Project of Foundations of Probability and Statistics

Paola Cavana, Julia Tsymbal

2022-2023

1 Introduction

Main goal of this analysis is to predict insurance costs for customers.

Link al dataset (https://www.kaggle.com/datasets/mirichoi0218/insurance?resource=download)

Columns:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- 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
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary’s residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

The formula of the BMI consists in the division of the weight of an adult subject, expressed in kilograms (kg), by the square of its height expressed in meters (m).

Category:

BMI range (kg/m2).

- Obesity class III (very serious): > 40,00.
- Obesity class II (severe): 35,01 to 40,00.
- Obesity class I (moderate): 30,01 to 35,00.
- Overweight: 25,01 to 30,00.
- Regular: 18,51 to 25,00.
- Slightly underweight: 17,51 to 18,50.
- Underweight: 16,01 to 17,50.
- Severe leanness (starvation): <16,01

1.1 Requirements

- install.packages(“ggplot2”)
- install.packages(“ggpubr”)
- install.packages(“GGally”)
- install.packages(“ggpairs”)
- install.packages(“wesanderson”)
- install.packages(“modest”)
- install.packages(“MASS”)
- install.packages(“dplyr”)
- install.packages(“caret”)
- install.packages(“tidyverse”)
- install.packages(“patchwork”)
- install.packages(“ggforce”)
- install.packages(“car”)

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.2
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.2.2
```

```
library(RColorBrewer)
library(wesanderson)
```

```
## Warning: package 'wesanderson' was built under R version 4.2.2
```

```
library(modest)
```

```
## Warning: package 'modest' was built under R version 4.2.2
```

```
library(MASS)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ tibble  3.1.8     ✓ purrr   0.3.5
## ✓ tidyr   1.2.1     ✓ stringr 1.4.1
## ✓ readr   2.1.3     ✓ forcats 0.5.2
```

```
## Warning: package 'tibble' was built under R version 4.2.2
```

```
## Warning: package 'tidyr' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
## Warning: package 'forcats' was built under R version 4.2.2
```

```
## ── Conflicts ─────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ purrr::lift()   masks caret::lift()
## ✖ dplyr::select() masks MASS::select()
```

```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.2.2
```

```
##
## Attaching package: 'patchwork'
##
## The following object is masked from 'package:MASS':
##
##     area
```

```
library(ggforce)
```

```
## Warning: package 'ggforce' was built under R version 4.2.2
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.2.2
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:purrr':
##
##     some
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

2 Data Exploration

we load the dataset:

```
insurance <- read.table("insurance.csv", header = TRUE, sep = ",")
```

Once the file is loaded, it is useful to carry out an exploratory analysis to observe its contents and features.

```
dim(insurance)
```

```
## [1] 1338    7
```

As we can see from the output, the dataset consists of 1338 rows and 7 columns.

```
colnames(insurance)
```

```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

Look at the head and tail of the datasets:

```
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
```

```
tail(insurance)
```

```
##      age    sex   bmi children smoker    region   charges
## 1333  52 female 44.70        3     no southwest 11411.685
## 1334  50   male 30.97        3     no northwest 10600.548
## 1335  18 female 31.92        0     no northeast  2205.981
## 1336  18 female 36.85        0     no southeast  1629.833
## 1337  21 female 25.80        0     no southwest  2007.945
## 1338  61 female 29.07        0    yes northwest 29141.360
```

With the following code we verify the data type of the variables.

```
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

We can observe that 2 variables are of type ‘numeric’, 3 variables are of type ‘char’ and 2 variables are of type ‘integer’.

We know that the typology ‘integer’ constitutes a subclass of the typology ‘numeric’, and has the characteristic to occupy less space in memory.

At this point we verify the presence of null or missing values.

```
insurance[rowSums(is.na(insurance)) > 0, ]
```

```
## [1] age      sex      bmi      children smoker   region   charges
## <0 rows> (or 0-length row.names)
```

The result of this function does not return lines, therefore inside the dataset there are no null or missing values.

3 Descriptive Statistics

Let’s proceed with the descriptive analysis of the dataset.

```
summary(insurance)
```

```
##       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
```

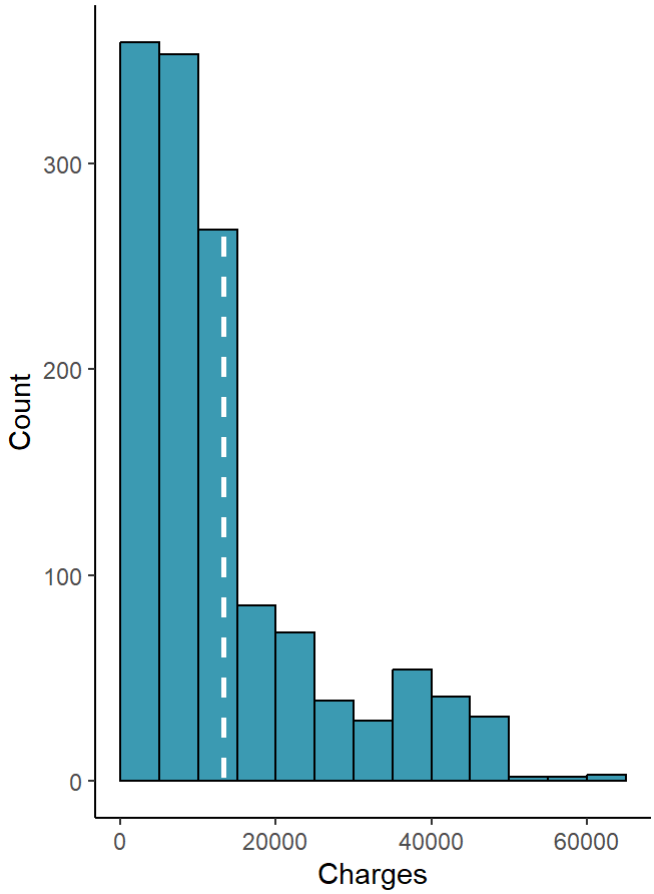
From the summary we can see some interesting features and pay attention to:

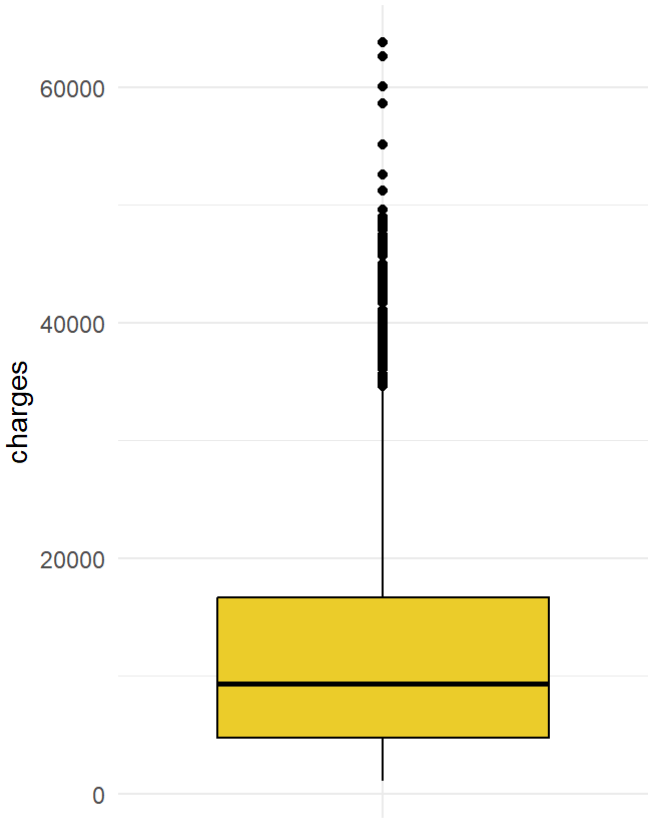
- The age of individuals varies from 18 to 64 years.
- The mean of bmi is 30.66 so we can say that the average individuals are in “Obesity class I (moderate)” class.
- Datasets individuals have an average child.
- From the charges summary we can notice that there is a big difference between min and max.

3.1 Variables

3.1.1 Charges

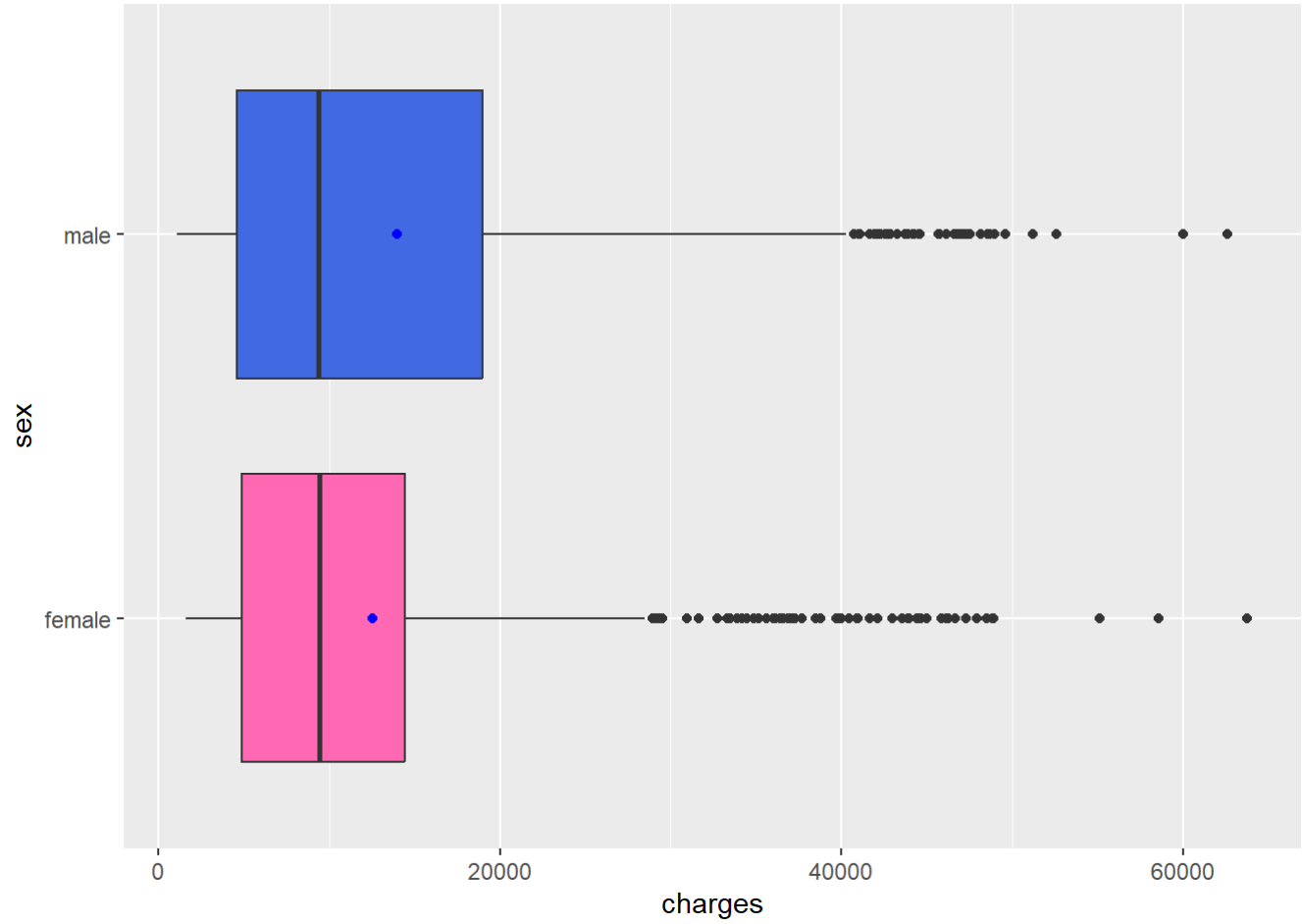
```
brx_Charges <- pretty(range(insurance$charges),
              n = nclass.Sturges(insurance$charges), min.n = 1)

Charges1 <- ggplot(insurance) +
  geom_histogram(aes(x=charges), fill = wes_palette("Zissou1")[1], color="black", breaks = br
x_Charges) +
  geom_vline(aes(xintercept=mean(charges)), color="white", linetype="dashed", linewidth=1) +
  labs(title="",x="Charges", y="Count") +
  theme_classic()

Charges2 <- ggplot(insurance, aes(x = "", y=charges)) +
  geom_boxplot(fill=wes_palette("Zissou1")[3], color="black") + labs(title = "", x = "", y =
"charges") +
  theme_minimal()

ggarrange(Charges1, Charges2,
          ncol = 2,
          nrow = 1)
```



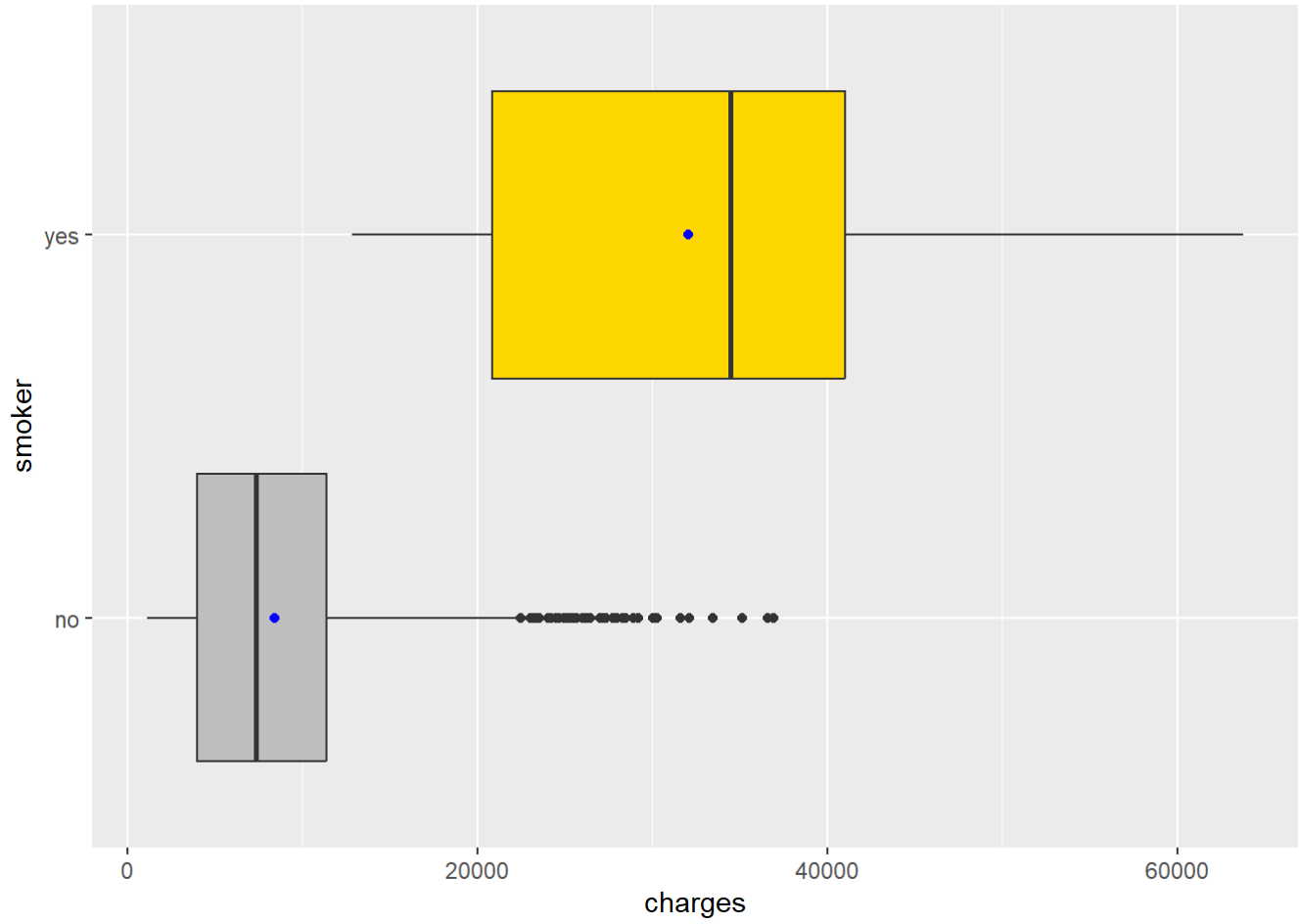



```
p2<-ggplot(insurance, aes(x=sex, y=charges)) +
geom_boxplot(fill=c('hotpink', "royalblue"))+
stat_summary(fun=mean, geom="point", color="blue")+
coord_flip()

p3<-ggplot(insurance, aes(x=smoker, y=charges)) +
geom_boxplot(fill=c('grey', "gold"))+
stat_summary(fun=mean, geom="point", color="blue")+
coord_flip()

p4<-ggplot(insurance, aes(x=region, y=charges)) +
geom_boxplot(fill=c('tan1', "tan2", 'tan3', "tan4"))+
stat_summary(fun=mean, geom="point", color="blue")+
coord_flip()

ggarrange(p2, p3, p4,
          ncol = 1,
          nrow = 1)
```

