Health Care Cost Prediction

Part I. Statistics and EDA

Tatiana Ilyasova

31 January 2022

Introduction

In this paper, I present the data analysis and the predictive modeling of a cost of individual health insurance. The work is divided into two parts. In the first part, I did statistics and exploratory data analysis. In the second part, I used the feature importance algorithm and built the different linear regression models. This work is not a presentation, but a notebook describing the very process of how data can be analyzed before building a model. Perhaps some of the charts and the tables are redundant. This is done to show different options. For the final presentation, I would change the number and the design of plots and tables. The dataset source: kaggle.com

The target variable is a cost of individual health insurance (the variable ‘charges’).

Columns

age: age of primary beneficiary (discrete)

sex: insurance contractor gender, female, male (categorical)

bmi: body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9 (continuous)

children: number of children covered by health insurance / number of dependents (discrete)

smoker: smoking (categorical)

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest (categorical)

charges: individual medical costs billed by health insurance (continuous, target)

Exploratory data analysis

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tidyverse) # to clean and manipulate a data + ggplot2
library(psych) # to get some useful descriptive statistic
library(corrplot) # to plot a correlation's matrix
library(gridExtra) # to arrange multiple plots on a page
library(lemon) # to work with legends and axis lines of 'ggplot2', facets, and some 'knitr'
               # extensions
library(knitr) # to create a better design of a report
options(scipen = 100, digits = 2)
```

```
# Load and verify the data as-is
df_insurance <- as_tibble(read.csv('insurance.csv', stringsAsFactors = T))
kable(head(df_insurance), align = 'rcrrccr')
```

age	sex	bmi	children	smoker	region	charges
19	female	28	0	yes	southwest	16885
18	male	34	1	no	southeast	1726
28	male	33	3	no	southeast	4449
33	male	23	0	no	northwest	21984
32	male	29	0	no	northwest	3867
31	female	26	0	no	southeast	3757

```
str(df_insurance)
```

```
## tibble [1,338 × 7] (S3: tbl_df/tbl/data.frame)
##  $ age     : int [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: int [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
```

```
# check the missing values and the duplicated observations
kable(data.frame(any_NA = sum(is.na(df_insurance)),
                 any_dupl = sum(duplicated(df_insurance))), align = 'll')
```

any_NA	any_dupl
0	1

The data contains 1338 observations with no missing values. The data includes one duplicate, but we cannot be sure if this is the same person or just another person with the same values without additional information, e.g., id-number. That why we do not exclude it from the data set. The ‘bmi’ and ‘charges’ are the continuous variables. The ‘age’ and ‘children’ are the discrete variables. The ‘sex’, ‘smoker’, and ‘region’ are the categorical variables.

Feature Engineering

Since the maximum number of children is five, it is better to transform this feature into a categorical one. This helps us to visualize relationships and see whether there are some relationships with the target and other features or not.

```
# Convert the variable 'children' from integer to factor
df_insurance$children <- as.factor(df_insurance$children)
```

For more convenient work with the data, we will create one more independent variable, BMI groups, named in the data frame ‘bmi_gr’. The easiest way is to use the common practice and separate the BMI index into four groups: underweight if the bmi is lower than 18.5, healthy weight if the bmi is between 18.5 and 24.9, overweight if the bmi is between 25 and 29.9, obesity if the bmi is larger than 30. source: https://www.forbes.com/health/body/bmi-chart-for-men-and-women/ (https://www.forbes.com/health/body/bmi-chart-for-men-and-women/)

```
# add a new column 'bmi_gr'
df_insurance <- df_insurance %>%
  mutate(bmi_gr = case_when(bmi <= 18.5 ~ 'Underweight',
         bmi > 18.5 & bmi < 25 ~ 'Healthy weight',
         bmi >= 25 & bmi < 30 ~ 'Overweight',
         bmi >= 30 ~ 'Obesity'))
# reorder the levels of bmi groups in order to get the right order from underweight
# to obesity, because R uses the alphabetical order by default
df_insurance$bmi_gr <- factor(df_insurance$bmi_gr,
                        labels = c('Underweight', 'Healthy weight',
                                   'Overweight', 'Obesity'),
                        levels = c('Underweight', 'Healthy weight',
                                   'Overweight', 'Obesity'))
# check if it is the right order now
head(df_insurance$bmi_gr)
```

```
## [1] Overweight     Obesity        Obesity        Healthy weight Overweight
## [6] Overweight
## Levels: Underweight Healthy weight Overweight Obesity
```

```
# check if the dataframe has changed
head(df_insurance, 3)
```

age	sex	bmi	children	smoker	region	charges	bmi_gr
19	female	28	0	yes	southwest	16885	Overweight
18	male	34	1	no	southeast	1726	Obesity
28	male	33	3	no	southeast	4449	Obesity

Summary statistics

```
# statistics
summary(df_insurance)
```

```
##       age         sex           bmi     children smoker           region
##  Min.   :18   female:662   Min.   :16   0:574    no :1064   northeast:324
##  1st Qu.:27   male  :676   1st Qu.:26   1:324    yes: 274   northwest:325
##  Median :39                Median :30   2:240               southeast:364
##  Mean   :39                Mean   :31   3:157               southwest:325
##  3rd Qu.:51                3rd Qu.:35   4: 25
##  Max.   :64                Max.   :53   5: 18
##     charges                 bmi_gr
##  Min.   : 1122   Underweight   : 21
##  1st Qu.: 4740   Healthy weight:224
##  Median : 9382   Overweight    :386
##  Mean   :13270   Obesity       :707
##  3rd Qu.:16640
##  Max.   :63770
```

```
round(prop.table(table(df_insurance$bmi_gr)), 2) # the proportions of each BMI groups
```

```
##
##    Underweight Healthy weight     Overweight        Obesity
##           0.02           0.17           0.29           0.53
```

```
round(prop.table(table(df_insurance$region)), 2) # the proportions of each region
```

```
##
## northeast northwest southeast southwest
##      0.24      0.24      0.27      0.24
```

```
round(prop.table(table(df_insurance$children)), 2) # the proportions of number of children
```

```
##
##    0    1    2    3    4    5
## 0.43 0.24 0.18 0.12 0.02 0.01
```

```
round(prop.table(table(df_insurance$smoker)), 2) # the proportions of smokers and non-smokers
```

```
##
##  no yes
## 0.8 0.2
```

The number of men and women is almost equal. The four different regions are also represented equally (only SW is a little bit larger). The ratio of smokers and non-smokers is 1 to 4. The number of people being overweight, underweight, and obese, is much larger than people with a healthy weight. The number of people having no children is 43%, and it is the majority, and the number of people having four or five children is 2% and 1% respectively.

```
# the statistical measures
# reorder the columns in order to place some of them in the begging of the table
describe(df_insurance, IQR = T) %>%
  relocate(vars,n, mean, median, sd, skew, kurtosis, se, everything())
```

	vars	n	mean	median	sd	skew	kurtosis	se	trimmed	mad	min	max	range	IQR
age	1	1338	39.2	39	14.05	0.06	-1.25	0.38	39.0	17.8	18	64	46	24.0

	vars	n	mean	median	sd	skew	kurtosis	se	trimmed	mad	min	max	range	IQR
sex*	2	1338	1.5	2	0.50	-0.02	-2.00	0.01	1.5	0.0	1	2	1	1.0
bmi	3	1338	30.7	30	6.10	0.28	-0.06	0.17	30.5	6.2	16	53	37	8.4
children*	4	1338	2.1	2	1.21	0.94	0.19	0.03	1.9	1.5	1	6	5	2.0
smoker*	5	1338	1.2	1	0.40	1.46	0.14	0.01	1.1	0.0	1	2	1	0.0
region*	6	1338	2.5	3	1.10	-0.04	-1.33	0.03	2.5	1.5	1	4	3	1.0
charges	7	1338	13270.4	9382	12110.01	1.51	1.59	331.07	11076.0	7440.8	1122	63770	62649	11899.6
bmi_gr*	8	1338	3.3	4	0.81	-0.84	-0.41	0.02	3.4	0.0	1	4	3	1.0

The data looks good. There are no extreme values such as age being equal to 150 or charges over one billion.

The age's mean and median are equal, and the skewness is near zero. The variable 'age' is normally distributed. The BMI's mean and median are equal, and the skewness is between -0.5 and 0.5. The variable 'BMI' is normally distributed. However, 53% of observations have obesity (BMI greater than 30), and mean and median of BMI is 30.66 and 30.43 respectively. The charges' mean is larger than the median, and the skewness is larger than one and is positive. The distribution of 'charges' is highly positive-skewed or right-tailed.

We calculate the average age, BMI and cost of medical insurance by gender and save these metrics into a new data frame.

```
# find the average of cost, bmi and age for men and women
# copy the data, compute the mean of 'charges', 'bmi', 'age' for men and women separately
avg_by_sex <- df_insurance %>%
  group_by(sex) %>%
  summarise(avg_cost = round(mean(charges),0),
            avg_bmi = round(mean(bmi),0),
            avg_age = round(mean(age),0),)
# compute the proportions of smoking men and smoking women
smoking_sex <- df_insurance %>%
  group_by(sex) %>%
  count(sex,smoker) %>%
  mutate(is_smoker = round(n/sum(n) * 100,2))  %>%
  filter(smoker == 'yes')
# add a column with the proportion of men and women who are smokers
avg_by_sex <- cbind(avg_by_sex, is_smoker_freq = smoking_sex$is_smoker)
avg_by_sex
```

sex	avg_cost	avg_bmi	avg_age	is_smoker_freq
female	12570	30	40	17
male	13957	31	39	24

In this data, the average age, BMI, and cost of medical insurance is almost equal when comparing men and women. The proportion of smoking men among all men is higher than that of women, but not by a lot.

Let’s examine the observations, having a minimum and a maximum cost of health insurance.

```
df_insurance %>%
  filter(charges == min(charges) | charges == max(charges))
```

age	sex	bmi	children	smoker	region	charges	bmi_gr
54	female	47	0	yes	southeast	63770	Obesity
18	male	23	0	no	southeast	1122	Healthy weight

As a result, we got that a non-smoking man, 18 years old, having a healthy weight, has the cheapest insurance, costing him ~1122 USD, meanwhile a smoking woman, 54 years old, having obesity, has the most expensive insurance, costing her ~63770 USD. They both have no children.

Now we look at potential outliers if they are any in the data.

```
outl_higher <- 16640 + 1.5 * 11899.63 # Q3 + 1.5IQR
outl_lower <- 4740 - 1.5 * 11899.63 # Q1 - 1.5IQR
kable(data_frame(outl_lower, outl_higher), align = c('ll'))
```

outl_lower	outl_higher
-13109	34489

```
# subset the data, including all the observations greater than the upper cutoff
out <- nrow(df_insurance %>%
       filter(df_insurance$charges > outl_higher)) # count number of observations
print(paste(round(out/nrow(df_insurance) * 100, 2), '%')) # % of outliers
```

