Insurance factors identification

Andrei Enescu

10/21/2020

Description

Background and Objective:

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.

Domain: Insurance

Dataset Description:

The insurance dataset holds 7 variables and the description of these variables are given below:

Attribute Description

Kilometers: Kilometers traveled per year
1: $<$ 1000
2: 1000-15000
3: 15000-20000
4: 20000-25000
5: $>$ 25000
Zone: Geographical zone
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
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: The number of insured in policy-years.

Claims: Number of claims
Payment: The total value of payments in Skr (Swedish Krona)

Analysis Tasks:

After understanding the data, you need to help the committee with the following by the use of the R tool:

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.

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.

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.

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. (Hint: Aggregate Dataset)

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.

Used libraries:

```
library(rio)
library(ggplot2)
```

Importing the data using RIO package

```
InsuranceDF <- import("Insurance_factor_identification.csv")
```

Checking data

```
head(InsuranceDF)
```

```
##   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
```

```
str(InsuranceDF)
```

```
## '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 ...
```

```
which(is.na(InsuranceDF))
```

```
## integer(0)
```

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.

```
str(InsuranceDF)
```

```
## '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 ...
```

```
summary(InsuranceDF)
```

```
##    Kilometres         Zone          Bonus            Make
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:3.000
##  Median :3.000   Median :4.00   Median :4.000   Median :5.000
##  Mean   :2.986   Mean   :3.97   Mean   :4.015   Mean   :4.992
##  3rd Qu.:4.000   3rd Qu.:6.00   3rd Qu.:6.000   3rd Qu.:7.000
##  Max.   :5.000   Max.   :7.00   Max.   :7.000   Max.   :9.000
##     Insured             Claims           Payment
##  Min.   :     0.01   Min.   :   0.00   Min.   :       0
##  1st Qu.:    21.61   1st Qu.:   1.00   1st Qu.:    2989
##  Median :    81.53   Median :   5.00   Median :   27404
##  Mean   :  1092.20   Mean   :  51.87   Mean   :  257008
##  3rd Qu.:   389.78   3rd Qu.:  21.00   3rd Qu.:  111954
##  Max.   :127687.27   Max.   :3338.00   Max.   :18245026
```

Observation

On the output we have a sumary of each column from the dataset.

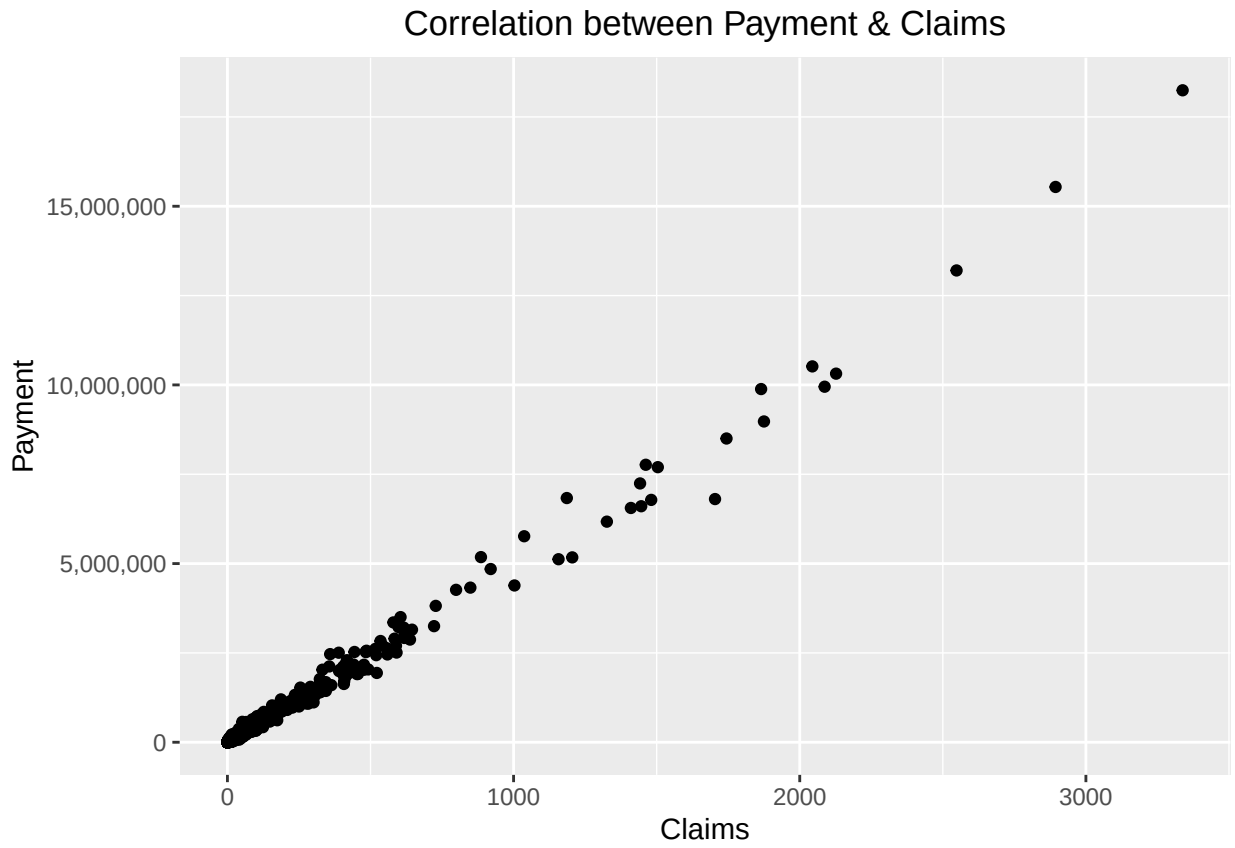
2. The total value of payment by an insurance company is an important factor to be monitored. So the committee has decided to find whether this payment is related to the number of claims and the number of insured policy years. They also want to visualize the results for better understanding.

Correlation between Payment & Claims

```
cor(InsuranceDF$Payment, InsuranceDF$Claims)
```

```
## [1] 0.9954003
```

```
ggplot(InsuranceDF,
       aes(x = Claims, y = Payment, label = Payment)) +
  geom_point(size = 1.5) +
  labs(title = "Correlation between Payment & Claims",
       x = "Claims",
       y = "Payment") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = scales::comma)
```

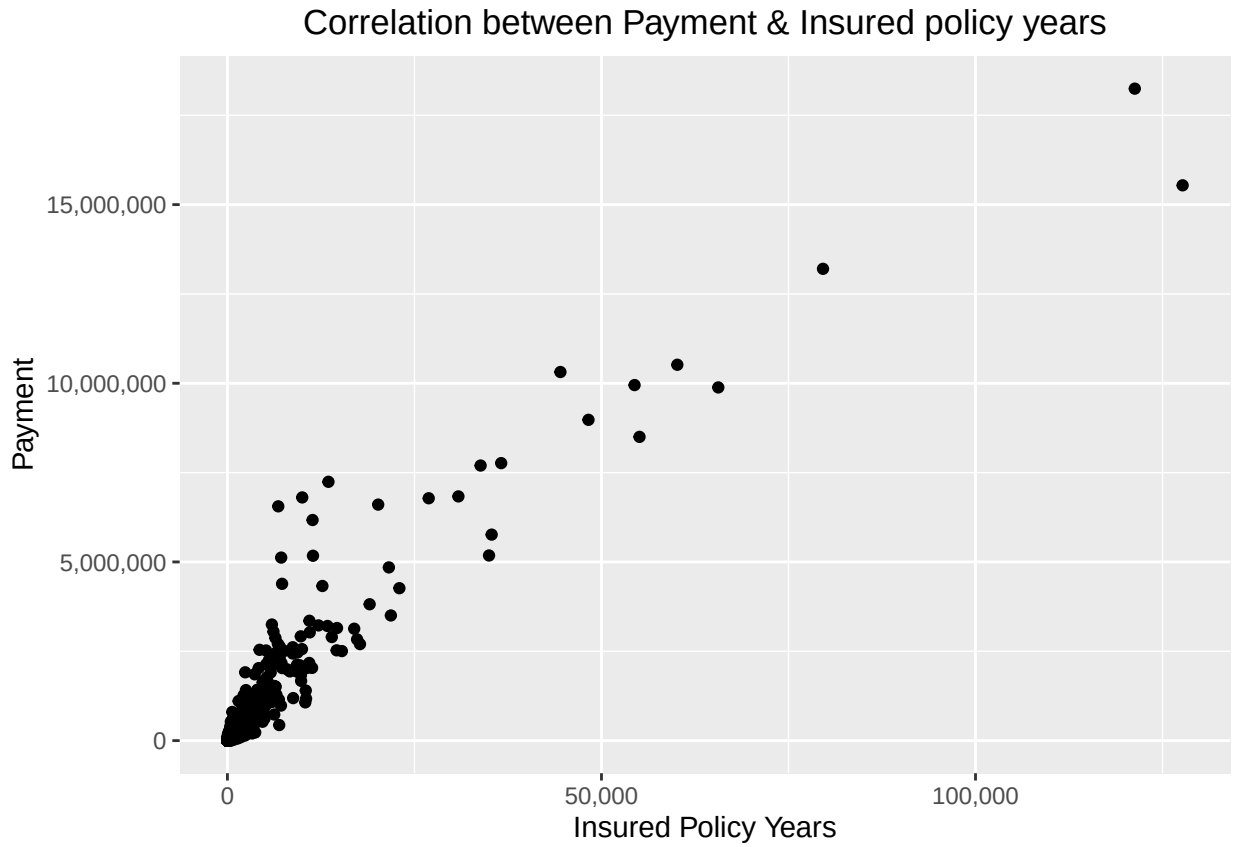


Correlation between Payment & No of insured policy years

```r
cor(InsuranceDF$Payment, InsuranceDF$Insured)
```

```
## [1] 0.933217
```

```r
ggplot(InsuranceDF,
       aes(x = Insured, y = Payment, label = Payment)) +
  geom_point(size = 1.5) +
  labs(title = "Correlation between Payment & Insured policy years",
       x = "Insured Policy Years",
       y = "Payment") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = scales::comma) +
  scale_x_continuous(labels = scales::comma)
```



Observation

From the correlation output we can observe that there is a strong positive correlation between Payment and Claims (99.5%) and also between Payment and Insured policy years (93.3%)
From both plots we observe that as Claims or Insured policy years increase, the Payment increases also

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.

```
fit1 <- lm(Payment ~ ., InsuranceDF)
summary(fit1)
```

```
##
## Call:
## lm(formula = Payment ~ ., data = InsuranceDF)
##
## 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
```

Observation

From the output we conclude the following:
* Kilometers, Zone, Insured and Claims are all significant variables and are affecting the Payment
* Make and Bonus, both have a higher P-value so they are affecting Payment in a lower manner

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.

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

```
##      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
```

```
KM <-
  apply(InsuranceDF[, c(5, 6, 7)], 2, function(x)
    tapply(x, InsuranceDF$Kilometres, mean))
KM
```

```
##     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
```

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

```
##     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
```

Observation

From the above outputs we can conclude:
* Zone 4 has the highest Claims and Payment
* Zone 7 has the lowest Insured, Claims and Payment
* Kilometer group 2 has the highest Claims and Payment
* Kilometer group 5 has the lowest Claims and Payment
* Bonus group 7 has the highest Insured, Claims and Payment

So for a new branch office, in order to have an increase in Payment they need to select Zone 4, Kilometers traveled per year: 1000 - 15000 (group 2) and seven years as Bonus

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.

```
fit2 <- lm(Claims ~ ., InsuranceDF)
summary(fit2)
```

```
##
## Call:
## lm(formula = Claims ~ ., data = InsuranceDF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -181.330   -3.196    0.887    3.755  231.782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.327e+00  1.436e+00   4.405 1.11e-05 ***
## Kilometres  -1.220e+00  2.462e-01  -4.956 7.75e-07 ***
## Zone        -7.697e-01  1.752e-01  -4.394 1.17e-05 ***
## Bonus       -4.339e-01  1.755e-01  -2.473  0.01349 *
## Make         4.402e-01  1.383e-01   3.182  0.00148 **
## Insured     -4.918e-03  1.735e-04 -28.349  < 2e-16 ***
## Payment      2.224e-04  9.762e-07 227.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.08 on 2175 degrees of freedom
## Multiple R-squared:  0.9937, Adjusted R-squared:  0.9936
## F-statistic: 5.685e+04 on 6 and 2175 DF,  p-value: < 2.2e-16
```

Observation

From the linear model we can conclude that all the independent variables have a high significance on the Claim variable.