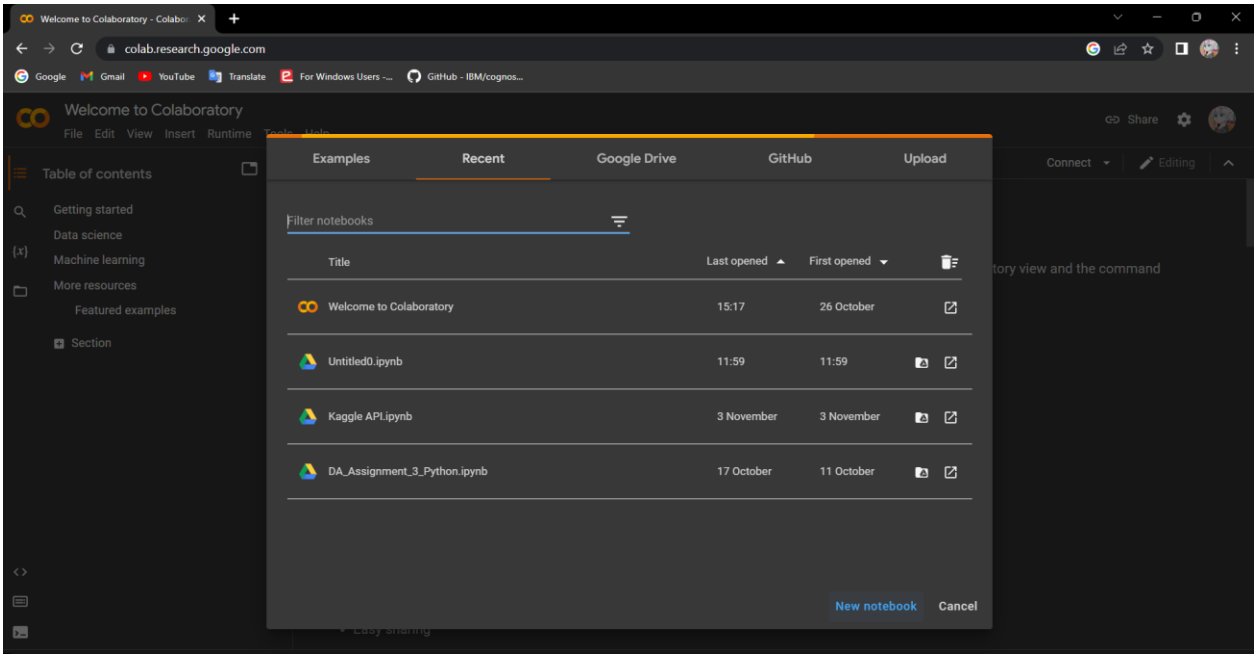
Date	03 November 2022
Team ID	PNT2022TMID30128
Project Name	Project - Global Sales Data Analytics

Sprint 1 - Dataset exploration and preparing the dataset

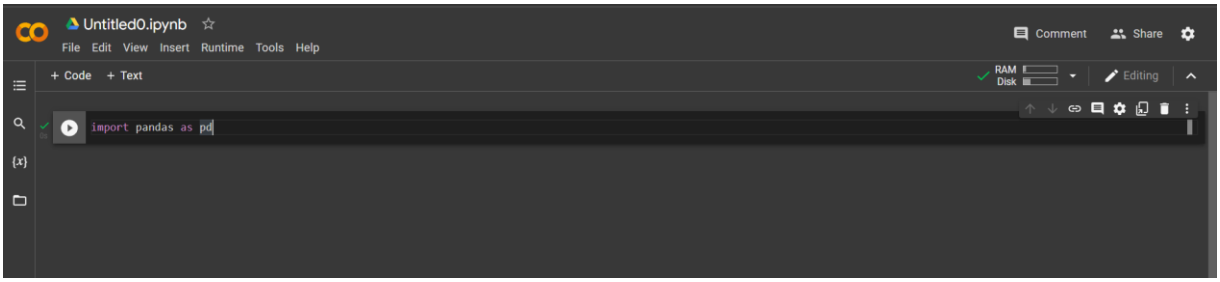
Dataset exploration: It is the process of finding what are the things that are present in the dataset.

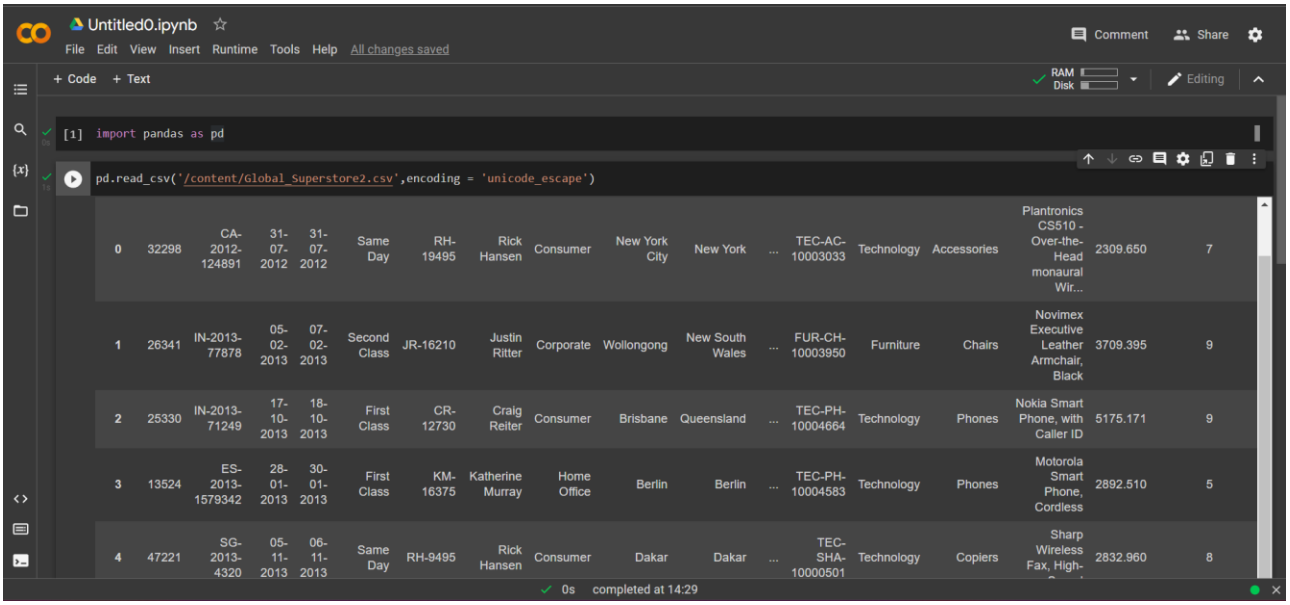
We have used Google collab to explore the dataset.

