


Data Wrangling with tidy

Open RStudio and create a new project under your Module 5 folder and call it **Mod5Assignment2**. For this assignment, you will be creating an R Markdown document that will include topics previously covered as well as the use of tidy to create tidy data that can be used to generate various visuals.

Once completed, all you need to do is submit the HTML document that is created.

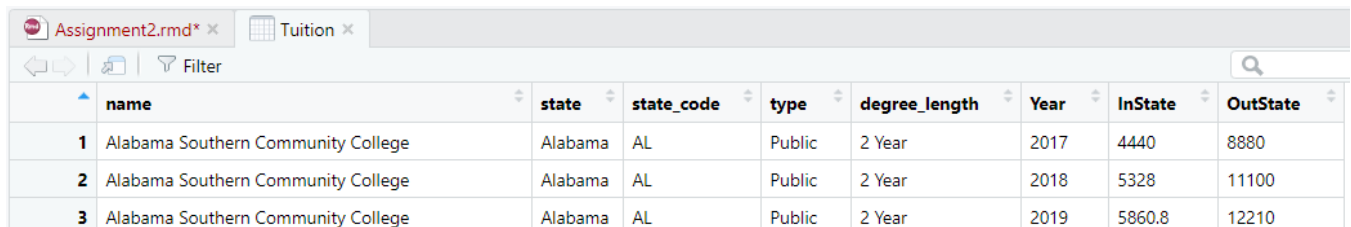
Create the R Markdown Document

- 1.) In RStudio, select *File -> New File -> Text File*. Save this file to your project as **Mod5Assign2Answer.rmd** (it is important to save with the .rmd extension as this saves the text file as an R Markdown file).
- 2.) Create a Header 1 with the title: **Module 5 - Assignment 2**
- 3.) Create a Header 2 with the title: **Last Name, First Name** (replace with your name)
- 4.) Create a Header 3 with the title: **More Data Wrangling**
- 5.) Click on the dropdown arrow next to the Knit icon  at the top of the R Markdown Pane in RStudio and select Knit to HTML. Notice: **FOR THIS PROJECT, YOU WILL CREATE AN HTML DOCUMENT WHEN YOU KNIT THE R MARKDOWN!!**
- 6.) Notice there is now an HTML document in your files for the project named **Mod5Assign2Answer.html**. This is the file you will be uploading later to Canvas.
- 7.) For this assignment, you will need to download the **tuition_cost.csv** file from Canvas. Save this in the same folder that you created the project in. Be sure to include the code to import both this document and load any necessary packages needed at the beginning of the R Markdown document.

Module 5: Assignment #2

Part 1 – Creating tidy data using tidyr

- 8.) Now that we have the data imported and packages loaded, we need to check to make sure the data is tidy. Remember, to be tidy data each row must represent a different observation but right now, rows actually represent multiple observations across a number of years.
- 9.) Create a Header 3 with the title: **Part 1 – Creating tidy data using tidyr**
- 10.) Create a new chunk of R code. Currently, tuition for both in-state and out of state are listed under the years 2017 – 2020. You will need to use pivoting to first create a new column (or name) called **Year** with the values for each year being called **tuition**. When doing this, you should save these changes to a new tibble called **Tuition**.
- 11.) Once you have created the new tibble, you will notice that there are 2 values for each year in the newly created **tuition** column. The first value is the in-state tuition and the second value is the out of state tuition with them separated by a “/” (e.g., 1000/5000). You will need to use the separate function to create two new columns: **InState** and **OutState**.
- 12.) You should now have 2 new columns (**InState** and **OutState**) which have replaced the **tuition** column. One last thing we need to do before moving forward is to change these two columns from character to a numeric. At this point, your file should now look like the following with over 11,000 observations and 8 variables:

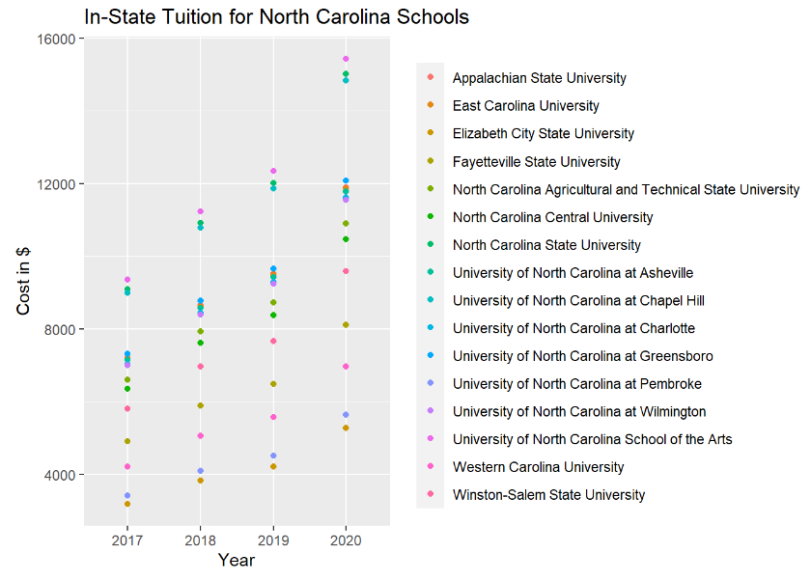


	name	state	state_code	type	degree_length	Year	InState	OutState
1	Alabama Southern Community College	Alabama	AL	Public	2 Year	2017	4440	8880
2	Alabama Southern Community College	Alabama	AL	Public	2 Year	2018	5328	11100
3	Alabama Southern Community College	Alabama	AL	Public	2 Year	2019	5860.8	12210

Part 2 – North Carolina Schools

- 13.) Create a new Header 3 with the title: **Part 2 – North Carolina Schools**
- 14.) Create a new chunk of R code. You are considering starting a new program at a public, 4 year university in the state of North Carolina. Create a new tibble called **Public** that contains just those types of schools (hint: use the filter command covered in previous modules). This should create file that will contain 64 observations.
- 15.) Within the same chunk of code, use ggplot to create a scatterplot of the universities similar to the image below (make sure to include a title for the plot, change the y-axis to text below and remove the legend title):

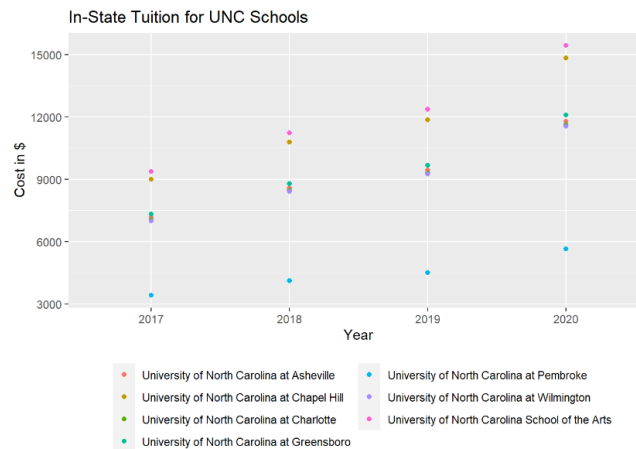
Module 5: Assignment #2



16.)Also, I would like you to create a plot of just the University of North Carolina schools including their satellite campuses. To do this, create a new tibble called **UNC** using the code below to filter only those schools in the UNC system:

```
UNC <- filter(Public,grep("University of North Carolina",name))
```

17.)Once you have created the new tibble called **UNC**, use ggplot to create the following scatterplot (remember to include a title for the plot, change the y-axis to text below and remove the legend title):



18.)You will notice the legend is at the bottom and text is a different size. This can be changed by including the following code in your ggplot:

```
theme(legend.position="bottom") +  
guides(colour = guide_legend(nrow = 4))
```

The legend.position code puts the legend at the bottom and guide_legend puts the universities in 4 rows.

Part 3 – Additional Universities

19.) Create a Header 3 with the title: **Part 3 – Additional Universities**

20.) You are also considering 2 other options about where you would like to attend:

- a. You enjoy skiing so you are considering moving to Colorado to go to a public, 4 year university. Since you will not be a resident, you will need to review the out of state tuition.
- b. You have a family member who lives in South Carolina and you are thinking of going to a public, 2 year university. You are able to use the address of your family member and will qualify for in state tuition.

21.) Using what you have done in Part 2 for all the public universities in North Carolina, create 2 additional scatterplots using ggplot2 of the Colorado universities and South Carolina universities. Remember that you will need to do some filtering to get just the universities needed and also match the formatting of the first plot in Part 2 (e.g., legends, titles, etc.)

22.) Finally, knit your R Markdown to HTML and upload the document to Canvas