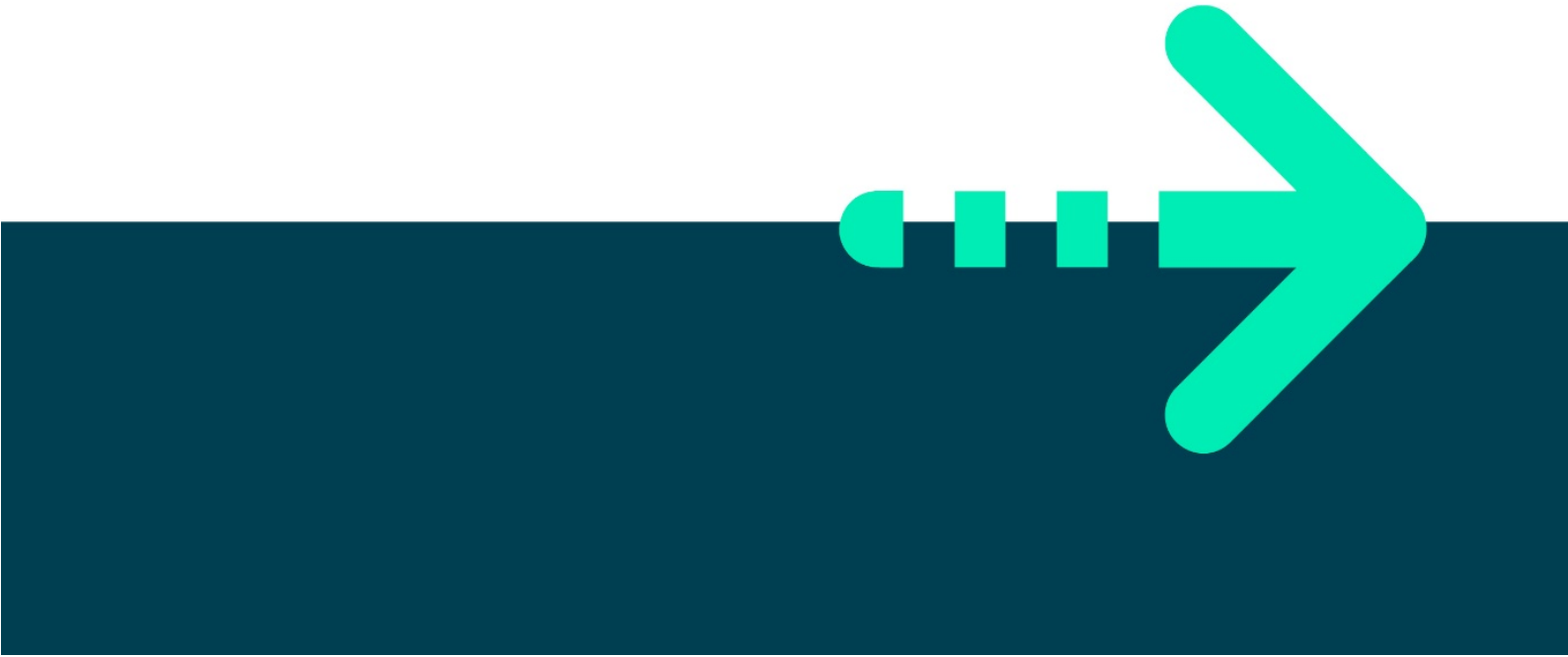
Activity 2: From web data to insights

Learner Guide



Contents

How to use this workbook....2
Introduction....3
KSBs....3
Scenario....3
Activity....4

How to use this workbook

Activity

Alongside this icon you'll find details of the group/individual activity or a point for everyone to discuss.

Useful tool

This icon indicates a technique that will help you put what you learn into practice.

Important idea or concept

Generally, this icon is used to draw your attention to ideas that you need to understand by this point in the course. Let your trainer know if you do not understand or see the relevance of this idea or concept.

Helpful hint

This icon guides you to tips or hints that will help you avoid the common pitfalls or to show you how to increase your effectiveness or efficiency in practising what you have learnt.

Key point

This icon is used to indicate something that practitioners in this field should know. It's likely to be one of the major things to remember from the course, so check you do understand these key points.

Reference material

When we have only touched briefly on a topic, this icon highlights where to look for additional information on the subject. It may also be used to draw your attention to International or National Standards or Web addresses that have interesting collections of information.

Definition

Where a word with a very specific definition (or one that could be described as jargon) is introduced, this will highlight that a definition is provided.

Warning

This icon is used to point out important information that may affect you and your use of the product or service in question.

Introduction

In this activity, you'll be introduced to Power BI to facilitate you in extracting unstructured data from the web. The objectives are to gather the web data, manipulate it, identify, and explain trends and patterns.

Here, we are going to start looking at Microsoft's PowerBI, a popular Business Intelligence product. As we start this example and open PowerBI, you should recognise the Microsoft Office style layout, which is very similar to Excel. Then, as we work through, you should recognise some of the processes and techniques we have discussed in the discover (online) learning pathway.

KSBs

This activity will cover the following apprenticeship skills elements:

- **S3**: Summarise and explain gathered data
- **S6**: Use tools and techniques to identify trends and patterns in data.

Scenario

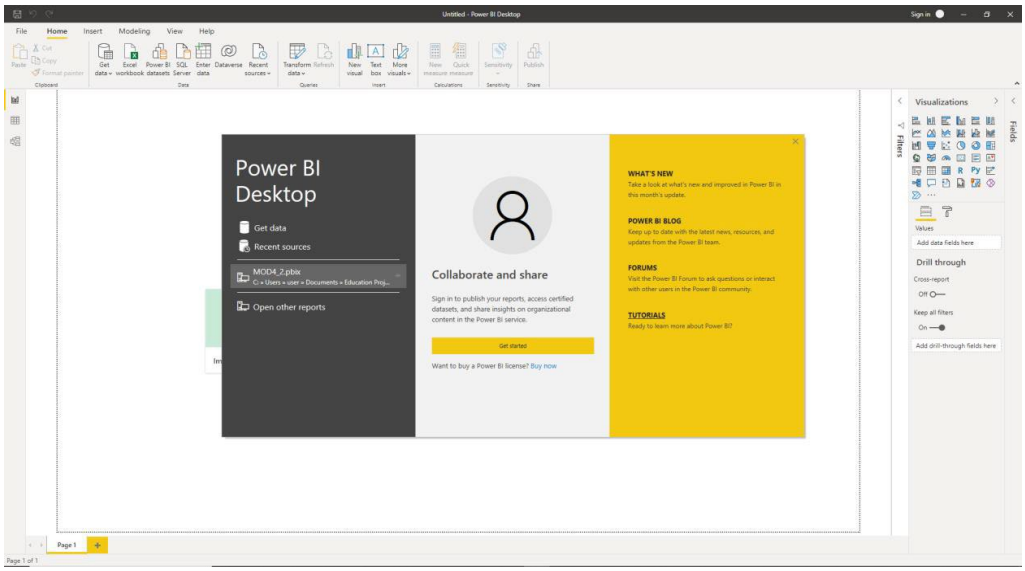
In a raffle at work, a friend has won a two night's stay in a London hotel of their choice. They would like to go to the theatre both nights that they are there and has sent you a link to a list of shows they would like to see. However, they are on a budget, so would like for you to advise them whereabouts in London they should stay in order to get the most out of their trip and to minimise travel costs while there. They would like to keep the cost down as much as possible but have set a fixed limit of £70.

Activity

Independent activity:

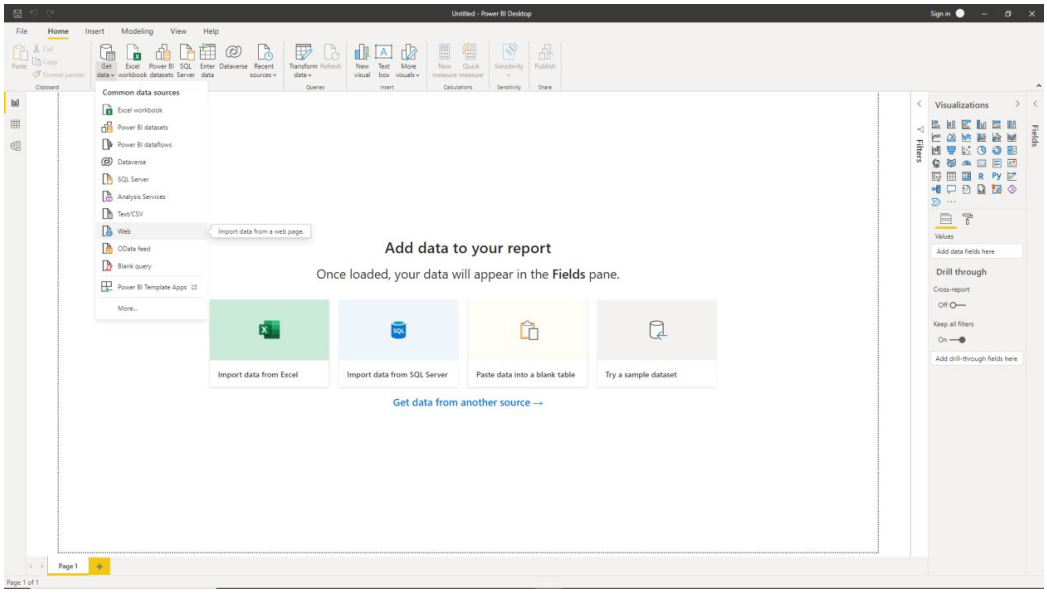
Extract web data

1. Open Microsoft PowerBI. We are going to close the dialog box that appears, but if you were to click Tutorials, you will find a series of courses giving you an in-depth exploration of PowerBI's functionality.



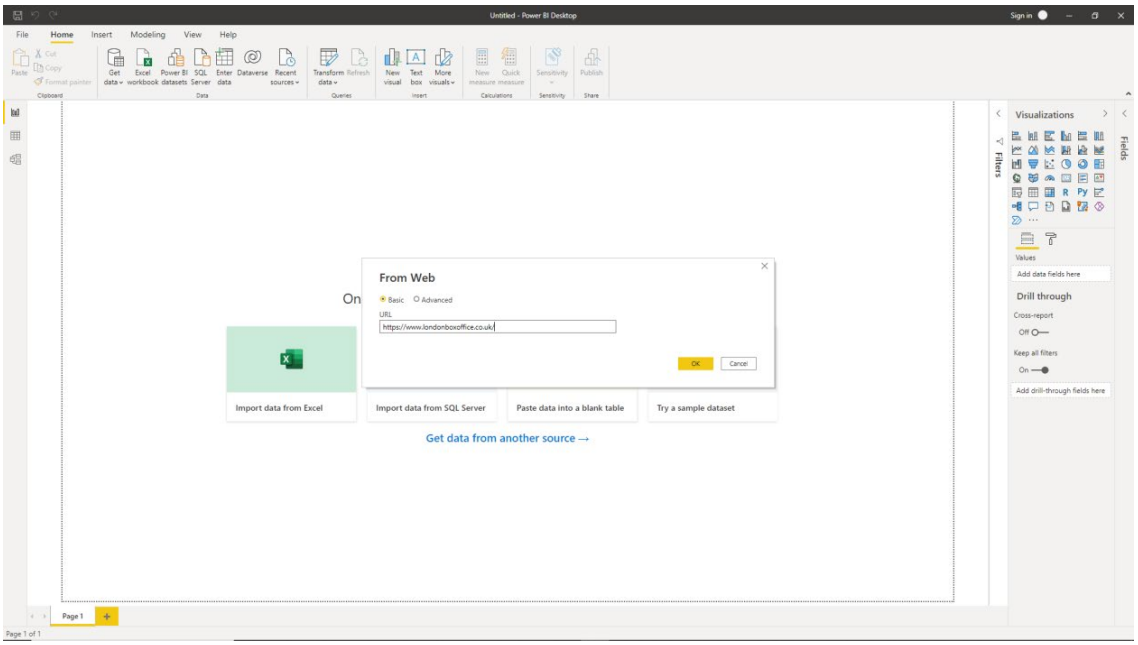
2. First of all, we are going to extract the unstructured data from the link that your friend has sent: https://www.londonboxoffice.co.uk/

