Q1.

The given dataset contains sample taken from insurance holders of 1471 patients records along with there are characteristics and decided premium.

1.1.Understanding insurance data set and its structure.

Exploring dimensions of given data set

```
> dim(data)
[1] 1471    8
> 
```

The given data set includes 1471 records with 8 variables. Following output shows the variable names.

```
> variable.names(data)
[1] "X"              "age"            "gender"         "bmi"            "num_kids"       "smoking_s
tatus"
[7] "district"       "premium"
> 
```

```
> attributes(data)
$names
[1] "X"              "age"            "gender"         "bmi"            "num_kids"
[6] "smoking_status" "district"       "premium"

$class
[1] "data.frame"
```

Below output shows the first six record and last six records of data set

```
> head(data)
  X age gender    bmi num_kids smoking_status district   premium
1 1  44 female 20.235        1            yes  badulla 19594.810
2 2  49 female 41.470        4             no   trinco 10977.206
3 3  29   male 35.500        2            yes  colombo 44585.456
4 4  57   male 34.010        0             no    galle 11356.661
5 5  36   male 28.880        3             no  badulla  6748.591
6 6  40 female 23.370        3             no  badulla  8252.284
```

```
> tail(data)
        X age gender    bmi num_kids smoking_status district   premium
1466 1466  24   male 26.790        1             no    galle 12609.887
1467 1467  46 female 28.900        2             no  colombo  8823.279
1468 1468  60 female 30.500        0             no  colombo 12638.195
1469 1469  58   male 35.700        0             no  colombo 11362.755
1470 1470  39 female 34.100        3             no  colombo  7418.522
1471 1471  62   male 30.875        3            yes    galle 46718.163
> 
```

When considerig the above firest and last few records, it can clearly see that “X” varible represent the number of the patiient, and there is no importancy of “X” varible for the analysis. So when doing descriptive and model fitting it should be remove the “X” variable.

Below output shows the data types of given data set. Gender and Smoking Status has two levels. District variable include 4 levels.

```
> str(data)
'data.frame':   1471 obs. of  8 variables:
 $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ age           : int  44 49 29 57 36 40 55 20 53 58 ...
 $ gender        : Factor w/ 2 levels "female","male": 1 1 2 2 2 1 2 2 2 1 ...
 $ bmi           : num  20.2 41.5 35.5 34 28.9 ...
 $ num_kids      : int  1 4 2 0 3 3 0 0 1 0 ...
 $ smoking_status: Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 1 1 1 ...
 $ district      : Factor w/ 4 levels "badulla","colombo",..: 1 4 2 3 1 1 1 1 2 4 ...
 $ premium       : num  19595 10977 44585 11357 6749 ...
>
```

Below output shows the summary of the given data set.

```
> summary(data)
       X               age            gender          bmi           num_kids     smoking_status
 Min.   :   1.0   Min.   :18.00   female:747   Min.   :16.82   Min.   :0.000   no :1188
 1st Qu.: 368.5   1st Qu.:26.00   male  :724   1st Qu.:26.60   1st Qu.:0.000   yes: 283
 Median : 736.0   Median :39.00                Median :30.50   Median :1.000
 Mean   : 736.0   Mean   :39.19                Mean   :30.92   Mean   :1.058
 3rd Qu.:1103.5   3rd Qu.:51.00                3rd Qu.:35.10   3rd Qu.:2.000
 Max.   :1471.0   Max.   :64.00                Max.   :53.13   Max.   :5.000
    district       premium
 badulla:347   Min.   : 1132
 colombo:356   1st Qu.: 4456
 galle  :378   Median : 9447
 trinco :390   Mean   :13119
               3rd Qu.:16069
               Max.   :62593
```

The given dataset does not have any missing values. The below output shows the dimension of dataset without missing values. Since dimension of dataset without missing values is same as dimension of original data set, there are no missing values in this data set.

```
> dim(data[complete.cases(data),])
[1] 1471    8
>
```

```
> any(is.na(data))
[1] FALSE
>
```

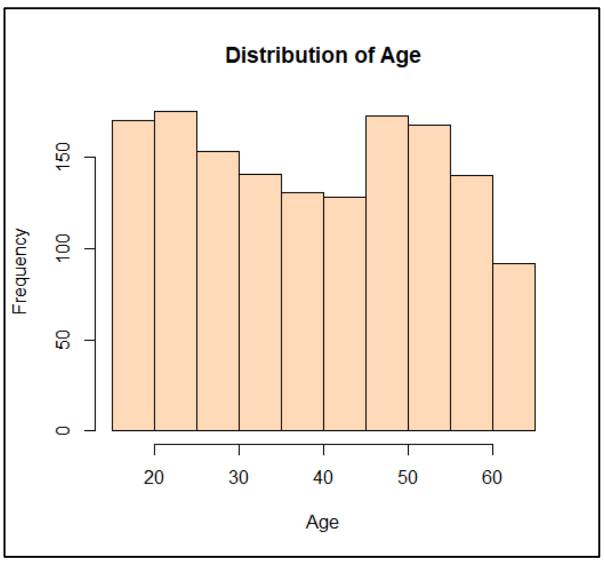
1.2.Exploring individual variables

1.2.1. Age

The patients in the data set has minimum of 18 year old, and maximum if 64 years old. Therefore, the range of age is 46 years. Average year of a patient is about 39 years and median is 39. So mean and median is very close to each other. Age has standard deviation of 14 years. It can be say that variance is considerably higher.

```
> summary(data$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   26.00   39.00   39.19   51.00   64.00 
> sd(data$age)
[1] 14.08868
```

Following figures shows the distribution of Age



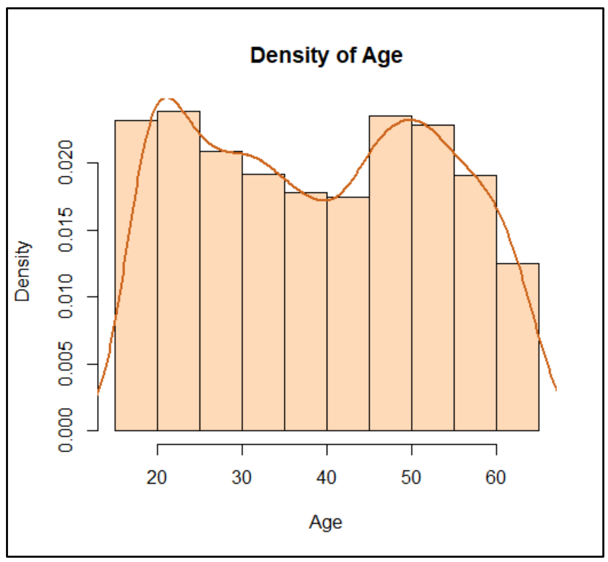


Figure 1.1: Histogram for Age Figure 1.2. Density plot for Age

When considering the above histogram in figure 1.1 it can be seen that frequency of age 20 -25 and 45-50 are high. Frequency of patients above 60 are quite small. As per figure 1.2. Density plot, it seems a bimodal distribution with two picks.

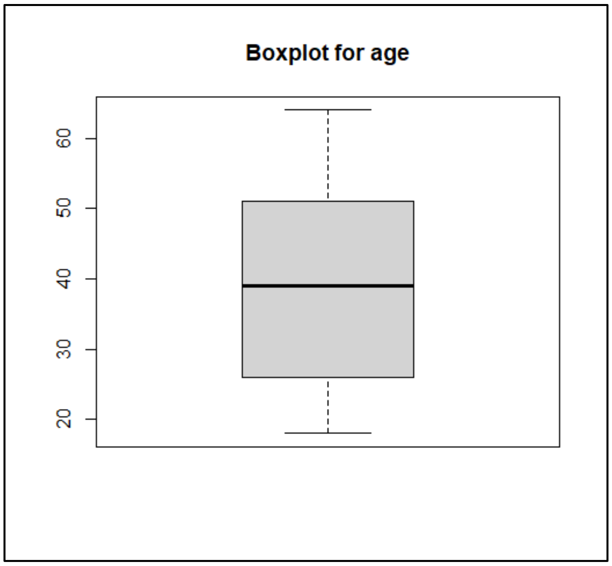


Figure 1.3: Box plot for Age

```
> boxplot.stats(data$age)
$stats
[1] 18 26 39 51 64

$n
[1] 1471

$conf
[1] 37.97011 40.02989

$out
integer(0)
```

25% of patients ages fall below the lower quartile 18 years. 75% of patients fall below the 51 years of age. 50% of patient’s age lies between 26 and 51. 50% of patient ages higher than the 64 years and lower than 18 years.

1.2.2. Gender

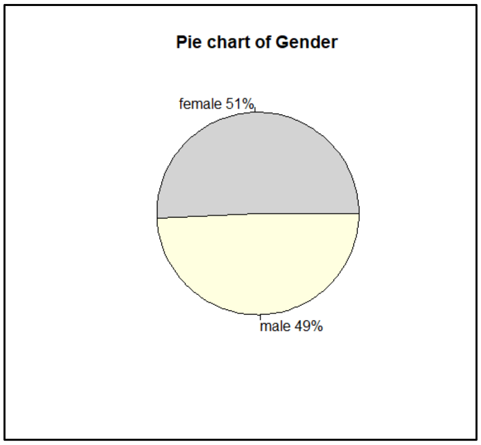
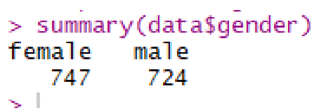


Figure 1.4: Pie chart of Gender



As per figure 1.4, there are 51% of female and 49% males are in the given data set. So the proportion of male and female are very much close to each other.

