Assignment 01:

Load the Iris dataset into a pandas DataFrame

Find the mean and median of the 'sepal_length' column.

1. Calculate the 75th percentile of the 'petal_width' column for each species in the Iris dataset.
2. Create a new column in the Iris DataFrame called 'sepal_area', which is the product of 'sepal_length' and 'sepal_width'.
3. Remove all rows in the Iris DataFrame where 'petal_length' is greater than twice the standard deviation of 'petal_length' for that species.
4. Normalize all numerical columns in the Iris DataFrame (except the 'species' column) using Min-Max scaling.
5. Find the three most common combinations of 'sepal_length', 'sepal_width', and 'petal_length' in the Iris dataset.
6. Group the Iris DataFrame by 'species' and find the row with the highest 'sepal_width' for each group.
7. Replace all negative values in the 'petal_width' column of the Iris DataFrame with the mean of the non-negative values in that column.
8. Calculate the correlation matrix for the 'sepal_length', 'sepal_width', 'petal_length', and 'petal_width' columns in the Iris dataset and find the feature with the highest absolute correlation with 'petal_width'.

Assignment 02:

Load the Titanic dataset (available in seaborn) into a pandas DataFrame

1. Find the average age of passengers for each class (1st, 2nd, and 3rd).
2. Create a new DataFrame that contains the count of male and female passengers in each age group (e.g., 0-10, 11-20, etc.).
3. Find the name and ticket number of the passenger(s) who paid the highest fare and survived the disaster.
4. Calculate the survival rate for passengers who were traveling alone (without any siblings, spouses, parents, or children) versus those who were traveling with family members.
5. For each passenger, calculate the age difference with the oldest sibling (if any) and the age difference with the youngest sibling (if any).
6. Find the most common deck letter (A, B, C, etc.) for each passenger class.
7. Group the Titanic DataFrame by 'Embarked' (port of embarkation) and find the percentage of passengers who survived in each group.
8. Calculate the correlation matrix for the 'Age', 'Fare', and 'Survived' columns in the Titanic dataset and find the feature with the highest absolute correlation with 'Survived'.
9. Create a new DataFrame that contains the 'Pclass', 'Sex', 'Age', and 'Fare' columns from the Titanic dataset and pivot it to have 'Pclass' as the index, 'Sex' as the columns, and 'Fare' as the values, with 'Age' as the weights.

Assignment 03:

Load the Planets dataset from seaborn

1. Scatter plot: Visualize the relationship between 'orbital_period' and 'mass' of the planets.
2. Bar plot: Display the count of planets discovered by each method.
3. Histogram: Visualize the distribution of 'distance' of planets from their respective stars.
4. Pair plot: Show pairwise relationships between 'orbital_period', 'mass', 'year', and 'distance'.
5. Violin plot: Compare the distribution of 'year' for different 'method' of planet discovery.
6. Swarm plot: Visualize the 'mass' of planets discovered by different 'method'
7. Heatmap: Create a heatmap to show the correlation matrix between numerical columns in the dataset.
8. Point plot: Compare the mean 'orbital_period' for each 'year' of planet discovery.
9. Bar plot with error bars: Show the mean 'mass' of planets for different 'method' of discovery with confidence intervals.

Assignment 04:

Load the Penguins dataset into a pandas DataFrame

1. Display the first 5 rows.
2. Calculate the average 'bill_length_mm' for each species of penguins.
3. the penguin with the highest 'body_mass_g' and display its species and other information.
4. Create a new DataFrame containing only the penguins with 'sex' as 'MALE' and 'island' as 'Torgersen'.
5. Calculate the correlation matrix for 'bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', and 'body_mass_g'.
6. For each species of penguins, find the mean, median, minimum, and maximum 'body_mass_g'.
7. Replace any missing values in the 'sex' column with the most frequent value in that column.
8. Create a new column in the DataFrame called 'bill_area', which is the product of 'bill_length_mm' and 'bill_depth_mm'.
9. Group the DataFrame by 'species' and calculate the average 'body_mass_g' and 'flipper_length_mm' for each species.
10. Calculate the total count of penguins for each 'island' and 'sex' combination.

Assignment 05:

Load the Diamonds dataset into a pandas DataFrame.

1. Calculate the average price for each cut of diamonds.
2. Find the diamond with the highest carat and display its details, including its cut, color, and clarity.
3. Create a new DataFrame containing only the diamonds with 'cut' as 'Ideal' and 'color' as 'D'.
4. Calculate the correlation matrix for 'carat', 'depth', 'table', and 'price' columns.
5. For each clarity grade of diamonds, find the mean, median, minimum, and maximum 'carat'.
6. Replace any missing values in the 'depth' column with the mean value of that column.
7. Create a new column in the DataFrame called 'volume', which is the product of 'x', 'y', and 'z' columns.
8. Group the DataFrame by 'cut' and 'color' and calculate the average price and carat for each group.
9. Calculate the total count of diamonds for each 'cut' and 'clarity' combination.

Assignment 06:

Load the dataset Assignment06.csv.

1. Here, the Area column is having the data in a mixed format of `alphabetical` and `int` characters. You need to remove the last two zeros out of the data present in the column. As you can see in the first record the 'Area' column is A100100 after conversion it will look like this: A1001
2. Filter the dataframe for records only for 2011 in a different dataframe.
3. Find the average value of 'geo_count' for the year 2022
4. Find the net average of 'ec_count' for year 2010,2011,2012
5. Plot KDE distribution of 'ec_count', for anzsic06 is F371.

Assignment 07:

Load the dataset Assignment07.csv

1. Find the most ordered item in the given dataframe.
2. Find the item, with the net highest number of quantity been sold.
3. Find the number of orders placed on '12-07-2011' and total quantity sold on this date.

4. Plot the number of orders placed with respect to each country.
5. Find the CustomerID of the person having the net highest number of quantity.