|  |  |  |
| --- | --- | --- |
| Wydział:  WIMiIP | **Piotr Wilkosz**  404121 | Gr. Proj.  2 |
| Kierunek:  Inżynieria  Obliczeniowa  Rok II | Temat:  **Analiza statystyczna rankingu uniwersytetów.** | Data  04.06.2021 |

1. **Cel projektu**

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

1. **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. 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** |
| ScoreRank | Ranking według kolumny ScoreResult. |
| University | Nazwa uniwersytetu. |
| Country | Kraj uczelni. |
| Number\_students | Ilość aktywnych studentów. |
| Numbstudentsper\_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 | Ocena wpływu z badań |
| Industry\_Income | Wynik w dochodach branży przemysłowej. |
| International\_Outlook | Wynik w światowym światopoglądzie |
| ScoreResult | Wynik podsumowujący |

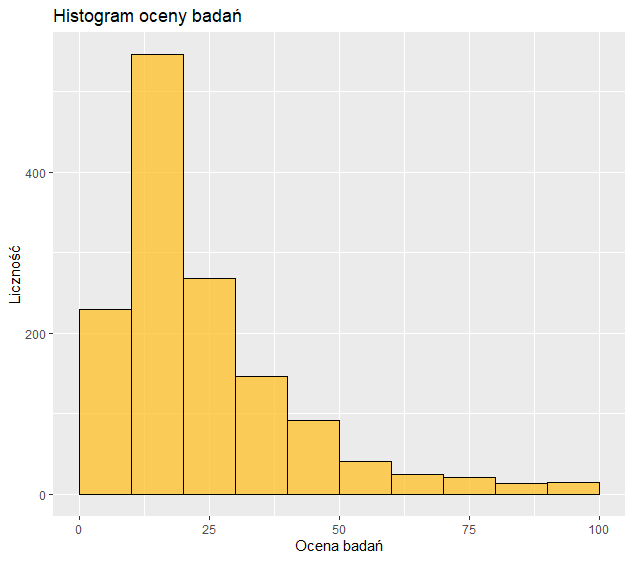
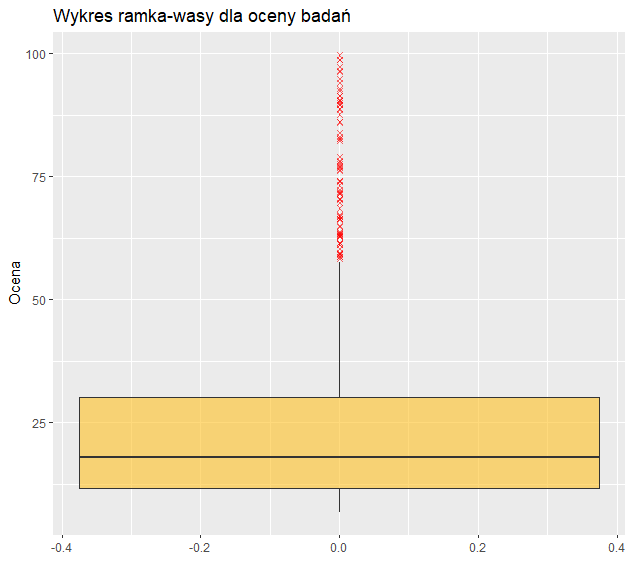
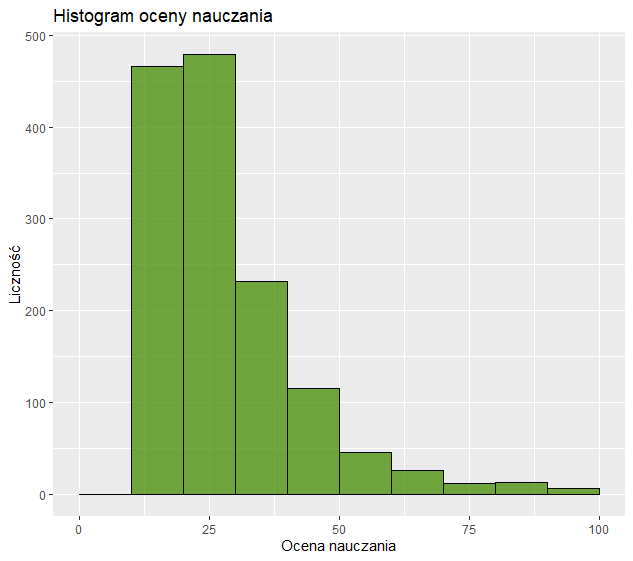
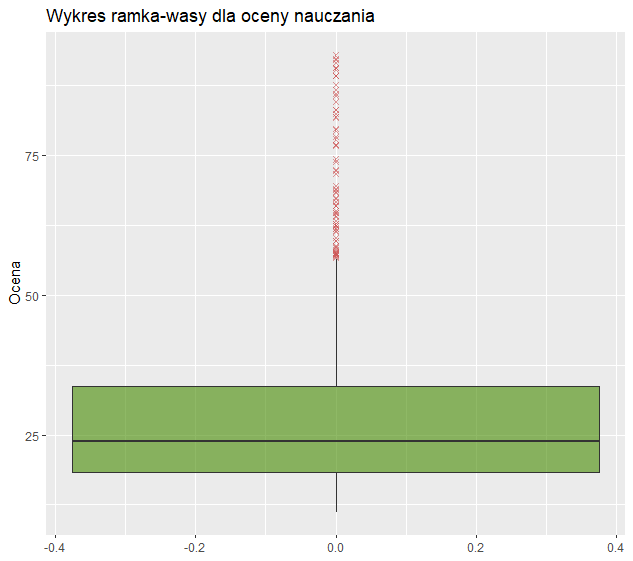
Wynik zawarty w zmiennej ScoreResult jest wynikową pozostałych wartości i pod uwagę bierze w 30% ocenę w nauczaniu, w 30% ocenę w badaniach, w 30% wynik wpływu z badań, w 7.5% wynik światowego światopoglądu oraz w 2.5% % wynik dochodów branży przemysłowej.

Do przeprowadzenia szczegółowej analizy statystycznej wytypowałem trzy zmienne:

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

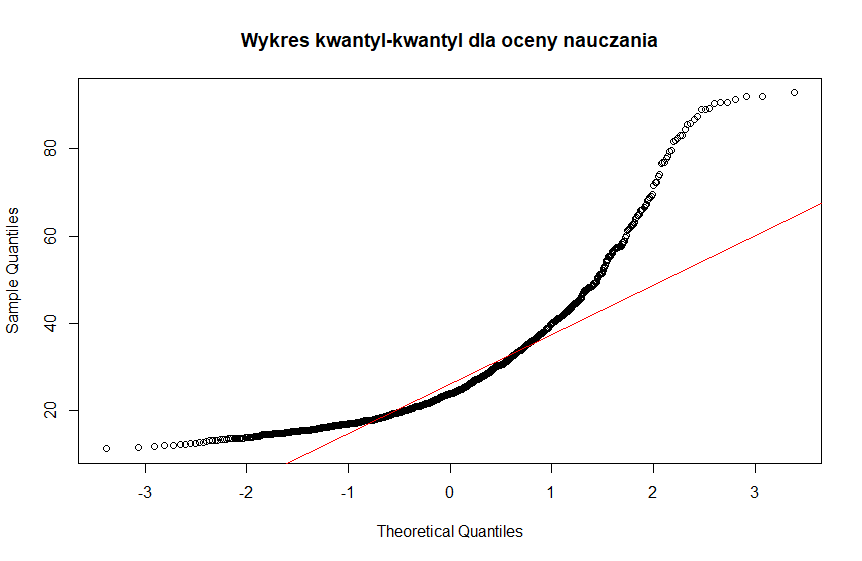
1. **Statystyka opisowa**

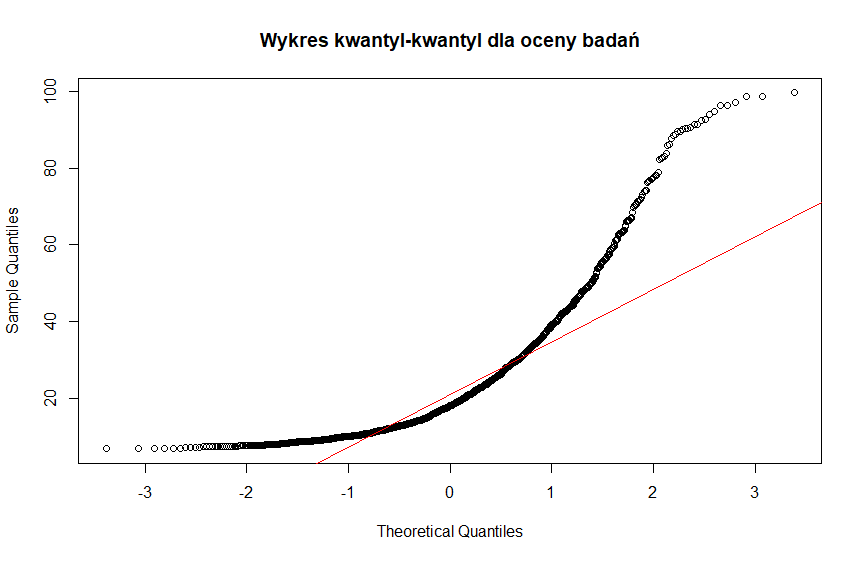
|  |  |  |
| --- | --- | --- |
|  | **Teaching** | **Research** |
| Minimum | 11.20 | 6.80 |
| Maximum | 92.80 | 99.60 |
| **Miary tendencji centralnej** | | |
| Średnia | 28.23 | 23.98 |
| Mediana | 23.80 | 18.00 |
| Moda | 16.70 | 10 |
| **Wskaźniki rozproszenia** | | |
| Skośność | 1.894433 | 1.814843 |
| 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 |

1. **Graficzna prezentacja danych**
2. **Rozkład normalny**Wynik testu Shapiro-Wilk’a dla zmiennych:  
   * 1. **Teaching:** p-value < 2.2 e-16
     2. **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 wykres Q-Q.

