Lab 1. Why William & Mary?

Due by noon on Friday, September 11th

"I look to the diffusion of light and education as the resource most to be relied on for ameliorating the condition, promoting the virtue, and advancing the happiness of man.

Letter from Thomas Jefferson to C.C. Blatchley, 1822

Laboratory in Brief:

The purpose of this laboratory is to think about ways to justify a decision using raw, unprocessed data, while also introducing the skills you'll need to do more advanced work later in the semester. You may be asking yourself, will I really need to defend my choice to attend William and Mary for another two weeks?" The goal of this assignment is not only to encourage you to consider in broad terms the college you have chosen and are now attending (which, by the way, was a good one), but also to begin thinking about how to obtain, manage and use data in support of this supposition.

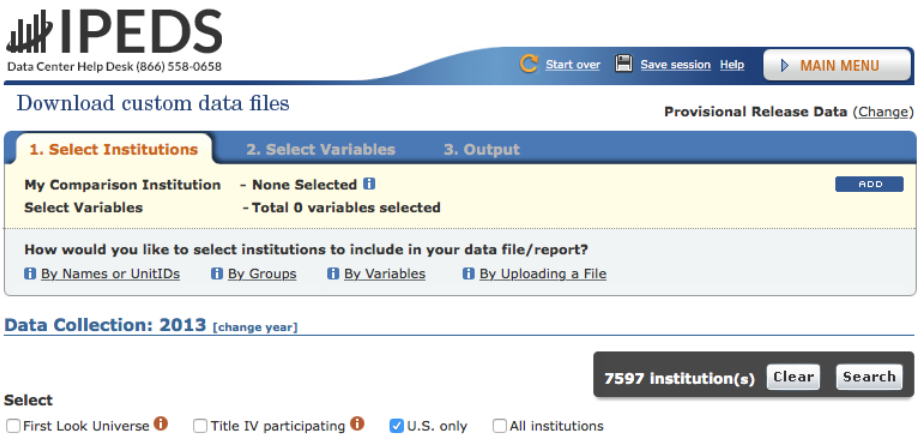
Goals of this Laboratory:

1. To use "a lot more data" to explain to yourself, parents, significant other, or your best friend that William and Mary really is "the best."
2. To begin learning how to use the Statistical Computing Language R while also understanding how to import data to R from a `CSV` file.
3. To visualize your findings for "System 1" (we'll be learning all about this in the lecture!).

Session 1: Monday, August 31st

Step by Step Instructions: First we Acquire the Data

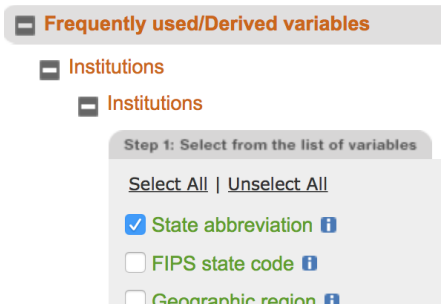
1. Let's download a file which gives statistics on colleges across the U.S. In your browser go to the following address and become familiar with the National Center for Education Statistics website, which we will be using in this lab.
   `https://nces.ed.gov/ipeds/datacenter/`
2. Find the **Download Custom Data Files** link on the left side of the webpages graphical user interface and select this button. Once you have followed this link, you will arrive at a page that inquires *"How would you like to select institutions to include in your data file/report?"* Three options are available, but you will want to choose the link for **By Groups**. A sub-window will appear and then also choose **EZ Group**.
3. Now you should arrive at a screen that asks for more detail about which institutions you want to select from the database. Select the **U.S. only button** and then the Search button (there were 7597 at the time I did the search). The screen should ap-



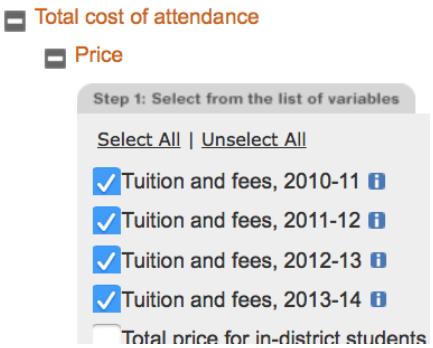
pear like this.

4. On the next page, look for the **Continue button** which is located below the **Download custom data files** box and after the statement *When you have finished selecting institutions* CONTINUE to Step 2 - Select Variables.

