# Determining When and Why to Use Univariate Analysis

Guillermo Fernández
DATA SCIENTIST

@guillermo_ai

# Summary

Get used to univariate analysis techniques

Understand when and why to use them

Understand the insight we can get from each technique

Perform univariate analysis techniques with Python

# Tracing a Knowledge Map

# Univariate Analysis Goal

**Summarize Observations**

To characterize data

**Numerically and Visually**

To represent information

# Types of Variables

**Quantitative**

Defined by numbers

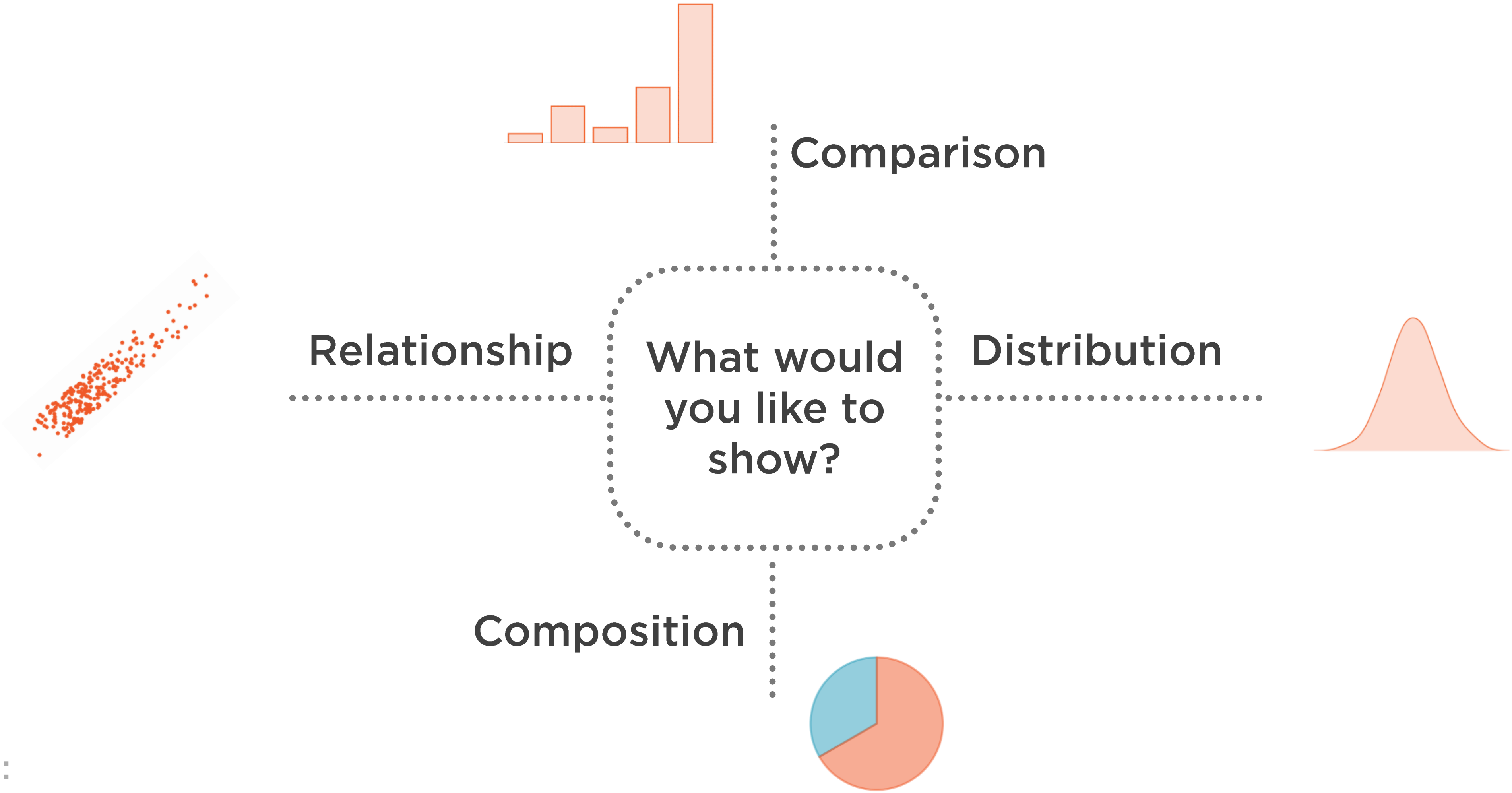
**Qualitative, Categorical or Nominal**

Defined by labels

**Chronological**

Defined by time

# Techniques Map



What would you like to show?

Comparison

Relationship

Distribution

Composition

Sources:
http://www.storytellingwithdata.com
https://visual.ly/blog/graphic-continuum
https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf

# Characterizing Data

# George Udny Yule Conditions

Independent of Observer

Depend on All Values of Series

Value Must Have a Concrete Meaning

Easy to Compute

Not Sensitive to Random Processes

# Measures of Central Tendency and Dispersion

**Local Concentration**

Mean, Median, Mode, Quantiles

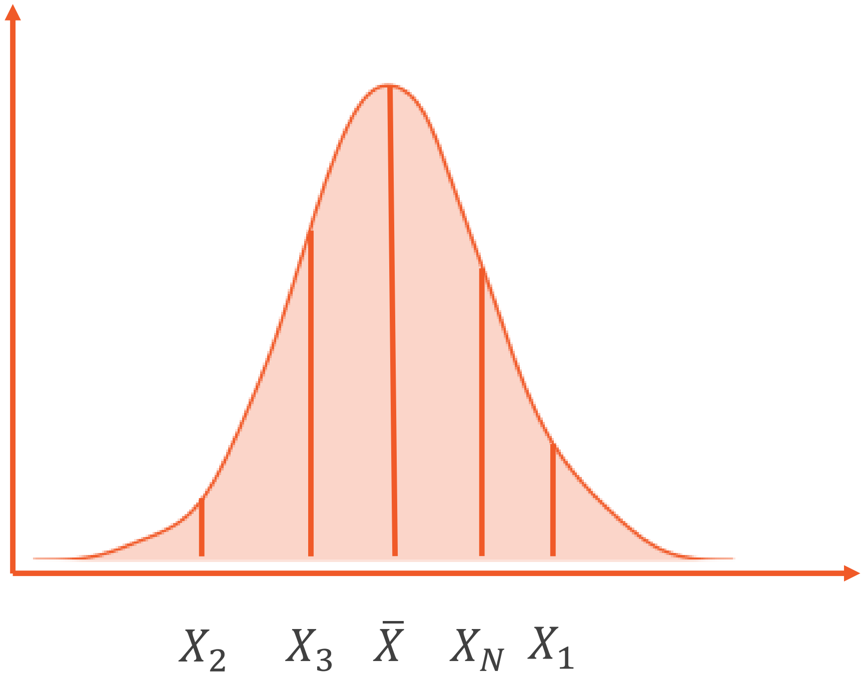
**Dispersion**

Standard deviation, Variance

**Shape**

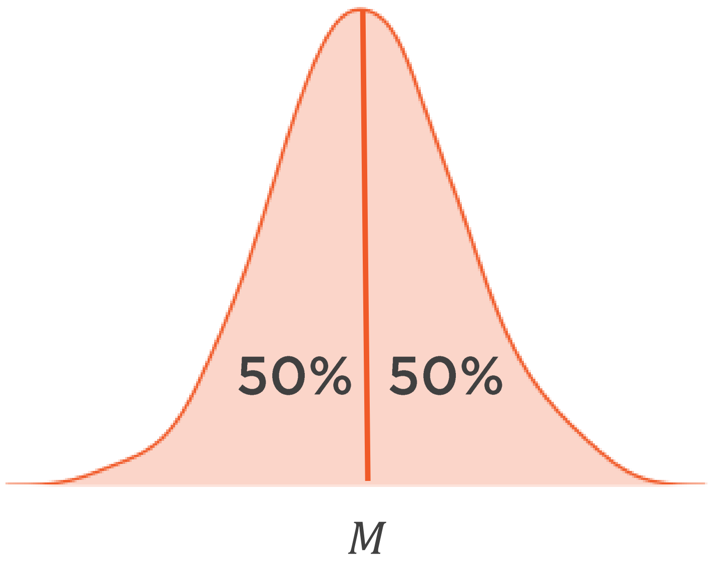
Skewness, Kurtosis

# Mean



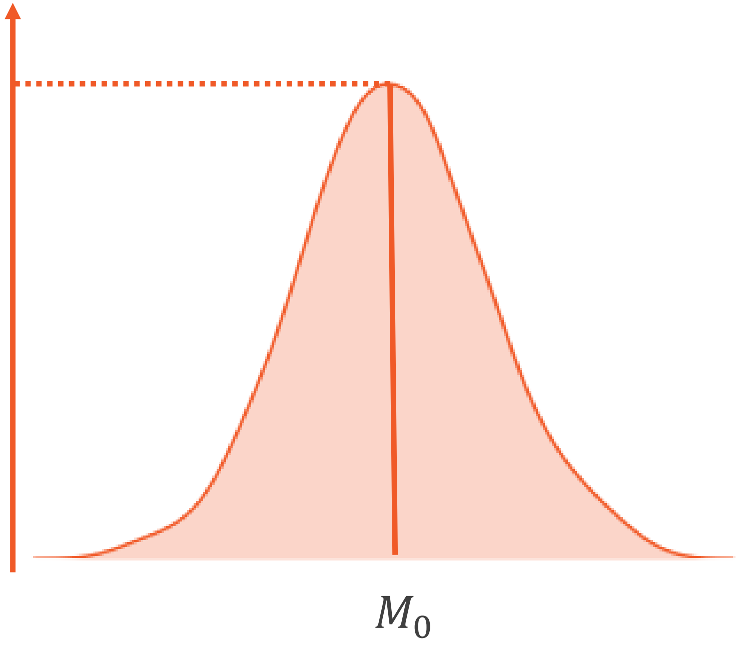
$$\bar{X} = \frac{\sum_{i}^{N} X_i}{N}$$

