**DS 2003: COMMUNICATION WITH DATA**

Midterm Instructions

**Important Dates:**

*Group Assignments:* by Tuesday, September 20th (in class)

Data Source Selection: Tuesday, September 27th (in class)

*Due Date:*Wednesday, October 5th @ 11:59pm

*In-Class Presentations:* Thursday, October 6th & Tuesday October 11th

**Learning Objectives:**

1. Demonstrate the ability to find a reputable data source.
2. Define questions that you would like to answer with the data.
3. Summarize the pertinent data using descriptive statistics.
4. Create graphs which accurately represent the data.
5. Describe your results using an R Markdown report.
6. Present to the class.

**Teams:** This project will be done in teams of 2-3

**Evaluation:**

* The entire project is worth 200 points (20% of your total grade)
* 150 points will come from your analysis/write-up
* 25 points will come from your presentation to the class
* 25 points will come from evaluation of effort and inclusivity from your teammates

**Instructions:**

1. Demonstrate the ability to find a reputable data source.

You will use [FiveThirtyEight](https://data.fivethirtyeight.com/) or [Kaggle](https://www.kaggle.com/datasets?fileType=csv) to discover a data source. It cannot be a dataset we have previously used in class. Choose something your group is interested in – don’t be afraid to pick unpopular topics.

Note 1: Database selection will be due on Tuesday 9/27 in class

Note 2: DO NOT plagiarize graphs from Kaggle, Github, or other open sources.

1. Define questions that you would like to answer with the data.

Come up with two questions that are feasibly answered within the dataset using visualization.

Example: How are x and y related? What are trends in x over a timeframe? What is the breakdown of x over different subclasses?

1. Summarize the pertinent data using descriptive statistics.

Your analysis should include a robust summary of the pertinent variables from your dataset. Use well formatted tables and data viz to show your summary descriptives

1. Create graphs which accurately represent the data.

You will need to create 2-3 highly quality graphs that accurately depict your data. Graphs should use ggplot2 and follow the guidelines set by Professor Cairo in “How Charts Lie” (pages 47 -50).

1. Describe your results using an R Markdown report.

Your questions, summary, graphs, and write-up will be housed in an R markdown and submitted in PDF file via Collab. This should look like a professional report.

[R Markdown: The Definitive Guide](https://bookdown.org/yihui/rmarkdown/markdown-syntax.html)

[R Markdown Help](https://rmarkdown.rstudio.com/lesson-1.html)

[Examples](https://rmarkdown.rstudio.com/gallery.html)

[Cheat Sheet](https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf)

Template

1. Present to the class.

Your group will do a **five slide** presentation. The first slide will consist of a short summary (2-3 sentences) of your dataset. The next 4 slides will consist of a question and graph answering that question.

Presentation will be 10 mins: 5-7 mins for presenting, the remaining time for questions. Presenting time CANNOT go over 7 mins. I suggest practicing.

[Template](https://myuva-my.sharepoint.com/:p:/g/personal/nak5dy_virginia_edu/EQIZuO1RCvtCobxRVDAY6OkBfg0tnhkqRtgaGMgoVC5vMg?e=QgHXuG)