Cleaning Document

Data Overview

Retail Superstore Dataset

There are four CSV files that will need to be cleaned before beginning analysis. The files are a Customers, Orders, Location, and Products file. I will break down the cleaning processes for each file separately to lay out the steps and issues for each file.

Why Clean Data?

Data cleaning is an important part of the data analysis process as the confidence in our results is dependent on the quality of our data. By ensuring our data is correct, complete, and relevant to the business problem/project at hand will lead to a correct data-driven decision.

Data Cleaning

Customers File

This Customers file contains two columns. A customer identification column and a customer name file containing first and last name. There are a total of 793 customer records.

- The naming conventions are going to be set now for every other dataset to set a consistent format and standard. I will convert all column names into a Title Case naming convention.
    - The formula used to make this Title case: =PROPER(A2)

- Using COUNTBLANK() function I found no record of blanks within the dataset. Alongside this I used the Remove Duplicate feature to check and remove duplicates. No duplicates were found.

- Exploring the Customer ID column, it follows a format of initials, their first and last name alongside a code set as their ID number. ID length follows a consistent field length of 8 characters.

- The Customer Name column contains various Text/Spelling errors such as inconsistent use of uppercase, proper case, and lowercase formats. There also seems to be various spacing issues in the names and incorrect spelling of the names.
    - To remove issues with spacing and inconsistent formatting, my options are TRIM or SUBSTITUTE. The function used: =PROPER(SUBSTITUTE(B2," ","")). I used this over trim because there were issues with spacing in the middle of a name

and using TRIM would not completely resolve the issue. The proper function is used to standardize the text.
    - To correct spelling mistakes, I used Spell-Check to fix any misspellings of names.

Location File

The Location file contains information of the customers within the United States such as Postal Code, CIty, and Country/Region information. There are a total of four columns and 631 records in this file.

- There are a total of 111 blanks. Postal Code has one blank, City has 51 blanks, and State has 60 blanks.

- The City column contains Text/Spelling errors such as various case formats.
    - Function used: Proper(trim(C2))

- Used Spell-Check to correct any misspelled City names

- To handle the blanks in the city column, I will use a XLOOKUP to accomplish the task by using the postal code to help determine the correct city and remedy all blanks/missing values.
    - Firstly, we need a list of cities where the blank values occur by filtering the column to show only blanks and copying it to another sheet.
    - After finding the associated city with each postal code, I created a list to which I can use a nested IF and XLOOKUP to fill in the missing values: IF(ISBLANK(C2),XLOOKUP(A2,Sheet2!$A$1:$A$51,Sheet2!$B$1:$B$51,,0),C2)
    - Formula Explanation: Starting from the outermost function, I will use a IF statement to test whether a cell, that being a cell within the City column, is blank or not. If it is not blank, it will return the original content of the cell. If the cell is blank, then it will run a XLOOKUP in which it will look at the cell within the Postal Code column. It will check for the same Postal Code in my list and return the associated city name.

- To handle the blanks in the State column, I will use a XLOOKUP to accomplish the task by using the city name to help determine the state name and thus remedy the blank/missing values.
    - Firstly, we need the list of states where the blank values occur by filtering the column to show only blanks and copying it to another sheet.
    - After finding the associated state with each city name, I created a list to which I can use a nested IF and XLOOKUP to fill in the missing values: =TRIM(IF(ISBLANK(D2),XLOOKUP(B2,Sheet3!$C$1:$C$59,Sheet3!$D$1:$D$59,,0),D2))
    - Formula Explanation: Starting from the outermost function, I will use a IF statement to test whether a cell, that being a cell within the State column, is blank

or not. If it is not blank, it will return the original content of the cell. If the cell is blank, then it will run a XLOOKUP in which it will look at the cell within the city column. It will check the city name and return the associated state with an exact match required.

Orders File

This event table contains 10 columns with transaction information such as the product price information, sales information, customer information, and shipping/order date information. There are a total of 10331 records in the dataset that spans from 2020 to 2023. Minimal cleaning was required for this event table and only needed to set the naming convention for the column names.

- Date format for the Order Date and Shipping Date column followed Day/Month/Year format. Excel however would not recognize this format and thus to standardize this, I have set all dates to follow a Month/Day/Year format. To accomplish I used a two step process:
    - Use Text to Columns feature and set the delimiter to “/” to extract each part of the date and put it into its own columns
    - Used this function to set the date format: =CONCAT(P2,"/",O2,"/",Q2)
    - Formula Explanation: Concat will join text strings together so I set the cells to follow the date format I needed.

Products File

This table contains four columns that contain product information such as ID, product name, product category and product sub category. There are a total of 1894 records or products that the retail company sells as a whole.

- The Product Name column contains various product names in various case formats. To standardize the names, I will use a Proper case format.
    - Function used: =PROPER(A2)

- The Product Name column also contained incorrect text/spelling errors for the product names. Removed any odd spacing in names and corrected any misspelling of the product names. Used Find and Replace to speed up corrections.

- The Category column contains Text/Spelling errors as well as missing values which I will tackle one at a time. Starting with the Text/Spelling errors, I discovered there are three categories, Technology, Furniture, and Office Supplies spelled in various formats. To standardize these names, I will set them to the appropriate category names.
    - Function used: =IFS(LEFT(D3,1) = "F", "Furniture",LEFT(D3,1) = "T","Technology",LEFT(D3,1) = "O","Office Supplies")
    - Formula Explanation: Using an IFS function, this function is similar to a IF however it will allow for multiple logical tests and values if true. By using a LEFT

to check if the first letter of the category meets the criteria, it will then set that cell to the corresponding category name,

- Sub Category contains 147 blank values. To resolve this issue, I used the Product Code to extract the sub category abbreviation and used Find and Replace to give the full Sub Category name associated with the product.

- Category contained 150 blank values. To resolve this issue, I used the Product Code to extract the sub category abbreviation and used Find and Replace to give the full Sub Category name associated with the product.

Cleaning Check

With the initial cleaning complete, I will now go over my cleaning to ensure that my data is correct and accurate and as well to double check the work before analysis.

Data Modeling

In this section, I will go over the Data Modelling conducted in Tableau. The data model I chose for the data is to follow a Star Schema. Why? The data is relatively small and with the added fact I only have three files makes the Star Schema the perfect choice. The event table is the Orders table with the Customers and Products table being the supplemental tables since they are tables in which to add more information to the events table.

I will use the relationship method to connect my data together because it will provide me the benefits of no data loss and flexible relationships. The cardinality will be set to Many to One. The reason for this configuration is because it will help to optimize the performance of Tableau. Whilst the data is not massive, the larger the datasets become, the more important optimizing our performance for Tableau will be. Another reason is that because the Customers and Products files contain unique records, we can be assured that the Many to One relationship is appropriate!