Project : Insurance Factors Identification
Domain : Insurance

Description:

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.The dataset contains 7 variables and their description is given below:
Kilometres : Kilometers travelled per year.
Zone : Geographical zones
Bonus : no claims bonus
Make : represents different common car models.
Insured : The number of insured in policy - years
Claims : Number of claims
Payment : Total value of payments in Swedish Krona

Objective:

The committee wants to do the following analysis in order to open a new branch and decide the right premiums for a certain set of situations. They are :

1. To do a descriptive analysis of the data collected in order to get insights on the data and to prepare for further analysis.
2. To find whether the payment is related to the number of claims and the number of insured in policy years and also visualize the results for better understanding.
3. 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. To find at what location, kilometer and bonus level the insured amount, claims, and payment gets increased.
5. To find whether the insured amount, zone, kilometer, bonus, or make affects the claim rates and to what extent.

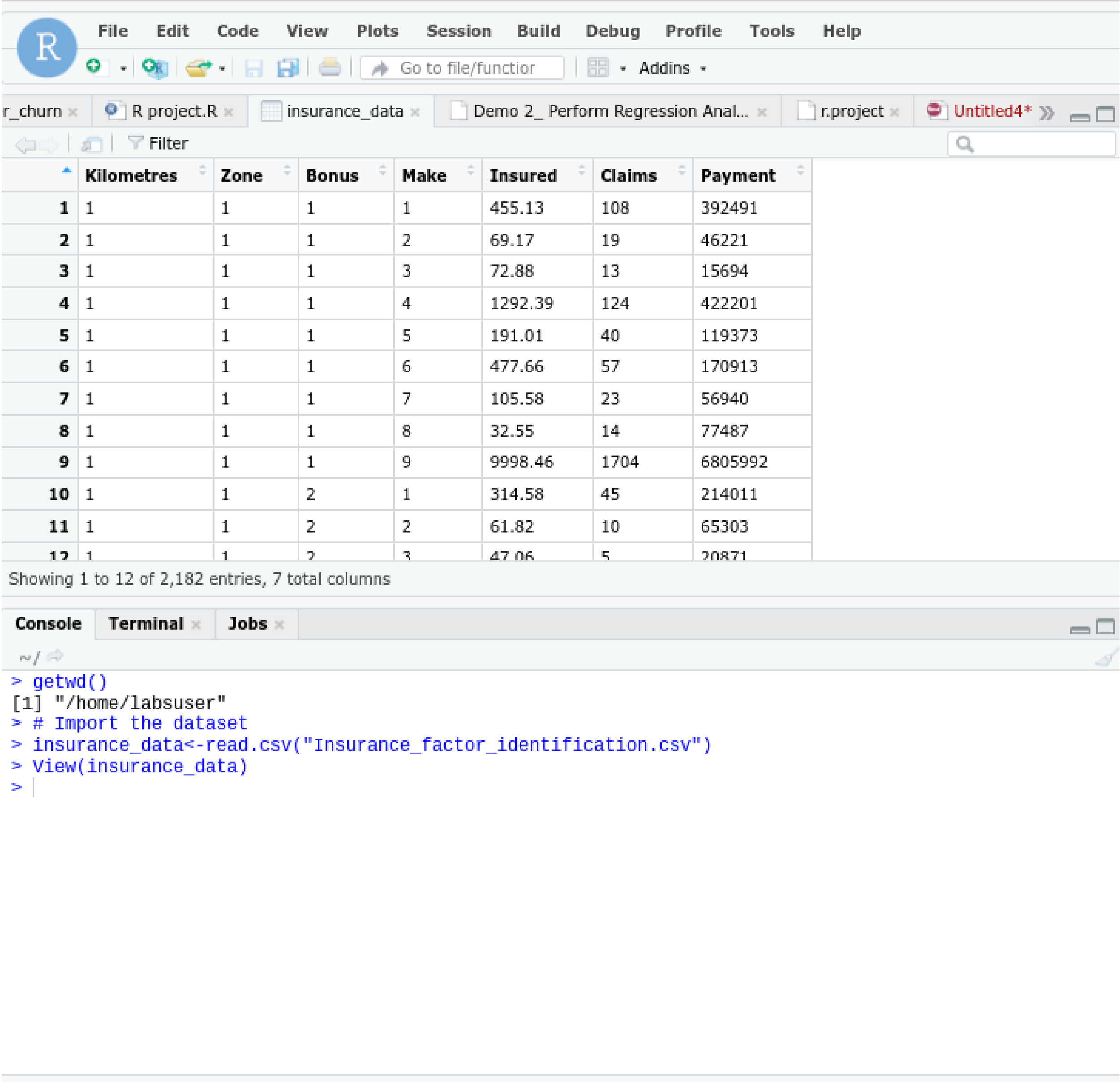
Codes:

```
setwd()
getwd()
# Import the dataset
insurance_data<-read.csv("Insurance_factor_identification.csv")
View(insurance_data)

# 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.
# statistical summary of the insurance data
summary(insurance_data)
str(insurance_data) # structure of the insurance data

# 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.
# correlation of claims and no. of insured years with payment
cor(insurance_data$Claims, insurance_data$Payment)
# correlation of Claims with Payment = 99.54%. They are positively correlated.
cor(insurance_data$Insured, insurance_data$Payment)
```

```
# correlation of insured with payment = 93.32%. They are positively correlated.

# Regression model to check the relation between Claims, Insured with dependent variable
# Payment.
result<-lm(formula = Payment~Insured + Claims, data = insurance_data)
summary(result)

# Visualizing the data
plot(insurance_data$Claims, insurance_data$Payment)
plot(insurance_data$Insured, insurance_data$Payment)

# 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.
# Regression model to check the relation between the dependent variable, Payments with all the
# independent variables.
result1 <- lm(formula= Payment~ .,data = insurance_data)
summary(result1)
# From the summary of the regression model, we can conclude that distance, location,Insured and
# Claims have very strong significance with the dependent variable, Payment. p-value of Kilometers
# = 1.18e-05, Zone = 0.027, Insured < 2e-16,  Claims < 2e-16. Bonus and Make have p-value>0.05.
# Hence they are not significant.

# 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 gets
# increased.
agg_kilometer<-aggregate(x=insurance_data[,5:7], by=insurance_data[c(1)], FUN=mean)
agg_kilometer
# At a distance of < 1000 Kilometer, the number of Insured in policy-years is maximum, but the
# Claims and Payments are higher in the Kilometer range of 1000-15000.


agg_location<-aggregate(x=insurance_data[,5:7], by = insurance_data[(2)], FUN=mean)
agg_location
# The Insured, the Claims and the Payments are maximum in Zone 4, ie, the rural areas in Southern
# Sweden.

agg_bonus<-aggregate(x=insurance_data[,5:7], by = insurance_data[(3)], FUN=mean)
agg_bonus
# At a bonus level of 7, the Insured, Claims and Payments are maximum.

# 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.
# Regression model to check the effect of insured amount, zone, kilometer, bonus,make on the claim
# rates
claim_result<-lm(formula=Claims~Insured+Zone+Kilometres+Bonus+Make, data = insurance_data)
summary(claim_result)
# The summary of the Regression model shows that all the independent variables have a strong
# impact on the Claim rates.
```

