INSTACART MARKET BASKET ANALYSIS

& Market Segmentation (Web Dataset)

Savla, Neel Girish

CIS 5250 Project Part 2

**A. Data Sets:**

**Instacart Market Basket Analysis**

URL: - <https://www.kaggle.com/c/instacart-market-basket-analysis/data>

**Web Data Set – Market Segmentation.**

URL :- <https://docs.google.com/spreadsheets/d/e/2PACX-1vTrZU_gXA57Vzl9T9grVKhapWLJobFXfFFTs7b6f8KYyf8gpwJRwHsrNlfRlV8-Y_bfRvEauN4ElSCx/pubhtml>

**B. Data Description:**

**Instacart Market Basket Analysis Data Description:**

The dataset is a relational set of files describing customer’s orders over time. The dataset contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, orders provided is between 4 and 100 of their orders, with the sequence of products purchased in each order. Data also provides the week and hour of day the order was placed, and a relative measure of time between orders.

There are different files/tables used in this project, those are, aisles, department, product order prior, product order train, orders & products. Each of them has unique Id. To describe them in detail, aisle table will have the aisle id and aisle number, department will have its Id and department name, product order prior will have record of customers prior order for each product, product order train has record of current items purchased, orders will tells, to which set (prior, train, test) an order belongs and product will have product description.

**Market Segmentation Data Description:**

Dataset contains 10,000 observations or we can call it transactions of 5 variables. Each row represents the transaction made by the reps. Each column contains the attributes of this dataset include: reps - Representative who are involved in the promotion and sale of the products in their respective region. products - There are 12 brands of products promoted by the company. qty - Quantity sold in unit’s revenue - Revenue generated for each transaction region - There are 4 regions - East, North, South and West India.

**C.** **Examine the Physical properties of the data:**

**Instacart Market Basket Analysis:**

Aisle.csv Table:

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Properties |
| Aisle\_id | Numeric | It contains values from 1 to 134. |
| Aisle | String | It contains unique values which describes each aisle. |

Department.csv Table:

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Properties |
| Department \_id | Numeric | It contains values from 1 to 21, which means there are 21 departments. |
| Department | String | It has values representing different departments like, frozen, bakery, etc. |

Order\_Products\_Prior.csv Table: Previously ordered data.

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Properties |
| Order\_id | Numeric | It has 3.42 million records. |
| Product\_id | Numeric | It has 5K recorded data. |
| Add\_to\_cart\_order | Numeric | Number of items added to cart. |
| Reordered | Numeric | It has values in 1 and 0, 1 represents yes, reordered and 0 represents not reordered. |

Order\_products\_train.csv Table: Currently ordered data.

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Properties |
| Order\_id | Numeric | It has 3.42 million records. |
| Product\_id | Numeric | It has 5K recorded data. |
| Add\_to\_cart\_order | Numeric | Number of items added to cart. |
| Reordered | Numeric | It has values in 1 and 0, 1 represents yes, reordered and 0 represents not reordered. |

Orders.csv

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Properties |
| Order\_id | Numeric | Its has 3.42 million data. |
| User\_id | Numeric | It has 2k user data. |
| Eval\_set | String |  |
| Order\_number | Numeric | It is categorized between 1 to 100 number. |
| Order\_hours\_of\_day | Numeric | Its shows hour data between 0 to 23 hours. |
| Days\_since\_prior\_order | String | It contains number of day before the order was done. |

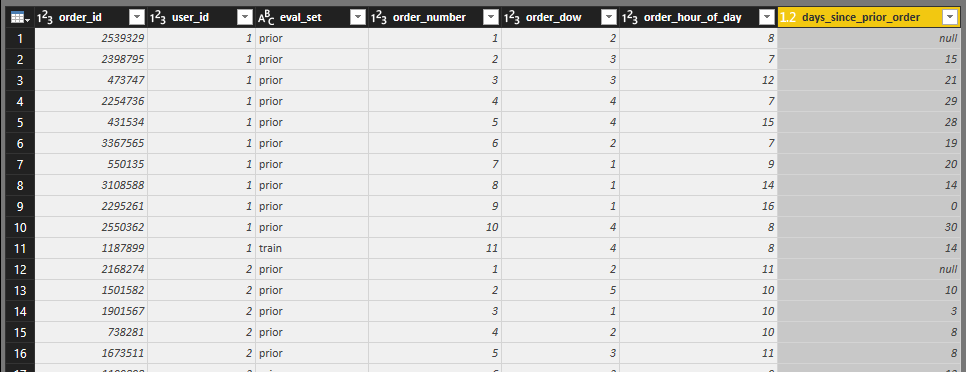
Products.csv Table:

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Properties |
| Product\_id | Numeric | It has 5K product data. |
| Product\_name | String | It has 49688 different product names. |
| Aisle\_id | Numeric | It contains values from 1 to 134. |
| Department\_id | Numeric | It contains values from 1 to 21, which means there are 21 departments. |

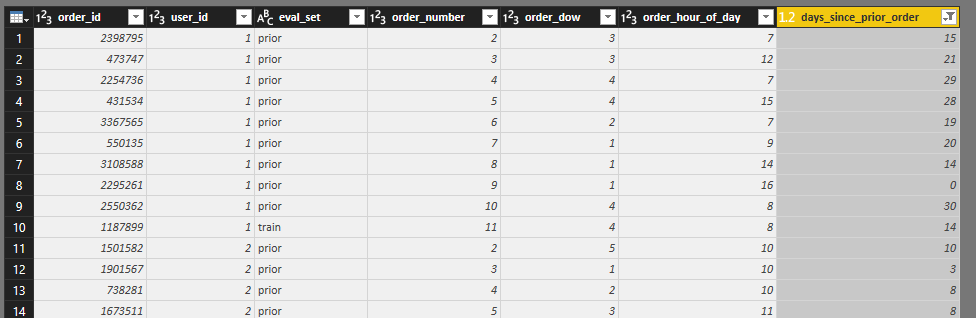
**D. Data Cleaning**:

1) **Missing values:** The data contains null value which means some values are missing in the last column name days\_since\_prior\_order. So, here we are removing the rows containing null value.

**Before:**

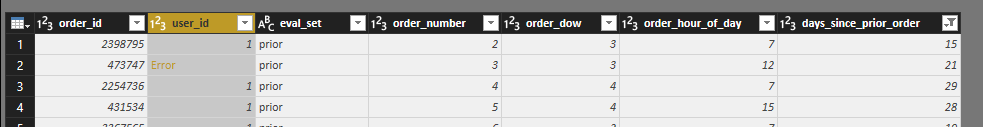


**After removing missing (null values):**

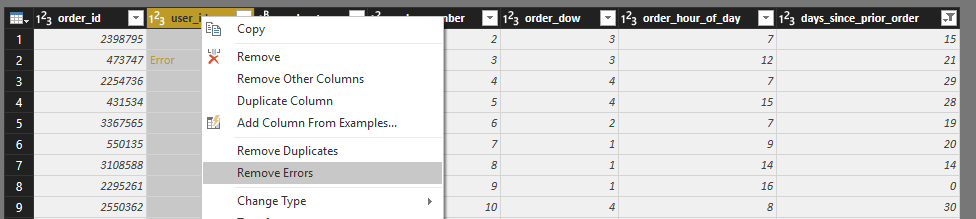


2) **Erroneous values:** Here, in this table the column user\_id, the data is of a similar type but if the data is not of the same type it is error.

**Before:**



**Step:**

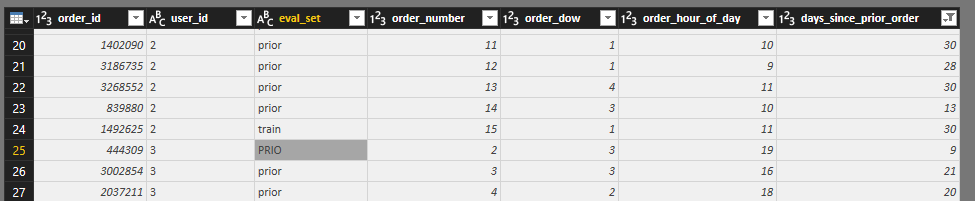


**After Entire row gets deleted:**

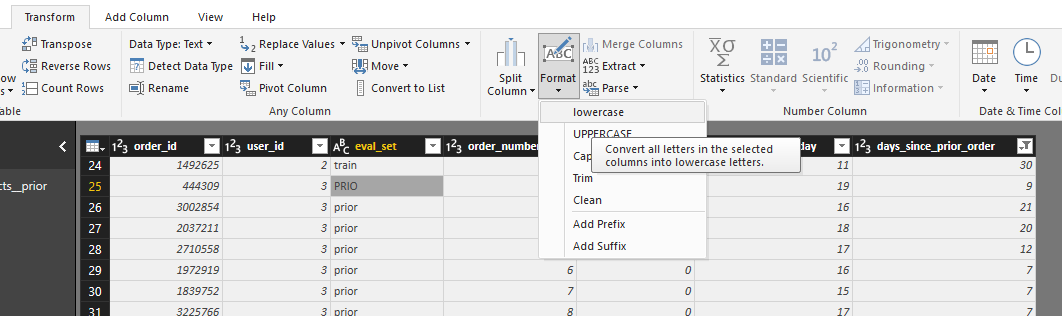


3) **Inconsistencies:** Here, the data in eval\_set column has to be in same format but there are some entries which has data in upper case, thus converting them all so similar format which is lower case, for better analysis.

**Before**:



Replacing inconsistent data in its original form.

