PROJEKT 2: PERSONALNE KOSZTY LECZENIA

Oliwia Stebelska, Gabriela Sumera, Weronika Pudło

1. **Sformułowanie celu projektu i stworzenie modelu reprezentującego badany problem.**

CHARAKTERYSTYKA ZBIORU:

Analizujemy zbiór, zawierający następujące dane dotyczące świadczeniobiorców korzystających z usług medycznych na terenie Stanów Zjednoczonych:

1. **Age:** zmienna ilościowa: wiek
2. **Sex:** zmienna jakościowa: płeć
3. **Bmi:** zmienna ilościowa: wskaźnik bmi
4. **Children:** zmienna ilościowa: ilość dzieci pokrytych ubezpieczeniem zdrowotnym.
5. **Smoker:** zmienna jakościowa: czy występuje nałóg w postaci palenia papierosów.
6. **Region:** zmienna jakościowa: region zamieszkania
7. **Charges:** zmienna ilościowa, objaśniana: personalny koszt leczenia.

CEL PROJEKTU:

Zbadanie poprzez analizę statystyczną zależności pomiędzy kosztem leczenia a cechami osoby, która podejmuje leczenie (świadczeniobiorcy).

Korzystając z danych oraz uzyskanych wniosków będziemy prognozować całkowite koszty leczenia pacjenta, w zależności od jego sytuacji życiowej.

PODSUMOWANIE DANYCH:

```
      age            sex            bmi          children     smoker
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
 Median :39.00                Median :30.40   Median :1.000
 Mean   :39.21                Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                Max.   :53.13   Max.   :5.000
       region       charges
 northeast:324   Min.   : 1122
 northwest:325   1st Qu.: 4740
 southeast:364   Median : 9382
 southwest:325   Mean   :13270
                 3rd Qu.:16640
                 Max.   :63770
```

Wybrany przez nas zbiór charakteryzował się brakiem pustych pól oraz zmiennych odstających.

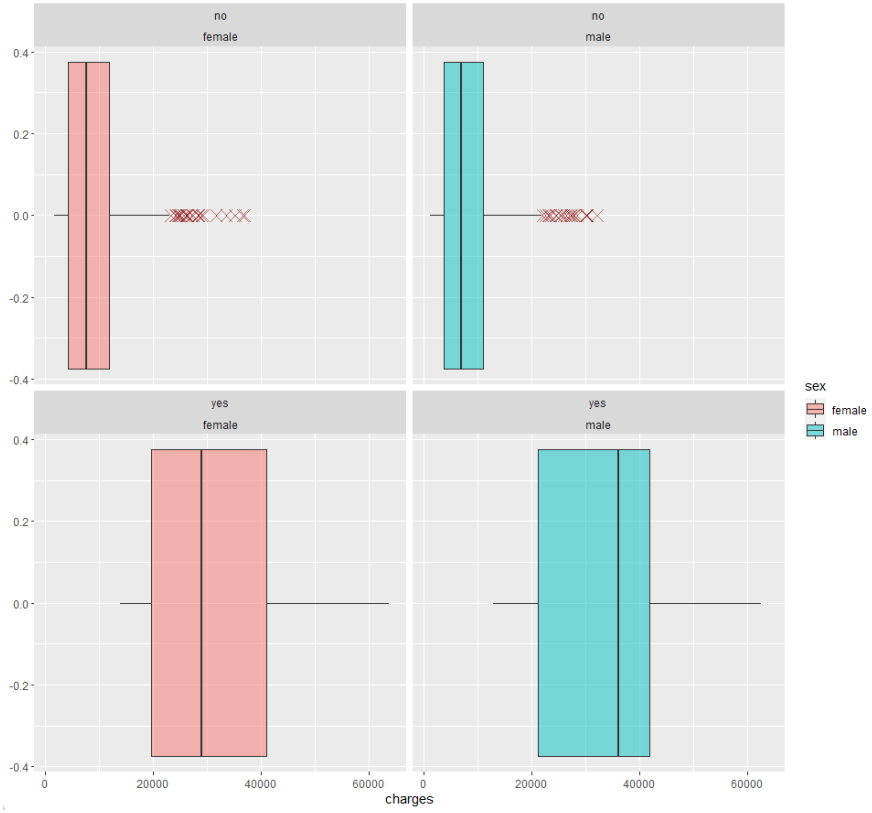
2. **Statystyczny opis struktury analizowanych cech, reprezentowanych przez zmienne liczbowe.**

STATYSTYKI OPISOWE DLA ZMIENNEJ CHARGES:

