-Imported all the necessary libraries and csv file for the analysis.
- there 9240 rows and 37 columns.
- from the above the information we can see that there are 7 numerical variables and rest 30 are categorical variables.

-From statistical info, we can see that there are outliers present mostly in 'totalVisits','Total Time Spent on Website' and 'Page Views Per Visit'.

1.Observations from EDA Process -

- Maximum lead conversion happened from Landing Page Submission.
- Major lead conversion in the lead source is from 'Google'
- Major lead conversion is from the Unemployed Group
- Major lead conversion from TotalVisits, Total Time Spent on Website, Page Views Per Visit
- Major conversion has happened from the emails that have been sent

After checking the counts of each variables to find insights :

- we will be removing these variables
    - Prospect ID - not required
    - Lead Number - not required
    - Country- not required
    - Receive More Updates About Our Courses- column only has 'No' doesn't makes sense to keep it.
    - Update me on Supply Chain Content - column only has 'No' doesn't makes sense to keep it
    - Get updates on DM Content - column only has 'No' doesn't makes sense to keep it
    - I agree to pay the amount through cheque -column only has 'No' doesn't makes sense to keep it
    - Magazine - column only has 'No' doesn't makes sense to keep it
- We will transform below columns of yes/no category to 1/0:
    - Do Not Email
    - Do Not Call
    - Search
    - Newspaper Article
    - X Education Forums
    - Newspaper
    - Digital Advertisement
    - Through Recommendations
    - a free copy of Mastering The Interview

2. Data Cleaning

1. Cleaning the dataset by removing the redundant variables/features.
2. After removing the redundant columns, we found that some columns are having label as ‘Select’ which means customer chose to not answer this question. Thus we would label null value to 'select' label.
3. Remove columns having more than 40% null values
4. Imputing missing values as per column data available

-we see there are columns having more than 40% missing values, so it is better to remove these columns as it imputing them could lead to bias predictions.

-dropping columns having missing values above 40%

-We get that the columns above, 'Last Activity', 'Tags' are provided by sales team. We will remove them before model building as the we don't a model having these features.

-Data is skwed, we are going to replace these labels (Facebook, bing, Click2call, Live Chat,Press_Release, Social Media, testone, WeLearn, blog, Pay per Click Ads, welearnblog_Home, youtubechannel, NC_EDM) in one label as 'Others'.

-we will deal with missing values by imputing missing values with max occuring label

-We will create another category for missing values as the count is very high and imputing missing values with median can lead to misleading results.

3. Data Transformation

1. Converting yes/no category columns to binary form 1/0.
2. to deal with columns having outliers will create bins for them.
3. will remove all the redundant and repeated columns.
4. create dummy variables

-As we can see there are outliers in 2 variables 'TotalVisits' and 'Page Views Per Visit'.

4. Data Preparation

1. Split the dataset into train and test dataset and scaled the datasets.
2. After this, we plot a heatmap to check the correlations among the variables.
3. check heatmap for highly correlated features.

-we couldn't find much which features are highly correlated and to drop thus we will now proceed with building our model and based on the p-values and VIFs, we will again check for correlation.

5. Building a Model

- We are going to use hybrid model creation using RFE and manual features selection.
- We will drop features having insignificant values one by one and create new models until all the features attain significant p-value<0.05 and vif-values < 4.1m.

ROC Curve Plotting

- ROC curve shows the trade off between True postive rate and False positive rate - means if sensitivity increases specificity will decrease.
- The curve closer to the left side border then right side of the border is more accurate.