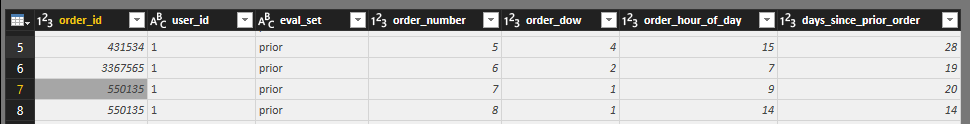


**After**:

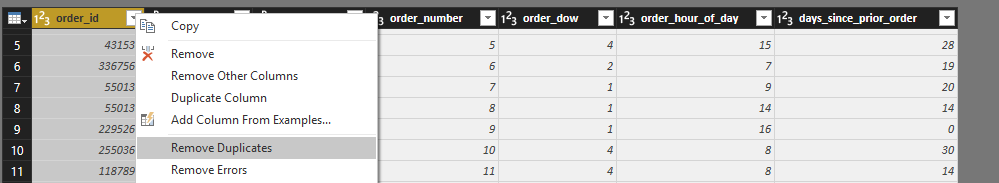


4) **Duplicate records**: Here, the order\_id has duplicate entry which is not correct the column should contain unique value as every order has its own unique id, thus removing duplicate value in this stage.

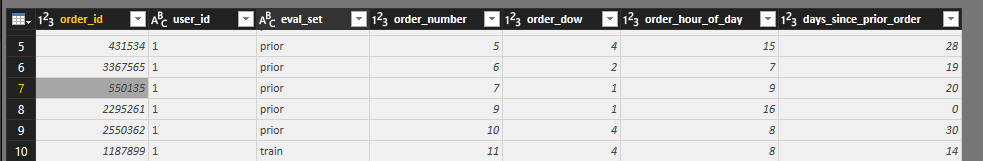
**Before**:



**Step**:

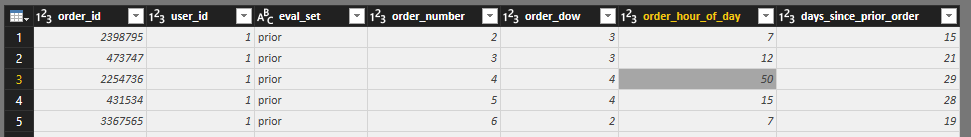


**After**:

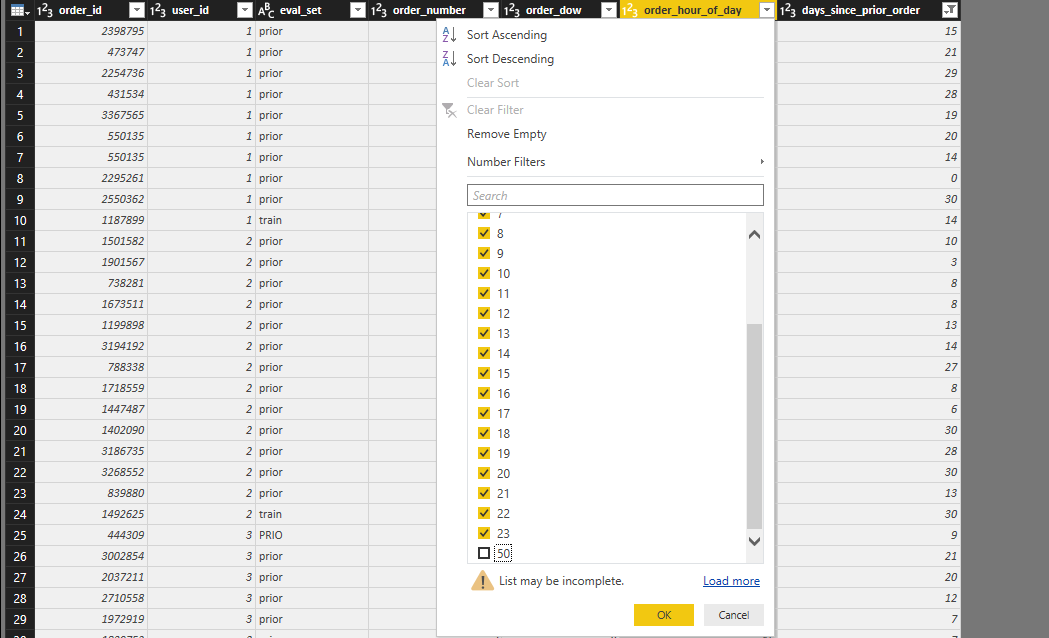


5) **Out of date**: Here the hours of day are from 0-23 the value cannot exceed 23.

