SCIENTIFIC DATA VISUALIZATION ACTIVITY-4

Name: Nikhitha Rayabarapu **Date: 09/23/2024**

Question 1:

1. Rename the store_nbr: I have renamed it to ‘nbr_str’ as shown below.

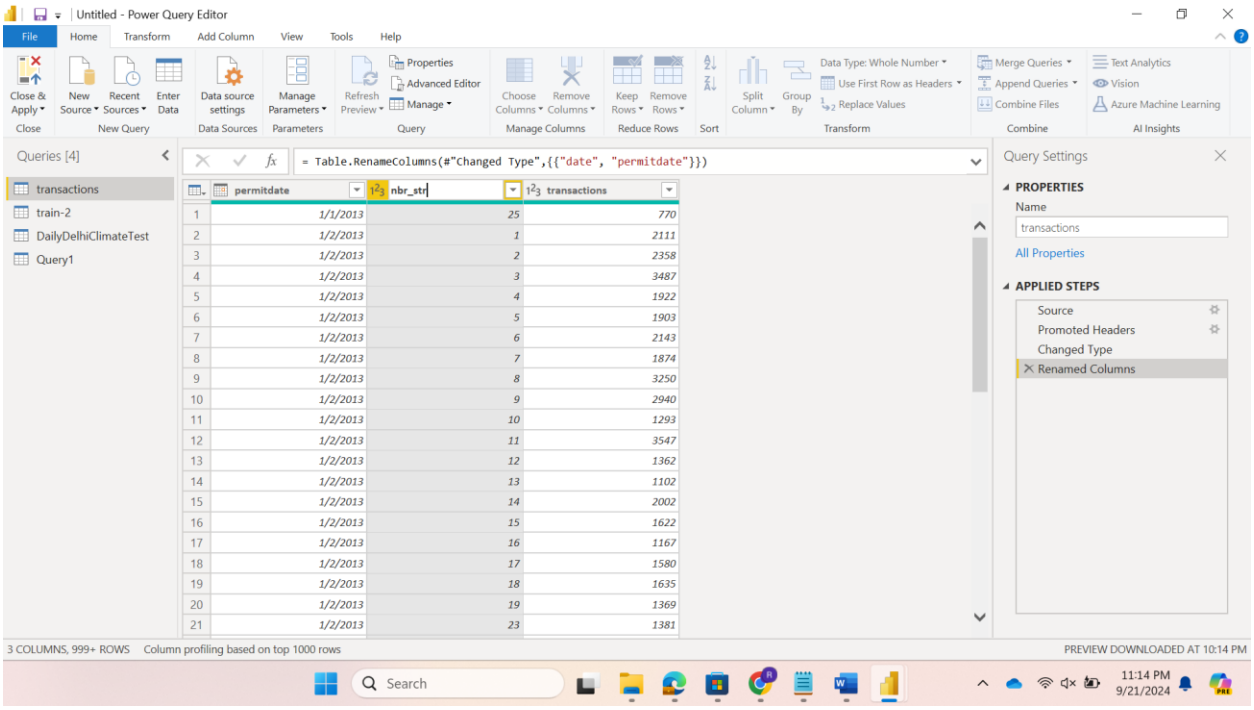

Fig 1.1

2. I have clicked on ‘Add Column’ on top bar and created the ‘Group By’ column with the formula as ‘=1’(fig 1.2). Then on right-clicking the column(fig 1.4), I have selected Group By and created Mean and Median functions for humidity attribute. The steps are as follows:

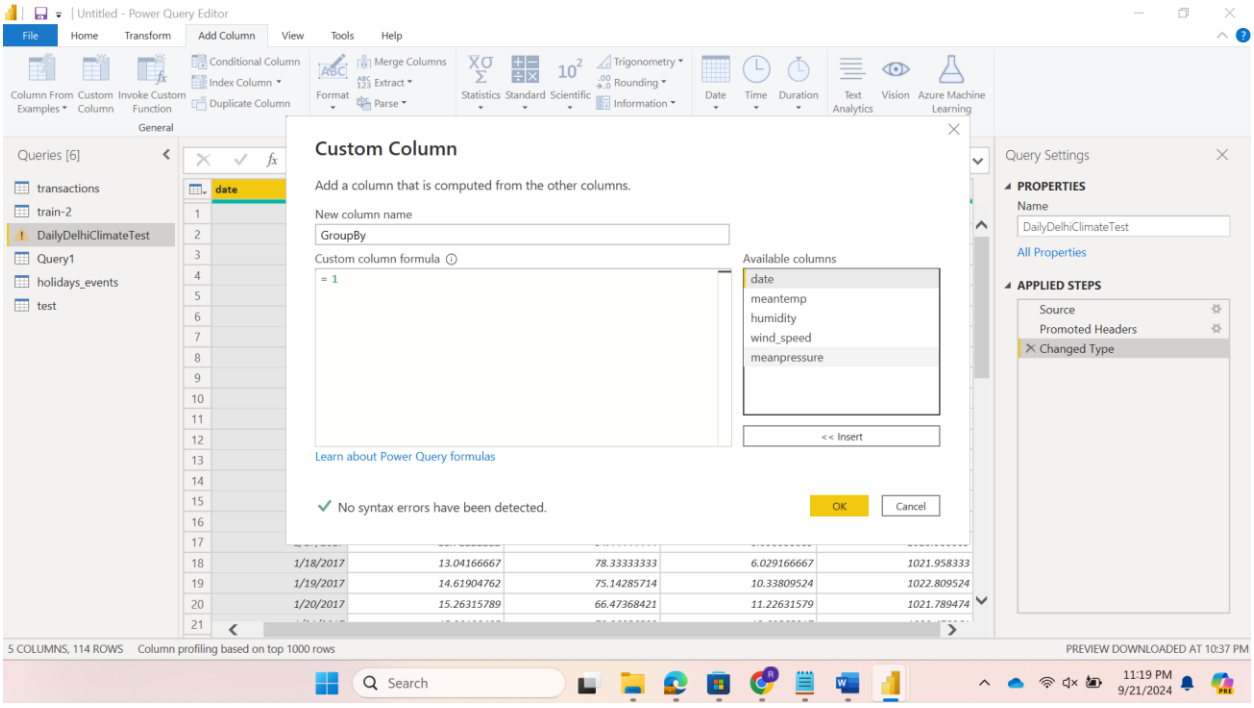
Untitled - Power Query Editor
Custom Column
Add a column that is computed from the other columns.
New column name
GroupBy
Custom column formula
= 1
Available columns
date
meantemp
humidity
wind_speed
meanpressure
<< Insert
Learn about Power Query formulas
No syntax errors have been detected.
OK
Cancel
Queries [6]
transactions
train-2
DailyDelhiClimateTest
Query1
holidays_events
test
Query Settings
PROPERTIES
Name
DailyDelhiClimateTest
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
5 COLUMNS, 114 ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 10:37 PM

Fig 1.2

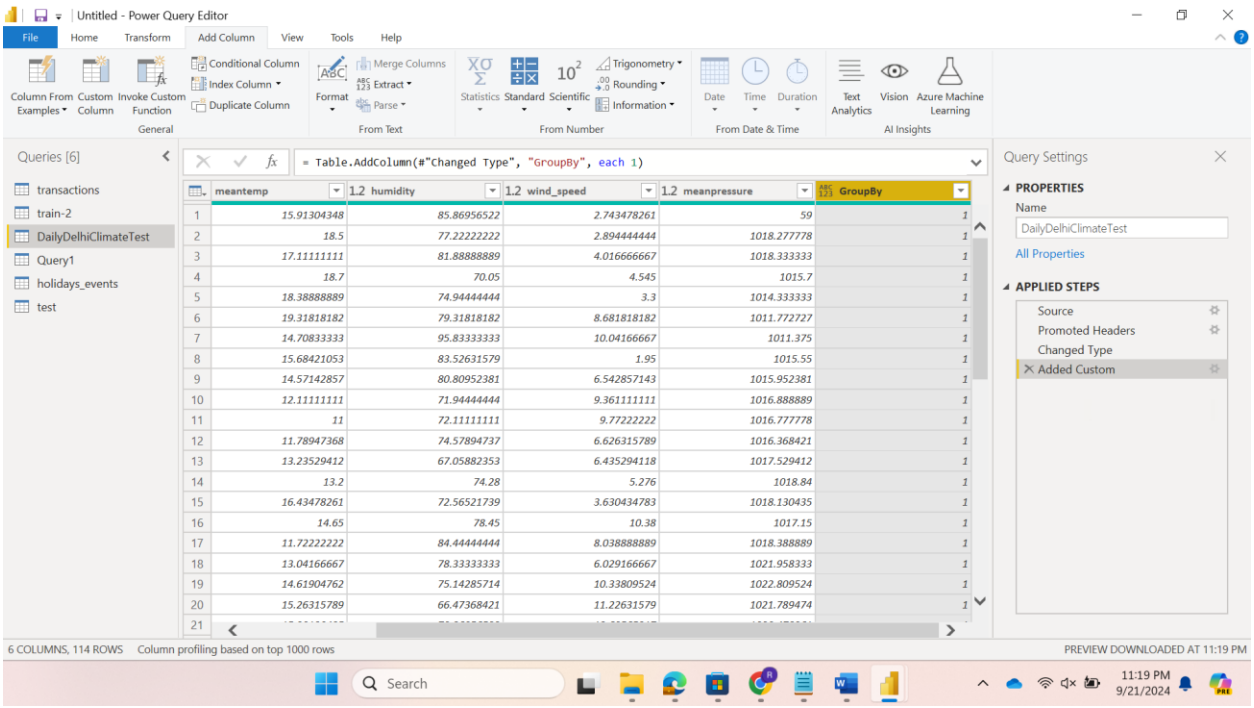
Untitled - Power Query Editor
= Table.AddColumn(#"Changed Type", "GroupBy", each 1)
meantemp
humidity
wind_speed
meanpressure
GroupBy
Queries [6]
transactions
train-2
DailyDelhiClimateTest
Query1
holidays_events
test
Query Settings
PROPERTIES
Name
DailyDelhiClimateTest
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
Added Custom
6 COLUMNS, 114 ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 11:19 PM

Fig 1.3

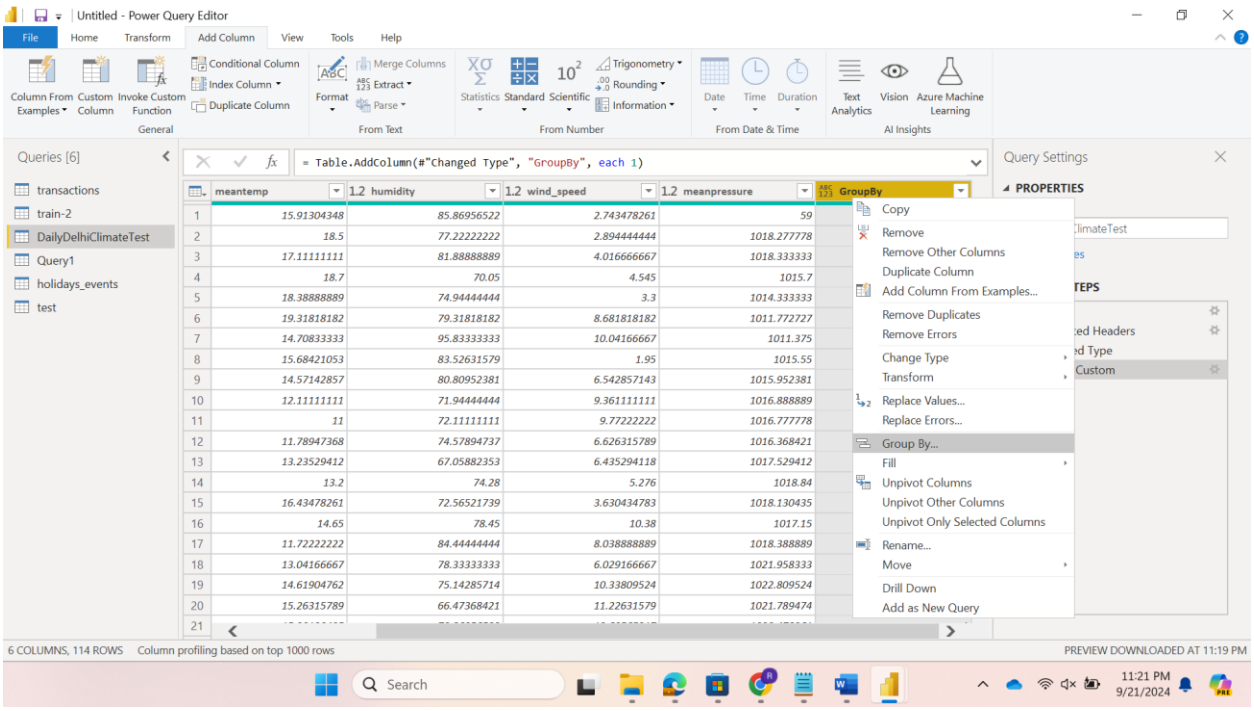

Fig 1.4

Creating Group By attribute to this table. First column name is `MeanValue` that gives the mean value of 'humidity' attribute. Next column is MedianValue that gives the median value of 'humidity' attribute.

