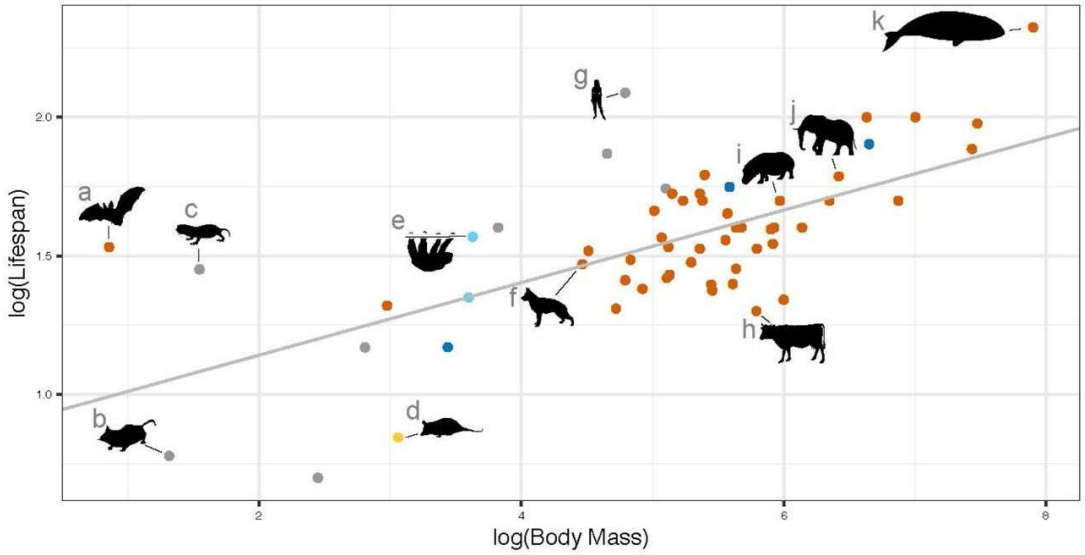
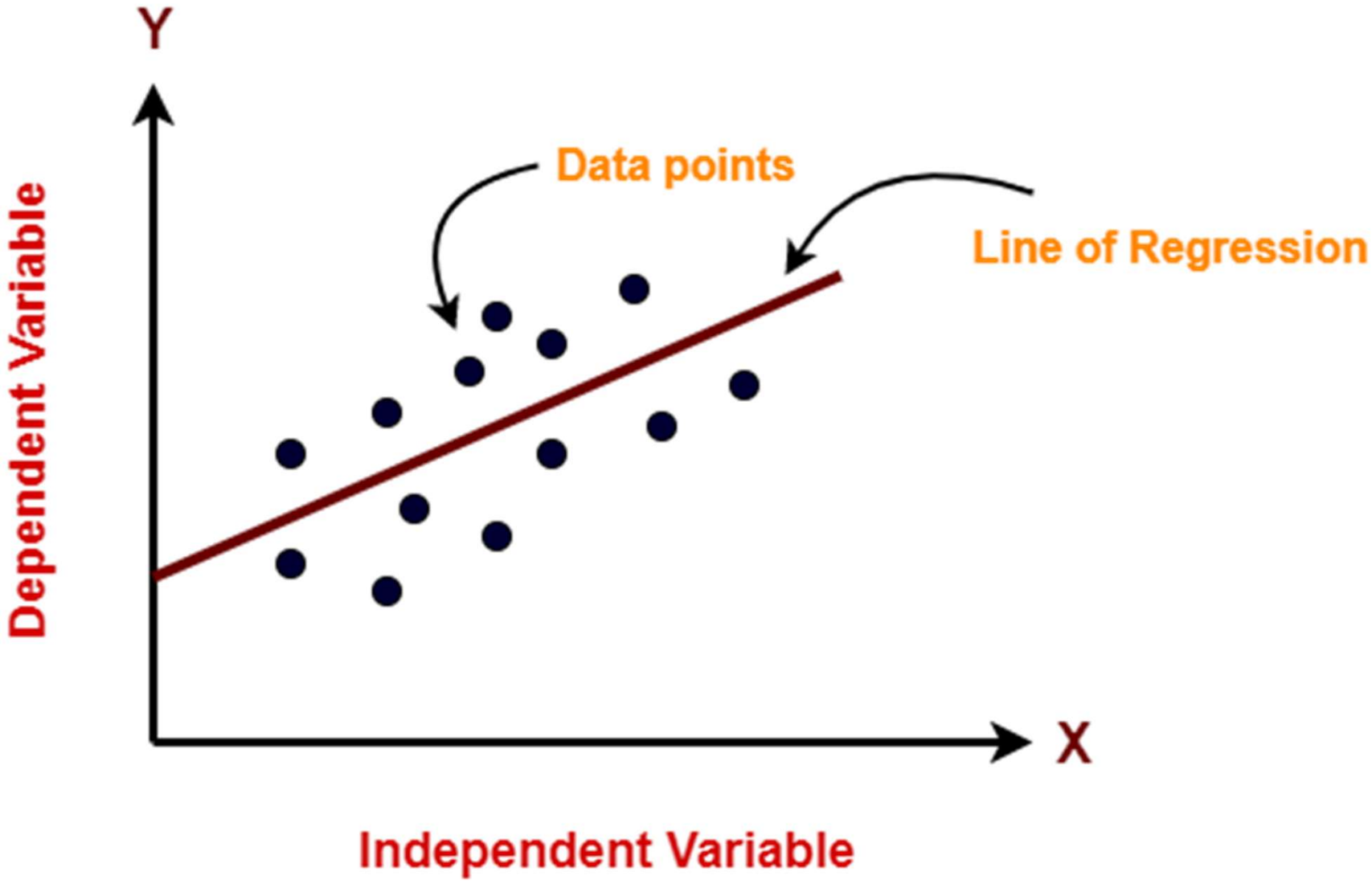
Linear Regression

