Name: **Date:** **Class:**

Pre-Activity Reading

Introduction

Just as important as how you collect your data is **how you analyze your data**. This activity takes you through some important data analysis concepts, which you practice in class using a real-world air quality data set and data analysis tool (Excel). Use the reading below to complete the *Pre-Activity Worksheet*.

Section 1: Statistics Review: Summarizing Data

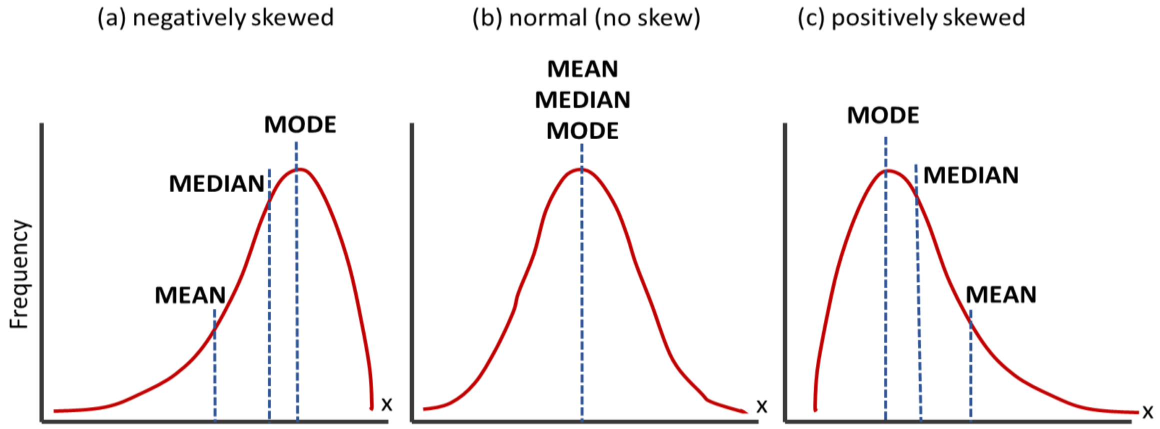
Basic Statistics

When you work with a data set, it is important to begin by considering the characteristics of the data set. Important characteristics include: "a measure of the center of the data, a measure of spread or variability, a measure of the symmetry of the data distribution, and perhaps estimates of extremes such as some large or small percentile" (Helsel and Hirsch 2002). These statistical parameters enable us to *characterize* and *quantitatively compare* different data sets. You are probably already familiar with some basic statistical parameters such as **mean**, **median**, **mode**, **maximum/minimum** and **range**. Refer to the *Excel Reference Sheet* for more in-depth definitions.

Sample Size, Data Distribution and Variance/Spread

Sample size, typically represented by the variable *n*, is the number of samples used to represent a **population**. For example, if you were administering a survey in your high school, you would want to survey enough students that your results were representative of the opinions of the entire school. In this example, the people you survey would be the *sample size*, and the student body would be the *population*. Sample sizes are used in science when it is too laborious or expensive to collect data from the entire population. When we collect air quality data, it is important to consider how representative our data is of the space around it. In a very simple environment (for example, a rural area with few point sources of pollution), our data might be representative of a larger area, however, in a complex environment like a city, that area would be smaller. In this air quality example, the sample size would be the number of monitoring locations in the given study area.

The diagram below illustrates the concept of **distribution.** A **histogram** is a type of graph used to plot data to assess the statistical distribution of a data set by plotting the recorded values on the x-axis vs. the frequency with which they occur on the y-axis. When the majority of measurements fall within the middle of the range, we say the data is normally distributed (b). If the majority of the data falls within either the high or low end of the range, we say the data distribution is skewed (a and c). Notice that in a normally distributed data set, the mean, median and mode are all the same. In a skewed data set distribution, the mode, median and mean fall in different places; they are not the same.



Brought to you by Engineering UNIVERSITY OF COLORADO BOULDER

Name: Date: Class:

Variability, or the range of a data set, is a measure used to compare one data set to another data set. Variability can be characterized by **standard deviation** and **interquartile range** (again, refer to the *Excel Reference Sheet* for more in-depth definitions). Data that has a larger variability, or range, has a higher standard deviation and a larger interquartile range than a data set with less variability.

