December 31, 2022

**Data Visualization with Excel**

**Learning Objectives**

* Identify the guiding principles for effectively using Excel
* Identify the basic components of a data set
* Define a relational database
* Identify when you would use a relational database
* Recall how data models function within a database
* Enable PowerPivot in Excel
* Create a table in Excel
* Create a relational data model using PowerPivot
* Build a PivotTable using PowerPivot and a relational data model
* Visualize data across multiple data sources using Pivot Charts
* Create and store calculated fields in PowerPivot
* Create PowerPivot and Power Charts using calculated fields
* Create PowerPivot and Power Charts using multiple data tables and filters

## Course goals and learning objectives

**Week 1: Preparing a Professional Excel**

* Identify the guiding principles for effectively using Excel
* Identify the basic components of a data set
* Define a relational database
* Identify when you would use a relational database
* Recall how data models function within a database
* Enable PowerPivot in Excel
* Create a table in Excel
* Create a relational data model using PowerPivot
* Build a PivotTable using PowerPivot and a relational data model
* Visualize data across multiple data sources using Pivot Charts
* Create and store calculated fields in PowerPivot
* Create PowerPivot and Power Charts using calculated fields
* Create PowerPivot and Power Charts using multiple data tables and filters

**Week 2: Advanced Scenario Analysis**

* Differentiate between Scenario Analysis, Sensitivity Analysis and Simulation
* Identify tools in Excel that perform “what if” analyses when building spreadsheet models
* Determine what variable is needed to achieve end results using Goal Seek
* Determine the optimal solution to a problem using Solver
* Recognize different values for variables and compare results using Scenario Manager
* Recognize the range of outcomes for a given range of a variable input using a one-way data table
* Determine the range of outcomes under two variables using a two-way data table
* Simulate multiple outcomes for a given objective using Excel

**Week 3: Data Visualization**

* Recall the importance of data visualization
* Identify the key principles of design for building a chart, table and graph
* Differentiate between effective and ineffective charts, tables and graphs
* Identify five types of charts in Excel
* Determine the appropriate chart to use to display the data with a data set
* Create the following types of charts in Excel: Bar or column chart, combination line and bar chart, Gantt chart, heatmap and 100% stacked chart

**Week 4: Dashboarding**

* Define the term dashboarding
* Recall examples of how dashboards are used in a business environment
* Identify the benefits of using a dashboard
* Differentiate between the three types of dashboards (Executive, Analytical, and Operational)
* Recall the five principles of design for creating a dashboard
* Create interactive dashboard components in Excel, including Form Controls, Group Boxes, Slicers and PowerPivot using Form Controls and Group Boxes
* Build an interactive Key Performance Indicator, or KPI, using Excel’s Form Controls, Conditional Formatting, and IF formulas
* Demonstrate using the camera tool
* Hide elements of Excel’s user display to present the user with a clean dashboard

# Common Terms Used in Excel

**Analysis ToolPak** If you need to develop complex statistical or engineering analyses, you can save steps and time by using the Analysis ToolPak. You provide the data and parameters for each analysis, and the tool uses the appropriate statistical or engineering macro functions to calculate and display the results in an output table. Instructions can be found at <https://support.office.com/en-us/article/Load-the-Analysis-ToolPak-6a63e598-cd6d-42e3-9317-6b40ba1a66b4>

**Analytical dashboards** More detailed and used to facilitate further ad-hoc analysis. Teams use these frequently to explore data sets that may be too complicated to review at the detailed line-item level.

**Array** A display or range of specific data.

**Bar or column chart** Also known as a column chart, has the height of each bar scaled to its value to allow for comparison against other bars (or values). This may be the most common type of graph used in business when visualizing discrete values.

**Boolean** An algebraic notation used to represent logical propositions.

**Cell Locking** A tool that is used extensively in Excel. It allows the user to anchor formulas to a cell, a row, or a column. By doing this, users are able to copy and paste formulas across workbooks efficiently. It is rare to find a finalized workbook that doesn’t have cell locking.

**CLEAN function** Removes all non-printable characters from text. The Clean function looks at a value in one cell, and removes non printable characters.

**Combination line and bar chart** Uses a bar chart in conjunction with a line chart to show how two variables relate to each other. This can be useful when trying to convey multiple series of related data in one visualization.

**CONCATENATE function** Concatenate function allows users to develop text strings from multiple inputs. Concatenate function returns a text string built from all input information (text, data).

**COUNTIF/COUNTIFS function** Allows you to count the number of occurrences in a range based on a specific variable. COUNTIF statements are mathematical functions that allow you to count a number of occurrences in a range based on single or multiple variables.

**Convert Text to Columns wizard** Can be used to separate simple cell content, such as first names and last names, into different columns. This function can delimit data based on any symbol or character such as a dash or underscore.

**Correlation** Provides a “normalized” measure of covariance – i.e., of how two variables change together. (“Normalization” of covariance results in a measure which we can use to meaningfully compare how two variables move together.)

**Covariance** A measure of how much two variables change together.

**Dashboarding** allows a user to quickly focus on the most important aspects of a system using the principles of visualization and data modeling.

**Database** A collection of data sets that are linked together through various relationships.

**Data set** A file that contains data that has been collected or is being collected live. For example, an Excel file that tracks your credit card purchases could be considered a data set.

**Data Tables** provide a useful tool to change inputs incrementally to our models – either one or two at a time – and determine the impact of these changes on modeled outcomes. We could use a one-way Data Table to answer the question “How sensitive is a capital project’s Net Present Value to changes in the required return (i.e., “discount rate”) between 1% and 10%?” Excel also provides two-way Data Tables, which allow us to change two variables in our model and observe the incremental impact of these changes on modeled outcomes.

**Data validation** Data Validation is a functionality that allows you to limit potential inputs from users. Data Validation allows a user to enter data or returns an error to the user through excel, based on inputs and limits imposed on a cell.

**Dependent variable** A dependent variable is one whose value is affected by another variable; it is often denoted as y.

**Embedded IF function** Embedding IF Statements involves combining multiple IF Statements into one set of tests to run additional formulas based on the outcomes of previous tests. The embedded IF can be triggered by a test either meeting the logical requirements or not meeting them. Embedded IF statements can incorporate additional formulas within structures.

**Excel ribbon** The Excel screen is designed as a ribbon to allow for easy access to the key features of the program. You can click through the ribbon or use the Alt key to toggle between different tabs of the ribbon and different buttons in each tab. The ribbon is entirely customizable - you can rearrange the tools in the ribbon.

**Executive dashboards** contain high-level information in summary form and usually cover a long-time horizon in order to communicate key organizational trends. We frequently see these in general managers, CEO’s and other core executive offices. This type of dashboard can help provide live insights into the business and help them identify how and where to focus their time. It can also be used to communicate corporate performance.

**Evaluate Formula** The evaluate formula is a tool in excel used to evaluate complex formulas one step at a time.

**FIND function** FIND function looks for a specific text string within another text string. FIND function returns the numerical position of text within a cell.

**Formula** Use formulas to make calculations in a cell based on a relationship with another cell or multiple cells.

**Formula bar** The names of cells begin with A1 in the top left. Columns are named by letters, and rows by numbers. The equals sign “=“begins a formula, and formulas can include numbers and other cells. You can have multiple sheets and workbooks open while Excel is running, and you can reference cells from different sheets in a single workbook.

**Formula logic** Each formula has unique logic that you must follow to make the formula calculate a result.

**Gantt chart** helps visualize related timelines. Gantt charts are ubiquitous in business, though often poorly implemented.

**Goal Seek** allows us to answer “what if” questions by starting, first, with the outcome of the analysis. We tell Excel what value we want to see as the output of a model and ask it to change one input value in our model repeatedly in order to achieve the stated output. For example, in a capital budgeting problem, if we wanted to know the discount rate we should use in order to achieve a specific Net Present Value, we can use Goal Seek to determine the discount rate to use in the NPV calculation.

**Heatmap** encodes values as colors allowing for patterns to be identified more easily.

**IF function** IF Statements allow you to compare different values or variables and return a specified value depending on the outcome of that test. The test must be structured as a Boolean, or a True/False Test. If no outputs are specified, the function will return the word “TRUE” if the test is true, and “FALSE” if the test is false.

**IF & AND function** IF + AND Statements facilitate the use of multiple logical tests in evaluating a data set, where all tests must be true. Each statement within the AND function must be structured as a unique test; you can’t simply list the values being compared. Similar to the IF Statement, the AND statement can be embedded within multiple functions to run more advanced tests.

**IF + OR function** IF + OR Statements facilitate the use of multiple logical tests in evaluating a data set, where any test can be true. Each statement within the OR function must be structured as a unique test; you can’t simply list the values being compared.

**Independent variable** A variable (often denoted by x ) whose variation does not depend on that of another.

**INDEX & MATCH** When used in conjunction, INDEX & MATCH allow you to build a more dynamic VLOOKUP. By using this methodology, you can return values, not only to the right of the lookup value, but also, to the left, above or below. INDEX function allows you to return a specified value within an array. MATCH function returns the numerical position of text within an array.

**Input Criteria** Information that must be listed in order for a formula to work properly.

**Internal data model (IDM)** IDM is the analytical engine that Excel uses behind the scenes to construct the power pivot. It is basically a database in which Excel organizes information. It can hold unlimited number of rows and columns.

**IRR function** Financial function that calculates the discount rate that would return an NPV of zero.

**LEFT function** LEFT function allows you to select a substring from a string. LEFT function returns the number of characters from a string, starting from the start of the string.

**LEN function** LEN function counts the number of characters from a provided text. LEN function returns the number of characters from provided text.

**Linear regression** Using linear regression, we can quantify the relationship between changes in the independent (or input) variable, and changes in the dependent (or outcome) variable. For example, let’s look at the relationship the variables – Y, X m and B – below:

* Y = mX + B
* This relationship could be read as “Y is equal to m multiplied by X and added to B.”

**MID function** MID function allows you to select a substring from a string. MID function returns the number of characters from a string starting from the start of the string, where you have defined start\_num.

**Named Ranges** Named ranges allow the user to give an alias to certain cells, rows, columns, or tables. These named ranges can be referenced when creating or building formulas.

**NPV function** Financial function that function determines the value of future dollars, in terms of today’s dollars, using a standard discount rate.

**Operational dashboards** usually cover shorter time frames in order to focus on specific operational goals. We see these frequently in operations where they are used to measure output and other key operational KPI’s.

**Paste Special** Use the Paste Special dialog to copy complex items from a Microsoft Office Excel worksheet and paste them into the same worksheet or another Excel worksheet using only specific attributes of the copied data, or a mathematical operation that you want to apply to the copied data.

**PEMDAS** order of operations Parentheses, Exponents, Multiplication and Division, and Addition and Subtraction Function.

**PivotCharts** Allows you to visualize data across multiple data sources.

**Pivot Table** A pivot table allows you to extract data from a large, detailed data set and transform static data into useful, easy-to-read information. Pivot tables allow you to organize vast amounts of data easily in many different ways. Pivot tables move and combine data fields easily to show different analyses. Customizable final Pivot Table display and format and creating your own formulas within a pivot table to increase your outcomes.

**PowerPivot** As the name implies, it is similar to pivot table feature in Excel; however, the PowerPivot is able to run pivot tables by merging several big datasets and run pivot tables off of the consolidated data.

**PROPER function** PROPER function allows you to quickly format text for proper nouns. PROPER function returns ‘properly’ formatted text where the first letter is capitalized and all other characters of a word are in lowercase.

**Relational database** A relational database is a database structured to recognize relationships among stored items of information. This type of database will have datasets that are linked through relationships stored within the data model. MySQL, Oracle Database, and SQL Server are all examples of relational database software. A relational database could also be any file that has data in it.

**R-Squared or the Coefficient of Determination** A number that indicates the proportion of the variance in one variable that is predictable from the other variable.

**Regression analysis** A method of analyzing how changes in input, or independent, variables affect the outcome, or dependent, variable.

**RIGHT function** RIGHT function allows you to select a substring from a string. RIGHT function returns the number of characters from a string, starting from the end of the string.

**Scenario Analysis** involves the definition of a specific state of the world, and the related impact on variables being analyzed, to capture and quantify outcomes. Scenario analysis helps us formalize one or more possible answers to questions about the future.

**Scenario Manager** allows us to save inputs to our model – and therefore related outputs – as named scenarios to facilitate asking these “what if” questions in our analyses. We may use Scenario Manager, for example, to capture outputs of multiple scenarios which answer the question “What is the Net Present Value under my best, moderate and worst-case views of the prospects for a given capital project?”

**Sensitivity analysis** involves changing the values of an input to a model or formula incrementally and measuring the related change in outcome(s).

**Simulation** involves iterating through possible values of one (or more) variables of a model in order to capture possible (and, ultimately, likely) outcomes. Simulation can be deterministic – in that we iterate over all possible values of a variable – or non-deterministic or random – in that we iterate over values based on some assumption of the probability of a variable taking on these values.

**Slope** The slope of the function is a number that describes both the direction and the steepness of the line graph of that function.

**Solver** is a powerful tool to help us mathematically optimize outcomes in our models to support scenario analysis such as “best case” or “worst case” analyses. Mathematical optimization can be thought of as determining the “best” (or “worst”) outcome, subject to defined constraints, from a group of alternative possible outcomes. For example, we may ask “What is the combination of capital projects (i.e., “best case”) I should pursue in order to maximize the Net Present Value of my capital allocation?” This is a simplified example, and the built-in Solver tool provided with Excel is very powerful, though we will barely scratch the surface of use cases for this potent optimization engine. We encourage you to explore Solver, and its many uses, independently to enhance your own spreadsheet analyses.

**Standard deviation** Measurement of the average dispersion of values in a dataset around their average value (i.e., how “spread out” the data are from their average value). It is related to variance, but is more frequently used to describe the average dispersion of a dataset.

**Statistical forecasting** The process of making predictions of the future based on past and present data and most commonly by analysis of trends.

**Statistical methods** Methods of collecting, summarizing, analyzing, and interpreting variable numerical data. Statistical methods can be contrasted with deterministic methods, which are appropriate where observations are exactly reproducible or are assumed to be so.

**Subtotal** The total of one set of a larger group of figures to be added.

**Subtotal function** The SUBTOTAL function returns a subtotal in a list or database. The Subtotal function has the ability to perform several mathematical functions such as sum, average, count, product, min, max and more. The SUBSTITUTE function substitutes new\_text for old\_text in a text string.

**TEXT Functions** text functions allow you to parse out data by a common delimiter in order to make data easier to filter.

**Traceability** Traceability is a feature in excel that helps users error check their work. Under the Traceability ribbon, users are able to trace the precedents and the dependents. They can also show formulas and use error checking.

**Trace Dependents** A tool used to find dependent cells in a formula.

**Trace Precedents** A tool used to find precedent cells in a formula.

**TRIM function** The TRIM function removes all spaces from a string (with the exception of single spaces between words).

**UPPER function** UPPER function allows you to quickly format text in uppercase. UPPER function returns uppercase formatted text.

**VALUE function** The VALUE function converts a text string that represents a number into a number. The text enclosed in quotation marks or a reference to a cell containing the text you want to convert.

**Variance** Measures how far, on average, a set of data values are “spread out” from their average, or “mean”.

**VLOOKUP** VLOOKUP functions allow to lookup a value you want to find in an Excel list or table. VLOOKUP function looks at a value in one column, and finds its corresponding value on the same row in another column. VLOOKUP stands for vertical, which means the data in the table must be arranged vertically, with data in rows. (For horizontally structured data, see HLOOKUP)

**XIRR function** XIRR function calculates the discount rate that would return an NPV of zero.

**XNPV function** Financial function that determines the value of future dollars in terms of today’s dollars using a standard discount rate for payments not equally spaced in time.

**Y-intercept** The point where the graph of a function (or, in our case, the graph of our relationship between our two variables) intersects with the y-axis.

Introduction to the Scenarios used in this course:

Throughout this course, we are going to be following two continuous, fictitious storylines from beginning to end. Our instructors will use these stories to teach you concepts in this course. In the first story, we will be helping a small startup airline connect their data sources, help them solve some issues. And in the second story, we'll be helping a boutique retailer organize their data, and then make better decisions by utilizing more advanced scenario and visualization technique.

