Homework #1

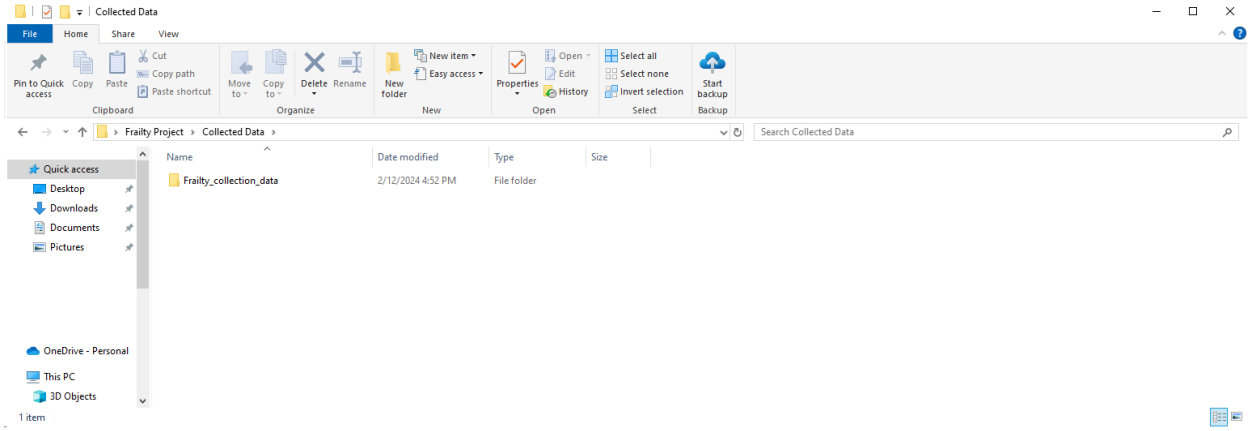
Nathan Bailey
Principles of Data Science

#1.

For designing a three stage reproducible workflow for the table that was given here is what I would do:

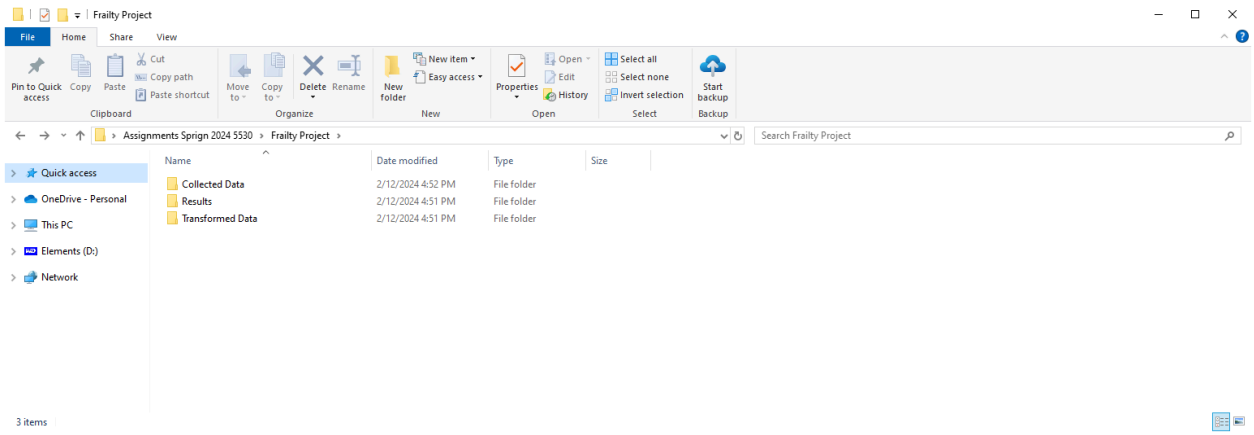
Stage 1

Have the data entered into a spreadsheet program(such as excel) and saved as a CSV file. Name the CSV properly and make sure it's in an easy to access location. The location should indicate that it is the collected raw data and hasn't been transformed in any way. (The name could be "Frailty_collection_data.csv")



Stage 2

Go through the raw data to find any missing data points and clean the data up so that it can be used to make data visualizations.
Never altering the original raw data but creating a new .csv file with the cleaned up data.
In this case the data in the table appears to already be cleaned up as nothing is missing and I would want to include whether both those with frailty symptoms and those without it.
These files would be stored in their respective folders separate from the raw data.



Stage 3

Using mainly R studio I would run multiple tests on the data , one such test would be a t-test frailty was associated with the other 4 data points, in particular “Grip Strength”.
I would construct multiple analysis models to compare the data storing the results in the results folder as well as creating a write-up of my analysis to make it easy to understand.