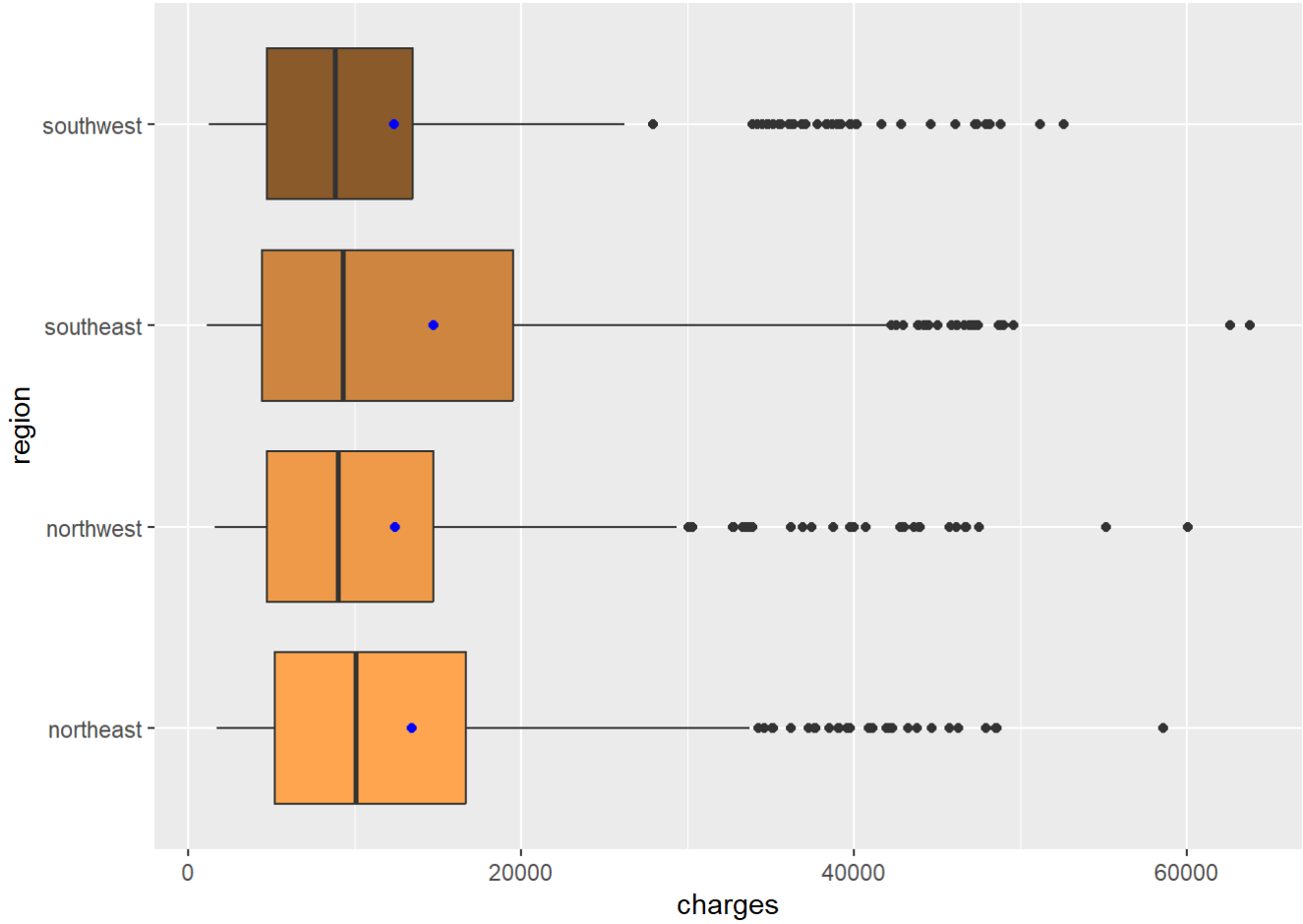
```
## $`1`
```


male
female
sex
0
20000
40000
60000
charges

```
##
## $`2`
```


yes
no
smoker
0
20000
40000
60000
charges

```
##
## $`3`
```




```
##
## attr(,"class")
## [1] "list"      "ggarrange"
```

```
ordered_Charges <- insurance[order(-insurance$charges), ]

head(ordered_Charges)
```

```
##      age    sex    bmi children smoker    region  charges
## 544   54 female 47.410        0    yes southeast 63770.43
## 1301  45   male 30.360        0    yes southeast 62592.87
## 1231  52   male 34.485        3    yes northwest 60021.40
## 578   31 female 38.095        1    yes northeast 58571.07
## 820   33 female 35.530        0    yes northwest 55135.40
## 1147  60   male 32.800        0    yes southwest 52590.83
```

```
tail(ordered_Charges)
```

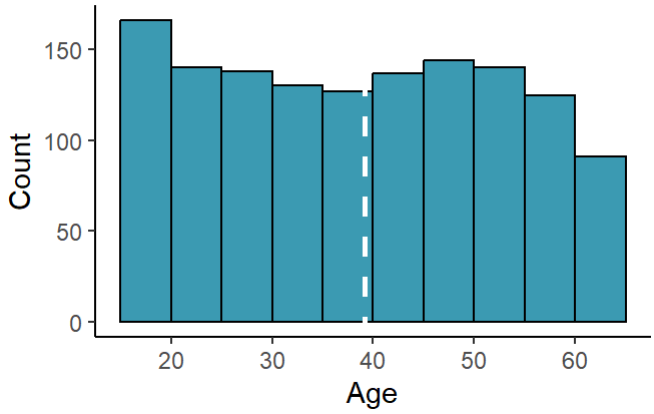
```
##      age  sex   bmi children smoker    region  charges
## 195   18 male 34.43        0     no southeast 1137.470
## 23    18 male 34.10        0     no southeast 1137.011
## 664   18 male 33.66        0     no southeast 1136.399
## 1245  18 male 33.33        0     no southeast 1135.941
## 809   18 male 30.14        0     no southeast 1131.507
## 941   18 male 23.21        0     no southeast 1121.874
```

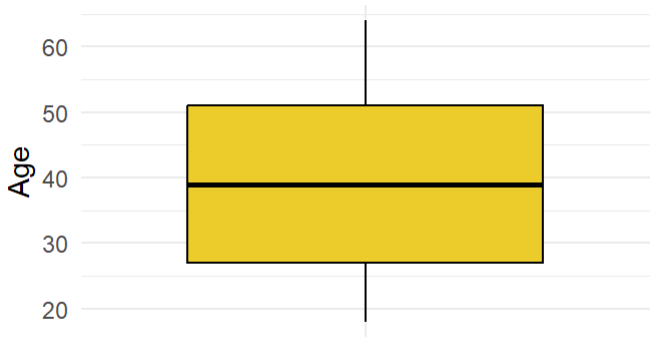
```
boxplot.stats(insurance$charges)
```

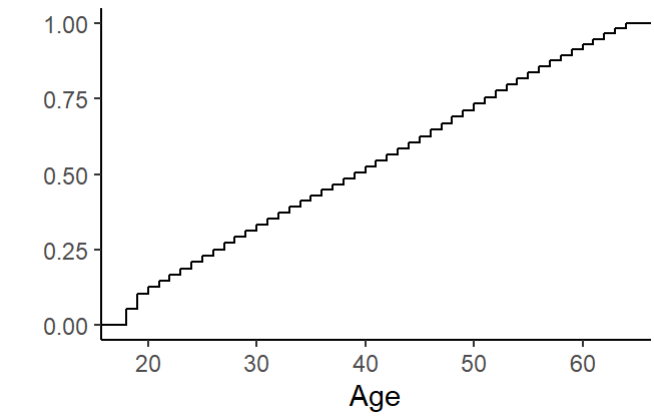
```
## $stats
## [1]  1121.874  4738.268  9382.033 16657.717 34472.841
##
## $n
## [1] 1338
##
## $conf
## [1] 8867.178 9896.888
##
## $out
##   [1] 39611.76 36837.47 37701.88 38711.00 35585.58 51194.56 39774.28 48173.36
##   [9] 38709.18 37742.58 47496.49 37165.16 39836.52 43578.94 47291.06 47055.53
##  [17] 39556.49 40720.55 36950.26 36149.48 48824.45 43753.34 37133.90 34779.61
##  [25] 38511.63 35160.13 47305.31 44260.75 41097.16 43921.18 36219.41 46151.12
##  [33] 42856.84 48549.18 47896.79 42112.24 38746.36 42124.52 34838.87 35491.64
##  [41] 42760.50 47928.03 48517.56 41919.10 36085.22 38126.25 42303.69 46889.26
##  [49] 46599.11 39125.33 37079.37 35147.53 48885.14 36197.70 38245.59 48675.52
##  [57] 63770.43 45863.21 39983.43 45702.02 58571.07 43943.88 39241.44 42969.85
##  [65] 40182.25 34617.84 42983.46 42560.43 40003.33 45710.21 46200.99 46130.53
##  [73] 40103.89 34806.47 40273.65 44400.41 40932.43 40419.02 36189.10 44585.46
##  [81] 43254.42 36307.80 38792.69 55135.40 43813.87 39597.41 36021.01 45008.96
##  [89] 37270.15 42111.66 40974.16 46113.51 46255.11 44202.65 48673.56 35069.37
##  [97] 39047.29 47462.89 38998.55 41999.52 41034.22 36580.28 35595.59 42211.14
## [105] 44423.80 37484.45 39725.52 44501.40 39727.61 48970.25 39871.70 34672.15
## [113] 41676.08 44641.20 41949.24 36124.57 38282.75 46661.44 40904.20 36898.73
## [121] 52590.83 40941.29 39722.75 37465.34 36910.61 38415.47 41661.60 60021.40
## [129] 47269.85 49577.66 37607.53 47403.88 38344.57 34828.65 62592.87 46718.16
## [137] 37829.72 36397.58 43896.38
```

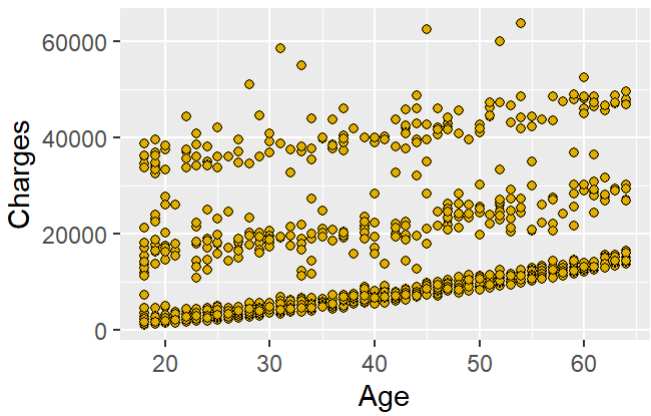
3.1.2 Age

```
brx_age <- pretty(range(insurance$age),
              n = nclass.Sturges(insurance$age), min.n = 1)

Age1 <- ggplot(insurance) +
  geom_histogram(aes(x=age), fill = wes_palette("Zissou1")[1], color="black", breaks = brx_ag
e) +
  geom_vline(aes(xintercept=mean(age)), color="white", linetype="dashed", linewidth=1) +
  labs(title="",x="Age", y="Count") +
  theme_classic()

Age2 <- ggplot(insurance, aes(x = "", y=age)) +
  geom_boxplot(fill=wes_palette("Zissou1")[3], color="black") + labs(title = "", x = "", y =
"Age") +
  theme_minimal()

Age3 <- ggplot(insurance, aes(age)) +
  stat_ecdf(geom="step") +
  labs(title="", y = "", x="Age") +
  theme_classic()

Age4 <- ggplot(insurance) +
  geom_point(aes(x=age, y=charges), shape=21, fill=wes_palette("Zissou1")[4], color="black")
 +
  labs(title="", y = "Charges", x="Age")

ggarrange(Age1, Age2, Age3, Age4,
          ncol = 2,
          nrow = 2)
```



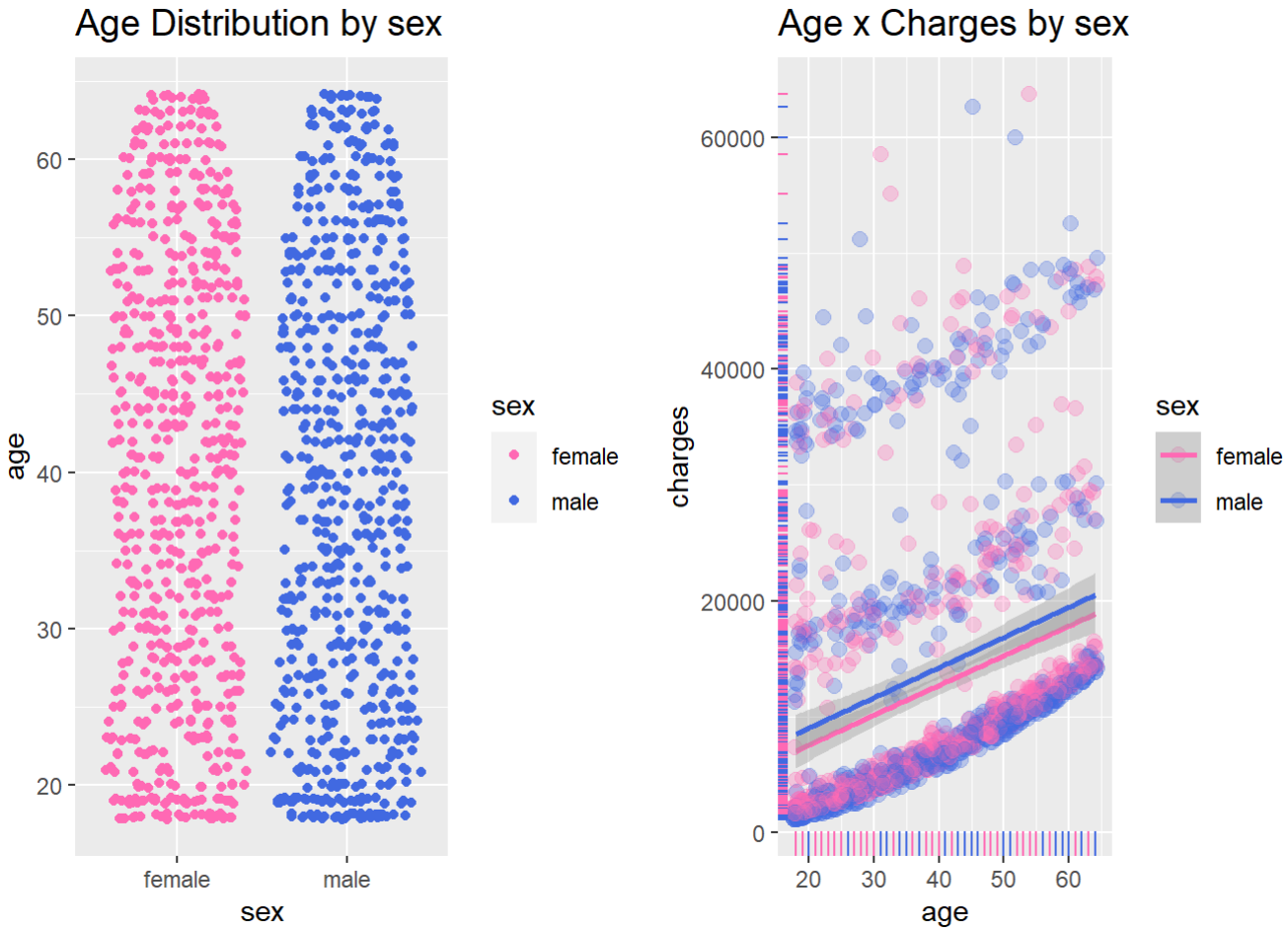






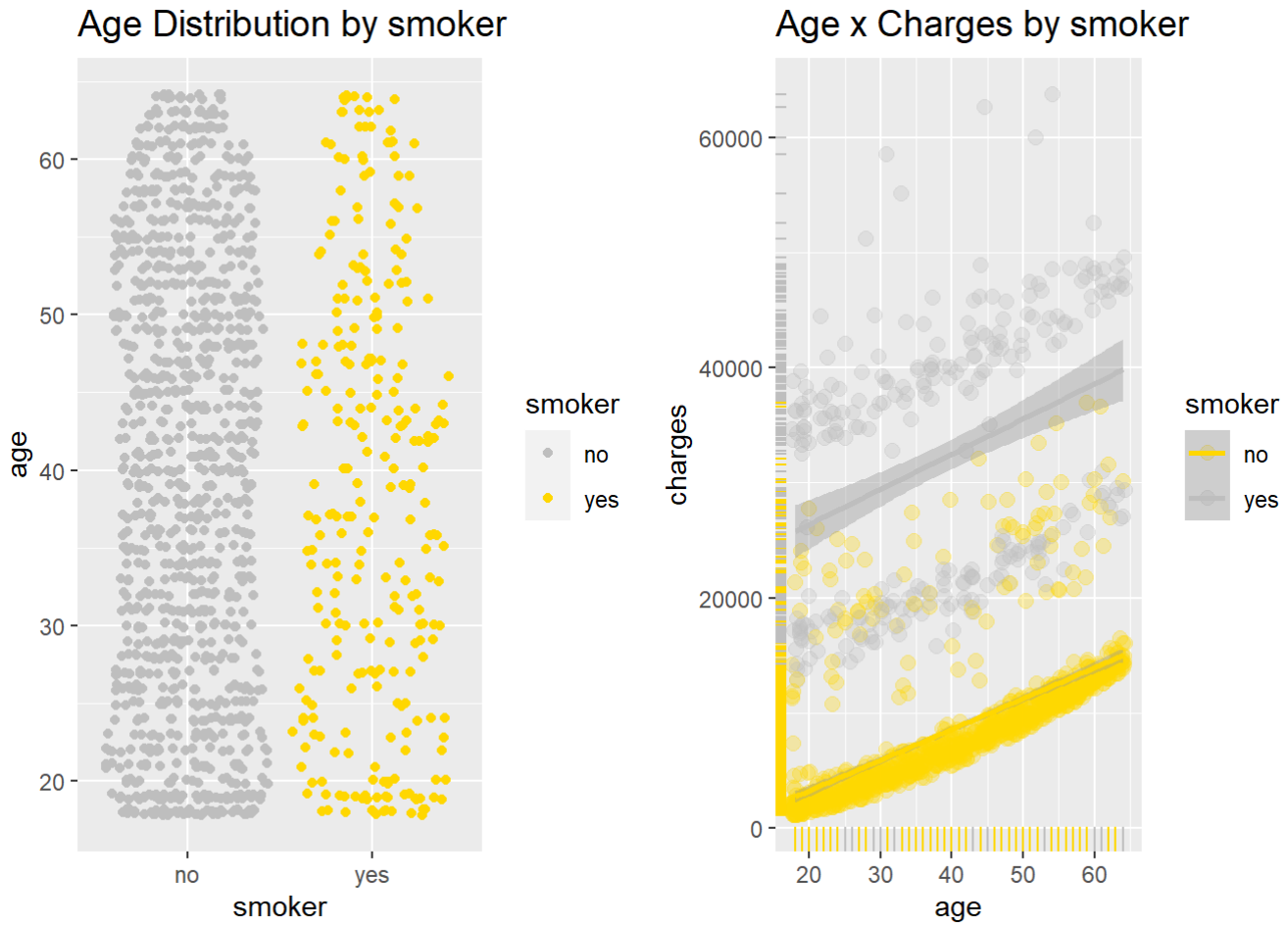
3.1.2.1 age & sex, charges

```
as1<-ggplot(insurance, aes(x=sex, y=age, color=sex)) +
geom_sina()+
scale_color_manual(values=c('hotpink', "royalblue"))+
labs(title="Age Distribution by sex")+
theme(plot.title = element_text(size=14))

as2<-ggplot(insurance, aes(x=age, y=charges, color= sex))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('hotpink', "royalblue"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Age x Charges by sex")+
theme(plot.title = element_text(size=14))

ggarrange(as1, as2,
          ncol = 2,
          nrow = 1)
```



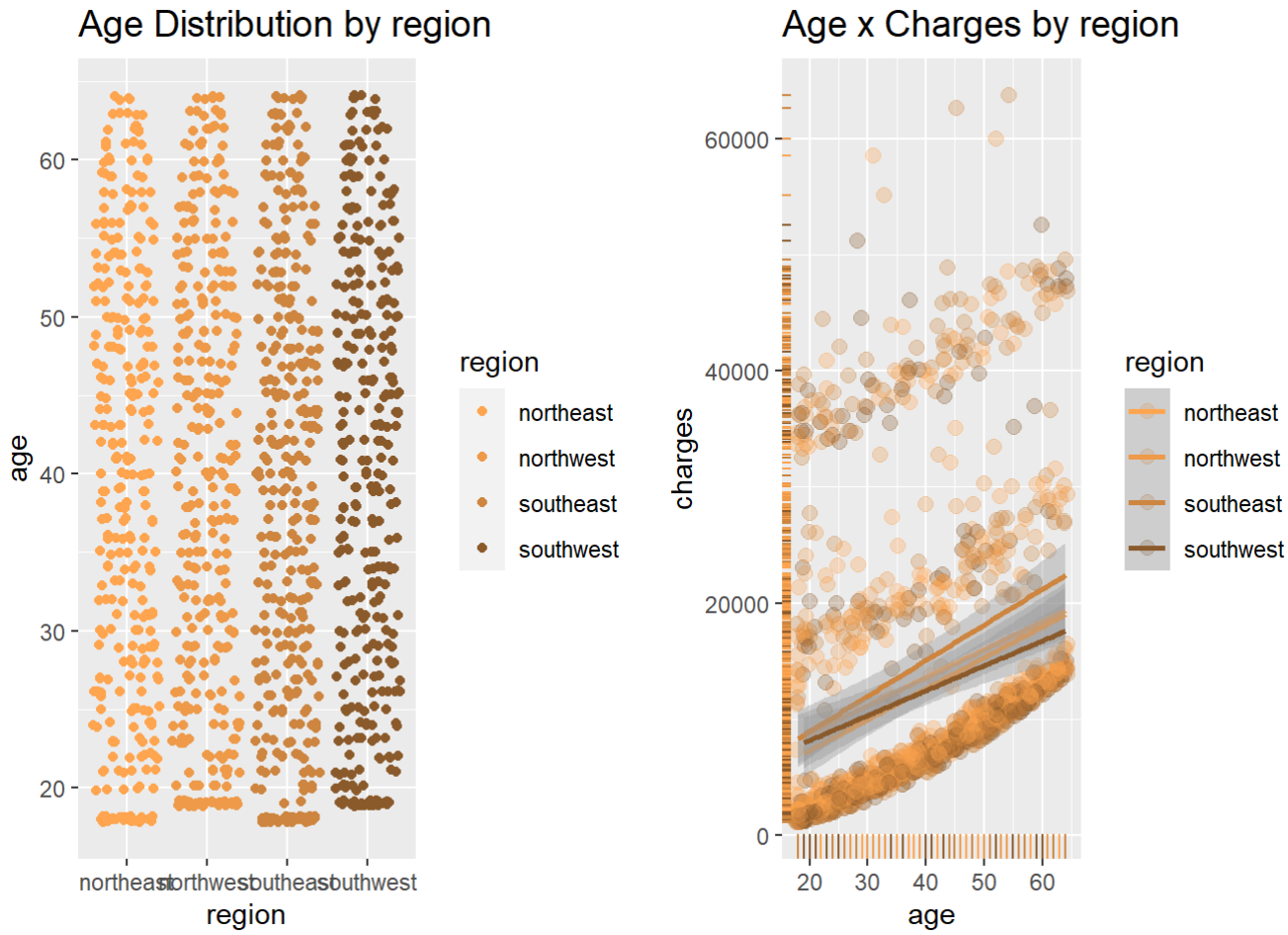
3.1.2.2 age & smoker, charges

```
ass1<-ggplot(insurance, aes(x=smoker, y=age, color=smoker)) +
geom_sina()+
scale_color_manual(values=c('grey', "gold"))+
labs(title="Age Distribution by smoker")+
theme(plot.title = element_text(size=14))

ass2<-ggplot(insurance, aes(x=age, y=charges, color= smoker))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('gold', "grey"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Age x Charges by smoker")+
theme(plot.title = element_text(size=14))

ggarrange(ass1, ass2,
          ncol = 2,
          nrow = 1)
```



3.1.2.3 age & region, charges

```
ar1<-ggplot(insurance, aes(x=region, y=age, color=region)) +
geom_sina()+
scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
labs(title="Age Distribution by region")+
theme(plot.title = element_text(size=14))

ar2<-ggplot(insurance, aes(x=age, y=charges, color= region))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Age x Charges by region")+
theme(plot.title = element_text(size=14))

ggarrange(ar1, ar2,
          ncol = 2,
          nrow = 1)
```



3.1.2.4 Conclusion age

```
ordered_Age <- insurance[order(-insurance$age), ]

head(ordered_Age)
```

```
##     age    sex    bmi children smoker    region  charges
## 63   64   male 24.700        1     no northwest 30166.62
## 95   64 female 31.300        2    yes southwest 47291.06
## 200  64 female 39.330        0     no northeast 14901.52
## 329  64 female 33.800        1    yes southwest 47928.03
## 336  64   male 34.500        0     no southwest 13822.80
## 379  64 female 30.115        3     no northwest 16455.71
```

```
tail(ordered_Age)
```

```
##      age    sex    bmi children smoker    region   charges
## 1284  18   male 30.030        1     no southeast  1720.354
## 1297  18   male 26.125        0     no northeast  1708.926
## 1316  18   male 28.310        1     no northeast 11272.331
## 1318  18   male 53.130        0     no southeast  1163.463
## 1335  18 female 31.920        0     no northeast  2205.981
## 1336  18 female 36.850        0     no southeast  1629.833
```

```
boxplot.stats(insurance$age)
```

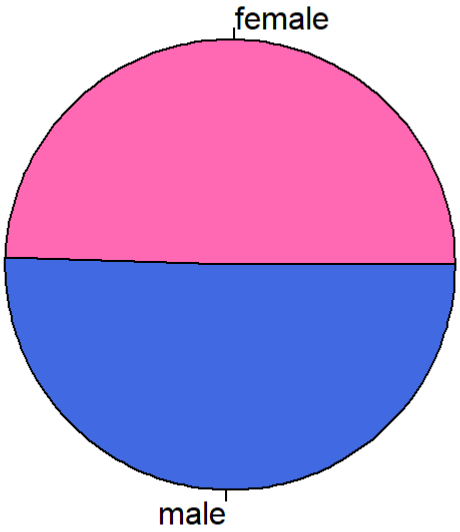
```
## $stats
## [1] 18 27 39 51 64
##
## $n
## [1] 1338
##
## $conf
## [1] 37.96333 40.03667
##
## $out
## integer(0)
```

3.1.3 Sex

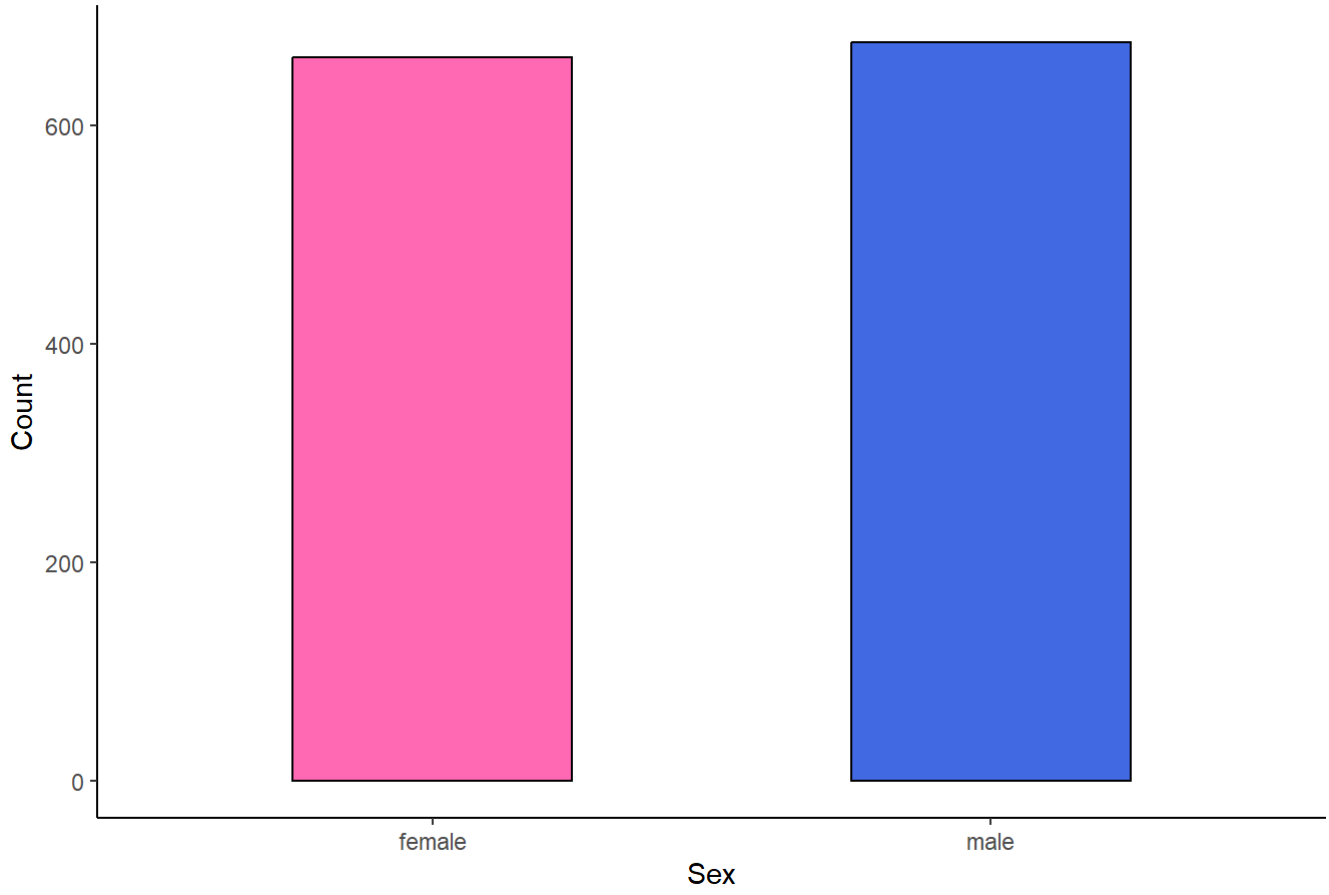
```
mode = function(){
    return(sort(-table(insurance$sex))[1])
}

mode()
```

```
## male
## -676
```

```
sex = insurance$sex

sex.freq = table(sex)

pie(sex.freq, col=c('hotpink', "royalblue"))
```




```
SexBar <- ggplot(insurance) +
  geom_bar(aes(x=sex), width = 0.5, fill=c('hotpink', "royalblue"), color="black") +
  labs(title="",x="Sex", y="Count") +
  theme_classic()

ggarrange(SexBar)
```



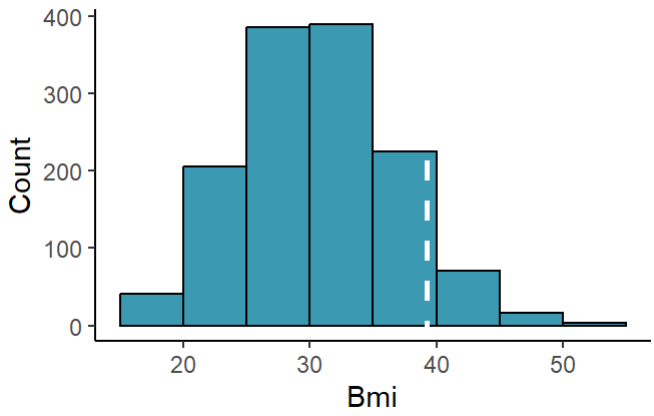

```
table(insurance$sex)
```

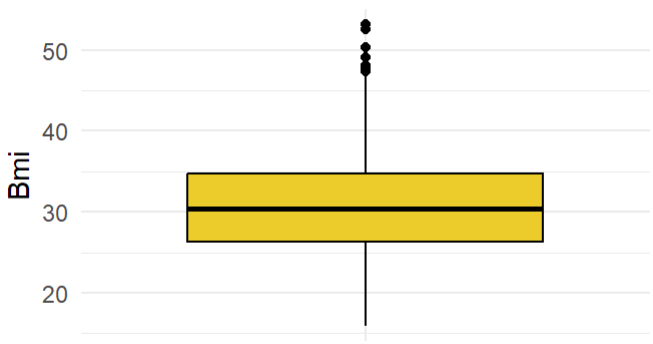
```
##
## female   male
##    662    676
```

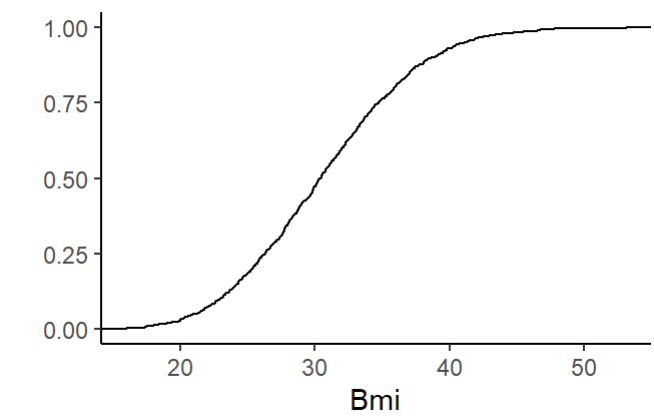
- There are 662 females and 676 males.

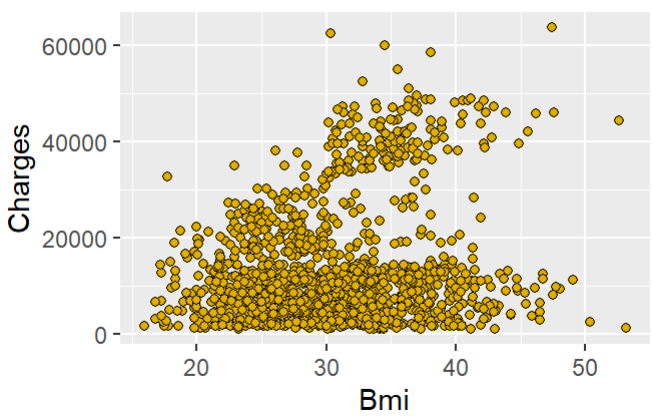
3.1.4 Bmi

```
brx_bmi <- pretty(range(insurance$bmi),
              n = nclass.Sturges(insurance$bmi), min.n = 1)

Bmi1 <- ggplot(insurance) +
  geom_histogram(aes(x=bmi), fill = wes_palette("Zissou1")[1], color="black", breaks = brx_bm
i) +
  geom_vline(aes(xintercept=mean(age)), color="white", linetype="dashed", linewidth=1) +
  labs(title="",x="Bmi", y="Count") +
  theme_classic()

Bmi2 <- ggplot(insurance, aes(x = "", y=bmi)) +
  geom_boxplot(fill=wes_palette("Zissou1")[3], color="black") + labs(title = "", x = "", y =
"Bmi") +
  theme_minimal()

Bmi3 <- ggplot(insurance, aes(bmi)) +
  stat_ecdf(geom="step") +
  labs(title="", y = "", x="Bmi") +
  theme_classic()

Bmi4 <- ggplot(insurance) +
  geom_point(aes(x=bmi, y=charges), shape=21, fill=wes_palette("Zissou1")[4], color="black")
 +
  labs(title="", y = "Charges", x="Bmi")

ggarrange(Bmi1, Bmi2, Bmi3, Bmi4,
          ncol = 2,
          nrow = 2)
```



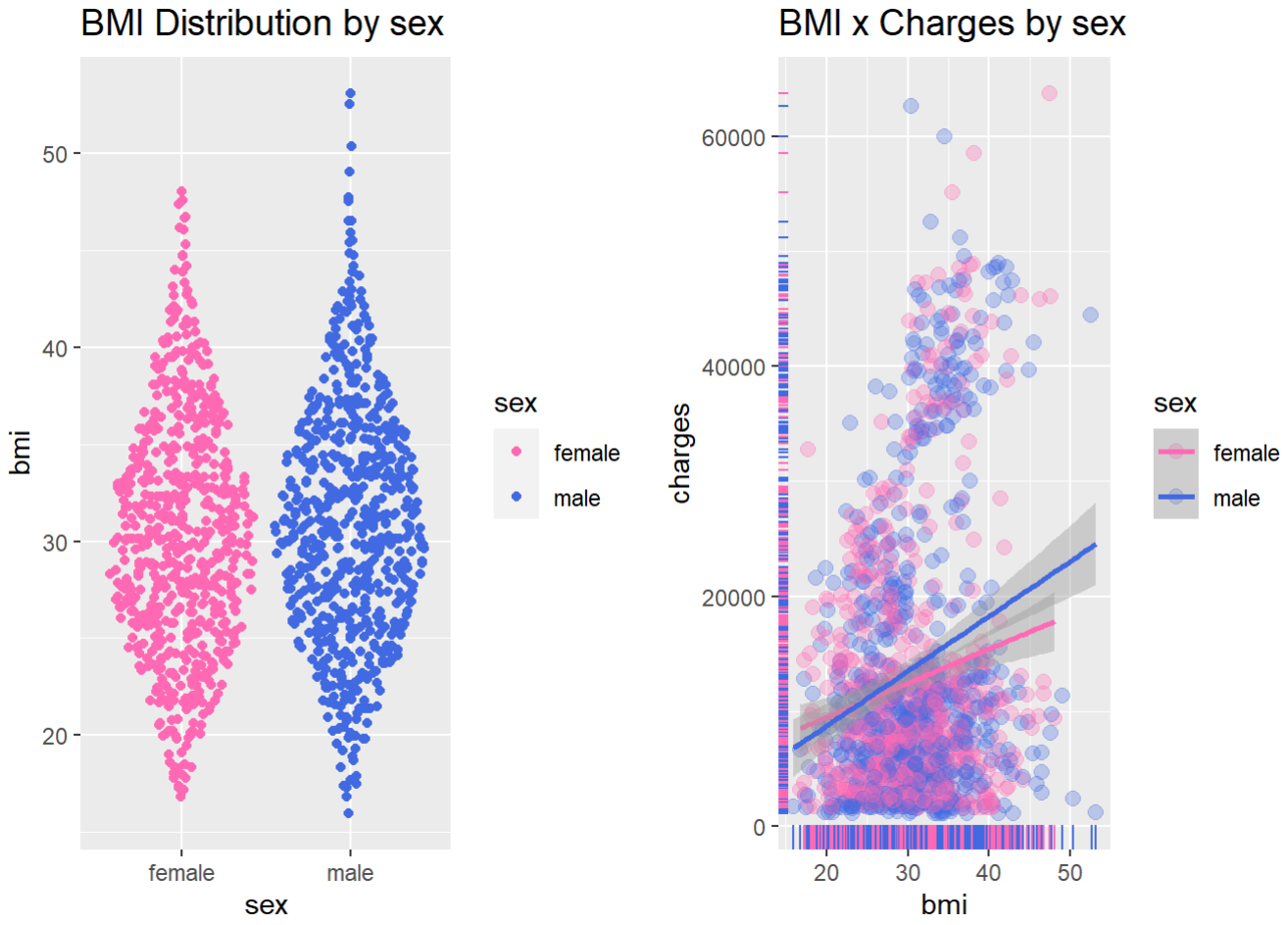






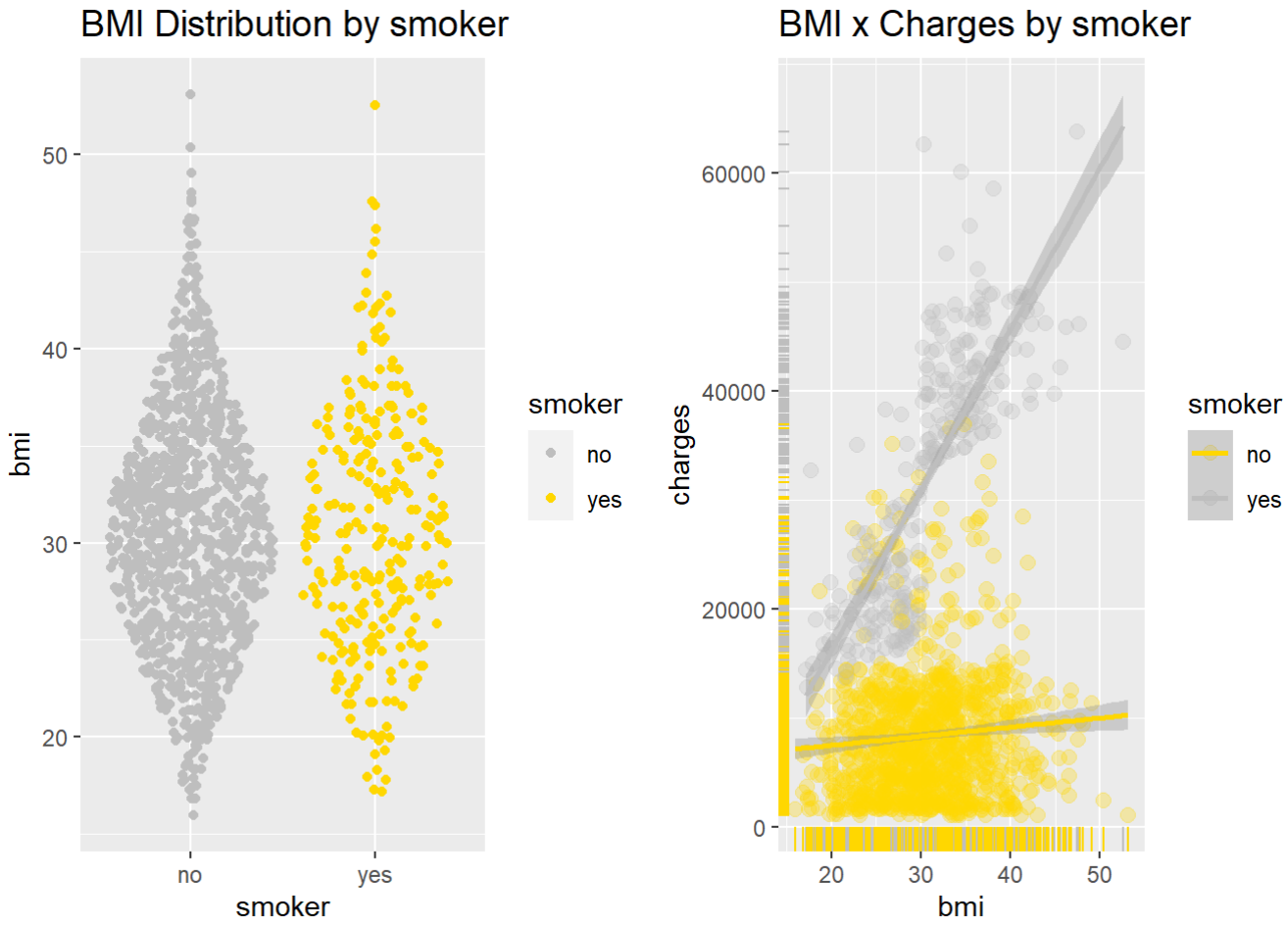
3.1.4.1 bmi & sex, charges

```
bs1<-ggplot(insurance, aes(x=sex, y=bmi, color=sex)) +
geom_sina()+
scale_color_manual(values=c('hotpink', "royalblue"))+
labs(title="BMI Distribution by sex")+
theme(plot.title = element_text(size=14))

bs2<-ggplot(insurance, aes(x=bmi, y=charges, color= sex))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('hotpink', "royalblue"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="BMI x Charges by sex")+
theme(plot.title = element_text(size=14))

ggarrange(bs1, bs2,
          ncol = 2,
          nrow = 1)
```



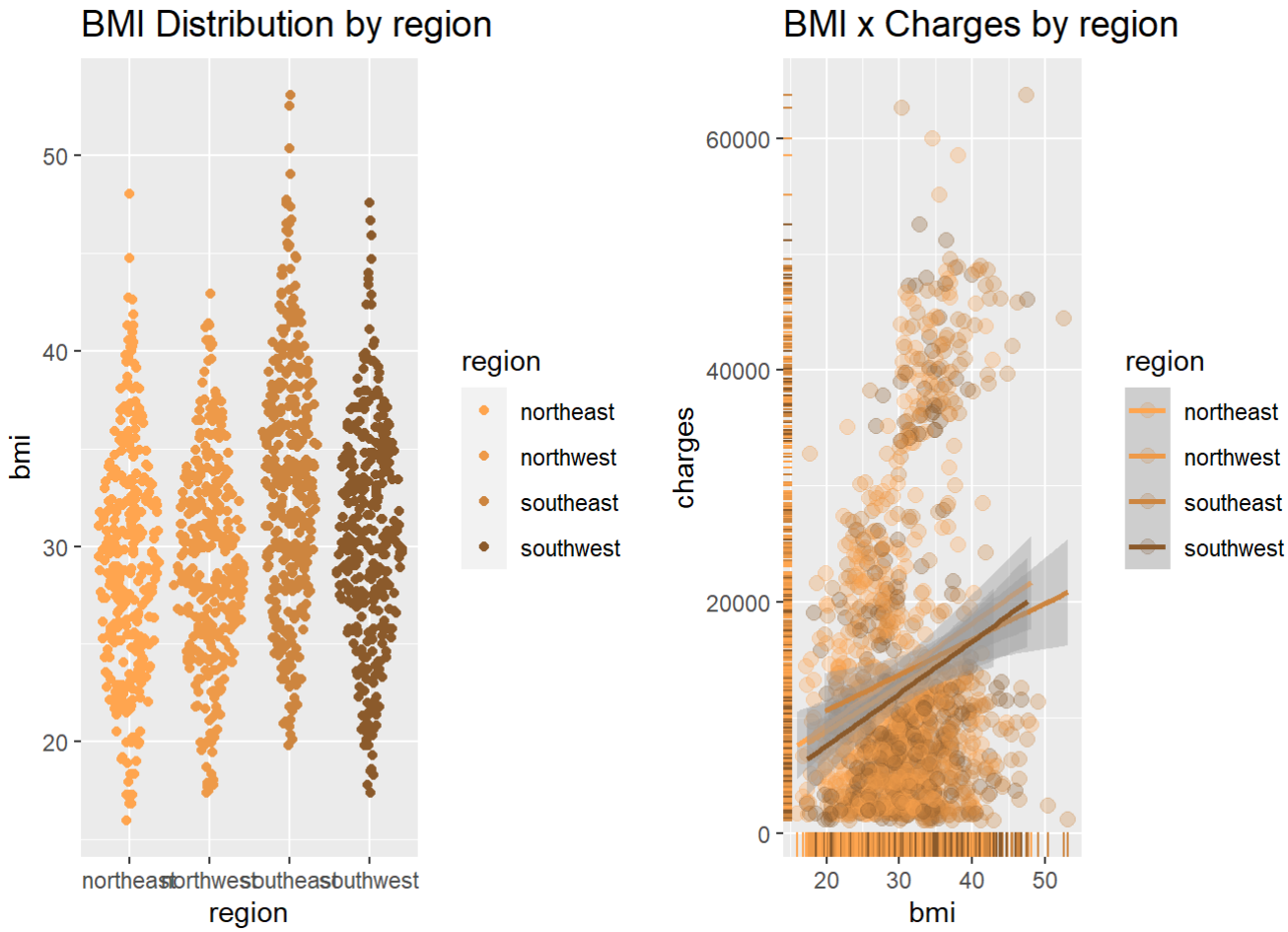
3.1.4.2 bmi & smoker, charges

```
bss1<-ggplot(insurance, aes(x=smoker, y=bmi, color=smoker)) +
geom_sina()+
scale_color_manual(values=c('grey', "gold"))+
labs(title="BMI Distribution by smoker")+
theme(plot.title = element_text(size=14))

bss2<-ggplot(insurance, aes(x=bmi, y=charges, color= smoker))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('gold', "grey"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="BMI x Charges by smoker")+
theme(plot.title = element_text(size=14))

ggarrange(bss1, bss2,
          ncol = 2,
          nrow = 1)
```



3.1.4.3 bmi & region, charges

```
br1<-ggplot(insurance, aes(x=region, y=bmi, color=region)) +
geom_sina()+
scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
labs(title="BMI Distribution by region")+
theme(plot.title = element_text(size=14))

br2<-ggplot(insurance, aes(x=bmi, y=charges, color= region))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="BMI x Charges by region")+
theme(plot.title = element_text(size=14))

ggarrange(br1, br2,
          ncol = 2,
          nrow = 1)
```



3.1.4.4 Conclusion bmi

```
ordered_Bmi <- insurance[order(-insurance$bmi), ]

head(ordered_Bmi)
```

```
##      age    sex   bmi children smoker    region   charges
## 1318  18   male 53.13        0     no southeast  1163.463
## 1048  22   male 52.58        1    yes southeast 44501.398
## 848   23   male 50.38        1     no southeast  2438.055
## 117   58   male 49.06        0     no southeast 11381.325
## 287   46 female 48.07        2     no northeast  9432.925
## 1089  52   male 47.74        1     no southeast  9748.911
```

```
tail(ordered_Bmi)
```

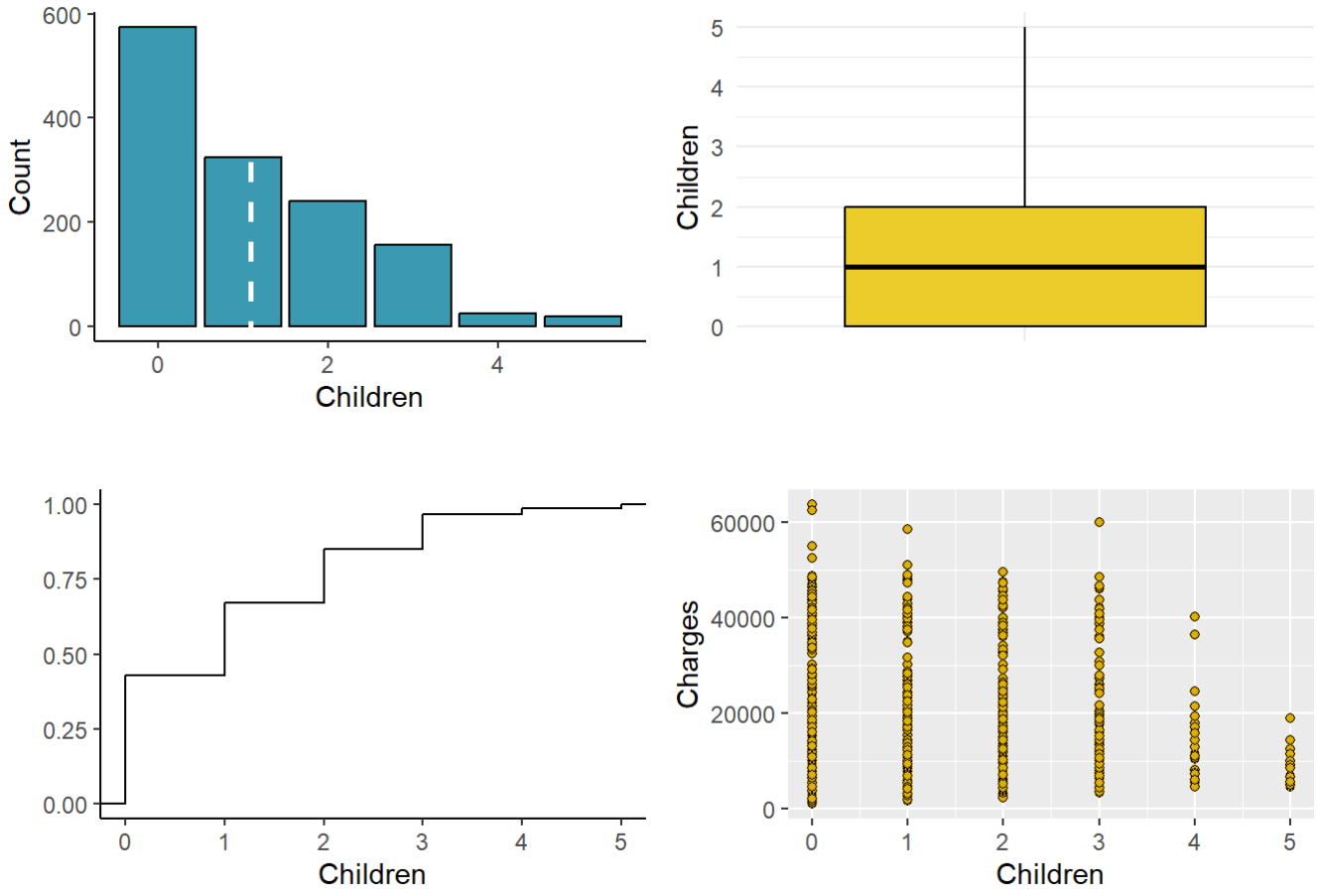
```
##      age    sex    bmi children smoker    region   charges
## 1030  37 female 17.290        2     no northeast  6877.980
## 1287  28 female 17.290        0     no northeast  3732.625
## 413   26 female 17.195        2    yes northeast 14455.644
## 429   21 female 16.815        1     no northeast  3167.456
## 1227  38   male 16.815        2     no northeast  6640.545
## 173   18   male 15.960        0     no northeast  1694.796
```

```
boxplot.stats(insurance$bmi)
```

```
## $stats
## [1] 15.96 26.29 30.40 34.70 46.75
##
## $n
## [1] 1338
##
## $conf
## [1] 30.03673 30.76327
##
## $out
## [1] 49.06 48.07 47.52 47.41 50.38 47.60 52.58 47.74 53.13
```

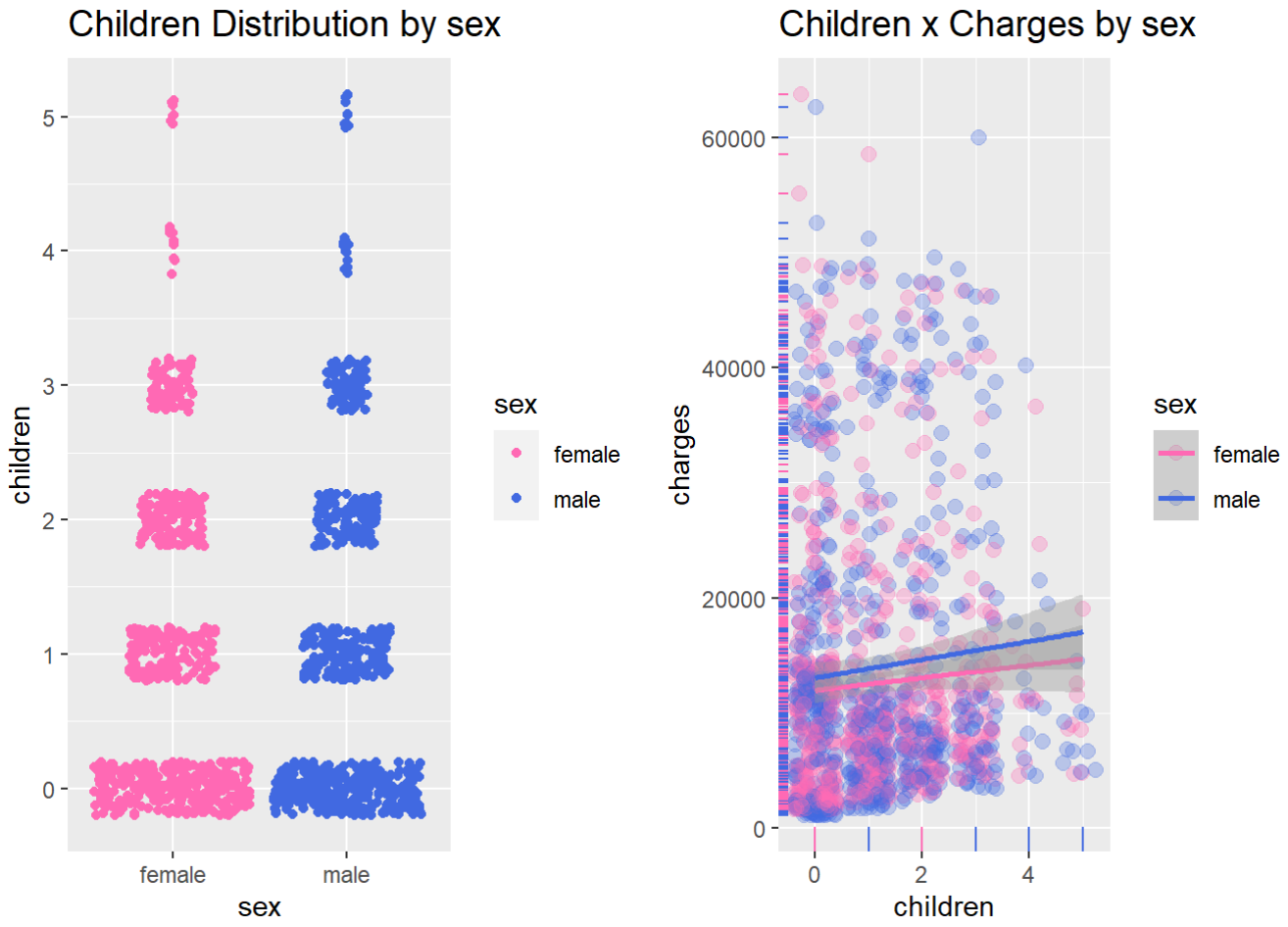
3.1.5 Children

```
Child1 <- ggplot(insurance) +
  geom_bar(aes(x=children), fill = wes_palette("Zissou1")[1], color="black") +
  geom_vline(aes(xintercept=mean(children)), color="white", linetype="dashed", linewidth=1) +
  labs(title="",x="Children", y="Count") +
  theme_classic()

Child2 <- ggplot(insurance, aes(x = "", y=children)) +
  geom_boxplot(fill=wes_palette("Zissou1")[3], color="black") + labs(title = "", x = "", y =
"Children") +
  theme_minimal()

Child3 <- ggplot(insurance, aes(children)) +
  stat_ecdf(geom="step") +
  labs(title="", y = "", x="Children") +
  theme_classic()

Child4 <- ggplot(insurance) +
  geom_point(aes(x=children, y=charges), shape=21, fill=wes_palette("Zissou1")[4], color="bla
ck") +
  labs(title="", y = "Charges", x="Children")

ggarrange(Child1, Child2, Child3, Child4,
          ncol = 2,
          nrow = 2)
```



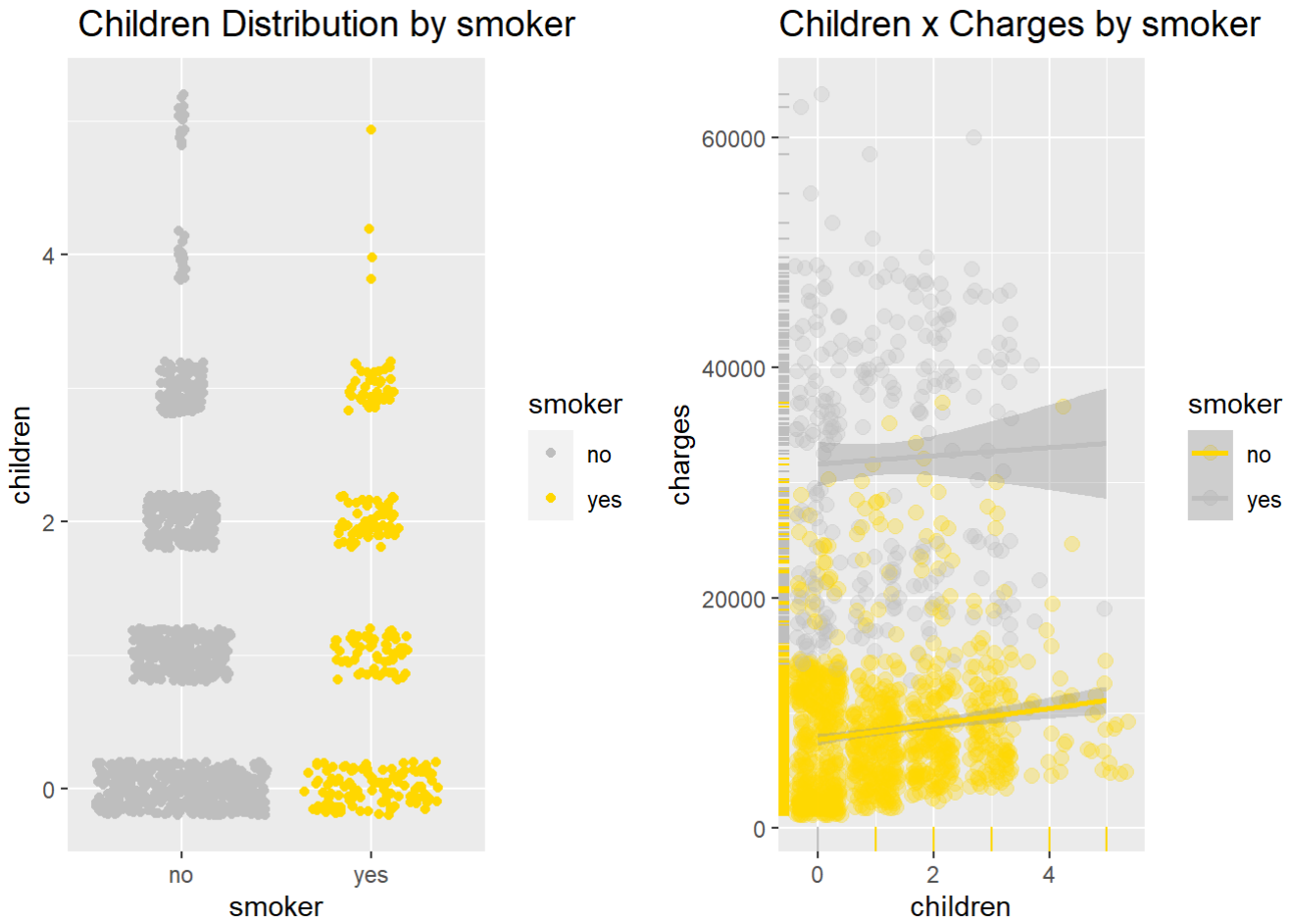
3.1.5.1 children & sex, charges

```
cs1<-ggplot(insurance, aes(x=sex, y=children, color=sex)) +
geom_sina()+
scale_color_manual(values=c('hotpink', "royalblue"))+
labs(title="Children Distribution by sex")+
theme(plot.title = element_text(size=14))

cs2<-ggplot(insurance, aes(x=children, y=charges, color= sex))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('hotpink', "royalblue"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Children x Charges by sex")+
theme(plot.title = element_text(size=14))

ggarrange(cs1, cs2,
          ncol = 2,
          nrow = 1)
```



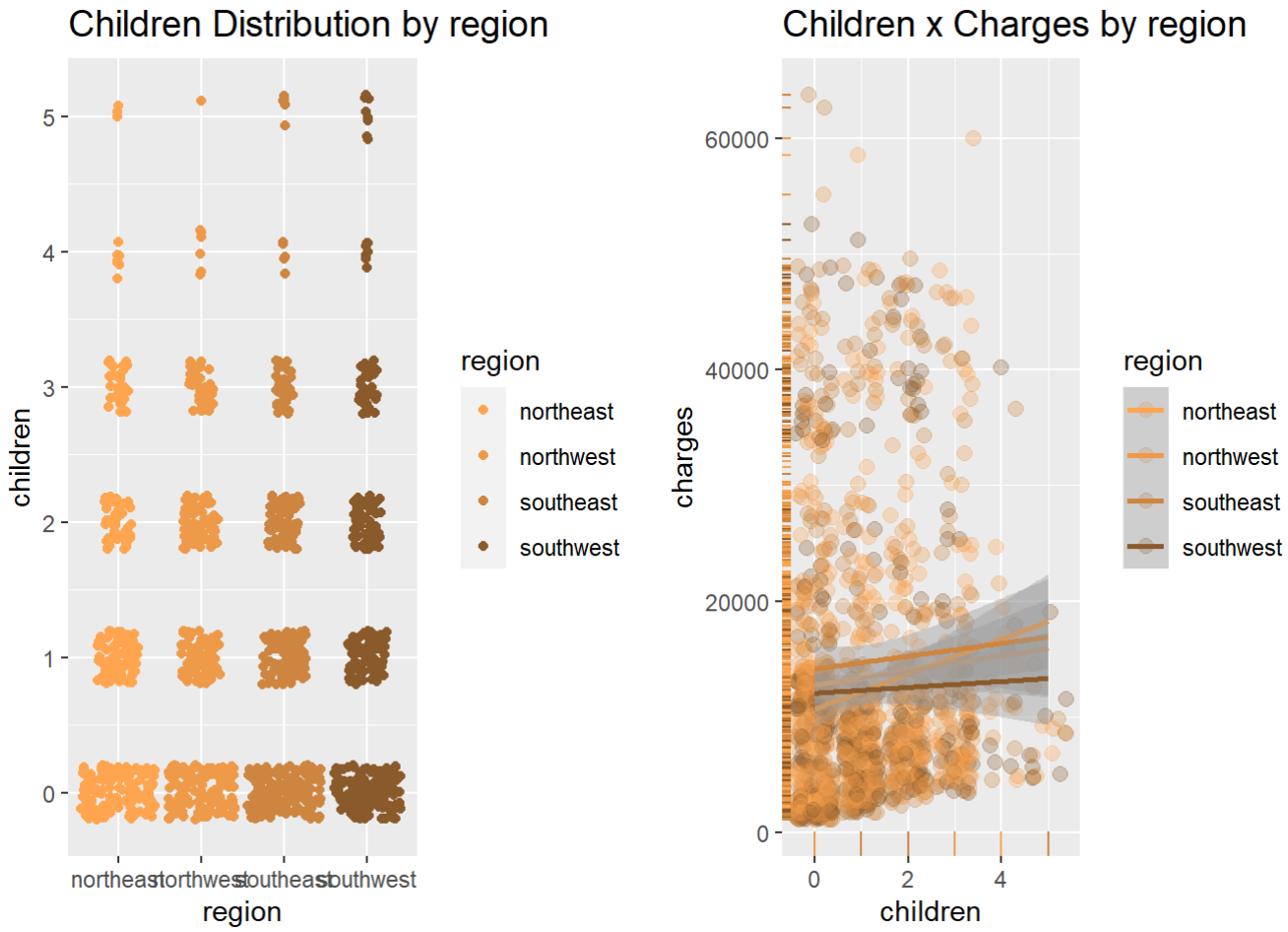
3.1.5.2 children & smoker, charges

```
css1<-ggplot(insurance, aes(x=smoker, y=children, color=smoker)) +
geom_sina()+
scale_color_manual(values=c('grey', "gold"))+
labs(title=" Children Distribution by smoker")+
theme(plot.title = element_text(size=14))

css2<-ggplot(insurance, aes(x=children, y=charges, color= smoker))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('gold', "grey"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Children x Charges by smoker")+
theme(plot.title = element_text(size=14))

ggarrange(css1, css2,
          ncol = 2,
          nrow = 1)
```



3.1.5.3 children & region, charges

```
cr1<-ggplot(insurance, aes(x=region, y=children, color=region)) +
geom_sina()+
scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
labs(title="Children Distribution by region")+
theme(plot.title = element_text(size=14))

cr2<-ggplot(insurance, aes(x=children, y=charges, color= region))+
geom_jitter(alpha=0.3, size=2.5)+
scale_color_manual(values=c('tan1', "tan2", 'tan3', "tan4"))+
geom_rug()+
geom_smooth(method=lm, formula=y~x)+
labs(title="Children x Charges by region")+
theme(plot.title = element_text(size=14))

ggarrange(cr1, cr2,
          ncol = 2,
          nrow = 1)
```



Function for mode computation

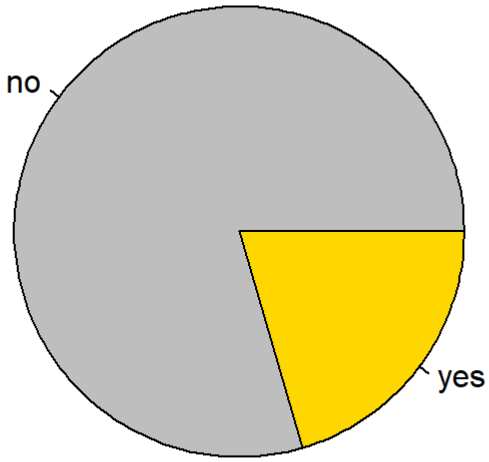
```
mode <- function(att){
    return(sort(-table(att))[1])
}
```

3.1.6 Smoker

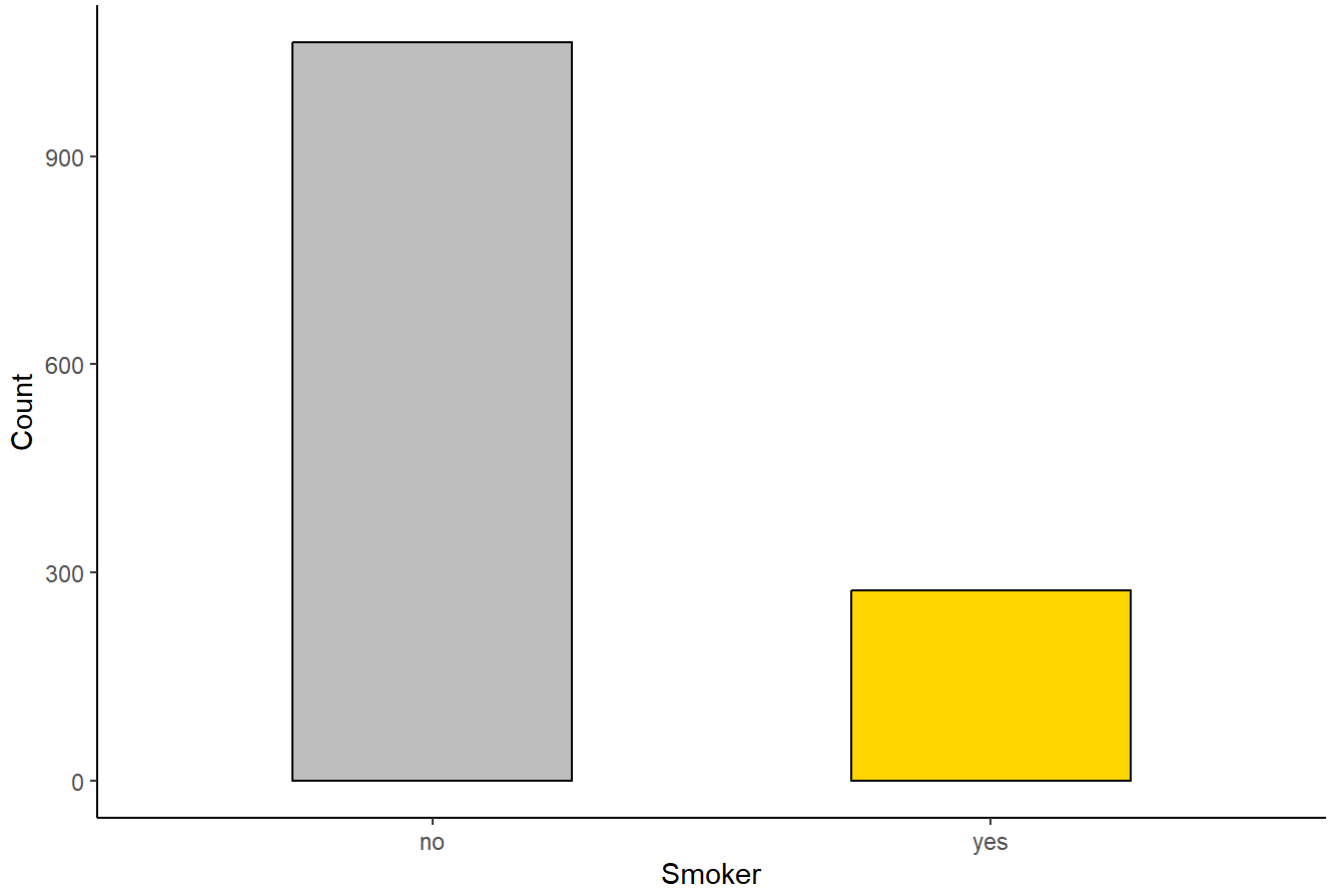
```
mode(insurance$smoker)
```

```
##    no
## -1064
```

```
smoker = insurance$smoker

smoker.freq = table(smoker)

pie(smoker.freq, col=c('grey', "gold"))
```




```
SmokerBar <- ggplot(insurance) +
  geom_bar(aes(x=smoker), width = 0.5, fill = c('grey', "gold"), color="black") +
  labs(title="",x="Smoker", y="Count") +
  theme_classic()

ggarrange(SmokerBar)
```




```
table(insurance$smoker)
```

```
##
##   no  yes
## 1064  274
```

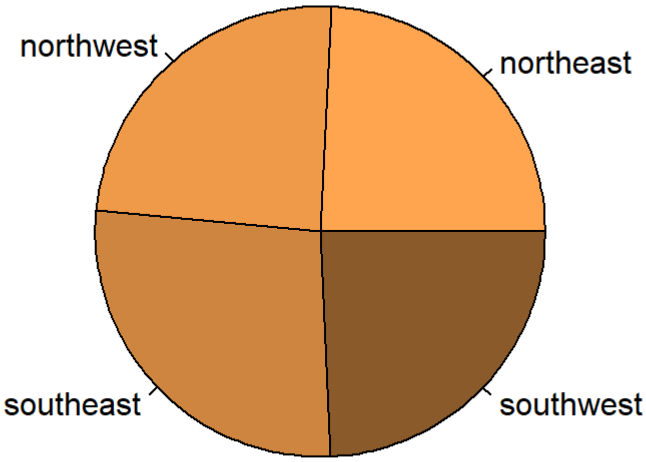
- 274 individals smoke and 1064 individuals doesn't smoke.

3.1.7 Region

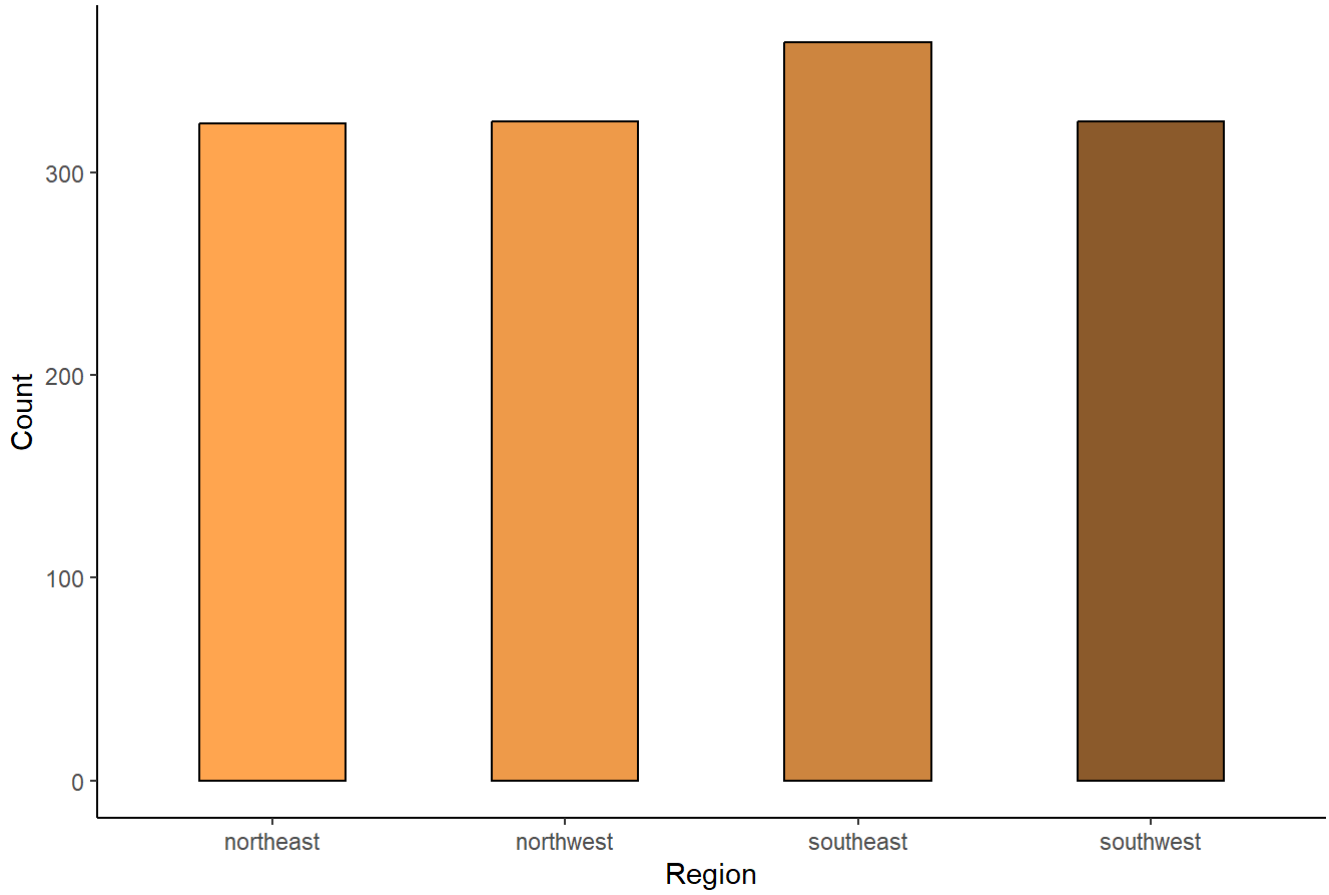
```
mode(insurance$region)
```

```
## southeast
##      -364
```

```
region = insurance$region

region.freq = table(region)

pie(region.freq, col = c('tan1', "tan2", 'tan3', "tan4"))
```




```
RegionBar <- ggplot(insurance) +
  geom_bar(aes(x=region), width = 0.5, fill=c('tan1', "tan2", 'tan3', "tan4"), color="black")
+
  labs(title="",x="Region", y="Count") +
  theme_classic()

ggarrange(RegionBar)
```



4 Tests

Is it the mean equal to 30?

```
t.test(insurance$bmi, mu = 30,
       alternative="two.sided") # (one-tailed test)
```

```
##
##  One Sample t-test
##
## data:  insurance$bmi
## t = 3.9792, df = 1337, p-value = 7.284e-05
## alternative hypothesis: true mean is not equal to 30
## 95 percent confidence interval:
##  30.33635 30.99045
## sample estimates:
## mean of x
##   30.6634
```

The null hypothesis is mean = 30. The p-value is less than a significance level of 5%. So we can reject the null hipothesis of mean = 30, and this means that the mean bmi is significantly different from 30.

Shapiro test to see if it's normally distributed.

```
shapiro.test(insurance$bmi)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insurance$bmi
## W = 0.99389, p-value = 2.605e-05
```

It is not normally distributed, so we use Wilcoxon test

```
wilcox.test(insurance$bmi, mu = 30,
            alternative="two.sided") # assuming the distribution is simetricaly aroud median
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  insurance$bmi
## V = 489736, p-value = 0.002205
## alternative hypothesis: true location is not equal to 30
```

With p-value is less then 5%, we can reject the null hypothesis of mean = 30. The mean is statistical different from 30.

Is the male bmi variance equal to the female bmi variance?

```
var.test(insurance[which(insurance$sex=="male"), "bmi"],
         insurance[which(insurance$sex=="female"), "bmi"])
```

```
##
##  F test to compare two variances
##
## data:  insurance[which(insurance$sex == "male"), "bmi"] and insurance[which(insurance$sex
== "female"), "bmi"]
## F = 1.0315, num df = 675, denom df = 661, p-value = 0.6892
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8861438 1.2004937
## sample estimates:
## ratio of variances
##           1.031475
```

The null hypothesis is same variance. The p-values greater then the significance level of 0.05 allows us to accept the null hipothesis of two normal samples with same variance.

t-test

```
t.test(insurance[which(insurance$sex=="male"), "bmi"],
       insurance[which(insurance$sex=="female"), "bmi"],
       alternative="two.sided",
       var.equal= TRUE)
```

```
##
##  Two Sample t-test
##
## data:  insurance[which(insurance$sex == "male"), "bmi"] and insurance[which(insurance$sex
== "female"), "bmi"]
## t = 1.6968, df = 1336, p-value = 0.08998
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08829755  1.21905646
## sample estimates:
## mean of x mean of y
##  30.94313  30.37775
```

p-value greater than 0.05 so we can accept the null hypothesis of equal average bmi (not significantly different) between men and women.

Shapiro Test for variable male

```
shapiro.test(insurance[which(insurance$sex=="male"), "bmi"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insurance[which(insurance$sex == "male"), "bmi"]
## W = 0.99305, p-value = 0.003133
```

Shapiro Test for variable male

```
shapiro.test(insurance[which(insurance$sex=="female"), "bmi"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insurance[which(insurance$sex == "female"), "bmi"]
## W = 0.99303, p-value = 0.003543
```

The p-value smaller then 0.05 for both cases so we reject the null hipothesis of normally distribution.

Wilcoxon test

```
wilcox.test(insurance[which(insurance$sex=="male"), "bmi"],
            insurance[which(insurance$sex=="female"), "bmi"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  insurance[which(insurance$sex == "male"), "bmi"] and insurance[which(insurance$sex
== "female"), "bmi"]
## W = 235332, p-value = 0.1014
## alternative hypothesis: true location shift is not equal to 0
```

p-value greater then 5 % allow us to accept the null hipothesis of median bmi equal (not significantly diferent) for man and women.

4.1 Anova Test

Test of the hypothesis used to determine whether there is a significant statistical difference between the averages of three or more groups of continuous data with respect to a category that differentiates them.

Smoker

```
anova<-aov(bmi ~ smoker,  data=insurance)
summary(anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## smoker         1      1    0.70   0.019  0.891
## Residuals   1336  49720   37.22
```

The p-value is 0.891, so we can't reject the null hypothesis of equal means.

```
shapiro.test(x=residuals(object = anova))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(object = anova)
## W = 0.9939, p-value = 2.68e-05
```

Kruskal Wallis

The Kruskal–Wallis test by ranks or one-way ANOVA on ranks is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. The parametric equivalent of the Kruskal–Wallis test is the one-way analysis of variance (ANOVA).

```
kruskal.test(bmi ~ smoker,  data=insurance)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  bmi by smoker
## Kruskal-Wallis chi-squared = 0.0064906, df = 1, p-value = 0.9358
```

The p-value is greater then 0.05 so we can conclude that there aren't significant differences between the treatment groups.

Region

```
anova2<-aov(bmi ~ region,  data=insurance)
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## region         3   4056  1352.0   39.49 <2e-16 ***
## Residuals   1334  45664    34.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Low p-value, so we can reject the null hypothesis of equal means.

```
shapiro.test(x=residuals(object = anova2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(object = anova2)
## W = 0.99549, p-value = 0.0005119
```

```
kruskal.test(bmi ~ region,  data=insurance)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  bmi by region
## Kruskal-Wallis chi-squared = 94.689, df = 3, p-value < 2.2e-16
```

4.2 Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related.

```
cor(insurance$charges, insurance$bmi)
```

```
## [1] 0.198341
```

There is no correlation between bmi and charges.

```
cor(insurance$charges, insurance$age)
```

```
## [1] 0.2990082
```

There is a low correlation between age and charges.

```
cor(insurance$charges, insurance$children)
```

```
## [1] 0.06799823
```

There is no correlation between children and charges

```
aggregate(cbind(age, bmi, charges) ~ region + smoker,
  data = insurance,
  mean
)
```

```
##      region smoker      age      bmi   charges
## 1 northeast     no 39.53696 29.33208  9165.532
## 2 northwest     no 39.16854 29.21268  8556.464
## 3 southeast     no 38.67399 33.44242  8032.216
## 4 southwest     no 40.18352 30.50787  8019.285
## 5 northeast    yes 38.23881 28.56522 29673.536
## 6 northwest    yes 39.32759 29.14043 30192.003
## 7 southeast    yes 39.73626 33.09670 34844.997
## 8 southwest    yes 36.10345 31.00517 32269.063
```

```
shapiro.test(insurance$charges)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insurance$charges
## W = 0.81469, p-value < 2.2e-16
```

```
shapiro.test(insurance$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  insurance$age
## W = 0.9447, p-value < 2.2e-16
```

Low p-values so we reject the null hypothesis of normality.

```
cor.test(insurance$charges, insurance$age, method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  insurance$charges and insurance$age
## z = 25.758, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.4753024
```

The low p-value would let us believe on a significant correlation. 0.47 can be considered a weak correlation.

5 Regression

A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables). You make this kind of relationships in your head all the time, for example when you calculate the age of a child based on her height, you are assuming the older she is, the taller she will be.

```
set.seed(703) #set seed (value) where value specifies the initial value of the random number
 seed.

training <- createDataPartition(insurance$charges, times=1, p=0.75, list=FALSE)

data_regr<-insurance[training,]
data_test<-insurance[-training,]

regr <- lm(charges ~ age+sex+bmi+children+smoker, data = data_regr)

summary(regr)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker, data = data_regr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11437  -2952  -1091   1345  29837
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11979.14    1115.72 -10.737  < 2e-16 ***
## age            257.12      13.99  18.380  < 2e-16 ***
## sexmale         58.07     389.10   0.149  0.88140
## bmi            315.78      31.72   9.955  < 2e-16 ***
## children       480.55     158.84   3.025  0.00255 **
## smokeryes    23519.69     478.62  49.141  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6144 on 1000 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7374
## F-statistic: 565.3 on 5 and 1000 DF,  p-value: < 2.2e-16
```

```
regr_pred<-predict(regr, data_test)

RMSE(regr_pred, data_test$charges)
```

```
## [1] 5852.901
```

```
R2(regr_pred, data_test$charges)
```

```
## [1] 0.7822309
```

P-value of F-statistics is very low, so at least one predictor is really related to the outcome. Besides the rmse of 5,853 and the r squared of 78%. Let’s pay atttention to the coefficients (column Estimate) and their significance levels: The intercept is -11979.14, and almost all the predictors (except sex) are significant, according to the p-values.

The interpretation of categorical variables, for example, “smoker”, can be done like this: ” the average charge increases 23519.69 if the individual smokes - with all other variables constant.

The coefficient value, when significant, is the average change on outcome with a unit increase on a predictor - with the others constant. While the correlation measures the force on relationship, the coefficient quantifies the relationship and allows predictions with an equation.

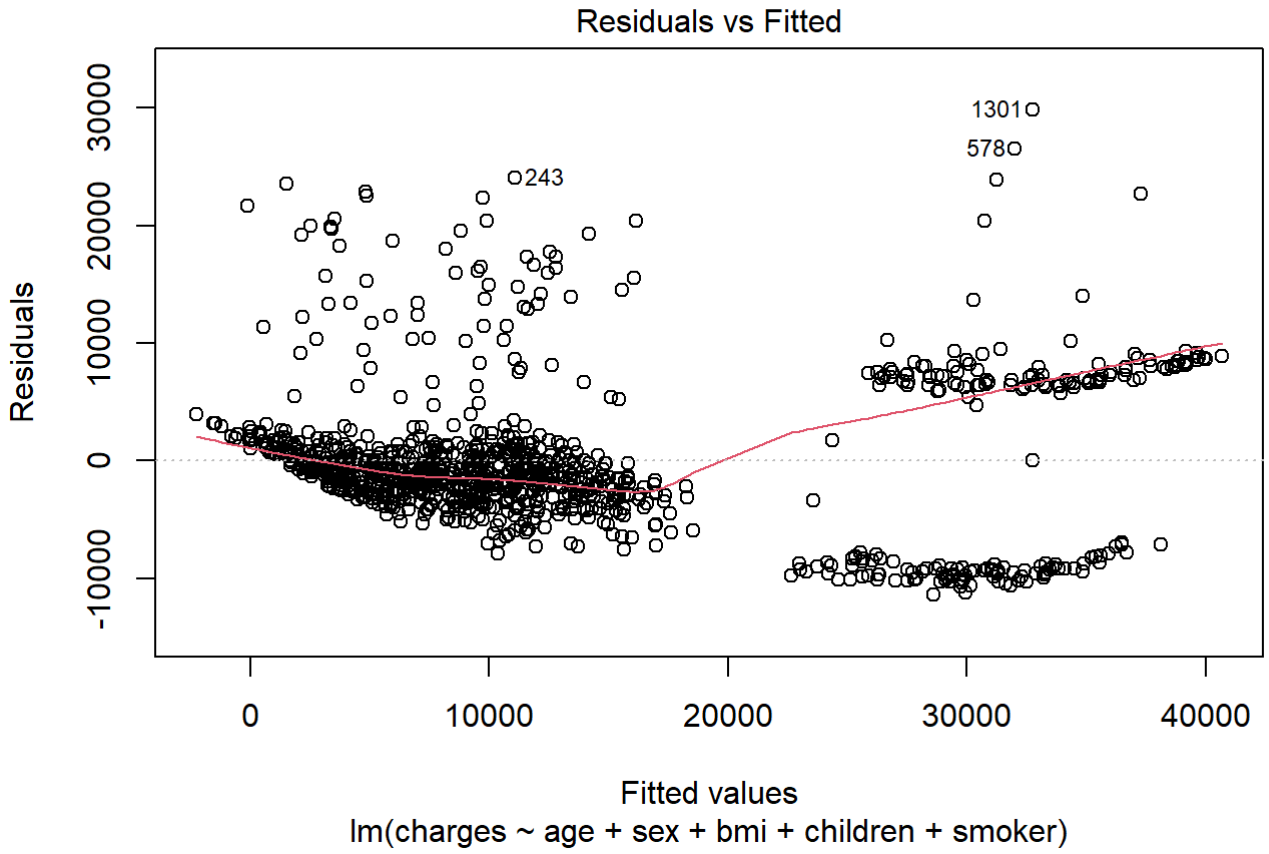
Here, for example, for each additional unit in age, the expected average cost is 257.12 higher, after controlling the others variables.

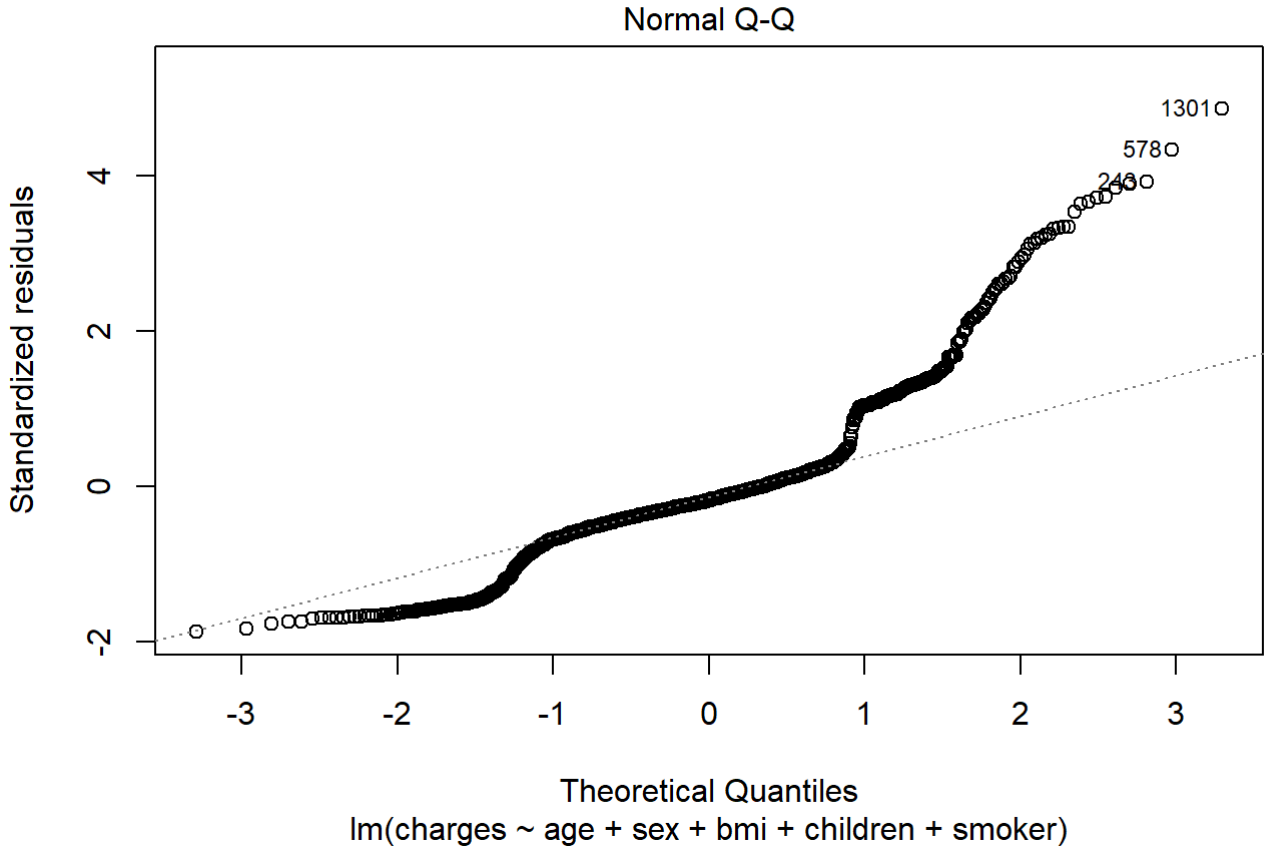
5.0.1 Assumptions of Linear Regression

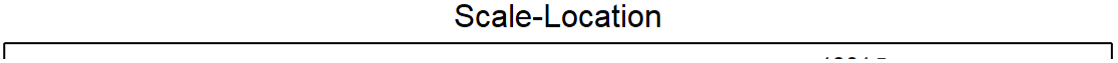
This method assumes that there is a linear relationship between predictors and outcome. The parameters will be estimated in order to give the best-fitted line, minimizing the sum of squared residuals (Ordinary Least Squares method). Linear regression has the advantage of interpretability, but there are some important checks to make.

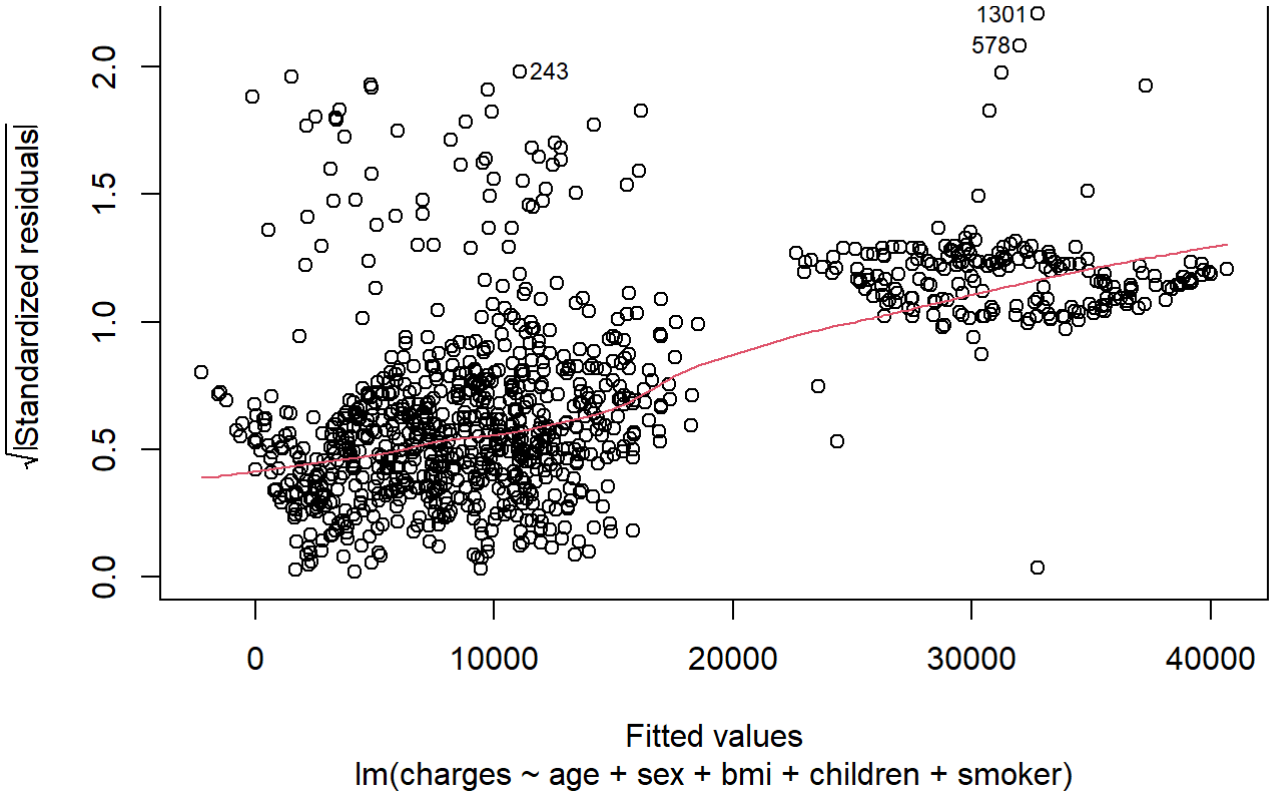
Let's check some assumptions of linear regression: - Linearity of data, linear relationship between x and y: Residuals x Fitted plot - Normality of residuals: Normal QQ plot - Homogeneity of residuals variance (homoscedasticity), residuals with constant variance: scale location plot

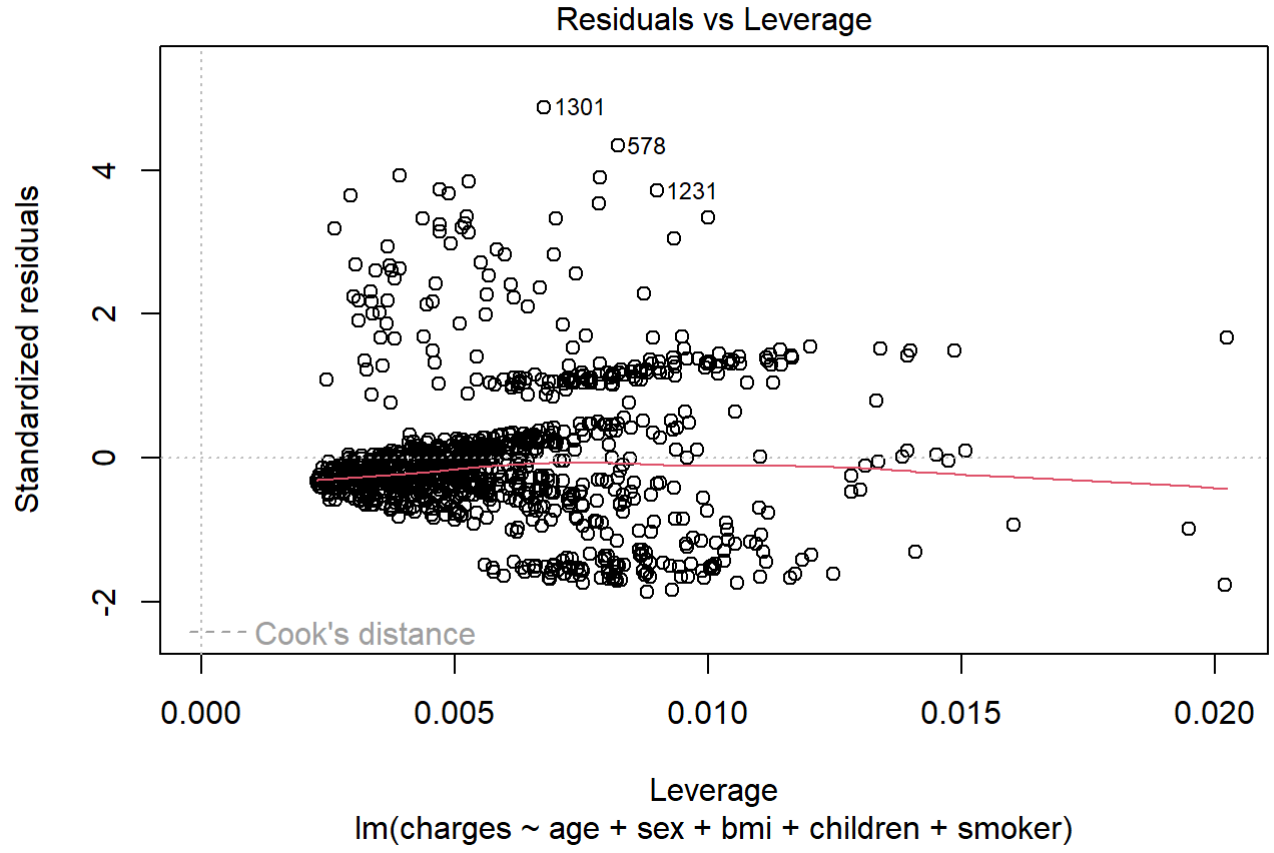
```
options(repr.plot.width=20, repr.plot.height=20)

par(mfrow=c(1,1))

plot(regr)
```


Residuals vs Fitted
Residuals
Fitted values
lm(charges ~ age + sex + bmi + children + smoker)
1301
578
243
-10000
0
10000
20000
30000
40000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ age + sex + bmi + children + smoker)
1301
578
-2
0
2
4
-3
-2
-1
0
1
2
3


Scale-Location





At Residuals x Fitted plot, the non horizontal line may indicate a non-linear relationship.

The Normal Q-Q plot is used to verify the normal distribution of the pattern residues. On the x-axis we find theoretical quantiles, on the y-axis we find standardized residues.We see that the residuals are not exactly on the straight line, indicating that they are not normally distributed.

Scale-Location plot allows to display on the x-axis the values of the model regression and on the y axis the square root of standardized residues. The non straight line indicates heteroscedasticity.

To investigate the presence of outliers in the residues we analyze the graph Residuals vs Leverage. Leverage scores are between 0 and 1. A high leverage score is therefore close to 1. In our case the residues are between leverage scores between 0.00 and 0.02. From the graph we observe how residues are contained in the range -2 to 4.

5.0.2 Durbin Watson Test

We use the Durbin Watson coefficient to check for autocorrelation between residues. This coefficient is between 0 and 4. Values close to 2 indicate that there is no autocorrelation.

```
durbinWatsonTest(regr)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.05048323      2.099114    0.11
##  Alternative hypothesis: rho != 0
```

Analyzing the test results we observe that the value of the Durbin-Watson coefficient is close to 2, that indicates the autocorrelation non-presence. To confirm this, the p-value is higher than the threshold value of 0.05; which does not allow us to reject the hypothesis that the residues are unrelated.

6 Conclusion

At this point, we can consider the analysis finished. Thanks to the development of this analysis. we explored which variables such as: age, sex, children, smoking, bmi index affect for price of insurance in the region where data was collected.

For each attribute we did basic statistical analysis which help us understand relationship and effect caused on insurance charges. We notice that if an individual smoke and have an high bmi index, charges increase as well. Other consideration is that, aging and smoking affects the increase of charges.

During this analysis we built a regression model. The goal of linear regression is to find an equation to show the mathematical relationship between predictors and outcome. As a result, the accuracy of this method is quite low, this indicates that the attributes of the dataset do not correlate linearly, but have a non-linear dependence. therefore, it will be appropriate to use non-linear models for further research.