Data Cleaning and Visualization:

1. Deleted all the columns that has null values more than 30%
2. Checking other features with null value graphically and see the no of select values.
3. Features that have combined null and select value more than 30% was dropped.
4. Features with ~ 15 % Null values are analyzed and found that they are Imbalance feature. Have one category value very high compared to others. Hence dropped them.
    - How did you hear about X Education
    - Specialization
    - City
5. For Features with less than 5 % null values dropped the null value rows.
6. Checked for any other category value with Select value. Found no features hence concluded the data cleaning.

Data Visualization:

- Did Numerical variable analysis- Observed that more the users spend time on the site and more the users visit the site has high chances of conversion.
- Checked Correlation and made same inference that no of visit and time spend is a good feature for prediction.

Data Preparation:

1. Converted categorical variable to dummy variable.
2. Converted features with Yes/No category to binary columns. Yes is 1 and NO is 0.
3. Split the data in Train and set with 70:30 ratio
4. Did Minmax scaling for Train set using Fit_transform
5. Did Minmax scaling for Train set using transform.

Model Building:

1. Feature selected using RFE selecting 20 features using Estimator as LogisticRegression()
2. Build model 1 using the RFE supported features .
3. Checked Summary and VIF.
   Observation:
    - PValue: Following features have p value more than 0.05 - Lead Source_Social Media,Lead Source_google
    - VIF : value still high for Lead Origin_Lead Add Form, Lead Source_Reference, Lead Source_Welingak Website

4. Build further models eliminating features with high pvalue and VIF and reached final model with 16 features with all P values and VIF Under control.
   Final Features:
   1. 'Do Not Email'
   2. 'TotalVisits'
   3. 'Total Time Spent on Website'
   4. 'Lead Origin_Lead Import'
   5. 'Lead Source_Olark Chat'
   6. 'Lead Source_Reference'
   7. 'Lead Source_Welingak Website'
   8. 'Last Activity_Converted to Lead'
   9. 'Last Activity_Email Bounced'
   10. 'Last Activity_Had a Phone Conversation'
   11. 'Last Activity_Olark Chat Conversation'
   12. 'Last Notable Activity_Email Link Clicked'
   13. 'Last Notable Activity_Email Opened'
   14. 'Last Notable Activity_Modified'
   15. 'Last Notable Activity_Olark Chat Conversation',
   16. 'Last Notable Activity_Page Visited on Website'

Model Evaluation:

- Predicted the Y value for Train data
- Created a data frame with the actual converted column, Leads score
- Set 0.5 cut off and found new converted. Accuracy and the sensitivity/recall has to be balanced. High False negative can cause the loss of customers who were potential leads. Hence the recall has to be high.
- Did cut off optimization using the multiple ways like Accuracy, Sensitivity and specificity trade off, Checked the sensitivity accuracy and sensitivity of all the threshold.
- Finalized the cut off as 0.33.

Model Evaluation in Test Data:

- Applied the same model in Test data.
- Predicted the Leads score.
- Applied cut off of 3.3 .
- Find the Summary of Train and test data:
  Train data Summary:
    - Accuracy : 0.7995591245473154
    - Sensitivity: 0.8205233033524121
    - Specificity: 0.7810499359795134
    - Precision: 0.7012578616352201
    - Recall : 0.8205233033524121

  Test Data Summary:

- Accuracy : 0.7818582445831803
- Sensitivity: 0.7906976744186046
- Specificity: 0.7768166089965398
- Precision: 0.6689478186484175
- Recall: 0.7906976744186046