



data oil
is the new

we need to find it,
extract it, refine it,
distribute it and
monetize it.

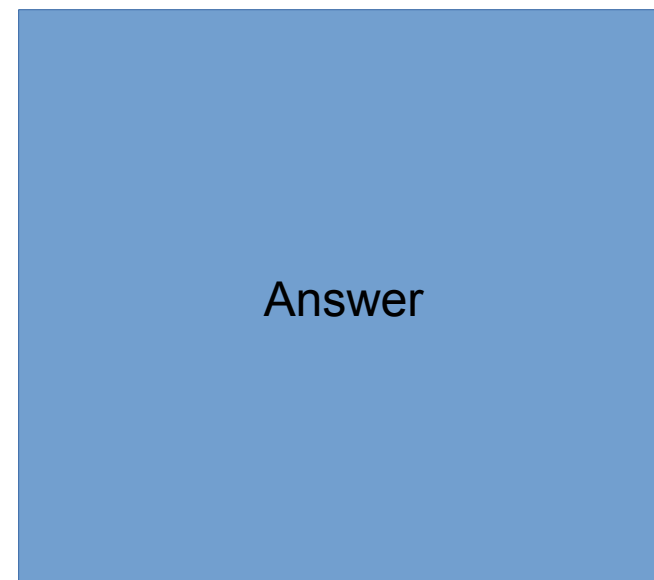
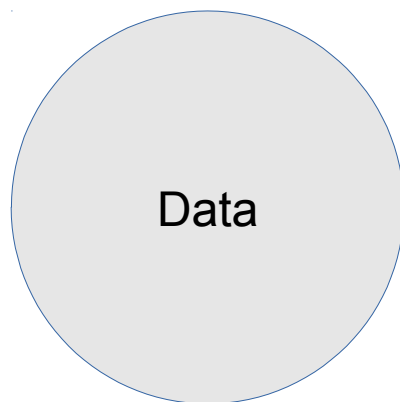
David Buckingham





The goal is to turn data into information, and
information into insight.

—Carly Fiorina

Route to find diamond form data is like moving in a maze



Qualitative Data	Quantitative Data
<p>Overview:</p> <ul style="list-style-type: none"> Deals with descriptions. Data can be observed but not measured. Colors, textures, smells, tastes, appearance, beauty, etc. Qualitative ? Quality 	<p>Overview:</p> <ul style="list-style-type: none"> Deals with numbers. Data which can be measured. Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc. Quantitative ? Quantity
<p>Example 1:</p> <p><i>Oil Painting</i></p>  <p>Qualitative data:</p> <ul style="list-style-type: none"> blue/green color, gold frame smells old and musty texture shows brush strokes of oil paint peaceful scene of the country masterful brush strokes 	<p>Example 1:</p> <p><i>Oil Painting</i></p>  <p>Quantitative data:</p> <ul style="list-style-type: none"> picture is 10" by 14" with frame 14" by 18" weighs 8.5 pounds surface area of painting is 140 sq. in. cost \$300

Data – Types of Variables

- **Quantitative variables** take numerical values whose "size" is meaningful. Quantitative variables answer questions such as **"how many?"** or **"how much?"**

For example, it makes sense to add, to subtract, and to compare two persons' weights, or two families' incomes: These are quantitative variables. Quantitative variables typically have measurement units, such as pounds, dollars, years, volts, gallons, megabytes, inches, degrees, miles per hour, pounds per square inch, BTUs, and so on.

- **Qualitative Variables:** Some variables, such as social security numbers and zip codes, take numerical values, but are not quantitative: They are **qualitative or categorical variables**.

The sum of two zip codes or social security numbers is not meaningful. The average of a list of zip codes is not meaningful.

