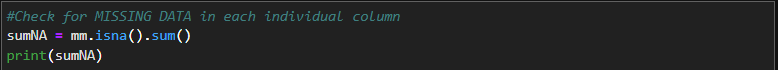
**Documentation Surrounding Project 2**

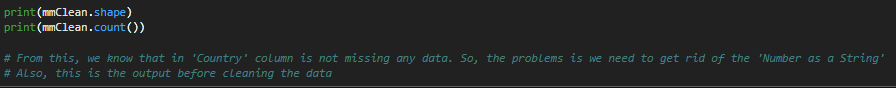
**Part 1: Cleaning the data:**

1. **Step 1: Checking for missing Data/ Finding Missing Data:**

* For step 1, I need to figure out how many columns that have missing data.
* I run the first code for checking how many rows should I have for each column
  + Ảnh có chứa văn bản, màn hình, bình phong

    Mô tả được tạo tự động
* Then I run the second code for determining the data is missing in specific columns
  + 
* Then I need to copy the data to a new DataFrame for cleaning data
  + 
* Then I display all the column’s type in the CSV file
  + Ảnh có chứa văn bản

    Mô tả được tạo tự động
* All of this to prepare for me to dig deeper to clean the data

1. **Step 2: Focus on cleaning the Data for columns: Country, Item Type, Order Priority, and Order ID**
   1. **Testing the erroneous data in “Country” column and fix it:**
   * Step 1: Determine the problem in the column, whether it is missing data or have dirty data by this code:
     + 
   * Step 2: Realize that the problem for this column is having dirty data. As the code above to see the type of the column, I know that the Data type of this column should be a string, so I “try” to convert all the data from this data to float, if they can convert, we can get it out and replace them to 'NULL', because only “Number as string” can converted to float. So, that’s how I have the code below:
     + Also, the code’s output below is the output for the first time I run the code. The reason I put this code here is because I want to show the differences between my outputs when I run the code multiple times.

Ảnh có chứa văn bản

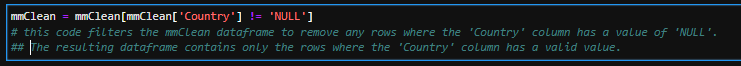
Mô tả được tạo tự động

* + - This is the same code as above. However, this is the output that showed when I ran the code for the 2nd time. The output only showed 2 lines, which is different from the output that I ran for the first time. The reason I put this code is because I want my project to show step-by-step on how the data change over the different times I run the code.

Ảnh có chứa văn bản

Mô tả được tạo tự động

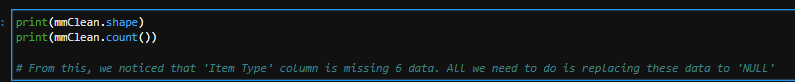
* + - Then I need to get all the data that is not “NULL”:



* + - Finally, I run this code for the 3rd time, and here is the output. As you can see, the “Number of erroneous Country: 0”. This is the final output that we want to see because we want to clear all the dirty data for this column. And I also return the total number of all the columns after I clean the data

Ảnh có chứa văn bản

Mô tả được tạo tự động

* 1. **Testing the erroneous data in "Item Type" Column and Fix it:**
  + Step 1: Determine the problem in the column, whether it is missing data or have dirty data by this code. From this, I know that this column is missing 6 data.
    - 
  + Step 2: Replace all the missing data to “NULL”.
    - This is the code for testing whether the data is missing or not. If missing, the blank spaces will be replaced by “NULL”.
    - Also, this is the very first code that shows the result below. The reason I put this code here is because I want to show the differences between my outputs when I run the code multiple times.

Ảnh có chứa văn bản

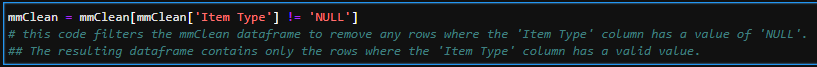
Mô tả được tạo tự động

* + This is the same code as above. However, this is the output that showed when I ran the code for the 2nd time. The output only showed 2 lines, which is different from the output that I ran for the first time. The reason I put this code is because I want my project to show step-by-step on how the data change over the different times I run the code.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + Then I need to get all the data that is not “NULL”:



* + Finally, I run this code for the 3rd time, and here is the output. As you can see, the “Number of invalid Items that changed to NULL: 0”. This is the final output that we want to see because we want to clear all the missing data rows for this column. And I also return the total number of all the columns after I clean the data

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + Also, to make sure, I run this code to return the first 15 rows that contain the row 11 that was a blank space, show that after clean, the row 11 is gone

Ảnh có chứa văn bản

Mô tả được tạo tự động

* 1. **Testing the erroneous data in "Order Priority" column and Fix it:**
  + Step 1: Determine the problem in the column, whether it is missing data or have dirty data by this code below. Then I realized that the column is missing 15 data

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + Step 2: Replace all the missing data to “NULL”.
    - This is the code for testing whether the data is missing or not. If missing, the blank spaces will be replaced by “NULL”.
    - Also, this is the very first code that shows the result below. The reason I put this code here is because I want to show the differences between my outputs when I run the code multiple times.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + - This is the same code as above. However, this is the output that showed when I ran the code for the 2nd time. The output is different compared with the output that I ran for the first time. The reason I put this code is because I want my project to show step-by-step on how the data change over the different times I run the code.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + - Then I need to get all the data that is not “NULL”:

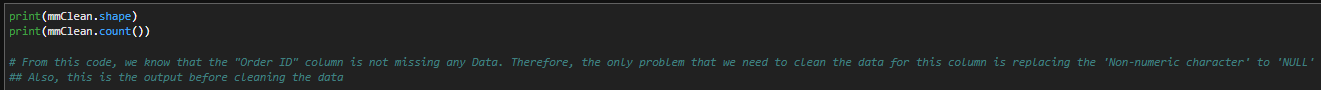


* + - Finally, I run this code for the 3rd time, and here is the output. As you can see, the “Number of invalid Items that changed to NULL: 0”. This is the final output that we want to see because we want to clear all the missing data rows for this column. And I also return the total number of all the columns after I clean the data

Ảnh có chứa văn bản

Mô tả được tạo tự động

* 1. **Testing the erroneous data in " Order ID" column and Fix it:**
  + Step 1: Determine the problem in the column, whether it is missing data or have dirty data by this code below. Then I realized that the column is not missing any Data. Therefore, the only problem that we need to clean the data for this column is replacing the 'Non-numeric character' to 'NULL'



* + Step 2: Replace all the missing data to “0”.
    - This is the code for testing whether the data is missing or not. If missing, the blank spaces will be replaced by “0”.
    - Also, this is the very first code that shows the result below. The reason I put this code here is because I want to show the differences between my outputs when I run the code multiple times.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + - In this code, we can see that the 'Non-numeric characters' have been found and replaced to 0, but the reason that the output showed "Number of erroneous values found and replaced with 0: 0" is because this is the second time that the code ran, and '0' is the number (int). This output belongs to the SECOND time I run the code above.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* + - Finally, I need to get all the data that is not “0”. And then I return the total number of all the columns after I clean all the dirty data.

1. **Step 3: Testing the data to see if it is clean/ready or not:**

* In this step, I need to test if there is the erroneous data that still exist in the data or not

1. **Step 4: Rewrite the Clean Data to a new CSV file.**

**Part 2: Exploratory Data Analysis with Reports & Visualizations**

* 1. **Get the top 10 countries we sell the most.**
  + To write this code, I need to use “groupby” and “count()” to gather the data that I want: group all the duplicate countries and count the number of sale transactions as a count of the order ID. Then I want to set the order of the Country in descending order. Finally, I need to return the top 10 of the countries that have the most sales.
  + Then, I need to use Matplotlib to create a chart to show these top 10 values by Country.
  1. **Answer the question by writing the results to a text file called MM\_Rankings.txt**
  + First, I need to create the string “Countries Most Sale Transaction”
  + Then I need to process the data to create a list of output lines
  + Then I need to concatenate the lists with the additional line
  + Finally, I write the output lines to a text file called MM\_Rankings.txt
  1. **Determine the count for how many online and offline orders we take**
  + First, I need to load the data into a pandas DataFrame
  + Group the data by Sales Channel and count the number of sales (Order ID) for each Sales Channel (Online and Offline). A sales transaction is represented by a row. Basically, in this code, we groupby the online and offline order type in 'Sale Channel' column
  + Then, each order type (online and offline) has multiple different order ID. So, after we groupby the online and offline order type, then we count sales transactions belonged to that order type
  + Finally, I print the DataFrames that I just made.
  1. **Determine the count of the different Order Priority types:**
  + Load the data into a pandas DataFrame
  + Group the data by Order Priority type and count the number of sales/transactions (Order ID) for each Order Priority (C,H,L,M). A sales transaction is represented by a row. Basically, in this code, we groupby the Order Priority type in 'Order Priority' column
  + Then, each Order Priority type (C,H,L,M) has multiple different order ID. So, after we groupby the Order Priority type, then we count sales transactions belonged to that type
  + Finally, I print the DataFrames that I just made.
  1. **Create a pie chart for 2A and 2B**
     + First, I return the DataFrame that belonged to the 2A,B part that will need to be used for creating the pie chart
     + Create a pie chart from 'oline\_offline\_order' DataFrame by using a pie chart
  2. **Add the results of the sales channel types and the order priorities to the file MM\_Rankings.txt.**
  + First, I print the DataFrame to see what I need to add in the text
  + Then, I create a string “Sales Channel/ Order Priority”
  + Then, I process the data to create a list of output lines
  + Then I concatenate the lists with the additional line
  + Then I write the output line to a text file MM\_Ranking.txt as a function ‘a’
  1. **Create a Boxplot using Seaborn showing the Total Profits DISTRIBUTION by Item Type:**
  + First, I need to create a boxplot using Seaborn by using sns.boxplot
  + Then, return the plot by plt.show()
  1. **Use Python to determine the sum of Total Profit by Item Type:**
  + First, load the data into a pandas DataFrame
  + Then I group the data by Item Type and sum the Total Profit for each Item.
  + Then I print the data as a new DataFrame
  1. **Now create a chart type of your choice (Seaborn or Matplotlib) showing the sums of the different Item Types: I choose to use Bar chart**
  + I need to determine the DataFrame that I want to use for making the bar chart
  + Then using python to create and show the bar chart using the DataFrame that we already determined
  1. **Rank the top 3 item types we did the most sales (brought in most profit) in to the least sales. (Use 'Total Profit' to determine this). Please list the item types and the amount of profit made from sales.**
  + First, load the data into a pandas DataFrame
  + Then, group the data by Item Type and sum the Total Profit for each Item.
  + Finally, I want to set them in descending order based on the Total Profit and then print the result.
  1. **Add the results of the top 3 item types to the file MM\_Rankings.txt**
  + First, I create the string “Highest Selling Items”
  + Then, I need to process the data to create a list of output lines
  + Then, I concatenate the lists with the additional lines
  + Then I write the output line to a text file MM\_Ranking.txt as a function ‘a’
  1. **Discuss what is being shown in the boxplots and do some business analytics around what sort of use this sort of chart might help in making decisions. Are there any unexpected results? Discuss them**
  + First, I need to return the boxplot again
  + Second, based on the data from boxplot, I have discussion for this boxplot

1. **Please determine the sum, average and maximum values for the 'Units Sold', 'Unit Cost', 'Total Revenue', 'Total Cost' and 'Total Profit'. Please put this in a report**
   1. **4A,B,C Produce the data above for the sum, average, maximum values of the requested columns.**
   * First, I load the CSV file into a DataFrame
   * Then I determined the column I want to sum, average, and maximum value
   * Then process the code for print the output
   1. **4D: Create two line plots using Seaborn or Matplotlib, one for the sums and one for both the averages and the maximums. DO NOT INCLUDE UNITS SOLD OR UNITS COST.**
   * First, I create a list of data that I need for making the line plots
   * Second, based on the list, I create a DataFrame from that list.
   * Finally, using the DataFrame that I just created, I create the line plot for the requirements.
   1. **4E: Now you will save these calculations below to a text file called MM\_Calc.txt.**
   * First, I need to determine the output that I want to put in the new text file
   * Then, I save the output (calculation) to a new text file called MM\_Calc.txt

**Part 3: Cross-Reference Statistics:**

* First, I need to load the input CSV file
* Then, I need to determine the columns that I need to use for this part
* Then I need to drop duplicate rows in the Region and Country columns by using code
* Get a list of unique regions
* Loop through the unique regions and store the countries for each region in a dictionary
* Convert the dictionary to a DataFrame
* Transpose the DataFrame so that the regions become the columns
* Write the DataFrame to a new CSV file