CENG 4512 Data Science and Analytics Midterm Homework
2020/2021 Fall
Online Shoppers Intention Analysis

Download Online Shoppers Intention dataset from
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

This dataset contains feature vectors belonging to 12,330 online sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset consists of 10 numerical and 8 categorical attributes.

Important! The '**Revenue**' attribute can be used as the **final class label**. This label is given for only comparative purposes. Here in this homework, you will assume you don't know the labels for shoppers.

Further details of the dataset can be found in https://link.springer.com/article/10.1007/s00521-018-3523-0

Tasks:

1. Your main focus for the first task should be discovering **hidden patterns** of online shoppers on this dataset using **data visualization and summary statistics**. You can work/relate any of the features but make sure you create advanced plots with **ggplot2** and use **dplyr** for data pre-processing. Make sure you evaluate your results.

   Important! Make sure you consider all features and notice some of them are characters or other types. You may convert some features for better analysis.
   Important! Some of the levels are not necessary for analysis and you may need to clean the dataset.
   Important! Lastly, decide on a segmentation of the dataset in this stage and follow up with the rest of the tasks. Explain the reasons of your segmentation.

2. Your main focus for the second task should be reducing dimensions using a **PCA analysis**. Make sure you use the findings in the previous step, and evaluate and visualize your results.

3. Your main focus for the third task should be **unsupervised learning**. Make sure you use the findings in the previous steps. Please consider different clustering approaches and different parameter settings. Please use Revenue labels to evaluate your results.

Details about submission:

1. You are supposed to submit a **statistical report** of your findings ideally via **RMarkdown** by 7th of Dec (24:00).
2. Please post your.html files. The name of the files should be your name_surname_hwmidterm (like eralp_dogu_hwmidterm.html) through DYS. Make sure your html file includes both code and results.