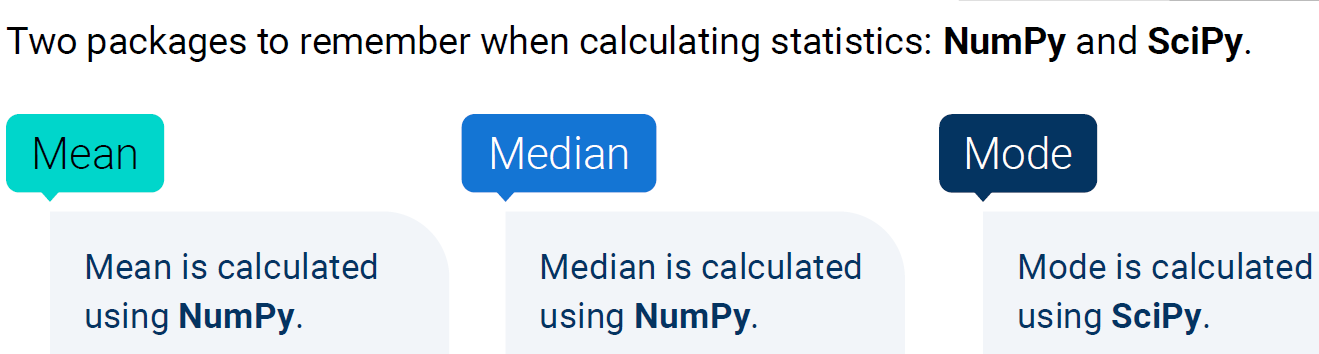
**10-16 notes**

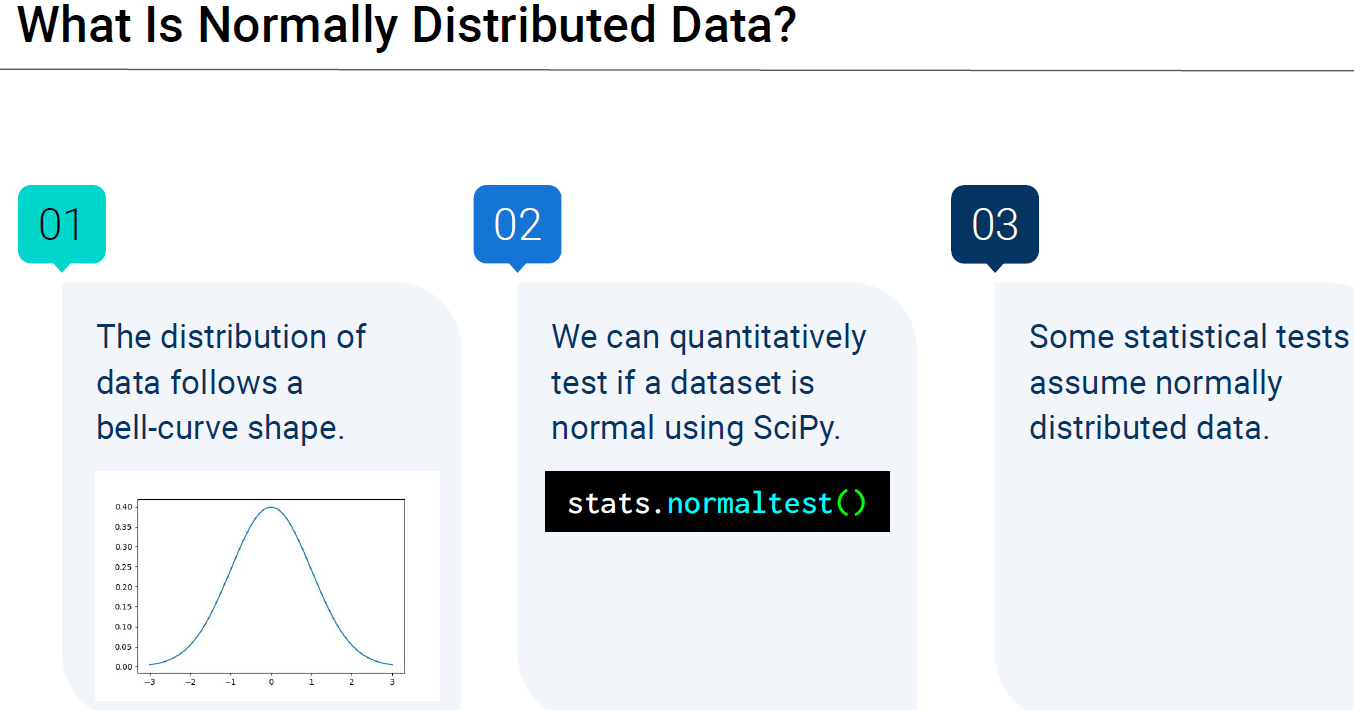
Introduction to Statistics

Goals:

By the end of this lesson, you will be able to:

* Calculate summary statistics such as mean, median, mode, variance, and standard deviation, by using Python.
* Plot, characterize, and quantify a normally distributed dataset with Python.
* Qualitatively and quantitatively identify potential outliers in a dataset.
* Differentiate between a sample and a population with respect to a dataset.
* Define and quantify correlation between two factors.
* Calculate and plot a linear regression with Python.





Normal data has a single bell.

Multi-modal data has multiple hills

Bi-modal data has 2 hills

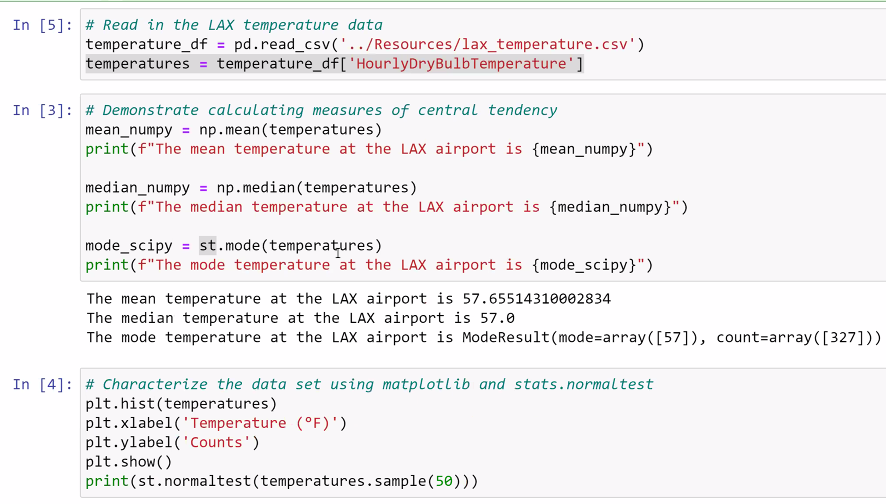
Skewed data has a hill with 2 different slope rates (one is sharper/shallower than the other)

* Hill leans to the side

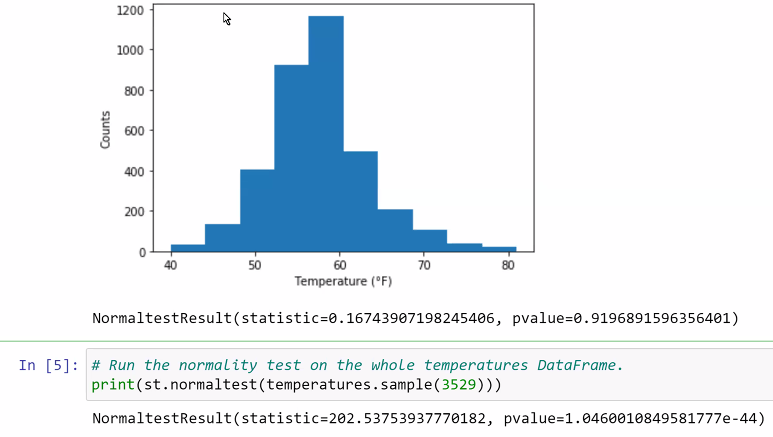
If you need to install something in Jupyter notebook, like matplotlib, you can type ! pip install matplotlib

* Exclamation point tells Jupyter notebook that you’re doing a command

**Samples solution:**



^ Mean, Median, and Mode are close together, so the data is fairly normal.



As sample size increases, data has a tendency to start to become more and more normal.

This data is not a good example of this.

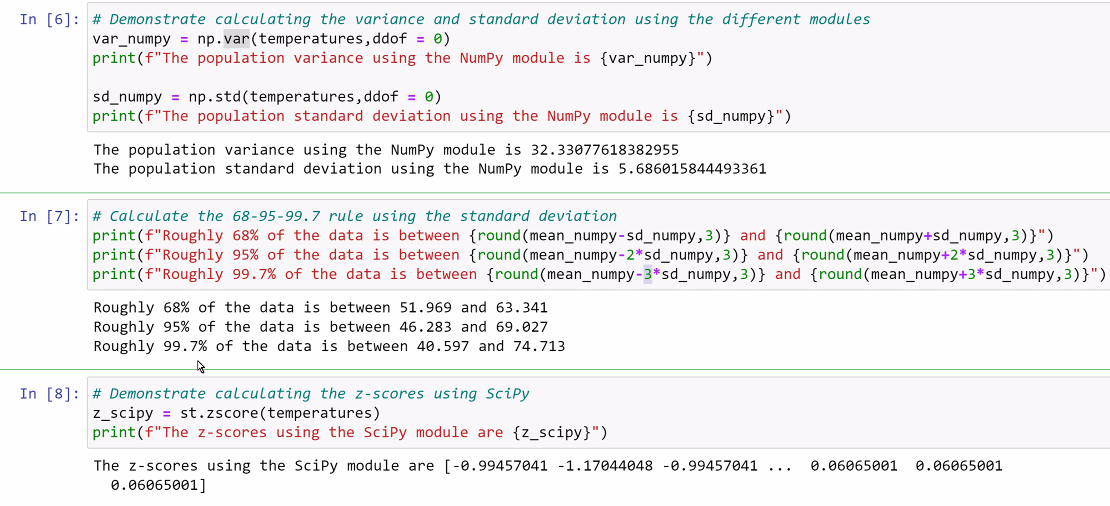
Graph above is based on sample of 50, but a p-value of 0.9197.

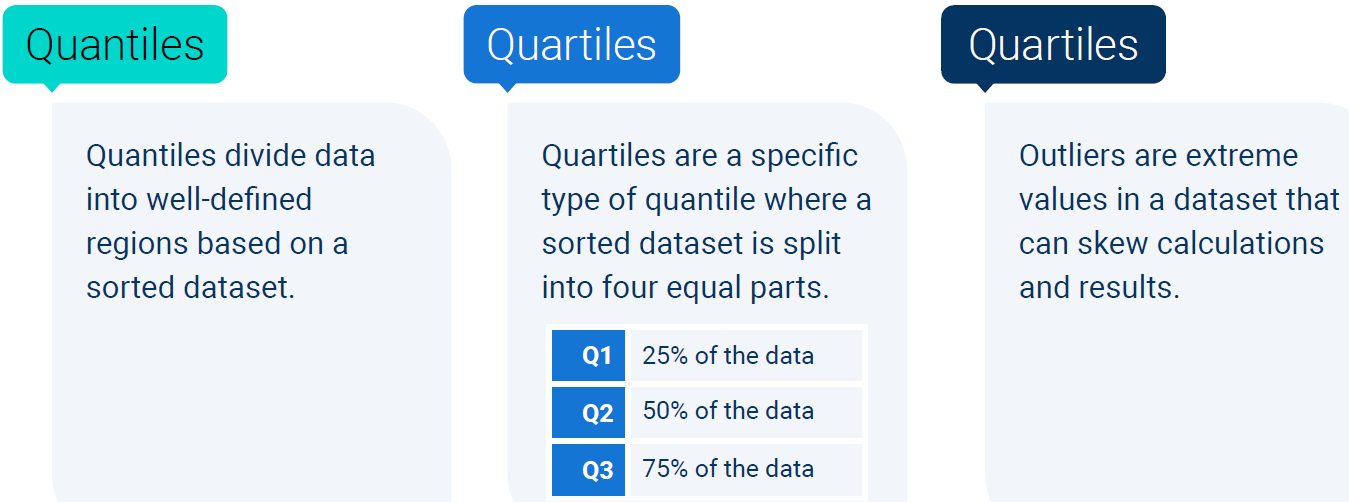
For the whole data set (3529 records) the p-value is 0.000…000104

(1.046 times 10 to the negative 44th power)

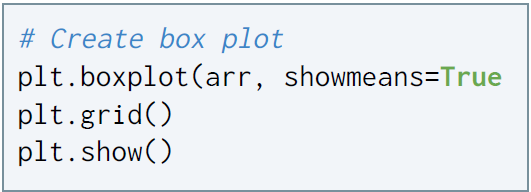
Small p-value = bad

Larger p-value = good

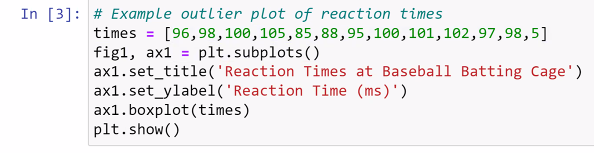


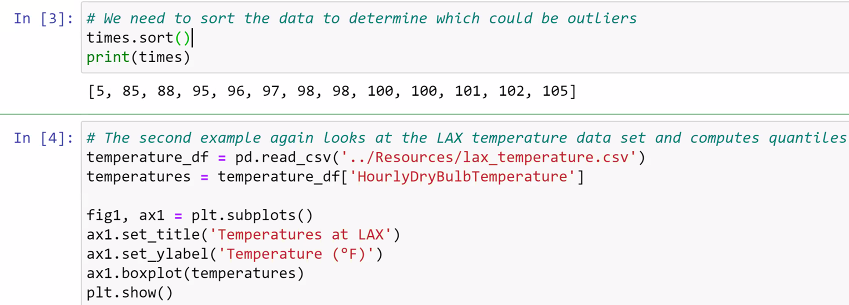


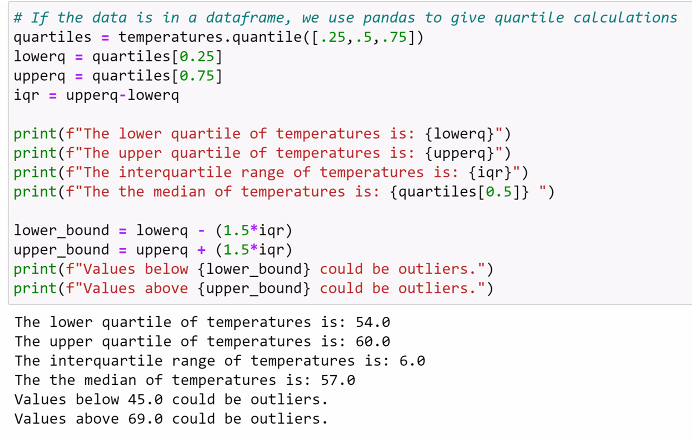
Exploratory data analytics: Is my data good enough to do any more in depth analysis?



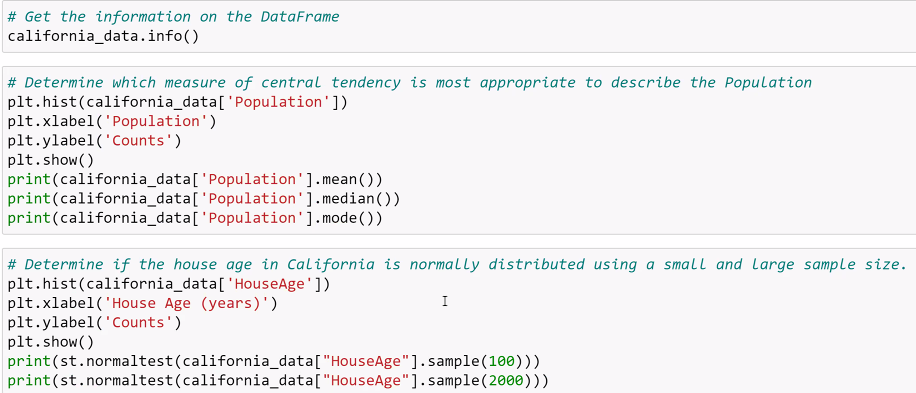
**Samples solution:**

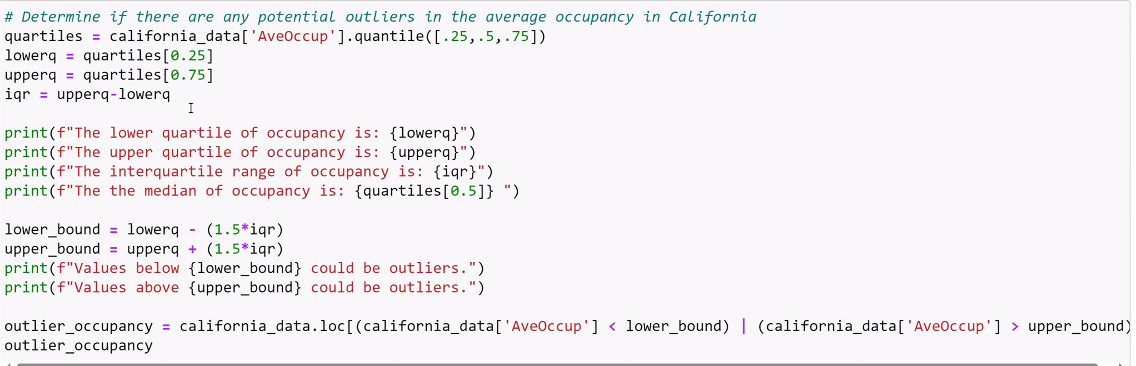




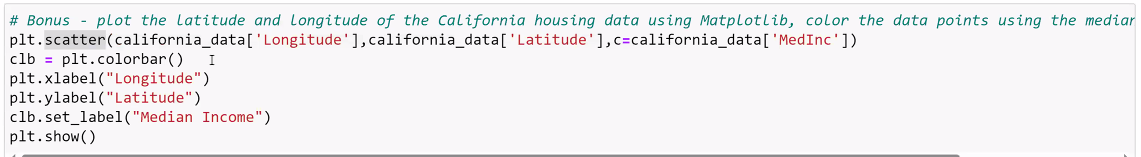


**Summary Stats solution:**



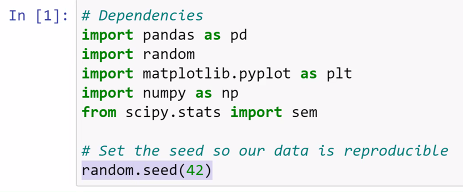


| <--- use this vertical line for “OR” It’s purple in the screenshot above

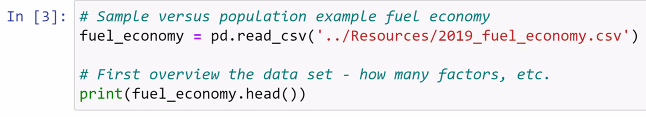


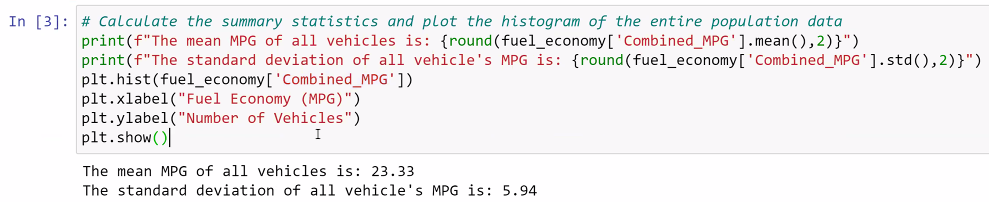
SEM means Standard Error of Mean

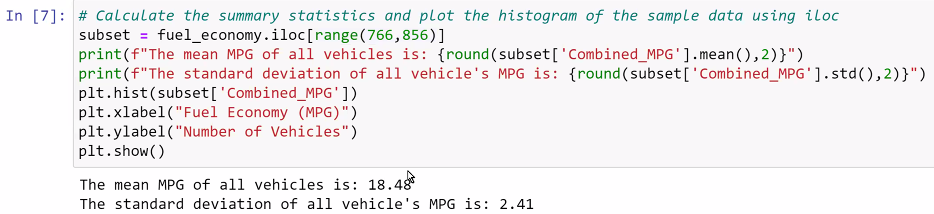
**Standard Errors solution:**



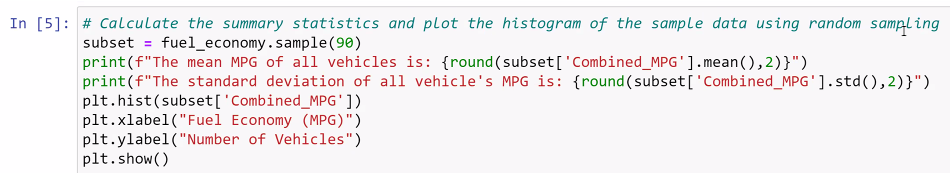
^ Set your seed for the random numbers. If not, you’ll get a different random set of numbers every time.



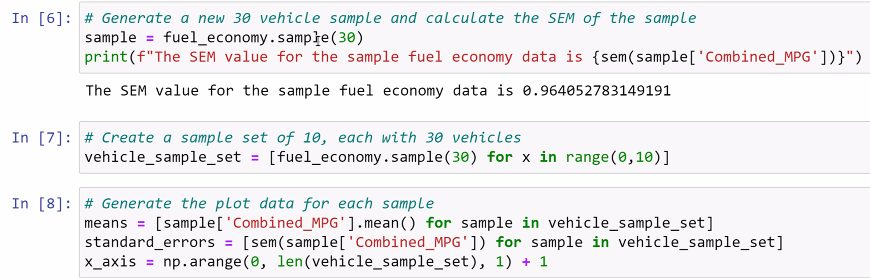




^ sample is a range of values based on iloc

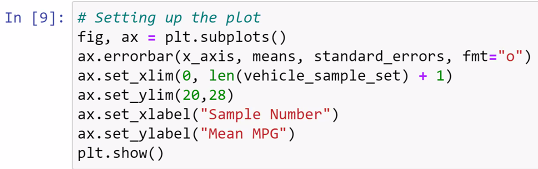


^ sample is random

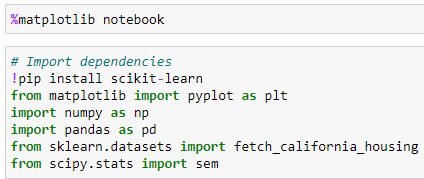


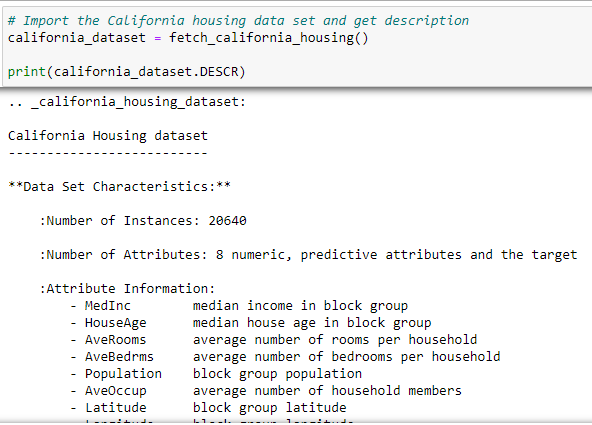
Small SEM is good, large SEM is bad. It shows how close the sample mean is to the population mean.

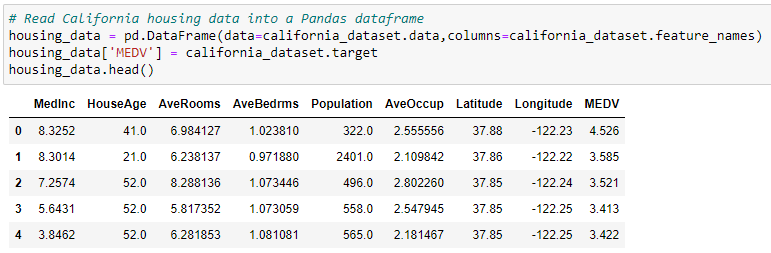
IN [7] generates 10 lists of 30 vehicles each



**Samples solution: Review me 8:30-9pm**

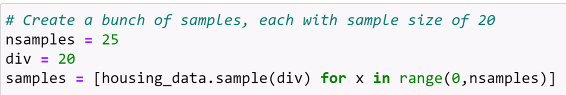


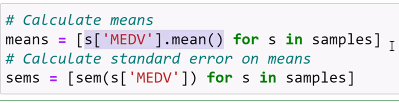




Independent variables are all columns except MEDV

MEDV is a DEPENDANT variable. In machine learning, you teach a machine with data that already exists, and predict dependent variables.

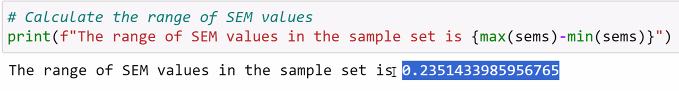


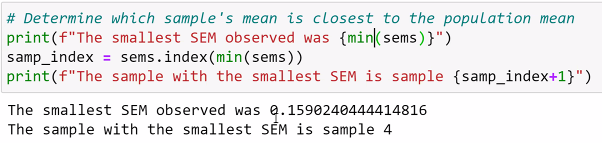


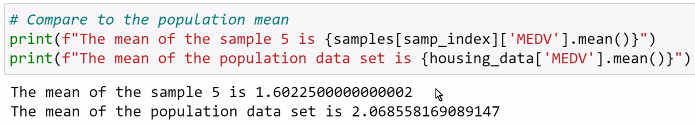


Errorbar function above is a specific graph. It NEEDS defined limits.

* PROFESSIONALLY, THIS IS USED ALL THE TIME





****

**About Correlation**

The Pearson correlation coefficient describes the variability between two factors, denoted by the variable *r*.

Pearson’s r is: **–1 ≤ r ≤ 1** always between -1 and 1

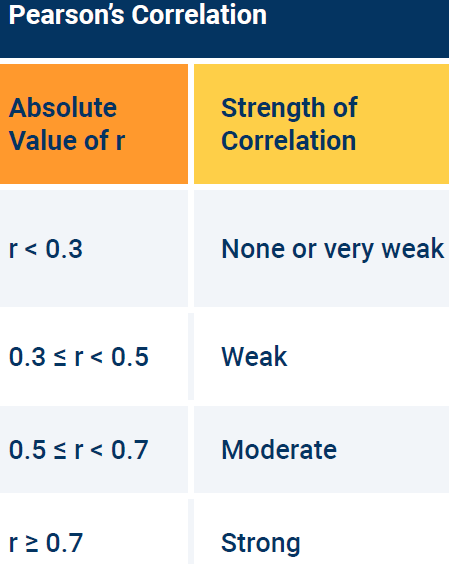
Positive correlation means line slopes ***up*** as you go left to right

Negative correlation means line slopes ***down*** as you go left to right

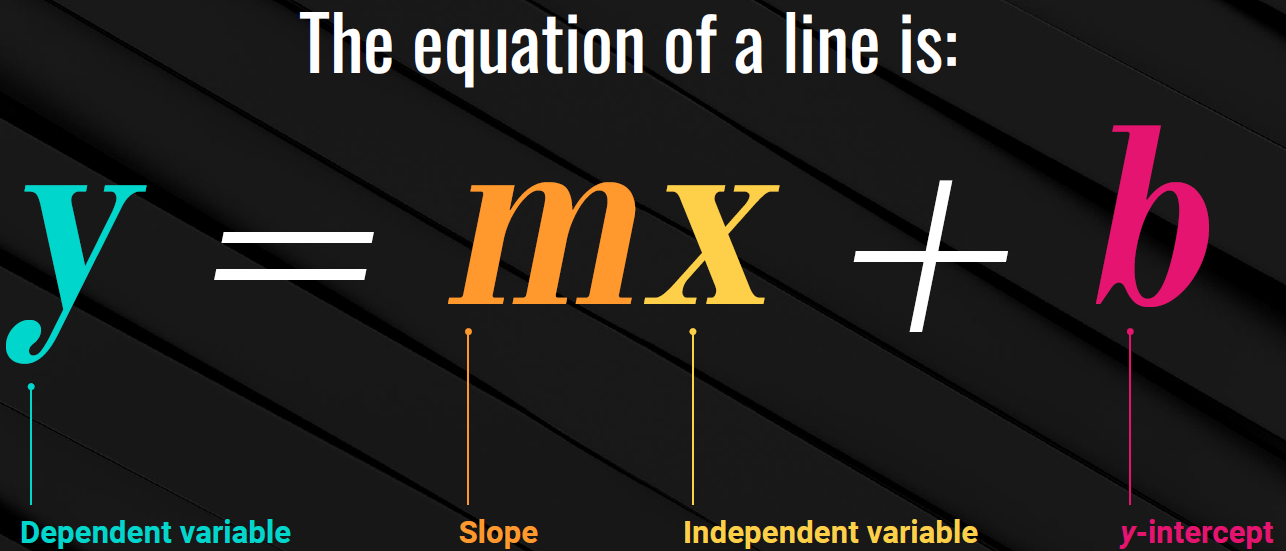


Carson’s wisdom: Malic acid makes things sour

* Harmless prank: a few drops of malic acid in Jello while you’re making it



**Linear Regression notes**



Also have y=mx+b+c

C is the error term 6:35pm in recording

Linear regression is a straight line through a scatter plot that can be used to predict where a new value may land on the y axis based on any value of x, or vice versa.



