

STATISTICS FUNDAMENTALS

Sri Kanajan

STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

- 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
- Create data visualizations - including: line graphs, box plots, and histograms- to discern characteristics and trends in a dataset
- Identify a normal distribution within a dataset using summary statistics and visualization
- ID variable types and complete dummy coding by hand

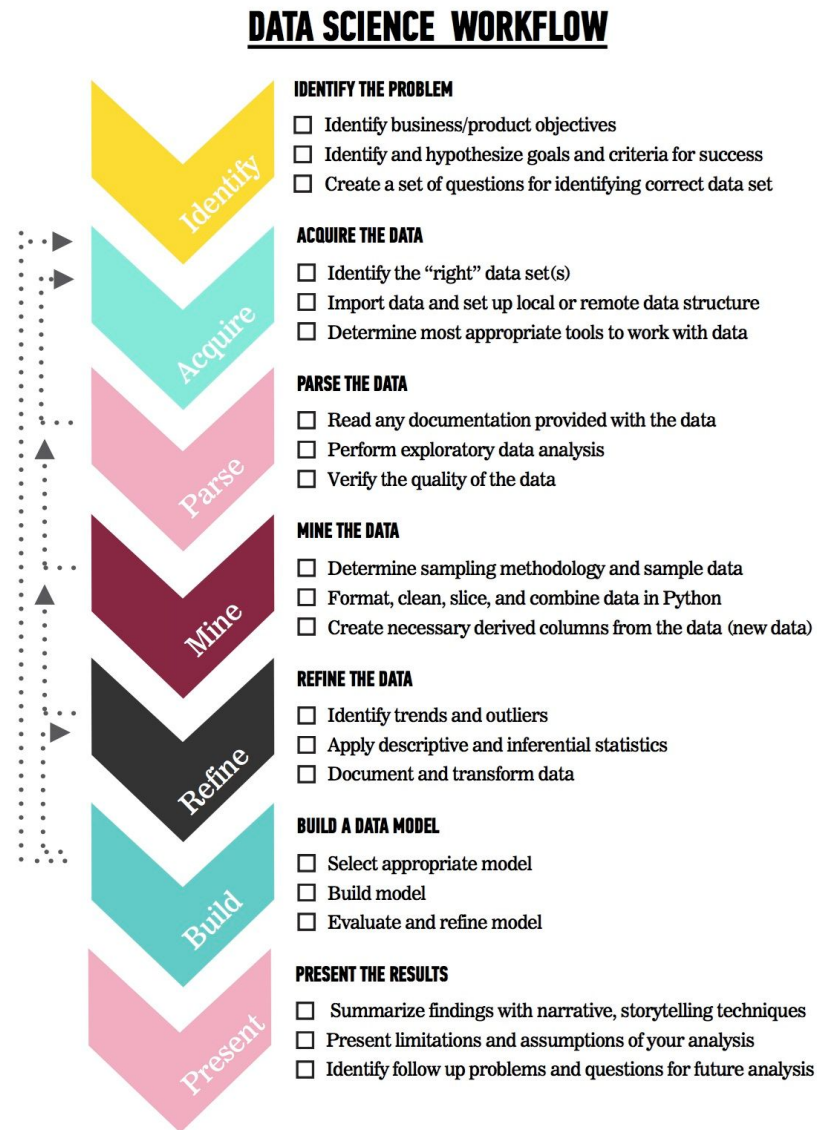
OPENING

STATISTICS FUNDAMENTALS

LET'S REVIEW THE DATA SCIENCE WORKFLOW

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



TODAY

- We're going to begin to talk about step 3: Parsing the Data
- We'll begin to talk about the fundamentals of Statistics

