# DSC424HomeWork1

Ronaldlee Ejalu

1/15/2021

## Load all the necessary libraries

```
library(readr)
library(tidyverse)

## -- Attaching packages -----------------------------------------------------------
------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v dplyr   1.0.2
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts --------------------------------------------------------------------
------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(gtsummary)

## Warning: package 'gtsummary' was built under R version 4.0.3

## #BlackLivesMatter

library(tableone)

## Warning: package 'tableone' was built under R version 4.0.3

library(broom)
library(dplyr) #dplryr calculations
library(corrplot) # Plot Correlations

## Warning: package 'corrplot' was built under R version 4.0.3

## corrplot 0.84 loaded

library(DescTools) # VIF Function

## Warning: package 'DescTools' was built under R version 4.0.3
```

## Read data into R studio

```
insurance_dataset <- read_csv("C:\\Users\\rejalu1\\OneDrive - Henry Ford
Health System\\DSC424\\Data Sets\\insurance_dataset.csv")

## Parsed with column specification:
## cols(
##   age = col_double(),
##   sex = col_character(),
##   gender_num = col_double(),
##   bmi = col_double(),
##   children = col_double(),
##   smoker = col_character(),
##   smoker_num = col_double(),
##   region = col_character(),
##   region_num = col_double(),
##   expenses = col_double()
## )
```

#view the 10 data observations

```
head(insurance_dataset)

## # A tibble: 6 x 10
##     age sex   gender_num   bmi children smoker smoker_num region
region_num
##   <dbl> <chr>      <dbl> <dbl>    <dbl> <chr>       <dbl> <chr>
<dbl>
## 1    19 fema~          0  27.9        0 yes             1 south~
4
## 2    18 male           1  33.8        1 no              0 south~
3
## 3    28 male           1  33          3 no              0 south~
3
## 4    33 male           1  22.7        0 no              0 north~
2
## 5    32 male           1  28.9        0 no              0 north~
2
## 6    31 fema~          0  25.7        0 no              0 south~
3
## # ... with 1 more variable: expenses <dbl>
```

#view the last 10 data observations

```
tail(insurance_dataset)

## # A tibble: 6 x 10
##     age sex   gender_num   bmi children smoker smoker_num region
region_num
##   <dbl> <chr>      <dbl> <dbl>    <dbl> <chr>       <dbl> <chr>
<dbl>
## 1    52 fema~          0  44.7        3 no              0 south~
```

```
4
## 2    50 male           1  31          3 no             0 north~
2
## 3    18 fema~          0  31.9        0 no             0 north~
1
## 4    18 fema~          0  36.9        0 no             0 south~
3
## 5    21 fema~          0  25.8        0 no             0 south~
4
## 6    61 fema~          0  29.1        0 yes            1 north~
2
## # ... with 1 more variable: expenses <dbl>
```

#data structure

```
str(insurance_dataset)
```

```
## tibble [1,338 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age       : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex       : chr [1:1338] "female" "male" "male" "male" ...
##  $ gender_num: num [1:1338] 0 1 1 1 1 0 0 0 1 0 ...
##  $ bmi       : num [1:1338] 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8
25.8 ...
##  $ children  : num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker    : chr [1:1338] "yes" "no" "no" "no" ...
##  $ smoker_num: num [1:1338] 1 0 0 0 0 0 0 0 0 0 ...
##  $ region    : chr [1:1338] "southwest" "southeast" "southeast"
"northwest" ...
##  $ region_num: num [1:1338] 4 3 3 2 2 3 3 2 1 2 ...
##  $ expenses  : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   gender_num = col_double(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   smoker_num = col_double(),
##   ..   region = col_character(),
##   ..   region_num = col_double(),
##   ..   expenses = col_double()
##   .. )
```

## summarized statistical data from the data set

```
summary(insurance_dataset)
```

```
##       age            sex              gender_num          bmi
##  Min.   :18.00   Length:1338        Min.   :0.0000   Min.   :16.00
##  1st Qu.:27.00   Class :character   1st Qu.:0.0000   1st Qu.:26.30  
##  Median :39.00   Mode  :character   Median :1.0000   Median :30.40  
##  Mean   :39.21                      Mean   :0.5052   Mean   :30.67  
##  3rd Qu.:51.00                      3rd Qu.:1.0000   3rd Qu.:34.70  
##  Max.   :64.00                      Max.   :1.0000   Max.   :53.10  
##     children        smoker            smoker_num        region         
##  Min.   :0.000   Length:1338        Min.   :0.0000   Length:1338       
##  1st Qu.:0.000   Class :character   1st Qu.:0.0000   Class :character  
##  Median :1.000   Mode  :character   Median :0.0000   Mode  :character  
##  Mean   :1.095                      Mean   :0.2048                     
##  3rd Qu.:2.000                      3rd Qu.:0.0000                     
##  Max.   :5.000                      Max.   :1.0000                     
##    region_num       expenses    
##  Min.   :1.000   Min.   : 1122  
##  1st Qu.:2.000   1st Qu.: 4740  
##  Median :3.000   Median : 9382  
##  Mean   :2.516   Mean   :13270  
##  3rd Qu.:3.000   3rd Qu.:16640  
##  Max.   :4.000   Max.   :63770
```

#check for any missing value #There are no missing values

```
sum(is.na(insurance_dataset))

## [1] 0
```

#get specific column index in R

```
as.data.frame(colnames(insurance_dataset))

##    colnames(insurance_dataset)
## 1                          age
## 2                          sex
## 3                   gender_num
## 4                          bmi
## 5                     children
## 6                       smoker
## 7                   smoker_num
## 8                       region
## 9                   region_num
## 10                    expenses
```

#distinct values of each factor column # gender_num, smoker_num, region_num

```
insurance_dataset.sex <- count(distinct(insurance_dataset), sex)
insurance_dataset.sex

## # A tibble: 2 x 2
##   sex        n
##   <chr>  <int>
## 1 female   662
## 2 male     675
```

```
insurance_dataset.smoker <- count(distinct(insurance_dataset), smoker)
insurance_dataset.smoker
```

```
## # A tibble: 2 x 2
##   smoker     n
##   <chr>  <int>
## 1 no      1063
## 2 yes      274
```

```
insurance_dataset.region <- count(distinct(insurance_dataset), region)
insurance_dataset.region
```

```
## # A tibble: 4 x 2
##   region        n
##   <chr>     <int>
## 1 northeast   324
## 2 northwest   324
## 3 southeast   364
## 4 southwest   325
```

## data cleaning

```
insurance.clean <- insurance_dataset %>%
  transmute(age = age
            , sex = as.factor(sex)
            , gender_num = gender_num
            , bmi = bmi
            , children = children
            , smoker = as.factor(smoker)
            , smoker_num = smoker_num
            , region = as.factor(region)
            , region_num = region_num
            , expenses = expenses) %>%
  mutate(
    sex = relevel(sex, ref = 'male')
    , smoker = relevel(smoker, ref = 'no')
    , region = relevel(region, ref = 'southeast')
  )

# Create a matrix for sex
sexdummies.matrix <- model.matrix(~insurance.clean$sex)

# Convert the model matrix into a data frame
sexdummies.frame <- data.frame(sexdummies.matrix)

# bind the data frame to data set
insurance.clean <- cbind(insurance.clean, sexdummies.frame)

# create a matrix for smoker
smokerdummies.matrix <- model.matrix(~insurance.clean$smoker)
```

```r
#Convert the model matrix into a data frame
smokerdummies.frame <- data.frame(smokerdummies.matrix)

#bind the data frame to data set
insurance.clean <- cbind(insurance.clean, smokerdummies.frame)

# create a matrix for region
regiondummies.matrix <- model.matrix(~insurance.clean$region)

# Convert the model matrix into a data frame
regiondummies.frame <- data.frame(regiondummies.matrix)

# bind the data frame to a data set
insurance.clean <- cbind(insurance.clean, regiondummies.frame)
```

## rename and select all the variables interest

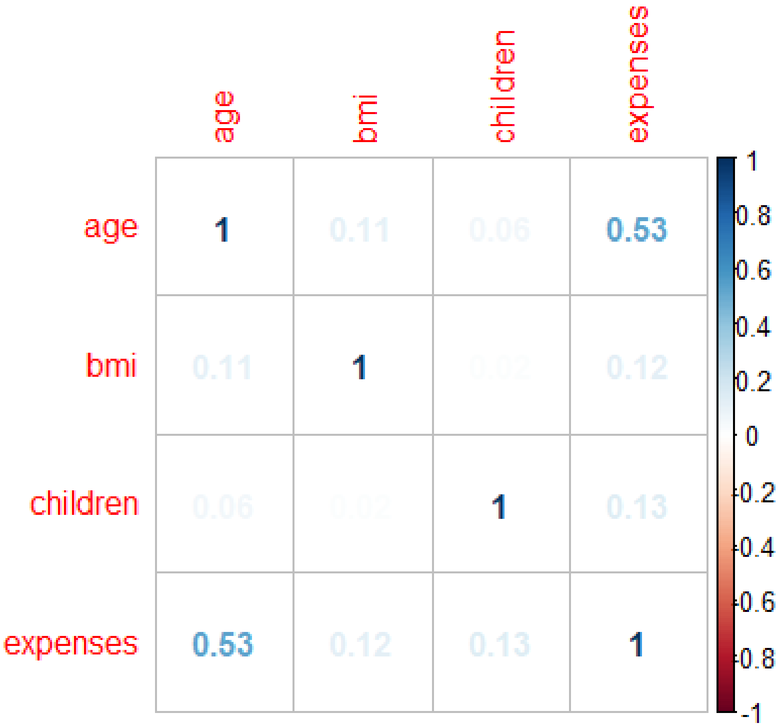
```r
insurancecleansed <- insurance.clean %>%
  select(age = age
           , gender_num = gender_num
           , bmi = bmi
           , children = children
           , smoker_num = smoker_num
           , region_num = region_num
           , expenses = expenses
           , sexfemale = insurance.clean.sexfemale
           , smokeryes = insurance.clean.smokeryes
           , northeast = insurance.clean.regionnortheast
           , northwest = insurance.clean.regionnorthwest
           , southwest = insurance.clean.regionsouthwest)
```

#extract out all numerical variables

```r
insurance.numvariables <- insurance_dataset[,c(1,4:5,10)]
```

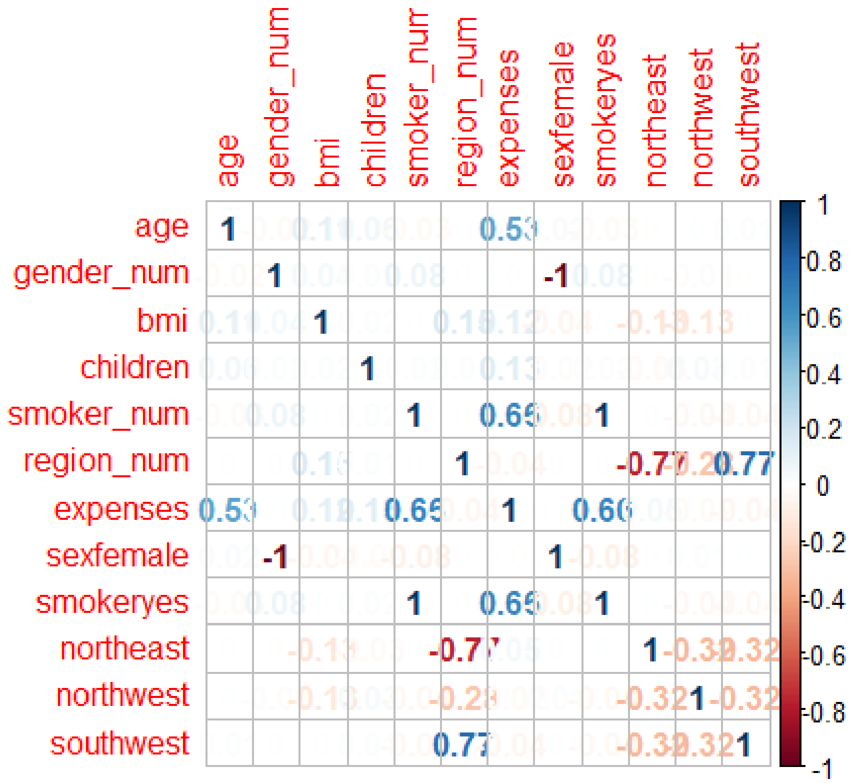
## check for multicollinearity amongst the numerical variables

```r
M <- cor(insurance.numvariables, method = "spearman")
M

##                 age        bmi   children  expenses
## age      1.00000000 0.10769164 0.05699222 0.5343921
## bmi      0.10769164 1.00000000 0.01558886 0.1194189
## children 0.05699222 0.01558886 1.00000000 0.1333389
## expenses 0.53439213 0.11941885 0.13333894 1.0000000

corrplot(M, method = "number")
```



## Check for multicollinearity amongst all the variables

```
m2 <- cor(insurancecleansed, method = "spearman")
corrplot(m2, method = "number")
```

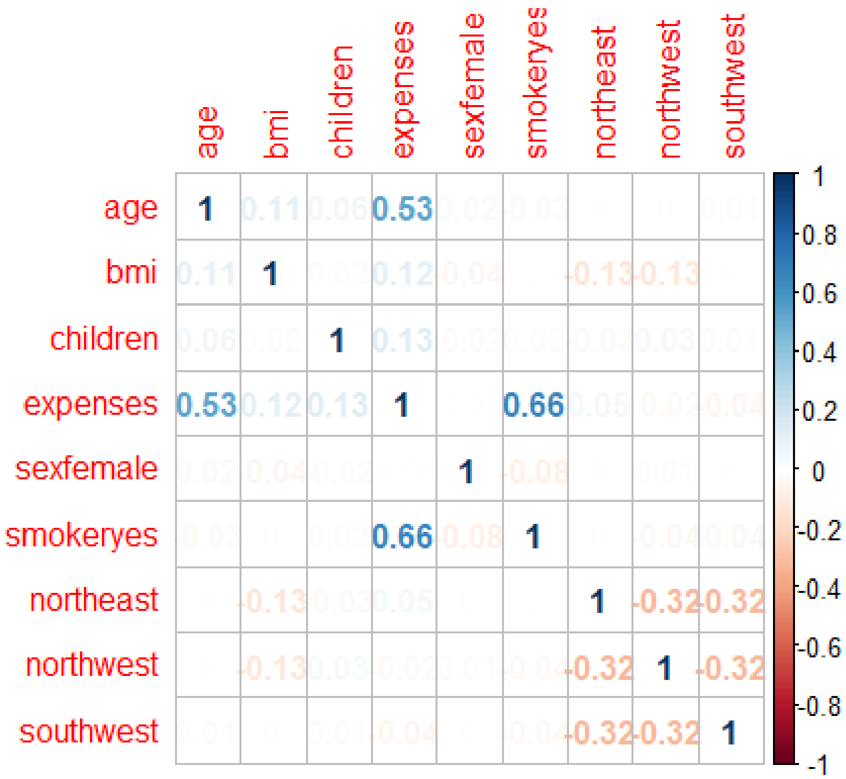


#Altering my data set after sensing multicollinearity in the original data # I select the variables of interest

```
insurancewithselectedvars <- insurancecleansed %>%
  select(age = age
            , bmi = bmi
            , children = children
            , expenses = expenses
            , sexfemale = sexfemale
            , smokeryes = smokeryes
            , northeast = northeast
            , northwest = northwest
            , southwest = southwest)
```

## Again check for multicollinearity

```
#summary(insurancewithselectedvars)
m3 <- cor(insurancewithselectedvars, method = "spearman")
#m3
corrplot(m3, method = "number")
```



```
model2 <- lm(expenses ~ ., data = insurancewithselectedvars)
summary(model2)

##
## Call:
## lm(formula = expenses ~ ., data = insurancewithselectedvars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13108.51    1090.51 -12.021  < 2e-16 ***
## age            256.84      11.90  21.586  < 2e-16 ***
## bmi            339.29      28.60  11.864  < 2e-16 ***
## children       475.69     137.80   3.452 0.000574 ***
## sexfemale      131.35     332.94   0.395 0.693255
## smokeryes    23847.48     413.14  57.723  < 2e-16 ***
## northeast     1035.60     478.68   2.163 0.030685 *
## northwest      682.81     478.95   1.426 0.154211
## southwest       76.29     470.64   0.162 0.871253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

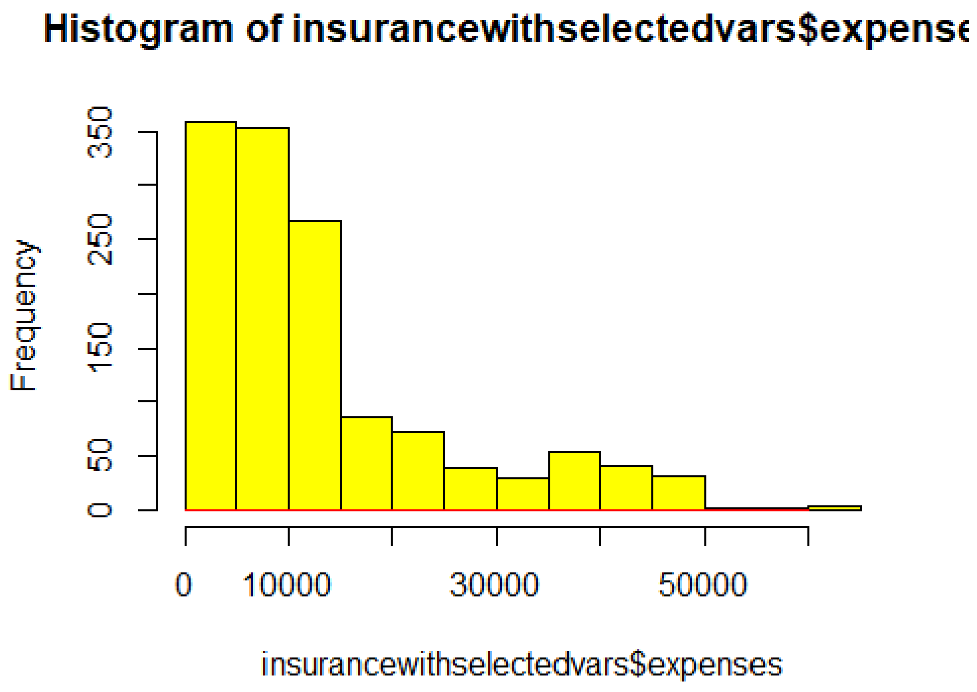
```
VIF(model2)
```

```
##       age       bmi  children sexfemale smokeryes northeast northwest 
southwest 
##  1.016843  1.106682  1.004008  1.008900  1.012067  1.531084  1.536030 
1.483177
```

## Explanatory analysis

#Histogram

```
hist(insurancewithselectedvars$expenses, col="yellow", freq=TRUE)
x <- seq(0, 60000, length.out = 50)
y <- with(insurancewithselectedvars, dnorm(x, mean(expenses), sd(expenses)))
lines(x, y, col="red")
```

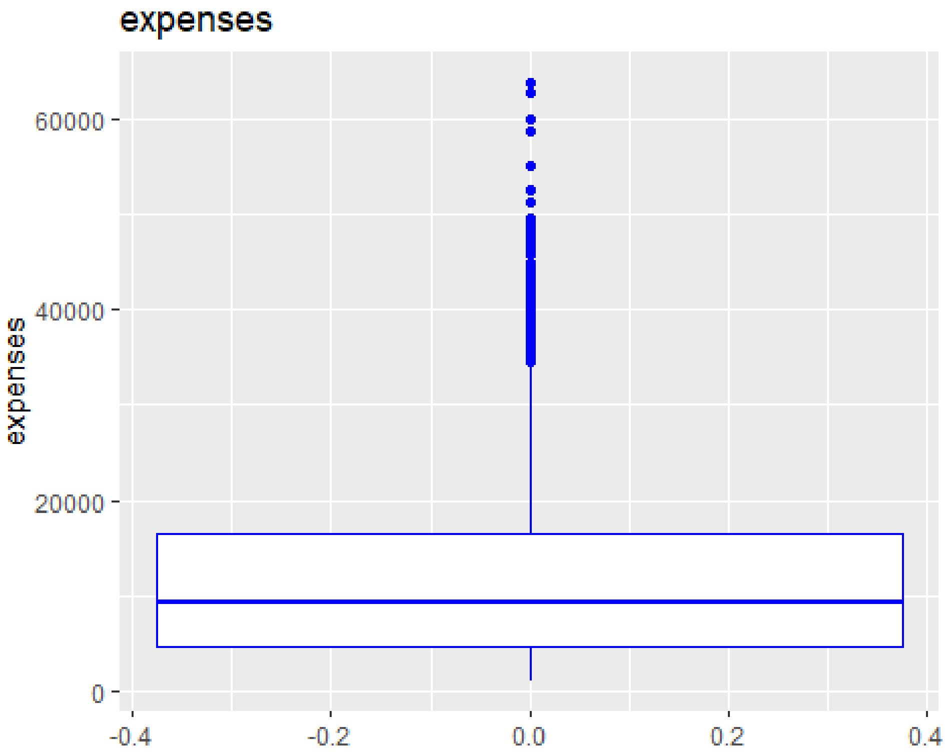


## Five - Number Summary for the Boxplot

```
summary(insurancewithselectedvars$expenses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1122    4740    9382   13270   16640   63770
```

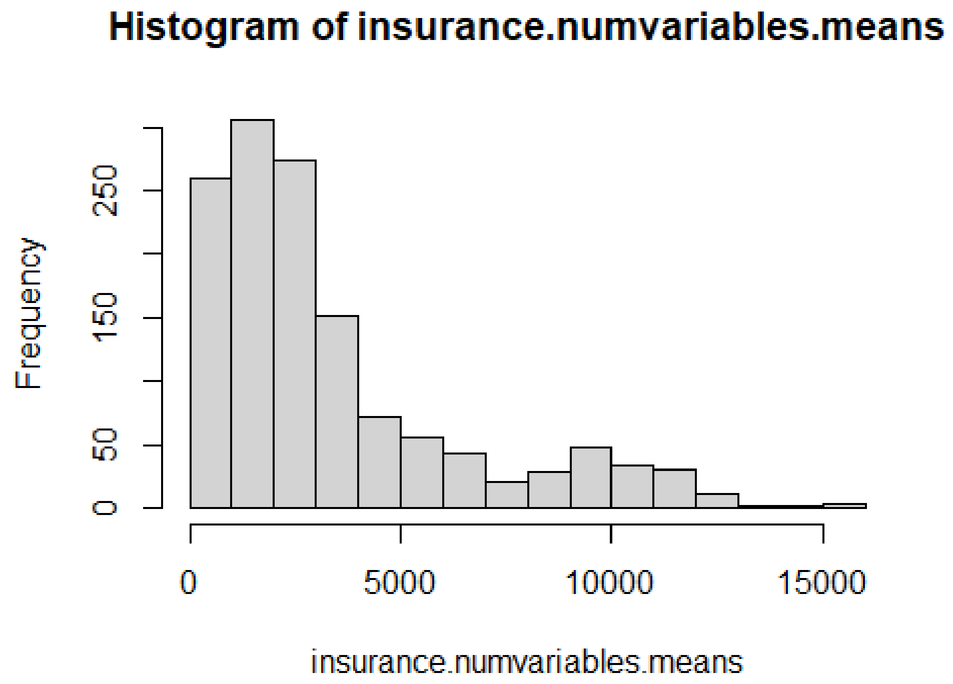
```
# Boxplots
insuranxebloxplot <-ggplot(insurancewithselectedvars, aes(y=expenses)) +
  geom_boxplot(col="blue") +
  labs(
    title="expenses",
    y="expenses")
insuranxebloxplot
```



```
ggsave("insuranxebloxplot.png")

## Saving 5 x 4 in image
```

## Return a vector with a mean value across each row of the insurance.numvariables data set

```
insurance.numvariables.means <- rowMeans(insurance.numvariables, na.rm=TRUE)
hist(insurance.numvariables.means)
```



#remove entries with the means greater than 5000

```
insurance.keep <- insurance.numvariables.means < 5000
```

## remove outliers from the original data frame

```
insuracedataset <- insurance_dataset[insurance.keep,]
```

## remove outliers from the numerical insurance data set

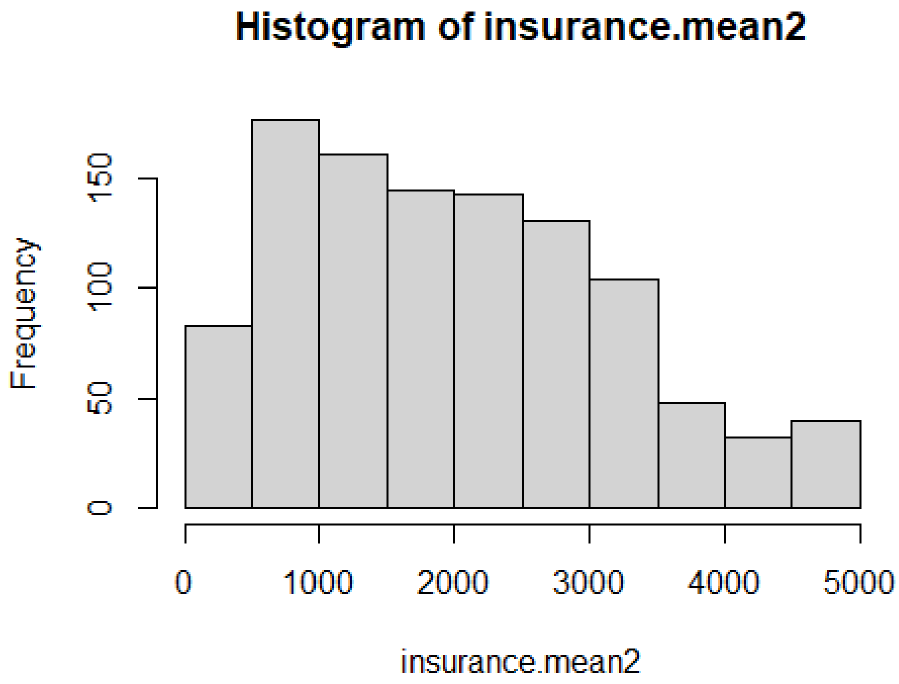
```
insurance.numvariables.withoutliers <-
insurance.numvariables[insurance.keep,]
```

#remove outliers from the insurance with selected vars data set

```
insurancewithselectedvars.withoutliers <-
insurancewithselectedvars[insurance.keep, ]
```

#plot the means with outliers removed

```
insurance.mean2 <- rowMeans(insurance.numvariables.withoutliers, na.rm =
TRUE)
hist(insurance.mean2)
```

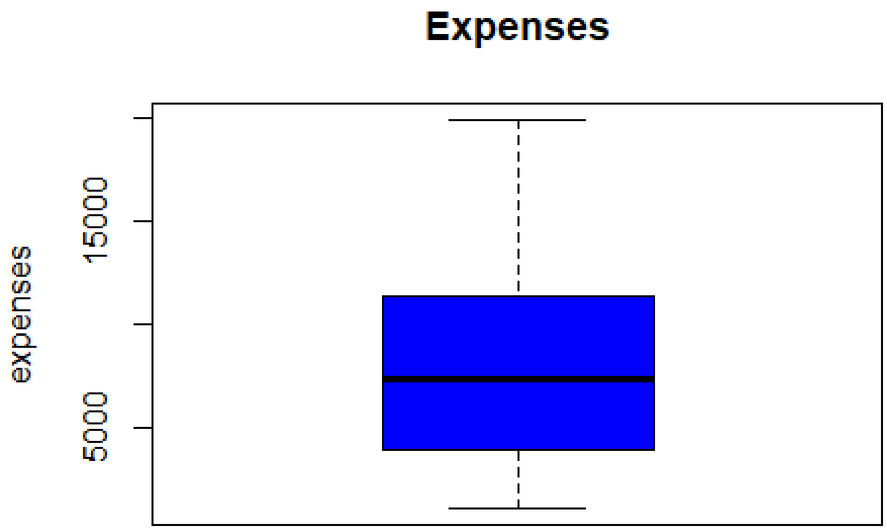


#five number summary

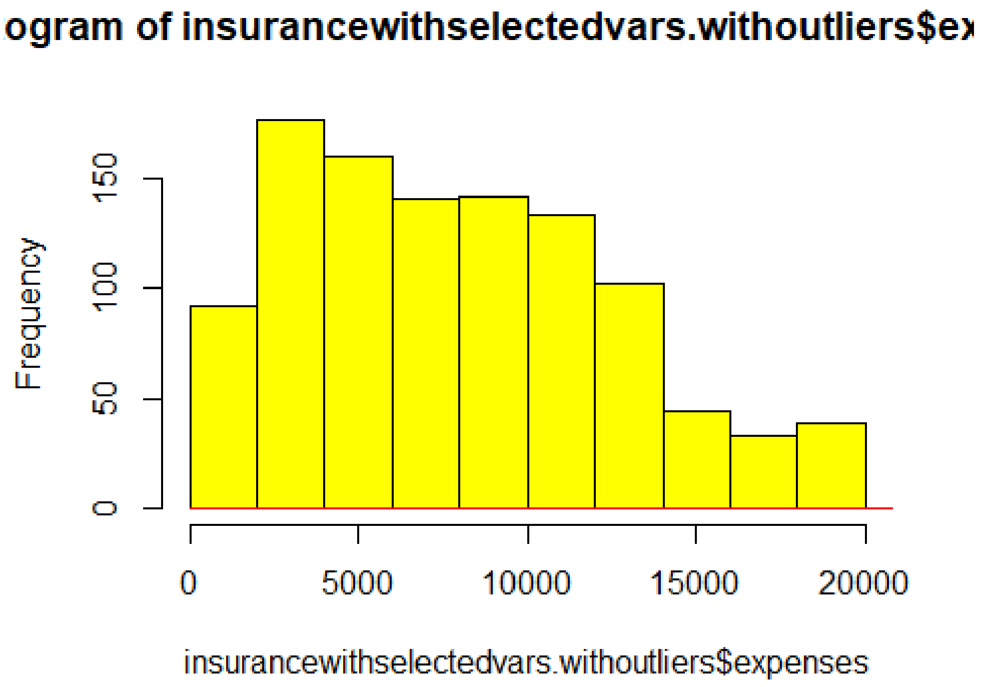
```
summary(insurancewithselectedvars.withoutliers$expenses,)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    3986    7345    7949   11363   19933
```

#box plot

```
boxplot(insurancewithselectedvars.withoutliers$expenses, col = "blue", main =
"Expenses", ylab = "expenses")
```



```
hist(insurancewithselectedvars.withoutliers$expenses, col="yellow",
freq=TRUE)
x <- seq(0, 60000, length.out = 50)
y <- with(insurancewithselectedvars.withoutliers, dnorm(x, mean(expenses),
sd(expenses)))
lines(x, y, col="red")
```



## model building after removing the outliers.

```
model3 <- lm(expenses ~ ., data = insurancewithselectedvars.withoutliers)
summary(model3)

## 
## Call:
## lm(formula = expenses ~ ., data = insurancewithselectedvars.withoutliers)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3400.6  -911.0  -512.9    77.5 15971.3 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4061.330    459.687  -8.835  < 2e-16 ***
## age           240.347      5.057  47.525  < 2e-16 ***
## bmi            36.218     12.147   2.982 0.002932 ** 
## children      476.477     56.228   8.474  < 2e-16 ***
## sexfemale     469.862    137.709   3.412 0.000669 ***
## smokeryes   13012.777    309.396  42.059  < 2e-16 ***
## northeast     609.983    201.904   3.021 0.002579 ** 
## northwest     358.805    199.860   1.795 0.072895 .  
## southwest    -192.527    195.892  -0.983 0.325921    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2243 on 1055 degrees of freedom
## Multiple R-squared:  0.7744, Adjusted R-squared:  0.7727
## F-statistic: 452.7 on 8 and 1055 DF,  p-value: < 2.2e-16
```

## Creating the automatic models

```
null = lm(expenses ~ 1, data = insurancewithselectedvars.withoutliers)
null
```

```
##
## Call:
## lm(formula = expenses ~ 1, data = insurancewithselectedvars.withoutliers)
##
## Coefficients:
## (Intercept)
##        7949
```

```
full = lm(expenses ~ ., data = insurancewithselectedvars.withoutliers)

#Forward regression
train_forward = step(null, scope = list(lower=null,
upper=full),direction="forward")
```

```
## Start:  AIC=17995.95
## expenses ~ 1
##
##             Df  Sum of Sq        RSS   AIC
## + age        1 8574022947 1.4953e+10 17516
## + smokeryes  1 5749769984 1.7777e+10 17700
## + children   1  548799328 2.2978e+10 17973
## + northeast  1   80416879 2.3446e+10 17994
## + sexfemale  1   80368634 2.3446e+10 17994
## <none>                    2.3527e+10 17996
## + southwest  1   15517598 2.3511e+10 17997
## + bmi        1    2601381 2.3524e+10 17998
## + northwest  1     890445 2.3526e+10 17998
##
## Step:  AIC=17515.7
## expenses ~ age
##
##             Df  Sum of Sq        RSS   AIC
## + smokeryes  1 9119040316 5.8336e+09 16516
## + children   1  375396047 1.4577e+10 17491
## + bmi        1  231703142 1.4721e+10 17501
## + northeast  1   71880163 1.4881e+10 17513
## + southwest  1   51420675 1.4901e+10 17514
## + sexfemale  1   44433177 1.4908e+10 17515
## <none>                    1.4953e+10 17516
## + northwest  1     280125 1.4952e+10 17518
```

```
##
## Step:  AIC=16516.21
## expenses ~ age + smokeryes
##
##              Df Sum of Sq        RSS   AIC
## + children   1 346670490 5486943917 16453
## + sexfemale  1  52197142 5781417265 16509
## + southwest  1  42744521 5790869886 16510
## + northeast  1  38271526 5795342880 16511
## + bmi        1  23782172 5809832234 16514
## <none>                   5833614407 16516
## + northwest  1   5818288 5827796119 16517
##
## Step:  AIC=16453.02
## expenses ~ age + smokeryes + children
##
##              Df Sum of Sq        RSS   AIC
## + sexfemale  1  54448427 5432495489 16444
## + southwest  1  51445266 5435498651 16445
## + northeast  1  45705618 5441238299 16446
## + bmi        1  24338028 5462605889 16450
## <none>                   5486943917 16453
## + northwest  1   5247104 5481696812 16454
##
## Step:  AIC=16444.41
## expenses ~ age + smokeryes + children + sexfemale
##
##              Df Sum of Sq        RSS   AIC
## + southwest  1  53224973 5379270517 16436
## + northeast  1  45666344 5386829146 16437
## + bmi        1  26777953 5405717536 16441
## <none>                   5432495489 16444
## + northwest  1   5317643 5427177847 16445
##
## Step:  AIC=16435.93
## expenses ~ age + smokeryes + children + sexfemale + southwest
##
##              Df Sum of Sq        RSS   AIC
## + bmi        1  25631742 5353638775 16433
## + northeast  1  21444599 5357825918 16434
## <none>                   5379270517 16436
## + northwest  1     21474 5379249043 16438
##
## Step:  AIC=16432.85
## expenses ~ age + smokeryes + children + sexfemale + southwest +
##     bmi
##
##              Df Sum of Sq        RSS   AIC
## + northeast  1  30061483 5323577292 16429
## <none>                   5353638775 16433
## + northwest  1    358961 5353279814 16435
##
## Step:  AIC=16428.86
## expenses ~ age + smokeryes + children + sexfemale + southwest +
##     bmi + northeast
##
##               Df Sum of Sq        RSS   AIC
## + northwest  1  16214040 5307363251 16428
## <none>                   5323577292 16429
##
## Step:  AIC=16427.61
## expenses ~ age + smokeryes + children + sexfemale + southwest +
##     bmi + northeast + northwest
```

```
summary(train_forward)
```

```
##
## Call:
## lm(formula = expenses ~ age + smokeryes + children + sexfemale +
##     southwest + bmi + northeast + northwest, data =
insurancewithselectedvars.withoutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3400.6  -911.0  -512.9    77.5 15971.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4061.330    459.687  -8.835  < 2e-16 ***
## age           240.347      5.057  47.525  < 2e-16 ***
## smokeryes   13012.777    309.396  42.059  < 2e-16 ***
## children      476.477     56.228   8.474  < 2e-16 ***
## sexfemale     469.862    137.709   3.412 0.000669 ***
## southwest    -192.527    195.892  -0.983 0.325921
## bmi            36.218     12.147   2.982 0.002932 **
## northeast     609.983    201.904   3.021 0.002579 **
## northwest     358.805    199.860   1.795 0.072895 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2243 on 1055 degrees of freedom
## Multiple R-squared:  0.7744, Adjusted R-squared:  0.7727
## F-statistic: 452.7 on 8 and 1055 DF,  p-value: < 2.2e-16
```

## using backward

```
train_backward = step(full, dierction="backward")
```

```
## Start:  AIC=16427.61
## expenses ~ age + bmi + children + sexfemale + smokeryes + northeast +
##     northwest + southwest
##
##               Df  Sum of Sq        RSS   AIC
## - southwest  1 4.8593e+06 5.3122e+09 16427
## <none>                    5.3074e+09 16428
## - northwest  1 1.6214e+07 5.3236e+09 16429
## - bmi        1 4.4725e+07 5.3521e+09 16435
## - northeast  1 4.5917e+07 5.3533e+09 16435
## - sexfemale  1 5.8566e+07 5.3659e+09 16437
## - children   1 3.6125e+08 5.6686e+09 16496
## - smokeryes  1 8.8989e+09 1.4206e+10 17473
## - age        1 1.1362e+10 1.6670e+10 17643
##
## Step:  AIC=16426.59
## expenses ~ age + bmi + children + sexfemale + smokeryes + northeast +
##     northwest
##
##               Df  Sum of Sq        RSS   AIC
## <none>                    5.3122e+09 16427
## - northwest  1 3.6238e+07 5.3485e+09 16432
## - bmi        1 5.2198e+07 5.3644e+09 16435
## - sexfemale  1 5.8136e+07 5.3704e+09 16436
## - northeast  1 8.4389e+07 5.3966e+09 16441
## - children   1 3.5932e+08 5.6715e+09 16494
## - smokeryes  1 8.9465e+09 1.4259e+10 17475
## - age        1 1.1374e+10 1.6687e+10 17642
```

```
summary(train_backward)
```

```
##
## Call:
## lm(formula = expenses ~ age + bmi + children + sexfemale + smokeryes +
##     northeast + northwest, data = insurancewithselectedvars.withoutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3391.4  -917.9  -508.4    76.4 15878.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4213.938    432.661  -9.740  < 2e-16 ***
## age           240.054      5.048  47.551  < 2e-16 ***
## bmi            38.442     11.934   3.221  0.00132 **
## children      475.040     56.208   8.452  < 2e-16 ***
## sexfemale     468.098    137.695   3.400  0.00070 ***
## smokeryes   13028.969    308.952  42.172  < 2e-16 ***
## northeast     711.222    173.648   4.096 4.53e-05 ***
## northwest     459.903    171.353   2.684  0.00739 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2243 on 1056 degrees of freedom
## Multiple R-squared:  0.7742, Adjusted R-squared:  0.7727
## F-statistic: 517.3 on 7 and 1056 DF,  p-value: < 2.2e-16
```

## using stepwise Regression

```
train_step = step(null, scope = list(upper=full), direction = "both")
```

```
## Start:  AIC=17995.95
## expenses ~ 1
##
##              Df  Sum of Sq        RSS   AIC
## + age         1 8574022947 1.4953e+10 17516
## + smokeryes   1 5749769984 1.7777e+10 17700
## + children    1  548799328 2.2978e+10 17973
## + northeast   1   80416879 2.3446e+10 17994
## + sexfemale   1   80368634 2.3446e+10 17994
## <none>                     2.3527e+10 17996
## + southwest   1   15517598 2.3511e+10 17997
## + bmi         1    2601381 2.3524e+10 17998
## + northwest   1     890445 2.3526e+10 17998
##
## Step:  AIC=17515.7
## expenses ~ age
##
##              Df  Sum of Sq        RSS   AIC
## + smokeryes   1 9119040316 5.8336e+09 16516
## + children    1  375396047 1.4577e+10 17491
## + bmi         1  231703142 1.4721e+10 17501
## + northeast   1   71880163 1.4881e+10 17513
## + southwest   1   51420675 1.4901e+10 17514
## + sexfemale   1   44433177 1.4908e+10 17515
## <none>                     1.4953e+10 17516
## + northwest   1     280125 1.4952e+10 17518
## - age         1 8574022947 2.3527e+10 17996
##
## Step:  AIC=16516.21
## expenses ~ age + smokeryes
##
##              Df  Sum of Sq        RSS   AIC
## + children    1 3.4667e+08 5.4869e+09 16453
## + sexfemale   1 5.2197e+07 5.7814e+09 16509
## + southwest   1 4.2745e+07 5.7909e+09 16510
## + northeast   1 3.8272e+07 5.7953e+09 16511
## + bmi         1 2.3782e+07 5.8098e+09 16514
## <none>                     5.8336e+09 16516
## + northwest   1 5.8183e+06 5.8278e+09 16517
## - smokeryes   1 9.1190e+09 1.4953e+10 17516
## - age         1 1.1943e+10 1.7777e+10 17700
```

```
##
## Step:  AIC=16453.02
## expenses ~ age + smokeryes + children
##
##              Df  Sum of Sq        RSS   AIC
## + sexfemale  1 5.4448e+07 5.4325e+09 16444
## + southwest  1 5.1445e+07 5.4355e+09 16445
## + northeast  1 4.5706e+07 5.4412e+09 16446
## + bmi        1 2.4338e+07 5.4626e+09 16450
## <none>                    5.4869e+09 16453
## + northwest  1 5.2471e+06 5.4817e+09 16454
## - children   1 3.4667e+08 5.8336e+09 16516
## - smokeryes  1 9.0903e+09 1.4577e+10 17491
## - age        1 1.1739e+10 1.7226e+10 17668
##
## Step:  AIC=16444.41
## expenses ~ age + smokeryes + children + sexfemale
##
##              Df  Sum of Sq        RSS   AIC
## + southwest  1 5.3225e+07 5.3793e+09 16436
## + northeast  1 4.5666e+07 5.3868e+09 16437
## + bmi        1 2.6778e+07 5.4057e+09 16441
## <none>                    5.4325e+09 16444
## + northwest  1 5.3176e+06 5.4272e+09 16445
## - sexfemale  1 5.4448e+07 5.4869e+09 16453
## - children   1 3.4892e+08 5.7814e+09 16509
## - smokeryes  1 9.0982e+09 1.4531e+10 17489
## - age        1 1.1694e+10 1.7127e+10 17664
##
## Step:  AIC=16435.93
## expenses ~ age + smokeryes + children + sexfemale + southwest
##
##              Df  Sum of Sq        RSS   AIC
## + bmi        1 2.5632e+07 5.3536e+09 16433
## + northeast  1 2.1445e+07 5.3578e+09 16434
## <none>                    5.3793e+09 16436
## + northwest  1 2.1474e+04 5.3792e+09 16438
## - southwest  1 5.3225e+07 5.4325e+09 16444
## - sexfemale  1 5.6228e+07 5.4355e+09 16445
## - children   1 3.5784e+08 5.7371e+09 16503
## - smokeryes  1 9.0883e+09 1.4468e+10 17487
## - age        1 1.1732e+10 1.7111e+10 17665
##
## Step:  AIC=16432.85
## expenses ~ age + smokeryes + children + sexfemale + southwest +
##     bmi
##
##              Df  Sum of Sq        RSS   AIC
## + northeast  1 3.0061e+07 5.3236e+09 16429
## <none>                    5.3536e+09 16433
## + northwest  1 3.5896e+05 5.3533e+09 16435
## - bmi        1 2.5632e+07 5.3793e+09 16436
## - southwest  1 5.2079e+07 5.4057e+09 16441
## - sexfemale  1 5.8632e+07 5.4123e+09 16442
## - children   1 3.5837e+08 5.7120e+09 16500
## - smokeryes  1 8.8877e+09 1.4241e+10 17472
## - age        1 1.1483e+10 1.6836e+10 17650
##
## Step:  AIC=16428.86
## expenses ~ age + smokeryes + children + sexfemale + southwest +
##     bmi + northeast
##
##             Df  Sum of Sq        RSS   AIC
## + northwest  1 1.6214e+07 5.3074e+09 16428
## <none>                    5.3236e+09 16429
## - southwest  1 2.4883e+07 5.3485e+09 16432
## - northeast  1 3.0061e+07 5.3536e+09 16433
## - bmi        1 3.4249e+07 5.3578e+09 16434
## - sexfemale  1 5.8535e+07 5.3821e+09 16439
## - children   1 3.6266e+08 5.6862e+09 16497
## - smokeryes  1 8.8970e+09 1.4221e+10 17472
## - age        1 1.1419e+10 1.6742e+10 17646
##
## Step:  AIC=16427.61
## expenses ~ age + smokeryes + children + sexfemale + southwest +
##     bmi + northeast + northwest
##
##             Df  Sum of Sq        RSS   AIC
## - southwest  1 4.8593e+06 5.3122e+09 16427
## <none>                    5.3074e+09 16428
## - northwest  1 1.6214e+07 5.3236e+09 16429
## - bmi        1 4.4725e+07 5.3521e+09 16435
## - northeast  1 4.5917e+07 5.3533e+09 16435
## - sexfemale  1 5.8566e+07 5.3659e+09 16437
## - children   1 3.6125e+08 5.6686e+09 16496
## - smokeryes  1 8.8989e+09 1.4206e+10 17473
## - age        1 1.1362e+10 1.6670e+10 17643
##
## Step:  AIC=16426.59
## expenses ~ age + smokeryes + children + sexfemale + bmi + northeast +
##     northwest
##
##             Df  Sum of Sq        RSS   AIC
## <none>                    5.3122e+09 16427
## + southwest  1 4.8593e+06 5.3074e+09 16428
## - northwest  1 3.6238e+07 5.3485e+09 16432
## - bmi        1 5.2198e+07 5.3644e+09 16435
## - sexfemale  1 5.8136e+07 5.3704e+09 16436
## - northeast  1 8.4389e+07 5.3966e+09 16441
## - children   1 3.5932e+08 5.6715e+09 16494
## - smokeryes  1 8.9465e+09 1.4259e+10 17475
## - age        1 1.1374e+10 1.6687e+10 17642
```

```
summary(train_step)
```

```
## 
## Call:
## lm(formula = expenses ~ age + smokeryes + children + sexfemale + 
##     bmi + northeast + northwest, data = 
insurancewithselectedvars.withoutliers)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3391.4  -917.9  -508.4    76.4 15878.8 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4213.938    432.661  -9.740  < 2e-16 ***
## age           240.054      5.048  47.551  < 2e-16 ***
## smokeryes   13028.969    308.952  42.172  < 2e-16 ***
## children      475.040     56.208   8.452  < 2e-16 ***
## sexfemale     468.098    137.695   3.400  0.00070 ***
## bmi            38.442     11.934   3.221  0.00132 ** 
## northeast     711.222    173.648   4.096 4.53e-05 ***
## northwest     459.903    171.353   2.684  0.00739 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2243 on 1056 degrees of freedom
## Multiple R-squared:  0.7742, Adjusted R-squared:  0.7727 
## F-statistic: 517.3 on 7 and 1056 DF,  p-value: < 2.2e-16
```

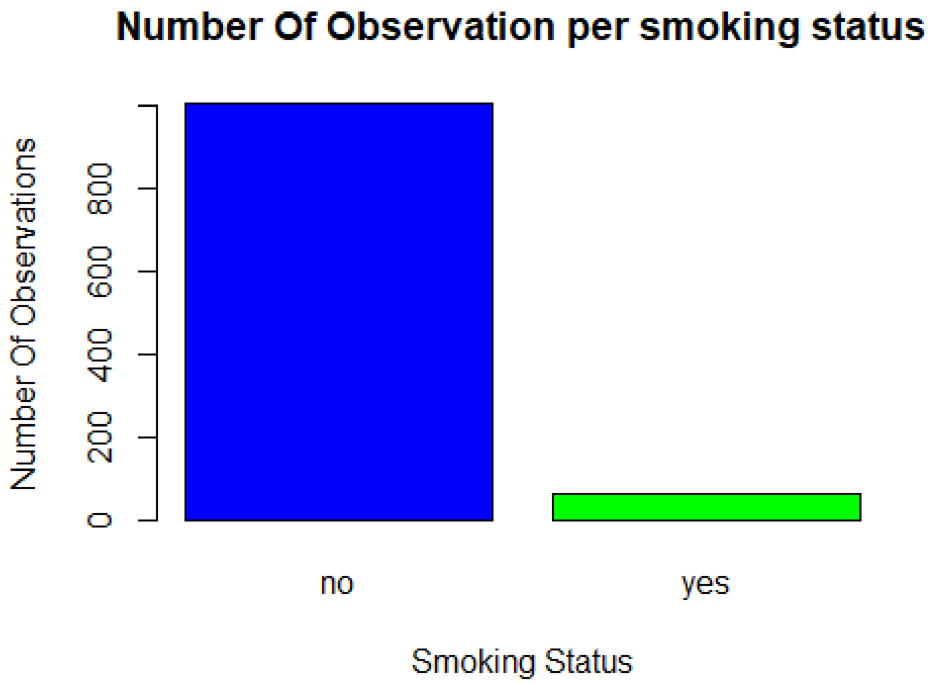
## Data Visualization

plot a box plot

```
counts <- table(insuracedataset$smoker)
counts
```

```
## 
##   no  yes 
## 1003   61
```

```
barplot(counts, main="Number Of Observation per smoking status",ylab="Number 
Of Observations", xlab="Smoking Status", col=c("blue","green"))
```

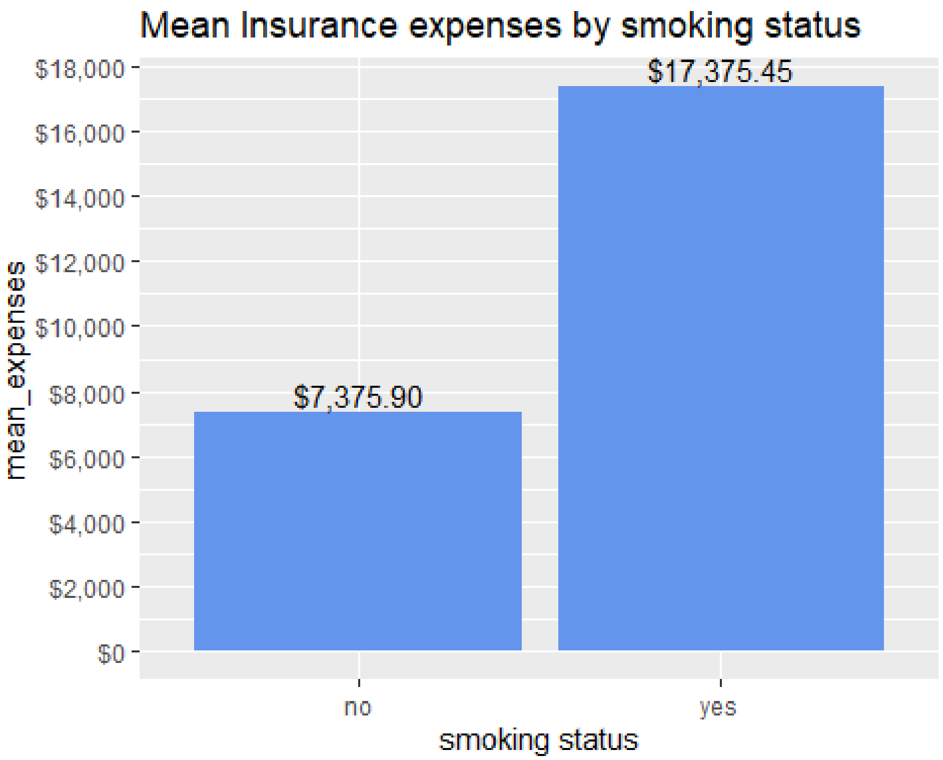


## calculate the mean expense by smoking status

## plot a bar chart

```
library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor

plotdata <- insuracedataset %>%
  group_by(smoker) %>%
  summarize(mean_expenses = mean(expenses))

## `summarise()` ungrouping output (override with `.groups` argument)

#plotdata

# plot the means
```

```r
ggplot(plotdata,
       aes(x = smoker,
           y = mean_expenses)) +
  geom_bar(stat = "identity",
           fill = "cornflowerblue") +
  geom_text(aes(label = dollar(mean_expenses)),
            vjust = -0.25) +
  scale_y_continuous(breaks = seq(0, 30000, 2000),
                     label = dollar
                     ) +
  labs(title = "Mean Insurance expenses by smoking status",
       x = "smoking status",
       y = "mean_expenses")
```


Mean Insurance expenses by smoking status
$18,000
$16,000
$14,000
$12,000
$10,000
$8,000
$6,000
$4,000
$2,000
$0
mean_expenses
$7,375.90
$17,375.45
no
yes
smoking status