SI 618 HW2

This homework is due September 15 right before class (3:59pm). Please turn in your Jupyter notebook (<uniqname>_si618_h2.ipynb and <uniqname>_si618_h2.html files) through Canvas.

The below files have been provided:

- invoices.json
- items.json
- purchases.json

They provided this data dictionary:

InvoiceNo: Invoice number. Nominal, a 6-digit integer uniquely assigned to each transaction.
StockCode: Product (item) code. Nominal, a 5-digit integer uniquely assigned to each distinct product.
Description: Product (item) name. Nominal.
Quantity: The quantities of each product (item) per transaction. Numeric.
InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
UnitPrice: Unit price. Numeric, Product price per unit in sterling.
CustomerID: Customer number. Nominal, a 5-digit integer uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.

A few notes from the company:

- If the InvoiceNo starts with the letter 'C', it indicates a cancellation. When conducting this analysis, we only want to analyze invoices that were shipped. (ie. not canceled)
- The datasets should be able to be merged, each row in the invoice table corresponds to multiple rows in the purchases table.
- To find out the description or unit cost of an item in the purchase table, the StockCode should be used to match up the product in the items table.
- They mentioned that they've been having a difficult time lately joining the items and purchases table, maybe there's something wrong with the columns?

Q1. [10 points] Describe the dataset.

1. Load the data files and display 5 random records from each file.
2. What fraction of invoices were shipped?
3. How many unique customers are there (regardless of shipped or not)?
4. What is the total number of unique items whose unit price is not more than 2?
5. How many missing/null values does each column in the data have?
6. Thinking ahead, how do you think you would join the different tables? Please share 2-3 sentences about your approach.

Q2. [15 points] Invoice Analysis

1. For each customer calculate how many shipped invoices they have placed. List the top 10 customers who have placed a shipped invoice in descending order.
2. Perform a similar calculation but instead of the number of invoices, calculate the average quantity of items per invoice for each customer. List the top 10 customers in descending order.
3. Based on 1 and 2, does it appear that the more invoices a customer has, the smaller the average size of an invoice? Explain your reasoning.

Hint: For 2.2, you may need to join two datasets together to answer the question.

Q3. [15 points] Item Analysis

1. What are the mean and median item-unit prices?
2. What % of items are below the average unit price ?
3. Generate a histogram of the unit prices. Select reasonable min/max values for the x-axis. State how the histogram supports the results from 1 and 2.

Q4. [25 points] Order Trends

1. What are the names of the top 10 items which had the most sales?
2. What are the stock codes of the top 10 most frequently ordered items by customers in descending order?
3. What are the top 5 invoices that generated the most revenue? (Revenue is calculated by marking up the unit price by 30%.)

Hint: When calculating the revenue, we suggest adding a new column on the dataframe.

Q5. [35 points] Customer Analysis

1. Discretize customers using dummy variables into five different segments (Q1,Q2,Q3,Q4,Q5) based on the 20^{th}, 40^{th}, 60^{th}, 80^{th} and 100^{th} percentiles of the total revenue they have generated for the company.
2. How much revenue is generated in total by each segment?
3. Using the pivot table function create a table that displays the number of customers for each country in each segment.

Hint: When calculating the segment, we suggest constructing a new dataframe as an intermediary step with the columns: CustomerID, Revenue, Segment.