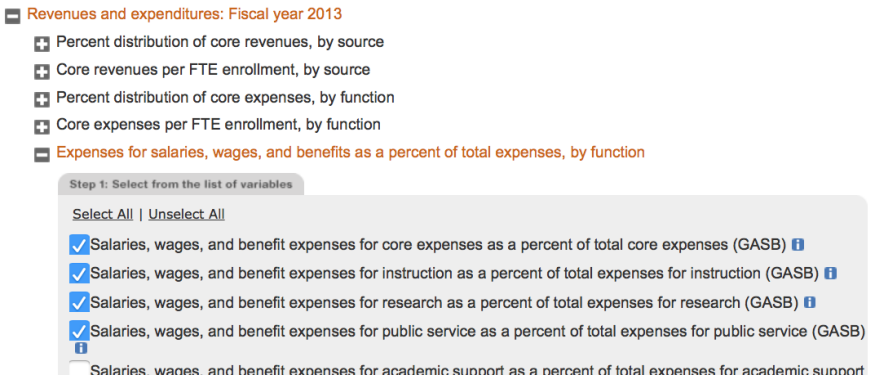
5. On the following page you will begin selecting the individual variables that will be used in your analysis. First find the **Frequently used/Derived variables** tab and expand that under **Institutions** until you see several check mark boxes



that selection. Choose **State abbreviation**. Then under **Total cost of attendance** and **price** select the check boxes for the first four



annual sets of **Tuition and fees**. Next click on the **Revenues and expenditures: Fiscal year 2013** and choose the **Expenses for salaries, wages and nbenefits as a percent of total expenses, by function**. Finally, check the first four boxes: **core expenses, instruction, research and**



public service.

6. Keep in mind that for this stage of the lab we will walk you through one very specific analysis, but to get the maximum grade on this assignment you will need to download more data to make your own case, so you may want to choose one or two additional variables you think are interesting now.

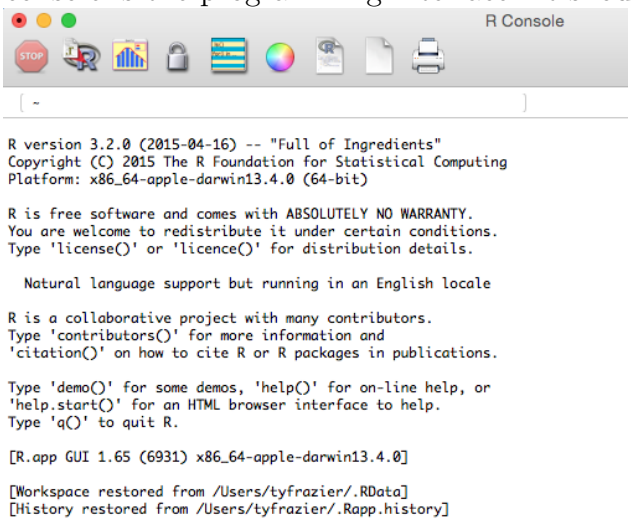
7. Finally, scroll back up to the top and click 

, then under the headline bar

Year 2013 download your data as a CSV.

