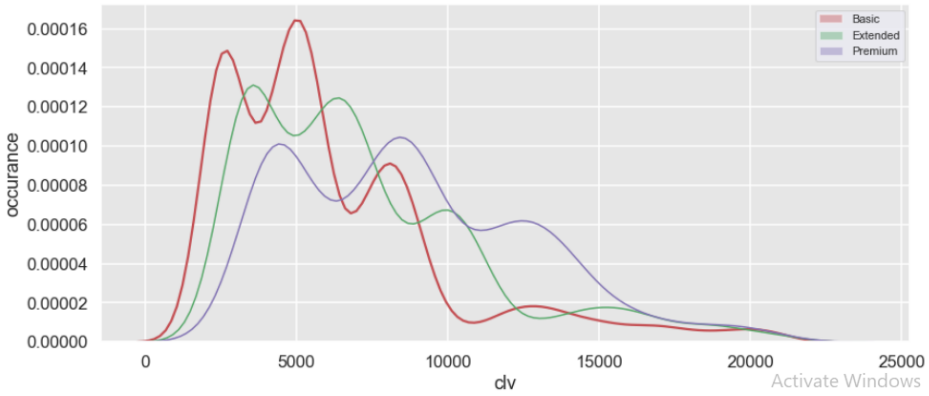
Vehicle Insurance Data Visualization report

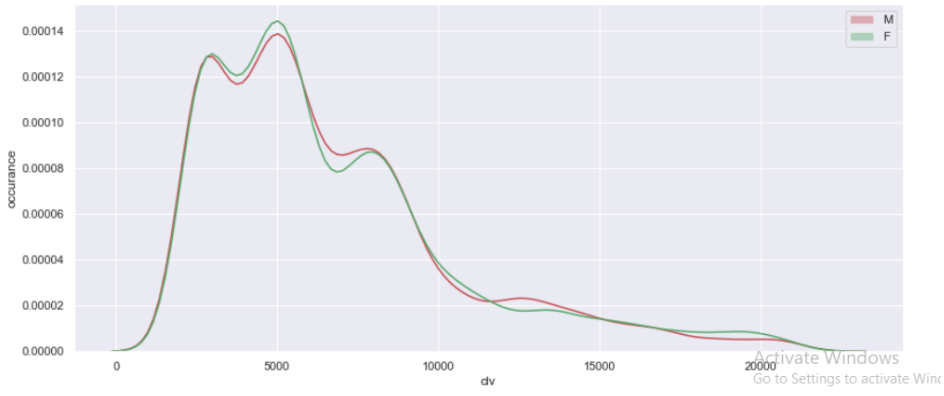
Introduction:

I am going to do data visualization and build an ML model on the Vehicle insurance dataset. This dataset contains multiple features according to the customer's vehicle and insurance type. The main objective is to increase the clv (customer lifetime value) which means clv is the target variable. This dataset is pretty clean already, a few outliers are there. Remove the outliers. The Target column is CLV (customer Live value). Customer lifetime value is the total worth to a business of a customer over the whole period of their relationship.

Dataset features:
CLV : It is a target column (customer Live value).
Education : Customer qualification.
Gender : Customer gender type.
Income : Customer Income.
Location : Customer Location.
Martial : Customer Martial status.
Monthly.Premium.Auto: car insurance premium is the amount you pay your insurance company on a regular basis.
Coverage: the amount of risk or liability that is covered for an individual or entity by way of insurance services.
Total.Claim.Amount: the sum payable at the maturity of an insurance policy.
Vehicle.Size : Size of Veicle.

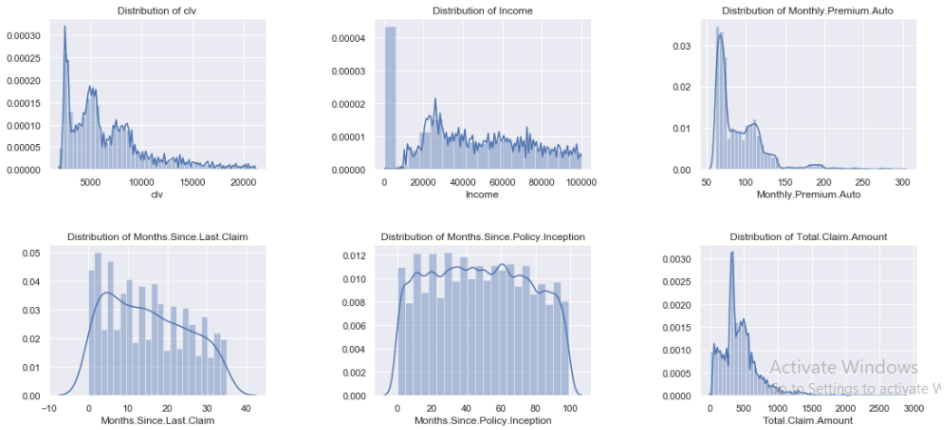


In this histogram I plot the data of Coverage over clv target variable to see the distribution between Basic, Extended, and Premium category. This graphs show that Basic have varience which is represented by red color. And Extended have second highest varience and Premium have third high varience over clv.



In this histogram I plot the data of Coverage over clv target variable to see the distribution between Basic, Extended, and Premium category. This graphs show that Basic have varience which is represented by red color. And Extended have second highest varience and Premium have third high varience over clv.

Distributions of Numeric Columns



clv (customer lifetime value)

Clv is a Target variable. The histogram of clv shows that the data has the highest peak at around 3000 and then the second highest peak at around 5000 and the third last highest peak is at around 8000 with average frequency. The Graph shows the distribution is right-skewed because the data has a tail on the right side. As the values of data increase the frequency of data decreases..

Income

Income is an independent variable that shows the monthly income of the customer. The histogram of Income shows that the data is approximately uniformly distributed. although the data has a higher peak at 0 income. But overall, the whole data is uniformly distributed.

Monthly.Premium.Auto

Monthly.Premium.Auto is the column of auto-renewal of monthly premium which is the independent variable. The histogram of Monthly.Premium.Auto shows that the data has the highest peak at around 80 and then the second highest peak at around 120 with average

frequency. The Graph shows the distribution of right-skewed because the data has a tail on the right side. As the values of data increase the frequency of data decreases.

Months.Since.Last.Claim

Months.Since.Last.Claim the column of monthly last claim by the customer. The histogram of Months.Since.Last.Claim shows that the data has the highest peak at around 5 and after 5 the distribution is going down. The Graph shows the distribution is right-skewed because the data has the highest peak on the left side and the lowest peak on the right side. As the values of data increase the frequency of data decreases.

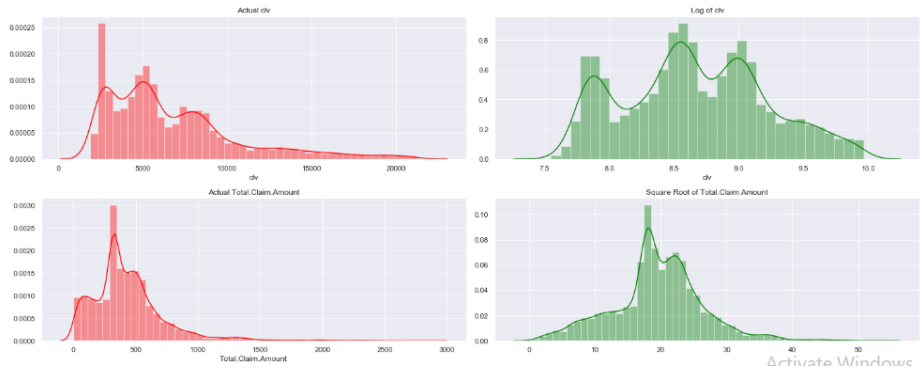
Months.Since.Policy.Inception

Months.Since.Policy.Inception is an independent variable that shows the monthly last Inceptions for vehicle insurance. The histogram of Months.Since.Policy.Inception shows that the data is highly uniformly distributed at each value of Months.Since.Policy.Inception the data is approximately equaly distributed.

Total.Claim.Amount

Total.Claim.The amount is the data of the total amount which was claimed by customers for insurance. The histogram of Total.Claim.Amount shows that the data has the highest peak at around 200 to 300. The Graph shows the distribution is right-skewed because the data has a tail on the right side. As the values of data increase the frequency of data decreases.

Log and Reciprocal Transformation



Skewness and Normality Test

```
_____________________________________________________________________
The normality test and Skewness of Actual clv
The Stat: 1842.5059574669185 and P value 0.0 of Actual clv
Skewness : 1.3292863591978683
---------------------------------------------------------------------
The normality test and Skewness of Log of clv
The Stat: 470.51482062202314 and P value 6.745355995084691e-103 of Log of clv
Skewness : 0.1620053528896739
_____________________________________________________________________
The normality test and Skewness of Actual Total.Claim.Amount
The Stat: 3204.169816526084 and P value 0.0 of Actual Total.Claim.Amount
Skewness : 1.7001788215164855
---------------------------------------------------------------------
The normality test and Skewness of Square Root of Total.Claim.Amount
The Stat: 165.31933857816122 and P value 1.2628790297638732e-36 of Square Root of Total.Claim.Amount
Skewness : 0.1070246478630503
_____________________________________________________________________
```


Hypothesis tests

H0: (null hypothesis): A variable follows a hypothesized distribution.

H1: (alternative hypothesis):A variable does not follow a hypothesized distribution.

Actual clv

The p-value (0.0000e-30) is much less than 0.05, Based on that evidence we succeed to reject the null hypothesis. This means we have piece of sufficient evidence to say that the true distribution of Actual clv is Not Normal Distribution. According to the Skew value (1.3292863591978683) of Actual clv the distribution is right-skewed.

Log of clv

The p-value (3.395998819099824e-27) is also much less than 0.05, Based on that evidence we succeed to reject the null hypothesis. This means we have piece of sufficient evidence to say that the true distribution of Log of clv is also Not Normal Distribution. According to the Skew value (0.1620053528896739) of Log of clv the distribution is little right-skewed. But if we compare it to the Actual clv column we can say that the Log Distribution of clv is very near to normal distribution

---

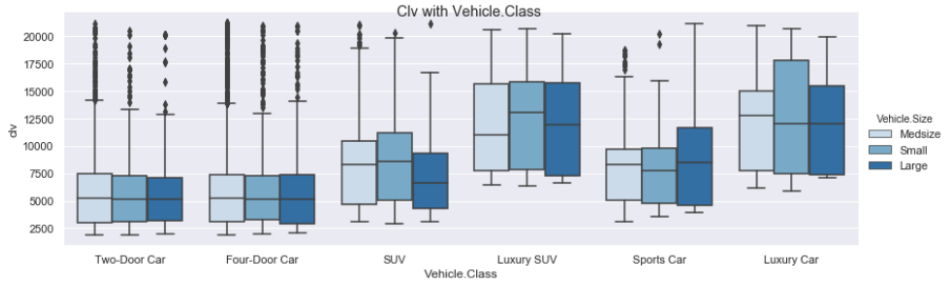
Actual Total.Claim.Amount

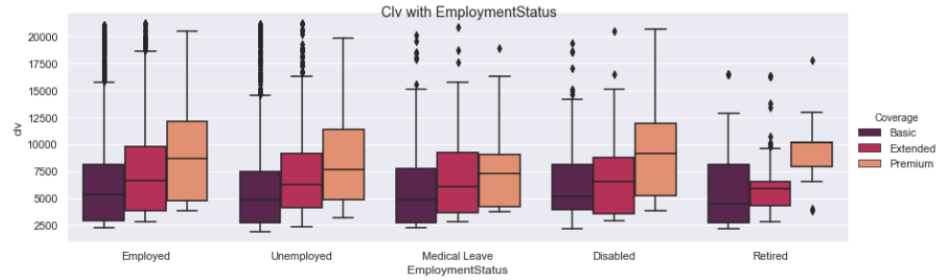
The p-value (0.0000e-40) is much less than 0.05, Based on that evidence we succeed to reject the null hypothesis. This means we have piece of sufficient evidence to say that the true distribution of the Actual Total.Claim.Amount is Not Normal Distribution. According to the Skew value (1.7001788215164855) of Actual Total.Claim.Amount the distribution is right-skewed.

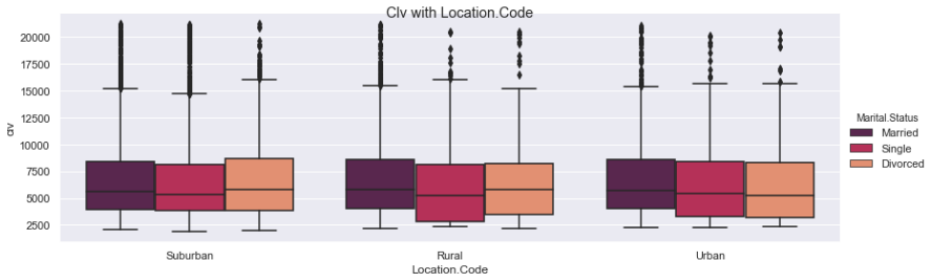
Squrae root of Total.Claim.Amount

The p-value (4.2111177636569615e-31) is also much less than 0.05, Based on that evidence we succeed to reject the null hypothesis. This meanswe have piece of sufficient evidence to say that the true distribution of Squrae root of the Total.Claim.Amount is also Not Normal Distribution. According to the Skew value (0.1070246478630503) of Squrae root of Total.Claim.Amount the distribution is less right-skewed. But if we compared it with the Actual Total.Claim.Amount, we can easily observed that the Log Distribution of Total.Claim.Amount is very near to the Normal Distribution.

Categorical variables Distribution Aanalysis







Clv with Vehicle.Class

Those customers who have a Luxury car have an average (12500) customer lifetime value (clv) and have a maximum value of 20000 clv. And those customers who have a Two-Door and Four-Door car have an

average of 5000 customer lifetime value (clv) which is lower than those customers who have Luxury cars and luxury SUV. And the customers who have a sports car have an average of 7500 clv.

Clv with EmploymentStatus

The customers with Premium Coverage have a high customer lifetime value (clv) which is more than 7500 clv and the customers with Extended Coverage have an average value of less than 7000 clv and the customers with Basic Coverage have an average value less than 5000. The distribution of EmploymentStatus is uniform which means every category is equally distributed and has not to impact on customer lifetime value (clv).

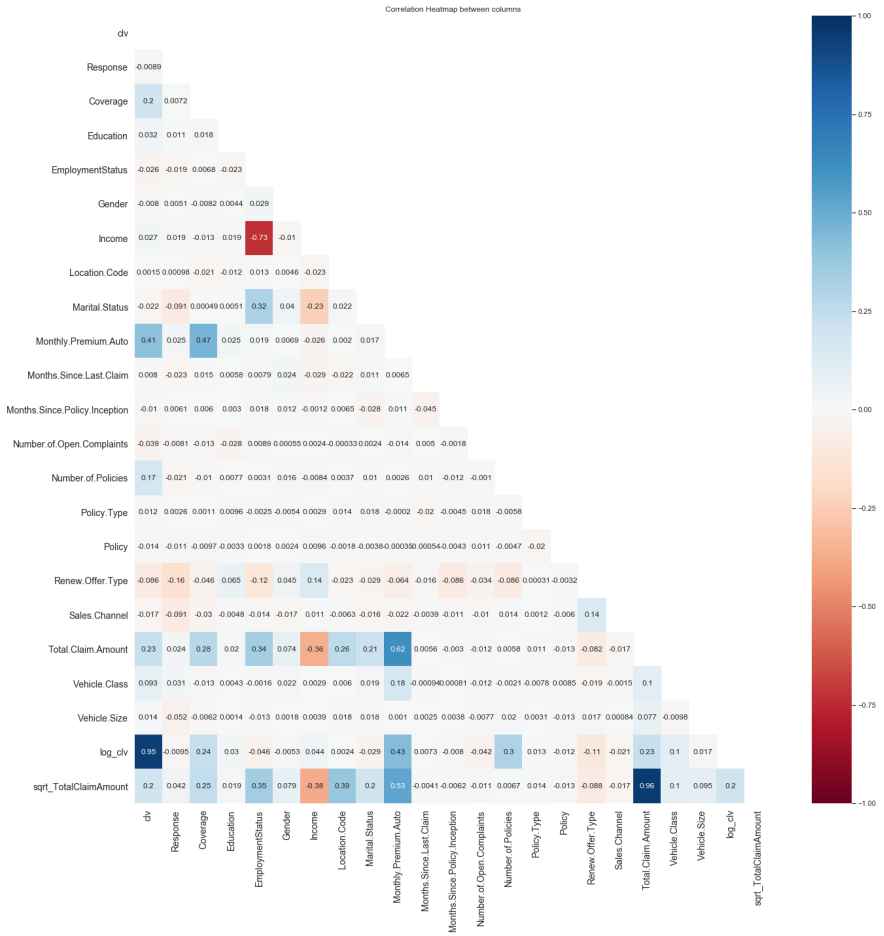
Clv with Location.Code

The Location.Code and Marital.Status also have an uniform distribution. The average clv of Location.Code and Marital.Status is 5000 and have an maximum 15000 clv and minimum 2500 clv in each category..

Income With Location.Code

Those customers who are from Rural and Urban area have an high income which is greator than and equal to 50000 and those who are form subrban area have an maximum income of 30000.

Correlation between Columns.



It can be useful in data analysis and modeling to better understand the relationships between variables. The statistical relationship between two variables is referred to as their correlation. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated.

• Positive Correlation: both variables change in the same direction.

• Neutral Correlation: No relationship in the change of the variables.

• Negative Correlation: variables change in opposite directions.

Positively Correlated variables

Coverage and Monthly.Premium.Auto are highly positively correlated and their correlation value is 0.47 which is near to one. It means they

depend on each other if the value of Coverage is the increased value of Monthly.Premium.Auto will also increase by 0.47 because their relationship is directly proportional to each other.

Clv and Monthly.Premium.Auto are positively correlated and their correlation value is 0.41 which is less near to one but their correlation is positive. It means they depend on each other if the value of Clv increases value Monthly.Premium.Auto will also increase by 0.41 because their relationship is directly proportional to each other.

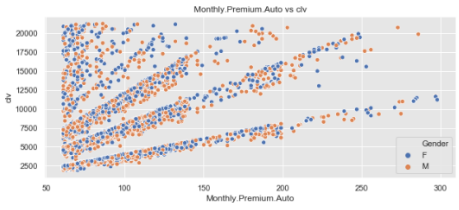
Total.Claim.Amount and Monthly.Premium.Auto are highly positively correlated and their correlation value is 0.62 which is near to one. It means they depend on each other if the value of Total.Claim.Amount has increased the value of Monthly.Premium.Auto will also increase by 0.62 because their relationship is directly proportional to each other.

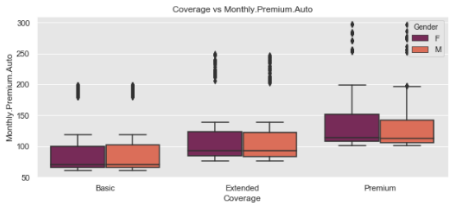
Negatively Correlated variables

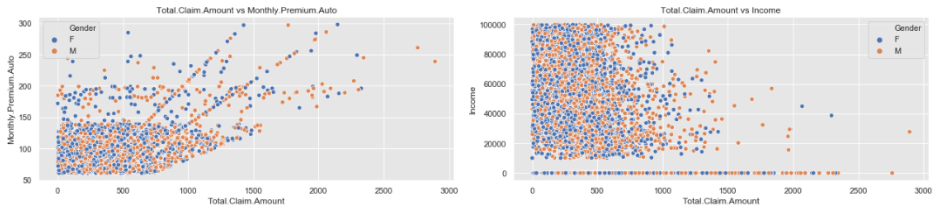
EmploymentStatus and Income are highly negatively correlated and their correlation value is -0.73 which is near to -1. It mean they are depend on each other if the value of EmploymentStatus is decreases the value of Income will increase by 0.73 and if the EmploymentStatus decreases the value of Income will decrease by 0.73 because their relation is inversly proportional to each other.

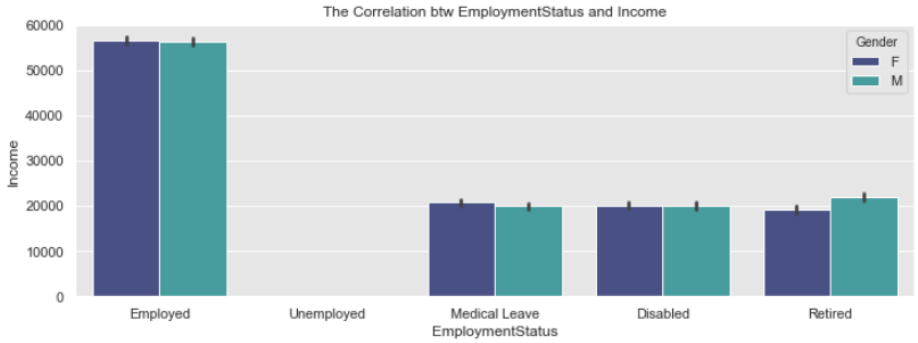
Neutrally Correlated variables

Location.Code and Income are neutrally correlated and their value is -0.023 which is very near to 0 form the left side. It mean they are not depend on each other if we change the value of one column then it will have an impact on other column.

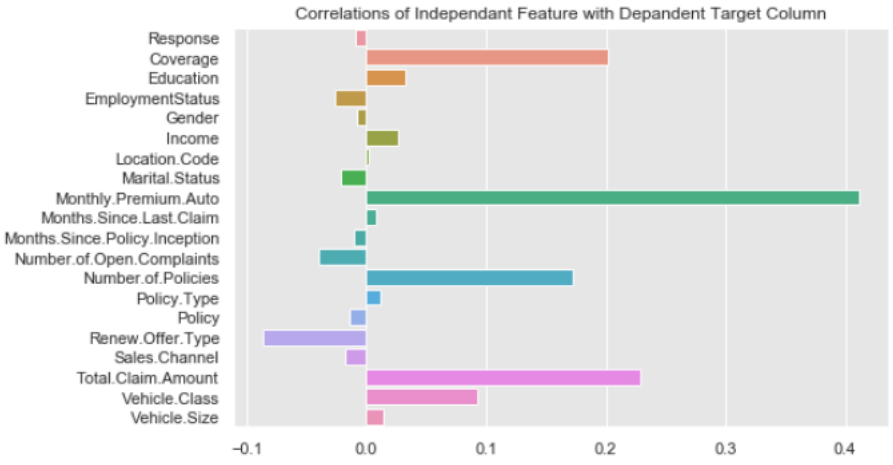


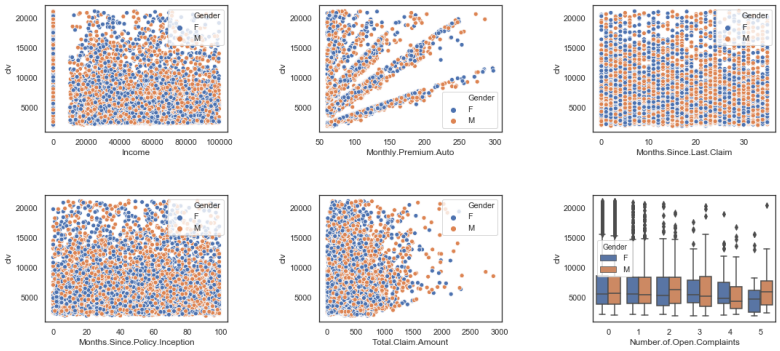






EmploymentStatus and Income are highly negatively correlated and their correlation value is -0.73 which is near -1. It means they are dependent on each other if the value of EmploymentStatus has decreased the value of Income will increase by 0.73 and if the EmploymentStatus decreases the value of Income will decrease by 0.73 because their relation is inversely proportional to each other. Those customers who are employed have high income than Medical Leave, Disabled, and Retired. And those customers who are unemployed have approximately equal to 0.