# Median



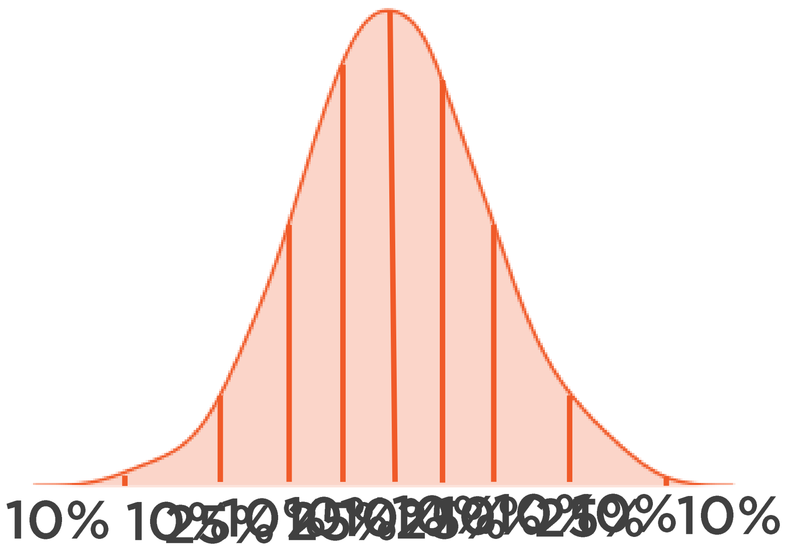
$$M = \left\{ \begin{array}{c} X_{p+1}\ \ \textit{if total observations are } 2p+1 \\ X_p\ \ \textit{if total observations are } 2p \end{array} \right\}$$

# Mode



$$M_0 = 3M - 2\bar{X}$$

# Quantiles, quartiles and deciles



$Q_1 > 25\%$

$Q_2 > 50\%$

$Q_3 > 75\%$

$D_1 > 10\%$

$D_2 > 20\%$

$D_3 > 30\%$

$D_4 > 40\%$

$D_5 > 50\%$

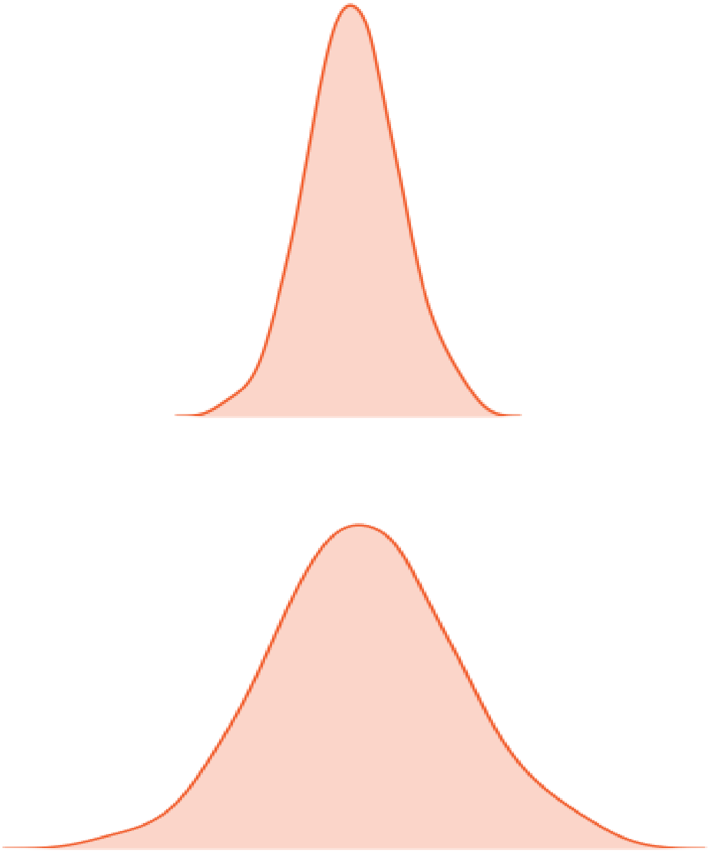
$D_6 > 60\%$

$D_7 > 70\%$

$D_8 > 80\%$

$D_9 > 90\%$

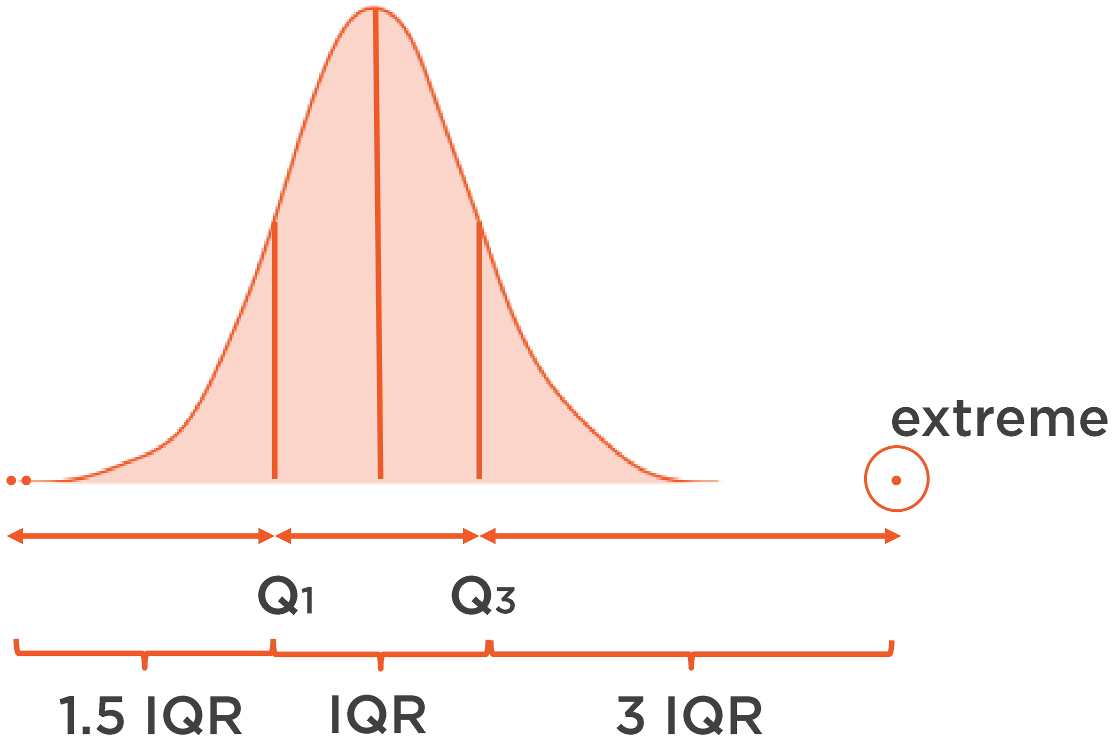
# Measures of Dispersion



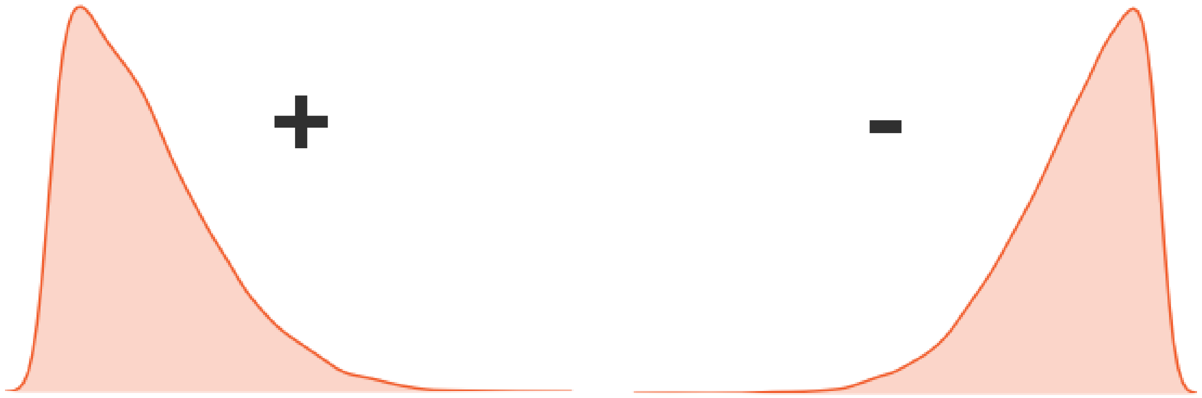
$$S = \sqrt{\frac{\sum_{i}^{N}(X_i - \bar{X})^2}{N}}$$

**Variance** $= s^2$

# Outliers

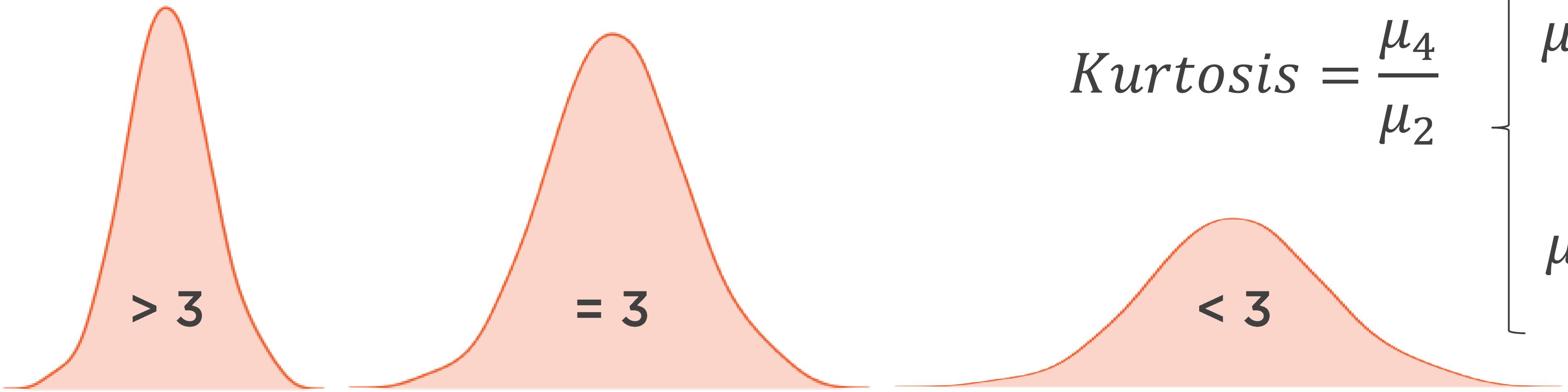


# Skewness and Kurtosis



$$Skewness = \frac{3(\bar{X} - M)}{s}$$

$$Kurtosis = \frac{\mu_4}{\mu_2} \begin{cases} \mu_4 = \dfrac{\sum_i^N (X_i - \bar{X})^4}{N} \\ \mu_2 = \dfrac{\sum_i^N (X_i - \bar{X})^2}{N} \end{cases}$$



Leptokurtic Mesokurtic Platykurtic

# Demo

**Learn different ways to compute measures of central tendency and dispersion with Python**

**Using Python packages**

- Statistics
- Pandas
- Numpy
- Scipy Stats

**Plot our first graph with Seaborn**

# Demo Tools

colab.research.google.com

# Visualization Libraries

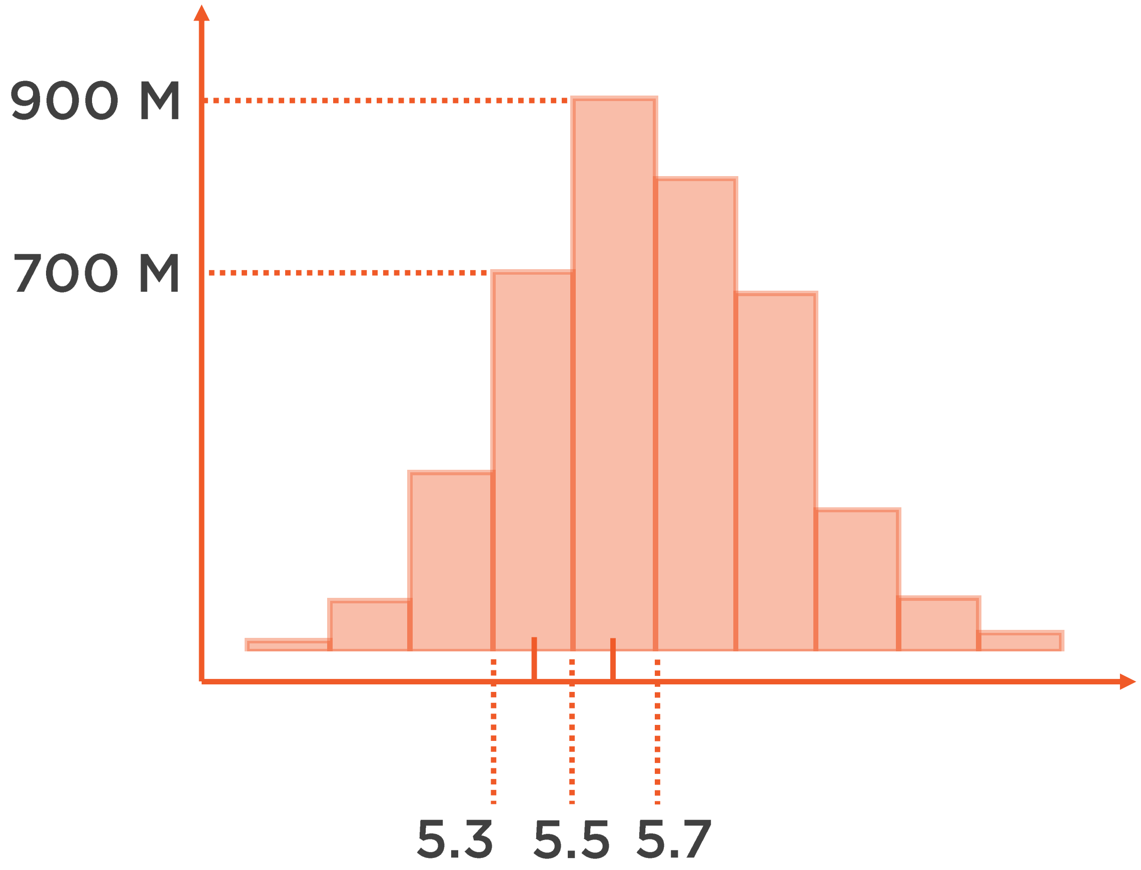
matplotlib

Version 3.0.3

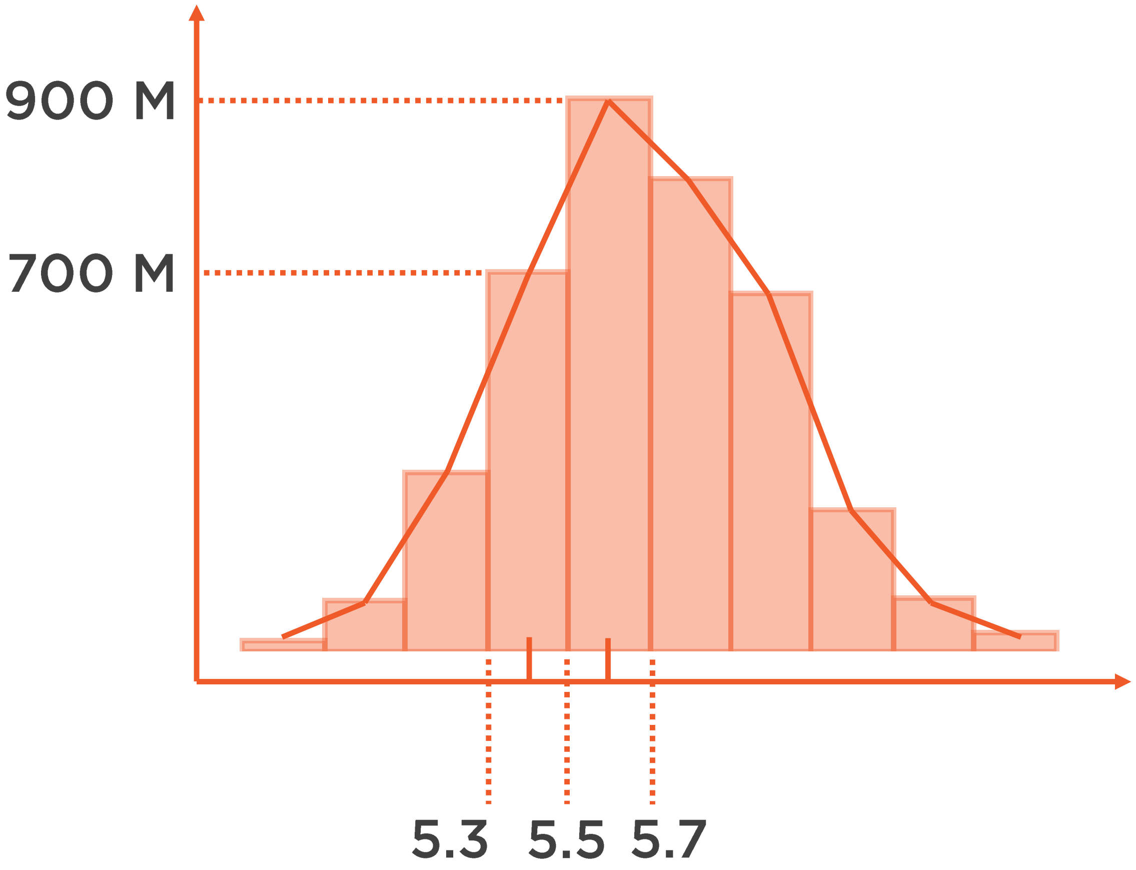
seaborn

# Univariate Distribution Plots

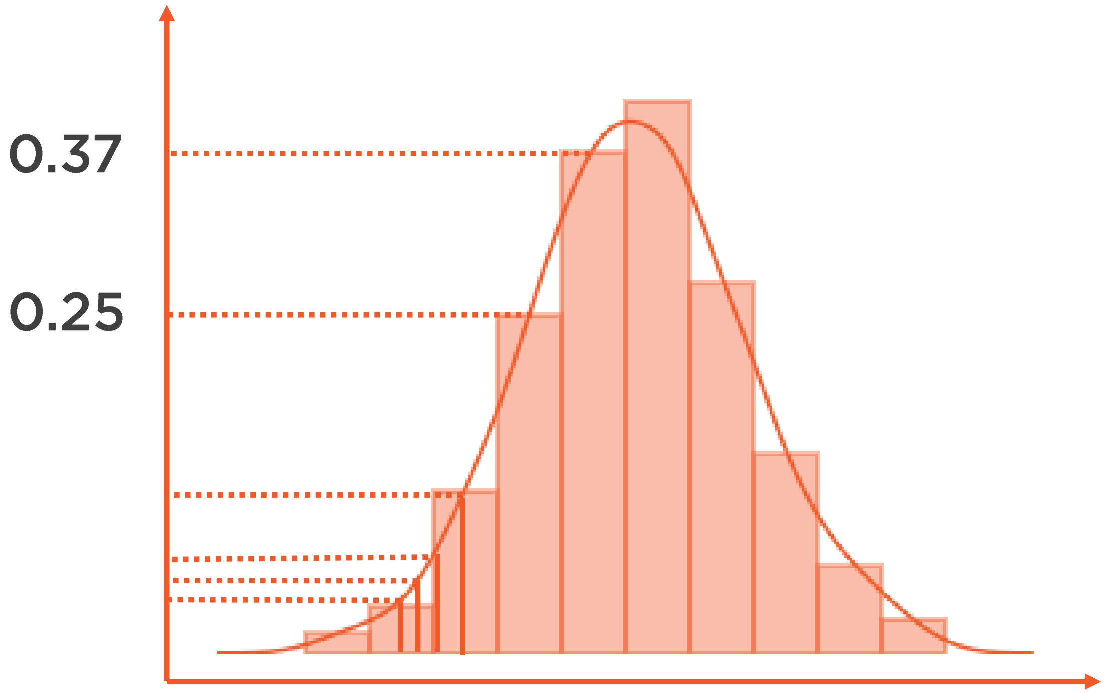
# Histogram



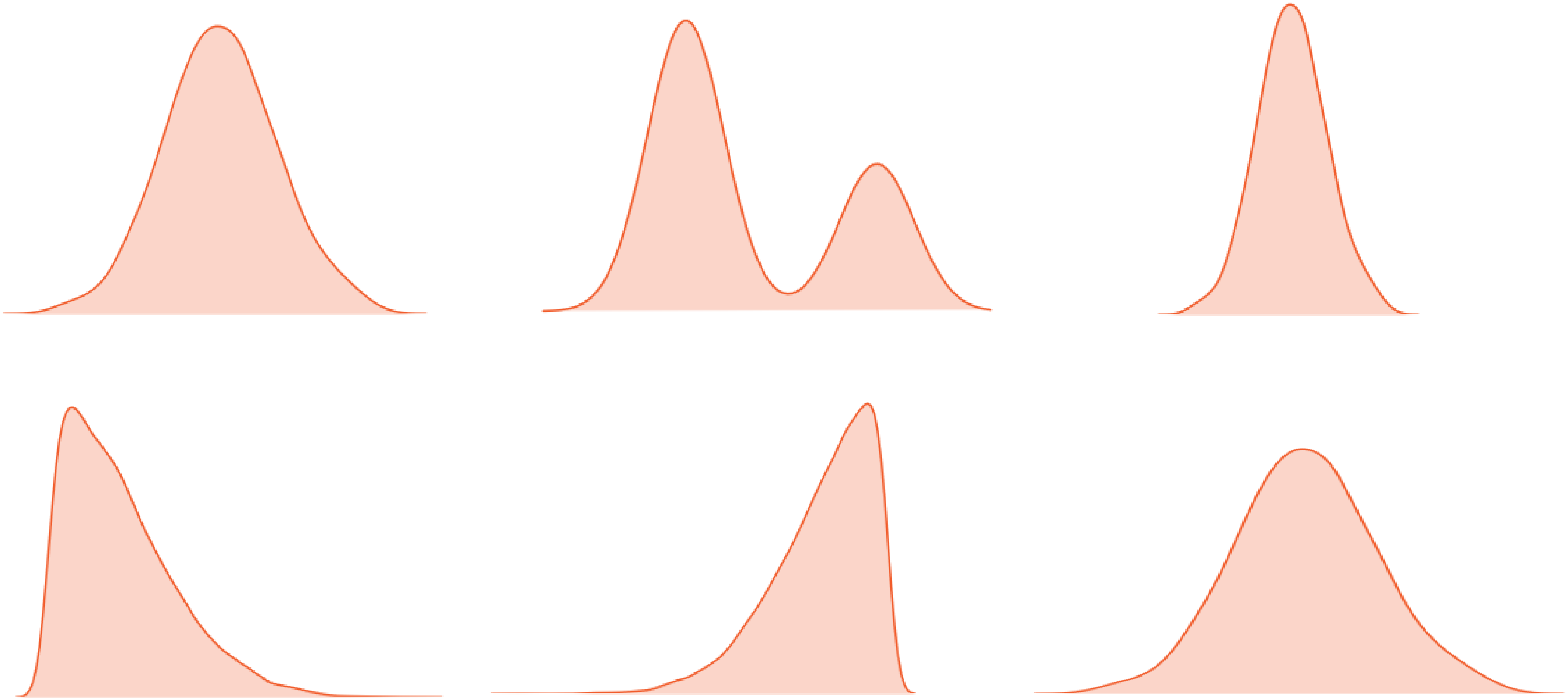
# Frequency Polygon



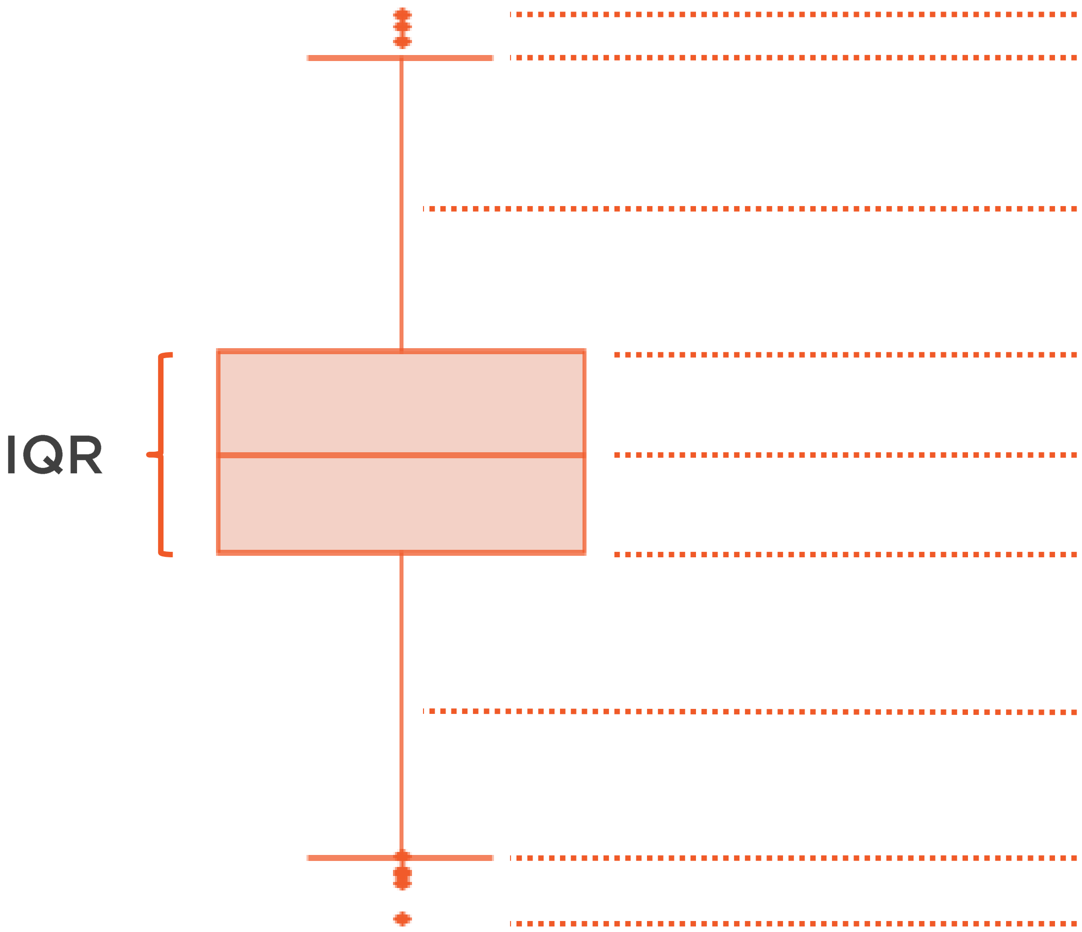
# Density Plot



# Types of Frequency Curves



# Box Plot



IQR

Outliers Extreme if > $Q_3$ + 3 IQR

Maximum $Q_3$ + 1.5 IQR

Whiskers

Third Quartile $Q_3$

Median or Second Quartile $Q_2$

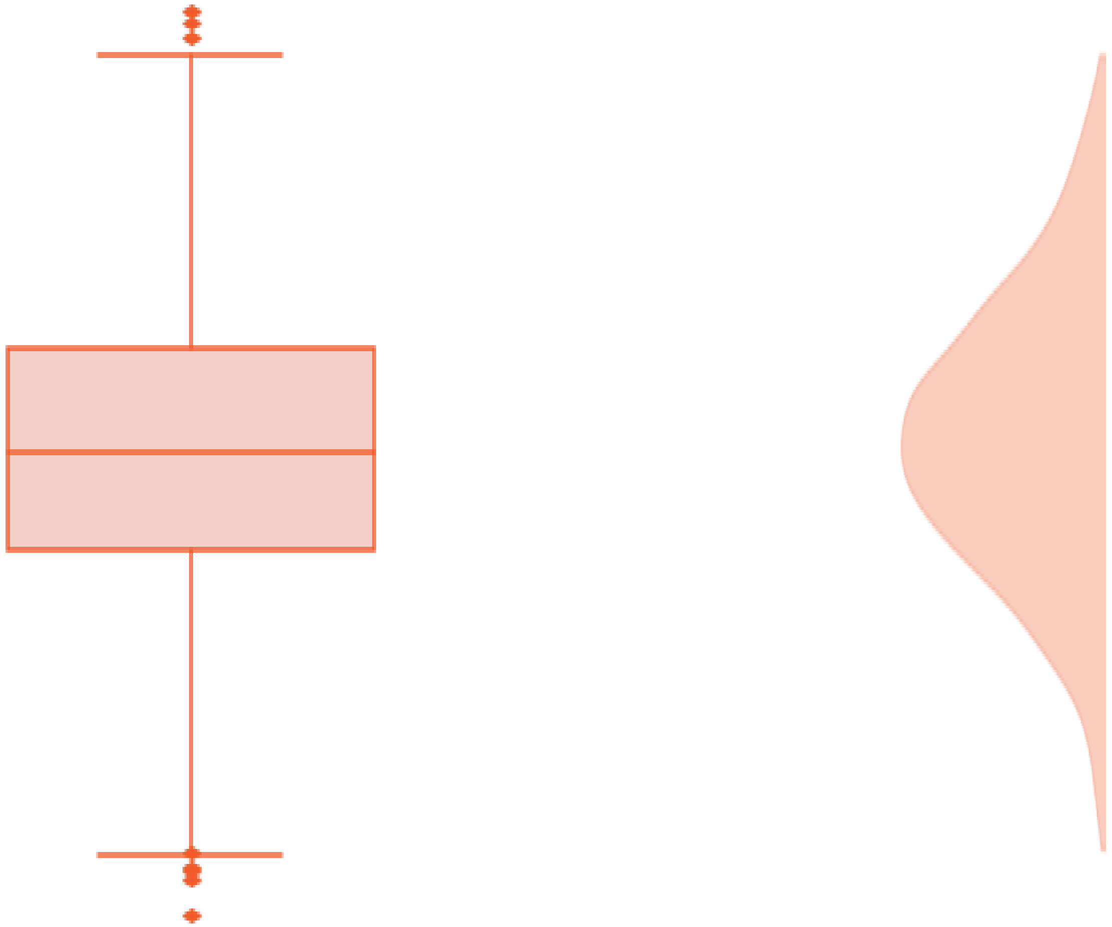
First Quartile $Q_1$

Whiskers

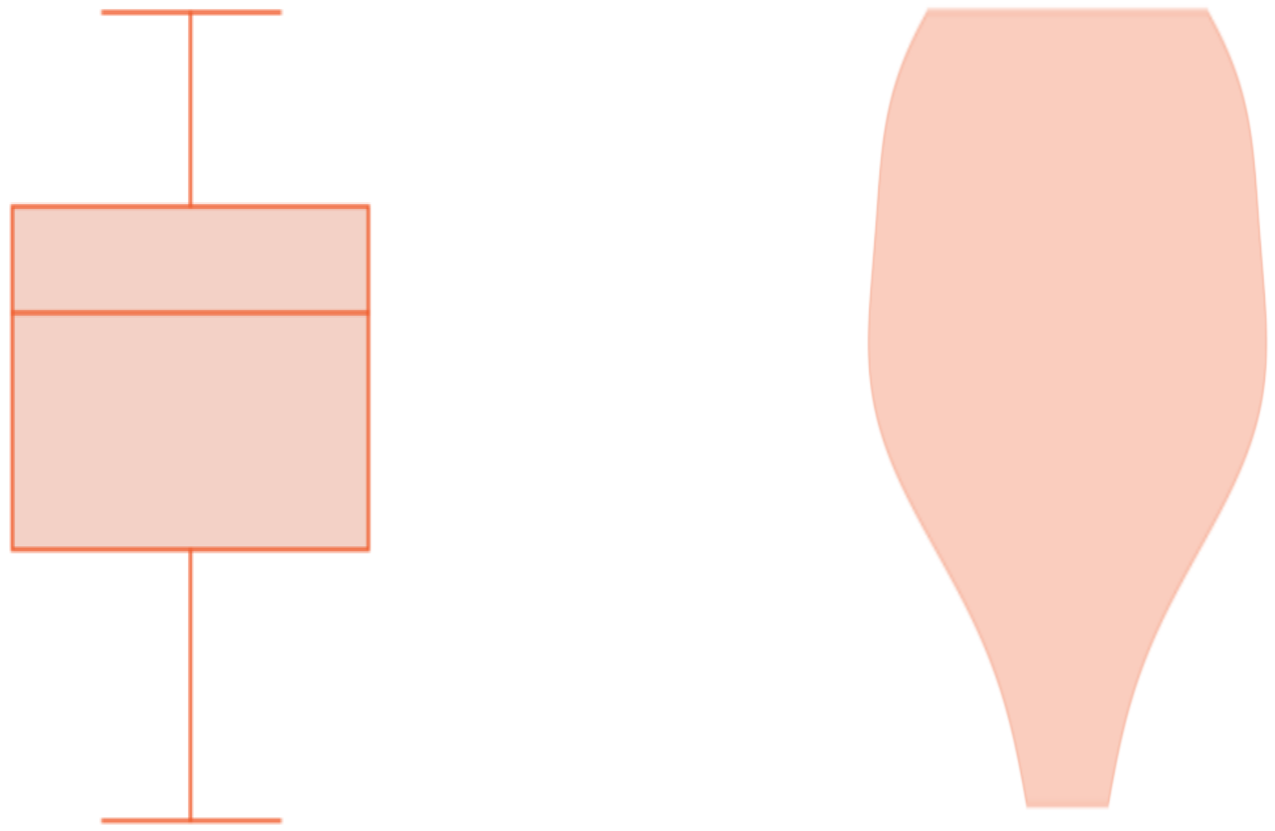
Minimum $Q_1$ - 1.5 IQR

Outliers Extreme if < $Q_1$ - 3 IQR

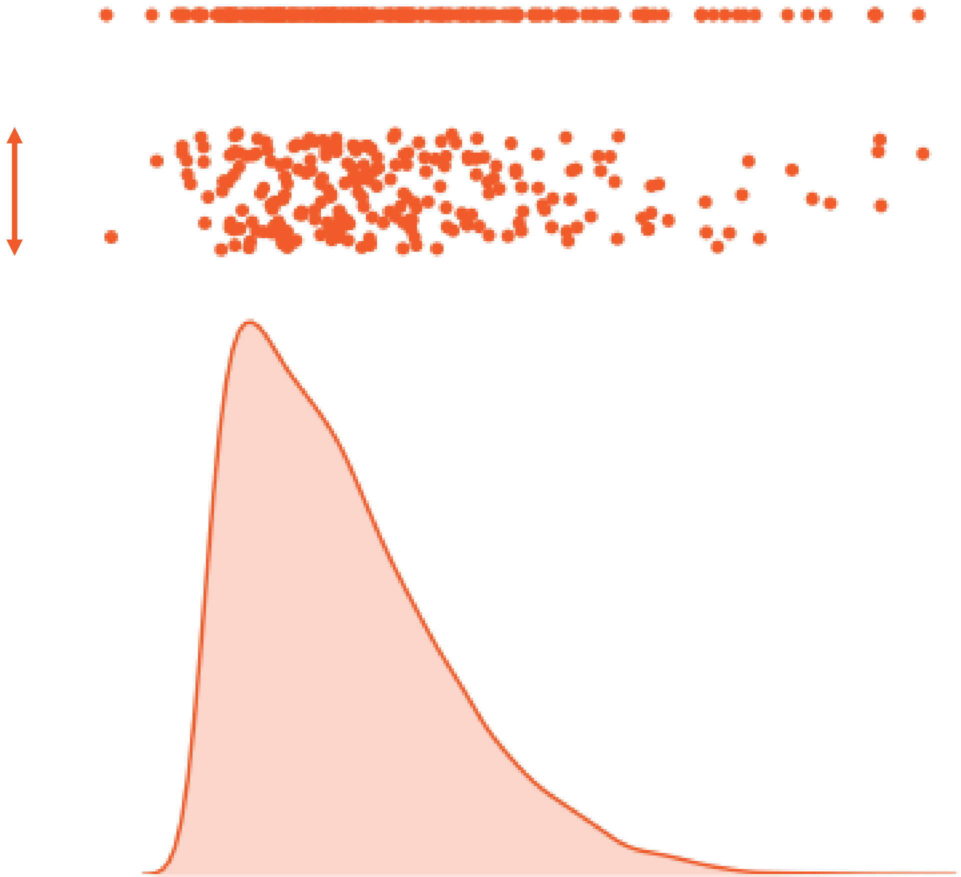
# Violin Plot



# Violin Plot

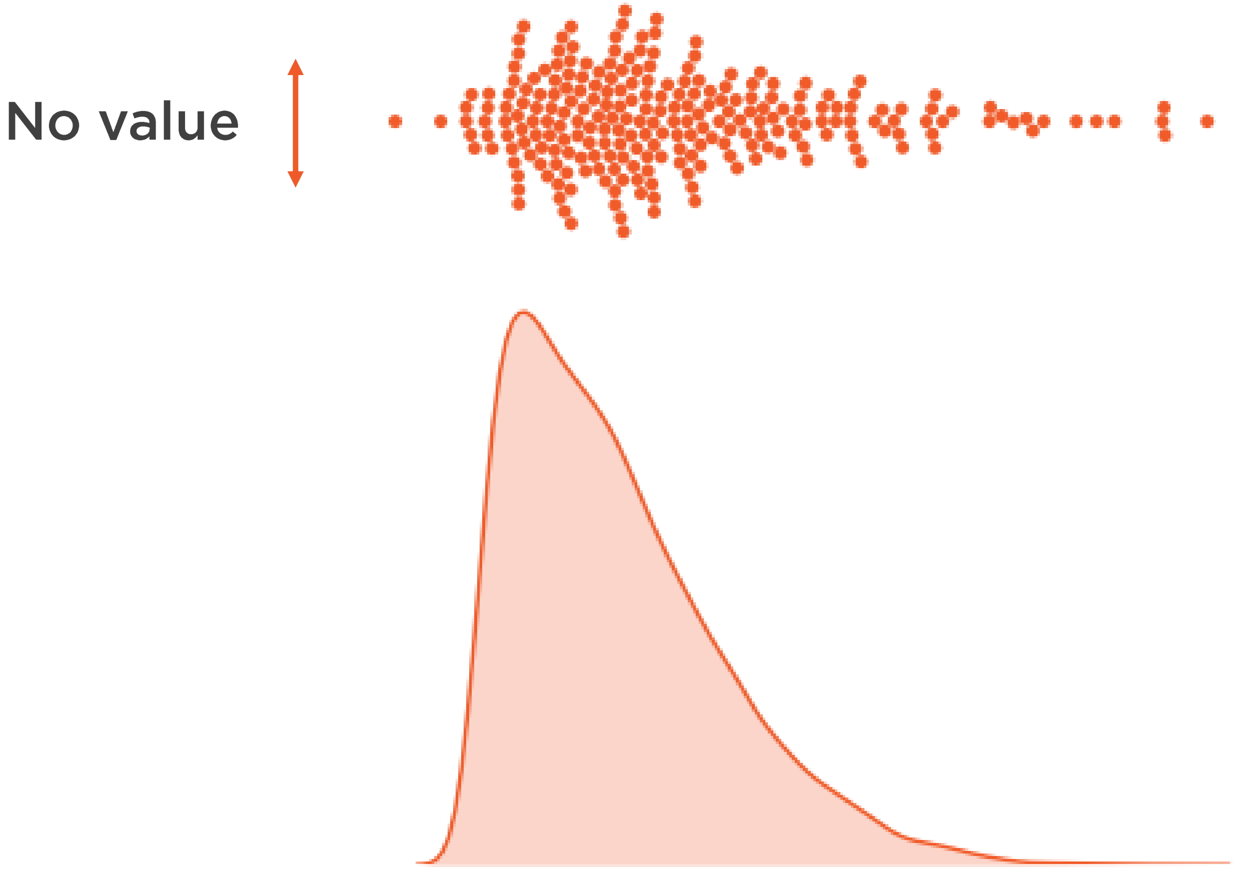


# Strip Plot



No value

# Swarm Plot



# Demo

Learn how to plot univariate distribution charts with Python

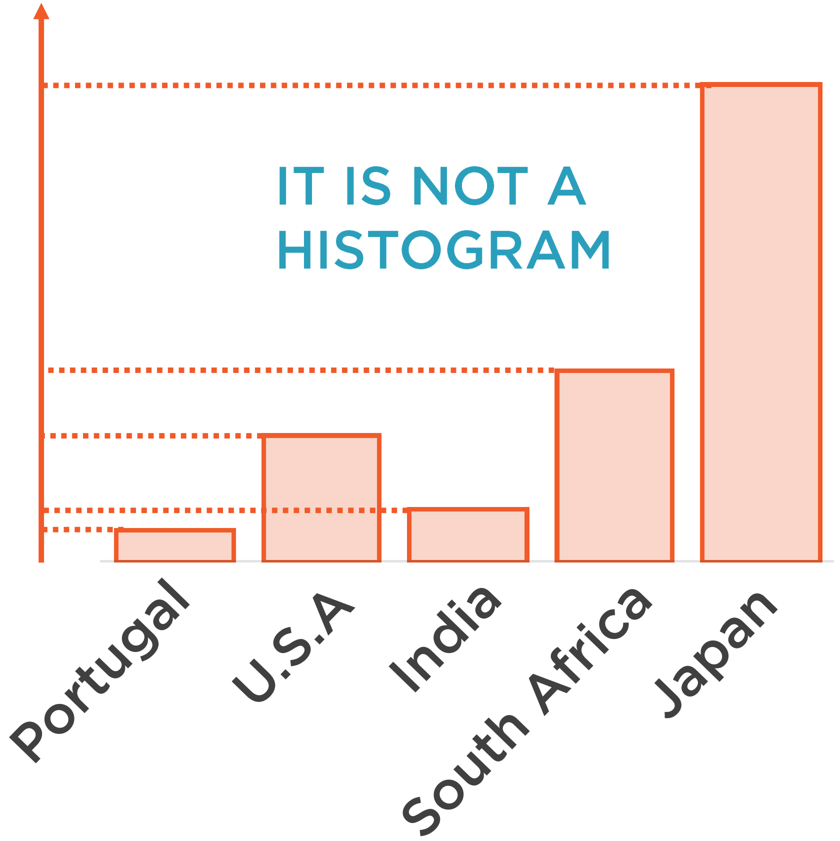
Using Python packages

- Matplotlib
- Seaborn

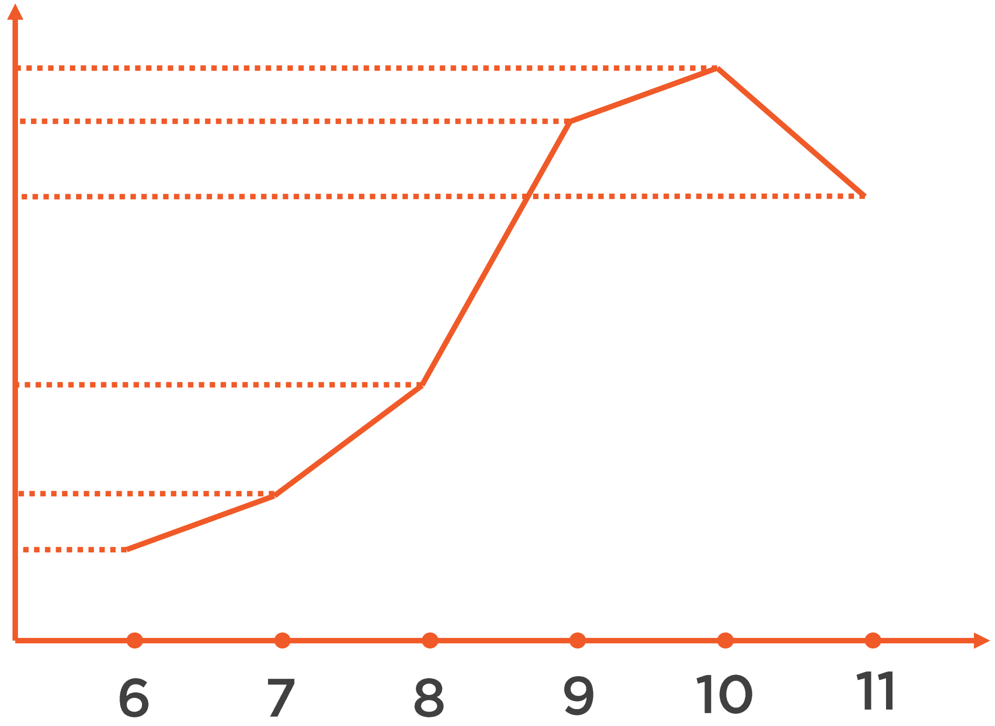
Learn how to customize some simple graph aesthetics

