Analiza kosztow medycznych w zaleznosci od parametrow czlowieka

404838, Dzmitry Mikialevich, czwartek 11^{30}
AGH, Wydzial Informatyki Elektroniki i Telekomunikacji
Rachunek prawdopodobienstwa i statystyka 2020/2021

Krakow, January 23, 2021

Ja, nizej podpisany(na) wlasnorecznym podpisem deklaruje, ze przygotowalem(lam) przedstawiony do oceny projekt samodzielnie i zadna jego czesc nie jest kopia pracy innej osoby.

Dzmitry Mikialevich

Contents

1 Introduction **2**

2 Summary of the report **2**

3 Data description **3**

4 Analysis of single variables **3**
4.1 Smoking . . . 3
4.2 Gender . . . 5
4.3 Age . . . 5
4.4 Testing BMI distribution . . . 9
4.5 Children . . . 14
4.6 Charges . . . 15

5 Testing **18**
5.1 Quick Theory Review on Testing . . . 18
5.2 Is BMI Normally Distributed? . . . 19
5.3 Do charges depend on gender? . . . 20

6 Interval Estimators for single variables **22**
6.1 BMI . . . 22

7 Dependencies between data samples **23**
7.1 Smoking and charges . . . 24
7.2 Gender and charges . . . 29
7.3 BMI and charges . . . 30

8 Building Regression a model **31**
8.1 First model . . . 31
8.2 Second model . . . 33
8.3 Third model . . . 34
8.4 Fourth model . . . 40
8.5 Summing up on models . . . 45

9 Conclusion **45**

1 Introduction

- Streszczenie i opis danych w raporcie sa napisane w jezyku polskim, natomiast pozostala czesc jest napisana w jezyku angielskim, jak bardziej wygodnym dla autora, tak i dla kompilatora sweave.
- Dla zapoznania sie ze szczegolami wykonania obliczen prosze o zapoznanie sie z raportem oraz dolaczonym plikiem mainScript.R, zawierajacym pelny opis wszystkich przeprowadzonych badan

2 Summary of the report

Raport powstal w oparciu o analize danych dotyczacych kosztow medycznych, zrobiony przez firme ubezpieczeniowa.
Jako wynik analizy, znalezione byly nastepujace zaleznosci:

Figure 1: Summary Plot

- Koszty leczenia liniowo zaleza od wieku
- Koszty leczenia zaleza od tego, pali osoba czy nie. W przypadku osob palacych z "normalnym:" BMI (okolo 30), koszty leczenia sa 2 razy wieksze niz dla osob niepalacych z normalnym BMI, natomiast obosy palace z wysokim albo bardzo nizkim BMI placa 4 razy wiecej niz osoby "zkwykle"
- Koszty leczenia zaleza od BMI, ale nie w takim wielkim stopniu, jak od palenia
- Koszty leczenia nie zaleza od plci.

Jako wynik badania powstaly kilka modeli, i nastepujaca byla wybrana przez autora jako najlepsza: model<- lm (charges age + smoker+ bmi + SmokerWithHighBMI,data), gdzie SmokerWithHighBMI to wartosc, wskazujaca, czy osoba pali i ma BMI>30, albo nie w przeciwnym przypadku. 8.3
Ten model daje p-value: < 2.2e-16, Adjusted R-squared: 0.8607, ale Residual Standard Error nie jest postaci normalnej. Z pewnym przyblizeniem, mozna sie zgodzic na taki model.

3 Data description

Dane do projektu pochodza ze strony `https://www.kaggle.com/mirichoi0218/insurance`. Skladaja sie z 1338 rekordow, zawierajacych nastepujaca informacje:

- age: wiek beneficjenta pierwotnego
- sex: plec kontrahenta ubezpieczeniowego, kobieta, mezczyzna
- bmi: Wskaznik masy ciala, zapewniajacy zrozumienie ciala, masy, ktore sa stosunkowo wysokie lub niskie w stosunku do wzrostu, obiektywny wskaznik masy ciala kg/m^2 na podstawie stosunku wzrostu do masy ciala, najlepiej 18,5 do 24,9
- children: Liczba dzieci objetych ubezpieczeniem zdrowotnym / Liczba osob na utrzymaniu
- smoker: Palenie
- region: obszar mieszkalny beneficjenta w USA, na polnocnym wschodzie, poludniowym wschodzie, poludniowym zachodzie i polnocnym zachodzie.
- charges: Indywidualne koszty leczenia rozliczane przez ubezpieczenie zdrowotne

4 Analysis of single variables

In this section we are going to perform analysis of single variables to discover their properties and to verify that data makes sens and is well-distributed.

4.1 Smoking

Let's take a look at plot of smokers:

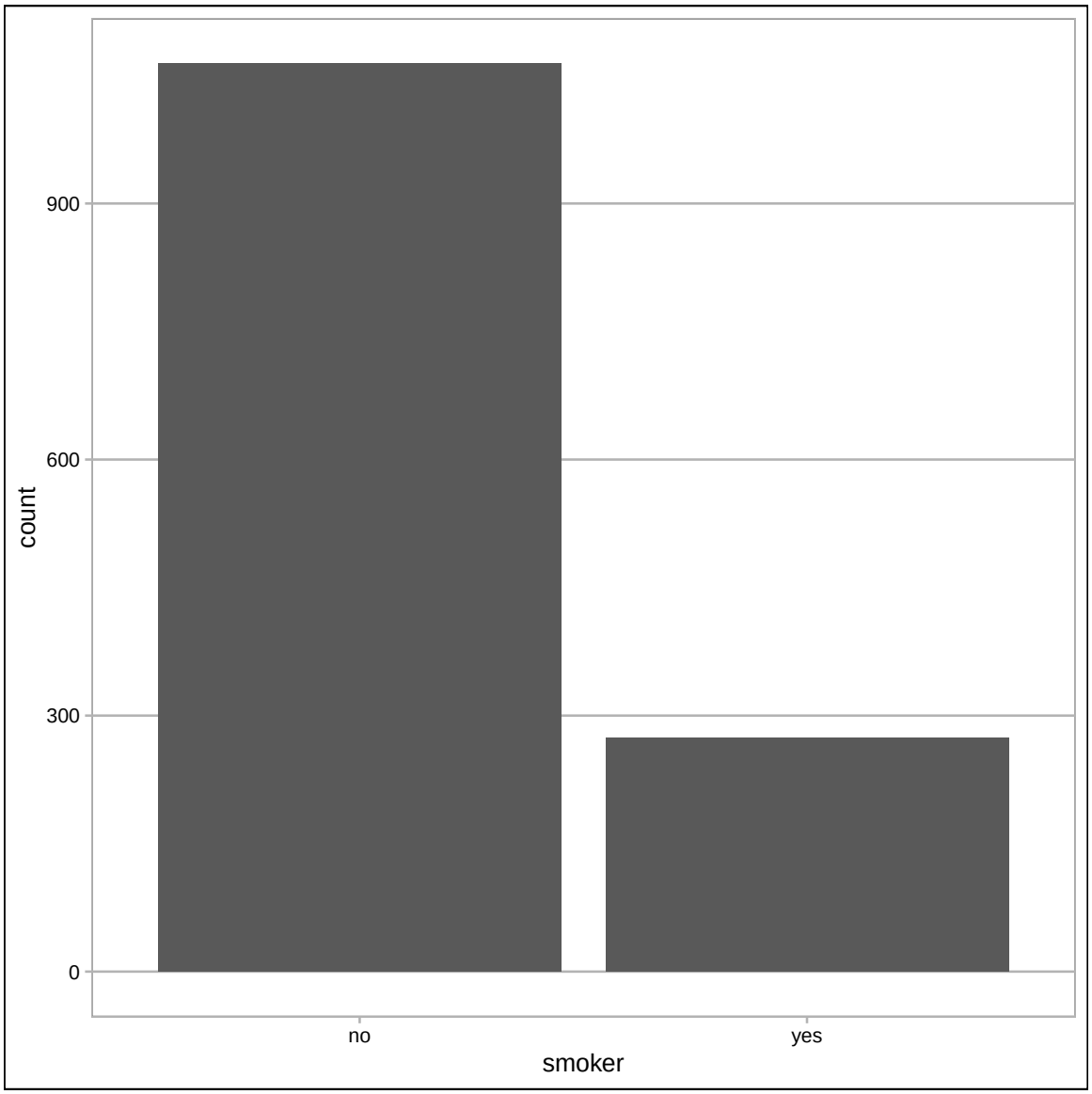

Figure 2: Plot of smokers

As we can see, there are more smokers, than non-smokers, which applies us to be more carefull while making decision

Point Estimation of Population Proportion

At this moment we can try to estimate the Proportion of smokers to non-smokers in America, having that small sample

```
> f = sum(data$smoker=='yes')
> n = length(data$smoker)
> f/n
```

```
[1] 0.2047833
```

The point estimate of smoking people proportion in survey is 20%, which is pretty close to our expectations (According to the CDC, as of 2015, a total of 15.1% of U.S. adults (16.7% of men and 13.6% of women) smoke)

4.2 Gender

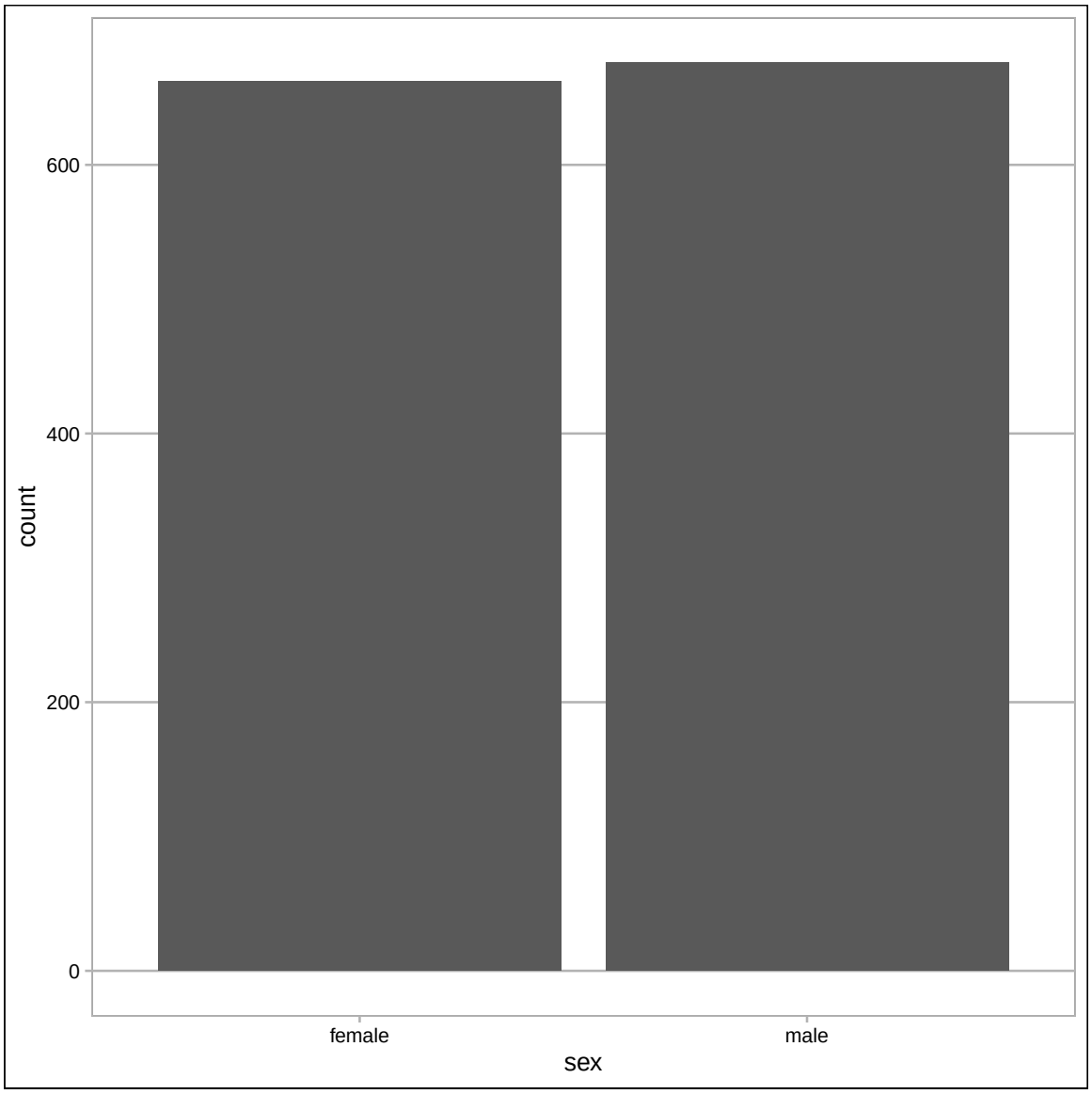

Figure 3: Plot of genders

As we can see we have almost equal distribution of man and women in survey

Point Estimation of Gender Proportion

```
>   f = sum(data$sex=='male')
>   n = length(data$sex)
>   f/n
```

```
[1] 0.5052317
```

As we can see, the point estimate of male gender proportion in survey is 50%, which also meets our expectations, meaning the data is accurate and well-distributed, comparing to population.

4.3 Age

Basic number properties

```
> describeBy(data$age)
```

```
   vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 1338 39.21 14.05     39   39.01 17.79  18  64    46 0.06    -1.25 0.38
```

Analysing which, we can outline these properties:

- **n** - number of items
- **mean** - mean of sample
- **sd** - standard deviation (a measure of the amount of variation)
- **median** - median of sample (50th percentile)
- **trimmed** - trimmed mean, in other words this value is more stable than the mean, because it is calculated like:
  1. Cut off 10% from left side of distribution
  2. Cut off 10% from right side of distribution
  3. Calculate average from the remaining 80%
- **mad** - mean absolute deviation (variability similar to the sum of squares)
- **min** - minimum value
- **max** - maximum value
- **range** - the difference between the max and min values
- **skew** - $A = \frac{\mu_3}{\sigma^3}$ a measure of asymmetry in the distribution
  1. $A == 0 \Rightarrow$ distribution is symmetrical
  2. $A > 0 \Rightarrow$ distribution has positive skew
  3. $A < 0 \Rightarrow$ distribution negative skew
- **kurtosis** - $Kurt[X] = \frac{\mu_4}{\sigma^4}$ a measure of the peakedness of the probability distribution
  1. $Kurt[X] == 0 \Rightarrow$ rounded peak of a normal distribution (Mesokurtic)
  2. $Kurt[X] > 0 \Rightarrow$ a sharper peak (Leptokurtic)
  3. $Kurt[X] < 0 \Rightarrow$ a flatter peak (Platykurtic)
- **se** - sample standard error

Average deviation

```
> avg.dev <- function(x)
+   mean(abs(x - mean(x)))
> c(avg.dev(data$age))

[1] 12.24893
```

Quantiles (minimum, lower-hinge, median, upper-hinge, maximum)

```
> fivenum(data$age)

[1] 18 27 39 51 64
```

(upper-hinge - lower-hinge)

```
> IQR(data$age)

[1] 24
```

Central, not absolute moments

```
> moment(data$age,0.25)
[1] 2.469482
> moment(data$age,0.5)
[1] 6.154262
> moment(data$age,0.75)
[1] 15.47053
> moment(data$age,1)
[1] 39.20703
```

Plots

Below we can see different types of plots to help us in analyzing the sample.

