1. Streszczenie raportu

Raport powstał w oparciu o dane dotyczące zawodników światowego rankingu tenisowego ATP/WTA. Zawierają między innymi informacje o wieku zawodnika w którym miał największe osiągnięcia oraz jakie osiąga wyniki w zależności od nawierzchni.

2. Opis danych

Dane do projektu pochodzą ze strony : https://www.kaggle.com/ramjasmaurya/tennis-atp-rankings-based-on-elo-scores?select=players+current+elo+rating+by+ATP.csv.

3. Analiza danych

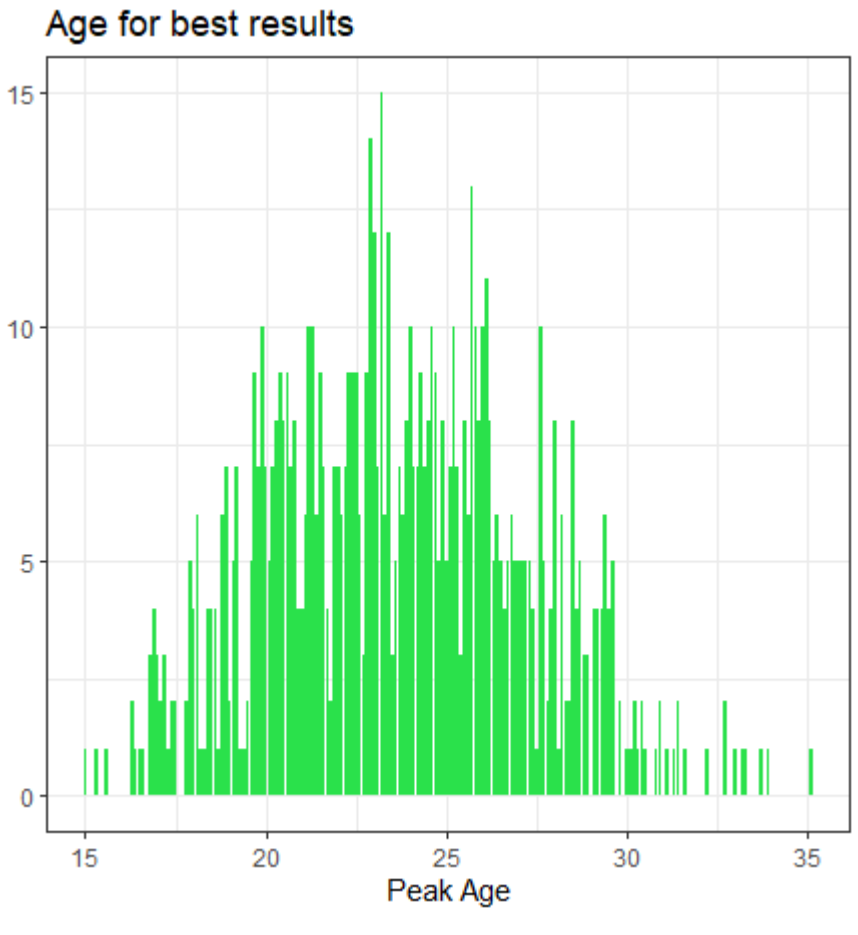
W pierwszym etapie sprawdzam czy dane nie zawierają null / wartości pustych.

```
colSums(is.na(ATP_main))
```

Wyliczam podstawowe dane dla wieku w którym zawodnik osiągał najlepsze wyniki (estymatory - średnia, wariancja, odchylenie standardowe)
- średnia: 23.7
- wariancja: 12.3
- odchylenie standardowe: 3.51

```
age <- sapply(ATP_main[2], mean) #
PAge <- sapply(ATP_main[10], mean)
Pvar <- sapply(ATP_main[10], var)
Psd <- sapply(ATP_main[10], sd) #
```

Tworzę histogram wieku w którym zawodnik osiągał najlepsze wyniki



Histogram symetryczny (z delikatnymi odchyleniami), dominanta jest zbliżona do estymowanej wartości oczekiwanej.

Dominanta:

```
tablica <- table(ATP_main[10])
tablica_s <- names(sort(tablica, decreasing=T))
dominant <- tablica_s[1]
```

Korzystam z rozkładu t-studenta w celu stworzenia przedziałów ufności.
Tworzę funkcję która będzie obliczać przedziały ufności dla wartości oczekiwanej w zależności od parametru alfa (1 - alfa -> poziom ufności).

```
range <- function(lvl) {
  alfa <- 1 - lvl
  return (round(PAge+c(-1,1)*Psd/sqrt(n)*qt(1-alfa/2, n-1),2))
}
```

