The price of being a smoker

Analysing insurance data of smokers and non-smokers

Federspiel Sven

2022-01-09

Introduction

For this Project we will have a look at the insurance data of smokers and non-smokers from https://www.kaggle.com/mirichoi0218/insurance which shows us how expensive smoking realy is.

The data

```
data <- read_csv("insurance.csv")
```

```
## Rows: 1338 Columns: 7

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

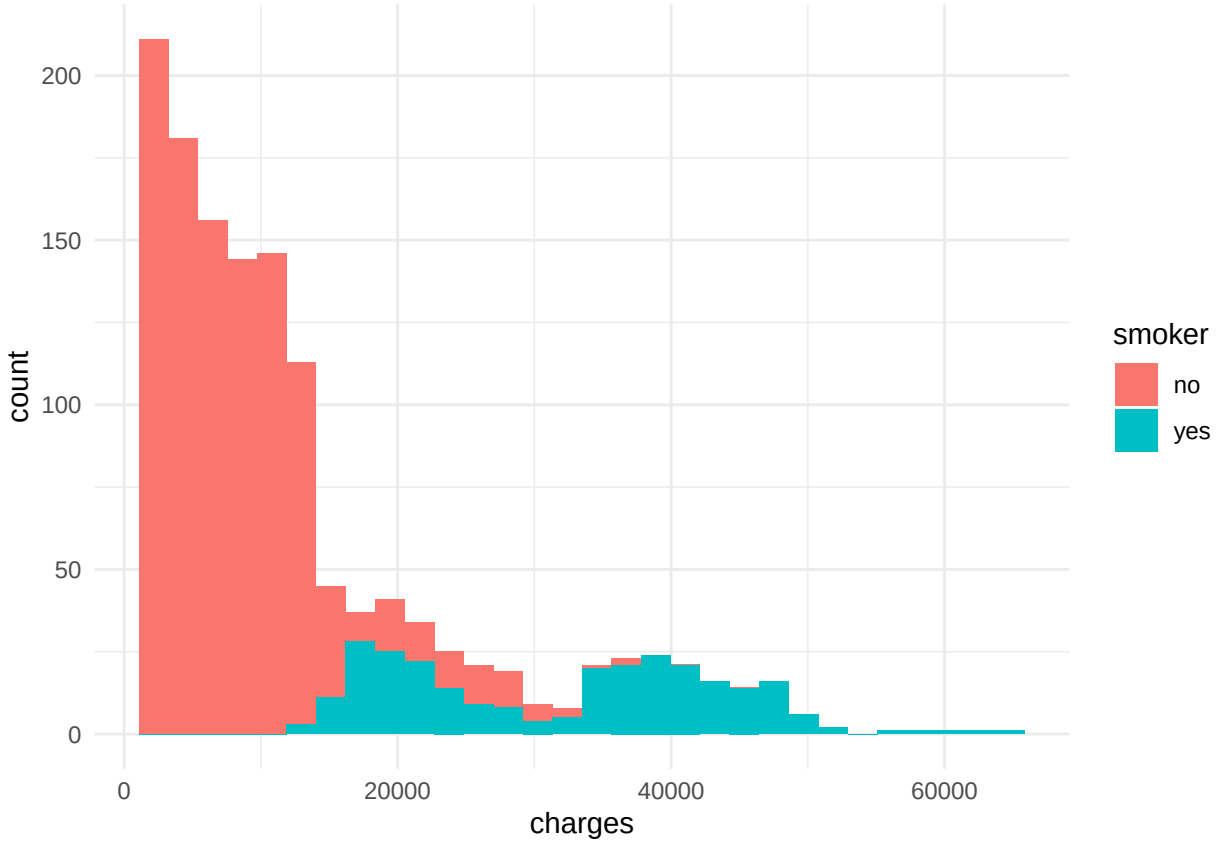
```
data
```

```
## # A tibble: 1,338 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.
##  7    46 female  33.4        1 no     southeast   8241.
##  8    37 female  27.7        3 no     northwest   7282.
##  9    37 male    29.8        2 no     northeast   6406.
## 10    60 female  25.8        0 no     northwest  28923.
## # ... with 1,328 more rows
```

Histogram

```
data %>%
 ggplot() +
  geom_histogram(aes(x = charges,fill = smoker)) +
  theme_minimal()
```

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



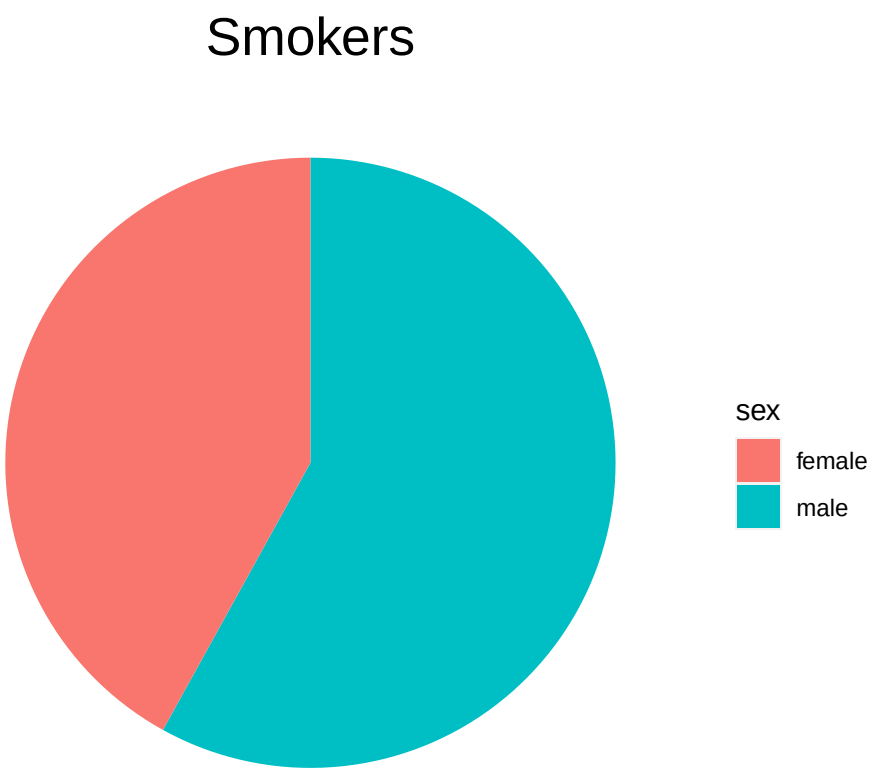
We can see that smokers have higher charges than non-smokers.

Pie chart

```
data2 = data.frame(sex = c("male","female"),
                   count = c(sum(data$smoker == "yes" & data$sex == "male"),
                             sum(data$smoker == "yes" & data$sex == "female")))

data2 %>%
  ggplot() +
  geom_bar(stat="identity",aes(x = "", y = count , fill=sex)) +
  coord_polar("y", start=0) +
```

```
  ggtitle("Smokers")+
  theme(axis.title = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        plot.title = element_text(hjust = 0.5, size = 20)
        )
```



We can see that the majority of smokers are male.

Scatterplot

```
data %>%
  ggplot(aes(x = charges ,y = bmi, color = smoker))+
  geom_point()+
  geom_smooth(method='lm', formula= y~x)+
  theme_minimal()
```

We can see that the bmi does not seem to have much impact on charges for non-smokers, but for smokers a higher bmi means higher charges.