3. Click the drop-down arrow below Home > Get Data. Select Web.

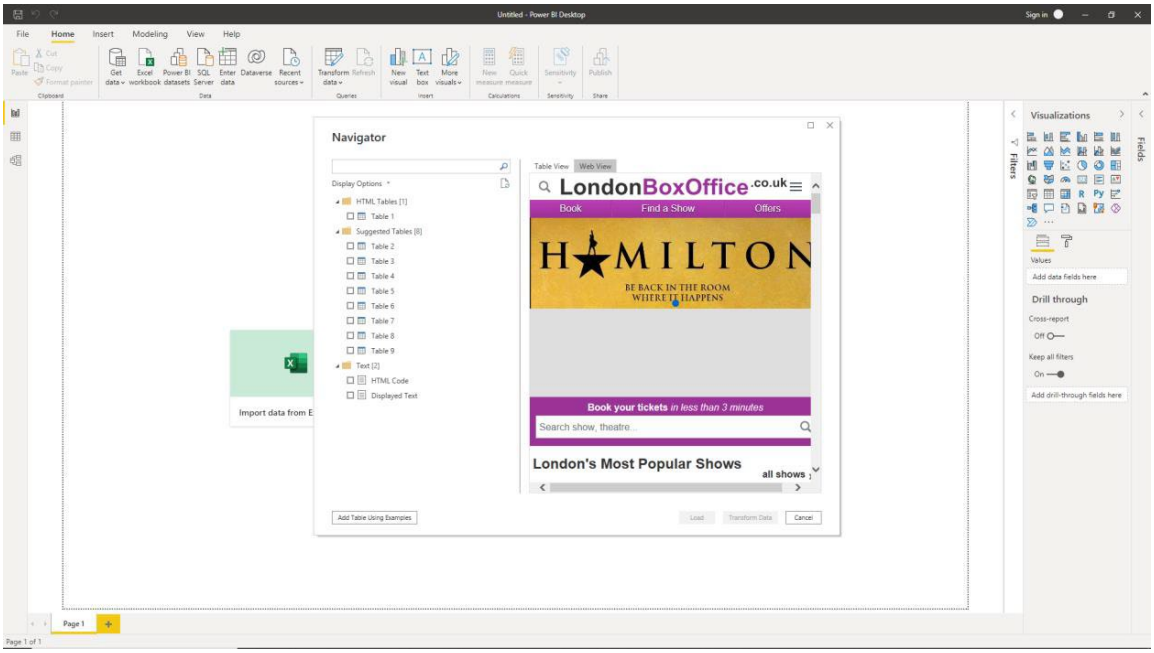


4. Copy and paste the following address into the pop-up window and click OK.

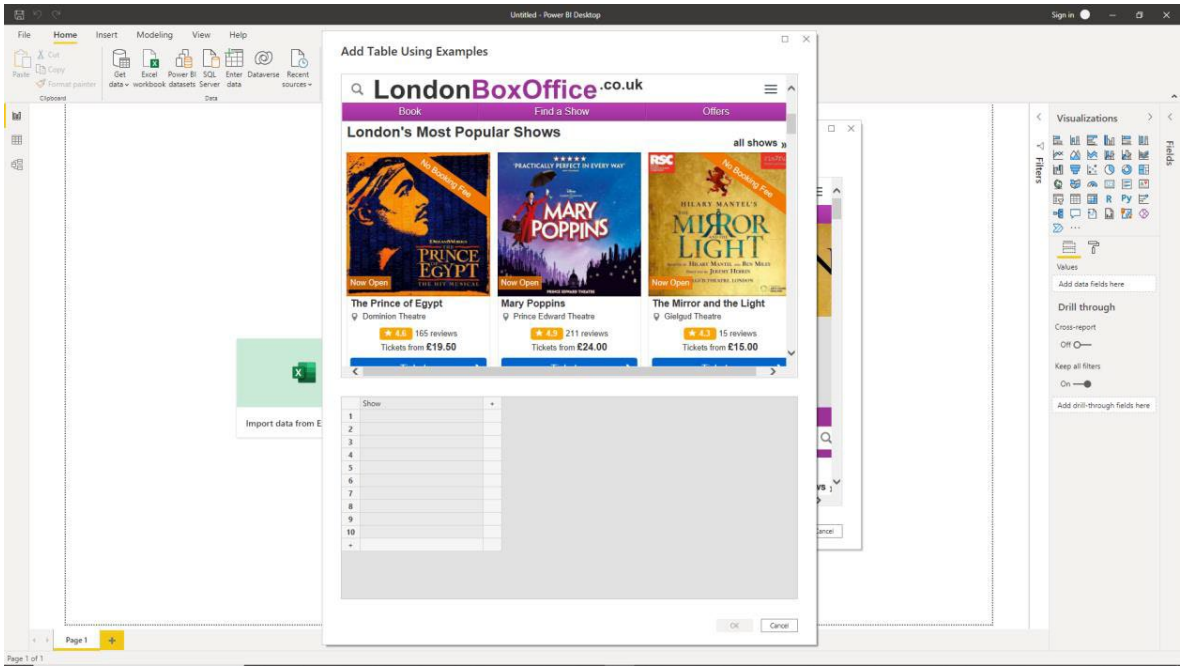
   https://www.londonboxoffice.co.uk/



5. You can examine the webpage in the Web View tab of the Navigator. Then click **Add Table using Examples**.

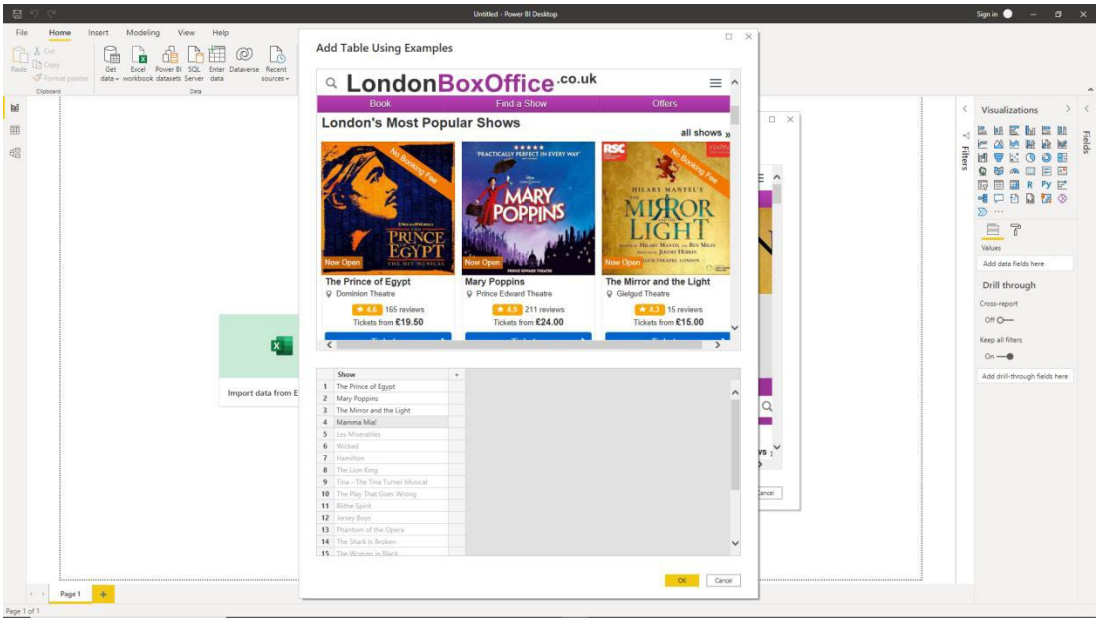


6. You can also see the Web View on the next screen, so can scroll to the data that we are interested in. We can also rename Column1 'Show' by double-14 clicking in the header.

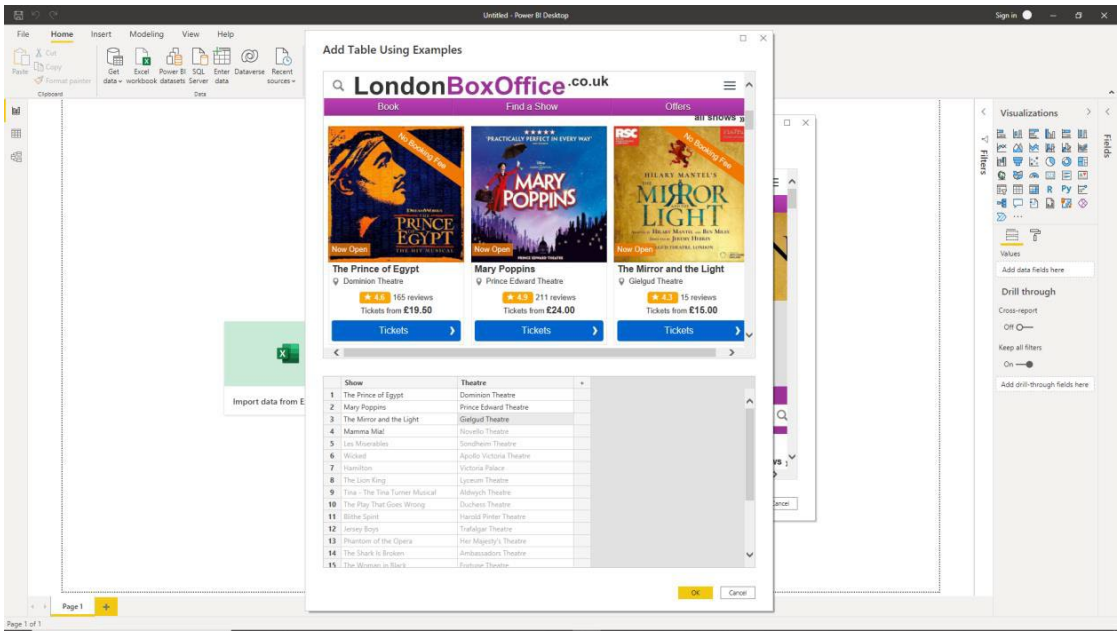


7. We can now give PowerBI examples of the data that we want to populate this column, in this case the production title. As you type, various autofill options will appear. Choose the title, then do the same for row 2. Tab away and the click a blank area in the form.

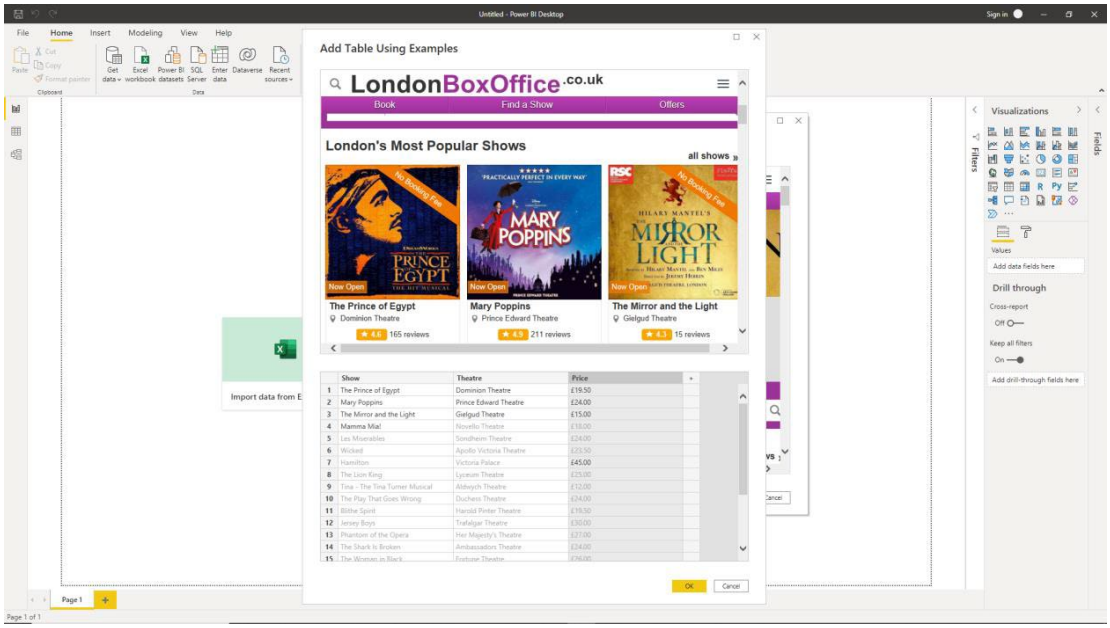
   You can see that the column should be automatically filled with the production titles. However, good practice dictates at the very least a cursory glance to check this is performing as expected.



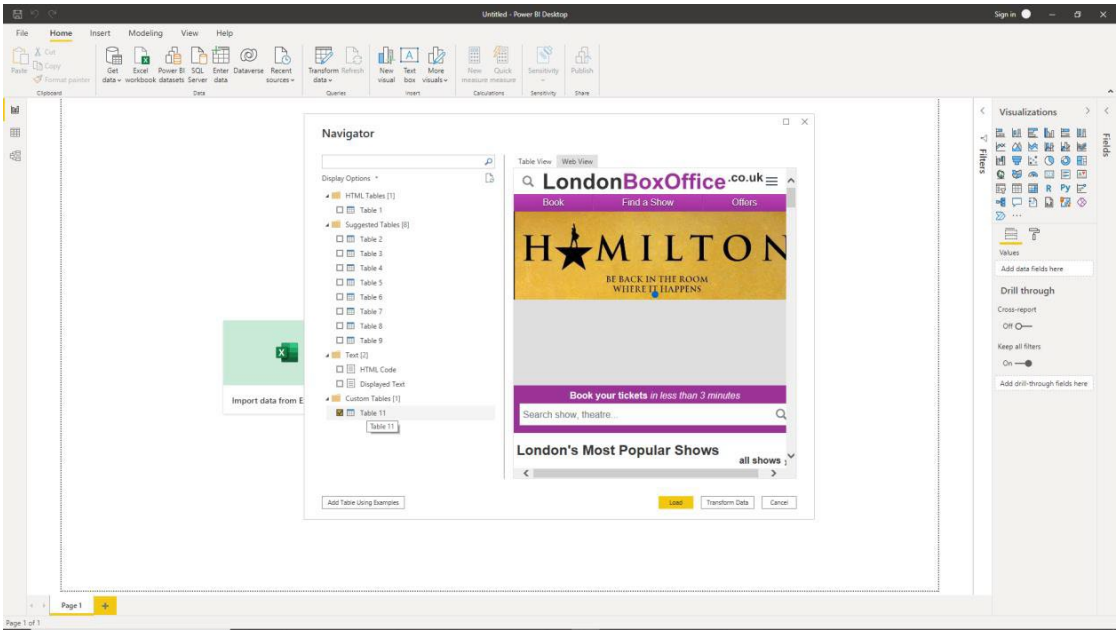
8. We can click '+' to insert a new column. Rename it 'Theatre', then similarly insert the first 2 rows of data for this column, double checking that the performances are tied to the appropriate theatre.



9. Repeat this process for Price. You can deal with any missing values by clicking in the cell and entering the correct value as you did to populate the first 2 rows. Once you are happy with the data, click OK.



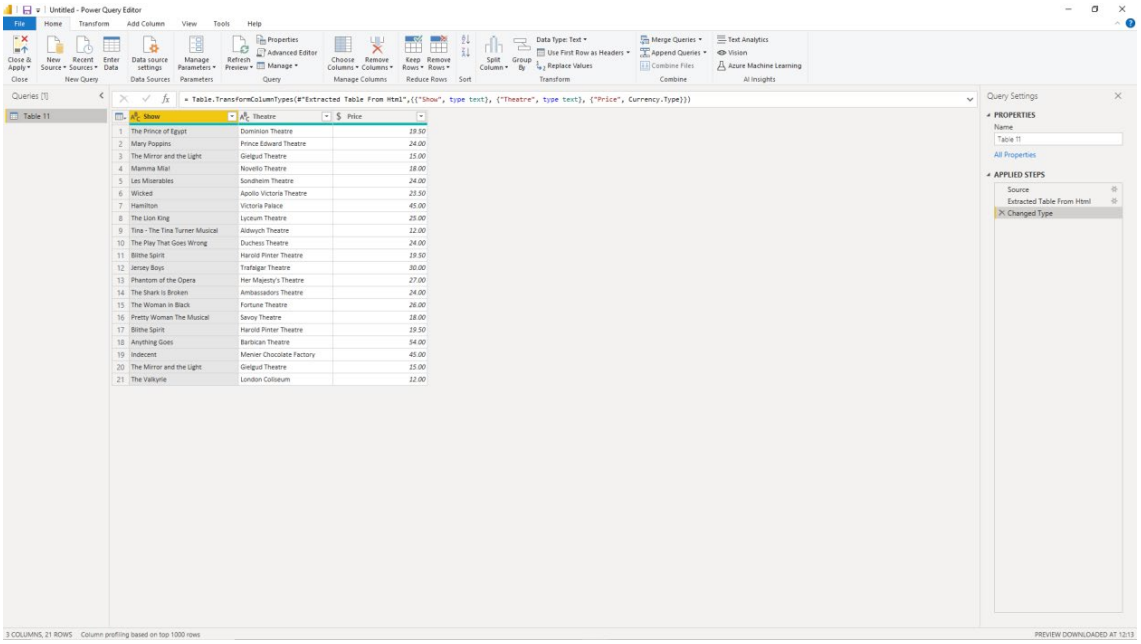
10. You should now be able to see a new table under Custom Tables, with a check box ticked. Now click Transform Data.



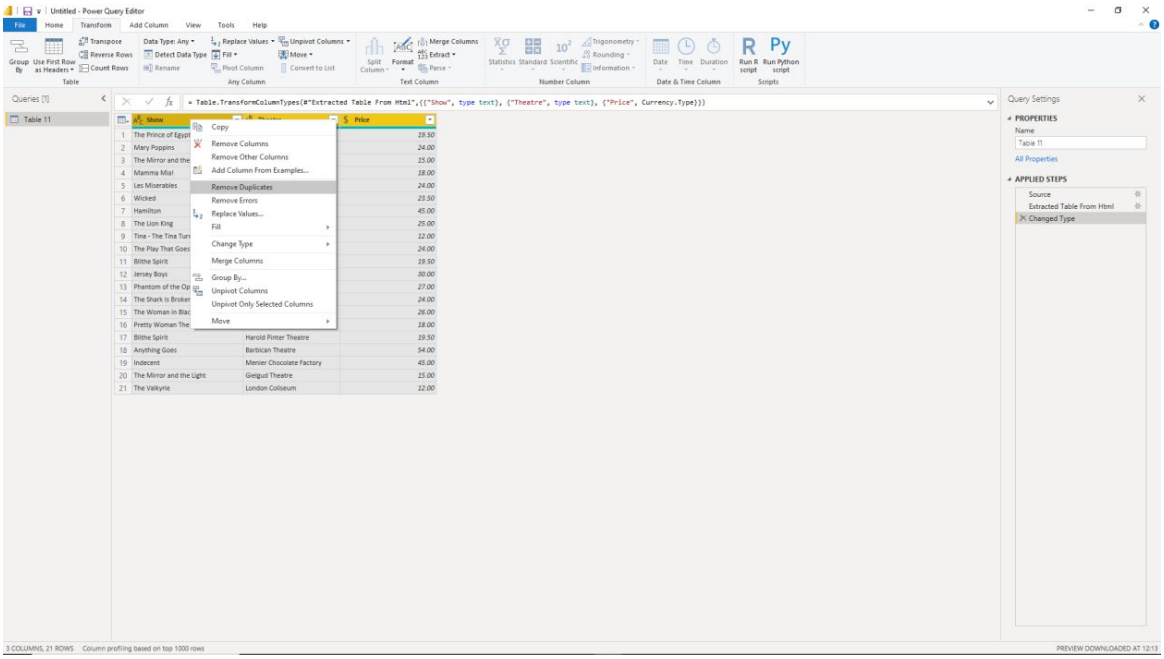
Independent activity:

Data transformation

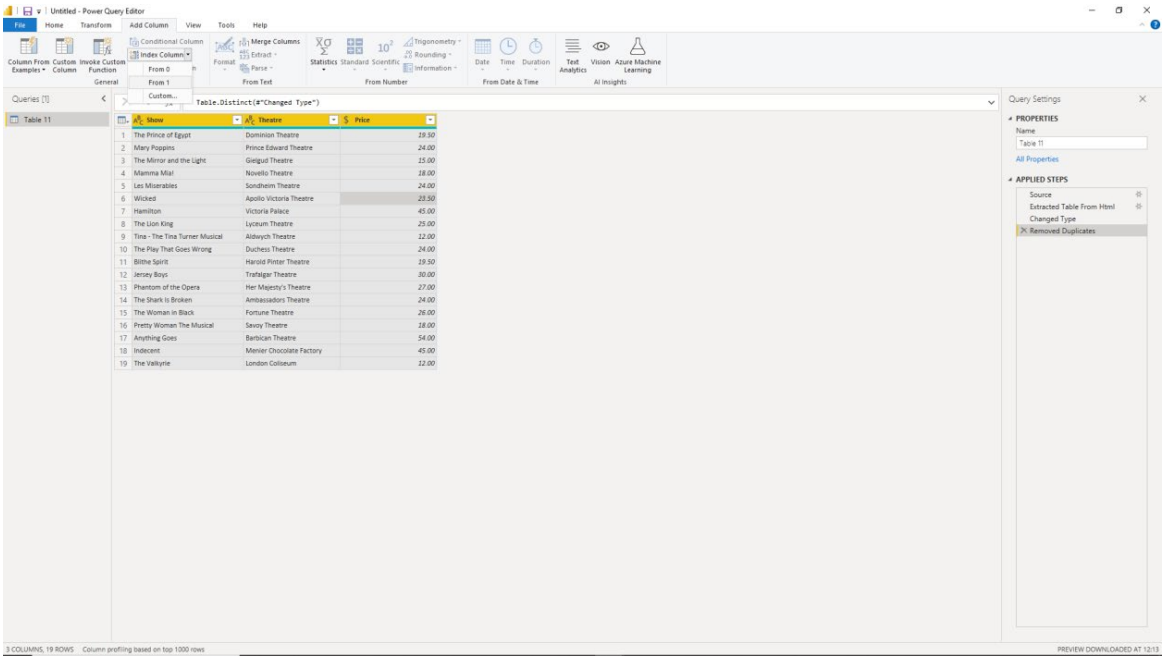
11. This will open Power Query that we have seen briefly before through Excel. Notice that it has recognised our Price column as a currency data type.



