Wydział: WIMiIP	**Piotr Wilkosz** 404121	Gr. Proj. 2
Kierunek: Inżynieria Obliczeniowa Rok II	Temat: **Analiza statystyczna rankingu uniwersytetów.**	Data: 07.06.2021

1. Cel projektu

Za cel analizy statystycznej obrano zbadanie wpływu jakości badań naukowych na jakość kształcenia na uniwersytecie.

2. Opis zbioru danych

Zbiór danych powstał na bazie światowego rankingu uniwersytetów „The World Universities Rankings". Ranking ten jest corocznie przygotowywany przez brytyjskie czasopismo „Times Higher Education". Zbiór ten zawiera informacje o rankingu poszczególnych światowych uniwersytetów wraz z szczegółowymi statystykami dotyczącymi działalności uczelni uzyskanych w roku 2020. Zbiór opisuje **14** atrybutów, które opisują **1396** obserwacji, wśród których znajdziemy między innymi takie informacje jak globalna ranga uczelni, ocena jakości kształcenia czy ilość studentów studiujących na danym uniwersytecie.

Dane zawarte w tabeli:

Nazwa zmiennej	**Opis**
Score*Rank*	Ranking według kolumny Score*Result*.
University	Nazwa uniwersytetu.
Country	Kraj uczelni.
Number_students	Ilość aktywnych studentów.
Numb*students*per_Staff	Stosunek liczby studentów do kadry.
International_Students	Odsetek studentów zagranicznych.
Percentage_Female	Odsetek kobiet.
Percentage_Male	Odsetek mężczyzn.
Teaching	Ocena w nauczaniu.
Research	Ocena w badaniach.
Citations	Wskaźnik wpływu badań w rozpowszechnianiu nowej wiedzy i pomysłów.
Industry_Income	Wynik w dochodach branży przemysłowej.
International_Outlook	Wynik międzynarodowych perspektw.
Score*Result*	Wynik podsumowujący.

Wynik zawarty w zmiennej **ScoreResult** jest wynikową pozostałych wartości i pod uwagę bierze w 30% ocenę w nauczaniu, w 30% ocenę w badaniach, w 30% wynik wpływu badań na społeczeństwo, w 7.5% wynik międzynarodowych perspektyw oraz w 2.5% % wynik dochodów branży przemysłowej.

Do przeprowadzenia szczegółowej analizy statystycznej wytypowałem trzy zmienne:

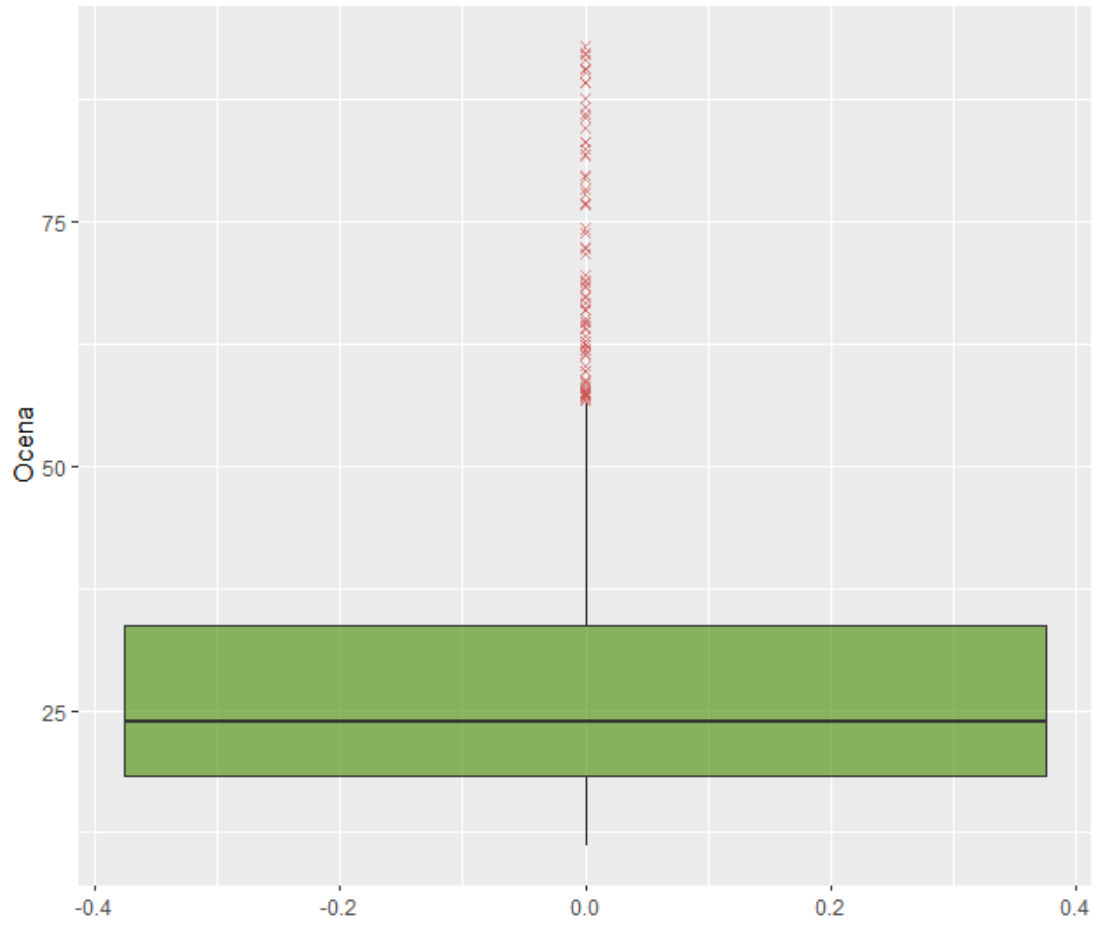
a. **Teaching** – zmienna ilościowa, wyrażająca ocenę kształcenia na uniwersytecie. Ocena ta składa się kilku składowych którymi są:
   i. ankieta dot. reputacji (nauczanie)
   ii. liczba tytułów doktora przypadających na nauczyciela akademickiego
   iii. liczba studentów przypadających na nauczyciela akademickiego
   iv. liczba nagród doktorskich/licencjackich
   v. przychód na nauczyciela akademickiego
b. **Research** – zmienna ilościowa wyrażająca ocenę prowadzonych badań uniwersyteckich. Ocena ta składa się kilku składowych którymi są:
   i. ankieta dot. reputacji uczelni (badanie)
   ii. zysk z prac badawczych (skalowany)
   iii. artykuły na pracownika naukowego i badawczego
   iv. dochód z badań publicznych / całkowity dochód z badań
c. **Country** – zmienna jakościowa określająca kraj, w którym znajduje się wybrany uniwersytet.