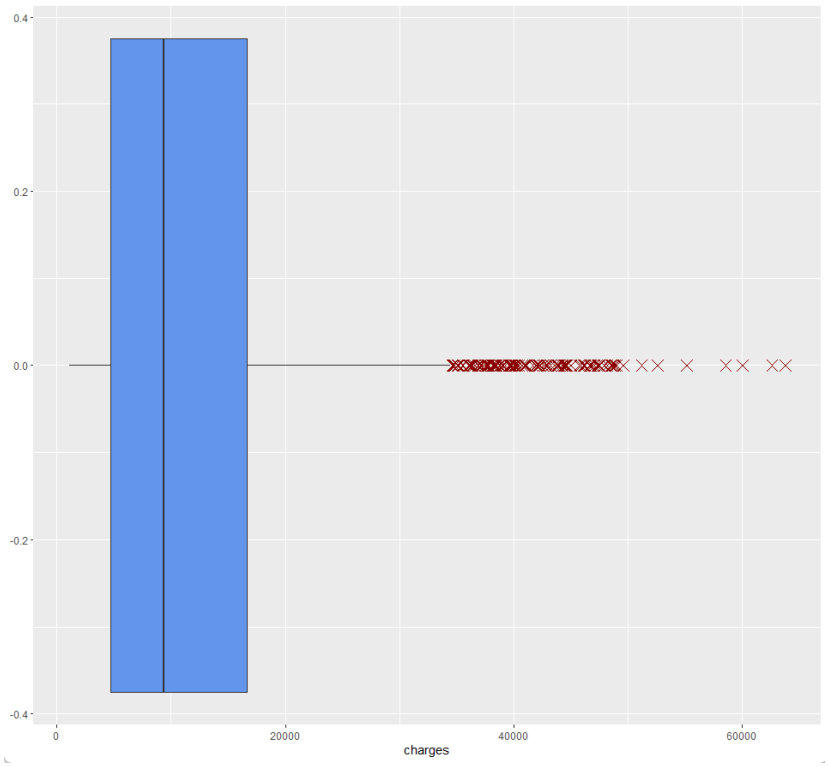
	srednia	odch.std	wariancja	Me	Mo	As	K
1	13270.42	12110.01	146652372	9382.033	1639.563	1.51418	4.595821

	kwantyle
0%	1121.874
25%	4740.287
50%	9382.033
75%	16639.913
100%	63770.428

	zakres
1	1121.874
2	63770.428



Rysunek 1 Skategoryzowany wykres ramka wąsy.



Rysunek 2 Wykres ramka wąsy.

CHARAKTERYSTYKA:

Średni koszt leczenia wynosi **13270.42$**, a **odchylenie standardowe** wynosi **12110.01$.** Statystyki ukazują duże zróżnicowanie wyników w naszym zbiorze.

Mediana wynosi **9382.033$.** Przez dużą wartość odchylenia standardowego mediana jest lepszą miarą w wypadku występowania dużych wartości odstających.

Patrząc na **rysunek 2** widoczne są liczne wartości odstające, które w sposób znaczący mogą zaburzyć interpretację danych. Można również zauważyć, że wpływają one między innymi na różnicę między średnią a medianą.

Minimalna, maksymalna wartość, kwantyl dolny i górny

Uwzględniając wszystkie dane można zauważyć, że różnica pomiędzy minimalną a maksymalną kwotą jest bardzo duża.

- **Przy 0 percentylu** dostajemy wartość równą 1121.874$. Oznacza to, że 100% kosztów jest równych lub większych niż 1121.874$, jest to nasze **minimum – kwantyl dolny.**
- **Przy 25 percentylu** dostajemy wartość równą 4740.287$. Oznacza to, że 25% kosztów jest mniejszych niż 4740.287$, a 75% kosztów jest równych lub większych niż 4740.287$.
- **Przy 50 percentylu** dostajemy wartość równą 9382.033$. Oznacza to, że 50% kosztów jest mniejszych niż 720 zł, a 50% kosztów jest równych lub większych niż 9382.033$. Jest to wartość mediany.

- **Przy 75 percentylu** dostajemy wartość równą 16639.913$. Oznacza to, że 75% kosztów jest mniejszych niż 16639.913$, a 25% kosztów jest równych lub większych niż 16639.913$.
- **Przy 100 percentylu** dostajemy wartość 63770.428$. Oznacza to, że wszystkie koszta są mniejsze bądź równe 63770.428$. Jest to nasze **maksimum – kwantyl górny.**

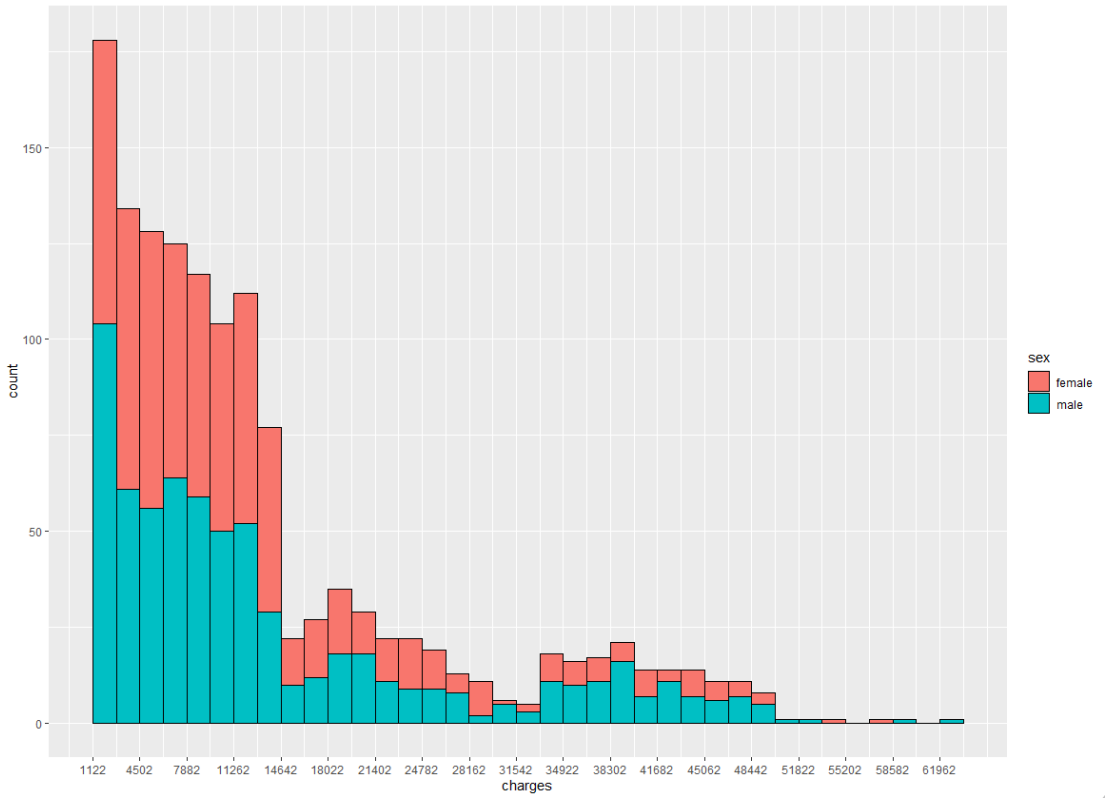
W nawiązaniu do wykresu ramka-wąsy, różnica pomiędzy medianą a wartością minimalną, jest zdecydowanie mniejsza niż pomiędzy medianą a wartością maksymalną.

Interpretacja skośności i kurtozy:

Poniższa interpretacja graficzna danych dowodzi, że mamy do czynienia z rozkładem **prawostronnie skośnym**. Potwierdzeniem jest również **miara skośności**, która określa kierunek i siłę asymetrii, dla zmiennej charges **wynosi 1.51418**, więc mamy do czynienia z rozkładem, który ma dłuższy prawy „ogon". Rozkład charakteryzuje się tym, że wartość średnia jest większa od mediany (co jest zgodne z naszymi wcześniejszymi statystykami opisowymi).

Kurtoza koncentruje się wokół średniej, gdy jest mniejsza od 0 – rozkład jest bardziej spłaszczony od normalnego a większa gdy rozkład jest bardziej wysmukły. Dla zmiennej *charges* kurtoza równa jest **4.595821**, sugeruje to występowanie **rozkładu wysmukłego.**

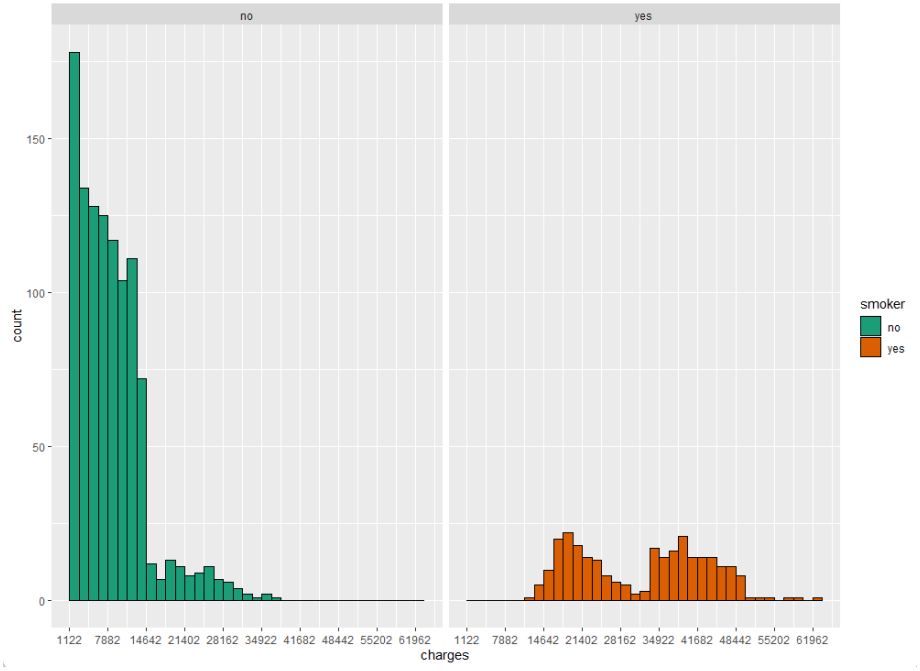
HISTOGRAM DLA ZMIENNEJ CHARGES Z PODZIAŁEM NA PŁEĆ:



Rysunek 3: Histogram 1

Przedstawione dane na histogramie informują nas o **kosztach leczenia** ze względu na **liczbę pacjentów z podziałem na płeć**. Zauważamy, że liczba kobiet jest porównywalna do liczby mężczyzn. Możemy wywnioskować, że płeć nie ma znacznego wpływu na koszty leczenia.

