Big Data Documentation

Release 2016 Fall

Paul Vincent Craven

Sep 25, 2017

Contents

CHAPTER 1

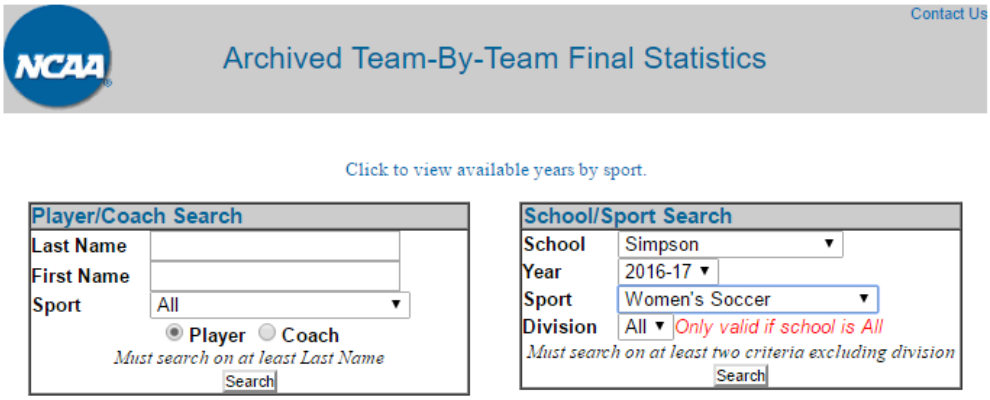
Tutorials

Working with Statistics Functions in Excel

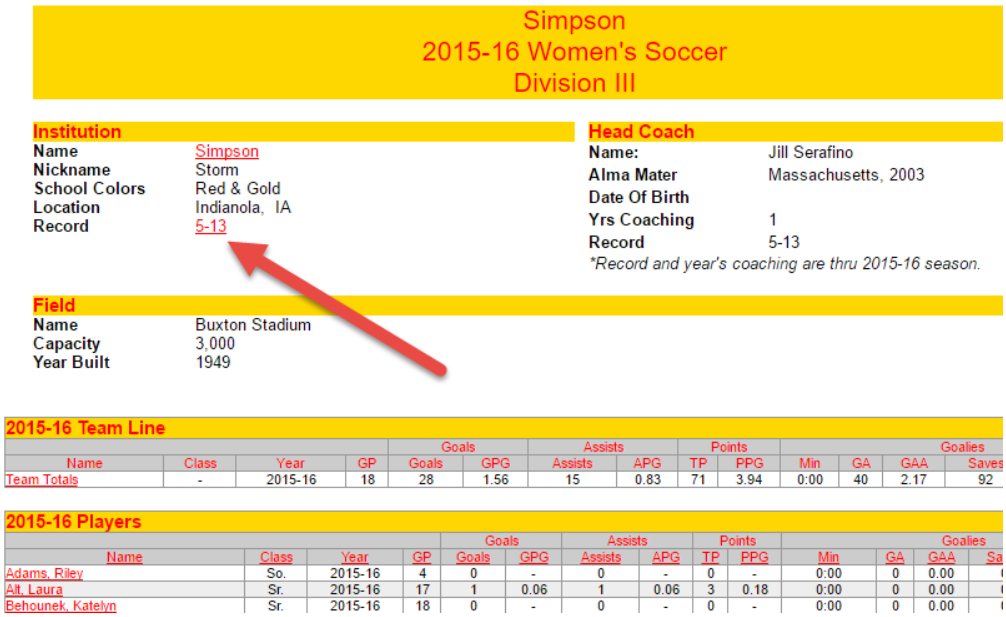
Start here:

http://web1.ncaa.org/stats/StatsSrv/careersearch

Select last year's women's soccer data for Simpson:



Hit 'Search'. Then select 'Simpson'. You should get a lot of data. But it isn't the data we want right now. Click on the link for the record:



Simpson
2015-16 Women's Soccer
Division III

Institution

Name	Simpson
Nickname	Storm
School Colors	Red & Gold
Location	Indianola, IA
Record	5-13

Head Coach

Name:	Jill Serafino
Alma Mater	Massachusetts, 2003
Date Of Birth	
Yrs Coaching	1
Record	5-13

**Record and year's coaching are thru 2015-16 season.*

Field

Name	Buxton Stadium
Capacity	3,000
Year Built	1949

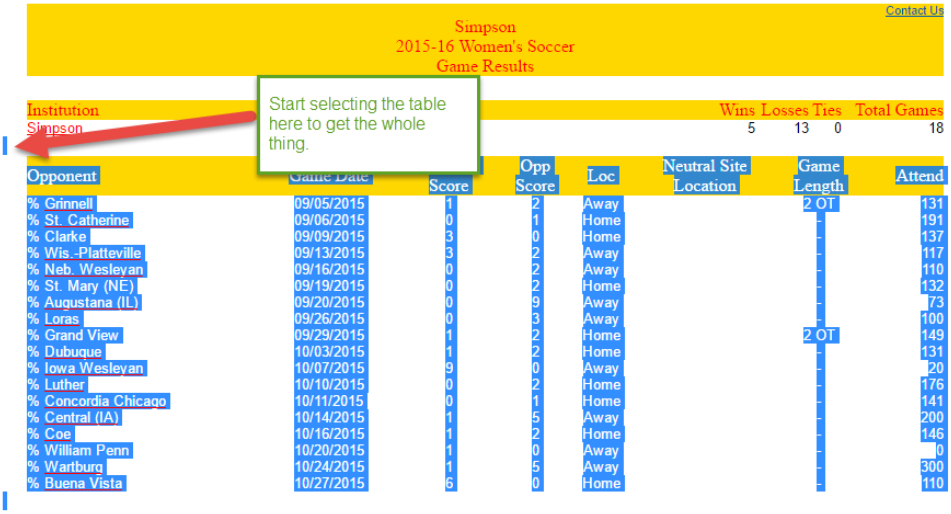
2015-16 Team Line

				Goals		Assists		Points		Goalies			
Name	Class	Year	GP	Goals	GPG	Assists	APG	TP	PPG	Min	GA	GAA	Saves
Team Totals	-	2015-16	18	28	1.56	15	0.83	71	3.94	0:00	40	2.17	92

2015-16 Players

				Goals		Assists		Points		Goalies			
Name	Class	Year	GP	Goals	GPG	Assists	APG	TP	PPG	Min	GA	GAA	Sa
Adams, Riley	So.	2015-16	4	0	-	0	-	0	-	0:00	0	0.00	
Alt, Laura	Sr.	2015-16	17	1	0.06	1	0.06	3	0.18	0:00	0	0.00	
Behounek, Katelyn	Sr.	2015-16	18	0	-	0	-	0	-	0:00	0	0.00	

To successfully select the table, you need to start copying the table before it begins. If you start in the table, it won't copy the table into Excel and it will be all one one line.



Contact Us

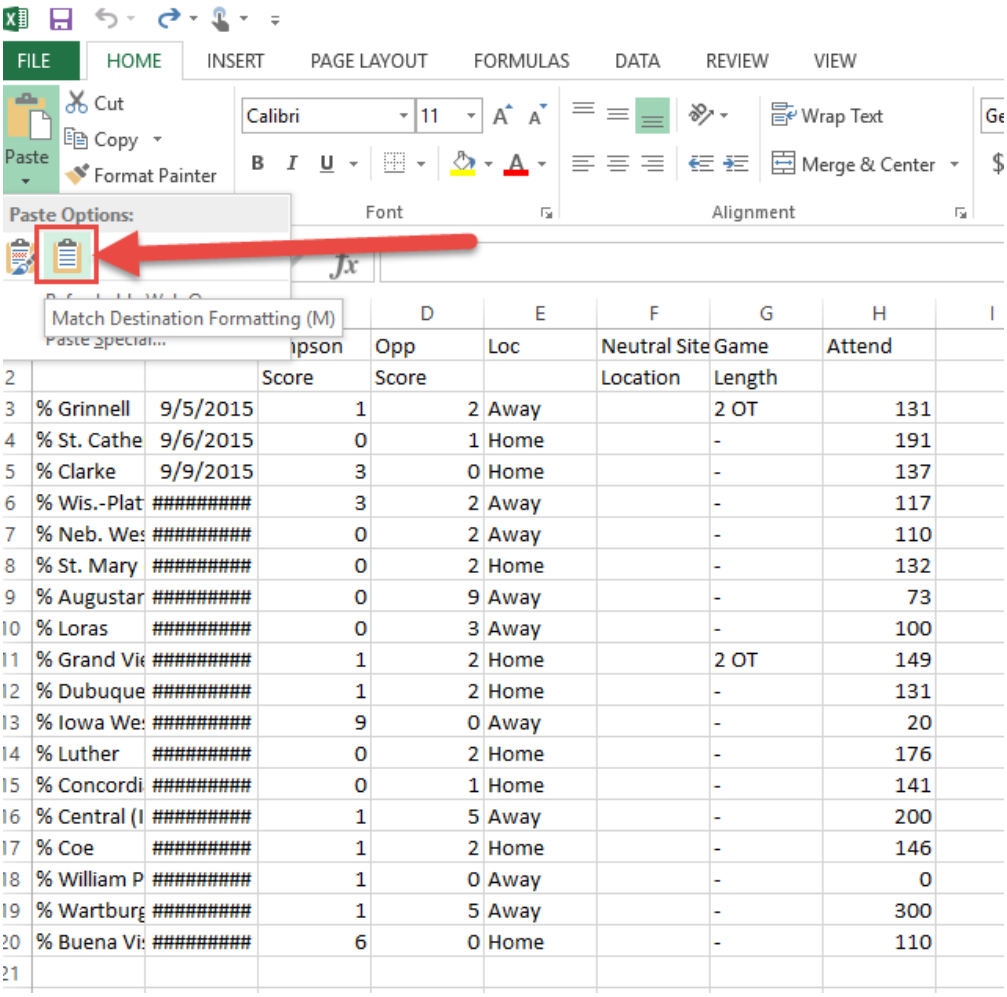
Simpson
2015-16 Women's Soccer
Game Results

Start selecting the table here to get the whole thing.

Institution	Wins	Losses	Ties	Total Games
Simpson	5	13	0	18

Opponent	Game Date	Score	Opp Score	Loc	Neutral Site Location	Game Length	Attend
% Grinnell	09/05/2015	1	2	Away		2 OT	131
% St. Catherine	09/06/2015	0	1	Home		-	191
% Clarke	09/09/2015	3	0	Home		-	137
% Wis.-Platteville	09/13/2015	3	2	Away		-	117
% Neb. Wesleyan	09/16/2015	0	2	Away		-	110
% St. Mary (NE)	09/19/2015	0	2	Home		-	132
% Augustana (IL)	09/20/2015	0	9	Away		-	73
% Loras	09/26/2015	0	3	Away		-	100
% Grand View	09/29/2015	1	2	Home		2 OT	149
% Dubuque	10/03/2015	1	2	Home		-	131
% Iowa Wesleyan	10/07/2015	9	0	Away		-	20
% Luther	10/10/2015	0	2	Home		-	176
% Concordia Chicago	10/11/2015	0	1	Home		-	141
% Central (IA)	10/14/2015	1	5	Away		-	200
% Coe	10/16/2015	1	2	Home		-	146
% William Penn	10/20/2015	1	0	Away		-	0
% Wartburg	10/24/2015	1	5	Away		-	300
% Buena Vista	10/27/2015	6	0	Home		-	110

Now that we've copied the data, paste it into Excel. We usually will want to "match the destination formatting" and not copy the extra style stuff from the web.



			D	E	F	G	H
			Opp	Loc	Neutral Site	Game	Attend
2		Score	Score		Location	Length	
3	% Grinnell	9/5/2015	1	2	Away	2 OT	131
4	% St. Cathe	9/6/2015	0	1	Home	-	191
5	% Clarke	9/9/2015	3	0	Home	-	137
6	% Wis.-Plat	#########	3	2	Away	-	117
7	% Neb. Wes	#########	0	2	Away	-	110
8	% St. Mary	#########	0	2	Home	-	132
9	% Augustar	#########	0	9	Away	-	73
10	% Loras	#########	0	3	Away	-	100
11	% Grand Vie	#########	1	2	Home	2 OT	149
12	% Dubuque	#########	1	2	Home	-	131
13	% Iowa Wes	#########	9	0	Away	-	20
14	% Luther	#########	0	2	Home	-	176
15	% Concordi	#########	0	1	Home	-	141
16	% Central (I	#########	1	5	Away	-	200
17	% Coe	#########	1	2	Home	-	146
18	% William P	#########	1	0	Away	-	0
19	% Wartburg	#########	1	5	Away	-	300
20	% Buena Vi	#########	6	0	Home	-	110
21							

- Show how to resize columns
- Show how to search/replace
- Show how to do cell references
- Show how to do equations
- Learn to do common stat functions

Mean	=AVERAGE(D4:D29)
Median	=MEDIAN(D4:D29)
Standard Deviation	=STDEV.P(D4:D29)
Mode	=MODE.MULT(D4:D29)
Other	
Max	=MAX(D4:D29)
Min	=MIN(D4:D29)
Sum	=SUM(D4:D29)
Count	=COUNT(D4:D29)

- Learn to create a new sheet

- Set a title
- Copy/paste cells
- Copy/paste while transposing

Tutorial: Outlining

Using an outline can save you time. Too many people skip outlining because they think it will cost them time. If it costs you time, you are doing it wrong.

Here's a way to do it:

Step 1: Create an outline-outline

Here's an outline you can start with for *every* paper.

- **Introduction**
  - Thesis
- Background information
- # words in paper / 20 = # outline points.
- Conclusion

Great! Now we just need to fill out that outline.

Step 2: Start the outline

Next, put together a thesis. You'll likely revise it, but get a start.

Also, list out the background information your paper will need to tell the reader.

- **Introduction**
  - Thesis: The threats that face cyber security have been helped and hindered by big data.
- **Background information**
  - What is big data?
  - What is cyber security?
- 20 points
- Conclusion

Step 3: Start filling out the outline

- **Introduction**
  - Thesis: The threats that face cyber security have been helped and hindered by big data.
- **Background information**
  - What is big data?
  - What is cyber security?

- Business (8 points)
- Government (7 points)
- Individual security (5 points)
- Conclusion

Step 4: More detail

Start filling in those points

- **Introduction**
  - Thesis: The threats that face cyber-security have been helped and hindered by big data.
- **Background information**
  - What is big data?
  - What is cyber-security?
- **Business**
  - Government has more services on-line
  - Protecting against cyber-espionage is now a thing
  - **Protect infrastructure**
    - Power
    - Water
    - Communications
    - Nuclear
  - etc.
- Government (Go ahead and list these now)
- Individual security (Go ahead and list these now)
- Conclusion (Do nothing here yet)

Step 5: Add in citations

Put in where you will use citations

- **Introduction**
  - Thesis: The threats that face cyber security have been helped and hindered by big data.
- **Background information**
  - What is big data? (Smith, 2012)
  - What is cyber security? (Whedon, 2014)
- **Business**
  - Government has more services on-line
  - Protecting against cyber-espionage is now a thing
  - **Protect infrastructure (Giles, 2014)**

        * Power
        * Water
        * Communications
        * Nuclear
    - etc.
- (etc)
- Conclusion

Step 6: Order everything

Spend time moving things around. Get things in the best possible order.

Make sure no point contradicts your thesis. If you have a lot of points contradicting your thesis, maybe revise the thesis?

Step 7: Write your conclusion

Write your conclusion. Talk about how your points support the thesis.

Avoid these mistakes:

- Do not simply restate your thesis. Again, show how your data/arguments support your thesis.
- Do not introduce anything new in the conclusion. This is not a good time for "But wait! There's more!"

Revise your thesis. Go back and spend time revising your thesis so that is fits with your whole paper.

Pivot Tables

When working with large data sets, Pivot Tables are a powerful tool for data analysis. Let's learn by example how they work.

To get started, download this Excel data file on colleges (NCES, 2016):

sc_101_college_data.xlsx

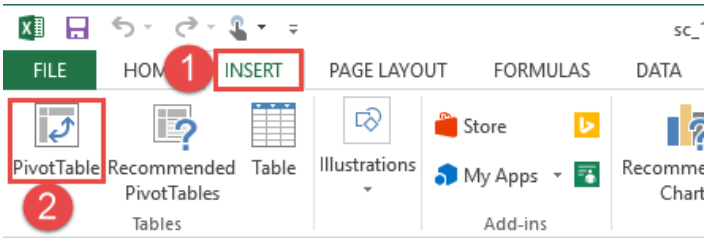
Go ahead and open it up. It has a lot of data. We can't tell too much about the data yet.

To start, we need to create the Pivot Table. Select cell A4, the top left corner of the data we want. Then hit Ctrl-A to select all. (Or ctrl-shift right arrow, followed by the down arrow.) It should look like this:

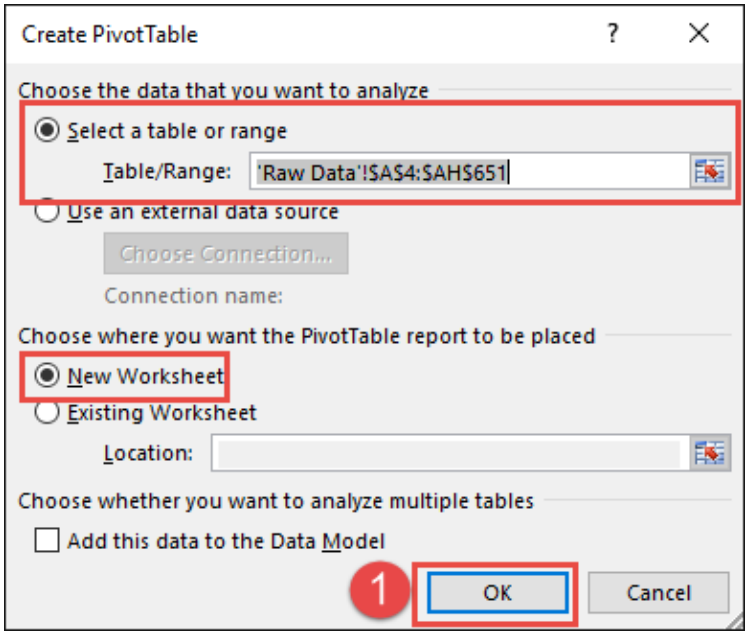
	A	B	C	D	E	
1	Source:National Center for Education Statistics, U.S. Department of Education					
2	Data downloaded from IPEDS data center on 09-26-2016					
3						
4	Institution name	State	Sector	Level	Highest level	Roon
5	Walden University	Minnesota	Private for-	Four or mo	Doctor's degree	
6	Allen College	Iowa	Private not-	Four or mo	Doctor's degree	
7	AIB College of Business	Iowa	Private not-	Four or mo	Bachelor's degree	
8	Briar Cliff University	Iowa	Private not-	Four or mo	Doctor's degree	
9	Buena Vista University	Iowa	Private not-	Four or mo	Master's degree	
10	Capri College-Dubuque	Iowa	Private for-	At least 2 b	At least 2, but less than 4 aca	
11	Capri College-Cedar Rapids	Iowa	Private for-	At least 2 b	At least 2, but less than 4 aca	
12	American College of Hairstyling-Cedar	Iowa	Private for-	At least 2 b	At least 2, but less than 4 aca	
13	Central College	Iowa	Private not-	Four or mo	Bachelor's degree	
14	Clarke University	Iowa	Private not-	Four or mo	Doctor's degree	

Raw Data

Next, select the "Insert" tab, followed by "Pivot Table"



For the next dialog, the defaults should be fine. We already selected the data, so that fills in for us. And we want to start our work on a new worksheet tab.

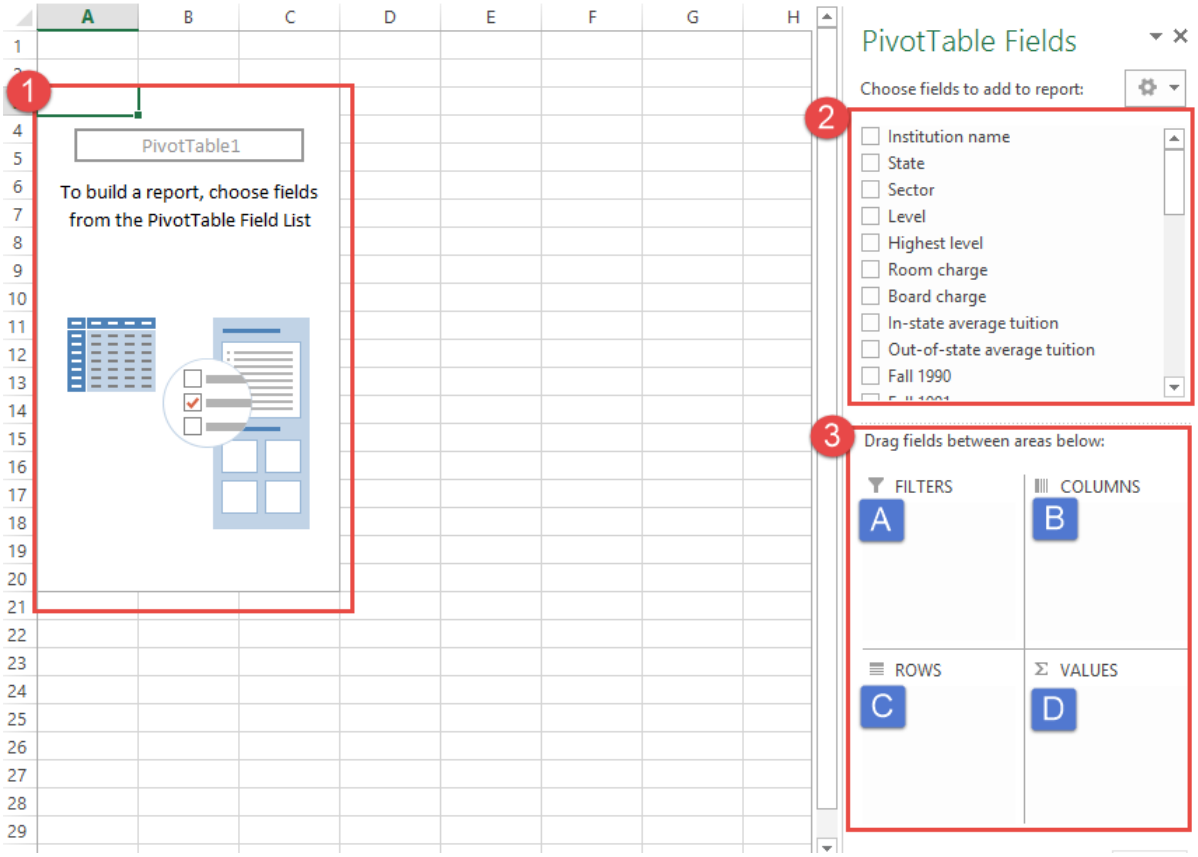


Next, Excel gives you something that seems about as clear as mud. Let's find the main parts:

1. This is where your results will go. We don't have any yet, so it is blank.
2. These are the columns in your report. We will take these and drag them into on of the boxes below.

3. These are where we will drag some of the fields from (2)]

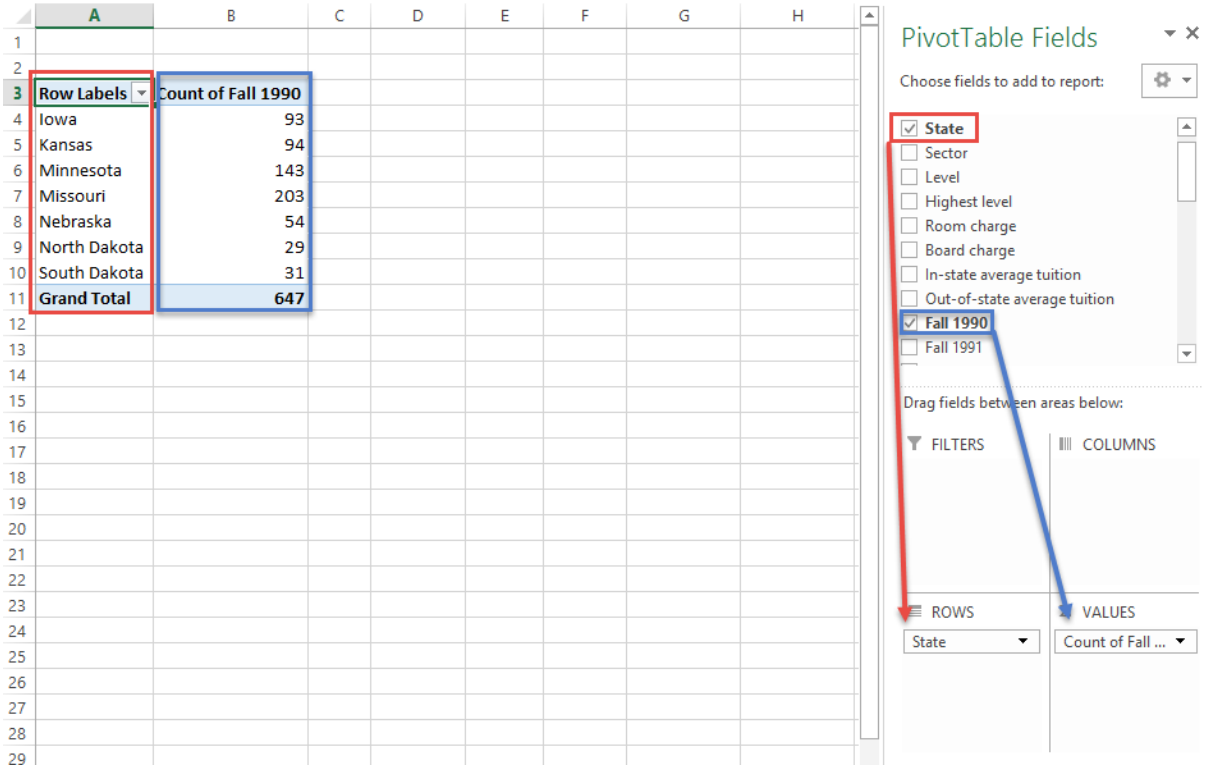
- A. Drag fields here to "filter" the data. If you only want to see colleges from Iowa, set up a filter.
- 2. This controls what shows in the columns. If you put "state" here, then each column will be a different state.
- 3. This controls what shows in the rows. If you put "state" here, then each row will be a different state.
- D. This shows what values will appear in the table. If we put "Sum of Fall 1990" we will see the number of students enrolled in the Fall of 1990.



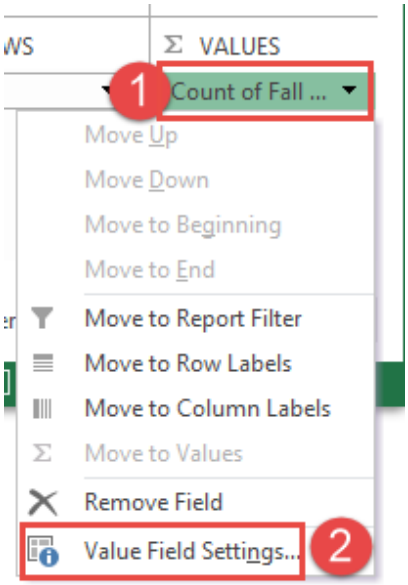
Next, let's take the "State" field and drag it into "rows." When you do this, you should see each state take a row.

After than, take "Fall 1990" and drag it to values. By default we get "count." "Count" will count the number of rows. Therefore if we have four rows with (4, 10, 0, 100) in them, we will get "4". Because there are four rows. It ignores the values.

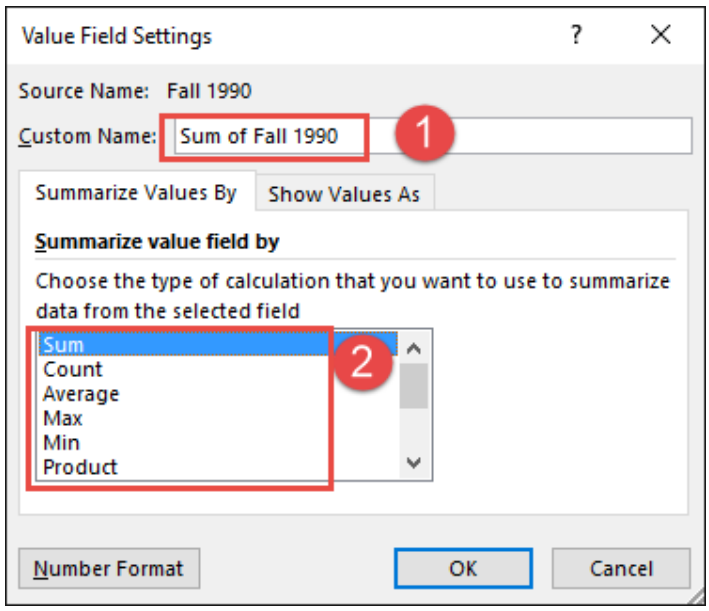
From the results we can see that Missouri has the most schools, while North Dakota has the fewest.



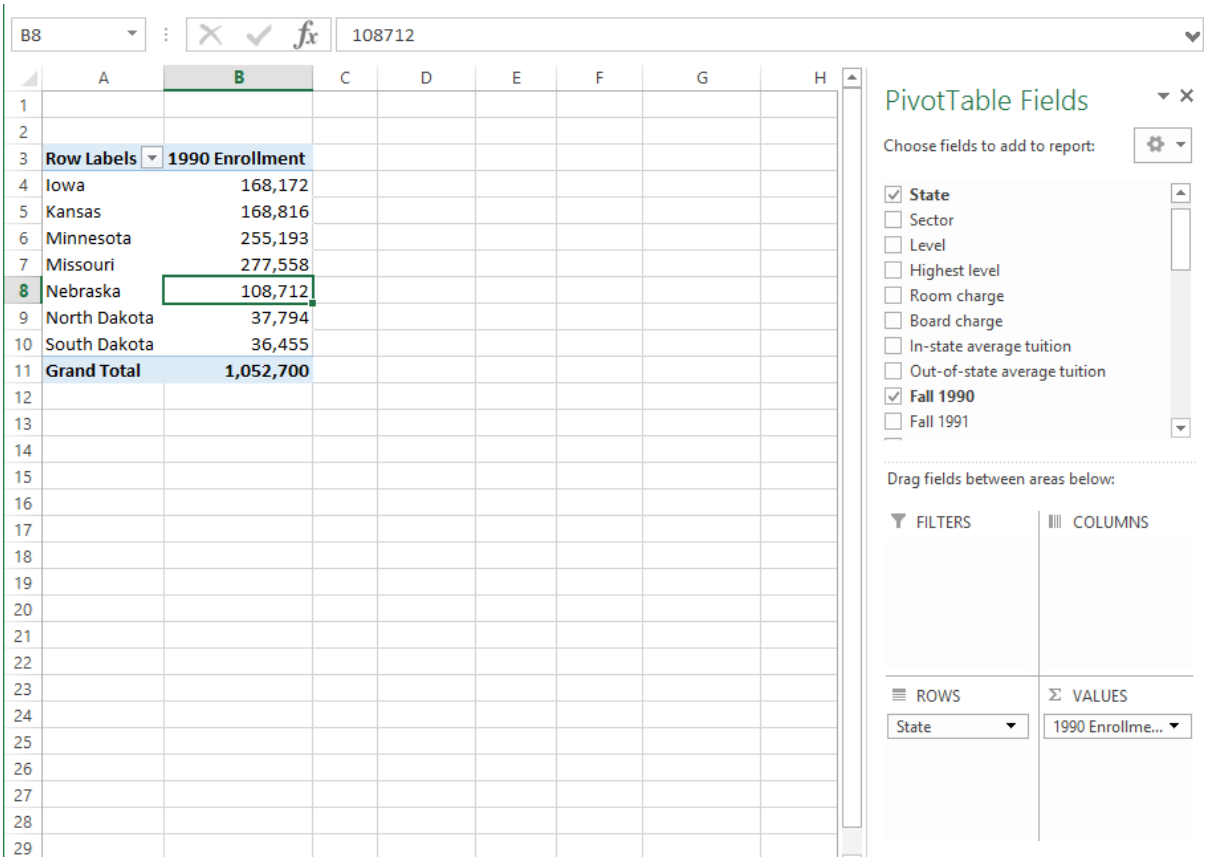
What if we don't want a "count" of schools? Select the count field (step 1 below) and then select "Value field settings" (step 2 below)



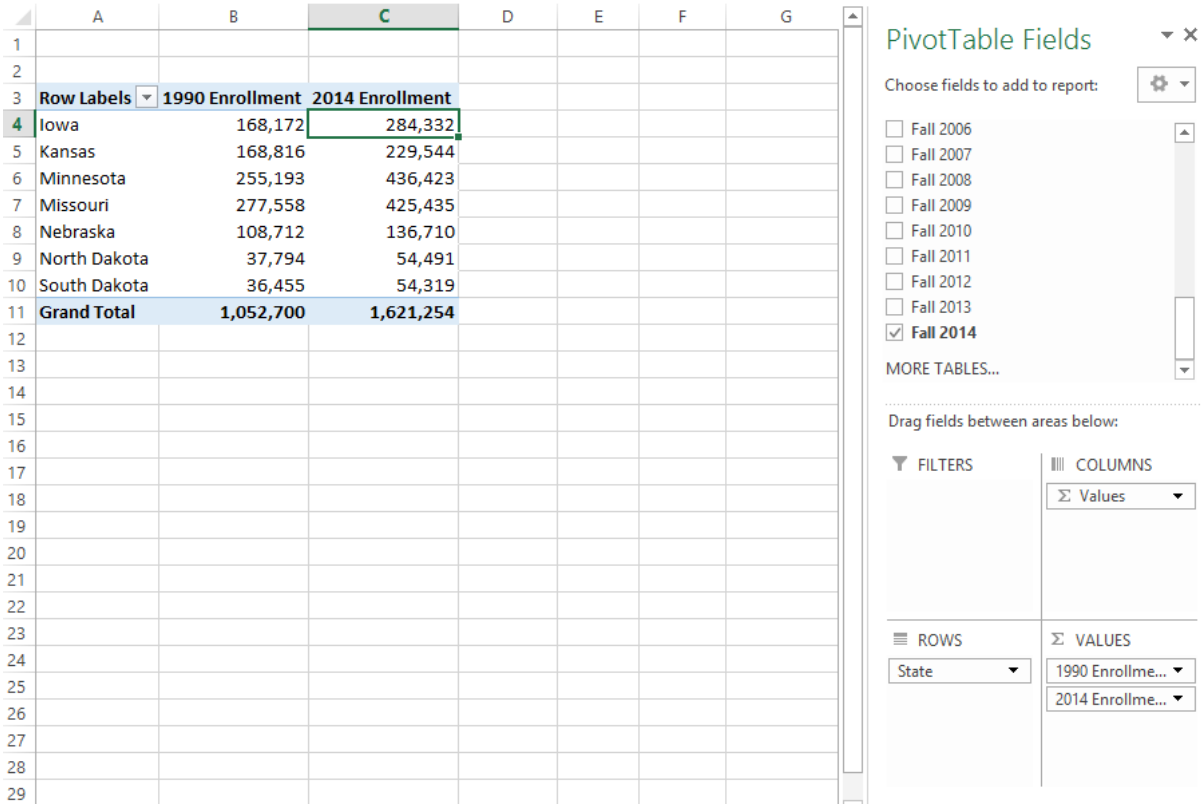
You can change the name of the field (step 1 below) and what we are calculating. If I select "sum" (step 2) I will get the total number of students in the state. If I select "Average" I'll get the average number of students. I can also change the number format. I changed the name, selected "sum," and changed the number format. See below:



And I'm rather happy with my result.

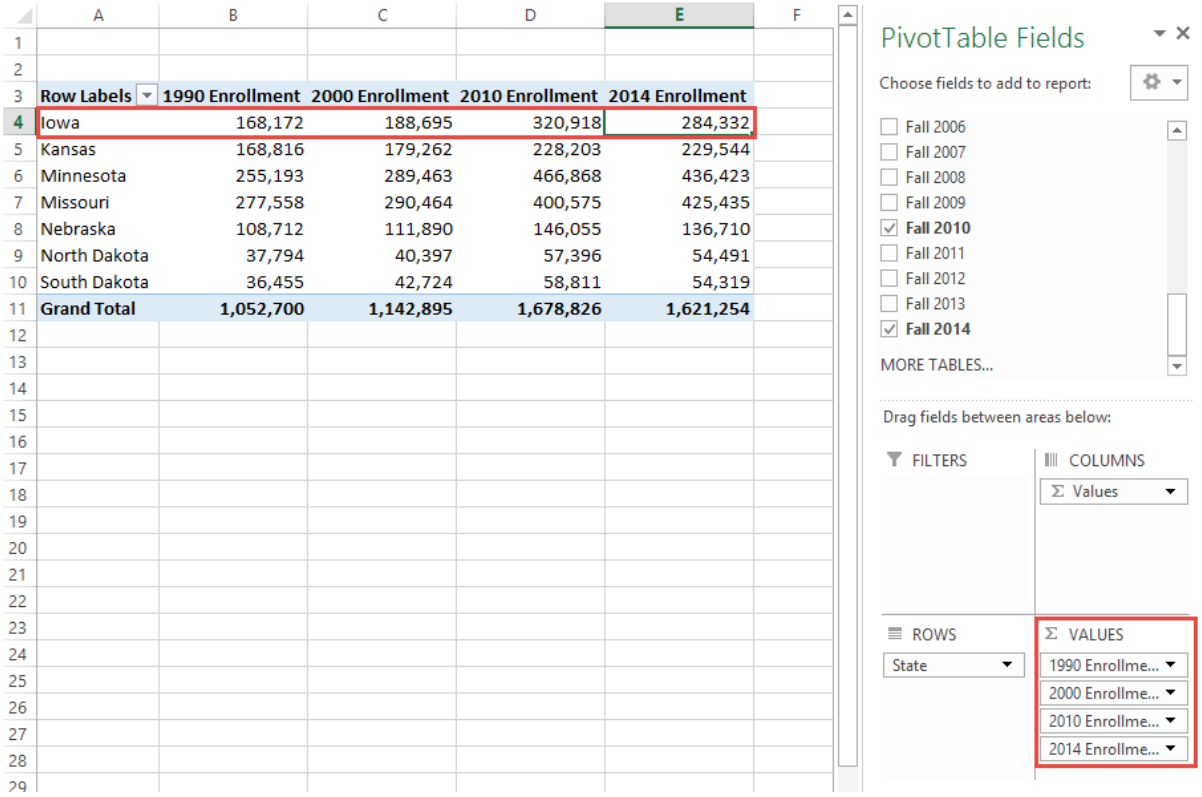


Let's expand this by adding in 2014 enrollment:



Row Labels	1990 Enrollment	2014 Enrollment
Iowa	168,172	284,332
Kansas	168,816	229,544
Minnesota	255,193	436,423
Missouri	277,558	425,435
Nebraska	108,712	136,710
North Dakota	37,794	54,491
South Dakota	36,455	54,319
Grand Total	**1,052,700**	**1,621,254**

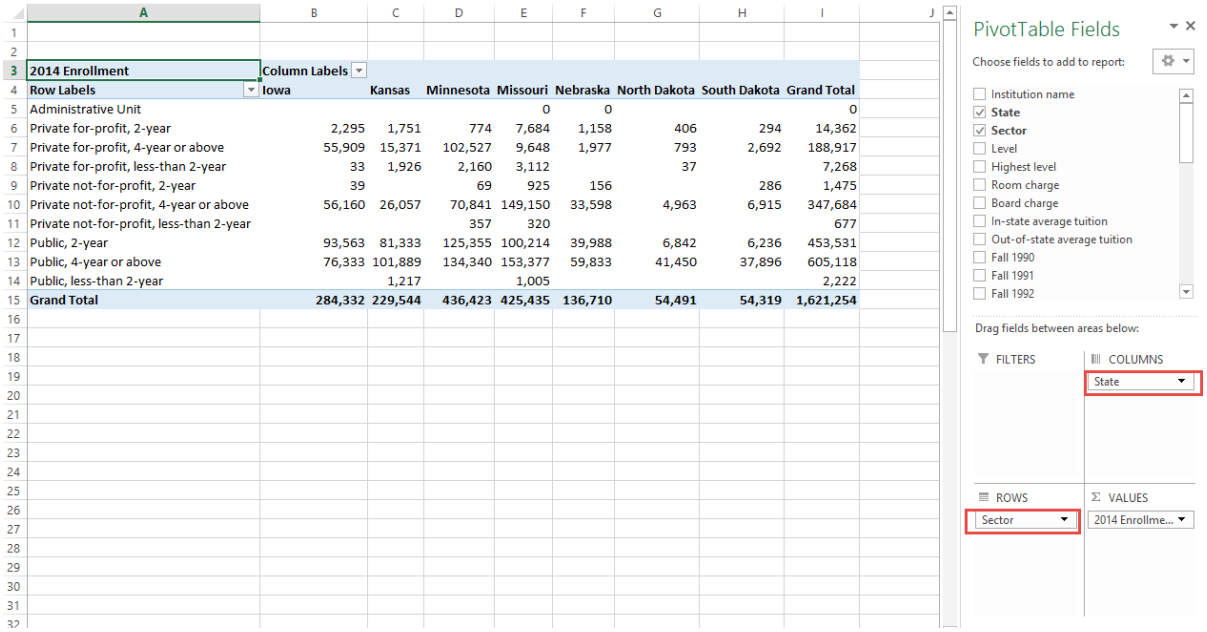
Why stop there? Let's add in 1990, 2000, 2010, and 2014 enrollment.



Row Labels	1990 Enrollment	2000 Enrollment	2010 Enrollment	2014 Enrollment
Iowa	168,172	188,695	320,918	284,332
Kansas	168,816	179,262	228,203	229,544
Minnesota	255,193	289,463	466,868	436,423
Missouri	277,558	290,464	400,575	425,435
Nebraska	108,712	111,890	146,055	136,710
North Dakota	37,794	40,397	57,396	54,491
South Dakota	36,455	42,724	58,811	54,319
Grand Total	**1,052,700**	**1,142,895**	**1,678,826**	**1,621,254**

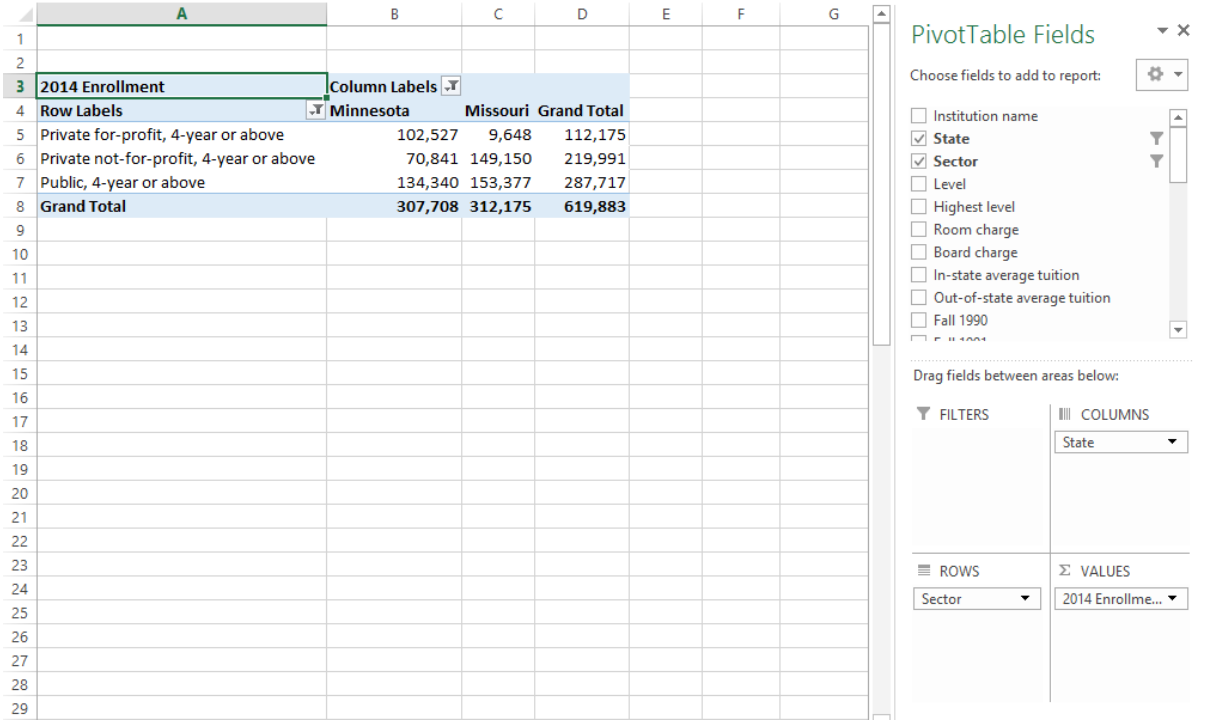
From this I can see that Iowa has taken a real hit in student enrollment while Missouri has done better.

You can make a more complex table. Here I am using both state, sector, and Fall 2014 enrollment:



2014 Enrollment	Column Labels							
Row Labels	Iowa	Kansas	Minnesota	Missouri	Nebraska	North Dakota	South Dakota	Grand Total
Administrative Unit				0	0			0
Private for-profit, 2-year	2,295	1,751	774	7,684	1,158	406	294	14,362
Private for-profit, 4-year or above	55,909	15,371	102,527	9,648	1,977	793	2,692	188,917
Private for-profit, less-than 2-year	33	1,926	2,160	3,112		37		7,268
Private not-for-profit, 2-year	39		69	925	156		286	1,475
Private not-for-profit, 4-year or above	56,160	26,057	70,841	149,150	33,598	4,963	6,915	347,684
Private not-for-profit, less-than 2-year			357	320				677
Public, 2-year	93,563	81,333	125,355	100,214	39,988	6,842	6,236	453,531
Public, 4-year or above	76,333	101,889	134,340	153,377	59,833	41,450	37,896	605,118
Public, less-than 2-year		1,217		1,005				2,222
Grand Total	284,332	229,544	436,423	425,435	136,710	54,491	54,319	1,621,254

I can spot some interesting information in this table. But the is too complex to easily spot what I'm looking at. I'll click on the 'down' arrows in the row and columns to add filters. That was I can narrow in on the interesting data:



2014 Enrollment	Column Labels		
Row Labels	Minnesota	Missouri	Grand Total
Private for-profit, 4-year or above	102,527	9,648	112,175
Private not-for-profit, 4-year or above	70,841	149,150	219,991
Public, 4-year or above	134,340	153,377	287,717
Grand Total	307,708	312,175	619,883

Look, for-profit colleges are clearly more popular in Minnesota! Given that Minnesota is more liberal than Missouri, I would have thought the opposite.

But an Excel file is not a report. I need to move this data into a report. I copied and pasted the table into MS Word. I adjusted a few fields to make the labels clearer. Then I wrote an explanation and citation around my data:

For-Profit Colleges in Minnesota and Missouri

The U.S. Department of Education makes a lot of college data publically accessible (NCES, 2016). Using this data, it is possible to see that about 33% of Minnesota students are enrolled in for-profit colleges. Interestingly, two states away only about 3% of Missouri students are enrolled in for-profit colleges.

2014 Enrollment	State		
Sector	Minnesota	Missouri	Grand Total
Private for-profit, 4-year or above	102,527	9,648	112,175
Private not-for-profit, 4-year or above	70,841	149,150	219,991
Public, 4-year or above	134,340	153,377	287,717
Grand Total	**307,708**	**312,175**	**619,883**

Bibliography

National Center for Education Statistics. "IPEDS Custom Data Reports." U.S. Department of Education, Web. 26 Sept. 2016.

That's it! I've used a Pivot Table to crunch 650 records of data with 14 fields each, and learn something I didn't know before.

Bibliography

National Center for Education Statistics. "IPEDS Custom Data Reports." U.S. Department of Education, Web. 26 Sept. 2016.

Tutorial: Processing Names

Read this article: How to Tell Someone's Age When All You Know Is Her Name.

Let's do something similar. Let's find when a person likely was born based on their age.

Getting the Data

You can get all the names of everyone born from the Social Security Administration here:

http://www.ssa.gov/oact/babynames/limits.html

We get the name, the count of people with that name, the sex, and the year born. But wait! Don't download the data yet. The data involves too many rows to easily process in Excel. Let's learn some new tools.

- Linux Computer - We will be using a computer running the Linux operating system that I've set up using Amazon's cloud computing tools. This computer already has the data on it.
- MobaXTerm - We need to get to that remote computer somehow. We will use a terminal program. A terminal program is like a web browser, but more primitive. It can receive characters and display characters. There are many terminal programs just like there are many web browsers. The terminal program we will use is called MobaXTerm. Macs are similar to Linux in many ways and already have a built-in program called "Terminal."
- Command-line tools - Into this terminal we will be typing commands that go to the computer that exists at Amazon. Rather than clicking on tools, we will be typing them in.

Wait, what is the source of my data? The source of this data is *not* MobaXTerm any more than Chrome is what you'd list for a source on a report. The data comes from the Social Security Administration. That is your source.

Command Line

Again, we will be using something called the "command line." Rather than using the mouse and menus, we will go old-school and type in commands. We can process more data this way.

Also, rather than use "Microsoft Windows" or "MacOS" we will be using a different operating system called "Linux."

Another thing that we *could* do with the command line is to create a file with all the commands we need. Then we could run that file, for each name. We don't have to recreate our work.

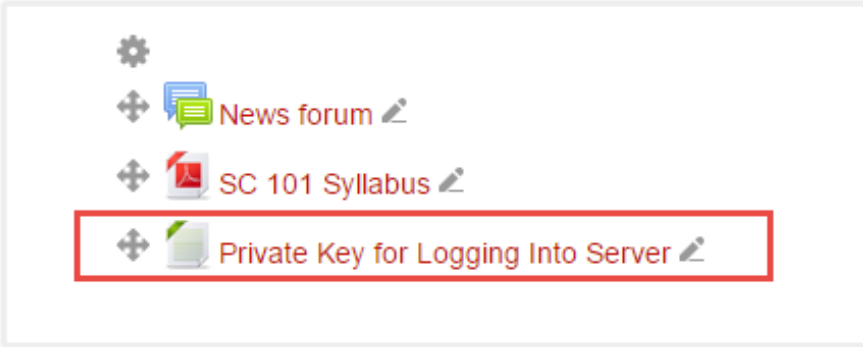
This isn't something you can do with a graphical interface.

To use the command like interface, we need to learn some commands.

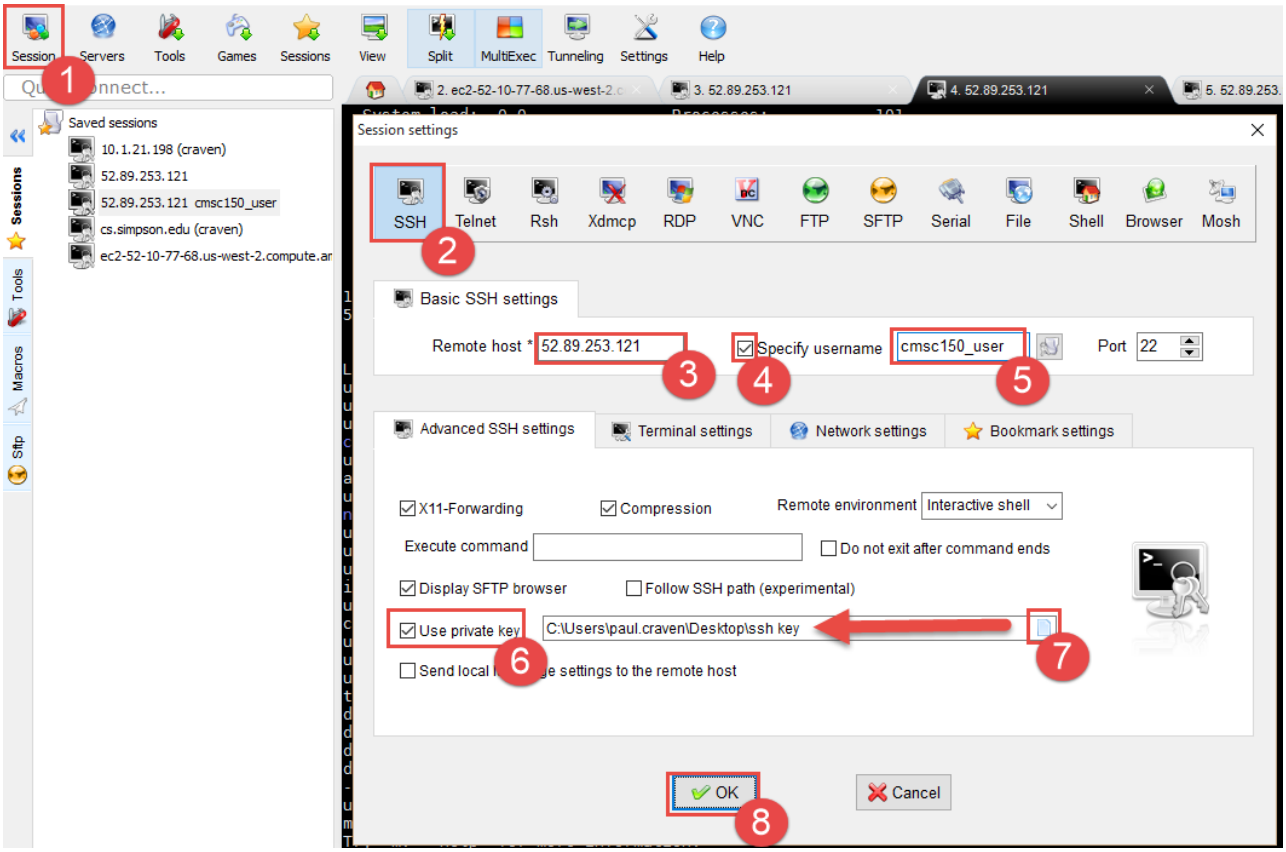
Here is a reference of commands we will need:

Command	Description
`ls`	List files in the current directory.
`ls -la`	List file details in the current directory.
`cd <dirname>`	Change directory
`cd ..`	Go up one directory
`cat <filename>`	List the entire contents of a file
`head <filename>`	List the first few lines of a file
`tail <filename>`	List the last few lines of a file
`pwd`	Show the current folders (the 'path')
`wc <filename>`	Count the words in a file
`egrep '<regex>' <filename>`	Pull out matching lines based on a regular expression
`egrep -c '<regex>' <filename>`	Count matching lines based on a regular expression
`sed 's/<reg>/replacement' <filename>`	Search and replace on a file
`<command> | <command>`	Pipe output from one command into another command
`<command> > file.txt`	Direct the output to a file

First, from the Scholar class website, find and download the "ssh key" file that that will let us log into out remote Linux server.



Save it somewhere on a flash drive or on your desktop. It will act as a really long password to the machine. Start up MobaXTerm. It is on the lab computers. You can also download it to your computer if you want. Create a new session. Fill out the session to look something like this:



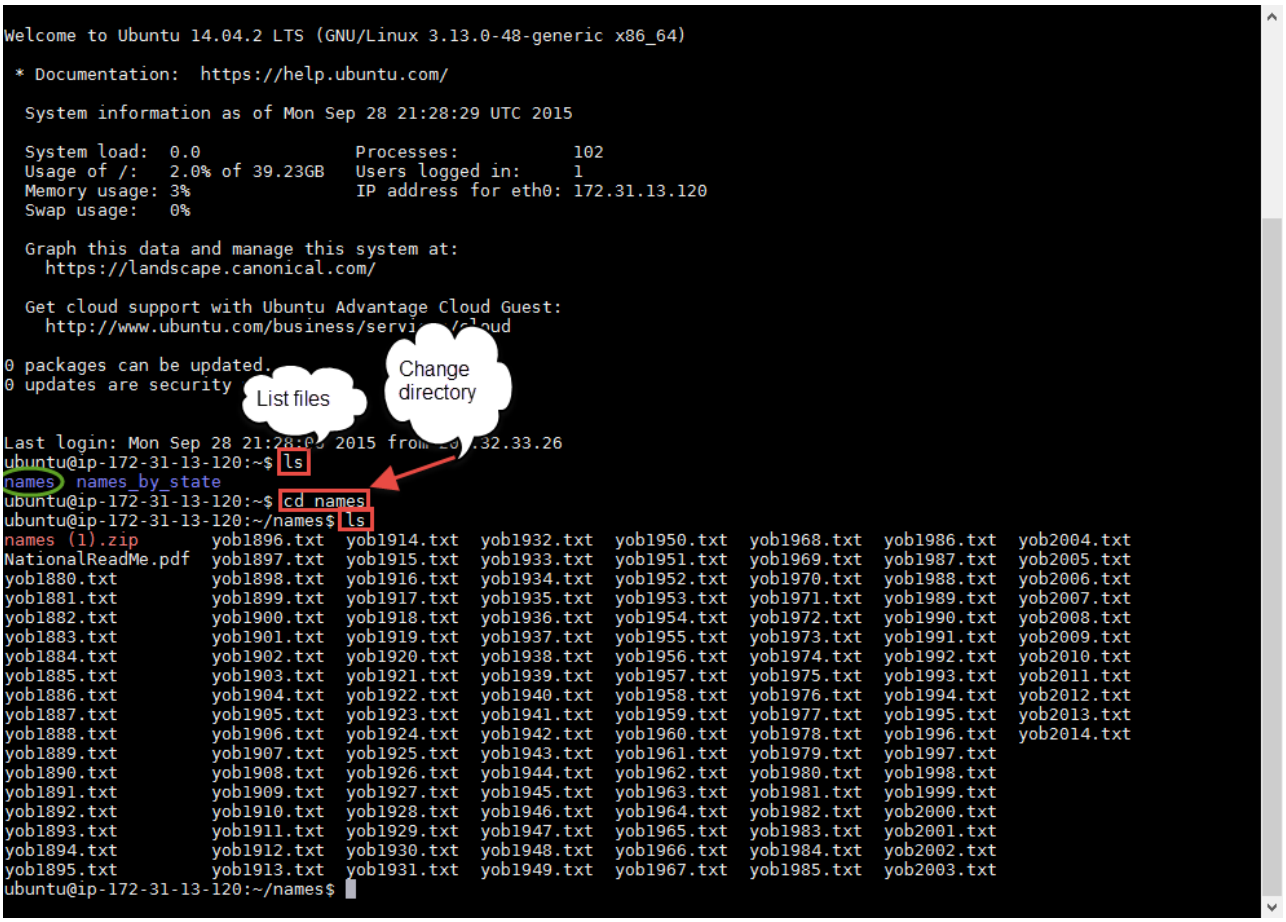
Except: use 52.27.55.158 as the "Remote host".

Go ahead and connect to the server.

Now we will do the following:

1. List all the files that are in the current folder using the `ls` command.
2. Change the folder we are in to the names folder using the `cd` command followed by the folder name.
3. Again, list all the files our new folder.

After doing this, your screen should look like what is below:



Each of these files contains all the names of births in the United States for that year, along with the count. It does NOT include a name if fewer than five people were born with that name.

We can see the contents of the file by using one of these commands:

- `cat <filename>` List the entire contents of a file. Bad idea because these are really big files. Hit Ctrl-C to stop the list if you do this anyway.
- `head <filename>` List the first few lines of a file
- `tail <filename>` List the last few lines of a file

That is a lot of data. How much data? Too much to easily count. Thankfully, there is a command that will count it for us.

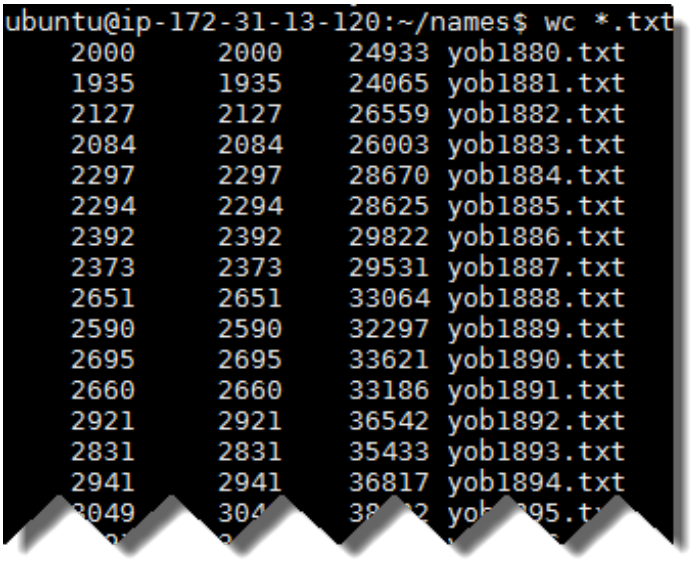
- `wc <filename>` Count the words in a file

Try it out:

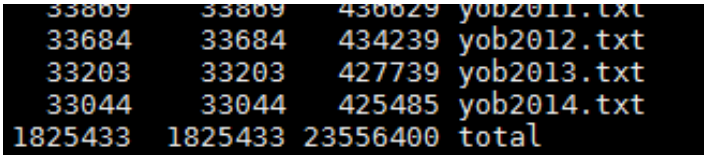
```
ubuntu@ip-172-31-13-120:~/names$ wc yob2014.txt
 33044  33044 425485 yob2014.txt
```

In my example, computer tells you there are 33,044 lines in the file. 33,044 words in the file. And 425,485 characters. You might get different results as you are working on updated data.

Many commands can work on multiple files. You can do this with a "wildcard." By typing wc *.txt it will run word-count on every single text file in this directory:

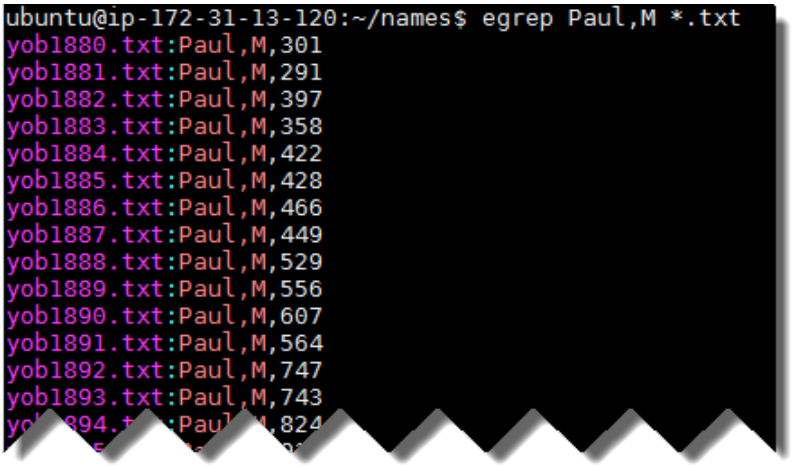


It also gives you a total at the end:



That's 1,825,433 lines we just counted.

We can use the `egrep` command to use a regular expression and pull out only the lines we are interested in. At this point, we won't even use a regular expression, we will just match text. Here I'm looking for every male named "Paul" born:



Try this with your name, or some other name you are interested in.

Copying the data

- Select the text in MobaXTerm by click-dragging
- It is automatically copied.
- Switch to Excel and paste

Now we need to get this data into a format we can use.

- Search and replace: Replace 'yob' with nothing.
- Search and replace: Replace '.txt:' with a comma.
- Select all the data.
- Click "Data" tab
- Click "Text to Columns"
- Our data is delimited (separated) by commas. So hit 'Next'
- Click 'Comma'
- Click 'Finish'
- Split using text-to-columns and search-and-replace

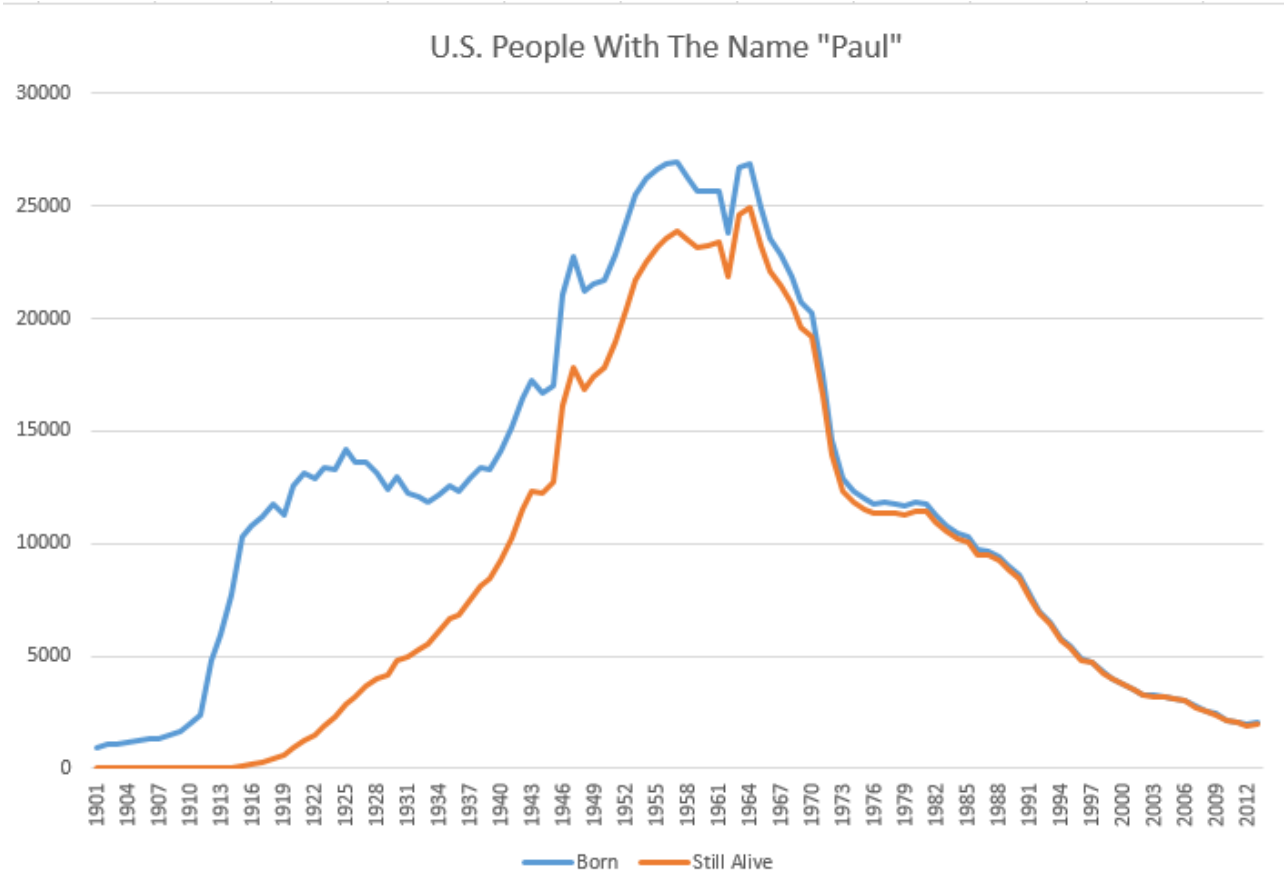
Figure out how many people are still alive today Use this data to calculate what percent of people born in a certain year are still alive. This data came from the Social Security Administration's (SSA's) life tables. It isn't all that accurate because the death rate is only for a person born in 2013, but it will work for our purposes here. If you plan on being an actuary, you'll likely use these tables a lot.

```
Age     Male Percent Alive      Female Percent Alive
0       1.00000 1.00000
1       0.99348 0.99462
2       0.99302 0.99425
3       0.99273 0.99403
4       0.99252 0.99387
5       0.99235 0.99373
6       0.99219 0.99361
7       0.99205 0.99351
8       0.99192 0.99341
9       0.99180 0.99331
10      0.99170 0.99322
11      0.99161 0.99312
12      0.99151 0.99303
13      0.99138 0.99291
14      0.99119 0.99278
15      0.99091 0.99262
16      0.99052 0.99243
17      0.99003 0.99220
18      0.98943 0.99194
19      0.98870 0.99165
20      0.98785 0.99132
21      0.98685 0.99095
22      0.98572 0.99054
23      0.98449 0.99010
24      0.98321 0.98963
25      0.98191 0.98915
26      0.98060 0.98864
27      0.97928 0.98811
28      0.97795 0.98755
29      0.97659 0.98697
30      0.97519 0.98635
31      0.97376 0.98569
32      0.97230 0.98500
33      0.97080 0.98426
34      0.96927 0.98348
35      0.96772 0.98265
```

```
36      0.96612 0.98176
37      0.96448 0.98081
38      0.96277 0.97979
39      0.96097 0.97870
40      0.95908 0.97753
41      0.95708 0.97627
42      0.95493 0.97491
43      0.95262 0.97343
44      0.95012 0.97182
45      0.94739 0.97004
46      0.94441 0.96810
47      0.94115 0.96597
48      0.93759 0.96364
49      0.93368 0.96109
50      0.92940 0.95829
51      0.92472 0.95524
52      0.91961 0.95193
53      0.91406 0.94834
54      0.90804 0.94449
55      0.90153 0.94038
56      0.89450 0.93598
57      0.88693 0.93126
58      0.87883 0.92623
59      0.87022 0.92090
60      0.86112 0.91526
61      0.85147 0.90927
62      0.84125 0.90287
63      0.83042 0.89600
64      0.81899 0.88858
65      0.80691 0.88054
66      0.79412 0.87177
67      0.78054 0.86223
68      0.76613 0.85187
69      0.75084 0.84069
70      0.73461 0.82864
71      0.71732 0.81561
72      0.69889 0.80147
73      0.67930 0.78616
74      0.65853 0.76961
75      0.63657 0.75177
76      0.61329 0.73244
77      0.58859 0.71148
78      0.56249 0.68888
79      0.53504 0.66467
80      0.50629 0.63880
81      0.47621 0.61114
82      0.44484 0.58159
83      0.41233 0.55016
84      0.37890 0.51694
85      0.34482 0.48205
86      0.31040 0.44565
87      0.27598 0.40796
88      0.24201 0.36933
89      0.20896 0.33017
90      0.17735 0.29104
91      0.14768 0.25257
92      0.12043 0.21542
93      0.09599 0.18027
```

```
94      0.07463 0.14775
95      0.05647 0.11839
96      0.04157 0.09267
97      0.02977 0.07083
98      0.02075 0.05285
99      0.01410 0.03852
100     0.00935 0.02745
101     0.00605 0.01909
102     0.00380 0.01292
103     0.00232 0.00850
104     0.00137 0.00541
105     0.00078 0.00333
106     0.00043 0.00197
107     0.00023 0.00112
108     0.00011 0.00061
109     0.00005 0.00031
110     0.00002 0.00015
111     0.00001 0.00007
112     0.00000 0.00003
113     0.00000 0.00001
```

Create a graph similar to the following for your name of choice:



We could actually process the data and create the graph completely with the command line. That's a bit more involved than this tutorial will get into. But know this: it is possible to automatically create one of these graphs for every single name.

CHAPTER 2

Assignments

Assignment 1: Citations

Use Data and Citations to Transform an Argument

Assignment

- Write a paragraph that states information
- Use at least two proper citations to back up that information
- Use MLA as your citation style.

Assignment Goals

- Start with an idea.
- Learn to add verifiable facts to your writing.
- Look for data bias.
- Cite the facts.
- Start viewing with suspicion any argument that doesn't do this.

Example

Here is a paragraph from a student paper:

> In the current age, traffic accidents have become a lot more common. This is because of the increase in amount of vehicles out on the roads and the new distractions that have come.

- The idea behind the prior paragraph was good.

- The wording isn't very good.
- Also, is the paragraph even correct?
- With 15 minutes of searching and 15 minutes of proper citation work, I was able to research it and transform the paragraph.

Here's my new paragraph, with citations:

> Even with cell phones, car accidents haven't gotten worse in the last ten years. Fatal traffic accidents decreased from 18.6 per 100,000 vehicles in 2003 to 12.2 per 100,000 vehicles in 2013 (NHTSA, 2014). Total accidents have been between 5.3 and 6.0 million per year from 2007 to 2013, and distraction related accidents have held steady between 16% and 17% of that number (NHTSA, 2011; NHTSA, 2015).
>
> "Fatality Analysis Reporting System." FARS Encyclopedia. National Highway Traffic, 2014. Web. 24 Aug. 2015.
>
> U.S. Department of Transportation and National Highway Traffic Safey Administration, ed. "Distracted Driving 2011." Traffic Safety Facts Apr. 2013.
>
> U.S. Department of Transportation and National Highway Traffic Safey Administration, ed. "Distracted Driving 2013." Traffic Safety Facts Apr. 2015.

Understanding Bias

- Wait, was my data biased?
- Bias does NOT have to be intentional
- For example, surveying students at Simpson, to be representative of all U.S. people 18-22, would introduce data bias.
- If the data is "hard numbers" that doesn't mean it is biased.
- If a website is .org, that does not mean it isn't biased.

Good Assignment Examples

Assignment 1:

> Ever since Lebron James entered the NBA there has been an argument about if he was better than the great Chicago Bulls player, Michael Jordan. Under initial inspection Michael Jordan seems to have the upper hand. Of the top 100 single season performances for scoring, Michael Jordan is listed 11 times, whereas Lebron James is only listed twice (Basketball Reference, 2015). On top of that Michael Jordan had 7 seasons where he scored more points than James did in his highest scoring season. Although scoring is a large part of the total game of basketball, defense also plays a major role in the game of basketball. In reference to rebounding Michael Jordan averaged 6.2 rebounds per game through 15 seasons (NBA.com, 2003). Lebron James is currently averaging 7.1 rebounds per game through 12 seasons (NBA.com, 2015).
>
> "LeBron James Career Stats." NBA Stats. NBA Media Ventures, LLC, 2015. Web. 1 Sept. 2015.
>
> "Michael Jordan Career Stats." NBA.com. NBA Media Ventures, LLC, 2003. Web. 1 Sept. 2015.
>
> "NBA & ABA Single Season Leaders and Records for Points." Basketball Reference. Sports Direct Inc. 2015. Web. 1 Sept. 2015.

Assignment 2:

> Previous studies have shown a relationship between having your cellphone within reach from you at night and how well someone sleeps. Harvard Medical School scientist have found specific wavelengths of light can suppress the slumber-inducing hormone melatonin in the brain(News Max Health, 2015). Wayne

Conn a sleep coordinator at Texas Medical Center says, "Using your phone right before bed is unhealthy, it overstimulates the brain, and will be harder to fall asleep and stay asleep" (Schmitt, 2014). So in conclusion if you want a better nights sleep then resist from using a cellphone before bed, and having it in reach while sleeping.

Schmitt, Madeline. "Cell Phones Believed to Cause Serious Sleep Problems." KXII RSS. 26 June 2014. Web. 01 Sept. 2015.

News Max Health. "Smartphones Can Cause Insomnia: Harvard Smartphones Can Cause Insomnia: Harvard." Bloomberg News, 7 Jan. 2014. Web. 1 Sept. 2015.

Bad Assignment Examples

Assignment 1:

Which Sport is More Dangerous? Football or Rugby?

In football, the injury rate per 1000 hours is 35.3. Most of those injuries occur in the leg area. Rugby's rater per 1000 hours is 69. That rate makes rugby the "most dangers" team sport in the world. Rugby also averages about 1.4 serious injuries per game.

Assignment 2:

League of Legends a sport?

League of Legends is a booming game that has player and fan base that is rising at extreme rates. This leads to the question, is League of Legends a sport? Some might think not but "Riot Games" the developer of League of Legends thinks otherwise. Riot Games reports increases in their player base over just a few years; it rose from 12 million players daily to 27 million in just over a year (Tassi, Jan. 27th, 2014) . The huge number of people that play League of Legends is really shown when compared to Call of Duty. The numbers for Call of Duty players combined monthly in the year 2011 only had 40 million players (Tassi, Jan. 27th, 2014). Many other things that are similar to professional sports include participation in tournaments, trading players during the off season, and a final championship that millions of people watch (Holly, Jul. 15th, 2013).

Bad Citation Examples

Don't use a URL:

http://simpsonathletics.com/sports/fball/2014-15/schedule

The dictionary is never a source:

"Stress." Merriam-Webster. Merriam-Webster, 2016. Web. 01 Sept. 2016.

You can probably guess at least a year:

2013-14 Memphis Grizzlies. (n.d.). Retrieved February 27, 2015, from http://www.basketball-reference.com/teams/MEM/2014.html

No clue where this came from:

List of United States university campuses by enrollment. (n.d.). Retrieved October 23, 2014.

Wait, you listed the publisher, but then said n.p.? And no date?

"Graduation Rate Trends 1999-2000 to 2009-2010." Education Week. N.p., n.d. Web. 03 Sept. 2015.

No way I'd find this again. It is basically saying "Here's this random thing I found on the Internet. I'm not even going to tell you where I found it."

"11 Facts About High School Dropout Rates." 11 Facts About High School Dropout Rates. N.p., n.d. Web. 03 Sept. 2015.

Oh, and never list Google as your source. That is like listing the library as your source.

Using Citation Generators

- There are many on-line tools to help generate citations.
- Be careful using the tools. Look at what they generate. Edit as needed.
- Try to avoid using n.d., or n.p. which indicates no publication date or publisher. That probably isn't a good source.
- Make sure that even without a URL you would be able to find the information.
- Ask the instructor if they want a URL or link in the citation.

Schedule for the Assignment

- Ten minutes to pick a question.
- Twenty minutes to pick data.
- Twenty minutes to create slides
- Prep presentation. Two PowerPoint slides, one paragraph and one citation slide.
- Some of this, you'll likely need to do outside of class. You will some time to review this on Thursday with myself, Maddy, and Christopher on Thursday.
- Due at end of class Thursday.
- We will go through them in a week. I won't associate names with paragraphs.

Sample Questions

- Who is the best player in sport <name sport>?
  - Is Ed Walsh as good as Cy Young? (baseball)
  - Is Aaron Rodgers the best quarterback in the NFL?
- Who do we get our oil from?
- What does the U.S. spend the most money on?
- What is the number one killer in the U.S.?
  - What are the most common ways of dying for twenty year olds?
- What is the best subject to major in at college?
- What is the best paying job in the U.S.?
- How likely is a student to drop out of college?
- Is the high school dropout rate increasing?
- Do statistics support that there is a "war on cops?"
- Is the rate of suicide increasing?

Assignment 2: Basic Excel

Open a new Excel document.

Create a information sheet, with your name, date, and a title. Name this tab “Information”.

Switch/create another tab.

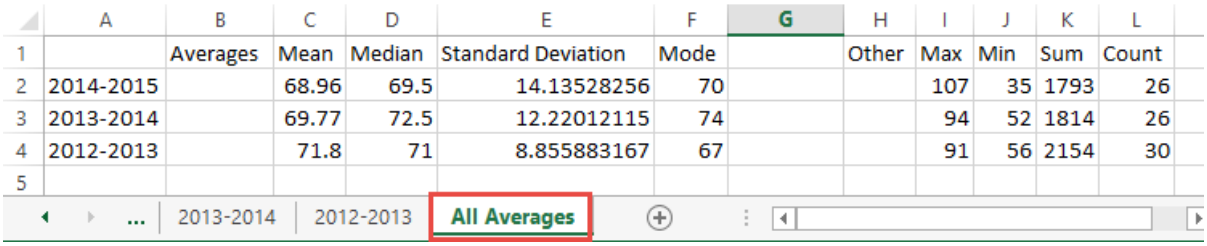
Pick a different sport than what we did in class as a tutorial. Use Excel to analyze the data. Remove unnecessary data. Set column width to proper values.

Do this for three years. Your result should look something like this:

Mens Soccer 2012/13				
Opponent	Game Date	Simpson Score	Opp Score	Loc
% Crown (MN)	9/1/2012	11	0	Home
% Bethel (MN)	9/2/2012	3	0	Home
% Grinnell	9/5/2012	2	0	Away
% Millikin	9/8/2012	5	0	Home
% Ill. Wesleyan	9/9/2012	0	1	Home
% Neb. Wesleyan	9/12/2012	2	3	Home
% Hendrix	9/14/2012	3	4	Away
% Texas-Tyler	9/15/2012	0	3	Neutral
% Culver-Stockton	9/20/2012	2	1	Away
% Coe	9/22/2012	2	0	Away
% Augustana (IL)	9/23/2012	2	1	Home
% Iowa Wesleyan	9/26/2012	5	2	Home
% Wartburg	10/4/2012	2	2	Home
% Dubuque	10/6/2012	2	3	Away
% Central (IA)	10/10/2012	5	0	Home
% Luther	10/13/2012	1	4	Away
% Amer. Inst. Business	10/16/2012	6	1	Home
% Cornell College	10/17/2012	5	1	Away
% Loras	10/20/2012	0	7	Home
% Buena Vista	10/23/2012	2	0	Away
% Wartburg	10/27/2012	1	0	Away
% Loras	10/31/2012	0	6	Away
	Average			
	Mean	2.772727	1.772727	
	Median	2	1	
	Std Deviation	2.521396	1.99845	
	Mode	2	0	
	Other			
	Max	11	7	
	Min	0	0	
	Sum	61	39	
	Count	22	22	

Info Sheet | 2014-15 | 2013-14 | **2012-13** | Final Averages

Finally, create a separate tab to hold the summary data:



	A	B	C	D	E	F	G	H	I	J	K	L
1		Averages	Mean	Median	Standard Deviation	Mode		Other	Max	Min	Sum	Count
2	2014-2015		68.96	69.5	14.13528256	70			107	35	1793	26
3	2013-2014		69.77	72.5	12.22012115	74			94	52	1814	26
4	2012-2013		71.8	71	8.855883167	67			91	56	2154	30
5												

... 2013-2014 2012-2013 All Averages

Double-check the rubric that is posted for this assignment. Make sure you have done the assignment in a way to get full points. Then upload the assignment on Scholar.

Assignment 3: Annotated Bibliography

You will be writing a report on 'Big Data' and its impact on *something*. You get to pick that something.

Remember, the topic *must* tie into Big Data. The most common grade-killing mistake with this paper is to write a paper that has nothing to do with Big Data, or only have the smallest link to Big Data.

Here are some prior paper titles students have written:

- Big Data Baseball: Defensive Sabermetrics
- Big Data + Fantasy Football = Success
- How Big Data is Changing Sports
- How Big Data is a Plus to Our Society
- Big Data in Football
- Big Data in Nike Sales
- Big Data and Insurance
- Amazon and Big Data
- Big Data and Personal Data Privacy
- What is Data Mining?
- Big Data Increasing Concert Ticket Prices?
- Big Data in Shopping
- Big Data in Computers
- Big Data and the Economy
- Fitbits and More - Personal Big Data

To get started, find some facts. Do some research and gather ideas. A great way to do this is to create an annotated bibliography for your report.

The bibliography must have:

- Five or more sources.
  - One source must be a book. You **must** identify in your annotated bibliography which of these is a book.
  - One source must be an article that isn't exclusively for the web.
  - One source must come from the web.
- Use MLA style to create a bibliographical entry.

- Annotate each entry. That is, write about:
  - What the resource says.
  - Talk about the relevance of the article to your paper. That is, how would you use it? What does it have to do with Big Data? You might have one source that doesn't have to do with Big Data, and just give background information, but if you don't see the tie-in of the source to Big Data, you need to keep looking.
  - Talk about the accuracy.
  - Talk about the bias of the article. People often forget this part. Don't. And remember, just because a paper has only numbers of facts does not make it unbiased. Just because a paper comes from a `.org` website doesn't mean it is unbiased. If you are unsure about this part, ask.
  - Talk about the quality of the source. The Washington Post is a higher quality source than BuzzFeed. Why?

Thursday will be "library day" to work on this.

- Meet at the library.
- Bring a notebook or notebook computer.
- Pick your topic. Pick a back-up topic. Pick a third topic if the first two fail.
- Spend time "prepping" so that you have the format of the bibliography all ready. That is, make sure your file is ready to start entering info, so you don't spend the time messing about with formatting rather than doing research.
- See the example annotated bibliography on Scholar.
- Try finding a few starter resources before you even get to class. This will give you more time to ask questions in class and get feedback. Otherwise you might be stuck with questions when you don't have anyone to answer them.
- Have you done enough prep-work to make sure that you've got questions? Great! then you are ready for Thursday's class.
- While at class on Thursday use Maddy, Christopher, Dr. Craven, and the library staff to help you in your research.
- Finish the bibliography outside of class, and upload to Scholar.

Next assignment will be to create an outline, then we will write the paper.

Grading rubric:

Includes name, date, title at top.	No *0 points*		Yes *1 points*		
Source 1	Missing, or missing much of the needed discussion. *0 points*	Missing two item, or two items aren't well stated. *1 points*	Missing an item, or an item isn't well stated. *2 points*	Contains discussion on relevance, accuracy, bias, quality *3 points*	
Source 2	Missing, or missing much of the needed discussion. *0 points*	Missing two item, or two items aren't well stated. *1 points*	Missing an item, or an item isn't well stated. *2 points*	Contains discussion on relevance, accuracy, bias, quality *3 points*	
Source 3	Missing, or missing much of the needed discussion. *0 points*	Missing two item, or two items aren't well stated. *1 points*	Missing an item, or an item isn't well stated. *2 points*	Contains discussion on relevance, accuracy, bias, quality *3 points*	
Source 4	Missing, or missing much of the needed discussion. *0 points*	Missing two item, or two items aren't well stated. *1 points*	Missing an item, or an item isn't well stated. *2 points*	Contains discussion on relevance, accuracy, bias, quality *3 points*	
Source 5	Missing, or missing much of the needed discussion. *0 points*	Missing two item, or two items aren't well stated. *1 points*	Missing an item, or an item isn't well stated. *2 points*	Contains discussion on relevance, accuracy, bias, quality *3 points*	
Has at least one book.	No *0 points*		Yes *3 points*		
Has at least one non-web (not including the book from above. Two books would be ok.)	No. *0 points*		Yes. *3 points*		
Has at least one web-base article.	No. *0 points*		Yes. *3 points*		
Spelling and grammar	Many mistakes, very difficult to read and	Several mistakes, distracts from understanding.	A few mistakes. *4 points*	One mistake *5 points*	No mistakes. *6 points*

Assignment 4: Outline

The goal of this assignment is to create an outline of the "Big Data" paper. First, read the *Tutorial: Outlining*.

You should include:

- Name, date, title at the top. Even if you submit it electronically, still put this up there.
- A draft **introduction**. Include a thesis statement. I should be able to easily find the thesis statement, and not get confused about which sentence it is. The thesis needs to involve Big Data. That is, anything that involves processing thousands or millions of data items.
- Background on the topic. Depending on how much background is needed, this might be part of the introduction, or the paper might need additional paragraphs.
- **Major Points**. You will likely need a major point for each paragraph. Each paragraph is about 50 words. So you'll need about 1000 / 50 = 20 points. So you should have 20 or so paragraph-worthy points. Some points might be worthy of multiple paragraphs. If so, you'll probably have a major point, and then minor points for each paragraph.
- Do your points have to do with Big Data? Make sure to clearly tie it in. Don't assume the reader will do so.
- Put **citations** in with major points the support. This is important! Put a proper MLA in-text citation next to the point in your outline where you plan on using it. This is the most frequently missed point.
- **Conclusion**. Don't just restate your thesis and your introductory paragraph. Figure out how to bring this paper to a strong close.
- Double check to make sure your points support the thesis.

Here's the grading rubric:

Has thesis statement	No *0 points*		Yes, but thesis is unclear *1 points*		Clear thesis *2 points*	
Talks about background	No *0 points*		Unclear what will be said about the background, or it is incomplete *1 points*		Yes *2 points*	
Has about 20 paragraph-worthy points	7 or fewer *0 points*	9 *1 points*	12 *2 points*	15 *3 points*	18 *4 points*	About 20 *5 points*
Organization	Unorganized *0 points*	Ok *1 points*	Good *2 points*	Great *3 points*		
Shows where citations will go	No *0 points*		A few *1 points*		Yes *2 points*	
Conclusion	None *0 points*		Unclear *1 points*		Yes *2 points*	

Assignment 5: Big Data Paper Version 1

Turn in your Big Data paper. Our "Writing Fellow" will proof the paper and offer suggestions.

While this version is not graded, it *must* be complete and turned in on-time. Otherwise there will be **no grade** for version 2.

- Include name, date, and title.
- Minimum 1,000 words for the body of the paper. This does not include the title or bibliography.
- Your thesis should be the last sentence of the first paragraph.
- Do not use first person references. Avoid second person references. No "I", "me", "you."
- Spell out numbers ten or less.
- Paragraphs should be about five sentences long. Follow P.I.E. when creating your paragraph.
- Include in-text citations. (Author, Year)
- Include your bibliography.
- Do not include the annotations in the bibliography.
- Make sure the tie-in to Big Data is clear.
- Proof-read the first paragraph multiple times. Say it out loud. This is the most important paragraph to get right.
- Check to make sure you do not introduce new topics in your conclusion.

Assignment 6: Big Data Paper Version 2

Turn in your Big Data paper.

This version will be graded, as long as you have turned in version 1.

- Do NOT turn in your paper with all the comments still on it.
- Include name, date, and title.
- Minimum 1,000 words for the body of the paper. This does not include the title or bibliography.
- Your thesis should be the last sentence of the first paragraph.
- Do not use first person references. Avoid second person references. No "I", "me", "you."
- Spell out numbers ten or less.
- Paragraphs should be about five sentences long. Follow P.I.E. when creating your paragraph.
- Include in-text citations. (Author, Year)
- Include your bibliography.
- Do not include the annotations in the bibliography.
- Make sure the tie-in to Big Data is clear.
- Proof-read the first paragraph multiple times. Say it out loud. This is the most important paragraph to get right.
- Check to make sure you do not introduce new topics in your conclusion.

Assignment 7: Big Data Paper 2 Pre-work

This is the research for the second report. You need to turn in:

- Initial data analysis (in MS Word format)
- An outline
- A bibliography
- Supporting Excel files

You can turn in the analysis, outline, and bibliography as one document.

The goal for the paper will be take some of the big data analysis techniques we learned, and apply them to learn something new.

Do not talk about what you could do. Actually do it. This has been the most frequent mistake made in the years past. Don't say "From this, we could determine what players we need to cut from our team." Instead say, "Because of x, y, and z, we should cut Bob, Fred, and George."

The final paper should have the following parts:

- **Introduction and thesis**
  - Because of the nature of this report, the main topic will be the processing of the data and its analysis. You may have one analysis on football and another analysis on smoking in this report. Or you could do both on football. Your goal is to show the reader what you can do with the data. Write your introduction accordingly.
  - Clearly label the thesis.
- **Text Processing**
  - Search up something out of the text files.
  - Create something new. Don't just re-create what we did with the name processing. At the very least, compare name trends in different states.
  - Describe what you are doing for processing, and how you are doing it.
  - Talk very specifically about the results.
  - If you want, you can feed this data into the graph and/or pivot table.
  - Do **not** do the exact same thing we did in class. Process different text files than the "name" text files. (You can compare names between states, that would be new.)
- **Graph**
  - Graph the data.
  - Label your x and y axis. Give the graph a title.
  - Describe in detail what this graph is showing. Don't just graph some random thing. Graph data that is informative.
  - Then describe why it matters, and what exactly you can derive from that data.
  - Not all data sets I've listed are good for pivot tables. For pivot tables you need data that has multiple categories on each line. Out of the *Data Sets*, I'd recommend looking at:
    - California ACT/SAT data
    - Minnesota Payroll Data
- **Pivot Table**

  - Use a pivot table.
  - Make one or more pivot tables that inform the reader.
  - Specify how you created the table. Where you got the data from.
  - Tell why the data you are showing matters and what you can do with it.
  - Create a pivot table that shows the power of what a pivot table can do. Don't just create a pivot table that is a recreation of the original table. Part of the purpose here is to show the user what a pivot table can do.
- Conclusion
- Bibliography
- Also - Upload your supporting Excel files.
- The paper should have 1,000 words or more. So 20-25 points in your outline.

However, this part of the assignment is just the pre-work. For this assignment I am looking for:

- Is the data analysis included in the outline? Make sure the data gets copied from your original sources into the MS Word document. I'll only look at the Word document, so if it only exists in an Excel document you won't get credit for it.
- Do you have a first version of your text processing?
- Do you have a first version of your pivot table?
- Do you have a first version of your graph?
- Do you have an outline showing what you will talk about?
- Do you show in the outline where you will cite items?
- Do you have at least one additional source that you've pulled into your outline discussion? (Not a raw data source.)
- Do you have a bibliography that cites the data that you used? And any other background info?s
- Do you also have the supporting Excel files you are turning in?

Assignment 8: HTML Assignment

Use the lesson on HTML and create your own web page. Make the web page about you, your dog, your favorite team, or whatever.

Use many of the elements we learned in class.

- HTML Tutorial
- CSS Tutorial
- Alternative site for HTML/CSS reference

Assignment 9: Big Data Paper 2

This is your second report.

Turn in a complete first draft. Our Writing Fellow will review the papers and give suggestions.

You **must** turn in a first draft to get a grade on the paper.

The goal for the paper will be take some of the big data analysis techniques we learned, and apply them to learn something new.

Do not talk about what you could do. Actually do it. This has been the most frequent mistake made in the years past. Don't say "From this, we could determine what players we need to cut from our team." Instead say, "Because of x, y, and z, we should cut Bob, Fred, and George."

The final paper should have the following parts:

- **Introduction**
  - Because of the nature of this report, the main topic will be the processing of the data and its analysis. You may have one analysis on football and another analysis on smoking in this report. Or you could do both on football. Your goal is to show the reader what you can do with the data. Write your introduction accordingly.
- **Text Processing**
  - Search up something out of the text files.
  - Create something new. Don't just re-create what we did with the name processing. At the very least, compare name trends in different states.
  - Describe what you are doing for processing, and how you are doing it.
  - Talk very specifically about the results.
  - If you want, you can feed this data into the graph and/or pivot table.
- **Graph**
  - Graph the data.
  - Label your x and y axis. Give the graph a title.
  - Describe in detail what this graph is showing. Don't just graph some random thing. Graph data that is informative.
  - Then describe why it matters, and what exactly you can derive from that data.
- **Pivot Table**
  - Use a pivot table.
  - Make one or more pivot tables that inform the reader.
  - Specify how you created the table. Where you got the data from.
  - Tell why the data you are showing matters and what you can do with it.
  - Create a pivot table that shows the power of what a pivot table can do. Don't just create a pivot table that is a recreation of the original table. Part of the purpose here is to show the user what a pivot table can do.
- Conclusion
- Bibliography
- Also - Upload your supporting Excel files.
- The paper should have 1,000 words or more.

Assignment 10: Salary Research Essay

Take a look at the PayScale Research Site.

Look at two careers you might be interested in. Write an essay on the data that you see on the site. Feel free to cut/paste images (ask how to use the 'snip' tool if you need) to help your essay.

- Put a name, date, and title at the top.
- What two jobs did you pick?
- What is the median salary? What is the range? What about bonus?
- How do the two jobs compare in salary?
- How does the location and experience affect the salary?
- What are the common "career paths" for the job(s) you are interested in?
- How many survey responses were part of the salary figures?
- What other interesting facts/figures are there about the job?

Paper 3 - Final Paper Details

Work for the final paper will be done in groups.

Assignments

Annotated Bibliography (3 sources, individual)	10 points	Tue 11/15 8 am
Research (individual)	10 points	Tue 11/22 8 am
Create draft	0 - required	Tue 11/29 8 am
Final presentations (group pres, ind grade)	10 points	Tue 12/6 and Thur 12/8
Create final paper (group, group grade)	30 points	Tue 12/13 10 am

Goal

Step 1: Create and present a paper related to an aspect of Big Data. I will be looking for the same types of things in Paper 3 that you did in Papers 1 and 2. There needs to be a research aspect to your paper. Do number crunching, dive into the books, come up with something new and interesting.

Below are some topics that students have successfully completed in prior years while taking this class:

- Fastest growing U.S. cities. Looking at city population by U.S. census data, and magazine listings they looked at fastest growing large and small cities. The also looked at job growth, economy, and more.
- Another paper looked at the spread of disease in pandemics. Graphs and data showed how cholera outbreaks changed week to week. The paper also looked at influenza outbreaks.
- One paper looked at coffee prices and oil prices to see if there was any correlation. It looked at which countries produced coffee and oil. There was a correlation with a few countries, but not with other countries. Mostly it showed that oil prices did not seem to affect coffee prices.
- "Living the ideal life." This paper looked at multiple sets of data and found that the ideal Iowa life would be:
  - Go to Grinnell College and major in Biology.
  - After completing your undergraduate studies, the ideal medical school is the University of Iowa Carver College of Medicine, specializing in being a Family and General Practitioner.
  - The ideal place to live is in Waukee in the neighborhood between NW 142nd Street and Hickman Road, in the zip code 50325.

  - The ideal house for sale in this neighborhood located at 2674 NW 152nd Street, Clive, IA. 50325.
  - Lastly, the ideal car to buy is the 2014 Toyota Prius C.
- Another paper looked at multiple sets of data to figure out which was the best U.S. state to live in.
- The last paper looked at the amount of people following football teams in social media, and see if that corresponded to how well the team's athletes were paid, game attendance, etc.
- Recruiting new Football by the numbers. What stats to look at, and then went into specific examples showing how to comb through lots of stats to find the best recruits.
- America's most popular sport - tracking the changing popularity of sports over time.
- How to write a hit song. Tracked stats on age of artist, album number, home state, producer, most popular words, number of words, tempo, and more.
- College football attendance - combed through a lot of records to track attendance over the years at top colleges.

Other topic ideas that people haven't done yet, but would be good:

- Presidents by the numbers. Pick a couple recent presidents and show how things changed during their presidency. (Provide a balanced look.)
- Energy production. Take a look at oil imports, exports, and alternative energy production over the years.
- Crime in the U.S. Track different crime rates based on location and type of crime.
- The changing demographics of the U.S. Look at the U.S. Census Bureau and show how the U.S. has changed over time.
- College Rankings: Create your own method of ranking colleges, using IPEDS and other databases.
- Where does the U.S. govt money go? Comb through current and historical U.S. budgets and figure out where the money went.
- Health - Take a look at the county-by-county breakdown in the U.S. on health and show what you find.

Step 2: To begin with, talk about different topic ideas in your group. Then go out and find resources for your paper. Split the labor and have each person work with three or more resources. Each person will annotate their bibliography, with specific emphasis given to how you will work the source into your paper. Find raw data sources as well. I'll grade each bibliography separately. Three sources is a minimum and probably won't get you an 'A.'

Step 3: Split the research. Have everyone do some number crunching, information gathering, and any other research. Each person will turn in their research results separately and be graded separately.

Step 4: Bring it together. Make a cohesive outline. Pull the research and sources into a final paper. Edit it and create a draft copy. Each person reviews the draft and improves it. Then submit and get the Writing Fellow's feedback. No draft, no grade for the final paper.

Step 5: Create a presentation. Use Presi or some other tool that allows each team member to contribute. Turn it in. All members get the same grade.

Step 6: Each team needs to present at least 10 minutes, and no more than 15 minutes. Grades are given individually, so if you don't talk, you get a zero.

Step 7: Turn in your final paper. All group members get the same grade.

Assignment 11 - Paper 3 - Annotated Bibliography

Each person in your group should come up with at least three data sources for your paper. These could be books, magazines, web articles, or raw data sources. If you have three people in your group, then you should end up with a minimum of nine data sources.

Coordinate with your group so that you don't have the same people researching the same sources.

Remember - doing the "minimum" does not mean that you get an "A". To get an A you need to do A work.

Each bibliography entry should provide:

- Citation
- Quick summary of the source material.
- How you could tie it into your paper. What does it have to do with the topic and how is it Big Data related?
- Quality of the material. Is it good?
- Reputation of the source. Does this come from a reputable source, like the U.S. Census Bureau, or does it come from some "unnamed source" on the web?
- Bias of the source. Does the source seemed biased? If it is just data, could there be sampling bias?

Sample annotated bibliography.

If you can't find enough source material, or if you can't solidly tie the sources to your thesis and Big Data, then rethink your topic.

Assignment 12 - Paper 3 - Research

Paper 3 - Research

Turn in an MS Word document with your research.

- Include name, date, title.
- Provide your data. You'll likely need to copy/paste it into the document. I don't want separate Excel documents uploaded. Present you data the way it would be in the paper.
- Note that in the rubric there are different levels for how much processing you did with your data. Let me know what you did for processing. If you used a pivot table, say so. If you used other tools, talk about them.
- Make your data look presentable
- Talk about possible bias in the data.
- Analyze the data. What does it mean? Explain it.
- Once you've got you data, you'll put the analysis of your group together and then writing a paper around it.

Research Rubric

Source	Source(s) for data of research not stated ***0 points***	Source(s) were clearly stated ***1 points***	Source(s) clearly stated, and possible bias was evaluated ***2 points***	
Analysis	No analysis done ***0 points***	Not much analysis done ***1 points***	Normal analysis done. Not particularly insightful. ***2 points***	Introspective and informative data analysis done. ***3 points***
Presentation	Data is unclear, presentation is poor. ***0 points***	Data is unclearly presented, or unattractively presented. ***1 points***	Data and conclusions are clear. Aesthetically pleasing presentation. ***2 points***	
Tools	No tools used ***0 points***	Basic Excel usage ***1 points***	Normal Excel usage ***2 points***	Advanced tools used (pivot tables, egrep, sql, or some other advanced tool) ***3 points***

CHAPTER 3

Data Sets

The easiest subjects to find data on are sports, the economy, health, education, and crime.

Meta-sets

- U.S. Government Open Data (Over 189,000 data sets)
- U.S. Census Bureau
- Kaggle Data Sets Lots of different data sets and some analysis tools as well.
- Google Public Data
- Amazon Public Datasets
- Awesome Public Datasets

Health

- HealthData.gov
- World Health Organization
- County Health Rankings
- **Center for Disease Control and Prevention (CDC)**
  - Sexually Transmitted Disease Morbidity
  - Trend Tables
- California Immunization Levels

Crime

- Law Enforcement Deaths
- People Killed by Police
- Uniform Crime Reporting Statistics
- Crime in the U.S. (Click on the year, then click on the report, then find the data tables.)

Education

- NCAA Graduation Success Rates
- National Center for Education Statistics (IPEDS)
- California ACT/SAT data

Economy

- US Budget, Deficit, and Debt
- US Oil Imports
- Minnesota Public Payroll Data

Entertainment

- Pokemon Stats

Sports

- College Football Statistics
- NCAA Statistics

CHAPTER 4

Large Text Data Sets

Names

Names for the whole U.S. are in the `names` folder. Names broken down on a state-by-state basis are in the `names_by_state` folder.

Original source is the Social Security Administration.

FEC

FEC Data

Contributions to political campaigns political/individual_contributions.txt (fields definitions)

Automotive

Complaints, defect investigations, and recalls for different cars. Source is the National Highway Traffic Safety Administration.

A list of fields and descriptions for the 'complaints' table.

Medicare

Medicare claims from Centers for Medicare & Medicaid Services

Weather

All of this data is in the `weather` folder.

The original source for the weather data is from NOAA (Site, FTP)

File info:

- weather/ghcnd-stations.txt – List of stations. Egrep what station you want.
- weather/*.dly – Weather files. Format description is here

Example processing weather data:

- `egrep TMAX USC00134063.dly | sed 's/^.\{11\}//' | sed 's/[09] .*/\1/' | sed 's/TMAX//' | sed 's/^[* 09]{4}/\1 /' | sed 's/^.{11}[09]/\1\.\2/' > ../craven.csv`
- `egrep TMAX USC00134063.dly` # Search the text file for 'TMAX* ' which is the temperature maximum
- `sed 's/^.\{11\}//'` # Use sed and regular expressions. From * the beginning of the line, remove 11 characters
- `sed 's/[0* 9] .*/\1/'` # Use sed to remove all of the line after the last temp reading.
- `sed 's/TMA* X//'` # Remove the TMAX from the line
- `sed 's/^[09]{4}/\1 /'` # Add a space after the year so it doesn't run into the month
- `sed 's/^.{11}[09]/\1\.\2/'` # Add a decimal into the number, because 345 is actually 34.5
- `> ../craven.csv` # Redirect to a file. Do it one directory up because there are way too many files here

CHAPTER 5

Example Data Analysis for Discussion

Entertainment

- John Goodman Is America's Greatest Supporting Actor
- The average color of every frame of a given movie, compressed into a single picture.
- TV finales that surprise/disappoint
- Popular Buzzfeed Clickbait Titles
- A Data Analysis of Board Game Rankings

Sports

- One Race, Every Medalist Ever. Video comparison of every 100 meter dash medalist since the 1896 Olympics.
- College Football Player Hometowns
- Every Shot Kobe Bryant Ever Made

History

- Plot millions of journal entries from 18th and 19th century ship logs, and you reveal a picture of ocean trade you've never seen before.
- 30 Most Edited Wikipedia Pages

Health

- Heart rate (bpm) during marriage proposal

- Animated Fertility Rates Over Time
- Bed Net Rates in Africa
- My Path to Sobriety
- Showering Rates

Transportation

- Animated Map of Ocean Shipping
- Plane Finder

Finance

- Emergency Bank Loans During Bailout
- Panama Papers
- Where the U.S. gets its oil from

Politics

- FiveThirtyEight Election Forecast
- Donald Trump is wrong that 'inner-city crime is reaching record levels' (9/1/16)
- Brexit Vote
- Republican vs. Democrat Occupations
- 100 Years of Presidential Elections
- Elections by region
- Rise of Partisanship

Society

- Ashley Madison
- Average Reddit Score vs. Number of Words
- Gender Breakdown of Jobs

Food

- Skittle Color Distribution

Science

- Radiation Levels
- Black Hole Sizes
- Live Earthquakes Map
- Who is Pirating Scientific Papers

Climate Change

- Changing opinions on climate change, from a CNN meteorologist

Misleading Charts

- That Map from The Washington Post About Female/Male Ratios Is Way Off. Here's a New One...
- The most misleading charts of 2015, fixed