Final Project Data Science with R and Python

PART A: Analysis

1. You will be taking an example data set from R or Python that is built in and you will use it to build some predictions (i.e., statistical learning). In other words, you will be doing some basic machine learning (or more advanced if you want to take on that challenge). You can decide which language to use. I highly encourage that people use R since that is the language we have been working in.
2. Here is a link to the example R Datasets that exist: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html. Here is a link to example Python datasets that exist
3. You can take any one of these datasets that interests you and you will do the following:

   You will conduct regression analysis (a form of supervised machine learning) and begin to derive your final conclusions. This will become the main portion of your final write up. The regression analysis should include the following:

- Statement of the data question and modeling objective (in this case, it should be prediction)
- Description of the data and response variable
- Exploratory data analysis - univariate, bivariate, multivariate
  - The EDA should include visualizations (with informative titles and axes labels) and summary statistics as you describe the main results from the visualizations.
- Brief explanation of the modeling approach and why you chose the model type. This should include the model type, model selection procedure, and interaction terms (if you decide to use them, you do not have to since we will not have talked much about them by the end of the course but we will cover this before the end).
- Output of the final model. *Note: The final model will be the result of numerous iterations and trying different models. You can include the other models you consider in the "Additional work" section at the end of the analysis write up.*
- Discussion of the assumptions for the final model
- Interpretations and interesting findings from the model coefficients
- Additional work of other models or analysis not included in the final model.

The regression analysis should be written as a report. Write your analysis in a .Rmd file called “analysis.Rmd” or a .ipynb called “analysis.ipynb” in Canvas. The logic used in R and Python is similar even if they have different syntax. If you feel more comfortable doing this in one language over another—by all means do that.

PART B: Presentation

The presentation will consist of two components: (1) Slide deck and (2) Video presentation OR write up.

Your slides should have:

- Title Slide
- Slide 1: Introduce the topic and motivation
- Slide 2: Introduce the data
- Slide 3: Highlights from Exploratory Data Analysis
- Slide 4: Final model
- Slide 5: Interesting findings from the model
- Slide 6: Conclusions + future work

You can use the software of your choice to create your slide deck. Save your slide deck as PDF or provide a link to view your slides online (e.g. in Google Slides)

Good luck and as always—feel free to ask Amber, Grace, or me (as final resort) for questions!

All my best,
Jose Aveldanes