Домашнее задание по курсу ‘Визуализация биомедицинских данных’

Юлия Матвиенко

2023-11-01

Загрузка необходимых библиотек

```
library(tidyverse)
library(ggplot2)
library(ggpubr)
```

Задание 1

```
insurance <- read.csv('insurance_cost.csv',
                      stringsAsFactors = T)
head(insurance)
```

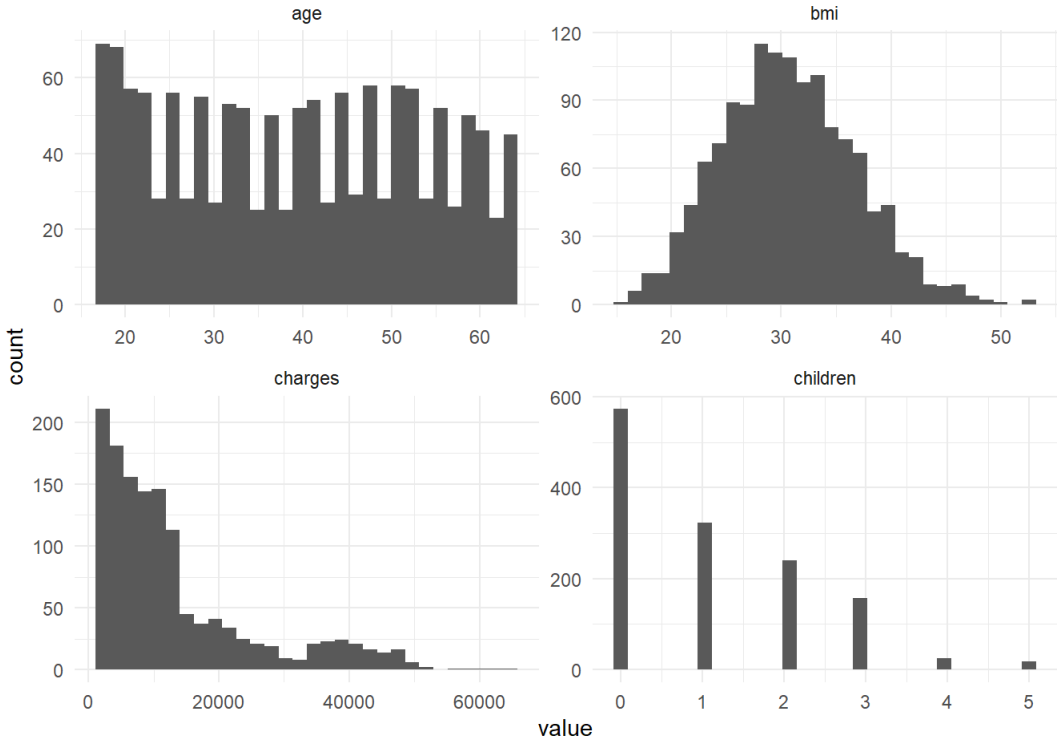
```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

```
summary(insurance)
```

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

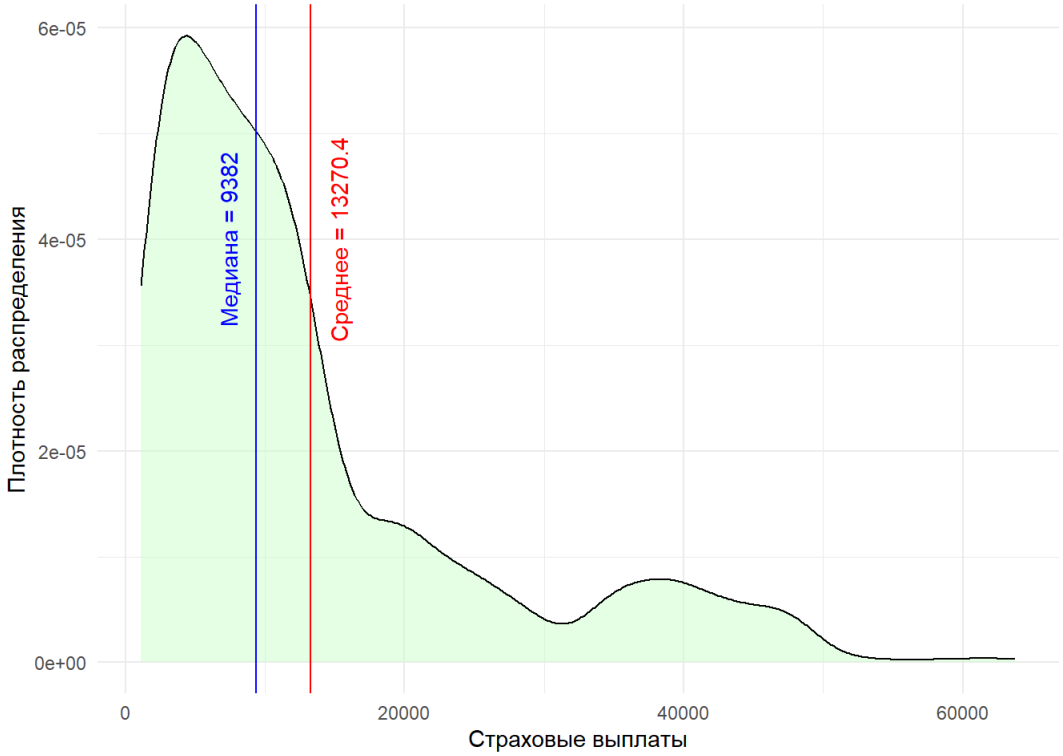
Задание 2