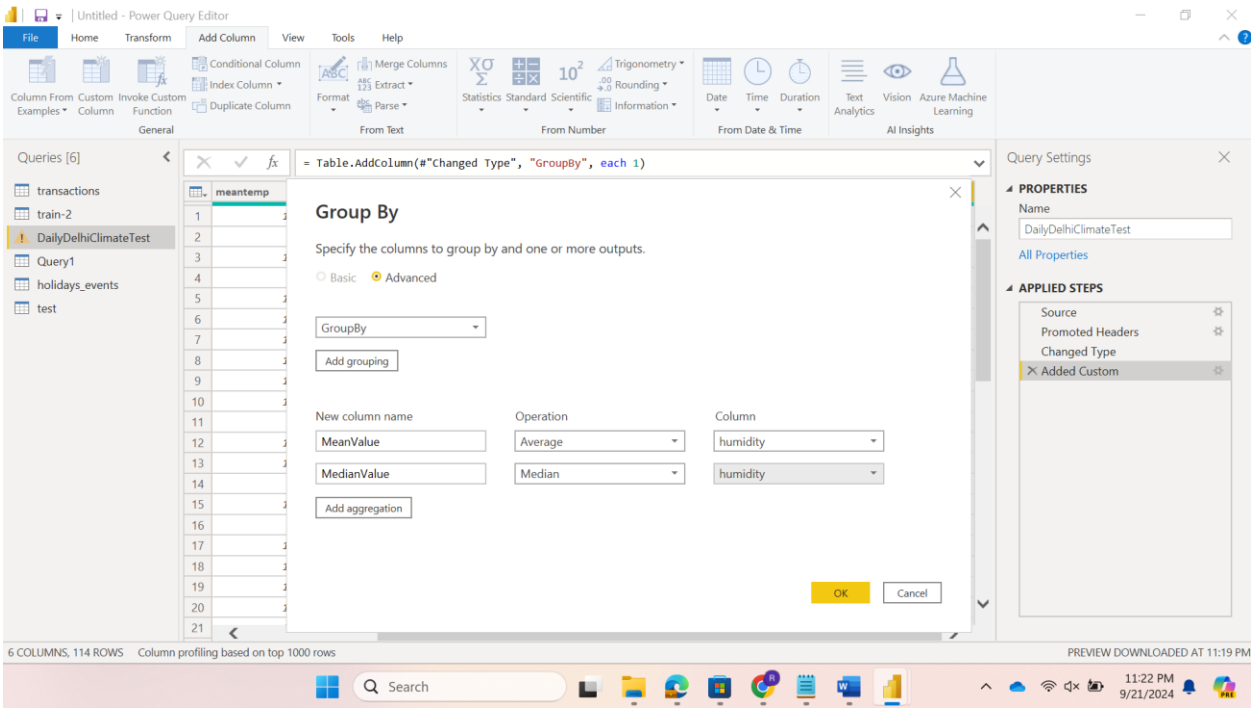

Fig 1.5

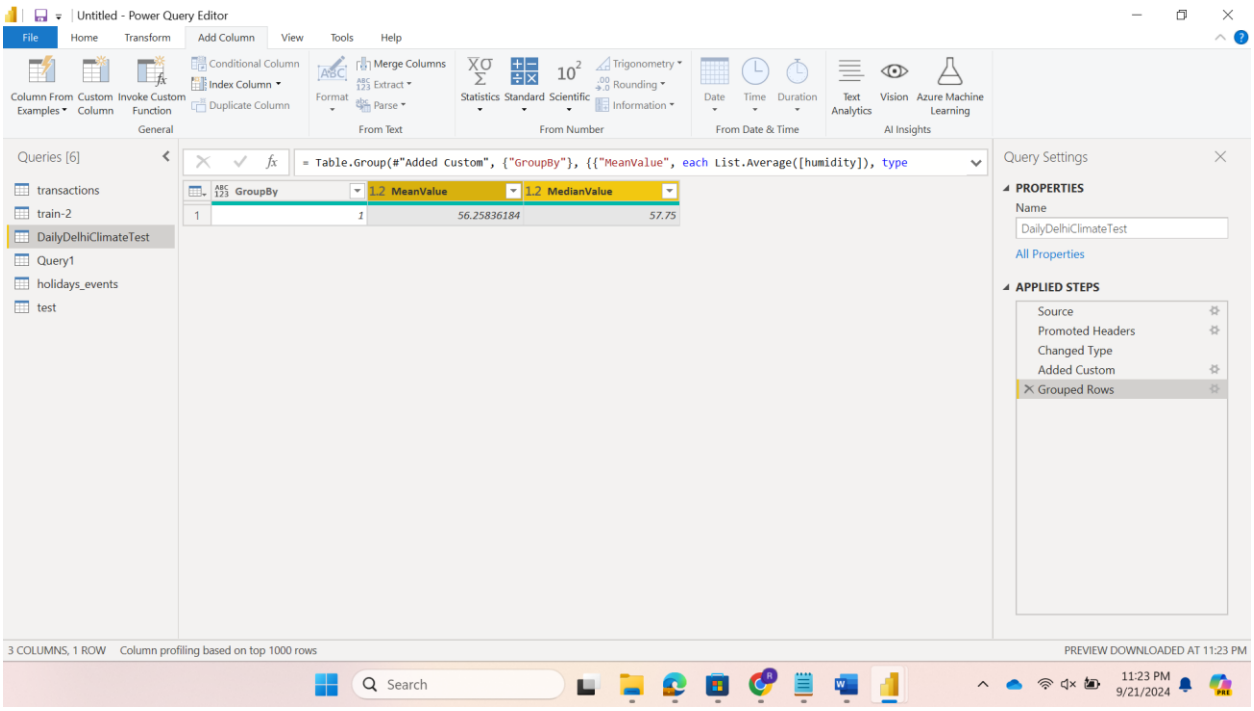

Fig 1.6

From the result (fig 1.6), we see Mean Value is approximately 56.26 and Median Value is 57.75.

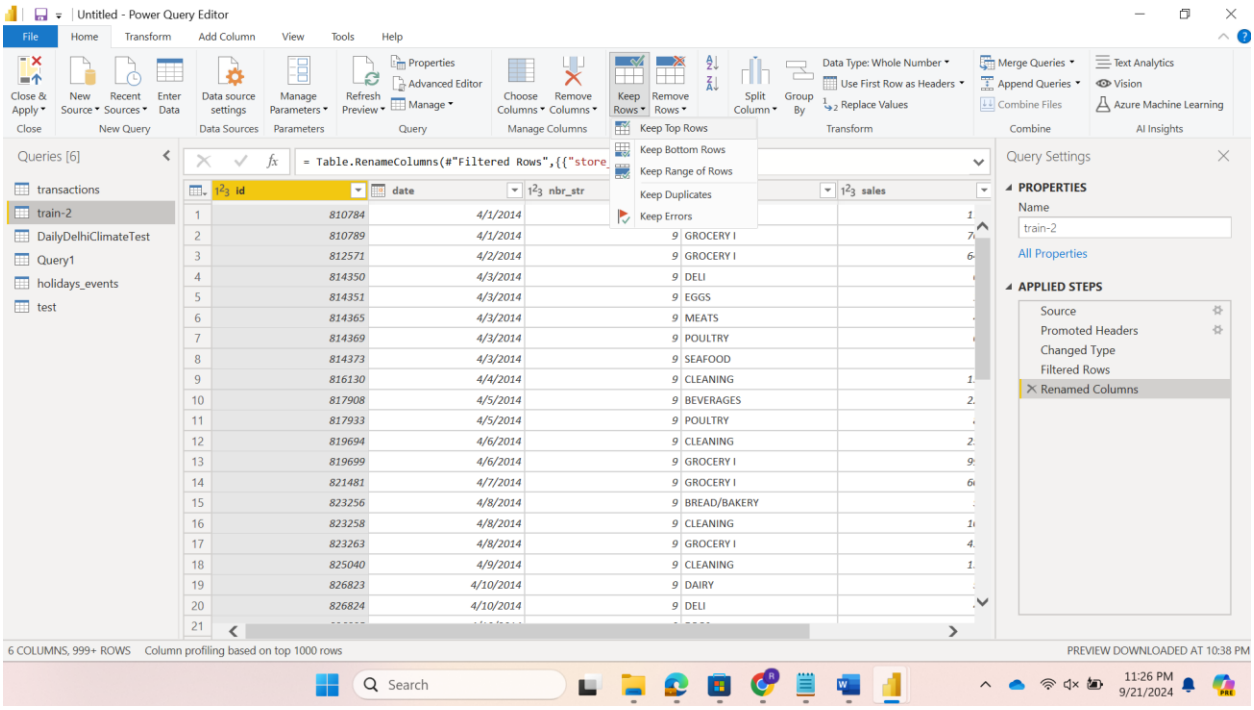

Fig 1.7

3. Query sort id <700000: Click on the drop-down menu of ‘Id’ attribute, Now select the ‘Number Filter’ -> ‘Less Than…’ as shown in fig 1.8

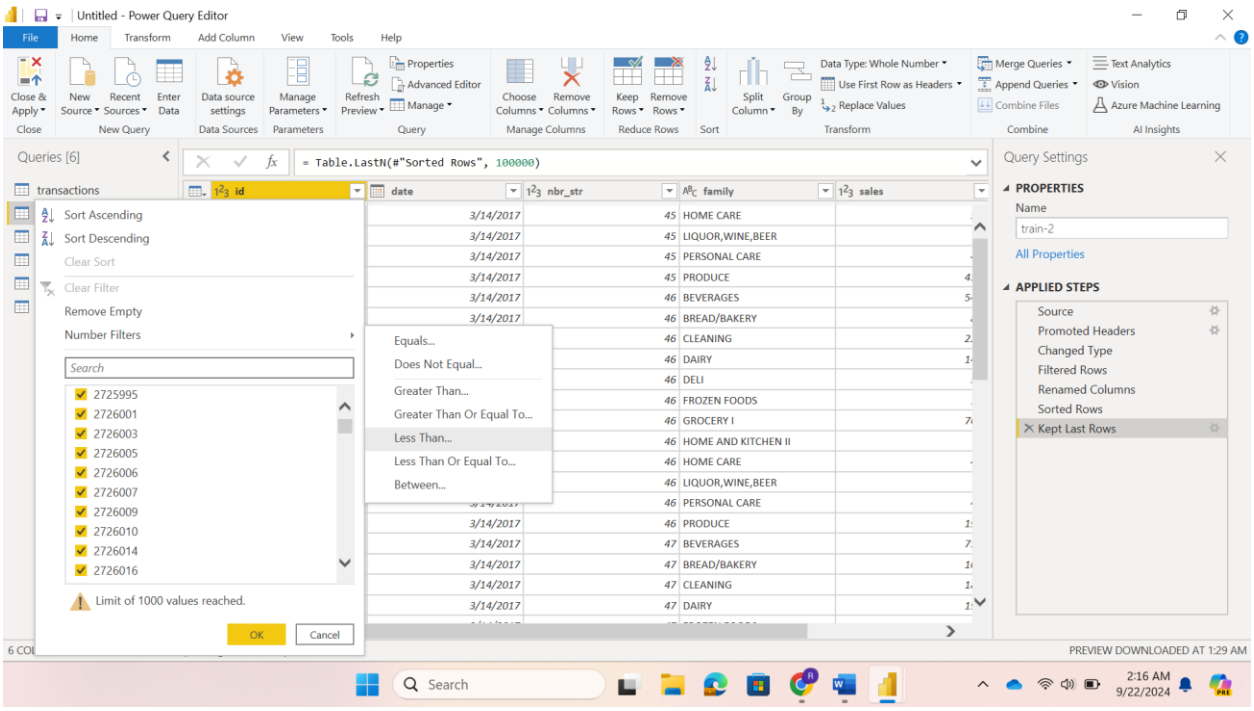

Fig 1.8

Now enter the value as 700000, as per the query (fig 1.9)

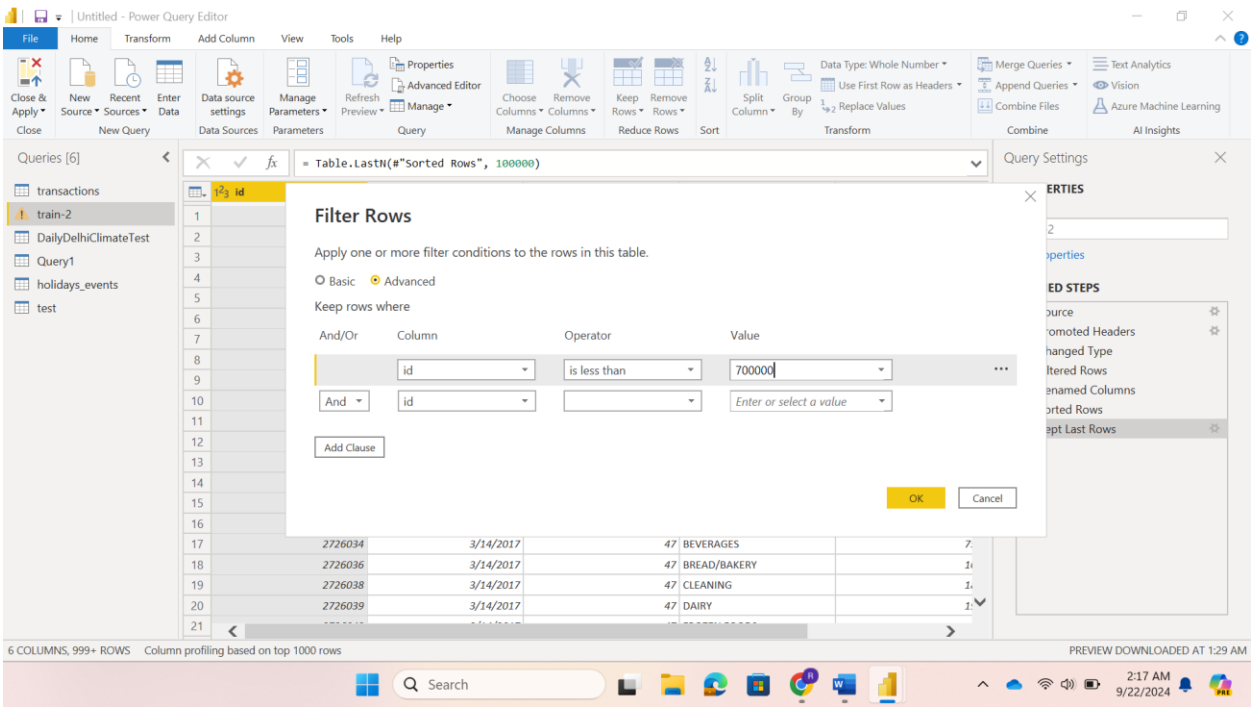

Fig 1.9

Now we get all rows with id<700000(fig 1.10).