1.2.2. BMI (Body Mass Index)

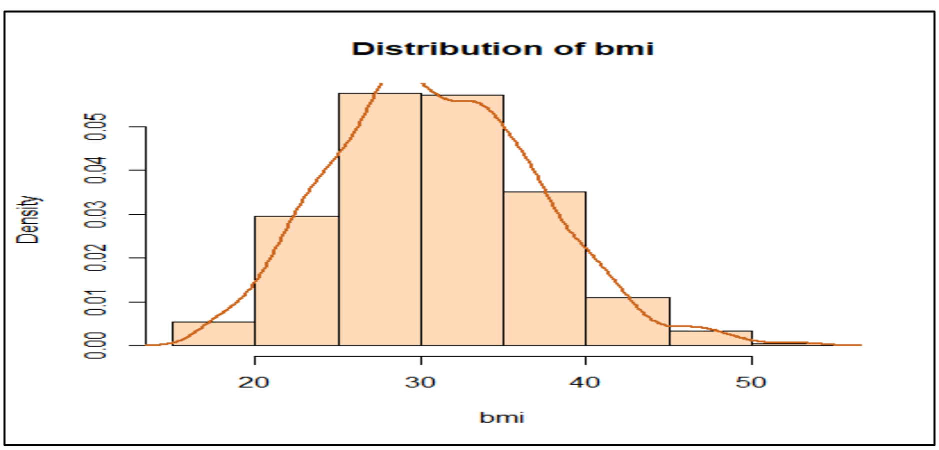


Figure 1.5: Distribution of BMI

```
> summary(data$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.82   26.60   30.50   30.92   35.10   53.13
>
```

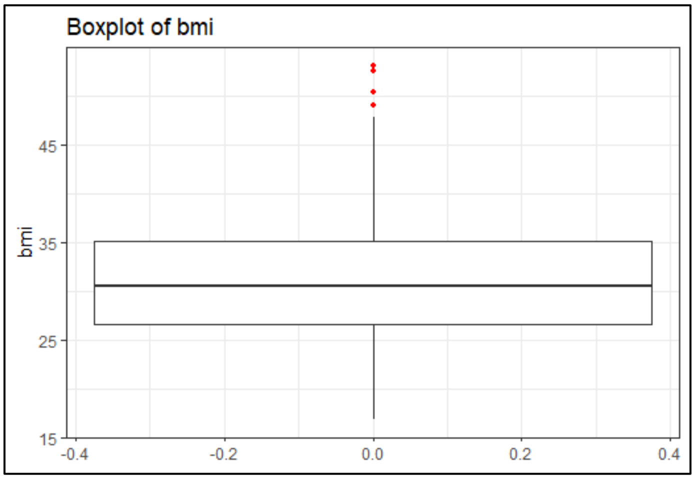


Figure 1.6: Boxplot for BMI

```
> boxplot.stats(data$bmi)
$stats
[1] 16.815 26.600 30.500 35.100 47.740

$n
[1] 1471

$conf
[1] 30.14984 30.85016

$out
[1] 49.06 53.13 50.38 52.58 49.06 49.06 53.13 49.06
```

25% of patients BMI value fall below the lower quartile 16.8. 75% of patients BMI fall below the 35.1. 50% of patient's BMI lies between 26.6 and 35.1 . 50% of patient BMI higher than the 47.74 years and lower than 16.81.

Below show the patients details that in outliers as BMI,

```
> data[which(data$bmi %in% outliers),]
        X age gender   bmi num_kids smoking_status district   premium bmi_ranges
142   142  58   male 49.06        0             no   trinco 11381.325      obese
359   359  18   male 53.13        0             no   trinco  1163.463      obese
489   489  23   male 50.38        1             no   trinco  2438.055      obese
639   639  22   male 52.58        1            yes   trinco 44501.398      obese
710   710  58   male 49.06        0             no   trinco 11381.325      obese
1246 1246  58   male 49.06        0             no   trinco 11381.325      obese
1247 1247  18   male 53.13        0             no   trinco  1163.463      obese
1391 1391  58   male 49.06        0             no   trinco 11381.325      obese
```

All the patients BMI values that consider as outliers are belong to trincomalee district, obese and male patients.

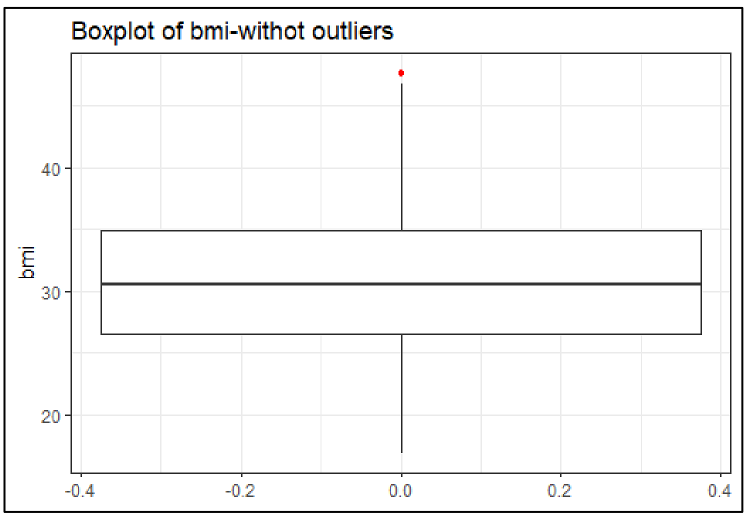


Figure: Boxplot for BMI after removing outliers

Above figure shows the box plot after removing the outliers from BMI. New data set was created by removing BMI outliers and will use if needed when fitting the models.

Further BMI value coded as categorical variable (BMI_ranges) as per below basis,

BMI for Adults

below 18.5 = Underweight

18.5-24.9 = Normal or Healthy Weight

25.0-29.9 = Overweight

30.0 or Above = Obese

(Source: https://en.wikipedia.org/wiki/Body_mass_index)

Following information shows the distribution of patients as per BMI ranges

```
bmi_range
under_weight normal_weight over_weight obese
          22           234         425   790
```

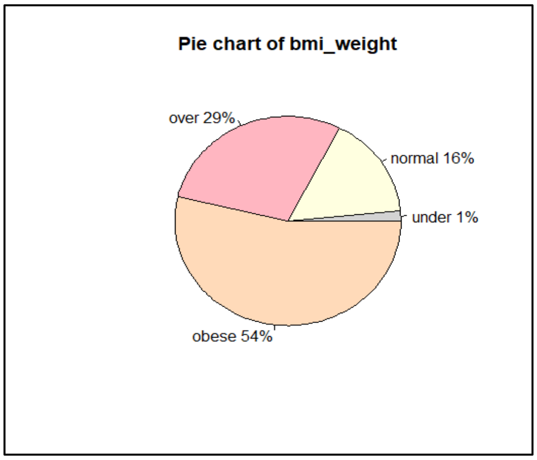


Figure 1.7: Pie chart for BMI_ranges

As per figure 1.7, 54% of the patients are obese, only 1% of them are under weight. Only 16% of the patients are having normal weight that is correct weight per height.

4. number of kids

```
A tibble: 6 x 2
num_kids counts
   <int>  <int>
       0    641
       1    367
       2    251
       3    173
       4     26
       5     13
```

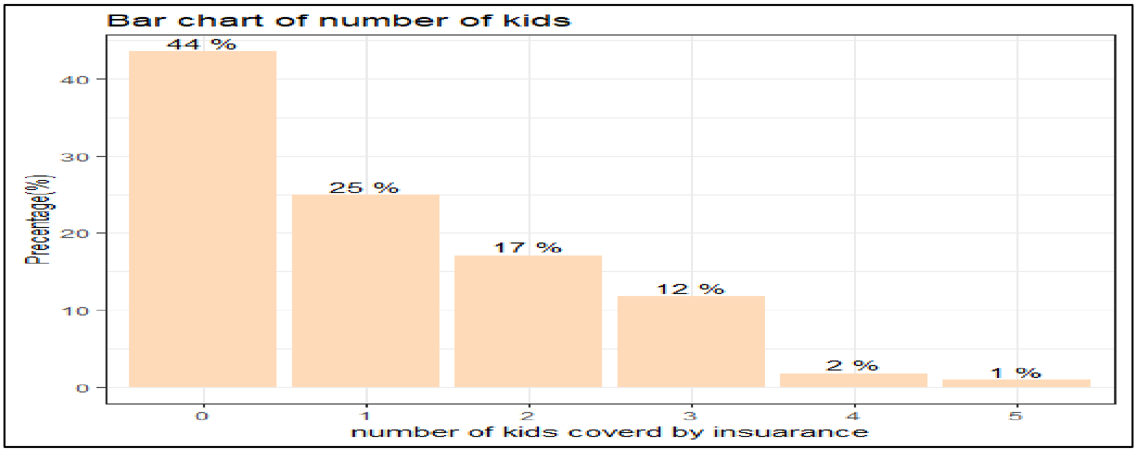


Figure 1.8: Bar chart for number of kids covering by insurance.

As per bar chart in figure 1.8, only 1% of the patients have 5 children coverage insurance. Majority (44%) of patient only have insurance that does not cover children.

Further number of kids covering is recoded to two categories as below,

Num kids =0 , then 0=child cover no

Num kids >0, then 1=child cover yes

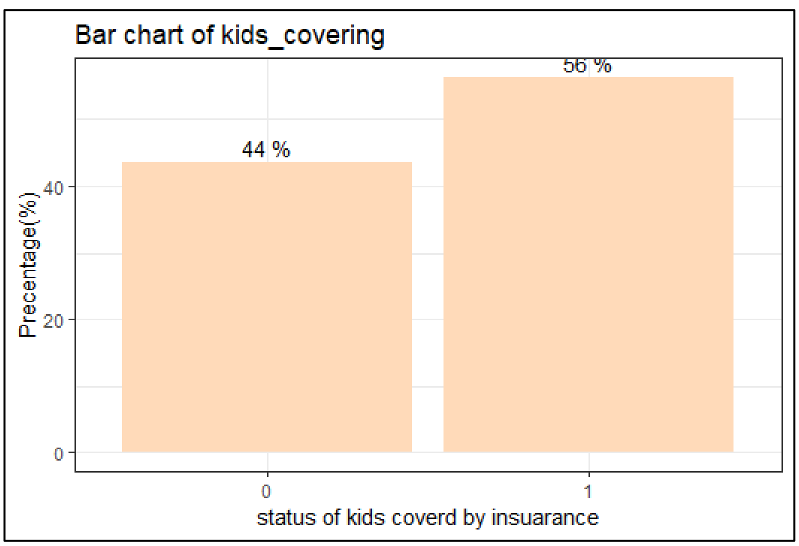


Figure 1.9 : Bar chart for status if covering children

As per figure 1.9, after recoded the number kids, 44% of patient have insurance without child coverage, 56% patient have insurance with children coverage.

