```r
#1 Loading our data as a dataframe

library(tidyverse)
```

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

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ purrr   0.3.4
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.0      ✔ stringr 1.4.1
## ✔ readr   2.1.2      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```r
hmodata<-data.frame(read_csv("Data.csv"))
```

```
## Rows: 7582 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (8): smoker, location, location_type, education_level, yearly_physical, ...
## dbl (6): X, age, bmi, children, hypertension, cost
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#2 Viewing basic attributes of our dataset

str(hmodata)
```

```
## 'data.frame':    7582 obs. of  14 variables:
##  $ X              : num  1 2 3 4 5 7 9 10 11 12 ...
##  $ age            : num  18 19 27 34 32 47 36 59 24 61 ...
##  $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
##  $ children       : num  0 1 3 0 0 1 2 0 0 0 ...
##  $ smoker         : chr  "yes" "no" "no" "no" ...
##  $ location       : chr  "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...
##  $ location_type  : chr  "Urban" "Urban" "Urban" "Country" ...
##  $ education_level: chr  "Bachelor" "Bachelor" "Master" "Master" ...
##  $ yearly_physical: chr  "No" "No" "No" "No" ...
##  $ exercise       : chr  "Active" "Not-Active" "Active" "Not-Active" ...
##  $ married        : chr  "Married" "Married" "Married" "Married" ...
##  $ hypertension   : num  0 0 0 1 0 0 0 1 0 0 ...
##  $ gender         : chr  "female" "male" "male" "male" ...
##  $ cost           : num  1746 602 576 5562 836 ...
```

```r
summary(hmodata)
```

```
##        X                  age             bmi           children    
##  Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000  
##  1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000  
##  Median :    24916   Median :39.00   Median :30.50   Median :1.000  
##  Mean   :   712602   Mean   :38.89   Mean   :30.80   Mean   :1.109  
##  3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.:2.000  
##  Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000  
##                                      NA's   :78                     
##     smoker            location         location_type      education_level   
##  Length:7582        Length:7582        Length:7582        Length:7582       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  yearly_physical      exercise           married           hypertension   
##  Length:7582        Length:7582        Length:7582        Min.   :0.0000  
##  Class :character   Class :character   Class :character   1st Qu.:0.0000  
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000  
##                                                           Mean   :0.2005  
##                                                           3rd Qu.:0.0000  
##                                                           Max.   :1.0000  
##                                                           NA's   :80      
##     gender               cost      
##  Length:7582        Min.   :    2  
##  Class :character   1st Qu.:  970  
##  Mode  :character   Median : 2500  
##                     Mean   : 4043  
##                     3rd Qu.: 4775  
##                     Max.   :55715  
## 
```

```r
head(hmodata,5)
```

```
##   X age    bmi children smoker      location location_type education_level
## 1 1  18 27.900        0    yes   CONNECTICUT         Urban        Bachelor
## 2 2  19 33.770        1     no  RHODE ISLAND         Urban        Bachelor
## 3 3  27 33.000        3     no MASSACHUSETTS         Urban          Master
## 4 4  34 22.705        0     no  PENNSYLVANIA       Country          Master
## 5 5  32 28.880        0     no  PENNSYLVANIA       Country             PhD
##   yearly_physical   exercise married hypertension gender cost
## 1              No     Active Married            0 female 1746
## 2              No Not-Active Married            0   male  602
## 3              No     Active Married            0   male  576
## 4              No Not-Active Married            1   male 5562
## 5              No Not-Active Married            0   male  836
```

```
tail(hmodata,5)
```

```
##          X age    bmi children smoker     location location_type
## 7578 13023  63 30.875        3    yes   NEW JERSEY         Urban
## 7579 54813  53 46.700        2     no PENNSYLVANIA         Urban
## 7580 64221  42 28.310        3    yes PENNSYLVANIA         Urban
## 7581 74732  33 27.000        2     no PENNSYLVANIA       Country
## 7582 13531  20 28.785        0     no     NEW YORK         Urban
##        education_level yearly_physical   exercise     married hypertension
## 7578 No College Degree              No Not-Active     Married            0
## 7579          Bachelor             Yes Not-Active Not_Married            0
## 7580          Bachelor              No     Active     Married            0
## 7581          Bachelor              No Not-Active Not_Married            0
## 7582          Bachelor              No     Active     Married            0
##      gender  cost
## 7578   male 25414
## 7579 female  6881
## 7580   male  9153
## 7581   male  4576
## 7582 female   270
```

```
#3 Viewing cost statistics to decide what cost to consider value as expensive
min(hmodata$cost)
```

```
## [1] 2
```

```
max(hmodata$cost)
```

```
## [1] 55715
```

```
mean(hmodata$cost)
```

```
## [1] 4042.961
```

```
median(hmodata$cost)
```

```
## [1] 2500
```

```
quantile(hmodata$cost)
```

```
##    0%   25%   50%   75%  100% 
##     2   970  2500  4775 55715
```

```
#4 Creation of a new column "cost_status" to categorize costs as 1,0 to get expensive based on o
ur prior analysis on cost statistics

hmodata$cost_status<- with(
hmodata, ifelse(cost>4800,"TRUE","FALSE"))
hmodata$cost_status<-as.factor(hmodata$cost_status)
```

```
#5 Checking for null values in the columns of the dataframe which have numeric data type

sum(is.na(hmodata$age))
```

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

```
sum(is.na(hmodata$bmi))#We see 78 null values
```

```
## [1] 78
```

```
sum(is.na(hmodata$children))
```

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

```
sum(is.na(hmodata$hypertension))#We see 80 null values
```

```
## [1] 80
```

```
sum(is.na(hmodata$cost))
```

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

```
#6 Data cleaning using na_interpolation on the columns which have null values

library(imputeTS)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
hmodata$bmi<-na_interpolation(hmodata$bmi)
hmodata$hypertension<-na_interpolation(hmodata$hypertension)
```

```
#7 Checking again for null values

sum(is.na(hmodata$age))
```

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

```
sum(is.na(hmodata$bmi))#We see 0 null values
```

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

```
sum(is.na(hmodata$children))
```

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

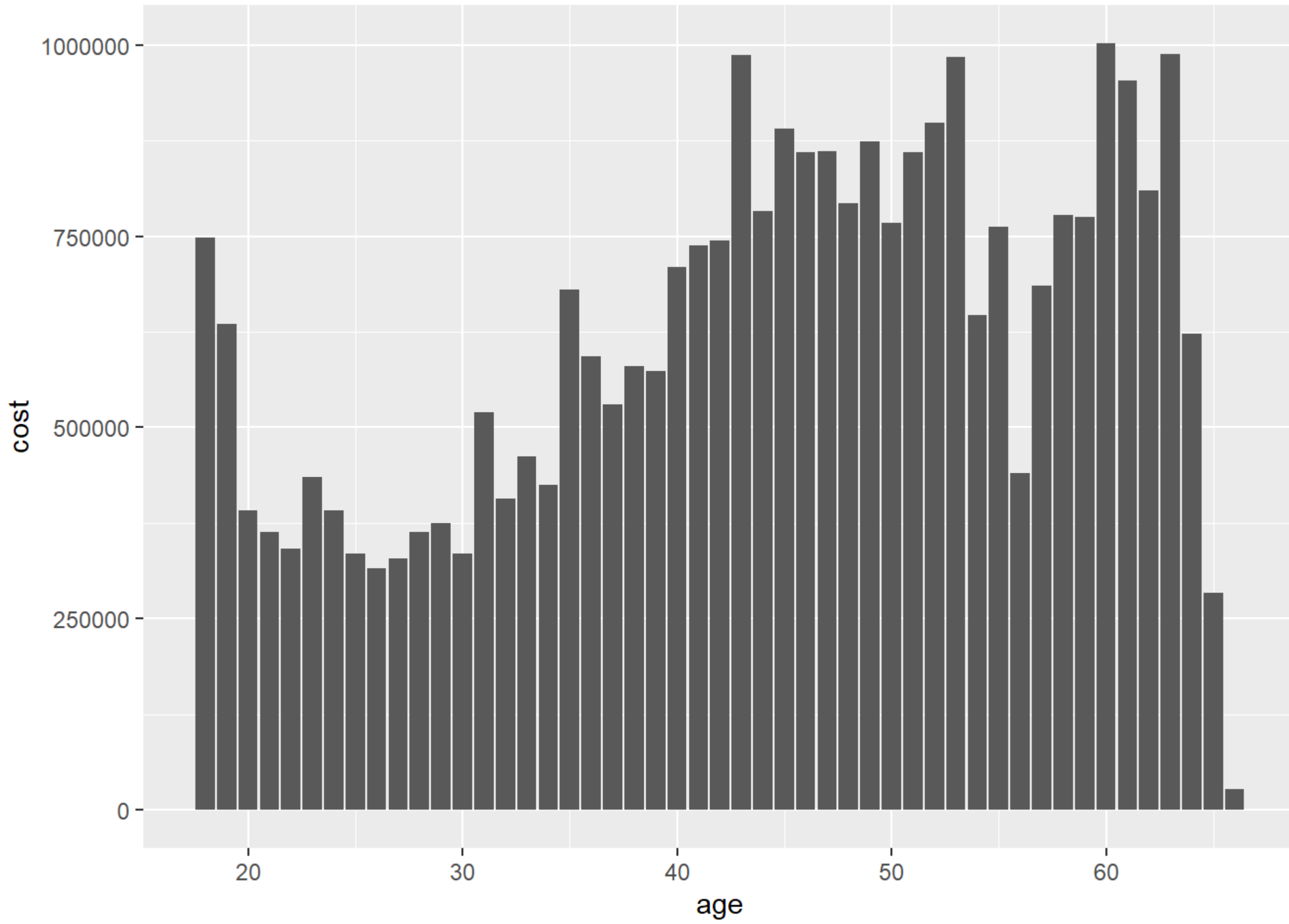
```
sum(is.na(hmodata$hypertension))#We see 0 null values
```

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

```
sum(is.na(hmodata$cost))
```

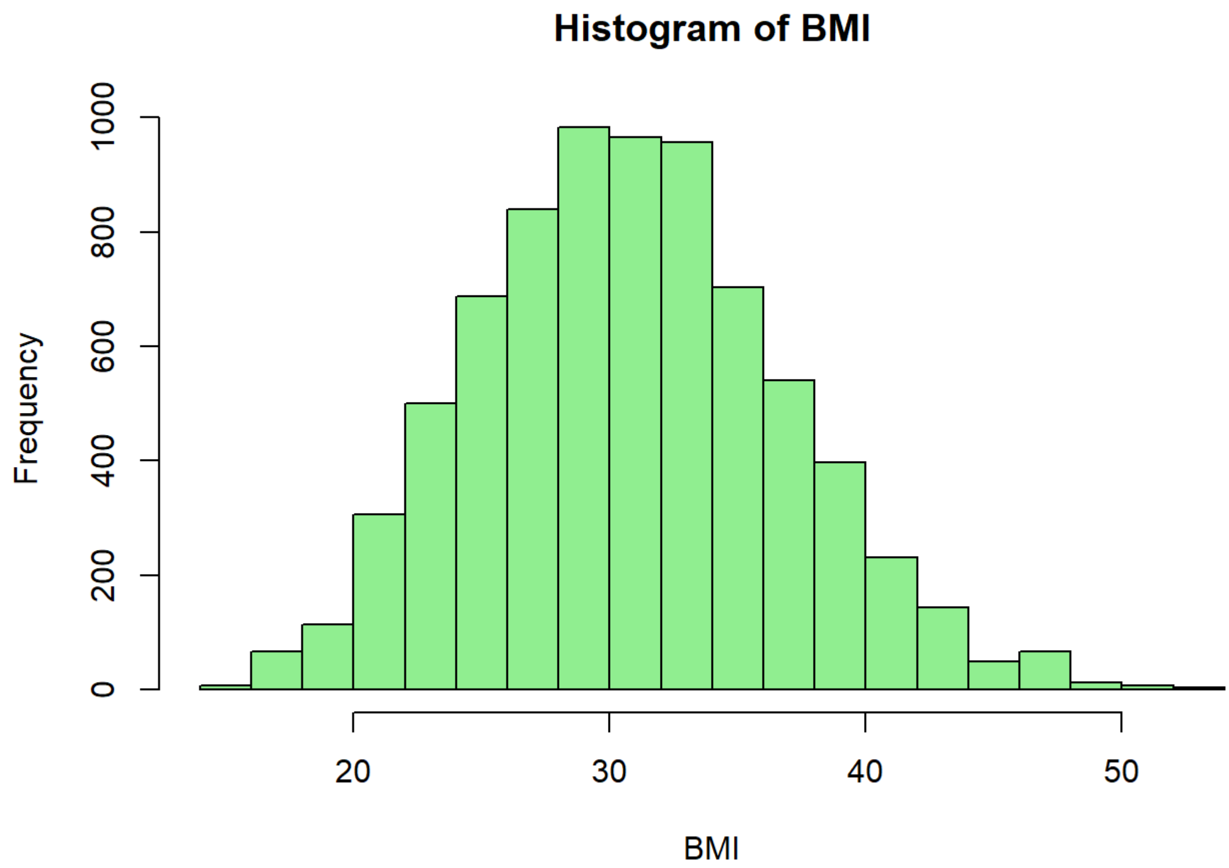
```
## [1] 0
```

```
#Analyzing dataset and visualizing for understanding

#8 Age vs Cost barplot
ggplot(hmodata,aes(x=age, y=cost)) +geom_bar(stat="identity")
```

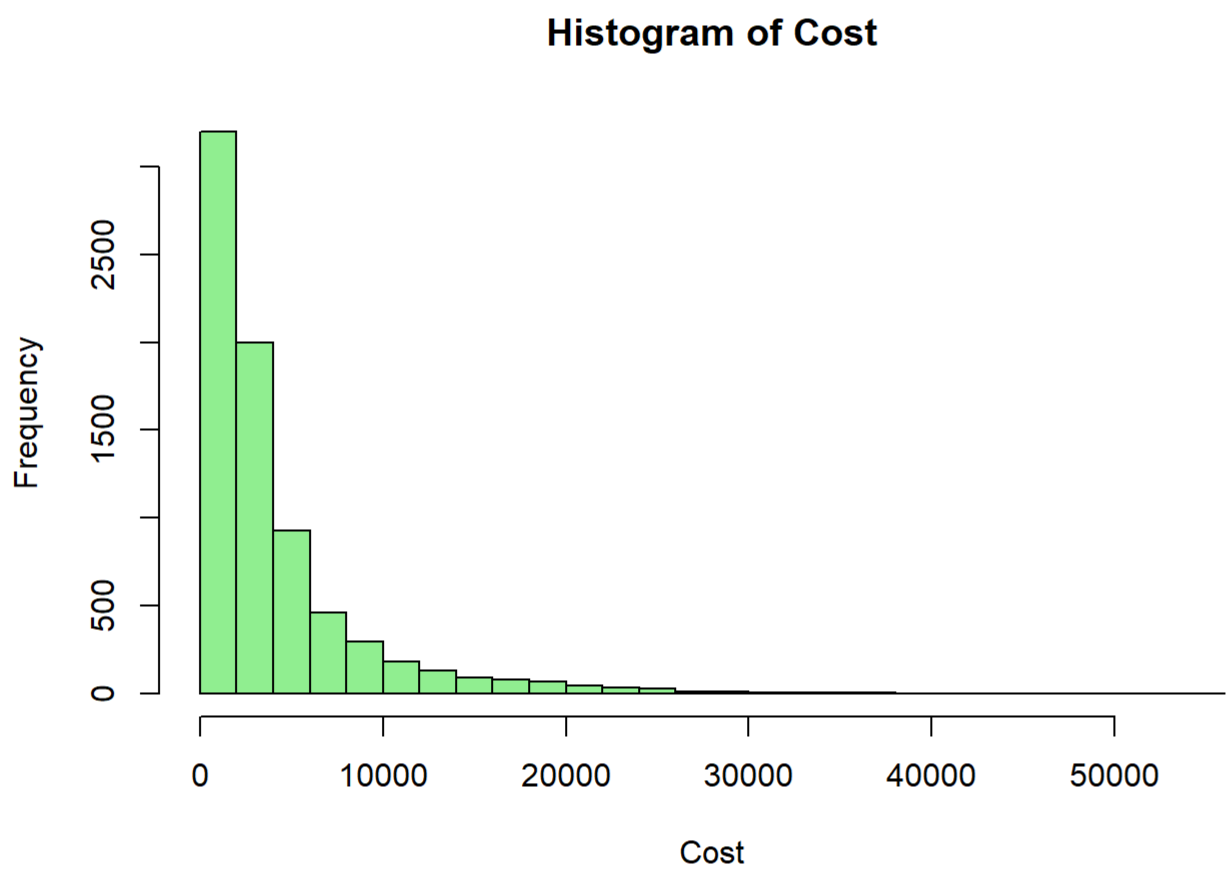



```
#Costs are initially high in teen years, and then dip down, and then gradually increase with age
```

```
#9 Generating histograms to see distribution of quantitative variables

hist(hmodata$bmi, breaks = 15, col = "light green", main = "Histogram of BMI", xlab = "BMI", yla
b = "Frequency")
```

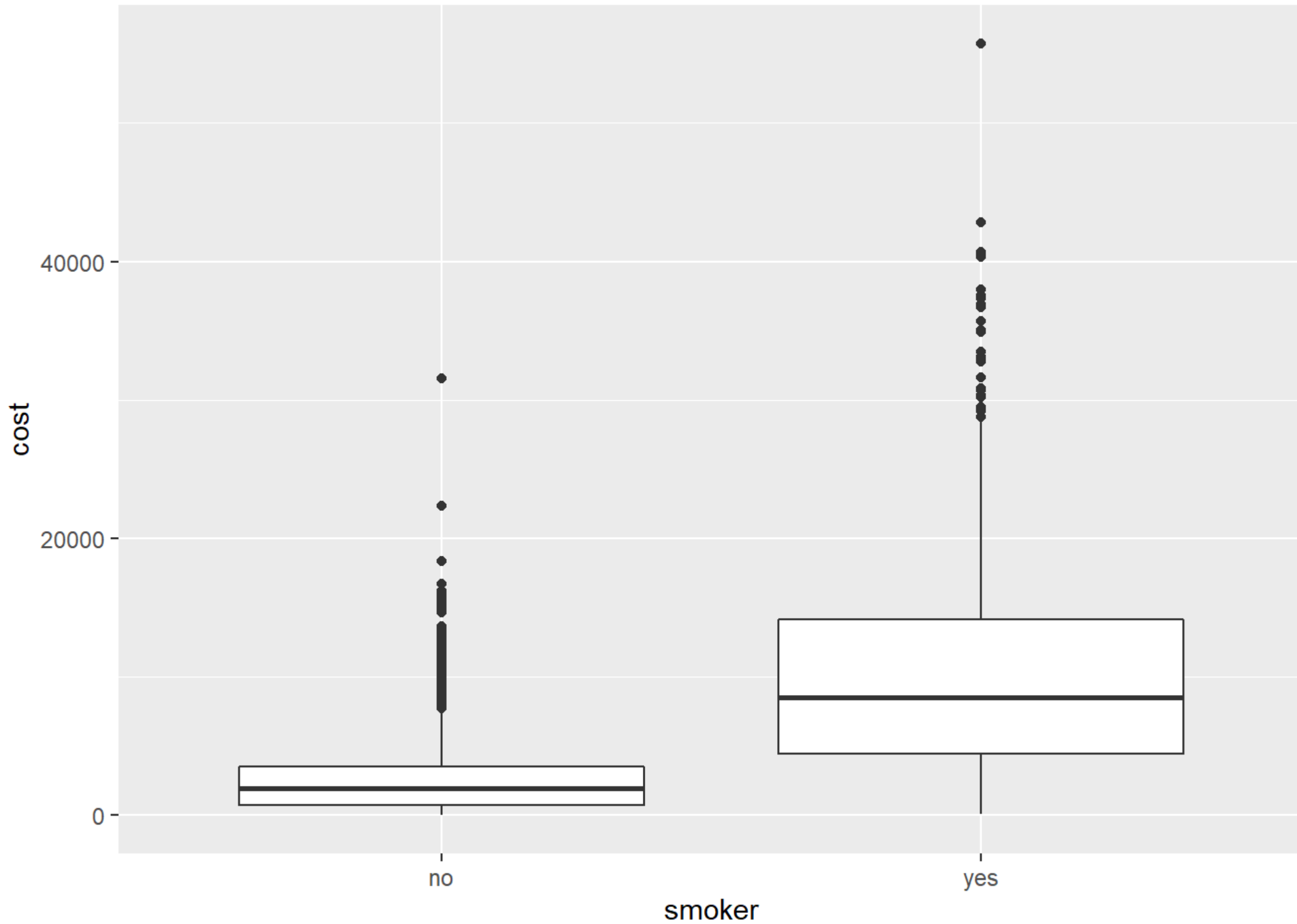



```
#We see a normal distribution here

hist(hmodata$cost, breaks = 20, col = "light green", main = "Histogram of Cost", xlab = "Cost",
ylab = "Frequency")
```



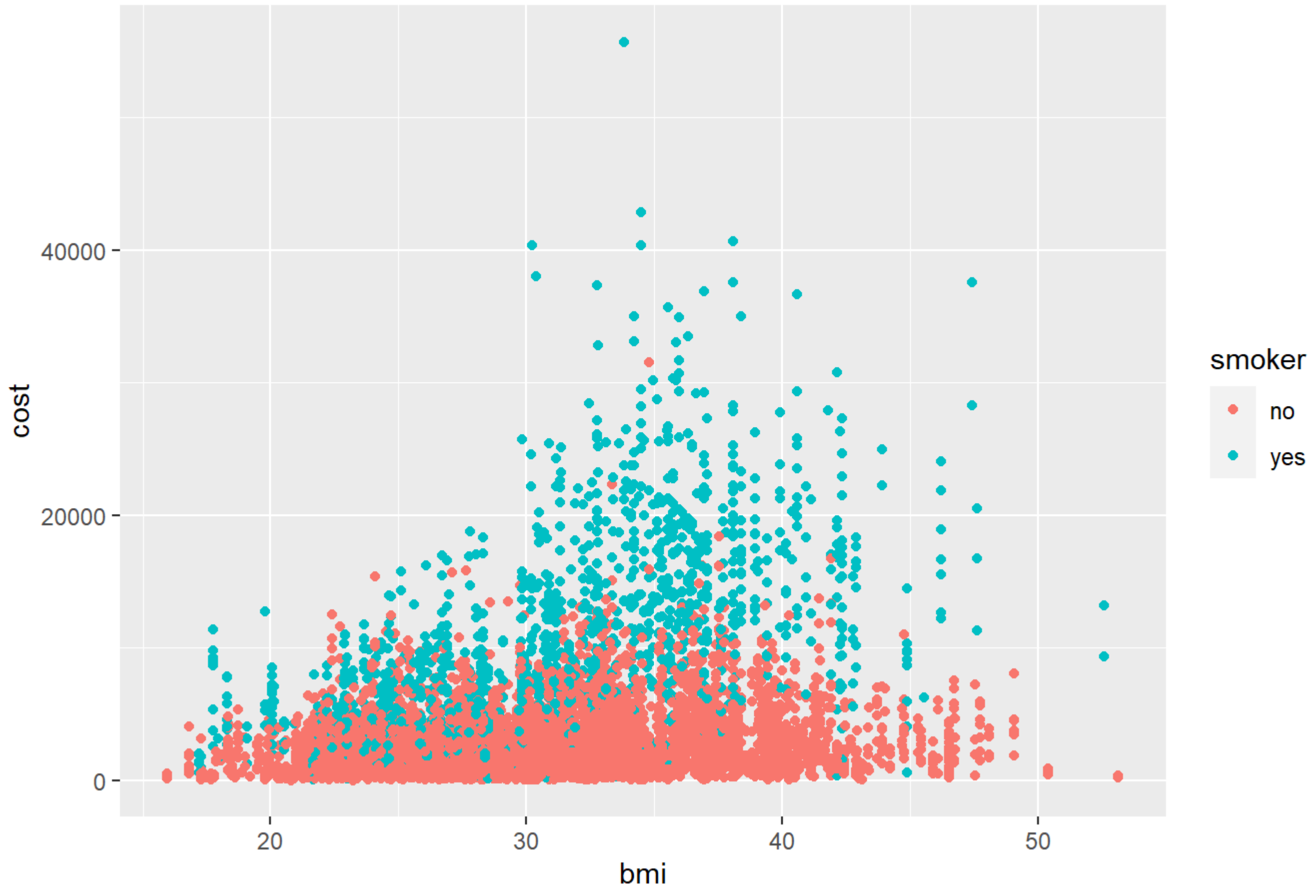

```
#We see a right skewed distribution, individuals with significantly higher cost have less freque
ncy
```

```
#10 Box plots to see any outliers


box_plot1 <- ggplot(hmodata, aes(x = smoker, y = cost)) + geom_boxplot()
box_plot1
```

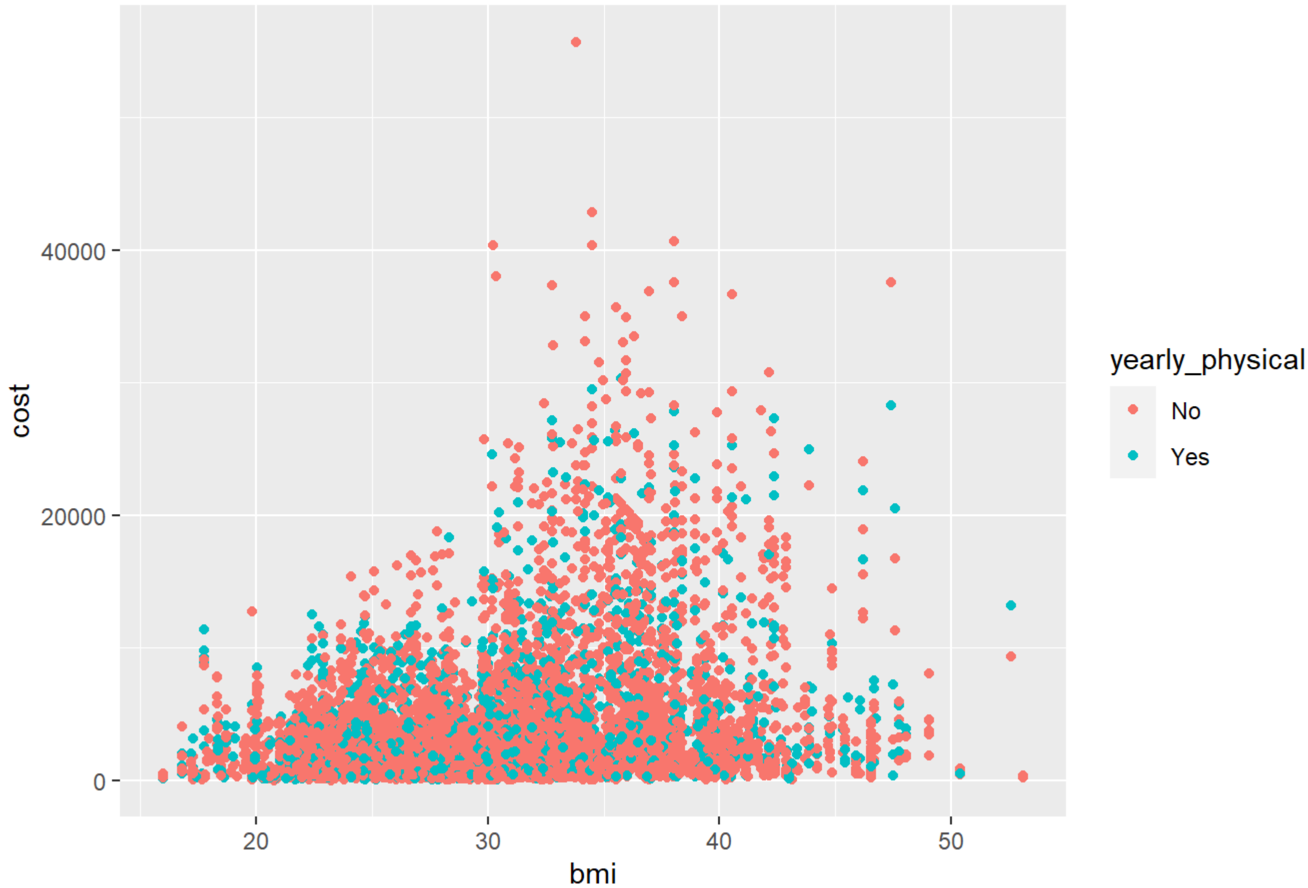



```
#Here we see that the costs for smokers are significantly higher than those for non smokers
```

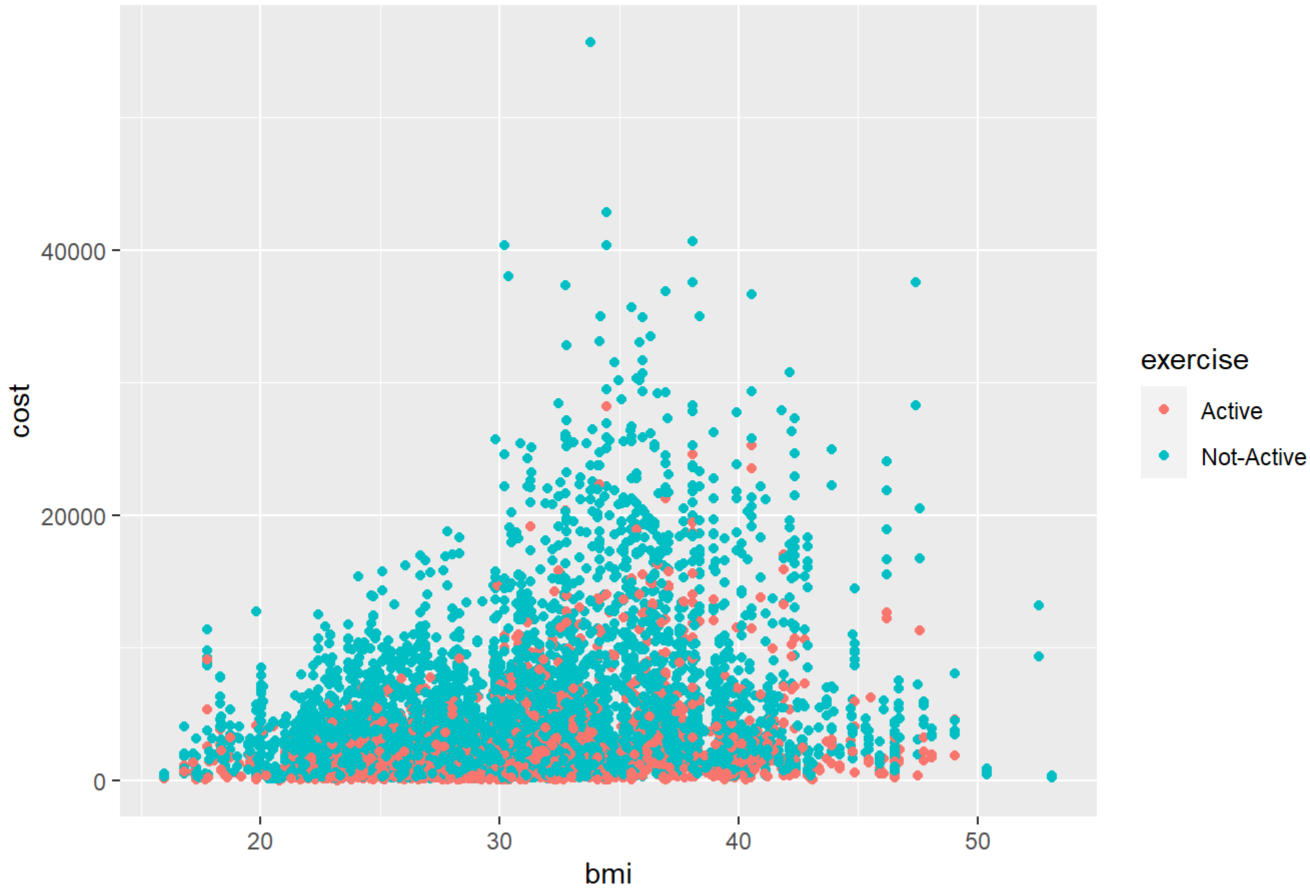
```
#11 Scatterplots
ggplot(hmodata)+geom_point(aes(x=bmi ,y=cost ,color=smoker))+
ylab('cost')+xlab('bmi')+ggtitle("")
```




```
ggplot(hmodata)+geom_point(aes(x=bmi ,y=cost ,color=yearly_physical))+
ylab('cost')+xlab('bmi')+ggtitle("")
```




```
ggplot(hmodata)+geom_point(aes(x=bmi ,y=cost ,color=exercise))+
ylab('cost')+xlab('bmi')+ggtitle("")
```




```
#12 Creating a duplicate dataset from the original dataset to use for model training

hmodata1 <- data.frame(hmodata)
```

```
#13 Predictive model svm
library(caret)
```

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

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(123)


hmodata_model <-data.frame(hmodata1)
#Creating duplicate dataset to utilize for prediction models

trainList <- createDataPartition(y=hmodata_model$cost_status,p=.70,list=FALSE)
#Creating data partition of our data frame to create a trainset for model training and a testset
for testing predictions

trainSet <- hmodata_model[trainList,]
testSet <- hmodata_model[-trainList,]

hmodata_svm1 <- train(cost_status ~ X+age+bmi+children+smoker+location_type+education_level+year
ly_physical+exercise+married+hypertension+gender, data = trainSet ,method = "svmRadial",trContro
l=trainControl(method ="none"), preProcess = c("center", "scale"))

predict_svm <- predict(hmodata_svm1, newdata=testSet)

confusionMatrix(predict_svm, testSet$cost_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  1655  267
##      TRUE     54  297
##
##                Accuracy : 0.8588
##                  95% CI : (0.8438, 0.8728)
##     No Information Rate : 0.7519
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5667
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9684
##             Specificity : 0.5266
##          Pos Pred Value : 0.8611
##          Neg Pred Value : 0.8462
##              Prevalence : 0.7519
##          Detection Rate : 0.7281
##    Detection Prevalence : 0.8456
##       Balanced Accuracy : 0.7475
##
##        'Positive' Class : FALSE
##
```

```
#SVM Model accuracy =85.88%
#SVM Model sensitivity =96.84%
```

```
#14 Prediction model ksvm

#install.packages("rio")
library(rio)
```

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

```
library(kernlab)
```

```
##
## Attaching package: 'kernlab'
```

```
## The following object is masked from 'package:purrr':
##
##     cross
```

```
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
library(rlang)
```

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

```
##
## Attaching package: 'rlang'
```

```
## The following objects are masked from 'package:purrr':
##
##     %@%, as_function, flatten, flatten_chr, flatten_dbl, flatten_int,
##     flatten_lgl, flatten_raw, invoke, splice
```

```r
library(caret)
set.seed(123)


hmodata_ksvm1<-ksvm(data= trainSet,cost_status~X+age+bmi+children+smoker+location_type+education
_level+yearly_physical+exercise+married+hypertension+gender, C=5, cross=3, prob.model=TRUE)

predict_ksvm <- predict(hmodata_ksvm1, newdata=testSet)

confusionMatrix(predict_ksvm, testSet$cost_status)
```

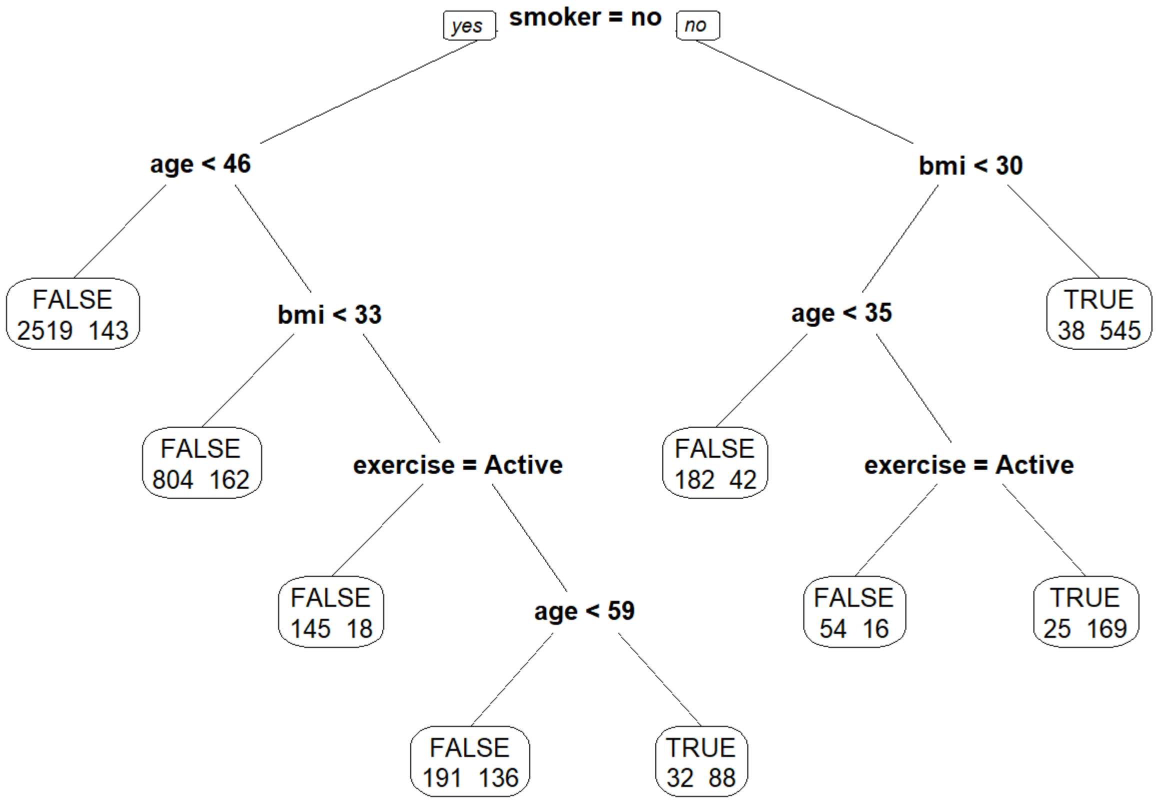
```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction FALSE TRUE
##      FALSE  1669  239
##      TRUE     40  325
##
##                Accuracy : 0.8773
##                  95% CI : (0.8631, 0.8905)
##     No Information Rate : 0.7519
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6269
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9766
##             Specificity : 0.5762
##          Pos Pred Value : 0.8747
##          Neg Pred Value : 0.8904
##              Prevalence : 0.7519
##          Detection Rate : 0.7343
##    Detection Prevalence : 0.8394
##       Balanced Accuracy : 0.7764
##
##        'Positive' Class : FALSE
##
```

```r
#KSVM Model Sensitivity 97.66%
#KSVM Model Accuracy 87.73%
```

```r
#15 Prediction Model training rpart tree

#install.packages('e1071', dependencies = TRUE)
#install.packages("rpart.plot")

library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.2
```

```
Treeplot<-rpart(cost_status ~ X+age+bmi+children+smoker+location_type+education_level+yearly_physical+exercise+married+hypertension+gender, data = trainSet, control = c(maxdepth = 5, cp=0.002))
prp(Treeplot, faclen = 0, cex = 0.8, extra = 1)
```




```
predict_tree <- predict(Treeplot, newdata=testSet, type = "class")

confusionMatrix(predict_tree, testSet$cost_status)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction FALSE TRUE
##      FALSE  1678  242
##      TRUE     31  322
## 
##                Accuracy : 0.8799
##                  95% CI : (0.8658, 0.893)
##     No Information Rate : 0.7519
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.632
## 
##  Mcnemar's Test P-Value : < 2.2e-16
## 
##             Sensitivity : 0.9819
##             Specificity : 0.5709
##          Pos Pred Value : 0.8740
##          Neg Pred Value : 0.9122
##              Prevalence : 0.7519
##          Detection Rate : 0.7382
##    Detection Prevalence : 0.8447
##       Balanced Accuracy : 0.7764
## 
##        'Positive' Class : FALSE
## 
```

```
#Tree Model Accuracy 87.99%
#Tree Model Sensitivity 98.19%
```