**Principles of Data Science**

**Assignment 1**

**Partha Nossam (16326619)**

Question 1:

Three stages of reproducible workflow.

1. Data Collection.
2. Data processing.
3. Data Analysis.

1. In data collection process, we need to collect the data from the source. Here we have the sample dataset. We need to save it as CSV file. Create a file with metadata and save the variable names.

A screenshot of a computer code

Description automatically generated

2. Data cleaning: in this process, we cleaned the null values, extra columns which are not used for business. Here, we can remove AGE column as it has negative correlation.

We keep the cleaned data in the data clean sub folder to avoid confusion between raw data and cleaned data.

A screenshot of a computer program

Description automatically generated

3.In data analysis, we want to predict a variable yes or no. We can use Logistic Regression classifier model to predict YES for frailty and NO for frailty. We can use different models based on our requirement like Linear regression, SVM.

Workflows:

Compute model

Logistic Regression model (to predict binary outcome)

Performance

tuning

output

Accuracy>95

run the model against data

Start

Identify the useful data

Data preprocessing

Analyze the dataset

Collection of data

Select diff data

NO

NO

YES

YES

Data Input and Processing Data Analyze