Please complete the project as *ONE PYTHON FILE ONLY.*

## Dataset Overview

For this project we are going to perform a number analytic tasks on the humanDetails.csv file. This dataset contains details about individuals such as their age, salary range, relationship, country and etc.

## Project Specification

The objective of this project is to provide an insight into the underlying pattern of the dataset. Please perform the following tasks:

1. Task1: For each known county use Work-class and Age to predict Income with the following setting:
   * 1. Any record associated with an unknown cell in work-class should be removed from the dataset.
     2. Some values in Age columns are represented as decade, e.g., 20s or 30s. You are required to clean these values by removing the s and convert it into an integer value.
     3. Income is either less or equal than 50k or greater than 50k.
     4. Cross validation using 5 folds and 20% test size.

For each country run decision tree classifier and try different values for the depth of the tree. Pick the depth that produces highest accuracy for test set. And finally pick those countries that their process of learning still suggest overfitting and visualize them using appropriate visualization technique to display the gap between training and test. In this task overfitting occurs if the gap between training and test is more than 20%.

Interpret the results.

1. Task2: Use hours-per-week, Occupation, Age and relationship to predict income. Apply *Decision Tree* and *K Nearest Neighbour* classifiers to compare their accuracy (test and training).
   1. Any unknown cell in the Occupation column should be filled with the Occupation with the highest frequency.
   2. Some values in Age columns are represented as decade, e.g., 20s or 30s. You are required to clean these values by removing the s and convert it into an integer value.
   3. One of the values of relationship attribute is *Other-relative* that needs to be removed from the dataset.
   4. Some values in hours-per-week attribute mentioned only once. Those values should be removed from the dataset.
   5. Income attribute should have two unique values.
   6. Use cross-validation with 5 folds.
   7. Try different values for the classifiers’ parameters.
   8. Visualize the results and interpret any pattern you find.

Interpret the results.

1. Task3: Use *age*, *fnlwgt*, *education-num* and *hours-per-week* to train a model with K-Means. Apply the following settings:
   1. Some values in Age columns are represented as decade, e.g., 20s or 30s. You are required to clean these values by removing the s and convert it into an integer value.
   2. Cluster the data into two clusters.

Reduce the number of features to two features. Use the first new feature as X coordinate and the second new feature as Y coordinate and visualize them twice using appropriate visualisation technique as follows:

* 1. The color of each individual in the first visualization is decided based on the value in *Income* attribute.
  2. The color of each individual in the second visualization is decided based on the label that is generated by the clustering algorithm.
  3. Visualize the two plots next to each other.

Interpret the results.

Note that visualization plots need to have proper labels and annotations.

Please use the python file template that is provided for you and complete the project in that file. Each task needs to be implemented as a separated function with interpretation as a comment below the function.

There should be correct task implementation (model training, accuracy reporting, visualization if needed and etc).