od średniej są rozrzucone pozostałe pomiary. Początkowo skorzystałem z definicji na wariancję, która jest kwadratem odchylenia standardowego.

$$Var(x) = EX^2 - (EX)^2$$

$Var(X)$ – wariancja,
EX^2 - moment drugiego rzędu,
EX - wartość oczekiwana (w Naszym przypadku średnia S).

Moment drugiego rzędu dla rozkładu dyskretnego można zapisać następująco:

$$\mathrm{E}X^2 = \sum_{i=1}^{n} {x_i}^2 * p_i = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 \qquad p_n = const.\ dla\ n\epsilon[1,n]$$

Licząc z definicji w pakiecie R Studio otrzymałem następujące wyniki:

$$\sigma_h \approx 9{,}22$$
$$\sigma_w \approx 11,99$$

Wyniki jaki otrzymałem posługując się wbudowaną funkcją „var” w R Studio po zaokrągleniu były takie same. Różnice jakie się pojawiały były rzędu 1E-3. W przypadku naszych danych błąd pomiaru jest na tyle mały, że nie wpływa na dalsze obliczenia.

Tak małe odchylenie od średniej, mówi nam, że pomiary nie są mocno rozproszone. Ostatnim krokiem w tej sekcji będzie policzenie współczynnika zmienności w celu potwierdzenia powyższego stanowiska. W odróżnieniu od odchylenia standardowego,
współczynnik zmienności zależy od wielkości arytmetycznej co daje nam dodatkową informację o tym jak duża jest wartość odchylenia na tle naszych obserwacji.

$$v = \frac{\sigma}{\mu} * 100\%$$

Po podstawieniu danych do wzoru otrzymałem $\sigma_w = 12.65\%$ oraz $\sigma_h = 4.64\%$ co tylko potwierdza prawdziwość wcześniejszych obserwacji.

Za pomocą funkcji „max” sprawdziłem maksymalne wartości:

- Wzrost 231 [cm]
- Waga 163 [kg]

Za pomocą funkcji „min” sprawdziłem minimalne wartości:

- Wzrost 160 [cm]
- Waga 60 [kg]

Miara asymetrii, histogram, wykres gęstości

Wygląd histogramu oraz wykresu gęstości możemy przewidzieć na podstawie współczynników miary asymetrii, czyli kurtozy oraz współczynnika asymetrii. Do obliczenia tych wartości (po uprzednim zaimportowaniu biblioteki „propagate”) użyłem wbudowanych w pakiecie R Studio funkcji: „skewness” oraz „kurtosis”. Wzory przydatne do teoretycznych obliczeń wyglądają następująco:

$$\gamma = \frac{E(X-EX)^3}{\sigma^3},$$
$$Kurt(X) = \frac{E(X-EX)^4}{\sigma^4}.$$

Dla analizowanych danych współczynniki wyniosły:

$$\gamma_h \approx -0{,}38$$
$$Kurt(X_h) \approx -0{,}12$$

$$\gamma_w \approx 0{,}27$$
$$Kurt(X_w) \approx 0{,}5$$

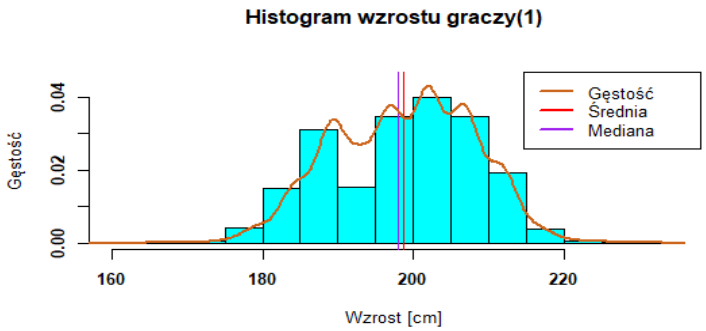
Dane odnośnie wzrostu zawodników:

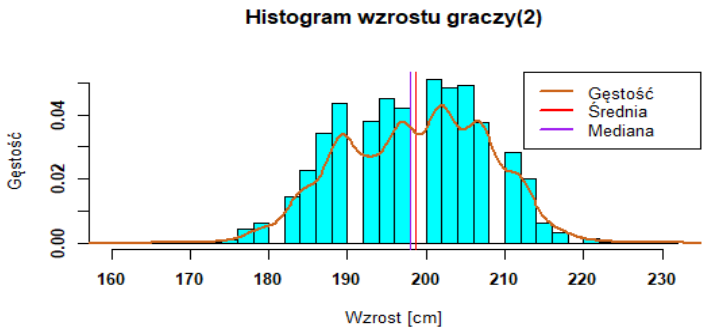
- Ujemna wartość parametru γ mówi o asymetrii lewostronnej.
- Wartość kurtozy porównujemy zazwyczaj z wartością dla rozkładu normalnego (dla podanego powyżej wzoru wynosi 3 jednak pakiet R podaję kurtozę w sposób unormowany, więc będziemy porównywać do zera). Na podstawie tego, że -0.12 < 0 możemy założyć, że otrzymany przez nas wykres będzie spłaszczony. Jest to tak zwany rozkład platokturyczny (ujemna wartość kurtozy).

