Homework One

MSDS 598 Spring 2022

Directions

- Submit a .ipynb notebook to Canvas.
- The Notebook should begin with a Markdown cell with your **Name** and the title of the Assignment, **Homework 1**. Failure to do so will result in points lost.
- Use Markdown Cells to **clearly** indicate which code answers which question and to answer short answer questions. Failure to do so will result in points lost.
- The filename for your notebook should be formatted like

  $$FirstName\_LastName\_AssignmentName.ipynb.$$

  Failure to do so will result in points lost.
- This is due on February 7th at Midnight Pacific time. Solutions will be posted one week after the due date.

For many of the following questions you will need to use various datasets. These can be found on Canvas under Week Two.

1. Consider an experiment where you randomly pick 10 people and write down their birthdays. Let X be the random variable that is 0 when nobody shares a birthday and is 1 when at least two people share a birthday in your sample.

(a) Describe the sample space of this experiment.

(b) Simulate this experiment and use this simulation to estimate the probability distribution of random variable X. Use at least 10,000 trials.

(c) What happens if I change my experiment to pick more people? Visualize this change using a few estimated probability distributions.

2. Consider the olympic athlete data from the January 31st lecture.

(a) Choose a country and set up a Hypothesis Test to decide whether a random sample of 100 athletes from that country have a mean *weight* different from the population mean.

(b) Calculate the p-value to determine whether to reject or fail to reject the null hypothesis with a significance level of $\alpha = 0.01$.

3. Import the Advertising dataset found on Canvas looking at dollars spent by businesses on advertising in various mediums along with the associated sales that quarter.

(a) What variables in this dataset appear linearly correlated? Pick a pair that look linearly correlated and calculate the R^2 value for a line of best fit using OLS.

(b) Explain in your own words what the relationship between these two variables using the line of best fit.

4. Consider the TV and sales variables.

(a) Are they linearly correlated? Are they correlated in some other way?

(b) How might we create (or *engineer*) a feature from the TV variable that was linearly correlated with sales?

5. Import the Penguins dataset we looked at in class on January 24th.

(a) Which other numerical variable predict bill length the best with a linear model? What is the R^2 for each of these?

(b) What if we create a separate linear model for each *species* to predict bill length? Does the overall effectiveness of our model improve? Use R^2 values to make your decision.

(c) Why might models get better if we create a separate linear model for each category? Can a model get "worse" if we do this?