CSCI 452 (Data Mining)
Data Exploration in R
75 pts

Overview

For this assignment, you'll explore a dataset and document your exploration process in an R Markdown file. When turning in your assignment, provide a .zip file containing your .rmd (R Markdown) file and the .html file you created when you performed the *knit* command in RStudio. Name your files *exploration.rmd* and *exploration.html*.

Here is a guide for creating an R Markdown document and knitting it into an .html document. Please check out some of the resources I have listed on our course website – there are lots of tutorials, blogs, etc out there to help you with questions that may arise.

For this assignment, you will pick one of the listed datasets (from FiveThirtyEight.com) and perform initial, exploratory data analysis. Here's a good example of the type of exploration and documentation that I'm looking for: (Vancouver Crime: https://rpubs.com/jtdoud/VanCrime)

Data sets (pick one):

- Hate Crimes (https://github.com/fivethirtyeight/data/tree/master/hate-crimes)
- Drug Use by Age (https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age)
- Airline Safety (https://github.com/fivethirtyeight/data/tree/master/airline-safety)
- College Majors (https://github.com/fivethirtyeight/data/tree/master/college-majors -- you can concentrate on recent-grads file but include others if you want to)

When loading your selected dataset, use a url that goes directly to the raw .csv. For example, if you're using College Majors, go to the link above, then click on the recent-grads.csv link, then click "Raw". Copy that url and use it in your R code, as in:

```
mydata =
read.csv("https://raw.githubusercontent.com/fivethirtyeight/data
/master/college-majors/recent-grads.csv")
```

Your report should include four parts:

1. (5 pts) Introduction – general info about the dataset and a code chunk showing any additional packages that are required to replicate results
2. (20 pts) Summary statistics
3. (25 pts) Basic visualizations – choose visualizations that are interesting to you and help you to better understand the data
4. (25 pts) Conclusion/Next Steps – this section should include information such as:
   a. Would anything have to be done to the data before starting to analyze it?
   b. What questions do you think it would be interesting to explore using the data?
   c. Is there extra information you'd like to (or would need to) have?

Notes:

- Be sure to display your R code in your report using code chunks (the gray boxes in the example). This should happen automatically as long as you don’t use `echo = FALSE`.
- Use the URL for the raw file in github so that I don’t have path/access issues
- If you decide to use libraries that are not included in the base installation of R, be sure to include a code chunk with those libraries (see the example linked to above). All packages must be available through CRAN installation in RStudio.

Create a .zip file containing your exploration.rmd file and your exploration.html file. Directly zip the files, do NOT zip a directory containing the files. Submit your .zip file as the Data Exploration lab to Autolab.