IE 8990 Distributed Machine Learning Techniques
Assignment 2
Due: 10/11/2021 11:59 PM

Use the **cruise_ship_info_missing.csv** file to answer the following questions:

Question 1: Perform some basic statistics with the dataset.

Question 2: Calculate the cardinality for each variable in the dataset. Can you reject any variables based on cardinality? (Hint: Look for cardinality of 1)

Question 3: Calculate the missing value percentage for each variable in the dataset. Can you eliminate any variables based on missing values? (Hint: Any value that has more than 10% missing values)

Question 4: Impute the continuous variable columns with median value. (Hint: Use 0.1 for relativeError)

Question 5: Use your subject knowledge to impute the **Cruise_line** variable (categorical variable)