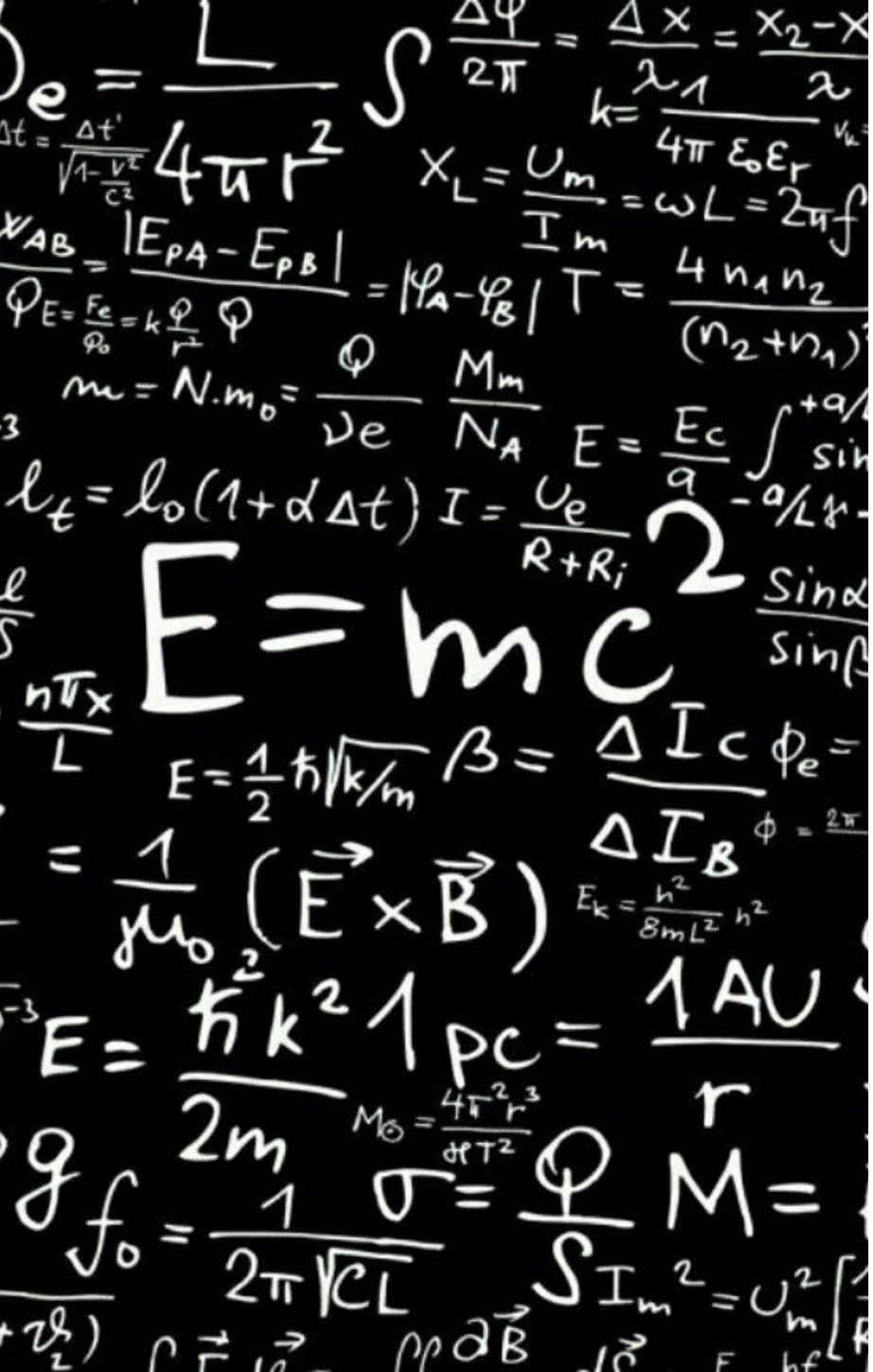
Day 1-2

- Day 1 Morning : Introduction to Machine Learning
- Day 1 Afternoon : Data Preparation with Python (Refresh)
- Day 2 Morning : Linear and Logistic Regression
- Day 2 Afternoon : Decision Trees and Random Forest, Nearest Neighbor



MACHINE LEARNING FIRST LOOK

Introducing Iris

- Iris is the name of a genus of flowers.
- There are 260-300 species of Iris.
- Can a machine classify the species of Iris, given a dataset?

Iris Dataset

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5	2	3.5	1	versicolor
3	6	2.2	4	1	versicolor
4	6	2.2	5	1.5	virginica
5	6.2	2.2	4.5	1.5	versicolor
6	4.5	2.3	1.3	0.3	setosa
7	5	2.3	3.3	1	versicolor
8	5.5	2.3	4	1.3	versicolor
9	6.3	2.3	4.4	1.3	versicolor
10	4.9	2.4	3.3	1	versicolor
11	5.5	2.4	3.8	1.1	versicolor
12	5.5	2.4	3.7	1	versicolor
13	4.9	2.5	4.5	1.7	virginica
14	5.1	2.5	3	1.1	versicolor
15	5.5	2.5	4	1.3	versicolor
16	5.6	2.5	3.9	1.1	versicolor

Attributes

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5	2	3.5	1	versicolor
3	6	2.2	4	1	versicolor
4	6	2.2	5	1.5	virginica
5	6.2	2.2	4.5	1.5	versicolor
6	4.5	2.3	1.3	0.3	setosa
7	5	2.3	3.3	1	versicolor
8	5.5	2.3	4	1.3	versicolor
9	6.3	2.3	4.4	1.3	versicolor
10	4.9	2.4	3.3	1	versicolor
11	5.5	2.4	3.8	1.1	versicolor
12	5.5	2.4	3.7	1	versicolor
13	4.9	2.5	4.5	1.7	virginica
14	5.1	2.5	3	1.1	versicolor
15	5.5	2.5	4	1.3	versicolor
16	5.6	2.5	3.9	1.1	versicolor

- Columns of the dataset
- It describes characteristics of each data.
- There are 2 types of attributes.
  - Numeric
  - Categorical

Numeric Attributes

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5	2	3.5	1	versicolor
3	6	2.2	4	1	versicolor
4	6	2.2	5	1.5	virginica
5	6.2	2.2	4.5	1.5	versicolor
6	4.5	2.3	1.3	0.3	setosa
7	5	2.3	3.3	1	versicolor
8	5.5	2.3	4	1.3	versicolor
9	6.3	2.3	4.4	1.3	versicolor
10	4.9	2.4	3.3	1	versicolor
11	5.5	2.4	3.8	1.1	versicolor
12	5.5	2.4	3.7	1	versicolor
13	4.9	2.5	4.5	1.7	virginica
14	5.1	2.5	3	1.1	versicolor
15	5.5	2.5	4	1.3	versicolor
16	5.6	2.5	3.9	1.1	versicolor

- Attributes that have numeric values.
  - aka numbers
- For example, this attribute contains information about sepal width of the flower.
- It contains numeric values of the width of the sepal.

