SUMMARY

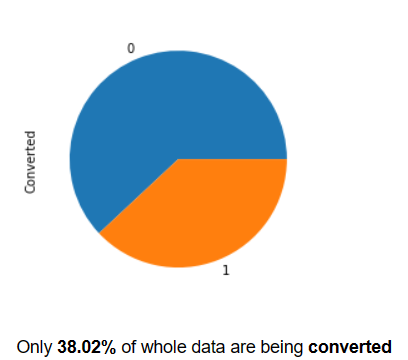
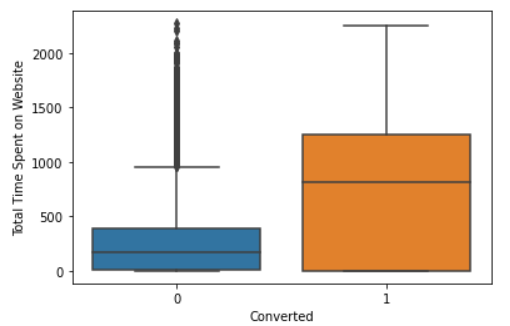
**Problem Statement:** X Education company, sells online courses to industry professional. Although it has great number of leads, the conversion rate is 30%. The management wants this conversion rate to become 80% or so. Additionally they want to attached a lead score against each leads between 0-100.

The need is to build a model to find ways to scale up the lead conversion rate.

**Model chosen**: Logistic Regression model

**Methodology and Steps:**

1. **Data Understanding**:- The dataset provided the information about how the customers visited the site, amount of time spent and the conversion rate along with many other parameters.
2. **Data Cleaning** :-
3. The categorical variables with label ‘select’ was replaced with NaN values because it was equivalent to null values.
4. Columns with high null values of more than 45% were dropped.
5. Some of the categorical columns as Tags, Country, Specialization was imputed with either the mode value or other labels.
6. Imbalanced columns were dropped.
7. Total Rows and columns before cleansing: 9240, 37
8. Total Rows and columns after cleansing: 9240, 30
9. **EDA:-**
10. Univariate, Bivariate and Multivariate analysis of converted, Total visits, Total Time Spent, Page views Per Visit was done.
11. Data imbalance was checked, only 38.02% of leads got converted. (graph below)
12. The target variable "Converted" has a medium correlation with variable "Total time spent on website".
13. Total visits and Page views per visit have a high correlation.

1. **DATA PREPARATION**:-
2. Binary Mapping and dummy features for categorical variables were created
3. Train & Test Sets in 70:30 ratio was done
4. Feature Scaling using Standardization.
5. **MODEL BUILDING**:-
6. Selected 15 features using RFE.
7. Variables with p-value> 0.05 were dropped.
8. Total 3 models were built with 14 variables having GOOD VIF
9. **MODEL EVALUATION**:-
10. Confusion matrix with cutoff value 0.3 was selected based on ROC curve.
11. Since the precision-recall view gave performance metrics around 75%, we preferred sensitivity-specificity score.
12. Evaluation metrics for train & test are very close to around 80%.

7. **CONCLUSION**:- The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion:

1-Lead Source Reference

2-Lead Source social media

3-Lead Source Olark Chat

Team Names:

1. Anupriya
2. Anwesha
3. Anuradha