**Data Analyst Report – PEI**

Data Preparation – I downloaded the three files from the websites mentioned below

Customers.xls **- [**[**https://easyupload.io/xkspfi**](https://easyupload.io/xkspfi)**]**

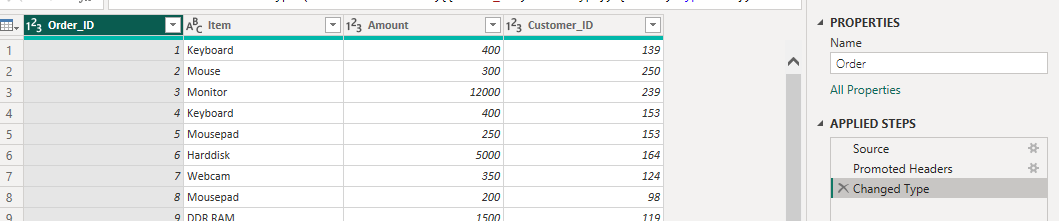
Orders.csv **- [**[**https://easyupload.io/vmbljr**](https://easyupload.io/vmbljr)**]**

Shippings.json **- [**[**https://easyupload.io/okfmxk**](https://easyupload.io/okfmxk)**]**

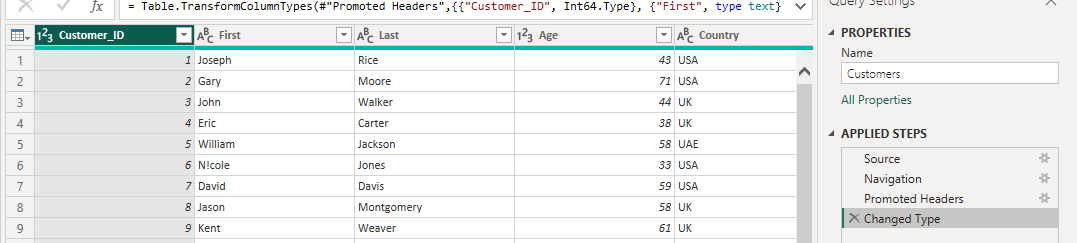
Initially, opened the files in excel for .csv and .xls and Notepad++ for JSON, to check for any inaccuracies or missing data and confirmed that there were none so no transformation or data manipulation was needed and hence loaded the 3 files into Power BI for creating charts and graphs.

Below are the screenshots for the transformations that were done to the 3 loaded tables

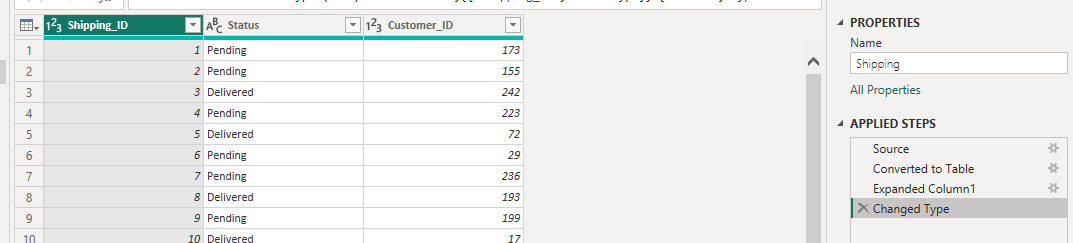
Orders Table



The transformations done to the Orders table is that the first row is treated as the Header for the column names and the Order\_ID type was changed from text to number.



Next is the Customers table and the changes done to that are similar to the Orders table and the first row is treated as the column names and the Customer\_ID type was changed from text to number



Finally is the Shipping table and since it is a JSON file, it is first converted into a table format with rows and columns separated and then the Shipping\_ID was changed from text to number format.

Accuracy - I was able to check initially for any accuracy errors but it looks like all the fields have some values and there are no duplicates in the customer table which should be a table with unique values.

Completeness – When it comes to completeness, I found that the data is lacking a number of key columns, eliminating which would not give us a good analysis. Also, there should be other files as well in addition to the customer, order and shipping csv and JSON files. Some of the key fields missing from those files are as below

Customer – State column, ZIP column, Creation Date column, Phone Number column

Order – Order Date column, Product ID column

Shipping – Shipping Date column, Shipping Address column, City, State & Zip column

Also, another table should have been made namely the Products table that would have columns like Product ID, Product Name, Description and Quantity

Reliability – The source data seems reliable but doesn’t seem to be the complete data. For a total of 250 unique customers, the Order table has only 160 unique Customers ID’s, while the shipping table has only 154 unique Customers ID’s which seems a bit low and thus it seems like this is a snapshot of the data and not the full data.

**Requirements for future datasets**

Based on the findings above and the shortcomings in the current dataset, below are the columns that should be there in the future datasets for better and accurate analysis:

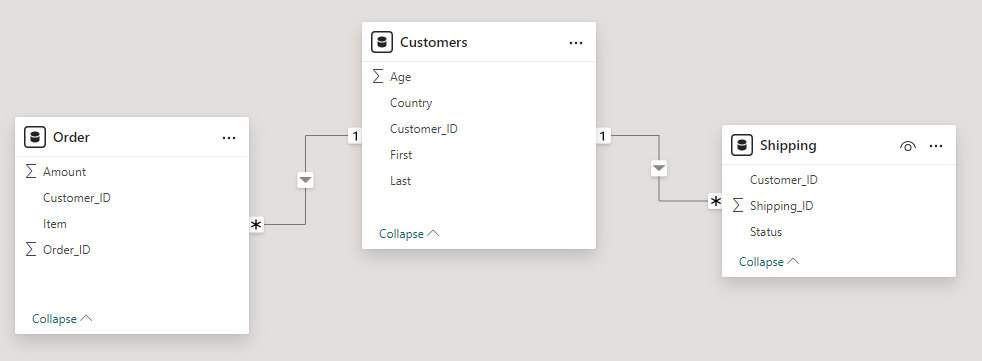
Customer – ID, First Name, Last Name, Address including (State, City, ZIP, Country), Phone Number, Email, Age

Order – Order ID, Product ID, Customer ID, Order Date, Amount

Product – Product ID, Product Name, Quantity, Price, Product Description

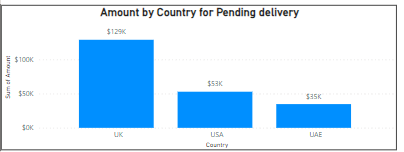
Shipping – Shipping ID, Shipping Date, Order ID, Shipping Address including (State, City, ZIP, Country)

**Existing Data Models**



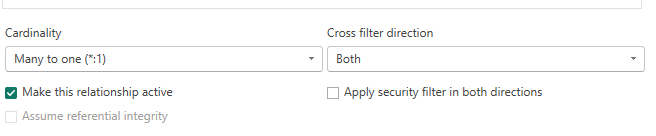
This is the present data model where the Customer table has unique customer’s information and the Shipping and Order table are joined to the customer table with a Many-to-One relationship. Based on the suggested future datasets the model would change a little where Shipping would have Order ID column and it would be connected to the Order table which in turn would be connected to the Customer table so we can figure out which customer actually placed a particular order.

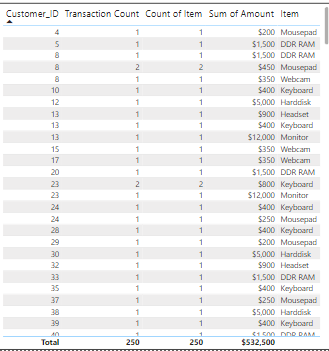
**Insights from the current data to the Sales team**

1. 

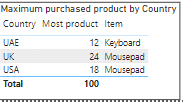
Above is the graph for the total amount spent with ‘Pending’ delivery status for each of the 3 countries, we can see that the UK has the most number of orders and hence higher number of ‘Pending’ deliveries while UAE has pretty low order count.

Also, for this particular graph I had to make the relationship between Customer and Shipping as “Bi-directional” to allow data flow from Order to Shipping to get the country and “Pending” status along with the total amount.

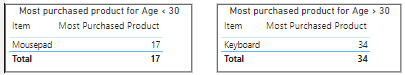


1. 

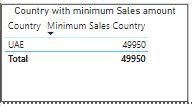
The above graph shows the total count of transactions and the item and since item count is 1 in each transaction and hence transaction and item count are identical. Some customers have more than one transaction but most of the customers have just 1 transaction which further states that the data may not be complete or it could be just a snapshot of the much bigger dataset.

1. 

The above graph gives us the maximum of the products purchased country wise and we can see that mousepad is most ordered by people from UK & USA and hence is a famous product for those 2 countries but keyboard on the other hand is famous in the UAE. Other products may not be marketed well in those countries and that could be something the marketing team could look at to further enhance the same of other items from the dataset.



From the above graph, we can clearly see that people who are younger say less than 30 desire to use the mousepad since they have an external mouse more often and find it more convenient to use, but people of age greater than 30 mostly prefer keyboard for typing since they are more accustomed to it. From this info, we can target sales of mousepad along with mouse and sell them as a combination for people who are less than 30 years old.

1. 

The above graph clearly shows that UAE is the country with the least sales and since that country has the least orders so it does make sense data wise. If we want to increase sales in this country then we could have some offers in place so that people place in more orders and also see if he site is getting enough traffic and if not then partner with Google to rank them higher so that people visit those sites more often for buying stuff.

**Insights to other team members**

Firstly, from what we saw about the data in the first place there are a lot of missing fields in the tables that we got and also there is a PRODUCT table completely missing so I would ask the data engineer to look into the issue and see if they can fix it and provide us with that data. If they are getting data from some external website then ask that team to provide the complete information.

Secondly, to the data scientists I would present the insights to them from the above graphs and ask if they see the same through their analysis by employing the Machine Learning & Data Science algorithms and if they found anything interesting and would ask them to share with our team and if useful we could share that with the sales team as well to help them make better decisions.

Thirdly, to the non-technical stakeholders I would be clear with them as to how the analysis was done and the kind of data we had and since this seems like a snapshot of the entire data so I would ask them to trust their intuition as well and not just trust the analysis entirely since less data is not very useful for important business decisions. I would ask our team to find data that is more complete and that accurate so that we could use that to do the analysis and then the other teams can be more trustful with the analysis that we will provide in the future.