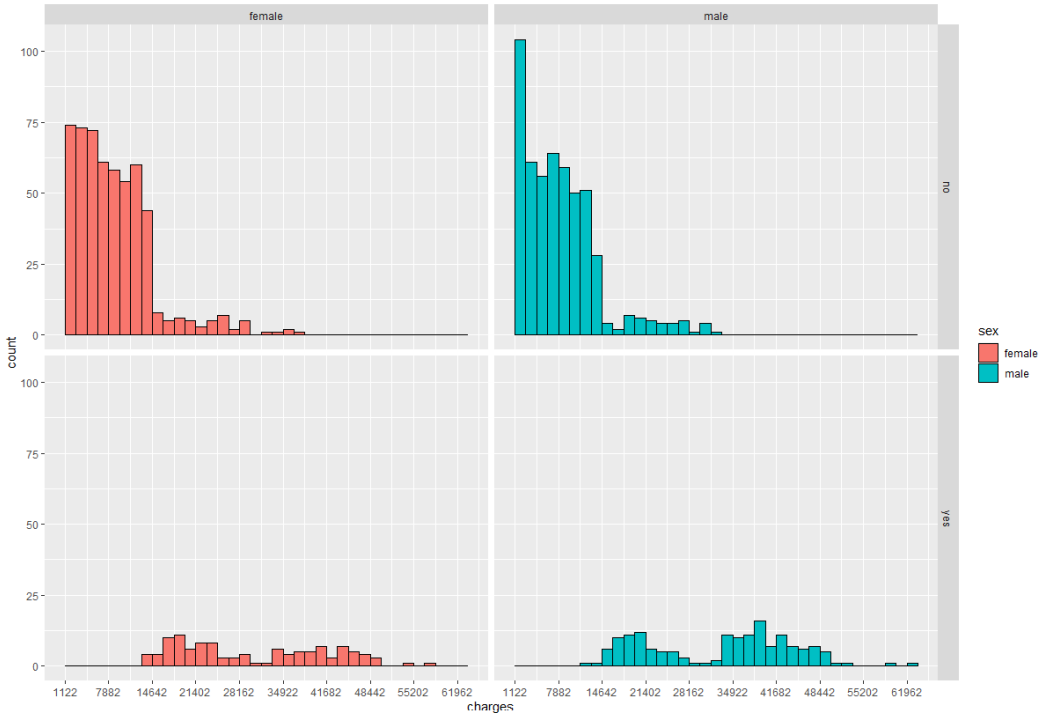
HISTOGRAM DLA ZMIENNEJ CHARGES Z PODZIAŁEM NA PALENIE:



Rysunek 4: Histogram 2

Przedstawione dane na 2 histogramie informują nas o liczbie pacjentów z podziałem na występowanie nałogu tytoniowego. Zauważamy, że pacjenci niepalący skupiają się w przedziale niższych kosztów leczenia, niż pacjenci palący, warto podkreślić, że liczba osób niepalących zdecydowania przewyższa liczbę osób palących biorących udział w badaniu.

HISTOGRAM SKATEGORYZOWANY W ZALEŻNOŚCI OD PŁCI I PALENIA:



Rysunek 5: Histogram 3

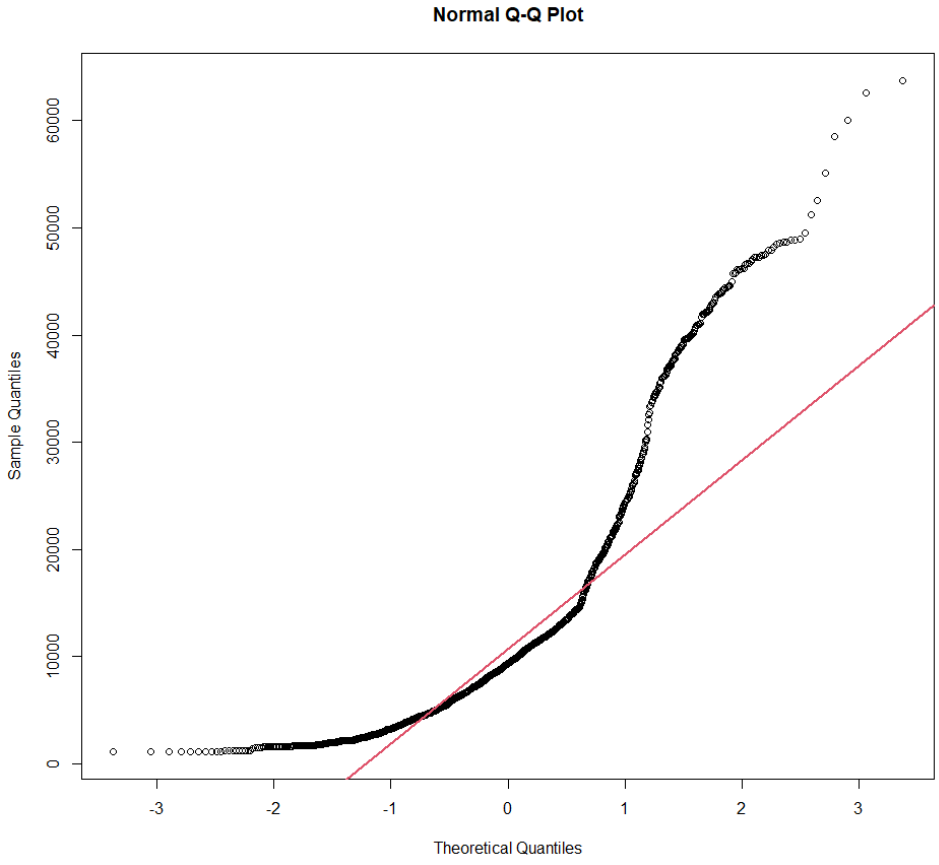
3. WNIOSKOWANIE STATYSTYCZNE:

Przedziały ufności dla dziewięciu losowych wartości zmiennej charges równych: [10355.641, 24227.337, 17352.680, 41097.162, 17179.522, 12928.791, 21978.677, 5253.524, 20177.671.]

```
> 18950.11-10200.64/sqrt(9)*qnorm(.975)
[1] 12285.81
> 18950.11+10200.64/sqrt(9)*qnorm(.975)
[1] 25614.41
```

$$P(\mathbf{12285.81} < \boldsymbol{\mu} < \mathbf{25614.41}) = 1 - \alpha = 0.95$$

Weryfikacja hipotezy o zgodności empirycznego rozkładu wybranej cechy z rozkładem normalnym:




```
        Shapiro-Wilk normality test

data:  dane$charges
W = 0.81469, p-value < 2.2e-16
```

H0: Zmienna charges ma rozkład normalny.