Step 1: Creating a new notebook

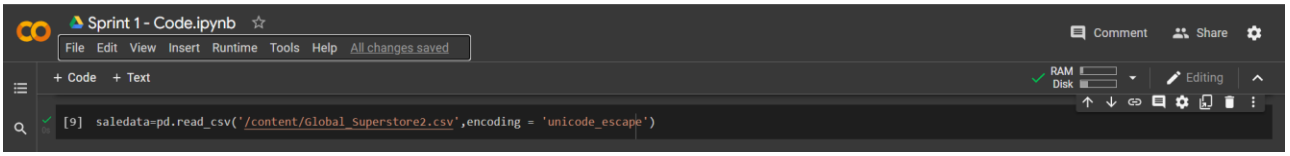


Step 2: Loading the data into jupyter notebook by using a python library called Pandas (for Python Data Analysis Library).





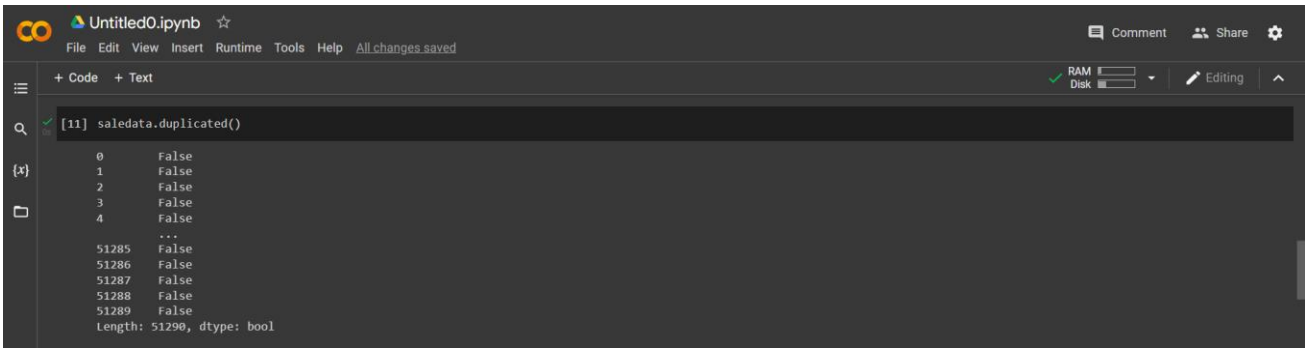
Step 3: Exploring the data by understanding the following things:



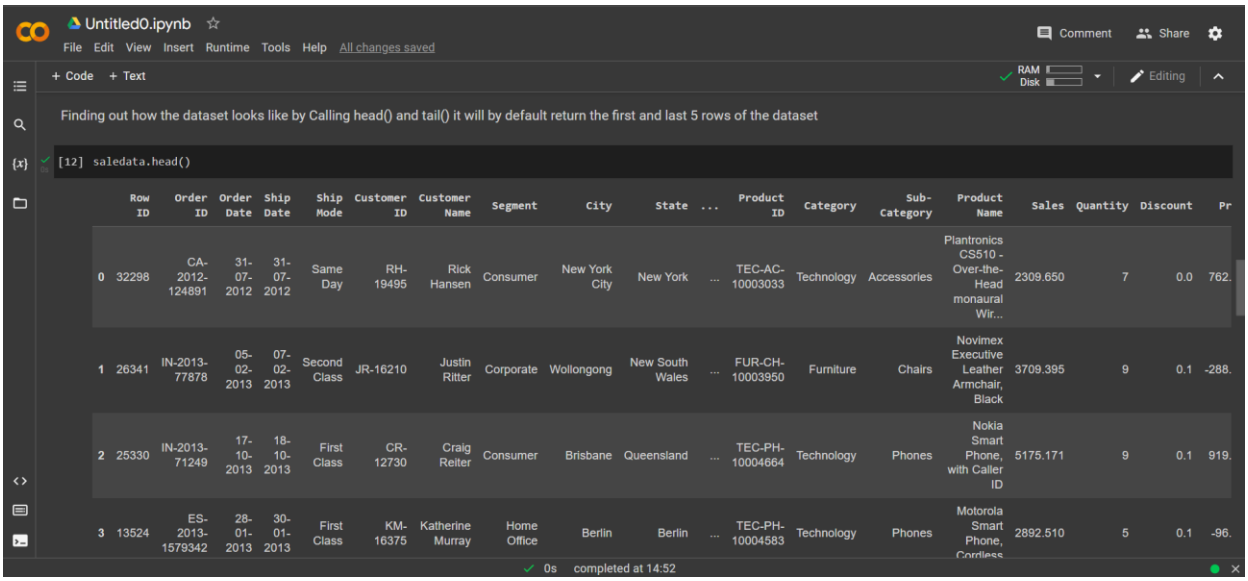
1. Finding out the total number of rows and columns in the dataset

Rows = 51290 **Columns =** 24

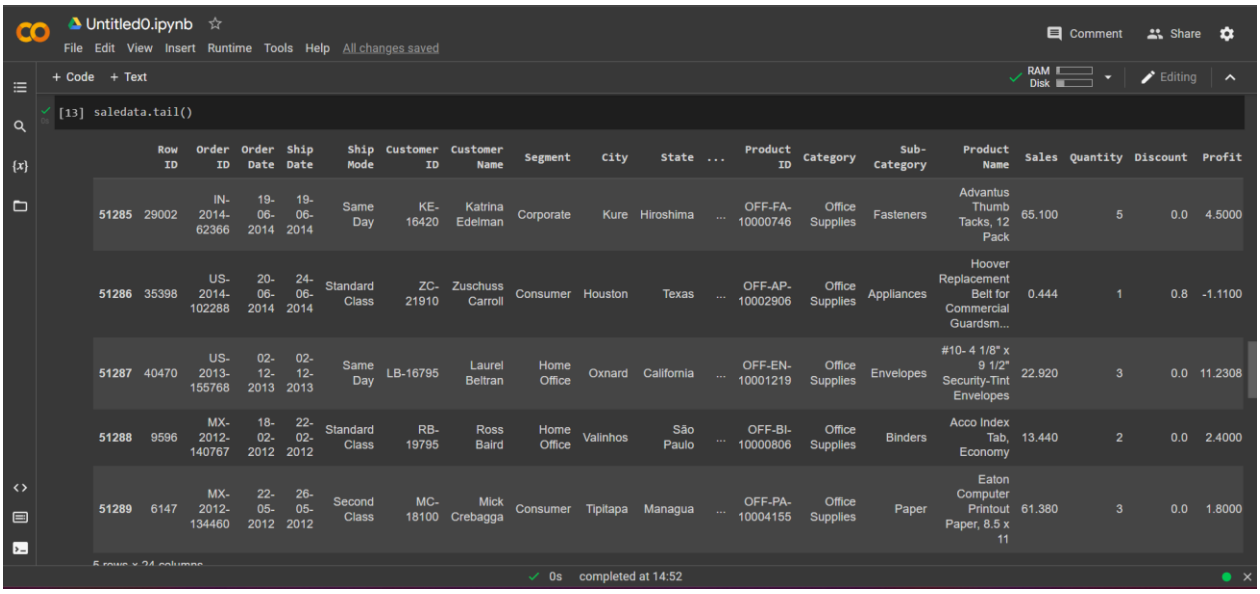
2. Finding if there is any duplicate value



