Analyze the report of Swedish Motor Insurance

Business Analytic Foundation with R Tools- Question

Abstract

In Sweden, all motor insurance companies apply identical risk arguments to classify customers, and thus their portfolios and their claims statistics can be combined. The data were compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The Committee was asked to look into the problem of analyzing the real influence on the claims of the risk arguments and to compare this structure with the actual tariff.

Presented by: Ankita Agarwal

Problem Statement:

The data gives the details of third-party motor insurance claims in Sweden for the year 1977. In Sweden, all motor insurance companies apply identical risk arguments to classify customers, and thus their portfolios and their claims statistics can be combined. The data were compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The Committee was asked to look into the problem of analyzing the real influence on the claims of the risk arguments and to compare this structure with the actual tariff.

Detailed description of the given dataset:

Variable	Description
Kilo-meters	Kilo-meters travelled per year 1: < 1000 2: 1000-15000 3: 15000-20000 4: 20000-25000 5: > 25000
Zone	Geographical zone 1: Stockholm, Goteborg, 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:	Number of insured in policy-years
Claims:	Number of claims
Payment:	Total value of payments in SKR (Swedish Krona)

To Analyze:

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.

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.

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.

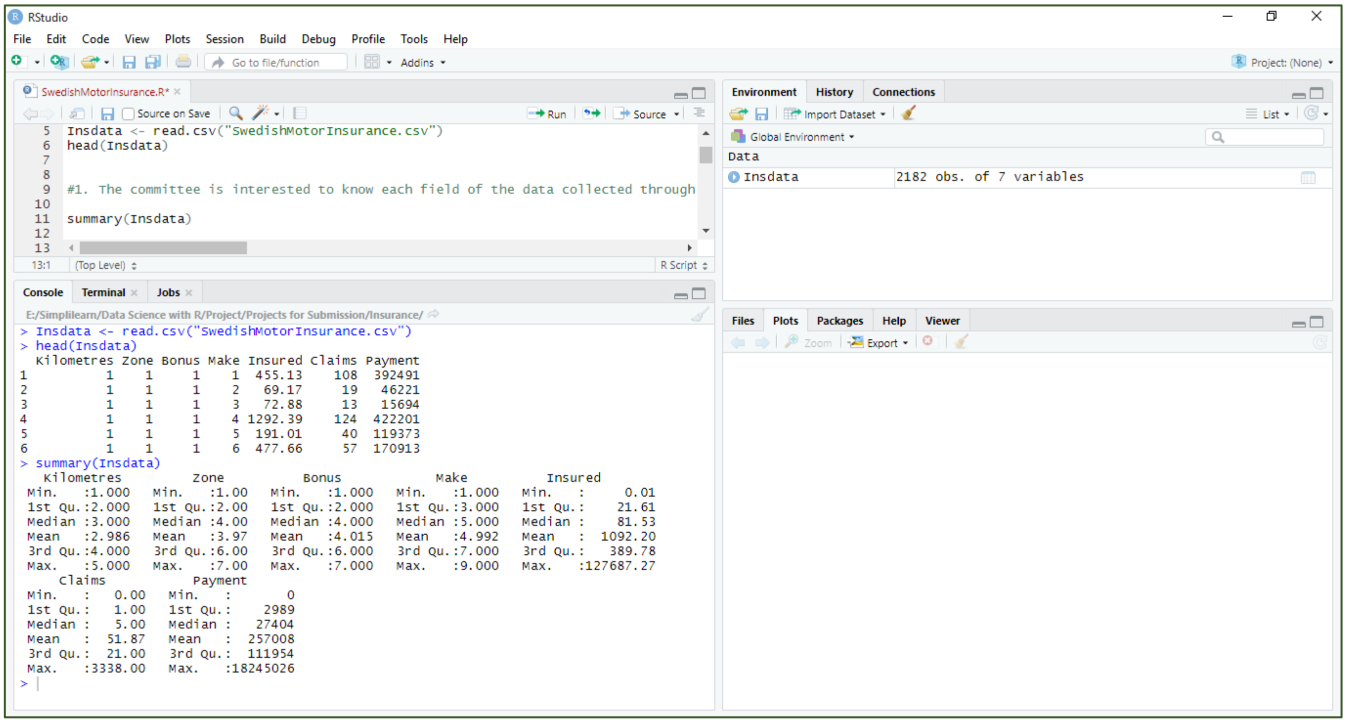
4. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilo-meter, and bonus level their insured amount, claims, and payment get increased. (Hint: Aggregate Dataset)

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, kilo meter, bonus, or make affects the claim rates and to what extent.

