Medical Charges: Data Analysis

Group - 7

25/11/2021

tinytex::install_tinytex

We begin with importing the ggplot2 and ggpubr packages required for data visualization. We also read the data from the .csv file containing insurance data.

```
library(ggplot2)
library(ggpubr)
library(corrplot)
```

```
## corrplot 0.91 loaded
```

```
insurance = read.csv("insurance.csv")
```

Let us now see the structure and summary of the given dataset.

```
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 ...
```

```
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  
```

Checking for Missing Values in the Data

```
f=function(x){any(is.na(x))}
check.na=apply(insurance,2,f);
check.na
```

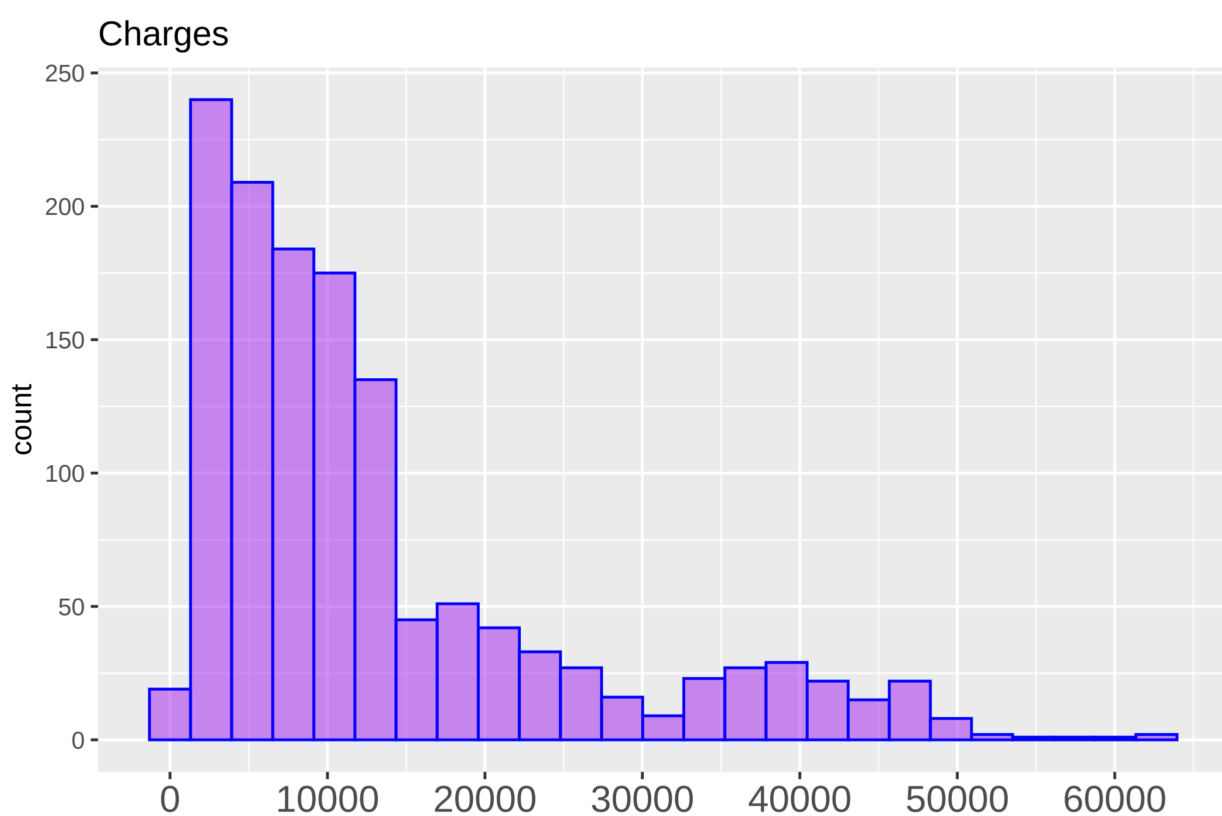
```
##      age      sex      bmi children   smoker   region  charges
##    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
```

As we can see in the above output, no column in the Insurance dataset contains empty value. Hence, there are no missing values in the dataset.

Exploratory Data Analysis

Clearly, charges is the output variable in this dataset. It gives us the insurance amount for a resident (a row of input variables in the dataset). Let us examine the distribution of charges using histogram.

```
insurance %>%
ggplot(aes(charges)) +
geom_histogram(color = "blue", fill = "purple",alpha = .5, bins = 25) +
scale_x_continuous(breaks = seq(0,66000,10000)) +
theme(axis.text.x = element_text(size = 14)) +
labs(title="Charges", x="")
```



The histogram plot shows that the distribution of charges is right-skewed. Let us confirm it by calculating the skewness of charges.