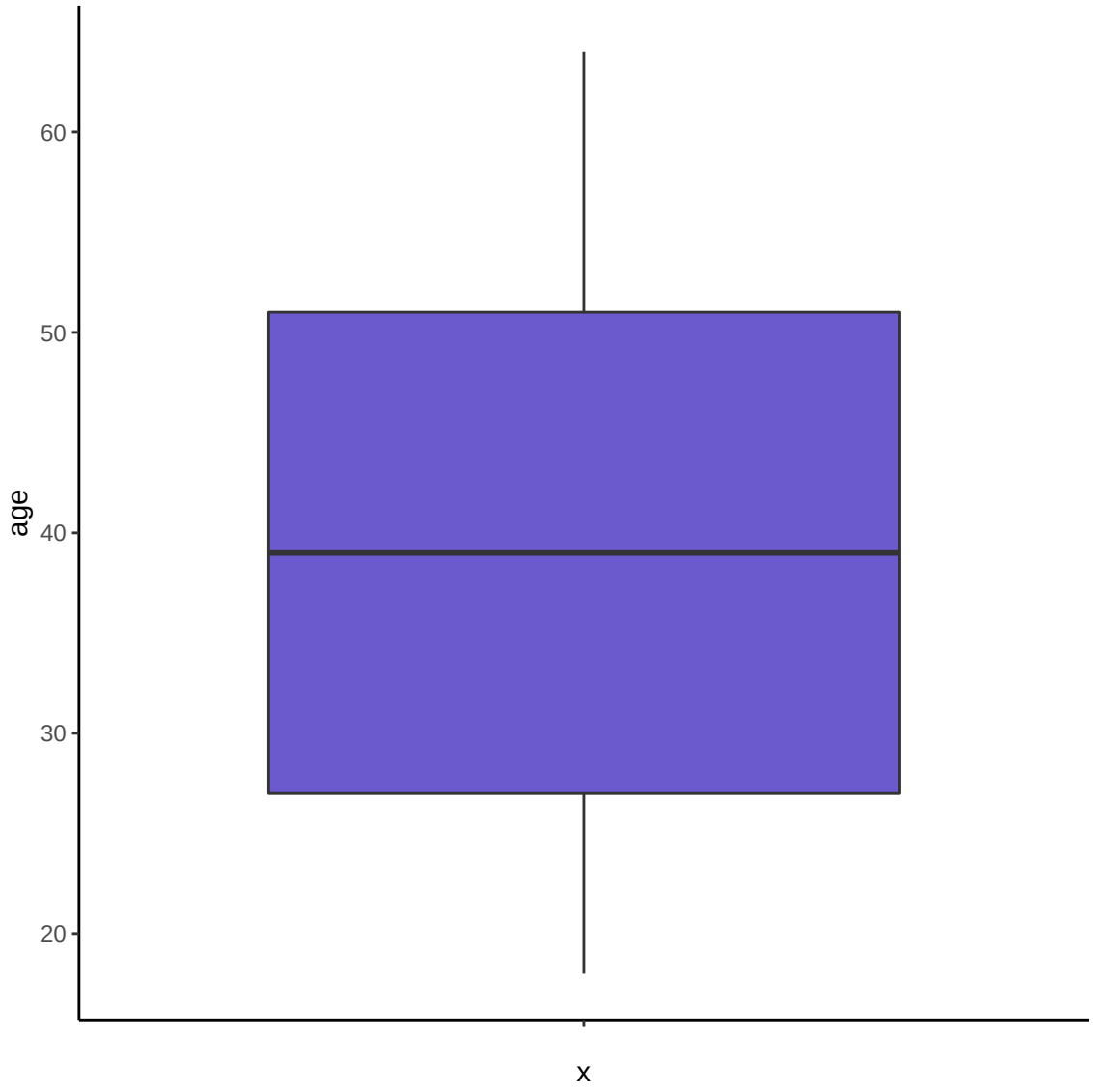

Figure 4: BoxPlot of age

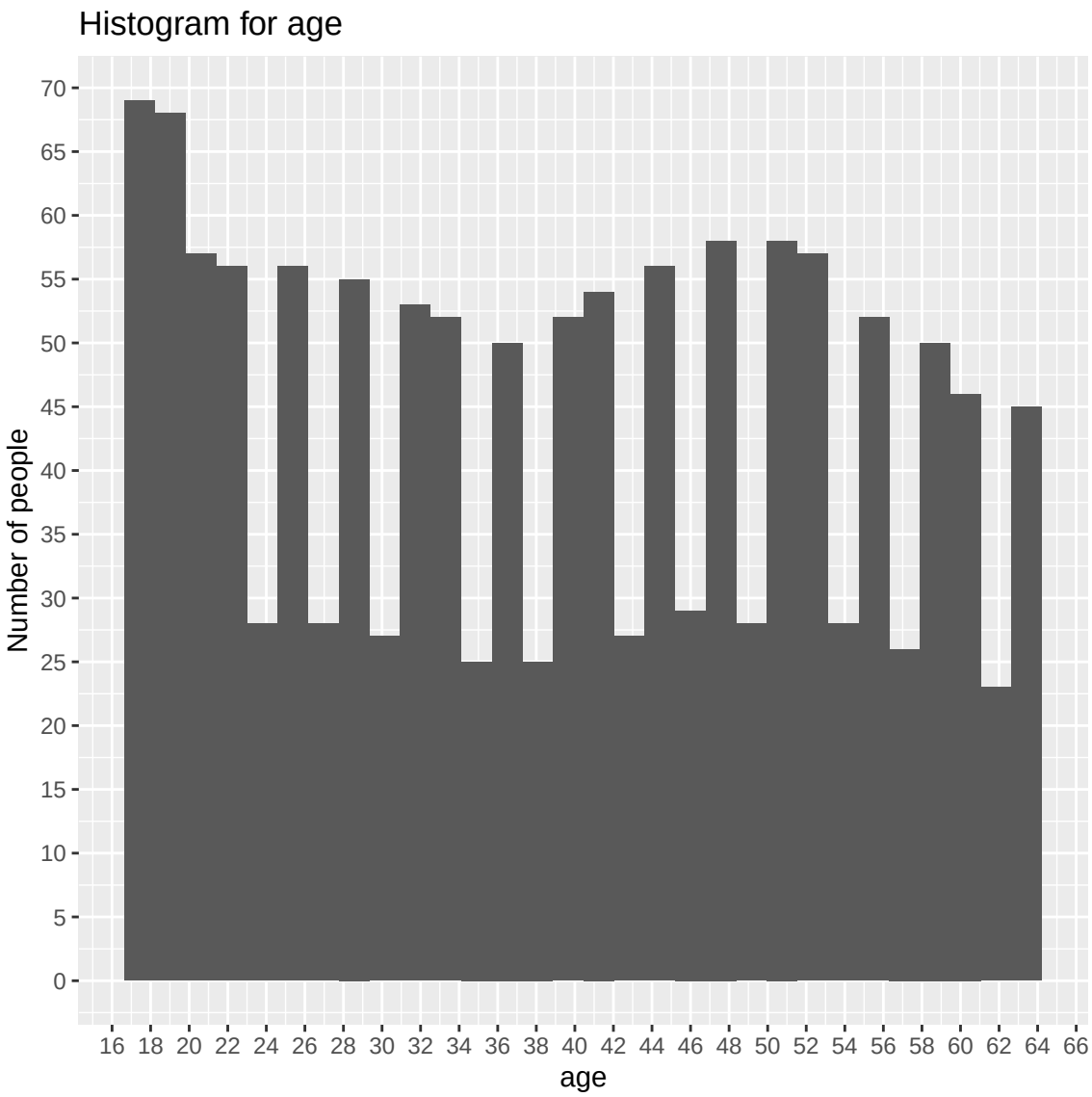

Figure 5: Histogram of age

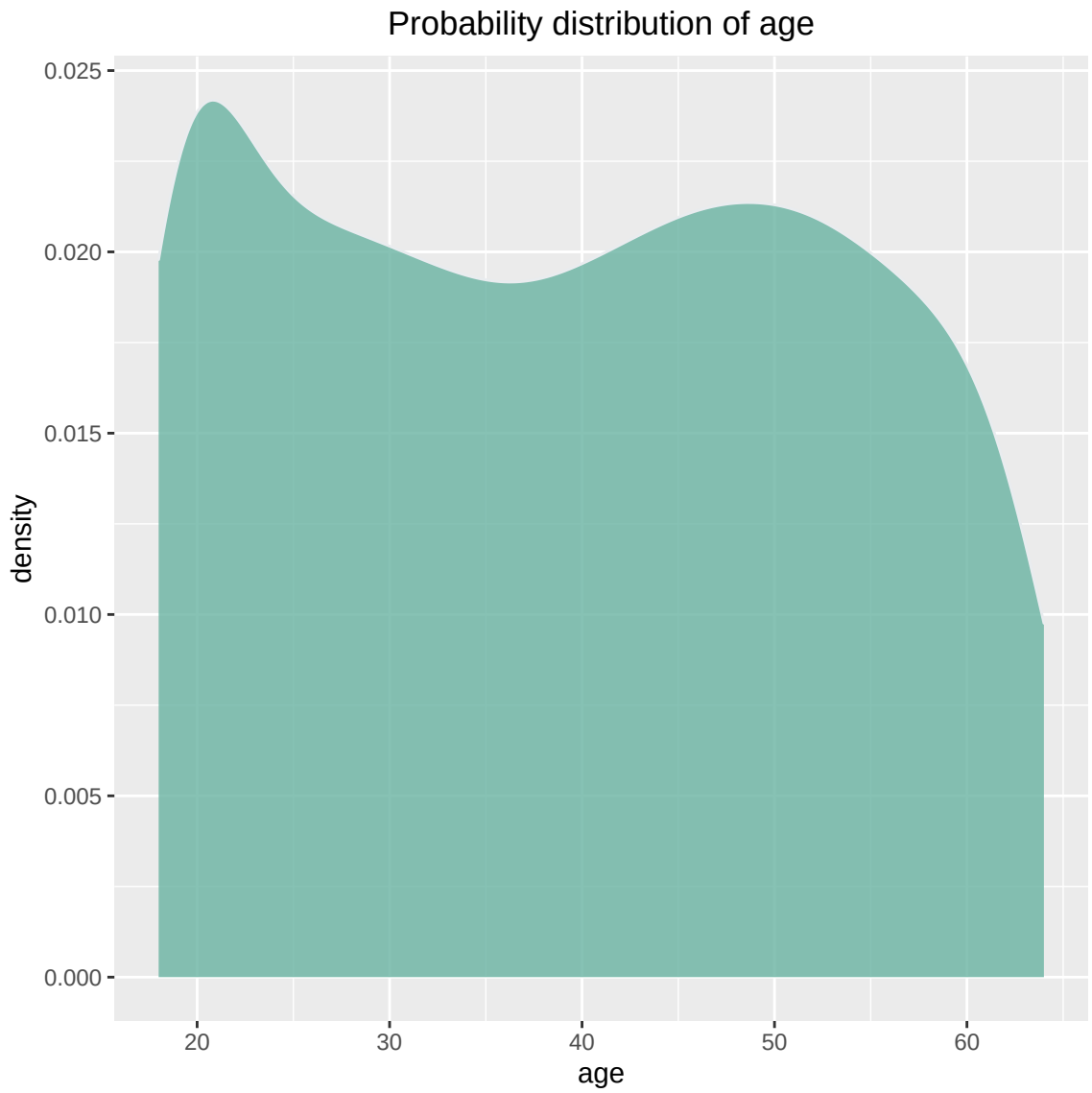

Figure 6: Probability distribution of age

As we can see, we've got evenly uniformly distributed variable, So next analysis can pretend to be accurate

4.4 Testing BMI distribution

Basic number properties

```
> describeBy(data$bmi)
```

```
   vars    n  mean  sd median trimmed mad   min   max range skew kurtosis   se
X1    1 1338 30.66 6.1   30.4    30.5 6.2 15.96 53.13 37.17 0.28    -0.06 0.17
```

Analysing which, we can outline the same properties as in previous one.

Average deviation

```
[1] 4.897871
```

Quantiles (minimum, lower-hinge, median, upper-hinge, maximum)

```
[1] 15.96 26.29 30.40 34.70 53.13
```

(upper-hinge - lower-hinge)

```
[1] 8.3975
```

Central, not absolute moments

```
[1] 2.344357
```

```
[1] 5.509917
```

```
[1] 12.98222
```

```
[1] 30.6634
```

Plots

Below we can see different types of plots to help us in analyzing the sample.

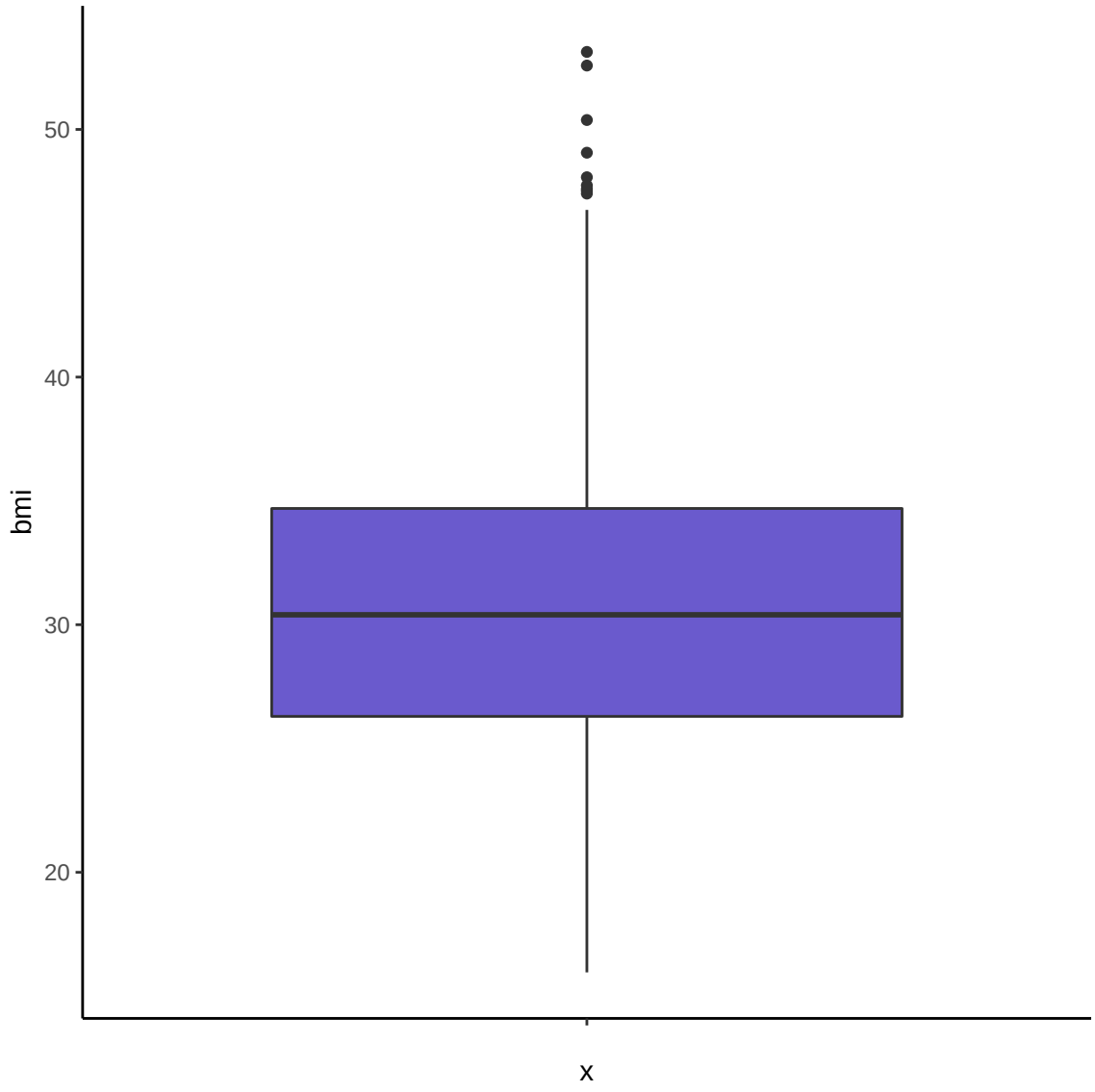

Figure 7: BoxPlot of bmi

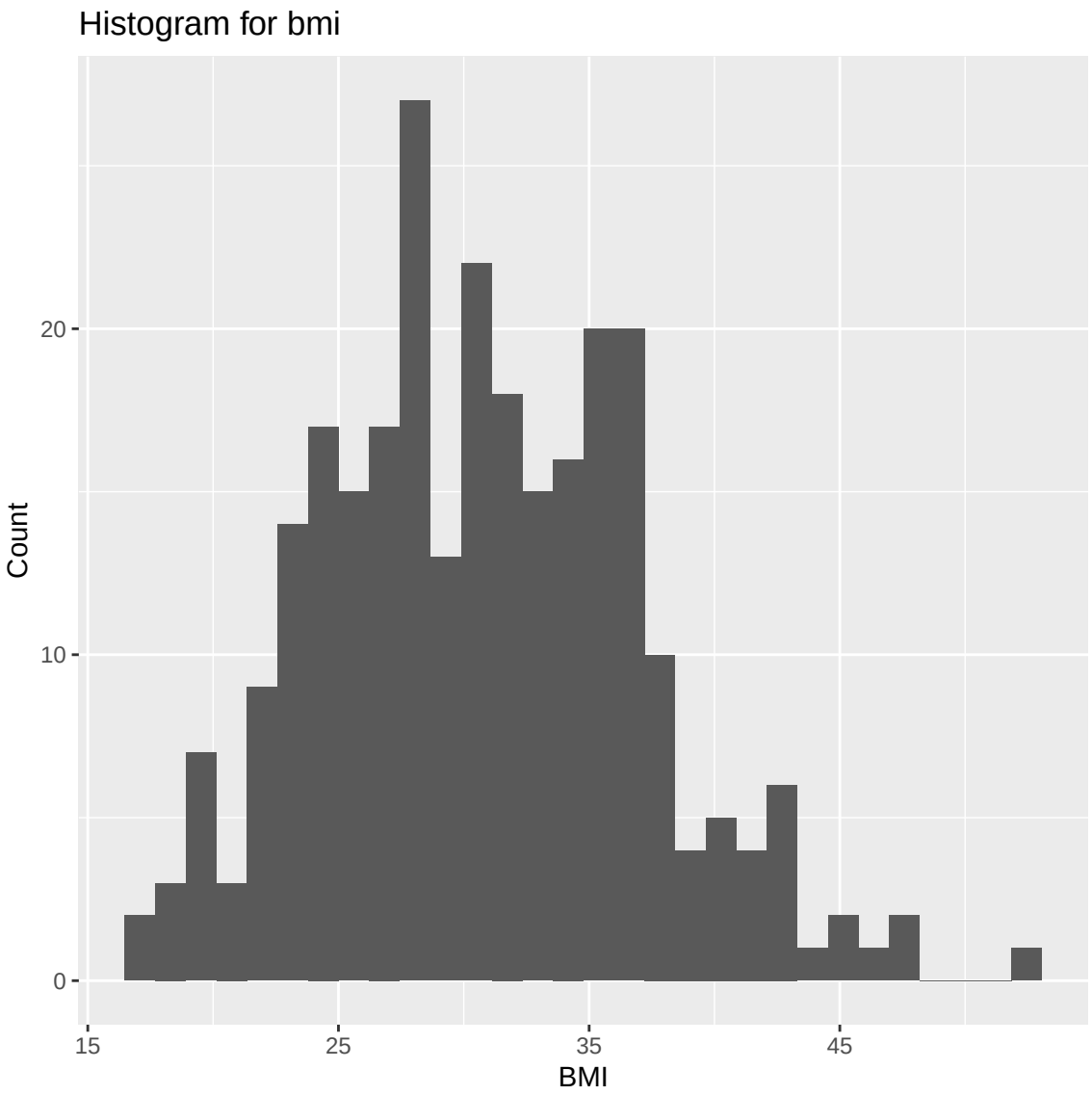

Figure 8: Histogram of bmi

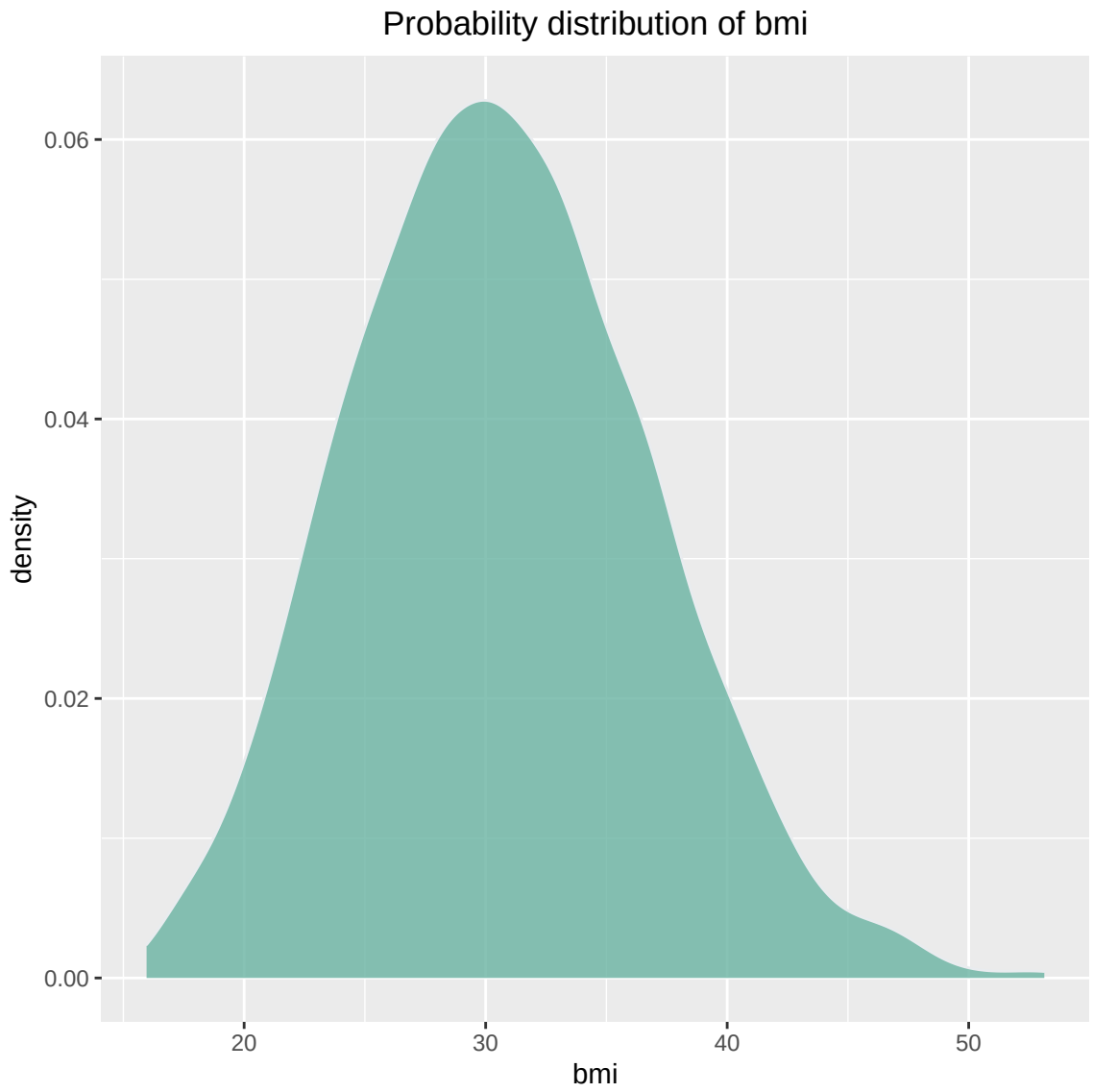

Figure 9: Probability distribution of bmi

As we can see, we've got something that looks like normal distribution, so let's take a closer look at it.

```
> ggplot( data, aes(x=bmi)) +
+   geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) +
+   ggtitle("Probability distribution of bmi") +
+   theme(plot.title = element_text(hjust = 0.5))
```