# Univariate Comparison Plots

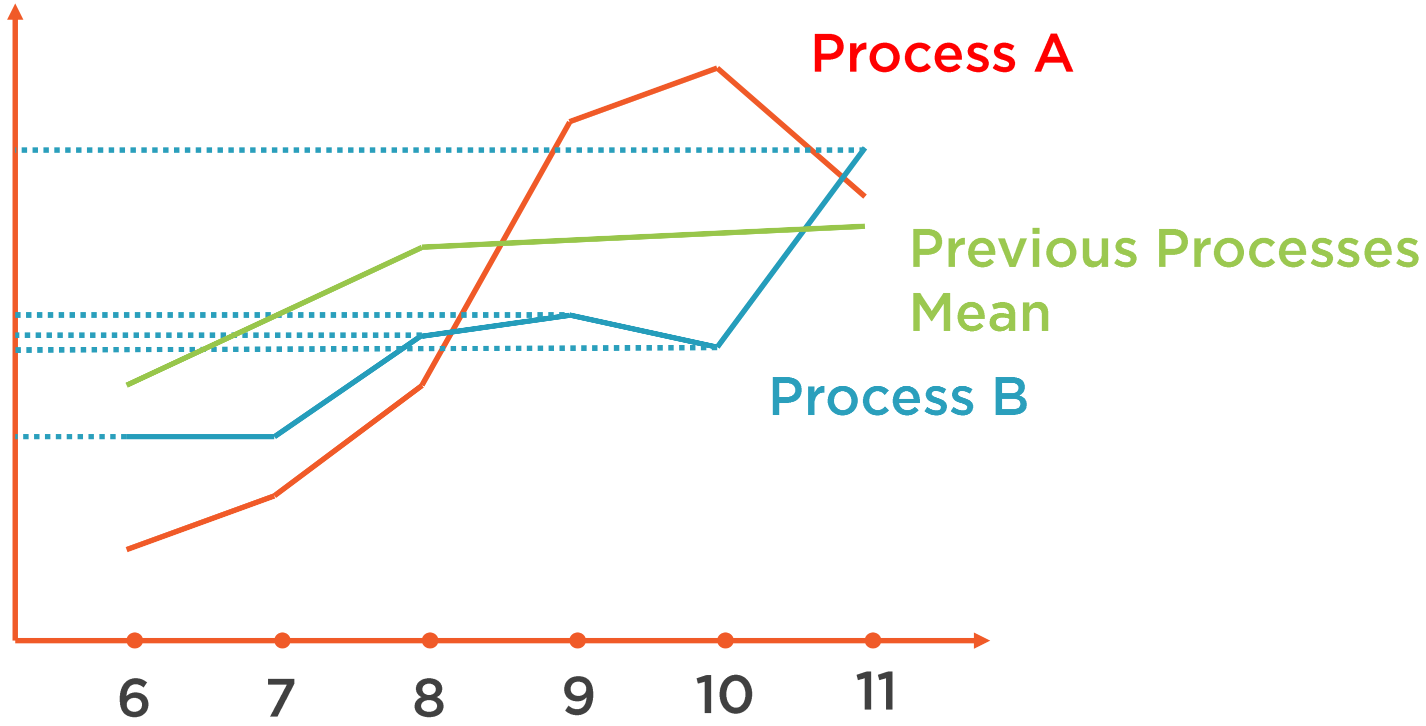
# Bar Diagram



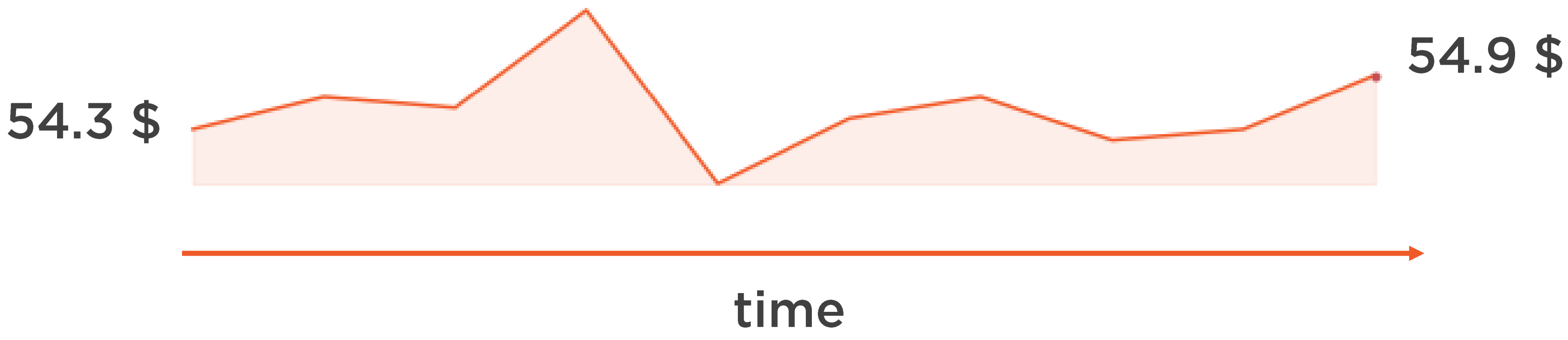
# Line Chart



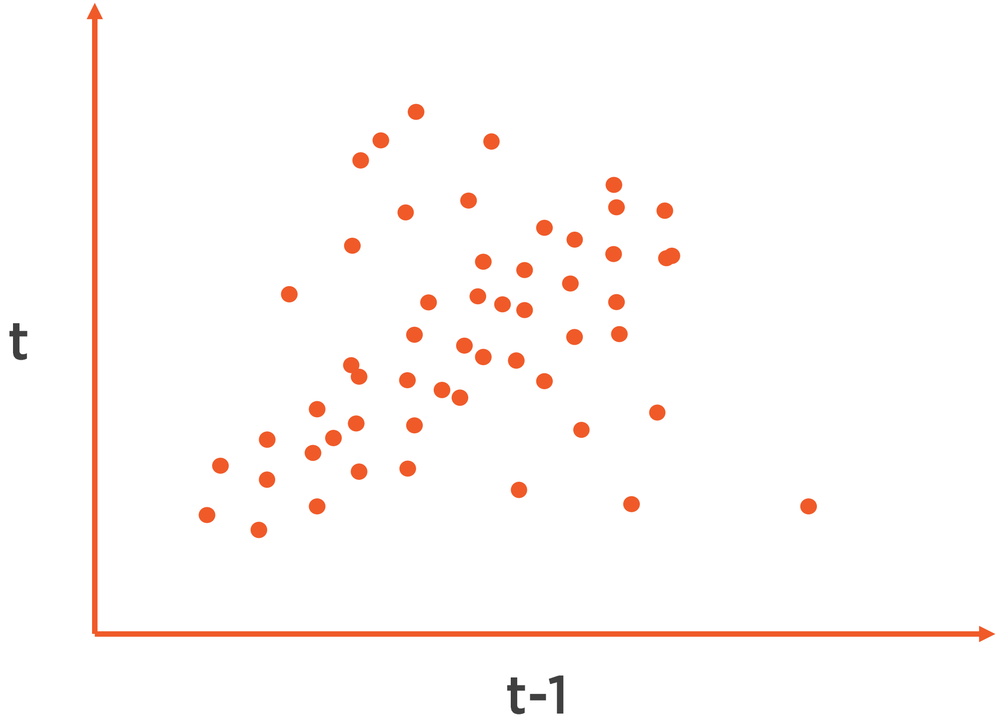
# Run Chart



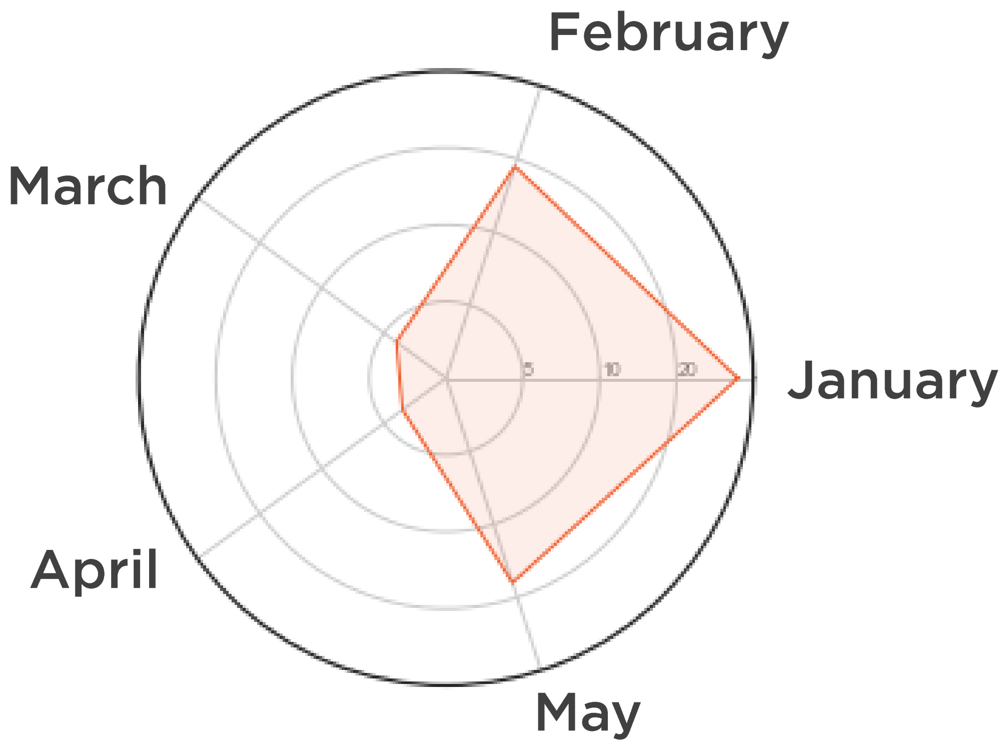
# Sparkline



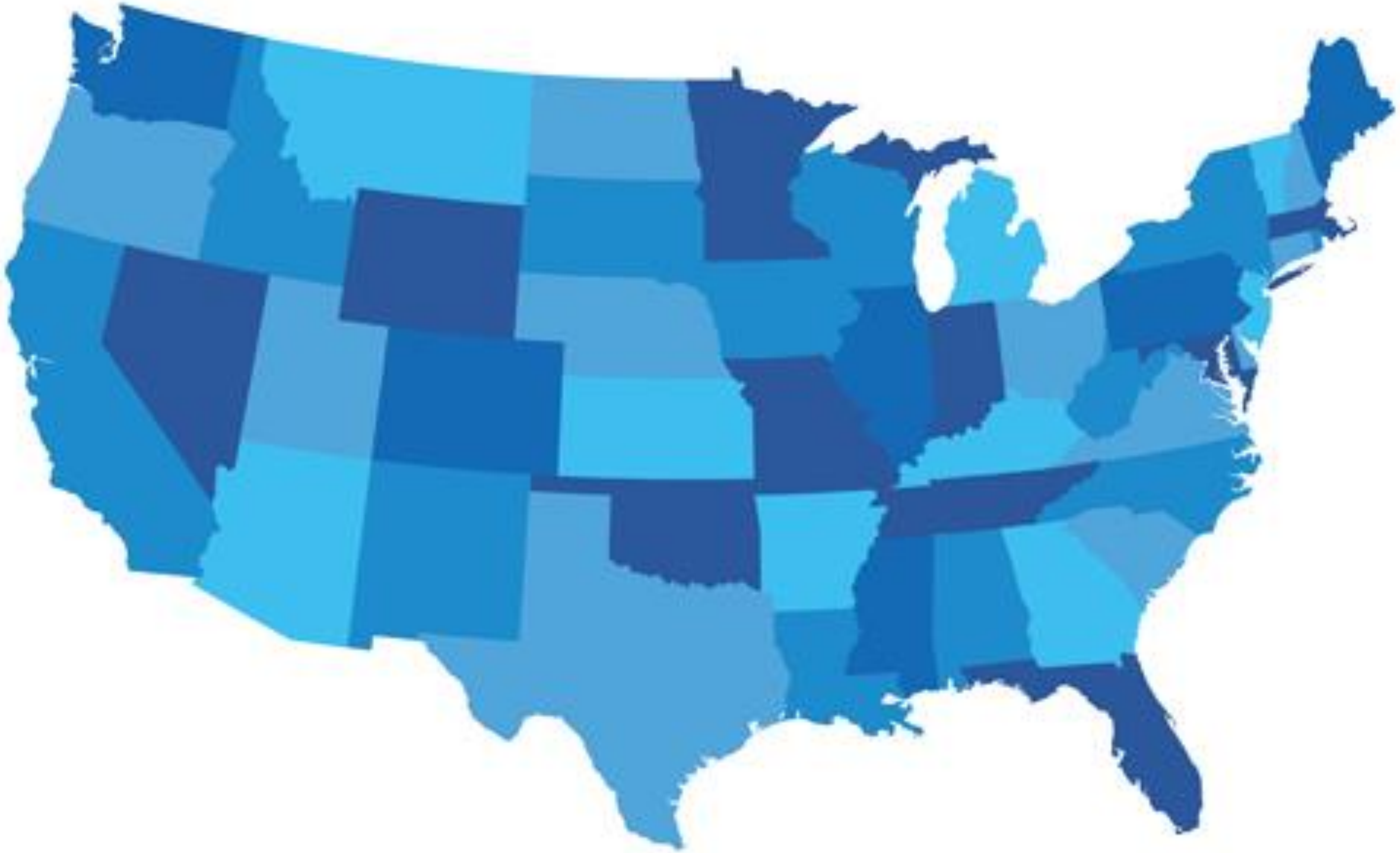
# Lag Plot



# Circular Area Chart



# Cartogram



Demo

**Learn how to plot univariate comparison charts with Python**

**Using Python packages**

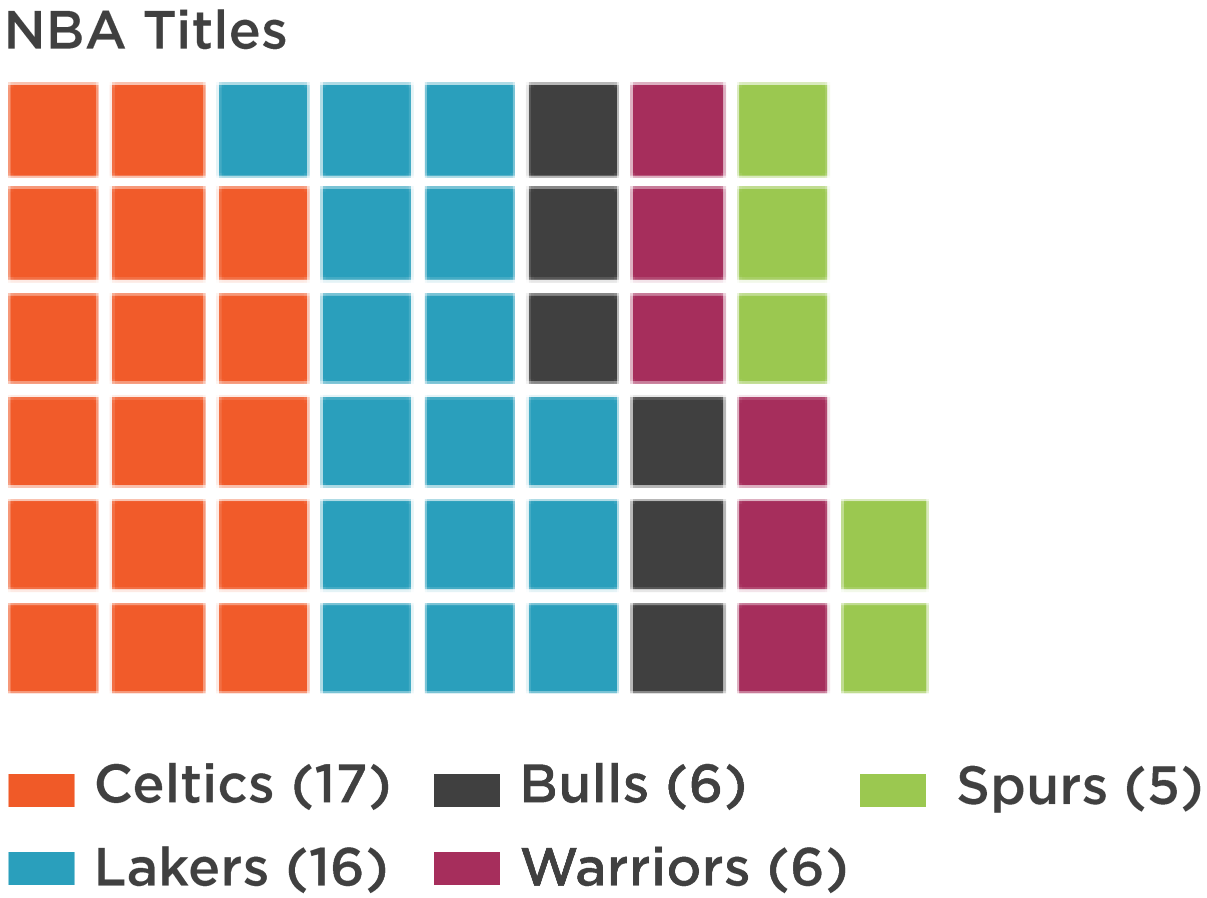
- Matplotlib
- Seaborn
- Pandas
- Geopandas
- Geoplot