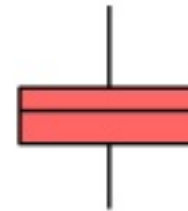
Qualitative and categorical variables typically do not have units. Qualitative or categorical variables—such as gender, hair color, or ethnicity—group individuals. Qualitative and categorical variables have neither a "size" nor, typically, a natural ordering to their values. They answer questions such as **"which kind?"** The values categorical and qualitative variables take are typically adjectives (for example, green, female, or tall). Arithmetic with qualitative variables usually does not make sense, even if the variables take numerical values. Categorical variables divide individuals into categories, such as gender, ethnicity, age group, or whether or not the individual finished high school

Statistics - Refresher

- **Statistics** is the science of collecting, organizing, summarizing, analyzing and interpreting data.
- **Descriptive Statistics:** When performing *descriptive statistics* you collect, organize, summarize, and graphically present data; then you are able to make conclusions about said data.
- **Inferential Statistics:** *Inferential statistics* are used when you want to make predictions and inferences about a larger group (a whole population) from data that was collected from a smaller group (a sample population)

- Descriptive Statistics

- ❑ When analyzing a graphical display, you can draw conclusions based on several characteristics of the graph.
- ❑ **You may ask questions such as:**
 - Where is the approximate middle, or center, of the graph?
 - How spread out are the data values on the graph?
 - What is the overall shape of the graph?
 - Does it have any interesting patterns?



Descriptive Statistics

Involves organizing, summarizing, and displaying data.

e.g. Tables, charts, averages



Inferential Statistics

Involves using *sample data* to draw conclusions about a *population*.



Descriptive Statistics

- Organise
- Summarise
- Simplify
- Describe and present data

Inferential Statistics

- Generalise from samples to populations
- Hypothesis testing
- Make predictions

Common Terms

- **Distribution:** The pattern of values in the data, showing their frequency of occurrence relative to each other.
- **Function:** A function is a relationship where each input number corresponds to one and only one output number
- **Model:** A model is a formula where one variable (response or outcome variable) varies depending on one or more independent variables (covariates). A model tries to establish a relationship among data points. One of the simplest models we can create is a **Linear Model** where we start with the assumption that the dependent variable varies linearly with the independent variable(s). Linear Model has a “constant” rate of change. An exponential Model has a “constant percent” rate of change. So if a population grows by 10 people per year(given the initial population as 100), it's a linear growth and the model will be:

$$P(t)=100+10t$$

But if a population grows by 10% each year(given the initial population as 100), it's an exponential growth and the model will be

$$P(t)=100(1+10\%)^t$$

A statistical model is a “mathematical” description of data

Measures of Central Tendency

Central tendency refers to the most typical value in a set of numbers

- **Median** is the half-way point of data. The median is the number that divides the (ordered) data in half—the smallest number that is at least as big as half the data. At least half the data are equal to or smaller than the median, and at least half the data are equal to or greater than the median. If the distribution is skewed, median is typically used to describe the center.
- **Mode:** The value that has highest frequency. Most frequently occurring value in the data set or the most popular value. It's the only measure of central tendency that can be used with nominal variables.
- **Mean:** The mean (more precisely, the arithmetic mean) is commonly called the average. It is the sum of the data, divided by the number of data. If there are outliers in data, mean can be strongly influenced. In such cases, median is more appropriate.

For qualitative and categorical data, the mode makes sense, but the mean and median do not

Definition of

Median

[more ...](#)

The middle number (in a sorted list of numbers).

10 11 13 15 16 23 26

middle number

To find the Median, place the numbers you are given in value order and find the middle number.

Example: find the Median of {13, 23, 11, 16, 15, 10, 26}.

Put them in order: {10, 11, 13, 15, 16, 23, 26}

The middle number is 15, so the median is 15.

(If there are two middle numbers, you average them.)

