Data Science

General Assembly Lecture 2

Instructor: Hamed Hasheminia

STATISTICS FUNDAMENTALS

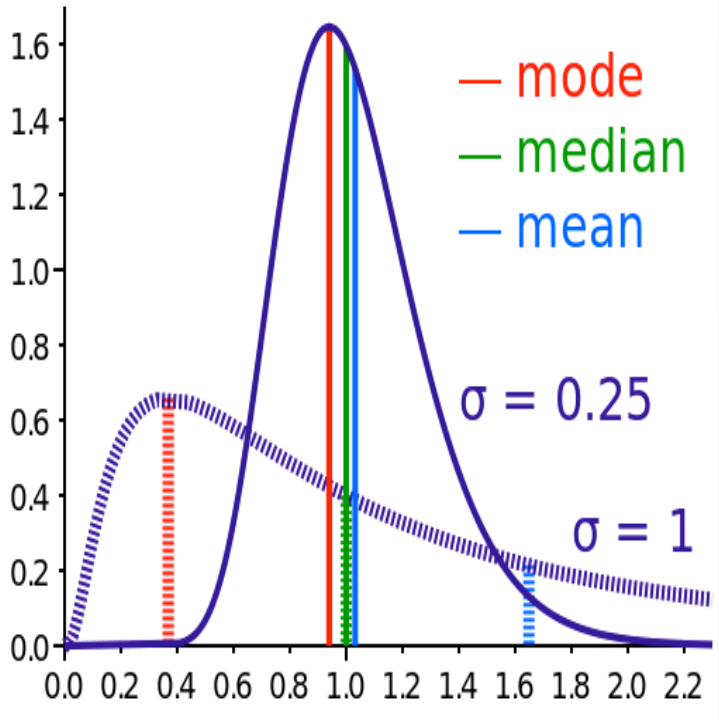
LEARNING OBJECTIVES

- Data Manipulation. Adding/removing columns and adding/removing observations
- Create data visualizations - including: boxplots, histograms- and scatterplots to discern characteristics and trends in a dataset
- Use NumPy and Pandas libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation, skewness coefficient
- Outliers
- Central Limit Theorem
- ID variable types and creating dummy variables

MEAN

- The mean of a set of values is the sum of the values divided by the number of values. It is also called the average.

$$\overline{X} = \frac{\sum X}{N}$$



MEAN EXAMPLE

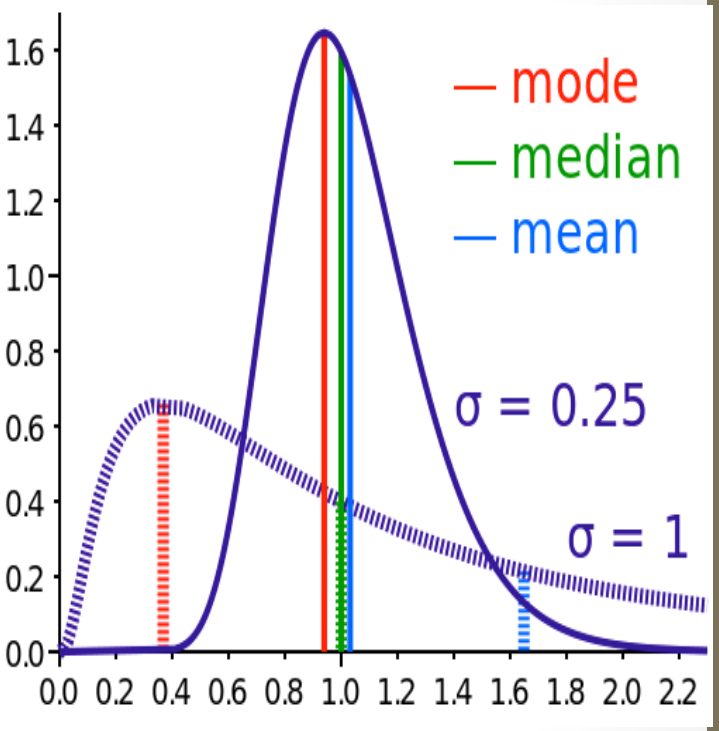
- Find the mean of 19, 13, 15, 25, and 18.

$$\frac{19 + 13 + 15 + 25 + 18}{5} = \frac{90}{5} = 18$$

MEDIAN

- The median refers to the midpoint in a series of numbers.

- To find the median
  - Arrange the numbers in order smallest to largest.
  - If there is an odd number of values, the middle value is the median.
  - If there is an even number of values, the average of the middle two values is the median.



MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

Ordered Values:

15, 19, 20, 29, 36

20 is the median

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

Ordered Values:

28, 37, 67, 75, 81, 92

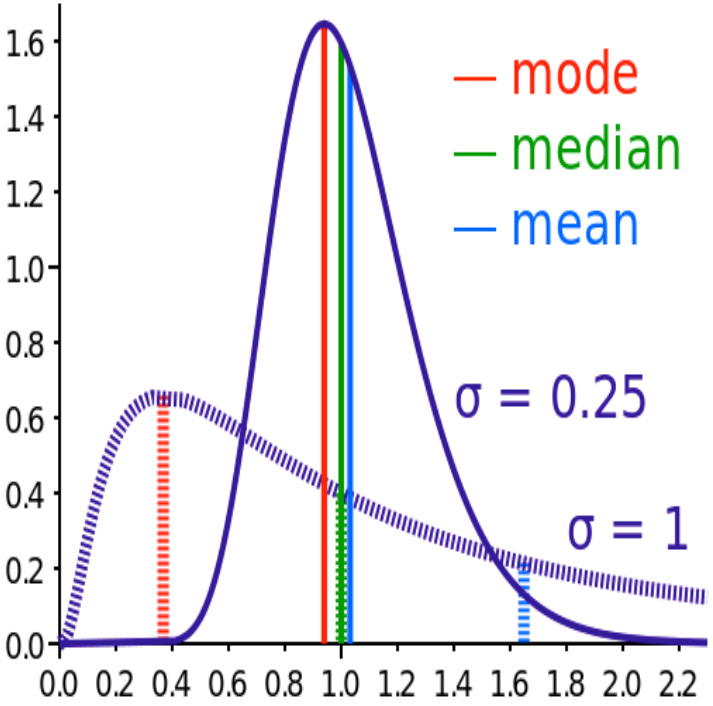
67 and 75 are the middle values.

$$\frac{67 + 75}{2} = \frac{142}{2} = 71$$

71 is the median.

MODE

- The mode of a set of values is the value that occurs most often.
- A set of values may have more than one mode or no mode.

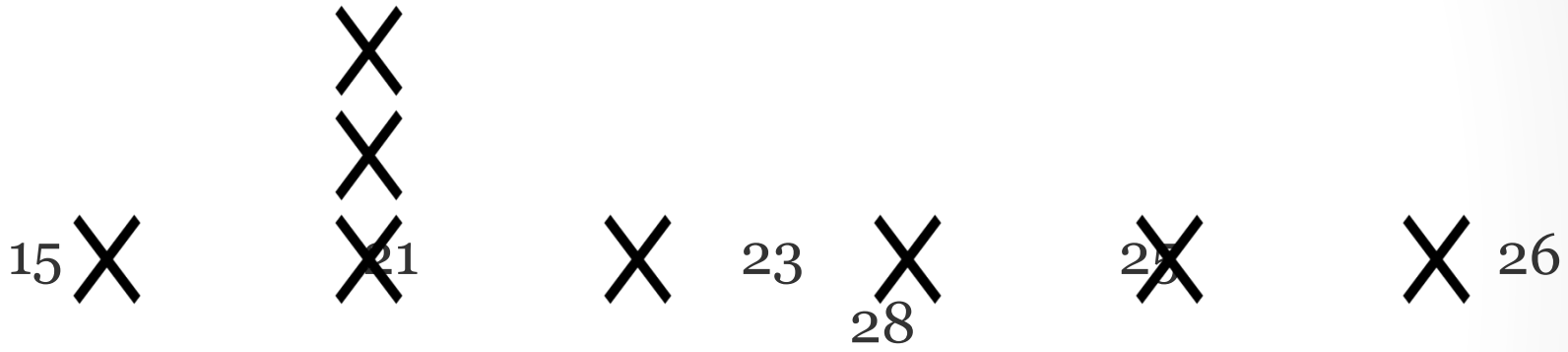


MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.

MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.



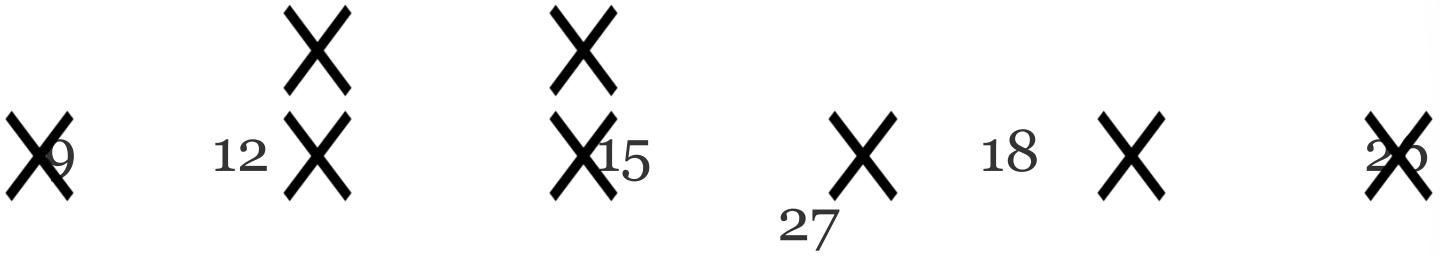
21 is the mode because it occurs most frequently

MODE EXAMPLE

- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.

MODE EXAMPLE

- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.



12 and 15 are the modes since the both occur twice.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

There is no mode since all values occur the same number of times.