Categorical Attributes

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5	2	3.5	1	versicolor
3	6	2.2	4	1	versicolor
4	6	2.2	5	1.5	virginica
5	6.2	2.2	4.5	1.5	versicolor
6	4.5	2.3	1.3	0.3	setosa
7	5	2.3	3.3	1	versicolor
8	5.5	2.3	4	1.3	versicolor
9	6.3	2.3	4.4	1.3	versicolor
10	4.9	2.4	3.3	1	versicolor
11	5.5	2.4	3.8	1.1	versicolor
12	5.5	2.4	3.7	1	versicolor
13	4.9	2.5	4.5	1.7	virginica
14	5.1	2.5	3	1.1	versicolor
15	5.5	2.5	4	1.3	versicolor
16	5.6	2.5	3.9	1.1	versicolor

- Attributes which values are categorical
  - aka discrete value
- For example, this attribute describes the species of each data point.
  - There are 3 groups (categories) of Iris: versicolor, virginica, and setosa.

Class Targets

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5	2	3.5	1	versicolor
3	6	2.2	4	1	versicolor
4	6	2.2	5	1.5	virginica
5	6.2	2.2	4.5	1.5	versicolor
6	4.5	2.3	1.3	0.3	setosa
7	5	2.3	3.3	1	versicolor
8	5.5	2.3	4	1.3	versicolor
9	6.3	2.3	4.4	1.3	versicolor
10	4.9	2.4	3.3	1	versicolor
11	5.5	2.4	3.8	1.1	versicolor
12	5.5	2.4	3.7	1	versicolor
13	4.9	2.5	4.5	1.7	virginica
14	5.1	2.5	3	1.1	versicolor
15	5.5	2.5	4	1.3	versicolor
16	5.6	2.5	3.9	1.1	versicolor

- Attributes which classify the data.
  - It tells the class of each data point.

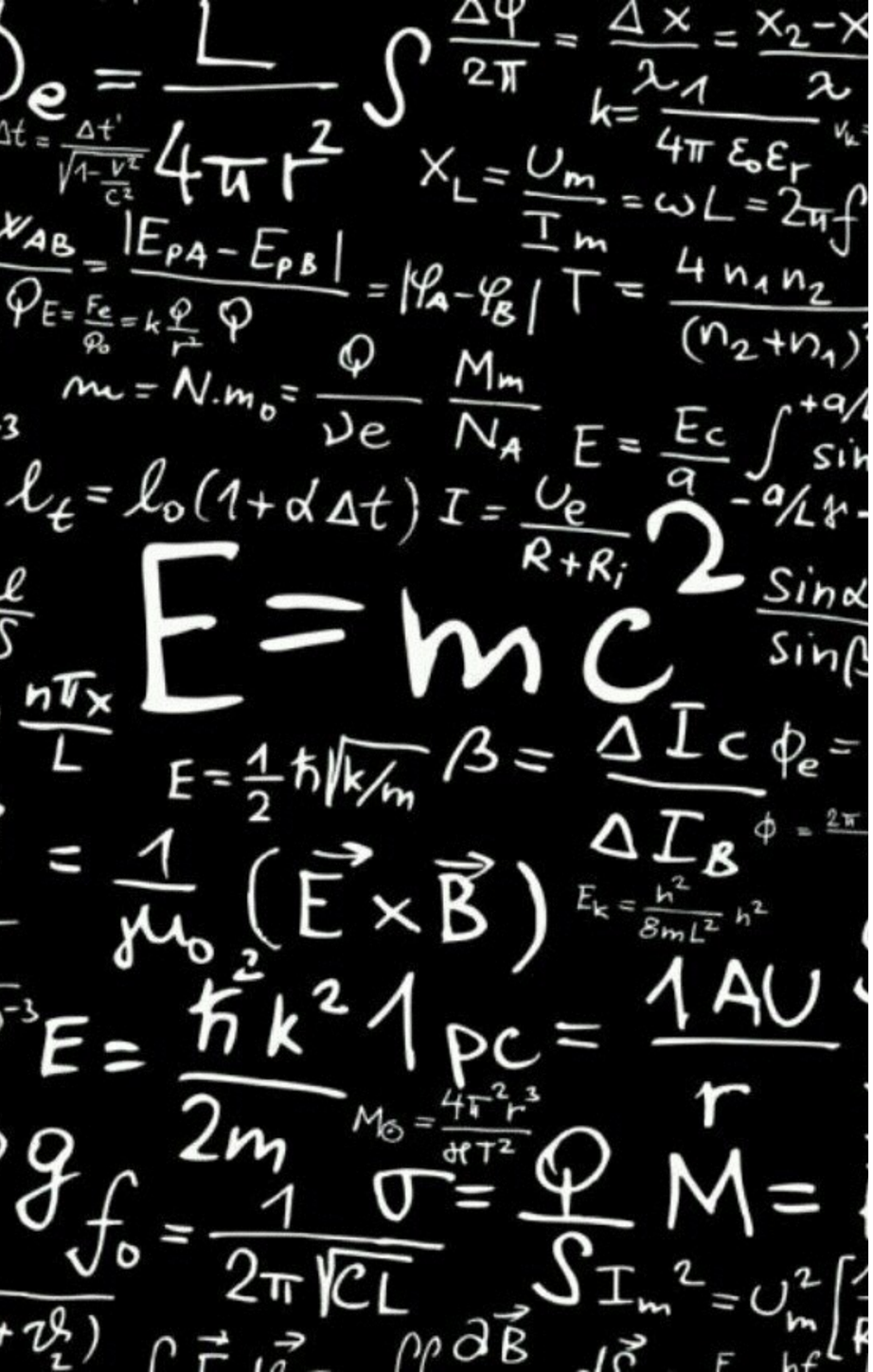
Instances

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5	2	3.5	1	versicolor
3	6	2.2	4	1	versicolor
4	6	2.2	5	1.5	virginica
5	6.2	2.2	4.5	1.5	versicolor
6	4.5	2.3	1.3	0.3	setosa
7	5	2.3	3.3	1	versicolor
8	5.5	2.3	4	1.3	versicolor
9	6.3	2.3	4.4	1.3	versicolor
10	4.9	2.4	3.3	1	versicolor
11	5.5	2.4	3.8	1.1	versicolor
12	5.5	2.4	3.7	1	versicolor
13	4.9	2.5	4.5	1.7	virginica
14	5.1	2.5	3	1.1	versicolor
15	5.5	2.5	4	1.3	versicolor
16	5.6	2.5	3.9	1.1	versicolor

- Rows of the dataset.
  - Each row describes one data point.
  - It tells the class of each data point.

