Insurance_project

Import dataset

```r
getwd()

## [1] "C:/Users/Gagan Dhakad/Downloads"

setwd("/Users/Gagan Dhakad/Downloads")
df <- read.csv("swedishmotorinsurance.csv")
head(df)

##   Kilometres Zone Bonus Make Insured Claims Payment
## 1          1    1     1    1  455.13    108  392491
## 2          1    1     1    2   69.17     19   46221
## 3          1    1     1    3   72.88     13   15694
## 4          1    1     1    4 1292.39    124  422201
## 5          1    1     1    5  191.01     40  119373
## 6          1    1     1    6  477.66     57  170913

str(df)

## 'data.frame':    2182 obs. of  7 variables:
##  $ Kilometres: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Zone      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Bonus     : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ Make      : int  1 2 3 4 5 6 7 8 9 1 ...
##  $ Insured   : num  455.1 69.2 72.9 1292.4 191 ...
##  $ Claims    : int  108 19 13 124 40 57 23 14 1704 45 ...
##  $ Payment   : int  392491 46221 15694 422201 119373 170913 56940 77487 68
05992 214011 ...
```

Notes:***Dataframe contain 7 variables that are numerical***

Question 1: ***The committee is interested to know each field of the data collected through Descriptive analysis to gain basic insights into the data set and to prepare For further analysis***

Answer: ***summary provide an information about variables values minimum, maximum, 1st,2nd(median), 3rd quartile and mean. we can see that payment and claims have minimum zero values but insured columns does not have minimum zero values that means the some observation where the car has been insured for a particuler amount of time. this result no payment or claims has been made for this car make, kilometres and zones***

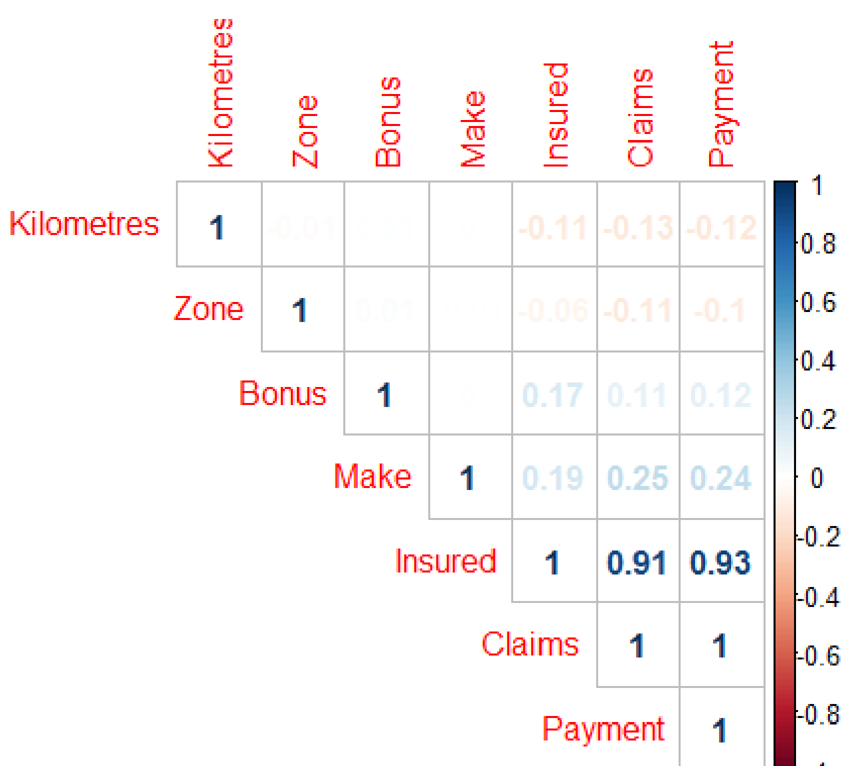
Descriptive Analysis of the dataset

```r
summary(df)
```

```
##    Kilometres          Zone           Bonus             Make      
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000  
##  1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:3.000  
##  Median :3.000   Median :4.00   Median :4.000   Median :5.000  
##  Mean   :2.986   Mean   :3.97   Mean   :4.015   Mean   :4.992  
##  3rd Qu.:4.000   3rd Qu.:6.00   3rd Qu.:6.000   3rd Qu.:7.000  
##  Max.   :5.000   Max.   :7.00   Max.   :7.000   Max.   :9.000  
##     Insured              Claims            Payment        
##  Min.   :     0.01   Min.   :   0.00   Min.   :       0  
##  1st Qu.:    21.61   1st Qu.:   1.00   1st Qu.:    2989  
##  Median :    81.53   Median :   5.00   Median :   27404  
##  Mean   :  1092.20   Mean   :  51.87   Mean   :  257008  
##  3rd Qu.:   389.78   3rd Qu.:  21.00   3rd Qu.:  111954  
##  Max.   :127687.27   Max.   :3338.00   Max.   :18245026
```

correlation matrix

```r
library(corrplot)

## corrplot 0.84 loaded

dff <- cor(df[, 1:7])

corrplot(dff, method = "number", "upper")
```



Notes: ***Correlation is used to test relationships between quantitative variables or categorical variables Here we see that the Insured-claims, Insured-Payment, Claims-Payment pairs are strongly corelated to each other they have higher correlation value***

type conversion of variables numerical into factor

```
mydata <- df

mydata$Kilometres <- as.factor(mydata$Kilometres)
mydata$Zone <- as.factor(mydata$Zone)
mydata$Bonus <- as.factor(mydata$Bonus)
mydata$Make <- as.factor(mydata$Make)


mydata$Kilometres <- factor(df$Kilometres, levels = c("1","2","3","4","5"))
levels(mydata$Kilometres) <- c("< 1000", "1000-15000","15000-20000", "20000-2
5000","> 25000")

str(mydata)
```

```
## 'data.frame':    2182 obs. of  7 variables:
##  $ Kilometres: Factor w/ 5 levels "< 1000","1000-15000",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ Zone      : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ Bonus     : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 2
...
##  $ Make      : Factor w/ 9 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 1
...
##  $ Insured   : num  455.1 69.2 72.9 1292.4 191 ...
##  $ Claims    : int  108 19 13 124 40 57 23 14 1704 45 ...
##  $ Payment   : int  392491 46221 15694 422201 119373 170913 56940 77487 68
05992 214011 ...
```

