Week 2 Exercises: Statistics for Data Science

Part 1: 2017 Appraisal Values

The file appraisal_2017.csv contains the 2017 appraised value (total_appr) and square footage (finished_area) for a random sample of 1000 houses in Davidson County.

Read this data into a dataframe named *appraisal*.

1. Create a scatterplot of total_appr vs finished_area.
2. By inspecting the scatterplot, describe the relationship between total_appr and finished_area. Is the direction of association positive or negative? Is the relationship linear? How strong is the relationship?
3. Do you see any points which might be considered outliers? Investigate those points.
4. Find the correlation between total_appr and finished_area. How strong is the relationship between the two variables?

Part 2: Penguins

The file pengiuns.csv contains the Palmer Penguins dataset, which contains size measurements for three penguin species observed on three islands in the Palmer Archipelago of Antarctica.

Read this dataset into a dataframe named *penguins*.

1. How many missing values are in this dataset? After checking this, use the dropna() method to remove any missing values.
2. Examine the distribution of penguin species by island. What do you find? (You might want to create a plot or two to aid in your analysis)
3. What do you find if you examine the distribution of weights by species?
4. Create a scatterplot of bill depth vs. bill length. What do you notice from the scatterplot? What is the correlation between bill depth and bill length?
5. Color the scatterplot from the previous question by species. What do you notice now?
6. Are there major differences in the distribution of species observed across the three years that the data was collected?
7. Inspect the relationship between body mass and flipper length. How is the relationship between these variables different than the one between bill length and bill depth that we observed above?
8. How does the distribution of body mass differ between male and female observations?

9. You can group by multiple variables simultaneously by passing a list into your groupby method. What do the differences in body mass between male and female penguins look like at a species level?