Data mining

Projekt Indywidualny

Wykonał:

Oleg Łyżwiński 305158

Warszawa 2024

Zad 1

Zbór danych Cars93 zawiera 93 wiersze i 27 kolumny. Informacje zawarte w kolumnach wykorzystywanych w tym zadaniu opisano poniżej:

- Min.Price to minimalna cena w 1000 $.
- MPG.city to spalanie w MPG(mil na US galon z rankingu EPA) w mieście.
- MPG.highway to spalanie w MPG na autostradzie.
- Weight to waga w funach.
- Origin to kraj produkcji określony jako non-USA or USA.
- Type wskaźnik z poziomami "Small", "Sporty", "Compact", "Midsize", "Large" and "Van".

Poniżej przedstawiono tabelę zawierającą wyżej opisane dane:

	Min.Price	MPG.city	MPG.highway	Weight	Origin	Type
1	12.9	25	31	2705	non-USA	Small
2	29.2	18	25	3560	non-USA	Midsize
3	25.9	20	26	3375	non-USA	Compact
4	30.8	19	26	3405	non-USA	Midsize
5	23.7	22	30	3640	non-USA	Midsize
6	14.2	22	31	2880	USA	Midsize
7	19.9	19	28	3470	USA	Large
8	22.6	16	25	4105	USA	Large
9	26.3	19	27	3495	USA	Midsize
10	33.0	16	25	3620	USA	Large
11	37.5	16	25	3935	USA	Midsize
12	8.5	25	36	2490	USA	Compact
13	11.4	25	34	2785	USA	Compact
14	13.4	19	28	3240	USA	Sporty
15	13.4	21	29	3195	USA	Midsize
16	14.7	18	23	3715	USA	Van
17	14.7	15	20	4025	USA	Van
18	18.0	17	26	3910	USA	Large
19	34.6	17	25	3380	USA	Sporty
20	18.4	20	28	3515	USA	Large

Następnie utworzono nowe zmienne: zużycie paliwa w litrach na 100 km w mieście i na autostradzie, waga samochodu w kg oraz cena wersji podstawowej samochodu w tys. PLN:

```
Cars93$Fuel_usage.city <- 3.8 / Cars93$MPG.city * 100 / 1.6
Cars93$Fuel_usage.highway <- 3.8 / Cars93$MPG.highway * 100 / 1.6
Cars93$Weight.kg <- Cars93$Weight * 0.4536
Cars93$Price.PLN <- Cars93$Price * 3.35
```

Poniżej przedstawiono wyznaczone dane:

Fuel_usage.city	Fuel_usage.highway	Weight.kg	Price.PLN
9.500000	7.661290	1226.988	53.265
13.194444	9.500000	1614.816	113.565
11.875000	9.134615	1530.900	97.485
12.500000	9.134615	1544.508	126.295
10.795455	7.916667	1651.104	100.500
10.795455	7.661290	1306.368	52.595
12.500000	8.482143	1573.992	69.680
14.843750	9.500000	1862.028	79.395
12.500000	8.796296	1585.332	88.105
14.843750	9.500000	1642.032	116.245
14.843750	9.500000	1784.916	134.335
9.500000	6.597222	1129.464	44.890
9.500000	6.985294	1263.276	38.190
12.500000	8.482143	1469.664	50.585

Podstawowe statystyki próbkowe dla danych opisujących cenę wersji podstawowej samochodu:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 24.79   40.87   59.30   65.36   78.06  207.37
```

Skośność i kurtoza wynosi:

kurtoza	3.0514181762633
skosnosc	1.48398211356445

W związku z niewystarczającą liczbą próbek nie wyznaczono skośności i kurtozy standaryzowanych. Jednak na podstawie uzyskanych danych możemy stwierdzić, że rozkład wartości Price.PLN ma przesunięty środek ciężkości w lewo(prawy ogon) oraz jest wyciągnięty do góry.

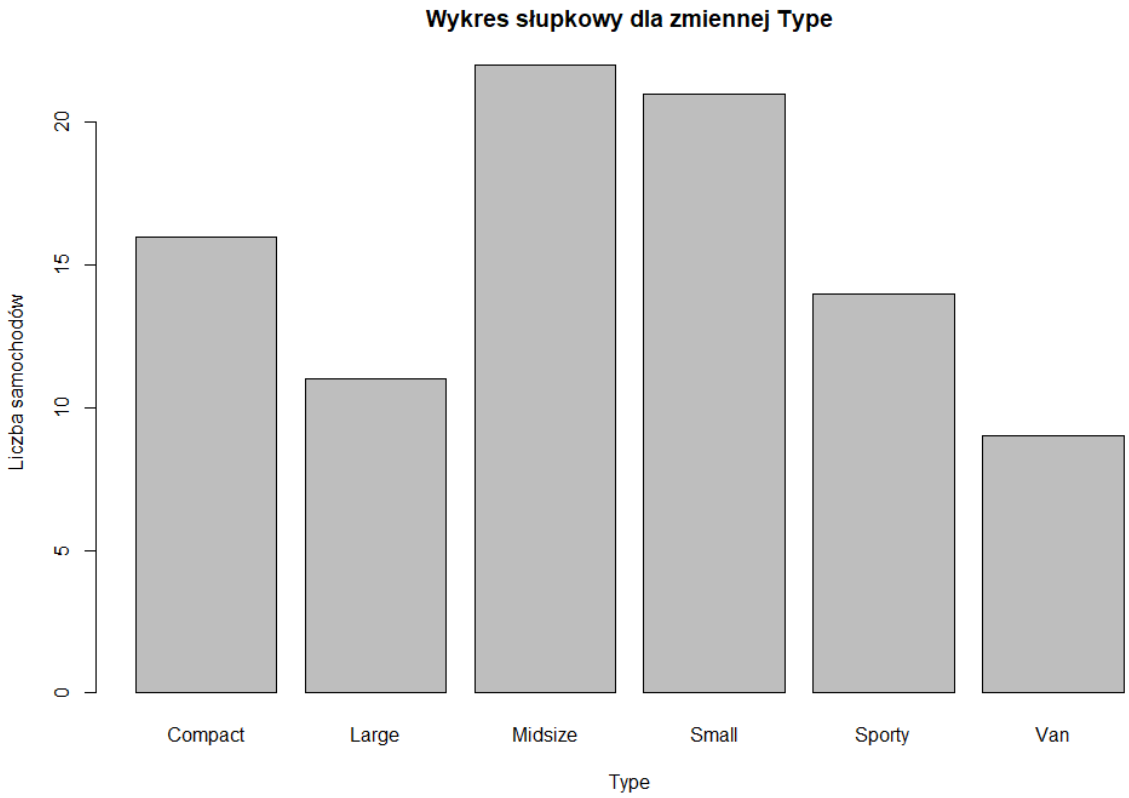
Centy rzędu 95 dla wartości Price.PLN wynosi:

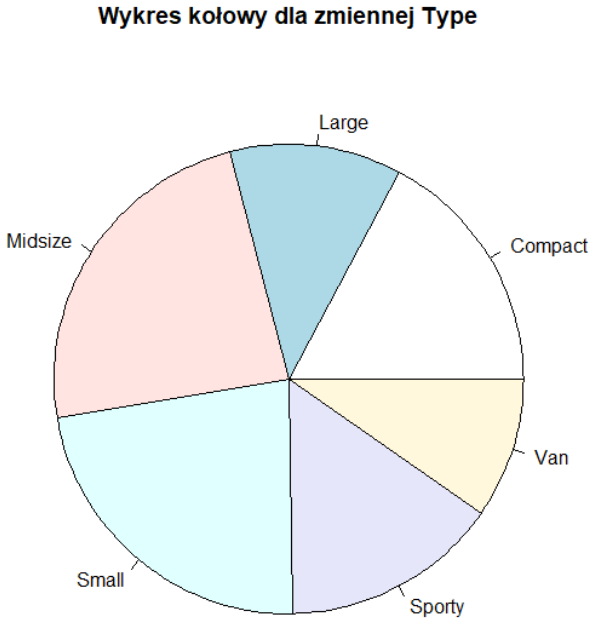
```
    95%
123.079
```

Dane powyżej 95 centyla:

```
    Manufacturer    Model Price.PLN
4           Audi      100   126.295
11      Cadillac  Seville   134.335
19     Chevrolet Corvette   127.300
48       Infiniti      Q45   160.465
59 Mercedes-Benz     300E   207.365
```

Wykresy Type

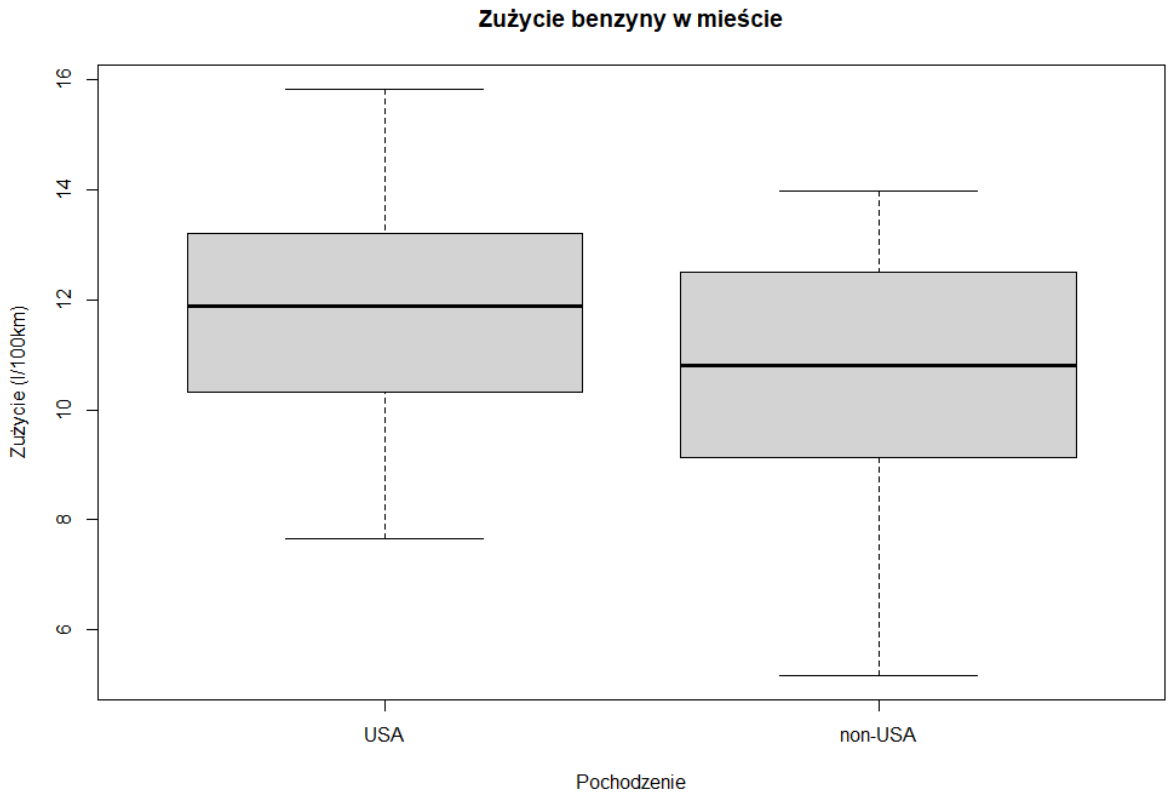




Na podstawie powyższych wykresów możemy zaobserwować, że najbardziej liczną grupę stanowią auta średniego rozmiaru, a najmniej liczną grupę Vany. Liczba samochodów sportowych to 14:

```
Sporty
    14
```

Zużycie benzyny w miescie wykresy pudełkowe:



Na podstawie wykresów pudełkowych możemy stwierdzić, że rozstęp miedzykwartylowy dla aut z USA jest przesunięty w górę, w stosunku do aut spoza USA, co oznacza, że część aut z USA spala w mieście więcej paliwa niż auta spoza USA.

Statystyki spalania w mieście aut z USA:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.661	10.326	11.875	11.709	13.194	15.833

Statystyki spalania w mieście aut spoza USA:

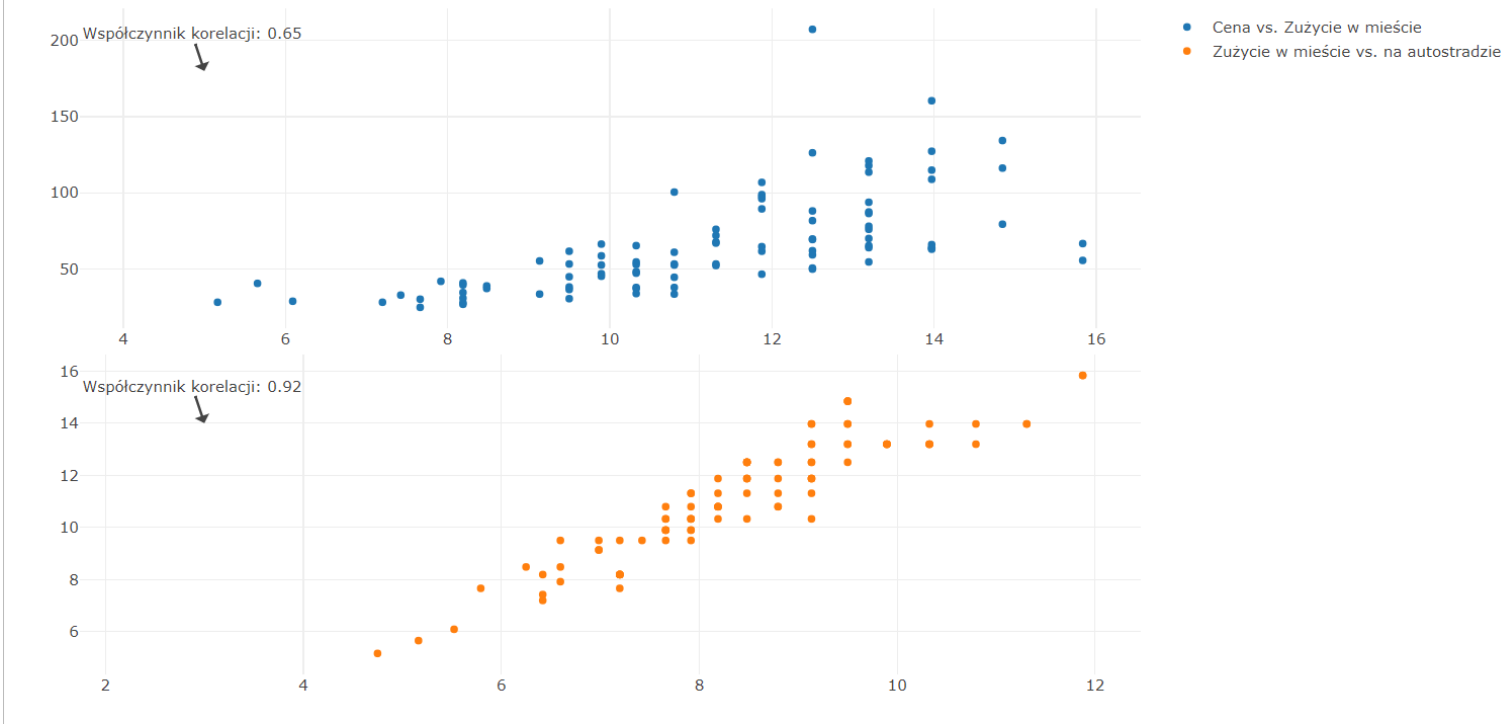
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.163	9.135	10.795	10.576	12.500	13.971

Większość aut amerykańskich znajduje się ponad medianą aut spoza USA, co pozwala wysnuć wniosek, że większość aut z USA spala więcej w mieście więcej niż wynik środkowy aut spoza USA.

Możemy zaobserwować, że dolny wąs dla aut spoza USA jest dłuższy, co sugeruje, że jakiś samochód ma bardzo niskie spalanie w mieście (5,163 l/100km).

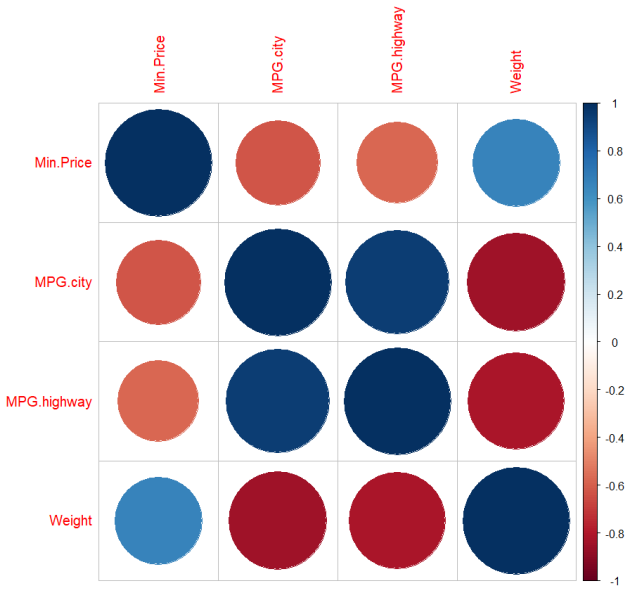
Podsumowując Auta spoza USA są statystycznie bardziej oszczędne niż auta z USA.

Poniżej przedstawiono wykres rozrzutu ceny podstawowej wersji samochodu od jego zużycia benzyny w mieście oraz wykres rozrzutu zużycia benzyny w mieście w funkcji zużycia benzyny na autostradzie.



Na podstawie powyższych zależności możemy stwierdzić, że rozrzutu ceny podstawowej wersji samochodu od jego zużycia benzyny w mieście są zmiennymi o znaczącej zależności liniowej ponieważ współczynnik korelacji Pearsona wynosi 0,65. Natomiast rozrzutu zużycia benzyny w mieście w funkcji zużycia benzyny na autostradzie stanowią zmienne o bardzo silnej zależności liniowej ponieważ współczynnik korelacji Pearsona wynosi 0,92.

Poniżej graficznie przedstawiono współczynniki korelacji Persona dla wykorzystywanych zmiennych:

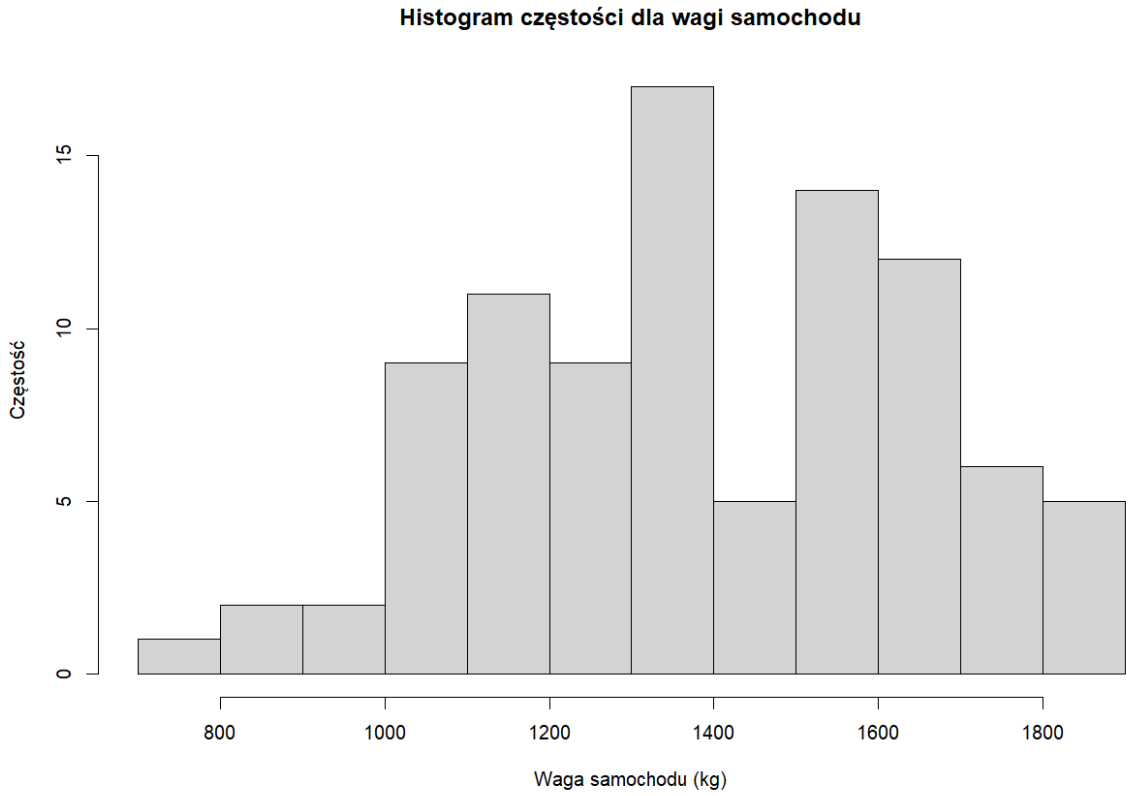


Utworzono więc model liniowy łączący zużycie benzyny w mieście i na autostradzie:

```
Call:
lm(formula = Fuel_usage.highway ~ Fuel_usage.city, data = Cars93)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99296 -0.37182 -0.02261  0.33517  1.31013

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.10228    0.28276   7.435 5.57e-11 ***
Fuel_usage.city  0.56527    0.02482  22.772  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5453 on 91 degrees of freedom
Multiple R-squared:  0.8507,    Adjusted R-squared:  0.8491
F-statistic: 518.5 on 1 and 91 DF,  p-value: < 2.2e-16
```

Możemy zauważyć, że model liniowy bardzo dokładnie opisuje te dane ponieważ p-value przyjmuje bardzo niską wartość rzędu 2.2e-16 i oczywiście została oceniona na ***.

Histogram wagi:



Możemy zaobserwować, że najwięcej aut zawiera się w przedziale od 1300 do 1400 kg. Najmniej liczną grupę stanowią auta warzące od 700 do 800 kg.

Zad 2

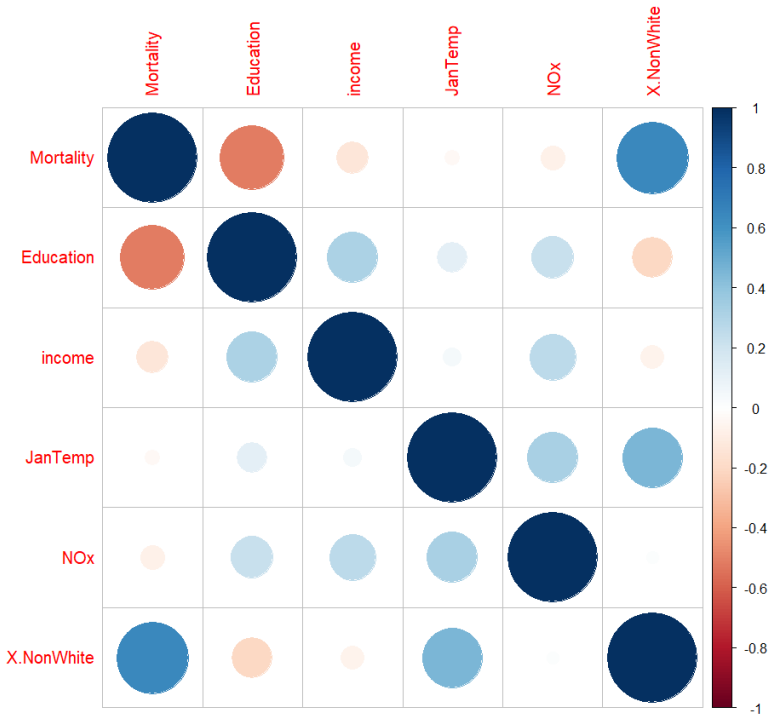
Wczytano plik airpollution.txt zawierający dane dotyczące związku pomiędzy zanieczyszczeniem powietrza i śmiertelnością w 60 miastach amerykańskich. Poniżej przedstawiono analizę statystyczną zmiennych Mortality, Education, NonWhite, income, JanTemp, JulTemp oraz NOx :

Mortality	Education	income	JanTemp	NOx	X.NonWhite
Min. : 790.7	Min. : 9.00	Min. : 40	Min. :12.00	Min. : 1.00	Min. : 0.80
1st Qu.: 898.4	1st Qu.:10.40	1st Qu.:29877	1st Qu.:27.00	1st Qu.: 4.00	1st Qu.: 4.95
Median : 943.7	Median :11.05	Median :32451	Median :31.50	Median : 9.00	Median :10.40
Mean : 940.3	Mean :10.97	Mean :32693	Mean :33.98	Mean : 22.60	Mean :11.87
3rd Qu.: 983.2	3rd Qu.:11.50	3rd Qu.:35384	3rd Qu.:40.00	3rd Qu.: 23.75	3rd Qu.:15.65
Max. :1113.2	Max. :12.30	Max. :47966	Max. :67.00	Max. :319.00	Max. :38.50

Wykonano analizę korelacji Pearsona dla badanych zmiennych:

	Mortality	Education	income	JanTemp	NOx	X.NonWhite
Mortality	1.00000000	-0.5110939	-0.13128724	-0.03030382	-0.07768549	0.64369115
Education	-0.51109395	1.0000000	0.31802310	0.11628379	0.22413065	-0.20877394
income	-0.13128724	0.3180231	1.00000000	0.04212376	0.26619955	-0.06866819
JanTemp	-0.03030382	0.1162838	0.04212376	1.00000000	0.32150613	0.45377412
NOx	-0.07768549	0.2241307	0.26619955	0.32150613	1.00000000	0.01941797
X.NonWhite	0.64369115	-0.2087739	-0.06866819	0.45377412	0.01941797	1.00000000

Graficzna prezentacja korelacji zmiennych:



Na podstawie współczynnika koelacji Pearsona możemy stwierdzić, że znacząca zależność liniowa cechuje zmienne Mortality i NonWhite, współczynnik Pearsona równy 0,64. Natomiast zmienne Mortality i Education, współczynnik Pearsona równy -0,51, cechuje umiarkowana zależność liniowa.

Poniżej przedstawiono dopasowanie modelu liniowego ze zmienną objaśnianą Mortality i zmienną objaśniająca NOx:

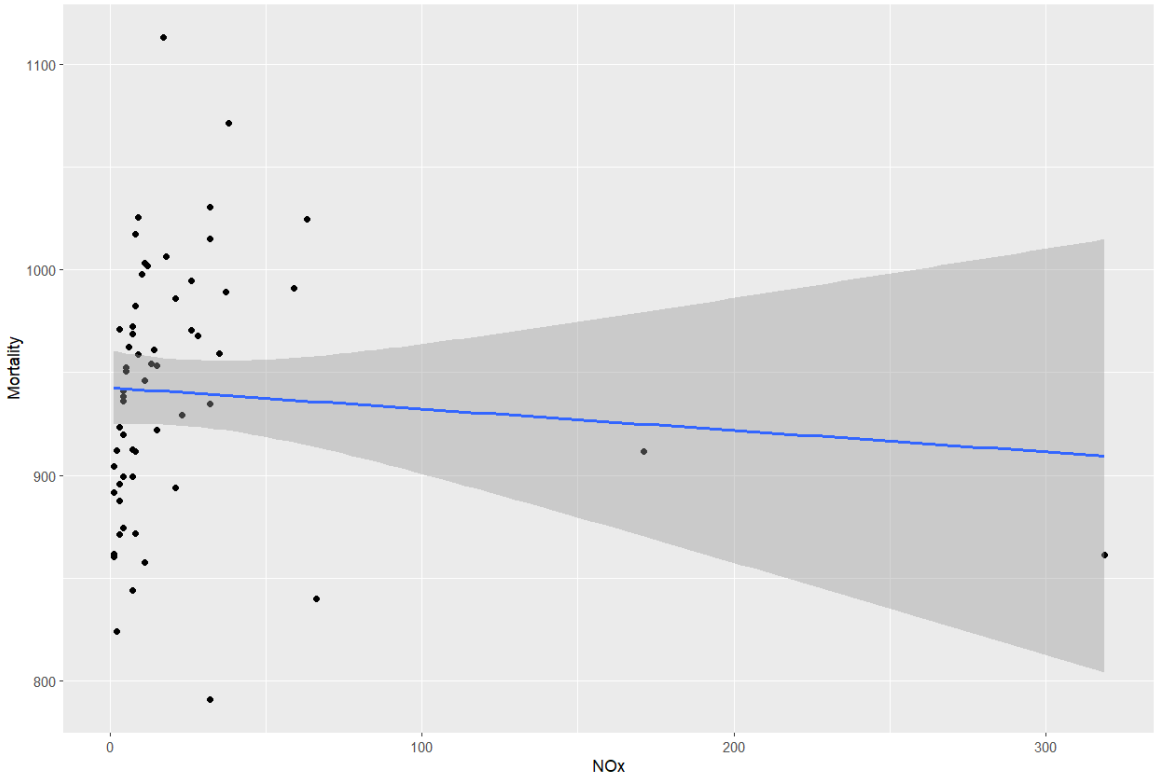
```
Call:
lm(formula = Mortality ~ NOx, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-148.639  -43.700    1.762   41.673  172.227

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 942.7052     9.0002 104.742   <2e-16 ***
NOx          -0.1043     0.1757  -0.593    0.555
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.56 on 58 degrees of freedom
Multiple R-squared:  0.006035,  Adjusted R-squared:  -0.0111
F-statistic: 0.3522 on 1 and 58 DF,  p-value: 0.5552
```

Równanie regresji liniowej opisującej te zmienne:

Mortality = -0,1043 NOx + 943,7052

Z 95% pewnością możemy stwierdzić, że współczynnik nachylenia prostej wynosi: -0,1043±0,1757. P-value wynosi, aż 0,5552, świadczy to o bardzo niskiej jakości modelu.

Niniejsza zależność liniową przedstawiono na poniższej charakterystyce:



Możemy zaobserwować, że 4 punkty wyraźnie odstają od pozostałych. Powoduje to znaczącą zmianę przebiegu prostej, przez co model słabo opisuje dane.

Poniżej przedstawiono dopasowanie modelu liniowego ze zmienną objaśnianą Mortality i zmienną objaśniającą log(NOx)

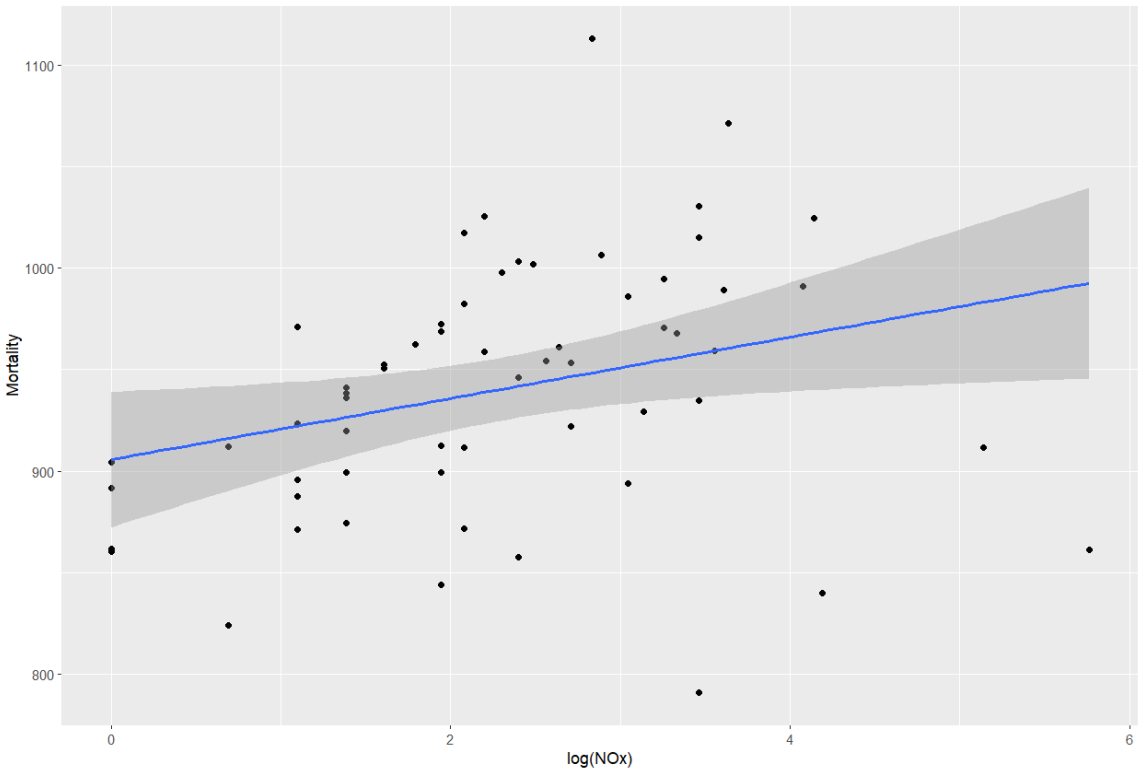
```
Call:
lm(formula = Mortality ~ log(NOx), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-167.212  -28.944    8.372   35.142  164.768

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  905.613     16.672  54.319   <2e-16 ***
log(NOx)      15.099      6.419   2.352   0.0221 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.96 on 58 degrees of freedom
Multiple R-squared:  0.0871,    Adjusted R-squared:  0.07136
F-statistic: 5.533 on 1 and 58 DF,  p-value: 0.02207
```

Równanie regresji liniowej opisującej te zmienne:

Mortality = 15,099 log(NOx) + 943,7052

Z 95% pewnością możemy stwierdzić, że współczynnik nachylenia prostej wynosi: 15,099±6,419. P-value wynosi, 0,0221 (*), świadczy to o średniej jakości modelu, jednak ich zależność jest znacząco wyższa niż w przypadku zależności Mortality i NOx. Możemy zaobserwować, że 4 punkty odstają od pozostałych danych i to one przesuwają środek ciężkości danych, należy je usunąć.

Niniejsza zależność liniową przedstawiono na poniższej charakterystyce:



Wyszukano obserwacje, których wartość bezwzględna residuów standaryzowanych przekraczających 2:

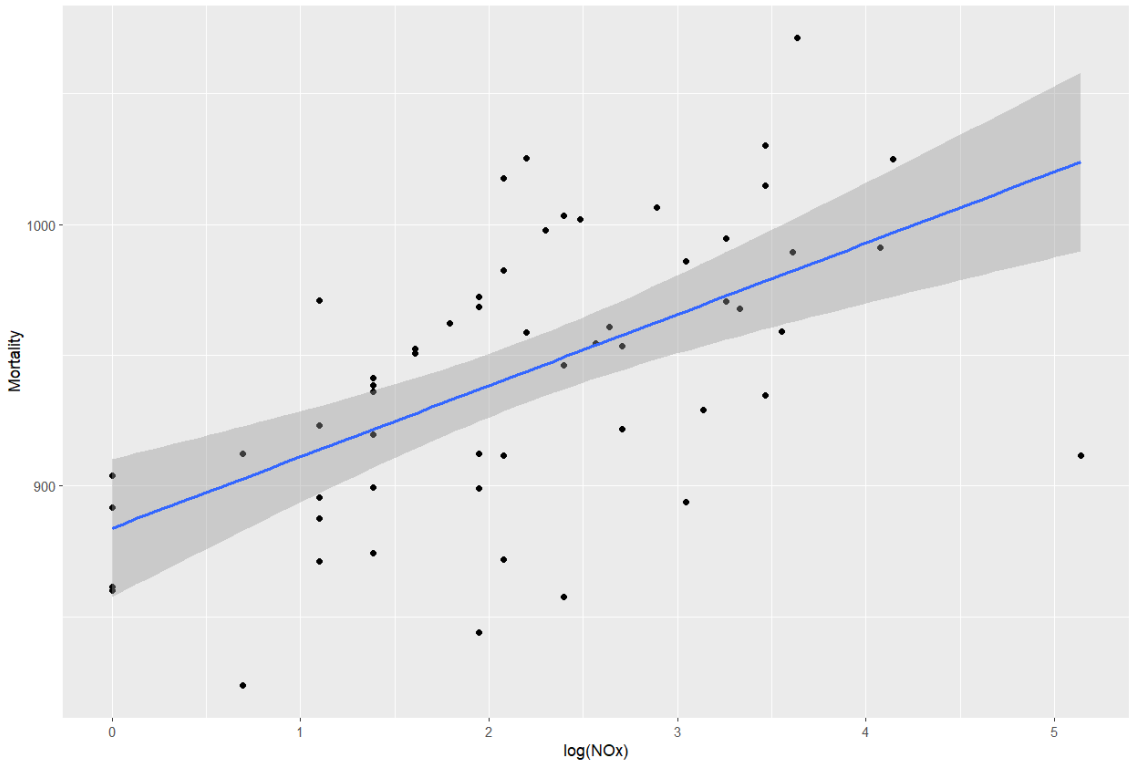
```
outliers <- which(abs(rstudent(model2)) > 2)
```

Poniżej przedstawiono dopasowanie modelu liniowego ze zmienną objaśnianą Mortality i zmienną objaśniającą log(NOx) po usunięciu obserwacji o dużych residuach standaryzowanych:

```
Call:
lm(formula = Mortality ~ log(NOx), data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max
-112.261  -25.015    6.104   28.504   88.297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  883.912     13.134  67.297  < 2e-16 ***
log(NOx)      27.238      5.378   5.064  5.1e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.73 on 54 degrees of freedom
Multiple R-squared:  0.322,     Adjusted R-squared:  0.3095
F-statistic: 25.65 on 1 and 54 DF,  p-value: 5.102e-06
```

Z 95% pewnością możemy stwierdzić, że współczynnik nachylenia prostej wynosi: 27,238±5,378. P-value wynosi, 5,102 e-6 (***), świadczy to dobrym opisywaniu zmiennej objaśnianej przez zmienną objaśniającą, a co za tym idzie poprawności modelu.

Niniejsza zależność liniową przedstawiono na poniższej charakterystyce:



Współczynnik determinacji dla modelu po usunięciu obserwacji o dużych residuach standaryzowanych (r_squared_new) oraz przed (r_squared_old):

r_squared_new	0.32201142269628
r_squared_old	0.0870955860032189

Możemy zaobserwować, że usunięcie punktów odstających z modelu spowodowało wzrost współczynnika R^2 z 0,0871 do 0,3220. Oznacza to że model po usunięciu przedstawia 32% danych.

Zad 3

Pliku savings.txt zawiera informacje dotyczące sytuacji ekonomicznej mieszkańców 50 krajów. Dane są wielkości uśrednione za lata 1960 - 1970:

- Country - nazwa kraju
- Savings - łączne oszczędności przypadające na osobę podzielone przez dochód netto
- pop15 - procent populacji poniżej 15 roku życia
- pop75 - procent populacji powyżej 75 roku życia
- dpi - dochód netto przypadający na jednego mieszkańca
- ddpi - tempo wzrostu dochodu (w %)

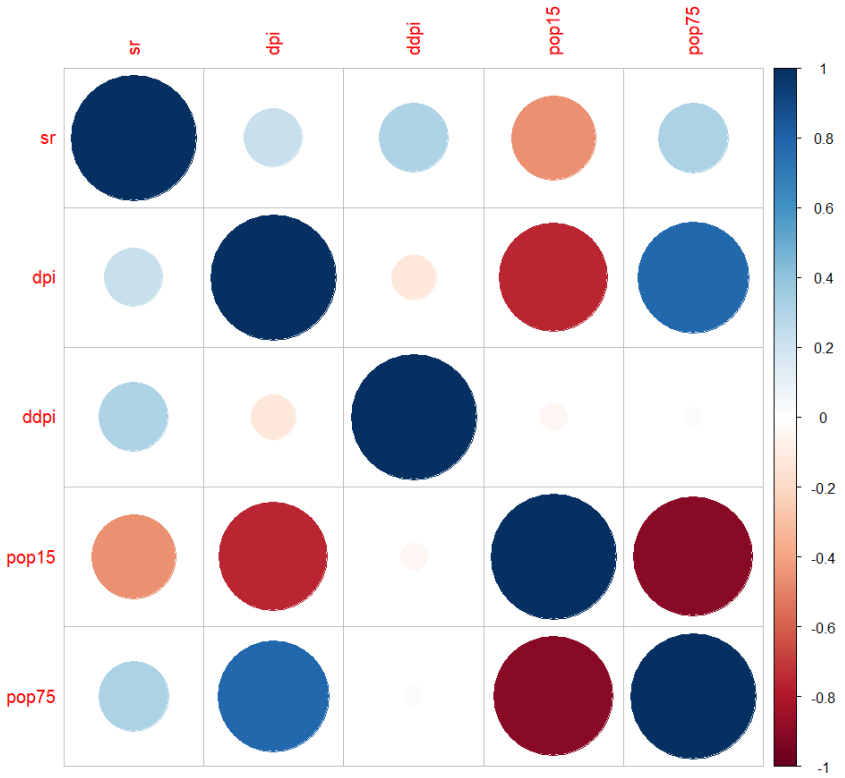
Analiza statystyczna danych zawartych w pliku savings.txt:

sr	pop15	pop75	dpi	ddpi
Min. : 0.600	Min. :21.44	Min. :0.560	Min. : 88.94	Min. : 0.220
1st Qu.: 6.970	1st Qu.:26.21	1st Qu.:1.125	1st Qu.: 288.21	1st Qu.: 2.002
Median :10.510	Median :32.58	Median :2.175	Median : 695.66	Median : 3.000
Mean : 9.671	Mean :35.09	Mean :2.293	Mean :1106.76	Mean : 3.758
3rd Qu.:12.617	3rd Qu.:44.06	3rd Qu.:3.325	3rd Qu.:1795.62	3rd Qu.: 4.478
Max. :21.100	Max. :47.64	Max. :4.700	Max. :4001.89	Max. :16.710

Wyznaczono również macierz korelacji:

	sr	dpi	ddpi	pop15	pop75
sr	1.0000000	0.2203589	0.30478716	-0.45553809	0.31652112
dpi	0.2203589	1.0000000	-0.12948552	-0.75618810	0.78699951
ddpi	0.3047872	-0.1294855	1.00000000	-0.04782569	0.02532138
pop15	-0.4555381	-0.7561881	-0.04782569	1.00000000	-0.90847871
pop75	0.3165211	0.7869995	0.02532138	-0.90847871	1.00000000

Poniżej przedstawiono graficzną prezentację macierzy korelacji:



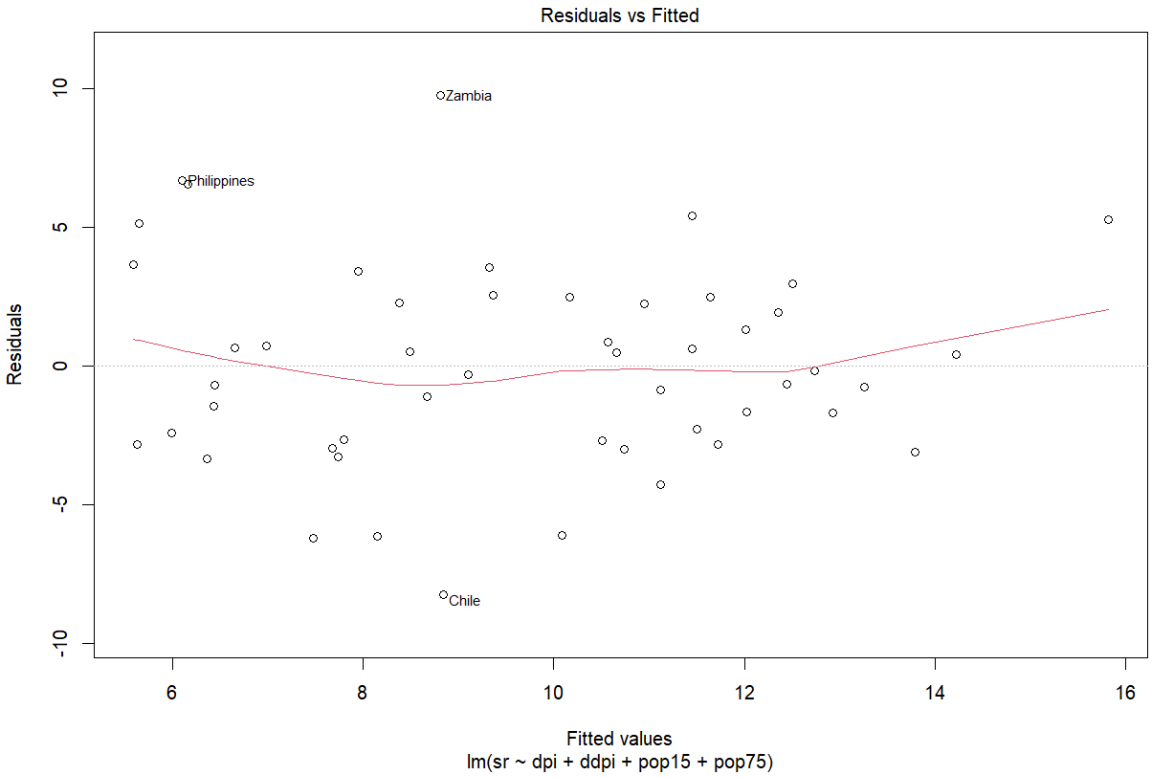
Silnie zależne liniowo są zmienne procent populacji poniżej 15 roku życia oraz powyżej 75 roku życia, ponieważ współczynnik korelacji Pearsona wynosi -0,91. Zmienna Savings najsilniej zależy od zmiennej pop15, współczynnik korelacji Pearsona wynosi -0,46

Dopasowanie modelu liniowego opisującego zależność Savings od dpi, ddpi, Pop15 i Pop75:

```
Call:
lm(formula = sr ~ dpi + ddpi + pop15 + pop75, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Na podstawie wartości współczynnika R^2 który wynosi 0,3385 oraz p-value wynoszącego 0,00079 (***) możemy stwierdzić, że model dobrze opisuje zmienną objaśnianą.

Poniżej przedstawiono wykres reszt dla wyżej powyższego modelu:



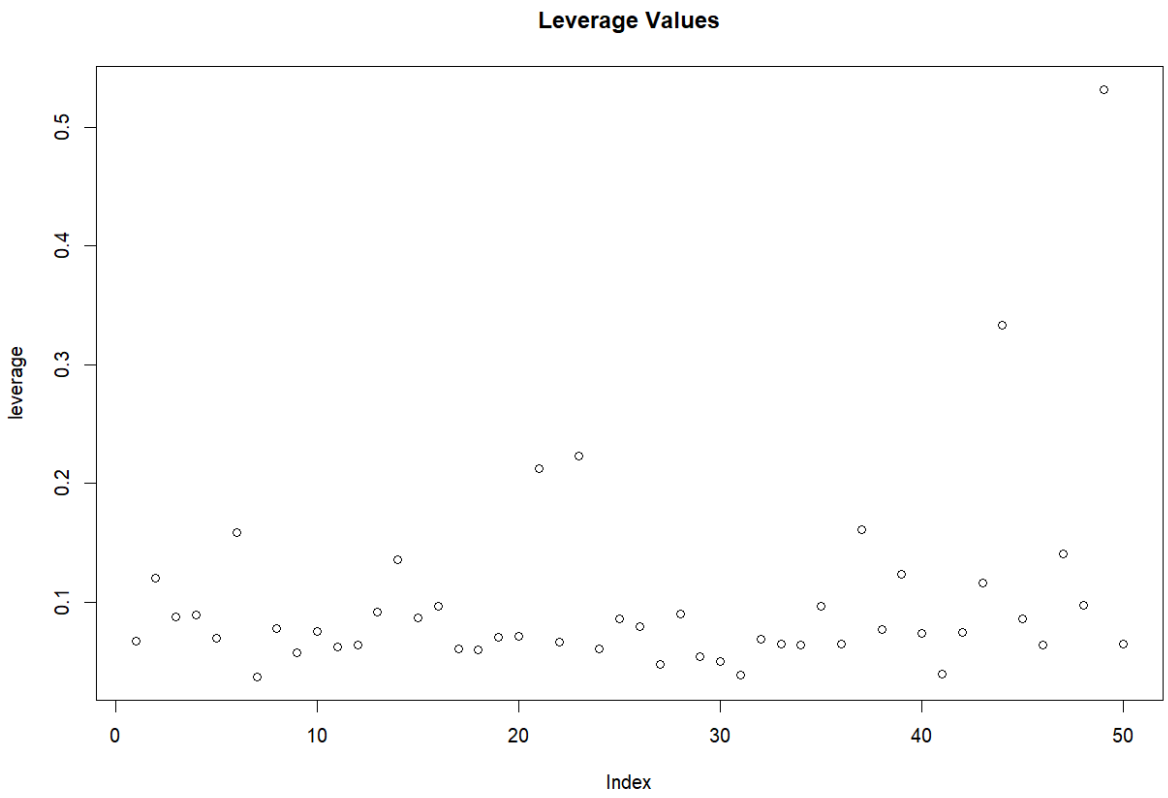
Spełnione jest założenie o liniowości oraz stałej wariancji, ponieważ czerwona linia znajduje się blisko linii przerywanej. Możemy również zauważyć Homoskedastyczność. Dane dla Zambii, Filipin i Chile mogą stanowić wartości odstające.

```
Kraj z największą wartością reszt : Zambia
Kraj z najmniejszą wartością reszt: Chile
```

Wyznaczono wartości dźwigni dla poszczególnych państw:

Australia	Austria	Belgium	Bolivia	Brazil
0.06771343	0.12038393	0.08748248	0.08947114	0.06955944
Canada	Chile	China	Colombia	Costa Rica
0.15840239	0.03729796	0.07795899	0.05730171	0.07546780
Denmark	Ecuador	Finland	France	Germany
0.06271782	0.06372651	0.09204246	0.13620478	0.08735739
Greece	Guatamala	Honduras	Iceland	India
0.09662073	0.06049212	0.06008079	0.07049590	0.07145213
Ireland	Italy	Japan	Korea	Luxembourg
0.21223634	0.06651170	0.22330989	0.06079915	0.08634787
Malta	Norway	Netherlands	New Zealand	Nicaragua
0.07940290	0.04793213	0.09061400	0.05421789	0.05035056
Panama	Paraguay	Peru	Philippines	Portugal
0.03897459	0.06937188	0.06504891	0.06425415	0.09714946
South Africa	South Rhodesia	Spain	Sweden	Switzerland
0.06510405	0.16080923	0.07732854	0.12398898	0.07359423
Turkey	Tunisia	United Kingdom	United States	Venezuela
0.03964224	0.07456729	0.11651375	0.33368800	0.08628365
Zambia	Jamaica	Uruguay	Libya	Malaysia
0.06433163	0.14076016	0.09794717	0.53145676	0.06523300

Poniżej przedstawiono wartości dźwigni na wykresie:



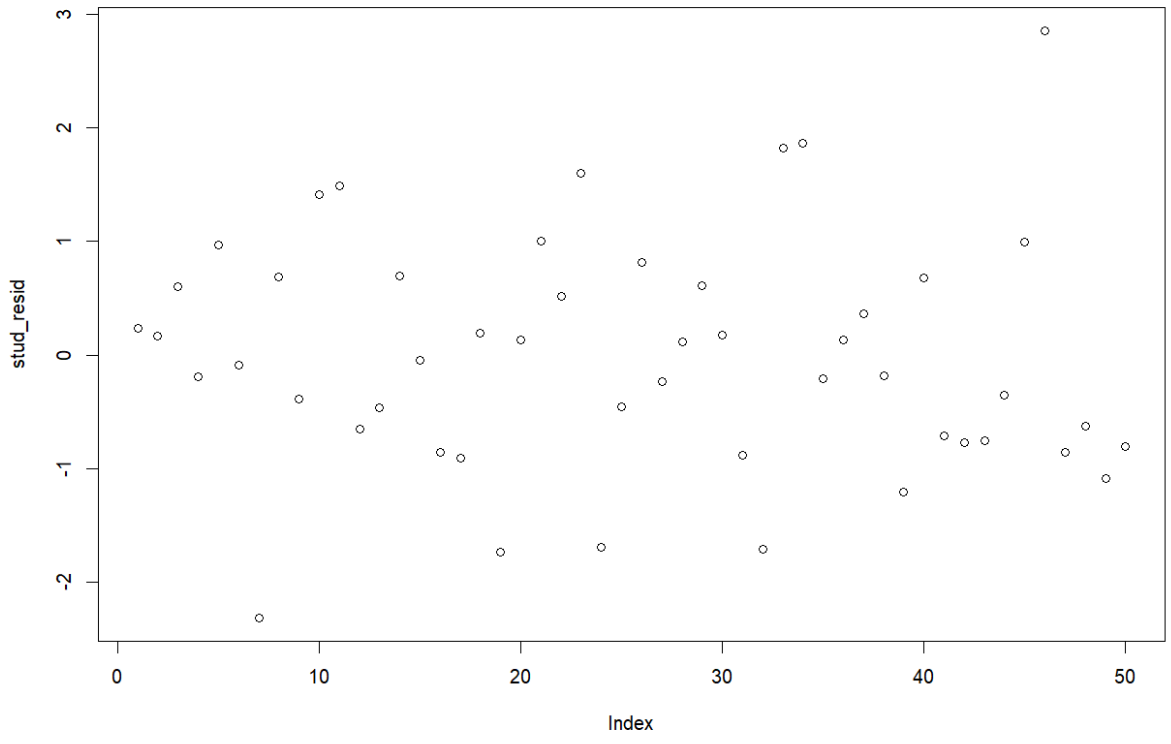
Za obserwacje odstające możemy przyjąć dane dla których wartość dźwigni jest większa od 2(k + 1)/n ,gdzie k oznacza liczbę zmiennych objaśniających, a n to liczba zmiennych. Obserwacje odstające:

Kraje z dużymi wartościami dźwigni:, United States, Libya,

Wyznaczono wartości reszt studentyzoweanych:

Australia	Austria	Belgium	Bolivia	Brazil
0.23271611	0.17095506	0.60655220	-0.19037831	0.96790816
Canada	Chile	China	Colombia	Costa Rica
-0.08983197	-2.31342946	0.69048169	-0.38946778	1.41731062
Denmark	Ecuador	Finland	France	Germany
1.48644473	-0.64957871	-0.45986445	0.69640933	-0.04918692
Greece	Guatamala	Honduras	Iceland	India
-0.85967533	-0.90854545	0.19051919	-1.73119989	0.13729730
Ireland	Italy	Japan	Korea	Luxembourg
1.00485886	0.52015744	1.60321582	-1.69103214	-0.45560591
Malta	Norway	Netherlands	New Zealand	Nicaragua
0.81227407	-0.23247367	0.11605663	0.61373189	0.17254242
Panama	Paraguay	Peru	Philippines	Portugal
-0.88147653	-1.70488128	1.82391409	1.86382587	-0.21040432
South Africa	South Rhodesia	Spain	Sweden	Switzerland
0.12996586	0.36714512	-0.18175853	-1.20293404	0.67532922
Turkey	Tunisia	United Kingdom	United States	Venezuela
-0.71138840	-0.76677907	-0.74959873	-0.35461507	0.99932569
Zambia	Jamaica	Uruguay	Libya	Malaysia
2.85355834	-0.85376418	-0.62253411	-1.08930326	-0.80489153

Poniżej przedstawiono wykres wartości reszt studentyzoweanych:



Jeśli wartość bezwzględna z reszty studentyzoweanej jest większa od dwóch to wtedy tą obserwację można podejrzewać o nietypowość.

Kraje z dużymi wartościami reszt studentyzowanych:, Chile, Zambia,

Przyjmuje się, że dla danej obserwacji, której współczynnik:

$$|DFFITS| > 2 \cdot \sqrt{\frac{k+1}{n}}$$

obserwację uznaje się za wpływową (k to liczba zmiennych objaśniających, a n to liczba obserwacji). Na tej podstawie dla badanych danych obserwacje wpływowe to:

```
Obserwacje wpływowe DFFITS:, Japan, Libya,
```

Przyjmuje się, że dla danej obserwacji, której współczynnik:

$$|DFBETAS| > \frac{2}{\sqrt{n}}$$

obserwację uznaje się za wpływową (n to liczba obserwacji). Na tej podstawie dla badanych danych obserwacje wpływowe to:

```
Obserwacje wpływowe DFBETAS:, Ireland, Japan, Libya
```

Za duża odległość Cook'a uważa się wartość tego współczynnika większą od 4/n. Na tej podstawie dla badanych danych obserwacje o dużym wpływie na obciążenie równania regresii to:

```
Duże odległości Cooka dla krajów:, Japan, Zambia, Libya
```

Przeprowadzono regresję dla danych z wyłączoną obserwacją o największej wartości odległości Cooke'a

Pierwotny model:

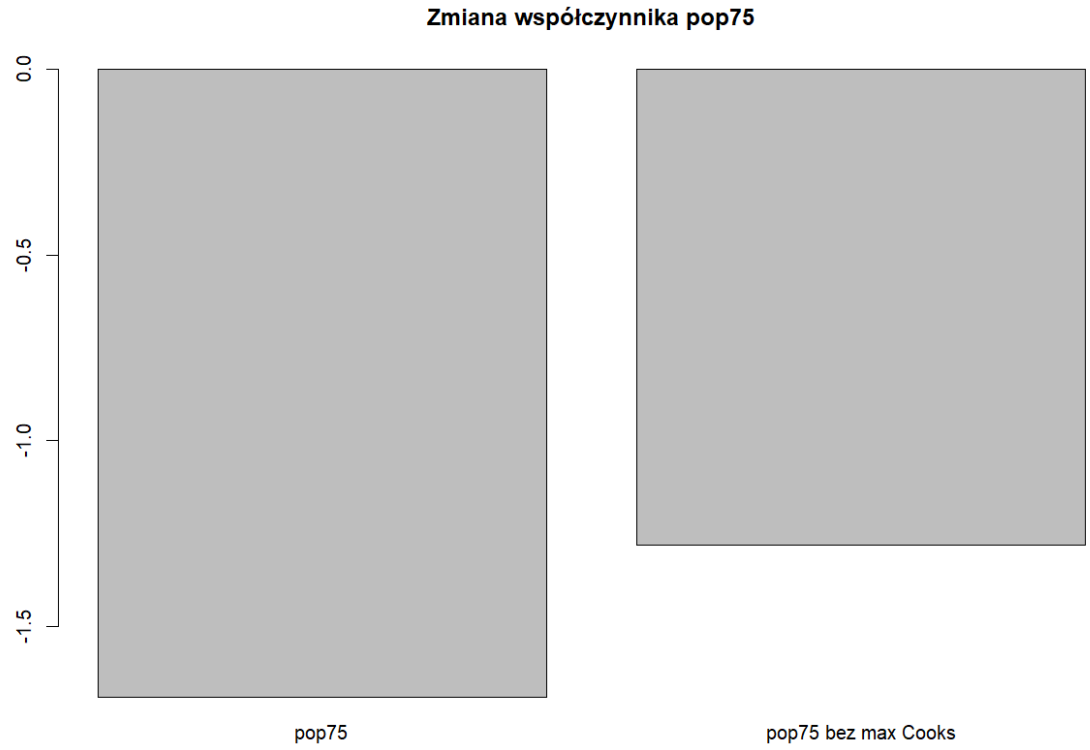
```
Call:
lm(formula = sr ~ dpi + ddpi + pop15 + pop75, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Zmienna dpi ma współczynnik p-value równy 0,72, jest to najwyższa wartość ze wsztsrkich zmiennych objaśniających, oznacza to że jest najmniej istotna. Jej wartość zawiera się w przedziale od 0.1 do 1, możemy więc uznać, że nie jest ona istotna dla tego modelu.

Po usunięciu obserwacji odstającej:

```
Call:
lm(formula = sr ~ dpi + ddpi + pop15 + pop75, data = data[-max_cook,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-8.0699 -2.5408 -0.1584  2.0934  9.3732

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.5240460  8.2240263   2.982  0.00465 **
dpi         -0.0003189  0.0009293  -0.343  0.73312
ddpi         0.6102790  0.2687784   2.271  0.02812 *
pop15       -0.3914401  0.1579095  -2.479  0.01708 *
pop75       -1.2808669  1.1451821  -1.118  0.26943
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.795 on 44 degrees of freedom
Multiple R-squared:  0.3554,    Adjusted R-squared:  0.2968
F-statistic: 6.065 on 4 and 44 DF,  p-value: 0.0005617
```

Poniżej przedstawiono wykresy słupkowe reprezentujące zmianę współczynnika pop 15 ora 75 dla obu modeli:



Wyznaczenie statystyk wpływu:

```
influence_stats <- influence.measures(new_model)
```

Na podstawie DFBETAS (miara wpływu obserwacji na poszczególne parametry modelu.):

```
Największy wpływ (pop15): Costa Rica
Największy wpływ (pop75): Ireland
```

Zad 4

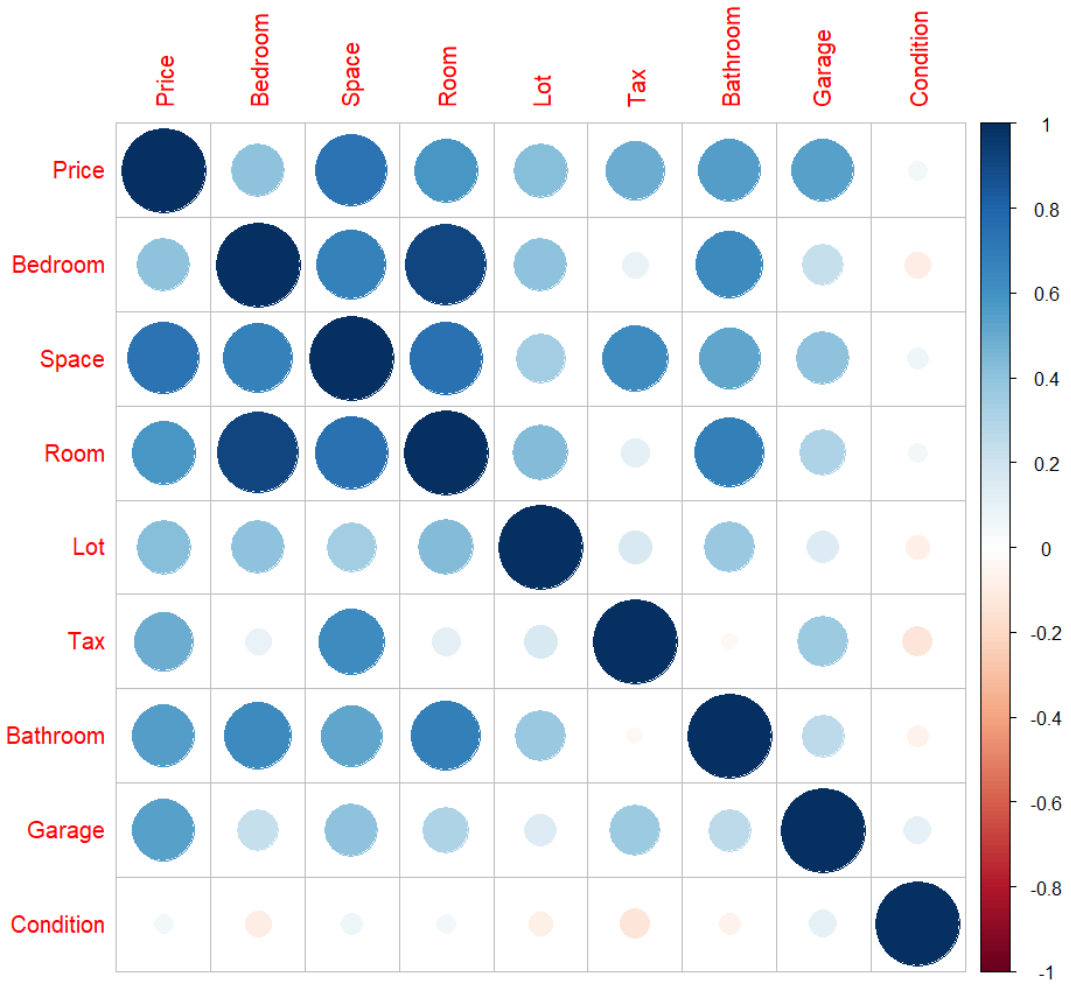
Dane zawarte w zbiorze realest.txt zawierają zmienne:

- cena domu na przedmieściach Chicago (Price)
- liczba sypialni (Bedroom),
- powierzchna w stopach kwadratowych (Space),
- liczba pokoi (Room),
- szerokość frontu dziadki w stopach (Lot),
- rocznego podatku od nieruchomości (Tax),
- liczba łazienek (Bathroom),
- liczba miejsc parkingowych w garażu (Garage),
- stan domu (Condition, 0-dobry, 1-wymaga remontu).

W celu analizy danych wykreślono macierz korelacji Pearsona:

	Price	Bedroom	Space	Room	Lot	Tax	Bathroom	Garage	Condition
Price	1.00000000	0.40684514	0.73592375	0.58061367	0.42278319	0.49666446	0.55282008	0.5439802	0.05236443
Bedroom	0.40684514	1.00000000	0.67530286	0.91752513	0.40425825	0.09589000	0.63279126	0.2394497	-0.09872838
Space	0.73592375	0.67530286	1.00000000	0.74007160	0.34216924	0.62179751	0.52515952	0.4048887	0.06870648
Room	0.58061367	0.91752513	0.74007160	1.00000000	0.43111109	0.11733545	0.68743357	0.3001060	0.05466723
Lot	0.42278319	0.40425825	0.34216924	0.43111109	1.00000000	0.16186163	0.37228112	0.1437620	-0.08428528
Tax	0.49666446	0.09589000	0.62179751	0.11733545	0.16186163	1.00000000	-0.03649755	0.3637449	-0.13027622
Bathroom	0.55282008	0.63279126	0.52515952	0.68743357	0.37228112	-0.03649755	1.00000000	0.2643710	-0.06653373
Garage	0.54398024	0.23944972	0.40488870	0.30010602	0.14376201	0.36374488	0.26437102	1.0000000	0.10456549
Condition	0.05236443	-0.09872838	0.06870648	0.05466723	-0.08428528	-0.13027622	-0.06653373	0.1045655	1.00000000

Poniżej przedstawiono graficzną prezentację macierzy korelacji:



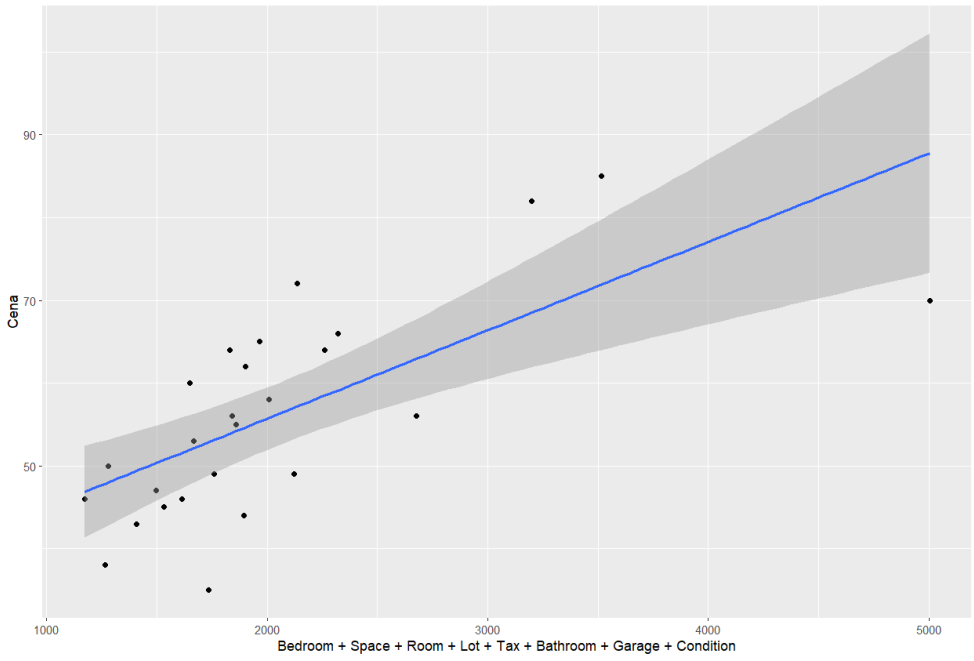
Na podstawie macierzy korelacji możemy zaobserwować, że zmienna Price najbardziej zależy od zmiennych: Space 0,74 oraz Room 0,58. Natomiast zmienna Bedroom najbardziej zależy od liczby Space, Room oraz Bathroom. Jest to oczywista zależność ponieważ większa liczba przestrzeni zazwyczaj wiąże się z większą liczbą pokoi, a co za tym idzie liczbą sypialni i łazienek.

Na podstawie danych zawartych w pliku wyznaczono model w którym zmienną objaśnianą była cena, a zmiennymi objaśnianymi pozostałe zmienne:

```
Call:
lm(formula = Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
    Garage + Condition, data = dane)

Residuals:
     Min       1Q   Median       3Q      Max
-12.7630  -4.0514   0.5389   2.3899  12.9855

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.712572   9.514111   1.441   0.1677
Bedroom     -7.756208   3.109374  -2.494   0.0232 *
Space        0.011626   0.008981   1.295   0.2128
Room         5.097706   2.764303   1.844   0.0827 .
Lot          0.228063   0.195434   1.167   0.2593
Tax          0.003374   0.006859   0.492   0.6291
Bathroom     5.718372   4.276867   1.337   0.1988
Garage       3.613603   2.064997   1.750   0.0982 .
Condition   -2.162027   4.137400  -0.523   0.6080
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.337 on 17 degrees of freedom
Multiple R-squared:  0.7688,    Adjusted R-squared:   0.66
F-statistic: 7.065 on 8 and 17 DF,  p-value: 0.0003757
```

Współczynnik determinacji modelu wynosi 0,7688, współczynnik p-value 0,00038. Świadczy to o dobrym opisywaniu danych przez model. Zmodyfikowany współczynnik determinacji modelu wynosi (0,66), wynika to z niewielkiego wpływu na model takich zmiennych jak Tax czy Lot. Model ten może posłużyć do predykcji danych spoza zbioru.

Poniżej przedstawiono graficzną prezentację tego modelu:



Zbadano wpływ zwiększenia liczby sypialni na cenę nieruchomości na podstawie współczynnika dla zmiennej Bedroom w wykonanym modelu:

```
Wpływ zwiększenia liczby sypialni o 1:  -7.756208
```

Oznacza to że cena domu maleje wraz ze wzrostem liczby sypialni, ponieważ współczynnik dla zmiennej Bedroom jest ujemny.

Obserwacja o największym wpływie na podstawie odległości Cooka:

	Price	Bedroom	Space	Room	Lot	Tax	Bathroom	Garage	Condition
6	44	4	897	7	25	960	2	1	0
8	70	3	2261	6	29	2700	1	2	0
11	85	8	2240	12	50	1200	3	2	0

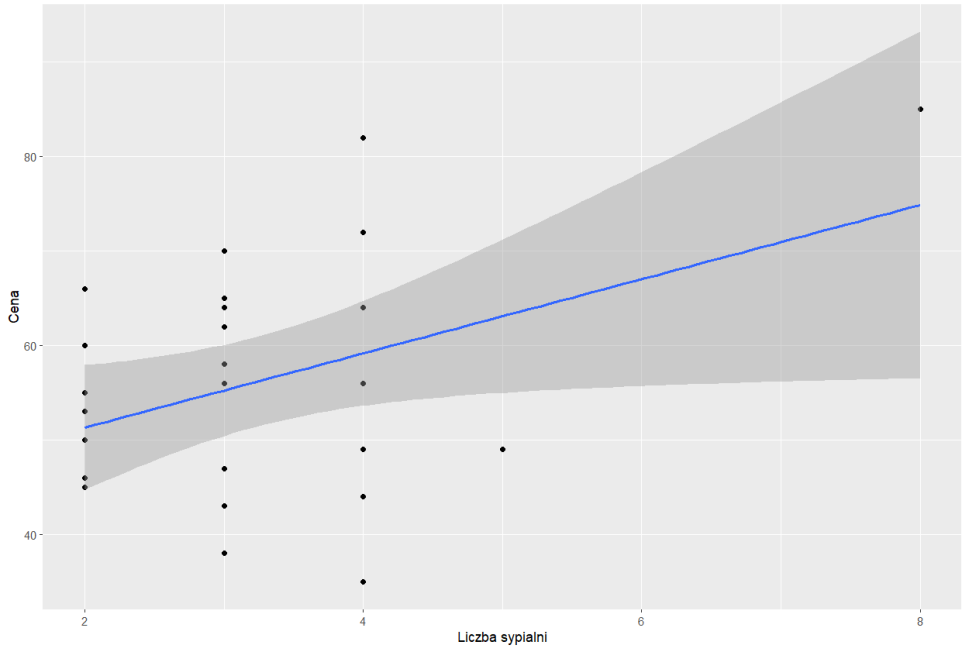
Spadek ceny przy wzroście liczby sypialni jest związany z obserwacjami które przy małej liczbie sypialni osiągnęły wysoką cenę. Obserwując wykres cen od sypialni (omówiony w dalszej części zadania) możemy zauważyć, że wraz ze wzrostem liczby sypialni rośnie rozstrzał cen nieruchomości.

Na podstawie danych zawartych w pliku wyznaczono model w którym zmienną objaśnianą była cena, a zmienną objaśniającą liczba sypialni:

```
Call:
lm(formula = Price ~ Bedroom, data = dane)

Residuals:
    Min      1Q  Median      3Q     Max
-24.170  -7.769   1.211   8.731  22.830

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.487      6.245   6.964 3.35e-07 ***
Bedroom        3.921      1.797   2.182   0.0391 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.73 on 24 degrees of freedom
Multiple R-squared:  0.1655,    Adjusted R-squared:  0.1308
F-statistic: 4.761 on 1 and 24 DF,  p-value: 0.03914
```

Na podstawie danych modelu możemy zauważyć, że zmienna Price nie zależy silnie liniowo od Bedroom, ponieważ współczynnik determinacji wynosi jedynie 0,1655. Ponadto współczynnik dla zmiennej Bedroom wynosi (3,921±1,797), jest obarczona sporym błędem standardowym. Na jej podstawie możemy wnioskować, że wraz ze wzrostem liczby sypialni cena domu wzrośnie. Jednak przy tym modelu nie bierzemy pod uwagę innych ważnych zmiennych wpływających na cenę nieruchomości.

Poniżej przedstawiono graficzną prezentację tego modelu:



Predykcja

Na podstawie predykcji wyznaczono cenę domu w dobrym stanie, z 3 sypialniami, o powierzchni 1500 stóp kwadratowych, z 8 pokojami, 40 stopami szerokości działki, 5 łazienkami, 1 miejscem w garażu i podatkiem w wysokości 1000 dolarów:

```
Przewidywana cena domu:  93.36735
```

Zad 5

Dane w pliku gala_data.txt zawierają informacje dotyczące 30 Wysp Galapagos:

Species - liczba gatunków żółwi na danej wyspie,

Endemics - liczba gatunków endemicznych,

Area - powierzchnia wyspy (w km2),

Elevation - najwyższe wzniesienie na wyspie (w m),

Nearest - odległość od najbliższej wyspy (w km),

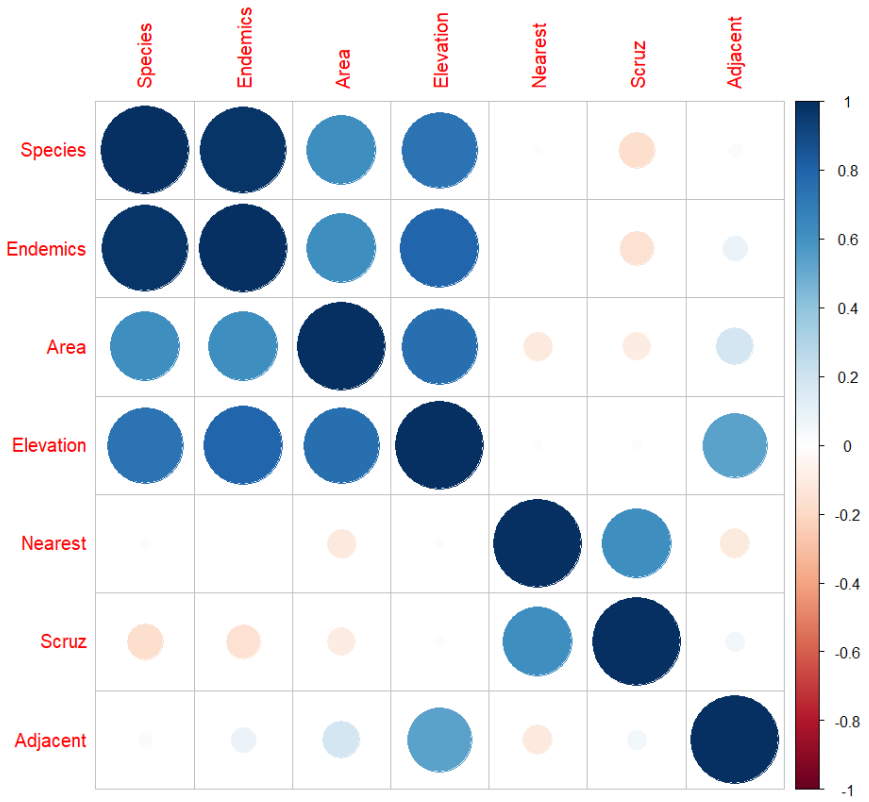
Scruz - odległość od wyspy Santa Cruz (w km),

Adjacent - powierzchnia najbliższej wyspy (w km^2).

W celu analizy danych wykreślono macierz korelacji Pearsona:

	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
Species	1.00000000	0.970876516	0.6178431	0.73848666	-0.014094067	-0.17114244	0.02616635
Endemics	0.97087652	1.000000000	0.6169791	0.79290437	0.005994286	-0.15426432	0.08265803
Area	0.61784307	0.616979087	1.0000000	0.75373492	-0.111103196	-0.10078493	0.18003759
Elevation	0.73848666	0.792904369	0.7537349	1.00000000	-0.011076984	-0.01543829	0.53645782
Nearest	-0.01409407	0.005994286	-0.1111032	-0.01107698	1.000000000	0.61541036	-0.11624788
Scruz	-0.17114244	-0.154264319	-0.1007849	-0.01543829	0.615410357	1.00000000	0.05166066
Adjacent	0.02616635	0.082658026	0.1800376	0.53645782	-0.116247885	0.05166066	1.00000000

Poniżej przedstawiono graficzną prezentację macierzy korelacji:

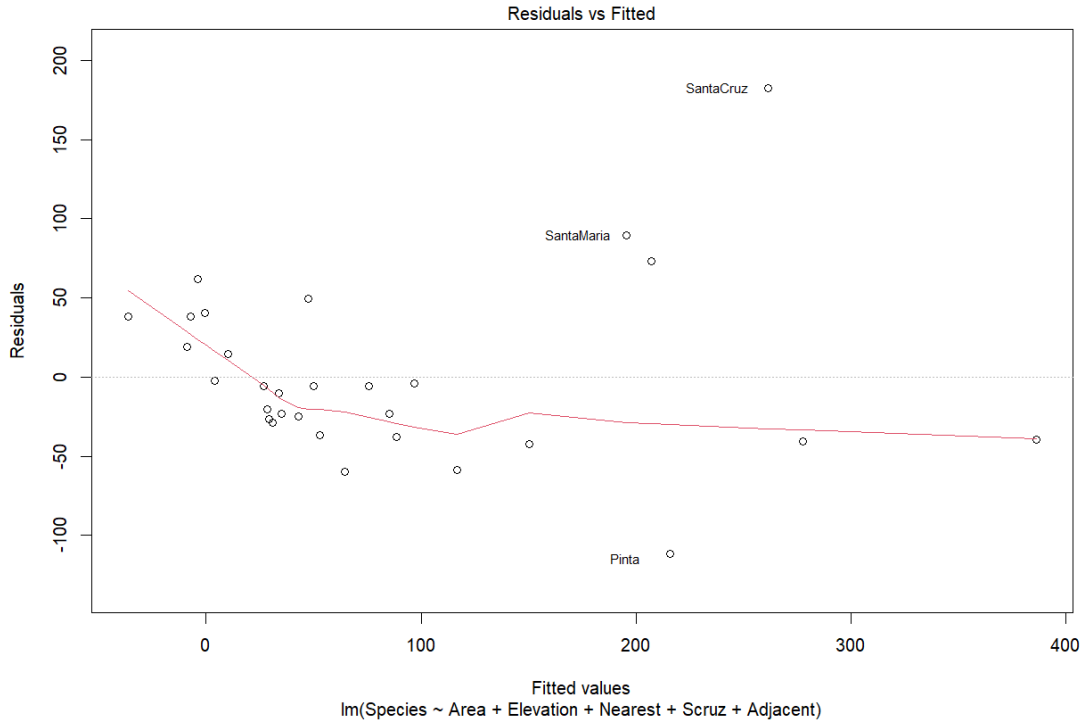


Na podstawie macierzy korelacji możemy zaobserwować, że zmienna Species najbardziej zależy od zmiennej: Endemics 0,97. Tak wysoka zależność tych zmiennych jest nie tylko korelacją, a wynikowością, ponieważ liczba gatunków żółwi zależy od liczby gatunków występujących na danym terenie.

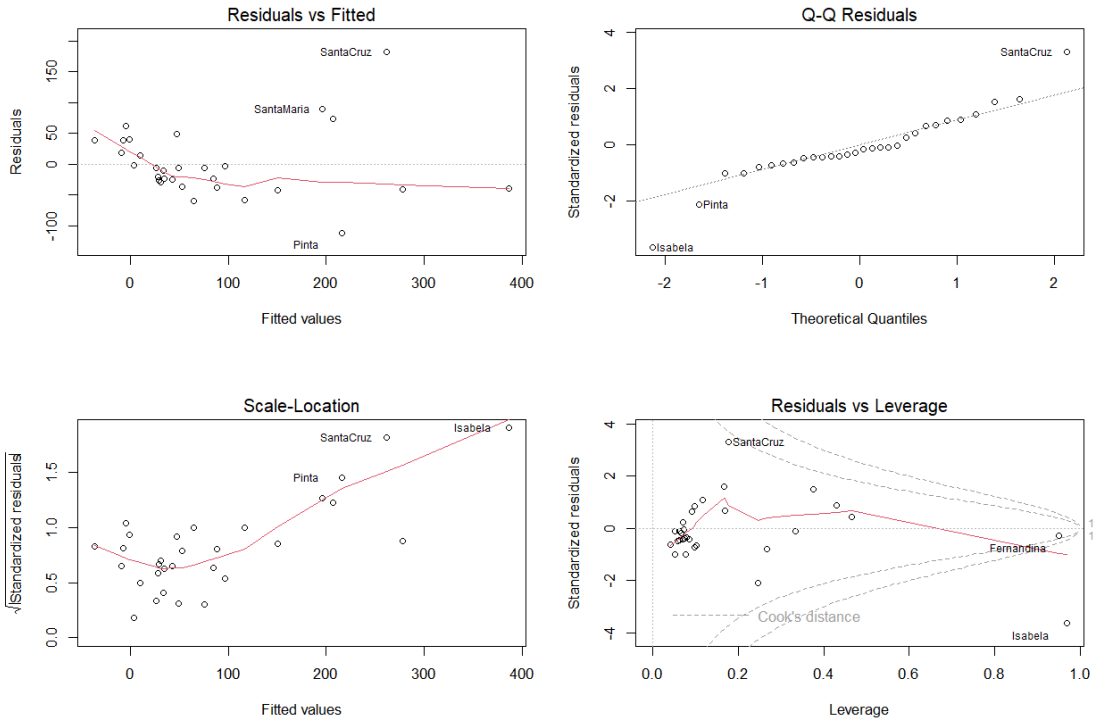
Na podstawie danych zawartych w pliku wyznaczono model w którym zmienną objaśnianą była liczba gatunków żółwi, a zmiennymi objaśnianymi pozostałe zmienne:

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = dane)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Dla wykonanego modelu wyznaczono wartości residuów w zależności od wartości przewidywanych:



Wyznaczono również pozostałe statystyki modelu:

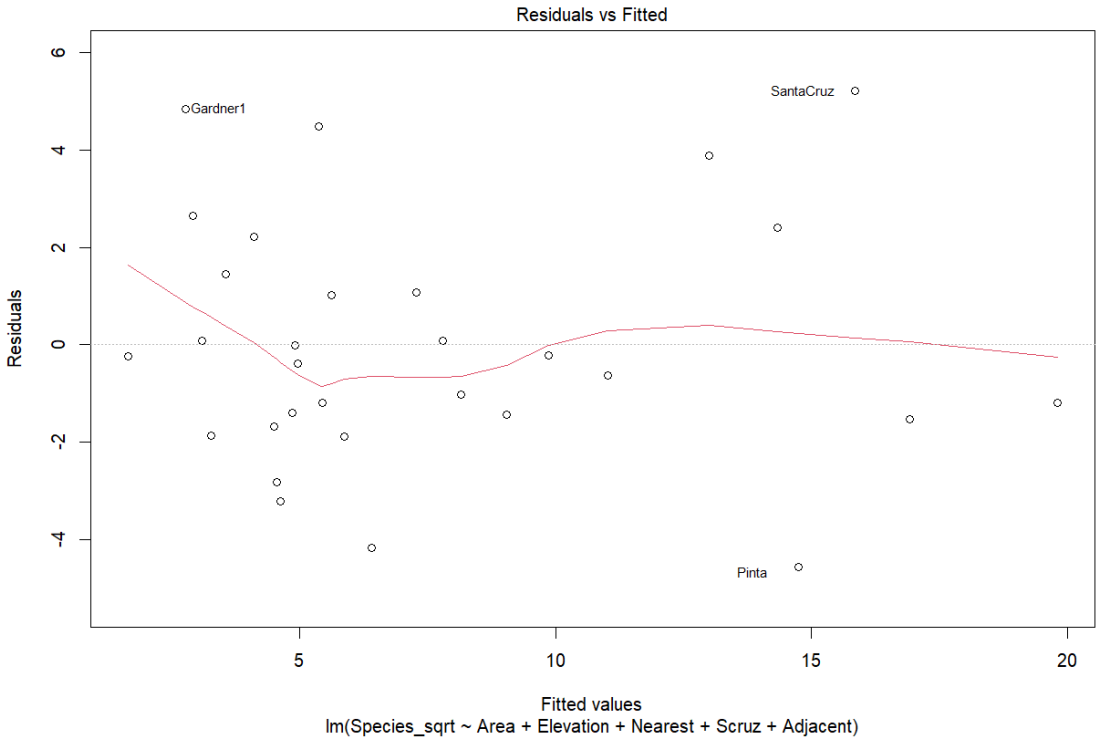


Na wykresie reszt oraz reszt standaryzowanych punkty nie są rozproszone losowo wokół linii horyzontalnej na wysokości 0, może to sugerować, że wariancja reszt zależy od przewidywanych wartości, średnia wariancja reszt znacząco odbiega od wartości 0 co jest naruszeniem założeń o liniowości oraz stałej wariancji. Wartość wariancji residuów zależy od wartości przewidywanych. Podważa to zasadność wykorzystywania modelu w celach predykcyjnych.

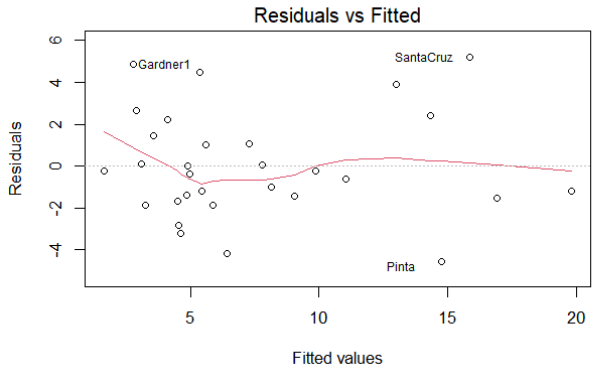
W celu usunięcia problemu zmiennej

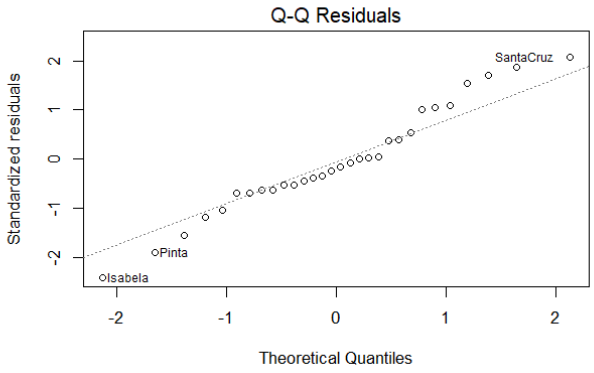
```
Call:
lm(formula = Species_sqrt ~ Area + Elevation + Nearest + Scruz +
    Adjacent, data = dane)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5572 -1.4969 -0.3031  1.3527  5.2110

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3919243  0.8712678   3.893 0.000690 ***
Area        -0.0019718  0.0010199  -1.933 0.065080 .
Elevation    0.0164784  0.0024410   6.751 5.55e-07 ***
Nearest      0.0249326  0.0479495   0.520 0.607844
Scruz       -0.0134826  0.0097980  -1.376 0.181509
Adjacent    -0.0033669  0.0008051  -4.182 0.000333 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.774 on 24 degrees of freedom
Multiple R-squared:  0.7827,    Adjusted R-squared:  0.7374
F-statistic: 17.29 on 5 and 24 DF,  p-value: 2.874e-07
```

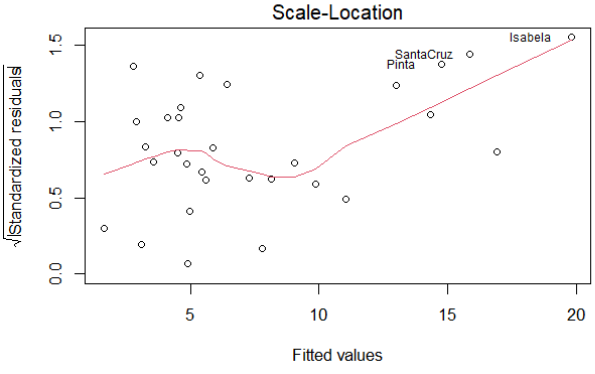
Dla wykonanego modelu wyznaczono wartości residuów w zależności od wartości przewidywanych:

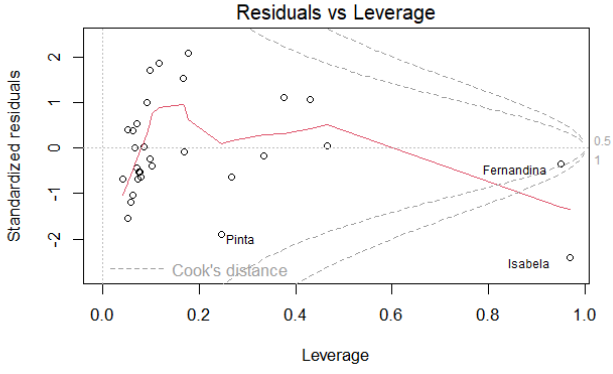


Wyznaczono również pozostałe statystyki modelu:









Na wykresie reszt oraz reszt standaryzowanych punkty są rozproszone losowo wokół linii horyzontalnej na wysokości 0, ponadto średnia wariancja reszt jest blisko wartości zerowej. Model ten spełnia więc założenie o liniowości oraz stałej wariancji. Wartość wariancji residuów nie zależy od wartości przewidywanych.

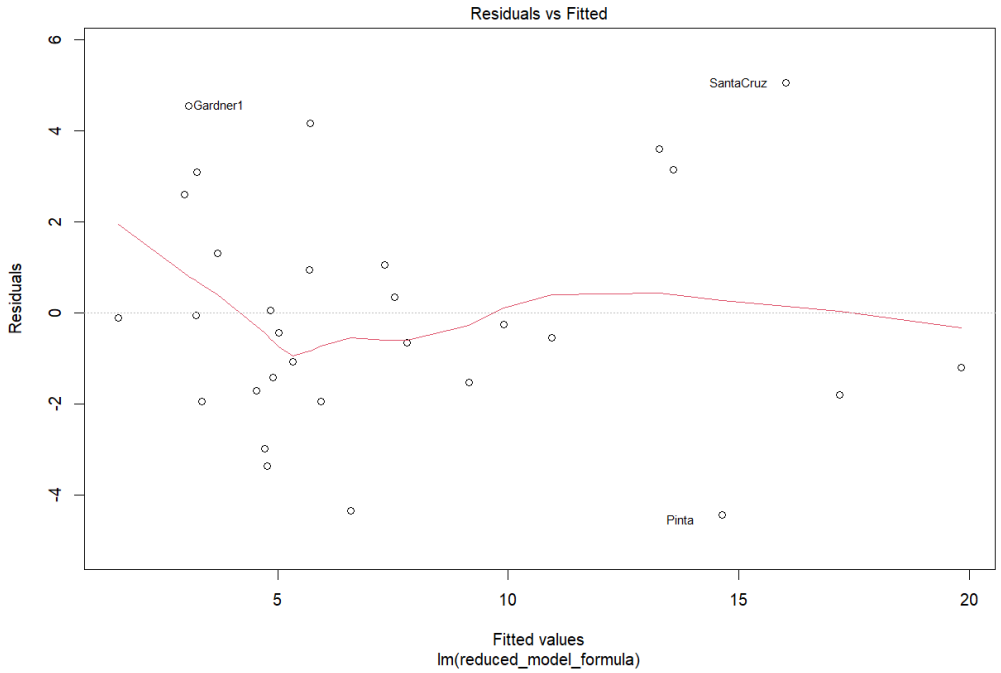
W celu poprawy jakości modelu, którego współczynnik determinacji wynosi 0,7827 usunięto zmienną objaśniającą, której współczynnik p-value był najmniejszy:

```
Zmienna o największym p-value:  Nearest
```

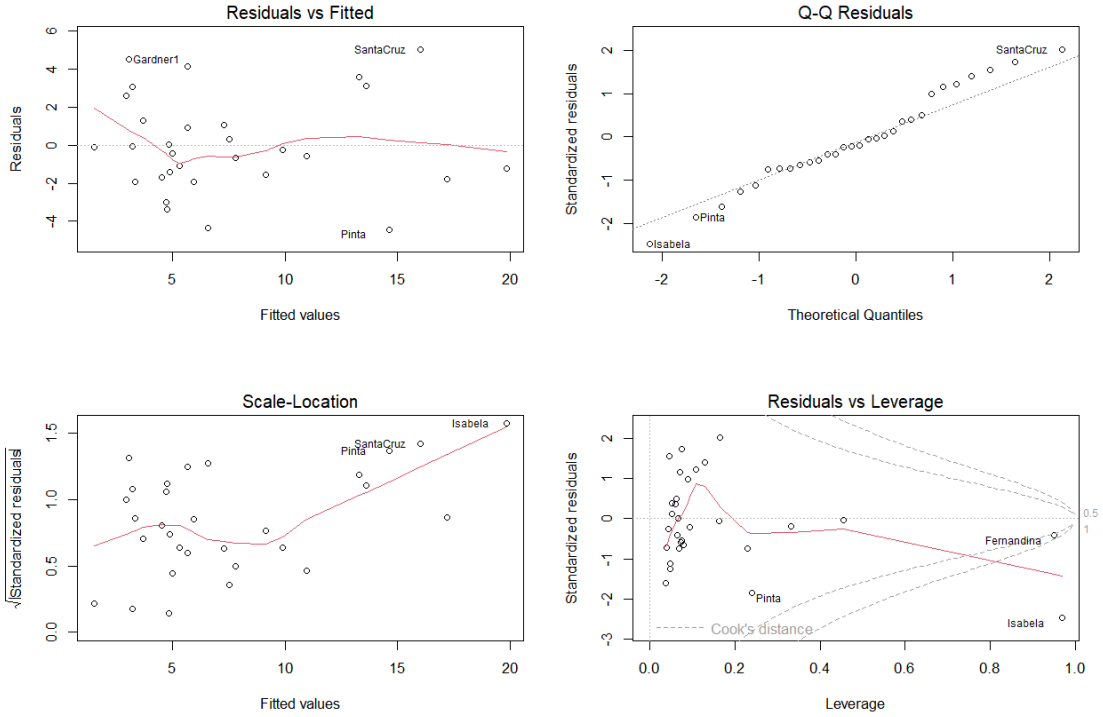
Dla danych po usunięciu zmiennej objaśniającej Nearest wyznaczono wartości residuów w zależności od wartości przewidywanych:

```
Call:
lm(formula = reduced_model_formula, data = dane)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4407 -1.6642 -0.3444  1.2431  5.0482

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.4114369  0.8576631   3.978 0.000525 ***
Area        -0.0020798  0.0009839  -2.114 0.044681 *
Elevation    0.0167745  0.0023387   7.173 1.62e-07 ***
Scruz       -0.0102950  0.0075309  -1.367 0.183784
Adjacent    -0.0034857  0.0007607  -4.583 0.000110 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.733 on 25 degrees of freedom
Multiple R-squared:  0.7802,    Adjusted R-squared:  0.7451
F-statistic: 22.19 on 4 and 25 DF,  p-value: 6.392e-08
```

Dla danych po usunięciu zmiennej objaśniającej Nearest wyznaczono wartości residuów w zależności od wartości przewidywanych:



Wyznaczono również pozostałe statystyki modelu:



Na wykresie reszt punkty są rozproszone losowo wokół linii horyzontalnej na wysokości 0, średnia wariancja reszt jest blisko wartości zerowej. Model ten spełnia więc założenie o liniowości oraz stałej wariancji. Wartość wariancji residuów nie zależy od wartości przewidywanych.

Porównanie współczynników determinacji wszystkich modeli:

```
Współczynnik determinacji modelu pierwotnego:  0.7658469
Współczynnik determinacji modelu spierwiastkowanego:  0.7826977
Współczynnik determinacji modelu zredukowanego:  0.7802496
```

Porównanie zmodyfikowanych współczynników determinacji wszystkich modeli:

```
Zmodyfikowany współczynnik determinacji modelu pierwotnego:  0.7170651
Zmodyfikowany współczynnik determinacji modelu spierwiastkowanego:  0.7374263
Zmodyfikowany współczynnik determinacji modelu zredukowanego:  0.7450896
```

Wartość współczynnika determinacji modelu spierwiastkowanego jest większa niż modelu zredukowanego ponieważ dodanie zmiennej objaśniającej do modelu zawsze zwiększa ten współczynnik niezależnie od przydatności tej zmiennej. Natomiast na podstawie współczynnika skorygowanego współczynnika determinacji możemy zaobserwować, że model ten lepiej opisuje dane po usunięciu obciążającej zmiennej objaśniającej. Współczynnik skorygowany bierze pod uwagę liczbę zmiennych objaśniających i dlatego dla ostatniego modelu przyjmuje on większą wartość. Na podstawie współczynnika skorygowanego model zredukowany daje lepsze dopasowanie do danych niż model pierwotny, nie ma to jednak dużego znaczenia ponieważ nie spełnia on założenia o liniowości oraz stałej wariancji.

Zad 6

Wczytano dane w pliku irys.txt, które zawierają informacje dotyczące 150 kwiatów irysów, które opisano 4 cechami: długość i szerokość patka, oraz długość i szerokość łodygi. Dodatkowo mamy atrybut decyzyjny (class) który przyjmuje 3 możliwe wartości: Iris-setosa, Iris-Versicolor oraz Iris-virginica, które równo dzielą zbiór po 50 obserwacji dla każdej z tych klas.

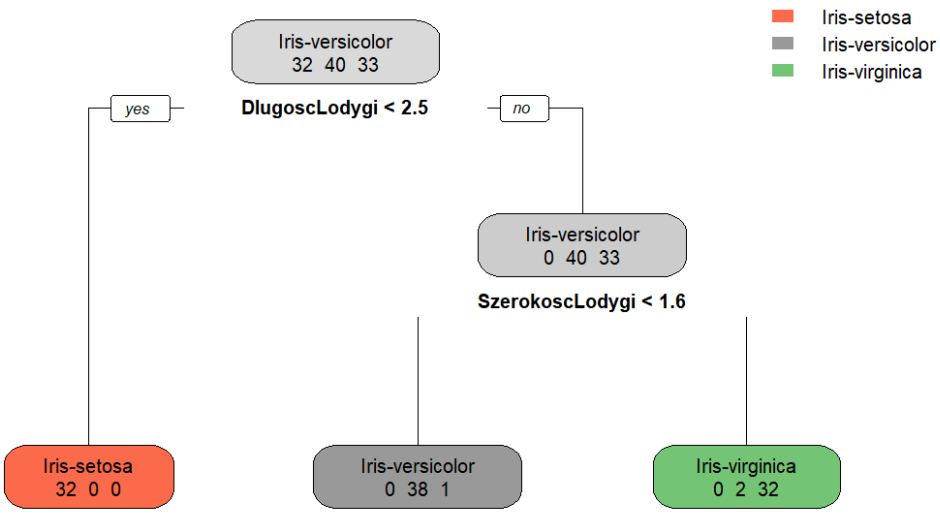
Zbiór danych podzielono na dwa zbiory danych:

- 70% dla zbioru uczącego
- 30% dla zbioru testowego

Poniżej przedstawiono model drzewa decyzyjnego:

```
n= 105

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 105 65 Iris-versicolor (0.30476190 0.38095238 0.31428571)
  2) DlugoscLodygi< 2.45 32  0 Iris-setosa (1.00000000 0.00000000 0.00000000) *
  3) DlugoscLodygi>=2.45 73 33 Iris-versicolor (0.00000000 0.54794521 0.45205479)
    6) SzerokoscLodygi< 1.6 39  1 Iris-versicolor (0.00000000 0.97435897 0.02564103) *
    7) SzerokoscLodygi>=1.6 34  2 Iris-virginica (0.00000000 0.05882353 0.94117647) *
```

Graficzna prezentacja drzewa decyzyjnego:



Reguły drzewa:

```
        Gatunek  Iris Iris Iris
    Iris-setosa [1.00  .00  .00] when DlugoscLodygi <  2.5
Iris-versicolor [ .00  .97  .03] when DlugoscLodygi >= 2.5 & SzerokoscLodygi <  1.6
 Iris-virginica [ .00  .06  .94] when DlugoscLodygi >= 2.5 & SzerokoscLodygi >= 1.6
```

Możemy zaobserwować, że pierwszy gatunek Iris-setosa klasyfikowany jest na podstawie długości łodygi gdy jest ona mniejsza od 2,45 (2,5 w prezentacji graficznej dane zostały zaokrąglone do jednego miejsca po przecinku). Kolejny gatunek Iris-vrsicolor spełnia warunek długość łodygi większa lub równa 2,45 oraz dla kolejnego węzła szerokość łodygi mniejsza od 1,6. Pozostałe obiekty zostały zaklasyfikowane do gatunku Iris-virginica

Błedy:

```
Root node error: 65/105 = 0.61905

n= 105

       CP nsplit rel error   xerror     xstd
1 0.49231      0  1.000000 1.000000 0.076556
2 0.46154      1  0.507692 0.723077 0.078389
3 0.01000      2  0.046154 0.061538 0.030177
```

Błąd dla całego zbioru wynosi 61,91%, natomiast błąd walidacji krzyżowej po pierwszym podziale zmalał do 72,31% wartości pierwotnej, a po drugim podziale jedynie o 6,15%

Macierz błędów dla zbioru testowego:

	prognozy		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	18	0	0
Iris-versicolor	0	7	3
Iris-virginica	0	2	15

Błędy: Trzy tulipany Iris - versicolor zostały zaklasyfikowane jako Iris - virginica oraz dwa tulipany Iris - virginica zostały zaklasyfikowane jako Iris - versicolor.

Liczba błędów modelu:

```
Ilość błędów:  5
```

W zbiorze testowym poprawnie rozpoznano:

```
Procent dobrze rozpoznanych gatunków:  88.88889 %
```

Zad 7

Wczytano plik irys.txt po czym przystąpiono do normalizacji. Dane znormalizowano zgodnie z wzorem:

$$a^{\star} = \frac{a - \min(a)}{\max(a) - \min(a)}$$

Tak przygotowany zbiór danych podzielono na dwa podzbiory danych:

- 70% dla zbioru uczącego
- 30% dla zbioru testowego

Przeprowadzono klasyfikację przy pomocy algorytmu k-NN dla 3-najbliższych sąsiadów.

Powyższą klasyfikację porównano z algorytmami k-NN o innej liczbie najbliższych sąsiadów: 2, 4, 5, 6, 7, 8, 9 i 11. Uzyskaną identyczną klasyfikację:

```
> k2 <- 2
> prognozy2 <- knn(dane_treningowe[, 1:4], dane_testowe[, 1:4], dane_treningowe$Gatunek, k = k2)
> cat("Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji? ", any(prognozy != prognozy2), "\n")
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE
> # Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji
> k2 <- 4
> prognozy2 <- knn(dane_treningowe[, 1:4], dane_testowe[, 1:4], dane_treningowe$Gatunek, k = k2)
> cat("Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji? ", any(prognozy != prognozy2), "\n")
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE
> # Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji
> k2 <- 5
> prognozy2 <- knn(dane_treningowe[, 1:4], dane_testowe[, 1:4], dane_treningowe$Gatunek, k = k2)
> cat("Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji? ", any(prognozy != prognozy2), "\n")
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE
> # Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji
> k2 <- 6
> prognozy2 <- knn(dane_treningowe[, 1:4], dane_testowe[, 1:4], dane_treningowe$Gatunek, k = k2)
> cat("Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji? ", any(prognozy != prognozy2), "\n")
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE

> k2 <- 7
> prognozy2 <- knn(dane_treningowe[, 1:4], dane_testowe[, 1:4], dane_treningowe$Gatunek, k = k2)
> cat("Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji? ", any(prognozy != prognozy2), "\n")
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE
> # Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji
> k2 <- 9
> prognozy2 <- knn(dane_treningowe[, 1:4], dane_testowe[, 1:4], dane_treningowe$Gatunek, k = k2)
> cat("Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji? ", any(prognozy != prognozy2), "\n")
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE
```

```
Czy zmiana liczby sąsiadów powoduje zmianę klasyfikacji?  FALSE
```

Powyższa zależność nie zawsze będzie identyczna, dużo zależy od podziału danych na zbiór uczący oraz testowy. W tym przypadku skorzystano z `set.seed(111)`

Macierz błędów dla zbioru testowego:

	prognozy		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	18	0	0
Iris-versicolor	0	9	1
Iris-virginica	0	2	15

Popełniono mniej błędów niż w przypadku drzewa decyzyjnego:

```
Ilość błędów:  3
```

Błędy: Jeden tulipany Iris - versicolor zostały zaklasyfikowane jako Iris - virginica oraz dwa tulipany Iris - virginica zostały zaklasyfikowane jako Iris - versicolor.

W zbiorze testowym poprawnie rozpoznano:

```
Dokładność klasyfikatora:  93.33333 %
```