# Задание

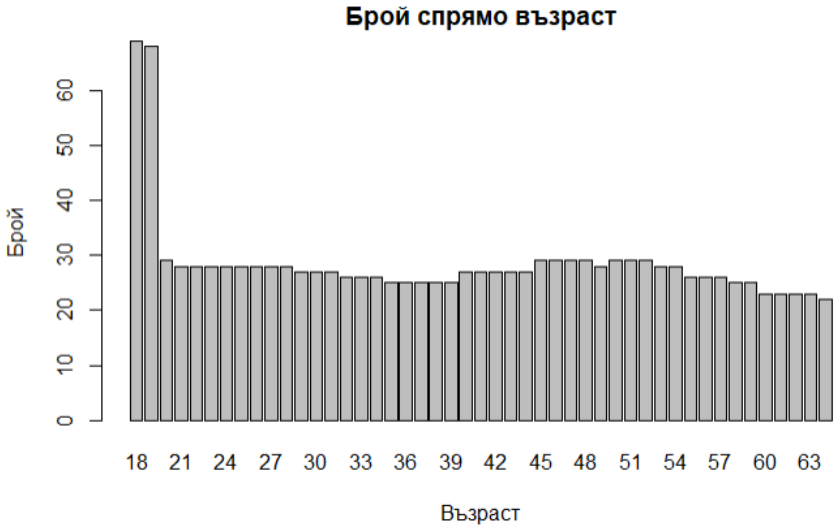
Да се извлече информация подходяща за обработка с линейна регресия. Данните да се анализират с апарата на линейната регресия и да се направят изводи по темата.

# Цел на проекта

Да се създаде линеен модел, който да предсказва каква е нужната застраховка на човек, ако се знаят неговите пол, възраст, регион, брой деца, BMI (body mass index) и дали е пушач. Да се направят изводи за това, кои са основните фактори при които се определя медицинския разход за един човек спрямо използваните данни.