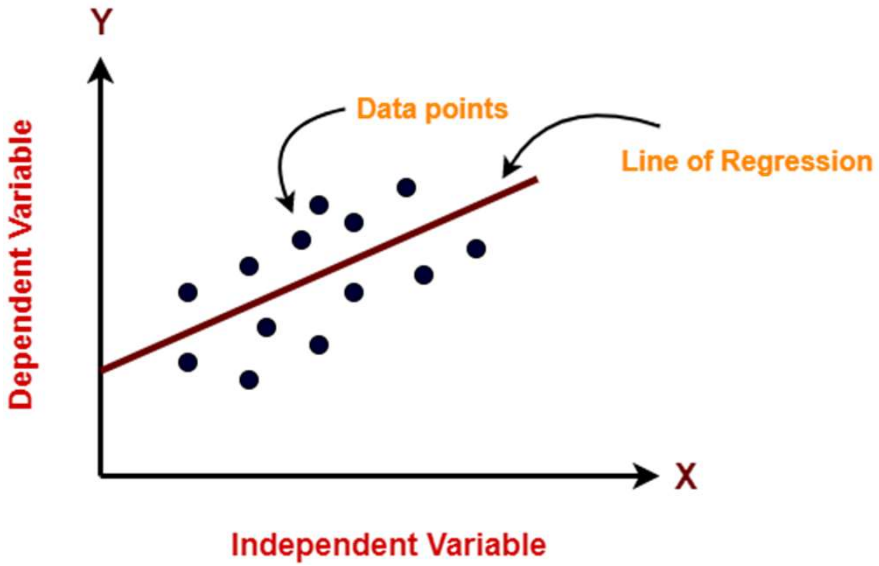




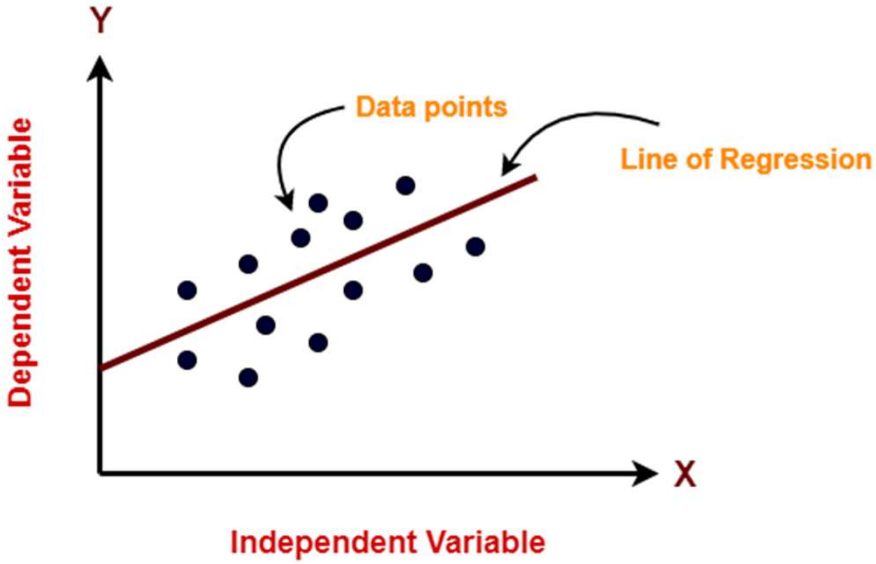
https://www.gatevidyalay.com/linear-regression-machine-learning-examples/

Linear regression in machine learning

- supervised learning
- simple to use and understand
- well-known pros and cons
- model finds the best fit linear line between the independent and dependent variable



Simple regression



slope *intercept*

$$y = m \cdot x + b$$

Occam’s razor

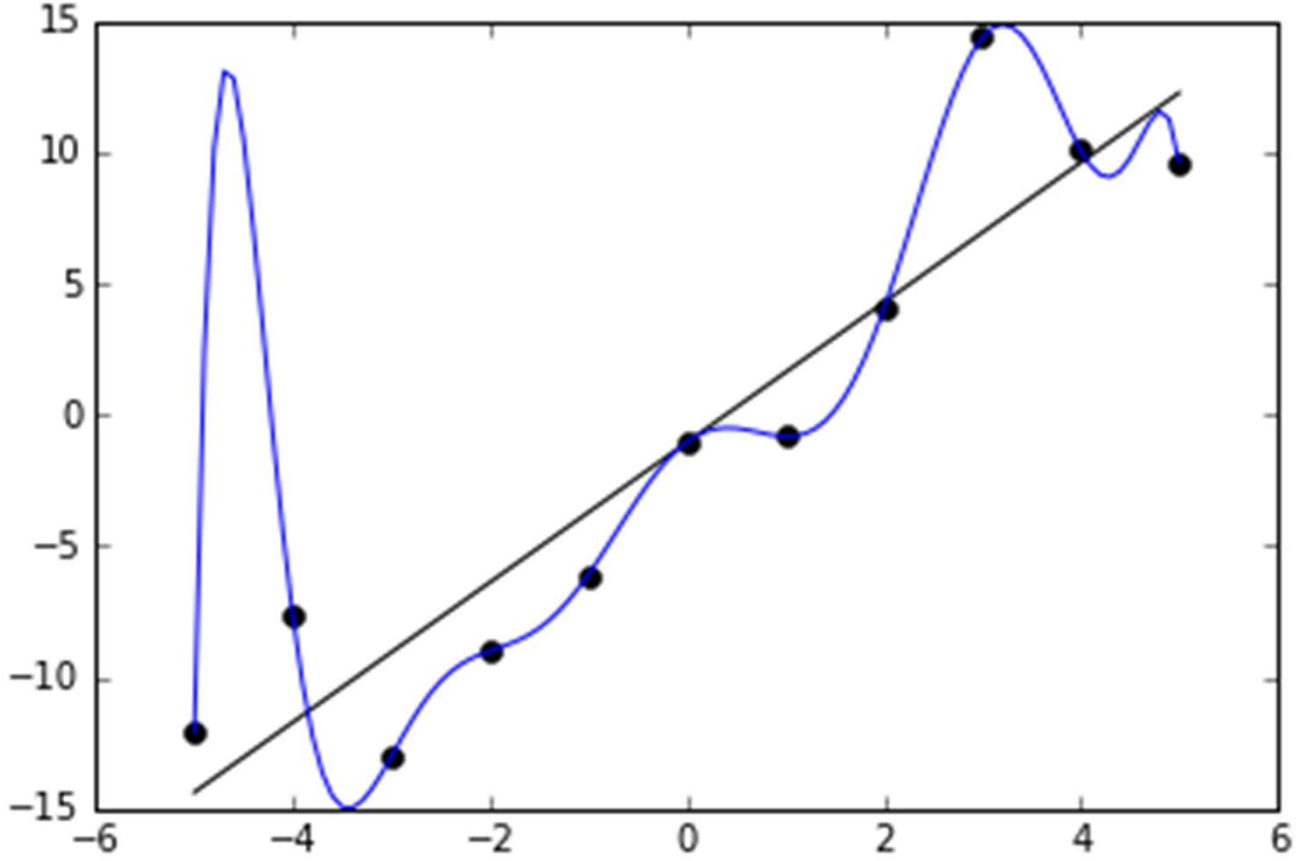
Occam’s razor

“Entia non sunt multiplicanda praeter necessitatem”

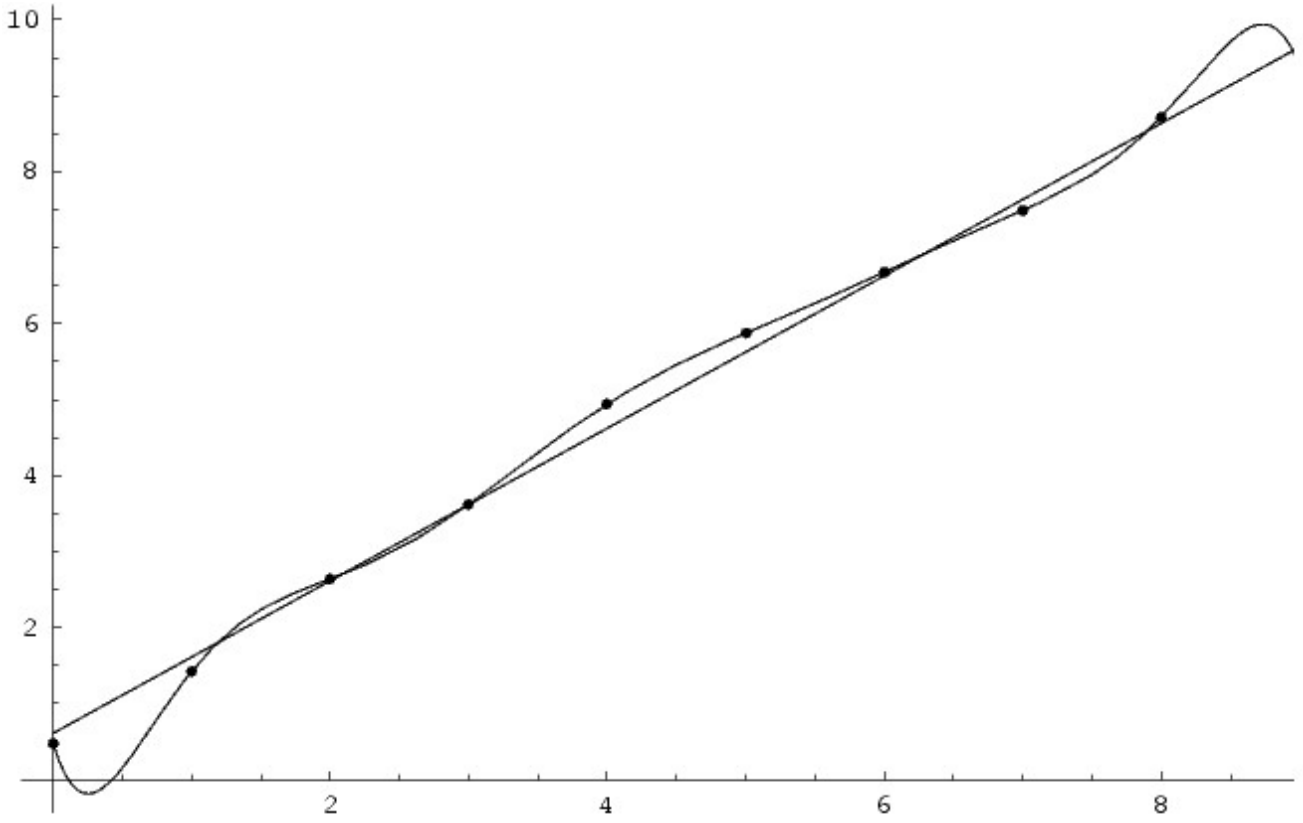
Occam’s razor

„Entities must not be multiplied beyond necessity”

Overfitting



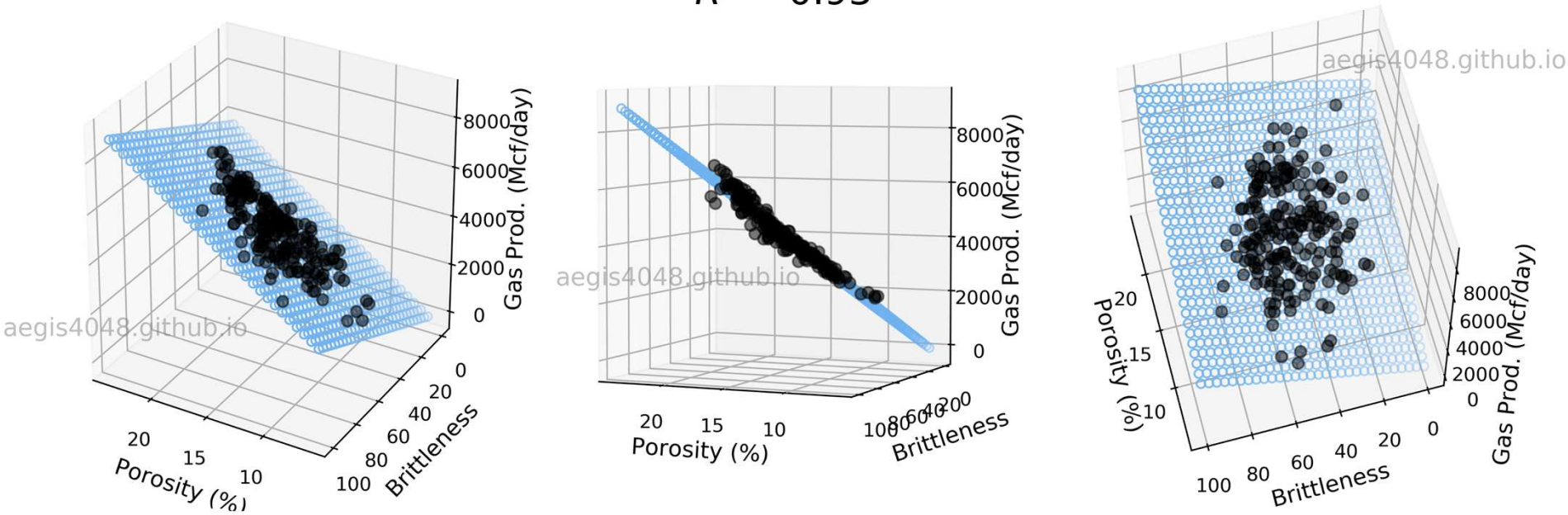
Overfitting



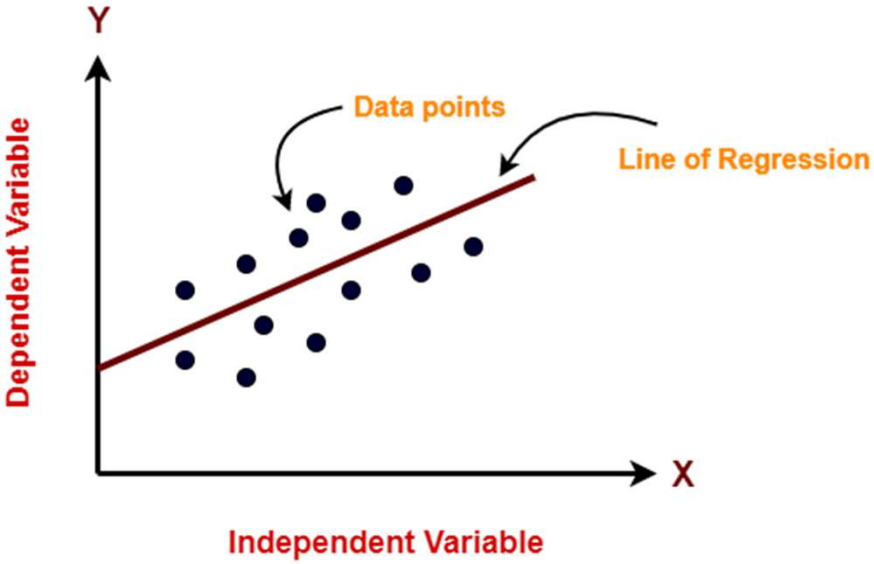
Multilinear regression

$$y = m_1 \cdot x_1 + m_2 \cdot x_2 + \ \ldots + m_n \cdot x_n + b$$

$$R^2 = 0.93$$



Simple regression



slope *intercept*

$$y = m \cdot x + b$$

sometimes it can be assumed as 0

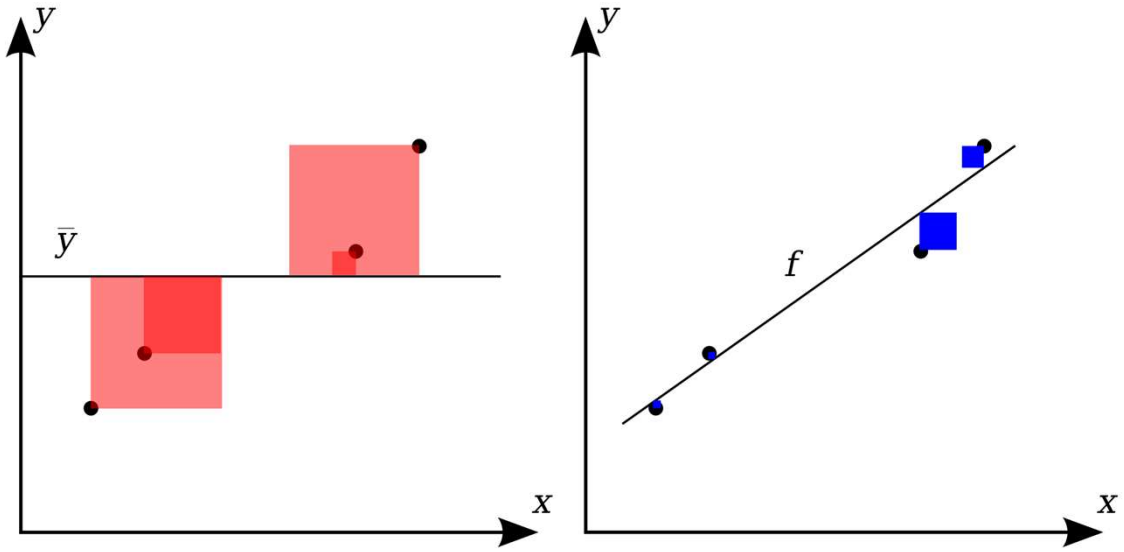
Coefficient of determination - R^2

$$SS_{res} = \sum_i (y_i - y_{pr\ \ i})^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

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

$$R^2 = 1 - FVU$$



unexplained variation

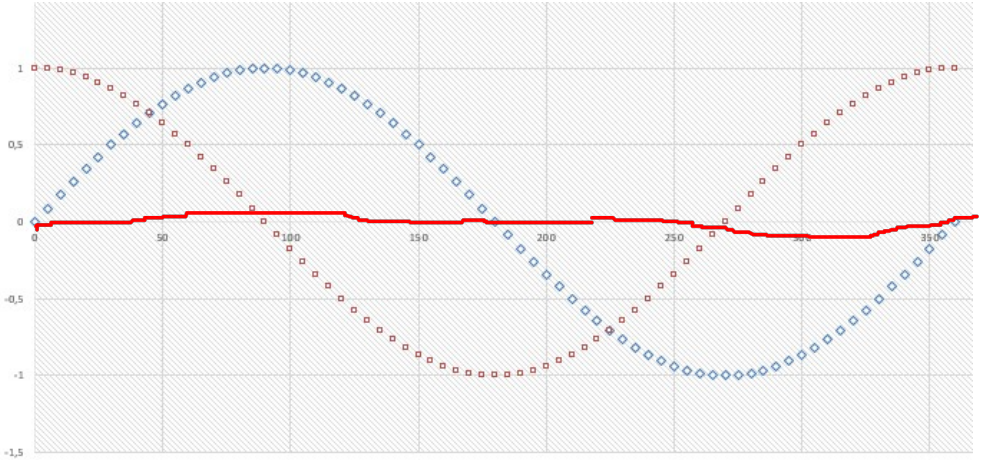
Coefficient of determination - R^2

- $unitless$
- $R^2 = 0 \rightarrow random\ prediction$
- $R^2 = 1 \rightarrow perfect\ predictive\ model$

https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/linear-regression/the-statistical-sommelier-an-introduction-to-linear-regression/quick-question-54/

Conditions for linear regression

- Relationship between indenpendent and dependent variable should be linear
- Data is not noisy
- Gaussian distribution of input and output data is better
- Rescaling input (standarization or normalization)



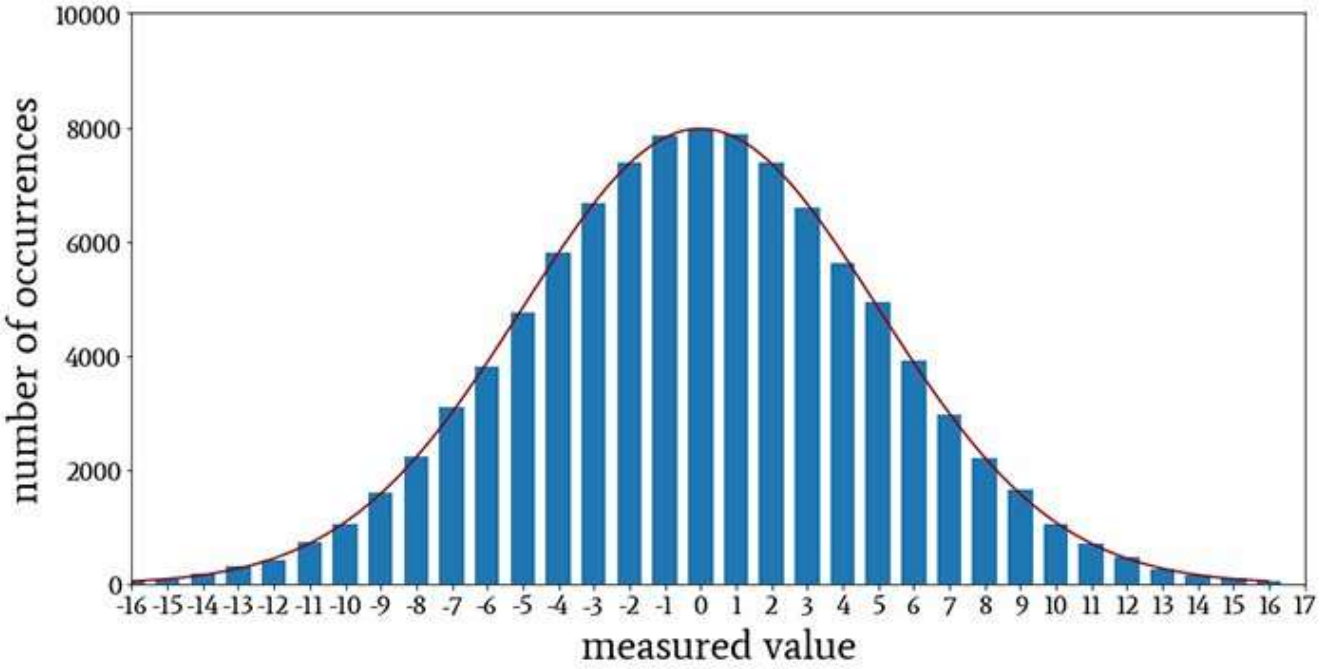
normalization vs standarization

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

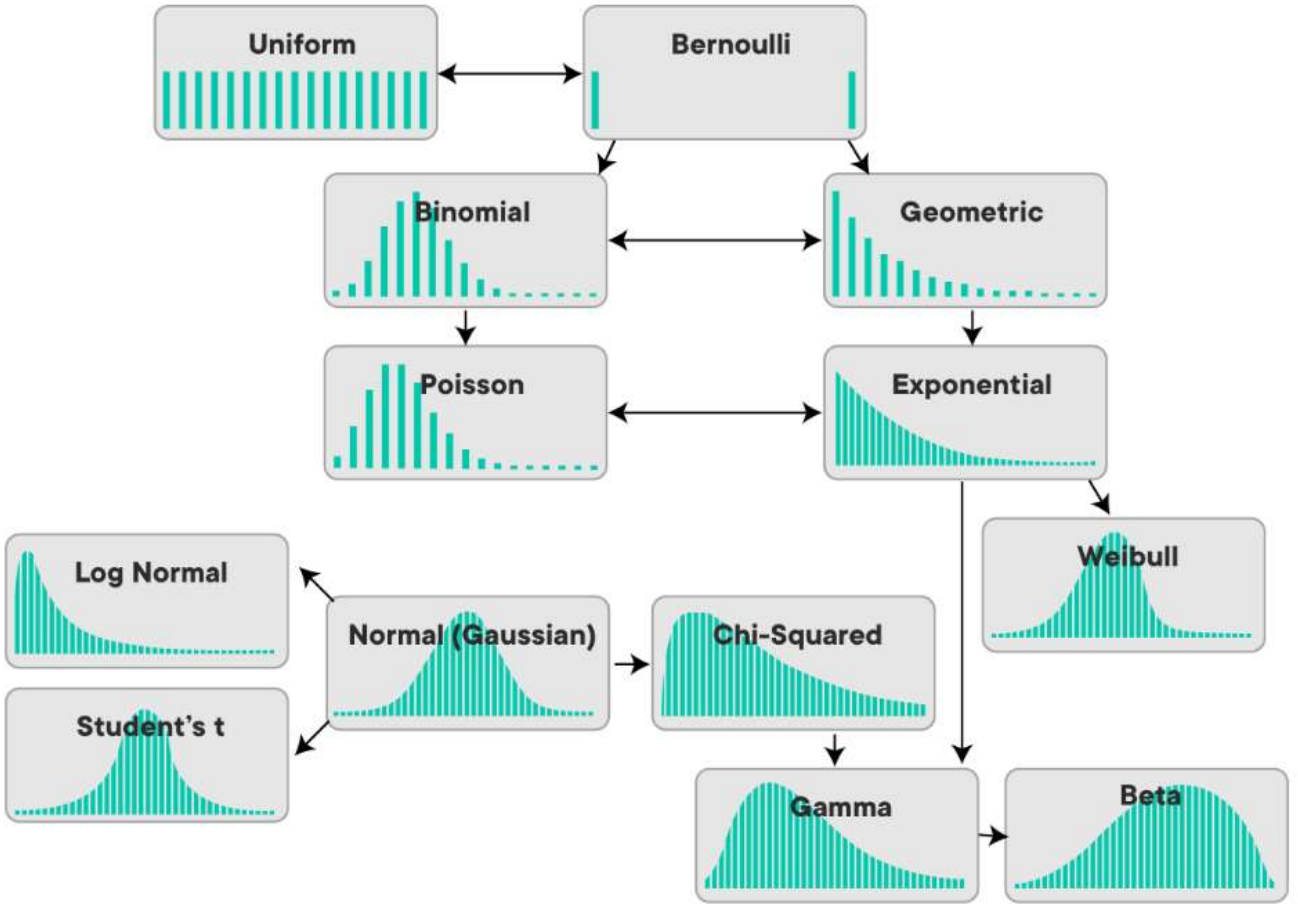
$$X_{new} = \frac{(X - \bar{X})}{SD(X)}$$

Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

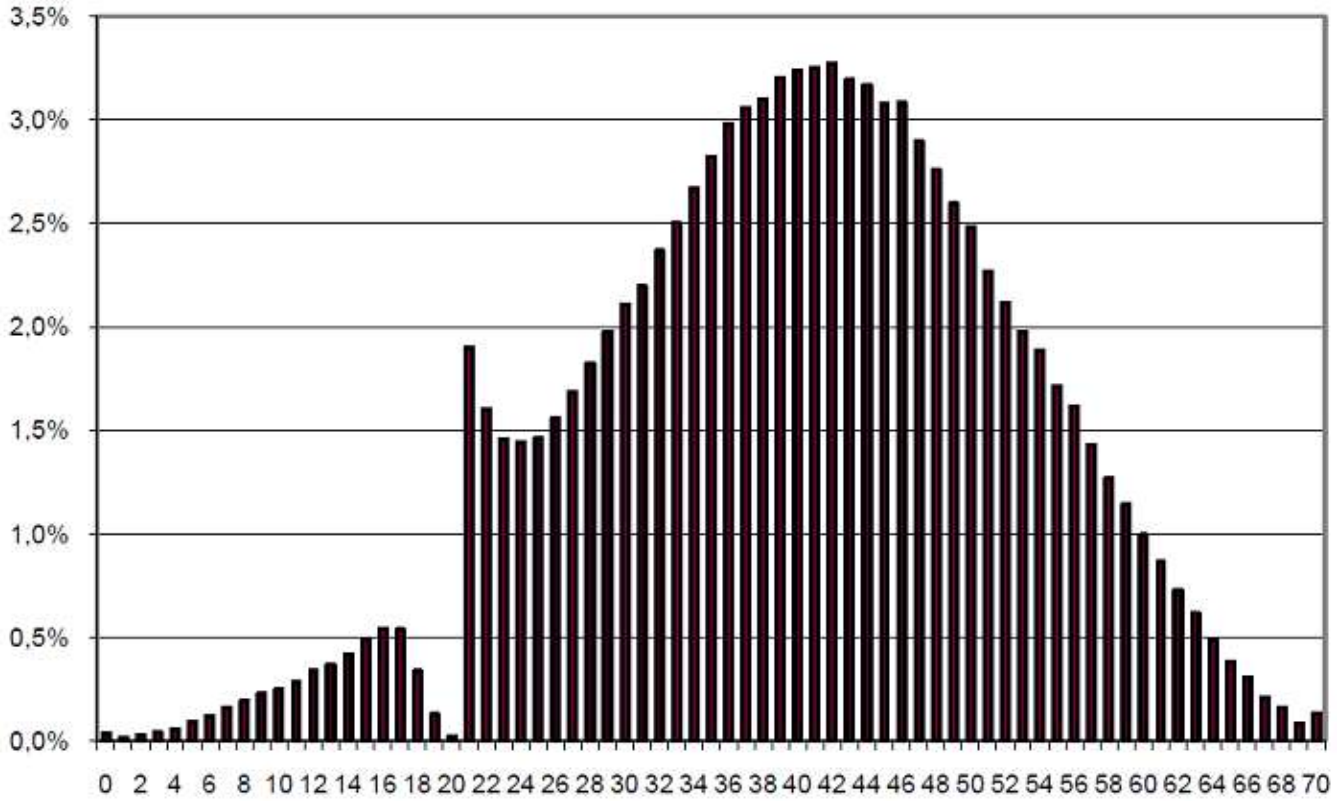
Gaussian distribution



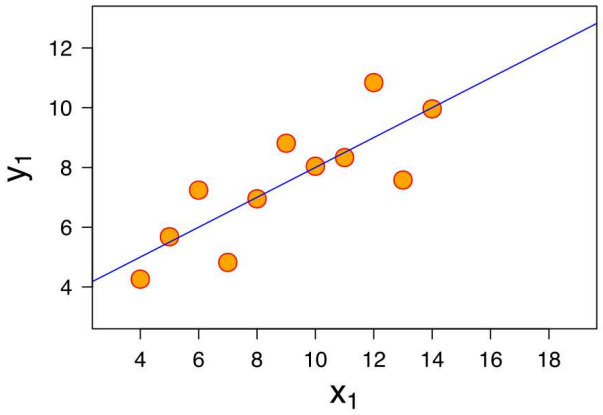
Kinds of distribution

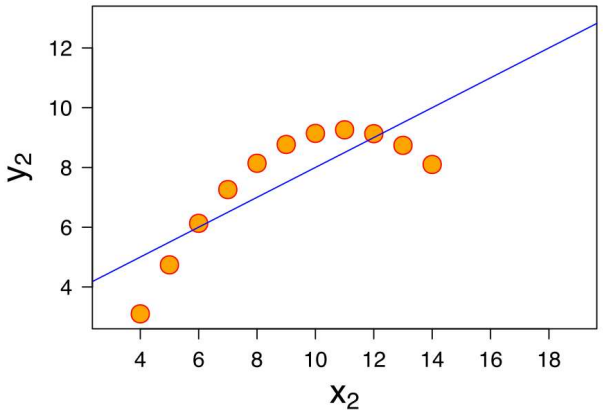


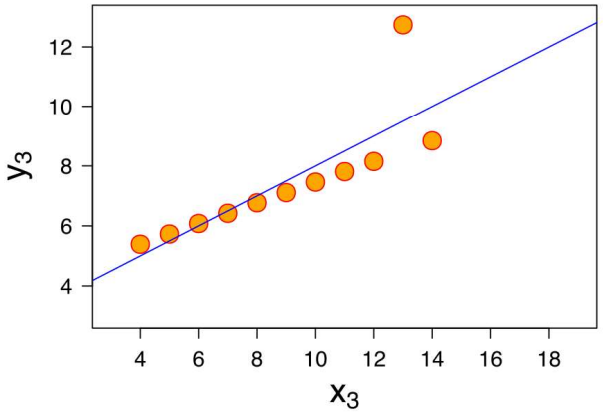
Distribution ?