INTRODUCTION

LAYING THE GROUND WORK

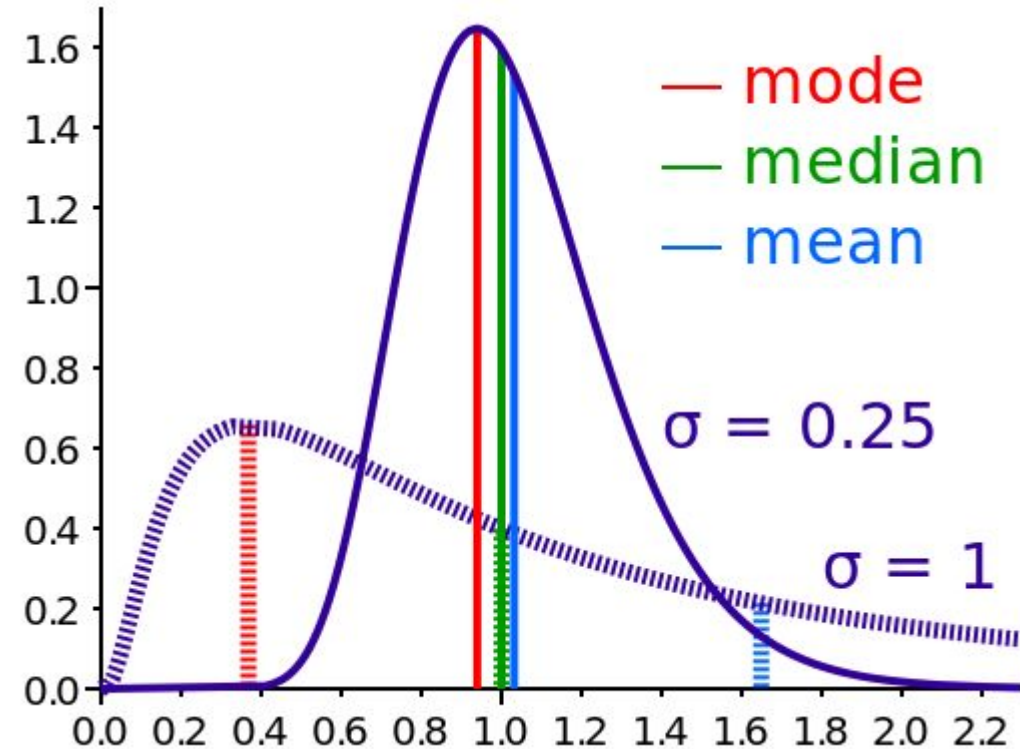
WE'RE GOING TO COVER SEVERAL TOPICS

- Mean
- Median
- Mode
- Max
- Min
- Quartile
- Interquartile Range
- Variance
- Standard Deviation
- Correlation

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.

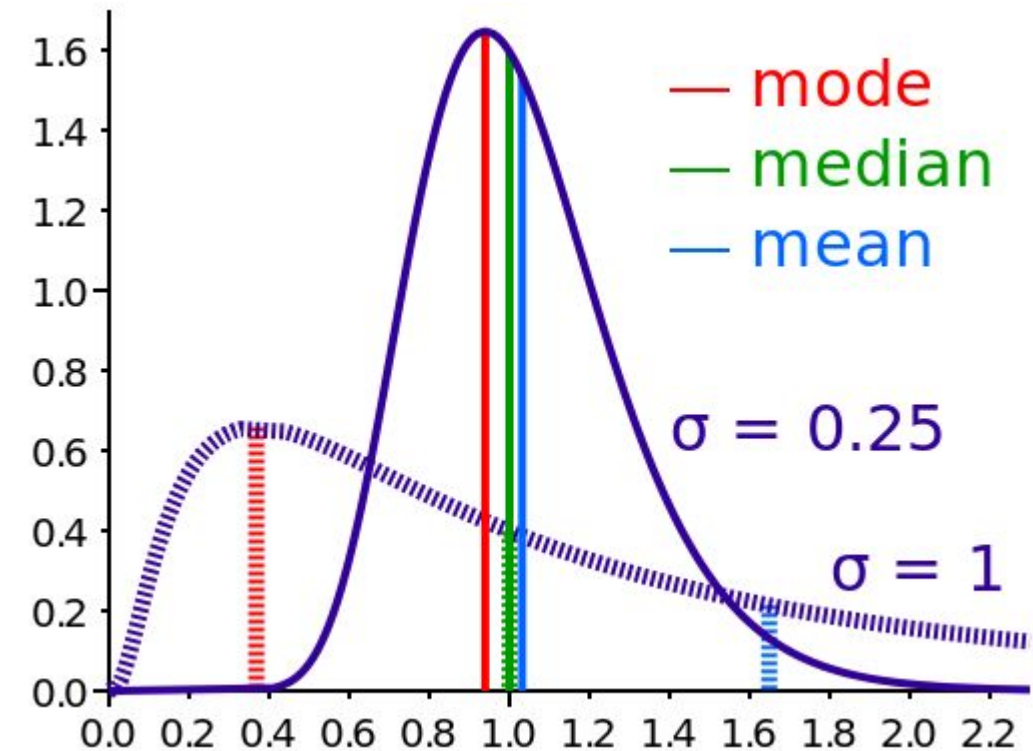
MEAN EXAMPLE

- 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.

MEDIAN EXAMPLE

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

Ordered Values:

28, 37, 67, 75, 81, 92

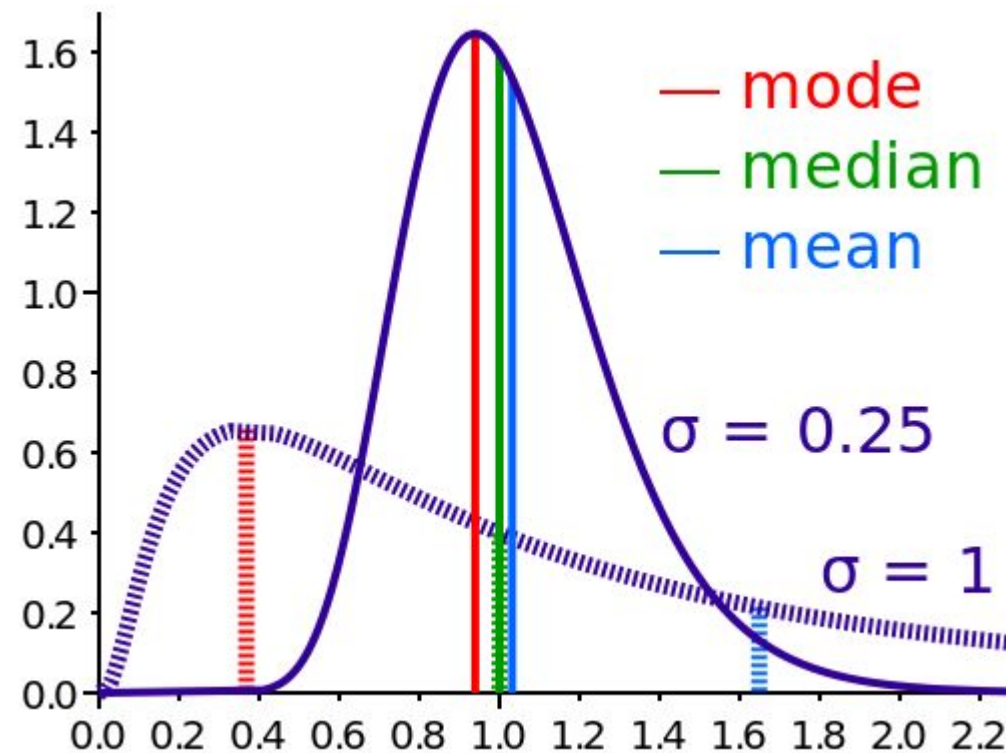
67 and 75 are the middle values.

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

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.

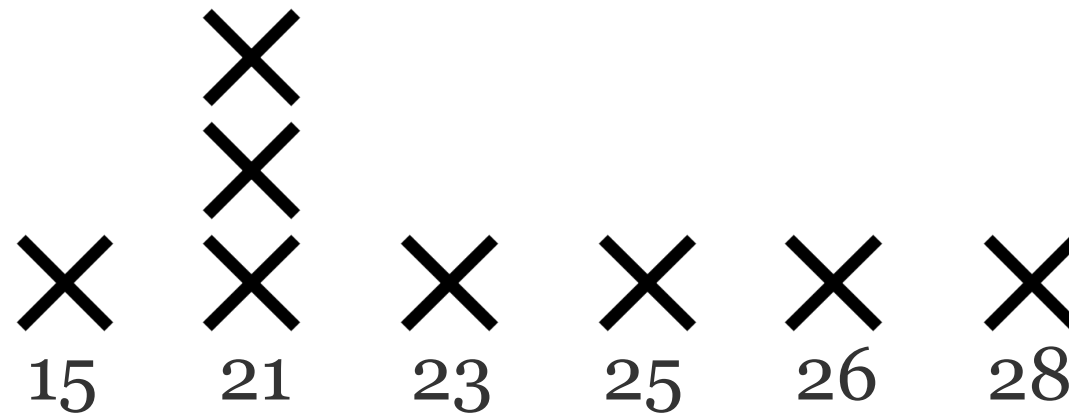


MODE EXAMPLE

- 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.



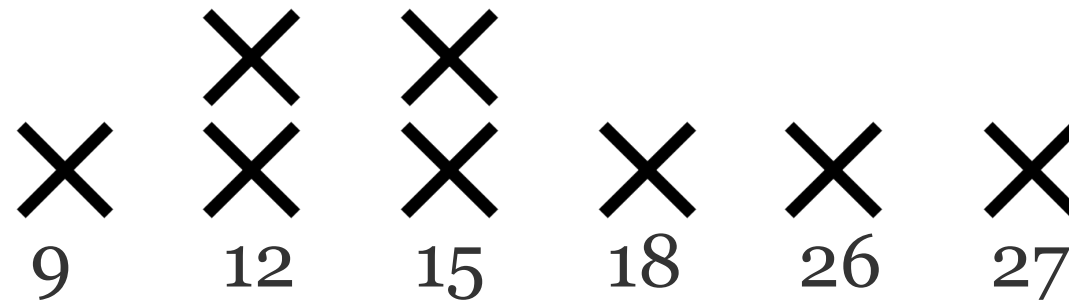
21 is the mode because it occurs most frequently

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.

CODEALONG

SUMMARY STATISTICS IN PANDAS

CODEALONG: SUMMARY STATISTICS IN PANDAS

- Open the starter-code notebook located in lessons/lesson-03/code/starter-code of the class repo.

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

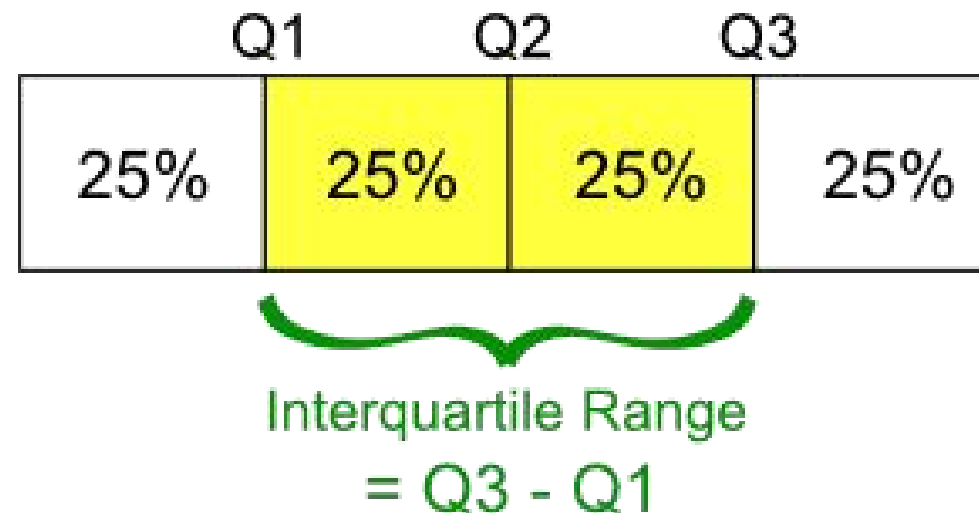
`.median()` - Compute median value

`.mode()` - Compute mode value

`.count()` - Count the number of observations

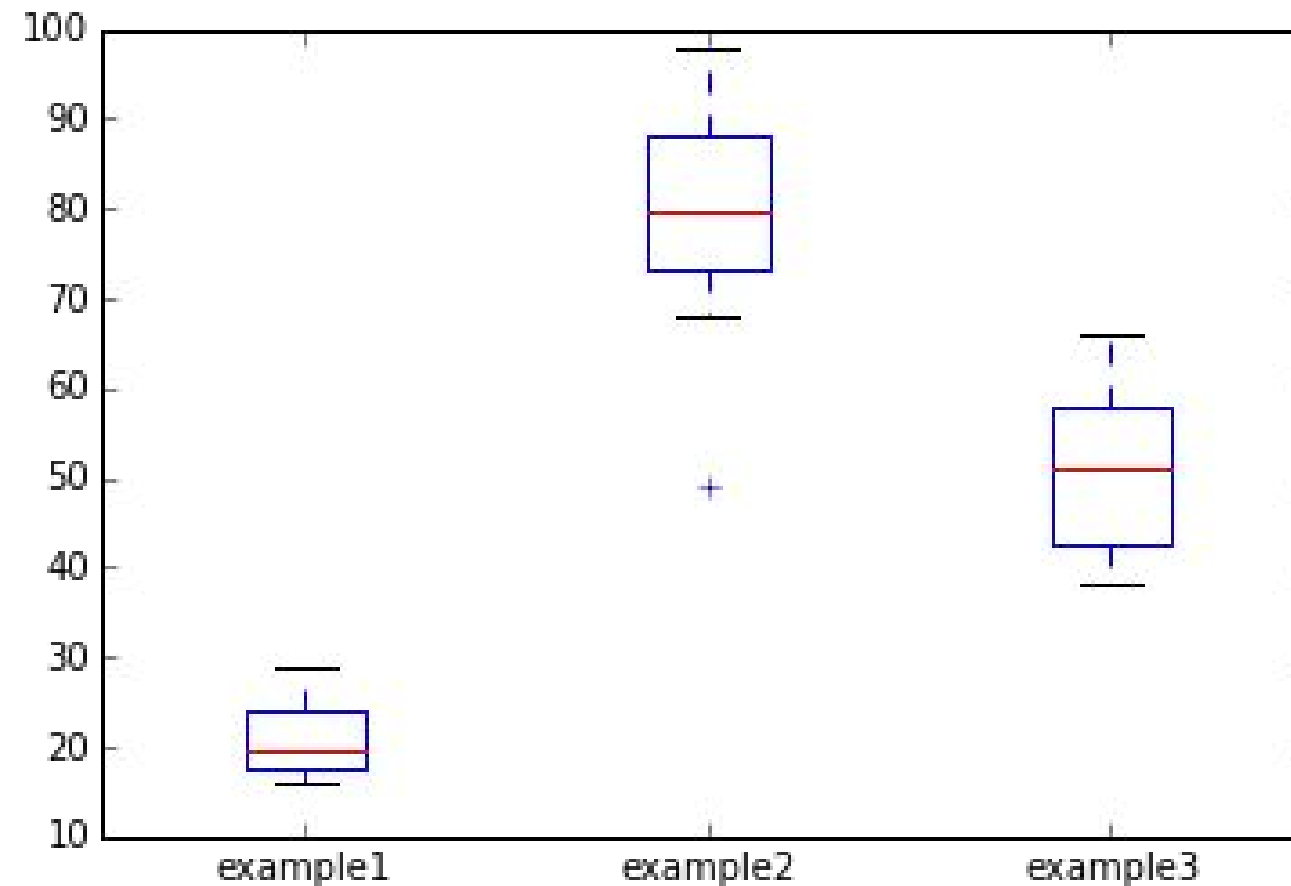
QUARTILES AND INTERQUARTILE RANGE

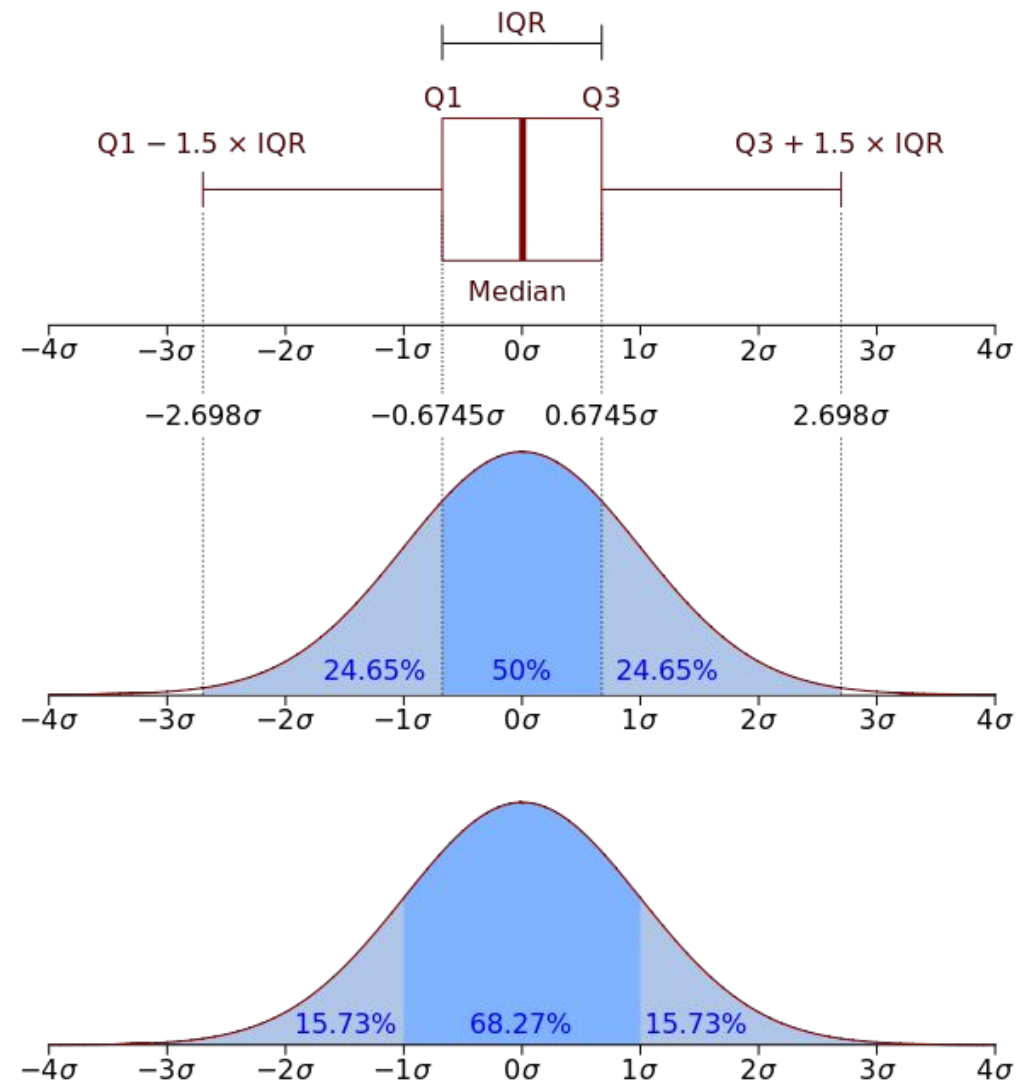
- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called first, second, and third quartiles, denoted Q_1 , Q_2 , and Q_3 , respectively.
- The interquartile range (IQR) is $Q_3 - Q_1$, a measure of variability.



CODEALONG PART 2: BOX PLOT

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





STANDARD DEVIATION

- Standard deviation (SD, σ for population, s for sample) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.
- Standard deviation is the square root of variance.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

CODEALONG PART 3: STANDARD DEVIATION & VARIANCE

- You can calculate variance and standard deviation easily in Pandas.

Methods include:

`.std()` - Compute Standard Deviation

`.var()` - Compute variance

`.describe()` - short cut that prints out count, mean, std, min, quartiles, max

CORRELATION

‣ The correlation measures the extent of interdependence of variable quantities. If continuous values, then use Pearson correlation.

‣ Assumptions of correlation:

- Linear relationship between the variables
- Mean and variance do not change over the range of x (stationary)
- Variable is continuous
- Absence of outliers

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

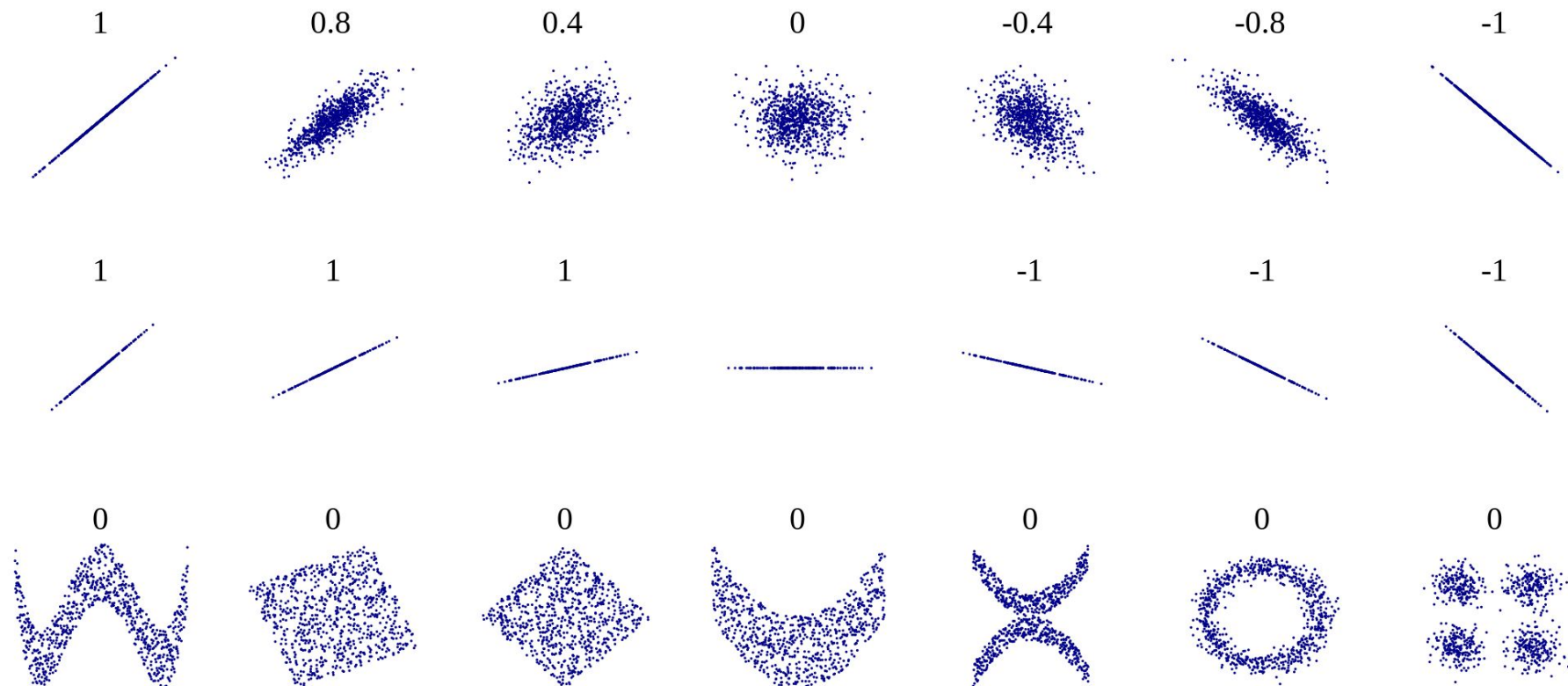
‣ Other types of correlation depending on the variable -

<http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/>

‣

CORRELATION

▸ Example correlation values



COLLINEARITY

- Two or more variables/features that linearly related to each other.
- E.g. $f_1 = f_2 * 3$ means that f_1 is colinear with f_2 OR $f_1 = f_2 + f_3$ means that f_1 is colinear with f_2 and f_3
- You can detect this by observing a very high correlation between the colinear features

CONTEXT

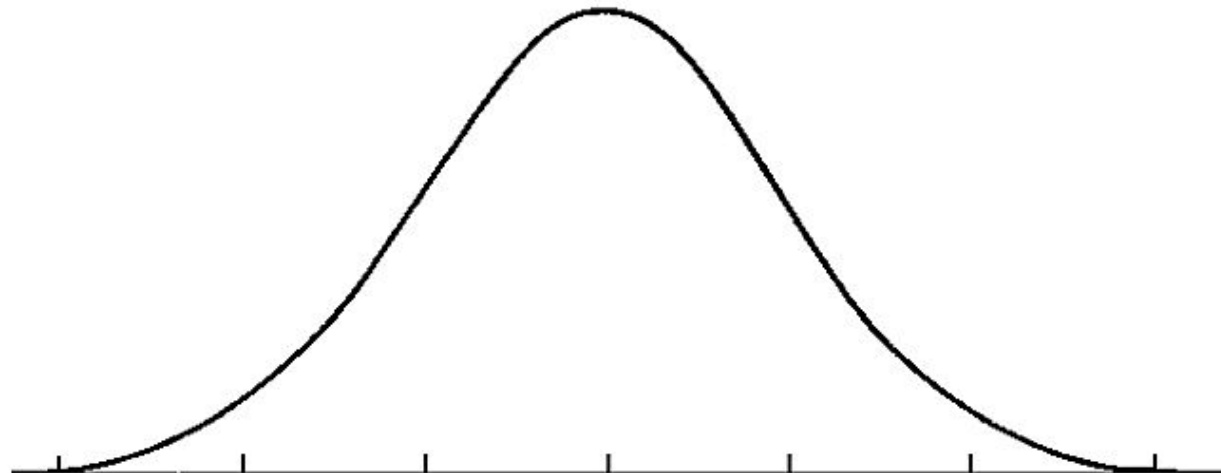
- For most projects, descriptive stats will come first. These help you get to know your dataset better.
- Sometimes, descriptive stats may be all you need to answer your question.

INTRODUCTION

IS THIS NORMAL?

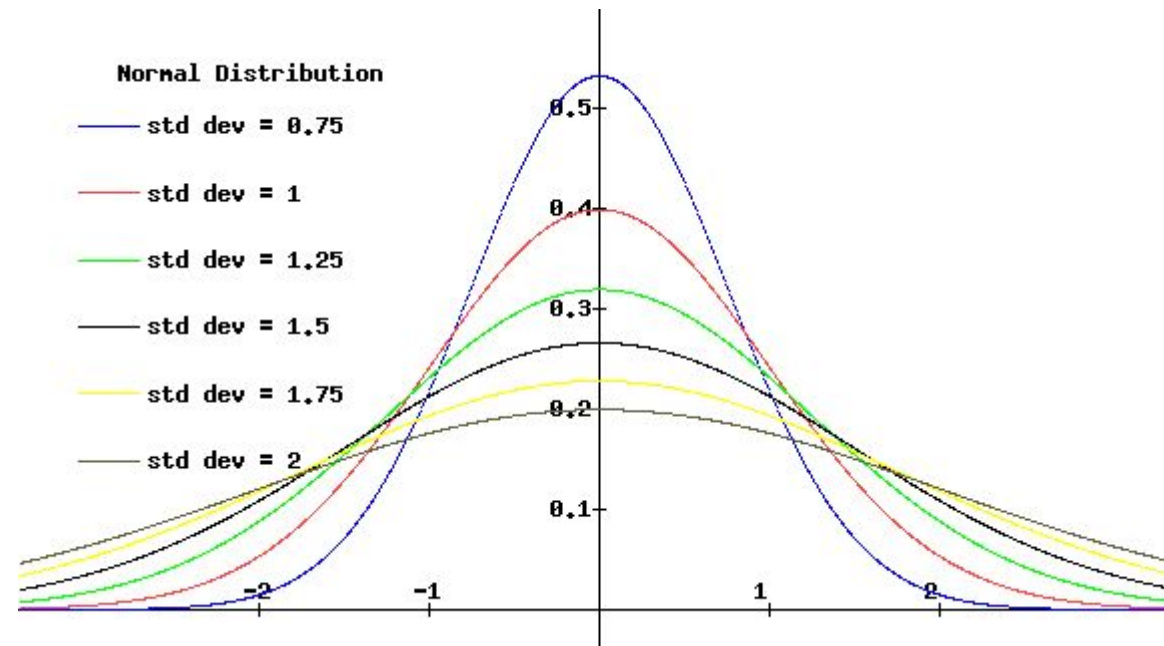
THE NORMAL DISTRIBUTION

- 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.



