Instructions

- This homework assignment is worth 63 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: February 3, 2023 by 11:59 pm.**

For this homework assignment, we will consider the `MarketingData.csv` data file. This data file contains information related to the annual spend amount of each of the 20,000 customers of a major retail company. The marketing team of the company used different channels to sell their goods and has segregated customers based on the purchases made using different channels, which are as follows:

- 0 → Retail
- 1 → Road Show
- 2 → Social Media
- 3 → Television

As the data scientist in charge, you are tasked with building a machine learning model that will be able to predict the most effective channel that can be used to target a customer based on the annual spend on the following seven products sold by the company: fresh produce, milk, grocery, frozen products, detergents, paper, and delicatessen. **In Python**, answer the following:

Exercise 1

(5 points) Upload the `MarketingData.csv` data file to your S3 bucket, and using the pandas library, read the `MarketingData.csv` data file and create a data-frame called `marketing_data`.

Exercise 2

(3 points) Report the number of observations in each of the marketing channels.

Exercise 3

(15 points) Create two visualizations that may show interesting relationships between the input variables and the target variable. Make sure to describe the visualizations.

Exercise 4

(40 points) Split the data into two data-frames (taking into account the proportion of 0s, 1s, 2s and 3s in `Channel`): `train` (80%) and `test` (20%). Then, do the following:

(i) Using the one-vs-rest strategy and one model of your preference, build the multi-class classification model. Then, predict the `channel` label on the `test` data-frame. Create the classification report.

(ii) Using the one-vs-one strategy and the same model from part (i), build the multi-class classification model. Then, predict the `channel` label on the `test` data-frame. Create the classification report.

(iii) Based on your results from parts (i) and (ii), what framework would you use to predict `channel`? Be specific.