```
#Поскольку в этом пункте не было требований по визуальному оформлению, вывела "графики для себя" - не очень красивые, зато сразу для всех количественных переменных

insurance %>%
  select(where(is.numeric)) %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x = value)) +
  geom_histogram() +
  theme_minimal() +
  facet_wrap(~ name, scales = "free")
```



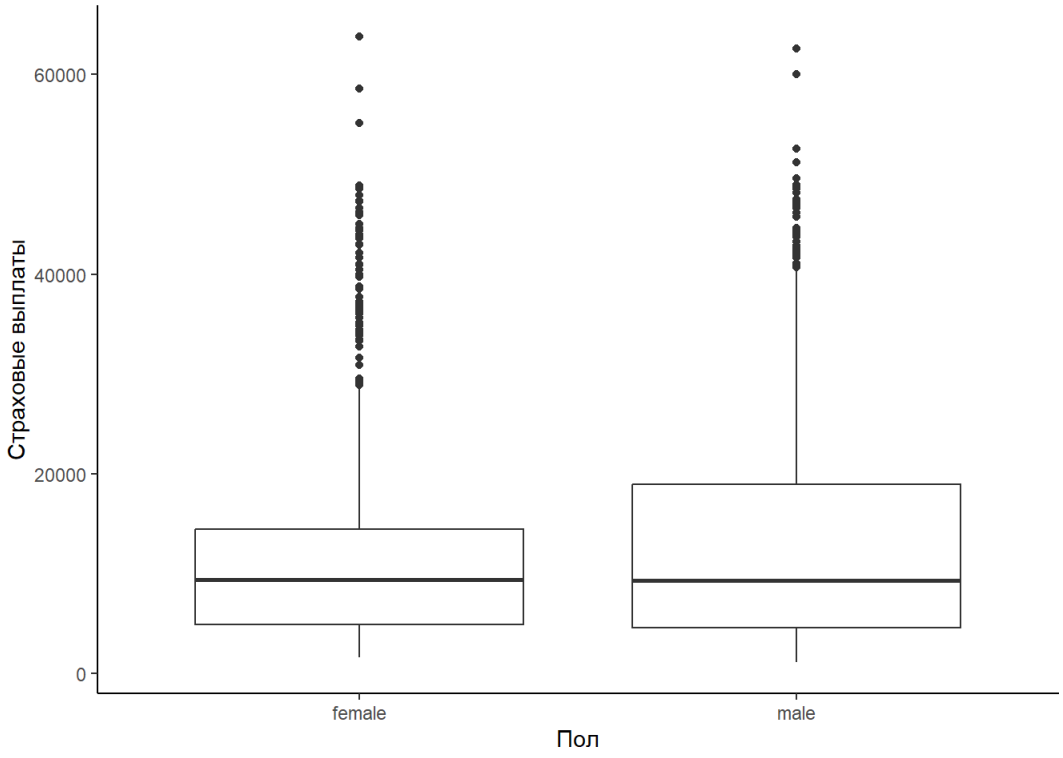
Задание 3

```
charges_mean <- round(mean(insurance$charges),1)
charges_median <- round(median(insurance$charges),1)

ch_density <- ggplot(data = insurance,
              aes(x = charges)) +
  geom_density(alpha = 0.5,
               fill = "#CCFFCC") +
  geom_vline(aes(xintercept = charges_mean),
             colour = "red") +
  annotate("text",
           x= charges_mean+2000,
           y=0.00004,
           label=paste0("Среднее = ", charges_mean),
           color = "red",
           angle = 90) +
  geom_vline(aes(xintercept = charges_median),
             colour = "blue") +
  annotate("text",
           x= charges_median-2000,
           y=0.00004,
           label=paste0("Медиана = ", charges_median),
           color = "blue",
           angle = 90) +
  theme_minimal() +
  labs(x = 'Страховые выплаты', y = 'Плотность распределения')

