Review:

1. Mean, Median, Mode,
2. Variance and Standard deviation,
3. Python: Arrays, Conditionals, Loops

Today's agenda

1. Probability
2. Regression
3. Python: Dataframes, regression

Statistics

Probability:

The likelihood of something happening.

Probability is a branch of mathematic that study/calculate this likelihood. Simply

$$Probability\ of\ event\ "A" = \frac{Number\ of\ ways\ event\ "A"\ can\ happen}{total\ number\ of\ possible\ outcome}$$

When tossing a coin the probability of the coin landing on head or tail is:

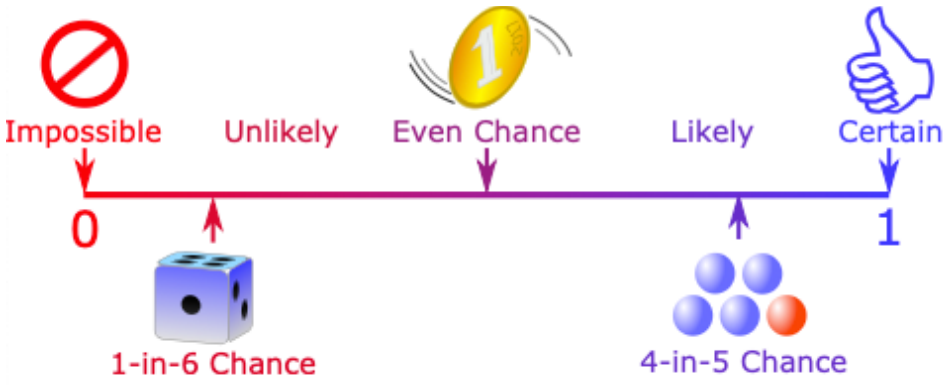
$$P(head) = \frac{1}{2} = 0.5$$

When Throwing a dice, the probability of it land on 6 is:

$$P(6) = \frac{1}{6}$$

Probability:

Give example of events with different probabilities,



Probability:

What is the probability of getting 2 same number when throwing a dice twice?

$$P(1\ and\ 1) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

What is the probability of getting same numbers when throwing 2 dices?

$$P(same\ numbers) = \frac{6}{36}$$

What is the probability the sum of the numbers on 2 dice to be 8?

$$P(add\ to\ 8) = \frac{5}{36}$$

Linear Regression

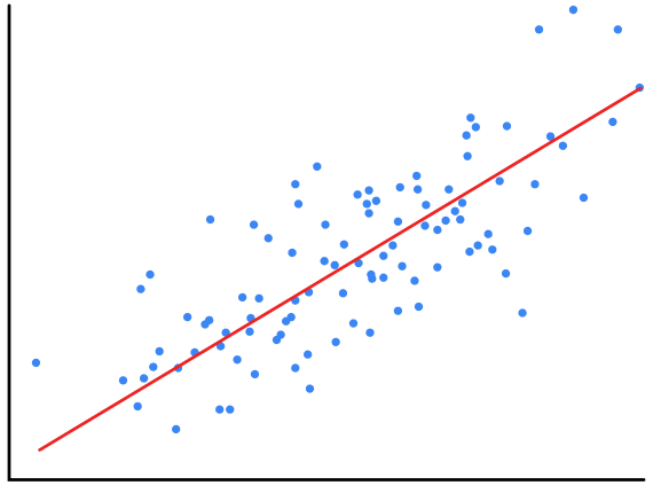
Linear Regression:

Is a linear relationship between a target and one or more variables (features).

A single line that best describe the relationship between the variable (or variables) and target.

If we have a single variable it is called “Simple Linear Regression”. If we have more than one variable it is called “Multiple Linear Regression”.

$$y = b_0 + b_1 x$$

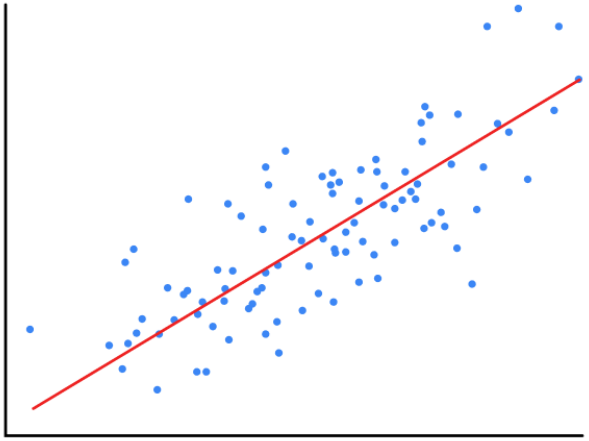


This is a line!

