Customer Lifetime Value

## Approach and Steps:

We try to analyze the data with the help of a **Linear Regression Model** in order to predict the variables that affect the Customer Lifetime Value of a company.

* We set the working directory accordingly.
* Call all the necessary library functions.
* We store the dataset inside data frame called “data”.
* In order to manipulate with the data, we create a backup called “data1”.
* We change the necessary factor variables with the help of as.factor().
* We conduct a basic exploration and find that the data has presence of extreme values. Hence, they need to be treated.
* We cap the outliers with the help of Quantile function and boxplot.
* We remove all the data that are unnecessary for of model prediction like State and Effective to date and store the remaining in data.final.
* In order to produce reproducible results, we run the set.seed()
* We split the data into train and test data having a 70:30 ratio.
* We run the regression on our train dataset, each time after removing variables or dummies that are insignificant.
* On our 4th iteration, we reach to our final model. The variables that remain significant are:
* Coverage
* Education
* Employment Status
* Gender
* Income
* Marital Status
* Monthly Premium Auto
* Number of Open Complaints
* Number of Policies
* Vehicle Class
* We check if there’s Multicollinearity in the model. Multicollinearity doesn’t exist.
* We run the model that we get from the train data into the test data and remove the insignificant variables/ values that have a very high p-value.
* We check for multicollinearity and find that it doesn’t exist.
* Then we run the tests for assumptions. We get the following results.
* **Multiple R-Squared :** 0.9294
* **Adjusted R-Squared :** 0.9287
* **Bp test p-value < 0.05**
* We use the Bp test to detect heteroscedasticity. We consider the null hypothesis that errors/residuals are homoscedastic, i.e. errors have constant variance. With the test result < 0.05 we reject the null hypothesis and conclude that the errors are heteroscedastic and does not have constant variance, which is not good for the model.
* **Dwt test statistic = 2.011192 and p-value >0.05**
* We use the dwt test to detect auto-correlation. Auto-correlation means there's correlation between the errors/residuals. As p-value is >0.05 we accept the null hypothesis that there is no Auto-correlation present.
* **ad test p-value < 0.05 and sf test p-value < 0.05**
* The Anderson-Darling test and Shapiro-Francia test are tests for normality, i.e. They test if the model has a normal or a near to normal distribution. In both cases, we consider the null hypothesis to be that the data is normal. As both tests give p-value < 0.05, we conclude that the data is not normally distributed.
* **MAPE value is 10.09%** which means that the average error of the model is 10.09% which means the model is good.

Based on the understanding of the model, we make the following business meaning.

* Basic Coverage is more profitable
* Educated married and employed customers are valuable and increase CLV
* Monthly premium increases CLV.
* No. of Open Complaints negatively impact CLV.
* Income and Gender has no effect on CLV.