ch_density
```

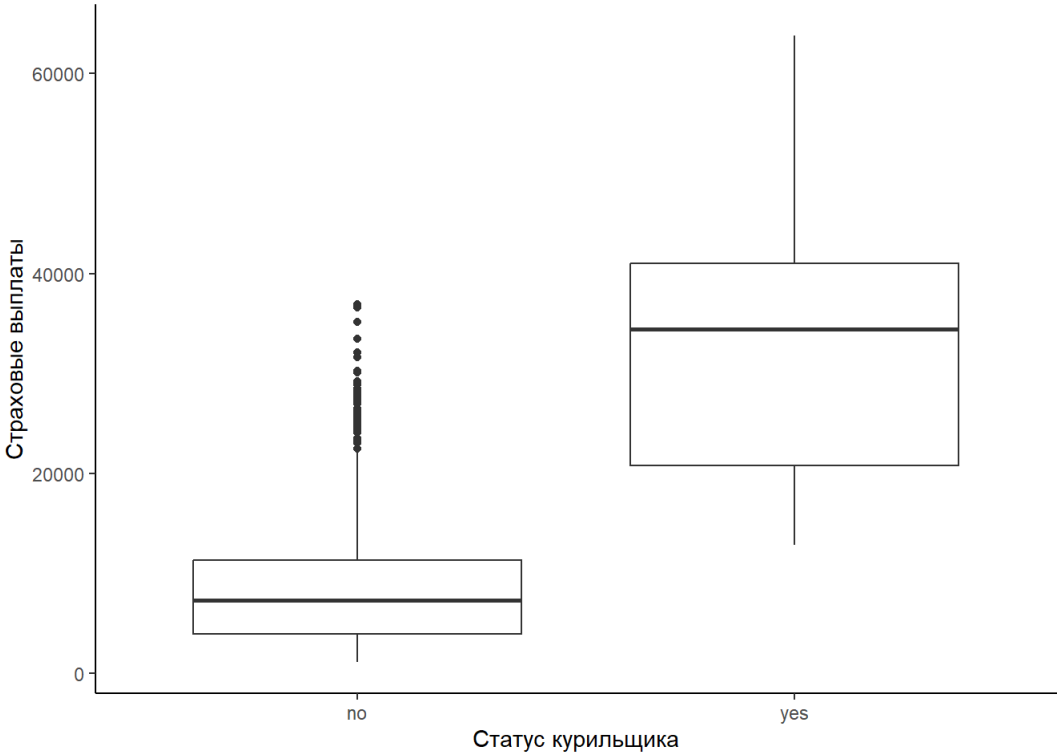


Задание 4

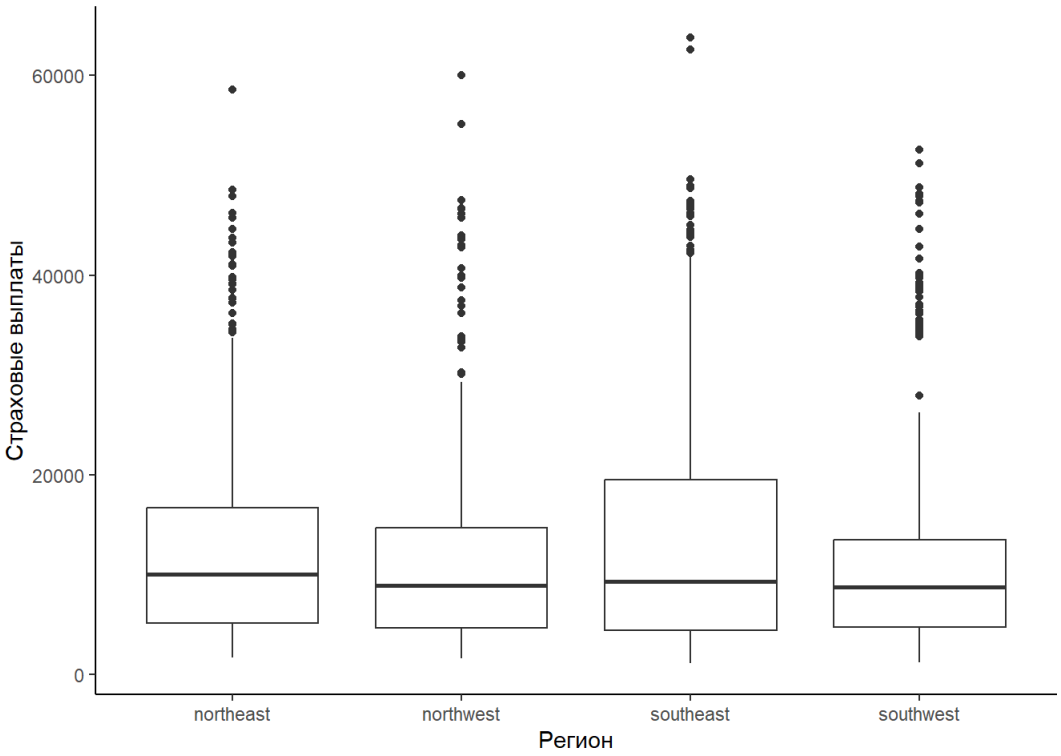
```
charges_by_sex <- ggplot() +
  geom_boxplot(data = insurance,
               aes(x = sex, y = charges)) +
  theme_classic() +
  labs(x = 'Пол', y = 'Страховые выплаты')

charges_by_sex
```




```
charges_by_sm <- ggplot() +
  geom_boxplot(data = insurance,
               aes(x = smoker, y = charges)) +
  theme_classic() +
  labs(x = 'Статус курильщика', y = 'Страховые выплаты')