5. Exploring smoking status

```
> levels(data$smoking_status)
[1] "no"  "yes"
>
```

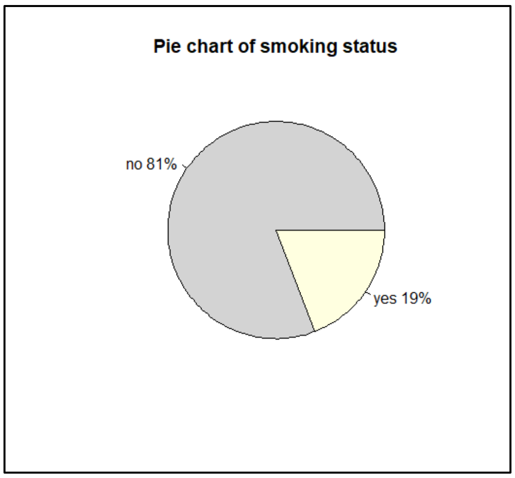


Figure 1.9: Pie chart of smoking status

81% of patients are nonsmokers. Only 19% of them are smoking

6. Exploring district

```
> levels(data$district)
[1] "badulla" "colombo" "galle"   "trinco"
>
```

```
> districtTable
     Var1 Freq
1 badulla  347
2 colombo  356
3   galle  378
4  trinco  390
>
```

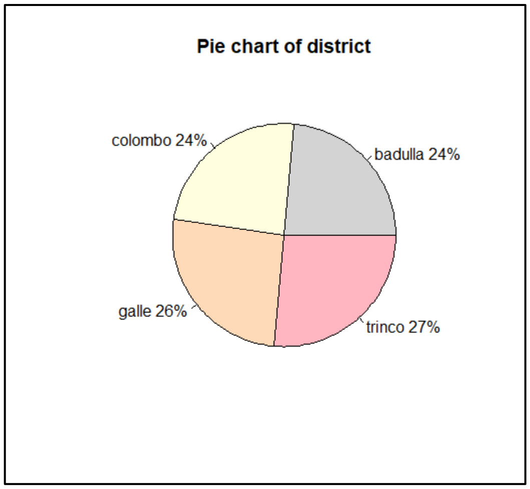


Figure 1.10: Pie chart of district

The distribution of patient among district are approximately same. It can be say that patents are equally represent their district.

7. Exploring premium

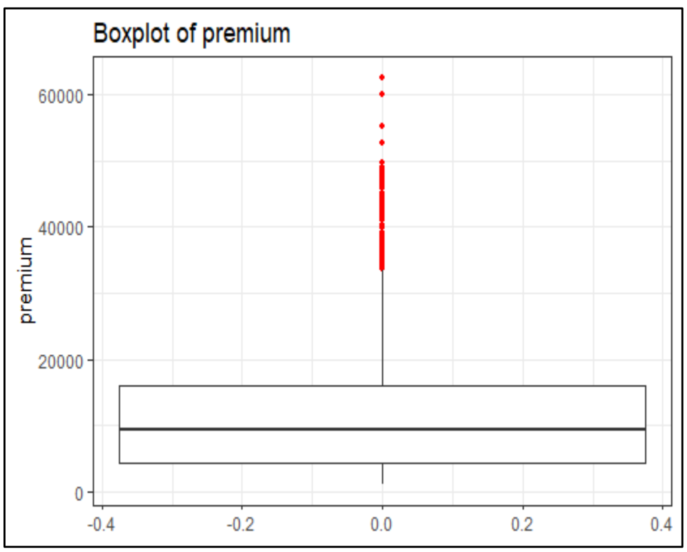
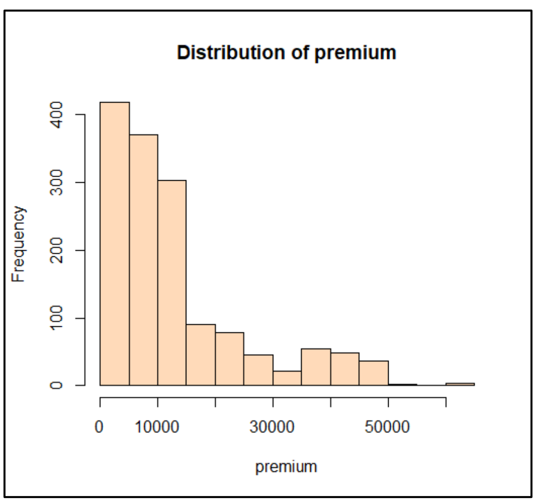


Figure 1.11: Boxplot for premium

```
> summary(data$premium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1132    4456    9447   13119   16069   62593 
> 
```

```
> boxplot.stats(data$premium)
$stats
[1]  1131.507  4456.092  9447.250 16069.085 33471.972

$n
[1] 1471

$conf
[1] 8968.846 9925.654

$out
  [1] 44585.46 47462.89 36021.01 46151.12 39241.44 42760.50 52590.83 37484.45 60021.40 40904.20
 [11] 48675.52 39774.28 36189.10 42560.43 36021.01 43753.34 41676.08 39983.43 42856.84 46255.11
 [21] 39871.70 42112.24 35147.53 45008.96 43943.88 38792.69 41034.22 48970.25 48549.18 42111.66
 [31] 34254.05 43813.87 43254.42 42760.50 43943.88 40974.16 42760.50 35147.53 43753.34 44400.41
 [41] 43753.34 38245.59 38709.18 34779.61 46718.16 43578.94 55135.40 41661.60 60021.40 36085.22
 [51] 37079.37 42112.24 38711.00 41949.24 47305.31 38792.69 36898.73 52590.83 41999.52 47403.88
 [61] 33732.69 46718.16 40932.43 36189.10 48173.36 38282.75 43813.87 37079.37 44501.40 36219.41
 [71] 38792.69 37742.58 33732.69 43813.87 45863.21 34617.84 36950.26 46661.44 38415.47 48173.36
 [81] 37607.53 62592.87 39774.28 39836.52 36085.22 40273.65 42760.50 40904.20 48824.45 36149.48
 [91] 36307.80 38711.00 36910.61 38711.00 41034.22 45863.21 45702.02 40941.29 34617.84 40103.89
[101] 42760.50 47496.49 62592.87 38711.00 38511.63 38126.25 36085.22 35160.13 44202.65 45008.96
[111] 44202.65 48549.18 41097.16 43896.38 35491.64 34838.87 47928.03 42983.46 36149.48 42969.85
[121] 39725.52 49577.66 40974.16 41661.60 38746.36 48675.52 48885.14 39774.28 35147.53 40932.43
[131] 46113.51 34254.05 47269.85 43753.34 45008.96 46113.51 34166.27 47291.06 39871.70 36307.80
[141] 39727.61 45008.96 46113.51 34439.86 38282.75 41919.10 46130.53 37607.53 36910.61 48675.52
[151] 36950.26 34779.61 33900.65 33907.55 39774.28 46718.16 39725.52 48173.36 43753.34 46718.16
```

25% of patients premium fall below the lower quartile 1131.5. 75% of patients premium fall below the 16069. 50% of patient's premium lies between 4456 and 16069 . 50% of patient's premium higher than the 33471.9 and lower than 1131.5.



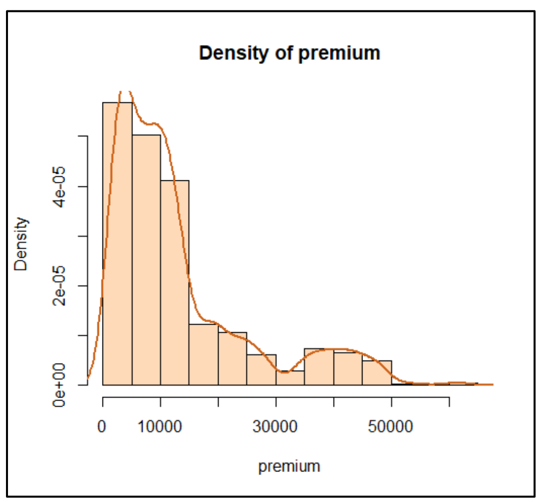
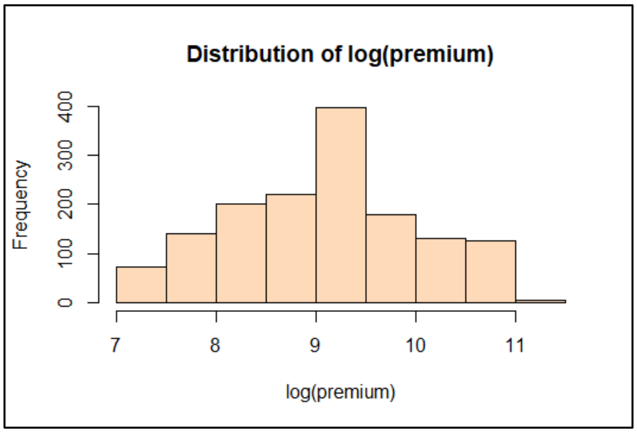


Figure 1.12: Histogram for premium   Figure 1.13: Density plot for premium

When considering the distribution of premium as per figure 1.12 and figure 1.13, it can see the distribution of premium is positively skewed.

Taking log transformation of premium and as per below plots the distribution become approximately normal.



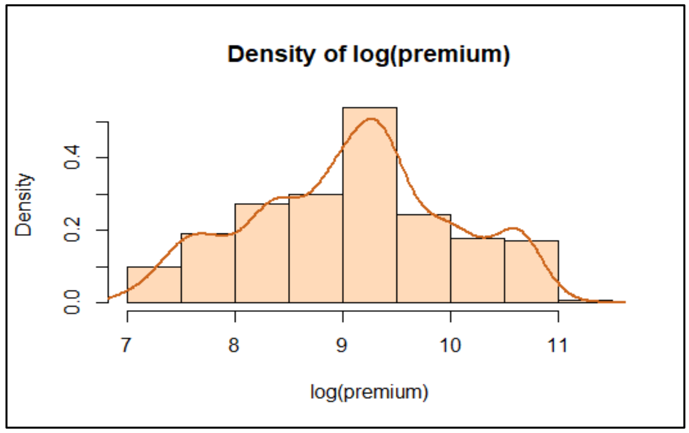


Figure 1.14: Histogram for log(premium)   Figure 1.15: Density plot for log(premium)

2. exploring multiple variables

2.1 Age vs BMI

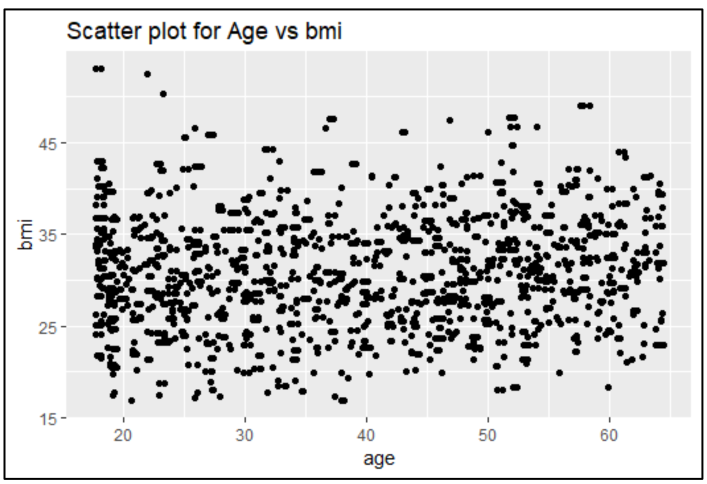


Figure 1.16 : Scatterplot for age vs BMI

As per above figure 1.16, there is no considerable relationship between age and BMI. Below correlation test also gives r= 0.08 which is not a strong correlation though the test is significant ($p<0.05$).

```
> res_agevbmi <-cor.test(data$age, data$bmi)
> res_agevbmi

	Pearson's product-moment correlation

data:  data$age and data$bmi
t = 3.2162, df = 1469, p-value = 0.001327
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.03264904 0.13415629
sample estimates:
       cor
0.08361958
```

To check the relation between age and BMI_ranges , spearman rank correlation test carried out.

```
> res_agevbmi_ranges

	Spearman's rank correlation rho

data:  data$age and data$bmi_ranges
S = 493990000, p-value = 0.008282
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.06881894
```

Above test also indicate the significant relationship but with poor correlation coefficient value.

2.2. Age vs Premium

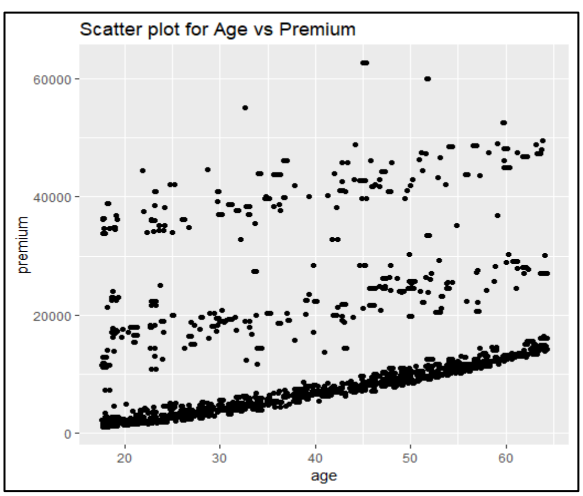
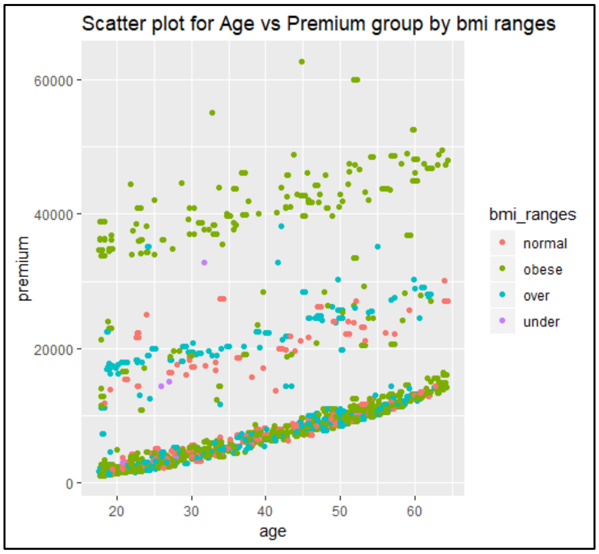


Figure 1.17 : Scatterplot for age vs premium

As per figure 1.17, it can be clearly see that there is a positive relationship between age and premium. In real world, also when age is high premium goes high. When considering the above scatterplot it can be there are three clusters in the premium and age.



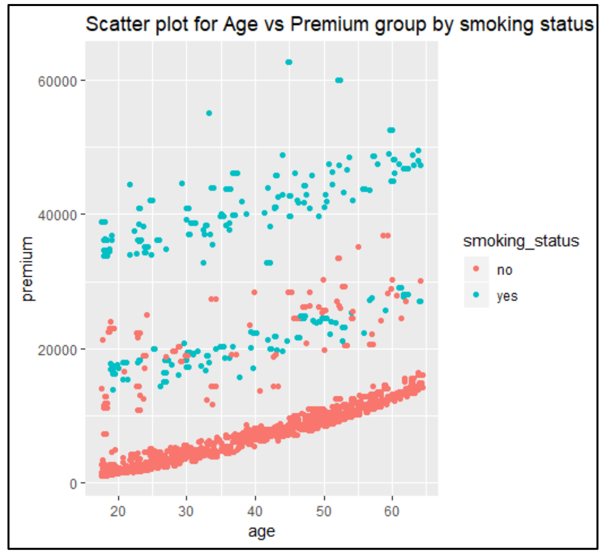


Figure 1.18: Scatter plot color by BMI_ranges

Figure 1.19: Scatterplot color by Smoking

The figure 1.18, shows the scatterplot for age vs premium separated by BMI_ranges. The three lines can see in the graph, the top line is consist with the patients with obese weight. The figure 1.19 shows the same scatter plot separated by smoking status. In there the top line observations are belong to patient with no smoking. Moreover, the below line consist with patients with smoking status yes.

As per person correlation result in below, the relationship is significant (P<0.05), and r=0.3 indicate there is somewhat strong relationship between age and premium.

```
> res_premiumvage

	Pearson's product-moment correlation

data:  data$age and data$premium
t = 12.357, df = 1469, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2598273 0.3524447
sample estimates:
      cor 
0.3068623 
```

Further to increase the linearity and overcome the clustering effect, examine the relationship between log(premium) and sqrt(age) transformation

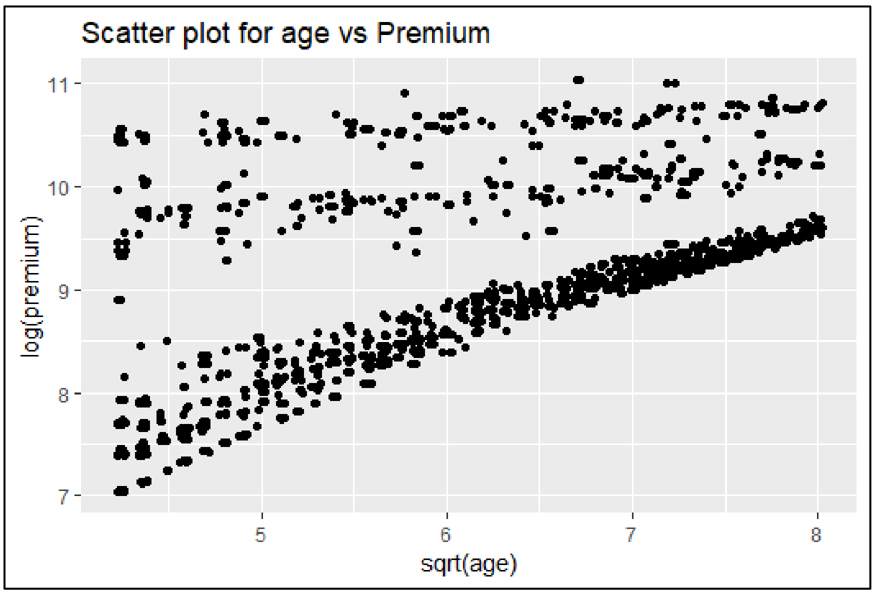


Figure 1.20: Scatter plot for log(premium) vs sqrt(age)

Figure 1.20 shows the scatterplot for log(premium) and sqrt(age), it can ve seen that the linearity of the relationship is improve than the figure 1.19. also correlation coefficient increase to 0.5

```
	Pearson's product-moment correlation

data:  sqrt(data$age) and log(data$premium)
t = 24.692, df = 1469, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5044316 0.5767252
sample estimates:
      cor 
0.5415789 
```

2.3.Gender vs smoking status

```
table(data$gender,data$smoking_status)

         no yes
female 619 128
male   569 155
```

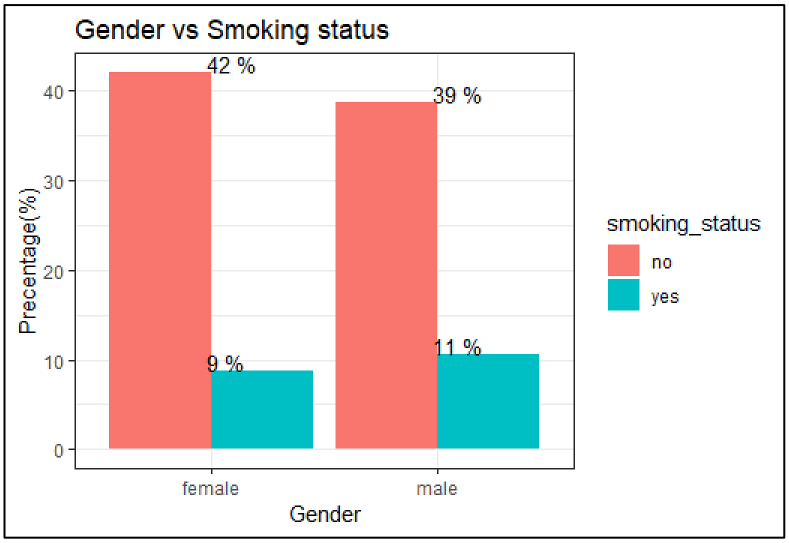


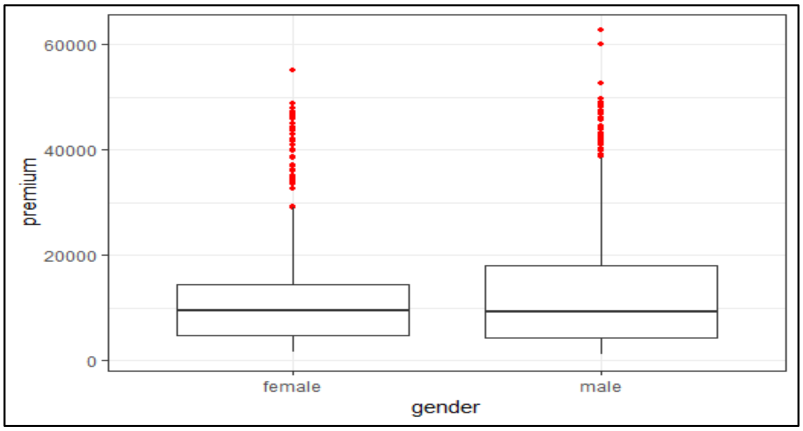
Figure 1.21: Gender vs Smoking status

```
> res_genderVsmoking

        Pearson's Chi-squared test with Yates' continuity correction

data:  data$gender_code and data$smoking_code
X-squared = 4.0511, df = 1, p-value = 0.04414
```

As per figure 1.21, 42% of the patients are female nonsmokers. 9% of patients are female smokers.

2.4. Gender vs premium

```
> aggregate(premium ~ gender, summary, data=data)
  gender premium.Min. premium.1st Qu. premium.Median premium.Mean premium.3rd Qu. premium.Max.
1 female     1622.188        4787.630       9549.565    12478.383       14453.740    55135.402
2   male     1131.507        4239.201       9382.033    13779.438       17942.106    62592.873
```



1.22.: box plot for gender vs premium

Above figure 1.22, shows the box plot for premium as gender wise. The premium range is higher for males than female.

```
> res_gendervsprm

        Spearman's rank correlation rho

data:  data$gender_code and data$premium
S = 521310000, p-value = 0.5067
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01732392
```

As per above spearman rank correlation test, the relationship is not significant (p>0.05). So there is no enough evidence to say that there is a relationship between gender and premium.

2.5. BMI vs smoking status

```
> aggregate(bmi ~ smoking_status, summary, data=data)
  smoking_status bmi.Min. bmi.1st Qu. bmi.Median bmi.Mean bmi.3rd Qu. bmi.Max.
1             no 16.81500    26.40000   30.49500 30.82493    34.80000 53.13000
2            yes 17.19500    26.99250   30.87500 31.34410    36.30000 52.58000
> table(data$smoking_status,data$bmi_ranges)

      normal obese over under
  no     188   637  345    18
  yes     46   153   80     4
>
```

Mean and median BMI value of smokers and nonsmokers are approximately same. As per below rank correlation test, there is no significant relationship between BMI and smoking status.

```
> res_bmivsmoking

	Spearman's rank correlation rho

data:  data$bmi and data$smoking_code
S = 515530000, p-value = 0.2793
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.02822499
```

2.6.BMI vs premium

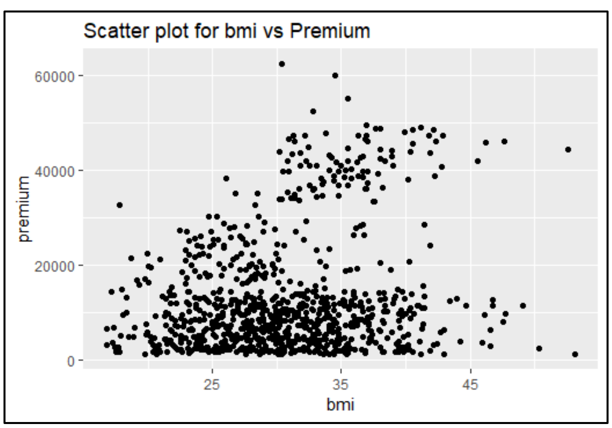


Figure1.23: Scatter plot BMI vs premium

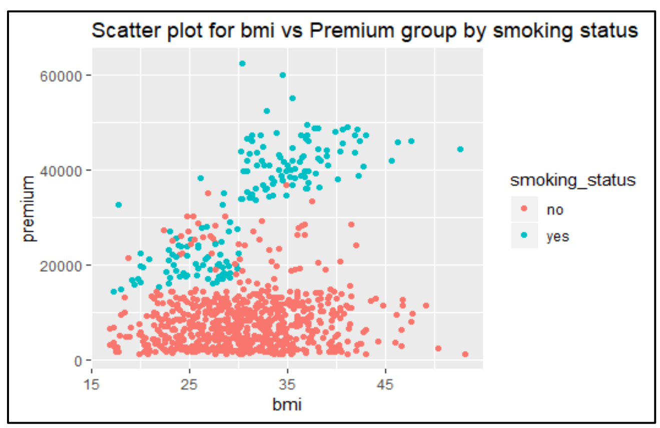


Figure1.24: Scatterplot BMI vs premium by Smoking status

As per figure 1.23 there is no strong evidence to say about strong liner relationship. But when BMI increase beyond 25 , then there is a increase of premium. As per figure 1.24, the BMI values of nonsmokers does not have clear increase trend with premium. But when considering smokers, there is an increase in premium with BMI.

```
> res_bmivspremium

        Pearson's product-moment correlation

data:  data$bmi and data$premium
t = 7.4662, df = 1469, p-value = 1.409e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1414774 0.2399698
sample estimates:
      cor
0.1912049
```

As per correlation test, there is a significant relationship between premium and BMI ($p<0.05$).

Correlation coefficient (r-0.19) indicate the positive linear relationship

2.7. Covariance and correlation in insurance data

Below is the Covariance and correlation matrix for insurance dataset

```
> cov(data[,c(2,4,5,8)])
                  age          bmi     num_kids      premium
age      1.984908e+02 7.318713e+00 8.651341e-01 5.228734e+04
bmi      7.318713e+00 3.859341e+01 4.264719e-02 1.436608e+04
num_kids 8.651341e-01 4.264719e-02 1.374811e+00 1.650874e+03
premium  5.228734e+04 1.436608e+04 1.650874e+03 1.462736e+08
>
```

```
> cor(data[,c(2,4,5,8)])
                age         bmi    num_kids   premium
age      1.00000000 0.083619579 0.052371107 0.3068623
bmi      0.08361958 1.000000000 0.005854804 0.1912049
num_kids 0.05237111 0.005854804 1.000000000 0.1164152
premium  0.30686227 0.191204855 0.116415173 1.0000000
>
```

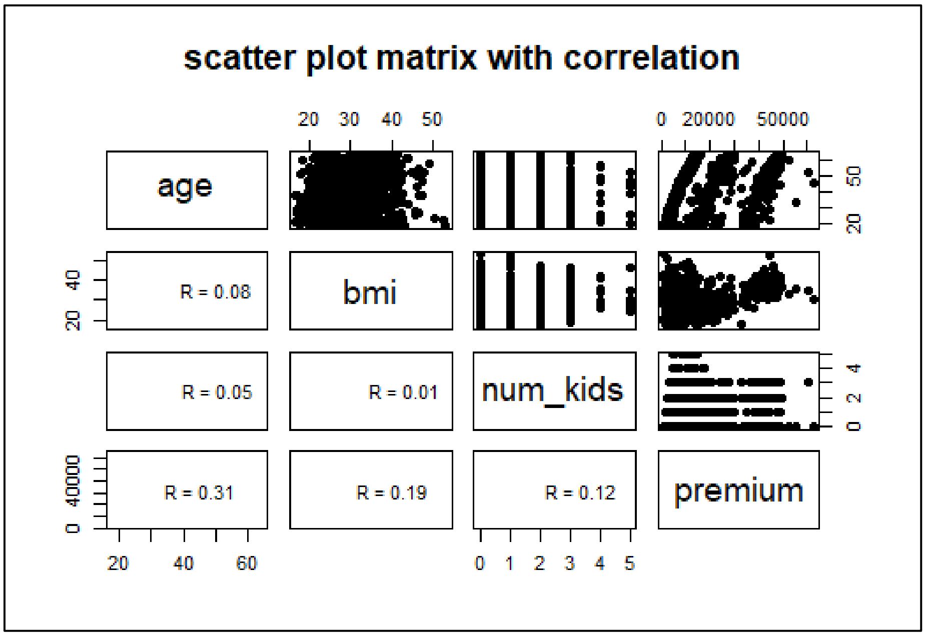


Figure 1.25: Correlation matrix for insurance data

b.

Since response variable is continues and there are more than one independent variables, use multiple regression model to predict the insurance data.

1.1. Frist fit the full model with all the possible variables

```
> summary(full.raw.model1)

Call:
lm(formula = premium ~ age + gender + bmi + num_kids + smoking_status +
    district, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-12523.0  -2664.5  -1067.6    994.3  29362.8

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -10305.09     953.34 -10.809  < 2e-16 ***
age                  252.98      11.32  22.354  < 2e-16 ***
gendermale           394.00     318.37   1.238 0.216089
bmi                  283.59      26.80  10.582  < 2e-16 ***
num_kids             515.96     135.65   3.804 0.000149 ***
smoking_statusyes  24141.21     403.24  59.868  < 2e-16 ***
districtcolombo     -973.16     459.13  -2.120 0.034209 *
districtgalle       -567.17     452.32  -1.254 0.210071
districttrinco      -994.14     461.69  -2.153 0.031461 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1462 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7483
F-statistic: 547.3 on 8 and 1462 DF,  p-value: < 2.2e-16
```

The estimated regression model is,

$\hat{Y}$= -10305.09 +252.98age + 394gendermale+ 283.59BMI+515.96num_kids+ 24141.21smoking_statusyes -973.16 district_colombo -567.17district_gall -994.14distric_trinco

The coefficient of gender male is not significant.

```
> summary(full.raw.model1)$r.squared
[1] 0.7496814
>
```

R-squared= 74.97%

Using forward selection to raw data

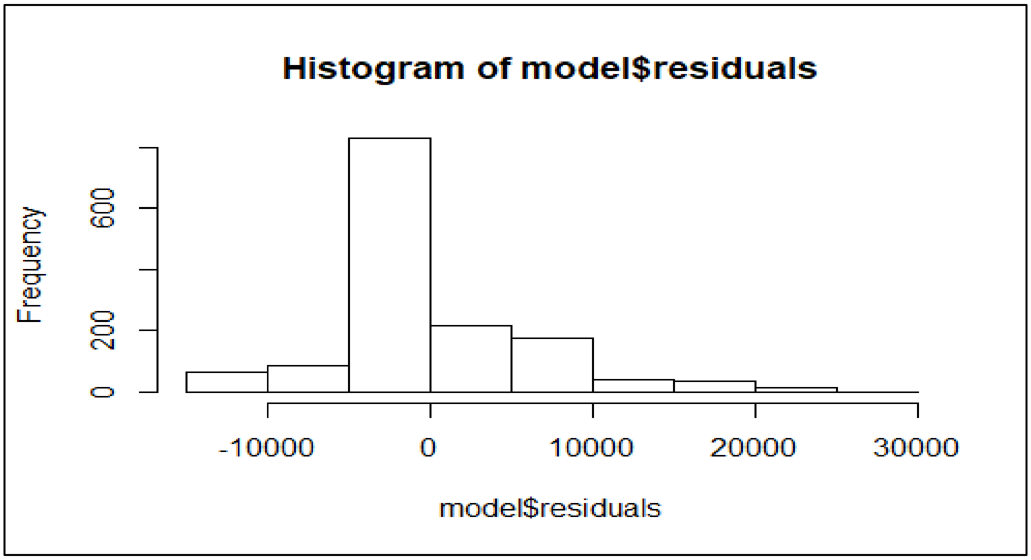
```
> summary(forwar.raw.model)

Call:
lm(formula = data$premium ~ smoking_status + age + bmi + num_kids +
    district, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-12314  -2669  -1100   1007  29547

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -10137.10     943.80 -10.741  < 2e-16 ***
smoking_statusyes  24166.75     402.78  59.999  < 2e-16 ***
age                  252.15      11.30  22.316  < 2e-16 ***
bmi                  285.63      26.75  10.677  < 2e-16 ***
num_kids             518.10     135.67   3.819  0.00014 ***
districtcolombo     -988.94     459.03  -2.154  0.03137 *
districtgalle       -589.97     452.02  -1.305  0.19204
districttrinco     -1002.15     461.73  -2.170  0.03013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6069 on 1463 degrees of freedom
Multiple R-squared:  0.7494,	Adjusted R-squared:  0.7482
F-statistic: 625.1 on 7 and 1463 DF,  p-value: < 2.2e-16
```