3. How the dataset looks like (Finding out how the variables are organized)

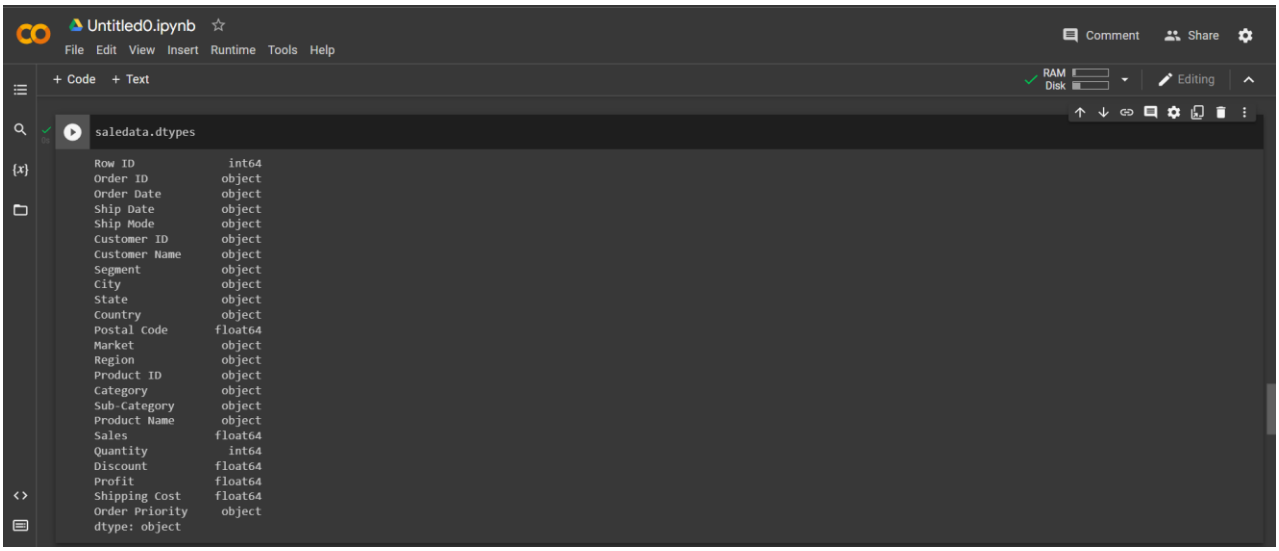


First five rows of the dataset

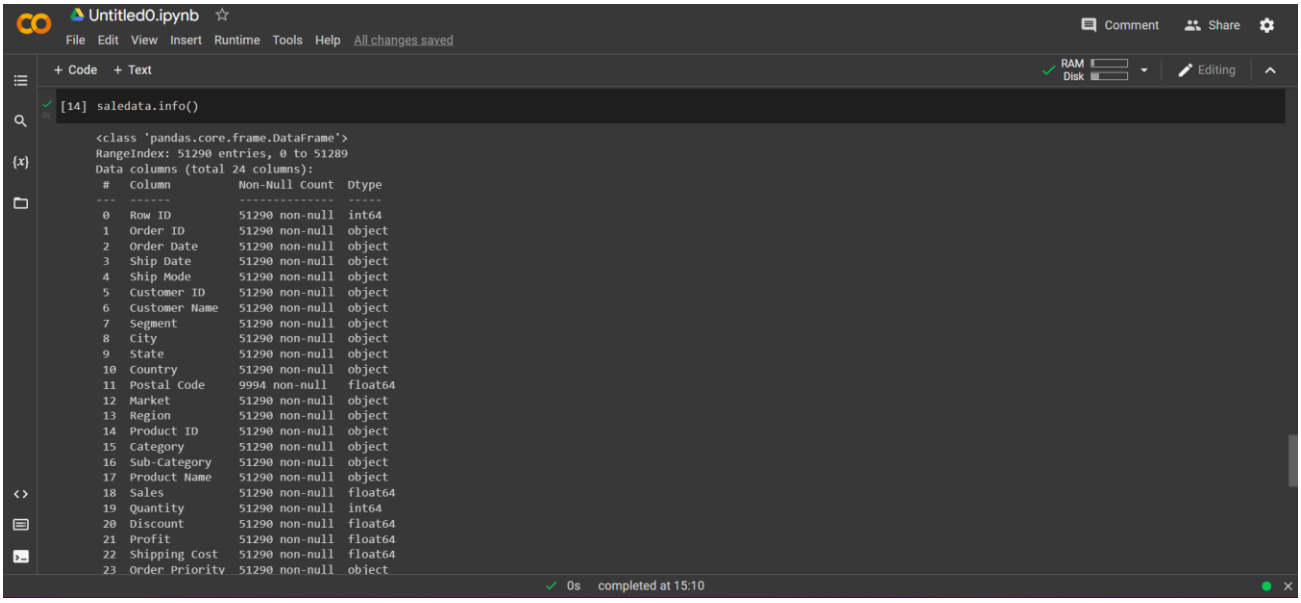


Last five rows of the dataset

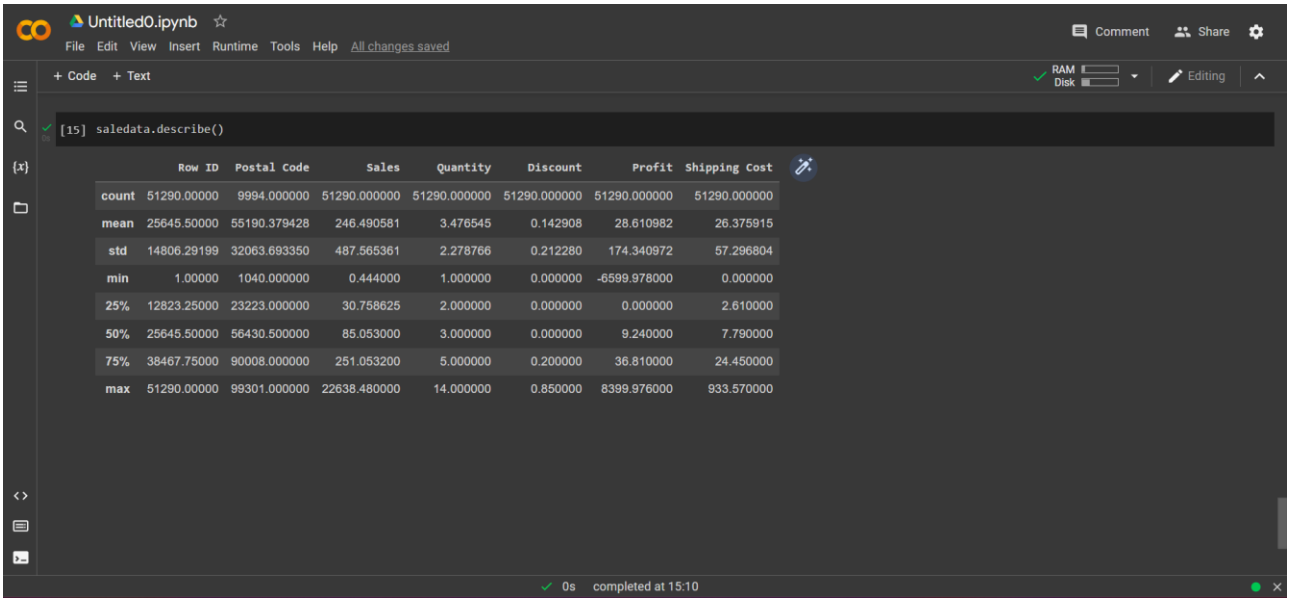
4. What kind of data types are present in the dataset



5. Finding out the overview information about the dataset



6. Finding out minimum, maximum and mean value of all numerical variables in the dataset.



	Row ID	Postal Code	Sales	Quantity	Discount	Profit	Shipping Cost
count	51290.00000	9994.000000	51290.000000	51290.000000	51290.000000	51290.000000	51290.000000
mean	25645.50000	55190.379428	246.490581	3.476545	0.142908	28.610982	26.375915
std	14806.29199	32063.693350	487.565361	2.278766	0.212280	174.340972	57.296804
min	1.00000	1040.000000	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	12823.25000	23223.000000	30.758625	2.000000	0.000000	0.000000	2.610000
50%	25645.50000	56430.500000	85.053000	3.000000	0.000000	9.240000	7.790000
75%	38467.75000	90008.000000	251.053200	5.000000	0.200000	36.810000	24.450000
max	51290.00000	99301.000000	22638.480000	14.000000	0.850000	8399.976000	933.570000

Each column in the dataset represents different information about the sale

- **Row ID -** It is used to uniquely identify a row in a table.
- **Order ID -** This ID is generated when the order is placed.
- **Order Date -** It shoe the date on when the order is Placed.
- **Ship Date -** The Shipment date of the product.
- **Ship Mode -** In which mode the shipment process is carried out.
- **Customer ID -** ID is generated when the customer Places the first order.
- **Customer Name -** It shows the name of the customer.
- **Segment -** It shows the segment of the customer.
- **City -** The city in which the customer lives.
- **State -** The state in which the customer resides.
- **Country -** It gives the country of the customer.
- **Market -** Market where the order is placed.
- **Region -** It gives the region of the market.
- **Product ID -** This is a unique ID generated for each Product.
- **Category -** It shows which category of the product.
- **Sub-Category -** Sub-Category to which the product Belongs.
- **Product Name -** The name of the product is mentioned.

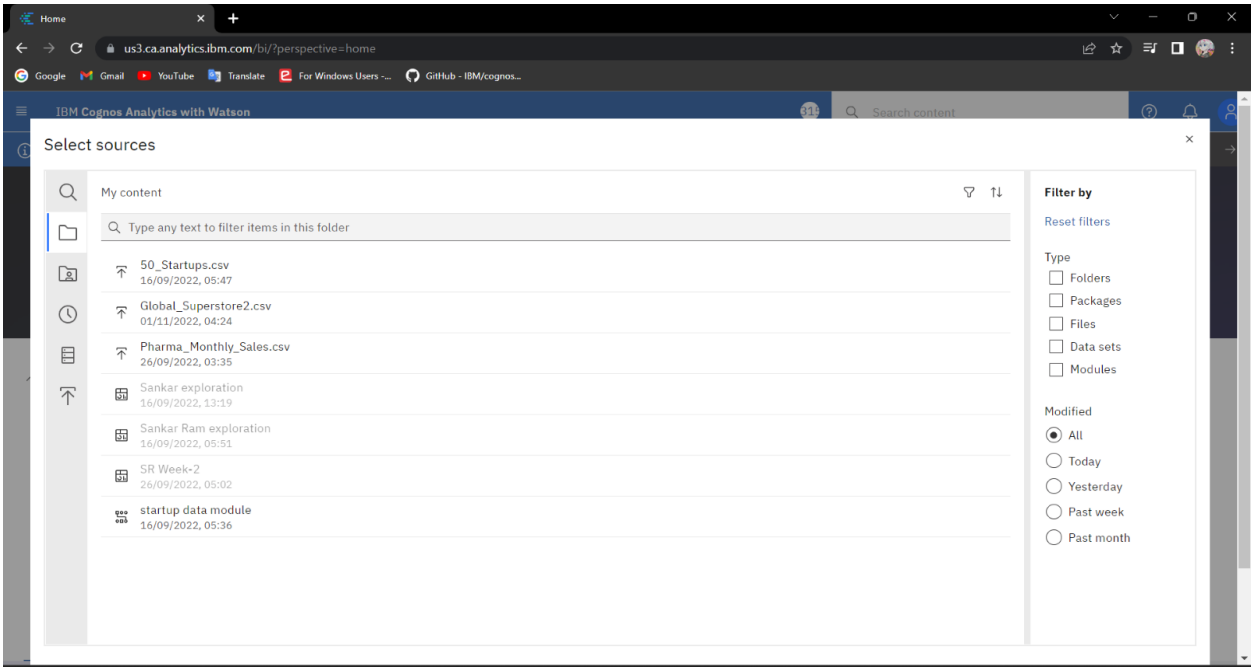
- **Sales -** It shows the sales details of the product.
- **Quantity -** It shows the number of products ordered.
- **Discount -** How much discount is provided.
- **Profit -** The profit earned by the retailer.
- **Shipping cost -** The cost of shipping.
- **Order Priority** - It shows the priority level of the order.

Preparing the dataset: It's the process of cleaning and transforming raw data prior to processing and analysis.

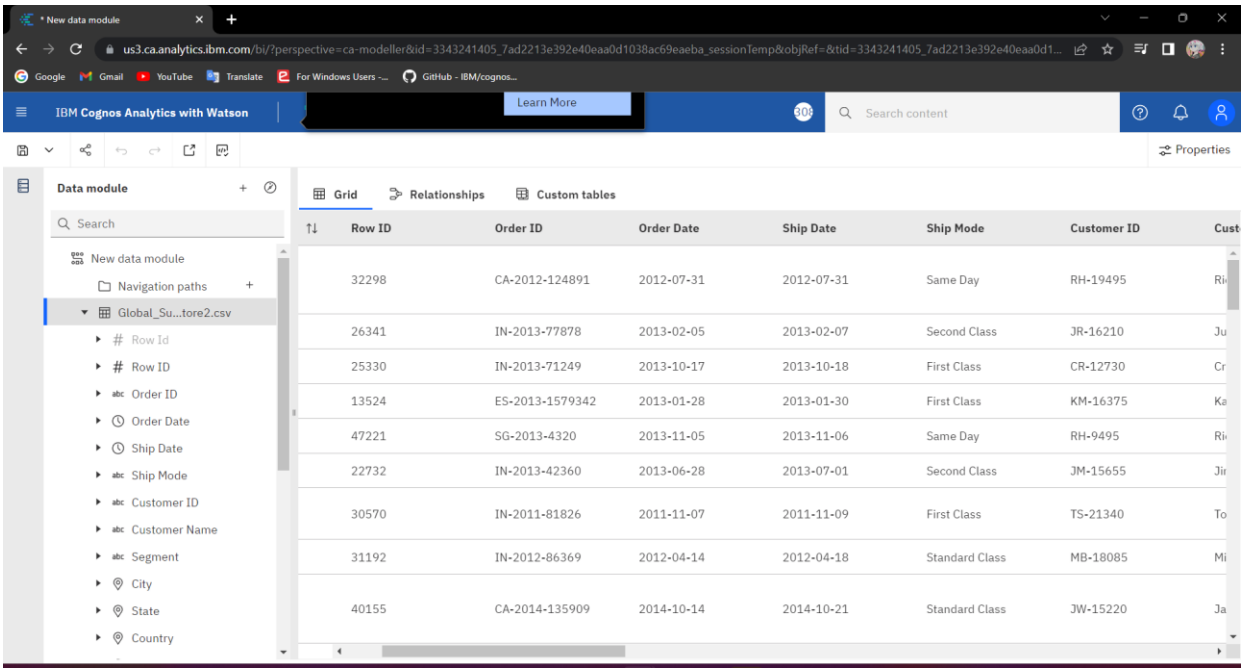
Once we load the data into IBM Cognos, we need to prepare the following:

- Prepare Calculations of Year, Month, Day fields and the related Navigation path.

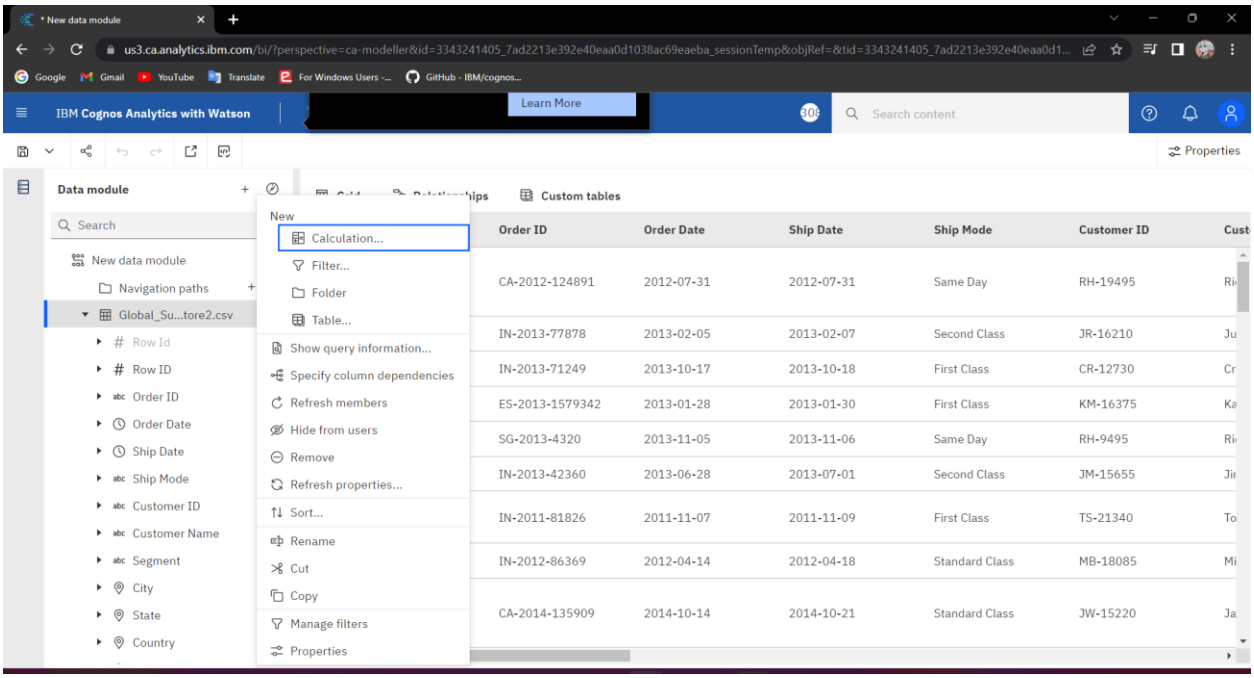
Select the dataset from the prepare data section.



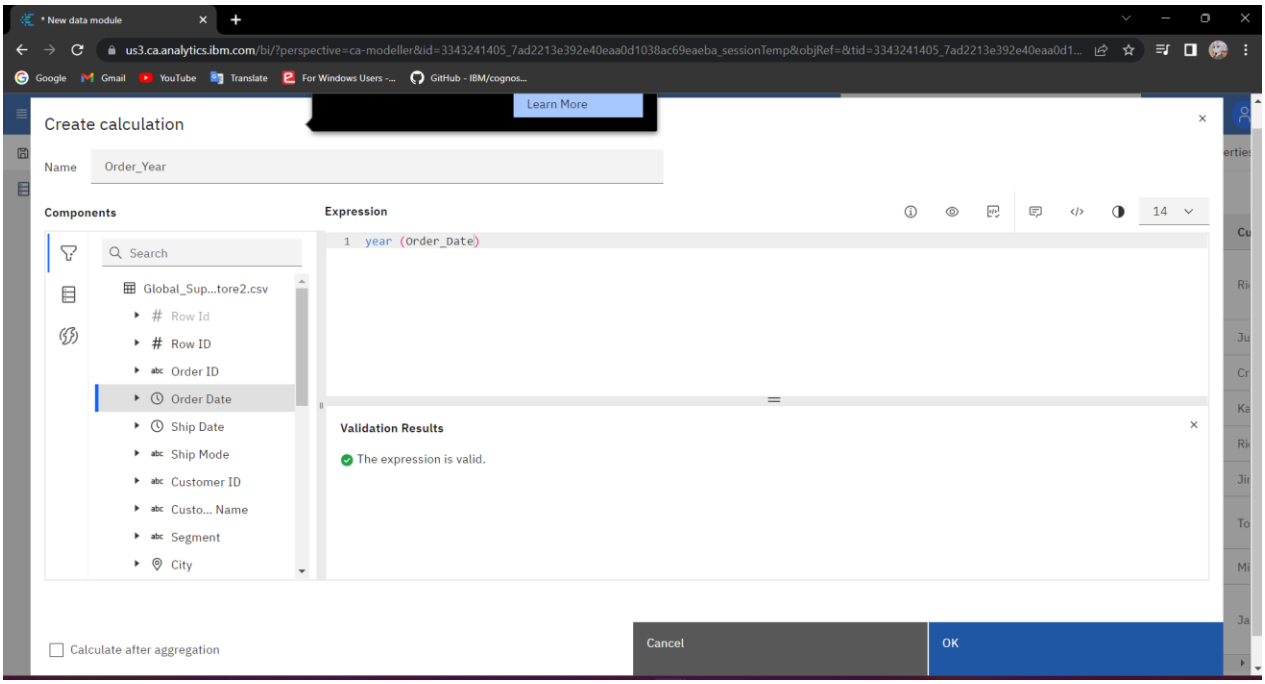
Once the dataset is loaded make the necessary calculations and navigation path



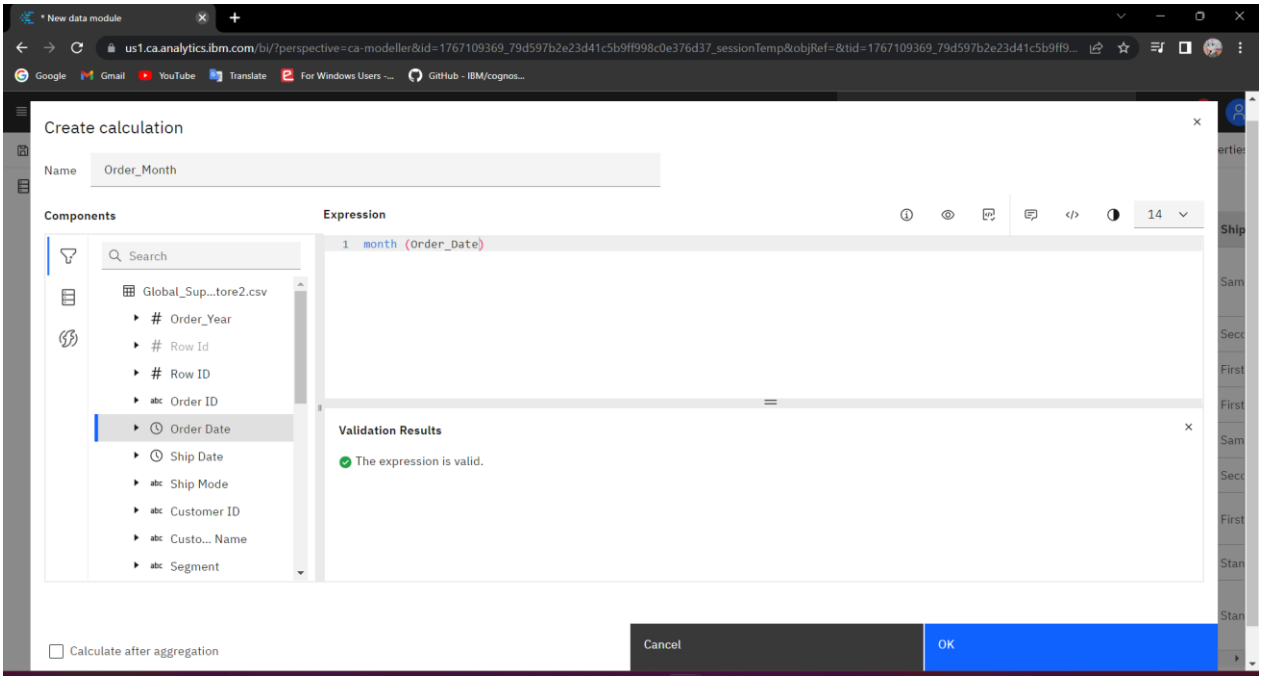
Creating new calculation



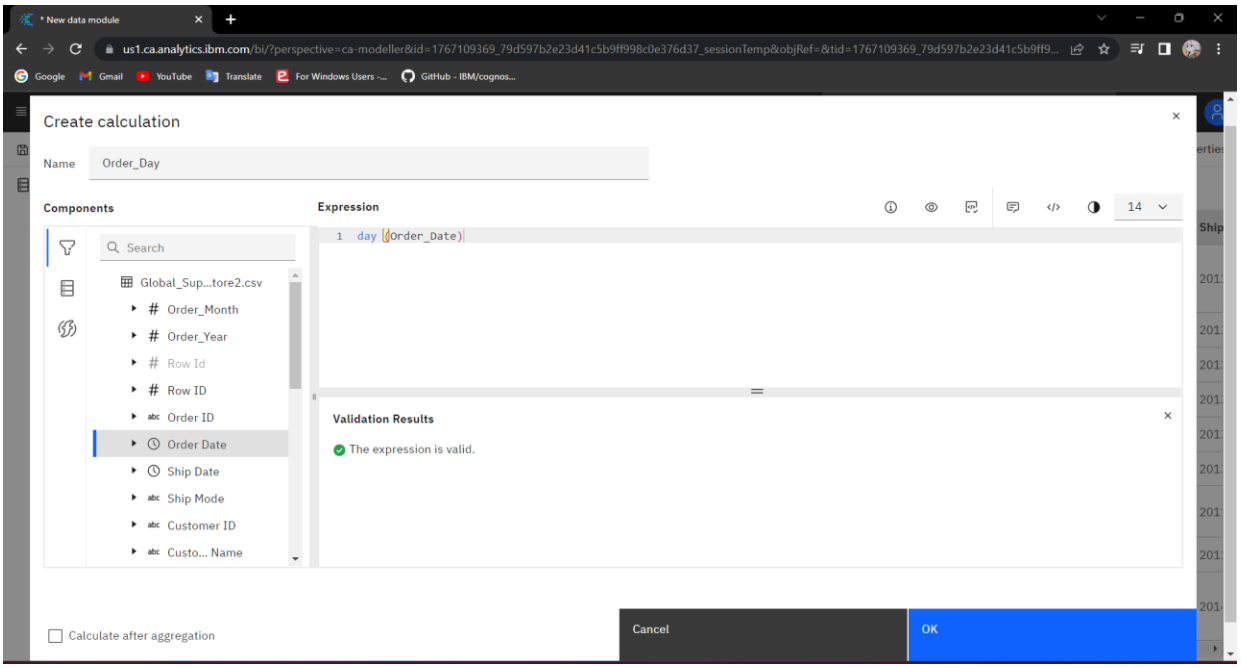
Creating “Order_Year” calculation



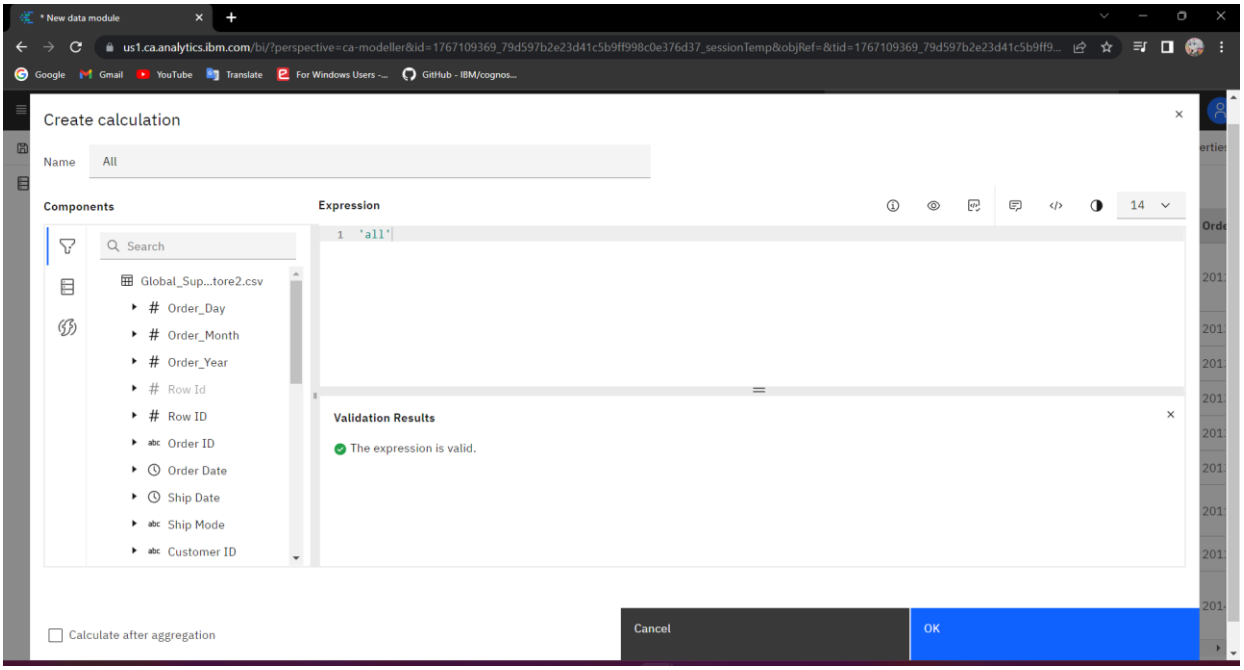
Creating “Order_Month” calculation



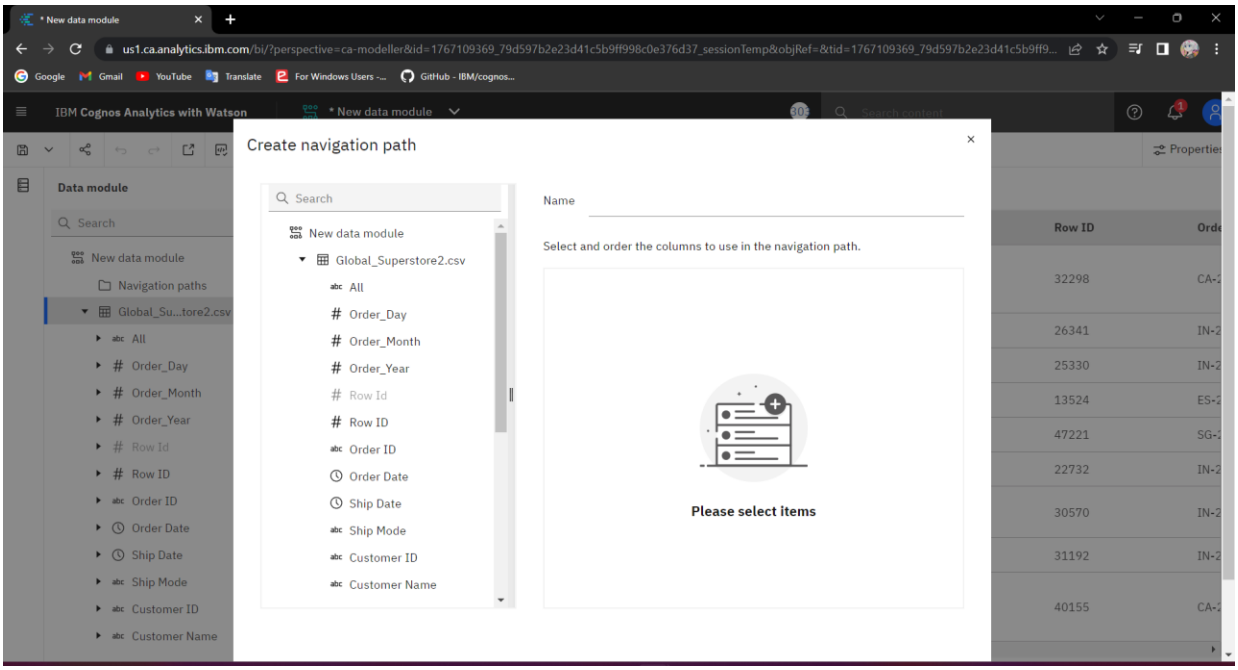
Creating “Order_Day” calculation

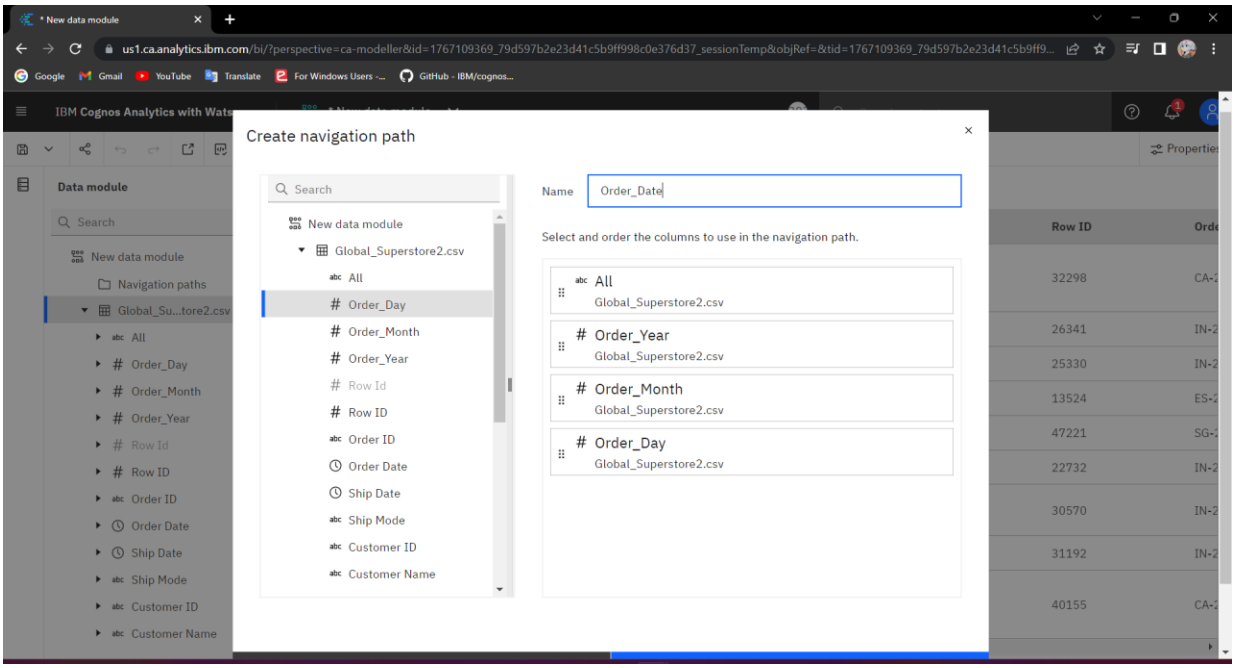


Creating “All” calculation for navigation path



Creating “Order_Date” navigation path

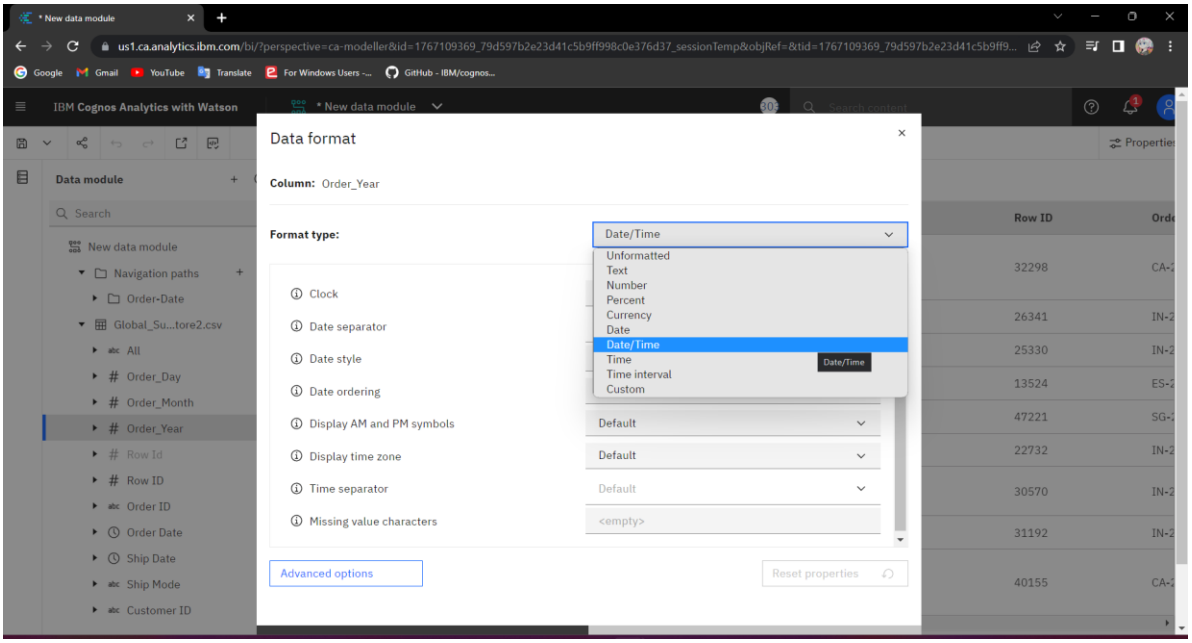




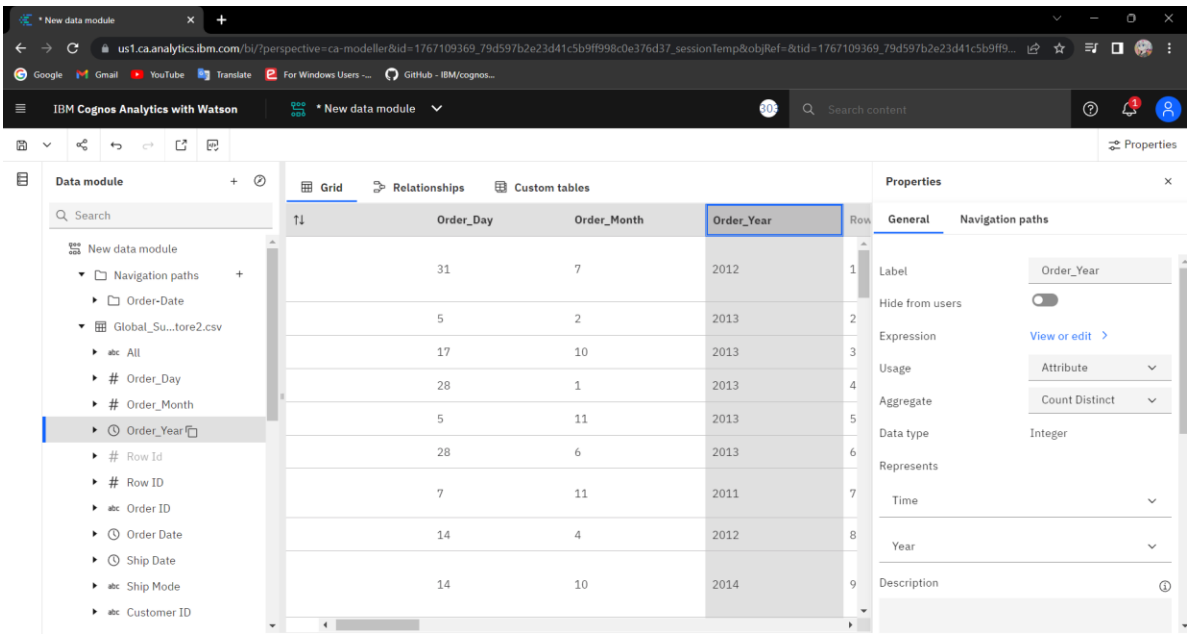
The order year, month and day are in numerical value, so we have to change it to date values.

The following steps are used to change from numerical value to date values

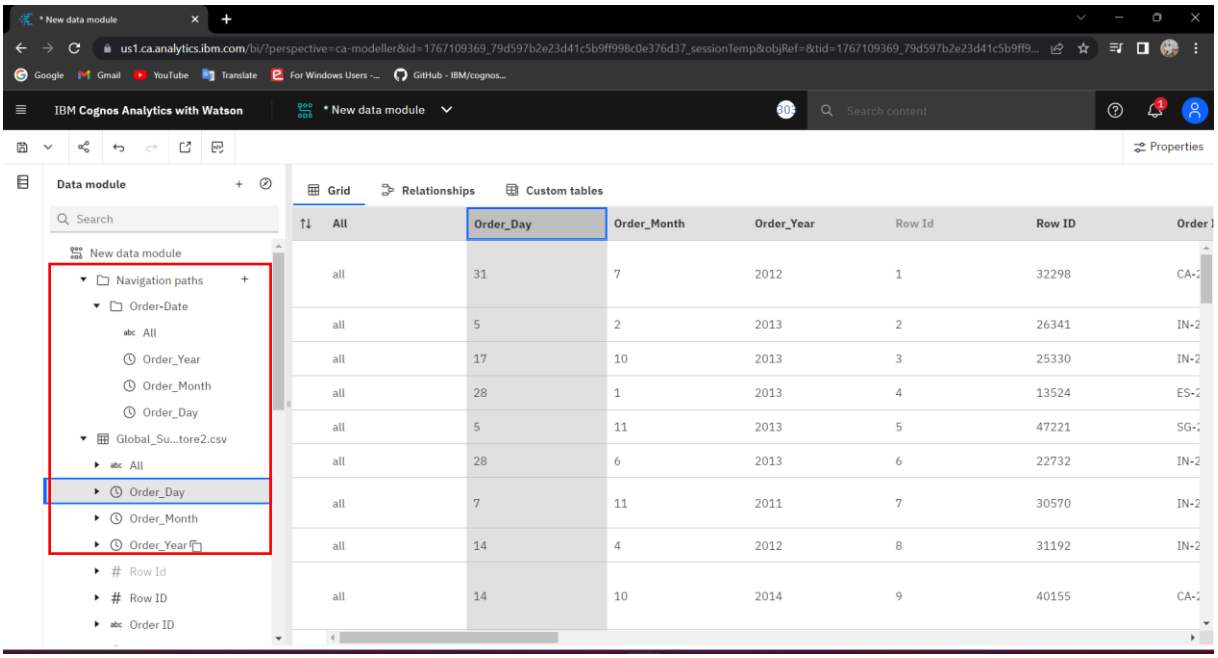
Step 1: Click on **Order_Year** and --> Format data and change the format type to “Date/Time” then click ok.



Step 2: Go to **Order_Year** properties and change the Usage to ‘Attribute’, Aggregate to ‘Count distinct’, Represents to ‘Time’ and change the display option to ‘Show members’.



Repeat steps 1 and 2 for **Order_Month** and **Order_Day** To change it to date values.



Now all the numerical values are changed into date value

Then save the dataset