Output Screenshots :

1. After importing, view the dataset

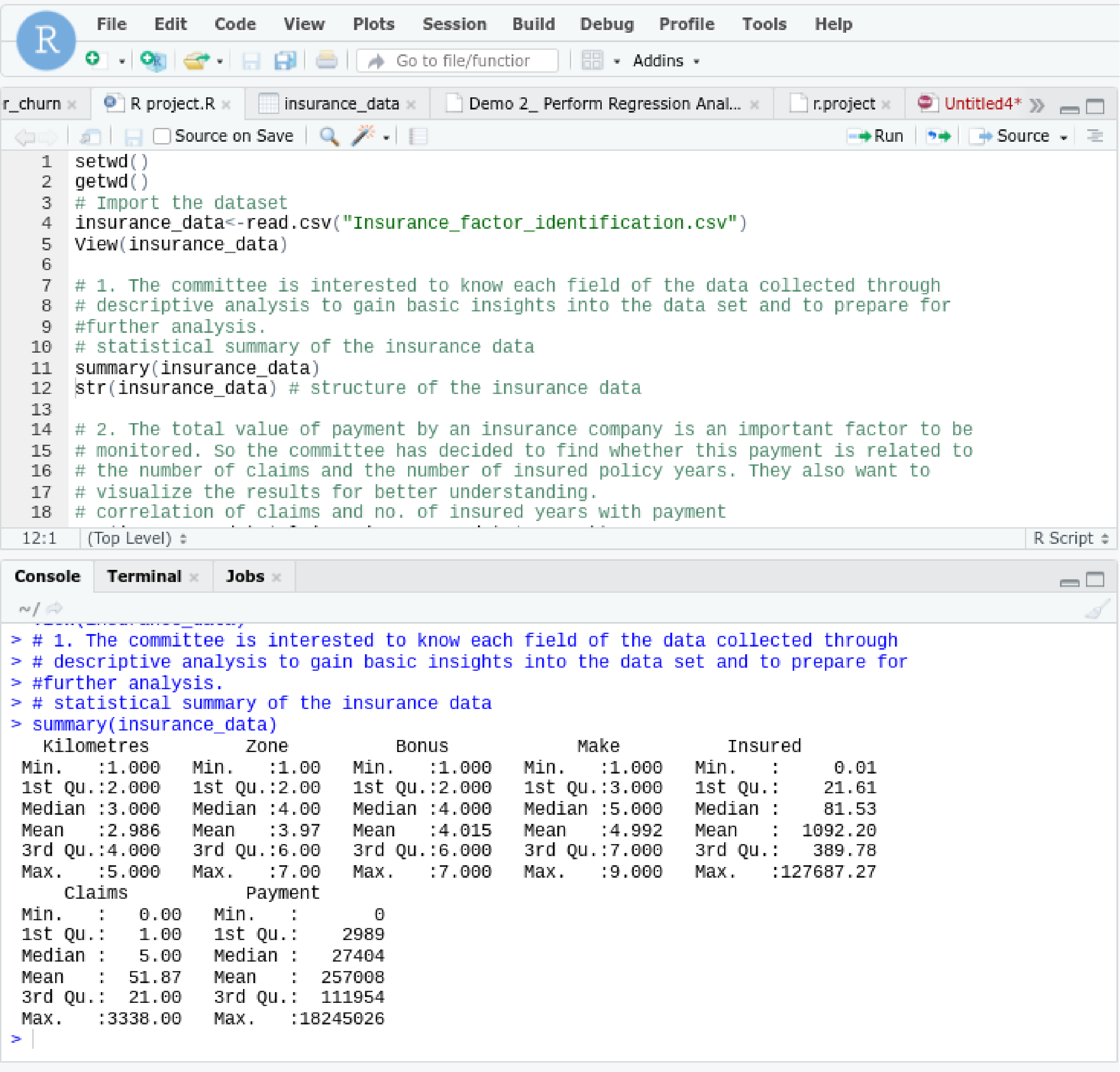


	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment
1	1	1	1	1	455.13	108	392491
2	1	1	1	2	69.17	19	46221
3	1	1	1	3	72.88	13	15694
4	1	1	1	4	1292.39	124	422201
5	1	1	1	5	191.01	40	119373
6	1	1	1	6	477.66	57	170913
7	1	1	1	7	105.58	23	56940
8	1	1	1	8	32.55	14	77487
9	1	1	1	9	9998.46	1704	6805992
10	1	1	2	1	314.58	45	214011
11	1	1	2	2	61.82	10	65303
12	1	1	2	3	47.06	5	20871

Showing 1 to 12 of 2,182 entries, 7 total columns

```
> getwd()
[1] "/home/labsuser"
> # Import the dataset
> insurance_data<-read.csv("Insurance_factor_identification.csv")
> View(insurance_data)
>
```

