Discussion Section #8
Due: To be submitted to CatCourses by 11:59pm.

Instructions:

As discussed on the Syllabus, Statistics is not a spectator sport! That's why a component of this course will be comfort in working with and analyzing real data. Lab this week will be playing with data frame in R.

Please follow along with your instructor as you go through an example with a simple data set on temperature in different cities.

During lab you will be encouraged to consider the following **three** different data sets:

1. College.csv: A data set collected with different colleges
2. US_County_Level_Presidential_Results_2008-2012-2016.csv: A set of Presidential Results by County in the US.
3. movies.csv: A data set scraped from IMDB of 6820 movies between 1986 and 2016 (220 movies chosen per year).

Each data sets is described in greater detail below.

Assignment:

For this assignment, you must answer the following for either the three data sets above, *or if you are so, inclined, you can consider only two of the instructor provided data sets and include one data set of your choosing for one of the three data sets above.*

We provide a sample code to help you get started. In this sample code, a sample plot is given for each data set. You can not use the sample plot as one of your plots in the work you submit.

1. (1 Point): Provide your R Code for all analyses
2. For each of the three data sets (3 Points per data set):
   - (1 Points = 0.5 Point Each Plot) Load data and make two different plots. (These can be histograms, box plots, scatter plots etc).
   - (2 Points = 1 Point Each Paragraph) For each plot, you must provide a short description of what you are seeing. This description must answer the following questions: (1) What are you actually plotting (both data and style of plot)? (2) Why did you make this plot? (3) What does this plot show us?

Note: If you do choose to analyze your own data set, you must also say where the data set came from and either include the dataset with your own submission or (if the file is larger than 1Gb) provide a link to where instructors/TAs can download it.

Instructor Provided Data Sets:

1. **US Colleges**
   (Filename: `College.csv`)

   This data set is prepared and curated as part of the popular textbook “Introduction to Statistical Learning” by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. This data set contains a variety of demographic information on a number of Colleges in the US. Originally, the data was collected from the 1995 US News and World Report and was used in the American Statistical Association Statistical Graphics Section’s 1995 Data Analysis Exposition.

   This data set contains 777 rows and 19 columns:

   - Private (Text): Public/private indicator
   - Apps (Integer): Number of applications received
   - Accept (Integer): Number of applicants accepted
   - Enroll (Integer): Number of new students enrolled
   - Top10perc (Integer): New students from top 10 % of high school class
   - Top25perc (Integer): New students from top 25 % of high school class
   - F.Undergrad (Integer): Number of full-time undergraduates
   - P.Undergrad (Integer): Number of part-time undergraduates
   - Outstate (Integer): Out-of-state tuition
   - Room.Board (Integer): Room and board costs
   - Books (Integer): Estimated book costs
   - Personal (Integer): Estimated personal spending
   - PhD (Integer): Percent of faculty with Ph.D.’s
   - Terminal (Integer): Percent of faculty with terminal degree
   - S.F.Ratio (Integer): Student/faculty ratio
   - perc.alumni (Integer): Percent of alumni who donate
   - Expend (Integer): Instructional expenditure per student
   - Grad.Rate (Integer): Graduation rate

2. **US Presidential Election Data by County for 2008, 2012, 2016**
   (Filename: `US_County_Level_Presidential_Results_2008-2012-2016.csv`)

   Every 4 years, the United States has a Presidential Election. Some enterprising folks on github/Twitter have pulled election results by state and county for the past 3 Presidential Elections. For each county and each Presidential Election year, we know: the total number of voters, the total number of voters for the Democrats, the number of voters for the Republicans and the number of voters for other candidates.

   This data set consists of 3,114 rows and 15 columns:

   - State Abbreviation (Text): The two letter postal code for the state.
   - County Name (Text): The name of the county

- Fips (Integer): A unique number designating the county state combination
- Total Number of Voters in 2008 (Integer)
- Number of Voters for Democrats in 2008 (Integer)
- Number of Voters for Republicans in 2008 (Integer)
- Number of Voters for Other Candidates in 2008 (Integer)
- Total Number of Voters in 2012 (Integer)
- Number of Voters for Democrats in 2012 (Integer)
- Number of Voters for Republicans in 2012 (Integer)
- Number of Voters for Other Candidates in 2012 (Integer)
- Total Number of Voters in 2016 (Integer)
- Number of Voters for Democrats in 2016 (Integer)
- Number of Voters for Republicans in 2016 (Integer)
- Number of Voters for Other Candidates in 2016 (Integer)

Study Link: https://github.com/tonmcg/US_County_Level_Election_Results_08-16

Note: The course instructor did some curation of the data set from the github. First, a number inconsistencies in the 2008 election results were corrected by referencing the true values reported for each county on Wikipedia. Second, the three election results were consolidated into a single csv file. Finally, because Alaska formally does not have counties, all results for Alaska were combined into a single row.

3. **Movies**

(Filename: `movies.csv`)

This data set was taken from a managed curated data set on Github of IMDB scraped data (https://github.com/danielgrijalva/movie-stats). The following description is taken from the Github page.

The data set consists of 6820 rows and 15 columns. Each movie has the following attributes:

- budget: the budget of a movie. (Integer, but note: some movies don't have this, so it appears as 0)
- company: the production company (Text)
- country: country of origin (Text)
- director: the director (Text)
- genre: main genre of the movie. (Text, 17 different options)
- gross: revenue of the movie (Integer)
- name: name of the movie (Text)
- rating: rating of the movie (R, PG, etc.) (Text, 13 different categories)
- released: release date (YYYY-MM-DD) (Text)
- runtime: duration (in minutes) of the movie (Integer)
- score: IMDb user rating (Float/Decimal)

- votes: number of user votes (Integer)
- star: main actor/actress (Text)
- writer: writer of the movie (Text)
- year: year of release (Integer)

Other Data Set Repositories:

As mentioned above, you may substitute **one** of the instructor provided data sets for a data set of your own choosing. If this is something you find interesting, then here's some data repositories you might find a good place to get started. (Note, your Video Project #2 will include the option of analyzing a data set of your own choosing. So here's an opportunity to start thinking more about it!)

- UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php
- Data.gov: https://www.data.gov/
- Datasets subReddit: https://www.reddit.com/r/datasets/
- Baseball: https://www.baseball-reference.com/
- American demographics from the Bureau of Labor Statistics https://www.bls.gov/
- US Census tutorials: https://www.census.gov/data/academy/courses.html
- Kaggle: https://www.kaggle.com/datasets
- Health Data: https://healthdata.gov/