```
## [1] "10.39 %"
```

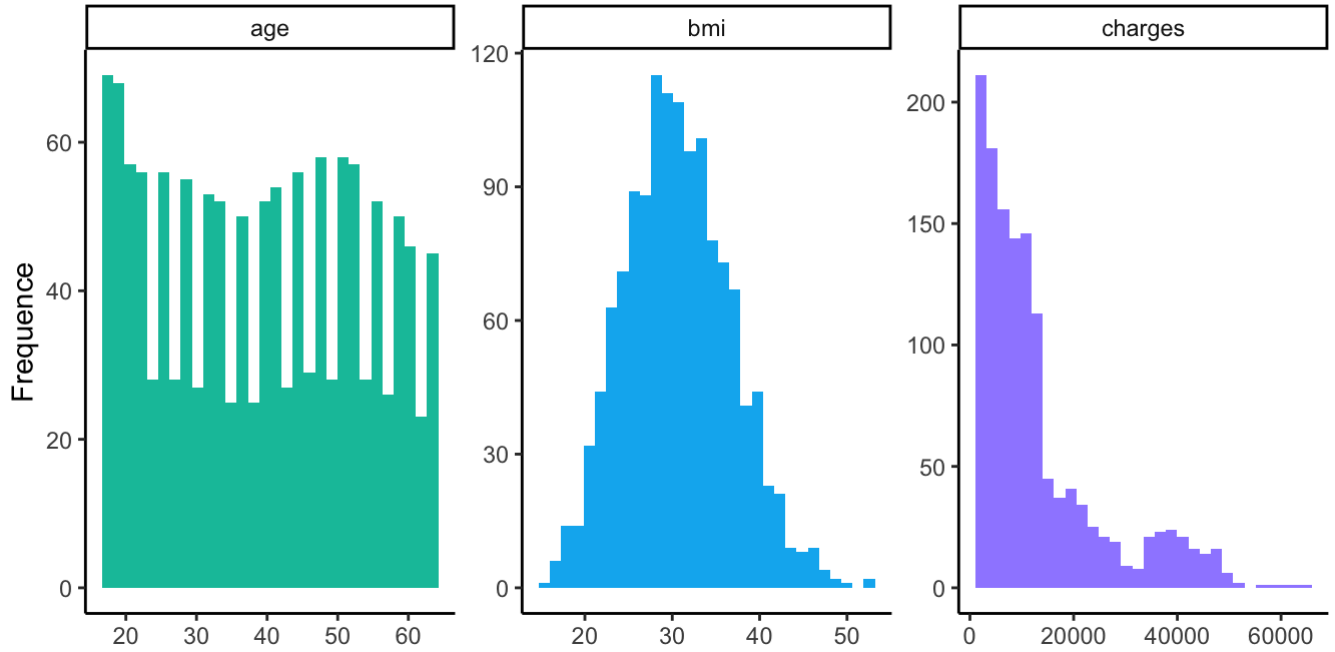
As expected, there are many observations on the right side of the distribution of the cost of medical insurance, and they are the extreme values. The total number of outliers is 139 observations which are 10,39% of the data.

Visualization

Cost of Medical Insurance

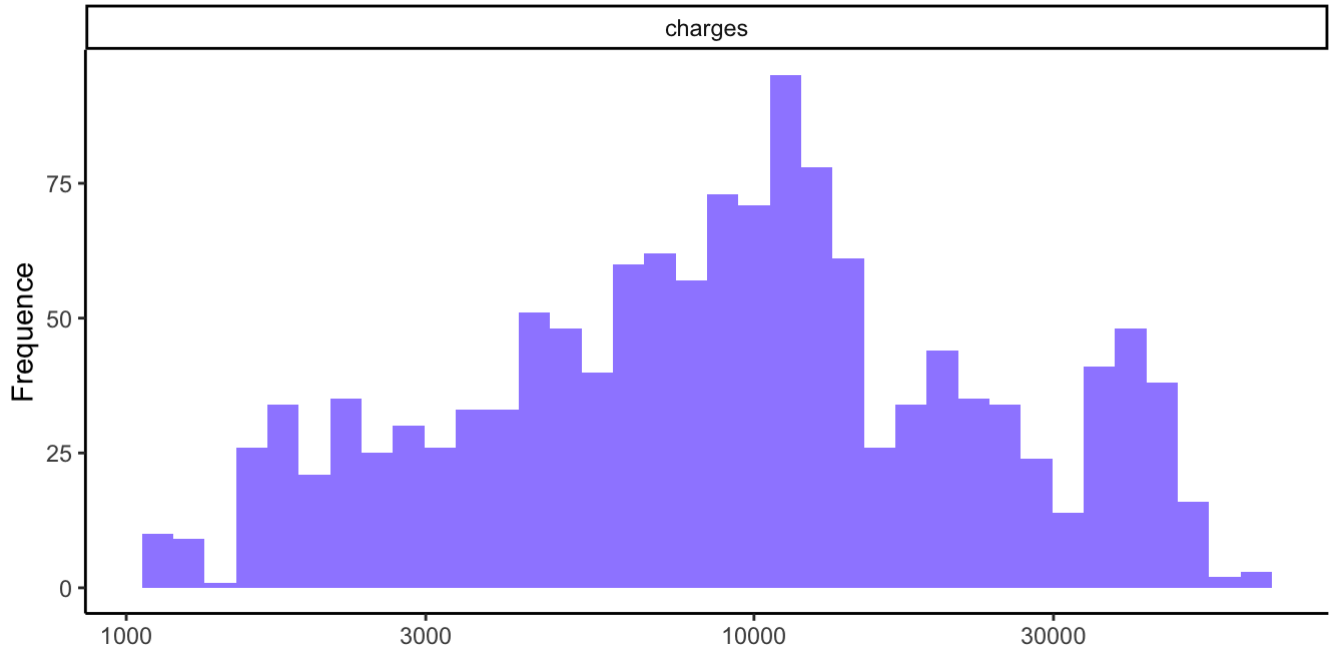
Let’s start with plotting distributions.

```
# copy the data and convert the categorical variables from factor to integer
# transform a new data frame so that all the columns gathered to one column 'Variables' and all the values
# to the second column ' Values'
hist_all <- df_insurance[,c(1,3,7)] %>%
  mutate_if(is.factor, as.integer) %>%
  pivot_longer(col = 1:3, names_to = 'Variables', values_to = 'Values')
# plot all the distributions
ggplot(hist_all, aes(Values)) +
  geom_histogram(aes(fill = Variables)) +
  scale_fill_hue(h = c(180, 270)) +
  facet_wrap(Variables~., scale = 'free') +
  theme_classic() +
  labs(y = 'Frequence', x = NULL) +
  theme(legend.position = 'none')
```



If we use a logarithmic scale, the distribution of the medical costs looks normal.

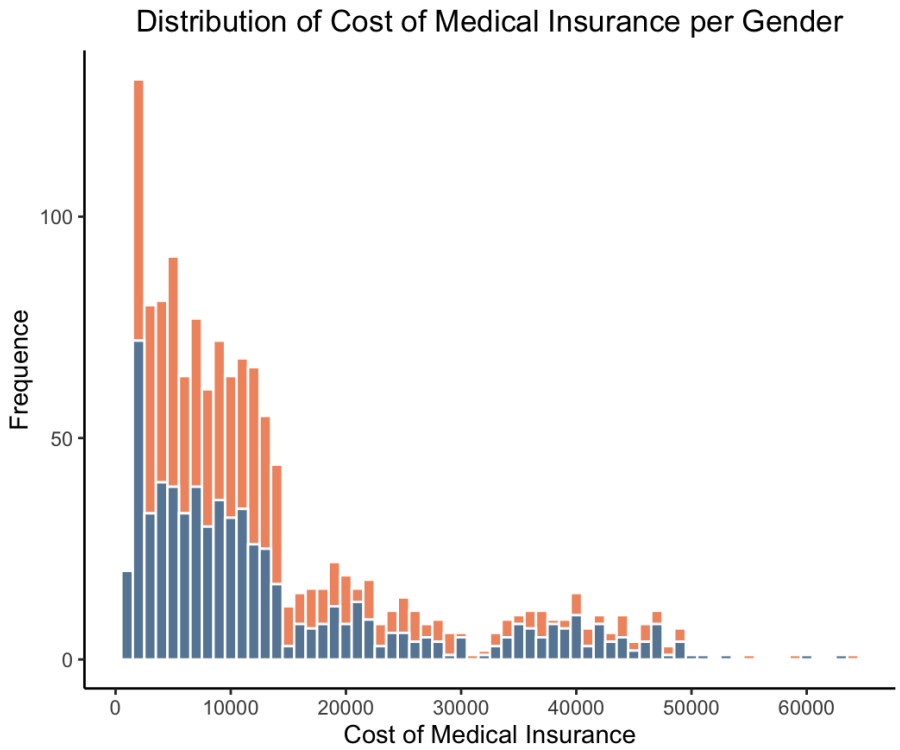
```
# copy the previous data frame and select only 'charges'
log_hist <- hist_all %>%
  filter(Variables == 'charges')
# plot the distribution
ggplot(log_hist, aes(Values)) +
  geom_histogram(aes(fill = Variables), binwidth = 0.05) +
  scale_x_log10() + # log10 scale
  scale_fill_hue(h = c(270, 270)) +
  facet_wrap(~Variables, scale = 'free') +
  theme_classic() +
  labs(y = 'Frequence', x = NULL) +
  theme(legend.position = 'none')
```

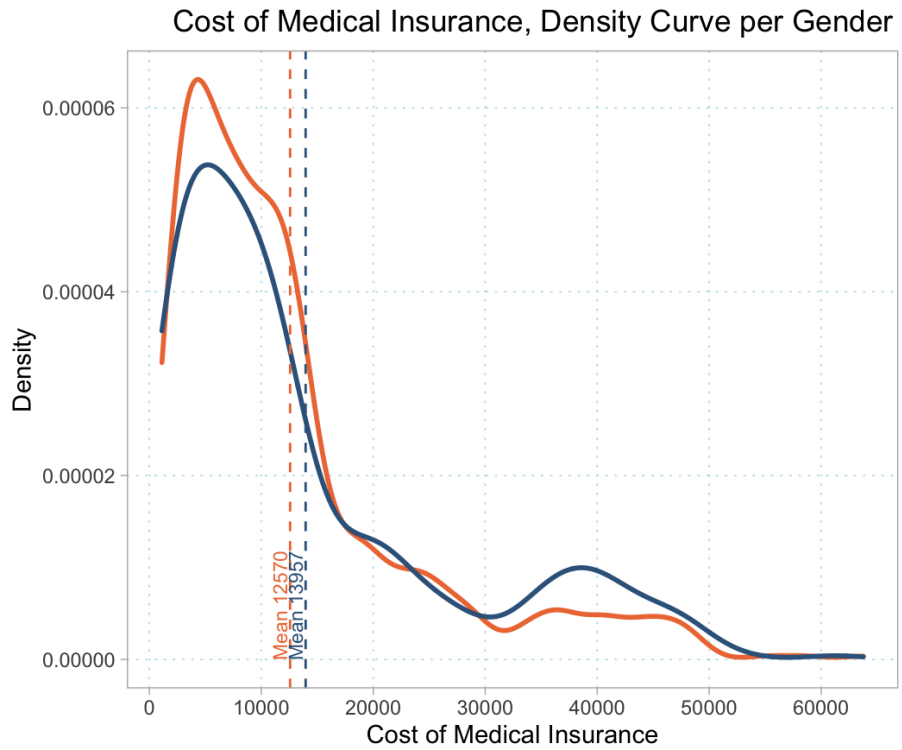


The target is ‘charges’. Let’s look at a distribution and a density curve of this variable again, but now we split the data by gender. Additionally, we add lines of the average cost of MI per gender on the density plot.

```
# choose and save the colors to use them in plots
sex_color <- c('sienna2', 'steelblue4')
bmi_color <- c('gold', 'skyblue', 'royalblue4','firebrick1')
```

```
# histogram
hist_charg <- ggplot(df_insurance, aes(charges)) +
  geom_histogram(aes(fill = sex), binwidth = 1000, alpha = 0.8, col = 'white') +
  scale_fill_manual(values = sex_color, name = NULL) +
  scale_x_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  labs(x = 'Cost of Medical Insurance', y = 'Frequence',
       title = 'Distribution of Cost of Medical Insurance per Gender') +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'bottom')
# density curves
density_charg <- ggplot(df_insurance, aes(charges, color = sex)) +
  geom_density(size = 1) +
  geom_vline(data = avg_by_sex, aes(xintercept = avg_cost, color=sex),
             linetype= 'dashed', size = 0.5) +
  scale_color_manual(values = sex_color, name = NULL) +
  geom_text(data = avg_by_sex, mapping = aes(x = avg_cost,
            label = paste('Mean', avg_cost), y = 0),
            angle = 90, size = 3, hjust = 0, vjust = -0.2) +
  scale_x_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_light() +
  labs(x = 'Cost of Medical Insurance', y = 'Density',
       title = 'Cost of Medical Insurance, Density Curve per Gender') +
  theme(plot.title = element_text(hjust = 0.9),
        panel.grid.major = element_line(size = 0.25,
                                        linetype = 'dotted', colour = 'lightblue'),
        panel.grid.minor = element_line(size = 0),
        legend.position = 'none')
# plot together
grid_arrange_shared_legend(hist_charg, density_charg, ncol = 2)
```

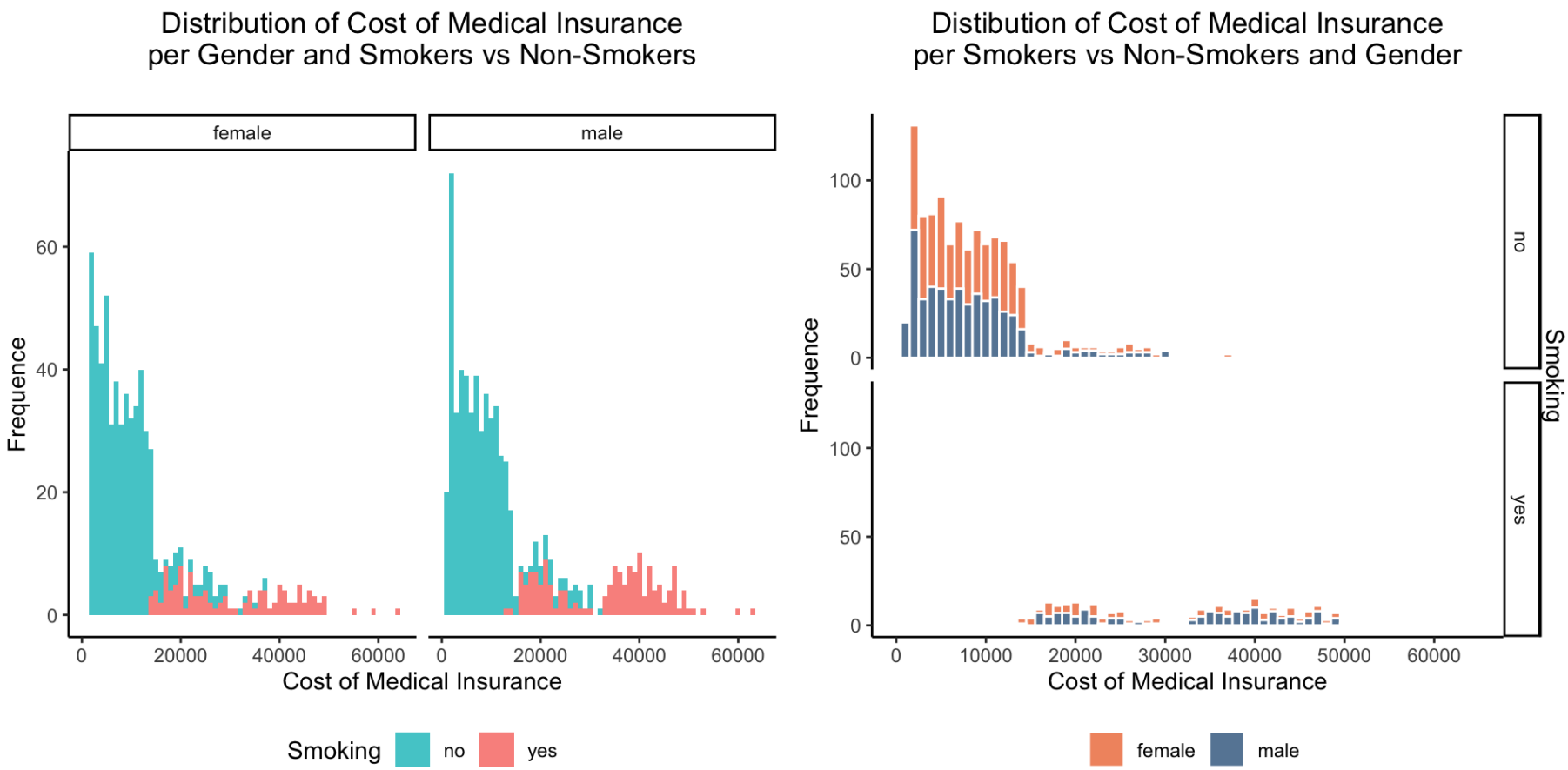




There is no significant difference between men and women.

Now we look at the distribution of the medical insurance costs by smokers vs non-smokers, and by men vs women. The difference between these two graphs is that in the first case we split the data by smoking, highlighting gender, and in the second one, we do contrariwise.