Definition of

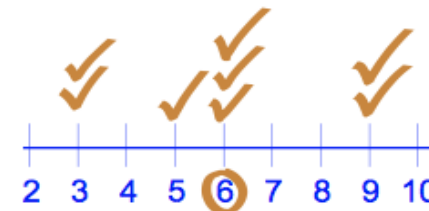
Mode

[more ...](#)

The number which appears most often in a set of numbers.

Example: in {6, 3, 9, 6, 6, 5, 9, 3} the Mode is 6 (it occurs most often).

6, 3, 9, 6, 6, 5, 9, 3



Percentiles:

Assume that the elements in a data set are rank ordered from the smallest to the largest.

The values that divide a rank-ordered set of elements into 100 equal parts are called **percentiles**.

Percentile: the value below which a percentage of data falls.

Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:



That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Quartiles:

The median of a data set is located so that 50% of the data occurs to the left of the median (and 50% of the data occurs to the right of the median). There is no reason to restrict our attention to the 50% level. For example, we can find a point where 25% of the data occurs on its left and 75% to its right. These points are known as the “first quartile” and “third quartile” respectively

Quartiles are the values that divide a list of numbers into quarters:

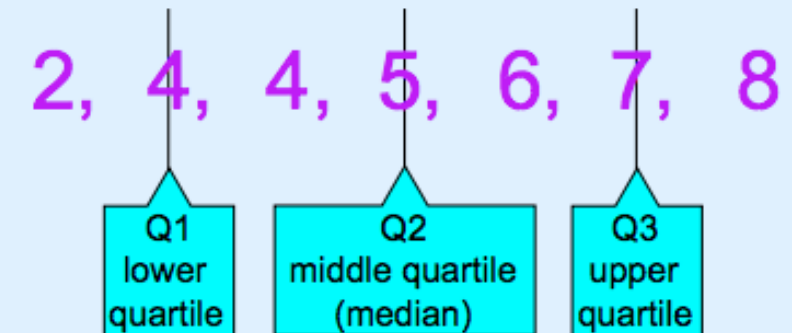
- Put the list of numbers **in order**
- Then cut the list into **four equal parts**
- The Quartiles are at the "cuts"

Like this:

Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

Cut the list into quarters:

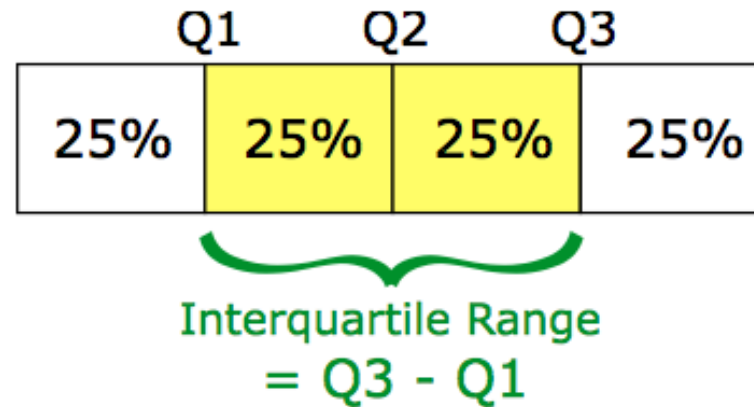


And the result is:

- Quartile 1 (Q1) = 4
- Quartile 2 (Q2), which is also the Median, = 5
- Quartile 3 (Q3) = 7

Interquartile Range

The "Interquartile Range" is from Q1 to Q3:

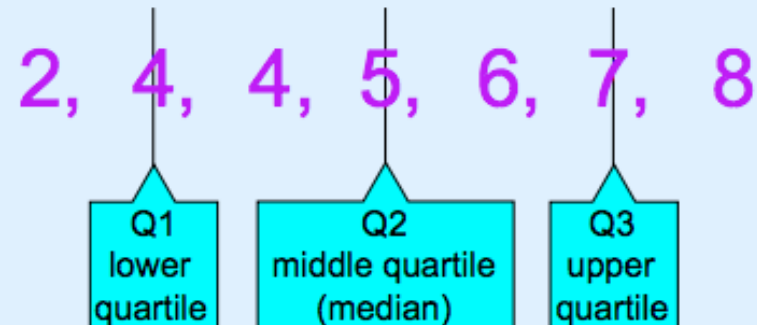


Difference between the 25th and 75th percentile.

