**Sales Dataset Cleaning Document**

# **Data Overview**

The dataset is Sales data. It contains 25 columns and 2824 rows. There is personal information in regards to the customer such as address, postal code, full name, and phone number if applicable. There is location data such as Country, State, and City. There is sales information such as quantity ordered, price, MSRP, Status of order. There is also information such as product category and product id.

# **Data Cleaning Log**

* The total number of blanks within the sheet is 4571. Looking deeper, there is 125 blanks within the Product Line Column, 363 blanks in the Customer Name column, 2521 in the Address 2 Column, 1486 in the State Column, and 76 blanks in the Postal Code column.
* There are also major issues with Inconsistencies within the dataset such as Labeling, Field Lengths, Formatting, Spelling Error, Text Errors, and Missing Values.
* Changed the headers to follow a more appropriate format. I used a combination of the PROPER, TRIM(), and SUBSTITUTE() function to make the headers in the proper case, addressing any trailing spaces, and replacing the underscores in the header names. I also applied a bold font for better readability.
* Columns in relation to currency such as Price Each, Sales, and MSRP were not given the correct number format. Changed from General to Currency format for a more appropriate number format. The Order Date column contained a “MM/DD/YYYY” format and also a timestamp. Timestamp was set at “00:00” for all entries in the Order Date column. I changed this number format to Short Date which removed the unnecessary timestamp.
* The following column names have been updated: Qtr ID, Month ID, Year ID. I removed the “ID” in the name to standardize the names a bit more using Find and Replace (CTRL + H). Set Find to “ID” and Replace it with “”. Can also use Substitute(text, “ID”,””) to accomplish this. I also removed columns with personal information. The columns were: Phone, Address line 1, Address Line 2, Postal Code, Customer Last Name, Customer First Name. Only Customer Name was kept as this contains the company name.
* Within the Status column, there were several misspellings of each status type. Using Find and Replace to amend the text errors. Used a combination of PROPER() and SUBSTITUTE() to amend odd spaces between the words and address case format.
* Within the Product Line column there was many Spelling/Text Errors. I used PROPER(TRIM()) to amend some of the text errors and case format. I used SpellCheck to autocorrect the misspelled words. Amended any Missing values by creating a reference table of all missing rows and identified the correct product line based on the Product Code and similar entries that had information. Using the Products Table, created a XLOOKUP that would correct the mislabeling of Products.
* Customer Name contained issues with Spelling/Text errors. Used TRIM() and PROPER() to address case format and unwanted spaces. Used SpellCheck to address most misspellings within the column. Remaining Misspellings not addressed by Spell Check was amended by Find and Replace (CTRL + H). Column also contained many missing values. To address these, used a XLOOKUP
  + =IF(ISBLANK(Q2),XLOOKUP(A2,Sheet1!$A$2:$A$292,Sheet1!$B$2:$B$292,,0),Q2)
  + This function took a bit more care to build because there was an issue when using the IF(ISBLANK()) statement. Cells that appeared to be blank actually were not “Blank”. Meaning that the function would not apply the xlookup because it contained some space that made it “not blank”. I corrected this by filtering for blanks and removing the spaces to make the cell truly blank. The function worked as it should.
* The “State” column contained entries with varying formats. States were referred to as a two character abbreviation or their full name. Most of the states were already in the former format so I decided to use this as the standard for all states. There were also missing/blank values within the column.
* To address the states with the full name I decided to create a reference table with the names in full and find their abbreviated name in two character format. Using Find and Replace to accomplish the standardization.
* To address the missing/blank values I filtered the sheet to show all entries where “State” was blank. I gathered the cities and created another reference table of all cities and find their associated state name in the two character format. To populate the entries, I used the following function: =IF(ISBLANK(R2), XLOOKUP(Q2, reference table, 2, FALSE),R2)
  + Explanation of formula: The formula is composed of an IF(),ISBLANK(), and XLOOKUP(). In essence, I want Excel to check if a cell within the “State” column is empty or blank. If true then I want Excel to execute a VLOOKUP to fill in the missing data with our reference table. However, if the cell is not empty or blank, then I want Excel to just return what was originally there.
  + Important to Note that due to this being International, not all countries have States in terms the way the US has “States”. Instead the Nation abbr is used.
* The “City” column contained names of the city in which the company was located. Most of the entries contained the full name of the city however one entry, “NYC” was in short name format. To standardize the data, using Find and Replace I set “NYC” as my find value and “New York City” as my replace value.
* A concerning column was the “Sales” column. The sale reported did not match a multiplication of “Order Quantity” and “Price Each”. To address this concerning issue, I came to the solution of changing the “Price Each” entry. This accomplishes not changing the reported sale number and also does not change the quantity ordered. Reason for changing the “Price Each” was because a column called “Deal Size” seems to be dependent on the reported sale number. I also do not want to change the reported quantity number as this could mean some orders have more or less of what was actually shipped.