Confidence interval value

```
quantile(mydata$Insured, .90)
```

```
##      90%
## 1688.186
```

```
quantile(mydata$Claims, .90)
```

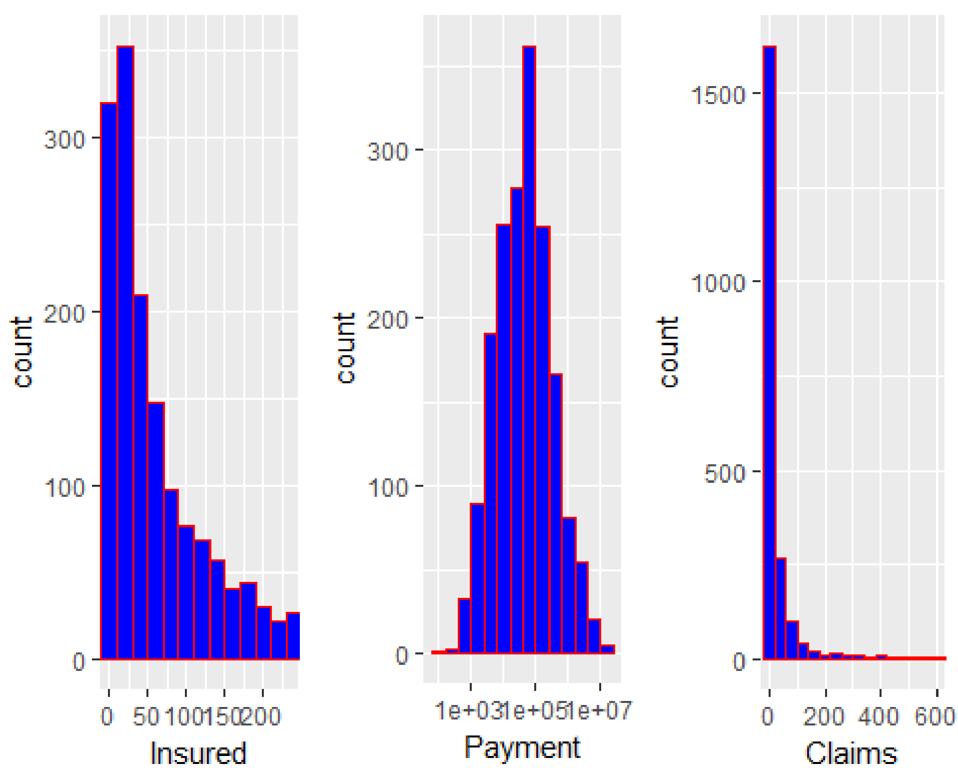
```
##  90%
## 87.9
```

```
quantile(mydata$Payment, .90)
```

```
##      90%
## 439777.7
```

Note: ***Here we calculate the value of 90% confidence interval of the variables and this value we apply to plotting the variable so that outlier does not effect the result of the plot***

```
library(ggplot2)
library(gridExtra)

h1 <- ggplot(data = mydata, aes(x = Insured)) +
      geom_histogram(binwidth=20, color = "red", fill = "blue") +
      coord_cartesian(xlim = c(0,235), expand = TRUE)

h2 <- ggplot(data = mydata, aes(x = Payment)) +
      geom_histogram(binwidth = 0.4, color = "red", fill = "blue")  +
      scale_x_log10()

h3 <- ggplot(data = mydata, aes(x = Claims)) +
      geom_histogram(binwidth = 40, color = "red", fill = "blue") +
      coord_cartesian(xlim = c(0,600), expand = TRUE)


grid.arrange(h1,h2,h3, ncol = 3)
```

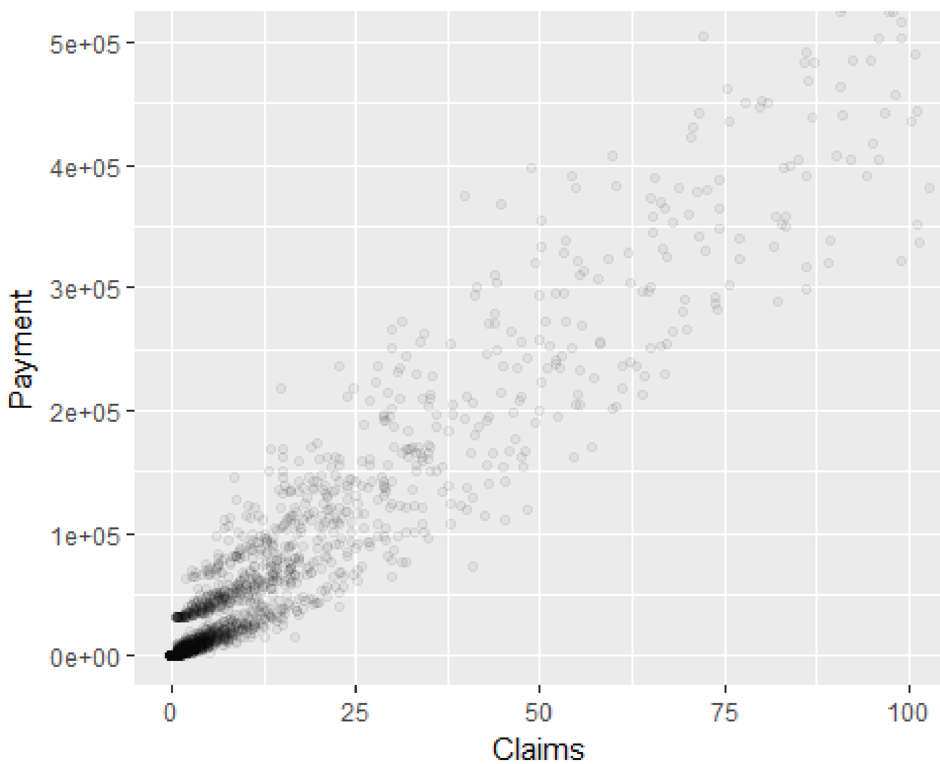


Notes: ***The histogram provide information that the most count of insured is below 200, and payment histogram we use log function for perfact distribution we see the distribution 1000 to 10000000, and Claims has higest numer of value below 100***

Question 2 :***The total value of payment by an insurance company is an important factor to be monitored. So the committee has decided to find whether this payment is related to number of claims and the number of insured policy years. They also want to visualize the results for better understanding***

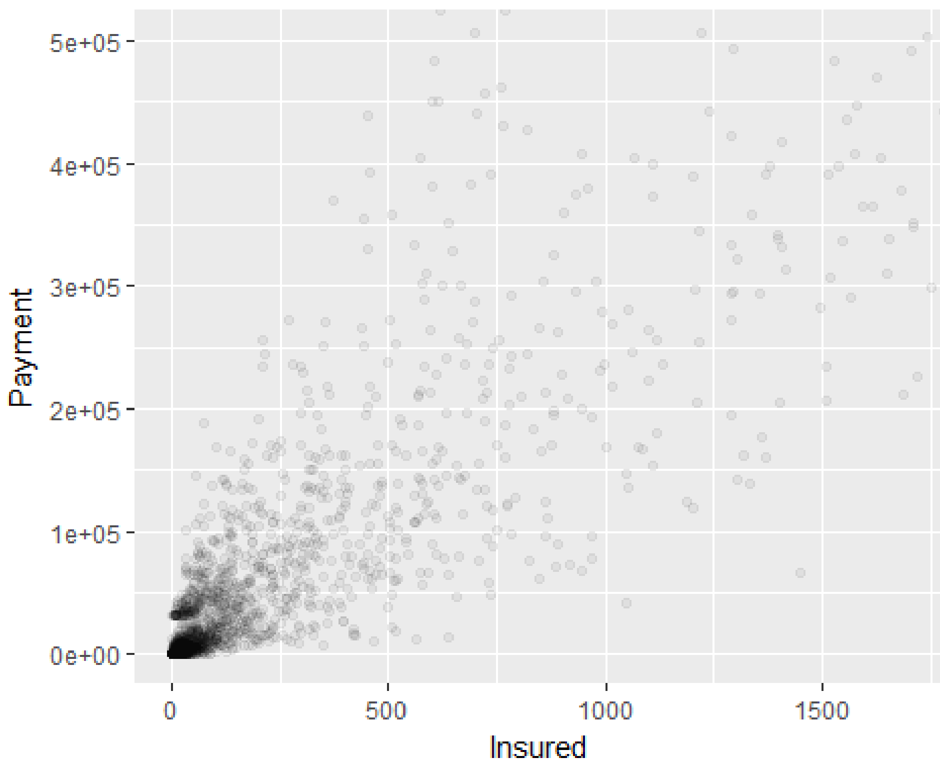
Answer:***The graph shows the strong positive relationship between variables, claims is approximate 100 % positve strongly corelated to payment that means if claims value is increase than value unit of payment is also linearly increase and the second graph shows that the value of Insured is 93% positive strongly corelated to the payment but we see that in second graph the value of variables is despressed in space means in a fitting line the variation is more compare to graph one***

```
ggplot(data = mydata, aes(x = Claims, y = Payment)) +
       geom_jitter( alpha = 0.05) +
       coord_cartesian(xlim = c(0,100), ylim = c(0,500000))
```



Notes: ***This give us a correlation of 0.91.This is a strong positive correlation with this plot. we can see that the bulk of the data lies below 1 lac and and claims are below 25 if Payment and Claims increase the value of both increase linearly but values are dispresed we have alpaha parameter for tansperancy so we set it alpha = 0.05 so the we clearly see the data points***

```
ggplot(data = mydata, aes(x = Insured, y = Payment)) +
       geom_jitter( alpha = 0.05) +
       coord_cartesian(xlim = c(0,1700), ylim = c(0,500000))
```



Notes: ***we see that the bulk of data is lie below the 5 lac and 300 Insured means most of the customers are in this tiny space***

Question 3 :***The committee wants to figure out the reasons for insurance payment increase and decrease. So they have decided to find whether distance, location, bonus, make, and insured amount or claims are affecting the payment or all or some of these are affecting it***

```
linear_model_1 <- lm(Payment ~ Insured+Claims+Make+Bonus+Zone+Kilometres, dat
a = mydata )

summary(linear_model_1)

##
## Call:
## lm(formula = Payment ~ Insured + Claims + Make + Bonus + Zone +
##     Kilometres, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -762236  -18278   -1588   16179  831273
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.346e+04  7.596e+03  -1.772 0.076604 .
## Insured                 2.809e+01  6.817e-01  41.206  < 2e-16 ***
## Claims                  4.289e+03  2.089e+01 205.288  < 2e-16 ***
```

```
## Make2                    -1.527e+04  6.306e+03  -2.422 0.015521 *
## Make3                    -1.283e+04  6.330e+03  -2.026 0.042851 *
## Make4                    -2.647e+04  6.359e+03  -4.162 3.28e-05 ***
## Make5                    -1.814e+04  6.312e+03  -2.873 0.004100 **
## Make6                    -1.863e+04  6.311e+03  -2.953 0.003185 **
## Make7                    -2.016e+04  6.329e+03  -3.186 0.001463 **
## Make8                    -1.005e+04  6.368e+03  -1.578 0.114664
## Make9                    -6.808e+03  7.004e+03  -0.972 0.331169
## Bonus2                    3.880e+03  5.626e+03   0.690 0.490513
## Bonus3                    4.371e+03  5.653e+03   0.773 0.439561
## Bonus4                    1.063e+03  5.663e+03   0.188 0.851057
## Bonus5                   -1.354e+03  5.648e+03  -0.240 0.810524
## Bonus6                    3.863e+03  5.624e+03   0.687 0.492220
## Bonus7                    1.536e+04  5.751e+03   2.670 0.007638 **
## Zone2                     1.402e+03  5.557e+03   0.252 0.800802
## Zone3                     3.908e+03  5.568e+03   0.702 0.482789
## Zone4                     3.314e+04  5.591e+03   5.927 3.58e-09 ***
## Zone5                     6.512e+03  5.614e+03   1.160 0.246247
## Zone6                     1.936e+04  5.598e+03   3.458 0.000555 ***
## Zone7                     4.971e+03  5.740e+03   0.866 0.386572
## Kilometres1000-15000      2.236e+04  4.707e+03   4.751 2.16e-06 ***
## Kilometres15000-20000     2.329e+04  4.703e+03   4.952 7.92e-07 ***
## Kilometres20000-25000     2.161e+04  4.746e+03   4.553 5.59e-06 ***
## Kilometres> 25000         2.150e+04  4.770e+03   4.507 6.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69670 on 2155 degrees of freedom
## Multiple R-squared:  0.9954, Adjusted R-squared:  0.9953
## F-statistic: 1.78e+04 on 26 and 2155 DF,  p-value: < 2.2e-16
```

Answer 3: ***The r square value is 99.5 % which tells that the variation of Payment based on Insured, Claims, Make, Bonus, Zone and Kilometres are very strong t-value for Insured, claims and Bonus are greater than 1.96 means that the variable are significant at 95% confidence level The p-value for the Insured, claims,Kilometres variable are less than 0.05 and hence the variable are found to be significant***

```
linear_model_2 <- lm(Payment ~ Insured+Claims + Kilometres, data =mydata)

summary(linear_model_2)
```

```
##
## Call:
## lm(formula = Payment ~ Insured + Claims + Kilometres, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -781850  -15220   -7769   14068  853620
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            -1.513e+04  3.425e+03  -4.417 1.05e-05 ***
## Insured                 2.850e+01  6.476e-01  44.011  < 2e-16 ***
## Claims                  4.295e+03  1.823e+01 235.537  < 2e-16 ***
## Kilometres1000-15000    2.226e+04  4.777e+03   4.660 3.35e-06 ***
## Kilometres15000-20000   2.371e+04  4.774e+03   4.965 7.40e-07 ***
## Kilometres20000-25000   2.267e+04  4.807e+03   4.716 2.56e-06 ***
## Kilometres> 25000       2.283e+04  4.829e+03   4.729 2.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 70740 on 2175 degrees of freedom
## Multiple R-squared:  0.9952, Adjusted R-squared:  0.9952 
## F-statistic: 7.48e+04 on 6 and 2175 DF,  p-value: < 2.2e-16
```

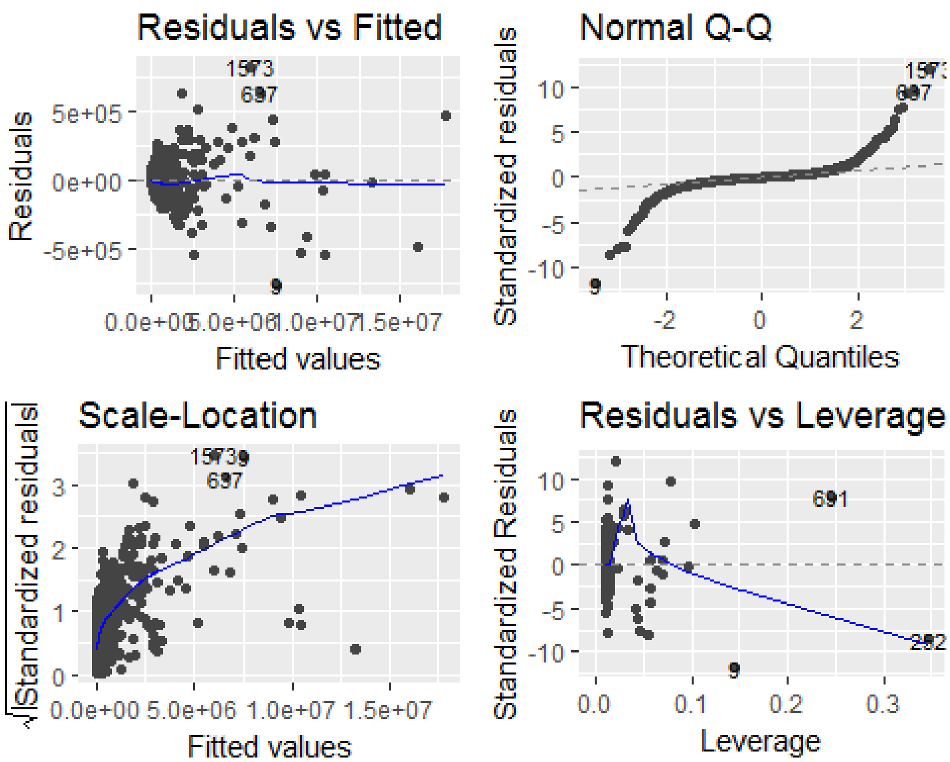
Answer 3: ***In this model we can say the value of R square is very tiny little bit reduce like 99.54 to 99.52 % that not affect our model means other variable except Insured, Kilometres and claims do not have much contribution The only thing is increase is F-static value That is means our model is perfectly competible with the data***

```
anova(linear_model_1, linear_model_2)
```

```
## Analysis of Variance Table
## 
## Model 1: Payment ~ Insured + Claims + Make + Bonus + Zone + Kilometres
## Model 2: Payment ~ Insured + Claims + Kilometres
##   Res.Df        RSS  Df   Sum of Sq      F    Pr(>F)    
## 1   2155 1.0461e+13                                      
## 2   2175 1.0885e+13 -20 -4.2404e+11 4.3675 3.703e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notes: ***with this comparesion we can see that Insured and claims contribute more significant information than other variable***

```
library(ggfortify)

autoplot(linear_model_1, label.size = 3)
```



Notes: ***In residual vs fitted plot shows if residuals have non-linear patterns we see that the residual are not perfactly spread around the horizonatal line, this plot is not have the fully non-linear relationship Q-Q plot shows if residuals are normally distributed,It's good if residuals are lined well on the straight dashed line.but in this plot the line would not be perfact scale-location provide spread location plot, the residuals begin to spread wider along the axis as it passes***

Question 4: ***The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilometer, and bonus level their insured amount, claims, and payment get increased***

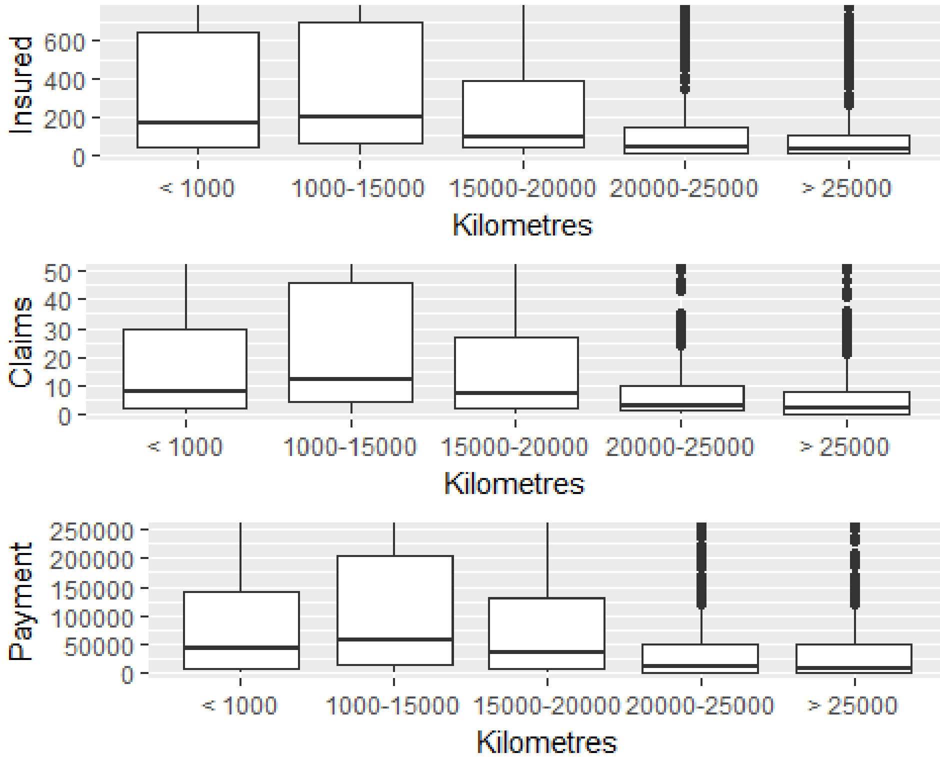
```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

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

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

```
kilometres_groups <- group_by(mydata, Kilometres)

fc_by_kilometres_groups <- summarise(kilometres_groups, mean_Insured = mean(I
nsured),
                                     mean_Claims = mean(Claims),
                                     mean_Payment = mean(Payment))

fc_by_kilometres_groups

## # A tibble: 5 x 4
##   Kilometres  mean_Insured mean_Claims mean_Payment
##   <fct>              <dbl>       <dbl>        <dbl>
## 1 < 1000             1838.        75.6      361899.
## 2 1000-15000         1824.        89.3      442524.
## 3 15000-20000        1082.        54.2      272013.
## 4 20000-25000         399.        20.8      108213.
## 5 > 25000             285.        18.0       93306.
```

Answer 4_1:***1000-15000 kilometers group has the maximum number of claims and payment but the insured number of years is lesser than the < 1000 kilometres group we can see that if the kilometres increase than the claims and number of insured in policy-years is decrease***

This thing we can also see using box plot they clearly show the result

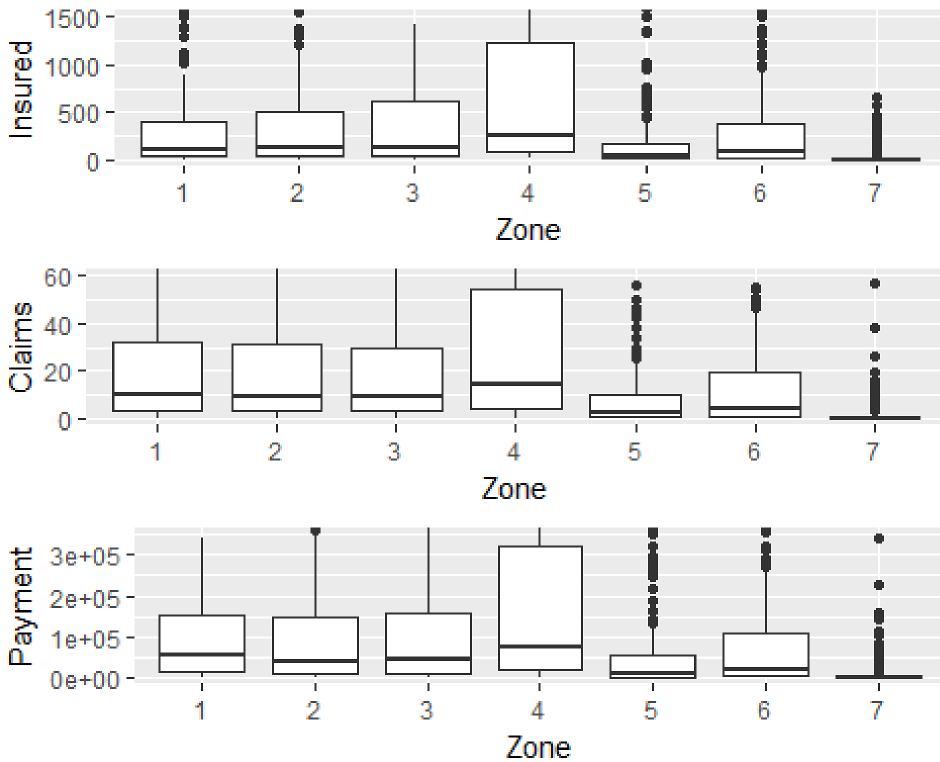
```
k1 <- ggplot(data = mydata, aes(x = Kilometres, y = Insured)) +
      geom_boxplot() + coord_cartesian(ylim = c(0,750))

k2<- ggplot(data = mydata, aes(x = Kilometres, y = Claims)) +
     geom_boxplot() + coord_cartesian(ylim = c(0,50))

k3 <- ggplot(data = mydata, aes(x = Kilometres, y = Payment)) +
      geom_boxplot() + coord_cartesian(ylim = c  (0,250000))


grid.arrange(k1,k2,k3, nrow =3)
```



Answer 4_2: ***we see that zone 4 has the highest number of claims and payment zone,that means the zone is lies 1000-20000 kilometers area because they also have highest number of claims and payment,and 1 to 4 have more insured years, claims and payments***

```
zone_groups <- group_by(mydata, Zone)

fc_by_zone_groups <- summarise(zone_groups, mean_Insured = mean(Insured),
                                  mean_Claims = mean(Claims),
                                  mean_Payment = mean(Payment))

fc_by_zone_groups

## # A tibble: 7 x 4
##   Zone  mean_Insured mean_Claims mean_Payment
##   <fct>        <dbl>       <dbl>        <dbl>
## 1 1           1036.        73.6       338519.
## 2 2           1231.        67.6       319922.
## 3 3           1363.        63.3       307551.
## 4 4           2689.       101.        537072.
## 5 5            385.        19.0        93002.
## 6 6            803.        32.6       175528.
## 7 7             64.9        2.11        9948.

z1 <- ggplot(data = mydata, aes(x = Zone, y = Insured)) +geom_boxplot() +
      coord_cartesian(ylim = c(0,1500))
```

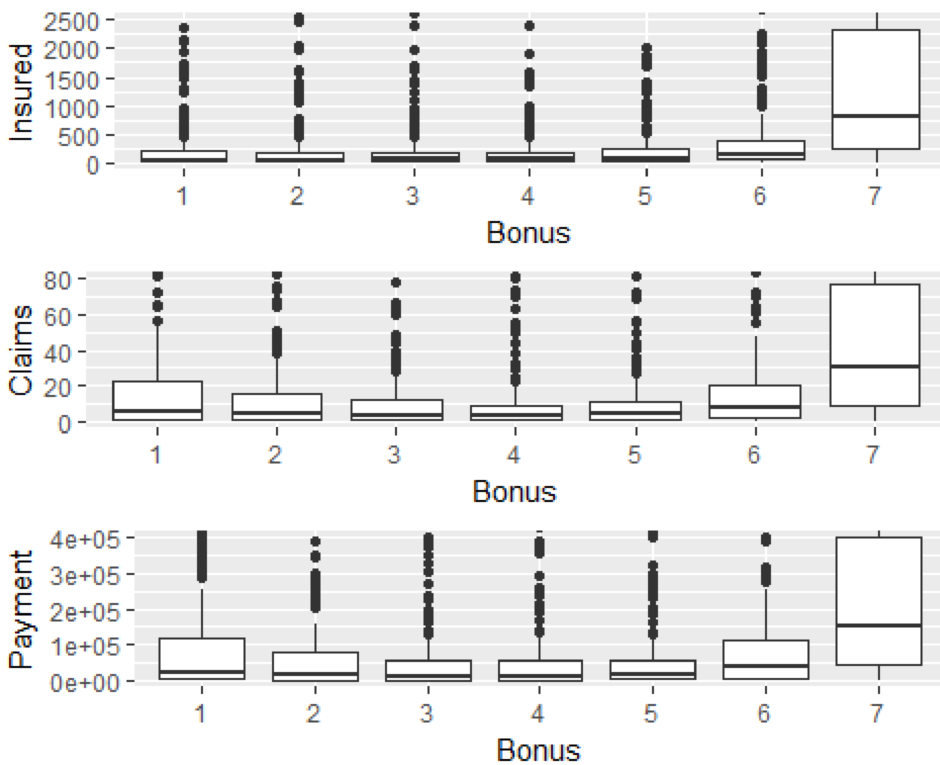
```
z2 <- ggplot(data = mydata, aes(x = Zone, y = Claims)) +geom_boxplot() +
      coord_cartesian(ylim = c(0,60))

z3 <- ggplot(data = mydata, aes(x = Zone, y = Payment)) +geom_boxplot() +
      coord_cartesian(ylim = c(0,350000))

grid.arrange(z1,z2,z3, nrow = 3)
```



Answer 4_3: ***The Bonus group-7 has the hghest number of claims, insured policy-years and payment as well, and all other groups not have much variation in a bonus aproximate pretty same***

```
bonus_groups <- group_by(mydata, Bonus)

fc_by_bonus_groups <- summarise(bonus_groups, mean_Insured = mean(Insured),
                                mean_Claims = mean(Claims),
                                mean_Payment = mean(Payment))


fc_by_bonus_groups

## # A tibble: 7 x 4
##   Bonus mean_Insured mean_Claims mean_Payment
##   <fct>        <dbl>       <dbl>        <dbl>
## 1 1             526.        62.5      282922.
## 2 2             451.        34.2      163317.
## 3 3             397.        25.0      122656.
## 4 4             360.        20.4       98498.
```

```
## 5 5              437.        22.8       108791.
## 6 6              806.        39.9       197724.
## 7 7             4620.       157.        819322.

b1 <- ggplot(data = mydata, aes(x = Bonus, y = Insured)) +
      geom_boxplot() + coord_cartesian(ylim = c(0,2500))

b2 <- ggplot(data = mydata, aes(x = Bonus, y = Claims)) +
      geom_boxplot() + coord_cartesian(ylim = c(0,80))

b3 <- ggplot(data = mydata, aes(x = Bonus, y = Payment)) +
      geom_boxplot() + coord_cartesian(ylim = c(0,400000))


grid.arrange(b1,b2,b3, nrow = 3)
```



Question 5 : ***The committee wants to understand what affects their claim rates so as to decide the right premiums for a certain set of situations. Hence, they need to find whether the insured amount, zone, kilometer, bonus, or make affects the claim rates and to what extent***

Answer:***we see in model_1 that the The r square value is 87 % which tells that the variation of response variable claims based on explantary variables Insured, kilometers, Make, Bonus, Zone and Kilometres are very strong The p-value for the Insured, Bonus, make variable are less than 0.05 and hence the variable are found to be significant variable kilometres have a p-value less than 0.05 except the 15000-***

20000 group A negative t-value simply indicates a reversal in the directionality of the effect, which has no bearing on the significance of the difference between groups we see in model_2 if we left some variables that has week correlation with claims, the result we get from model_1, the r square value is 86% which is pretty good means varible have impact on claims but not high we see in model_2 The p-value for the Insured, Kilometres and make variable are less than 0.05 and hence the variable are found to be significant

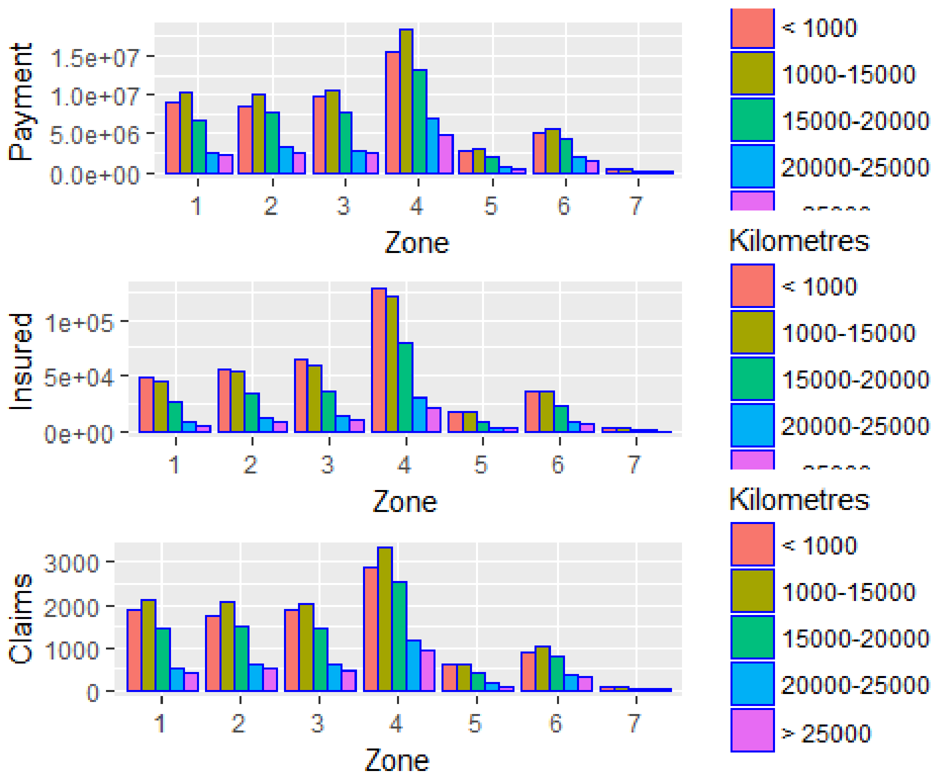
```
cor(df[,1:5], df$Claims)

##                  [,1]
## Kilometres -0.1284519
## Zone       -0.1146872
## Bonus       0.1051024
## Make        0.2532120
## Insured     0.9103478

mul_model_1 <- lm(Claims ~ Kilometres + Zone + Bonus + Make + Insured, data =
mydata)

summary(mul_model_1)

##
## Call:
## lm(formula = Claims ~ Kilometres + Zone + Bonus + Make + Insured,
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -983.95  -16.36    0.06   14.09 1222.44
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.130e+01  7.679e+00   9.284  < 2e-16 ***
## Kilometres1000-15000  1.423e+01  4.843e+00   2.938 0.003341 **
## Kilometres15000-20000 8.060e-01  4.848e+00   0.166 0.867982
## Kilometres20000-25000 -1.317e+01 4.884e+00  -2.697 0.007057 **
## Kilometres> 25000    -1.309e+01  4.910e+00  -2.666 0.007737 **
## Zone2                -1.165e+01  5.724e+00  -2.036 0.041887 *
## Zone3                -1.983e+01  5.724e+00  -3.464 0.000543 ***
## Zone4                -2.059e+01  5.747e+00  -3.583 0.000347 ***
## Zone5                -3.574e+01  5.737e+00  -6.230 5.60e-10 ***
## Zone6                -3.416e+01  5.724e+00  -5.969 2.79e-09 ***
## Zone7                -4.461e+01  5.839e+00  -7.641 3.23e-14 ***
## Bonus2               -2.533e+01  5.775e+00  -4.385 1.21e-05 ***
## Bonus3               -3.334e+01  5.784e+00  -5.765 9.35e-09 ***
## Bonus4               -3.679e+01  5.784e+00  -6.361 2.44e-10 ***
## Bonus5               -3.614e+01  5.771e+00  -6.263 4.55e-10 ***
## Bonus6               -2.950e+01  5.763e+00  -5.119 3.35e-07 ***
## Bonus7               -2.374e+01  5.907e+00  -4.019 6.03e-05 ***
```

```
## Make2                 -1.375e+01  6.494e+00  -2.117 0.034346 *
## Make3                 -1.727e+01  6.515e+00  -2.651 0.008088 **
## Make4                 -1.911e+01  6.543e+00  -2.921 0.003523 **
## Make5                 -1.278e+01  6.501e+00  -1.966 0.049478 *
## Make6                 -1.514e+01  6.498e+00  -2.330 0.019899 *
## Make7                 -1.611e+01  6.515e+00  -2.473 0.013469 *
## Make8                 -1.813e+01  6.553e+00  -2.767 0.005712 **
## Make9                  1.180e+02  6.759e+00  17.451  < 2e-16 ***
## Insured                2.924e-02  3.122e-04  93.649  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.83 on 2156 degrees of freedom
## Multiple R-squared:  0.8746, Adjusted R-squared:  0.8732
## F-statistic: 601.7 on 25 and 2156 DF,  p-value: < 2.2e-16
```

```
mul_model_2 <- lm(Claims ~   Make + Kilometres + Insured, data = mydata)

summary(mul_model_2)
```

```
##
## Call:
## lm(formula = Claims ~ Make + Kilometres + Insured, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1012.04    -9.14    -1.59     8.79  1272.12
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.049e+01  5.709e+00   3.590 0.000338 ***
## Make2                 -1.353e+01  6.687e+00  -2.024 0.043122 *
## Make3                 -1.696e+01  6.708e+00  -2.529 0.011508 *
## Make4                 -1.845e+01  6.736e+00  -2.739 0.006220 **
## Make5                 -1.243e+01  6.694e+00  -1.857 0.063389 .
## Make6                 -1.503e+01  6.691e+00  -2.247 0.024757 *
## Make7                 -1.567e+01  6.708e+00  -2.336 0.019577 *
## Make8                 -1.725e+01  6.745e+00  -2.558 0.010598 *
## Make9                  1.162e+02  6.932e+00  16.769  < 2e-16 ***
## Kilometres1000-15000   1.417e+01  4.987e+00   2.842 0.004529 **
## Kilometres15000-20000  9.604e-01  4.992e+00   0.192 0.847459
## Kilometres20000-25000 -1.253e+01  5.026e+00  -2.493 0.012727 *
## Kilometres> 25000     -1.221e+01  5.050e+00  -2.418 0.015673 *
## Insured                2.952e-02  3.044e-04  96.981  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.96 on 2168 degrees of freedom
## Multiple R-squared:  0.8663, Adjusted R-squared:  0.8655
## F-statistic:  1081 on 13 and 2168 DF,  p-value: < 2.2e-16
```

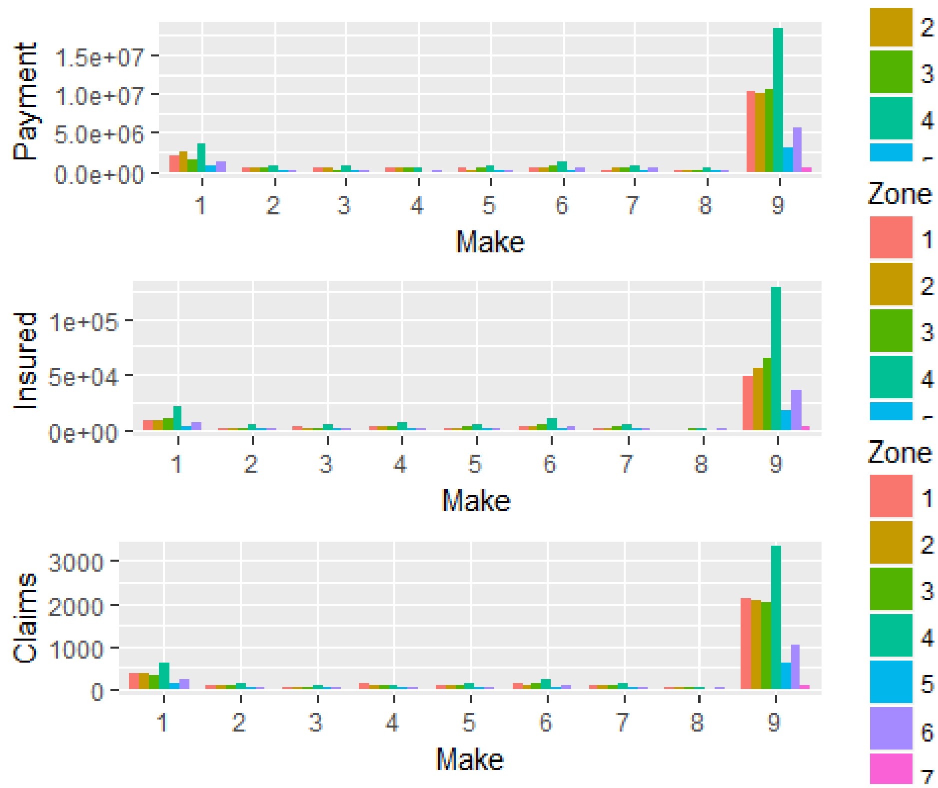
```
g1 <- ggplot(data = mydata, aes(x = Zone, y = Payment )) +
      geom_bar(stat = "identity", aes(fill = Kilometres), position = "dodge",
col = "blue")

g2 <- ggplot(data = mydata, aes(x = Zone, y = Insured )) +
      geom_bar(stat = "identity", aes(fill = Kilometres), position = "dodge",
col = "blue")

g3 <- ggplot(data = mydata, aes(x = Zone, y = Claims )) +
      geom_bar(stat = "identity", aes(fill = Kilometres), position = "dodge",
col = "blue")

grid.arrange(g1,g2,g3, nrow =3)
```



Notes: ***Using This bar graph we can say that zone-4(1000-15000) has highest number of Claims, Insured and payment and zone-1,2,3 have aprroximate same result not much variation in claims, payment and Insured***

```
m1 <- ggplot(data = mydata, aes(x = Make, y = Payment )) +
      geom_bar(stat = "identity", aes(fill = Zone), position = "dodge")

m2 <- ggplot(data = mydata, aes(x = Make, y = Insured )) +
      geom_bar(stat = "identity", aes(fill = Zone), position = "dodge" )

m3 <- ggplot(data = mydata, aes(x = Make, y = Claims )) +
      geom_bar(stat = "identity", aes(fill = Zone), position = "dodge")
```

```
grid.arrange(m1,m2,m3, nrow = 3)
```



Notes: ***The result we get in this graph is car model-1 is mostly use in zones except other model-9, and other model-9 car is mostly used in zones they have highest number of Claims, Insured and payment compare to other***