Department of Statistics: Analytics Workshops

Day: Wednesdays every other week (9/27, 10/11, 10/25, 11/8, 11/29)
Time: 7-9 p.m. (tentative)
Location: 2006 Sheridan Road, B02 (tentative)
Contact: Bailey Pleva
Email: baileypleva2018@u.northwestern.edu

Overview: The purpose of these workshops is to engage undergraduate students (as well as others who may be interested) in applications of statistics through analytics in a collaborative environment.

Format of Workshops: For each workshop, you will be presented with a dataset and an assigned task. Please bring your laptop fully charged, as it isn't guaranteed you will have access to a power outlet. You will be spending most of the time working on the assigned task with others (Bring friends!). At the end of each workshop, you will be asked to share you work to GitHub.

Prerequisite Knowledge: These workshops are open to all skill levels. However, because work will involve programming in either Python or R environments, it is important to have at least a basic knowledge of programming. This means at least being familiar with variable assignment, data types, writing simple loops, and writing simple functions. Also, it is important to have at least a basic understanding of statistical concepts and visualization. Those who are unfamiliar with either language or are unfamiliar with the integration of statistics into programming are still encouraged to come, but to prepare beforehand by looking through resources in https://www.kaggle.com/wiki/Tutorials. Northwestern's library also has plenty of helpful books and articles for all levels.

GitHub: Because the goal is to make these workshops collaborative environments, it is really important to have a GitHub account. We want to be able to share files with each other to learn the different ways we can work through a dataset/problem. If you do not yet have an account, there is a tutorial to familiarize with how to use Github and why you would want to at https://guides.github.com/activities/hello-world/.

Code and Readability: Because you are producing code for others to see, you should make sure that it is readable. This means adding plenty of comments to your code to explain what you are doing and following style guidelines. Below are Google's style guides for R and for Python:

- R: https://google.github.io/styleguide/Rguide.xml
- Python: https://google.github.io/styleguide/pyguide.html

Useful Tools: Below are a list of packages/libraries, tools, and guides you might find useful for work in and outside of the workshops. There are a lot of packages/libraries and guides not mentioned which also might be useful. Feel free to share them.

- IDEs
  - R: RStudio, Red-R, Visual Studio
  - Python: IPython/Jupyter Notebook, Spyder, Atom, Rodeo, nteract
- Visualization
  - R: Base Graphics, ggplot2, lattice, plotly, shiny, rmarkdown
  - Python: matplotlib, seaborn, ggplot, Bokeh, plotly
- Processing
  - R: dplyr, tidyr, magrittr, data.table, reshape2
  - Python: numpy, pandas
- Modeling
  - R: car, randomforest, caret, bigrf, cba, rankcluster
  - Python: scikit-learn, mlpy, statsmodels, scipy
- Guides
  - "R for Data Science" – Grolemund, Garrett; Wickham, Hadley
    - http://r4ds.had.co.nz/index.html
    - Complete (free) book on how to do data science with R. Even if you are familiar with data science, there are helpful chapters on specific topics.
  - swirl
    - http://swirlstats.com/
    - swirl is actually a package in R that teaches you how to program and fundamentals of data science within your console. If you want to learn R, or even brush up on your skills, try it out.
  - "The Ultimate Python Seaborn Tutorial: Gotta Catch 'Em All"
    - https://elitedatascience.com/python-seaborn-tutorial
    - A lot of people probably already know matplotlib. Seaborn is an interface for matplotlib that arguably makes your visuals 10x better.
  - "An introduction to data cleaning with R" – de Jonge, Edwin; van der Loo, Mark
    - https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
    - Data cleaning and preparation is essential for analysis. This is a short guide on how to do both in R. R Bloggers, Analytics Vidhya, and KDNuggets are all great, and there are plenty more.

- **Feedback is Appreciated**: This is something new for the Department of Statistics, and it is something we would like to continue for the future. If you have any sort of recommendations for improvement, feedback on what you enjoy and don't enjoy, ideas for skills to work on, datasets you would like to provide, etc., please feel free to email me, baileypleva2018@u.northwestern.edu. The main thing that will determine the value these workshops provide is the level of participation. Come ready to think, talk, explore, and, most importantly, have fun!