

Assignment 1

Ambati Sai Aravind







16336124

1)

a) Data collection:

The data is collected and stored in a suitable format.

Folder structure:

Name	Size	Kind
>  analysis_result	--	Folder
>  clean_data	--	Folder
▼  raw_data	--	Folder
 Frailty_raw_data.csv	241 bytes	CSV Document
 README.txt.rtf	2 KB	RTF Document
>  src	--	Folder

- Gathered the data and stored it in a file with .csv format in order to perform the analysis.
- README.txt file contains the metadata about the dataset.

Frailty dataset

This dataset have 10 observations and contains 5 columns namely

- Height column in represented in terms of inches
- Weight column is represented in terms of pounds
- Grip strength is represented in terms of kg
- Age column representing the age of the individual in years
- Frailty column represents the categorization of individual based on weakness

[README.txt file content]

Stage 2 – Data Processing:

Gathered data will be processed here in order to transform the data to be analysis ready.

- Preprocessing of data is performed using the scripts.

The image displays two screenshots of a Jupyter Notebook titled 'preprocessing_data.ipynb'. The first screenshot shows the initial data loading and inspection. The second screenshot shows the preprocessing steps, including column renaming, categorical value replacement, and dropping rows with missing values.

Files: sample_data, Frailty_raw_data.csv, clean_data.csv

Code:

```
[8] import pandas as pd
```

```
[9] #fetching raw data file
raw_data=pd.read_csv('/content/Frailty_raw_data.csv')
```

```
[10] #printing raw data
raw_data.head()
```

	Height (Inches)	Weight (Pounds)	Age	Grip strength	Frailty
0	65.8	112	30	30	N
1	71.5	136	19	31	N
2	69.4	153	45	29	N
3	68.2	142	22	28	Y
4	67.8	144	29	24	Y

```
[11] #Preprocessing the data to perform the analysis
```

```
[12] #updating the column names with the related units of measure
raw_data.columns=['Height(in)', 'Weight(lb)', 'Age', 'Grip_strength(kg)', 'Frailty']
raw_data.columns

# Changing the categorical values into numerical class values
raw_data['Frailty'].replace(['Y', 'N'],
                             [0, 1], inplace=True)
```

Files: sample_data, Frailty_raw_data.csv, clean_data.csv

```
[11] #Preprocessing the data to perform the analysis
```

```
[12] #updating the column names with the related units of measure
raw_data.columns=['Height(in)', 'Weight(lb)', 'Age', 'Grip_strength(kg)', 'Frailty']
raw_data.columns

# Changing the categorical values into numerical class values
raw_data['Frailty'].replace(['Y', 'N'],
                             [0, 1], inplace=True)
```

```
#dropping of the rows if na occurrences
clean_data=raw_data.dropna()
clean_data.head()
```

	Height(in)	Weight(lb)	Age	Grip_strength(kg)	Frailty
0	65.8	112	30	30	1
1	71.5	136	19	31	1
2	69.4	153	45	29	1
3	68.2	142	22	28	0
4	67.8	144	29	24	0

```
[14] #Saving the clean data into the csv file
clean_data.to_csv('clean_data.csv')
print('successfully saved the data')
```

successfully saved the data

- Initially the data is loaded into the script and then certain preprocessing steps have been performed like renaming with suitable column names and converting the categorical values into numerical values making them more efficient for analysis. Validated if there are any records with NA occurrences. And the final obtained clean data which is ready to perform analysis is stored as a separate file.

Folder Structure:

>	analysis_result	--	Folder
▼	clean_data	--	Folder
	clean_data.csv	243 bytes	CSV Document
▼	raw_data	--	Folder
	Frailty_raw_data.csv	241 bytes	CSV Document
	README.txt.rtf	2 KB	RTF Document
▼	src	--	Folder
	analysis.ipynb	31 KB	Document
	preprocessing_data.ipynb	15 KB	Document

Stage 3 – Data Analysis:

- Data analysis is performed by running the scripts to find the pattern of relation between the fields of the dataset.

analysis.ipynb ☆

File Edit View Insert Runtime Tools Help All changes saved

Files

- ..
- sample_data
- clean_data.csv

```
[5] import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

```
[6] from plotly.tools import FigureFactory as FF
```

```
[7] data=pd.read_csv('/content/clean_data.csv')
```

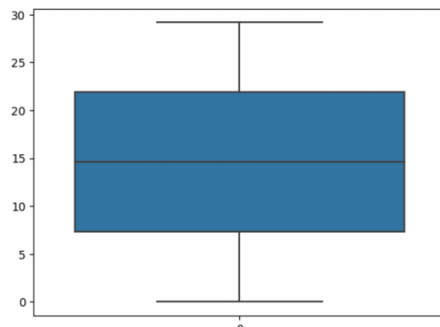
```
[8] data.head()
```

Unnamed: 0	Height(in)	Weight(lb)	Age	Grip_strength(kg)	Frailty
0	0	65.8	112	30	1
1	1	71.5	136	19	1
2	2	69.4	153	45	1
3	3	68.2	142	22	0
4	4	67.8	144	29	0

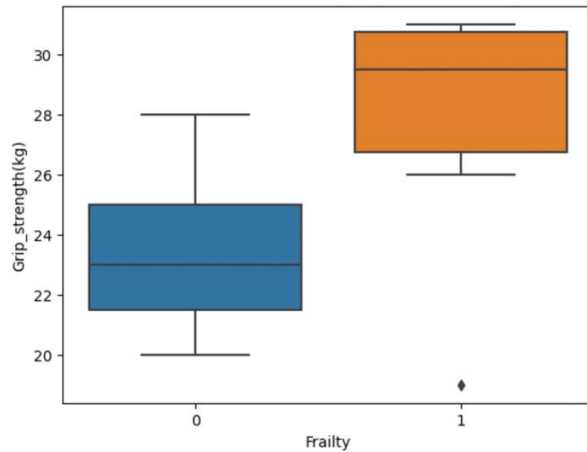
```
[9] #Performing the 2 sample t-test
t2 = stats.ttest_ind(data['Weight(lb)'],data['Frailty'])
print('analysis:',t2)
```

analysis: TtestResult(statistic=29.155389914737384, pvalue=1.3273210493059615e-16, df=18.0)

- In this stage analysis is performed on the processed clean data.
- Two sample t-test analysis is done between weight and frailty and results are plotted.



- box plot graph is generated with the X-axis having Frailty and the Y-axis having Grip strength data.



Analysis Result: The box plot representation clearly shows that greater the Grip Strength, the stronger will be the individual person.

Folder Structure:

Name	Size	Kind
analysis_result	--	Folder
result.png	12 KB	PNG image
ttest_result.png	7 KB	PNG image
clean_data	--	Folder
clean_data.csv	243 bytes	CSV Document
raw_data	--	Folder
Frailty_raw_data.csv	241 bytes	CSV Document
README.txt.rtf	2 KB	RTF Document
src	--	Folder
analysis.ipynb	31 KB	Document
preprocessing_data.ipynb	15 KB	Document