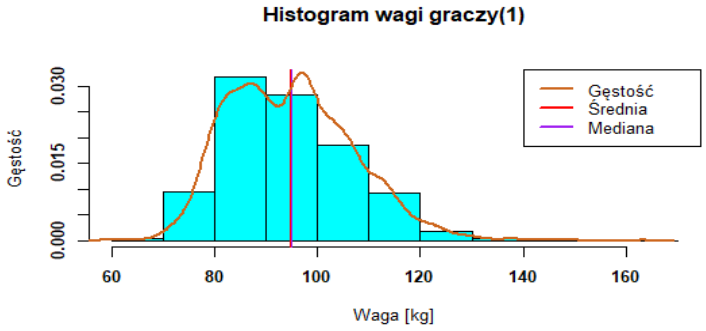
Dane odnośnie wzrostu zawodników:

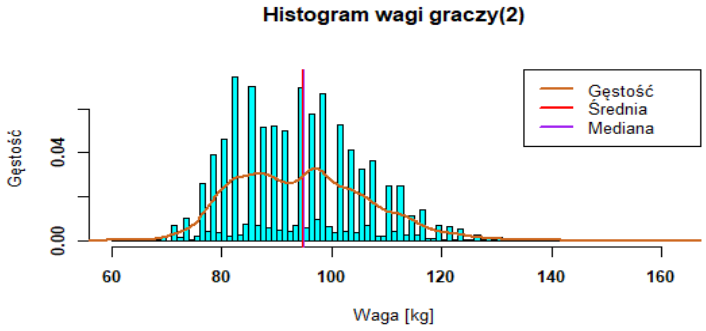
- Dodatnia wartość parametru γ mówi o asymetrii prawostronnej.
- Na podstawie tego, że 0.5 > 0 możemy założyć, że otrzymany przez nas wykres będzie wysmukły.

W tej sekcji skorzystamy z dwóch rodzajów histogramów, na których zaznaczyłem gęstość oraz miary środka. Te z numerem jeden to właściwe histogramy, których będę używał w dalszej analizie. Histogramy z numerem dwa posłużą nam tylko do wyciągnięcia kolejnych wniosków.









Nasuwa się pytanie, dlaczego to akurat wykresy po lewej stronie są histogramami, których chcę w przyszłości używać. Oba z histogramów z większą ilością danych posiadają wady, które mogłyby utrudnić dalszą analizę. W wypadku wykresu dla wzrostu graczy są to puste klasy, zaś problem z wykresem wagi graczy polega na zbyt dużej ilości przedziałów co sprawia, że dane są kompletnie nieczytelne i mogą łatwo doprowadzić do dużej pomyłki (ponieważ obserwacje o dużej gęstości przeplatają się z obserwacjami o małej gęstości). Ponadto drugi histogram dla wagi graczy mocno rozbiega się z gęstością.

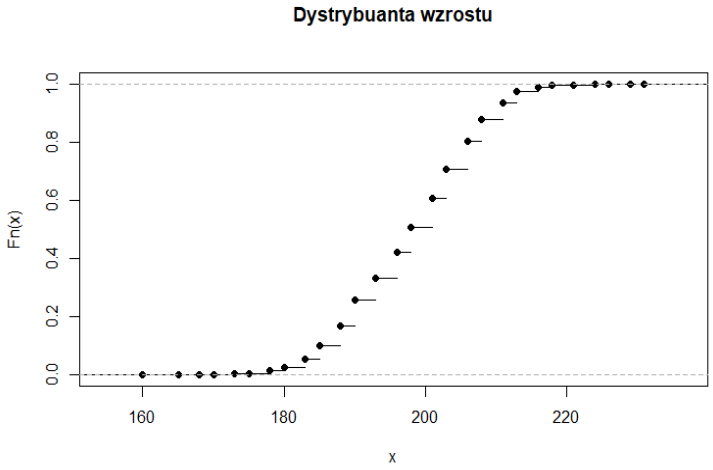
Przejdźmy do histogramów z numerem 1. Gęstość na wykresach nie pokrywa się idealnie z histogramem, jednak jesteśmy w stanie przewidzieć co powoduje tą sytuację w obu przypadkach.