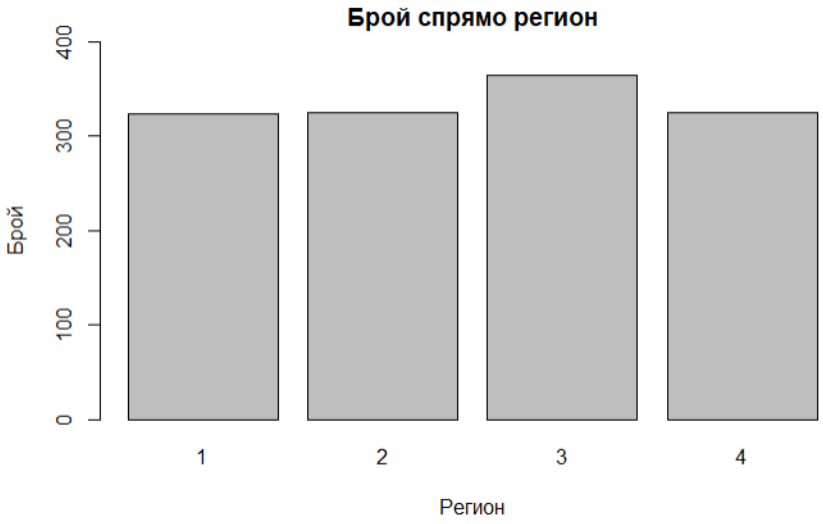
# Източник на данни

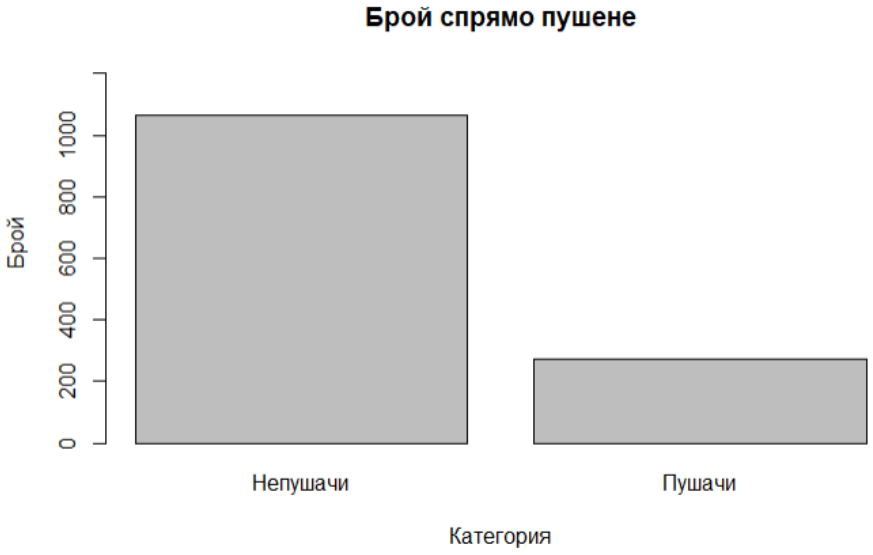
Данните са извлечени от платформата Keggle.com. Извадката има 1338 наблюдения, като са посочени в тях пол, възраст, регион, брой деца, BMI (body mass index) и дали е пушач.  Общ преглед на данните.

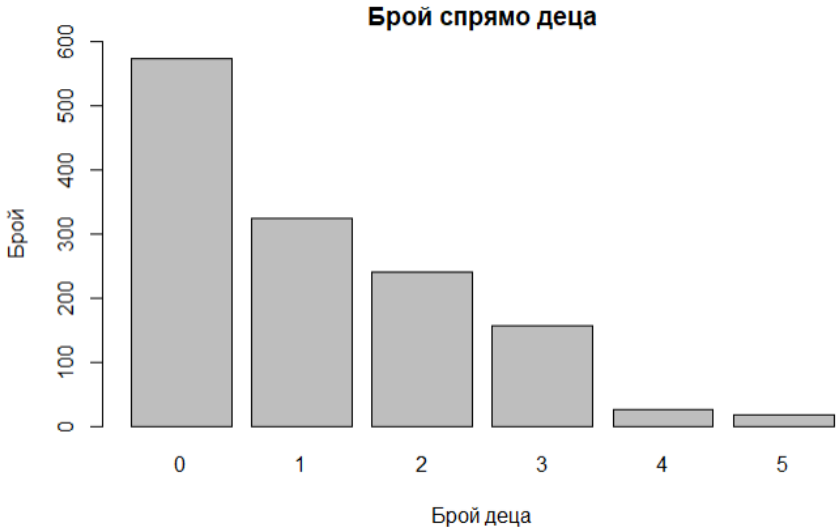



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00   27.00   39.00   39.21   51.00   64.00
```

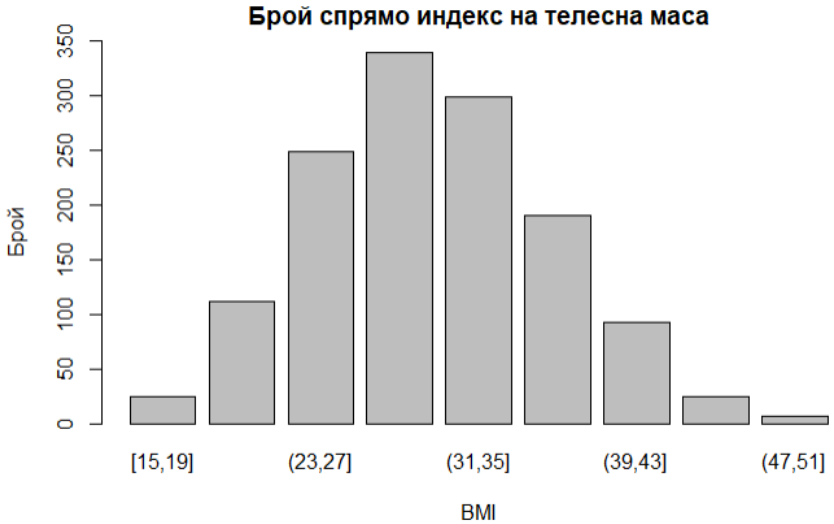

Брой спрямо пол
Брой
0
100
200
300
400
500
600
700
1
2
Пол


Брой спрямо регион
Брой
0
100
200
300
400
1
2
3
4
Регион

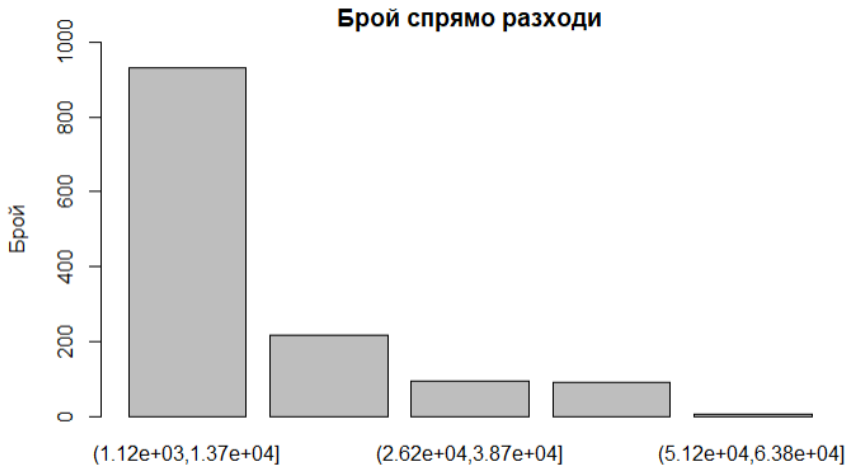

Брой спрямо пушене
Брой
0
200
400
600
800
1000
Непушачи
Пушачи
Категория




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.095   2.000   5.000
```




