Notes:

- **3 Rules of Data Analysis:**
  1. Make a picture
  2. Make a picture
  3. Make a picture

- **Making Piles:**
  - Pile together things that seem to go together to see how they distribute across different categories.
    - You can **count** things in a category and make a **frequency table** which records the totals and the category names.
    - **Proportion** – fraction of the data in each category. Divide the counts by the total number of cases.
    - **Percentage** – multiply the proportion by 100
    - **Relative Frequency Table** – displays the percentages rather than the counts
      - both types of tables show the **distribution** of a categorical variable since they tell how frequently each occurs.

- **Making a Picture:**
  - Tests show that our eyes are more impressed by the **area** on a graph rather than the **length**.
  - **Area principle**: the area occupied by a part of the graph should correspond to the magnitude of the value it represents.
  - **Bar chart**: displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison
  - **Relative Frequency bar chart**: replace the counts with percentages
  - **Pie chart**: shows the whole group of cases as a circle.
    - slice the circle into pieces whose size is proportional to the fraction of the whole in each category
  - Before you make a bar chart or a pie chart, always check the Categorical Data Condition: The data are counts or percentages of individuals in categories.

- Doing Statistics Right, means you have to THINK.
- To Look at 2 categorical variables together, arrange the counts in a two-way table.
- **Contingency table:** table that shows how the individuals are distributed along each variable, contingent on the value of the other variable.
  - **Marginal distribution:** when the frequency distribution of one of the variables is in the margins of a contingency table.
  - When the distribution of one variable is the same for all categories or another, we say they are **independent** (no relationship between them).
- **Cell:** each block of a table that gives the count for a combination of values of the two variables.
- Always be sure to ask PERCENT OF WHAT?
- **Conditional distribution:** shows the distribution of one variable for just the individuals who satisfy some condition on another variable.
- **Segmented Bar Chart:** treats each bar as a “whole” and divides it proportionally (page 29).

- **Don’t violate the area principle.** Even it is prettier.
- **Keep it honest.**
- **Don’t confuse similar-sounding percentages:**
  - Percent of passengers who were both first class and survived
  - Percent of first class passengers who survived
  - Percent of survivors who were in first class
- **Don’t forget to look at the variables separately too.**
- **Be sure to use enough individuals.**
- **Don’t overstate your case.**
  - It is rare for two variables to be entirely independent
- **Simpson’s Paradox:** be careful when you average across different levels of a second variable. It’s always better to compare percentages or other averages within each level of the other variable. The overall average may be misleading.