# Analysis of Insurance Factor Identification report of Swedish Motor Insurance

**Using R-Tools**

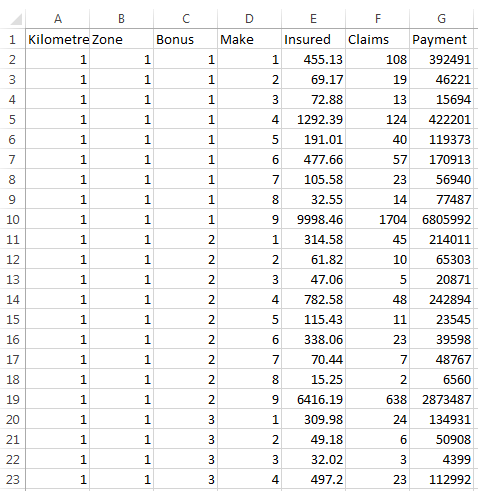
## Problem Statement:

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

Here we show the proper description of the data “Insurance\_factor\_identification” in tabulation form as follows and renamed it as “insurance” in R for analysis.

|  |  |
| --- | --- |
| Variable | Description |
| Kilometre | Kilometres travelled per year  1: <1000 2: 1000-15000 3: 15000-20000 4: 20000-25000 5: >25000 |
| Zone | Geographical zone  1: Stockholm, Goteborg, and Malmo 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 |
| Payments | Total value of payments in SKR (Swedish Krona) |

Dataset :-



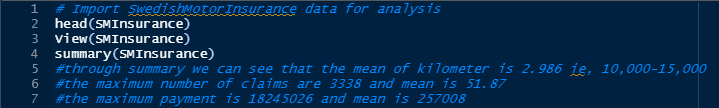
The data\_set is “Insurance\_factor\_identification” data consisting of 2182 observations of 7 variables i.e the dimension of the dataset is 2182x7.

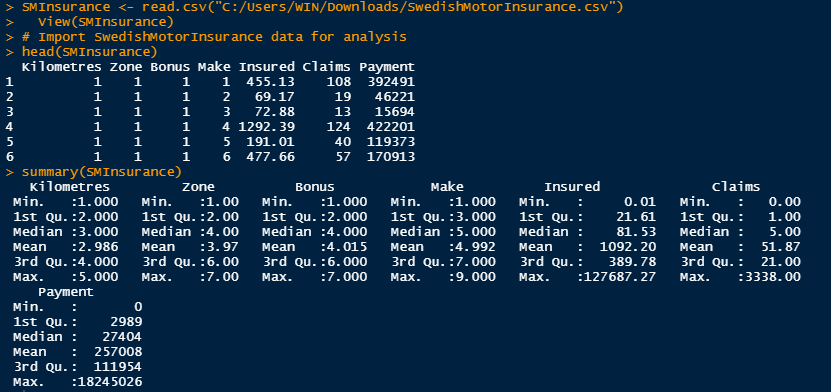
1. **The committee is interested to know each field of the data collected through descriptive analysis to gain basic insights into the data set and to prepare for further analysis.**

Solution:

We start by importing the dataset and find out the head, summary of the data we’ll be working on.

Code :-





Observation:-

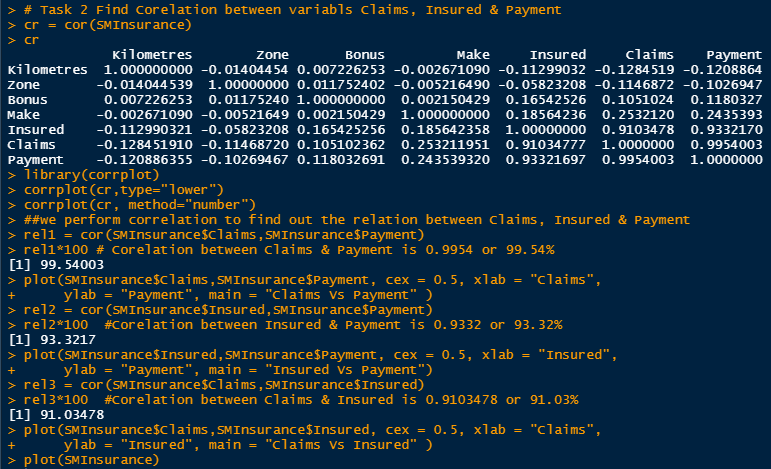
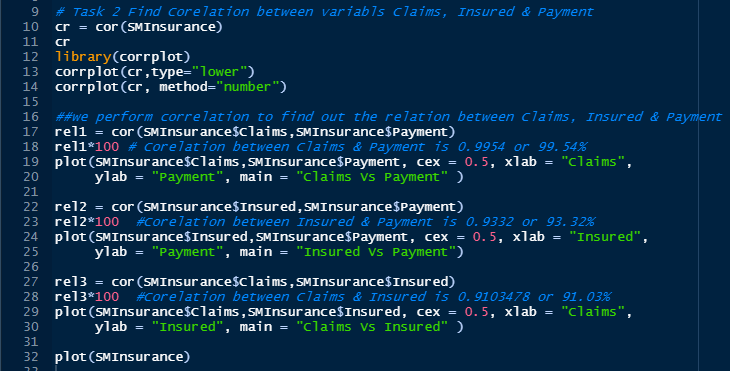
* The result gives us the median, mean, quartile, max and min values. Through summary we can see that the mean of kilometre is 2.986 ie, 10,000-15,000
* The maximum number of claims are 3338 and mean is 51.87 .The maximum payment is 18245026 and mean is 257008.
* We can conclude that **there is some Cars which are insured but there has been no insurance claimed during proceeding years.** Minimum number of cars 0.01 which are insured but have Zero claims so Zero payments.

1. **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.**

# Solution :-

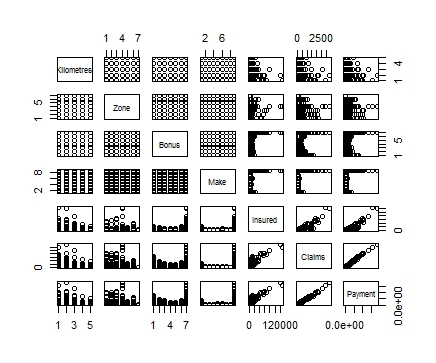
As per problem statement "Payment” is important factor so it is dependent variable. We have to find corelation between other two Independent variables “Claims” and “Number of Insured Policy years”.

Code :-



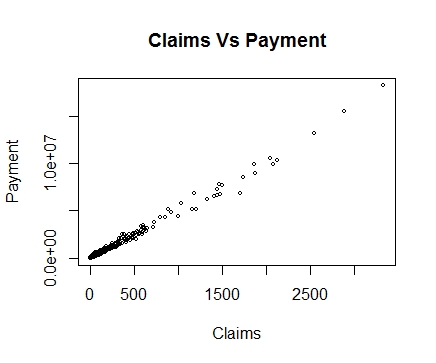
**Visualizations:-**

|  |
| --- |
| Visualization shows Scatter plot of Corelation between all variables.  Graph between Claim Vs Payment & Insured Vs Payment is Linear |



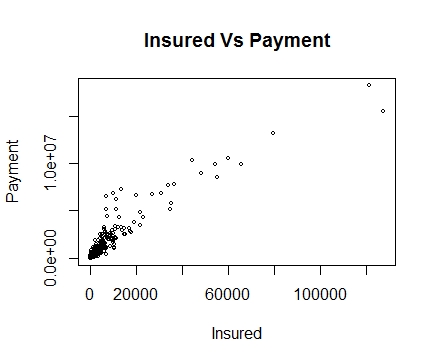
* This Corelation chart shows Corelation between all variables.

1. Graph between Claims Vs Payment



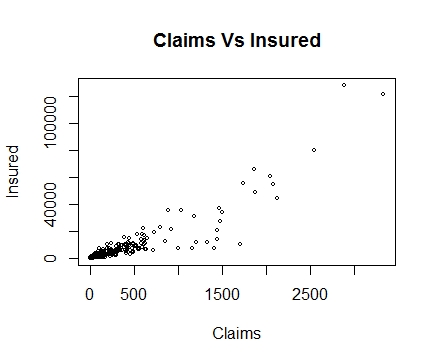
* There is highly correlation between Claims & Payment i.e 99.54% . also graph is showing linearity.

(II) Graph between Insured Vs Payment

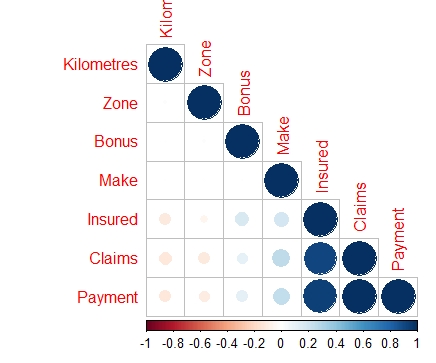


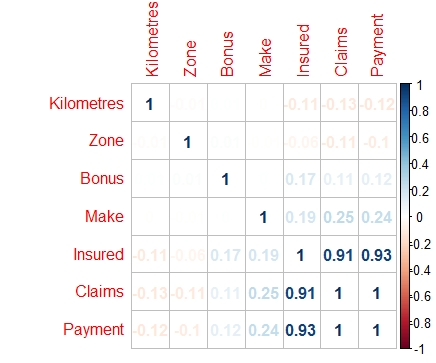
* There is highly correlation between Insured & Payment i.e 93.32% . also graph is showing linearity

(III) Graph between Claims Vs Insured



* There is highly correlation between Claims & Insured i.e 91.03% . also graph is showing linearity.

**For better visualization of corelation :-**

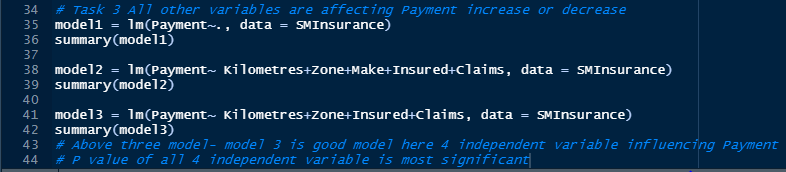
****

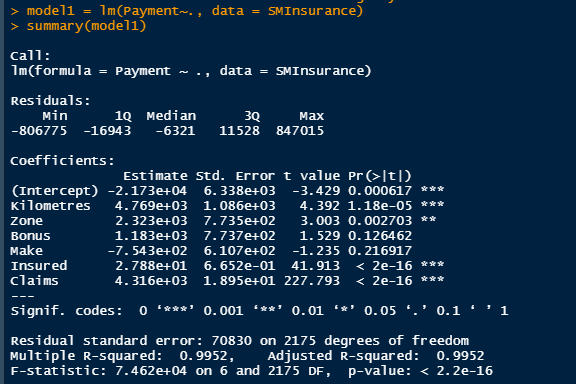
# The Payment factor is highly correlated to Claims and the number of Insured Policy Years so we conclude that as the Claims and the No of policy year increases the payment tends to increase consistently.

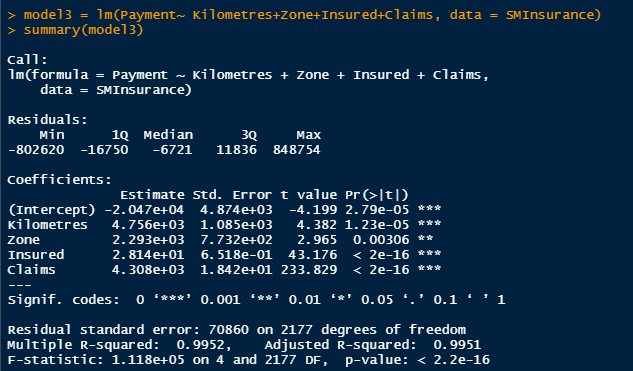
* Also from Plot it is seen that the Dark Blue colour denotes the high degree of correlation between Insured vs Payments (93.32%) and Claim vs Payments (99.54 approx equals to ‘1’).

1. **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.**

**Code :-**

****

****

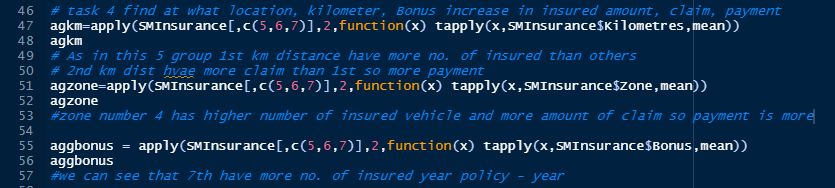


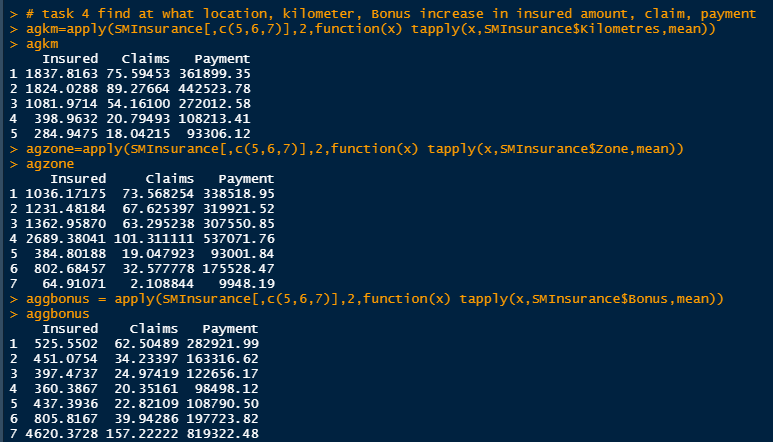
Observation:-

* As per linear regression model3 is best fit model which indicate that there are 4 independent variables (Kilometer, Zone, Insured, Claims) which is highly affecting dependent variable Payment. It has low P-value which indicate that these 4 variables are highly significant. With increasing or decreasing of these 4 variables subsequently increase or decrease in Payment.
* Other Two independent variable Make & Bonus not affecting that much to Payment having high P-Value.
* The R-Squared value of model3 is very good 0.9952 nearly 1 also difference R-Squared & Adjusted R-Squared is 0.001 nearly Zero which indicate model is Best fit model.

1. **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 get increased. (Hint: Aggregate Dt)**

**Code :-**



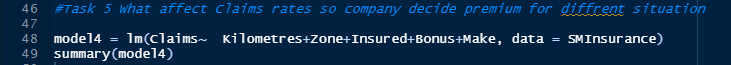
****

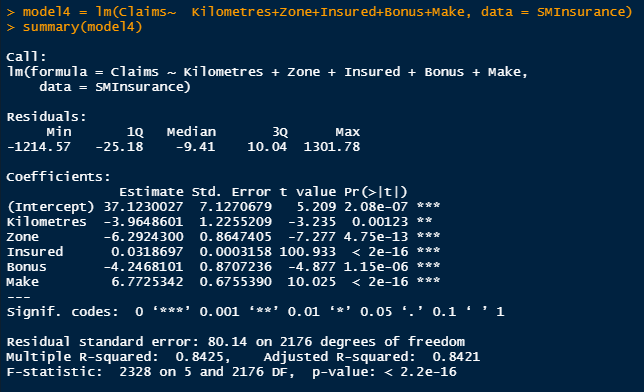
Observation:-

* We find mean value of Insured, Claims & Payments based on kilometer, Zone and Bonus group all variables.
* Group 1 in Km have more Insured than Group 2 also have lower claims and payments than Group 2. So, from range 1000 to 15000 km Claim and payment both are increased.
* We can see based on Zone wise first 4 groups have more insured year but group 4 have more no. of Insured highest Claim so more Payments.
* In group of Bonus group 7 have highest no. of Insured & more payment also.

1. **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.**

**Code :-**





Observation:-

* We build model4 by Regression Analysis which indicate that dependent variable “Claim rate” is affected by 4 independent variables “Kilometer, Zone, Insured, Bonus, Make”
* We conclude that from summary these 5 Independent variables have low P-Value these variable highly significant & having R-Squared value 0.8425 and Adjusted R-Squared value 0.8421 difference is 0.0004 close to zero.
* **So, these independent variables i) Kilometer ii) Zone iii) Insured iv) Bonus v) Make is affect “Claim rate” by 84.25%**