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.96   26.30   30.40   30.66   34.69   53.13
```




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1122    4740    9382   13270   16640   63770
```

# Разработка на проекта

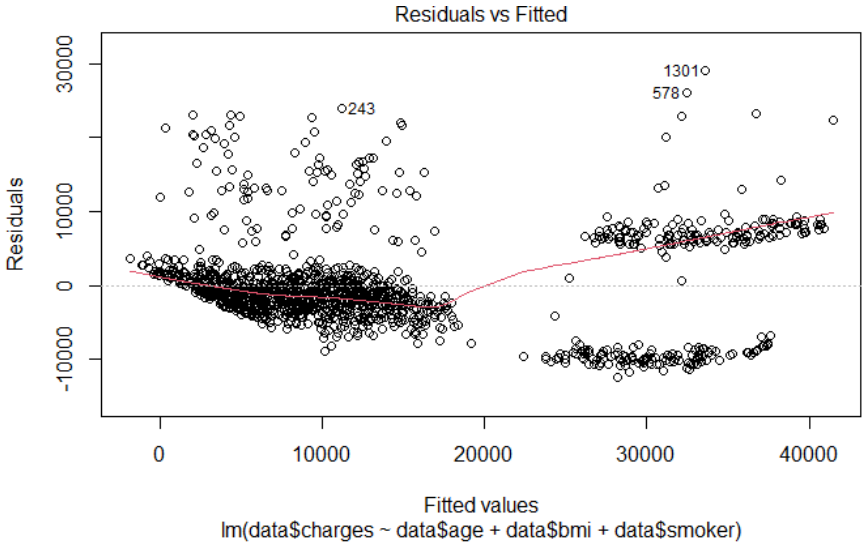
Данните се извличат и се форматират подходящо – нечислените данни ги преобразуваме в числени.

