CEBD 1260 – Spring 2019 – Team Project

Selection of a Dataset of Interest

- **List of team members**
  Ho Tak So (Frank So)
  Ricardo Luis da Costa Rocha

- **Link to the github repository**
  https://github.com/coastrock/CEBD1260_team_project.git

- **Link to the dataset**
  https://www.kaggle.com/mehdidag/black-friday

- **1-2 Paragraph(s) describing the dataset**
  Descriptions given by Kaggle:
  "Dataset of 550 000 observations about the black Friday in a retail store, it contains different kinds of variables either numerical or categorical. It contains missing values"
  "The dataset is a sample of the transactions made in the store that wants to know better the customer purchase behaviour against different products. Specifically, it is a regression problem where they are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables."
  "Classification problem can also be settled in this dataset since several variables are categorical".
  This dataset is also particularly convenient for clustering and maybe find different clusters of consumers within it."

- **2-3 bullet points with potential problems/opportunities**

  1. Use case: Black Friday consumers are primarily single young male
     We could check if this is true analyzing this dataset
  2. "Predicting the age of the consumer"
     This prediction could be interesting for the store increase its offers in products according costumers age
  3. "Predicting the category of goods bought"
     This prediction could be interesting for the store improve the most worthy categories of goods bought