Fourth Industrial Revolution (4IR) Summer School
Data Preparation – Day 3 Exercises

Question 1

Access the given URL (https://github.com/awesomedata/awesome-public-datasets) to get access to the dataset available in this page. Use the given "Clone with HTTPS" link to bring the data in your colab. The given dataset consists of 12 variables and 891 data points.

- First, write a python program that access the csv file in the given Zip file.
- Read the csv file to a dataframe called "org_data".
- Use the describe() function to provide a statistical information about the variables in the dataset.
- Extract four columns (Name, Sex, Age and Embarked) to create a new dataframe called "sub_data". Use the "Name" column as an index of the "sub_data" dataframe.
- Use the following conditions to extract a subset of the data stored in "sub_data" datfarme:
- All male passengers under 18 who are in the Embarked class Q or S.
- Export all the last two dataframes in csv files (use the same names given to the dataframes)

Question 2

Write Pandas program to read the given "stud_scores.csv" file to a dataframe called "scores".

- Sort the data ascendingly using the final project score and keep the results in a new dataframe "sorted_scores".

- Add a new column “Total” to the “scores” dataframe to keep the total scores for each student.

- Rank the dataframe according to the new “Total” column.

Question 3

Two different datasets for wart treatment are presented in “UCI Machine Learning Repository”.

- Cryotherapy Dataset (https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+)

- Immunotherapy Dataset (https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset)

Read the two datasets to dataframes, then find the correlation matrix for each dataset.

Although the datasets have the same variables, the dataset has an additional variable “induration_diameter”. Thus, drop the additional column from the Cryotherapy Dataset. Then merge the two datsets together in one dataframe.

Finally, find the correlation matrix again for the merged dataset and compare it with the two generated earlier.

Setting Up Kaggle in Google Colab (Useful link: https://towardsdatascience.com/setting-up-kaggle-in-google-colab-ebb281b61463)