Statystyki Opisowe

Wprowadzenie:

1. Celem analizy było zbadanie statystyk dla dwóch cech horsepower (hp), czyli liczba koni mechanicznych, moc silnika, oraz rear axle ratio (drat) czyli ilość obrotów, które wał napędowy musi wykonać, aby oś obróciła się jeden raz. Zostały sprawdzone podstawowe parametry oraz zależność między dwoma cechami.
2. Celem analizy było sprawdzenie zależności między oczekiwaną długością życia, a liczbą populacji danego kraju.

Teoria:

Pojęcia, które będą używane w dalszej analizie: **mean** (średnia) - średnia arytmetyczna wszystkich wartości, **median** (mediana) - wartość środkowa w uporządkowanym zbiorze danych, **quartile** (kwartyle) - wartości cechy w proporcjach co 25%, **kurtosis** (kurtoza) - miara koncentracji wyników, wskazuje na zakres występowania wartości odstających, **skewness** (skośność) - rozbieżność między wartością średnią, a centrum danego rozkładu, **variance** (wariancja) - średnia arytmetyczna kwadratów odchyleń wartości od średniej, **standard deviation** (odchylenie standardowe) - miara rozproszenia wokół średniej, **normal distribution** (rozkład normalny) - ciągły rozkład symetryczny, **kernel density distribution** (rozkład gęstości jądrowej) - metoda estymacji funkcji gęstości prawdopodobieństwa, wygładza histogram, **rug plot** - wizualizacja rozkładu danych w formie krótkich pionowych kresek, **korelacja** - analiza korelacji, sprawdza czy istnieją zależności między zmiennymi.

Dane:

Zbiór danych mtcars pochodzi z amerykańskiego magazynu Motor Trends z 1974. Zawiera on statystyki w 11 parametrach na podstawie 32 samochodów.

Zbiór danych gapminder pochodzi z niezależnej non-profit Szwedzkiej organizacji. Posiada dane statystyczne na temat m. in. wartości GDP na przestrzeni lat.

Analiza: zadanie 1

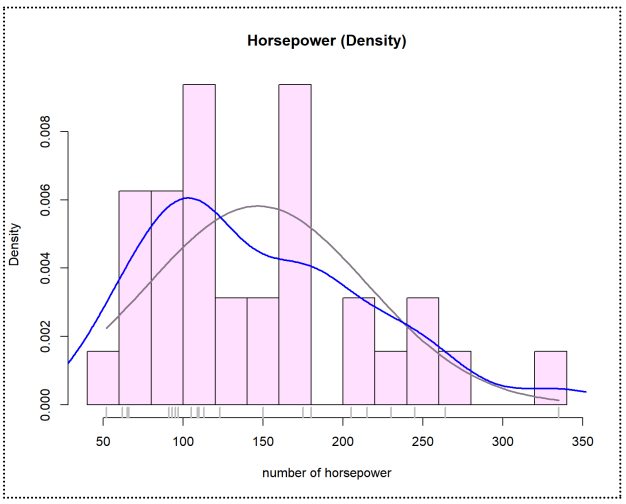
1. Statystyki opisowe
   Na podstawie zbioru danych obliczono podstawowe wartości tj. minimalna wartość, maksymalna wartość, średnia, mediana oraz 1 i 3 kwartyl.

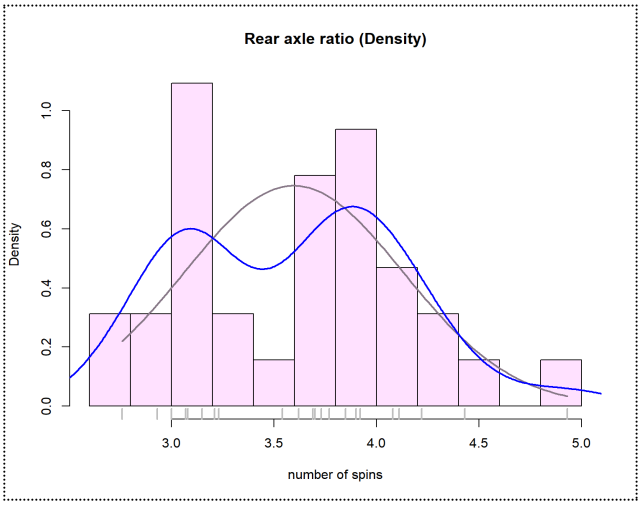
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
hp	52.0	96.5	123.0	146.7	180.0	335.0
drat	2.760	3.080	3.695	3.597	3.920	4.930

2. Analiza rozkładu danych

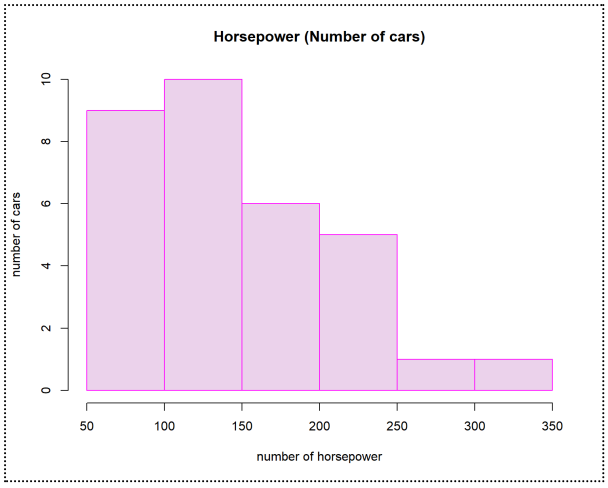
	kurtosis	skewness	variance	sd	range
hp	-0.1355511	0.7260237	4700.867	68.56287	283
drat	-0.7147006	0.2659039	0.2858814	0.5346787	2.17

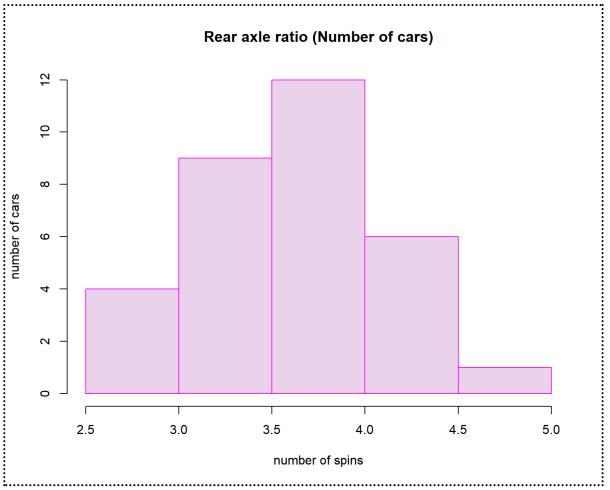
Wartość hp ma lekko ujemną kurtozę - oznacza to typ platykurtyczny (K<0) czyli intensywność wartości ekstremalnych jest mniejsza niż w przypadku rozkładu normalnego. Dodatkowo, współczynnik skośności jest większy od zera, co wykazuje prawostronną skośność, wartość średniej jest większa od mediany. Cecha hp charakteryzuje się dużą wariancją oraz odchylenie standardowe - wartości są daleko od średniej. W przypadku cechy drat dostrzegamy bardziej ujemną kurtozę oraz mniejszą prawostronną asymetrię. Wariancja i odchylenie natomiast są mniejsze. Wyniki są lepiej widoczne na poniższych wizualizacjach gęstość oraz zależnościach między ilością samochodów a wartością danej cechy:





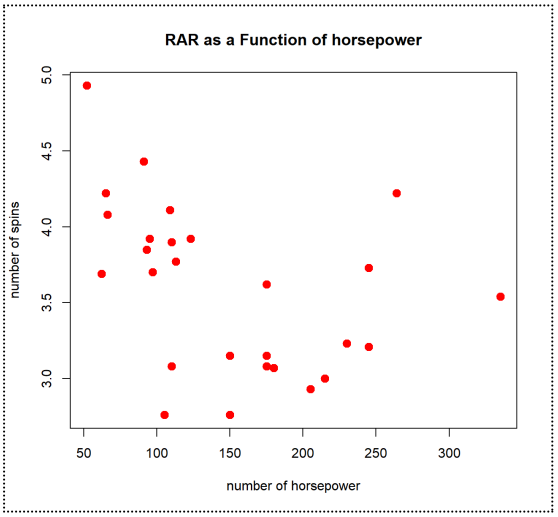
Ciemnoróżowa linia pokazuje rozkład normalny, natomiast niebieska linia rzeczywisty rozkład gęstości. Szare linie reprezentują rozkład danych.