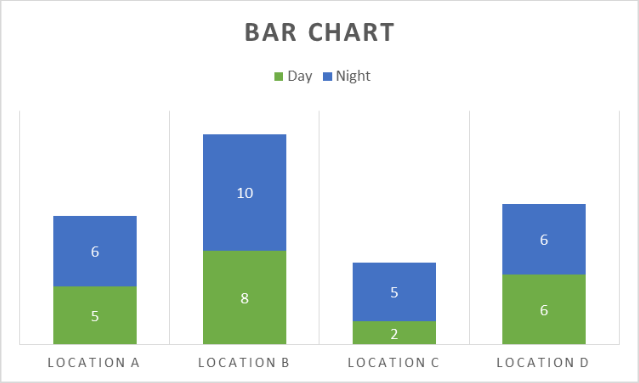
Uncertainty and Outliers

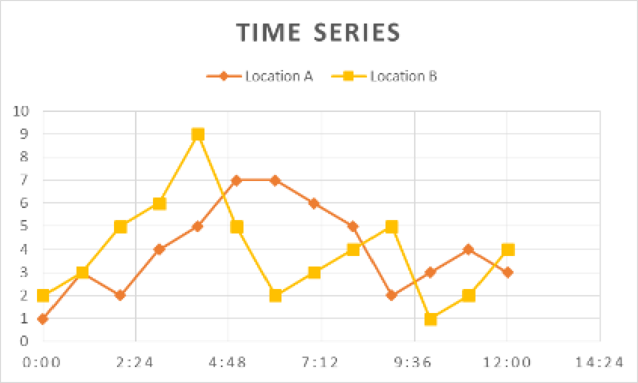
Uncertainty is something you will not have to worry about too much for this project, but keep in mind error is always associated with scientific measurements. For example, typical error for a low-cost CO_2 sensor is +/- 10 ppm, which means a measurement of 400 ppm may actually be between 390-410 ppm.

Spread can be highly affected by **outliers**. An outlier is a data point that is far from the majority of other data points in the data set. Outliers can be created by uncertainty in measurement equipment or produced by a problem with data collection or processing. If you choose to remove data points because they are outliers, you must be able to explain why. Sometimes outliers are mathematically defined as points that are more than two standard deviations away from the mean.

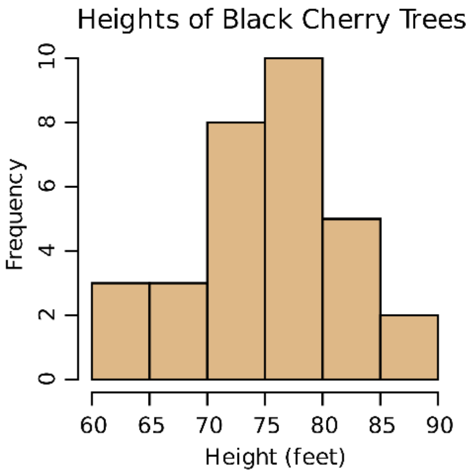
Section 2: Visualizing Data by Graphing

Visualizing data in plots can help you see patterns you were not aware of, or see your data in a new light.





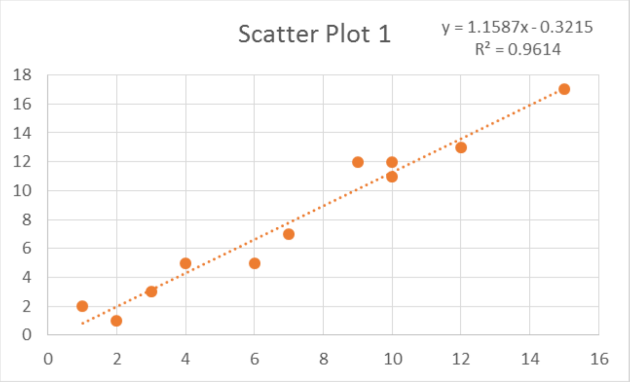
Bar charts are ideal for summarizing data or comparing data sets, for example, comparing data averages from different locations or recorded during different activities. **Time series** are useful for analyzing trends over time for one or multiple data streams.