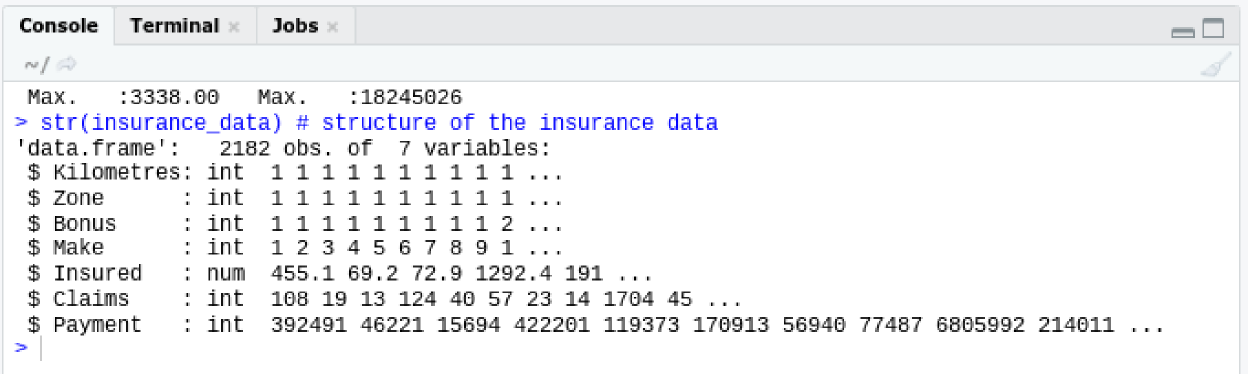
2.Descriptive analysis of the data.



```r
setwd()
getwd()
# Import the dataset
insurance_data<-read.csv("Insurance_factor_identification.csv")
View(insurance_data)

# 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.
# statistical summary of the insurance data
summary(insurance_data)
str(insurance_data) # structure of the insurance data

# 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.
# correlation of claims and no. of insured years with payment
```

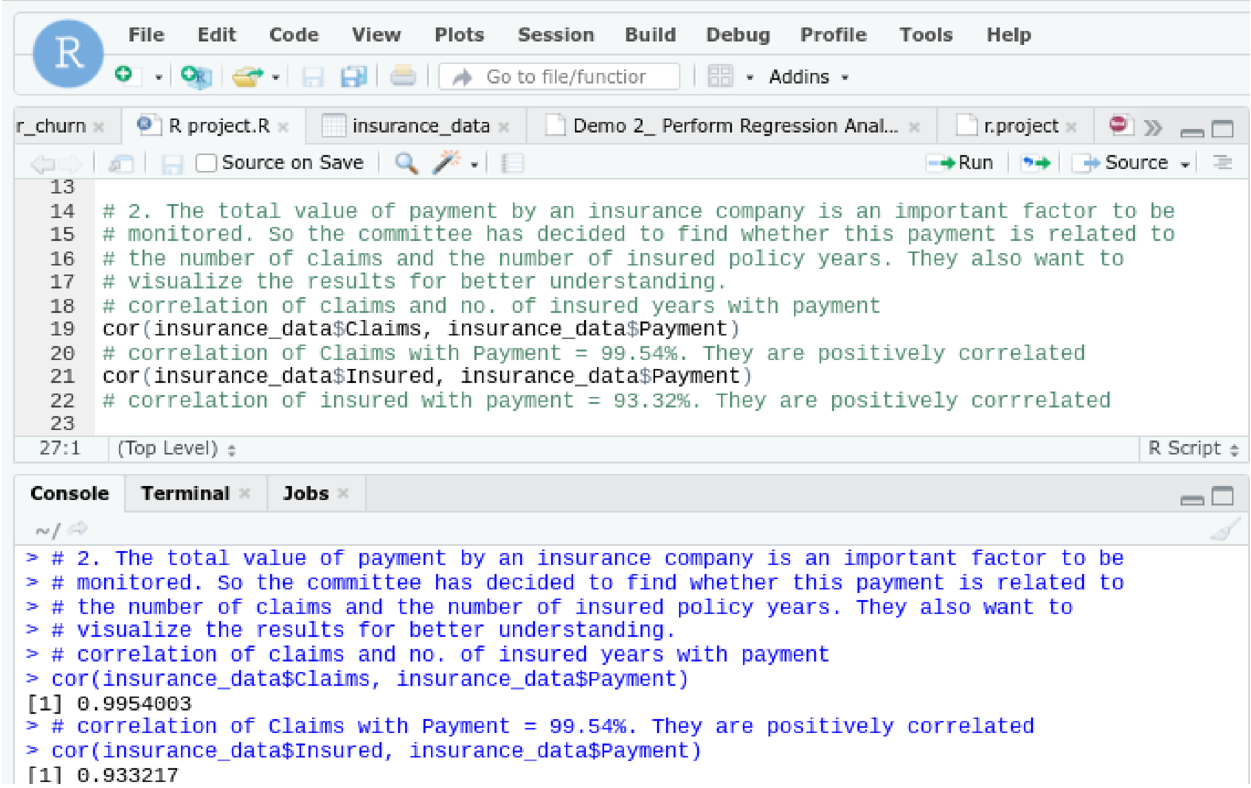
```
> # 1. The committee is interested to know each field of the data collected through
> # descriptive analysis to gain basic insights into the data set and to prepare for
> #further analysis.
> # statistical summary of the insurance data
> summary(insurance_data)
   Kilometres         Zone          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  
> 
```

3. Structure of the data



```
 Max.   :3338.00   Max.   :18245026  
> str(insurance_data) # structure of the insurance data
'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 6805992 214011 ...
> 
```

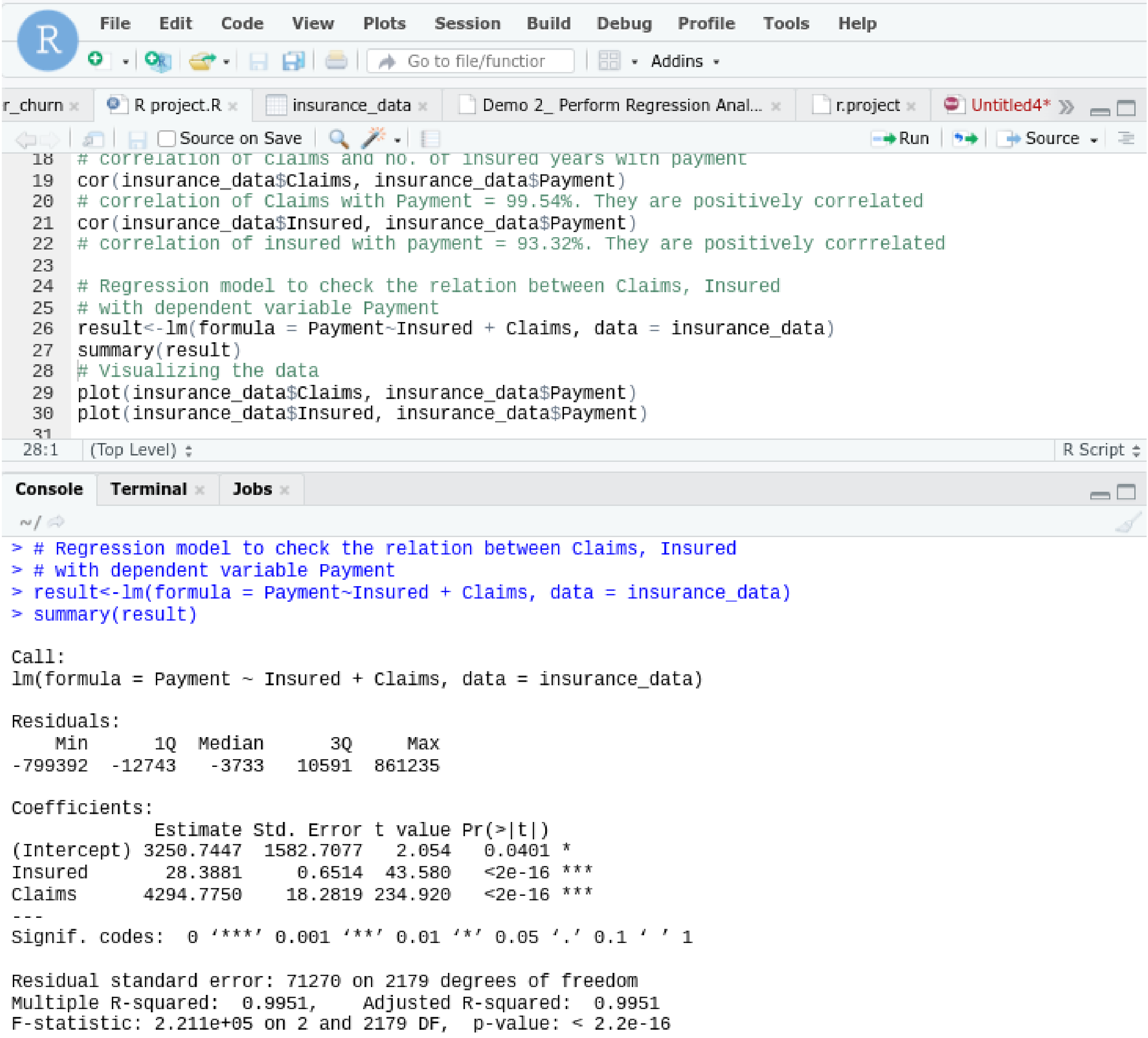
4. Relationship of payment with Claims and Insured.



```
# 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.
# correlation of claims and no. of insured years with payment
cor(insurance_data$Claims, insurance_data$Payment)
# correlation of Claims with Payment = 99.54%. They are positively correlated
cor(insurance_data$Insured, insurance_data$Payment)
# correlation of insured with payment = 93.32%. They are positively corrrelated
```

```
> # 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.
> # correlation of claims and no. of insured years with payment
> cor(insurance_data$Claims, insurance_data$Payment)
[1] 0.9954003
> # correlation of Claims with Payment = 99.54%. They are positively correlated
> cor(insurance_data$Insured, insurance_data$Payment)
[1] 0.933217
```

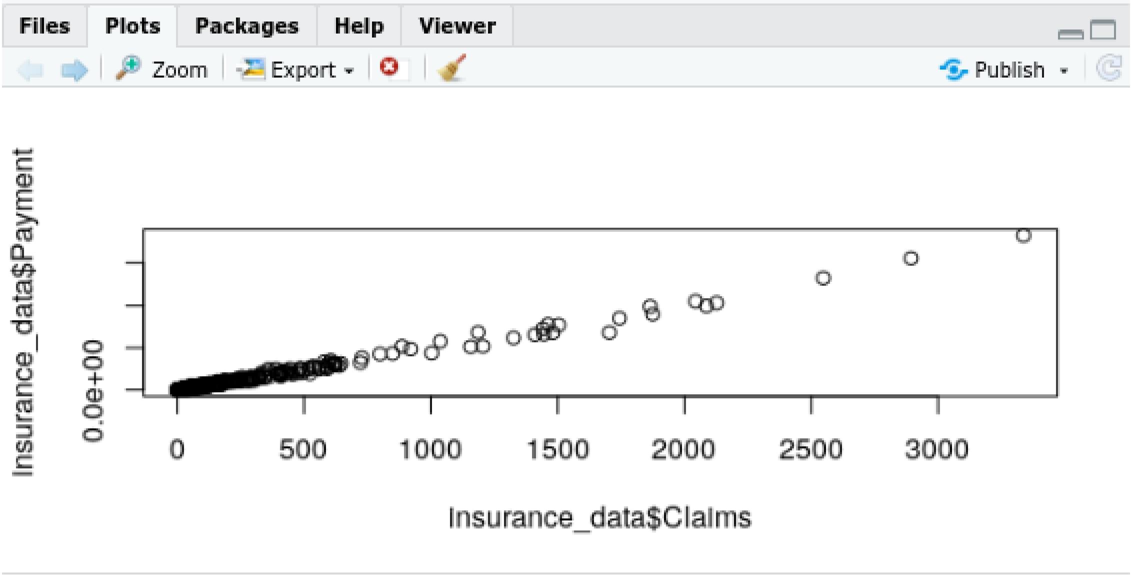
5.RegressionAnalysis

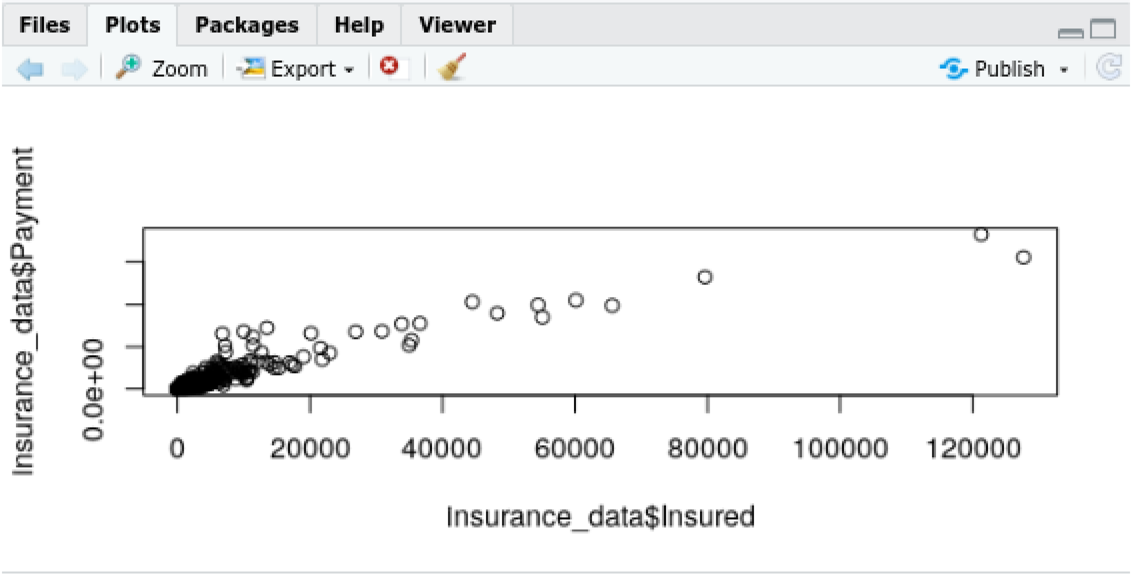


```r
# correlation of claims and no. of insured years with payment
cor(insurance_data$Claims, insurance_data$Payment)
# correlation of Claims with Payment = 99.54%. They are positively correlated
cor(insurance_data$Insured, insurance_data$Payment)
# correlation of insured with payment = 93.32%. They are positively corrrelated

# Regression model to check the relation between Claims, Insured
# with dependent variable Payment
result<-lm(formula = Payment~Insured + Claims, data = insurance_data)
summary(result)
# Visualizing the data
plot(insurance_data$Claims, insurance_data$Payment)
plot(insurance_data$Insured, insurance_data$Payment)
```

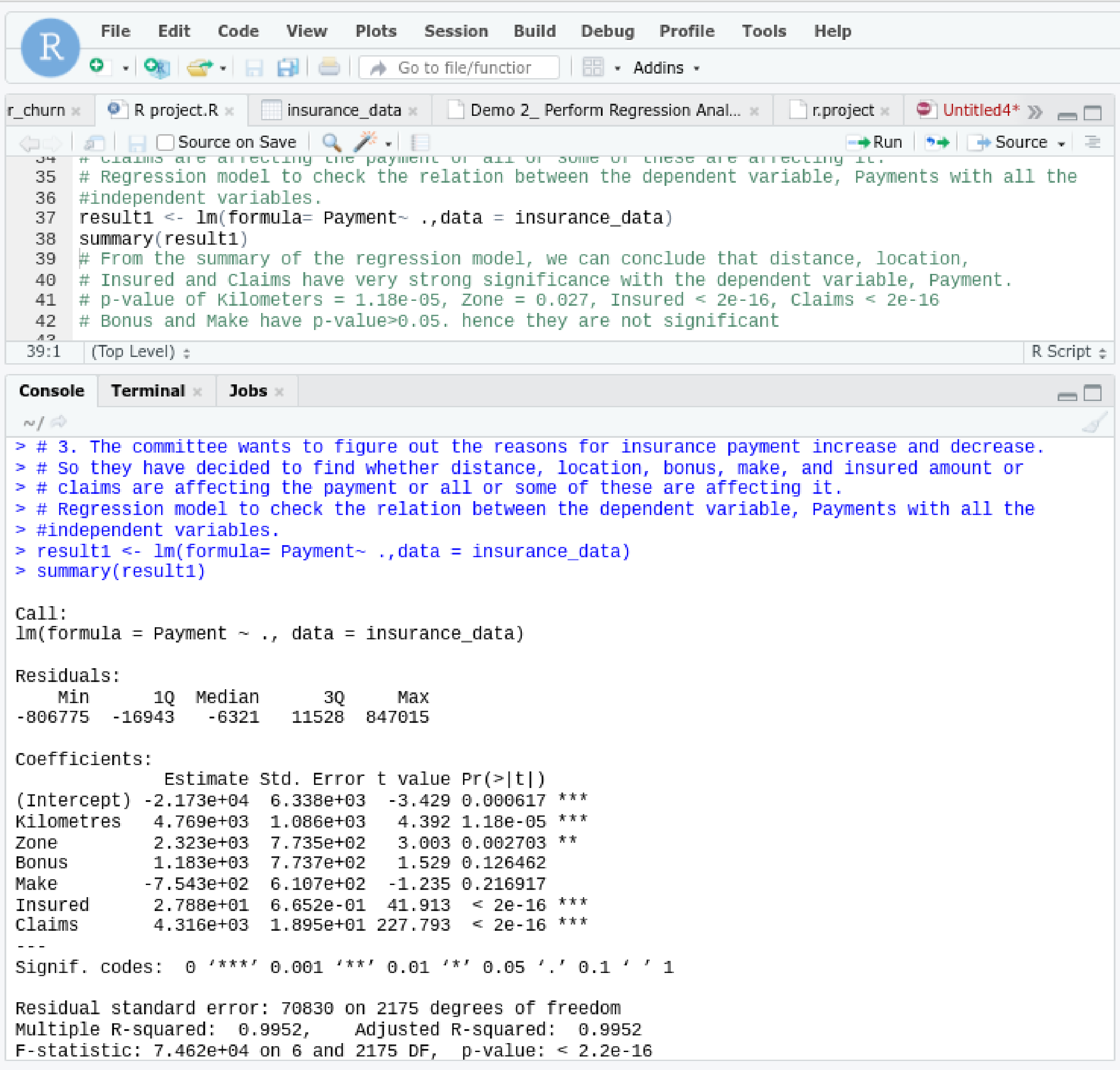
```
> # Regression model to check the relation between Claims, Insured
> # with dependent variable Payment
> result<-lm(formula = Payment~Insured + Claims, data = insurance_data)
> summary(result)

Call:
lm(formula = Payment ~ Insured + Claims, data = insurance_data)

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 *  
Insured       28.3881     0.6514  43.580   <2e-16 ***
Claims      4294.7750    18.2819 234.920   <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
```

