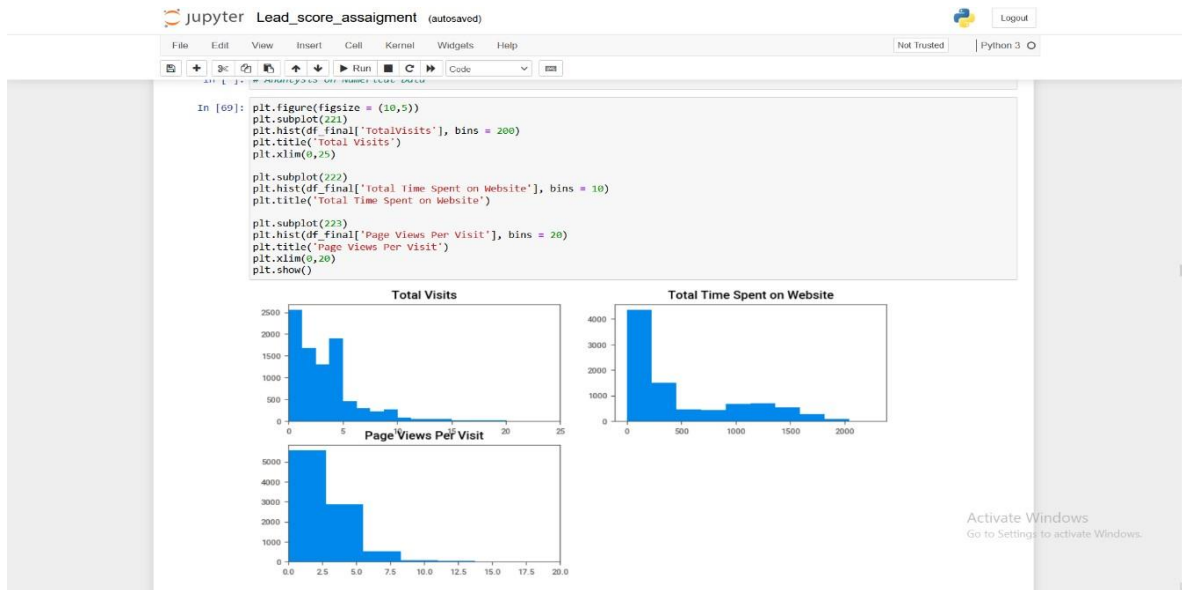



SUMMARY

- **Data checks and data cleaning is performed (missing value imputation, removing duplicate data and other kinds of data redundancies, etc.) in the following ways:**
 1. Use Sweet viz library we got to know that there is select value which would be filter is select so replace that by nan and lower case select 's s alphabet.
 2. Checked for distinct value and remove them as they cannot be used for analysis.
 3. Handle duplicate Value
 4. We have dropped columns which has more than 35% row are null.
 5. We have replaced few null values by not provided as those columns are required for further analysis.
 6. For other countries we renamed them with outside of India as we have very less result for outside of India country.
- **In the second step we did EDA (exploratory data analysis) univariate and bivariate analysis and got some valuable insights**



- **Dummy variable for categorical columns**

Also make sure that first column of dummy variable has been dropped as it is obvious.










 jupyter leadsssss (autosaved)

Not Trusted


Python 3

Logout

File Edit View Insert Cell Kernel Widgets Help



Code



```
In [74]: # Dummy Variable
df_final.loc[:, df_final.dtypes == 'object'].columns

Out[74]: Index(['Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call',
               'Last Activity', 'Country', 'Specialization',
               'What is your current occupation',
               'What matters most to you in choosing a course', 'Search',
               'Newspaper Article', 'X Education Forums', 'Newspaper',
               'Digital Advertisement', 'Through Recommendations',
               'A free copy of Mastering The Interview', 'Last Notable Activity'],
              dtype='object')
```

```
In [75]: dummy = pd.get_dummies(df_final[['Lead Origin', 'Specialization', 'Lead Source', 'Do Not Email', 'Last Activity', 'What is your c
# Add the results to the master dataframe
df_final_dum = pd.concat([df_final, dummy], axis=1)
df_final_dum
```

Out[75]:

	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	...	Last Notable Activity_Form Submitted on Website	Last Notable Activity_Had a Phone Conversation	Last Notable Activity_Modified	Last Activi Conv
0	API	Clark Chat	No	No	0	0.0	0	0.00	Page Visited on Website	not provided	...	0	0	1	
1	API	Organic Search	No	No	0	5.0	874	2.50	Email Opened	India	...	0	0	0	
2	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.00	Email Opened	India	...	0	0	0	
3	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.00	Unreachable	India	...	0	0	1	
4	Landing Page Submission	Google	No	No	1	2.0	1428	1.00	Converted to Lead	India	...	0	0	1	
...
9235	Landing Page Submission	Direct Traffic	Yes	No	1	8.0	1845	2.67	Email Marked Spam	Outside_India	...	0	0	0	
9236	Landing Page Submission	Direct Traffic	No	No	0	2.0	238	2.00	SMS Sent	India	...	0	0	0	

- **Building The Model using Train and Test split**

- We have split the data between Train and test in proportion of 70:30
- Target Variable is converted and rest of them are factors variable.
- With the help of MinMaxScaler scaling "TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'".
- Use Logistic regression
- Create the model with 15 Variable for RFE
- Removed all the variable having P values greater than 0.05.
- We have dropped "Last Notable Activity_Had a Phone Conversation", "What is your current occupation_Housewife" and "What is your current occupation_Other" one by one and run the model.
- Made the predication on Test data
- We have achieved accuracy around 81.02%
- In the ROC curve we could deduce the optimal point near to .37.

- **We made the following analysis from the model.**

Below are the factors which has high impact on conversion rate.

1. The total time spent on Website.
2. If the source is following below
 - 2.1 Google
 - 2.2 Direct Traffic
 - 2.3 Organic Search
3. Last Activity
 - SMS
 - Olark chat conversation
4. When the current occupation is a working professional.

- **In order to handle the future challenges**

- Make website more attractive as most leads are converting through Website
- Last activity is coming through Olark chat and SMS
- Mostly working professional are attractive so tailor make the courses as per their needs for targeting the potential clientele.
- We can also some education benefits for unemployed or students to churn higher rate and provide them more information regarding the course and how it can help them in pocketing higher package.
- They Can create the chat robots for genuine calls and chat and they can do further analysis based on data in chat robots.