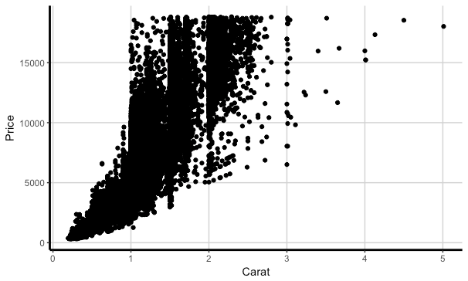
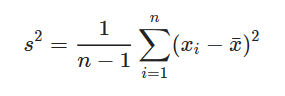
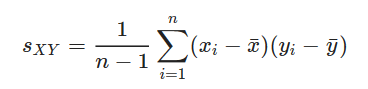
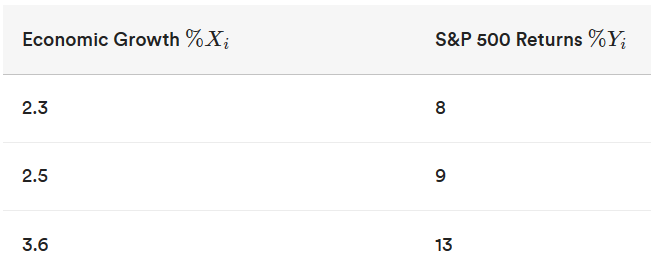
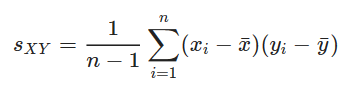
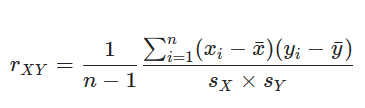
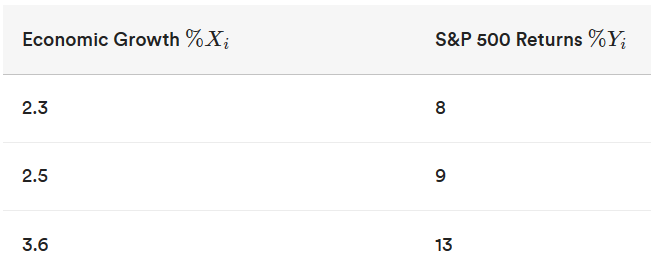
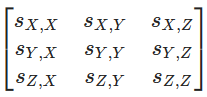
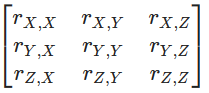
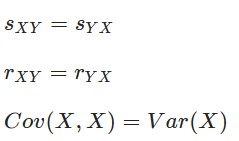
Unit 1-4: Revisiting Correlation and Covariance

* Correlation and Covariance
  + Here's a nifty trick for you: If you know that two variables are highly correlated — for example, house prices and number of bedrooms — you can estimate the value of one variable based on the other. So we could estimate that Petunia's five-bedroom home would be valued higher than Elmer's two-bedroom home next door.
  + That trick comes to you courtesy of your new data science buddies — correlation and covariance.
* Distributions
  + Recall that a distribution is the set of all values a variable can take on and how frequently each value occurs.
  + Distributions help us to arrive at many conclusions. For example:
    - Given a set of flight arrival times, we can approximate what percentage of all flights will arrive late.
    - Given a set of exam scores, we can estimate how many students will score below 60 percent.
    - Given a set of taxi fares, we can look at the spread of taxi fares in a city.
  + Good stuff for sure, but distributions don't tell the whole story.
* Distributions vs. Relationships
  + There are certain questions a distribution cannot answer. For example:
    - What is the effect of the airline on whether or not a flight will arrive late?
    - How well do exam scores from the previous semester predict exam scores for the next semester?
    - How do taxi fares differ at various times of the day?
  + To answer any of these, we examine the relationships between two variables, instead of the distribution of one.
* Bivariate Statistics
  + When we seek to describe a distribution of data, we calculate a number of summaries that quantify the center or spread of a distribution, such as mean, median, and standard deviation. These are called **univariate statistics** because they are statistics calculated from only one variable.
  + **Bivariate statistics** on the other hand, are statistics calculated from two variables. The two most widely known bivariate statistics are correlation and covariance.
    - The underlying question we try to answer is: "As one variable changes, how does the other change?"
    - Correlation and covariance get the job done.
  + The scatterplot below is from a data set on diamonds. Using the fields for carat size and price, we can see that, as the carat size increases, so does the price.
    - 
    - Note: We can also see a few outliers from this scatterplot that should be looked into further.
* Variance Review
  + Recall that the formula for the variance of a sample is given by:
  + 
  + Remember that variance quantifies the spread of a variable's distribution. We find, roughly, the average distance between every observation and the mean.
  + Covariance follows a similar process.
* Covariance
  + The covariance formula for a sample is given by:
  + 
  + We're still using (*x*i−) to quantify change, but now we're looking at how X and Y change together instead of just X by itself.
  + We even use s to denote covariance, but add a subscript XY to indicate that we're calculating the covariance between variables X and Y.
    - Note: You may, alternatively, see this written as .
  + Let's walk through an example of how to calculate covariance. Imagine we have the following table, which details economic growth and S&P 500 returns.
    - 
    - Find the sample mean for x and y.
      * x = 2.8
      * y = 10
    - For each value x, find x - the mean of x.
      * (2.3−2.8)=−0.5,(2.5−2.8)=−0.3,(3.6−2.8)=0.8
    - For each value y, find y - the mean of y.
      * (8−10)=−2,(9−10)=−1,(13−10)=3
    - Multiply the first value x (-0.5) with the first y value (-2), the second x value with the second y value, and so on.
      * (−0.5∗−2)=1,(−0.3∗−1)=0.3,(0.8∗3)=2.4
    - Sum the output of the values from Step 4 and divide by n−1.
      * (1+0.3+2.4)/(3−1)=(3.7/2)=1.85
* Interpreting the Covariance
  + Note that covariance measures the strength and direction of the **linear relationship** between X and Y.
  + Any conclusions we draw about relationships are limited to linear relationships. We'll get into this more specifically in a bit.
  + The covariance can take on any real value: positive, zero, or negative.
    - Values greater than zero indicate that the variables have a positive linear relationship
    - Values less than zero indicate that the variables have a negative linear relationship.
    - Values close to zero indicate little to no linear relationship between the two variables.
  + Beyond this, however, it can be difficult to interpret covariances.
  + Let's say we have two pairs of variables. The first pair of variables has a sample covariance of 2 and the second pair of variables has a covariance of 5.
    - We might be inclined to say that the second pair of variables has a stronger linear relationship than the first pair of variables, but that wouldn't be quite correct.
    - Covariance, much like standard deviation and variance, relies heavily on the original scale of the variables being compared. If we have two variables X and Y, where X is centimeters and Y is in liters, then *s*XY​ is expressed with the units "centimeters times liters."
    - Sounds difficult to interpret? We think so too! But fear not — correlation is a measure that helps to sidestep these issues.
* Pearson's Correlation
  + Recall that the covariance is given by:
    - 
  + Pearson's correlation coefficient, *r*XY​ — which is often simply referred to as "correlation" — is given by:
    - 
  + The correlation is calculated by taking the covariance and scaling it by the standard deviations of X and Y.
  + Let's walk through an example using the same covariance data from earlier. First, give it a try on your own to calculate the Pearson's correlation of the data below:
    - 
    - 
      1. Calculate (*x*i−)
      2. Calculate (*y*i−)
      3. Multiply the values of Step 1 and Step 2 for each row of the data and sum down. (3.7)
      4. Square the values in Step 1 and sum down. (0.98)
      5. Square the values in Step 2 and sum down. (14)
      6. Multiply Steps 4 and 5. (0.98∗14)=13.72
      7. Take the square root of Step 6. (3.704)
      8. Divide Step 3 by Step 7. (3.7/3.704)=.999
  + Interpreting Pearson's Correlation
    - By scaling covariance by the standard deviations of X and Y, the correlation is no longer based on the original units or scale of X and Y. In fact, correlation is a unitless measure. As a result, it's a lot easier for us to interpret and compare correlations.
    - Pearson's *r*XY​ measures the strength and direction of the linear relationship between two variables but is bounded between -1 and 1.
* Correlation and Multicollinearity
  + Correlation is important for determining linear relationships between two variables, but having too many overlapping correlated variables can actually make a machine learning model go kaput.
  + This leads to an idea in data science called **multicollinearity**, which exists when there are a high number of intercorrelations or inter-associations among independent variables.
  + Going back to our example involving home prices and number of bedrooms, imagine we also have a data point for the number of bathrooms in a given home. You might imagine that a house with more bedrooms will also have more bathrooms, which in turn leads to higher home prices. Since bedrooms and bathrooms are two independent variables that are highly correlated, including the number of bathrooms into our overall model may not be necessary.
  + To avoid multicollinearity, data scientists use feature selection, which is the process of narrowing down the number of features in a dataset, in order to strictly focus on those that are important and necessary.
  + In our example, if we know that home price is our dependent variable, we can test the correlation of all our independent variables (e.g. the number of bedrooms and the number of bathrooms) and only select one of the two independent variables into our model. We only need to select one or the other because they are highly correlated with each other; therefore, both are not necessary in order to suitably determine home prices.
* Interpreting Pearson's Correlation
  + If rXY = -1 then X and Y have a perfectly linear negative relationship.
  + If rXY = 1 , then X and Y have a perfectly linear positive relationship.
  + If rXY = 0, then X and Y have no linear relationship.
  + Values of rXY that are close to zero have a weak linear relationship, whereas values that are farther from zero have a stronger linear relationship.
* Correlation and Covariance Matrices
  + Suppose that we have three variables, X, Y, and Z, and want to view their covariances. The covariance matrix for variables X, Y, and Z is given by:
    - 
    - This is convenient because we can look at many covariances at once, instead of needing to manually calculate a covariance for each pair of variables.
    - This structure generalizes for p>1 variables.
  + Correlation and covariances are one-number summaries of our data. When calculating the correlation or covariance for many variables at once, it's customary to use a matrix.
    - A matrix is a rectangular array of numbers.
  + Similar to the covariance matrix, the correlation matrix for variables XXX, YYY, and ZZZ is given by:
    - 
    - This is convenient because we can look at many correlations at once, instead of needing to type out the command for correlation for each pair of variables.
    - This matrix also generalizes for p>1 variables.
* Correlation and Covariance in Python
  + In Python, calculating correlation and covariance matrices is pretty simple.
    - x = [1,2,3,4,5]
    - y = [6,5,4,3,2]
    - import numpy as np
    - print(np.cov(x,y) ) ## Prints the covariance matrix.
    - print(np.corrcoef(x,y)) ## Prints the correlation matrix.
* Other Correlation Measures
  + Throughout this lesson, we spent a lot of time learning about the Pearson's correlation. Although this is typically the most popular type of correlation to measure, there are other ways of calculating correlation too. Some additional methods include the Kendall rank correlation and Spearman's rank correlation.
  + Kendall Rank Correlation:
    - The Kendall approach is used to measure the ordinal association between two measured quantities. Specifically, it measures the strength of dependence between two variables. For example, you may want to determine if there is a correlation between two track runners finishing in 1st and 2nd place.
  + Spearman's Rank Correlation:
    - Spearman's correlation is used to measure the degrees of association between two variables. A question that might be asked with Spearman correlation would be, "Is there a statistically significant relationship between track runner age and how they place in a given race?"
* Real World: Correlation and Covariance
  + What did we learn?
    - Correlation is used in many real world contexts. Correlation can be used to compare the linear relationship between two pairs of variables, as correlation is unitless.
    - Correlation has more statistical relevance - but covariance is the basis for a number of statistical operations.
    - When calculating covariance, it is inadvisable to compare the relationships between two pairs of variables if the units are different.
* Properties of Correlation and Covariance
  + There are some additional properties of correlation and covariance that we haven't discussed yet. These include:
    - 
    - These facts can be helpful in a number of practical and statistical contexts, which we'll cover later!
* Recap
  + Both correlation and covariance measure the strength and direction of linear relationships between two variables.
  + Values near zero indicate little to no linear relationship between the two variables, values greater than zero indicate a positive linear relationship, and values less than zero indicate a negative linear relationship.
  + Correlation is more commonly used to compare two pairs of variables, whereas covariance is more important in a statistical context.
  + The NumPy functions np.cov() and np.corrcoef() take in a series of quantitative variables and output covariance and correlation matrices, respectively.