price prediction

```
library(readxl)
data <- read_excel("insurance.xlsx")
head(data)
```

```
## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

```
summary(data)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   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
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

#check for any missing data

```
colSums(is.na(data))
```

```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

#plot the distribution for insurance charge #Plot indicates right-skewed distribution.There is little peak in the distribution around 40k.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

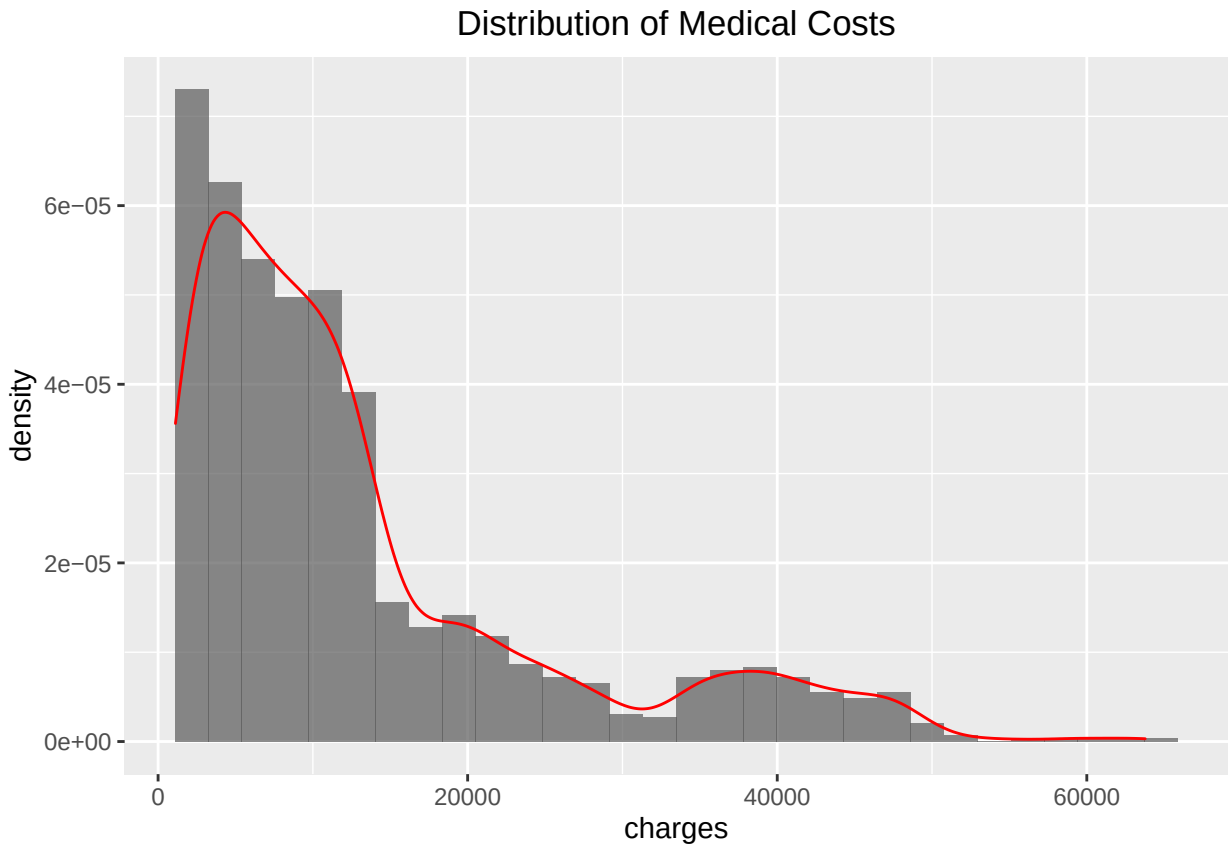
```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

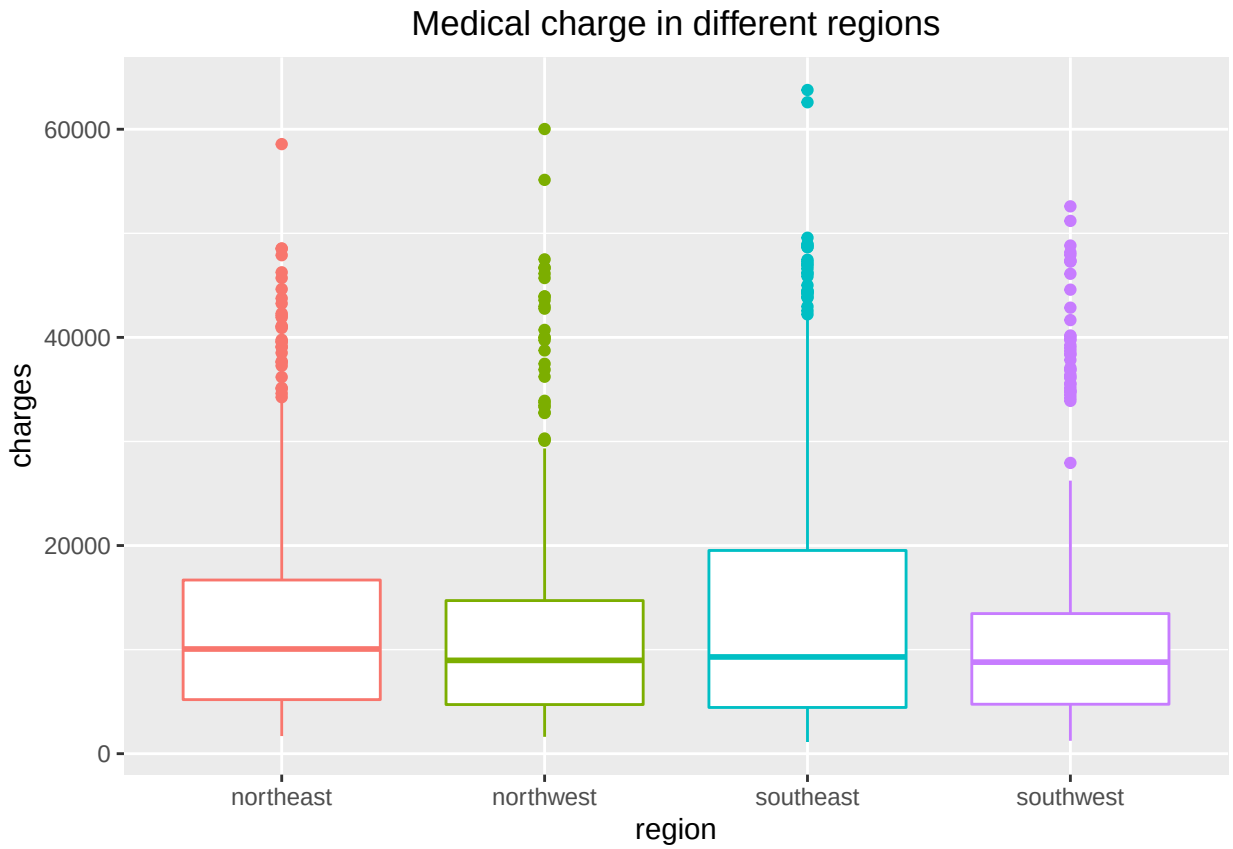
```
ggplot(data, aes(charges)) +
    geom_histogram(aes(y=..density..), alpha = 0.7)  +
    geom_density(col = "red") +
    labs(title = "Distribution of Medical Costs") +
    theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



#Insurance charge in different regions #The Southeastern region has a higher spread of cost and average medical costs is higher in southeastern reigon when compared to the other regions.Southweast has the lowest spread of cost and lowest average mdeical cost compared to other reigons.

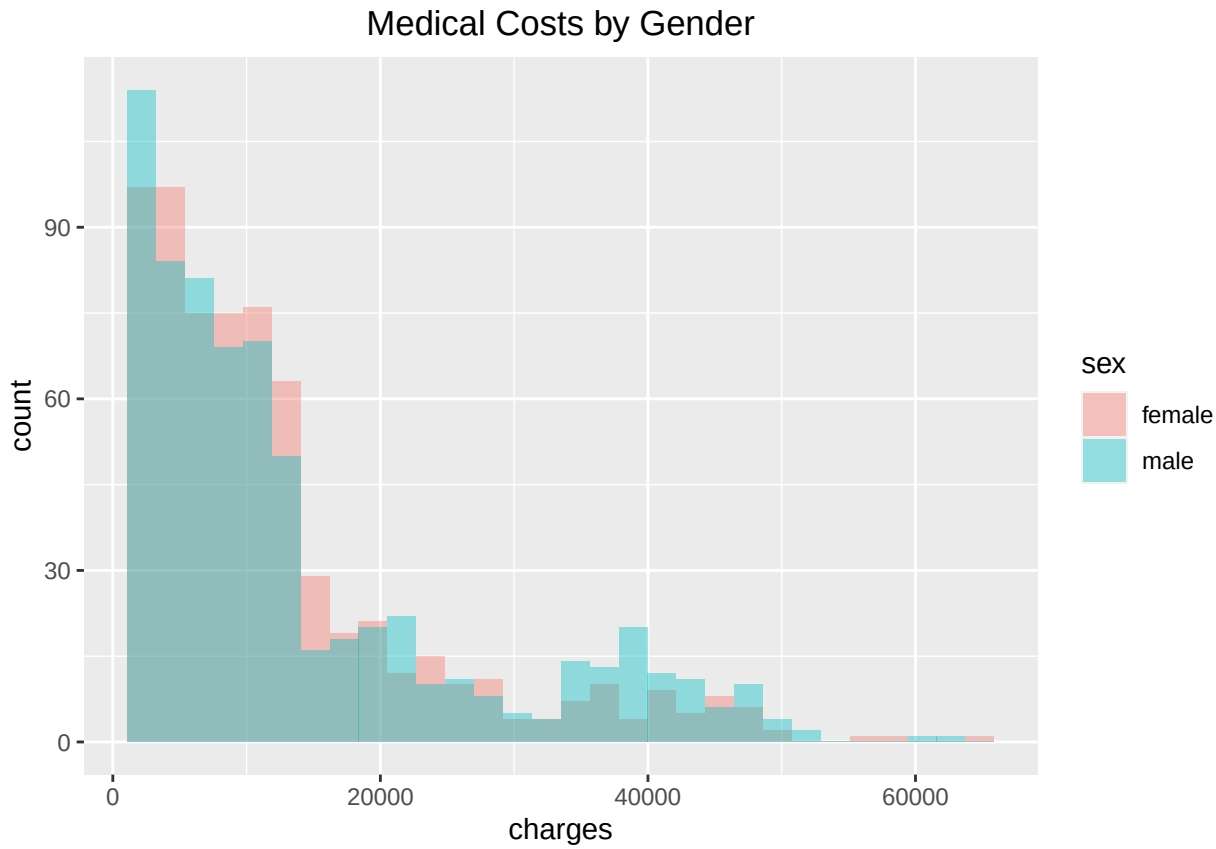
```
ggplot(data, aes(x = region, y = charges, color = region)) +
  geom_boxplot() +
  labs(title = "Medical charge in different regions") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "none")
```



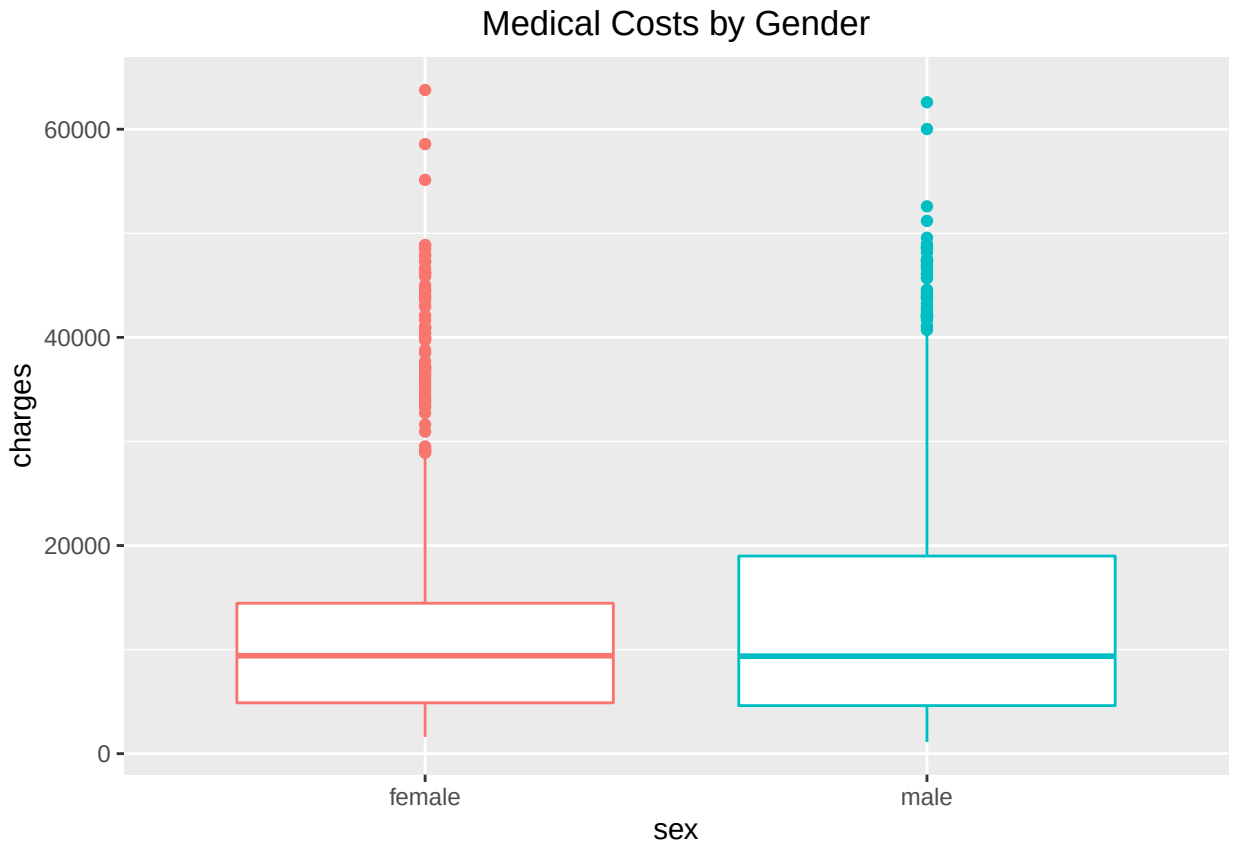
#Medical Costs by Gender. #Histogram indicates medical costs distributions for men and women look pretty similar, the box plot distribution of men is a little more spread compared to women.

```
ggplot(data, aes(x = charges, fill = sex)) +
  geom_histogram(position = "identity", alpha = 0.4) +
  labs(title = "Medical Costs by Gender") +
    theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data, aes(x = sex, y = charges, color = sex)) +
  geom_boxplot() +
  labs(title = "Medical Costs by Gender") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "none")
```



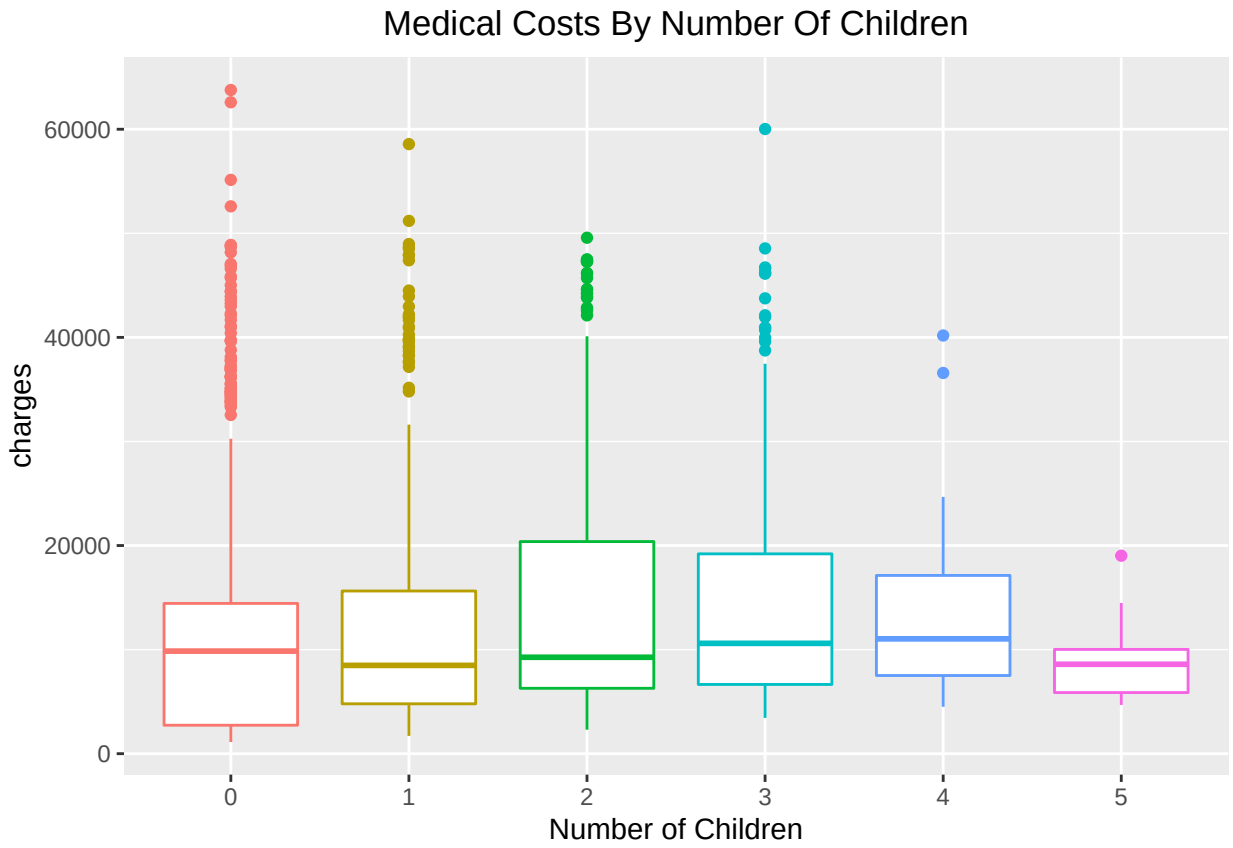
#Medical cost grouped by number of children. Average medical cost increases if there as 2/3 children.Average medical cost is less when there are 5 children.

```
data %>%
    group_by(children) %>%
    summarise(median = median(charges), mean = mean(charges),total = n()) %>%
    arrange(desc(mean))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 4
##   children median   mean total
##      <dbl>  <dbl>  <dbl> <int>
## 1        3 10601. 15355.   157
## 2        2  9265. 15074.   240
## 3        4 11034. 13851.    25
## 4        1  8484. 12731.   324
## 5        0  9857. 12366.   574
## 6        5  8590.  8786.    18
```

```
ggplot(data, aes(x = as.factor(children), y = charges, color = as.factor(children))) +
  geom_boxplot() +
  labs(title = "Medical Costs By Number Of Children",
      x = "Number of Children") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "none")
```

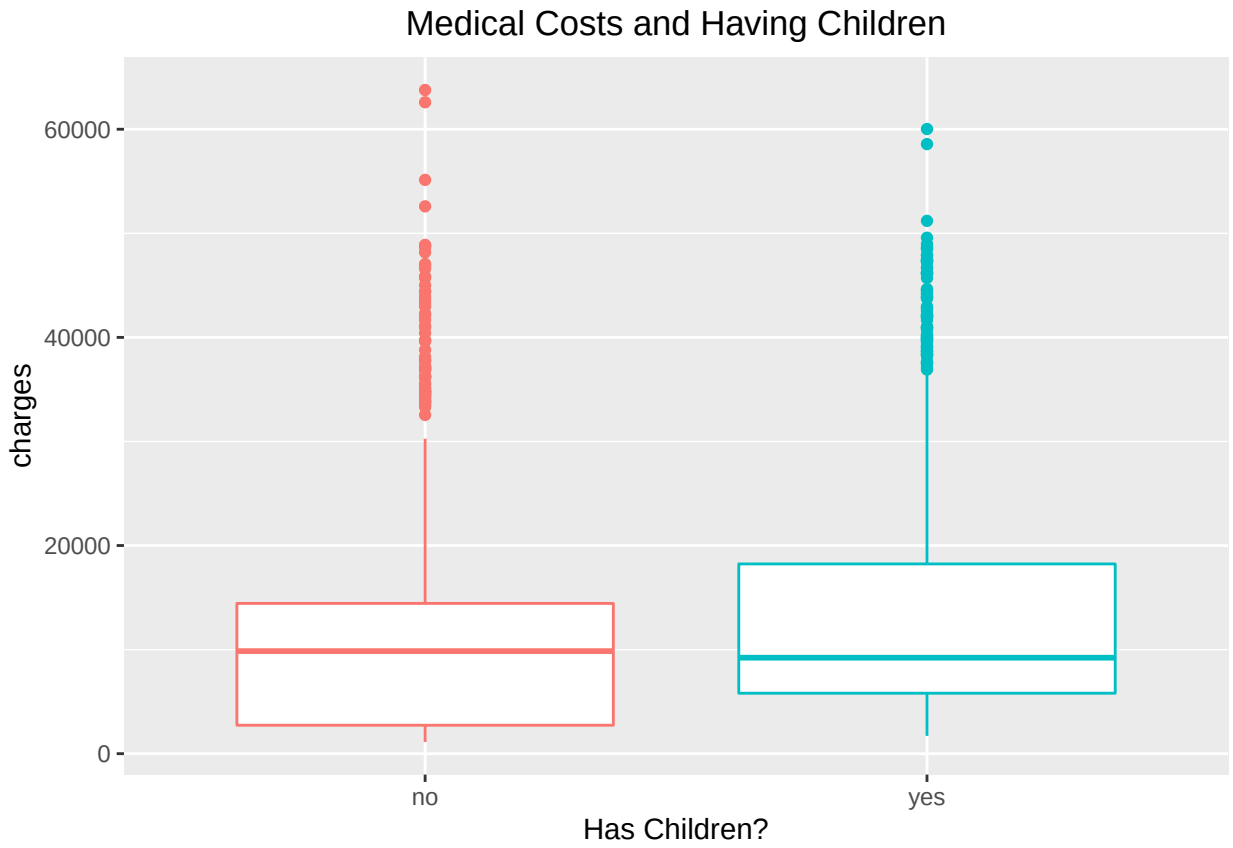


#Medical cost difference children vs no children #there's not much difference between the two group's distributions

```
data <- data %>%
  mutate(has_children = ifelse(children > 0, "yes", "no"))
data$has_children <- as.factor(data$has_children)

summary(data$has_children)
```

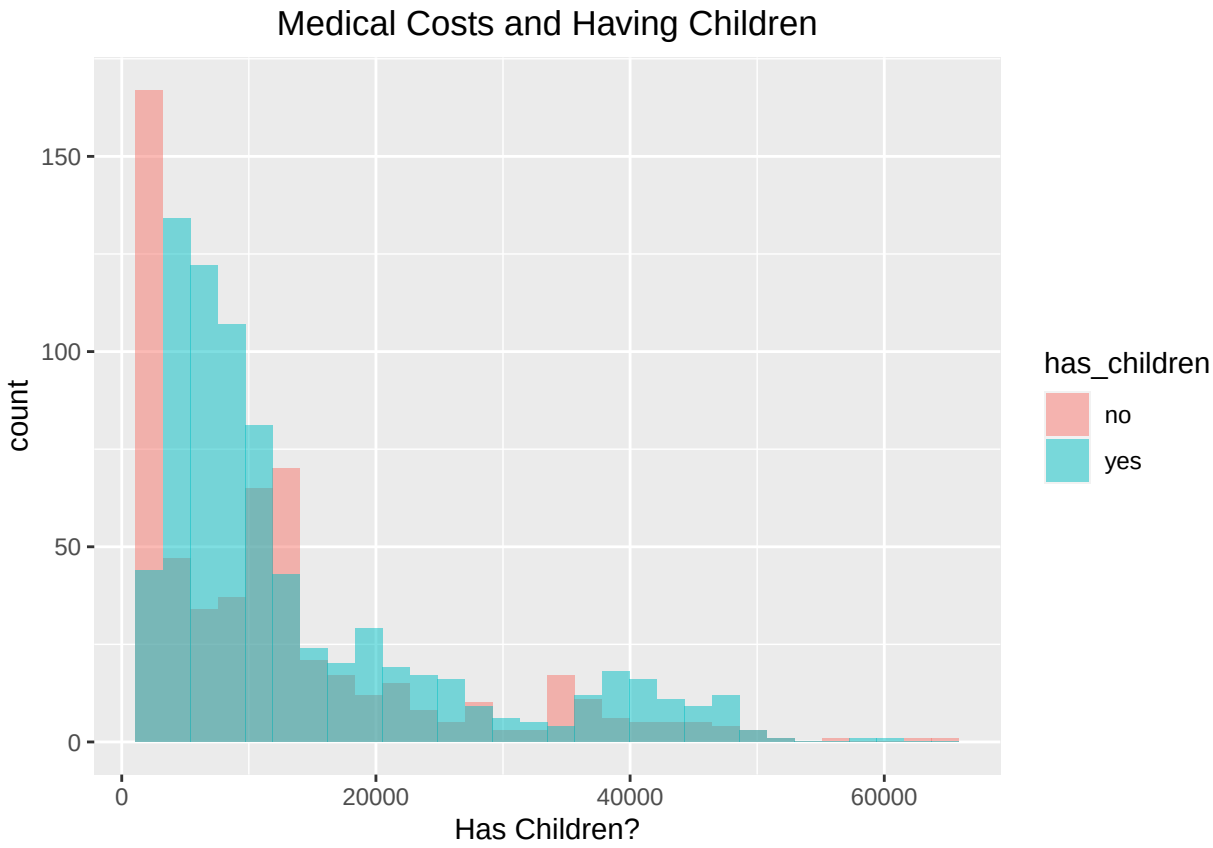
```
##  no yes
## 574 764
```

```
cc1 <- ggplot(data, aes(y = charges, x = has_children, color = has_children)) +
  geom_boxplot() +
  labs(title = "Medical Costs and Having Children",
       x = "Has Children?") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "none")

cc2 <- ggplot(data, aes(x = charges, fill = has_children)) +
  geom_histogram(position = "identity", alpha = 0.5) +
  labs(title = "Medical Costs and Having Children",
       x = "Has Children?") +
    theme(plot.title = element_text(hjust = 0.5))

plot(cc1)
```



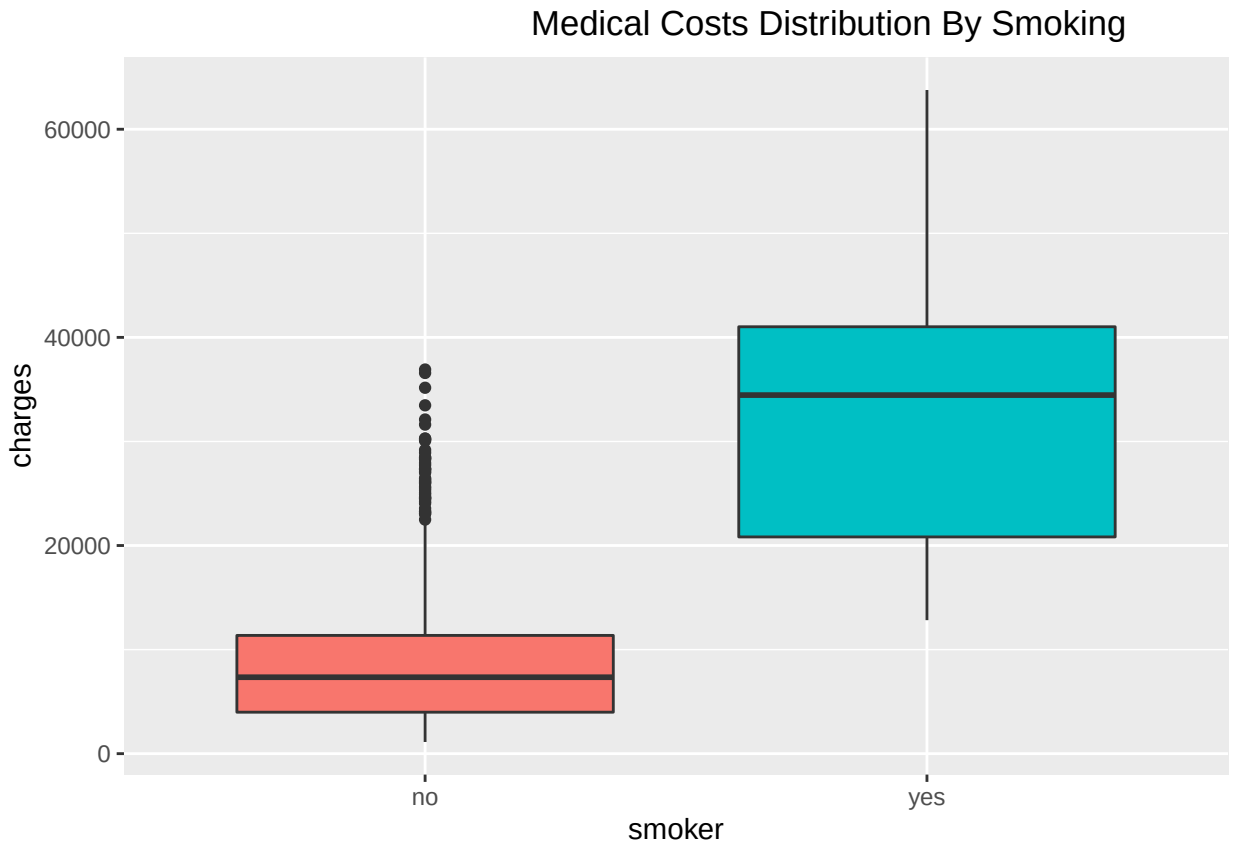

```
plot(cc2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



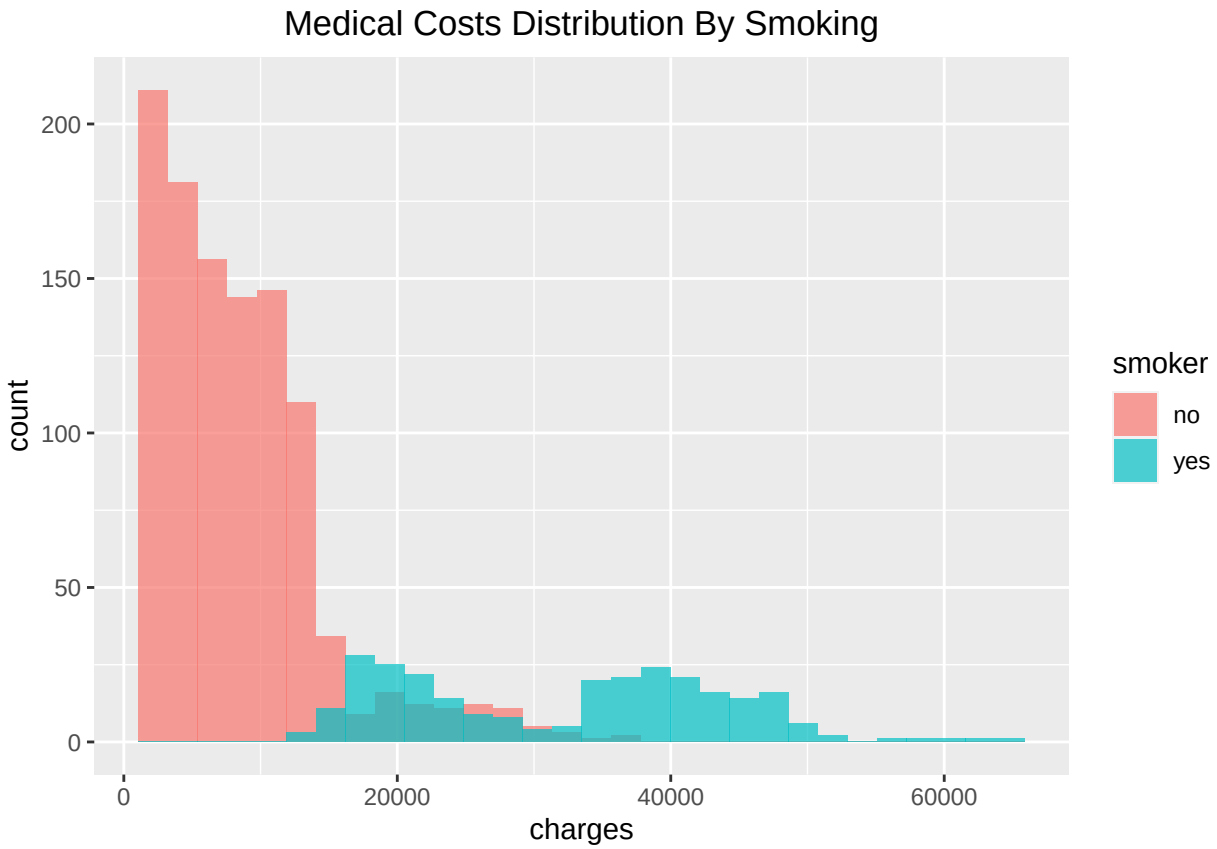
#Medical cost difference smoker vs non-smoker # Smokers have an increase medical expense compared to non-smokers. Smoker's distribution looks bimodal, with one mode around 20k and another around 40k.

```
cs1 <- ggplot(data, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot()+
  labs(title = "Medical Costs Distribution By Smoking") +
    theme(plot.title = element_text(hjust = 0.8),
          legend.position = "none")

cs2 <- ggplot(data, aes(charges, fill = smoker)) +
  geom_histogram(alpha = 0.7, position = "identity")+
  labs(title = "Medical Costs Distribution By Smoking") +
    theme(plot.title = element_text(hjust = 0.5))

plot(cs1)
```



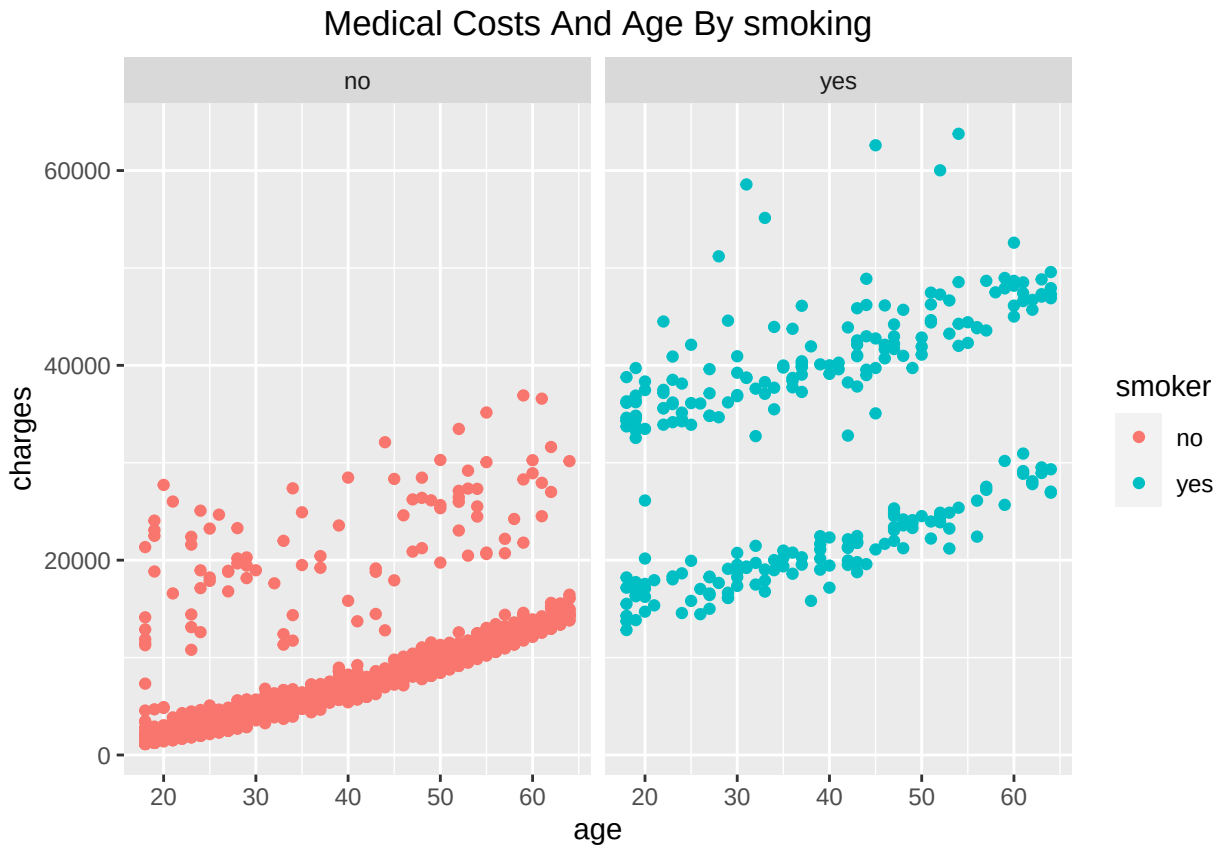

```
plot(cs2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



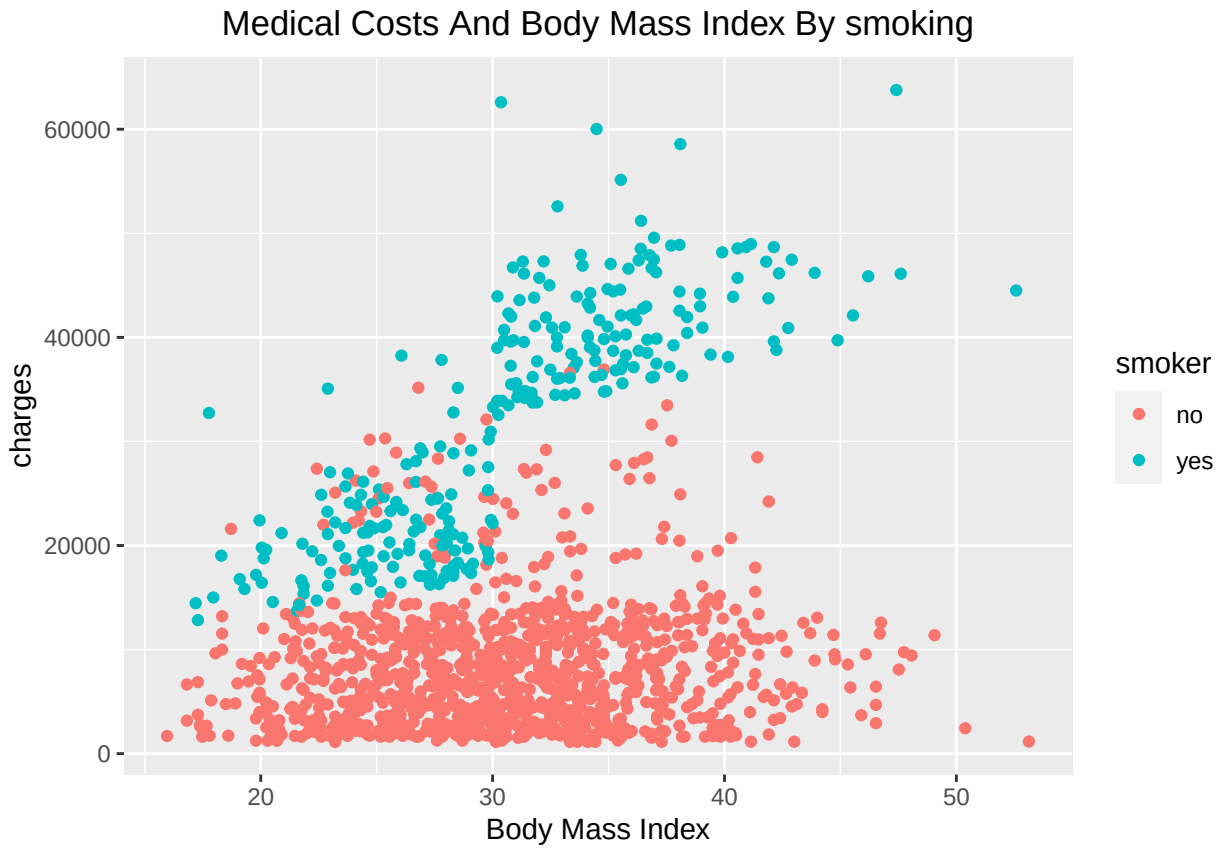
#Smoker's medical costs get higher with age when compared to non-smokers.

```
ggplot(data, aes(x = age, y = charges, color = smoker)) +
  geom_point() +
  facet_wrap(~smoker) +
  labs(title = "Medical Costs And Age By smoking") +
    theme(plot.title = element_text(hjust = 0.5))
```

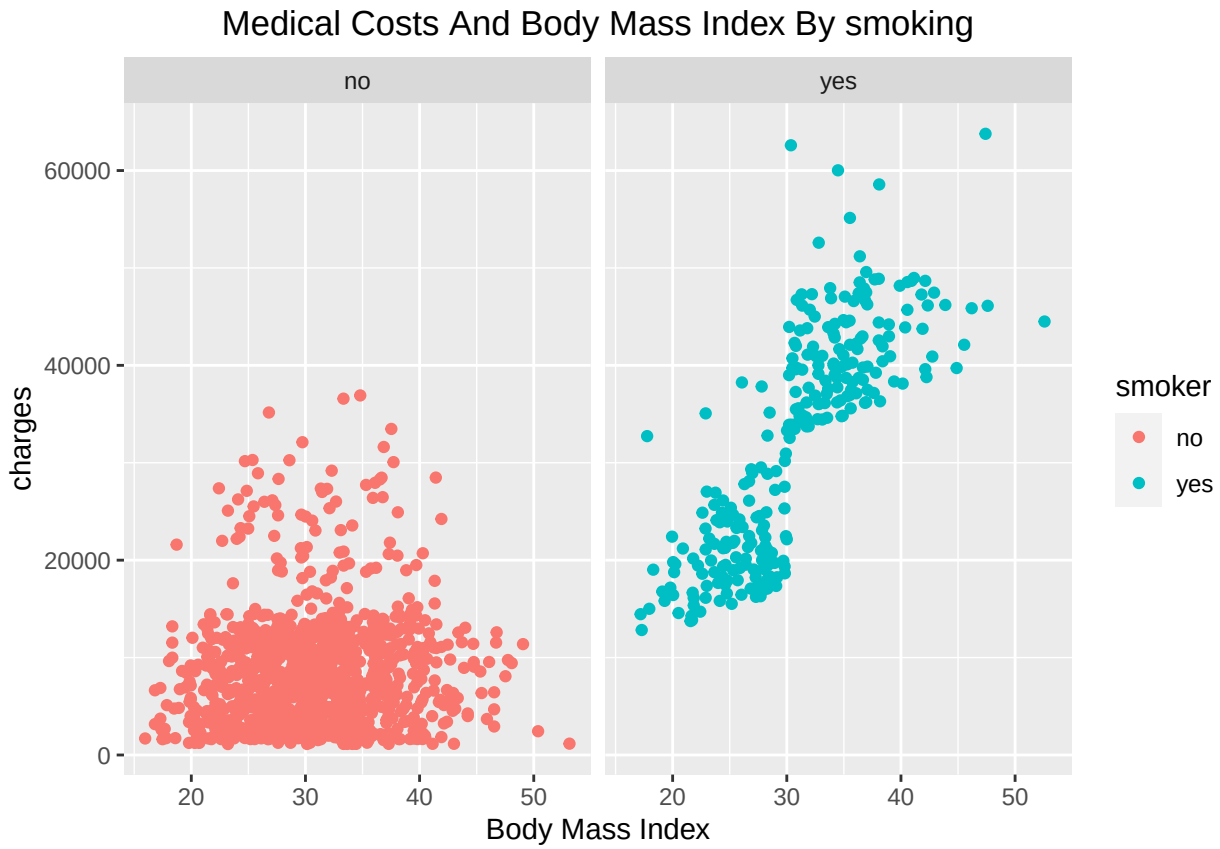


#Smoker's medical costs increase more rapidly when their body mass increases.

```
cm1 <- ggplot(data, aes(x = bmi, y = charges, color = smoker)) +
  geom_point() +
  labs(title = "Medical Costs And Body Mass Index By smoking",
       x = "Body Mass Index") +
    theme(plot.title = element_text(hjust = 0.5))

cm2 <- ggplot(data, aes(x = bmi, y = charges, color = smoker)) +
  geom_point() +
  labs(title = "Medical Costs And Body Mass Index By smoking",
       x = "Body Mass Index") +
    theme(plot.title = element_text(hjust = 0.5)) +
    facet_wrap(~smoker)


plot(cm1)
```


Medical Costs And Body Mass Index By smoking
charges
60000
40000
20000
0
20
30
40
50
Body Mass Index
smoker
no
yes

```
plot(cm2)
```



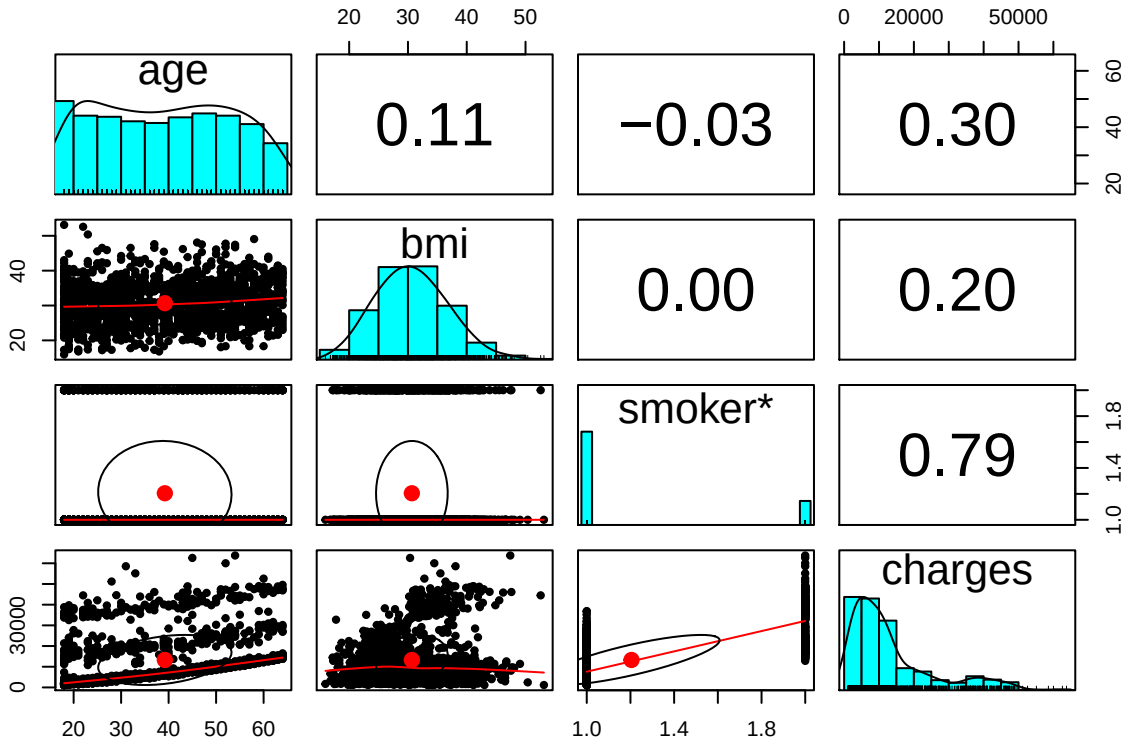
#Correlation between variables

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(data[c("age", "bmi", "smoker", "charges")])
```




```
library("cowplot")
```