**Principles of Data Science**

**Assignment 1**

**Student name: Lingojigari Bharathchandra**

**Student ID : 16337200**

1.

The three stages of Reproducible workflow are given below:

Stage 1 : Data Collection :- The data is collected from the available sources.

Stage 2 : Data Cleaning :- Looking at the data available, we can be able to know where the data is unclear or any changes to be made, for example adding values to null values will help in the processing of data.

Stage 3 : Data Analysis :- To figure out what can be analyzed, we have to look for features that are present in data and plot them for visualization , as it helps in understanding the unique features.

The given data :

The raw data :

Chart

Description automatically generated with medium confidence

Cleaned data with added BMI :

Table

Description automatically generated with medium confidence

Visualizations :

The following scatter plot is the comparison of Age and the BMI. We can see that the BMI is varying with the age.

The histogram is the analysis of BMI and height(in inches). We can see the BMI is fluctuating with the height.

The line plot is comparing Weight of the person measured in pounds versus BMI. BMI is parallel to the Weight. If the weight increases, BMI increases and vice versa.

2.

The Three stages of Workflow :

Stage 1 : Data Collection :- The data is collected from StudentPerformance.csv

Stage 2 : Data Processing :- The raw data is processed into a format that can be used for the analysis. For example, The total column and Rank is added so that we can give a better plots for understanding.

Stage 3 : Data analysis :- The plots are driven by features. The insights are easily seen through the graphs.

The given Data :-

Table

Description automatically generated

The Cleaned and processed Data :-

Table

Description automatically generated with medium confidence

Visualizations :-

The Below plot is a histogram plotted by gender versus total marks. Its been seen that the female marks are par higher than males’ marks.

The below plot is histogram plotted by the total marks vs race/ethinicity. It is seen that Group E has most total marks and Group A has lowest total marks.

This plot is between parental level of education and marks. Its seen that parents having masters degree has their children with most marks and those who are high school has lowest marks.

Reading vs Math

The below plot is reading marks vs math marks . its seen that most people have them both in common and have the similar marks.

Rank vs Math

The below plot is math marks vs rank. The rank is more since the math marks are more. It means the rank is less if math marks are more.