Data Science with R

Project: Insurance factors identification

Submitted by: Sagar Samaria

Oct-21-2020

- **Project Description**

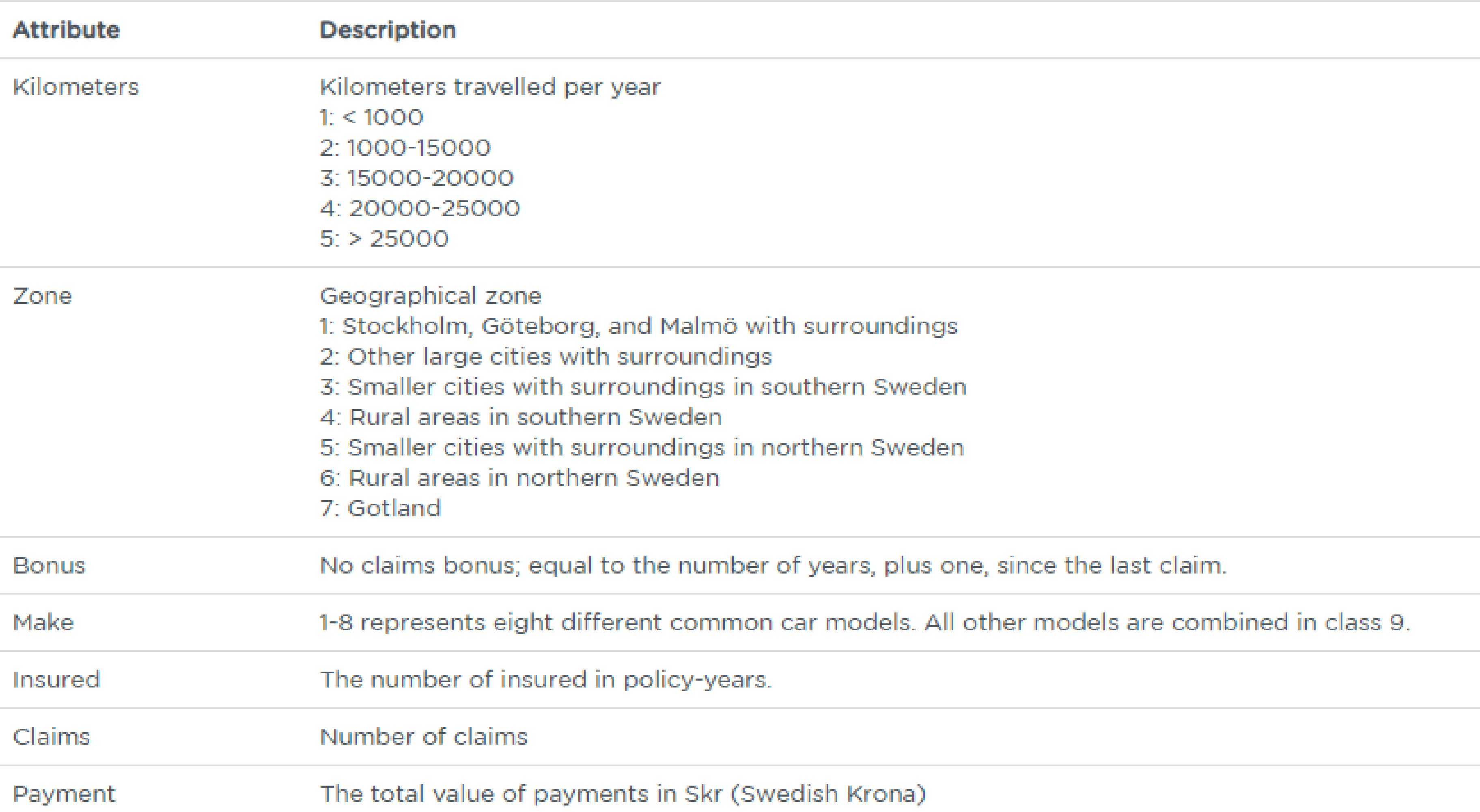
Data is provided for a third party motor insurance claims in Sweden for the year 1977. The insurance companies apply identical risk arguments to classify customers, and then combines the portfolios and claims statistics.

- **Scope and Objective**

To analyze the real influence on the claims of the risk arguments and to compare this structure with the actual tariff.

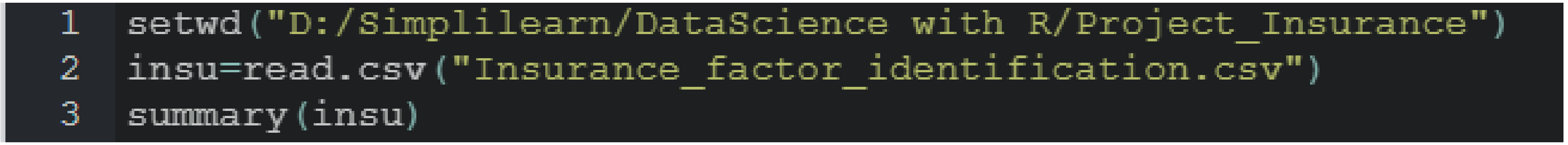
- **Variables provides**

The insurance dataset holds 7 variables and the description of these variables are given below:

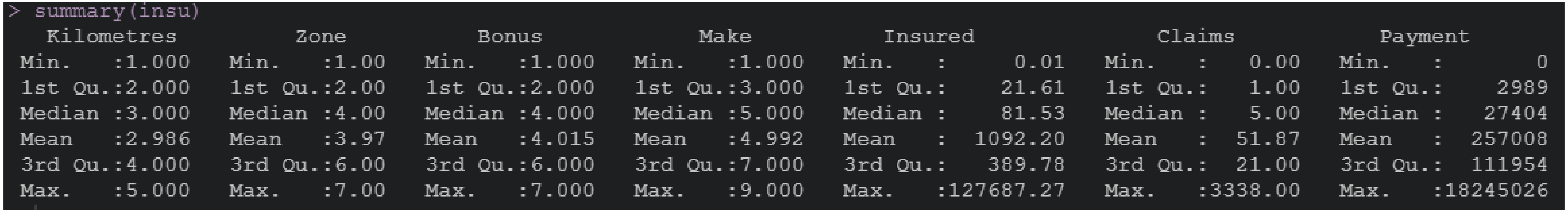


Attribute	Description
Kilometers	Kilometers travelled per year 1: < 1000 2: 1000-15000 3: 15000-20000 4: 20000-25000 5: > 25000
Zone	Geographical zone 1: Stockholm, Göteborg, and Malmö with surroundings 2: Other large cities with surroundings 3: Smaller cities with surroundings in southern Sweden 4: Rural areas in southern Sweden 5: Smaller cities with surroundings in northern Sweden 6: Rural areas in northern Sweden 7: Gotland
Bonus	No claims bonus; equal to the number of years, plus one, since the last claim.
Make	1-8 represents eight different common car models. All other models are combined in class 9.
Insured	The number of insured in policy-years.
Claims	Number of claims
Payment	The total value of payments in Skr (Swedish Krona)

Q-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.



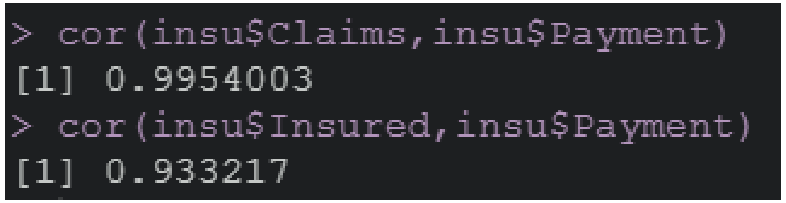
```r
setwd("D:/Simplilearn/DataScience with R/Project_Insurance")
insu=read.csv("Insurance_factor_identification.csv")
summary(insu)
```



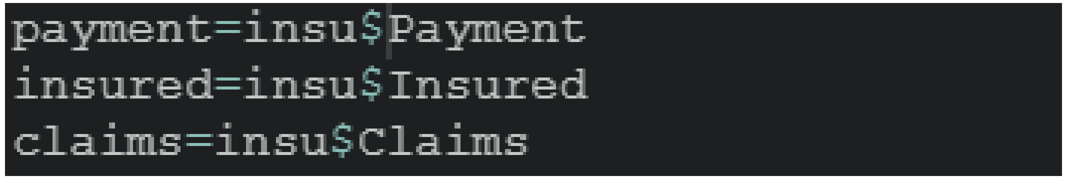
```
> summary(insu)
   Kilometres          Zone            Bonus             Make          Insured              Claims            Payment        
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :     0.01   Min.   :   0.00   Min.   :       0  
 1st Qu.:2.000   1st Qu.:2.00   1st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

- A list variable insu is used to read the database Insurance_factor_identification.csv and the summary command summarizes the variable providing insights about the database.
- There are some null values in the column Claims and Payment which means that no Claim or Payment has been made for that datapoint(s).

Q-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 the number of claims and the number of insured policy years. They also want to visualize the results for better understanding.

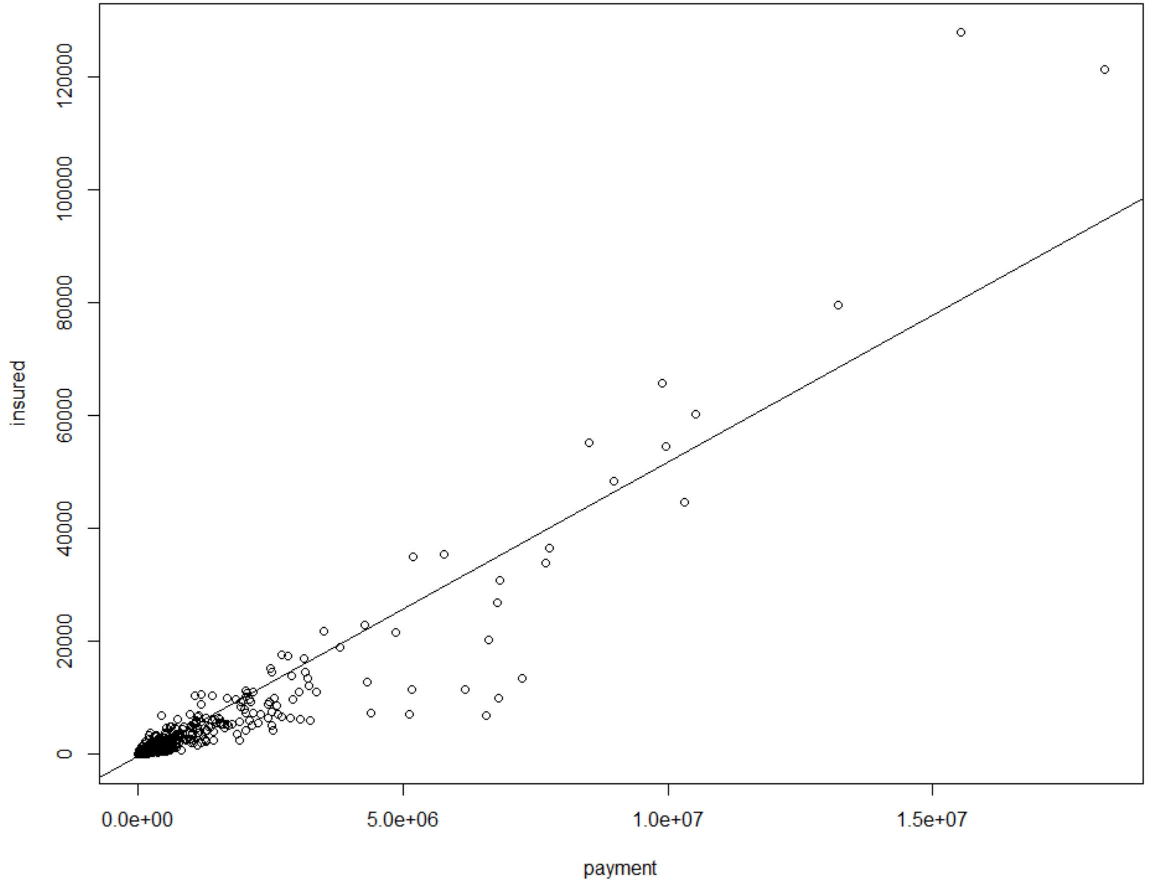


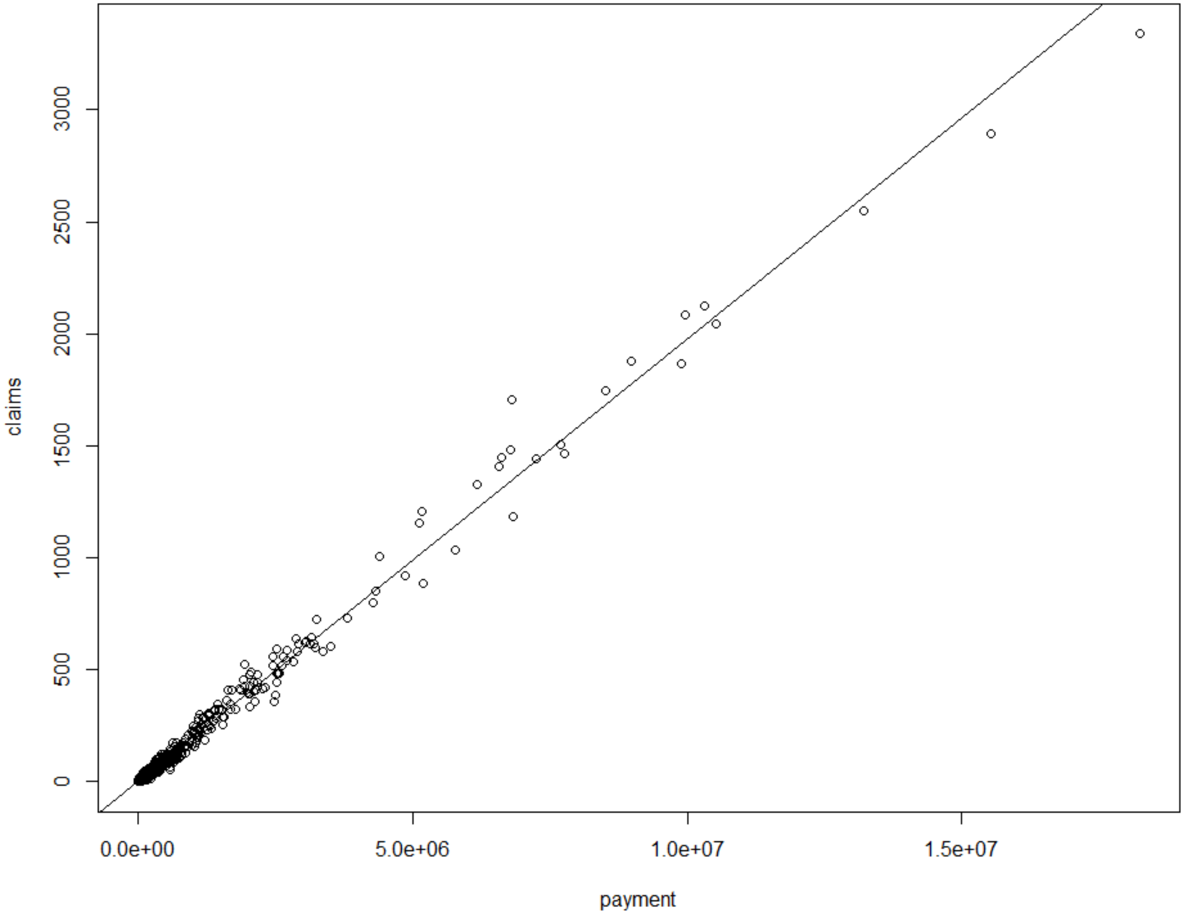
```
> cor(insu$Claims,insu$Payment)
[1] 0.9954003
> cor(insu$Insured,insu$Payment)
[1] 0.933217
```



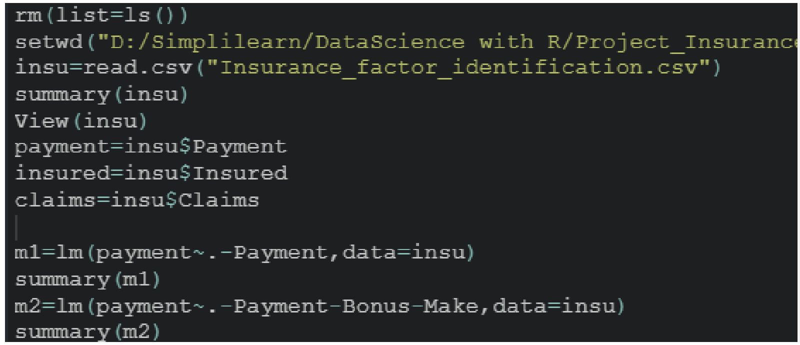
```
payment=insu$Payment
insured=insu$Insured
claims=insu$Claims
```

- Positive and very correlation is seen between payment and Claims, and between payment (99.5%)

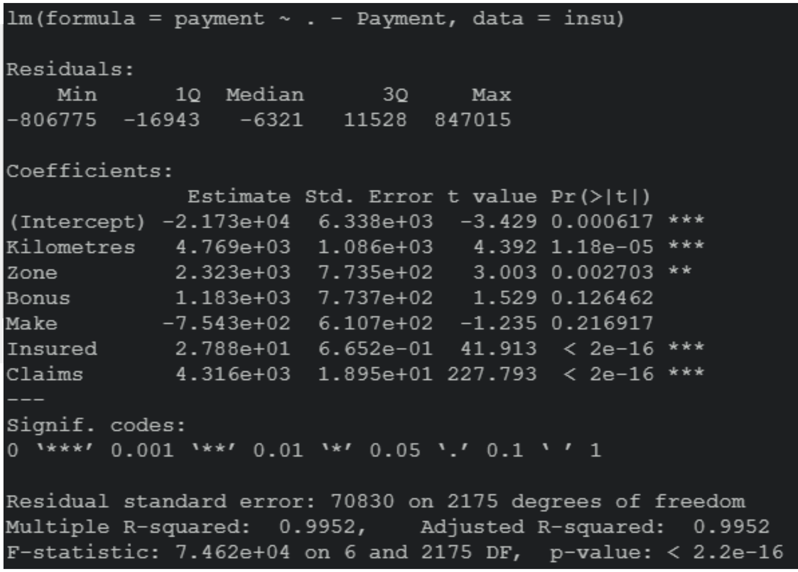




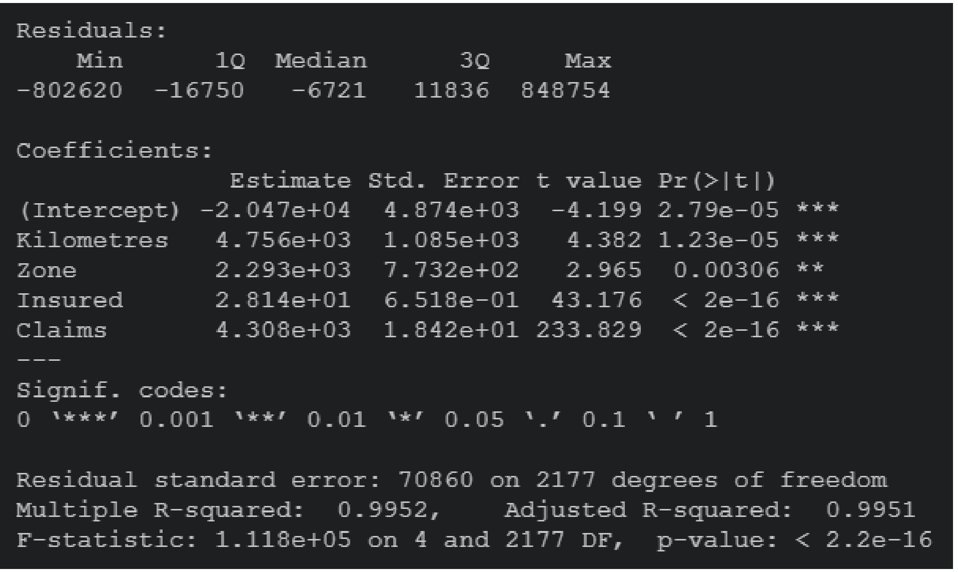
Q3. The committee wants to figure out the reasons for insurance payment increase and decrease. So they have decided to find whether distance, location, bonus, make, and insured amount or claims are affecting the payment or all or some of these are affecting it.



```
rm(list=ls())
setwd("D:/Simplilearn/DataScience with R/Project_Insuranc
insu=read.csv("Insurance_factor_identification.csv")
summary(insu)
View(insu)
payment=insu$Payment
insured=insu$Insured
claims=insu$Claims

m1=lm(payment~.-Payment,data=insu)
summary(m1)
m2=lm(payment~.-Payment-Bonus-Make,data=insu)
summary(m2)
```



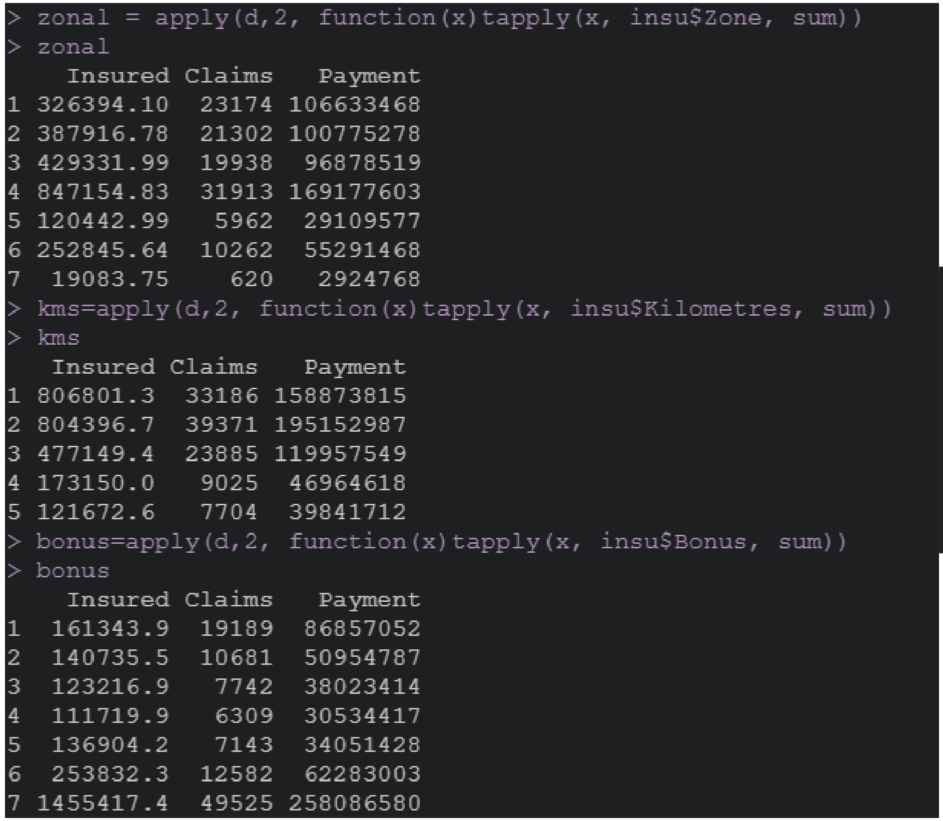
```
lm(formula = payment ~ . - Payment, data = insu)

Residuals:
    Min      1Q  Median      3Q     Max 
-806775  -16943   -6321   11528  847015 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.173e+04  6.338e+03  -3.429 0.000617 ***
Kilometres   4.769e+03  1.086e+03   4.392 1.18e-05 ***
Zone         2.323e+03  7.735e+02   3.003 0.002703 ** 
Bonus        1.183e+03  7.737e+02   1.529 0.126462    
Make        -7.543e+02  6.107e+02  -1.235 0.216917    
Insured      2.788e+01  6.652e-01  41.913  < 2e-16 ***
Claims       4.316e+03  1.895e+01 227.793  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70830 on 2175 degrees of freedom
Multiple R-squared:  0.9952,	Adjusted R-squared:  0.9952 
F-statistic: 7.462e+04 on 6 and 2175 DF,  p-value: < 2.2e-16
```



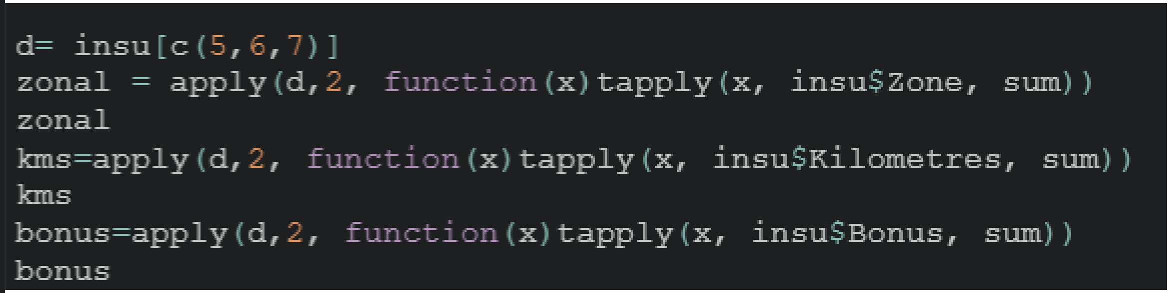
```
Residuals:
    Min      1Q  Median      3Q     Max 
-802620  -16750   -6721   11836  848754 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.047e+04  4.874e+03  -4.199 2.79e-05 ***
Kilometres   4.756e+03  1.085e+03   4.382 1.23e-05 ***
Zone         2.293e+03  7.732e+02   2.965  0.00306 ** 
Insured      2.814e+01  6.518e-01  43.176  < 2e-16 ***
Claims       4.308e+03  1.842e+01 233.829  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70860 on 2177 degrees of freedom
Multiple R-squared:  0.9952,	Adjusted R-squared:  0.9951 
F-statistic: 1.118e+05 on 4 and 2177 DF,  p-value: < 2.2e-16
```

- Kilometers, Zone, Insured, Claims are significant variables whose values will affect the Payment variable
- Deleted insignificant variables Bonus and Make from the model insu to create model (m2)
- Data shows a low p value hence the model (m2) is good for linear regression,
- R2 value and adjusted R2 values is very very close, hence the model is good for linear regression

Q4. 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 gets increased. (Hint: Aggregate Dataset)



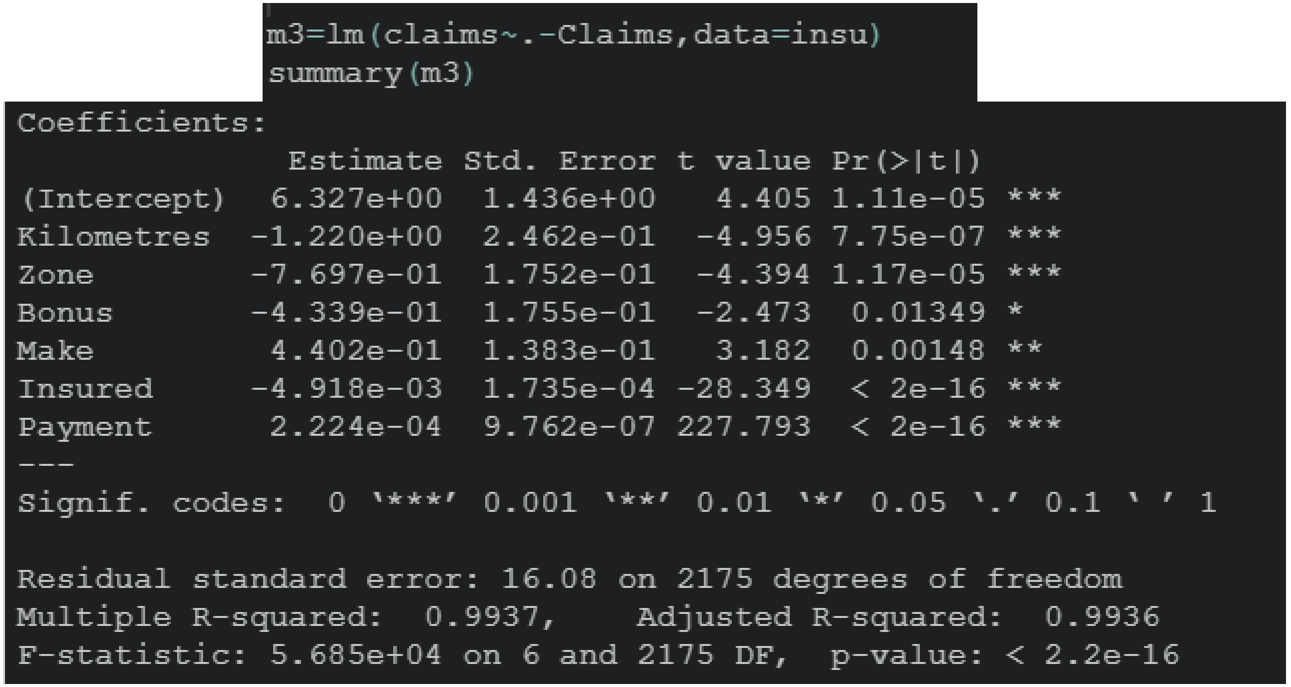
```
> zonal = apply(d,2, function(x)tapply(x, insu$Zone, sum))
> zonal
    Insured Claims   Payment
1 326394.10  23174 106633468
2 387916.78  21302 100775278
3 429331.99  19938  96878519
4 847154.83  31913 169177603
5 120442.99   5962  29109577
6 252845.64  10262  55291468
7  19083.75    620   2924768
> kms=apply(d,2, function(x)tapply(x, insu$Kilometres, sum))
> kms
   Insured Claims   Payment
1 806801.3  33186 158873815
2 804396.7  39371 195152987
3 477149.4  23885 119957549
4 173150.0   9025  46964618
5 121672.6   7704  39841712
> bonus=apply(d,2, function(x)tapply(x, insu$Bonus, sum))
> bonus
    Insured Claims   Payment
1  161343.9  19189  86857052
2  140735.5  10681  50954787
3  123216.9   7742  38023414
4  111719.9   6309  30534417
5  136904.2   7143  34051428
6  253832.3  12582  62283003
7 1455417.4  49525 258086580
```



```
d= insu[c(5,6,7)]
zonal = apply(d,2, function(x)tapply(x, insu$Zone, sum))
zonal
kms=apply(d,2, function(x)tapply(x, insu$Kilometres, sum))
kms
bonus=apply(d,2, function(x)tapply(x, insu$Bonus, sum))
bonus
```

- Aggregate the data for Kilometers, Bonus and Zone by Insured, Claims, Payment
- The highest payment, claims and insured are in Zone-4 and Zone 7 has the lowest Insured, Claims and Payment
- Zone-2 in Kilometer wise has the highest claims and payments
- Group-7 has the highest Insured, Claims and Payment

Q5. 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, kilometre, bonus, or make affects the claim rates and to what extent.



```
m3=lm(claims~.-Claims,data=insu)
summary(m3)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.327e+00  1.436e+00   4.405 1.11e-05 ***
Kilometres  -1.220e+00  2.462e-01  -4.956 7.75e-07 ***
Zone        -7.697e-01  1.752e-01  -4.394 1.17e-05 ***
Bonus       -4.339e-01  1.755e-01  -2.473  0.01349 *
Make         4.402e-01  1.383e-01   3.182  0.00148 **
Insured     -4.918e-03  1.735e-04 -28.349  < 2e-16 ***
Payment      2.224e-04  9.762e-07 227.793  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.08 on 2175 degrees of freedom
Multiple R-squared:  0.9937,    Adjusted R-squared:  0.9936
F-statistic: 5.685e+04 on 6 and 2175 DF,  p-value: < 2.2e-16
```

- Using dependent variable claims and others as independent variable created a linear regression model (m3)
- Data shows a low p value hence the model (m3) is good for linear regression
- R2 value and adjusted R2 values is very very close, hence the model is good for linear regression
- Independent variables are highly significant and have a strong influence on the Claims variable

R Worksheet

```
rm(list=ls())
setwd("D:/Simplilearn/DataScience with R/Project_Insurance")
insu=read.csv("Insurance_factor_identification.csv")
summary(insu)
View(insu)
payment=insu$Payment
insured=insu$Insured
claims=insu$Claims

m1=lm(payment~.-Payment,data=insu)
summary(m1)
m2=lm(payment~.-Payment-Bonus-Make,data=insu)
summary(m2)

cor(insu$Claims,insu$Payment)
cor(insu$Insured,insu$Payment)

plot(payment,insured)
plot(payment,claims)
abline(lm(insured~payment))
abline(lm(claims~payment))

library(dplyr)
d= insu[c(5,6,7)]
zonal = apply(d,2, function(x)tapply(x, insu$Zone, sum))
zonal
kms=apply(d,2, function(x)tapply(x, insu$Kilometres, sum))
kms
bonus=apply(d,2, function(x)tapply(x, insu$Bonus, sum))
bonus

m3=lm(claims~.-Claims,data=insu)
summary(m3)
```