Premium insurance for policyholders using Linear Regression with R

Column details - age: age of primary beneficiary sex: gender- female, male bmi: Body mass index, providing an understanding of the 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. children: Number of children covered by health insurance smoker: Yes NO region: the policyholder's residential area in the US, northeast, southeast, southwest, northwest. charges: Individual medical costs billed by health insurance

Now, since we got a brief introduction about the dataset, we will now begin with the coding. So let's dive in. We will first load the data set in R and process it: We will predict which of the above category of the person would be responsible to make him the premium insurance holder. The person who will be charged more would be the premium policyholder.

```
getwd()

## [1] "C:/Users/badal/Documents"
```

Install Required packages.

install.packages("psych") install.packages("tidyverse") install.packages("corrplot") install.packages("knitr") install.packages("gridExtra")

load library

```
library(psych)

## Warning: package 'psych' was built under R version 3.6.1

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.1

## -- Attaching packages --------------------------------------- tidyverse 
1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   1.0.0
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6.1

## Warning: package 'tibble' was built under R version 3.6.1

## Warning: package 'tidyr' was built under R version 3.6.1
## Warning: package 'readr' was built under R version 3.6.1

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'stringr' was built under R version 3.6.1

## Warning: package 'forcats' was built under R version 3.6.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()

library(knitr)

## Warning: package 'knitr' was built under R version 3.6.1

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.1

## corrplot 0.84 loaded

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.6.1

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

Read file

```
insurance <- read.csv('C://Users/badal/Desktop/datset_/insurance.csv')
head(insurance)

##   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

describe(insurance)

##            vars    n     mean       sd   median  trimmed      mad      min
## age           1 1338    39.21    14.05    39.00    39.01    17.79    18.00
## sex*          2 1338     1.51     0.50     2.00     1.51     0.00     1.00
## bmi         3 1338    30.66     6.10   30.40    30.50    6.20   15.96
## children    4 1338     1.09     1.21    1.00     0.94    1.48    0.00
## smoker*     5 1338     1.20     0.40    1.00     1.13    0.00    1.00
## region*     6 1338     2.52     1.10    3.00     2.52    1.48    1.00
## charges     7 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.87
##                max    range  skew kurtosis     se
## age         64.00    46.00  0.06    -1.25   0.38
## sex*         2.00     1.00 -0.02    -2.00   0.01
## bmi         53.13    37.17  0.28    -0.06   0.17
## children     5.00     5.00  0.94     0.19   0.03
## smoker*      2.00     1.00  1.46     0.14   0.01
## region*      4.00     3.00 -0.04    -1.33   0.03
## charges  63770.43 62648.55  1.51     1.59 331.07
```

```
str(insurance)
```

```
## '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 ...
```

The dataset has 7 variables, and 1338 cases.

```
summary(insurance)
```

```
##       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
```

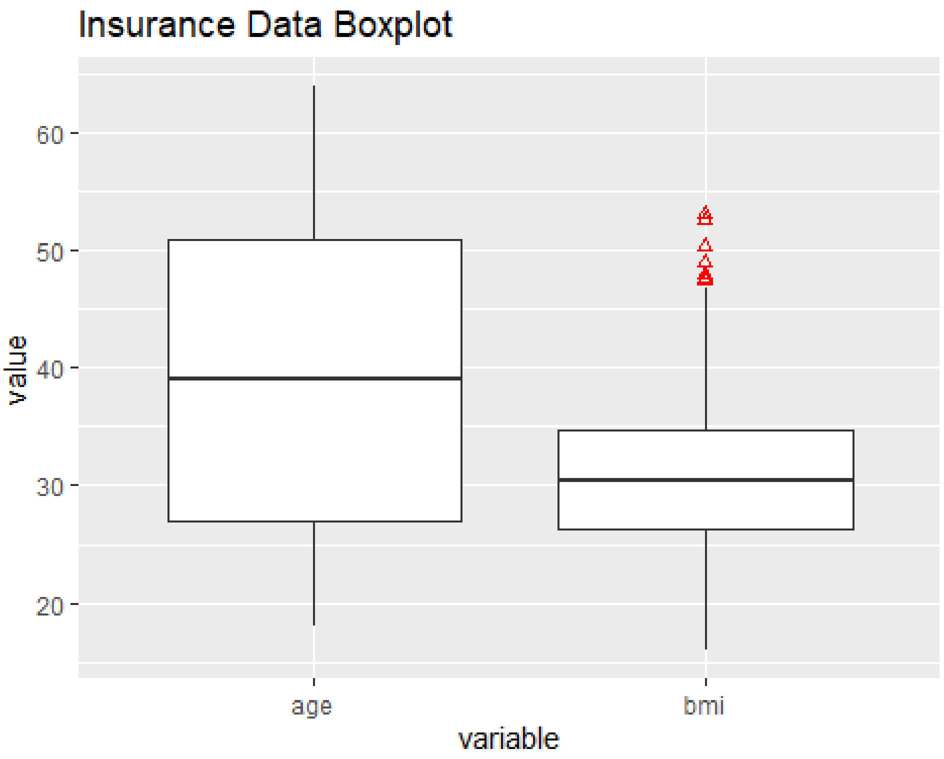
```
any(is.na(insurance))
```

```
## [1] FALSE
```

No missing values present in the dataset.

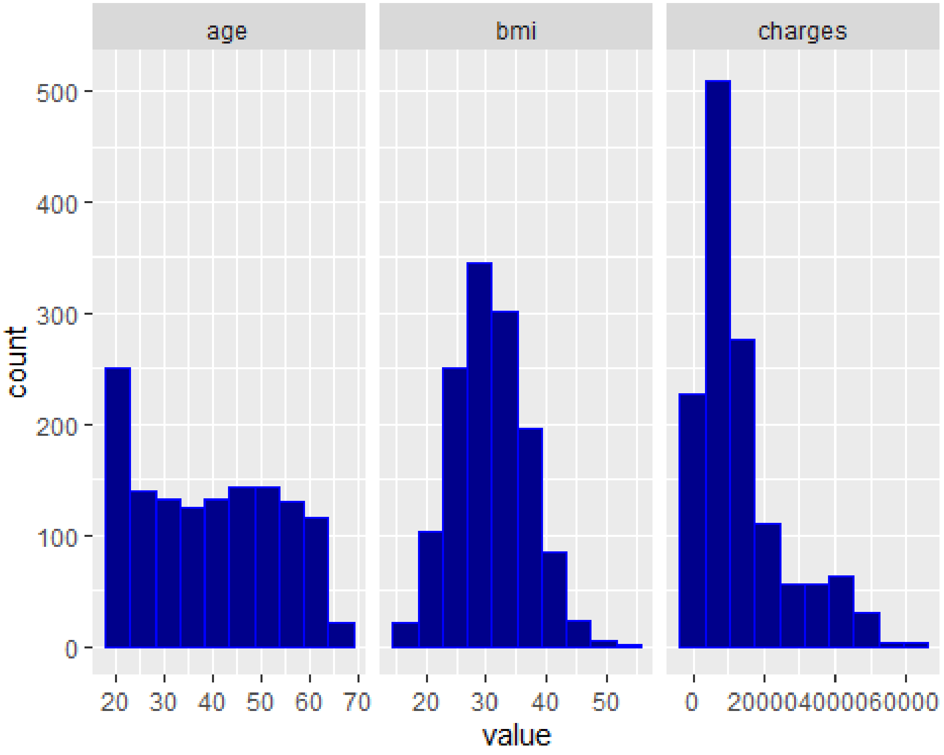
Box plot

```
insurance_boxplot <- insurance %>%
  select(c(1, 3)) %>%
  gather()

boxplot <- ggplot(insurance_boxplot, aes(x = key, y = value)) +
  labs(x = "variable", title = "Insurance Data Boxplot") +
  geom_boxplot(outlier.colour = "red",fill="white", outlier.shape = 2)

boxplot
```



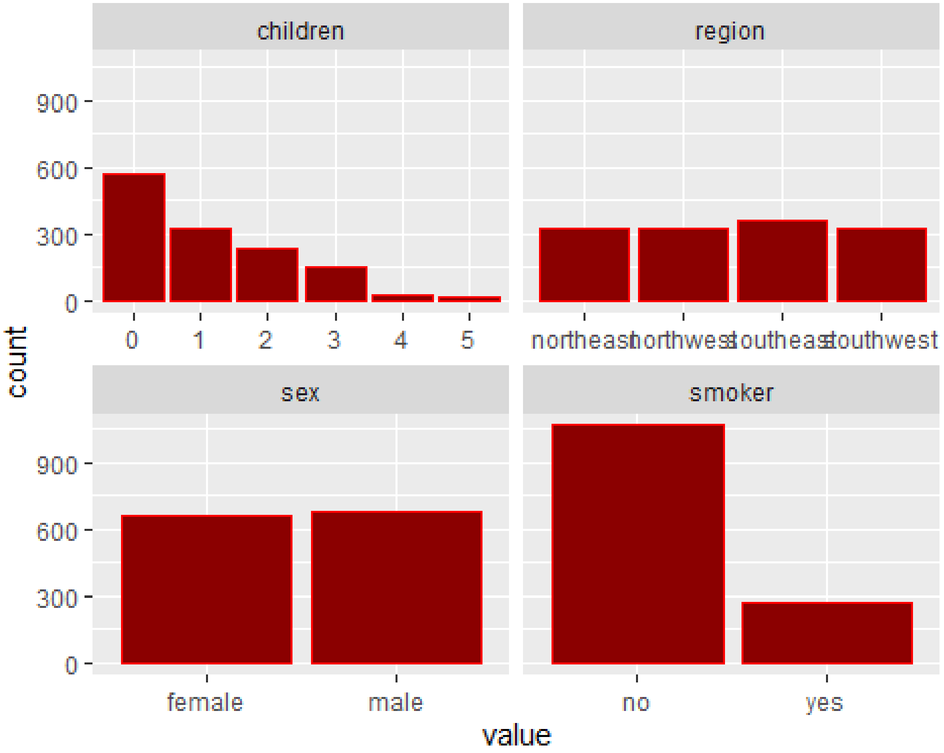
Histogram

```
insurance_hist <- insurance %>%
  select(c(1, 3, 7)) %>%
  gather()

hist <- ggplot(data = insurance_hist, mapping = aes(x = value)) +
  geom_histogram(bins = 10, color="blue", fill="darkblue") +
  facet_wrap(~key, scales = 'free_x')

hist
```



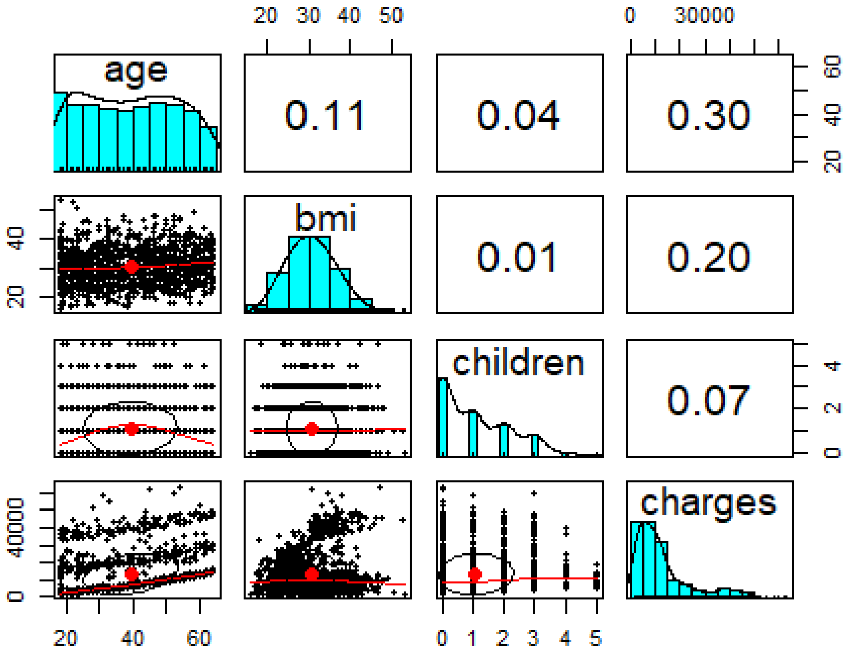
Bar chart

```
insurance_bar <- insurance %>%
  select(c(2, 4:6)) %>%
  gather()

## Warning: attributes are not identical across measure variables;
## they will be dropped

barchat <- ggplot(data = insurance_bar,mapping = aes(x = value), colorspaces)
+
  geom_bar(colour= "red" , fill= "darkred") +
  facet_wrap(~key, scales = 'free_x')

barchat
```

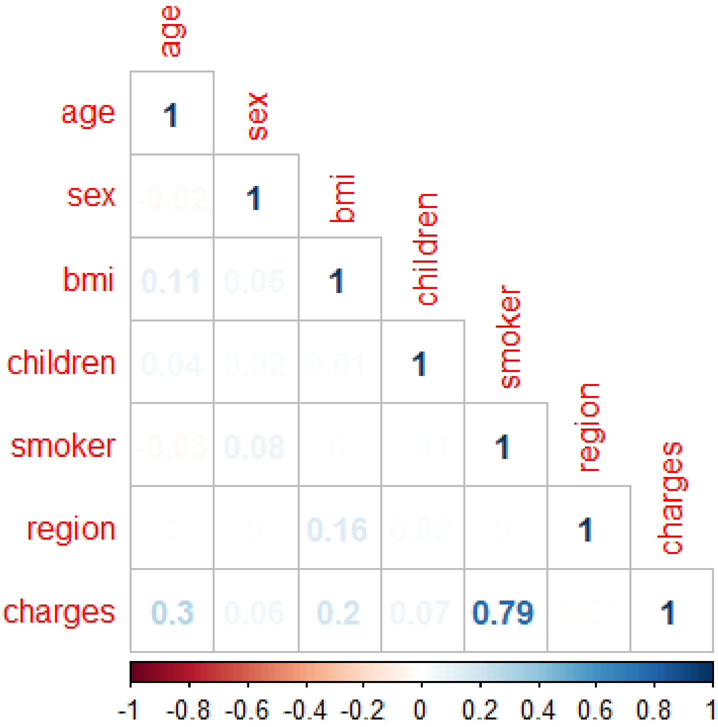


correlation

```
pairs.panels(insurance[c("age", "bmi", "children", "charges")])
```



```
Corr_ins <- mutate_all(insurance,
                       funs(as.numeric))

## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

corrplot(cor(Corr_ins), method = "number",
         type = "lower")
```



Correlation with dependent variable

```
corp <- apply(Corr_ins[, -7], 2, function(x)
  cor.test(x, y=Corr_ins$charges)$p.value)

cor_table <- cor(Corr_ins[, -7], Corr_ins$charges)
```

```r
kable(cbind(as.character(corp),cor_table),
      col.names = c("P value", "Correlation with dependent variable"))
```

	P value	Correlation with dependent variable
age	4.88669333171859e-29	0.299008193330648
sex	0.0361327210059298	0.0572920622020254
bmi	2.45908553511669e-13	0.198340968833629
children	0.0128521285201365	0.0679982268479048
smoker	8.2714358421744e-283	0.787251430498477
region	0.82051783646525	-0.00620823490944446

Model_1

```r
model_1 <- lm(formula = charges ~ .,
         data = insurance)
summary(model_1)

##
## Call:
## lm(formula = charges ~ ., data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## 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: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

model_2

```r
model_2 <- lm(formula = charges ~ . -sex -region,
         data = insurance)
summary(model_2)
```

```
## 
## Call:
## lm(formula = charges ~ . - sex - region, data = insurance)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11897.9  -2920.8   -986.6   1392.2  29509.6 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
## age            257.85      11.90  21.675  < 2e-16 ***
## bmi            321.85      27.38  11.756  < 2e-16 ***
## children       473.50     137.79   3.436 0.000608 ***
## smokeryes    23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489 
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Anova testing

```
anova(model_2, model_1)
```

```
## Analysis of Variance Table
## 
## Model 1: charges ~ (age + sex + bmi + children + smoker + region) - sex - 
##     region
## Model 2: charges ~ age + sex + bmi + children + smoker + region
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   1333 4.9078e+10                           
## 2   1329 4.8840e+10  4 238917273 1.6253 0.1654
```

The first model perform better than the newer model, so we will use model_1 for our predictions.

```
set.seed(1234) #setting seed to reproduce result of random sampling
train <- insurance %>%
  sample_frac(., size = 0.8, replace = F)
test <- anti_join(insurance, train)
```

```
## Joining, by = c("age", "sex", "bmi", "children", "smoker", "region",
"charges")
```

```
model_1 <- lm(formula = charges ~ .,
             data = train)
predicted.charges <- predict(object = model_1,
                             newdata = test, type = "response")
```

```r
results.df <- data.frame(cbind(actuals = test$charges, predicted =
predicted.charges))

results.df <- results.df %>%
  mutate(error = results.df$actuals - results.df$predicted) %>%
  round(., 2)
results.df <- results.df %>%
  mutate( error_percent =
paste0(round(results.df$error/results.df$actuals*100,2),"%"))

kable(head(results.df))
```

actuals	predicted	error	error_percent
3866.86	5841.86	-1975.00	-51.08%
27808.73	35836.03	-8027.30	-28.87%
39611.76	31919.85	7691.91	19.42%
1837.24	819.40	1017.83	55.4%
2395.17	2181.64	213.53	8.92%
13228.85	15968.60	-2739.76	-20.71%

```r
sprintf("The Average percent error is: %s%%",
round(mean(results.df$error/results.df$actuals*100), 2))
```

```
## [1] "The Average percent error is: -19.99%"
```

Our model was able to predict the premium insurance for policy holders with a mean difference of ~19%.

While sex and region have no major contributors to the model, the model without those variables actually performed slightly worse. therefore, if region was further broken down by state, it may provide more accuracy.

Result: smoker is highly correlated with charges - however, a smoker is very likely to have a higher premium.