HW1

Mariia Firuleva

22 10 2022

Load libraries

```
library(data.table)
library(ggplot2)
library(RColorBrewer)
library(dplyr)
library(tidyr)
library(ggpubr)
```

1

```
df <- fread('insurance_cost.csv') %>%
  mutate_if(is.character, as.factor)
knitr::kable(df %>% head())
```

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

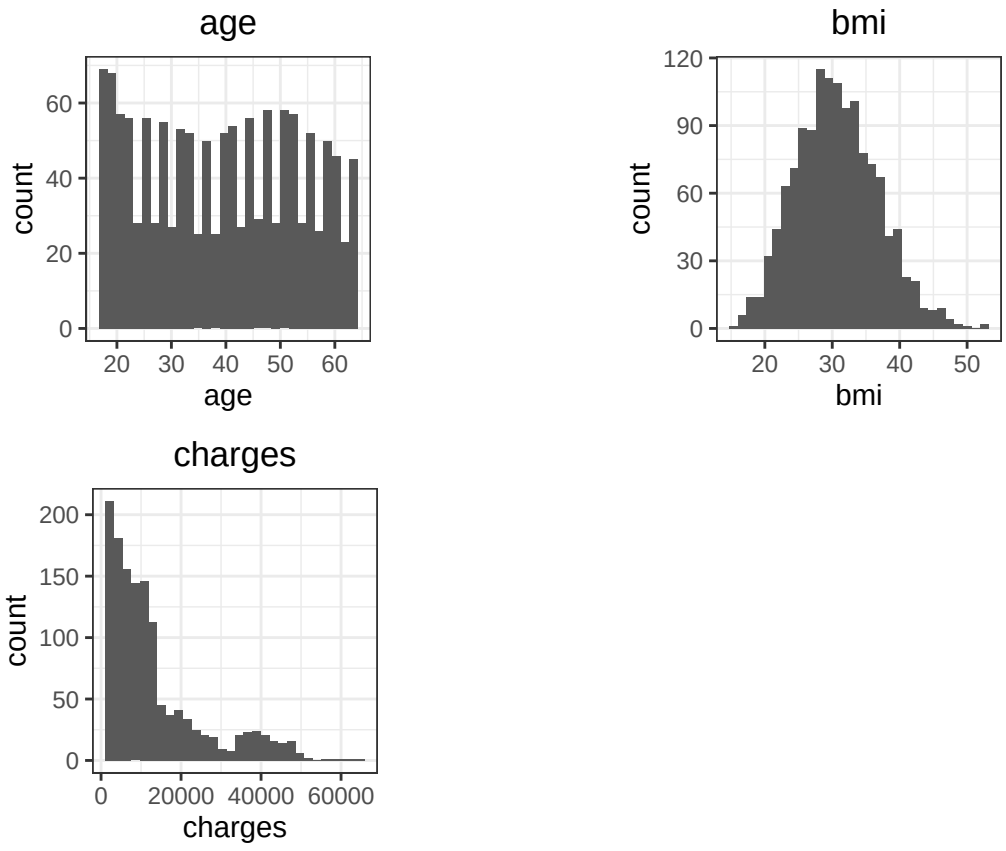
```
knitr::kable(table(df$children))
```

Var1	Freq
0	574
1	324
2	240
3	157
4	25
5	18

```
df <- df %>%
  mutate(children = factor(children))
```

2

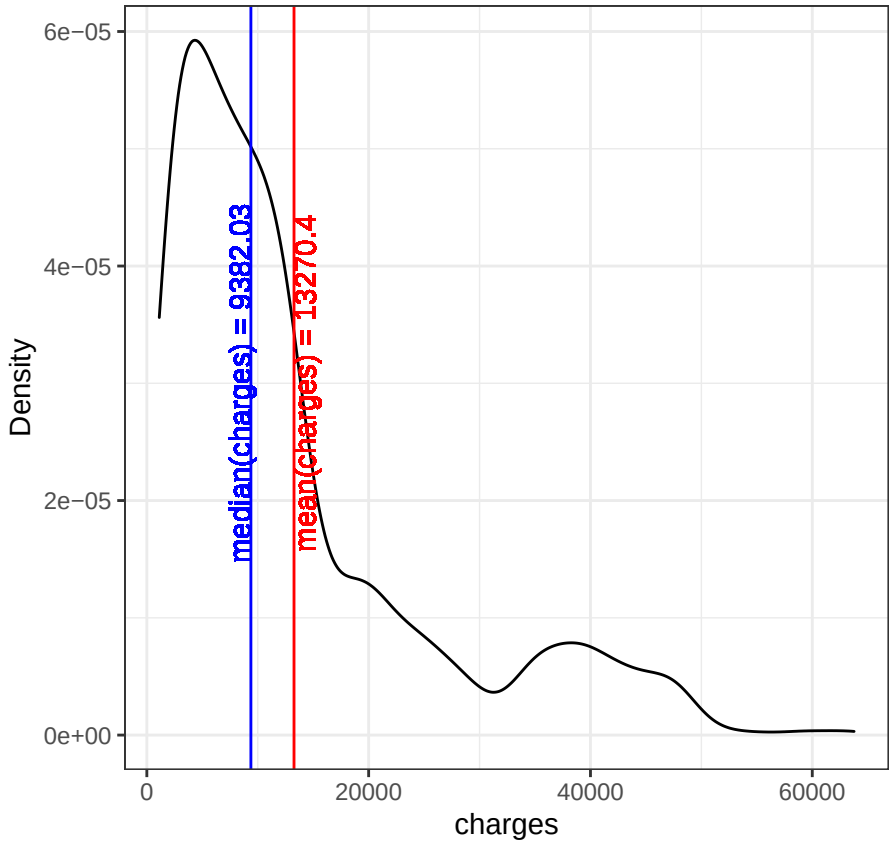
```
get_hist <- function(df, variable) {
  ggplot(df, aes_string(x = variable))+
    geom_histogram()+
    theme_bw()+
    theme(aspect.ratio = 1,
          plot.title = element_text(hjust = 0.5))+
    ggtitle(sprintf('%s', variable))
}

hists <- lapply(colnames(df %>% select_if(is.numeric)), function(col) get_hist(df, col))
ggarrange(plotlist = hists)
```



3

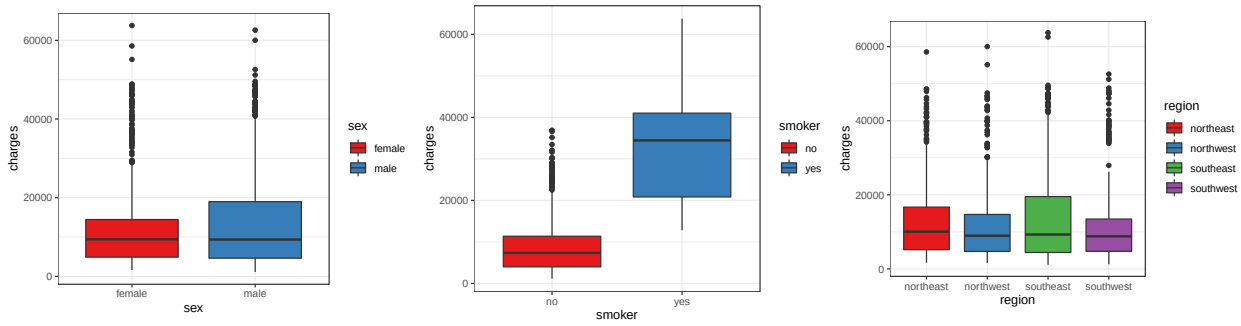
```
p3 <- ggplot(data = df, aes(x = charges))+
  geom_density(alpha = 0.5)+
  geom_vline(xintercept = mean(df$charges), color='red')+
  geom_text(aes(x=mean(df$charges) + 1e3, label=sprintf("mean(charges) = %g", mean(df$charges)),
                y=0.00003), colour="red",
            angle=90, text=element_text(size=12)) +
  geom_vline(xintercept = median(df$charges), color='blue')+
  geom_text(aes(x=median(df$charges) - 1e3, label=sprintf("median(charges) = %g", median(df$charges)),
```

```
                y=0.00003), colour="blue",
            angle=90, text=element_text(size=12)) +
  theme_bw()+
  ylab('Density')+
  theme(aspect.ratio = 1)
p3
```



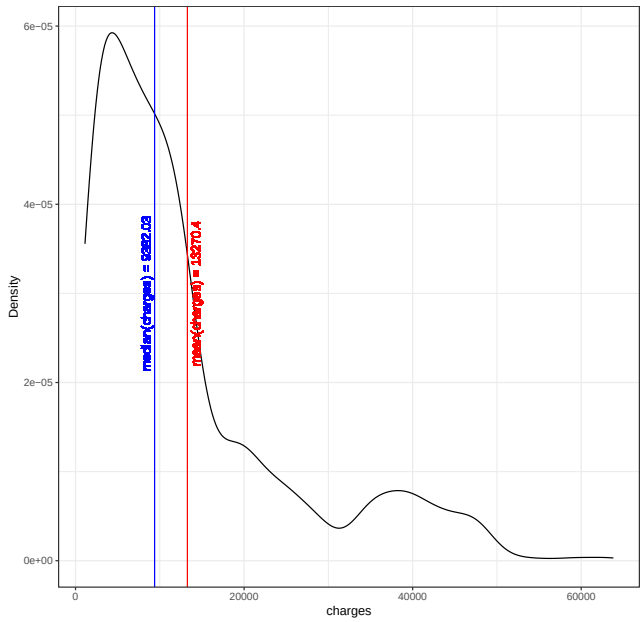
4

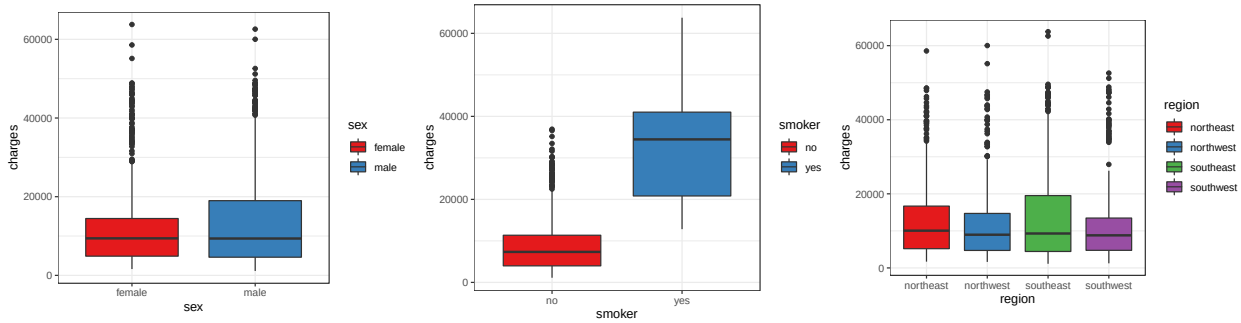
```
get_bxplt <- function(df, x, y) {
  ggplot(data = df, aes_string(y = y, x = x, fill = x))+
    geom_boxplot()+
    theme_bw()+
    scale_fill_brewer(palette = 'Set1')+
    theme(aspect.ratio = 1)
}

bxplts <- lapply(c('sex', 'smoker', 'region'), function(x)
  get_bxplt(df = df, x = x, y = 'charges'))
p4 <- ggarrange(plotlist = bxplts, ncol = 3)
p4
```



5

```
ggarrange(p3, p4, nrow = 2)
```

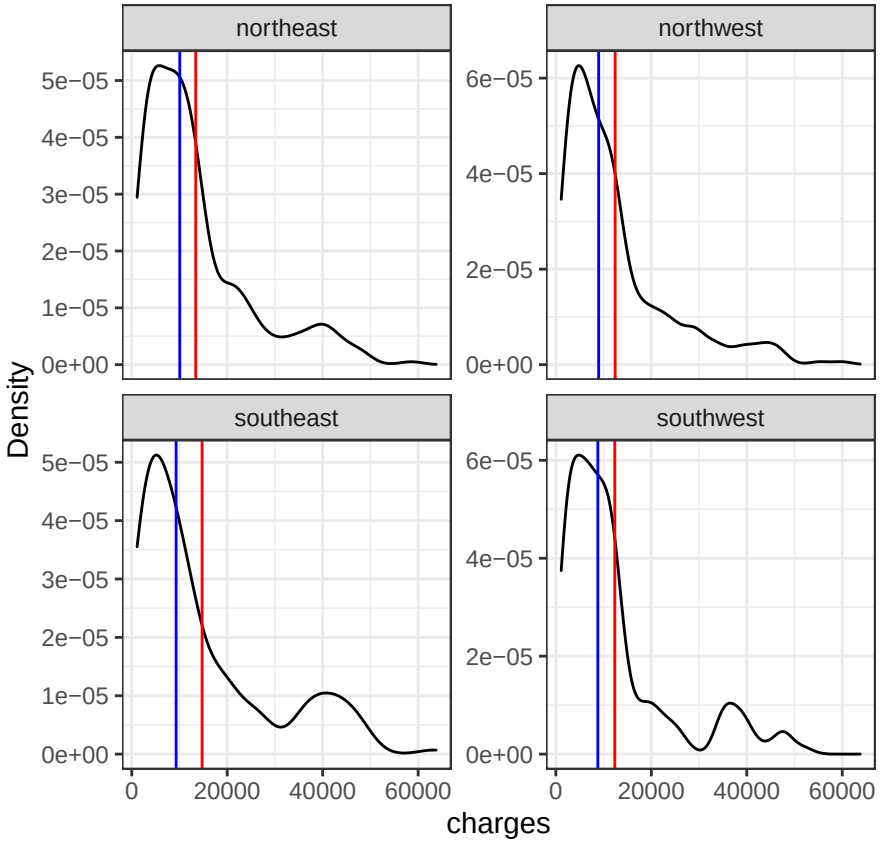




6

```
mean_median_df <- df %>%
  select(charges, region) %>%
  group_by(region) %>%
  summarise(mean=mean(charges), median=median(charges))


ggplot(df, aes(x = charges))+
  geom_density()+
  facet_wrap(~ region, scales='free_y')+
  geom_vline(data = mean_median_df, aes(xintercept = mean), color="red")+
```

```
  geom_vline(data = mean_median_df, aes(xintercept = median), color="blue")+
  theme_bw()+
  ylab('Density')+
  theme(aspect.ratio = 1)
```

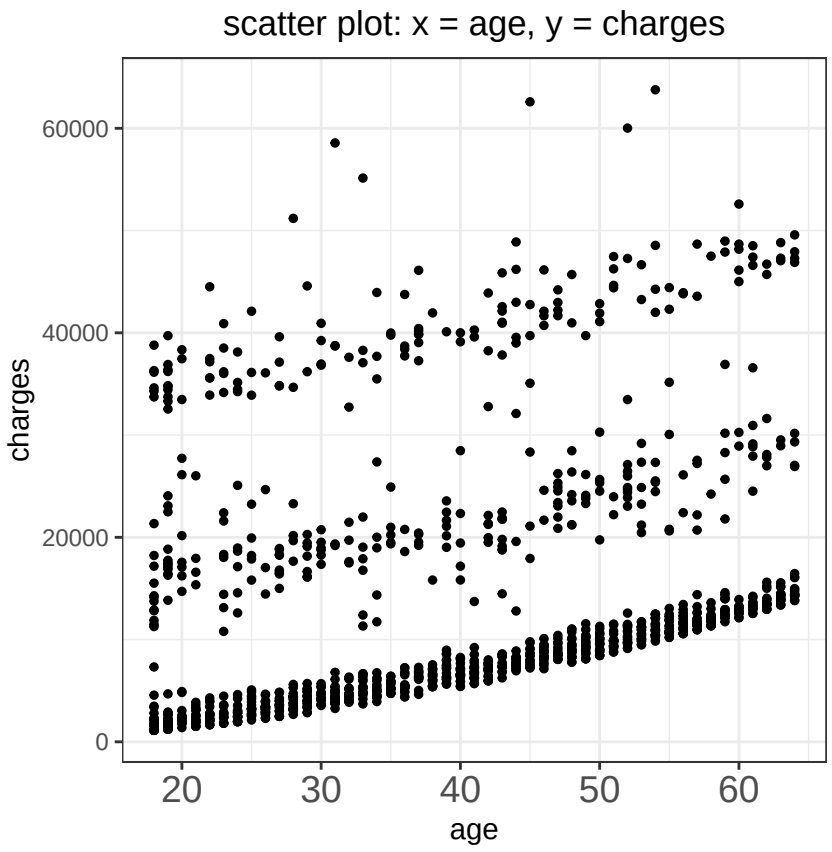



```
knitr::kable(mean_median_df)
```

region	mean	median
northeast	13406.38	10057.652
northwest	12417.58	8965.796
southeast	14735.41	9294.132
southwest	12346.94	8798.593

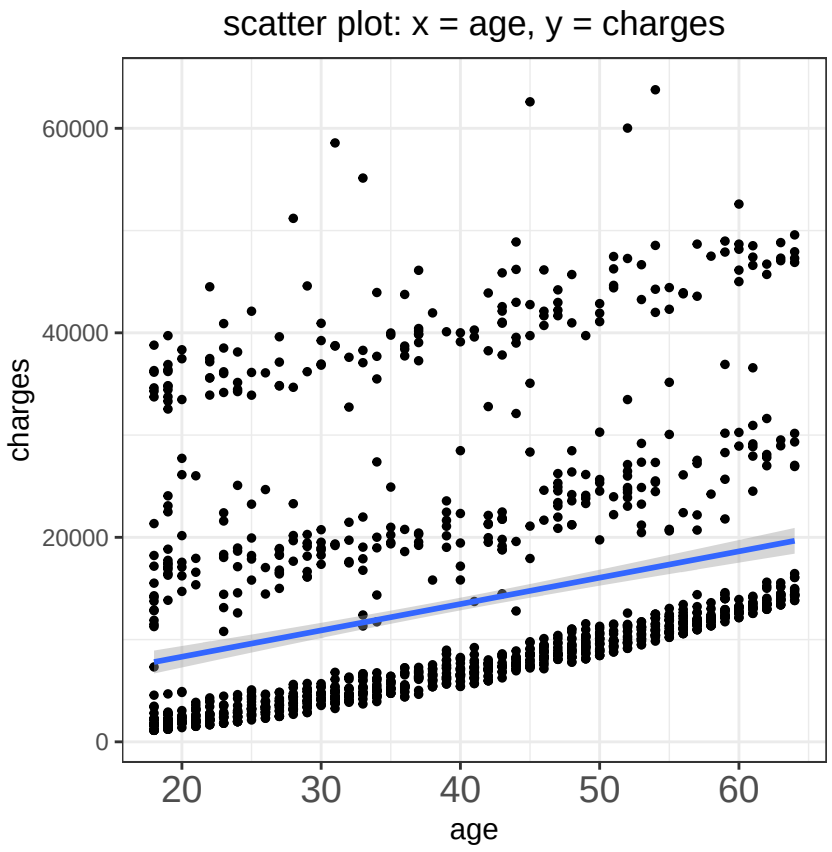
7

```
ggplot(df, aes(x = age, y = charges))+
  geom_point(size=1)+
  theme_bw()+
  ggtitle('scatter plot: x = age, y = charges')+
  theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
        plot.title = element_text(hjust = 0.5))
```



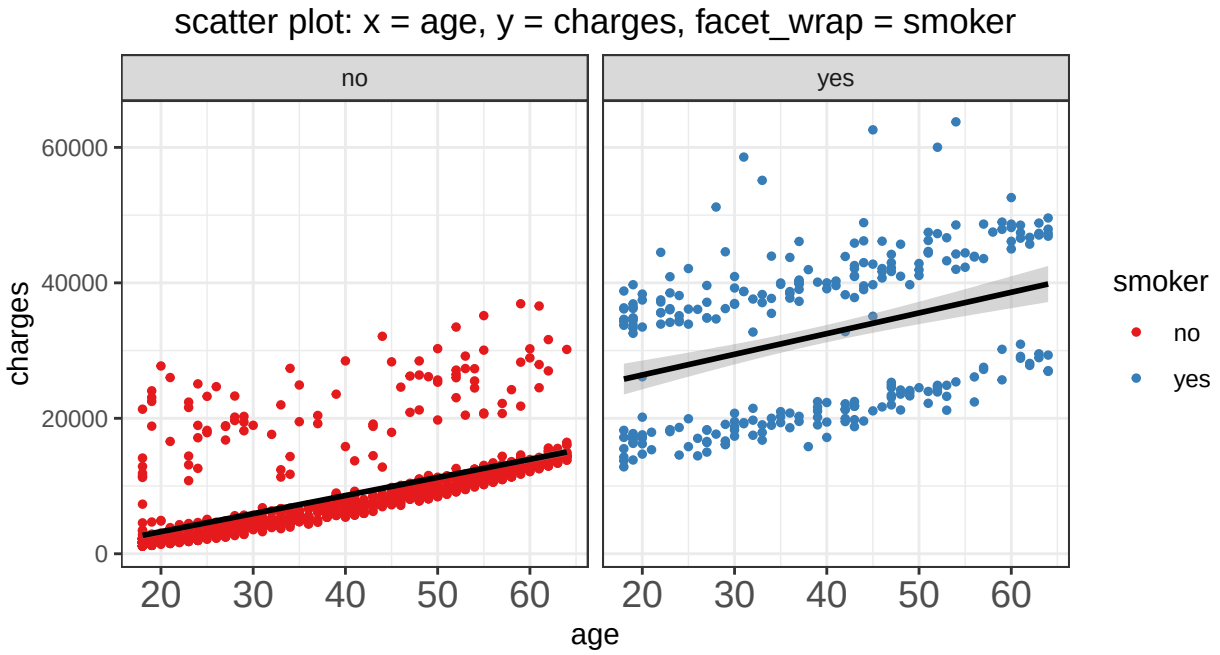
8

```
ggplot(df, aes(x = age, y = charges))+
  geom_point(size=1)+
  geom_smooth(method = lm) +
  theme_bw()+
  ggtitle('scatter plot: x = age, y = charges')+
  theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
        plot.title = element_text(hjust = 0.5))
```



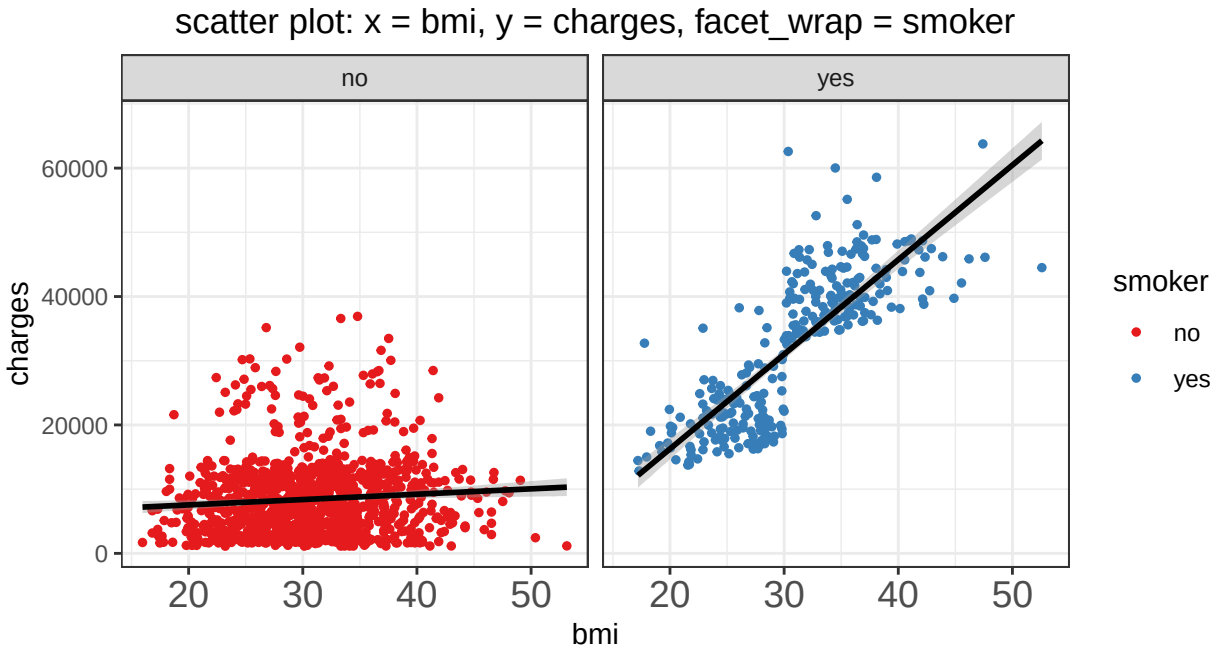
9

```
ggplot(df, aes(x = age, y = charges, col=smoker))+
    geom_point(size=1)+
    facet_wrap(.~smoker)+
    geom_smooth(method = lm, colour="black") +
    scale_color_brewer(palette = 'Set1')+
    theme_bw()+
    ggtitle('scatter plot: x = age, y = charges, facet_wrap = smoker')+
    theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
          plot.title = element_text(hjust = 0.5))
```



10

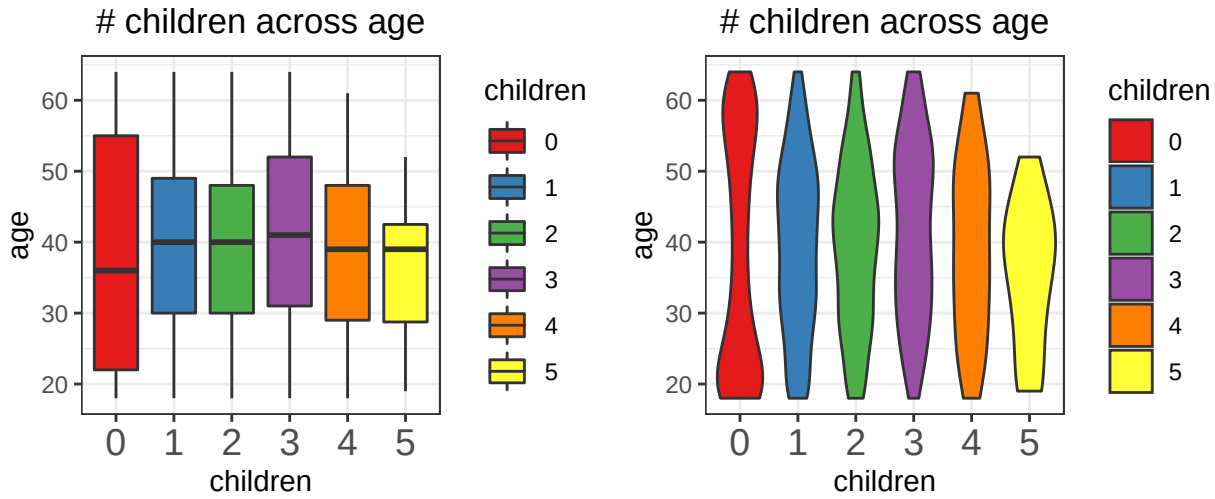
```
ggplot(df, aes(x = bmi, y = charges, col=smoker))+
    geom_point(size=1)+
    facet_wrap(.~smoker)+
    geom_smooth(method = lm, colour="black") +
    scale_color_brewer(palette = 'Set1')+
    theme_bw()+
    ggtitle('scatter plot: x = bmi, y = charges, facet_wrap = smoker')+
    theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
          plot.title = element_text(hjust = 0.5))
```



11

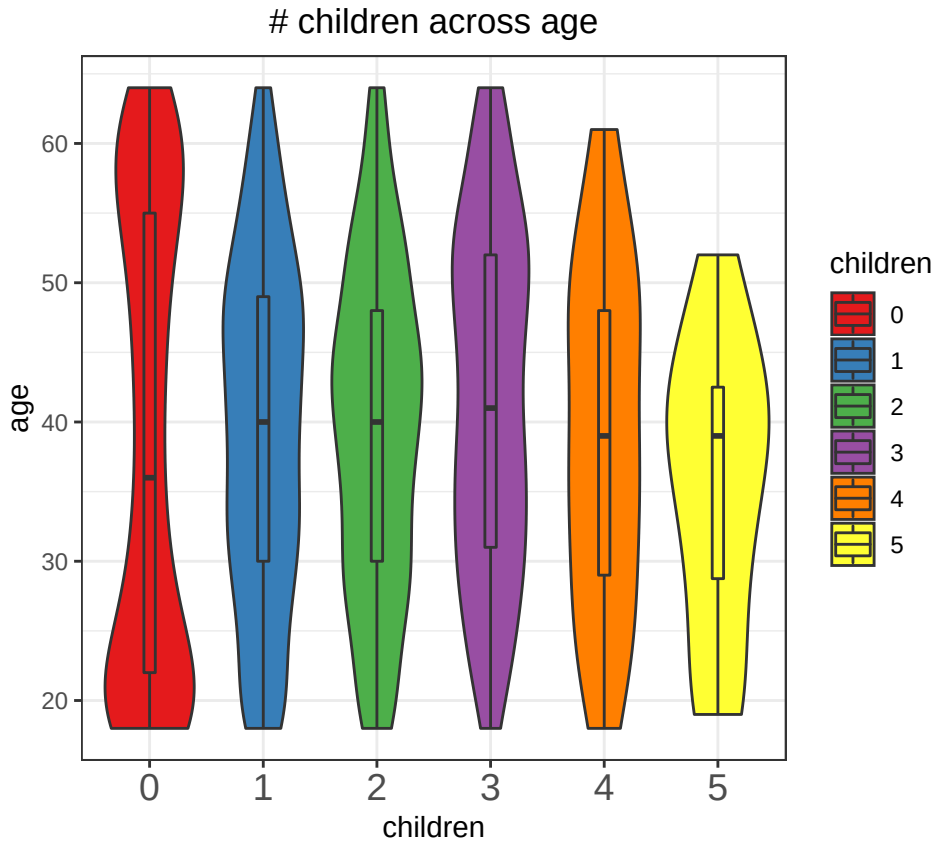
How number of children is distributed across the age?

```
p1 <- ggplot(df, aes(x = children, y = age, fill=children))+
  geom_boxplot()+
  theme_bw()+
  ggtitle('# children across age')+
  scale_fill_brewer(palette='Set1')+
  theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
        plot.title = element_text(hjust = 0.5))

p2 <- ggplot(df, aes(x = children, y = age, fill=children))+
  geom_violin()+
  theme_bw()+
  ggtitle('# children across age')+
  scale_fill_brewer(palette='Set1')+
  theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
        plot.title = element_text(hjust = 0.5))

ggarrange(p1, p2)
```



We can use boxplots and violin plots as well. Violin plots is better, since we can see the exact information about target variable distribution. However, the perfect option is the combination of boxplot & violinplot at the same layout

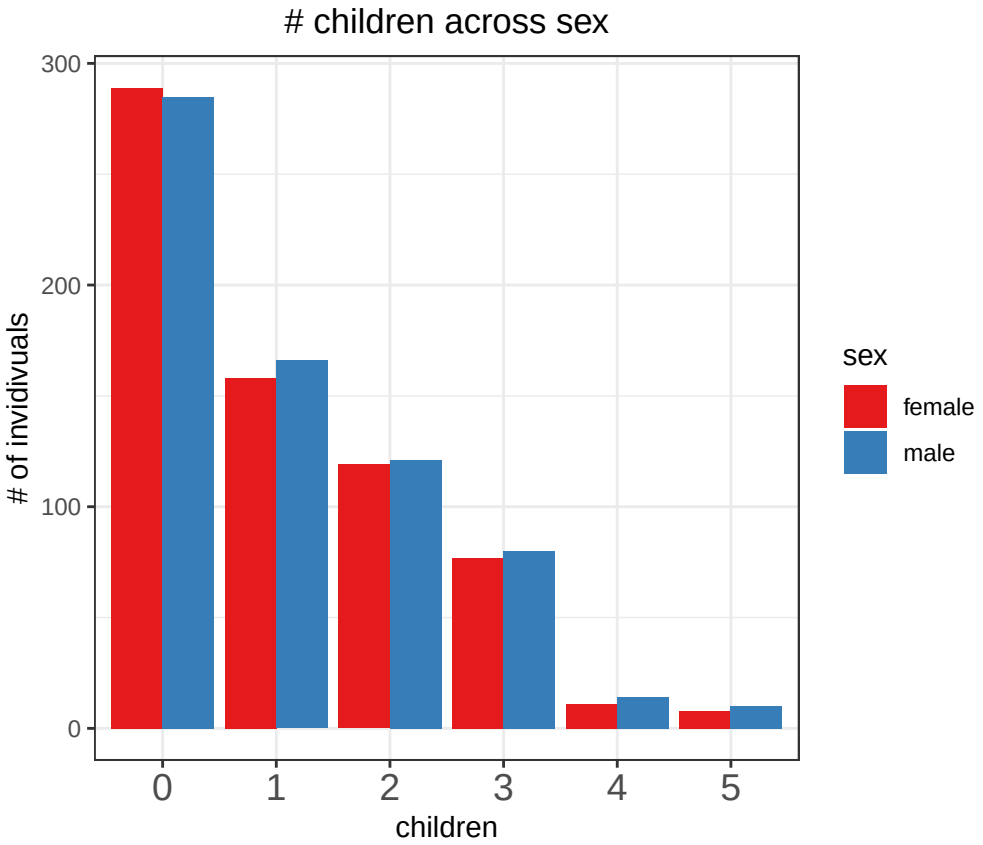
```
p2+geom_boxplot(width=0.1)
```



12

How number of children is distributed across females and males?

```
ggplot(df, aes(children, ..count..))+
  geom_bar(aes(fill = sex), position = "dodge")+
  theme_bw()+
  ggtitle('# children across sex')+
  scale_fill_brewer(palette='Set1')+
  ylab('# of invidivuals')+
  theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
        plot.title = element_text(hjust = 0.5))
```

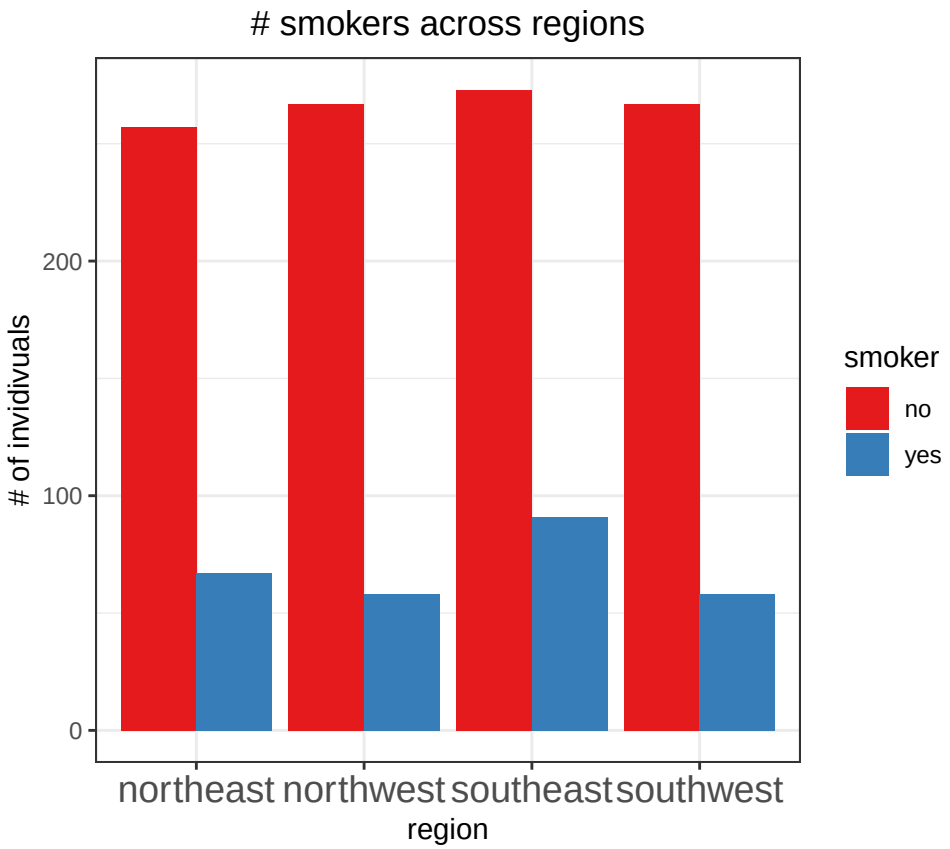


Barplots suitable for the visualization of categorical variables.

13

How smokers are distributed across regions (in terms of "raw" counts)?

```
ggplot(df, aes(region, ..count..))+
  geom_bar(aes(fill = smoker), position = "dodge")+
  theme_bw()+
  ggtitle('# smokers across regions')+
  scale_fill_brewer(palette='Set1')+
  ylab('# of invidivuals')+
  theme(aspect.ratio = 1, axis.text.x = element_text(size = 14),
        plot.title = element_text(hjust = 0.5))
```

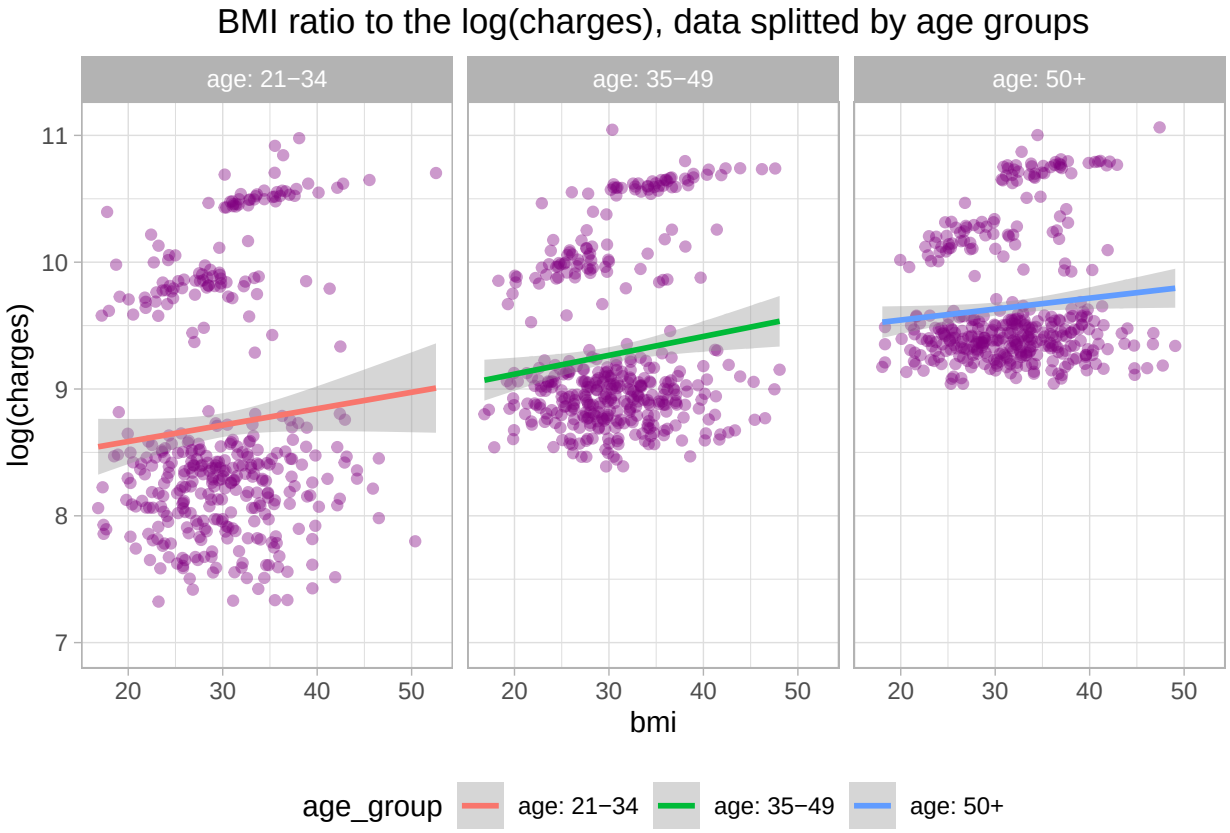


Since "raw counts" are comparible between different regions, we can inrepret this plot without switching to the proportions (however, such type of questions should be answered using proportions, since number of observations can vary between different groups).

Barplots suitable for the visualization of categorical variables.

14

```
df <- df %>%
  filter(age >= 21) %>%
  mutate(age_group = ifelse(age <= 34, 'age: 21-34',
                            ifelse(age > 34 & age < 50, 'age: 35-49', 'age: 50+')))

ggplot(df, aes(x = bmi, y = log(charges)))+
  geom_point(alpha=0.4, color='#800080')+
  facet_wrap(.~age_group)+
  ylim(7, max(log(df$charges)))+
  geom_smooth(aes(group = age_group, color=age_group), method="lm")+
  theme_light()+
  theme(legend.position = 'bottom', plot.title = element_text(hjust = 0.5))+
  ggtitle('BMI ratio to the log(charges), data splitted by age groups')
```


BMI ratio to the log(charges), data splitted by age groups
age: 21−34
age: 35−49
age: 50+
log(charges)
11
10
9
8
7
20
30
40
50
bmi
age_group
age: 21−34
age: 35−49
age: 50+