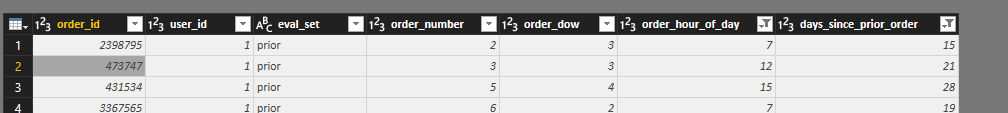
Before:



**Step:**



**After**:

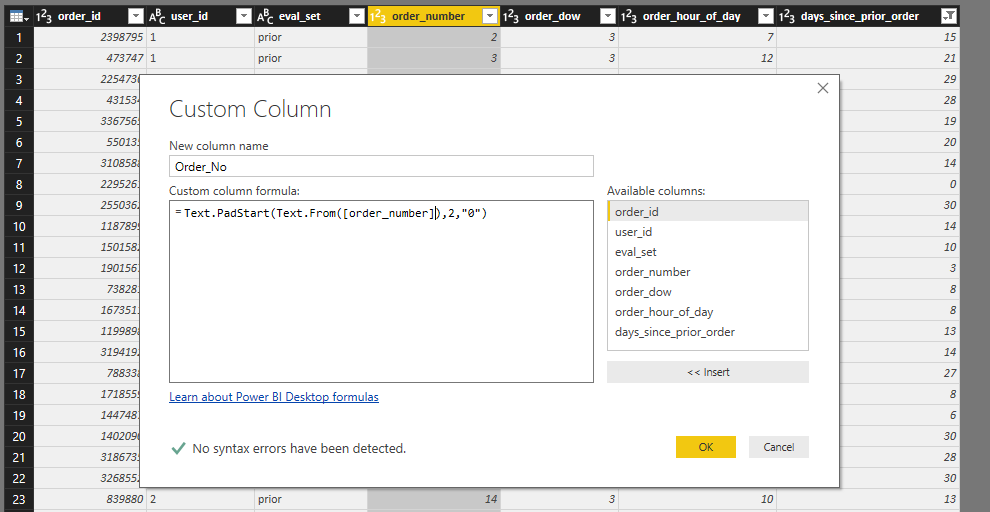


6) **Leading or trailing issues**: Here, the order number has numbers in one digit and 2 digit, so for getting them in a unified form we are converting every number into two digits by padding a “0” leading every single digit number.

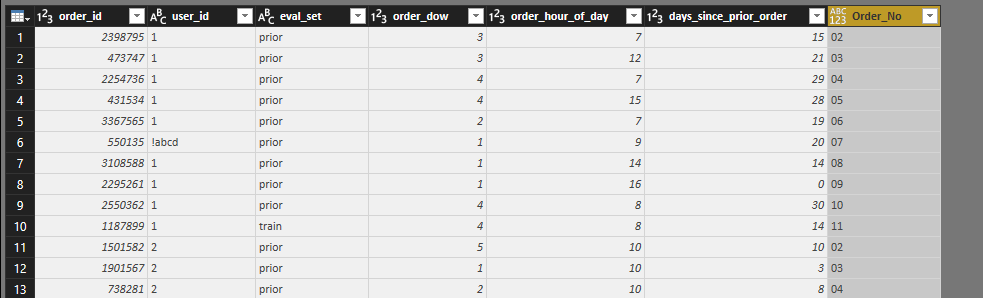
**Before**:



Created a new custom column and used the formula: **Text.PadStart(Text.From([Month]),2,”0″)**



**After** :

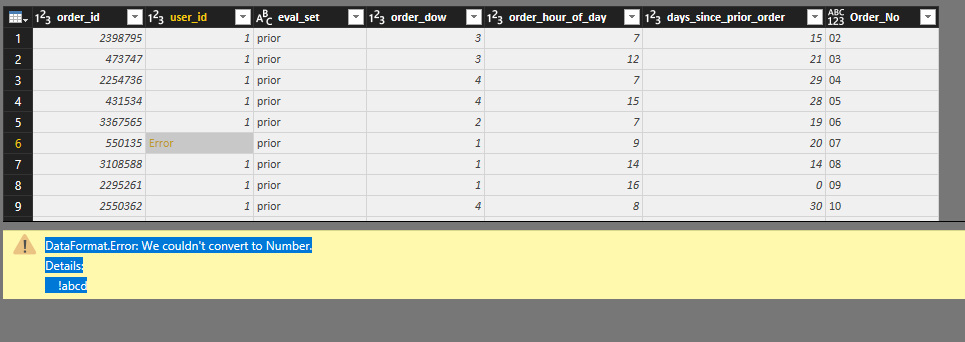


7) **Date format:**

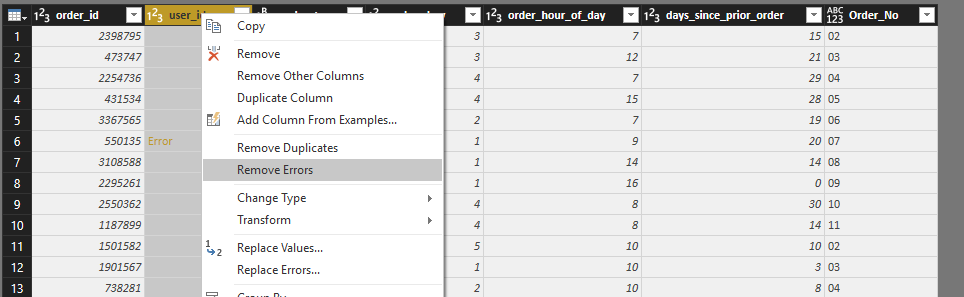
Don’t have any table containing date data in it.

