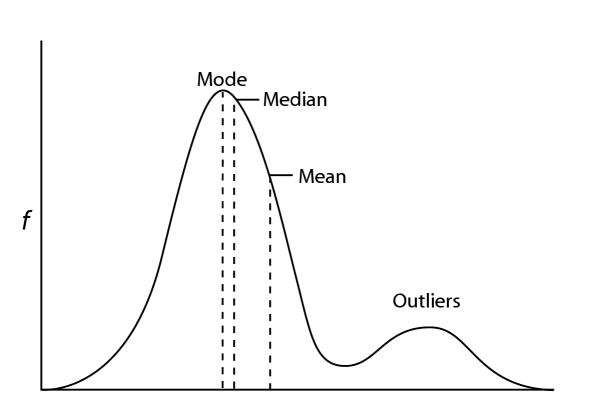
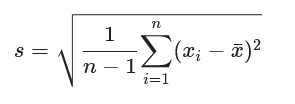
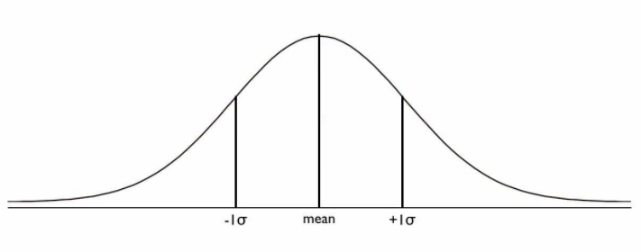
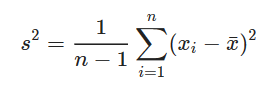
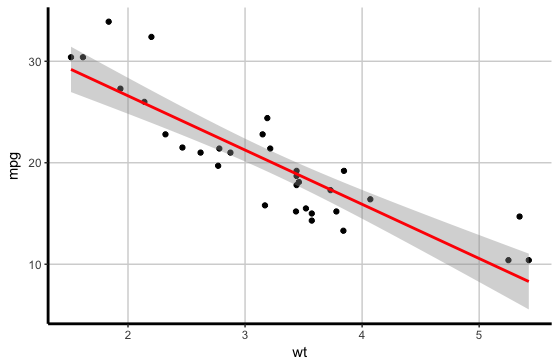
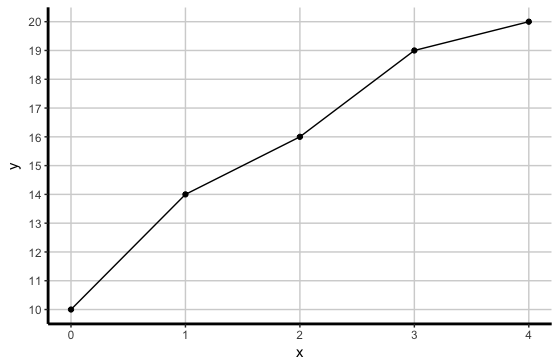
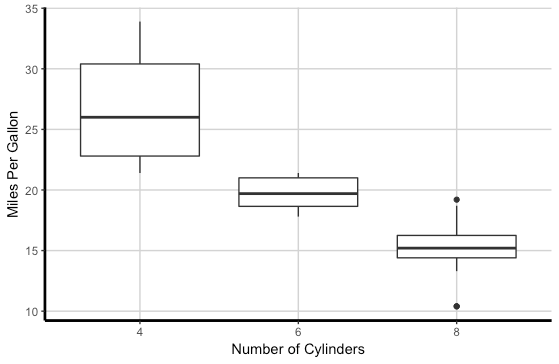
Unit 1-2: Measures of Central Tendency and Dispersion (Descriptive Statistics)

* Descriptive Statistics
  + Data sets are powerful. They can predict the future (will my product be a hit?) and reveal secrets about the past (just how big was the iceberg that sank the Titanic?). But sometimes they're not much to look at.
  + This is where descriptive statistics come in!
  + This branch of statistics provides us with ways of understanding and summarizing data — and presenting visually appealing results. Let's learn how, shall we?
* Descriptive vs. Inferential Statistics
  + A quick consideration to note before we proceed:
    - **Descriptive statistics** focus on summarizing, describing, and understanding the data we observe.
    - **Inferential statistics** focus on generalizing the results from a sample to a larger population.
  + In this lesson we'll concentrate on descriptive statistics.
* Distributions
  + The distribution of a variable describes all possible values of the variable, as well as how likely that variable is to take on each of those values.
  + There are often two main questions we are interested in regarding distributions:
    - What is the center of the distribution?
    - What is the spread around the center of the distribution?
* Measures of Central Tendency
  + Here's where we'll remember three old pals from high school: mean, median, and mode. They are all ways we can describe the center of a distribution.
  + The **mean** of a distribution is the average value. To calculate the mean, sum each observation and divide by the total number of observations.
* Measures of Central Tendency: Practice
  + For a list of numbers [4,6,5,8,6,6,4] what is the **mean**? Give it a try.
    - The mean is (4+6+5+8+6+6+4)/7=39/7The graph below shows the mean, median, and mode plotted out.
    - 
    - Note that we have a few outliers, which make our graph skew to the right. Therefore, the mean is slightly to the right of our highest point (mode).
  + The **median** of a distribution is the middle value. To calculate the median, sort the values from smallest to largest and then identify the center.
    - If there are an odd number of observations, simply select the middle value. If there are an even number of observations, average the two middle values.
  + The **mode** of a distribution is the most frequent value. If one value in the data set occurs more frequently than all other values, it's the mode.
    - For the list of numbers [4,6,5,8,6,6,4] the mode is 6.
    - For the list of numbers [4,6,5,8,6,4,5] note that the values 4, 5, and 6 all occur exactly twice! In this case, these values are all the mode
* Measures of Spread
  + The **range** of a distribution is the maximum value minus the minimum value. This is the simplest way to measure the spread of the distribution.
    - For the data set [4,6,5,8,6,6,4] the maximum value is 8 and the minimum value is 4. Therefore, the range is 8−4=4
  + The **standard deviation** of a distribution is the average approximate distance from each observation to the mean.
    - The formula for standard deviation of a sample of data is:



* + - Consider the data set [3, 4, 5]. To find the standard deviation:
      * 1) Calculate the mean.
        + 4
      * 2) Take the difference of each observation and the mean.
        + The differences are -1, 0, and 1)
      * 3) Square each result from Step 2.
        + The squared results are 1, 0, and 1
      * 4) Sum each result from Step 3.
        + The sum is 2
      * 5) Divide the result of the previous step by (n-1), where n is the number of observations.
        + 2 / (3-1) = 2/2 = 1
      * 6) Take the square root of this result.
        + The square root of 1 is 1
        + The standard deviation of the data set [3,4,5] is 1
      * Using the above example, our mean value would be 4 and our lowercase sigma represents one standard deviation to the right or left of the mean
        + 
  + Luckily, we've already found the **variance** — we just haven't explicitly said it yet. In our previous list of six steps, try stopping after Step 5.
    - The formula for variance for a sample is:
    - 
    - The image below represents a linear model using the MTCars data set. To record the variance, imagine a vertical line between each point on the graph to the predicted value (red line).
    - Now, take that difference from every point on the graph, square those values, sum them together, and then divide the result by (n-1).
    - 
* Percentiles
  + Percentiles are an effective means of summarizing a distribution. You've likely heard of percentiles before — for example, a parent might say, "My child is in the 85th percentile for height."
  + If we take a set of data sorted from smallest to largest and divide it into 100 equally sized groups, the 99 values that divide the data into these equally sized groups are called percentiles.
  + For example, the first percentile is the value that splits the bottom 1 percent from the top 99 percent. The second percentile is the value that splits the bottom 2 percent from the top 98 percent. And so on.
  + In day-to-day usage, there are some percentiles that are particularly important:
    - 50th percentile: Divides the bottom 50 percent from the top 50 percent; also called the median.
    - 25th percentile: Divides the bottom 25 percent from the top 75 percent; also called the first quartile.
    - 75th percentile: Divides the bottom 75 percent from the top 25 percent; also called the third quartile.
  + In the data set [10,14,16,19,20] the median is 16, the 25th percentile is 12, and the 75th percentile is 19.5.
    - Note: When calculating the 25th and 75th percentiles, we exclude the median from the calculation.
      * 
  + The following is a process for calculating the values of the 25th and 75th percentiles.
    - Remember that the 50th percentile is equivalent to the median!
    - 1) Order the values in numerical order: [10, 14, 16, 19, 20].
    - 2) Split the values into two groups: those less than the median and those greater than the median ([10, 14] and [19, 20]).
    - 3) To calculate the 25th percentile, we find the midpoint of the first group. Sum the values of the first group, [10, 14], divide by the number of values in that group. In this case: (10 + 14) / 2 = 24 / 2 = 12.
    - 4) To calculate the 75th percentile, we find the midpoint of the second group. Sum the values of the second group, [19, 20], divide by the number of values in that group. In this case: (19 + 20) / 2 = 39/ 2 = 19.5.
    - 5) The 50th percentile is equal to the median value — in this case, 16.
  + Recall that we often summarize the center and spread of a distribution. We can use percentiles for both!
  + As we've discussed previously, we use the median (50th percentile) as one measure of center. We can use the interquartile range (which is calculated as the 75th percentile minus the 25th percentile) as our measure of spread.
  + Interquartile range can be solved using the following:
    - Q1 is the "middle" value in the first half of the rank-ordered data set.
    - Q2 is the median value in the set.
    - Q3 is the "middle" value in the second half of the rank-ordered data set.
  + If we have the sorted list [1, 3, 5, 7, 9], the interquartile range (IQR) can be solved like so:
    - Q1 = 2
    - Q2 = 5
    - Q3 = 8
    - IQR = (Q3 - Q1) = (8 - 2) = 6
  + As you can imagine, we can calculate lots of different measures of center and spread using percentiles; these are just some of the most common.
* Skewness
  + When discussing the spread of a distribution, we might be interested in more than just how far values extend beyond the center. We might want to know how the distribution itself is shaped. Does the distribution seem symmetrical around the center or is it skewed?
  + If approximately 50 percent of the distribution is less than the mean and approximately 50 percent of the distribution is greater than the mean, we call that distribution symmetric. Otherwise, we call the distribution skewed.
  + By definition, one guideline is to see if the mean and median are close to one another.
    - If the mean is approximately equal to the median, the distribution is usually symmetric.
    - If the mean is significantly different from the median, we say that the distribution is skewed.
  + If the distribution is skewed, it can either be left skewed or right skewed. We can identify whether it is left or right skewed based on the direction of its "tail."
    - If the tail appears on the left side, we say the distribution is left skewed. (This occurs when the mean is less than the median.)
    - If the tail appears on the right side, we say the distribution is right skewed. (This occurs when the mean is greater than the median.)
* Skew and Measures of Center
  + Which value best represents the middle of a distribution? We can answer that question using "measures of center."
  + For example, a skewed distribution suggests the presence of outliers. Outliers are unusually large or small values.
  + Outliers can significantly affect the mean of a distribution. As a result, for a skewed distributions, we might want to use the median instead of the mean to summarize a "typical" value.
  + Conversely, a symmetric distribution suggests the minimal influence of outliers. For symmetric distributions, the mean is typically used as the measure of center.
  + By using a box plot, we can see many of our descriptive statistics at once, including outliers. Let's look at an example calculating the average miles per gallon of an eight-cylinder car.
    - In the chart below, the black dots beyond the box for eight cylinders represent outliers in our data.
    - In other words, an eight-cylinder car will average about 15 miles per gallon. However, there was one instance when an eight-cylinder car only got 10 miles per gallon. Conversely, one eight-cylinder car was beyond the average and got about 19 miles per gallon.
    - 
* Dealing With Outliers
  + As an analyst or data scientist, it's up to you to determine what to do with outliers. In some cases, you may want to keep outliers, and in other cases you may want to remove them.
  + For example, if we want to know the average net worth of residents in Washington State, our data will be heavily skewed because of billionaire Bill Gates. In a case like this, you may want to remove all values outside of x standard deviations from the mean, with x being a value on which you'll need to make a decision.
  + As another example, imagine we work for a company that tracks the internal temperature of engines from trains and airplanes.
    - It may be important for us to consider outliers because they can have severe implications on business decisions and potentially people's safety!
    - In this case, you could build logic into a model that says, if the temperature is greater than a certain degree, it should send an alert.
* Real-World Recommendation
  + While there are certain conditions in which the mean or median might be a better representation of the "center" of a distribution, as a practical note, it's always a good idea to report both to provide a more complete picture.