charges_by_sm
```

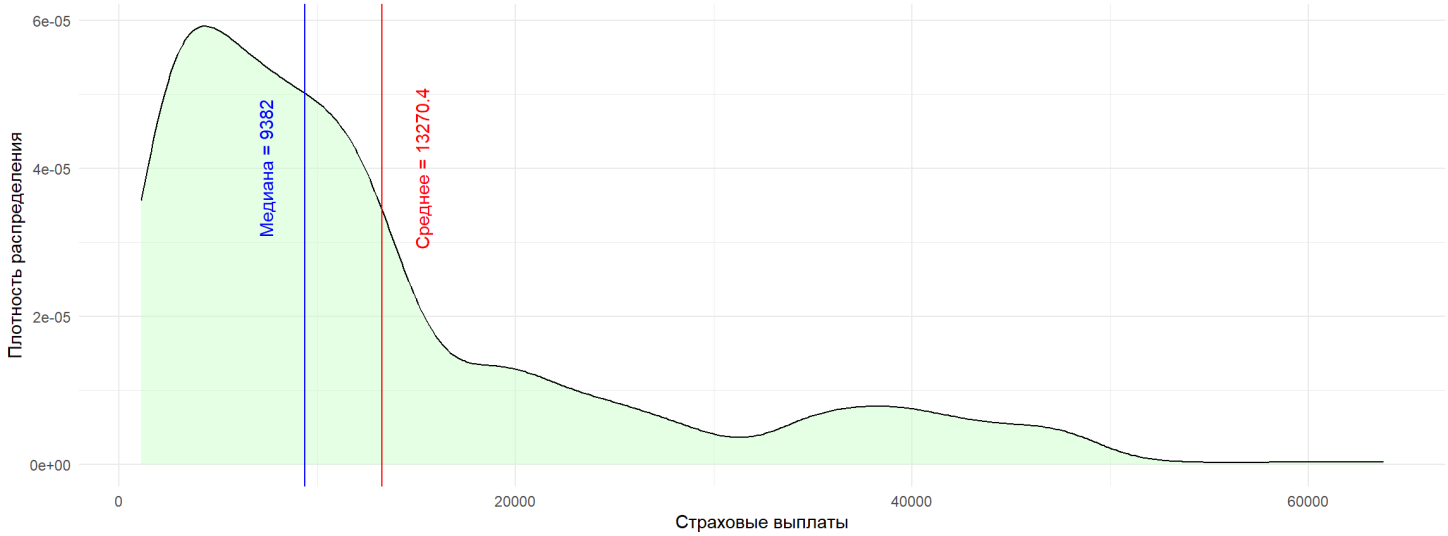


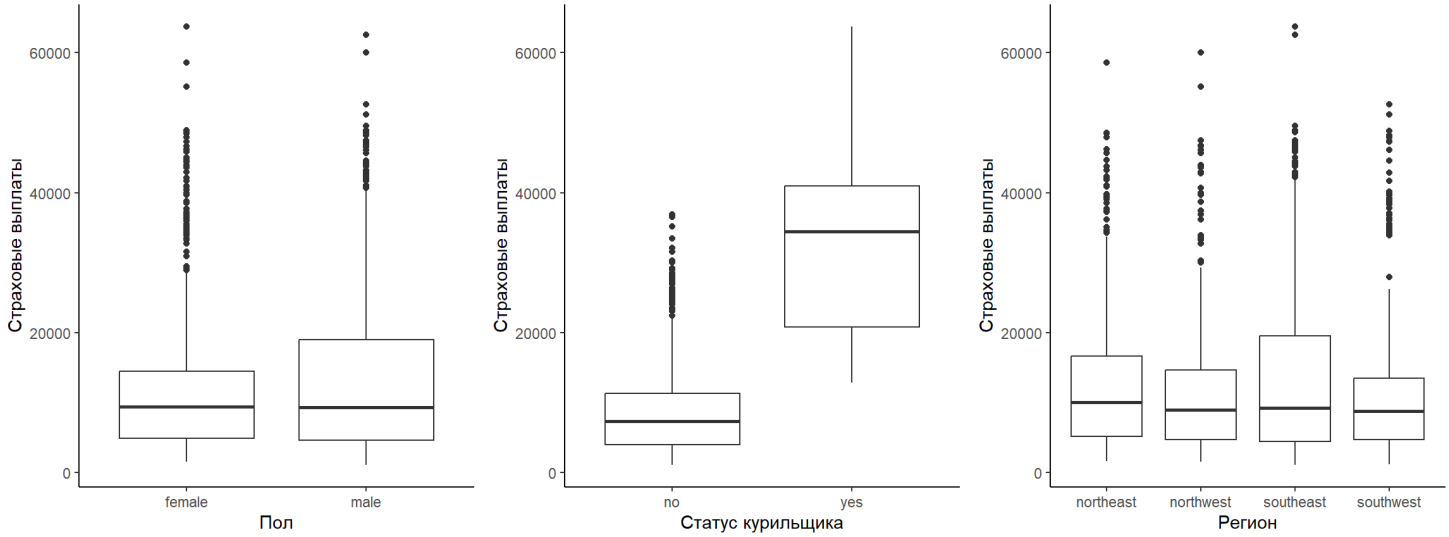

```
charges_by_reg <- ggplot() +
  geom_boxplot(data = insurance,
               aes(x = region, y = charges)) +
  theme_classic() +
  labs(x = 'Регион', y = 'Страховые выплаты')

