Preparing Data for Analysis in Power BI

Prepared By

Said Fawzy

Manager of Information Center
Arab Contractors

Table of Content:

Contents

Introduction

لما كانت قوة أى مؤسسة تكمن فى سرعة إتخاذها القرار ومواجهة تغيرات السوق والمنافسة ونقل المؤسسة من مكانة إلى أخرى أفضل عن طريق الحصول على أكبر عائد ، وتقليل التكاليف ومعرفة مواطن القوة والضعف فيها وفى البيئة التى تحيطها ، ولأن البيانات والمعلومات هى القوة الاساسية الداعمة لاتخاذ اى قرار باى مؤسسة ، ولما كانت البيانات فى حد ذاتها لا تعين على اتخاذ القرار الصحيح ، ولكن يلزم لها ان تجيب على الاسئلة المحددة والدقيقة و الدائرة فى اذهان متخذى القرار حتى يتثنى لهم الرؤية الواضحة للامور على اساس متين يعول عليه بعد إعدادها وتقديمها فى صورة تقارير ومخططات واضحة جليه.

وعليه فقد قمت بإعداد هذه الدورة كمقدمة فى إعداد البيانات فى برنامج الـ Power BI حرصت فيه ان يكون مبسطا ومركزا ويعطى صورة للدارس الذى يرغب فى الاستمرار فى هذا المجال من معرفة اساسيات إعداد البيانات للتحليل فى هذا البرنامج.

وقد قمت باستعرض مراحل البرنامج كاملة من كومبيوتر المستخدم واستخراج البيانات واعداد التقارير حتى رفع التقرير الى السحابة الإليكترونية ، ومشاركته مع افراد المؤسسة ، وذلك قبل البدء فى اول مرحلة من مراحل البرنامج وهى استيراد وتحميل واعداد البيانات للتحليل وهو موضوع هذا البرنامج التدريبى.

وقد حرصت على ان تكون المادة العلمية مليئة بالصور التوضيحية من البرنامج حسب اخر اصدر له فى يناير 2024 وذلك حتى تكون مرجعا رسوميا للمتدرب يلجأ اليه عند الحاجة.

وقد قمت بتقسيم كل فصل الى :

- عرض مبدئى لفكرة الفصل ومثال عملى يشاهده الدارس لتطبيق الفكرة.
- اسئلة لاختبار المعرفة حتى نتأكد من فهم الدارس للنقاط التى القيت عليه.
- تدريب عملى على بيانات مقدمة للتحليل والاعداد وبيانات يتم استيرادها من الانترنت.

واتمنى ان تكون هذع المادة العلمية داعما لكل من اراد تعلم اساسيات البرنامج ، واتمنى ان يعيينى المولى عز وجل على استكمال باقى المواد العلمية لهذا البرنامج فى القريب العاجل

وفقنا الله الى ما يحبه ويرضاه

مهندس سعيد فوزى محمد هدى

مدير مركز المعلومات – مدير الجودة

إدارة العطاءات

المقاولون العرب

القاهرة 22 يناير 2024

Chapter 1: Introduction to Power BI

What is Power BI?

Power BI is software to create & publish reports and data stories from your datasets. You can make highly interactive, engaging and powerful reports, dashboards or visuals with Power BI. You can connect to any data (Excel files, SQL databases, BI warehouses, Cloud data, APIs, web pages and more), mashup the data, link one table with others, create *clickable* visualizations and then share them with your audience securely through Power BI.

How is Power BI different from Excel?

- **Power BI allows rich, immersive and interactive experiences** out-of-box. You can click on a bar in bar chart & other visuals respond to the event and highlight or filter relevant data. You can show graphs & visuals that are very tricky (or impossible) to reproduce in Excel like maps, pictures and custom visuals.
- **Power BI works with large data sets. There** is no artificial limit of 1mn rows in Power BI. You can hookup to a business data set and analyze any volume of data. The limit depends on what your computer (or Power BI server) can process.
- **Share and read reports easily. You** can create reports in Power BI and share them in formats that are universal (*i.e.* browser pages or apps). This means your boss need not have Excel or Power BI installed to enjoy the beautiful reports you create.
- **Power BI is for story telling** while Excel is for *almost anything.* We can use Excel to simulate pendulum motion, calculate Venus's orbit, model a start-up business plan or many other things. Power BI is mainly for data analysis & story telling. If you try to replicate a large, intricate financial model or optimization problem with Power BI, you will either fail or suffer miserably. On the other hand, if you use Power BI for making reports, running cool analysis algorithms (clustering, outlier detection, geo-spatial patterns etc.) you will wow your colleagues and bosses.

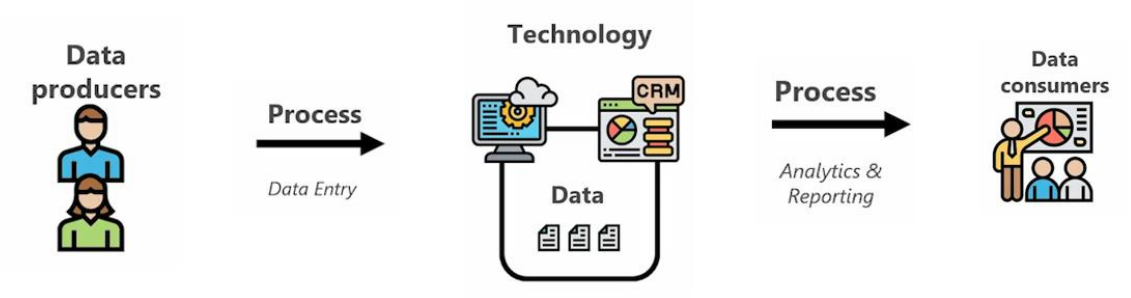
Data Life Cycle

Business Process

Some examples of business processes are things like sales inventory management and payroll. So, payroll has processes around it like time sheet entry and then invoicing. But a lot of it really starts around the time sheets that people enter and all the different tracking you're doing and all the downstream processes that end up coming out of it.

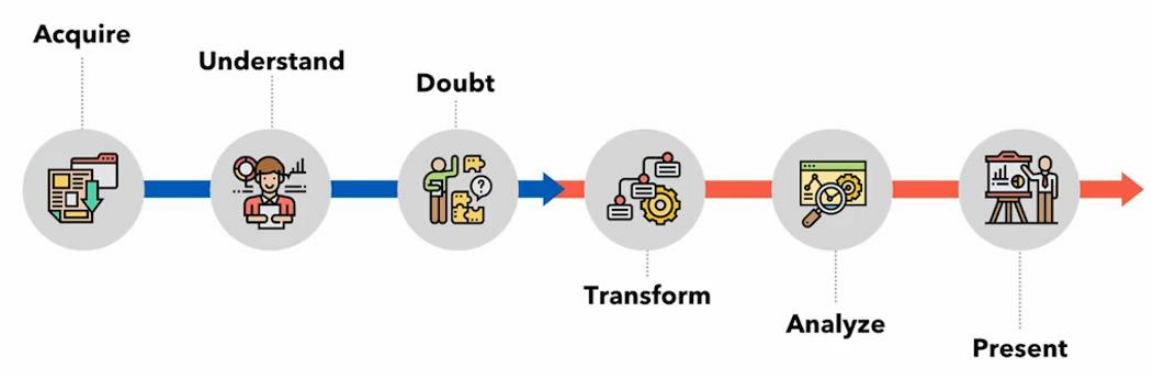
- Pick one of those or a business process that you have in your organization and frame it around this data lifecycle.

Data Cycle in your organization



- Typically, there's **some type of technology** that we have in our organization that is used to help manage our business process. So, we may have **spreadsheets**, **CRM,** we may have an **ERP** all depending on how large our organization is and how complex the business process is.
- We have **data producers**; these are people that are in our organization that help manage that business process. And then those people end up having some type of **process** around that where they do data entry. So at the end of the day, we have data producers that are in our organization and they're using some type of process to enter data into our systems.
- So after that process is followed and hopefully we have a standard and consistent and ultimately simple process because when we have those things in place we typically find that organizations end up with higher quality data. So, if the process is understood and is followed then the result in data that gets stored in our technologies is usually of a higher degree of quality. And you may have heard the term **garbage in garbage out**.
- The next thing that happens is we have **data consumers** on the other side that say Hey we'd like to how she measured this business process. How well are we doing in terms of sales? What are our inventory levels? Do we need to replenish our inventory? What are our time sheets looking like? Are we logging as many hours in a month as we thought we might be we on track to hit our revenue targets?
- So, there's all kinds of **business questions** that come out of any one of these processes here.
- So, the data consumers are typically the ones that are asking those questions. And sometimes the data producers and data consumers are the same people. And then the data consumers want to use some type of process around analytics and reporting to actually go ahead and measure that business process so we can see how well it's actually performing. Okay.

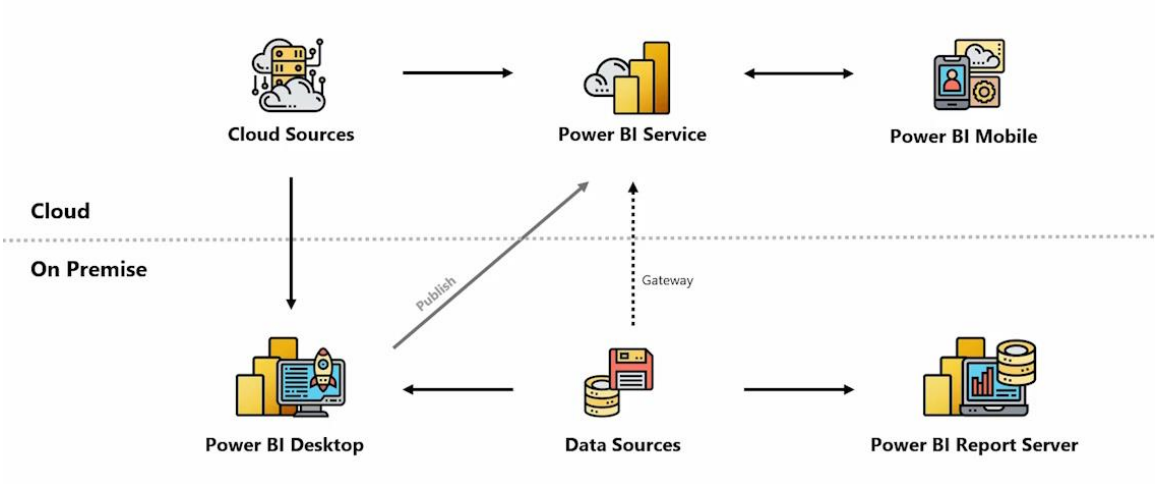
Analytics process flow



- It is the general stages of answering a business question.
    - **step number one when a business question is asked**, you are trying to understand where the data might be in an organization that supports that. So, you must go through some type of **acquisition step**. Maybe you understand that data, maybe you don't.
    - The second step is to go go ahead and try **to understand** what that data is telling us.
    - Then often we run into a phase of **doubt**. So, we get the data we start working with and it doesn't quite seem to be adding up to what we'd expect. So, there's some doubting that will happen. And ultimately, we end up going back and working with those subject matter experts to really understand what's happening.
    - Next what ends up happening is we go through some type of **transformation** process. So maybe that data's not quite in the form that we want for reporting or maybe there are some quality issues that are found in that data. We want to go ahead and do some cleansing on that.
    - Then we go through after our data in a report ready state, and we start doing some **analysis** on it to ultimately try and answer that business question.
    - Once we have our findings in place then we're ready to go off and start **presenting** the findings of our business questions to our business users.
- It's highly iterative. We're moving back and forth between these different circles all the time as we learn new things, new business questions are being asked and we learn different things about our data.

Power Bi Environment

Power BI Architecture



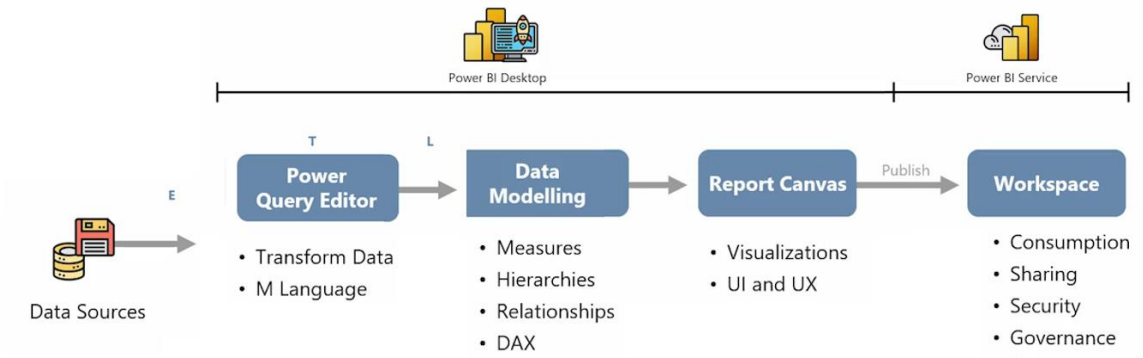
There are separate **two key areas of our services**:

- **On premise**:
  - things installed on a desktop or a server room maybe in your organization.
  - Power BI desktops first connect into data sources and ingest them into the Power BI desktop.
  - We then bring that data in, go through some transformations and get it into a report ready state.
  - And finally building some reports that we can use for consumption.
- **On cloud:**
  - Once that's complete, we want to go through a **publishing** process.
  - We take our work that we've done on the power BI desktop and publish it into the **power BI service** that resides in the cloud.
  - At this point in time the work that you've published will be available through a **mobile device** and almost instantaneously it will be available to you in power BI mobile. And if you have access to the power BI mobile app on your preferred mobile device then you can access your content through there as well.
  - The next one is a **gateway** up is we somehow need a way to get our refresh data daily or whatever your schedule happens to be from our sources on premise up into the cloud.
  - So, the gateway is essentially going to be that transportation highway that allows the Power BI service to connect back down into your on-premises data sources. And bring the new data into the service on your scheduled refresh basis.
  - If you have **data in the cloud,** you don't need the data gateway. You can just connect directly to those cloud-based sources.

- **Power BI report server**:

- if you actually want to do Power BI work and build things up but never publish it in the cloud and take advantage of some of the features in Power BI you can utilize the Power BI report server.

Power BI End to End Service



- In the **Power BI desktop,** we will work with the power query editor. doing some data modeling, end up building some reporting.
- And ultimately when we're done with those processes, we will take our work and publish it into the **Power BI service** and put our work into a **workspace** in that service.
- we first need to go in and find our **data sources** that's answer our business question we will connect then go ahead and **extract** that data from that data source.
- Once we bring it into the power query editor from an extract perspective then go through a **transformation** step and transform data, using the **M language** that is inside the power query editor.
- Once we go through that step of transforming our data and getting it ready for reporting we will then **load** that data over into the data model.
- Once our data is loaded into the data model and we have data that is almost ready for reporting we will go through a process of adding extra value to that data model. We will build relationships between our tables. Will perhaps add hierarchies to our data model which we'll do in this course. We will then go through and build some **DAX** expressions building some measures.
- Once we have our data model ready for reporting then we will go ahead to the **report canvas** and start doing some **visualizations**. And take those visualizations and craft stories. And hopefully stories that are compelling, easy to understand and satisfy our business questions.
- Once we have put stuff on the report canvas, we will then **publish** our work into the **Power BI service**. Because that is where people will go ahead and consume our data. We will work on sharing our data out.

Power Bi Licenses

	Not in Premium capacity	Premium capacity
Free	Use as a personal sandbox where you create content for yourself and interact with that content. A free license is a great way to try out the Power BI service. You can't consume content from anyone else or share your content with others	Interact with content assigned to Premium capacity and shared with you. Free, Premium per-user, and Pro users can collaborate without requiring the free users to have Pro accounts.
Pro	Collaborate with Premium per-user and Pro users by creating and sharing content.	Collaborate with free, Premium per user, and Pro users by creating and sharing content.

Source: https://docs.microsoft.com/en-us/power-bi/consumer/end-user-license

Download Power PI Desktop

1. You can download directly from Power BI Site:
    - https://powerbi.microsoft.com/en-us/downloads/
    - Advantage: You can apply to many computers.
2. You Can Install form **Microsoft Store**.
    - Advantage: Automatic update

Chapter 2: Your first Project in Power BI

Power BI Components

- Power BI Desktop,
- Power BI Apps,
- and Power BI Service.

Power BI Desktop

- is a Windows-based desktop application that is mainly used by data analysts or report designers to clean, transform, and load data, create a data model, design reports, and publish these reports. Power BI Desktop uses Power BI connector to access various data types and data sources.
- **Connectors** allow you to read data from various sources. This includes resources located in the local file system, such as Microsoft Excel, or PDF documents. Conventional database systems hosted on internal servers called on-premises databases, Cloud-based databases, and even external enterprise applications, and application programming interfaces, or API's.

Power BI service

- is the Cloud-based BI service or Software as a service part of Power BI. It is used by report users and administrators. Power BI apps is the native mobile application of Power BI.

- It's available on iOS, android, and Windows. With these components and interfaces, Microsoft's Power BI enables users from various disciplines, such as Report Designers, administrators, and business users, to use a product according to their roles.

Power BI Workflow

- the order in which you use these components is known as a workflow.
- A Power BI workflow can be described as the steps taken with data to create, publish, and share.
  1. A typical workflow in Power BI often starts with **the creation of a report** in **Power BI Desktop**. Report designers and developers are primarily responsible for this task.
  2. Once the report is ready, you **publish** it to the **Power BI service**, where administrators can assign permissions, and specific users can consume the reports.

Now, let's examine each step of the workflow in more detail.

- **Create** is about importing data and creating a report. This step is when you import your data sources into **Power BI Desktop**, clean, transform, and load your data to have targeted data for your reports. Use your filter data to create a report and analyze and present your data using various visualizations and charts in your reports.
- Then you move on to the **publish** step of the workflow, where you publish reports and create dashboards. That means you publish a report to the Power BI service and share your data with others by creating dashboards and use different visualizations and filters to make your data more understandable in your dashboard.
- The final step of this workflow is **sharing**. In this step, you share dashboards with users and manage access to your data. Share your dashboards with the users needed to make it easier to collaborate on projects. Manage access to your data by ensuring that dashboard's have different user permission levels.
- This is also where you consider **mobile usage**. For instance, using Power BI mobile apps, you can view and interact with reports and dashboards that have content pinned from reports anytime and anywhere. You can use different features of the mobile apps to explore and share your data from different perspectives.

In summary, a typical Microsoft Power BI workflow sequences the requirements needed to choose data sources and types in step 1, and then step 2 is used to visualize the data. The third and final workflow step presents the resulting reports and dashboards to cater to different user types and their requirements. Using such a workflow, you combine different types of data from many sources using various components, such as Power BI Desktop, Power BI service, and Power BI apps.

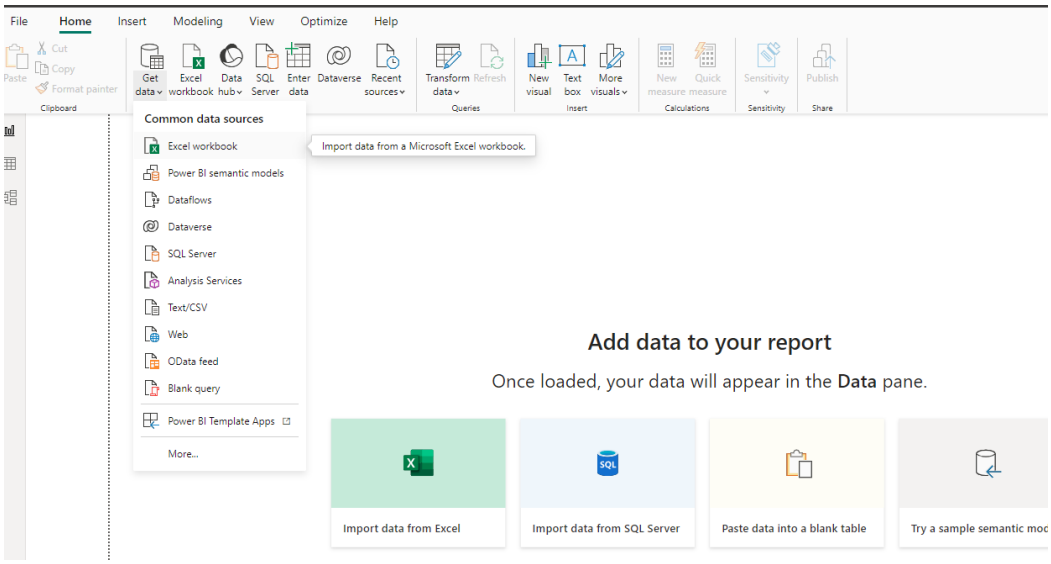
Question

Which of the following is the SaaS (Software as a Service) based web application of Microsoft Power BI?

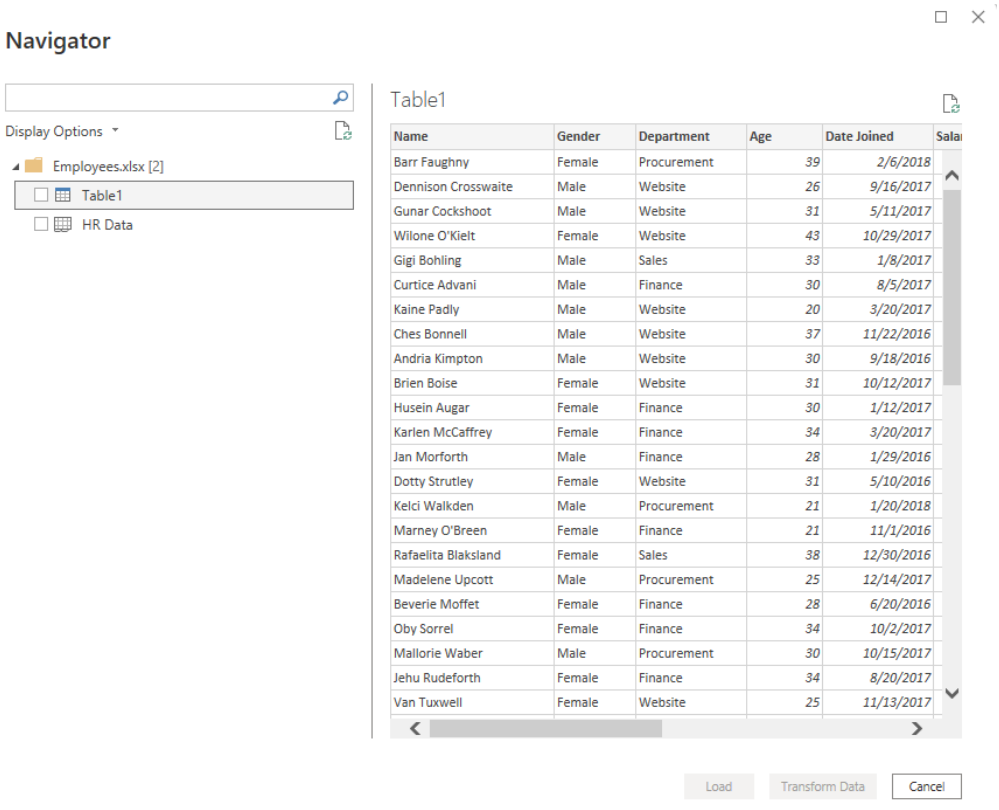
A. Power BI Service
B. Power BI Apps
C. Power BI Desktop

Exercise 1 A: Getting Data

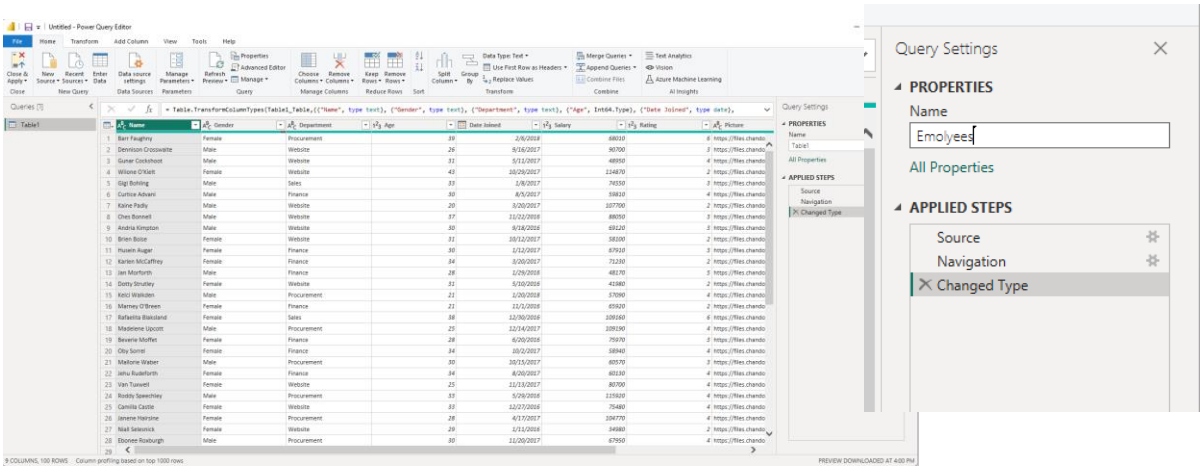
1. Use the Excel file: **Employees.xlsx** in Lab folder.
2. Explore the data in the workbook and .
3. Notice you have a worksheet with the Name **HR** and a table with name **table1**.
4. Explore columns you have.
5. Open Power BI.
6. Close the **splash** screen.
7. From **Home** Ribbon in **Data** Group Select **Get Data**.



8. Select Excel Workbook.
9. Browse to your Employees workbook to open.
10. In the **Navigation Pane** notice that on the left you have the tables that are available in the source and when you click a table you get a preview on the right side.
11. You have a sheet and a table with the same date.
12. Select **table1**.

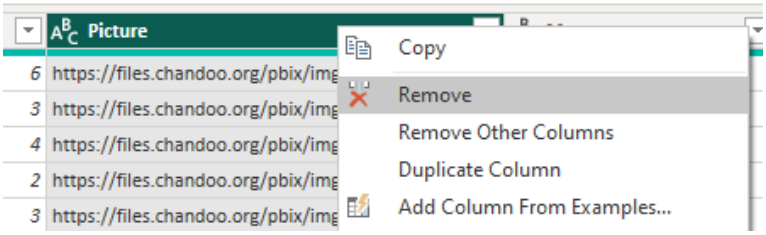


13. You can select Loade to **load** to your data model directly.
14. But select **Transform** to go to power query

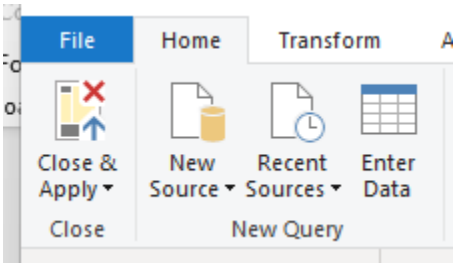


15. The Power Query is opened in a separate window.
16. Explore Power Query.
17. You have your table in the middle of the screen.
18. You have Ribbons to help you to transform data.
19. On the left you have Queries Pane.
20. In right you have Query settings with Properties and the Applied Steps.
21. Change the name of your Query to **Employees**.

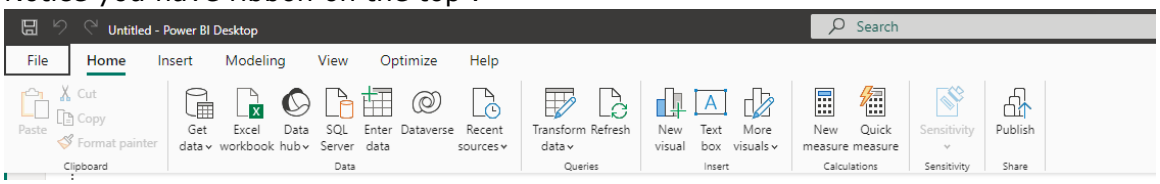
22. We do not need the Picture column.
23. Right click then choose to remove.



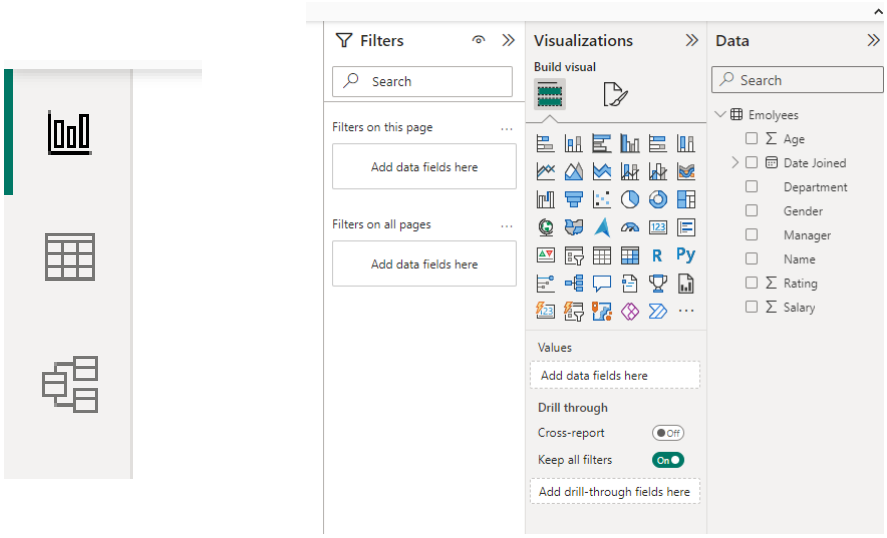
24. Notice that you have the step added to the **Applied steps**.
25. You can undo what you have done by just deleting the step from the Applied steps.
26. Do and redo your removing column.
27. You can now click close and Apply in the Home tab to close power query and get back to the Power BI.



28. You get back to your Power BI File.
29. Notice you have ribbon on the top .

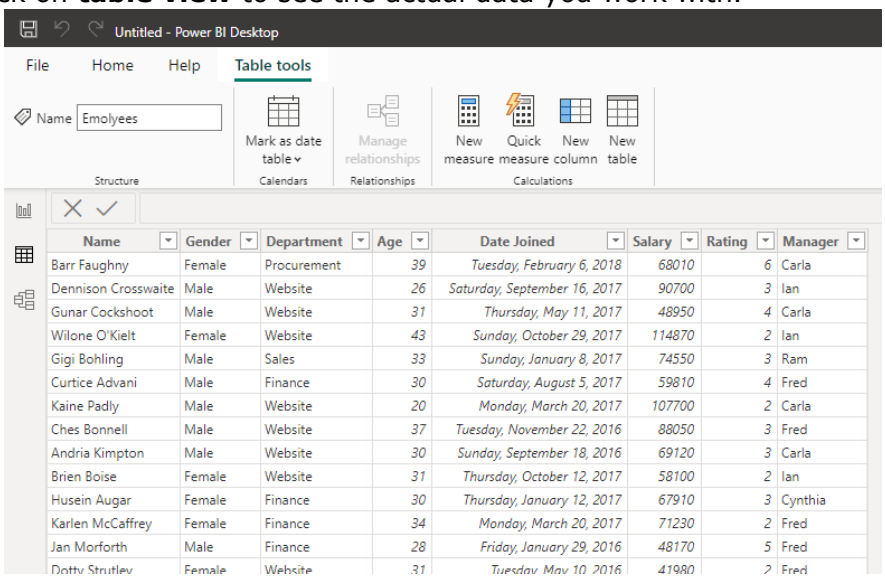


30. You have Pans on the right : Data, Visualization , and Filter.
31. And on the left you have the three icons of the 3 view of power BI which are : **Report View ,table View and Model view.**



32. Report view is where you do your visulization.

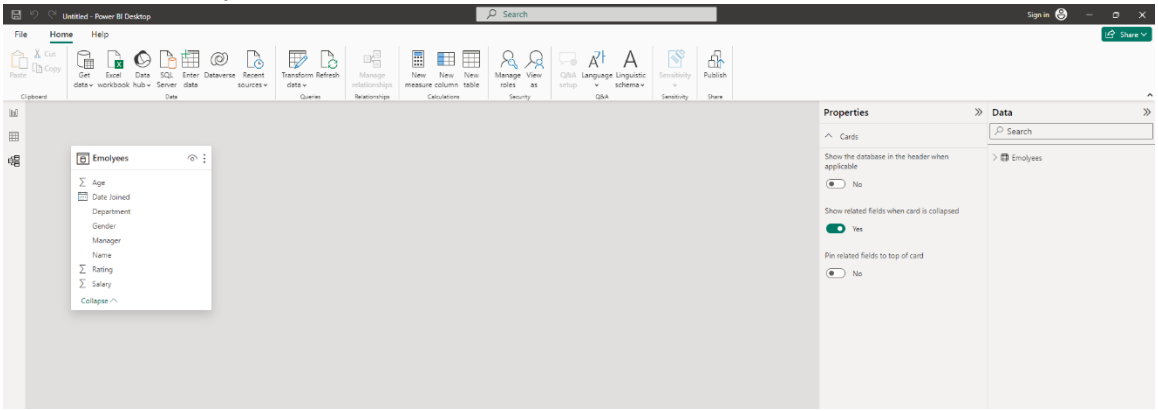
33. Click on **table view** to see the actual data you work with.



Name	Gender	Department	Age	Date Joined	Salary	Rating	Manager
Barr Faughny	Female	Procurement	39	Tuesday, February 6, 2018	68010	6	Carla
Dennison Crosswaite	Male	Website	26	Saturday, September 16, 2017	90700	3	Ian
Gunar Cockshoot	Male	Website	31	Thursday, May 11, 2017	48950	4	Carla
Wilone O'Kielt	Female	Website	43	Sunday, October 29, 2017	114870	2	Ian
Gigi Bohling	Male	Sales	33	Sunday, January 8, 2017	74550	3	Ram
Curtice Advani	Male	Finance	30	Saturday, August 5, 2017	59810	4	Fred
Kaine Padly	Male	Website	20	Monday, March 20, 2017	107700	2	Carla
Ches Bonnell	Male	Website	37	Tuesday, November 22, 2016	88050	3	Fred
Andria Kimpton	Male	Website	30	Sunday, September 18, 2016	69120	3	Carla
Brien Boise	Female	Website	31	Thursday, October 12, 2017	58100	2	Ian
Husein Augar	Female	Finance	30	Thursday, January 12, 2017	67910	3	Cynthia
Karlen McCaffrey	Female	Finance	34	Monday, March 20, 2017	71230	2	Fred
Jan Morforth	Male	Finance	28	Friday, January 29, 2016	48170	5	Fred
Dotty Strutley	Female	Website	31	Tuesday, May 10, 2016	41980	2	Fred

34. Click on **Model view** to see Your data model

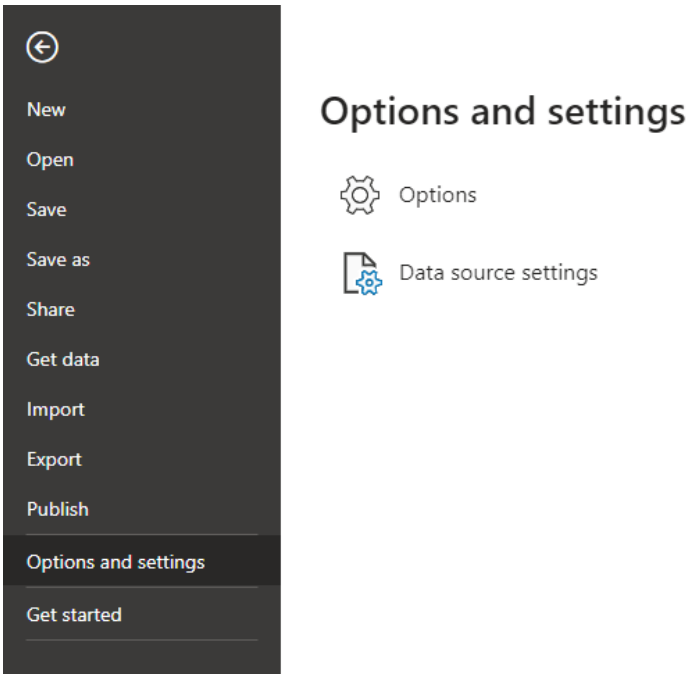
35. Notice is only made of one table this time.



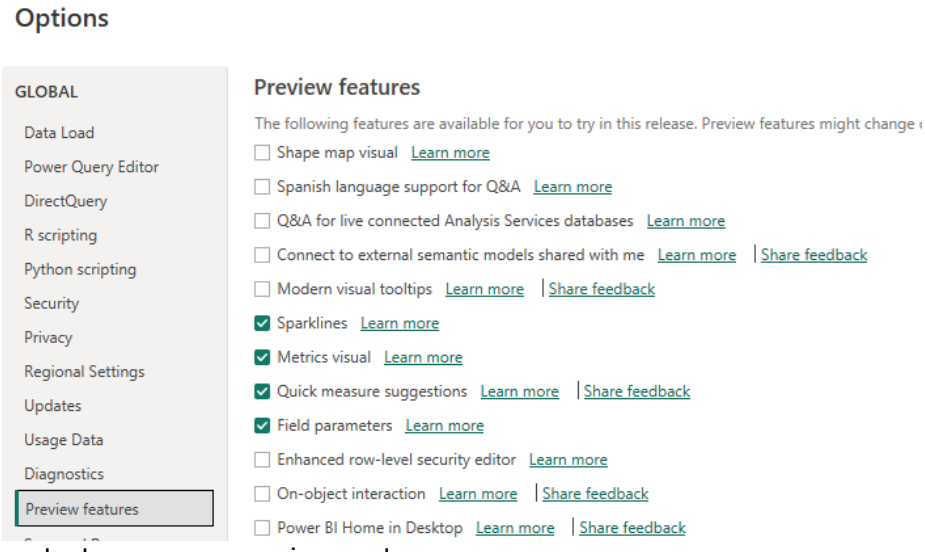
36. Go back to Report View.
37. Notice that is an old View of Power BI.

Exercise 1B: Change Power BI Settings to on Object Interaction

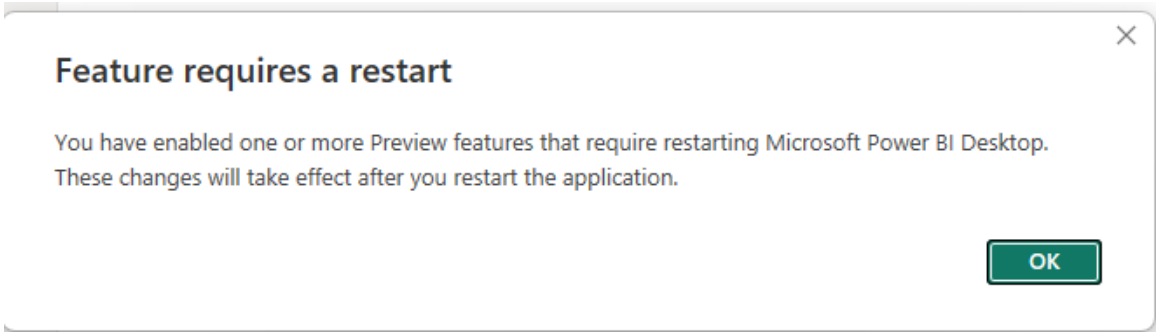
38. Let us move our view to On Object Interaction View.
39. First Save your File as **My First Power BI Report.pbix.**
40. Go to File ➔ Option and settings ➔ Options.



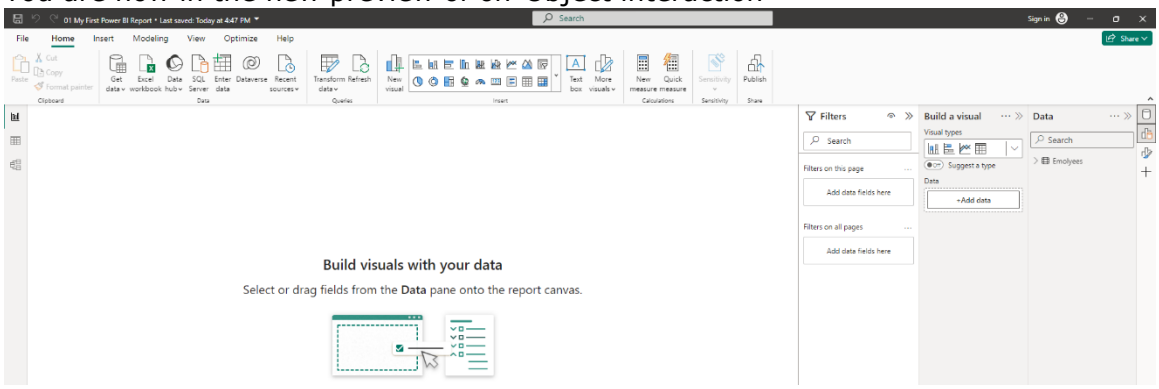
41. In Google ➔ Preview Features ➔ select **On-Object Interaction**



42. You got a restart message requirement.



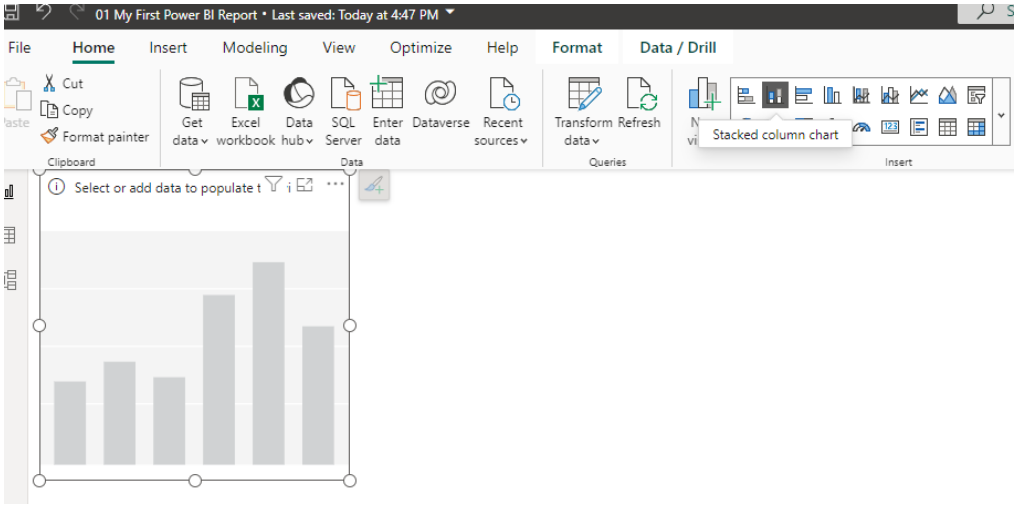
43. Close and reopen Power BI.
44. You are now in the new preview of on-Object interaction



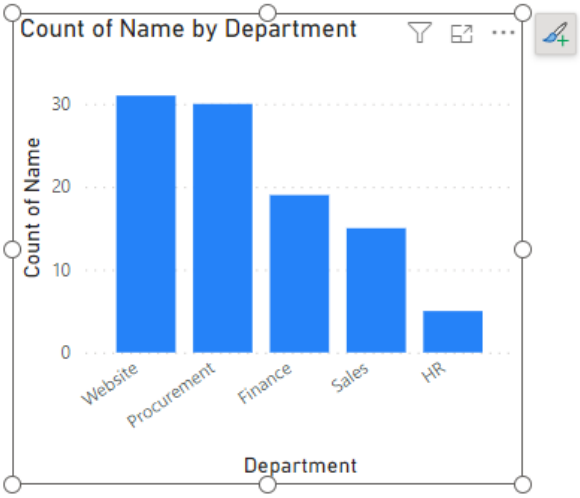
45. Notice that Visuals are now on the top and you have a new interface.

Exercise 1C: Creating Bar Chart

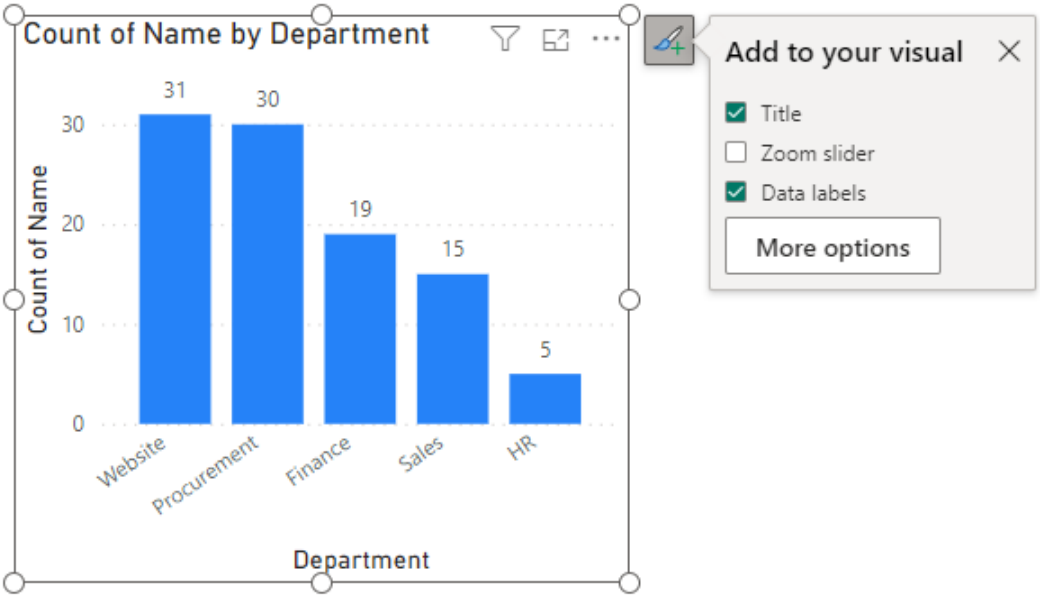
46. Select Stacked column visual to appear in your report Canvas.



47. Your boss wants to know How many employees in each department.
48. Drag Department from data pane to X-axis in Build pane , and drag name to Y-axis (it will then count).
49. You have now the count of employees in each department.



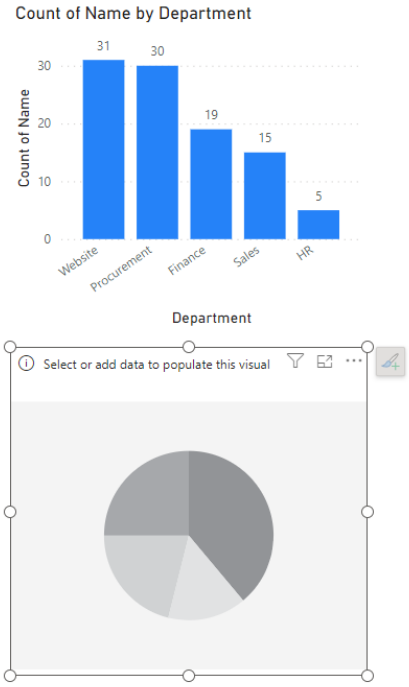
50. Click on the icon on the right of the Visual (add and remove to your visual) and select Data Labels.
51. The Number of employees is now Appear on top of each column.

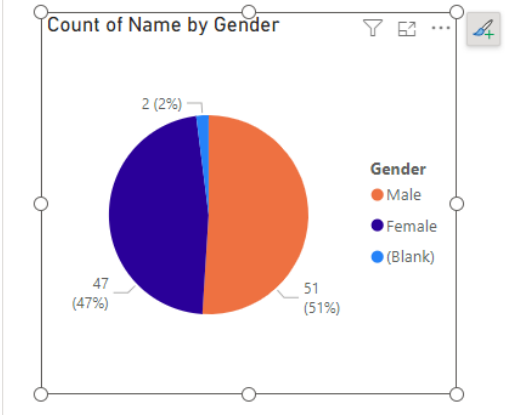


Exercise 1 D: Create Pie Chart

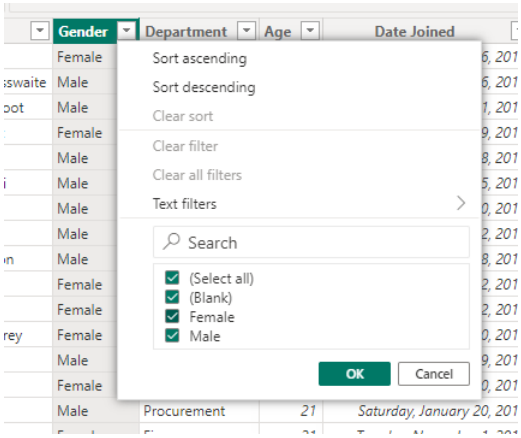
52. We want to what is the ratio between men and women in the company
53. The suitable visual here is Pie Chart.
54. First Click on a free space in Canvas.
55. Then click on Pie from the visual so it is in your Report.

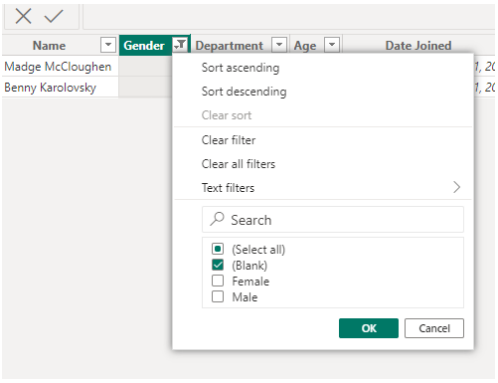
56. Drag Ginder to legend and Names to Values, so Power BI count them.





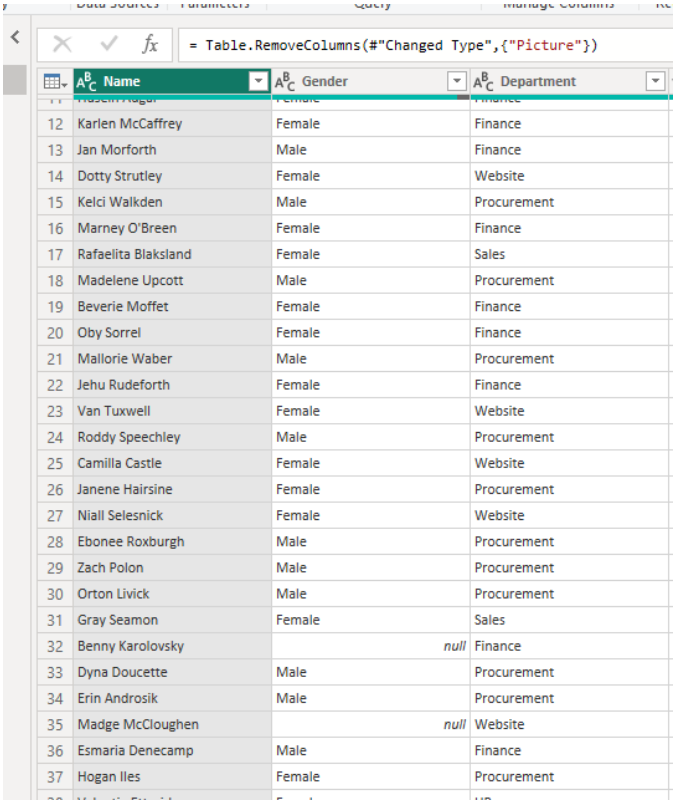
57. Notice you have 3 Gender in the legend **Male**, **female** and **Blank**.
58. That means you have a missing value in this column.
59. Go to **table view** and check the **Gender** column.



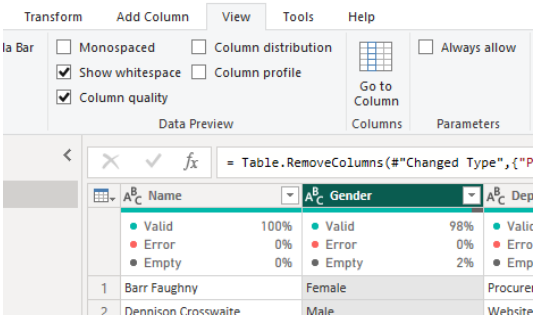


60. Notice you have blank value.
61. You can filter the view to see them.
62. Try to change the value of those two rows, you cannot.
63. To transform the data, you must use power query editor.
64. From the **home** tab in data ribbon chose **Transform data** to open power query again.

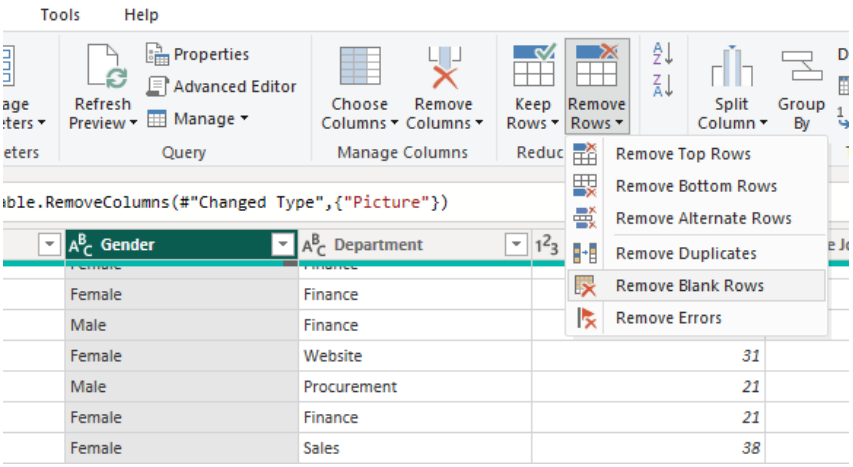
65. Notice you have two null values in the column.
66. Also, if you go to **view** tab ➔ **Data Preview** ➔ check **column quality**.
67. You get the % of **valid, error** and **empty** values.



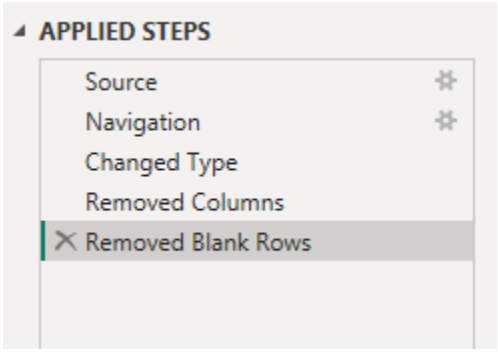
	Name	Gender	Department
12	Karlen McCaffrey	Female	Finance
13	Jan Morforth	Male	Finance
14	Dotty Strutley	Female	Website
15	Kelci Walkden	Male	Procurement
16	Marney O'Breen	Female	Finance
17	Rafaelita Blaksland	Female	Sales
18	Madelene Upcott	Male	Procurement
19	Beverie Moffet	Female	Finance
20	Oby Sorrel	Female	Finance
21	Mallorie Waber	Male	Procurement
22	Jehu Rudeforth	Female	Finance
23	Van Tuxwell	Female	Website
24	Roddy Speechley	Male	Procurement
25	Camilla Castle	Female	Website
26	Janene Hairsine	Female	Procurement
27	Niall Selesnick	Female	Website
28	Ebonee Roxburgh	Male	Procurement
29	Zach Polon	Male	Procurement
30	Orton Livick	Male	Procurement
31	Gray Seamon	Female	Sales
32	Benny Karolovsky	*null*	Finance
33	Dyna Doucette	Male	Procurement
34	Erin Androsik	Male	Procurement
35	Madge McCloughen	*null*	Website
36	Esmaria Denecamp	Male	Finance
37	Hogan Iles	Female	Procurement



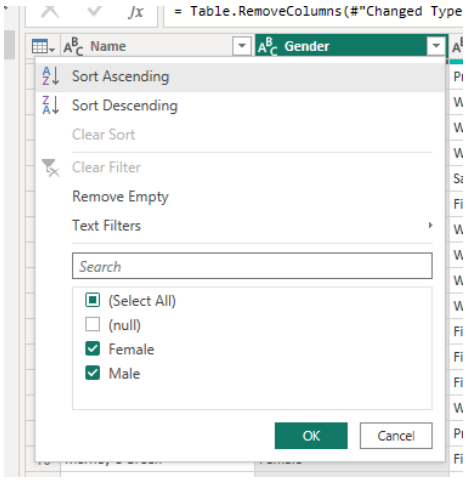
68. You must decide now.
69. You can delete those rows.
70. Click Home ➔ Remove Rows ➔Remove Blank Rows.

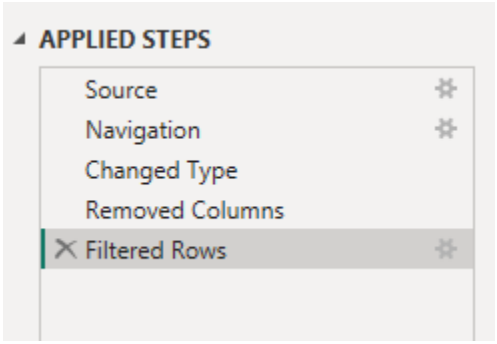


71. It is not a good decision.
72. Go and delete that step from applied steps pan in the right.



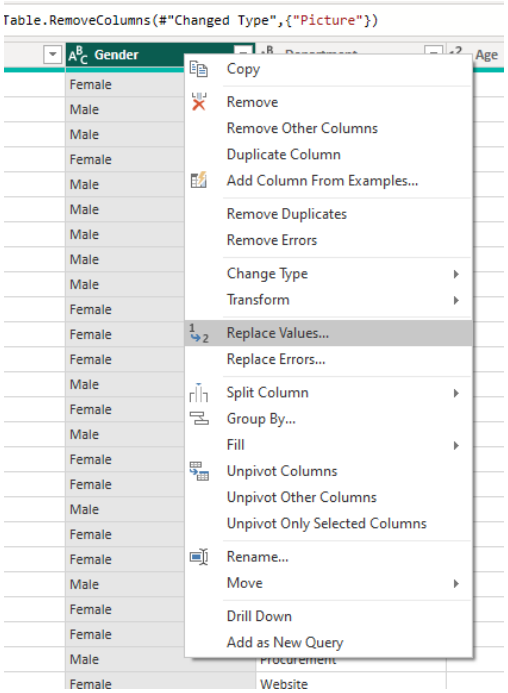
73. You can do the same thing by filtering the value in the Gender column.



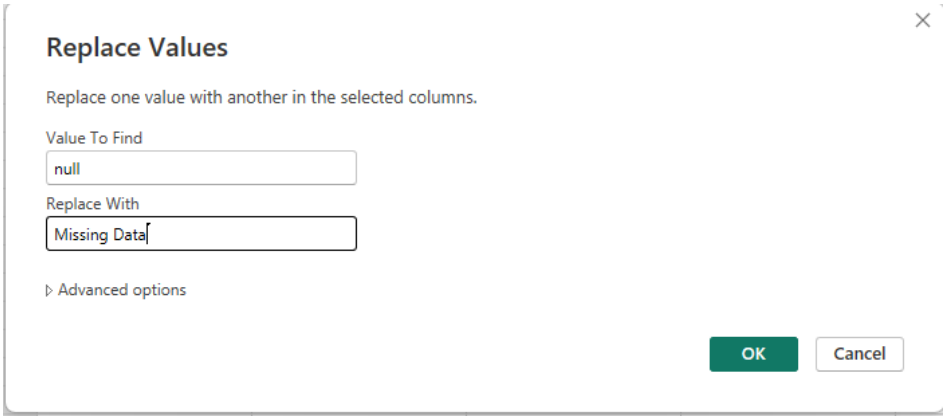


74. Also, it is not a good decision, go and delete this step too.
75. The best thing is to put value in this blank like "**missing data"**.

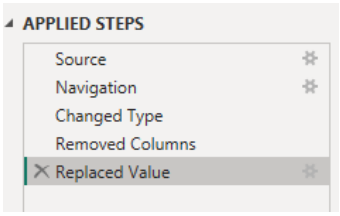
76. Right click the Gender column and choose Replace Value.

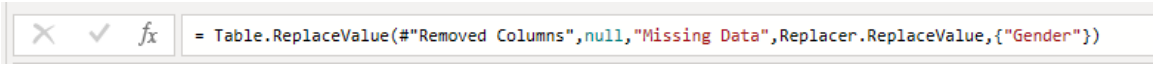


77. In the replace value dialogue box replace null with missing data.



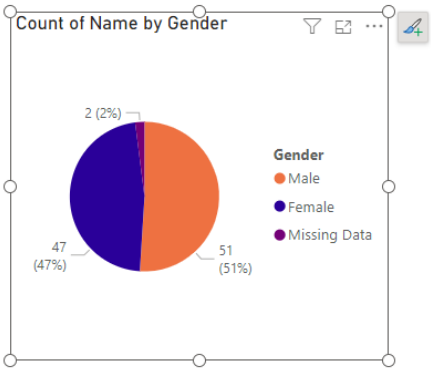
78. It is a chance to see the applied step pan.
79. Click every step and watch the corresponding **M language** line in the formula bar.





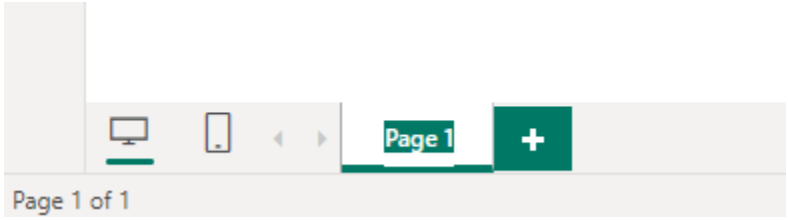
80. You can also go to **Home** tab ➔ **Query** Group ➔ **advanced editor**.
81. You can here see and edit the steps manually in **M language**.
82. From **home** tab click **Close & Apply** to go back to power BI.
83. Notice now it appears in **table view** and **repot view** in the Pie chart.

Niall Selesnick	Female	Website
Ebonee Roxburgh	Male	Procurement
Zach Polon	Male	Procurement
Orton Livick	Male	Procurement
Gray Seamon	Female	Sales
Benny Karolovsky	Missing Data	Finance
Dyna Doucette	Male	Procurement
Erin Androsik	Male	Procurement
Madge McCloughen	Missing Data	Website
Esmaria Denecamp	Male	Finance

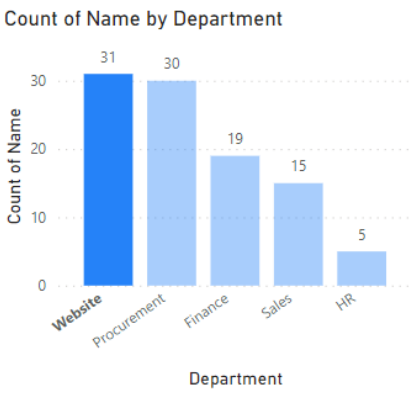


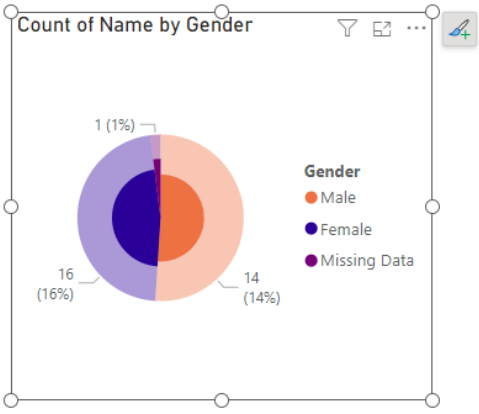
Exercise 1 E: Visuals interactivity

84. Now you have two visuals in your report.
85. It is Page one of your report.
86. Double click on the page 1 in bottom and rename it to **Employees**.



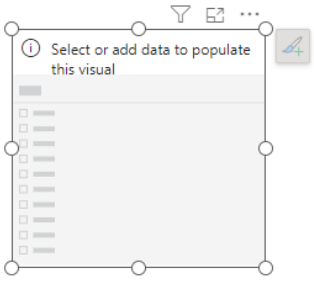
87. You can also click on + button and add a new page.
88. Click on Column Website department in the column chart.
89. You will see that pie chart reflects that and show only value of that department only.



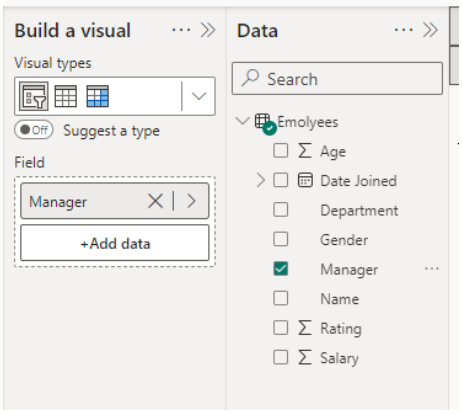


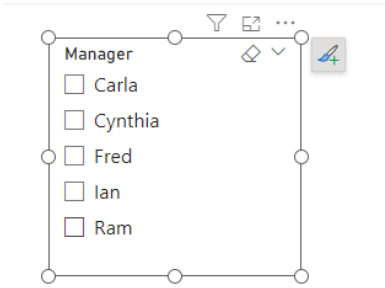
90. Go and click on every department to see their values in pie chart.
91. Now we want to filter by manager

92. You can add visual slicer and add manger to it.

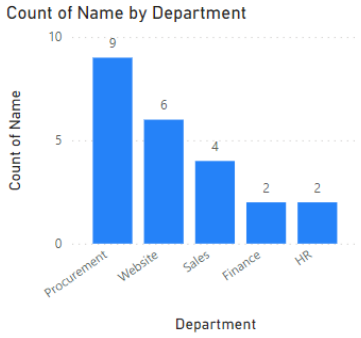


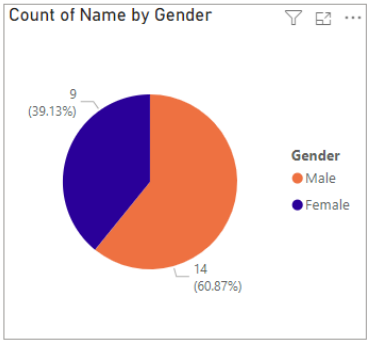
93. Make sure you always click on empty part of canvas before you add any new visual.

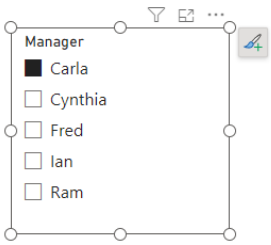




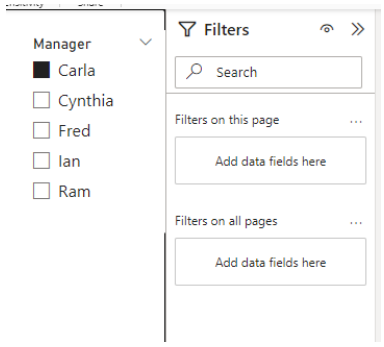
94. Now you can filter the two visuals with your new slicer.







95. Notice you can do that by using filter pane to save space, but slicers are so intuitive for the users of your report.

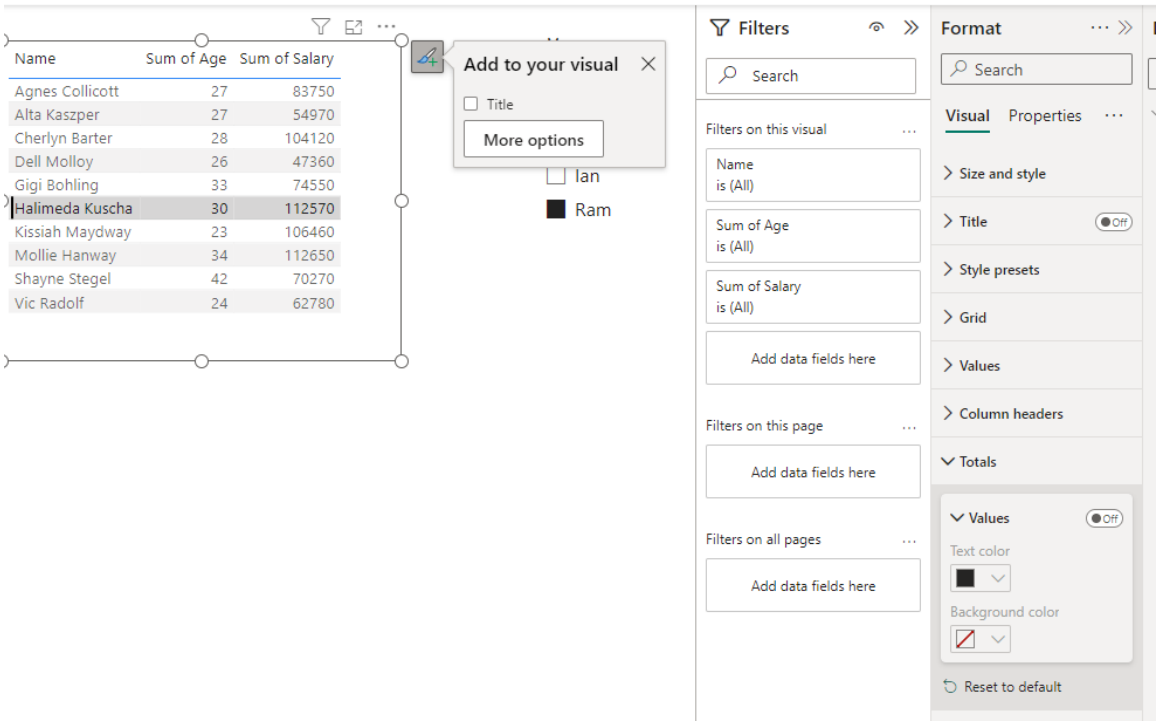


Exercise 1F: Creating table and format results

1. Your boss liked your report, but it asked you to see the employees of each manager.
2. Add a table to your report
3. In Data pane click on Name, age, rating and salary.
4. You must now have your table like this.

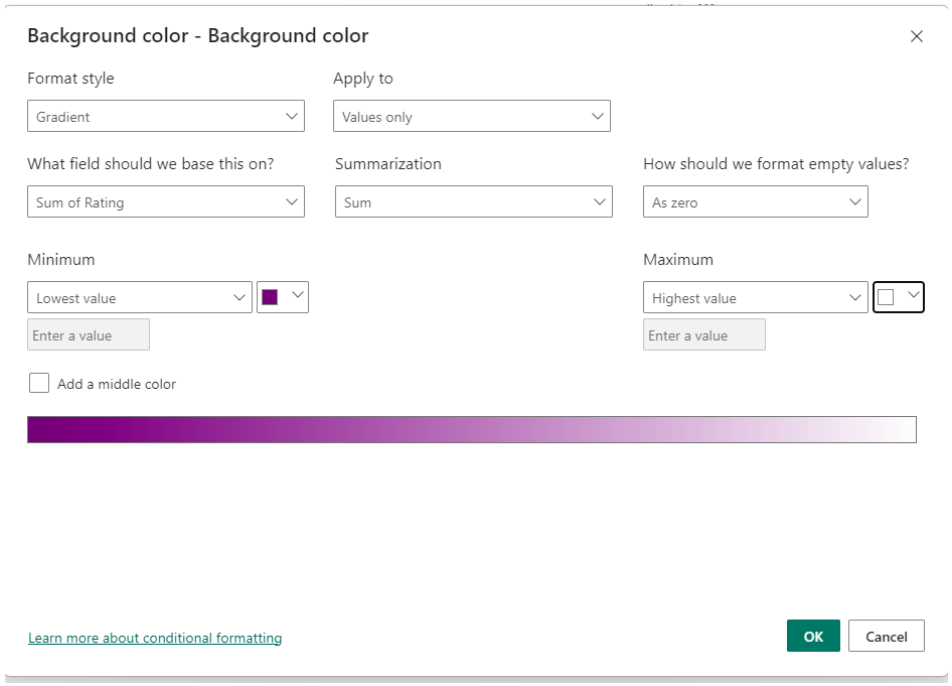
Name	Sum of Age	Sum of Salary
Agnes Collicott	27	83750
Alta Kaszper	27	54970
Cherlyn Barter	28	104120
Dell Molloy	26	47360
Gigi Bohling	33	74550
Halimeda Kuscha	30	112570
Kissiah Maydway	23	106460
Mollie Hanway	34	112650
Shayne Stegel	42	70270
Vic Radolf	24	62780
Total	**294**	**829480**

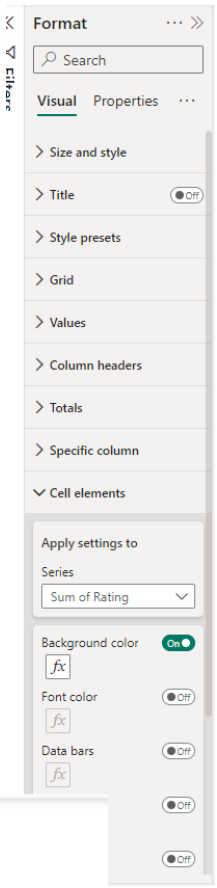
5. Go to icon in right of table (format) ➔More Options ➔Visual tab ➔Totals ➔ Values and make it **off**.



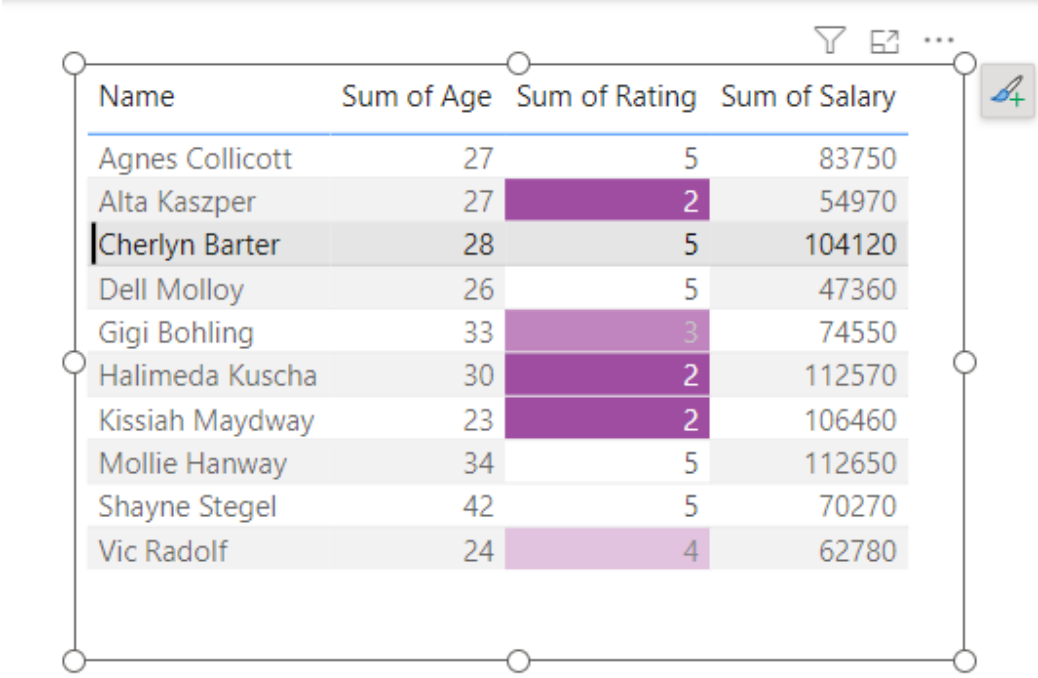
6. Notice you have search bar in format pane you can search for what you want.
7. Your boss is happy now he can see every manager employee.
8. He asked you to see which employees have low ratings.
9. He wanted to having conditional formatting for rating column.
10. Go to Format pan ➔ Visual➔Cell elements➔Apply setting to ➔ choose rating ➔Make back ground color **on** and thin click the fx icon.