Final model from forward selection using step function,

$\hat{Y}$= -10137.10 + 24166.75smoking_status_yes + 252.15age +285.63BMI+518.10num_kids - 988.94district_colombo -589/97district_gall -1002.15distric_trinco

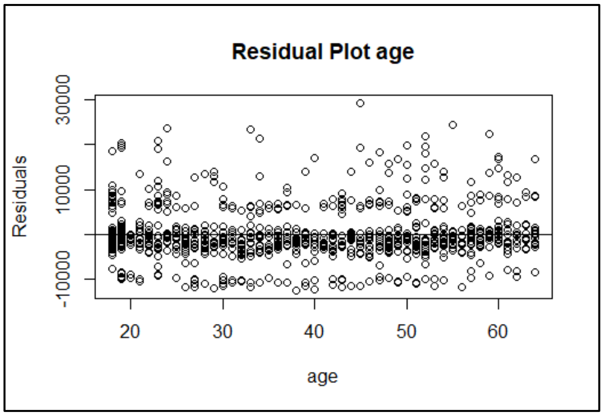
Checking model assumptions

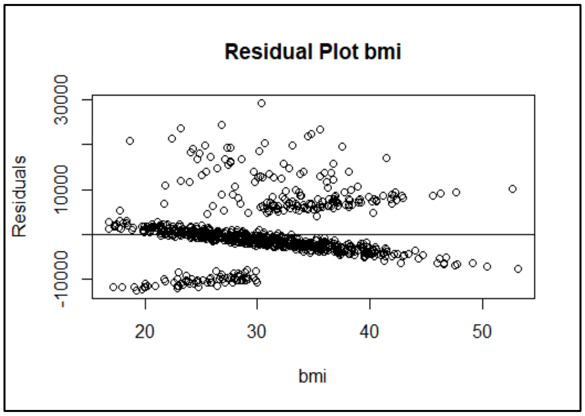
1. Normality of the residual



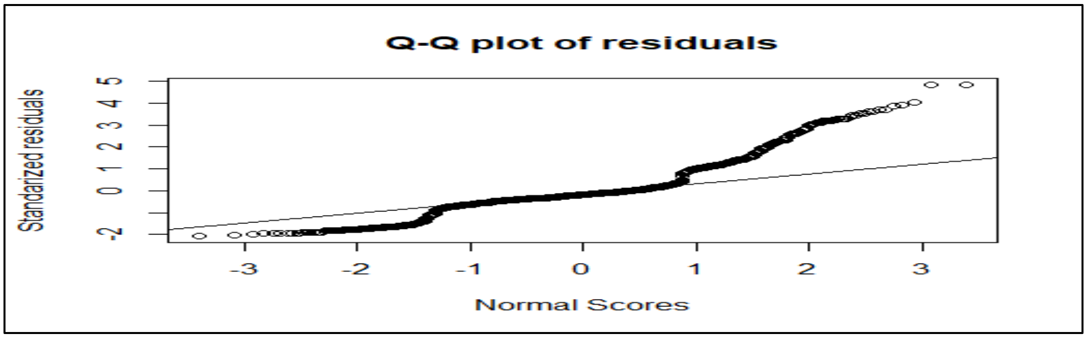
Above graph shows the shape of the distribution of residuals. It can be seen that shape is slightly skewed to right.

2.Residuals plot for independent variable

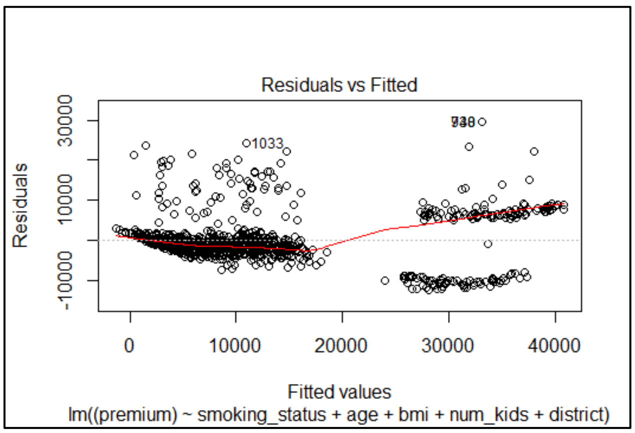


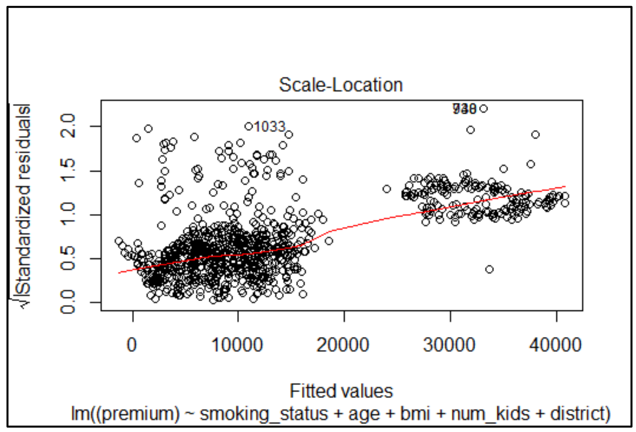


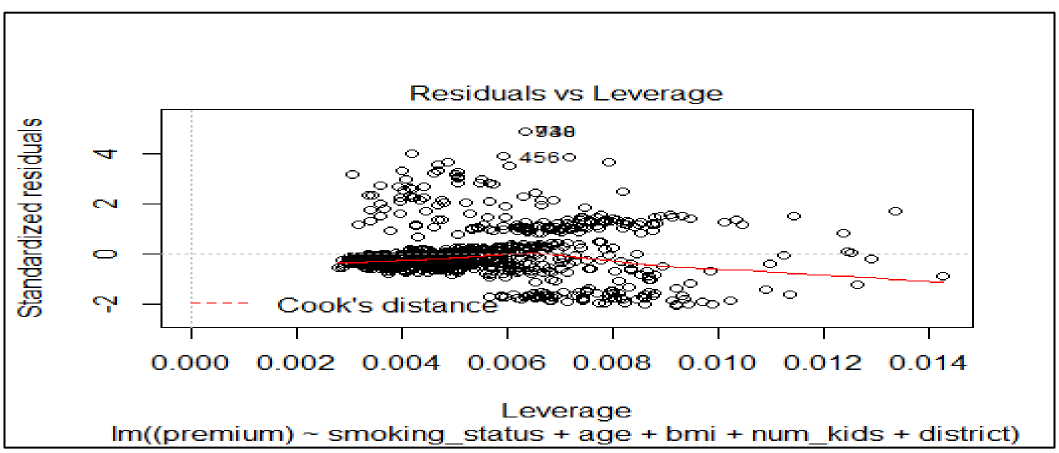
3.Normal probabolty plot of residuals for comparing residuals with normaly distribuiton.



As per above Q-Q plot, may point are not falls closer to the straight line, so the normality assumption is violated.







Below shapiro test carried out to find normality of residuals

```
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.88099, p-value < 2.2e-16

>
```

H0: Data are normally distributed
H1: Data are not normally distributed

p-value < 2.2e-16
p-value <0.05, we do not reject H1 at 5% level
Therefore, errors are not normally distributed

4.RMSE

To validate the outcome use RMSE, first Make a data frame with premium and fitted values

```
> head(premium_resid)
  data.premium fitted.value     residuls
1    19594.810    31422.277 -11827.46732
2    10977.206    15133.789  -4156.58297
3    44585.456    31529.255  13056.20057
4    11356.661    13360.021  -2003.36059
5     6748.591     8743.751  -1995.15984
6     8252.284     8178.547     73.73712
```

```
> RMSE1
[1] 6052.142
>
```

Smaller the RMSE is better

5. Autocorrelation

To check the auto correlation Durbin-Watson test carried out.

```
> dwtest(model)

	Durbin-Watson test

data:  model
DW = 1.9816, p-value = 0.3609
alternative hypothesis: true autocorrelation is greater than 0
```

Since DW=1.98 (between 1.5 and 2.5), so we can say that there is no autocorrelation. Since p value is greater than 0.05 the test statistic not significant, therefore we don't have enough evidence to sat that there is a autocorrelation.

6. Multicolinearity

To check whether there are relationships among independent variables use Variance influential factor (VIF)

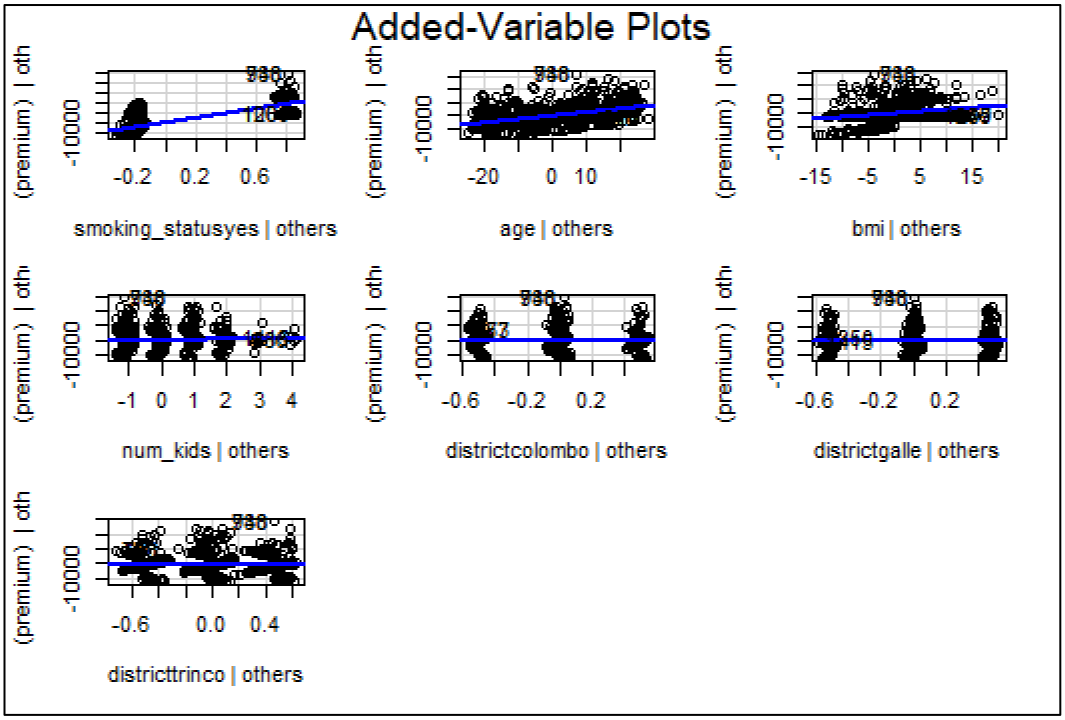
```
> vif(model)
                   GVIF Df GVIF^(1/(2*Df))
smoking_status 1.006809  1        1.003399
age            1.011543  1        1.005755
bmi            1.102507  1        1.050003
num_kids       1.009978  1        1.004977
district       1.100773  3        1.016131
> 
```

Since VIF values are lower than 5, it can be say that there is no multiclinerty exist

7.Bonferonni p-value for most extreme observations

```
> outlierTest(model)# Bonferonni p-value for most extreme obs
    rstudent unadjusted p-value Bonferonni p
740 4.922815         9.4982e-07    0.0013972
938 4.922815         9.4982e-07    0.0013972
> 
```

8. Added variable plot for check influential observation



9. Non-constant error variance test

```
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 254.1776, Df = 1, p = < 2.22e-16
>
```

C. Improving the above model

a. Take the log transformation of premium

```
> summary(forward.raw.model1)

Call:
lm(formula = log(premium) ~ smoking_status + age + num_kids + 
    district + bmi, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.01619 -0.19474 -0.05331  0.05319  2.09189 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        7.1816192  0.0718995  99.884  < 2e-16 ***
smoking_statusyes  1.5264673  0.0306845  49.747  < 2e-16 ***
age                0.0343491  0.0008608  39.904  < 2e-16 ***
num_kids           0.1026647  0.0103351   9.934  < 2e-16 ***
districtcolombo   -0.1281334  0.0349695  -3.664 0.000257 ***
districtgalle     -0.0960053  0.0344357  -2.788 0.005373 ** 
districttrinco    -0.1568315  0.0351749  -4.459 8.88e-06 ***
bmi                0.0081362  0.0020381   3.992 6.87e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4623 on 1463 degrees of freedom
Multiple R-squared:  0.7488,    Adjusted R-squared:  0.7476 
F-statistic:   623 on 7 and 1463 DF,  p-value: < 2.2e-16
```

$\text{Log}(\hat{Y})$= 7.18+ 1.52smoking_status_yes + 0.03age +0.008BMI+0.102num_kids - 0.12district_colombo -0.096district_gall -1002.15distric_trinco

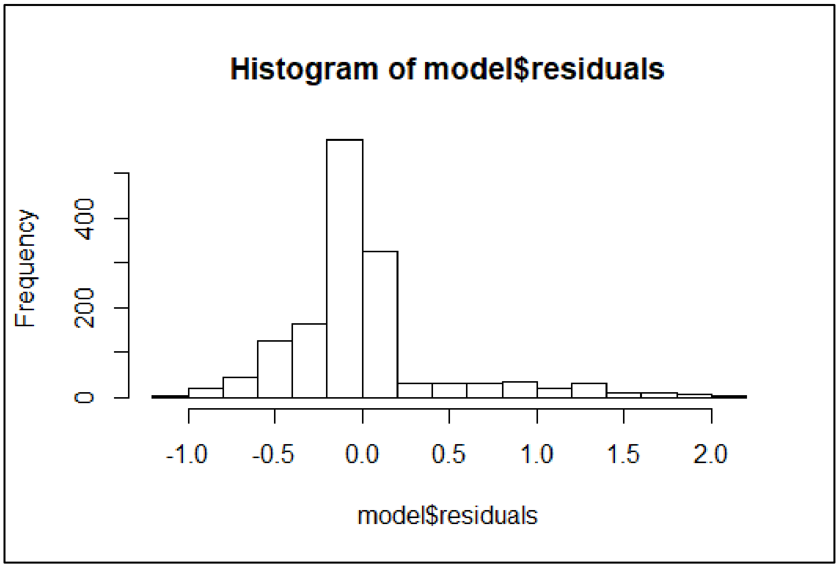


Figure: Distribution of residuals

Shape of the above distribution is slightly skewed to rigth

Residuals vs independent variables

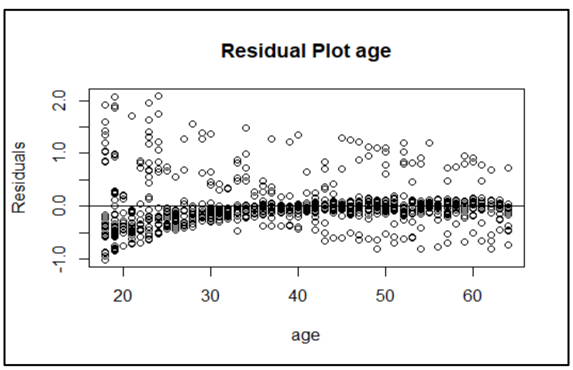


Figure: Residual plot age

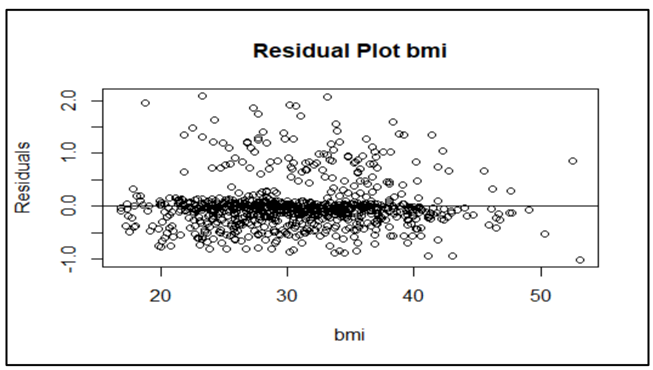
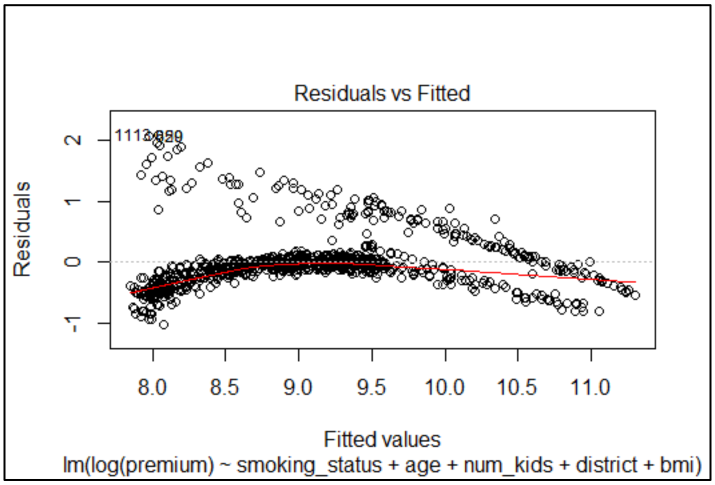
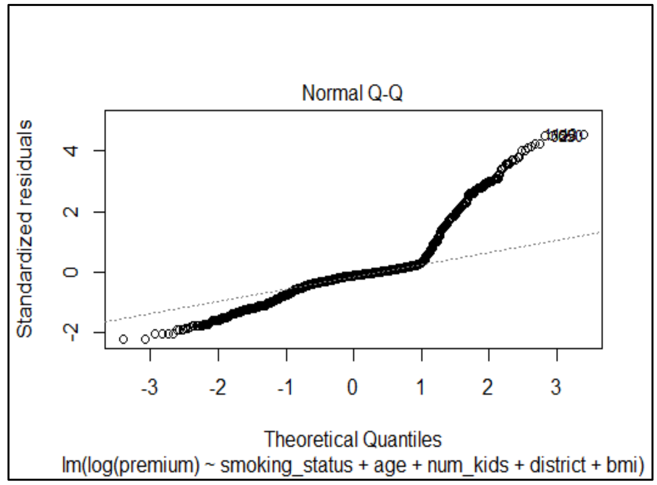
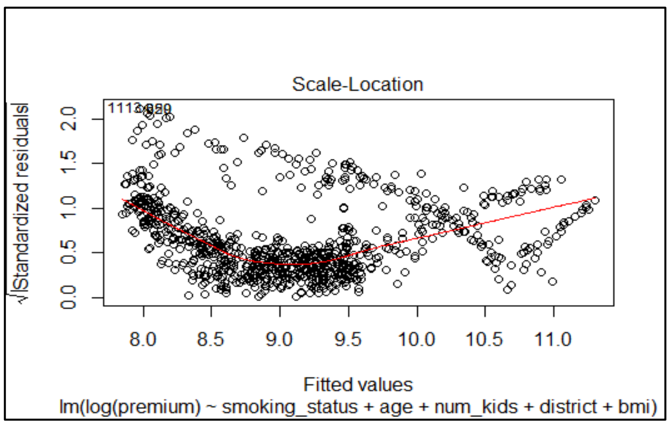


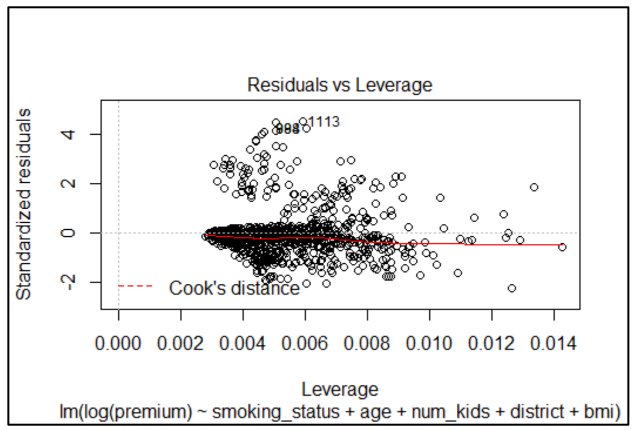
Figure: Residual plot BMI

As per above residual plots, it can be seen that residuals plot BMI is improve than the previous model in part b.