It describes the **middle 50%** of the observations.

To calculate it just **subtract Quartile 1 from Quartile 3**, like this:

Example:



The **Interquartile Range** is:

$$Q3 - Q1 = 7 - 4 = 3$$

Measures of Dispersion

These measure the extent of variability in data. Range, interquartile range and standard deviation are the three commonly used measures of dispersion.

Range: Difference between the largest and smallest observation in the data.

Standard Deviation: It is the measure of **spread of data about the mean**. It measures roughly how far off the entries are from their average. It tells us how the data is spread out. **The more the SD, the more spread out data is.** Since its simply a measure, it can't be negative.

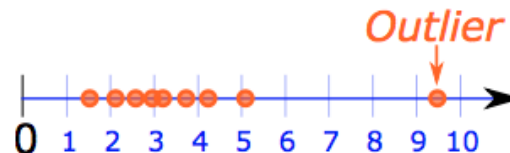
$$s = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

When you add a constant to a list of values, the average also adds up by constant but the SD doesn't change. If you multiply by a constant, the new average and new SD also get multiplied by that constant.

- **Variance:** Mean of Squared deviations. Or simply, it's the square of Standard deviation.
- **Outlier:** An outlier is a data point that lies outside the general range of the data. In the presence of outliers, the mean of the dataset will be significantly affected. In such cases, median makes for sense.

Outlier < Q1 - 1.5*(IQR)

Outlier > Q3 + 1.5*(IQR)



Box and Whisker Plot:

It's a visual representation of Min, Max, Median and quartiles on a single graph.

Its mainly used for identifying outliers easily.

4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11

Put them in order:

3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

Cut it into quarters:

3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18

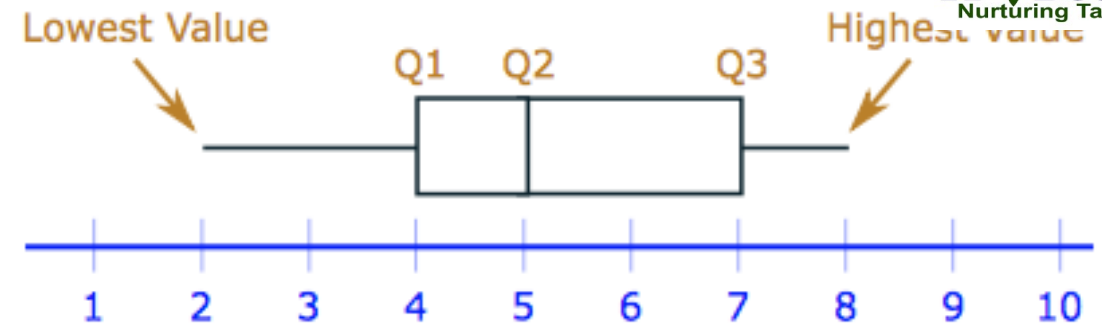
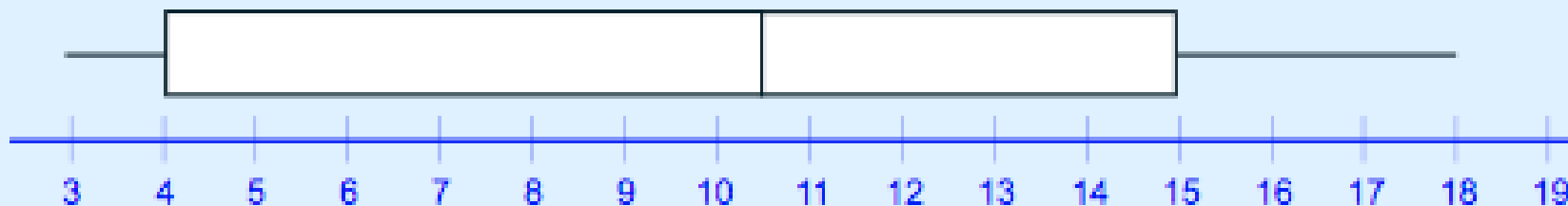
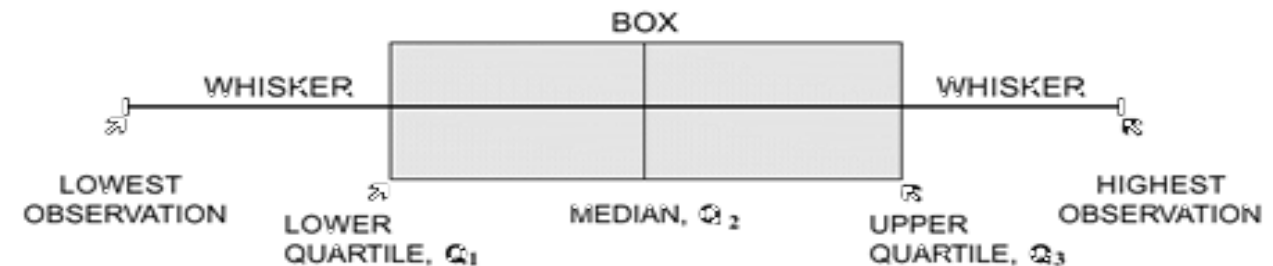


Figure 1. Box and whisker plot



And the **Interquartile Range** is:

$$Q3 - Q1 = 15 - 4 = 11$$

$$\text{Outlier} < 4 - 1.5 \times 11 = 4 - 16.5 = -12.5$$

$$\text{Outlier} > 15 + 1.5 \times 11 = 15 + 16.5 = 31.5$$

Some more Terms...

Significance of SD: SD gives you an insight that how much your data is spread out. With the help of SD you can compare 2 datasets more effectively. If the average of 2 data sets is same, it does not mean that the SD will be same. E.g 99,100,101 and 0, 100, 200 have same mean i.e 100 but they have different standard deviations. The SD of (99,100,101) is only 1 but the SD of (0,100,200) is 100 which is very large.

Lets say the average starting salary in a company is 80000\$. Would you consider joining it? There may be few outliers which may have skewed the average. Additionally, if you know that SD is 2000\$, you may consider joining it.

Definition of

Z-score

[more ...](#)

Z Score: A z-score is the measure of the number of standard deviations a particular data point is away from the mean i.e how many standard deviation away from mean is the observed value. Its also called Z-value

$$Z = \frac{\text{Deviation from mean}}{\text{Standard Deviation}}$$

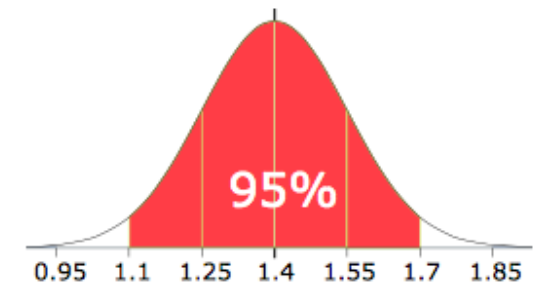
How many standard deviations a value is from the mean.

In this example, the value 1.7 is 2 standard deviations away from the mean of 1.4, so **1.7 has a z-score of 2.**

Similarly 1.85 has a z-score of 3.

So to convert a value to a Standard Score ("z-score"):

- first subtract the mean,
- then divide by the standard deviation



Covariance

Variance and Standard Deviation only operate on 1 dimension so that you could only calculate the standard deviation for each dimension of the data set independently of the other dimensions. There should be a measure to find out how much the dimensions vary from the mean *with respect to each other*. Covariance is such a measure. Covariance is always measured between 2 dimensions.

If you calculate the covariance between one dimension and itself, you get the variance.

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)}$$

Correlation

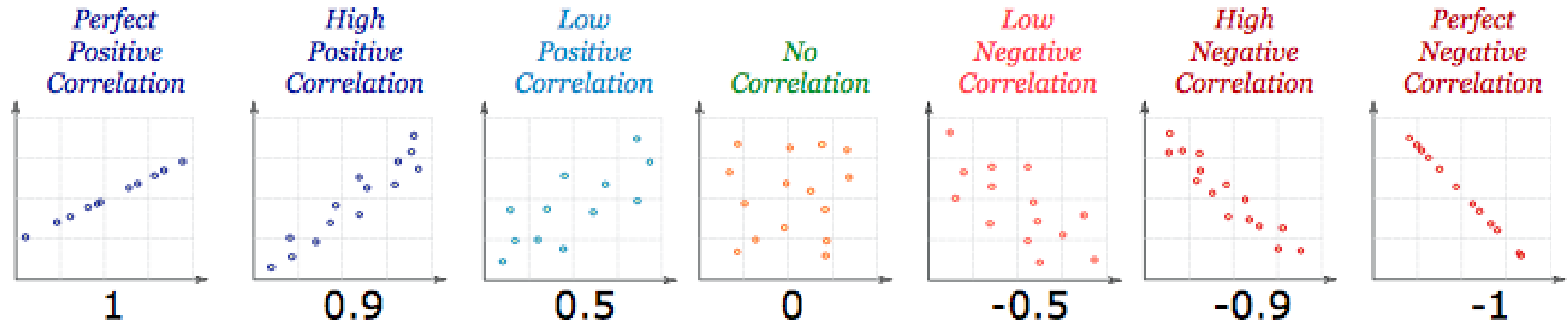
Correlations are mathematical relationships between variables. **Correlation Coefficient (r)** is a number between -1 and 1. It measures linear association i.e how tightly the points are clustered about a straight line. The correlation is said to be linear if the data points lie in an approximately straight line.

A correlation between two variables doesn't necessarily mean that one caused the other or that they're actually related in real life. A correlation between two variables means that there's some sort of mathematical relationship between the two. This means that when we plot the values on a chart, we can see a pattern and make predictions about what the missing values might be. What we don't know is whether there's an actual relationship between the two variables, and we certainly don't know whether one caused the other, or if there's some other factor at work.

Correlation = Covariance(X,Y) / SQRT(Var(X)* Var(Y))

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

Here we look at **linear correlations** (correlations that follow a line).



"Correlation Is Not Causation"

A common saying is "Correlation Is Not Causation" ... which says that a correlation does **not** mean that one thing causes the other.

There can be many reasons the data has a good correlation.

(And it **may** be true that one causes the other, we need to think carefully.)

Multi-collinearity refers to the situation when 2 independent variables are highly correlated. Multi-collinearity generally degrades the performance of linear regression model.

Multi-collinearity means that several variables are essentially measuring the same thing. It doesn't add to the predictive capability of the model and it may make the model fit less well. Since you are predicting an outcome, you want your factors to be independent. Correlation indicates two or more factors are providing your model with similar data which will decrease the model's ability to accurately predict.

Example: Predicting home prices. Square feet and number of bedrooms could be two of your factors considered. But logically you could see how these two measurements would be correlated; likely positive correlation. What if a home you want to predict for only has one room but the sqft of a 5 bedroom home? Your model is 'expecting' 5 bedrooms and that bedrooms add value to the home. Your model will predict price using one room but not as accurately as it would if the bedrooms only slightly varied from your model. Your model would more accurately predict the price if, in this example, bedrooms were removed AND the regression model was created again.