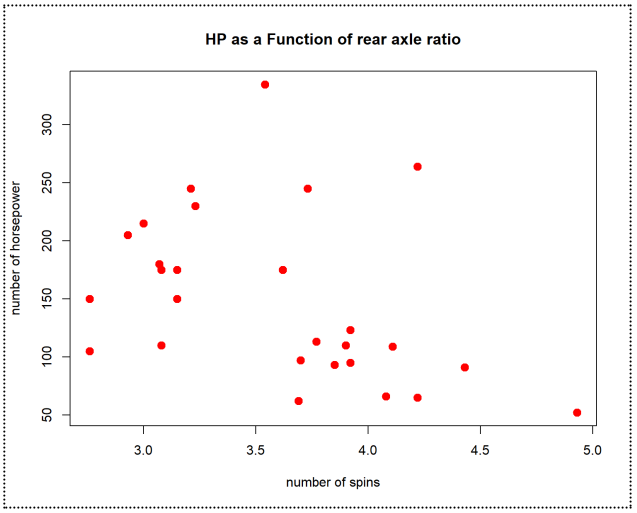




Na podstawie wykresów możemy stwierdzić, że cecha Horsepower (hp) charakteryzuje się większą zmiennością, a cecha Rear Axle Ratio (drat) jest bardziej stabilna.

3. Korelacja





Na powyższych wykresach widzimy (od lewej) zależność liczby obrotów od liczby koni mechanicznych oraz zależność liczby koni mechanicznych od liczby obrotów. Na pierwszy rzut oka nie są widoczne żadne wyraźne zależności. Dlatego, wykonany został **test korelacji liniowej Pearsona**, badający związki liniowe, w którym zwiększenie wartości jednej cechy powoduje proporcjonalne zwiększenie drugiej cechy. Wartość korelacji wynosi **-0.4488** co oznacza umiarkowaną negatywną korelację. P-value, czyli wartość wskazująca prawdopodobieństwo, że taka korelacja może zostać uzyskana przypadkowo wynosi **0.009989**. Wartość $p < 0.05$ co oznacza, że korelacja jest **istotna statystycznie**.

Podsumowanie:
Cecha hp wykazuje dużą zmienność, ma asymetryczny rozkład - znaczną prawostronną skośność.
Wartości cechy drat charakteryzuje mniejsza intensywność wartości ekstremalnych niż w rozkładzie normalnym oraz mała wariancja.
Test korelacji liniowej Pearsona wykazał umiarkowaną negatywną korelację.

Analiza: zadanie 2

1. Relacja między oczekiwaną długością życia a populacją

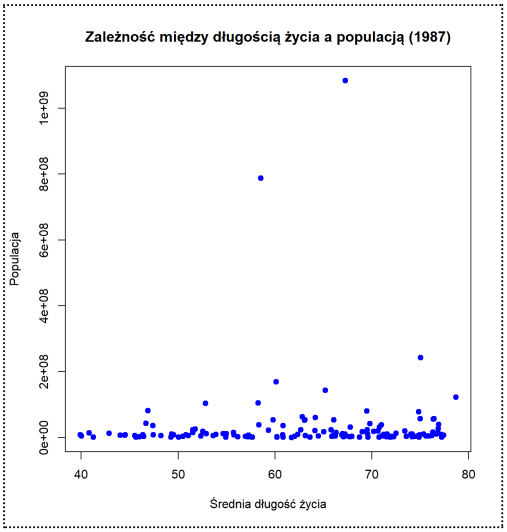
Na podstawie wykresu możemy zauważyć, że wraz z zwiększającą się populacją, lekko zwiększa się długość życia. Pomimo tej dodatniej relacji kraje o bardzo dużej populacji tj. Chiny (67 lat) lub Indie (58) mają mniejszą oczekiwaną długość życia od małych krajów np. Kuwejt (74 lat).