Learning Species

- Given the dataset, can a machine learn the species of Iris?
- Lab 1: First look at Machine Learning



DATA PREPROCESSING

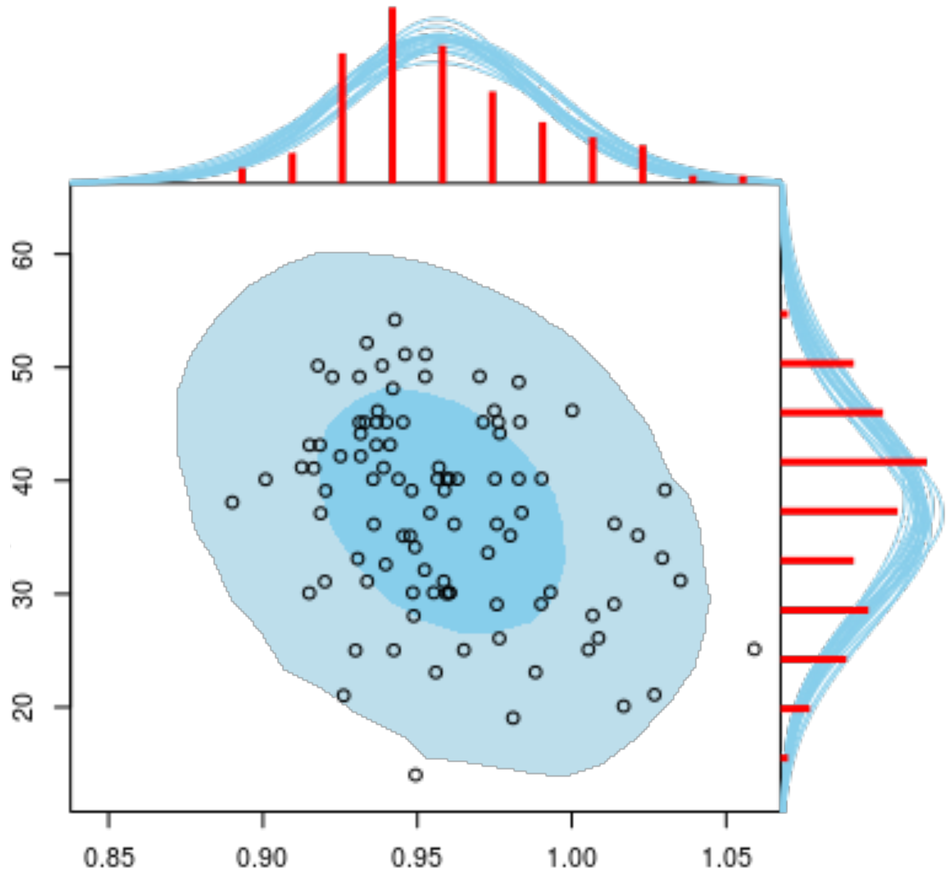
DATA PREPROCESSING

- **Data Cleaning**: Real-world data is dirty – incomplete and inconsistent. We will need to perform cleaning operations such as filling in missing values, removing some outliers.
- **Data Integration**: We might need to pull data from several sources and make sure they are consistent.
- **Data Transformation**: We might need to transform data to fit machine learning model assumption such as scaling and standardizing.
- **Feature Extraction**: We need to transform data to help our machine learning model to learn from it, such as image processing.

MODEL ASSUMPTION

It’s important to be aware of model assumptions, when you use them.

- For example, regression has the following assumptions:
  - X and Y variables are numeric
  - X and Y variables have normal distributions
  - All variables must have the same variance



DATA PREPROCESSING

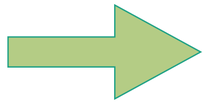
In this lecture, we will cover the most basic data preprocessing work that you will need to do in most cases:

- One-hot encoding
- Feature scaling
- Dealing with skewed data
- Dealing with missing data
- Removing outliers

ONE-HOT ENCODING

- When we have categorical data and our model requires numerical data, we need to transform text to numbers to feed into the model.
- A one hot encoding operation transform categorical data into columns of 0 and 1.

Sample	Color
1	Red
2	Green
3	Blue
...	...



Sample	Is_Red	Is_Green	Is_Blue
1	1	0	0
2	0	1	0
3	0	0	1
...	...	...	...

FEATURE SCALING

- Imagine if x1 and x2 are not similar in scale
- For example
  - x1 = number of bedrooms (0-20)
  - x2 = area of the house (24-3000 sqm)
- This means the scale of your theta will also be different.

WHY SCALING

- Some algorithms like regression models and neural networks are very sensitive to scales.
- For example, if you want to compare the impact of number of bathrooms (scale 1-4) to the impact of areas (scale 24-250 sqm) on house prices, how do you make sure you are comparing apples or oranges? You need to rescale!

