Texas Christian University
COSC 40023 Spring 2025
Assignment 1

Due: Feb 3rd, Monday, at 11:30 PM (Late submission NOT accepted)
Submission (two files, no compression): `assignment1.py` and `assignment1.R` through TCU Online

Download the dataset called "Customer_Data.csv", and then preprocess the data in both Python and R. Two files should be created named `assignment1.py` and `assignment1.R`, respectively. Here are some additional requirements.

1. Take a look at the raw data. We decided to choose the last column as a dependent variable, and the rest of the columns as independent variables. Based on our decision, what problem are we trying to solve through machine learning? Is the problem a regression problem or a classification problem? Put your answers as comments at the very beginning of `assignment1.py`
2. Missing ages should be replaced by the median of the ages, and missing salaries should be replaced by the mean of the salaries.
3. Categorical data should be encoded properly.
4. `1/3` of the data should go to the test set. In addition, `random_state` must be set to `0` in Python, and `seed` must be set to `123` in R.
5. Feature scaling should be conducted using normalization.
6. Have sufficient single-line and multi-line comments.