6.Visualization





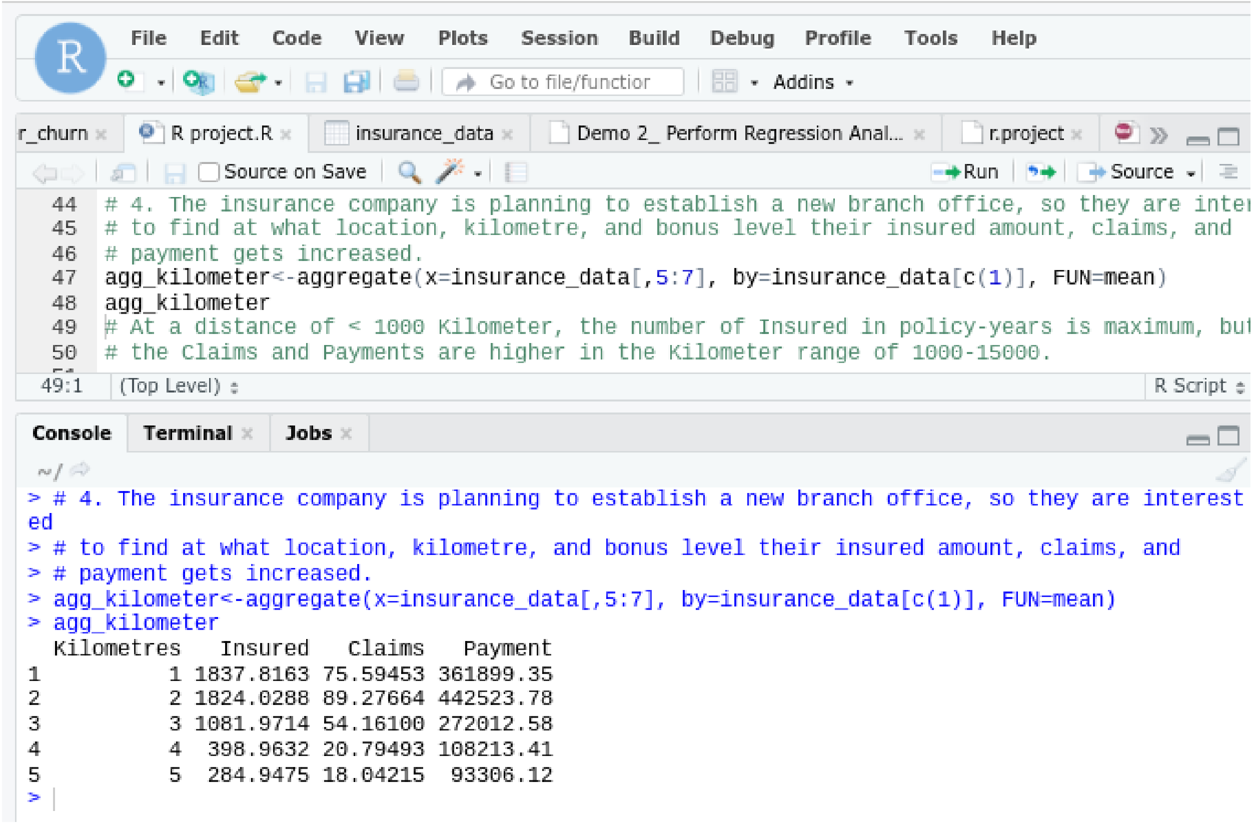
7. Regression model to check the relationship of Payment with all the independent variables.



```
# claims are affecting the payment or all or some of these are affecting it.
# Regression model to check the relation between the dependent variable, Payments with all the
#independent variables.
result1 <- lm(formula= Payment~ .,data = insurance_data)
summary(result1)
# From the summary of the regression model, we can conclude that distance, location,
# Insured and Claims have very strong significance with the dependent variable, Payment.
# p-value of Kilometers = 1.18e-05, Zone = 0.027, Insured < 2e-16, Claims < 2e-16
# Bonus and Make have p-value>0.05. hence they are not significant
```

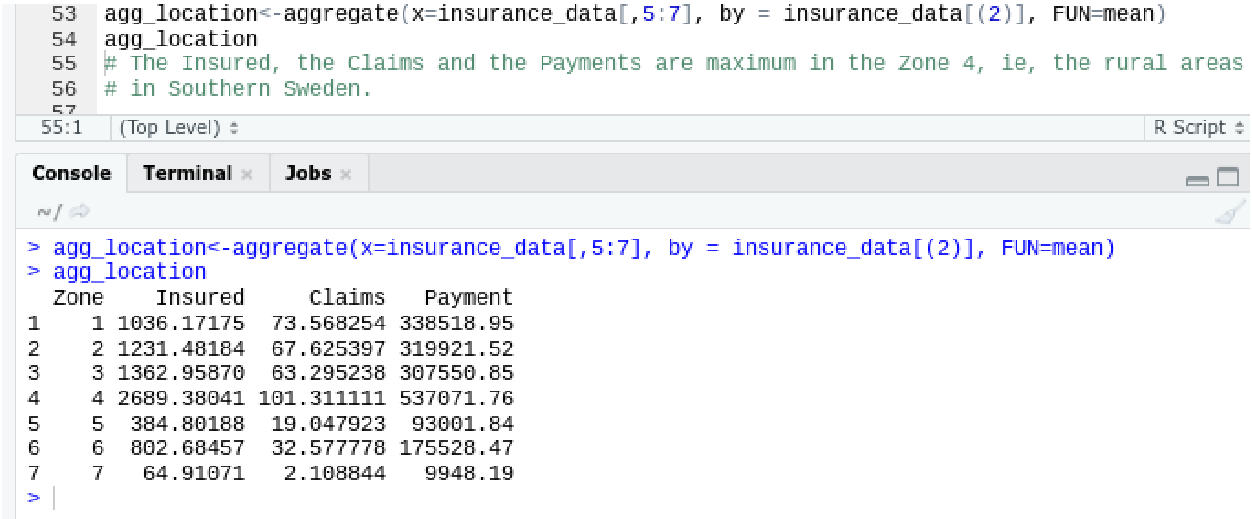
```
> # 3. The committee wants to figure out the reasons for insurance payment increase and decrease.
> # So they have decided to find whether distance, location, bonus, make, and insured amount or
> # claims are affecting the payment or all or some of these are affecting it.
> # Regression model to check the relation between the dependent variable, Payments with all the
> #independent variables.
> result1 <- lm(formula= Payment~ .,data = insurance_data)
> summary(result1)

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

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
```

