

# DS342 - Data Analytics

# Lecture 6

## Business Intelligence (BI)

### Tools for Data Analysis



# Chapter Contents

- ▶ 4.1 Introduction
- ▶ 4.2 Importing Data into Excel with Power Query
- ▶ 4.3 Data analysis with Power Pivot
- ▶ 4.5 Data Cleansing

# 4-1 Introduction (slide 1 of 2)

- ▶ Data analysis has become more important in the business world.
- ▶ Microsoft and other software companies have created better tools for analyzing data, also known as creating **business intelligence (BI)**.

## 4-1 Introduction (slide 2 of 2)

- ▶ Microsoft add-ins for Excel: Power BI suite
  - Power Query
  - Power Pivot
  - Power View
  - Power Map
- ▶ These are known as **self-service BI** add-ins, meaning people who need to analyze data for their jobs no longer need to rely on IT departments to get the results they need right away.

# 4-2 Importing Data into Excel with Power Query

(slide 1 of 7)

- ▶ Data sets to be analyzed are often stored somewhere else, *not* in Excel workbooks.
  - You must learn methods for importing data into Excel.
  - The most recent tools are better and easier to use than ever.
  - These tools are often referred to as **Power Query**.

# 4-2a Introduction to Relational Databases

(slide 2 of 7)

- ▶ Data sets discussed in previous chapters were **single-table databases**, also known as **flat files**, where all the data are held in a single rectangular table.
- ▶ More generally, a database is a set of related tables, where each table is a data set of the type discussed in the previous chapters.
- ▶ Because the tables are related, this type of database is called a **relational database**.

# 4-2a Introduction to Relational Databases

(slide 3 of 7)

- ▶ Two questions:
  - Why are data often stored in multiple tables rather than in a single large table?
  - How are the tables related?

# 4-2a Introduction to Relational Databases

(slide 4 of 7)

- ▶ If all of the data was stored in a single large table, this would inflate the size of the database and also introduces the possibility of errors.
- ▶ Because every individual is given a unique line, if one individual's information changes, like their address, this change must be made everywhere that person exists in the database.
  - The solution is to split the database into separate tables that are related through the use of keys.

# 4-2a Introduction to Relational Databases

(slide 5 of 7)

- ▶ Until recently, all multi-table relational databases were stored in a **relational database management system (RDBMS)**.
  - This is a software that creates and manages databases.
  - Example: Microsoft Access, which is a desktop system where your database files can reside locally on your hard drive.
- ▶ Data are usually stored in a RDBMS, but data analysts typically want to perform their analyses in Excel.
  - Technologies have been developed to import data from a RDBMS into Excel.
  - These technologies involve one or more queries, where a **query** specifies exactly which data you want to import.

# 4-2a Introduction to Relational Databases

(slide 6 of 7)

- ▶ Queries are so common that a language for queries was developed about 30 years ago.
  - This language is called **Structured Query Language (SQL)**.
  - The power of SQL is that, with minor modifications, it can be used to query *any* RDBMS.
  - However, it presents a hurdle for the many Excel users who do not know SQL, so Microsoft and other companies have been trying to make querying easier by creating more transparent tools. (**Power Query**)

# 4-2a Introduction to Relational Databases

(slide 7 of 7)

- ▶ It has always been possible to create multi-table relational databases directly in Excel.
  - Each table is a single-table database.
  - To relate them, you can use VLOOKUP functions.
  - This VLOOKUP method is frequently used, but it is cumbersome.
  - Also, tables might be too large to fit into Excel's (approximately) one-million row limit.

## 4-2b Excel's Data Model (slide 1 of 3)

- ▶ When you run a query to import data into Excel, you must decide where you want to store the data.
  - Until recently, there were 2 basic options:
    - Into an Excel table
    - Into a pivot table (or pivot chart) report
  - The first option is often a good one unless the size of the imported data is larger than the size of an Excel worksheet.
  - The second option makes sense if your eventual goal is to analyze the data with pivot tables or pivot charts.

## 4-2b Excel's Data Model (slide 2 of 3)

- ▶ There is now a third option, a **Data Model** in Excel.
  - A Data Model essentially mimics the behavior of a RDBMS, but all within Excel.
  - A Data Model stores related tables in a special type of compressed format
    - **not in an Excel worksheet** – for use in creating pivot table reports.
      - This means that you can store all your data, even millions of rows of data, inside an Excel workbook and then use Excel's powerful tools to analyze the data.
      - No VLOOKUPs are necessary, and once the data are in Excel, there is no further need for Access or any other RDBMS.

## **4-2b Excel's Data Model** (slide 3 of 3)

- ▶ Before Excel's relatively new Data Model, relational data had to be imported from Access or other relational database packages into Excel for analysis.
- ▶ Now relational databases can be stored in Excel directly, where they can be analyzed.
- ▶ In short, a Data Model is a relational database that is stored into Excel.

# Example 4.1: Importing Related Database



## Tables (slide 1 of 17)

- ▶ **Objective:** To use PowerQuery to import the Contoso tables and the Stores and Geography files into a Data Model in Excel.
- ▶ **Solution:** The following steps can be used to obtain the results in the **Contoso Import Finished.xlsx** file.

# Example 4.1: Importing Related Database



## Tables (slide 2 of 17)

Solution:

1. Open a new workbook in Excel. If the Developer tab isn't visible, right-click any ribbon, select Customize the Ribbon, and check the developer item in the right pane of the resulting dialog box. The Developer ribbon has useful tools for programmers, but it also provides an easy way to get to Excel's add-ins lists.

# Example 4.1: Importing Related Database Tables (slide 3 of 17)



Solution:

2. Click COM Add-Ins on the Developer ribbon and check the Microsoft Power Pivot for Excel item. This creates a Power Pivot tab and associated ribbon. You will need Power Pivot to “see” the imported data in the following steps.