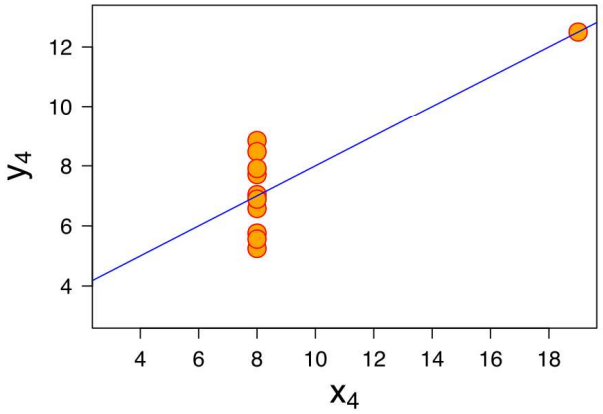


Traps of linear regression



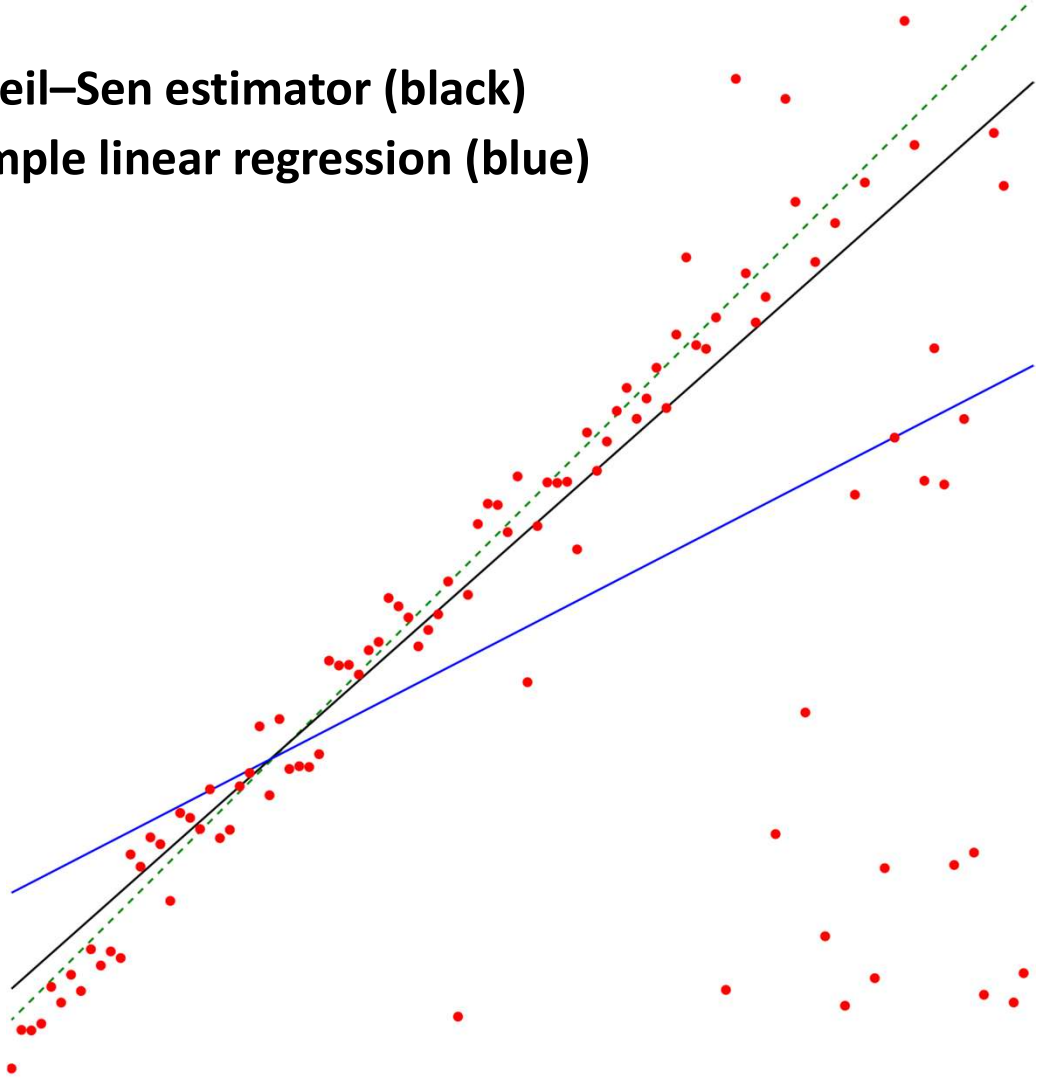






Other prediction methods

- **Theil–Sen estimator (black)**
- **simple linear regression (blue)**



Importing data example

```
import pandas as pd
file = 'file_name.csv'
df = pd.read_csv(file)
```

Importing data example

```
import pandas as pd
file = 'file_name.csv'
df = pd.read_csv(file)

print(df.head(10))
```

Linear regression – sklearn

- sklearn - - - - > linear_model
- your_regression = linear_model.LinearRegression()
- your_model = your_regression.fit(X, y)

dependent variable
independent variable } *training data set*

creating model of your data set

only if fit_intercept = true is set

- print(your_model.coef_)

Contains slope and intercept

Contains value of R^2

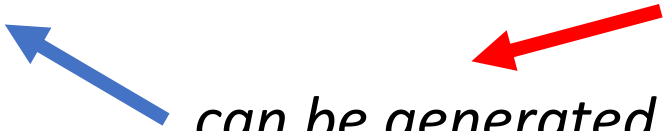
- print(your_model.score(X,y)