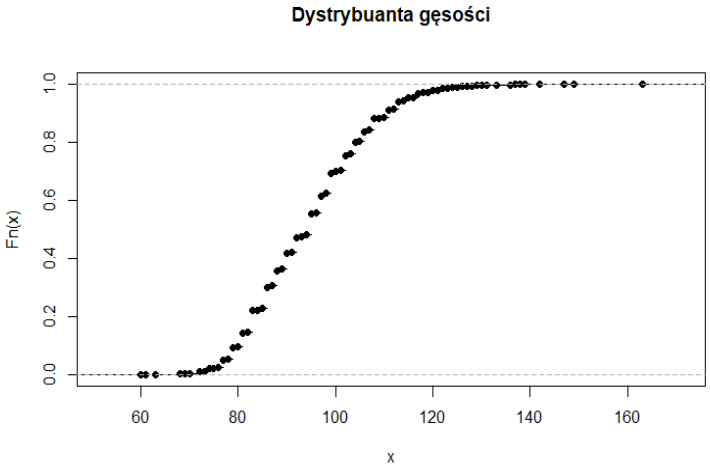
1. Histogram wzrostu zawodników dość mocno rozchodzi się z gęstością na przedziale (190, 195). Spójrzmy na histogram wzrostu numer 2, aby zrozumieć co implikuje takie zachowanie. Jak widać histogram wzrostu graczy byłby bardzo zbliżony do rozkładu normalnego, gdyby nie wcześniej wspomniane puste klasy, które zaniżają średnią dla przedziału. Warto także zapamiętać podobieństwo (po wyłączeniu pustych wartości) tego rozkładu do rozkładu normalnego.

2. Histogram wagi zawodników składa się z 10 słupków (niewidoczne na brzegach to pojedyncze obserwacje), tymczasem wektor z informacjami o wadze zawodników posiada aż 75 unikalnych obserwacji. Ilość słupków, którymi dysponujemy nie pozwala nam skutecznie dopasować histogramu do gęstości. Ponadto, gdy rozpatrzymy każdą obserwację osobno zauważymy, że te z dużą gęstością przeplatają się z tymi o ledwo zauważalnej gęstości. Taki rozkład zwiększa zagrożenie pojawienia się dużych zaokrągleń w wypadku, gdy mamy do czynienia z dużymi i nagłymi zmianami. Przykładem może być przedział (110, 120), w którym pojawiło się dużo więcej obserwacji o niskiej gęstości, co spowodowało dość mocne odchylenie gęstości od histogramu.

Kształt powyższych histogramów potwierdza poprzednie wnioski. W wypadku histogramu wzrostu na pierwszy rzut oka nie widać, czy wykres jest asymetryczny, jednak wystarczy przypomnieć sobie wcześniej omówiony spadek na przedziale (190, 195) oraz zauważyć, że pierwsze obserwacje o zauważalnej gęstości pojawiają się dopiero w 4 tym przedziale, aby rozwiać wątpliwości. Puste klasy sprawiają, że wykres jest bardziej spłaszczony niż wysmukły. W przypadku histogramu wagi jeden z mniejszych słupków, po którym od razu następuje największy utrudnia nam odgadnięcie z jaką asymetrią mamy do czynienia, dlatego musimy zdać się na nasze obliczenia i założyć, że jest prawostronnie asymetryczny.

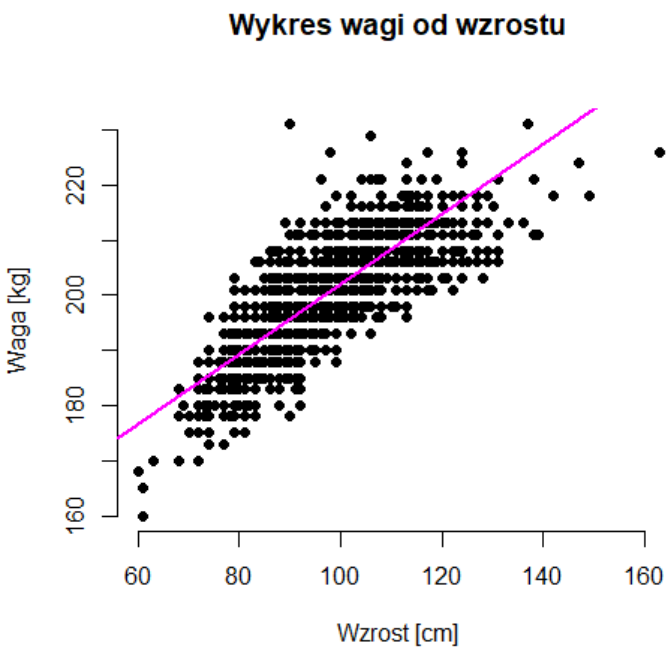
Obliczyłem także dystrybuanty badanych zmiennych i wstawiłem ich wykresy poniżej.