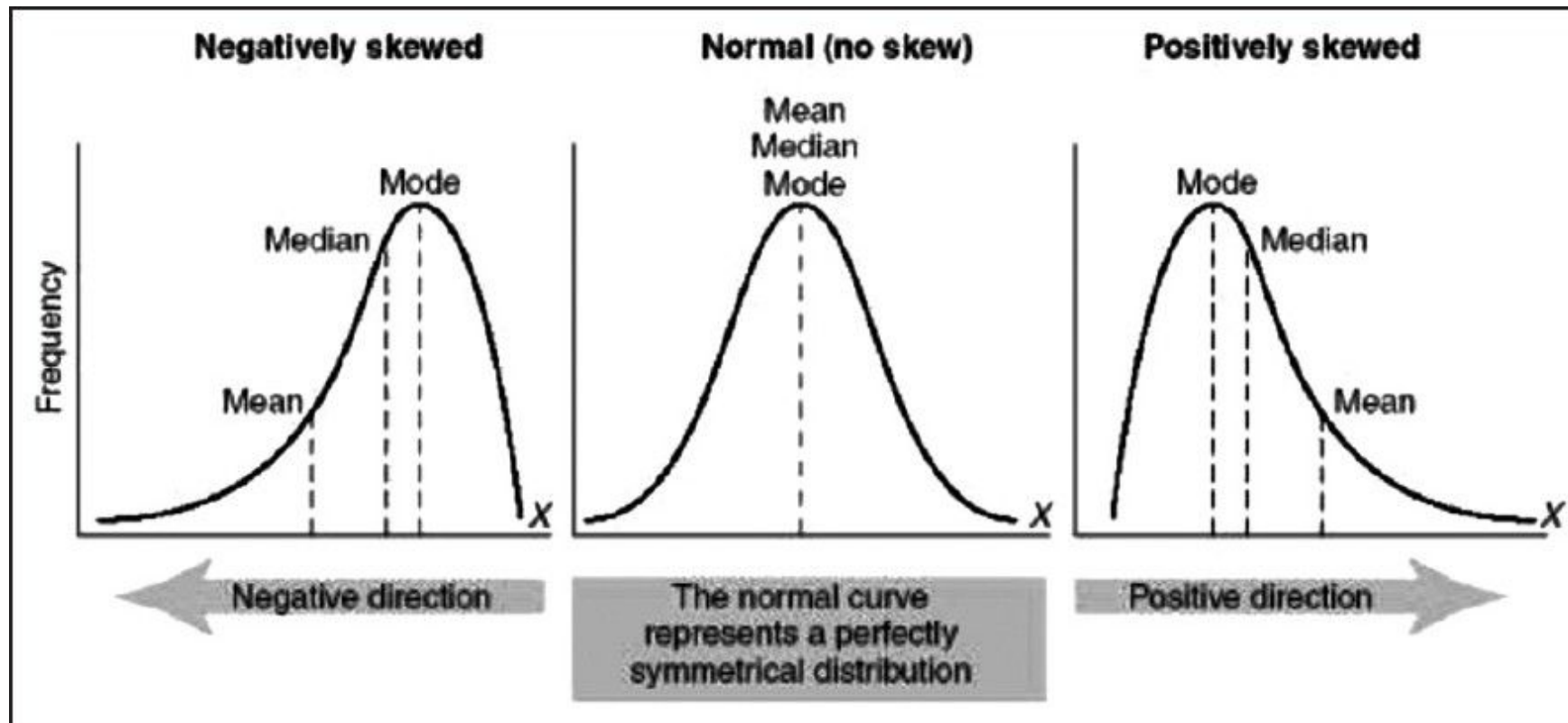
THE NORMAL DISTRIBUTION

- 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 is tall and narrow.



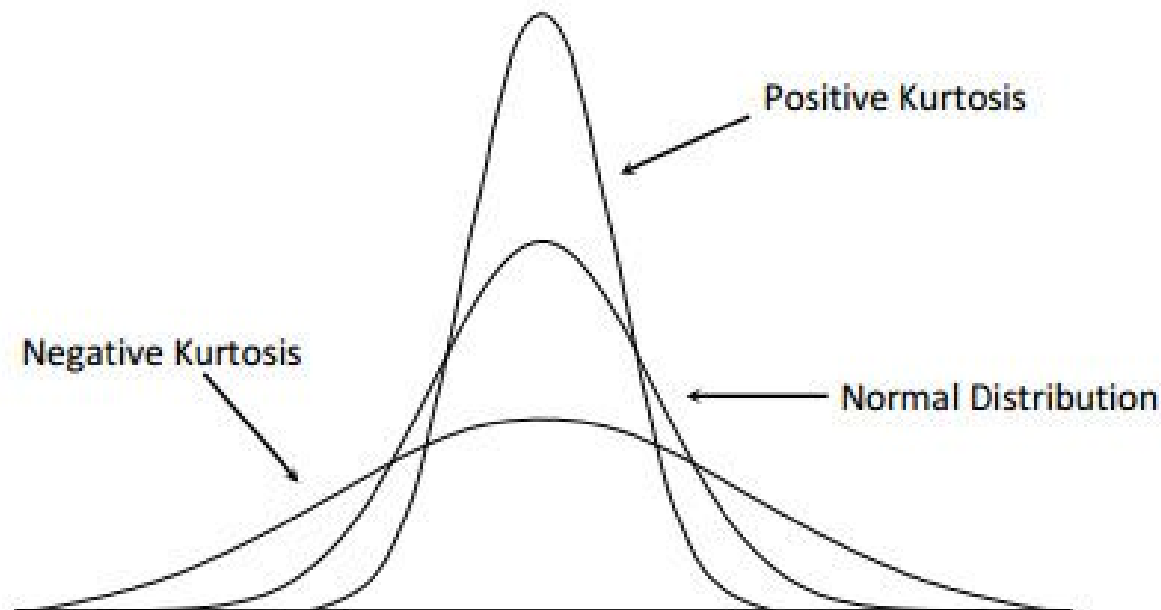
SKEWNESS

- 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.



KURTOSIS

- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.
- Datasets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails.



DETERMINING THE DISTRIBUTION OF YOUR DATA

- Follow along as we walk through this in an iPython Notebook.

DEMO

DEALING WITH CATEGORICAL DATA

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

- But this implies an ordered relationship - is urban twice suburban?
That 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 $\text{area_urban}=0$ and $\text{area_suburban}=0$, 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
<code>female</code>	<code>0</code>
<code>male</code>	<code>1</code>

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

LESSON

CREDITS

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET