## Introduction to PivotTables

* Create a PivotTable: Insert > PivotTable > Select data > OK
* The primary “Canvas” of the PivotTable is the Field menu. We will drag-and-drop to get data configurations
* Drag Name to Rows, then move over to Columns
  + This is a nice shorthand way to remove duplicates!
* Generally you use PivotTables not just to list categories but to *measure* something by those categories, that’s where the Values come in
  + Drag Sales Qty to Values.
  + We can do the same with Sales $. We might also want to format the results
  + If you want to rename these figures, simply type over them
* One of the coolest things about PivotTables is how easy it is to re-format it
  + Click anywhere inside the PivotTable to then head to the Design tab to change colors, add totals and so forth
* Drag Date to the Rows area of the PivotTable (below Name)
  + In general I’m not a fan of the “compact” PivotTable view, so let’s change it:
    - PivotTable Options > Display > Classic PivotTable layout.
    - Repeat the Names if you wish by right-clicking on any of them > Field Settings > Layout & Print > Repeat item labels.
    - Which Pivot style do you prefer? Which might be better for which circumstances?   
      Table

      Description automatically generated
* How can you “Pivot” this data by date? Play around with the different options here
* Next let’s experiment with Filters. Add Region to the Filters area. You have now worked with every quadrant of the PivotTable, nice work!
* Time to work with calculated fields!
  + Click anywhere in PivotTable > Analyze > Fields, Items, & Sets
  + Insert Calculated Field > $ per Unit = Sales $ / Sales Qty
  + Drag it to the PivotTable Values
  + There are also lots of ways to express these values as % of Grand or Row total, etc.
    - Right click on the value of interest > Show Values as …
* How can we maintain this PivotTable for future use? What if there is new data?
  + The default way is to right-click on the PivotTable > Refresh
  + But this will *only* update the data that’s within the *Data Source*
    - Check what this is with PivotTable Analyze > Change Data Source
    - The data source does *not* pick up our appended Week 2 data!   
      Graphical user interface, application, table, Excel

      Description automatically generated
  + A better way to store the data for PivotTables will be as *Tables*. Click anywhere inside the order data and select Ctrl + T.
    - You will see that the dataset is now nicely formatted in its own right – You can check out these options in the Table Design menu.
    - I am going to give the Table a sensible name like sales:  
      Table

      Description automatically generated
    - Now, go back into the PivotTable data source and change the reference to the table:   
      Graphical user interface, text, application, email

      Description automatically generated

Nice work creating, designing and modifying PivotTables and their data sources!

## PivotCharts and visualizations

Let’s first “hack” a PivotChart to make overlapping histograms:

* Drag price to Axis, cd to Legend and Count of Index to Value

Graphical user interface, application, table, Excel

Description automatically generated

* It’s going to look pretty bad right now, let’s fix that:
  + Right-click on the price along the Axis > Group > OK
  + Right-click any of the resulting bars on the PivotChart
    - Set Series Overlap to 100%
    - Set Gap Width to 0%
* Probably still need to adjust the groups in the PivotChart, go ahead and do that.

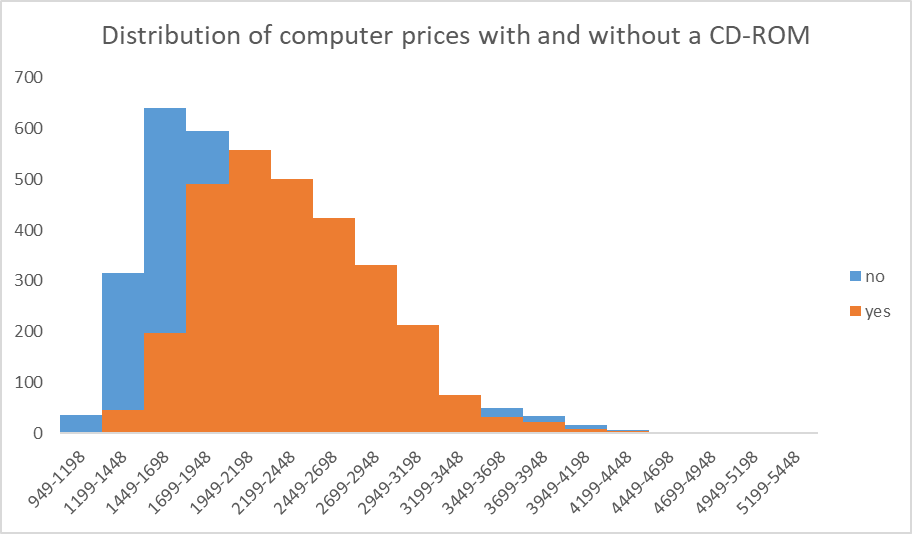
Your completed analysis should look something like this:

Chart, histogram

Description automatically generated

Right-click one of the labels and select “Hide all Field Buttons on Chart.” Also, click on any of the horizontal jailbars and press Delete on keyboard. We don’t need this chartjunk!

We should however have a chart title, so click on the plus sign on upper-right of chart, select Chart Title and give the chart a name:



Now let’s try a couple of PivotCharts using time series data.

* Insert a PivotChart based on the housing starts data.
  + Place Date on the Row Labels and Sum of starts on the Values
* By default this will roll up the data by Year.
  + We can “drill down” by right-clicking on the Date field and selecting Expand/Collapse > Expand Entire Field:  
    Graphical user interface, application, Excel

    Description automatically generated
  + You can also change the groupings by right-clicking and selecting Group. For example, this will aggregate the data by quarter only, regardless of year:   
    Graphical user interface

    Description automatically generated
* Go ahead and clear some of the chartjunk off this thing! Add a good chart title too.
* Now let’s create something to compare the percentage of total housing starts by quarter for each year.
  + Move Years to the Rows, Quarters to Columns (get rid of Dates) and Sum of starts to Values
  + Right-click any of the Values > Show Values As > % of Column Total
  + Insert > Recommended Charts > 100% Stacked Column:  
    Graphical user interface, application, table

    Description automatically generated
  + Clean up the chartjunk!
  + What happens if you right-click on the data > Select Data > Switch Row/Column?

PivotChart exercises

* When you create the bar chart of bills by day, you may notice that the days aren’t in a great order:

Graphical user interface, application, table

Description automatically generated

* There are a couple of ways to fix this:
  + First, you can type directly over the entries to put them in the desired order
  + The more elegant way will be to create a custom list:
    - Go to File > Options > Advanced > Edit Custom Lists (way toward the bottom)   
      Graphical user interface, text, application, email

      Description automatically generated
    - Go to NEW LIST > Add > then add your entries and click OK   
      Graphical user interface, application

      Description automatically generated
    - Now when you create the PivotTable it should automatically display the values in that order.

PivotChart exercises

Here are some extra hints and context along with the answers about the questions. First, let’s set up all the data and so forth to answer these questions:

* Convert Teams and Franchises into Tables – not totally necessary to complete this, but a good habit!
* Create the following columns in Teams that will help answer the PivotTable questions:

|  |  |
| --- | --- |
| Column name | Formula |
| >=1973? | =[@yearID]>=1973 |
| >=1903? | =[@yearID]>=1903 |
| FranchiseName | =XLOOKUP([@franchID], franchises[franchID], franchises[franchName]) |
| LgWin-01 | =IF([@LgWin]="Y", 1,  IF([@LgWin] = "N", 0, " ")) |
| WSWin-01 | =IF([@WSWin]="Y", 1,  IF([@WSWin] = "N", 0, " ")) |

1. Plot the following: the franchises with the 10 most World Series victories since 1903 versus those franchises' total number of World Series appearances.
   * **Warning:** While the Teams table does include a column called “Name” for the team’s name, this is *not* the same as the *Franchise* name. (For example, the Philadelphia Athletics and the Oakland Athletics are 2 different team names but the same franchise). Find the Franchise name in the TeamsFranchises table.
   * **Hint:** PivotTables do not allow for Label or Value filters if a filter is placed in the Filter area. But they do work in a Row or Column. Create your own filter by using conditional logic on the source data to flag the rows that are in 1903 or later.
   * **Hint:** A team only *appeared* in the World Series, but didn’t win it, if their LgWin value is set to Y but their WSWin value is N.

Create a PivotTable with WSWin equal to Y and >=1903? Equal to TRUE. Then place FranchiseName on the Rows, Sum of WSWin-01 (these are WS Wins) and LgWin-01 (these are WS Appearances) on Values.

Sort the data by WS Wins by clicking on Row Labels. You can also use this menu to get the top 10 values  
Chart

Description automatically generated

1. The late 1980s were a “golden age” for stolen bases. But shortly afterward, a new statistically driven approach to baseball became popular, as described in Michael Lewis’s *Moneyball.* One claimed effect of this shift was a drop in stolen bases. Is this actually true? Plot the average number of stolen bases per game from 1990 to 2017.

* Place yearID in Rows and filter >=1990, Add FranchiseName to filter and Avg Stolen Bases to the Values.

1. Do wild card teams ever win the World Series? Build a PivotTable that identifies the wild card teams that have won a league championship and/or the World Series. Break out your PivotTable by league, team name, and year.
   * **Hint:** Fields exist in the dataset to flag whether a team was a League winner or a World Series winner. Use this information to build your PivotTable, with the caveat that PivotTables can't calculate based on non-numerical data. What you can do, however, is use conditional logic to convert No’s and Yes’s to 0’s and 1’s. You can then use *this* data for PivotTable calculations.

Filter by WCWin = Y and LgWin-01 = 1. You can now add yearID, lgID and name to the Rows and Sum of LgWin-01 and Sum of WSWin-01 to the Values.

1. It has been speculated that the National League (NL) has had stronger pitching statistics ever since the 1973 introduction of the Designated Hitter in the American League (AL). Plot the average team ERA of AL teams and NL teams since 1973.
   * **Hint:** Because Filters in the Filter area do not allow for Row or Value filters, use conditional logic to flag the rows in the source data that are "since 1973."

Solution: Filter by the >=1973 = TRUE flag, set yearID to rows, lgID to columns and average ERA to Values. Visualize as a linechart. A lower ERA is desirable so this looks to be true, for the most part!

## Power Pivot & Excel Dashboards

Before doing anything else, make sure you see a Power Pivot tab on your home ribbon!

If you do not, you need to load it. [Follow these instructions to do so](https://support.microsoft.com/en-us/office/start-the-power-pivot-add-in-for-excel-a891a66d-36e3-43fc-81e8-fc4798f39ea8).

**Loading to Power Pivot**

We’ve got three workbooks in the data folder:

* orders.xlsx
* people.xlsx
* returns.xlsx

We want to load each of these into Power Pivot, via Power Query.

* Data > Get Data > From File > From Excel Workbook
* Navigate to the first file > Click OK
* We are *not* going to transform this in Power Query right now, so click Load > Load to:  
  A screenshot of a computer

  Description automatically generated with medium confidence
* Select Only Create Connection > Add this data to the Data Model > OK  
  Graphical user interface, text, application

  Description automatically generated
* Do the same for the other two tables.

**Viewing the data in Power Pivot**

* Go back to the Power Pivot tab and click Manage under Data Model
* We are now in the Power Pivot Editor. Let’s look around a bit.
  + First thing you will notice is that, like Power Query, while you can click around and see your data, you can’t just type over it.
  + Unlike Power Query, you should generally be making *minimal changes* to your tables in Power Pivot.
    - One difference is formatting the data. It’s nice to do it in Power Pivot because the PivotTables will then match that formatting.
    - For example, we can format
    - To learn the difference between formatting data in Power Pivot versus defining data types in Power Query, [check out this post](https://excelguru.ca/do-data-types-matter-in-power-query/)

Graphical user interface, application, table, Excel

Description automatically generated

Now we’ll go to Diagram View, this will show a *schema* of the database. Drag the relationships between these variables to create them.

Graphical user interface

Description automatically generated with low confidence

Fact vs dimension tables. The dimension tables are kind of like lookup tables.

Still in Diagram View, let’s click on PivotTable up at the top:

Graphical user interface, application, Word

Description automatically generated

We have a Power PivotTable!

It’s going to look very similar to a regular PivotTable, here are some big differences:

* We can use data from multiple tables on the same Pivot! For example, drag Person and Sum of Sales onto the PivotTable. Notice that Sales is even pre-formatted for you, nice!   
    
  Graphical user interface

  Description automatically generated
* One particular measure that is NOT available in a regular PivotTable but is in the Power PivotTable is the Distinct Count. To understand the difference let’s drag Count of Order ID vs Distinct Count of Order ID to the PivotTable. Distinct Count is down at the bottom. What is the difference?  
    
  Graphical user interface, application

  Description automatically generated

**Implicit vs Explicit Measures**

While using the PivotTable to change aggregation types is very familiar, you do NOT want to do this in Power Pivot. Instead, we are going to do this “explicitly” writing our own explicit measures. And this is where DAX comes in.

Go back to the Power Pivot editor and click Calculation Area on/off a couple of times. This area at the bottom is actually like a workbook that you can write formulas in!  
  
Graphical user interface, application, table, Excel

Description automatically generated

To see this in action, let’s go back to the Sales column in orders. We will go to the cells below this column and use the AutoSum feature for right now. Let’s take the Average of Sales:

Graphical user interface, application, table, Excel

Description automatically generated

We can go over to Quantity and take the Sum as well.

This is like the “Macro Recorder” way of creating these measures. It’s easy, but not always scalable and extensible. For example, it’s considered a better practice to *explicitly* state the name of the table where a column is coming from in the formula, with AutoSum does not do.

Let’s fix that. You will see this notation is identical to Excel tables.   
  
Graphical user interface, application, table, Excel

Description automatically generated

**Creating calculated measures**

Close out of the Power Pivot editor and head back to your PivotTable. You should now see these explicit measures in your Power PivotTable. You could format Quantity to thousands if you’d like:

Graphical user interface, application

Description automatically generated

You can learn more about these measures by going to the Power Pivot tab of the ribbon > Measures > Manage Measures. Click New:

Graphical user interface, text, application, Word

Description automatically generated

Let’s create a Profit margin measure:

Graphical user interface, application

Description automatically generated

You should be starting to see something like this:

Graphical user interface, application

Description automatically generated

**Time intelligence**

This is all great to know, but to really dig into most business insights we want to be able to compare across time periods. Power Pivot makes this relatively easy.

Let’s go back to Diagram View in the editor, then go to the Design tab of the ribbon. Click on Date Table > New:

Diagram

Description automatically generated

We will now create a relationship between Order Date in orders and Date in Calendar. Your schema should now look like this:

Graphical user interface, application

Description automatically generated

Go back to your Power PivotTable. We are now going to create some different date measures to analyze this data. To make best use of time and date functionality in Power Pivot, you should have a date table.

Go back to the Create Measures area and create a Total Sales measure first:

Graphical user interface, text, application

Description automatically generated

We will now create a MTD Sales column:   
  
Graphical user interface, application

Description automatically generated

Clear out your existing PivotTable; let’s add Date Hierarchy on the Rows and Total Sales MTD in the Values:

Graphical user interface, application, table, Excel

Description automatically generated

Expand the years out to see sales each month, day and so forth.   
  
Graphical user interface, application, table, Excel

Description automatically generated

Let’s create a YTD now:

Graphical user interface, application

Description automatically generated

We can also get a measure to compare the current month or year to the previous month or year:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Drag those into the PivotTable and you should see something like this:­­­­­

Graphical user interface, application, table

Description automatically generated

### Additional Power Pivot analysis & visualization with nycflights13

### Additional Power Pivot analysis & visualization with nycflights13

First step will be to create the data model using the four tables provided. You can see a list of all the relationships in the data model by going to Design > Manage Relationships.

I am also going to add a flight\_id number in the flights table by using the PQ editor.

Graphical user interface, application

Description automatically generated

Insert the 3D Map

By default, all the data loaded to the Data Model will be available to use in the 3D map.

* Insert > 3D Maps > New Tour

To create the map, you *must* place lat and lon in the location fields. I am going to drag average distance to Height and then tzone to Category

You can see if you hover over the points of the map you can see some information about the datapoints. I am going to customize these with the Data Card:

* Click under Layer Options > Data Card > Customize
* I am going to add Destination (name) and clean some other stuff up on this data card/tooltip like so:  
  Graphical user interface

  Description automatically generated
* I’d like to format the average distance a little nicer so it’s not rounded to all these crazy formats. This is why it’s important to use *explicit* measures!
  + Go back to the
  + This chart will be saved in the 3D Maps area as a “Tour.” Click on Tour 1 to the left to rename it something else like “Average Distances”
  + Now create an explicit average distance measure:   
    Graphical user interface, text, application, email

    Description automatically generated
  + Go back to the map and use this measure in Height. You can also make some tweaks to this “tour,” for example I am going to place the count of overall flights in the Tooltips.
  + What other Tours can you think of making here?
    - These maps can even be animated – drag time\_hour to the Time area for that.

## Cancellation analysis

The first step is to create a 1/0 flag indicating whether the flight was cancelled. If you look at the flights data, you will see that some records have a missing value for flight arrival time. I am going to assume this means that these flights never arrived, i.e. were cancelled!

Creating the cancellation column

* While it’s possible to create calculated columns in Power Pivot, generally Power Query is going to the faster, more stable option. So go to Data > Queries & Connections > Flights query > Edit
  + Create a flight\_cancelled column: Add Column > Conditional Column. If the arrival time is null, then flag as 1, otherwise it’s a 0  
    A picture containing graphical user interface

    Description automatically generated

There are lots of interesting analyses that you could perform concerning the cancellation rates. For example, this plot visualizes cancel rates over time by originating airport:

Chart, line chart

Description automatically generated

Looks like it’s not just your data -- [LaGuardia is *not* the place to be for winter travel](https://www.thrillist.com/news/new-york/worst-airports-us-ranked-canceled-flights)!

One thing that I was particularly interested in was whether regional airlines appear to have a higher cancellation rate – I created a custom TRUE/FALSE flag in the airlines table to group the records by regional/non-regional airlines and got the following results:



It *definitely* looks like more regional than non-regional flights get cancelled. But how can I confirm this using the power of statistical inference?

I am going to use the Chi-square test which will compare the actual vs expected values in the data for a significant difference.

* The actual values can be found by counting up the number of Flight IDs that were cancelled and not cancelled for the regional and non-regional carriers:

Graphical user interface, application

Description automatically generated

Next, copy-paste Values directly beneath the “live” PivotTable and use formulas to derive what these categories would be with just a straight proportion. You can then use the CHISQ.TEST() function to compare the two sets of data:

Graphical user interface, text, application

Description automatically generated

The p-value < .05 indicates that with 95% confidence, there is a difference in cancel rates b/w regional and non-regional carriers.

If you would like to learn more about how to add an extra level of statistical rigor to your thinking, be sure to take the Business Statistics in Excel course!