H1: Zmienna charges nie ma rozkładu normalnego.

p-value < 2.2e-16, $\alpha = 0{,}05$

p-value < α, mamy podstawy do odrzucenia H0, zmienna charges **nie ma rozkładu normalnego**.

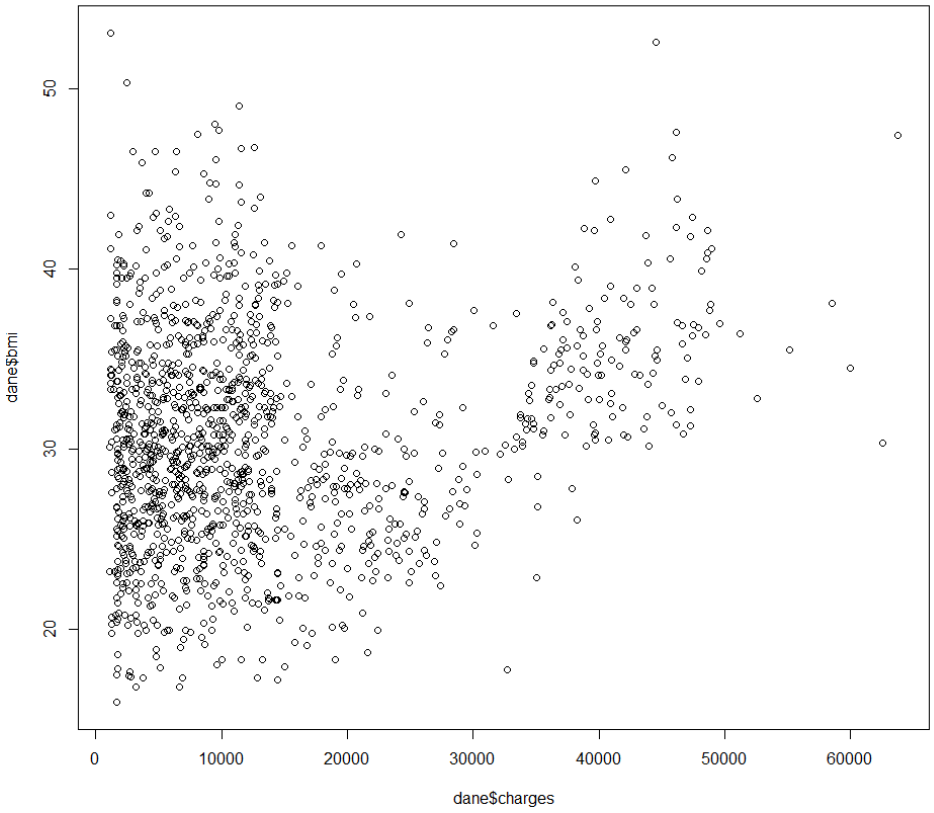
Związek korelacyjny pomiędzy badanymi zmiennymi:

Współczynnik korelacji dla zmiennych **charges oraz bmi** wynosi **0.198341**, co oznacza że występuje zależność liniowa rosnąca oraz korelacja słaba.

```
> cor.test(dane$charges,dane$bmi)

        Pearson's product-moment correlation

data:  dane$charges and dane$bmi
t = 7.3966, df = 1336, p-value = 2.459e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1463052 0.2492822
sample estimates:
      cor
0.198341
```