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

Linear regression – sklearn

- Predictions with ready models:
  - y_prediction = model.predict(x_pred)

reshaping might be needed



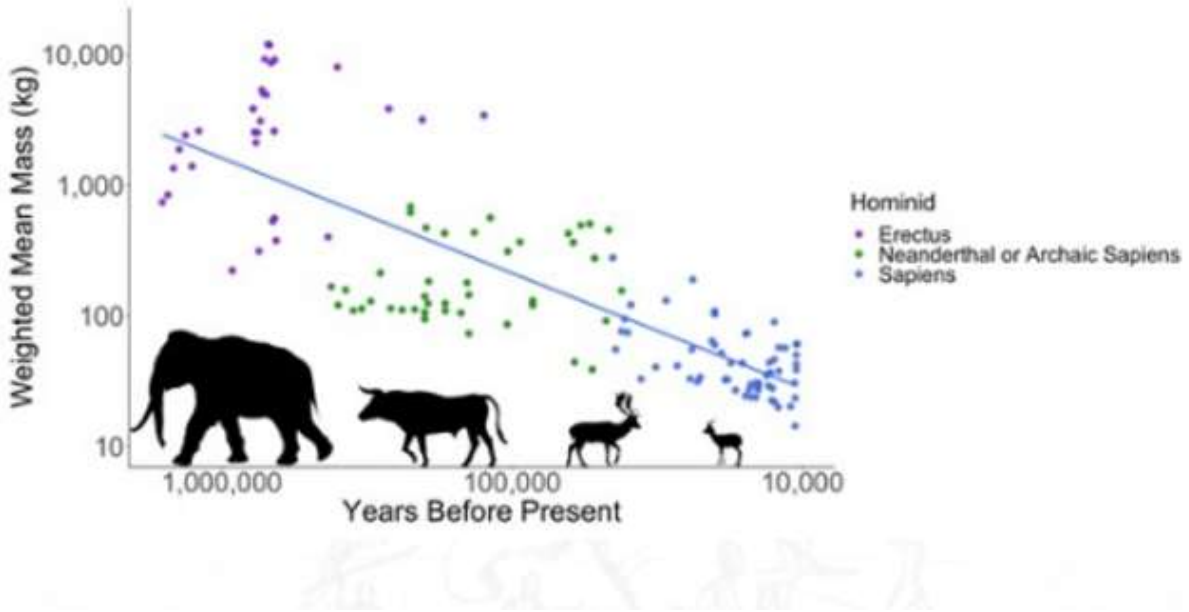
can be generated via np.linspace method

or

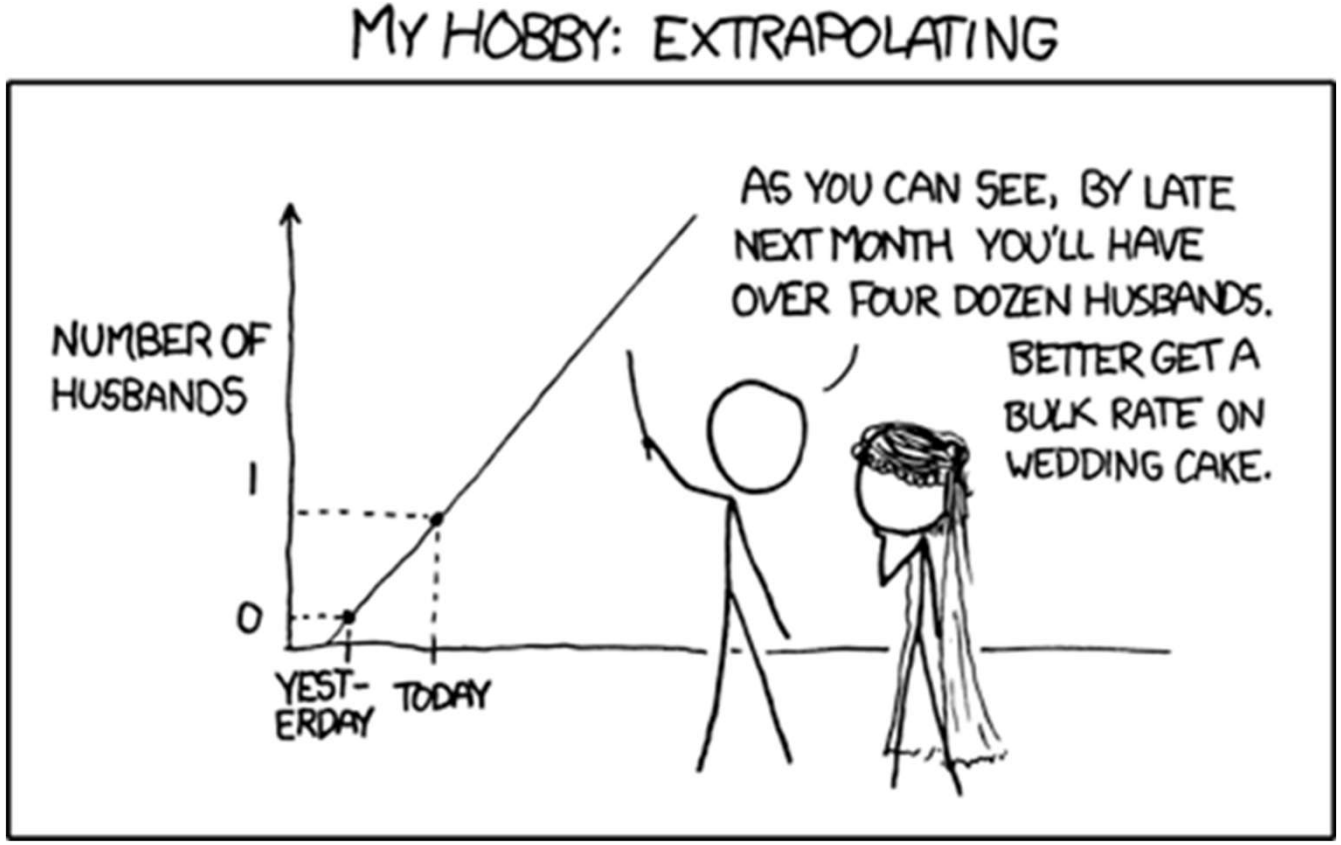
list comprehension

- Result can be shown using matplotlib

interpolation vs extrapolation



interpolation vs extrapolation



Zbiór danych

- https://www.kaggle.com/kumarajarshi/life-expectancy-who

- https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression