simplilearn
your pace, your place

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.

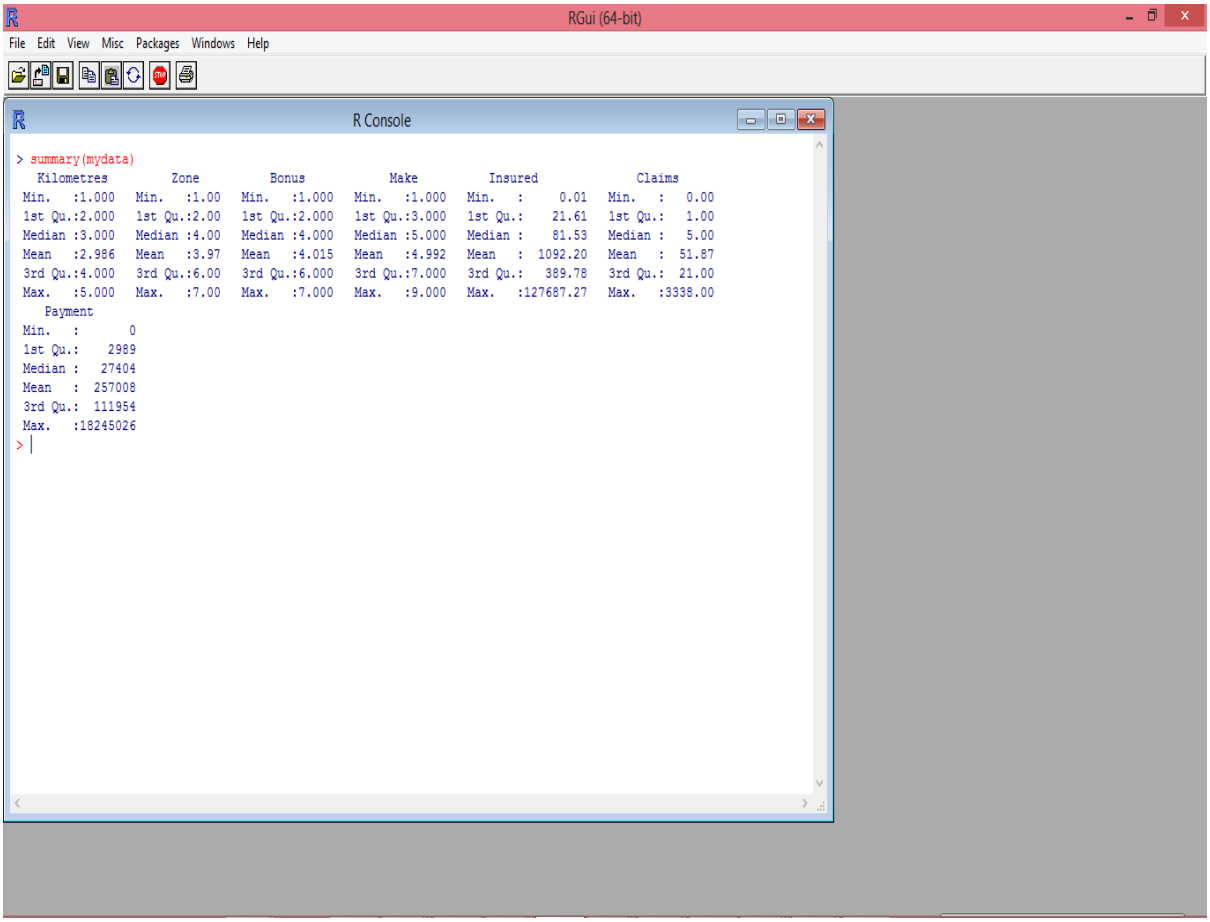
   Since the dataset is large, we need to find a pattern of the data. In order to find the distribution of data, perform the summary statistics.

   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.



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.

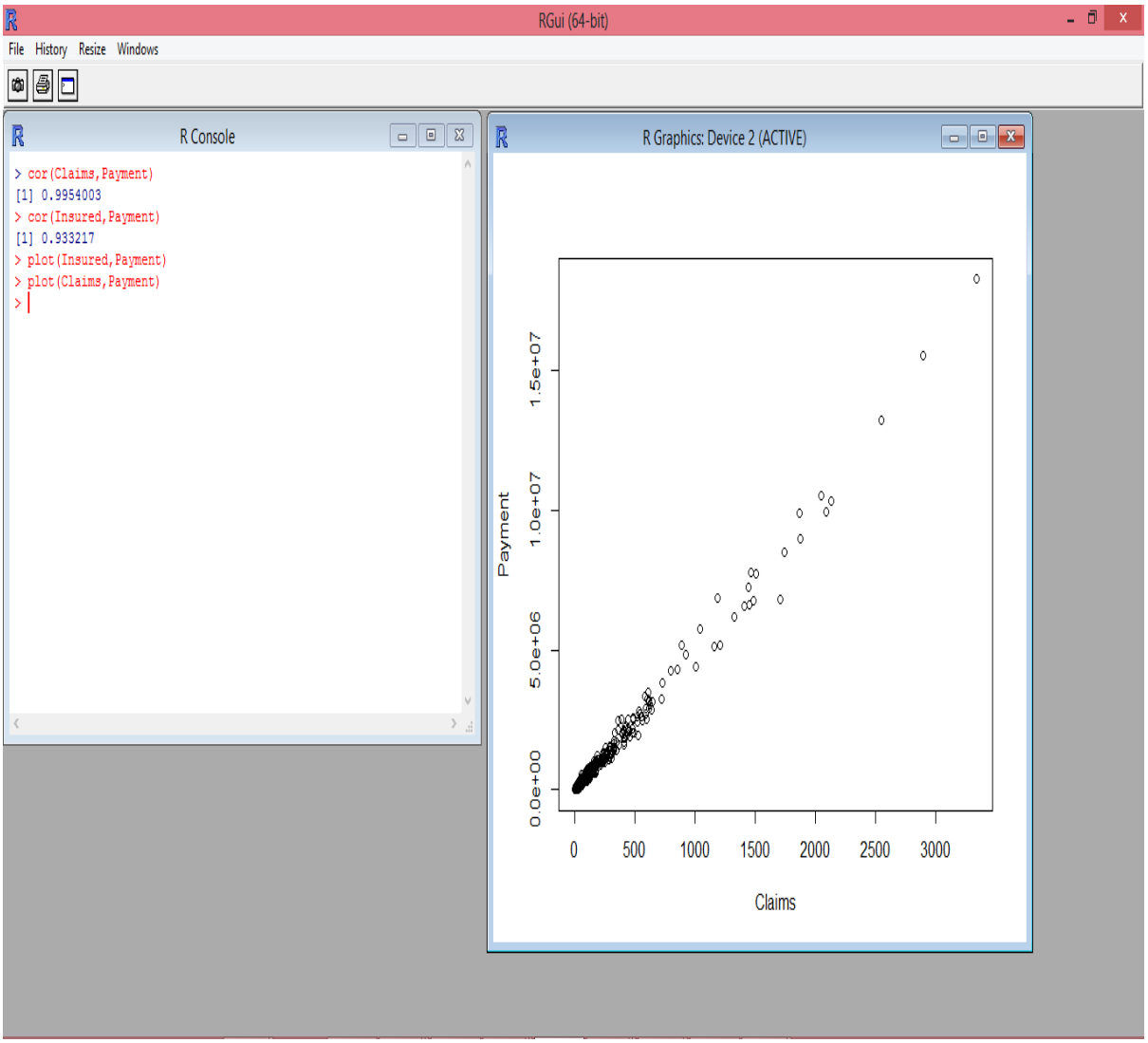
   In order to find the relationship of *insured* and *claims* variables with *payment* variable, perform correlation function. This would help us in finding the relationship—whether it is positively or negatively related. To view a graphical representation, perform a scatter plot.

   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.

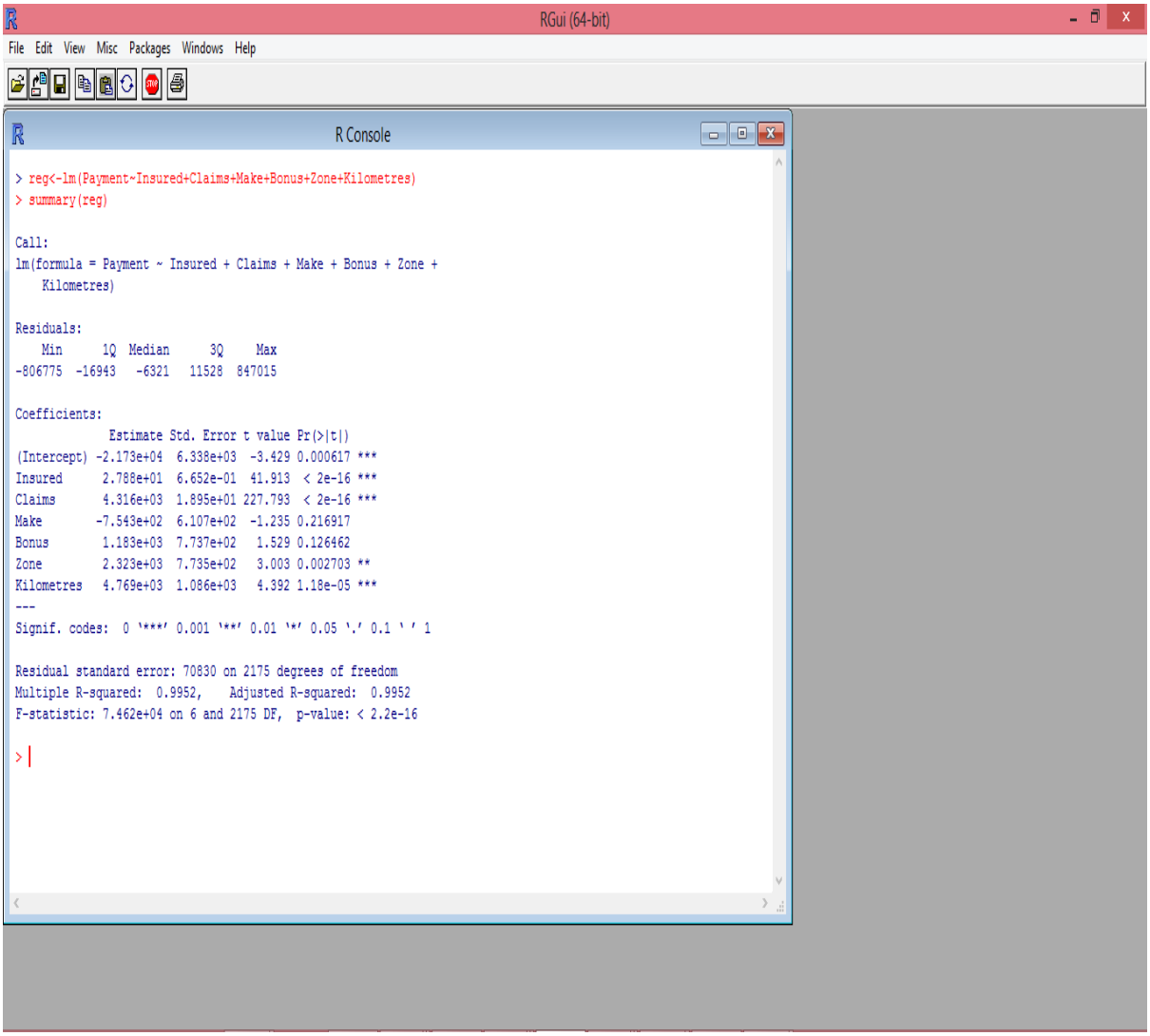
   Independent variable: insured, claims, make, bonus, zone, and kilometers
   Dependent variable: payment

   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.



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.

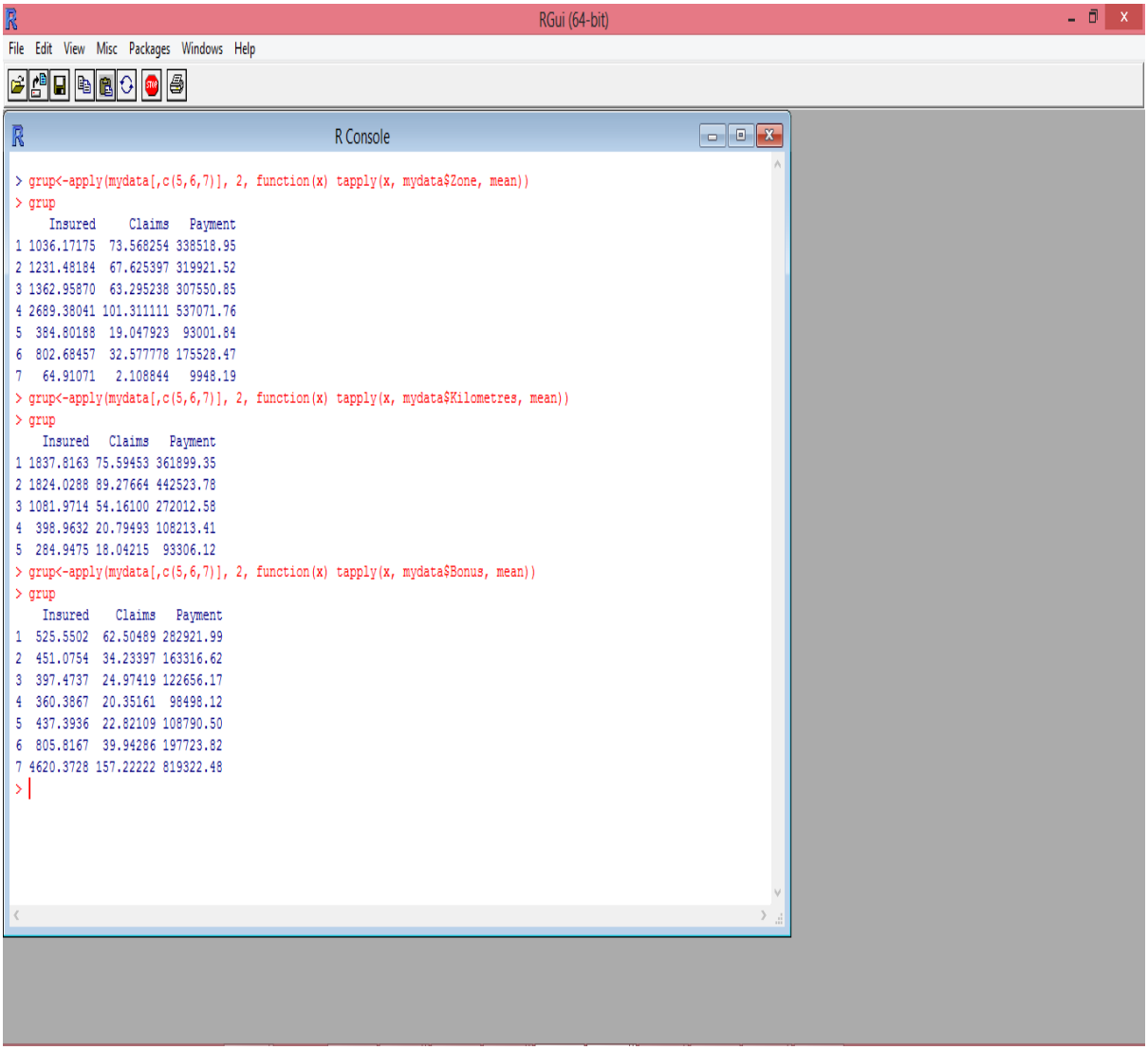
   Code:

   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:

a. Zone 4 has the highest number of claims, and thus payment as well.
b. Zones 1-4 have more insured years, claims, and payments.
c. 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.
d. There is not much variation in groups of bonus except for 7 with unusually high number of insured years, claims, and payments.



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.

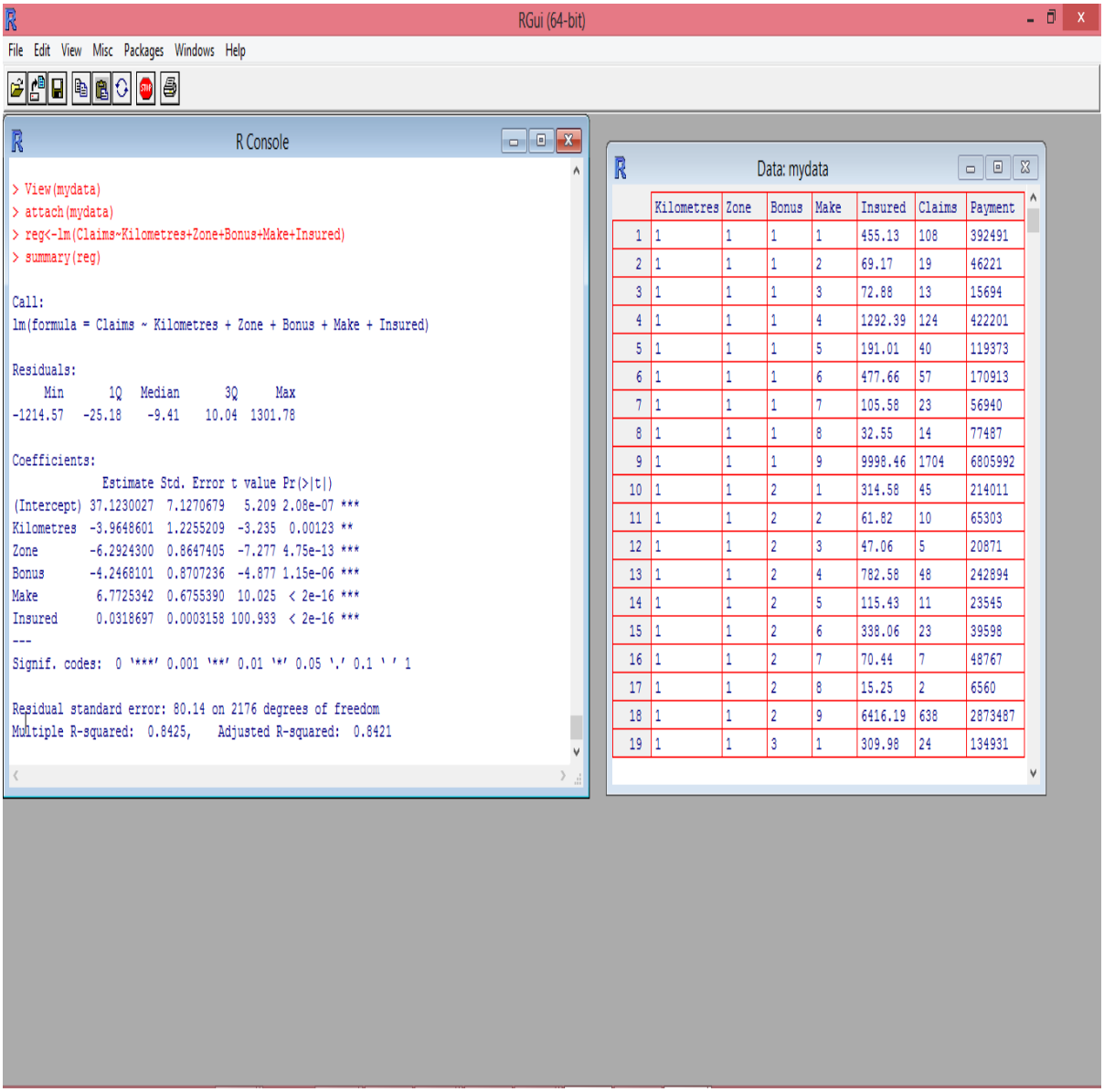
   In order to find the dependency of claim variable by other variables build a linear regression model.

   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.



	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