Feature Selection

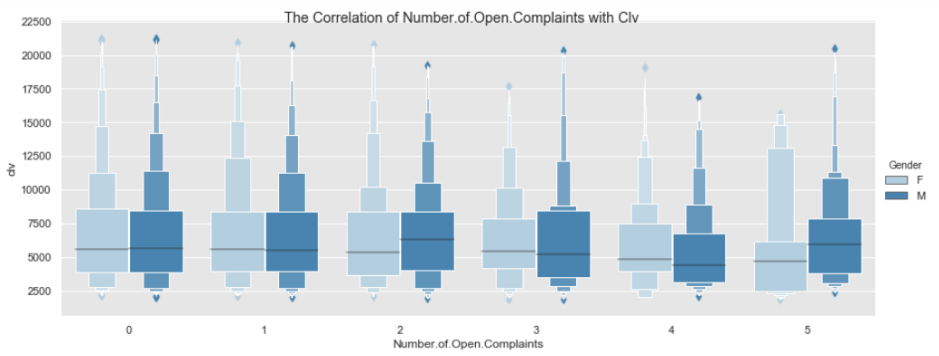
By analyzing the Bar and Scatter plots we can easily see the correlation of independent columns with Target variable clv.

Let's first talk about **Income** correlation with clv. The bar plot shows that the correlation between Income and clv is (0.027) which is too low. so that we can say that the correlation is neutral because the scatter plot of Income and clve shows the uniform behavior. For model training Neutral variables are not good so that we can eliminate income for training.

Now let's talk about **Monthly.Premium.Auto**. The correlation between Monthly.Premium.Auto and clv are highly positive which is (0.41). So, Monthly.Premium.Auto is an important feature for Model training. As you can see that the scatter plot of Monthly.Premium.Auto and clv show many linear lines which are good high Monthly.Premium.Auto values have high clv values.

The correlation of **Months.Since.Last.Claim** and **Months.Since.Policy.Inception** with clv are (0.008) and (-0.01) which are neutral. so, these columns are not correlated with Target variable clv so we can eliminate them for better Prediction.

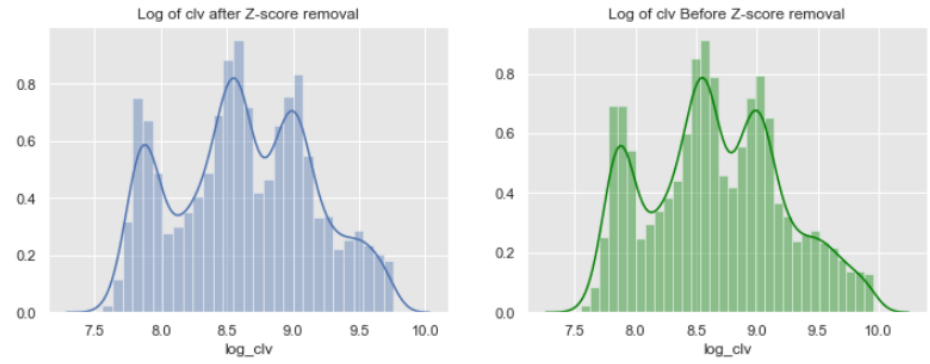
The correlation of **Total.Claim.Amount** with clv is (0.23) which is also positive. so this feature is important for feature selection.



The correlation of Number.of.Open.Complaints with clv is (-0.039) which is near to neutral. so, this columns is not correlated with Target variable clv so we can eliminate it for better Prediction. The Boxen plot shows that the average is approximately same in all category of Number.of.Open.Complaints which is 5000 clv.

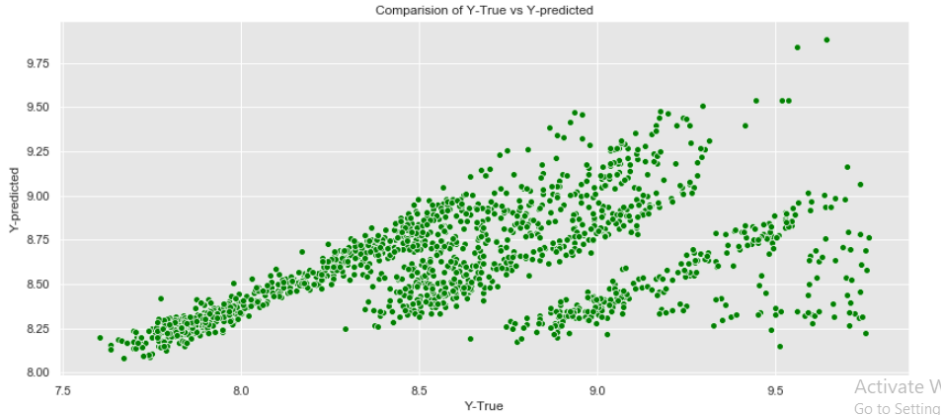
Removing the outliers by using Standard Deviation Method

One of the most commonly used tools in determining outliers is the Z-score. Z-score is just the number of standard deviations away from the mean that a certain data point is.



LinearRegression Model:

linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable.Continuous predictor variables are also called covariates, and categorical predictor variables are also called factors.

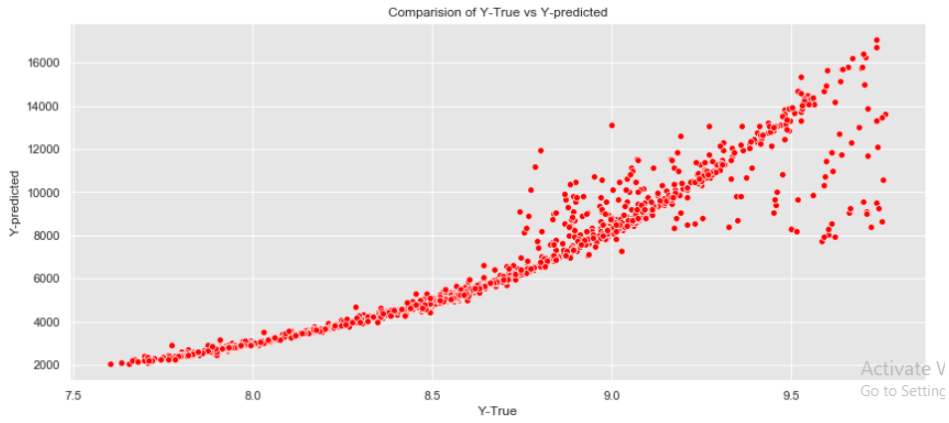


This Graph shows the performance of Linear Regression Model by comparing the Actual Y values with their log predicted values. The graph shows the irregular straight lines which is not a high performance. The Accuracy of Linear Regression Model is 31% and the error is 69% The Accuracy is too Low and the Error is too high.

RandomForestRegressor:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression . ... A Random Forest operates

by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.



This Graph shows the performance of Random Forest Regressor Model by comparing the Actual Y values with their log predicted values. The graph shows the approximately smooth curve with a high Accuracy of 96% and with the error of 0.033% which is too low.

Conclusion

Hypothesis Testing

The Hypothesis Tests proved that clv (customer lifetime value), Monthly.Premium.Auto, Months.Since.Last.Claim, and Months.Since.Last.Claim are right-skewed because the data has a tail on the right side. As the values of data increase the frequency of data decreases. And Income and Months.Since.Policy.Inception columns are uniformaly distributed because they are equally distributed and have the same Frequency.

Correlation Testing

By plotting Correlation plots and values i proved that the Coverage, Monthly.Premium.Auto, Number.of.Policies, and Total.Claim.Amount columns are positively correlated to the Target Variable and their relation is directly proportional to the CLV column.The Number.of.Open.Complaints and Renew.Offer.Type are negatively correlated to the Target Variable and their relation is inversely proportional to the CLV column