**Car Insurance Data Analysis**

**Data Overview**

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.

**Dataset Variables**

**Kilometres -** Kilometres travelled per year

1: < 1000

2: 1000-15000

3: 15000-20000

4: 20000-25000

5: > 25000

**Bonus -** No claims bonus; equal to the number of years, plus one, since the last claim

**Zone -** Geographical zones

1: Stockholm, Göteborg, 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

**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 Swedish Krona.

**Objective-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.**

**Code & Output::**

setwd("C:/Insurance\_Data\_Analysis")

DS <- read.csv("SwedishMotorInsurance.csv")

View(DS)

|  |
| --- |
| 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 |
|  |
| |  | | --- | | **Objective-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.** | |

**Code & Output::**

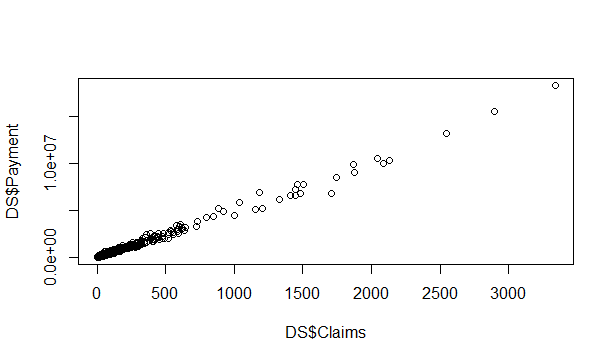
cor(DS$Claims,DS$Payment)

> cor(DS$Claims,DS$Payment)

[1] 0.9954003

* Value is highly significant

plot(DS$Insured,DS$Payment)



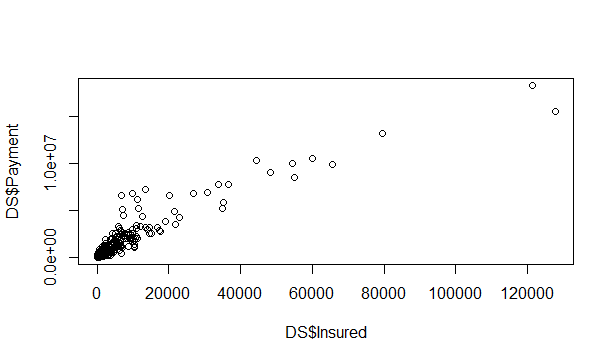
cor(DS$Insured,DS$Payment)

> cor(DS$Insured,DS$Payment)

[1] 0.933217

* Value is highly significant

plot(DS$Insured,DS$Payment)



**Objective-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.**

**Code & Output::**

lineModel <- lm(Payment ~ ., data = DS)

summary(lineModel)

> lineModel <- lm(Payment ~ ., data = DS)

> summary(lineModel)

Call:

lm(formula = Payment ~ ., data = DS)

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

* This shows that except Bonus and Make all other variables are significant and are influencing the payment.
* I used backward elimination method to build our final liner model.

*Payment ~ Claims + Insured + Zone + Kilometres*

**Objective-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 get increased.**

**Code & Output::**

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

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

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

> ZoneResult <- apply(DS[,c(5,6,7)],2, function(x)tapply(x, DS$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(DS[,c(5,6,7)],2, function(x)tapply(x, DS$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(DS[,c(5,6,7)],2, function(x)tapply(x, DS$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

**Findings –**

1. Zone 4 has the highest value for Insured, Claims and Payments.
2. Kilometre group 1 and 2 has the highest value for Insured, Claims and Payments.
3. Bonus group 7 has the highest value for Insured, Claims and Payments.

**Objective-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.**

**Code & Output::**

Model1 <- lm(Claims ~ Kilometres + Zone + Bonus + Make + Insured, data=DS)

summary(Model1)

> Model1 <- lm(Claims ~ Kilometres + Zone + Bonus + Make + Insured, data=DS)

>

> summary(Model1)

Call:

lm(formula = Claims ~ Kilometres + Zone + Bonus + Make + Insured,

data = DS)

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

F-statistic: 2328 on 5 and 2176 DF, p-value: < 2.2e-16