Linear Regression:

This relationship may not be true for a large set of datapoints!

$$y = b_0 + b_1 x + \epsilon$$



x is the independent variable ("feature" or simply the "variable")

y is the dependent variable ("output" or "target")

b_0 is the intercept

b_1 is the slope

ϵ is the error term

Linear Regression:

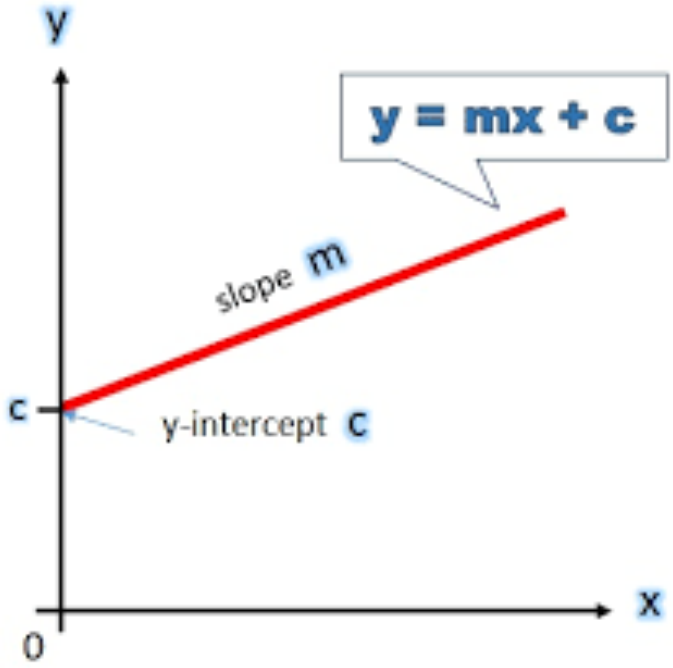
As the name suggest linear regression explain the relationship simply by a line.

A line need an intercept and a slope.

$$y = b_0 + b_1 x$$

b0 is the intercept and

b1 is the slope



Interactive line:
https://www.desmos.com/calculator/59qdbtnlzy

Linear Regression:

Regression model estimates the nature of the relationship between the independent and dependent variables.

- Change in dependent variables that results from changes in independent variables.
- Strength of the relationship.
- Statistical significance of the relationship.

Examples for simple linear regression:

1. Relationship between the Square footage and price of the house, discuss?
2. Relationship between the age and price of the house, discuss?
3. Relationship between oil production and its price, discuss?

Linear Regression:

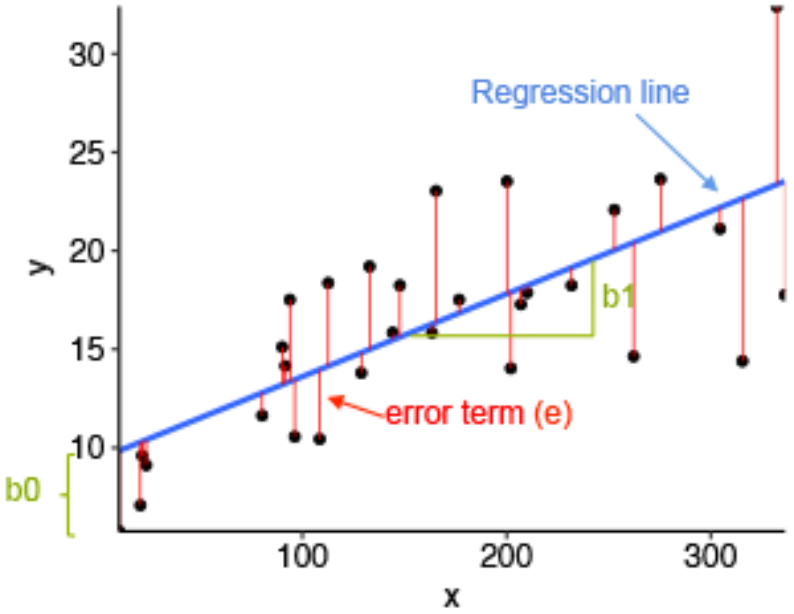
Residuals:

After we fit the line through the data points, it is highly likely that not all datapoints fall exactly on the line. There will be scattered all around the line (below or above the line), so as a results there is a vertical distance between the line and the datapoint, this distance is referred to as residual.

$$\epsilon = y - \hat{y}$$

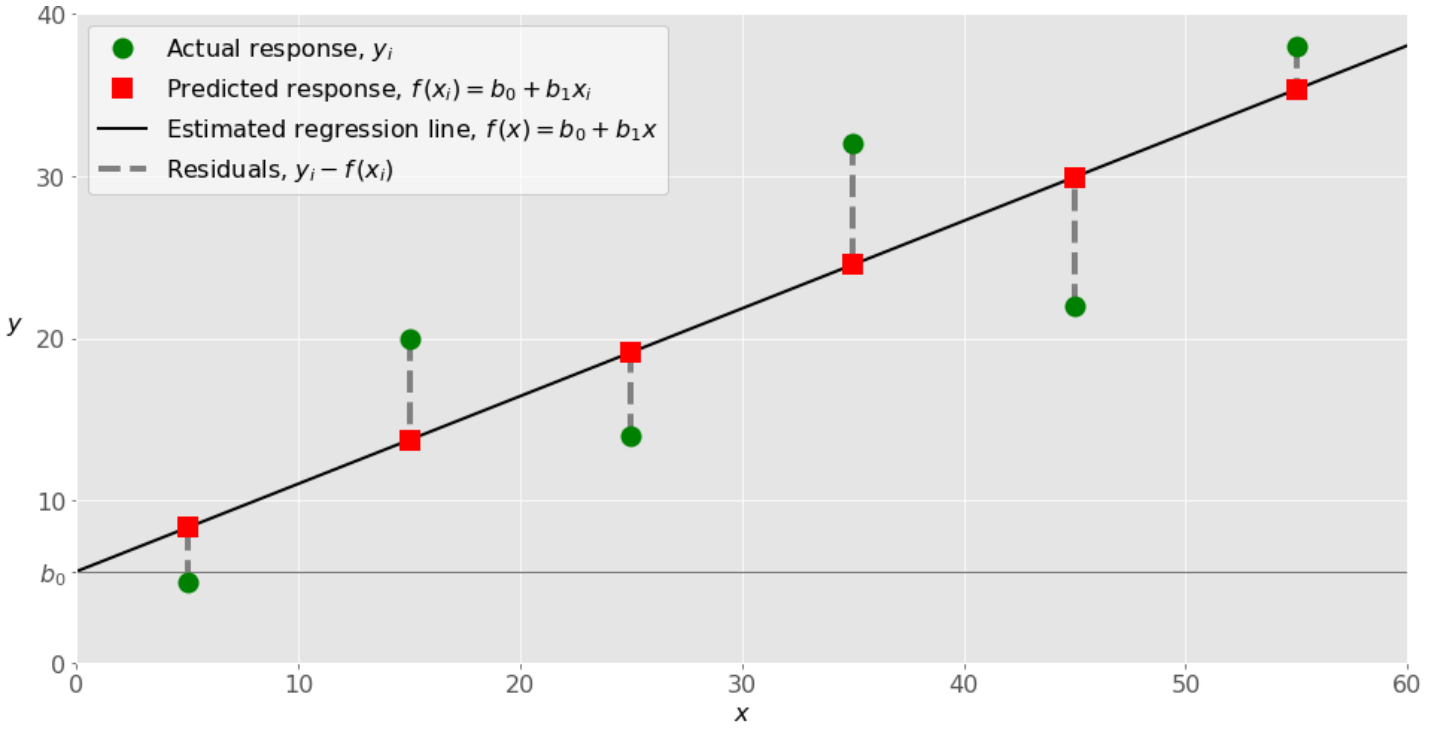
y observed value (true value)

$\hat{y}$ Predicted value (by the line)



Linear Regression:

Residuals:



Linear Regression:

How linear regression finds the best line?

There are several methods/formulation, one mostly used is ordinary least squares (OLS).

Least-square:

Minimize the sum of squared residuals (difference between the actual and predicted values)

$$\epsilon = y - \hat{y}$$

$$y = b_0 + b_1 x + \epsilon \;\Rightarrow\; \epsilon = y - b_0 - b_1 x$$

Lets say we have *n* datapoints, OLS find the minimum $f(b_0, b_1)$, where

$$f(b_0, b_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y - \hat{y})^2 = \sum_{i=1}^{n} (y - b_0 - b_1 x)^2$$

This happens where we "**Fit**" a model! (you will see more on this!)

Linear Regression: Evaluation

We say a model fits the data very well if the differences between the observed values and the model's predicted values are very small and not biased.

R-squared (Coefficient of determination):

R-squared measure how close are the observed vales to the fitted line.

R-squared = Explained variation / Total variation

R-squared always take a value between 0% and 100%,

0% indicates a very bad fit (the fitted model explain none of the variability of data around its mean).

100% indicates that the model gives a perfect fit of the data (the fitted model explain all of the variability of data around its mean).

Linear Regression: Evaluation (R-squared) cont.

$$S_{tot} = \sum_{i=1}^{n} (\mu - \hat{y})^2$$

$$S_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

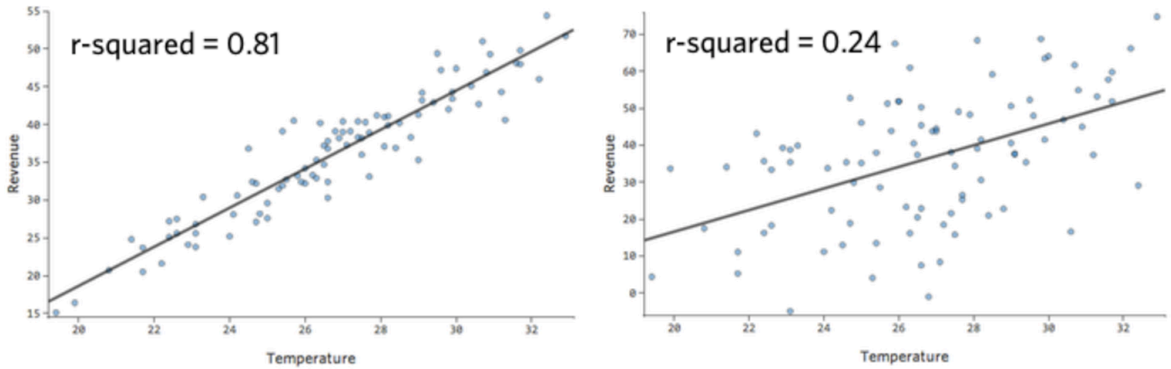
$$R^2 = 1 - \frac{S_{res}}{S_{tot}}$$

Where μ is the mean

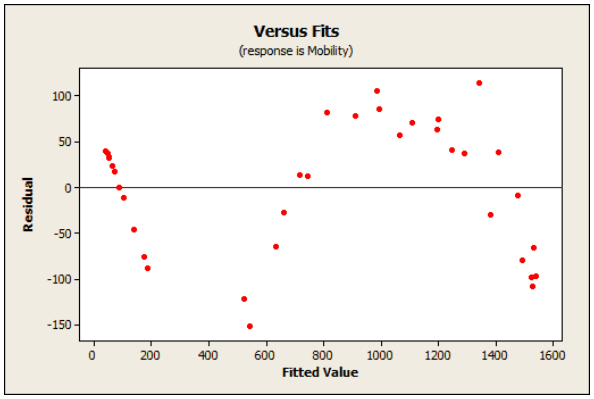
y observed value (true value)

$\hat{y}$ Predicted value (by the line)

Linear Regression: Evaluation

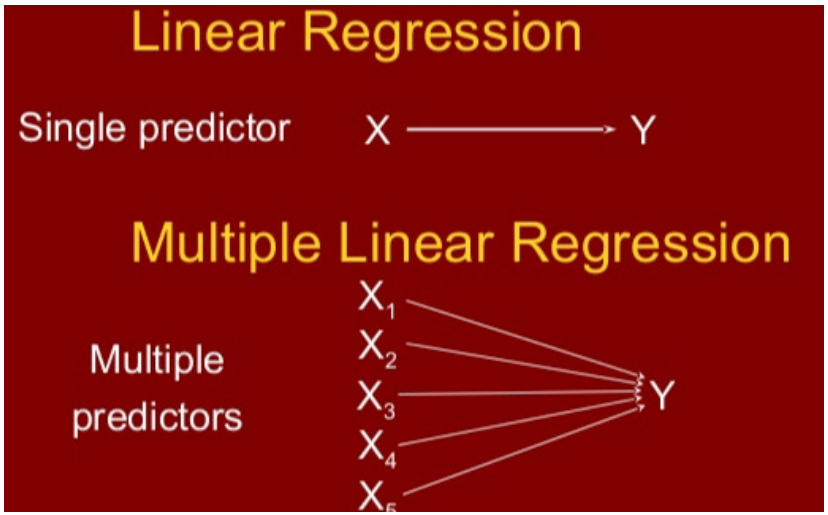


High R-squared not always means good fit, always check the residuals and fit plots!



Linear Regression:

Simple vs multiple linear Regression:



- **Simple Linear Regression – one independent variable.**

$$y = b_0 + b_1x_1$$

- **Multiple Linear Regression – multiple independent variables.**

$$y = b_0 + b_1x_1 + b_2x_2 \ldots + b_nx_n$$

2^{nd} independent variable and weight (coefficient)

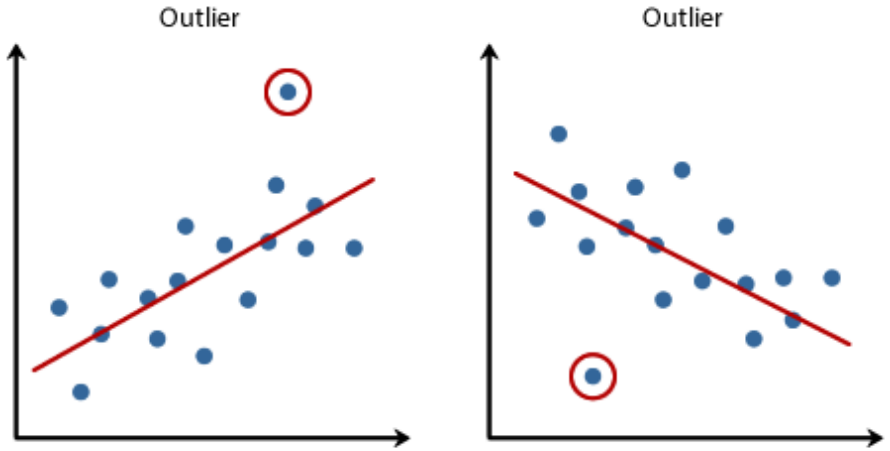
n^{th} independent variable and weight (coefficient)

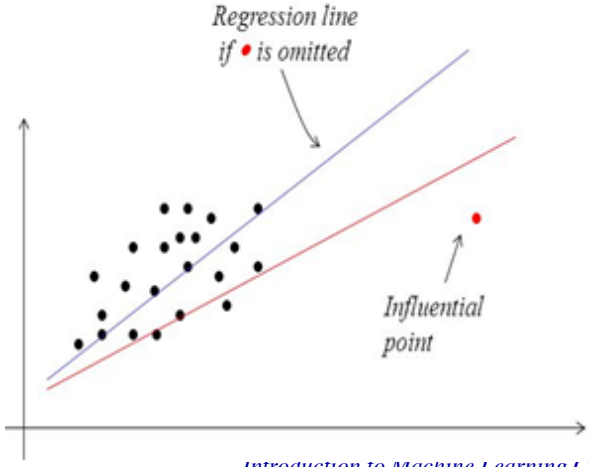
Linear Regression: Evaluation, outliers

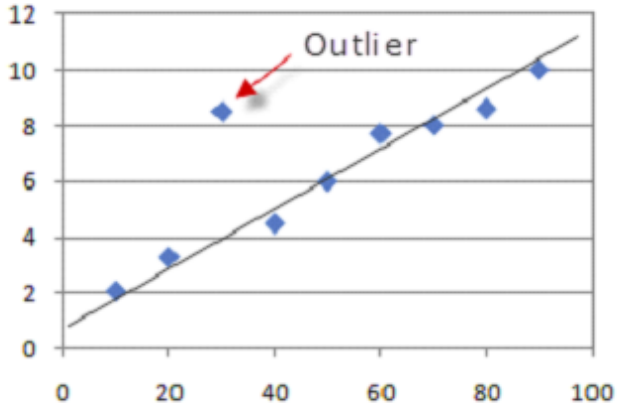
Outliers:

Datapoint that are far away from the rest of the datapoints.