```
# plot the distribution by gender
hist_sm_gen <- ggplot(df_insurance, aes(charges)) +
  geom_histogram(aes(fill = smoker), binwidth = 1000, alpha = 0.8) +
  scale_fill_hue(direction = -1, name = 'Smoking') +
  facet_grid(~ sex) +
  theme_classic() +
  labs(x = 'Cost of Medical Insurance', y = 'Frequence',
       title = 'Distribution of Cost of Medical Insurance\nper Gender and Smokers vs Non-Smokers\n') +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'bottom')
# plot the distribution by smoking
hist_gen_sm <- ggplot(df_insurance, aes(charges)) +
  geom_histogram(aes(fill = sex), binwidth = 1000, alpha = 0.8, col = 'white') +
  scale_fill_manual(values = sex_color, name = NULL) +
  scale_x_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_y_continuous(sec.axis = sec_axis(~ . , name = 'Smoking', breaks = NULL, labels = NULL)) +
  theme_classic() +
  facet_grid(smoker ~.) +
  labs(x = 'Cost of Medical Insurance', y = 'Frequence',
       title = 'Distibution of Cost of Medical Insurance\nper Smokers vs Non-Smokers and Gender\n') +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'bottom')
# plot two plots together
grid.arrange(hist_sm_gen, hist_gen_sm, ncol = 2)
```



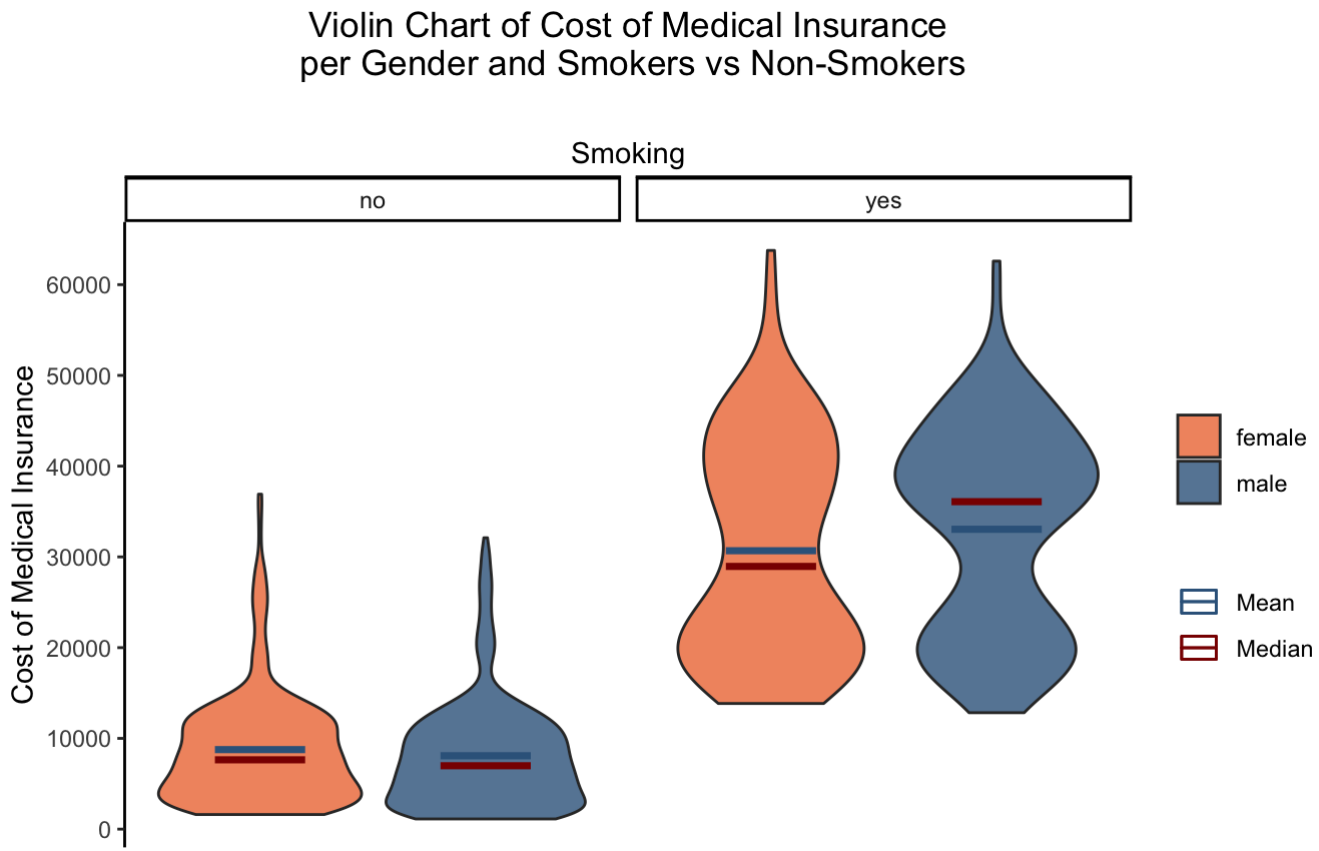
Here we see that there is the significant difference. The non-smoking persons have not been charged greater than about 40 000 USD. At the same time, the smoking persons have not been charged less than about 13 000 USD. There is no difference between men and women. Let's compute the exact upper and lower cutoffs of costs for smokers and non-smokers.

```
# select only the smokers and the minimum cost
sm_lower <- df_insurance[,c(5,7)] %>%
  filter(smoker == 'yes') %>%
  slice_min(charges)
# select only the non-smokers and the maximum cost
non_upper <- df_insurance[,c(5,7)] %>%
  filter(smoker == 'no') %>%
  slice_max(charges)
# combine these two measures to a table
kable(rbind(sm_lower, non_upper), align = 'c')
```

smoker	charges
yes	12829
no	36911

In addition, we plot a violin chart, which is a hybrid of a box plot and a kernel density plot. We add the mean and median by each segment.

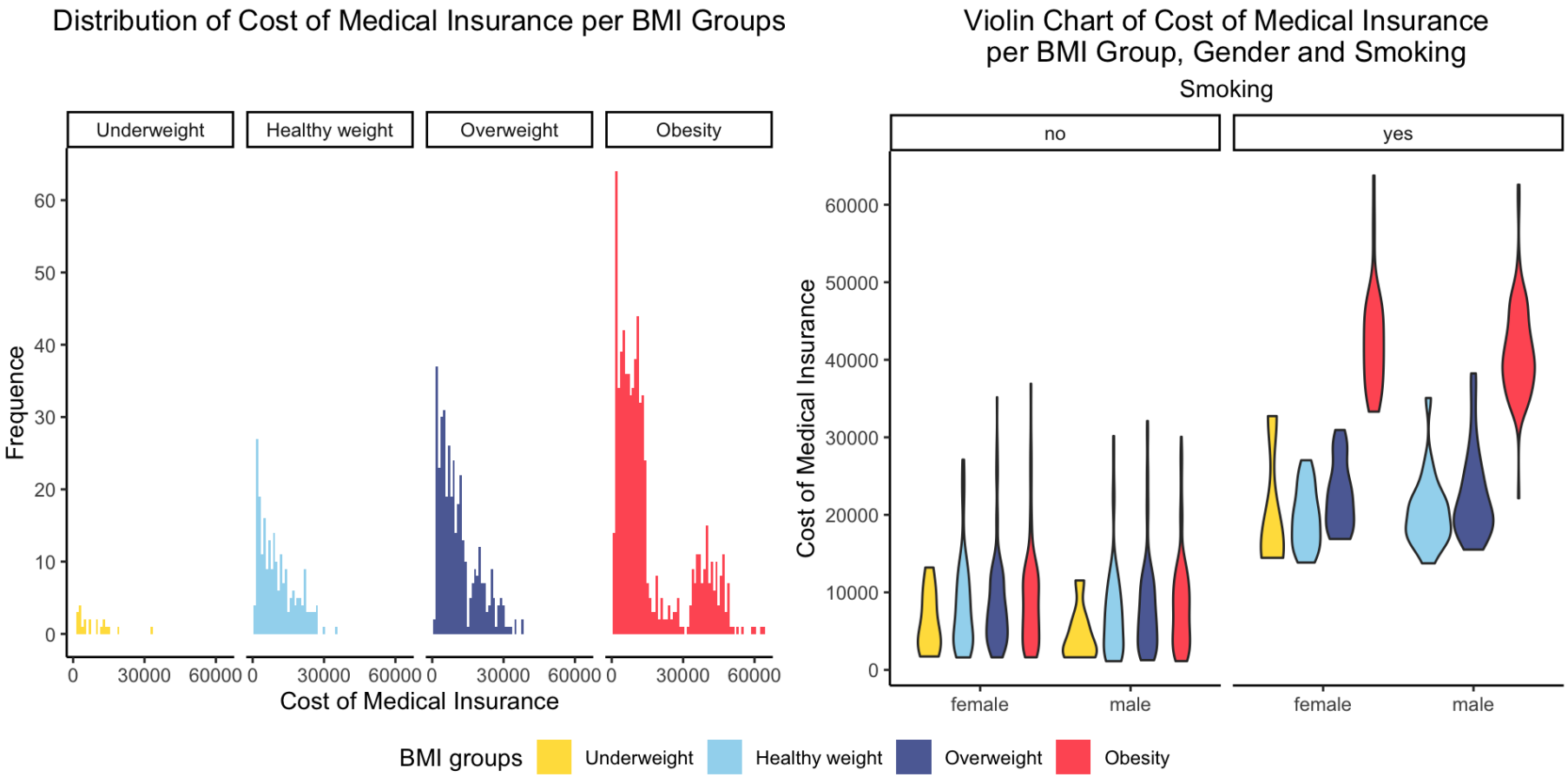
```
# violin plot of cost of MI by gender and smoking
ggplot(df_insurance, aes(sex, charges)) +
  geom_violin(aes(fill = sex), alpha = 0.8) +
  stat_summary(fun = median, geom = 'crossbar', width = 0.4, aes(color = 'Median')) + # add median
  stat_summary(fun = mean, geom = 'crossbar', width = 0.4, aes(color = 'Mean')) + # add mean
  scale_fill_manual(values = sex_color, name = NULL) +
  scale_colour_manual(values = c('steelblue4', 'red4'), name = NULL) + # customize the colors of mean and median
  facet_grid(~ smoker) + #split data by smoking
  theme_classic() +
  labs(y = 'Cost of Medical Insurance', x = 'Smoking',
       title = 'Violin Chart of Cost of Medical Insurance\n per Gender and Smokers vs Non-Smokers\n') +
  scale_x_discrete(position = 'top') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x.top = element_blank()) +
  guides(colour = guide_legend(order = 2),
            fill = guide_legend(order = 1))
```



Comparing men and women, we can not see any significant difference. However, as we have already seen, there is the significant difference between smokers and non-smokers.

Let's look at the distribution of the cost of medical insurance, splitting the data by BMI groups.

```
# histogram of cost of MI by BMI groups
hist_bmi_gr <- ggplot(df_insurance, aes(charges)) +
  geom_histogram(aes(fill = bmi_gr), binwidth = 1000, alpha = 0.8) +
  scale_fill_manual(values = bmi_color, name = 'BMI groups') + # highlight by BMI groups
  facet_grid(~ bmi_gr) + # split by BMI groups
  scale_x_continuous(breaks = seq(0, 65000, by = 30000)) +
  scale_y_continuous(breaks = seq(0, 65, by = 10)) +
  theme_classic()  +
  labs(x = 'Cost of Medical Insurance', y = 'Frequence',
       title = 'Distribution of Cost of Medical Insurance per BMI Groups\n\n') +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'none')
# violin plot of cost of MI by gender, smoking, BMI groups
violin_bmi_gr <- ggplot(df_insurance, aes(sex, charges)) +
  geom_violin(aes(fill = bmi_gr), alpha = 0.8) + # highlight by BMI groups
  scale_fill_manual(values = bmi_color, name = 'BMI groups') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  facet_wrap(~ smoker) + # split by smoking
  theme_classic() +
  labs(y = 'Cost of Medical Insurance', x = NULL,
       title = 'Violin Chart of Cost of Medical Insurance\nper BMI Group, Gender and Smoking',
       subtitle = 'Smoking') +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'bottom')
# plot together
grid_arrange_shared_legend(hist_bmi_gr, violin_bmi_gr, ncol = 2)
```



Only one BMI group, “Underweight”, has the significant difference between men and women. It seems that there are no smoking men or just a few, being underweight in this data. There is a significant difference in how the costs are scattered in each group. According to the graph, only people having obese have medical insurance, costing them more than 35 000 USD. Let’s compute this and look if it is true.