3. Statystyka opisowa

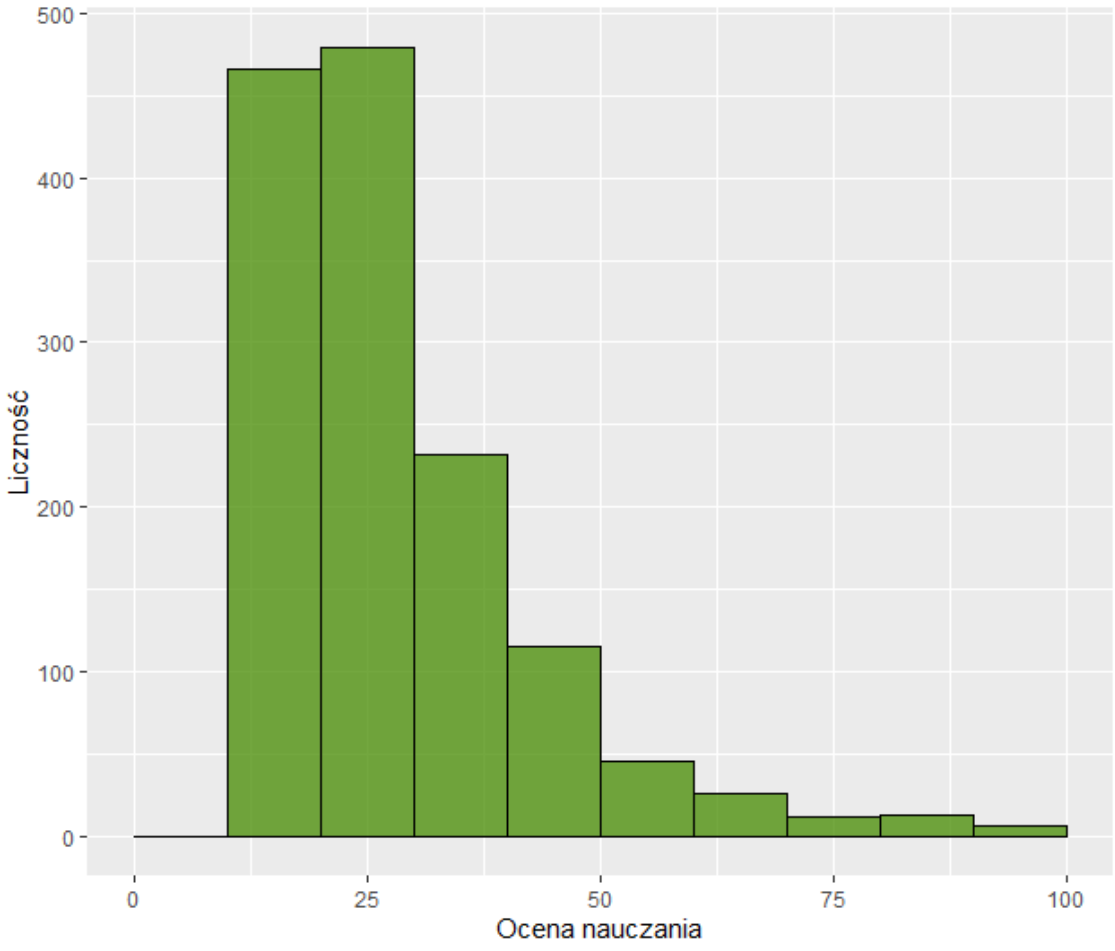
	Teaching	**Research**
Minimum	11.20	6.80
Maximum	92.80	99.60
Miary tendencji centralnej		
Średnia	28.23	23.98
Mediana	23.80	18.00
Moda	16.70	10
Wskaźniki rozproszenia		
Skośność	1.894433	1.814843
Odchylenie standardowe	14.14955	17.53704
Wariancja	200.2098	307.5479
Dolny kwartyl	33.60	11.60
Górny kwartyl	18.30	30.10
Rozstęp	81.6	92.8
Kurtoza	7.274186	6.494914

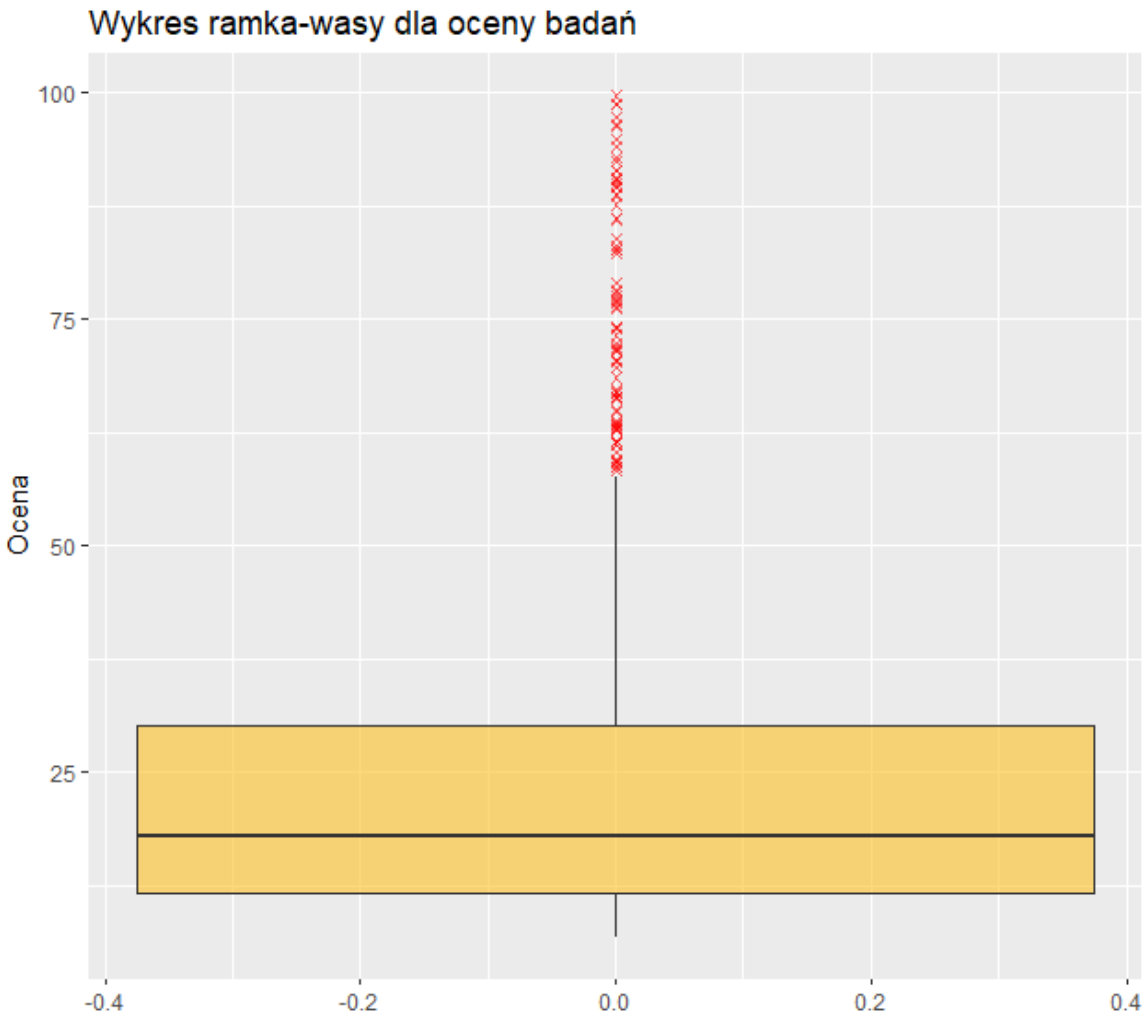
4. Graficzna prezentacja danych

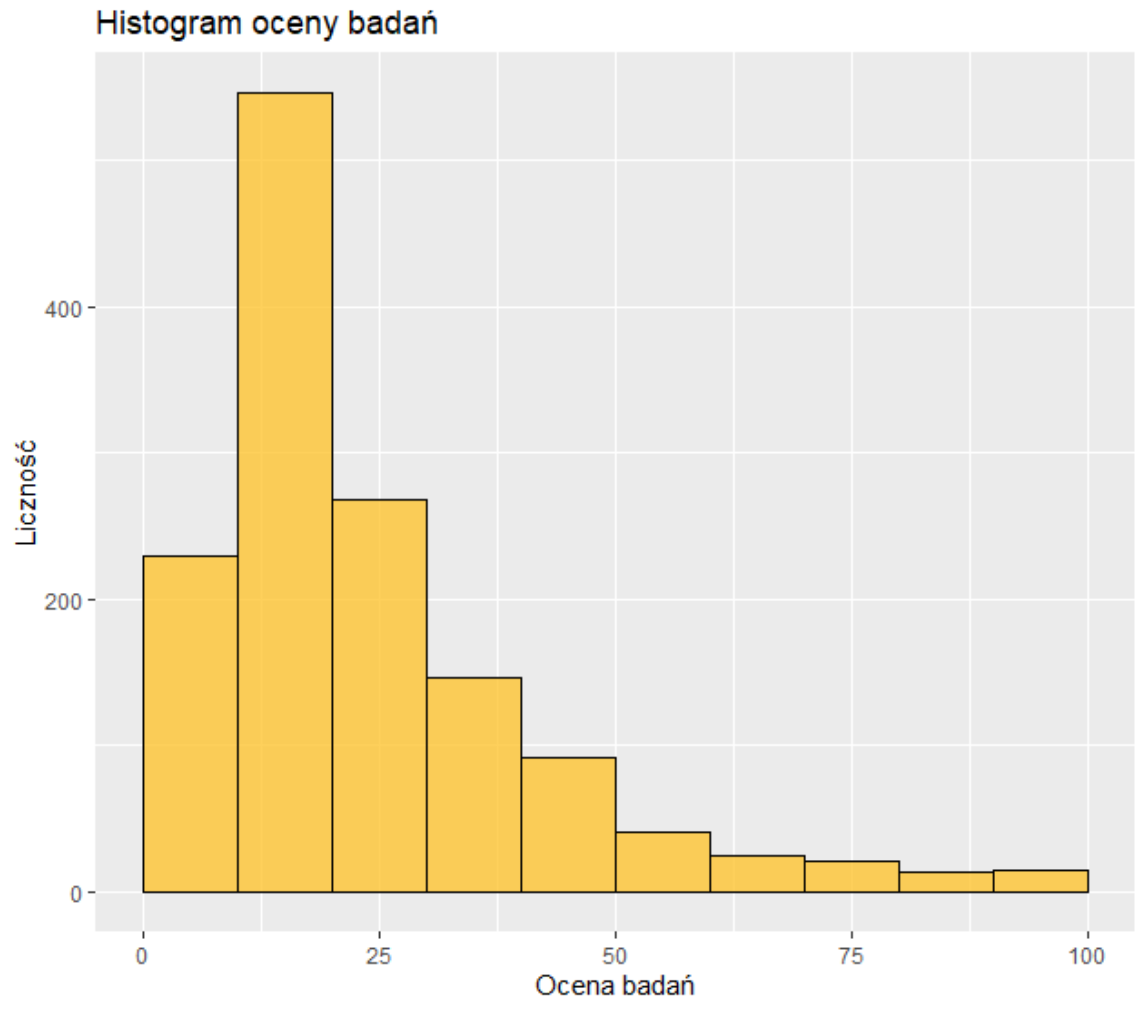
Wykres ramka-wasy dla oceny nauczania

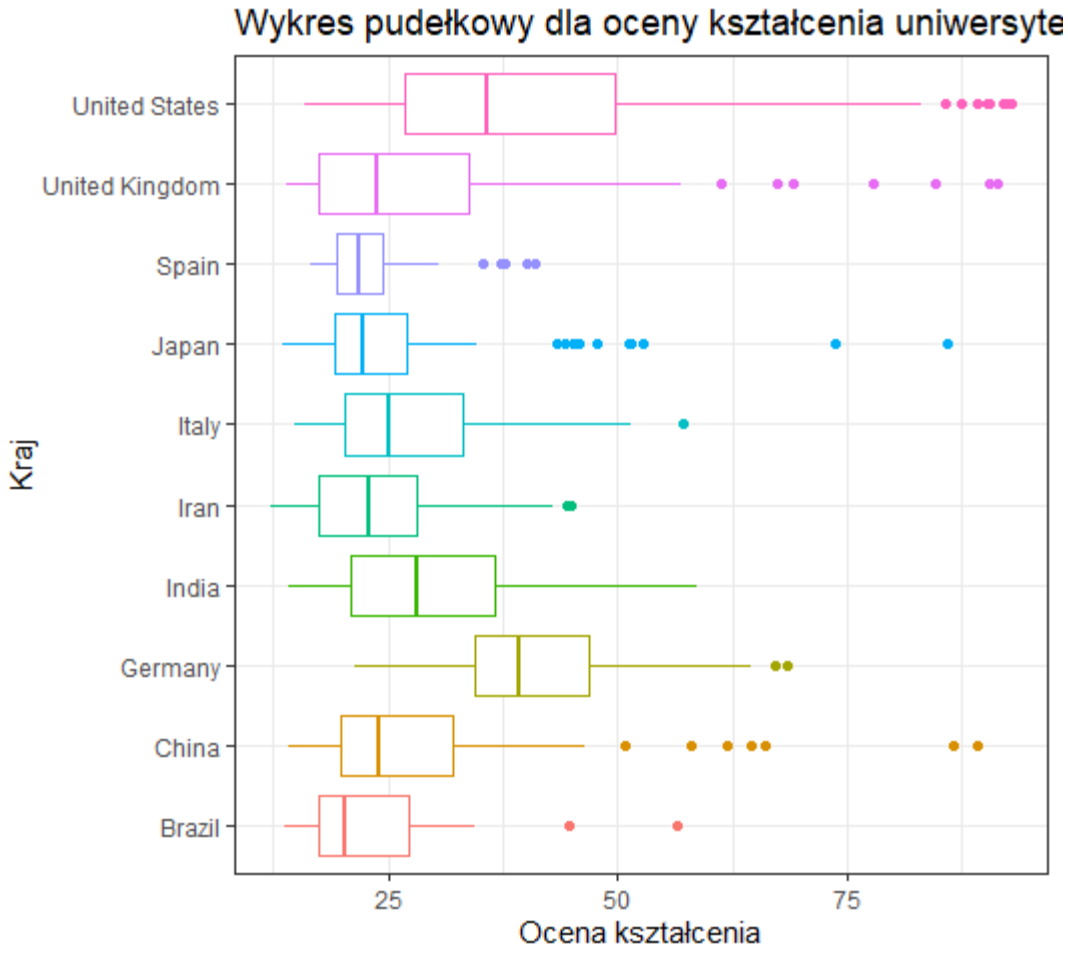


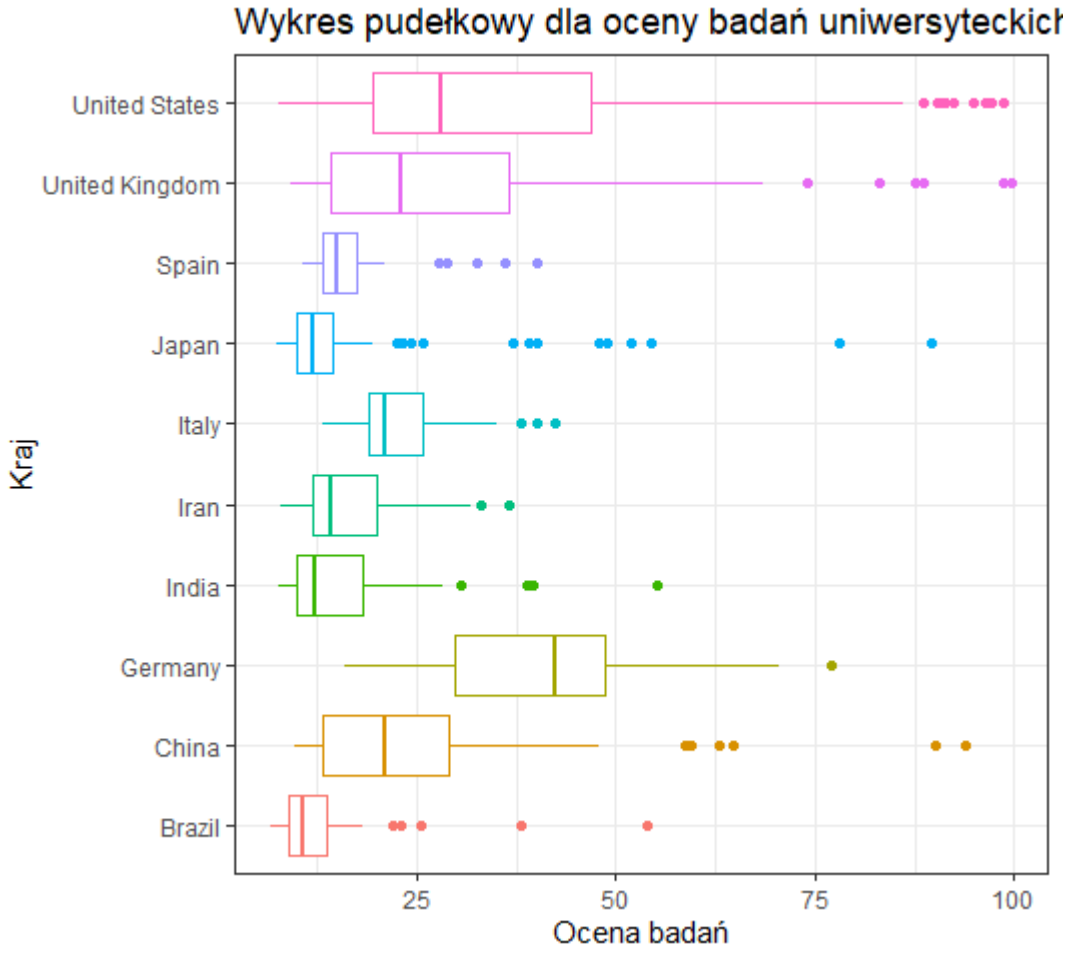
Histogram oceny nauczania




Wykres ramka-wasy dla oceny badań
Ocena
100
75
50
25
-0.4
-0.2
0.0
0.2
0.4


Histogram oceny badań
Liczność
400
200
0
0
25
50
75
100
Ocena badań


Wykres pudełkowy dla oceny kształcenia uniwersyte
United States
United Kingdom
Spain
Japan
Italy
Iran
India
Germany
China
Brazil
Kraj
25
50
75
Ocena kształcenia


Wykres pudełkowy dla oceny badań uniwersyteckich
United States
United Kingdom
Spain
Japan
Italy
Iran
India
Germany
China
Brazil
Kraj
25
50
75
100
Ocena badań

5. Rozkład normalny

Wynik testu Shapiro-Wilk'a dla zmiennych:

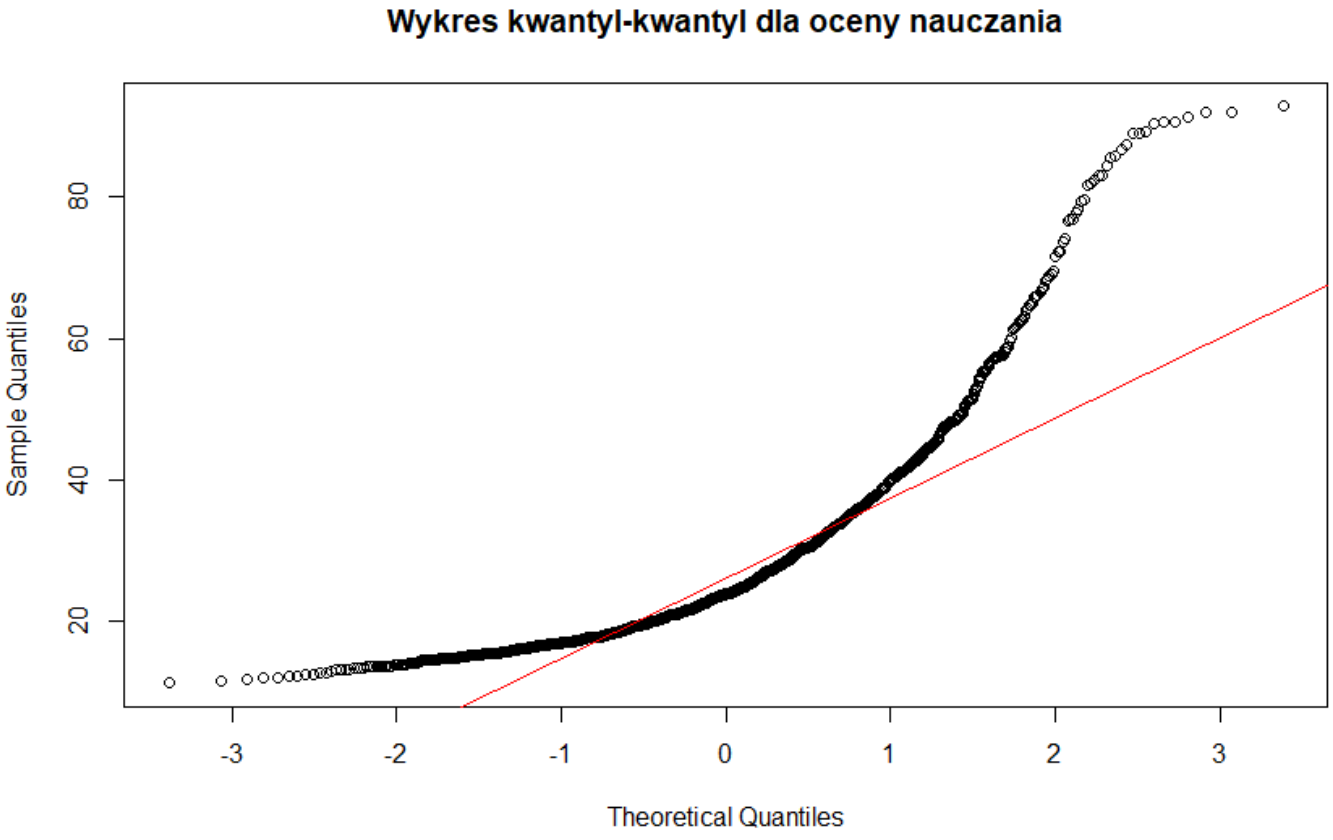
```
        Shapiro-Wilk normality test

data:  data$Teaching
W = 0.81575, p-value < 2.2e-16
```

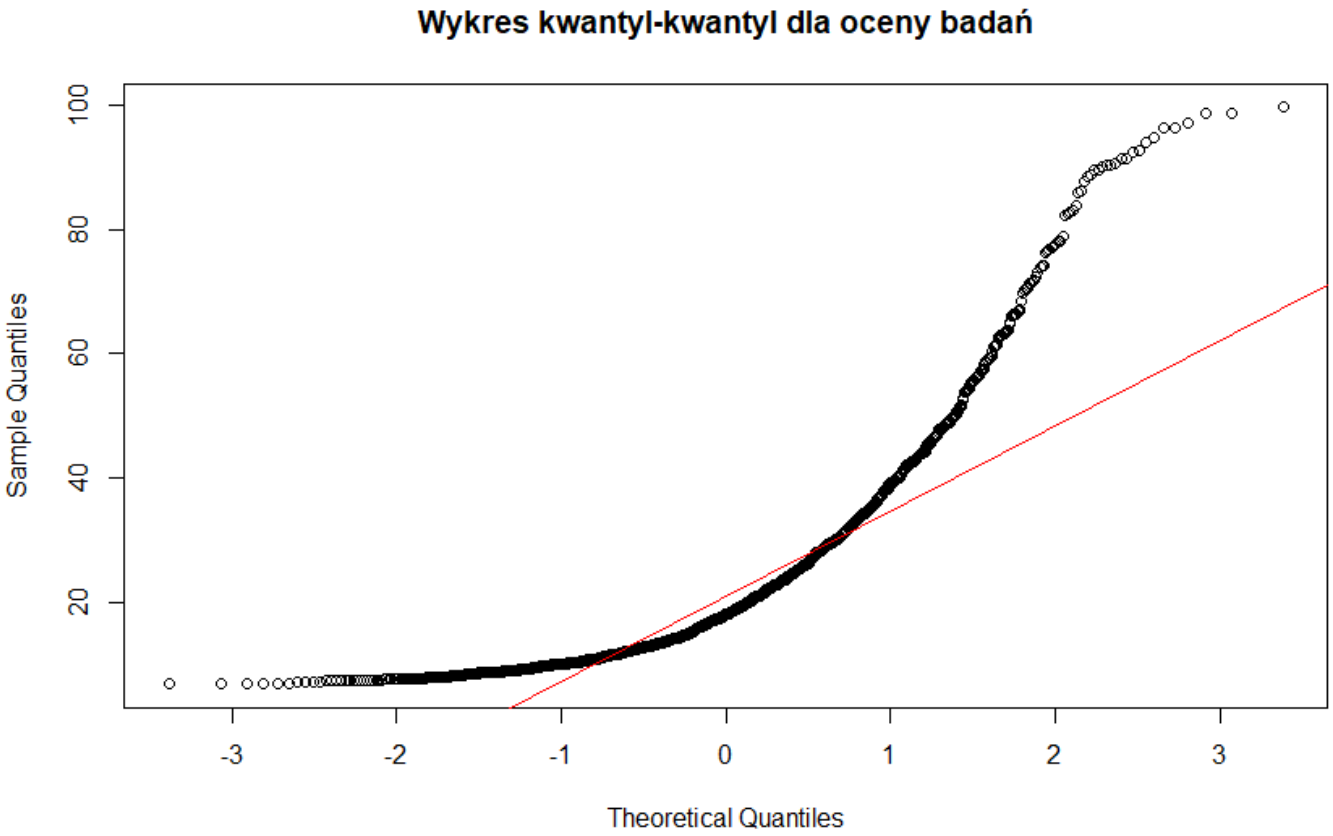
```
        Shapiro-Wilk normality test

data:  data$Research
W = 0.80301, p-value < 2.2e-16
```

i. **Teaching:** p-value < 2.2 e-16
ii. **Research:** p-value <2.2 e-16

H0: m=0 – reszty mają rozkład normalny
H1: m≠0 – reszty nie mają rozkładu normalnego

W obu przypadkach **p-value** < 0.05, a więc odrzucam hipotezę **H0** i przyjmuję **H1**. Rozkład nie jest normalny. W celu potwierdzenia wyniku przeprowadzam i analizuję wykresy kwantyl-kwantyl.





W obu przypadkach wartości znacznie odstają od krzywej teoretycznej. Wartości nie stanowią rozkładu normalnego.

6. Przedziały ufności

Przy **95%** pewności ustalam, że

a. Średnia ocena nauczania na uczelni znajduje się w przedziale od **27.49** do **28.97**.
b. Średnia ocena badań na uczelni znajduje się w przedziale od **23.06** do **24.90**.

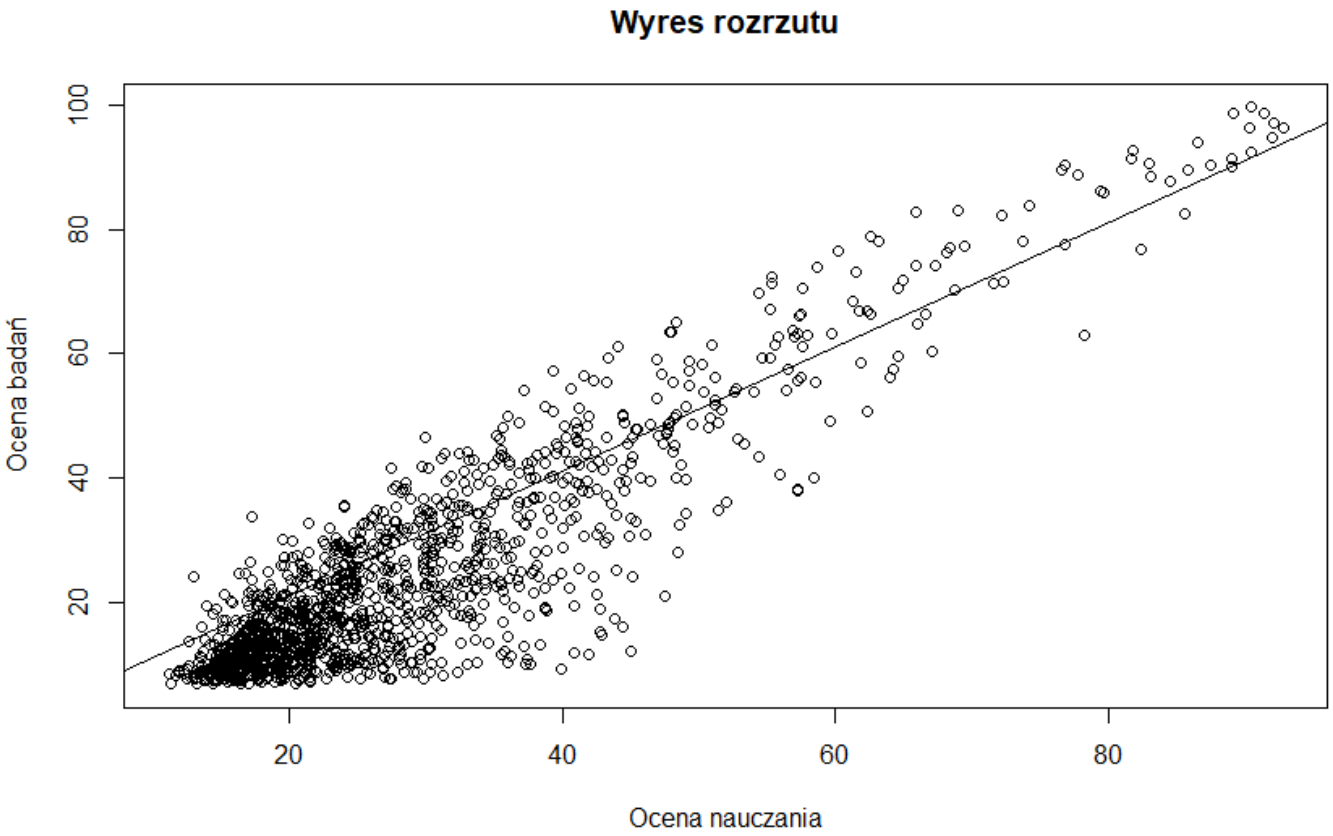
```
95 percent confidence interval:
 27.48619 28.97197
```

```
95 percent confidence interval:
 23.06042 24.90190
```

Oznacza to, że średnio na **20** przypadków **jedna uczelnia** znajdzie się poza wskazanym przedziałem.

7. Korelacja pomiędzy typowanymi zmiennymi

Do modelu regresji liniowej przyjmuję jakość nauczania jako zmienną objaśnianą oraz jakość badań jako zmienną objaśniającą.



H0: Współczynnik korelacji równy 0 w zbiorowości generalnej.

H1: Współczynnik korelacji różny od 0.

```
        Pearson's product-moment correlation

data:  data$Teaching and data$Research
t = 77.4, df = 1394, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8902906 0.9101423
sample estimates:
      cor
0.9006855
```

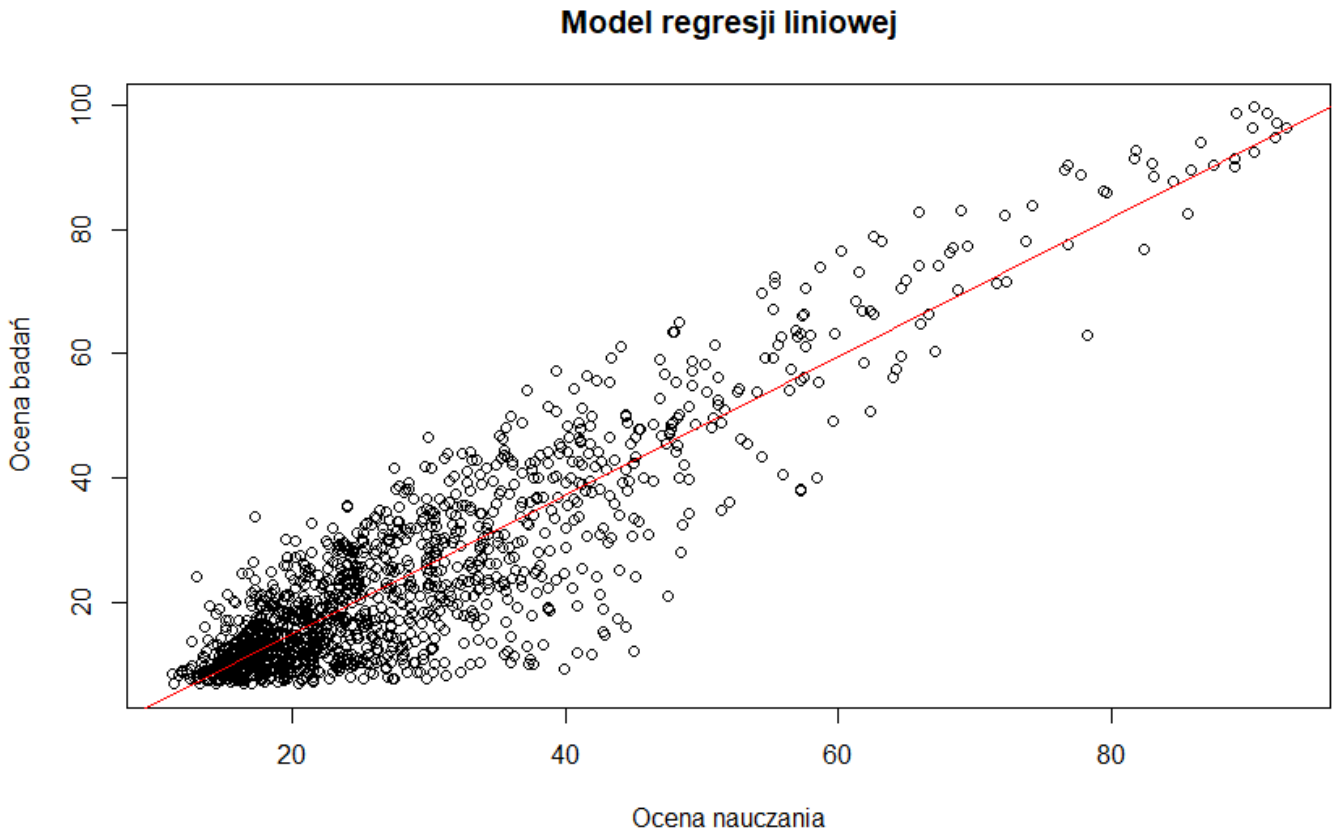
W pierwszym kroku weryfikuję jakość dopasowania poprzez test korelacji Pearson'a. Wynik testu wykazał wartość **p-value** poniżej **2.2 * 10^-16**. Wartość ta jest mniejsza od **α** wynoszącej **0.05**, więc nie jest spełniona hipoteza **H0**. Przyjmuję hipotezę **H1**. Korelacja między zmiennymi nie jest równa **0**. Współczynnik korelacji wynosi **0.90**. Zauważyć można stały przyrost drugiej zmiennej, można przypuszczać, że między zmiennymi istnieje zależność liniowa.

```
Call:
lm(formula = data$Teaching ~ data$Research)

Coefficients:
  (Intercept)  data$Research
      10.8018         0.7267
```

W związku posługując się metodą najmniejszych kwadratów dopasowano prostą opisaną poniższym równaniem:

$$\mathbf{y} = 0.726707\mathbf{x} + 10.801803$$

$$\mathbf{Teaching} = 0.726707\mathbf{Reseach} + 10.801803$$



Test F Fishera-Snedecora

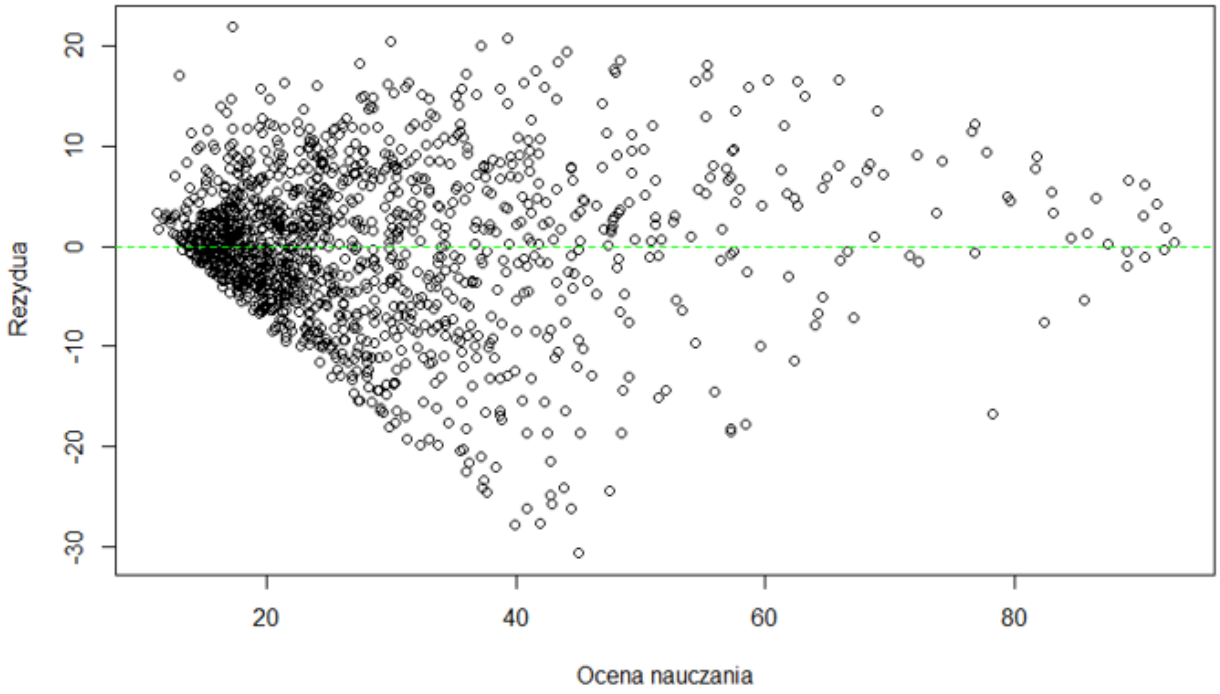
H0: m=0 – brak liniowej zależności pomiędzy zmiennymi.
H1: m≠0 – istnieje liniowa zależność pomiędzy zmiennymi.

```
        F test to compare two variances

data:  data$Teaching and data$Research
F = 0.65099, num df = 1395, denom df = 1395, p-value = 1.41e-15
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5861028 0.7230548
sample estimates:
ratio of variances
         0.6509873
```

p-value < α, więc odrzucam **H0** i uznaję **H1**.

8. Analiza reszt modelu regresji

a. Losowość odchyleń



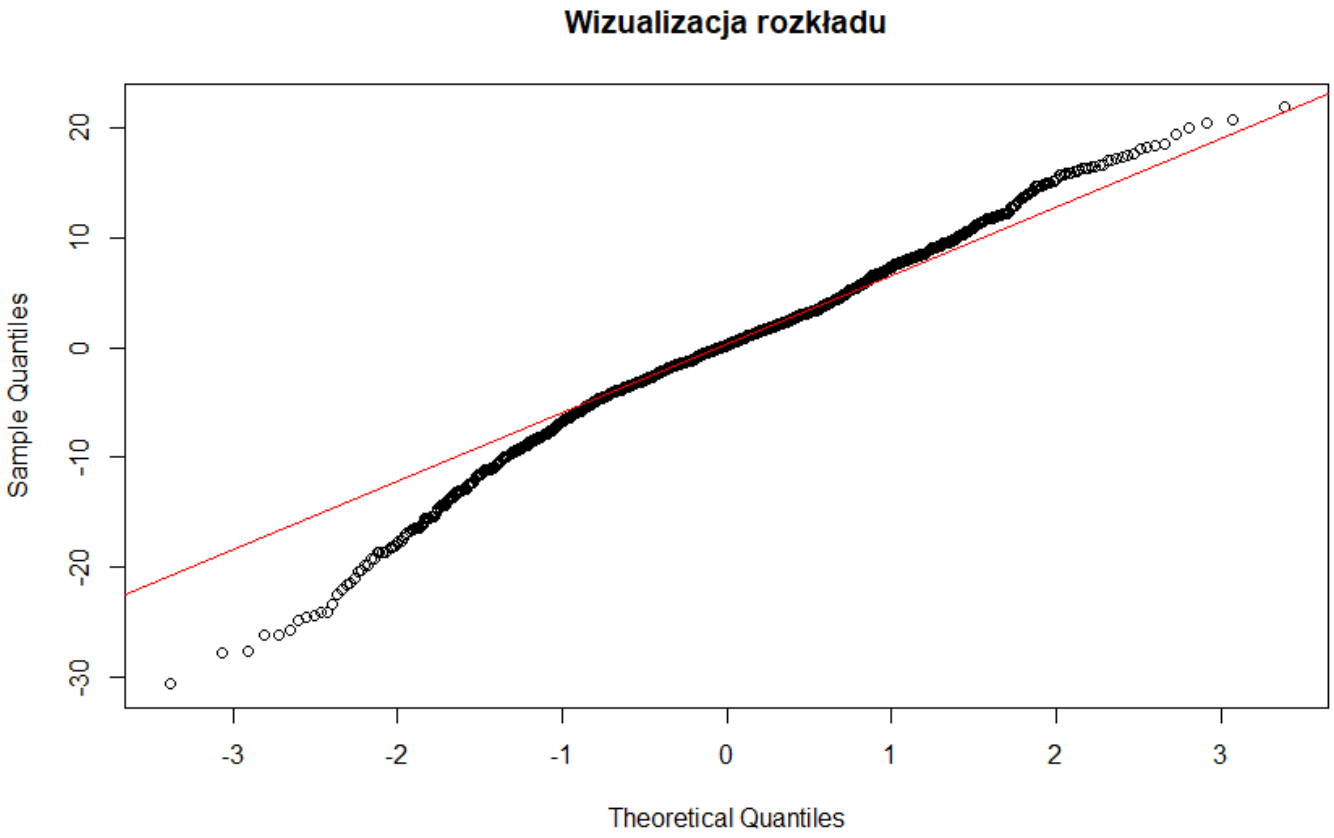
Zmienne układają losowo się poniżej i powyżej krzywej teoretycznej.

b. Rozkład normalny

H0: m=0 – reszty mają rozkład normalny.
H1: m≠0 – reszty nie mają rozkładu normalnego.

```
        Shapiro-Wilk normality test

data:  model1$residuals
W = 0.98525, p-value = 1.01e-10
```

Wynik testu Shapiro-Wilk'a na rezyduach pokazał, że **p-value** = **1.01e-10**. Wartość ta jest niższa od współczynnika **α**, przez co odrzucam hipotezę **H0**. Przyjmuję **H1**, więc rozkład nie jest normalnym co potwierdza poniższy test kwantyl-kwantyl.



c. Nieobciążalność reszt

```
> mean(model1$residuals)
[1] -1.145128e-15
```

Średnia z rezyduum wynosi **-1.145128e-15**. Wartość ta jest bardzo bliska zeru. Oznacza to, że reszty nie są obciążone.

d. Homoscedastyczność

H0 - występuje stale rozproszenie reszt.
H1 - występuje heteroscedastycznosc.

```
        studentized Breusch-Pagan test

data:  model1
BP = 95.585, df = 1, p-value < 2.2e-16
```

Przeprowadzam test Breuscha-Pagana, że **p-value** wynosi poniżej **2.2e-16**. Wartość ta jest niższa od **α,** więc reszty nie mają stałego rozproszenia. Oznacz to, że zbiór jest heteroscedastyczny.

9. Interpretacja modelu regresji

```
Residual standard error: 7.622 on 1394 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.8111
```

Weryfikacja modelu wykazała, że korelacja jest istotna statystycznie i jest ona korelacją bardzo wysoką. Współczynnik korelacji jest dodatni, więc wraz z wzrostem oceny badań wzrasta również ocena nauczania. Współczynnik determinacji $\mathbf{r^2}$ = **0.8111**, a więc **81%** zmienności oceny badań na ocenę nauczania wyjaśniono przez model regresji liniowej.

10. Wnioski

Z statystyki opisowej wybranych zmiennych można wywnioskować, że dane są mało rozproszone od średniej. Dla obu zmiennych możemy zauważyć stosunkowo wysoką wartość skośności. W związku z tym również większe jest odchylenie standardowe i wariancja. Wskazuje to na spore rozrzucenie wielkości wokół średniej. Dla zmiennych zachodzi również stosunkowo wysoka różnica pomiędzy średnią a medianą. W obu przypadkach średnia jest większa od mediany, co wraz z wysoką, dodatnią skośnością mówi o rozkładzie prawostronnym zbiorów.

W przypadku histogramu zarówno dla wartości punktowej jakości badań uniwersyteckich jak i jakości kształcenia widoczny jest rozkład prawostronnie skośny. Mówi to o spadku ilości uniwersytetów wraz z wzrastającą wartością oceny – badań i nauczania. Wraz z medianą pokazuje to obecną, w której według rankingu największa ilość uniwersytetów prezentuje niską jakość zarówno badań jak i kształcenia. Oba wykresy nie są symetryczne.

Wykres ramka wąsy utwierdza wywnioskowaną spadek ilości ocen. W przypadku wykresu ramka wąsy dla obu danych ilościowych widzimy znaczną ilość wartości odstających nad górnym wykresem. Dane są bardzo rozproszone, co można było również zauważyć na podstawie mediany, mody, wartości minimalnych i maksymalnych. Wartości odstające i ekstremalne odpowiedzialne są za wybitne uniwersytety osiągające wyjątkowe wyniki, dlatego są traktowane jak poprawne i nie zostały wykluczone. W obu przypadkach mediana znajduje się w dolnej części pudełka.

Podsumowując uzyskane wyniki pozwalają twierdzić, iż poziom badań na uniwersytecie ma znaczący wpływ na poziom nauczania uniwersyteckiego.