

Intermediate Excel Workshop
Empirical Reasoning Center
erc.barnard.edu
erl@barnard.edu
May 3, 2017

Co-Sponsored by Barnard Career Development

1 Introduction

This guide has been designed to accompany the Intermediate Excel workshop at the Empirical Reasoning Center. The example dataset was compiled by Walt Hickey at fivethirtyeight.com and contains information on 1,794 films released from 1970 to 2013. His article, “The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women” examines the budgets and revenues of films that pass the Bechdel test. The Bechdel test is a popular method of measuring how female-friendly a movie is. To pass the test: 1) there must be two named female characters, 2) the two women must talk to each other, and 3) the conversation cannot be about a man.

This workshop builds off of the Introduction to Excel Workshop. For review consult the introductory step-by-step guide.

The topics covered in this workshop include:

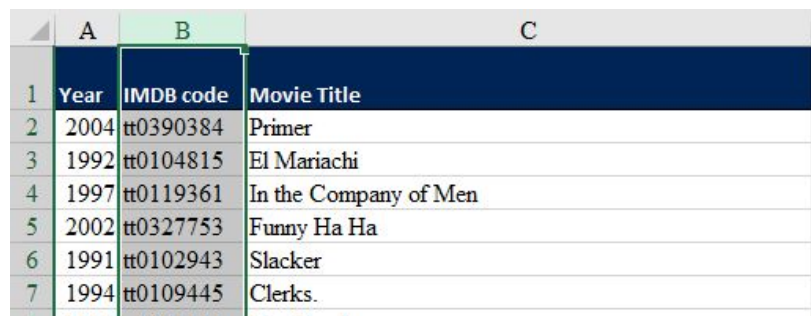
- Deleting & Inserting columns
- Sorting & Filtering data
- Searching with VLOOKUP
- Conditional statements
- Nested formulas
- Pivot Tables

2 Deleting Unnecessary Variables

When working with large datasets in Excel it is very important to stay organized. Spreadsheets can become difficult to manage when you have more columns than you can see at once. Therefore, it is useful to delete columns/variables when you know you will never need them.

2.1 Delete IMDB code variable

We have no need for the IMDB code variable in column B. To delete the entire column at once we have to highlight the entire column. To do this, click on the column letter, literally the B. You should then see the entire column is selected.

A screenshot of an Excel spreadsheet. Column B is highlighted in green. The spreadsheet has three columns: A (Year), B (IMDB code), and C (Movie Title). The rows are numbered 1 through 7. The data in column B includes IMDB codes for various movies.

	A	B	C
1	Year	IMDB code	Movie Title
2	2004	tt0390384	Primer
3	1992	tt0104815	El Mariachi
4	1997	tt0119361	In the Company of Men
5	2002	tt0327753	Funny Ha Ha
6	1991	tt0102943	Slacker
7	1994	tt0109445	Clerks

Figure 1: Select the IMDB Code Column

Now, in the Home tab, you should see a “Delete” button in the “Cells” group. You cannot press backspace or delete on your keyboard.

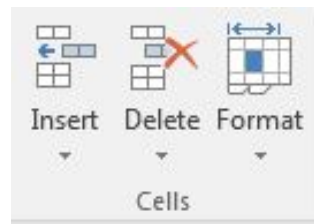
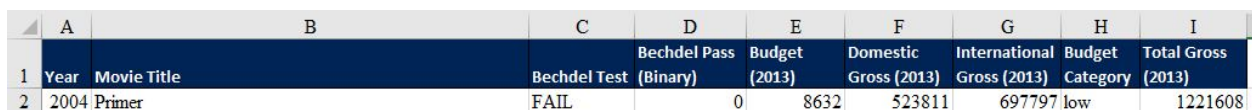


Figure 2: Delete

Then you should see the IMDB code column disappear and the Movie Title variable move to the left into column B.

A screenshot of an Excel spreadsheet after deleting column B. The columns are now A (Year), B (Movie Title), C (Bechdel Test), D (Bechdel Pass (Binary)), E (Budget (2013)), F (Domestic Gross (2013)), G (International Gross (2013)), H (Budget Category), and I (Total Gross (2013)). The data in column B is now the movie titles.

	A	B	C	D	E	F	G	H	I
1	Year	Movie Title	Bechdel Test	Bechdel Pass (Binary)	Budget (2013)	Domestic Gross (2013)	International Gross (2013)	Budget Category	Total Gross (2013)
2	2004	Primer	FAIL	0	8632	523811	697797	low	1221608

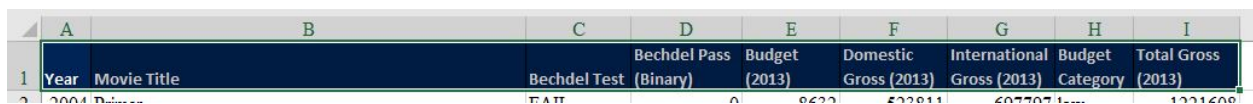
Figure 3: Data Without IMDB Code

3 Sorting

There are many reasons why you may want to sort your data. You may want a list of alphabetized names, or you may want observations organized by date so you can enter more variables.

Currently, the dataset is sorted by budget. We are going to sort this data set by year and then movie title. When you sort on two variables, the dataset is first sorted according to the first variable. But what happens when two observations have the same value? If we sort by year then there are five movies from 1971, what order should they be listed in? By adding a second variable, you tell Excel exactly what to do. For all of the movies released in the same year, we want them alphabetized by Movie Title.

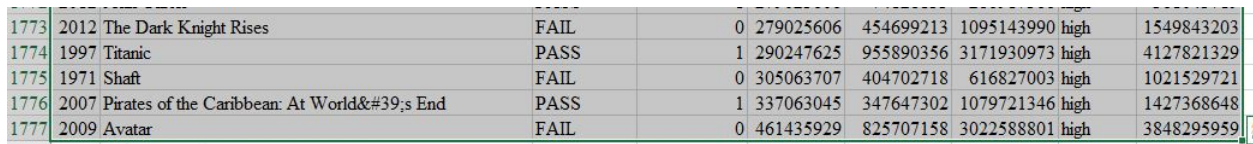
To do this, we need to highlight all of the data including the variable names. Any cells that are NOT selected will remain in the exact order they are now. There are many ways to highlight the data. We will click on cell A1. Then we will press CTRL + Shift + (the right arrow). This will highlight the cells from A1 to J1.



	A	B	C	D	E	F	G	H	I
1	Year	Movie Title	Bechdel Test	Bechdel Pass	Budget	Domestic	International	Budget	Total Gross
			(Binary)	(2013)	(2013)	Gross (2013)	Gross (2013)	Category	(2013)

Figure 4: Highlight the First Row

Now we want to highlight all of the rows below those cells. Now press CTRL + Shift + (the down arrow). Now all of the data should be selected.



1773	2012	The Dark Knight Rises	FAIL	0	279025606	454699213	1095143990	high	1549843203
1774	1997	Titanic	PASS	1	290247625	955890356	3171930973	high	4127821329
1775	1971	Shaft	FAIL	0	305063707	404702718	616827003	high	1021529721
1776	2007	Pirates of the Caribbean: At World's End	PASS	1	337063045	347647302	1079721346	high	1427368648
1777	2009	Avatar	FAIL	0	461435929	825707158	3022588801	high	3848295959

Figure 5: Highlight All of the Data

Now we want to sort the data. In the Home tab, in the Editing group, click on the arrow next to the “Sort & Filter” button. Then choose the “Custom Sort” option.



Figure 6: Sort & Filter

You should see the following pop-up window.

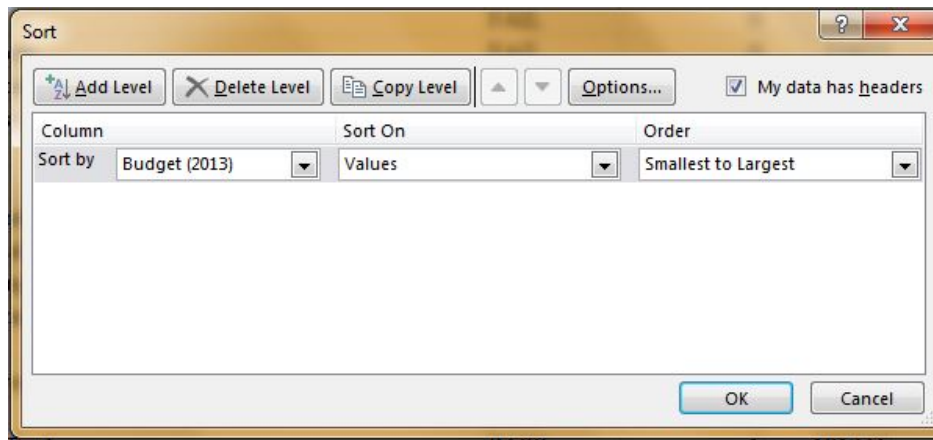


Figure 7: Sort & Filter Pop-Up Window

The first thing you want to do is check “My list has headers” so your variable names are not sorted like an observation. When alphabetizing movie titles, you don’t want to move the “Movie Title” variable name to the Ms; it should stay in row 1.



Figure 8: My List Has Headers

Then you want to choose “Year” under the Column drop down menu. Then you want to click the plus sign to add a second variable to sort by. In the second row, choose “Movie Title” under the Column drop down menu. Then press “OK.”

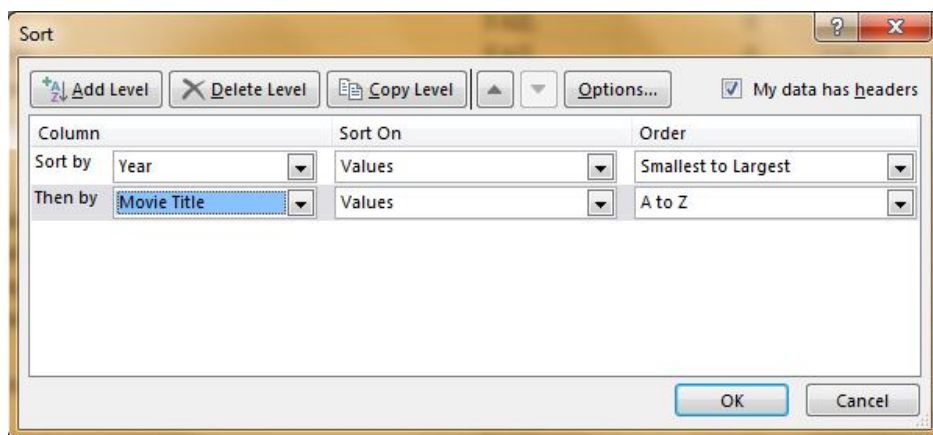


Figure 9: Sort by Year and Movie Title

Now you should notice that the dataset has been sorted.

4 Conditional Statements

A conditional statement is an if-then statement. That is, if a certain condition is met, do A; if that condition is NOT met, do B. This is often how binary variables are made.

4.1 Profitable Variable

We are interested in differentiating between profitable and unprofitable movies. We are going to create this binary variable using the “if” function. A movie was profitable if the total gross revenue is greater than or equal to the budget; otherwise, the movie was unprofitable.

We are going to name this variable in cell J1 – “Profitable.” In cell J2 we are going to enter the formula. This is a function, so it starts with an “=.” The function is called if, so the formula starts “=if(” and should display a helpful pop-up. This function requires 3 inputs: the condition, what to display if the condition is met, and what to display if the condition is not met.

The first input is the condition, which is that total gross is greater than or equal to the budget. So the formula is now “=if(I2>=E2,” with a comma to separate each input. The next input is called “value if true,” so what should this variable equal if cell I2 exceeds or equal cell E2. Because we want this to be a binary variable, the value 1 indicates true. So the formula is now “=if(I2>=E2,1,” with another comma. The last input is called “value if false,” so what should this variable equal if cell E2 exceeds cell I2. In binary variables, the value 0 indicates false. So the final formula is “=if(I2>=E2,1,0)” and press enter to see the first result.

	A	B	C	D	E	F	G	H	I	J
1	Year	Movie Title	Bechdel Test (Binary)	Bechdel Pass	Budget (2013)	Domestic Gross (2013)	International Gross (2013)	Budget	Total Gross (2013)	Profitable
2	1970	Beyond the Valley of the Dolls	PASS		5997631	53978683	53978683	low	107957366	=if(I2>=E2,1,0)

Figure 10: Profitable Formula

Now we want to apply this formula to the rest of the column. To do this, select cell J2 and move the cursor to the bottom right corner of the cell. When the cursor looks like a small, black cross, double click the mouse. This applies the formula until the next blank cell of data.

J
Profitable
1
1
1
0
1
1

Figure 11: Profitable Formula

5 Inserting a Variable

While, the Profitable variable identifies which movies were profitable, it does not provide information on the size of the profit. For that, we would need to calculate the profit. However, I want the profit variable to be between the Total Gross and Profitable variables.

I will do this by inserting a blank column to the left of the Profitable variable. To do this highlight column J by clicking on the J.

	A	B	C	D	E	F	G	H	I	J
1	Year	Movie Title	Bechdel Test	Bechdel Pass (Binary)	Budget (2013)	Domestic Gross (2013)	International Gross (2013)	Budget Category	Total Gross (2013)	Profitable
2	1970	Beyond the Valley of the Dolls	PASS	1	5997631	53978683	53978683	low	107957366	1
3	1971	Escape from the Planet of the Apes	FAIL	0	14386286	70780525	70780525	low	141561050	1
4	1971	Shaft	FAIL	0	305063707	404702718	616827003	high	1021529721	1
5	1971	Straw Dogs	FAIL	0	143862856	59412143	64760273	high	124172416	0
6	1971	The French Connection	FAIL	0	12659931	236848653	236848653	low	473697306	1
7	1971	Willy Wonka & the Chocolate Factory	FAIL	0	17263543	23018057	23018057	medium	46036114	1

Figure 12: Highlight Profitable Column

Then in the Home tab, in the Cell group, click on the “Insert” button.

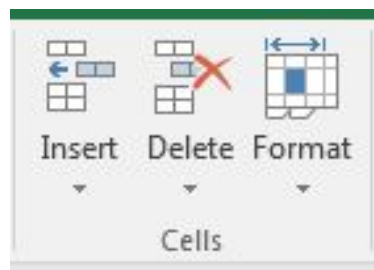


Figure 13: Insert

You should see a blank column J and the Profitable variable now in column K.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	Movie Title	Bechdel Test	Bechdel Pass (Binary)	Budget (2013)	Domestic Gross (2013)	International Gross (2013)	Budget Category	Total Gross (2013)		Profitable
2	1970	Beyond the Valley of the Dolls	PASS	1	5997631	53978683	53978683	low	107957366		1

Figure 14: Inserted a Blank Column

5.1 Calculate Profit

Now in column J we want to calculate a Profit variable. In cell J1 label the variable “Profit.” In cell J2 enter the formula “=I2-E2” and apply it to the rest of the column.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	Movie Title	Bechdel Test	Bechdel Pass	Budget	Domestic	International	Budget	Total Gross	Profit	Profitable
2	1970	Beyond the Valley of the Dolls	PASS	(Binary)	(2013)	Gross (2013)	Gross (2013)	Category	(2013)		
					1	5997631	53978683	53978683	low	107957366	=I2-E2

Figure 15: Calculate Profit

6 Nested Formulas

What if you need to satisfy two conditions? For example, categorical variables with more than two categories would require two conditions. As well, some definitions of binary variables would require two conditions.

It is easier to make a profitable movie when the budget is low. So, we are interested in making a stricter definition of profitable – “successful.” The successful variable should equal “successful” if the movie was profitable and the budget category is not low, “unsuccessful” otherwise.

Start by labeling this variable in cell L1, “successful.” In cell L2, we will start the formula with “=if(” however we need to specify that we have two conditions. To do this, we will nest another function inside this “if” function. The “And()” function allows us to list two conditions that both must be true. The formula is now “=if(and(” followed by the two conditions. The two conditions are that the profitable value is “1” and the budget category is not low. The formula is now “=if(and(K2=1,H2<>“low”).” The “<>” indicates not equal to. Now that the logical condition is done, we have to enter value if true and value if false like the profitable variable. However, rather than a binary variable we can use text.

So, if the condition is true the movie is successful, otherwise unsuccessful. So the formula should be “=if(and(K2=1,H2<>“low”), “successful”, “unsuccessful”).” The text must always be within quotation marks. Press enter and apply the formula to the rest of the column.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year	Movie Title	Bechdel Test	Bechdel Pass	Budget	Domestic	International	Budget	Total Gross	Profit	Profitable	Successful			
2	1970	Beyond the Valley of the Dolls	PASS	(Binary)	(2013)	Gross (2013)	Gross (2013)	Category	(2013)						
					1	5997631	53978683	53978683	low	107957366	101959735	=if(and(K2=1,H2<>“low”), “successful”, “unsuccessful”)			

Figure 16: Calculate “Successful”

7 Filtering

Often you may want to limit your analyses to a subset of your data. That is, you may want to exclude some observations. You can do this using filtering. In this case, we want to filter out all of the movies that were unsuccessful.

To do this, highlight all of your data. Then, click on the “Sort & Filter” button and choose “Filter.”

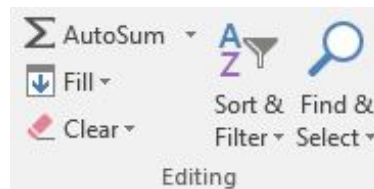


Figure 17: Sort & Filter

You should see an arrow for a drop-down menu next to each variable label. Each of these drop-down menus allows you to filter the data by a variable value.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Yes	Movie Title	Bechdel Test	Pass	Budget	Domestic	International	Budget	Total Gross	Profit	Profitable	Successful

Figure 18: Filter Menus

To filter out the unsuccessful movies, click on the arrow next to the Successful variable. You should see the possible values that this variable could take. By unchecking unsuccessful, all rows for unsuccessful movies will be hidden.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Yes	Movie Title	Bechdel Test	Pass	Budget	Domestic	International	Budget	Total Gross	Profit	Profitable	Successful
4	1971	Shaft	FAIL	0	305063707	404702718	616827003	high	1021529721	716466014	1	successful
7	1971	Willy Wonka & the Chocolate Factory	FAIL	0	17263543	23018057	23018057	medium	46036114	28772571	1	successful
9	1972	The Godfather	FAIL	0	39004975	752051643	1496119403	medium	2248171046	2209166071	1	successful
11	1973	High Plains Drifter	FAIL	0	82329139	82329139	82329139	high	164658278	82329139	1	successful
13	1973	The Exorcist	PASS	1	62926730	1074306128	2111900435	medium	3186206563	3123279833	1	successful

Figure 19: Filter Out “Unsuccessful” Movies

To remove the filter click on the “Sort & Filter” button and unclick the filter option.

7.1 Calculations with Filtered Data

The filter feature only changes what data is visible. Functions using filtered data would still include the hidden data because the data is still present, just not visible. However, you can copy and paste filtered data to its own worksheet or use the limited autocalculations.

8 Searching for Particular Values

Sorting and filtering your data can be helpful for organization. However, sometimes you just need to look up a particular value, and if you have a lot of data then sorting and filtering will still take time. The function “VLOOKUP” allows you to search your data by one variable and returns the value of another variable. For example, we want to name a movie title and learn if it passed or failed the Bechdel test without scrolling through 1,700 rows of data.

To do this, open a new worksheet by clicking on the plus at the bottom of the window next to existing worksheets.

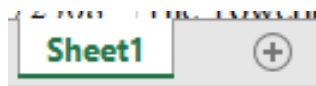


Figure 20: Open a New Worksheet

Now, let’s label where our input (movie title) and output (Bechdel test result) will go. So type in cell A1 “Movie Title:” and in cell A2 “Bechdel Test Result:.” The goal is to type the movie title into cell B1 and have the Bechdel test result appear in cell B2. To do this, we will use the VLOOKUP function in cell B2.

	A	B	C
1	Movie Title:		
2	Bechdel Test Result:		
3			

Figure 21: VLOOKUP Set-Up

The VLOOKUP function requires four inputs: the value you want to search for, the table to search in, the column number of the value you want to return, and how exact the match needs to be. In this example, the value you want to search for will be the movie title that is entered in cell B1. The table to search will be the data in Sheet 1 starting with column B and ending with column L. The VLOOKUP function requires that the first (left-most) column of the table be the column that would contain the searching value (movie title). The column number of the value you want to return is 2. This input does not refer to the column letter of the variable (Bechdel Test). This input is asking for how many columns from the left IN THE TABLE YOU PROVIDED AS THE PREVIOUS INPUT is the variable you want to return for the searched for movie title. The Bechdel Test variable is in column C, but is the second column in the table beginning with column B. Therefore, this function will find the row with the given movie title and return the value in the Bechdel Test column for that row. The last input should be 0 if the given search value (movie title) has to be an exact match – capitalization, punctuation, and spelling. This input should be 1 if the given search

value can be an approximate match – small differences.

Putting all of that information together would give you the formula to enter in cell B2:

`=VLOOKUP(Sheet2!B1,Sheet1!B2:L1777,2,0).`

	A	B	C	D	E
1	Movie Title:				
2	Bechdel Test Result:	=VLOOKUP(Sheet2!B1,Sheet1!B2:L1777,2,0)			
3					

Figure 22: VLOOKUP Formula

After entering the formula, you will notice that cell B2 says “#N/A” which is an error message that means that the formula can’t find what it has been asked to look for. This makes sense because the cell B1, which should contain a movie title is blank!

	A	B
1	Movie Title:	
2	Bechdel Test Result:	#N/A

Figure 23: VLOOKUP Error Message

If we enter the name of a movie in this dataset then this error message should go away and display the Bechdel Test Result instead. So, in cell B1 enter the movie title, The Golden Compass. Then, cell B2 should indicate that this movie passed the Bechdel Test.

	A	B	C
1	Movie Title:	The Golden Compass	
2	Bechdel Test Result:	PASS	

Figure 24: VLOOKUP Result

9 Pivot Tables

Similar to the built-in functions, Excel can help automate tables.

9.1 Example I

The first step is the click on the pivot table option in the Insert tab.

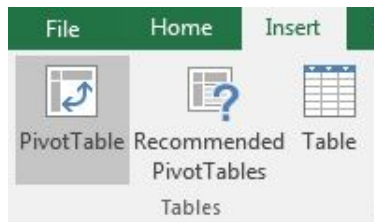


Figure 25: Pivot Table

Then the “Create PivotTable” window should pop up where you select the data to make the pivot table. We want the pivot table to be placed in another sheet, so we should click on the option for “New Worksheet.”

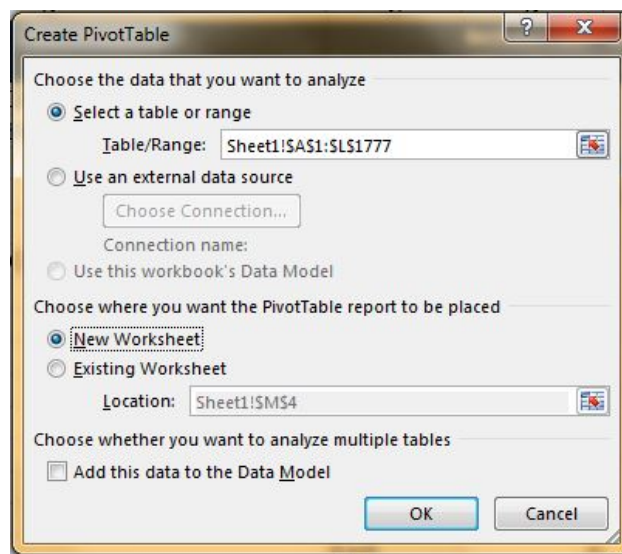


Figure 26: Create PivotTable

Once we have selected data and the location to make the pivot table, we can click “OK.” We will get a blank pivot table.

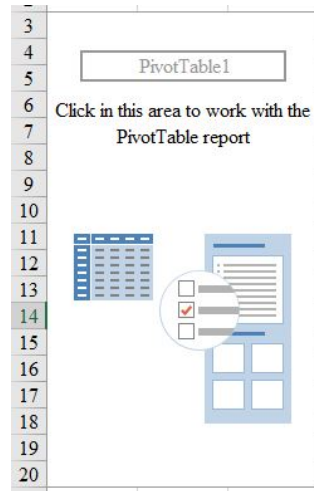


Figure 27: Empty Pivot Table

On the right of the excel window we could see a variety of options we could choose for the pivot table. We start from the top with the pivot table fields. We could choose different fields to add to our pivot table. These fields are all the column titles from the data we chose. (That is why we want to have clear and informative variable labels, so it is easier for us to make the pivot table.)



Figure 28: Pivot Table Fields

In this case, we want Year as the row labels and Bechdel Test as the column labels. In order to do this we can just drag the fields to the areas below.

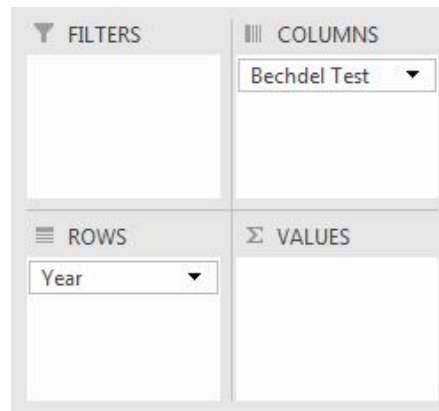


Figure 29: Pivot Table Quadrants

Now if we look at the pivot table, we have row and column labels displayed.

	Column Labels		
Row Labels	FAIL	PASS	Grand Total
1970			
1971			
1972			
1973			

Figure 30: Pivot Table with Labels

The next step would be to choose a value that we want to summarize. The possible values that we could summarize would be budget, revenues, or profit. In this case, we want to summarize profit. Thus we drag “Profit” down to the Values area.

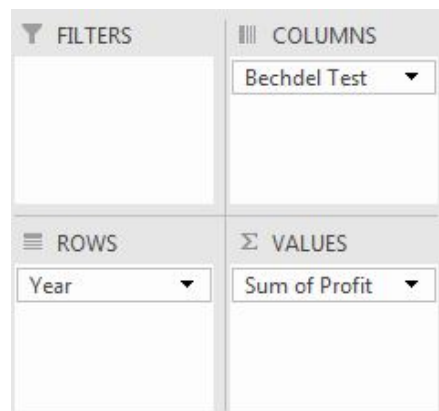


Figure 31: Pivot Table Values

This table shows the total sum of profit for all movies by year and Bechdel Test result. We could also choose to display the average profit instead of the sum. To do this, click on “Sum of Profit” in the Values area. In the drop-down menu, select “Value Field Settings.” In the pop-up window, click “Summarize Values By.” Then from the drop down menu, choose “Average.”

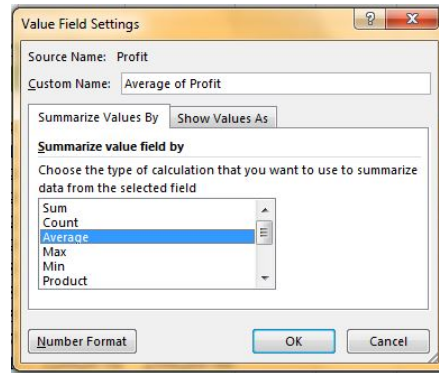


Figure 32: Pivot Table Values

Click “OK,” and you will get the pivot table showing the average profit by year and Bechdel Test result.

Average of Profit Column Labels			
Row Labels	<input checked="" type="checkbox"/> FAIL	PASS	Grand Total
1970		101959735	101959735
1971	262752056.8		262752056.8
1972	2209166071	4544658	1106855365
1973	810634087.3	3123279833	1273163236
1974	490747673.8	295401697	434934537.6
1975	1080183066		1080183066
1976	436789033.8	144629106.7	327229061.1
1977	1444887799	130538356.5	1069359387
1978	594800307.2	1013072078	699368249.8

Figure 33: Final Pivot Table

You can copy and paste pivot tables to format them for a report. Or you could create a chart!

9.2 Example II

You can also continuously edit and change your pivot table. Let's create a pivot table to compare the number of films by budget category and Bechdel Test result.

In this case, we want Bechdel Test as the row labels and Budget Category as the column labels. Now we simply want to know how many observations fall into each category. For this, we can change the calculation from average to count.

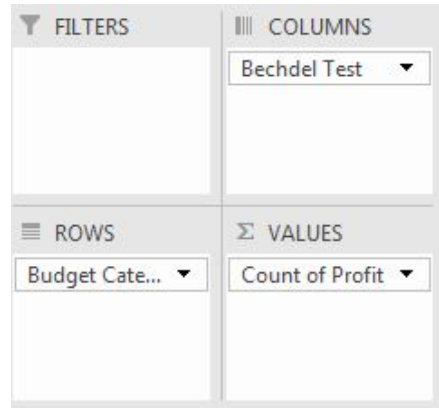


Figure 34: Pivot Table Quadrants II

Then the final Pivot Table should look like the following figure.

Count of Profit	Column Labels			
Row Labels	high	low	medium	Grand Total
FAIL	300	212	470	982
PASS	149	226	419	794
Grand Total	449	438	889	1776

Figure 35: Final Pivot Table II