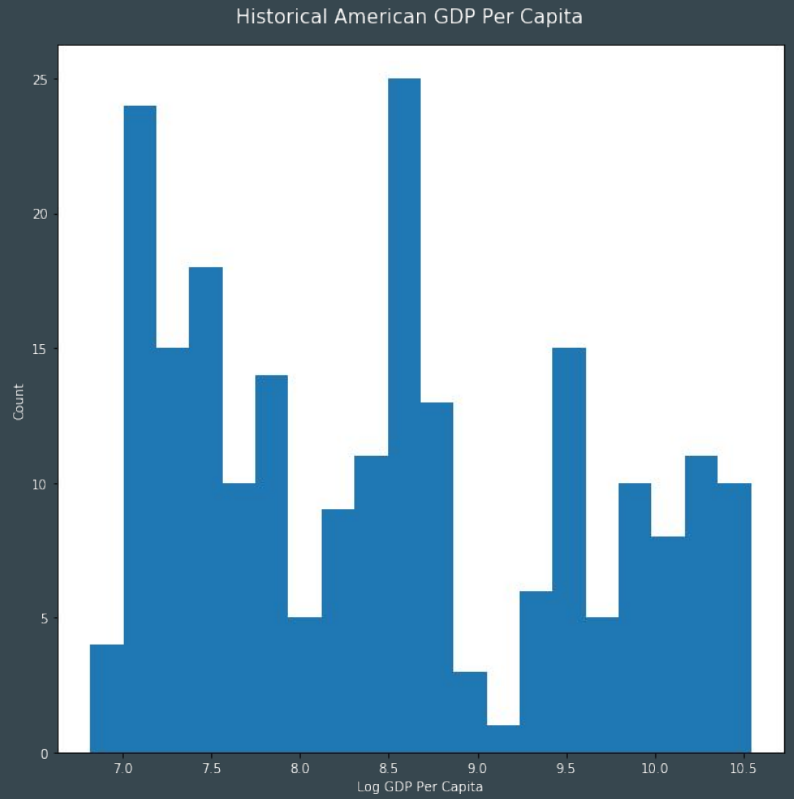
Week Ten: Relationships Between Variables

CS 217

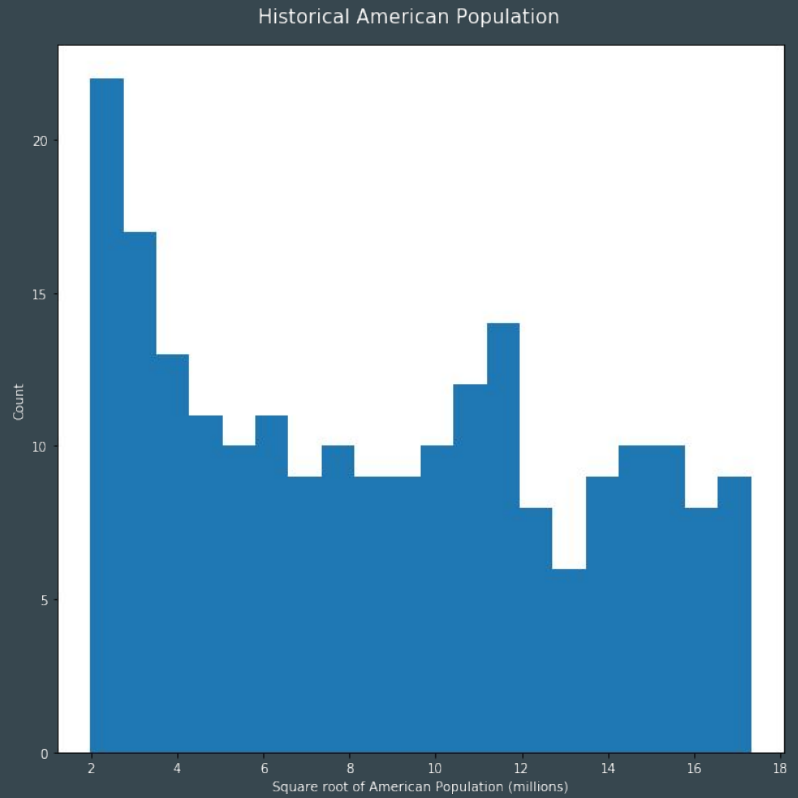
One Variable

- Thus far, everything that we've looked at has been in the context of a single variable
- We can look at a distribution of the annual history of GDP per capita in America and visualize it.



One Variable

- Thus far, everything that we've looked at has been in the context of a single variable
- Or we can look at the distribution of the annual history of the population in America and visualize it.



Two Variables

- But what about the relationship between these two variables?

Covariance

- **Covariance** is a measure of the tendency of two continuous variables to vary together
- For instance, if the X variable is greater than the mean of X while the Y variable is greater than the mean of Y, the two variables will have a positive covariance.

$$Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\sigma^2 = \frac{(X-\mu)^2}{N}$$

Covariance

- **Covariance** is a measure of the tendency of two continuous variables to vary together
- Or if the X variable is less than the mean of X while the Y variable is less than the mean of Y, the two variables will have a positive covariance.

$$Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\sigma^2 = \frac{(X-\mu)^2}{N}$$

Covariance

- **Covariance** is a measure of the tendency of two continuous variables to vary together
- If the X variable is less than the mean of X while the Y variable is greater than the mean of Y, or vice versa, the two variables will have a negative covariance.
- Note the similarity between the formula for covariance and the formula for variance.

$$Cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\sigma^2 = \frac{(X-\mu)^2}{N}$$

Correlation

- **Correlation** scales the covariance value between -1 and 1 by diving the covariance value by the product of the standard deviations of each variable
- If correlation is positive, we know that when one variable is high, the other tends to be high
- If correlation is negative, we know that when one variable is low, the other variable tends to be low

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Correlation

- **Correlation** scales the covariance value between -1 and 1 by diving the covariance value by the product of the standard deviations of each variable
- If correlation is positive, we know that when one variable is high, the other tends to be high
- If correlation is negative, we know that when one variable is low, the other variable tends to be low

Correlation	Meaning
-1	Perfect Negative Correlation
-0.7	Strong Negative Correlation
-0.5	Moderate Negative Correlation
-0.3	Weak Negative Correlation
0	No linear relationship
0.3	Weak positive correlation
0.5	Moderate positive correlation
0.7	Strong Positive Correlation
1	Perfect Positive Correlation

Correlation

- Say we have the heights and weights of five people. What is the covariance and correlation of their height and weight?

	Height	Weight
Person 1	66”	125
Person 2	70”	160
Person 3	76”	240
Person 4	62”	250
Person 5	69”	155

Correlation

- Say we have the heights and weights of five people. What is the covariance and correlation of their height and weight?

	Height	Weight
Person 1	66”	125
Person 2	70”	160
Person 3	76”	240
Person 4	62”	250
Person 5	69”	155
Expected Value	68.6”	186

Correlation

- Say we have the heights and weights of five people. What is the covariance and correlation of their height and weight?

	Height	Weight	Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1	66”	125			
Person 2	70”	160			
Person 3	76”	240			
Person 4	62”	250			
Person 5	69”	155			
Expected Value	68.6”	186			

Correlation

- The covariance here is 17.4

	Height	Weight	Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1	66”	125	- 2.6”	-61	
Person 2	70”	160	1.4”	-26	
Person 3	76”	240	7.4”	54	
Person 4	62”	250	-6.6”	64	
Person 5	69”	155	0.4”	-31	
Expected Value	68.6”	186			

Correlation

- The covariance here is 17.4

	Height	Weight	Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1	66”	125	- 2.6”	-61	158.6
Person 2	70”	160	1.4”	-26	-36.4
Person 3	76”	240	7.4”	54	399.6
Person 4	62”	250	-6.6”	64	-422.4
Person 5	69”	155	0.4”	-31	-12.4
Expected Value	68.6”	186			

Correlation

- The covariance here is 17.4

	Height	Weight	Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1	66”	125	- 2.6”	-61	158.6
Person 2	70”	160	1.4”	-26	-36.4
Person 3	76”	240	7.4”	54	399.6
Person 4	62”	250	-6.6”	64	-422.4
Person 5	69”	155	0.4”	-31	-12.4
Expected Value	68.6”	186			**17.4**

Correlation

- The standard deviations are 4.63 and 49.74

			Height - E(Height)	Weight - E(Weight)	
Person 1			- 2.6”	-61	
Person 2			1.4”	-26	
Person 3			7.4”	54	
Person 4			-6.6”	64	
Person 5			0.4”	-31	
Standard Deviation			4.63	49.74	

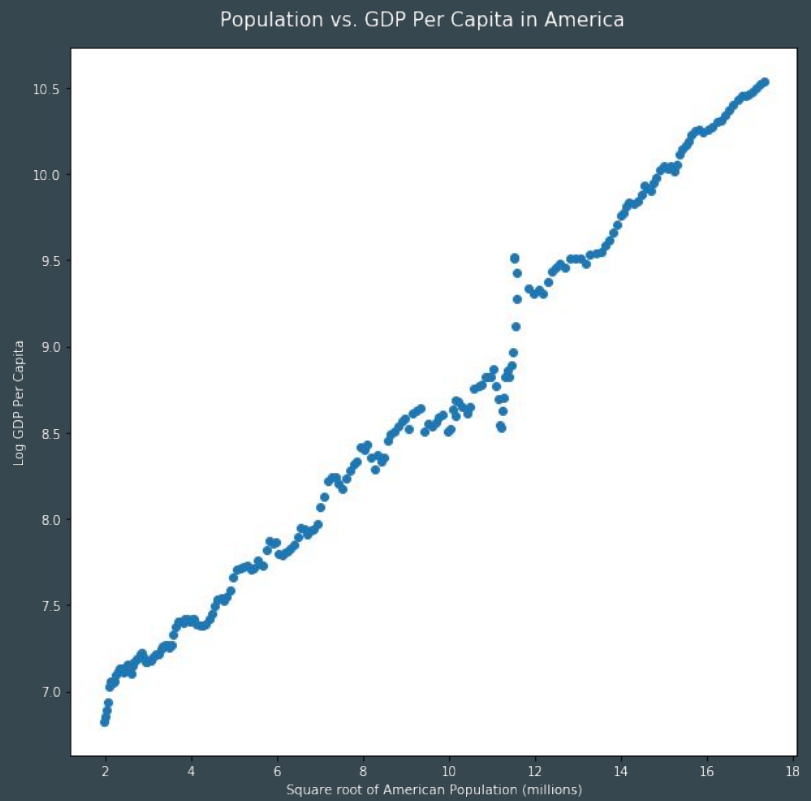
Correlation

- Correlation is 17.4 / (4.63 * 49.74) = 0.0755, which is an extremely weak positive correlation

			Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1			- 2.6”	-61	158.6
Person 2			1.4”	-26	-36.4
Person 3			7.4”	54	399.6
Person 4			-6.6”	64	-422.4
Person 5			0.4”	-31	-12.4
Standard Deviation			4.63	49.74	**17.4**

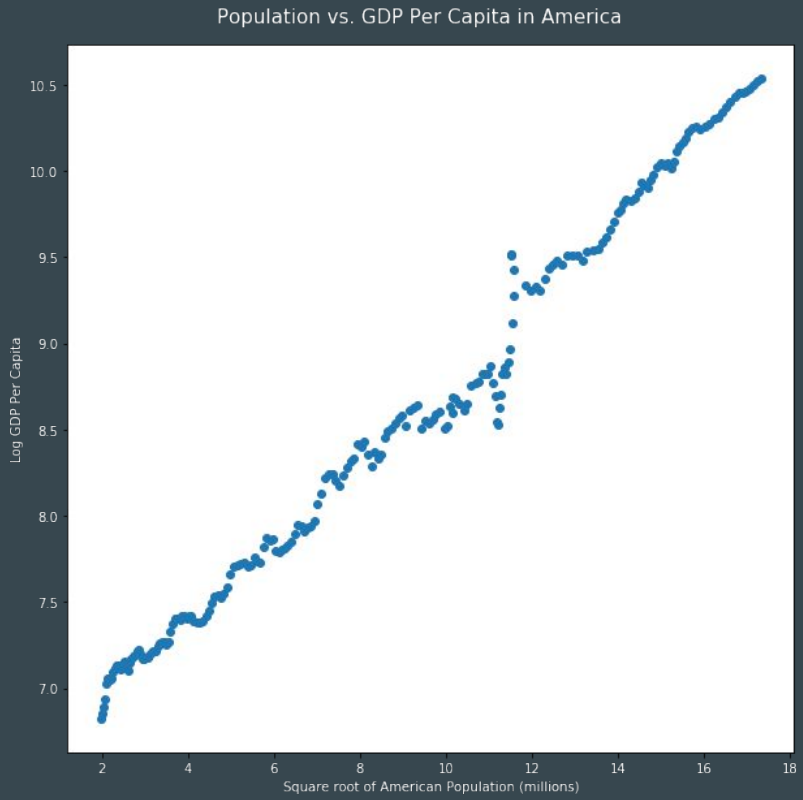
Scatterplots

- A **scatterplot** is a great way to visualize the relationship between two variables and should be done before calculating the covariance and correlation
- To the right is the relationship between population and GDP per capita in American history.
- What does the relationship look like here?



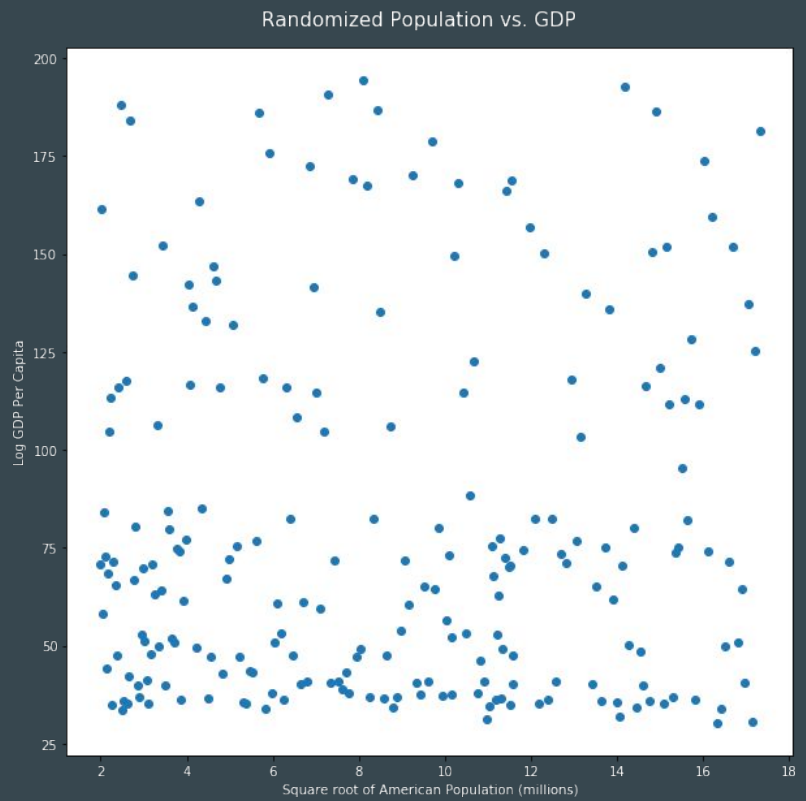
Scatterplots

- The two variables look like they have a **very strong linear correlation,** and indeed, the correlation here is 0.9934.

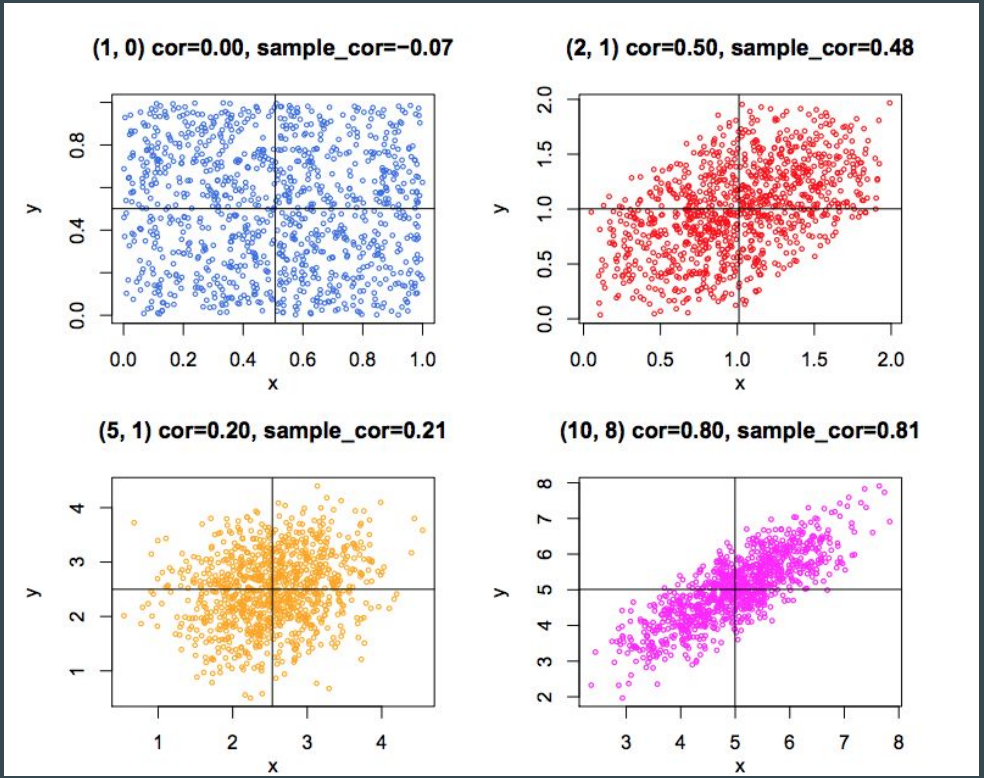


Scatterplots

- These two variables, on the other hand, look like they have no correlation
- Indeed, the correlation here is 0.047



Scatterplots



Hypothesis Testing

- Since you all love hypothesis testing, there is a way to implement it for correlation values
- While correlation values closer to -1 or 1 are indicative of a stronger correlation, we can use hypothesis testing to test whether there is a correlation at all
- This is especially useful for small datasets

$$t = \frac{r*\sqrt{n-2}}{\sqrt{1-r^2}}$$

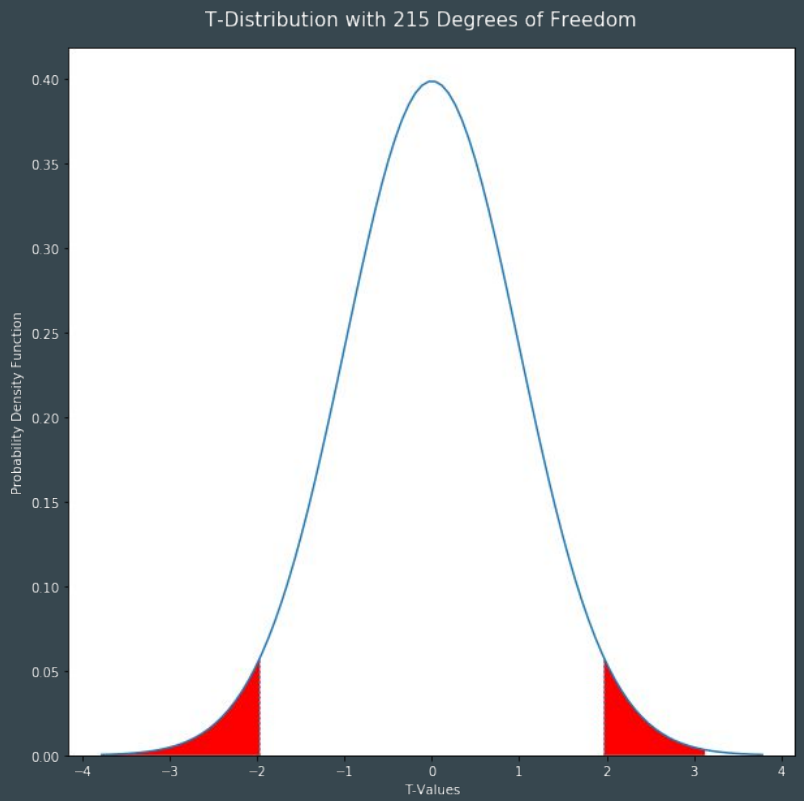
Hypothesis Testing

- Using the formula, we can find the t-value from the r value and the length of the respective distributions (n)
- We can then see if the t-value is in the rejection region for a ‘standard’ t-distribution with a mean of 0, standard deviation of 1, and n - 2 degrees of freedom
- The test is two-tailed, and the significance level at which you will test is up to you.

$$t = \frac{r*\sqrt{n-2}}{\sqrt{1-r^2}}$$

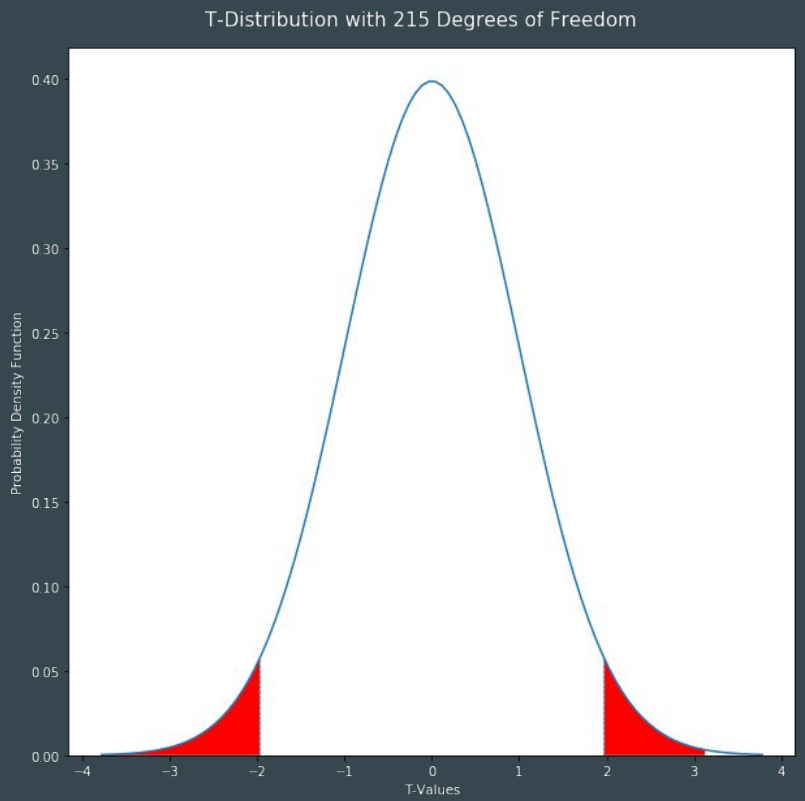
Hypothesis Testing

- For our example of population vs. GDP, there were 217 data points.
- A T-Distribution with a mean of 0, standard deviation of 1, and 215 degrees of freedom has rejection regions at the 0.05 significance level of less than -1.97 and greater than 1.97.
- The null hypothesis is that there is no correlation and the alternate hypothesis is that there is a correlation.



Hypothesis Testing

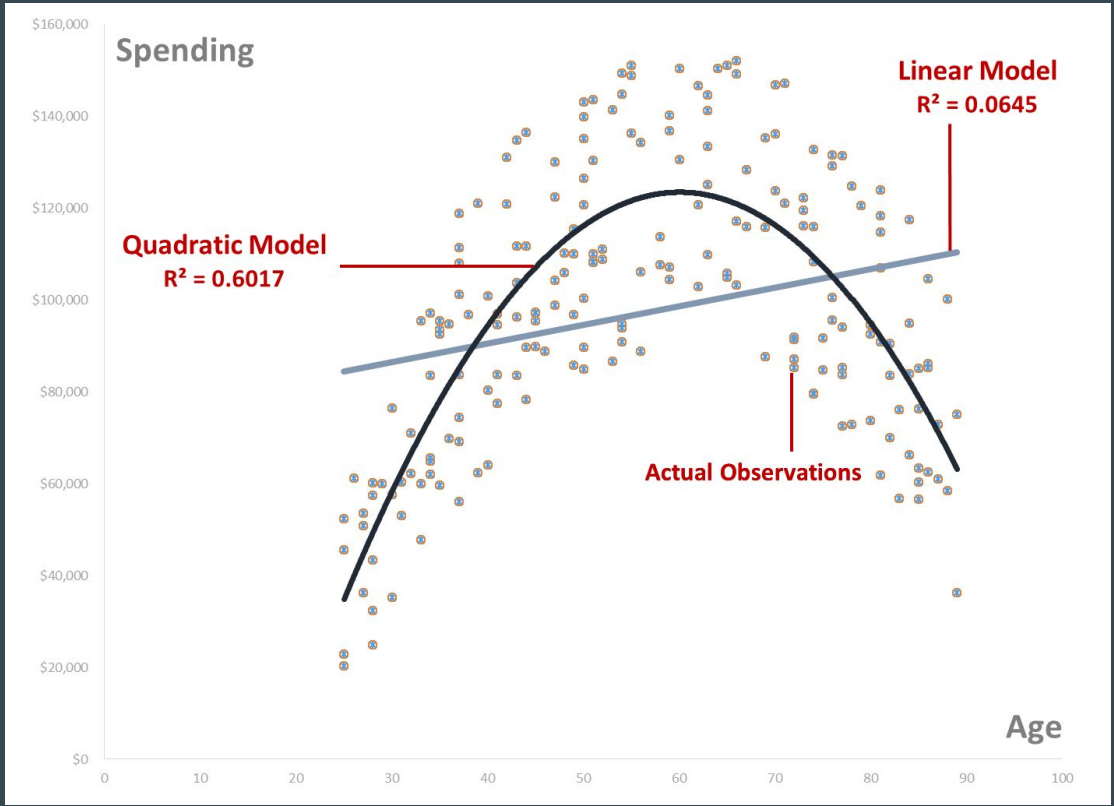
- Our first correlation of 0.9934 has a T-value of 127.27, so we can clearly reject the null hypothesis that there is no correlation.
- Our second correlation of 0.04 has a T-value of 0.69, so we fail to reject the null hypothesis that there is no correlation.



Linear Relationships

- While Pearson's correlation metric is a great way of measuring **linear relationships**, it doesn't capture non-linear relationships as well.

Linear Relationships

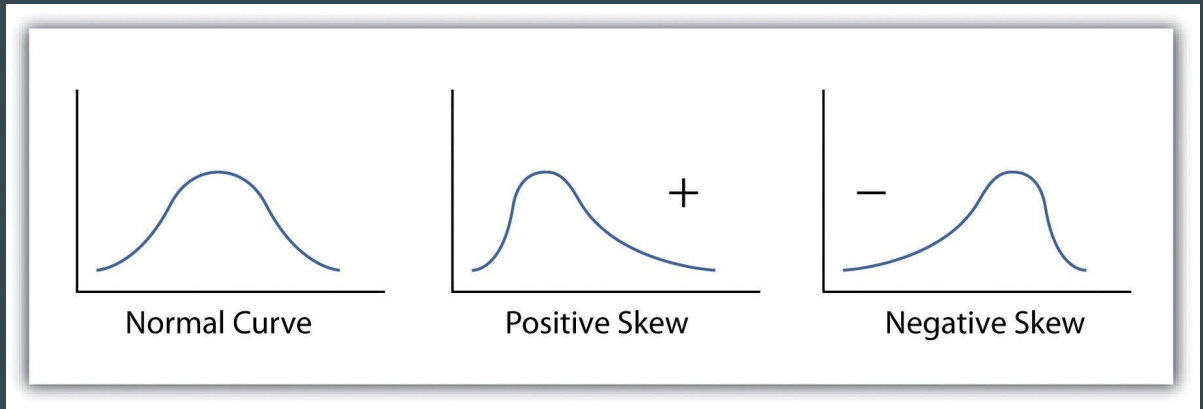


Linear Relationships

- While Pearson's correlation metric is a great way of measuring **linear relationships**, it doesn't capture non-linear relationships as well.
- There are two ways to solve this. First, we can **transform** our data to make the relationship linear.

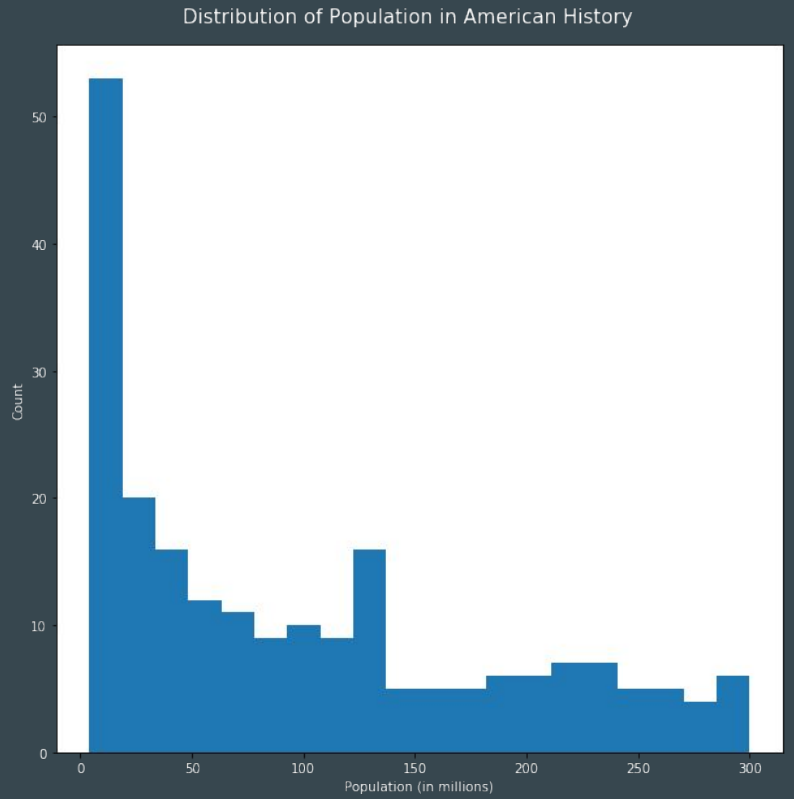
Transformation

- Transformation in this case largely involves **unskewing** skewed data.
- A continuous distribution where more data is clustered around the lower end of the distribution is **positively skewed**.
- A continuous distribution where more data is clustered around the higher end of the distribution is **negatively skewed**.



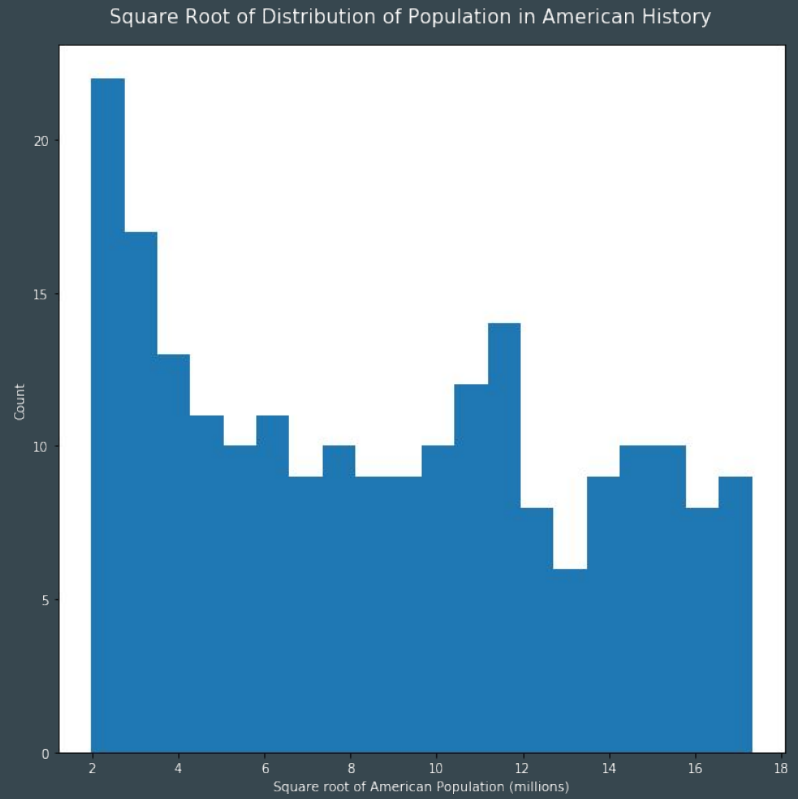
Transformation

- Some common ways to transform data that is **positively skewed** are taking the log of the data, the square root of the data, or the square cube of the data.
- Generally we want to normalize the data, though it is up to us to experiment with different methods to ensure this happens.



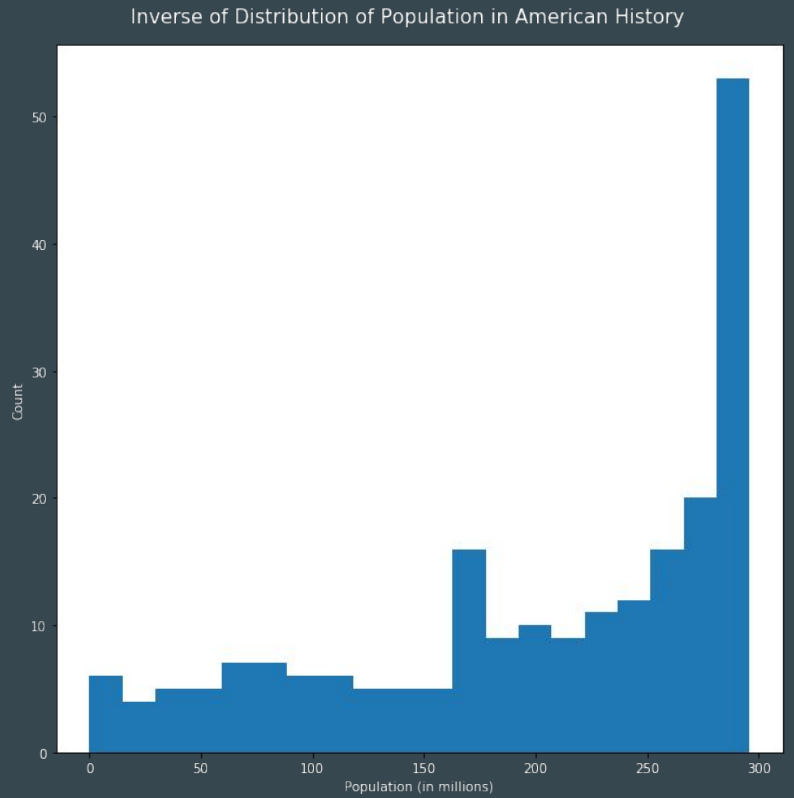
Transformation

- To the right we have now taken the square root of the original distribution.
- While there is still somewhat of a skew, the distribution is much closer to normal.



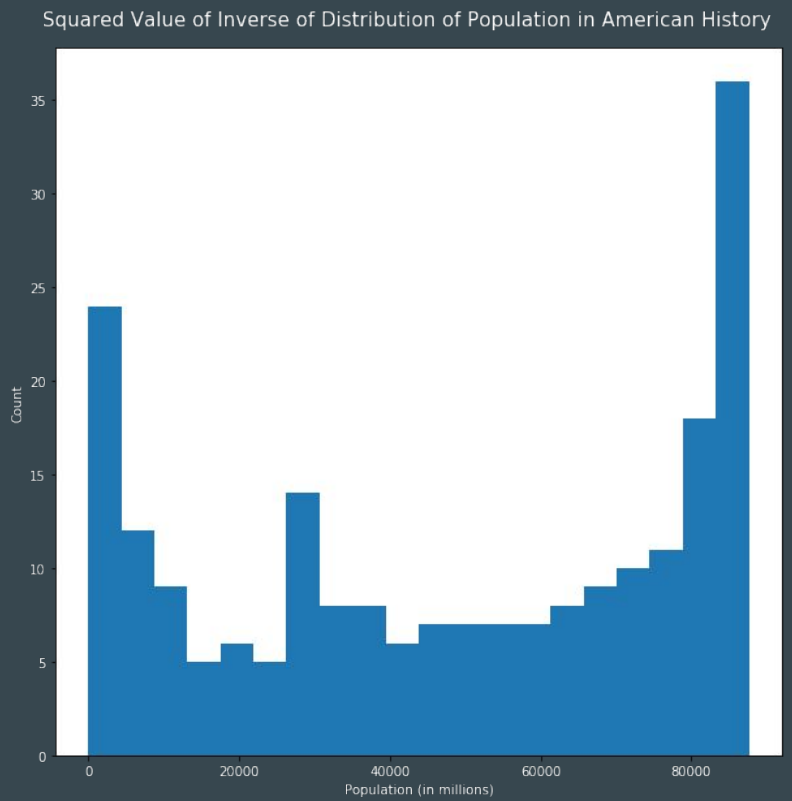
Transformation

- Some common ways to transform data that is **negatively skewed** are taking the exponential value, or squaring or cubing the data.
- Generally we want to normalize the data, though it is up to us to experiment with different methods to ensure this happens.



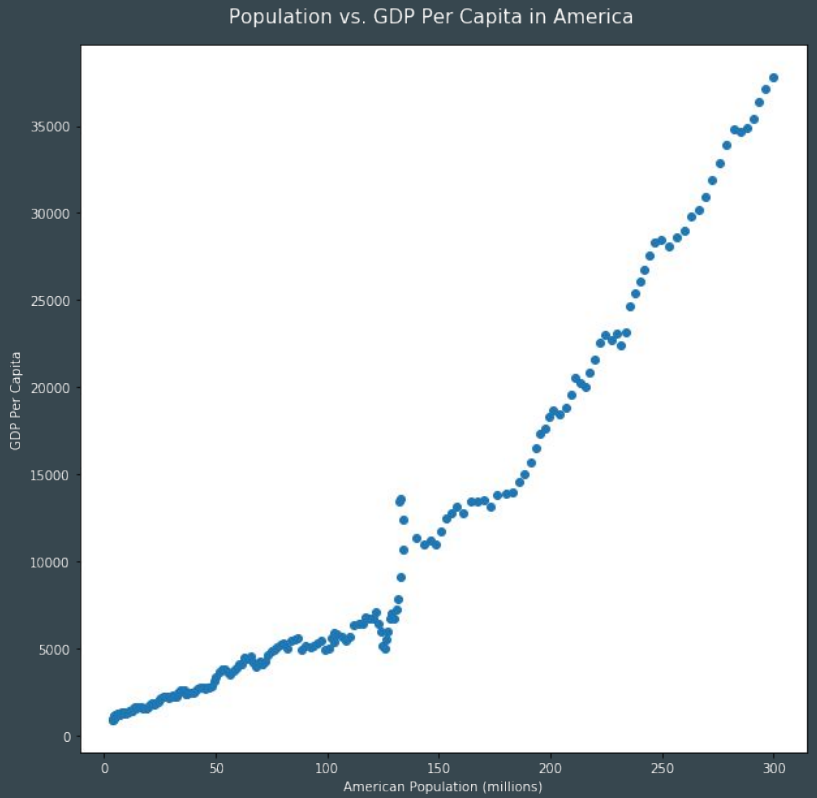
Transformation

- To the right we have now squared the original distribution.
- While there is still somewhat of a skew, the distribution is much closer to normal.



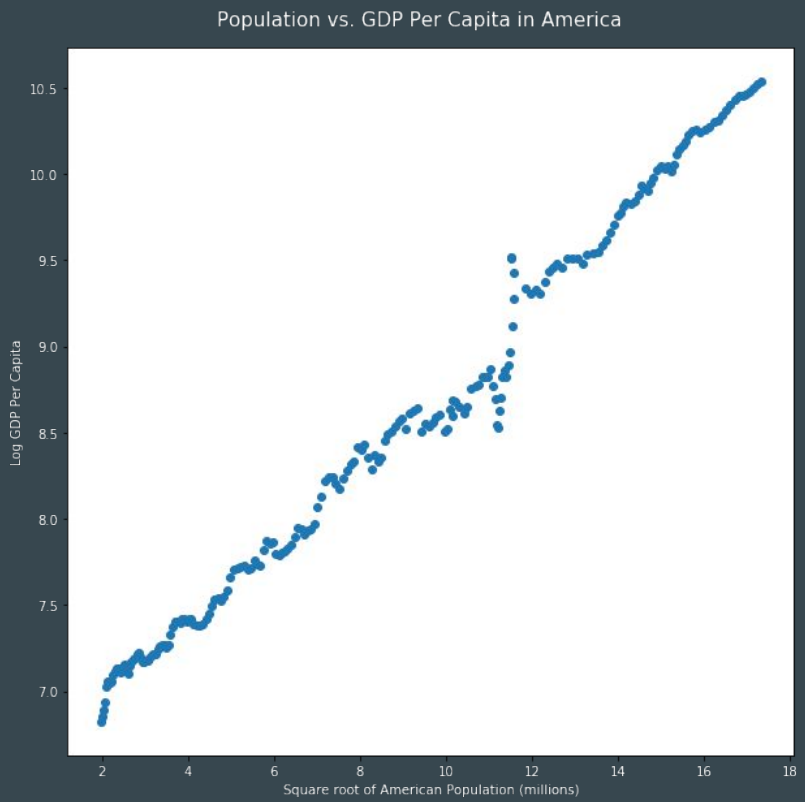
Transformation

- To the right is the relationship between population and GDP in America, with both variables not being transformed.
- While this isn't as extreme as the example shown earlier, note that the true relationship appears to be nonlinear and most of the data points are clustered at the lower end of the distribution



Transformation

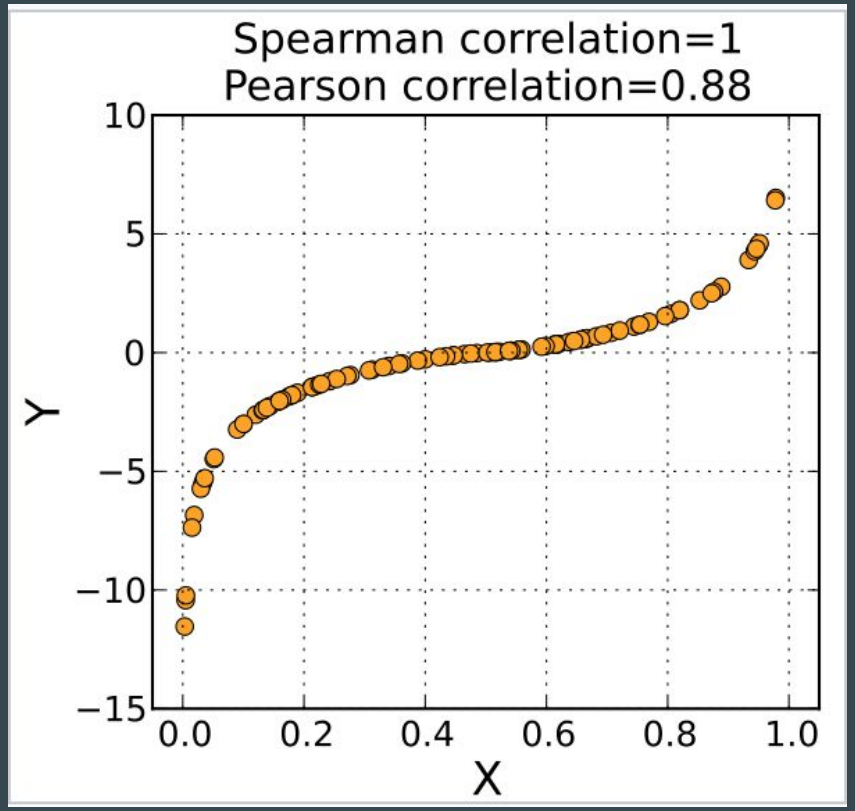
- After transformation of both variables, there is a clear linear relationship between the two variables, with data points across the distribution.
- When we do linear regression next week, these assumptions will come into play.

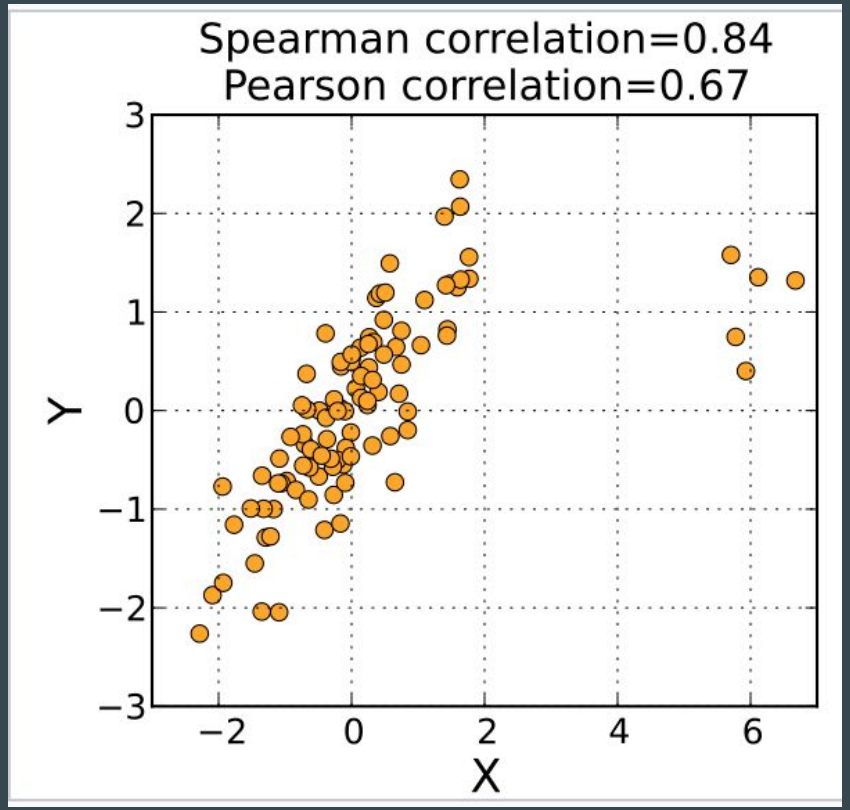


Correlation

- **Spearman's Rank Correlation** is an alternative way of calculating correlation that mitigates the effect of outliers and skewed distributions
- Rather than calculate the correlation of the values of two variables, we calculate the correlation of the **ranks** of two variables
- We calculate the correlation (and covariance) in the same manner we did previously, just with our ranked variables rather than the original variables.
- This method is not as susceptible to outliers, and does not require the relationship between the two distributions to be linear.

Correlation





Correlation

- Say we have the heights and weights of five people. What is the covariance and correlation of their height and weight?

	Height	Ranked Height	Weight	Ranked Weight
Person 1	66”	4	125	5
Person 2	70”	2	160	3
Person 3	76”	1	240	2
Person 4	62”	5	250	1
Person 5	69”	3	155	4
Expected Value	68.6”	2.5	186	2.5

Correlation

- Here, the covariance is 0, meaning that the correlation is also 0.

	Ranked Height	Ranked Weight	Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1	4	5	1	2	
Person 2	2	3	-1	0	
Person 3	1	2	-2	-1	
Person 4	5	1	2	-2	
Person 5	3	4	0	1	
Expected	3	3			

Correlation

- Here, the covariance is 0, meaning that the correlation is also 0.

	Ranked Height	Ranked Weight	Height - E(Height)	Weight - E(Weight)	Height Diff * Weight Diff
Person 1	4	5	1	2	2
Person 2	2	3	-1	0	0
Person 3	1	2	-2	-1	2
Person 4	5	1	2	-2	-4
Person 5	3	4	0	1	0
Expected	3	3			0

Common Mistakes with Correlations

- Over time, ice cream consumption is correlated with the rate of pool drownings. Eating ice cream will make you drown in the pool!

Common Mistakes with Correlations

- Over time, ice cream consumption is correlated with the rate of pool drownings. Eating ice cream will make you drown in the pool!
- Both of these things happen at the same type of year, but one has nothing to do with the other. Correlation **does not** imply causation!

Common Mistakes with Correlations

- People who formerly smoked are more likely to die of lung cancer than people who currently smoke. If you're a smoker, DON'T QUIT.

Common Mistakes with Correlations

- People who formerly smoked are more likely to die of lung cancer than people who currently smoke. If you're a smoker, DON'T QUIT.
- When lifelong smokers find out they have lung cancer, they quit smoking and become former smokers. This is called **reverse causation.**

Common Mistakes with Correlations

- Golfers are more prone to heart disease, cancer, and arthritis than the general population. Golf is bad for you!

Common Mistakes with Correlations

- Golfers are more prone to heart disease, cancer, and arthritis than the general population. Golf is bad for you!
- Golfers tend to be older than the general population, and older people are more prone to these diseases than the rest of the general population. This is called **omitted variable bias**

Causation

- Given these common fallacies, how do we actually prove evidence of causation?
- The best way is a **randomized control trial**, where subjects are assigned randomly to two groups: a **treatment group** and a **control group**
  - This is used in laboratory sciences, medicine, and a few other disciplines
  - It is much harder to replicate outside of these instances for both ethical, financial and practical reasons
- A **natural experiment** is when the split between a treatment and control group occurs naturally, such as examining differences between people from different countries or education levels
  - Of course, the other fallacies we spoke about earlier may apply