8) **Other data types**: Here, the user\_id has numeric data type but contains a value “!abcd” which is not numeric removing the other data type content.

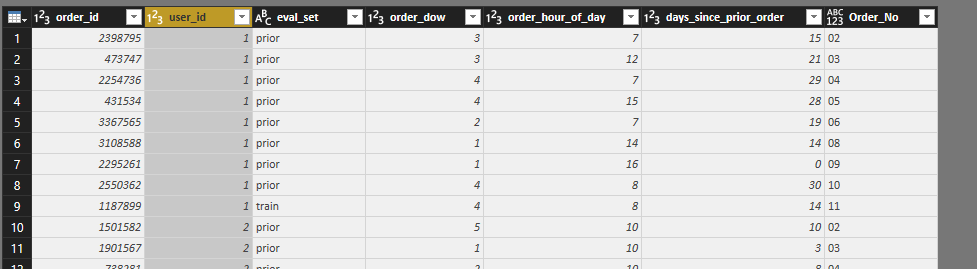
**Before**:



**Step**:

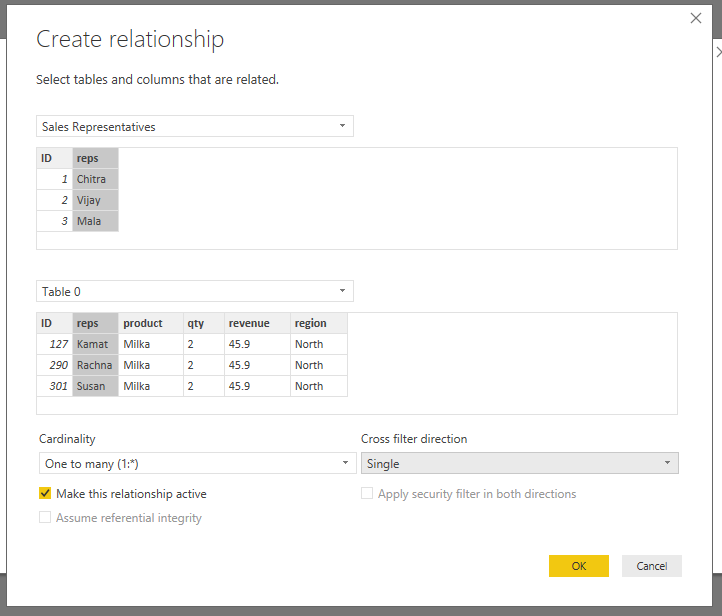


**After**:

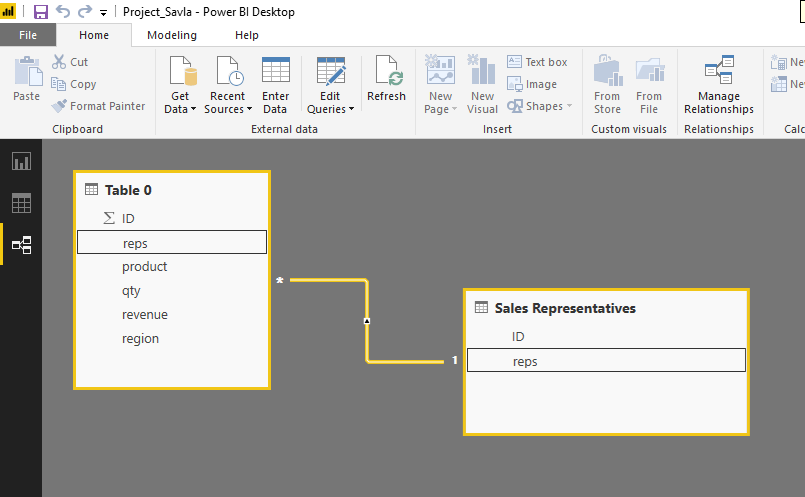


**E. Relationship between the Sales Representatives.csv data and web data.**

**Step**:

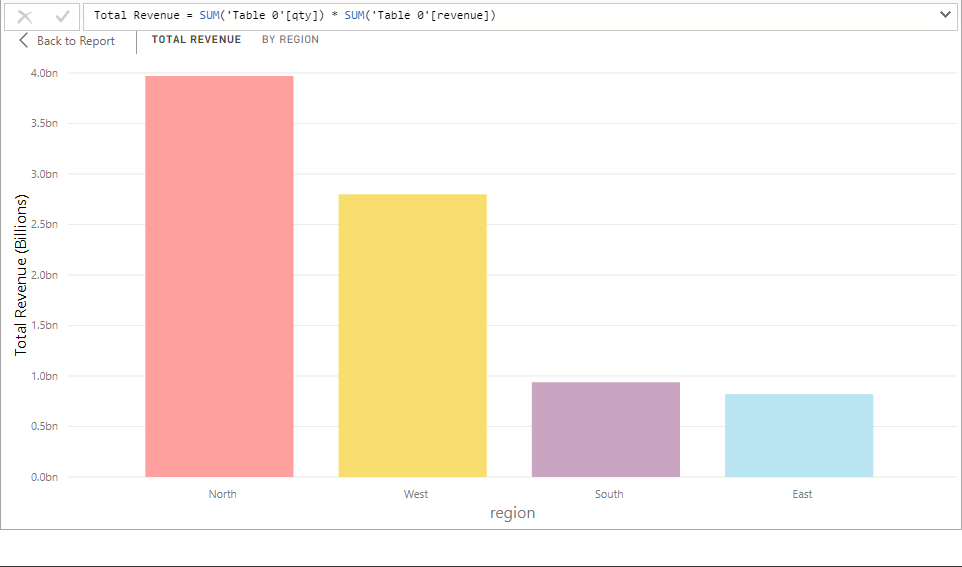


**Relationship**: In this relationship the table named Sales Representatives is the excel file and the table named Table 0 is the table generated from web. In the relation below the reps (i.e Representatives name) column is the primary key holding all the unique values and the reps (i.e Representatives name) column in Table 0 is the foreign key. This relation is used to get help in visualizing certain scenarios such as highest sales by the sales representatives, which representative generated highest revenue and for which product and also in what region. Thus, table relationships will help in further analysis. It is a one to many relation the reps by Sales Representatives table is one and reps by Table 0 which is web data is many.



**F. 1 Measure:** Here, a measure is created named as ‘Total Revenue’ which is a sum of column quantity(qty) \* revenue, the reason behind creating this measure is to get the overall revenue and which will help in analyzing one of scenarios to find out total revenue with respect to region (North, West, East, South). By analyzing the total revenue with respect to region we can find out the region which generates the highest revenue and the region that generates the lowest revenue which according to the graph below is North with the highest revenue and East with the Lowest revenue. Thus, we can predict the future needs of improvements in the region according to total revenue they generated.

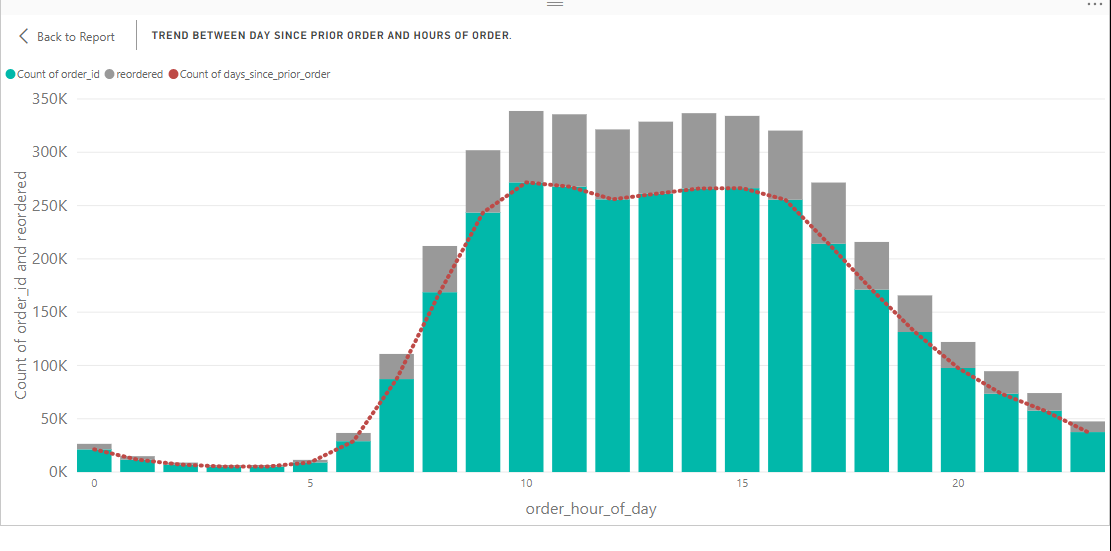
Formula: Total Revenue = SUM(‘Table 0’[qty]) \* SUM(‘Table 0’[revenue])



**G. Visuals and Dashboard:**

G.1. **Trend between Prior order and hours of order?**

Here, in the graph below shows information of the days since the prior order was made with respect to the product reordered and on the basis of count of order\_id. The x axis represents the order hour of data which contains data range from 0 to 23 as there are 24 hours in a day. The y axis represents count of order id and the count of products reordered. The green color represents the count of order id to find the hour of day when the maximum of orders is made, which according to the graph is the 10th hour of the day which is like 10am when the maximum orders are placed. The grey color bar represents the products reordered, where we can see that the reordering is done more in the hour when the ordering is done more and least, we the order hour is least. The red dotted line represents the count of days from when the previous order was done. Thus, showing the trend between previous order and reorder with respect to hours of order and count of order id.



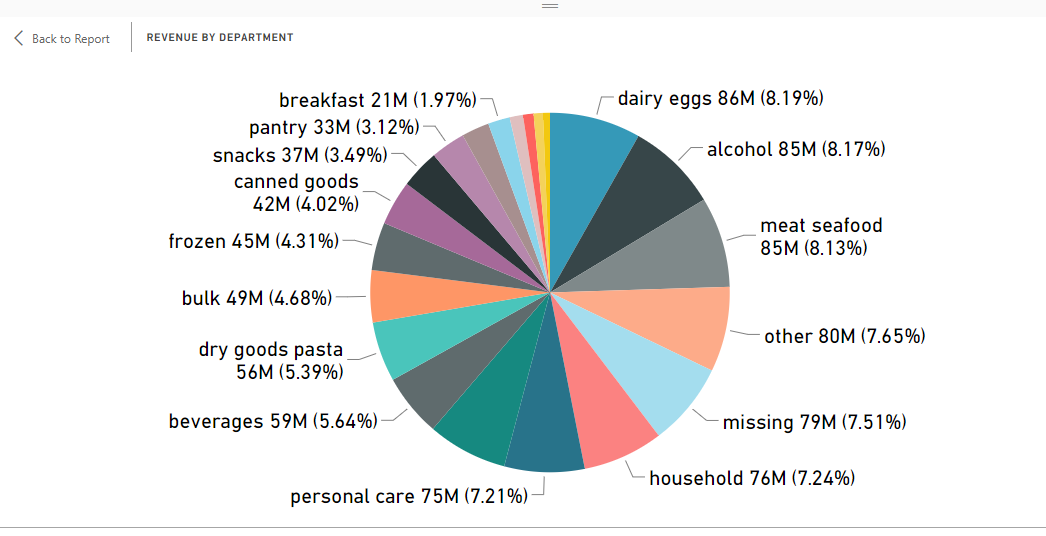
G.2. **Qty sold by sales representatives id?**

Here, the graph below represents the quantity of products sold by the sales representatives by considering their id numbers. The X axis shows the Id numbers ranging from 0 to 100, 100 to 200 and so on where as the Y axis represents the quantity of products sold. According to the graph, the reading is maximum ranging from Id between 0 to 100 which the reading is lowest for the id reordered between range 100 to 200. Thus, the sales representatives ranging from id 0 to 100 are have the best sales and the sales representatives ranging from 100 to 200 have some of the lowest sales.

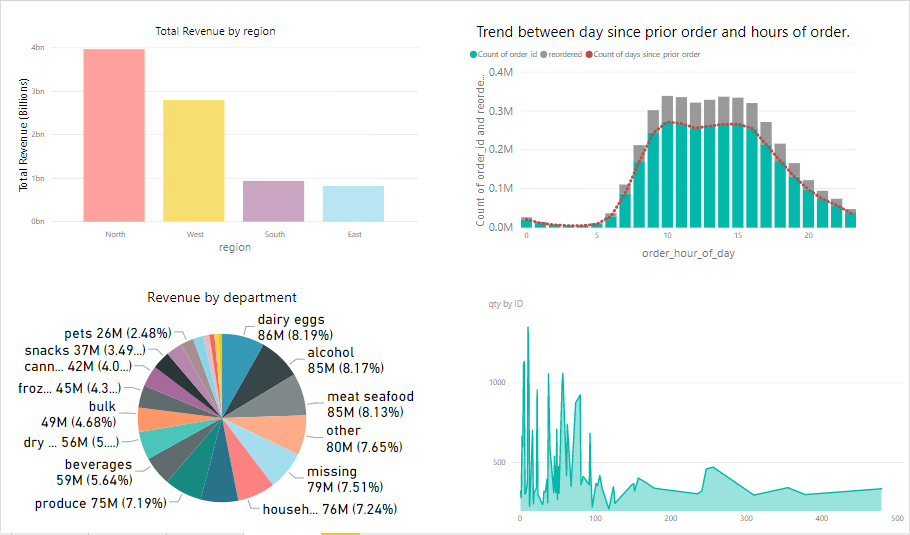
A picture containing text

Description automatically generated

G.3. **Which department has the highest revenue generated**?

