Question 1

Data Preparation

You have the following data files: Customers.csv, Items.csv, order_item.csv, orders.csv

Explore the Data

- Print for each table (in one operation per table) the column names, types, and number of null values (and more)
- Print for each table (in one operation per table) the column names, minimum values, maximum values, etc.
- Print the first 5 rows for each table

Data Cleaning

In the items table, there are products that appear multiple times (with the same name) and it was decided to keep products with the same name, but keep the product with the most non-null fields, and then delete the rest.

Note: (Follow the example shown in the class and implement it here).

Method of Operation:

- Add a new column counting how many non-null values exist in each row
- Sort by product name + the new column created, from largest to smallest
- Remove duplicates by product name (keep only the first occurrence)

Updating the order_item table:

- Replace the product IDs in the order_item table with the ID of the product that has the most non-null fields. For example:

48	Camera	Photography Equipment	2039.08	471	Panasonic	TRUE
28	Camera		695.71		Sony	FALSE

Here, we see that for the product Camera with ID 28, we should replace all mentions in the order_item table of product ID 28 with ID 48.

Delete from the customers table all customers who have fewer than 3 orders. Delete from the order_item table all sales that were made in the previous step.

Data Analysis

Use the following data files: Customers.csv, Items.csv, order_item.csv, orders.csv

1. What is the average price of an item?
2. Which customer has purchased the most products?
3. Add a total_price column to the order_item table, calculating the purchase cost (product price * quantity)
4. Display each purchase and its total price
5. Find what is the most expensive purchase? The cheapest? The average?
6. From the customers table, create a PIVOT table and display each person by index, and gender as columns. Show how many customers belong to each gender within that nationality. For example (partial table):

gender nationallity	Agender	Bigender	Female	Genderfluid	Genderqueer
Argentina	1	0	10	1	0
Brazil	0	0	6	0	0

Research and present in a graph the following details:

Recommendation: Remove rows with empty values

1. Distribution of customers by gender
2. Distribution of purchase quantities by nationality
3. Histogram of customer distribution by age
4. Graph showing the quantity of new customers who joined, by year
5. Graph showing the quantity of sales by months
6. Histogram showing the distribution of purchase quantities by age (each purchase should be calculated in a single row. No need to count product quantities, etc.) - Is there more/less purchasing in certain age groups?
7. *Challenge: Purchase costs by nationality. That is - we'll sum up the purchase costs that were made for all products for each customer, combine all purchases from the same nationality - then we'll see how much each nationality spends

Question 3

Questions:

1. df.dropna

a. Removes duplicates from columns

b. Removes rows that contain missing values (NaN)

c. Replaces missing values with a default value

d. Removes the index column from the table

2. df['city'].unique()

a. Returns the name of the city that appears most frequently

b. Returns a list of all cities in the table

c. Returns the number of different values in the 'city' column

d. Removes duplicate values from the 'city' column

3. df['gender'].replace({'M': 'Male', 'F': 'Female'})

a. Converts values to numbers according to a defined order

b. Replaces the values of the gender column according to the conversion map

c. Marks rows where gender is M or F

d. Connects between different values in the gender column

4. df.sample(5)

a. Selects the 5 smallest values in the table

b. Returns all values of the sample columns

c. Selects 5 random rows from the table

d. Creates a copy of the table with only 5 columns

5. df['score'].fillna(0)

a. Deletes all values in the score column

b. Marks rows where the value is zero

c. Fills the missing values in the score column with zeros

d. Converts all values to zeros

6. df.drop_duplicates

a. Removes rows containing missing values

b. Replaces duplicate values with the average value

c. Removes duplicate rows in the table

d. Removes all unique values in the table

7. df['price'].nlargest(3)

a. Returns the 3 smallest values in the price column

b. Returns the 3 missing values in the price column

c. Returns the 3 most common values in the price column

d. Returns the 3 largest values in the price column

8. df['city'].value_counts()

a. Displays the values in the column in random order

b. Returns the number of occurrences of each value in the column

c. Removes duplicates in the city column

d. Converts textual values to numbers

9. df['quantity'] > 100

a. Checks which rows in the quantity column are exactly equal to 100

b. Marks only the rows where the quantity is equal to zero

c. Creates a Boolean series of True/False according to the condition

d. Changes the values in the column to True

10. df.groupby('category').mean()

a. Divides the table by category and calculates the average for numeric columns

b. Removes duplicate values in the category column

c. Returns only the category that appears most frequently

d. Sorts the categories according to the number of their occurrences

Submission Instructions:

There are 2 repositories to create in GITHUB:

- One for today's solution
  - There will be NO COMMIT after today's date
- Second for the rest of the solutions
  - There will be NO COMMIT after Friday

For submitting answers to the American questions – create a text file with the question number + answer number It's also worth noting the text of the answer in words, to prevent confusion.

For example:

Question 18 - answer b "duplicate removal"

You should send the link to the repository by email:

pythonai250824+EXAMAILIB@gmail.com

Please note in the email:

- First name
- Last name
- Course number

Good luck!