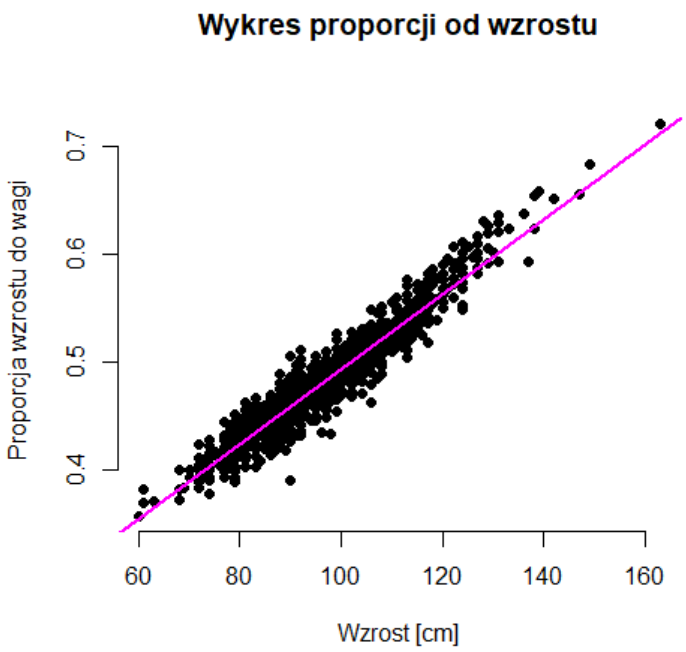




Wykres rozrzutu

Kolejnym punktem w sprawozdaniu będzie sprawdzenie wykresu rozrzutu w celu znalezienia zależności między zmiennymi. Korzystam tutaj z wbudowanej funkcji „plot”.



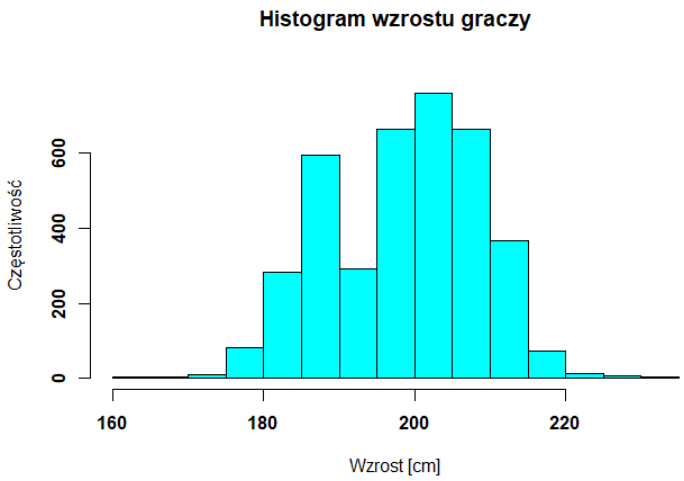


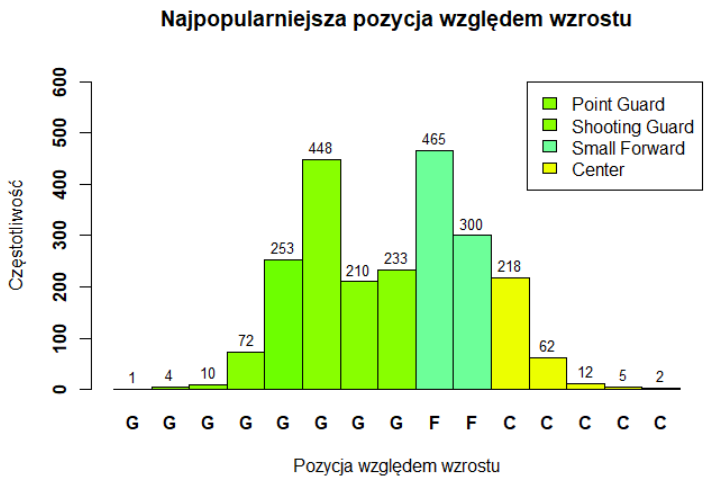
Po wygenerowaniu pierwszego wykresu zauważyłem, że waga i wzrost rosną liniowo jednak panuje tam spore rozproszenie, dodatkowo występuje kilka wartości odstających przez co ciężko dokonać regresji liniowej. Postanowiłem unormować ten wykres poprzez zastąpienie wagi proporcją wagi do wzrostu. Zaszyfrowane w ten sposób dane są wciąż proste do odczytania, a my widzimy, że występuje tutaj liniowa zależność między danymi.

Najpopularniejsze pozycje graczy względem wzrostu i wagi

W tej sekcji skupię się na przeanalizowaniu grup, na które zostali podzieleni zawodnicy względem wagi i wzrostu w celu sprawdzenia jaka pozycja jest najpopularniejsza dla graczy o konkretnej sylwetce. Powszechną wiedzą jest fakt, że budowa zawodnika gra ogromną rolę w jego predyspozycjach do grania na określonej pozycji. Postanowiłem sprawdzić, jak wygląda ta zależność wśród zawodników NBA. Należy pamiętać, że sposób w jaki analizuję dane może w łatwy sposób doprowadzić do wyciągnięcia fałszywych wniosków ze względu na zły dobór punktów granicznych w histogramie dotyczącym pozycji graczy. Wykresy przedstawione poniżej były wielokrotnie sprawdzane, wersje, które zamieszczam w raporcie uznałem za najbardziej dokładne, ponieważ pokrywają się z histogramami wzrostu, wagi oraz proporcji wagi do wzrostu bez widocznych przekłamań. Przykładowo przy innym doborze parametrów mogłoby się okazać, że najpopularniejsza pozycja na przedziale od 190 do 195 cm liczy aż 400 obserwacji, kiedy na podstawie histogramu wzrostu widzimy, że graczy na tym przedziale jest tylko lekko ponad 200. Jest tak z uwagi na to, że graczy o wzroście 190cm jest aż 334, a histogram wzrostu składa się z przedziałów prawostronnie zamkniętych, więc 190 wlicza się do obserwacji z przedziału od 185 do 190.

Pierwszym krokiem było przeanalizowanie wzrostu zawodników pod kątem pozycji, na której grają.

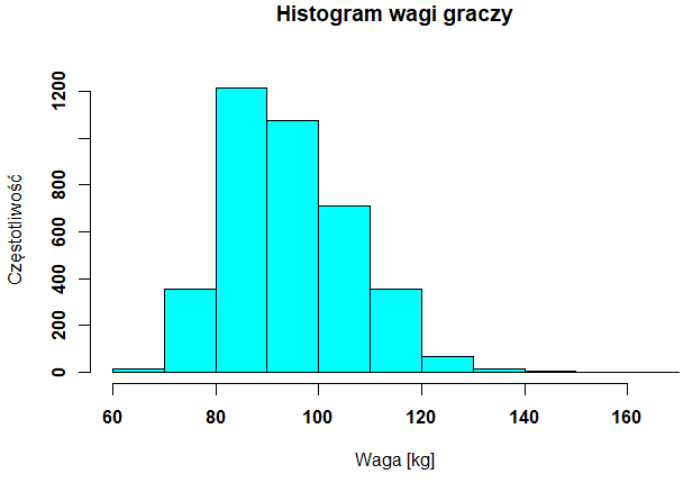


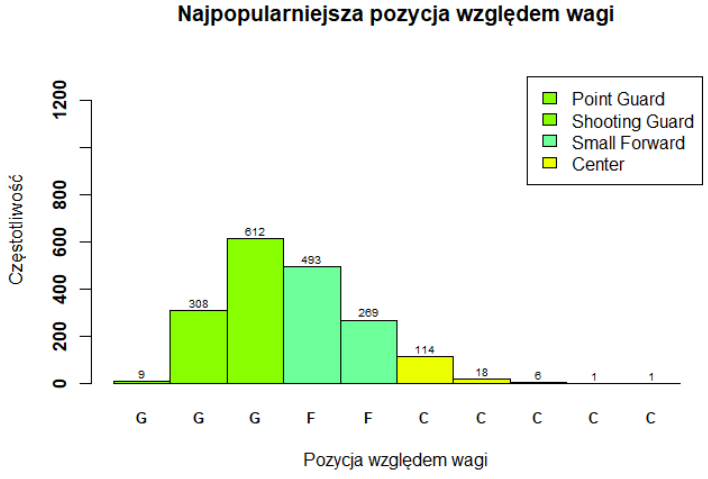


Na podstawie tych danych możemy dojść do następujących wniosków:

- Zawodnicy niscy (do 185cm) są całkowicie zdominowani przez rozgrywających i rzucających obrońców.
- Najpopularniejsza pozycja na przedziale (185-190cm) to także rozgrywający lub rzucający obrońca jednak tutaj jest to zdecydowanie niższy procent zawodników, bo około 75%.
- Gracze na przedziale (190-195cm) to głównie rozgrywający i rzucający obrońcy. Mamy tutaj do czynienia z kolejnym wzrostem ich występowania, ponieważ procentowo ich ilość wygląda podobnie jak wypadku pierwszych 5 kolumn (wzrost do 185cm).
- Przedział od 200 do 210cm jest zdominowany przez skrzydłowych, jednak w wypadku przedziału 205-210cm jest to zaledwie 50% graczy o takim wzroście, więc nie ma tutaj mowy o takiej dominacji jak w wypadku rozrywających, czy też rzucających obrońców.
- Powyżej 210cm najczęściej występująca pozycja to center.
- Wszystkie pozycje powyżej 215cm to center.