```
df_insurance %>%
  filter(bmi_gr != 'Obesity' & charges > 35000) %>% # select all BMI groups, execpt 'Obesity' & costs more than 35 000 US
        D
  arrange(charges)
```

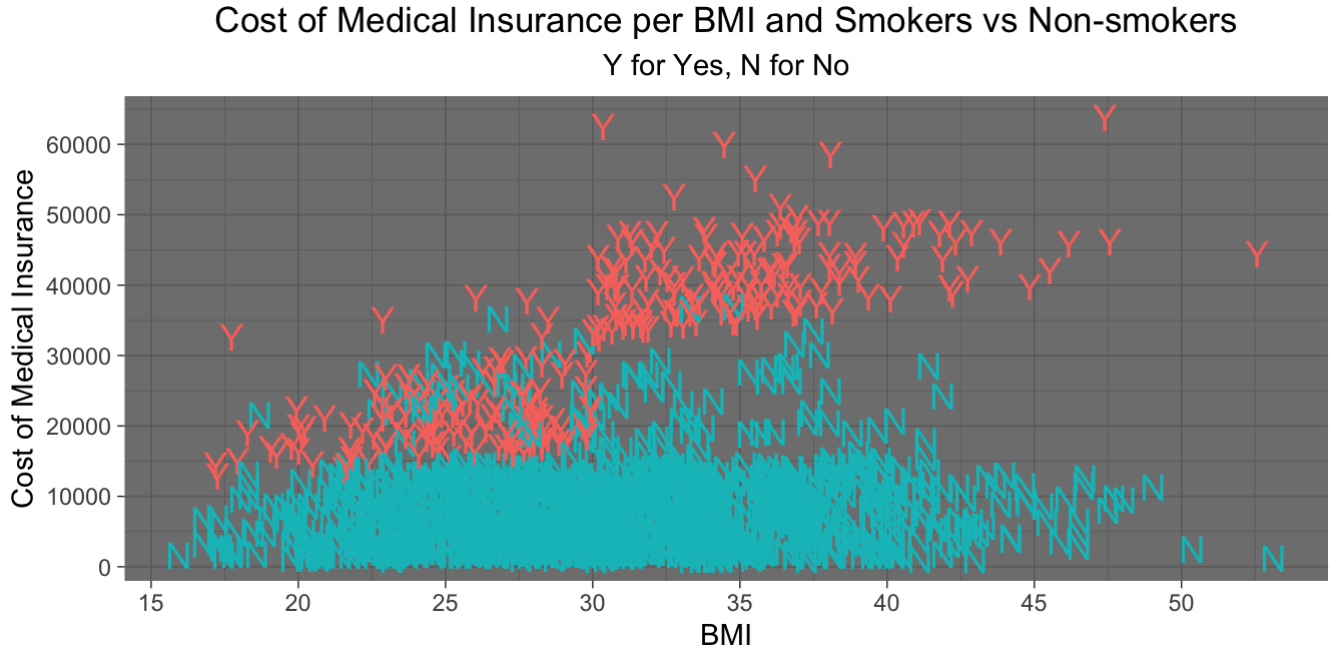
age	sex	bmi	children	smoker	region	charges	bmi_gr

age	sex	bmi	children	smoker	region	charges	bmi_gr
45	male	23	0	yes	northeast	35069	Healthy weight
24	male	28	0	yes	northeast	35148	Overweight
55	female	27	1	no	southwest	35160	Overweight
43	male	28	0	yes	southwest	37830	Overweight
42	male	26	1	yes	southeast	38246	Overweight

Five persons have been charged greater than 35 000 USD. Four of them have overweight, and only one man is a healthy weight, but he is a smoker. Moreover, four persons are smokers.

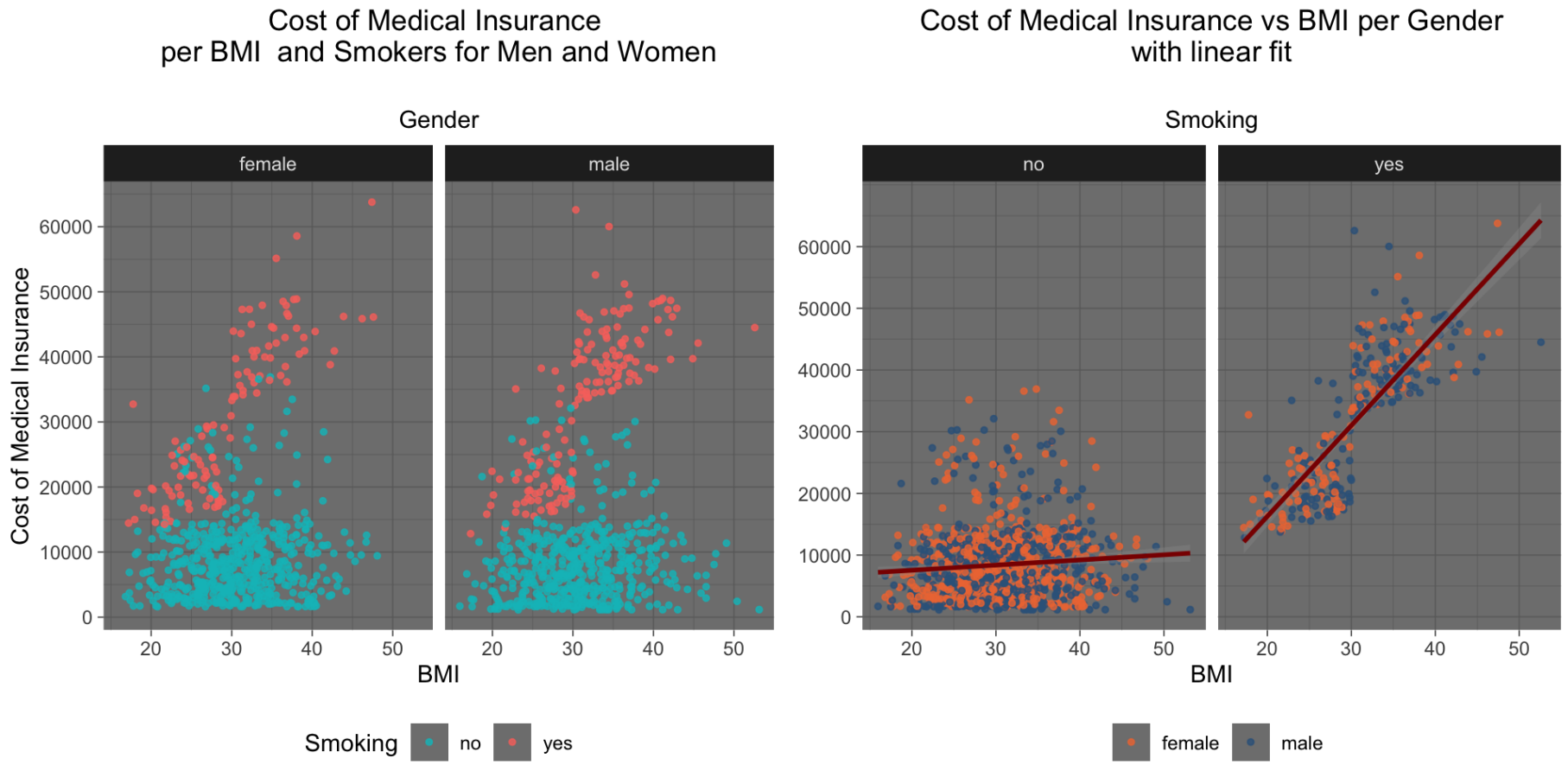
Now we plot the continuous variable BMI, and highlight smokers and non-smokers with color and the symbols ‘Y’ and ‘N’.

```
ggplot(df_insurance, aes(bmi, charges)) +
  geom_point(aes(color = smoker, shape = smoker, size = 0.7)) +
  scale_color_hue(direction = -1) +
  scale_shape_manual(values = c(78,89)) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_x_continuous(breaks = seq(0, 60, by = 5)) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'BMI',
       title = 'Cost of Medical Insurance per BMI and Smokers vs Non-smokers',
       subtitle = 'Y for Yes, N for No') +
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'none')
```



Here we see an insightful pattern. The graph shows that there is a linear relationship between cost and BMI when it comes to the smokers. Let’s look at the same relationship, but this time we split the data by gender on one plot, and by smoking on the second one.

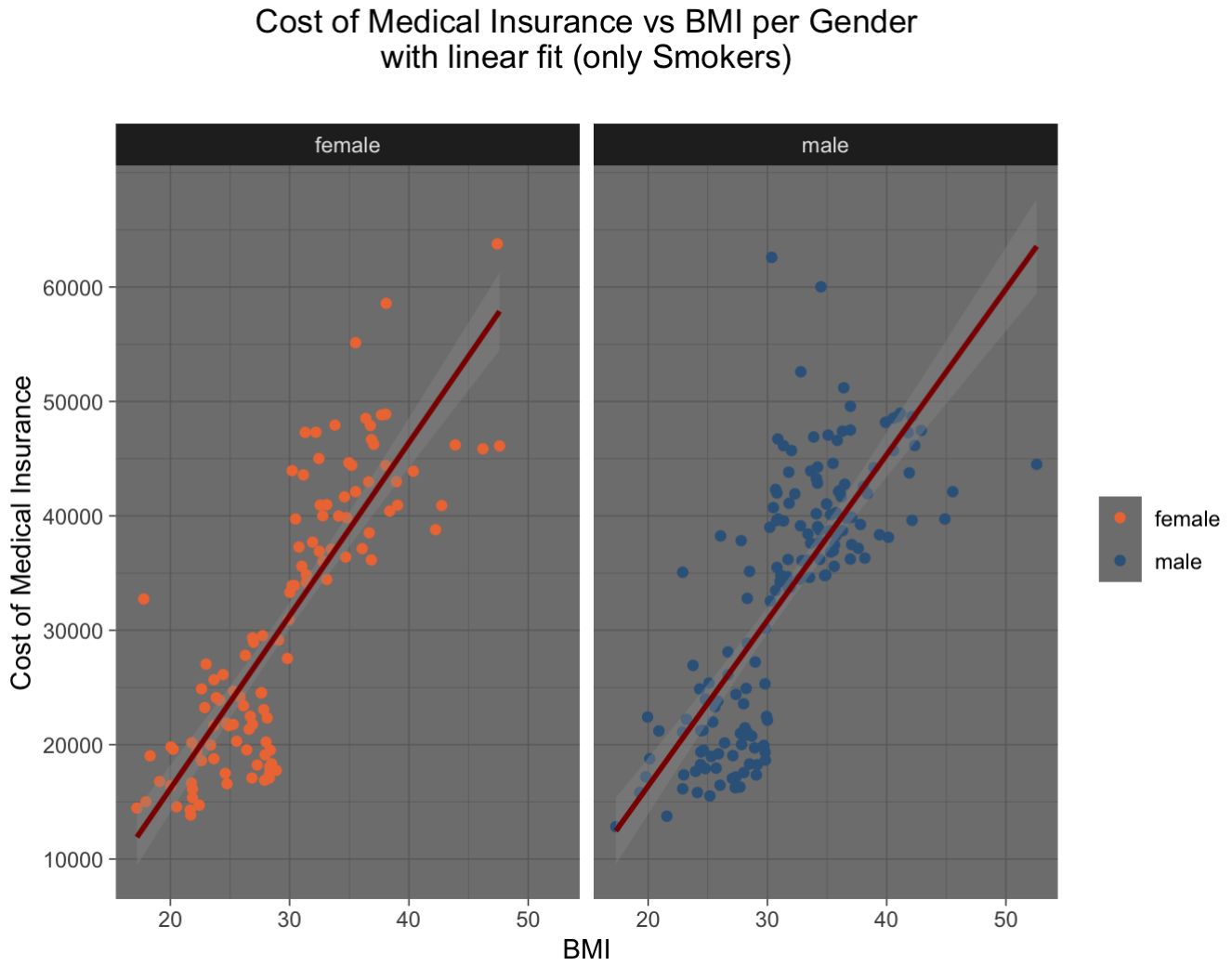
```
plt_bmi_sm <- ggplot(df_insurance, aes(bmi, charges)) +
  geom_point(aes(color = smoker), alpha = 0.8, size = 1) +
  scale_color_hue(direction = -1, name = 'Smoking') +
  facet_grid(~ sex) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'BMI',
       title = 'Cost of Medical Insurance \nper BMI  and Smokers for Men and Women\n',
       subtitle = 'Gender') +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'bottom')

plt_bmi_gen <- ggplot(df_insurance, aes(bmi, charges)) +
  geom_point(aes(color = sex), alpha = 0.8, size = 1) +
  scale_color_manual(values = sex_color, name = NULL) +
  stat_smooth(method = lm, se = T, formula = y ~ x, color = 'red4') +
  facet_grid(~ smoker) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_dark() +
  labs(y = NULL, x = 'BMI',
       title = 'Cost of Medical Insurance vs BMI per Gender\nwith linear fit\n',
       subtitle = 'Smoking') +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = 'bottom')

grid.arrange(plt_bmi_sm, plt_bmi_gen, ncol = 2)
```



There is no difference between men and women, but we see the linear relationship between the cost and BMI among the smokers.

Let's subset the data frame, selecting only the smoking men and women, and plot only selected observations.

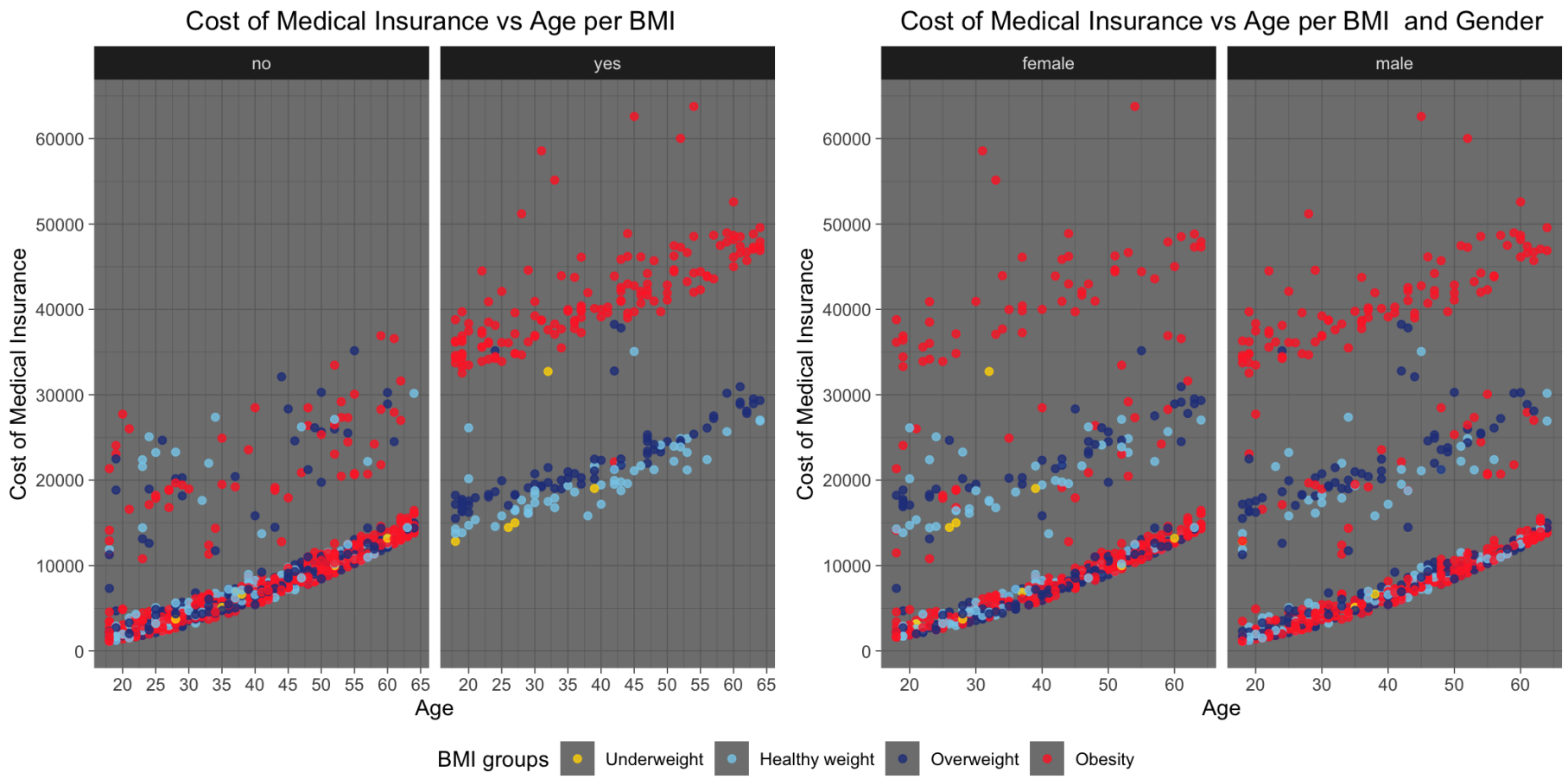
```
# subset all the smokers and save them to a new data frame
smoker_bmi <- df_insurance[,c(2,3,5,7)] %>%
  filter(smoker == 'yes')

ggplot(smoker_bmi, aes(bmi, charges)) +
  geom_point(aes(color = sex)) +
  scale_color_manual(values = sex_color, name = NULL) +
  stat_smooth(method = lm, se = T, formula = y ~ x, color = 'red4') +
  facet_grid(~ sex) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_dark()  +
  labs(y = 'Cost of Medical Insurance', x = 'BMI',
       title = 'Cost of Medical Insurance vs BMI per Gender\nwith linear fit (only Smokers)\n') +
  theme(plot.title = element_text(hjust = 0.5))
```



Yes, there is a strong linear correlation between cost and BMI among the smokers in both gender groups without any difference between men and women. Both groups look identically.

The next continuous variable is age. Let's look at the relationship between the cost of medical insurance vs age per BMI group on one plot, and per gender on the second one. We plot two graphs together.

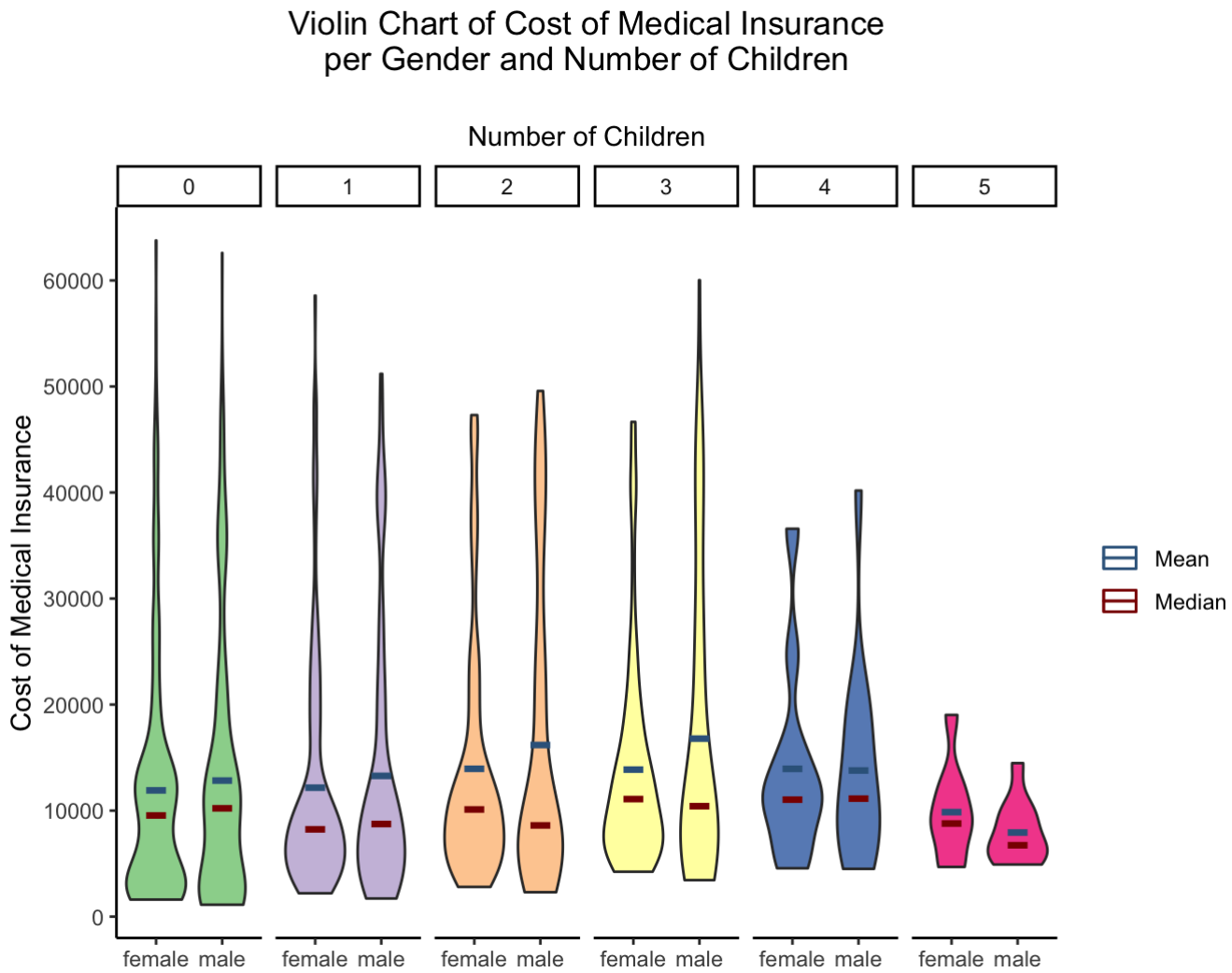
```
# scatter plot cost vs age, highlight by bmi groups
plt_age <- ggplot(df_insurance, aes(age, charges)) +
  geom_point(aes(color = bmi_gr), alpha = 0.8) +
  scale_color_manual(values = bmi_color, name = 'BMI groups') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
    scale_x_continuous(breaks = seq(0, 70, by = 5)) +
  theme_dark() +
  facet_grid(~ smoker) +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance vs Age per BMI ') +
  theme(plot.title = element_text(hjust = 0.5))
# scatter plot cost vs age, split by gender
plt_age_gend <- ggplot(df_insurance, aes(age, charges)) +
  geom_point(aes(color = bmi_gr), alpha = 0.8) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_color_manual(values = bmi_color, name = 'BMI groups') +
  theme_dark() +
  facet_grid(~ sex) +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance vs Age per BMI  and Gender') +
  theme(plot.title = element_text(hjust = 0.5))
# plot two plots together
grid_arrange_shared_legend(plt_age, plt_age_gend, ncol = 2)
```



As far for BMI groups, it is obvious that all the extreme values (greater than 34489.44 USD) are the observations, having obesity. There are a few observations, that have been charged close to the upper outliers cutoff. We have already extracted these observations above. There was only one smoking man, being a healthy weight. However, we can not conclude that all the observations having obesity, have been charged more. As we see here and have seen on the distribution plot, there is a bunch of people having obesity, are represented in all the percentiles.

How does the distribution between the cost of medical insurance by gender and number of children look? We plot a violin chart and look at it.

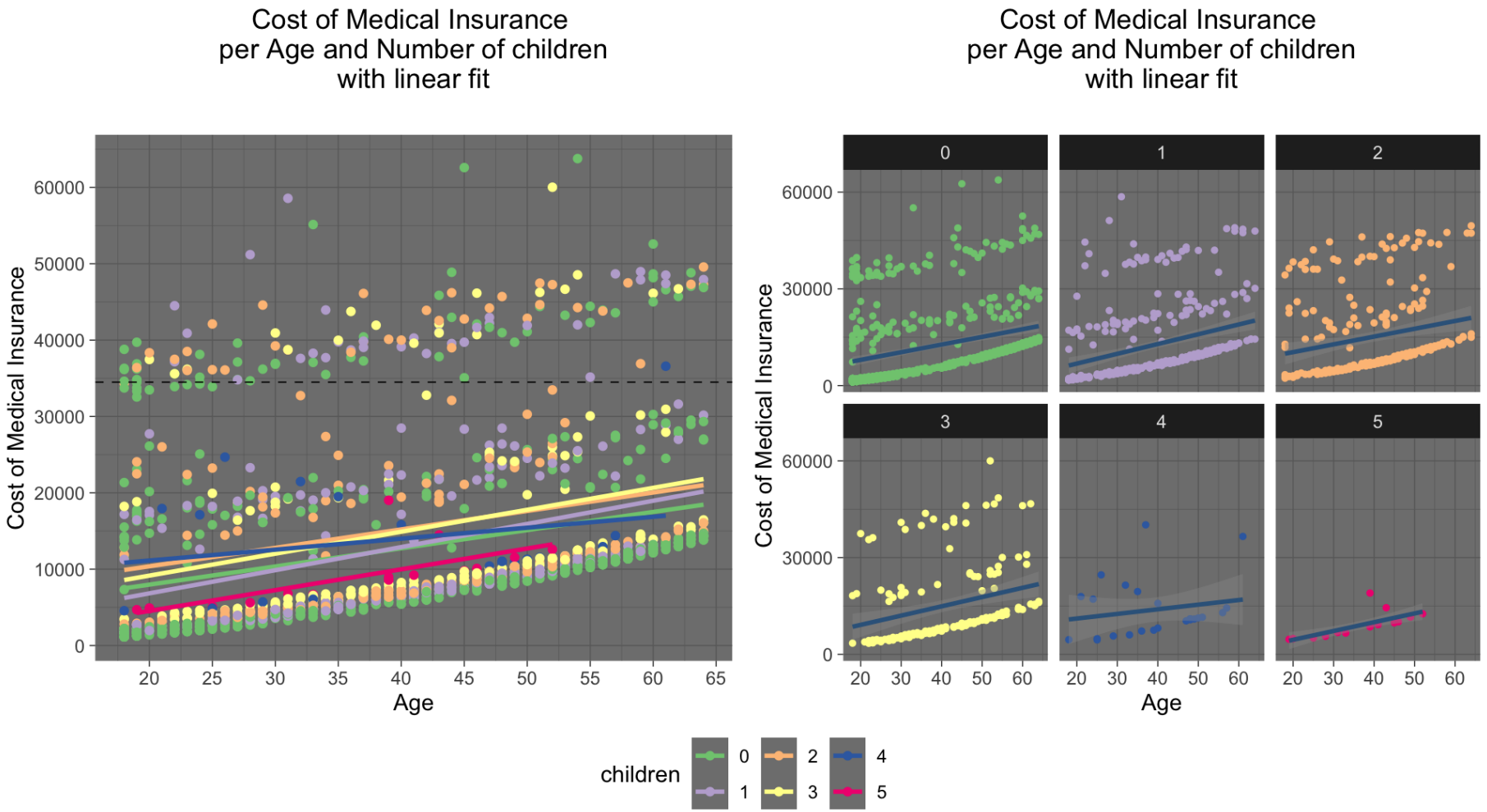
```
ggplot(df_insurance, aes(sex, charges)) +
  geom_violin(aes(fill = children), alpha = 0.8) +
  facet_grid(~ children) +
  stat_summary(fun = median, geom = 'crossbar', width = 0.3, aes(color = 'Median')) +
  stat_summary(fun = mean, geom = 'crossbar', width = 0.3, aes(color = 'Mean')) +
  scale_colour_manual(values = c('steelblue4', 'red4'), name = NULL) +
  scale_fill_brewer(palette = 'Accent') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  labs(x = NULL, y = 'Cost of Medical Insurance',
       title = 'Violin Chart of Cost of Medical Insurance\nper Gender and Number of Children\n',
       subtitle = 'Number of Children') +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5)) +
  guides(fill = "none")
```



There is some difference between men and women having two, three, or five children, but not by a lot.

We plot two scatter plots to look if we can find a relationship between the cost of medical insurance and age, highlighting with color the number of children.

```
plt_age_child <- ggplot(df_insurance, aes(age, charges, color = children), size = 3) +
  geom_point() +
  scale_color_brewer(palette = 'Accent') +
  stat_smooth(method = lm, se = F, formula = y ~ x) +
  geom_hline(yintercept = outl_higher, linetype= 'dashed', size = 0.3) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_x_continuous(breaks = seq(0, 65, by = 5)) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance \nper Age and Number of children\nwith linear fit\n') +
  theme(plot.title = element_text(hjust = 0.5))

plt_age_child_sep <- ggplot(df_insurance, aes(age, charges)) +
  geom_point(aes(color = children), size = 1) +
  scale_color_brewer(palette = 'Accent') +
  stat_smooth(method = lm, se = T, formula = y ~ x, color = 'steelblue4') +
  facet_wrap(children~.) +
  scale_y_continuous(breaks = seq(0, 65000, by = 30000)) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance \nper Age and Number of children\nwith linear fit\n') +
  theme(plot.title = element_text(hjust = 0.5))

grid_arrange_shared_legend(plt_age_child, plt_age_child_sep, ncol = 2)
```



Here we see a linear relationship between the cost of MI and age. Also, it seems that the number of children matters, we see an interaction effect. Although, the proportion of observations having four or five children is very small, only 3 % of the data!

A few observations have four children and medical insurance costing them more than upper outliers cutoff. Let's compute it and look at whether these people are smokers or/and have problems with weight.

```
df_insurance %>%
  filter(charges > outl_higher & children == '4')
```

age	sex	bmi	children	smoker	region	charges	bmi_gr
37	male	34	4	yes	southwest	40182	Obesity
61	female	33	4	no	southeast	36580	Obesity

As a result, we got two persons: a smoking man and a non-smoking woman. They are both obese.

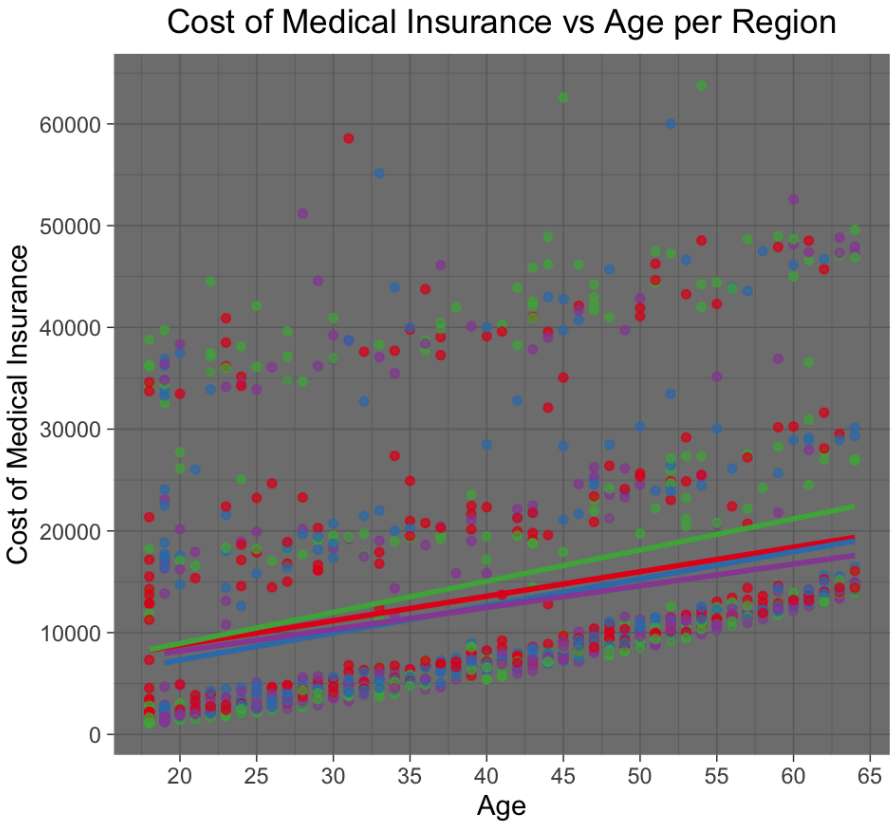
It would be useful to transform the attribute ‘children’ to dummy variable and look at the relationship again. We create a new data frame and recode the variable ‘children’ as ‘0’ if an observation has no children, and as ‘1’ if he or she has any number of children.

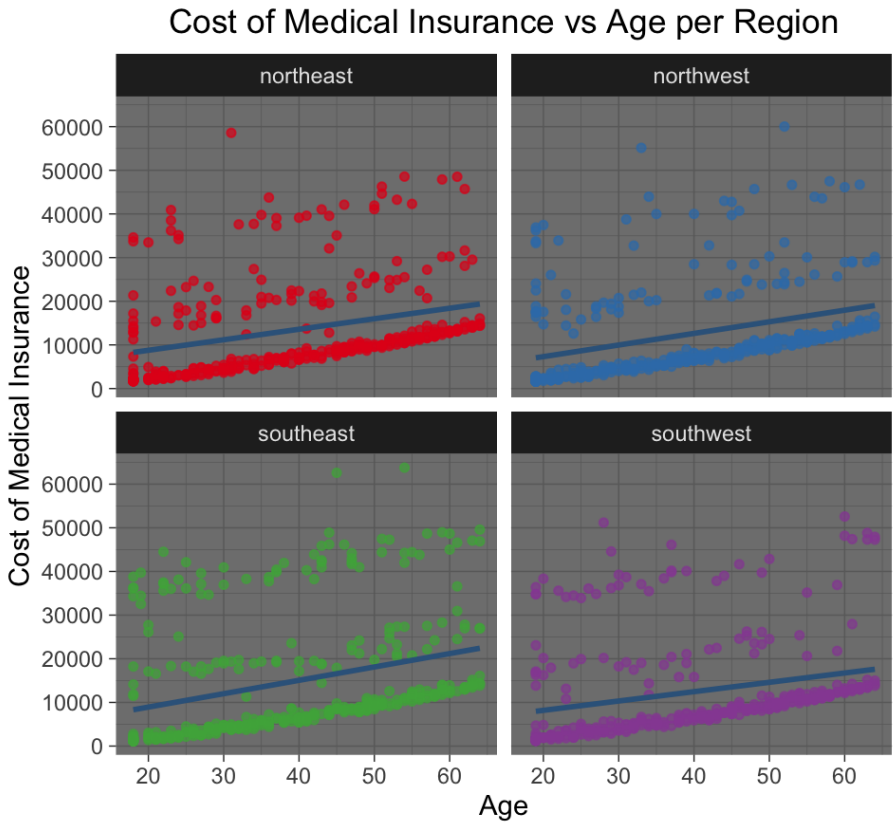
```
df_children <- df_insurance
df_children$children[df_children$children != '0'] <- '1'
ggplot(df_children, aes(age, charges, color = children), size = 3) +
  geom_point() +
  scale_color_brewer(palette = 'Accent', name = 'Children?', labels = c('No', 'Yes')) +
  stat_smooth(method = lm, se = F, formula = y ~ x) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_x_continuous(breaks = seq(0, 65, by = 5)) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance per Age \nand Children as an Indicator Variable\nwith linear fit\n') +
  theme(plot.title = element_text(hjust = 0.5))
```

Well, we can not say that this has changed the picture drastically. We will look if the number of children is a statistically significant feature when we build a model later.

Now we look at a relationship between the cost of MI and age, as well as between the cost of MI and BMI per region.

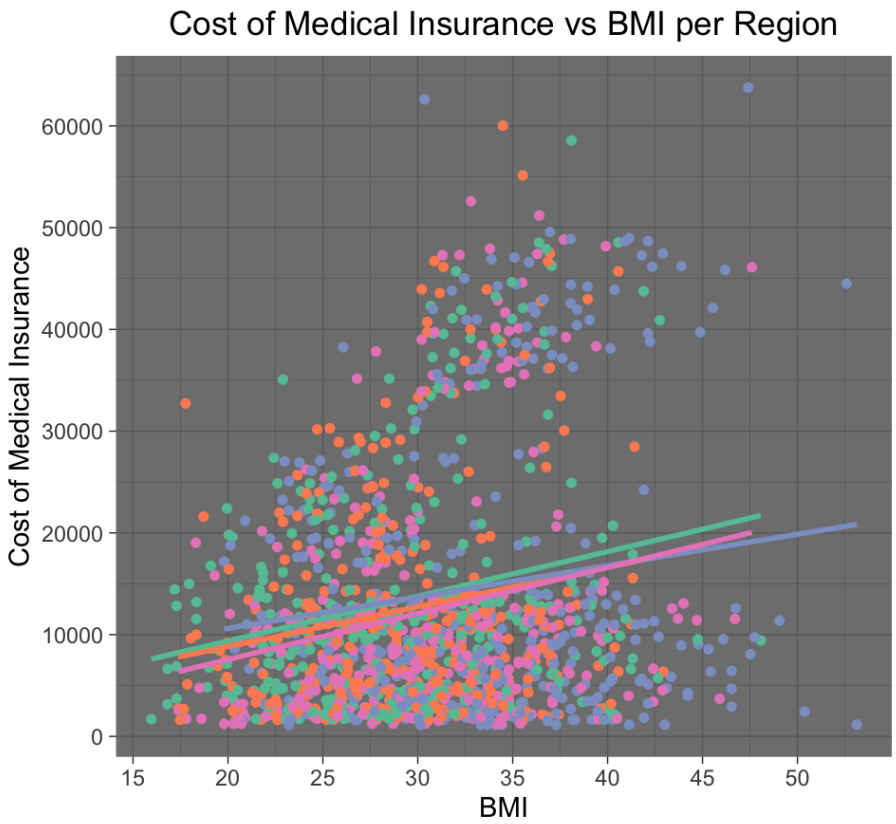
```
plt_age_reg <- ggplot(df_insurance, aes(age, charges, color = region)) +
  geom_point(alpha = 0.7, size = 1.3) +
  scale_color_brewer(palette = 'Set1', name = NULL) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_x_continuous(breaks = seq(0, 70, by = 5)) +
  stat_smooth(method = lm, se = F, formula = y ~ x) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance vs Age per Region') +
  theme(plot.title = element_text(hjust = 0.5))

plt_age_reg_sep <- ggplot(df_insurance, aes(age, charges, color = region)) +
  geom_point(size = 1.3, alpha = 0.7) +
  scale_color_brewer(palette = 'Set1', name = 'Region') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  facet_wrap(region ~.) +
  stat_smooth(method = lm, se = F, formula = y ~ x, color = 'steelblue4') +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'Age',
       title = 'Cost of Medical Insurance vs Age per Region') +
  theme(plot.title = element_text(hjust = 0.5),
        legend.title = element_text(hjust = 0.8))

grid_arrange_shared_legend(plt_age_reg, plt_age_reg_sep, ncol = 2)
```

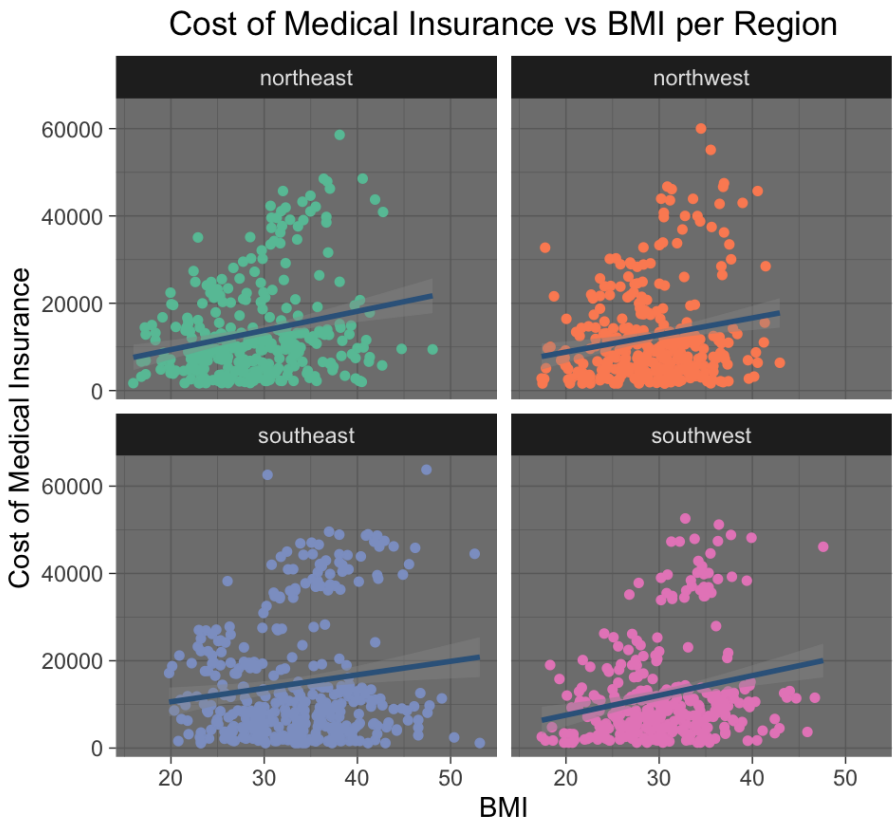

Cost of Medical Insurance vs Age per Region
Cost of Medical Insurance
Age
0
10000
20000
30000
40000
50000
60000
20
25
30
35
40
45
50
55
60
65


Cost of Medical Insurance vs Age per Region
northeast
northwest
southeast
southwest
Cost of Medical Insurance
Age
0
10000
20000
30000
40000
50000
60000
20
30
40
50
60


northeast
northwest
southeast
southwest

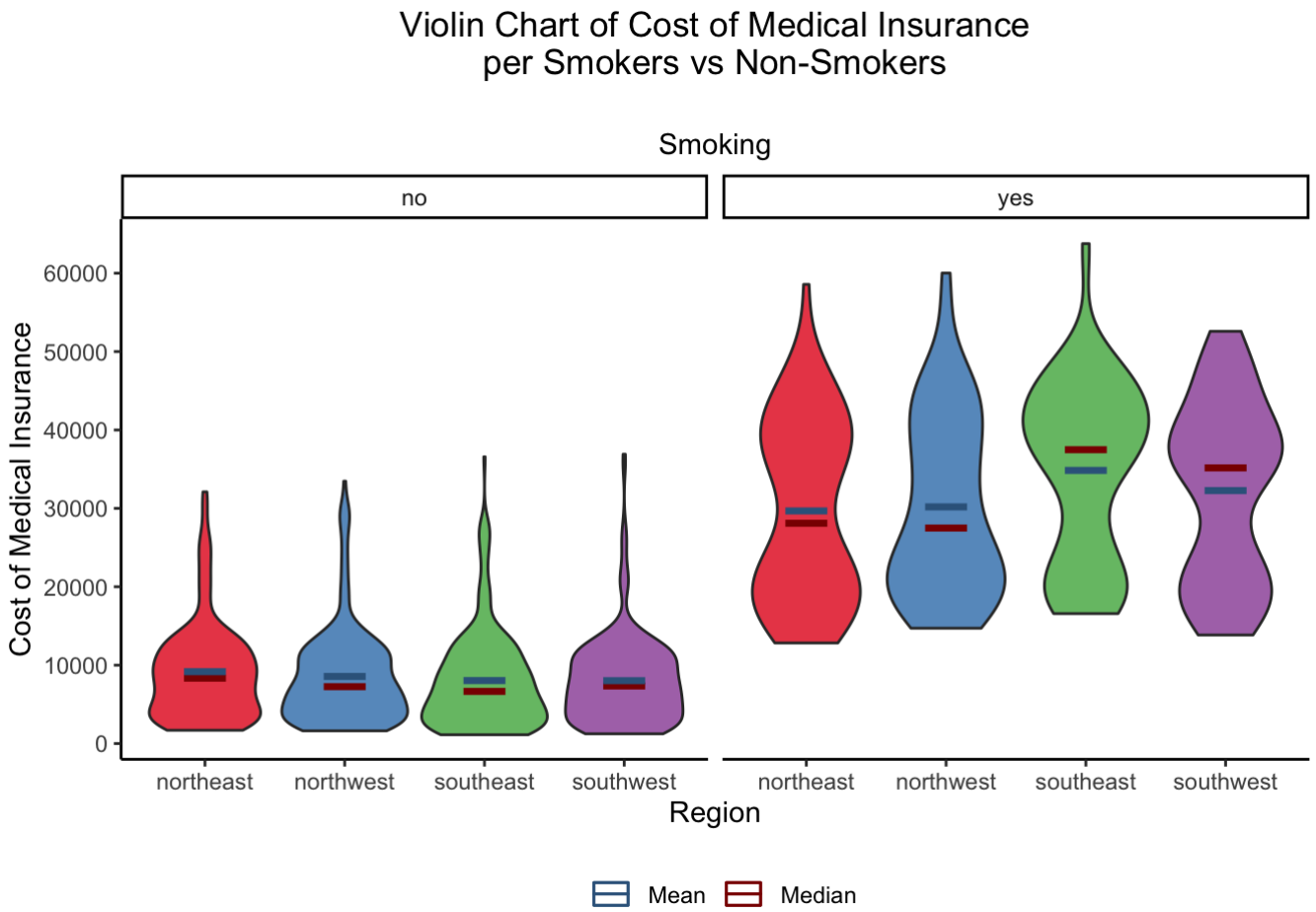
```
plt_bmi_reg <- ggplot(df_insurance, aes(bmi, charges, color = region)) +
  geom_point(size = 1.3) +
  scale_color_brewer(palette = 'Set2', name = NULL) +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  scale_x_continuous(breaks = seq(0, 60, by = 5)) +
  stat_smooth(method = lm, se = F, formula = y ~ x) +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'BMI',
       title = 'Cost of Medical Insurance vs BMI per Region') +
  theme(plot.title = element_text(hjust = 0.5))

plt_bmi_reg_sep <- ggplot(df_insurance, aes(bmi, charges)) +
  geom_point(aes(color = region), size = 1.3) +
  facet_wrap(region ~.) +
  scale_color_brewer(palette = 'Set2') +
  stat_smooth(method = lm, se = T, formula = y ~ x, color = 'steelblue4') +
  theme_dark() +
  labs(y = 'Cost of Medical Insurance', x = 'BMI',
       title = 'Cost of Medical Insurance vs BMI per Region') +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = 'none')

grid_arrange_shared_legend(plt_bmi_reg, plt_bmi_reg_sep, ncol = 2)
```





All the regions look very similar to each other. Let's plot a violin chart of the cost of MI by region and smokers.

```
ggplot(df_insurance, aes(region, charges)) +
  geom_violin(aes(fill = region), alpha = 0.8) +
  stat_summary(fun = median, geom = 'crossbar', width = 0.3, aes(color = 'Median')) +
  stat_summary(fun = mean, geom = 'crossbar', width = 0.3, aes(color = 'Mean')) +
  scale_colour_manual(values = c('steelblue4', 'red4'), name = NULL) +
  scale_fill_brewer(palette = 'Set1', name = 'Region') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  facet_wrap(~smoker) +
  labs(x = 'Region', y = 'Cost of Medical Insurance',
       title = 'Violin Chart of Cost of Medical Insurance\nper Smokers vs Non-Smokers\n',
       subtitle = 'Smoking') +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5), legend.position = 'bottom') +
  guides(colour = guide_legend(order = 2),
            fill = 'none')
```

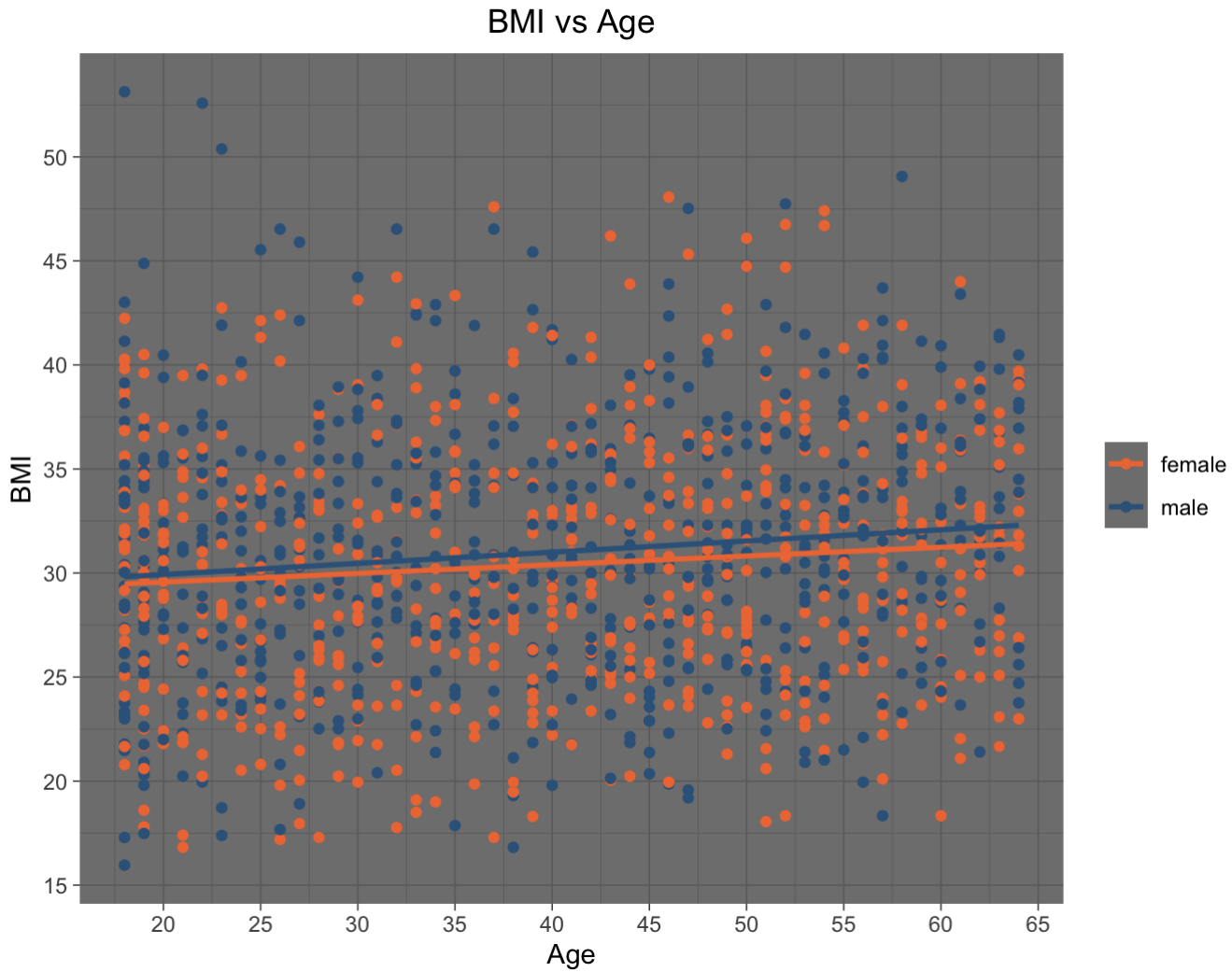


There is no difference between the regions among the non-smokers. We see a slight difference between the regions among the smokers. The southwest and southeast regions are similar and have the higher mean and median than the northwest and northeast regions. There are more smokers in the south.

Scatter Plot BMI vs Age

Is there any relationship between BMI and age? Let’s look at a scatter plot.

```
ggplot(df_insurance, aes(age, bmi, color = sex)) +
  geom_point(size = 1.5) +
  scale_color_manual(values = sex_color, name = NULL) +
  scale_shape_discrete(name = 'BMI groups') +
  scale_y_continuous(breaks = seq(10, 60, by = 5)) +
  scale_x_continuous(breaks = seq(0, 70, by = 5)) +
  stat_smooth(method = lm, se = F, formula = y ~ x) +
  theme_dark() +
  #facet_grid(~ sex) +
  labs(x = 'Age', y = 'BMI', title = 'BMI vs Age') +
  theme(plot.title = element_text(hjust = 0.5))
```

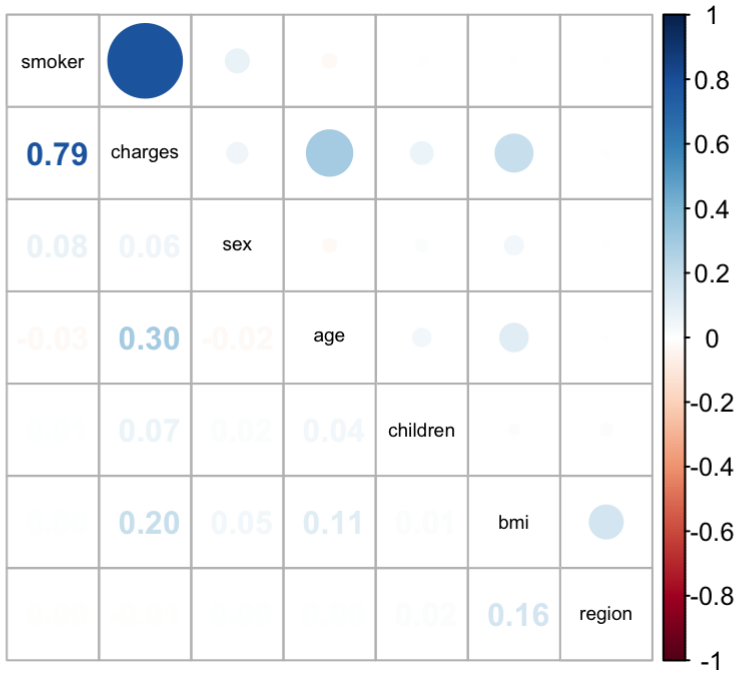


There is no relationship between BMI and age, both for men and women.

Correlation Matrix

While we've already looked at all the relationships between the target and attributes, as well as between different attributes, it's always good to compute correlation coefficients. We plot a mixed matrix where we can see both the correlation coefficients and their visualizations in the form of circles.

```
# copy the data and convert all factor variables to integer
df_insurance_cor <- df_insurance[,c(-8)] %>%
  mutate_if(is.factor, as.integer)
corrplot.mixed(cor(df_insurance_cor), order = 'AOE', tl.col = 'black', tl.cex = 0.6)
```

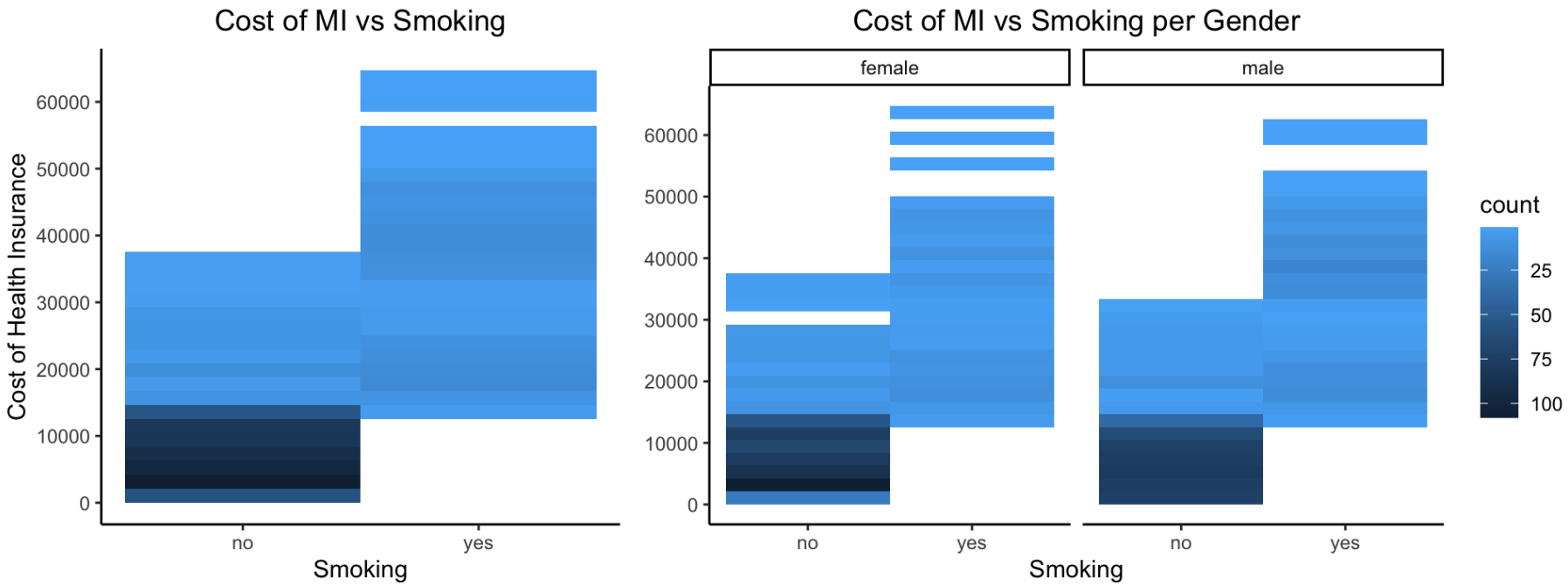


As expected, there is a strong correlation between the cost of medical insurance and smoking, 0.79. The correlation between the cost and BMI, as well as between the cost and age, is weak, being 0.2 and 0.3 respectively.

Heatmap of 2d Bin Counts

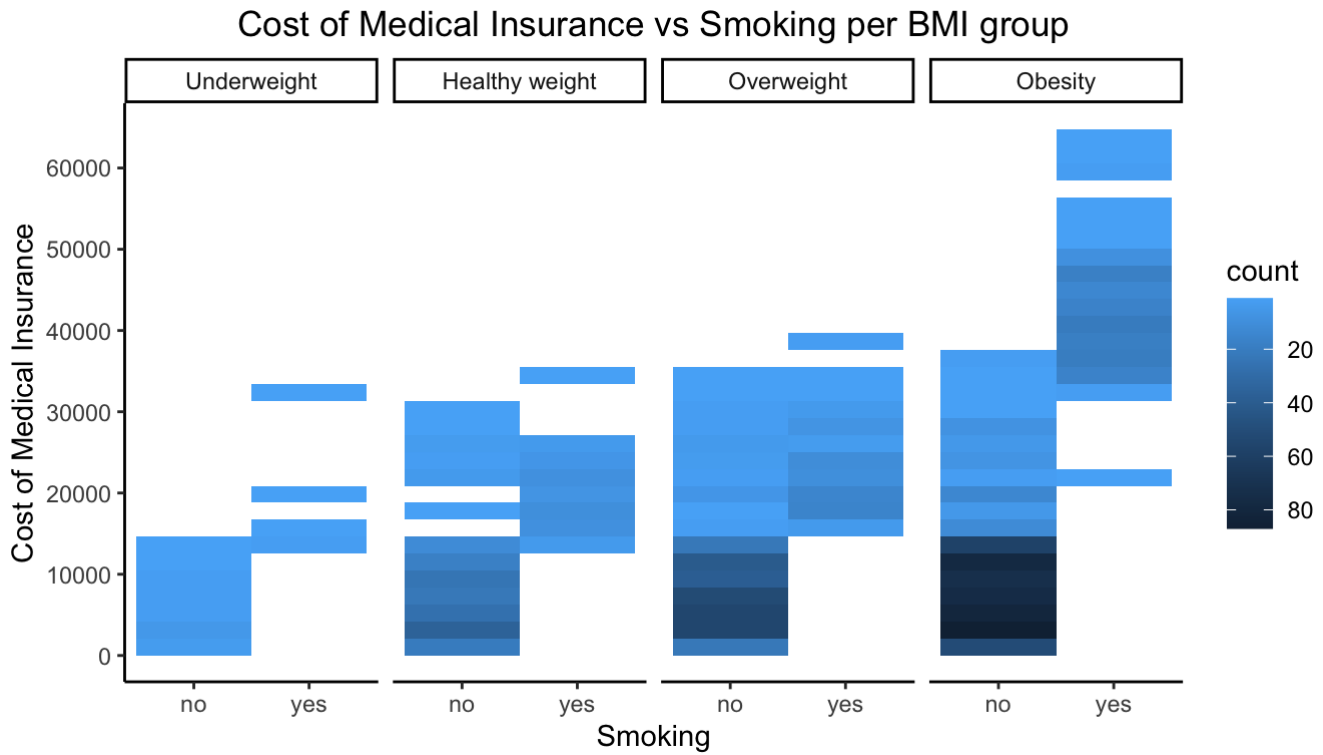
Lastly, we use a heatmap of 2d bin counts plot to have a look at the target and some features again. This kind of plot is very useful in presence of overplotting. We start with plotting the variable ‘charges’ by smoking, as well as by smoking and gender. The number of observations satisfying one cluster counts in each rectangle, and then maps the number of cases to the rectangle’s fill. We use the number of bins set by default.

```
sm_pl <- ggplot(df_insurance, aes(smoker, charges)) +
  geom_bin_2d() +
  scale_fill_gradient(trans = 'reverse') + # reverse the color palette
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  labs(x = 'Smoking', y = 'Cost of Health Insurance',
       title = 'Cost of MI vs Smoking') +
  theme(legend.position = 'none', plot.title = element_text(hjust = 0.5))
sm_sex_pl <- ggplot(df_insurance, aes(smoker, charges)) +
  geom_bin_2d() +
  facet_grid(~sex) +
  scale_fill_gradient(trans = 'reverse') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  labs(x = 'Smoking', y = NULL,
       title = 'Cost of MI vs Smoking per Gender') +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(sm_pl, sm_sex_pl, ncol = 2, heights = c(3,1), widths = c(2,3)) # adjust the figures' proportions on the plan
       e
```



Now we add BMI groups and remove gender.

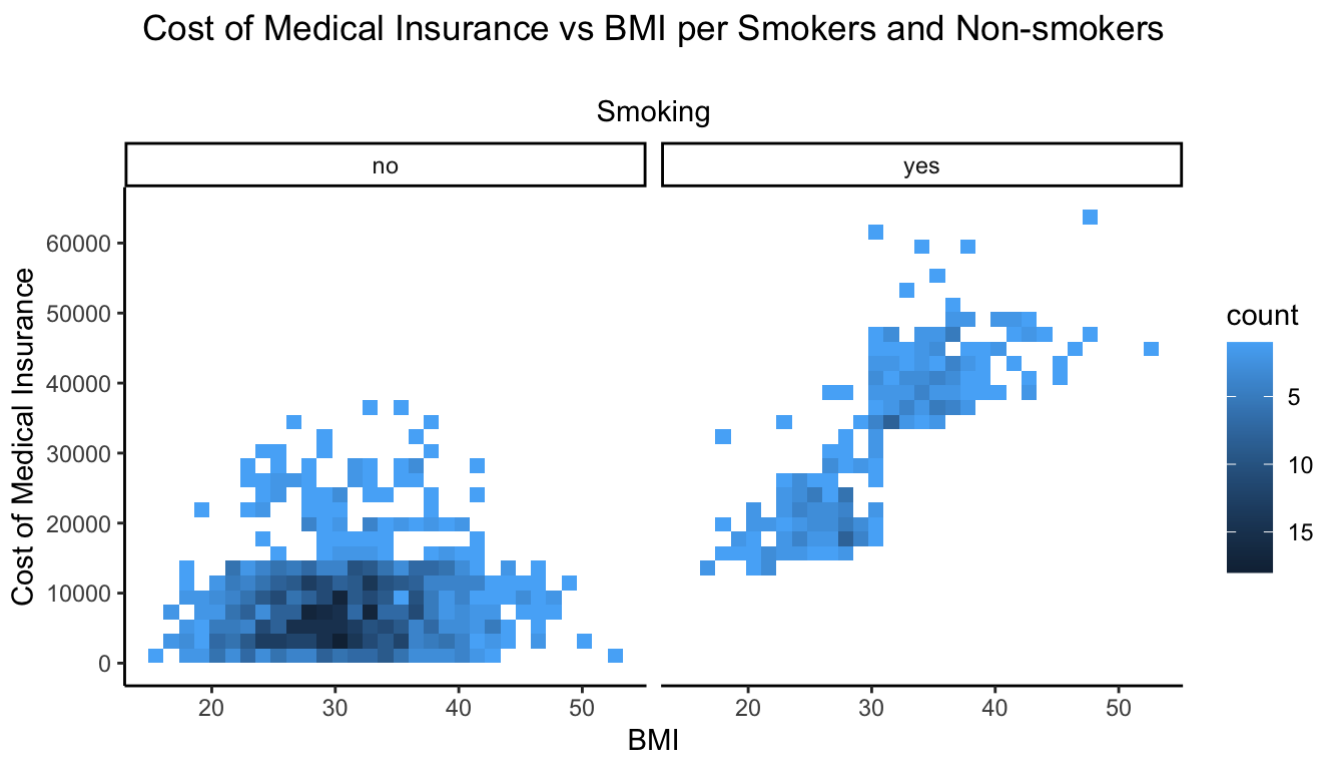
```
ggplot(df_insurance, aes(smoker, charges)) +
  geom_bin_2d() +
  facet_grid(~bmi_gr) +
  scale_fill_gradient(trans = 'reverse') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  labs(x = 'Smoking', y = 'Cost of Medical Insurance',
       title = 'Cost of Medical Insurance vs Smoking per BMI group') +
  theme(plot.title = element_text(hjust = 0.5))
```



Smoking and obesity are not only harmful to health, but also to the wallet!

On this plot we count the number of the observations by smoking with BMI as the continuous variable.

```
ggplot(df_insurance, aes(bmi, charges)) +
  geom_bin_2d() +
  facet_grid(~smoker) +
  scale_fill_gradient(trans = 'reverse') +
  scale_y_continuous(breaks = seq(0, 65000, by = 10000)) +
  theme_classic() +
  labs(x = 'BMI', y = 'Cost of Medical Insurance',
       title = 'Cost of Medical Insurance vs BMI per Smokers and Non-smokers\n',
       subtitle = 'Smoking') +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



What could be done more?

It would be useful to define groups by age. I looked at different segments used in statistical research. But I thought that they were not suitable for this case. This sample represents the age from 18 to 64. If we divide into groups of 10 years, we will get too many segments. It could have been divided into groups of 18-23 as juniors, 23-60 as middle, and 60-64 as seniors, but then we would get very uneven segments, which would distort the statistics. In any case, this is something that can be explored further.

What is next?

1. Feature importance algorithm to see which features have the impact on the target.
2. To build different linear regression models: multiple linear regression, robust linear regression, and polynomial linear regression, and evaluate their performance.