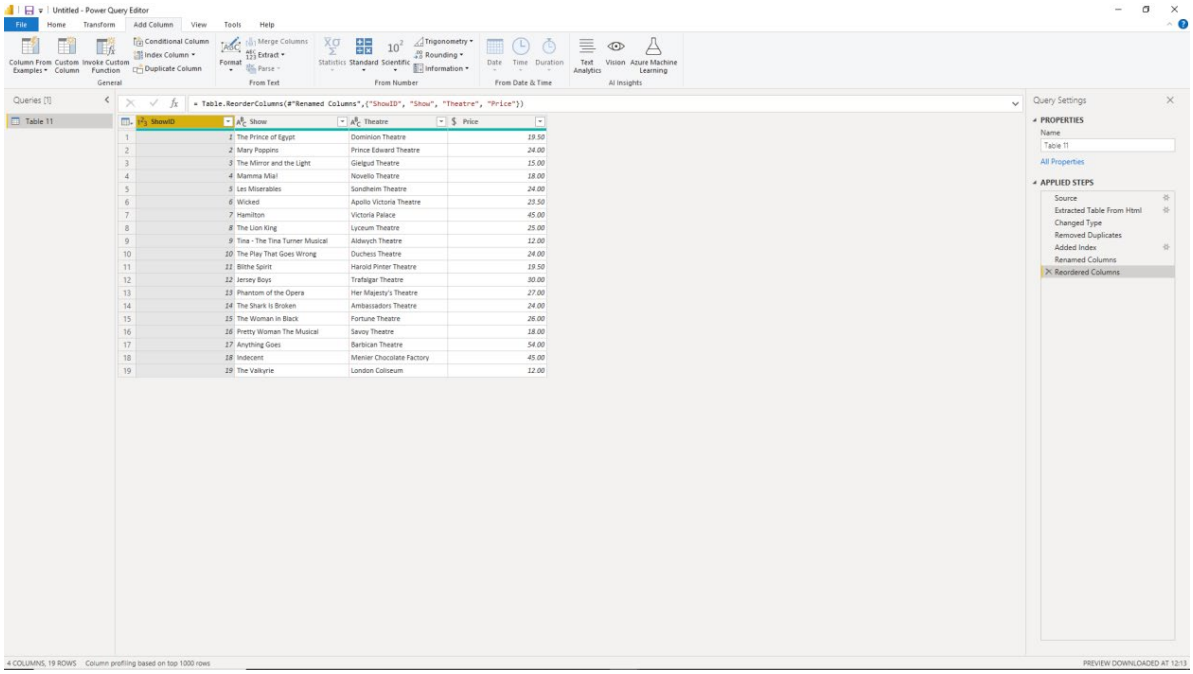
12. The data we have seems reasonably clean. However, there are a few duplicated records in the list. Press Ctrl-A to select all the data, then in a column header, right-click and select remove duplicates.



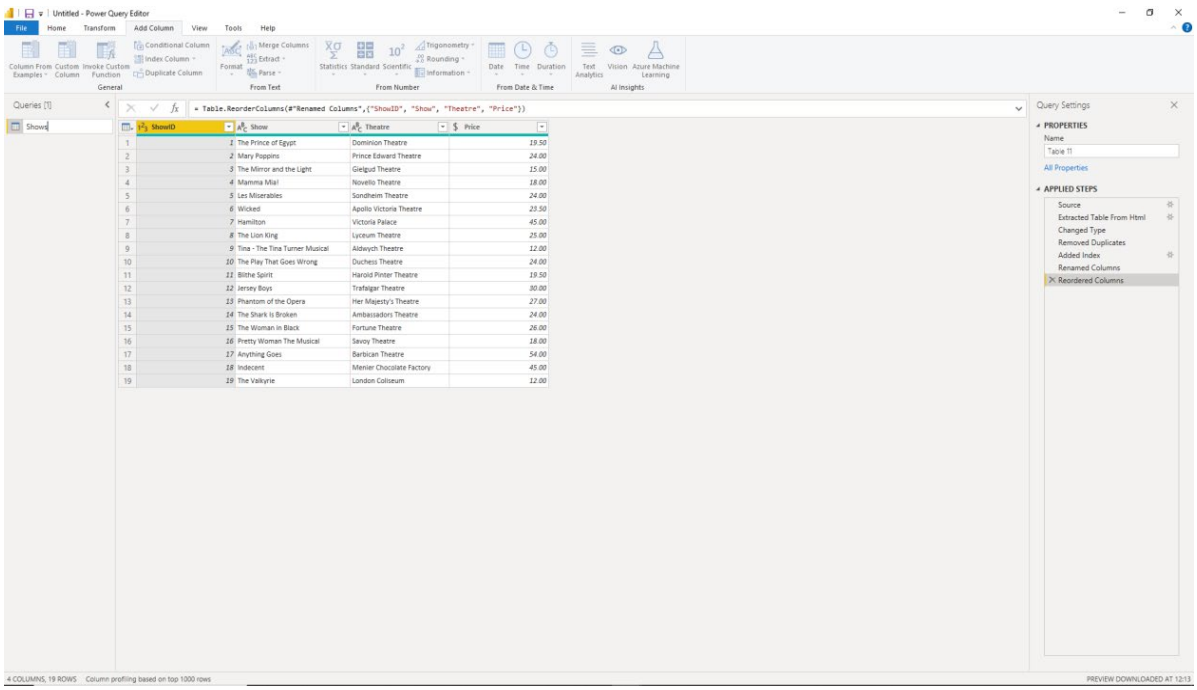
13. Finally, for this table, we want to add a Primary Key by navigating to Add Column, clicking the drop-down arrow next to Index Column, and selecting ‘from 1’.



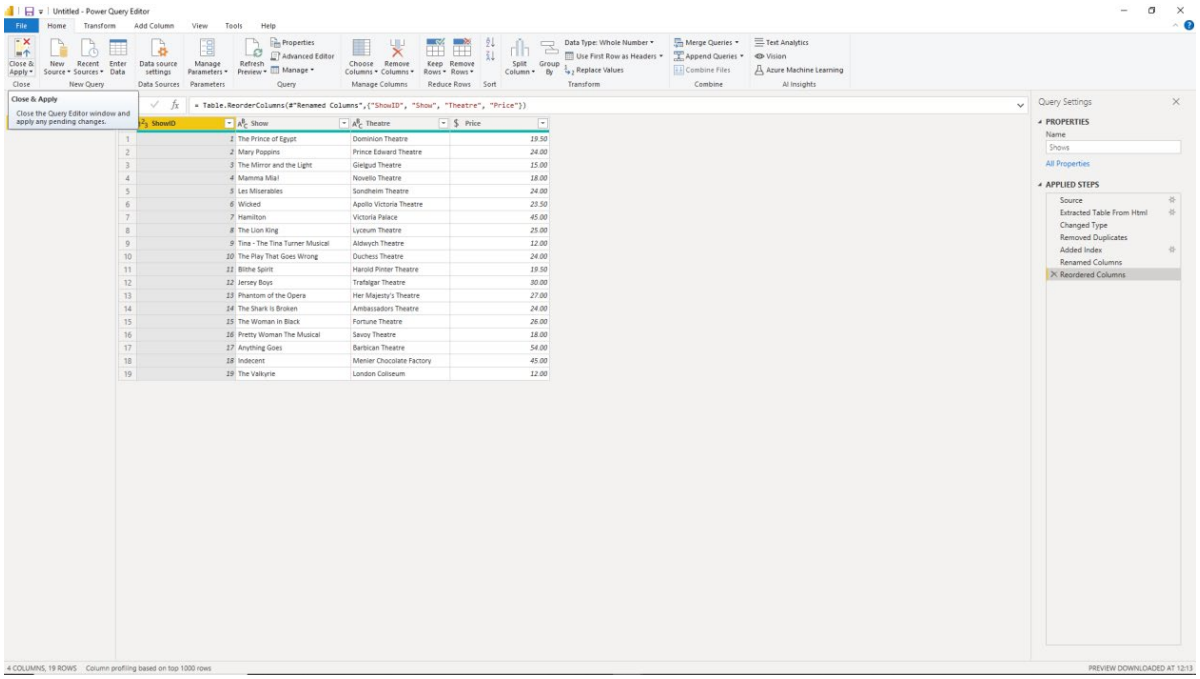
14. You can then Double-click the header and rename the column ShowID. Then click and drag the header across to the left of the table to move the column.



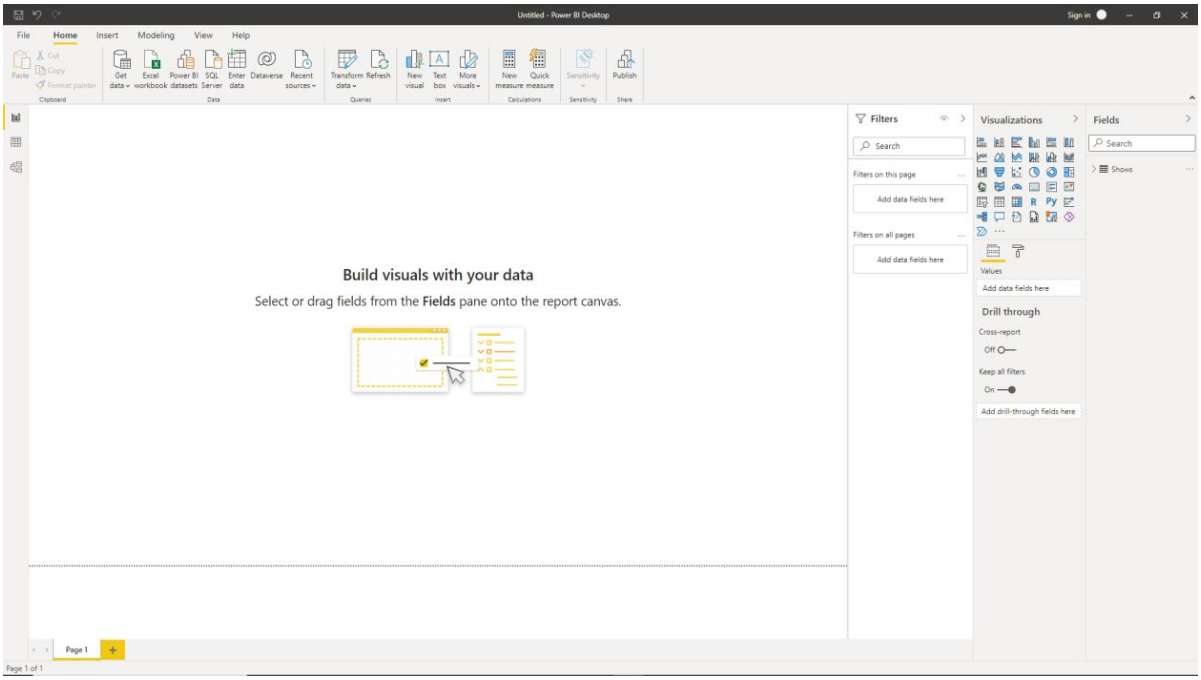
15. Double-click the table name in the queries panel at the far left of the screen and rename it ‘Shows’.



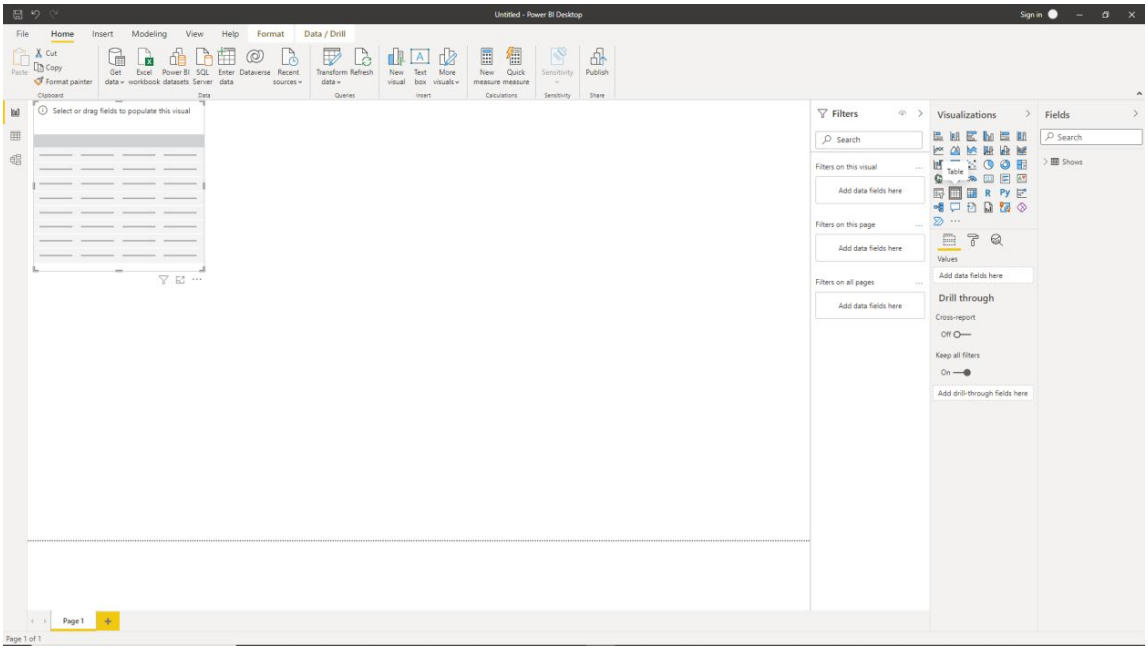
16. Navigate to Home and click Close and Apply.



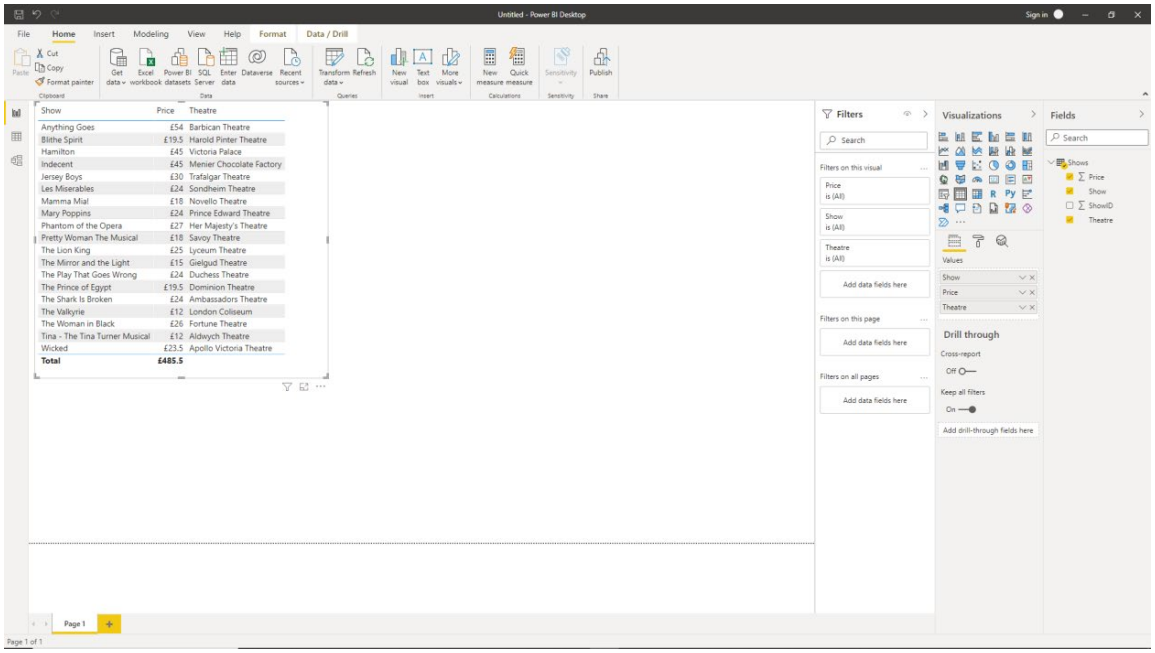
17. Now we can start to build a simple dashboard to display all the data we need for decision-making.



18. Let's start by applying some statistical techniques as we did previously in module 4. Click the Table icon in the Visualisations Pane to Insert one into our canvas.

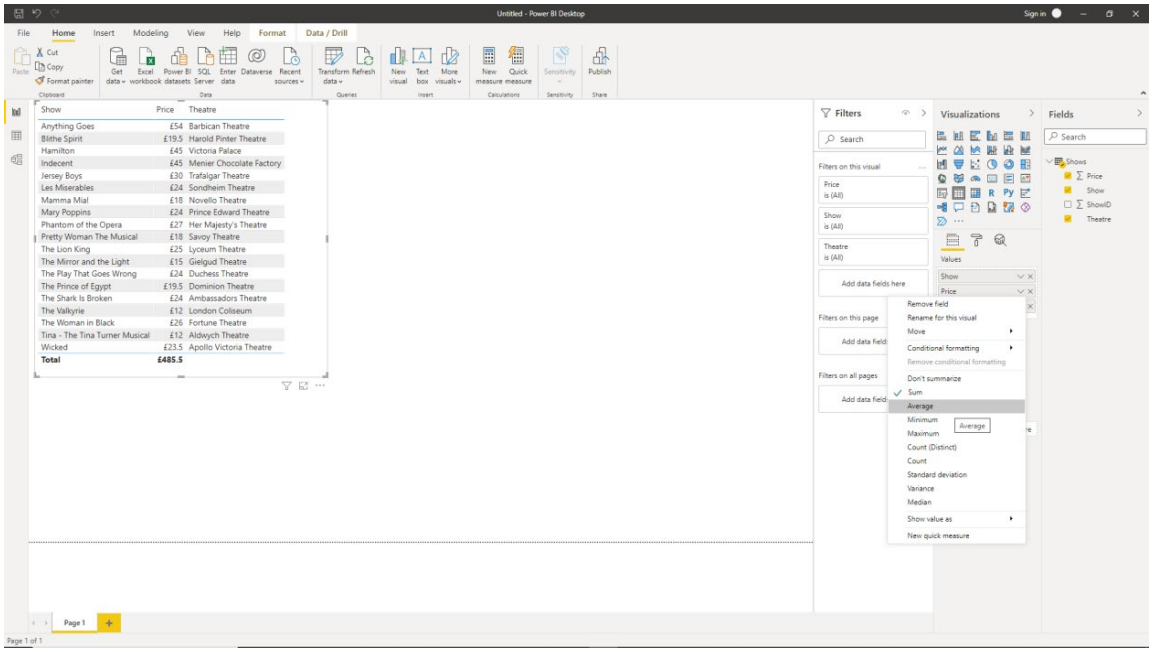


19. Expand Shows in the Fields Pane (on the far right) to display our column headings. Then click the select boxes for Price, Show, and Theatre, or drag them into the visual.



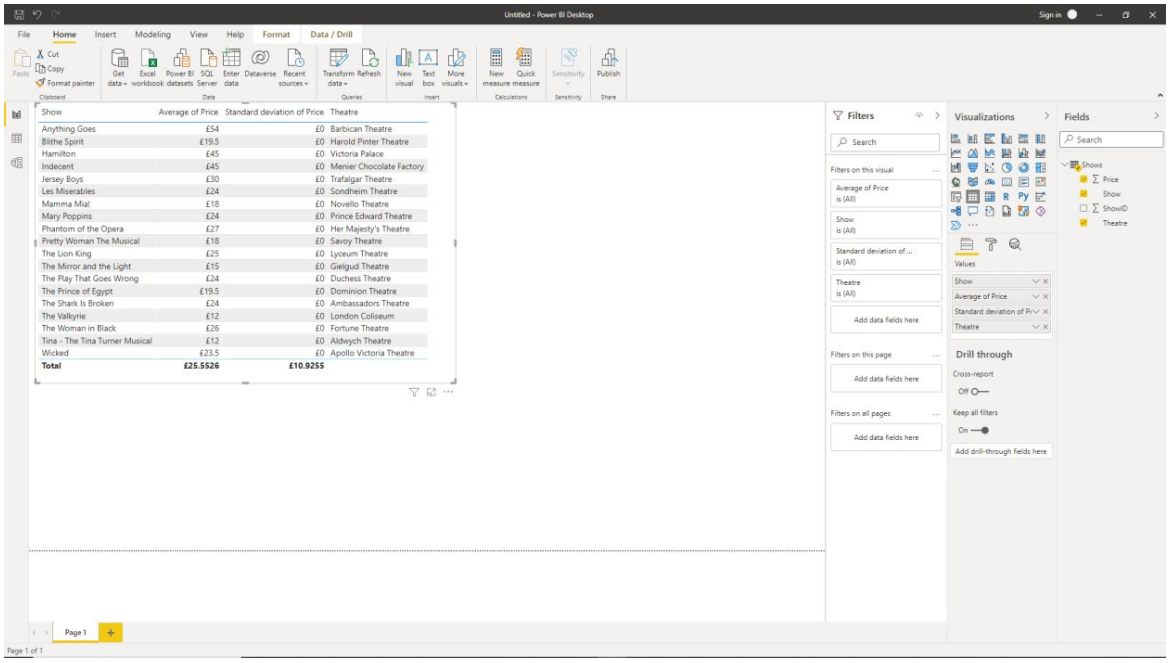
20. You can resize the table box by dragging the edges, and re-order the columns by dragging the values in the Visualisations Pane. We can see that PowerBI has automatically summed our price column, however, that isn't very useful for our needs. What may be more useful is to find the mean of the prices.

    Click the drop-down arrow next to Price in Values section of the Visualisation Pane and select Average.



21. Notice that while that row is still entitled Total, it now displays the Mean. Another important statistical technique we have used before is Standard Deviation.

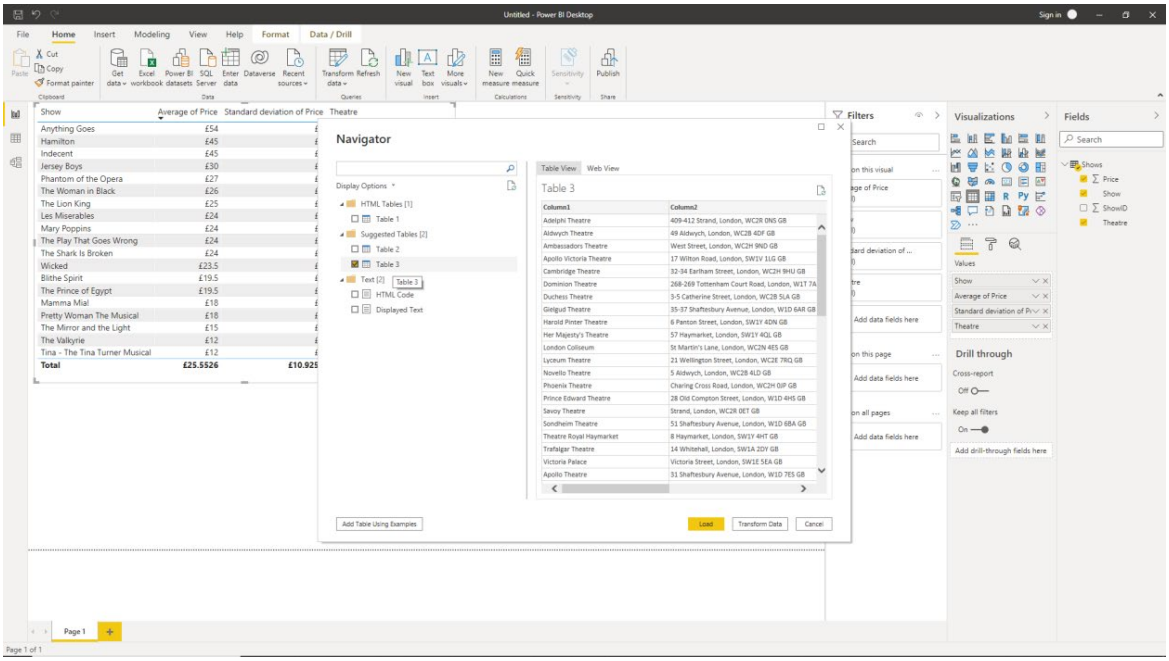
    Drag another example of Price from the Fields Pane to underneath Average of Price in the Values section of the Visualisations Pane, and again click the drop-down arrow, this time selecting Standard Deviation.



22. Your friend wants to know where these shows will be in order to book the most convenient hotel. You search around on the website your friend sent and find the following page which you should paste into the Home > Get Data > From web dialog box.

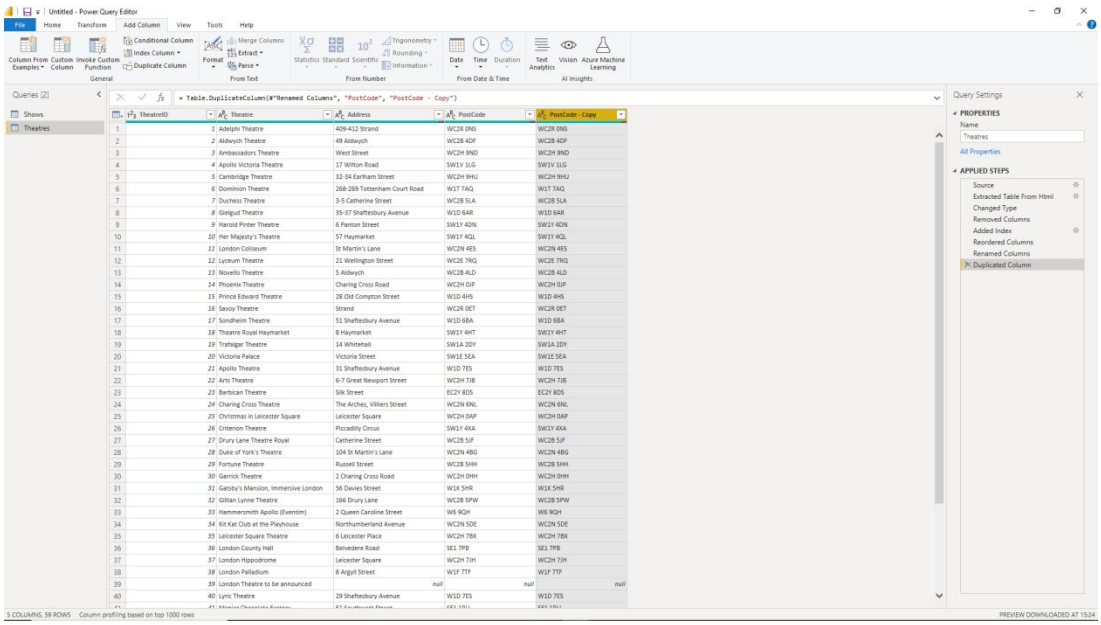
    https://www.londonboxoffice.co.uk/london-theatres

Select suggested table3 as below to transform in Power Query.



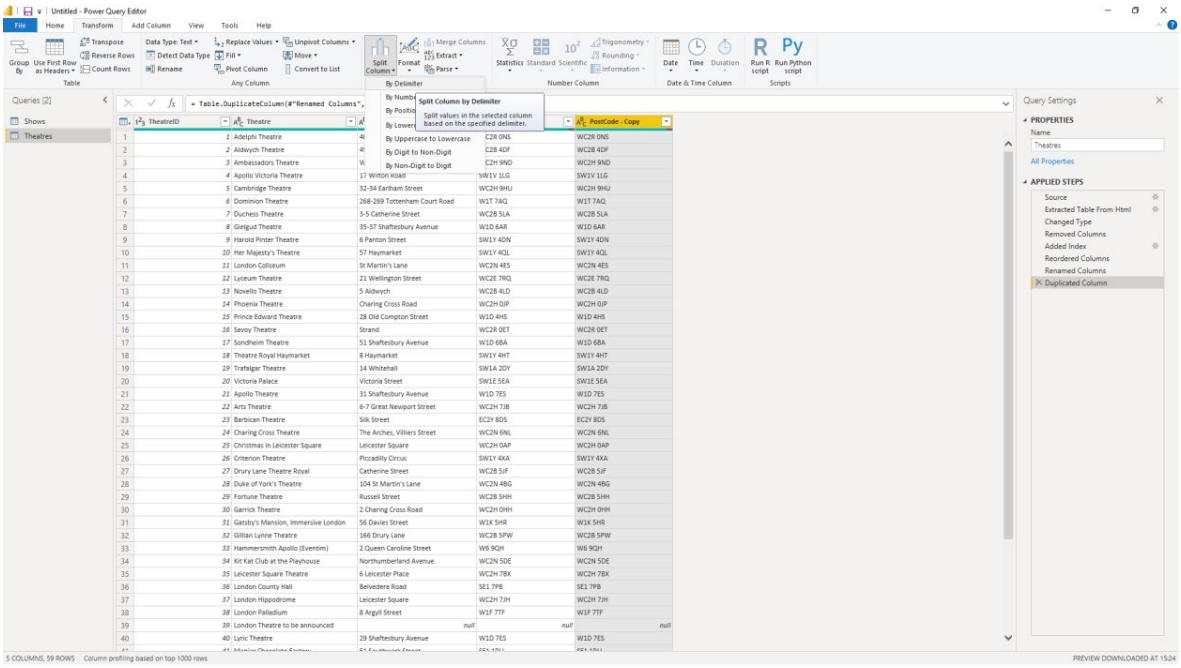
23. In Power Query, rename the table Theatres, remove columns 2, 3, 5, and 7 by holding down Ctrl on your keyboard to select all four columns, and right click a header and select Remove columns. Add an Index column as before and drag this column to be the first column. Rename the remaining columns TheatreID, Theatre, Address, and PostCode.

24. Right-click the column heading for PostCode and duplicate the column.

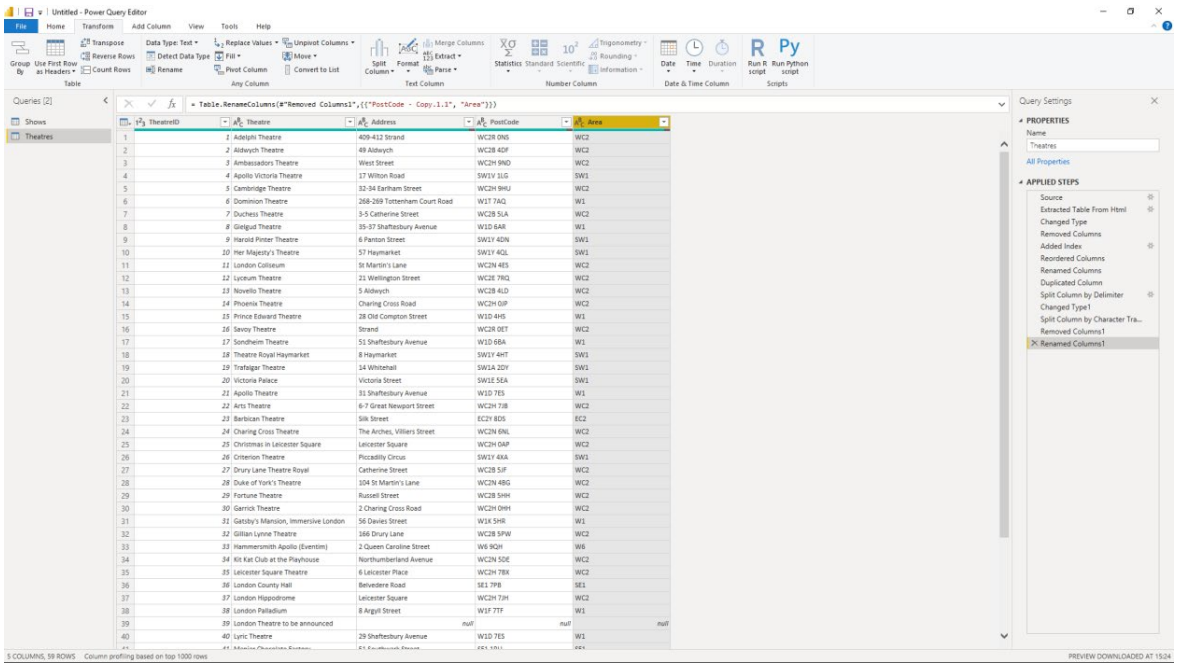


25. Split column PostCode – Copy by the right most space delimiter,

then split column PostCode - Copy.1 by digit to non-digit. Then, remove the two right most columns.



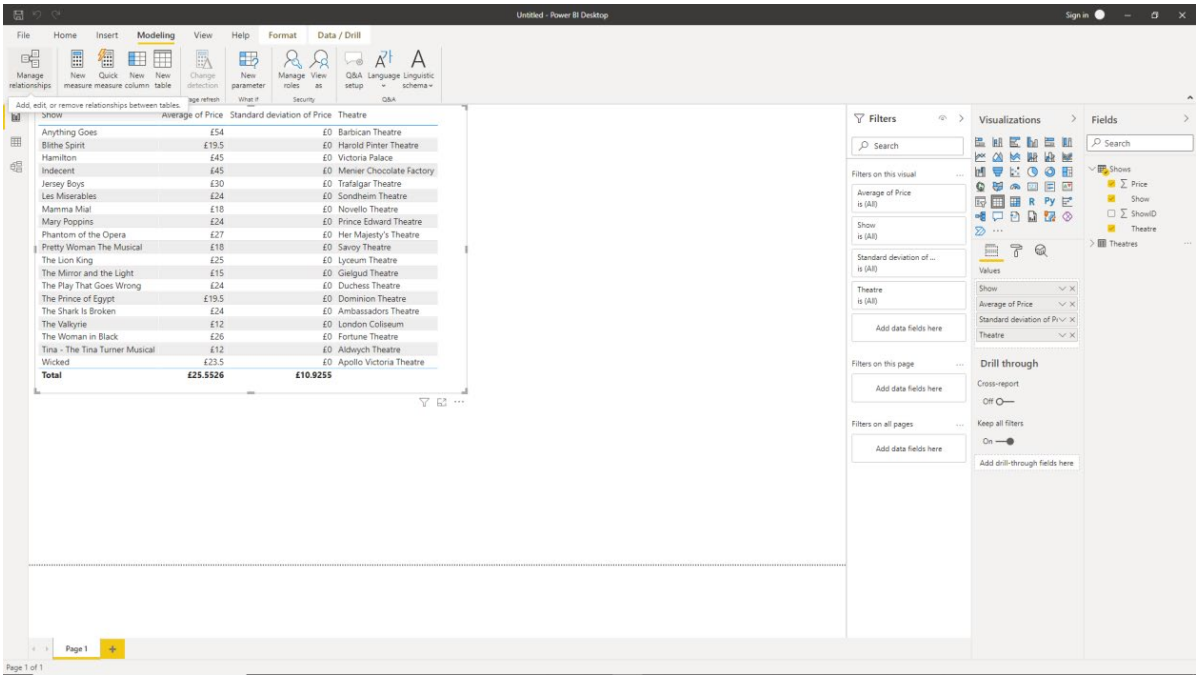
26. Finally, rename the newest column ‘Area’, then Close and Apply.



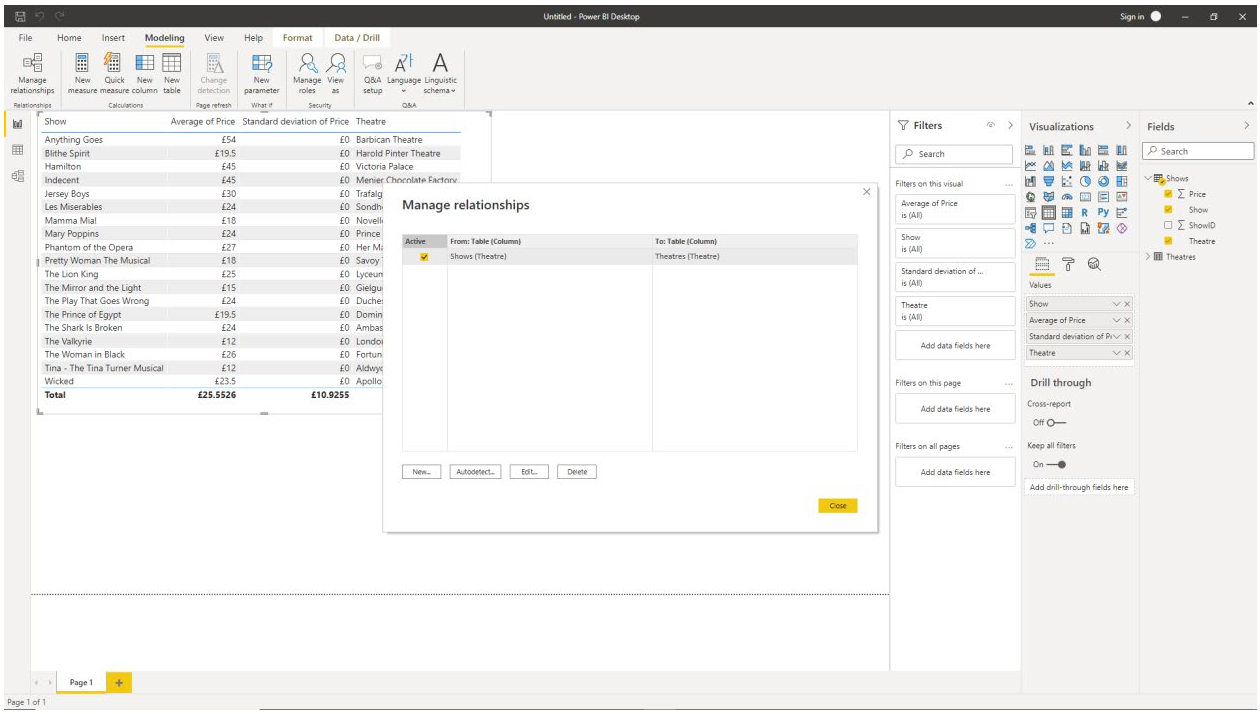
Independent activity:

Data modelling

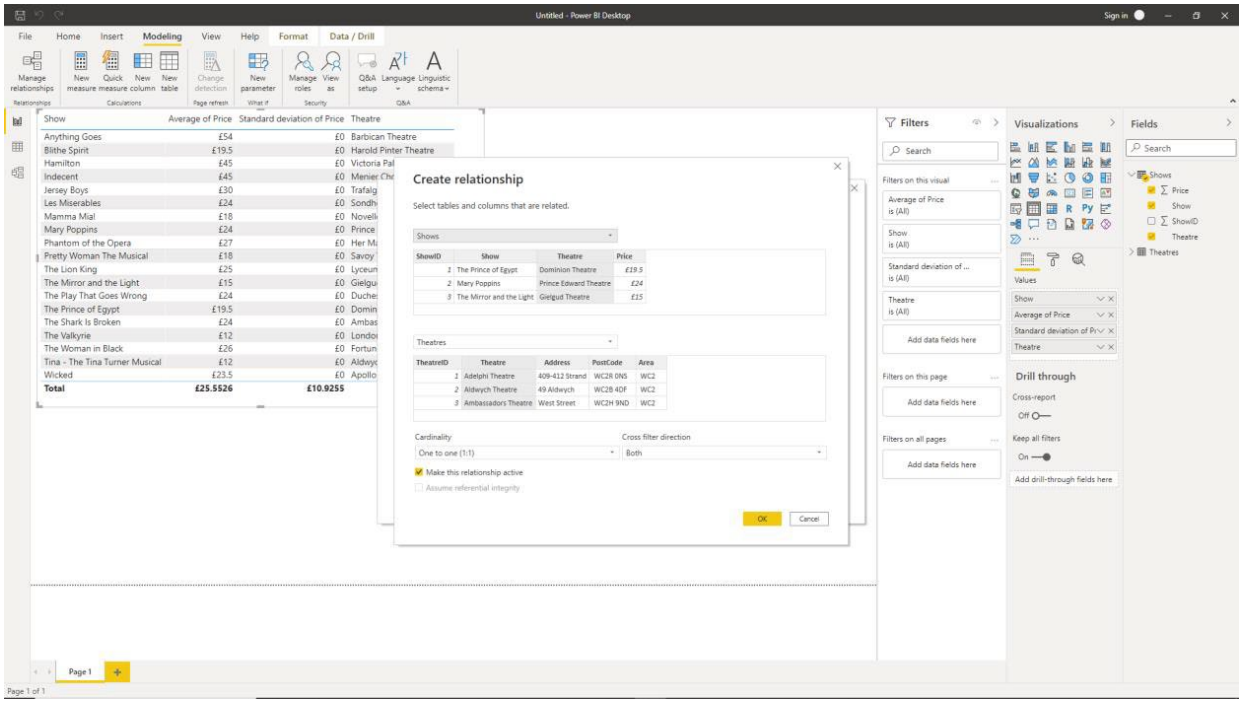
27. Now we have two tables, we can start creating a data model. Navigate to the Modelling ribbon > Manage relationships.



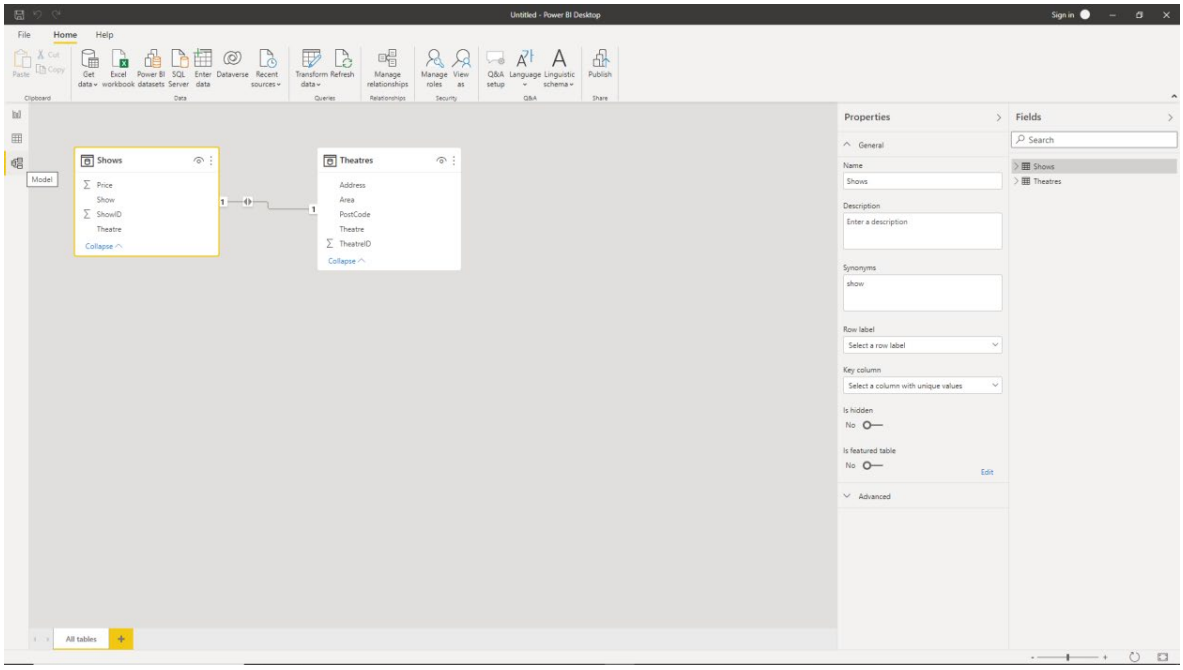
28. Here you can see that it has autodetected the relationship between Theatre in Shows and Theatre in Theatres.



29. If that relationship hadn't been autodetected, you can add relationships manually by using the new button and specifying relationships. You can then OK and close back to the dashboard.



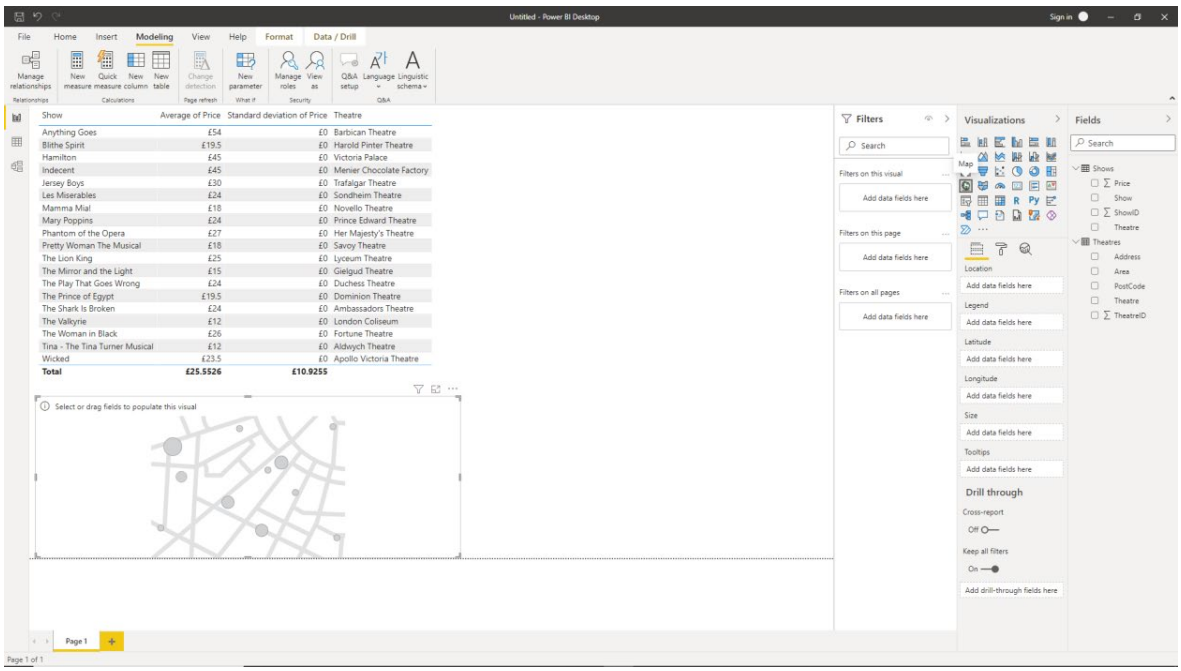
30. You can also view and manage your data model by clicking the bottom most icon on the far left of the screen. Navigate back to the dashboard with the topmost icon.



Independent activity:

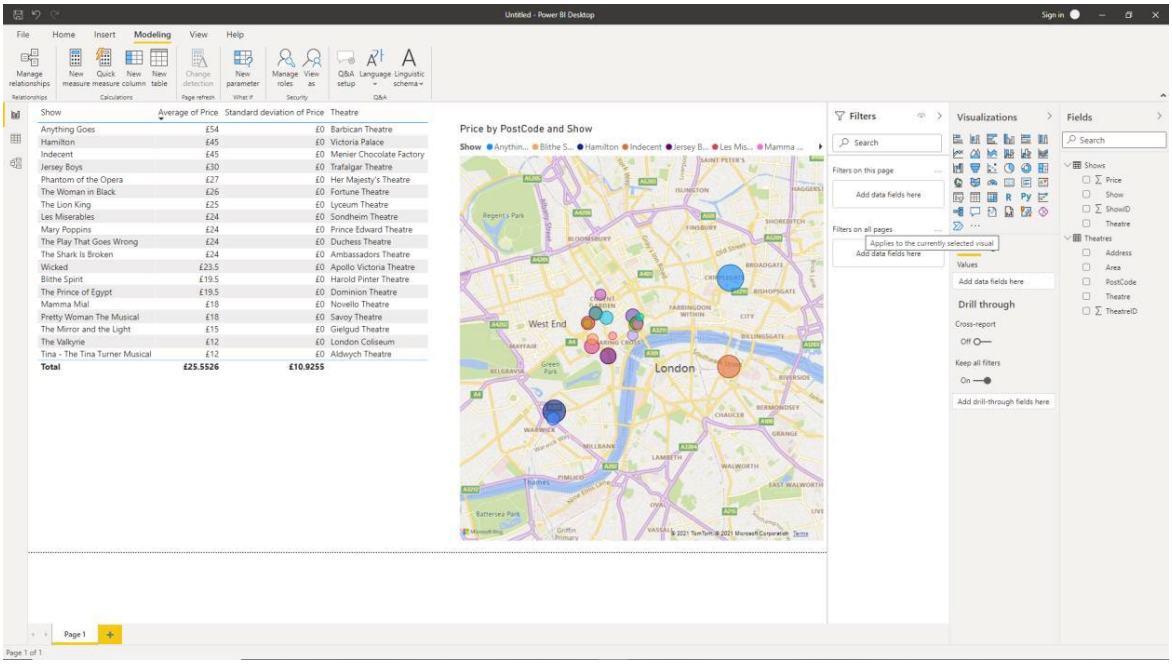
Map visualisation

31. Click away from your Table visualisation and insert a map visualisation.

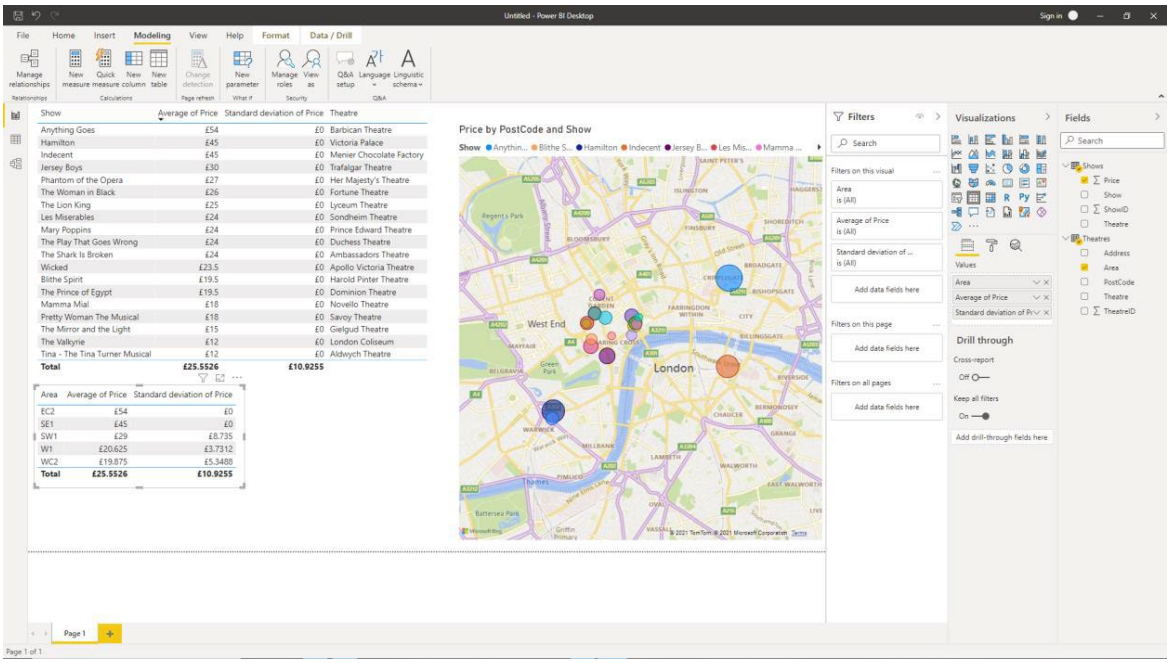


32. Drag into the map Postcode from Theatres and Price from Shows. You will see the map update with the locations of the theatres with PowerBI automatically sizing by Price. Drag the Show field into the Legend section of the Visualisations Pane, to show on the map which performance is where. Resize and reposition the map over to the right of the dashboard.

    What trends, patterns, or insights present themselves in your results?

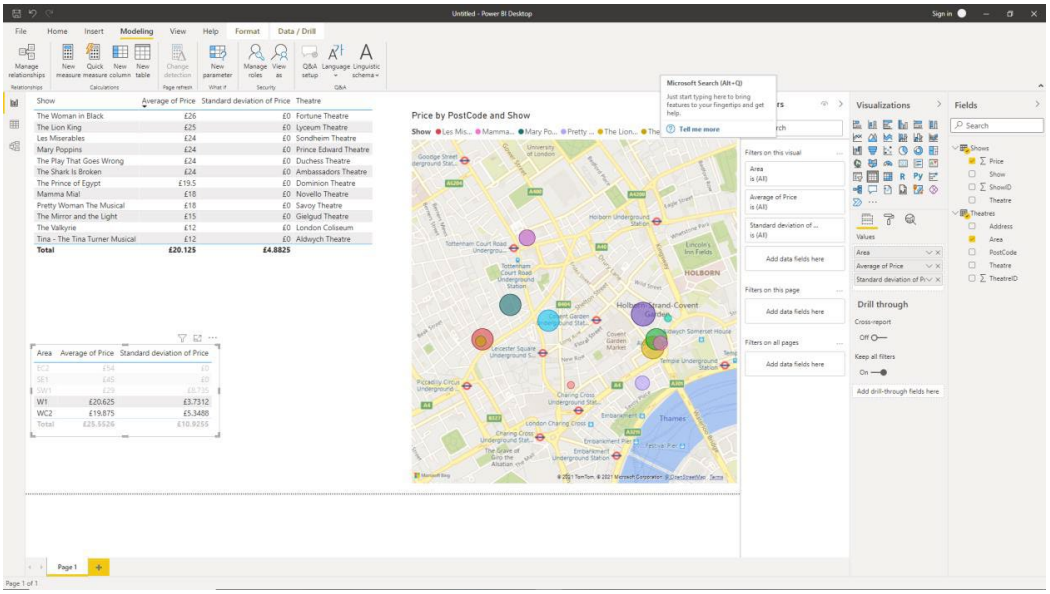
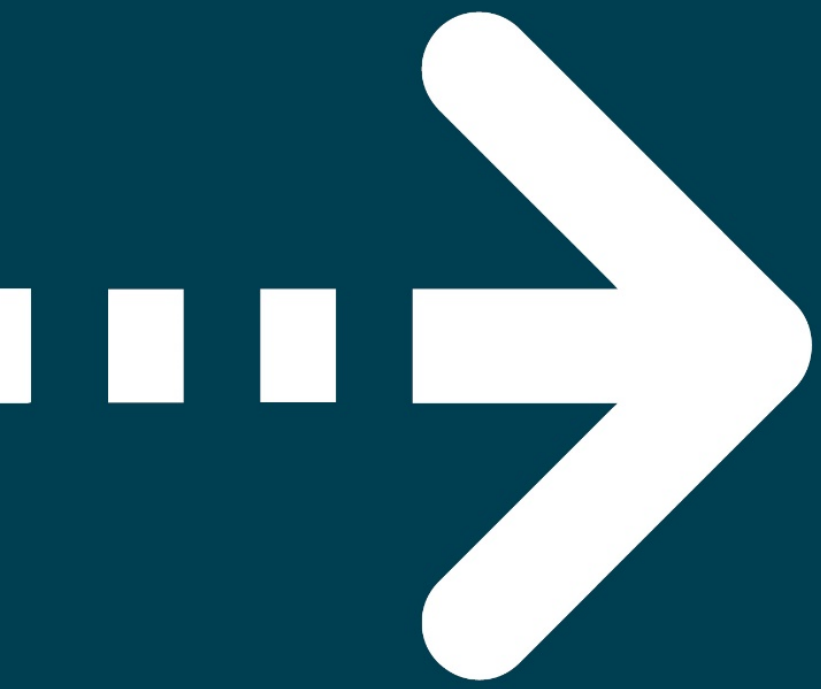


33. In your dashboard, insert a new table visualisation and populate it with fields, Area and two instances of Price. Then in the Values section of the Visualisation Pane set one Price to Average and one to Standard Deviation as before.



34. By selecting the cheapest area by average, or ctrl-selecting multiple, we can quickly focus on just the information useful to us.

    Think of the ways in which this could be improved, for instance, showing ratings or accessibility. Can you add hotels to the map to help your friend in their search?



35. Save your work as evidence and submit to your DLC via your Bud account.