11. In conditional formatting dialogue box change color for min and max.





12. You can also format font color

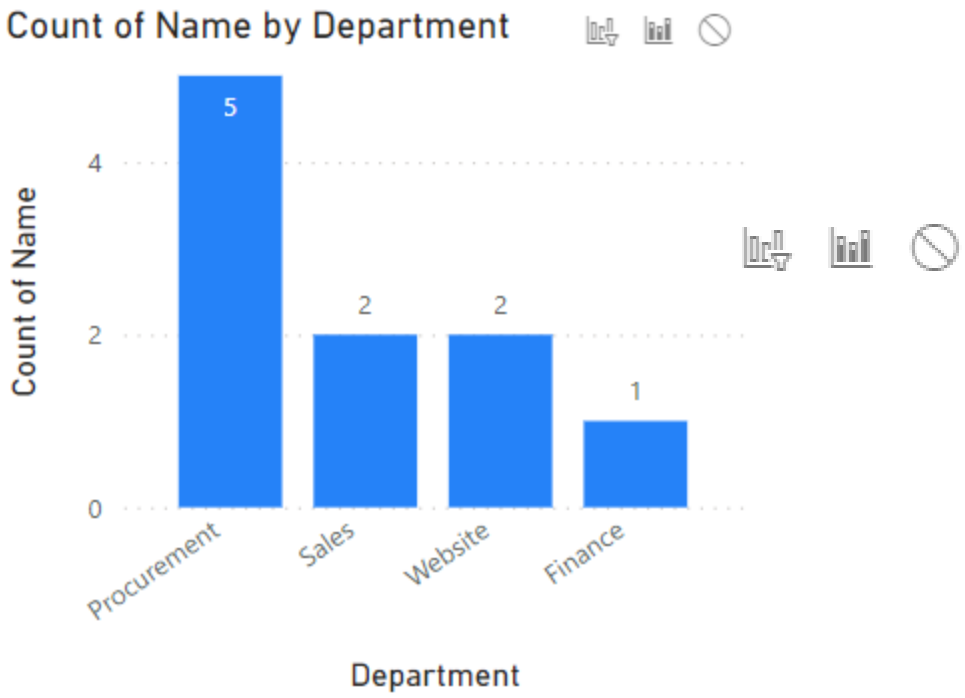


Name	Sum of Age	Sum of Rating	Sum of Salary
Agnes Collicott	27	5	83750
Alta Kaszper	27	2	54970
Cherlyn Barter	28	5	104120
Dell Molloy	26	5	47360
Gigi Bohling	33	3	74550
Halimeda Kuscha	30	2	112570
Kissiah Maydway	23	2	106460
Mollie Hanway	34	5	112650
Shayne Stegel	42	5	70270
Vic Radolf	24	4	62780

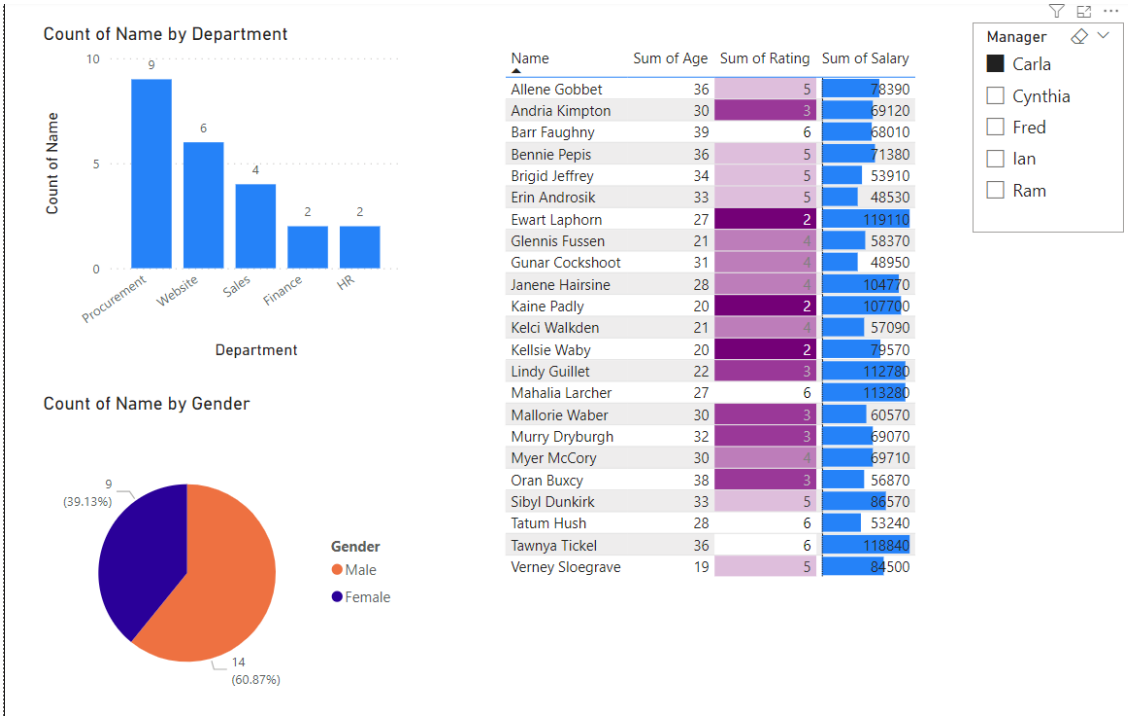
13. Your boss wanted to highlight the high salary.
14. Chose salary this time and use the data bars options and make it on.
15. You must have your table now like this

Name	Sum of Age	Sum of Rating	Sum of Salary
Agnes Collicott	27	5	83750
Alta Kaszper	27	2	54970
Cherlyn Barter	28	5	104120
Dell Molloy	26	5	47360
Gigi Bohling	33	3	74550
Halimeda Kuscha	30	2	112570
Kissiah Maydway	23	2	106460
Mollie Hanway	34	5	112650
Shayne Stegel	42	5	70270
Vic Radolf	24	4	62780

16. Notice that table also when you click on an employee it filters the other two charts which has no meaning.
17. Stop that as follow as you select the table chart:
    - Go to Format tab ➔Indicators group ➔ edit indicators.



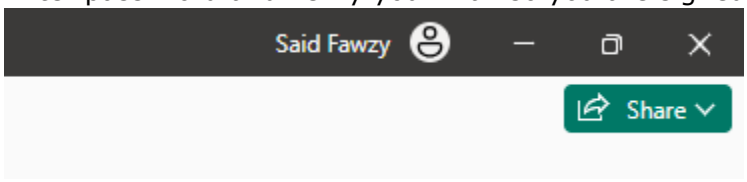
    - This shows new icons on top right of other visuals.
    - They are Filter, Highlight, **non**.
    - Chose non for all other visuals.
    - Click edit indicators again to remove these icons.
18. Now your final report like this:

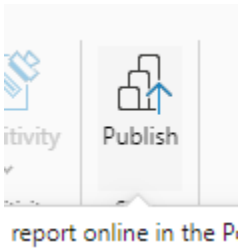


19. You can click any column Header to sort.
20. Save your work.

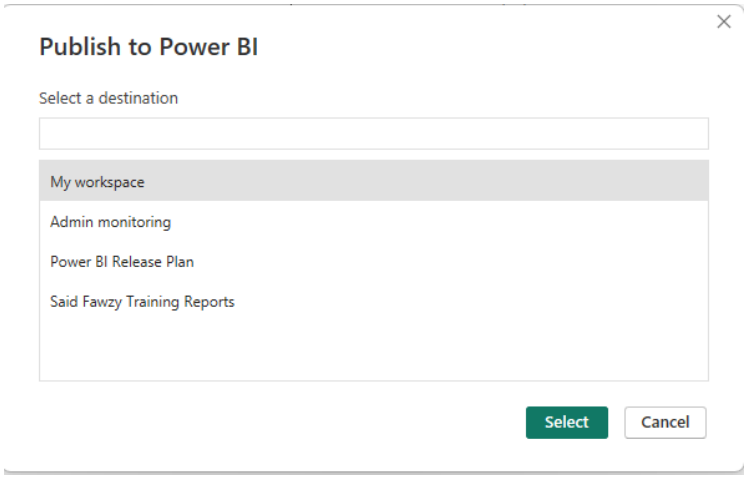
Exercise 1G: Publish and share your report.

21. Click sign in in top right of the screen.
22. Enter your business mail or developer mail and press continue.
23. Enter pass word and verify your mail so you are signed in

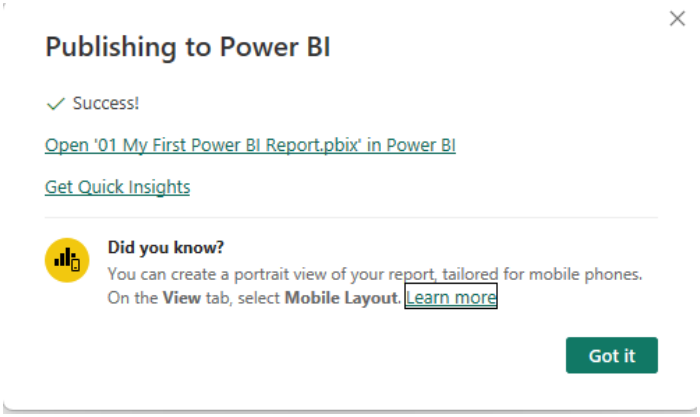


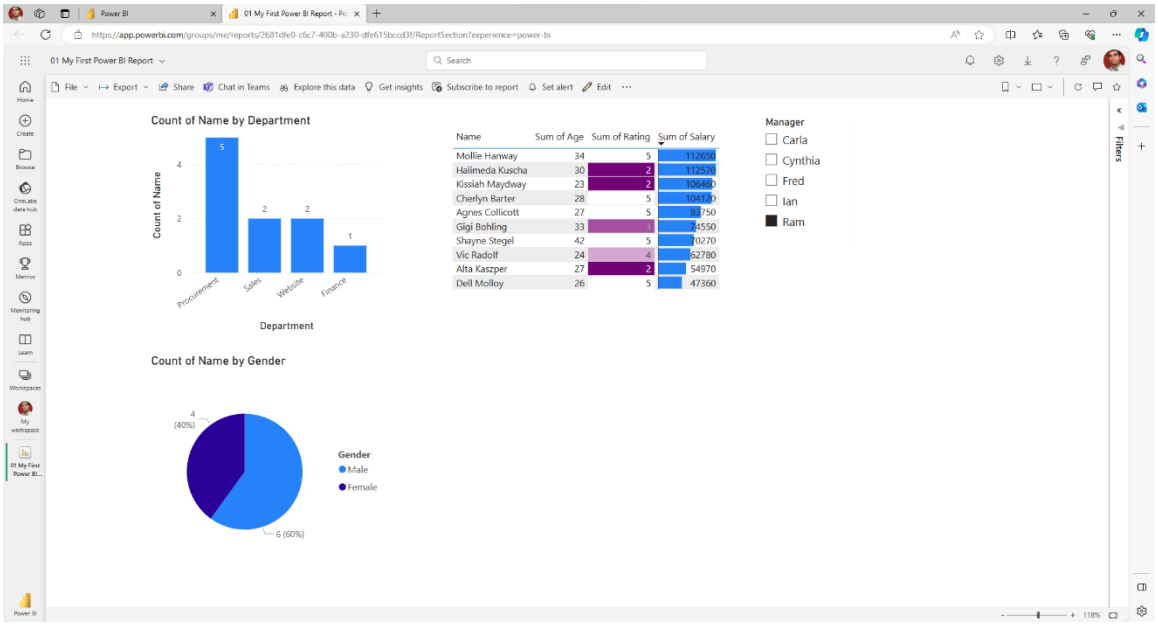


24. on home tab click **publish**.
25. Select your destination workspace



26. Then click the link to go to your report





27. Now your report is published in your workspace, and you can share it if you have pro account

Knowledge Chek

Question 1

Which product has strong reporting features and is typically used to begin a workflow in Power BI?

A. Microsoft Power BI Apps
B. Microsoft Power BI Service
C. Microsoft Power BI Desktop

Question 2

You want to publish your report and share your data with others by creating dashboards. Which of the following products would you use to accomplish this?

A. Microsoft Power BI Desktop
B. Microsoft Power BI Service
C. Microsoft Power BI Apps

Question 3
True or False: The typical workflow in Microsoft Power BI starts with the creation of a report in Power BI Desktop.

A. True
B. False

Chapter 3: Connecting to Data Sources

The ETL (Extract, Transform, Load) process

- Have you ever tried to solve a jigsaw puzzle when the pieces are scattered everywhere, and you don't even know if those pieces belong to the same puzzle?
- That's what it can feel like as a data analyst tasked with extracting insights from data that spread across multiple sources, formats, and structures.
- Not to worry, there's a way to solve this problem. The **Extract**, **Transform**, **Load**, or **ETL** process.

ETL Components

- **ETL** stands for extract, transform and load, the names given to the three main steps in the ETL process.
- This process involves taking raw data from various sources, preparing it for analysis, and loading it into a repository or data storage and management system.

Extract

- **Extract** is the first step in the ETL process, which involves retrieving and extracting raw data from different sources, such as databases, files, or other data storage systems. For example:
    - Customer Relationship Management, or **CRM**.
    - Enterprise resource planning system, or **ERP**
    - Spreadsheets.
- The extraction process involves pulling the data from these different sources.
- Then, you consolidate it into an easily accessible central location, often a temporary intermediate storage location known as the **staging area** and prepare it for further processing in the next step.

Transforming

- Once the data is extracted, the second step is to transform it.
- Transforming the data involves cleaning, structuring, and enriching the data to make it more suitable for analysis.
- This may involve:
    - removing duplicates,
    - handling missing values,
    - creating new calculated fields,
    - converting data types, and
    - standardizing measurement units. let's say that the sales and marketing data is in US dollars. But the manufacturing and purchasing data is in different currencies, depending on where in the world the sales or purchase take

place. As part of transforming the data, you may need to convert all the currency values into a standard unit of measurement, in this case US dollars, to ensure consistency.

Load

- The third and last step involves loading the transformed data into the final storage system, typically a data warehouse. Where it can be readily accessed and analyzed, for example, using tools like Power BI.
- Depending on the organization's needs, the loading process can be a one-time event or scheduled to run regularly.

ETL Benefits

- The ETL process ensures that the data analyzed is **accurate**, **clean**, and **consistent**, which in turn **supports informed decision making**.
- This process offers many benefits, including:
    - **Data integration**: ETL helps integrate data from different sources, providing a unified view of an organization's data, making it easier for analysts to perform analysis and derive insights.
    - **Data Quality**: ETL processes involve data cleansing and validation, which significantly improve data quality.
    - **Data consistency**: By transforming data into a standardized format, ETL ensures consistency across various datasets, enabling analysts to easily compare and analyze data from different sources.
    - **Enhanced performance**: By aggregating, summarizing, or indexing data during the transformation process, ETL can improve query performance and reduce the load on data analysis systems.
    - **Data governance**: ETL can support data governance initiatives by helping organizations maintain a single source for their data, ensuring that everyone has access to the same accurate information.
    -

ETL With Power BI

- Power BI is just one tool that comes equipped with built in ETL capabilities, enabling you to connect to many different data sources, transform your data using Microsoft Power Query, and load it into the Power BI data model.
- Power Query is a powerful ETL tool within Power BI, providing a graphical interface and formula language, called **M**, to perform various data transformation tasks.
- With Power Query, you can extract data from multiple sources, clean and structure it, and load it into Power BI for creating reports and visualizations.

Question

In the ETL process, which step involves retrieving raw data from different sources, such as databases and files?

A. Transform
B. Visualize
C. Load
D. Extract

Data sources that you can connect to in Power BI.

- **Flat files**: are a common type of data source that can be used for ETL or extract, load, and transform in Power BI.
    - Examples of flat files include **CSV**, **TXT**, and Microsoft **Excel** files.

Relational data sources

- such as **SQL Server**, **MySQL**, and **Oracle** Databases.
- commonly used by large organizations because they provide a high level of reliability, data integrity, and security.

- **NoSQL databases:**
    - such as **MongoDB** and **Cassandra**
    - becoming increasingly popular for ETL in Power BI.
    - These databases are designed to store and manage large volumes of **unstructured** or **semi-structured** data, making them ideal for use in a wide range of applications.

Combining Data Sources

- Power BI has the flexibility to connect to a wide range of data sources.
- By combining data from various sources such as sales figures, inventory, production, and supplier information, your department could gain valuable insights into customer behavior, product performance, and supplier performance.
- Combining data sources can benefit different stakeholders in a business by providing valuable insights into customer behavior, product performance, and supplier performance.
- This information can be used to make **informed decisions**, leading to improved supply chain management, reduced costs, and increased customer satisfaction.
- Data integration can be a daunting task, especially when you are working with multiple data sources that have varying formats, structures, and quality levels.
- The combination of these sources can often lead to inconsistencies and errors, making it difficult to derive meaningful insights and make informed decisions.
- But you don't need to worry. Tools like Power BI simplify the process of
    - combining data from different sources,
    - reducing the time and effort required to create a comprehensive view of your data.
- It is designed to be user friendly and accessible even for non-technical users, with
    - an intuitive interface and
    - drag and drop functionality that makes it easy to create reports and visualizations.
- Power BI also allows you to customize your reports and visualizations to suit your company's specific needs.
- You can choose from a wide range of pre-built templates and visualizations or create your own custom designs.
- This flexibility makes it easy to create reports that are tailored to the unique needs of your business.

- It also enables collaboration by allowing you to **share** your reports and visualizations with colleagues, clients, or stakeholders by sharing reports or embedding them in websites or apps.
- This collaborative approach can improve communication and ensure that everyone is working with the same data, ultimately driving business success.
- Combining data sources is a great method of providing valuable information that can lead to improved supply chain management, reduced costs, increased customer satisfaction, and ultimately drive business success.
- Tools like Power BI, with its built-in data connections, can simplify the process of combining data from different sources, reducing the time and effort required to create a comprehensive view of your business.

Question

You are setting up Power BI connectors for Adventure Works. What kind of services can the company use Power BI to connect to?

A. Power BI connectors only link to Microsoft services
B. Power BI connectors are limited to local computers or personal accounts.
C. Power BI connects to many external apps and cloud services.

Connecting to flat data source

- Every day, businesses generate large amounts of data. But where do they store it all?
- many organizations store and export data as files, such as flat files.

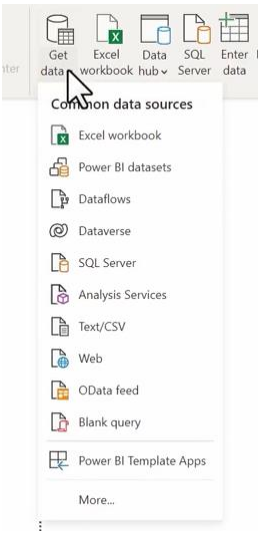
what is a flat file?

- A flat file is a file type that contains a single data table, with a uniform structure for every row of data, and does not have hierarchies.
- Some examples of flat files include:
    - comma separated value or CSV files,
    - delimited text or TXT files and
    - fixed width files. Additionally, output files from various applications such as Microsoft Excel Workbooks, can also be classified as flat files.

How to set up a flat data source

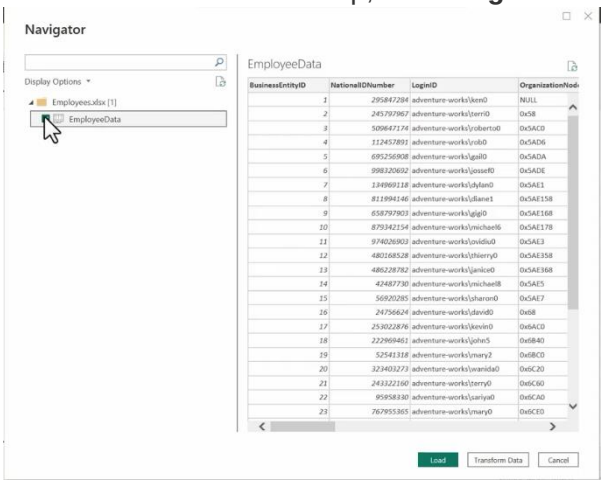
- The first step is to determine which file location you need to use to export the data.
- The file location is important, because when it is **changed**, Power Bi will not be able to refresh the data. This can cause errors, such as file not found, or Data source not found.
- Once you have located your file, you can proceed in Power Bi to display available data sources in the **Home** group of the Power Bi desktop ribbon.

- Select the **Get Data** button option, or down arrow to open the Common Data Sources list.



- If the data source you want isn't listed under Common Data Sources, select more to open the Get **Data dialog box**.

- If you need an Excel data source for example, select the Excel workbook you want, and select Open.
- When your file is connected to Power Bi desktop, the **Navigator window** opens.

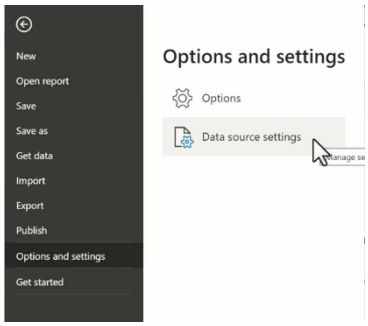


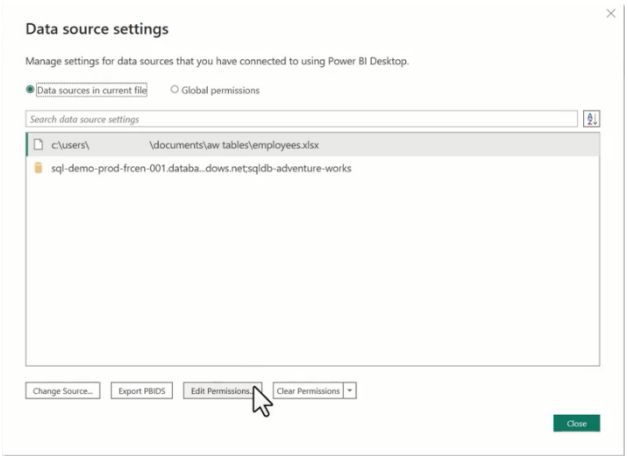
- This window displays the tables available in your data source. The excel file in this example.
- You can select a table to preview its contents, and to ensure that the correct data is loaded into the model.
- After selecting the checkbox of the table that you want to bring into Power Bi, it activates the Load button.
- Now you can select the Load button to import your data into the Power Bi dataset.

change the location of your source file

- In case you need to change the location of your source file for a data source during development. Or if your file storage location changes, you'll need to update your **connection strings** in Power Bi, to keep your reports up to date.

- To do this in Power Bi desktop:
    - select File in the menu bar,
    - then select Options and Settings from the File menu.
    - And now, select Data Source Settings from the options and settings menu.
    - You can also change or clear the permissions, by selecting Edit or Clear permissions, respectively.
    - Permissions cover the privacy level and credentials used for connecting to a data source.





- Remember that any structural changes to the file, can break the reporting model. So it's important to reconnect to the same file with the same file structure.
- By following these steps, you'll be able to ensure, that your report uses the most accurate and UpToDate information available.

Question

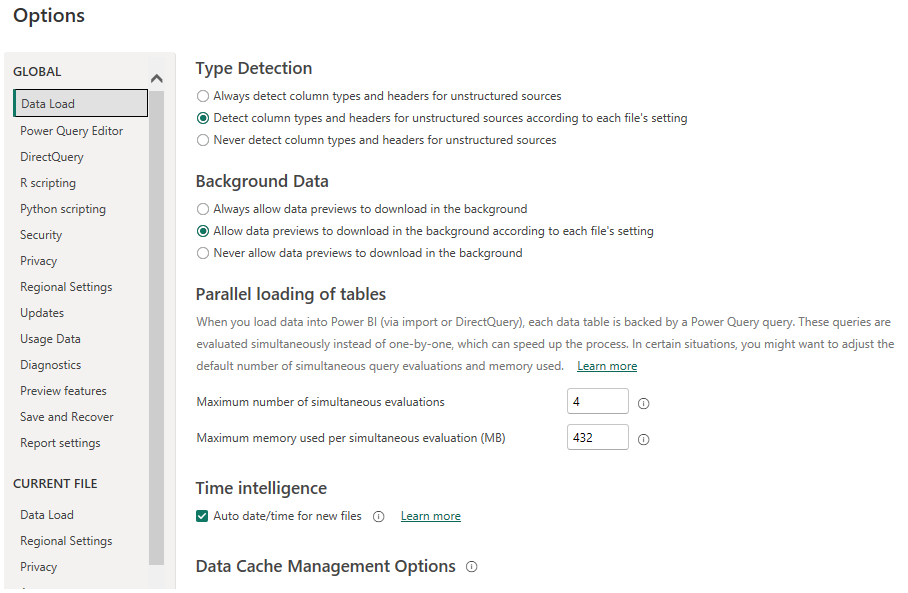
What should you do if you need to change the location of your source file in Power BI?

A. Select the correct option.
B. Clear permissions.
C. Update the connection string.
D. Create a new connection string and leave the old one unchanged.

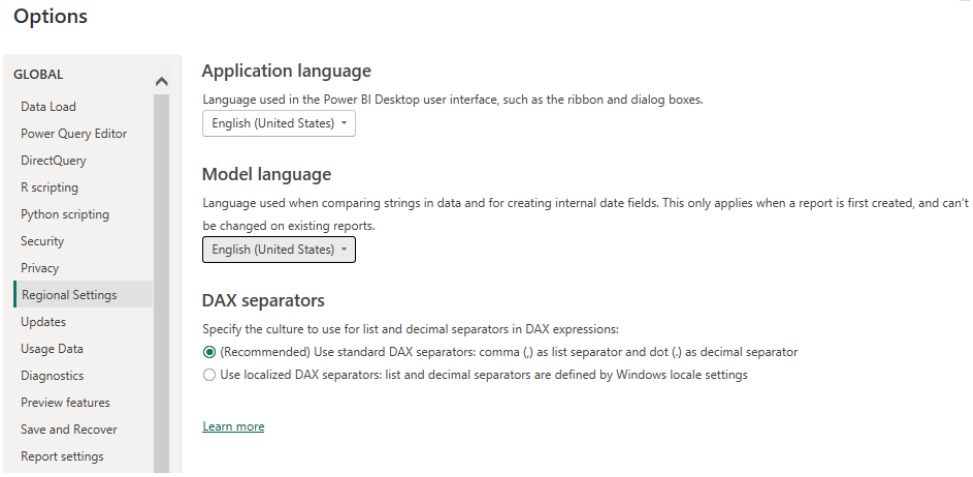
Exercise 2: Preparing Settings of Project File

1. Open Power BI.
2. Go to File ➔Options and Settings➔Options.

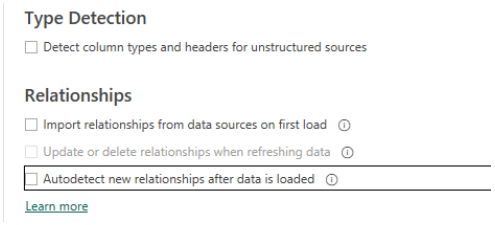
3. In option screen you have two Sections on left: Global and Current File.



4. In the Global section change in regional settings:
    a. Application Language= English United States.
    b. Model Language = English United States.

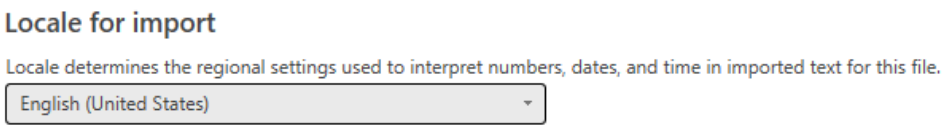


5. In Current File ➔Data Load Unselect detect column types and relationship



(That is because we want to do that manually)

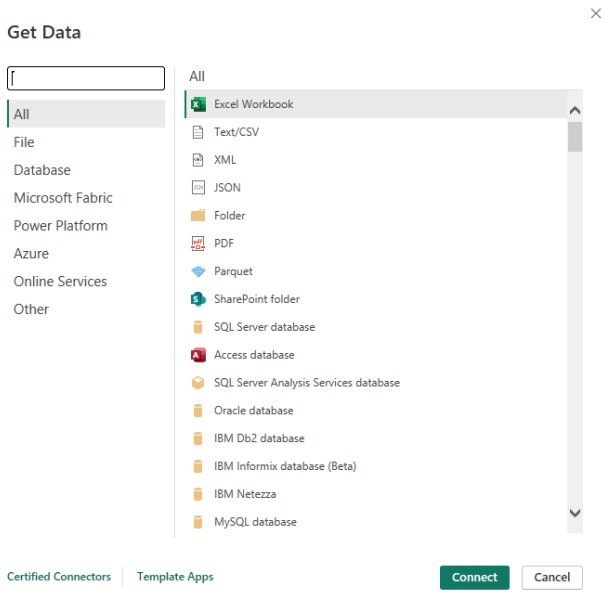
6. In Current File ➔Regional Settings Change Local for import to English Unitd States.



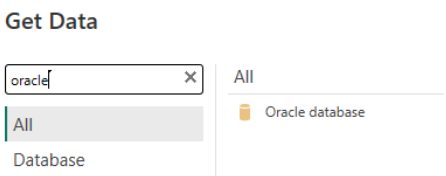
7. Click Ok you will get a message that you have to restart your application.
8. Save your file first as **Transforming Data Project.pbix**.
9. Close and reopen your saved file.

Exercise 3: Connecting to Data Source

1. Use the file: **Transforming Data Project.pbix** that you have created in Exercise 2
2. Use the Get Data Button on Home tab.
3. You can either:
    a. Click in upper part ➔ Get the full options of connecting to data sources.



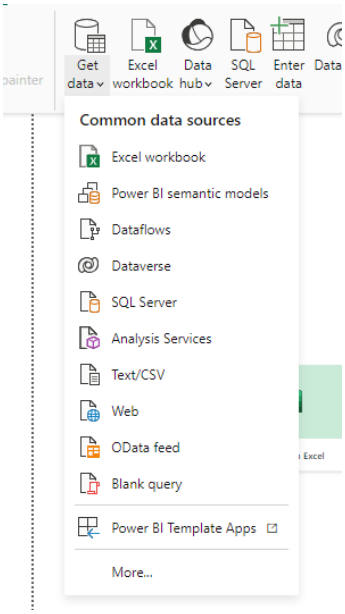
    b. You can search for the data source you want to connect to in the search box.
    c. Write "oracle" in the search box, the list is filtered only to Oracle Database connector.



    d. Try write "Excel" and you will get the Excel connector.
    e. Click the X icon on search to infilter the list.

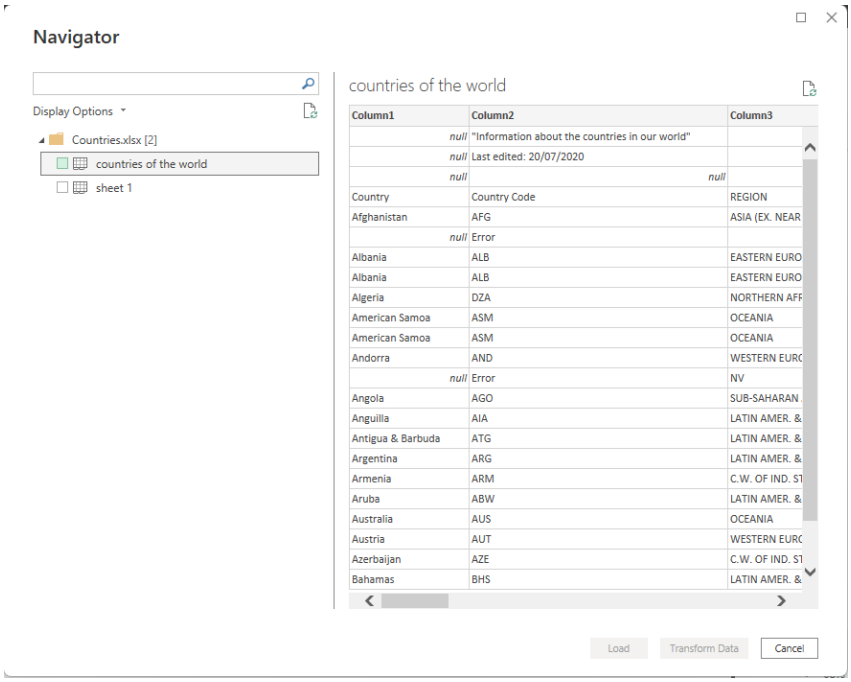
f. You can also use the Categories available in the left pan to get only those connectors of this category.
g. Notice "**All**" Categories is selected.
h. Click on **File** Category to filter the right pane for only file type connectors you can use.
i. Click on **Database** Category and see what databases you can connect to.

4. Or you can Click on the **Arrow** down the **Get Data** Button to get a list of the most common data sources.

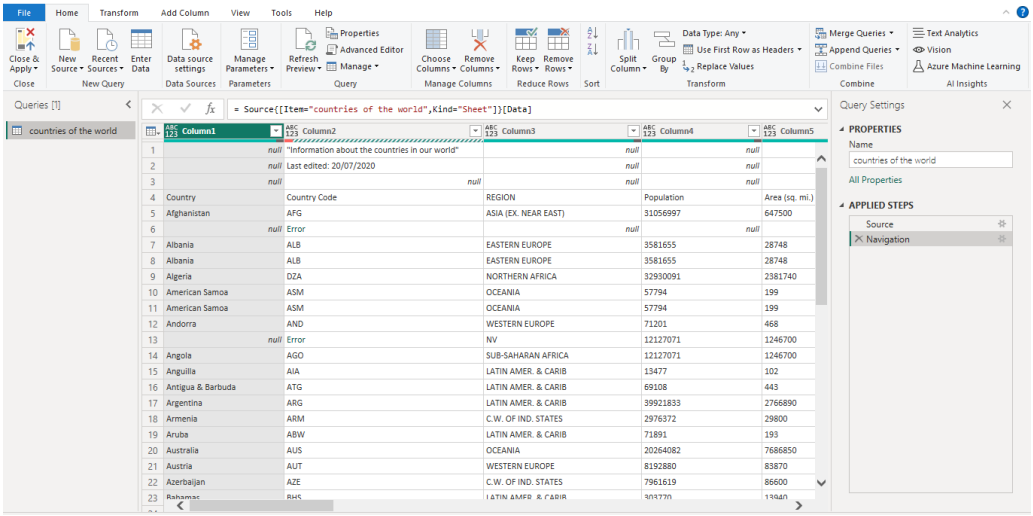


5. If you click on the more option at the bottom of the list, you get the complete connectors list again.
6. We want to connect to an excel file.
7. Notice it has a separate icon you can use directly from the Home tab because it is a very common data source.
8. Explore the data file **Countries.xlsx** file in your Files Folder and open in Excel.
9. Use Excel data source connector and connect to **Countries.xlsx** file in your Files Folder.

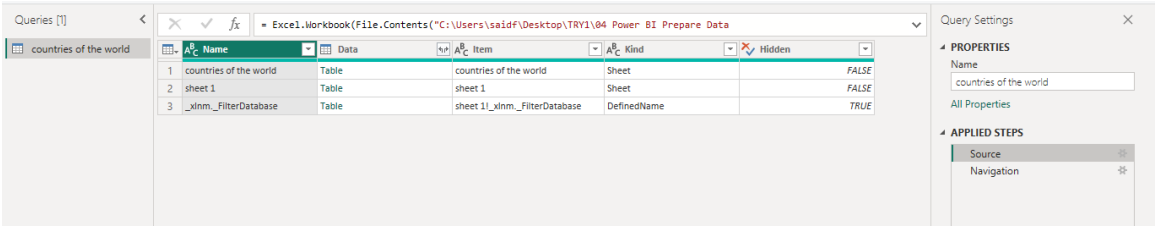
10. In the **Navigation Pane** that appears, on the left you have all data tables you have in the source, when you click one you can preview on the right.



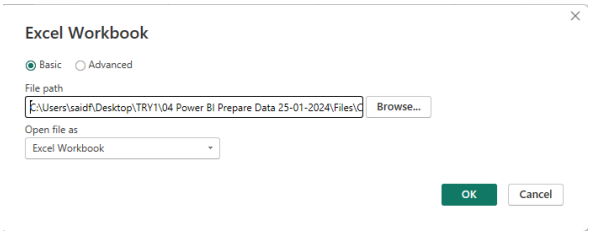
11. Select check box next to **Countries of the world** worksheet.
12. Notice there are many problems with that data in the right.
13. Click **Transform Data** button to go to **Power Query** editor.



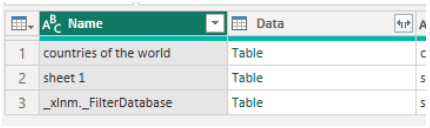
14. On the **Queries** pane you have only one query.
15. On the **Applied steps** there are only two steps, **source** and **navigation**.
16. Select **source** step and you will see the step in **M language** on formula bar.



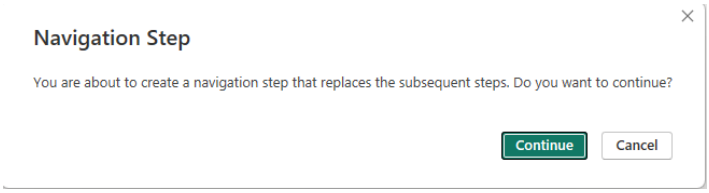
17. You can edit the step in the **formula bar** and change the source file path.
18. Alternatively, you can click the **Gear icon** next to the step and change the path graphically.



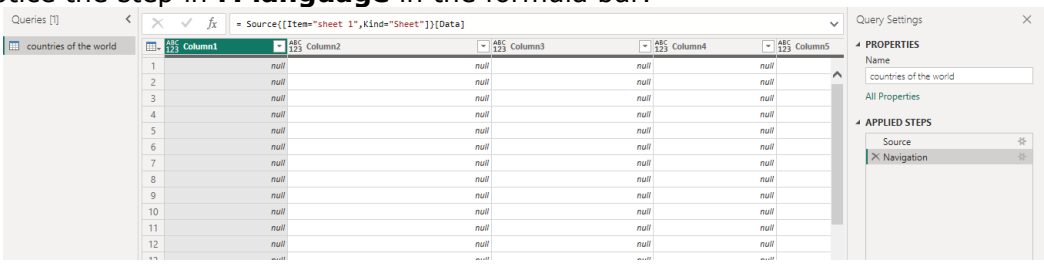
19. Click Cancel.
20. You can change the path and the type from the drop-down box.
21. While you are in source step, the next step is that we navigate to specific sheet in the Excel workbook (which is **Countries of the world**).
22. You can change the navigation to other sheet.
23. Click on the **table** link in **Data** Column next to **sheet1.**



24. You will get a warning message that you are changing the next navigation step.



25. Click continue to confirm.
26. You jump into navigation step with sheet1 data which is an empty one.
27. Notice the step in **M language** in the formula bar.



28. To correct that, click on the Gear icon next to the Navigation step and change the navigation to Countries of the world sheet and press OK.

29. Can you see how applied steps are powerful to modify your transformation any time?

Knowledge Check

Question 1

What is the difference between a dataset and a data source in Power BI?

A. A dataset is the same as a data source, but with fewer features.
B. A dataset is a container that holds some of the data from a data source, whereas a data source is where the data actually comes from.
C. A dataset and data source are the same thing.

Question 2

What is the maximum size of an Excel workbook that can be uploaded to Power BI?

A. 1 GB
B. 10 GB
C. 100 GB

Question 3

What are the types of workbooks that Power BI supports? Select all that apply:

A. Workbooks with ranges or tables of data.
B. Workbooks with connections to external data sources.
C. Workbooks with shapes and images
D. Workbooks with data models.

Question 4

What **is** the primary purpose of the Transform step in the ETL process?

A. To clean, structure, and enrich the data to make it more suitable for analysis.
B. To load the transformed data into the final storage system.
C. To extract data from multiple sources.
D. To analyze and visualize the data.

Question 5

What does source data refer to? Select all that apply.

A. Data that has been analyzed and refined for specific purposes.
B. Pre-processed data used for analysis and decision-making.
C. Raw, unprocessed information collected, stored, and managed by an organization.
D. The initial input used as the basis for further processing, transformation, and analysis.

Chapter 4: Transforming Data

Why data needs to be transformed

- Data transformation can involve different activities, such as cleaning, merging, and profiling data.

components of data transformation

- Before you can start working with that data, you need to clean and transform the raw data to ensure its accuracy and consistency.
- You learned that data may come from different sources. However, the data from these sources may contain inconsistencies that make accurate analysis difficult.
- Data from different sources can be untidy, incomplete, and inconsistent, making it difficult to draw meaningful insights.

- That's why data transformation is a crucial step. It helps you prepare data for analysis.

Some of the inconsistencies you may find in data

- let's say you are working on an analysis related to products in an e-commerce database. For this task, you need some relevant fields for your report.
- However, the table has hundreds of fields, making designing the report difficult. A data transformation would be when you **include certain columns** from the data and **exclude others** before loading for analysis and reporting.
- Another transform example would be selecting fields and transforming by **merging** them, such as in the customer table with fields for the first and last name. You want to display them as a single full name field by merging fields with a space between.

what data cleaning is?

- Data that is not structured is more flexible in terms of rules and therefore more likely to be disorganized and require cleaning.
- You may not encounter as clean data as you would expect in Excel data or in data organized using delimiter symbols such as angle brackets or commerce. In such cases, the data should have a preliminary examination to identify **incorrect** data or **separate rows** where content refers to the **same values**.
- like **ware house** how it's written as two words and **warehouse** has in one word.
- You can resolve these inconsistencies by passing them through filters with specific rules.
- This examination is referred to as **data cleaning**.
- Another data issue you may encounter is the need to **merge** or append multiple data sources.

Question

Data transformation is the process of preparing data for analysis.

A. True
B. False

Introduction to Power Query and its interface

- Power Query is part of Power BI Desktop, allowing for seamless data preparation within the Power BI environment.
- Power Query is a data transformation and data preparation tool allowing you to connect, clean, and transform data from a wide range of sources.
- It ensures that your data is ready for analysis, enabling you to create insightful visualizations and reports.

Exploring Power Query

- Features that Power Query can help with:
    - **Data connectivity**, Power Query connect to various data sources both on-premises and the Cloud directly within Power BI Desktop. You can access data from traditional databases as well as file-based sources.
    - **Data extraction and transformation:** Power Query interface allows you to extract and transform data with ease. During the extraction process, you can filter, sort, and apply custom transformations, ensuring that you import only the required data.

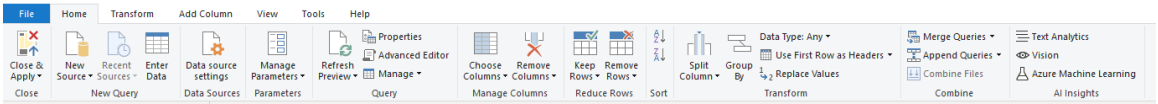
- **Power Query Editor**: in Power BI within Power BI Desktop, which provides a graphical user interface or GUI for designing and managing queries.
- Tabs such as **Home**, **Transform**, **Add Column**, and **View** have data manipulation tools.
- **Applied steps**: Power Query records each transformation as an applied step allowing you to review, modify, or delete any step. This ensures that your data transformations are transparent and easily modifiable.
- **Performance and scalability:** Power Query handles large datasets efficiently using various techniques that optimize performance and reduce memory usage.

Starting Power Query

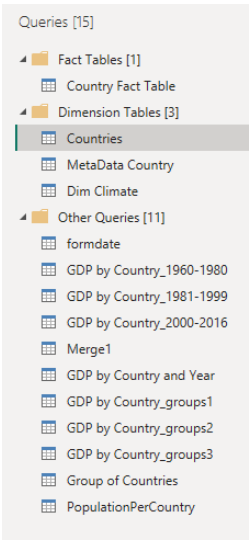
- To get started, you'll need to import your data into Power BI using Power Query. To begin the import, you must add a data source in the Power BI Desktop.
- In the **Home** tab, Select, "**Get Data**" to choose a data source.
- The power query editor opens in a separate Power BI window where you can apply various data transformations, such as removing columns, changing data types, and filtering data.
- Next, you need to **load** the data, select your data source and configure the connection settings if necessary.
- Select "**Transformed Data**" to open the Power Query Editor.

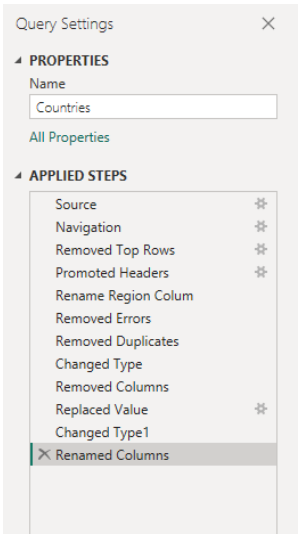
Navigating in Power Query.

- The Power Query Editor has several key areas:
  - **Ribbon**: The ribbon is the set of toolbars at the top of the window and helps you quickly find the commands that you need to complete your tasks.
  - The ribbon tabs, such as **home**, **transform**, **add column**, and **view**, contain commands and tools for data transformation and manipulation.

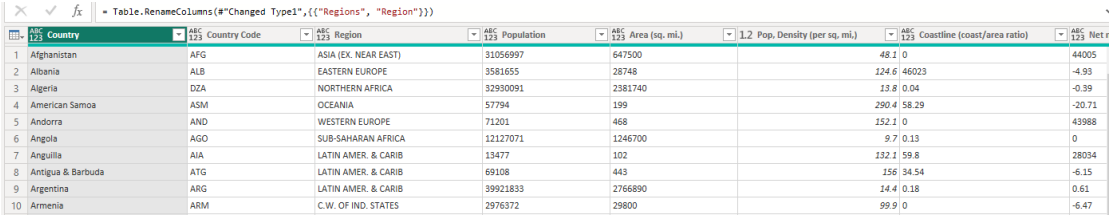


  - **The query's pane:** is located on the left side of the editor. The queries pane displays a list of all the queries in your project.
  - Select a query to view or edit its applied steps and data preview. This pane is where you can manage and navigate between different queries in your project.
  - By selecting a query, you can view the data and the applied steps associated with it, helping you keep track of your work and maintain organization in your project.





-
- **Applied steps section**: on the right pane below the ribbon.
- It displays the sequence of transformations applied to the selected query.
- Select a step to view the data state at that point or delete, re-order or modify steps as needed.
- The applied steps section provides a visual representation of the transformations applied to your data, making it easier to understand the changes made.
- By reviewing the applied steps, you can identify errors, redundancies, or inefficiencies in your data transformations.
- **Data preview**: in the center of the power query window.
- The data preview pane displays a preview of your data as it appears after the applied transformations.
- You can interact with the data by sorting, filtering, or changing the datatype of columns.
- This pane enables you to review your data at different stages of the transformation process, helping you to get your transformations accurate and effective before loading the data into the data model. Question



	Country	Country Code	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net r
1	Afghanistan	AFG	ASIA (EX. NEAR EAST)	31056997	647500	48.1	0	44005
2	Albania	ALB	EASTERN EUROPE	3581655	28748	124.6	46023	-4.93
3	Algeria	DZA	NORTHERN AFRICA	32930091	2381740	13.8	0.04	-0.39
4	American Samoa	ASM	OCEANIA	57794	199	290.4	58.29	-20.71
5	Andorra	AND	WESTERN EUROPE	71201	468	152.1	0	43988
6	Angola	AGO	SUB-SAHARAN AFRICA	12127071	1246700	9.7	0.13	0
7	Anguilla	AIA	LATIN AMER. & CARIB	13477	102	132.1	59.8	28034
8	Antigua & Barbuda	ATG	LATIN AMER. & CARIB	69108	443	156	34.54	-6.15
9	Argentina	ARG	LATIN AMER. & CARIB	39921833	2766890	14.4	0.18	0.61
10	Armenia	ARM	C.W. OF IND. STATES	2976372	29800	99.9	0	-6.47

-

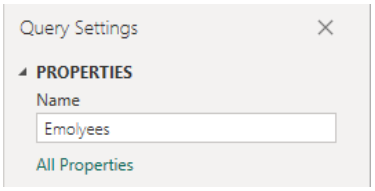
Question

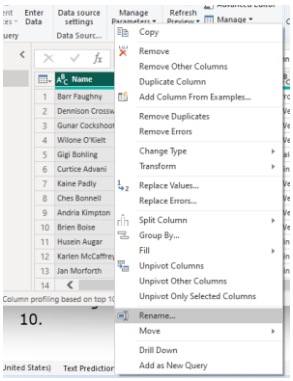
What is the primary purpose of Power Query in Power BI?

A. To automate the process of sharing reports and dashboards.
B. To create insightful visualizations and reports.
C. To facilitate seamless data preparation for analysis and visualization.
D. To predict future trends and patterns in the data.

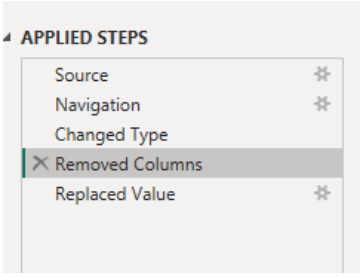
Exercise 4: Exploring Power Query

1. Open file: My first Power PI **Project.pbix** you have created in Exercise 1.
2. Remember that we have three views.
3. Go ahead and explore them.
4. Go to Report view.
5. Open Power Query: Home ➔Queries group➔Transform data.
6. On the top you see your Ribbon tabs.
7. On the left is Queries pane that shows your tables.
8. You can change the name of any query by double clicking or in the property pane in the right.
9. You can rename the columns either right click and choose Rename or double clicking the header.



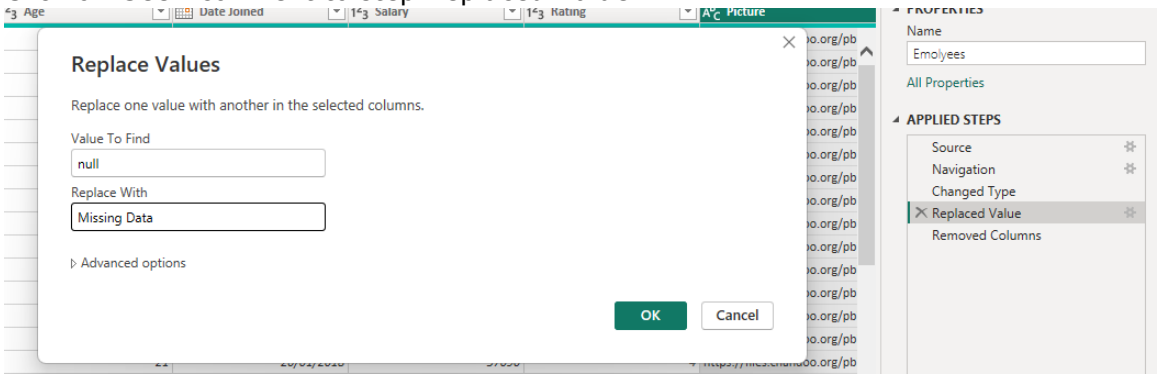


10. In the right you have the applied steps pane.
11. You can move to a step to see how was your data look at this point.



12. Click on Changed Type step so you can see your data before removing picture column.
13. You can see that column still there.
14. Click on Remove columns step the column is not here.
15. Click on Removed columns step and delete this step.
16. The column now appears at the final step.
17. Again, go and delete this column and the step now will be the last step.
18. Notice that some steps have a Geer icon beside.
19. That means you can modify this step.

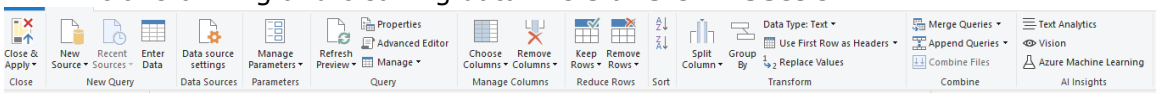
20. Click on Geer icon next to step Replaced Value.



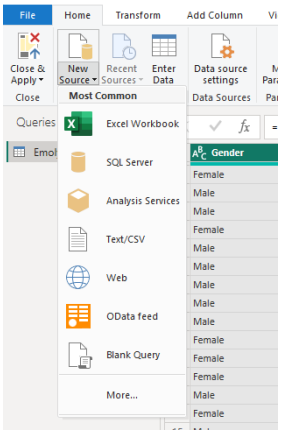
21. Change **Missing Data** into **No Value** and Click OK.
22. You can also rename the step into a meaningful name.
23. Right Click **Replaced Value** step and choose **Rename**.
24. Rename the step into **Replace null values in Gender**.

Exploring Ribbon Tabs

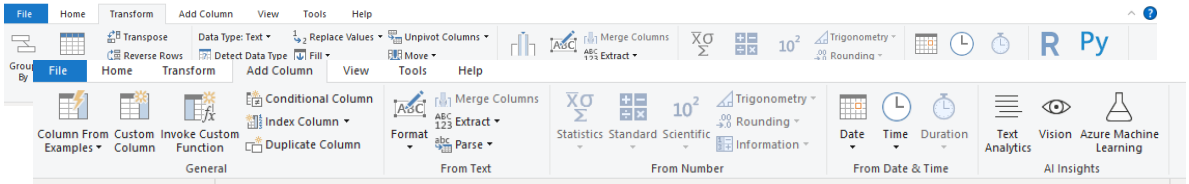
25. In **Home** tab You have many Groups of sections that mostly used when transforming and cleaning data like **transform section**



26. You can Get new data **Home ➔New Query** Group **➔New Source**.

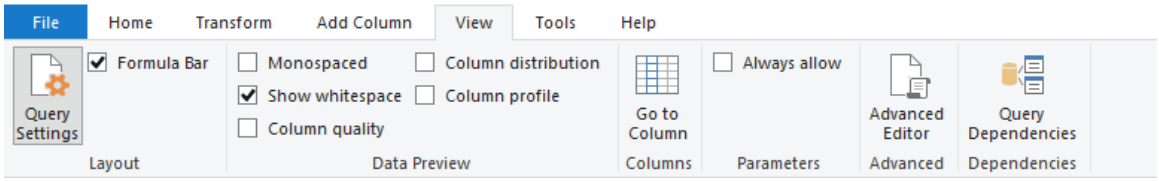


27. In the transform tab you have many options to transform your data.



28. The Add column tab allows you to add calculated new columns to help you in your analysis that are not originally in the source data.

29. The **View** tab can show and hide some section in your data view like for example profiling your data in **Data Preview** Group



30. We will cover those tabs through our training course.
31. Now click on **Close & Apply** to go back to your Power BI Report.
32. Notice that changes you have made is applied.
33. Notice in the Pie chart has the **No Value** instead of **Missing Data**.
34. Save and close your file.

The Applied Steps list

- Thanks to the Applied Steps List in Power Query, you can easily undo and reorder steps without losing progress.
- In the Power Query Editor, you'll find the Applied Steps List on the right pane below the ribbon. It has all the steps you've performed on your data presented in the order of application.
- The Applied Steps List is a visual representation of the transformations applied to your data.
- By reviewing the Applied Steps, you can identify errors, redundancies or inefficiencies in your data transformations.
- To view the data state at a specific point in the process, select the corresponding step in the Applied Steps List.
- The Applied Steps List makes it easy to correct a mistake or change your mind or undo a transformation.

Undo a step

- To undo a step, simply select the X icon next to the step to remove.
- Power Query will automatically revert the data to the state it was in before that step was applied.
- Please note that removing a step will also remove all subsequent steps in the list, as they are dependent on the previous transformations.

Reordering Steps

- To reorder steps, select and drag the step you'd like to move to a new position in the list.
- Power Query will update the data accordingly applying the transformations in the new sequence.
- You should note that reordering steps might affect the results of subsequent transformations.
- Review your data and the Applied Steps List to check everything.

Modifying a Step

- Suppose you need to modify a step, just select the Gear icon next to the step.
- This opens a settings window to edit the transformation parameters.
- When changed, select OK to apply the update.

- As with reordering steps, modifying a step might affect subsequent transformations. Always review your data and the Applied Steps List to ensure everything is as expected.

Adding a New Step

- To add a new step, use the Power Query Editor ribbon to choose a transformation such as filtering or sorting.
- When you perform a new data transformation, it's added to the Applied Steps List.

Adding Filters

- With the Power Query editor, you can also add filters.
- Filtering is the process of narrowing down your data set by displaying only the rows that meet specific criteria.
- It helps focus on a particular subset of data, remove unwanted data that may affect your analysis, or simplify your data set for better readability.

Steps to add a filter:

- In the Power Query Editor, select the column header for the column you want to filter.
- This highlights the entire column.
- With a column selected, select the small down arrow next to the column header. This opens a drop-down menu with filtering options such as text filters, number filters or date filters, depending on the data type in the column.
- Choose the type of filter and select OK.
- Notice the new filtering step has been added to the Applied Steps List.

Sorting Data

- You can also sort your data set.
- Sorting is the process of arranging your data in a specific order, either ascending or descending.
- Sorting organizes data based on specific attributes such as alphabetical order, numerical values, or chronological order, helping to identify the highest or lowest values in a data set.

Steps to filter Column

- Select the column header for the column you want to sort.
- In the home tab of the ribbon, find the **Sort group**. Choose sort ascending **A to Z** or sort descending **Z to A**. To sort the selected column in ascending or descending order. The data is sorted based on your chosen sorting order.
- Check the Applied Steps List to ensure the new sorting step is added.

Renaming an Applied Steps

- Finally, for better organization and readability, you can rename any step in the Applied Steps List.
- Just right click the step you'd like to rename and select **Rename**.
- Enter a new descriptive name for the step and press Enter.
- Renaming steps helps keep track of transformations, making it easier to navigate and understand the data transformation process.

Question

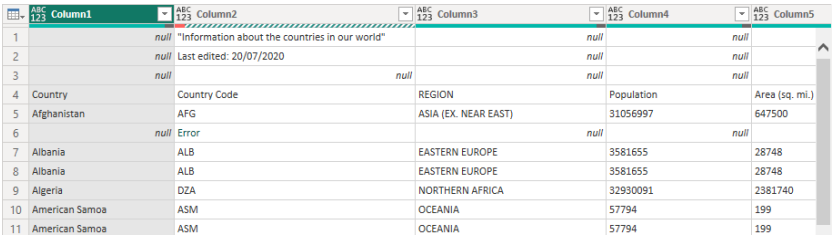
What is the purpose of the **Applied Steps** section in the Power Query Editor?

A. To provide a graphical user interface for designing and managing queries.
B. To preview the data after the applied transformations.
C. To show the sequence of transformations applied to the selected query.
D. To display a list of all the queries in your Power BI project.

Exercise 5: Editing Rows

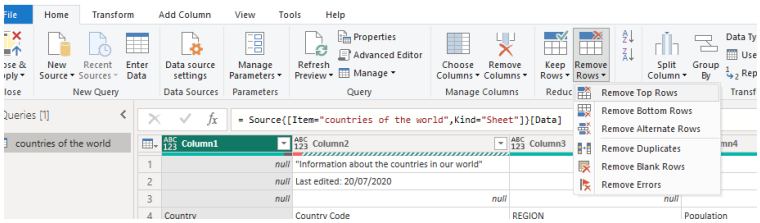
Remove Top Rows

1. Notice that we have rows from the Excel file that were a Heading and has no data in it.

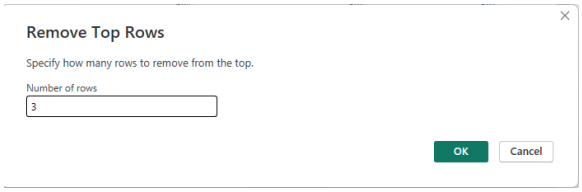


	Column1	Column2	Column3	Column4	Column5
1	null	"Information about the countries in our world"	null	null	
2	null	Last edited: 20/07/2020	null	null	
3	null	null	null	null	
4	Country	Country Code	REGION	Population	Area (sq. mi.)
5	Afghanistan	AFG	ASIA (EX. NEAR EAST)	31056997	647500
6	null	Error	null	null	
7	Albania	ALB	EASTERN EUROPE	3581655	28748
8	Albania	ALB	EASTERN EUROPE	3581655	28748
9	Algeria	DZA	NORTHERN AFRICA	32930091	2381740
10	American Samoa	ASM	OCEANIA	57794	199
11	American Samoa	ASM	OCEANIA	57794	199

2. We do not need those rows, let us get rid of them.
3. Home➔Reduce Rows group➔Remove Rows➔Remove top rows.



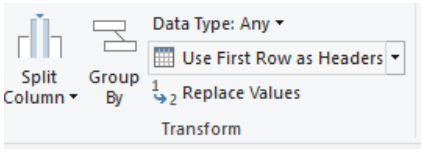
4. In the remove top rows dialogue box write 3 and press OK.



5. Your top three rows now are deleted, and a step **Removed top rows** is added to your applied steps.
6. And you can click on the Gear icon to change the number you have applied.
7. Try to go to Navigation step and get back to the Remove top rows step to see the change you have made in your data.

Change Column Header

8. Notice that the headers are column1, column2, ….
9. We want to use the top row we have reached to now as column names.
10. Home➔ Transform group ➔Use First Row as Headers.

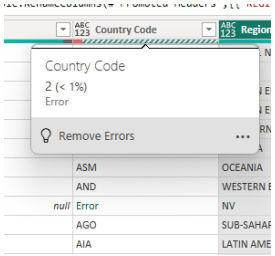


11. Now your first row is up to be your headers.
12. Double click on **the REGION** column and change its name to **Region**.
13. You see an Applied Step "Renamed Columns".
14. Go and Rename the step to be "Rename Region Column".
15. Notice here you do not have Gear Icon/

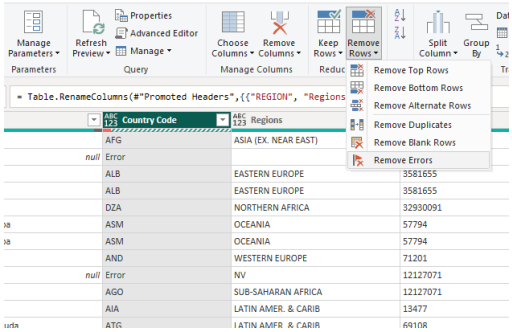
16. So, if you want to change the name of the column you have to do that in formula bar in M language.
17. Rename it Regions and press Enter.
18. You can Change the Query name in either double clicking in the Query pane on left, or in Name Property in Setting on the right.
19. Rename your Query name to **Countries**.

Remove Error and Empty Rows

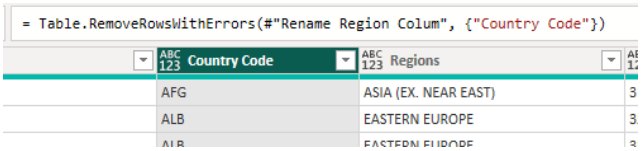
20. Notice in **Country Code** column we have errors, and we have **red** indicator in the column header when we hover on it says you have 2 errors in this column and that is represent less tan 1% of the data.



21. Notice that we have no data completely on those two rows.
22. So we decided to delete those rows.
23. First make sure you have selected the **Country Code** column only.
24. Then from Remove Rows➔Remove Errors.

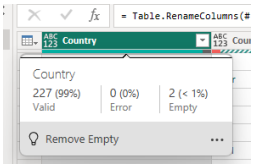


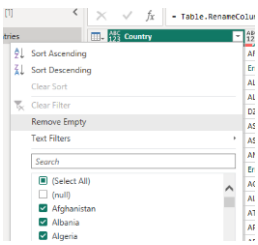
25. Notice an applied step is added, the M language code in formula bar, and that Country Code column has no red error indicator anymore and a green line under the column name indicates that there is no error.

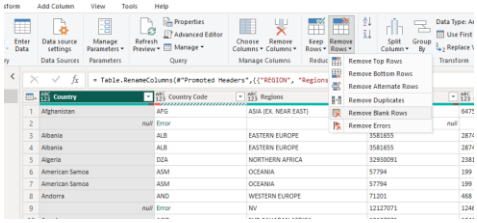


26. We can get the same result doing it in another way.

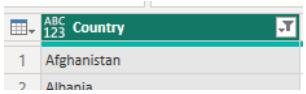
27. First remove the last step **Removed Errors**.
28. Notice that the two errors in the **Country Code** have null values in other columns.
29. Click the arrow in the Country Header it will show all values in the column including the null value.
30. You can either:
    - Unselect null value from the lest.
    - Select remove Empty from the to.







31. Remove the Null Rows and a **Filtered Rows** Step is added, and a **Funnel icon** appears on the column.



Remove the Duplicate Rows

32. Notice that we have some duplicate rows in the data like **Albania**.
33. Select Country column.
34. Home➔Remove Rows➔Remove Duplicates.
35. Just Hover over these options and see the message:

    "*Remove rows containing duplicated values in the current selected column.*"

36. Select Country Column and Apply the step.

Data types in Power BI

- Data types are defined at the column level. The values contained within a given column are configured to align with the designated data type of the column.
- Every data type has some specific transformations and options that can be applied. As mentioned before, the Transform and Add column tabs and the column filter options are primarily used for these transformations.
- You can use this reading as a cheat sheet of data types available in Power BI. Additionally, the linked pages in this reading provide additional valuable information about Power BI data types. By the end of this reading, you'll be able to identify data types in Power BI.
- When examining data types, the best approach is to categorize them into groups such as number types, date or time types, and other types such as text types. This way, you can easily identify the unique properties of each data type and recognize differences between types.

Data types Groups

Number types

- **Decimal number**: This data type can handle numbers with fractional values as well as whole numbers. The maximum precision (number of digits in a number after the

decimal point) that the decimal number type can represent is 15 digits. The decimal separator can be anywhere in the number. For example, 99, 99.50, and 99.20930 are all valid decimal numbers. One example could be the price of a watch, $99.99. In another example, you could use the 15 digits of the decimal number data type to store the first 15 digits of the mathematical constant pi, which is 3.141592653589793.

- **Fixed decimal number**: The decimal separator always has four digits to its right and allows for 19 digits of significance. This data type is useful in cases where rounding might introduce errors. For instance, 99.0000, 99.5000, and 99.2093 are all valid fixed decimal numbers.
- **Whole number**: This is an integer type that has no digits to the right of the decimal place. It has 19 digits of positive or negative whole numbers, such as -10, 0, and 103. Its range is between *-2^63+1 and 2^63-2.*

Date and Time types

- **Date/time**: Represents both a date and a time value. Dates between the years 1900 and 9999 are supported. This data type is useful for keeping date and time data together. For example, a spreadsheet with Purchase Date or Order Date columns.
- **Date**: This data type represents just a date with no time portion. This data type is useful when you need only the date element of your records, such as birth date or contract date.
- **Time**: This represents just time data with no date portion. This data type is useful when you need only the time part of your dataset, such as an activity start hour or end hour.
- **Date/time/timezone**: This represents a UTC Date/Time with a time-zone offset. UTC, or Coordinated Universal Time, is the primary time standard by which the world regulates clocks and time.
- **Duration**: This data type represents the length of time. This data type is useful when measuring or calculating the time difference, such as Activity Duration or Sleep Time.

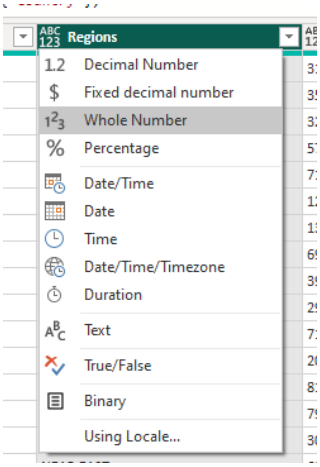
Other data types

- **Text**: This is a Unicode character data string. This can be strings, numbers, or dates represented in a text format. The maximum length of this data type can be 536,870,912 bytes. Or 268,435,456 Unicode characters. Unicode is an international character encoding standard that assigns a unique number to every character across languages and scripts.
- **True/false**: This is also known as Boolean data type, which has either a True or a False value.
- **Binary**: This represents any data with a binary format. For example, a non-human readable format is represented by ones and zeros. Binary files can contain diverse types of data. For instance, image or video files serve as binary files that are intended for computer systems to interpret.

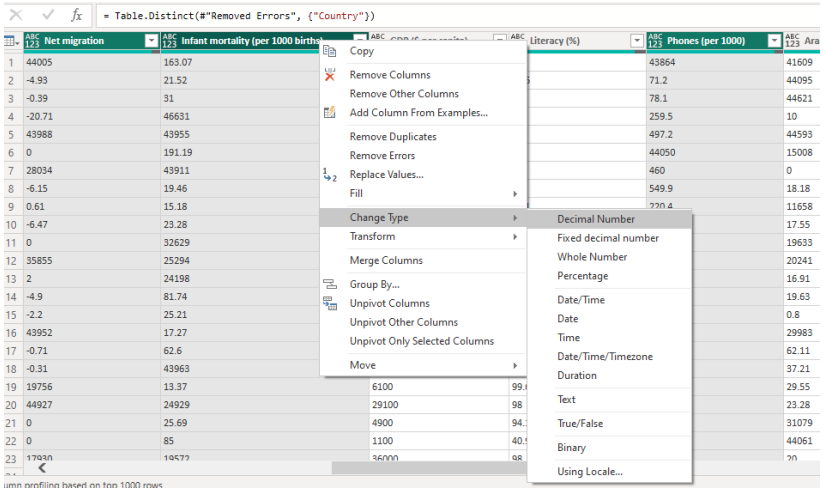
Exercise 6: Changing Data Types

1. Each Column its data must be saved in only one data type.
2. The data type symbol appears on the left of each column header.

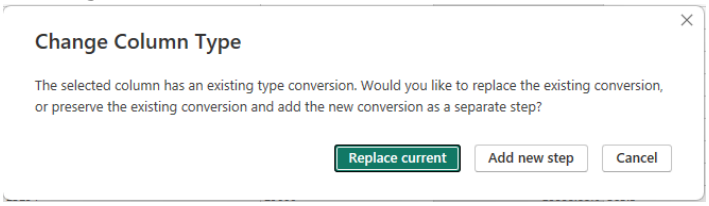
3. You can click to explore other data types, or to change the column data type.



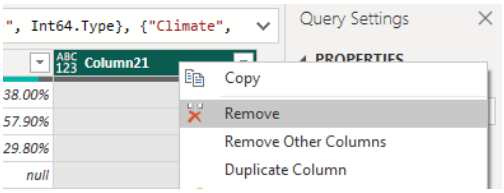
4. Power BI treats every data type differently in calculation and visualization.
5. So, it is important to choose the correct data type for each column.
6. Power BI usually **automatically detect** columns data types when loading data to power query editor.
7. But as you remember we have **disabled** this option in our settings in **Exercise 2**.
8. In real life you should enable this option, so it makes your life easier.
9. But we have disabled this option to do it manually.
10. Undetected Column has the symbol (ABC123)
11. Change the data types as follows:
12. Change **Area** to **Whole Number**.
13. Change **Coastline** to **decimal Number**.
14. Select columns (**Net Irrigation** , **Infant Mortality**, **Phones**) right click and chose change type to **decimal Number.**

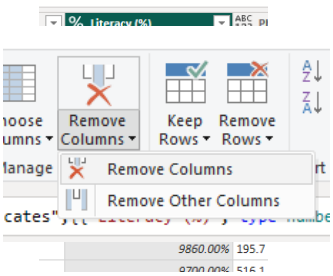


15. Change **Literacy** column type to **%.**
16. Notice that it gives us the wrong result as it multiplies the Number by **100**.
17. Try to change the column to **decimal**.



18. Power query asks you if you want to replace the step or add a new one.
19. Choose to replace current.
20. Select **GDP**, **Climate**, **Deathrate** and make them whole Number.
21. Select **Agriculture**, **Industry**, **Service** to percentage.
22. The last column **column21** has no data so remove it.
23. Right click and choose **Remove** or select Home ➔Mange Column



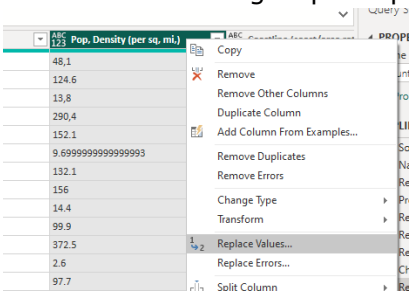


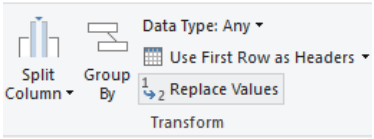
group➔**Remove columns**.

24. Select **Arable**, **Other**, **Birthrate** and change to **Decimal**.
25. Select **Country, Country Code, Regions** change to **Text**.
26. Change **Population** to **Whole Number**.
27. Notice that **Pop Density** column has a problem, sometimes the decimal is written as a (,) and sometimes as a (.).
28. Try to change to **decimal,** the (,) will be ignored and **48,1** became **481**.
29. Delete the applied step.
30. Power BI interpret (,) or (.) as a decimal separator depending on the settings of your local.
31. So, you must change (,) into (.) first before you change the type to decimal.

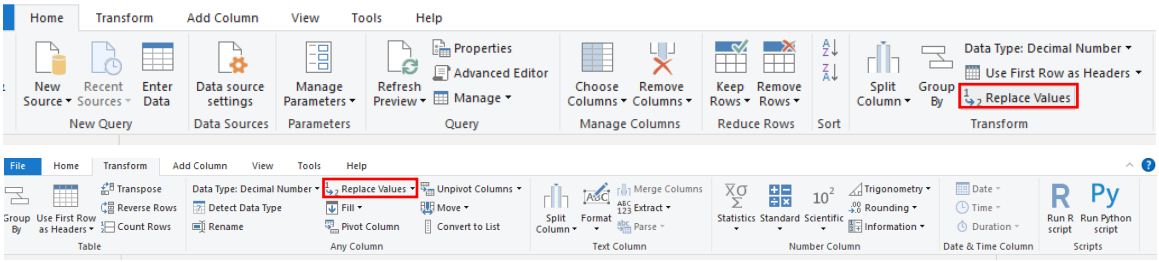
Replace Values

32. Select the **Pop Density** Column.
33. Either:
    i. right click and chose **Replace Values** or
    ii. Home➔Transform group➔Replace Values.



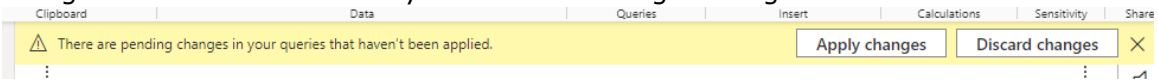


34. Replace Value of (,) to (.).
35. Now you can change the column type to **Decimal**.
36. Notice that **Replace Values** are in both **Home** and **Transform** tabs.

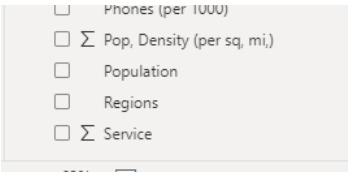


Close and Apply

37. We have made a lot of changes to our data.
38. But all that are in power query.
39. We must move our changes back to power BI Model.
40. Home ➔Close & Apply.
41. You get back to power BI and all your changes has been applied to your data model.
42. Get back to power query again and change the column **Regions** to **Region**.
43. This time click the arrow of Close & Apply and choose only **Close**.
44. You get back to Power BI but you have a warning message.



45. And, your change is not applied to the column.



46. Now you can click the **Apply Changes** button to get your changes applied.
47. Alson Notice that you can save your project in both Power Query Editor and in Power BI window by clicking the Desk icon.

If there are changes that did not apply to your project while you still in **Power Query**, you will be promoted to apply changes or only save your project without applying it.

Knowledge Chek

Question 1

Which of the following operations are steps in the data transformation process? Select all that apply.

A. Cleaning data
B. Creating insights from data
C. Shaping data
D. Removing data

Question 2

Which of the following data types are part of the number type group? Select all that apply.

A. Binary

B. Whole number
C. Fixed decimal number
D. Text

Question 3
Which one of the following features are used to track, re-order or delete the steps completed in Power Query?

A. Properties
B. New Source
C. Applied Steps
D. Queries

Question 4
Which of the following options can be used for Power Query Optimization? Select all that apply.

A. Choose the right data types for columns.
B. Filter rows in the queries.
C. Choose only the columns that you will use in the data model.

Chapter 5: Working With Columns

Benefits of working with Columns

- A common data manipulation you'll encounter is working with columns. Working with columns in Power Query in Power BI is an essential skill for data analysts and professionals who regularly deal with data.
- One of the main benefits of learning to work with columns is efficient data preparation.
- **Eliminating** unimportant or repetitive columns allows you to concentrate on the most crucial data for your analysis. Minimizing the data set size and streamlining the data structure for easier manipulation and quicker processing.
- Another benefit of working with columns is **improved data readability** and interpretation. Removing unnecessary columns helps declutter your dataset, making it easier to read and understand.
- **Renaming** columns with more descriptive names helps you quickly identify the purpose and content of each column.
- One other benefit of working with columns is that it **allows for enhanced data analysis and reporting**. By focusing on the most relevant columns, you can produce more accurate and meaningful analysis. This allows you to deliver actionable insights to your team and organization, leading to better decision making.
- Finally, working with columns means **time and resource savings**. Efficiently removing and renaming columns in Power Query can save you a significant amount of time during the data preparation stage. This means you can devote more time to analyzing the data and generating insights.
- By streamlining your data preparation process, you also **reduce the computational resources required** to process your data. This can lead to **faster analysis** and in some cases, **cost savings**, particularly when working with cloud-based services that charge based on resource usage.

Remove columns:

- In the Power Query Editor locate the column you want to remove. To select a single column, select its header. If you need to select multiple columns, hold down the keyboard control key and click on multiple column headers to remove.
- With the columns you want selected, you're ready to proceed. Right click any of the selected column headers. In the context menu that appears, select **Remove Columns**. The selected columns are removed from your dataset.
- You will notice a new step, **Removed columns** appears in the **applied steps list** on the right pane reflecting the updated data state.

Rename columns

- In the Power Query Editor, select the header of the column to rename, right click the header of the selected column. In the context menu, select Rename. A text box appears. Type in a new column name. Press Enter to save the change.
- Again, you'll notice the new step in the applied steps list.

Promote header rows

- The first thing is to identify which row in your dataset contains the headers. In most cases, this is the first row. If your dataset has additional information or metadata above the headers, you may need to scroll down to find the appropriate row.
- Now you can promote the header row. Once you've identified the header row on the ribbon, use the **Home** tab to locate the **transform group**. Select **use first row as headers**. This promotes the first row to be used as column headers replacing the existing headers.
- Note, if the header row isn't the first row, you'll need to **remove any rows above** the header row before promoting it. To do this, **select the rows you want to remove** by selecting the row numbers on the left side of the editor. Then, on the ribbon in the Home tab, select **remove rows**.
- You will notice a new step removed rows in the applied steps list on the right pane reflecting the updated data state.

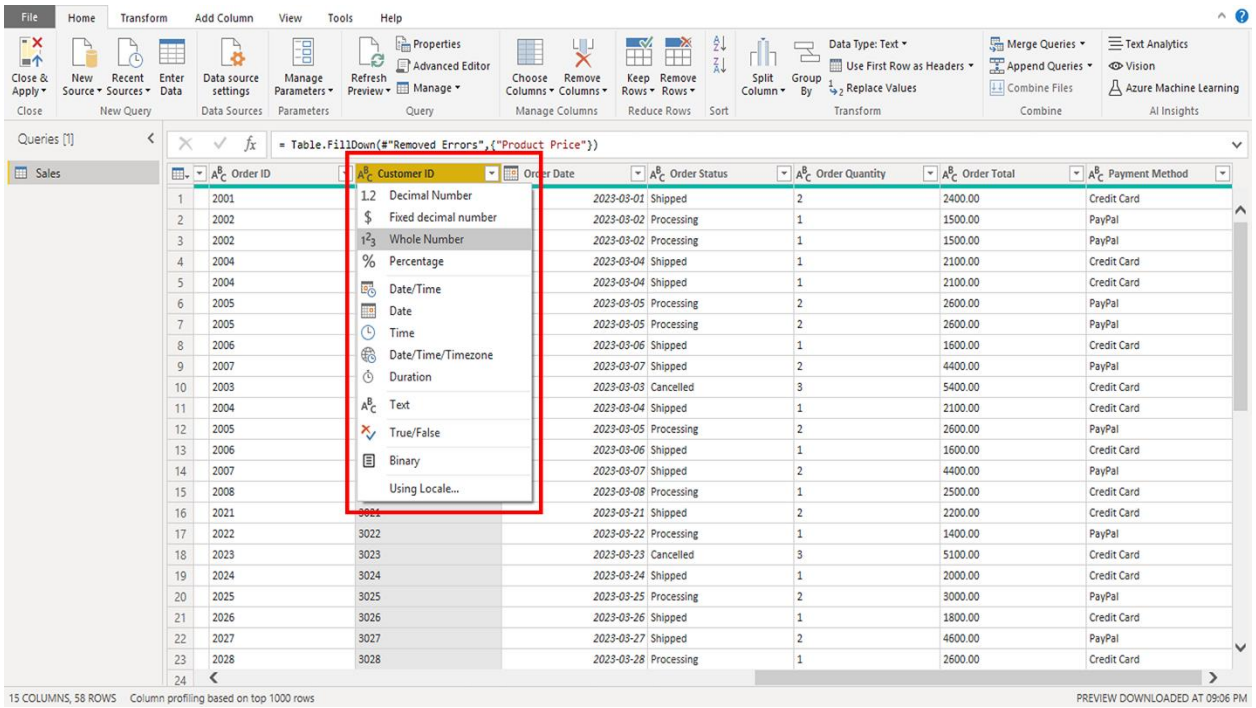
Question

How does removing unnecessary columns from a dataset benefit the data analysis process?

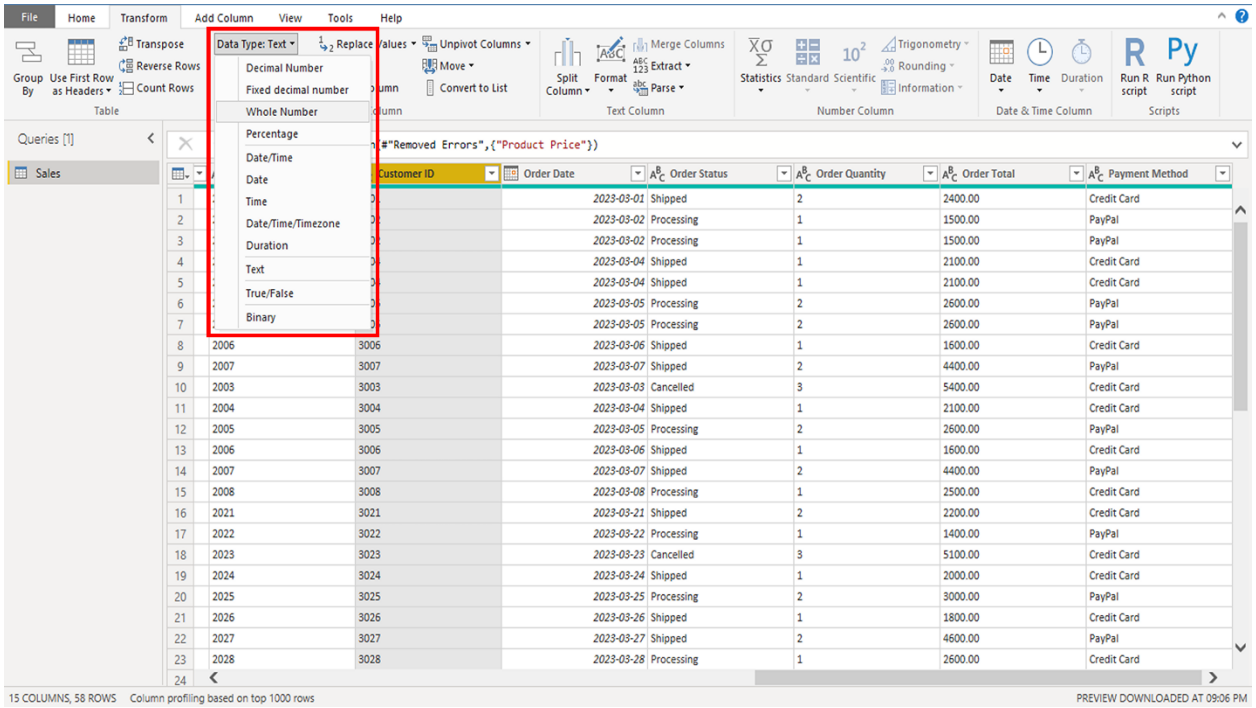
A. It reduces the dataset size, making it easier to manipulate and process.
B. It changes the structure of the dataset entirely.
C. It creates new columns with more relevant data.
D. It makes the dataset look more visually appealing.

Change a data type

1. On the left side of the column header, select the **data type** icon and then select the correct data type from the drop-down list.



2. Alternatively, in the **Transform** tab, select **Data Type** and then select the correct data type from the list.

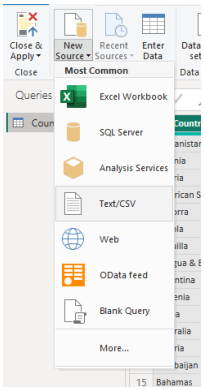


2. When you save this change, this step is called **Changed Type** and is reiterated every time the data is refreshed.

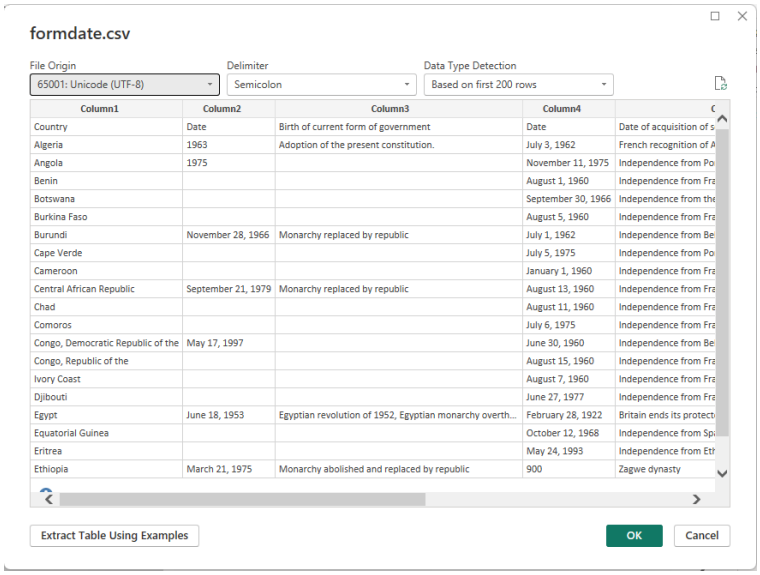
Exercise 7: Connecting to CSV File

1. Now we want to connect to a new data source.

2. This time it is a csv file.
3. Explore the file: **formdate.scv** in **Excel** and then **Notepad**.
4. Open Power Query.
5. You can connect to files from Power Query too.
6. Click on **New Source** from **Home** tab and choose **Text/CSV**.



7. Browse to **formdate.xls** file in your **Files** folder.
8. Notice that you do not have only **OK** button that is because you are already in Power Query.



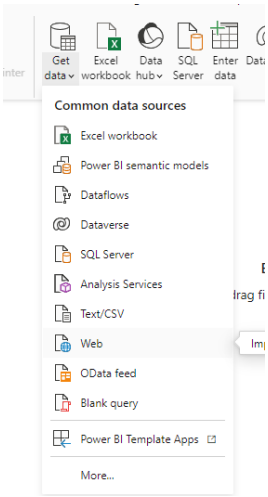
Column1	Column2	Column3	Column4	
Country	Date	Birth of current form of government	Date	Date of acquisition of s
Algeria	1963	Adoption of the present constitution.	July 3, 1962	French recognition of A
Angola	1975		November 11, 1975	Independence from Po
Benin			August 1, 1960	Independence from Fra
Botswana			September 30, 1966	Independence from the
Burkina Faso			August 5, 1960	Independence from Fra
Burundi	November 28, 1966	Monarchy replaced by republic	July 1, 1962	Independence from Be
Cape Verde			July 5, 1975	Independence from Po
Cameroon			January 1, 1960	Independence from Fra
Central African Republic	September 21, 1979	Monarchy replaced by republic	August 13, 1960	Independence from Fra
Chad			August 11, 1960	Independence from Fra
Comoros			July 6, 1975	Independence from Fra
Congo, Democratic Republic of the	May 17, 1997		June 30, 1960	Independence from Be
Congo, Republic of the			August 15, 1960	Independence from Fra
Ivory Coast			August 7, 1960	Independence from Fra
Djibouti			June 27, 1977	Independence from Fra
Egypt	June 18, 1953	Egyptian revolution of 1952, Egyptian monarchy overth...	February 28, 1922	Britain ends its protect
Equatorial Guinea			October 12, 1968	Independence from Sp
Eritrea			May 24, 1993	Independence from Eth
Ethiopia	March 21, 1975	Monarchy abolished and replaced by republic	900	Zagwe dynasty

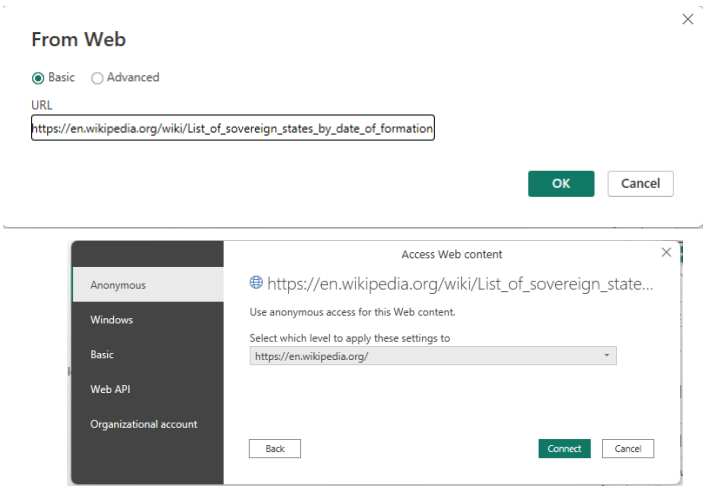
9. You have many options here.
    a) **File Origin**: is the type of language in the file.
    b) **Delimiter**: it is how to separate column in each row or the file.
    c) **Data Type detection**: it is how the software detects the type of each row.
10. The software has detected that the separator is **semicolon**.
11. Try to change the separator to **comma➔** the detection is not correct.
12. Explore other separators that can be used.
13. Get the separator again to be **semicolon**.
14. Explore Data Type detection it is the first **200 rows** by default.
15. Click OK to load the file to Power Query.
16. Explore the table you have loaded.

17. The table gives information about every country, when the current government was borne and what its independent day.
18. Close & Apply and Save your file.

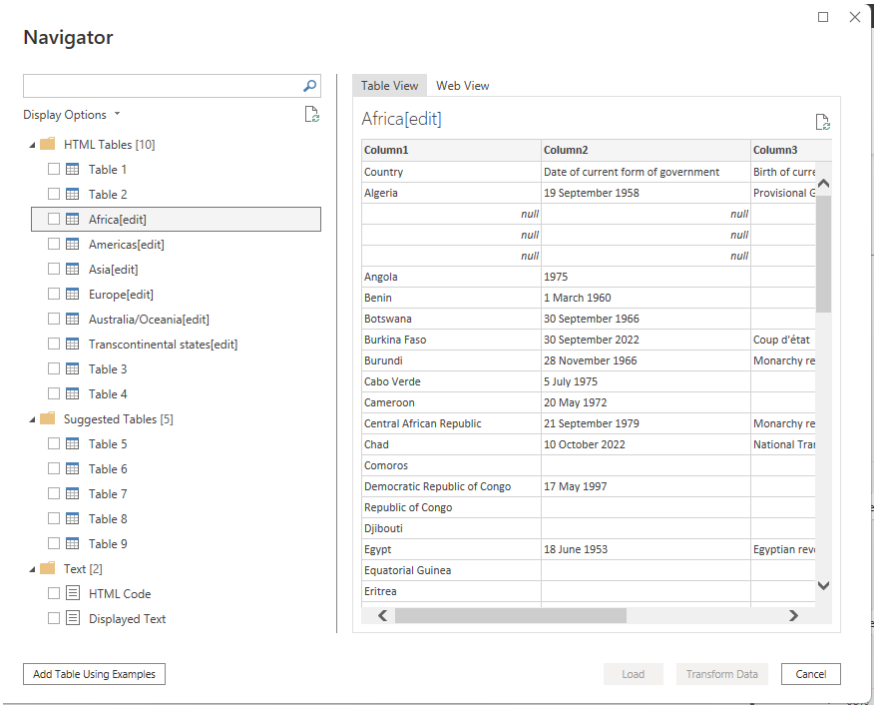
Exercise 8: Connect to a Web Page

1. In Google Search Type: "list of sovereign states by date of formation".
2. The first result will be the on Wikipedia:
3. https://en.wikipedia.org/wiki/List_of_sovereign_states_by_date_of_formation
4. Explore the table that in the web page.
5. We want to connect to this table in this page.
6. **Copy** the URL of the page.
7. Go to Power BI.
8. Click Get Data ➔Common Data Source➔Web.
9. In the From Web Dialog Box Paste the URL of the Page.
10. Click OK.
11. You go to another dialogue box click to connect Anonymous.
12. Click Coonect.





13. You will go to the **Navigator** again in the left you see all tables that Power BI can detect in the webpage.



14. You can explore each table till you find the one you want.
15. You can shift between **table view** and **Web View** on the right to see the original webpage.
16. On the left you will find that Power BI has **HTML tables**, **Suggested Tables**.
17. That is to make it easier for you to choose.
18. Select **Africa** table and click **Transform**.
19. Because the Webpage is always changing, we will use our CSV File through our exercises.
20. Right Click the new query and **delete**.
21. Now it is time to Transform the new data we have loaded into Power Query.
22. From left pane chose formdate query.
23. Home ➔Transform group➔Use First Row as Headers.

24. Notice that in column **Date_1** there is mixed data type , sometimes they write a date sometimes only the year.
25. Try to change its type to Date ➔you get an error.
26. Remove the step.
27. And let us see what we should do for that in Next Exercise.
28. Close & Load and save your file.
29. Do not worry if you receive an error.

Common data errors

- Before you begin to transform data in Power BI, you must first make sure that your dataset is accurate and reliable. Otherwise, you risk producing data analysis results that are incorrect.
- There are several types of errors that commonly occur in data sets.

Scenario

You Company recently produced a large dataset containing data on customers and sales. The marketing department plans to use this dataset to generate insights into the business and to help the business grow.

	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Sales
1	TrailBlazer 1000	None	958	5	300	287400
2	TrailBlazer 2000	Low	53,4	10	$7	17525,97
3	SpeedMaster 1000	Low	918	10	300	269892
4	SpeedMaster 2000	Low	1774	10	125	215097,5
5	Explorer 1000	Low	866	250	$345	9976,32
6	Explorer 2000	Medium	null	fifty	15	7908,75
7	GravityMaster 1000	2	1679	260	350	552391
8	GravityMaster 2000	Medium	588	120	20	10936,8
9	Pathfinder 1000	Medium	1366	260	20	25134,4
10	Pathfinder 2000	Medium	973	10	20	2013-10-01
11	Voyager 1000	High	2072	260	15	27972
12	Voyager 2000	High	six hundred	5	15	8936,4
13	Adventurer 1000	High	2641	10	20	45953,4
14	Adventurer 2000	High	1727	5	7	10396,54
15	EnduroMaster 1000	High	663	120	125	70443,75
16	EnduroMaster 2000	None	2146	5	7	15022
17	FatTrail 1000	Low	703,75	3	12	17166,6
18	FatTrail 2000	1	1728	10	300	508032
19	CrossRider 1000	Low	1901	10	12	22127,64
20	CrossRider 2000	Low	349	250	350	117264
21	DuoExplorer 1000	Medium	2861	120	15	40769,25
22	DuoExplorer 2000	Medium	727	260	350	239183
23	E-Mountain 1000	Medium	3244,5	250	null	36208,62

However, one of the data analysts believes that there are errors in the data set. These are common errors Adventure Works must identify and remedy before analysis.

Common errors

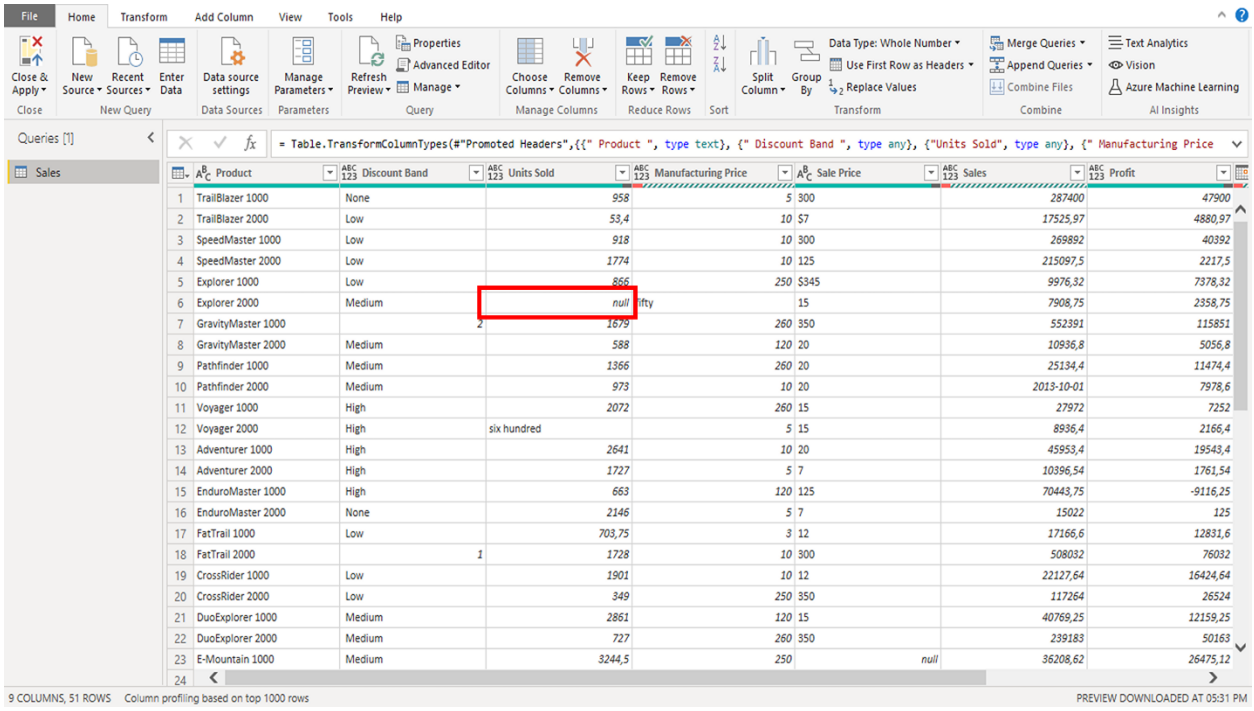
There are three main types of errors that you'll encounter as a data analyst. These are:

- Missing or null values
- Duplicate rows
- Inconsistent data types.

You must be able to identify instances of these errors in your datasets. If the errors are not identified, then their inclusion will lead to inaccurate, skewed, and inflated results. They can also give rise to extra, unnecessary storage and processing requirements.

Missing or null values

- A missing or null value occurs when data is absent or unavailable for certain cells or records within a dataset.
- For example, in the following datasheet, for the **Sales Price** column, the cell content on row **6** states **NULL**, indicating that there is no value in this location.

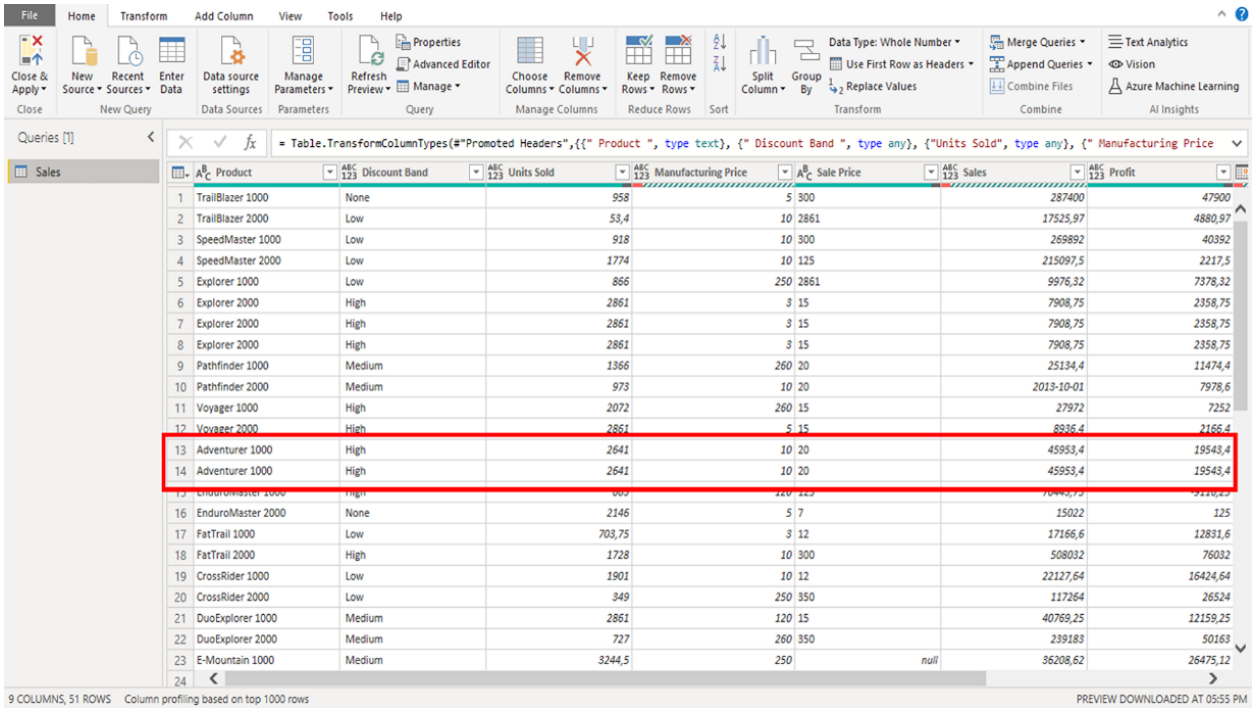


	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Sales	Profit
1	TrailBlazer 1000	None	958	5	300	287400	47900
2	TrailBlazer 2000	Low	53,4	10	$7	17525,97	4880,97
3	SpeedMaster 1000	Low	918	10	300	269892	40392
4	SpeedMaster 2000	Low	1774	10	125	215097,5	2217,5
5	Explorer 1000	Low	866	250	$345	9976,32	7378,32
6	Explorer 2000	Medium	null	fifty	15	7908,75	2358,75
7	GravityMaster 1000	2	1679	260	350	552391	115851
8	GravityMaster 2000	Medium	588	120	20	10936,8	5056,8
9	Pathfinder 1000	Medium	1366	260	20	25134,4	11474,4
10	Pathfinder 2000	Medium	973	10	20	2013-10-01	7978,6
11	Voyager 1000	High	2072	260	15	27972	7252
12	Voyager 2000	High	six hundred	5	15	8936,4	2166,4
13	Adventurer 1000	High	2641	10	20	45953,4	19543,4
14	Adventurer 2000	High	1727	5	7	10396,54	1761,54
15	EnduroMaster 1000	High	663	120	125	70443,75	-9116,25
16	EnduroMaster 2000	None	2146	5	7	15022	125
17	FatTrail 1000	Low	703,75	3	12	17166,6	12831,6
18	FatTrail 2000	1	1728	10	300	508032	76032
19	CrossRider 1000	Low	1901	10	12	22127,64	16424,64
20	CrossRider 2000	Low	349	250	350	117264	26524
21	DuoExplorer 1000	Medium	2861	120	15	40769,25	12159,25
22	DuoExplorer 2000	Medium	727	260	350	239183	50163
23	E-Mountain 1000	Medium	3244,5	250	null	36208,62	26475,12
24							

It's important to scan your dataset for missing or null values before you perform data analysis. The inclusion of these values can lead to incorrect calculations, skew statistical results, or generate misleading insights.

Duplicate rows

- Duplicate rows are instances in a dataset when two or more rows have identical values across all columns. This error often occurs because of data entry errors, glitches within the system, or data that's been merged from multiple sources.
- For example, the dataset contains identical records in rows **13** and **14**. Most likely, this occurred because the dataset was created by merging two different spreadsheets that contained an overlap of data. Both instances of this data have now merged into one spreadsheet leading to duplication.

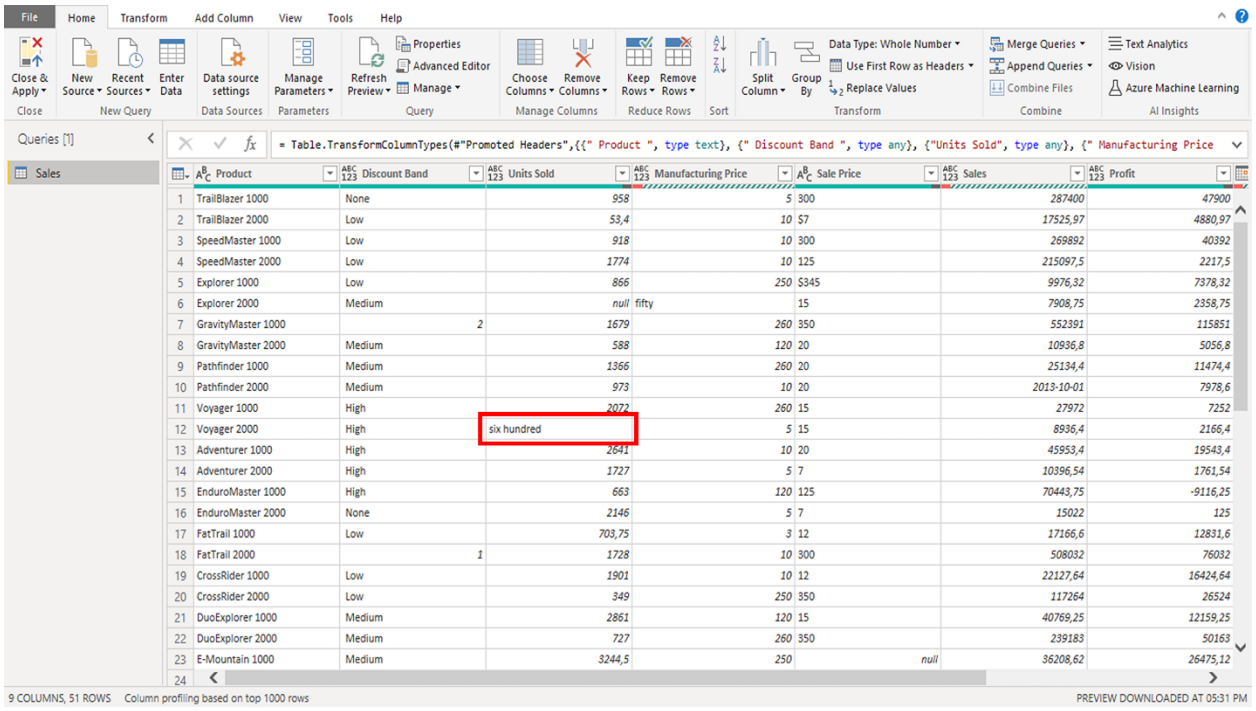


	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Sales	Profit
1	TrailBlazer 1000	None	958	5	300	287400	47900
2	TrailBlazer 2000	Low	53,4	10	2861	17525,97	4880,97
3	SpeedMaster 1000	Low	918	10	300	269892	40392
4	SpeedMaster 2000	Low	1774	10	125	215097,5	2217,5
5	Explorer 1000	Low	866	250	2861	9976,32	7378,32
6	Explorer 2000	High	2861	3	15	7908,75	2358,75
7	Explorer 2000	High	2861	3	15	7908,75	2358,75
8	Explorer 2000	High	2861	3	15	7908,75	2358,75
9	Pathfinder 1000	Medium	1366	260	20	25134,4	11474,4
10	Pathfinder 2000	Medium	973	10	20	2013-10-01	7978,6
11	Voyager 1000	High	2072	260	15	27972	7252
12	Voyager 2000	High	2861	5	15	8936,4	2166,4
13	Adventurer 1000	High	2641	10	20	45953,4	19543,4
14	Adventurer 1000	High	2641	10	20	45953,4	19543,4
16	EnduroMaster 2000	None	2146	5	7	15022	125
17	FatTrail 1000	Low	703,75	3	12	17166,6	12831,6
18	FatTrail 2000	High	1728	10	300	508032	76032
19	CrossRider 1000	Low	1901	10	12	22127,64	16424,64
20	CrossRider 2000	Low	349	250	350	117264	26524
21	DuoExplorer 1000	Medium	2861	120	15	40769,25	12159,25
22	DuoExplorer 2000	Medium	727	260	350	239183	50163
23	E-Mountain 1000	Medium	3244,5	250	null	36208,62	26475,12
24							

- You must make sure that you resolve all instances of data duplication before processing your dataset. If left unresolved, these errors can **inflate** the size of the dataset. This inflation could then skew your results.
- Such errors could also lead to unnecessary **storage** because your storage solutions need to host data that your projects don't require. Or they could give rise to extra processing overheads because your software needs to process large amounts of unnecessary data.

Inconsistent data types

- Inconsistent data types occur when values within a single column contain different types of data.
- For example, row **12** of the **Units Sold** column in the dataset contains inconsistent data types. The data types for cells of the **Units Sold** column should all be numeric. Instead, the column has a mix of numeric and text data types.



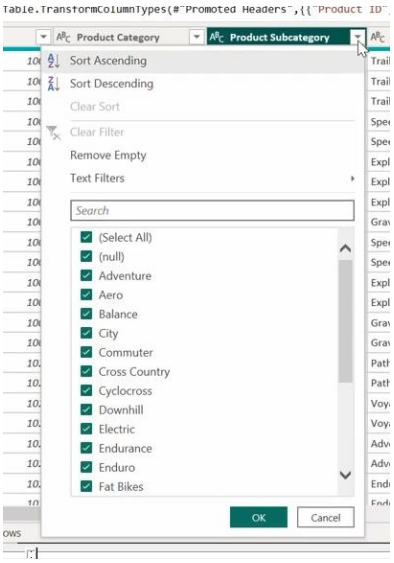
- It's important to identify and resolve any inconsistent data types within your dataset. If they remain in the dataset, they can cause calculations to misbehave, which can lead to errors in results.

Activity: Dealing with errors in Power Query

- When analyzing your data, you need to ensure accuracy and reliability, but datasets often contain errors that lead to inaccurate results.
- Using Power Query, you can fix many common dataset errors.

Example of Dealing with Errors

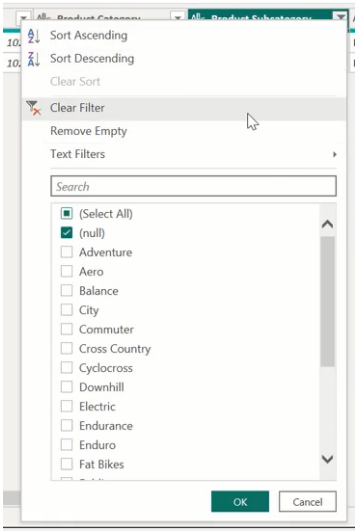
- You can use the **Sales.xlsx** file in your Files Folder to explore data.
- We have a list of bicycle products and key information about each product like name, price, weight, category, and description.
- However, several of these rows contain null or missing values, these errors need to be resolved before the data can be analyzed.
- To systematically identify missing or null values, select the drop-down arrow in the column header for the variable you're examining. This opens a filter menu used to filter the data in the column based on specific criteria.
- The filter menu contains options like empty or null. Available options depend on the data type of the column.



-
- **Empty** refers to **blank cells** in text columns, **null** refers to missing values in **numeric** or date columns.
- Select the appropriate option to filter and display rows that contain missing or null values in the selected column.
- Inspect the data table in the editor and identify any rows with missing or null values. In this dataset, two rows contain missing values, row 16 and row 17 have a missing value in the product subcategory column.

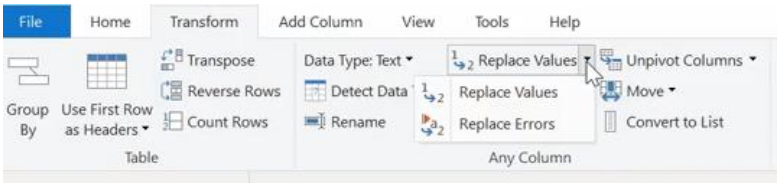
	Product ID	Product Category	Product Subcategory	Product Name	Product Description
1	1021	Mountain Bikes	null	Pathfinder 1000	Agile trail bike for all skill leve
2	1022	Mountain Bikes	null	Pathfinder 2000	High-performance trail bike

- Now that you've identified the values, you can resolve them.
- There are three ways to resolve missing values:
  - you can **replace them with default values**,
  - **replace them with values from another column**, or
  - **remove the rows containing missing values**.
- Clear your filter.

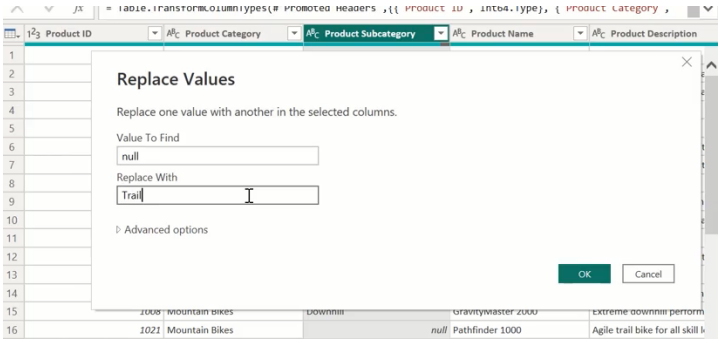


Replace Value

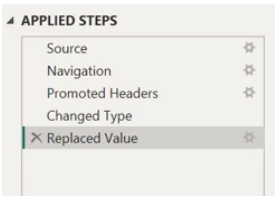
- For Our Example, the best approach is to replace its missing values with **default values**. Logically, the default values can represent the missing data without distorting the analysis or visualizations.
- First, in the ribbon at the top of the editor, select the **Transform** tab. You use this tab to access the tools and functions for modifying and transforming the data.
- Next, select the **Replace Values** button, then select Replace Values from the drop-down menu. You use this option to replace specific values in a column with a new value, in this case, you can replace all null or missing values.



-
- A **replace values dialog box** appears on screen, it has a text box labeled **Value** to Find where you specify the value you want Power Query to identify and replace.
- The aim is to find missing or null values in the product subcategory column. So, in the Value to Find box, you can write **null**.
- Below the Value to Find box, there's another text box labeled **Replace With**. This is where you type the new value you want to replace the missing or null values with. The new value should be consistent with the column's data type, which is text.
- So, let's replace the missing values in the product subcategory with the text value **trail**, which represents the default category for **trail bikes**.



- Finally, select OK to confirm and make the change.
- When you select the OK button in the Replace Values dialog box, Power Query scans the sheet for the values you've instructed it to identify. It then replaces each instance of these values based on the criteria you specified in the Replace With box.



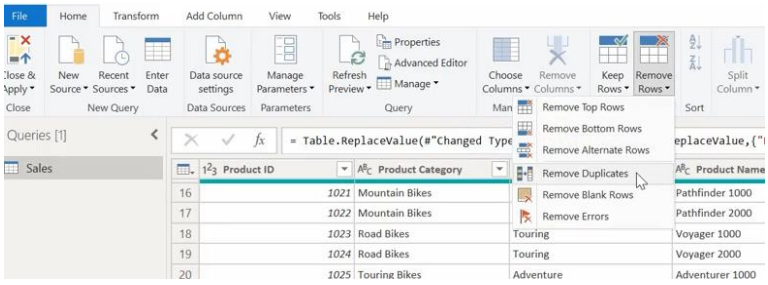
-
- You can review a history of all data transformation operations you've applied to the dataset in Applied Steps pane on the right-hand side of the Power Query editor window.

Remove Duplicates

- There are still duplicate rows errors present. The entries in rows 22 to 24 are duplicates of other records in the sheet, and identical records also exist in rows 25 to 27.

21	1026	Touring Bikes	Adventure	Adventurer 2000	Premium adventure tourir
22	1027	Mountain Bikes	Enduro	EnduroMaster 1000	Endurance-focused mount
23	1028	Mountain Bikes	Enduro	EnduroMaster 2000	High-performance enduro
24	1026	Touring Bikes	Adventure	Adventurer 2000	Premium adventure tourir
25	1027	Mountain Bikes	Enduro	EnduroMaster 1000	Endurance-focused mount
26	1028	Mountain Bikes	Enduro	EnduroMaster 2000	High-performance enduro
27	1041	Mountain Bikes	Fat Bikes	FatTrail 1000	All-terrain fat bike
28	1042	Mountain Bikes	Fat Bikes	FatTrail 2000	High-performance fat bike
29	1043	Road Bikes	Cyclocross	CrossRider 1000	Versatile cyclocross bike

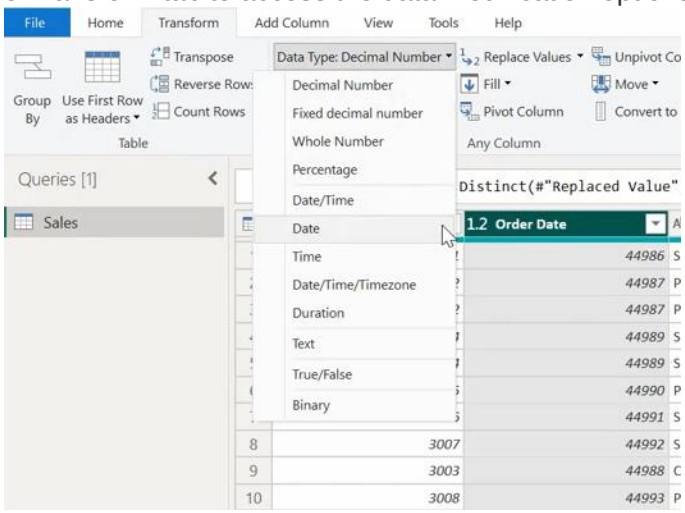
- On the Home tab, access the data manipulation functions. From these functions, select the **Remove Rows** option and a drop-down menu appears, select **Remove Duplicates** from the options.



- Power query analyzes the dataset and finds rows that have identical values in the selected columns, it then removes all but one instance of each group of duplicates.

Inconsistent data type

- That's good progress. Just one final error left in the dataset, inconsistent data types. in the form of order dates, let's fix this final error.
- The inconsistent data is in the column Order Date, select the column header to select and apply changes to the entire column.
- Next, select the Transform tab to access the data modification options. Select the



- Data Type button, then select the **Date** data type from the drop-down menu. This converts all values in the column to the selected data type, meaning all data types in the column are now consistent.

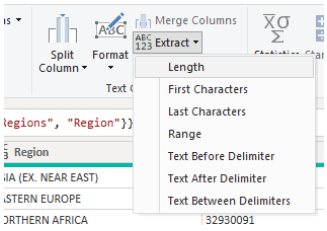
Question

Which of these issues in Power Query within Power BI is related to the presence of empty cells in your dataset?

A. Duplicate rows
B. Missing or null values
C. Inconsistent data types
D. Data entry errors

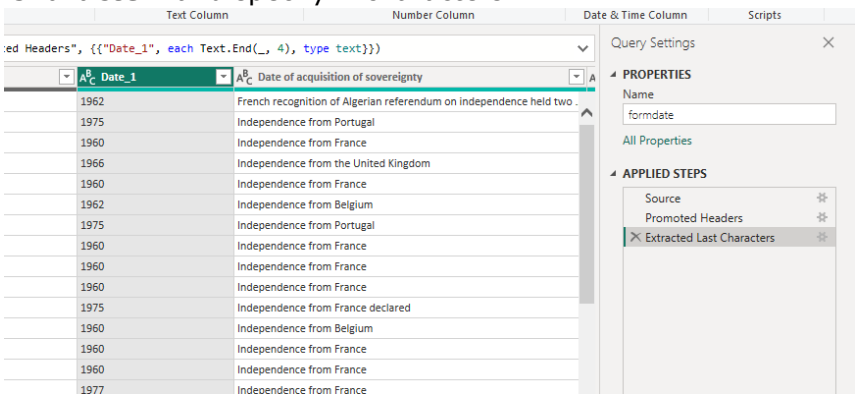
Exercise 9: Extracting Text

1. We want now to learn how to extract certain information from a column.
2. For example, we want to Extract only the year from the Date_1 column.
3. Notice that we have Extract Option in the Transform tab.
4. Transform➔Text Column ➔ Extract.
5. Click to see what options you have and try them.

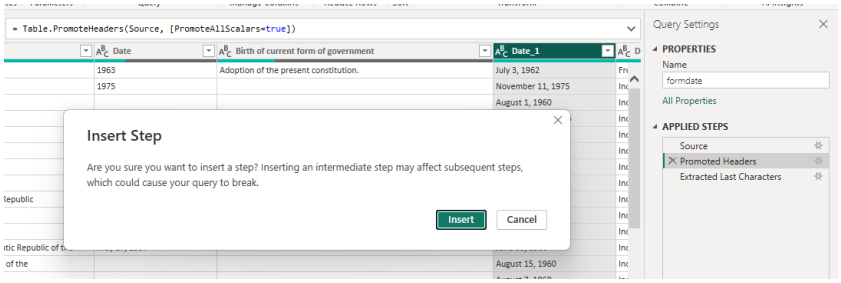


6. Select the **Date_1** column and try **Length** option.
7. This option gives you a Number representing the length of the text.

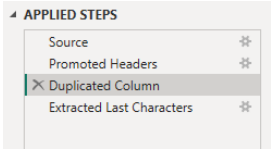
8. Delete Applied step.
9. We want the year, notice it comes always after comma (,).
10. Try the option "**Text After Delimiter**".
11. In the dialogue box write "**,**" as a delimiter.
12. Notice the result is Ok for some value.
13. But some values have no comma, so no value were extracted.
14. So, this is not a good option.
15. Delete the step.
16. Try option "**First Character**" and write **7** as number of characters.
17. It gives us nothing important.
18. Delete the step.
19. We think if we started from the end and extracted the last 4 characters will do and it will work in Case only year were written and even there are only 3 characters.
20. Try "**Last Character**" and specify **4** characters.



21. Wow it has worked this time.
22. But It would be better if we left the original column too.
23. Jump back a step (**Prompted Headers**).
24. You will see how the column was before extracting the year from.
25. Now right click the column Header and select Duplicate column.
26. You will get a warning that you are inserting a step.
27. Click Insert.



28. Now you have the column duplicated inserted before the Extracting the column.

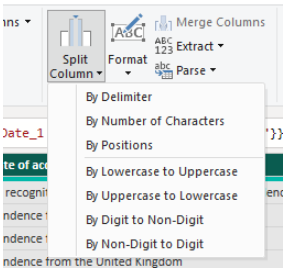


29. Go to the last step.

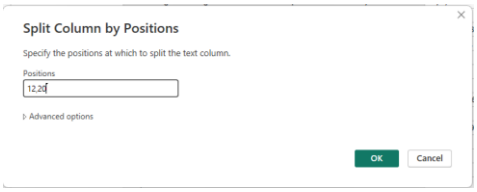
30. Rename the column that has the Year only "**Year of formation**".
31. Change its data type into **Whole Number.**
32. Rename the column that you have duplicated "**Date_1 Copy"** to "**Date of formation**".

Exercise 10: Split Column

1. We want to know how to split a column.
2. Look at the column "**date of acquisition of sovereignty**".
3. You can see it ends with the name of the country.
4. Let us try the option Transform➔Text Column group➔Split Column.
5. As you can see it is so difficult to extract a country from.
6. Select that column first.
7. Click on Split Column and see its options.

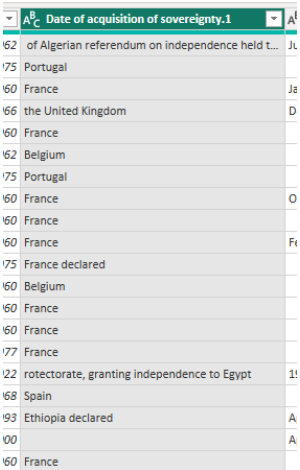


8. Try selecting **Positions** option.
9. Try 12,20 as two positions (start at letter 12 and stop at 20).

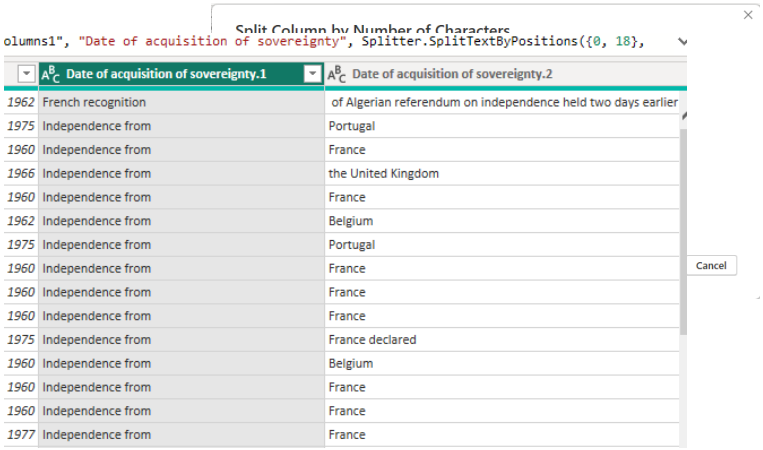


10. Go back step and go forth to see what happened.

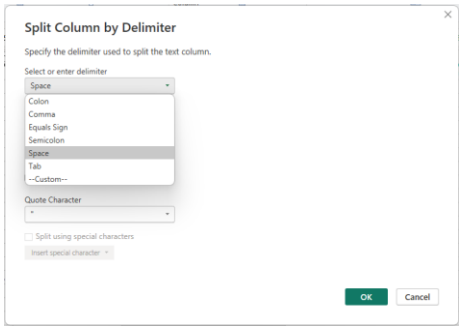
11. The column has split into two columns start at letter number 12 and stopped at number 20 in first column and the rest in the second column.
12. Click the Gear icon on the step and change 12,20 to 18.



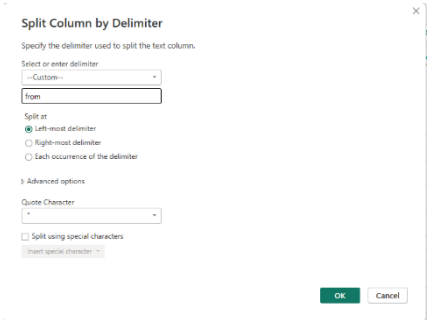
13. The result is better it started at 18 and till the end in one column.
14. But not all countries came.
15. Delete the last step and let us try something else.
16. Try "**Number of Characters**" options.
17. Try Number 18 and **once as far as left as possible**.

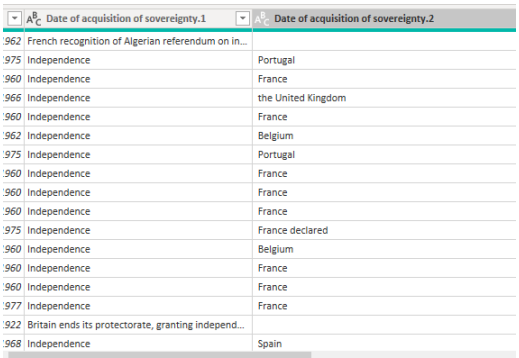


18. We got the similar result of two columns.
19. Only 18 Characters in 1st column and the rest in 2nd column.
20. That is not what we want.
21. Delete this step and try "**By Delimiter**" Option.



22. You can chose where starting splitting(space, comma, ….).
23. But there is also Custom.
24. If you look at the data you can find that the country always comes after the word "**from**".
25. Let us try this as a custom delimiter.

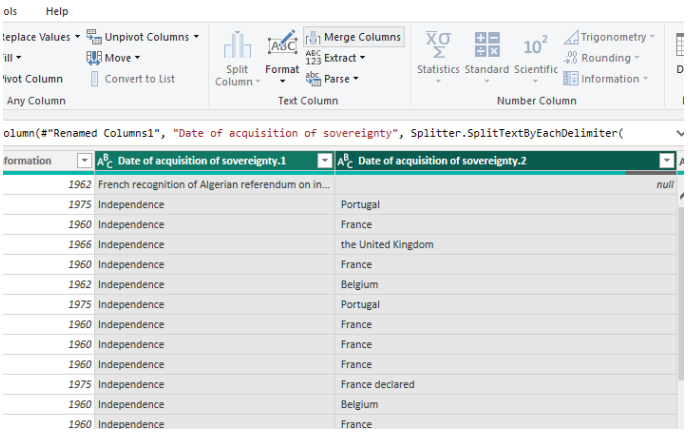


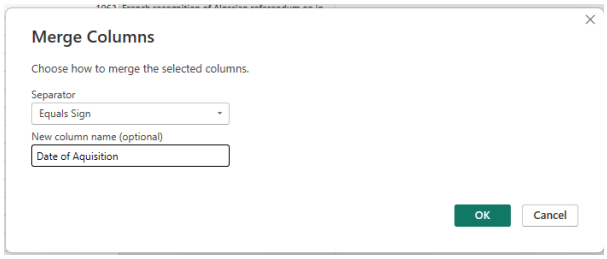


26. You will have a better result now.

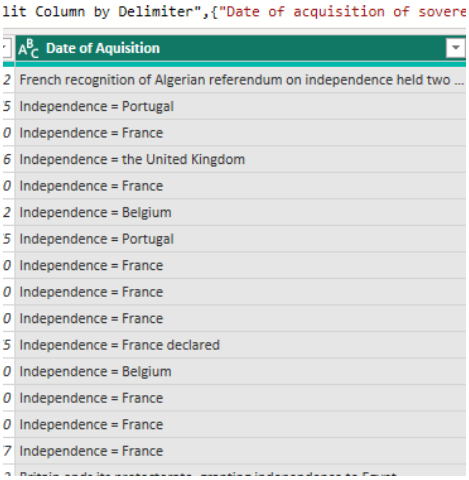
Merge columns

27. If you want to do the opposite and merge the result to get the original one.
28. First select the two columns.
29. Transform➔Text Column group➔Merge column.
30. Then select a separator (=) and give a name to the merged column (Date of Acquisition).

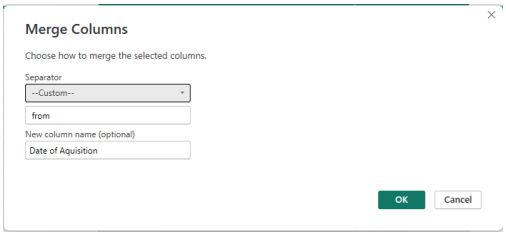


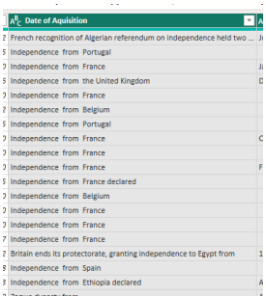


31. You will get the result of one column.



32. If you click the Gear icon of the step and change the separator to custom and make it "**from**" ,you will get the result of the original column.

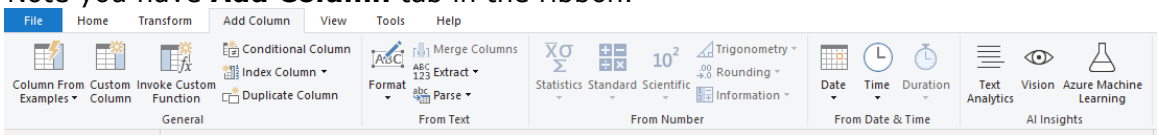




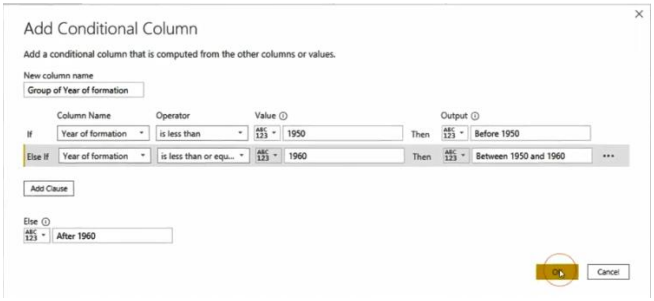
Exercise 11: Creating A Conditional Column

1. Let us now examine how to add new columns that are based on information in existing columns.

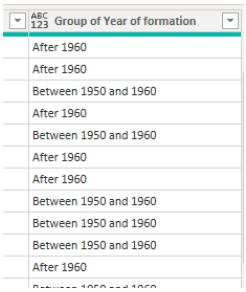
2. Note you have **Add Column** tab in the ribbon.



3. Let us start with Add Column➔General group➔Conditional Column.
4. We have created a column "**Year of formation**" containing the year.
5. We want to make groups of years (before 1950 or before 1960 or after 1960.
6. Click Conditional column and add the condition like the following:

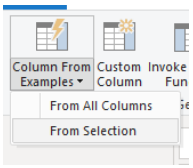


7. The result is grouping the years in 3 groups.

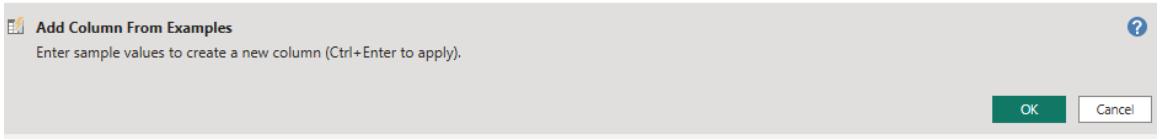


Exercise 12: Creating a Column from Examples

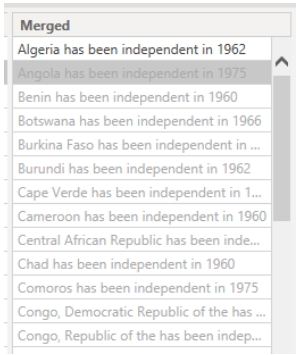
1. Let us create a column in an intelligent way.
2. Check **Create a column from Example** , you have two options:



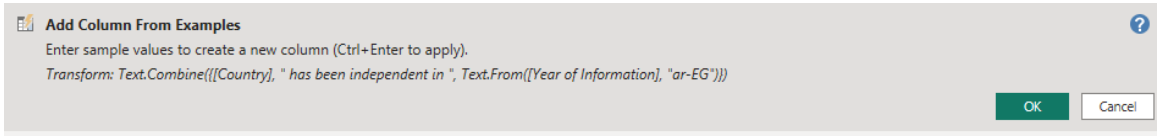
   a. From all columns.
   b. From Selection.
3. We want our new column to be based on a selection.
4. So first we select the columns.
5. Then click Column from Examples ➔From Selection.
6. For example, we want to have a column based on **Country** and **Year of formation**.
7. For example: Algeria has become independent in 1962.
8. Select Columns: **Country**, **Year of formation**.
9. Then click Column from Examples ➔From Selection.
10. You will have a new column **Column1** at the end and a message at the top



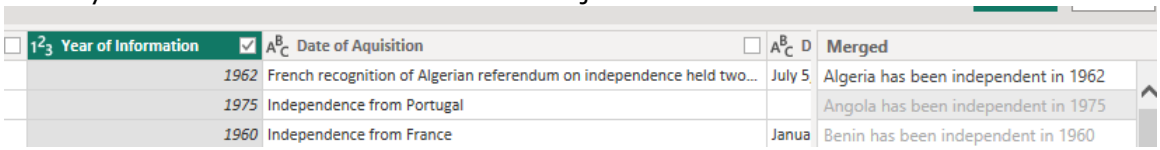
11. We will enter an example and Power BI will detect what we want.
12. In the first cell in the new column type: "**Algeria has become independent in 1962**".
13. Press Enter.
14. You will find that the whole column has been completed the same way.



15. And the message in the top has been changed to inclume **M langauge code** that create the new column.

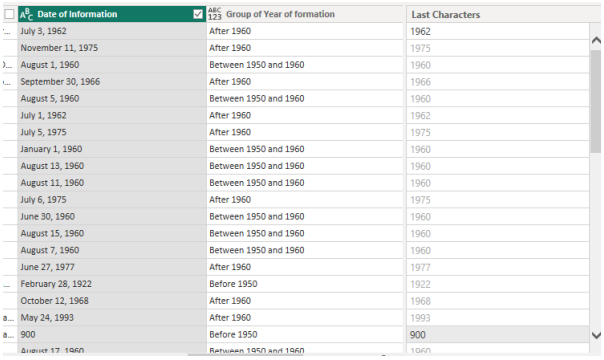


16. If you want to modify the pattern you can click on the cell again and update, and if you want to include another column just click on the check box beside.



17. If you are satisfied with the result, click **OK** on the message on the top to accept.
18. Now we have the column we wanted.
19. Do you remember that we were trying to extract **Year** from column **Date of formation**?
20. Let us try to do the same thing again but this time by using Column from Example this time.
21. First select **Date of formation** column.
22. Add column ➔ General group ➔ Column from Example ➔ from selection.
23. A new column appears at the end.
24. In the first cell of the column write **1962** and press enter.
25. Notice you have all years extracted but line 19 is empty.
26. The data in that line was only 900.

27. This time you can manually write 900 in the cell.

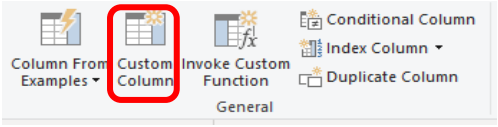


Date of Information	Group of Year of formation	Last Characters
July 3, 1962	After 1960	1962
November 11, 1975	After 1960	1975
August 1, 1960	Between 1950 and 1960	1960
September 30, 1966	After 1960	1966
August 5, 1960	Between 1950 and 1960	1960
July 1, 1962	After 1960	1962
July 5, 1975	After 1960	1975
January 1, 1960	Between 1950 and 1960	1960
August 13, 1960	Between 1950 and 1960	1960
August 11, 1960	Between 1950 and 1960	1960
July 6, 1975	After 1960	1975
June 30, 1960	Between 1950 and 1960	1960
August 15, 1960	Between 1950 and 1960	1960
August 7, 1960	Between 1950 and 1960	1960
June 27, 1977	After 1960	1977
February 28, 1922	Before 1950	1922
October 12, 1968	After 1960	1968
May 24, 1993	After 1960	1993
900	Before 1950	900

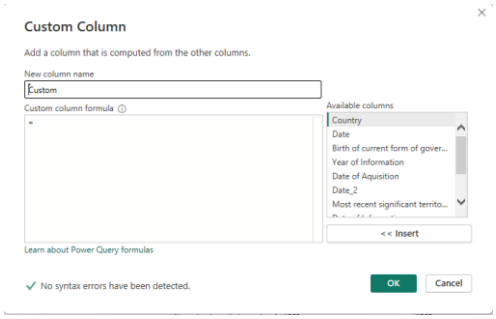
28. Click Ok to accept.

Create a custom Column

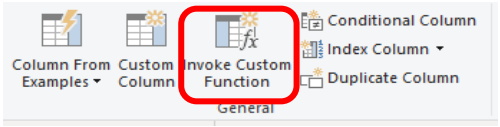
29. Notice that you have an option to create a custom column.



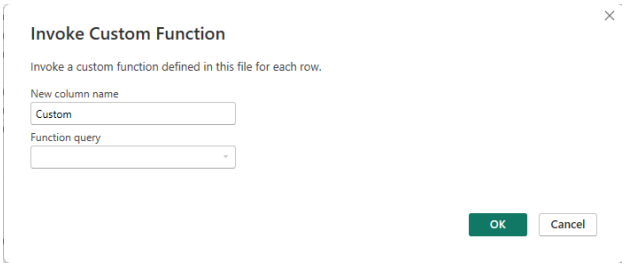
30. When you click it allows you to write **M Language** to create a column.



31. That is too advanced for now.
32. You also have the Invoke custom function option.



33. Here you must create a custom query first and then create a column based on it.



34. That is also a very advanced topic for now.

Knowledge Check

Question 1
Several columns in your worksheet contain missing or null values. Which of the following options must you type in the **Value to Find** field to locate these values?

A. Missing
B. Null
C. 0

Question 2
What steps did you take to address inconsistencies in date columns in your worksheet? Select all that apply.

A. You changed the data type of the column to **Text** by clicking on the **DataType** icon next to the column name.
B. You dropped duplicate rows by selecting **Remove Duplicates** inside the **Remove Rows** menu.
C. You changed the data type of the column to **Date** by clicking on the **Data Type** icon next to the column name.
D. You replaced any empty values with a default date using the **Replace Values** tool.

Question 3
True or false: Once you completed all the data cleaning steps, you clicked **Apply** or **Close & Apply** to apply all the transformations you made.

A. False
B. True

Chapter 6: Advanced Data Transformation

The Importance of data combination

- The first reason for combining data is that it allows you to **consolidate information from various sources** or tables into a single table. This consolidation can provide a unified view of the data, making it **easier to analyze and gain insights**.
- The next reason you would combine tables is to create **relationships**. Combining tables is crucial for establishing relationships between related data. In Power BI, relationships between tables are used to create meaningful visualizations and enable interactive analysis. By combining tables, you can link data points across different tables based on **common fields or keys**.
- Combining tables also enables you to **enrich your data by adding additional information.** For example, you may have a table with client details and another table with product information. By combining these tables, you can create a comprehensive dataset that includes both client and product details, allowing for a more comprehensive analysis.
- Another reason to combine data is that **it provides a broader scope for analysis**. By merging multiple tables, you gain **deeper insights** by analyzing data from different angles.
- Lastly, combining tables help **simplify data management** in Power BI. Instead of working with multiple separate tables, having a single consolidated table reduces

complexity and makes it **easier to handle data updates**, **refreshments**, and **maintenance tasks**.

Ways to Combine Data

- Now that you understand the reasons why it is important to combine data, let's look at the ways to do it.
- In Power BI, there are two ways to combine data: **append** and **merge**.

Append Queries

- When you append queries, you're adding rows of one table or query to another table or query. By adding multiple lists one below the other, you will see **an increase in the number of rows**.
- Say for instance, you have two separate classes, class A and class B, the need to take an exam together. To do this, you have to combine the 20 students in class A with 20 students in class B resulting in a combined class list of 40 students.

Merging Queries

- On the other hand, when merging queries, you consolidate data from multiple tables into a single entity by leveraging a shared column between the tables.
- For example, data with specific contents such as **gender**, **category**, and **city** is stored in different independent tables and referenced by main tables that require this information.
- This allows you to use this information within a specific context, enables easy data classification and ensures data integrity.

Question

Which one of the following describes the reason for Adio's request to combine two different sales datasets together?

A. Enriching data
B. Creating relationships
C. Consolidating information
D. Enhancing analysis

What is a join?

- To combine the data of two tables with different column structures, you need to specify the method in which the two tables should be combined, this is known as a join.
- Join is when you merge or combine data from different places to create a bigger and a more complete dataset. It helps you view all the information in one place, putting puzzle pieces.
- Your manager has tasked you to list all products, but there are category names and indicate which category has the most products.
- During your investigation, you notice that category data is referenced to a table called **categories**. It is also being used by the common columns named **category key**.
- On closer inspection, you notice the row with a category key of one has a category name of bikes and the row with a category key of two has a category name of accessories. Your conclusion is that any row with a value of one in the category key column has bikes as the product category.

- One of the key usage areas of joins is merging the two tables in this manner and matching related data by using the relationship. One of the key usage areas of joins is merging two or more tables and matching related data by using the relationship.
- Joining data is essential for Power BI data analysts because it enables you to combine information from different sources, giving you a complete picture of the data.
- Joining data can help you validate data accuracy, make informed decisions, and perform advanced analysis.
- Joining data also empowers you to gain a holistic understanding, uncover valuable insights, and make data-driven conclusions.
- Overall, join is a powerful technique that enhances your data analysis capabilities and allows you to unlock the full potential of your data.
- When you merge queries, you're combining the data from multiple tables into one based on a column that is common between the tables.
- Merge with join allows you to match related data, integrate data, explore relationships. When you append queries, you are adding rows of data to another table or query.
- Append with join helps you to ensure consistency and allow you to expand your existing dataset.

Question

Which of the following options can be considered as the purposes of join operation? Select all that apply.

A. Integrating data
B. Ensuring consistency
C. Exploring relationships
D. Creating insights from data.

Join keys

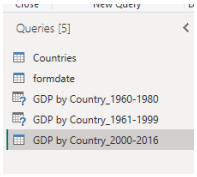
- At this point, it would be beneficial for you to gain an understanding of the table structures where you store the data.
- Tables are used for storing data and may need to be interconnected for mutual data exchange. Predefined and content-specific field values such as gender, city, country, category, type, department and status information can be stored in independent tables. When needed, you can read from these tables.
- For example, if you want to define a product, you will store fields like product name, code, description, and price directly in the product table itself, while reading fields like category and status from separate tables that hold predefined and independent values. It is during this reading process that the concept of join keys will become clearer.
- For example, let's say in the *Category* table, a row with a **CategoryKey** of 1 has a **CategoryName** of "Bikes," and a row with a **CategoryKey** of 2 has a **CategoryName** of "Accessories." If you want to assign the Category of "Accessories" to a product in the *Product* table, all you need to do is select the **CategoryKey** as 2. This way, you have chosen a well-defined value from an existing set for a field with known content like category or status. Using join keys prevents difficulties that may arise from typing the product category incorrectly or using a value that could convey the same meaning. If you have thousands of products, if there is a typing error or a different naming convention in the category field, it may be impossible to find those products listed under the Accessories category. Here is where the join key method provides a crucial solution for classification and categorization.

Constraints

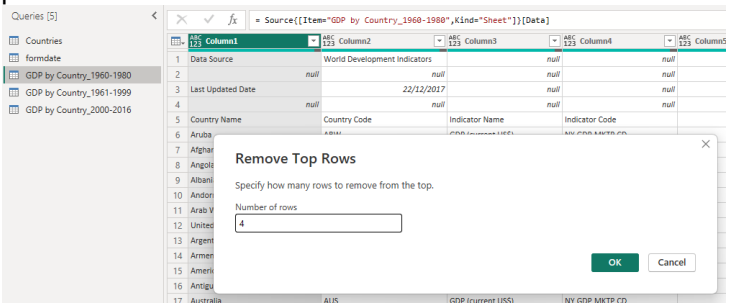
- There are two critical points in the usage of join keys. The first one is that the key field in the reference table must be unique.
- For example, in the *Categories* table, the **CategoryKey** field value is unique at the row level. A value used in one row cannot be used again in another row.
- The other critical point is that the type of the common column between the tables to be merged must be the same. For instance, if the *Product* table is to be merged with the *Categories* table using the **CategoryKey** field, both column types must be the same.

Exercise 13: Merging Queries

1. We want to have more information about each country's GDP.
2. Explore the data in the file "**GDP by country 1960-2016.xlsx**" in your **Files** folder.
3. You have here information about every country's GDP from 1960 till 2016.
4. Notice you have 3 worksheets in the workbook:
    - From 1960-1980
    - From 1951-2016
    - From 2000-2016
5. We want to find a way to combine all that information in one table.
6. We can use the **merge queries** option in Power Query to do that.
7. But first let us connect those 3 worksheets into Power Query.
8. Use the **New Source** button in Power Query and load the 3 worksheets from the Excel file.
9. Now you must have queries available in project.

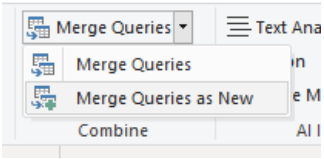


10. Select each query of the three one by one.
11. Use Home ➔Reduce Rows group➔Remove Rows➔Remove top Rows.
12. Remove top 4 rows.

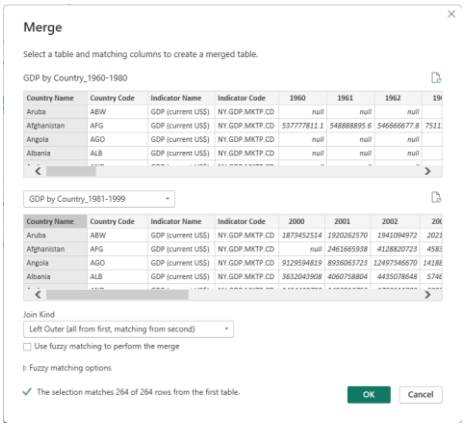


13. Then set the first row as a header.
14. Home➔Transform group➔Use first row as headers.
15. Now we can merge query to a new query.
16. Select the first Query **1960-1980**.

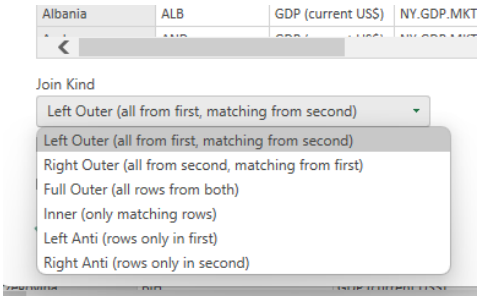
17. In Home➔Combine group➔Merge Queries➔Merge Queries as New



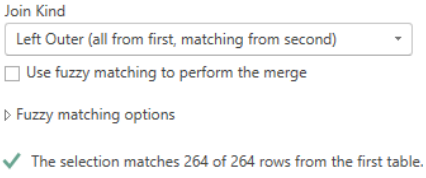
18. That will give us a new table have two tables merged.



19. Select your 2nd table **1981-1999**.
20. Select the **matching columns** that merge the two tables.
21. Here choose the **Country Name** in both tables (also **Country Code** will do).
22. Go and explore what type of joins are available in the Joint **Kind** drop-down box.

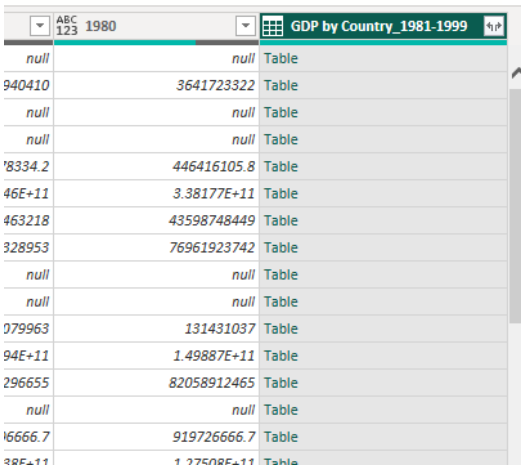


23. Notice that you will get 264 rows merged if you used this kind of joint.

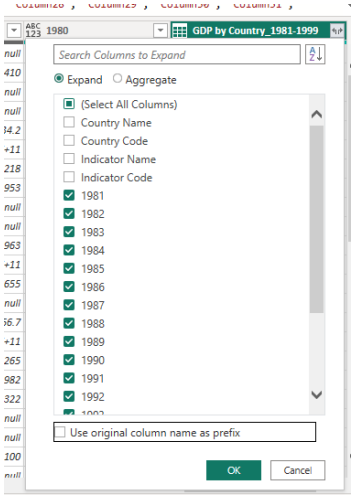


24. Try all 6 other types of joints and see how many rows you will get.
25. Because you have 264 rows in both tables matched you will always get the same result whatever type of joint you used.
26. Click Ok to get a new query "**Merged1**".
27. Review your new table.
28. table starts with column from 1960 till 1980.

29. Then you have empty columns.
30. Those columns were there from the start.
31. Select them and right click and select **remove columns**.
32. Notice we have something weird.
33. Just a column with word **table** in each cell.

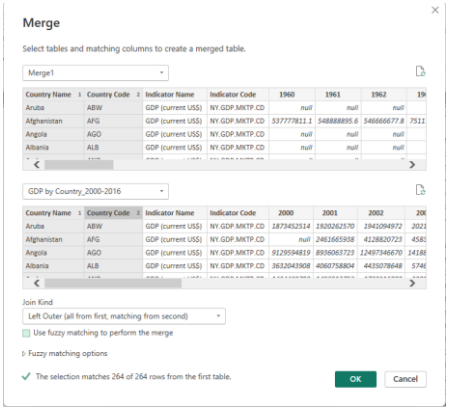


34. Click on the **two arrows** on the right of the new column header.
35. select the columns you want to include from the second table.
36. Also uncheck using the second table name as a prefix.



37. Click **Ok** to get the columns you want.
38. Now you can repeat the same steps to merge table "**Merge1**" you have just created and table "**2000-2016**".
39. You can use **Merge Queries** to get the result.
40. But for ease use **Merge Queries as New** and get new table "Merge2".
41. This time use 2 matches column, this will give you the same result but sometimes you will need to do that.

42. Use **CTRL** button to select more than one column in both query.



43. Also select only columns from **2000 to 2016** in the new query "**Merge2**".
44. Now you have final query "Merge2" that have 4 columns of data and the rest columns are GDP from 1960 till 2016.

Join types

- A join type in Microsoft Power BI refers to how tables of data are related to each other in the software. The joins are important because they determine how data is consolidated from multiple sources into a single view.
- Understanding join types and their implications is crucial to building accurate, efficient, and meaningful data models in Power BI
- . Let's say we have two tables, one on the **left** for sales and one on the right for **countries**. The sales table has three columns, date, countryID, and units. The countries table has two columns, ID and country.
- The sales table, **countryID** column can be used as **a join key** with the **ID** column of the countries table.
- Now let's explore each join type and how they combine data.

left outer join

- If a left outer join is used, all rows in the left table are kept and the matching rows from the right table are merged in.
- If the left table is missing columns that the right table has, the columns are included as part of the merge.
- It is important to note that if there is no match for a row between the tables, default or null values will be used for columns where matching data is unavailable. In this scenario, the resulting table will have the columns from the left table; date, countryID, and units, along with the country name column.
- Since the right table did not have a countryID of four, the country name is null.

Eight Outer Join

- A right outer join works similarly to the left outer join, except that all rows in the right table are kept and the matching rows from the left table are merged in.
- Again, if the right table is missing columns that the left table has the columns are included as part of the merge.
- Similarly, if there is no match for a row between the tables, default or no values will be used for columns where no matching data is available.

- In our scenario, the resulting table will have date, countryID, units, and country name.

Full Outer Join

- The full outer join is used when you want to retrieve all records from both tables, regardless of whether they have matching values in the join condition.
- In this scenario, since the right table has an ID of four and the left table does not have a corresponding entry with a countryID of four, a row is created with a country name for ID four, and with null values in all other columns. In the previous video, what is a join? You used full outer joins and append with joins by matching related data.

Inner Join

- For inner join, only matching rows from both left and right tables are merged together.
- This join type is helpful when you want to focus only on the sales that have corresponding data in another table and exclude any sales data that don't match.

- As a data analyst, you often come across the requirement to combine data from different tables or datasets related to sales and product tables. This is where merging operations, specifically join types become crucial.

- Keep in mind that you should choose the combination types based on how you choose them. Taking into account the specific needs of the analysis. The choice of join type will impact the inclusiveness of the data in your analysis. It's important to consider your analysis objectives and the specific requirements of your project.

- Each join type serves a different purpose and selecting the appropriate one, ensures that you obtain the desired results set for your analysis of order and order details data.

Question

Which type of JOIN operation includes only the matching records from both joined tables?

Select the correct option.

A. INNER JOIN
B. LEFT OUTER JOIN
C. FULL OUTER JOIN

Unpivot and pivot columns

- Unpivot and pivot operations are data transformation techniques that you can use to reshape and restructure data in Power BI.

Unpivot

- The unpivot operation refers to the transformation of data from a **wide** format with multiple columns to a **narrow** format with fewer columns by reshaping the data structure.
- It involves converting column headers into row values resulting in a more structured and standardized representation of the data.

- The unpivot operation is useful in data analysis supporting data normalization by organizing data in a tabular format.
- This facilitates analysis, variable comparison, and data aggregation and summary as related information is consolidated into a single column.
- Transforming data from a wide to a narrow structure can also enable data compatibility and integration with other systems or tools that require a narrow format..

Pivot

- On the other hand, the pivot operation refers to the transformation of data from a **narrow** format with fewer columns to a **wide** format with multiple columns by reorganizing the data structure.
  It enables data analysts to convert rows into columns based on specific criteria or values.
  This operation is often used to:
  - summarize and aggregate data,
  - create cross-tabulations, and
  - represent data in a more structured easy-to-understand way for analysis and reporting.

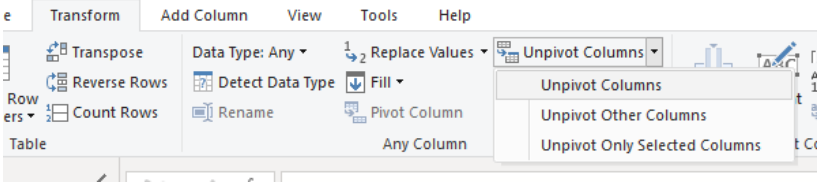
Question

Which of the following statements describes the pivot operation in Power BI? Select all that apply.

A. The pivot operation converts data from a narrow format to a wide format by reorganizing the data structure.
B. The pivot operation is used for data aggregation and summarization by converting rows into columns.
C. The pivot operation supports data normalization by converting column headers into row values.
D. The pivot operation involves transforming data from a wide format to a narrow format.

Exercise 14 Pivoting and Unpivoting

1. We want to make our data easy and have just one column for the year and what was the GPD in this year.
2. That would make our visualization and analysis easier.
3. First, we do not need columns "**Indicator Name**" and "**Indicator Code**".
4. So, select and Remove.
5. We want now to unpivot our table.
6. Select all year columns (1960 till 2016).
7. Transform➔Any Column group➔Un Pivot Columns.

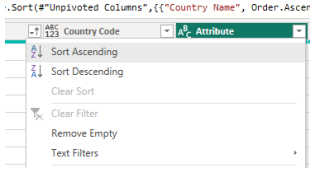


8. Power Query gives you a new two columns.
   - **Attribute**: the names of the column we had before (Year).

- **Value**: the value associated with that attribute (Year).

	Attribute	Value
	100% Valid 0% Error 0% Empty	100% Valid 0% Error 0% Empty
	1994	1330167598
	1995	1320670391
	1996	1379888268
	1997	1531843575
	1998	1665363128
	1999	1722798883
	2000	1873452514
	2001	1920262570

9. You can go and **sort** the **country** column than the **Year** Column.
10. This will give you Country sorted alphabetically then Year sorted inside.
11. That is just to show you are doing well till now.

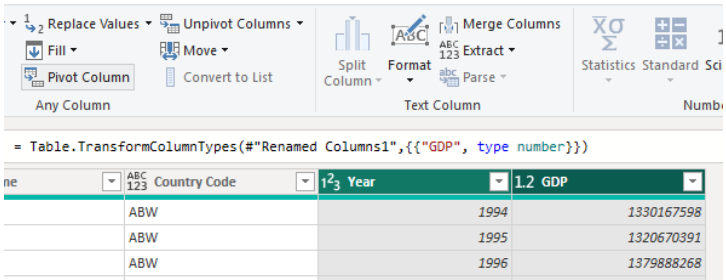


	Country Name	Country Code	Attribute	Value
1	Afghanistan	AFG	1960	537777811.1
2	Afghanistan	AFG	1961	548888895.6
3	Afghanistan	AFG	1962	546666677.8
4	Afghanistan	AFG	1963	751111191.1
5	Afghanistan	AFG	1964	800000044.4
6	Afghanistan	AFG	1965	1006666638
7	Afghanistan	AFG	1966	1399999967
8	Afghanistan	AFG	1967	1673333418
9	Afghanistan	AFG	1968	1373333367
10	Afghanistan	AFG	1969	1408888922

12. Delete this **sorting** step.
13. As you can see selecting all the other columns was so tedious.
14. Let us delete **unpivoted columns** step and redo it another way.
15. This time do not select the columns you want to unpivot.
16. Select the columns "**Country Name**" and "**Country Code**".
17. Transform➔Any Column group➔Un Pivot Columns➔Unpivot other coulmns.
18. You will get the same result.
19. Change the new column names and their data type:
    - Attribute➔Year (Whole Number).
    - Value➔GDP (Decimal Number).
20. Also change the table name **Merge2 ➔GDP by Country and Year.**

Try Pivoting

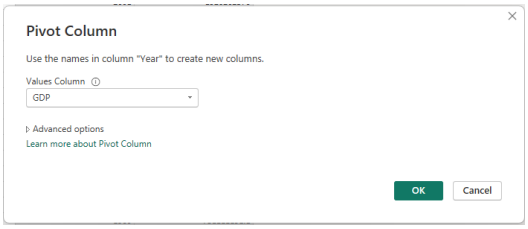
21. Let us try pivoting which is the opposite of unpivoting.
22. You must pivot only two columns.
23. Try to select 3 columns, you will find the **pivot column** button is unavailable.
24. Select now **Year** and **GDP** columns.
25. Transform➔Any Column group➔Pivot Column.



26. Confirm GDP as your Value Column and click OK.
27. You will get the table as before pivoted.
28. Delete this Step.

Append Tables

- By now, you know that there are two ways to combine data in Power BI, **append** and **merge**.
- When **merging** queries, you consolidate data from multiple tables into a single entity



  by leveraging a shared column between the tables.
- When you **append** queries or tables, you add rows from one or more tables to another query or table.
- Before you append the queries, you have to format the data of both files to ensure they have:
  - **an equal number of columns** and
  - that the columns have **the same names** and **data types**.

If you don't have an equal number of columns or different column names:

- the extra columns will be added to the most **right** of the query by preserving their values in the originating query and setting **null** values for the matching new query.
- This may be confusing. So try to have an equal number of columns with the same column titles.

Merging Queries Steps:

- On the Power Query Editor ribbon, navigate to the Home ribbon tab and select the **Append Queries** dropdown menu.
- You can:
  - select **Append Queries as New** to create a new query or table from the Appended output, or
  - select **Append Queries** to merge the rows from an existing table into another.
- If you select **Append Queries as New**, you will create a new **master table**.
- This selection displays the Append window where you can select the tables you want to combine from the available tables section and add them to the Tables to append section.
- When you select OK, a master table is created that contains the sales data of both tables.

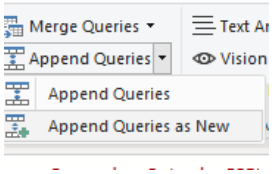
Question

Which of the following is the operation of putting two or more tables or queries in one master table together?

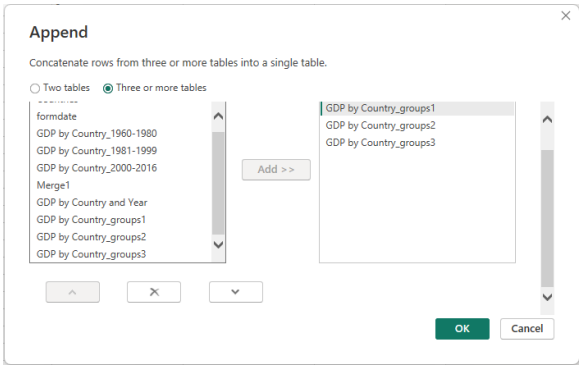
A. Appending
B. Merging
C. Combining

Exercise 15: Appending Queries

1. Merging was to combine columns.
2. Appending means to combine rows.
3. To do this the tables must have the same columns.
4. We want to load 3 tables.
5. This time they are text files.
6. Go and load the three files from Files folder: (**GDP by Country_groups1, GDP by Country_groups2, GDP by Country_groups3**).
7. Explore the file first in Notepad.
8. Connect files one by one to load into Power Query.
9. Home➔New Source➔Text/CSV.
10. Notice the three tables have the same structure.
11. We want to combine those three queries into one.
12. Home➔Combine group➔Append Query➔Append Queries as New.

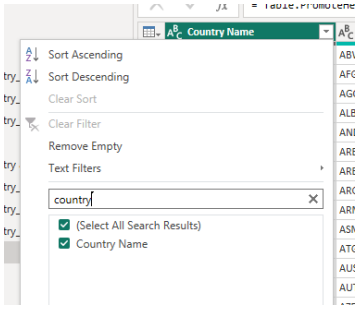


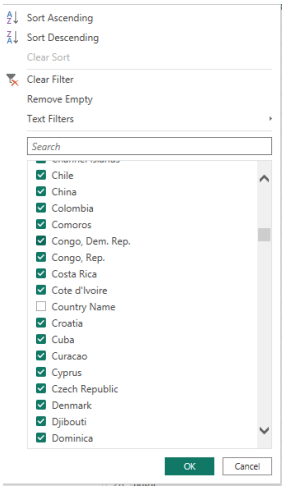
13. Select three or more tables option.
14. Select your 3 tables.
15. Notice that you have options in the bottom to **delete, move up** or **move down** your tables.
16. Click OK to get new Query **Append1** have all information in one single table.
17. Select the two empty columns and remove.
18. If you go back to the original tables, you will find that in each one there were a **header** row that was treated as **data.**



19. You must remove those rows from the new **Append1** table.

20. Select **Append1** Query.
21. Click **use first row as headers** button.
22. Click in the down arrow of the first column (**Country Name**) to see what data



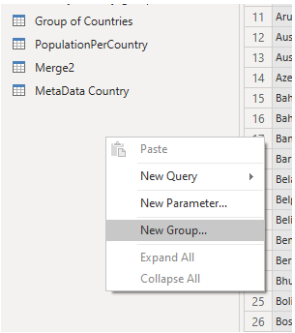


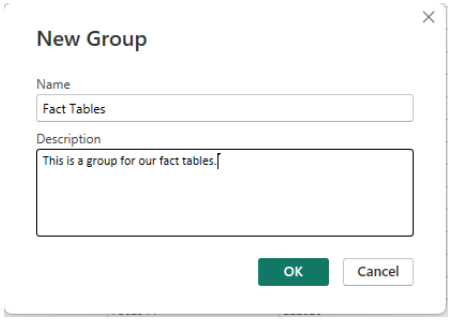
inside it.

23. Search by country word to filter it will give you the value of Country Name.
24. Uncheck the box beside the country name to filter all that rows.
25. And do not forget to click the X right the search box to remove the country filter.
26. Rename **Append1** query as **Group of Countries**.
27. Close and apply your changes and save your file.

Exercise 16: Organizing Queries in Groups

1. We want to keep our queries more organized.
2. We can use Groups and move certain queries to each group.
3. Right click any empty space in Queries Pane in the right and chose New Grope
4. Give it a name **Fact Tables** and a description.

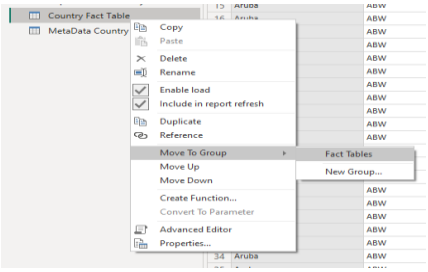




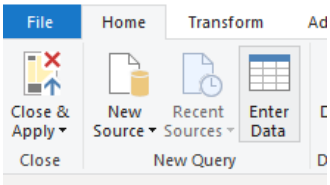
5. Move **Country Fact Table** to the new group either drag and drop or right click and choose **Move to Group**.
6. Create another group and name it "Dimension Tables" and move tables "**Countries**" and "**MetaData Country**" into it.

Exercise 17: Entering Data Manually

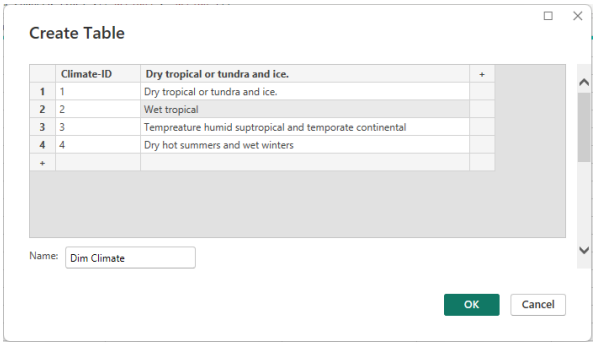
1. Look at **Countries** table.
2. You have a column called **Climate**.
3. Its data is Numbers (1.2.3,4).
4. We have searched on the internet and recognized Climate ID mean.
    1. Dry tropical or tundra and ice.



    2. Wet tropical.
    3. Temperature humid subtropical and temperate continental.
    4. Dry hot summers and wet winters.
6. It would be better to have this information in a Dimensional table.
7. So we will not have to enter those long lines many time in our fact table.
8. We could create this table manually.
9. Home➔ New Query ➔Enter Data

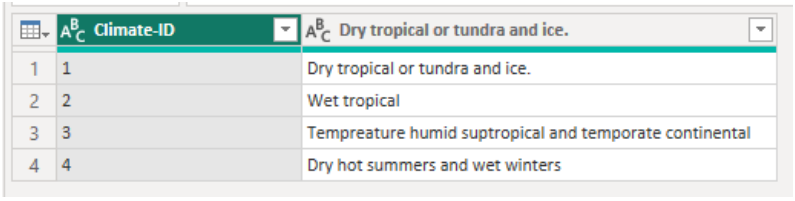


10. Enter your data in the table.



11. Click OK to get the new table added.
12. Move it to **Dimension Tables** group.
13. What if you want to edit this table? where?
14. You can click the **Gear Icon** on right of Source step of this Query.

15. This will open the edit table window.



	Climate-ID	Dry tropical or tundra and ice.
1	1	Dry tropical or tundra and ice.
2	2	Wet tropical
3	3	Tempreature humid suptropical and temporate continental
4	4	Dry hot summers and wet winters

Knowledge check

Question 1

Which feature allows you to combine related data between differently structured data sources in Power Query?

A. Merging
B. Grouping
C. Appending

Question 2

Which of the following can be considered as a purpose of merging data with joins? Select all that apply:

A. Exploring Relationships
B. Expanding Data
C. Integrating Data
D. Matching Related Data

Question 3

True or False. The full outer join is useful when you want to retrieve all the records from both tables, regardless of whether they have matching values in the join condition.

A. True
B. False

Question 4

You import 4 Microsoft Excel tables named *Sales*, *Product*, *Reseller* and *Employee* into Power Query.

Sales contains the following columns:

- **SalesOrderNumber**
- **OrderDate**
- **ProductKey**
- **ResellerKey**

- **EmployeeKey**
- **SalesTerritoryKey**
- **Quantity**
- **Unit Price**
- **Sales**
- **Cost**

Your manager asked you to list ***Sales*** data with the descriptive information from the ***Product***, ***Reseller*** and ***Employee*** tables for the columns which have the suffix "**Key**". What should you do to accomplish this task? Select all that apply:

A. Join *Sales* and *Product* tables based on the **ProductKey** column.
B. Merge the *Sales* table with the *Product*, *Reseller* and *Employee* tables respectively.
C. Join *Sales* and *Reseller* tables based on the EmployeeKey column.
D. Check the column types of (**ProductKey**), (**ResellerKey**) and (**EmployeeKey**) in the *Sales*, *Product*, *Reseller* and *Employee* tables.

Question 5

You import two Microsoft Excel tables named *Product* and *Categories* into Power Query. There are 319 rows in the ***Product*** table. Nine of the total rows in the *Product* table do not have Categories data, so the **CategoryKey** of these rows has **NULL** values.

Your manager asked you to list Product data by showing their category names including the rows which have NULL values in **CategoryKey** column. What should you do to accomplish this task?

A. Merge *Product* and *Categories* tables based on **CategoryKey** column by choosing **Left Outer Join** in the join kind dropdown.
B. Merge *Product* and *Categories* tables based on **ResellerKey** column.
C. Merge *Product* and *Categories* tables based on **CategoryKey** column by choosing **Inner Join** in the

Chapter 7: Data Profiling

Introduction to Data Profiling and Statistical Analysis

- Before analyzing any data set, it is important to examine and evaluate the data you are working with. Analyzing the data without evaluating its accuracy, completeness and alignment with the objectives can lead to misleading results.
- When examining a data set for the first time, there are several aspects you should look at, especially for numerical fields.
- You should check these characteristics for each numerical field:
    - minimum or min,
    - maximum or max,
    - average or mean,
    - frequently occurring values or mode and
    - standard deviation.
- The best way to start assessing data is with data you can immediately troubleshoot.

Example

- Imagine you are reviewing a data set that has an **age** field. For instance, there could be someone in the data set with an age of **200**, which would be extremely unlikely to be true. If so, there may be an **outlier** in the data.
- Look at the **minimum** and **maximum** values, such as appearing between **21** and **77**. These are realistic ages, unlike 200.
- The concept of **distribution of data** refers to **how the data points are spread or arranged within a data set**. It describes the **pattern** or **shape** of the data when plotted on a graph.
- Understanding the distribution of data is crucial in data analysis because it helps you gain insights into the **central tendency**, **variability**, and overall characteristics of the data.

Outliers

- The formal definition of an outlier in statistics is **a data point that significantly deviates from other observations**.
- Outlier data can be handled by applying a technique called **min, max scaling** or **normalization**.
- The aim is to adjust the mean and standard deviation of the data proportionately while preserving the ratio of the distance between outlier data and other data points.
- Analyzing the distribution allows you to make informed decisions, identify outliers, and choose appropriate statistical techniques for further analysis.
- There are situations where there may be values in the data set that skew the average.
- For example, there may be examples close in age. Let's say there are three individuals aged 80 and above.
- If you solely rely on the average to evaluate the distribution, these outliers can mislead you by increasing the average. In this case, it would be appropriate to examine the distribution more closely.
- When taking a closer look at the data, you may find that the distribution is normal, but the three records mentioned in the example are outliers.

Standard deviation.

- Standard deviation is **a statistical measure that quantifies the amount of variation or dispersion in a data set**.
- It provides a way to understand how individual data points differ from the mean or average of the data set.
- The main objective here is to prevent outliers from causing deviations in your analysis results, minimizing their impact.

Distribution of data.

- The balanced distribution of data points that fall outside the outliers is another factor that affects data quality and your analysis results.
- It is important for descriptive variables such as age, gender, income status, occupation, city and neighborhood to represent as many diverse groups as possible and be evenly distributed among others.
- If not, a cluster of records that closely resemble each other will lead to narrow intervals when defining norms which will mislead your analysis.

- Profiling and statistically analyzing data, including examining its **distribution**, **min max**, **mean** and **mode** values detecting **outliers**, if any, and **normalizing** outliers, ensuring that the data represents the entirety of the data set, are the key elements that demonstrate data quality.
- Considering these factors will enhance the accuracy and quality of analysis and predictions made with this data.

Question

Which column characteristic gives the most frequently repeated value in selected records? Select the correct option.

A. Mode
B. Average
C. Min

Profiling Data in Power BI

- Your Company recently conducted a field **survey** to increase sales and collected potential customer data.
- This resulted in an **Excel** file containing information such as **age**, **gender**, **occupation**, **income level**, **address**, and **phone number** of prospective customers.
- Since the survey data was collected manually, it was not subjected to any validation.
- Therefore, before analyzing the data, it is necessary to confirm that the data is:
    - **valid**,
    - within the desired **ranges** and
    - **quantities**, and **exhibits** a good **distribution**.

- Before starting analysis on any data set, it is important to examine the data by examining various aspects such as:
    - completeness,
    - accuracy,
    - uniqueness, and
    - consistency.
- Data profiling enables the identification of potential **issues** and **anomalies** within the dataset.
- This proactive approach allows you to make **informed decisions** about:
    - data cleaning,
    - transformation and
    - enrichment,
- ultimately leading to **improved data quality**.
- Additionally, data profiling facilitates:
    - effective data **exploration** and **visualization** by providing insights into data patterns, relationships, and trends.
- It empowers users to:
    - discover hidden insights,
    - uncover data inconsistencies, and
    - make data-driven decisions with confidence.

Unique and Distinct

- Before delving into data profiling tools, let's first consider two important factors in data profiling, unique and distinct.
- In Power BI,
    - **Unique**: total number of values that only appear once.
    - **Distinct**: total number of different values, regardless of how many of each you have.

Profiling Tools in Power Query

- Microsoft power BI offers the following two profiling tools in the Power Query editor:
    - column quality, and
    - column distribution.

Column quality

- Column quality focuses on **valid**, **error** and **empty** rows on each column, allowing you to validate your row values.
- The column quality feature labels values in rows in five categories:
    - **Valid**, shown in ➔ green,
    - **Error**, shown in ➔ red,
    - Empty, shown in ➔ dark gray,
    - Unknown, shown in ➔ dashed ------- green, indicates when there are errors in a column, the quality of the remaining data is unknown.
    - Unexpected error shown in ➔ dashed --------red.
- These indicators are displayed directly **underneath the name of the column** as part of a small bar chart.
- The number of records in each column quality category is also displayed as a **percentage**.
- By **hovering over** any of the columns, you are presented with a **numerical distribution** of the quality of values throughout the column.
- Additionally, selecting the **Ellipses button** opens some **quick action** buttons for operations on the values.

Column Distribution

- Column distribution provides a set of visuals underneath the names of the columns that showcase the frequency and distribution of the values in each of the columns.
- The data in these visualizations is sorted in **descending** order from the value with the highest frequency.
- By hovering over the distribution data in any of the columns, you get information about the overall data in the column, with **distinct count** and **unique** values.
- You can also select the **Ellipses button** and choose from a menu of available operations.

Example

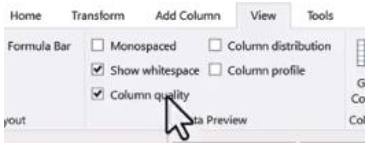
- Imagine that you have a selection of bike accessories that are supplied by four different suppliers, **supplier A**, **supplier B**, **supplier C**, and **Supplier D**.
- In this case, there are **four distinct suppliers**.
- Now imagine you have two bikes, each with a unique supplier to any other bikes you currently stock. These would be considered **two unique suppliers**.

Column Profiling

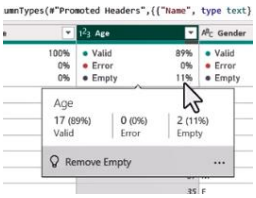
- Another type of profiling in Power BI is column profile.
- Column profile provides column statistics such as **minimum**, **maximum**, **average**, **frequently occurring values**, and **standard deviation**, and in addition, **value distribution on the selected column**.
- This is very important when assessing data to detect anomalies and outliers.

Apply Data Profiling in Power BI

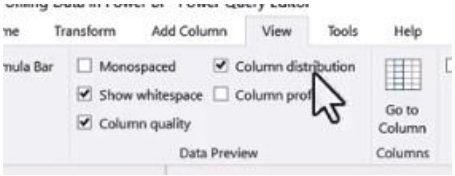
- Your company conducted a field survey to increase sales and collected potential customer data.
- This survey resulted in an Excel file containing information such as **age, gender, occupation, income level, address,** and **phone number** of prospective customers. The survey data was collected manually. It was not subjected to any validation.
- Therefore, before analyzing the data, it is necessary to confirm that the data is valid, within the desired ranges and quantities, and exhibits a good distribution.
- Check the **Column Quality** checkbox inside the **Data Preview** group of **view** to assess column quality.



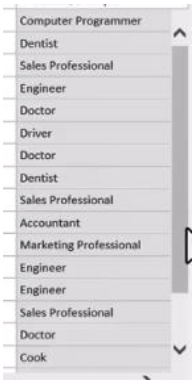
- In the **age** column, **89%** of the values are valid, **0%** of the values are error, and **11%** of the values are empty rows.



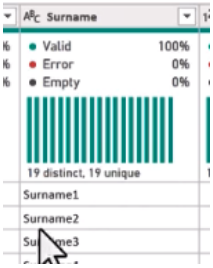
- To assess **column distribution** for the Occupation column, on the **View** tab, from inside the **Data Preview** group, check **column distribution**.



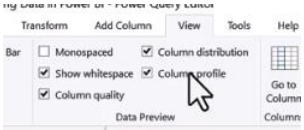
- Note that there are **9 distinct** values and **2 unique** values.



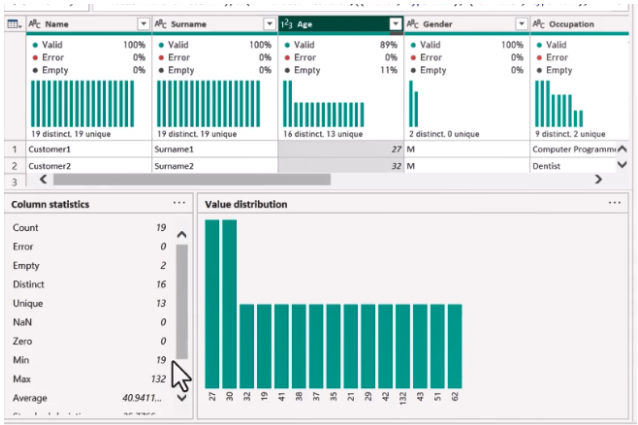
- **Computer programmer** and **accountant** are the occupations which appear only once.
- For each column, note that **if all the row values are distinct, then unique and distinct amounts will be equal**.
- For example, you can see that there are **19 distinct** and **19 unique** values for the surname column.



- Select the **age** column and then check the **Column Profile** checkbox.



- Note that maximum value for age column is **132**, which is not acceptable.



- Examine the **minimum**, **maximum**, **average** and other column statistics and review the value distribution chart.

Question

Which menu item gives the distinct and unique row values amounts for a selected column?

A. Column Distribution
B. Column Profile
C. Column Quality

Using the data profiling tools

Now, we will consolidate this knowledge and examine how we can detect and analyze them using Microsoft Power BI tools.

Microsoft Power BI offers data profiling tools in Power Query Editor as follows:

- Column quality
- Column distribution
- Column profile

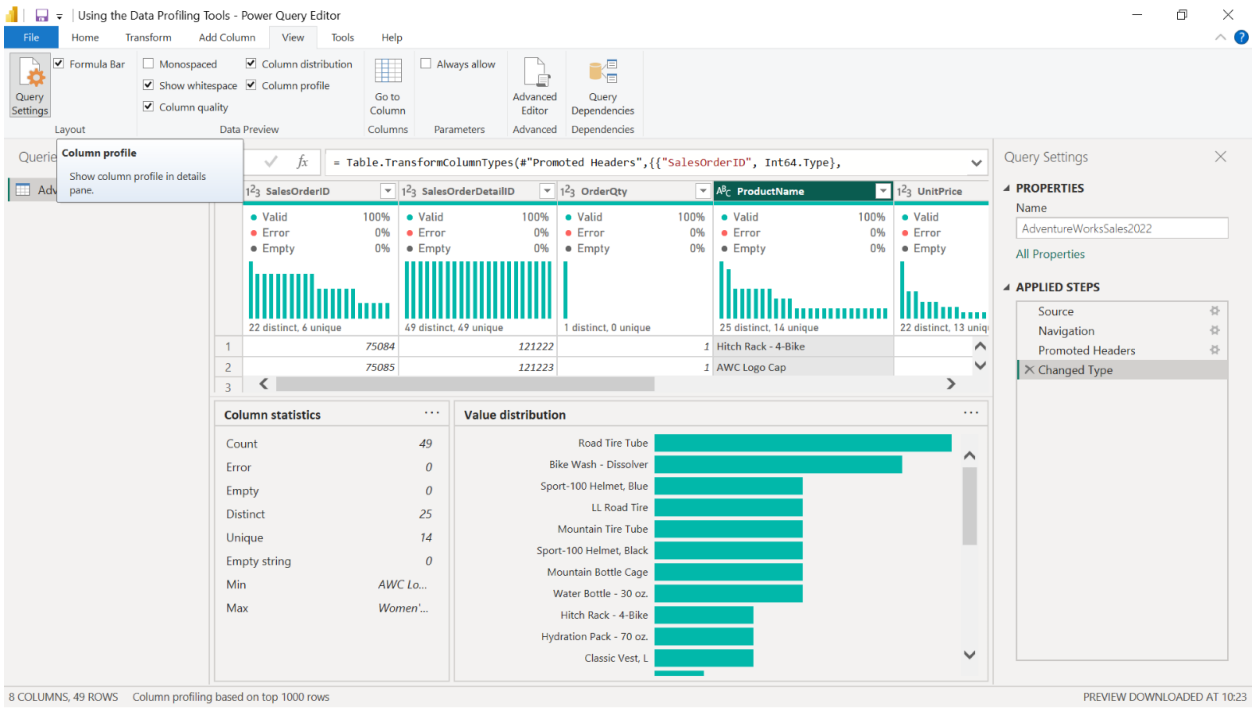
These 3 data profiling tools are inside the **Data Preview** group when the **View** ribbon tab is selected. By default, **Power Query performs data profiling on the initial 1,000 rows of your data.**

To include the entire dataset in the profiling process, you can modify the column profiling settings by checking the lower-left corner of your editor window.

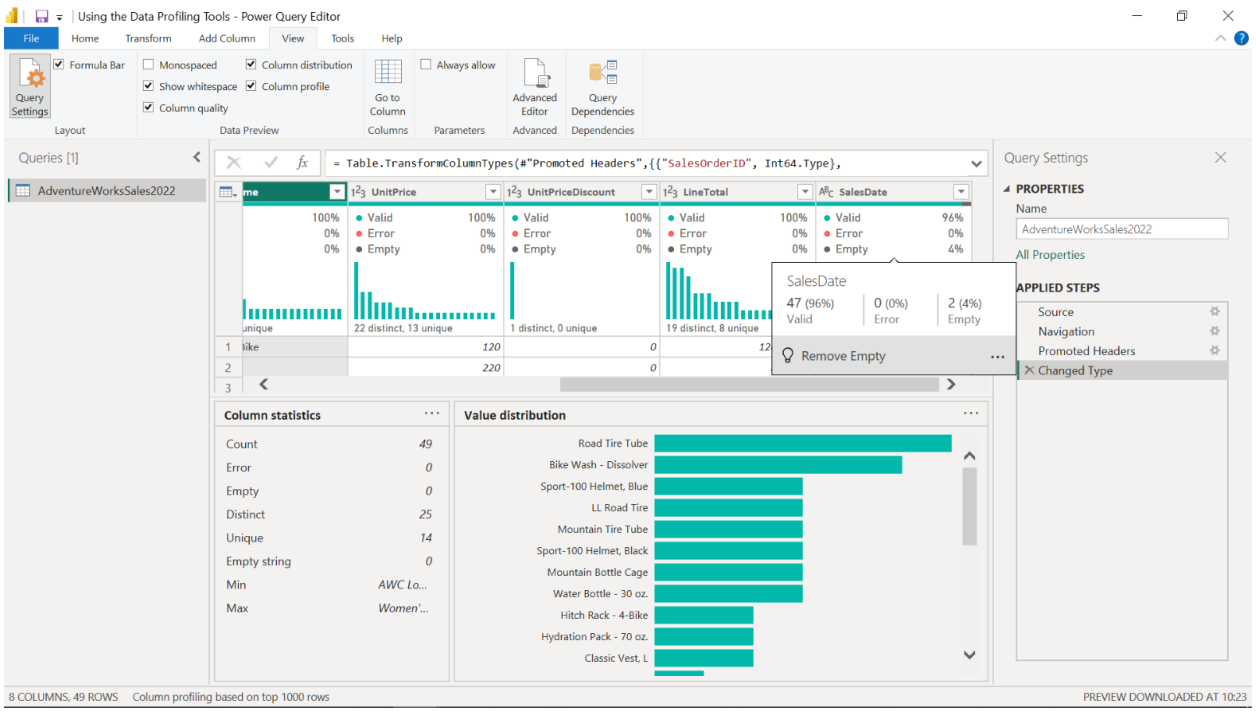
Column quality

There are 3 main labels for column quality,

- Valid, shown in green.
- Error, shown in red.
- Empty, shown in dark grey.



By simply hovering over any of the columns, you can instantly view the numerical distribution of values within the column, providing insights into the quality of data. Moreover, by clicking on the ellipsis button (...), you gain access to a set of convenient action buttons for performing operations on the values. Column quality allows you to validate your row values and find out the empty rows.



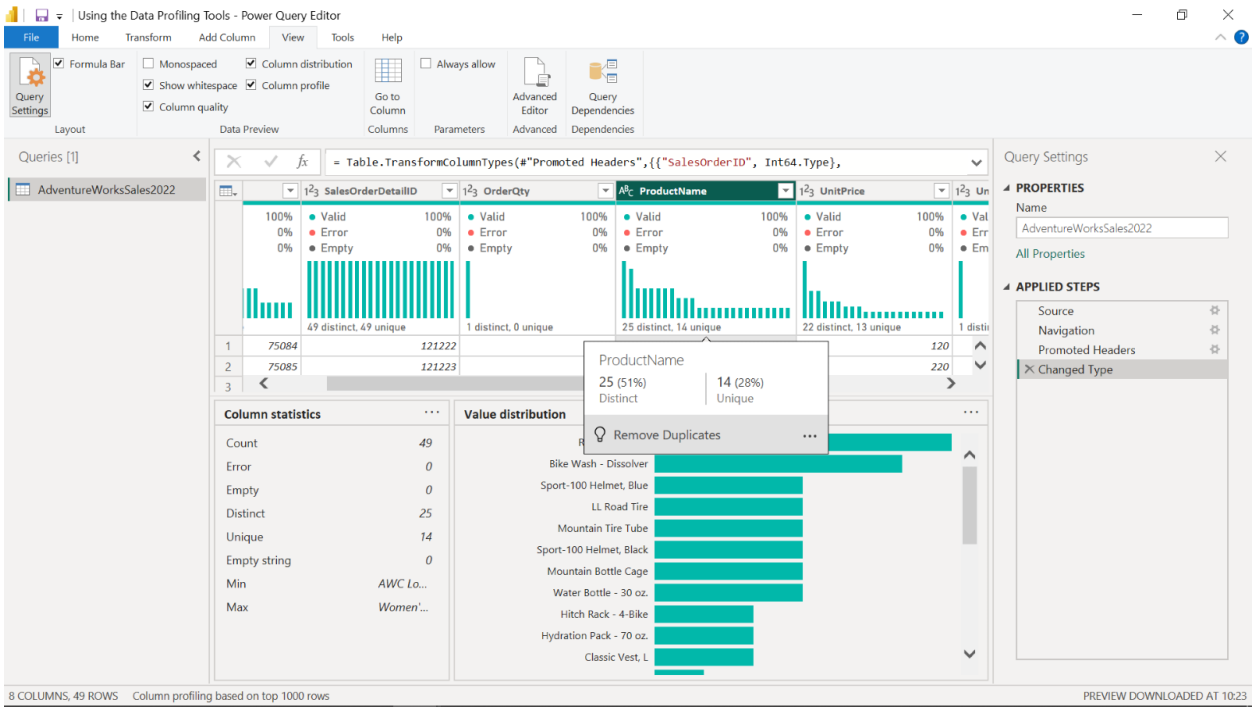
Column distribution

By activating the **Column distribution** feature, you get **Distinct** amount, **Unique** amount, f**requency** and d**istribution** of the values under the validity information. As mentioned before, **Distinct** is known as “total number of different values”, whereas **Unique** is described as “total number of values that only appear once”.

Distinct works by checking each value in the dataset and if the value has not yet been seen, the distinct count is incremented by 1.

Unique works by finding the total number of values that only appear once in the dataset.

By simply hovering over the distribution data in any column, you can quickly access valuable information about the data within that column, including distinct count and unique values. Additionally, by clicking on the ellipsis button and exploring the menu options, you can conveniently perform various operations on the data.

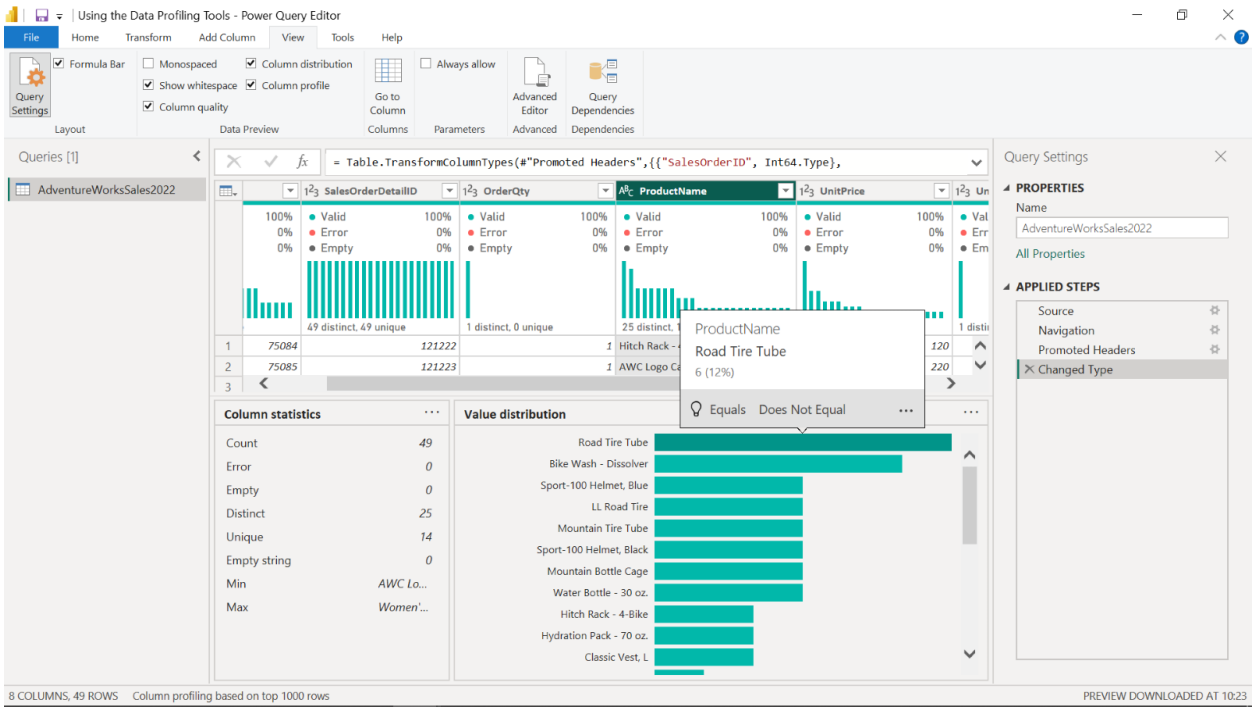


Column profile

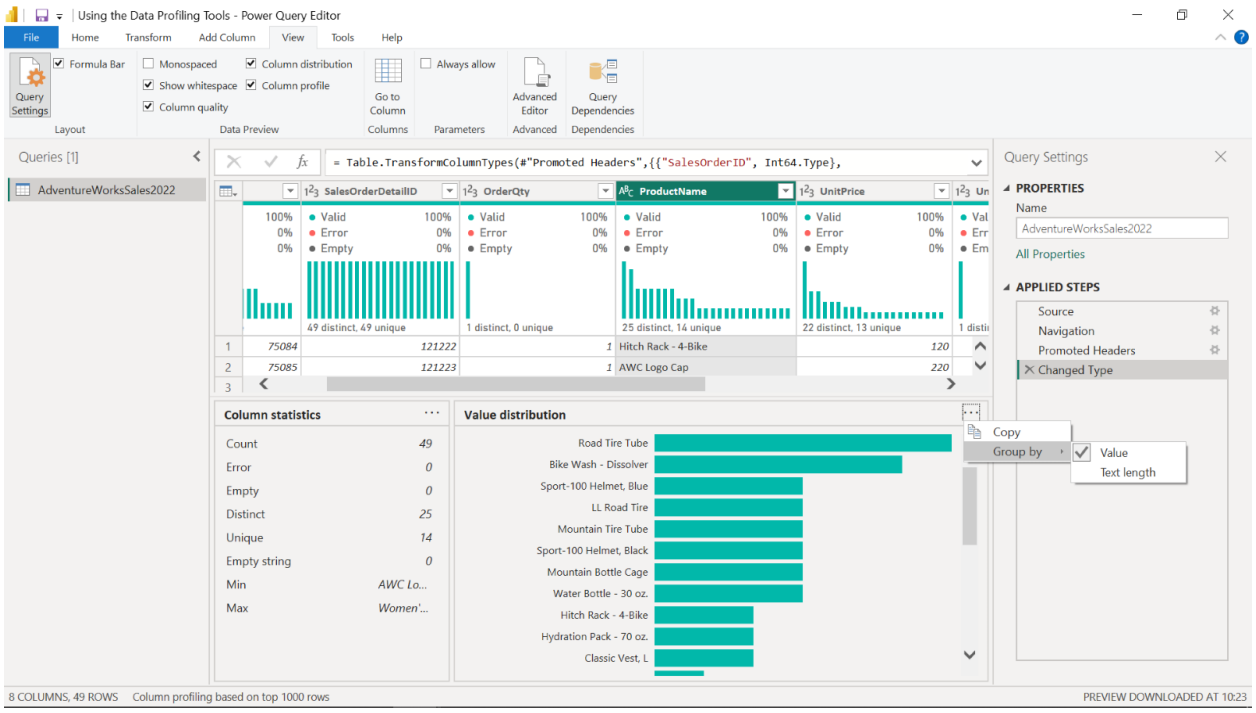
Column profile gives you column statistics and value distribution on the selected column. Some of the available statistics are minimum, maximum, average, count, empty, unique, distinct and empty. Here are the short definitions of the column statistics in Power BI column profiling.

- Minimum: The smallest value in a column.
- Maximum: The largest value in a column.
- Average: The arithmetic mean of all the values in a column.
- Count: The total number of values in a column.
- Empty: The number of empty or null values in a column.
- Unique: The total number of values that only appear once in a column.
- Distinct: The total number of different values in a column.

Statistics and value distribution give valuable insights for detecting anomalies and outliers. The formal definition of an outlier in statistics is a data point that significantly deviates from other observations. An outlier refers to an individual data point or a group of data points that deviates significantly from the remaining data set. On the other hand, an anomaly represents a single point or a group of points that exhibit considerable distance from other points in the multi-dimensional feature space.



By hovering over any of the bars in the chart, you can easily access detailed information about that specific segment. Furthermore, right-clicking on a bar presents a range of available transformations specific to that value. Additionally, by selecting the ellipsis button (...) located in the upper-right corner of the value distribution chart, you can utilize the "Group by" feature to group your chart values based on various options such as value, parity, sign, or text length.

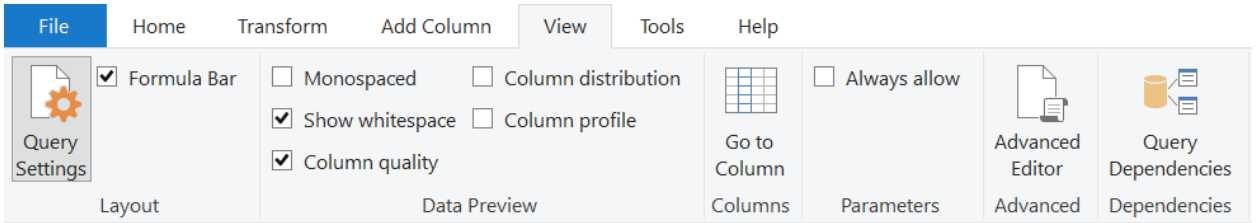


Exercise 18: Profiling a dataset

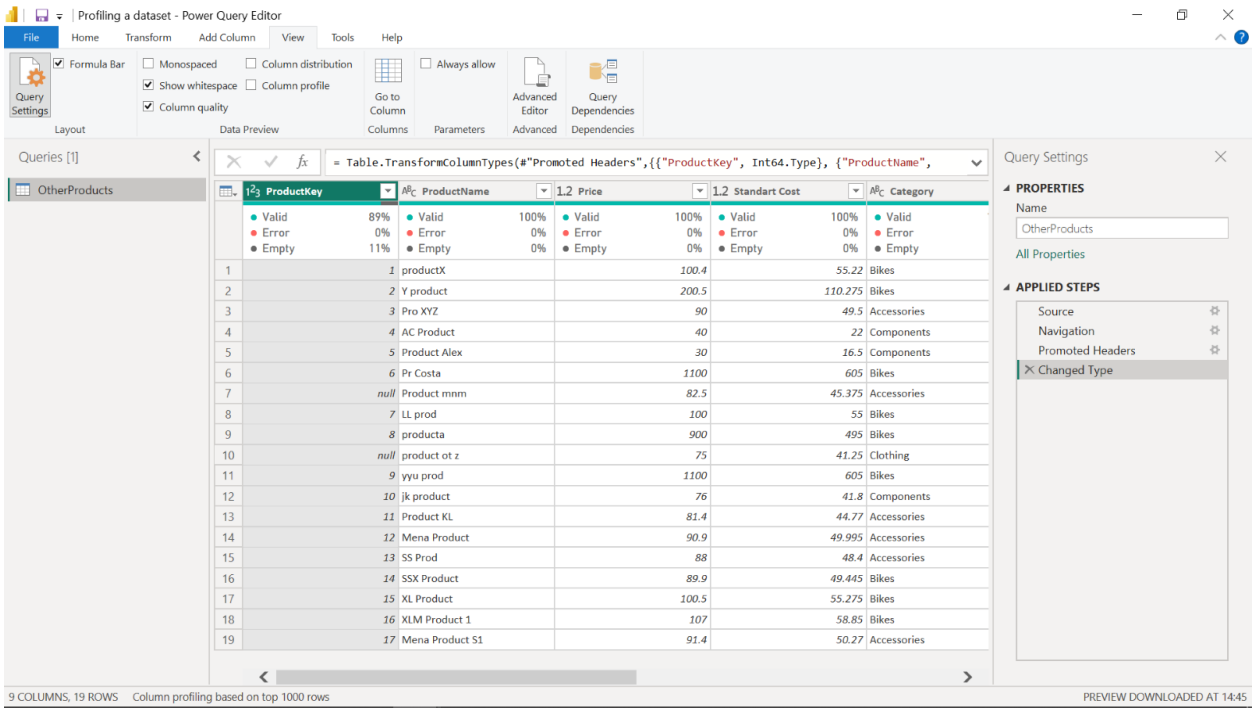
1. Use file: **Other Company Products.xlsx** from your Files Folder
2. Create new project and save it as **Profiling Company Products.pbix.**
3. Navigate to the **Home** ribbon tab at the top of the **Power BI** window.
4. Select the **Excel Workbook** button in the **Data** group, in the middle of the toolbar.
5. Select *Other Company Products.xlsx* and select **Transform Data** in the opened window.
6. The **Power Query Editor** window opens. You can now begin profiling the data.

Detect empty values in ProductKey column

1. There are empty values in **ProductKey** column.
2. To detect empty and invalid values, you need to assess column quality, on the **View** ribbon tab, from inside the **Data Preview** group, and select the **Column Quality** checkbox. **The column quality feature allows you to easily** determine the percentage of valid, error, or empty values found in columns.

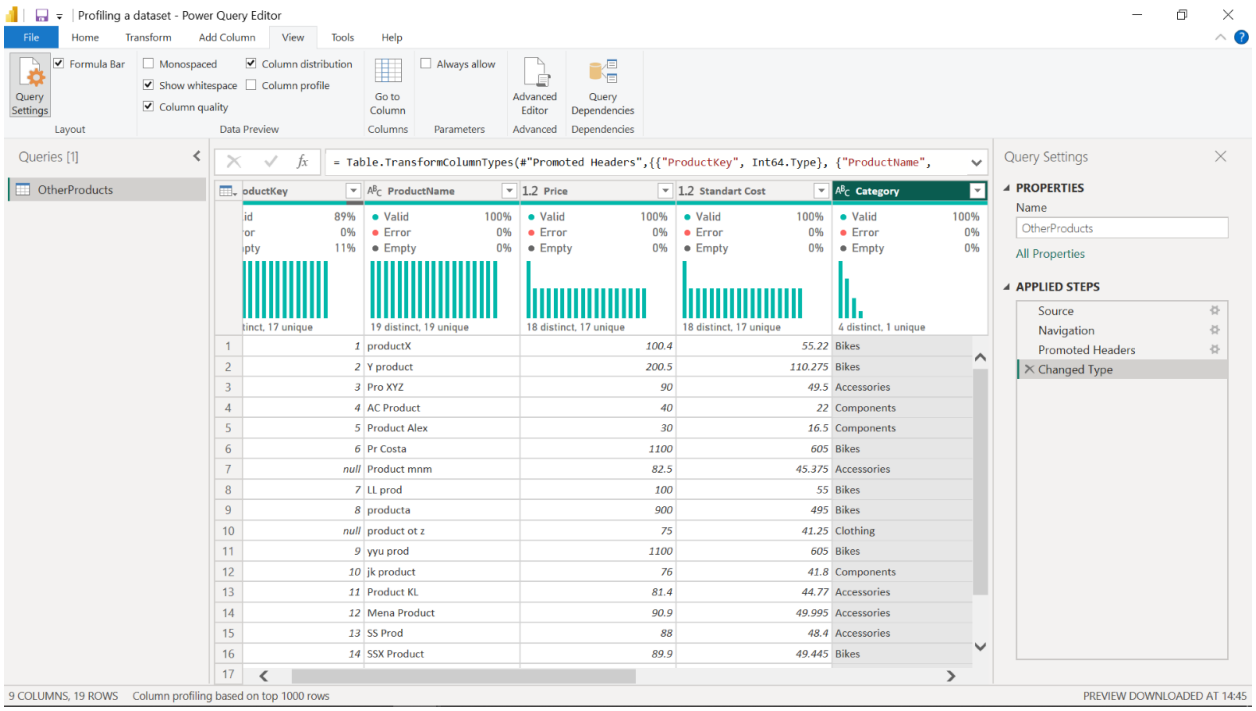


1. Note that 89% of the values are **Valid**, 0% of the values are **Error** and 11% of the values are **Empty** rows in the **ProductKey** column.

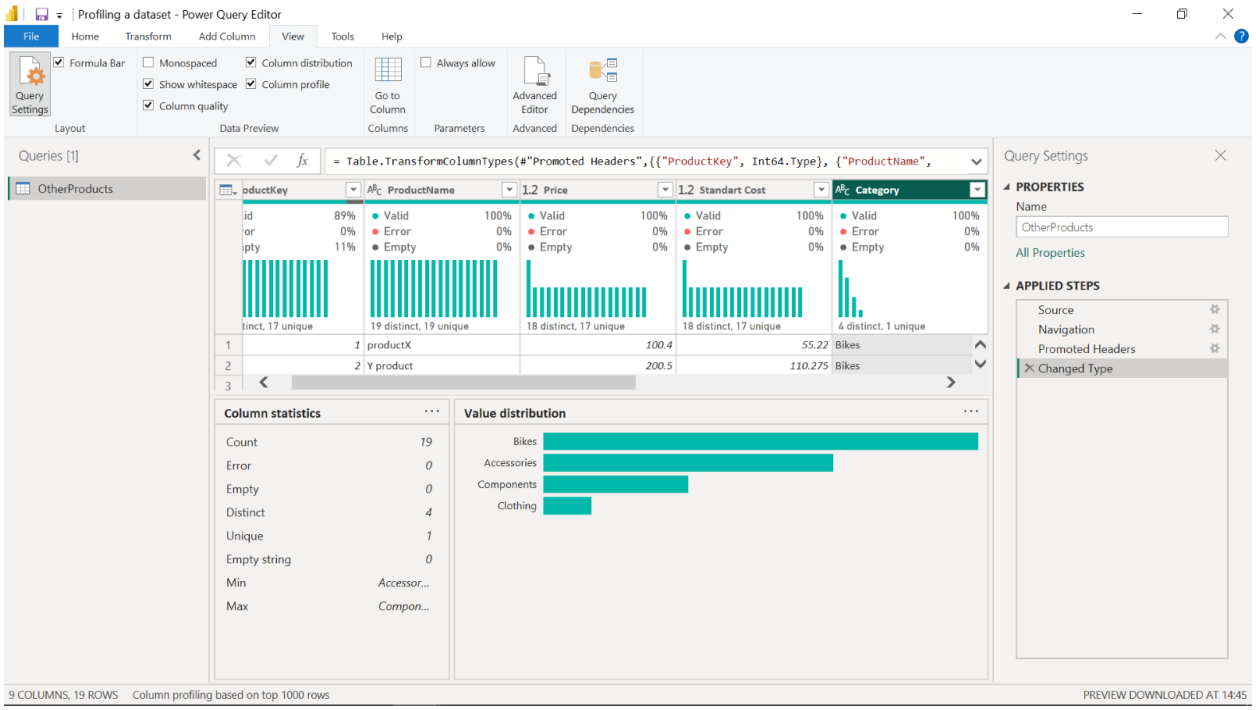


Assess the distribution of product categories

1. There are many categories in the **Product** list, and you need to find out how the data is distributed by the category data.
2. To assess column distribution, on the **View** ribbon tab, from inside the **Data Preview** group, check the **Column Distribution** checkbox. Note that there are 4 **distinct** values and 1 **unique** value.

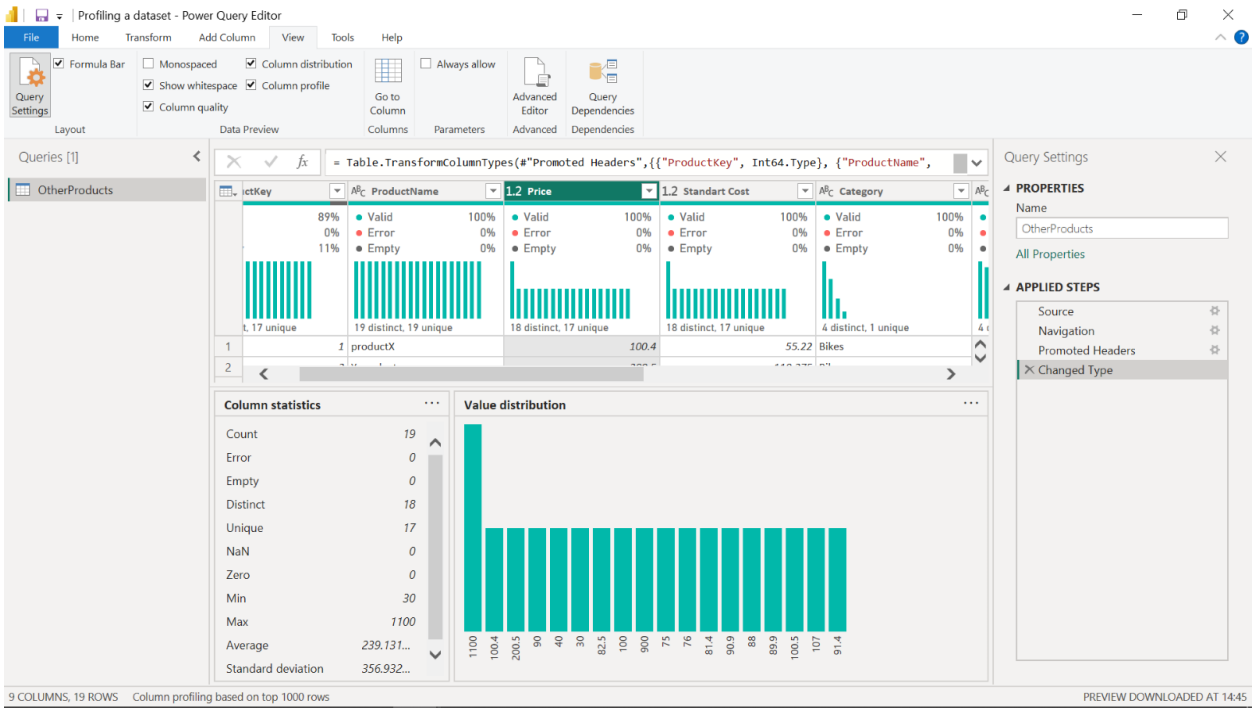


1. Check the **Column Profile** checkbox, while keeping **Column Distribution** checkbox as checked. Note that there are 9 Bikes, 6 Accessories, 3 Components and 1 Clothing categories.



Detect potential anomalies in the price column

1. You have to assess the **Price** column in **Product** list and you need to find out min, max, mean values and the distribution of the values. To detect potential anomalies and assess column distribution for the **Price** column, on the **View** ribbon tab, from inside the **Data Preview** group, check **Column Profile** while keeping **Column Distribution** checkbox as checked.
2. Note that min value is 30, max value is 1100, and average is 239.13 for the **Price** column.
3. There are 18 distinct and 17 unique values, which means there are 2 products with the same price.
4. When you assess the value distribution, it can be considered as normal distribution and there are some outliers like 30, 40, 900 and 1100. The aim of this assessment is to find a potential anomaly in the **Price** column.



Knowledge Check

Question 1

Which of the following Power Query menu items provides the user with column information like the number of empty and distinct rows and rows with errors?

A. Column Quality
B. Column Profile
C. Column Distribution

Question 2

Which of the following is defined as a data point that significantly deviates from other observations?

A. Standard Deviation
B. Outlier
C. Anomaly

Question 3

True or False: Distinct is known as "total number of different values", regardless of how many of each we have. Unique is known as "total number of values that only appear once". In this case, for fields with Primary Key or Unique Constraint defined, the values of Unique and Distinct will be equal.

A. True
B. False

Question 4
You import an Excel table named EmployeeData2023 into Power Query. You removed all other columns except Country.

The Country column has the following 10 row values:

- USA
- France
- France
- Ireland
- England
- England
- USA
- USA
- Spain
- France

What are the unique and distinct values of this column?

A. 1 unique and 5 distinct
B. 3 unique and 7 distinct
C. 2 unique and 8 distinct
D. 2 unique and 5 distinct

Question 5
You need to identify if data in a column contains Empty values. Which of the following can be used to quickly identify this? Select all that apply.

A. Column Distribution
B. Column Profile
C. Column Quality

Chapter 8: Practice and Final Project

Practice 1: Merge GDP table with Population Data

1. We want to import population data for each country.
2. We want to merge this data with that data of GDP in one table.
3. The data is in a CSV file called: **PopulationPerCountry.csv** in your Files.
4. Follow the steps below:
    a. Connect to the data source.
    b. Rename the new table **Population**.
    c. Remove the top rows and use first row as headers.
    d. unpivot columns.
    e. remove the two columns we do not need "**Indicator Name"** and "**Indicator Code**".
    f. rename columns to "**Population**", "**Year**" and "**Country Name**".
    g. Change the data type if necessary.
    h. Merge the new table "**Population**" with "**GDP by Country and Year**" Via Columns "**Year**" and "**Country Name**".

i. Better to use **Right Outer Joint**? Why?
j. Rename your **Merge2** table to "**Country Fact Table**".

5. You will find the solution of the practice in the Lab Solution Folder.

Practice 2: Adding Dimension Table

1. We want to add more information about each country.
2. This type of table is called Dimension Table.
3. We will discuss that in the upcoming course.
4. Explore the file: **MetaData Country.csv** in your File table.
5. It contains information about each country.
6. Load the table into Power Query.
7. Use First Row as a header.
8. Remove the last empty column.

Final Project

1. The files you will use in the final project are in the **Final Project** Subfolder under Files Folder**.**
2. Use file **Project.pbix** file and do the following tasks:
3.

Tasks

1. Create a new Power BI file.
2. Connect the 3 csv source files:

List of Orders

1. Use first Row as Headers
2. Try to change the data type of the column "Order Date" to Date - see what the problem is and remove this applied step again.
3. Create a column from examples to format the Order Date column in the structure "YYYY-MM-DD" – for example 2018-04-01.
4. Now change the data type of that new column again to Date.
5. Remove the Order Date column and then rename your Custom column to "Order Date".
6. Create another column from Examples to extract the weekday from your newly created "Order Date" column.

Order Details

1. Use first Row as Headers
2. Change the data type of the columns accordingly.

Sales Target

1. Use First Rows as Headers
2. Change the data type of the "Target" column accordingly.

Final Project Solution

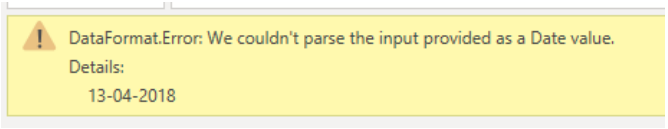
1. Connect to the 3 CSV files (**List of Orders, Order Details, Sales target**).
2. Open Power Query Editor.

List of Orders

3. Select List of Orders table.
4. Use First row as headers.

Change Order Date Data Type

5. Look at Order Date Column.
6. Notice that the date format is written as (DD/MM/YYYY).
7. While our format is American (YYYY/MM/DD).
8. Try to change type to date ➔You receive an error.

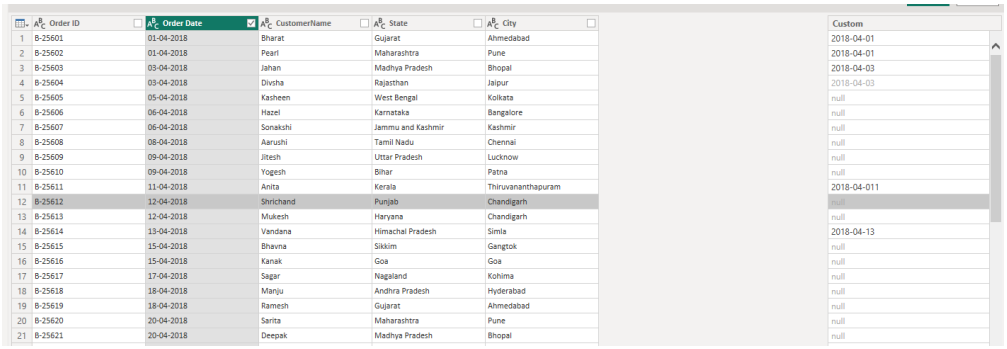


9. Click on the error to see.
10. Delete the step.
11. Cancel the last two steps and let us find a solution.
12. We can do a column by Example.
13. Select **Order Date** column.
14. Add Column➔General group➔Column from Example➔Clomn from Selection.
15. In the first column write the date in the format you want (**2018-04-01**) and press enter.
16. Nothing happened.
17. So, till now Power BI do not understand our pattern.
18. Keep going in second cell write (**2018-04-01**) and press enter.
19. Power BI now understands the pattern, but there is still error in the result.
20. In 3rd cell write (**2018-04-03**) and press enter.
21. Notice it goes well till it reaches line it added the day as 011

	Order ID	Order Date	CustomerName	State	City	Custom
1	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	2018-04-01
2	B-25602	01-04-2018	Pearl	Maharashtra	Pune	2018-04-01
3	B-25603	03-04-2018	Jahan	Madhya Pradesh	Bhopal	2018-04-03
4	B-25604	03-04-2018	Divsha	Rajasthan	Jaipur	2018-04-03
5	B-25605	05-04-2018	Kasheen	West Bengal	Kolkata	2018-04-05
6	B-25606	06-04-2018	Hazel	Karnataka	Bangalore	2018-04-06
7	B-25607	06-04-2018	Sonakshi	Jammu and Kashmir	Kashmir	2018-04-06
8	B-25608	08-04-2018	Aarushi	Tamil Nadu	Chennai	2018-04-08
9	B-25609	09-04-2018	Jitesh	Uttar Pradesh	Lucknow	2018-04-09
10	B-25610	09-04-2018	Yogesh	Bihar	Patna	2018-04-09
11	B-25611	11-04-2018	Anita	Kerala	Thiruvananthapuram	2018-04-011
12	B-25612	12-04-2018	Shrichand	Punjab	Chandigarh	2018-04-012
13	B-25613	12-04-2018	Mukesh	Haryana	Chandigarh	2018-04-012
14	B-25614	13-04-2018	Vandana	Himachal Pradesh	Simla	
15	B-25615	15-04-2018	Bhavna	Sikkim	Gangtok	

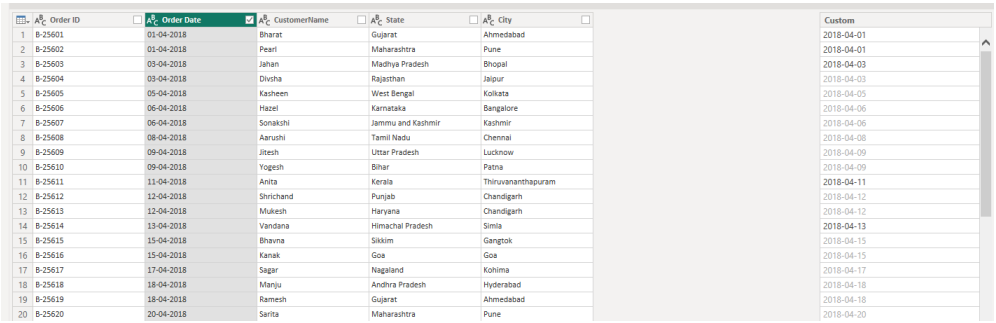
22. In line 14 try to enter the value again (2018-04-13) and press enter.

23. You got null as the pattern of a day 011 is not understood.



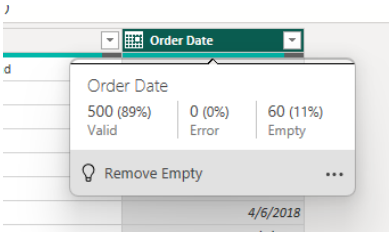
	Order ID	Order Date	CustomerName	State	City	Custom
1	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	2018-04-01
2	B-25602	01-04-2018	Pearl	Maharashtra	Pune	2018-04-01
3	B-25603	03-04-2018	Jahan	Madhya Pradesh	Bhopal	2018-04-03
4	B-25604	03-04-2018	Divsha	Rajasthan	Jaipur	2018-04-03
5	B-25605	05-04-2018	Kasheen	West Bengal	Kolkata	null
6	B-25606	06-04-2018	Hazel	Karnataka	Bangalore	null
7	B-25607	06-04-2018	Sonakshi	Jammu and Kashmir	Kashmir	null
8	B-25608	08-04-2018	Aarushi	Tamil Nadu	Chennai	null
9	B-25609	09-04-2018	Jitesh	Uttar Pradesh	Lucknow	null
10	B-25610	09-04-2018	Yogesh	Bihar	Patna	null
11	B-25611	11-04-2018	Anita	Kerala	Thiruvananthapuram	2018-04-011
12	B-25612	12-04-2018	Shrichand	Punjab	Chandigarh	null
13	B-25613	12-04-2018	Mukesh	Haryana	Chandigarh	null
14	B-25614	13-04-2018	Vandana	Himachal Pradesh	Simla	2018-04-13
15	B-25615	15-04-2018	Bhavna	Sikkim	Gangtok	null
16	B-25616	15-04-2018	Kanak	Goa	Goa	null
17	B-25617	17-04-2018	Sagar	Nagaland	Kohima	null
18	B-25618	18-04-2018	Manju	Andhra Pradesh	Hyderabad	null
19	B-25619	18-04-2018	Ramesh	Gujarat	Ahmedabad	null
20	B-25620	20-04-2018	Sarita	Maharashtra	Pune	null
21	B-25621	20-04-2018	Deepak	Madhya Pradesh	Bhopal	null

24. Go and correct 011 to be 11 only in the day in line 11 and press enter.

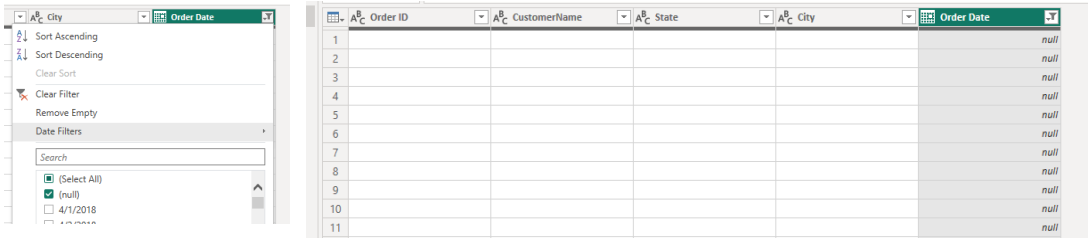


	Order ID	Order Date	CustomerName	State	City	Custom
1	B-25601	01-04-2018	Bharat	Gujarat	Ahmedabad	2018-04-01
2	B-25602	01-04-2018	Pearl	Maharashtra	Pune	2018-04-01
3	B-25603	03-04-2018	Jahan	Madhya Pradesh	Bhopal	2018-04-03
4	B-25604	03-04-2018	Divsha	Rajasthan	Jaipur	2018-04-03
5	B-25605	05-04-2018	Kasheen	West Bengal	Kolkata	2018-04-05
6	B-25606	06-04-2018	Hazel	Karnataka	Bangalore	2018-04-06
7	B-25607	06-04-2018	Sonakshi	Jammu and Kashmir	Kashmir	2018-04-06
8	B-25608	08-04-2018	Aarushi	Tamil Nadu	Chennai	2018-04-08
9	B-25609	09-04-2018	Jitesh	Uttar Pradesh	Lucknow	2018-04-09
10	B-25610	09-04-2018	Yogesh	Bihar	Patna	2018-04-09
11	B-25611	11-04-2018	Anita	Kerala	Thiruvananthapuram	2018-04-11
12	B-25612	12-04-2018	Shrichand	Punjab	Chandigarh	2018-04-12
13	B-25613	12-04-2018	Mukesh	Haryana	Chandigarh	2018-04-12
14	B-25614	13-04-2018	Vandana	Himachal Pradesh	Simla	2018-04-13
15	B-25615	15-04-2018	Bhavna	Sikkim	Gangtok	2018-04-15
16	B-25616	15-04-2018	Kanak	Goa	Goa	2018-04-15
17	B-25617	17-04-2018	Sagar	Nagaland	Kohima	2018-04-17
18	B-25618	18-04-2018	Manju	Andhra Pradesh	Hyderabad	2018-04-18
19	B-25619	18-04-2018	Ramesh	Gujarat	Ahmedabad	2018-04-18
20	B-25620	20-04-2018	Sarita	Maharashtra	Pune	2018-04-20

25. All data now is correct.
26. Click the OK button in the upper message to accept.
27. Now change the data type in the custom column into Date type, it changes with no error this time.
28. Remove the Original **Order Date** Column and rename the **Custom** column to **Order Date**.
29. Notice you have gray indicator in the column. When you hover, you find it is empty values.



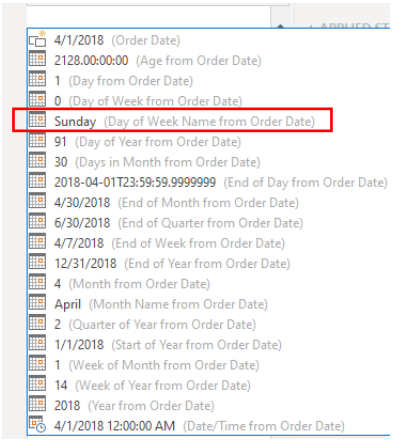
30. First Go and see what are those value are.
31. Click in the down arrow and select the null value.



32. They are all empty rows.
33. First delete the filtering step.
34. Then click the down arrow again and select **Remove Empty**.

Add New Column Weekday

35. Select Order Date column.
36. Use column by example from section.
37. Try to type in 1st cell you will find the Power Bi is so intelligent and give you Sunday as an option as it detects it is a date



38. Select Sunday and press enter.
39. Click OK to accept.
40. You now have the new column.
41. You can also achieve the same result by:
    - First Select Order Date column.
    - Add Column ➔From Date & Time group ➔ Date➔ Day➔Name of Day

Order details

1. Use First Row as a header.
2. Change **Account** and **Profit** Columns type to **Whole Number**.

Sales Target

3. Use First row as headers.
4. Change **Target** Column type to **whole number**.