caoston-2

Hussin. Almustafa

2022-10-07

Medical Cost Personal Dataset

Introduction :

For this project, I will be applying machine learning techniques on the/ Medical Cost Personal Dataset/The data are downloadable from The link to the data set

the data explain the cost of a small sample of USA population Medical Insurance Cost based on some attributes depicted on "Content".my github repo

the Goal of the project : The purpose of this project is to predict the medical expenses .

Methods:

after downloading the data i,m going to explore it, extract the Information from it , and cleaning the data ,then prepare the data for analysis ,Visualization ,then start Training and resampling several models.

##1-data loding:

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
library(dplyr)
```

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

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

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

```
library(caretEnsemble)
library(kernlab)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.7     v stringr 1.4.1
## v tidyr   1.2.0     v forcats 0.5.1
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::alpha()    masks kernlab::alpha()
## x ggplot2::autoplot() masks caretEnsemble::autoplot()
## x purrr::cross()      masks kernlab::cross()
## x tidyr::expand()     masks Matrix::expand()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x tidyr::pack()       masks Matrix::pack()
## x tidyr::unpack()     masks Matrix::unpack()
```

```
library(reshape2)
```

```
##
## Attaching package: ’reshape2’
##
## The following object is masked from ’package:tidyr’:
##
##     smiths
```

```
library(RColorBrewer)
library(data.table)
```

```
##
## Attaching package: ’data.table’
##
## The following objects are masked from ’package:reshape2’:
##
##     dcast, melt
##
## The following object is masked from ’package:purrr’:
##
##     transpose
##
## The following objects are masked from ’package:dplyr’:
##
##     between, first, last
```

```
library(caret)
```

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

```
##
## Attaching package: ’caret’
##
## The following object is masked from ’package:purrr’:
##
##     lift
```

2- Data Exploration:

```
set.seed(123)
##Reading Data
dat <- read.csv("C:/Users/Almustafa/Documents/med.csv")

str(dat)
```

```
## ’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 see the data is a data frame with 1338 observations and 7 variables, the variable - charges - as response variable

The column descriptions look like this:

age: age of primary beneficiary

sex: 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 ˆ 2)

children: Number of children

smoker: Smoking Yes or No

region: the beneficiary’s residential area in the US . charges: Individual medical costs billed by health insurance.

check the summary of variables :

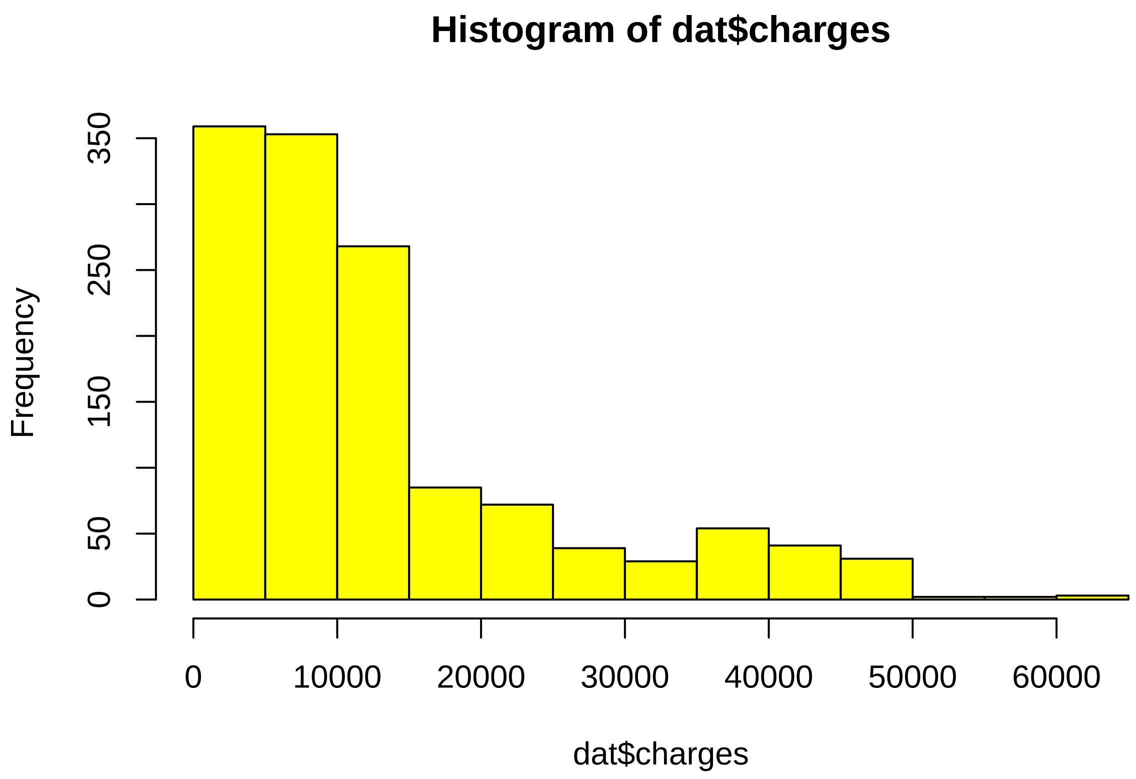
```
summary(dat)
```

```
##       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
```

we can see that there is a difference between mean & median in the variable charges

##Let's see the distribution of charges

```
hist(dat$charges,col = "yellow")
```



the plot shows that the majority of people in our data have yearly medical expenses between 0 - 15,000$.

look at the categorical variable distribution:

```
table(dat$sex)
```

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

```
table(dat$smoker)
```

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

```
table(dat$region)
```

```
##
## northeast northwest southeast southwest
##       324       325       364       325
```

find the relationship among variables

```
cor(dat[c("age","bmi","children","charges")])
```

```
##                age       bmi   children    charges
## age      1.0000000 0.1092719 0.04246900 0.29900819
## bmi      0.1092719 1.0000000 0.01275890 0.19834097
## children 0.0424690 0.0127589 1.00000000 0.06799823
## charges  0.2990082 0.1983410 0.06799823 1.00000000
```

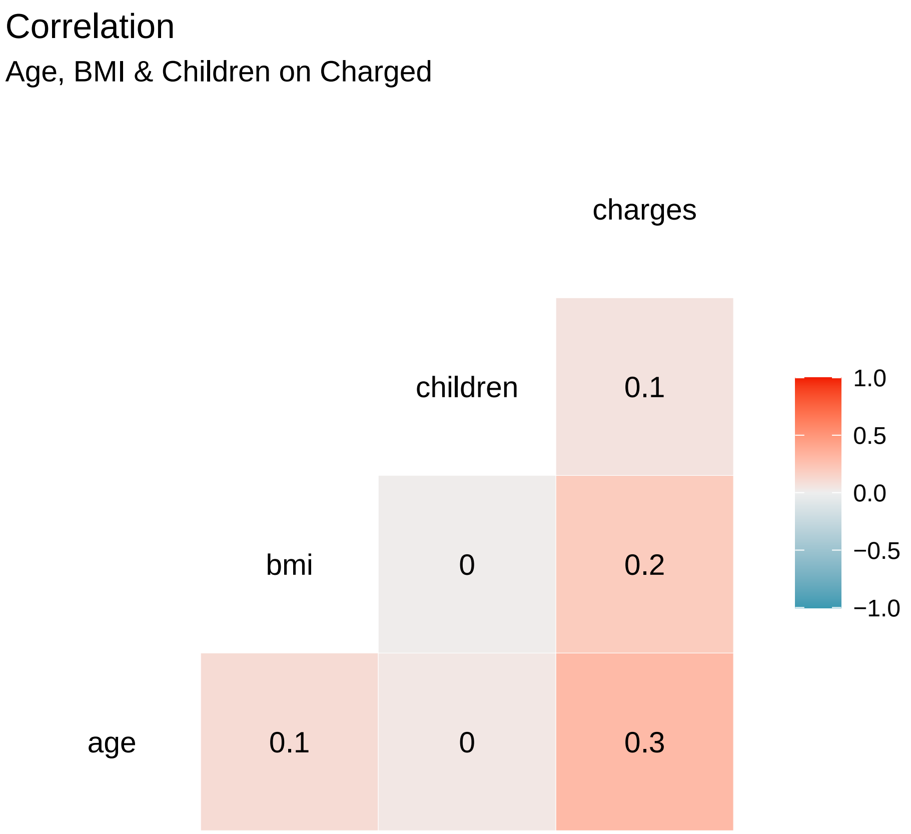
there is a perfect correlation between a variable .There appears to be a weak positive association between age and BMI, which means that with age, body mass tends to increase. There is also a moderately positive relationship between age and expenses, BMI and expenses, and children and expenses. These associations indicate that with increasing age, body mass and number of children, the expected cost of insurance increases

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggcorr(dat, label = T, color = "black", size = 4)+
  labs(title = "Correlation ",
       subtitle = "Age, BMI & Children on Charged ")
```

```
## Warning in ggcorr(dat, label = T, color = "black", size = 4): data in column(s)
## 'sex', 'smoker', 'region' are not numeric and were ignored
```



the plot shows that age has the highest correlation with charges

3- data cleaning :

```
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
sum(is.na(dat))
```

```
## [1] 0
```

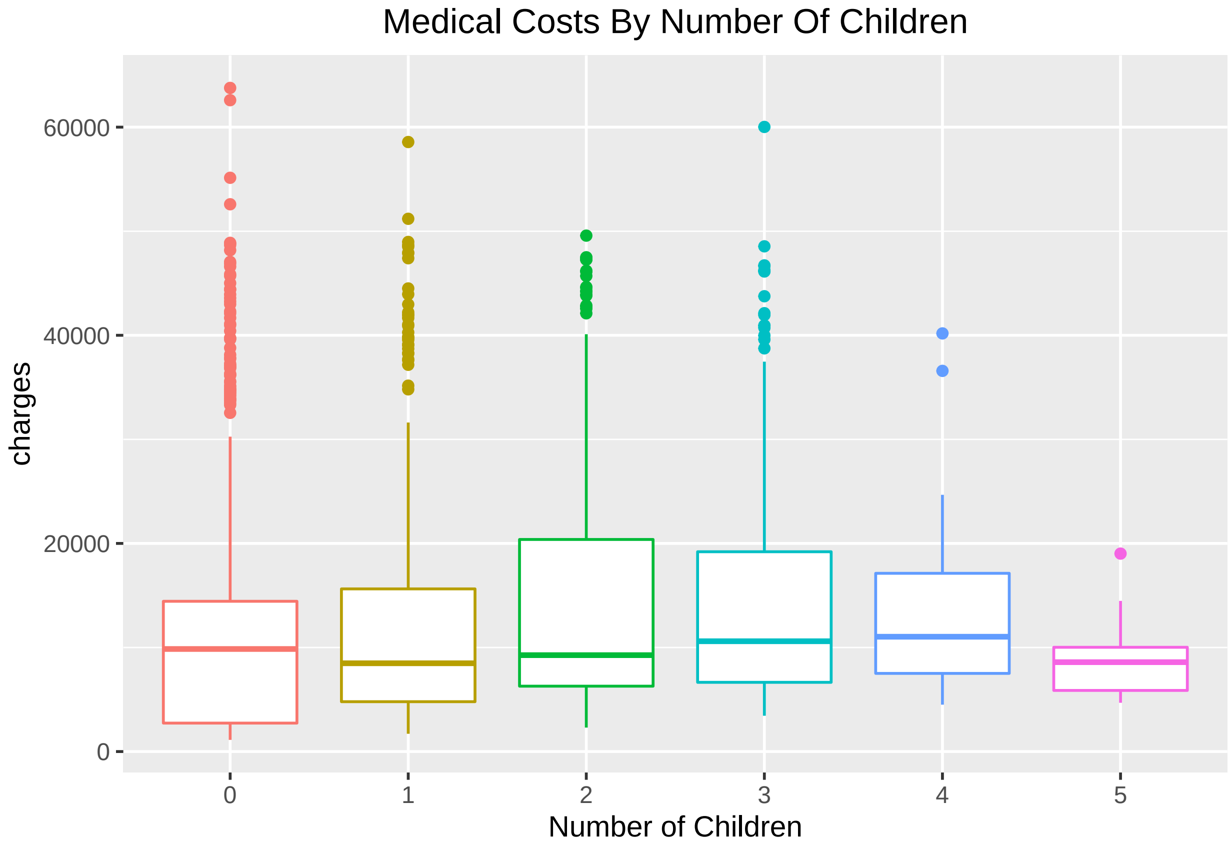
there is no Na in our data,

removing empty rows and col,s:

```
dat <-dat %>% remove_empty(whic=c("rows"))
dat <-dat %>% remove_empty(whic=c("cols"))
```

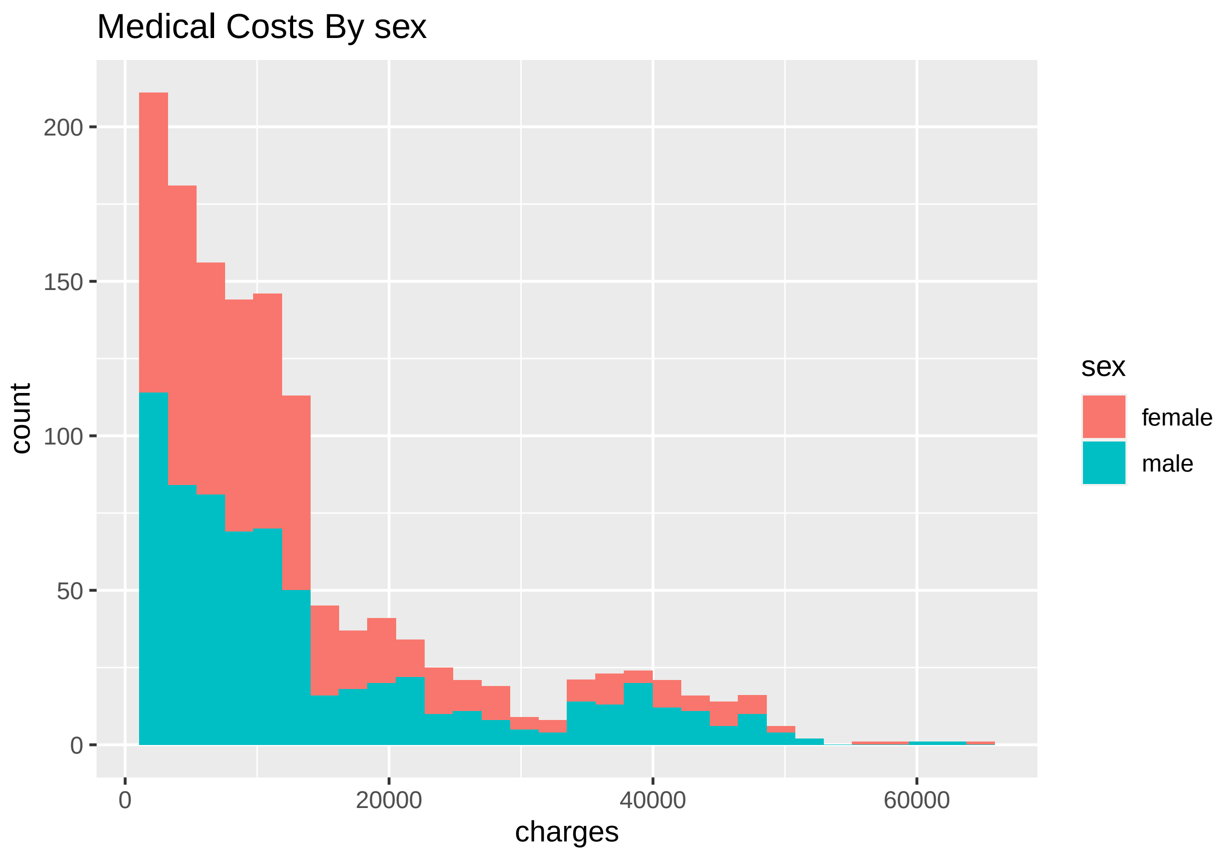
data visualisation :

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



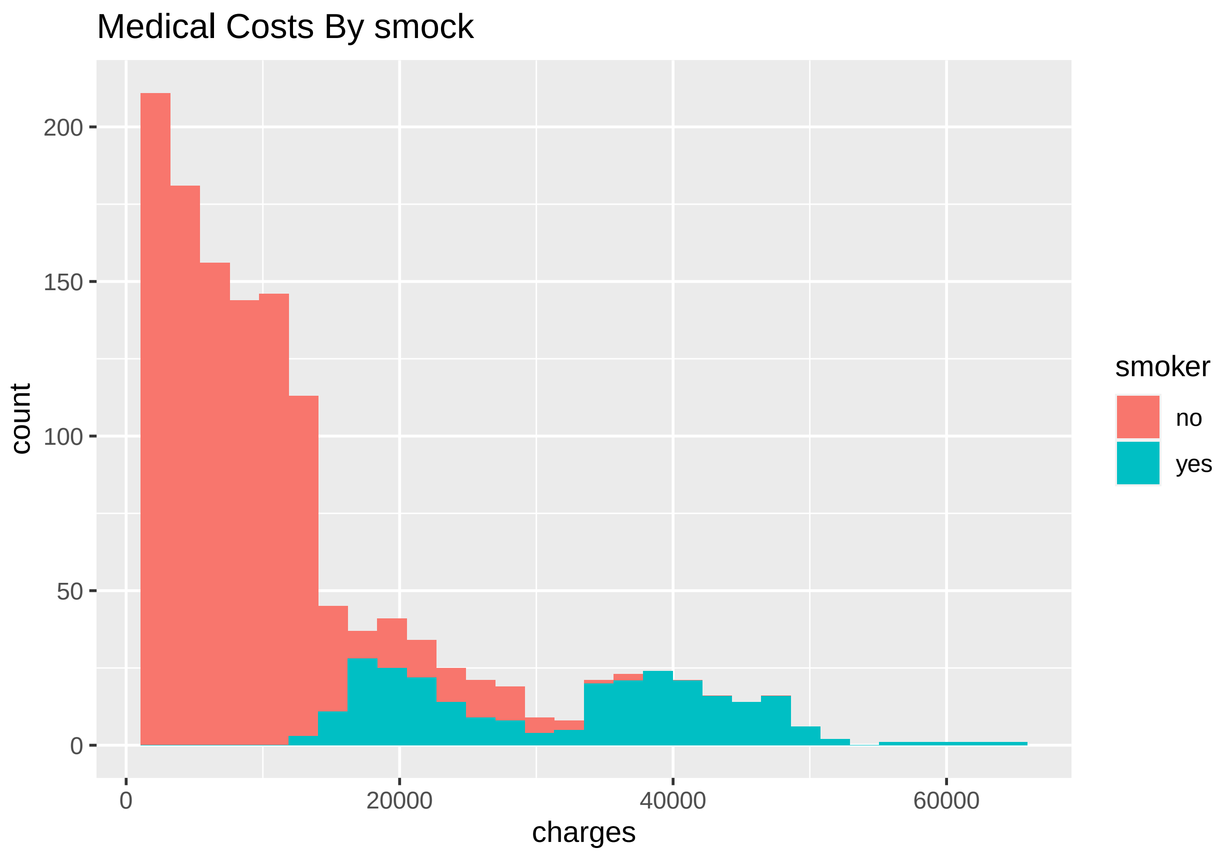
```
   dat%>%
  ggplot(aes(charges,fill=sex,binwidth=30))+
  geom_histogram()+
  labs(title = "Medical Costs By sex")
```

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



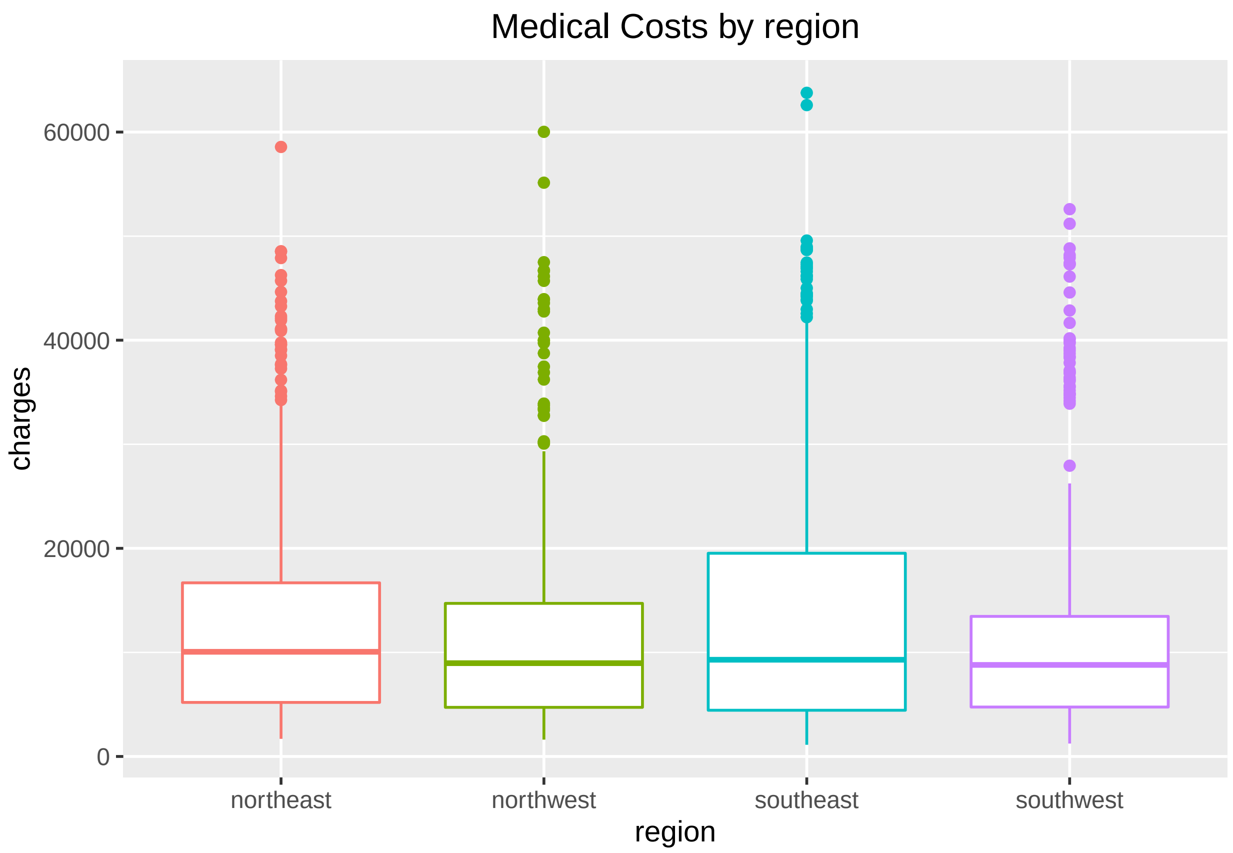
```
  dat%>%
  ggplot(aes(charges,fill=smoker, binwidth=30))+
  geom_histogram()+
  labs(title = "Medical Costs By smock")
```

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

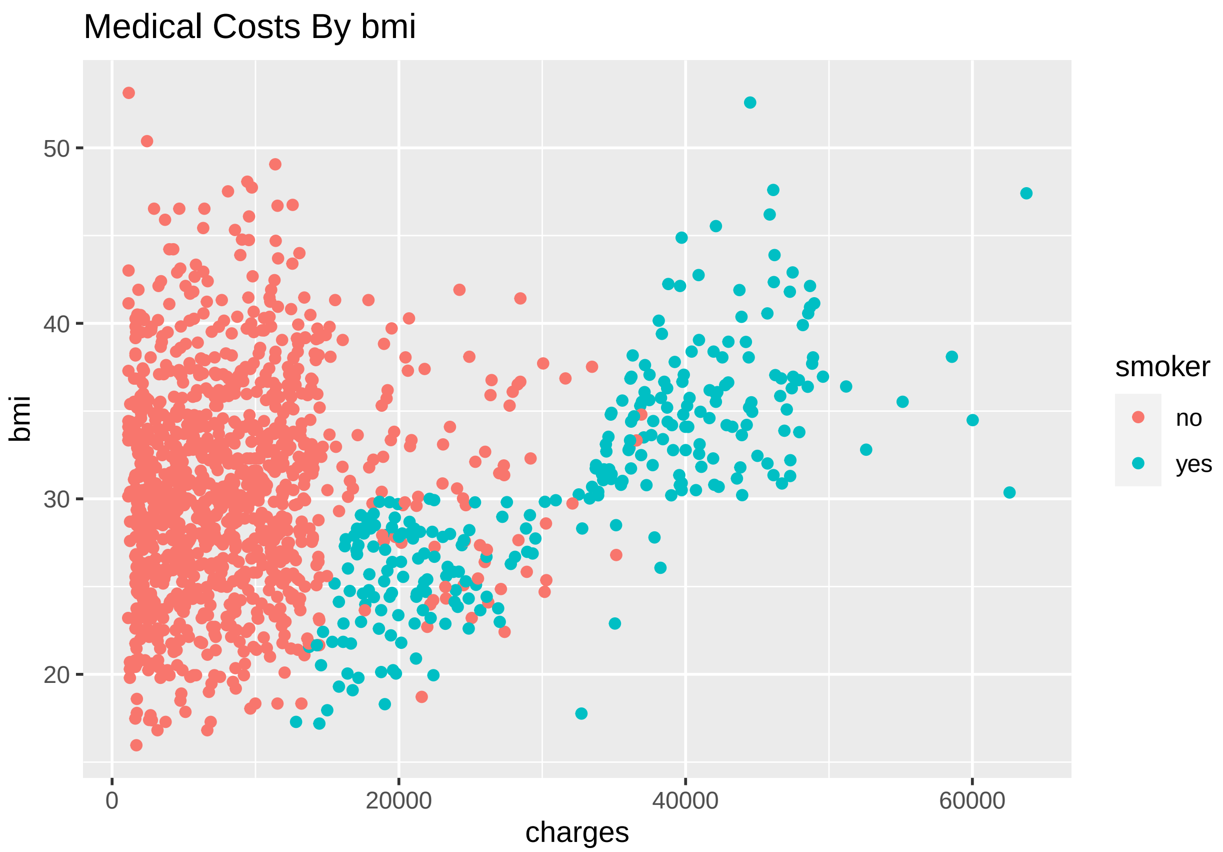


Smokers have more charges than non-smokers.

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



```
dat%>%
 ggplot(aes(charges,bmi,color= smoker,fill=smoker))+
 geom_point()+
 labs(title = "Medical Costs By bmi")
```



modeling :

```
library(caret)
set.seed(2002)
y<- dat$charges

test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)

dat_train <- dat %>% slice(-test_index)
dat_test <- dat  %>% slice(test_index)
```

model-1 Linear Regression:

```
m <- mean(dat_train$charges)
m
```

```
## [1] 13251.17
```

Our root mean squared error is:

```
sqrt(mean((m - dat_test$charges)^2))
```

```
## [1] 12351.59
```

using the function lm()To fit a linear regression model to data

```
fit <- lm(charges~ ., data = dat_train)
fit$coef
```

```
##     (Intercept)             age         sexmale             bmi        children
##     -12396.0881        256.8043       -374.6309        354.1554        469.4450
##       smokeryes regionnorthwest regionsoutheast regionsouthwest
##      23742.9607       -118.3729      -1093.7706       -704.4017
```

```
 y_hat <- predict(fit, dat_test)
sqrt(mean((y_hat - dat_test$charges)^2))
```

```
## [1] 6024.599
```

We can see that this does indeed provide an improvement over our guessing approach.

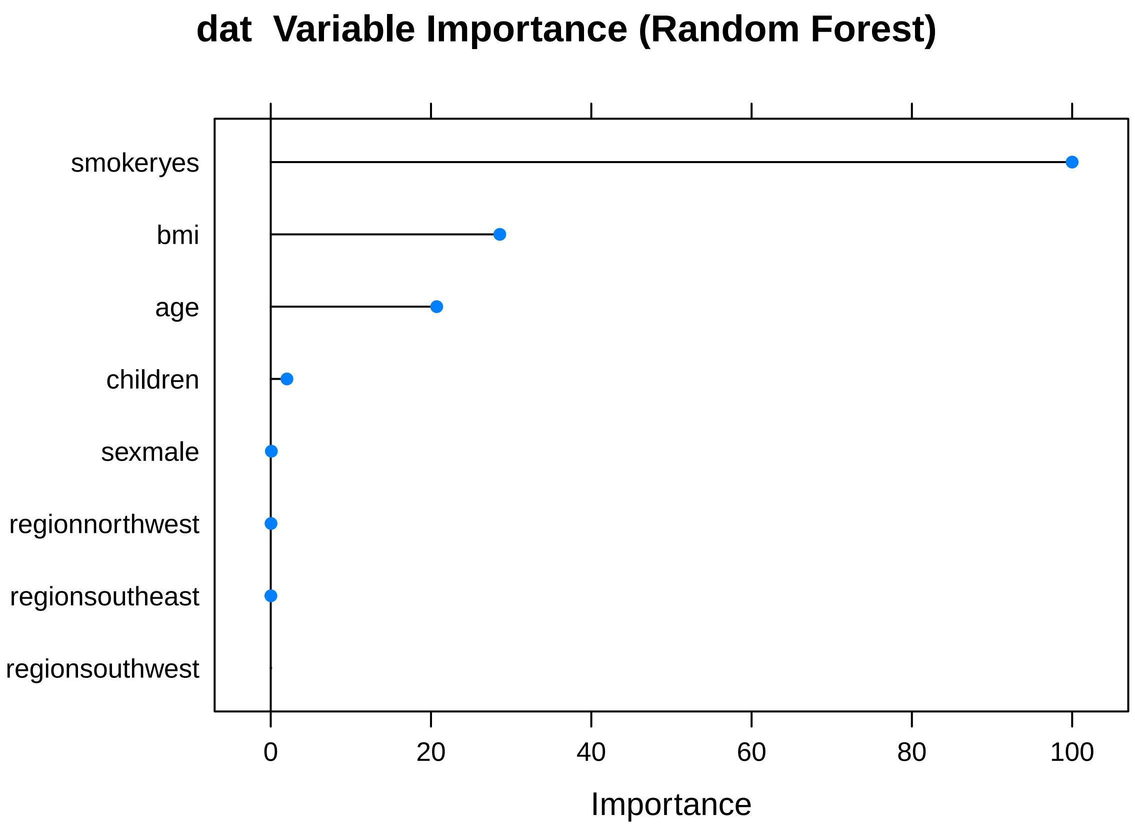
model-2 Random Forest :

```
rf <- train(charges ~., data = dat_train, method = "rf")
rf
```

```
## Random Forest
##
## 1070 samples
##    6 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1070, 1070, 1070, 1070, 1070, 1070, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     5345.568  0.8306883  3673.149
##   5     4894.491  0.8326070  2761.640
##   8     5106.759  0.8192830  2884.123
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 5.
```

Variable importance plot from random forest

```
varimp_RF <- varImp(rf)
plot(varimp_RF, main = "dat  Variable Importance (Random Forest)")
```



we use the random forest to predict on our test data.

```
fit <- predict(rf,dat_test)
fit[1:8]
```

```
##         1         2         3         4         5         6         7         8 
##  9065.898  3360.915  2727.388  4193.344 14689.996 37739.937  4264.770  5125.904
```

by comparing the models we can see that we get better performance with higher R-Squared and lower MAE. using random forest algorithm .