

Analyze the report of Swedish Motor Insurance

Business Analytic Foundation with R Tools- Solutions



Solutions

Disclaimer: In Business Analytics, there are different ways of solving the same set of problems, we are just presenting one. Feel free to explore other ways of answering these questions.

1. The committee is interested to know each field of the data collected through descriptive analysis to gain basic insights into the dataset and to prepare for further analysis.

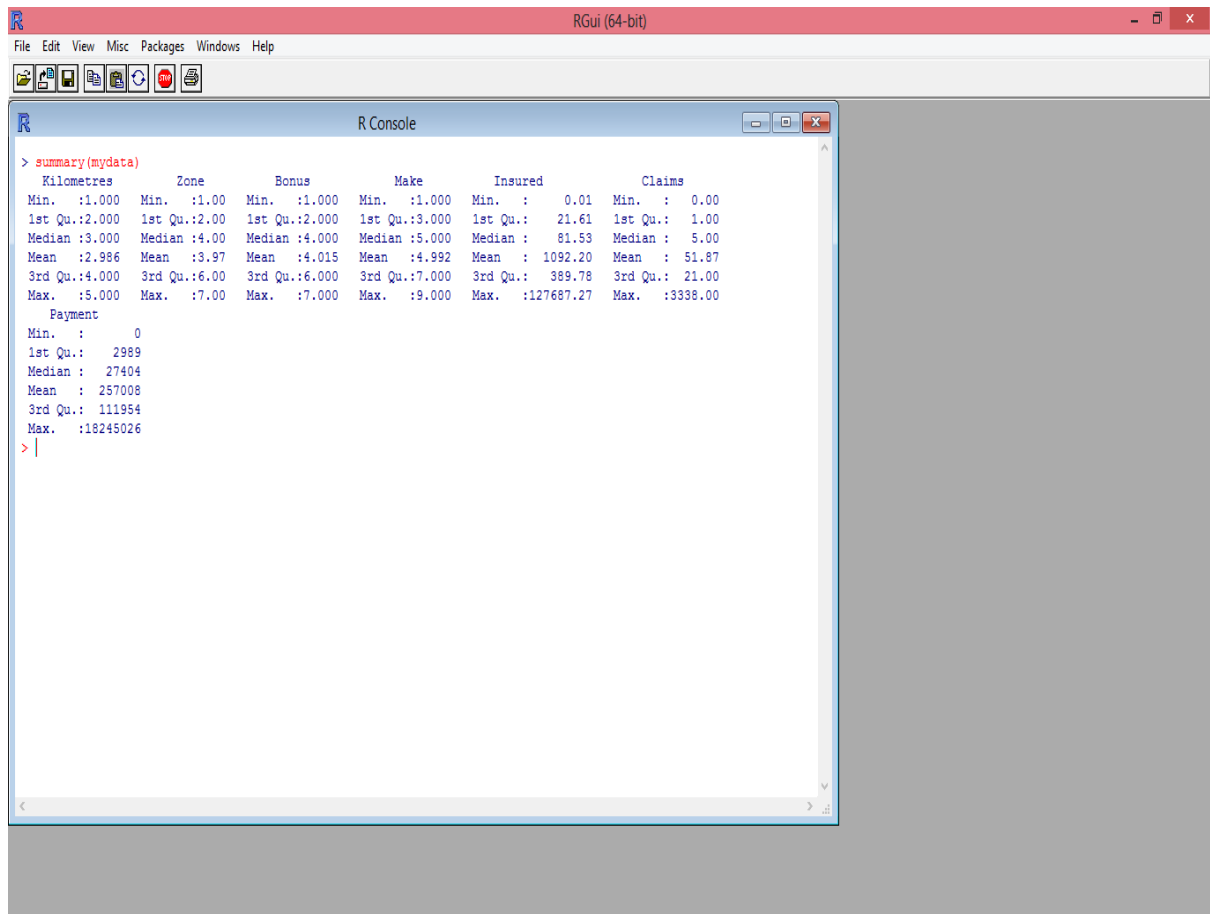
Since the dataset is large, we need to find a pattern of the data. In order to find the distribution of data, perform the summary statistics.

Code:

```
Summary(mydata)
```

Result:

The results provide the minimum and maximum values. It also provides the mean and median values of all variables. From this you can understand the spread of data. We can see that claims and payment also have null or zero values, however the insured column does not have a zero value. This specifies that there are few entries where the car has been insured for a given period of time. However, no claim or payment has been made for that combination of car make, zone, and kilometres.



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

      Payment
Min.    :  0
1st Qu.: 2989
Median : 27404
Mean   : 257008
3rd Qu.: 111954
Max.   :18245026
> |
```

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.

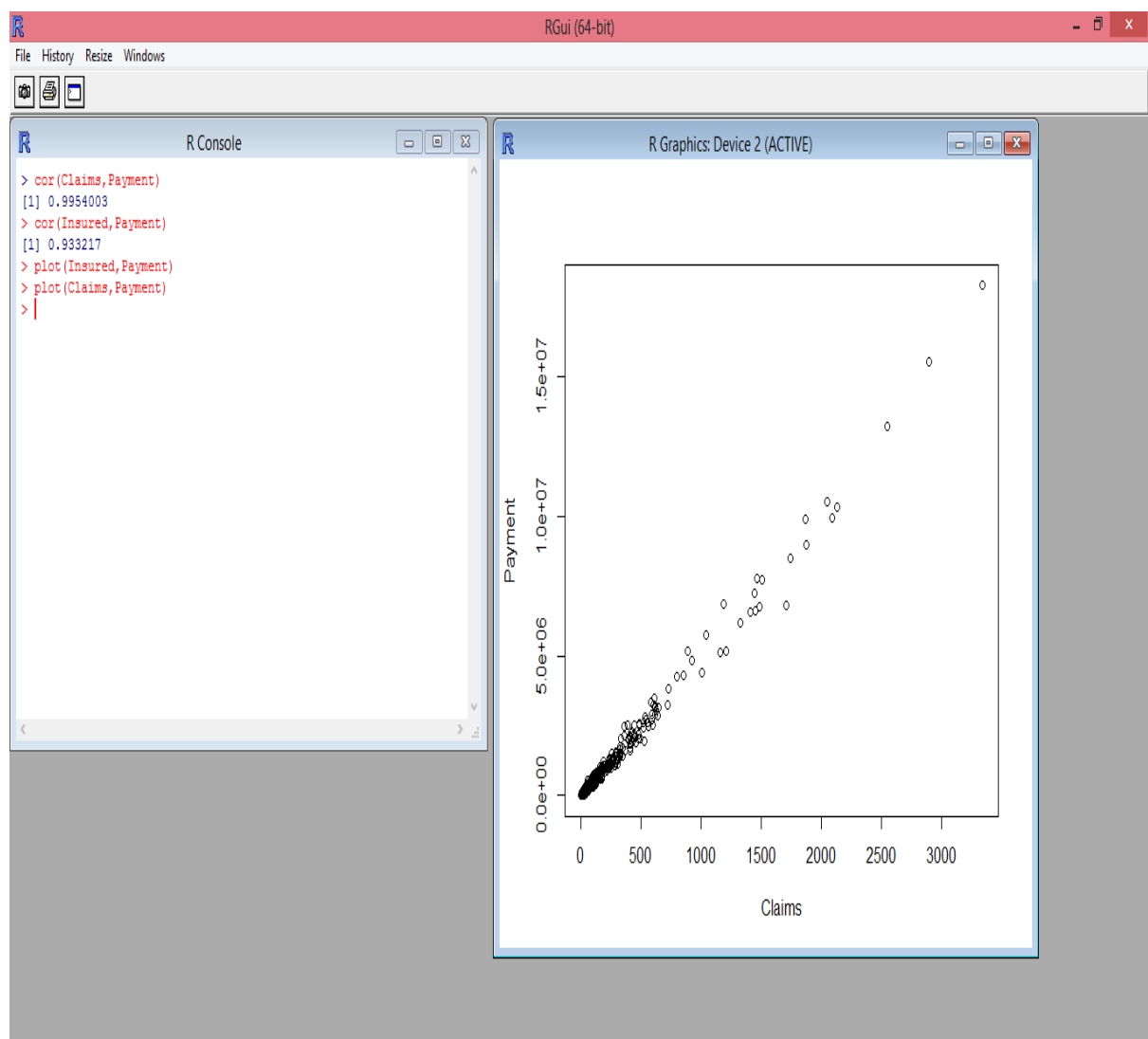
In order to find the relationship of *insured* and *claims* variables with *payment* variable, perform correlation function. This would help us in finding the relationship—whether it is positively or negatively related. To view a graphical representation, perform a scatter plot.

Code:

```
cor(Claims,Payment)
cor(Insured,Payment)
plot(Insured,Payment)
plot(Claims,Payment)
```

Result:

The results show that *claims* is 99 percent positively correlated with *payment* and *insured* is 93 percent positively correlated with *payment*. The scatter plot shows that the relationship between the variables are strong as there is a linear trend in the graph, that is, as the value of claims increases, the payment value also increases and the same trend will occur for the insured and the 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 them are affecting it.

In order to find the impact of all the variables on the payment variable, build a linear regression model.

Independent variable: insured, claims, make, bonus, zone, and kilometers

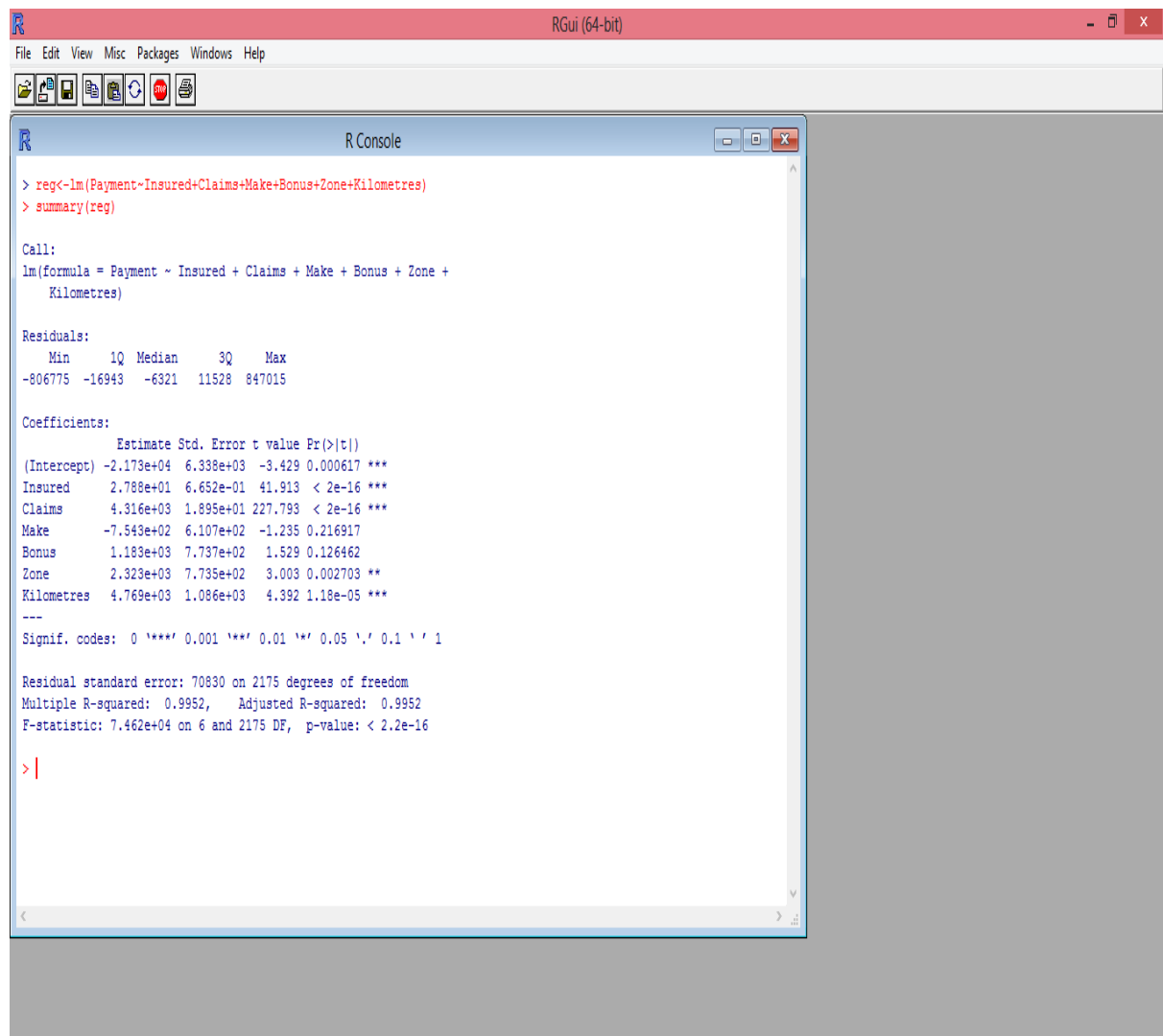
Dependent variable: payment

Code:

```
reg<-lm(Payment~Insured+Claims+Make+Bonus+Zone+Kilometres)
```

Result:

The result shows the intercept value and estimated values of all independent variables. From this we can derive the regression line and this would help us in predicting the future payment values. The high p-value of the make and bonus show that they do not make much impact on payment, as compared to all other variables.



```

RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

> reg<-lm(Payment~Insured+Claims+Make+Bonus+Zone+Kilometres)
> summary(reg)

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

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 ***
Insured      2.788e+01  6.652e-01  41.913 < 2e-16 ***
Claims       4.316e+03  1.895e+01  227.793 < 2e-16 ***
Make        -7.543e+02  6.107e+02  -1.235 0.216917
Bonus        1.183e+03  7.737e+02   1.529 0.126462
Zone         2.323e+03  7.735e+02   3.003 0.002703 **
Kilometres   4.769e+03  1.086e+03   4.392 1.18e-05 ***
---
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

> |

```

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

In order to find the mean value of insured, payment, and claims based on zone, kilometre, and bonus variables, group all the result variables based on individual categorical variables.

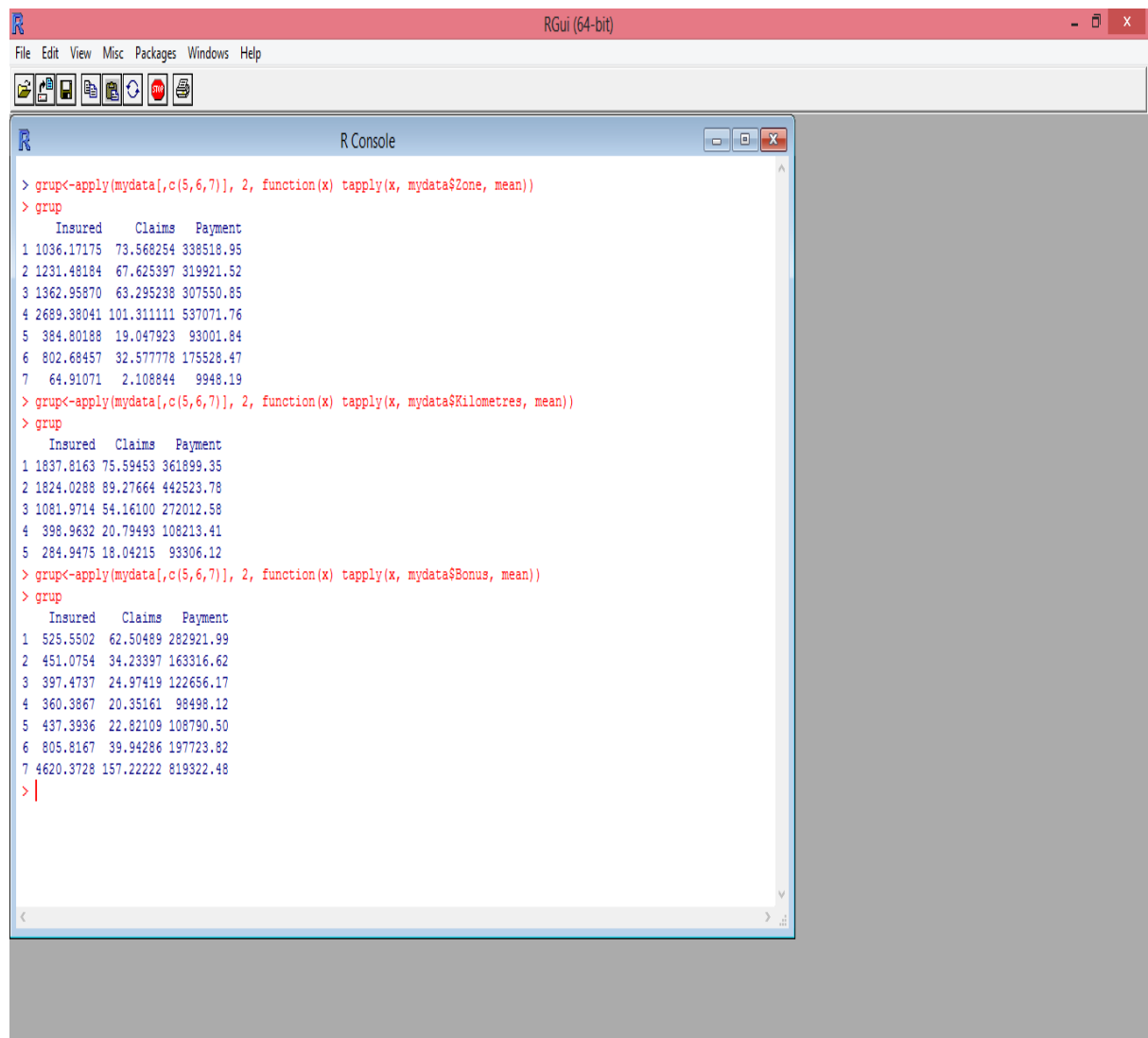
Code:

```
grup<-apply(mydata[,c(5,6,7)], 2, function(x) tapply(x, mydata$Zone,
mean))
grup<-apply(mydata[,c(5,6,7)], 2, function(x) tapply(x, mydata$Kilometres,
mean))
grup<-apply(mydata[,c(5,6,7)], 2, function(x) tapply(x, mydata$Bonus,
mean))
```


Result:

The following observations can be made from the results:

- Zone 4 has the highest number of claims, and thus payment as well.
- Zones 1-4 have more insured years, claims, and payments.
- Kilometer group 2 has the maximum payments. Though the insured number of years is lesser than kilometre 1, the claims and payments are higher for group 2.
- There is not much variation in groups of bonus except for 7 with unusually high number of insured years, claims, and payments.



```

RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

> grup<-apply(mydata[,c(5,6,7)], 2, function(x) tapply(x, mydata$Zone, mean))
> grup
      Insured   Claims  Payment
1 1036.17175  73.568254 338518.95
2 1231.48184  67.625397 319921.52
3 1362.95870  63.295238 307550.85
4 2689.38041 101.311111 537071.76
5  384.80188  19.047923  93001.84
6  802.68457  32.577778 175528.47
7   64.91071   2.108844   9948.19
> grup<-apply(mydata[,c(5,6,7)], 2, function(x) tapply(x, mydata$Kilometres, mean))
> grup
      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
> grup<-apply(mydata[,c(5,6,7)], 2, function(x) tapply(x, mydata$Bonus, mean))
> grup
      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
>

```

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.

In order to find the dependency of claim variable by other variables build a linear regression model.

Code:

```
reg<-lm(Claims~Kilometres+Zone+Bonus+Make+Insured)
summary(reg)
```

Dependent variable: claims

Independent variable: kilometres, zone, bonus, make, and insured

Result:

The results provides the intercept and estimated value and this in turn shows that all the p values of independent variables, such as kilometres, zone, bonus, make, and insured are highly significant and are making an impact on the claims.

The screenshot displays the RGui (64-bit) interface. The R Console window shows the execution of the following commands:

```
> View(mydata)
> attach(mydata)
> reg<-lm(Claims~Kilometres+Zone+Bonus+Make+Insured)
> summary(reg)
```

The output of the `summary(reg)` command is as follows:

```
Call:
lm(formula = Claims ~ Kilometres + Zone + Bonus + Make + 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 ***
Kilometres  -3.9648601   1.2255209  -3.235  0.00123 **
Zone        -6.2924300   0.8647405  -7.277  4.75e-13 ***
Bonus       -4.2468101   0.8707236  -4.877  1.15e-06 ***
Make         6.7725342   0.6755390  10.025 < 2e-16 ***
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
```

The Data: mydata window displays the following table:

	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
13	1	1	2	4	782.58	48	242894
14	1	1	2	5	115.43	11	23545
15	1	1	2	6	338.06	23	39598
16	1	1	2	7	70.44	7	48767
17	1	1	2	8	15.25	2	6560
18	1	1	2	9	6416.19	638	2873487
19	1	1	3	1	309.98	24	134931