CIS 8392
Topics in Big Data Analytics

#Assignment 1

Yu-Kai Lin

Assignment 1

Step 1. Find a messy data from the Internet (google public data and you will see many websites that you can use to find such data)

- The data must violate at least one of the tidy data principles
- Any data format (csv, txt, excel, ...) is okay, as long as you can read it into R. However, you need to convert and save the data as a csv file and include it in your submission.
- Each student will find a unique data. No two students can use the same data.
- Once you find a data, double check that it is not used by another student : CIS 8392 Assignment 1 Data Singup Sheet
- Once you are certain that no other student uses the same data, sign up your data in the data signup sheet. After you signed up your data, take a screenshot (preferably include the date/time) of the sign-up sheet for your own record in case someone else modifies your record.

Assignment 1

Step 2. Use R Markdown to achieve the following:

1. Specify author, date, and title in the YAML metadata of your document
2. Describe the data source, background, characteristics, variables, etc.
3. Load the data into R. Depending on the data format, you need to find an appropriate way to import the data.
4. Show and explain why the data is not tidy. Don't use data that is already tidy.
5. Tidy up the data using `dplyr` and/or `tidyr`
6. Explain why the data is tidy now
7. Create two different & meaningful data visualization out of the tidy data using `ggplot`. (It is not enough to just change one variable in the axis.)
8. Identify the patterns in each plot and explain why they are meaningful

---

Resources to learn R Markdown:

- https://r4ds.had.co.nz/r-markdown.html
- https://rmarkdown.rstudio.com/lesson-1.html
- http://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf
- R Markdown: The Definitive Guide by Yihui Xie, J.J. Allaire, and Garrett Grolemund

Assignment 1

Here are some additional notes about writing a RMarkdown report. Violating these rules may lead to a lower grade.

- **Put the data in the same folder as your Rmd file**. Whenever we run/knit an RMarkdown file, it uses the folder with the Rmd file as the working directory.
- Read the data in your Rmd code chunk **using relative path**. If you use an absolute path, I will not be able to knit the Rmd file to an html file from my end.
    - **In previous semesters, typically about 50% of students did not follow the above two points and lost 5 points in their first assignment which could be easily avoided.**
- You will lose 5 points if for any reason (input path, error in code, etc.) the Rmd file cannot be knitted to an html file.
- All tables (any output of a data frame) must be formatted using kable in your R Markdown report.
- Distinguish headings (## heading) and normal text. We should not put all the text in headings.
- Do not put your discussions/explanations in code chunk. Write them as normal text.
- Do not use `include=FALSE` or `echo=FALSE` in your code chunk. I need to read your code. You may use `message=FALSE`, `warning=FALSE` to suppress messages/warnings.
- Do not write an excessively long line of code. Break it into multiple lines to improve readability.

Assignment 1

Step 3. Knit the R Markdown file (.Rmd) to an HTML file

Step 4. The Rmd, HTML, and csv files must follow the following naming rule:

- `Assignment1-YourLastName-Title With Six Words Or Less.FileExtension`
  - For example:
  - Assignment1-Lin-Twitter Data Wrangling and Visualization.Rmd
  - Assignment1-Lin-Twitter Data Wrangling and Visualization.html
  - Assignment1-Lin-Twitter Data Wrangling and Visualization.csv

Step 5. If the csv file is larger than 5MB, remove some rows such that the file size is 5MB or less

Step 6. Submit the three files (individually) to iCollege

- In other words, do NOT put them in a zip (or other archive) file

Assignment 1

Due by the beginning of next class

Extra credit: the student who has the best report (determined by the instructor) will be given 5 extra points towards the final grade

- Submissions that are too similar would not be considered for the extra credit

Grading is based on the following:

- Grading is based on the submitted files on iCollege. Do not wait till the last minutes before the deadline. You will lose 10 points for late submission. You will receive 0 point if you submit your assignment via email.
- Whether the submission matches the record in the data sign-up sheet
- Whether all required files were submitted to iCollege on time, following the naming rule.
- Whether the size of the CSV file is less than 5MB
- Whether the Rmd file is syntactically correct and can render the html file
- Whether the report has a professional format and style (succinct and yet provides adequate and clear discussions about the data and the plots)
- Whether the report meets the requirements specified in Step 2

FAQs (for this as well as future assignments)

1. **If my submission is time-stamped at 5:30pm (say, uploaded at 5:30:05pm), is it still a late submission?**

    - **Answer: Yes. Regardless which courses you are taking, assignment deadlines on iCollege are always treated as a sharp deadline (5:30pm means 5:30:00.000pm on iCollege). You need to submit your work *before* deadline.**

2. **If I uploaded a wrong file, can I update my submission with the correct file?**

    - **Answer: If it is still before the submission deadline, you can update the file as many times as you want, and only the latest version will be graded. If it is after the submission deadline, you cannot resubmit your work. Please ensure that you upload the correct file when you are done.**

3. **Once the assignment is graded, can I resubmit the assignment based on the instructor's feedback to improve my grade?**

    - **Answer: No. You are welcome to rework your assignment following the instructor's feedback, but we do not accept re-submission and the grade will not be updated.**