Now We Import the Data to R

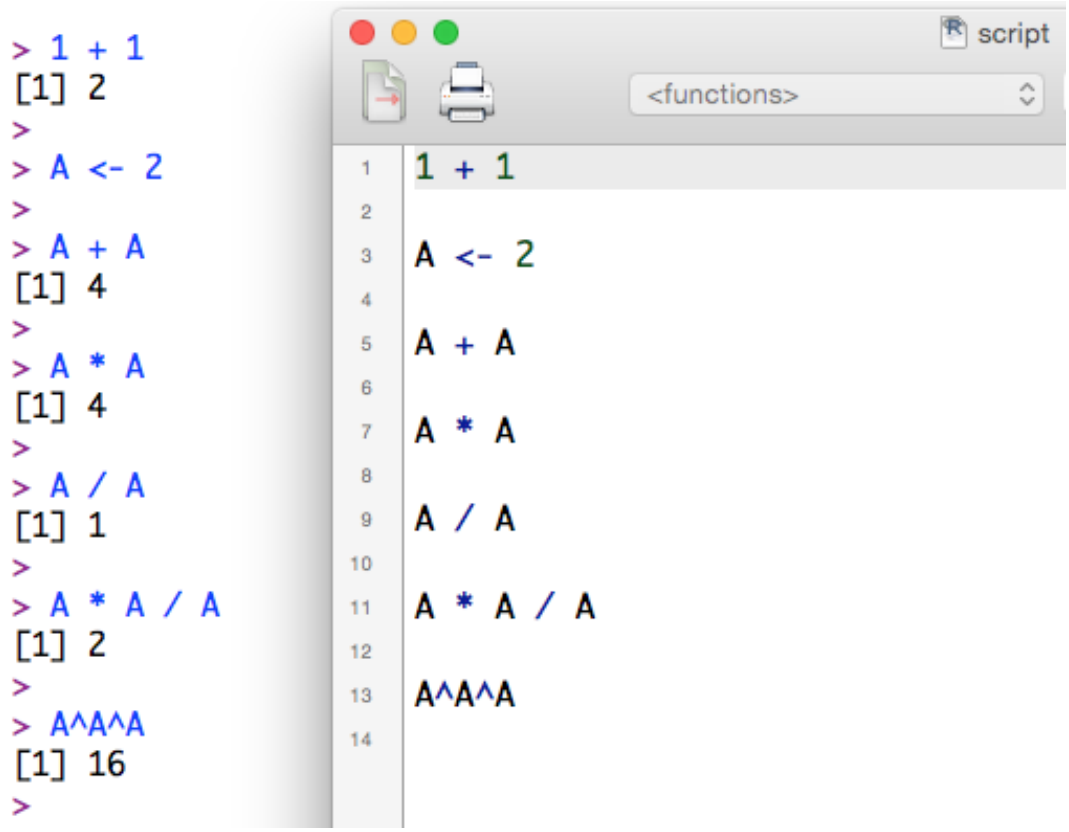
8. Open R or R Studio

9. Identify where the console is the programming interface. It should look something like the following. 

Go ahead try entering a few commands. Try entering the following commands in the R console.

```
> 1 + 1
[1] 2
> A <- 2
> A + A
[1] 4
> A * A
[1] 4
> A / A
[1] 1
> A * A / A
[1] 2
> A^A^A
[1] 16
```

10. Now lets create a script file. In the R user interface find the drop down menu for **file** and choose **New Document**. Creating a script file enables us to do two things. First we can save our code and reuse it later. Second can use our script document as the source of which commands we will send to the console for execution. Go ahead and write the same code in your script file from item 9 and then use the **command enter** key combination in order to send those commands to the console.



Be sure to save the file and give it a name. I recommend just calling it "lab1.csv" in order to keep it simple. Using your William & Mary "H" drive is a good idea.

11. Since now you have some basics under your belt, lets go ahead and load the CSV file/folder we downloaded from the National Center for Education Statistics. First copy the file into your William & Mary H:\\ drive so it doesn't get deleted. To do this, open up "My Computer", go to "Downloads", and copy the file you downloaded (click on it and push "ctrl+c"). Now, go back to My Computer, choose your H drive, and paste it (ctrl+v on your keyboard). While in this case, you probably won't have to work with zip files, in the future you will likely need to learn how to extract files from a zip. I recommend you google "how to extract a zip file" if you need help, but you can also raise your hand to ask an instructor or TA. This stuff can get complicated, so never hesitate to ask if you get lost!

12. Back in R, in the window at the top, type in the following commands below.

```
R> setwd("H:\textbackslash\textbackslash")

R> collegeData \textless - read.csv("lab1.csv")

R> view(collegeData)

R> str(collegeData)

R> names(collegeData)

R> dim(collegeData)

R> nrow(collegeData)
```

13. Click on each line and then click "run" like we did before to see each line in action, or select the entire script and click "run" to run it all at once.

Session 2: Wednesday, September 2nd

Step by Step Instructions:

1. For starting out on day 2, let's actually run a some analysis in R. For example, let us consider the question, how many colleges are in each state? When you type in `names(collegeData)`, you'll notice one of the names is `HD2013.State.abbreviation` - one of the columns you requested when you did the download. To see how many colleges are in each state, type in:

   ```
   R> table(collegeData$HD2013.State.abbreviation)
   ```

2. Now, let's use a different command to identify the average tuition across all colleges:?

```
R> summary(collegeData$DRVIC2013.Tuition.and.fees..2013.14)
```

3. We can also do the same thing, but only looking at Virginia schools ? does Virginia have a higher average tuition or lower than the rest of the country?:

```
R> VA_colleges <- collegeData[which(collegeData$HD2013.State.abbreviat
R> summary(VA_colleges$DRVIC2013.Tuition.and.fees..2013.14)
```

4. What if we want to get really clever, and identify only those institutions in VA that are above the national mean? Is William and Mary above the national mean?

```
R> AboveUSMeanVA <- collegeData[which(collegeData$HD2013.State.abbrevi
R> View(AboveUSMeanVA)
```

5. You can read through each entry to figure out what William and Mary's tuition is, or you could type this:

```
R> collegeData[which(collegeData$institution.name=="College of Willia
```

6. You can use similar approaches to find the cheapest and most expensive schools in the country:

```
R> min(collegeData$DRVIC2013.Tuition.and.fees..2013.14, na.rm=TRUE)?R
```

Analysis

7. Now, we're going to do some basic analysis that you can include on your revised infographic. Note the goal of this R tutorial is not to create beautiful visualisations ? though R can do that. Rather, we want to extract data you can then use in visme to improve your original infographic. Let's do something very simple to start: how does the percentage increase of William and Mary's tuition compare to (a) all schools, and (b) Virginia schools from 2010 to the 2013-2014 school year?

8. First, let's calculate this for William and Mary (type in WMchange) at the end.

```
R> WMTuition_2014 <- collegeData[which(collegeData$institution.name==
R> WM_change <- WMTuition_2014 / WMTuition_2010
```

9. Now, let's calculate the average change for every other school. This is the first time we're going to add our own new ?column? to the data frame. To see how this works, first type in the below command, which will show you the current set of columns (i.e., the data available for each college):?

```
R> names(collegeData)
```

10. Now, let's create a new one that tells us how much tuition has changed. Note this is the same thing we did for just William and Mary, but for every school:

```
R> collegeData$All_change <- collegeData$DRVIC2013.Tuition.and.fees..
R> names(collegeData)
```

11. Note now there is a new "name" in collegeData, which represents a new column. Let's summarize the change column now:

```
R> summary(collegeData$All_change)
```

12. Let's do a comparison between the tuition in 2014 and the percent of that money spent on instructors salary. First, let's look it up for William and Mary:

```
R> collegeData[which(collegeData$institution.name=="College of Willia
```

13. And second, let's compare that to the national average:

```
R> summary(collegeData$DRVF2013.Salaries..wages..and.benefit.expenses
```

14. Finally, let's plot out all schools tuition in 2014 contrasted to the % of tuition they spend on instructors salary.

```
R> plot(collegeData$DRVF2013.Salaries..wages..and.benefit.expenses.fo
```

15. That on it's own isn't particularly helpful, as it's hard to see patterns with outliers and without knowing which dot is William and Mary. It's also probably more fair to compare to only Virginia schools, so let's drop out everything that isn't Virginia and label W&M in an appropriate color. First, re-use the old VAcolleges dataset you made:

```
R> plot(VA_colleges$DRVF2013.Salaries..wages..and.benefit.expenses.fo
```

16. Now, let's label just W&M ? save this figure (click the ?export? button above the chart) as you will be turning it in!

```
R> WM_plot <- collegeData[which(collegeData$institution.name=="College
points(WM_plot$DRVF2013.Salaries..wages..and.benefit.expenses.for.inst
```

Session 3: Monday, September 7th

Step by Step Instructions:

1. Go back to Step 1 from Session 1 and choose at least one variable that was not included in this analysis. Follow all the steps from Session 1.

2. Conduct similar analysis from Session 2, plot your variable and save the output.

Session 4: Wednesday, September 9th

Step by Step Instructions: Lab Questions

1. Calculate which two states have the most colleges, and how many does each one have?
2. Calculate the mean tuition cost for the US and for Virginia? Does Virginia have a higher or lower average tuition cost than the rest of the USA? What is this difference?
3. Calculate the 2010-2011, 2011-2012, 2012-2013, and 2013-2014 Tuition and Annual Percent Change for William & Mary, all Universities in Virginia, and all Universities in the USA.
4. Create a graph showing USD invested in instruction as compared with tuition costs and highlight William & Mary. Create a similar chart considering the mean of these two values for all Universities in Virginia and all Universities in the USA.
5. Calculate what are the most and least expensive Universities in Virginia and the USA.
6. Attach a printed copy of your "visme" visualisation to this document, or submit it electronically. What variable did you include that was not explicitly a part of this lab? What challenges did you have in retrieving it, and how did you use it to illustrate William and Mary is an exceptional (or, crazy!) choice of where to get your degree?

Stretch Goals
Made it this far? Want to try to go even farther? Just want to learn? Try any of the following to really impress us, it won't count for your grade, but it'll give you a leg up on future assignments:

- In your download, ask for the "latitude" and "longitude" columns, then plot colleges following this tutorial: `http://www.milanor.net/blog/?p=594`
- Examine how US schools compare to international schools.
- Make similar comparisons with a different dataset ? i.e., try

Final Output for Submission

Due by noon on Friday, November 13th:

... Make certain the Final Report meets the following criteria.

- Single spaced with block paragraph style and 10 to 12 font, preferably New Times Roman, Arial, Courier or a similar font.
- x to x pages in length
- Output from Session 2
- Output from Session 3

- Answers to Questions Considered during Session 4

Grading

This lab will be graded based on the six deliverables identified on the final page of this lab. The first five questions are worth 50% of your grade, and are based on your capability to work through the steps below (i.e., if you follow the steps, you'll get the right answer!). The last question ? question 6 ? asks you to update your visualisation you created last week using not only the datasets referenced in this lab, but asks you to download at least one additional variable to use in your update. The process of critically deciding what data to use, how to analyse it, and how to visualise it will be worth the remaining 50% of your grade.