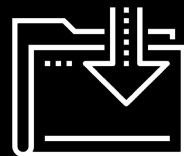




Introduction to Statistics

Data Boot Camp

Lesson 5.4



Class Objectives

By the end of today's class, you will be able to:



Calculate summary statistics such as mean, median, mode, variance and standard deviation using Python.



Plot, characterize, and quantify a normally distributed dataset using Python.



Qualitatively and quantitatively identify potential outliers in a dataset.



Differentiate between a sample and a population in regards to a dataset.



Define and quantify correlation between two factors.



Calculate and plot a linear regression in Python.

Don't worry! Class will not be painful.

... but there is a lot to do
And this will help you with your homework!



We Will Build on Concepts You Already Know





What is a measure of central tendency?

Measure of Central Tendency = Center of a Dataset

Three most common measures are **mean**, **median**, and **mode**.



Mean is the sum of all values divided by the number of elements in a dataset.



Median is the middle value in a sorted dataset.



Mode is the most frequently occurring value(s) in a dataset.

Measures of Central Tendency in Python

Two packages to remember when calculating statistics are **NumPy** and **SciPy**.



Mean is calculated using **NumPy**.



Median is calculated using **NumPy**.



Mode is calculated using **SciPy**.



When new data comes along,
you must plot it!

Why Plot Data?

01

To determine if the data is normally distributed.

02

To determine if the data is multimodal.

03

To characterize clusters in the dataset.

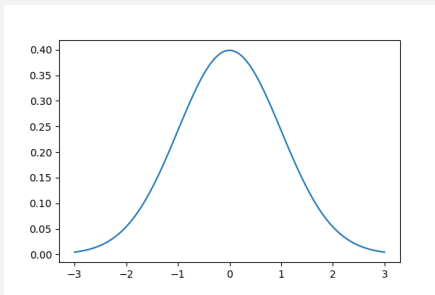
What Is Normally Distributed Data?

01

Measurements in a dataset are obtained independent of one another.

02

The distribution of data follows a bell curve shape.



03

We can quantitatively test if a dataset is normal using SciPy.

```
stats.normaltest()
```



What are **variance** and
standard deviation?

Variance & Standard Deviation Describe Variability of Data



Variance is the measurement of how far each value is away from the mean of the dataset.

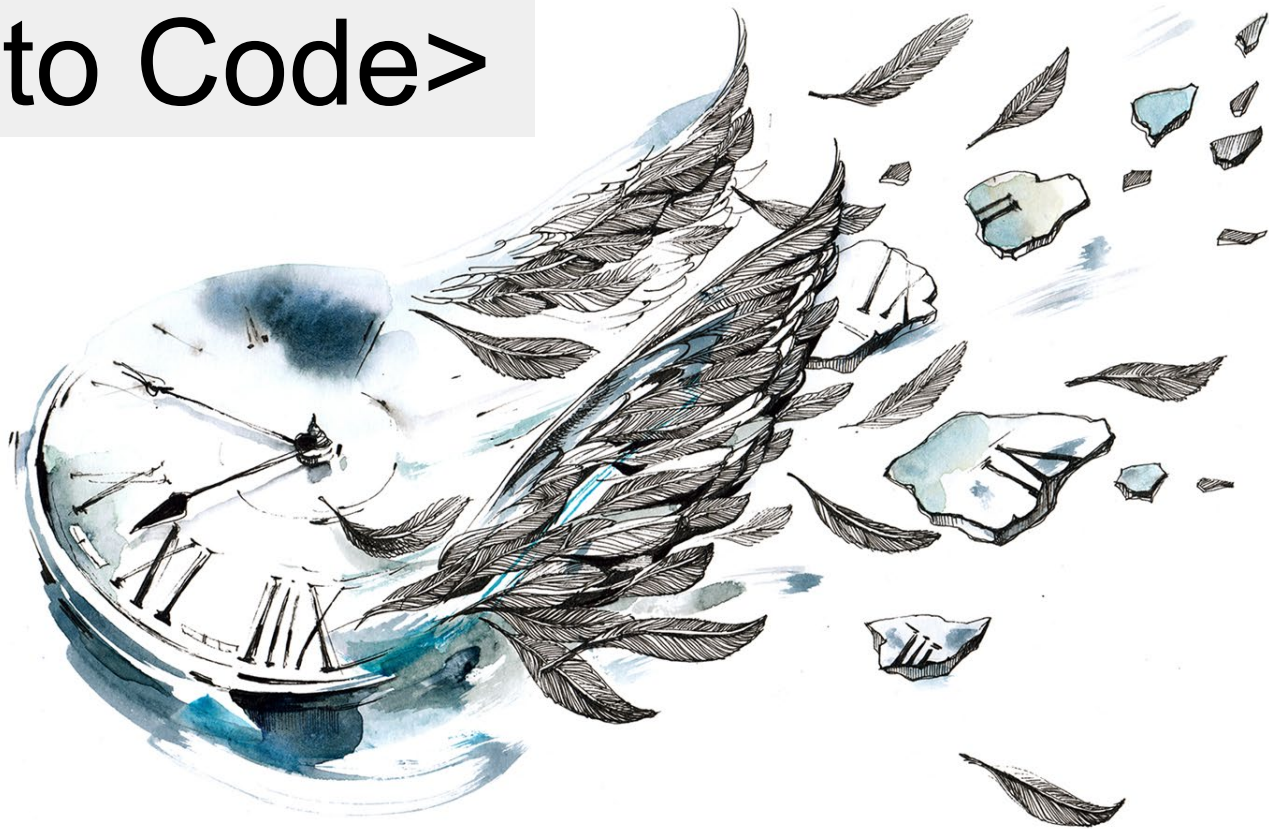


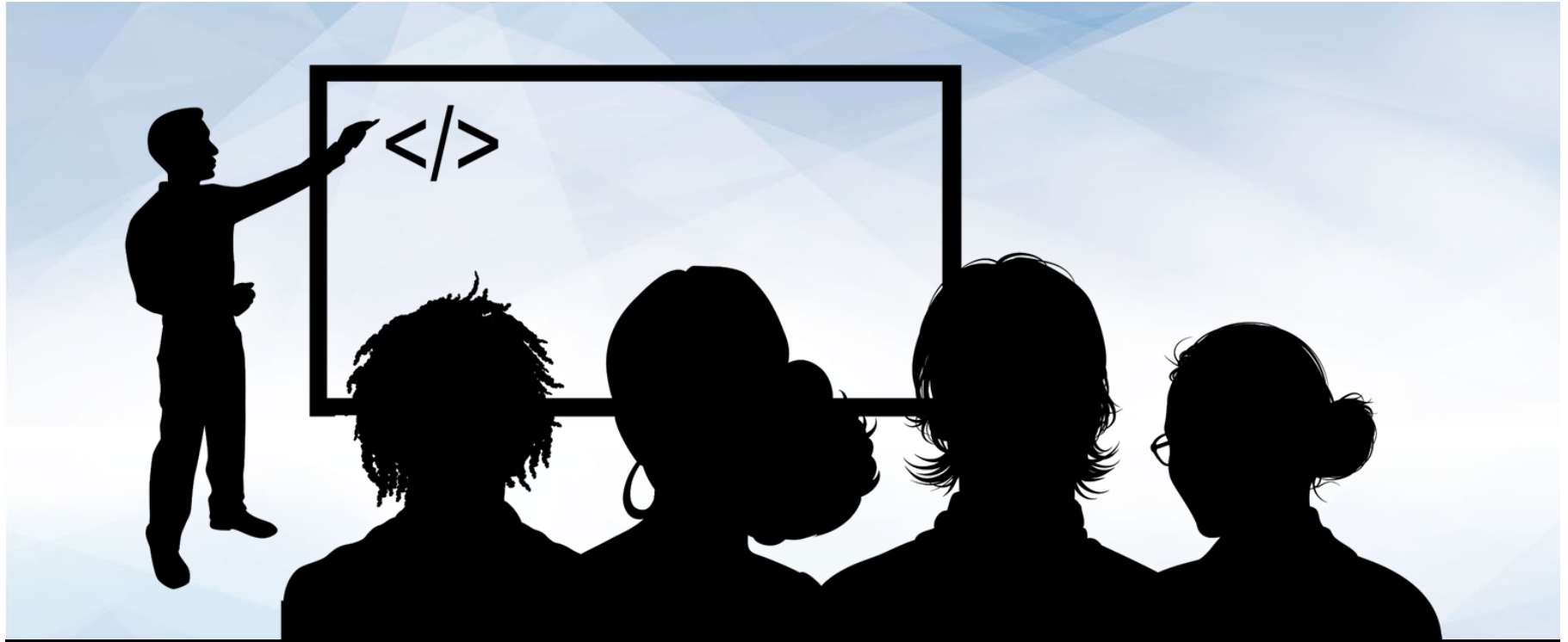
Standard deviation is the square root of variance.



In Python, both variance and standard deviation are calculated using the NumPy module.

<Time to Code>





Instructor Demonstration

Quantiles, Quartiles, and Outliers



What are quantiles, quartiles,
and outliers ?

Quantiles, Quartiles, and Outliers Describe a Dataset

01

Quantiles divide data into well-defined regions based on a sorted dataset.

02

Quartiles are a specific type of quantile where a sorted dataset is split into four equal parts.

Q1: 25% of the data

Q2: 50% of the data

Q3: 75% of the data

03

Outliers are an extreme value in a dataset that can skew calculations and results.

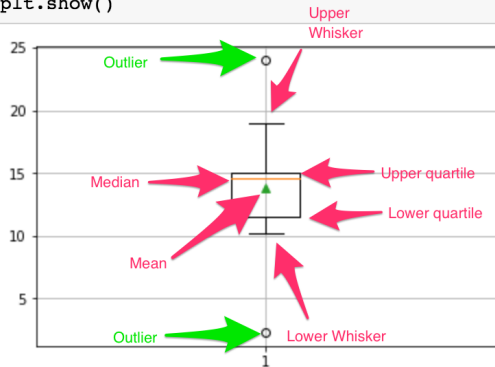
How to Identify Potential Outliers

01

Qualitatively

Use **box and whisker plots** to visually identify potential outlier data points.

```
# Create box plot
plt.boxplot(arr, showmeans=True)
plt.grid()
plt.show()
```



02

Quantitatively

Determine the outlier boundaries in a dataset using the **“1.5 IQR” rule**.

- IQR is the interquartile range, or the range between the 1st and 3rd quartiles.
- Anything **below** $Q1 - 1.5 \text{ IQR}$ could be an outlier.
- Anything **above** $Q3 + 1.5 \text{ IQR}$ could be an outlier.

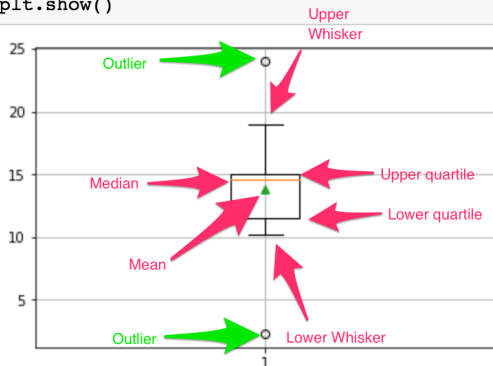
How to Identify Potential Outliers in Python

01

Qualitatively

Use Matplotlib's `pyplot.boxplot` function to plot the box and whisker.

```
# Create box plot
plt.boxplot(arr, showmeans=True)
plt.grid()
plt.show()
```