Analysis and Interpretations:

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.

To get a basic insight into the data, we do a 5-point summary analysis for the data set.



```
Insdata <- read.csv("SwedishMotorInsurance.csv")
head(Insdata)


#1. The committee is interested to know each field of the data collected through

summary(Insdata)
```

E:/Simplilearn/Data Science with R/Project/Projects for Submission/Insurance/

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

Data

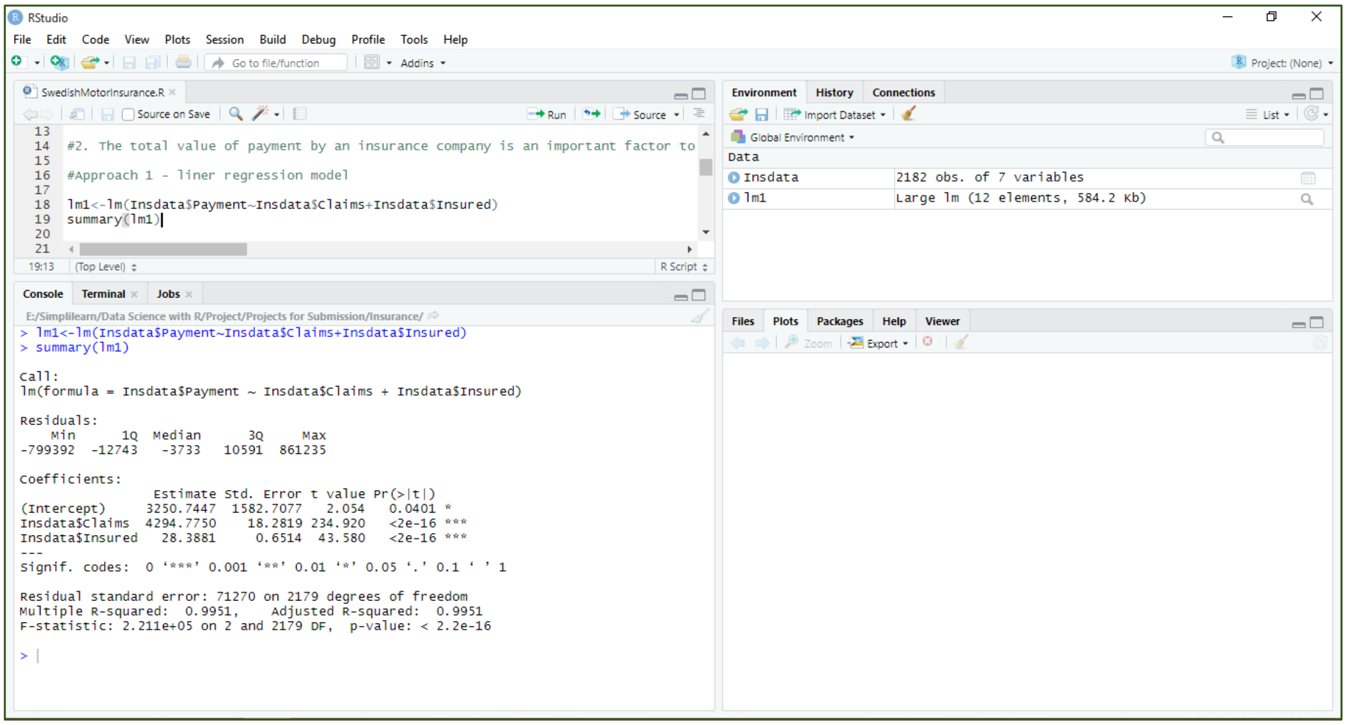
Insdata   2182 obs. of 7 variables

Interpretation:

By looking at the 5-point summary of the data, we can safely say that the range between the minimum and maximum data for Claims and Payment is very large with 0 being the least and 3338 and 18245026 being the maximum respectively. We do see a few 0s in these variables though all the cars have been insured telling us that some cars have never claimed for the Insurance.

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.

In order to check the dependencies of the variables to each other (payment with claims and payment with number of policy years) we can either create a linear regression model or check the correlation between the variables.



```
#2. The total value of payment by an insurance company is an important factor to

#Approach 1 - liner regression model

lm1<-lm(Insdata$Payment~Insdata$Claims+Insdata$Insured)
summary(lm1)
```

E:/Simplilearn/Data Science with R/Project/Projects for Submission/Insurance/

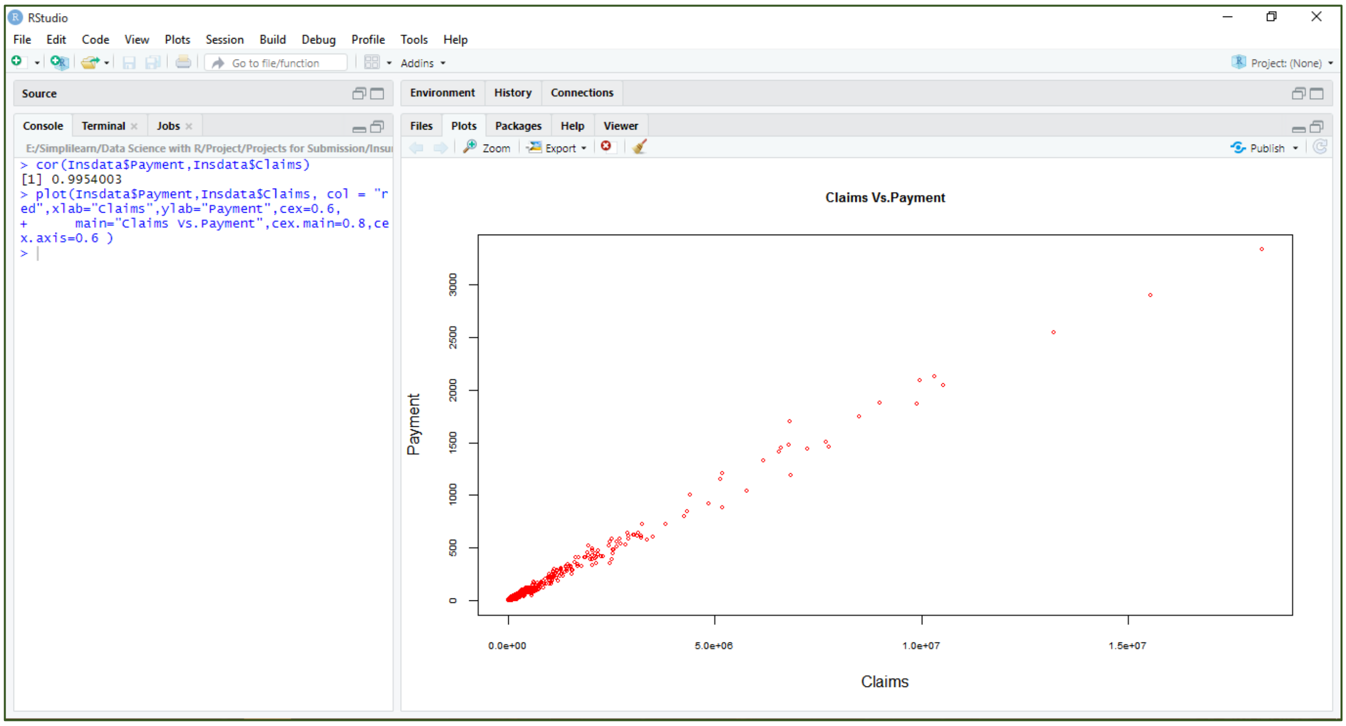
```
> lm1<-lm(Insdata$Payment~Insdata$Claims+Insdata$Insured)
> summary(lm1)

Call:
lm(formula = Insdata$Payment ~ Insdata$Claims + Insdata$Insured)

Residuals:
    Min      1Q  Median      3Q     Max 
-799392  -12743   -3733   10591  861235 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3250.7447  1582.7077   2.054   0.0401 *  
Insdata$Claims  4294.7750    18.2819 234.920   <2e-16 ***
Insdata$Insured   28.3881     0.6514  43.580   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71270 on 2179 degrees of freedom
Multiple R-squared:  0.9951,	Adjusted R-squared:  0.9951 
F-statistic: 2.211e+05 on 2 and 2179 DF,  p-value: < 2.2e-16

> 
```

Data

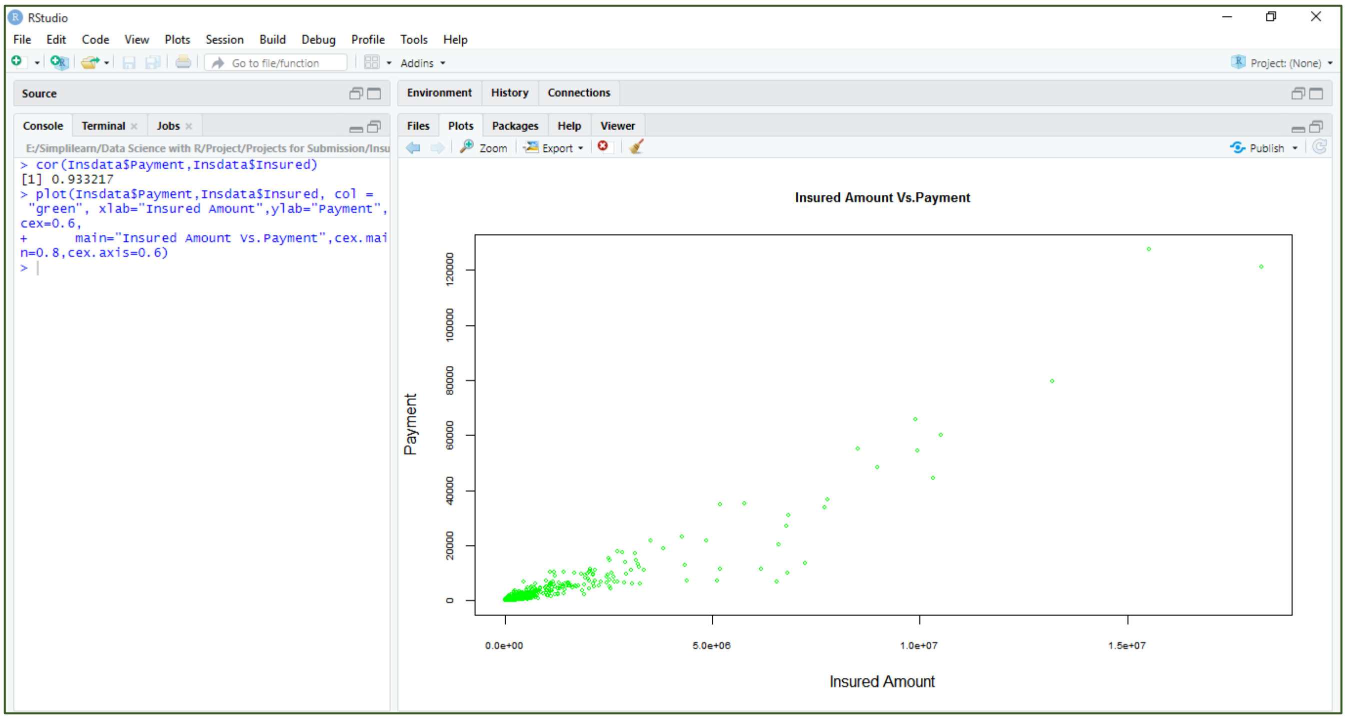
Insdata	2182 obs. of 7 variables
lm1	Large lm (12 elements, 584.2 Kb)

Interpretation:

Here, p-value for both claims and number of years insured is way less than the 0.05 we assume while creating any linear model signifying that payments is positively dependent on both the variables.



```
> cor(Insdata$Payment,Insdata$Claims)
[1] 0.9954003
> plot(Insdata$Payment,Insdata$Claims, col = "r
ed",xlab="Claims",ylab="Payment",cex=0.6,
+       main="Claims Vs.Payment",cex.main=0.8,ce
x.axis=0.6 )
> 
```