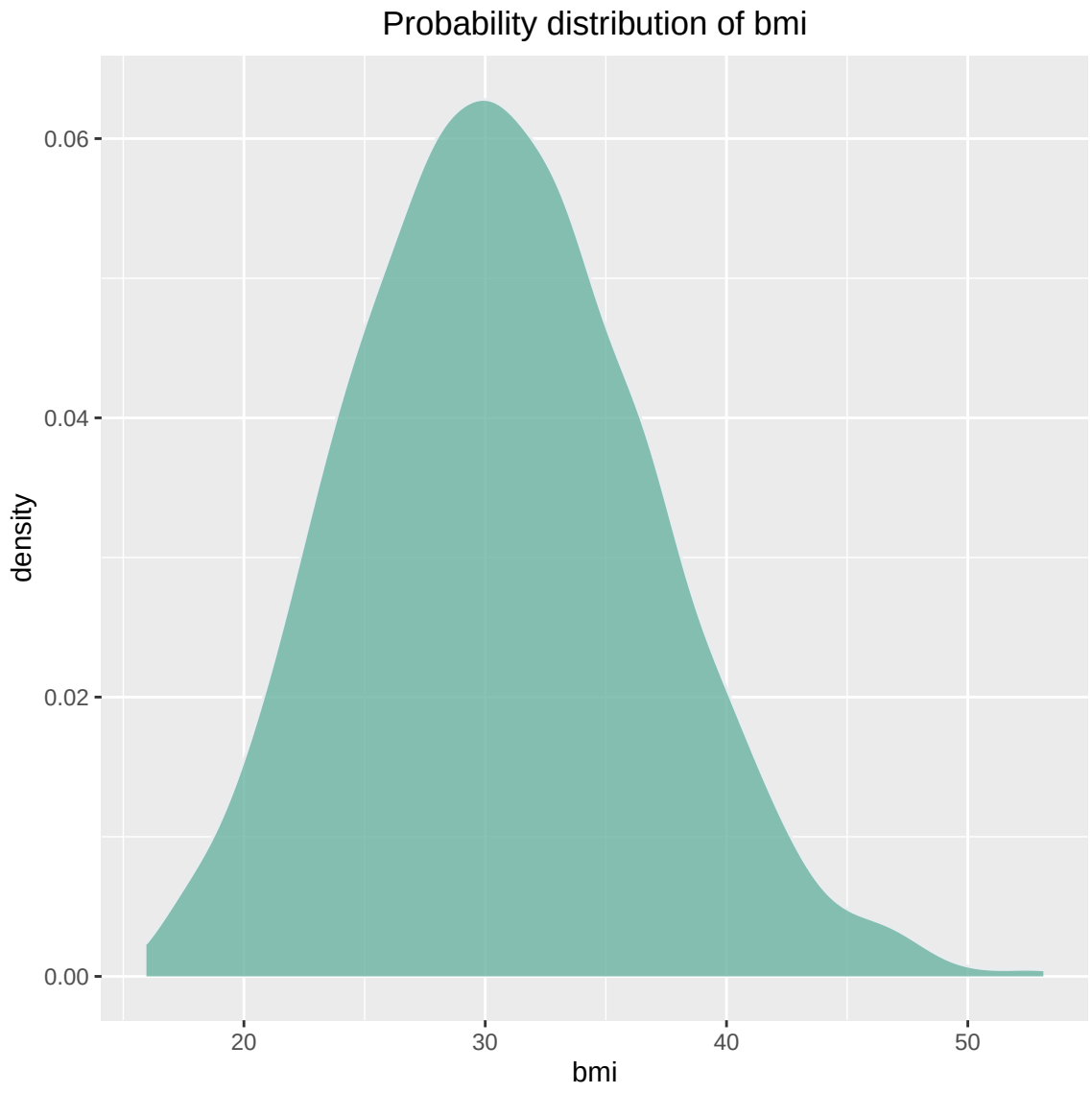

Figure 10: Density plot of BMI

Visual methods -> Density plot of BMI

The plot looks like normal, so continue our analyzing, now using qqplot:

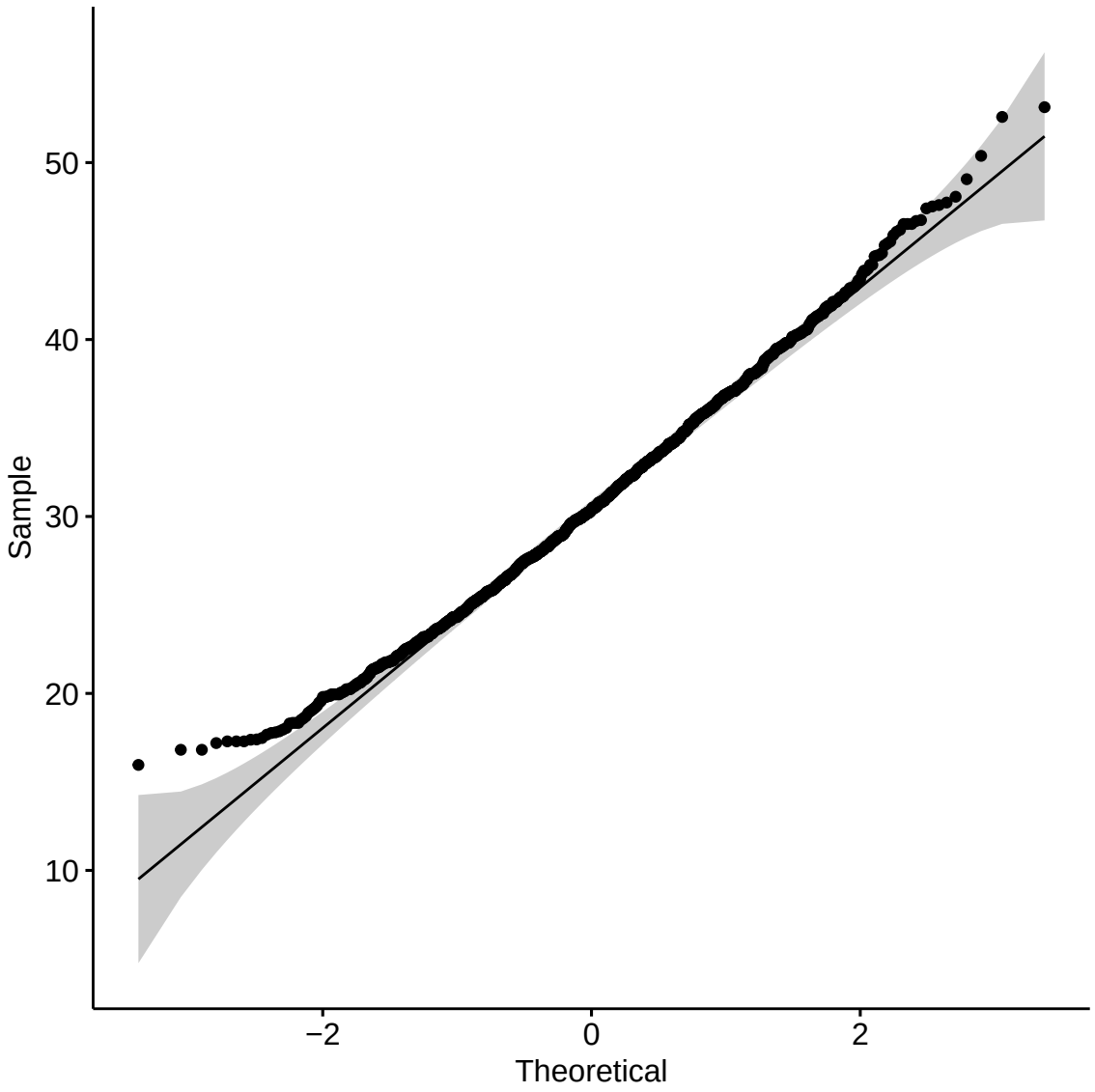

Figure 11: Plot of BMI

As we can see there, at the begining there is small deviation from line of normal distribution, but still it's worth testing.

4.5 Children

Let's see whether charges depend on children

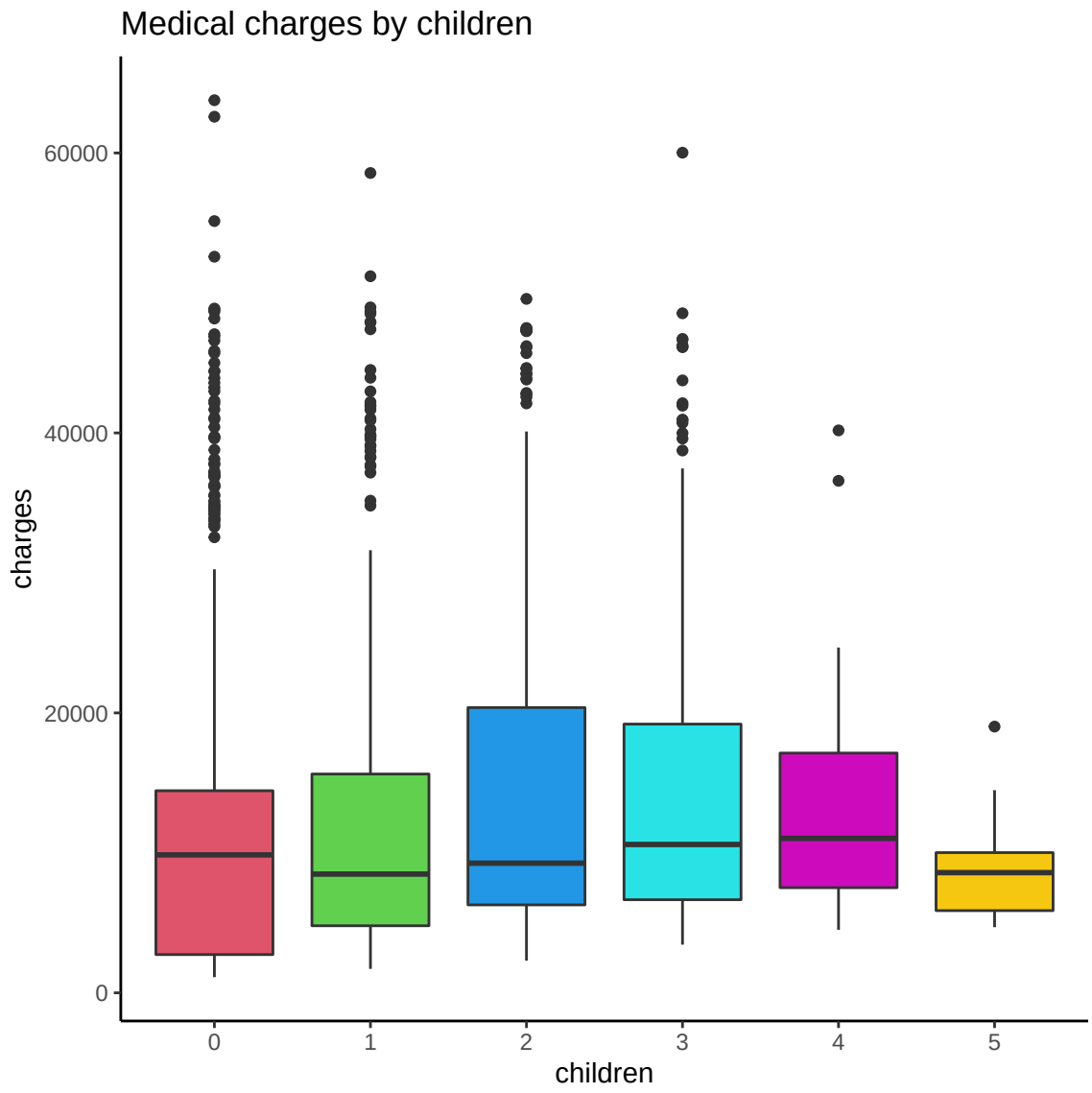

Figure 12: Charges on children

4.6 Charges

Basic number properties

```
> describeBy(data$charges)

   vars    n     mean       sd  median  trimmed     mad     min      max
X1    1 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.87 63770.43
      range skew kurtosis     se
X1 62648.55 1.51     1.59 331.07
```

Average deviation

```
[1] 9091.127
```

Quantiles (minimum, lower-hinge, median, upper-hinge, maximum)

```
[1]  1121.874  4738.268  9382.033 16657.717 63770.428
```

(upper-hinge - lower-hinge)

```
[1] 11899.63
```

Central, not absolute moments

```
[1] 9.982301
```

```
[1] 104.8336
```

```
[1] 1154.389
```

```
[1] 13270.42
```

Plots

Below we can see different types of plots to help us in analyzing the sample.

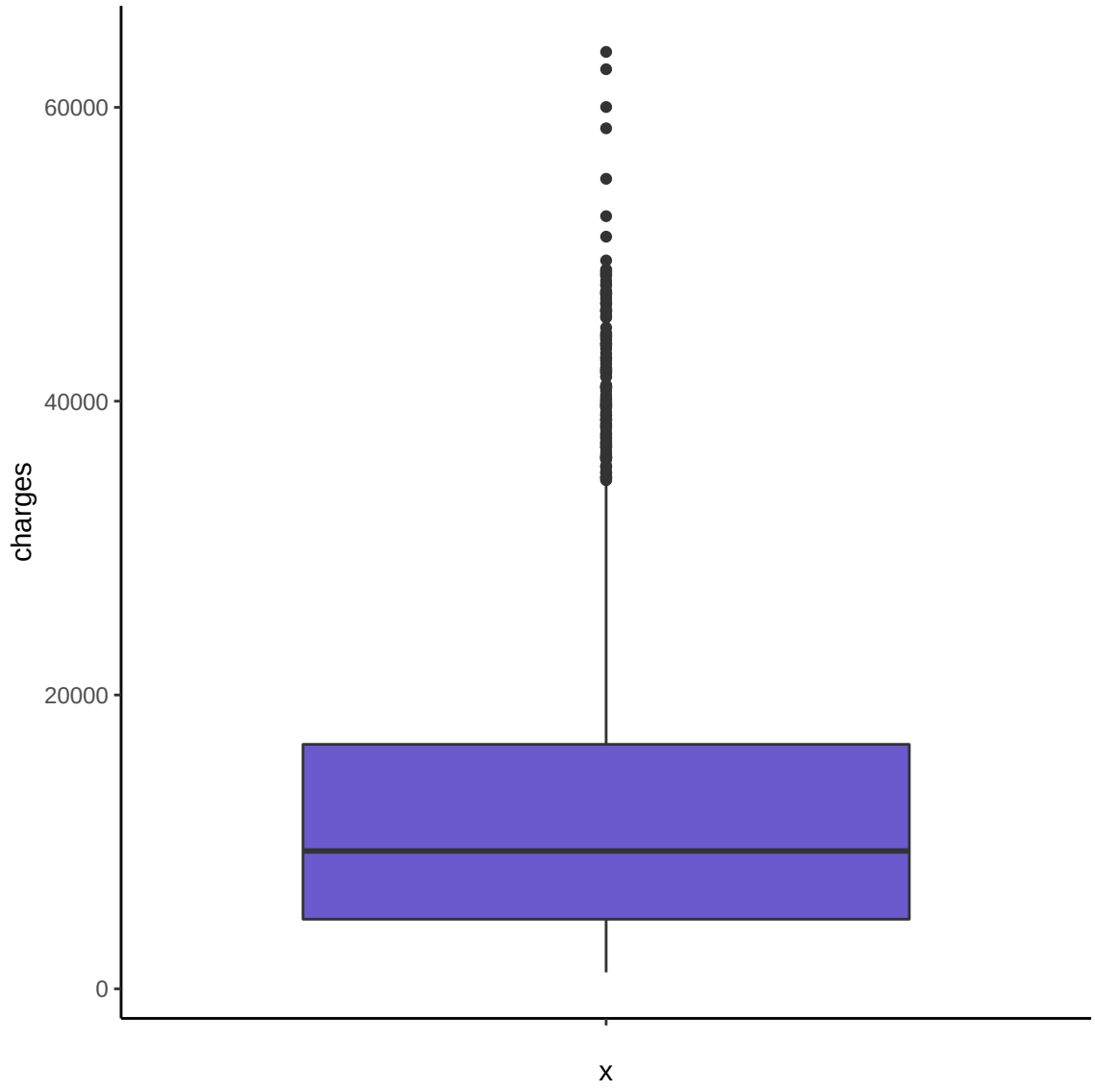

Figure 13: BoxPlot of charges

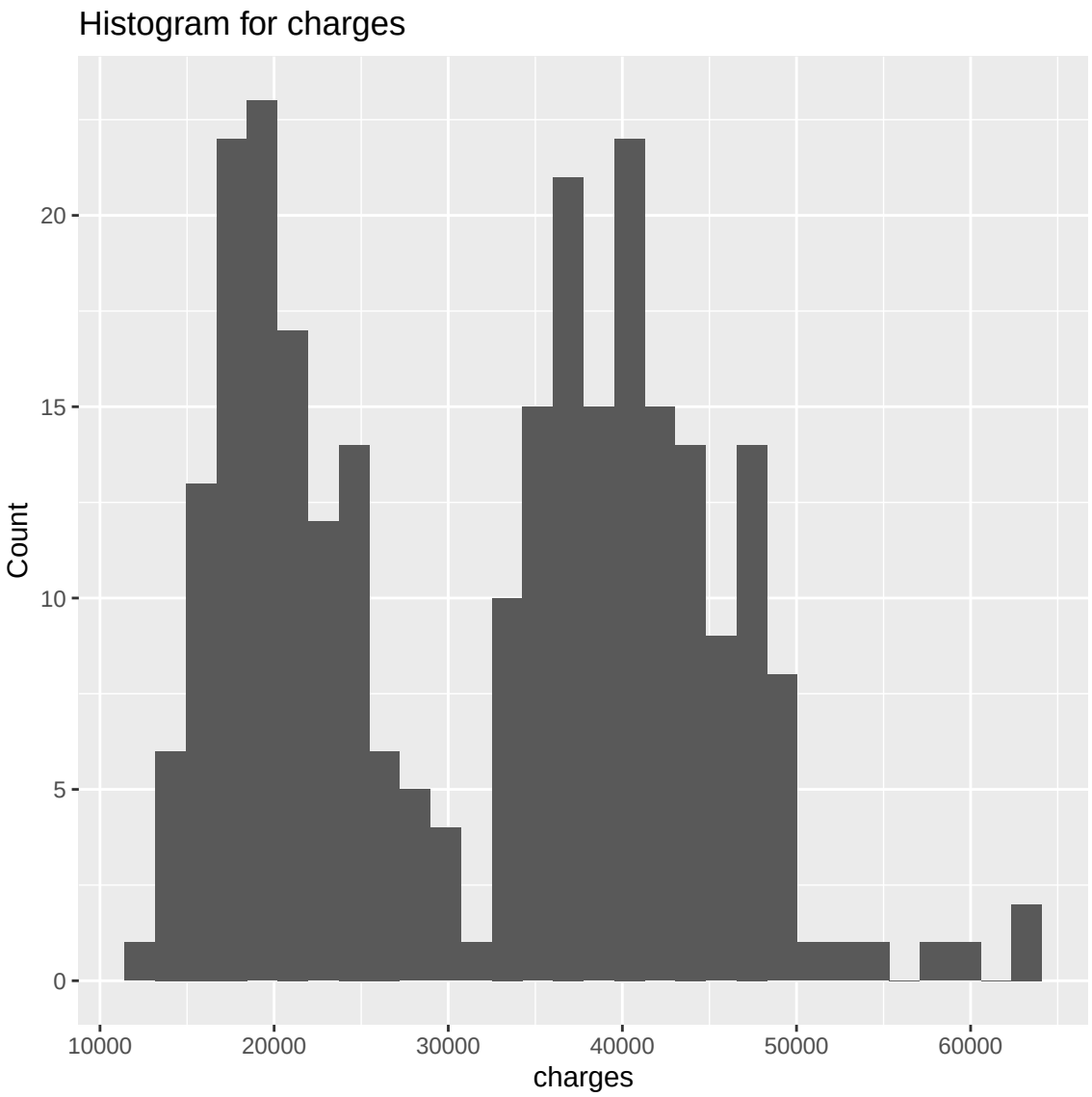

Figure 14: Histogram of charges

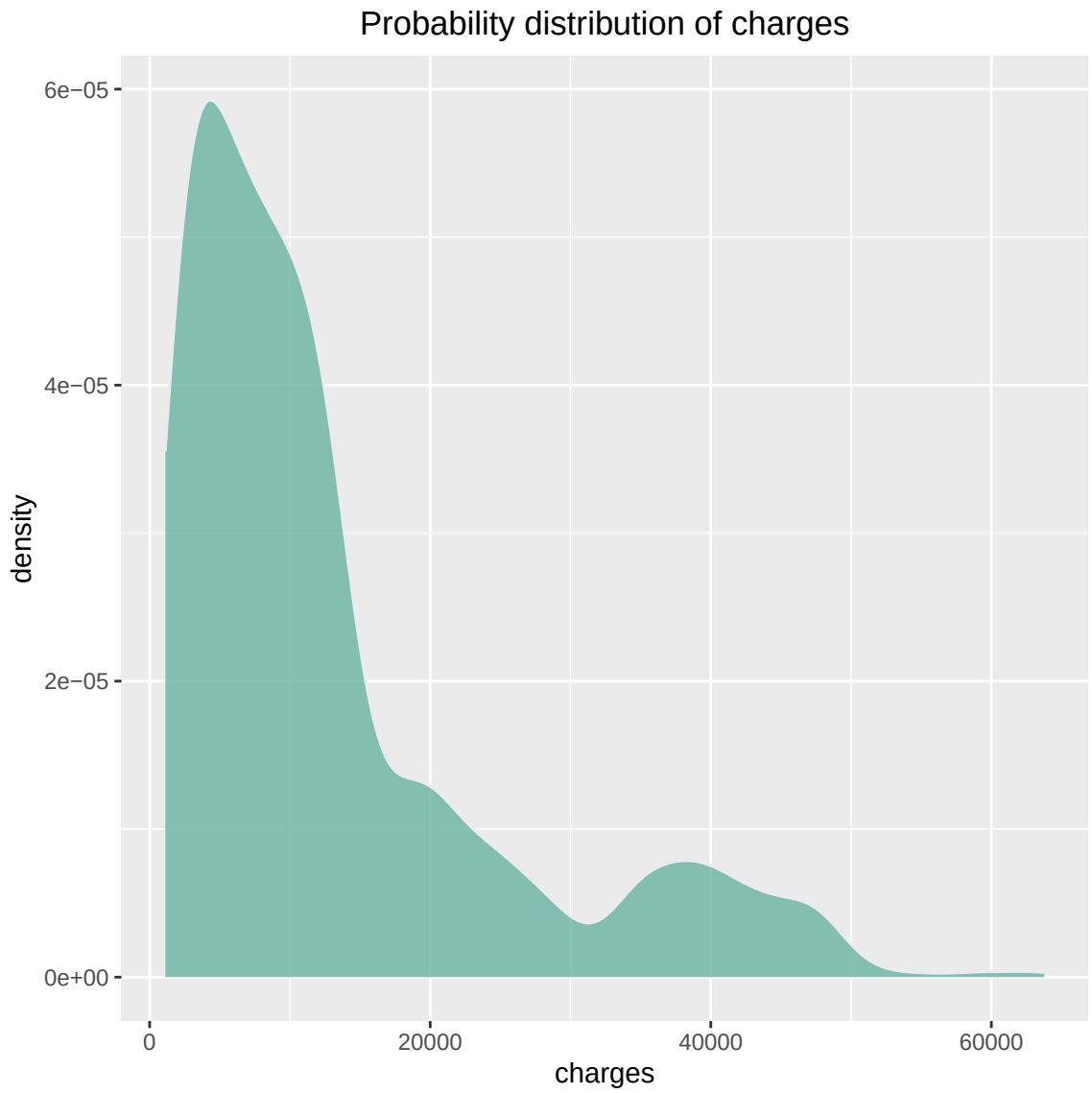

Figure 15: Probability distribution of charges

5 Testing

5.1 Quick Theory Review on Testing

Two-samples t-test

In t-test, the null hypothesis is that the mean of the two samples is equal. So the alternative hypothesis is that the means are different or, $|m_1 - m_2| > 0$ So basically, we want to take or reject the null hypothesis with some confidence interval (the range of values within which the difference may lie) Also, t-test gives us a $p-value$, probability of us, making wrong decision Having small $p-value$ suggests having small probability for null-hypothesis being true.

Shapiro-Wilk's method

Method, based on correlation between the data and the corresponding "normal points" Null-hypothesis is that distribution is normal and we reject the null hypothesis if p < 0.05, meaning that distribution is more likely not normal. Wilk's test should not be significant to meet the assumption of normality.

One-sample t-test

Assumptions:

- Population is normally distributed
- Independent samples

- Random sample via all population distribution
- Continuous

Defining null-hypothesis, we assume that mean of our population is equal to a hypothezed value

5.2 Is BMI Normally Distributed?

In the case of BMI, we can see some outliers, and it can be a point where we stop testing and say that the model doesn't meet the Assumptions, but we will go further.

First of all let's apply Shapiro-Wilk's test (assuming that Assumptions are met)

```
> shapiro.test(data$bmi)

        Shapiro-Wilk normality test

data:  data$bmi
W = 0.99389, p-value = 2.605e-05
```

Having p-value = 2.605e-05, we can reject the null hypothesis, that distribution is normal, applying from that non-normality of our sample distribution.

Also we can try t-test, with default arguments:

```
> t.test(data$bmi, y = NULL,
+        alternative = c("two.sided", "less", "greater"),
+        mu = 0, paired = FALSE, var.equal = FALSE,
+        conf.level = 0.95)

        One Sample t-test

data:  data$bmi
t = 183.93, df = 1337, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 30.33635 30.99045
sample estimates:
mean of x
  30.6634
```

Having tested with t.test (p-value $<$ 2.2e-16) we can conclude that it's highly significant that BMI is not distributed normally

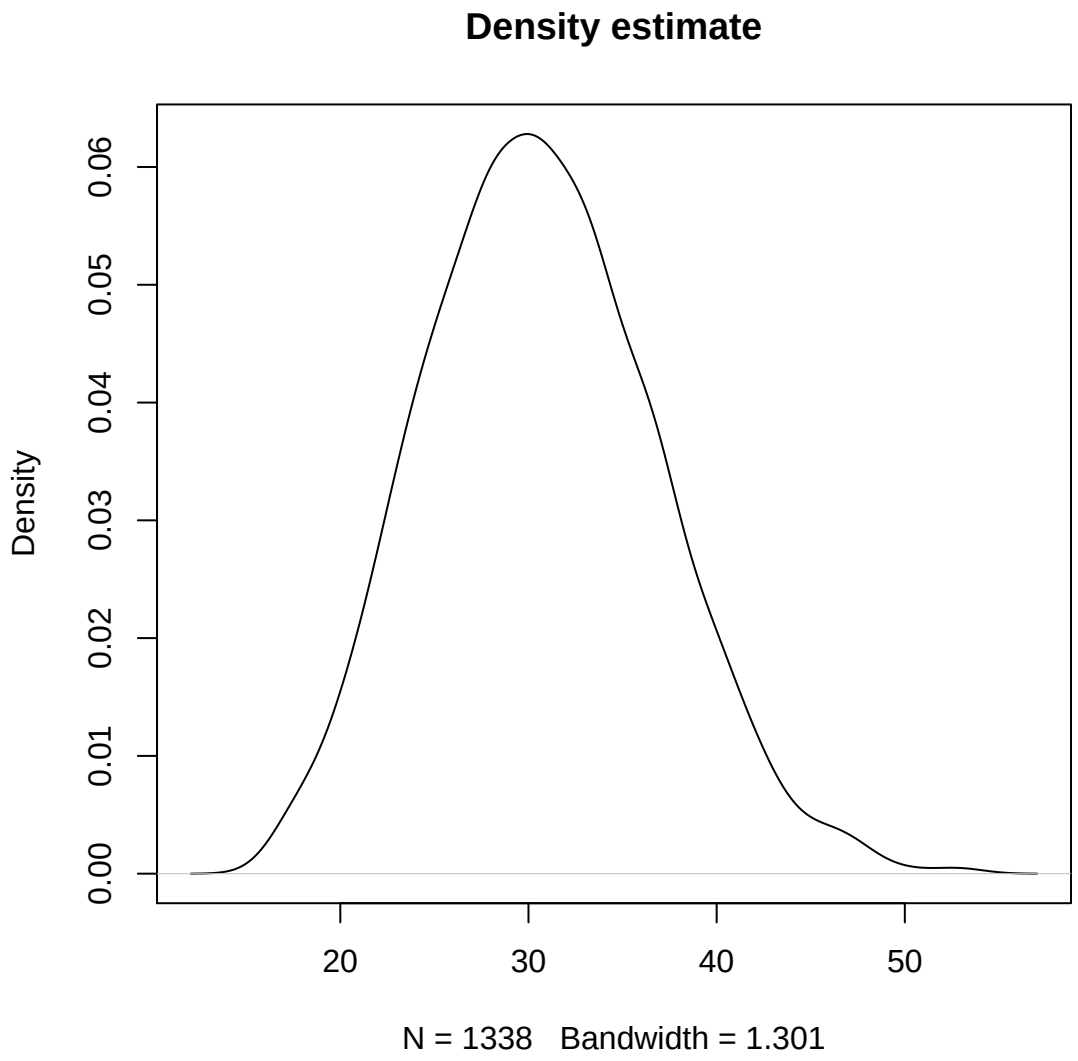

Figure 16: Dencity estimaate

Density Estimate of data

5.3 Do charges depend on gender?

```
> mans_charge <- subset(data,sex=="male" & smoker=='no')
> females_charge <- subset(data,sex=="female"& smoker=='no')
```

As we see on boxplots below, visually they have the same distribution, and median

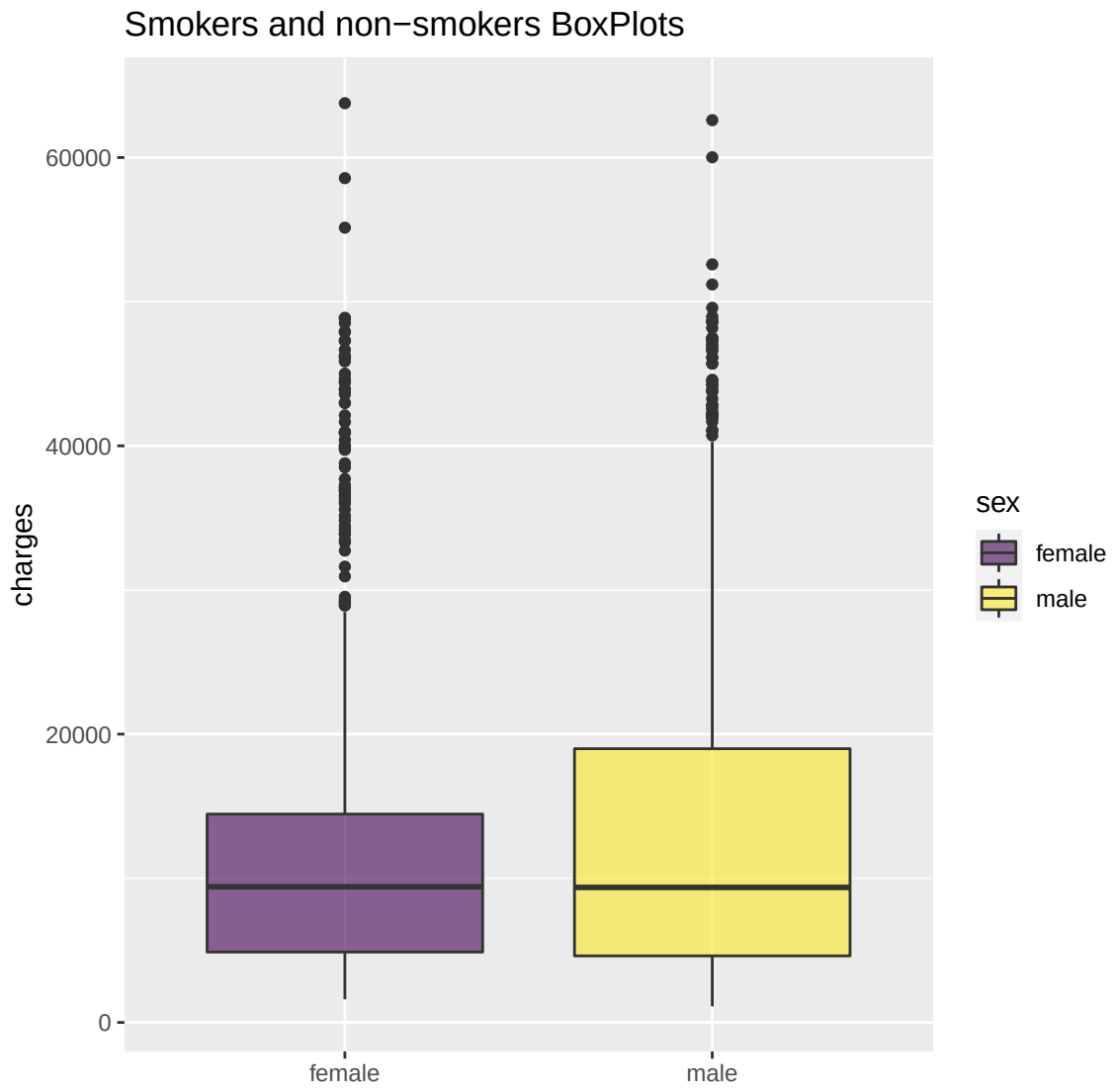

Figure 17: BoxPlot of Smokers and non-smokers

Let's show it using t-test

First check for assumptions: Distribution is not normal, so we need to stop in there, but we will continue due to insterest, understanding that we can't conclude anything from this testing

```
> test <- t.test( mans_charge$charges, females_charge$charges
+ )
> test

        Welch Two Sample t-test

data:  mans_charge$charges and females_charge$charges
t = -1.8396, df = 1061, p-value = 0.0661
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1395.16882    44.98368
sample estimates:
mean of x mean of y
 8087.205  8762.297
```

So because p-value is < 0.07 we could assume (if distribution was normal)right that they are from same population. Which would mean that it doesn't matter which gender is the person, whose charge we are analysing.

