

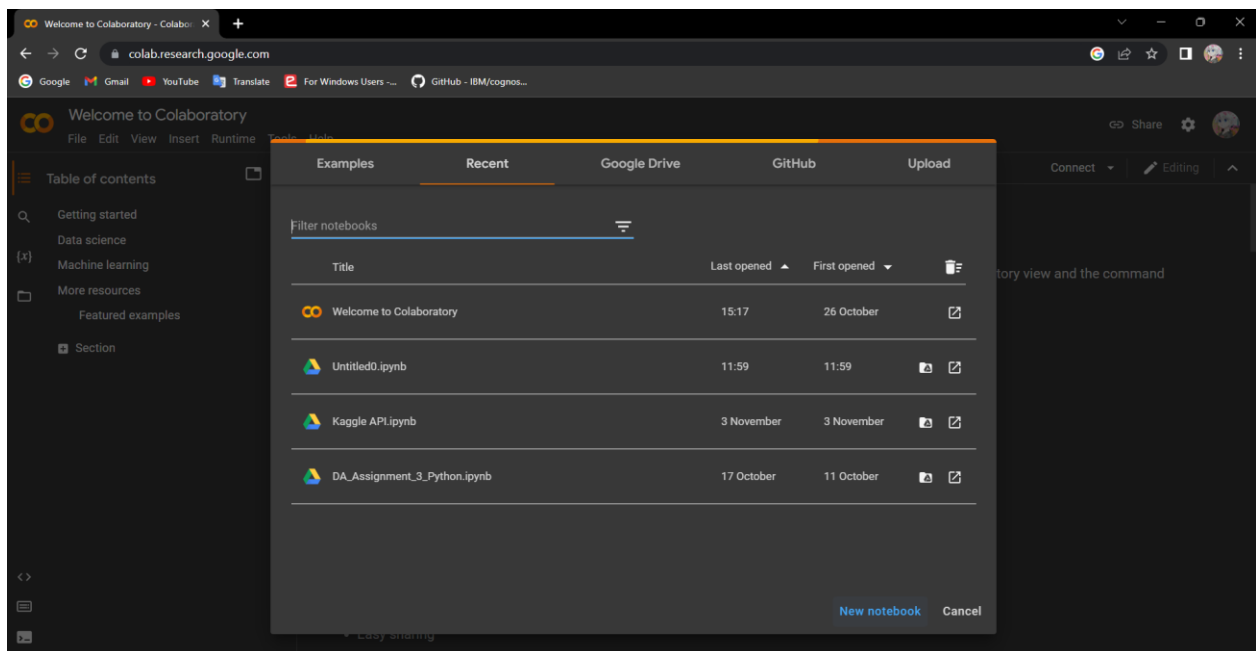
Date	03 November 2022
Team ID	PNT2022TMID30128
Project Name	Project - Global Sales Data Analytics

Sprint 1 - Dataset exploration and preparing the dataset

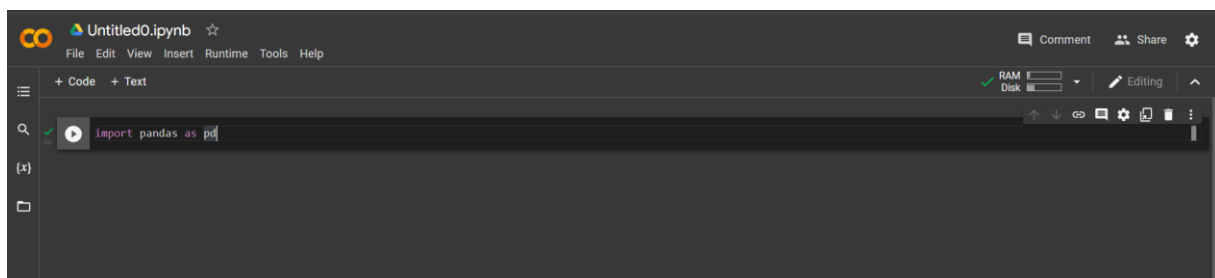
Dataset exploration: It is the process of finding what are the things that are present in the dataset.

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).



The screenshot shows a Jupyter Notebook titled 'Untitled0.ipynb'. The code cell contains the following Python code:

```
[1] import pandas as pd

pd.read_csv('/content/Global_Superstore2.csv', encoding = 'unicode_escape')
```

The output displays the first five rows of the dataset as a table:

	0	1	2	3	4
	32298	26341	25330	13524	47221
CA-2012-124891	31-07-2012	05-02-2013	17-10-2013	28-01-2013	05-11-2013
Same Day	Second Class	First Class	First Class	First Class	Same Day
RH-19495	JR-16210	CR-12730	KM-16375	KM-16375	RH-9495
Rick Hansen	Justin Ritter	Craig Reiter	Katherine Murray	Katherine Murray	Rick Hansen
Consumer	Corporate	Consumer	Home Office	Home Office	Consumer
New York City	Wollongong	Brisbane	Berlin	Berlin	Dakar
New York	New South Wales	Queensland	Berlin	Berlin	Dakar
TEC-AC-10003033	FUR-CH-10003950	TEC-PH-10004664	TEC-PH-10004583	TEC-PH-10004583	TEC-SHA-10000501
Technology	Furniture	Technology	Technology	Technology	Technology
Accessories	Chairs	Phones	Phones	Phones	Copiers
Plantronics CS510 - Over-the-Head monaural Wir...	Novimex Executive Leather Armchair, Black	Nokia Smart Phone, with Caller ID	Motorola Smart Phone, Cordless	Motorola Smart Phone, Cordless	Sharp Wireless Fax, High-
2309.650	3709.395	5175.171	2892.510	2892.510	2832.960
7	9	9	5	5	8

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

The screenshot shows a Jupyter Notebook titled 'Sprint 1 - Code.ipynb'. The code cell contains the following Python code:

```
[9] saledata=pd.read_csv('/content/Global_Superstore2.csv', encoding = 'unicode_escape')
```

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

The screenshot shows a Jupyter Notebook titled 'Untitled0.ipynb'. The code cell contains the following Python code:

```
saledata.shape
```

The output displays the shape of the dataset as a tuple:

```
(51290, 24)
```

Rows = 51290 Columns = 24

2. Finding if there is any duplicate value

The screenshot shows a Jupyter Notebook titled 'Untitled0.ipynb'. The code cell contains the following Python code:

```
[11] saledata.duplicated()
```

The output displays a boolean array indicating whether each row is a duplicate of a previous row. The first five rows are marked as False, followed by an ellipsis, and then rows 51285 through 51289 are also marked as False. The output ends with 'length: 51290, dtype: bool'.