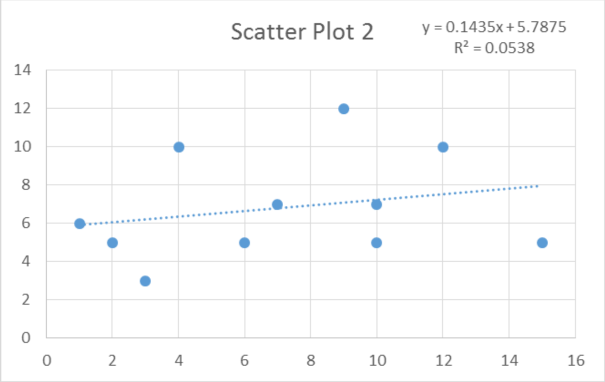


The plot at the left is a **histogram**. As explained above, histograms display the frequency of measurements over repeated sampling. For example, this plot indicates that 10 trees were measured to have a height between 75 and 80 feet. Histograms help you visualize the distribution of data (that is, normal vs. skewed).

Section 3: Comparing Data Sets

Data visualization, or graphing, can be a useful tool to compare data sets. By plotting two variables against each other in a **scatter plot**, we can determine whether or not a statistical relationship exists between the two variables. Essentially, we are looking for patterns, the simplest being a **linear relationship** ($y = m*x+b$). In each of the scatter plots below, a linear relationship or model has been "fitted" to *x* and *y* variables (this is also known as **linear regression**). In linear regression, we take a linear model ($y = m*x+b$) and solve for the coefficients (*m* and *b*) that result in a line plotted through the data points with the least distance possible between the points and the line.





Scatter Plot 1 provides an example of two variables that appear to be related; Scatter Plot 2 is an example of two variables that do not appear to be related. We know this because the points plot closely to the line, or model, in Scatter Plot 1, and the R^2 value is high, indicating that the linear model is a "good fit" and can serve to explain the current data set or predict future data.

In Scatter Plot 2, the points are far away from the line and the R^2 value is low. **R^2** (pronounced R-squared), also called the **coefficient of determination**, indicates how well the data fit the statistical model (in this case the linear relationships). $R^2 = 1$ indicates that the data fits the model perfectly, whereas $R^2 = 0$ indicates that the data does not fit the model and that no relationship may exist. The examples above demonstrate data that "fits" the model well (Scatter Plot 1), and data that does not (Scatter Plot 2).

In air quality, two pollutants may demonstrate a linear relationship if they come from the same source. You could also look for relationships between pollutants and environmental conditions, for example temperature, which can help you to further understand your data.

Section 4: The Power of Data Visualization

With software advancements, we now have really interesting ways to visualize data that can help viewers immediately connect to the message behind the data. For example, the two maps on the following page show $PM_{2.5}$ and NO_2 levels across the globe. Look for the following similarities and differences in the map.

Similarity:

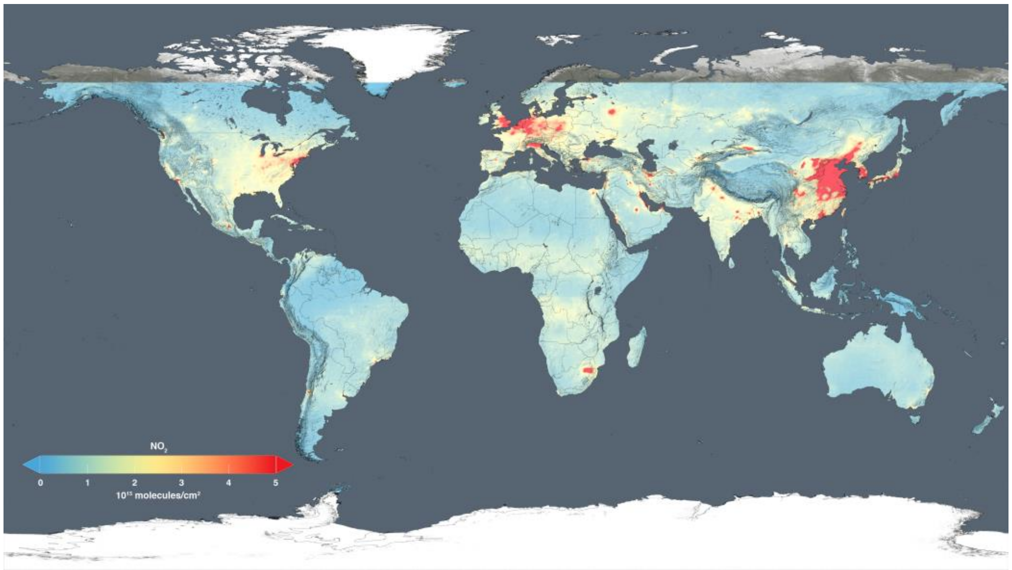
- China appears to have both high $PM_{2.5}$ and high NO_2, which is likely due to its rapid industrialization with little implementation of control technologies to improve air quality

Differences:

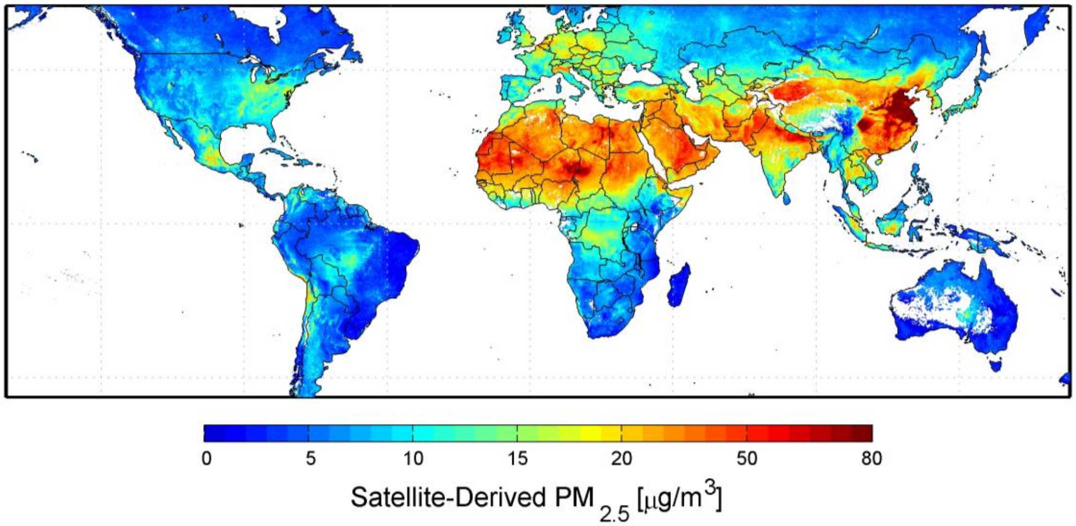
- Look at the high levels of $PM_{2.5}$ in northern Africa and the Arabian Peninsula. These are not very industrial areas, so we do not see the NO_2, but we do see the $PM_{2.5}$ due to the large expanse of desert and dust coming from the deserts.

- Alternatively, many areas of Europe have high NO_2. In Europe, $PM_{2.5}$ is fairly well controlled, but the large number of diesel cars results in elevated NO_2 levels.

Map of nitrogen dioxide levels across the globe:



Map of PM2.5 levels across the globe:



Another way to visualize and communicate data is via “**infographics**,” which are graphics with pictures, graphs, statistics and words that are designed to quickly convey information or data to readers.

////

Image sources:

Map of NO_2: 2005-2014 NASA,Wikimedia Commons https://commons.wikimedia.org/wiki/File:15-233-Earth-GlobalAirQuality-2014NitrogenDioxideLevels-20151214.jpg

Map of PM2.5m: 2010 Aaron van Donkelaar, Dalhousie University, NASA, Wikimedia Commons https://commons.wikimedia.org/wiki/File:483897main_Global-PM2.5-map.JPG