**Data Cleaning and formatting**

The XYZ Company Manager wants a clean data from You being Analyst and he suggested following changes to be reflected in the dataset and wants you to do implement following changes along with a presentation (ppt), so do create a new ppt and paste the results in it.

**Product Lookup table**

1. Confirm that column headers have been promoted and that all data types are correct.
2. ProductCost and ProductPrice are Changed to Currency and Product Size is removed as got multiple data formats from Product Lookup Table.
3. Add a new column to extract all the characters before dash (“-”)in the Product SKU column, and name it “SKU Type”.
4. Extract all the characters before 2nd dash (“-”), instead of the first.
5. Replace zero (0) in the Product Style Column with “NA”
6. Close and load to your Data Model.
7. From ProductName remove the size part and put it in another column and put provide a size no for text data and NA if the size no is not available. (Red=42, Black=40, Blue=44, M=40, L=42, XL=44)
8. Take ProductColor and provide a Ascii code/hexa codes for every color (for multicolor take #FE8080).
9. There is 50 percent discount on each item (find it on each ProductPrice) and rename the Column name as Discounted Price .
10. He want you to do few calculations such as:

i) What is the average product cost?

ii) How many Colors do we sell or product in?

iii) How many distinct customers do we have ?

iV) What is the maximum annual customer Income?

V) Return the tables to original state.

**Customer Lookup Table**

1. Change Male with “M” and Female with “F” from Gender Column.
2. Create a new Column and provide the Port no as 443 for annual income 170000 and for other annual income provide a port no of 80. (Annual Income).
3. The Manager wants a new column created from “Email Column”which only holds domain name (Adventure Works).

**Solution:** Duplicate Email Column as “Domain Name”: Go to **Add column** Section and then click on Email Column and then click on Duplicate. Here in Domain Name, I want to keep only Adventure works so go to Transform section and use delimiter from Extract Section. Use delimiter from Extract Section, Text After Delimiter. Replace(Value to find – Replace with Space)

1. It even hides the Email Address for these Customers by putting XYZ for these customers and for others use the same email id.   
   Hint: Use Condition on AnnualIncome.
2. Merge the Prefix+FirstName+LastName, so for that click on all the 3 columns (using CTRL), click Add Column (to create a new column) and then click Merge.
3. Merge MaritalStatus and Gender Column and then Split that Column again in MaritalStatus and GenderColumn ().
4. Create a column named Parents and based on conditions, get me the result as given below.

A screenshot of a computer

Description automatically generated

**Sales Data 2022 Table**

1. From OrderNumber Column, only remove the text part as Numbers are sufficient to provide the Information (Transform Section).
2. Find the sum of ItemsOrdered using Group by: Perform Group By ProductKey on Sum of OrderedQuantity, (Group By in Transform Section)
3. Perform Group By Index on Sum of OrderedQuantity (Group By in Transform Section)
4. Applying Condition on OrderQuantity and then creating another column(Quantity Type) based on the given if else condition🡪 Add Column🡪 Conditional Column

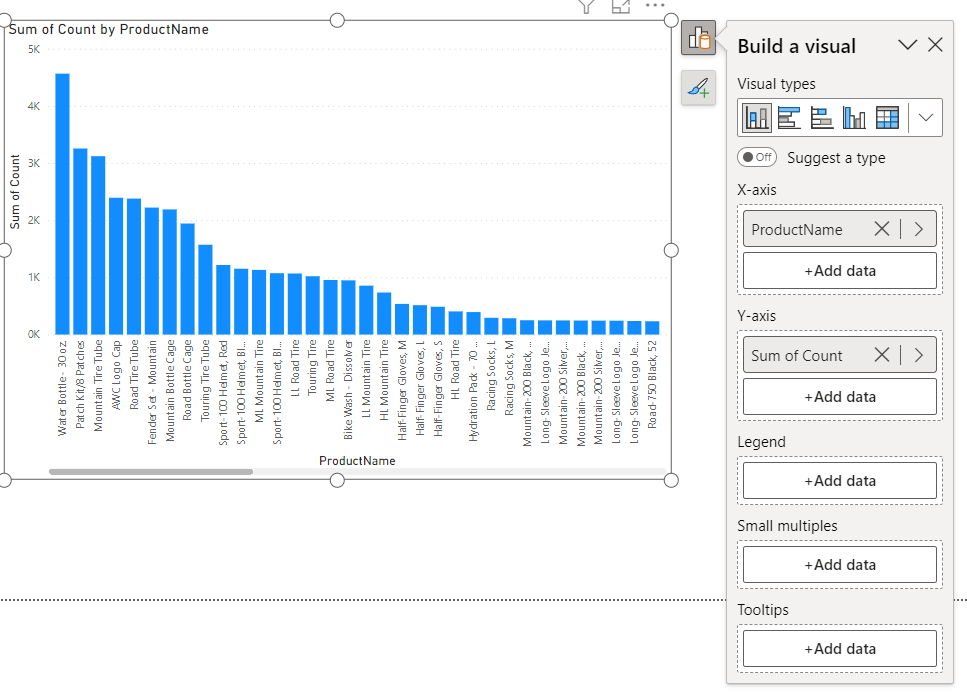
**Sales (Unpivot Demo)**

1. Unpivot North, South and Central region and rename it as Region and Item Sold.
2. Pivot the Region Vs Item Sold from Sales (Unpivot Demo)

Do it wilt Tables:

1. Provide Index No with all the tables.
2. Provide a appropriate column name to all the new columns that you have created.

Create this Chart



Others:

1) https://en.wikipedia.org/wiki/List\_of\_automobile\_sales\_by\_model

2) https://www.kaggle.com/datasets/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring