Reyn Okinaga
March 15, 2023

Python Project – Marvel Mart Project Write-up

To start off with the "Data Cleaning" section of the project, first I needed to figure out what missing data and incorrect data we had in the txt file. I first started with looking at a more depth look of the data with looking at the missing value as well as the data types of the columns. I found that there was missing "Item Types" and "Order Priority" columns. Then, I went and grouped the "Country", "Item Type", "Order Priority", and "Order ID" columns individually to see if there was any incorrect data. I found that there were incorrect data types in the "Country" and "Order ID" columns. After looking at the missing and incorrect data, I needed to fix it. I created a copy of the original txt file to edit and clean up the data. I filled the missing values with "NULL". For the "Country" column, we converted the incorrect data to float values, as there were some strings with numbers in the data, and replaced it with "NULL". For the "Order ID" column, I converted the incorrect data to floats, as there were strings in the data, to find the strings in the data and replace them with 0.0. I checked again after replacing the data and shows that we have successfully corrected the data. Then I removed all the rows that have been altered.

Onto the next part of "Exploratory Data Analysis with Reports and Visualizations" section, we were asked to get the top ten countries with most sales. By grouping the countries together, counting the rows (sales), and using "nlargest" code, we were able to get the top ten countries with most sales. We then made a graph and wrote our results in a txt file called MM_Rankings. For the data of online/offline sales and order types, I grouped the columns individually and got the count for it. I made pie charts to represent that info and wrote the results in the txt file. Then, I made a boxplot of the item types by total profit and then gathered that information with the sum of total profits to create a bar chart. By using the nlargest code again, I was able to return the top 3 item types that gained the most sum of total profit and wrote the information into the txt file. Lastly, I got the sum, mean, and max values of the unit cost, units sold, total revenue, total profit, and total cost of the data. However, I created one dictionary containing all values to help append the information into the txt file and another one that didn't contain units sold and unit cost to use for the line graphs.

For the last part of the "Cross-Reference Statistics", I first needed to make a dictionary of lists with the countries and regions. First, I created a list of set from the regions columns with only unique values and then created an open dictionary to fill out later. Inside the open dictionary, I used a for loop for the list and created another list where the countries match with the regions. After, the dictionary has the keys of the regions and then the values have the list of countries. Lastly, I created a series first with pandas dataframe with the dictionary and then input the series data into the csv file. The index is removed to remove the numbered rows and the header is shown as the header as it is the regions.