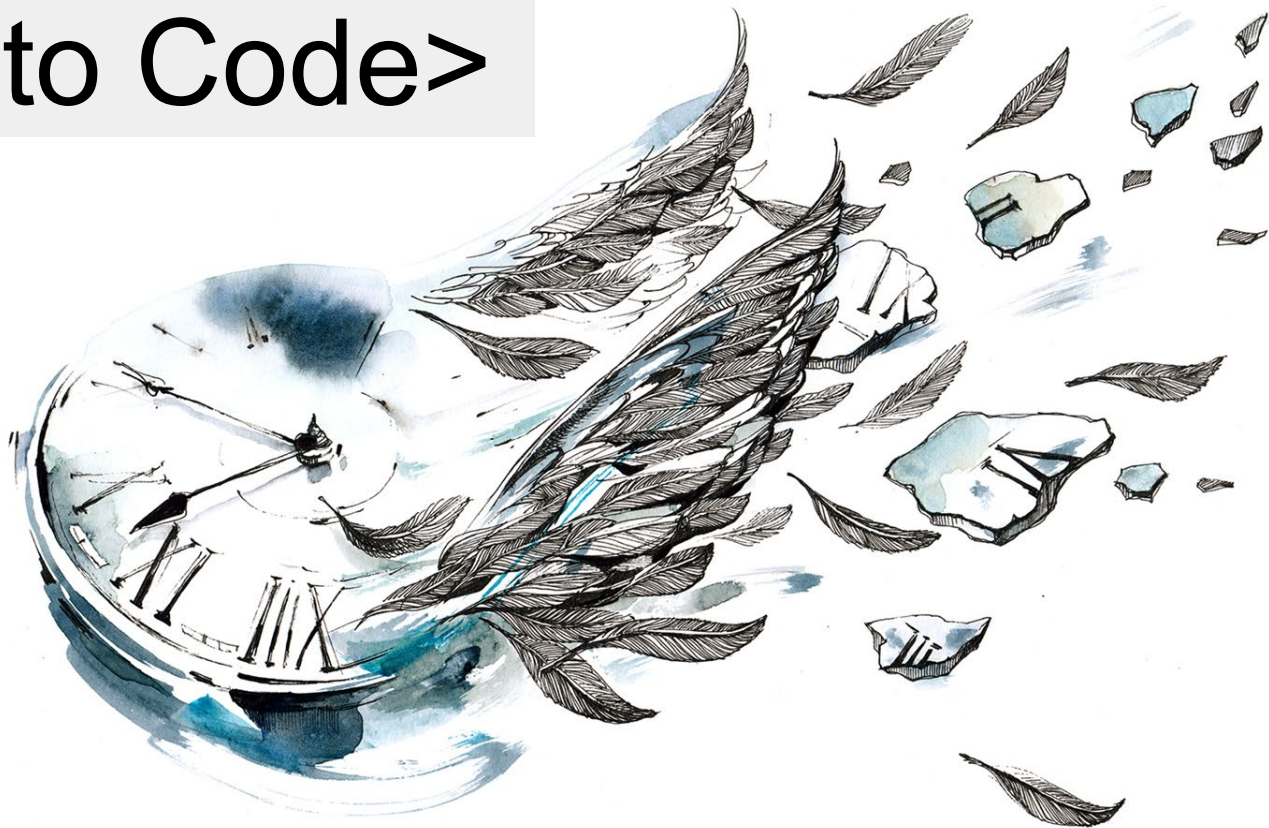


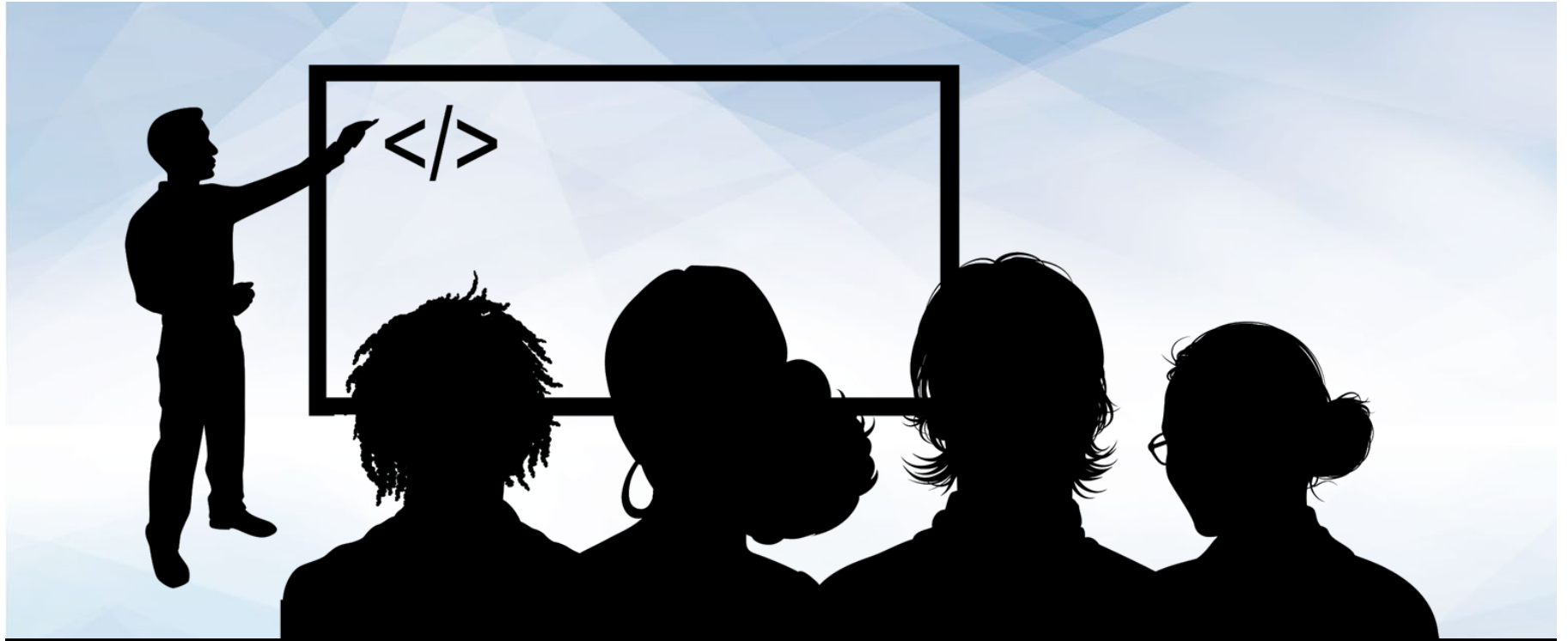
02

Quantitatively

- Use Pandas' `series.quantile` function to calculate the quantile.
- Calculate the outlier boundaries.


<Time to Code>





Instructor Demonstration

Sample, Population, and SEM



Let's think about
the following
scenario...

