Linear Regression

Ashish Toppo

Dataset Link: Medical Cost Personal Datasets

Using this data set I need to predict the cost billed by a health insurance company based on some parameters.
Description:
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

Reading of the dataset and importing necessary libraries

```
library(tidyverse)
library(reshape2)
library(caTools)
library(corrplot)
library(MLmetrics)
library(ggfortify)
data_0 <- read.csv("insurance.csv", stringsAsFactors = TRUE) #converting all strings to factors
```

```
#peeping into data
head(data_0)
```

```
##   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
```

```
# dimension of the dataset
dim(data_0)
```

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

```
# structure of the dataset
str(data_0)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Factors: are the data objects which are used to categorize the data and store it as levels.
I'll be converting variables to factors if there comes a need to. For now I'll be jumping into analysis.

Summary Statistics

```
summary(data_0)
```

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                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
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Observations:
1. The insurance is only provided to adults only(age<=18).
2. The maximum age is noted to be 64 which shows that the insurance isn't covered for senior citizens.source
3. The data set has got nearly same proportion for both male and female.
4. Columns like "bmi" ,"region" and "children" are worth checking using visualization techniques.
5. Even the regions are somewhat in similar proportions.

Missng Values and their proportions by columns

```
null_cols <- colSums(is.na(data_0))/nrow(data_0)*100
null_cols[null_cols!=0]
```

```
## named numeric(0)
```

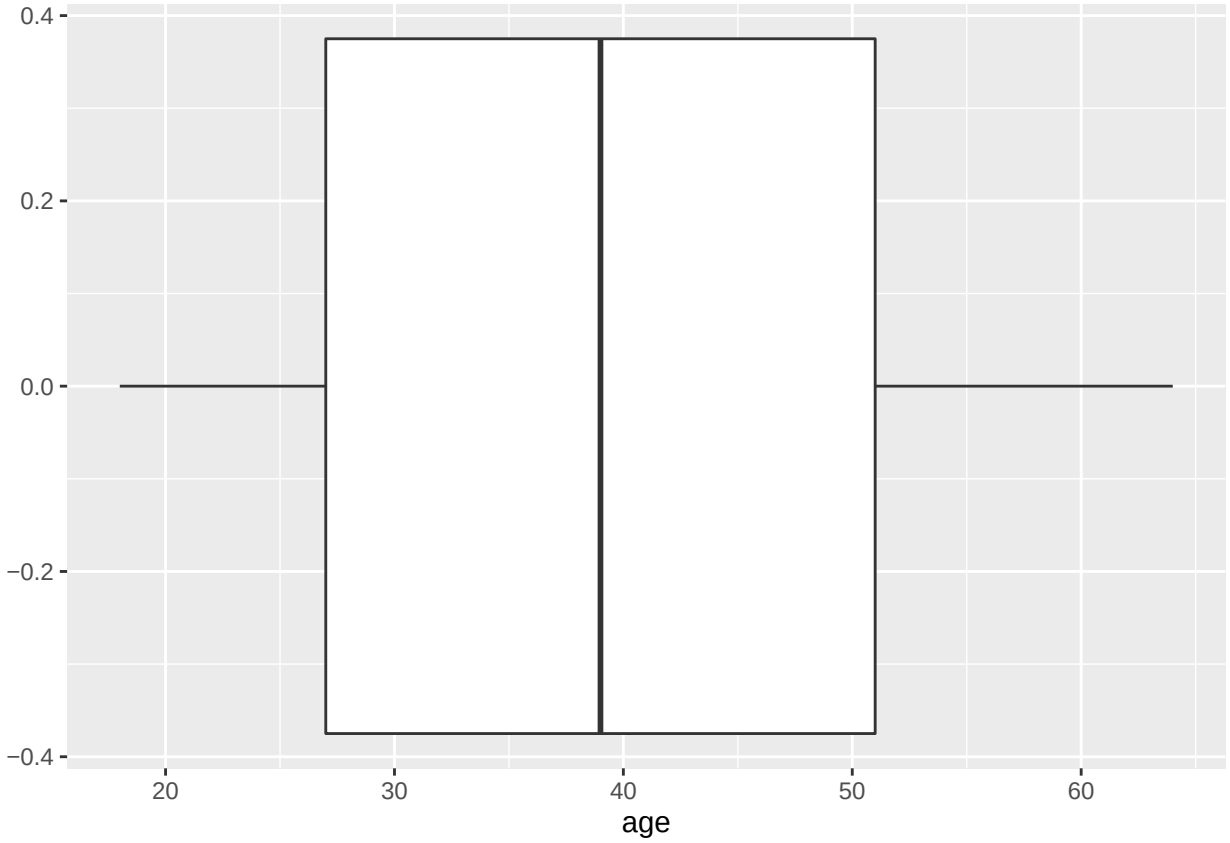
We are lucky to not have any missing data in any of the columns.

Looking for outliers in numeric columns and their analysis

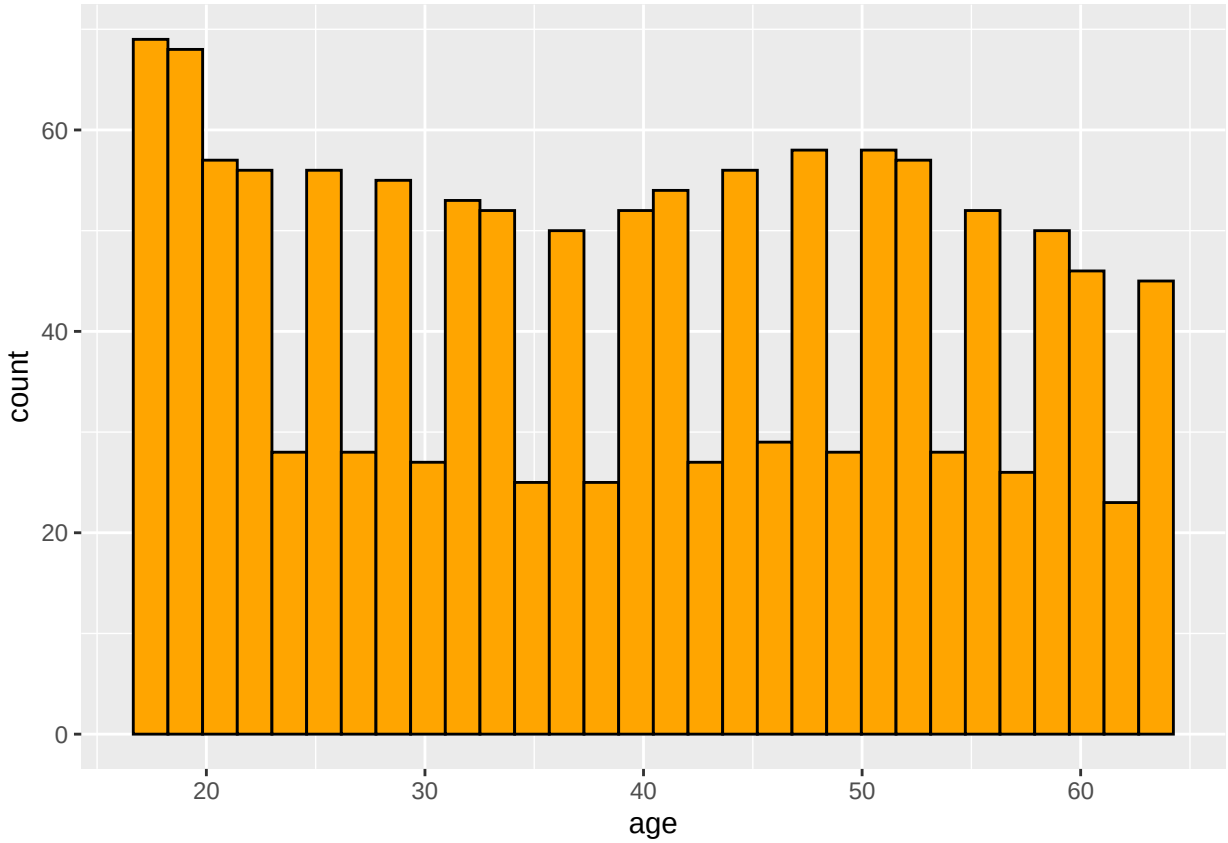
```
box_dist <- function(x,y){
  #` a custom function to plot histogram and box plot for a column
  print(x %>% ggplot(aes_string(y))+geom_boxplot())
  print(x %>% ggplot(aes_string(y))+geom_histogram(fill="orange", color="black"))
}
```

For age

```
box_dist(data_0,"age")
```




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



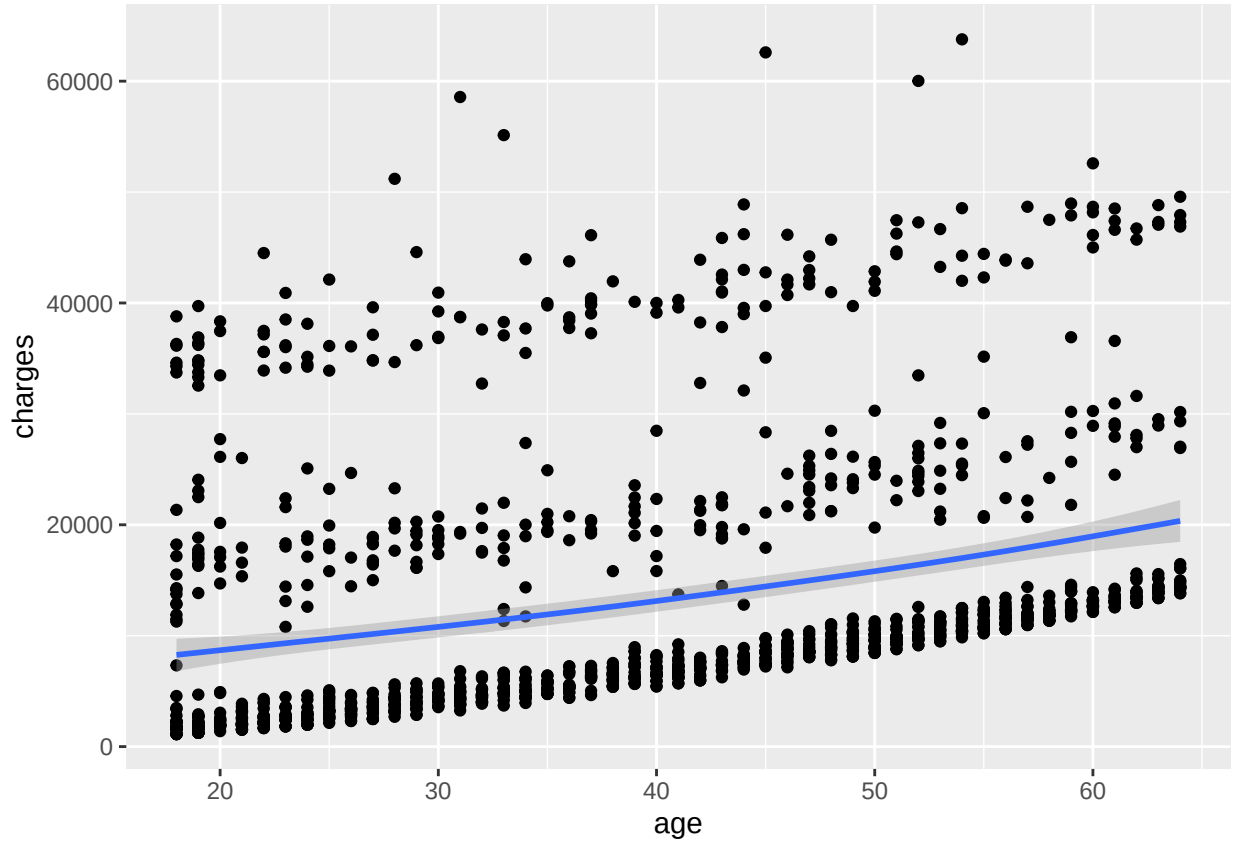
Observations
1. There aren't any outliers.
2. Looking at the distribution we can see that:
* Highest number of insurances was purchased by people below the age of 20, even though people don't buy insurances at such early ages so its worth checking which gender or region has such customers.
* We can see a **WAVY** curve which reduces at ages near 35 and near 64, which can be thought as people normally marry in early thirties and people becoming health concerned at the early fifties.
* There isn't any other pattern seen with regards to age.

Doing some Analysis on age

Q *Linear relation b/w age and charges*

```
linearity <- function(df,c1,c2){
  #`a function to produce scatter plot and checking correlation b/w independent and dependent
  print(df %>% ggplot(aes_string(c1,c2))+geom_point()+geom_smooth())
  print(cor(df[c1],df[c2]))
}
linearity(data_0,"age","charges")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
##       charges
## age 0.2990082
```

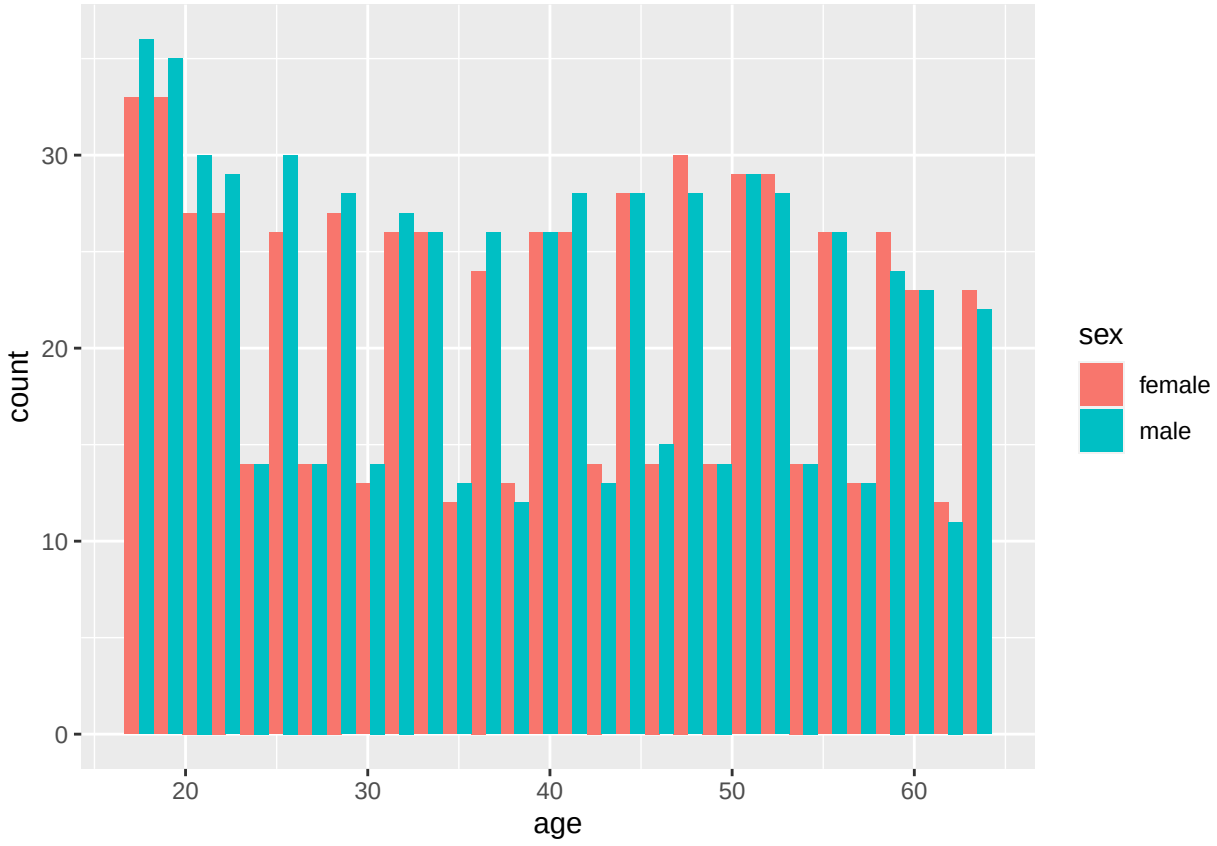
Observations

1. There is a linear relation b/w age(independent) and charges(dependent) but weak correlation.

Q *Is gender a factor for getting life insurance and does age matters*

```
data_0 %>% ggplot(aes(age))+geom_histogram(aes(fill=sex),position = "dodge")
```

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



Observations

1. Men tend to buy insurance at early ages(age <30).
2. There is an increase in no. of women choosing insurance after age 40 compared to men. It will be worth checking if regions or bmi play a role in this trend.

Q *Looking why there is increase in no. of women getting insured after age 40*

```
#checking if a particular region causes this rise
data_0 %>% filter(age>40) %>%  group_by(region) %>% summarise(table = table(sex))
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 8 x 2
## # Groups:   region [4]
##   region    table
##   <fct>     <table>
## 1 northeast 79
## 2 northeast 76
## 3 northwest 80
## 4 northwest 73
## 5 southeast 83
## 6 southeast 89
## 7 southwest 79
## 8 southwest 78
```

Observations
The above table might be confusing but it's simple as sex(factor) has got levels female, male and first is female.
1. So every odd index for each region, counts the no. of females with age>40.
2. So it seems that region doesn't plays a role in this as regions has got (79,80,83,79) which is almost in similar proportion.

```
# lets check if bmi is cause for this increase
data_1 <- data_0 #to not tamper original data
# converting bmi's to categories
data_1$bmi <- cut(data_1$bmi, breaks = c(-Inf,24.9,30,Inf), labels = c("healthy","overweight","obese"))
data_1 %>% filter(age<40) %>% group_by(bmi) %>% summarize(table=table(sex))
```

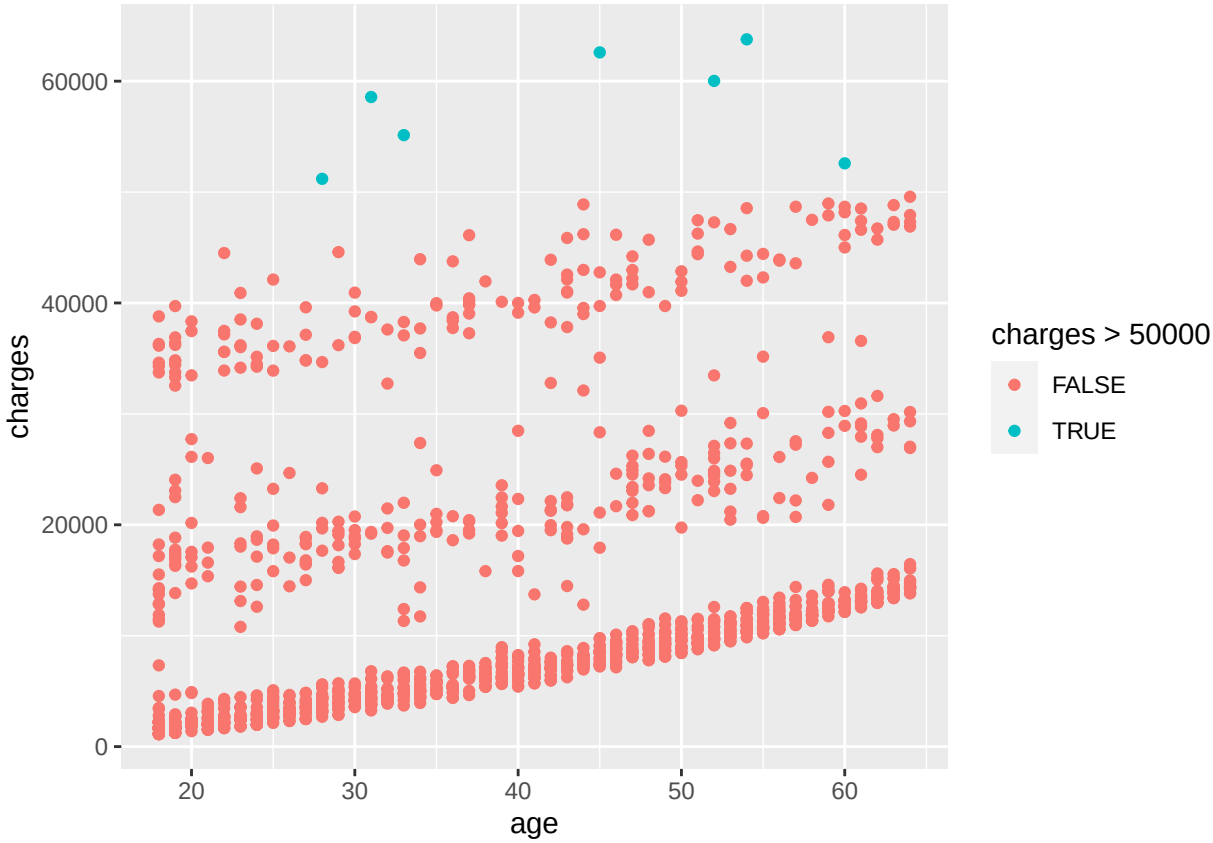
```
## `summarise()` has grouped output by 'bmi'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 6 x 2
## # Groups:   bmi [3]
##   bmi        table
##   <fct>      <table>
## 1 healthy     81
## 2 healthy     65
## 3 overweight  95
## 4 overweight 106
## 5 obese      152
## 6 obese      175
```

Observations
1. In the above table every odd index for each bmi category is count of females.
2. It can be seen that BMI is a possible cause for this rise in no. of women getting insured after age 40, they might gain weight as it's highly likely that they become mothers by this age which alters bmi.
3. Still compared to men women are healthy even after the age of 40(compare first two rows)

Q *Are age and charges correlated*

```
data_0 %>% ggplot(aes(age,charges)) + geom_point(aes(color=charges>50000))
```



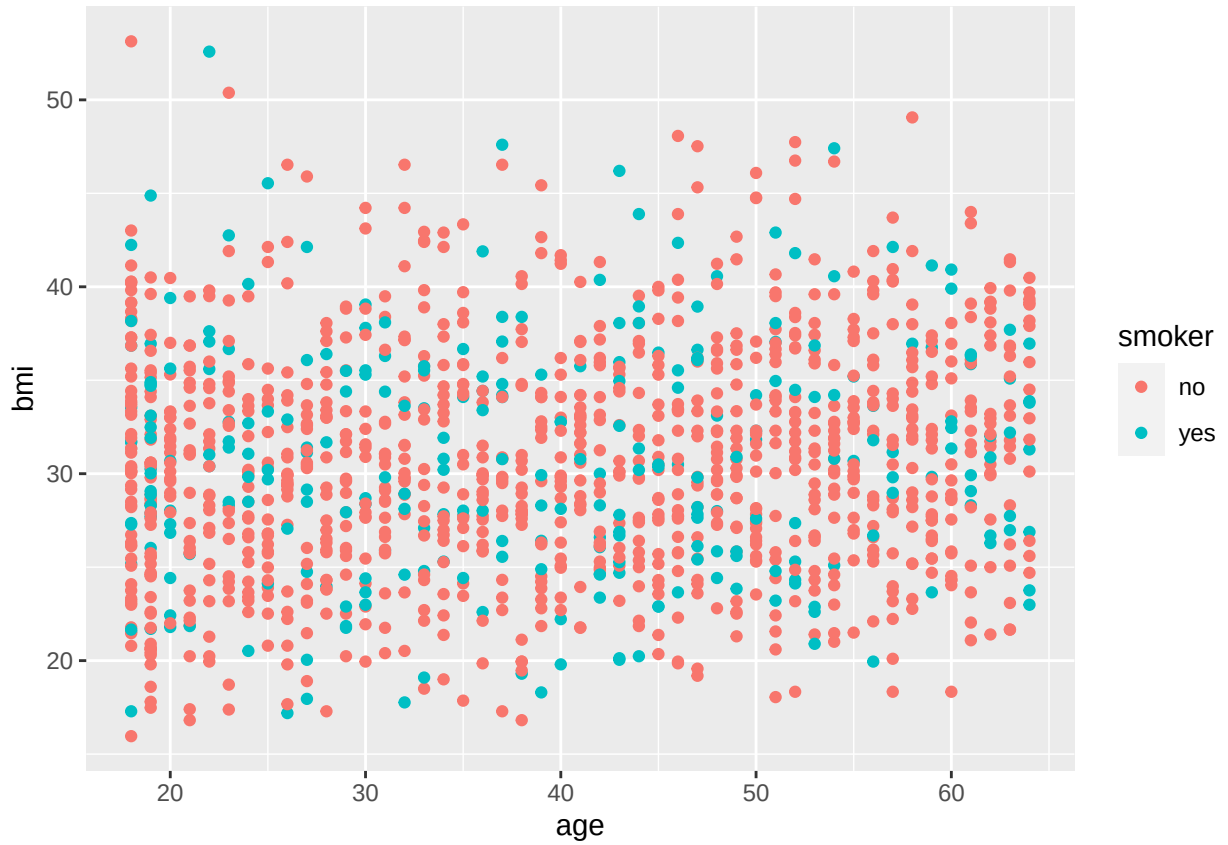
Observations

1. There is a linear trend which can be observed so age influences the charges as when people age they are prone to getting ill.
2. There are few patients which are charged more than 50000 for some people b/w 25 to 60, maybe the got some serious illness. A look in BMI and checking if that person is a smoker might get us some insight.

Q *Does BMI related to age and smoking*

```
data_0 %>% ggplot(aes(age,bmi)) + geom_point(aes(color=bmi>24.9))# bmi <24.9 is overweight <30 is obese
```


```
data_0 %>% ggplot(aes(age,bmi)) + geom_point(aes(color=smoker))
```



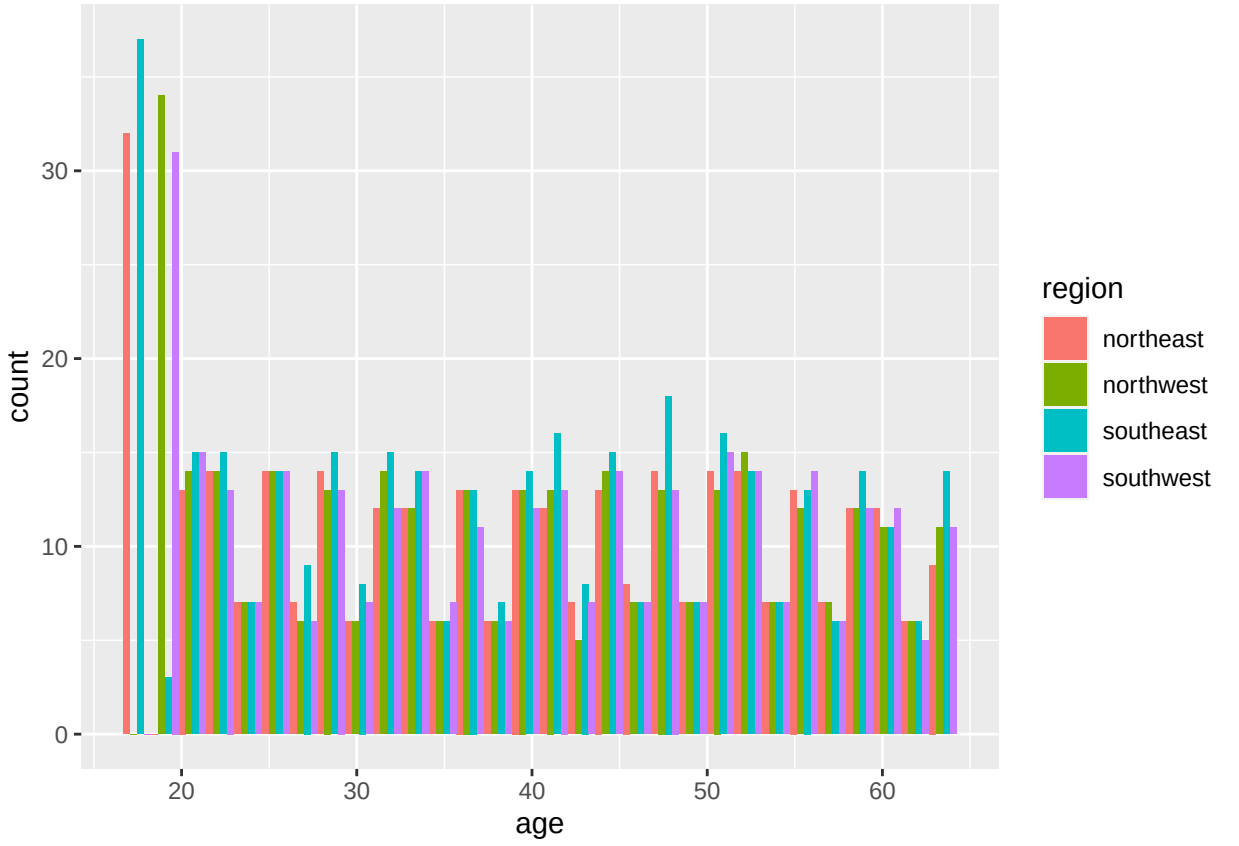
Observations

1. Many people are not having a healthy bmi(<24.9) as seen in the plot, they are either overweight(bmi<24.9) or obese(bmi<30).
2. BMI isn't affected by age but affected by life style as the points are spread evenly for all the ages. Even some people have bmi above 50 at top left for age < 25.
3. It's hard to tell whether smoking affects the bmi as people of all the ages smoke as per data.

Q *People from which region and ages choose to get insured*

```
data_0 %>% ggplot(aes(age))+geom_histogram(aes(fill=region),position = "dodge")
```

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

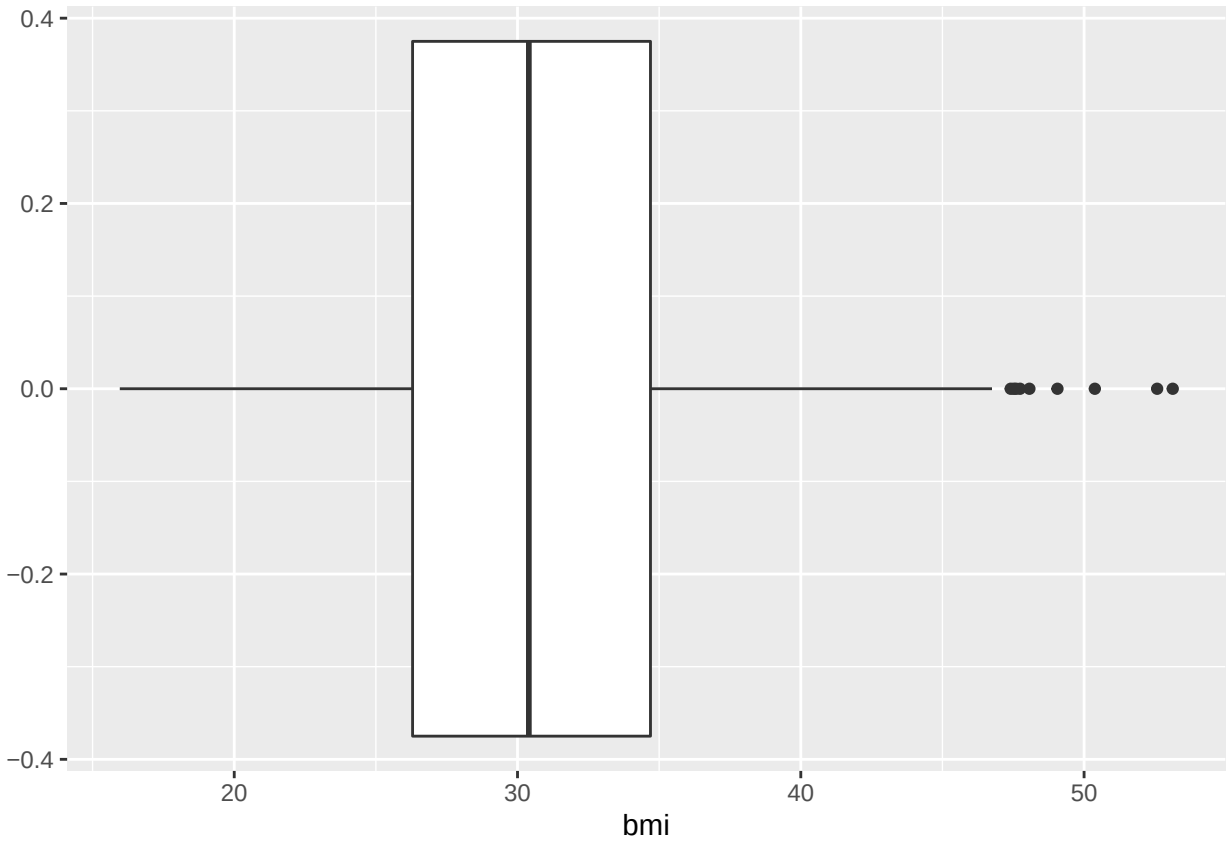


Observation

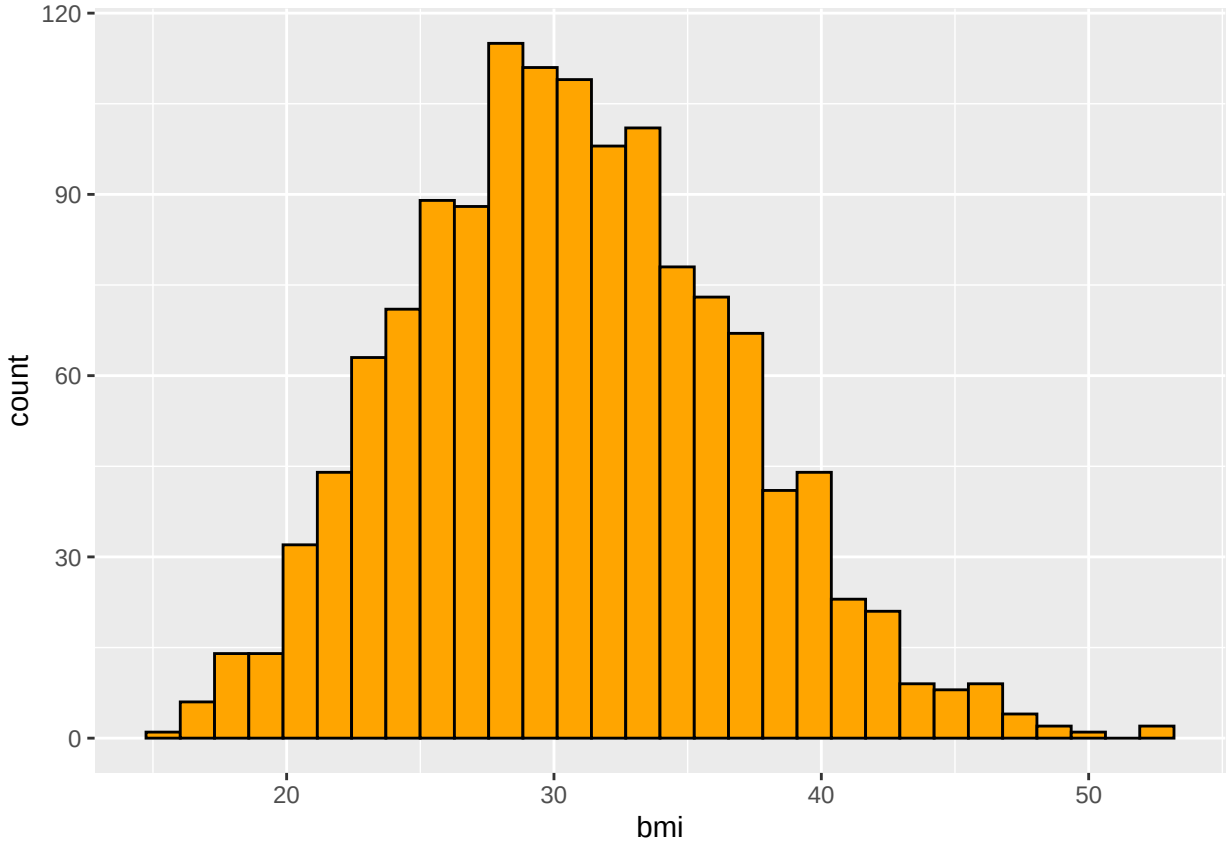
1. People from southwest region doesn't get insured early (age<=19).
2. The southeast region has got the highest no. of people insured at every age as seen by the **BLUE BARS**. It might be that the data has got highest no. of points belonging to southeast region.

For bmi

```
box_dist(data_0,"bmi")
```




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



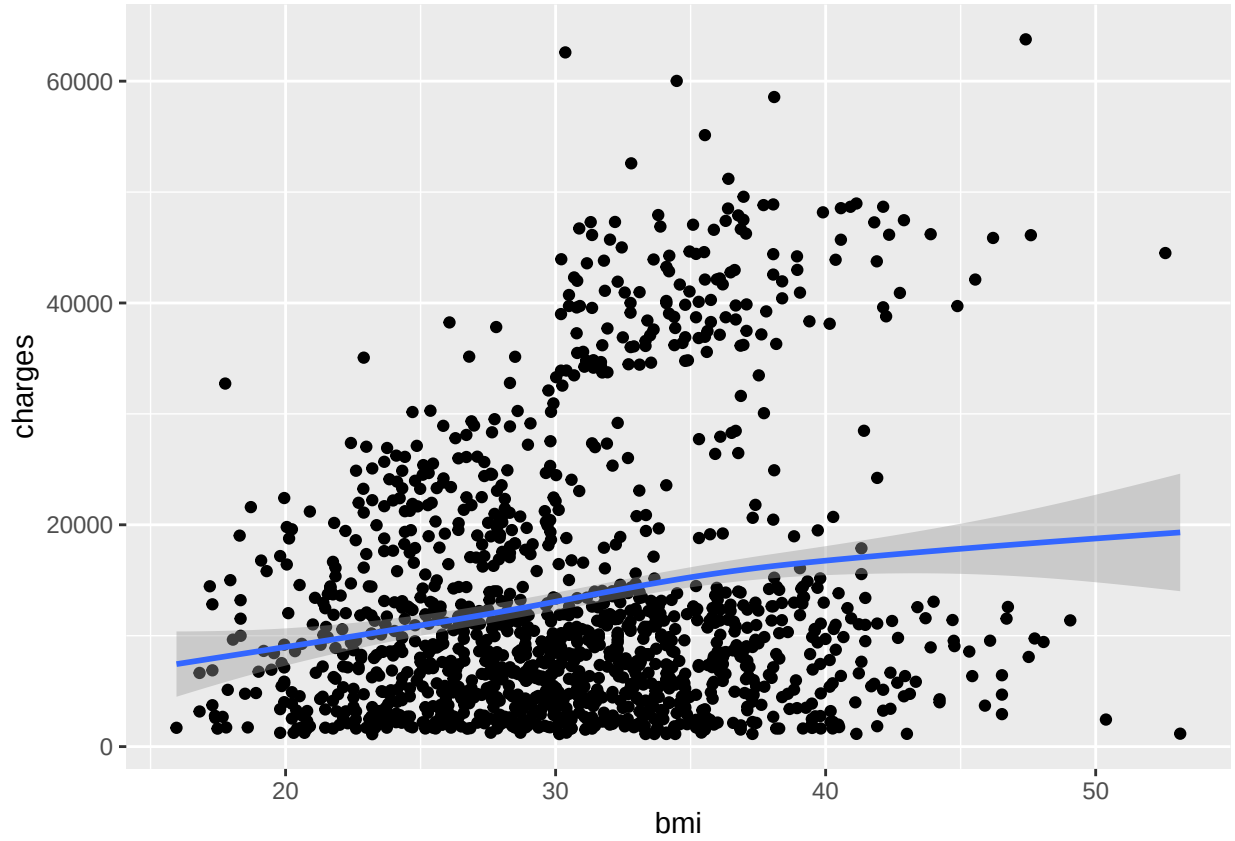
Observations
There are outliers as observed previously in (age vs bmi). Let's first check if bmi is related to cost of insurance.

Doing some Analysis on bmi

Q *Linear relation b/w bmi and charges*

```
linearity(data_0,"bmi","charges")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

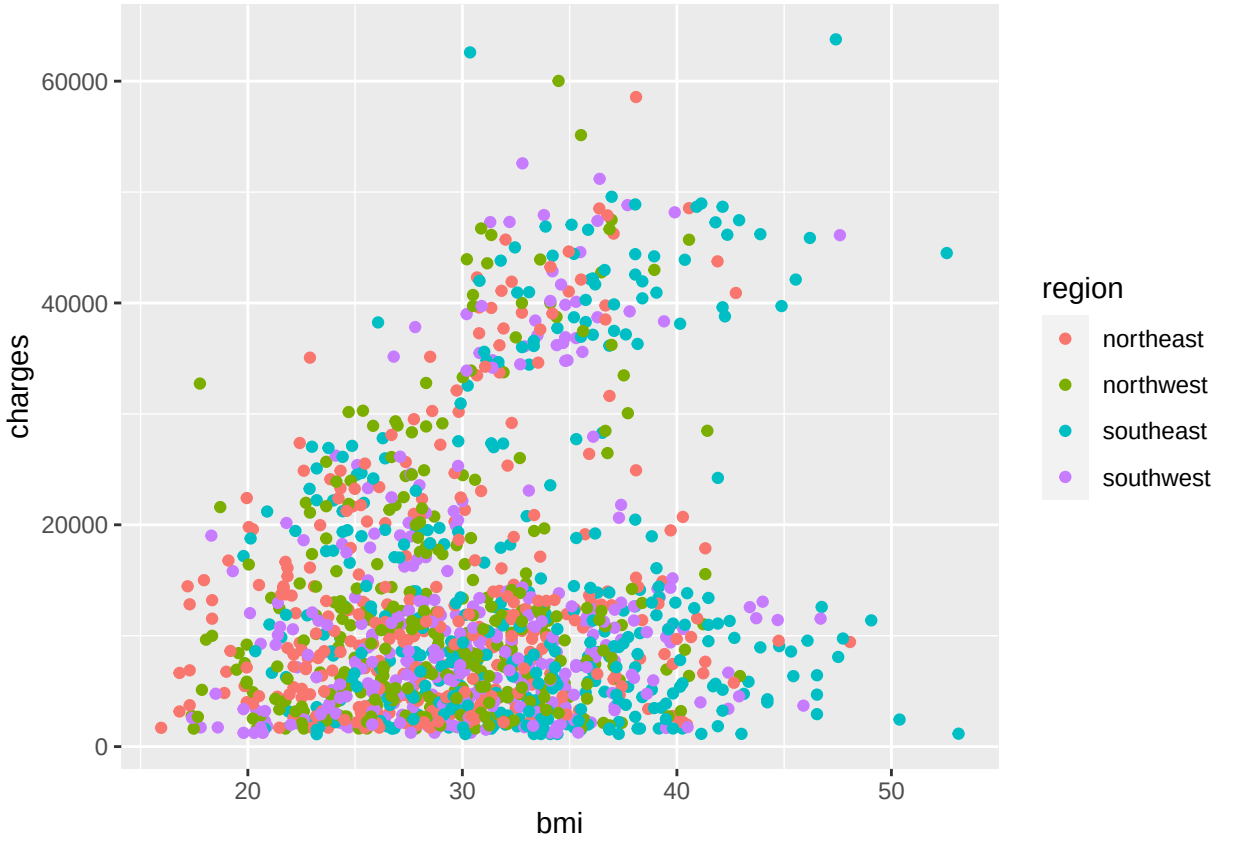



```
##      charges
## bmi 0.198341
```

Observations
1. There is a linear relationship(almost) b/w bmi(independent) and charges(dependent) but too weak correlation(<.2).

Q *Are bmi and charges related and does region affects this relationship*

```
data_0 %>% ggplot(aes(bmi,charges))+geom_point(aes(color=region)) #outliers seem to be above bmi of 47
```



Observations
1. The outliers(bmi<47) surprisingly comes from the southeast region(blue extreme points to right).
2. Most of the points (especially blue which represent southwest region) has got points greater than bmi of 40. So highest degree of obese people are from southeast region.
So, It can be seen that i can remove the outliers in bmi as it won't affect much as it has points to preserve the trend that most points above bmi of 40 belong to **southeast region**.
3. It can be seen that the charges tend to drastically increase for some people with bmi<30(obese) so yes BMI influences the charges.
4. It is such that the people's bmi within 25 to 35 has got most of the points. It may be that the bmi of 30 is the threshold for a person to be obese and a person's might be having health issues with bmi closer to 30 or after 30.

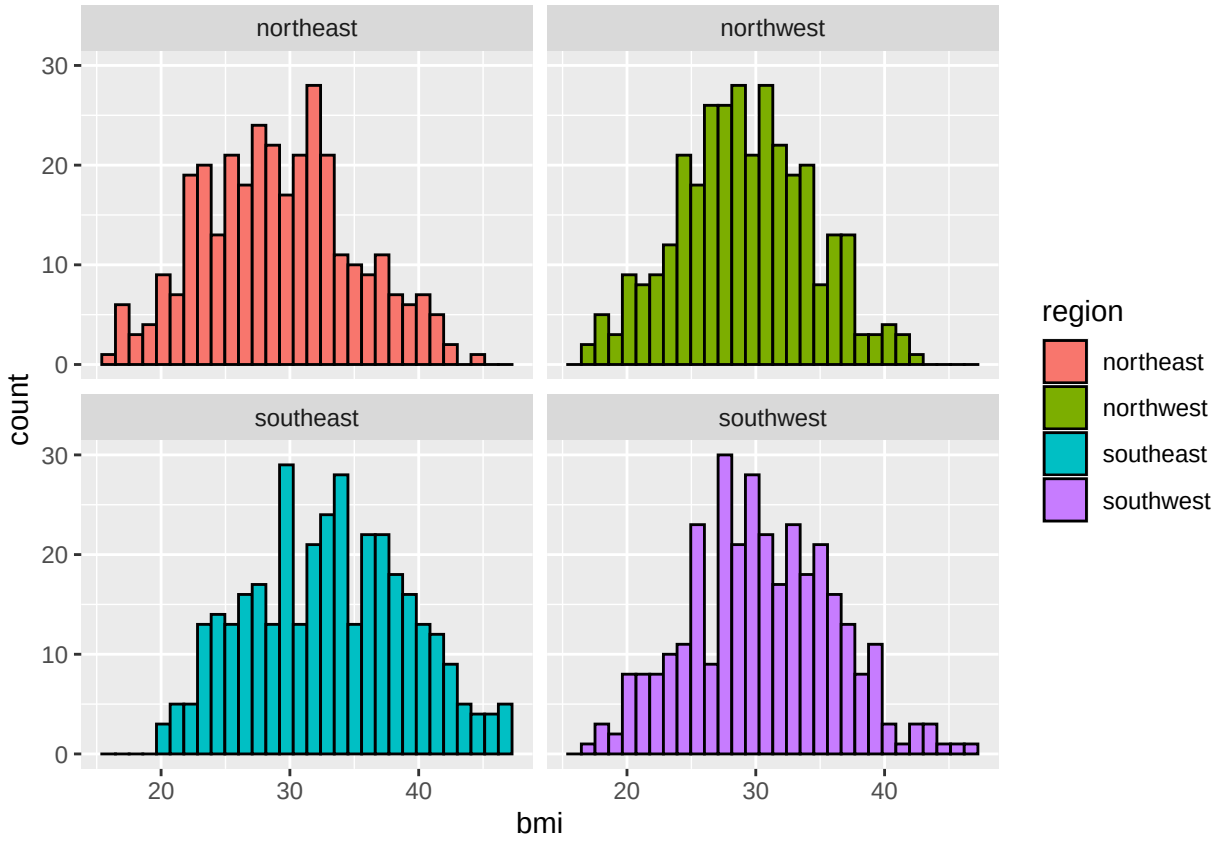
```
# removing outliers in bmi
data_2 <- data_0 %>% filter(!bmi>47) #not tampering original data
# no. of rows removed
nrow(data_0)-nrow(data_2)
```

```
## [1] 9
```

Q *What type of bmi's does each region have*

```
data_2 %>% ggplot(aes(bmi))+geom_histogram(aes(fill=region),color="black")+facet_wrap(~region)
```

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

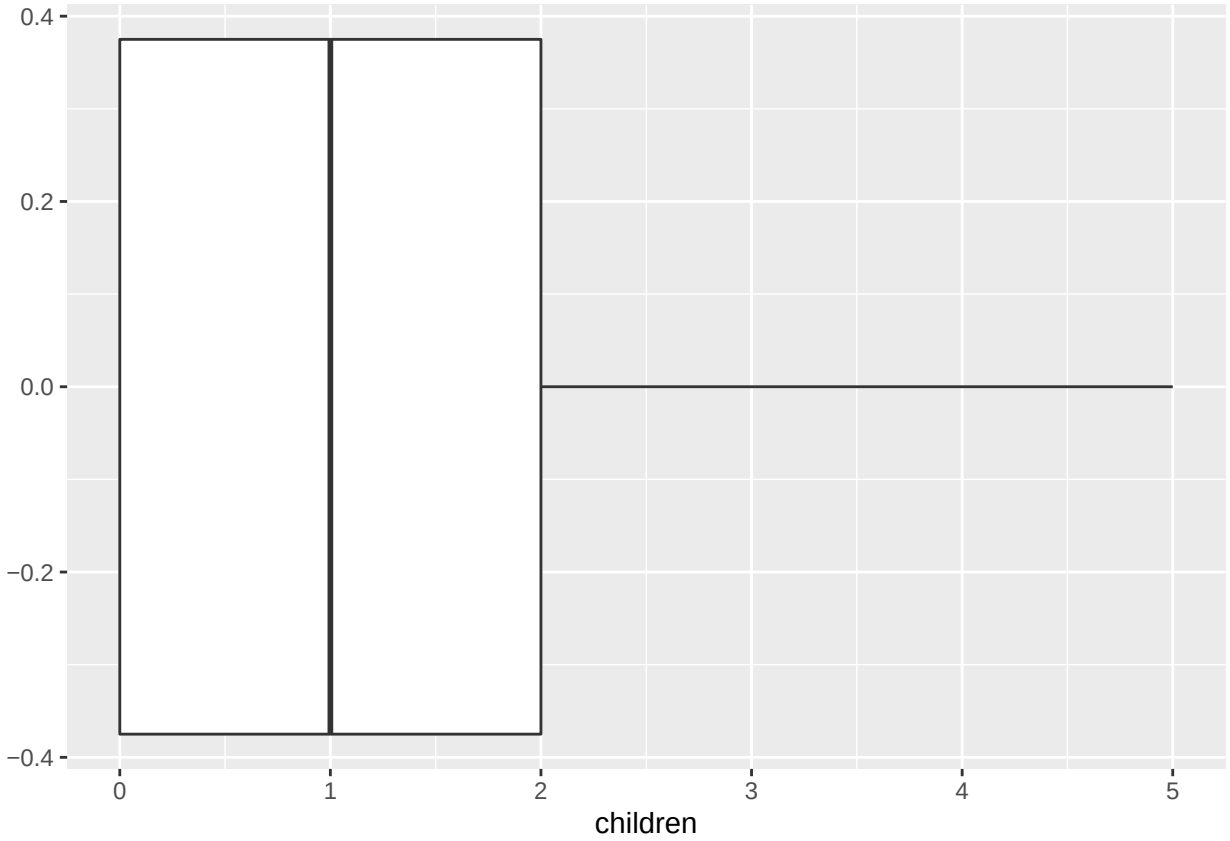


Observations

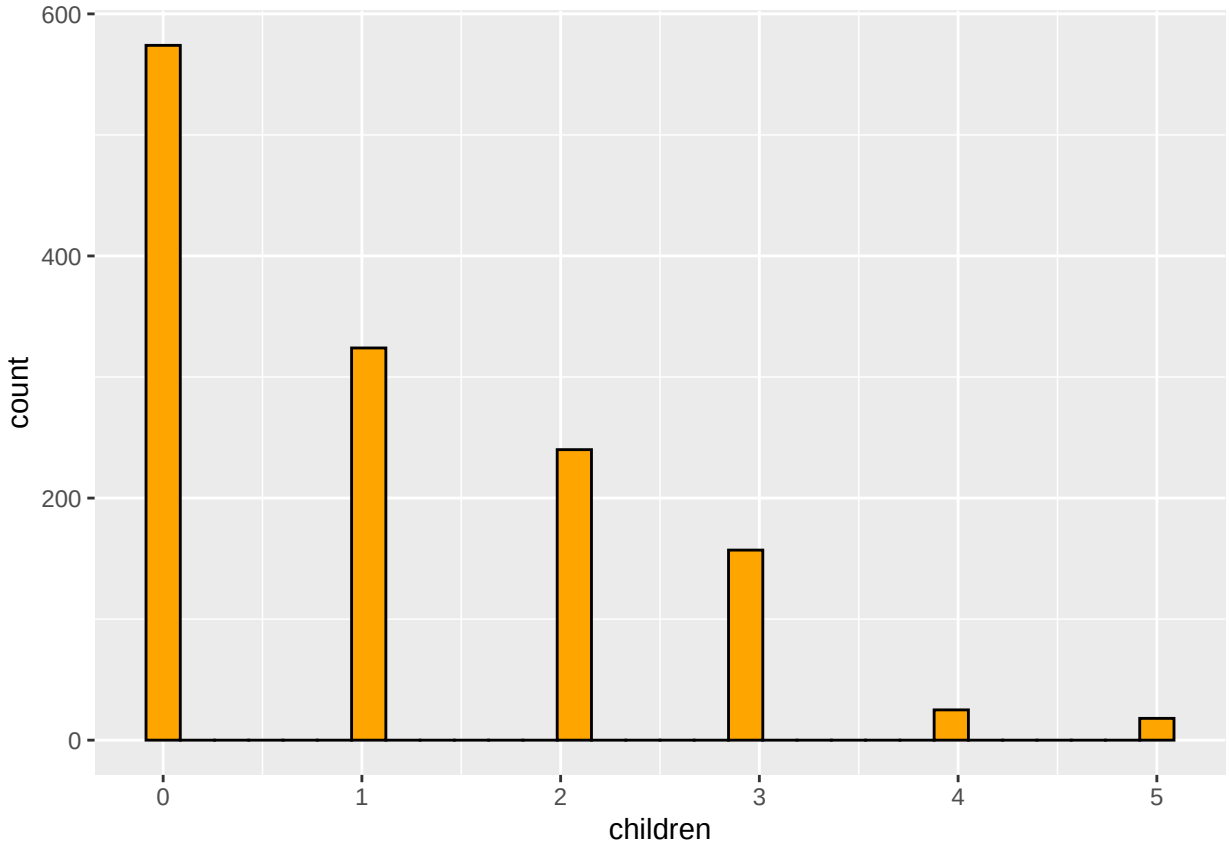
1. Southeast region doesn't have underweight people(bmi<18.5).
2. All the regions have high no. of obese people.

For Children

```
box_dist(data_0,"children")
```




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



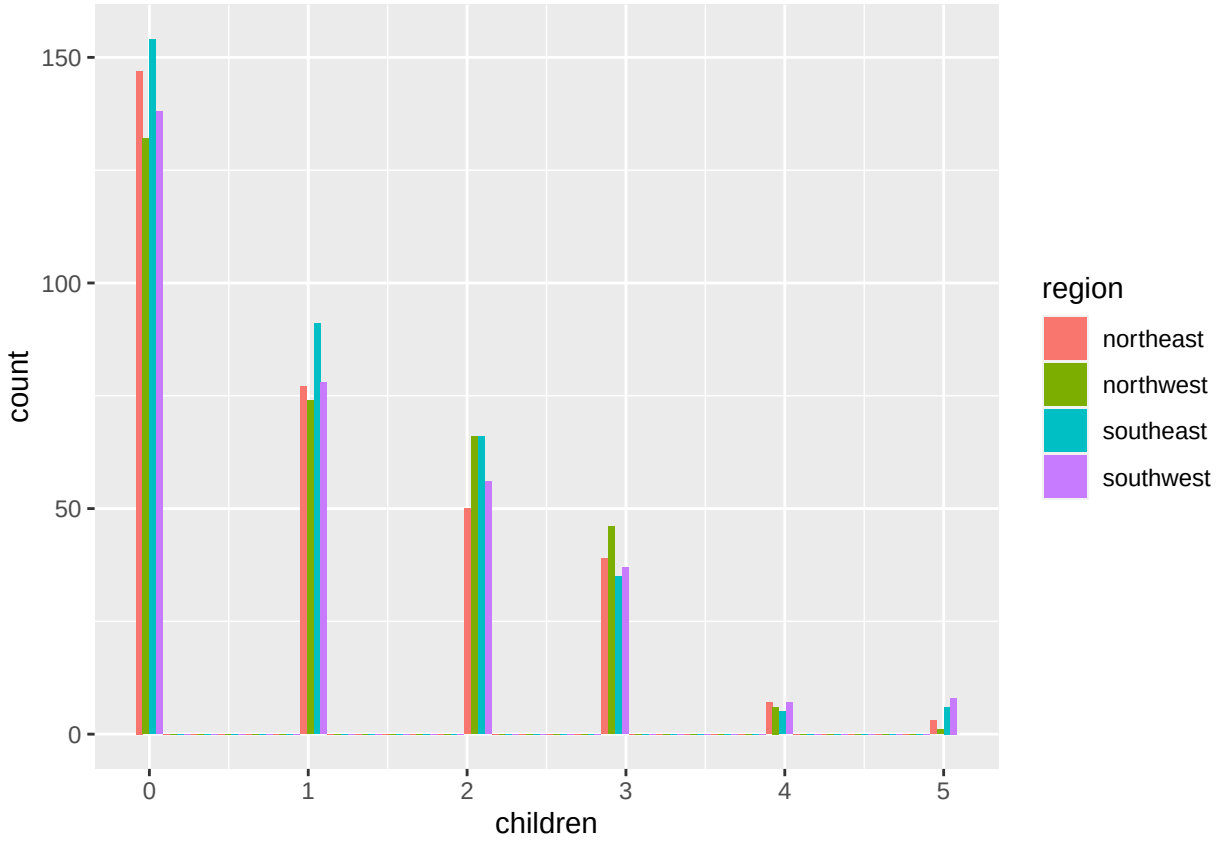
Observations
1. No outliers, hence no need for removal or transformation(s).

Some analysis

Q *No. of children per region*

```
data_2 %>% ggplot(aes(children))+geom_histogram(aes(fill=region),position = "dodge")
```

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

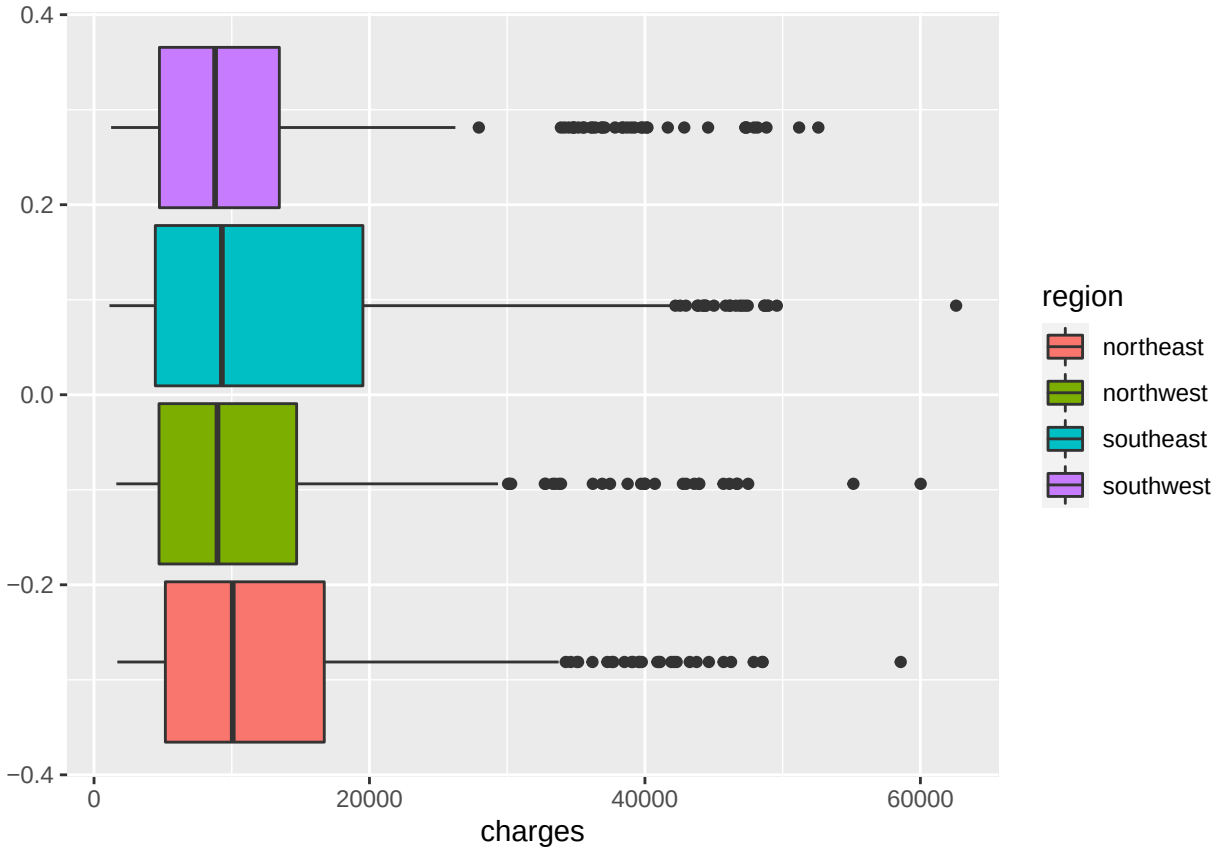


Observations

1. Southeast region has the maximum people with no children followed by northeast region.
2. Northwest region has the minimum people with 5 no. of children.
3. Southwest region has the maximum people with children <=2.

Q *How much of charges does each region make*

```
data_2 %>% ggplot(aes(charges))+geom_boxplot(aes(fill=region))
```

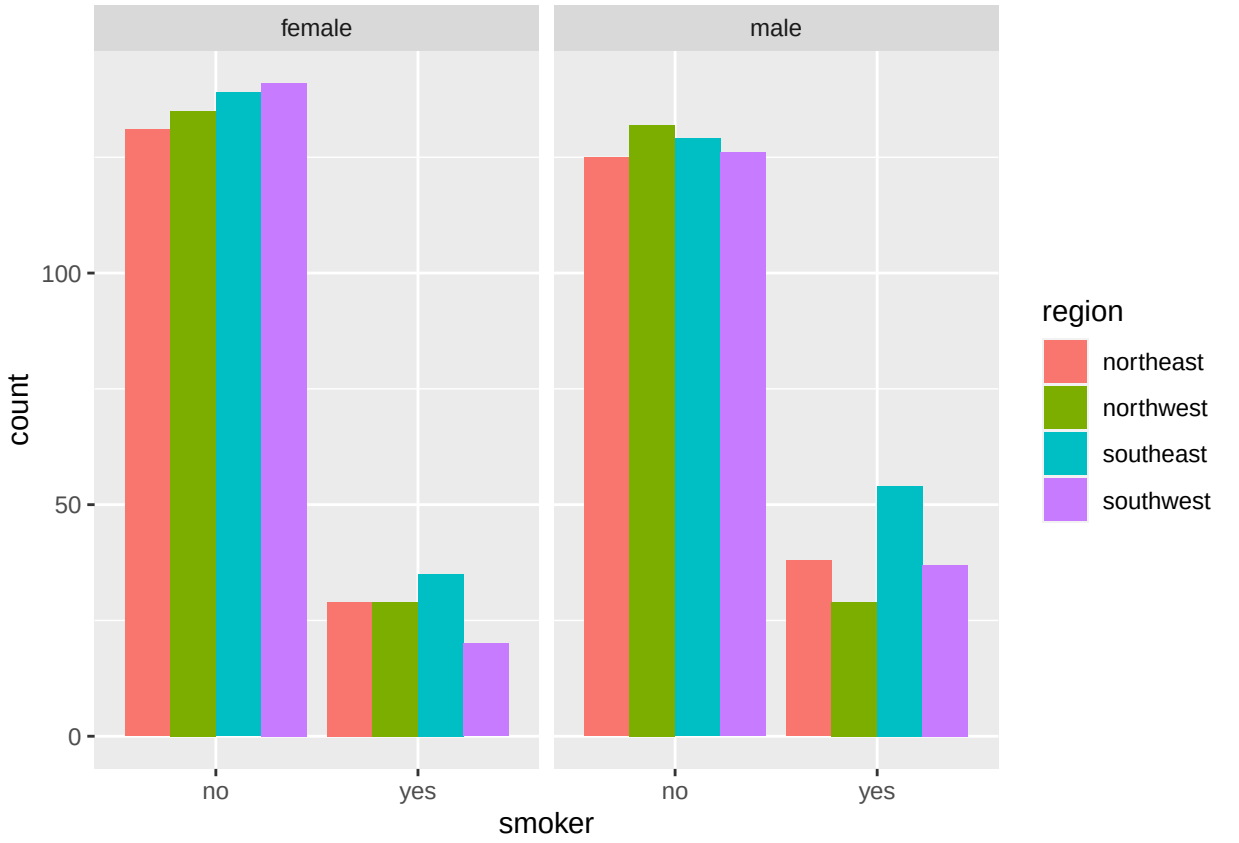


Observations

1. Only Southwest region has not crossed the charge of 55000.
2. Southeast region is the most paying as the range of Q3 and Q2 is high compared to others.
3. There are many outliers in other regions compared to southeast region.

Q *Which region has most no. of smokers based on gender*

```
data_2 %>% ggplot(aes(smoker))+geom_bar(aes(fill=region),position = "dodge")+facet_wrap(~sex)
```



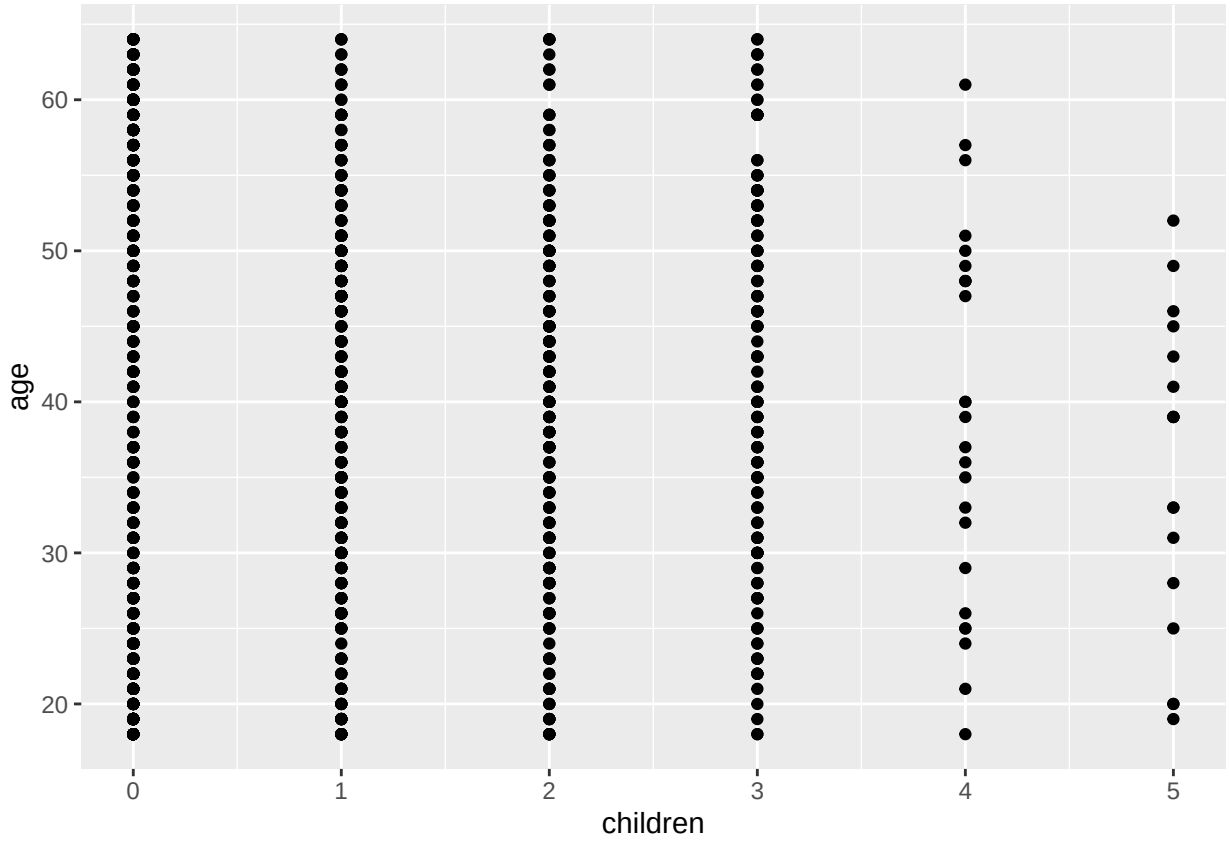
Observations
1. Southeast region has the maximum people who are smokers followed by northeast region(both male and female).
2. Northwest has least no. of male smokers and southwest has least no. of female smokers.

Q *Linear regression b/w children and charges*

```
linearity(data_0,"children","age")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```



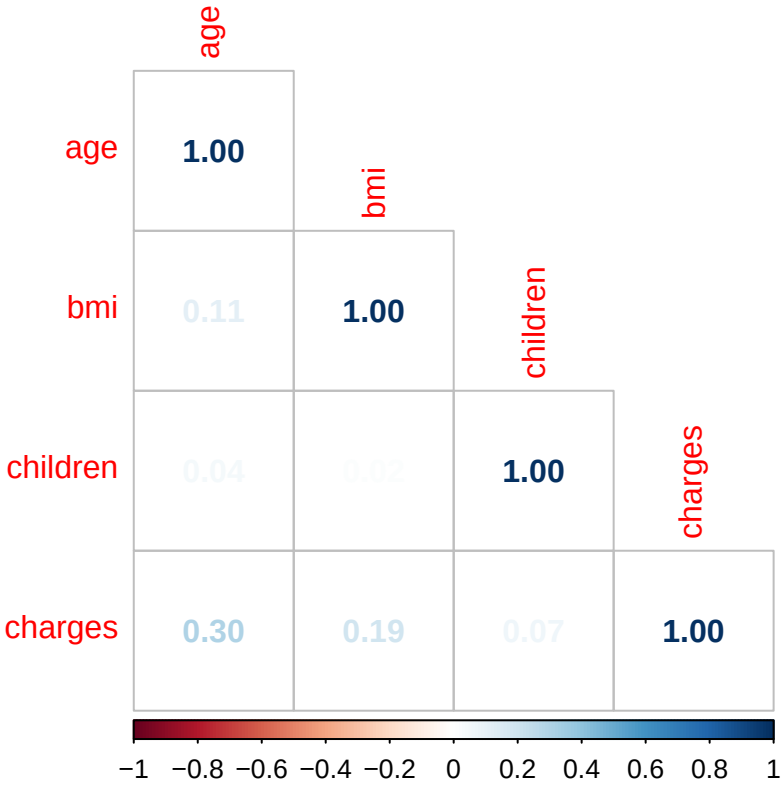

```
##                age
## children 0.042469
```

Observations

1. There is a no linear relationship b/w no. of children and charges and correlation is ~0. so it's better to not consider this variable.

Checking for collinearity

```
# getting the indexes of all numerical columns
num_idx <- which(!grepl("factor|character",sapply(data_2,class)))
M <- cor(data_2[num_idx])
corrplot(M,method = "num",type = "lower")
```



Observations

1. There is weak correlation(s) among independent variables. So no need to remove any columns or do any step wise relation check for variables.

Splitting and Modelling

```
sample <- sample.split(data_2$charges,.7)
train <- data_2 %>% filter(sample==TRUE)
test <- data_2 %>% filter(sample==FALSE)
#verifying the dimensions
dim(test) #test set
```

```
## [1] 399   7
```

```
dim(train) #train set
```

```
## [1] 930   7
```

Model

```
model <- lm(charges~age+bmi,train)
predict_train <- predict(model,train[c("age","bmi")])
predict_test <- predict(model,test[c("age","bmi")])
```

Evaluating the performance of the model

```
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13002  -7064  -4916   7636  47895
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8288.72    2081.90  -3.981 7.39e-05 ***
## age           250.26      26.12   9.583  < 2e-16 ***
## bmi           386.18      62.26   6.202 8.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11230 on 927 degrees of freedom
## Multiple R-squared:  0.1361, Adjusted R-squared:  0.1342
## F-statistic: 73.01 on 2 and 927 DF,  p-value: < 2.2e-16
```

```
# RMSE of test
RMSE(predict_test,test$charges)
```

```
## [1] 11432.84
```

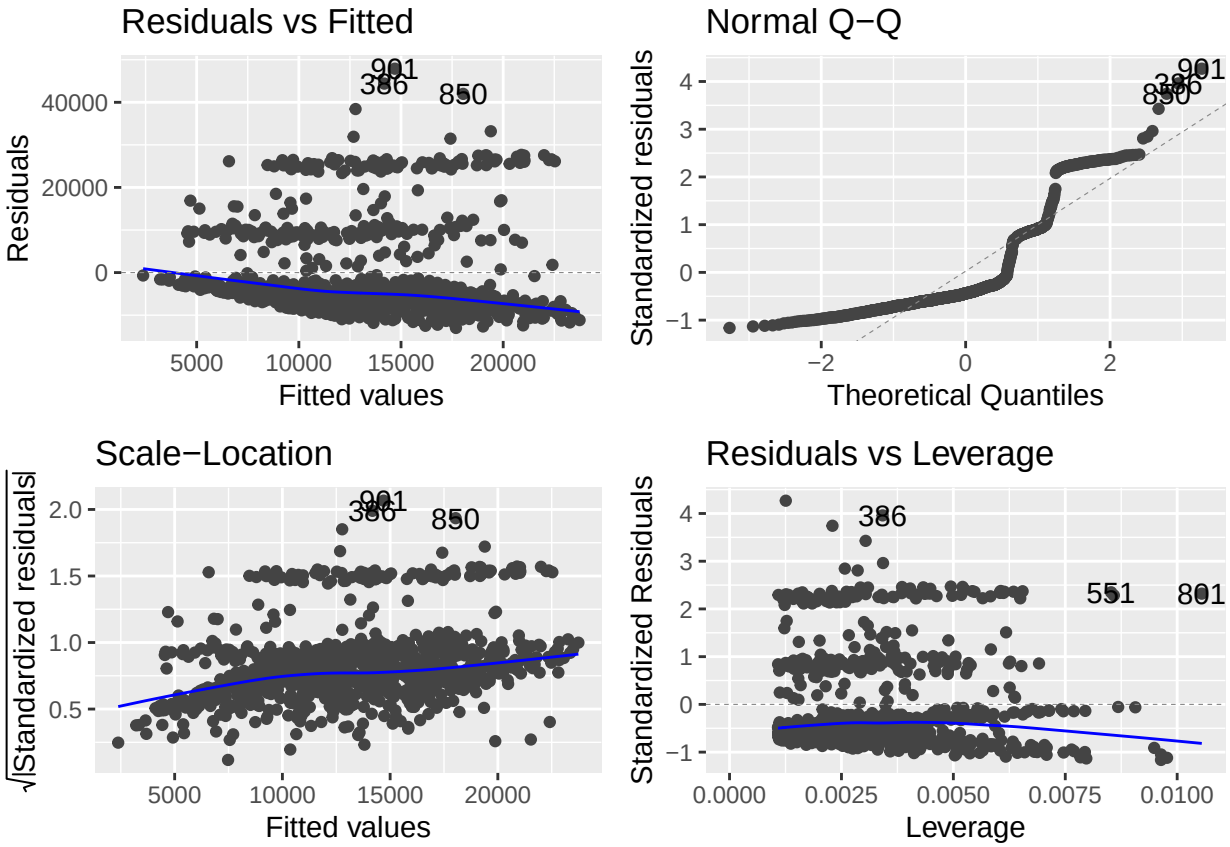
```
# RMSE of train
RMSE(predict_train,train$charges)
```

```
## [1] 11211.54
```

```
# R2 Score
R2_Score(predict_test,test$charges)
```

```
## [1] 0.06548813
```

```
library(ggfortify)
autoplot(model)
```



Observations

1. The r-squared score is low as we see that the assumptions are not fulfilled for linear regression.
2. While looking at the correlation b/w *each independent variable* and *the dependent variable* which we found to be very low and we didn't consider one of the variables out of a total of 3 variables which leave us with only 2.
3. The model would have performed better if it satisfied the linearity assumption.
4. It's highly unlikely to find dataset which satisfies the assumptions of the linear regression.
5. So it's much better to move or to check tree based models.
6. I specially worked on this data to specifically show how tedious can be working with linear regression model but if a good dataset is found it too will perforam well.