According to the chart below there are in total 21 different departments and based on the data a pie chart is generated which represents the revenue generated by different departments. Thus, by reading the chart we can see that the highest revenue generated by the department which is the dairy and eggs department which is around 86Millon and which comprises a share of 8.19% between all of the other departments, one more interesting reading is that the department of alcohol and meat, seafood is one million less than the highest revenue generating department. The lowest revenue generating department is breakfast department which generates only 21million and hold a share of only 1.97% between all other departments. Thus, it can be predicted that the services and products with the department holding highest revenue id the best and the lowest revenue needs improvement.



**G) 2. Dashboard:**



**Storytelling:**

Staring with the first chart at the left top corner the chart is a bar chart representing the total revenue in billions in different region of business. The X axis represents the total revenue calculated by a formula and Y axis represents the regions (north: red, west: yellow, south: purple, east: blue). The interesting reading from this chart is the maximum and minimum revenue generated by regions. The maximum revenue generated is by Northern region while the second highest revenue generated is by western region, the third one is the southern region and the last and the lowest revenue generated is by the eastern region.(1)

Moving on the next chart at the right top corner, is about the trend of orders placed prior and the orders reorder. The chart is developed using some data such as order id, hours of day, day since prior order and reorder. The x axis represents the order hour of data which contains data range from 0 to 23 as there are 24 hours in a day. The y axis represents count of order id and the count of products reordered. The green color represents the count of order id to find the hour of day when the maximum of orders is made, which according to the graph is the 10th hour of the day which is like 10am when the maximum orders are placed. (2) The grey color bar represents the products reordered, where we can see that the reordering is done more in the hour when the ordering is done more and least, we the order hour is least. The red dotted line represents the count of days from when the previous order was done. The interesting reading from the chart are, higher the order in an hour of day, higher is the reordering and also a trend of prior order can be noticed within.

Moving on to the third chart which is at the left bottom of the dashboard, the graph is about the revenue each department generates. As per the graph the departments are categorized in 21 different types. Now, according to the revenue data, a pie chart is created where according to the different revenue values the part of pies are divided into parts. Some of the pie parts are big which means more revenue is generated which some are small which means low revenue is presented, thus having different shapes according to the contribution of revenue generated. According to the pie chart reading, the highest revenue generated by the department which is the dairy and eggs department which is around 86Millon and which comprises a share of 8.19% between all of the other departments, one more interesting reading is that the department of alcohol and meat, seafood is one million less than the highest revenue generating department.(3) The lowest revenue generating department is breakfast department which generates only 21million and hold a share of only 1.97% between all other departments. (4) Thus, it can be predicted that the services and products with the department holding highest revenue id the best and the lowest revenue needs improvement.

Finally, the last chart at the right bottom is about the sales representatives and their contributions towards sales. The sales representatives are divided into range according to their id’s which means the ids are ranging from 0-100, 100-200 and so on. The most interesting thing found while reading this chart is that the Id ranging between 0-100 have some of the highest sales which means that the sales representatives in this range are good at their work, while the range from 100 -200 have some minimum sales records which means the sales representatives are lacking back in their work. The rest of the range of id represents not highest and not lowest sales which means they need to work hard in case of being one of the highest sales representative sales.

Thus, concluding the story by summarizing the charts like the first one calculating the total revenue with respect to region of business, the second chart representing the trend of orders placed prior and the orders reorder with respect to order ids and hours of the day, the third chart represents the revenue with respect to departments and the last chart represents the sales representatives and their sales.

**H. The Purpose Map:**

|  |  |  |
| --- | --- | --- |
| EXPLANATORY | EXHIBITORY | EXPLORATORY |
|  |  |  |

**Explanatory**: Explanatory in this context essentially means we as visualizers will provide the viewer with a visual portrayal of the subject’s data and will also take some responsibility to bring key insights to the surface, rather than leave the prospect of interpreting the meaning of the information entirely to the viewer.

**Exhibitory**: Visualizations are found in the final separate ‘experience’ category within the latitude of the purpose map. They are characterized by being neither explicitly explanatory nor exploratory.

**Exploratory**: Visualizations differ from explanatory visualizations in that they are focused more on helping the viewer or more specifically in this case the user finds their own insights.

**Reference:**

1. <https://www.forbes.com/sites/bizcarson/2017/11/08/amazon-whole-foods-deal-future-of-instacart-grocery-delivery/#5cfd2a516d5a>
2. <https://jungleworks.com/how-instacart-works-makes-money-revenue-business-model/>
3. <https://www.digitalcommerce360.com/2017/03/08/instacart-itemizes-its-costs-even-as-it-adds-400-million-in-funding/>
4. <https://www.axios.com/instacart-ceo-talks-about-his-34-billion-company-1513300823-d9ba9059-bca4-4eca-ae40-a717737a49f9.html>