Predicting the City Election

Weeks before Election Day, a local newspaper wants to predict the winner of the mayoral election. The newspaper will poll voters for their intended candidate. Consider the following:

- It would be prohibitively expensive to poll all voters.
- It is logistically impossible to know who will actually go out to vote on Election Day.
- Therefore, the newspaper must predict the outcome of the election using data from a *subset* of the population.



This is a **sample dataset**
versus a **population dataset** .

Population Dataset vs. Sample Dataset

01

Population Dataset

- Dataset containing all possible elements of an experiment or study.
- In statistics, “population” does not mean “people.”
- Any complete set of data is a population dataset.

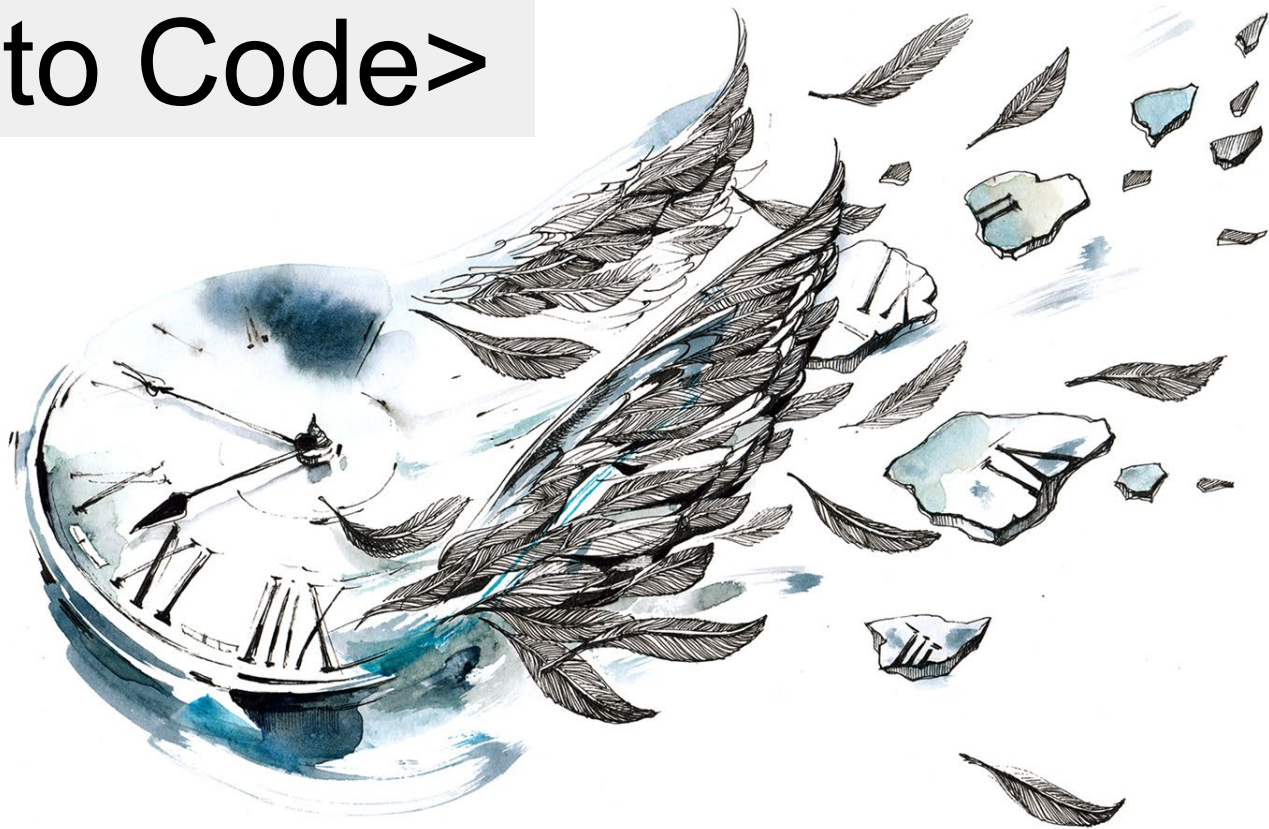
02

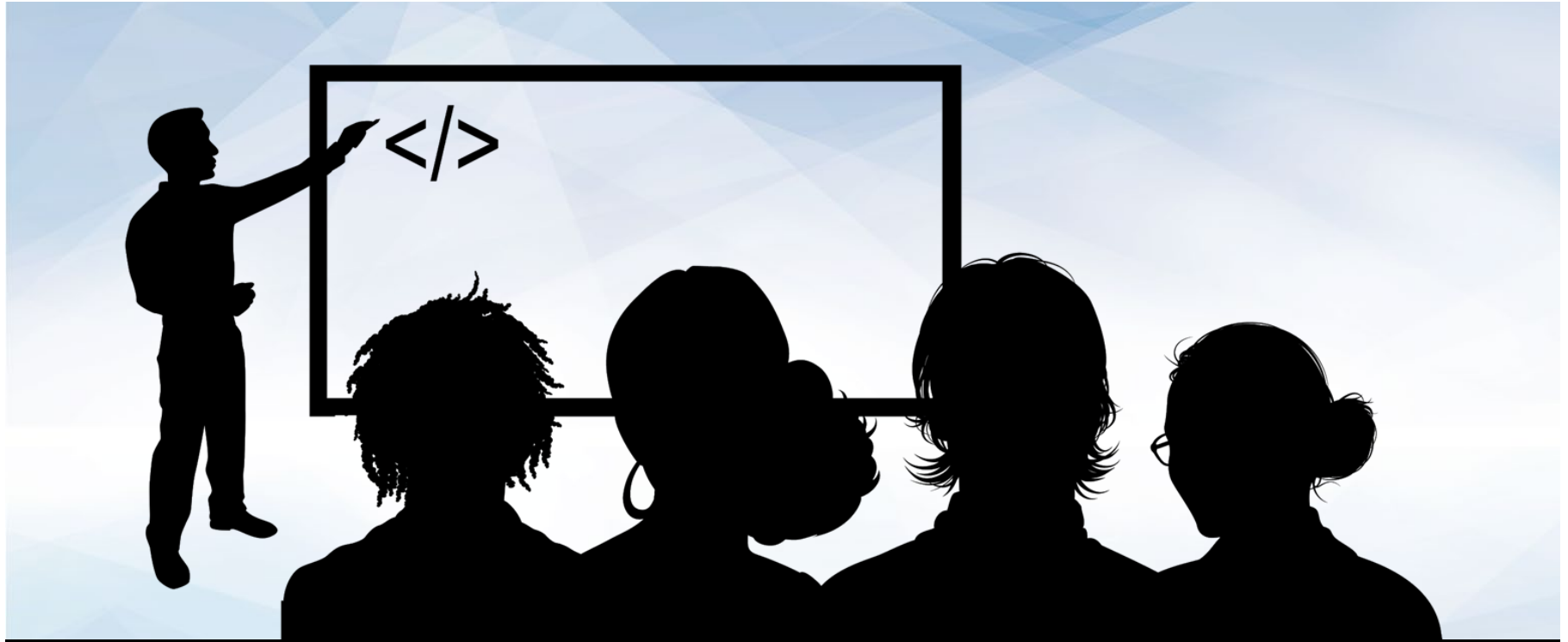
Sample Dataset

- A subset of population data.
- A sample dataset can be selected randomly from the population or selected with bias.