8.



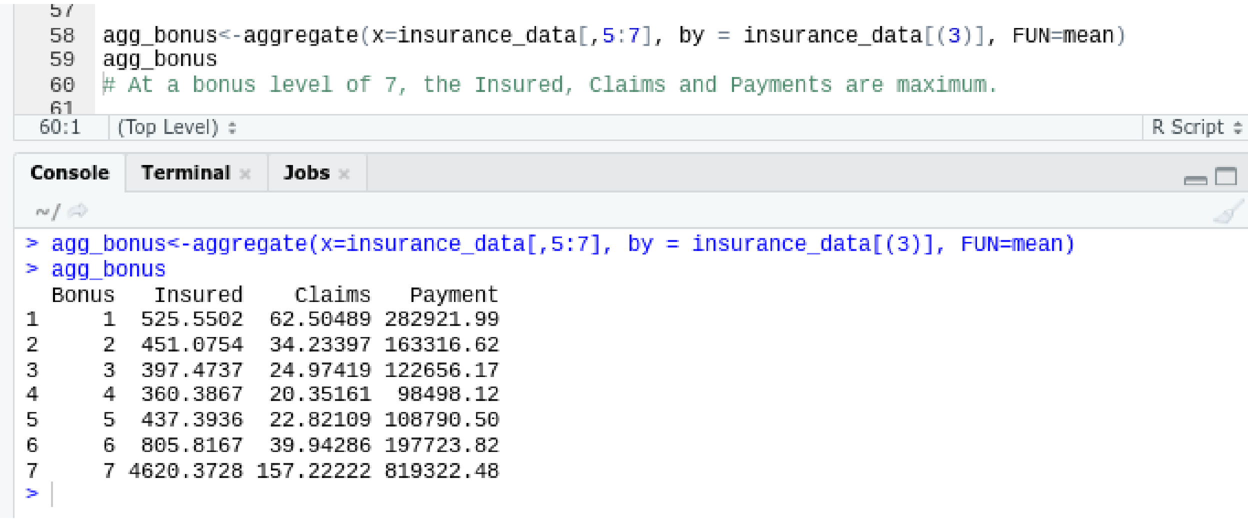
```r
# 4. The insurance company is planning to establish a new branch office, so they are inter
# to find at what location, kilometre, and bonus level their insured amount, claims, and
# payment gets increased.
agg_kilometer<-aggregate(x=insurance_data[,5:7], by=insurance_data[c(1)], FUN=mean)
agg_kilometer
# At a distance of < 1000 Kilometer, the number of Insured in policy-years is maximum, but
# the Claims and Payments are higher in the Kilometer range of 1000-15000.
```

```
> # 4. The insurance company is planning to establish a new branch office, so they are interest
ed
> # to find at what location, kilometre, and bonus level their insured amount, claims, and
> # payment gets increased.
> agg_kilometer<-aggregate(x=insurance_data[,5:7], by=insurance_data[c(1)], FUN=mean)
> agg_kilometer
  Kilometres   Insured   Claims   Payment
1          1 1837.8163 75.59453 361899.35
2          2 1824.0288 89.27664 442523.78
3          3 1081.9714 54.16100 272012.58
4          4  398.9632 20.79493 108213.41
5          5  284.9475 18.04215  93306.12
> 
```



```r
agg_location<-aggregate(x=insurance_data[,5:7], by = insurance_data[(2)], FUN=mean)
agg_location
# The Insured, the Claims and the Payments are maximum in the Zone 4, ie, the rural areas
# in Southern Sweden.
```

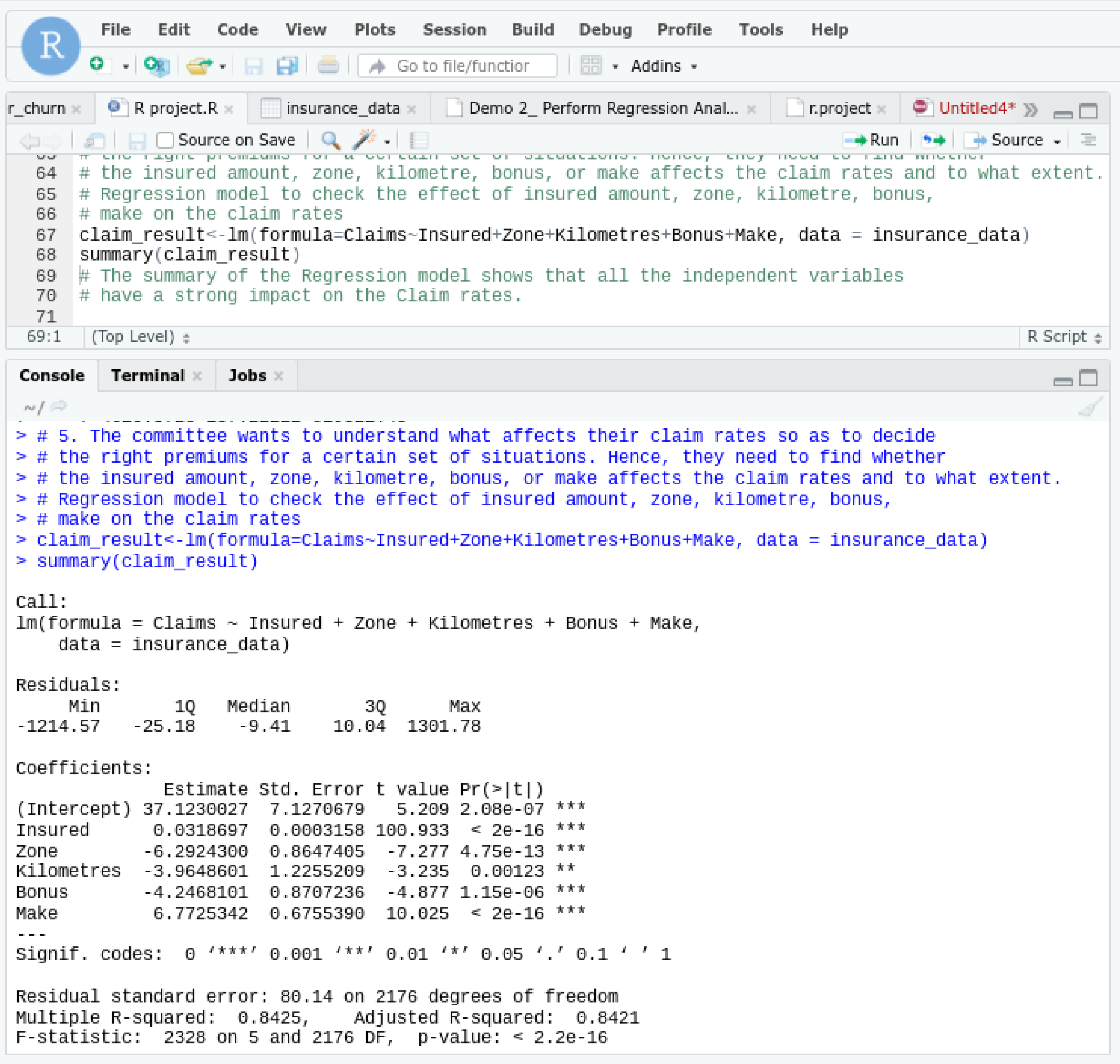
```
> agg_location<-aggregate(x=insurance_data[,5:7], by = insurance_data[(2)], FUN=mean)
> agg_location
  Zone    Insured     Claims   Payment
1    1 1036.17175  73.568254 338518.95
2    2 1231.48184  67.625397 319921.52
3    3 1362.95870  63.295238 307550.85
4    4 2689.38041 101.311111 537071.76
5    5  384.80188  19.047923  93001.84
6    6  802.68457  32.577778 175528.47
7    7   64.91071   2.108844   9948.19
> 
```



```
agg_bonus<-aggregate(x=insurance_data[,5:7], by = insurance_data[(3)], FUN=mean)
agg_bonus
# At a bonus level of 7, the Insured, Claims and Payments are maximum.
```

```
> agg_bonus<-aggregate(x=insurance_data[,5:7], by = insurance_data[(3)], FUN=mean)
> agg_bonus
  Bonus   Insured    Claims   Payment
1     1  525.5502  62.50489 282921.99
2     2  451.0754  34.23397 163316.62
3     3  397.4737  24.97419 122656.17
4     4  360.3867  20.35161  98498.12
5     5  437.3936  22.82109 108790.50
6     6  805.8167  39.94286 197723.82
7     7 4620.3728 157.22222 819322.48
> 
```

9. Regression model to check the effect of distance, location, insured, make and bonus on Claims.



```
# the insured amount, zone, kilometre, bonus, or make affects the claim rates and to what extent.
# Regression model to check the effect of insured amount, zone, kilometre, bonus,
# make on the claim rates
claim_result<-lm(formula=Claims~Insured+Zone+Kilometres+Bonus+Make, data = insurance_data)
summary(claim_result)
# The summary of the Regression model shows that all the independent variables
# have a strong impact on the Claim rates.
```

```
> # 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, kilometre, bonus, or make affects the claim rates and to what extent.
> # Regression model to check the effect of insured amount, zone, kilometre, bonus,
> # make on the claim rates
> claim_result<-lm(formula=Claims~Insured+Zone+Kilometres+Bonus+Make, data = insurance_data)
> summary(claim_result)

Call:
lm(formula = Claims ~ Insured + Zone + Kilometres + Bonus + Make, 
    data = insurance_data)

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 ***
Insured      0.0318697  0.0003158 100.933  < 2e-16 ***
Zone        -6.2924300  0.8647405  -7.277 4.75e-13 ***
Kilometres  -3.9648601  1.2255209  -3.235  0.00123 ** 
Bonus       -4.2468101  0.8707236  -4.877 1.15e-06 ***
Make         6.7725342  0.6755390  10.025  < 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
```

Analysis:

- The data consists of 2182 observations of 7 variables.
- Correlation of Claims with Payment = 99.54%. They are positively correlated. Correlation of Insured with Payment = 93.32%. They are positively correlated. R squared value of 0.99 indicates that 99% of variation in the dependent variable is explained by the variation of the independent variable.
- From the summary of the Regression model, we can conclude that distance, location, Insured and Claims have very strong significance with the dependent variable, Payment. p-value of Kilometers = 1.18e-05, Zone = 0.027, Insured < 2e-16, Claims < 2e-16. ‘ Bonus’ and ‘Make’ have p-value>0.05. Hence they are not significant. Adjusted R squared is 99.52%.

- At a distance of < 1000 Kilometer, the number of Insured in policy-years is maximum, but Claims and Payments are higher in the Kilometer range of 1000-15000.

- The Insured, the Claims and the Payments are maximum in Zone 4, ie, the rural areas in Southern Sweden.

- At a bonus level of 7, the Insured, Claims and Payments are maximum.

- The summary of the Regression model shows that the independent variables, Insured, Zone, Kilometers, Bonus and Make have a strong impact on the Claim rates. The adjusted R squared is 84.21%.