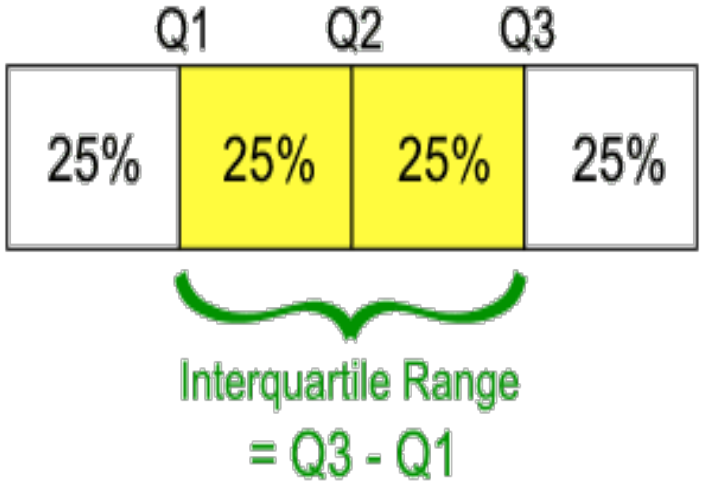
CODEALONG PART 1: BASIC STATS

- We can use Pandas to calculate the mean, median, mode, min, and max.

```
Methods available include:
.min() - Compute minimum value
.max() - Compute maximum value
.mean() - Compute mean value
.median() - Compute median value
.mode() - Compute mode value
.count() - Count the number of observations
```

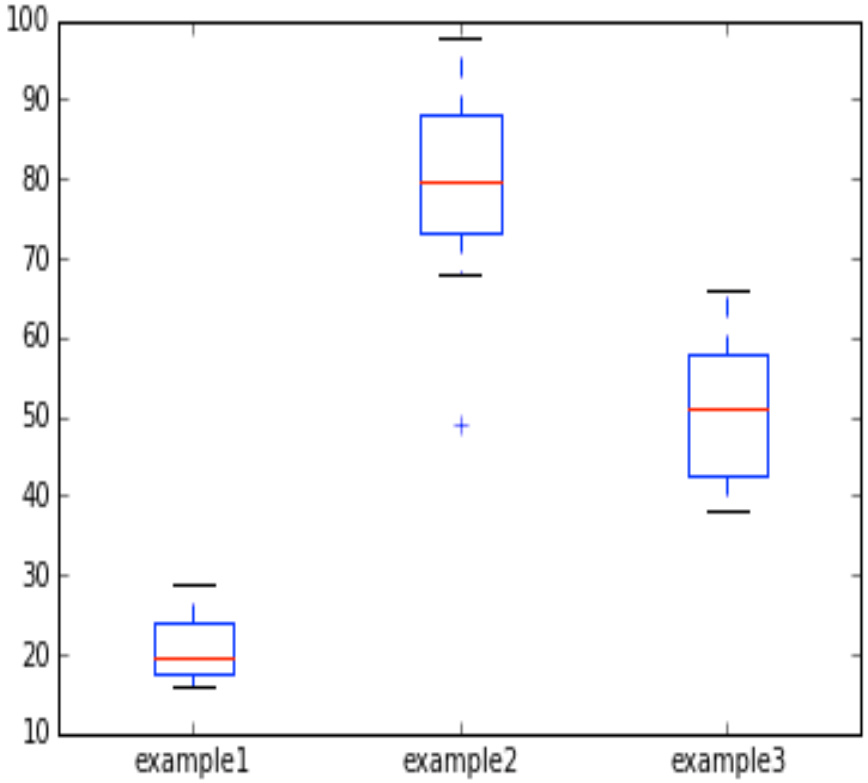
QUARTILES AND INTERQUARTILE RANGE

- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called first, second, and third quartiles, denoted Q1, Q2, and Q3, respectively.
- The interquartile range (IQR) is Q3 - Q1, a measure of variability.



CODEALONG PART 2: BOX PLOT

- Box plots give a nice visual of min, max, mean, median, and the quartile and interquartile range.



Five-Number Summary

- The *five-number summary* of a distribution reports its median, quartiles, and extremes (maximum and minimum).

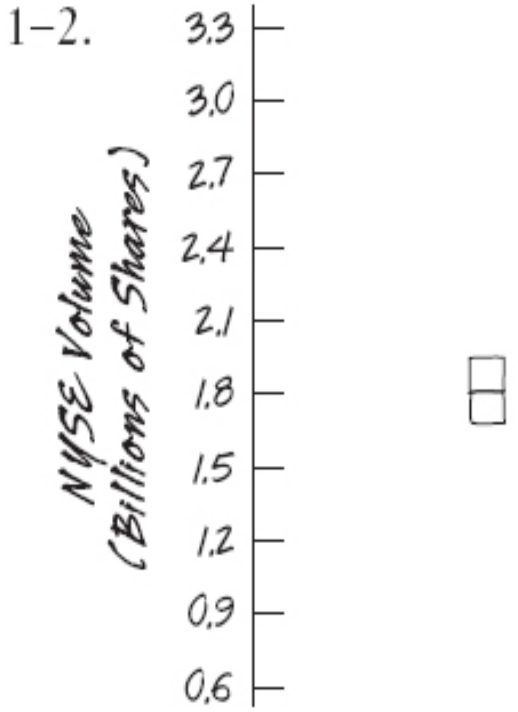
Max	3.287
Q3	1.972
Median	1.824
Q1	1.675
Min	0.616

Boxplots

- Once we have a five-number summary of a variable, we can display that information in a *boxplot*. To make a boxplot:

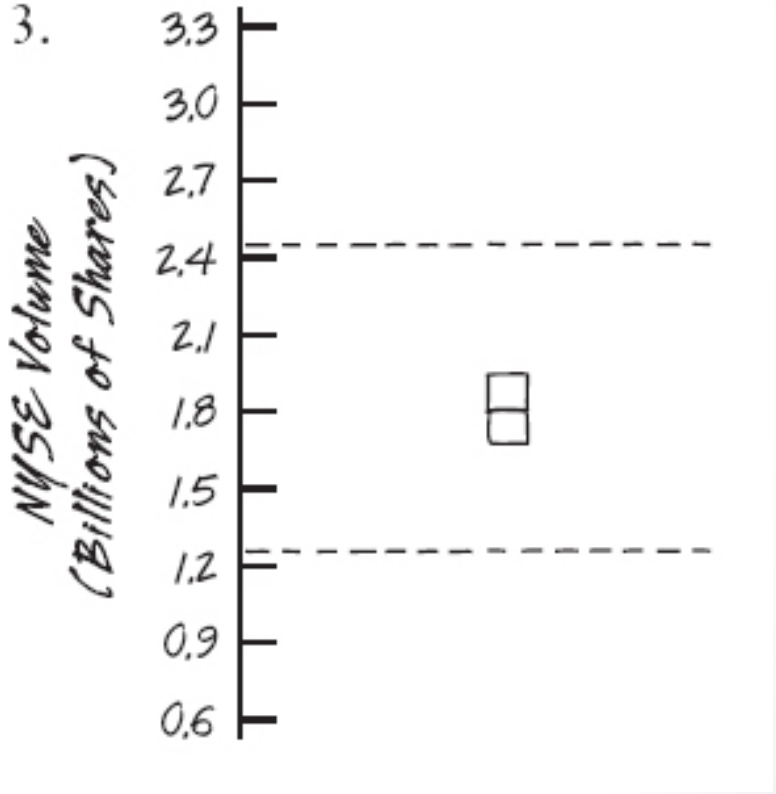
1) Draw a single vertical axis spanning the extent of the data.

2) Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box



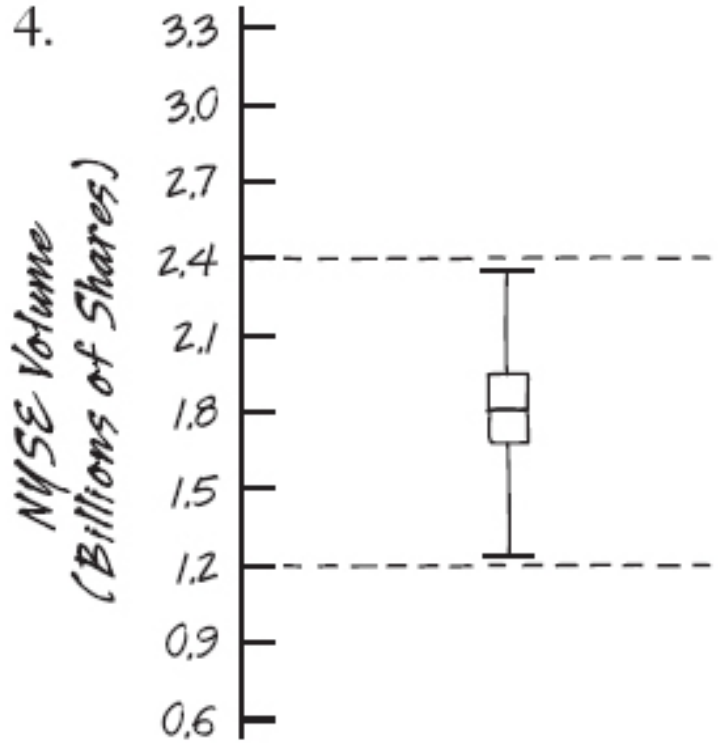
Boxplots

3) Erect (but don't show in the final plot) "fences" around the main part of the data, placing the upper fence 1.5 IQRs above the upper quartile and the lower fence 1.5 IQRs below the lower quartile.



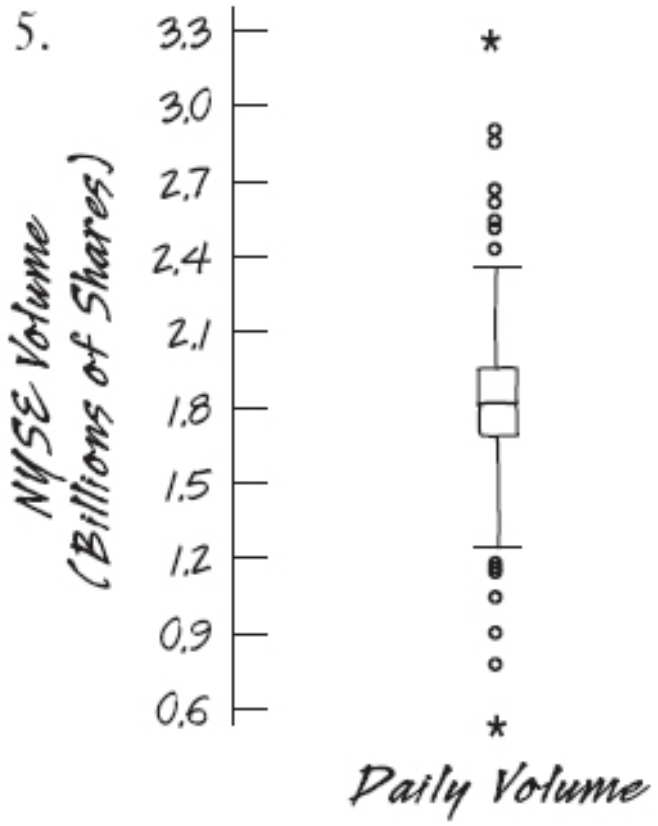
Boxplots

4) Draw lines (whiskers) from each end of the box up and down to the most extreme data values found *within* the fences.



Boxplots

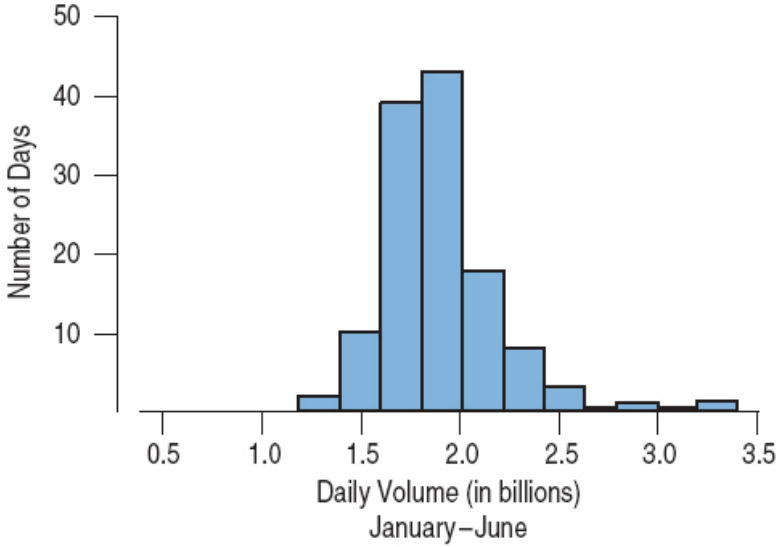
5) Add any outliers by displaying data values that lie beyond the fences with special symbols.

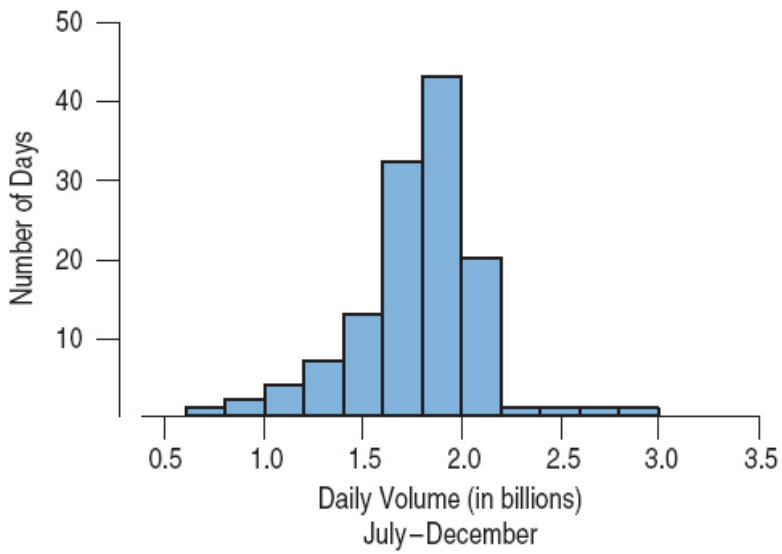


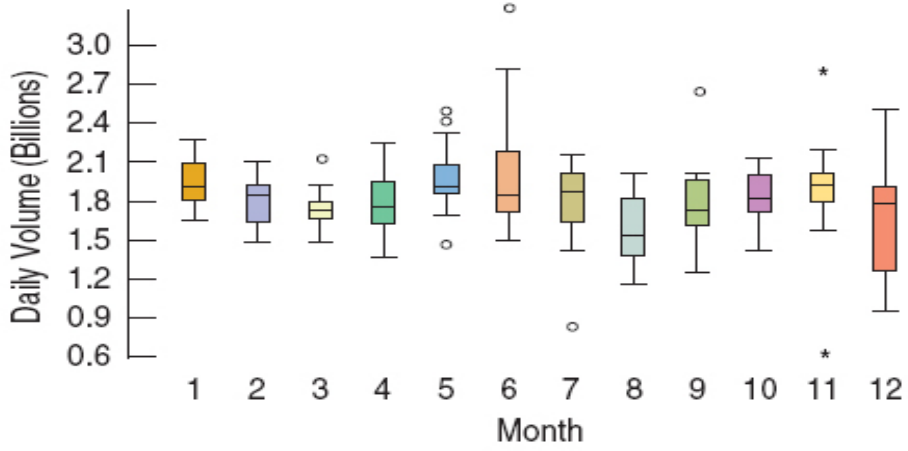
Boxplots

- The center of a boxplot shows the middle half of the data between the quartiles – the height of the box equals the IQR.
- If the median is roughly centered between the quartiles, then the middle half of the data is roughly symmetric. If it is not centered, the distribution is skewed.
- The whiskers show skewness as well if they are not roughly the same length.

Comparing Groups



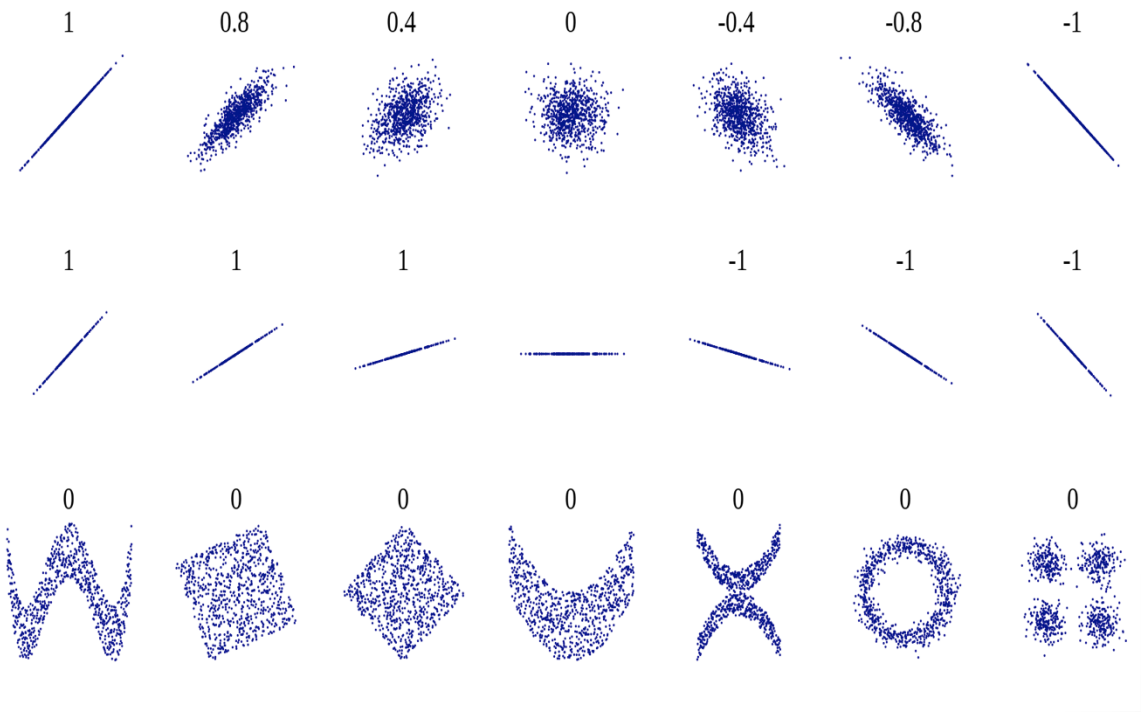




CORRELATION

- The correlation measures the extent of linear interdependence of variable quantities.

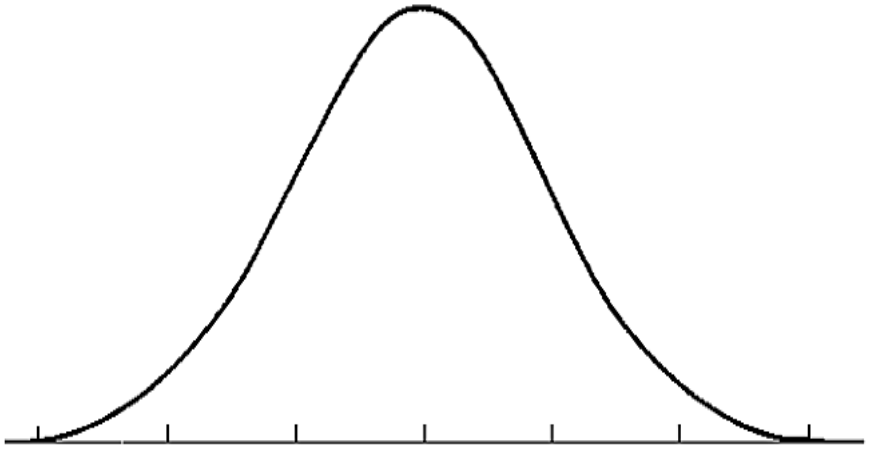
- Example correlation values



IS THIS NORMAL?

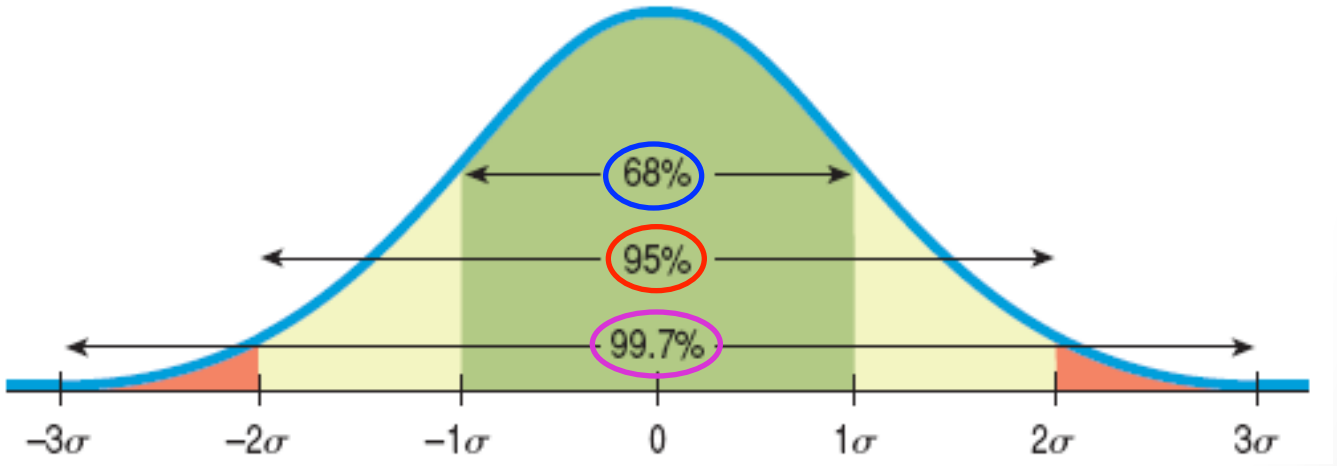
THE NORMAL DISTRIBUTION

- A normal distribution is often a key assumption to many models.
- The normal distribution depends upon the *mean* and the *standard deviation*.
- The *mean* determines the center of the distribution. The *standard deviation* determines the height and width of the distribution.



The *68-95-99.7 Rule* (the *Empirical Rule*)

- In bell-shaped distributions, about 68% of the values fall within one standard deviation of the mean, about 95% of the values fall within two standard deviations of the mean, and about 99.7% of the values fall within three standard deviations of the mean.



Practice with Normal Distribution Calculations

- **Example 1**: Each Scholastic Aptitude Test (SAT) has a distribution that is roughly unimodal and symmetric and is designed to have an overall mean of 500 and a standard deviation of 100.
- Suppose you earned a 600 on an SAT test. From the information above and the 68-95-99.7 Rule, where do you stand among all students who took the SAT?

Practice with Normal Distribution Calculations

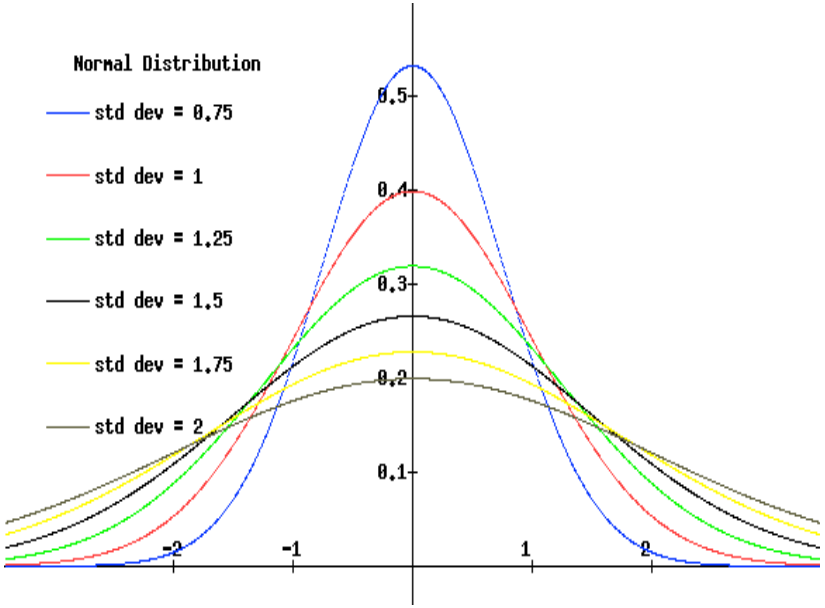
- **Example 1** (continued): A score of 600 is 1 SD above the mean. That corresponds to one of the points in the 68-95-99.7% Rule.
- About 32% (100% – 68%) of those who took the test were more than one SD from the mean, but only half of those were on the high side.
- So about 16% (half of 32%) of the test scores were better than 600.

Practice with Normal Distribution Calculations

- **Example 2**: Assuming the SAT scores are nearly normal with $N(500,100)$, what proportion of SAT scores falls between 300 and 600?

THE NORMAL DISTRIBUTION

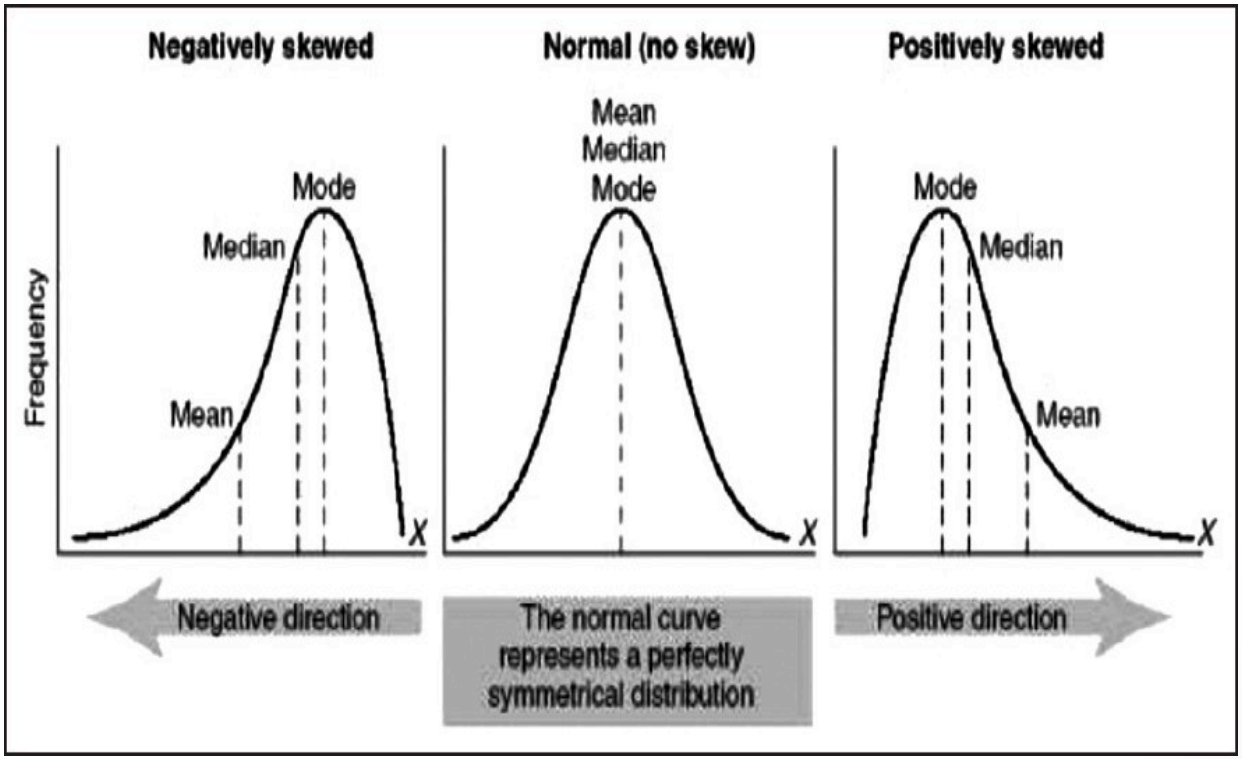
- Normal distributions are symmetric, bell-shaped curves.
- When the standard deviation is large, the curve is short and wide.
- When the standard deviation is small, the curve it tall and narrow.



SKEWNESS

- Skewness is a measure of the asymmetry of the distribution of a random variable about its mean.

- Skewness can be positive or negative, or even undefined.



DEMO

CLASSES

CLASS/DUMMY VARIABLES

- Let's say we have the categorical variable `area`, which takes on one of the following values: `rural`, `suburban`, and `urban`.

- We need to represent these numerically for a model. So how do we code them?

CLASS/DUMMY VARIABLES

- How about `0=rural`, `1=suburban`, and `2=urban`?

CLASS/DUMMY VARIABLES

- No, that implies that `urban` is *twice* `suburban`, an ordered relationship. This doesn't make sense.

- However, we can represent this information by converting the one `area` variable into two new variables, `area_urban` and `area_suburban`.

CLASS/DUMMY VARIABLES

- We'll draw out how categorical variables can be represented without implying order.
- First, let's choose a reference category. This will be our "base" category.
- It's often good to choose the category with the largest sample size and a criteria that will help model interpretation. If we are testing for a disease, the reference category would be people without the disease.

CLASS/DUMMY VARIABLES

- Step 1: Select a reference category. We'll choose `rural` as our reference category.
- Step 2: Convert the values `urban`, `suburban`, and `urban` into a numeric representation that does not imply order.
- Step 3: Create two new variables: `area_urban` and `area_suburban`.

CLASS/DUMMY VARIABLES

- Why do we need only two dummy variables?

rural	urban	suburban

- We can derive all of the possible values from these two. If an `area` isn't `urban` or `suburban`, we know it must be `rural`.

- In general, if you have a categorical feature with k categories, you need to create k-1 dummy variable to represent all of the information.

CLASS/DUMMY VARIABLES

- Let's see our dummy variables.

	area_urban	area_suburban
rural	0	0
suburban	0	1
urban	1	0

- As mentioned before, if we know `area_urban=0` and `area_suburban=0`, then the area must be `rural`.

CLASS/DUMMY VARIABLES

- We can do this for a `gender` variable with two categories: `male` and female.

- How many dummy variables need to be created?

CLASS/DUMMY VARIABLES

- # of categories - 1 = 2 -1 = 1

CLASS/DUMMY VARIABLES

- We will make `female` our reference category. Thus, `female=0` and `male=1`.

	gender_male
female	0
male	1

- This can be done in Pandas with the `get_dummies` method.