- In statistics, a population is a complete dataset that contains all possible elements of a study or experiment. In this scenario, the population dataset would be the voting habits of all eligible voters in the city.
- In statistics, a sample is a subset of a population dataset, where not all elements of a study or experiment are collected or measured. In this scenario, the sample dataset is the 1,000 eligible voters polled across the city.
- In data science, the concept of sample versus population does not strictly apply to people and/or animals. Any comprehensive dataset is considered a population, and any dataset that is a subset of a larger data set is considered a sample.

<Time to Code>



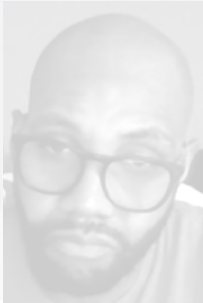


Instructor Demonstration

Correlation Conundrum

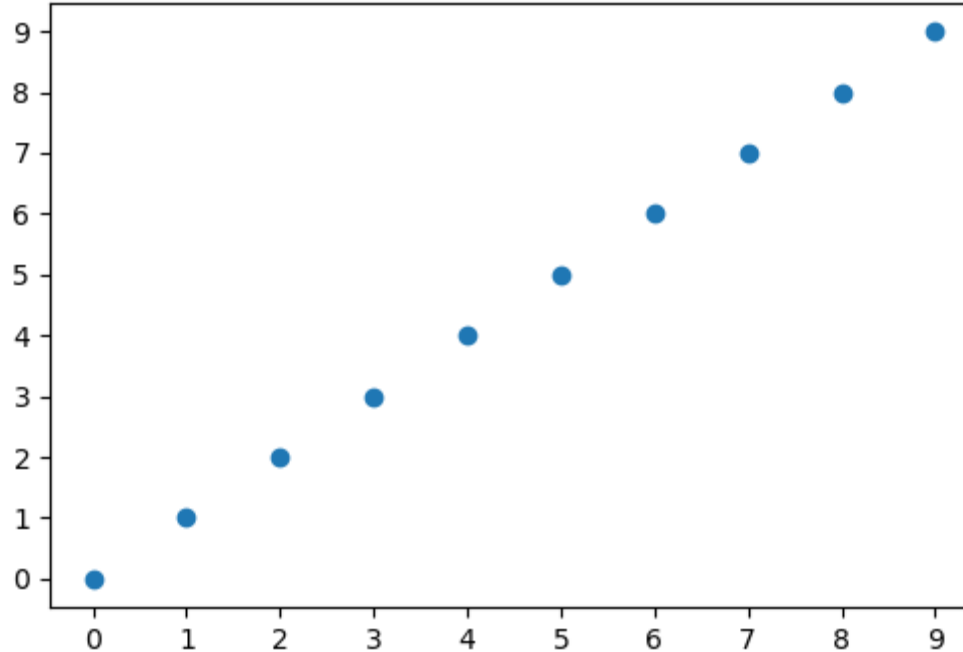


Correlation describes the question, “Is there a relationship between A and B?”



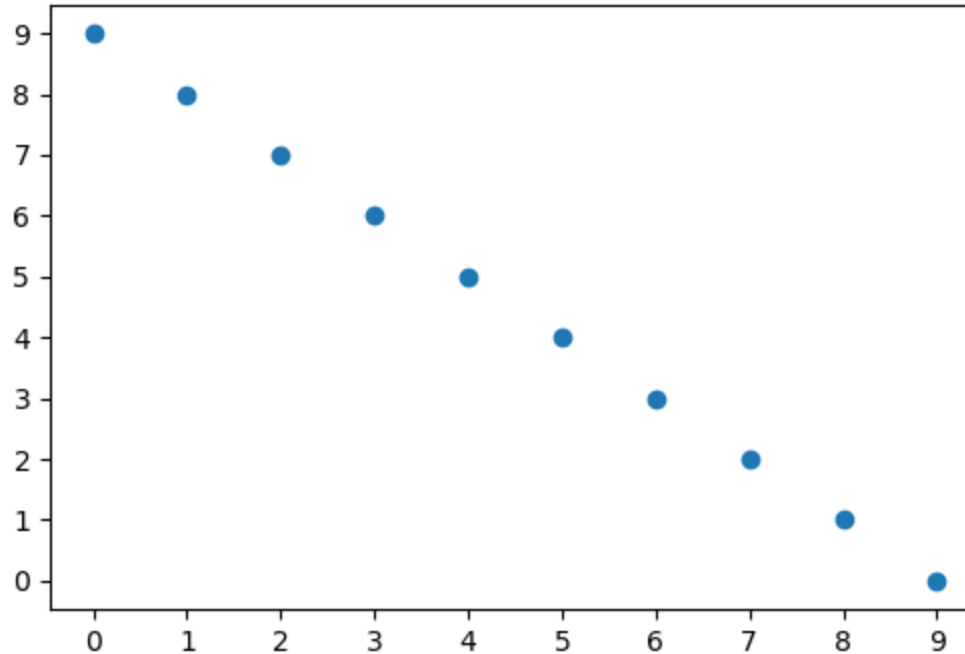
Often in data analysis we will ask the question “Is there a relationship between Factor A and Factor B?” This concept is known in statistics as correlation.

Positive Correlation



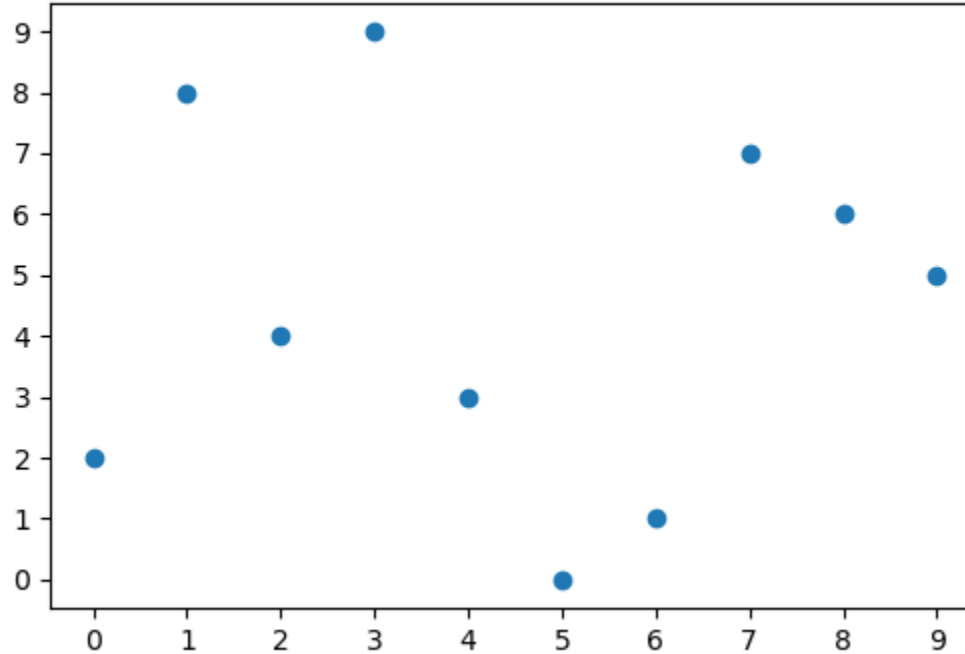
- This is an example of a positive correlation. When two factors are positively correlated, they move in the same direction. When the factor on the x-axis increases, the factor on the y-axis increases as well.

Negative Correlation



- This is an example of a negative correlation. When two factors are negatively correlated, they move in opposite directions. When the factor on the x-axis increases, the factor on the y-axis decreases.

No Correlation

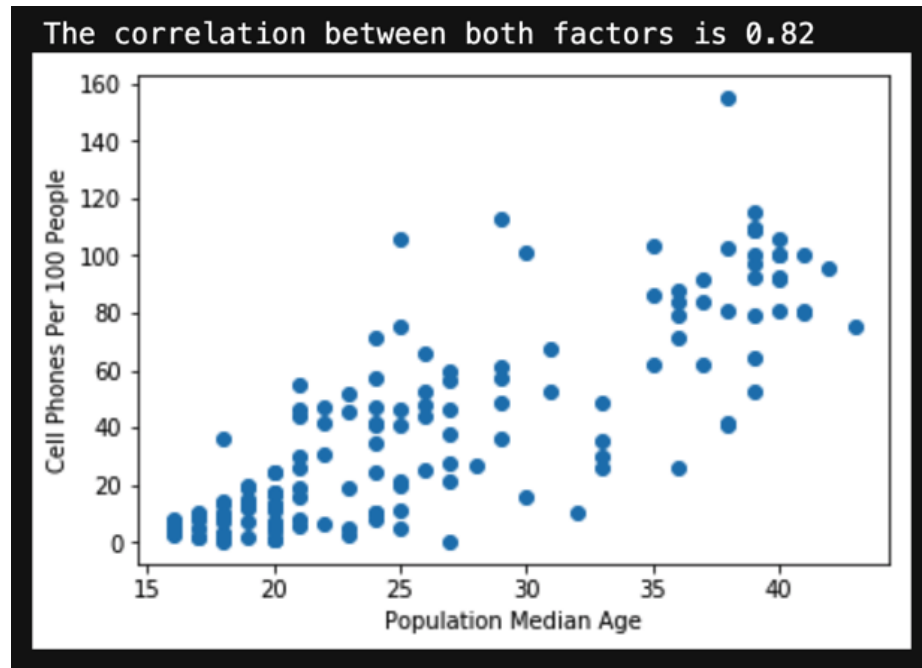


- This is an example of two factors with no correlation. When two factors are not correlated, their values are completely independent between one another.
- With real-world data, it can be difficult to determine if two factors are correlated. Therefore, we must be able to quantify the correlation between two factors.

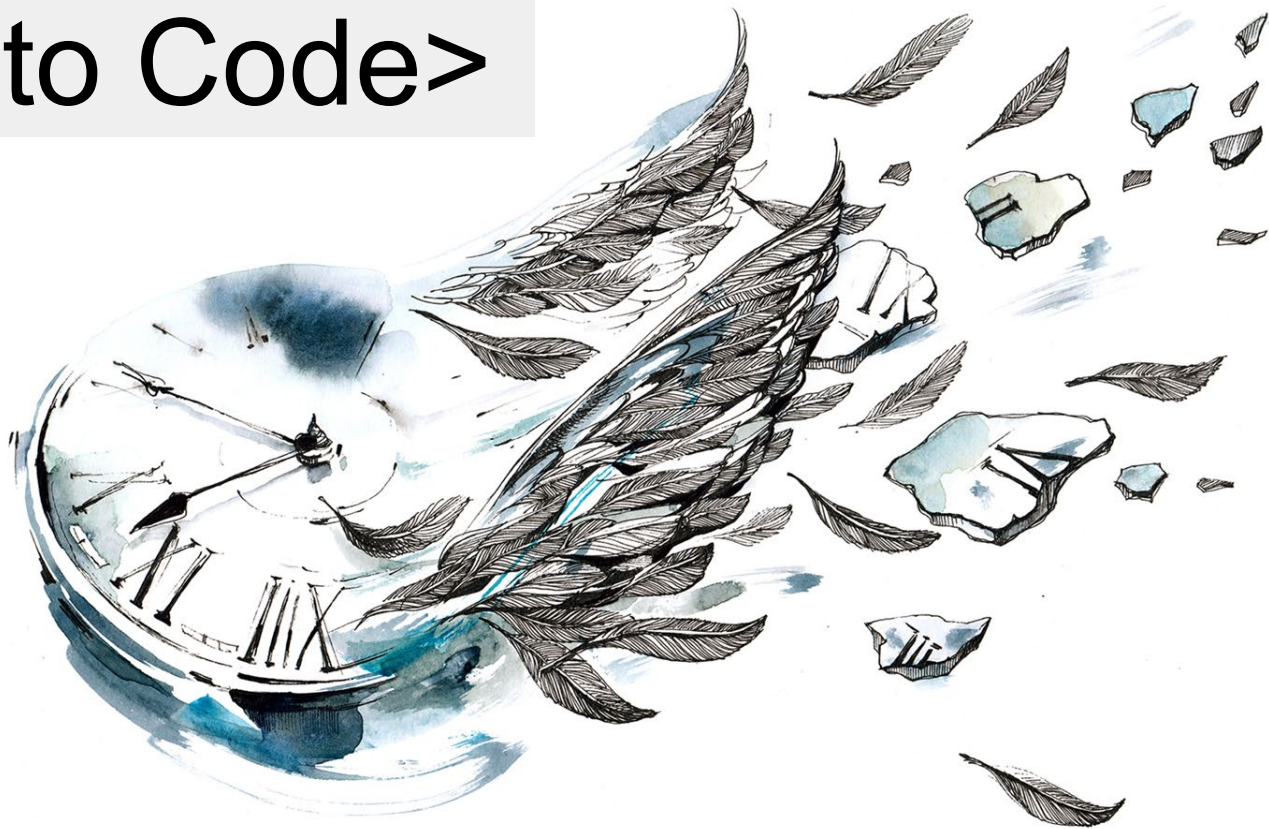
Pearson's Correlation Coefficient

In statistics, we quantify correlation using Pearson's r .

- Pearson's correlation coefficient describes the variability between two factors, denoted by the variable r .
- Pearson's r is $-1 \leq r \leq 1$
 - -1 indicates perfect negative correlation.
 - 1 indicates perfect positive correlation.
 - 0 indicates no correlation.
- Real-world data is never perfect.

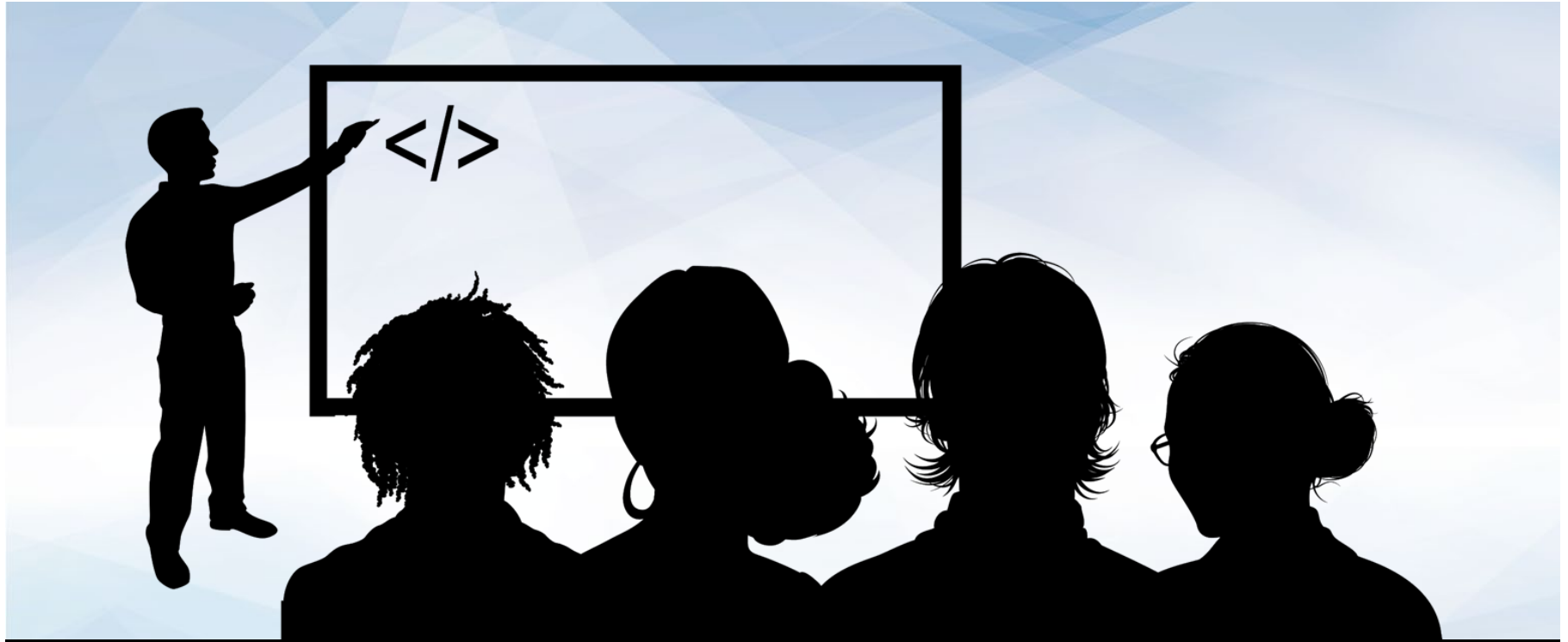


<Time to Code>





Time's Up! Let's Review.



Instructor Demonstration

Fits and Regression



What is the equation of a line?



The equation of a line is:

$$y = mx + b$$

The Equation of a Line

- The equation of a line defines the relationship between x-values and y-values.
- When it comes to variables in the equation, we refer to the `x` in the equation as the independent variable, and the `y` as the dependent variable.
- Additionally, the slope of a line is denoted as `m` in the equation and the y-intercept is denoted as `b` in the equation.
- Knowing the slope and y-intercept of a line, we can determine any value of `y` given the value for `x`. This is why we say `y` is dependent on `x`.

$$y = mx + b$$

The diagram shows the equation $y = mx + b$ in a large, black, serif font. Below the equation, there are three red arrows pointing upwards to specific parts of the equation. The first arrow points to the y and is labeled "Dependent variable" below it. The second arrow points to the m and is labeled "Slope" below it. The third arrow points to the b and is labeled "y-intercept" below it. Additionally, there are two red arrows pointing upwards to the x and the $+$ sign, with the label "Independent variable" centered below them.

Dependent variable

Slope

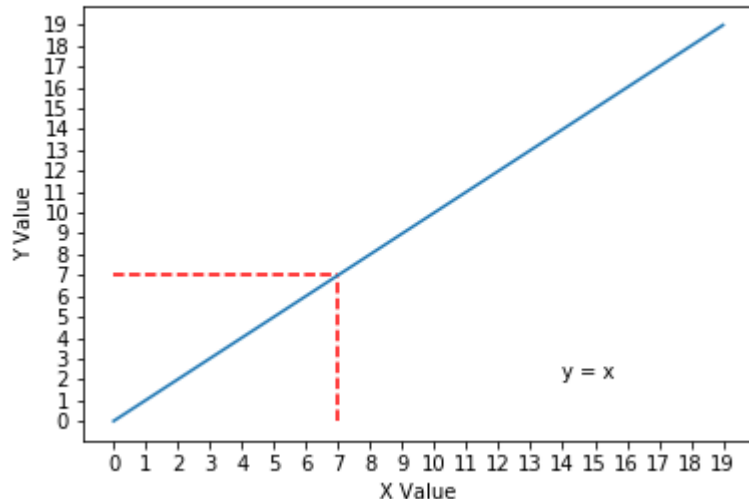
Independent variable

y-intercept

The equation of a line determines y values given x

In this example:

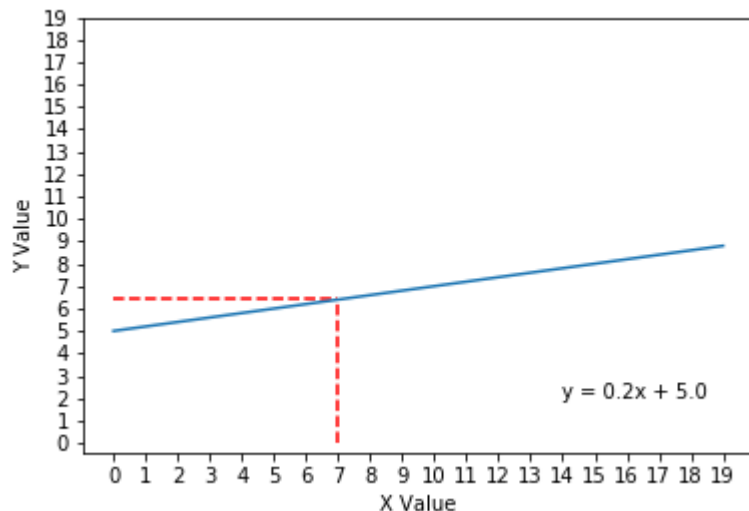
- Slope = 1
 - y -intercept = 0
 - Whatever x is, the value of y is the same.
- This first plot is considered the ideal linear relationship of y and x , where the x and y values are the same value.
 - In this plot, the equation for line is $y = x$ because the slope is equal to 1 and the y -intercept is equal to 0.
 - If we look at the x -value of 7 (denoted by the vertical dashed line), the corresponding y -value is also 7 (denoted by the horizontal dashed line).



The Equation of a Line Determines y Values, Given x

In this example:

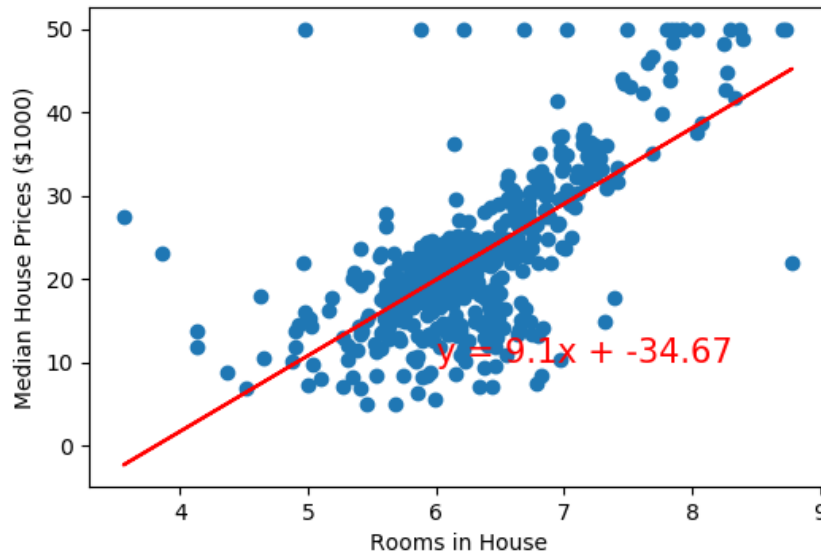
- Slope = 0.2
 - y -intercept = +5
 - If $x = 7$, then $y = 6.4$
- In this linear relationship between x and y , the slope is much smaller, but the y -intercept is much larger.
 - If you plug an x -value of 7 into the equation, the resulting y -value is 6.4. Demonstrate this to the class.
 - This idea of relating x -values and y -values using the equation of a line is the general concept of linear regression.



Linear Regression Fits the Equation of a Line to Real World Data

Linear regression...

- Predicts the values of factor B, given values from factor A.
- Estimates where data points that were not measured might end up if more data was collected.
- Is used to predict housing prices, stock market, weather, etc.



- Linear regression is used in data science to model and predict the relationship between two factors.
- Although this may sound similar to correlation, there is a big difference between the two concepts: correlation quantifies if Factor Y and Factor X are related, while regression predicts Factor Y values given values from Factor X.
- By fitting the relationship of two factors to a linear equation, linear regression allows us to predict where data points we did not measure might end up if we had collected more data.
- Linear regression is a truly powerful tool; it provides us the means to predict house prices, stock market movements, and the weather based on other data.
- We will not dive into the mathematical details of linear regression—rather, we will focus on how to use SciPy's [linregress function](#) to perform a linear regression, and visualize the linear regression using Matplotlib.

<Time to Code>

