Data Cleaning:

Dataset information:

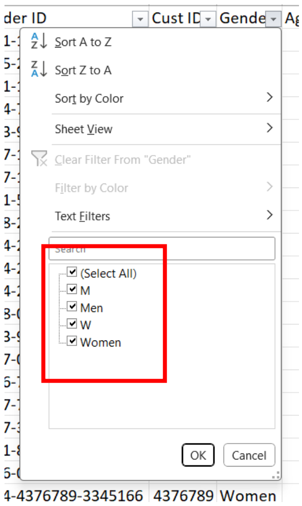
This dataset has information about the company's products, sales & revenue etc.

1. **Order ID** – a unique order id for each order
2. **Cust ID** – a unique customer id for each customer
3. **Gender** – gender of the customer (Men & Women)
4. **Age** – age of the customer
5. **Date** – date when the order was placed
6. **Status** – status of the order, there are 4 statuses (1. Delivered, 2. Refunded, 3. Returned, 4. Cancelled)
7. **Channel** – online platforms such as Amazon, Myntra, Flipkart etc. these are the channels where Vrinda store sells its products. (there are 7 different online channels)
8. **SKU** – Stock-keeping unit
9. **Category** – category of the product (there are 8 different categories)
10. **Size** – size of the product (mainly there are clothing products & their sizes are – S, L, M, XL etc.)
11. **Qty** – quantity of the ordered products
12. **currency** – currency it is in Indian rupees, because Vrinda store operates in India
13. **Amount** – amount of each product
14. **ship-city** – name of the city, where the product will be delivered
15. **ship-state** – name of the state, where the product will be delivered
16. **ship-postal-code** – postal code of the city, where the product will be delivered
17. **ship-country** – country is India
18. **B2B** – I don't know about this column (maybe it indicates about who is the buyer a business or a customer) this column contains only two values "TRUE" & "FALSE"

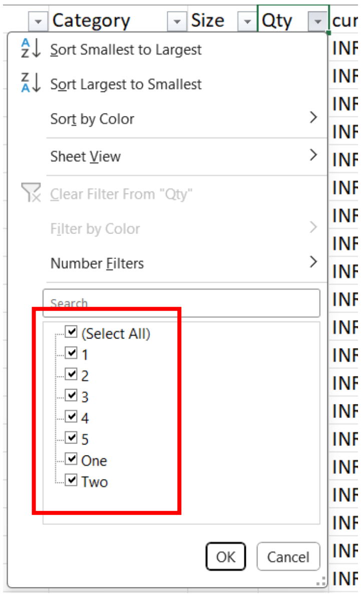
This dataset has 18 columns & 31047 rows and it does not contain any missing values.

Issues with the dataset:

1. **Gender** – This column has 2 values Men and Women, but there is "M" instead of "Men", and "W" instead of "Women".



2. **Qty** – This column has 5 values (1, 2, 3, 4 & 5), but there is “One” instead “1” and “Two” instead “2”.



☑ (Select All)
☑ 1
☑ 2
☑ 3
☑ 4
☑ 5
☑ One
☑ Two

Solving the issues:

1. **Gender** – we will use *find & replace* to replace the values: “M” to “Men” and “W” to “Women”
2. **Qty** – we will use *find & replace* to replace the values: “One” to “1” and “Two” to “2”

Feature Transformation:

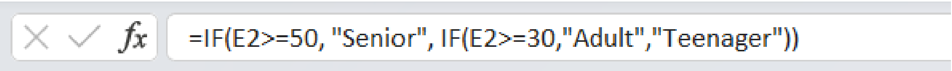
Here we have to transform some features/columns in this dataset, so that we can effectively solve the business problems.

1. **Age** – we will create **“Age Group”** column from this **Age** column, and the new column contains 3 values *“Adult”, “Senior”* and *“Teenager”.*

2. **Date** – we will extract *Month Name* from this **Date** column and name the new column as **“Month”**

Transformation:

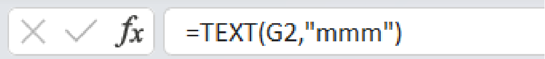
1. **Age** – To create **Age Group** column form the **Age** column, we have to use this function;



```
=IF(E2>=50, "Senior", IF(E2>=30,"Adult","Teenager"))
```

This IF function will make 3 categories like if the person’s age is greater than 50 then we will assign as “Senior”, if the person’s age is greater than 30 & bellow 50 then we will assign “Adult” and who’s age is less than 30 we will assign as “Teenager”, because our data is not very big so we will assign Teenager for those who’s age is less than 30. You can change the formula according to your needs or the problem you are solving.

2. **Date** – To extract month name from **Date** column we will use this function;



```
=TEXT(G2,"mmm")
```

This function will give us first 3 letters of the month like if the month name is “January” so we will get “Jan” etc.

If you have any suggestions or find any mistakes, then you feel free to share your thoughts, I would love to engage with you.
Thank you…!