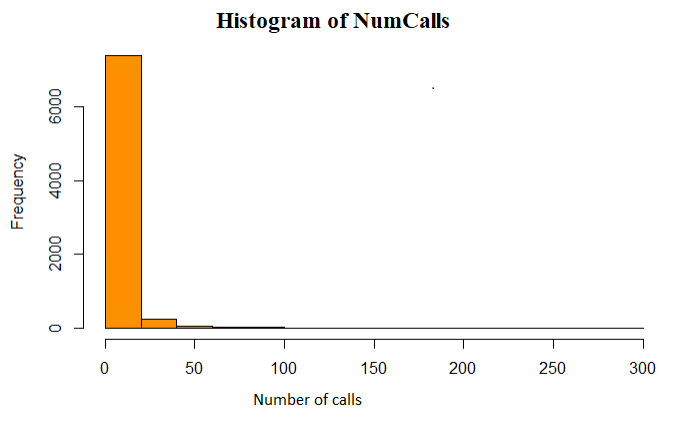
**Description of the task**

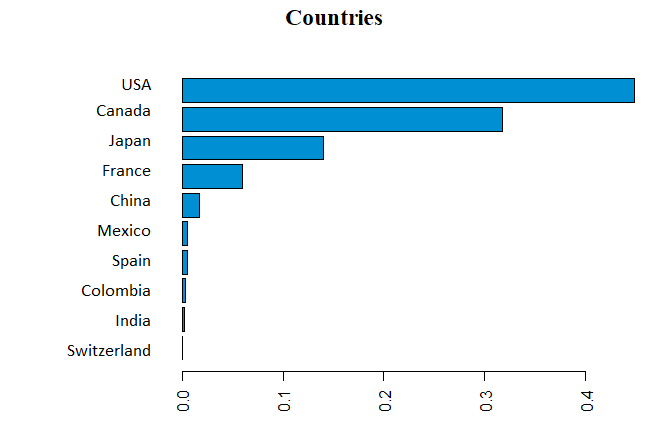
In a spreadsheet Customers2.xlsx each row represents a customer and his activity (how many phone conversations he had, how many times he downloaded a product brochure, how many times he visited the company website, etc.). The last column says if he eventually purchased the product or not.

The environment should be: Jupyter Notebook, Python 3, scikit-learn, pandas. Add other libraries, if needed. Add comments to the code.

1. Conduct some Exploratory Data Analysis (your decision, what you want to include).
2. For one of the numerical columns (e.g. NumCalls) crate a histogram showing how many times number of calls is between 0 and 19, between 20 and 39, between 40 and 59, etc. (this is sample graph from another dataset, where there was over 7000 NumCalls between 0 and 19).



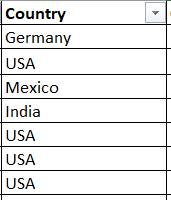
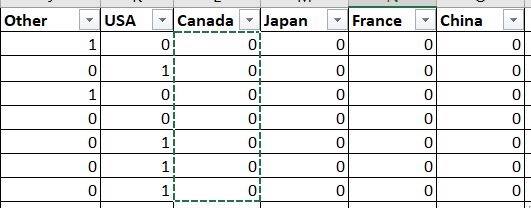
1. Create a bar chart showing what is a percentage of a specific country in the column Countries (this graph is to just show look and feel, the actual percentages will be different):



1. Add a column TotalInteractions that contains sum of the columns NumCalls, NumEmails, NumDownloads, NumEvents, and NumForms.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NumCalls | NumEmails | NumDownloads | NumEvents | NumForms | TotalInteractions |
| 33 | 19 | 1 | 0 | 1 | 54 |
| 0 | 1 | 0 | 1 | 0 | 2 |
| 4 | 6 | 2 | 0 | 0 | 12 |
| 0 | 10 | 0 | 0 | 1 | 11 |
| 7 | 0 | 0 | 0 | 0 | 7 |
| 10 | 1 | 0 | 1 | 0 | 12 |
| 5 | 0 | 0 | 0 | 0 | 5 |
| 1 | 0 | 0 | 0 | 0 | 1 |

1. Replace column “Country” with one column per country:

 => 

Actually, do it only for top 5 countries:

|  |
| --- |
| USA |
| Canada |
| Japan |
| France |
| China |

All other countries add to the column **Other**

4. Create classification machine learning model that predicts last column MadeAPurchase.

5. Show some metrics of your choice (e.g. precision, recall, ROC).

6. Preform some sample hyper-parameter tuning.