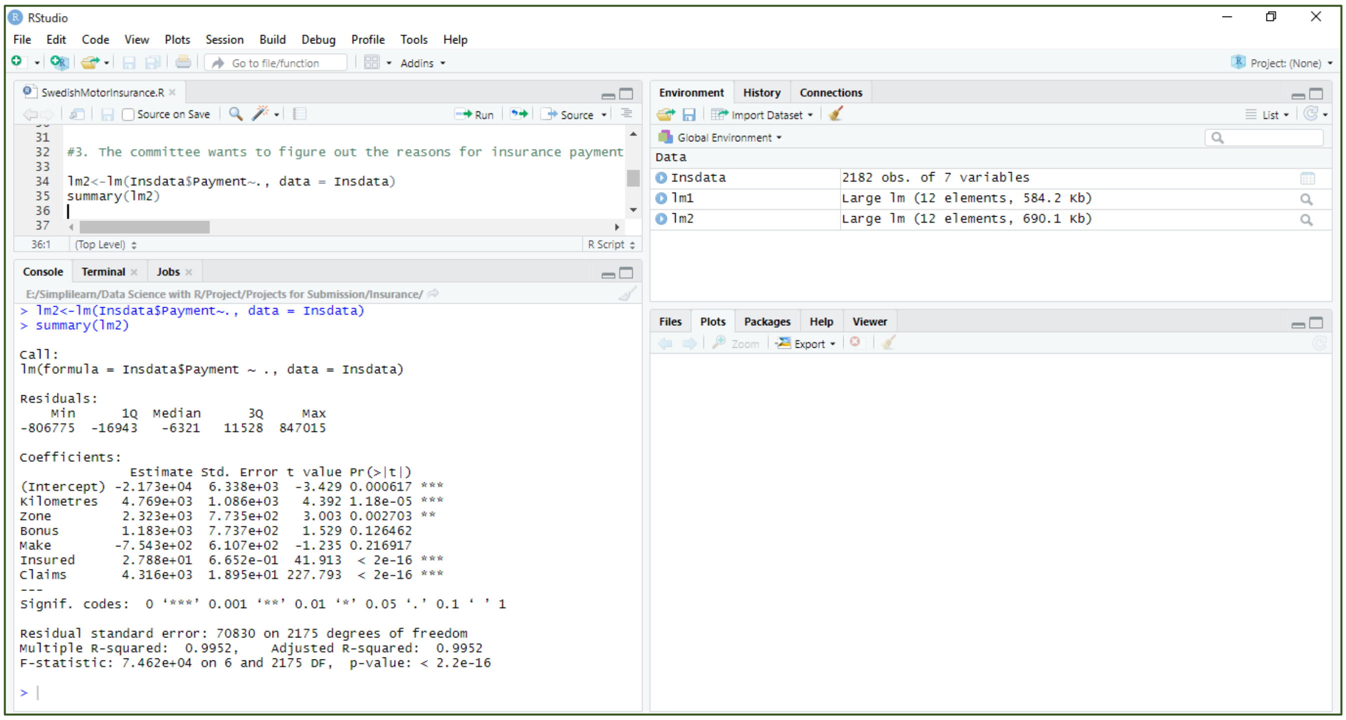
```
> cor(Insdata$Payment,Insdata$Insured)
[1] 0.933217
> plot(Insdata$Payment,Insdata$Insured, col =
 "green", xlab="Insured Amount",ylab="Payment",
cex=0.6,
+       main="Insured Amount Vs.Payment",cex.mai
n=0.8,cex.axis=0.6)
> 
```

Interpretation:

As seen in the plots above, both variables, claims and number of years insured form an almost positive linear correlation curve with payment signifying a direct dependencies between the variables.

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.

To check the dependencies of distance, location, bonus, make, insured amount and claims on payment we create a linear regression model.



```
#3. The committee wants to figure out the reasons for insurance payment

lm2<-lm(Insdata$Payment~., data = Insdata)
summary(lm2)
```

```
> lm2<-lm(Insdata$Payment~., data = Insdata)
> summary(lm2)

Call:
lm(formula = Insdata$Payment ~ ., data = Insdata)

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

> |
```

Data	
Insdata	2182 obs. of 7 variables
lm1	Large lm (12 elements, 584.2 Kb)
lm2	Large lm (12 elements, 690.1 Kb)

Interpretation:

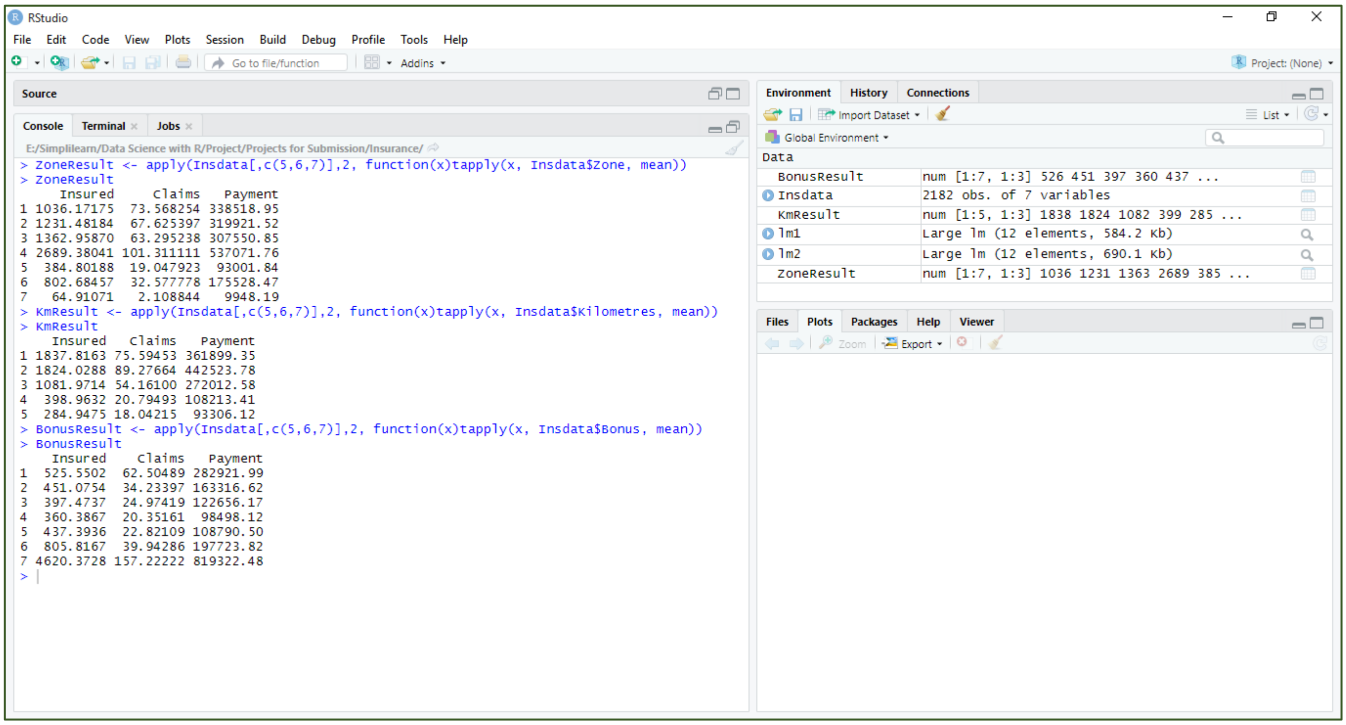
As we know, when p-value is less than 0.05, which we assume while creating a linear regression model, we consider that variable to be significant in the model.

Here, we can see that Bonus and Make have very high p-values of 0.13 and 0.22. Hence, we can say that they are not significant variables in this data. This means that these variables do not make much difference to the payment.

However, Kilometers (distance), Zone (region), number of years insured and claims have significantly less p-value and contribute or effect payment.

4. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilo-meter, and bonus level their insured amount, claims, and payment get increased.