FEATURE SCALING

normalization with max, min, mean

$$x_1 := \frac{x_1}{max(x_1)}$$

$$x_1 := \frac{x_1 - mean(x_1)}{max(x_1)}$$

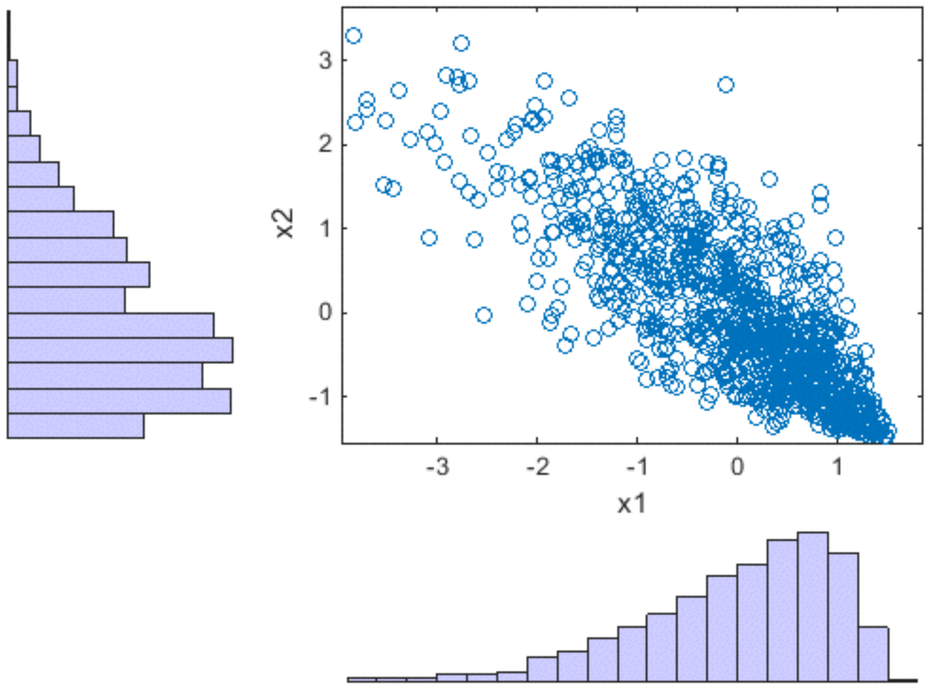
$$x_1 := \frac{x_1 - mean(x_1)}{max(x_1) - min(x_1)}$$

standardization a.k.a. z-score

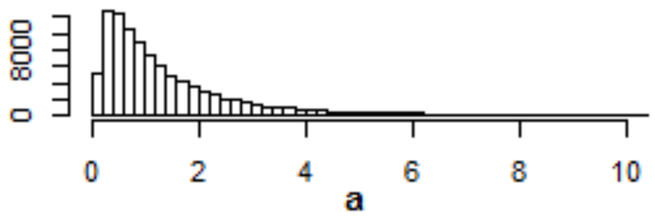
$$x_1 := \frac{x_1 - mean(x_1)}{std(x_1)}$$

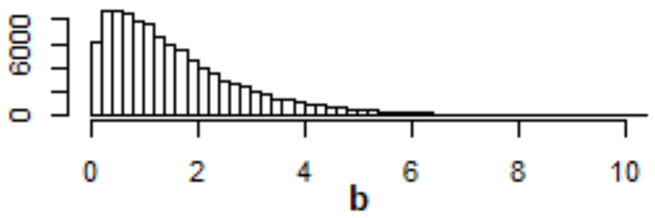
What's the range of these rescaled variables?

