Python Libraries - Moed A Exam

you are a data analyst in electronic shop and you got the following datasets:

1. **Items.csv file** → Contain data about all the items that have been sold in the store.

   The csv has the following fields:

   a. item_id → Unique id for each item
   b. item_name → The name of the item
   c. item_category → What category this item is associate to
   d. item_price → The price of the item in $
   e. stock_quantity → How much quantity from this item left in the store stock
   f. item_brand → What is the brand of the item
   g. item_availability → Show true if the item available and false if not

2. **customers.csv file** → Contain data about all customers that joined between the dates: 01.01.2020 and 01.01.2023

   The csv has the following fields:

   a. id → Unique id for each customer
   b. first_name
   c. last_name
   d. gender → Customer gender (Male, Female, Agender, etc…)
   e. age
   f. nationality → Customer nationality
   g. joining_date → The customer joining date
   h. phone_number
   i. email

3. **orders.csv file** → Contain data about all orders been made during 2023 year

   The csv has the following fields:

   a. order_id
   b. customer_id → Pointing to the customer id that created the order
   c. delivery_address
   d. delivery_days → How much days passed from the creation of the order until the delivery arrived to the customer
   e. order_date
   f. payment_method → How the customer paid for the order
   g. order_source → Where the customer created the order (in store or online)

4. **order_item.csv file** → Contain data about all items in each order appear in the orders.csv file

   The csv has the following fields:

   a. id
   b. order_id → Pointing to the associated order id
   c. item_id → Pointing to the associated item id
   d. quantity → How much quantity from that item was requested in the order

Data Preparation:

Make sure to use copy() of the original datasets in all data preparation actions.

1. Handle duplicate data in item dataset:
    a. In item dataset duplicate data is one of the following:
        i. Same item name and same item brand.
        ii. Same item name and no brand vs with brand (meaning item has the same name but 1 row with brand and 1 without.
    b. In case you find duplicate data, remove the item with less information. Make sure you change associations in other data sets to the duplicated item id so your data will still be accurate.
    For example → If you choose to remove item_id 3 because it's duplicated with item_id 5, make sure to change any association in other data sets from item_id 3 to item_id 5.

2. Handle duplicate data in customer dataset:
    a. In customer dataset duplicate data is on of the following:
        i. Customer with the same email address
    b. In case you find duplicate data, remove the customer with less information. Make sure you change associations in other data sets to the duplicated customer id so your data will still be accurate.

3. Handle missing data in all datasets:
    a. In case the missing data in a specific column is above 5% fill the missing data with a default valid value of your choice

b. In case the missing data is below 5% remove the row from your calculations and make sure to adjust your other datasets accordingly.
c. In case you found a row with missing data with a mandatory column (like id, name, etc…) remove this row from your dataset.

Data analysis:

Use the copy() datasets from your data preparation answer and answer the following questions

1. Explore the **customers.csv** dataset and answer the following questions, base your answers with data calculations and visualizations if needed:
    a. Count the number of customers by gender, show a dataframe with each gender type and how many customers we have from that type.
    Plot bar chart to visualize your result.
    b. Plot the customer age distribution with histogram chart.
    c. Find what is the year with most joining customers to the company.
    d. Examine whether there's a prevailing trend in customer joining over the timeframe from 2020 to 2023. Has there been a consistent increase or decrease in the number of customers joining over the years, or does the data suggest a more sporadic or unpredictable pattern?
    e. Identify if there's a particular month that consistently sees a higher number of new customers. Is there a specific time of

year when more customers tend to join, suggesting some kind of seasonality in customer sign-ups across all the years in this dataset?

2. Explore the **items.csv** dataset and answer the following questions, base your answers with data calculations and visualizations if needed:
    a. Count the number of items by category, show a dataframe with each category and how many items are associated with that category.
    Plot pie chart to visualize your result.
    b. Create a new dataframe containing the item with largest quantity in stock and the lowest quantity in stock, in case you put default values for missing data don't use those items in your calculations.
    c. Calculate the mean quantity in stock of all items, in case you put default values for missing data don't use those items in your calculations.

3. Explore the **orders.csv** dataset and answer the following questions, base your answers with data calculations and visualizations if needed:
    a. Calculate the amount of customers who have made just a single purchase, those who have made two, three, four purchases, and so forth.
    Plot bar chart to visualize your result.
    b. Find the top 5 customers (customer id, first name and last name) that ordered the most orders.

    c. Calculate for each payment method how many customers are paying with it.
       Plot pie chart to visualize your result.

4. Use **all datasets** and unser the following questions:
    a. Add a new column to the orders.csv dataset that show the total price of each order.
    b. Use the total_price column you calculated in the previous exercise and show what is the max, min and mean of the orders total price.
    c. Investigate if there is a correlation between the number of items in order and the delivery time (more items meaning more delivery time?)
       Use a scatter plot to support your answer.
    d. Find how many customers never created any order.
    e. Find what are the top 5 items that has been ordered the lowest (in your calculations consider the item quantity in the order)

Bonus → Decision making (10 Points):

1. The management started in August 2023 a marketing campaign in the USA country to market their website.

   Was that campaign useful or not?

   Base your answer on valid calculations and visualizations of your choice.

2. During August 2023 The management provided a discount prices for those who are using gift cards. Did it increase the sales or not? Base your answer on valid calculations and visualizations of your choice.

ECOM SCHOOL

המכללה למקצועות הדיגיטל וההייטק