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 4.1.1
```

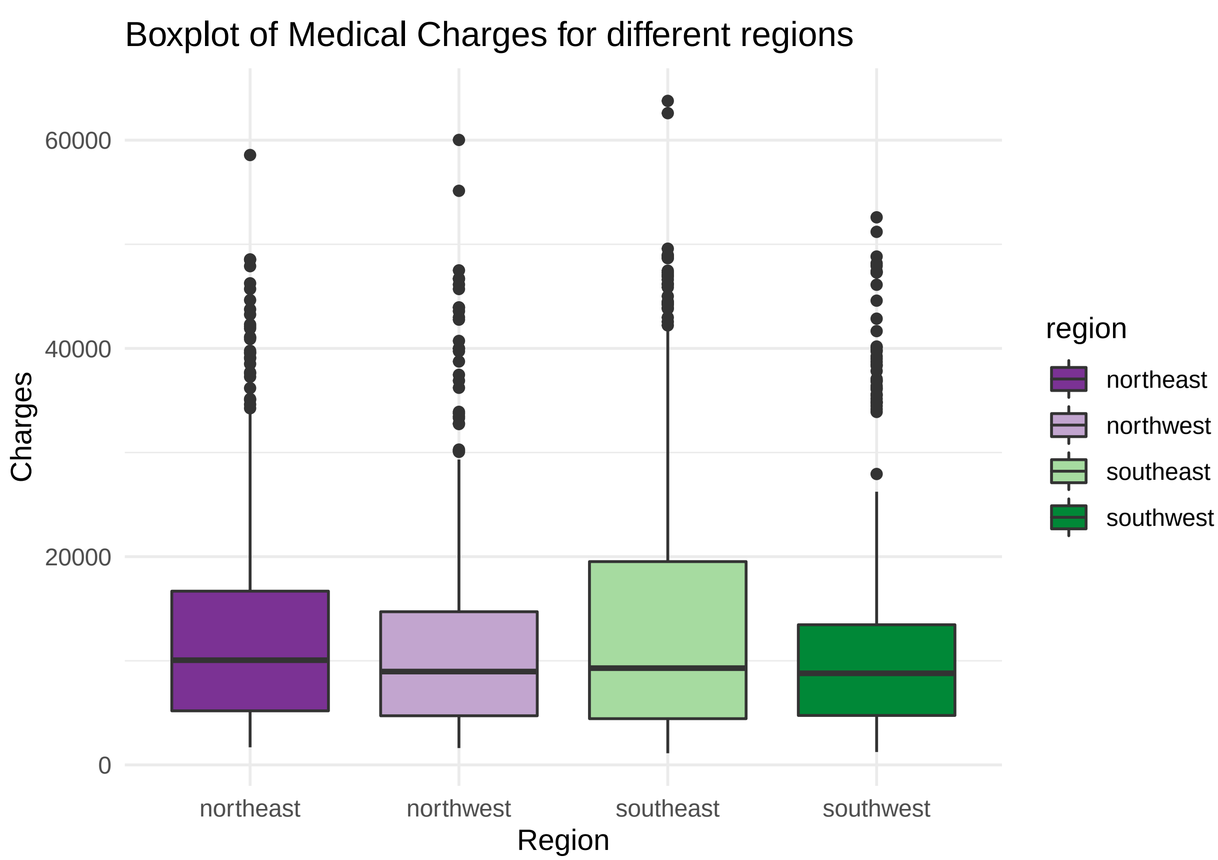
```
charges = insurance$charges
skewness(charges)
```

```
## [1] 1.51418
```

As expected, the skewness measure is positive. It indicates that the charges are positively (right) skewed.

Let us begin by visualizing the relation between the charges incurred by a resident and their region.

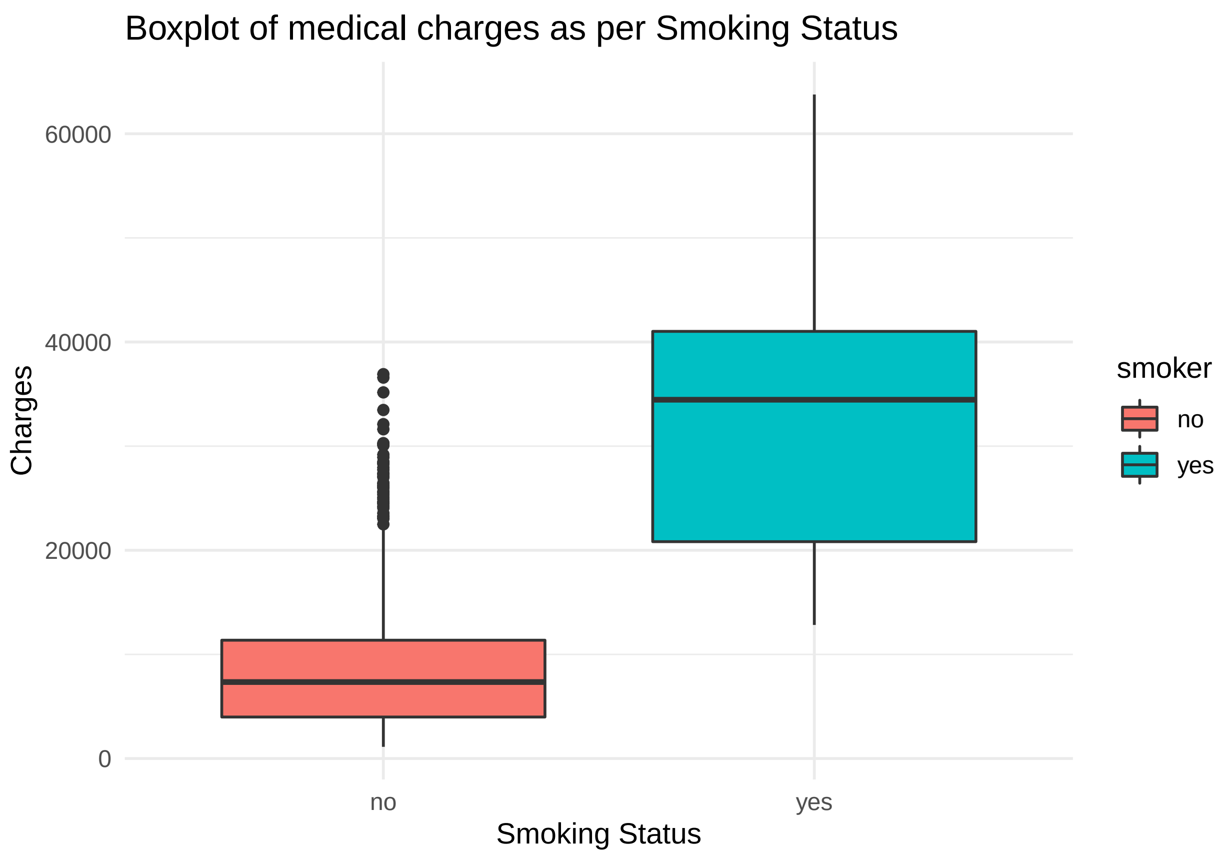
```
ggplot(insurance) +
  aes(x = region, y = charges, fill = region) +
  geom_boxplot(shape = "circle") +
  scale_fill_brewer(palette = "PRGn", direction = 1) +
  labs(
    x = "Region",
    y = "Charges",
    title = "Boxplot of Medical Charges for different regions"
  ) +
  theme_minimal()
```



These boxplots indicate that the average medical cost is similar for all the regions. That is, a resident's region does not have much impact on the medical cost incurred.

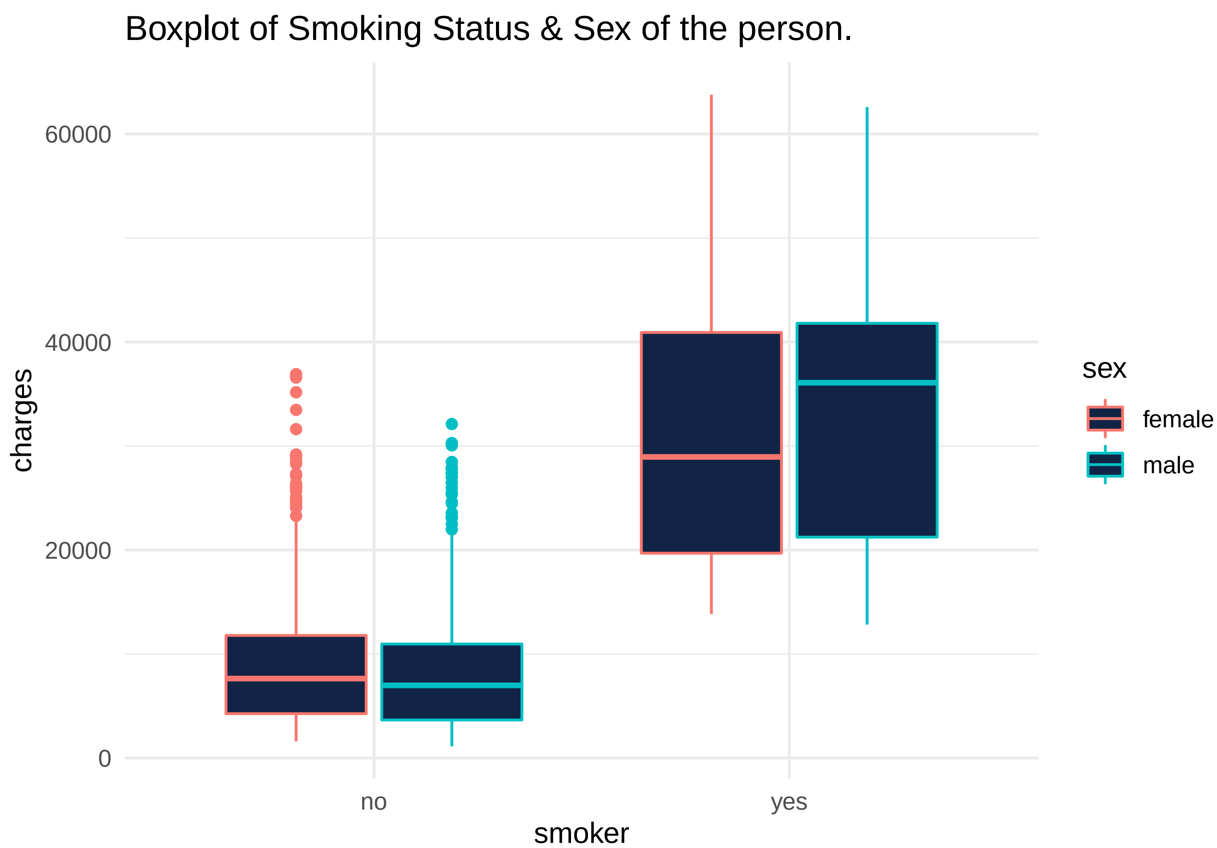
Let us now examine if smoking affects the charges for a person.

```
ggplot(insurance) +
  aes(x = smoker, y = charges, fill = smoker) +
  geom_boxplot(shape = "circle") +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Smoking Status",
    y = "Charges",
    title = "Boxplot of medical charges as per Smoking Status"
  ) +
  theme_minimal()
```



Clearly, the residents who smoke have higher charges than those who do not. This observation is further supported by separating boxplots for male and female residents.

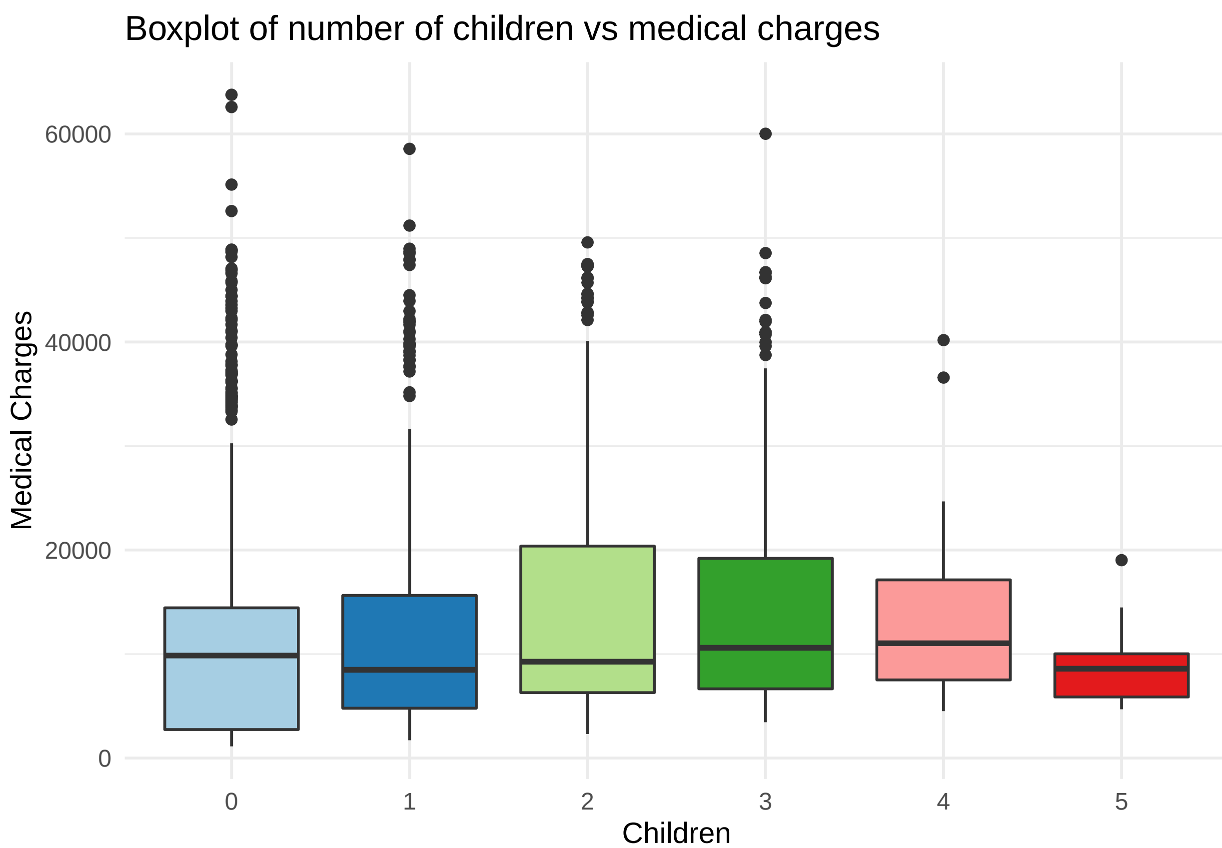
```
ggplot(insurance) +
  aes(x = smoker, y = charges, colour = sex) +
  geom_boxplot(shape = "circle", fill = "#112446") +
  scale_color_hue(direction = 1) +
  labs(
    title = "Boxplot of Smoking Status & Sex of the person."
  ) +
  theme_minimal()
```



Even in the population that smokes, the male residents incur higher charges than femalle residents.

Now, we will check if the number of children a person has affects their medical charges.

```
# encoding number of childern as categorical variable
insurance$children =  as.factor(insurance$children)
ggplot(insurance) +
  aes(x = children, y = charges, fill = children) +
  geom_boxplot(shape = "circle") +
  scale_fill_brewer(palette = "Paired", direction = 1) +
  labs(
    x = "Children",
    y = "Medical Charges",
    title = "Boxplot of number of children vs medical charges"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

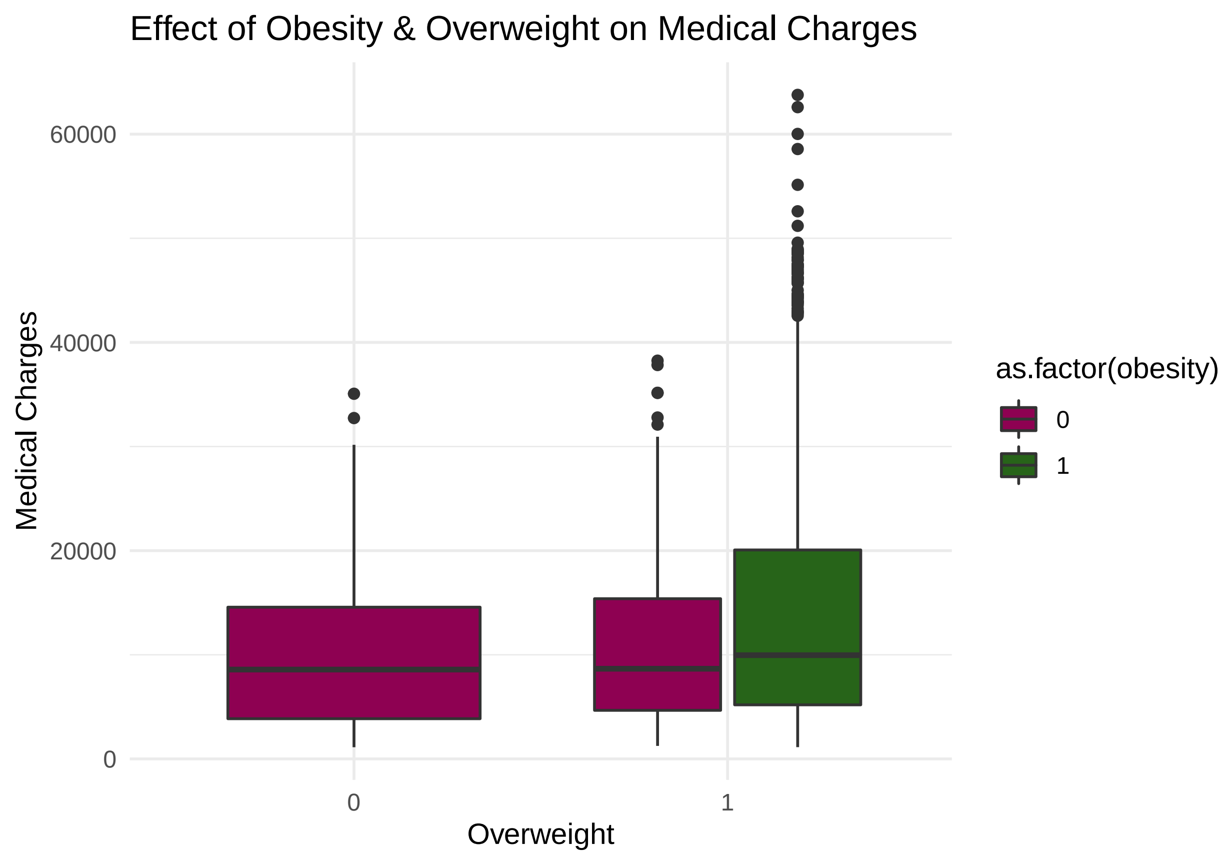


We see that the mean charges for a person is the least when they have five children.

Impact of BMI on insurance charges:

```
obesity = ifelse(insurance$bmi >= 30,1,0)
overweight = ifelse(insurance$bmi >= 25,1,0)
insurance_updated = cbind(insurance, as.factor(obesity), as.factor(overweight))

ggplot(insurance_updated) +
  aes(
    x = `as.factor(overweight)`,
    y = charges,
    fill = `as.factor(obesity)`
  ) +
  geom_boxplot(shape = "circle") +
  scale_fill_manual(
    values = c(`0` = "#8E0152",
    `1` = "#276419")
  ) +
  labs(
    x = "Overweight",
    y = "Medical Charges",
    title = "Effect of Obesity & Overweight on Medical Charges"
  ) +
  theme_minimal()
```



Relation between charges and age :

Let us first classify the population into different age groups: less than 30 years, 31 to 40 years, 41 to 50 years, 51 to 60 years and 61 to 70 years

```
summary(insurance$age)
```

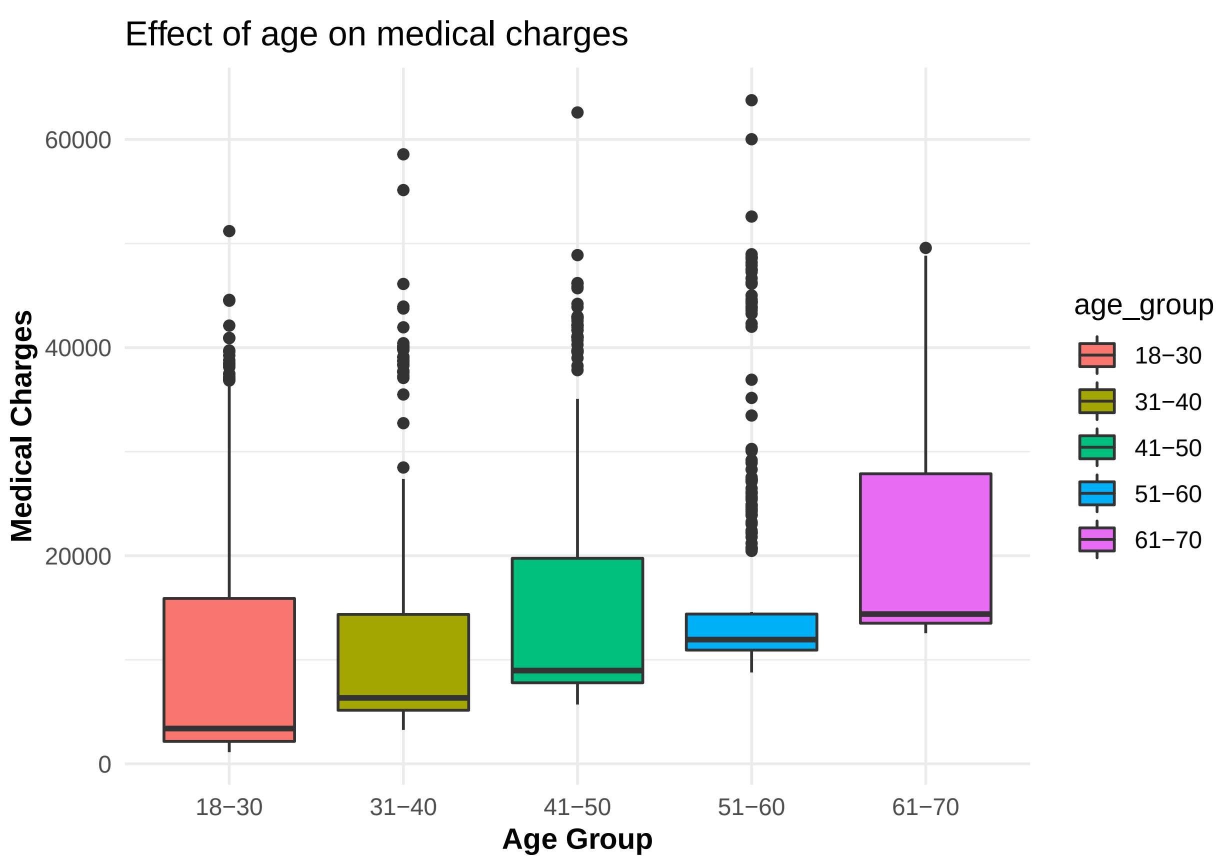
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   27.00   39.00   39.21   51.00   64.00
```

```
# we will bucket people as per age group and see how the medical charges trend in those groups.
age_group = vector("character")
for(i in 1:nrow(insurance)){
  if(insurance$age[i] <= 30){
    age_group[i] = "18-30"
  }else if (insurance$age[i] <= 40){
    age_group[i] = "31-40"
  }else if (insurance$age[i] <= 50){
    age_group[i] = "41-50"
  } else if(insurance$age[i] <= 60){
    age_group[i] = "51-60"
  }else if (insurance$age[i] <= 70){
    age_group[i] = "61-70"
  }
}
table(age_group)
```

```
## age_group
## 18-30 31-40 41-50 51-60 61-70
##   444   257   281   265    91
```

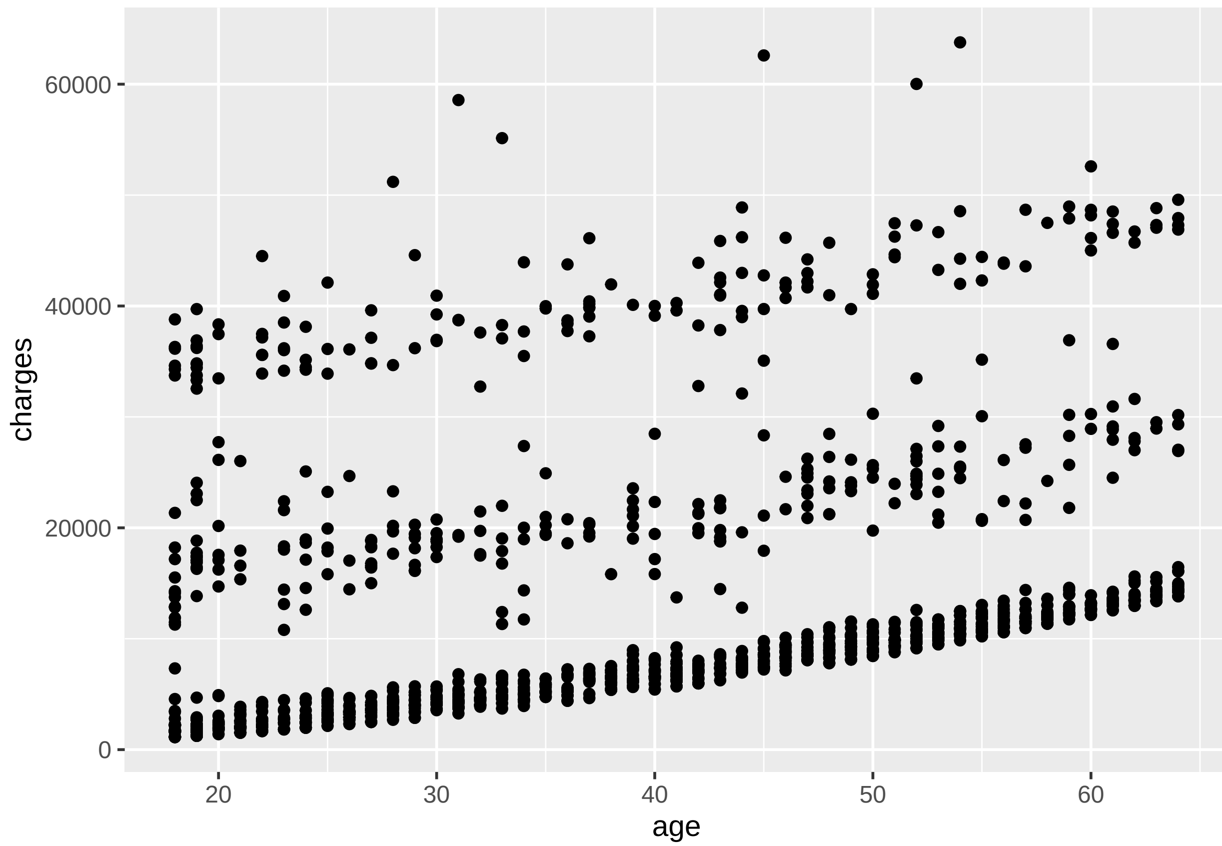
```r
age_group = as.factor(age_group)
insurance_updated = cbind(insurance_updated, age_group)

ggplot(insurance_updated) +
  aes(x = age_group, y = charges, fill = age_group) +
  geom_boxplot(shape = "circle") +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Age Group",
    y = "Medical Charges",
    title = "Effect of age on medical charges"
  ) +
  theme_minimal() +
  theme(
    axis.title.y = element_text(face = "bold"),
    axis.title.x = element_text(face = "bold")
  )
```



The boxplots indicate that the average medical charges are higher for the population in higher age groups.

```
ggplot(insurance_updated, aes(x=age, y=charges)) + geom_point()
```



Correlation Between Different Columns in the Dataset

```
# make correlation table
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: Factor w/ 6 levels "0","1","2","3",..: 1 2 4 1 1 1 2 4 3 1 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```
insurance$children = as.numeric(insurance$children)
# encoding yes as 1 and no as 0
# female as 0 and male as 1
insurance$smoker= ifelse(insurance$smoker == "yes", 1,0)
insurance$sex = ifelse(insurance$sex == "male", 1 , 0)
# encoding southwest as 1, southeast as 2, northwest as 3, northeast as 4
region_encode = vector("numeric")
for(i in 1:nrow(insurance)){
```

```r
  if(insurance$region[i] == "southwest"){
    region_encode[i] = 1
  }
  else if (insurance$region[i] == "southeast"){
    region_encode[i] = 2
  }
  else if(insurance$region[i] == "northwest"){
    region_encode[i] = 3
  }else if (insurance$region[i] == "northeast"){
    region_encode[i] = 4
  }
}
insurance_updated = cbind(insurance, region_encode)
head(insurance_updated)
```

```
##   age sex    bmi children smoker    region   charges region_encode
## 1  19   0 27.900        1      1 southwest 16884.924             1
## 2  18   1 33.770        2      0 southeast  1725.552             2
## 3  28   1 33.000        4      0 southeast  4449.462             2
## 4  33   1 22.705        1      0 northwest 21984.471             3
## 5  32   1 28.880        1      0 northwest  3866.855             3
## 6  31   0 25.740        1      0 southeast  3756.622             2
```

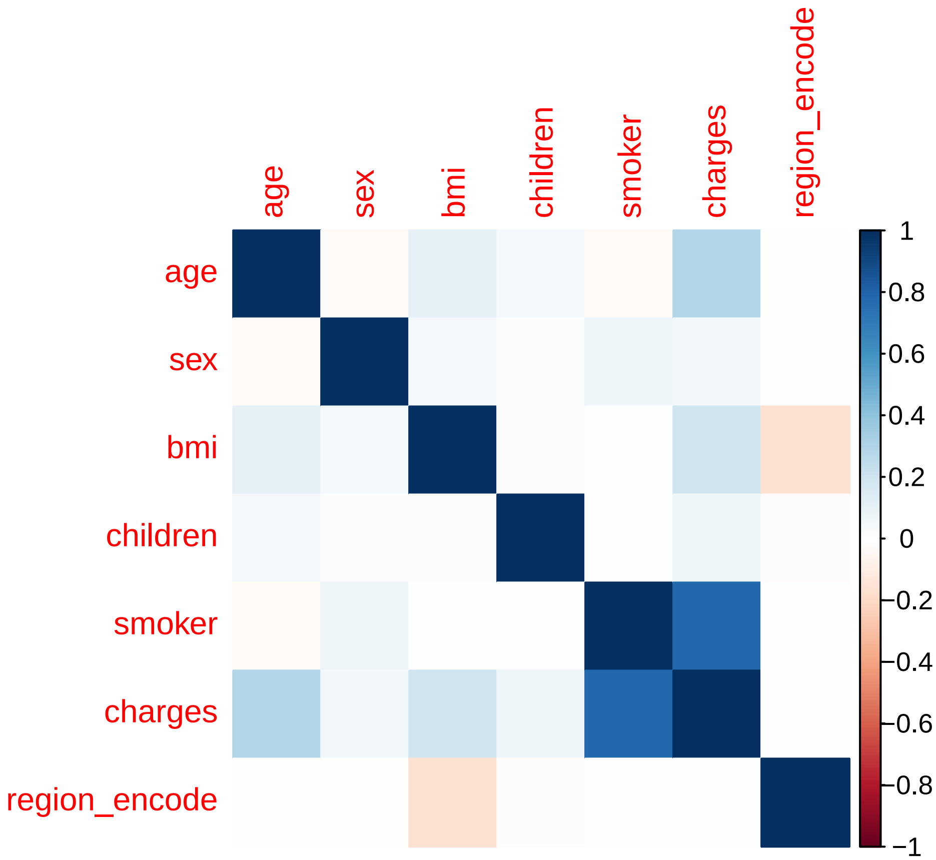
```r
colnames(insurance_updated)
```

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

```r
cor(insurance_updated[,-6])
```

```
##                       age          sex          bmi    children       smoker
## age            1.000000000 -0.020855872  0.109271882  0.04246900 -0.025018752
## sex           -0.020855872  1.000000000  0.046371151  0.01716298  0.076184817
## bmi            0.109271882  0.046371151  1.000000000  0.01275890  0.003750426
## children       0.042468999  0.017162978  0.012758901  1.00000000  0.007673120
## smoker        -0.025018752  0.076184817  0.003750426  0.00767312  1.000000000
## charges        0.299008193  0.057292062  0.198340969  0.06799823  0.787251430
## region_encode -0.002127313 -0.004588385 -0.157565849 -0.01656945  0.002180682
##                   charges region_encode
## age           0.299008193  -0.002127313
## sex           0.057292062  -0.004588385
## bmi           0.198340969  -0.157565849
## children      0.067998227  -0.016569446
## smoker        0.787251430   0.002180682
## charges       1.000000000   0.006208235
## region_encode 0.006208235   1.000000000
```

```r
corrplot(cor(insurance_updated[,-6]), method = "color")
```



The correlation plot indicates that charges are mildly correlated to age and BMI of a person, and strongly correlated to whether they smoke or not. We classify the population into obese and overweight to see how the charges fare for both the categories.

Building Models to Predict the Charges

We can now build models to predict the charges for a resident.

```
# Now we build the model.
# Split the data into train & test.
set.seed(123)
ID = 1:nrow(insurance)
insurance = cbind(ID, insurance)
train = sample(ID,1000)
test = ID[-train]
train = insurance[train,-1]
test = insurance[test,-1]
lr = lm(charges~., data = train)
summary(lr)
```

```
##
## Call:
## lm(formula = charges ~ ., data = train)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -11170  -2981  -1011   1592  30019
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -12871.52    1204.11 -10.690  < 2e-16 ***
## age                243.81      14.24  17.125  < 2e-16 ***
## sex               -300.42     394.34  -0.762 0.446338
## bmi                364.01      33.77  10.779  < 2e-16 ***
## children           605.71     161.65   3.747 0.000189 ***
## smoker           24004.41     479.10  50.103  < 2e-16 ***
## regionnorthwest   -769.50     559.20  -1.376 0.169106
## regionsoutheast   -886.49     564.73  -1.570 0.116794
## regionsouthwest  -1028.69     562.73  -1.828 0.067844 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6182 on 991 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7469
## F-statistic: 369.6 on 8 and 991 DF,  p-value: < 2.2e-16
```

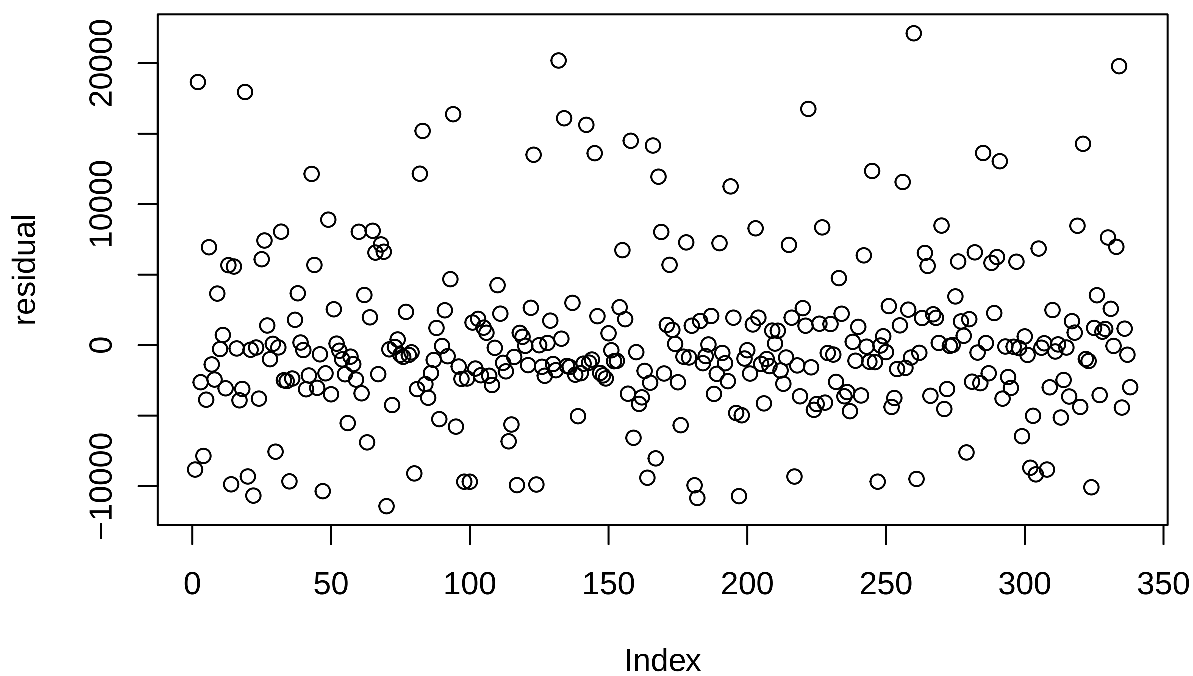
We remove the following variables from our original model as they are insignificant: **sex** and **region**.

```
lr2 = lm(charges~ age + bmi + children + smoker, data = train )
summary(lr2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11520.0  -3070.2   -940.7   1611.6  29653.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13332.66    1157.80 -11.516  < 2e-16 ***
## age            244.73      14.22  17.214  < 2e-16 ***
## bmi            351.29      32.28  10.883  < 2e-16 ***
## children       600.60     161.56   3.718 0.000212 ***
## smoker       23993.17     477.12  50.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6184 on 995 degrees of freedom
## Multiple R-squared:  0.7478, Adjusted R-squared:  0.7468
## F-statistic: 737.7 on 4 and 995 DF,  p-value: < 2.2e-16
```

The adjusted R square decreases by a very small margin, but we are able to simplify the model by removing two variables.

```r
c = lr2$coefficients
p = vector("numeric")
for (i in 1:nrow(test)){
  p[i] = c[1] + test$age[i]*c[2]+test$bmi[i]*c[3]+ test$children[i]*c[4]+test$smoker[i]*c[5]
}
residual = (test$charges-p)
plot(residual)
```




```r
#The RMSE is:
sqrt(mean(residual^2))
```

```
## [1] 5743.834
```

PREDICTION 1 : 19-year old female, smoker, overweight and with no children, resident of the Southwest region of the USA.

```r
# Prediction 1
predict(lr2, data.frame(age = 19, bmi = 27.9 , children = 0, smoker = 1))
```

```
##        1
## 25111.34
```

PREDICTION 2 : 55-year-old man, smoker, non-obese and without children, resident of the Northeast region of the USA.

```r
# Prediction 2
predict(lr2, data.frame(age = 55,bmi = 25, children = 0, smoker = 1))
```

```
##        1
## 32902.81
```

PREDICTION 3 : 70-year-old woman, non-smoker, obese and with two children, resident of the southeastern USA.

```r
# Prediction 3
predict(lr2, data.frame(age = 70, bmi = 35, children = 2, smoker = 0))
```

```
##        1
## 17294.65
```

PREDICTION 4 : 22x-year-old woman, smoker, non-obese and with 4 children, resident of the northwestern region of the USA.

```r
# Prediction 4
predict(lr2, data.frame(age = 22,bmi = 23, children = 4,smoker = 1))
```

```
##        1
## 26526.59
```

Conclusion

We carried out data analysis using library and also linear regression on charges and other input variables. We found that:

1. The variable **smoker** affects charges the most. That is, a person who smokes is likely to incur higher charges than those who do not.
2. The BMI of a person mildly affects the charges. If a person is in obese category, they are expected to incur higher charges than other overweight and underweight residents.
3. On an average, people with 5 children are charged less than others.
4. Region and sex of a resident do not affect the charges significantly. Hence, these input variables are not of much interest while studying the insurance data.

We built two models: by filtering on the basis of significance.

1. Regressing charges on all the input variables
2. Regressing charges on all the input variables **except sex and region** .

In the second model, the fit is not impacted adversely as the adjusted R^2 value decreases slightly from 0.7469 to 0.7468.