APPLIED ECONOMENTRICS AND TIME SERIES ANALYSIS – S24
BUAN 6312.004

GROUP 10
FINAL REPORT

THE PREDICTION OF CUSTOMER LIFE-TIME VALUE OF AN AUTO INSURANCE COMPANY

THE UNIVERSITY
OF TEXAS AT DALLAS

AMRUTHA SOMASHSKEHAR [AXS220079]
NISHANTH GOPENATHAN [NXG220009]
SAI MADHAN MUTHYAM [SXM220216]
SAI SAMPATH PATRO YELLUMAHANTI [SXY210061]
SRIRAM KOUSHIK MAJETI [SXM220076]
VYSHALI CHINTAPALLI [VXC220004]

ABSTRACT:

To support customer-centric methods and strategic decision-making, this initiative recognizes the crucial need to anticipate Customer Lifetime Value (CLV) in the auto insurance business. Leveraging a linear regression model, the study integrates diverse factors—including the number of policies, monthly premium, total claim amount, number of open complaints, and marital status—to forecast CLV accurately. The model accommodates potential non-linear relationships by incorporating log transformations for continuous variables, enhancing its predictive capability. The project's significance lies in its potential to empower insurers with actionable insights into CLV dynamics, enabling them to optimize resource allocation, refine marketing strategies, and prioritize customer retention efforts. Ultimately, the project benefits insurers by facilitating informed decision-making and maximizing revenue and customers by fostering personalized and value-driven interactions with insurance providers.

INTRODUCTION:

The auto insurance industry is characterized by intense competition and a constant quest to maintain customer loyalty while maximizing long-term profitability. Central to achieving these goals is the concept of Customer Lifetime Value (CLV), which provides insurers with a holistic understanding of the enduring value each customer brings to their business. However, accurately predicting CLV poses a significant challenge due to the multitude of factors influencing customer behavior and value generation. In response, this project aims to develop a predictive model for CLV in the auto insurance sector, leveraging advanced statistical techniques and incorporating diverse customer attributes.

By constructing a robust framework for CLV prediction, this project seeks to revolutionize how insurers approach customer relationship management, marketing strategy, and resource allocation. Insights gained from the predictive model can inform strategic decision-making processes, enabling insurers to prioritize investments, optimize marketing campaigns, and allocate resources more effectively. Furthermore, the outcomes of this project are expected to benefit both insurers and customers.

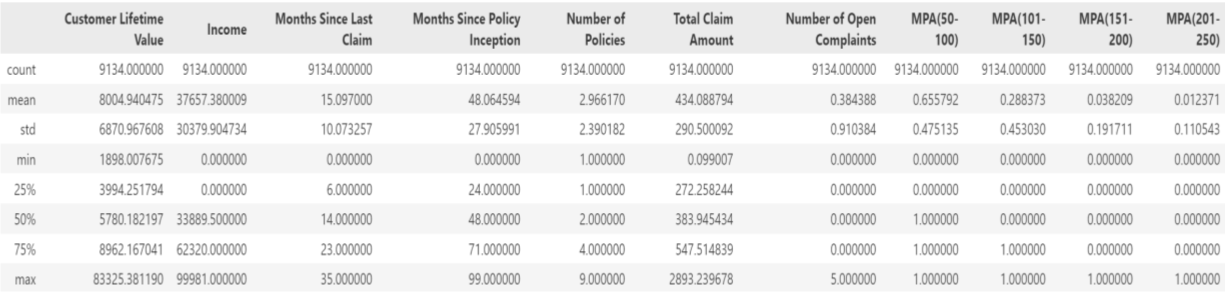
In subsequent sections, this report will delve into the methodology employed for CLV prediction, the data sources utilized, the results obtained, and the implications of findings for the auto insurance industry. Through rigorous analysis and interpretation, we aim to illuminate the intricate dynamics of CLV and its implications for insurers and customers, contributing to advancements in this critical domain and facilitating informed decision-making in the auto insurance sector.

DATA ANALYSIS:

Our data set has been taken from Kaggle, and it belongs to the UCI ML Repository. Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners.

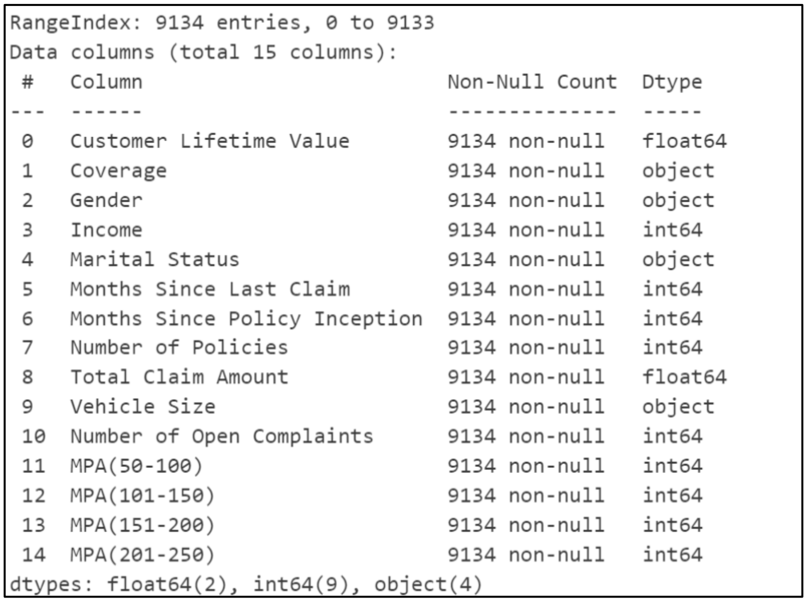
The dataset contains customer information from 5 different American states along with details regarding their vehicle insurance policies. And holds valuable insights for clustering customers based on their behavior, facilitating targeted marketing strategies for new insurance policies.

Summary Statistics of the numerical variables.



	Customer Lifetime Value	Income	Months Since Last Claim	Months Since Policy Inception	Number of Policies	Total Claim Amount	Number of Open Complaints	MPA(50-100)	MPA(101-150)	MPA(151-200)	MPA(201-250)
count	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000
mean	8004.940475	37657.380009	15.097000	48.064594	2.966170	434.088794	0.384388	0.655792	0.288373	0.038209	0.012371
std	6870.967608	30379.904734	10.073257	27.905991	2.390182	290.500092	0.910384	0.475135	0.453030	0.191711	0.110543
min	1898.007675	0.000000	0.000000	0.000000	1.000000	0.099007	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3994.251794	0.000000	6.000000	24.000000	1.000000	272.258244	0.000000	0.000000	0.000000	0.000000	0.000000
50%	5780.182197	33889.500000	14.000000	48.000000	2.000000	383.945434	0.000000	1.000000	0.000000	0.000000	0.000000
75%	8962.167041	62320.000000	23.000000	71.000000	4.000000	547.514839	0.000000	1.000000	1.000000	0.000000	0.000000
max	83325.381190	99981.000000	35.000000	99.000000	9.000000	2893.239678	5.000000	1.000000	1.000000	1.000000	1.000000

Variable Description

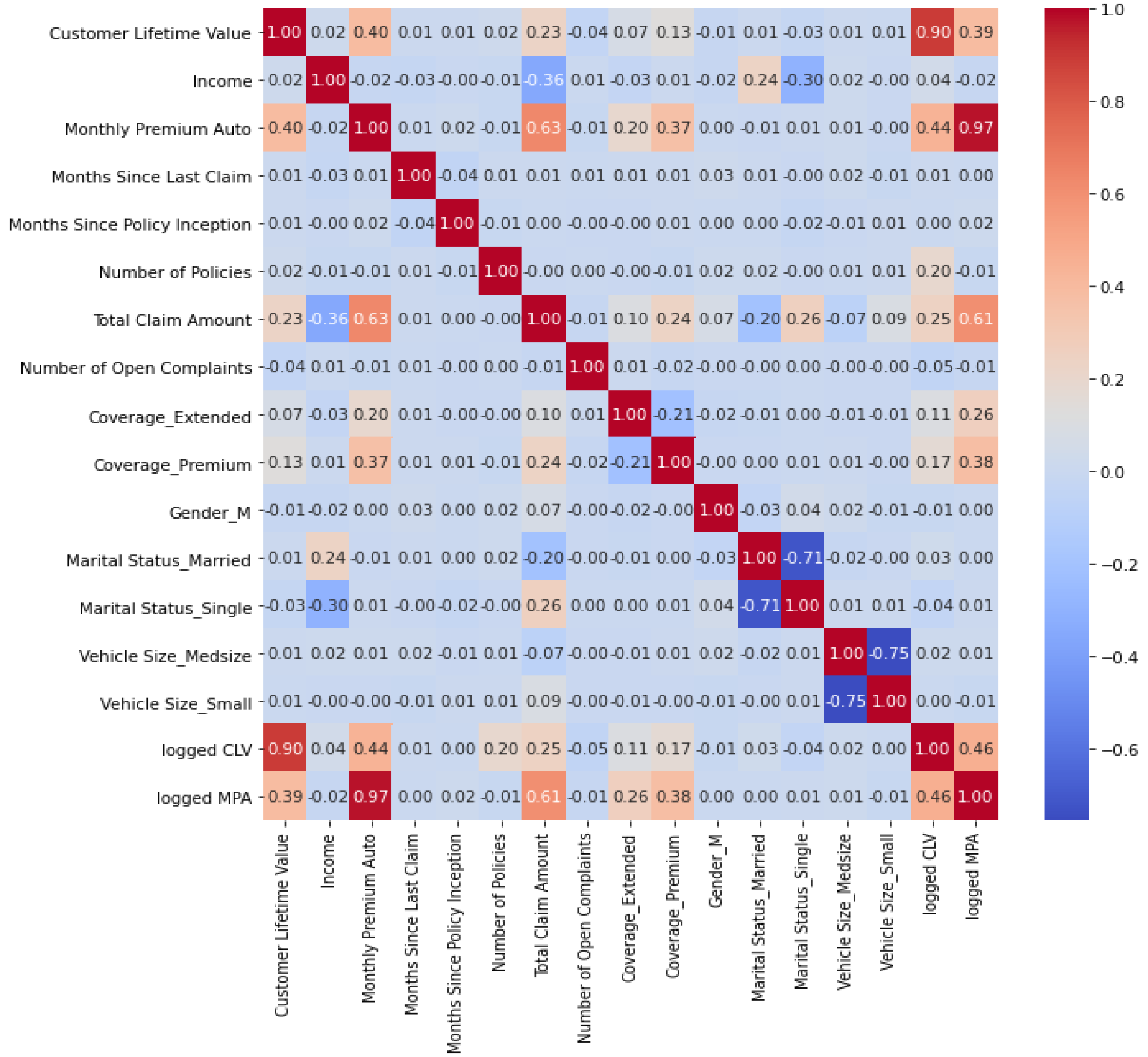


```
RangeIndex: 9134 entries, 0 to 9133
Data columns (total 15 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   Customer Lifetime Value        9134 non-null   float64
 1   Coverage                       9134 non-null   object 
 2   Gender                         9134 non-null   object 
 3   Income                         9134 non-null   int64  
 4   Marital Status                 9134 non-null   object 
 5   Months Since Last Claim        9134 non-null   int64  
 6   Months Since Policy Inception  9134 non-null   int64  
 7   Number of Policies             9134 non-null   int64  
 8   Total Claim Amount             9134 non-null   float64
 9   Vehicle Size                   9134 non-null   object 
 10  Number of Open Complaints      9134 non-null   int64  
 11  MPA(50-100)                    9134 non-null   int64  
 12  MPA(101-150)                   9134 non-null   int64  
 13  MPA(151-200)                   9134 non-null   int64  
 14  MPA(201-250)                   9134 non-null   int64  
dtypes: float64(2), int64(9), object(4)
```

Exploratory Data Analysis

Here comes the basic data analysis and visualization part that displays the relationship of the Y variable and the individual X variables.

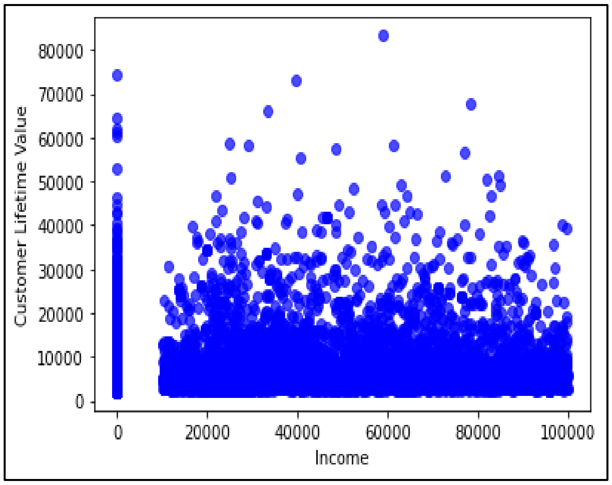
Correlation Matrix of all the variables.

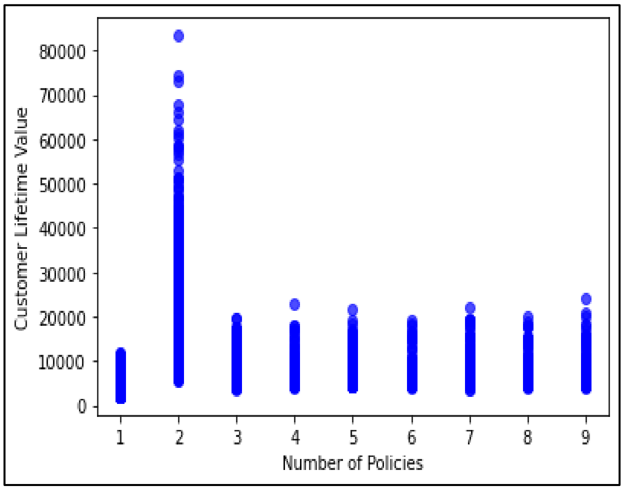


	Customer Lifetime Value	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Policies	Total Claim Amount	Number of Open Complaints	Coverage_Extended	Coverage_Premium	Gender_M	Marital Status_Married	Marital Status_Single	Vehicle Size_Medsize	Vehicle Size_Small	logged CLV	logged MPA
Customer Lifetime Value	1.00	0.02	0.40	0.01	0.01	0.02	0.23	-0.04	0.07	0.13	-0.01	0.01	-0.03	0.01	0.01	0.90	0.39
Income	0.02	1.00	-0.02	-0.03	-0.00	-0.01	-0.36	0.01	-0.03	0.01	-0.02	0.24	-0.30	0.02	-0.00	0.04	-0.02
Monthly Premium Auto	0.40	-0.02	1.00	0.01	0.02	-0.01	0.63	-0.01	0.20	0.37	0.00	-0.01	0.01	0.01	-0.00	0.44	0.97
Months Since Last Claim	0.01	-0.03	0.01	1.00	-0.04	0.01	0.01	0.01	0.01	0.01	0.03	0.01	-0.00	0.02	-0.01	0.01	0.00
Months Since Policy Inception	0.01	-0.00	0.02	-0.04	1.00	-0.01	0.00	-0.00	-0.00	0.01	0.00	0.00	-0.02	-0.01	0.01	0.00	0.02
Number of Policies	0.02	-0.01	-0.01	0.01	-0.01	1.00	-0.00	0.00	-0.00	-0.01	0.02	0.02	-0.00	0.01	0.01	0.20	-0.01
Total Claim Amount	0.23	-0.36	0.63	0.01	0.00	-0.00	1.00	-0.01	0.10	0.24	0.07	-0.20	0.26	-0.07	0.09	0.25	0.61
Number of Open Complaints	-0.04	0.01	-0.01	0.01	-0.00	0.00	-0.01	1.00	0.01	-0.02	-0.00	-0.00	0.00	-0.00	-0.00	-0.05	-0.01
Coverage_Extended	0.07	-0.03	0.20	0.01	-0.00	-0.00	0.10	0.01	1.00	-0.21	-0.02	-0.01	0.00	-0.01	-0.01	0.11	0.26
Coverage_Premium	0.13	0.01	0.37	0.01	0.01	-0.01	0.24	-0.02	-0.21	1.00	-0.00	0.00	0.01	0.01	-0.00	0.17	0.38
Gender_M	-0.01	-0.02	0.00	0.03	0.00	0.02	0.07	-0.00	-0.02	-0.00	1.00	-0.03	0.04	0.02	-0.01	-0.01	0.00
Marital Status_Married	0.01	0.24	-0.01	0.01	0.00	0.02	-0.20	-0.00	-0.01	0.00	-0.03	1.00	-0.71	-0.02	-0.00	0.03	0.00
Marital Status_Single	-0.03	-0.30	0.01	-0.00	-0.02	-0.00	0.26	0.00	0.00	0.01	0.04	-0.71	1.00	0.01	0.01	-0.04	0.01
Vehicle Size_Medsize	0.01	0.02	0.01	0.02	-0.01	0.01	-0.07	-0.00	-0.01	0.01	0.02	-0.02	0.01	1.00	-0.75	0.02	0.01
Vehicle Size_Small	0.01	-0.00	-0.00	-0.01	0.01	0.01	0.09	-0.00	-0.01	-0.00	-0.01	-0.00	0.01	-0.75	1.00	0.00	-0.01
logged CLV	0.90	0.04	0.44	0.01	0.00	0.20	0.25	-0.05	0.11	0.17	-0.01	0.03	-0.04	0.02	0.00	1.00	0.46
logged MPA	0.39	-0.02	0.97	0.00	0.02	-0.01	0.61	-0.01	0.26	0.38	0.00	0.00	0.01	0.01	-0.01	0.46	1.00

The correlation matrix depicts correlation between Customer Lifetime Value and Monthly Premium Auto, Total Claim Amount and Coverage Premium.

Customer lifetime Value VS Income and Number of policies

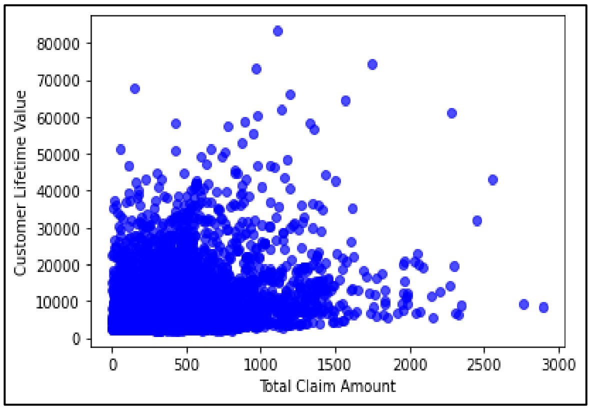


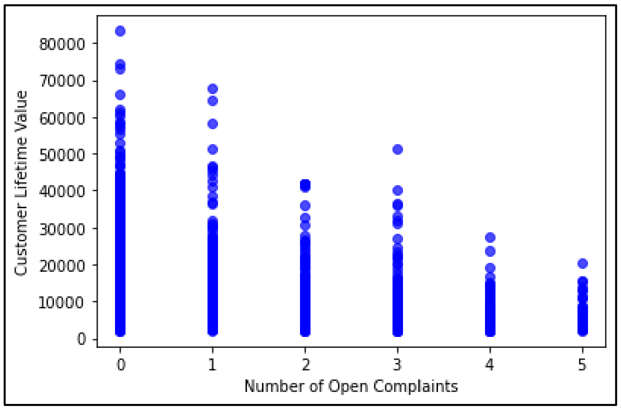


Customer Lifetime Value & Income: The above plot shows that CLV values are randomly scattered across various income groups. Since, there is no identifiable pattern, there is no explainable relation between the CLV and customer income.

Customer Lifetime Value & Number of Policies: The depicted plot indicates that there is no significant correlation between the quantity of policies purchased by customers and their Customer Lifetime Value (CLV). However, it is observed that customers who purchase or maintain two policies contribute the highest CLV to the company.

Customer Lifetime Value VS Total Claim Amount and Number of Open Complaints

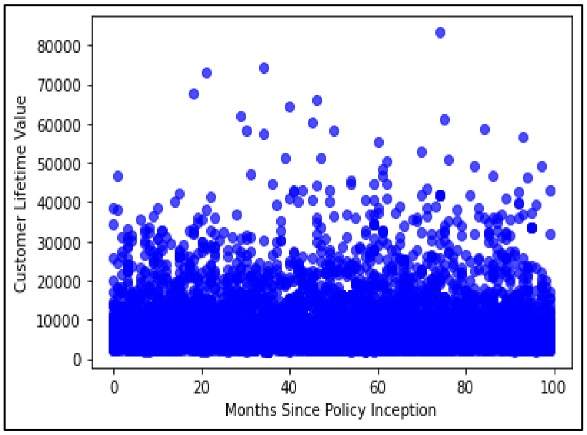


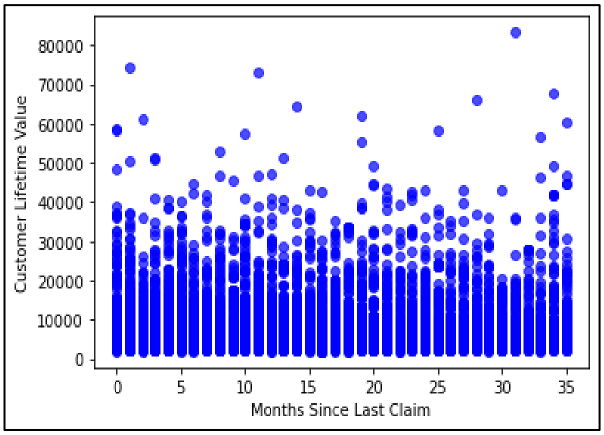


Customer Lifetime Value & Total Claim Amount: Based on the above plot, there appears to be a slight correlation between CLV and Total Claim Amount. However, it's worth noting that this relationship doesn't yield any significant insights.

Customer Lifetime Value & Number of Open Complaints: The above depicted figure reveals a clear pattern of decreasing CLV as the number of Open Complaints increases. There is a notable correlation between these two variables, suggesting that the number of Open Complaints (NoOC) can be utilized to explain variations in CLV.

Customer Lifetime Value VS Months Since Policy Inception and Months since last claim

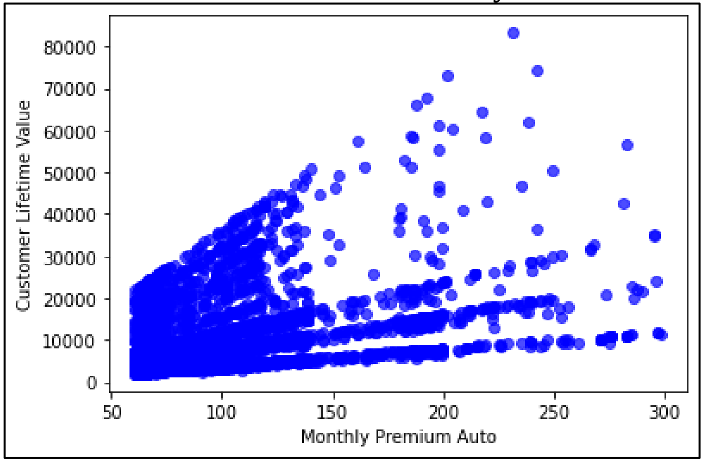




Customer Lifetime Value & Months Since Policy Inception: The above scatter plot shows no clear trend. Hence, there is no noticeable relationship between CLV and Months Since Policy Inception.

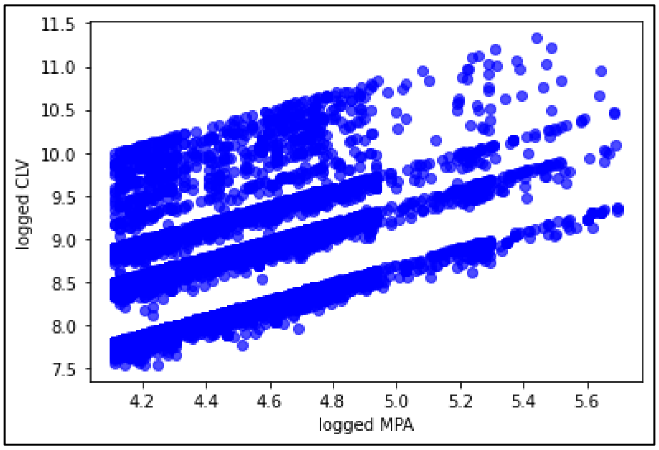
Customer Lifetime Value & Months since last claim: There is no noticeable pattern of correlation between CLV and Months since last claim. Hence, variation CLV may not be explained by Months Since last claim.

Customer Lifetime Value & Monthly Premium Auto

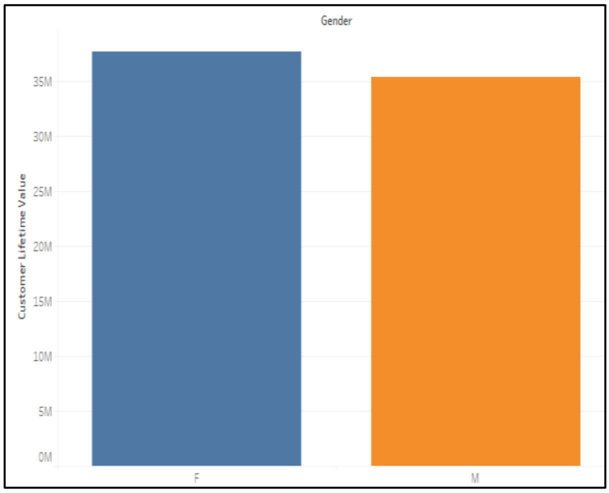


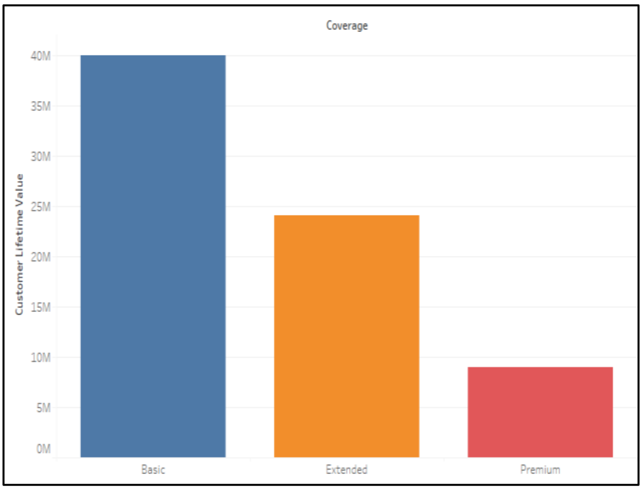
It appears that there is a significant correlation between Customer Lifetime Value (CLV) and the Monthly Premium paid by customers for their auto insurance. Customers who pay higher premiums seem to contribute more CLV to the company. But as the scale of the values are largely differed, let’s try log transforming the values for better understanding.

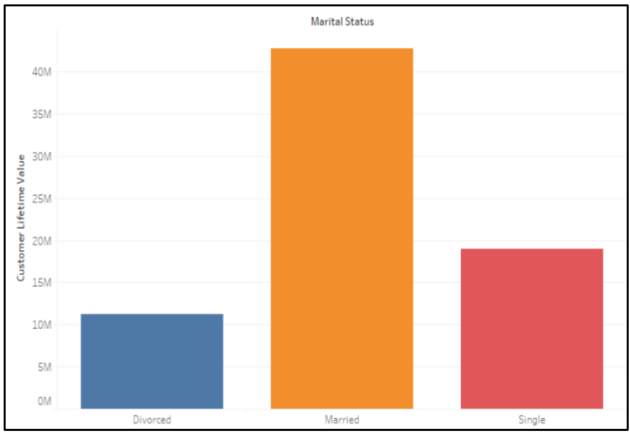
Logged (Customer Lifetime Value) & Logged (Monthly Premium Auto)

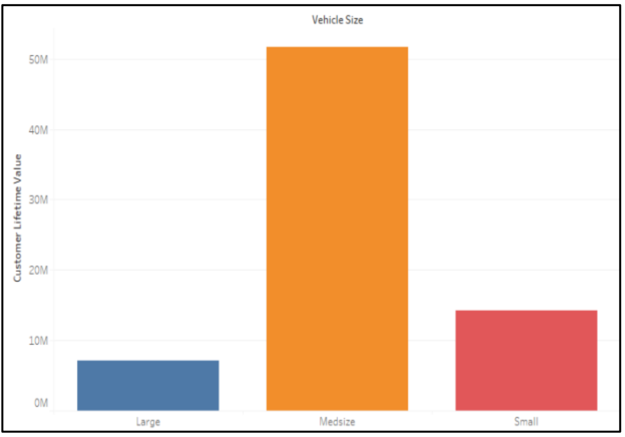


It seems there's a notable relation between Customer Lifetime Value (CLV) and the Monthly Premium customers pay for their auto insurance. Those paying higher premiums appear to contribute more CLV to the company. This suggests that customers opting for pricier plans may receive better coverage or services, leading to greater CLV. Understanding this link can inform targeted marketing strategies aimed at retaining customers willing to pay higher premiums, potentially enhancing the company's profitability in the long term.









From all the above graphs we conclude: Gender does not have a significant impact on Customer Lifetime Value. Customer with Basic coverage plans generate higher CLV for the company. Married customers tend to have a greater contribution to the insurance company. Customer with policies for midsize vehicles generate the highest CLV for the company.

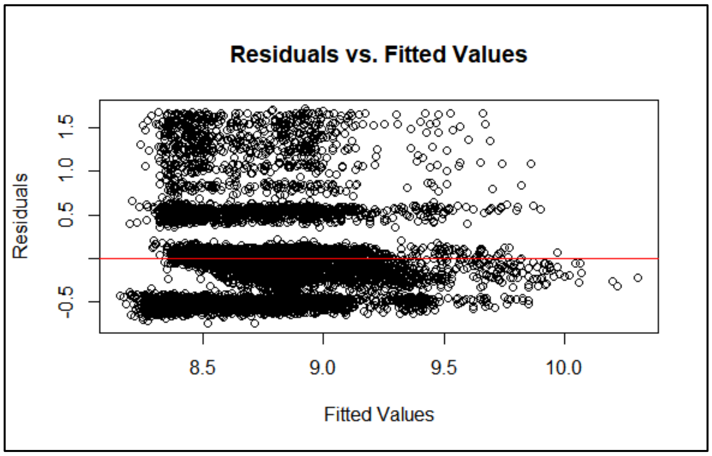
METHODOLOGY:

ASSUMPTION 1: LINEARITY IN PARAMETERS (MLR1)

Implication: The model is predicated on a direct, linear correlation between the predictors and the outcome variable, represented as a straight line in a two-dimensional space when holding all other variables constant.

Linearity Check:

Explaining Linearity from graph: The residuals do not display any clear systematic pattern (e.g., a curve or a wave) that deviates from the horizontal line. This suggests that the model does not suffer from obvious non-linearity issues. This proves that our model is Linear in Parameters.



Interpretation: The plot above represents a Residuals vs. Fitted Values plot, which is commonly used in regression analysis to assess the linearity and homoscedasticity (constant variance) of residuals.

ASSUMPTION 2: INDEPENDENCE OF ERRORS (MLR2)

Explanation: Independence of errors stipulates that the residuals (errors) of the regression model do not influence each other. That is, the error term for any given observation should be unrelated to the error term of any other observation.

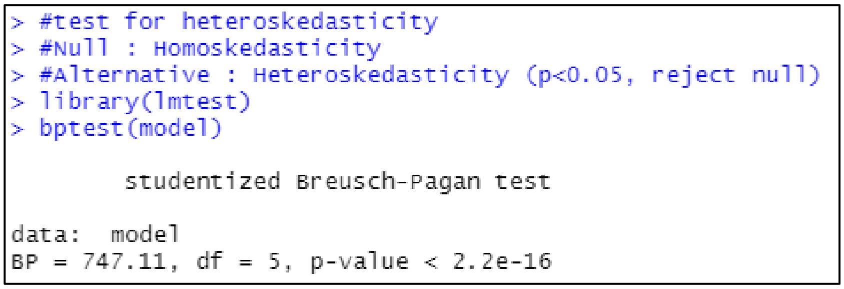
Implication: There should be no autocorrelation, which means the residuals from different observations should not exhibit any pattern or correlation with each other across sequential observations.

ASSUMPTION 3: HOMOSCEDASTICITY (CONSTANT VARIANCE OF ERRORS) (MLR3)

Explanation: This assumption demands that the variance among the residuals be constant across all levels of the independent variables. The spread or scatter of the residuals should remain even, regardless of the predictor values. If the variance of the residuals varies significantly across the range of values for an independent variable or over time, it's known as heteroscedasticity.

Implication: The residuals should scatter uniformly around the regression line throughout the range of predictor values, indicating consistent error variance.

Heteroskedasticity Check:



```
> #test for heteroskedasticity
> #Null : Homoskedasticity
> #Alternative : Heteroskedasticity (p<0.05, reject null)
> library(lmtest)
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 747.11, df = 5, p-value < 2.2e-16
```

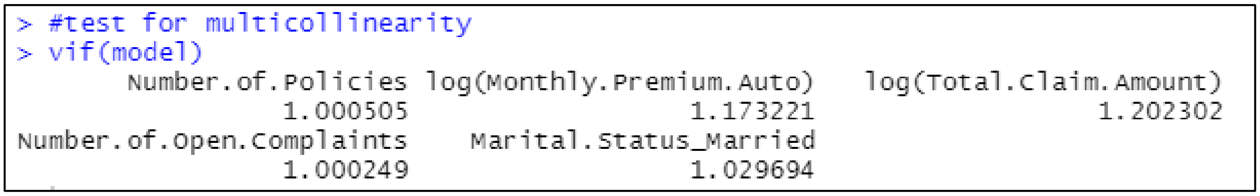
Interpretation: Since the p-value is significantly less than the typical alpha level of 0.05, we reject the null hypothesis of homoskedasticity. This indicates that there is statistically significant heteroskedasticity in the model residuals.

ASSUMPTION 4: NO PERFECT MULTICOLLINEARITY (MLR4)

Explanation: The regression model assumes that no independent variable is a perfect linear combination of any other. Perfect multicollinearity occurs if one or more predictors can be precisely predicted from the others within the model.

Implication: If perfect multicollinearity exists, it can render the regression coefficients unstable, causing them to vary significantly with minor changes in the model or data, thus complicating interpretation.

Multicollinearity Check:



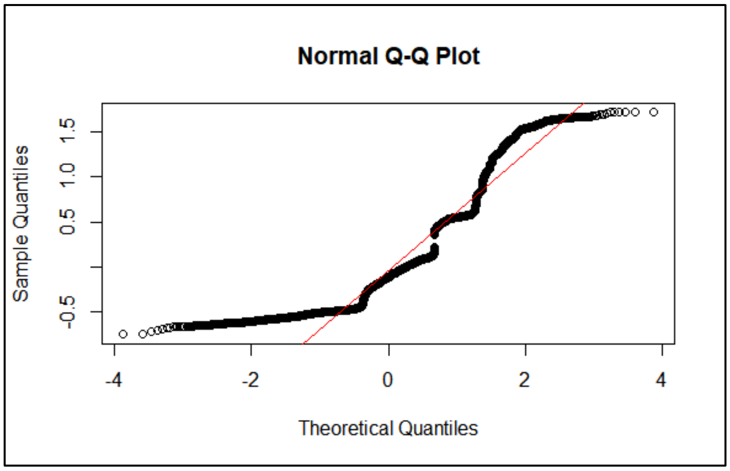
```
> #test for multicollinearity
> vif(model)
       Number.of.Policies log(Monthly.Premium.Auto)   log(Total.Claim.Amount)
                 1.000505                  1.173221                  1.202302
Number.of.Open.Complaints    Marital.Status_Married
                 1.000249                  1.029694
```

Variance Inflation Factor (VIF) value < 5: Generally, a VIF value below 5 suggests that there is no severe multicollinearity among the variables. This is an indicator that the variable does not share a substantial amount of variance with other independent variables in the model.

ASSUMPTION 5: NORMALITY OF ERRORS (MLR5)

Explanation: This assumption requires the residuals of the regression model to follow a normal distribution. The normality of residuals is crucial for the validity of various statistical tests, including hypothesis tests and the construction of confidence intervals.

Normality of Errors Check:



Interpretation: The plot above represents a Normal Q-Q (Quantile-Quantile) Plot, which is used in statistics to visually assess whether a dataset likely follows a normal distribution. The graph plots the quantiles of the sample data against the quantiles of a standard normal distribution. **Theoretical Quantiles:** These are the expected values of observations from a normal distribution, sorted from least to greatest.

- **Sample Quantiles:** These are the sorted values from the dataset being analyzed.
- **Red Line:** This line represents what the data points would follow if the sample distribution was exactly normal.

Deviation in Tails:

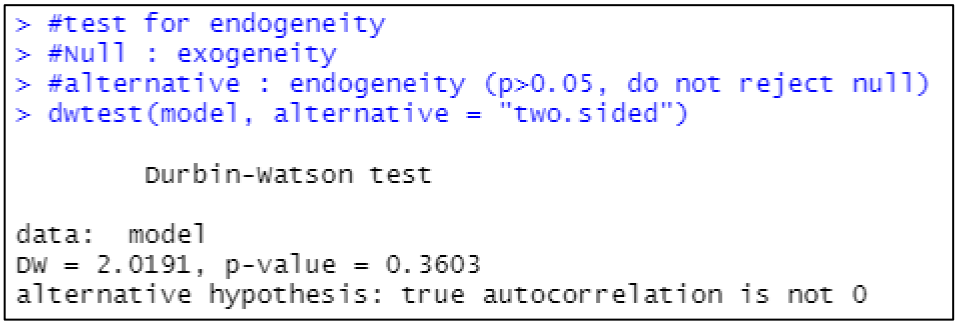
- **Left Tail (Lower Quantiles):** The data points slightly deviate from the line, suggesting the distribution might have lighter tails than a normal distribution.
- **Right Tail (Upper Quantiles):** The more noticeable deviation at the upper end suggests the presence of outliers or heavy tails compared to what would be expected under normality.
- **Center of the Plot:** In the middle range, the data points closely adhere to the red line, suggesting that the central part of the distribution is approximately normal.

ASSUMPTION 6: NO ENDOGENEITY (MLR6)

Explanation: This assumption posits that all independent variables are exogenous; that is, they are not influenced by the model's error terms. Endogeneity occurs when predictors are correlated with the error term, potentially due to omitted variables, measurement errors, or simultaneous causality.

Implication: A violation of this assumption leads to biased and inconsistent coefficient estimates, which misrepresents the true relationship between the dependent and independent variables.

Endogeneity Check:



```
> #test for endogeneity
> #Null : exogeneity
> #alternative : endogeneity (p>0.05, do not reject null)
> dwtest(model, alternative = "two.sided")

        Durbin-Watson test

data:  model
DW = 2.0191, p-value = 0.3603
alternative hypothesis: true autocorrelation is not 0
```

Test Statistic (DW): The Durbin-Watson statistic ranges from 0 to 4, where:

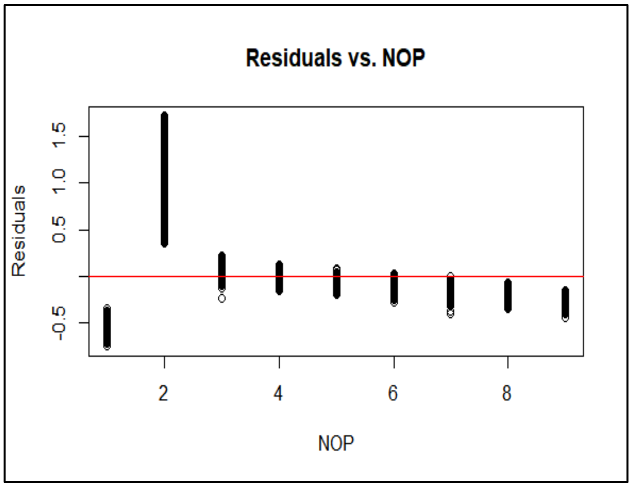
- A value of 2 indicates no autocorrelation.
- Values approaching 0 suggest positive autocorrelation.
- Values approaching 4 suggest negative autocorrelation.

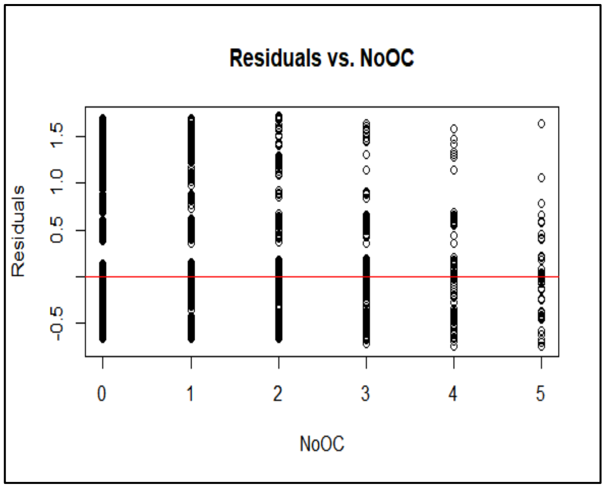
Specifics of Output:

- **DW = 2.0191:** This value is very close to 2, suggesting that there is little to no autocorrelation in the residuals of your model.
- **p-value = 0.3603:** This high p-value supports the null hypothesis of no autocorrelation (since p > 0.05), indicating that there is no significant evidence of autocorrelation in the residuals at the conventional 5% significance level.

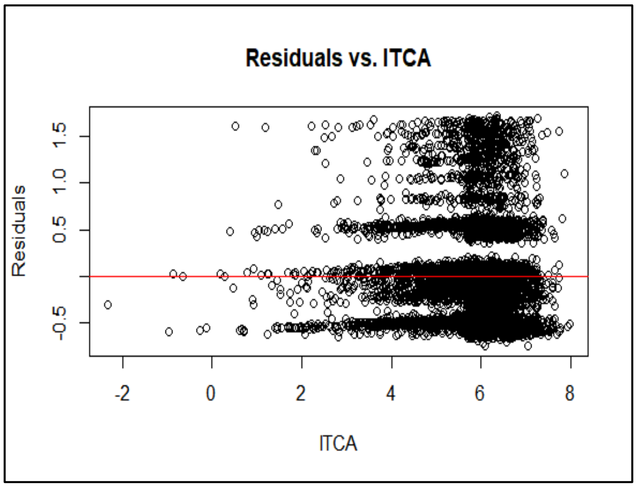
Independence Check:

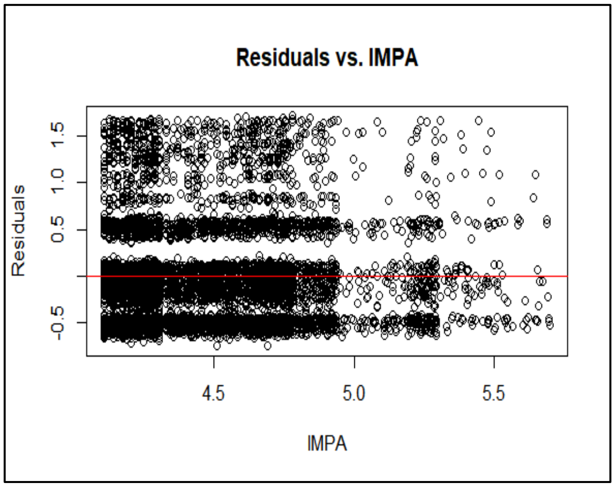
Additionally, each of the variables is plotted against residuals, to check for the patterns.



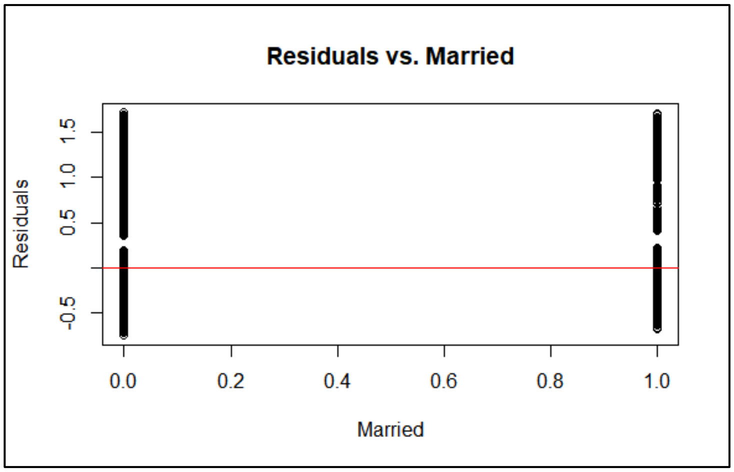


Interpretation: The plots above are "Residuals vs. NOP" (Number of Policies) and "Residuals vs. NoOC" (Number of Open Complaints). These plots are used in regression diagnostics to check for specific relationships between residuals (errors) and each predictor variable, which can give insights into potential issues with the model such as non-linearity, heteroscedasticity, or outliers.



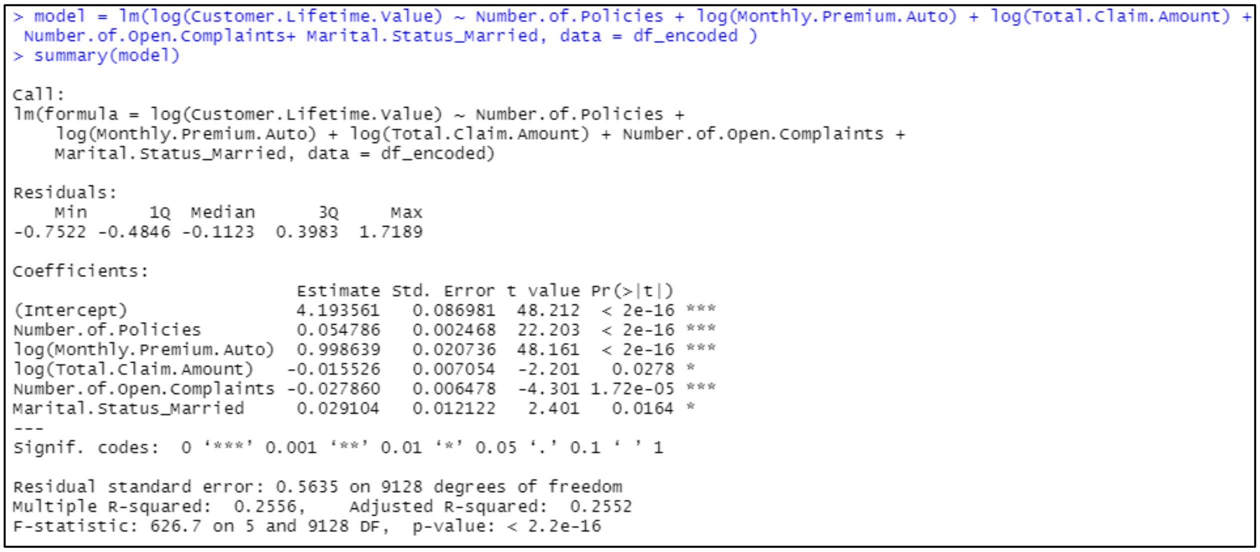


Interpretation: The above plots represents "Residuals vs. ITCA" and "Residuals vs. IMPA" are useful for evaluating assumptions in your regression model such as linearity, homoscedasticity, and the identification of potential outliers.



Interpretation: The plot clearly shows two distinct groups of residuals based on the marriage status. For both groups (Married = 0 and Married = 1), the residuals are distributed around the zero line, which indicates that the mean of the residuals is approximately zero in both cases. This is a good sign as it suggests that the model does not consistently overestimate or underestimate values for either group.

RESULTS:



```
> model = lm(log(Customer.Lifetime.Value) ~ Number.of.Policies + log(Monthly.Premium.Auto) + log(Total.Claim.Amount) +
 Number.of.Open.Complaints+ Marital.Status_Married, data = df_encoded )
> summary(model)

Call:
lm(formula = log(Customer.Lifetime.Value) ~ Number.of.Policies +
    log(Monthly.Premium.Auto) + log(Total.Claim.Amount) + Number.of.Open.Complaints +
    Marital.Status_Married, data = df_encoded)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7522 -0.4846 -0.1123  0.3983  1.7189

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.193561   0.086981  48.212  < 2e-16 ***
Number.of.Policies         0.054786   0.002468  22.203  < 2e-16 ***
log(Monthly.Premium.Auto)  0.998639   0.020736  48.161  < 2e-16 ***
log(Total.Claim.Amount)   -0.015526   0.007054  -2.201   0.0278 *
Number.of.Open.Complaints -0.027860   0.006478  -4.301 1.72e-05 ***
Marital.Status_Married     0.029104   0.012122   2.401   0.0164 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5635 on 9128 degrees of freedom
Multiple R-squared:  0.2556,    Adjusted R-squared:  0.2552
F-statistic: 626.7 on 5 and 9128 DF,  p-value: < 2.2e-16
```

FORMULA:

log(CLV) = 4.194+ 0.999*log(MPA) – 0.016*log(TCA) + 0.055*NoP – 0.0279*NoOC + 0.0291*Marital.Status_Married

The above output is from a linear regression model that explains the logarithm of Customer Lifetime Value (CLV) based on several predictors: the number of policies (NoP), logarithm of Monthly Premium Auto (MPA), logarithm of Total Claim Amount (TCA), the number of open complaints (NoOC), and a binary variable indicating marital status (Marital.Status_Married). Below is an interpretation of the regression summary and the coefficients for the model:

Model Summary:

Formula: The regression model predicts the logarithm of Customer Lifetime Value based on the specified predictors. The model uses log-transformations for Monthly Premium Auto and Total Claim Amount to likely address skewness and non-linear relationships.

Residuals: The range of residuals from the minimum of -0.7522 to a maximum of 1.7189 indicates the spread of errors around the predicted values. A median close to zero (-0.1123) suggests the model's predictions are generally centered around the actual data points.

Coefficients:

- **Intercept (4.194):** When all other variables are zero, the expected log (CLV) is approximately 4.194.
- **log (Monthly.Premium.Auto) (0.999):** This suggests a near one-to-one relationship where a 1% increase in MPA is associated with a 0.999% increase in CLV, holding other factors constant. This is a very significant predictor.
- **log (Total.Claim.Amount) (-0.016):** A 1% increase in the Total Claim Amount is associated with a 0.016% decrease in CLV, which might suggest that higher claims are associated with slightly lower lifetime value.
- **Number of Policies (0.055):** Each additional policy is associated with a 5.5% increase in CLV, suggesting that customers with more policies tend to have higher lifetime values.
- **Number of Open Complaints (-0.0279):** Each additional open complaint is associated with a 2.79% decrease in CLV, indicating dissatisfaction or issues might lead to lower lifetime value.
- **Marital.Status_Married (0.0291):** Being married is associated with a 2.91% increase in CLV compared to not being married.

Significance: All predictors are statistically significant, with p-values less than 0.05, indicating strong evidence against the null hypothesis of no effect. The significance codes in R (`***`, `**`, `*`) further underscore the strength of these relationships, with the most significant being the logarithm of Monthly Premium Auto.

Fit of the Model:

R-squared (0.2556): Approximately 25.56% of the variability in log (CLV) is explained by the model. While this is a moderate amount, it suggests other factors not included in the model also influence CLV.

Adjusted R-squared (0.2552): This adjusts the R-squared for the number of predictors in the model, showing a similar value, which is good as it indicates minimal loss of explanatory power when adjusting for the number of variables.

F-statistic: The F-statistic and its very small p-value (< 2.2e-16) indicate that the model is statistically significant overall, meaning it provides a better fit to the data than an intercept-only model.

CONCLUSION

In summary, our analysis reveals that our model explains approximately 25.56% of the variation in the data, as indicated by the R2 value. This suggests that the selected variables within the model are moderately effective in predicting Customer Lifetime Value (CLV). The adjusted R2 value, which is very close to the R2 value, further supports the notion that the chosen predictors are robust in explaining the variability in CLV.

Furthermore, the statistical significance of all coefficients, as evidenced by their p-values being less than the conventional threshold of 0.05, reinforces the reliability of our model's predictions.

However, it's important to note that our analysis also uncovers issues related to heteroskedasticity, which arises from the nonlinear nature of the relationship between the predictors and CLV. Additionally, the presence of omitted variable bias is a concern, as important variables such as driving experience, age, education, and alcohol consumption are not included in our dataset. This omission may result in biased estimates and reduced predictive accuracy.

To address these limitations and improve model performance, future research could explore advanced techniques such as XG Boost, Random Forest, and Neural Networks. These methods have the potential to capture complex relationships in the data more effectively and may yield better predictive outcomes. By incorporating additional variables and employing more sophisticated modeling approaches, we can enhance our understanding of CLV determinants and improve the accuracy of our predictions.