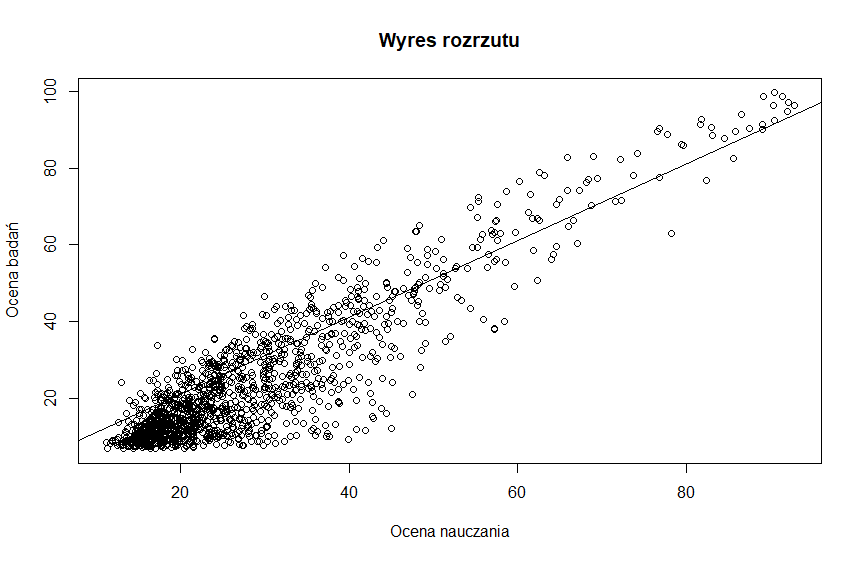




W obu przypadkach wartości znacznie odstają od krzywej teoretycznej. Wartości nie stanowią rozkładu normalnego.

1. **Przedziały ufności**Przy 95% pewności ustalam, że
   1. Średnia ocena nauczania na uczelni znajduje się w przedziale od 27.48619 do 28.97197
   2. Średnia ocena badań na uczelni znajduje się w przedziale od 23.06042 do 24.90190

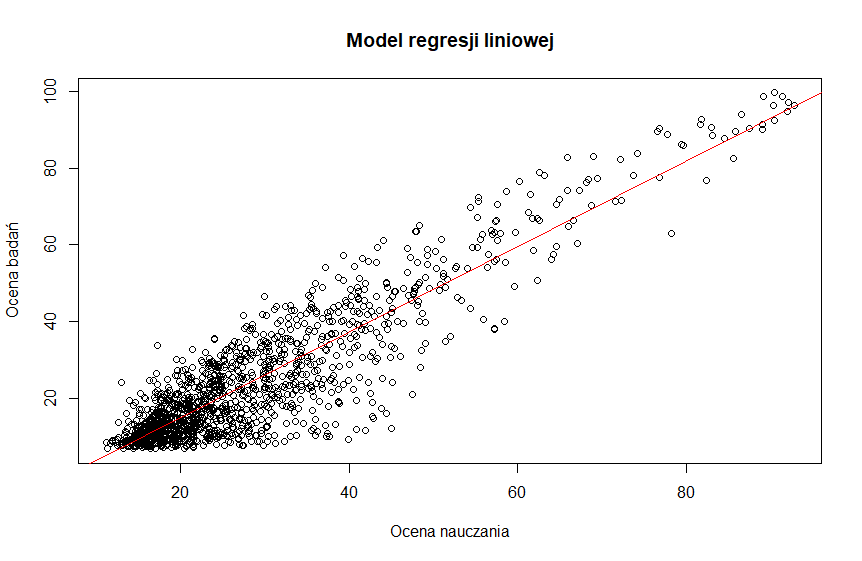
Oznacza to, że średnio na 20 przypadków jedna uczelnia znajdzie się poza wskazanym przedziałem.

1. **Korelacja pomiędzy typowanymi zmiennymi**  
     
   Do modelu regresji liniowej przyjmuję jakoś nauczania jako zmienną objaśnianą oraz jakość badań jako zmienną objaśniającą.  
     