wynik dla 90% : 23.45 23.86

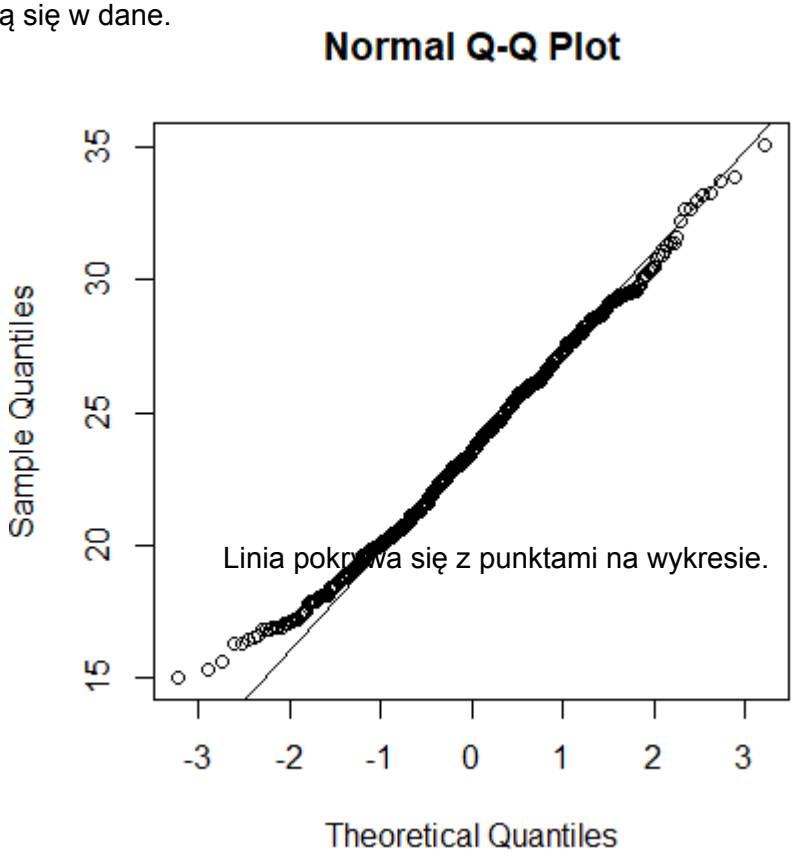
Korzystam z rozkładu X^2 w celu stworzenia przedziałów ufności.
Tworzę funkcję która będzie obliczać przedziały ufności dla wariancji w zależności od parametru alfa (1 - alfa -> poziom ufności).

```
varRange <- function(lvl) {
  alfa <- 1 - lvl
  return (round(sqrt(Psd*n/qchisq(c(1-alfa/2,alfa/2), n-1)),2))
}
```

wynik dla 90%: 1.80 1.96

i -> list wieku zawodników w którym osiągali najlepsze wyniki w życiu

Tworzę normalny graf kwantyli dla listy i (qqplot i qqline) dorysowuję prostą wpasowującą się w dane.



- Stawiam Hipotezę H0 : rozkład jest rozkładem normalnym.

Hipotezę będę sprawdzał stosując test Shapiro-Wilk'a (korzystam z funkcji bibliotecznej)

```
> shapiro.test(i)

        Shapiro-Wilk normality test

data:  i
W = 0.99418, p-value = 0.004097
```

Ze względu na to że p-value jest mniejsze od 0.05 odrzucam hipotezę H0.

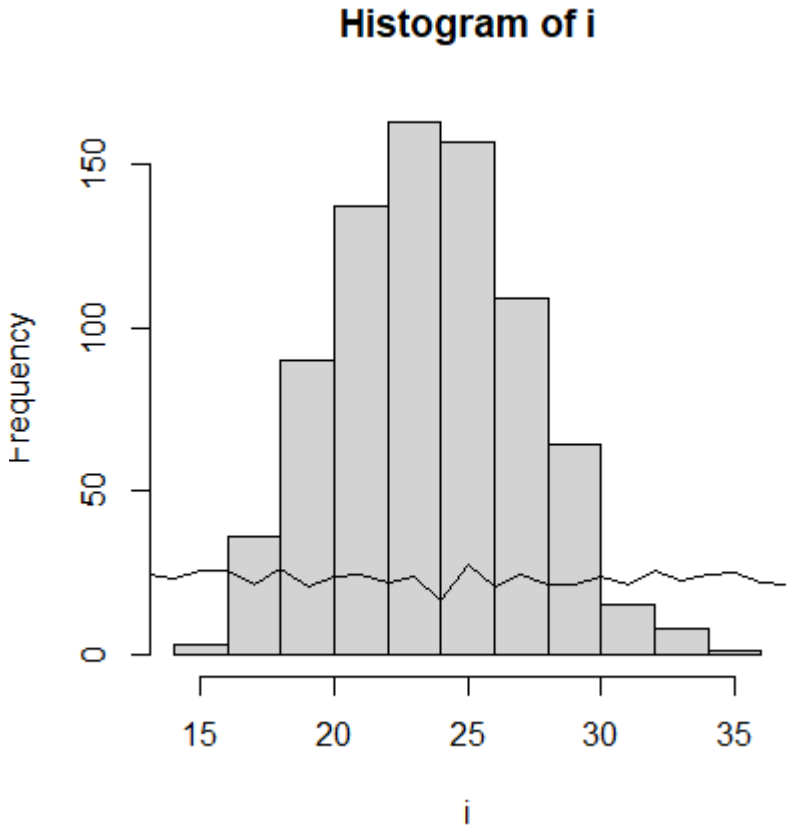
- Stawiam Hipotezę H1: rozkład jest rozkładem t Studenta.

Wykorzystam funkcję biblioteczną t.test(i).

```
data:  i
t = 188.55, df = 782, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 23.41080 23.90338
sample estimates:
mean of x
 23.65709
```

Ponownie p-value jest bardzo małe dlatego tę hipotezę również muszę odrzucić.

Ponownie tworzę histogram lat przy czym tym razem zaokrąglam wiek do całości



4. Wnioski

- Pomimo, że wykresy przypominały rozkład normalny testy pokazały że niem nie jest.
- Zawodnicy osiągali najlepsze wyniki będąc pomiędzy 20 a 30 lat.
- Można założyć że ciało ludzkie osiąga szczyt wydajności około 24 czwartego roku życia (średnia).