First, let's explore our airline example. Throughout the student exercise, you will be playing the role of data analyst for a small startup airline operating out of the Chicago, Illinois area in the United States.

The airline's management team has a strong airlines operations background, but don't have much financial or data analytics experience. Many of their systems are disparate and not connected. In your interview, you are shocked to find out that the airline cannot even pull, from a singular location, details about aircraft that are flying specific routes and seating capacity of those aircraft. But, instead, would need to go and open two separate data files. The airline is hoping that you can help them solve their issues by not only connecting their data sources, but also helping them make some large strategic decisions that they need to make as they continue to grow and position themselves in the market. For our second story, you will be working on your own after having the techniques demonstrated in the instructional videos. You now have been hired by a rapidly expanding company that manufactures clothing. The company has been around a long time, and has operated under the strong gut feeling of the founder.

However, the founder's daughter is taking a larger role in the business. She's combining her father's strong gut market feel with her analytical analysis to drive a boom in sales.

This sales growth has strained the business infrastructure and many of the supporting processes. You've been hired to help the company continue to enhance their analytical decision-making.

The guiding principles for effectively using Excel:

**Principle 1: Be client ready**

remember that your work is a deliverable used to communicate your findings. I call it being client ready. Your work may be used by someone else down the road, or you may need to reacquaint yourself with the analysis down the line. Headers, footers, and a clean layout can not only help you organize your thought process, but really elevate your work.

**Principle 2: Begin with the end in mind**

**Principle 3: Start with a template**

**Principle 4: Separate the data**

Each data set you work with should live independently from non-associated data. Throughout this course, we will learn techniques to help bridge relationships between your data, and extract the relevant attributes we need for our analytics.

**Principle 5: Limit hard coding**

limit hard coding as much as possible. Avoid using hard numbers in your analysis, and instead, define a variable and then reference that variable when needed. This helps to enable consistent assumptions across your work, and makes it much more flexible and easier to update.

**Principle 6: Break down large, complex formulas**

We could try to write one long formula, or piece of code, but generally, it's better to break it down to its individual components. Not only will it be easier to update and transfer, but it limits errors and helps someone follow the story of your data.

**Principle 7: Do a thorough review**

Once your work is done, do a thorough review. When I receive or finish a draft of any model, I always assume there are mistakes. Thorough reviews are essential.

But I also look to build in error checks and other validation techniques to let the data check itself, by calculating the results in two ways to verify that they

come out aligned.

**Principle 8: Ask for help or look online**

If you find something that is taking too long to complete, chances are that someone has run into the same problem before. Leverage your network, or search for that online. There is likely a blog or a video out there that will help you quickly figure out the syntax for a formula, or help you construct your analysis.

**Principle 9: Use the keyboard**

We'd like to encourage our analyst to use as many shortcuts as possible. There are dozens of keyboard shortcuts, and taking the time to learn them now will dramatically reduce the time it takes for you to do your work. Don't start by memorizing random shortcuts, but learn to make a language. If you start finding ourself using your similar command, stop and learn the shortcut, and add it on your technical vocabulary.

## Data Visualization:

Data and analytics are all about bringing client challenges to life so we can overcome them by taking data, whether structured, or unstructured, and turning it into usable information and insights. Once we gather the data and perform your analysis, you need to be able to present that data in a way that makes sense to your end user. In other words, you have to visualize the data. Visualization is all about making complex insights simple.

For centuries, people have depended on visual representations such as charts, and maps, to understand information more easily and quickly. As more and more data are collected and analyzed, decision makers at all levels of the organization welcome data visualization software the enables them to see analytical results presented visually. It helps them to find relevance among millions of variables, communicate concepts and hypotheses to others and even predict the future.

**Visualization is important because**

* the brain processes visual information faster.
* Users can spot patterns or trends which are not obvious in a flat structure.
* Information can be interpreted by any audience.
* Decision making is faster and cheaper.

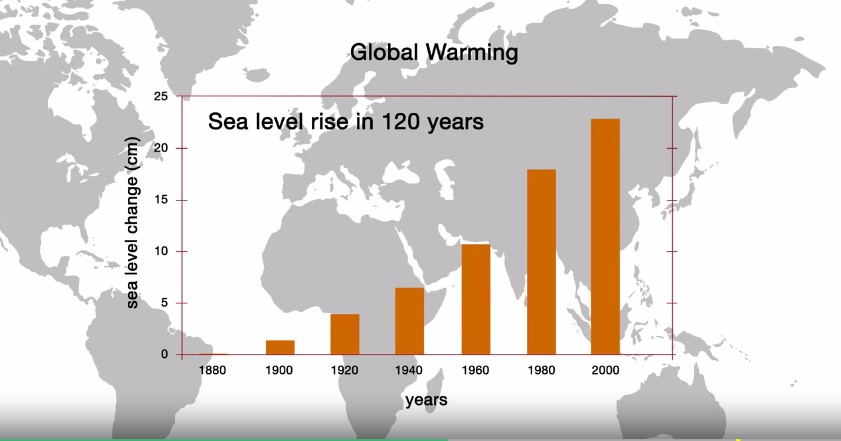
Being able to visualize data allows you to present an analysis so that anyone can understand what the information means.

Working with Data Sets [Data Model]:

As we start to work with data, we need to understand three basic concepts.

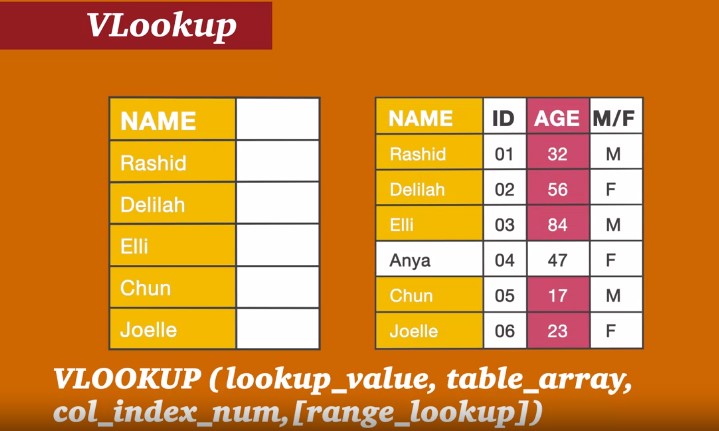
1. What is a data set?
2. How can we combine data sets?
3. What is a database?
4. What is a data set and why is it useful?

A data set is a file that contains data that has been collected or is being collected live. For instance, an Excel file that tracks our credit card purchases could be considered a data set. Data sets are useful as they are a core component of the data driven decision making process.



Let’s say a global warming theory is based on data analytics. The sea levels continue to rise every year, and having collected that data for years, scientists are able to make inferences from the data. What if I have two data sets that share common elements? Can I combine them?

In course two of this specialization, we use the VLOOKUP function to look up values from one table to combine it with another. VLOOKUP’s up to work because there are matching values between two tables. Those pieces of data are linked by that value.



And finally, what is a database? A database is a collection of data sets that are linked together through various relationships. We find databases and data models throughout the business world. Databases comprise core data infrastructure that supports the modern business. Today, we find everything is tracked and stored. This can range from financial data to how many times the door on floor 12 was unlocked by a given employee. All of this information needs to be stored and linked to other data sets using relationships. For the example previously mentioned, you may be linking the door unlock data set to the employee security data set, which may be linked to the individual employee ID at the company.

Introduction to the Components of Data Sets and the Relational Databases:

One of the most common types of databases you will come across is a rational database. In this video I'm going to talk about what a relational database is, and why you would want to use this type of database. So, what is a relational database.?

A relational database is a database structured to recognize relationships among stored items of information. This type of database will have data sets that are linked through relationships stored within the data model.

MySQL, Oracle and SQL Server are all examples of relational database software.

The next big question is, why would someone be interested in using a relational database? Today's organizations all have access to huge amounts of data. And the amount of information collected on a daily basis continues to grow. Saving data provides organizations, the ability to automate their tasks and improve decision making. Organizations use data to make decisions based on quantitative analysis instead of a gut feel. This can lead to better decisions, but it is also the keystone for automation.

In real business environments, many different data sets are needed to be combined to drive business insights. For example, business analytics can be used on customer data to predict future purchasing behaviors. Or analyzing inventory trends can lead to automated ordering algorithms. Given how important databases are in today's business environment, it is crucial to understand how data is stored in these databases. The concept of a data model is key to understanding this.

Let's take a look at data models. What is a data model? And how might a data model function within a database?

A data model will tell us what data can be stored, where it can be stored, and how to store the data. It will also prevent bad data from entering the database.

A data model states the rules under which the database will operate. Other places where you might have heard of a data model is in XML. XML uses a hierarchical data model, which means that data is stored in such a manner that it looks like a tree. Every tree has branches. And every branch has leaves.

Relational data bases use a relational data model. This means that the data is stored in rows rather than in tree like structures. Finally, why is relational data useful in today's world? Relational data allows for tables to connect and share data. Using this model, duplicate data will not be collected because data is typically only stored once in a relational model. You can relate one row of data to many other rows of data. In other models, that one row of data will have to be repeated many times. This matters when a database becomes large because maintaining it will be more difficult. Relational databases strive to keep the database size smaller and prevent duplicate rows of data. In this video I gave you a high-level overview of relational databases and why would use them.

Introduction to Keys and Forming Data Tables:

Storing data within a database is important. But as we learned in data cleansing, it is also important to understand how the data is stored. Taking things one step further, it is also important to understand the relationships between the data stored in a database. Let’s talk about the architecture of a database. We are going to look at **tables, rows, functions and columns**, and talk about what they are and how they are used within a relational database.

**Tables and used within a relational database**:

A table is used to store data within a database and should be thought of as a dataset. Typically, a database would be comprised of many different tables. In the case of a relational database, usually there will be many tables that have some relation to each other.

**Row and its function within a relational database**:

A row is an entry of data into a table. Databases are comprised of tables and tables are comprised of rows. Each row is an entry within a table. Row might share a relationship with a row in another table, just like tables might have a relationship to one another.

**Column**:

A column is an element within a row. In the table setup, the column dictates what type of information is stored within all rows within the table. For example, if we have a data set that includes hospital patient data, the first column could be the name of a patient, the second the age of that patient within the row, and the third column is the patient's address. Frequently, we refer to these columns as attributes. Databases are comprised of tables, tables are comprised of rows and rows are comprised of columns, also known as attributes. Now that I have introduced you to the architecture of a database, let's take a look at a graphical representation of a database to illustrate these concepts. The graphical representation of row, table, and column.



**Keys**:

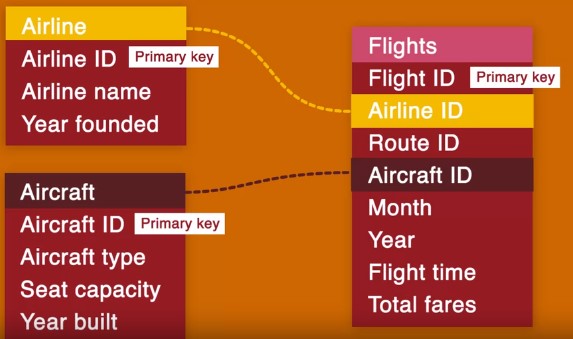
* We used keys to establish the relationships
* Unlock further information about a matching key within a new table
* Two types of keys to unlock these relationships: primary keys and foreign keys

**Primary Key**:

It is an attribute that mandates a value will be unique throughout the entire table.

**Foreign Key**:

It is an attribute within a table that provides a link between data in two tables.



**Introduction to joins:**

Joins allow two different rows from different tables to come together to become one large row containing the value from both rows.

* Inner join
* Outer join
* Left join
* Right join
* Outer join – right outer join
* Outer join – left outer join

**Overview of SQL and Other database tools**:

When you have a database, it is important that we have a tool that allows us to interact with the data. A common term you will hear while interacting with database is SQL, Structured Query Language, this is nothing more than a programming language designed to interact with databases, and the data contained within the databases.

SQL stands for Structured Query Language. It is used to give commands within a relational database, and perform the following functions on the data. Create tables to store data within rows. Update tables so that the data is entered into, and maintained within the database. Read data from tables so that analysis can be performed. Delete data from tables so the database can be maintained.

And in relational database software, we'll use their own version of SQL. Hence, SQL vary slightly from database software to database software. But they're essentially the same languages. Here are a few examples of software you will use SQL with.

Oracle Database, MySQL Database, Toad for SQL Server and Microsoft Access.

We won't be using SQL programming language, but we will use Microsoft Excel pivot table functionality, to extract data from the database.

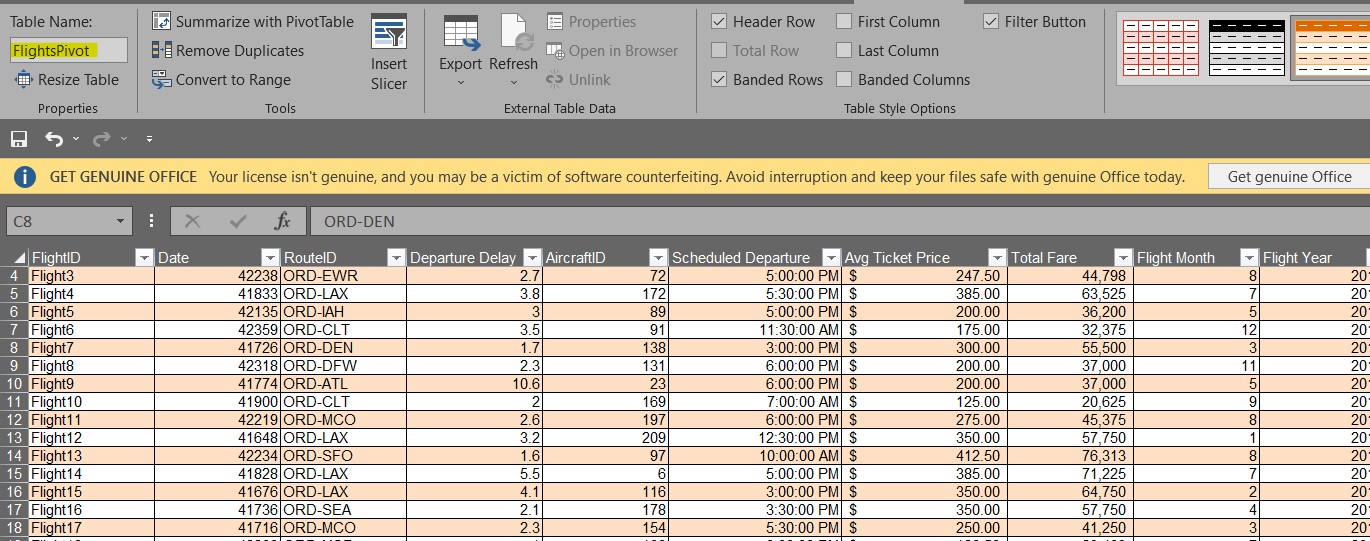
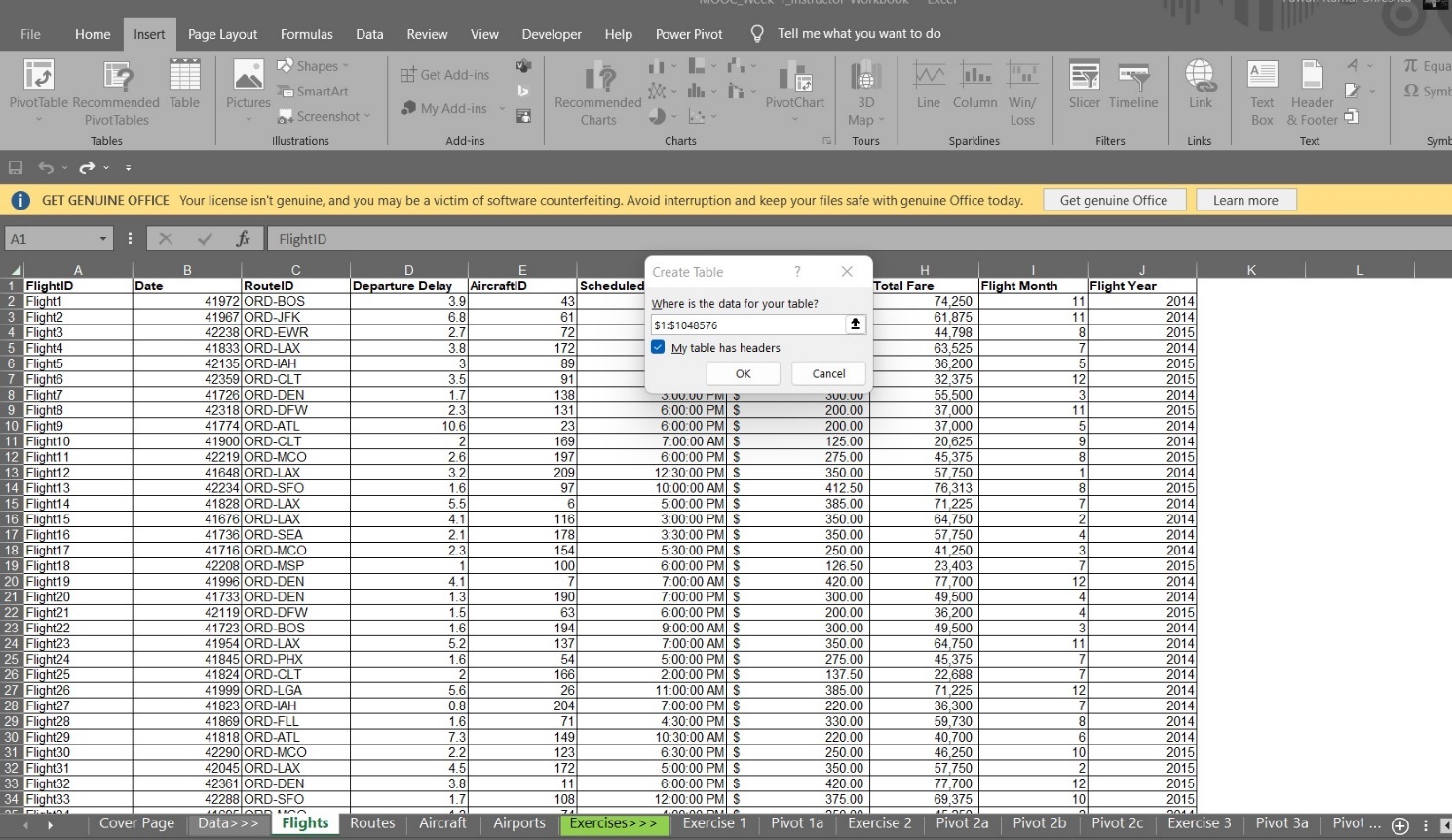
We are aware, that there are more powerful database tools out there but throughout this course, we are going to be using **Excel**, as it is flexible and can be used to demonstrate a range of concepts. It is also very common in the market and highly accessible. Excel will be a great tool to practically apply all that we have just learned.

Linking Data Tables and Building the Relational Model:

**Objective:**

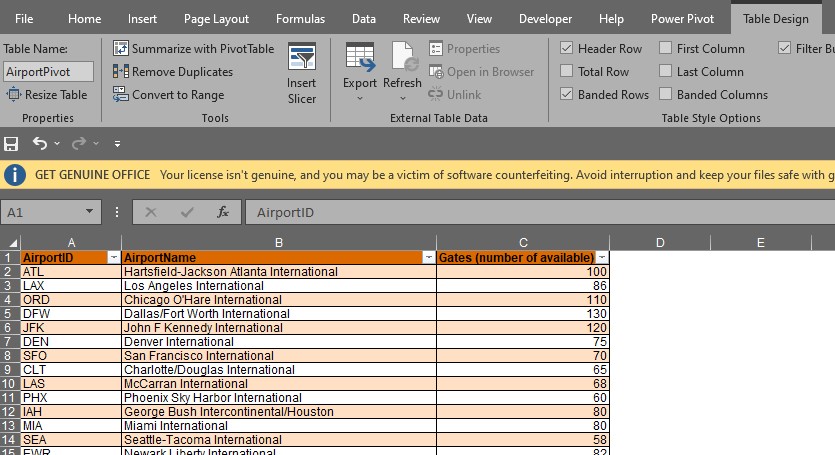
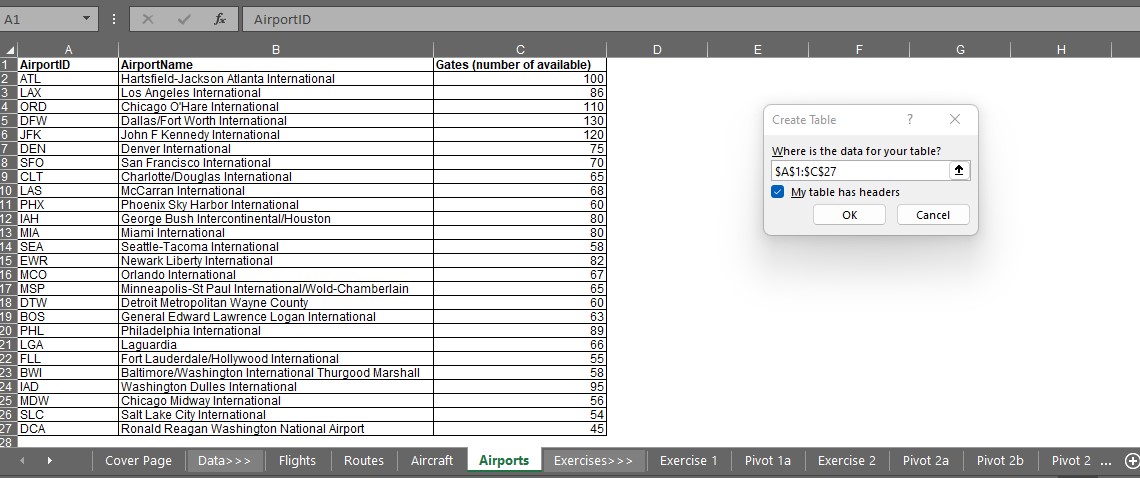
To create a relational data model using PowerPivot.

We’re using PowerPivot to organize our data into a relational data model. PowerPivot creates data models that can be built out into PivotTables and Pivot Charts.



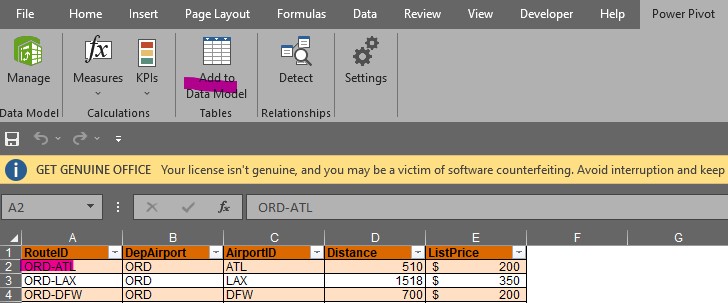
For AircraftPivot:

Select all data and CTRL+T and go to design and change the table name to AircraftPivot.

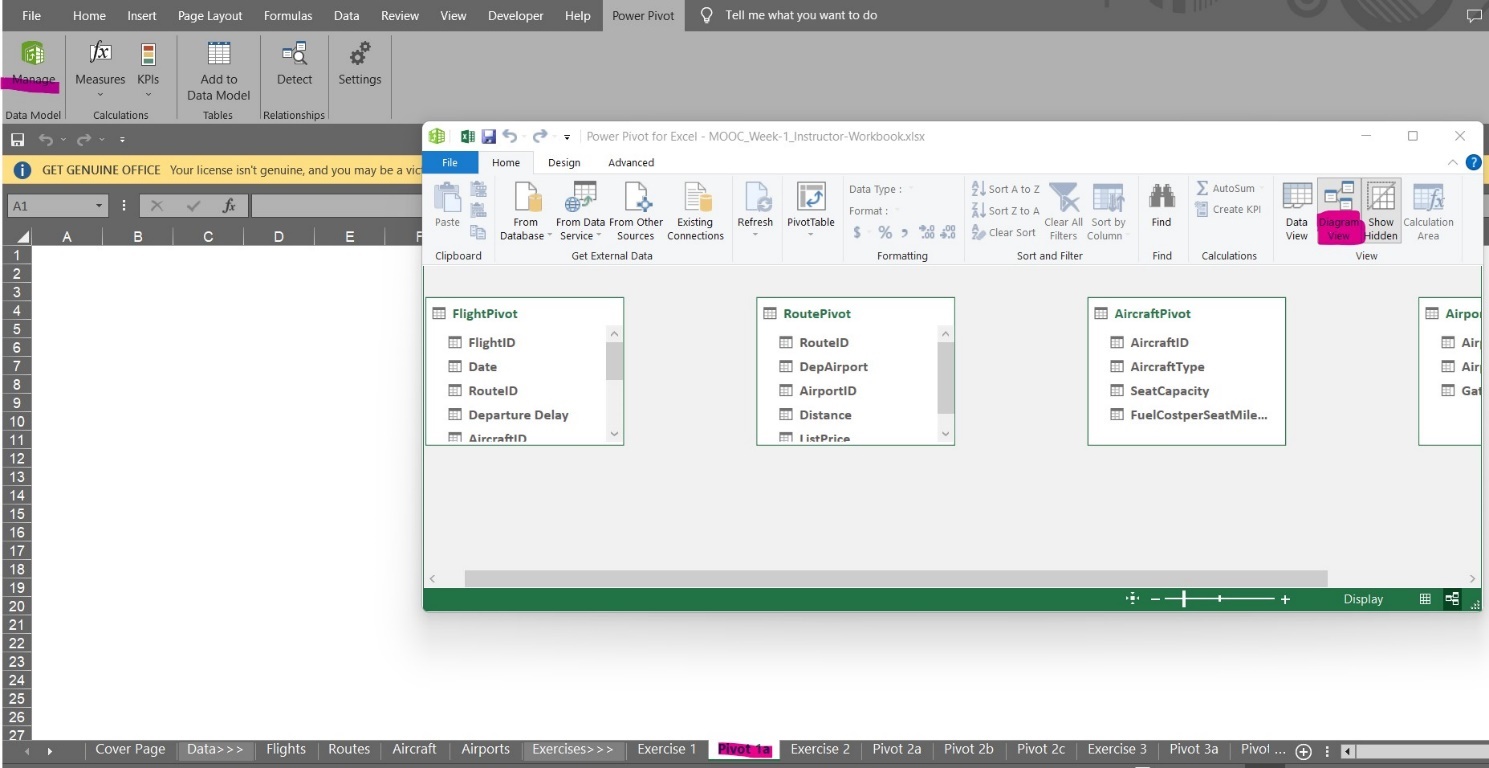


Now we have created our four data tables – FlightPivot, RoutePivot, AircraftPivot and AirportPivot.

And it’s time to complete our data model in PowerPivot. First, we must load each individual table into the PowerPivot tool in order to add them to our data model. Let’s start with the Flights table. First, we’re going to navigate with table and then click on a cell in the table. On the PowerPivot tab, click add to data model. We need to do this for all three tables as well.

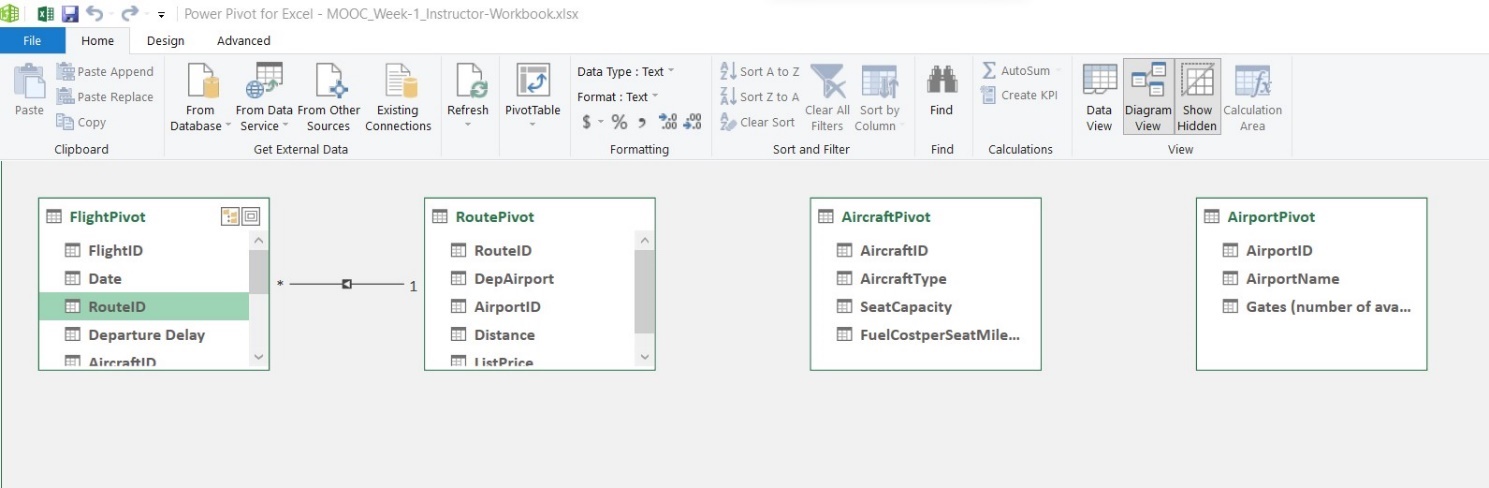


Now, we’re added all 4 of the tables to our data model, we need to form the required relationships between the tables to create our PowerPivot.

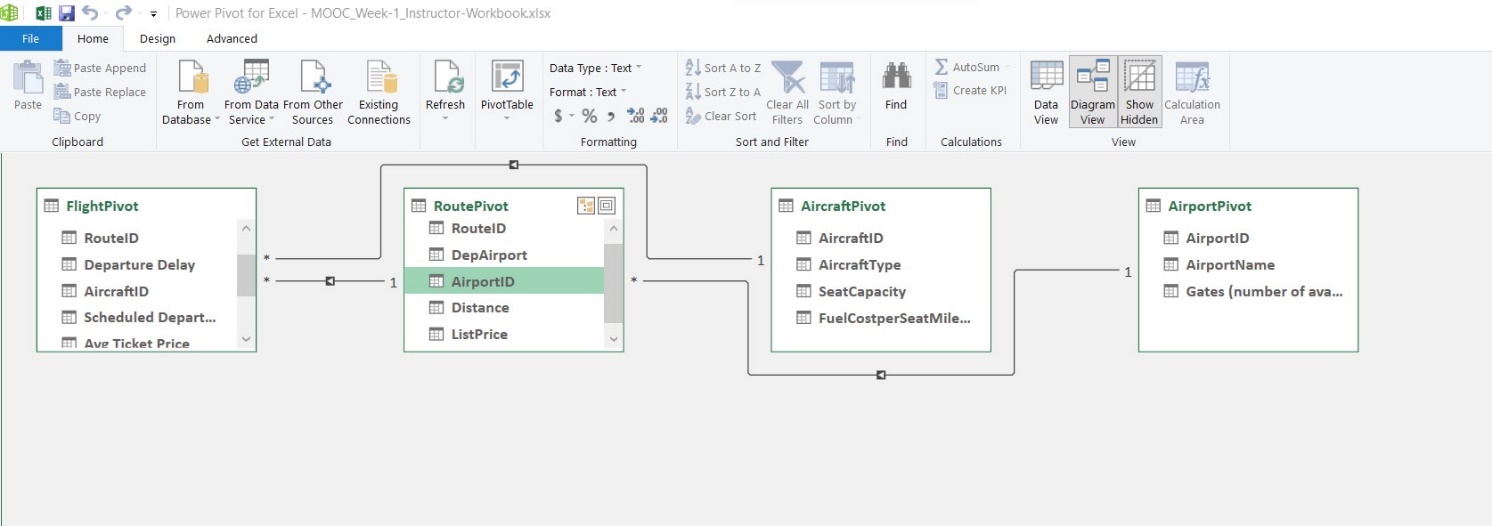


This enables us to see each of the tables we have previously saved within the data model. For each table, we need to identify how it links to the other tables by dragging lines between the tables to create relationships. In order to be linked in PowerPivot, they must have the exact same name.

We can have RouteID in both Flights and Routes Pivot which means that this is a foreign key that we can use to link the two tables. Clicking RoutID on FlightPivot table and dragging it onto the Rout ID in the routes tables.



As we can see, this creates a link in the diagram showing that these two fields have been linked.



As we can see, the diagram now has multiple lines on it, representing all of the relationships we just created.

Using PowerPivot to visualize data:

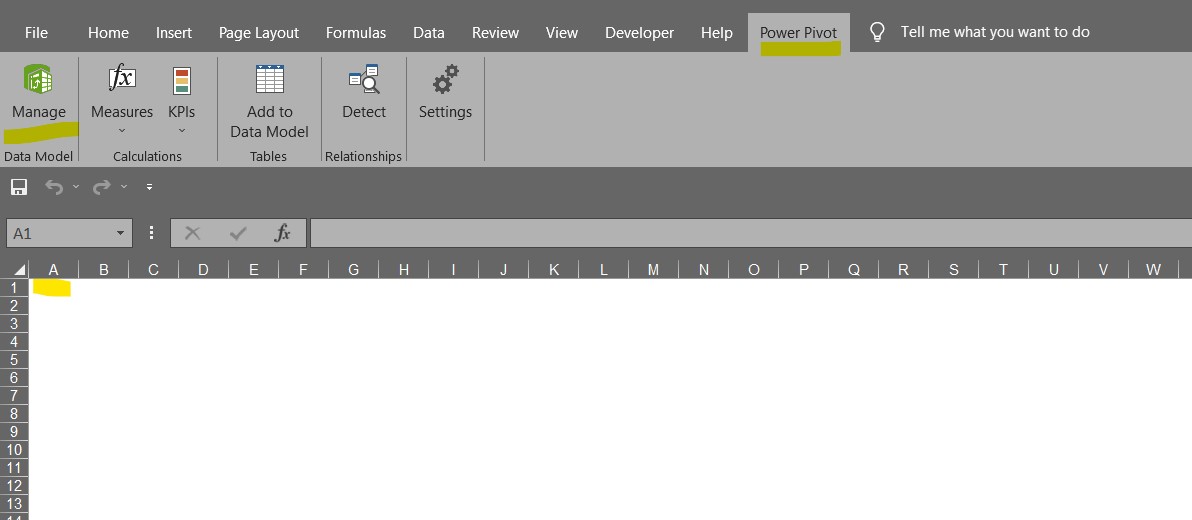
Let’s start with question 2a.

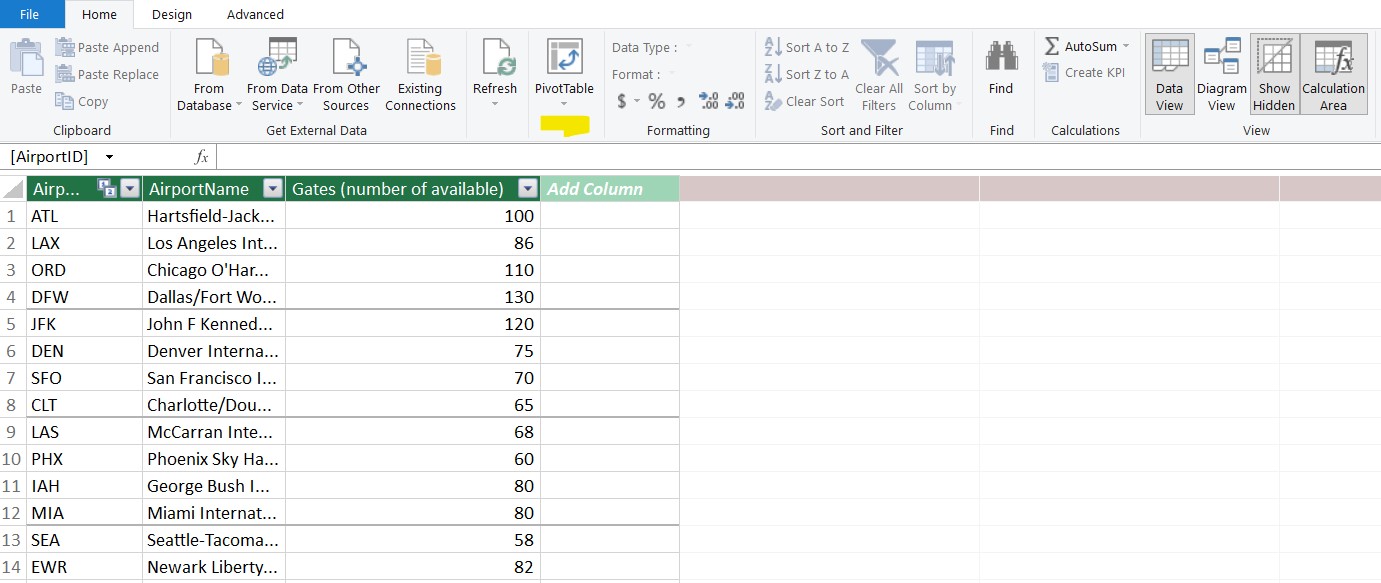
**How many flights use the A319 aircraft**?

As we know from previous exercise that flight and aircraft data are stored in 2 separate tables. However, we used the AircraftID field to link these 2 tables together in the last exercise.

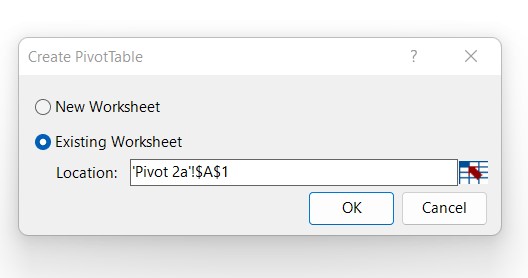
We’re going to place our PivotTables on tabs 2a, 2b and 3c as we can see in the workbook that provided. Let’s navigate Pivot 2a. To setup a pivot table using PowerPivot, select the cell we would like to place our table in. A cell going to be A1.

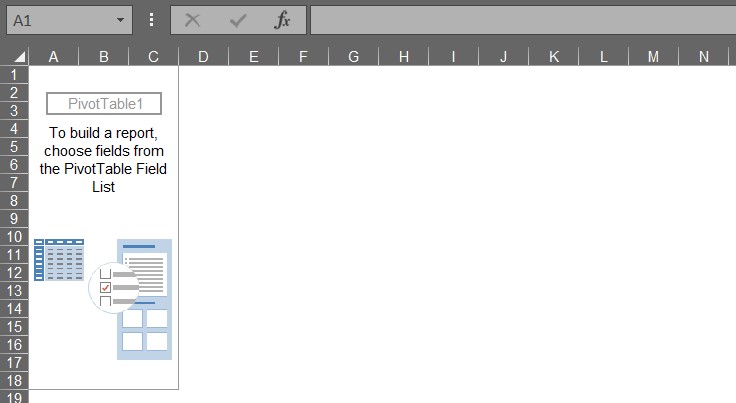
Go to PowerPivot tab and click on manage then click on Pivot table.





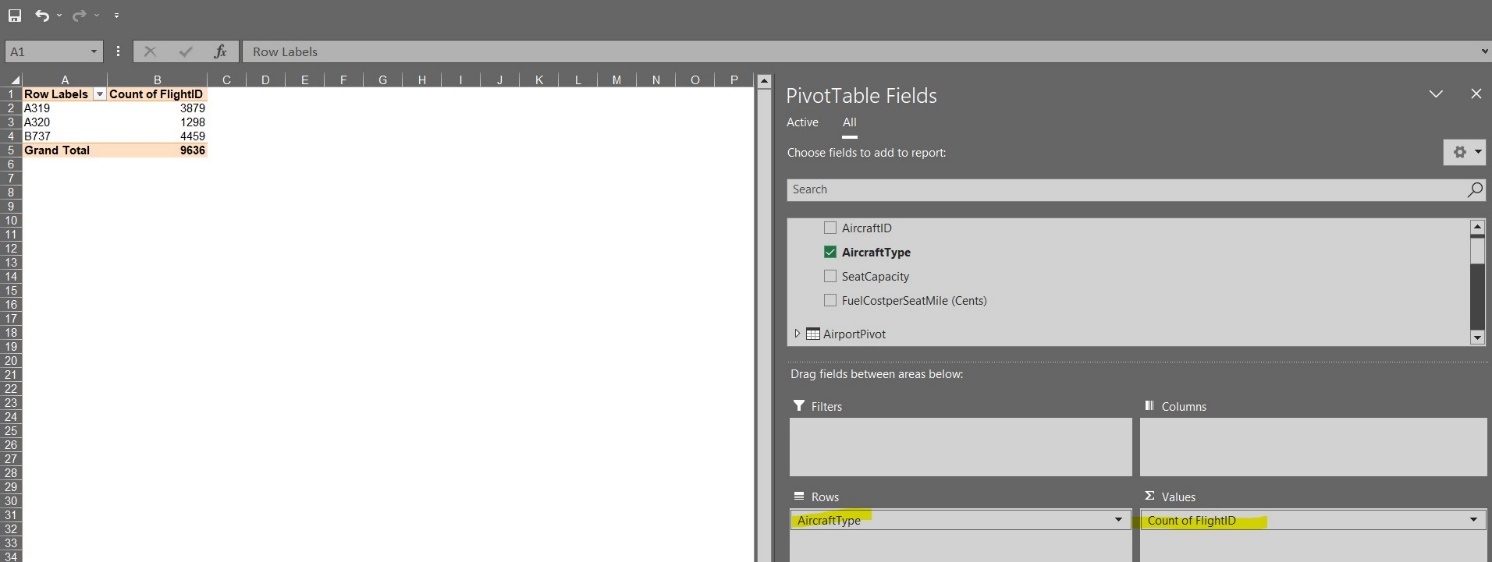
Choose Pivot Table and choose existing workbook option and ok.





PivotTables created with PowerPivot operate similar to regular PivotTables, except for one main feature: Data can be sliced and diced across multiple tables. Now expand FlightPivot and AircraftPivot.

Now we can see the two fields that we are looking for. Click FlightID and drag it to the VALUES section and click AircraftType and drag it to the ROWS section.

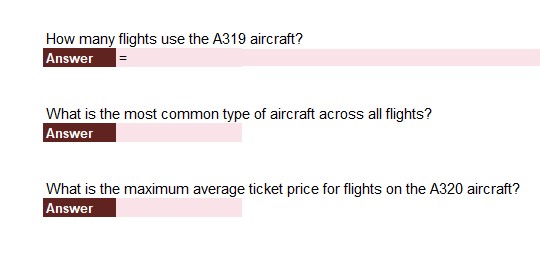


And we see a pivot table form.

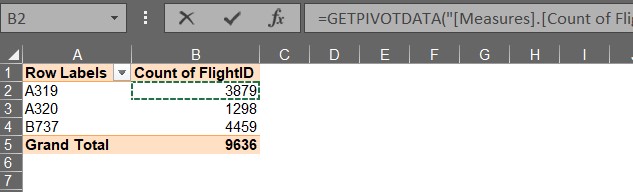
The Pivot Table is showing a **count of the flight IDs** which is what we want to show since we are trying to determine the number of flights using the A319 aircraft.

If we see cell B2, they were a total of **3879 flights using A319 aircraft**.

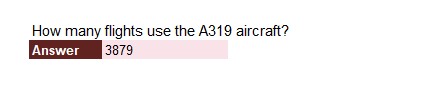
So, let’s go and link the value in the pivot table to the answer on exercise 2 tab. Select the cell V13, and type =.



Next, we’ll navigate the tab with our Pivot 2a and select the value in the cell B2.



And press Enter key.



If the data tables are updated, we could refresh the pivot which would automatically update the answer in this lab.

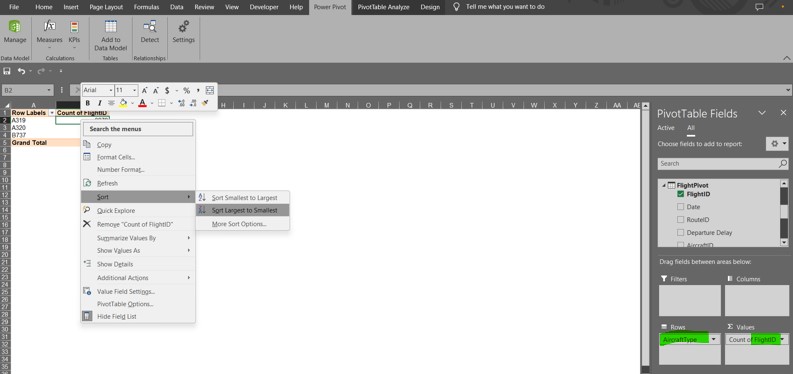
Let’s look at the second question,

**What is the most common type of aircraft across all flights**?

The pivot table required will be very similar to the one used in question 2a but we'll also need to use the sort function to find the most commonly used aircraft. We will begin by navigating to the pivot 2b tab and opening power pivot.

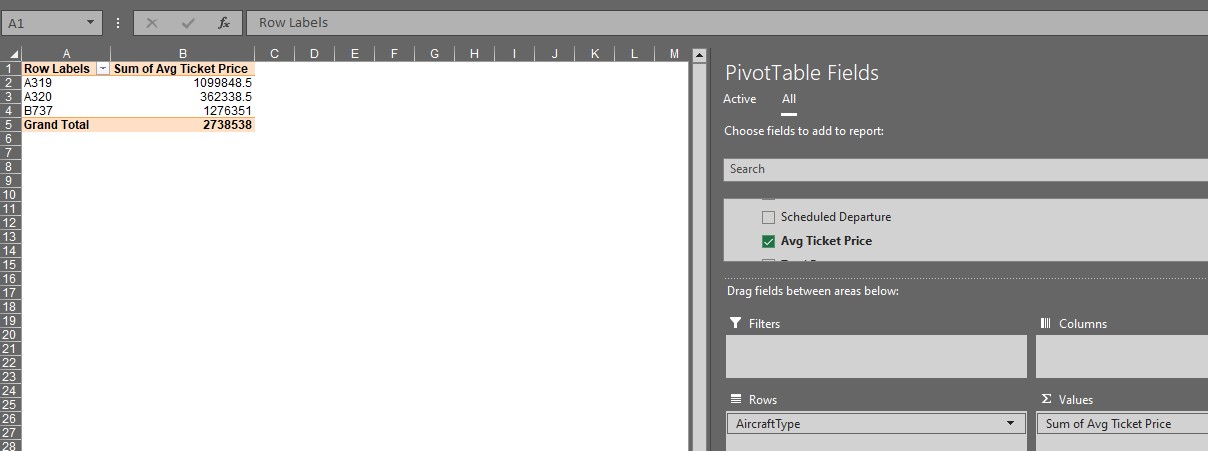
After selecting the pivot table option, we can select existing worksheet and then OK.

Similar to last time, where we'll drop the FlightID and Aircraft Type into the ROWS and VALUES section. We'll need to solve the selection to find our solution. Select the Pivot table, right click and select Sort Largest to Smallest. They should give us our desired answer.

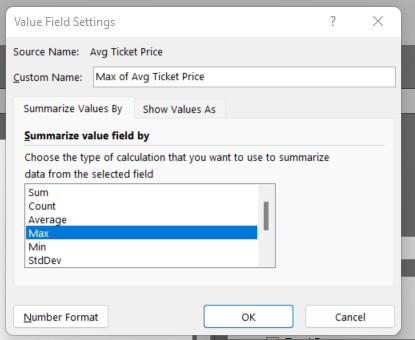
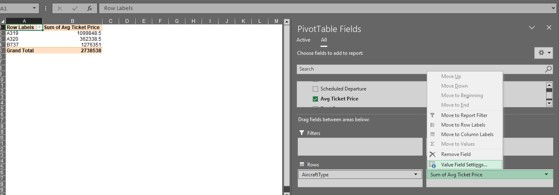


The next problem is a little trickier as we must now use a different type of value calculation to find our answer. Let's set up our pivot table in section 2C. By again, going through the power pivot tab.

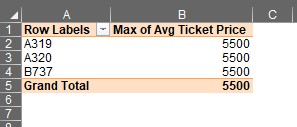
Since we are interested in the type of aircraft and average ticket price. Let's drag these fields into the rows and values sections.



We're looking for the maximum average ticket price for the A320F graph. Click that field under the value section and choose value field setting. Select Max and click OK. We will now see the maximum ticket price for all aircrafts.



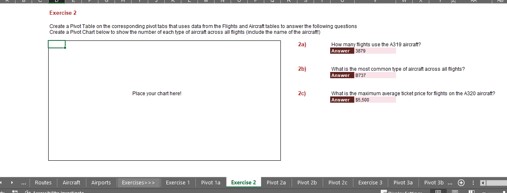
Now we can see the maximum price for all aircraft is $5500.





Another feature of PowerPivot is PivotCharts. PivotCharts allow us to visualize data across multiple data sources. We need to **visualize the number of flights flown for each aircraft type.**

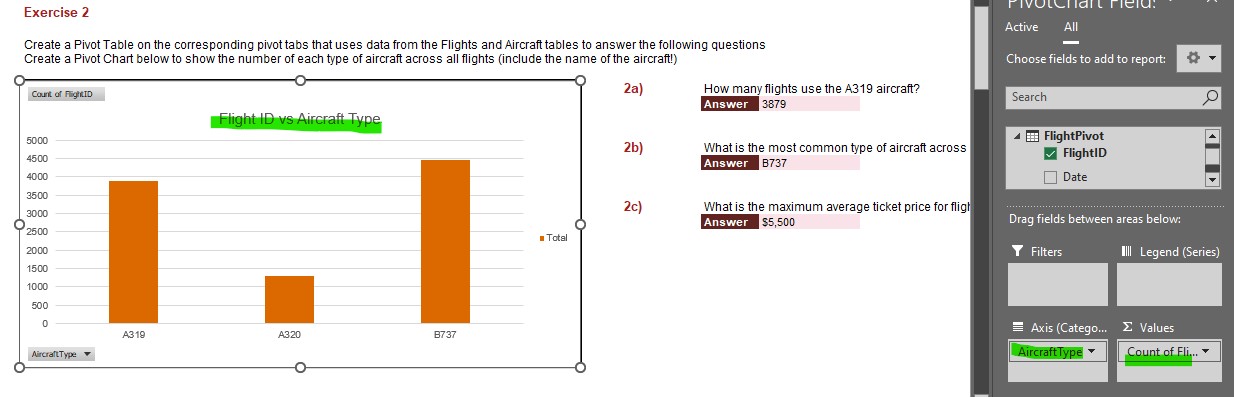
To do this we’re going to select the cell D12 in the Exercise 2 tab and PowerPivot and click manage.



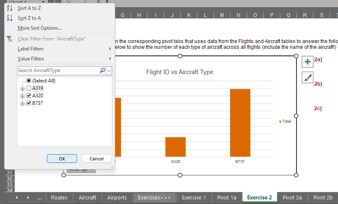
And select PivotTable option and we’re going to select PivotChart. Now adjust the height and width of the chart. And select values for the various axises and measurements similar to creating a Pivot Table. For this exercise we need to **find a number of flights flown for each aircraft.**

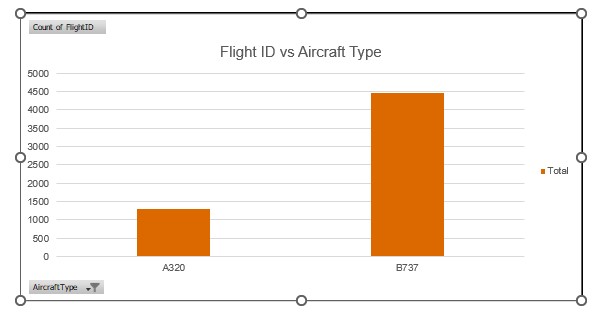
So, we drag the aircraft type attribute from our aircraft pivot data set into axis section. Then we drag our flight ID attribute from our flight pivot dataset into our values section.

We can now see a chart detailing our chosen information.



Finally, take note of the filter in the bottom right-hand corner of the chart. This can be used to manipulate the chart to filter on the values another options. Let’s select only A320 and B737.





By far this now, we learned how to use PowerPivot and the relational data the relational data model to build pivot tables and pivot charts to solve problems.

**Calculated Fields**:

For this exercise, the CFO of the airline has identified two major values that will need to be calculated in the future to better understand rates and adjust to a changing business environment.

One, the cost per mile for each aircraft in dollars. Two, the total revenue for each flight, assuming that there is now an additional 10% tax on all fares. With a few simple Excel formulas, we can leverage existing fields in our tables to address these changes.

Let's start with the first question where we are being asked, which aircraft has the highest fuel cost per mile?

In order to answer this question, we're going to start by creating a field that will calculate the total cost per mile for each aircraft.

If we look in the aircraft data table, we see that we have the fuel cost per seat per mile in column D. But this only gives us a cost for a seat and not for the entire aircraft. However, the aircraft table also provides a seat capacity of each aircraft in column C.

If we combine information in these two columns, we'll be able to calculate the total fuel cost per mile for all of the aircrafts.

We must also consider the fact that the fuel cost per mile is measured in cents, which is expected due to its low value. However,

we're going to want to show the total cost per aircraft in dollars, since cents aren't a standard measure.

To do this we'll need to take the total cost in cents and divide that by 100. Since there are 100 cents in each dollar.

Assuming that each flight is full, we can calculate the fuel costs per mile by multiplying fuel costs per seat per mile by the seat capacity of the airplane, and then dividing that value by 100 to get the total cost in dollars. We're going to **be using a Calculated Field to perform this calculation, and then we'll use this field as input to a pivot table which will give us the analysis we're looking for**. So, let's get started.

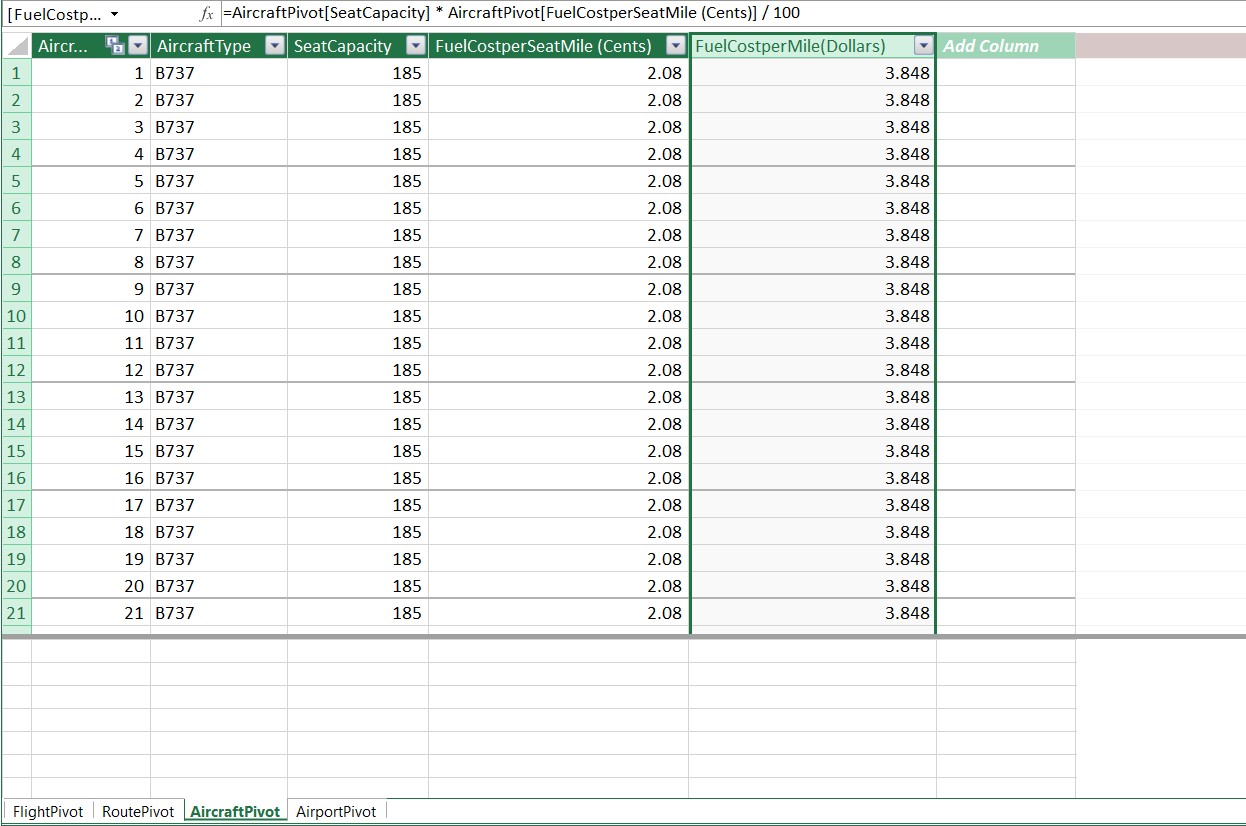
Total cost in dollar = SeatCapacity \* FuelCostperMile / 100

We're going to start by creating the Calculated Field by navigating to the PowerPivot tab and selecting Manage. Within the Data View section, there is an option that allows us to select the data table that we want to add the Calculated Field to.

The values we want to use are both in the aircraft table. So let's go ahead and select AircraftPivot from the list.

There is an add column option on the right side which contains a formula bar that can be used to input the calculations similar to how regular formulas work in Excel. So let's go ahead and click on the formula bar to start editing the formula.

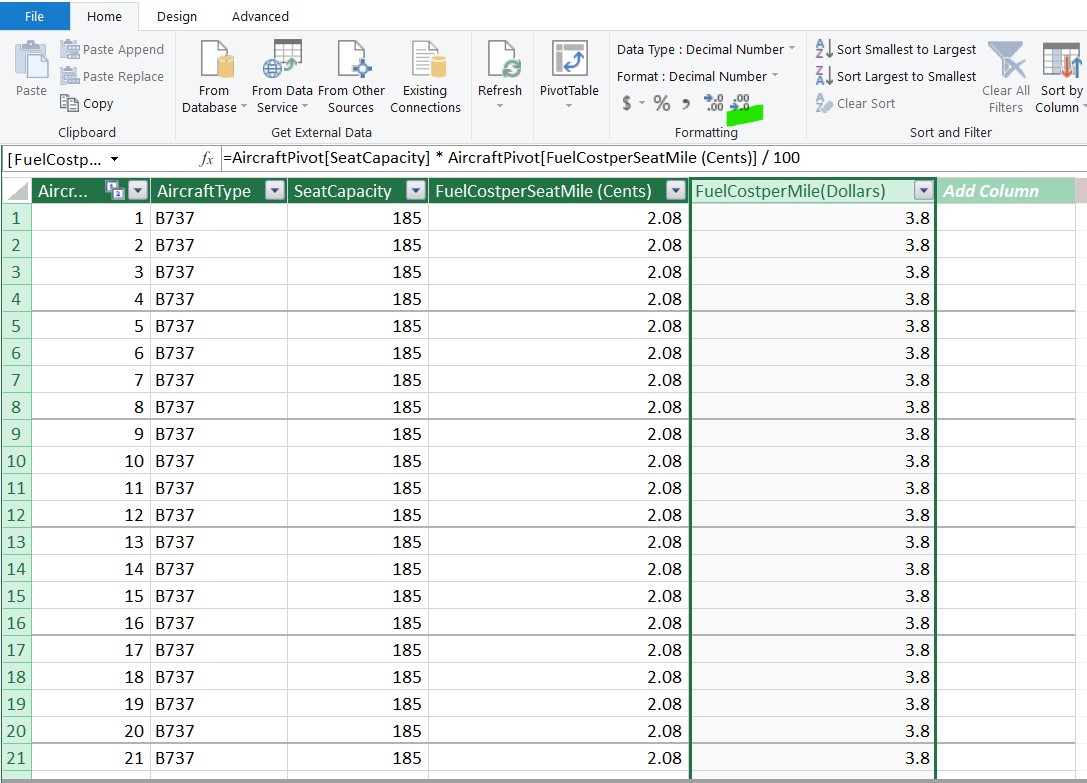
Remember, we want to multiply the SeatCapacity field by the FuelCostperSeatMile field. And then we need to divide this by 100 in order to get a value in dollars instead of cents. So, we're going to type equals and we'll select the SeatCapacity field which populates in our formula. Next, we input multiply and select the FuelCostperSeatMile field. Now we just need to add division by 100 and we are done with our formula.



Next, we need to provide a new name for the new field. Let's call it FuelCostPerMile(Dollars), since it's consistent with the naming of the other fields.

We're almost done with creating the Calculated Field. All that remains is making sure the field is formatted correctly.

To do this, let's navigate to the formatting tab. And select Decrease Decimal, which allows us to adjust the Calculated Field to two decimals.



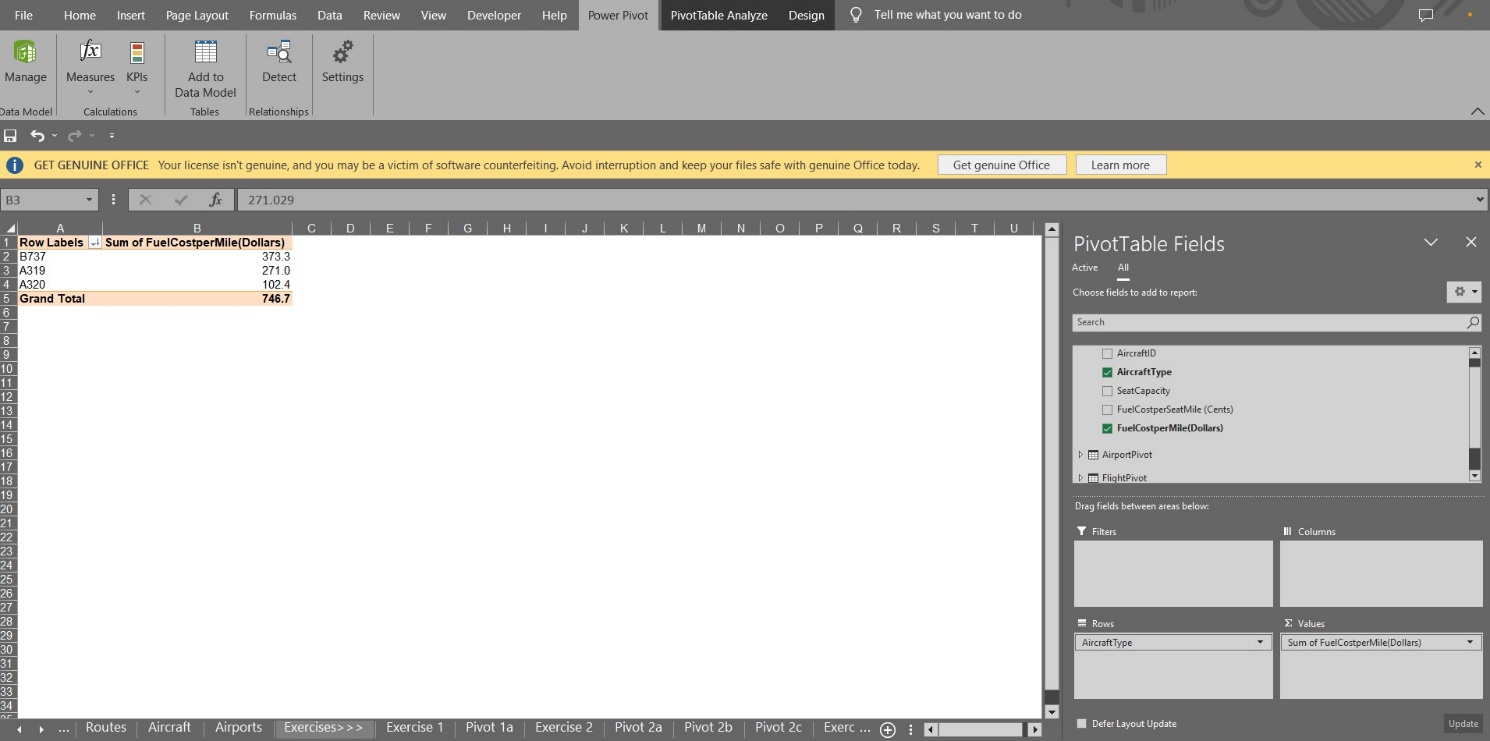
Now that we have created the new field which calculates the total fuel cost per mile for each aircraft, we are going to create a Pivot Table using this data to determine which aircraft has the highest fuel cost per mile. Let's create our Pivot Table in a separate tab, just like we did for exercise two. So, let's head to the tab named Pivot 3a.

We want to place the Pivot Table in the top left corner, so let's select cell A1 before navigating to the PowerPivot tab on the ribbon.

Here we select the Pivot Table option. Brings up a pop-up window where we're going to select the existing worksheet for the location of the table. And then we click OK to create the table.

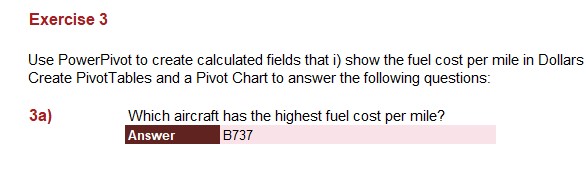
Next, we're going to drag our new Calculated Fields into the values section which provides us with a total sum of the cost across all aircraft types.

We want to know the cost for individual airlines, so let's drag the aircraft type into the rows section of the Pivot Table. Since, we want to know which aircraft has the highest fuel cost per mile, we're going to right-click on the Pivot Table, and select sort from largest to smallest in the menu.



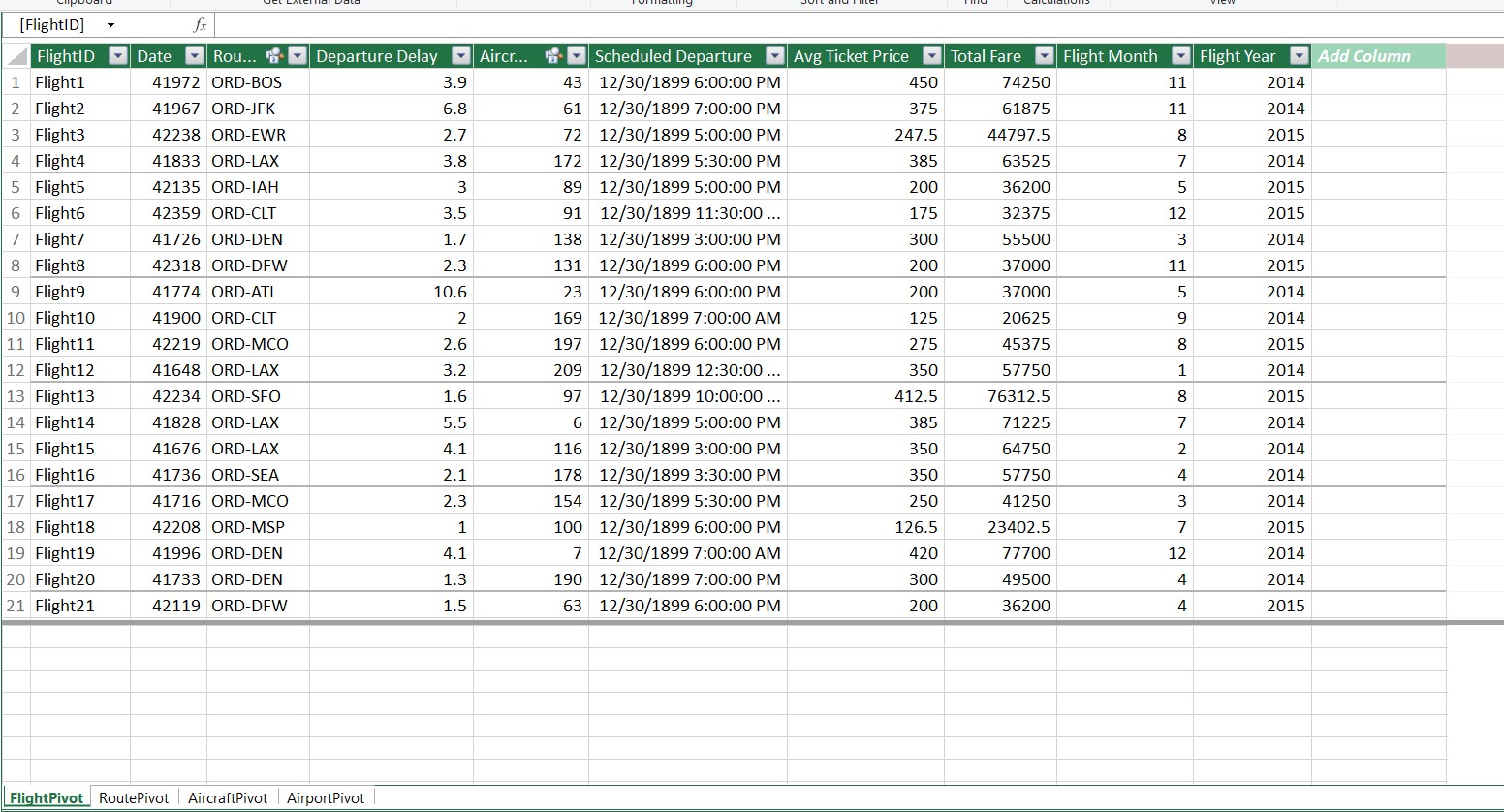
We can now see that the B737 aircraft has the highest fuel cost per mile, among the aircrafts flown by our company.

Let's go back to the exercise three tab and put that answer into cell G13.

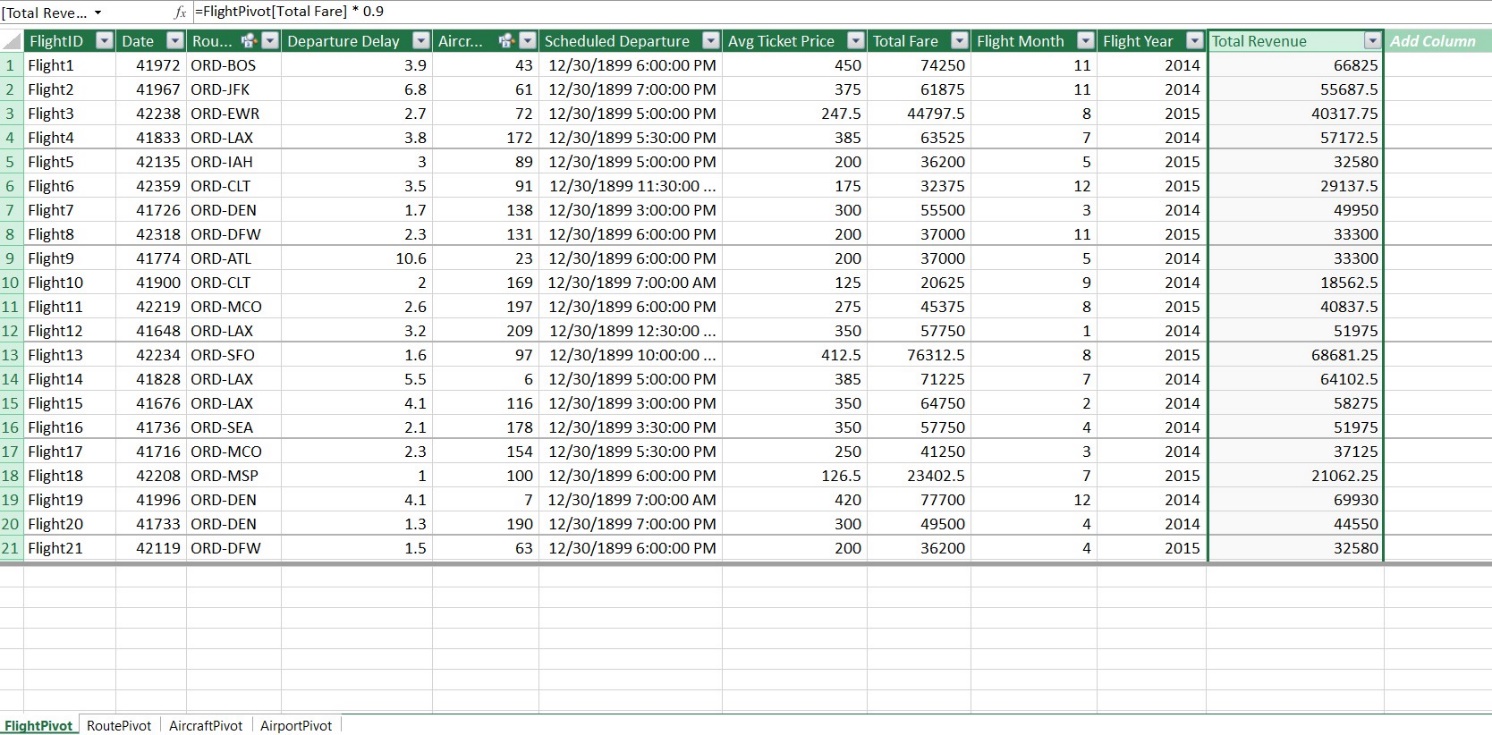


As part of the next question in exercise 3, we're being asked to determine which flights have the highest total revenue.

If we look at the Flights table, we notice that we have the total fare for each flight in column H. We are going to create a new Calculated Field that allows us to calculate the total revenue by applying a 10% tax to the total fares collected for each flight. So, let's head back to the exercise three tab and let's get started. Just like before, we start by navigating to the PowerPivot Tab, and selecting Manage. Within the data view section, there's a list of available data tables. Since we want to calculate the total revenue based on the total fares that can be found in the Flights table, we're going to select FlightsPivot from the list.

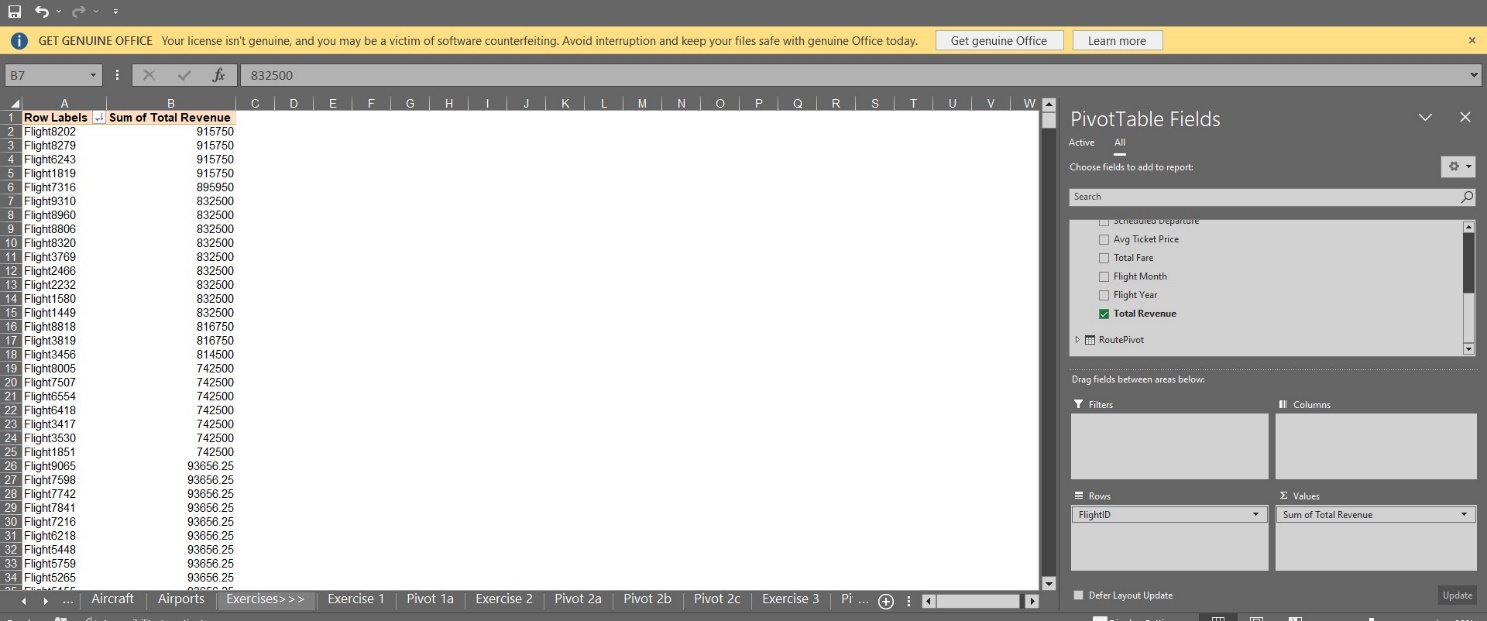


Next, we're clicking the formula bar on the right side of the window to start entering our formula. Remember, the total revenue is going to be equal to the total fares minus a 10% tax. Hence, we are going to multiply the total fare by 0.9. So, let's type equals and then we select the total fare field which populates in our formula. Next, we add multiple type .9 and press enter to close the formula.

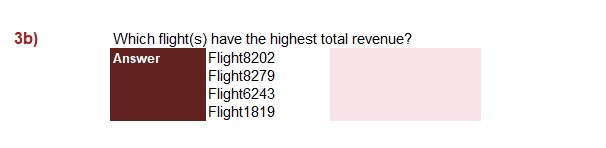


Then we are going to name this new field Total Revenue.

Now that we have created our Calculated Field, which provides us with the total revenue per flight, taking into account of the 10% new tax, we are going to create a pivot using this data in order to determine which flights have the highest total revenue. We're going to create our pivot table in a separate tab again, so let's head to the tab named Pivot 3B. We want to place the pivot table in the top left corner, so let's select cell A1 before navigating to the PowerPivot tab on the ribbon. Here we select the Pivot Table option. Which brings up a pop-up window where we're going to select Existing Worksheet for the location of the table and then we click OK to create the table. Next, we're going to drag the Total Revenue Field into the value section which provides us with a total sum of all revenue across all flights. But we want to know the total revenue on a flight-by-flight basis, so let's drag the Flight ID into the row section of our pivot table. As you can see this produces a large list of flights and their corresponding revenues. We want to see which flight has the largest revenue. So, let's right click on the pivot table and select Sort from Largest to Smallest in the menu.



We can now see that there are 4 flights that all have the highest revenue. So let's go ahead and copy these values. Next, we're going to navigate back to the Exercise 3 tab, and we're going to paste our answer into cell G17.



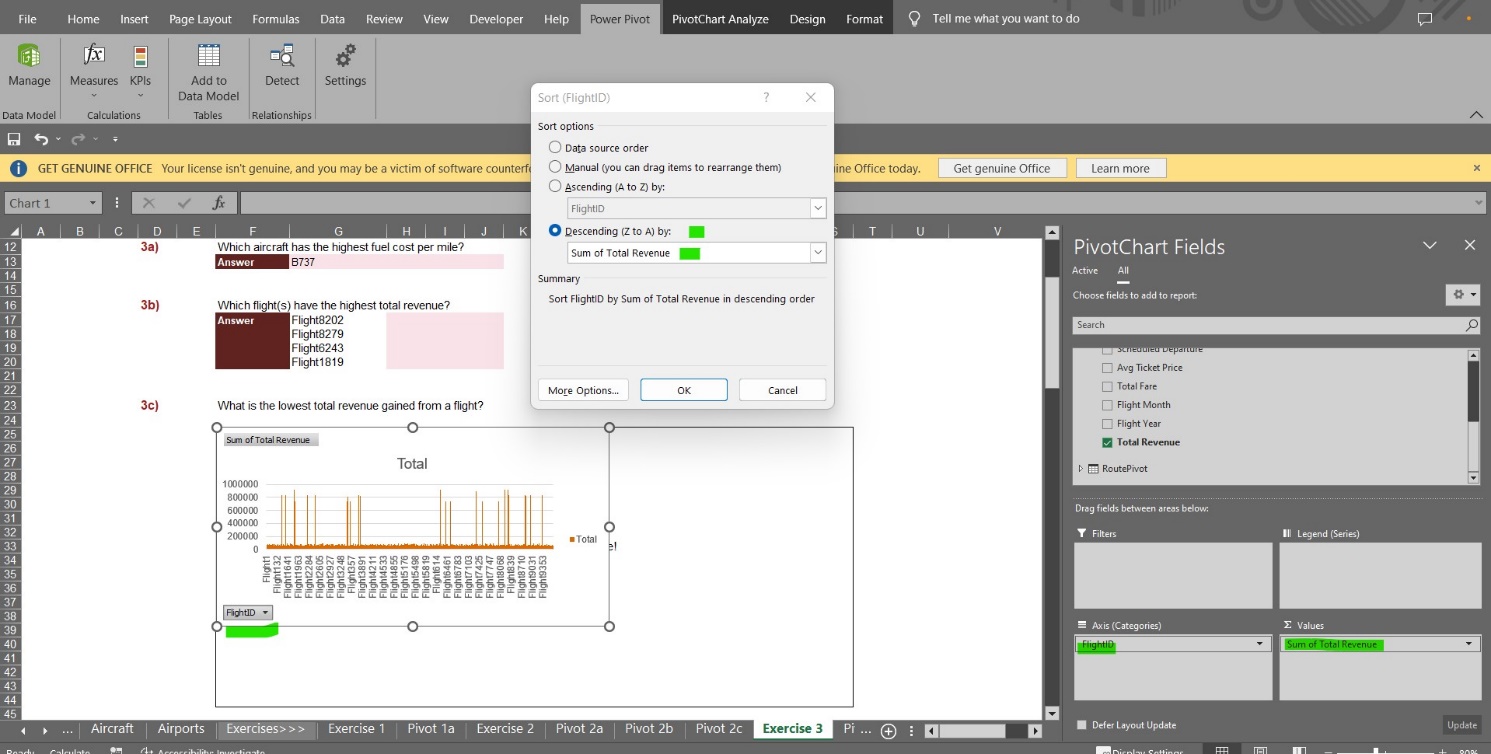
The final question in Exercise 3 asks us to provide a visual representation of the revenue distribution across all flights.

This chart is going to rely on the Total Revenue Field that we just created. So we don't need to create a new Calculated Field.

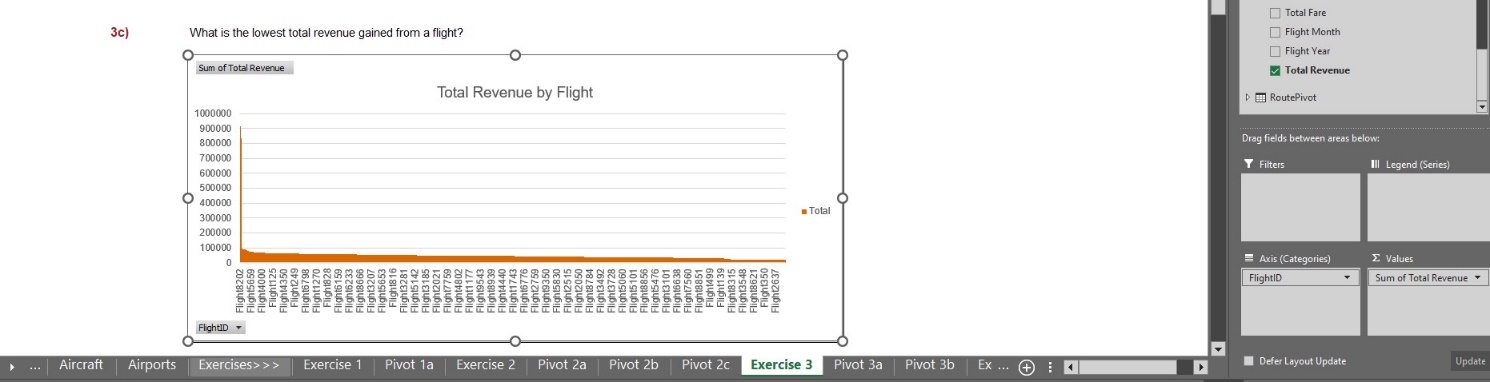
Instead of using this data to create a pivot table, we are going to use it to create a pivot chart. So let's get started.First,

similar to what we did with the pivot tables, we are going to center the cell that we would like to place our chart in. Here, we have been

provided with the chart area. Next, we navigate to the Power- Pivot tab on the ribbon. Select the PivotChart option. We are going to drive the flight attribute into our access section. And the total revenue attribute into a value section, since we are using this field as a value of measurement. The chart that we just created provides us with a basic view of the various revenue levels across flights. To sort this from largest to smallest, select the drop-down in the bottom left corner that says Flight ID. And select More Sort Options from the menu. This brings up a pop-up window that allows us to sort the data in the chart. Select Descending Z to A. And pick Sum of Total Revenue from the list. Click OK to confirm.



Finally, we are going to change the name of the chart by selecting the placeholder text, and replacing it with Total Revenue by Flight. We can now see that there are a few flights with very high total revenue amounts with a significant drop-off after this level.



We learn how to create and store Calculated Fields in PowerPivot and then use these new fields to create PowerPivots, and PowerCharts.

Multiple Data Tables and Filters:

Let's go ahead and navigate to the tab named Pivot 4a. In order to insert our PivotTable, we're going to go to the Powerpivot tab And select Pivot Table.

We want to insert the pivot on this tab. So let's select existing worksheet and then we are going to click on OK.

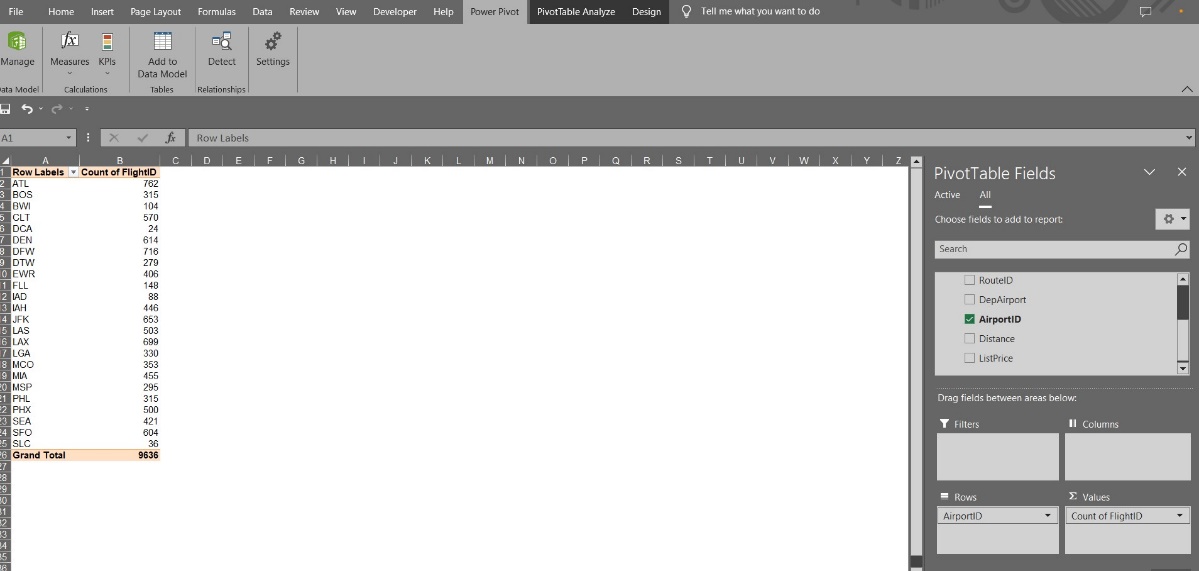
We're looking for the number of flights on a specific route. So, we're going to count the FlightIDs from the flights PivotTable since each FlightID represents a unique flight.

Let's go ahead and drag this field in the value section. Making sure to check that we are counting the values.

This gives us the total number of flights across all routes; however we just want to see the number of flights on the ORD LAX route.

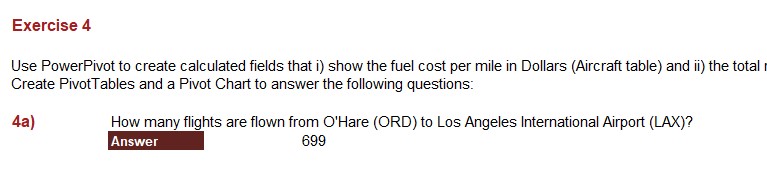
Remember, our routes table only contains flights leaving from Chicago.

This means that we focus purely on the arrival airport which is listed as the airport ID in the Routes PivotTable.



As you can see, we now have a pivot table that shows the number of flights for each of the airports that our company flies to. We're interested in the total number of flights going to LAX which can be found in row 16.

We're going to select this answer. Copy it. And then move back to the Exercise 4 tab where we paste answer into cell G13.



So, we’ve done question 1.

let's move to Question 4B. Here we're being asked which route accounts for the lowest percentage of total revenue? Remember how we created a calculator field for the total revenue in exercise three that took into account 10% tax. We're going to use that field here to determine the total revenue for each of the routes. So, let's head to the Pivot 4B tab and select PivotTable in the Power Pivot tab on the ribbon.

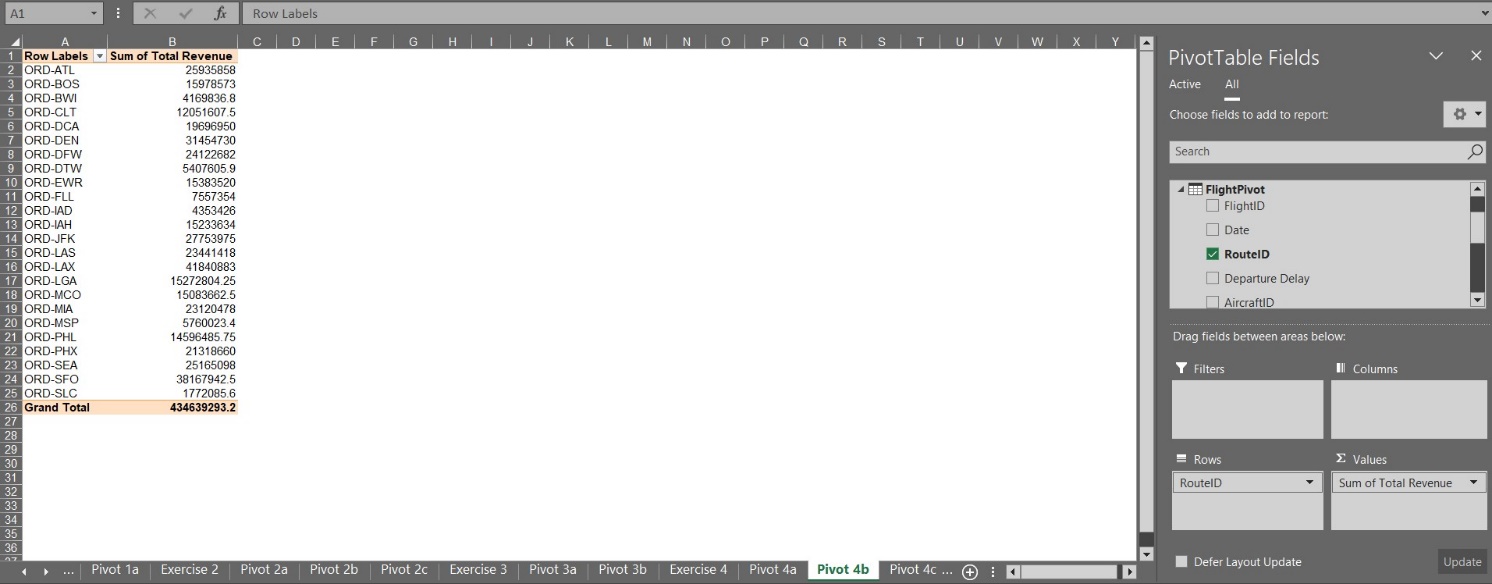
Just like before, we want to place the pivot in this sheet. So, we're going to select Existing worksheet and click OK.

Next, we're going to take the total revenue field from the flights pivot and drag it into the value section.

Make me show that we sum the values.

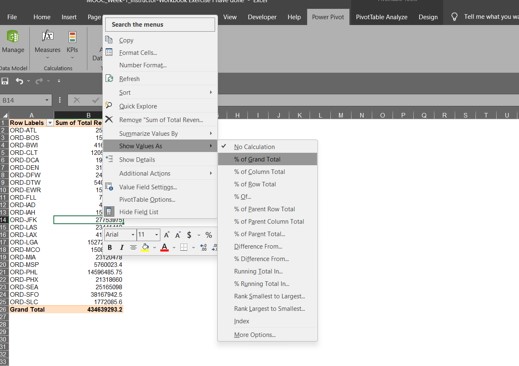
We want to look at the revenue on a route-by-route basis. So, let's grab the RouteID field from the Route PivotTable and drag it into the raw section.

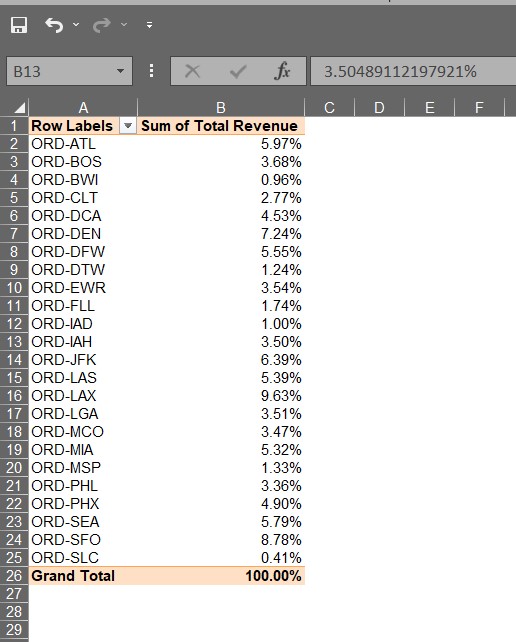
We now have a table showing the total revenue for each of the routes.



However, this does not yet answer the question since we are asked to determine which route accounts for the lowest percentage of total revenue.

So, let's start by changing the pivot table to show the revenue values are percentages. We do this by right-clicking on the PivotTable and selecting Show Values As followed by selecting Percentage of Grand Total. As we can see, the table now shows us what percentage of the total revenue is coming from each of the routes.



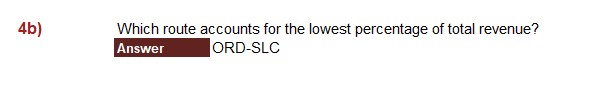


In order to determine the route that accounts for this lowest percentage, let's right-click on the PivotTable again and select Sort from lowest to highest.



Now we can see that the Chicago to Salt Lake City route is the one bringing in the least amount of revenue.

So, let’s go back to the exercise 4 tab and input the answer in the cell G17.



The next question asks us to find for the most popular month to fly to Fort Lauderdale or FLL as it is designated by its airport code. Answering this question is going to require the airport ID from Routes PivotTable along with the month information from the Flights PivotTable. Remember that the months in the table are stored as numbers with January being month number one and so on and so forth. Let's get started. Just like before we're going to place the PivotTable in the tab associated with the question. So Let's head to the Pivot4C tab.

As usual, we're going to select the PivotTable option from the Power Pivot tab on the ribbon.

Next, we select Existing Worksheet and then click OK to create the PivotTable.

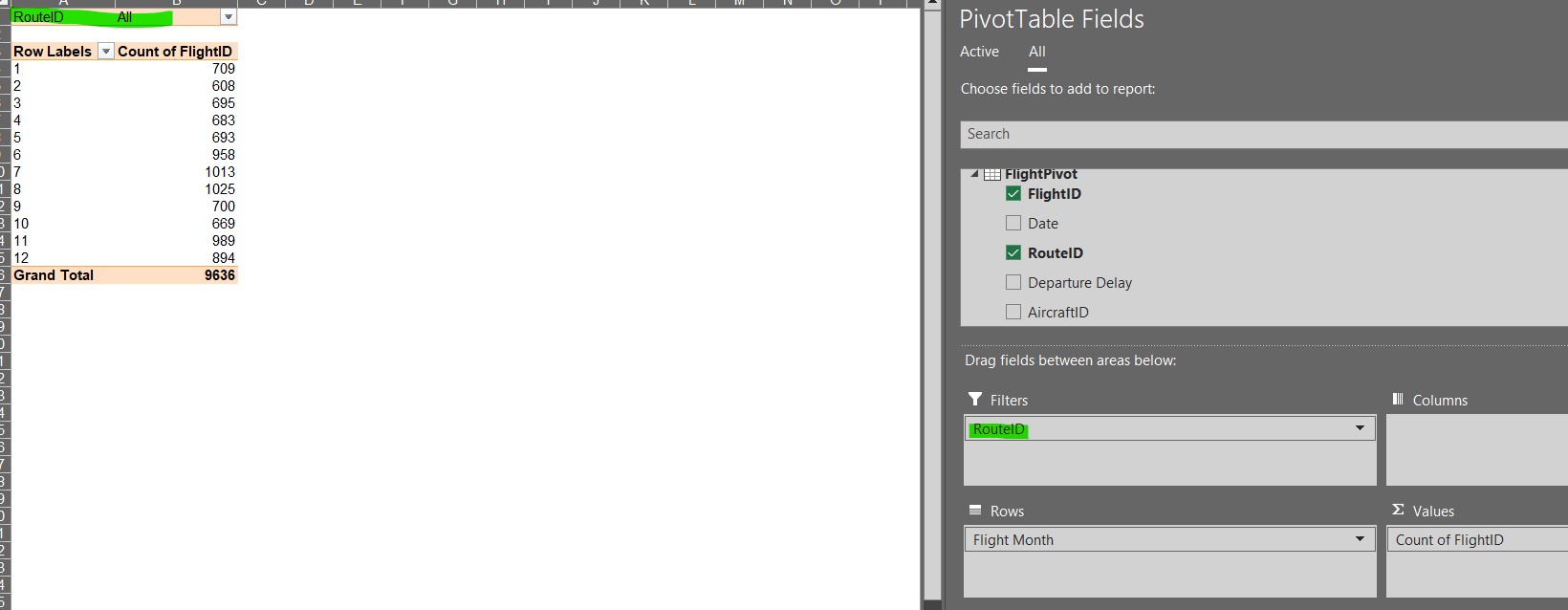
We want to count the number of flights heading to each destination. So, let's go ahead and select the FlightID field from the Flights Pivot And drag it into the value section, making sure that we're counting the IDs to get the number of flights.

We want to see the number of flights on a monthly basis. So, let's grab the months reel from the Flights Pivot and drag it into the row section.



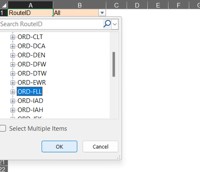
Our PivotTable is now showing us the number of flights for each month across all routes flown by the airline. Since we are trying to determine the most popular month specifically for the Chicago to Fort Lauderdale route. Wwe're going to go ahead and filter this table using the RouteID field from the Routes PivotTable.

Let's select the RouteID field and drag it into the filter section.

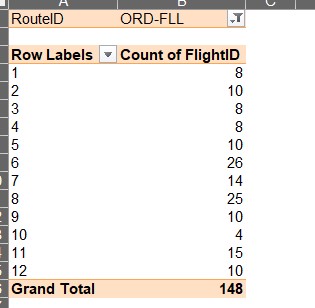


See how this adds a filter above our PivotTable.

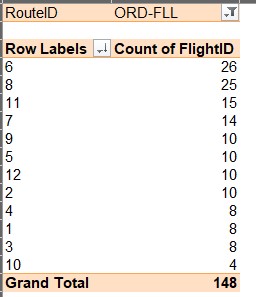
We're going to click on the filter we just created and select the ORD FLL route from the drop down.



Now our PivotTable is showing us the number of flights on the Chicago to Fort Lauderdale route on a monthly basis.

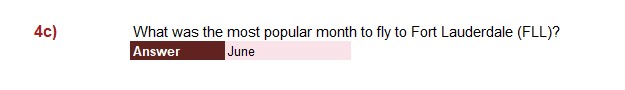


We want to know which month is the most popular. So, let's go ahead and solve the values in column B by right-clicking, selecting Sort and picking the Largest to Smallest option.



As we can see, month number 6 has the most flights.

Hence, we go back to the Exercise 4 tab and input that June was the most popular month in cell G21.



Finally, we're going to create a pivot chart to show the number of flights per month on the Chicago-Philadelphia route. We've been provided with the chart area where we are going to place this graph. So, let's click on cell F26.

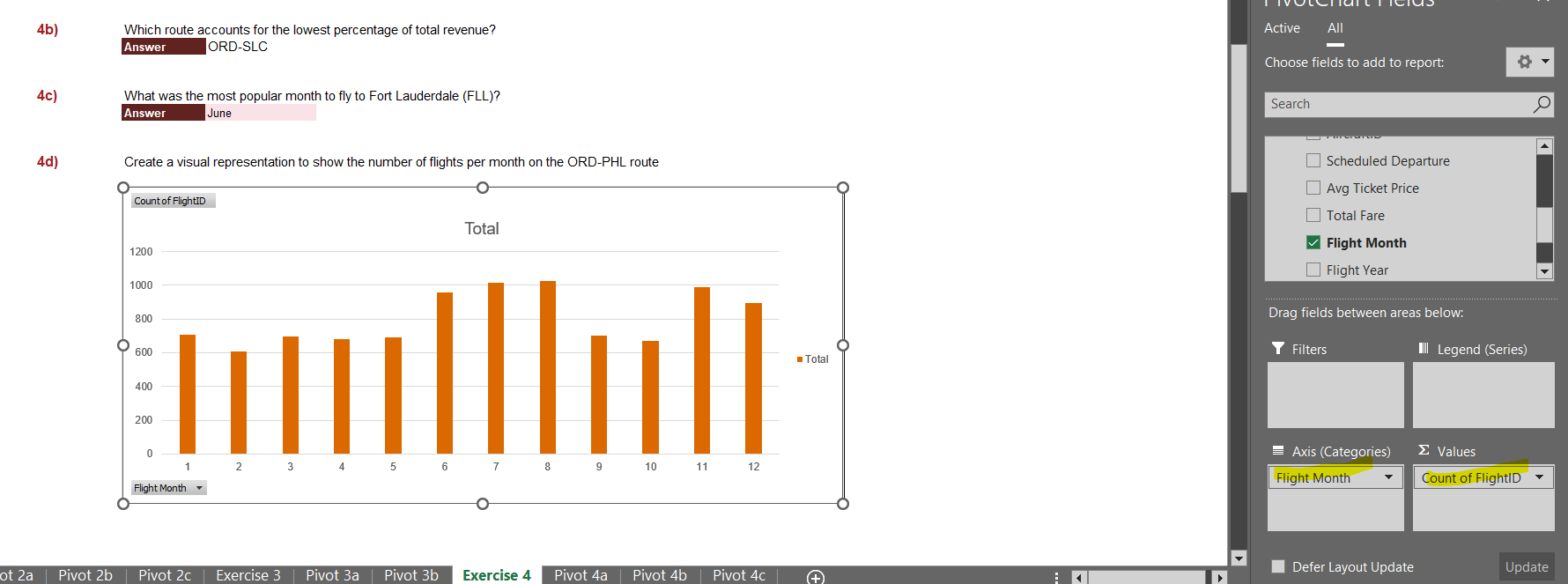
And navigate to the PowerPivot tab on the ribbon.

Where we're going to select the PivotChart option.

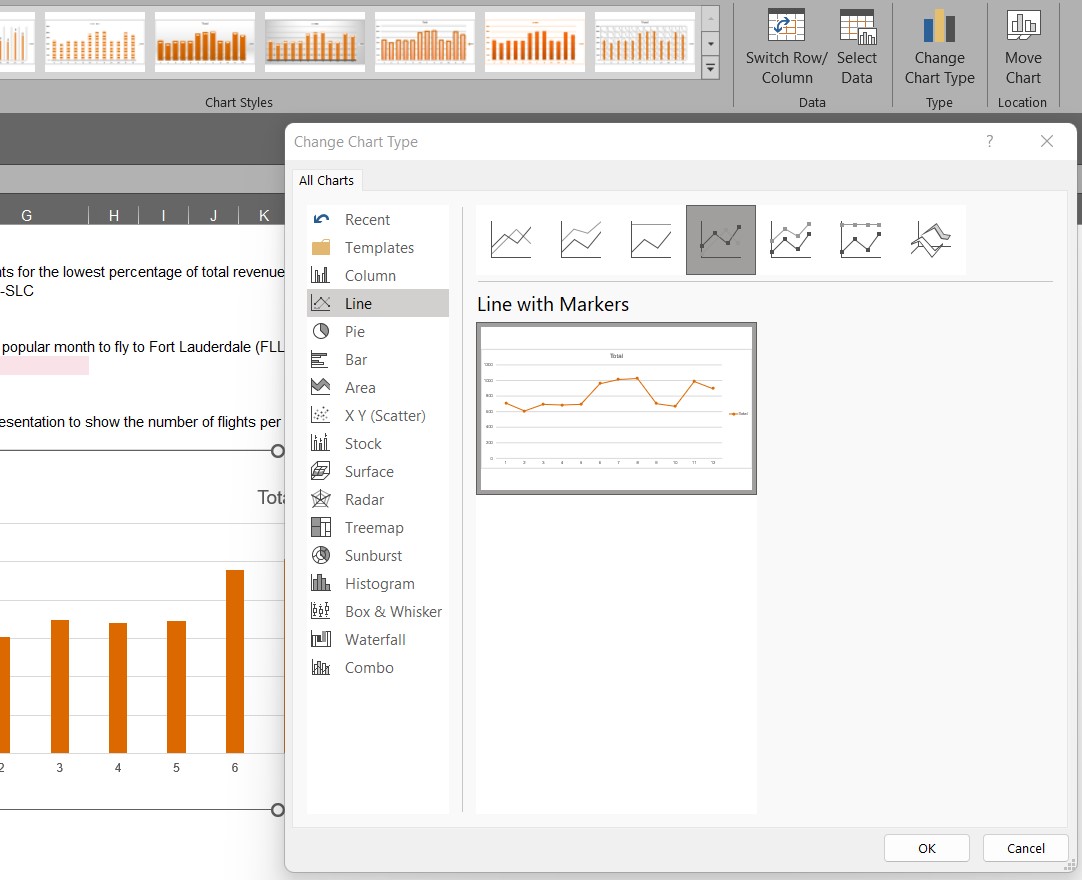
We want to show the number of flights per month. So, we are going to grab the month attribute from the Flights PivotTable.

And drag it to the Access section.

We also said the FlightID attribute from the same table and drop it into the values section, making sure we are counting the values.

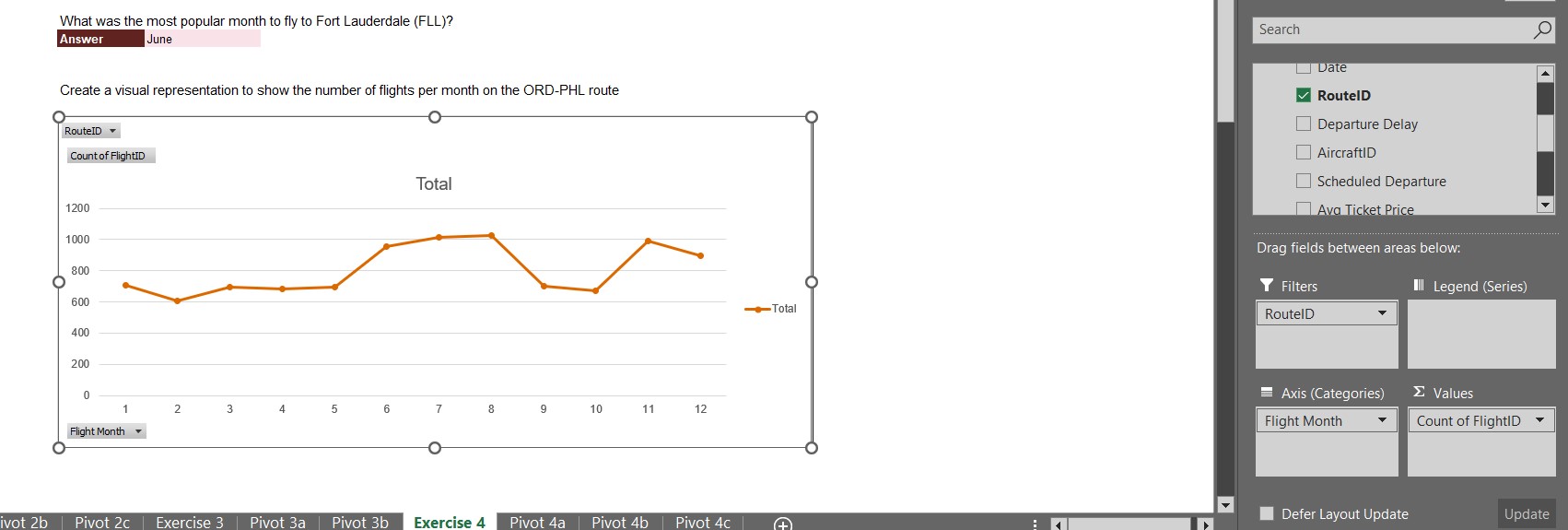


The chart we just created provides us with an overview of the number flights per month. But since we assuring data that changes over time, a line graph will be a more appropriate representation. So, let's go ahead and update the chart type. This is done by clicking on the Chart, go into the Design option and selecting Change Chart Type.

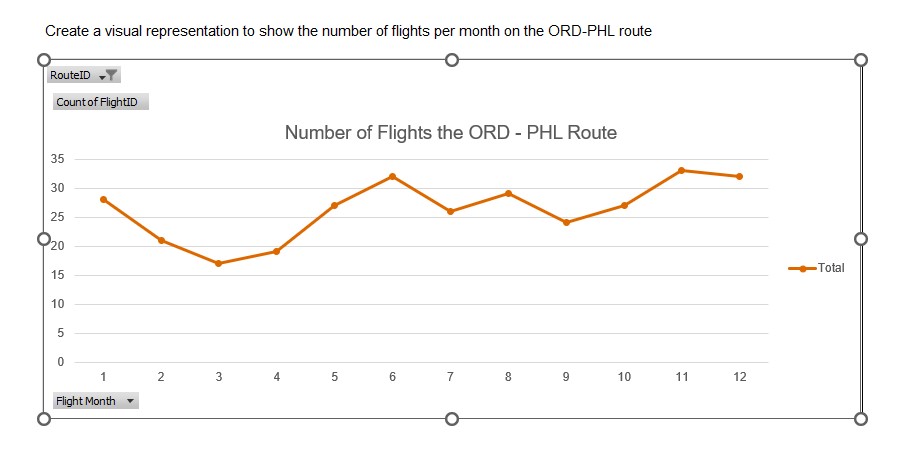


Remember, we only want to show the number of flights per month that are flying the Chicago/Philadelphia route. So, we are going to need to filter the data that is being plotted.

We do this by selecting the RouteID and dragging it into the Filter section.



As you see, this creates a filter in the upper corner of the graph. Let's click on this filter and select ORD-PHL.



Finally, we're going to change the name of the chart by selecting the placeholder text and replacing it with Number of Flights Flying the ORD- PHL Route.

We can now see how the number of flights on the Chicago Philadelphia route are distributed over the course of the year. **This route is clearly more popular in June and November, whereas March is the least popular month**.

We learned how to create Power Pivots and Power Charts using multiple data tables and filters.