UP2044856

EDA Report (Practical Data Science Coursework)
All graphs have been calculated **over a 12-month period**
And all relevant **outliers** have been removed from **the Claims Variable**

Introduction

The problem that has been presented to me is to obtain a high level of understanding of what is impacting the Claims and Payment from the risk categories that have been supplied.

I have begun my work by performing a comprehensive descriptive analysis using histograms and box and whisker plots.

The histograms have been used to show the mean number of claims against the claims variable and shows which variable have the highest and lowest values allowing us to understand at a higher level how everything is impacting each other, this has been repeated three different times which different variables, but the methods used remain the same.

I then used box and whisker plots to show the level of dispersion among the variables obtaining a greater understand of how the values are dispersed.

I then found out that our dataset is very skewed leading me to use the correlation analysis of Spearman over Pearson this has been explained below. I then began my correlation analysis and produced scatter plots and showing my line of best fit, then creating my assumptions and highlighted key ideas and impacts.

Finally produced a linear regression model and was able to show which variables had the highest and lowest level of impact on the Claims and Payment variable.

Data Cleansing

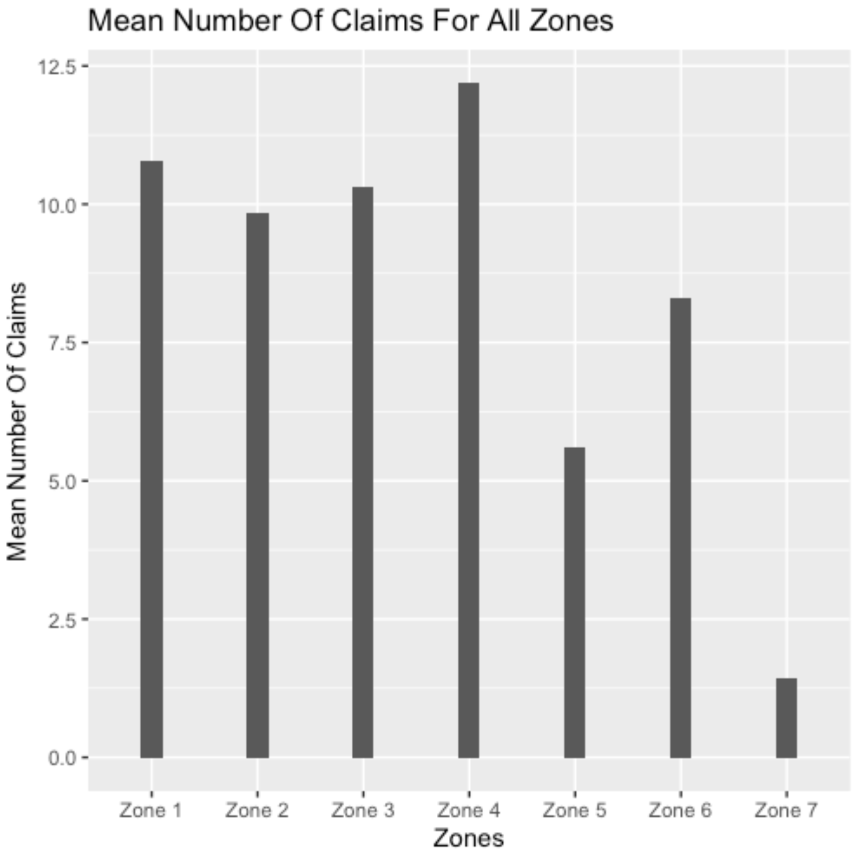
Outliers are data points that differ from other observations. They appear due to many different reasons, for example, data entry errors, measurement errors, data processing errors, or they may simply be a natural occurrence but may not necessarily truly reflect the true meaning of the data. These extreme values affect the variance and standard deviation of data distribution.

We will apply the Interquartile Rule to identify these outliers, we essentially want the subset of our dataset such that it falls within the range,

$$\mathrm{Q}_3 - 1.5 \cdot \mathrm{IQR} < X < \mathrm{Q}_3 + 1.5 \cdot \mathrm{IQR}$$

Applying this, we have removed 329 entries from our dataset which could have affected the integrity of our data. Now that we have sanitised our data, we are now ready to analyse it.

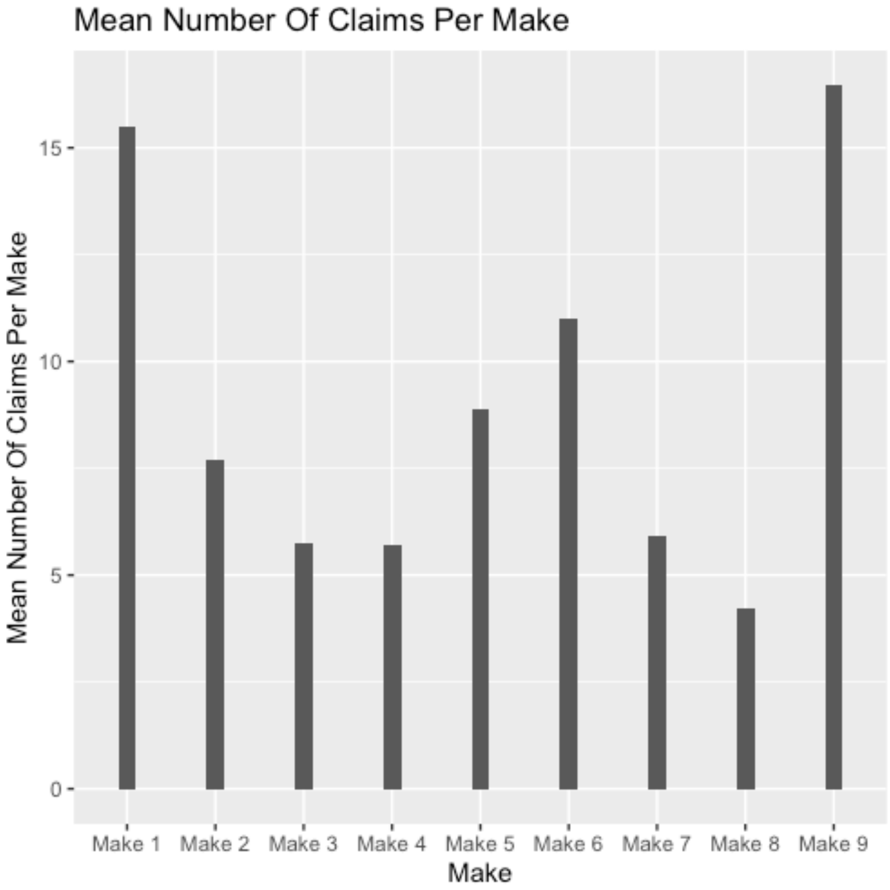
Central tendencies



Zone #	Mean
Zone 4	12.18455
Zone 1	10.78346
Zone 3	10.29804
Zone 2	9.832669
Zone 6	8.310714
Zone 5	5.619377
Zone 7	1.43299

Assumptions

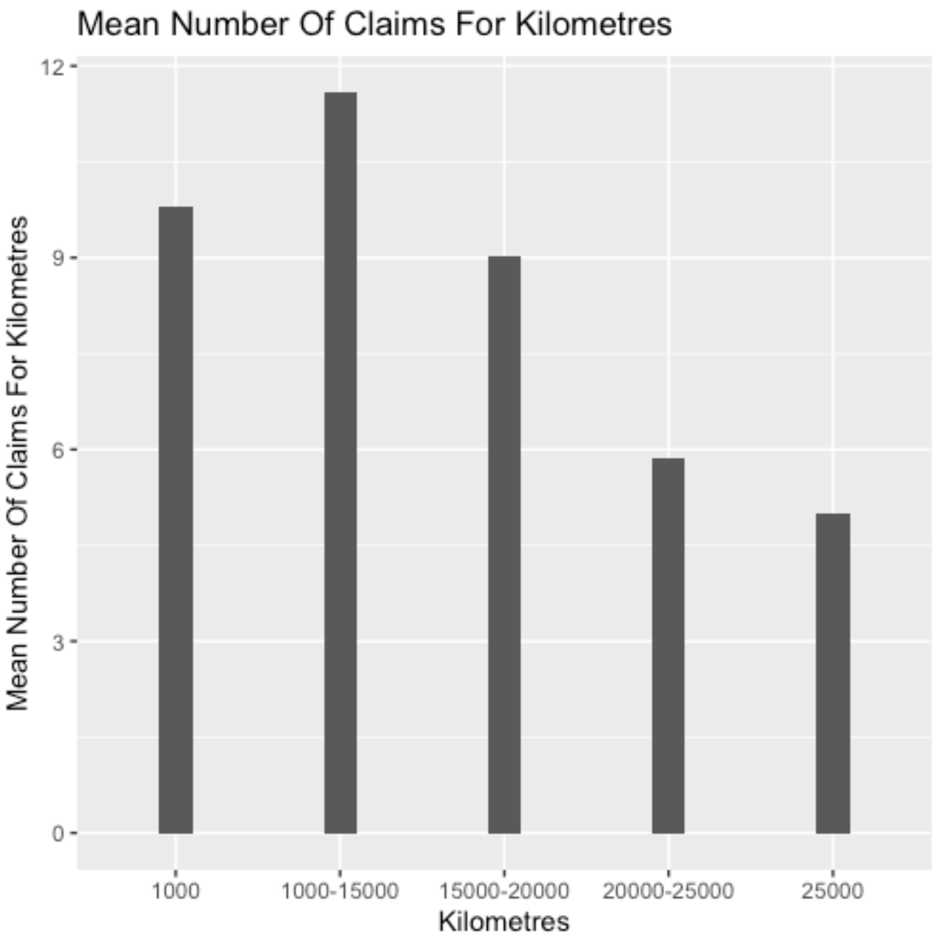
- Zone 4 (Rural areas in southern Sweden) has the highest mean number of claims this could be because the infrastructure has not been as well maintained eg: lights, roading, stop signs and so on.
- Zone 7 (Gotland) has the lowest number of claims as its population is far lower than that of southern Sweden, the population of Sweden is **10.23 million (2019)** in contrast to the population of Gotland is **only 58,464.**



Make #	Mean
Make 9	16.45614
Make 1	15.47514
Make 6	11.00455
Make 5	8.869565
Make 2	7.709402
Make 7	5.931034
Make 3	5.757447
Make 4	5.724891
Make 8	4.238298

Assumptions

- The reason for Make 9 being the highest is perhaps that all other car models are included in this variable, there are roughly 60 different car manufactures and being that only 8 of them are put into other variables means that the rest (52) are put into the Make 9 variable, therefore, making it the highest as there is a lot more variability.
- A possible reason for Make 8 being the lowest in terms of the mean number of claims is because these are cars that are owned by the average person who values reliability over anything else, therefore, resulting in the fewest number of claims.

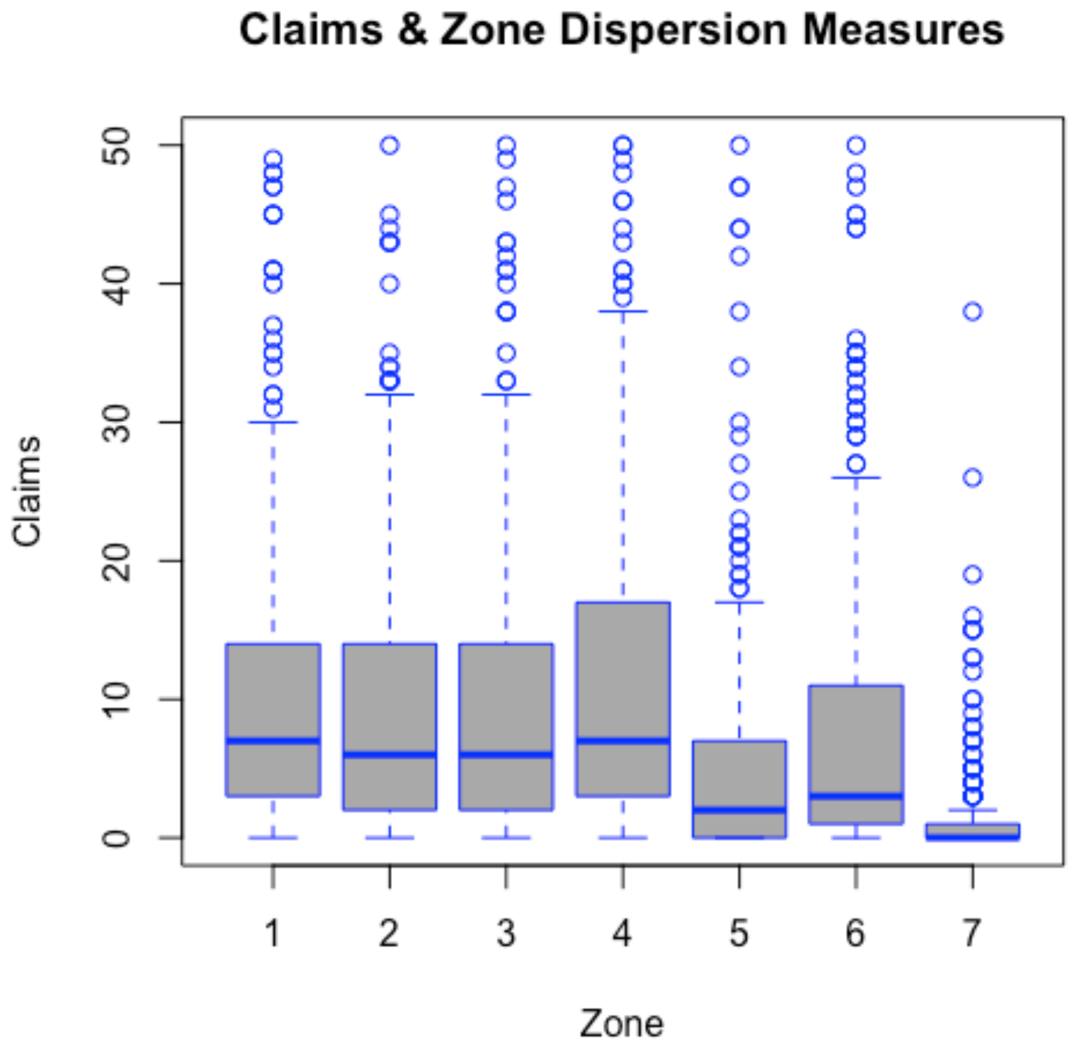


Kilometers	Mean
1000-15000	11.57771
1000	9.793872
15000-20000	9.030055
20000-25000	5.858586
25000	4.98977

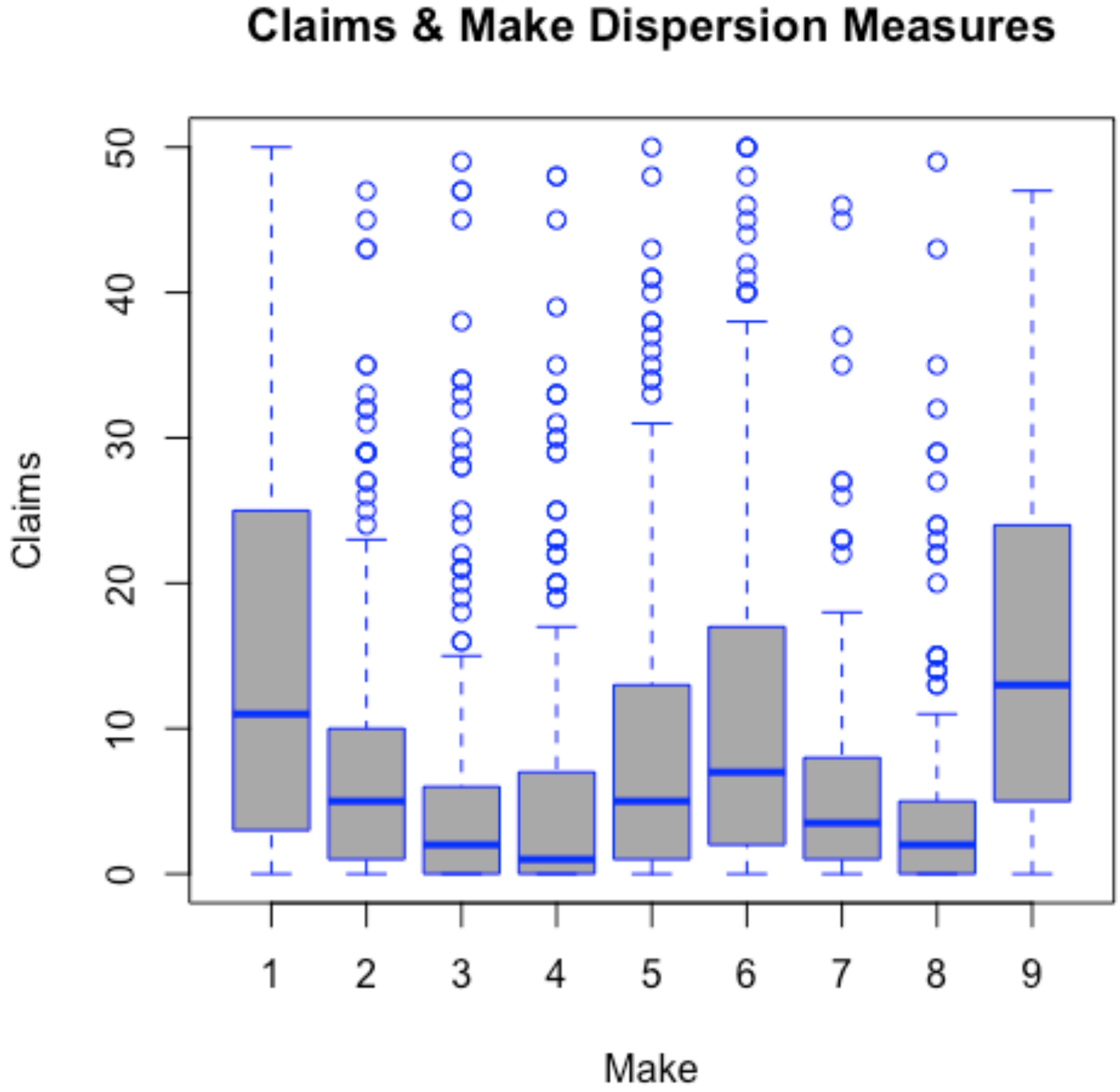
Assumptions

- We can assume that the reason for the highest number of claims for 1000-15000 variable is that drivers within this category are new drivers and have not yet obtained the level of driving skill required to have much fewer accidents while on the road
- We can also assume that the reason for the variable 25000 having the fewest number of claims is that these drivers are the most experienced out of the bunch and therefore resulting in the fewest number of claims over a 12 months

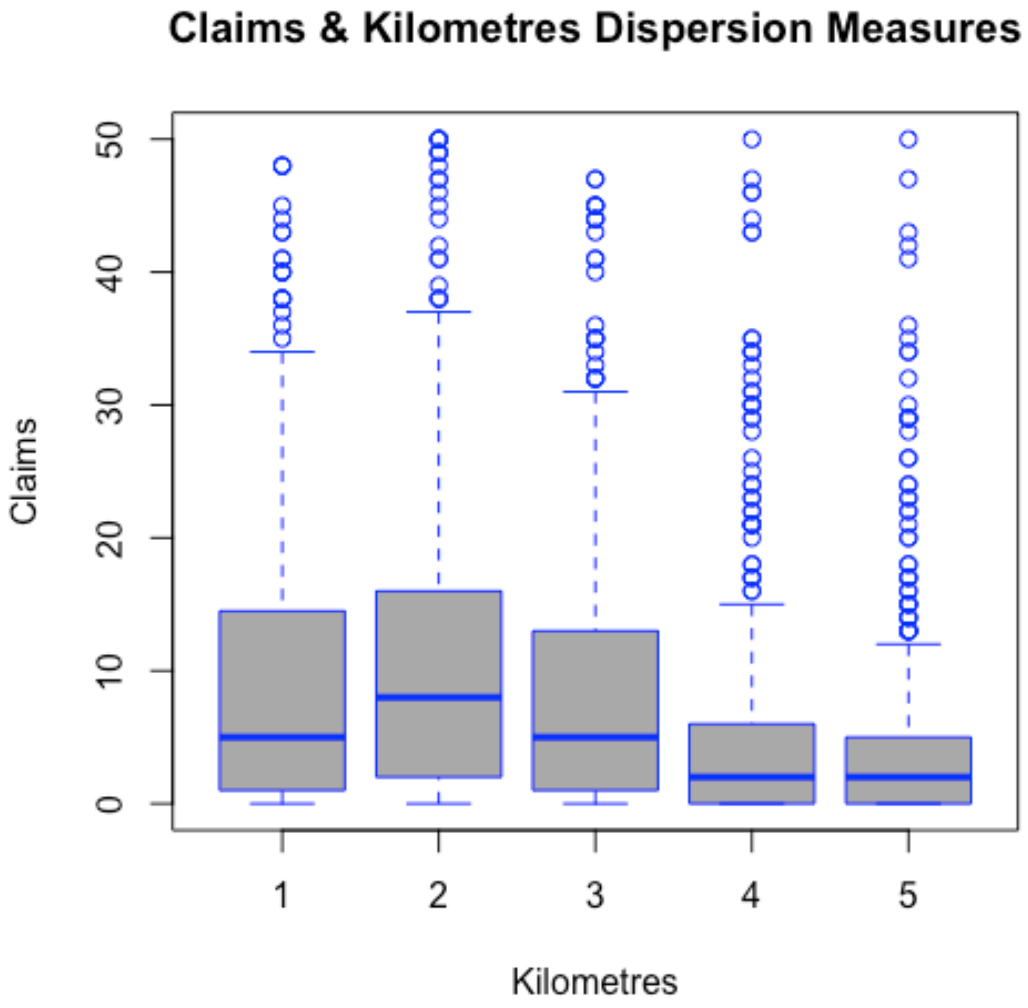
Dispersion Measures



	Min	Q1	Median	Mean	Q3	Q3-Q1
Zone 4	0	3	7	12.18	17	14
Zone 2	0	2	6	9.8	14	12
Zone 3	0	2	6	10.3	14	12
Zone 1	0	3	7	10.78	14	11
Zone 6	0	1	3	8.3	11	10
Zone 5	0	0	2	5.6	7	7
Zone 7	0	0	0	1.4	1	1



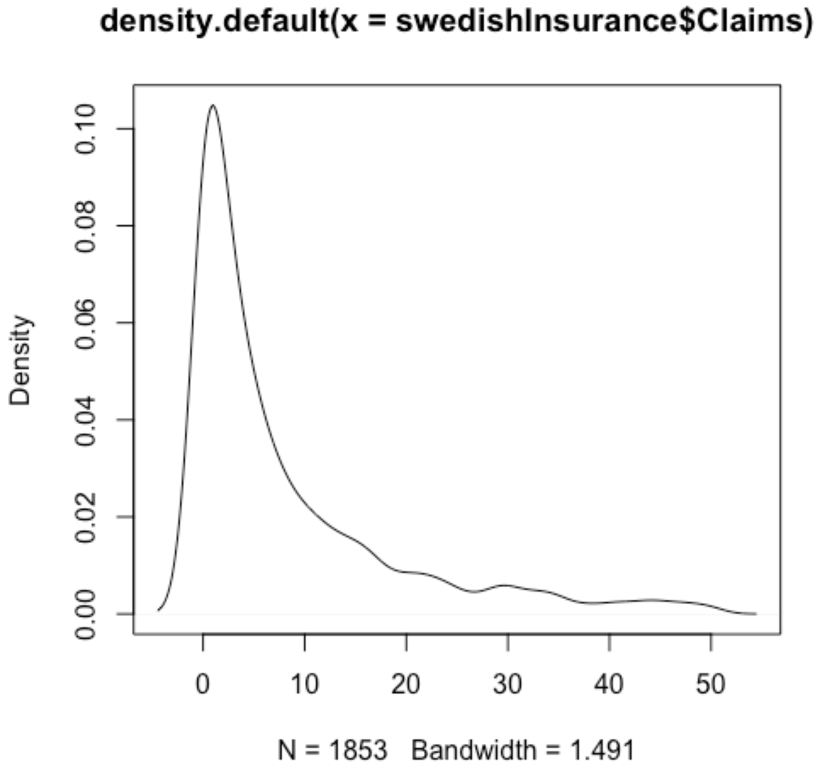
	Min	Q1	Median	Mean	Q3	Q3-Q1
Make 1	0	3	11	15.4	25	22
Make 9	0	5	13	16.4	24	19
Make 6	0	2	7	11	17	15
Make 5	0	1	5	8.8	13	12
Make 2	0	1	5	7.7	10	9
Make 4	0	0	1	5.7	7	7
Make 7	0	1	3.5	5.9	8	7
Make 3	0	0	2	5.7	6	6
Make 8	0	0	2	4.2	5	5



	Min	Q1	Median	Mean	Q3	Q3-Q1
Kilometers 2	0	2	8	11.58	16	14
Kilometers 1	0	1	5	9.7	14.5	13.5
Kilometers 3	0	1	5	9.03	13	12
Kilometers 4	0	0	2	5.8	6	6
Kilometers 5	0	0	2	4.99	5	5

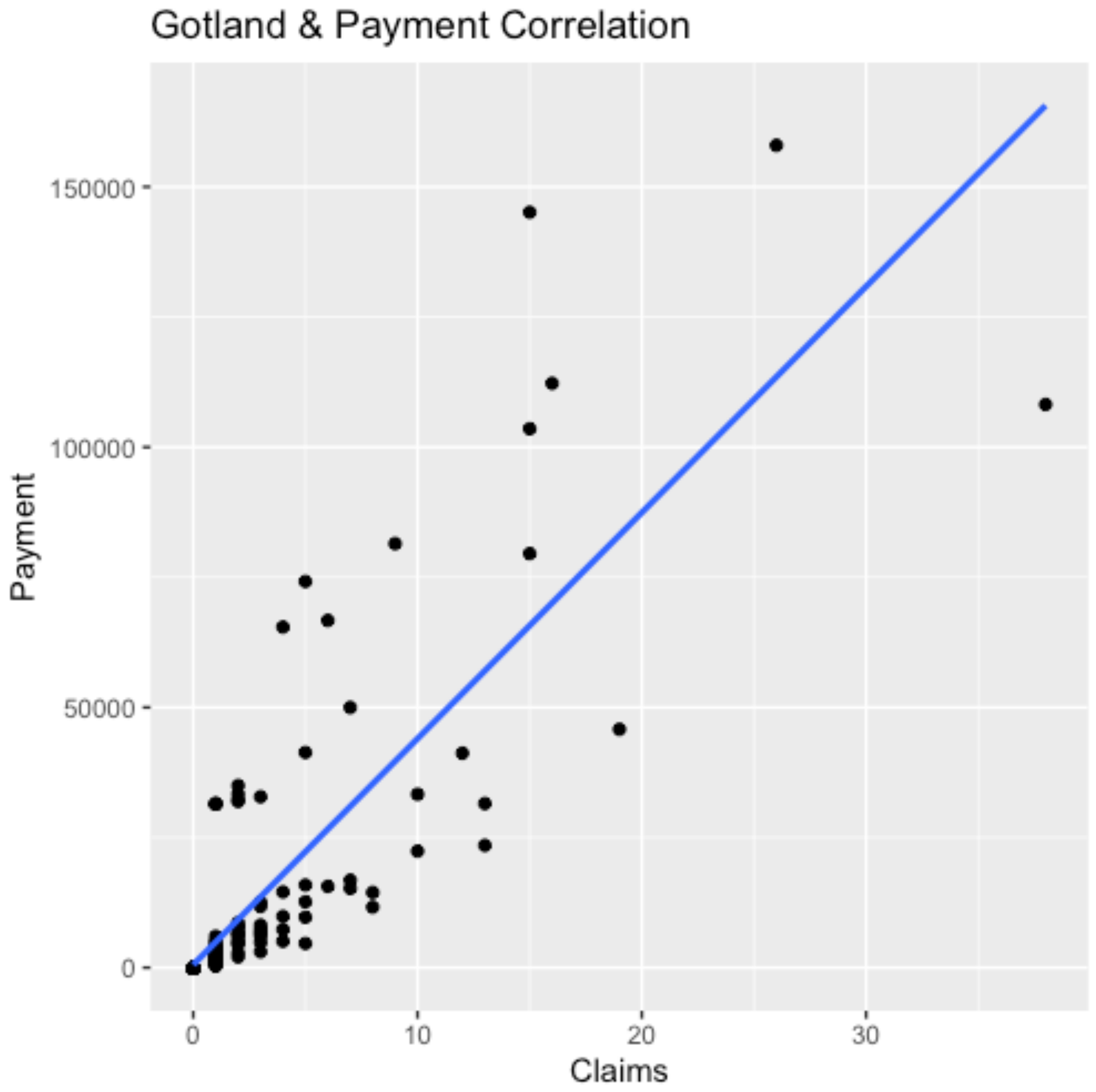
CORRELATION ANALYSIS

Distribution of the Claims from the Coursework Dataset



- Shapiro-Wilk normality test
- data: swedishInsurance$Claims
- W = 0.74958, p-value < 2.2e-16

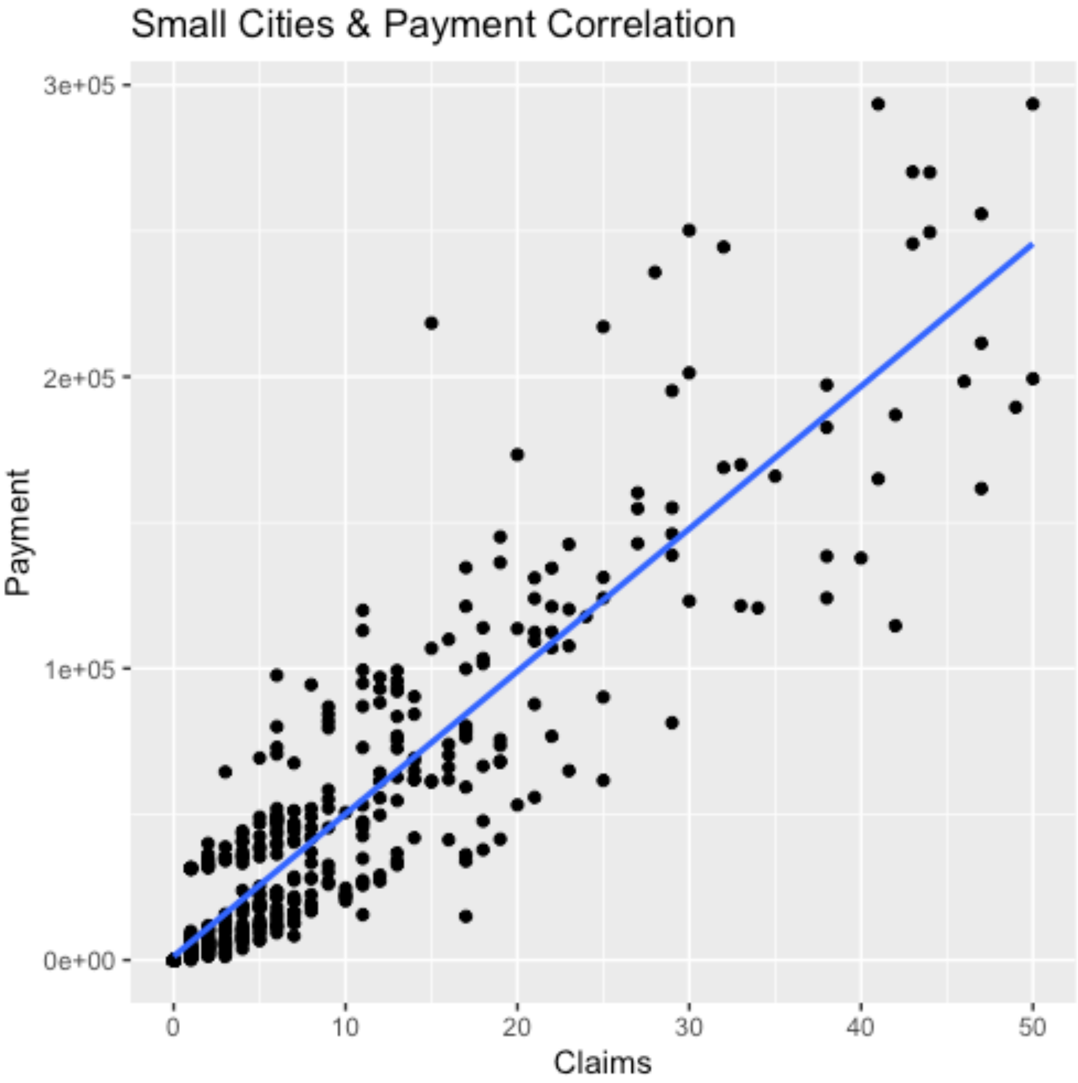
- Our density plot is **positively** skewed for the distribution of **Claims**
- p-value < 2.2e-16 as you can see our P-value is **smaller** than 2.2e-16 meaning that is it **negligible** as the value is so small
- But if the P-value was **more** than 0.05 then it means the data is **normally** distributed
- So as mentioned above this is the reason why I have gone with **Spearman** instead of Pearson's for my correlation analysis



Gotland (Zone 7 CA = **0.9869706**)

Assumptions

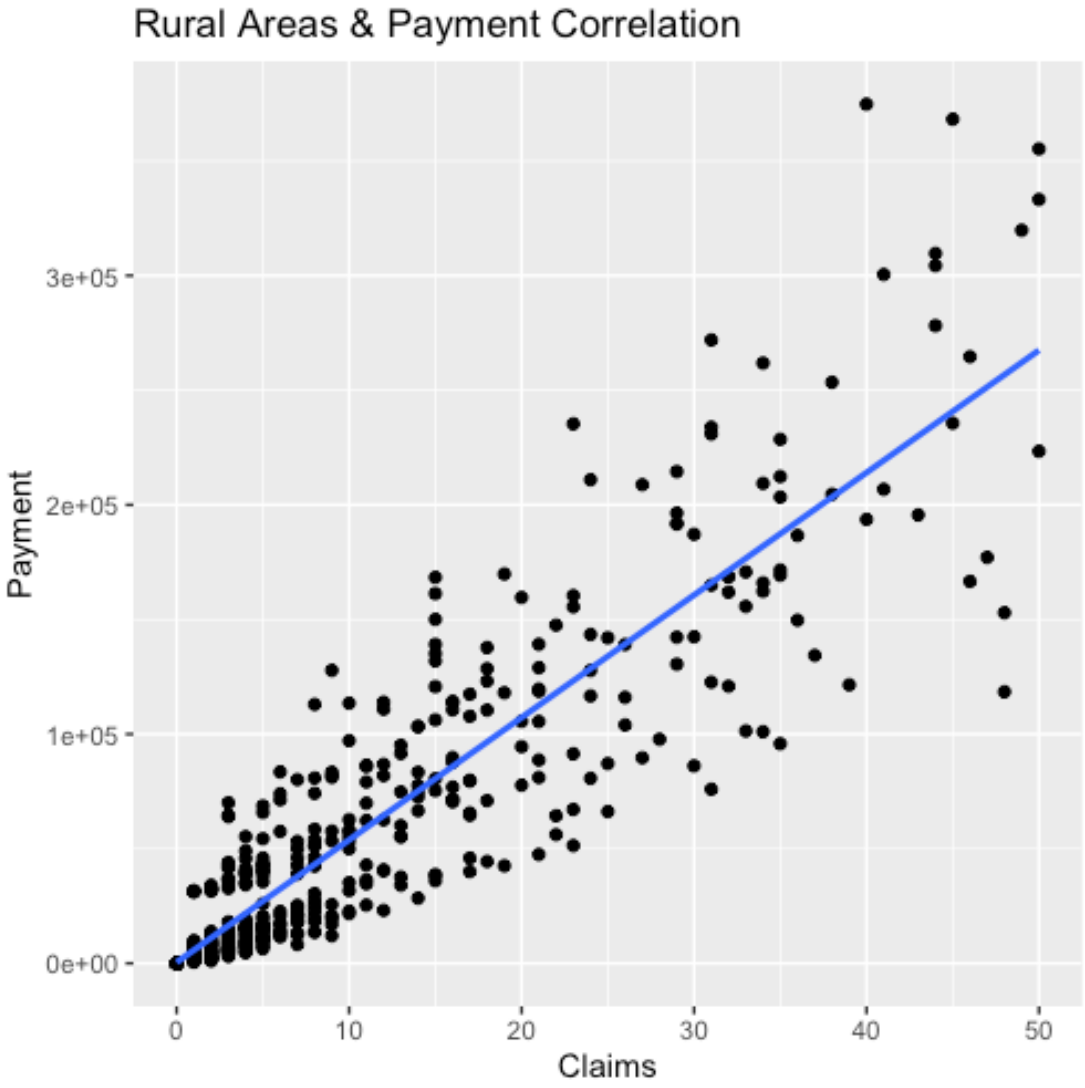
- Gotland had the highest number of claims this is more than likely a direct correlation due to the number of people that live on the island of Gotland (58,464).
- Due to the number of people and the claims it seems that although that their people living in Gotland the claims there were made when we compare the payment to the payment for this zone the correlation is very high.
- This could be due to several things, infrastructure, location, laws, and many more things that we could make an assumption off



- Small Cities (Zone 3 and Zone 5 CA = **0.92643**)

Assumptions

- Small Cities had the second number of claims, this could be because more people are living in these areas but due to the size of the area, this is heavily impacting the claims.
- Another reason is that these areas have much many more cars.
- Perhaps people living in these cities the vast majority drive but due to the size of where they live the likelihood of claims is much higher as we can see from the Correlation analysis

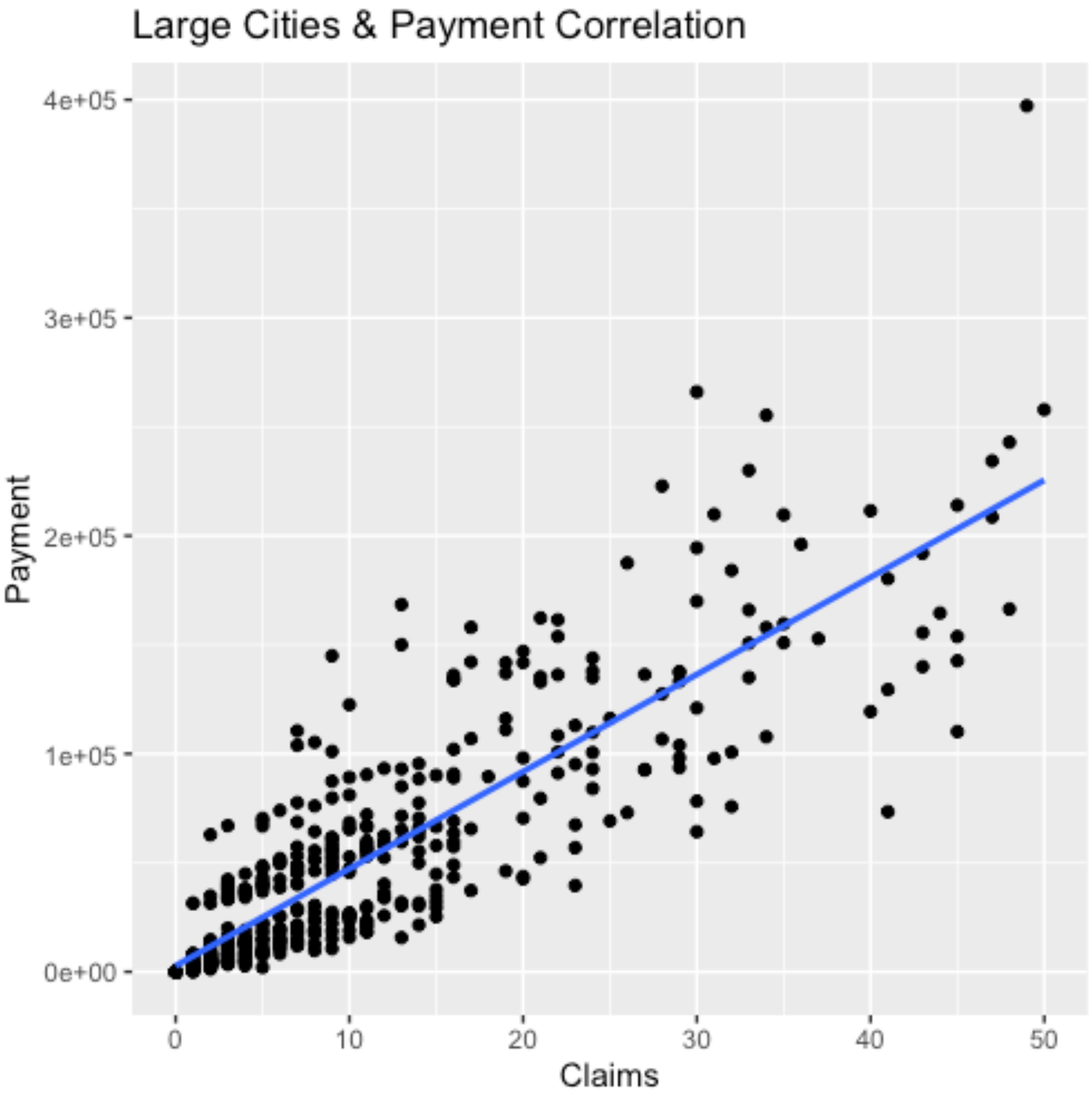


- Rural Areas (Zone 4 and Zone 6 CA = **0.9225104**)

<u>Assumptions</u>

- We can assume that the reason for Rural areas having the third-highest Correlation value is because of not only the high number of claims made but the fact that the areas that are highly correlated are Zone 4 and Zone 6 which are the rural areas of Sweden

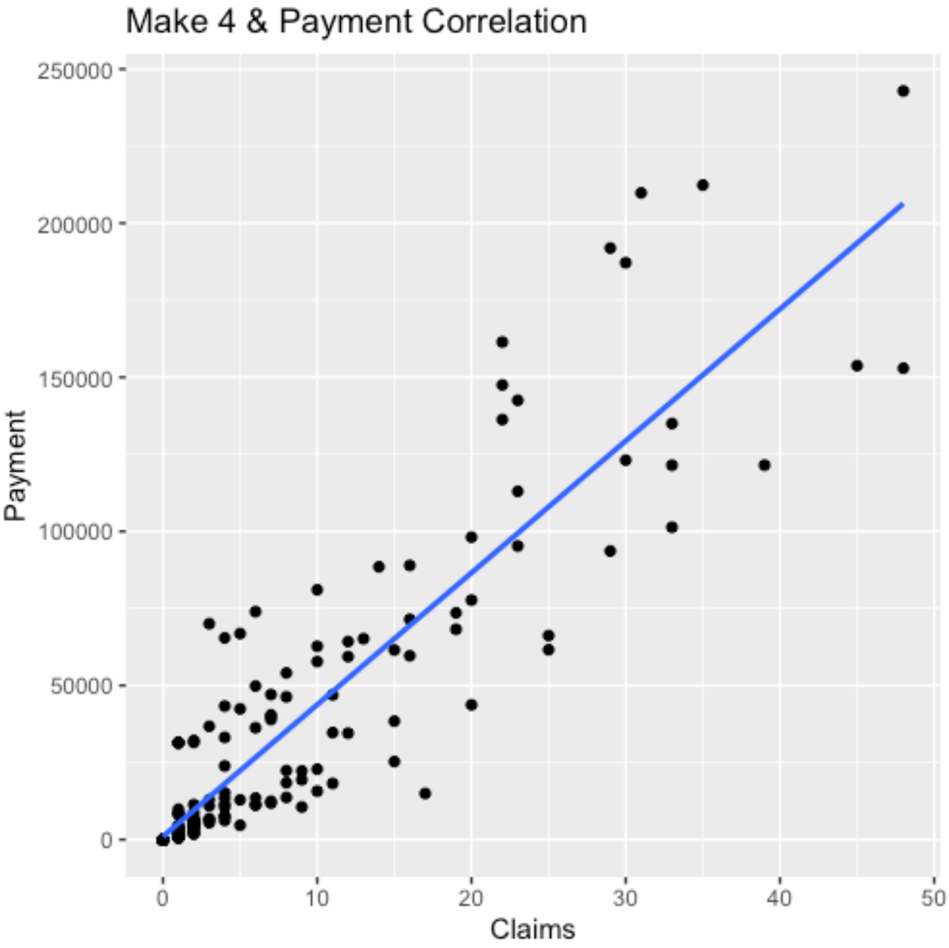
- There could be several things that could happen in the rural areas of Sweden, a number of them could be due to the infrastructure, weather, perhaps free-roaming animals in the roads, the rough terrain these a few things that could be heavily impacting the number of claims made from those areas hence our high correlation value



- Large Cites (Zone 1 and Zone 2 CA = **0.9014782**)

<u>Assumptions</u>

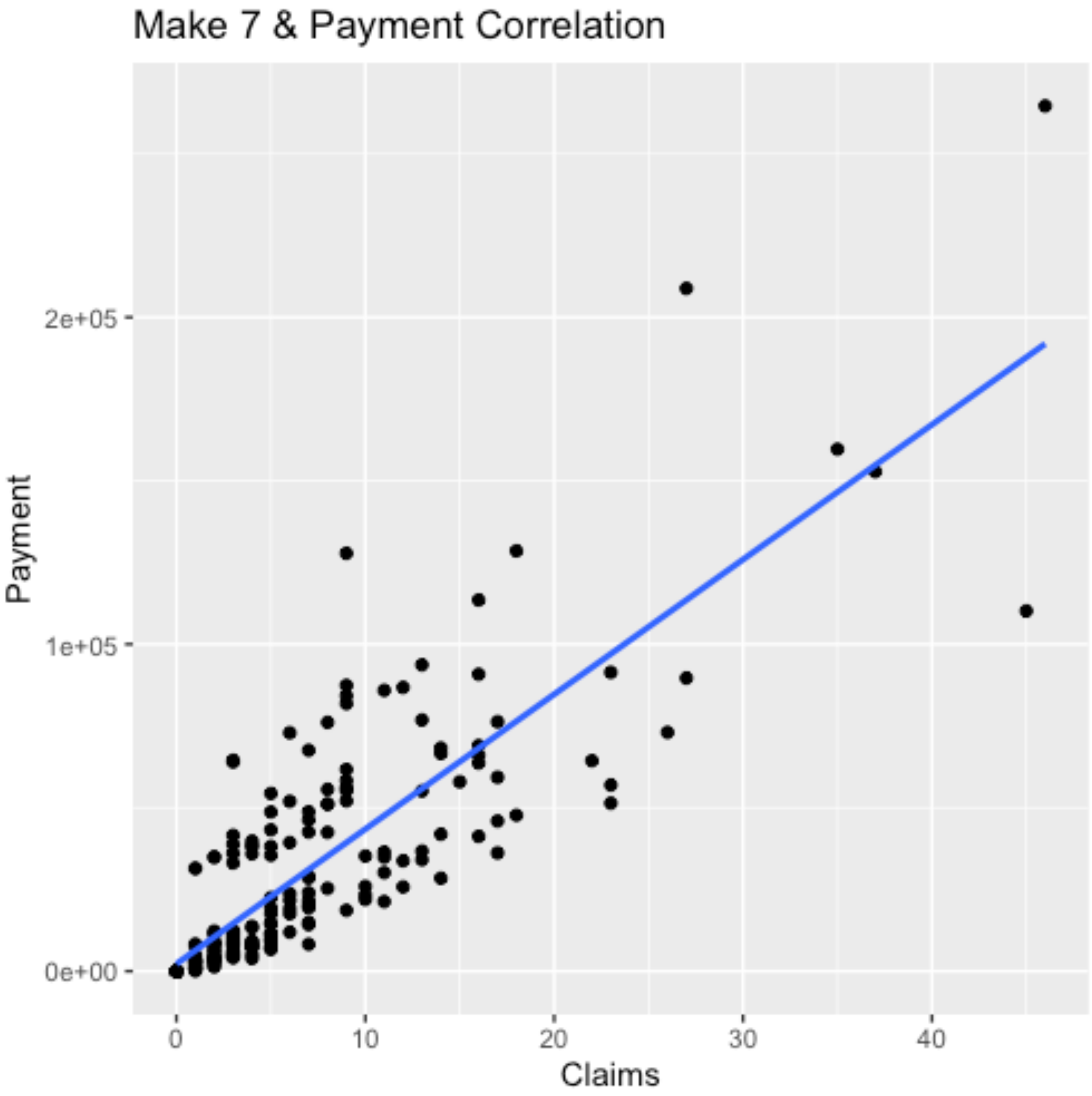
- Finally large cities had the lowest value in terms of the correlation analysis
- This could be because less people drive and take public transport
- Being that the cities are so big they are therefore less congested resulting in fewer claims
- Further adding to the above the infrastructure in terms of the roadworthy ness of the roads and the stoplights are very heavily maintained resulting in fewer claims for people living in these areas



Make #	Correlation Value
Make 4	0.9654472
Make 1	0.9374646
Make 2	0.9333129
Make 3	0.9332873
Make 9	0.9283497
Make 6	0.9254517
Make 5	0.9174462
Make 8	0.9075048
Make 7	0.9056136

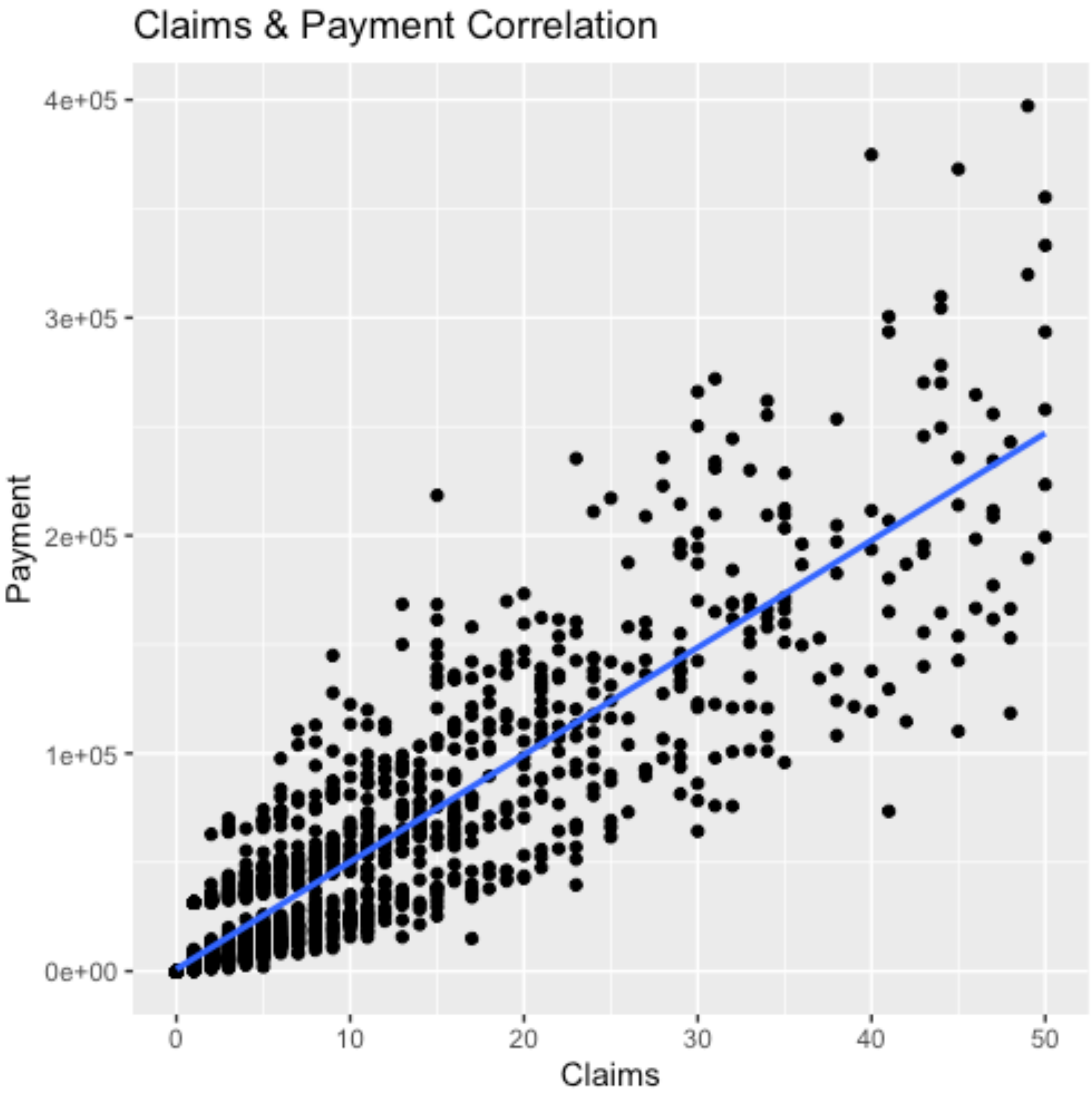
Assumptions

- Make 4 had the highest value of CA = **0.9654472** which was the highest number of claims for the Make 4 variable
- We can see from the line of best fit that the values have a lot of variances but due to the CA value of 0.9654472 we know that the variables claims and payment within Make 4 are highly correlated and this makes in particular has the highest correlation value.

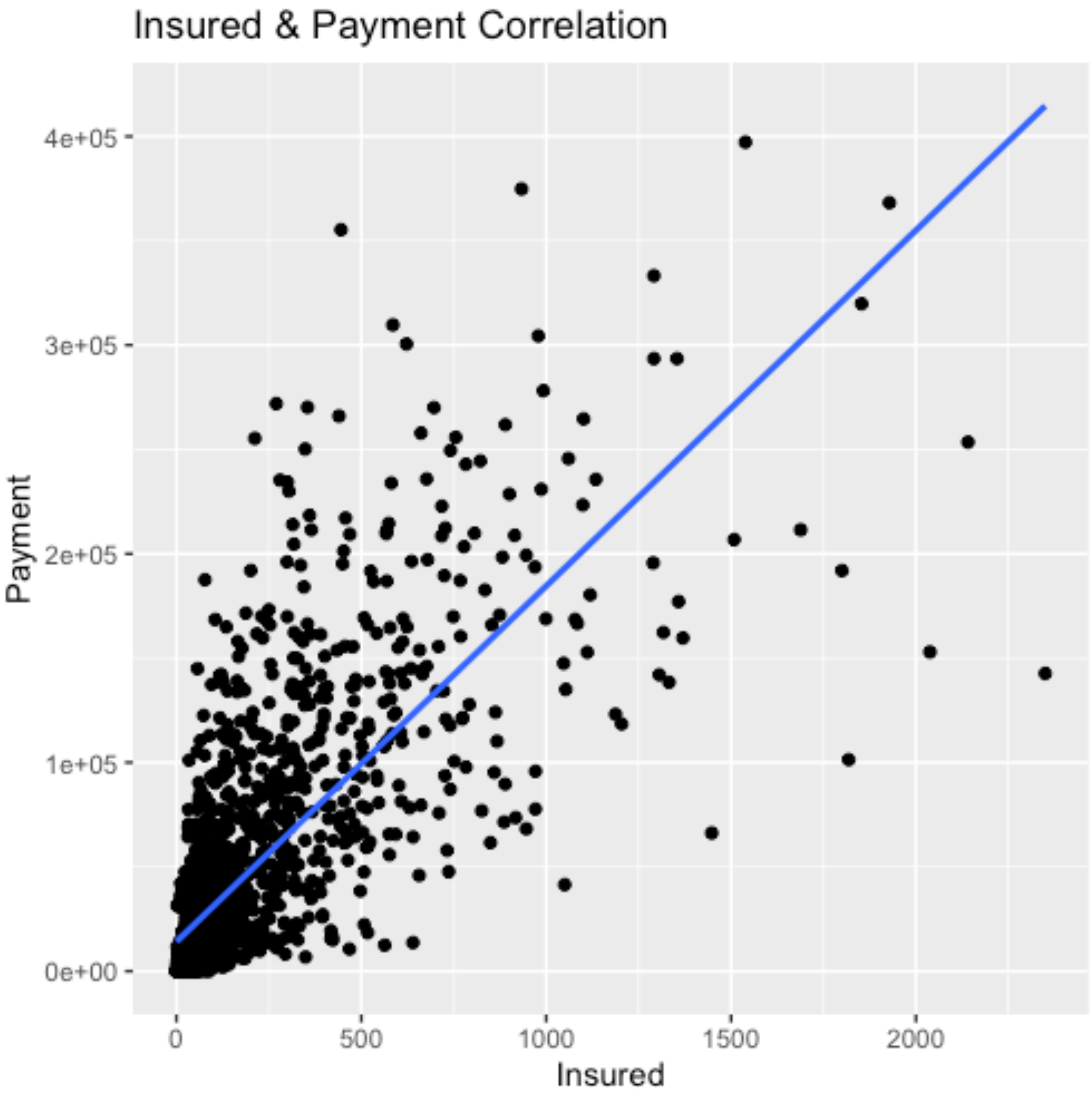


Assumptions

- Make 7 had the lowest value of **CA with = 0.9056136**
- While the correlation between the Make 7 claims and the payment the variables are very spread out and varied showing that while when the claims increase so does the payment, but the claims are not going to be consistent
- To further add to the above the claims themselves within the Make 7 have a lot of variance within them but are still highly correlated once put against the payment variable but within our dataset, this is the lowest.

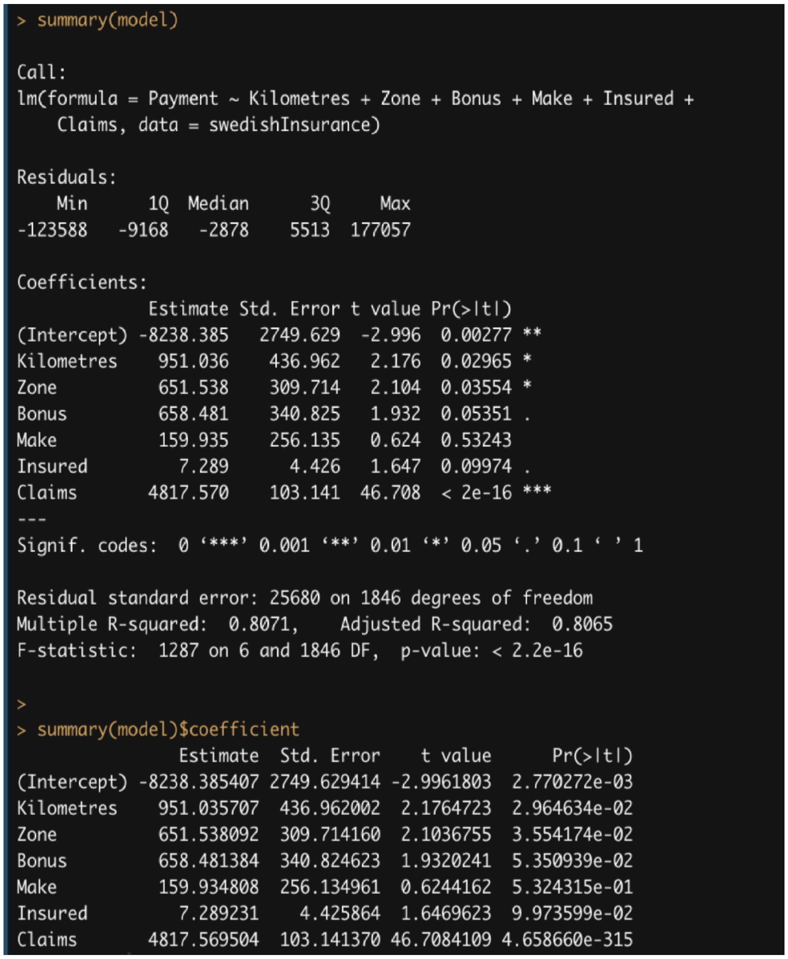


- We can see from our correlation analysis value of 0.9388519 that the claims variable and the payment variable are highly correlated resulting in a strong positive relationship, as the claims increase so does the payment made by the Swedish insurance company.



- We can see from our correlation analysis value of 0.8458319 that the insured variable and the payment variable are highly correlated resulting in a strong positive relationship, as the number of insured increase so does the payment made by the Swedish insurance company.

Payment Regression Model



```
> summary(model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-123588   -9168   -2878    5513  177057

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8238.385   2749.629  -2.996  0.00277 **
Kilometres    951.036    436.962   2.176  0.02965 *
Zone          651.538    309.714   2.104  0.03554 *
Bonus         658.481    340.825   1.932  0.05351 .
Make          159.935    256.135   0.624  0.53243
Insured         7.289      4.426   1.647  0.09974 .
Claims       4817.570    103.141  46.708  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25680 on 1846 degrees of freedom
Multiple R-squared:  0.8071,    Adjusted R-squared:  0.8065
F-statistic:  1287 on 6 and 1846 DF,  p-value: < 2.2e-16


>
> summary(model)$coefficient
                Estimate  Std. Error    t value      Pr(>|t|)
(Intercept) -8238.385407 2749.629414 -2.9961803  2.770272e-03
Kilometres    951.035707  436.962002  2.1764723  2.964634e-02
Zone          651.538092  309.714160  2.1036755  3.554174e-02
Bonus         658.481384  340.824623  1.9320241  5.350939e-02
Make          159.934808  256.134961  0.6244162  5.324315e-01
Insured         7.289231    4.425864  1.6469623  9.973599e-02
Claims       4817.569504  103.141370 46.7084109 4.658660e-315
```

Variable	Std. Error
Insured	4.425864
Claims	103.141370
Make	256.134961
Zone	309.714160
Bonus	340.824623
Kilometers	436.962002

Variable	P-Value
Kilometers	2.770272e-03
Zone	2.964634e-02
Bonus	3.554174e-02
Insured	5.324315e-01
Make	5.350939e-02
Claims	9.973599e-02

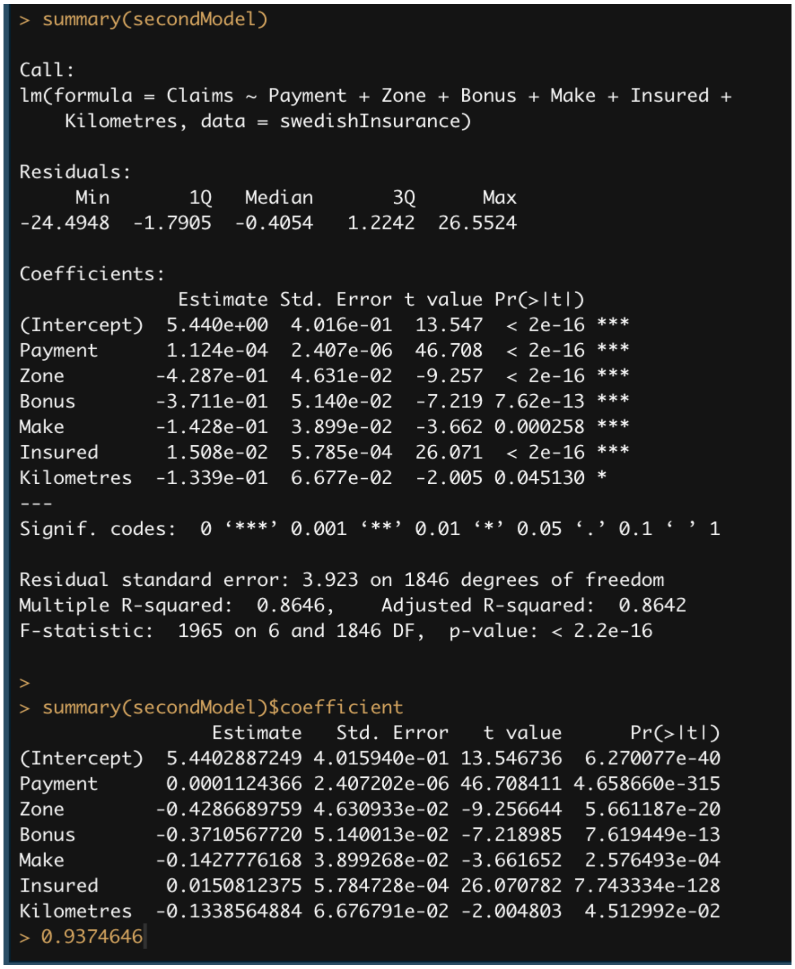
- Std.Error: the standard error of the coefficient estimates. As we can see the **Insured** variable is the smallest as this represents the accuracy of the coefficient. Because **Insured** has the smallest Std.Error this shows that it has the highest level of confidence regarding the estimate. Furthermore, **Kilometres** in regard to the Std.Error this shows that it has the lowest level of confidence.

- Pr(>|t|): The p-value corresponding to the t-statistic. Because **Kilometres** has the smallest p-value, it shows it is the most significant estimate, additionally **Claims** has the highest p-value this shows that this has the least significant estimate for our model.

Model Equation

Payment = -8238.38 + 951.03 Kilometers + 651.53 Zone +
658.48 Bonus + 159.93 Make + 7.28 Insured + 4817.56 Claims

Claims Regression Model



```
> summary(secondModel)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-24.4948  -1.7905  -0.4054   1.2242  26.5524 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.440e+00  4.016e-01  13.547  < 2e-16 ***
Payment      1.124e-04  2.407e-06  46.708  < 2e-16 ***
Zone        -4.287e-01  4.631e-02  -9.257  < 2e-16 ***
Bonus       -3.711e-01  5.140e-02  -7.219 7.62e-13 ***
Make        -1.428e-01  3.899e-02  -3.662 0.000258 ***
Insured      1.508e-02  5.785e-04  26.071  < 2e-16 ***
Kilometres  -1.339e-01  6.677e-02  -2.005 0.045130 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.923 on 1846 degrees of freedom
Multiple R-squared:  0.8646,	Adjusted R-squared:  0.8642 
F-statistic:  1965 on 6 and 1846 DF,  p-value: < 2.2e-16


> 
> summary(secondModel)$coefficient
                 Estimate   Std. Error   t value      Pr(>|t|)
(Intercept)  5.4402887249 4.015940e-01 13.546736  6.270077e-40
Payment      0.0001124366 2.407202e-06 46.708411 4.658660e-315
Zone        -0.4286689759 4.630933e-02 -9.256644  5.661187e-20
Bonus       -0.3710567720 5.140013e-02 -7.218985  7.619449e-13
Make        -0.1427776168 3.899268e-02 -3.661652  2.576493e-04
Insured      0.0150812375 5.784728e-04 26.070782 7.743334e-128
Kilometres  -0.1338564884 6.676791e-02 -2.004803  4.512992e-02
> 0.9374646
```

Variable	Std. Error
Payment	2.407202e-06
Make	3.899268e-02
Zone	4.630933e-02
Bonus	5.140013e-02
Insured	5.784728e-04
Kilometers	6.676791e-02

Variable	P-Value
Make	2.576493e-04
Kilometers	4.512992e-02
Payment	4.658660e-315
Zone	5.661187e-20
Bonus	7.619449e-13
Insured	7.743334e-128

- Std.Error: the standard error of the coefficient estimates. As we can see the **Payment** variable is the smallest as this represents the accuracy of the coefficient. Because **Payment** has the smallest Std.Error this shows that it has the highest level of confidence regarding the estimate. Furthermore, **Kilometres** in regard to the Std.Error this shows that it has the lowest level of confidence.

- Pr(>|t|): The p-value corresponding to the t-statistic. Because **Make** has the smallest p-value, it shows it is the most significant estimate, additionally **Insured** has the highest p-value this shows that this has the least significant estimate for our model.

Model Equation

Claims = 5.44 + 0.0011 Payment + - 0.42 Zone + - 0.37 Bonus
+ - 0.14 Make + 0.01 Insured + - 0.13 Kilometers

Conclusion

From my exploratory data analysis, I have been able to show which variables have the highest or the lowest impact on the claims & Payment variable and further have been able to show which risk categories have the highest and lowest level of impact on the claims. Going forward with this piece of presented information the Swedish Committee is now able to make the necessary adjustment to how they protect their customers depending on the risk categories that have been supplied by the committee.