Summer 2025: DSA 5620 ICP 4

1. Creating a DataFrame from a given dictionary

```
data = {
    'ID': np.arange(1, 1000001),  # 1 million IDs
    'Value': np.random.rand(1000000),  # 1 million random values
    'Category': np.random.choice(['A', 'B', 'C', 'D'], size=1000000)  # Random categories
}
```

2. Output first 10 rows.
3. Access a column “Value”
4. Modify columns in the DataFrame with names (ID number, Random value, Choice) and show output for first five rows.
5. Run the below given code by removing bugs and errors.

```
import pandas as pd
pd.set_option('display.max_rows', None)
#pd.set_option('display.max_columns', None)
student_data = pd.DataFrame({
    'school_code': ['s001','s002','s003','s001','s002','s004'],
    'class': ['V', 'V', 'VI', VI, 'V', 'VI'],
    'name': ['Alberto Franco','Gino Mcneill','Ryan Parkes', 'Eesha Hinton', 'Gino Mcneill', 'David Parkes'],
    'date_Of_Birth ': ['15/05/2002','17/05/2002','16/02/1999','25/09/1998','11/05/2002','15/09/1997'],
    'age': [12, 12, 13, 13, 14, 12],
    'height': [173, 192, 186, 167, 151, 159],
    'weight': [35, 32, 33, 30, 31, 32],
    'address' ['street1', 'street2', 'street3', 'street1', 'street2', 'street4']},
    index=['S1', 'S2', 'S3', 'S4', 'S5', 'S6'])
print("Original DataFrame:")
print(student_data)
print('\nSplit the said data on school_code, class wise:')
result = student.groupby(['school_code', 'class'])
for name,group in result:
    print("\nGroup:")
    print(name)
    print(group)
```

6. Read the provided CSV file ‘data.csv’. https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing
7. Show the basic statistical description about the data.
8. Check if the data has null values.
    a. Replace the null values with the mean
9. Select at least two columns and aggregate the data using: min, max, count, mean.
10. Filter the dataframe to select the rows with calories values between 500 and 1000.
11. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.
12. Create a new “df_modified” dataframe that contains all the columns from df except for “Maxpulse”.
13. Delete the “Maxpulse” column from the main df dataframe
14. Convert the datatype of Calories column to int datatype.
15. Using pandas create a scatter plot for the two columns (Duration and Calories).

** Follow the rubric guidelines.

Submission Guidelines:

1. Once finished document your code and make sure all parts if the assignments are completed.
2. Push your code to your GitHub repo and update the ReadMe file, add your info.
3. Submit the assignment.
4. Present your work in class time to proof the execution and completesubmission.

After class submission:

1. Once finished document your code and make sure all parts if the assignments are completed.
2. Push your code to your GitHub repo and update the ReadMe file, add your info.
3. Submit the assignment before the deadline.
4. Record a short video (1~3) minute, proof of execution and complete assignment.
5. Add video link to ReadMe file.