To find at what level the claims and payment increases, we need to find the categorical values of location, kilometer and bonus.



```
E:/Simplilearn/Data Science with R/Project/Projects for Submission/Insurance/
> ZoneResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Zone, mean))
> ZoneResult
     Insured     Claims   Payment
1 1036.17175  73.568254 338518.95
2 1231.48184  67.625397 319921.52
3 1362.95870  63.295238 307550.85
4 2689.38041 101.311111 537071.76
5  384.80188  19.047923  93001.84
6  802.68457  32.577778 175528.47
7   64.91071   2.108844   9948.19
> KmResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Kilometres, mean))
> KmResult
    Insured   Claims   Payment
1 1837.8163 75.59453 361899.35
2 1824.0288 89.27664 442523.78
3 1081.9714 54.16100 272012.58
4  398.9632 20.79493 108213.41
5  284.9475 18.04215  93306.12
> BonusResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Bonus, mean))
> BonusResult
    Insured    Claims   Payment
1  525.5502  62.50489 282921.99
2  451.0754  34.23397 163316.62
3  397.4737  24.97419 122656.17
4  360.3867  20.35161  98498.12
5  437.3936  22.82109 108790.50
6  805.8167  39.94286 197723.82
7 4620.3728 157.22222 819322.48
> 
```

```
Data
  BonusResult   num [1:7, 1:3] 526 451 397 360 437 ...
  Insdata       2182 obs. of 7 variables
  KmResult      num [1:5, 1:3] 1838 1824 1082 399 285 ...
  lm1           Large lm (12 elements, 584.2 Kb)
  lm2           Large lm (12 elements, 690.1 Kb)
  ZoneResult    num [1:7, 1:3] 1036 1231 1363 2689 385 ...
```

Interpretation:

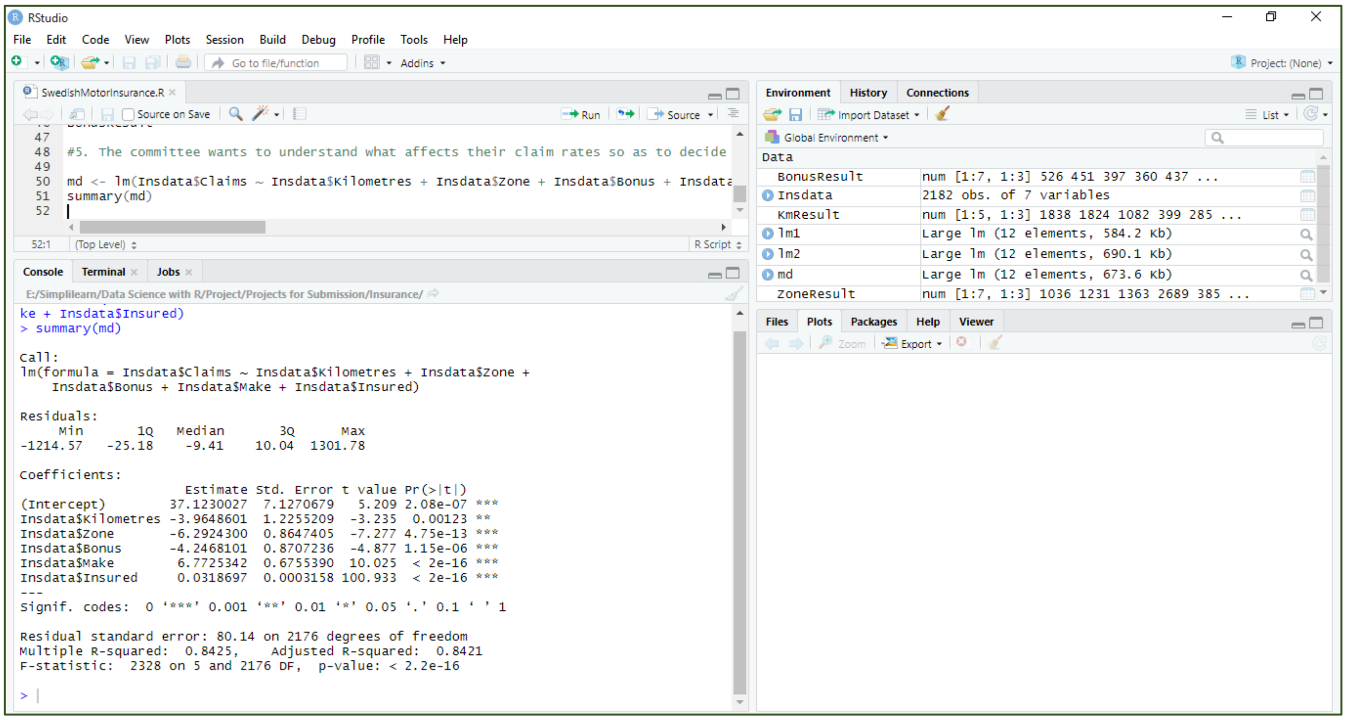
From the above, we can see that Zone 4 has the maximum number of insured years, claims and thus payment whereas Zone 7 has the least number of insured years and claims making payments in the zone the least. Zone 1-4 have the maximum number of insured years, claims and corresponding payments.

Kilometer group 5 has the least number of insured years, claims and payment while group 1 has the maximum number of insured years, claims and thus payments.

Bonus group 7 has a radically high number of insured years, claims and subsequent payments as compared to other groups.

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, kilo meter, bonus, or make affects the claim rates and to what extent.

To understand what variable affects the claim rates and the extent of it we create a linear regression model.



```
#5. The committee wants to understand what affects their claim rates so as to decide

md <- lm(Insdata$Claims ~ Insdata$Kilometres + Insdata$Zone + Insdata$Bonus + Insdata
summary(md)
```

```
ke + Insdata$Insured)
> summary(md)

Call:
lm(formula = Insdata$Claims ~ Insdata$Kilometres + Insdata$Zone +
    Insdata$Bonus + Insdata$Make + Insdata$Insured)

Residuals:
     Min       1Q   Median       3Q      Max
-1214.57   -25.18    -9.41    10.04  1301.78

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        37.1230027  7.1270679   5.209 2.08e-07 ***
Insdata$Kilometres -3.9648601  1.2255209  -3.235  0.00123 **
Insdata$Zone       -6.2924300  0.8647405  -7.277 4.75e-13 ***
Insdata$Bonus      -4.2468101  0.8707236  -4.877 1.15e-06 ***
Insdata$Make        6.7725342  0.6755390  10.025  < 2e-16 ***
Insdata$Insured     0.0318697  0.0003158 100.933  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80.14 on 2176 degrees of freedom
Multiple R-squared:  0.8425,    Adjusted R-squared:  0.8421
F-statistic:  2328 on 5 and 2176 DF,  p-value: < 2.2e-16

> 
```

```
Data
  BonusResult   num [1:7, 1:3] 526 451 397 360 437 ...
  Insdata       2182 obs. of 7 variables
  KmResult      num [1:5, 1:3] 1838 1824 1082 399 285 ...
  lm1           Large lm (12 elements, 584.2 Kb)
  lm2           Large lm (12 elements, 690.1 Kb)
  md            Large lm (12 elements, 673.6 Kb)
  ZoneResult    num [1:7, 1:3] 1036 1231 1363 2689 385 ...
```

Interpretation:

As we know, when p-value is less than 0.05, which we assume while creating a linear regression model, we consider that variable to be significant in the model.

Here, we can see that all independent variables have p-values less than 0.05 and are quite significant and thus, have a strong influence on claims.

Programming Codes:

```
#Reading Insurance Data

Insdata <- read.csv("SwedishMotorInsurance.csv")
head(Insdata)


#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.

summary(Insdata)

#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.

#Approach 1 - liner regression model

lm1<-lm(Insdata$Payment~Insdata$Claims+Insdata$Insured)
summary(lm1)

#Approach 2 - Find correlation between variables

cor(Insdata$Payment,Insdata$Claims)
plot(Insdata$Payment,Insdata$Claims, col = "red",xlab="Claims",ylab="Payment",cex=0.6,
    main="Claims Vs.Payment",cex.main=0.8,cex.axis=0.6 )

cor(Insdata$Payment,Insdata$Insured)
plot(Insdata$Payment,Insdata$Insured, col = "green", xlab="Insured
Amount",ylab="Payment",cex=0.6,
    main="Insured Amount Vs.Payment",cex.main=0.8,cex.axis=0.6)


#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.

lm2<-lm(Insdata$Payment~., data = Insdata)
summary(lm2)
```

#4. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilo-meter, and bonus level their insured amount, claims, and payment get increased. (Hint: Aggregate Dataset)

```
ZoneResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Zone, mean))
ZoneResult

KmResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Kilometres, mean))
KmResult

BonusResult <- apply(Insdata[,c(5,6,7)],2, function(x)tapply(x, Insdata$Bonus, mean))
BonusResult
```

#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, kilo meter, bonus, or make affects the claim rates and to what extent.

```
md <- lm(Insdata$Claims ~ Insdata$Kilometres + Insdata$Zone + Insdata$Bonus + Insdata$Make +
Insdata$Insured)
summary(md)
```

---The End---