Avery Hultgren

3/10/2022

BUAN 4210

Eric Lloyd

**Marvel Mart Python Project**

* **Intro**

For this project I was tasked with cleaning a large csv file, creating visualizations, different forms of analysis, and cross reference statistics with the data. The project itself was split into three separate sections based around these tasks, data cleaning, exploratory data analysis, and cross-reference statistics.

* **Data Cleaning**

For this section I started out by using .info and .isna().sum() to learn what areas had wrong or missing data, from there I went to work on each specific column that had a problem, which were the Country, Order Type, Order ID, and Item Type columns, each with either integers/strings where they should not be, things that inputted in the wrong areas (putting ‘clothes’ in a date column for example), or null cells. The main way I look for errors was using a try and except to try to convert each cell into a float, this was useful for finding integers as well as finding strings, I would then in the except area make it so if anything for example was not able to be converted into a float in the Order ID column it would be turned into a 0 if it wasn’t an integer, and a count at the beginning of the code would go up 1 and the wrong data would be printed.

Then after doing a verson of this for each column that had any errors, I made it so that each row for these columns that were flagged by the past codes would be deleted on a new clean copy of the data as due to our large csv file and the small amount of erroneous data it was of little cost.

* **Exploratory Data Analysis**

Now that the data in the csv file was cleaned, I used the clean copy to figure out different things about the data such as the top countries by sale number using both tables and pie charts to show the data found and appending to a text file with the results, creating other visualizations for each other separate columns and some of the connections between them, finding the distribution of the total profit of different types of items sold, showing the total profit by item type, as well as showing the average, max, and sum of the total cost, profit, revenue, unit cost and units sold, and then making graphs showing this data.

* **Cross-Reference Statistics**

The last part of the project asked me to get a list of all the regions they sell to as well as the countries in those regions with no duplicates in either, for this part I first grouped the region column by the country column and then used unique to make it so there would be no duplicates, then I turned it into a dictionary. After turning that specific group by into a dictionary with zero duplicates I passed it into a data frame and used transpose to make the columns and rows change so that the regions column would be the headers with the countries in each column of those regions, then I created a csv file of it using to\_csv.