charges_by_reg
```



Задание 5

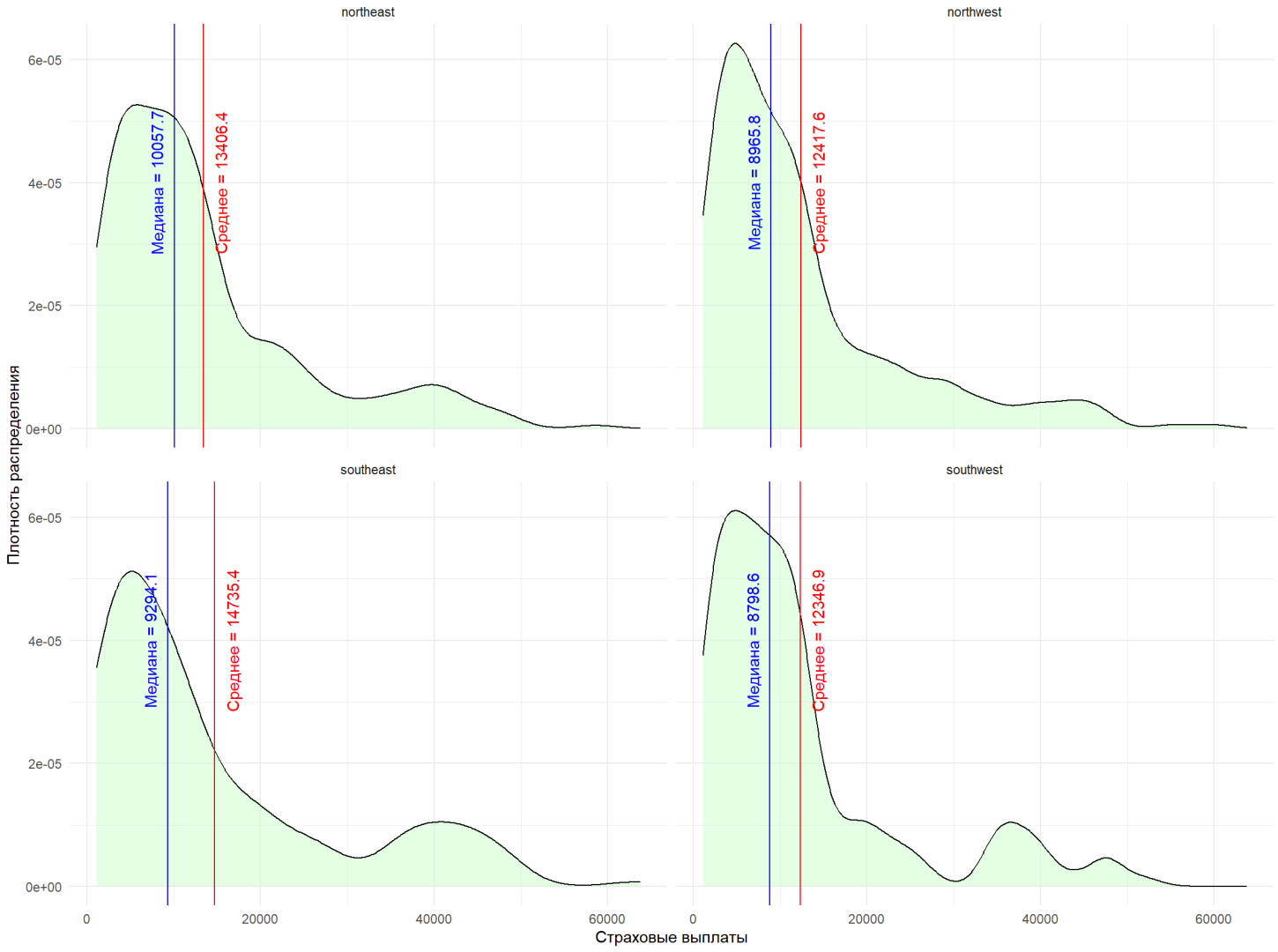
```
ggarrange(ch_density,
          ggarrange(charges_by_sex, charges_by_sm, charges_by_reg, ncol = 3),
          nrow = 2) +
  ggtitle("Характеристики суммы страховых выплаты")
```


Медиана = 9382
Среднее = 13270.4
Плотность распределения
6e-05
4e-05
2e-05
0e+00
0
20000
40000
60000
Страховые выплаты


Страховые выплаты
60000
40000
20000
0
female
male
Пол
no
yes
Статус курильщика
northeast
northwest
southeast
southwest
Регион

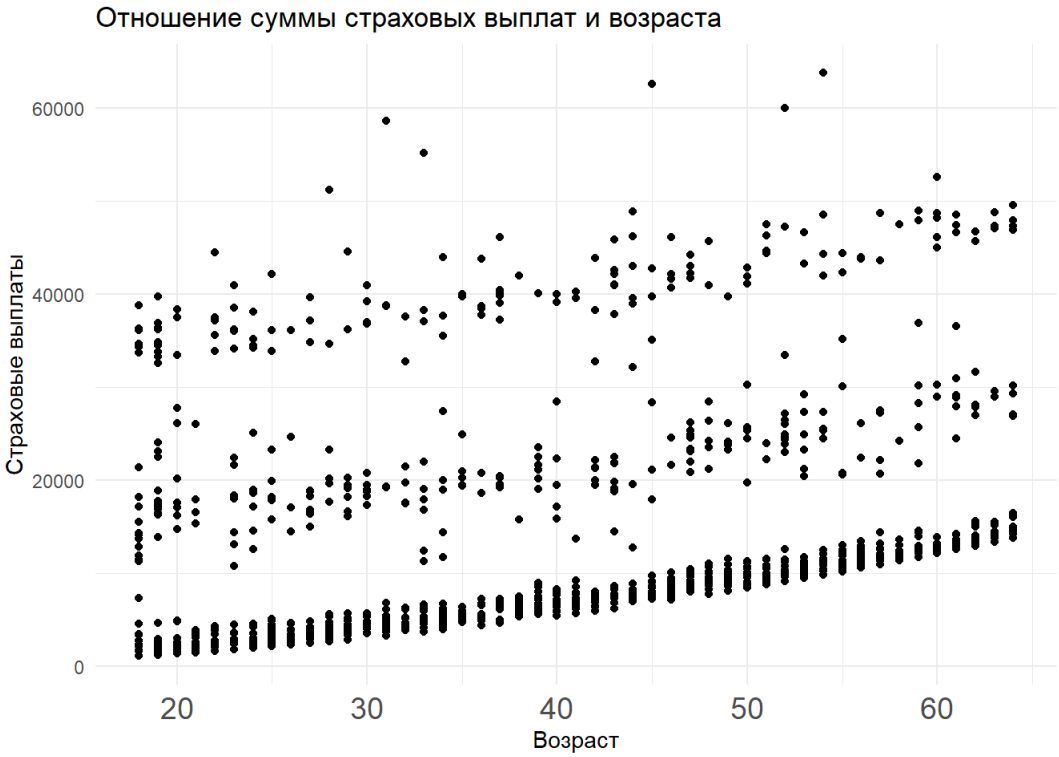
Задание 6

```r
xmean <- insurance %>%
   group_by(region) %>%
   summarise(value = round(mean(charges),1),
             stat = "Mean")

xmedian <- insurance %>%
   group_by(region) %>%
   summarise(value = round(median(charges),1),
             stat = "Median")

xint <- rbind(xmean, xmedian)