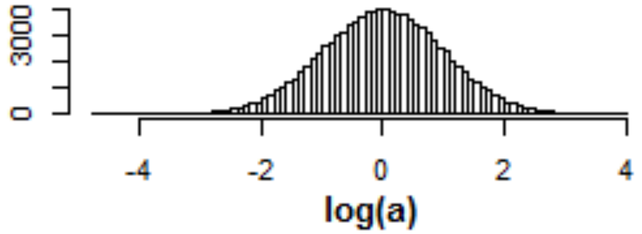
SKEWED DATA

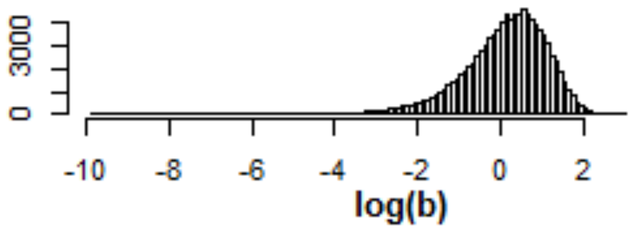


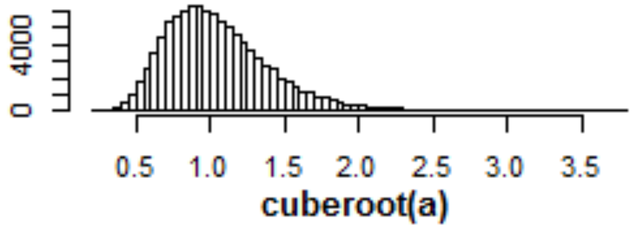
CORRECTING SKEWED DATA

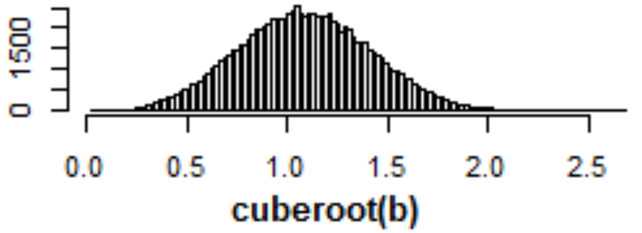












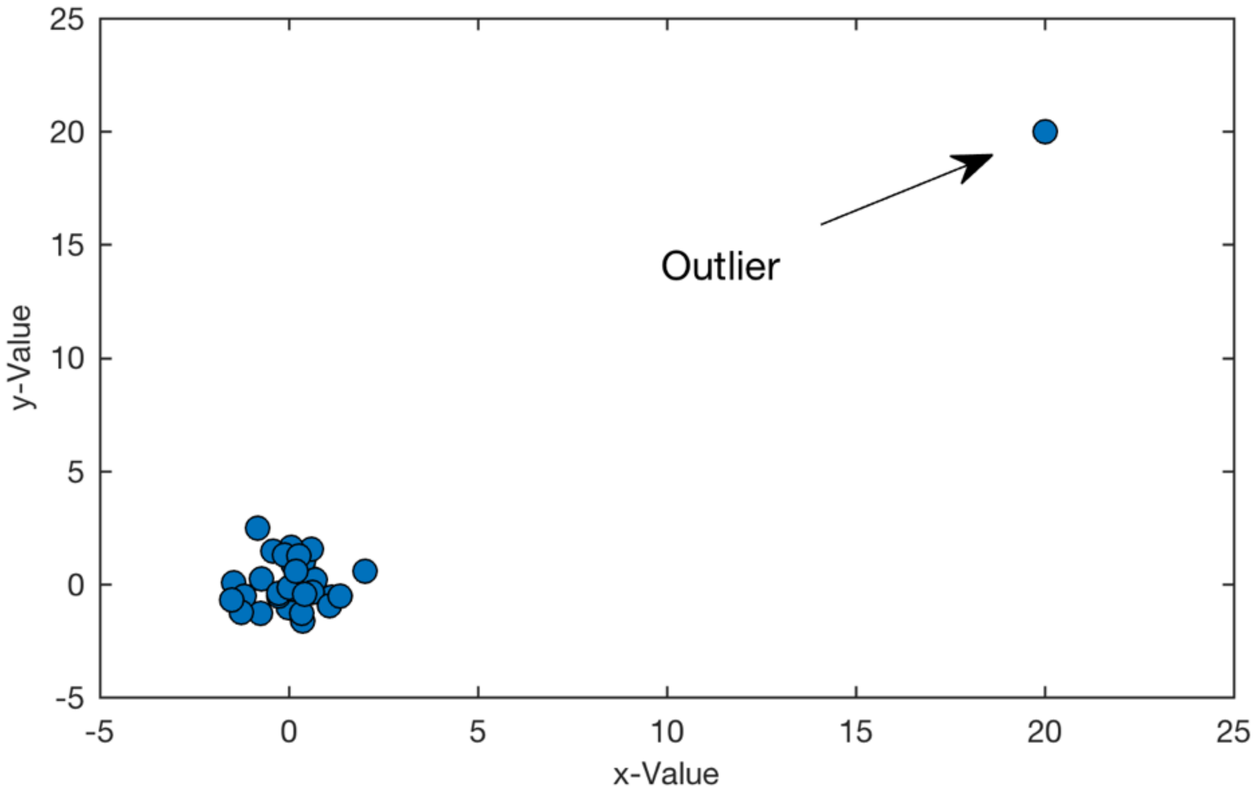
MISSING DATA

- If data are assumed to be missing at random, we may simply ignore the data
  - You went to all houses and randomly for some houses, you took time to do detailed house measurement
  - In this case, you simply remove the missing data from the analysis (cut the whole row or column)
  - Be aware that missing data might not be as random as you think

MISSING DATA

- Mean or mode substitution
  - Missing income? Fill it with average income?
  - Don’t know if the patient is left-handed or right-handed, assume right-handed, because it’s more common.

OUTLIERS



- Outliers are those data points that are way different from the rest.
- Outliers can be misleading.
- For example, in this picture, without outliers you see no correlations in the data, but with outliers, correlations start to emerge, but those correlations are unreal!

WHAT CAUSE OUTLIERS?

- Measurement errors (mistyped data)
- Sensor errors
- Freaky circumstances (events that are unlikely, but happen, for example, a gigantic product sales due to liquidation).

REMOVING OUTLIERS

- Make a scatterplot or histogram plot of your data and look for extreme values.
- Assume that data has a Gaussian distribution and look for values more than 2-3 standard deviations from the mean.
- Filter out outliers candidate from training dataset and see if model's performance improve.

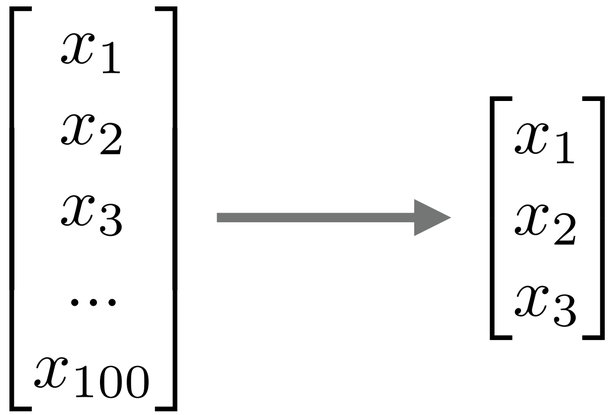
OUTLIER REMOVAL ALGORITHM

- Train model with all the data
- Look at data points with largest residual errors (say 10% of the data)
- Remove those data points from the dataset
- Retrain the model, do it over and over.

DIMENSIONALITY REDUCTION

Dimensionality Reduction: take very high-dimensional features and transform them into lower-dimension features without losing the quality of the data.

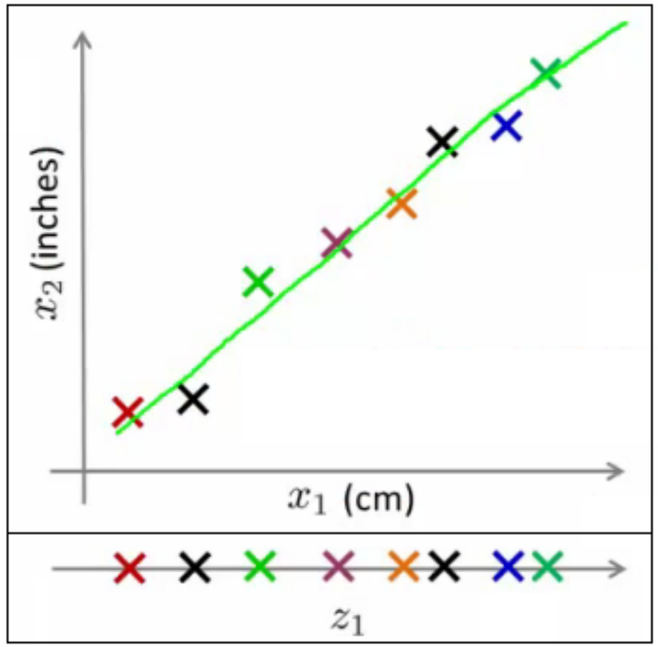
This is usually done by optimizing a cost function that quantifies goodness of components.



