Online Shoppers Intention

Problem Statement:

Based on given data of visitors browsing for online shopping, build different clusters to know whether person is only browsing and visiting multiples pages or also generating revenue for the shoppers as well. Analyse and compare the clusters formed with the existing Revenue Column.

Data Set Information:

The dataset consists of feature vectors belonging to 12,330 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.

Attribute Information:

The dataset consists of 10 numerical and 8 categorical attributes.
The 'Revenue' attribute can be used as the class label.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories.

The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another. The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date.

Citation / Reference:

Please use the below link to cite this dataset:
Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018).

https://link.springer.com/article/10.1007/s00521-018-3523-0

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Expected Approach/Outcomes:

- Perform required cleaning to bring the uniformity in the data.
- Carry-out uni-variate, Bi-variate and Multti-varaiate analysis to understand the data relationships.
- Perform required missing value treatment
- Perform Outlier treatment if required
- Perform appropriate scaling
- Perform required encoding techniques
- Build the different cluster models.
- Analyse the optimum number of cluster using appropriate techniques.
- Make the appropriate business interpretation using the cluster centroids.
- Perform the EDA on cluster groups to understand the cluster characteristics.
- Perform PCA and apply clustering on top of it. Comment whether PCA is really helping the clustering process.
- Also try different graphs to visualize the clusters and its characteristics.