**Learn how to customize some simple graph aesthetics**

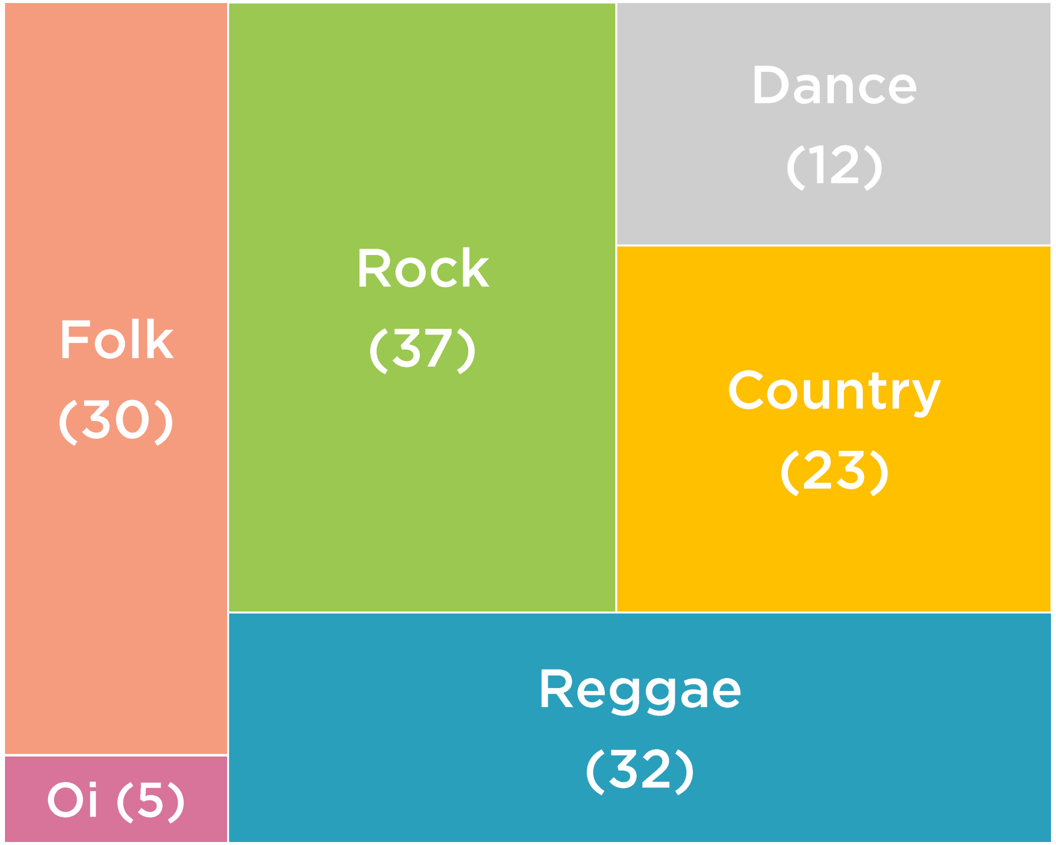
# Univariate Composition Plots

# Pie Chart

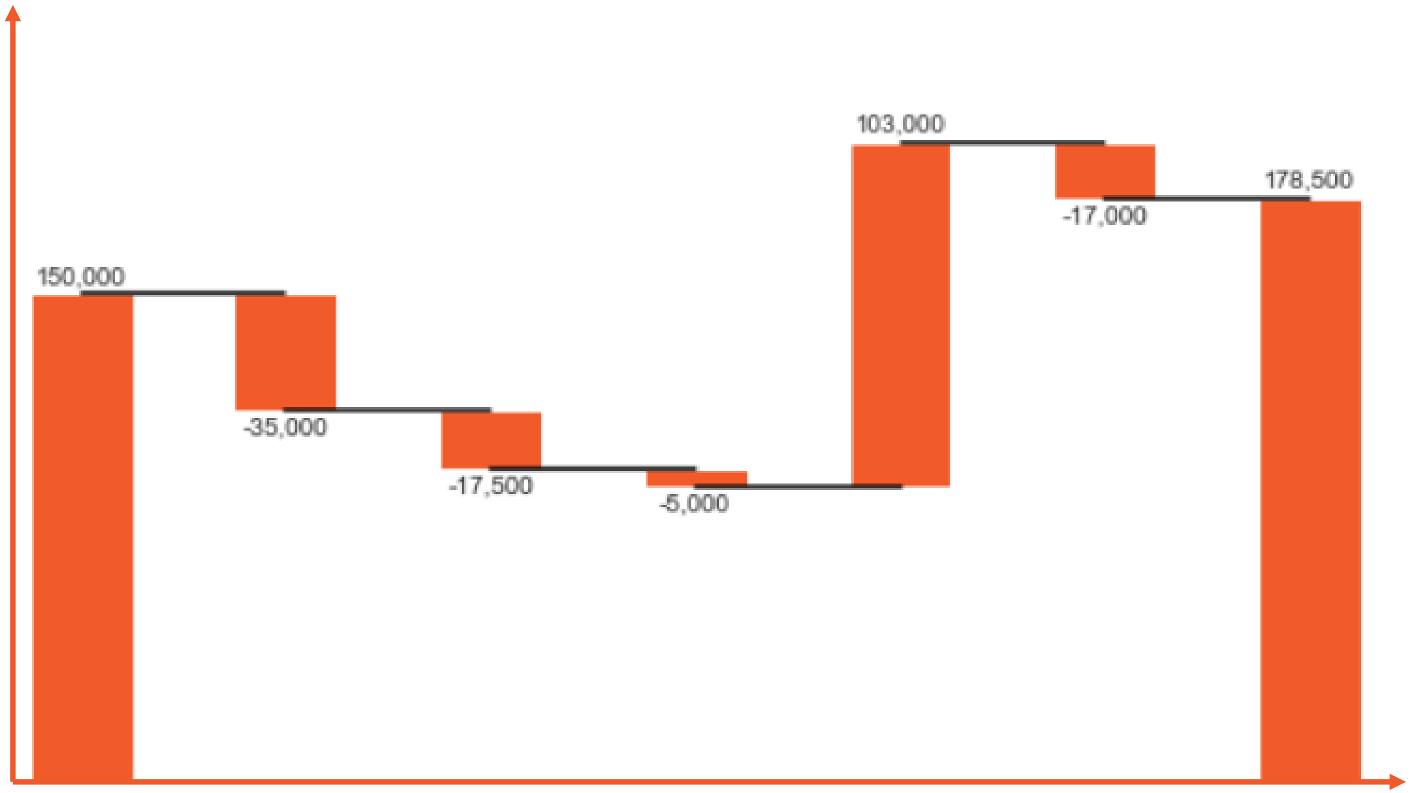
# Waffle Chart



# Tree Map



# Waterfall Chart



# Demo

**Learn how to plot univariate composition charts with Python**

**Using Python packages**

- Matplotlib
- Seaborn
- Pandas
- PyWaffle
- Squarify

**Learn how to customize some simple graph aesthetics**

# Univariate Analysis Tests

# Hypothesis Testing

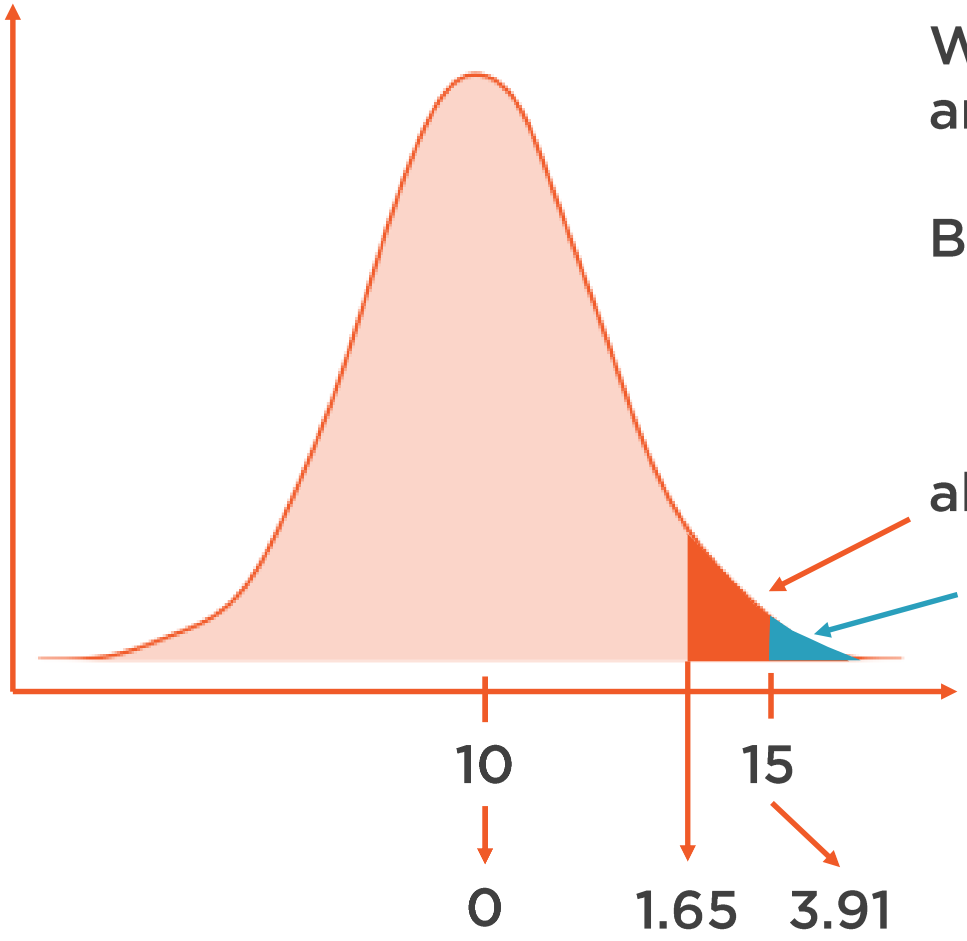
Formulate the null hypothesis (accepted fact)

State alternate hypothesis (chance)

State the rejection region (alpha level)

Test if the observed scenario is statistically significant

# Hypothesis Testing: T-test



We observe 30 pieces with mean size 15 and standard deviation 2

But we think they should have a mean of 10

alpha = 0.05

p-value

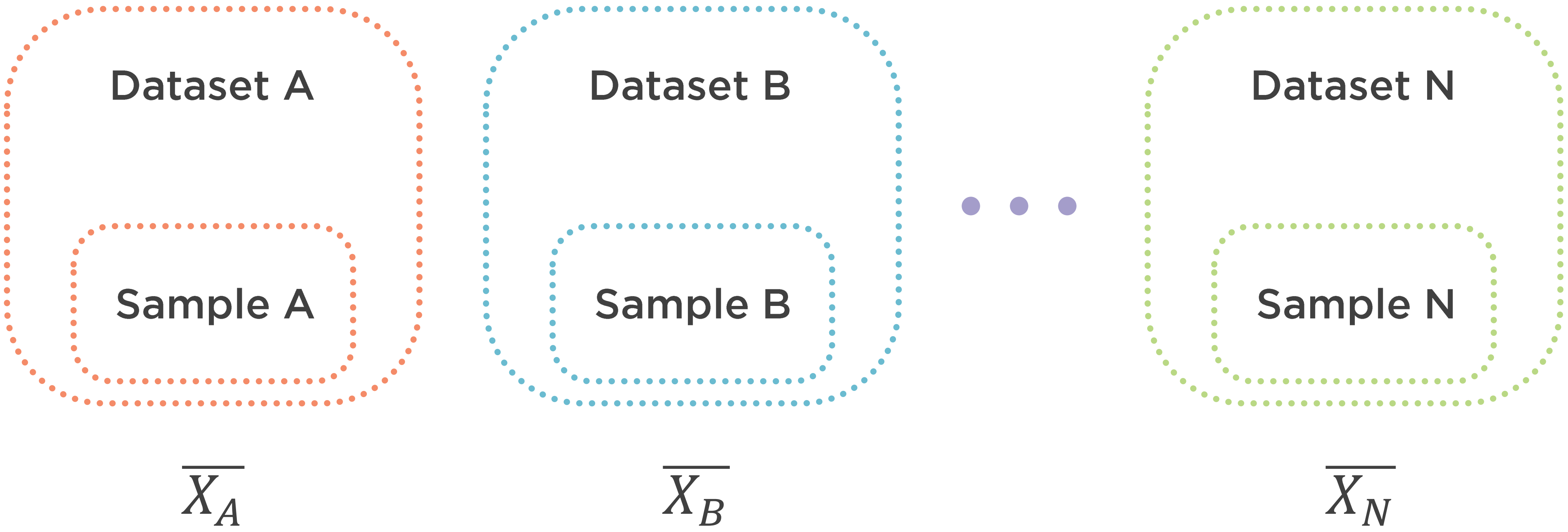
$H_0$: mean is 10
$H_1$: mean is > 10

$$Z = \frac{15 - 10}{2 / \sqrt{30}} = 3.91$$

# ANOVA – Analysis of Variance

H$_0$: $\mu_A = \mu_B = \cdots = \mu_N$

H$_1$: at least one mean is different from the others



# Assumptions

**Normality**

Datasets must behave with a normal distribution

**Homoscedasticity**

Variance of datasets should be homogeneous

**Independent Observations**

Datasets must be independent from each other

# Demo

Learn how to perfom a quick hypothesis and ANOVA tests

Using Python package

- Scipy