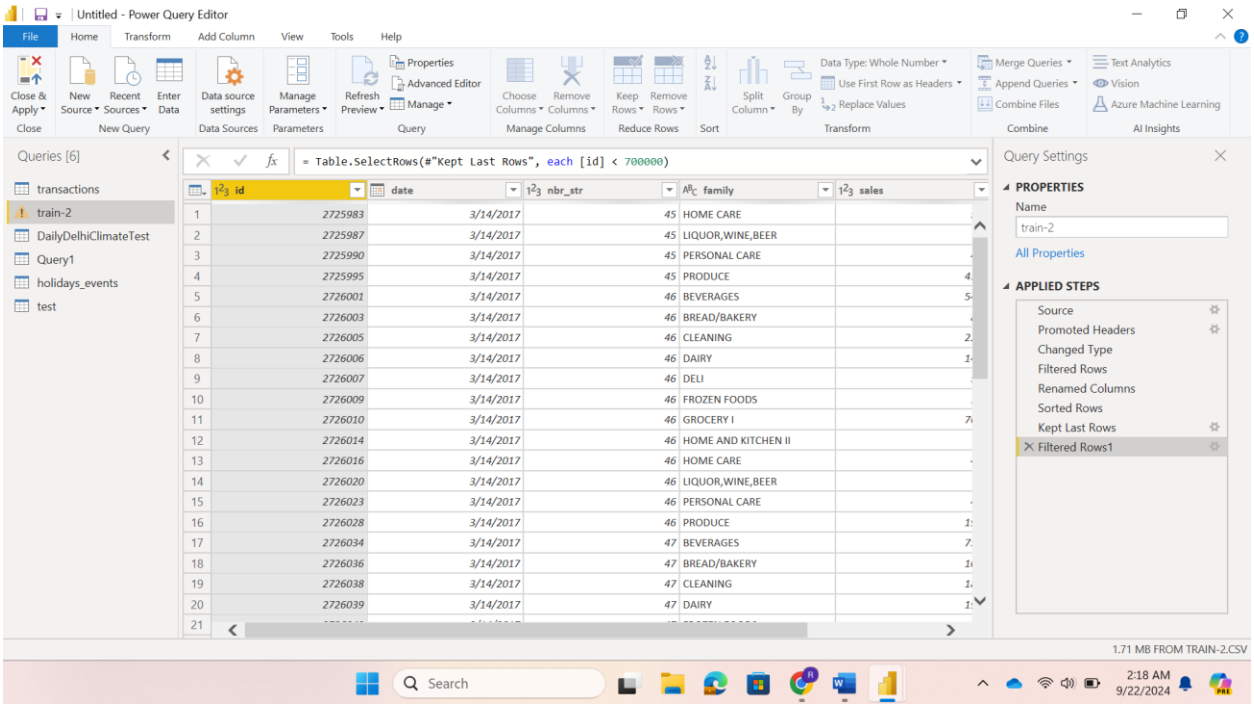

Fig 1.10

Difficulties faced while doing this task:

The train-2.csv is a huge file and it takes a lot of time (9-10 minutes) to load. You will face this issue if you are using PowerBI application through myuntlab. However, using the PowerBI DESKTOP application in your laptop, the loading can be done faster(5-10 seconds). Not only loading the data, but also while doing sorting and filtering ID's in train-2.csv. I have faced this issue. Another thing I have faced is at step 11, i got an error called no "Changed Type" then I realized that we need to add a new step and then execute this query. Hence after adding the new step I managed to solve this issue.

Question: Filter the Dates in transaction in descending order:

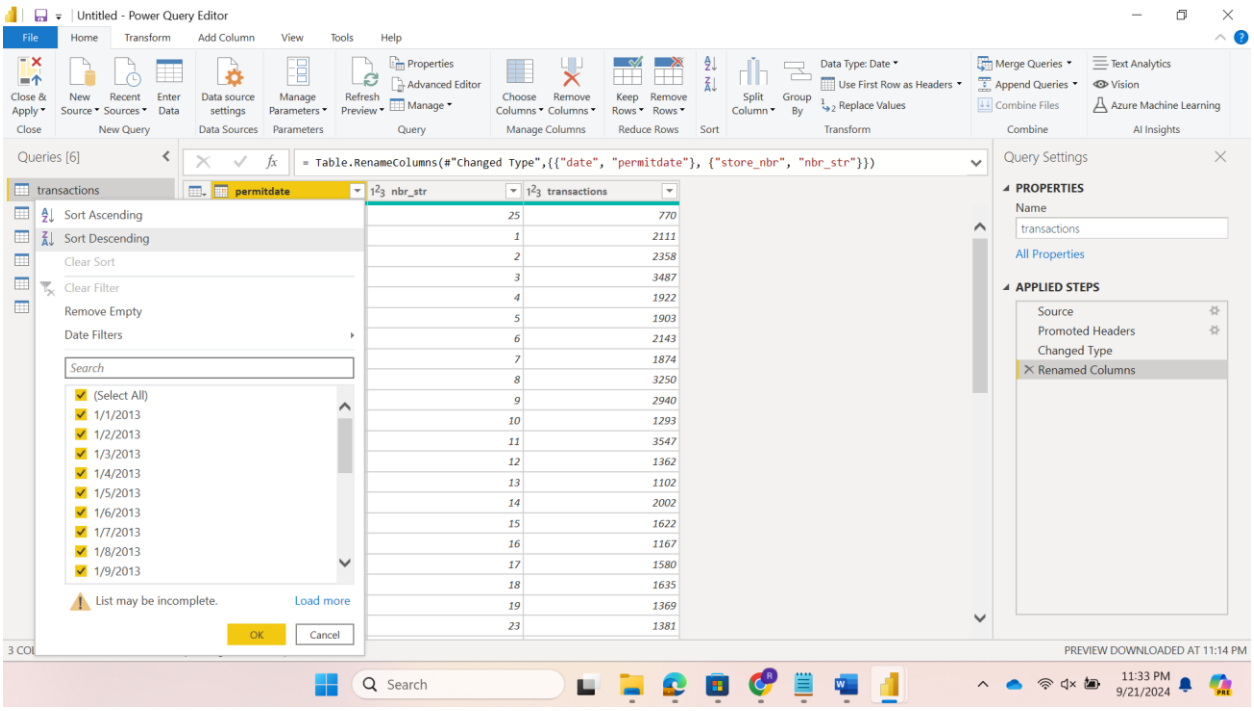

Fig 1.11

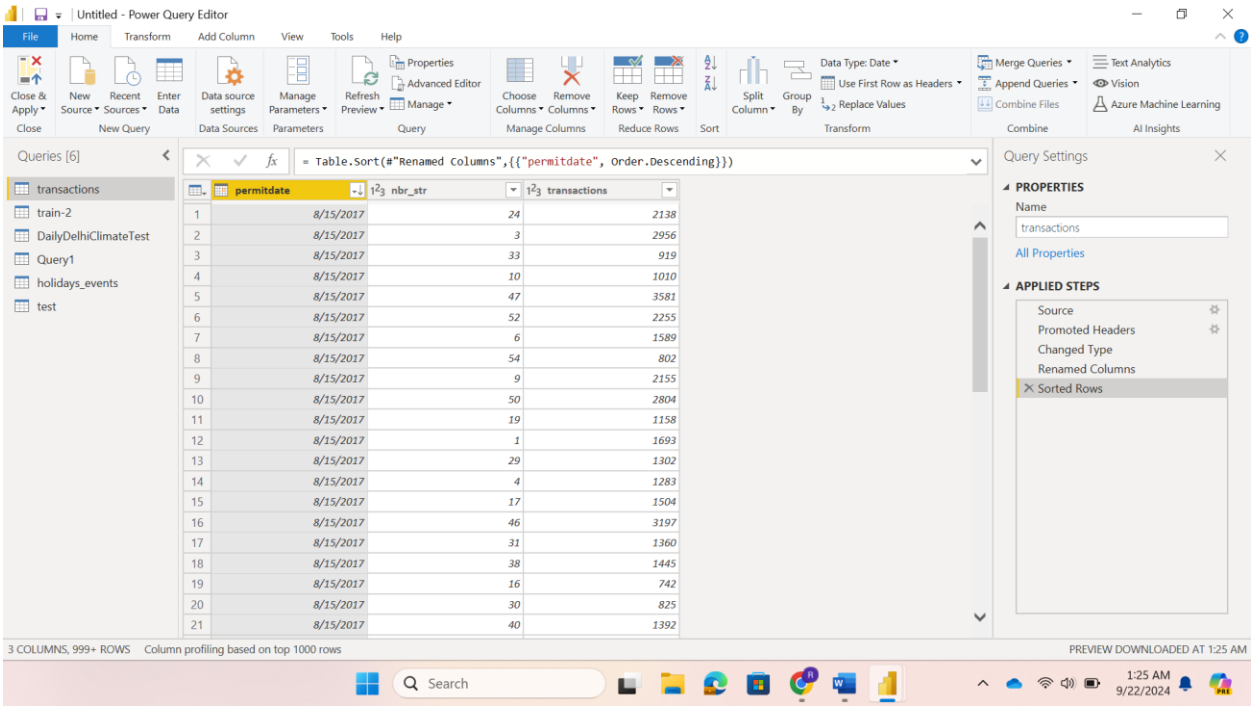

Fig 1.12

Explanation: Click the drop-down symbol at ‘id’ column name and click on the ‘Sort Descending’ option. Now all the dates are sorted in descending order or use the following query: Table.Sort(#“Renamed Columns”, {{“permitdate”, Order.Descending}})

Question 2:

There is one relationship between train: id and test: id that is a one-on-one relationship since Id value is unique.

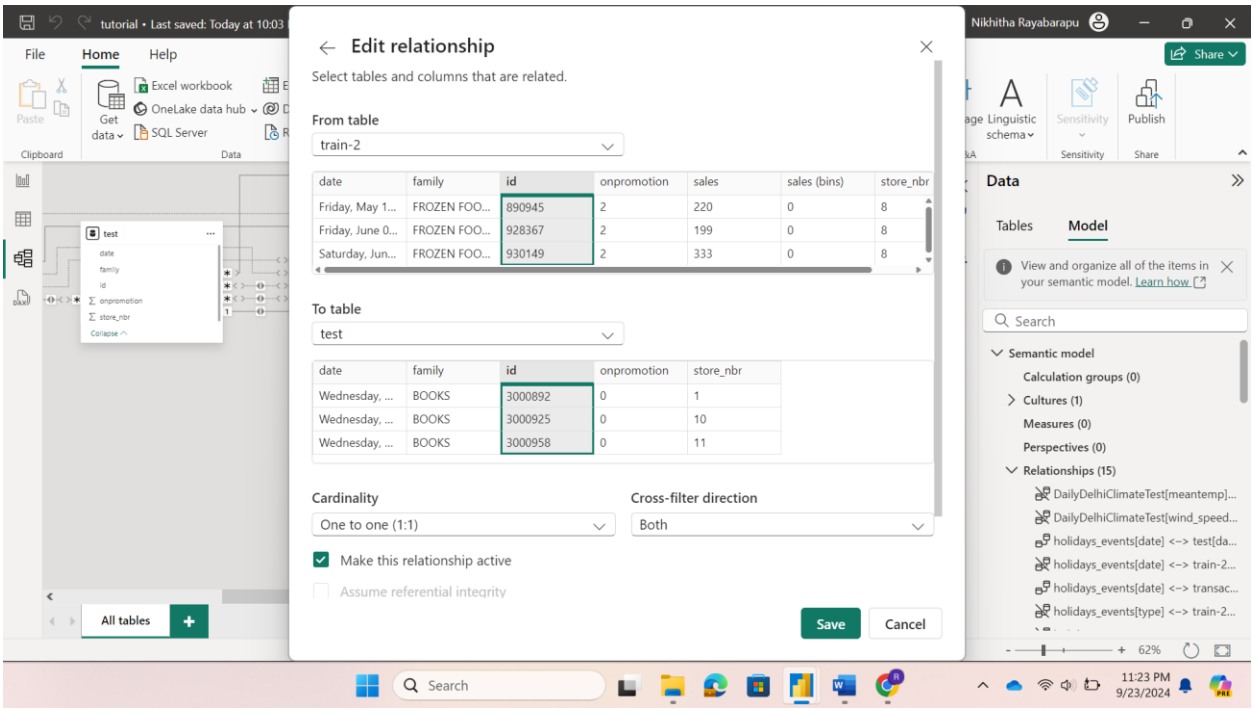

Fig 2.1

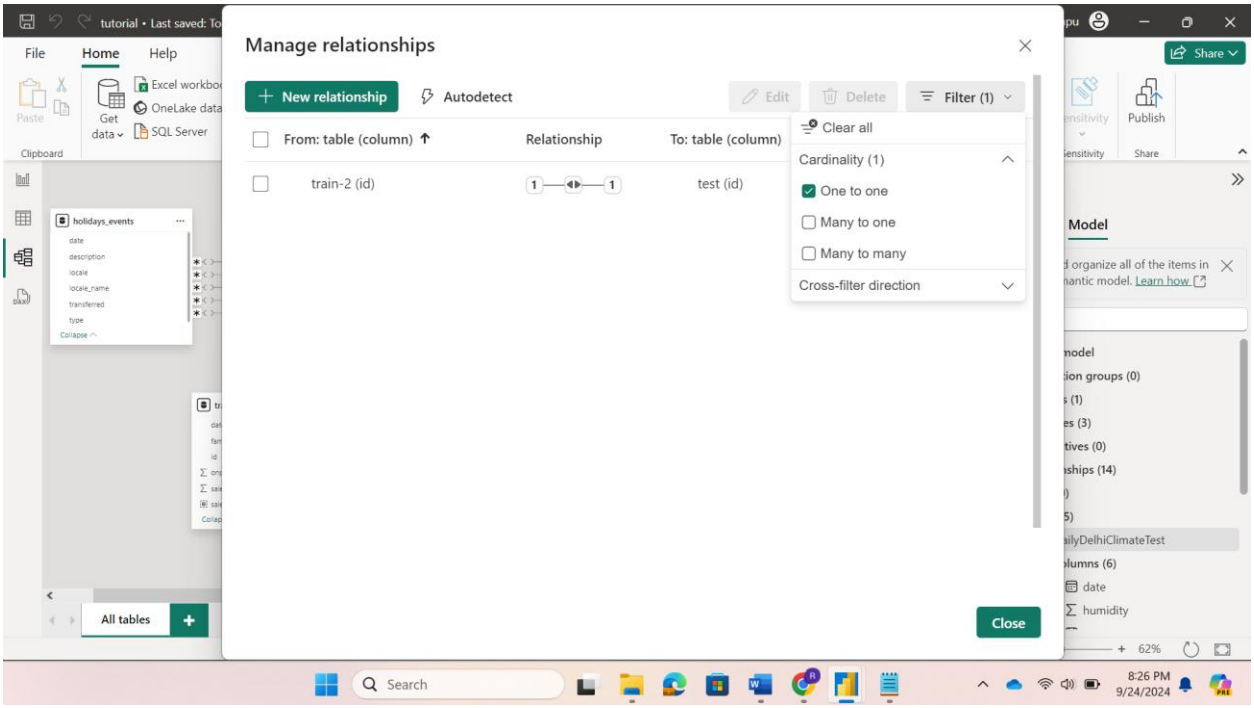

Fig 2.2

There are a total of **4 possible active relationships** in the given data(fig 2.2). The rest of the relationships are inactive.

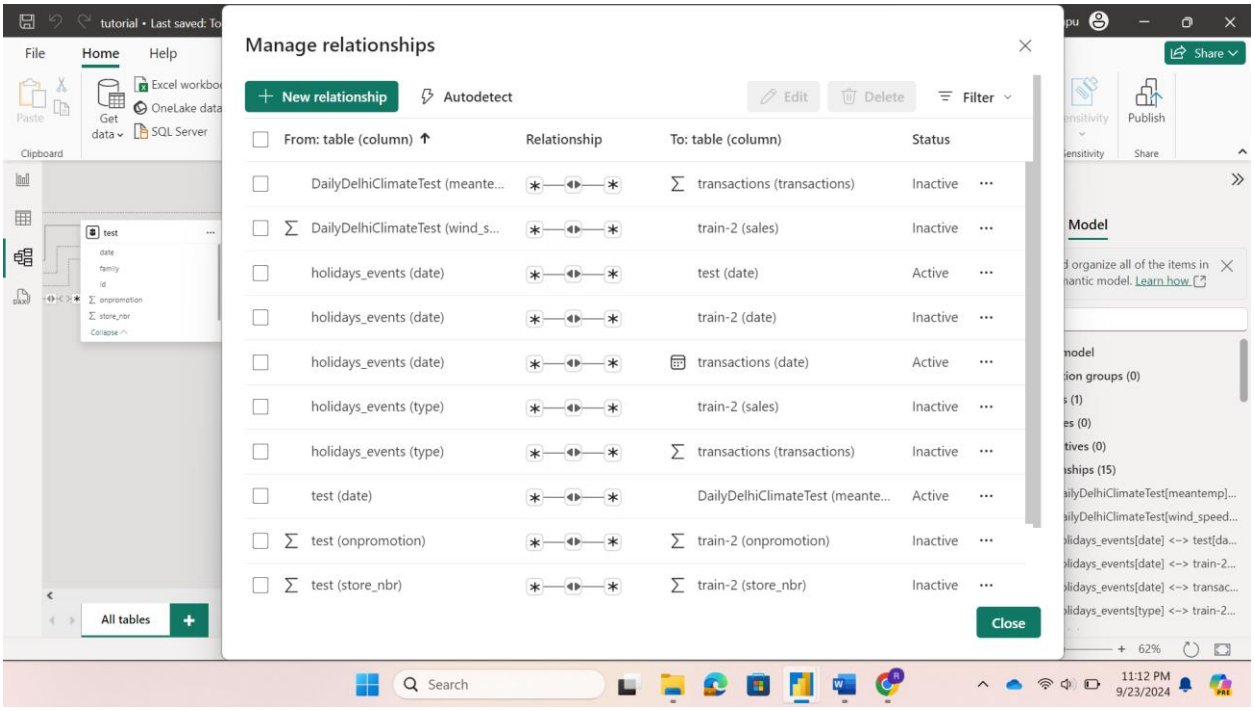

Fig 2.3

Model View after finding out all the possible relationships in my data.

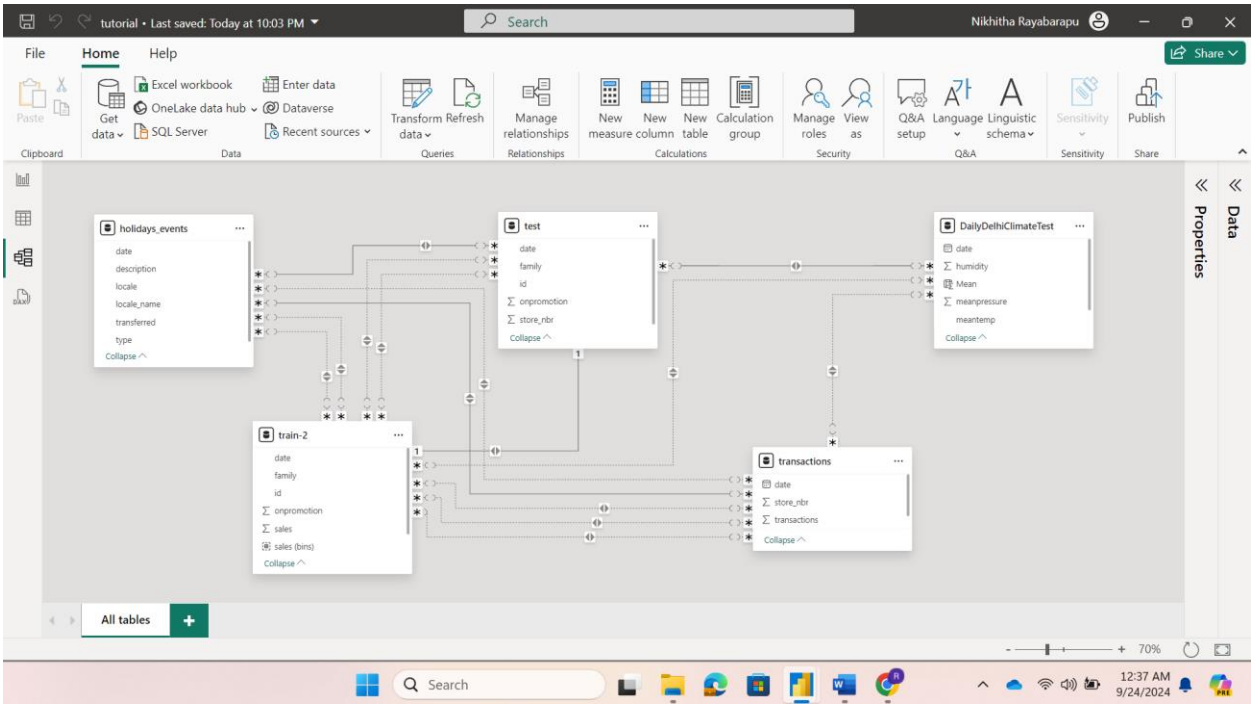

Fig 2.4

Will cardinality of the relationship really affect the relationship and the way the attributes interact with the other?

Ans:

Yes, the way attributes from different tables interact with one another in a data model is greatly influenced by the cardinality of a connection. Cardinality affects the form and behavior of data relationships by defining the number of entries in one table that correspond to records in another.

Please answer the following Questions

A. What happens if you select cross-filter direction as single and filter based on the attributes?

Ans: Then the filter will move only in one direction between the two related tables.(if single is selected). It means any filtering based on attributes from one table will effect the other table, but not the other way around

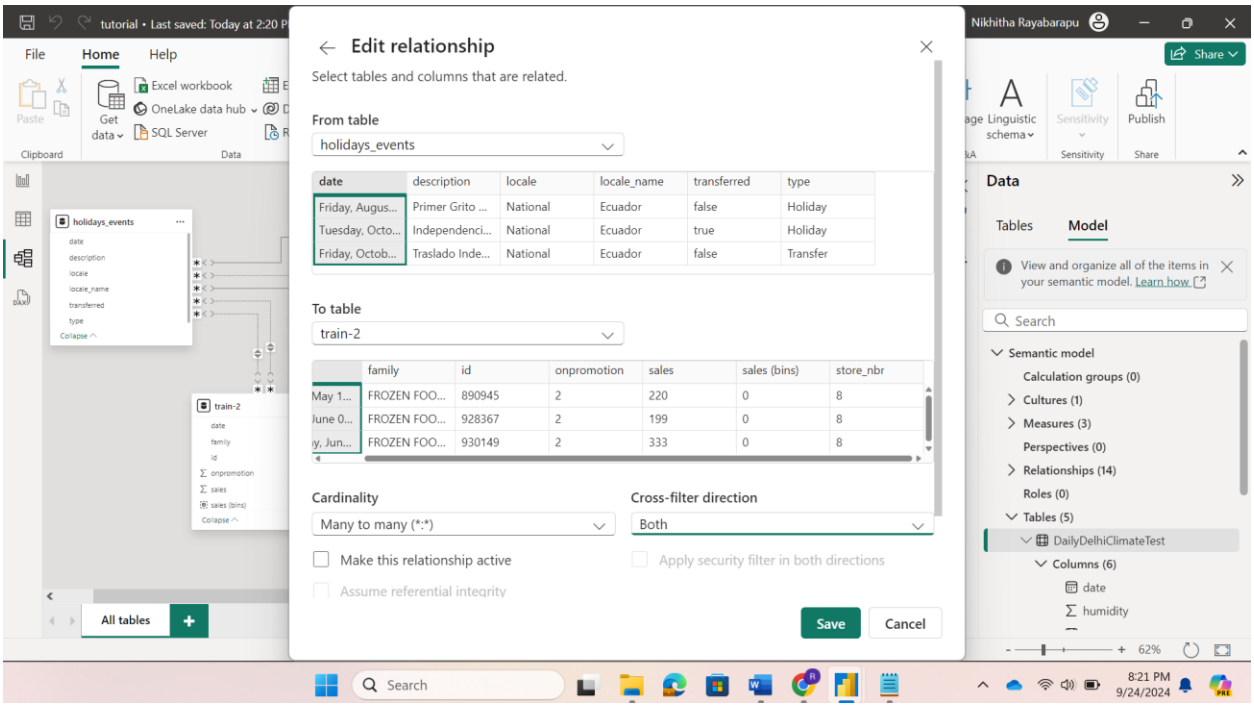

Fig 2.5

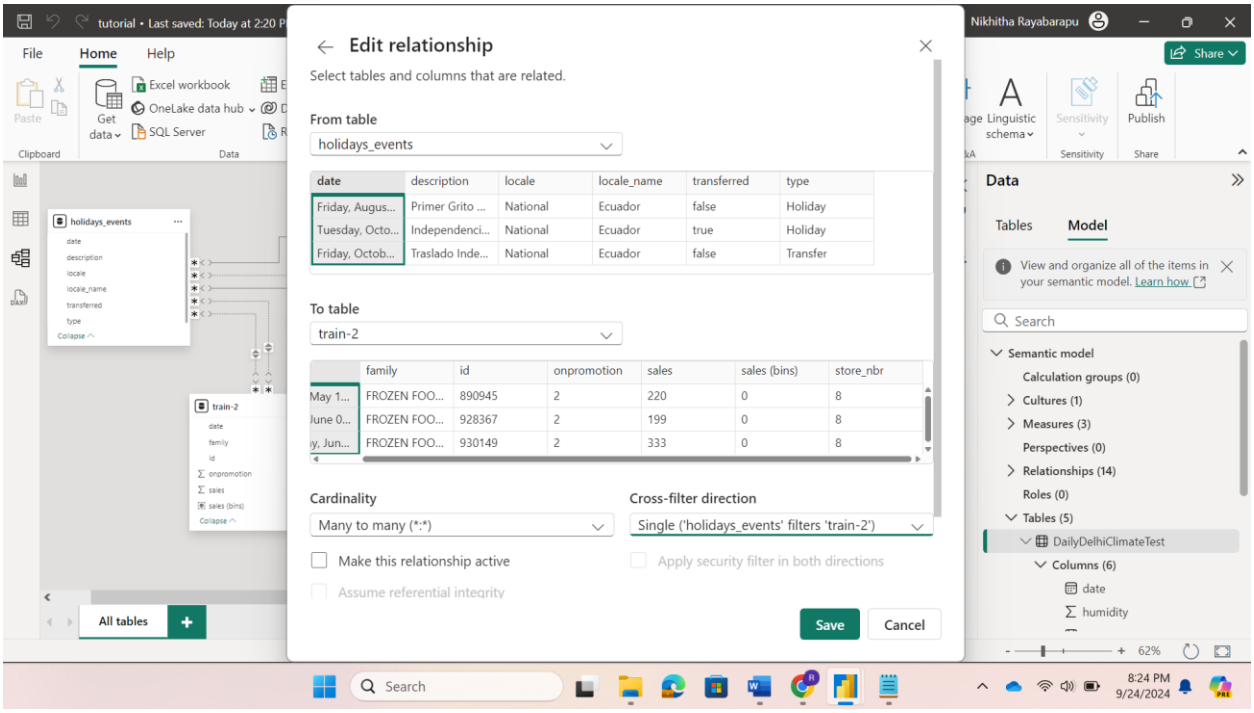

Fig 2.6

B. Can we use multiple relations in the table and set them as active? Please explain your opinion.

Having multiple active relationships between the same table can cause uncertainty and confusion(also called “ambiguity”) over which relationship to apply when doing computations. This results in inaccurate results. Hence, by having only one active relationship, the model remains consistant, while also using inactive relationships as needed for different circumstances.

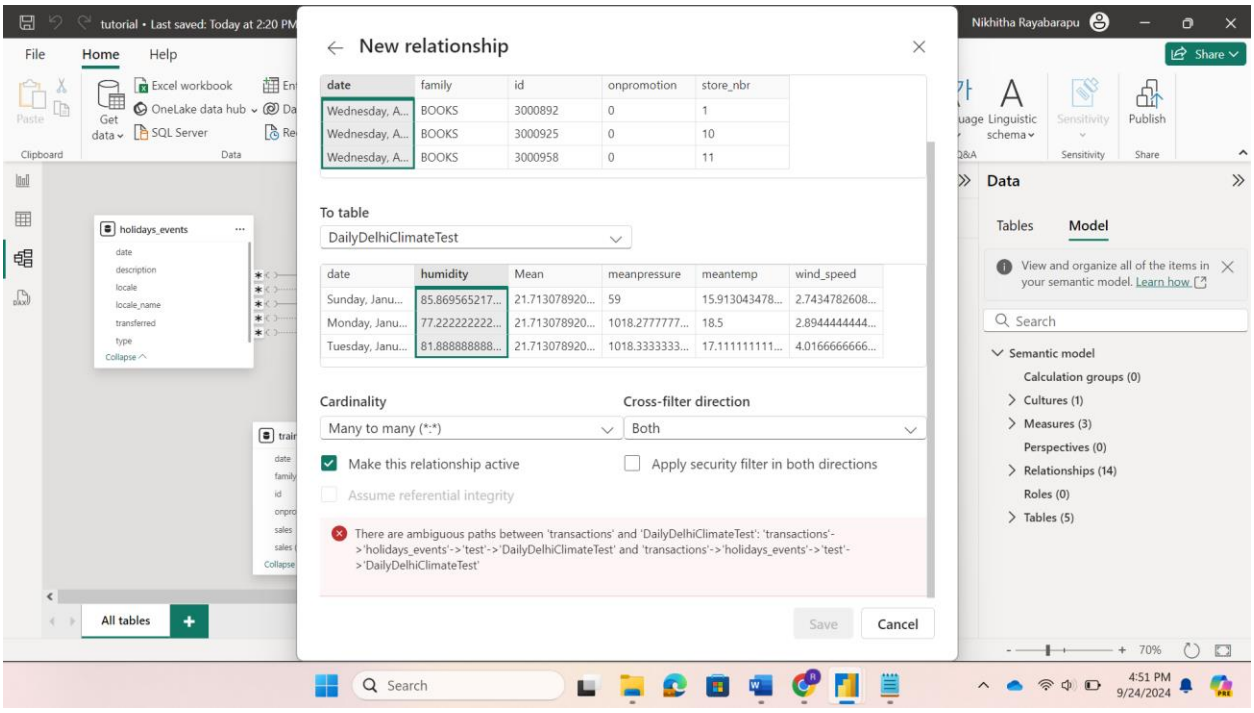

Fig 2.7

In fig 2.7, you can see the warning Power BI is giving if we try to make multiple active relationships.

Question 3:

Select the attributes Transferred, Type and Date from Holidays_events and transactions from transactions table.

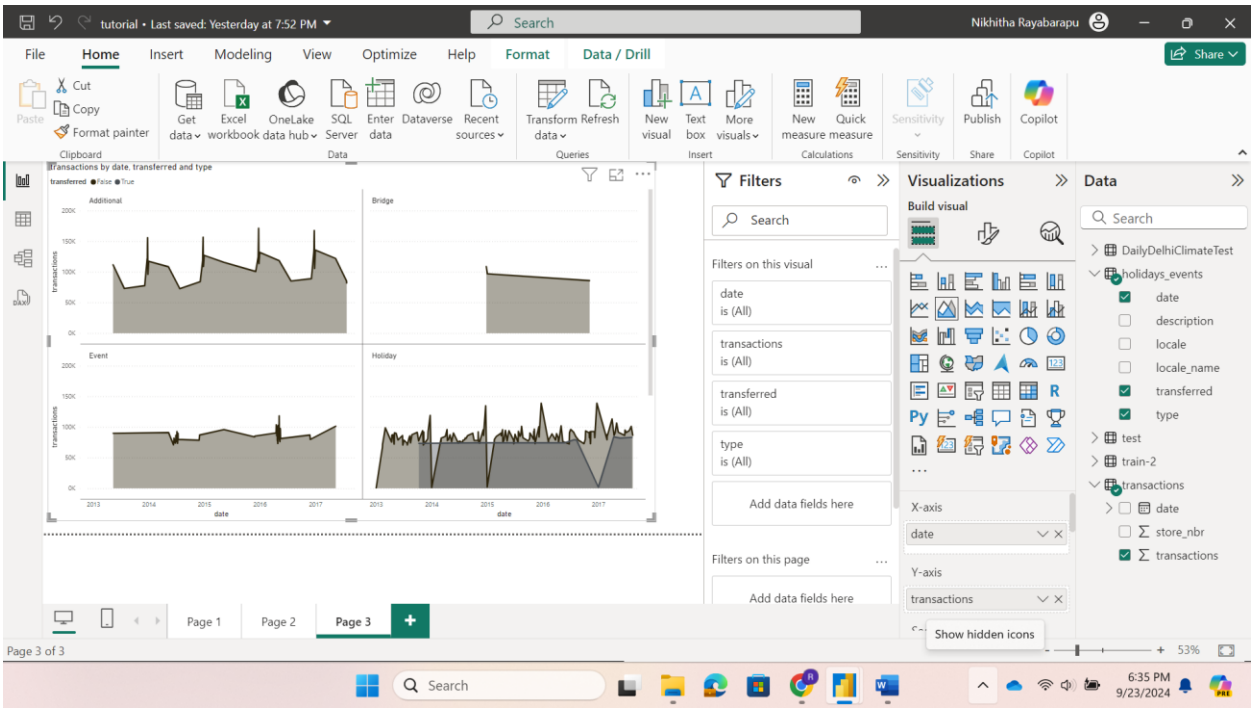

Fig 3.1

Explanation:

The above graph (Fig.3.1) shows many transactions done on different holiday categories over the time from 2013 to 2017. The categories include additional, bridge, event, etc. The graph also shows whether or not the holiday was transferred. The X axis shows the timeline and the Y axis indicates the number of transactions ranging from 0 to 200,000. The transactions in the additional holidays are consistent with periodic peaks, most likely associated with recurrent holidays. The holiday panel has the most variations with many peaks and troughs in the number of transactions. The light Gray area depicts the non-transferred holidays and the dark Gray area depicts the transferred one.

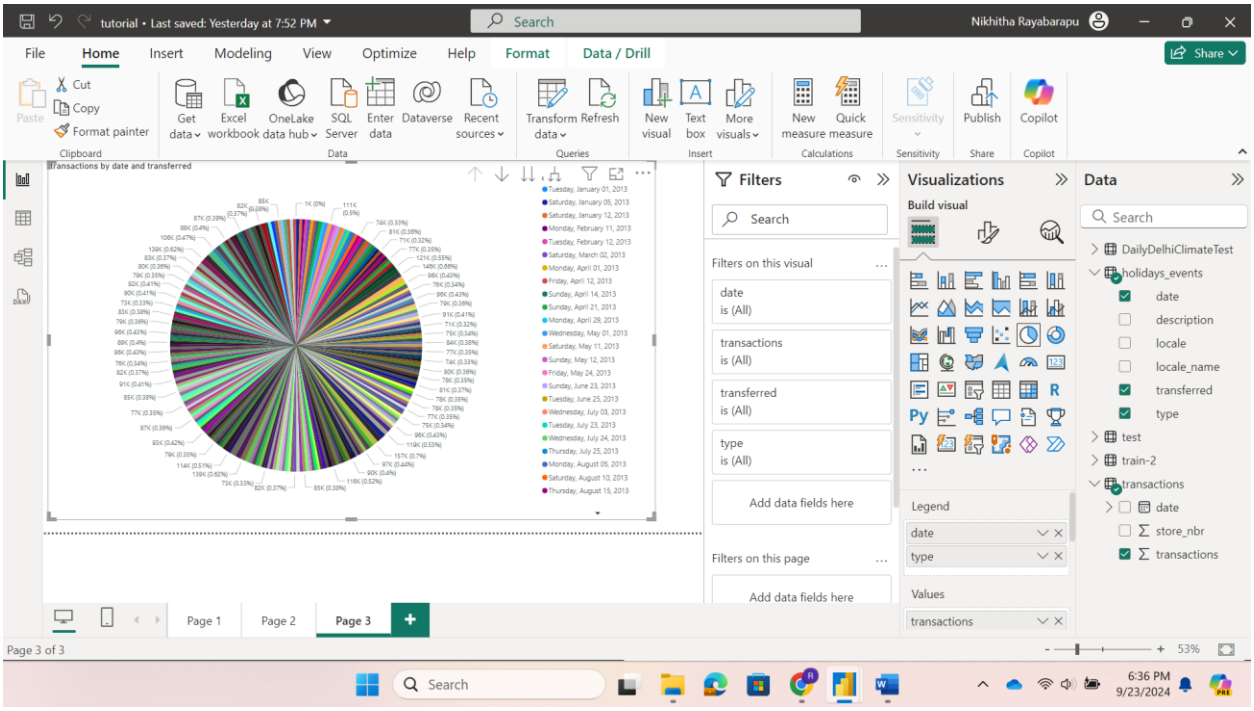
tutorial • Last saved: Yesterday at 7:52 PM
Search
Nikhitha Rayabarapu
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Publish
Share
Copilot
Transactions by date and transferred
Filters
Search
Filters on this visual
date is (All)
transactions is (All)
transferred is (All)
type is (All)
Add data fields here
Filters on this page
Add data fields here
Visualizations
Build visual
Legend
date
type
Values
transactions
Data
Search
DailyDelhiClimateTest
holidays_events
date
description
locale
locale_name
transferred
type
test
train-2
transactions
date
store_nbr
transactions
Page 1
Page 2
Page 3
Page 3 of 3
53%
Search
6:36 PM
9/23/2024

Fig 3.2

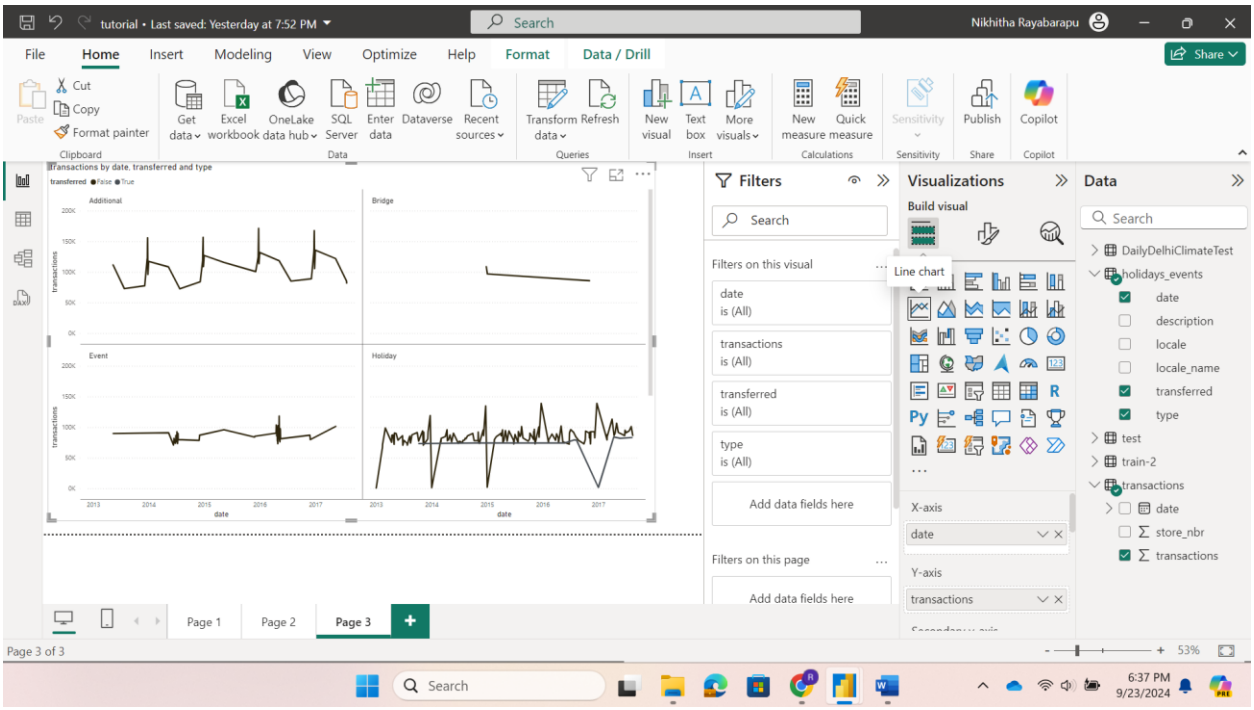
tutorial • Last saved: Yesterday at 7:52 PM
Search
Nikhitha Rayabarapu
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Publish
Share
Copilot
Transactions by date, transferred and type
transferred False True
Additional
Bridge
Event
Holiday
transactions
date
Filters
Search
Filters on this visual
date is (All)
transactions is (All)
transferred is (All)
type is (All)
Add data fields here
Filters on this page
Add data fields here
Visualizations
Build visual
Line chart
X-axis
date
Y-axis
transactions
Data
Search
DailyDelhiClimateTest
holidays_events
date
description
locale
locale_name
transferred
type
test
train-2
transactions
date
store_nbr
transactions
Page 1
Page 2
Page 3
Page 3 of 3
53%
Search
6:37 PM
9/23/2024

Fig 3.3

Explanation: This graph (fig 3.3) is divided into several panels: first by the category of holiday, Additional, Bridge, Event, and Holiday, and second by a variable that indicates whether a holiday was transferred or not. It displays transaction trends over time. Each panel represents transactions for a certain type of holiday between 2013 to 2017. The transactions in the Additional panel are characterized by periodic peaks; this would mean that there is a constant increase in the volume of transactions at a certain point in time, probably correlated to recurrent extra holidays. The Bridge panel is more regular, with fewer oscillations and low volumes of transactions. In this panel, the light-grey area covers those holidays not transferred, while the dark-grey-shaded area indicates the amount of holiday transferred.

Take another page and select date, sales, and store_nbr from train and use a clustered column chart and explain both the attribute's roles in the visualization and explain your findings.

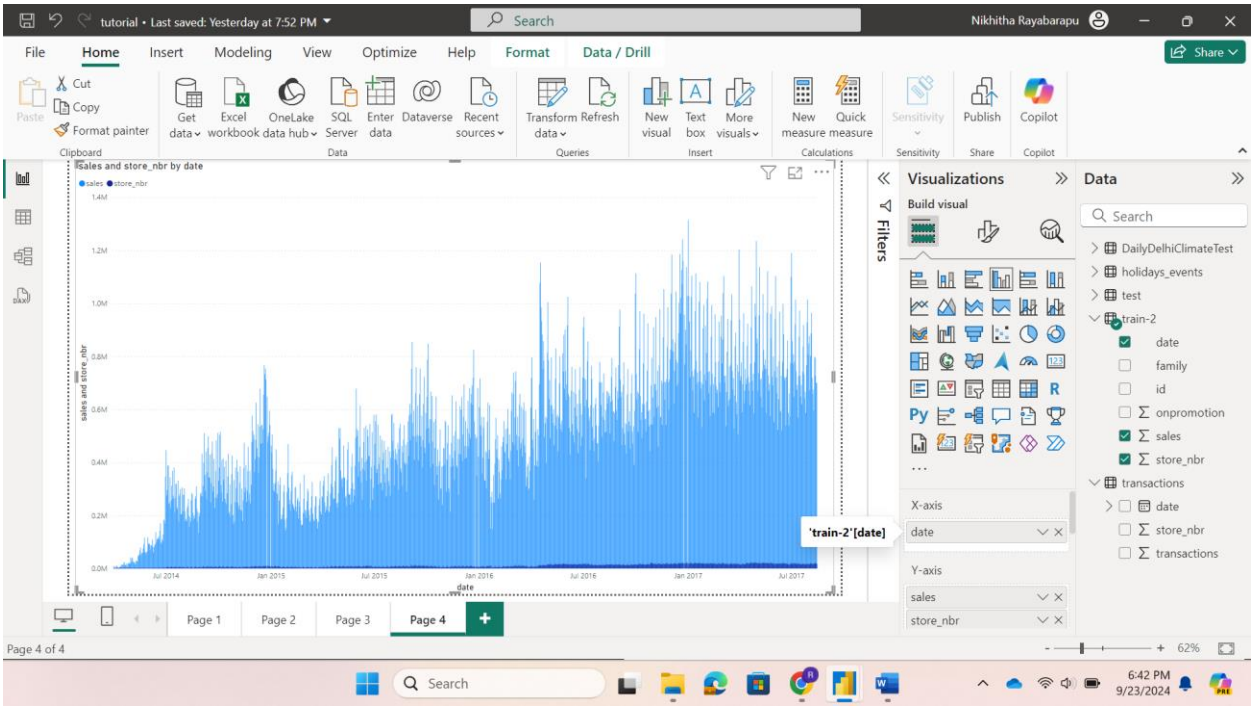

Fig 3.4

Explanation:

This graph takes date on the sales and store_nbr on y-axis. We can see that there is a notable increasing trend in sales over time that is from 2013 to 2017 whereas store_nbr it's really small and how to differentiate from sales in this chart. This maybe due to the values of sales and store_nbr not being in the same range of the scale. Sales values are extremely high and store_nbr values are really small.

Display the Visual Table for the above attributes in separate pages.

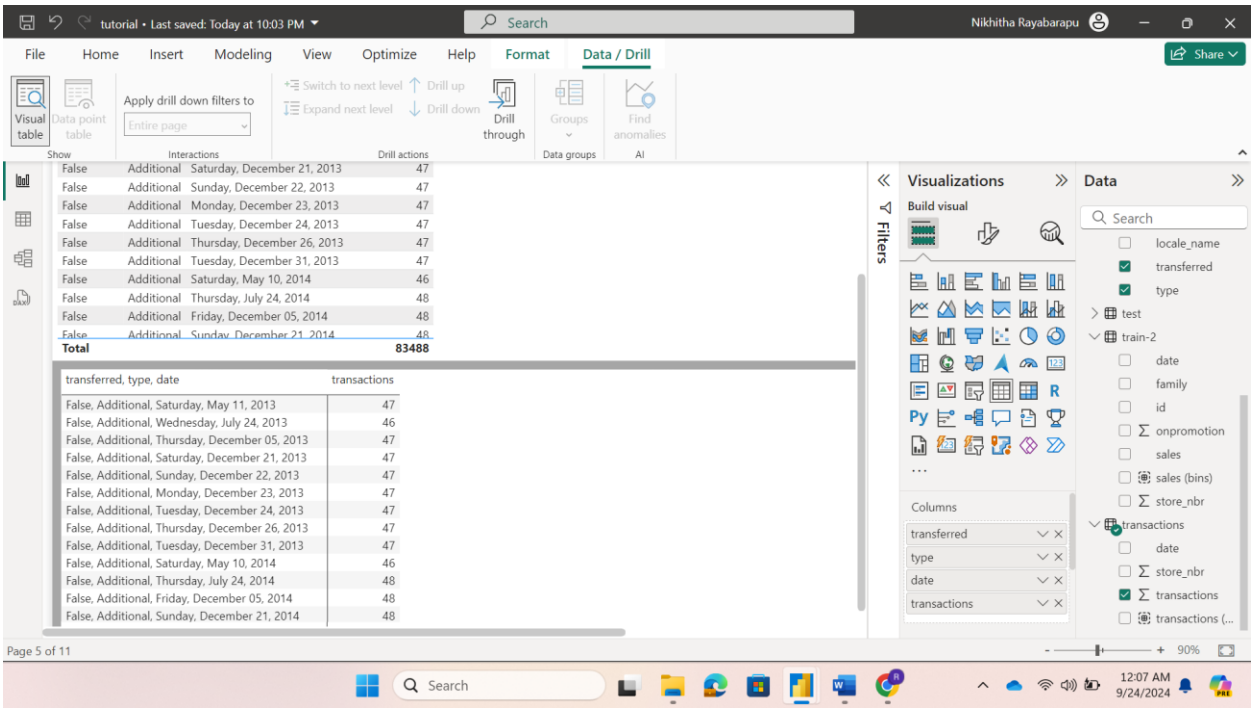

Fig 3.5

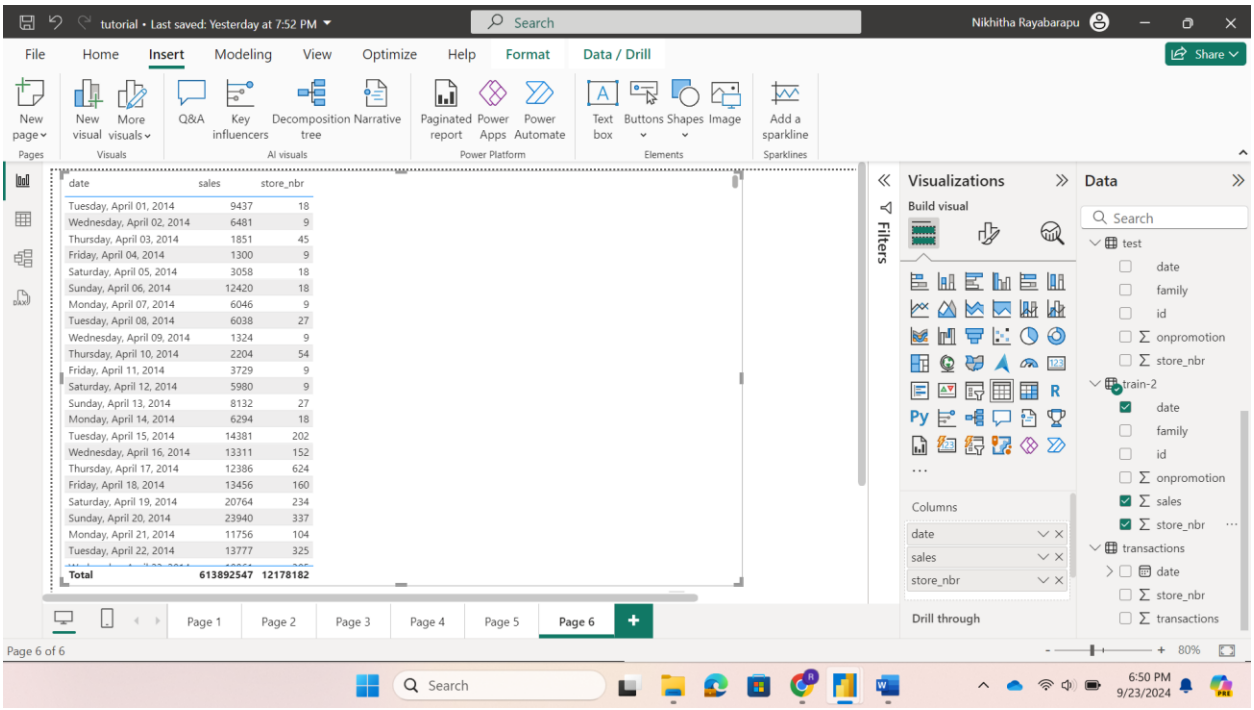

Fig 3.6

A. What is the Highest sales in 2016 and How much compared to the 2nd Highest?

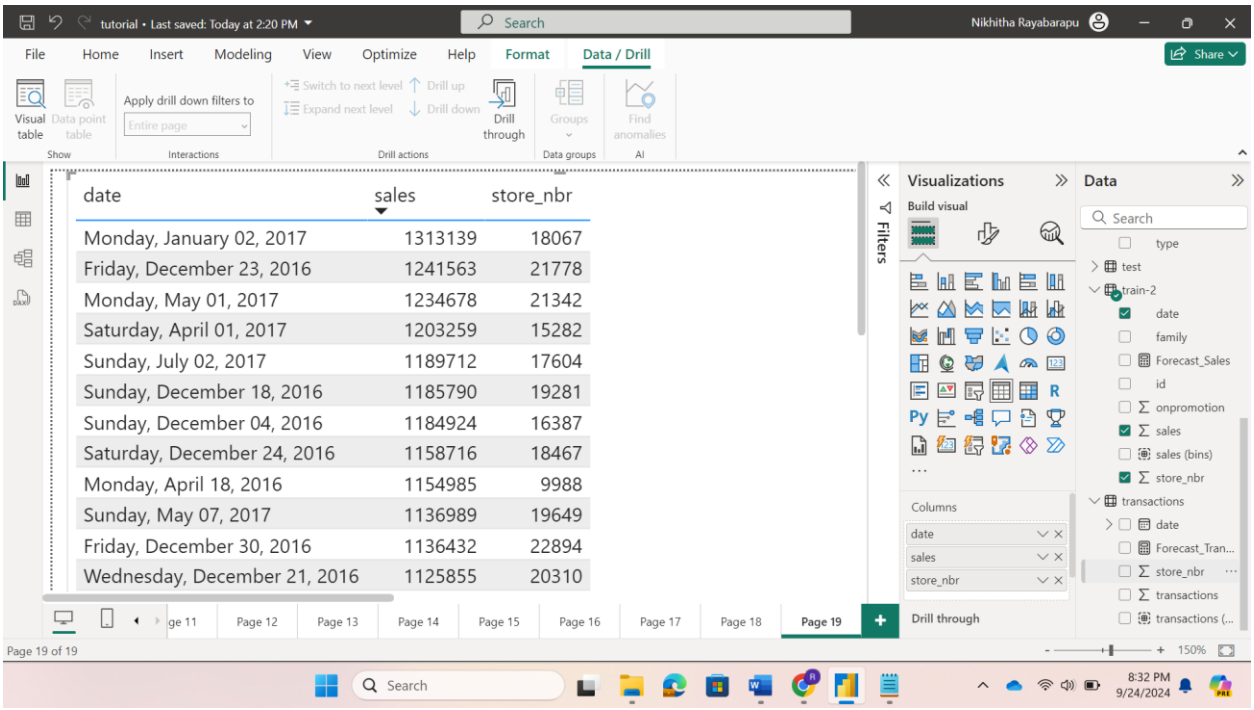

Fig 3.7

The highest sales in 2016 were 1241563. 2nd highest sales in 2016 is 1185790, the difference is 55773.

B. Why do the attributes work differently for different visualizations? (Explain the question based on the attributes given in the above question)

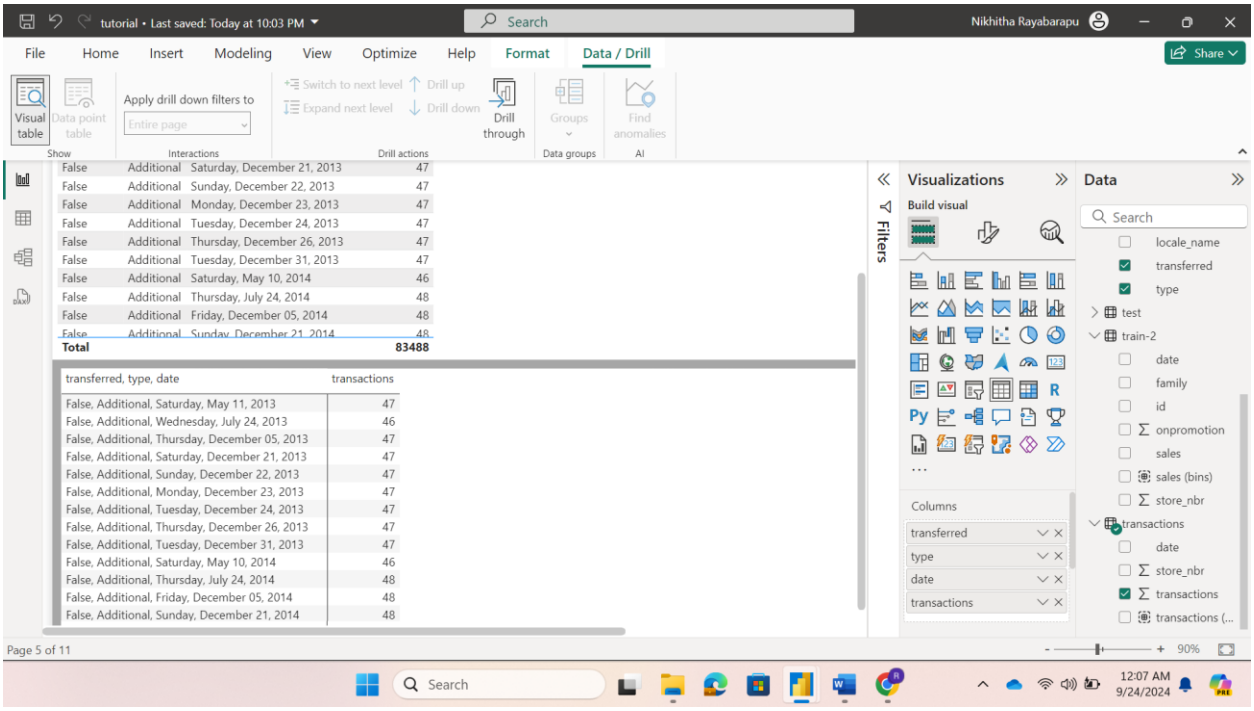

Fig 3.8

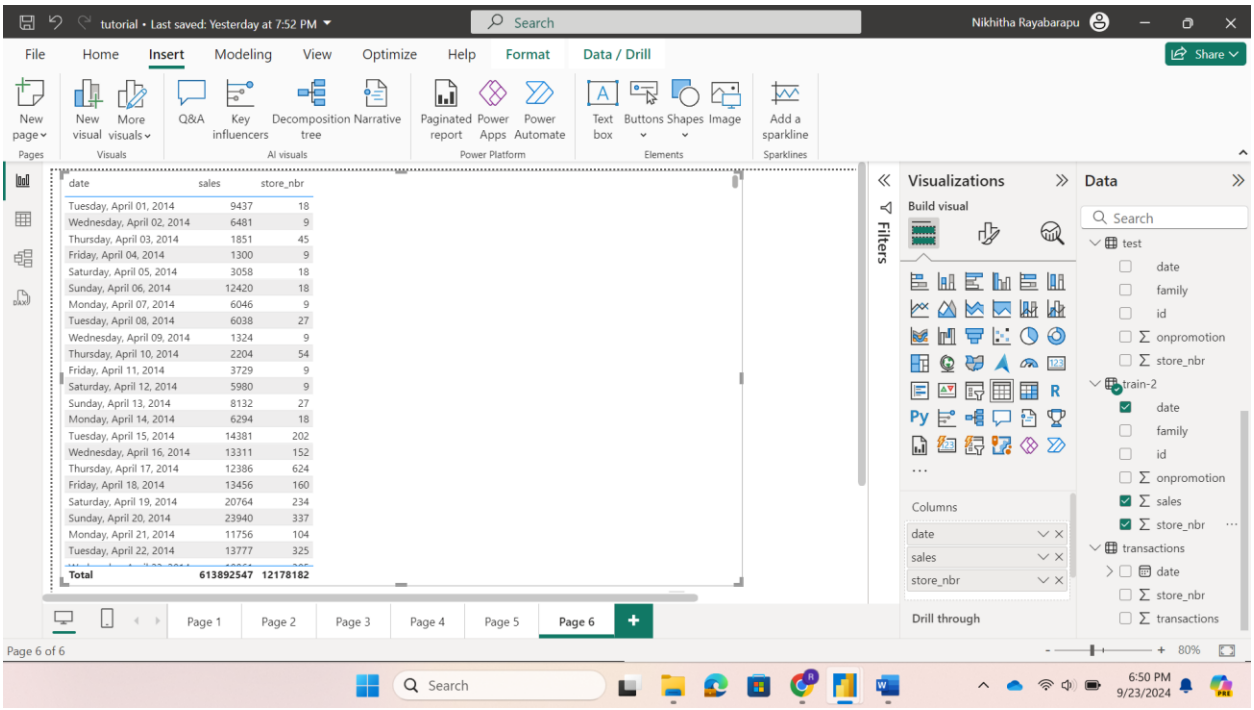

Fig 3.9

Explanation: This is because each attribute is designed to serve a specific purpose, and they interact based on the relationships we have designed. For example, in the above two visualizations, we have date as a common attribute. For the first table (fig 3.8) the date attribute is used to define different types of holidays and number of transactions and whether the holiday is transferred or not. The same attribute when used in the second table (fig 3.9), defines the number of sales and store_nbr, Since the two tables are related.

Question 4:

Use transactions to use grouping to the attribute instead of Sales.

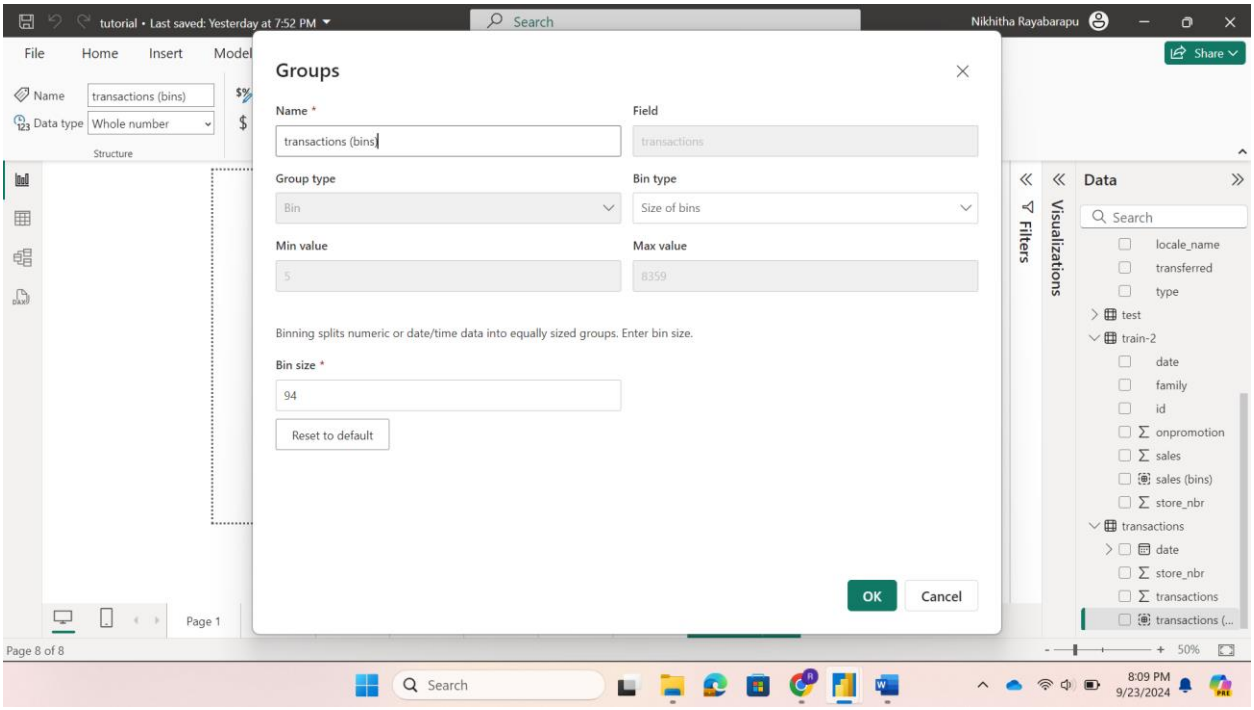

Fig 4.1

Select all the attributes in Transactions table to visualize and remove quarters from the date in the X axis.

With Quarter:

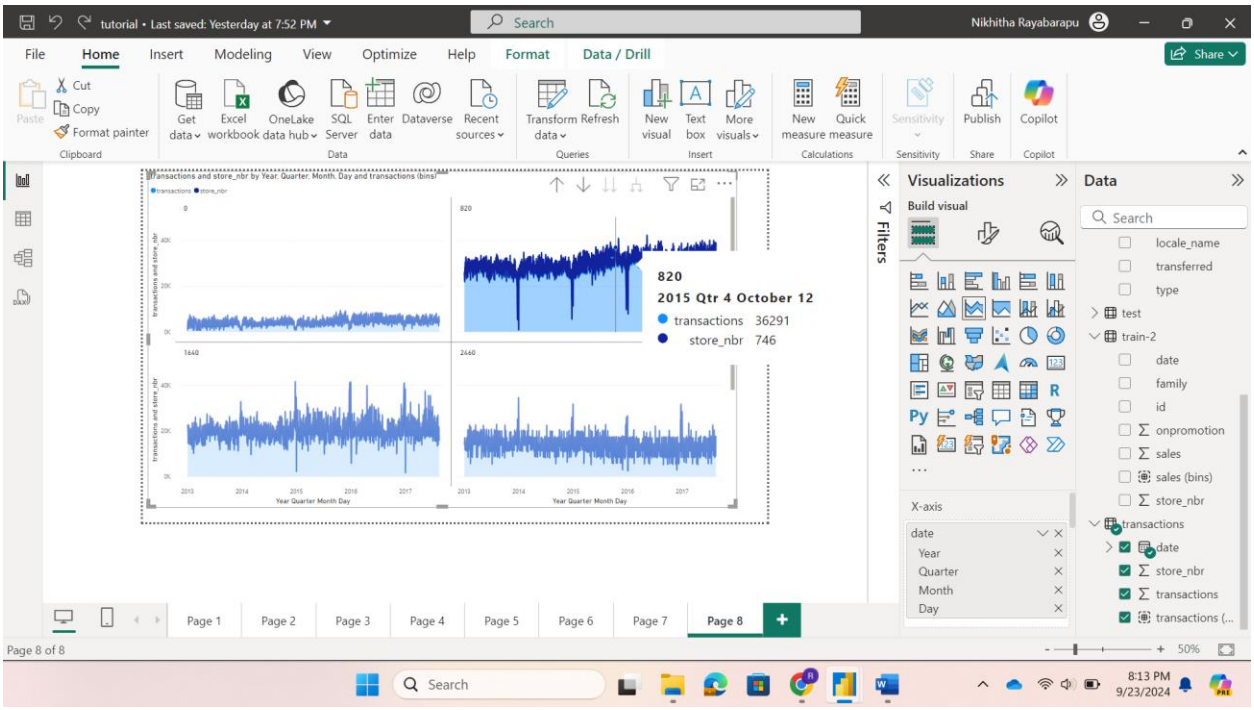

Fig 4.2

Without Quarter:

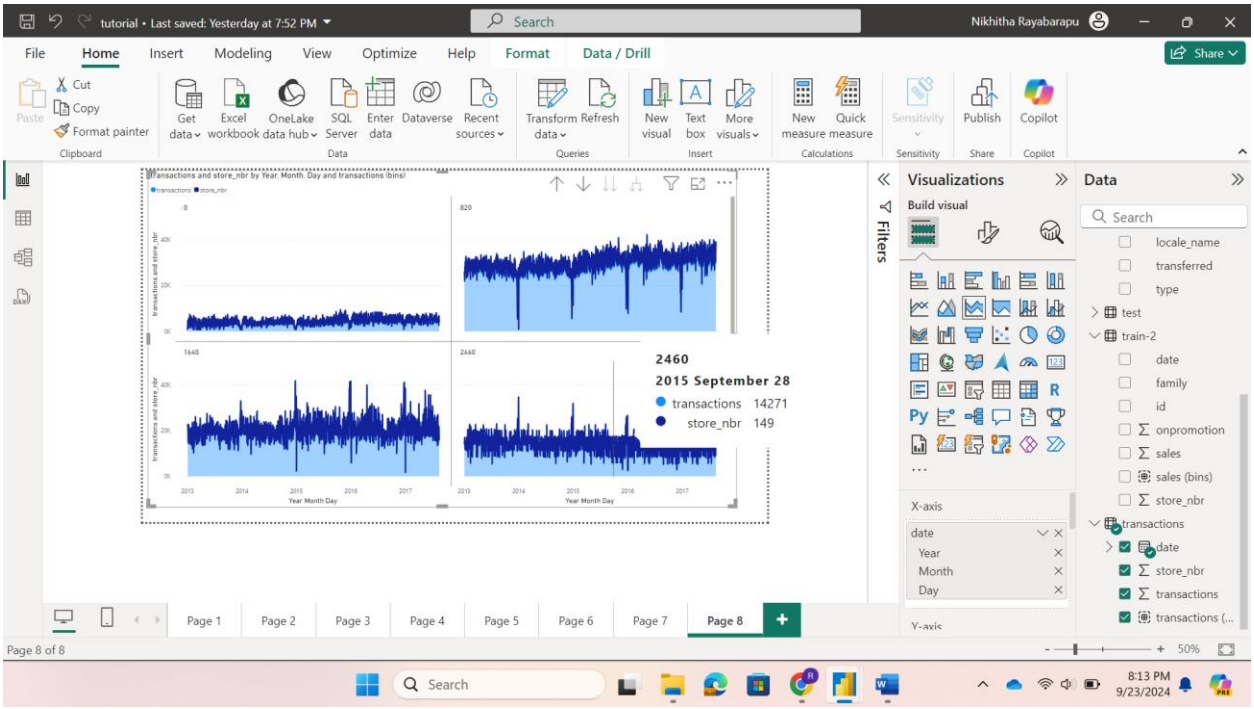

Fig 4.3

Note the changes and compare the above attributes using a line graph and Explain your understanding in detail in about 100 – 200 Words.

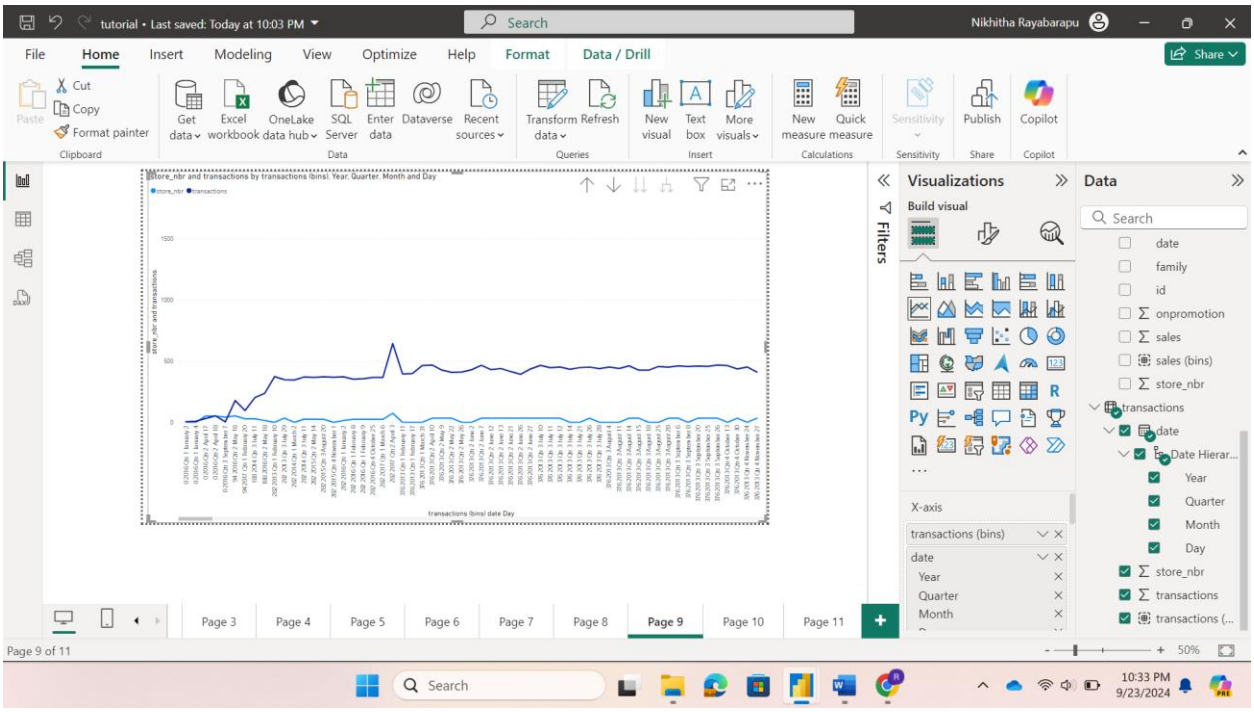

Fig 4.4

Explanation: There are not any significant changes I have observed when I have removed the quarter. It does not impact the overall data. The light blue color denotes the store_nbr attribute and violet color line defines the transaction attribute. The values of 'store_nbr' are lower compared to the values of 'transactions'. We have significantly higher transactions and store_nbr on March 6th. You can see the peak standing out on that day. The line graph is almost constant with little to no flauctuations. The two lines denoting transactions and store_nbr are almost parallel to each other, this indicates that both these attributes are directly proportional to each other.

A. Just click on sum of transactions and change it to count and explain why it decreases.

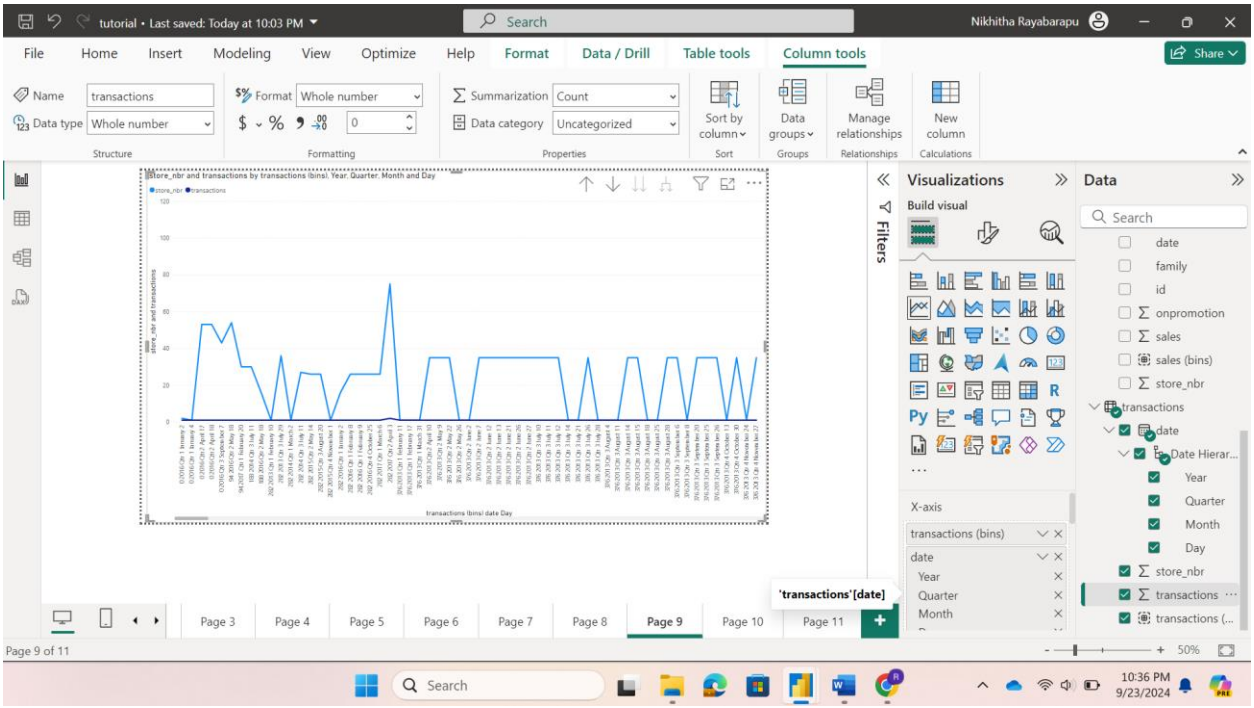

Fig 4.5

Explanation:

The sum aggregate will return the sum of all the transactions and the count returns the no. of transactions. Generally, values of sum aggregate is larger with higher no. of observations, but count is always less than the sum. That's why when we plot by taking "count" instead of "sum" the values lessens. It has also effected the "store_nbr" attribute.

Question 5:

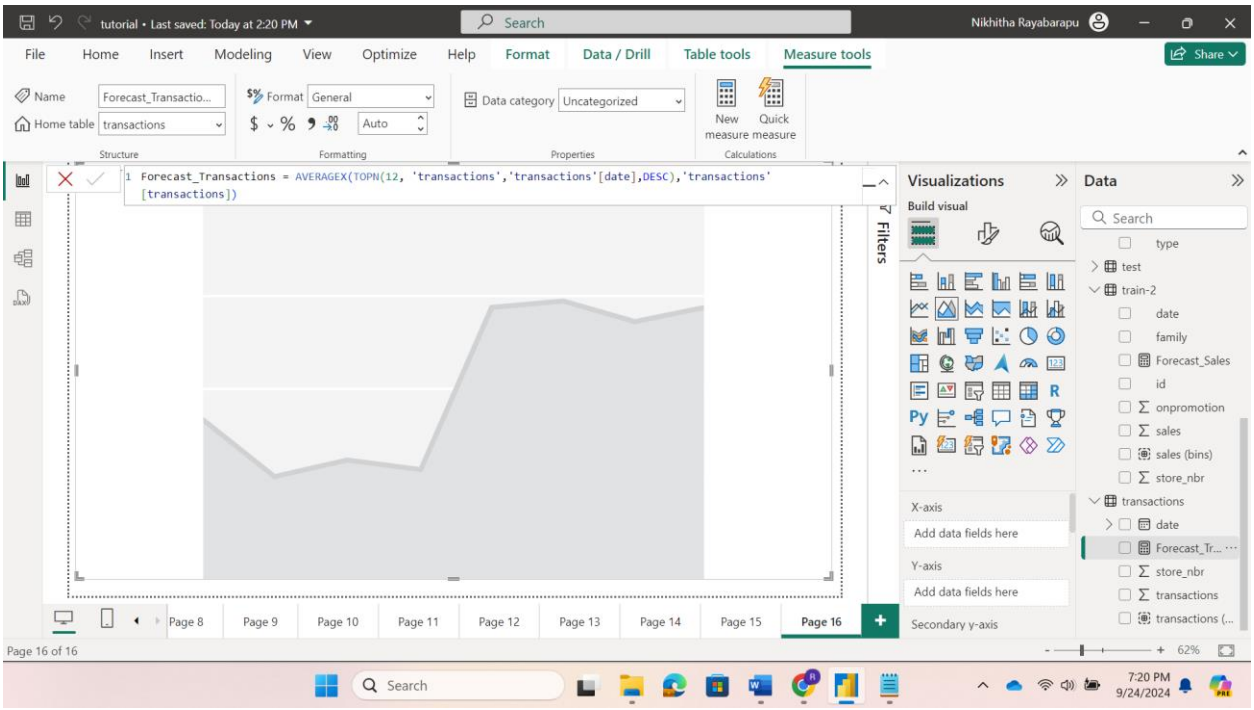

Fig 5.1

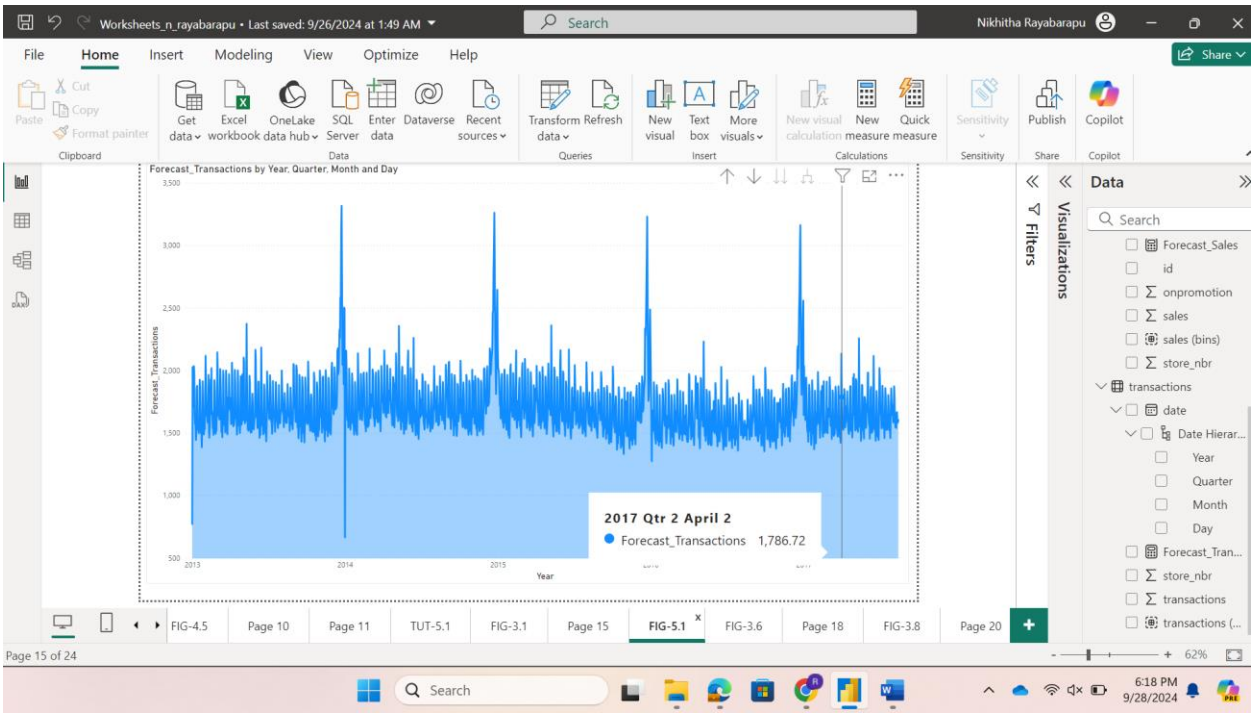

Fig 5.2

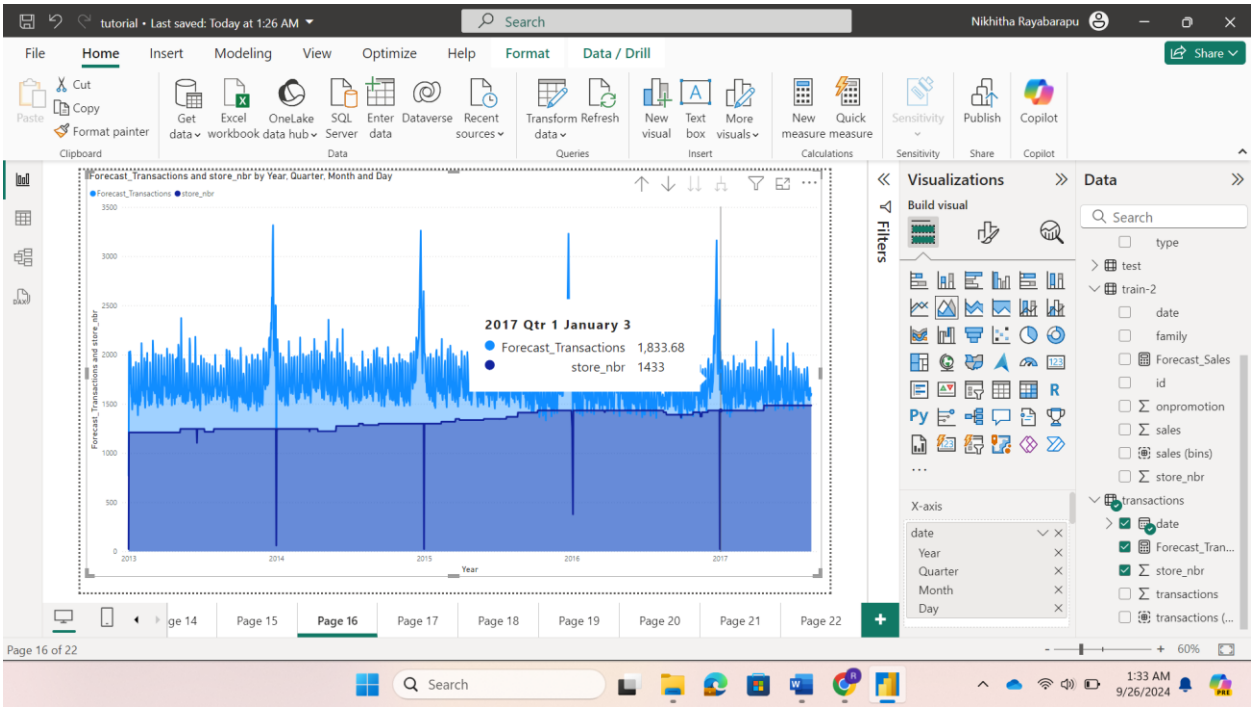

Fig 5.3

Answer the Following Questions

A. Explain how Time series forecasting is important and why.

Ans: Time series forecasting plays a critical role in assisting businesses and organizations in making well-informed decisions by predicting future values based on previous data. By examining patterns including trends, seasonality, and cyclical behavior, time series forecasting offers valuable insights into upcoming events.

B. Explain your Complete Understanding of the Activity in 250- 300 words.

In this tutorial, I was able to acquire the knowledge necessary to utilize the Power BI Desktop to load, transform, and visualize data. I have learned how to use the power query editor in Power BI. I have learned basic data manipulation techniques like sorting and filtering rows, creating queries, and creating groups by columns in task one. Coming to task 2, I have learned to create all possible and important relationships between different attributes in the given tables. I have learned to visualize various graphs based on different attributes end tables some tasks 3. It also involved looking at the visual table to find insights. Task 4 involved creating attributes and grouping the attributes. From Task 5, I have understood how to utilize the time-series forecasting process. I have understood that these capabilities are essential to data analysis for businesses across several sectors.