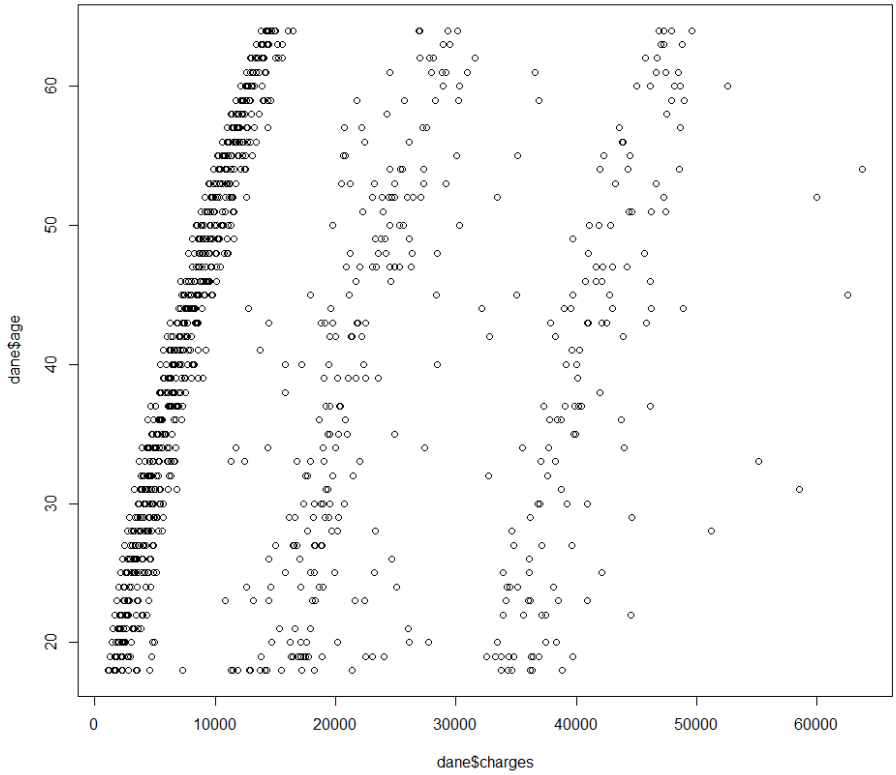


Współczynnik korelacji dla zmiennych **charges oraz age** wynosi **0.2990082**, co oznacza że również występuje zależność liniowa rosnąca oraz korelacja słaba.

```
> cor.test(dane$charges,dane$age)

        Pearson's product-moment correlation

data:  dane$charges and dane$age
t = 11.453, df = 1336, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2494139 0.3470381
sample estimates:
      cor
0.2990082
```



MODEL REGRESJI:

```
Call:
lm(formula = charges ~ ., data = dane2)

Residuals:
   Min     1Q Median     3Q    Max
-13884  -6994  -5092   7125  48627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6916.24    1757.48  -3.935 8.74e-05 ***
age           239.99      22.29  10.767  < 2e-16 ***
bmi           332.08      51.31   6.472 1.35e-10 ***
children      542.86     258.24   2.102   0.0357 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11370 on 1334 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.1181
F-statistic: 60.69 on 3 and 1334 DF,  p-value: < 2.2e-16
```

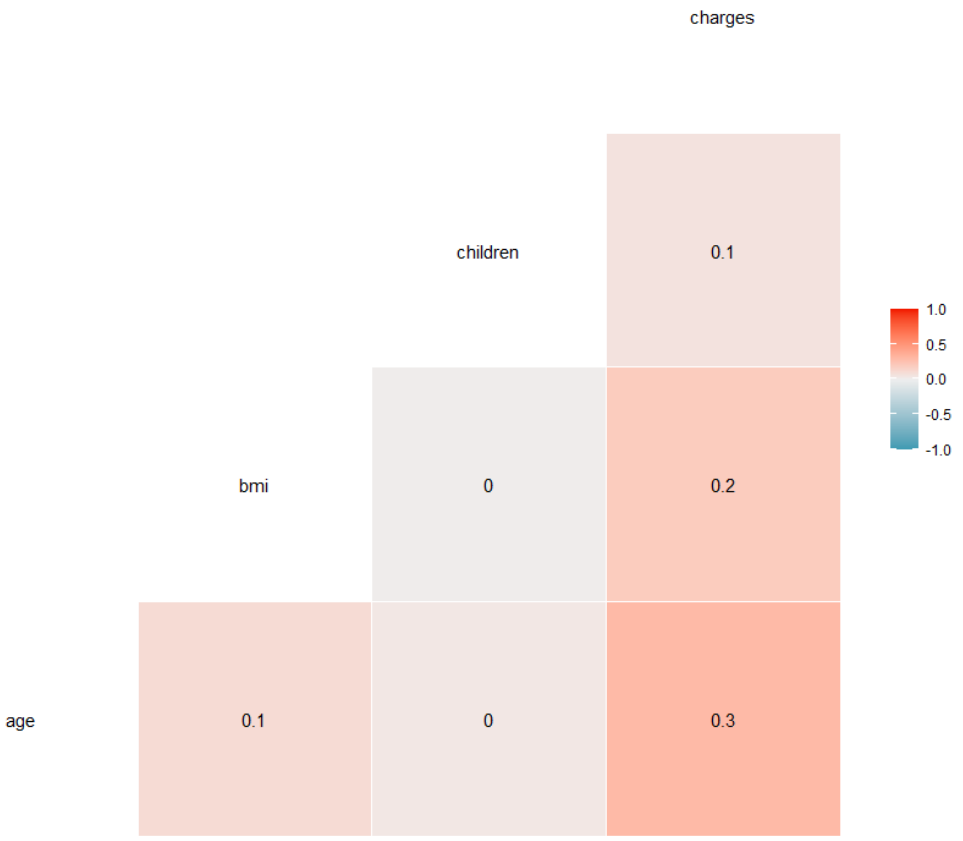
RÓWNANIE REGRESJI WIELORAKIEJ:

CHARGES = 239.9945*AGE + 332.0834*BMI + 542.8647*CHILDREN – 6916.2433

```
> model$coefficients
(Intercept)         age         bmi    children
 -6916.2433    239.9945    332.0834    542.8647
>
```

Na zmienną charges oddziałuje **więcej niż jedna zmienna**, określamy więc macierz korelacji:

MACIERZ KORELACJI:

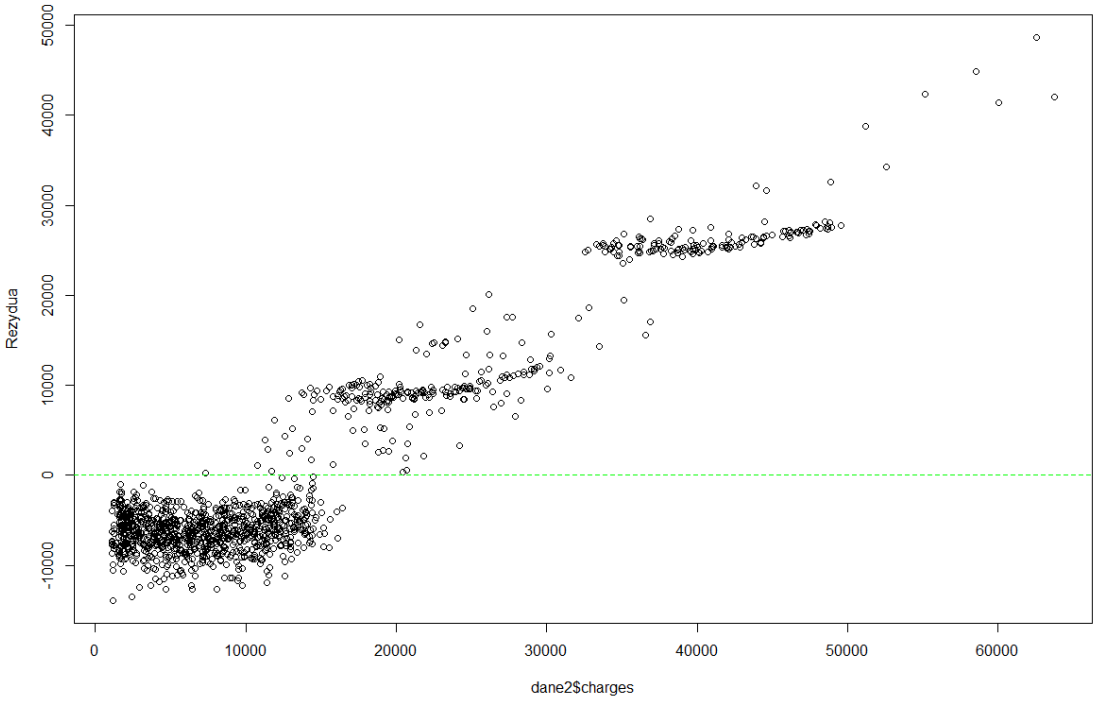


ISTOTNOŚĆ KORELACJI:

```
> cbind(cor.test(dane2$charges,dane2$age)$p.value,
+       cor.test(dane2$charges,dane2$bmi)$p.value,
+       cor.test(dane2$charges,dane2$children)$p.value)
             [,1]         [,2]       [,3]
[1,] 4.886693e-29 2.459086e-13 0.01285213
> |
```

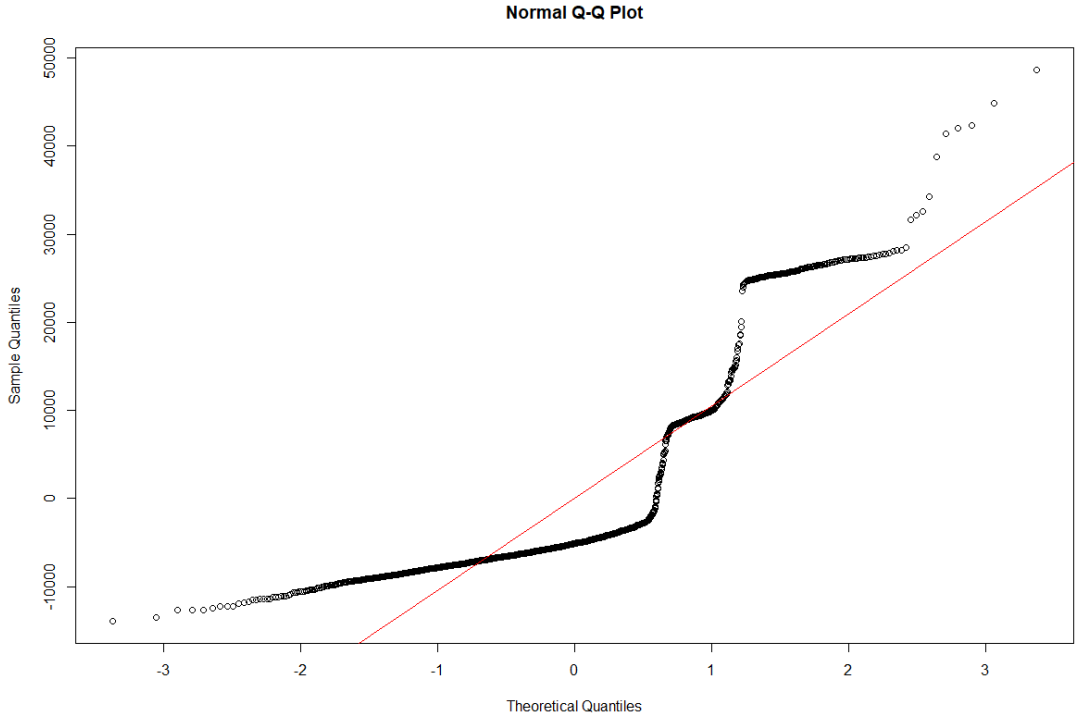
ANALIZA RESZT:

1. LOSOWOŚĆ ODCHYLEŃ:



Z powyższego wykresu możemy odczytać, że reszty układają się poniżej i powyżej krzywej teoretycznej, lecz nie w sposób losowy.

2. NORMALNOŚĆ ROZKŁADU RESZT:



```
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.75201, p-value < 2.2e-16
```

H0: Reszty mają rozkład normalny.

H1: Reszty nie mają rozkładu normalnego.

P-value < α<0,05: odrzucamy H0**, reszty nie mają rozkładu normalnego**.

3. NIEOBCIĄŻNOŚĆ RESZT:

```
> mean(model$residuals)
[1] 4.284631e-13
```

E(εi) ≠ 0, wartość oczekiwana reszt **nie jest równa 0.**

4. HOMOSCEDASTYCZNOŚĆ:

```
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 134.26, df = 3, p-value < 2.2e-16
```

H0: Występuje stałe rozproszenie reszt.

H1: Występuje heteroscedastyczność.

P-value < α<0,05: odrzucamy H0, nie występuje stałe rozproszenie reszt, **występuje heteroscedastyczność.**

5. NADMIAROWOŚĆ:

Współczynnik współliniowości:

```
> vif(model)
     age      bmi children
1.013816 1.012152 1.001874
```

Wszystkie zmienne mieszczą się w przedziale: 1<VIF<10, co oznacza **nieznaczną współliniowość predyktorów**, w tym przypadku wartości są **bardzo zbliżone do 1.**

INTERPRETACJA MODELU:

- Współczynnik determinacji: R^2**= 0.1201.**
- **12%** zmienności zmiennej *charges* wyjaśniono przez model regresji wielorakiej.
- Korelacja istnieje, lecz jest dosyć **słaba.**
- Korelacje są dodatnie, wraz z ich wzrostem, wzrasta zmienna charges.
- Współczynniki korelacji są bardzo małe, wykazują korelacje nikłą.
- Każdy **wzrost wieku o 5 lat** oznacza zwiększenie kosztów leczenia o około **1200$,** przy niezmieniającej się wartości bmi oraz liczbie dzieci.
- Każdy **spadek wskaźnika bmi o jedną jednostkę** oznacza spadek kosztu leczenia o **332.0834$,** przy niezmieniającym się wieku oraz liczbie dzieci.
- Każdy **wzrost liczby dzieci** objętych ubezpieczeniem zdrowotnym o jedno, zwiększa koszty leczenia **o 542.8647$,** przy niezmieniającym się wieku oraz wartości wskaźnika bmi.

PREDYKCJE:

1. Predykcja kosztów leczenia w zależności od **wieku pacjenta**:

```
> nowe <- data.frame(age=26, bmi=25, children =2)
> predict(model,nowe)
       1
8711.426
> nowe <- data.frame(age=55, bmi=25, children =2)
> predict(model,nowe)
       1
15671.27
```

Dla pacjenta w wieku 26 lat koszty leczenia będą prawie 2 razy niższe niż dla pacjenta starszego, o tym samym wskaźniku bmi oraz liczbie dzieci pokrytych opieką zdrowotną.

2. Predykcja kosztów leczenia w zależności od **wskaźnika bmi**:

```
> nowe <- data.frame(age=40, bmi=37, children =0)
> predict(model,nowe)
       1
14970.62
> nowe <- data.frame(age=40, bmi=20, children =0)
> predict(model,nowe)
       1
9325.203
```

Pacjent o bmi równym 37 zapłaci za leczenie o 5000$ więcej od pacjenta z niższym bmi, równym 20.

3. Predykcja kosztów leczenia w zależności od **liczby dzieci:**

```
> nowe <- data.frame(age=30, bmi=22, children =1)
> predict(model,nowe)
      1
8132.29
> nowe <- data.frame(age=30, bmi=22, children =4)
> predict(model,nowe)
       1
9760.884
```

Pacjent posiadający jedno dziecko zapłaci za leczenie niewiele mniej niż pacjent posiadający czwórkę dzieci objętych ubezpieczeniem zdrowotnym.

PODSUMOWANIE WYNIKÓW ANALIZ:

Dane dla których opracowywałyśmy analizę statystyczną dotyczyły kosztów leczenia dla pacjenta o zdefiniowanych wcześniej cechach.

Zmienną objaśniającą, dla której prowadziliśmy analizę statystyczną była *charges*, czyli powyżej wspomniany koszt leczenia.

Jedną ze wspomnianych cech jest uzależnienie od wyrobów tytoniowych. Patrząc na wszystkie przedstawione powyżej wykresy, pacjent palący statystycznie będzie mieścił się w przedziale wyższych kosztów leczenia, niż w przypadku osoby niepalącej. Możemy domniemywać, że ma to dużo wspólnego z kondycją zdrowia, które jest narażone na znaczne pogorszenie przez stosowanie używki.

Kolejną analizowaną cechą jest wskaźnik BMI, jest to stosunek masy ciała do wzrostu do kwadratu. U analizowanej przez nas grupie pacjentów minimalne BMI wynosiło 15,96 co według tabeli BMI wskazuje na lekką niedowagę, a maksymalne BMI wynosiło 53,13 co wskazuje na poważną otyłość. Średnia wartość wskaźnika BMI (30,66) według tabeli wskazuje na otyłość. Analizując zbiór, możemy zauważyć, że przeciętna badana osoba ma nadwagę, co wpływa na wzrost kosztów leczenia przy rosnącym wskaźniku BMI.

Ostatnią analizowaną zmienną jest wiek pacjenta. Z analizy wynika, że im starszy jest pacjent, tym więcej zapłaci za koszty leczenia, może być to związane z występowaniem chorób pojawiających się na starość.