ggplot() +
 geom_density(data = insurance,
              aes(x = charges),
              alpha = 0.5,
              fill = "#CCFFCC") +
 geom_vline(data = xint,
            aes(xintercept = value, color = stat)) +
 geom_text(data = xmean,
           aes(x = value+2000,
               y = 0.00004,
               label=paste0("Среднее = ", value),
               color = 'mean',
               angle = 90)) +
  geom_text(data = xmedian,
           aes(x = value-2000,
               y = 0.00004,
               label=paste0("Медиана = ", value),
               color = 'median',
               angle = 90)) +
 scale_color_manual(values = c('red', 'red', 'blue', 'blue')) +
 theme_minimal() +
 theme(legend.position = 'none') +
 labs(x = 'Страховые выплаты', y = 'Плотность распределения') +
 facet_wrap(. ~ region, nrow = 2)
```



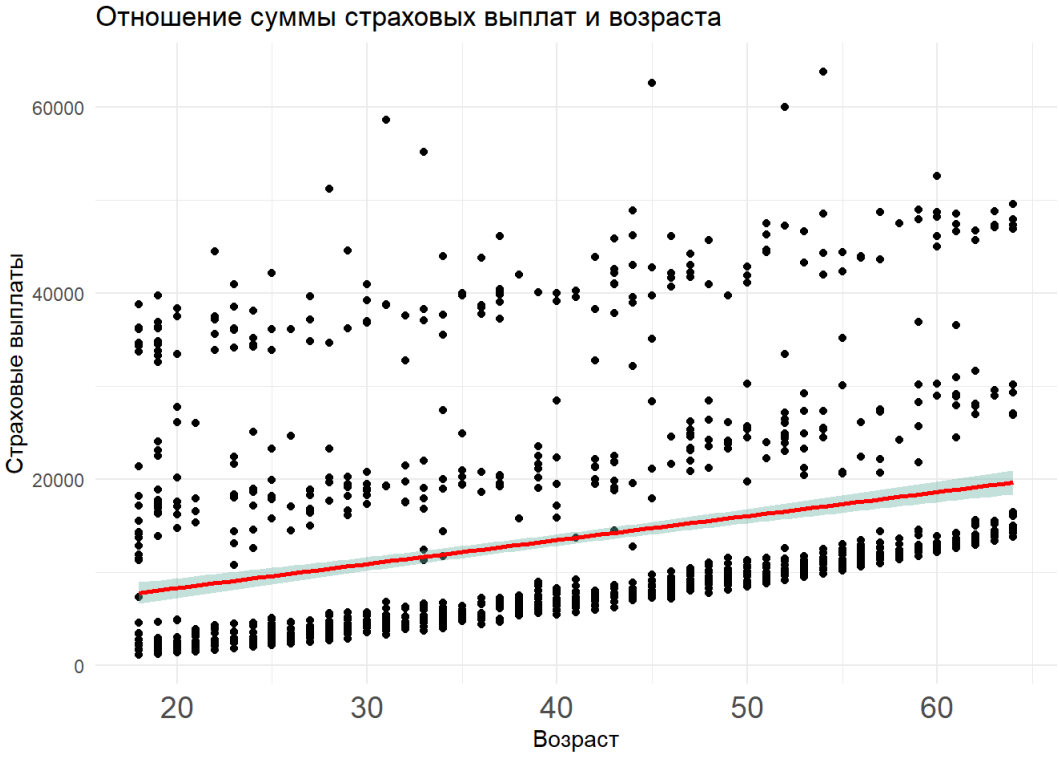
Задание 7

```
scatter <- ggplot(data = insurance,
       aes(x = age, y = charges)) +
  geom_point() +
  theme_minimal() +
  labs(x = "Возраст", y = "Страховые выплаты", title = "Отношение суммы страховых выплат и возраста") +
  theme(axis.text.x = element_text(size=14))

