

# 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.

The screenshot shows the Microsoft Power Query Editor interface. A table is displayed with three columns: 'permitdate', 'nbr\_str', and 'transactions'. The 'nbr\_str' column is highlighted with a yellow border. In the 'APPLIED STEPS' pane on the right, the 'Renamed Columns' step is listed under the 'Changed Type' section. The 'Name' field in the properties pane is set to 'transactions'. The status bar at the bottom indicates 'PREVIEW DOWNLOADED AT 10:14 PM' and shows the date '9/21/2024' and time '11:14 PM'.

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:

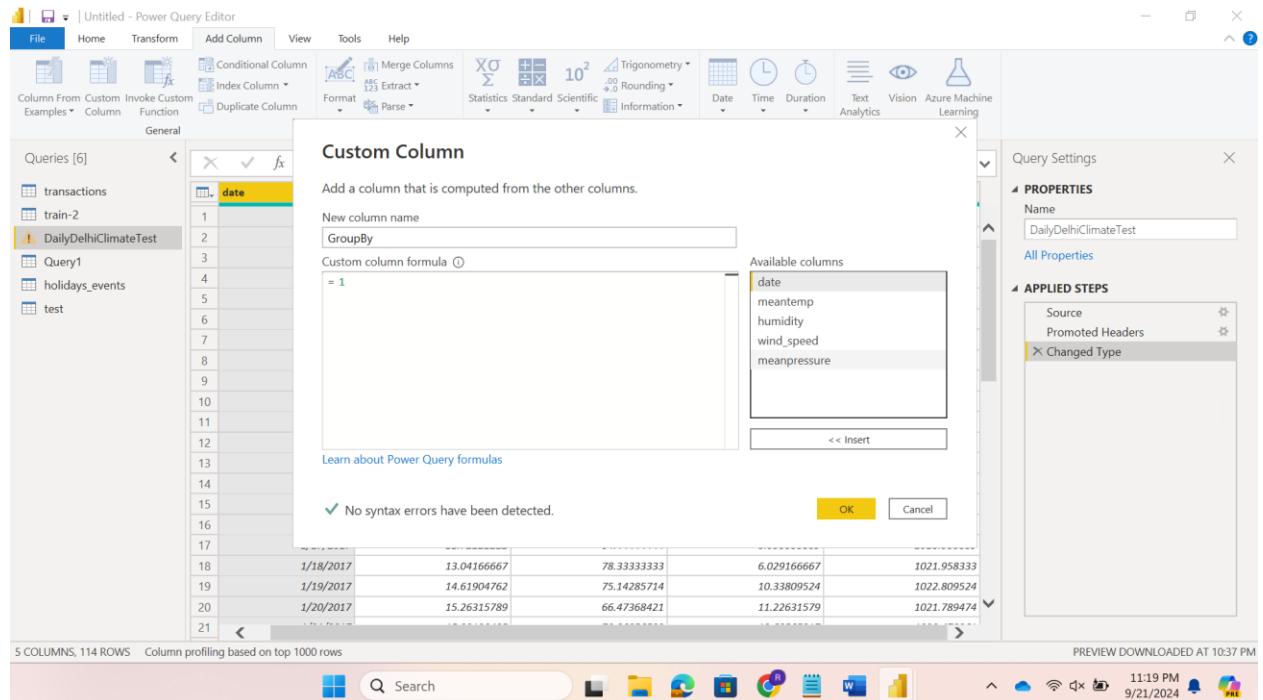


Fig 1.2

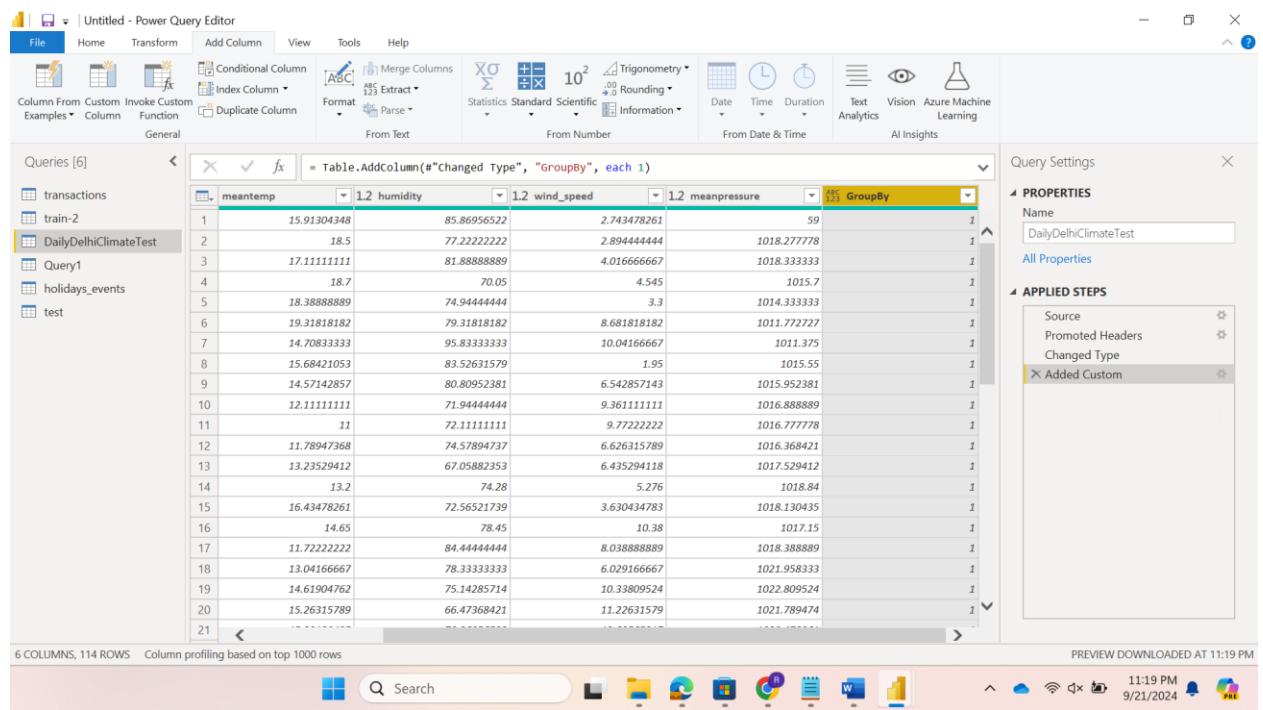


Fig 1.3

The screenshot shows the Power Query Editor interface. A table named 'meantemp' is displayed with four columns: 'meantemp', '1.2 humidity', '1.2 wind\_speed', and '1.2 meanpressure'. The 'humidity' column is currently selected. A context menu is open at the top of the column, with 'Group By...' highlighted under the 'Transform' section. The 'Properties' pane on the right shows the query name as 'DailyDelhiClimateTest'. The status bar at the bottom indicates 'PREVIEW DOWNLOADED AT 11:19 PM'.

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.

The screenshot shows the Power Query Editor interface with the 'Group By' dialog box open. The dialog box allows specifying columns to group by and one or more outputs. In this case, it is grouped by 'humidity'. Two aggregations are defined: 'MeanValue' using the 'Average' operation and 'MedianValue' using the 'Median' operation. The 'OK' button is visible at the bottom right of the dialog box. The status bar at the bottom indicates 'PREVIEW DOWNLOADED AT 11:19 PM'.

Fig 1.5

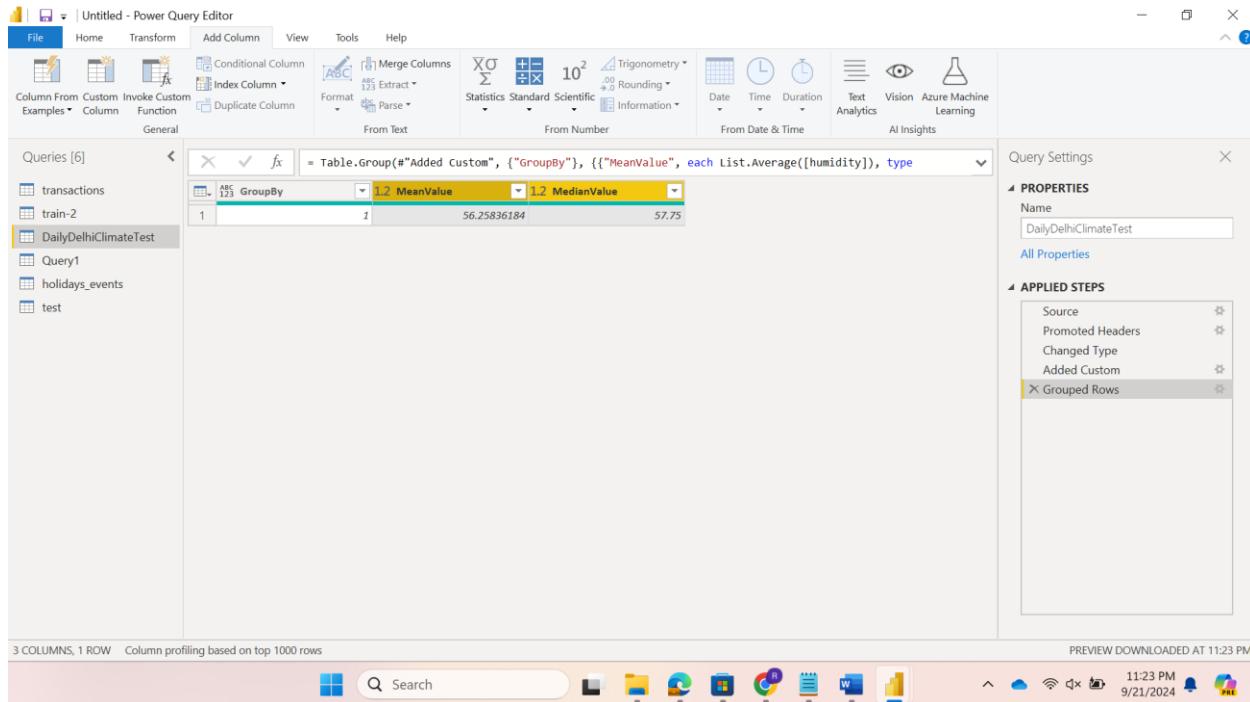


Fig 1.6

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

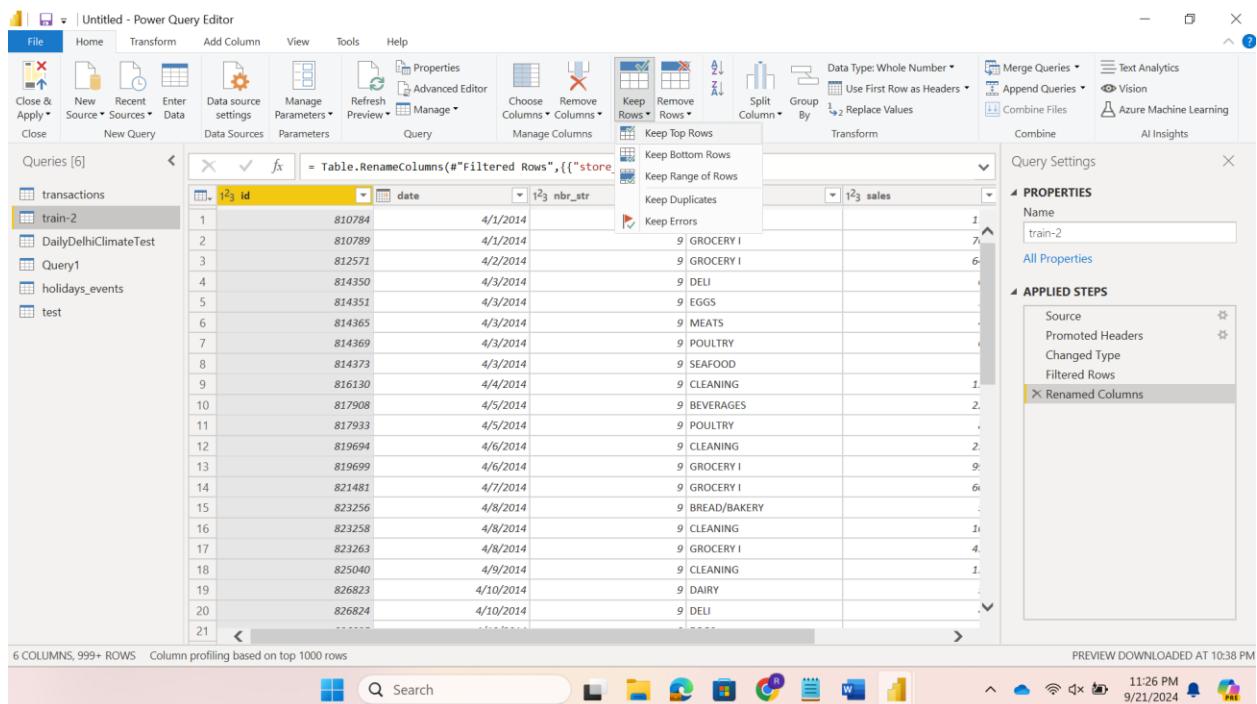


Fig 1.7

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

The screenshot shows the Power Query Editor interface with the 'transactions' query selected. The 'id' column is highlighted. A context menu is open over the 'id' column header, with the 'Number Filters' option selected. A sub-menu is displayed, showing various comparison operators: Equals..., Does Not Equal..., Greater Than..., Greater Than Or Equal To..., Less Than..., Less Than Or Equal To..., and Between... . The 'Less Than...' option is highlighted. The main table view shows data rows with columns: date, nbr\_str, family, and sales.

Fig 1.8

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

The screenshot shows the Power Query Editor interface with the 'train-2' query selected. The 'id' column is highlighted. A 'Filter Rows' dialog is open, showing the condition: 'And/Or' set to 'And', 'Column' set to 'id', 'Operator' set to 'is less than', and 'Value' set to '700000'. Below the dialog, the main table view shows data rows with columns: date, nbr\_str, family, and sales.

Fig 1.9

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

The screenshot shows the Microsoft Power Query Editor interface. The 'File' tab is selected. In the 'Queries [6]' pane, 'train-2' is the active query. The main area displays a table with columns: id, date, nbr\_str, family, and sales. A formula bar at the top shows the formula: = Table.SelectRows(#"Kept Last Rows", each [id] < 700000). The 'APPLIED STEPS' pane on the right lists the steps taken: Source, Promoted Headers, Changed Type, Filtered Rows, Renamed Columns, Sorted Rows, Kept Last Rows, and Filtered Rows1 (which is highlighted). The status bar at the bottom indicates '1.71 MB FROM TRAIN-2.CSV'.

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:

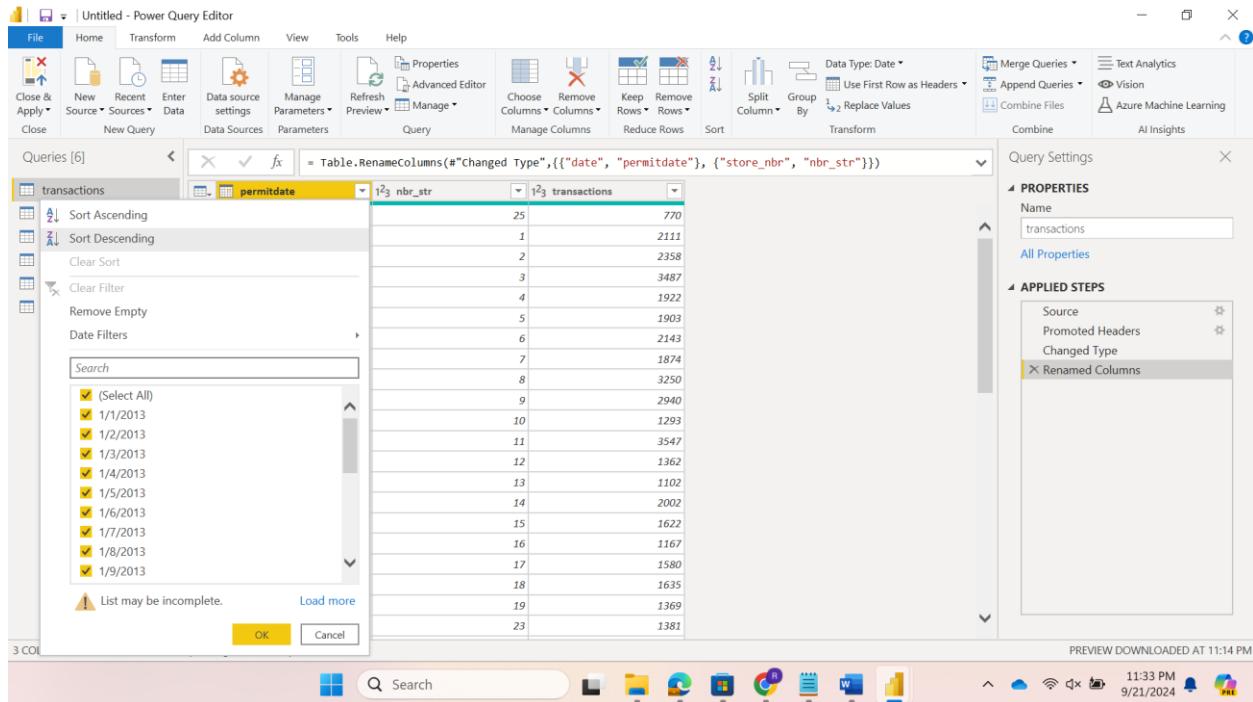


Fig 1.11

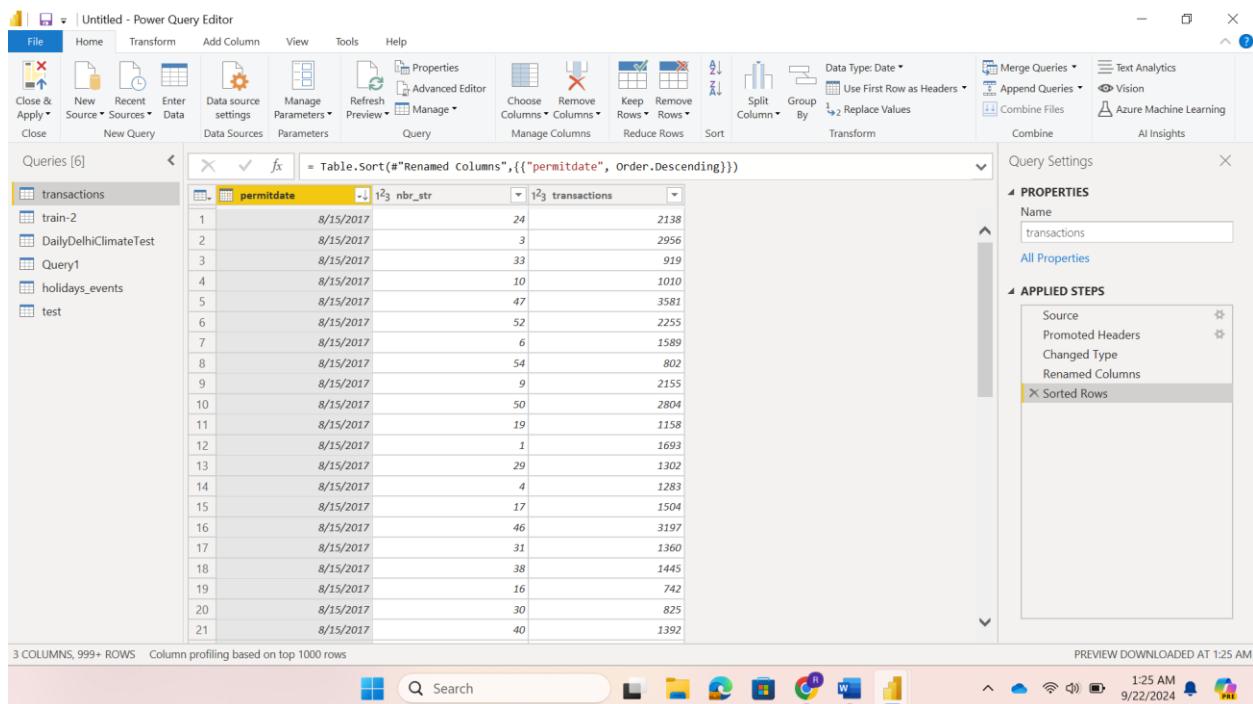


Fig 1.12

**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.**

From table: train-2

date	family	id	onpromotion	sales	sales (bins)	store_nbr
Friday, May 1...	FROZEN FOO...	890945	2	220	0	8
Friday, June 0...	FROZEN FOO...	928367	2	199	0	8
Saturday, Jun...	FROZEN FOO...	930149	2	333	0	8

To table: test

date	family	id	onpromotion	store_nbr
Wednesday, ...	BOOKS	3000892	0	1
Wednesday, ...	BOOKS	3000925	0	10
Wednesday, ...	BOOKS	3000958	0	11

Cardinality: One to one (1:1)      Cross-filter direction: Both

Make this relationship active       Assume referential integrity

Save      Cancel

Fig 2.1

+ New relationship      Autodetect

From: table (column) ↑      Relationship      To: table (column) ↓

train-2 (id)      1 —> 1      test (id)

Cardinality (1)

One to one

Many to one

Many to many

Cross-filter direction

Close

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.

The screenshot shows the 'Manage relationships' dialog in Power BI. It lists 15 possible relationships across various tables. Out of these, 4 are marked as 'Active' and 11 are 'Inactive'. The active relationships are:

- holidays\_events (date) ↔ test (date)
- transactions (date) ↔ holidays\_events (type)
- test (date) ↔ DailyDelhiClimateTest (meantemp)
- train-2 (sales) ↔ transactions (transactions)

Fig 2.3

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

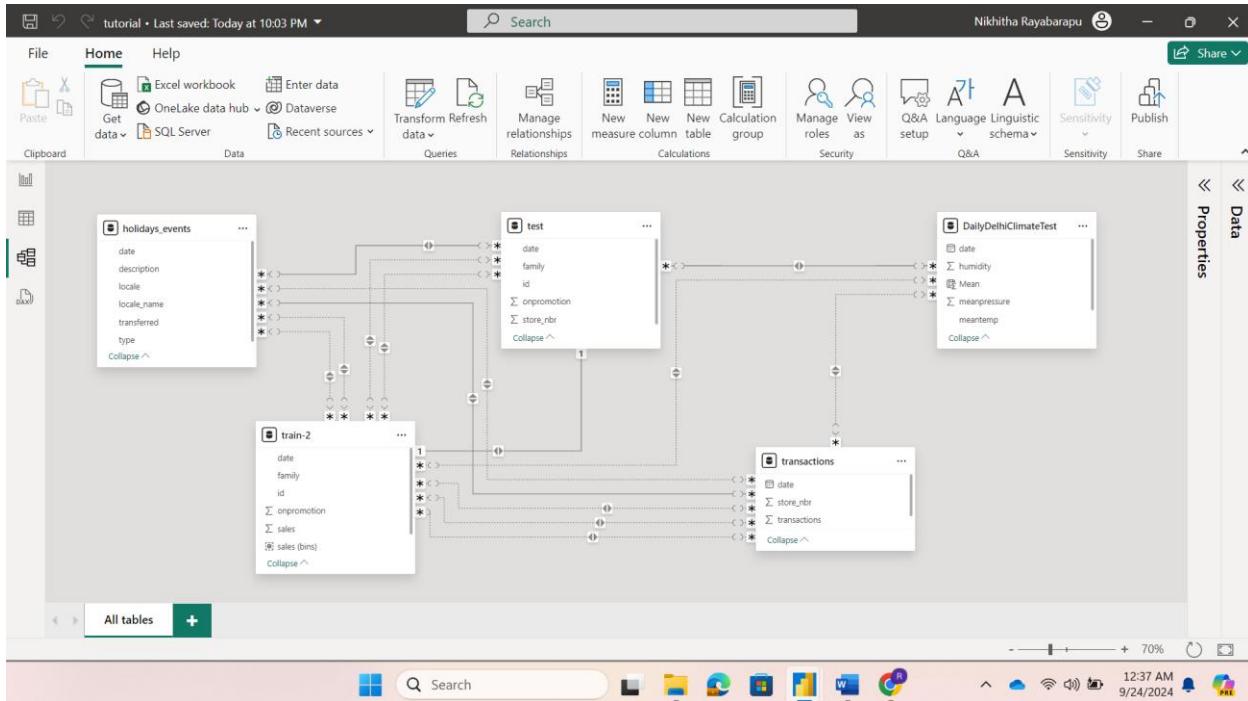


Fig 2.4

**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

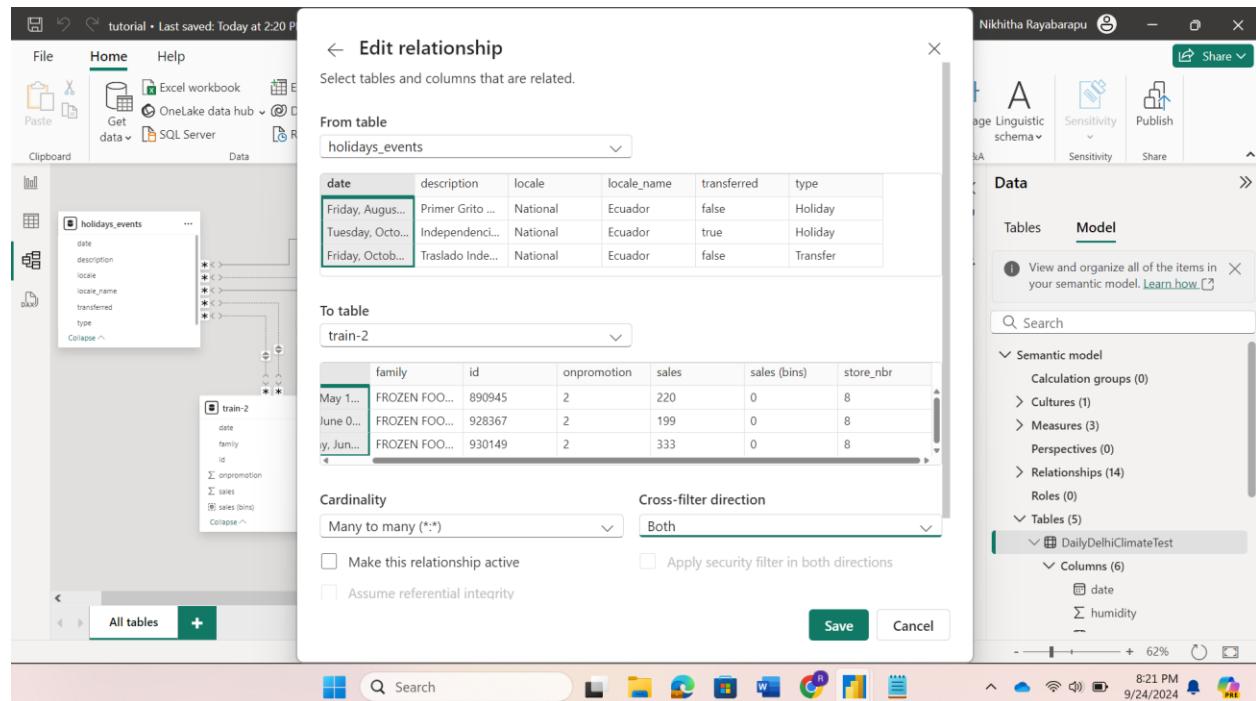


Fig 2.5

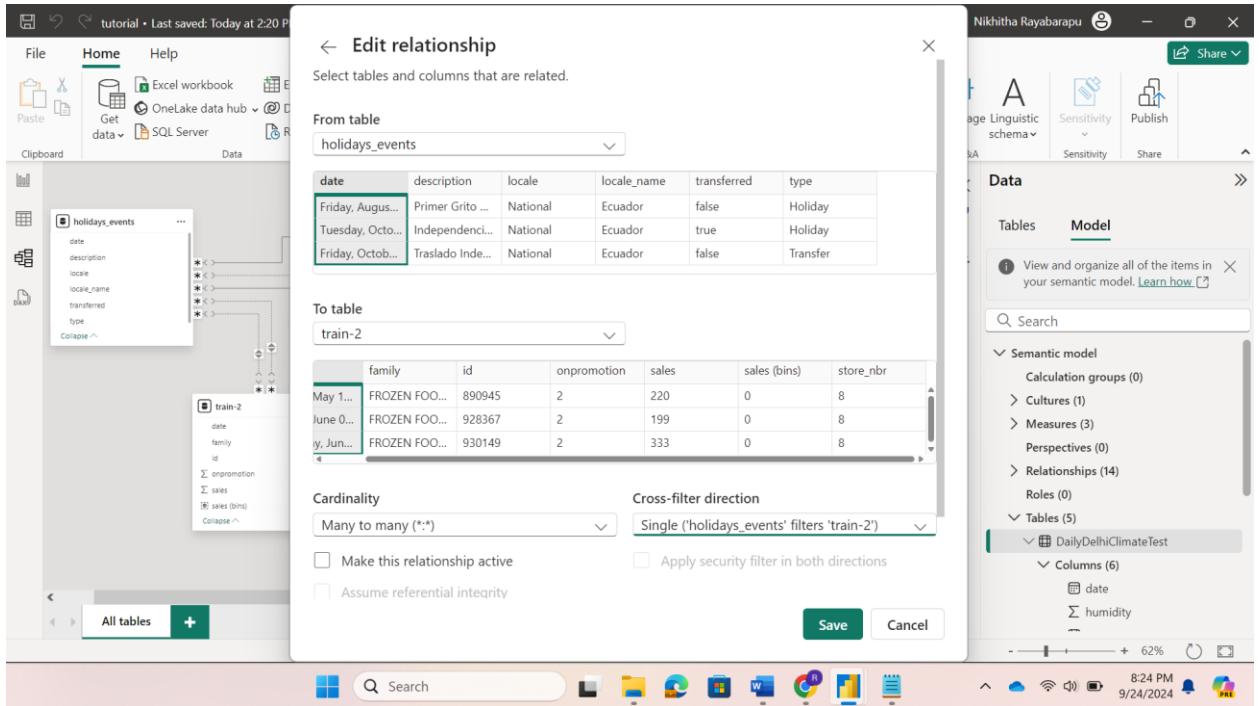


Fig 2.6

## 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.

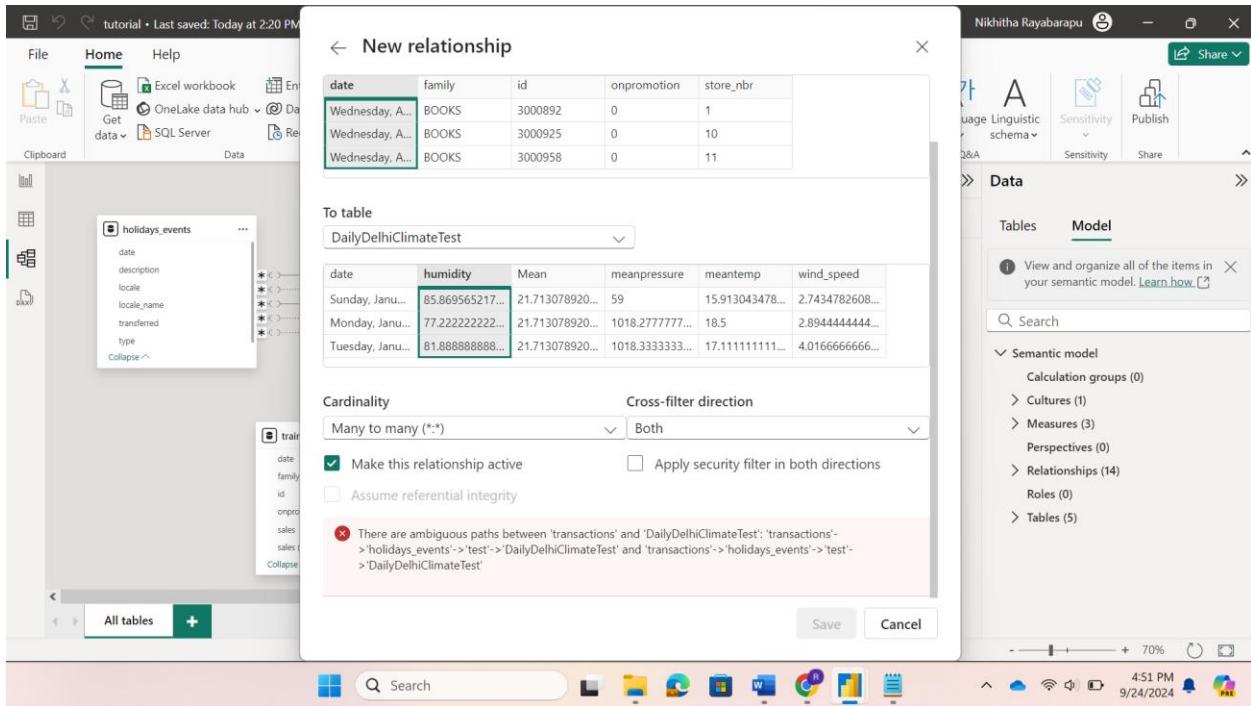


Fig 2.7

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 table.

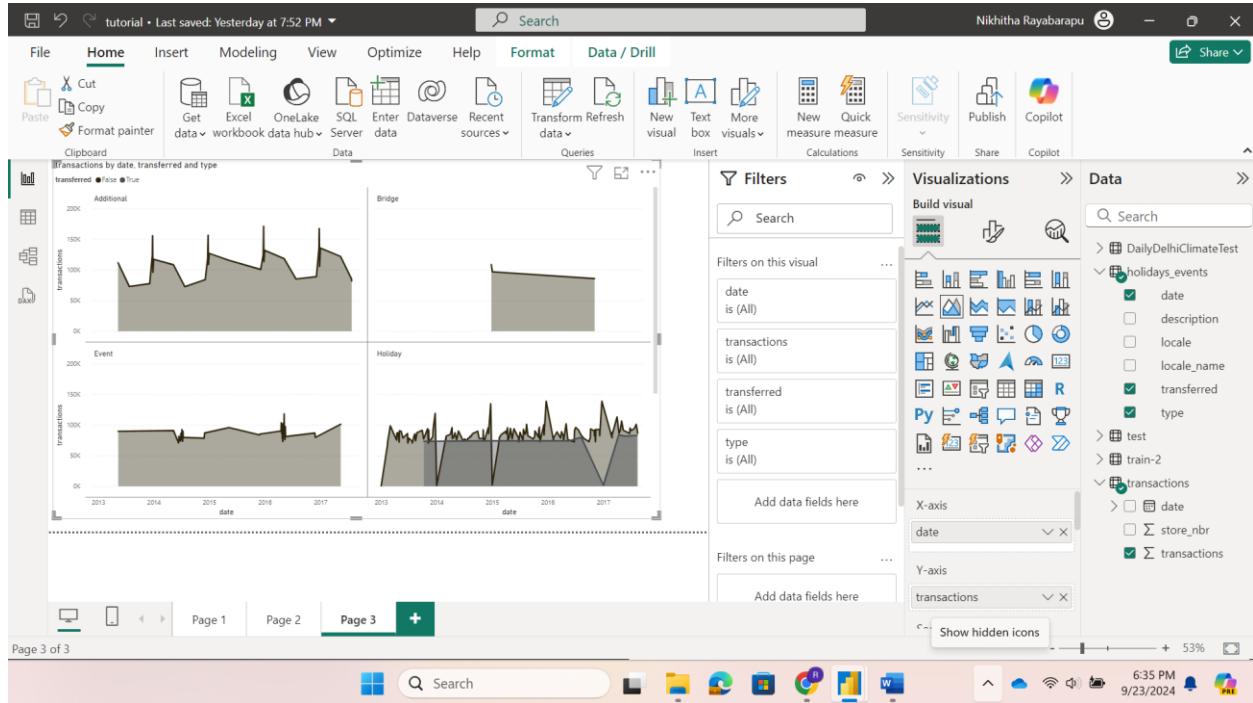


Fig 3.1

### 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.

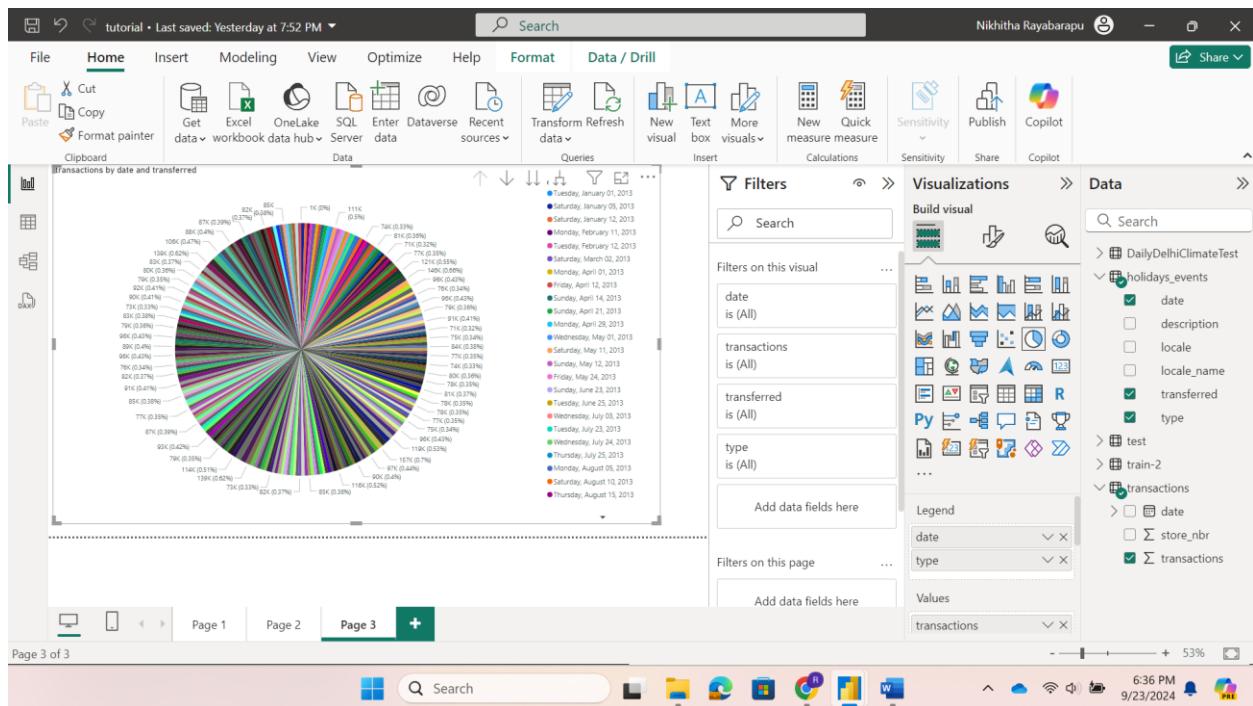


Fig 3.2

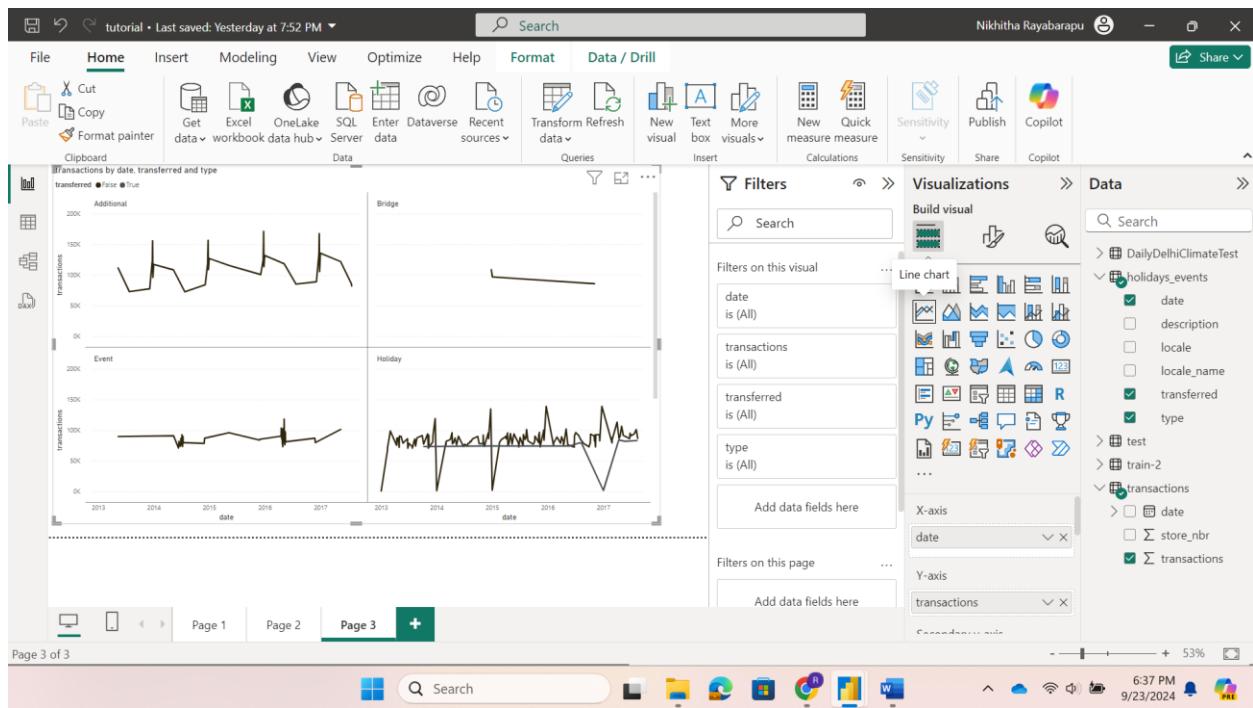


Fig 3.3

**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.**

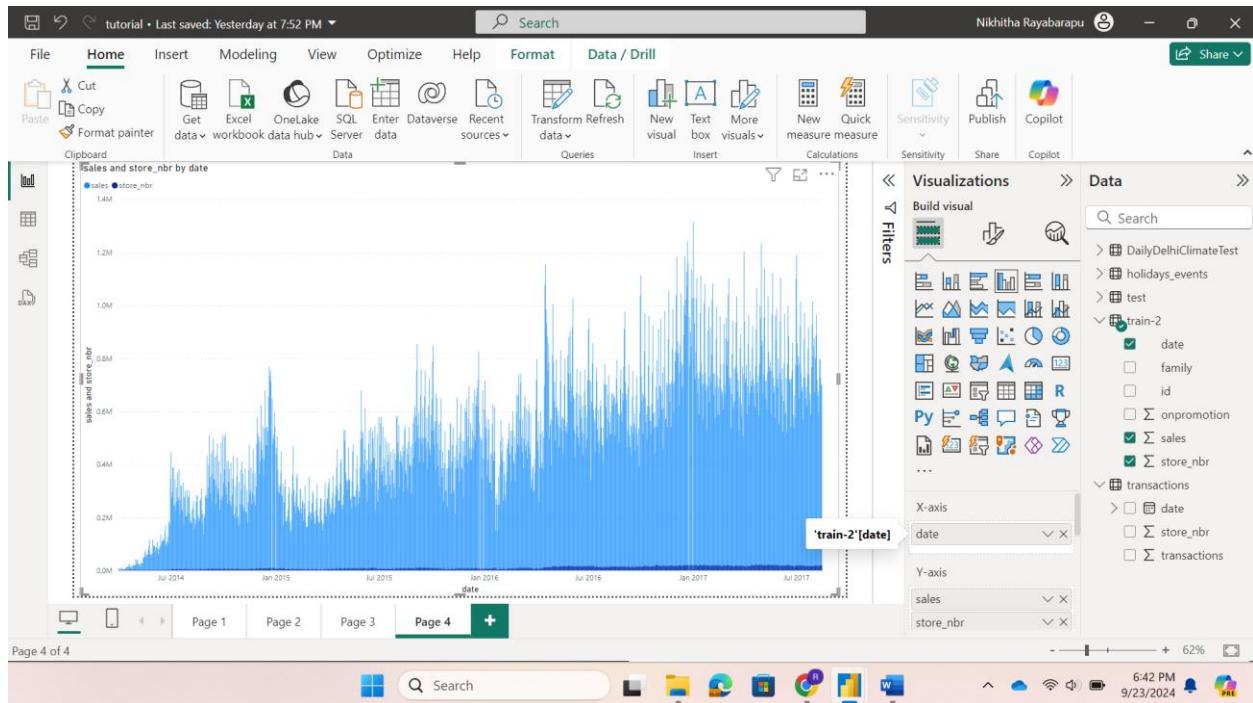


Fig 3.4

### 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.**

The screenshot shows the Power BI desktop interface. The ribbon is at the top with tabs: File, Home, Insert, Modeling, View, Optimize, Help, Format, and Data / Drill. The 'Data / Drill' tab is selected. On the left, there's a 'Visual table' icon and a dropdown menu for 'Apply drill down filters to' set to 'Entire page'. The main area displays a table with columns: Show, Interactions, and Drill actions. The data rows show various dates and counts (e.g., False, Additional, Saturday, December 21, 2013; 47). A total row at the bottom shows 83488. Below the table is a section titled 'transferred, type, date' with a list of items. To the right is a 'Filters' pane with sections for 'Visualizations' and 'Data'. The 'Data' pane shows a search bar and a list of columns: locale\_name, transferred, type, test, train-2, date, family, id, onpromotion, sales, store\_nbr, transactions, and several summation columns like  $\Sigma$  sales and  $\Sigma$  store\_nbr. The status bar at the bottom indicates 'Page 5 of 11' and the system clock '12:07 AM 9/24/2024'.

Fig 3.5

This screenshot shows the Power BI desktop interface with the 'Insert' tab selected in the ribbon. The main area displays a paginated table with columns: date, sales, and store\_nbr. The data shows daily sales figures from April 1, 2014, to April 22, 2014. A total row at the bottom shows 613892547 and 12178182. Below the table is a 'Pages' section with buttons for Page 1 through Page 6, and a 'Drill through' button. To the right is a 'Filters' pane similar to Fig 3.5, showing the same column list and expanded sections for 'test', 'train-2', and 'transactions'. The status bar at the bottom indicates 'Page 6 of 6' and the system clock '6:50 PM 9/23/2024'.

Fig 3.6

## A. What is the Highest sales in 2016 and How much compared to the 2<sup>nd</sup> Highest?

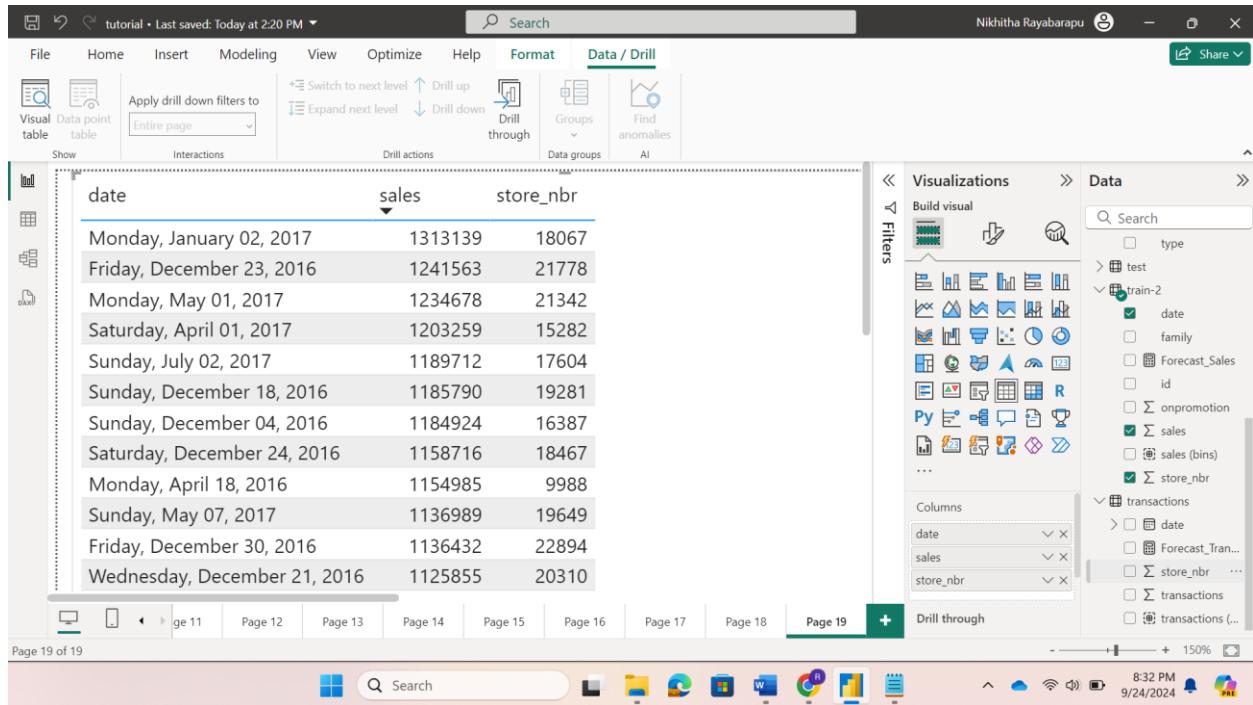


Fig 3.7

The highest sales in 2016 were 1241563. 2<sup>nd</sup> 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)

The screenshot shows a Microsoft Power BI interface. The top navigation bar includes File, Home, Insert, Modeling, View, Optimize, Help, Format, and Data / Drill. The Data / Drill tab is selected. On the left, there's a 'Visual table' icon and a 'Data point table' icon. A dropdown menu titled 'Apply drill down filters to' is open, showing 'Entire page'. Below this is a table with columns: 'Show' (checkbox), 'Interactions' (text), 'Date' (text), and 'Drill actions' (button). The table lists several rows of data. To the right of the table is a 'Filters' section and a 'Visualizations' pane containing various chart icons. The 'Data' pane on the far right shows a search bar and a list of columns: locale\_name, transferred (checked), type (checked), date, family, id, Σ onpromotion, sales, sales (bins), Σ store\_nbr, transactions (checked), date (checked), Σ store\_nbr (checked), Σ transactions (checked). At the bottom, the status bar shows 'Page 5 of 11' and the system tray shows the date and time as 9/24/2024 at 12:07 AM.

Fig 3.8

The screenshot shows a Microsoft Power BI interface with the 'Insert' tab selected. The top navigation bar includes File, Home, Insert, Modeling, View, Optimize, Help, Format, and Data / Drill. The Data / Drill tab is selected. On the left, there are sections for 'Pages' (New page, New visual, Visuals), 'Visuals' (New visual, More visual, Q&A, Key influencers, Decomposition, Narrative, Power Platform), and 'AI visuals' (Text box, Buttons, Shapes, Image, Sparkline, Sparklines). Below these is a table with columns: date, sales, and store\_nbr. The table lists data from April 1, 2014, to April 22, 2014, with totals at the bottom. To the right of the table is a 'Filters' section and a 'Visualizations' pane. The 'Data' pane on the far right shows a search bar and a list of columns: date, family, id, Σ onpromotion, Σ store\_nbr, test (checked), date, family, id, Σ onpromotion, Σ store\_nbr, train-2 (checked), date, family, id, Σ onpromotion, Σ sales (checked), Σ store\_nbr (checked), Σ transactions (checked), date (checked), Σ store\_nbr (checked), Σ transactions (checked). At the bottom, the status bar shows 'Page 6 of 6' and the system tray shows the date and time as 9/23/2024 at 6:50 PM.

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.

The screenshot shows the 'Groups' dialog box in Power BI. The 'Name' field is set to 'transactions (bins)' and the 'Field' dropdown is set to 'transactions'. Under 'Group type', 'Bin' is selected, and under 'Bin type', 'Size of bins' is selected. The 'Min value' is 5 and the 'Max value' is 8359. Below this, a note says 'Binning splits numeric or date/time data into equally sized groups. Enter bin size.' The 'Bin size' input field contains '94'. At the bottom right of the dialog are 'OK' and 'Cancel' buttons. To the right of the dialog, the 'Data' pane is visible, showing various fields like locale\_name, transferred, type, date, family, id, etc. The 'Visualizations' pane is also partially visible.

Fig 4.1

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

With Quarter:

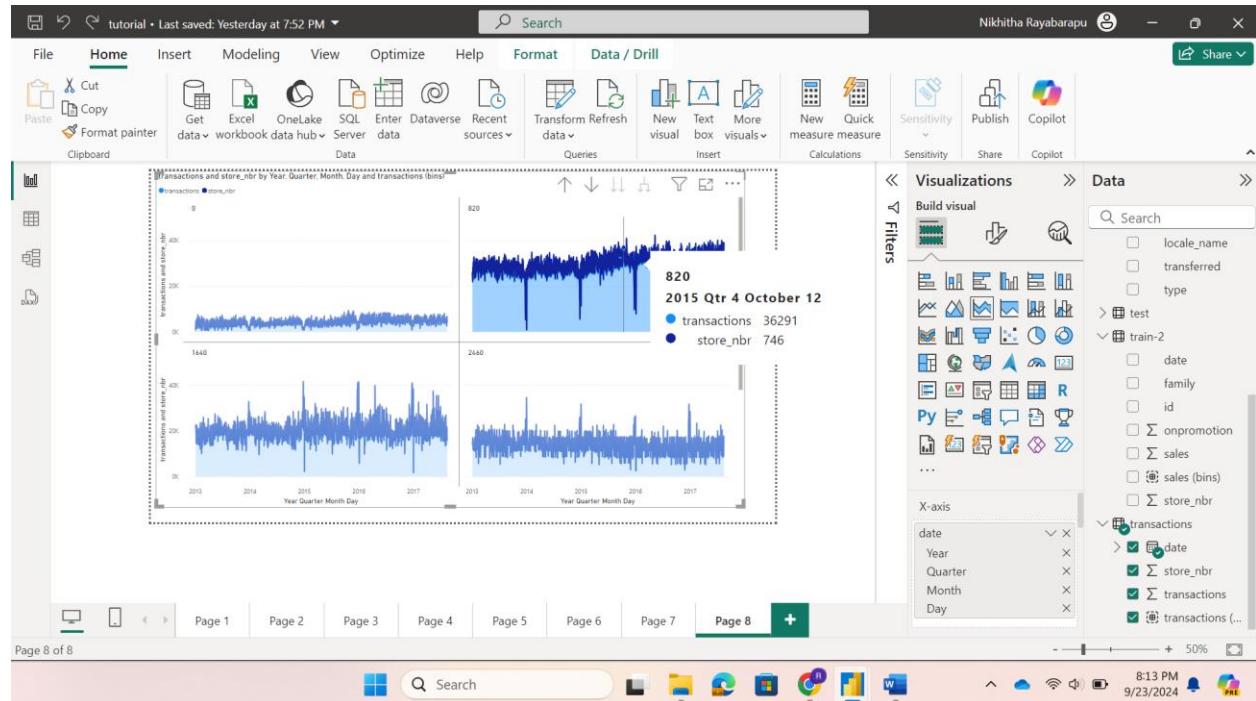


Fig 4.2

Without Quarter:

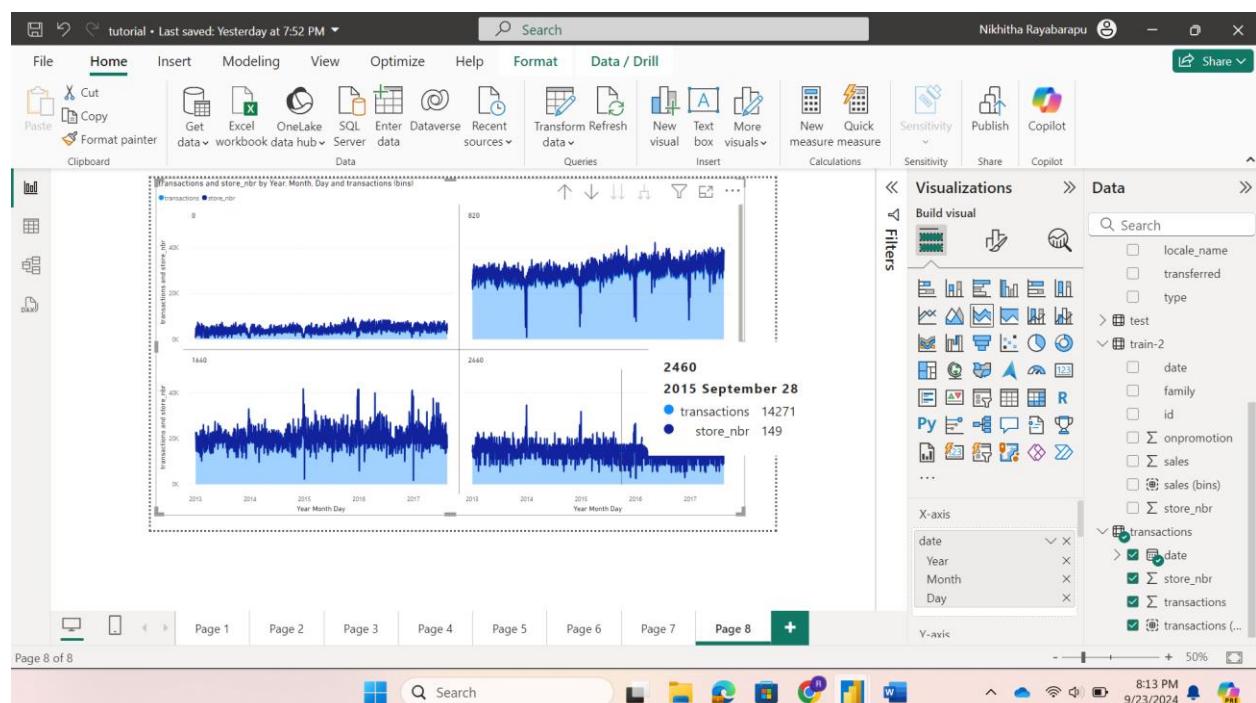


Fig 4.3

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

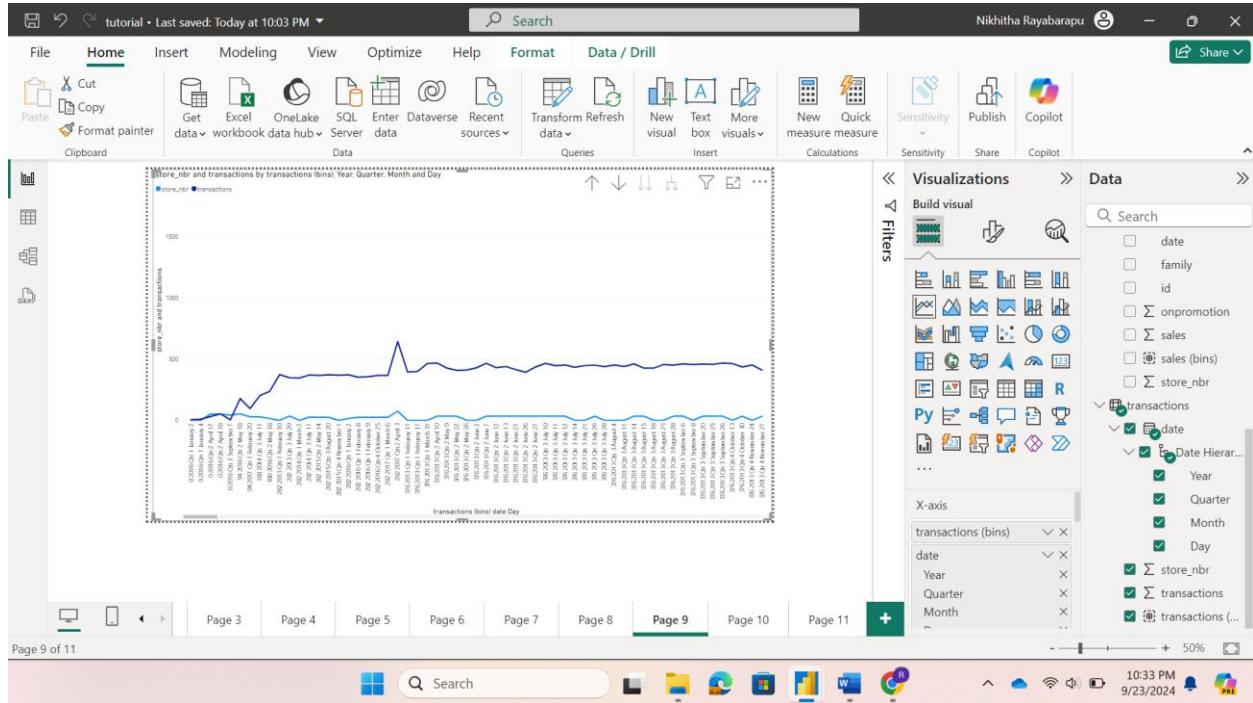


Fig 4.4

**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 6<sup>th</sup>. You can see the peak standing out on that day. The line graph is almost constant with little to no fluctuations. 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.**

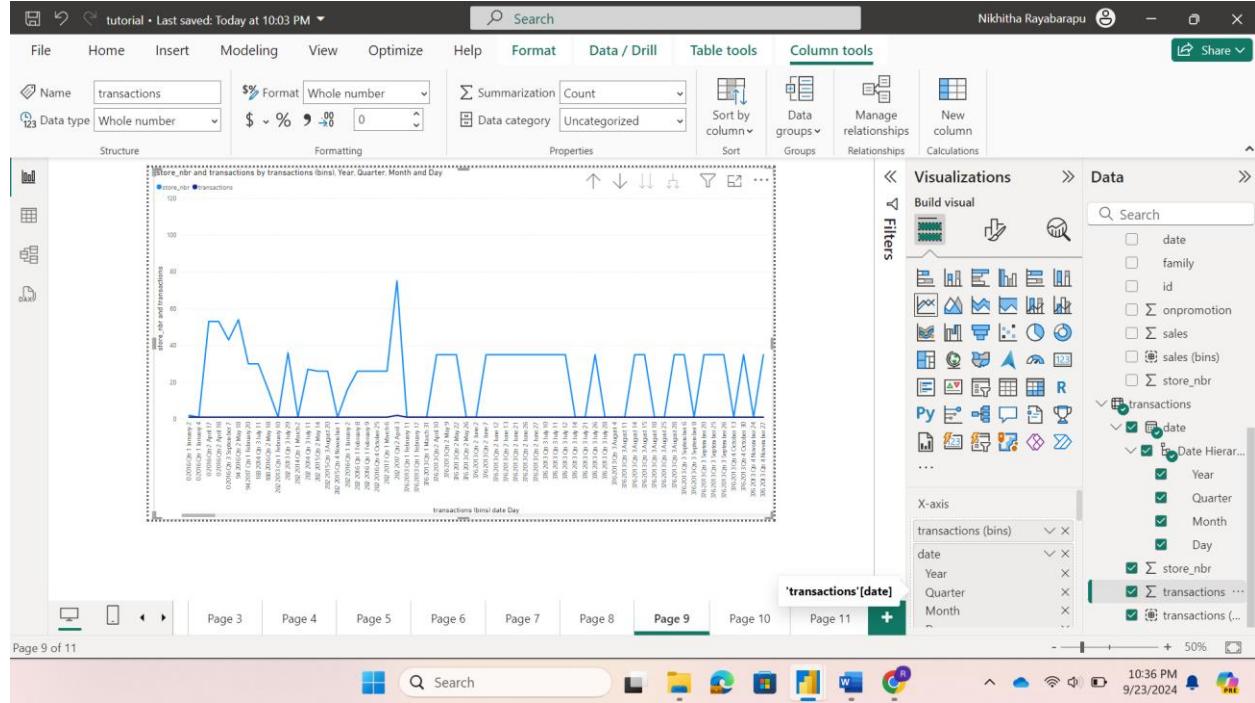


Fig 4.5

**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:

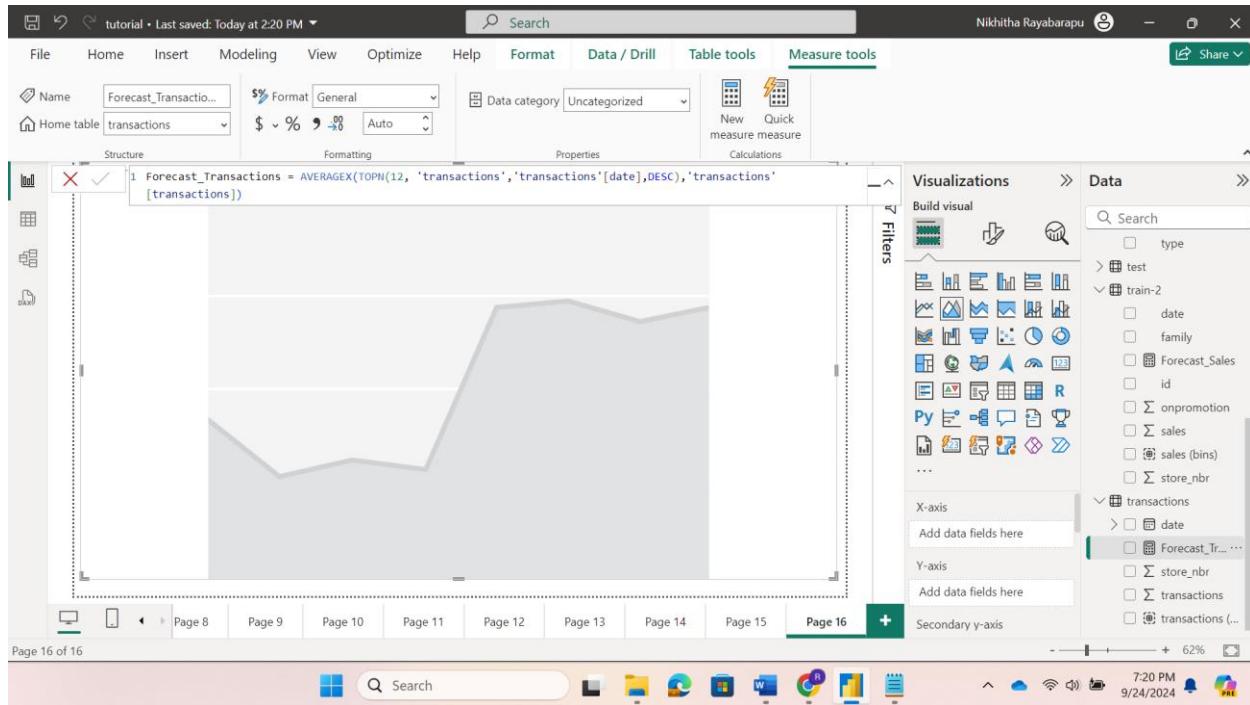


Fig 5.1

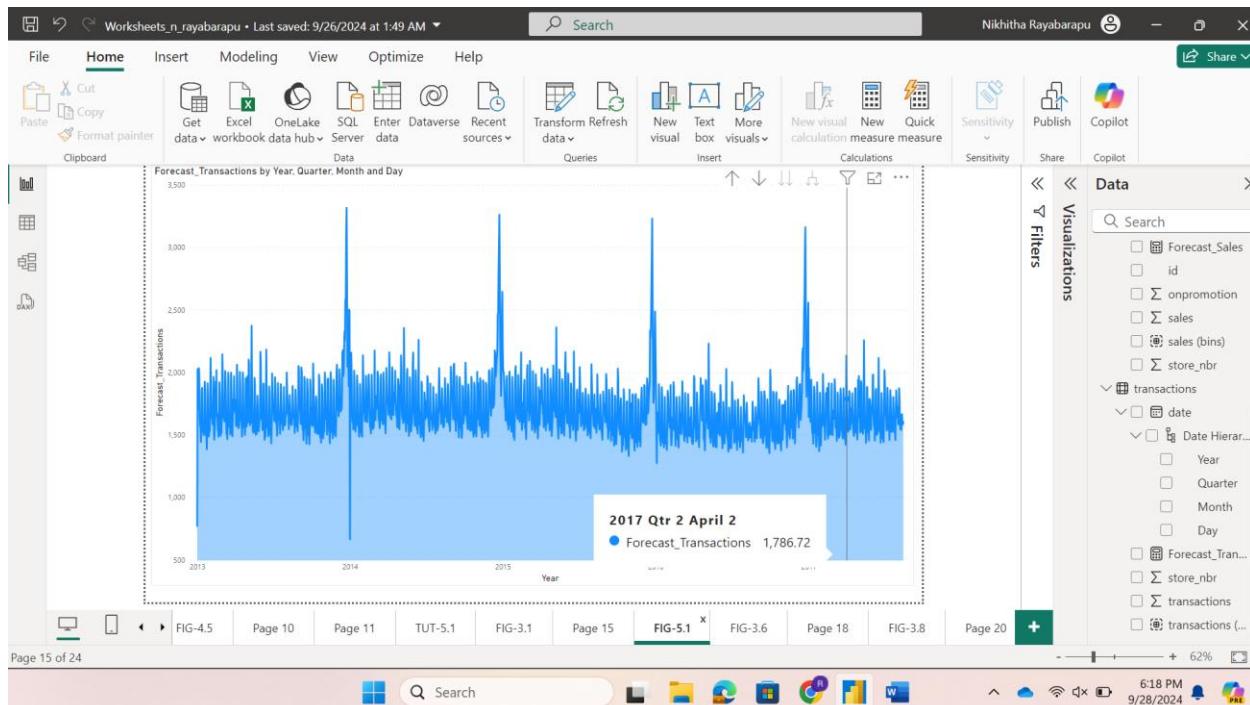


Fig 5.2

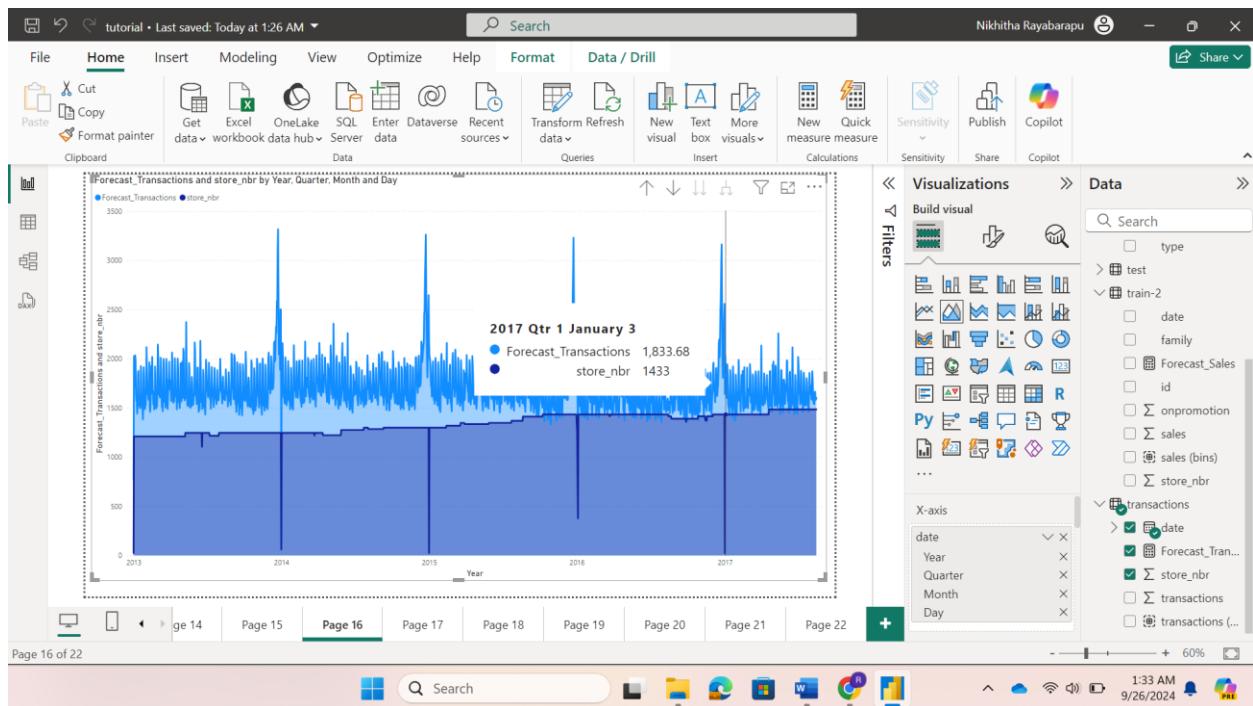


Fig 5.3

## 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 and 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.