Project 2 – Wholesale Customer
Due on Thu 10th March 11:59pm EST

Context: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories.

Dataset Source: Excel File attached on eClass

Task: Goal of this project is to best describe the variation in the different types of customers that a wholesale distributor interacts with.

Implementation:

- Perform EDA and any data cleaning if necessary.
- Implement Feature Scaling to Normalize the data(compare the histogram/KDE for MinMaxScaler and StandardScaler). Choose one of the Scaler to proceed ahead and provide reasoning as to why it was selected?
- Find optimal number of features using RFECV and show the plot between Number of features selected vs Cross validation score (use channel as target variable)
- Implement KMeans Clustering for K=2 to K=15 and based on elbow method identify what is the optimum number of clusters
- Implement PCA with number of original features to answer how much variance is explained by first 2 components and by first 4 components and visualize the clusters in the data
- Implement XGBoost Classifier with 5 Fold CV and report the performance metrics

Submission Instructions: Please just submit one jupyter notebook containing all the code and make use of markdown cells to include the comments, answers, reasoning, analysis, etc.

Note: Name of your file should be your "Project2-id_Firstname_Lastname.ipynb"