scatter
```



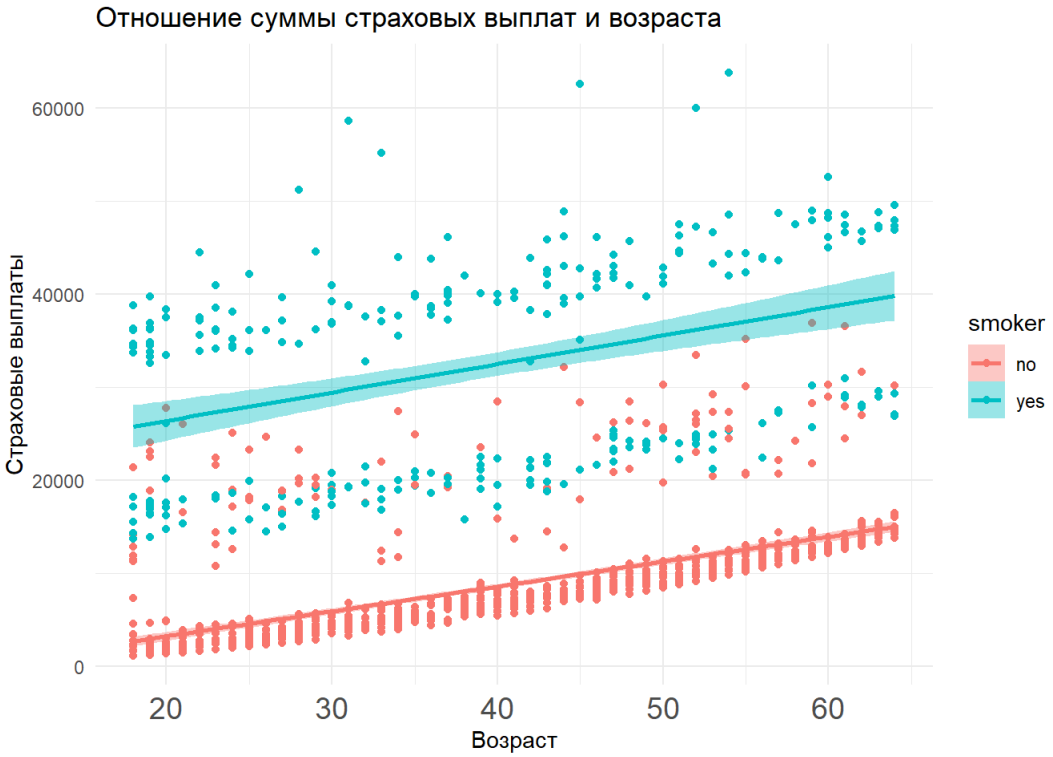
Задание 8

```
scatter +
  geom_smooth(method=lm,
              color="red", fullrange = T,
              fill="#69b3a2",
              se=TRUE)
```



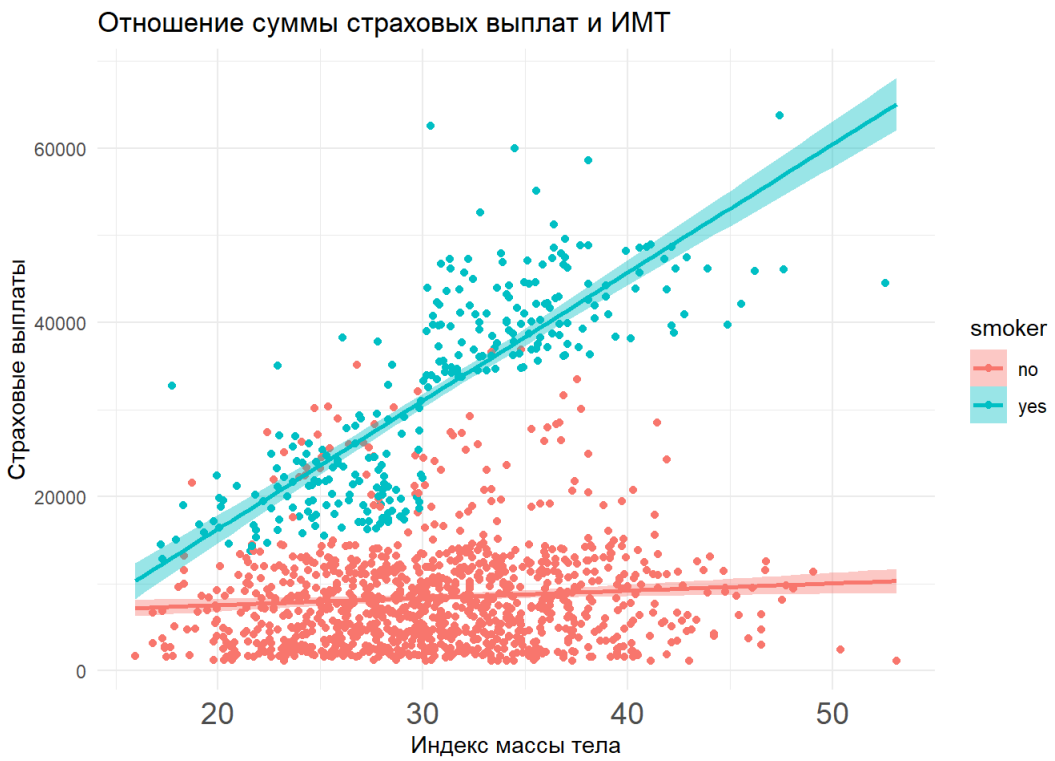
Задание 9

