



JULY 24, 2021


DATA ANALYTICS & MANAGEMENT

ASSIGNMENT 1: ANALYSIS OF A MOTOR INSURANCE COMPANY

BRIAN COLLINS

Student ID: 2078876

Word Count (excl. tables): 1606



Introduction

The purpose of this report is to analyse Swedish third-party motor insurance claims. This report will describe the data contained therein and provide descriptive statistics of the underlying variables before conducting a correlation analysis. Finally, this report will produce a linear regression analysis to create models that can be applied in estimating claim numbers and payment values to aid in financial and resource planning.

Dataset Overview

Kilometres – distance groups in order from smallest to largest, represented as an integer. Distance groups 1 and 2 exhibit the most significant average claims, with mean claims per group decreasing for distance groups 3-5. This is likely due to fewer, more professional drivers displaying markers of decreased risk (Figure 1).

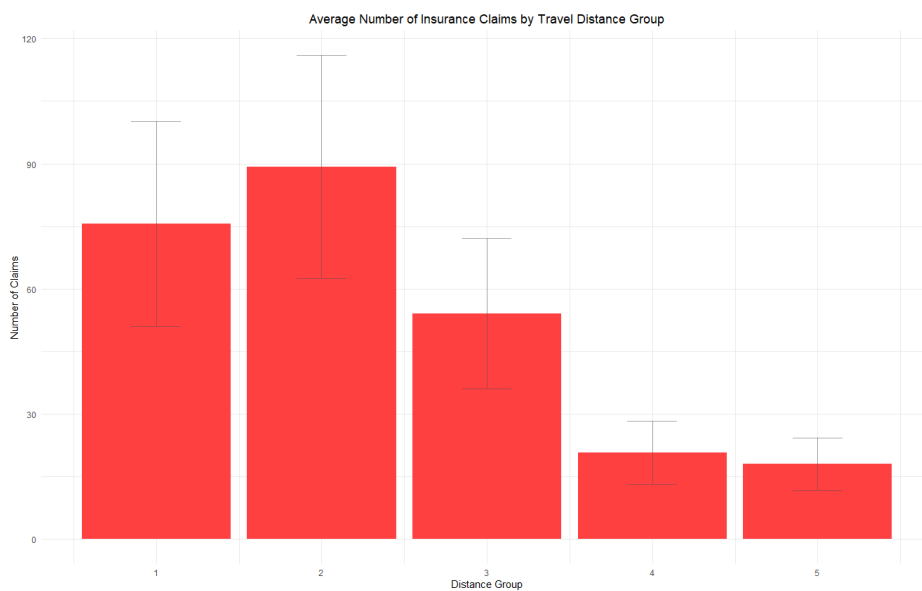


Figure 1: Bar chart showing the average number of claims by distance group

Zone - In a broadly descending order of size, zone-clustering regions of the country are represented as an integer (figure 2). Zones 1-4 present the largest share of the number of claims, with the average number of claims reducing with distance from Sweden's largest cities. However, zone 4, representing rural areas in southern Sweden, displays a disproportionately large number of claims. This is potentially due to the combination of faster roads and wildlife found in the agricultural south of the country.

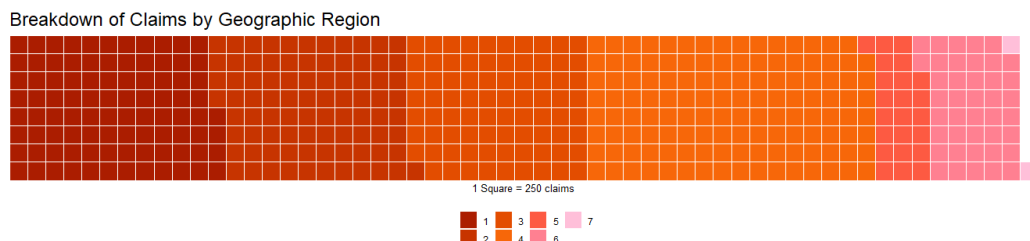


Figure 2: Waffle chart of claims by Zone, one square = 250 claims

Figure 4 illustrates that the average value of payments follows a broadly similar trend to the number of claims; again 4th Zone displays a significant average value. This may be because of the prevalence of expensive agricultural equipment found in this region, with claims for even minor accidents resulting in significant payments.

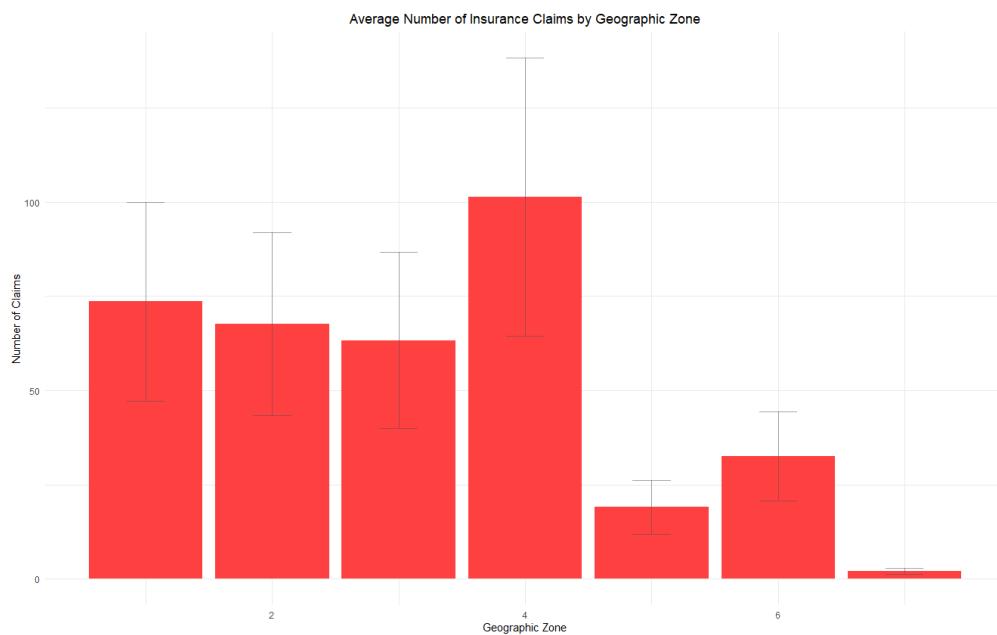


Figure 3: Average number of claims by geographic Zone.

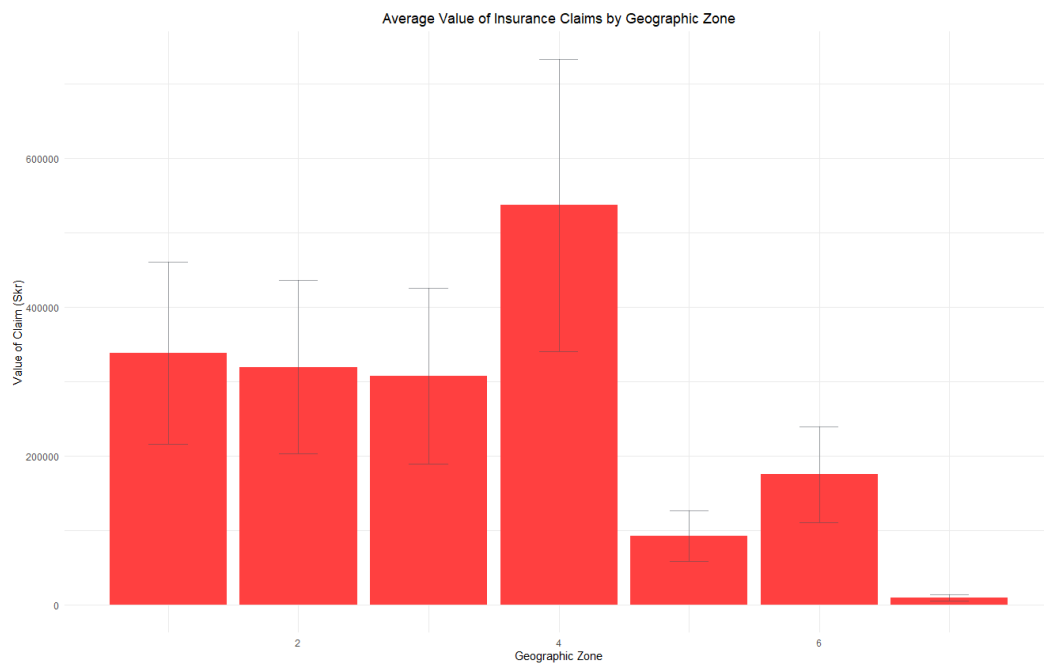


Figure 4: Average value of claims by geographic Zone

Bonus – representing the number of years no claims bonus (NCB) since the last claim, represented as an integer. Figure 5 presents the average number of claims by years NCB, displaying increased risk for drivers in their first year. The number of claims rises for drivers in the later clusters, logically due to the number of drivers in group 7, illustrated in Table 1.

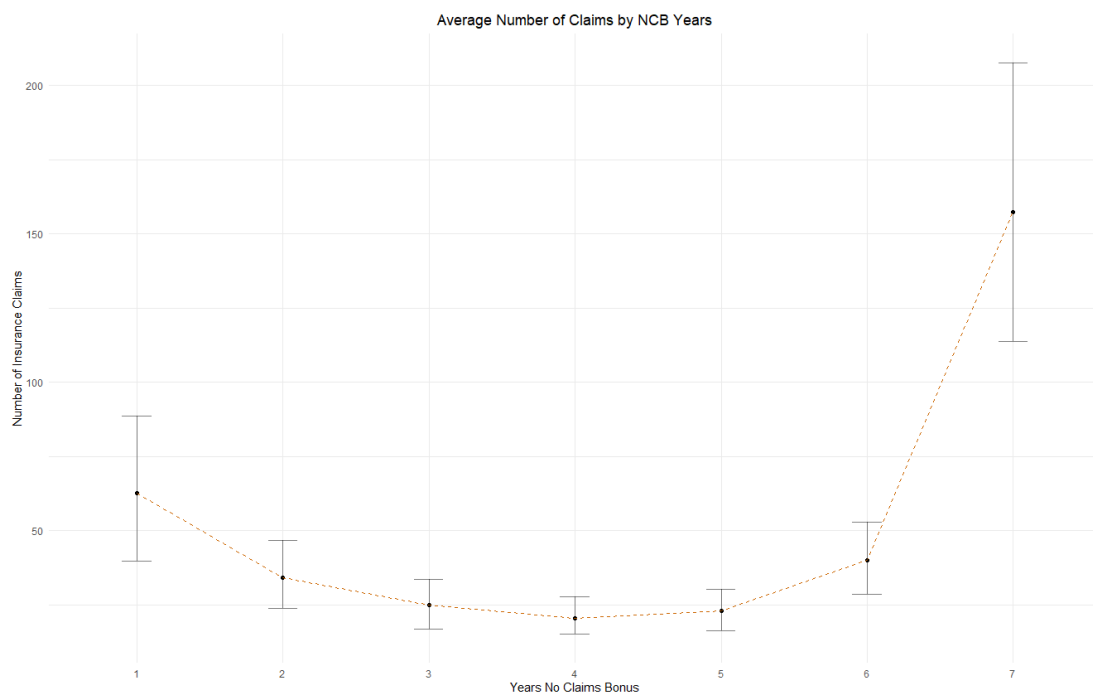


Figure 5: Average number of claims by years NCB

Bonus	Avg. Insured	Avg. Claims	Avg. Payment (Skr)
1	525.55	62.50	282,921.99
2	451.08	34.23	163,316.63
3	397.47	24.97	122,656.17
4	360.39	20.35	98,498.12
5	437.39	22.82	108,790.50
6	805.82	39.94	197,723.82
7	4620.37	157.22	819,322.48

Table 1: Average values for insured, claims and payment grouped by NCB

Make – Integers representing the eight most popular manufacturers; other vehicles are clustered into group nine. Figure 6a illustrates the significant number of vehicles that fall outside of the central cluster of 8 manufacturers, with several outliers falling in that category concerning claims. Figure 6b presents an exploded view of the boxplots facilitating an understanding of vehicle claims by Make, illustrating that the number of claims is significantly higher for manufacturer group 1.

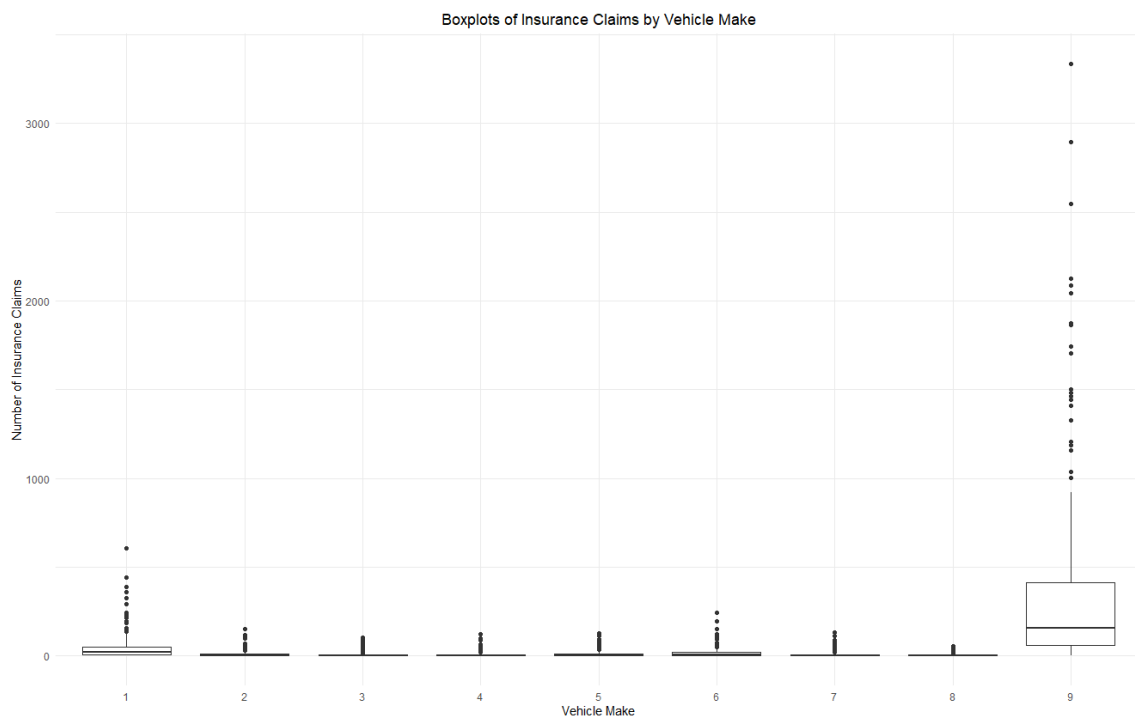


Figure 6(a): Boxplot showing claim number quartiles, whiskers incl. outliers

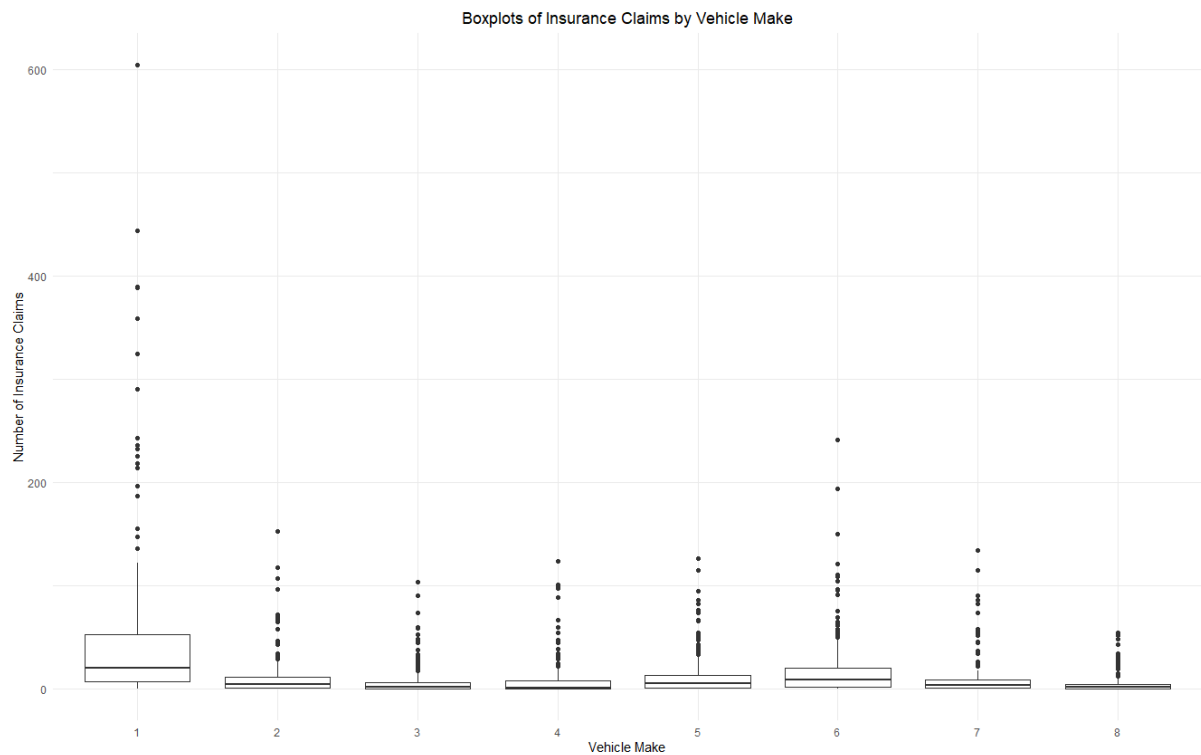


Figure 6(b): Boxplot showing claim number quartiles, whiskers excl. outliers

Insured – Numerical data representing the number of the insured policy years. Figure 7 displays a scatter of insured policy years against payments, clearly illustrating a positive linear relationship between variables, also mirrored between claims and Payment, as seen in figure 8.

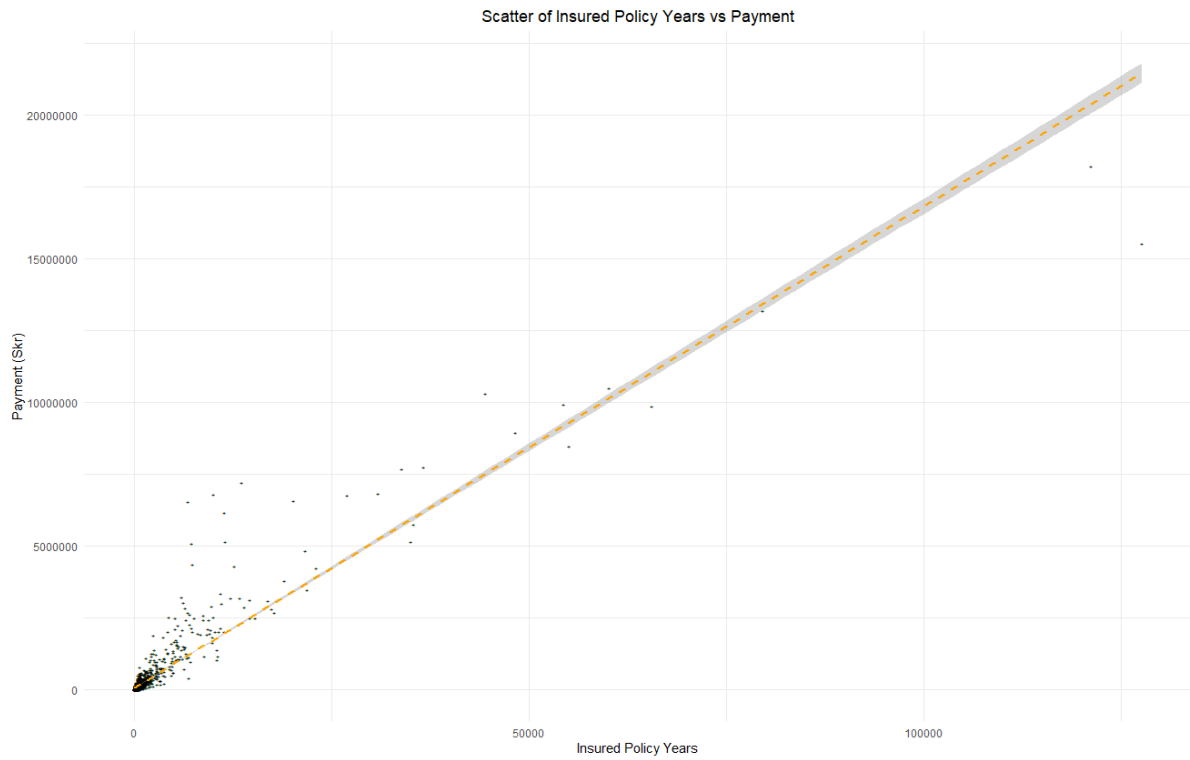


Figure 7: Scatter graph of insured policy years and Payment

Claims – Integer representing the number of claims made against insurance policies.

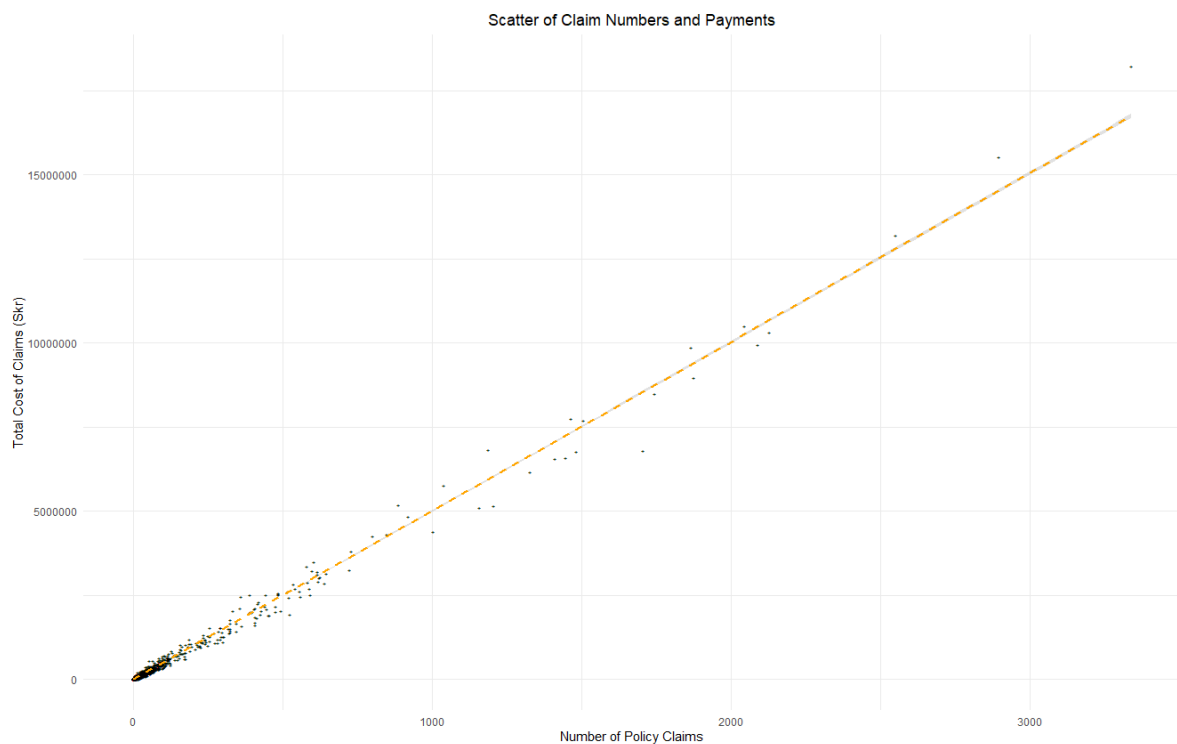


Figure 8: Scatter of policy claims vs payments

Payment – Total value of insurance payouts made against policies in Swedish Krona (Skr), represented as an integer.

Descriptive statistics for non-categorical variables can be found in Table 2, measures of central tendency (Table 3) in addition to frequency distribution plots for continuous variables, can be seen in figures 9, 10 & 11. In all cases, the distribution of continuous variables are significantly positively skewed with a leptokurtic distribution.

	Min	Max	Range	Sum	Med	Mean	CI of Mean (95%)	Var	Std. Dev
Insured	0.01	127,687.27	127,687.26	2,383,170.08	81.525	1,092.1953	237.6659	32,048,690.03	5,661.1562
Claims	0	3,338	3,338	113,171	5	51.8657	8.4682	40,687.2039	201.7107
Payment	0	18,245,026	18,245,026	560,790,681	27,403.5	257,007.6448	42,707.4271	1.03486E+12	1,017,282.586

Table 2: Descriptive statistics for insurance dataset

Variable	Skewness	Kurtosis
Payment	9.11	108.66
Claims	8.57	93.15
Insured	13.95	249.92

Table 3: Measures of Central Tendency

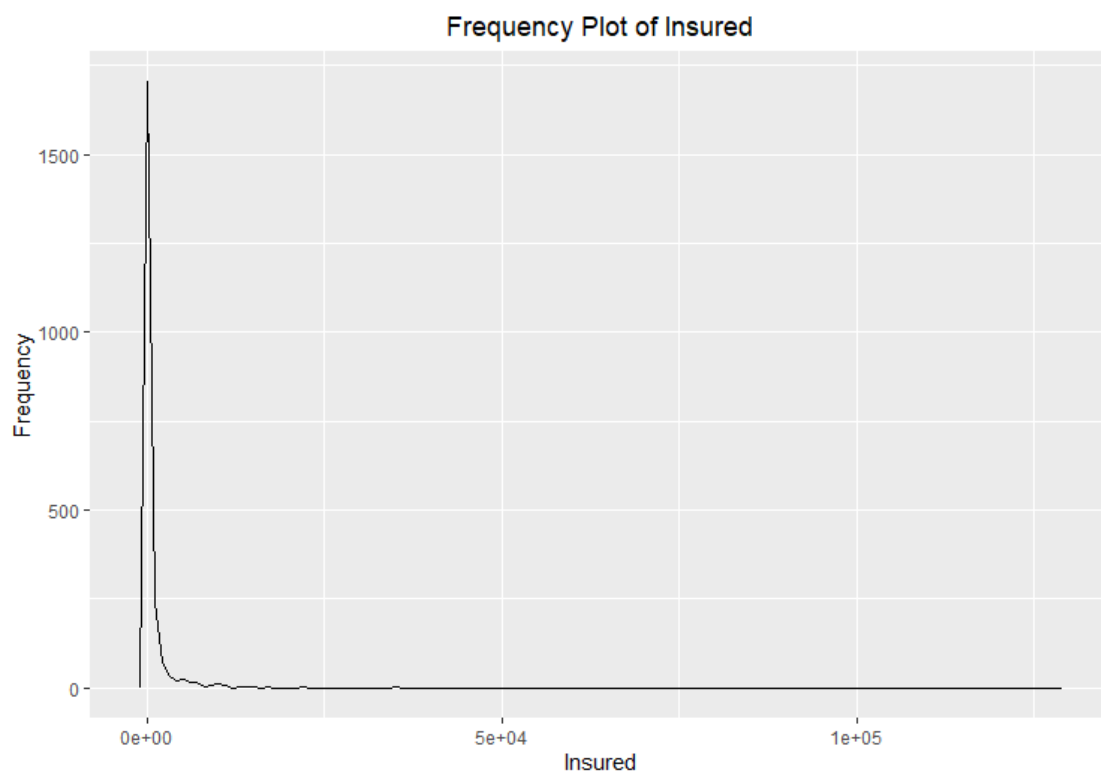


Figure 9: Frequency Distribution of Insured

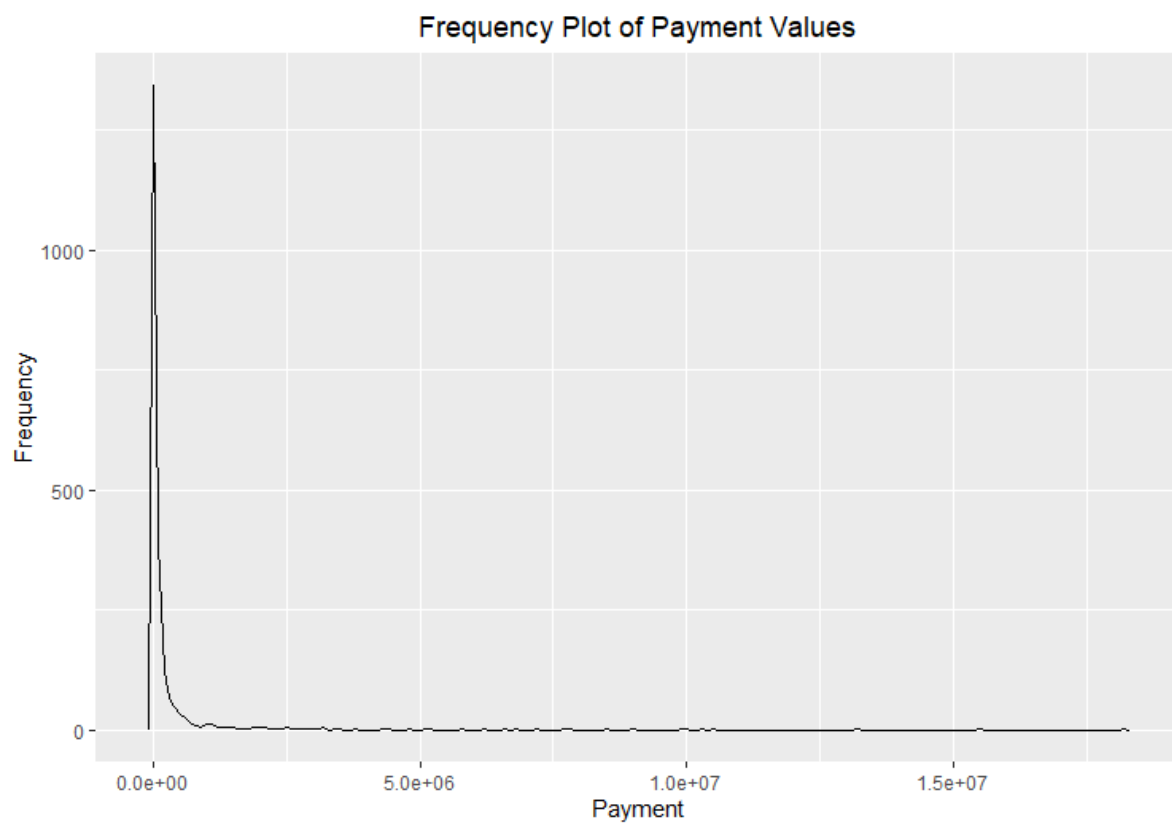


Figure 10: Payment

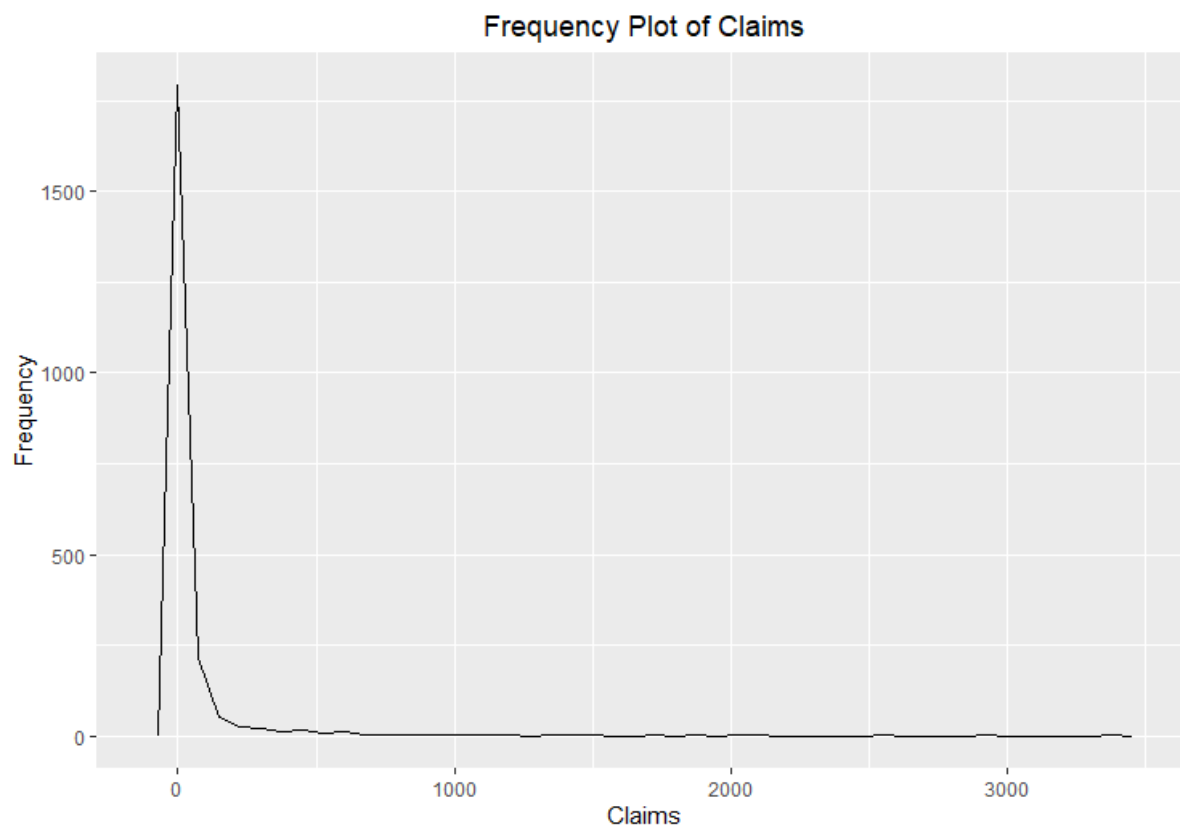


Figure 11: Frequency Distribution of Claims

Establish whether there is a preexisting relationship between insurance payments amount and other tracked variables.

Method

Before undertaking any statistical correlation testing, it was first necessary to perform a visual analysis of key variables using scatter plots to determine the underlying relationships. However, these scatter plots were only helpful in comparing parametric data, and as such, Kilometers, Zone, Bonus, and Make, as ordinal data are more straightforward to infer through bar charts.

A Shapiro-Wilk was used to confirm the probability of data having been sampled from a normally distributed population. The test indicated that this probability was $2.2e-16$ (below the 0.05 confidence interval) for all variables. The null hypothesis, stating the data was sampled from a normally distributed population, can be rejected. With this, Pearson could be rejected as an appropriate test of correlation between the data. As a result of this, Spearman's Rho was selected given the previously established monotonic relationship, the non-normalcy of the data distribution, and variable types.

Analysis

Table 4 represents a matrix of Spearman's Correlation Coefficients between all variables in the dataset. Payment has a strong positive correlation with both claim number and the number of insured in policy years, with values of 0.9030 and 0.9624, respectively. Bonus and Make also both had weak positive correlations with Payment, with 0.2021 and 0.1182, respectively. Finally, Kilometres and Zone both had moderate negative correlations of -0.2418 and -0.3634, respectively.

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment
Kilometres	1	-0.01393	0.007206	-0.00268	-0.32902	-0.26426	-0.24218
Zone	-0.01393	1	0.011674	-0.00519	-0.32006	-0.38682	-0.36345
Bonus	0.007206	0.011674	1	0.002157	0.351141	0.197773	0.202058
Make	-0.00268	-0.00519	0.002157	1	0.111041	0.112388	0.118209
Insured	-0.32902	-0.32006	0.351141	0.111041	1	0.933337	0.903032
Claims	-0.26426	-0.38682	0.197773	0.112388	0.933337	1	0.962443
Payment	-0.24218	-0.36345	0.202058	0.118209	0.903032	0.962443	1

Table 4: Spearman's Rho correlation coefficients

P-Values for the correlation analysis can be found in Table 5. In all cases, calculated values were less than 0.05, and as such, the null hypothesis, stating that no relationship is present, can be rejected.

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment
Kilometres	NA	0.515516	0.736568	0.900551	0	0	0
Zone	0.515516	NA	0.58573	0.80857	0	0	0
Bonus	0.736568	0.58573	NA	0.919784	0	0	0
Make	0.900551	0.80857	0.919784	NA	1.99E-07	1.41E-07	3.06E-08
Insured	0	0	0	1.99E-07	NA	0	0
Claims	0	0	0	1.41E-07	0	NA	0
Payment	0	0	0	3.06E-08	0	0	NA

Table 5: Spearman's Rho correlation p-values

Conclusion

These correlations allow us to outline the following hypotheses:

1. As the number of policy claims increases, so to do the total value of payouts made.
2. The number of insured policy years for a given Zone, distance, Bonus, and Make displays a significant positive correlation with the value of insurance payouts.
3. Smaller distance groups display higher risk markers, likely due to the smaller number of professional drivers covering the longer distances.
4. Insurance payments increase with proximity to Sweden's largest cities. However, the correlation coefficient is comparatively weak because of the significant number of claims and the value of payments in rural Zone 4 (figures 3 & 4).
5. There is a weak positive correlation for both Bonus and Make with Payments. However, this is likely due to large payments values associated with the largest categories of both variables.

Determine which combination of variables provides the most accurate prediction of insurance payouts and claims to assist in financial modelling

Method

Initially, a generalised regression model was created using all potential independent variables for both targets. This model served as a basis for a stepwise approach to determining possible models for further analysis. Outputs from the stepwise approach were then compared to determine which predictor variables were best suited for implementation.

The first phase of testing was based on establishing multicollinearity between independent variables in the regression model by ascertaining the variance inflation factor for each model. Where high levels of multicollinearity existed, models were further split into further testing models. These additional testing models were compared both in terms of their residuals, R^2 (incl. adjusted), F-Statistic, and Cooks Distance to ascertain the effect of outliers within the dataset, alongside a final ANOVA comparison of both models.

Based on the comparison of fit described above, the most accurate model was selected for further use.

Analysis

Payment Model

Direction	Suggested Variables
Both & Forward	Claims + Insured + Kilometres + Zone + Bonus – AIC 48744
Backward	Kilometres + Zone + Bonus + Insured + Claims – AIC 48744

From this, a testing model was selected. (Claims + Insured + Kilometres + Zone + Bonus)

<i>R²</i>	<i>F-Statistic</i>	<i>P-Value</i>	<i>Mean VIF</i>
0.9952	8.952e+04 on 5 and 2176 DF	< 0.00000000000000022	3.048265

Insured Model: *Insured + Kilometres + Zone + Bonus*

<i>R²</i>	<i>F-Statistic</i>	<i>P-Value</i>	<i>Mean VIF</i>	<i>% obs Residual >2.58</i>	<i>Std. Cooks Distance Outliers</i>
0.8748	3802 on 4 and 2177 DF	< 0.00000000000000022	1.023519	1.6	2

Claims Model: *Claims + Kilometres + Zone + Bonus*

<i>R²</i>	<i>F-Statistic</i>	<i>P-Value</i>	<i>Mean VIF</i>	<i>% obs Residual >2.58</i>	<i>Std. Cooks Distance Outliers</i>
0.9912	6.116e+04 on 4 and 2177 DF	< 0.00000000000000022	1.022149	0.82	3

Claims

Direction	Suggested Variables
Backward, Forward & Both	Payment + Insured + Kilometres + Zone + Make + Bonus – AIC 12128

<i>R²</i>	<i>F-Statistic</i>	<i>P-Value</i>	<i>Mean VIF</i>
0.9937	5.685e+04 on 6 and 2175 DF	< 0.00000000000000022	3.436533

Insured Model: *Insured + Kilometres + Zone + Bonus*

<i>R²</i>	<i>F-Statistic</i>	<i>P-Value</i>	<i>Mean VIF</i>	<i>% obs Residual >2.58</i>	<i>Std. Cooks Distance Outliers</i>
0.8425	2328 on 5 and 2176 DF	< 0.00000000000000022	1.034239	1.56	2

Payment Model: *Payment + Kilometres + Zone + Bonus*

<i>R</i> ²	<i>F</i> -Statistic	<i>P</i> -Value	Mean VIF	% obs Residual >2.58	Std. Cooks	Distance Outliers
0.9913	4.972e+04 on 5 and 2176 DF	< 0.00000000000000022	1.044052	2.11	3	

Conclusion

The claims model (Claims + Kilometres + Zone + Bonus) provides a superior prediction of the target variable when modelling for payments. Modelling is statistically significant, and error is kept within acceptable limits, with 0.82% of observations having a standardised residual > 2.58. Figure 12 shows a plot of model residuals against predicted values. When interpreting this scatter, it is crucial to consider that a random distribution around zero indicates a normal distribution.

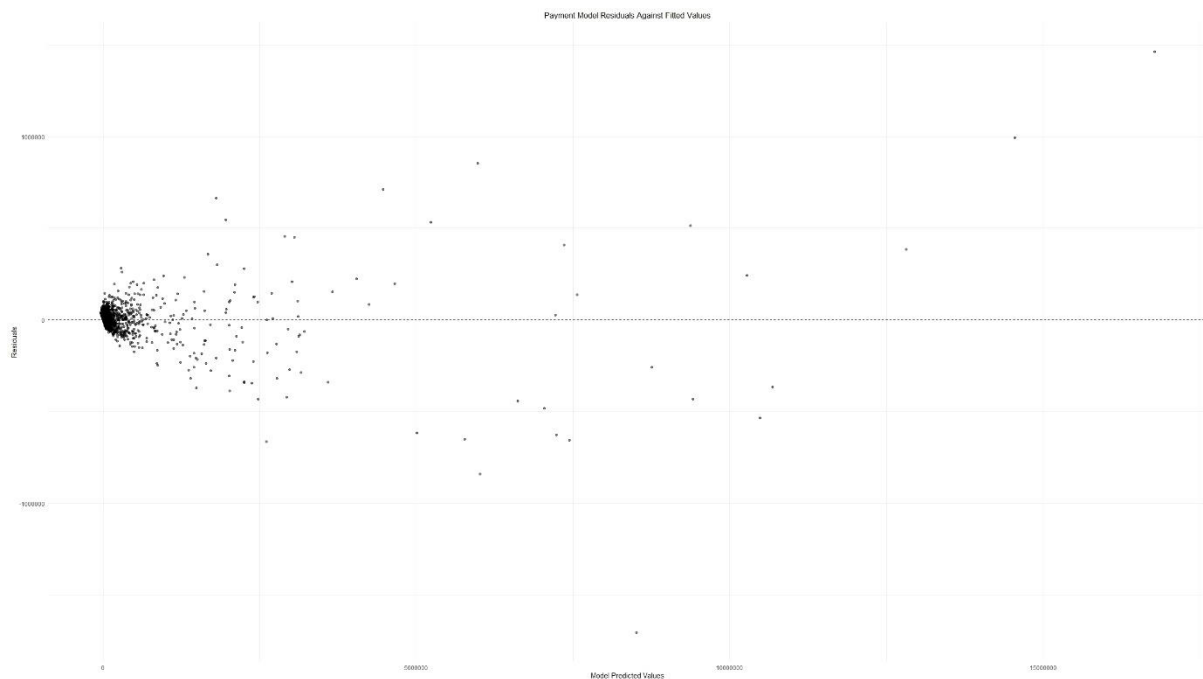


Figure 12: Plotted residuals for Final Payment Model

In the case of claims modelling, however, no clear model can be identified. A contributing factor to this lies in the significant outlier values of two observations representing low mileage, vehicle type and rural areas in southern Sweden. This could indicate an error in the underlying data, or potentially given the prominence of the agricultural industry in southern Sweden, could be valid given the prevalence of expensive farm equipment representing significant insured values.

Concerning outlier values raised by the Cook's Distance test in both cases (available in appendix 6 – 9), these observations could be removed to reduce their impact on the model. However, if the values were valid, the impact of this on financial planning could be significant. Therefore further investigation into the validity of these observations is necessary.

Finally, while the payment model is undoubtedly more accurate in terms of its residuals between predicted and observed values (Figure 13), the inclusion of Payment as an independent variable is only helpful where that data exists. When conducting financial forecasting for future potential payments, the Insured model would still provide statistically significant and accurate predictions based on residual plots of original data (Figure 14).

Concerning plotting residuals, this clearly illustrates an underlying heteroskedasticity in the data, whereby error does not remain consistent throughout the range of predicted values. The target variable's associated data were transformed using different methods to account for this, with no success. Full details of this and associated residual plots are available in appendix 1.

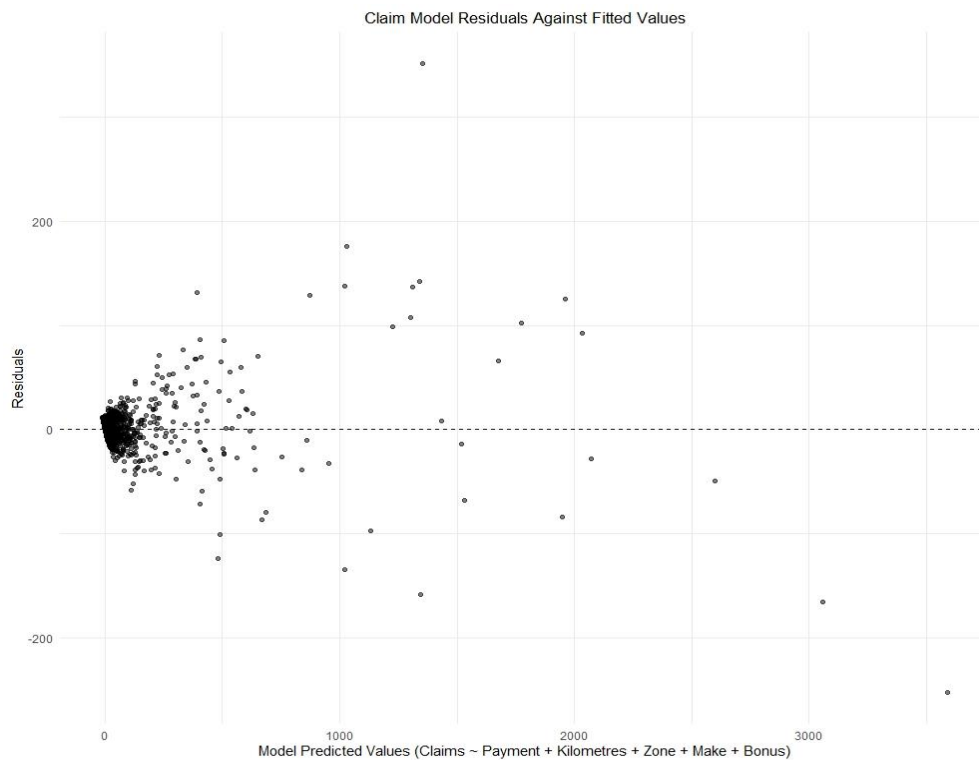


Figure 13: Plotted residuals for Final Claims Model

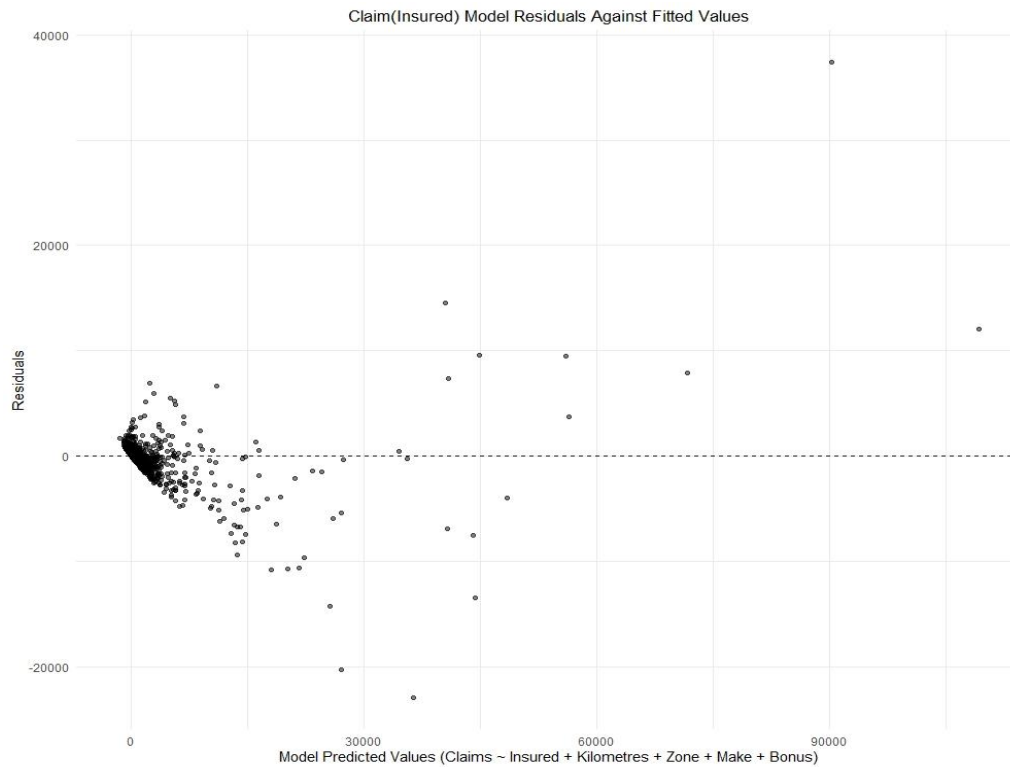


Figure 14: Plotted residuals for Insured (Test) Model

The committee plans to extend their coverage over a few more cities/areas shortly and predict their payments and the number of claims.

Method

In order to predict claim numbers and payment values, a new model needed to be created using Zone, Kilometre, Bonus, Make & Insured, given that not all independent variables in the models outlined above had values supplied (*Input values for the prediction can be seen in Appendix 6*). Fit analysis for this model can be seen in Appendix 2 (Payment Model) and Appendix 3 (Claims Model). The same caveats outlined in the work above concerning outlier values apply here.

Analysis

Predicted values given data points can be seen below in Table 6.

Location	Claims	Payment
Vittangi	156.8249	805192.1
Halmstad	374.0307	1867116
Uppsala (lower)	569.7284	2951236
Uppsala (higher)	822.0087	4260746

Table 6: Predicted values for claim and Payment using Appendix 4 values

Conclusion

The impact that these new values would have on their associated groups can be seen in Table 7. Values were established for both the lower and upper bound of the predicted Uppsala Insured variable, and the actual final value will likely reside somewhere within this range.

Location	Current Claim	Predicted New Claims	Claim Increase %age	Current Payment	Predicted New Payment	Payment Growth %age
Vittangi	3	156.8249	51.27496	8813	805192.1	90.36413
Halmstad	1157	374.0307	-0.67672	5121951	1867116	-0.63547
Uppsala (lower)	2	569.7284	283.8642	1916	2951236	1539.311
Uppsala (higher)	2	822.0087	410.0044	1916	4260746	2222.772

Table 7: Current, predicted, new total and percentage increase values based on predictions

The insurance company plans to establish a new branch office, so they are interested in finding at what location, kilometre, and bonus level their insured amount, claims, and payment gets increased. Use developed models to do this task.

Tables 8, 9 & 10 display coefficient values for independent variables in regression models focused on predicting Claims, Payments and Insured values, respectively. These coefficient values illustrate to what degree each predictor affects the outcome if all other predictors remain constant. However, while these values facilitate understanding how changes in an independent variable impact the target variable, they are still representative of the unit of measure of the underlying independent variable. As such, standardised beta represents an easier to grasp indication of the importance of a predictor variable in the underlying model, illustrating the number of standard deviations by which the dependent variable will change when the predictor changes by one standard deviation. Each table has been ordered by the absolute standardised beta value to illustrate the importance of each variable in its associated model.

Independent Variable	Model Coefficient	Standardised Beta
Payment	0.000196617	0.99159103
Zone	-1.296011876	-0.01277861
Bonus	-1.184765672	-0.01175021
Make	0.908853427	0.01165606
Kilometres	-1.236444403	-0.00864551

Table 8: Coefficients and standardised coefficients for claims regression model

Independent Variable	Model Coefficient	Standardised Beta
Claims	5024.4	0.996258187
Bonus	6680.12	0.013136659
Zone	5886.98	0.011509448
Kilometres	5158.24	0.007151628

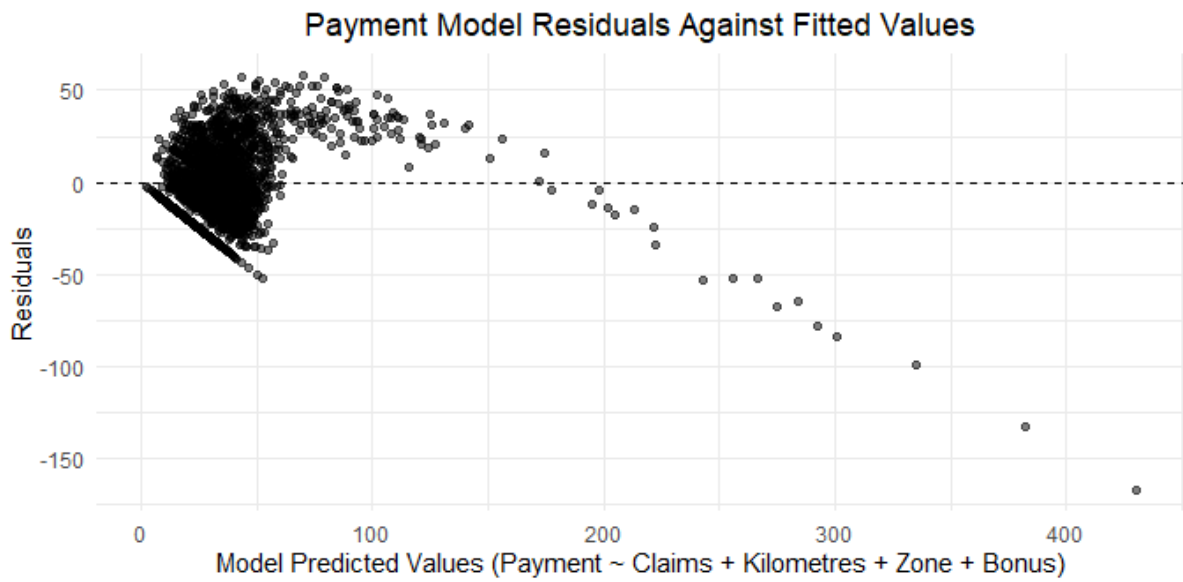
Table 9: Coefficients and standardised coefficients for payment regression model

Independent Variable	Model Coefficient	Standardised Beta
Payment	0.0161175	2.89623981
Claims	-55.3689629	-1.9728323
Bonus	87.9712821	0.03108693
Make	-44.3630424	-0.0202723
Kilometres	-66.4985759	-0.01656732

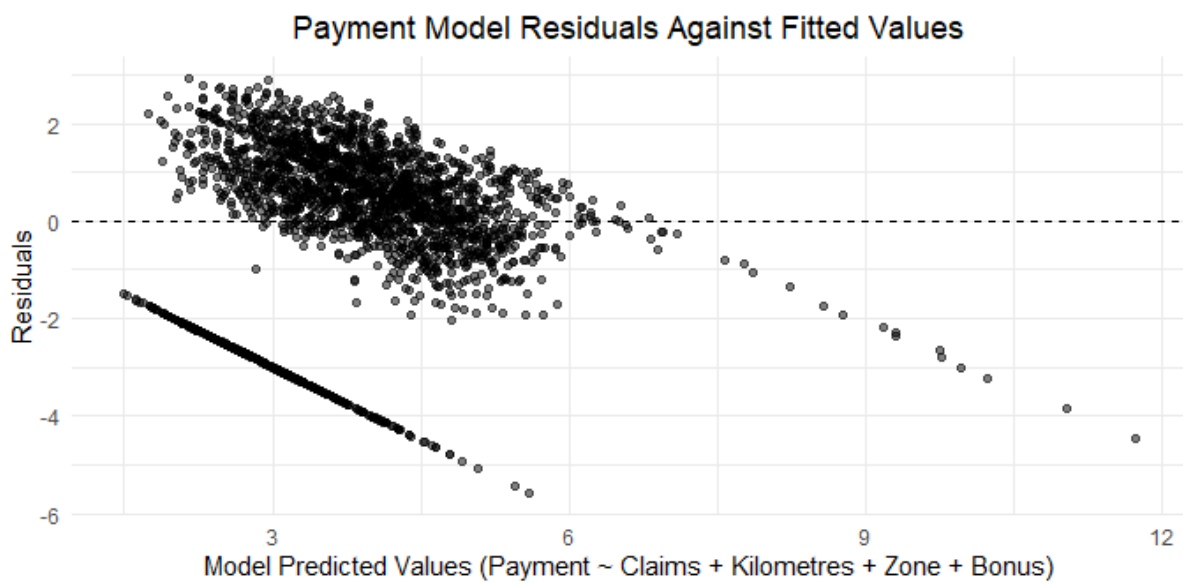
Table 10: Coefficients and standardised coefficients for insured regression model

Appendix 1: Data Transformation

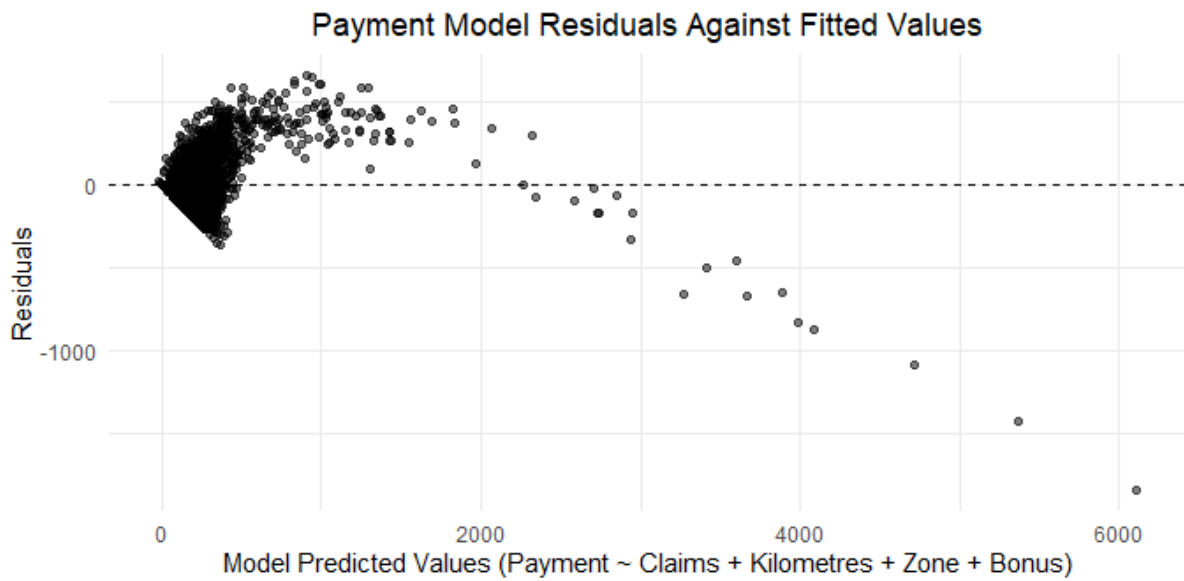
Cube Root:



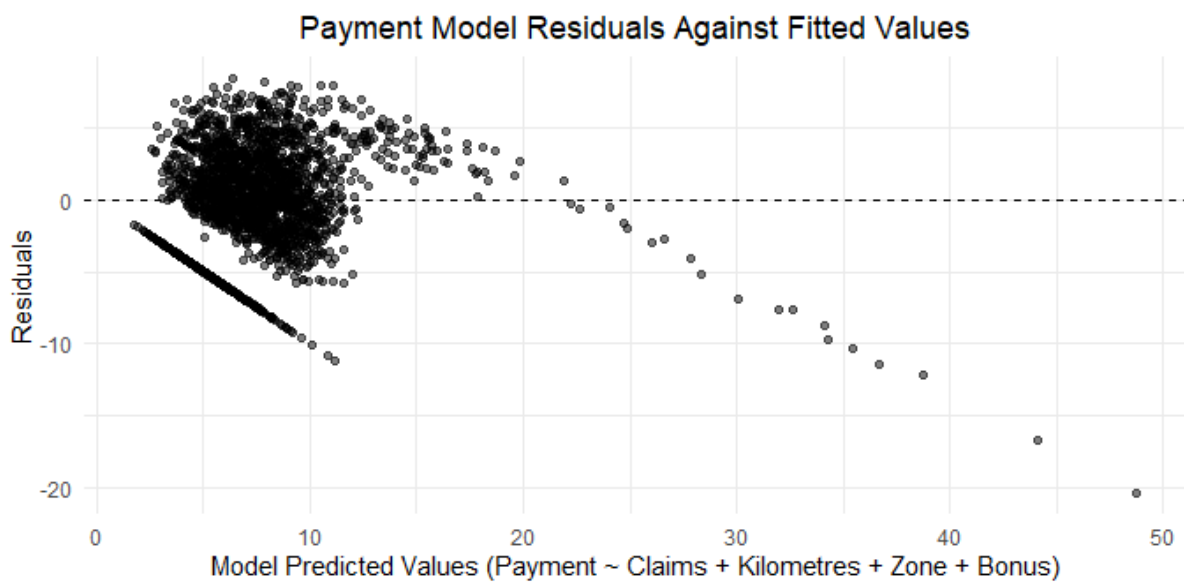
Log 10:



Square Root:



Tukey:



Appendix 2: Payment Prediction Model Fit Analysis

1. $R^2 = 0.8798$, $F = 3186$ on 5 and 2176 DF, $p = < 0.000000000000000022$
2. Mean VIF: 1.034239
3. 1.56% observations standard residual > 2.58
4. Two outliers

Appendix 3: Claim Prediction Model Fit Analysis

1. $R^2 = 0.8425$, $F = 2328$ on 5 and 2176 DF, $p = < 0.000000000000000022$
2. Mean VIF: 1.034239
3. 1.8% observations standard residual > 2.58
4. Two outliers

Appendix 4: Additional Coverage Data

Location	Zone	Kilometre	Bonus	Make	Insured
Vittangi	5	2	2	3	4621
Halmstad	3	2	1	9	9500
Uppsala	2	4	4	3	17500:25416

Appendix 5: Assumptions

1. All work assumes that the data represents a snapshot of a single point in time. Insured represents the total number of insured years of all policies that fulfil the criteria of the categorical data. Claims and payment are, by extension, the total number of claims made against those policies and any insurance pay-outs value.
2. For the most part, given the purpose of the underlying data, Payment will, unless explicitly stated otherwise, be the primary target variable. The assumption is that this data will be utilised in financial planning for the insurance company.
3. The difference between zone 4 rural and Zone 6 rural is that zone 4's land use will primarily be focused on agricultural uses given the position further south of the arctic circle and permafrost. As such, crops can be grown in these locations. This agricultural land use will typically change the types of vehicles on the roads in these areas and, as such, the risk factors. Zone 6, on the other hand, will primarily be forestry. The locations will also result in zone 4 being far more populous than zone 6.
4. Prediction task refers to insured amount, "4621 insured amount", in the case of task 1. For this assignment, I have presumed that this refers instead to 4621 insured policy years.
5. Outliers identified by Cook's distance should not be unilaterally removed from the analysis until further investigation into the validity of those observations. Once their validity has been confirmed, a business decision would need to be made concerning the likelihood of repeated observations of this magnitude. Only then could a decision be made as to whether these observations should remain within the prediction model.

Appendices

Appendix 6: Payment -Insured Model Outliers

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment	I_Standard_Residual	I_Large_Residual	I_Cooks_Distance
252	1	4	7	9	127687.3	2894	15540162	-18.844	0	21.64339
691	2	4	7	9	121293.1	3338	18245026	-6.73262	0	2.429282

Appendix 7: Payment - Claims Model Outliers

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment	C_Standard_Residual	C_Large_Residual	C_Cooks_Distance
9	1	1	1	9	9998.46	1704	6805992	-18.1313	0	2.312769
252	1	4	7	9	127687.3	2894	15540162	10.90511	1	2.433458
691	2	4	7	9	121293.1	3338	18245026	16.34377	1	7.619212

Appendix 8: Claims - Insured Model Outliers

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment	I_Standard_Residual	I_Large_Residual	I_Cooks_Distance
252	1	4	7	9	127687.3	2894	15540162	-17.35	TRUE	15.58674
691	2	4	7	9	121293.1	3338	18245026	-7.92266	TRUE	2.85325

Appendix 9: Claims - Insured Model Outliers

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment	P_Standard_Residual	P_Large_Residual	P_Cooks_Distance
9	1	1	1	9	9998.46	1704	6805992	18.86908	TRUE	3.255891
252	1	4	7	9	127687.3	2894	15540162	-9.30181	TRUE	1.92333
691	2	4	7	9	121293.1	3338	18245026	-14.5289	TRUE	3.294818