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

Untitled0.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

Finding out how the dataset looks like by Calling head() and tail() it will by default return the first and last 5 rows of the dataset

```
[12] saledata.head()
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	...	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Pr
0	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	...	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the-Head monaural Wir...	2309.650	7	0.0	762.
1	26341	IN-2013-77878	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales	...	FUR-CH-10003950	Furniture	Chairs	Novimex Executive Leather Armchair, Black	3709.395	9	0.1	-288.
2	25330	IN-2013-71249	17-10-2013	18-10-2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland	...	TEC-PH-10004664	Technology	Phones	Nokia Smart Phone, with Caller ID	5175.171	9	0.1	919.
3	13524	ES-2013-1579342	28-01-2013	30-01-2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin	...	TEC-PH-10004583	Technology	Phones	Motorola Smart Phone, Corflless	2892.510	5	0.1	-96.

0s completed at 14:52

First five rows of the dataset

Untitled0.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

```
[13] saledata.tail()
```

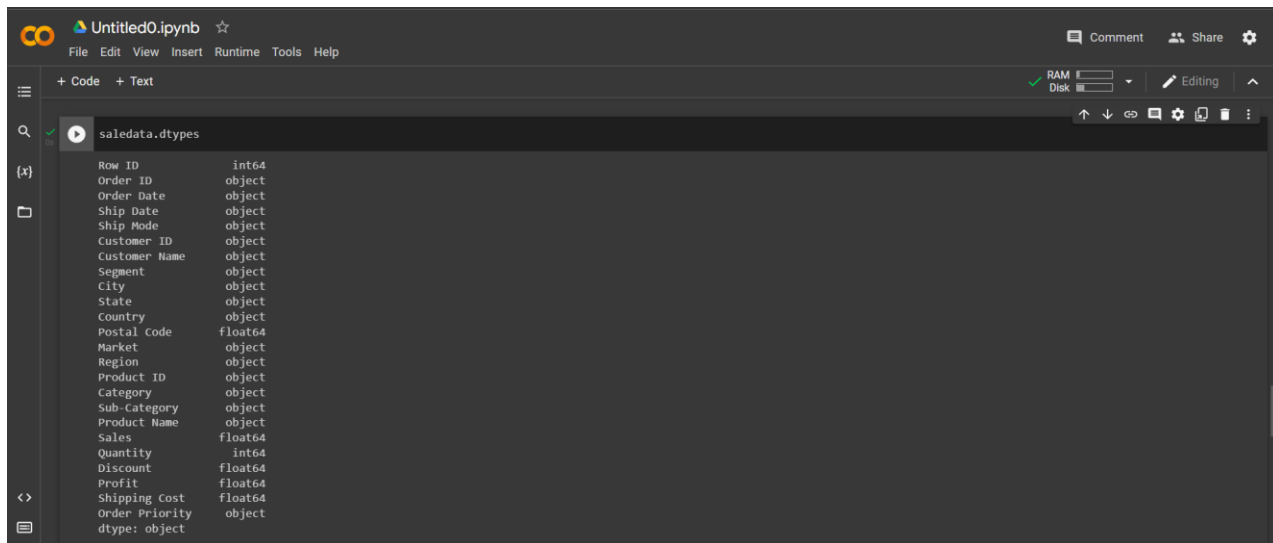
	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	...	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
51285	29002	IN-2014-62366	19-06-2014	19-06-2014	Same Day	KE-16420	Katrina Edelman	Corporate	Kure	Hiroshima	...	OFF-FA-10000746	Office Supplies	Fasteners	Advantus Thumb Tacks, 12 Pack	65.100	5	0.0	4.5000
51286	35398	US-2014-102288	20-06-2014	24-06-2014	Standard Class	ZC-21910	Zuschuss Carroll	Consumer	Houston	Texas	...	OFF-AP-10002906	Office Supplies	Appliances	Hoover Replacement Belt for Commercial Guardsm...	0.444	1	0.8	-1.1100
51287	40470	US-2013-155768	02-12-2013	02-12-2013	Same Day	LB-16795	Laurel Beltran	Home Office	Oxnard	California	...	OFF-EN-10001219	Office Supplies	Envelopes	#10- 4 1/8" x 9 1/2" Security-Tint Envelopes	22.920	3	0.0	11.2308
51288	9596	MX-2012-140767	18-02-2012	22-02-2012	Standard Class	RB-19795	Ross Baird	Home Office	Valinhos	São Paulo	...	OFF-BI-10000806	Office Supplies	Binders	Acco Index Tab, Economy	13.440	2	0.0	2.4000
51289	6147	MX-2012-134460	22-05-2012	26-05-2012	Second Class	MC-18100	Mick Crebagga	Consumer	Tiptapa	Managua	...	OFF-PA-10004155	Office Supplies	Paper	Eaton Computer Printout Paper, 8.5 x 11	61.380	3	0.0	1.8000

5 rows x 20 columns

0s completed at 14:52

Last five rows of the dataset

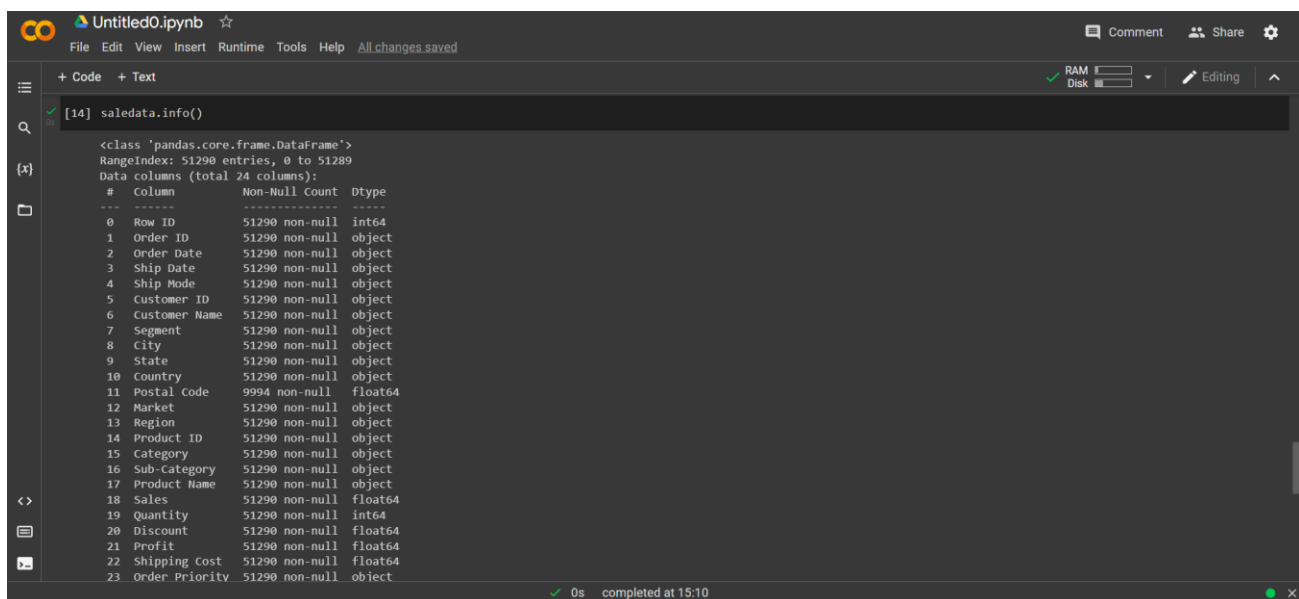
4. What kind of data types are present in the dataset



The screenshot shows a Jupyter Notebook interface with a single code cell. The code cell contains the command `saledata.dtypes`, which has been executed. The output displays the data types for each column in the dataset. The columns and their corresponding data types are as follows:

Column	Data Type
Row ID	int64
Order ID	object
Order Date	object
Ship Date	object
Ship Mode	object
Customer ID	object
Customer Name	object
Segment	object
City	object
State	object
Country	object
Postal Code	float64
Market	object
Region	object
Product ID	object
Category	object
Sub-Category	object
Product Name	object
Sales	float64
Quantity	int64
Discount	float64
Profit	float64
Shipping Cost	float64
Order Priority	object
dtype:	object

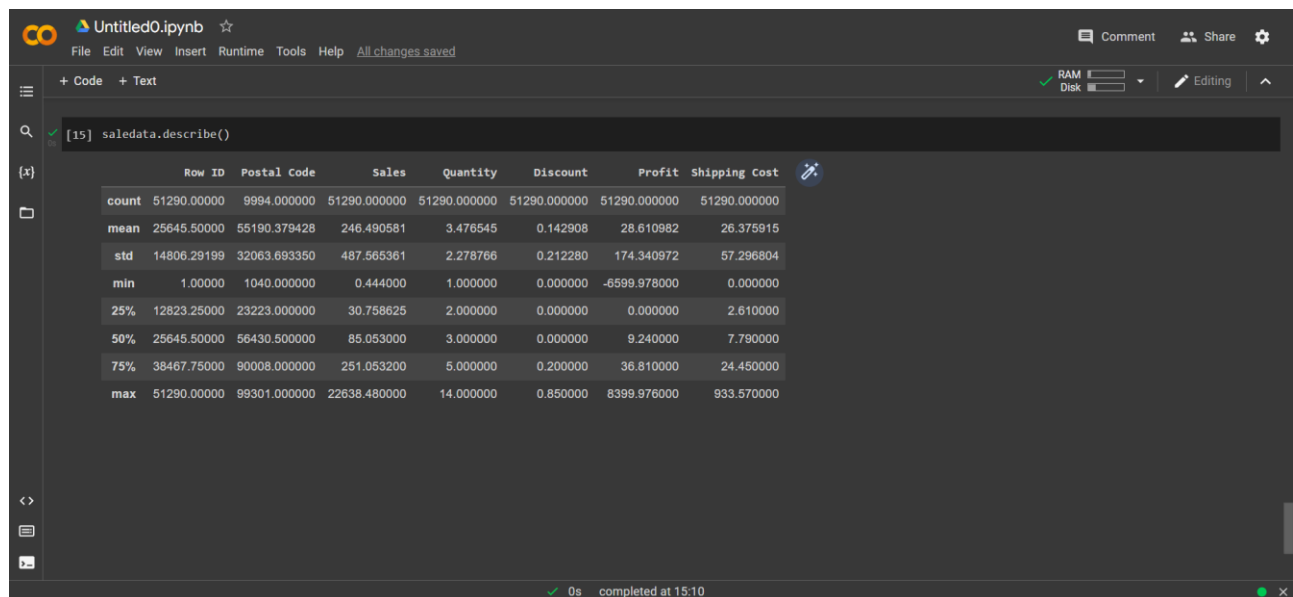
5. Finding out the overview information about the dataset



The screenshot shows a Jupyter Notebook interface with a single code cell. The code cell contains the command `[14] saledata.info()`, which has been executed. The output provides a detailed overview of the dataset, including the number of entries, the number of columns, and the data type for each column. The output is as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   Row ID                51290 non-null int64   
 1   Order ID              51290 non-null object  
 2   Order Date            51290 non-null object  
 3   Ship Date             51290 non-null object  
 4   Ship Mode             51290 non-null object  
 5   Customer ID           51290 non-null object  
 6   Customer Name         51290 non-null object  
 7   Segment              51290 non-null object  
 8   City                  51290 non-null object  
 9   State                 51290 non-null object  
10  Country               51290 non-null object  
11  Postal Code           9994 non-null  float64
12  Market                51290 non-null object  
13  Region                51290 non-null object  
14  Product ID            51290 non-null object  
15  Category              51290 non-null object  
16  Sub-Category          51290 non-null object  
17  Product Name          51290 non-null object  
18  Sales                 51290 non-null float64
19  Quantity              51290 non-null int64   
20  Discount              51290 non-null float64
21  Profit                51290 non-null float64
22  Shipping cost         51290 non-null float64
23  Order Priority         51290 non-null object
```

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



```
[15] salesdata.describe()
```

	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 shows 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.

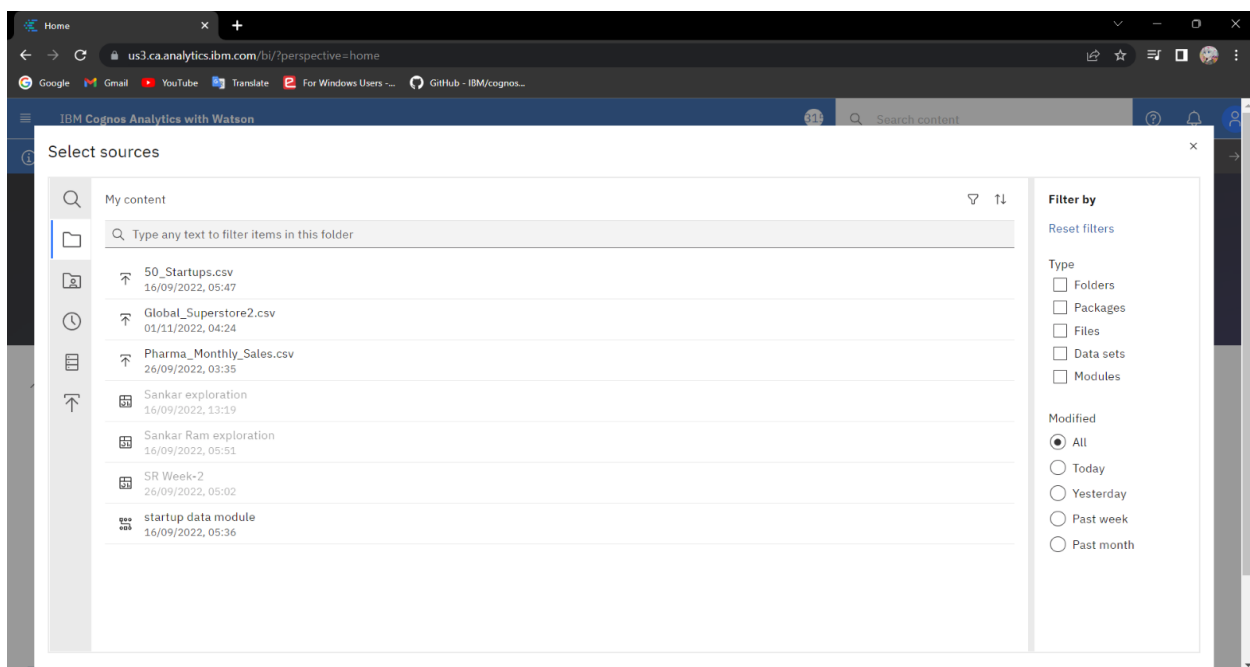
- **Sales** - It shows the sales details of the product.
- **Quantity** - It shows the number of products ordered.
- **Discount** - How much discount is provided.
- **Profit** - The profit earned by the retailer.
- **Shipping cost** - The cost of shipping.
- **Order Priority** - It shows the priority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

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

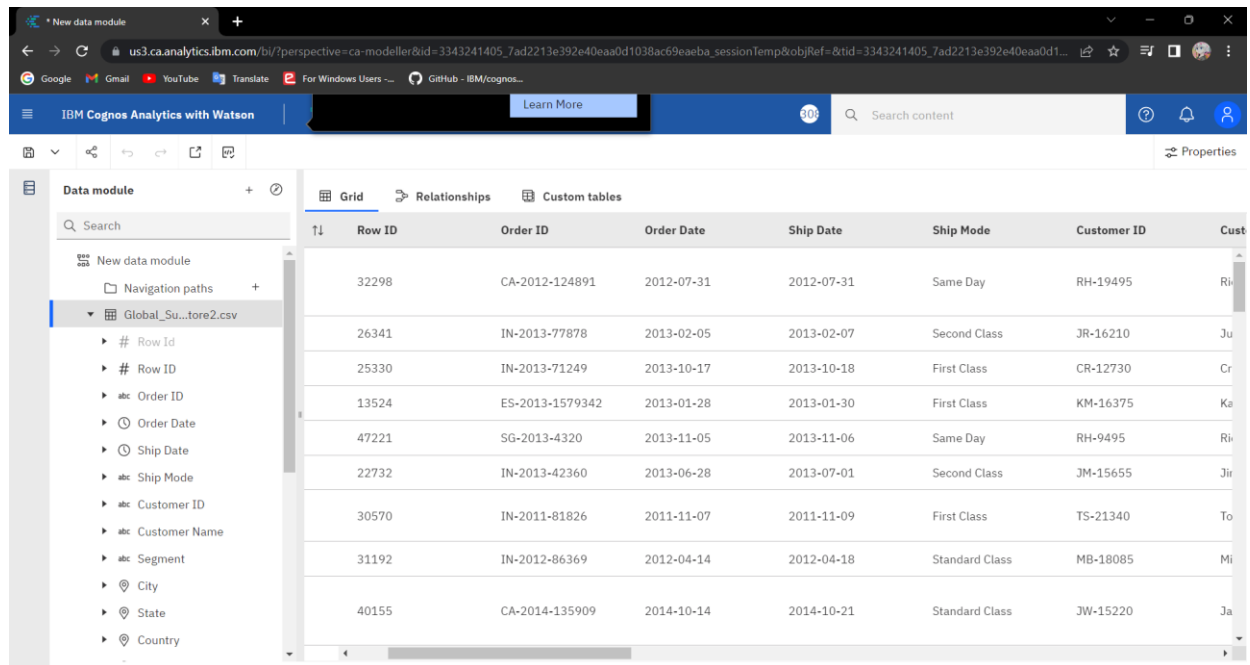
Once we load the data into IBM Cognos, we need to prepare the following:

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

Select the dataset from the prepare data section.



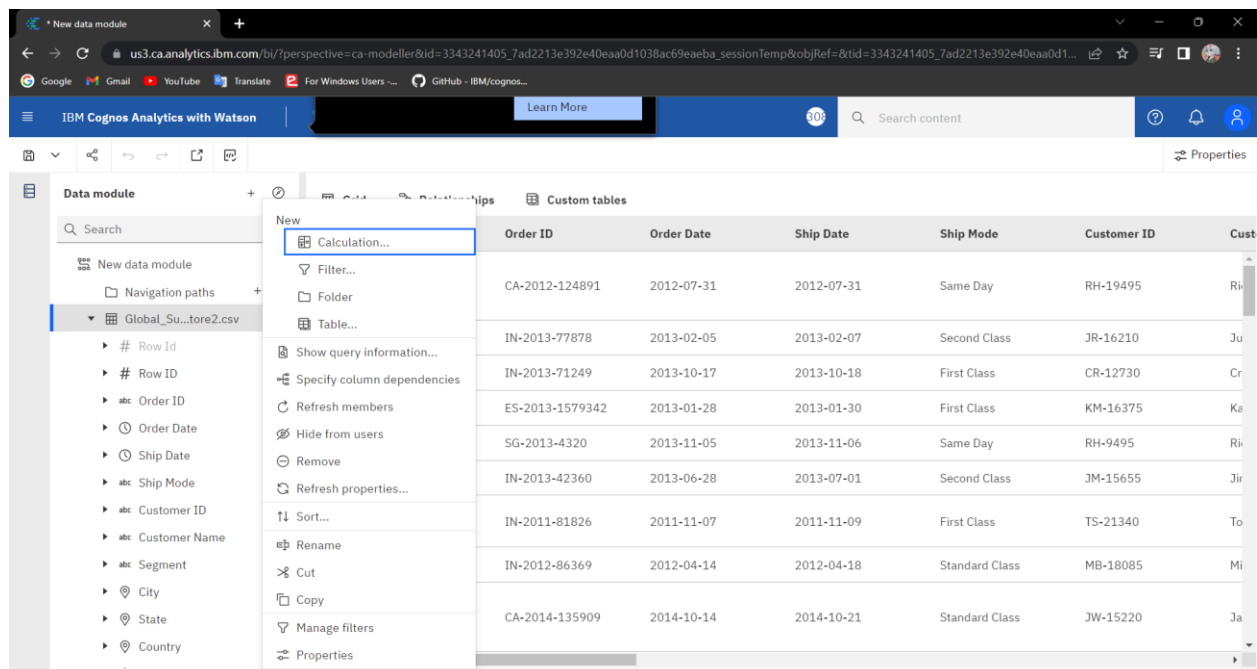
Once the dataset is loaded make the necessary calculations and navigation path



The screenshot shows the IBM Cognos Analytics interface. On the left, the 'Data module' pane displays a tree view with 'Global_Su...tore2.csv' selected. The main area shows a grid view of the dataset with columns: Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, and Customer Name. The data is sorted by Row ID in descending order.

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name
32298	CA-2012-124891	2012-07-31	2012-07-31	Same Day	RH-19495	Ri...
26341	IN-2013-77878	2013-02-05	2013-02-07	Second Class	JR-16210	Ju...
25330	IN-2013-71249	2013-10-17	2013-10-18	First Class	CR-12730	Cr...
13524	ES-2013-1579342	2013-01-28	2013-01-30	First Class	KM-16375	Ka...
47221	SG-2013-4320	2013-11-05	2013-11-06	Same Day	RH-9495	Ri...
22732	IN-2013-42360	2013-06-28	2013-07-01	Second Class	JM-15655	Jir...
30570	IN-2011-81826	2011-11-07	2011-11-09	First Class	TS-21340	To...
31192	IN-2012-86369	2012-04-14	2012-04-18	Standard Class	MB-18085	Mi...
40155	CA-2014-135909	2014-10-14	2014-10-21	Standard Class	JW-15220	Ja...

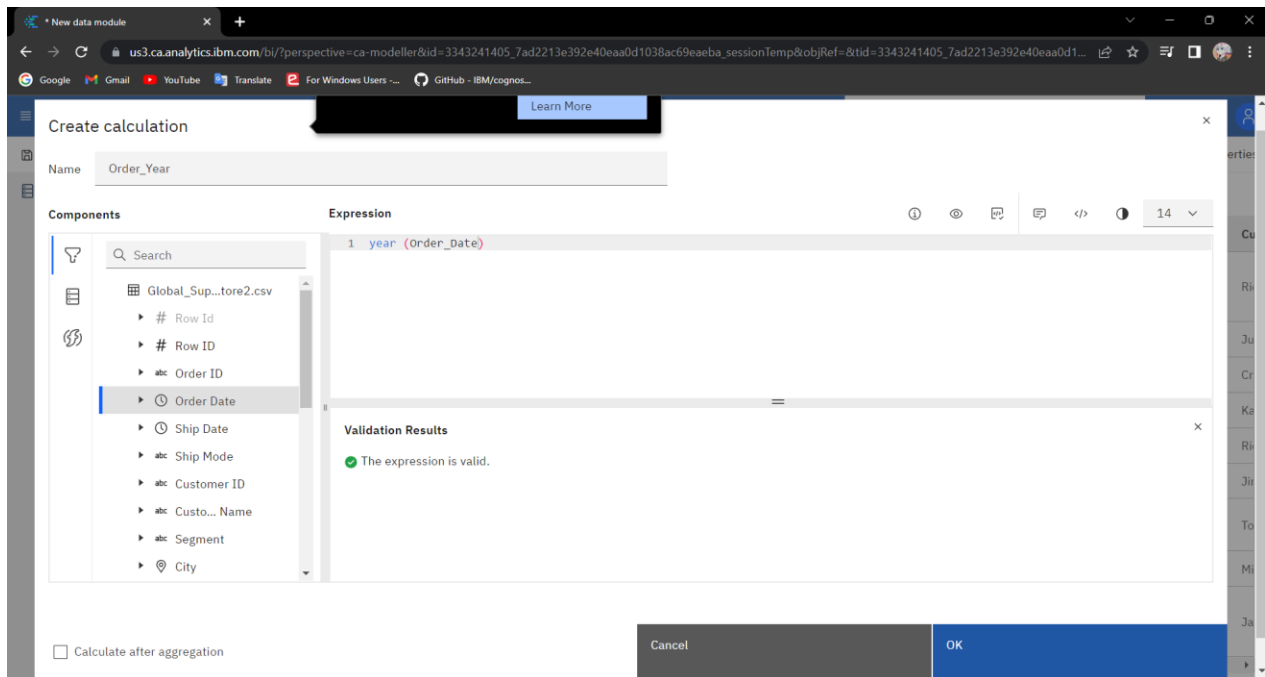
Creating new calculation



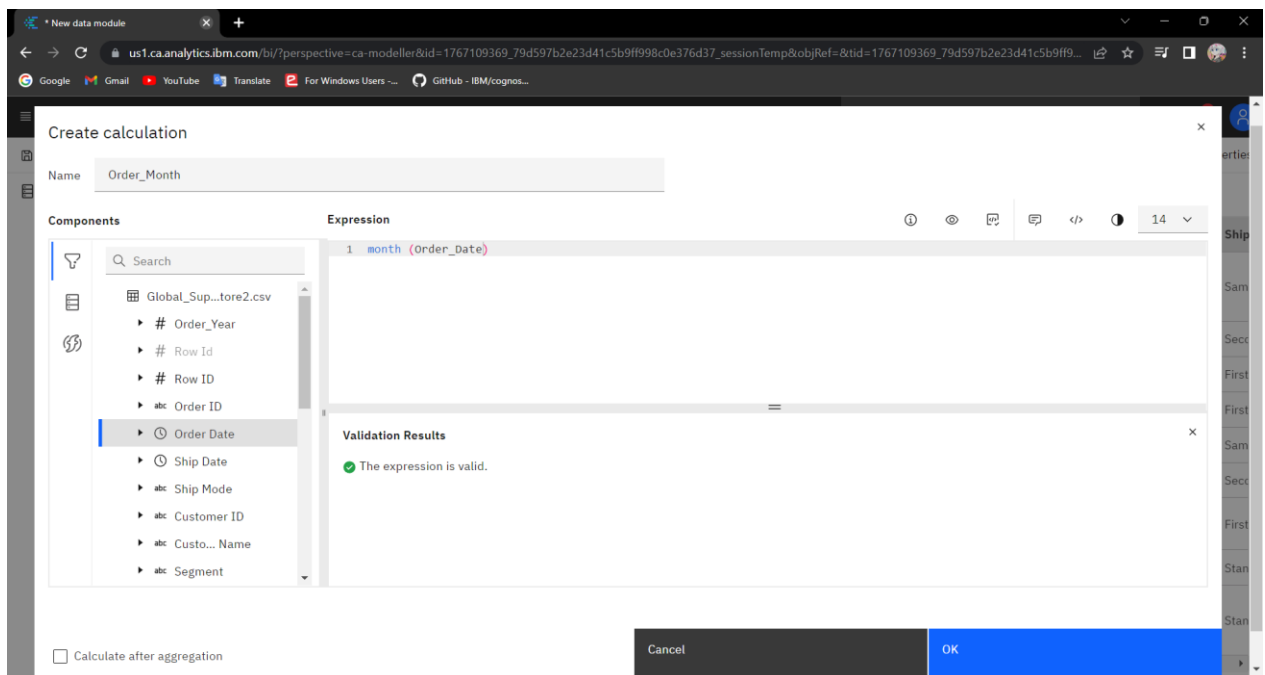
The screenshot shows the IBM Cognos Analytics interface with the 'New' menu open. The 'Calculation...' option is highlighted. The background shows the same data module grid view as the previous screenshot.

Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name
CA-2012-124891	2012-07-31	2012-07-31	Same Day	RH-19495	Ri...
IN-2013-77878	2013-02-05	2013-02-07	Second Class	JR-16210	Ju...
IN-2013-71249	2013-10-17	2013-10-18	First Class	CR-12730	Cr...
ES-2013-1579342	2013-01-28	2013-01-30	First Class	KM-16375	Ka...
SG-2013-4320	2013-11-05	2013-11-06	Same Day	RH-9495	Ri...
IN-2013-42360	2013-06-28	2013-07-01	Second Class	JM-15655	Jir...
IN-2011-81826	2011-11-07	2011-11-09	First Class	TS-21340	To...
IN-2012-86369	2012-04-14	2012-04-18	Standard Class	MB-18085	Mi...
CA-2014-135909	2014-10-14	2014-10-21	Standard Class	JW-15220	Ja...

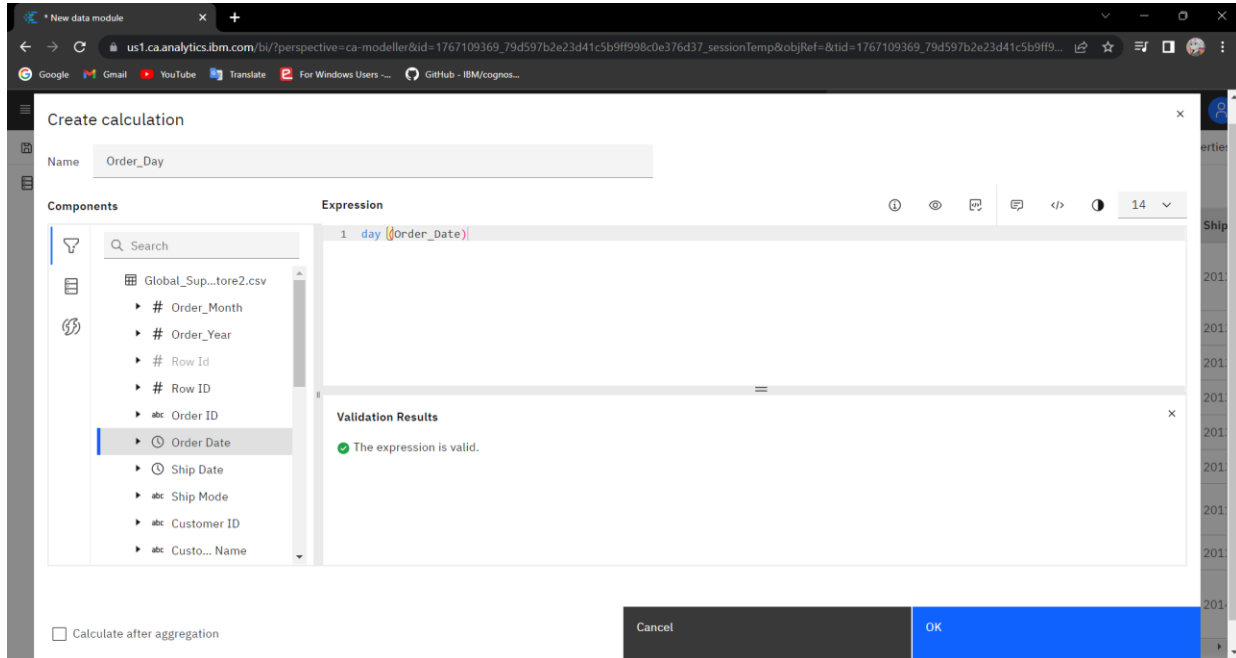
Creating “Order_Year” calculation



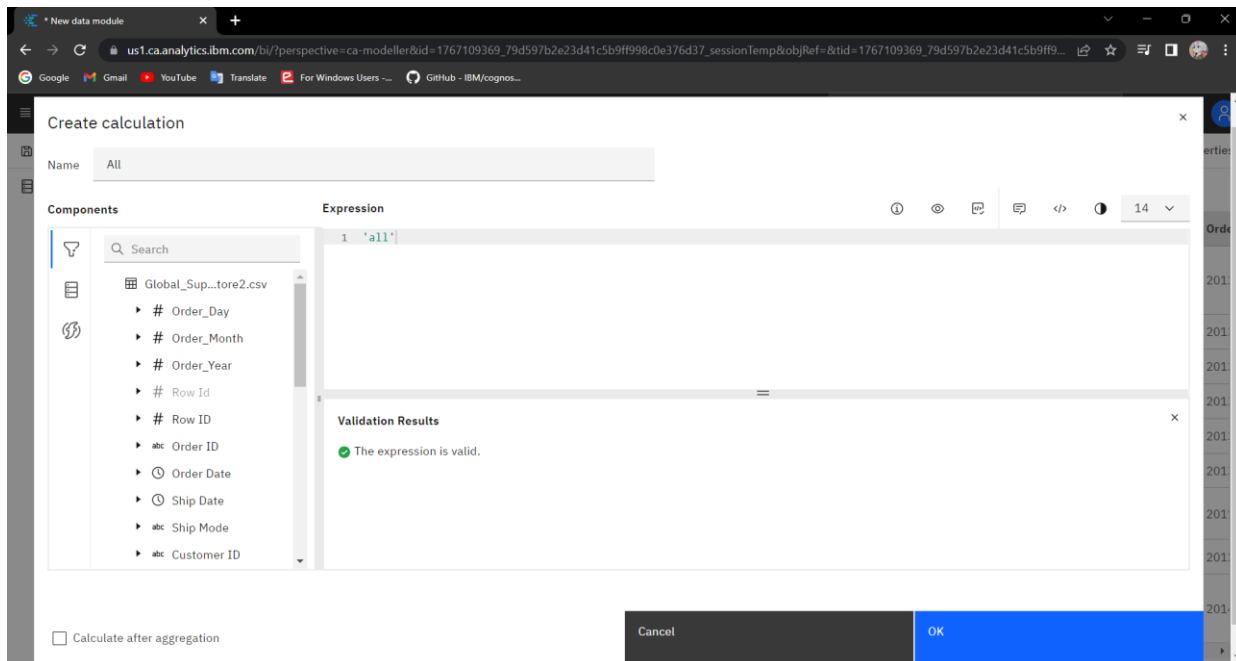
Creating “Order_Month” calculation



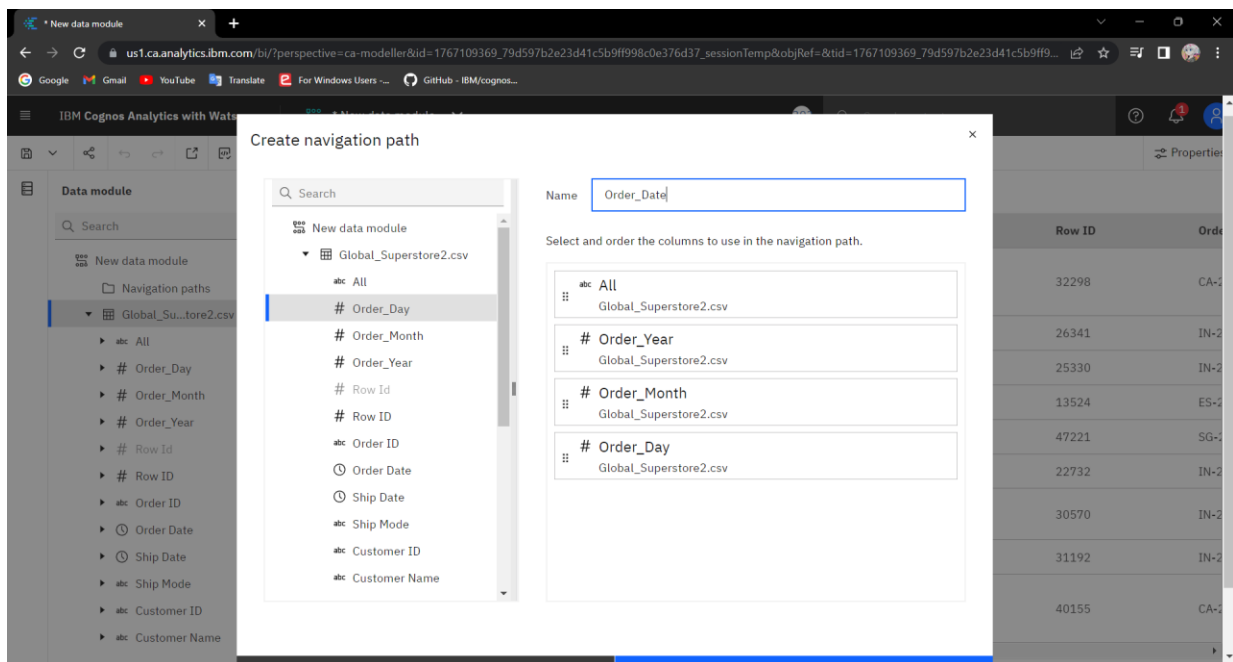
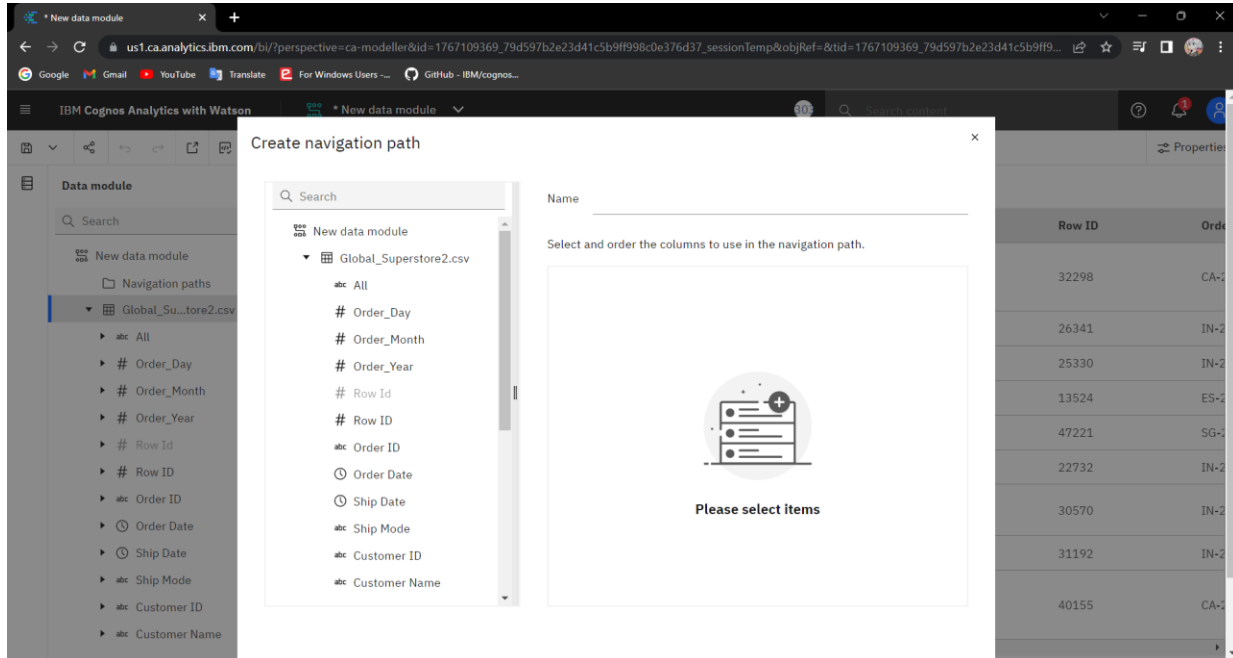
Creating “Order_Day” calculation



Creating “All” calculation for navigation path



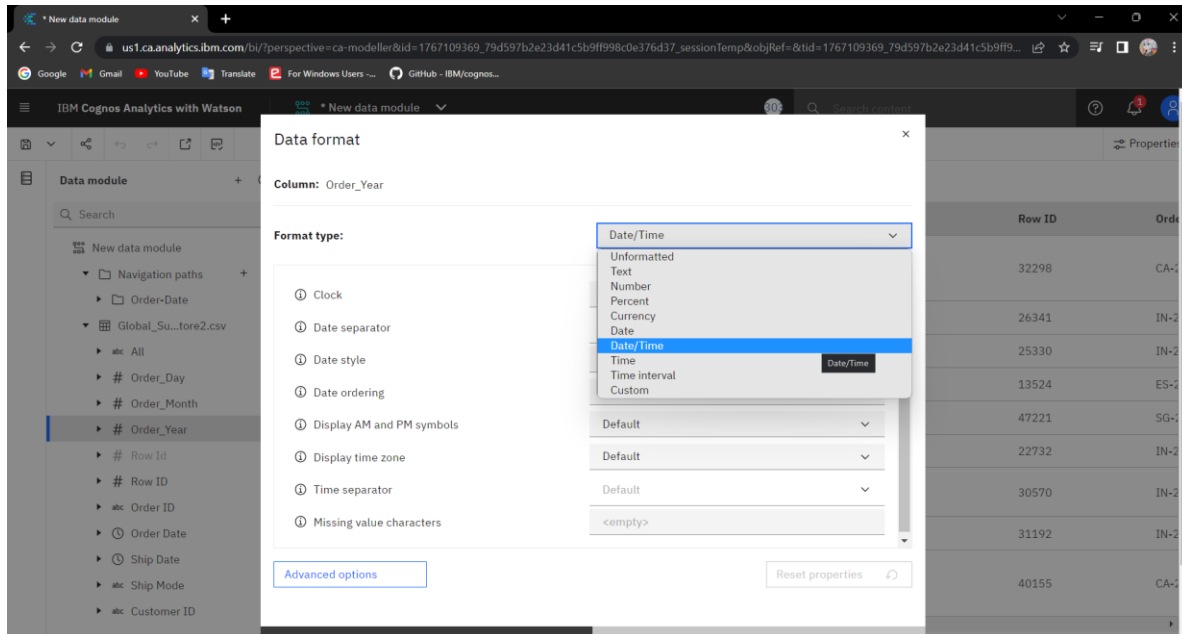
Creating “Order_Date” navigation path



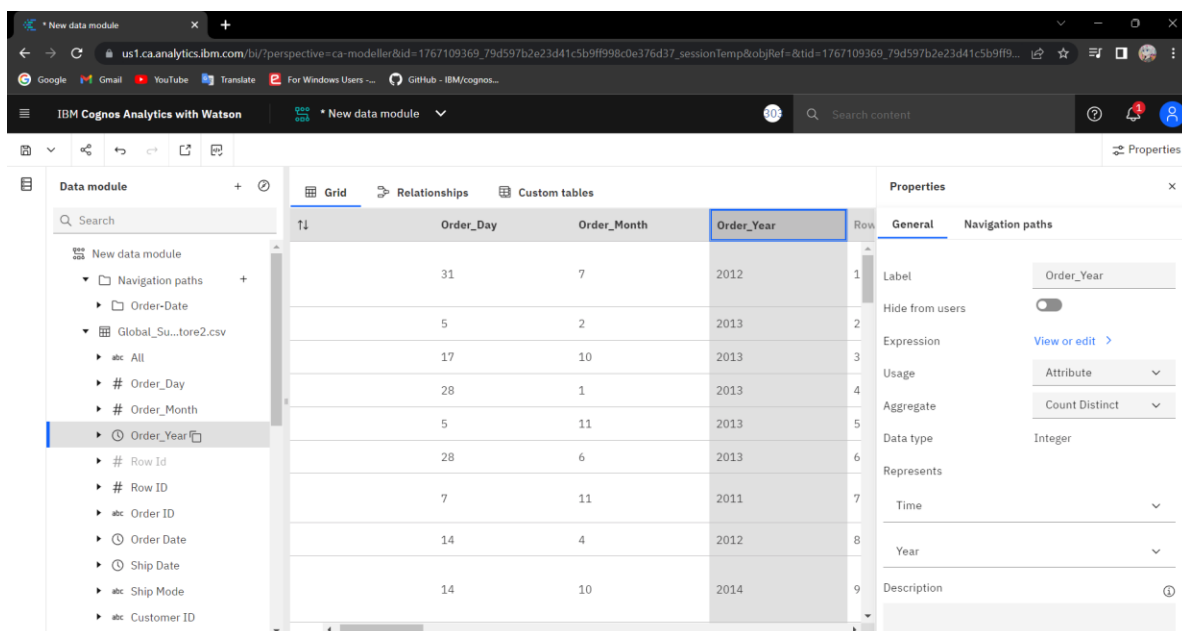
The order year, month and day are in numerical value, so we have to change it to date values.

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

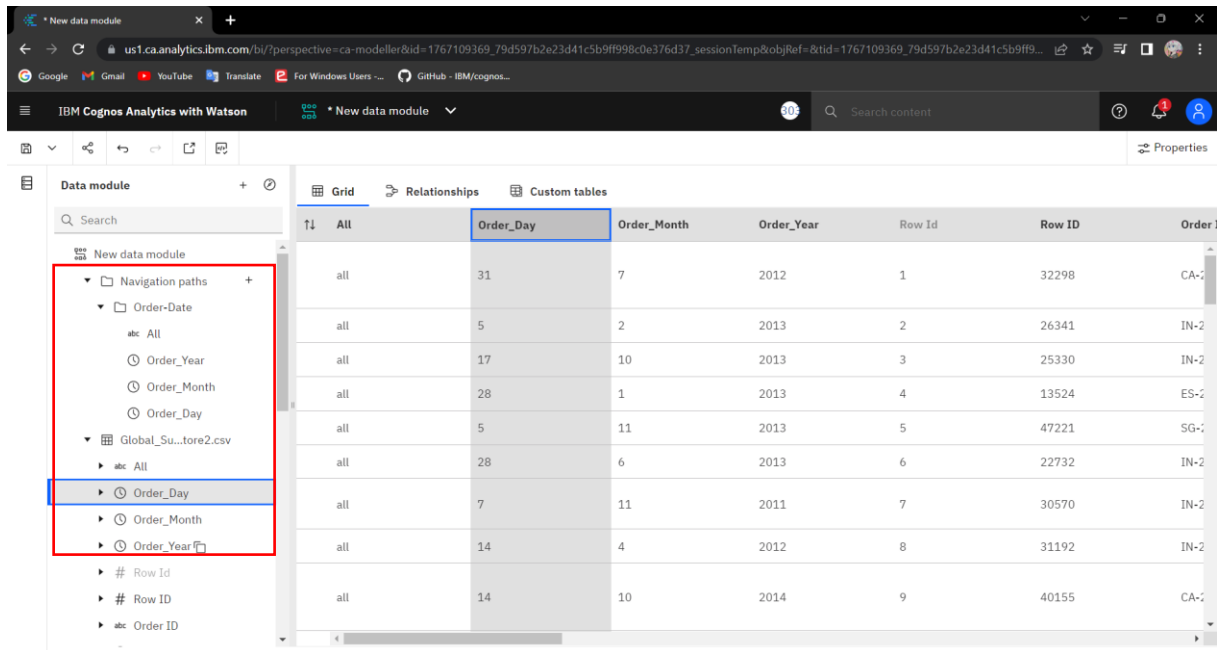
Step 1: Click on **Order_Year** and --> Format data and change the format type to “Date/Time” then click ok.



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.

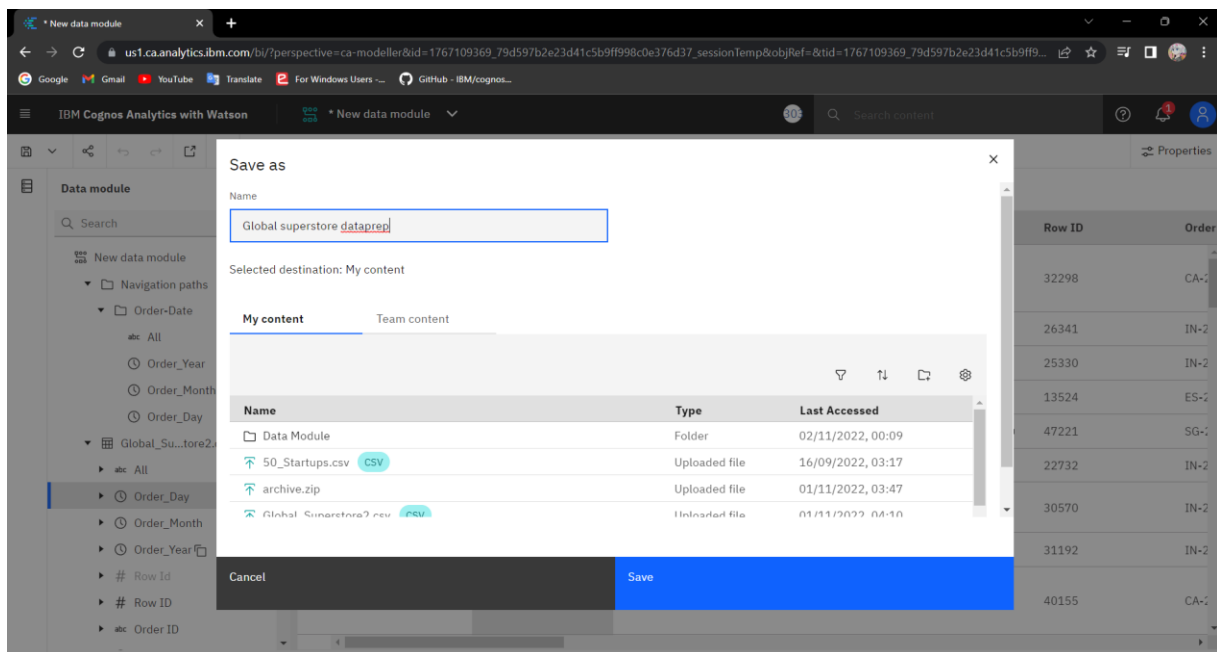


The screenshot shows the IBM Cognos Analytics interface. On the left, a navigation pane is open, showing a tree structure under 'New data module'. The 'Order-Date' folder is expanded, and 'Order_Day' is selected. The main area displays a data table with columns: Order_Day, Order_Month, Order_Year, Row Id, Row ID, and Order ID. The data rows show numerical values for Order_Day, Order_Month, and Order_Year.

Order_Day	Order_Month	Order_Year	Row Id	Row ID	Order ID
31	7	2012	1	32298	CA-2
5	2	2013	2	26341	IN-2
17	10	2013	3	25330	IN-2
28	1	2013	4	13524	ES-2
5	11	2013	5	47221	SG-2
28	6	2013	6	22732	IN-2
7	11	2011	7	30570	IN-2
14	4	2012	8	31192	IN-2
14	10	2014	9	40155	CA-2

Now all the numerical values are changed into date value

Then save the dataset



The screenshot shows the 'Save as' dialog box in IBM Cognos Analytics. The 'Name' field contains 'Global superstore dataprep'. The 'Selected destination' is 'My content'. The 'My content' tab is active, showing a list of existing data modules and files.

Name	Type	Last Accessed
Data Module	Folder	02/11/2022, 00:09
50_Startups.csv	Uploaded file	16/09/2022, 03:17
archive.zip	Uploaded file	01/11/2022, 03:47
Global Superstore new	Uploaded file	01/11/2022, 04:10