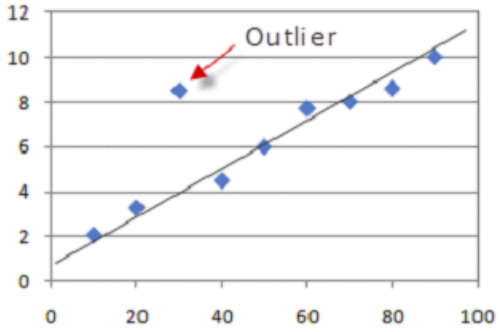
They are unusual values and may cause that the model deviate from reality and lead to wrong conclusions!







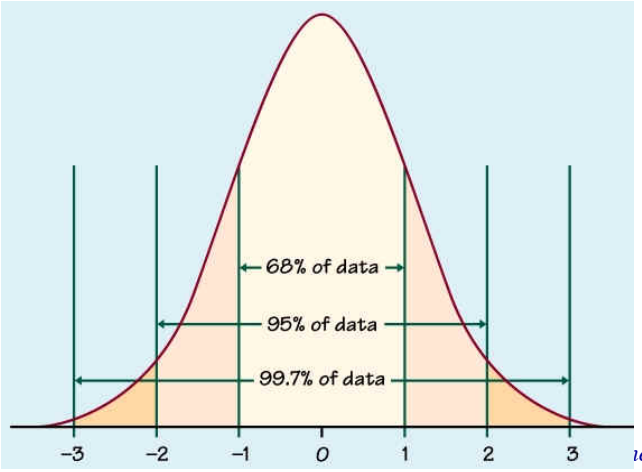
Linear Regression: Evaluation, outliers

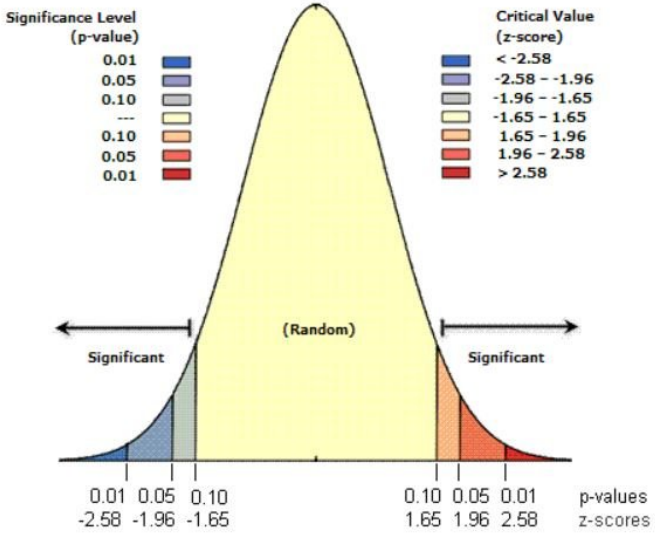


How to detect outliers:

1. Sort your dataset, and look for the min and max values
2. Visualise your dataset (scatter plot, Boxplot…)
3. Uses Z-scores: this is the number of standard deviation above or below the mean

$$z = \frac{x - \mu}{\sigma}$$





Linear Regression: Evaluation, outliers

How to detect outliers:

4. Outlier Fences using the Interquartile Range
5. Hypothesis Tests
6. Your knowledge of the database is the key

Python:

Pandas Dataframes:

From pandas: “Two-dimensional, size-mutable, potentially heterogeneous tabular data.”

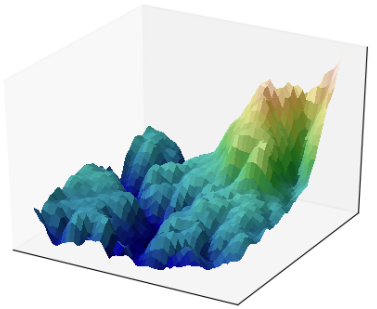
- Possible to change
- Different data types

https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html

https://www.earthdatascience.org/courses/intro-to-earth-data-science/scientific-data-structures-python/pandas-dataframes/

Python:

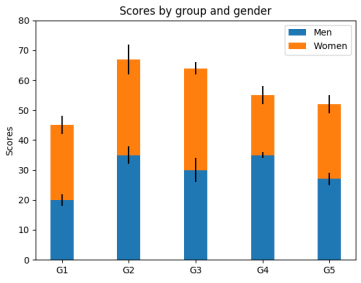
matplotlib

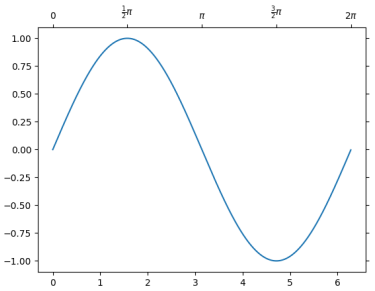


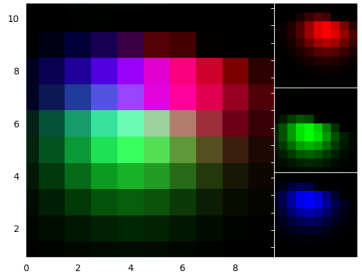
Matplotlib:

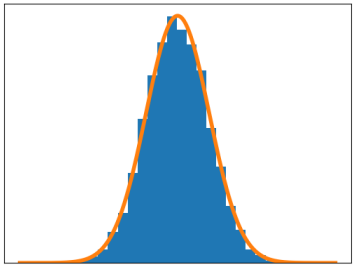
Matplotlib is a python plotting library for creating high quality 2D/3D figures.

https://matplotlib.org









Python:

Matplotlib:

https://matplotlib.org

Colors: https://matplotlib.org/tutorials/colors/colors.html

Markers: https://matplotlib.org/3.1.3/api/markers_api.html

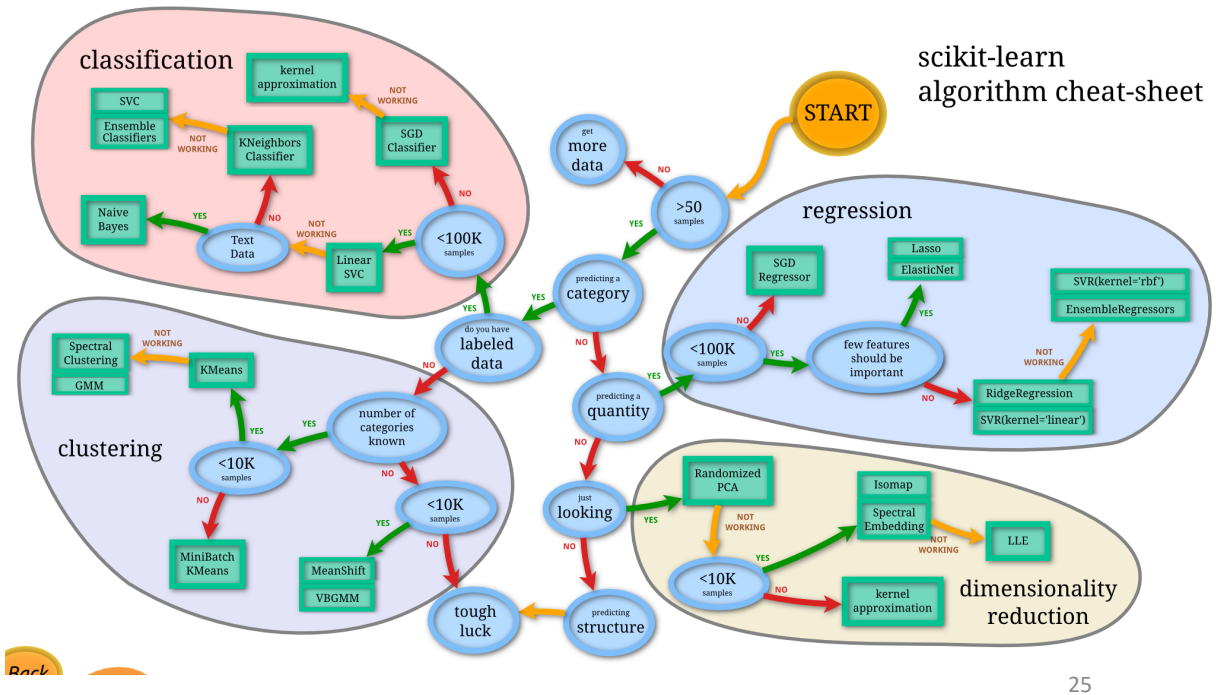
Linestyles: https://matplotlib.org/3.1.0/gallery/lines_bars_and_markers/linestyles.html

https://matplotlib.org/2.1.1/api/_as_gen/matplotlib.pyplot.plot.html

Python:

Scikit learn:

A free library in Python that has great features for clustering, classification, regression and dimensionality reduction.



Python:

Scikit learn: Linear Regression

For almost every model we do, we follow these:

Sklearn docs:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html