   H0: Współczynnik korelacji równy 0 w zbiorowości generalnej  
   H1: Współczynnik korelacji różny od 0(SIGMA XD)  
     
   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 ALFA wynoszącej 0.05, więc nie jest spełniona hipoteza H0. Przyjmuję hipotezę H1. Korelacja między zmiennymi nie jest równa 0. Współczynnik korelacji wynosi 0.9006855. Zauważyć można stały przyrost drugiej zmiennej, można przypuszczać, że między zmiennymi istnieje zależność liniowa.

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

**y = 0.726707x + 10.801803**

**Teaching = 0.726707 \* Reseach + 10.801803**

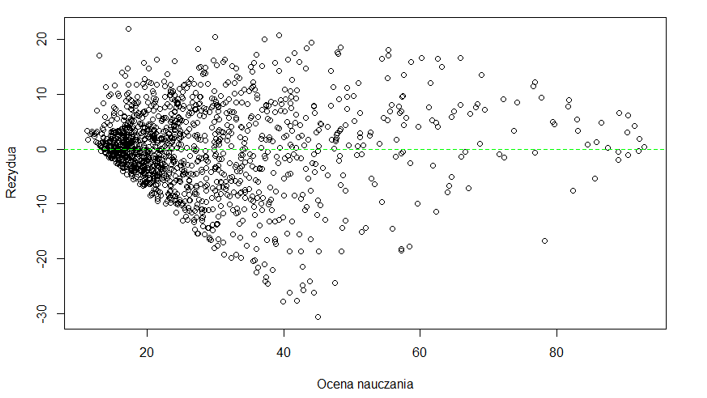


Test F Fishera-Snedecora

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

P < ALFA => odrzucam H0 i uznaję H1.

1. **Analiza reszt modelu regresji**
   1. **Losowość odchyleń**

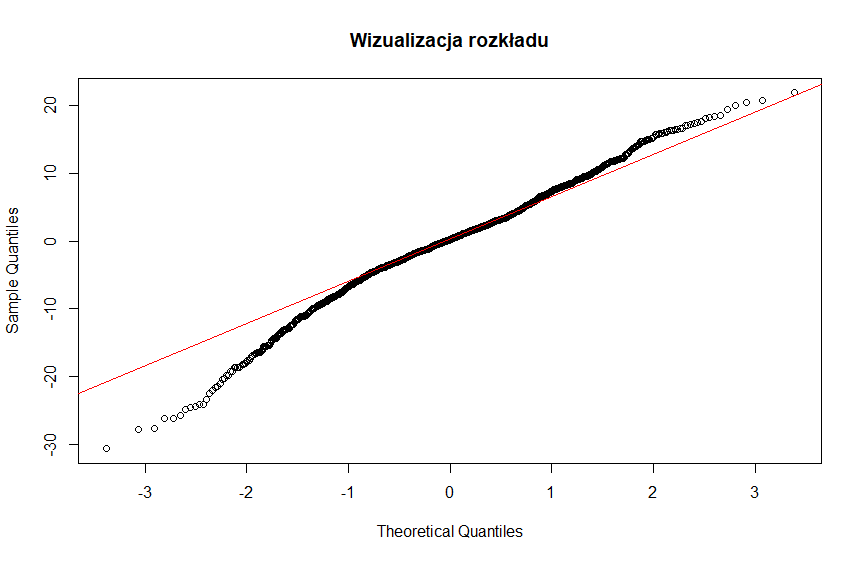


Zmienne układają losowo się poniżej i powyżej krzywej teoretycznej. Reszty nie są obciążone.

* 1. **Rozkład normalny**

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

Wynik testu Shapiro-Wilk’a na rezyduach pokazał, że p-value = 2.2e-16. Wartość ta jest niższa od współczynnika ALFA, przez co odrzucam hipotezę H0. Przyjmuję H1, więc rozkład nie jest normalnym co potwierdza poniższy test Q-Q.



* 1. **Nieobciążalność reszt**

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

* 1. **Homoscedastyczność**

H0 - występuje stale rozproszenie reszt

H1 - występuje heteroscedastycznosc

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

1. **Analiza reszt modelu regresji**

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 r^2 = 0.8111, a więc 81% zmienności oceny badań na ocenę nauczania wyjaśniono przez model regresji liniowej.

1. **Wnioski**

Z uzyskanych wyników statystyki opisowej oceny badań

#To USE

The first isn't easy to evaluate precisely, but there are approximately 20000 Universities, worldwide. Of those, about 7000 are in the USA. Now that varies between huge State universities and much smaller institutions.

Liczba obserwacji zależy od dostępności danych w danym kraju. We have good access to state supplied data from the UK which is why w we have so many UK universities.

Karsltand Univeristy sa międzynarodowi, National University of the South

Też argentyna