Następnie pod tym samym kątem przeanalizowałem dane dotyczące wagi zawodników.

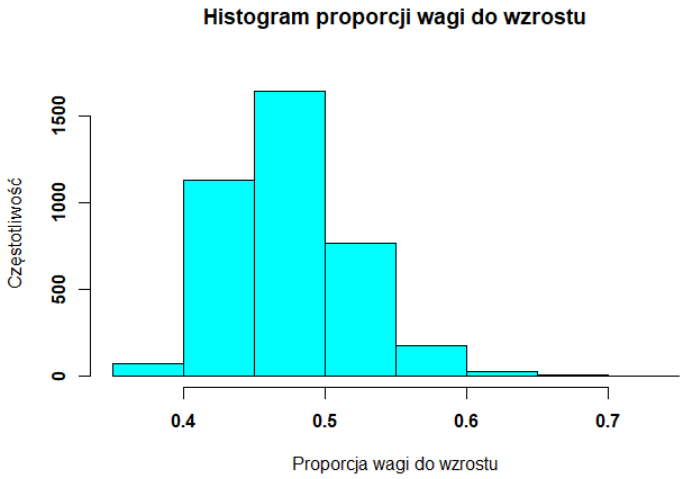


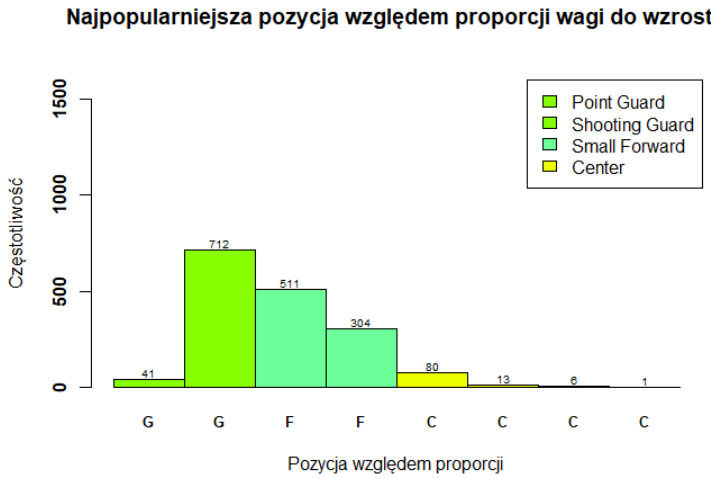


Na podstawie wykresów możemy zauważyć, że:

- Gracze do 80kg są zdominowani przez rozgrywających oraz rzucających obrońców.
- Gracze ważący od 80 do 90kg to w połowie rozgrywający oraz rzucający obrońcy, widać że panuje tutaj większa różnorodność.
- Skrzydłowi to około 70% zawodników ważących od 90 do 110kg.
- Najpopularniejsza pozycja dla wagi od 110 do 130 kg to center, widać jednak, że na tych przedziałach panuje zróżnicowanie i nie tylko centrzy potrafią grać wykorzystując przewagę swojej masy.
- Przedział od 130 do 160kg to wyłącznie centrzy. Jest tak prawdopodobnie dlatego, że przewaga masowa nad przeciwnikiem jest najbardziej pożądana na tej właśnie pozycji.

Ostatnim krokiem będzie przeanalizowanie jak wygląda zależność wagi do wzrostu, ponieważ wcześniejsza analiza nie daje nam pełnej informacji o sylwetce graczy. Z dwóch osobnych histogramów ciężko wywnioskować, czy gracze wysocy to jednocześnie Ci najcięższi, czy może jednak model wysokiego gracza zakłada, że jest to osoba wyjątkowo szczupła.





Łącząc te dane z wcześniejszymi wykresami możemy zauważyć, że:

- Rozgrywających i rzucających obrońców o:
    1. wzroście z przedziału od 160 do 195cm jest 998,
    2. wadze do 60 do 90kg jest 929,
    3. proporcji wagi do wzrostu <= 0.45 jest 753.

    Zawodnicy grający na tych pozycjach to około 70% najszczuplejszej grupy zawodników NBA. Możemy też założyć, że większość z nich mieści się w obu grupach: z punktu 1 oraz 2.

- Skrzydłowych o:
    1. wzroście z przedziału od 200 do 210cm jest 765,
    2. wadze do 90 do 110kg jest 762,
    3. proporcji wagi do wzrostu od 0.45 do 0.55 jest 815.

    Zawodnicy grający na tych pozycjach to około 40% zawodników NBA z grupy o najbardziej zbalansowanej sylwetce. Możemy też założyć, że większość z nich mieści się w obu grupach: z punktu 1 oraz 2, jest to jednak sylwetka dużo częściej osiągana przez osoby, nie należące do obu tych grup (w tym wypadku osób o rozważanej proporcji jest więcej niż osób, które mogłyby przynależeć do obu grup).

- Centrów o:
    1. wzroście z przedziału od 195 do 231 jest 299,

2. wadze do 110 do 160kg jest 140,
3. proporcji wzrostu od 0.55 do 0.75 jest 100.

Zawodnicy grający na tych pozycjach to około 60% grupy zawodników NBA o największej wadze. Możemy też założyć, że większość zawodników z drugiej grupy mieści się także w pierwszej. Widzimy, że mimo powszechnej wiedzy o tym, że to center jest zazwyczaj osobą o dużo większym współczynniku wagi do wzrostu to w grupie panuje spore zróżnicowanie.

.

Kwartyle oraz wykres pudełkowy

Jako jeden z ostatnich etapów analizy danych policzę kwartyle, (pierwszego, drugiego i trzeciego rzędu) a następnie sprawdzę ich zgodność z wykresem pudełkowym.
Kwartyl drugiego rzędu to inaczej mediana (funkcja 'median'), z definicji dla parzystej ilości obserwacji wzór na medianę prezentuje się następująco:

$$M = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}.$$

Innymi słowy *M* dla parzystej ilości posortowanych obserwacji to po prostu średnia arytmetyczna dwóch środkowych wyrazów. Do wyliczenia kwartyli korzystam z funkcji „quantile". Dla naszych obserwacji wyniki prezentują się następująco:

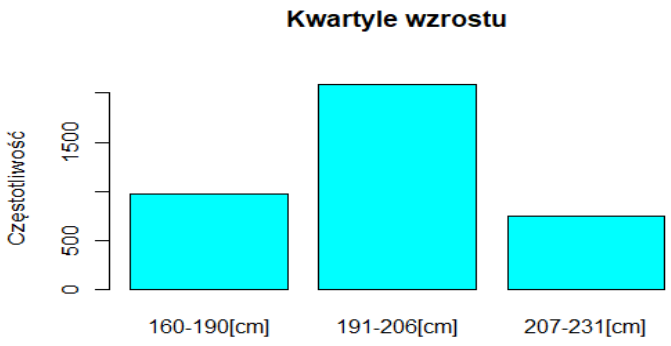
- Dla danych dotyczących wzrostu:

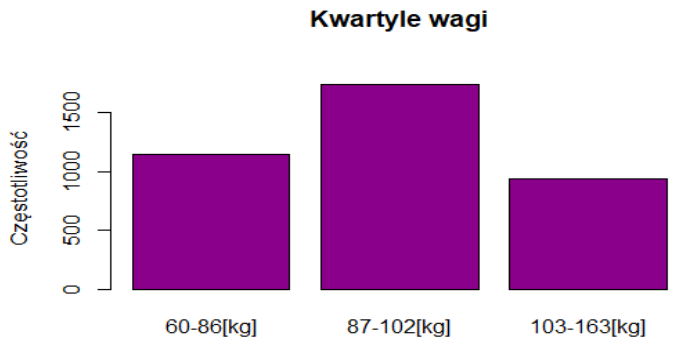
$$Q_{h1} = 190$$
$$Q_{h2} = M = 198$$
$$Q_{h3} = 206$$

- Dla danych dotyczących wagi:

$$Q_{w1} = 86$$
$$Q_{w2} = M = 95$$
$$Q_{w3} = 102$$

Kwartyl *Q1* można policzyć również licząc medianę dla pierwszej połowy posortowanych obserwacji, analogicznie kwartyl *Q3* dla drugiej połowy obserwacji. Licząc w taki sposób otrzymałem wyniki identyczne jak powyżej. Zauważamy tutaj, że w obu przypadkach mediana jest zbliżona do wartości oczekiwanej (198.69 dla wzrostu oraz 94.78 dla wagi) co tylko potwierdza wcześniejsze obserwacje odnośnie w miarę równomiernie rozłożonych danych.



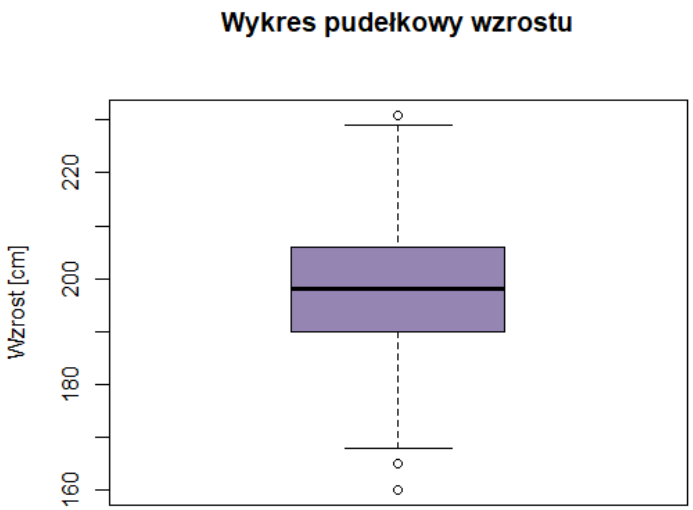


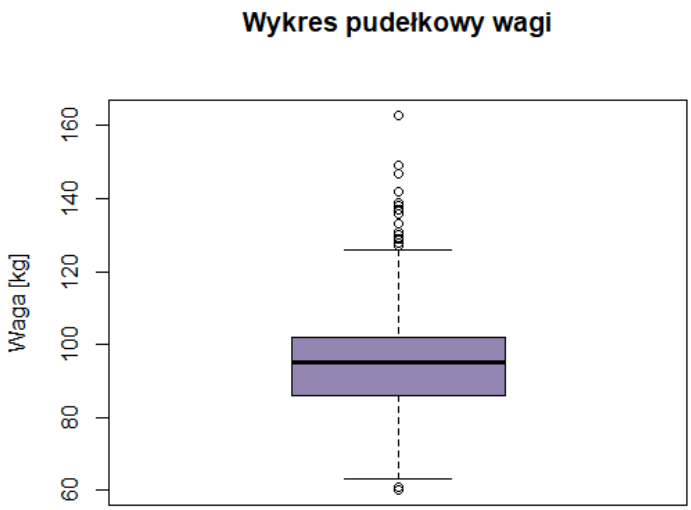
Policzyłem także rozstęp międzykwartylowy, który jest różnicą pomiędzy trzecim, a pierwszym kwartylem:

$$Q_{h3} - Q_{h1} = 206 - 190 = 16$$

$$Q_{w3} - Q_{w1} = 102 - 86 = 16$$

W obu przypadkach różnica w porównaniu ze średnią lub medianą jest niewielka, należy jednak zauważyć, że w wypadku zestawienia wag zawodników z tymi miarami jest dwa razy większa niż w wypadku ich wzrostów.





Wygenerowałem wykres pudełkowy, aby dodatkowo sprawdzić poprawność moich wyników. Jak widać obliczone przez nas mediany pokrywają się z środkowymi liniami w pudełku, co potwierdza, że zostały dobrze policzone. Obserwacji, które odstają jest znacznie więcej w wypadku danych związanych z wagą zawodników, co zgadza się z obliczonym powyżej rozstępem międzykwartylowym.

Podsumowanie

Mimo, że dane które wybrałem nie okazały się najciekawsze udało mi się dojść do kilku ciekawych wniosków dotyczących zarówno analizowania danych jak i analizowanych danych. Postaram się przytoczyć tutaj najważniejsze z nich, w kolejności w jakiej pojawiały się w raporcie.

- Współczynnik zmienności dla wagi zawodników NBA jest niemal trzykrotnie większy niż współczynnik zmienności dla wzrostu.
- Warto przeanalizować histogramy z różnie dobranymi przedziałami, ponieważ może nam to dostarczyć dodatkowych informacji.
- Można próbować unormować wykres rozrzutu poprzez dodanie współczynnika, na który składają się obie zmienne. W rezultacie możemy uzyskać bardziej czytelny wykres, który mówi nam o relacji panującej pomiędzy badanymi zmiennymi.
- Sposób w jaki dobierzemy skraje punkty w histogramie ma ogromne znaczenie, ponieważ źle dobrane mogą całkowicie zmienić wygląd naszego wykresu.
- W przypadku wzrostu mamy do czynienia z delikatną asymetrią lewostronną co oznacza, że większa ilość obserwacji jest zauważalna powyżej przeciętnej
- W przypadku wagi mamy do czynienia z delikatną asymetrią prawostronną co oznacza, że większa część graczy posiada wagę poniżej przeciętnej.

- Sylwetka ma ogromne znaczenie w NBA. Najczęstsze pozycje dla wzrostu, wagi oraz proporcji tych dwóch zmiennych układają się we wszystkich przypadkach od rozgrywającego lub rzucającego obrońcy przy niskich wartościach, poprzez skrzydłowych przy średnich wartościach, aż do centrów, którzy przeważają w górnych progach zarówno przy wzroście jak i przy wadze.
- Dane nie są mocno rozproszone, jednak waga zawodników wacha się dwukrotnie bardziej niż wzrost co widać po rozstępie międzykwartylowym oraz wykresach pudełkowych.
- Centrzy wcale nie dominują grupy osób z najwyższą proporcją wagi do wzrostu tak mocno jak nam się wydaje.