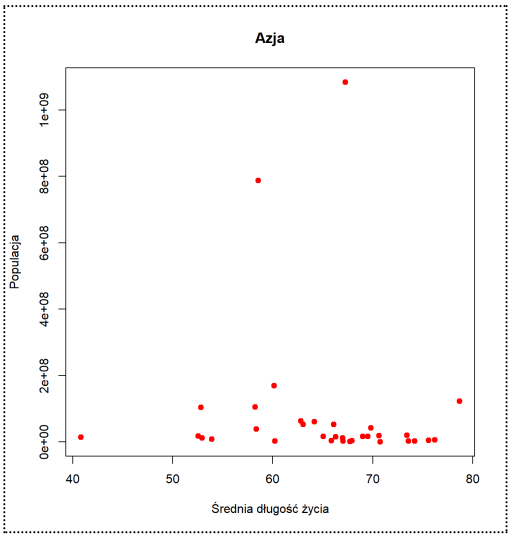
2. Korelacja

W wyniku przeprowadzenia testu korelacji Pearsona otrzymaliśmy wynik **0.0331**, co potwierdza bardzo słabą pozytywną korelację między zmiennymi. Wartość p wynosi **0.6961** czyli jest większa niż 0.05, czyli **nie możemy stwierdzić statystycznie istotnej korelacji.**

Podsumowanie:

Nie można stwierdzić zależności między długością życia i wielkością populacji danego kraju.





Bibliografia:
https://www.badgertruck.com/heavy-truck-information/what-is-axle-ratio/
https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/mtcars
https://cran.r-project.org/web/packages/gapminder/readme/README.html
https://www.gapminder.org/

Kod:

```
library(datasets)
head(mtcars)

summary(mtcars$hp)
summary(mtcars$drat)

kurtosis(mtcars$hp)
kurtosis(mtcars$drat)

skewness(mtcars$hp)
skewness(mtcars$drat)

var(mtcars$hp)
var(mtcars$drat)

sd(mtcars$hp)
sd(mtcars$drat)

max(mtcars$hp)-min(mtcars$hp)
max(mtcars$drat)-min(mtcars$drat)

hist(mtcars$hp,
    xlab = "number of horsepower",
    ylab = "number of cars",
    main = "Horsepower (Number of cars)",
    border = "magenta",
    col = "thistle2")

hist(mtcars$hp,
       xlab = "number of horsepower",
       main = "Horsepower (Density)",
       col = "thistle1",
       breaks = 14,
    freq = FALSE)

curve(dnorm(x, mean = mean(mtcars$hp),
sd = sd(mtcars$hp)),
     col = "thistle4",
     lwd = 2,
     add = TRUE,
     from = min(mtcars$hp),
     to = max(mtcars$hp))

lines(density(mtcars$hp), col = "blue", lwd
= 2)
rug(mtcars$hp, lwd =2, col = "gray")

hist(mtcars$drat,
    xlab = "number of spins",
    ylab = "number of cars",
    main = "Rear axle ratio (Number of
cars)",
    border = "magenta",
    col = "thistle2")

hist(mtcars$drat,
    xlab = "number of spins",
    main = "Rear axle ratio (Density)",
    col = "thistle1",
    breaks = 14,
    freq = FALSE)

curve(dnorm(x, mean =
mean(mtcars$drat), sd = sd(mtcars$drat)),
     col = "thistle4",
     lwd = 2,
     add = TRUE,
     from = min(mtcars$drat),
     to = max(mtcars$drat))

lines(density(mtcars$drat), col = "blue",
lwd = 2)
rug(mtcars$drat, lwd =2, col = "gray")

plot(mtcars$drat, mtcars$hp,
    pch = 19,
    cex = 1.3,
    col = "red",
    main = "HP as a Function of rear axle
ratio",
    ylab = "number of horsepower",
    xlab = "number of spins")

plot(mtcars$hp, mtcars$drat,
    pch = 19,
    cex = 1.3,
    col = "red",
    main = "RAR as a Function of
horsepower",
    ylab = "number of spins",
```

```
    xlab = "number of horsepower")

cor.test(mtcars$hp, mtcars$drat)

# Zadanie 2

install.packages("gapminder")
install.packages("dplyr")
library(gapminder)
library(dplyr)

View(gapminder)
gap_1987 <- filter(gapminder, year ==
1987)

plot(gap_1987$lifeExp, gap_1987$pop,
    main = "Zależność między długością
życia a populacją (1987)",
    xlab = "Średnia długość życia",
    ylab = "Populacja",
    pch = 19, col = "blue")

asia <- filter(gap_1987, continent ==
"Asia")
plot(asia$lifeExp, asia$pop,
    main = "Azja",
    xlab = "Średnia długość życia",
    ylab = "Populacja",
    pch = 19, col = "red")

cor.test(gap_1987$lifeExp,
gap_1987$pop)
```