Q-Q plot has improve than model in part b.

```
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.831, p-value < 2.2e-16
```

As per above test results, w is significant , so there is evidencr to say that residual in not normal.

Transforming and recoding with Log(premium) , sqrt(age), BMI_ranges

```
> summary(forward.raw.model1)

Call:
lm(formula = log(premium) ~ smoking_status + sqrt(age) + as.factor(kids_cover) +
    district + as.factor(bmi_ranges), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83958 -0.23181 -0.06894  0.08085  2.11373

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.02109    0.11909  50.561  < 2e-16 ***
smoking_statusyes       1.53118    0.03076  49.776  < 2e-16 ***
sqrt(age)               0.41976    0.01050  39.972  < 2e-16 ***
as.factor(kids_cover)1  0.20035    0.02454   8.165 6.85e-16 ***
districtcolombo        -0.12783    0.03513  -3.639 0.000283 ***
districtgalle          -0.09959    0.03453  -2.884 0.003984 **
districttrinco         -0.15053    0.03487  -4.317 1.69e-05 ***
as.factor(bmi_ranges)2  0.08922    0.10379   0.860 0.390158
as.factor(bmi_ranges)3  0.13056    0.10194   1.281 0.200511
as.factor(bmi_ranges)4  0.21944    0.10120   2.168 0.030287 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4637 on 1461 degrees of freedom
Multiple R-squared:  0.7476,	Adjusted R-squared:  0.7461
F-statistic: 480.9 on 9 and 1461 DF,  p-value: < 2.2e-16

>
```

Log(Ŷ)= 6.02+ 1.53smoking_status_yes + 0.4(sqrt(age)) +0.2kids_coverYes -

0.12district_colombo -0.096district_gall -

15distric_trinco++0.08BMI_normal+0.13BMI_over+0.2BMI_obese

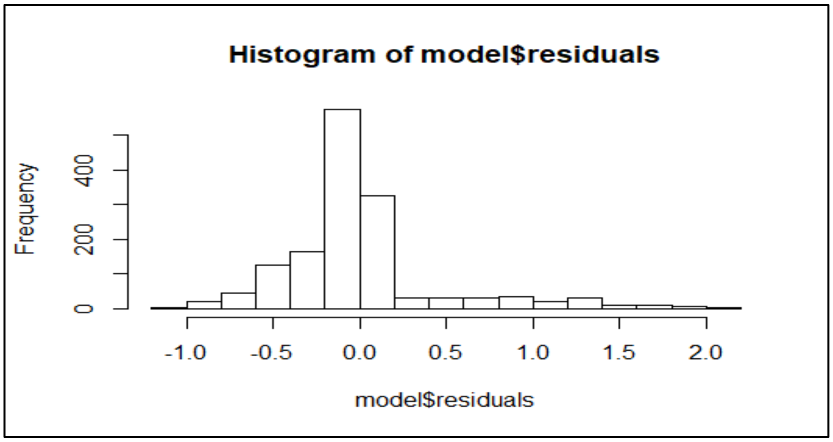
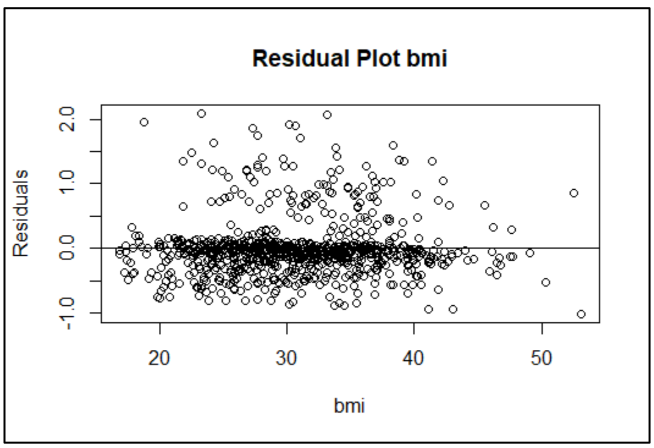
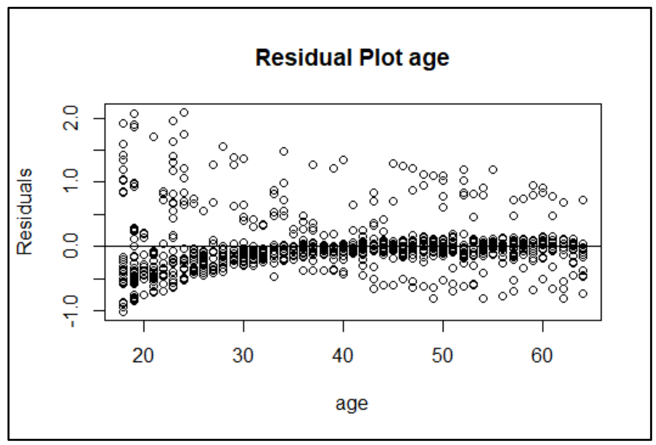


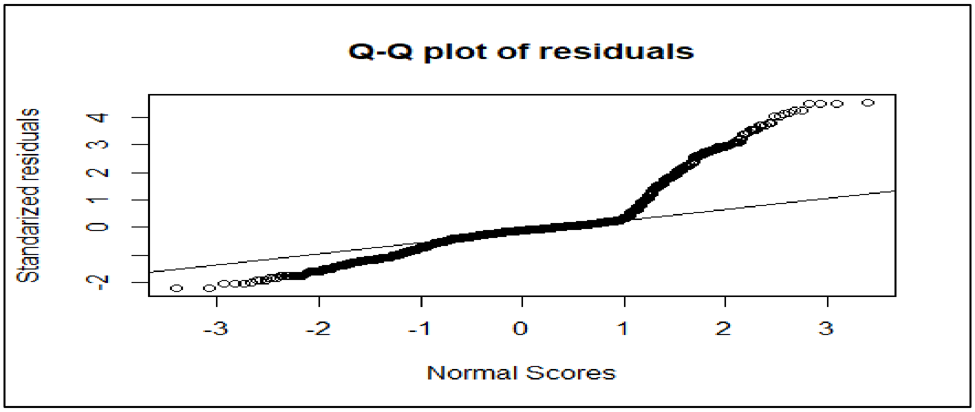
Figure: residual plot

Shape of the residual distribution has no considerable different with previous models





Distribution of residual plots of independent variables are same as previous figures. Scattered around zero and no any considerable liberality.



Q-Q plot is also same as previous figures.

```
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.831, p-value < 2.2e-16
```

After recoding and log transformation, the shapiro test is significant, so ther is evidence to say that residuals are not normally distributed.

e.

Split the insurance data as 80% training data & 20% test data

Build the below model using training dataset,

```
> summary(forwardmodel)

Call:
lm(formula = log(premium) ~ smoking_status + sqrt(age) + as.factor(kids_cover) + 
    district + as.factor(bmi_ranges), data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.80080 -0.23563 -0.08275  0.07635  2.14823 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.99164    0.13158  45.535  < 2e-16 ***
smoking_statusyes       1.49766    0.03502  42.769  < 2e-16 ***
sqrt(age)               0.42717    0.01179  36.235  < 2e-16 ***
as.factor(kids_cover)1  0.22925    0.02757   8.316 2.51e-16 ***
districtcolombo        -0.13282    0.03974  -3.342 0.000858 ***
districtgalle          -0.10515    0.03869  -2.718 0.006674 ** 
districttrinco         -0.14967    0.03901  -3.837 0.000131 ***
as.factor(bmi_ranges)2  0.05967    0.11220   0.532 0.594938    
as.factor(bmi_ranges)3  0.12287    0.11064   1.111 0.266990    
as.factor(bmi_ranges)4  0.17780    0.10968   1.621 0.105293    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4664 on 1166 degrees of freedom
Multiple R-squared:  0.7445,	Adjusted R-squared:  0.7426 
F-statistic: 377.6 on 9 and 1166 DF,  p-value: < 2.2e-16

> 
```


```
> summary(forwardmodel)

Call:
lm(formula = log(premium) ~ smoking_status + sqrt(age) + as.factor(kids_cover) + 
    district + as.factor(bmi_ranges), data = trainingData)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.80080 -0.23563 -0.08275  0.07635  2.14823 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.99164    0.13158  45.535  < 2e-16 ***
smoking_statusyes       1.49766    0.03502  42.769  < 2e-16 ***
sqrt(age)               0.42717    0.01179  36.235  < 2e-16 ***
as.factor(kids_cover)1  0.22925    0.02757   8.316 2.51e-16 ***
districtcolombo        -0.13282    0.03974  -3.342 0.000858 ***
districtgalle          -0.10515    0.03869  -2.718 0.006674 ** 
districttrinco         -0.14967    0.03901  -3.837 0.000131 ***
as.factor(bmi_ranges)2  0.05967    0.11220   0.532 0.594938    
as.factor(bmi_ranges)3  0.12287    0.11064   1.111 0.266990    
as.factor(bmi_ranges)4  0.17780    0.10968   1.621 0.105293    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4664 on 1166 degrees of freedom
Multiple R-squared:  0.7445,	Adjusted R-squared:  0.7426 
F-statistic: 377.6 on 9 and 1166 DF,  p-value: < 2.2e-16

> 
```

Predict the log(premium) using build model. AS below create a data frame with actual (premium) in test data set, and predicted log(premium) and residuals

```
> head(premium_resid)
   X.testData.premium.  exp.predict_premium.
6             8252.284             7960.474
9            10065.413            10484.073
10           24227.337            10646.860
12            7650.774             9576.680
14           19214.706             5531.188
17            9617.662            11664.799
>
```

```
> RMSE1=sqrt(mean((testData$premium-exp(predict_premium))**2))
> RMSE1
[1] 8168.036
>
```