```
ggplot(data = insurance,
       aes(x = age, y = charges, color = smoker, fill = smoker, group = smoker)) +
  geom_point() +
  geom_smooth(method=lm,
              fullrange = T,
              se=TRUE) +
  theme_minimal() +
  labs(x = "Возраст", y = "Страховые выплаты", title = "Отношение суммы страховых выплат и возраста") +
  theme(axis.text.x = element_text(size=14))
```



Задание 10

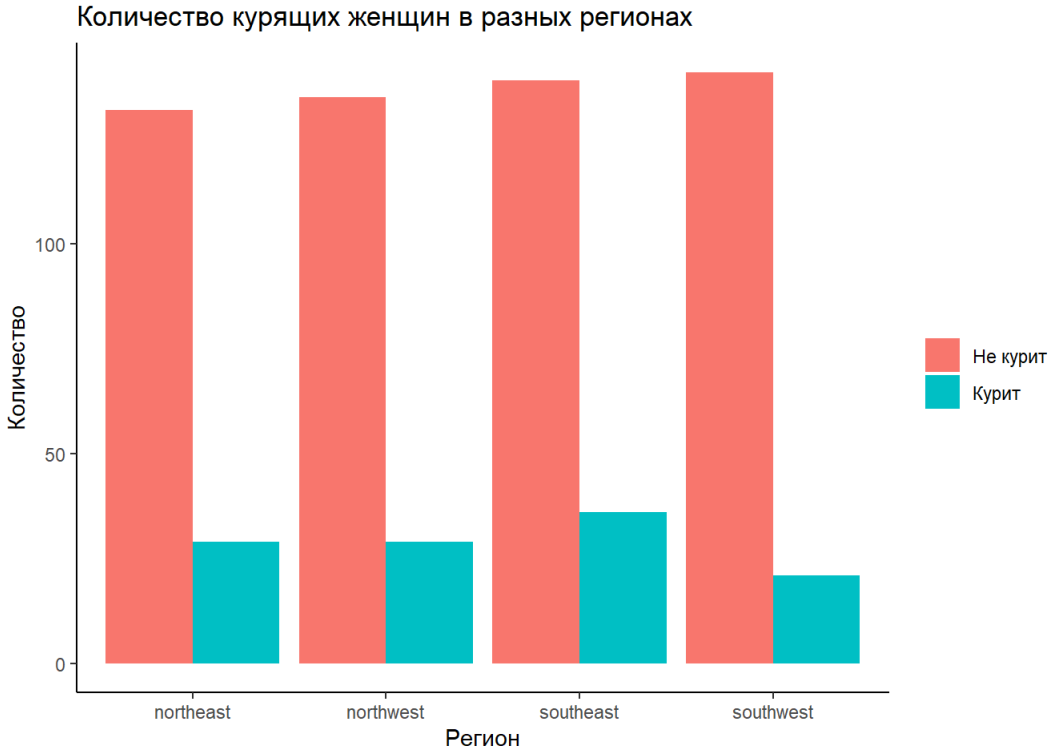
```
ggplot(data = insurance,
       aes(x = bmi, y = charges, color = smoker, fill = smoker, group = smoker)) +
  geom_point() +
  geom_smooth(method = lm,
              fullrange = T,
              se = TRUE) +
  theme_minimal() +
  labs(x = "Индекс массы тела", y = "Страховые выплаты", title = "Отношение суммы страховых выплат и ИМТ") +
  theme(axis.text.x = element_text(size=14))
```



Задание 11

Вопрос 1. Как распределено количество курящих/некурящих женщин в разных регионах? Мы хотим посмотреть распределение номинативной переменной, с разбивкой по другой номинативной переменной. Целесообразно использовать geom_bar, а разбивку по второй переменной выполнить с использованием разных цветов.

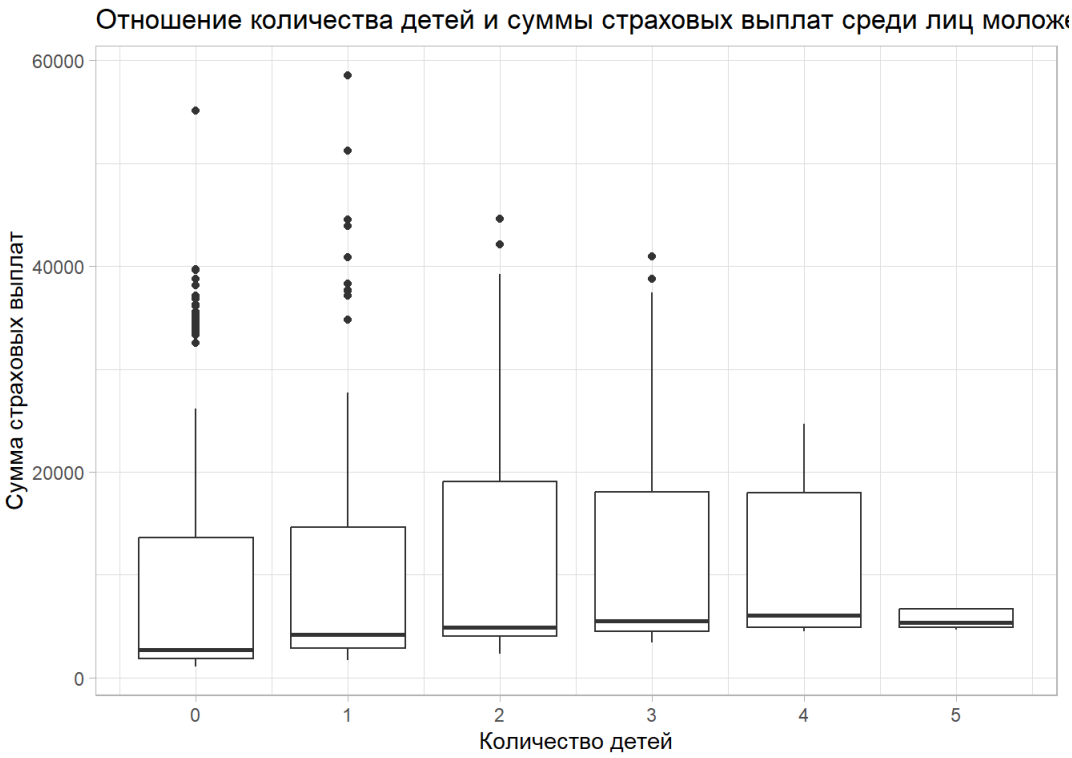
```
insurance %>%
  filter(sex == "female") %>%
  ggplot() +
  geom_bar(aes(x = region, fill = smoker),
           position = "dodge") +
  theme_classic() +
  labs(x = "Регион", y = "Количество", title = "Количество курящих женщин в разных регионах") +
  scale_fill_discrete(labels = c('Не курит', 'Курит')) +
  theme(legend.title = element_blank())
```



Задание 12

Вопрос 2. Связано ли количество детей с суммой страховых выплат за год среди лиц моложе 35 лет? В этом контексте целесообразно провести группировку по количеству детей и построить график боксплот.

```
insurance %>%
  filter(age < 35) %>%
  ggplot() +
  geom_boxplot(aes(x = children, y = charges, group = children)) +
  scale_x_continuous(breaks = seq(0, 5, by = 1)) +
  theme_light() +
  labs(x = "Количество детей", y = "Сумма страховых выплат", title = "Отношение количества детей и суммы страховых выплат среди лиц моложе 35 лет")
```



Задание 13

Вопрос 3. Как соотносится ИМТ и сумма страховых выплат среди мужчин? Поскольку нам необходимо оценить отношение между двумя количественными переменными, целесообразно использовать scatter-plot.

```
insurance %>%
  filter(sex == "male") %>%
  ggplot() +
  geom_point(aes(x = bmi, y = charges)) +
  theme_classic2() +
  labs(x = "Индекс массы тела", y = "Страховые выплаты", title = "Отношение ИМТ и суммы страховых выплат среди мужчин")
```

Задание 14

```
insurance %>%
  mutate(age_group = case_when(
    age < 35 ~ "age: 21-34",
    age >= 35 & age < 50 ~ "age: 35-49",
    age >= 50 ~ "age: 50+"
  )) %>%
  ggplot() +
  geom_point(aes(x = bmi, y = log(charges)),
             color = "#663399",
             alpha = 0.5) +
  geom_smooth(aes(x = bmi, y = log(charges), color = age_group),
              method = lm,
              fullrange = T,
              se = TRUE,
              alpha = 0.2) +
  facet_grid(. ~ age_group) +
  theme_minimal() +
  theme(legend.position = "bottom") +
  ggtitle("Отношение индекса массы тела к логарифму трат по возрастным группам")
```