# Example 4.1: Importing Related Database Tables (slide 4 of 17)



Solution:

3. On the Data ribbon, click Get Data, then From Database, and then From Microsoft Access Database. Maneuver to the **Contoso.accdb** file and click Import.

# Example 4.1: Importing Related Database Tables (slide 5 of 17)



Solution:

4. This opens the Navigator window. Fill it in as shown, checking the Select multiple items option and checking all five tables. When any table is selected, you see a preview of its data in the right pane.

# Example 4.1: Importing Related Database Tables (slide 6 of 17)

The screenshot shows the Microsoft Access Navigator dialog box. On the left, there's a tree view of tables in the 'ContosoSales.accdb' database. The 'Sales' table is selected and highlighted with a green border. The main area displays the 'Sales' table with the following data:

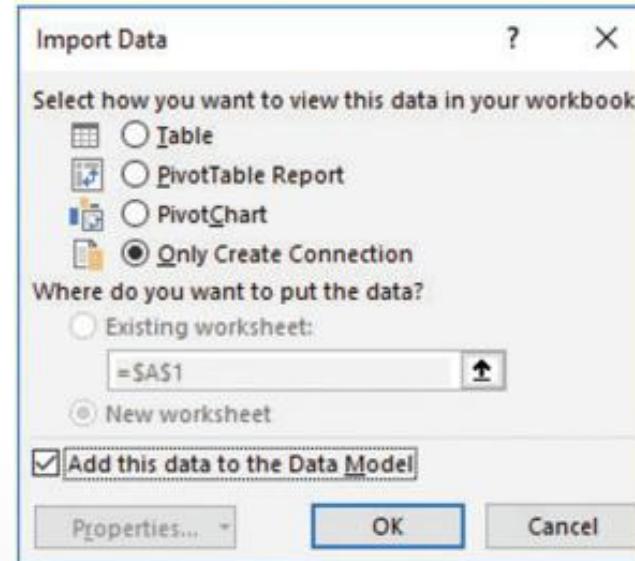
SalesKey	DateKey	StoreKey	ProductKey	ReturnAmount
7578	10/15/2008 12:00:00 AM	143	919	
7580	2/9/2008 12:00:00 AM	5	585	
7582	1/2/2009 12:00:00 AM	307	840	
7584	11/28/2007 12:00:00 AM	199	1480	
7588	6/1/2007 12:00:00 AM	200	452	
7589	6/27/2007 12:00:00 AM	200	325	
7590	11/2/2007 12:00:00 AM	281	511	
7591	5/23/2008 12:00:00 AM	139	1648	
7595	4/6/2009 12:00:00 AM	85	1352	
7596	2/18/2008 12:00:00 AM	200	589	
7597	5/21/2007 12:00:00 AM	310	1237	
7599	6/20/2009 12:00:00 AM	69	1374	
7601	12/19/2007 12:00:00 AM	309	357	
7605	9/25/2008 12:00:00 AM	307	1024	
7606	1/6/2007 12:00:00 AM	115	603	
7607	3/20/2008 12:00:00 AM	146	1532	
7608	3/21/2008 12:00:00 AM	261	189	
7609	4/4/2007 12:00:00 AM	310	712	
7610	4/18/2007 12:00:00 AM	306	899	
7613	3/18/2007 12:00:00 AM	180	1334	
7614	1/29/2008 12:00:00 AM	45	1455	
7616	6/26/2007 12:00:00 AM	75	842	
7617	6/28/2007 12:00:00 AM	119	952	

At the bottom of the dialog box, there are buttons for 'Select Related Tables', 'Load', 'Edit', and 'Cancel'.

# Example 4.1: Importing Related Database Tables (slide 7 of 17)

Solution:

- From the Load dropdown list, select Load To. This opens the Import Data dialog box. Fill it out as shown, selecting Only Create Connection and checking the bottom Data Model option, and click OK.



# Example 4.1: Importing Related Database Tables (slide 8 of 17)



Solution:

6. A Query & Connections pane with five queries, one for each table, will open in Excel, but when the data have finished downloading, you won't see any data. Read on.

# Example 4.1: Importing Related Database Tables (slide 9 of 17)

Solution:

- Because you have filled out the Import Data dialog box, the data don't appear in an Excel table or pivot table. Instead they are stored in a Data Model. To see the Data Model, select Manage from the Power Pivot ribbon. This opens the Power Pivot window, where you can see the imported data. This looks like an Excel workbook, with a tab for each of the five tables, but it is quite different.

# Example 4.1: Importing Related Database Tables (slide 10 of 17)

The screenshot shows the Power Pivot for Excel interface. The ribbon at the top has tabs for File, Home, Design, and Advanced. The Design tab is selected, showing icons for Add, Insert Function, Create Relationship, Manage Relationships, Table Properties, Mark as Date Table, Date Table, and Undo/Redo. Below the ribbon is a data grid with the following columns: SalesKey, DateKey, StoreKey, ProductKey, ReturnAmount, TotalCost, SalesAmount, and Add Column. The data grid contains 12 rows of sales data. At the bottom of the grid, there are tabs for Sales, ProductSubcategories, Dates, Products, and ProductCategories, with Sales being the active tab. The status bar at the bottom shows "Record: 1 of 33,565".

	SalesKey	DateKey	StoreKey	ProductKey	ReturnAmount	TotalCost	SalesAmount	Add Column
1	7588	6/1/2007 1...	200	452	\$0.00	\$1,682.10	\$3,299.25	
2	7589	6/27/2007 ...	200	325	\$0.00	\$2,254.70	\$4,990.00	
3	7648	1/15/2007 ...	200	167	\$0.00	\$562.88	\$1,086.75	
4	7652	4/27/2007 ...	200	385	\$0.00	\$831.00	\$1,630.00	
5	7653	8/20/2008 ...	200	826	\$0.00	\$103.44	\$202.80	
6	7673	6/12/2009 ...	200	777	\$0.00	\$141.80	\$278.00	
7	64	1/19/2009 ...	200	1400	\$0.00	\$446.85	\$969.93	
8	7079	11/1/2009 ...	200	221	\$0.00	\$3,580.98	\$7,667.20	
9	7087	8/13/2009 ...	200	504	\$0.00	\$1,727.52	\$4,866.40	
10	7089	12/29/2009 ...	200	1029	\$0.00	\$2,584.14	\$7,480.00	
11	7096	10/4/2009 ...	200	310	\$0.00	\$1,524.40	\$2,990.00	
12	7118	9/29/2008	200	367	\$0.00	\$3,988.80	\$7,661.00	

# Example 4.1: Importing Related Database Tables (slide 11 of 17)



Solution:

7. Did you notice the two buttons circled at the bottom right? Those are toggles between data and diagram views.

# Example 4.1: Importing Related Database



## Tables (slide 12 of 17)

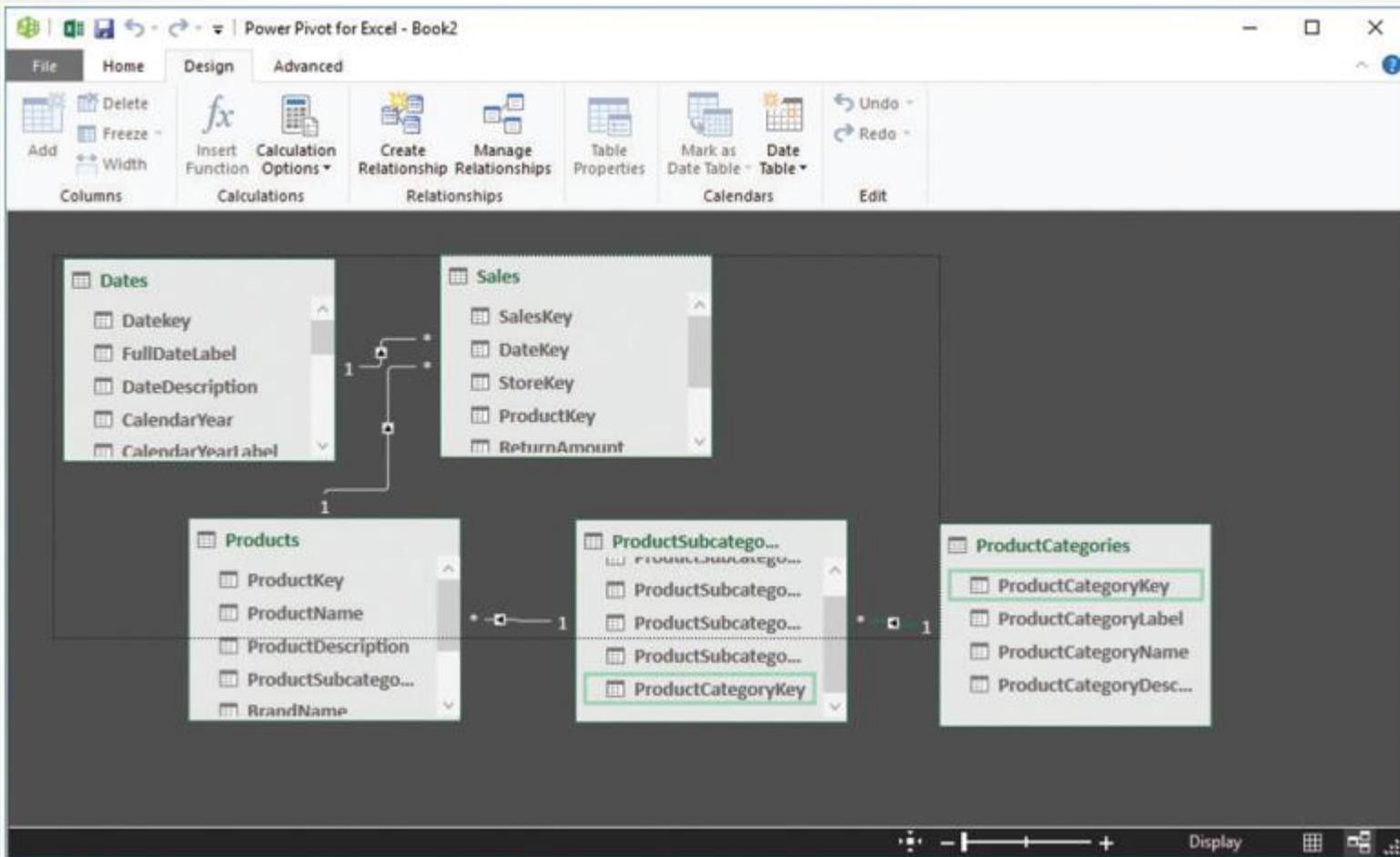
Solution:

8. Click the diagram button at the bottom right and then move/resize the tables appropriately to see the relationship diagram.

Although these tables are related in the Access source file through primary and foreign keys, they aren't automatically linked with the arrows shown in the figure. However, you can drag from the obvious key fields to create these arrows and hence relate the tables.

- ▶ For example, drag from DateKey in the Dates table to DateKey in the Sales table to relate these two tables.

# Example 4.1: Importing Related Database Tables (slide 13 of 17)



# Example 4.1: Importing Related Database



## Tables (slide 14 of 17)

Solution:

9. Back into the Excel window, click From Text/CSV on the Data ribbon, maneuver to the **Geography.csv** file, click Load To, and again fill out the dialog box. Then repeat with the Stores data. This time, select From File from the Get Data dropdown list, choose From Workbook, and maneuver to the **Sales.xlsx** files. Once again, fill out the dialog box.

# Example 4.1: Importing Related Database

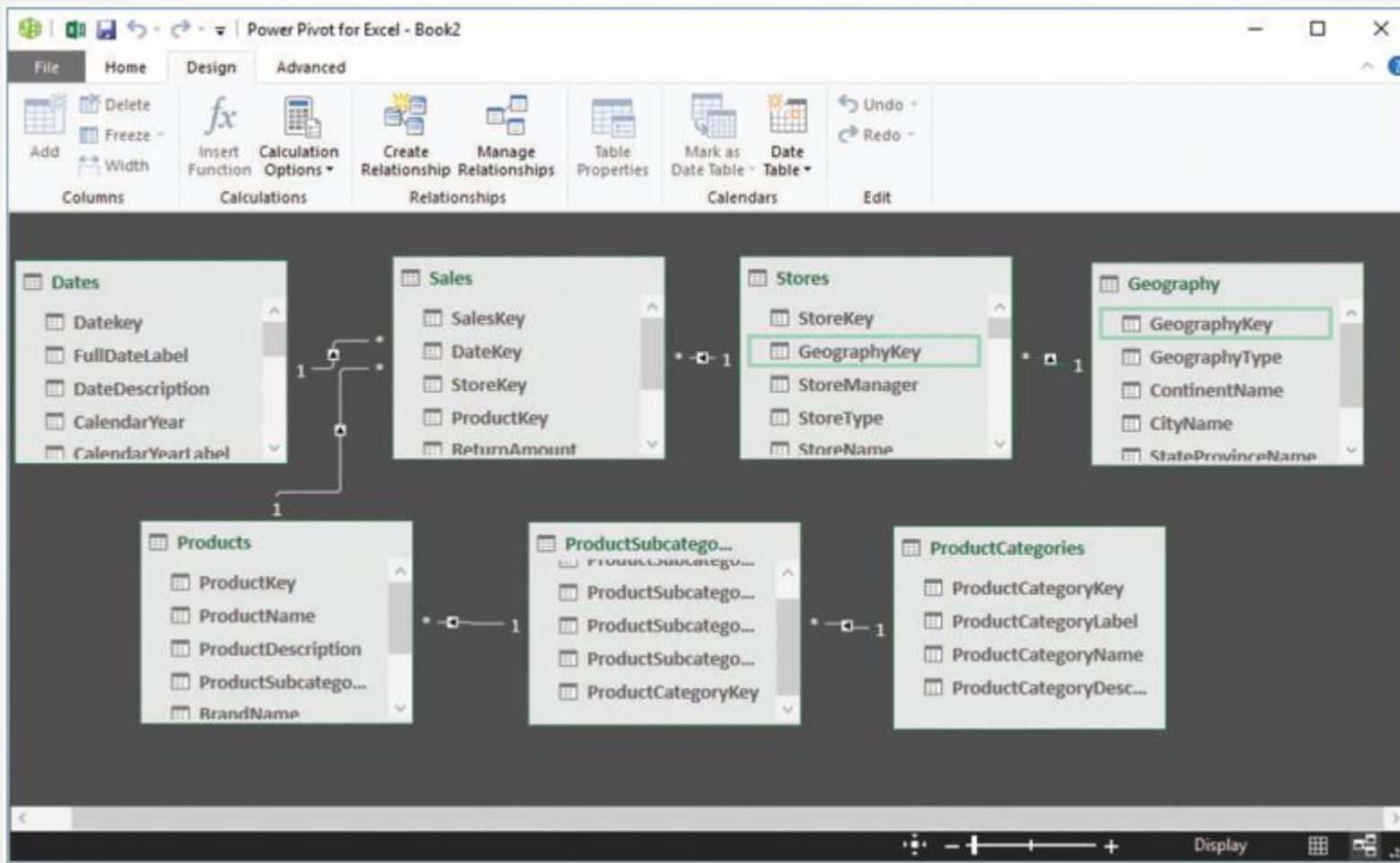


## Tables (slide 15 of 17)

Solution:

10. Open diagram view in the Power Pivot window. You will see the new Geography and Stores tables, but they won't be related to anything. To relate them, drag StoreKey in the Sales table to StoreKey in the Stores table, and then drag GeographyKey in the Stores table to GeographyKey in the Geography table.

# Example 4.1: Importing Related Database Tables (slide 16 of 17)



# Example 4.1: Importing Related Database



## Tables (slide 17 of 17)

### ▶ Date Table

- Most Data Models will include a separate Dates table, even if dates of transactions are already listed in another table.
- The purpose of a Dates table is to list all possible dates in many possible ways.
  - Year, Quarter, Month number, Month name, Day of the week, Whether it is a weekday, Whether it is a holiday, and so on.

# Example 4.2: Creating A Data Model with A Pivot Table (slide 1 of 11)



- ▶ **Objective:** To create a Data Model by creating a pivot table from the baseball data and, in the process, to relate the two tables.
- ▶ **Solution:** The following steps are required. (You can see the results in the **Baseball Salaries with Team Info1 Finished.xlsx** file.)

# Example 4.2: Creating A Data Model with A Pivot Table

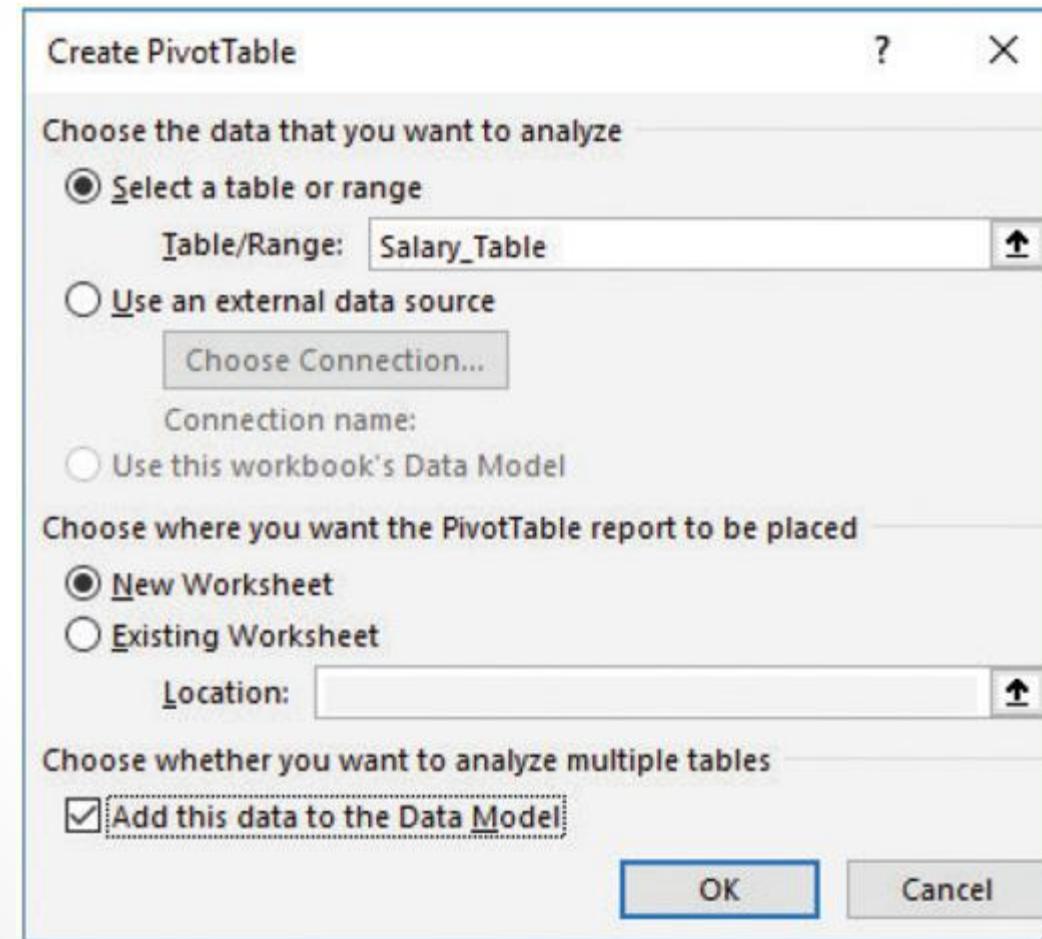
(slide 2 of 11)



Solution:

1. Open the **Baseball Salaries with Team Info1.xlsx** file.
2. Select any cell in the Salaries table; select PivotTable on the Insert ribbon; fill out the resulting dialog box, making sure to check the Data Model option near the bottom; and click OK.

# Example 4.2: Creating A Data Model with A Pivot Table (slide 3 of 11)



# Example 4.2: Creating A Data Model with A Pivot Table

(slide 4 of 11)



Solution:

3. As usual, a new sheet with a blank pivot table will open, but something else happens. If you open the Power Pivot window, you will see that the table from the Salaries sheet has been added to a new Data Model.
4. Back in the pivot table sheet in Excel, drag Salary to the Values area.

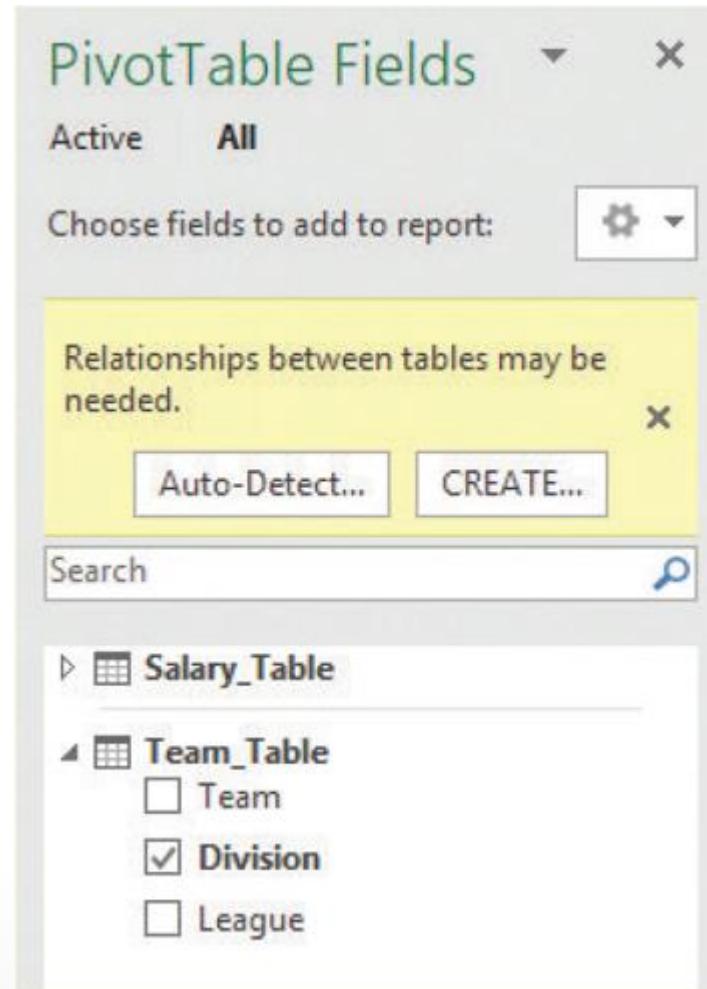
# Example 4.2: Creating A Data Model with A Pivot Table (slide 5 of 11)



Solution:

5. At this point, the “Active” link at the top of the PivotTable Fields pane is active (bold-faced), and only the fields in the Salary\_Table are listed. But now click the “All” link. All fields in both tables are then listed.

# Example 4.2: Creating A Data Model with A Pivot Table (slide 6 of 11)



# Example 4.2: Creating A Data Model with A Pivot Table (slide 7 of 11)



Solution:

6. Expand the Team\_Table list and check the Division field, the goal being to find the sum of salaries in each of the six divisions. However, because the two tables are not yet related, you will see a warning in yellow. Also, the pivot table values will be wrong because Excel is unable to combine data from two tables that aren't related.

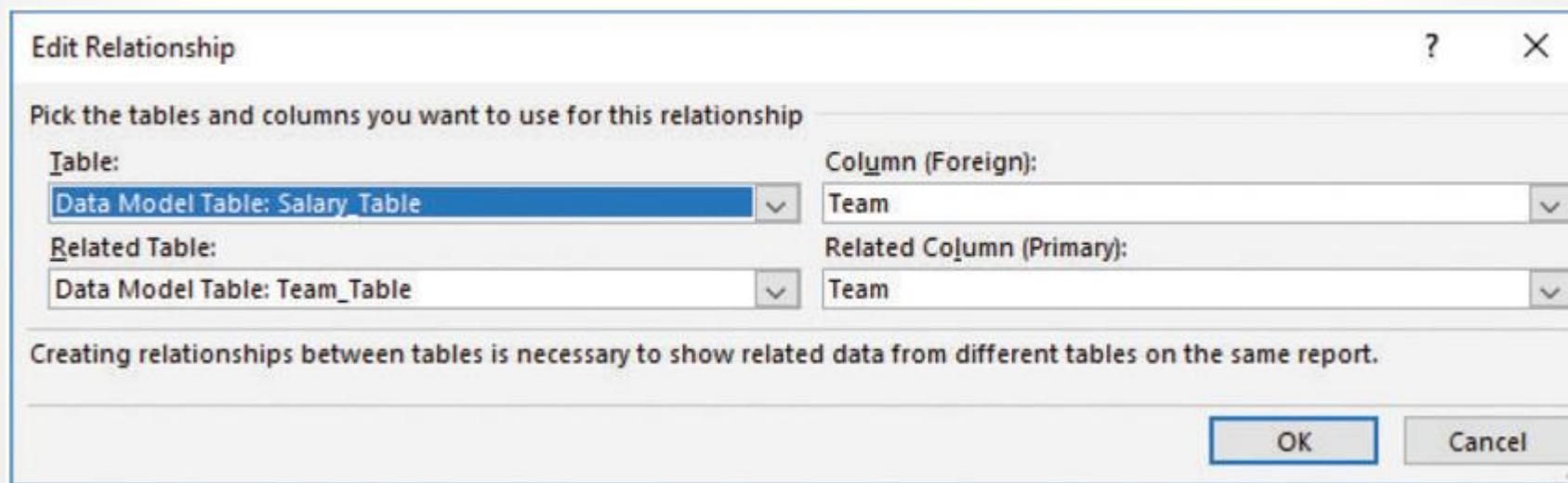
# Example 4.2: Creating A Data Model with A Pivot Table (slide 8 of 11)



Solution:

7. To fix the problem, click the Auto-Detect button. In this case, Excel has the intelligence to recognize that the Team fields in the two tables serve as key fields, so it uses them to relate the tables. Nevertheless, it lets you examine the relationship.  
To do this, click the Manage Relationships button on the resulting dialog box and then click Edit to see the dialog box.

# Example 4.2: Creating A Data Model with A Pivot Table (slide 9 of 11)



# Example 4.2: Creating A Data Model with A Pivot Table

(slide 10 of 11)



Solution:

8. To see what else has happened, open the Power Pivot window again. In diagram view, you will see both tables, linked by their key fields. That is, both tables are now part of the workbook's Data Model. The only difference between this example and the previous example is that the Data Model has now been created through pivot table operations, not through importing external data.

# Example 4.2: Creating A Data Model with A Pivot Table (slide 11 of 11)



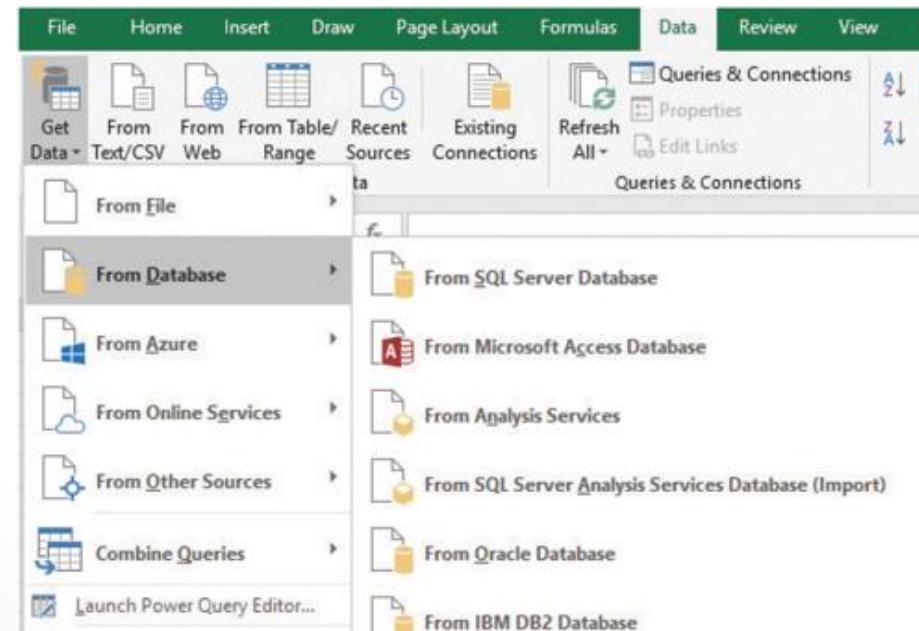
Solution:

9. Want a surprise? To be safe, save your work. Then delete the Salaries and Team sheets. Now manipulate the pivot table any way you like. The pivot table will still work fine! This is because the pivot table's source is the Data Model, and the original data sheets are no longer necessary.

# 4-2c Creating and Editing Queries (slide 1 of 2)

## ▶ Power Query

- The tools for querying with Power Query are in the Get & Transform Data group on the data ribbon.



# 4-2c Creating and Editing Queries (slide 2 of 2)

- ▶ Three important strengths of Power Query
  - You can import data from more sources than ever before.
  - Once you locate the data of interest, you can use the new Query Editor to “shape” the data to be imported.
  - Your queries are stored in step-by-step fashion so that they can be modified and reused later.

# Example 4.1: Importing Related Database Tables (Continued) (slide 1 of 2)



- ▶ **Objective:** To learn how to “shape” a query with Power Query.
- ▶ **Solution:** If you select From Microsoft Access Database and then maneuver to **Contoso.accdb** file, you get the Navigator window, where you can select the tables of interest. For now, select all 5 tables as before. However, instead of loading the data right away, click the Edit button. This opens the Query Editor. This lets you preview the data, but more importantly, it lets you “shape” the data to be imported.

# Example 4.1: Importing Related Database Tables (Continued) (slide 2 of 2)

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with columns: ProductKey, ProductName, and ProductDescription. The table contains 22 rows of data, mostly for Contoso MP3 players. The 'Applied Steps' pane on the right shows a single step named 'Navigation'. The status bar at the bottom indicates '9 COLUMNS, 999+ ROWS' and 'PREVIEW DOWNLOADED AT 2:45 PM'.

ProductKey	ProductName	ProductDescription
1	Contoso 512MB MP3 Player E31 Silver	512MB USB driver plays MP3 and WMA
2	Contoso 512MB MP3 Player E31 Blue	512MB USB driver plays MP3 and WMA
3	Contoso 1G MP3 Player E100 White	1GB flash memory and USB driver plays MP3 and WMA
4	Contoso 2G MP3 Player E200 Silver	2GB flash memory, LCD display, plays MP3 and WMA
5	Contoso 2G MP3 Player E200 Red	2GB flash memory, LCD display, plays MP3 and WMA
6	Contoso 2G MP3 Player E200 Black	2GB flash memory, LCD display, plays MP3 and WMA
7	Contoso 2G MP3 Player E200 Blue	2GB flash memory, LCD display, plays MP3 and WMA
8	Contoso 4G MP3 Player E400 Silver	4GB flash memory and FM Radio, LCD Display with 7-Color Backlight, o...
9	Contoso 4G MP3 Player E400 Black	4GB flash memory and FM Radio, LCD Display with 7-Color Backlight, p...
10	Contoso 4G MP3 Player E400 Green	4GB flash memory and FM Radio, LCD Display with 7-Color Backlight, p...
11	Contoso 4G MP3 Player E400 Orange	4GB flash memory and FM Radio, LCD Display with 7-Color Backlight, o...
12	Contoso 4GB Flash MP3 Player E401 Blue	1.8" color LCD, play MP3, WMA and Video MTV, and share JPG
13	Contoso 4GB Flash MP3 Player E401 Black	1.8" color LCD, play MP3, WMA and Video MTV, and share JPG
14	Contoso 4GB Flash MP3 Player E401 Silver	1.8" color LCD, play MP3, WMA and Video MTV, and share JPG
15	Contoso 4GB Flash MP3 Player E401 White	1.8" color LCD, play MP3, WMA and Video MTV, and share JPG
16	Contoso 8GB Super-Slim MP3/Video Player M800 White	2" color LCD, Touchpad, Plays music, video, photos and text
17	Contoso 8GB Super-Slim MP3/Video Player M800 Red	2" color LCD, Touchpad, Plays music, video, photos and text
18	Contoso 8GB Super-Slim MP3/Video Player M800 Green	2" color LCD, Touchpad, Plays music, video, photos and text
19	Contoso 8GB Super-Slim MP3/Video Player M800 Pink	2" color LCD, Touchpad, Plays music, video, photos and text
20	Contoso 8GB MP3 Player new model M820 Black	2" LCD with blue-white LED, 320x240-pixel, plays music, video, photos...
21	Contoso 8GB MP3 Player new model M820 Blue	2" LCD with blue-white LED, 320x240-pixel, plays music, video, photos...
22	Contoso 8GB MP3 Player new model M820 Yellow	2" LCD with blue-white LED, 320x240-pixel, plays music, video, photos...

## 4-3 Data Analysis with Power Pivot

- ▶ This section discusses the most publicized member of Microsoft's Power BI tools, **Power Pivot**.
  - The Data Model and Power Pivot work hand in hand for powerful data analysis.
  - Make sure Power Pivot is loaded by checking its item in the COM Add-ins list.
  - It is easy to go back and forth between the Excel window and the Power Pivot window. Both can be open at the same time.

# 4-3a Basing Pivot Tables on a Data Model

- If Power Pivot is loaded, you can click the PivotTable dropdown arrow in the Power Pivot window to see the options in this figure.

The screenshot shows the Power Pivot for Excel window with the title bar "Power Pivot for Excel - Contoso Import.Finished.xlsx". The ribbon tabs are "File", "Home", "Design", and "Advanced". The "PivotTable" tab is selected in the ribbon. The main area displays a table of sales data with columns: SalesID, ProductID, ReturnAmount, TotalCost, SalesAmount, and Add Column. A dropdown menu is open over the "PivotTable" tab, listing various chart and table options:

- Chart and Table (Horizontal)
- Chart and Table (Vertical)
- Two Charts (Horizontal)
- Two Charts (Vertical)
- Four Charts
- Flattened PivotTable

At the bottom of the window, there is a ribbon bar with tabs: Sales, ProductSubcategories, Dates, Products, ProductCategories, Geography, and Stores. The "Sales" tab is selected. The status bar at the bottom shows "Record: 1 of 33,565".

# 4-5 Data Cleansing (slide 1 of 2)

- ▶ When you study statistics in a course, the data sets you analyze have usually been carefully prepared by the textbook author or your instructor. For that reason, they are usually in good shape.
- ▶ You cannot count on real-world data sets to be so perfect. This is especially the case when you obtain data from external sources such as the Web.
- ▶ There can be all sorts of problems with the data, and it is your responsibility to correct these problems before you do any serious analysis.
- ▶ This initial step, called **cleansing data**, can be tedious, but it can prevent totally misleading results later on.

## 4-5 Data Cleansing (slide 2 of 2)

- ▶ Cleansing data requires careful detective work to uncover all possible errors that might be present.
- ▶ Once an error is found, it is not always clear how to correct it. A case in point is missing data.
- ▶ For example, some respondents to a questionnaire, when asked for their annual income, might leave this box blank.
- ▶ How should you treat these questionnaires when you perform the eventual data analysis?

# Example 4.6: Customer Data with Errors



(slide 1 of 6)

- ▶ **Objective:** To find and fix errors in this company's data set.
- ▶ **Solution:**
  - We purposely constructed this data set to have a number of problems, all of which you might encounter in real data sets.
  - We begin with the Social Security Number (SSN). Presumably, all 1500 customers are distinct people, so all 1500 SSNs should be different. How can you tell if they are?

# Example 4.6: Customer Data with Errors



(slide 2 of 6)

1. Sort on the SSN column.
2. Once the SSNs are sorted, enter the formula  
**=If(B3=B2,1,0)** in cell J3 and copy this formula down column J. This formula checks whether two adjacent SSNs are equal.
3. Enter the formula **=SUM(J3:J501)** in cell J2 to see if there are any duplicate SSNs. As you can see, there are two pairs of duplicate SSNs.

# Example 4.6: Customer Data with Errors

(slide 3 of 6)

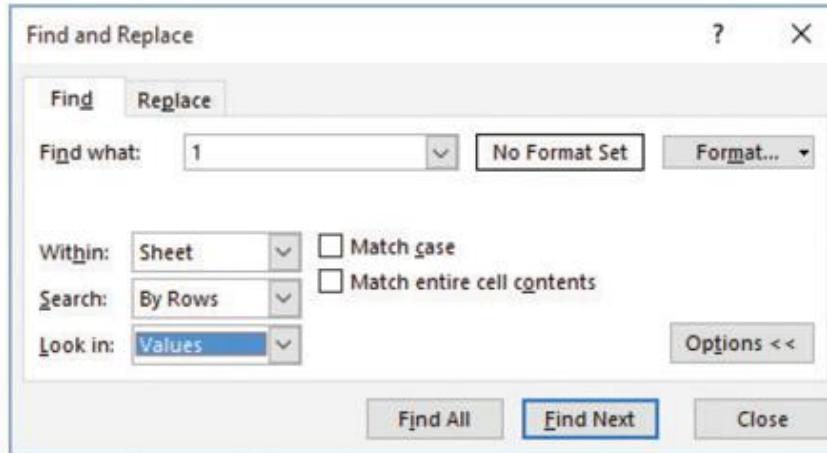
	A	B	C	D	E	F	G	H	I	J
1	Customer	SSN	Birthdate	Age	Region	Cred Card User	Income	Purchases	Amount Spent	Duplicates
2	681	001-05-3748	03/24/36	63	North	0	159700	1	530	2
3	685	001-43-2336	08/21/63	35	North	0	149300	4	1750	0
4	62	001-80-6937	12/27/54	44	West	1	44000	4	2020	0
5	787	002-23-4874	01/31/76	23	North	0	153000	3	1330	0
6	328	004-10-8303	10/19/76	22	West	1	49800	4	1940	0
7	870	004-39-9621	10/13/57	41	South	0	138900	2	1010	0

# Example 4.6: Customer Data with Errors



(slide 4 of 6)

4. To find the duplicates, select the range from cell J3 down and select Find from the Find & Select dropdown menu on the Home ribbon, with the resulting dialog box filled in as shown. In particular, make sure the bottom box has Values selected.



# Example 4.6: Customer Data with Errors



(slide 5 of 6)

5. Click the Find Next button twice to find the offenders.  
Customers 369 and 618 each have SSN 283-42-4994,  
and customers 159 and 464 each have SSN 680-00-  
1375.
  
- ▶ At this point, the company should check the SSNs of  
these four customers, which are hopefully available from  
another source, and enter them correctly here.

# Example 4.6: Customer Data with Errors

(slide 6 of 6)

- ▶ Some other problems that can occur in a data set:
  - Missing birthdates that were simply entered as 9999, which were then formatted as dates.
  - Negative ages
  - Text entries that begin with a space.
  - Entries that are interpreted as text rather than numbers.
  - Entering income as a two-digit number, assuming that the units are thousands.