Methods:

Principle component analysis, singular value decomposition, independent component analysis

DIMENSIONALITY REDUCTION



PCA

- Starting with high-dimensional data
- Find new axes where data points can be projected on
- Get data on a new lower dimension form

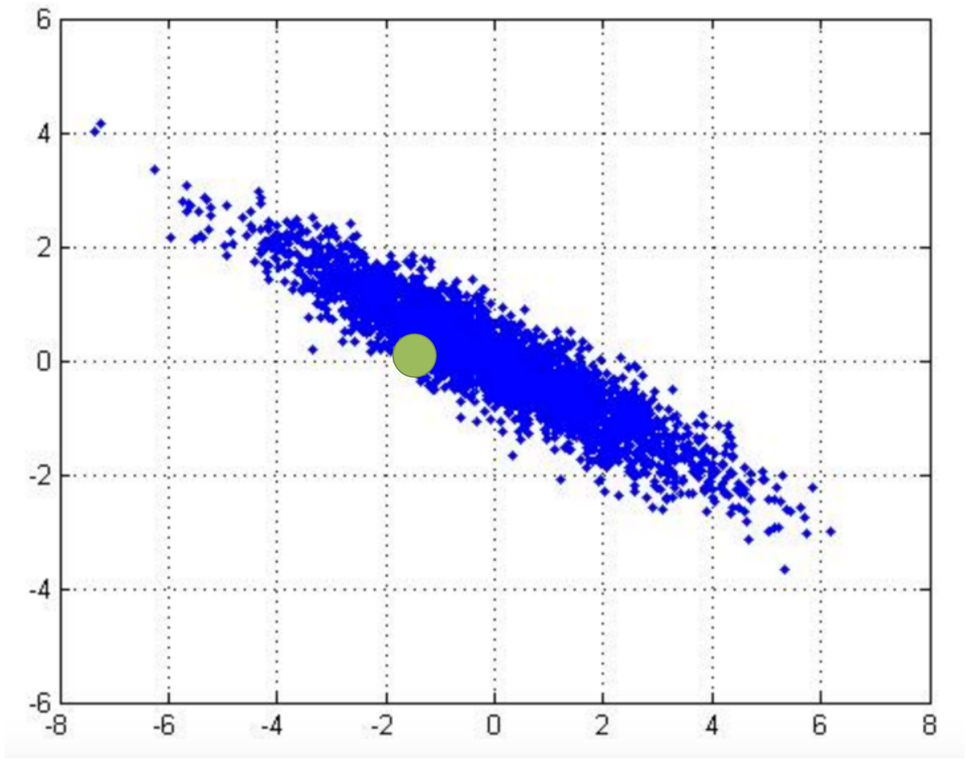
More Realistic Example

We collected restaurant ratings from 1M people rating 1000 restaurants (this dataset has 1000x1000 dimensions). We don't want to build model in 1000 dimensions (too computationally expensive). We will need PCA for this one.

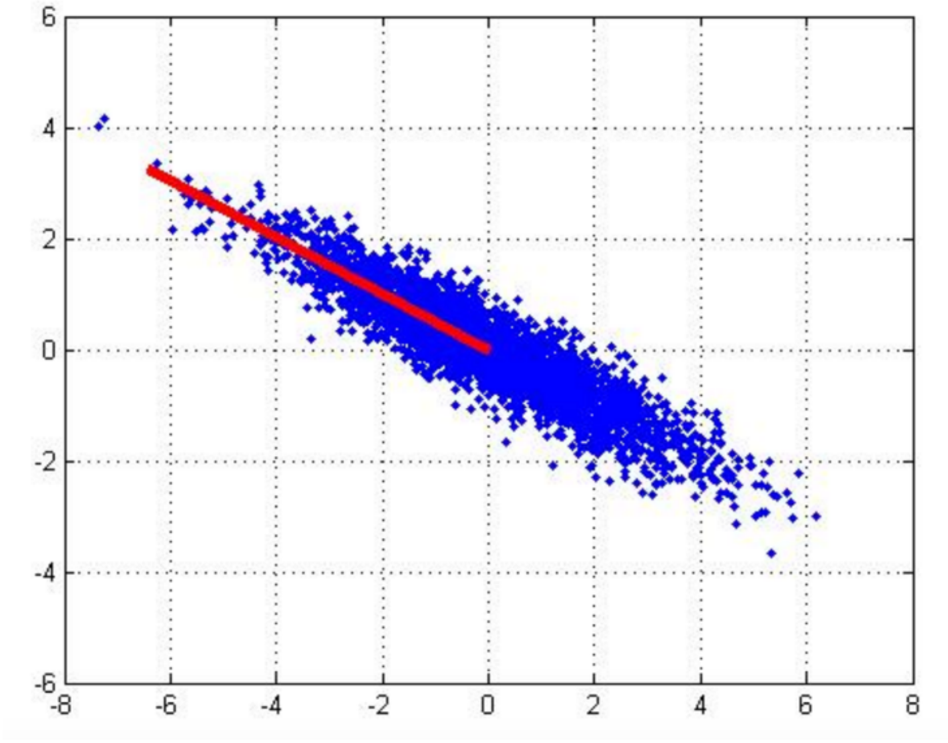
WHY PRINCIPAL COMPONENT ANALYSIS

- Visualization
- More efficient use of resources
- Noise and outlier removal
- Faster processing by machine learning algorithm

PRINCIPLE COMPONENT ANALYSIS ALGORITHM



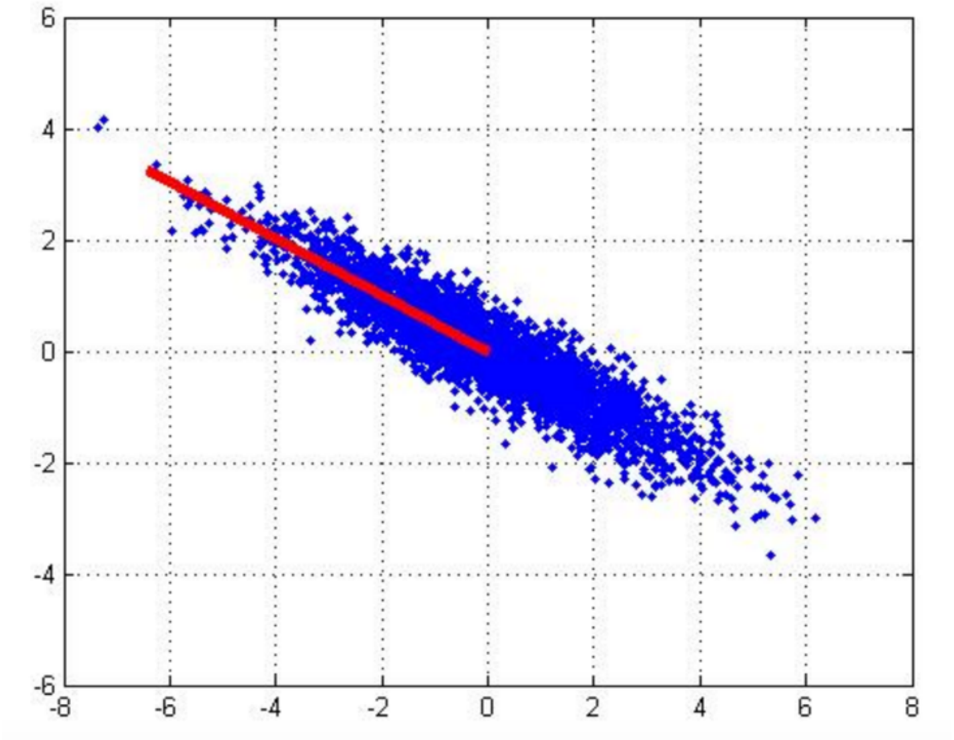
Start at the center of the data



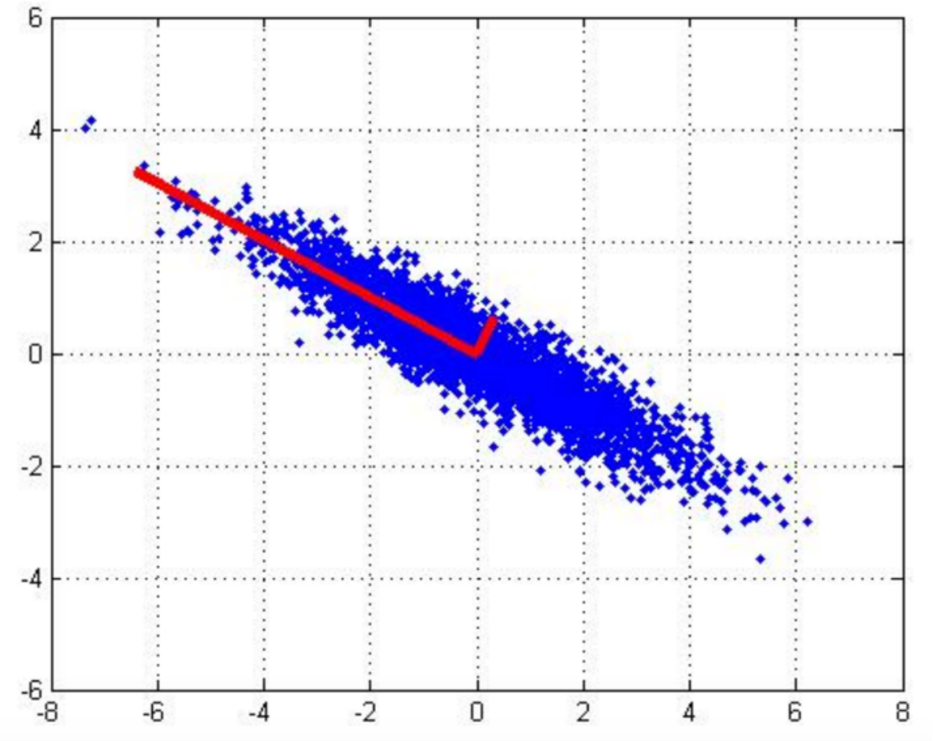
Find the dimension which maximizes variance.
That's the first principal component!

Slide of Matt Gormley

PRINCIPLE COMPONENT ANALYSIS ALGORITHM



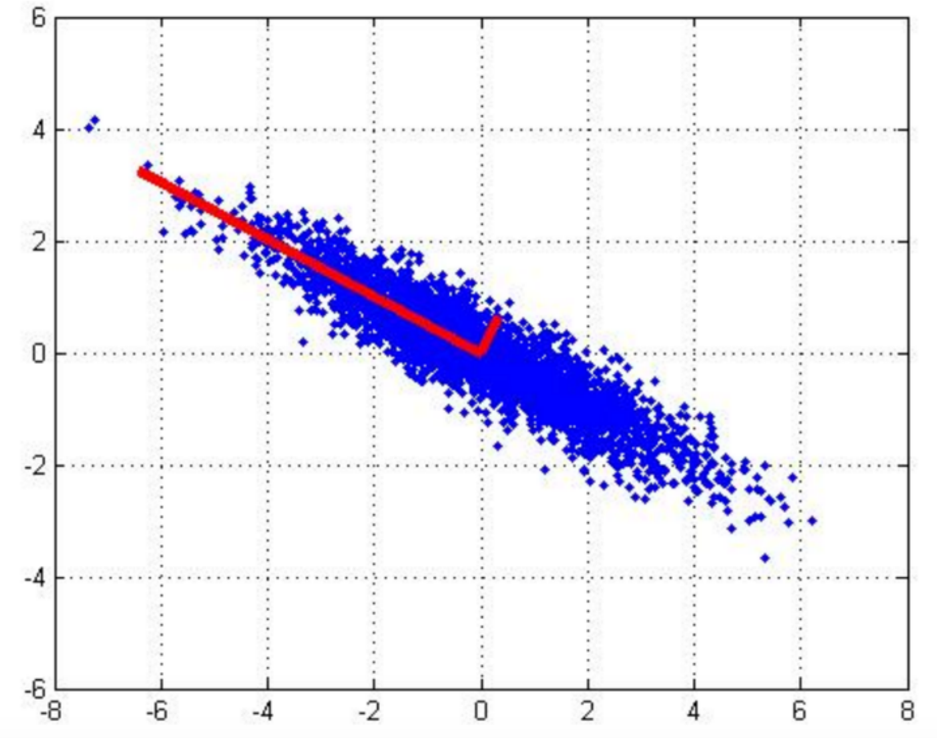
Find the dimension which maximizes variance. That's the first principal component!

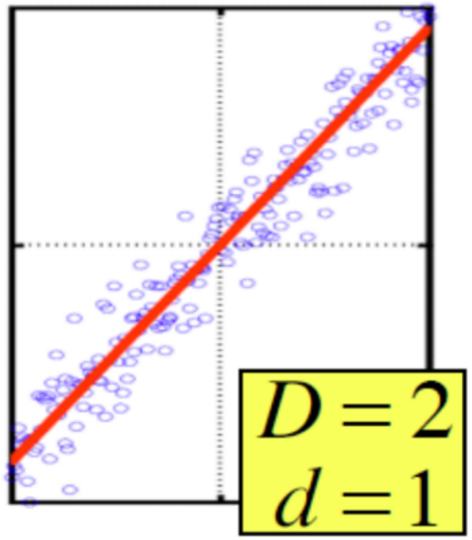


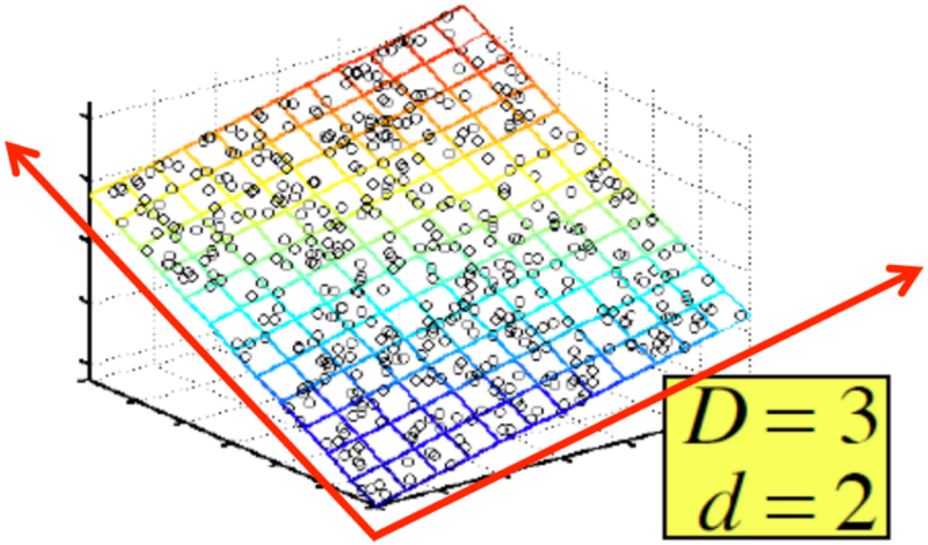
Find another dimension that is perpendicular to the first dimension, and maximize variance. That's the second principal component!

Slide of Matt Gormley

PRINCIPLE COMPONENT ANALYSIS ALGORITHM







Find another dimension that is perpendicular to the first dimension, and maximize variance. That's the second principal component!

For N dimensions, continue until you reach the N component.

Slide of Matt Gormley

MORE ADVANCED TECHNIQUES

- How to preprocess data with R:

  https://machinelearningmastery.com/pre-process-your-dataset-in-r/

- Missing value treatment:

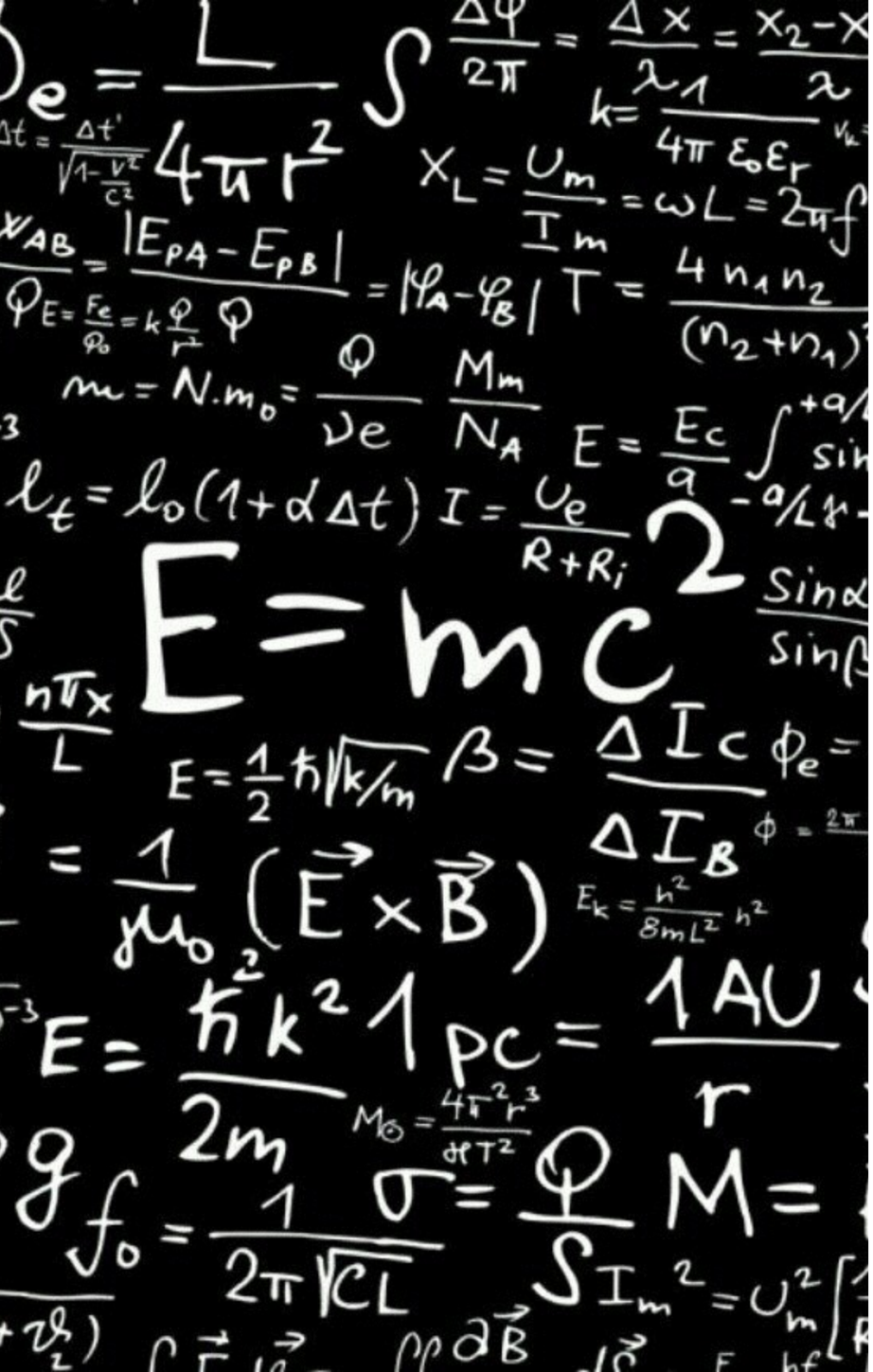
  https://www.r-bloggers.com/missing-value-treatment/

- How to handle missing data with Python:

  https://machinelearningmastery.com/handle-missing-data-python/

- How to identify outliers in your data:

  https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/



DATA PREPROCESSING LAB