**The Report Questions**

Part 1: Cleaning the Data

It came to our attention that some of the data was either incorrectly entered or missing entirely! This is going to throw off our calculations if it is left unchecked. We were grateful to discover none of it happened in our accounting columns of the data for that would be very detrimental but we aren't sure where the missing/incorrect data is elsewhere. Here's what we do know to help your investigation into finding the missing/incorrect data. Of the columns that we have, we know the missing/incorrect data comes from these columns:

* Country   (either missing OR will be a number as a string)
* Item Type   (either missing OR won't be a valid Item Type from the other ones listed)
* Order Priority   (either missing OR won't be a valid priority code)
* Order ID   (either missing OR won't be a number)  
    
  If you find incorrect/missing data and its text type for that column, change it to NULL. If you find incorrect/missing data and its a number type for that column, change it to 0. (or 0.0 if its a float). Test for missing values FIRST then if you find the ones that are missing, you don't have to test those for incorrect values. **You need to change the values, then rewrite to a new CSV file called Marvel\_Mart\_Sales\_clean.csv so it can be used later with the correct values.**

*(Note from Instructor: While it is possible and acceptable to produce a non-pandas using solution, I would suggest that you use pandas Dataframes for this. It makes a difficult job  much easier. The guide to how to do that is here and some other resources:*

* [*https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/ (Links to an external site.)*](https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/)
* [*https://www.geeksforgeeks.org/python-read-csv-using-pandas-read\_csv/ (Links to an external site.)*](https://www.geeksforgeeks.org/python-read-csv-using-pandas-read_csv/)
* [*https://www.geeksforgeeks.org/saving-a-pandas-dataframe-as-a-csv/ (Links to an external site.)*](https://www.geeksforgeeks.org/saving-a-pandas-dataframe-as-a-csv/)

***You will need to google things to get this done. Part of being a programmer is knowing how to find solutions on the internet and adapting them.****)*

Part 2: General Statistics

First, we would like you to get us some general statistics from the data. I suggest you create a dictionary of lists with the keys being the heads of the CSV file and the list attached to it being all of the values for that heading. Duplicates should be included here.

1. Produce the following and write it to a text file called **Marvel\_Mart\_Rankings.txt**. Be sure to use append so that you can append data rather than writing over top of the previous data. Be sure to include a newline between each append to the file. When writing to the file, please output in a text form such as:  
     
   Countries Most Sales:  
   Country 1: (sales number)  
   Country 2: (sales number)  
   ...  
   (Answer question) "The country we should build our shipping center is \_\_\_\_\_\_ because \_\_\_\_..."   
     
   (A) We want to know which countries we sell the most to so we can pick a new location to build a shipping center. Rank the Top 10 countries we sell to the most to least along with the number of sales we've had with that country. We have shipping centers in Trinidad and Tobago, Guinea, and Maldives right now. Which country should we build a shipping center in based on most sales and lack of shipping center? Please justify your reasoning.  
     
   (B) Provide a count for how many online and offline orders we take. Which do we take more of? You do not need to justify your answer.  
     
   (C) Rank the top 3 years we did the most sales (brought in most profit) in to the least sales. (Just the years, not the whole dates).  Use the Order Date, not the Ship Date. Please list the years and the amount sold. Answer the question "Which year did we sell the most in?"  
     
   *(Note from instructor: doing large number sums with floats in Python usually produces scientific notation but we don't want that. You can turn that off by putting the following line under the import statements at the start of the script:* pd.set\_option('display.float\_format', lambda x: '%.3f' % x)
2. Now you will save these calculations below to a text file called **Marvel\_Mart\_Calc.txt**. When writing to the file, please output format such as:  
     
   Sums:  
   Units Sold: (Number)  
   Unit Cost: (Number)  
   Total Revenue: (Number)  
   Total Cost: (Number)  
   Total Profit: (Number)  
   (Newline)  
   Averages:  
   Units Sold: (Number)  
   ...  
   (Newline)  
   Maximums:  
   Units Sold:  
   ...  
     
   (A) Produce the data above for the sum of each one.  
   (B) Produce the data above for the average of each one. (Average Units Sold, Average Cost, etc)  
   (C) Produce the data above for the maximum of each one. (Max Units Sold, Max Cost, etc.)

Part 3: Cross-Reference Statistics

For this part you will be cross-referencing the data in the CSV file and the getting an output and writing it to a new CSV file.

1. We need you to get a list of the Regions and then the countries we sell to in that region. Please be sure no duplicates Regions or countries exist. Please return this as a dictionary of lists with the keys of the dictionary being the name of each Region and the list attached to that being all the countries we sell to for that region. You may also return it as a Series of Lists (although I found that to be harder). Finally, if you want to use an alternate method with pandas dataframes, that will be accepted as well. Be sure your output is easy to read and your code makes sense.  
   Write this out to a CSV file called Countries\_By\_Region.csv.  
     
   **(Be careful here as there is a header row when you convert the csv to a dictionary. If you end up getting the header row in your final result, just remove it. You are free to do it however you want as long as in the end its a dictionary of lists (or Series of Lists), the keys being the Regions and then the list for that key being the countries that is sold to, with no duplicates. And then print it to the csv file.)**