Assignment 4

Data wrangling is done to process the collected data in various formats to be able to analyze them in a better way.

Data Gathered:

The data I was able to find was – Average height by country for males and females 2022 & Average literacy rate by country for males and females 2022

The 2 sources I found on web are Kaggle & World Population Review

Kaggle - https://www.kaggle.com/datasets/majyhain/height-of-male-and-female-by-country-2022

World population review - https://worldpopulationreview.com/country-rankings/literacy-rate-by-country

After combining the data:

In order to work and visualize the data, we need to combine the 2 datasets, making sure no information is lost. So, first we combine the data, make sure that there are no missing values. Combining of the data was done by using the data wrangling method – 'merge'. There is a common column between both the datasets – 'country'.

However, before finalizing this as the final data, the missing values should be removed. Otherwise, it will be an issue and we will not get proper values.

Part of the new data selected:

Data to be worked on or used as a training data is usually a part of the entire data – so we select some part of the data. I selected a few rows and 2 columns.

Transforming the data: - Using clustering & k means

I used the clustering technique k means to predict the data ad then plot the output.

Visualize the Data:

I used matplotlib library and scatter plot to visualize the prediction. We see how using the 100 rows for the particular columns, the new data is predicted for the rest of the data. One issue with this is that all the rows & columns are not inter dependent. However, we can see from the output, that the clustering

was done in a pretty effective way. There are 4 clusters used for clustering. And the data has been divided into clusters depending on the column values (average male and female height). Using the training data, the test data gets divided into the clusters. If the clustering was not done properly, we would see a mix of colors all over the place. However, as we can see, with the increase in average eight value, the cluster the row is divided into is changed. Any country with average height more than 162.5 is divided into cluster 2 or 3 depending on the x-axis value as well. That means the average value of male& female for a country is used together (as a 2d array we can think), to divide the values into clusters.