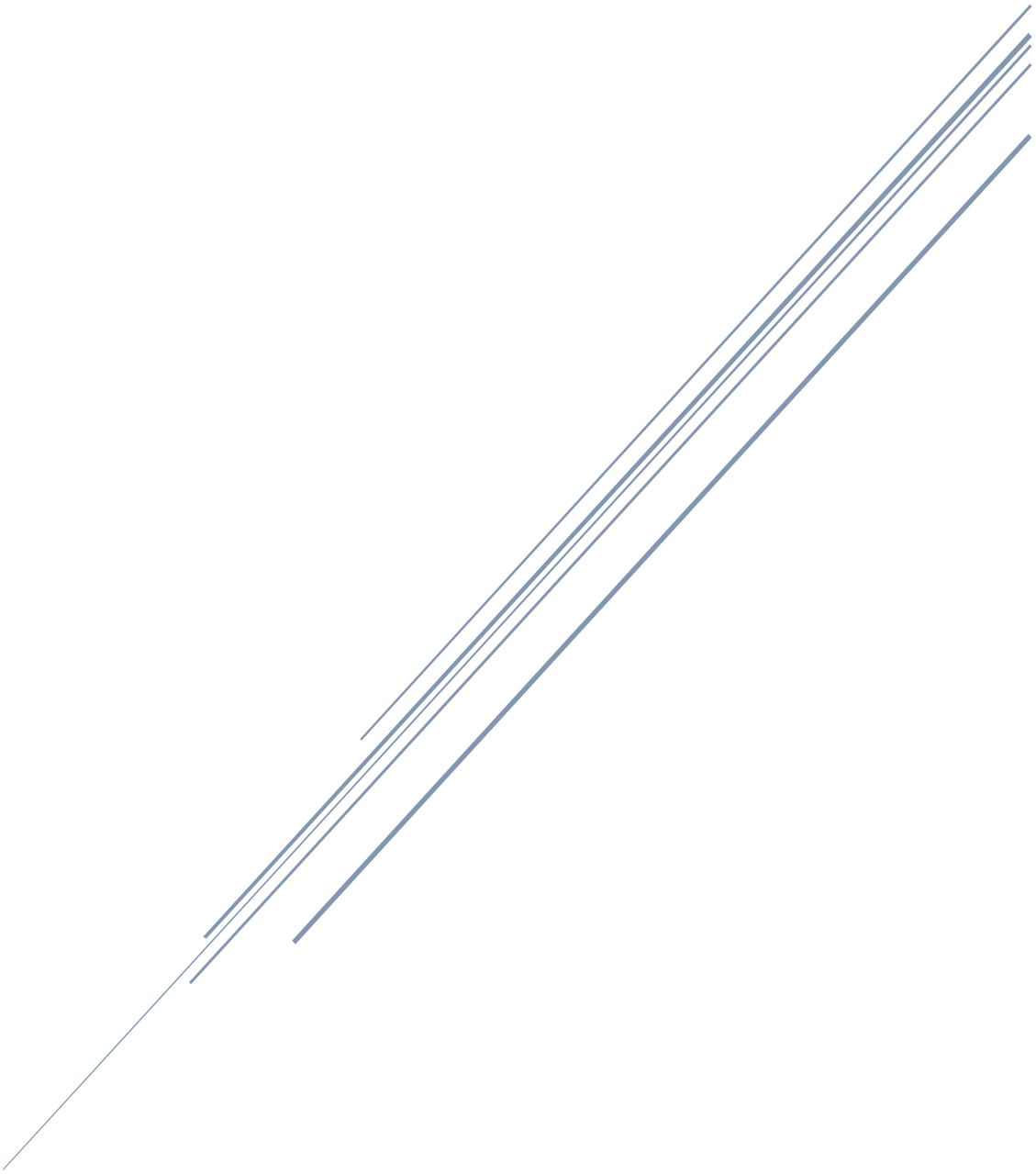
DATA ANALYSIS OF A MOTOR INSURANCE COMPANY



Objectives

a) Provide comprehensive descriptive analysis.
b) Perform correlation analysis between variables and provide conclusions.
c) Establish reasons for insurance payment increase and decrease by developing a regression model and test this model.
d) Establish what affects claim rates so as to decide the right premiums for a certain set of situations. Develop an appropriate regression model and test it.
e) The company is planning to extend their coverage over a few more cities/areas in near future and would like to predict their payments and number of claims. The below scenarios have been given to get an idea of the future.

    Case 1: Vittangi (A small city in the north), 8500 km travel per year, Bonus for 2 years, type 3 cars with 4621 insured amount.

    Case 2: Halmstad and outskirt, type 9 cars with average 12500 km travel per year, no claim bonus, average 9500 insured amount

    Case 3: Uppsala (A large city), average 22300 km travel per year, estimation between 17500 to 25416 insured amount, type 3 car, 4 years bonus.

Objective A

Descriptive Analysis

Check for any missing values to ensure the dataset is complete, no missing values in this case.

```
> colSums(is.na(Data))
Kilometres       Zone      Bonus       Make    Insured     Claims    Payment
         0          0          0          0          0          0          0
```

The below descriptive analysis performed by using the "summary" function in R and shows the central tendency measures of different variables in the motor insurance dataset.

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

Change the categorical values to factors, as per the below screenshot.

```
> Data$Kilometres <- as.factor(Data$Kilometres)
> Data$Zone <- as.factor(Data$Zone)
> Data$Bonus <- as.factor(Data$Bonus)
> Data$Make <- as.factor(Data$Make)
> summary(Data)
 Kilometres Zone    Bonus       Make       Insured             Claims           Payment
 1:439      1:315   1:307   1      :245   Min.   :     0.01   Min.   :   0.00   Min.   :       0
 2:441      2:315   2:312   2      :245   1st Qu.:    21.61   1st Qu.:   1.00   1st Qu.:    2989
 3:441      3:315   3:310   9      :245   Median :    81.53   Median :   5.00   Median :   27404
 4:434      4:315   4:310   5      :244   Mean   :  1092.20   Mean   :  51.87   Mean   :  257008
 5:427      5:313   5:313   6      :244   3rd Qu.:   389.78   3rd Qu.:  21.00   3rd Qu.:  111954
            6:315   6:315   3      :242   Max.   :127687.27   Max.   :3338.00   Max.   :18245026
            7:294   7:315   (Other):717
```

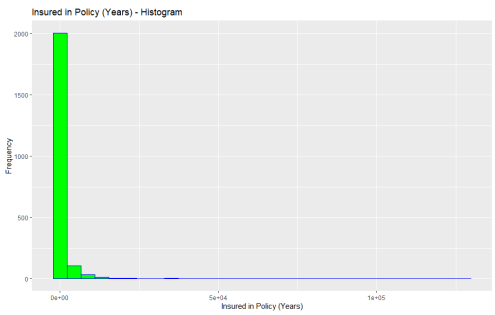
I have used the pastecs package to see further descriptive analysis that shows different measures of dispersion such as range, variance, and standard deviation.

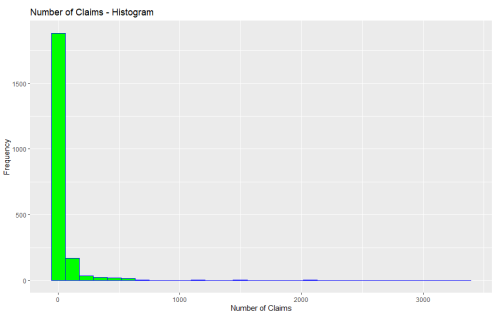
```
> stat.desc(Data)
          Kilometres Zone Bonus Make         Insured        Claims      Payment
nbr.val           NA   NA    NA   NA      2182.000000   2182.000000 2.182000e+03
nbr.null          NA   NA    NA   NA         0.000000    385.000000 3.850000e+02
nbr.na            NA   NA    NA   NA         0.000000      0.000000 0.000000e+00
min               NA   NA    NA   NA         0.010000      0.000000 0.000000e+00
max               NA   NA    NA   NA    127687.270000   3338.000000 1.824503e+07
range             NA   NA    NA   NA    127687.260000   3338.000000 1.824503e+07
sum               NA   NA    NA   NA   2383170.080000 113171.000000 5.607907e+08
median            NA   NA    NA   NA        81.525000      5.000000 2.740350e+04
mean              NA   NA    NA   NA      1092.195270     51.865720 2.570076e+05
SE.mean           NA   NA    NA   NA       121.193065      4.318188 2.177781e+04
CI.mean           NA   NA    NA   NA       237.665936      8.468192 4.270743e+04
var               NA   NA    NA   NA  32048690.027080  40687.203877 1.034864e+12
std.dev           NA   NA    NA   NA      5661.156245    201.710694 1.017283e+06
coef.var          NA   NA    NA   NA         5.183282      3.889095 3.958180e+00
```

Visualising the Data

Histograms

The descriptive analysis told us that all of these continuous variables have very low third quartile values which shows that the data is clustered heavily around the left tail (positively skewed). The fact that the mean value is higher than Q3 for all of these continuous variables shows that all three have high value outliers that are impacting the mean insured years, claims per year and total payments (SEK) – as shown in the histograms below.

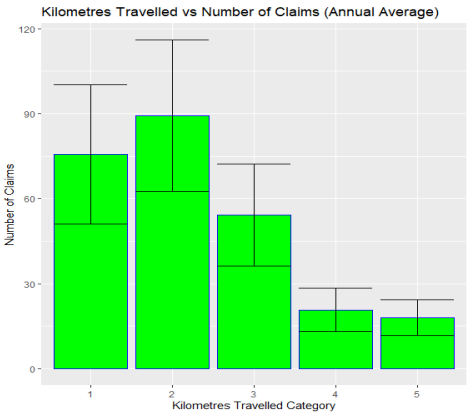


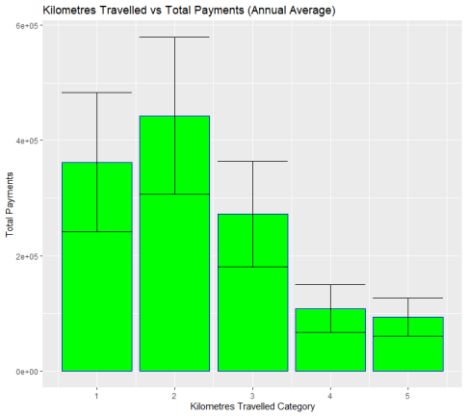


Bar Charts

Kilometres vs Claims & Payment

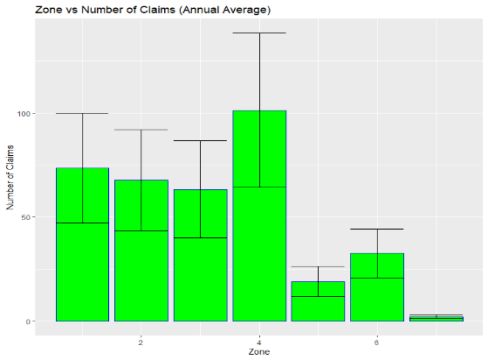
- Customers who travelled between 1,000-1,500 kilometres per year have the highest number of insurance claims and insurance payments and those who travelled greater than 25,000 kilometres have the least number of claims and lowest value of insurance payments.

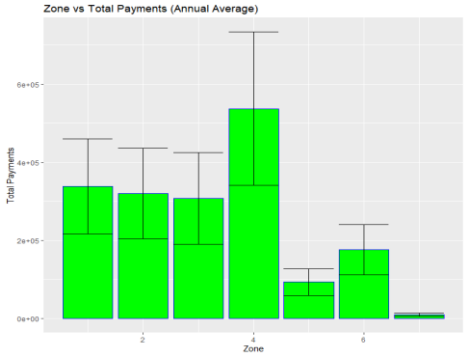




Zone vs Claims & Payment

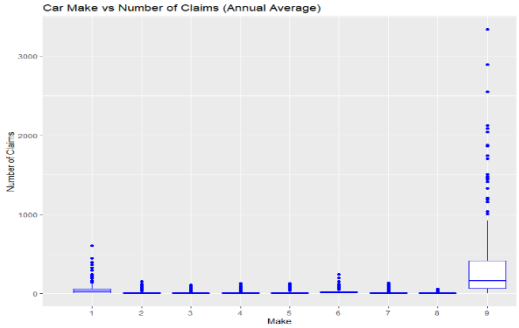
- The mean number of claims and payments are greatest in rural areas in southern Sweden and lowest in Gotland.
- It is to be expected that the Zone with the greatest number of claims would also have the highest value of payments and vice versa.

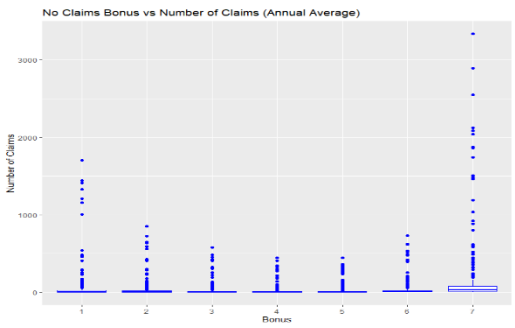




Box Plots

- Customers with 8 years no claims (category 7) have the largest range of claims made by customers, and also the highest average/mean.
- Customers who drive a category 9 car are most likely to make a claim on their insurance policy.



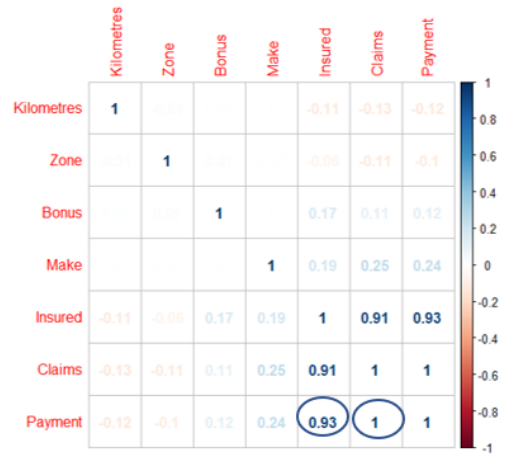


Objective B

Correlation Analysis

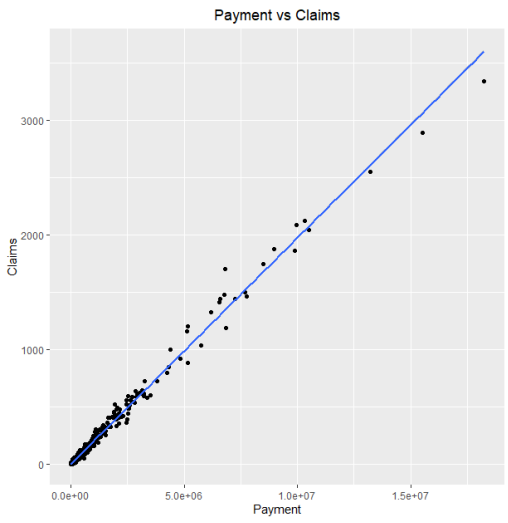
The below tables shows the correlation between different variables using Pearson's method. Payment and Claims have the strongest relationship with a positive correlation of 99.54%. Therefore, we can conclude that the number of claims has the biggest impact of the total value of payments (SEK). Payment and Insured have the second strongest relationship with a positive correlation of 93.32, and we can therefore state that the number of insured in policy -years has the next biggest impact on the total value of payments (SEK).

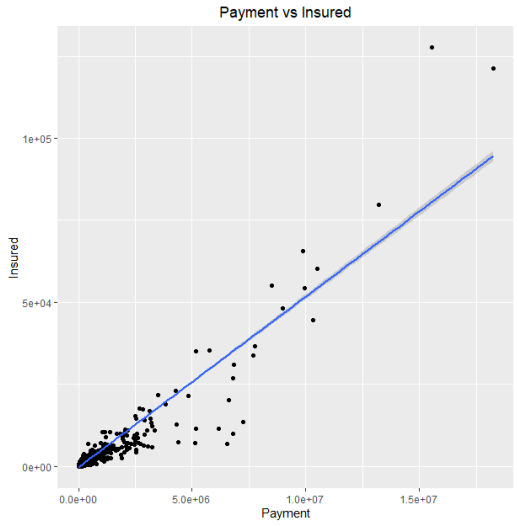
The general rule of thumb is anything with over 50% correlation is strong, so it is clear that 99.54% and 93.32% correlation signal very strong positive relationships i.e., when one variable increases, so does the other.




```
> Hmisc::rcorr(as.matrix(Data))
           Kilometres  Zone Bonus  Make Insured Claims Payment
Kilometres       1.00 -0.01  0.01  0.00   -0.11  -0.13   -0.12
Zone            -0.01  1.00  0.01 -0.01   -0.06  -0.11   -0.10
Bonus            0.01  0.01  1.00  0.00    0.17   0.11    0.12
Make             0.00 -0.01  0.00  1.00    0.19   0.25    0.24
Insured         -0.11 -0.06  0.17  0.19    1.00   0.91    0.93
Claims          -0.13 -0.11  0.11  0.25    0.91   1.00    1.00
Payment         -0.12 -0.10  0.12  0.24    0.93   1.00    1.00

n= 2182


P
           Kilometres Zone   Bonus  Make   Insured Claims Payment
Kilometres            0.5120 0.7358 0.9008 0.0000  0.0000 0.0000
Zone       0.5120            0.5832 0.8076 0.0065  0.0000 0.0000
Bonus      0.7358     0.5832        0.9200 0.0000  0.0000 0.0000
Make       0.9008     0.8076 0.9200        0.0000  0.0000 0.0000
Insured    0.0000     0.0065 0.0000 0.0000         0.0000 0.0000
Claims     0.0000     0.0000 0.0000 0.0000 0.0000         0.0000
Payment    0.0000     0.0000 0.0000 0.0000 0.0000 0.0000
```

We can also see from the scatter graph below that "Payments vs Claims" and "Payments vs Insured" have a linear relationship, which graphically displays the strong positive correlation between these variables.





Objective C

Dependent Variable: Payment

Independent Variable: Kilometres, Zone, Bonus, Make, Insured and Claims

P-Values

All of the p-values (independent vs dependent variable) are < 0.05 which shows that the correlation with the outcome is significant and that the null hypothesis is false or should be rejected.

```
P
           Kilometres Zone   Bonus  Make   Insured Claims Payment
Kilometres            0.5120 0.7358 0.9008 0.0000  0.0000 0.0000
Zone       0.5120             0.5832 0.8076 0.0065  0.0000 0.0000
Bonus      0.7358     0.5832         0.9200 0.0000  0.0000 0.0000
Make       0.9008     0.8076 0.9200        0.0000  0.0000 0.0000
Insured    0.0000     0.0065 0.0000 0.0000         0.0000 0.0000
Claims     0.0000     0.0000 0.0000 0.0000 0.0000         0.0000
Payment    0.0000     0.0000 0.0000 0.0000 0.0000  0.0000
```

Correlation

We need to take a look at multicollinearity before we develop a regression model. A very strong correlation of 0.91 between independent variables Insured and Claims which is a problem because independent variables should be independent. Because the degree of correlation between these variables is high, it can cause problems when we fit the model and interpret the results.

Developing Model

Stepwise Regression – Both Ways

- Tables below show stepwise regression (both ways) when excluding claims and insured variables respectively:

```
> model2 <- lm(Payment ~ Insured + Make + Bonus + Zone + Kilometres, data = Data)
> #stepwise regression - both ways
> ols_step_both_p(model2)
```

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Insured	addition	0.871	0.871	159.6460	62096.0948	365607.0389
2	Make	addition	0.876	0.876	68.9820	62009.8401	358369.7060
3	Zone	addition	0.878	0.878	27.7880	61969.4484	354986.8429
4	Bonus	addition	0.880	0.879	9.2270	61950.9767	353406.6203
5	Kilometres	addition	0.880	0.880	6.0000	61947.7416	353064.0240

```
> model <- lm(Payment ~ Claims + Make + Bonus + Zone + Kilometres, data = Data)
> ols_step_both_p(model)
```

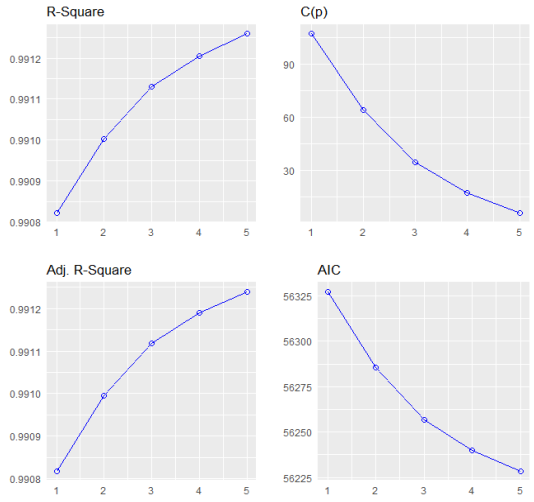
Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Claims	addition	0.991	0.991	107.2110	56327.3480	97481.7111
2	Bonus	addition	0.991	0.991	63.9120	56285.6605	96532.8442
3	Zone	addition	0.991	0.991	34.5720	56256.9152	95877.1425
4	Make	addition	0.991	0.991	17.5430	56240.0338	95485.1382
5	Kilometres	addition	0.991	0.991	6.0000	56228.4955	95211.2467

Model lm(Payment ~ Claims + Make + Bonus + Zone + Kilometres, data = Data) has a higher R-Square, Adj R-Square and lower AIC which suggests a better fit thus will be the model to test.

Testing the Model

lm(Payment ~ Claims + Make + Bonus + Zone + Kilometres, data = Data)

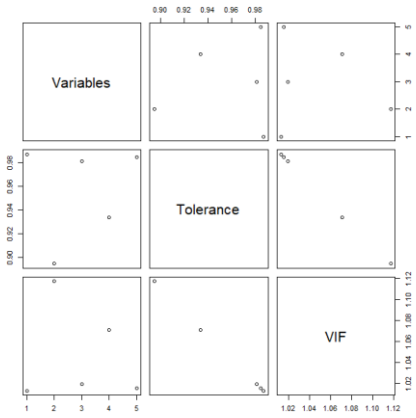



```
> regressionmodel <- lm(Payment ~ Claims + Bonus + Zone + Make + Kilometres, data = Data)
> summary(regressionmodel)

Call:
lm(formula = Payment ~ Claims + Bonus + Zone + Make + Kilometres,
    data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-1709435   -19823      329    21690  1436850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52209.42    8463.40  -6.169 8.18e-10 ***
Claims       5036.73      10.68 471.408  < 2e-16 ***
Bonus        6557.21    1025.69   6.393 1.98e-10 ***
Zone         6009.06    1033.00   5.817 6.87e-09 ***
Make        -3653.30     815.59  -4.479 7.88e-06 ***
Kilometres   5370.36    1459.30   3.680 0.000239 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 95210 on 2176 degrees of freedom
Multiple R-squared:  0.9913,    Adjusted R-squared:  0.9912
F-statistic: 4.936e+04 on 5 and 2176 DF,  p-value: < 2.2e-16
```

- R-Squared value for model is 99.1% meaning that this model nearly explains all the variation in the response variable around its mean.
- The problem with R-Squared is that this value will always increase with the more variables you add to it.
- AIC penalises the model for having more variables. The lower the AIC the better the model fits, therefore if we remove either the Insured or Claims variable due to multicollinearity it would appear as though a model excluding the "Insured" variable is a better fit.
- T-values have a score greater 2 or less than -2, which shows that we have confidence in these variables as predictors.
- F-statistic: 14.74 on 4 and 661 DF, p-value: 1.541e-11

Multicollinearity

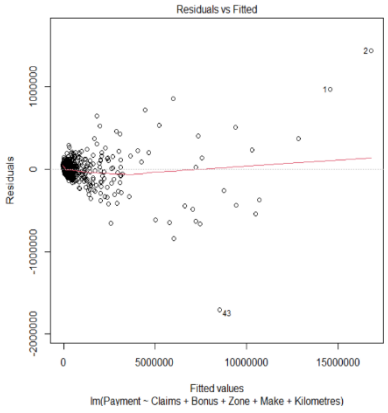
- We have already removed one variable due to multicollinearity which was established from very high correlation between independent variables.
- Test for any further multicollinearity by measuring the Variance Inflation Factor (VIF).
- The graph shows that all 5 variables have a VIF score of ~ 1, and the mean of all 5 variables is 1.04.
- Therefore, there is no multicollinearity issue with this model.

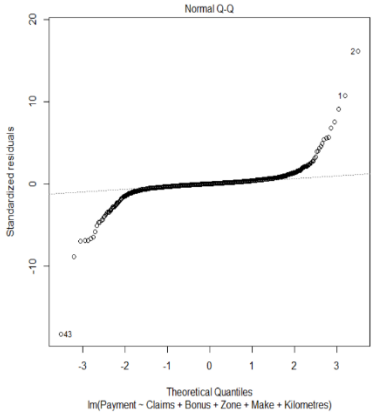


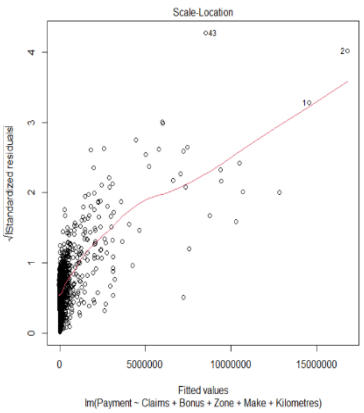
Residuals

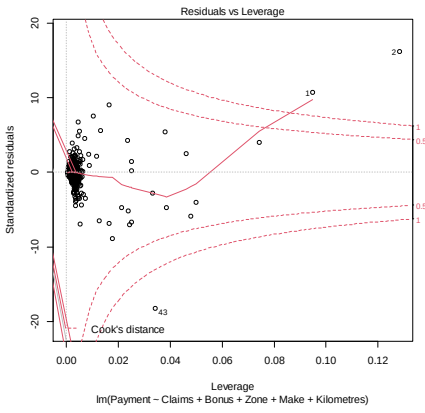
```
> summary(LargeResids)
   Mode   FALSE    TRUE
logical    2112      70
> summary(VeryLargeResids)
   Mode   FALSE    TRUE
logical    2134      48
> summary(LargestResids)
   Mode   FALSE    TRUE
logical    2149      33
> (70/(2112+70))*100
[1] 3.208066
> (48/(2134+48))*100
[1] 2.199817
> (33/(2149+33))*100
[1] 1.512374
```

- 70 cases (3.2%) lie outside the limits of 1.96 and -1.96
- 48 cases (2.2%) lie outside the limits of 2.58 and -2.58
- 33 cases (1.5%) lie outside the limits of 3.29 and -3.29
- Model is normally distributed as 96.8% of our data is within 1.96 standard deviations from the mean.





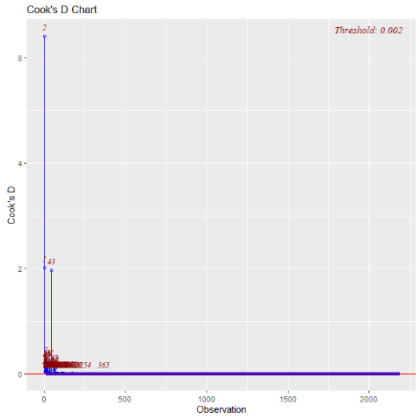




Influential Cases – Cooks Distance

```
> Data[which(Data[,9]>1),]
# A tibble: 3 x 9
  Kilometres  Zone Bonus  Make Insured Claims  Payment Rstandard CooksDistance
       <dbl> <dbl> <dbl> <dbl>   <dbl>  <dbl>    <dbl>     <dbl>         <dbl>
1          1     4     7     9  127687.   2894 15540162      10.7          2.02
2          2     4     7     9  121293.   3338 18245026      16.2          6.41
3          1     1     1     9    9998.   1704  6805992     -18.3          1.97
```

- Three influential cases with a Cook's distance of greater than 1 i.e., points that are negatively affecting the regression model.
- I have decided not to remove these outliers from the data due to the small number of cases (3).



Conclusion

From testing model "lm(Payment ~ Claims + Make + Bonus + Zone + Kilometres, data = Data)" we have identified the variables that have an impact on the dependent variable (Payment). The testing completed has allowed me to make the conclusion that this model is a good fit.

The estimate or B values in the model summary table below tells us the degree to what a change in each predictor will impact the dependent variable. The b-values tell us how much the Payment will increase/decrease with as the predictor variables increase by one unit. This here answers the question in regard to the reasons for insurance payment increase and decrease.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52209.42    8463.40  -6.169 8.18e-10 ***
Claims       5036.73      10.68 471.408  < 2e-16 ***
Bonus        6557.21    1025.69   6.393 1.98e-10 ***
Zone         6009.06    1033.00   5.817 6.87e-09 ***
Make        -3653.30     815.59  -4.479 7.88e-06 ***
Kilometres   5370.36    1459.30   3.680 0.000239 ***
```

Objective D

Dependent Variable: Claims

Independent Variable: Kilometres, Zone, Bonus, Make, Insured

Note: payment variable is not considered here.

P-Values

All of the p-values shown below (independent variables vs “Claims” variables) are < 0.05 which shows that the correlation with the outcome is significant and that the null hypothesis is false or should be rejected.

```
P
           Kilometres Zone   Bonus  Make   Insured Claims Payment
Kilometres            0.5120 0.7358 0.9008 0.0000  0.0000 0.0000
Zone       0.5120             0.5832 0.8076 0.0065  0.0000 0.0000
Bonus      0.7358     0.5832        0.9200 0.0000  0.0000 0.0000
Make       0.9008     0.8076 0.9200        0.0000  0.0000 0.0000
Insured    0.0000     0.0065 0.0000 0.0000         0.0000 0.0000
Claims     0.0000     0.0000 0.0000 0.0000 0.0000         0.0000
Payment    0.0000     0.0000 0.0000 0.0000 0.0000  0.0000
```

Correlation

All Independent variables appear to be independent because the degree of correlation between these variables is low as shown in the correlation matrix below. Therefore, there is no need to remove any variables from the model.

```
> Hmisc::rcorr(as.matrix(Data))
           Kilometres  Zone Bonus  Make Insured Claims Payment
Kilometres       1.00 -0.01  0.01  0.00   -0.11  -0.13   -0.12
Zone            -0.01  1.00  0.01 -0.01   -0.06  -0.11   -0.10
Bonus            0.01  0.01  1.00  0.00    0.17   0.11    0.12
Make             0.00 -0.01  0.00  1.00    0.19   0.25    0.24
Insured         -0.11 -0.06  0.17  0.19    1.00   0.91    0.93
Claims          -0.13 -0.11  0.11  0.25    0.91   1.00    1.00
Payment         -0.12 -0.10  0.12  0.24    0.93   1.00    1.00
```

Developing Model

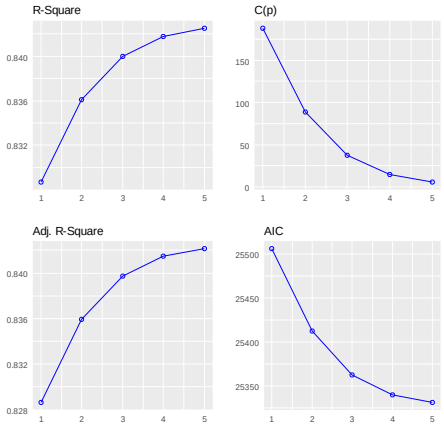
lm(Claims ~ Insured + Make + Bonus + Zone + Kilometres, data = Data)

Stepwise Regression – Both Ways

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Insured	addition	0.829	0.829	188.1980	25506.0646	83.4959
2	Make	addition	0.836	0.836	88.7210	25412.4218	81.7047
3	Zone	addition	0.840	0.840	37.1490	25362.1860	80.7510
4	Bonus	addition	0.842	0.841	14.4670	25339.6920	80.3175
5	Kilometres	addition	0.842	0.842	6.0000	25331.2215	80.1434

Testing the Model

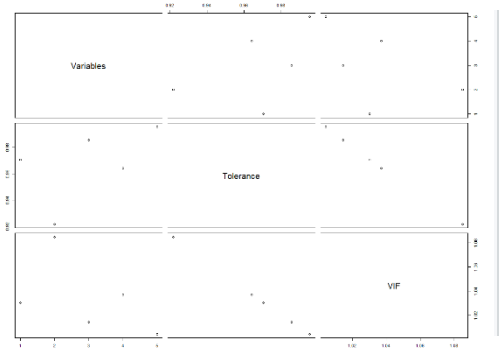



```
> summary(model3)

Call:
lm(formula = Claims ~ Insured + Make + Bonus + Zone + Kilometres,
    data = Data)

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 ***
Insured      0.0318697  0.0003158 100.933  < 2e-16 ***
Make         6.7725342  0.6755390  10.025  < 2e-16 ***
Bonus       -4.2468101  0.8707236  -4.877 1.15e-06 ***
Zone        -6.2924300  0.8647405  -7.277 4.75e-13 ***
Kilometres  -3.9648601  1.2255209  -3.235  0.00123 **
---
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
```

- Adjusted R-Squared - explains 84.21% of the variation in the response variable around its mean. This is a very high number an indicates a good fitting model.
- T-Value - we can see from the table above, that all t-values have a score greater 2 or less than -2, which shows that we have confidence in these variables as predictors.

Multicollinearity

There is no multicollinearity in this model as we can see from the graph/table below as the VIF scores are all around 1 which indicates no multicollinearity.

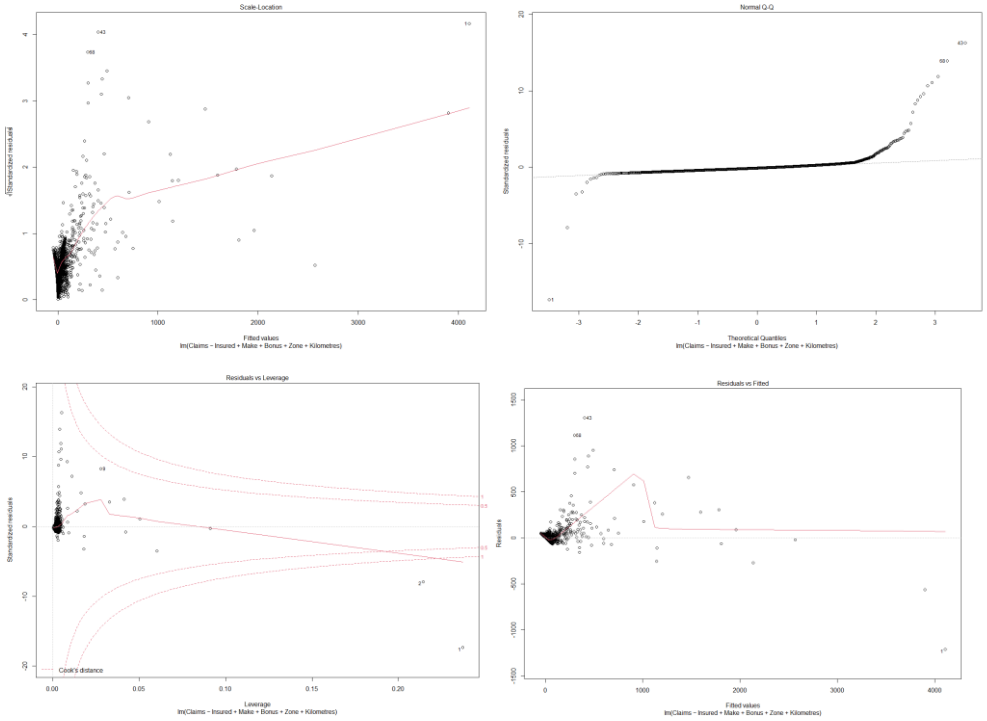



```
> ols_vif_tol(model3)
   Variables Tolerance      VIF
1    Insured 0.9216779 1.084978
2       Make 0.9642857 1.037037
3      Bonus 0.9705897 1.030301
4       Zone 0.9956379 1.004381
5 Kilometres 0.9857096 1.014498
```

Residuals

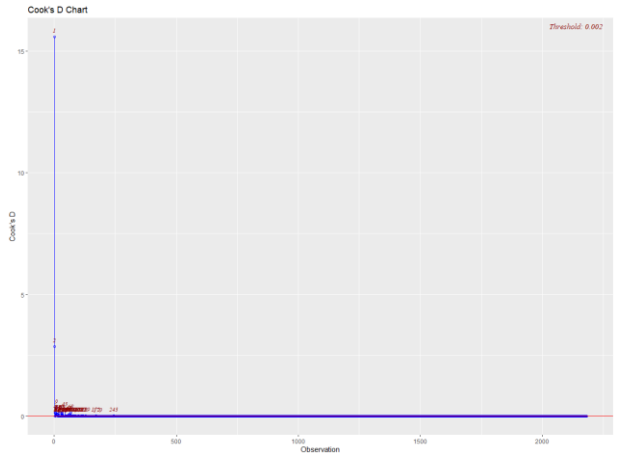
```
> summary(LargeResids2)
   Mode    FALSE     TRUE
logical     2134       48
> summary(VeryLargeResids2)
   Mode    FALSE     TRUE
logical     2148       34
> summary(LargestResids2)
   Mode    FALSE     TRUE
logical     2155       27
> 48/(48+2134)
[1] 0.02199817
> 34/(34+2148)
[1] 0.01558203
> 27/(27+2155)
[1] 0.01237397
```

- 48 cases (2.2%) lie outside the limits of 1.96 and -1.96
- 34 cases (1.6%) lie outside the limits of 2.58 and -2.58
- 27 cases (1.2%) lie outside the limits of 3.29 and -3.29
- From this information we can determine that our model is normally distributed as 97.8% of our data is within 1.96 standard deviations from the mean.



Influential Cases - Cooks Distance

Only two variables with a Cooks distance score of greater than 1, and I have therefore decided not to remove these from the population due to the small number.



Conclusion

From the summary of model "lm(Claims ~ Insured + Make + Bonus + Zone + Kilometres, data = Data) shown below, we can see the b-values which helps us understand what affects the claim rates. It is clear that a unit change in the Insured and Make predictors have a positive impact the number of claims. However, Bonus, Zone and Kilometres all have negative b-values and thus each unit change in these predictor variables have a negative impact on the number of claims.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.1230027  7.1270679   5.209 2.08e-07 ***
Insured      0.0318697  0.0003158 100.933  < 2e-16 ***
Make         6.7725342  0.6755390  10.025  < 2e-16 ***
Bonus       -4.2468101  0.8707236  -4.877 1.15e-06 ***
Zone        -6.2924300  0.8647405  -7.277 4.75e-13 ***
Kilometres  -3.9648601  1.2255209  -3.235  0.00123 **
```

Objective E

Case Interpretation

Case #	Kilometres	Zone	Bonus	Make	Insured
1	2	5	3	3	4,621
2	2	3	1	9	9,500
3A	4	2	5	3	17,500
3B	4	2	5	3	25,416

Plug in values to regression models (below) developed in part C & D using case interpretation numbers and respective b-values established from the regression analysis.

Predicted Claim Regression Model

lm(Claims ~ Insured + Make + Bonus + Zone + Kilometres, data = Data)

Predicted Payment Regression Model

lm(Payment ~ Claims + Make + Bonus + Zone + Kilometres, data = Data

Case 1 Interpretation

Claim Predicted Value

37.123 + (0.032 * 4621) + (6.773 * 3) + (-4.247 * 3) + (-6.292 * 5) + (-3.964 * 2) = 153

Payment Predicted Value

-52209.42 + (5036.73 * 153) + (-3653.30 * 3) + (6557.21 * 3) + (6009.06 * 5) + (5370.36 * 2) = 767,908

```
> Case1PredictedClaims <- 37.123 + (0.032 * 4621) + (6.773 * 3) + (-4.247 * 3) + (-6.292 * 5) + (-3.964 * 2)
> Case1PredictedClaims <- round(Case1PredictedClaims)
> Case1PredictedClaims
[1] 153
> Case1PredictedPayment <- -52209.42 + (5036.73 * Case1PredictedClaims) + (-3653.30 * 3) + (6557.21 * 3) + (6009.06 * 5) + (5370.36 * 2)
> Case1PredictedPayment
[1] 767908
```

Case 2 Interpretation

Claim Predicted Value

37.123 + (0.032 * 9500) + (6.773 * 9) + (-4.247 * 1) + (-6.292 * 3) + (-3.964 * 2) = 371

Payment Predicted Value

-52209.42 + (5036.73 * 371) + (-3653.30 * 9) + (6557.21 * 1) + (6009.06 * 3) + (5370.36 * 2) = 1,818,863

```
> Case2PredictedClaims <- 37.123 + (0.032 * 9500) + (6.773 * 9) + (-4.247 * 1) + (-6.292 * 3) + (-3.964 * 2)
> Case2PredictedClaims <- round(Case2PredictedClaims)
> Case2PredictedClaims
[1] 371
> Case2PredictedPayment <- -52209.42 + (5036.73 * Case2PredictedClaims) + (-3653.30 * 9) + (6557.21 * 1) + (6009.06 * 3) + (5370.36 * 2)
> Case2PredictedPayment
[1] 1818863
```

Case 3A Interpretation

Claim Predicted Value

37.123 + (0.032 * 17500) + (6.773 * 3) – (4.247 * 5) –(6.292 * 2) – (3.964 * 4) = 568

Payment Predicted Value

-52209.42 + (5036.73 * 567) + (-3653.30 * 3) + (6557.21 * 5) + (6009.06 * 2) + (5370.36 * 4) = 2,863,979

```
> Case3APredictedClaims <- 37.123 + (0.032 * 17500) + (6.773 * 3) + (-4.247 * 5) + (-6.292 * 2) + (-3.964 * 4)
> Case3APredictedClaims <- round(Case3APredictedClaims)
> Case3APredictedClaims
[1] 568
> Case3APredictedPayment <- -52209.42 + (5036.73 * Case3APredictedClaims) + (-3653.30 * 3) + (6557.21 * 5) + (6009.06 * 2) + (5370.36 * 4)
> Case3APredictedPayment
[1] 2863979
```

Case 3B Interpretation

Claim Predicted Value

37.123 + (0.032 * 25416) + (6.773 * 3) – (4.247 * 5) –(6.292 * 2) – (3.964 * 4) = 821

Payment Predicted Value

-52209.42 + (5036.73 * 567) + (-3653.30 * 3) + (6557.21 * 5) + (6009.06 * 2) + (5370.36 * 4) = 4,138,272

```
> Case3BPredictedClaims <- 37.123 + (0.032 * 25416) + (6.773 * 3) + (-4.247 * 5) + (-6.292 * 2) + (-3.964 * 4)
> Case3BPredictedClaims <- round(Case3BPredictedClaims)
> Case3BPredictedClaims
[1] 821
> Case3BPredictedPayment <- -52209.42 + (5036.73 * Case3BPredictedClaims) + (-3653.30 * 3) + (6557.21 * 5) + (6009.06 * 2) + (5370.36 * 4)
> Case3BPredictedPayment
[1] 4138272
```

Case Results

The below table shows the prediction for claims and payments for the 3 given scenarios:

Case #	Claims	Payment
1	153	767,908.00
2	371	1,818,863.00
3A	568	2,863,979.00
3B	821	4,138,272.00

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