6 Interval Estimators for single variables

6.1 BMI

We will use The One Sample t Test -> determines whether the sample mean is statistically different from a known or hypothesized population mean. Assumptions where discussed previously, so let's just check them:

- **Independent** - OK
- **Random** - OK
- **Continuous** - OK
- **Normally distributed** - According to Shapiro-Wilk's test, formally, we don't have normally distributed BMI, so we can't apply t-test, but because difference between our distribution and normal is acceptably small, let's assume normality

```
> t.test(data$bmi)

        One Sample t-test

data:  data$bmi
t = 183.93, df = 1337, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 30.33635 30.99045
sample estimates:
mean of x
  30.6634
```

Our confidence interval is [30.33635, 30.99045], mean of x is 30.6634 and p-value $<$ 2.2e-16, meaning, that we can reject null-hypothesis, that sample and population having same mean, and accept alternative, that means are different and mean of population lies in [30.33635, 30.99045] with probability of 95 percent

Now let's make some calculus by ourselves and see, can we get the same result:
Math part:

- **n** - number of elements in sample
- $\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i)$ - mean of sample
- $\sigma = \sqrt{\sigma^2}$ - standard deviation
- $\sigma_x = \frac{\sigma}{sqrt(n)}$ - standard error
- $z = \Phi(0.025)$ - critical value Z, normal distribution in 0.025 (As we want to get 95% interval of confidence, we need to take 2.5% from left and right)
- $[l = \mu - z * \sigma_x, r = \mu + z * \sigma_x]$ - Confidence interval

R part:

```
> n <- length(data$bmi)
> mu <- mean(data$bmi)
> s <- sd(data$bmi)
> err <- s/sqrt(n)
> z <- qnorm(0.025, lower.tail = F)
> lower_ci <- mu - z*err
> upper_ci <- mu + z*err
> interval_estimation <- c("estimate" = mu, "lower95%" = lower_ci, "upper95%" = upper_ci)
> round(interval_estimation, digits = 5)

estimate lower95% upper95%
30.66340 30.33664 30.99015
```

7 Dependencies between data samples

Now let's check for dependencies. First of all build Correlogram to find out correlation dependencies

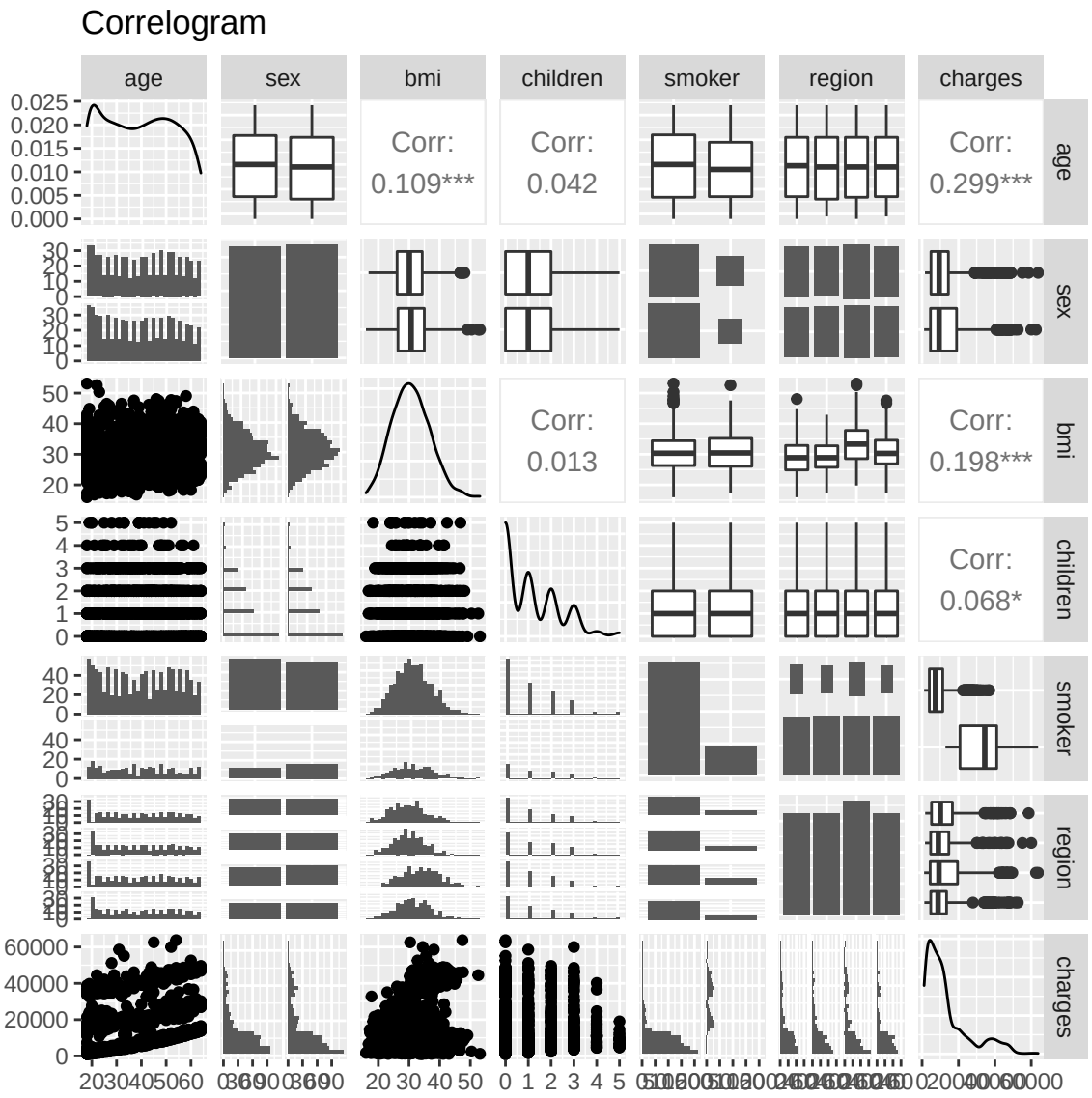

Figure 18: Correlogram

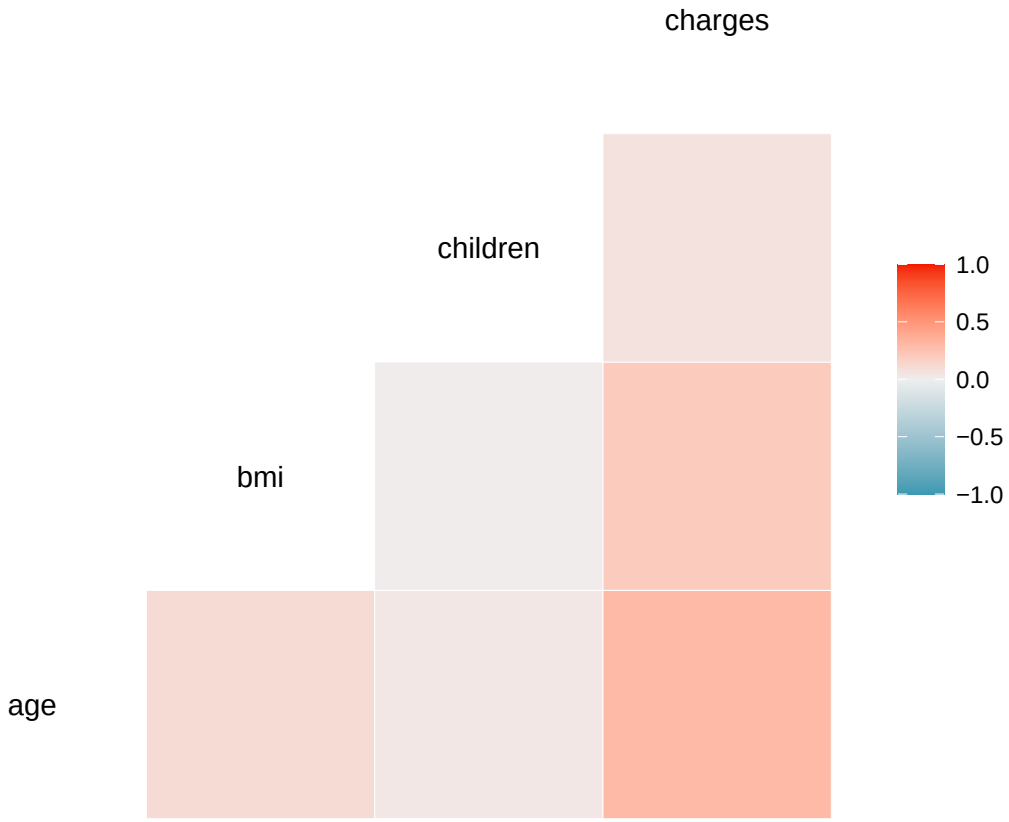

Figure 19: Correlation

As we can see, the age has the highest correlation with charges (0.299)***, Also BMI has (0.198)*** correlation, which can give us some expectancies to Future, where *** means Pr(>|t|) close to 0

7.1 Smoking and charges

Looks like we have more smokers than non-smokers, lets have a look How it affects out charge statistics

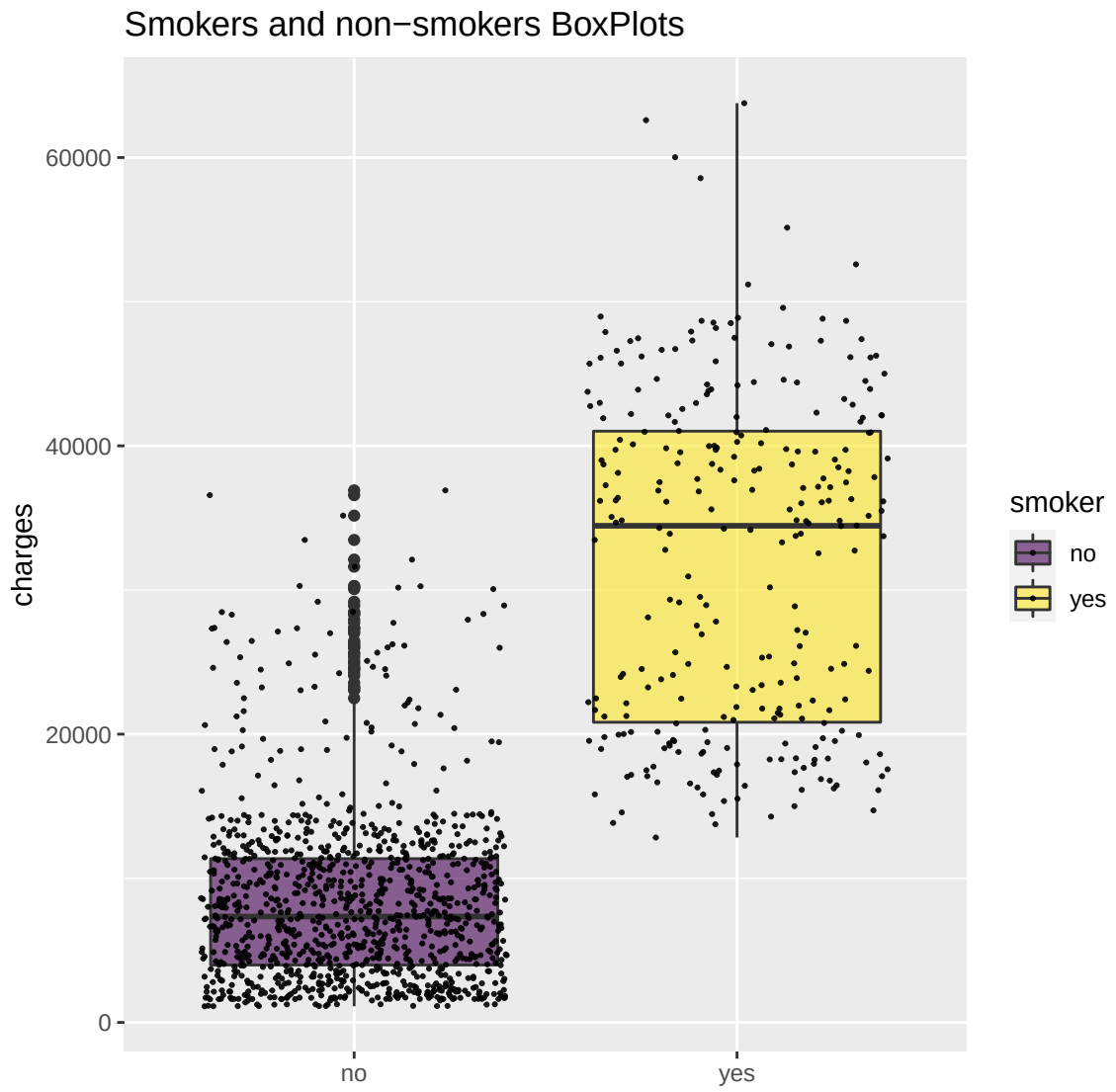

Figure 20: BoxPLot on smoking

Looks like more data about non-smokers doesn't change the picture, but we can trust it more And we can visually divide data from smokers into 2 groups, lets find out why

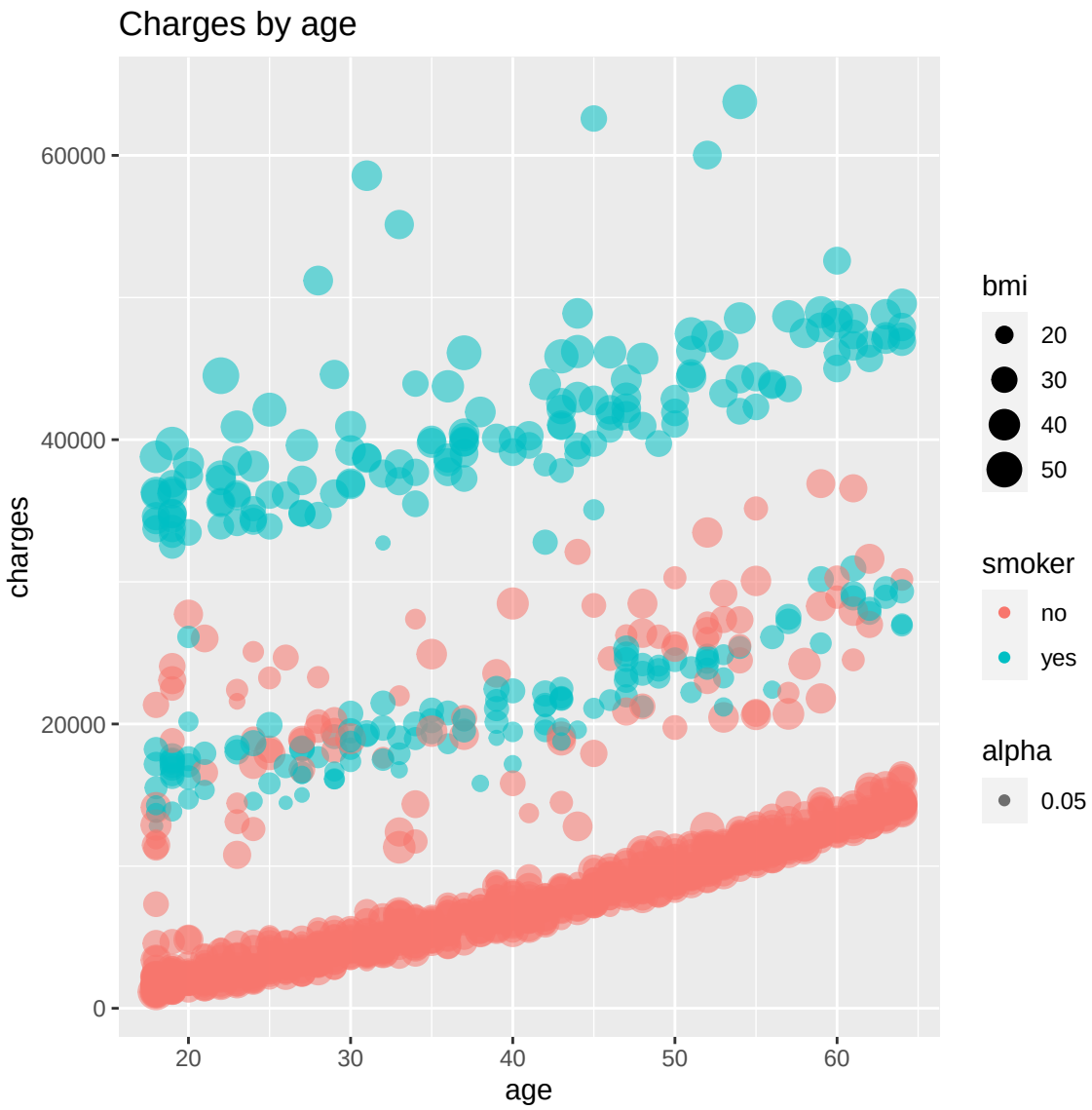

Figure 21: Age and Charges

From that we can find out that smokers with high BMI pay more than smokers with low Also there looks like a linear dependency between age and charges, but we will find it out later

So lets plot the age and the charges for non-smokers

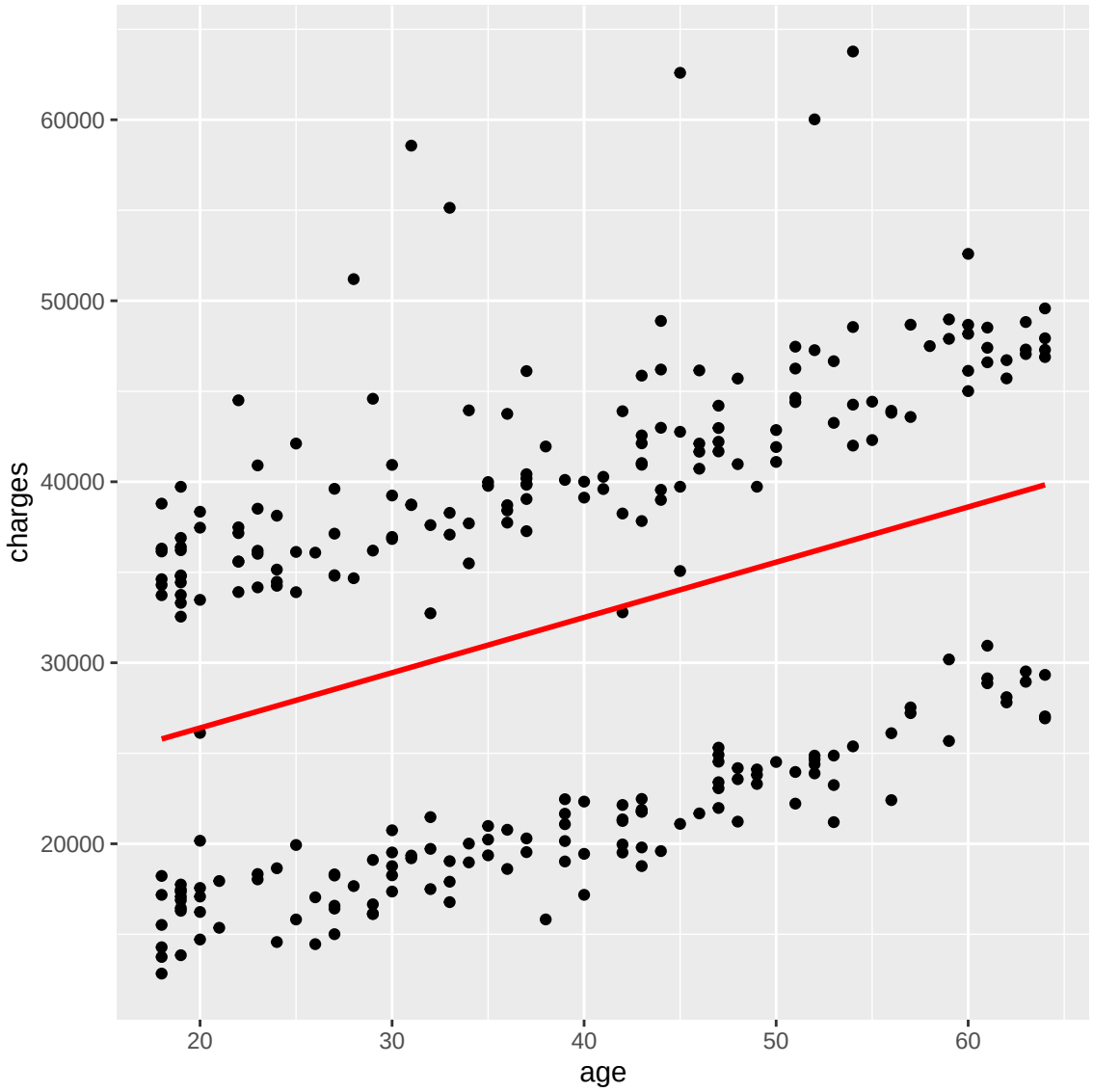

Figure 22: Plot of age and charges

So we can see that charges are greater, as a line-dependency is visually rising faster
Now lets for our interest find out at what age is the most 'smoker' age

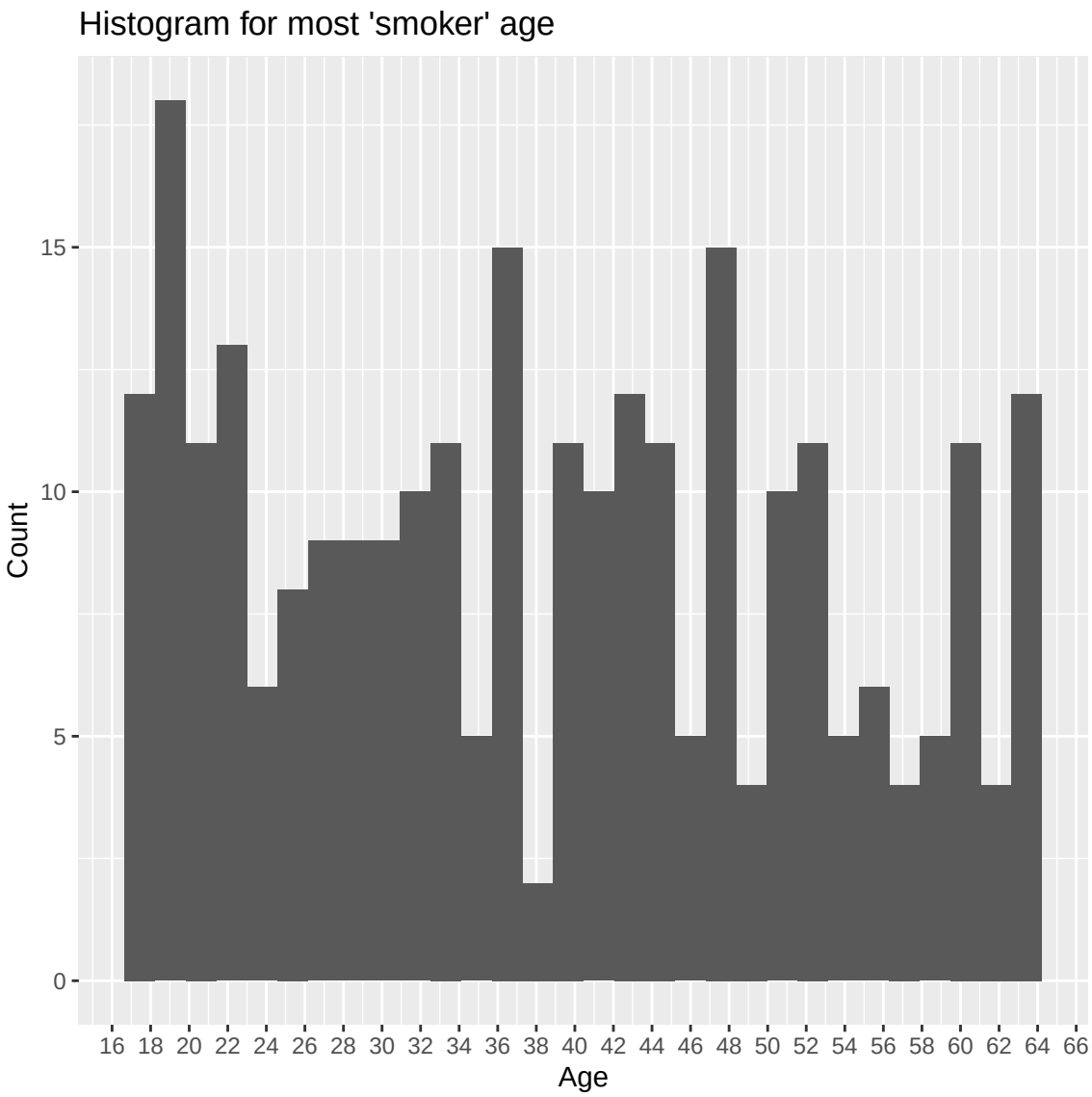

Figure 23: Histogram for most 'smoker' age"

So around 19 is the most smoking age (based on our data)

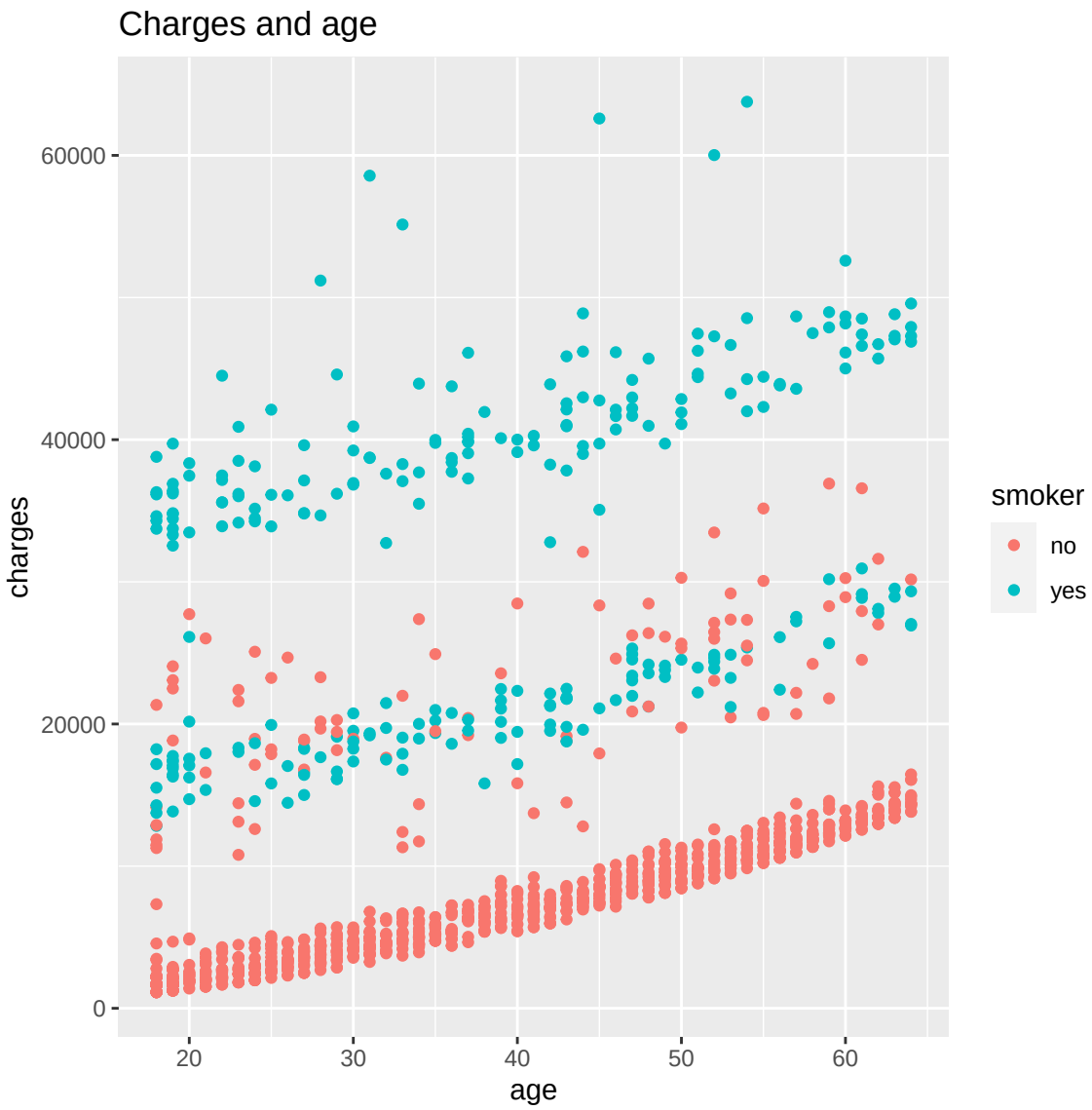

Figure 24: Charges on age

Summing up on smoking

So based on the analysis we can say, that smokers in average pay more for treatment then non-smokers, smokers can be divided into two groups - those with normal BMI (30) and those with high BMI, the second group is more affected by deceases and in the end pays more for treatment. (Or insurance is more in our case)

7.2 Gender and charges

Let's see is gender connected with charges

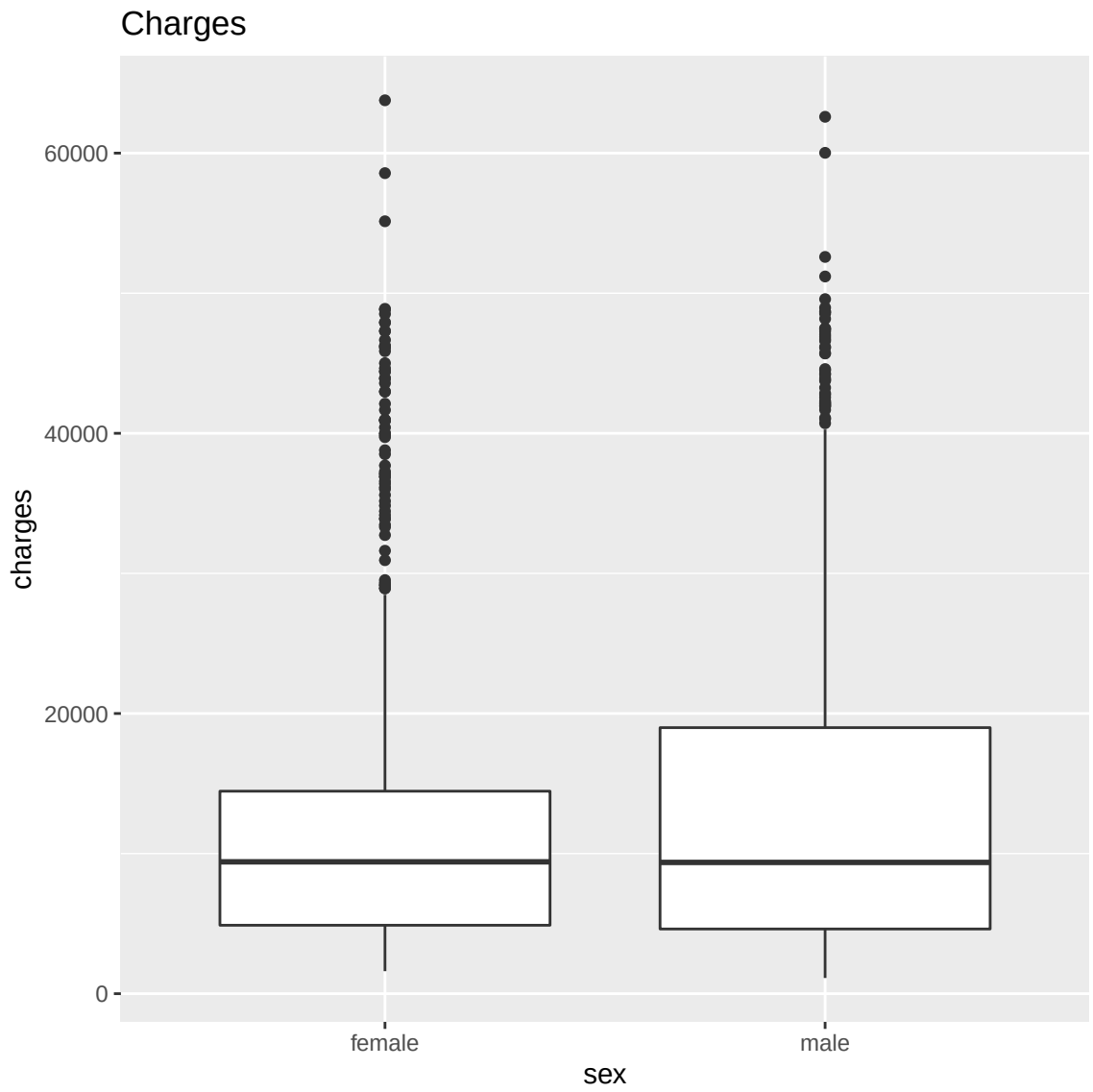

Figure 25: Gender and charges

So it's not correlated, and gender doesn't affect charges

7.3 BMI and charges

Summing up we have

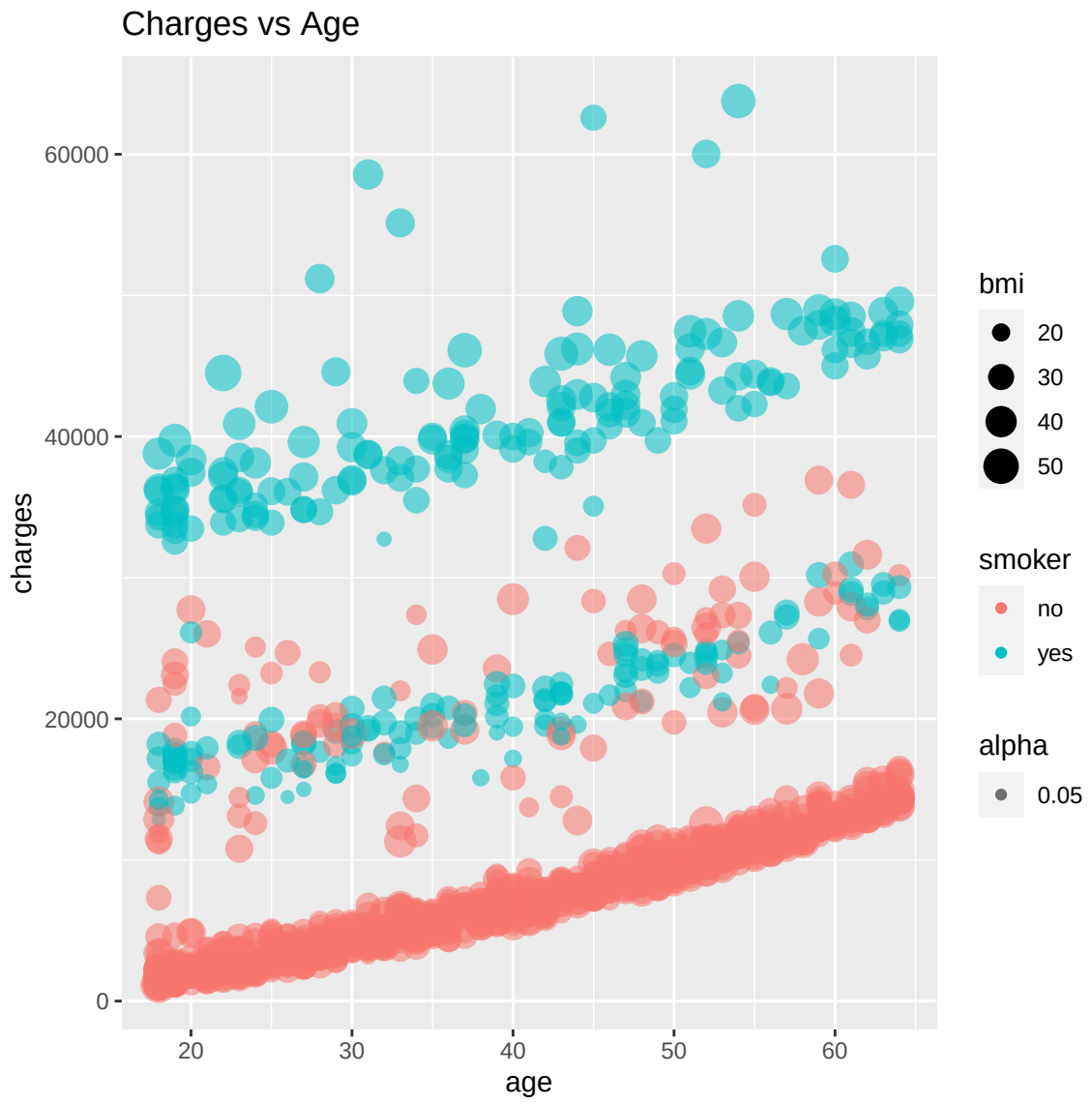

Figure 26: Age and Charges

```
> cor(data$bmi,data$charges)
```

```
[1] 0.198341
```

Correlation is 0.198 which means that there is a positive correlation between two variables, but it is weak and likely unimportant.

Let's divide bmi data into obese and not (More than 30 bmi is obese)

```
> bmiMoreOrLessThan30 <- ifelse(data$bmi>=30,"yes","no")
> ggplot(data = data,aes(bmiMoreOrLessThan30,charges)) + geom_boxplot(fill = c(2:3)) + ggtitle("Obesity")
```

As we can see there is no big difference in insurance costs, but those with high BMI has more outliers, which means hard deseases etc.

8 Building Regression a model

Let's build linear model using all possible variables and analyse the result

8.1 First model

```
> model_all <- lm(charges ~ age + sex + bmi + children + smoker,data)
> summary(model_all)
```

```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11837.2  -2916.7   -994.2   1375.3  29565.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12052.46     951.26 -12.670  < 2e-16 ***
age            257.73      11.90  21.651  < 2e-16 ***
sexmale       -128.64     333.36  -0.386 0.699641
bmi            322.36      27.42  11.757  < 2e-16 ***
children       474.41     137.86   3.441 0.000597 ***
smokeryes    23823.39     412.52  57.750  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7497,        Adjusted R-squared:  0.7488
F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Analyse the results for group of data to figure out which needs to be removed And which relationships are the strongest Let's remind what those columns mean, and apply it to our results:

- **Standard Error** - measures the average amount that the coefficient estimates vary from the actual average value of our response variable
  - **Expectations** -> lower number relative to estimation coefficients
  - **Reality** -> For age and bmi this values are low enough
- **T Value** -> Is a measure of how many standard deviations our coefficient estimate is far away from 0.
  - **Expectations** -> If it is far away from 0, we can reject null-hypothesis (declare that relationship) exists
  - **Reality** -> For age,bmi and smokers it seems that some kind of relationship exists, but The strongest one for a first look is with smokers (Need to mention that we are talking about Relationship with Charges)
- **Reality**
  - **Expectations** -> lower number relative to estimation coefficients
  - **Reality** -> For age and bmi this values are low enough
- **Pr(> |t|)** -> Relates to the probability of observing any value equal or larger than t. In other words, indicates, the possibility of value/relation been observed by chance
  - **Expectations** -> less than 0.05
  - **Reality** -> For all except gender is less than 0.05, which satisfies us
- **Residual standard error** -> measure of quality of linear regression fit the data (In other words it is the average amount that the response (age/bmi etc.) will deviate from the true regression line.)
  - **Expectations** -> Lower, comparing with estimate, better. Also expecting it to be normal
  - **Reality** -> Only for smokers we can see low difference, or about 25% error on guessing
- **Multiple R-squared** - > Measure of how well the model is fitting the actual data.
  - **Expectations** -> Close to 1
  - **Reality** -> 74% , not bad, but can be better
- **Adjusted R-squared** ->Adjusts Multiple R-squared for the number of variables considered
  - **Expectations** -> Close to 1
  - **Reality** -> 0.7488 = 74%

- **F-statistics** -> indicator of whether there is a relationship between our predictor and the response variables.
  - **Expectations** -> Further from 1 the better (comparing with data size and predictors)
  - **Reality** -> 798

8.2 Second model

So to start with, I'd remove not really suitable sex and children

```
> model<- lm(charges ~ age + bmi + smoker, data)
> summary(model)

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

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) -11676.83     937.57  -12.45   <2e-16 ***
age            259.55      11.93   21.75   <2e-16 ***
bmi            322.62      27.49   11.74   <2e-16 ***
smokeryes    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
```

And now check weather Residual standard error is following normal distribution

```
> res <- resid(model)
> qqnorm(res)
> qqline(res)
```

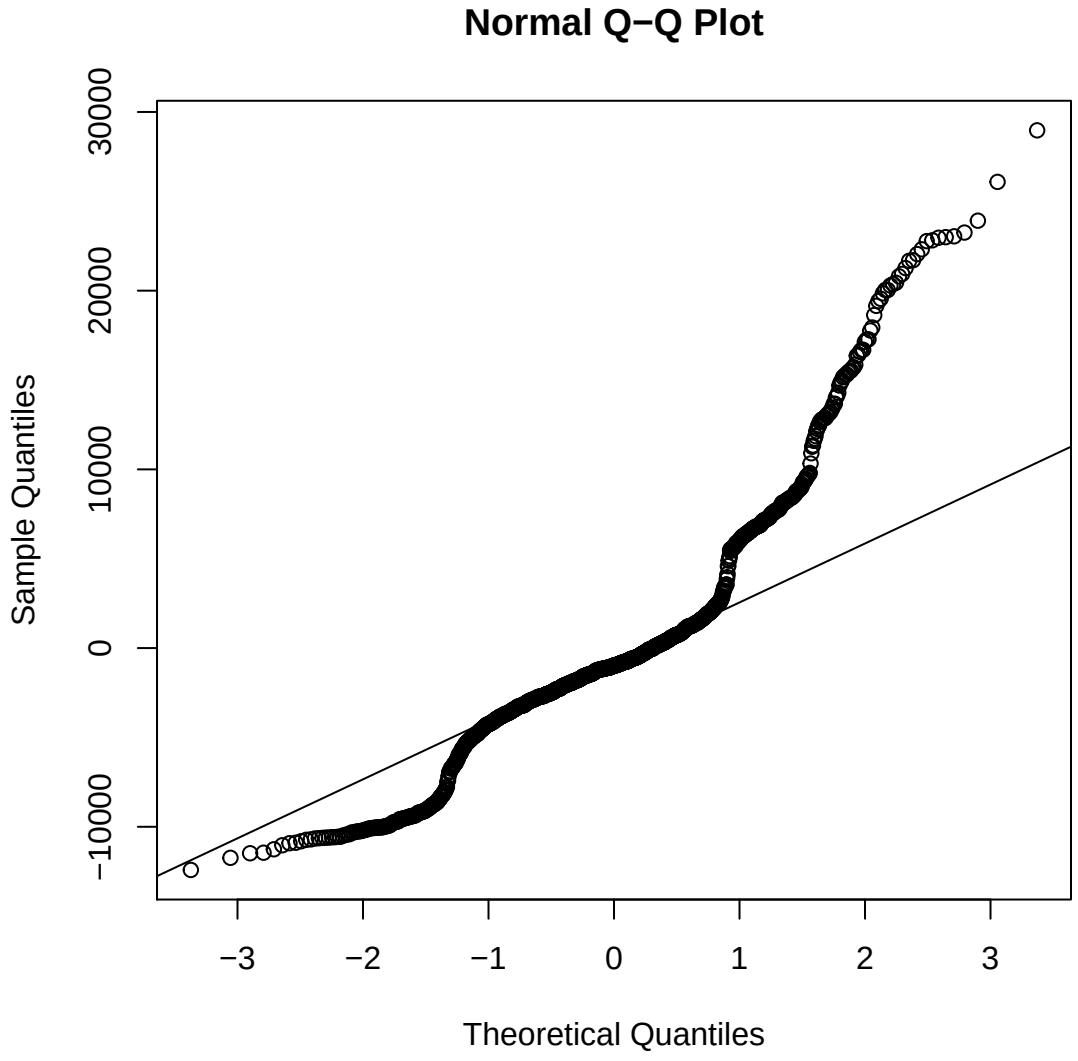

Figure 27: Residual Standard Error

As we can see, Residual standard error is not following normal distribution, and R-squared can still be better.

8.3 Third model

Now I suggest looking at Smoking closely, because it affects charges the most, But as we saw in previous analysis, smokers can be devided into two categories Those with low and high BMI. I suggest us doing that, by adding new variable -> SmokerWithHighBMI

```
> data$SmokerWithHighBMI <- ifelse(data$bmi>30
+   & data$smoker=="yes","yes","no")
> describeBy(data$charges,data$SmokerWithHighBMI)

 Descriptive statistics by group
group: no
   vars    n   mean     sd  median trimmed     mad     min      max    range
X1    1 1194 9842.6 7142.3 8338.75 8863.75 6219.68 1121.87 38245.59 37123.72
   skew kurtosis    se
X1 1.19     1.16 206.7
```

```
------------------------------------------------------------
group: yes
   vars   n     mean      sd   median  trimmed    mad      min      max
X1    1 144 41692.81 5829.16 40918.31 41225.72 5651.2 32548.34 63770.43
      range skew kurtosis     se
X1 31222.09 1.06     1.78 485.76
```

Now build the model

```
> model<- lm (charges ~ age  + smoker+ bmi + SmokerWithHighBMI,data)
> summary(model)

Call:
lm(formula = charges ~ age + smoker + bmi + SmokerWithHighBMI,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5296.9 -1973.2 -1257.8  -398.7 24230.8

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3516.435    738.288  -4.763 2.12e-06 ***
age                    266.328      8.858  30.068  < 2e-16 ***
smokeryes            13593.479    435.742  31.196  < 2e-16 ***
bmi                     47.674     22.031   2.164   0.0306 *
SmokerWithHighBMIyes 19506.680    590.858  33.014  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4521 on 1333 degrees of freedom
Multiple R-squared:  0.8611,        Adjusted R-squared:  0.8607
F-statistic:  2065 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Analysing as in previous model, we can come to Adjusted R-squared is 85,8%, which is higher than in previous model which gives us the possibility to say that data is fitting the model Errors are lower, t-values are greater, Pr($>|t|$) is slightly bigger on average, which, makes our results more random, but for all except BMI it's still less than 0.05, which is normal

Now lets build Residual standard error and check if is Following normal distribution

```
[1] 26326.32
```

```
[1] 26357.51
```

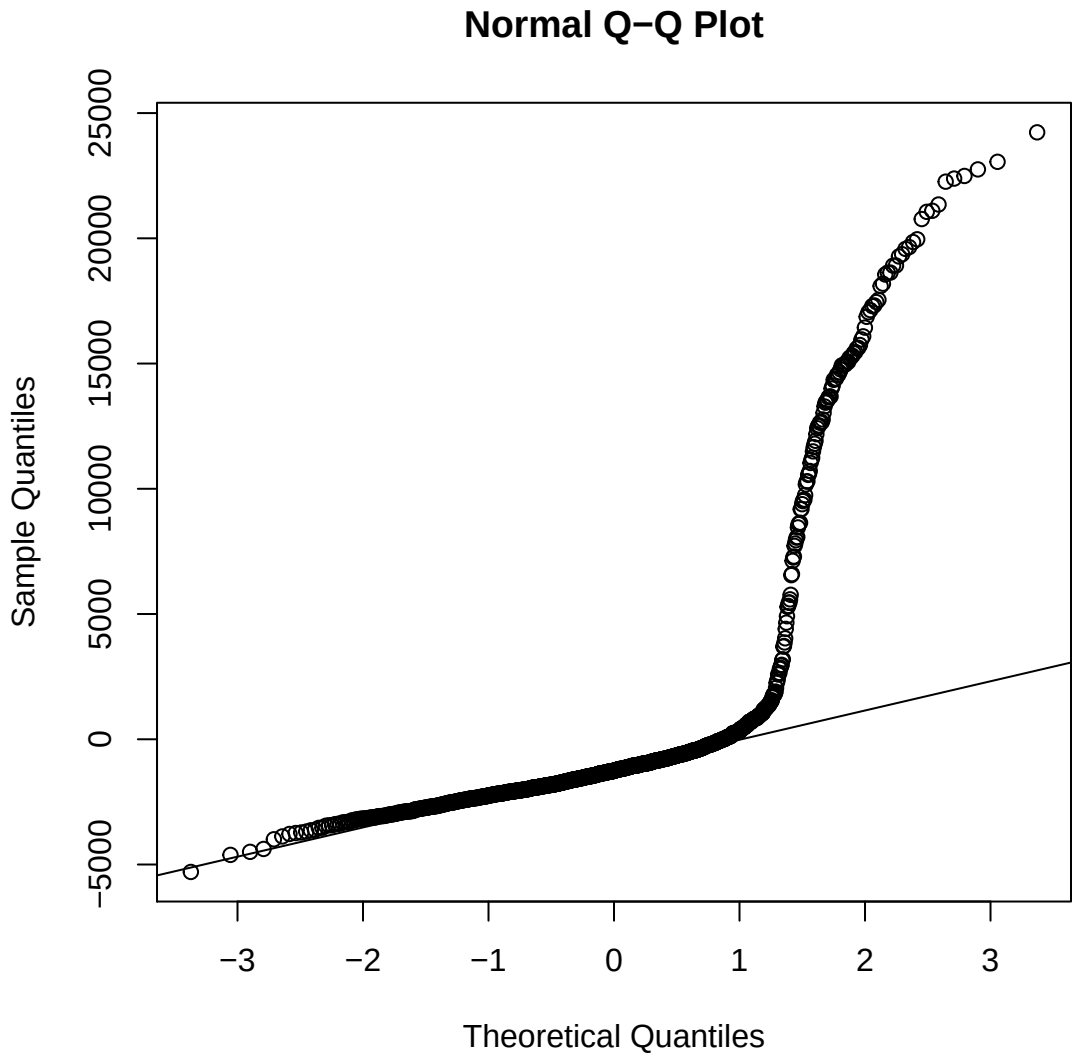

Figure 28: Residual standard error

Let's also check for plots of our model parameters
Interpreting of results:

- Residuals vs Fitted -> Shows if residuals have non-linear patterns. In our case we have a bit clustered left side which means that there are Some outliers not covered by model
- Normal Q-Q -> Shows if residuals are normally distributed As we can see there is a part when data stops following the normal distribution Which means that our model doesn't cover all the outliers, and Error while defining them can be significant !!! Which means that we need to work More on model
- Scale-Location -> The assumption of equal variance Data is not spread randomly on the line, which is not quite good
- Residuals vs Leverage -> This plot helps us to find influential cases (i.e., subjects) if any. Watch if Cook's distance is high

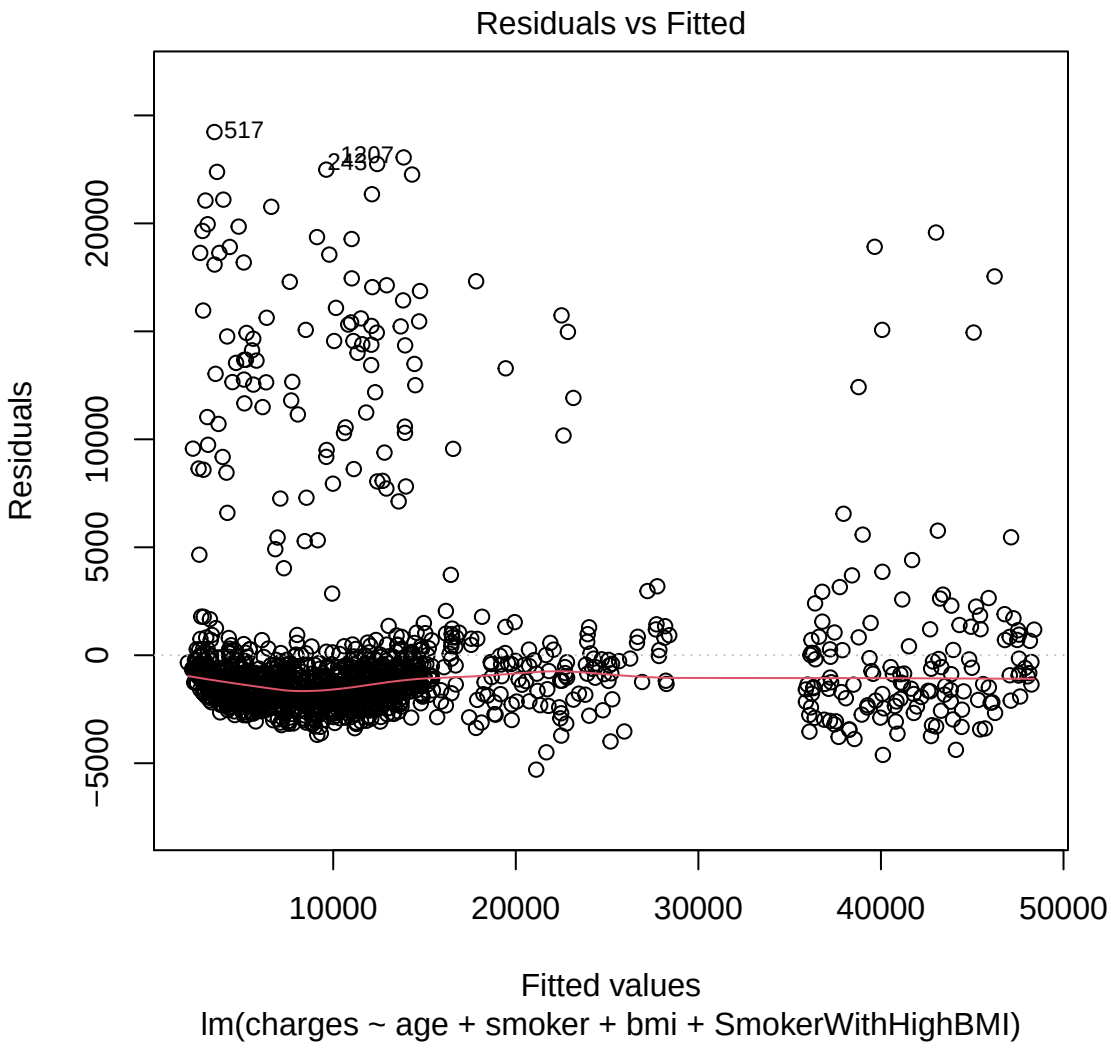

Figure 29: Residuals vs Fitted

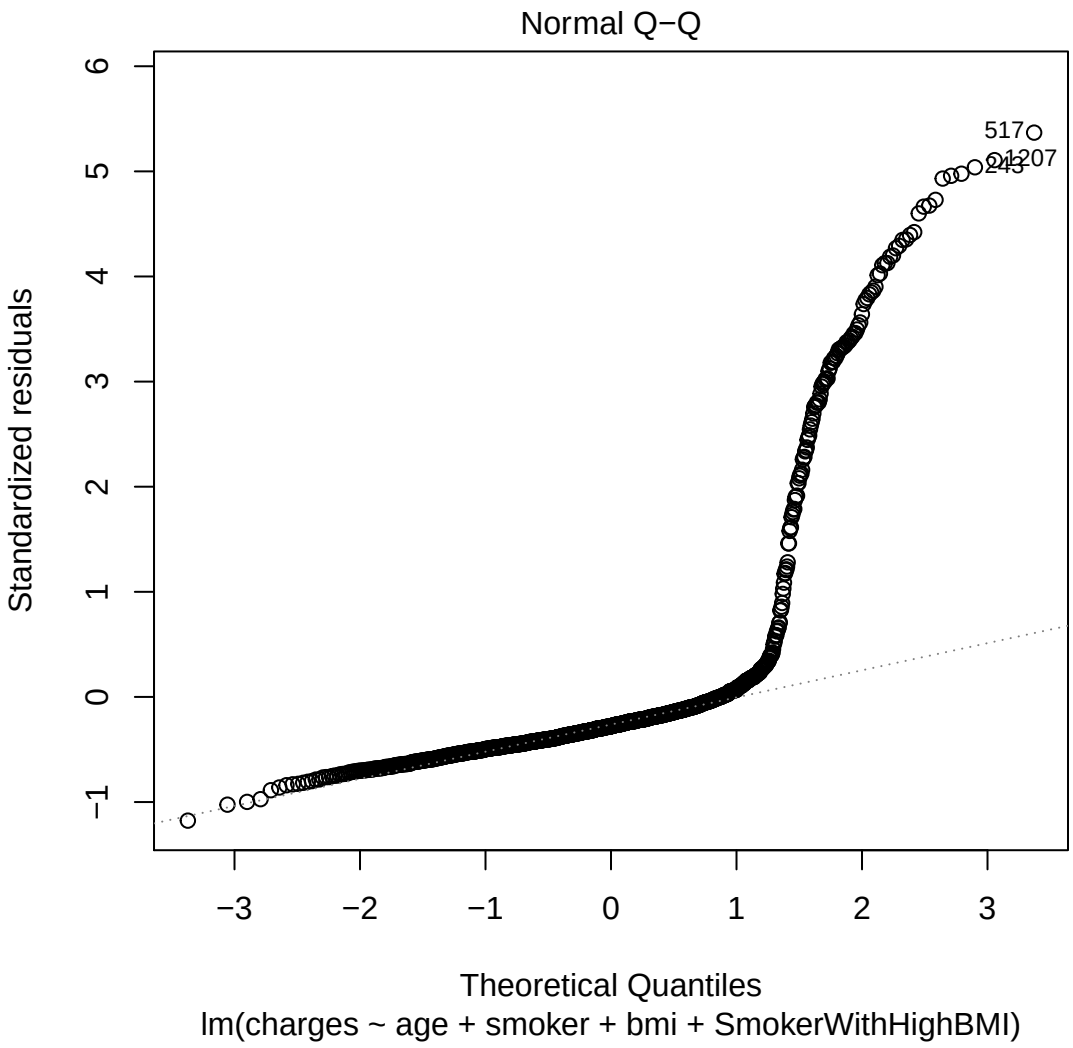

Figure 30: Normal Q-Q

Figure 31: Scale-Location

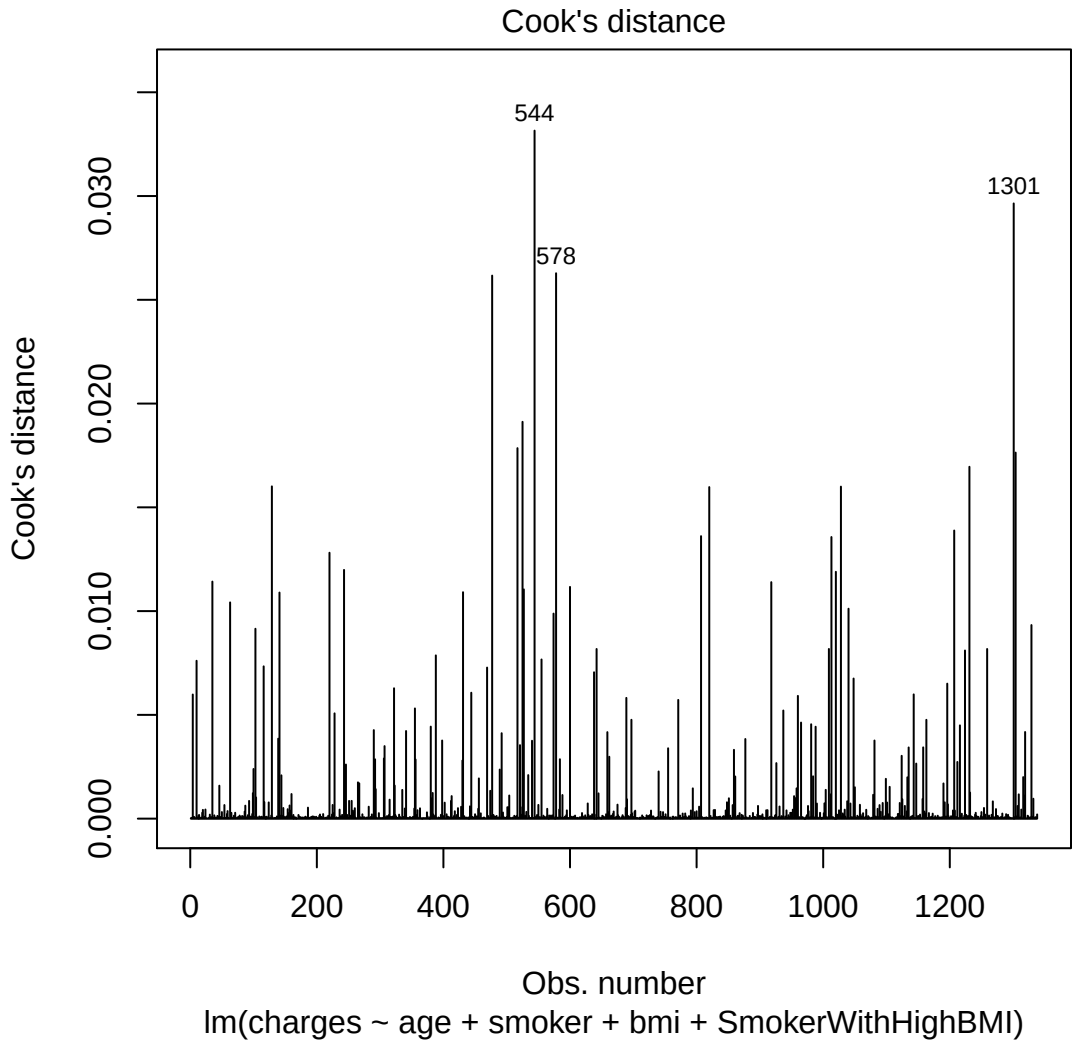

Figure 32: Residuals vs Leverage

Because we are not realy satisfied with result, let's search more:

8.4 Fourth model

```
> model<- lm (log(charges) ~ age  + age^2 +
+               smoker+ bmi + SmokerWithHighBMI + children,data)
> summary(model)

Call:
lm(formula = log(charges) ~ age + age^2 + smoker + bmi + SmokerWithHighBMI +
    children, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90987 -0.17960 -0.04327  0.04700  2.12107

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.2433744  0.0707134  102.43   <2e-16 ***
age                  0.0349966  0.0008428   41.52   <2e-16 ***
smokeryes            1.2155367  0.0414294   29.34   <2e-16 ***
```

```
bmi                  0.0018016  0.0020946    0.86     0.39
SmokerWithHighBMIyes 0.6248198  0.0561753   11.12   <2e-16 ***
children             0.1020739  0.0097599   10.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4298 on 1332 degrees of freedom
Multiple R-squared:  0.7824,        Adjusted R-squared:  0.7816
F-statistic: 957.7 on 5 and 1332 DF,  p-value: < 2.2e-16
```

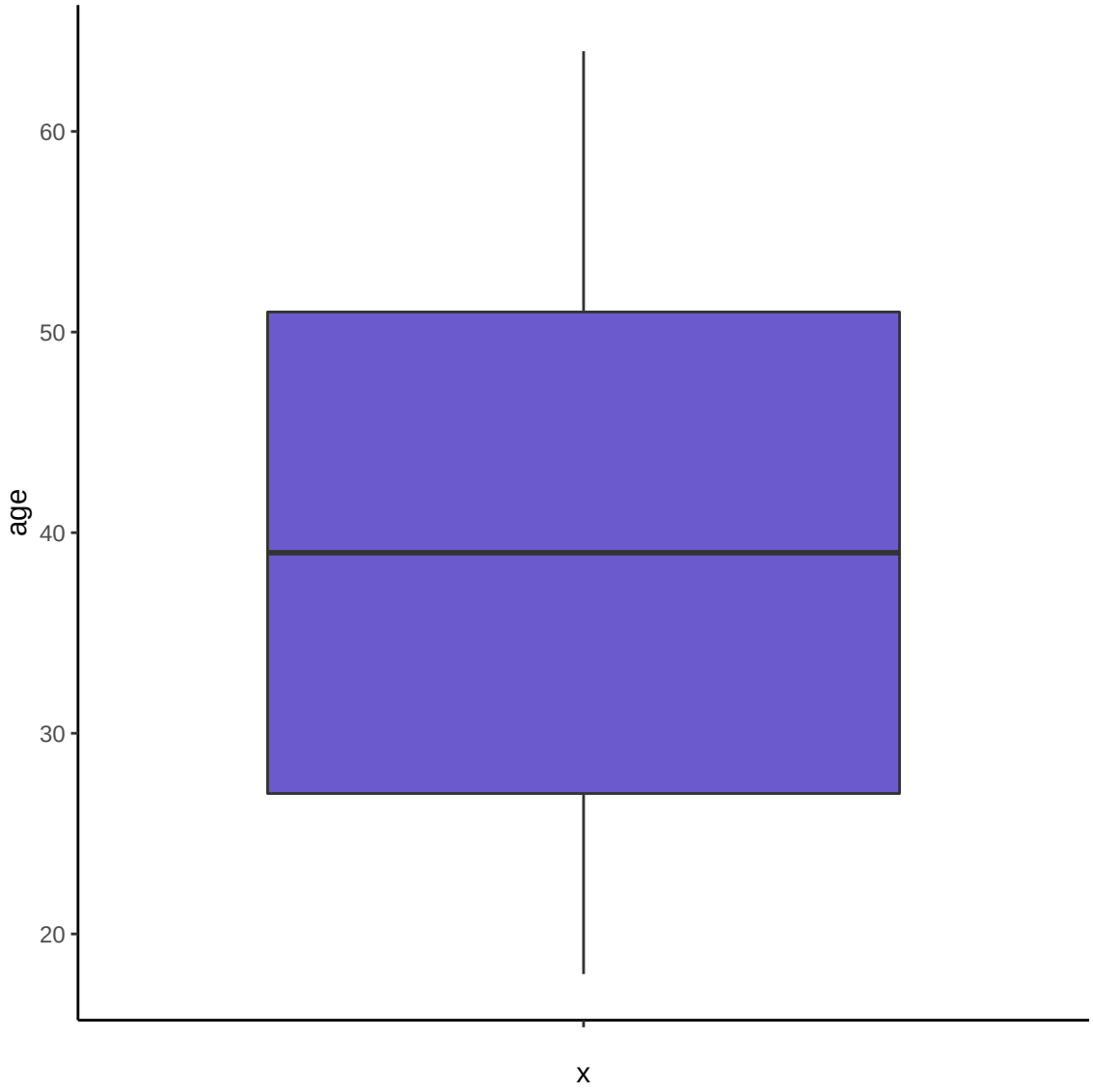

Figure 33: Residuals vs Fitted

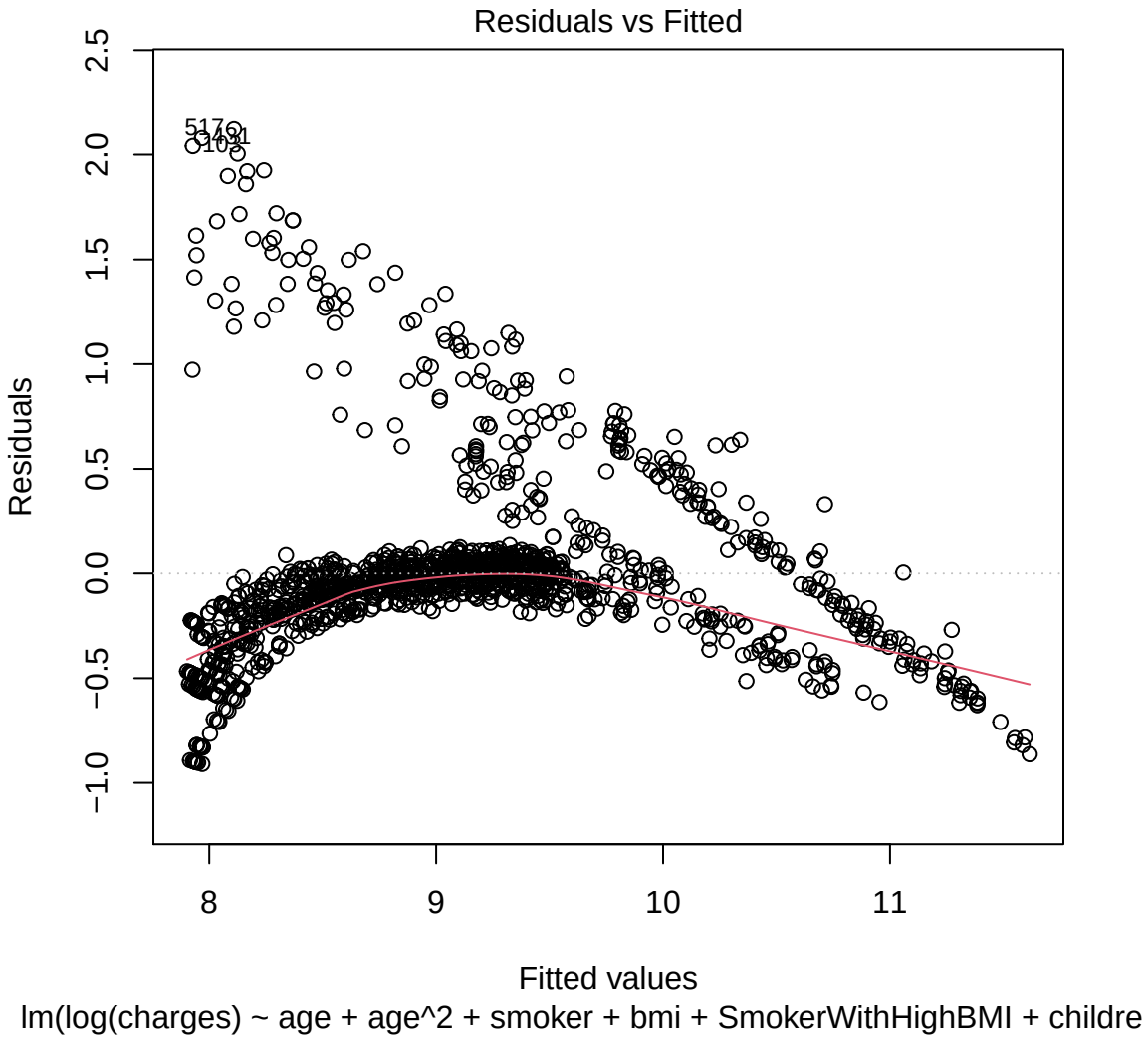

Figure 34: Residuals vs Fitted

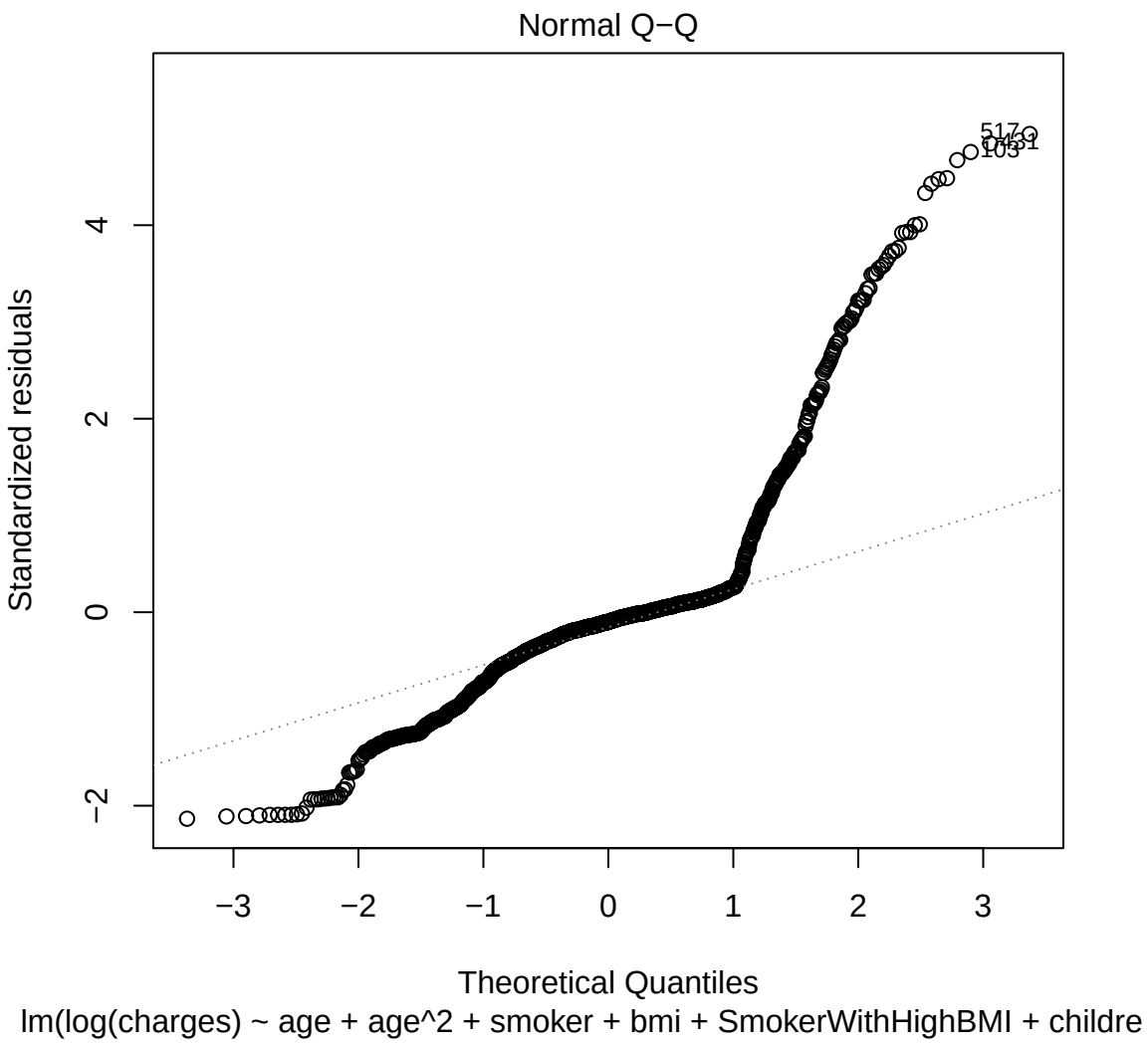

Figure 35: Normal Q-Q

Figure 36: Scale-Location

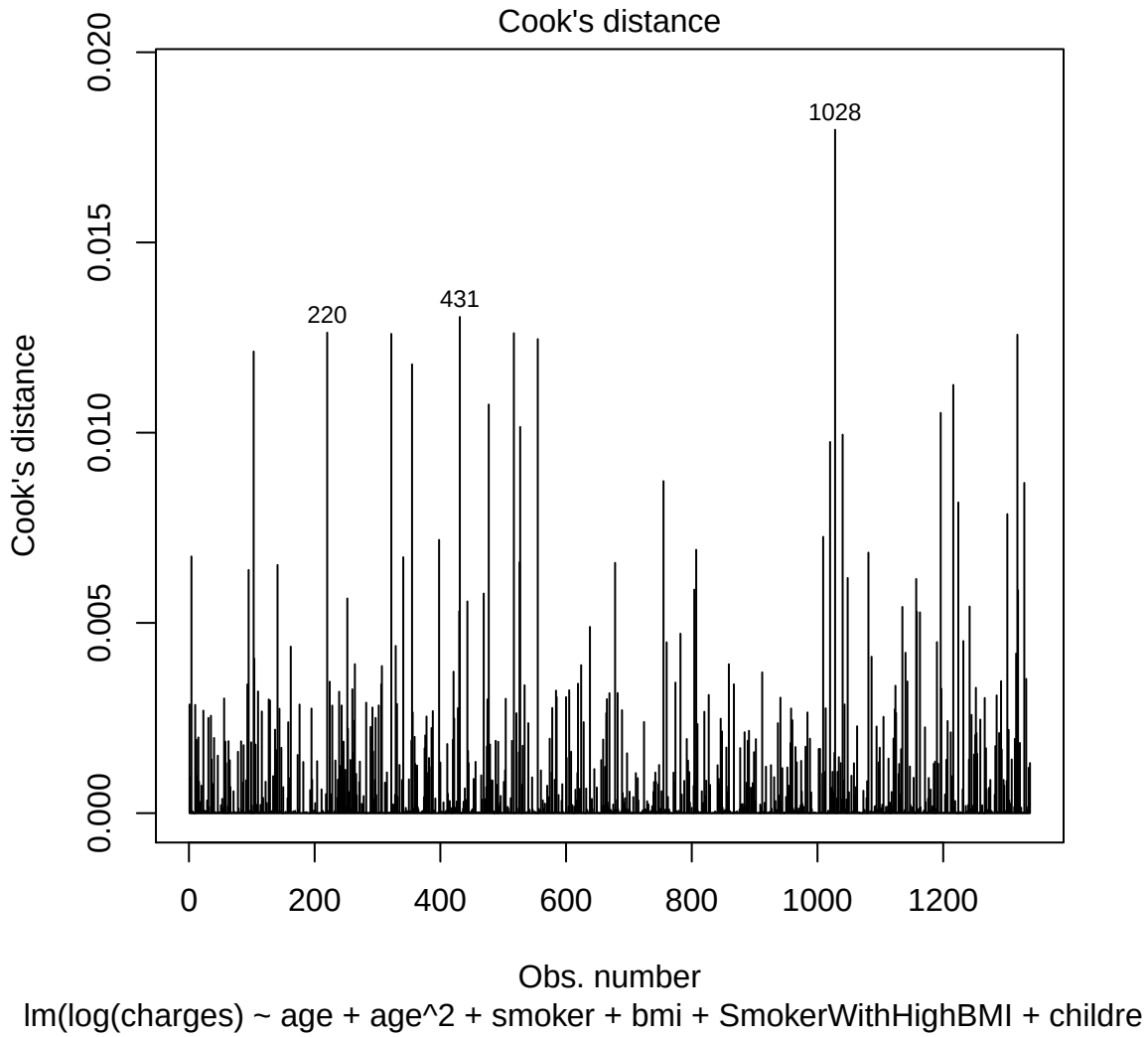

Figure 37: Residuals vs Leverage

As we can see, data on graphics is better distributed, but we still have problems with Normal Q-Q, which is not normal Residuals are essentially the difference between the actual observed response values. So we need them to be normally distributed across 0

8.5 Summing up on models

Choosing the best model from those, that we have built, I would choose the Third One 8.3. Because it has averagely better results comparuing to others, but still it doesn't fit all the data, espessially those with high BMI.

9 Conclusion

As we can see on the Figure 26 there are next dependencies

- Charges depend lineary on Age
- Charges depend on smoking status, you pay, 2 or 4 (if you also have high BMI) times more than those with middle BMI (around 30) and non-smoking
- Charges depend on BMI, but BMI affects less
- Charges doesn't depend on gender