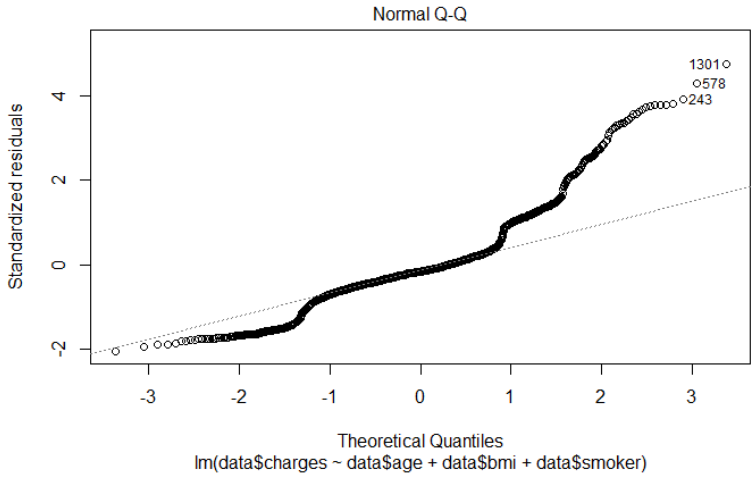
Преди да създаването на проекта преглеждаме корелациите между различните данни, за да определим, кои са най-подходящи за използване за модела.

```
                age          sex         bmi    children       smoker       region      charges
age      1.000000000 -0.020855872 0.109271882 0.04246900 -0.025018752  0.002127313  0.299008193
sex     -0.020855872  1.000000000 0.046371151 0.01716298  0.076184817  0.004588385  0.057292062
bmi      0.109271882  0.046371151 1.000000000 0.01275890  0.003750426  0.157565849  0.198340969
children 0.042468999  0.017162978 0.012758901 1.00000000  0.007673120  0.016569446  0.067998227
smoker  -0.025018752  0.076184817 0.003750426 0.00767312  1.000000000 -0.002180682  0.787251430
region   0.002127313  0.004588385 0.157565849 0.01656945 -0.002180682  1.000000000 -0.006208235
charges  0.299008193  0.057292062 0.198340969 0.06799823  0.787251430 -0.006208235  1.000000000
```

В случая особено влияние оказва пушенето, възрастта и BMI. Създаваме модел спрямо тези предиктори, който ще ни служи за сравнение по-късно.

```
Call:
lm(formula = data$charges ~ data$age + data$bmi + data$smoker)

Residuals:
     Min       1Q   Median       3Q      Max
-12415.4  -2970.9   -980.5   1480.0  28971.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -35500.51    1060.50  -33.48   <2e-16 ***
data$age       259.55      11.93   21.75   <2e-16 ***
data$bmi       322.62      27.49   11.74   <2e-16 ***
data$smoker  23823.68     412.87   57.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6092 on 1334 degrees of freedom
Multiple R-squared:  0.7475,    Adjusted R-squared:  0.7469
F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```



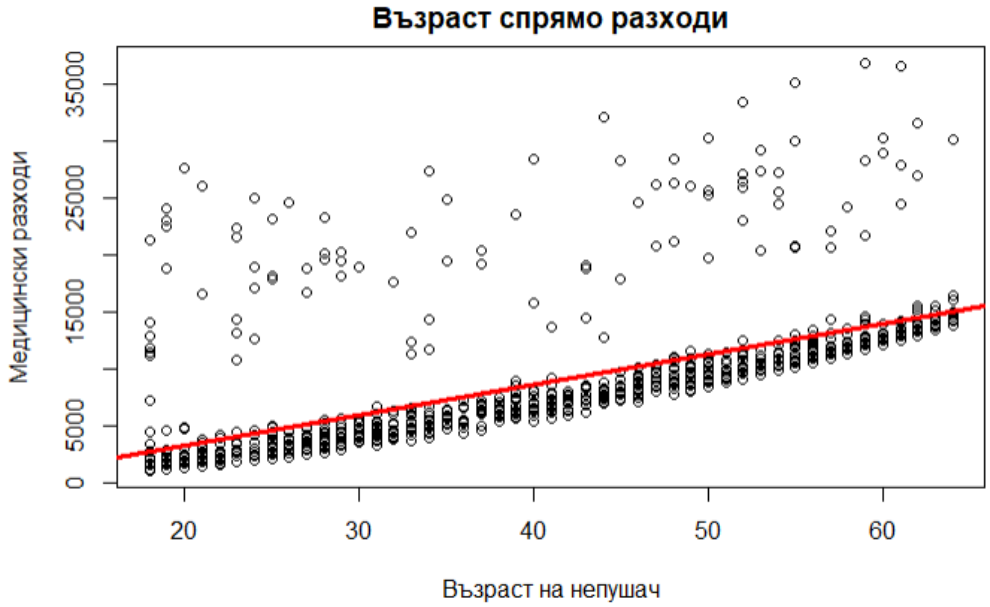
От първоначалния модел може да се забележи, че не приближава много лошо данните. От Q-Q plot-а може да се установи, че в модела има повече от едно разпределение (Glen_b, 2014). Нека разделим извадката на две части – пушачи и непушачи. Да разгледаме корелацията на извадката без пушачи.

| | age | sex | bmi | children | region | charges |
|---|---|---|---|---|---|---|
| age | 1.00000000 | -0.022338060 | 0.12263798 | 0.033395332 | 0.01160057 | 0.62794678 |
| sex | -0.02233806 | 1.000000000 | 0.01911866 | 0.002090212 | -0.01057069 | -0.05632016 |
| bmi | 0.12263798 | 0.019118656 | 1.00000000 | 0.019208413 | 0.14341283 | 0.08403654 |
| children | 0.03339533 | 0.002090212 | 0.01920841 | 1.000000000 | 0.01693369 | 0.13892870 |
| region | 0.01160057 | -0.010570687 | 0.14341283 | 0.016933689 | 1.00000000 | -0.07331625 |
| charges | 0.62794678 | -0.056320161 | 0.08403654 | 0.138928705 | -0.07331625 | 1.00000000 |

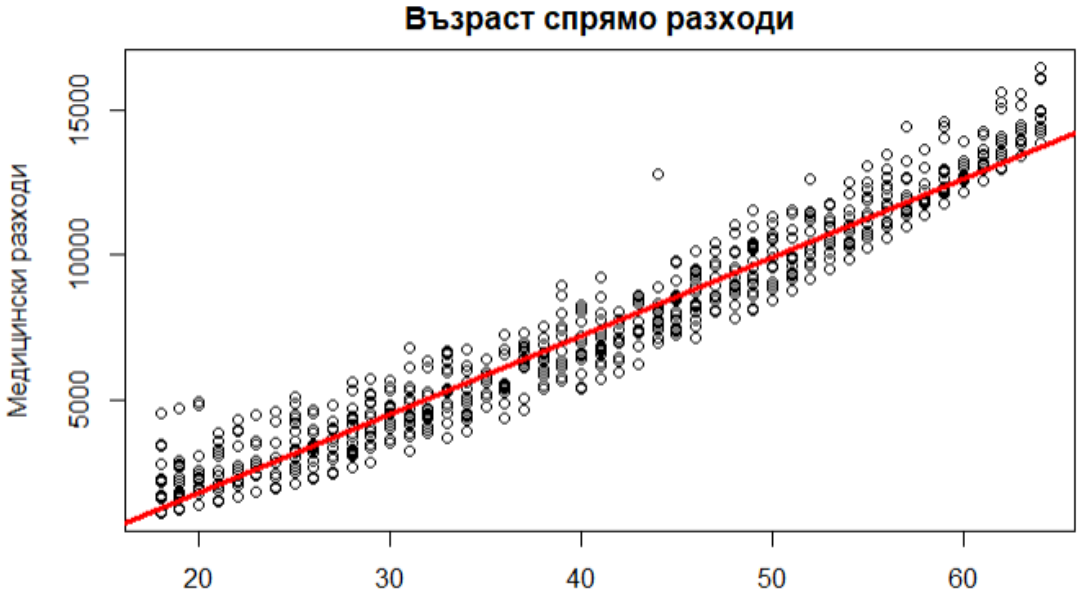
Тук важна роля играе възрастта и за това само нея ще използваме като предиктор. Създадения модел има видя:

```
Call:
lm(formula = data2$charges ~ data2$age)

Residuals:
    Min      1Q  Median      3Q     Max
-3182.9 -1948.6 -1363.8  -665.2 24470.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2091.42     425.10   -4.92    1e-06 ***
data2$age     267.25      10.16   26.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4667 on 1062 degrees of freedom
Multiple R-squared:  0.3943,    Adjusted R-squared:  0.3937
F-statistic: 691.4 on 1 and 1062 DF,  p-value: < 2.2e-16
```

Този модел има една идея по-добро поведение спрямо първоначални, но отново има съществени разминавания между предсказаните от модела стойности и тези на извадката. Това може да го забележим и от графиката:




Може да забележим, че повечето хора попадат, върху регресионната права, но една част са доста извън нея. Това не се дължи на фактор породен от извадката, който не сме включили в модела, следователно може да считаме, че е външен фактор и да изключим тези стойности. След премахване на излишните данни новия модел придобива вида:



```
Call:
lm(formula = data222$charges ~ data222$age)

Residuals:
    Min      1Q  Median      3Q     Max
-1795.8  -594.3   -66.7   499.2  4508.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3558.801     80.706   -44.1   <2e-16 ***
data222$age   269.257      1.929   139.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 841.4 on 964 degrees of freedom
Multiple R-squared:  0.9529,    Adjusted R-squared:  0.9528
F-statistic: 1.948e+04 on 1 and 964 DF,  p-value: < 2.2e-16
```

Който вече има много по-хубаво поведение, от предишните два модела до сега.

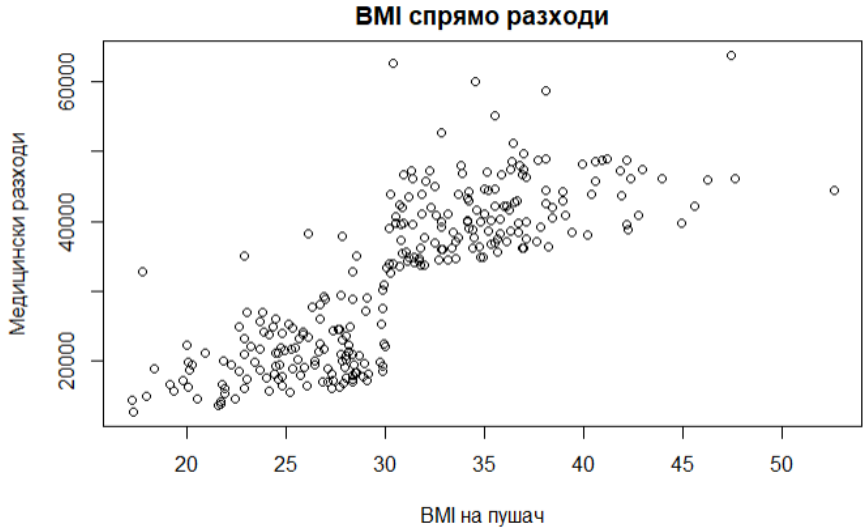
Да се върнем към данните касаещи пушачите. Нека видим при тях кои са основните фактори за медицински разходи.

| | age | sex | bmi | children | region | charges |
|---|---|---|---|---|---|---|
| age | 1.000000000 | -0.005758137 | 0.05967388 | 0.08118329 | -0.03654818 | 0.36822444 |
| sex | -0.005758137 | 1.000000000 | 0.14834993 | 0.07690698 | 0.06694318 | 0.10122606 |
| bmi | 0.059673882 | 0.148349931 | 1.00000000 | -0.01261916 | 0.21219790 | 0.80648061 |
| children | 0.081183289 | 0.076906981 | -0.01261916 | 1.00000000 | 0.01512611 | 0.03594501 |
| region | -0.036548182 | 0.066943177 | 0.21219790 | 0.01512611 | 1.00000000 | 0.12898348 |
| charges | 0.368224444 | 0.101226064 | 0.80648061 | 0.03594501 | 0.12898348 | 1.00000000 |

При тях се забелязва, че възрастта и bmi оказват основно влияние. Нека разгледаме модел, който приближава медицинския разход по възраст и bmi.

```
Call:
lm(formula = data1$charges ~ data1$age + data1$bmi)

Residuals:
     Min       1Q   Median       3Q      Max
-14604.4  -4315.1   -240.5   3638.0  29316.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22367.45    1931.86  -11.58   <2e-16 ***
data1$age      266.29      25.06   10.63   <2e-16 ***
data1$bmi     1438.09      55.22   26.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5754 on 271 degrees of freedom
Multiple R-squared:  0.7532,    Adjusted R-squared:  0.7514
F-statistic: 413.6 on 2 and 271 DF,  p-value: < 2.2e-16
```

Може да подобрим модела на същия принцип на, който подобрихме и оригиналния. Да разгледаме зависимостите между предикторите и разходите.


Забелязва се оформяне на две групи при възраст/разходи. Тя се дължи на факта, че висок индекс на телесна маса води до по-високи разходи – bmi>30 се счита за затлъстяване, както показва и втората графика. За това нека разделим извадката на две части – тези с наднормено тегло и тези с нормално. Тези с поднормено тегло са малко и може да ги премахнем без да повлияят на модела. След филтрациите нека погледнем корелациите между данните:

- за тези с наднормено тегло

```
                age          sex         bmi    children      region     charges
age      1.000000000 -0.008614303  0.01631004  0.10726897 -0.04578534  0.66720887
sex     -0.008614303  1.000000000 -0.01451813  0.09095388  0.03433412 -0.09487546
bmi      0.016310043 -0.014518133  1.00000000  0.09159318  0.10834756  0.37900759
children 0.107268969  0.090953877  0.09159318  1.00000000 -0.05664248  0.15278129
region  -0.045785340  0.034334120  0.10834756 -0.05664248  1.00000000 -0.06103978
charges  0.667208865 -0.094875458  0.37900759  0.15278129 -0.06103978  1.00000000
```

- за тези с нормално тегло

```
                age          sex         bmi    children      region     charges
age      1.000000000 -0.03792989  0.003642967  0.06479168 -0.08626855  0.90897462
sex     -0.037929888  1.00000000  0.168043398  0.08918650  0.05985445 -0.04950460
bmi      0.003642967  0.16804340  1.000000000 -0.09109829  0.09799376  0.33661886
children 0.064791684  0.08918650 -0.091098292  1.00000000  0.12714566  0.13365551
region  -0.086268546  0.05985445  0.097993760  0.12714566  1.00000000 -0.06662219
charges  0.908974620 -0.04950460  0.336618859  0.13365551 -0.06662219  1.00000000
```

От тук може да заключим, че съществено влияят възрастта и bmi. Така получаваме моделите:

- За пушачите наднормено тегло получаваме:

```
Call:
lm(formula = data11$charges ~ data11$age + data11$bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-1932.5  -804.5  -228.4   412.1  7078.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13929.480   1098.161   12.68   <2e-16 ***
data11$age    275.822      8.357   33.01   <2e-16 ***
data11$bmi    458.335     29.418   15.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1419 on 135 degrees of freedom
Multiple R-squared:  0.9084,    Adjusted R-squared:  0.9071
F-statistic: 669.5 on 2 and 135 DF,  p-value: < 2.2e-16
```

- За пушачите с нормално тегло получаваме:

```
Call:
lm(formula = data12$charges ~ data12$age + data12$bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-2051.3  -642.4    18.4   522.6  5921.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1015.312    858.051  -1.183    0.239
data12$age    264.927      6.784  39.054   <2e-16 ***
data12$bmi    456.996     31.869  14.340   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1005 on 116 degrees of freedom
Multiple R-squared:  0.9373,    Adjusted R-squared:  0.9362
F-statistic: 867.5 on 2 and 116 DF,  p-value: < 2.2e-16
```

Сравнявайки крайните модели с началните може да кажем, че крайните са по-близки до данните, съответно и по-добри ги приближават. Но за сметка на това, използваме 3 модела вместо един.

# Заключения:

- Не може да се създаде линеен модел, който да използва цялата информация на куп, за да приближи качествено данните. Но за сметка на това може да се направят наблюдения как си влияят данните и да се конструират модели базирани на части от извадката.
- Пушенето е съществения фактор, който обуславя какви са медицинските разходи на един човек. В съвкупност с наднормено тегло нещата става и по-скъпи.
- При добро поддържане на телесна маса медицинските разходи се увеличават с течение на възрастта, но това не е единствен фактор, разбира се.

# Източници на информация

## Библиография

Choi, M. (2018). *Medical Cost Personal Datasets*. Извлечено от Keggle: https://www.kaggle.com/mirichoi0218/insurance

Glen_b. (5 6 2014 г.). *How to interpret a QQ plot*. Извлечено от Stack Exchange: https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot