# Instructions to Run Code from DOS prompt:

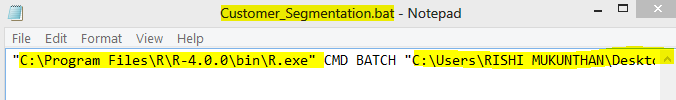
**List of Coding Files that are submitted:**

**R:**

1. Customer\_Segmentation\_EDA.R
2. Customer\_Segmentation.R
3. Customer\_Segmentation\_EDA.bat
4. Customer\_Segmentation.bat

For running R scripts files from the windows command prompt, a batch file (.bat) need to be prepared. I have attached the .bat files for the 2 R scripts that I am attaching.

You can just run the .bat file which will open the command prompt and start executing the R script. Before running the .bat file, edit the .bat file path with the corresponding R installation path and path of R script to be run.

* Example: My R installation/R.exe file path is **"C:\Program Files\R\R-4.0.0\bin\R.exe"**
* R script path: **"C:\Users\RISHI MUKUNTHAN\Desktop\Data Science\Projects\Edwisor Customer Segmentation\Final\R\Customer\_Segmentation.R"**

**What each R script does?**

**Customer\_Segmentation\_EDA.R** – This is the R script where I have performed Exploratory Data Analysis for the Credit Card dataset. The Rout file would contain the R console outputs and the EDA\_R.pdf contains all the plots generated to explore the distribution of variables and boxplots for outlier analysis.

Make sure to edit the setwd path inside the R script and place the input data set in the current working directory in your system.

**Input**: credit-card-data.csv

**Output:** Missing\_perc\_R.csv, EDA\_R.pdf, Customer\_Segmentation\_EDA.Rout

**Customer\_Segmentation.R** – This is the R script where I have performed data preprocessing, performed PCA (Dimension Reduction), implemented K-Means clustering algorithm.

The Rout file would contain the R console outputs.

Customer Segmentation R Plots.pdf contains all the plots generated including the cluster analysis results, final results interpretation plots etc.

Clustering\_Output\_R.csv contains the output data set with labeled classes for each observation based on the clusters identified.

**Input:** credit-card-data.csv

**Output:** Clustering\_Output\_R.csv, Customer Segmentation R Plots.pdf, Customer\_Segmentation.Rout

**Python:**

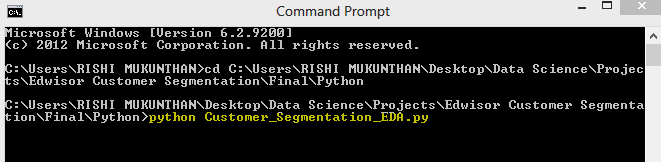
1. **Customer\_Segmentation\_EDA.ipynb**
2. **Customer\_Segmentation.ipynb**
3. **Customer\_segment\_Prediction\_Model.ipynb**
4. **Customer\_Segmentation\_EDA.py**
5. **Customer\_Segmentation.py**
6. **Customer\_Segment\_Prediction.py**
7. **Customer\_Segment\_Prediction\_RawData.py**

**File 1.** The jupyter notebook file where I have performed EDA.

**File 2.** The jupyter notebook file where data preprocessing and modelling has been done on the given dataset.

**File 3.** Additional work I did to implement logistic regression model to train on the labeled data I generated after clustering.

**File 4.** Is the .py script which is to be executed from the windows command prompt by running with the below command. Open command prompt from the path where you have downloaded the code files in your system and run the python command to run the .py scripts.



While running the .py script for each EDA plots there will be popup window with the image which needs to be closed to proceed with the execution of the script.

**Input:** credit-card-data.csv

**Output:** 35 image png files of EDA

**File 5.** The python script which contains the data preprocessing and modelling code.

**Input:** credit-card-data.csv

**Output:** 14 png files with plots of PCA, K means and final results interpretation, Clustering\_Output\_Python.csv

**File 6.** The python script takes in the labeled output data from clustering and trains on the data. Training and evaluation metrics like confusion matrix and accuracy score will be printing in the cmd console.

It outputs the scaler and model pickle file.

**Input:** Clustering\_Output\_Python.csv

**Output:** customer\_segment\_pred\_model.pkl, customer\_segment\_pred\_scaler.pkl

**File 7.** This python script can take any new future customer data which is in the credit-card-data.csv format and predict the label/group the customer observation belongs to.

**Input:** credit-card-data.csv , customer\_segment\_pred\_model.pkl, customer\_segment\_pred\_scaler.pkl

**Output:** data\_predicted.csv

**Python Libraries and their version used in the project:**

fancyimpute 0.5.4

joblib 0.13.2

matplotlib 3.3.0

numpy 1.19.0

pandas 1.0.4

scikit-learn 0.23.1

seaborn 0.9.0

tensorflow 2.0.0

**Libraries loaded in R:**

"ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced", "C50", "dummies", "e1071", "Information", "MASS", "rpart", "gbm", "ROSE", 'sampling', 'DataCombine', 'inTrees’, ‘tidyverse’, ‘funModeling’, ‘Hmisc’,’ factoextra’,’ dplyr’