Linear Regression

Garanti BBVA Technology

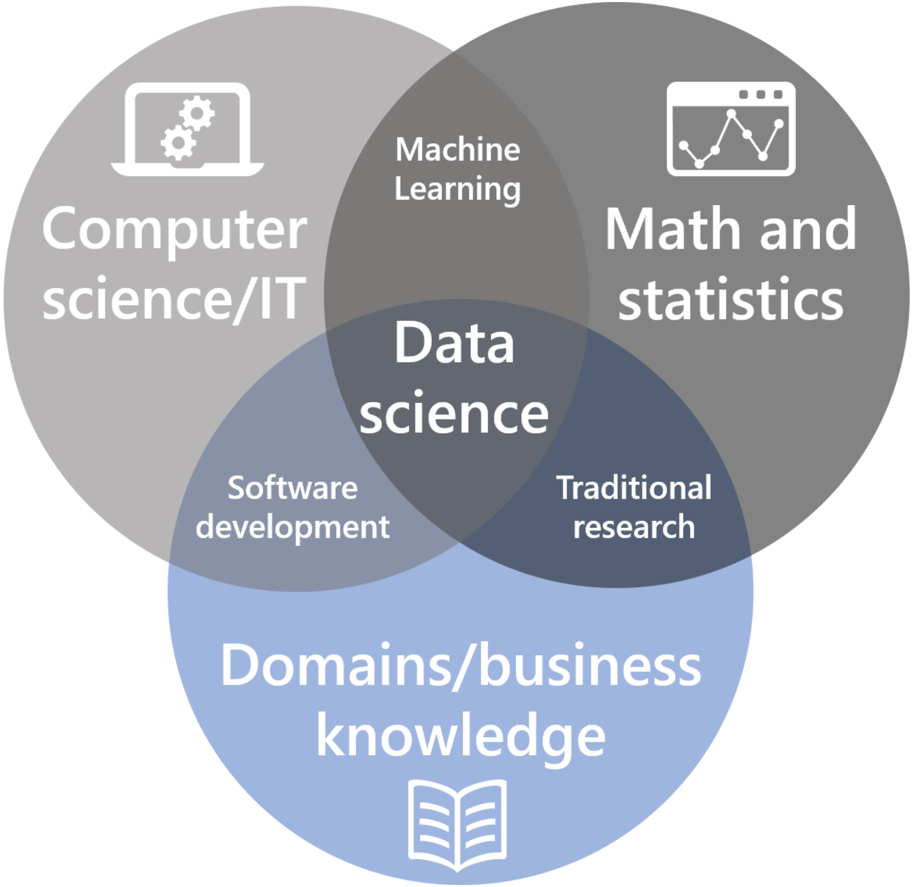
AGENDA

Garanti BBVA Technology

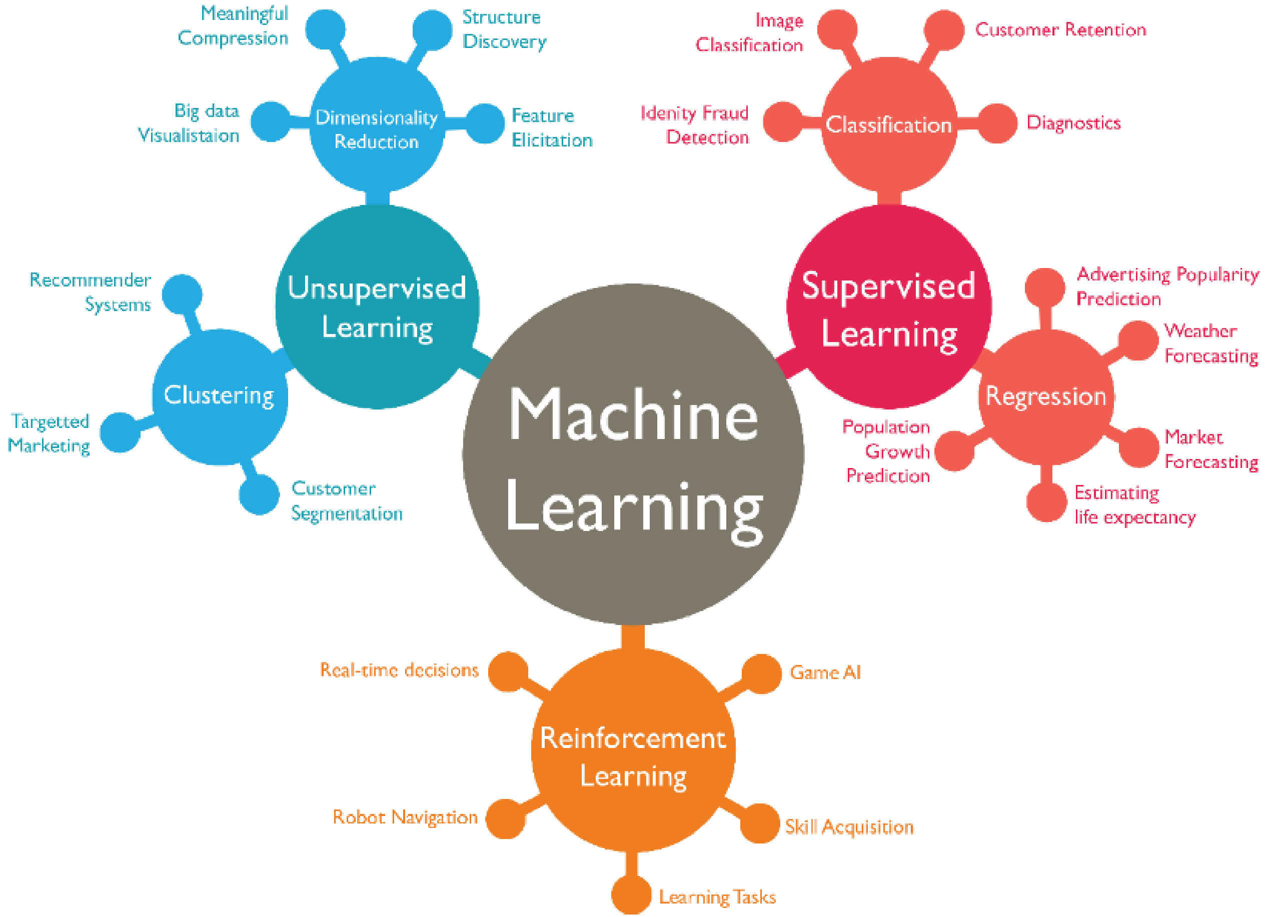
01

Linear Regression

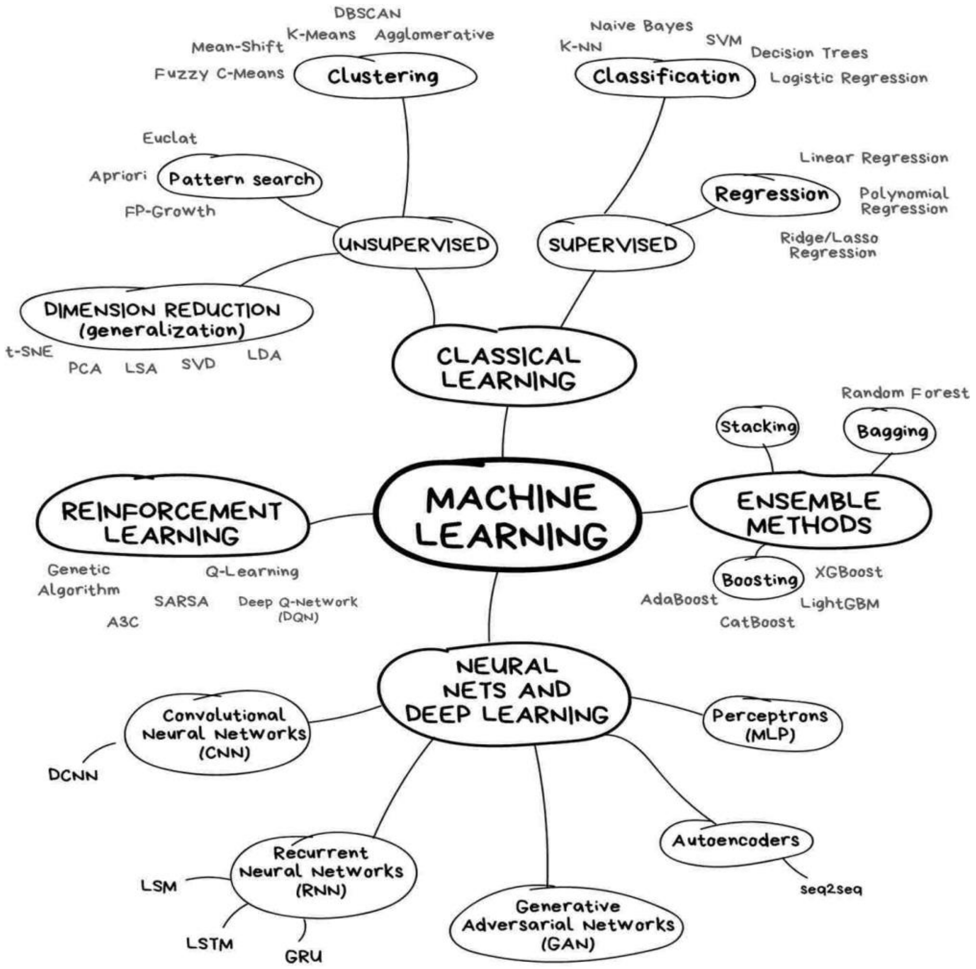
Hierarchy - Reminder



Machine Learning Branches - Reminder



Machine Learning Algorithms



Regression: for What?

Used when predicting a continuous dependent variable from number of independent variables.

Insights on consumer behavior

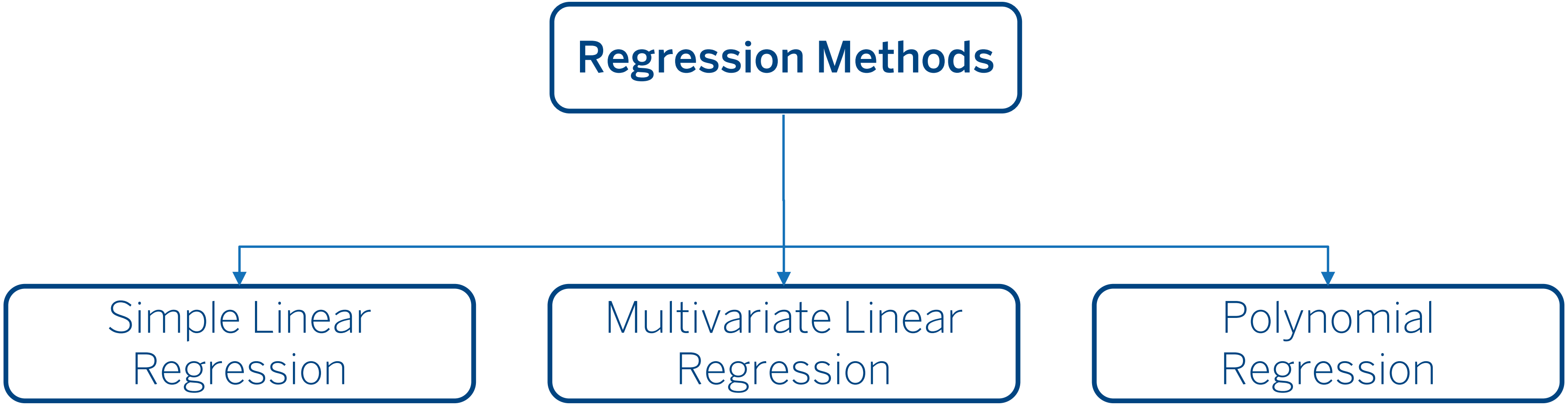
Understanding business

Evaluating market trends

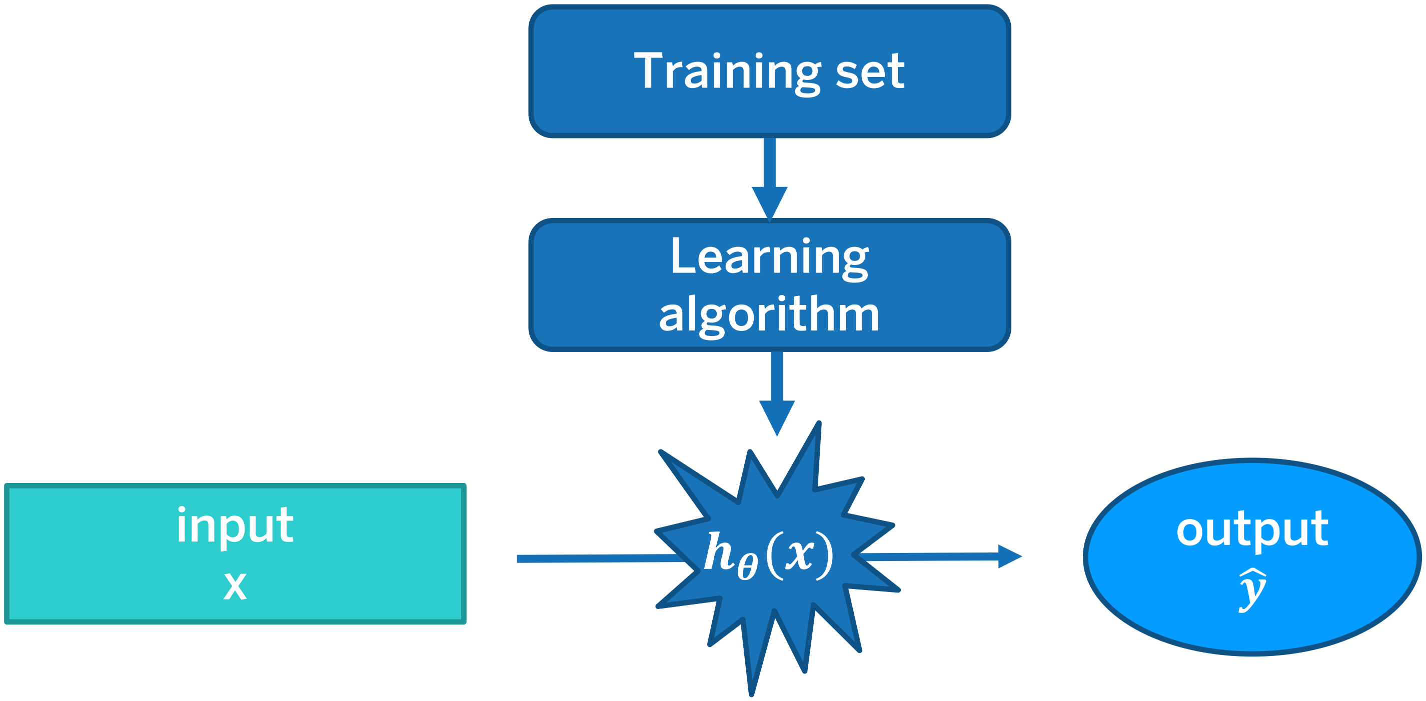
Examples from Finance Sector:

- Income Prediction
- Customized Interest Rate
- Customized Insurance Pricing

Regression Methods



Model Representation



$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$$

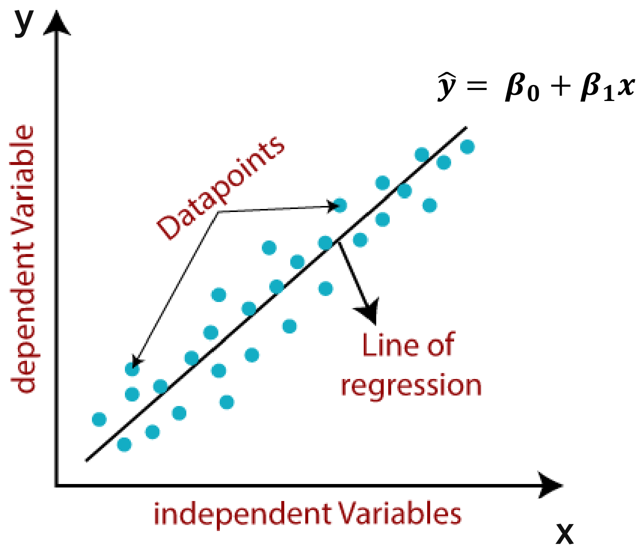
Univariate Linear Regression: Model Representation

Linear Regression

Linear Regression is a model that allows to estimate the value of a **quantitative** (numerical) variable as a linear function of the input variables or predictors.

$$\hat{y} = \beta_0 + \beta_1 x$$

where
$\hat{y}$: model estimate for the variable y
x : the input variable or predictor
β_0: the model estimate when x = 0
β_1: variable weight (slope)



Multivariate Linear Regression: Model Representation

Linear Regression

Linear Regression is a model that allows to estimate the value of a **quantitative** (numerical) variable as a linear function of the input variables or predictors.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ \ldots$$

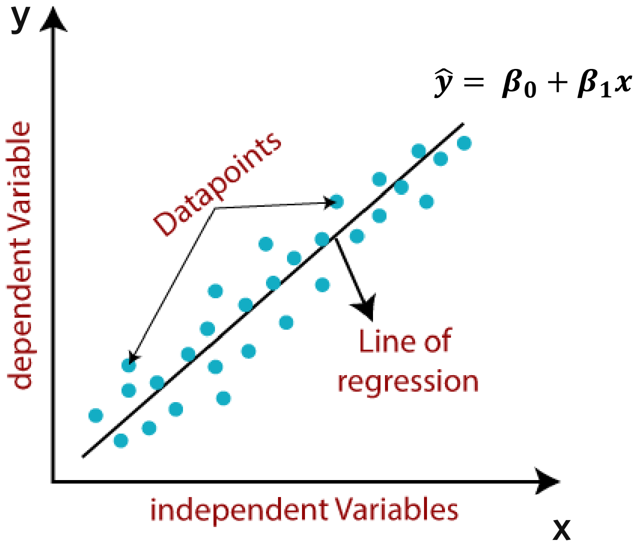
where

$\hat{y}$: model estimate for the variable y

x_i: the input variables or predictors

β_0: the model estimate when x = 0

β_j: variable weights



Objective of Model

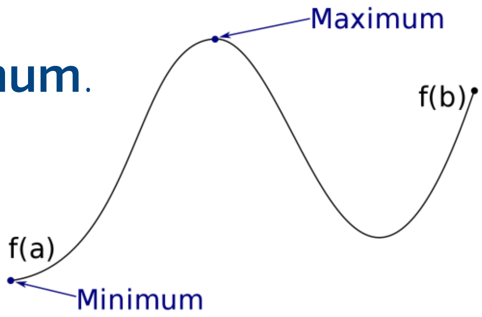
Objective Function;
Define anything that we are optimizing.

- Cost for a company
- Total profit

That function is optimal at a specific points X1, X2 etc.

What we do:

- Finding X1, X2 for which $\boldsymbol{h\theta}\ (\boldsymbol{x})$ is **minimum** or **maximum**.



Objective of Model

Linear Regression

- Intercept($\boldsymbol{\beta_0}$)
- Slope($\boldsymbol{\beta_x}$)

$$\hat{y} = \boxed{\boldsymbol{\beta_0}} + \boxed{\boldsymbol{\beta_1}}\boldsymbol{x_1} + \boxed{\boldsymbol{\beta_2}}\boldsymbol{x_2} + \boxed{\boldsymbol{\beta_3}}\boldsymbol{x_3} + \ \dots$$

Two main methods in finding regression parameters:

1) OLS(Ordinary Least Square)

- Non iterative
- Analytical solution (mathematical operations)

2) Gradient Descent

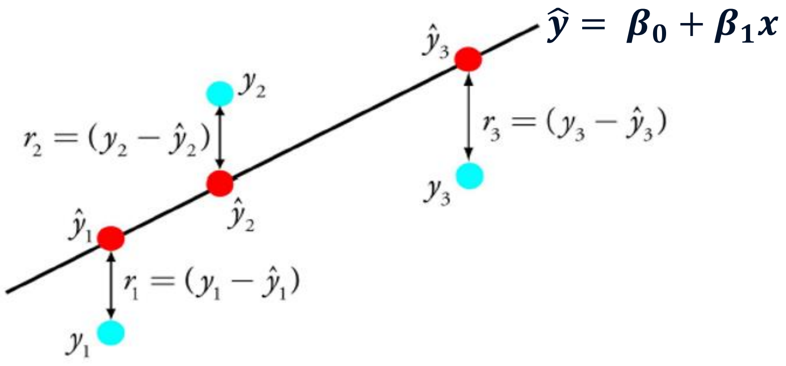
- Iterative
- Optimization method

Objective of Model

Linear Regression

Maximizing the similarity means **minimizing** difference.

Our goal is to develop the model that minimize the distance between actual and predicted output.

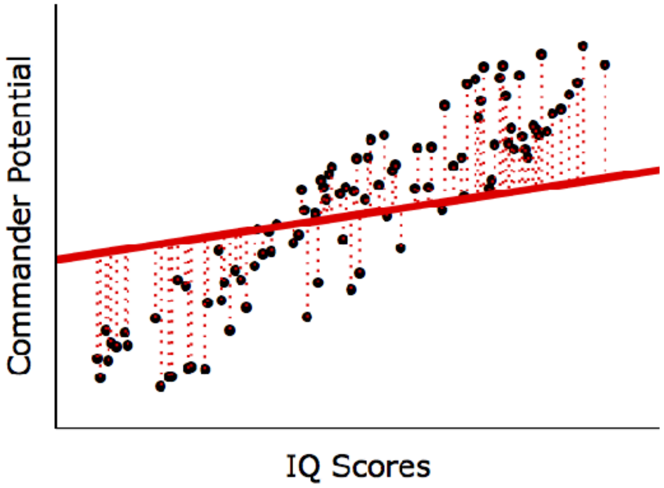


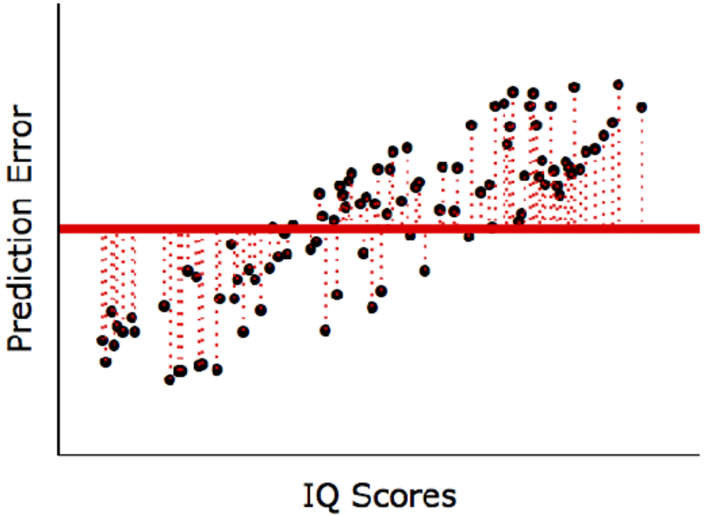
$$\mathrm{SS}_{\mathrm{residual}} : \sum_{i=1}^{n} \underbrace{(\hat{y}_i - y_i)^2}_{r_i^2}$$

r_i: residuals

Regression Model (Residuals & Error)

Linear Regression





Regression Model Optimization

Linear Regression(OLS)

The method of least squares chooses the values for $\boldsymbol{\beta_0}$, and $\boldsymbol{\beta_1}$ to minimize the sum of squared errors:

$$SS_{residual} = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{m} ((\beta_0 + \beta_1 x_i) - y_i)^2$$

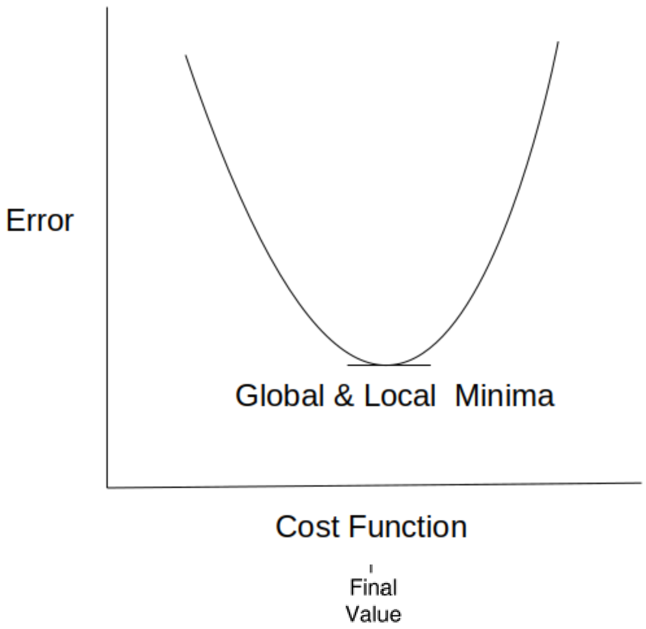
Using calculus, we obtain estimating formulas for $\boldsymbol{\beta_0}$, and $\boldsymbol{\beta_1}$:

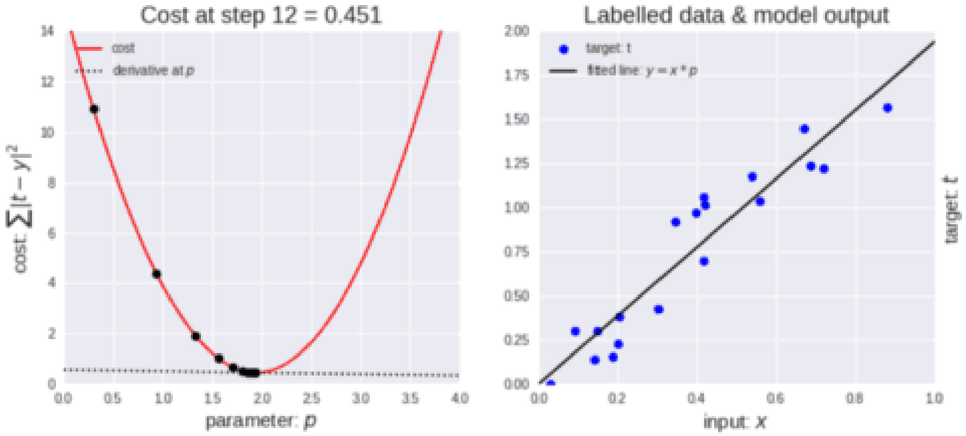
$$\beta_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x}_i)^2} = \frac{\text{Covariance}}{\text{Variance}}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Regression Model Optimization

Linear Regression - Whaf if we use *Gradient Descent* ?





Cost Function:

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}\underbrace{(\hat{y}_i - y_i)^2}_{r_i^2} = \frac{1}{2m}\sum_{i=1}^{m}(\beta_0 + \beta_1 x_i - y_i)^2$$

r_i: residuals

https://youtu.be/sDv4f4s2SB8

Garanti BBVA Technology

02

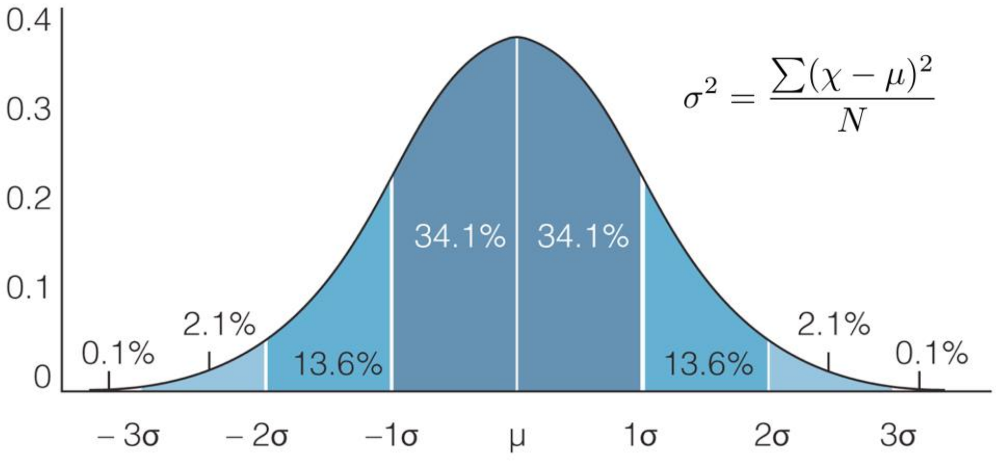
Terminology & Assumptions

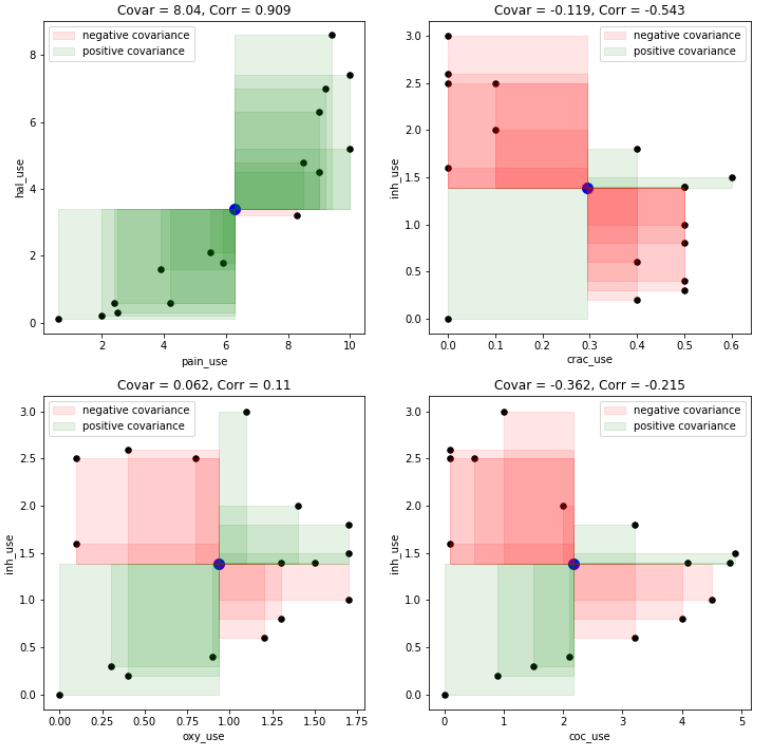
Terminology

Covariance

Correlation

Variance





Terminology
Covariance

The metric evaluates how much the variables change together.

- **Positive covariance**: Tend to move in the same direction.
- **Negative covariance**: Tend to move in inverse directions.
- Covariance (-Inf, + Inf)

Stockbroker

ABC company stock

BIST 100

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Cov(X,Y) = 9280$$

$$Cov(X,Y) = -56$$

Terminology

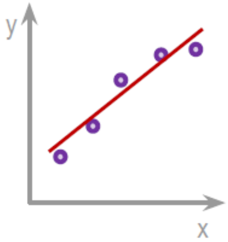
Pearson Correlation

How *linearly* 2 numerical variables behave

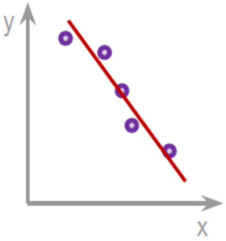
It shows the **direction of movement** of the variables and the **strength** of the relationship.

- Pearson Correlation [-1 to 1]

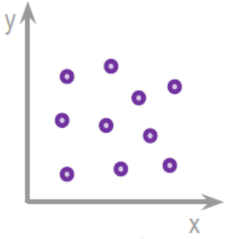
$$Corr(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



$\rho > 0$



$\rho < 0$



$\rho \sim 0$

It does not give cause-effect relationship !

Terminology

Covariance - Correlation

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Corr(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cov1 = X:TL , Y:USD

Cov2 = X:TL , Y:kg

Can not be comparable !

Independent from units.

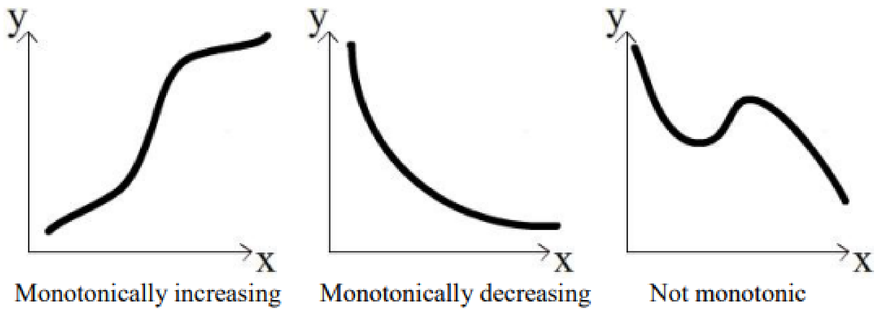
Correlation is just normalized covariance.

Can be comparable !

Terminology

Spearman's Rank Correlation

Spearman’s correlation coefficient is a statistical measure of the **strength of a monotonic relationship** between paired data ([-1 to 1]).

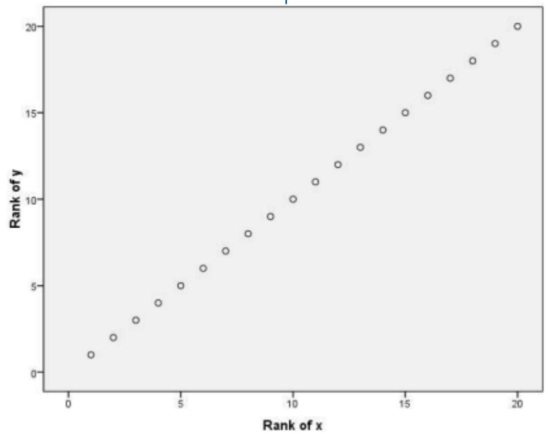


No Normality assumption anymore !

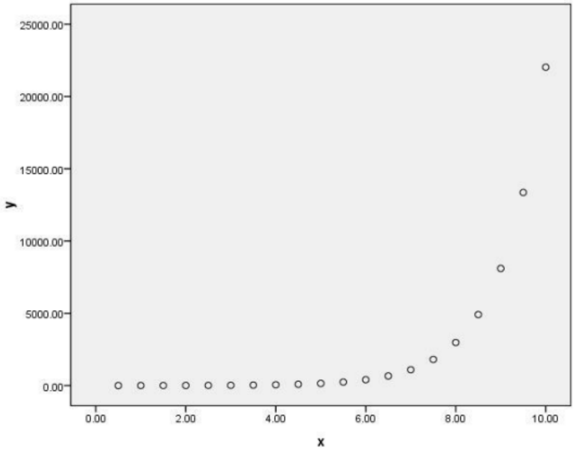
Example:

X=[10, 20, 30, 40, 1000] → [1.0, 2.0, 3.0, 4.0, 5.0]
Y= [−70,−1000,−50,−10,−20] → [2.0, 1.0, 3.0, 5.0, 4.0]

Linear relationship:



Monotonic relationship:



Potential Problems with the Model

Linear Regression is a *simple regression* model which offers certain **advantages**:

- Results interpretability
- Ease of use
- Low computational cost

Limitations of Linear Regression:

- **Simplistic in some cases**

Not great data that has not a linear relationship between Y and X.

- **Sensitivity to outliers**

Observation that is away from the major cluster of points have a squared impact.

- **Prone to poor performance**

Due to the various assumption, hard to capture structure of data.

- **Hard to tune in complex models**

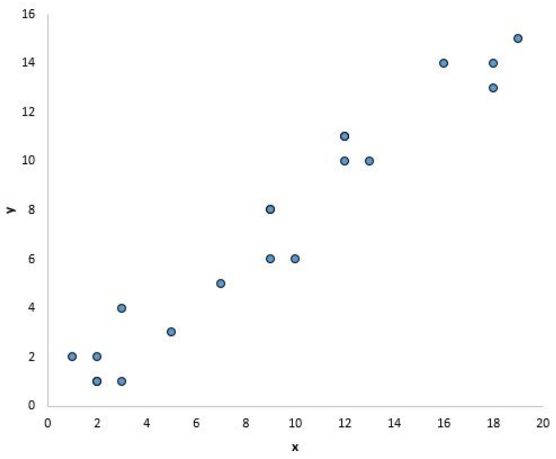
Too complex with many parameters and less data.

Assumptions of Linear Regression – Y and X

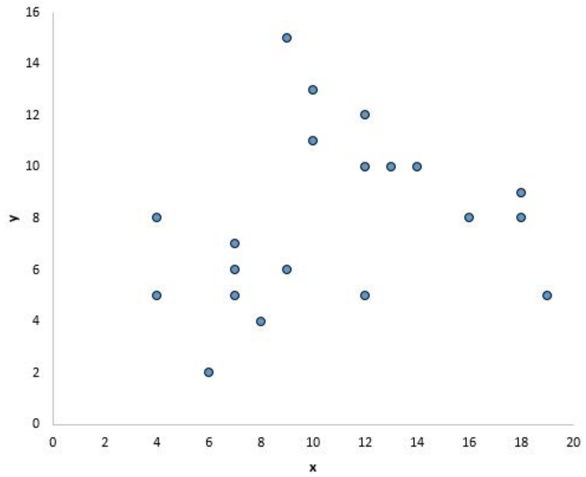
Assumption : Linear Relationship

There is a linear relationship between the independent variable **X**, and the independent variable **y.**

How to determine if this assumption is met:

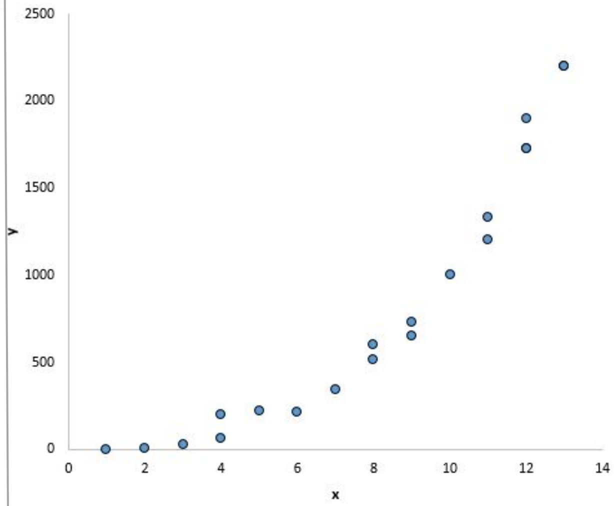


Linear relationship



No linear relationship

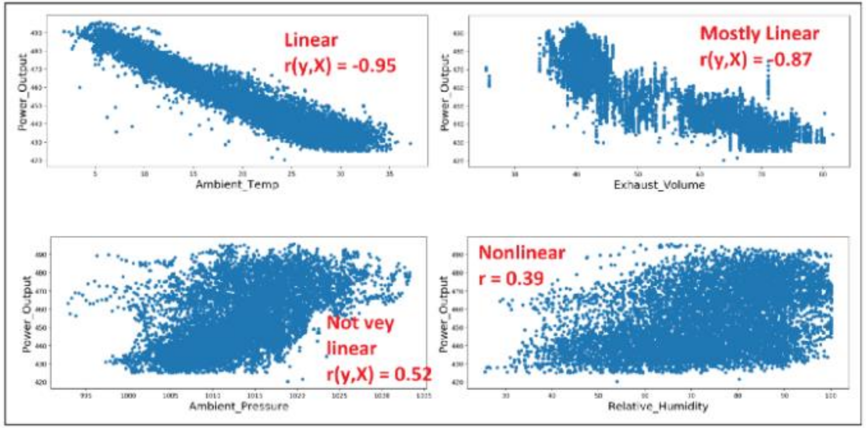
- Apply a nonlinear transformation to the independent and/or dependent variable.



Relationship, but not linear

Assumptions - Linear functional form

- There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s).
- A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X.
- An additive relationship suggests that the effect of X on Y is independent of other variables.

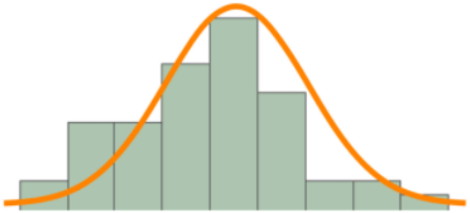


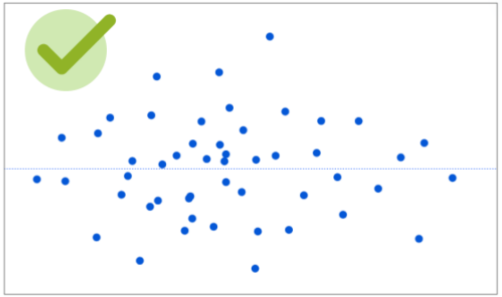

```
Ambient_Temp         -0.948128
Exhaust_Volume       -0.869780
Ambient_Pressure      0.518429
Relative_Humidity     0.389794
Power_Output          1.000000
Name: Power_Output, dtype: float64
```

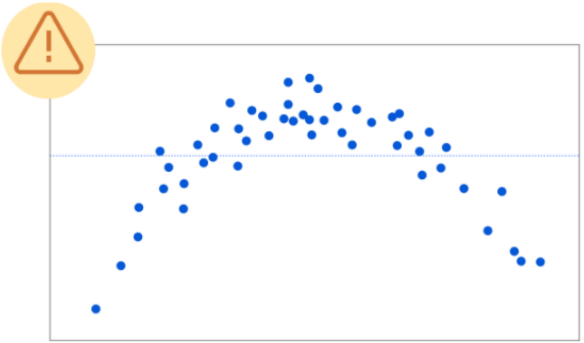
Assumptions - Residuals

The residuals are asumed to

- **be approximately normally distributed (with a mean of zero)**
- have a constant variance (*homoscedasticity*)
- be independent of one another (*no autocorrelation*)



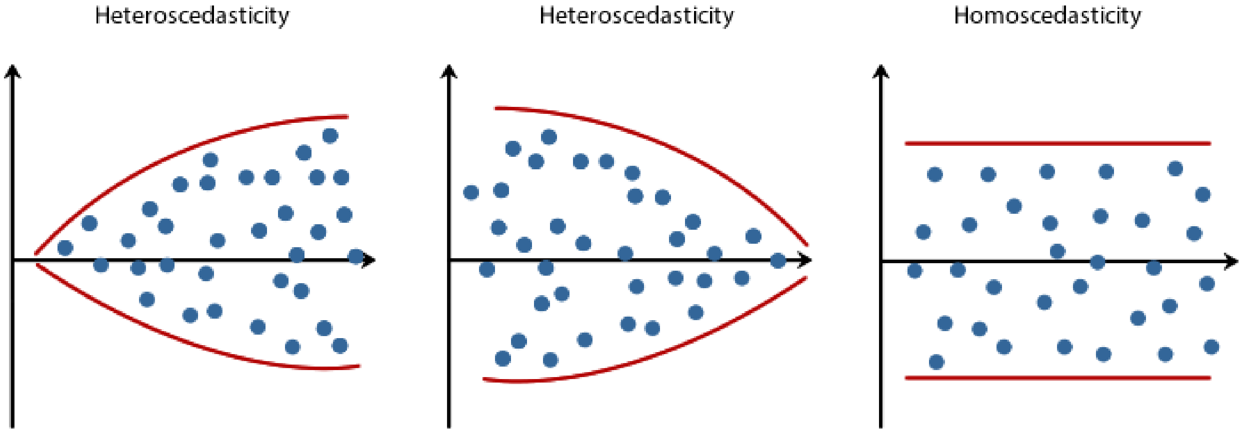




Assumptions - Residuals

The residuals are asumed to

- be approximately normally distributed (with a mean of zero)
- **have a constant variance (*homoscedasticity*)**
- be independent of one another (*no autocorrelation*)



Assumptions of Linear Regression - **Residuals**

Assumption : Equal Variances (Homoscedasticity)

Residuals have constant variance at every level of **X** known as *homoscedasticity*. When this is not the case, it is called as *heteroscedasticity*.

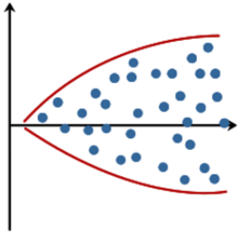
Example:

Family Income **(X)**

Luxury Spending **(Y)**

Low Family Income **; Error variation is low.**

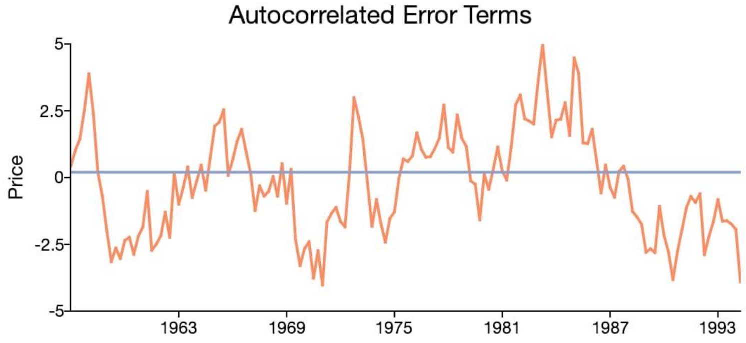
High Family Income **; Error variation is high**

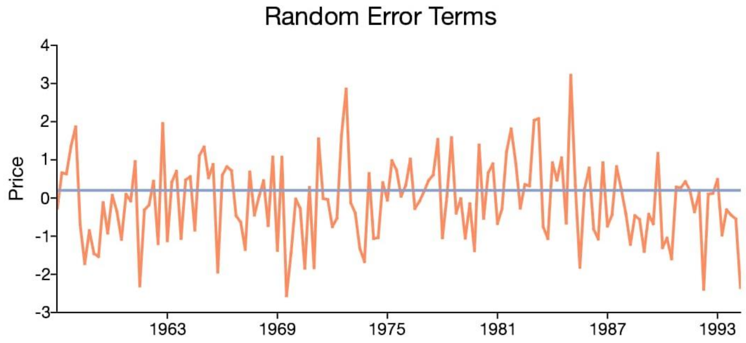


Assumptions - Residuals

The residuals are asumed to

- be approximately normally distributed (with a mean of zero)
- have a constant variance (*homoscedasticity*)
- **be independent of one another (*no autocorrelation*)**





03

Polynomial Regression

Polynomial Regression: Model Representation

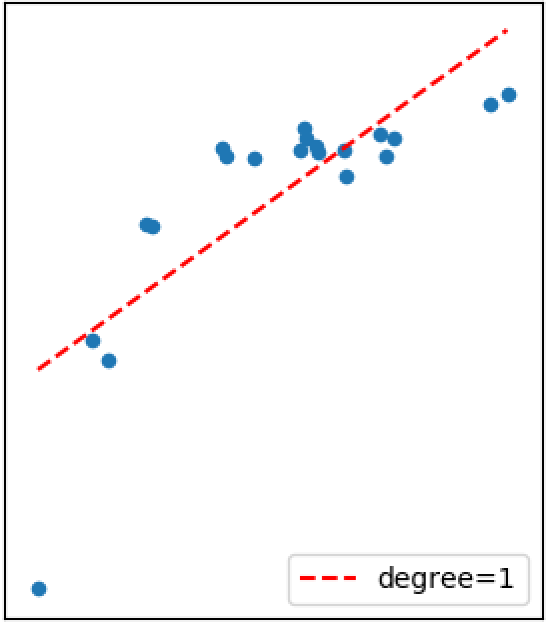
The behavior of the hypothesis function can be changed to represent our data better. We can create additional features based on x:

Quadratic Function: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$

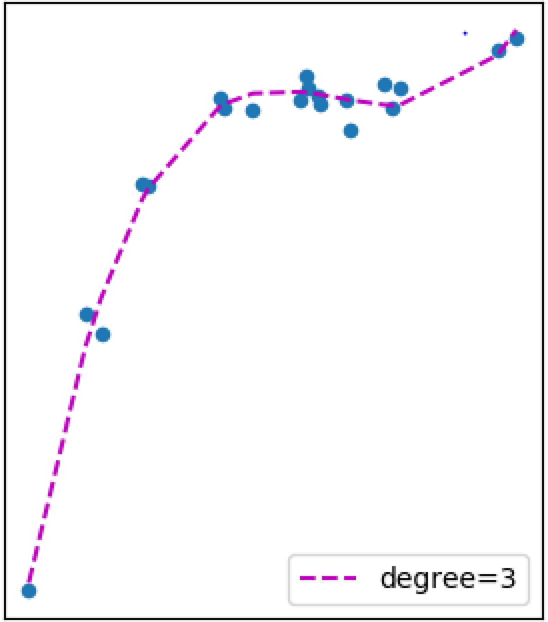
Cubic Function: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^3$

Square Root Function: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_1}$

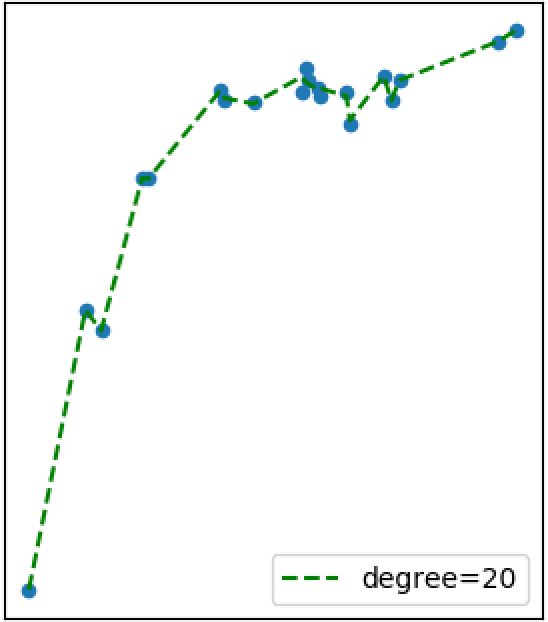
Polynomial Regression



Underfit
High Bias
Low Variance

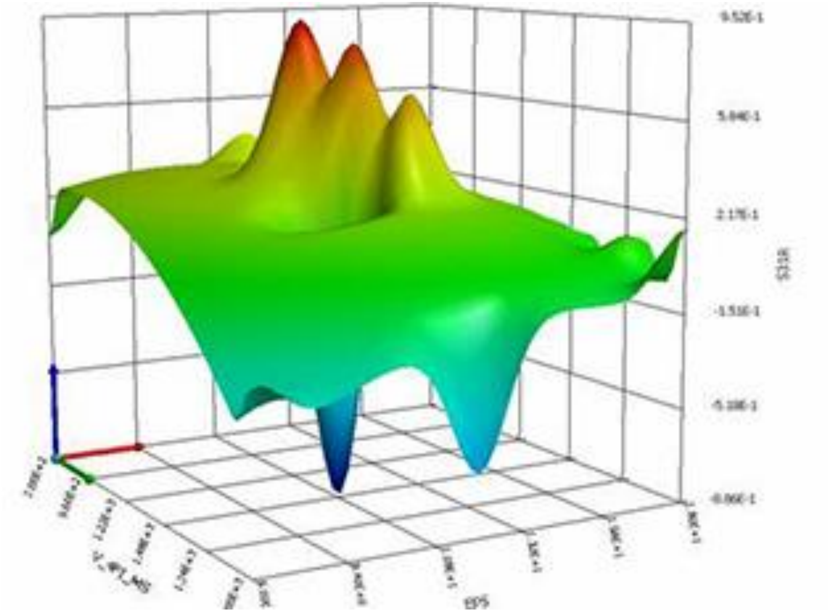


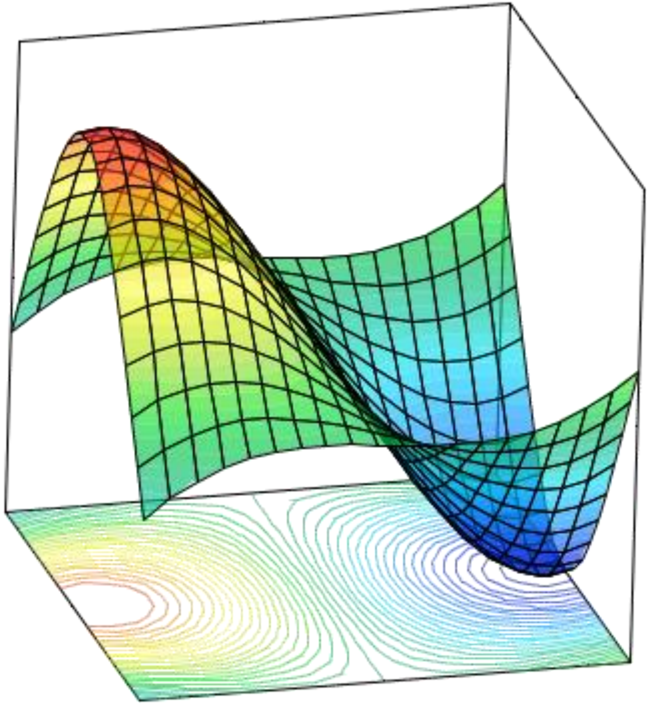
Correct Fit
Low Bias
Low Variance



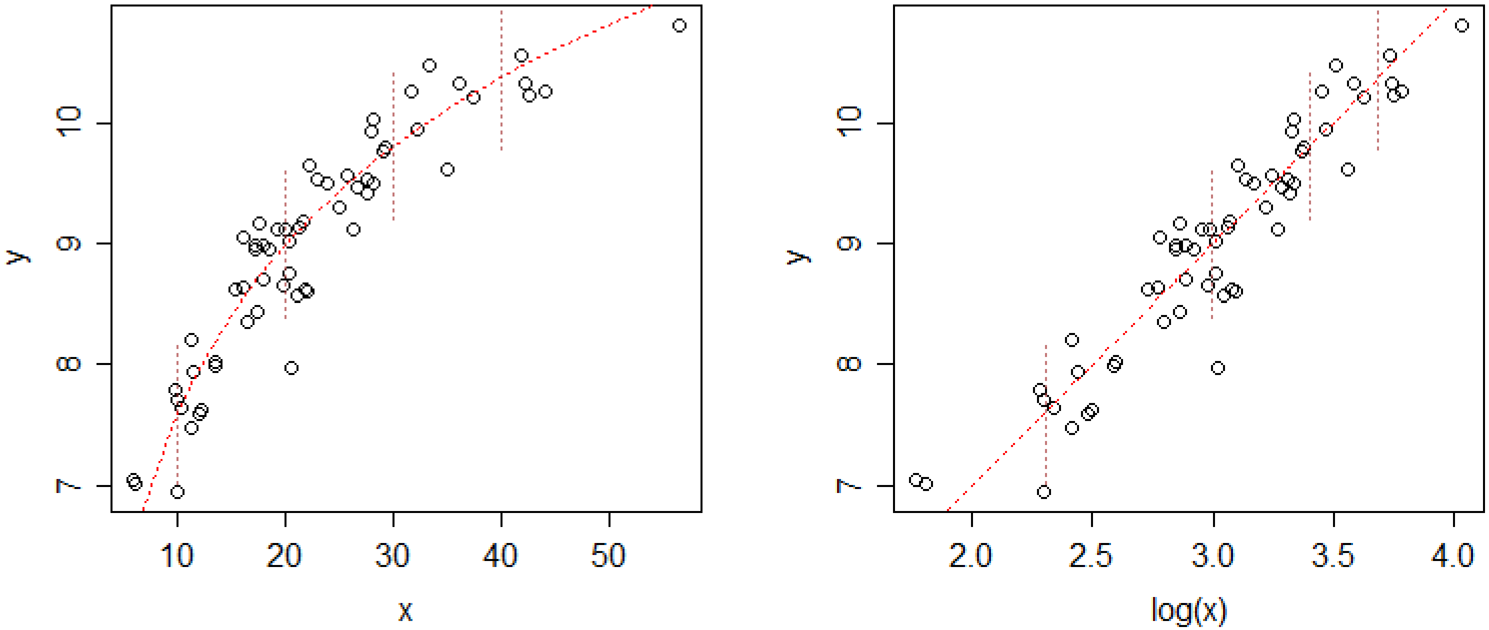
Overfit
Low Bias
High Variance

Polynomial Regression Optimization





Polynomial Regression & Logarithmic Data



04

Outliers

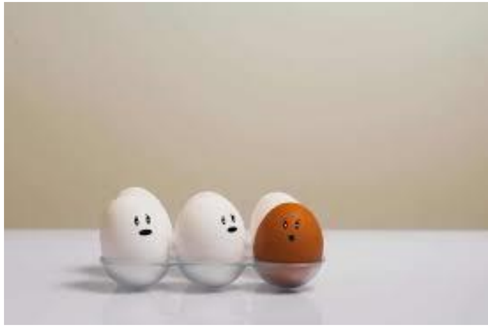
Data discrepancies

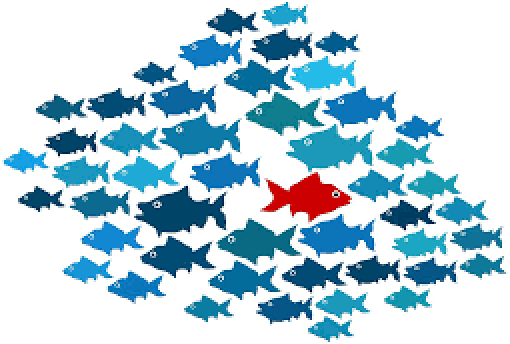
Strange values

We may find values in a dataset that just don't "fit«

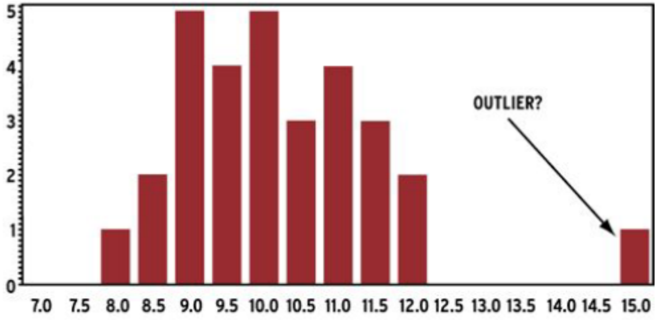
It may be because they are outside acceptable or admissible values

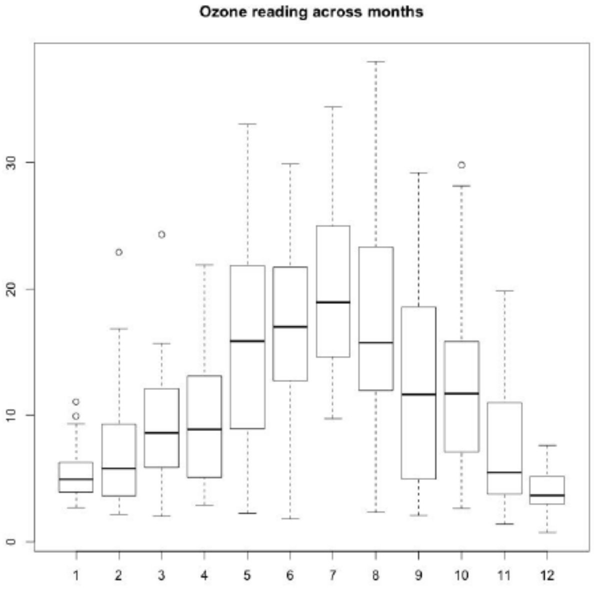
Atypical values or **Outliers**

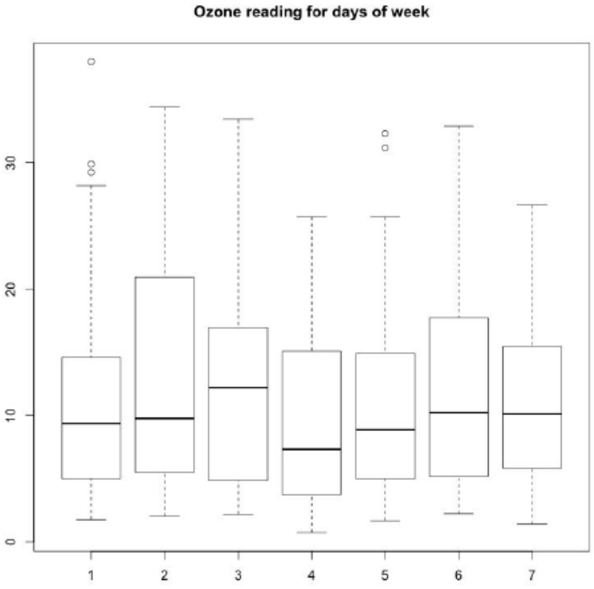




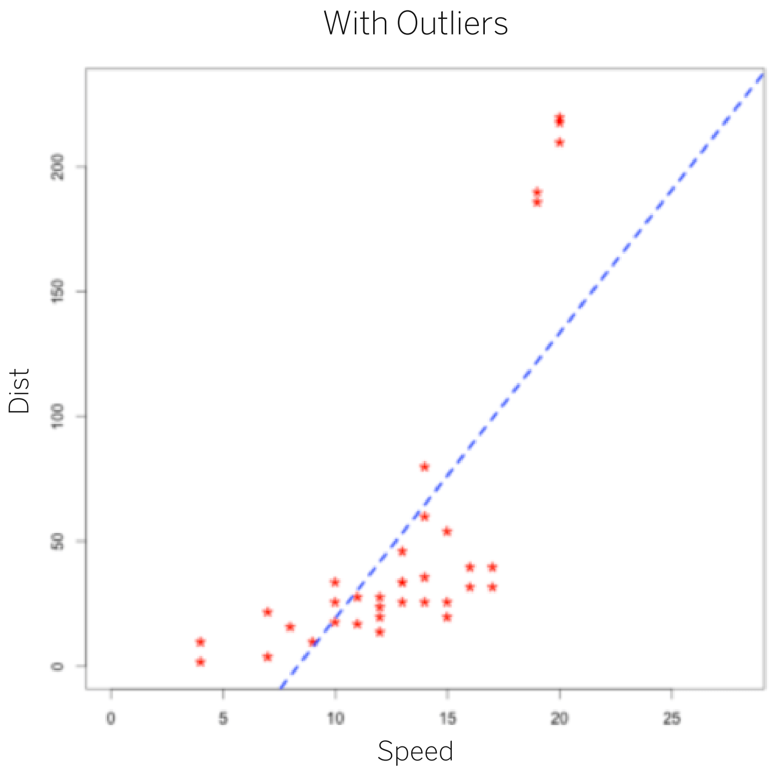
Outliers

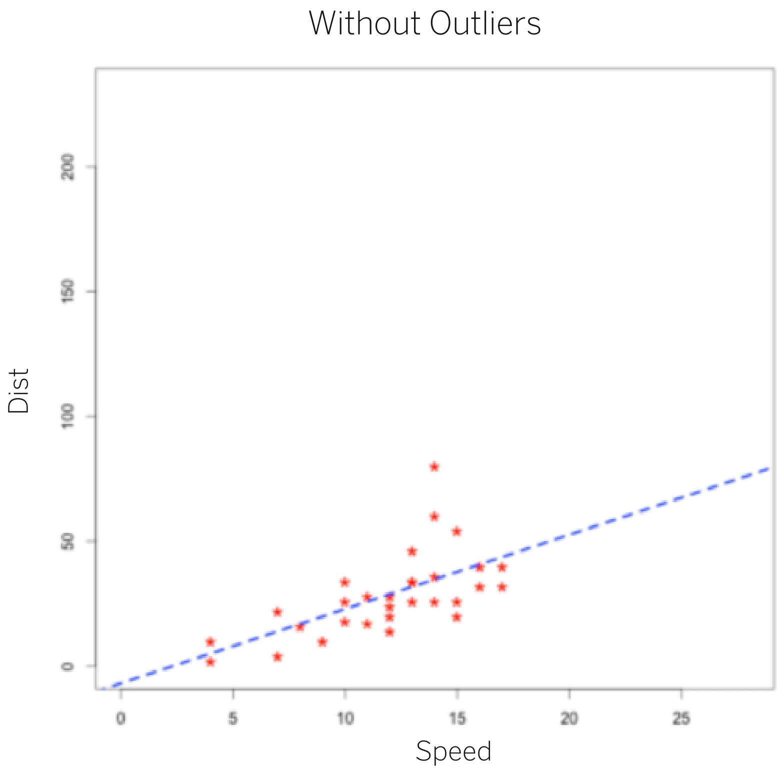

OUTLIER?
0
1
2
3
4
5
7.0 7.5 8.0 8.5 9.0 9.5 10.0 10.5 11.0 11.5 12.0 12.5 13.0 13.5 14.0 14.5 15.0


Ozone reading across months
0
10
20
30
1 2 3 4 5 6 7 8 9 10 11 12


Ozone reading for days of week
0
10
20
30
1 2 3 4 5 6 7

Outliers


With Outliers
Dist
Speed
0
50
100
150
200
0
5
10
15
20
25


Without Outliers
Dist
Speed
0
50
100
150
200
0
5
10
15
20
25

05

Balancing Bias And Variance

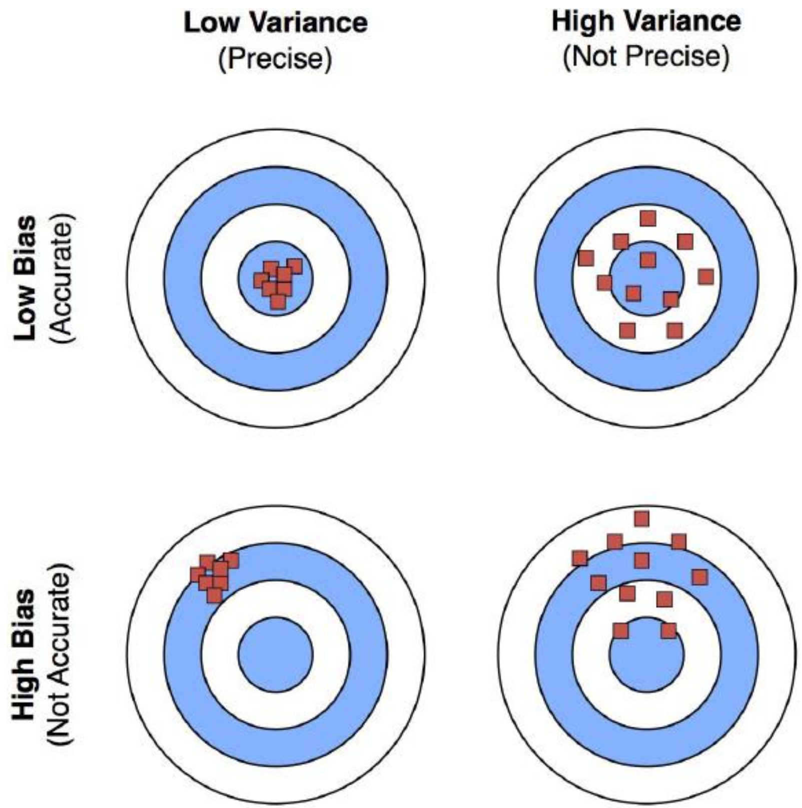
Bias vs. Variance

Bias: An error caused by the **difference between the model prediction and the correct value.**

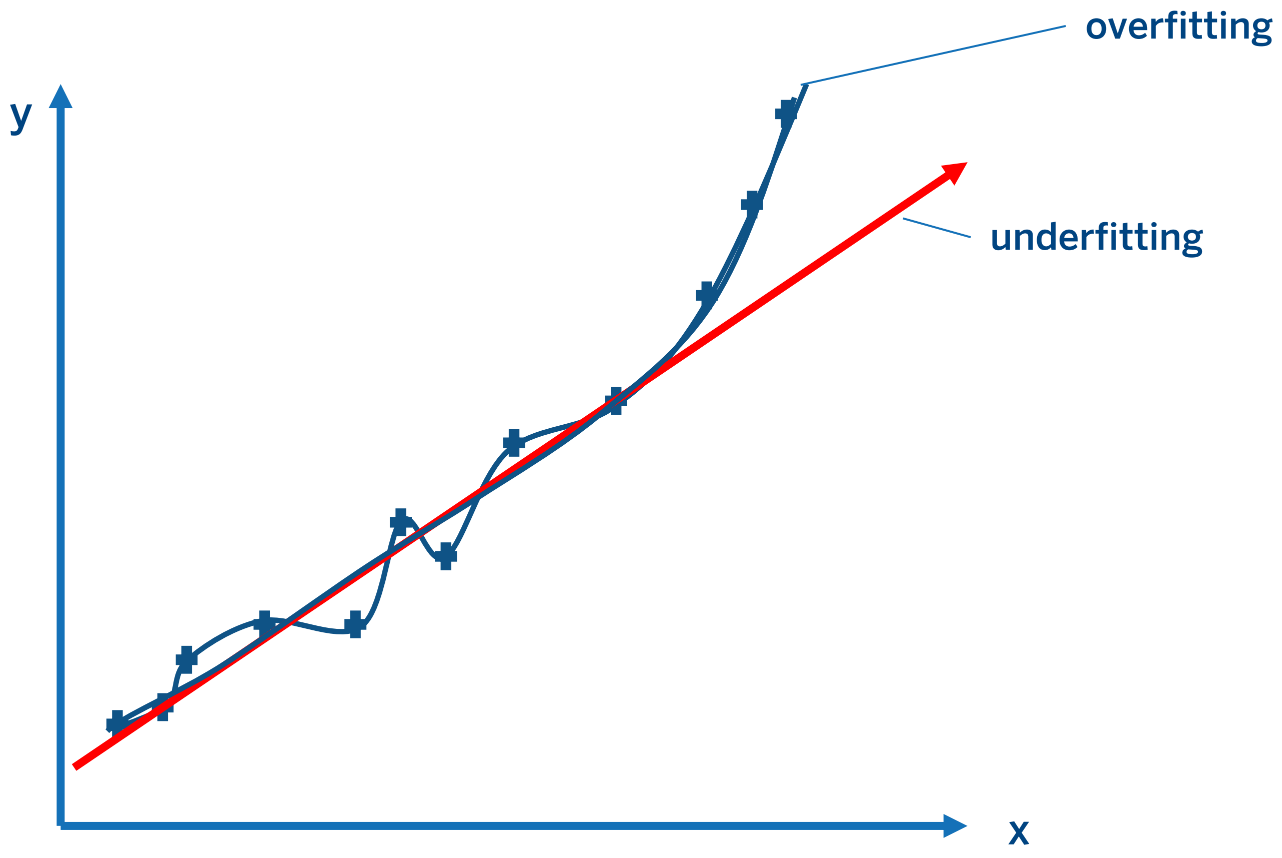
- It is minimized by increasing the model coplexity.

Variance: An error caused by the **sensitivity of the model to minor variations in the training data.**

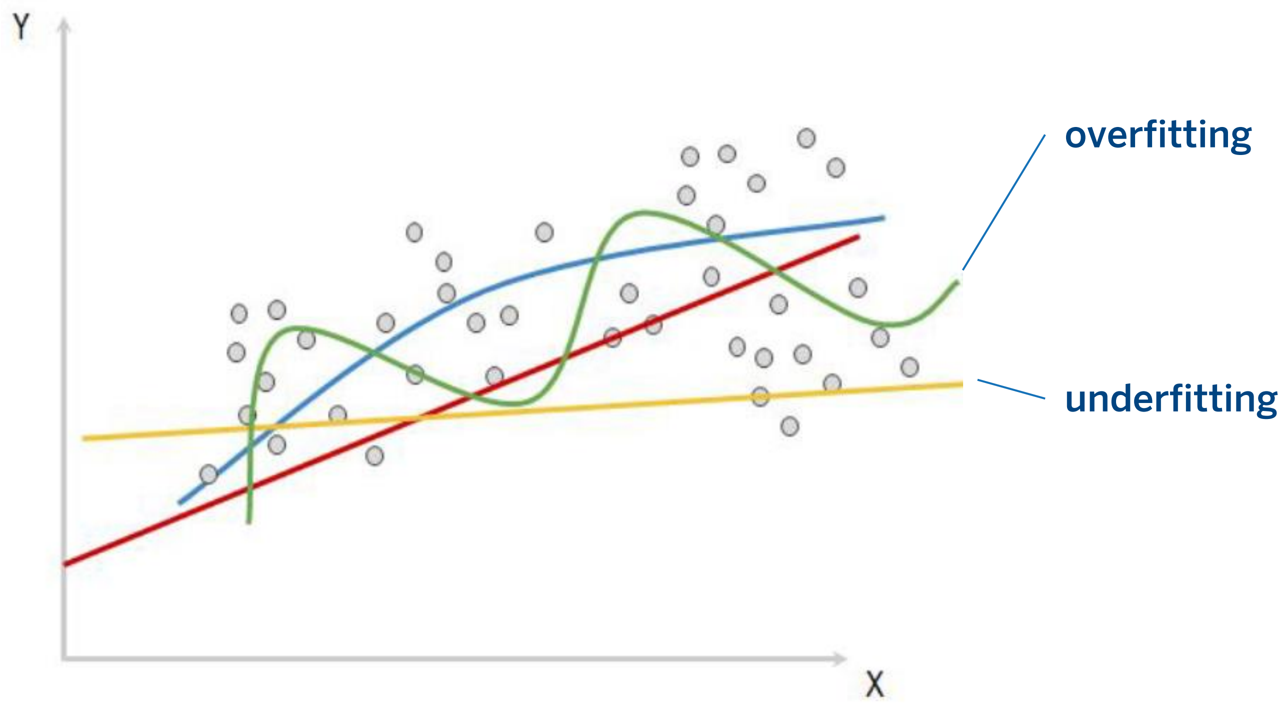
- It is minimized by decreasing the complexity of the model.



Regression Models



Bias vs. Variance



Bias: An error caused by the difference between the model prediction and the correct value.

Variance: An error caused by the sensitivity of the model to minor variations in the training data.

A problem inherent to the modeling process

All the models must balance the bias and the variance

The predictions made with the model have a combined error of:
Bias + Variance + Irreducible error

- A bias error occurs because the model is too simple.
- A variance error occurs because the model is too complex.
- Balanced model = minimizes the sum of bias and variance errors

06

Data Transformation

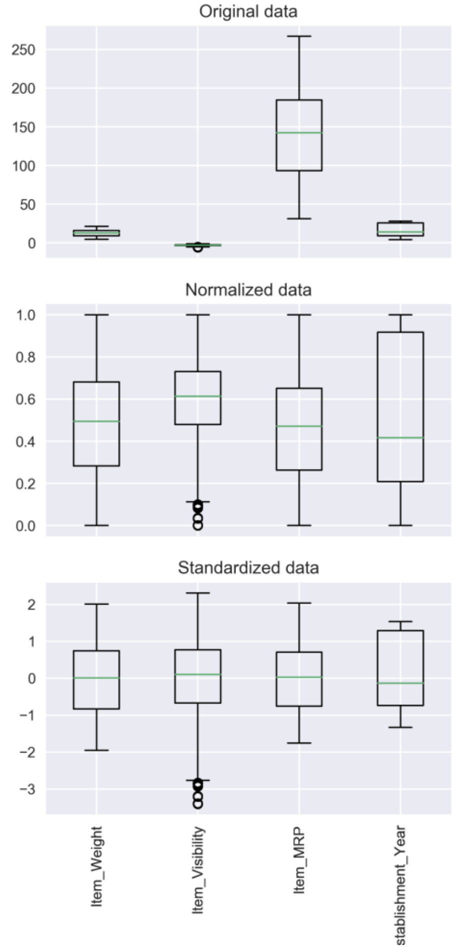
Normalization vs. Standardization

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

Mean $$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

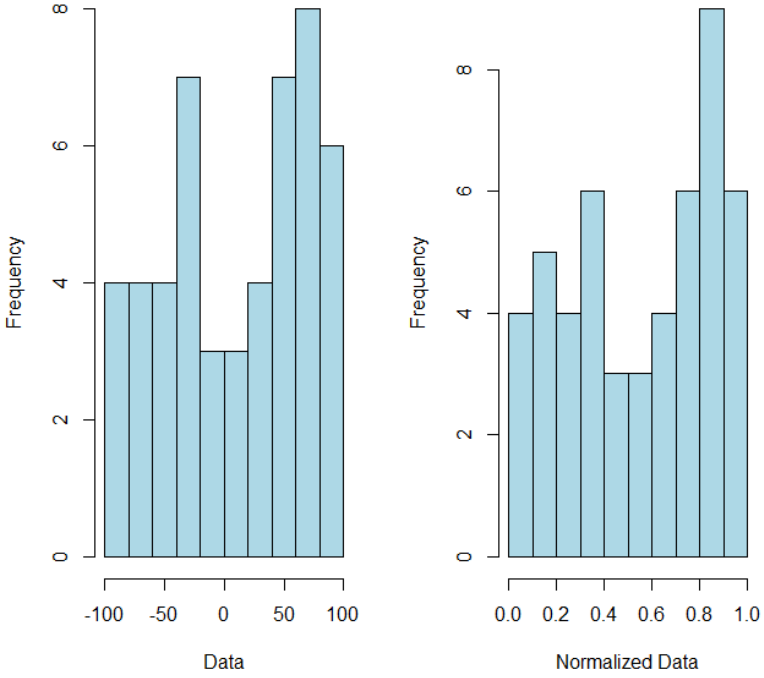
Standard Deviation $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

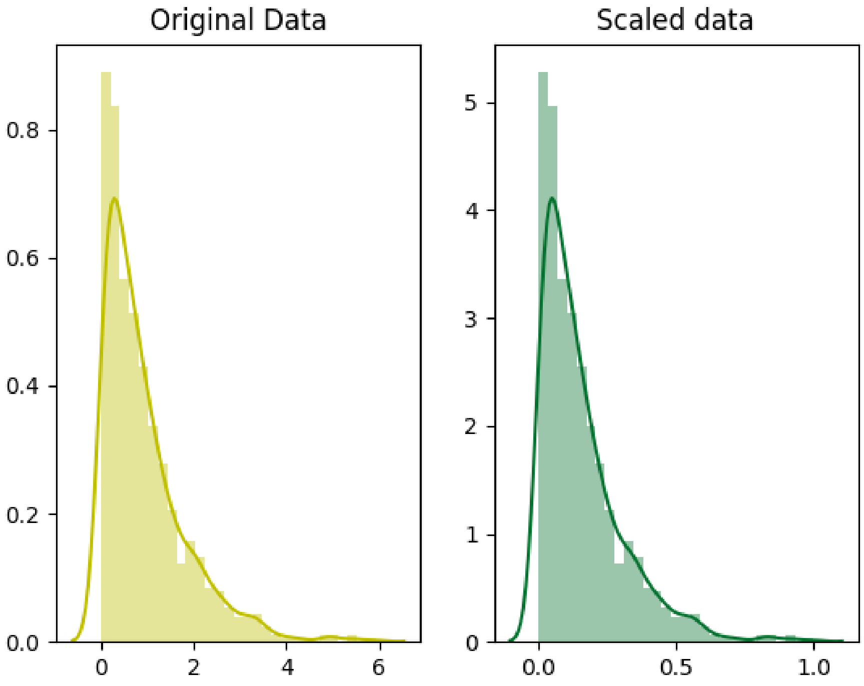
Min-Max Scaling $$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$



Normalization

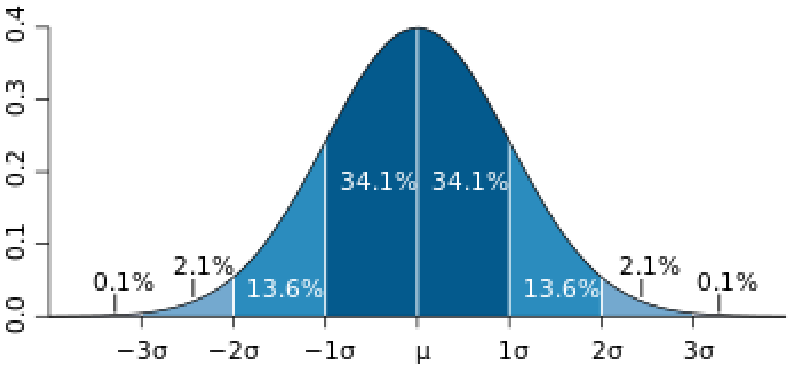
$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

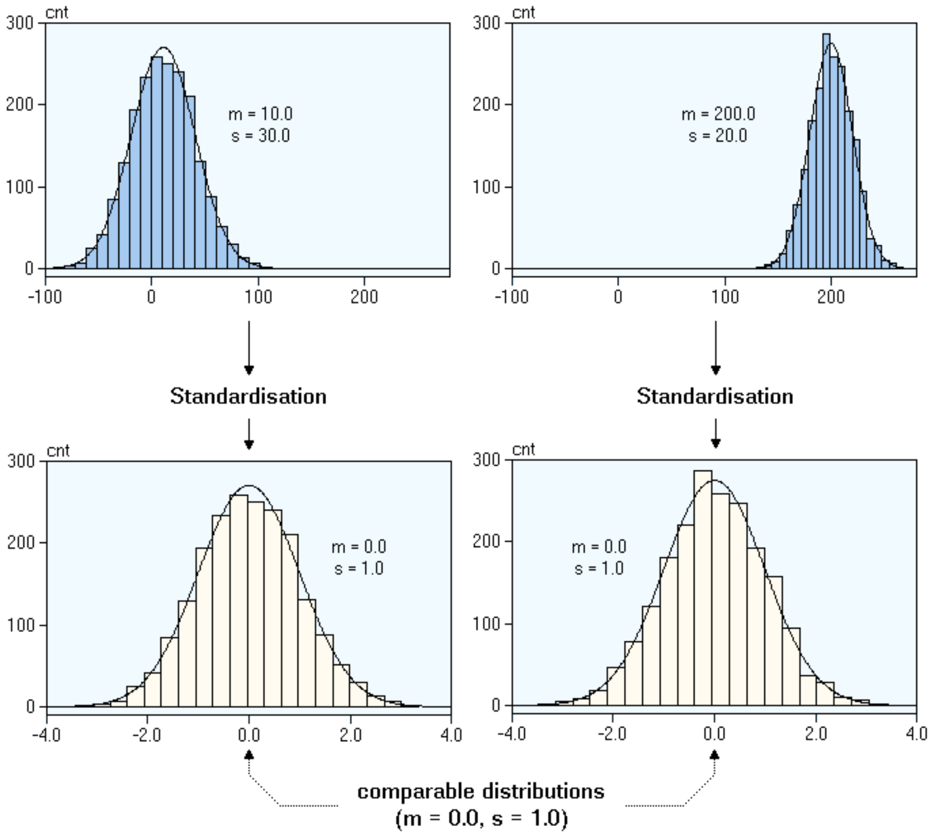




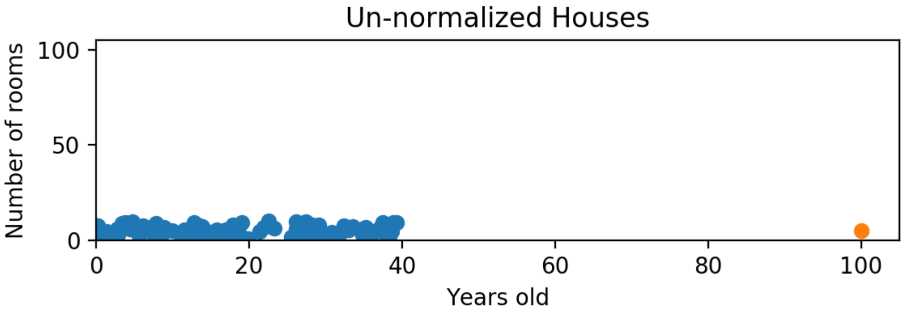
Standardization

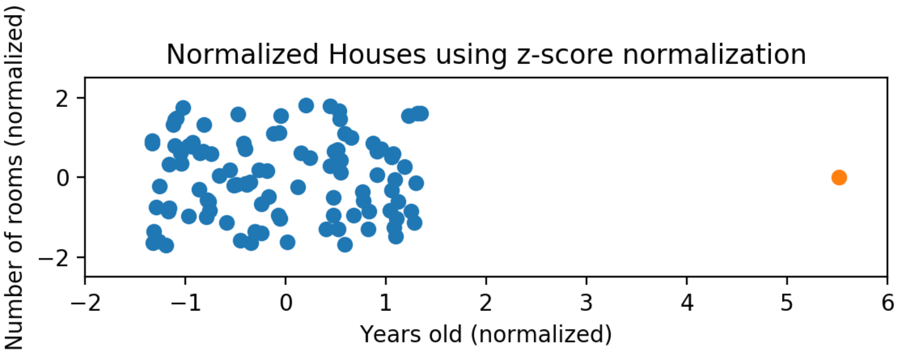
$$z = \frac{x - \mu}{\sigma}$$

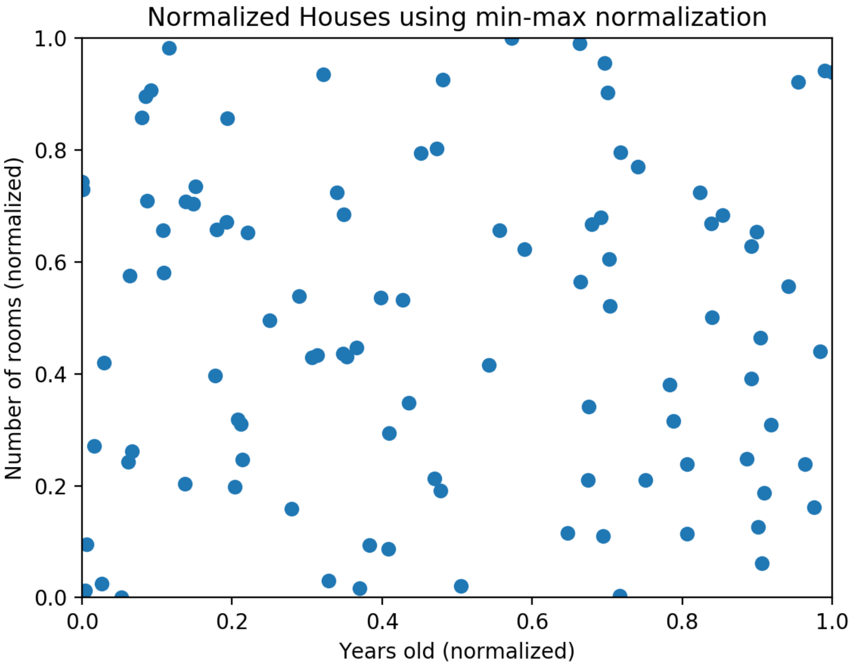




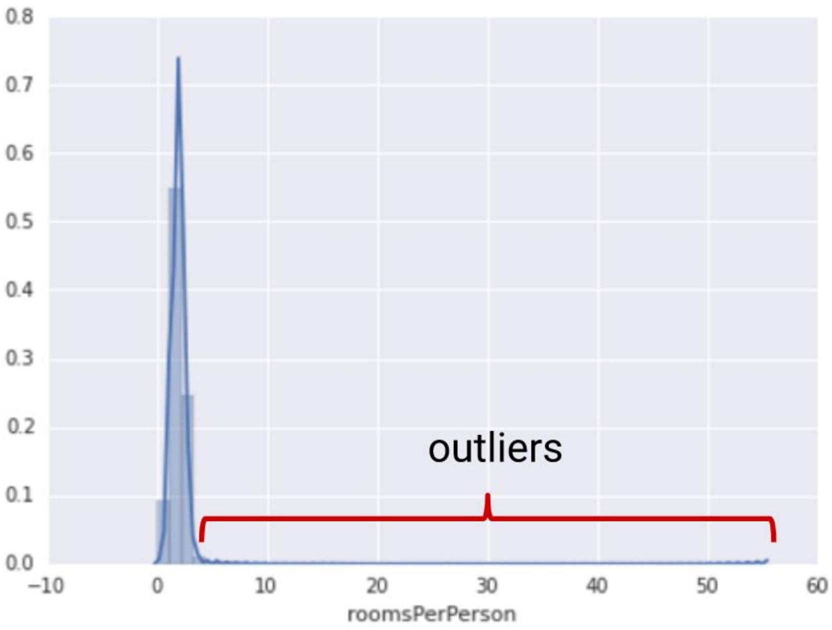
Normalization or Standardization



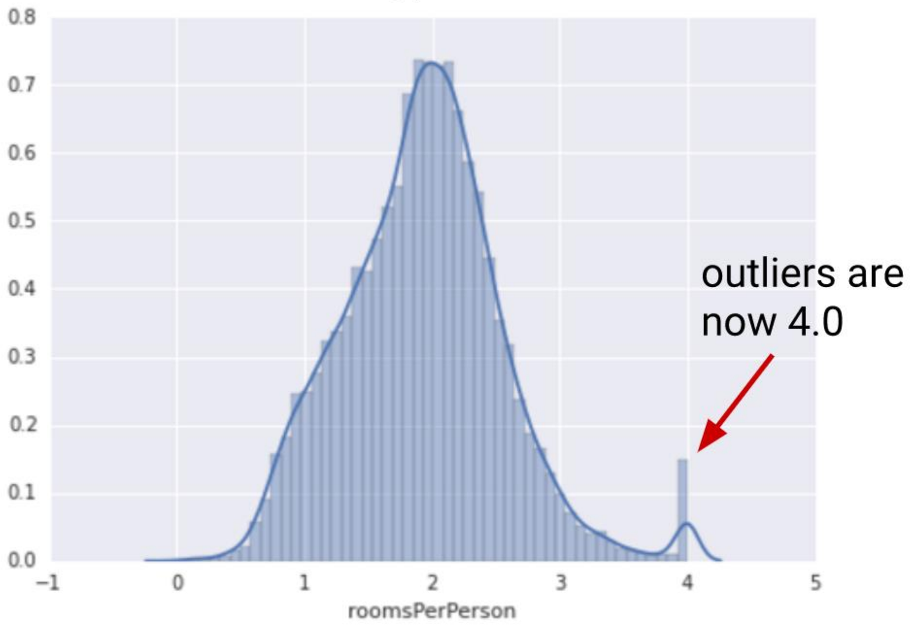




Capping Data



Same feature, capped to a max of 4.0



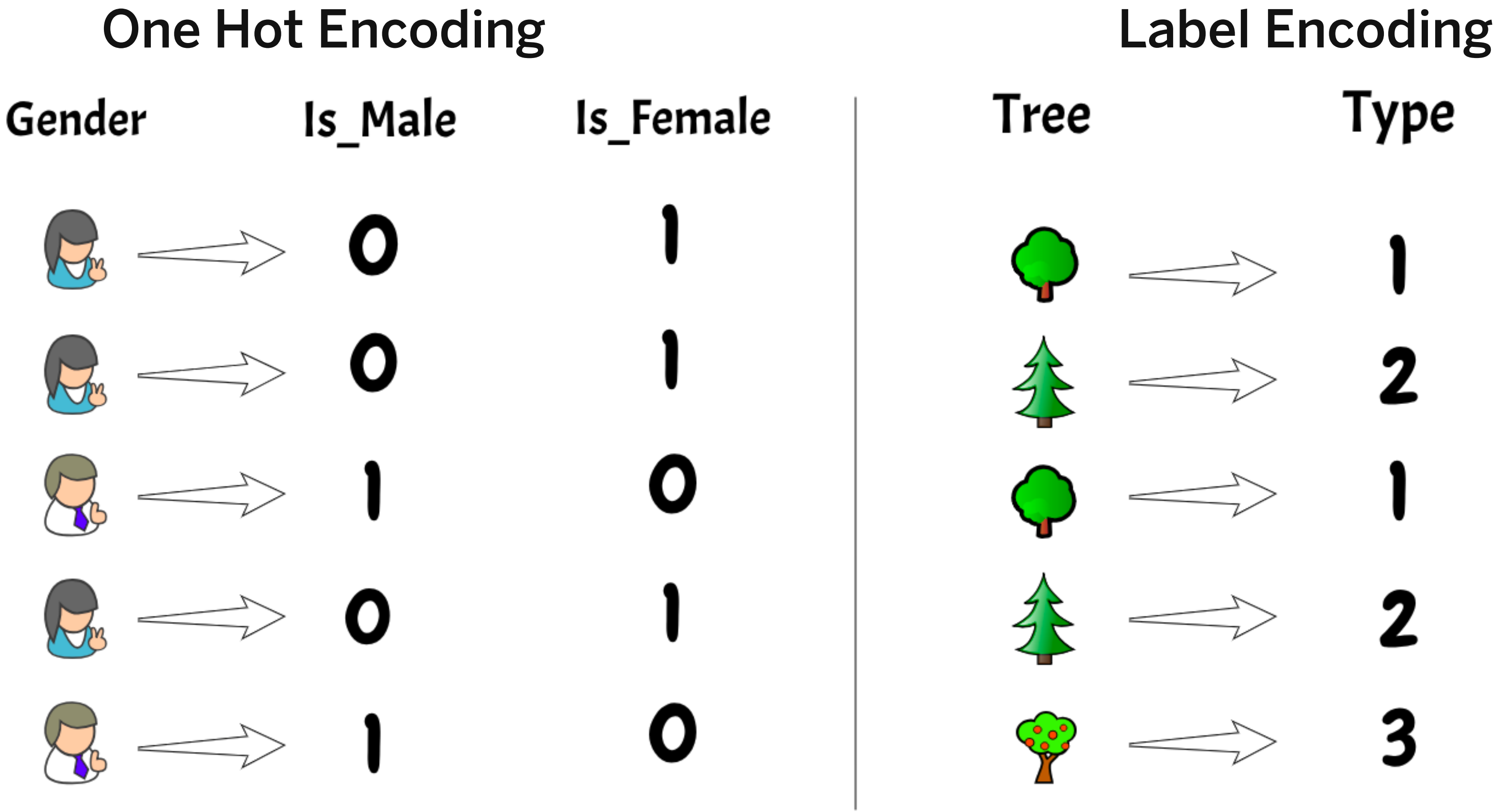
Handle Missing Values

1.Deleting Rows with missing values
2.Impute missing values for continuous variable (Mean, Mode, Median)
3.Impute missing values for categorical variable (Mode or New)
4.Prediction of missing values

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Handle Categorical Data



Handle Categorical Data

Label & One Hot Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Handle Categorical Data

One Hot Encoding Example 1

BRIDGE-TYPE (TEXT)	BRIDGE-TYPE (NUMERICAL)
Arch	0
Beam	1
Truss	2
Cantilever	3
Tied Arch	4
Suspension	5
Cable	6

BRIDGE-TYPE (TEXT)	BRIDGE-TYPE (Arch)	BRIDGE-TYPE (Beam)	BRIDGE-TYPE (Truss)	BRIDGE-TYPE (Cantilever)	BRIDGE-TYPE (Tied Arch)	BRIDGE-TYPE (Suspension)	BRIDGE-TYPE (Cable)
Arch	1	0	0	0	0	0	0
Beam	0	1	0	0	0	0	0
Truss	0	0	1	0	0	0	0
Cantilever	0	0	0	1	0	0	0
Tied Arch	0	0	0	0	1	0	0
Suspension	0	0	0	0	0	1	0
Cable	0	0	0	0	0	0	1

Handle Categorical Data

One Hot Encoding Example 2

SAFETY-LEVEL (TEXT)	SAFETY-LEVEL (NUMERICAL)
None	0
Low	1
Medium	2
High	3
Very-High	4

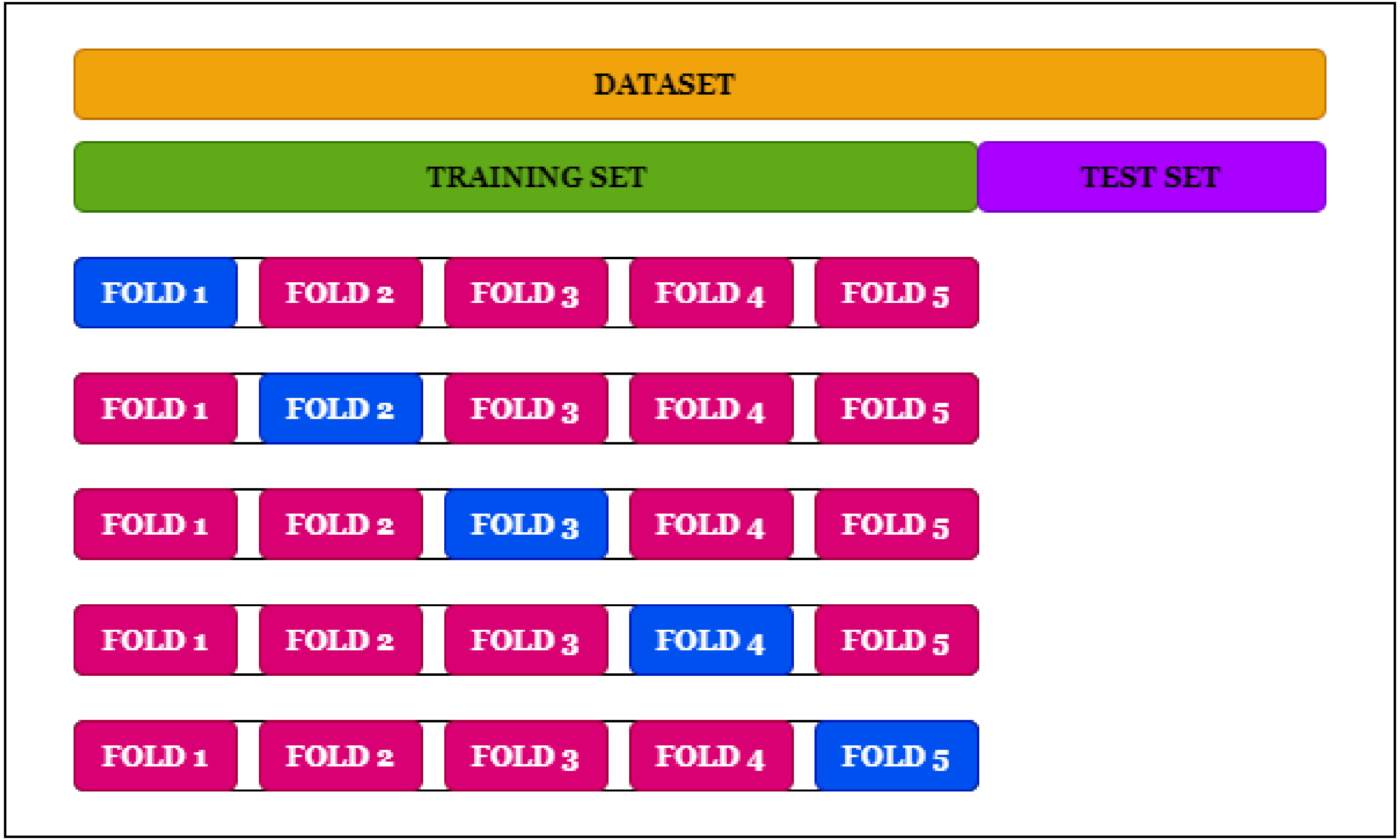
SAFETY-LEVEL (TEXT)	SAFETY-LEVEL (None)	SAFETY-LEVEL (Low)	SAFETY-LEVEL (Medium)	SAFETY-LEVEL (High)	SAFETY-LEVEL (Very High)
None	1	0	0	0	0
Low	0	1	0	0	0
Medium	0	0	1	0	0
High	0	0	0	1	0
Very-High	0	0	0	0	1

Garanti BBVA Technology

07

Splitting Data: Train & Test

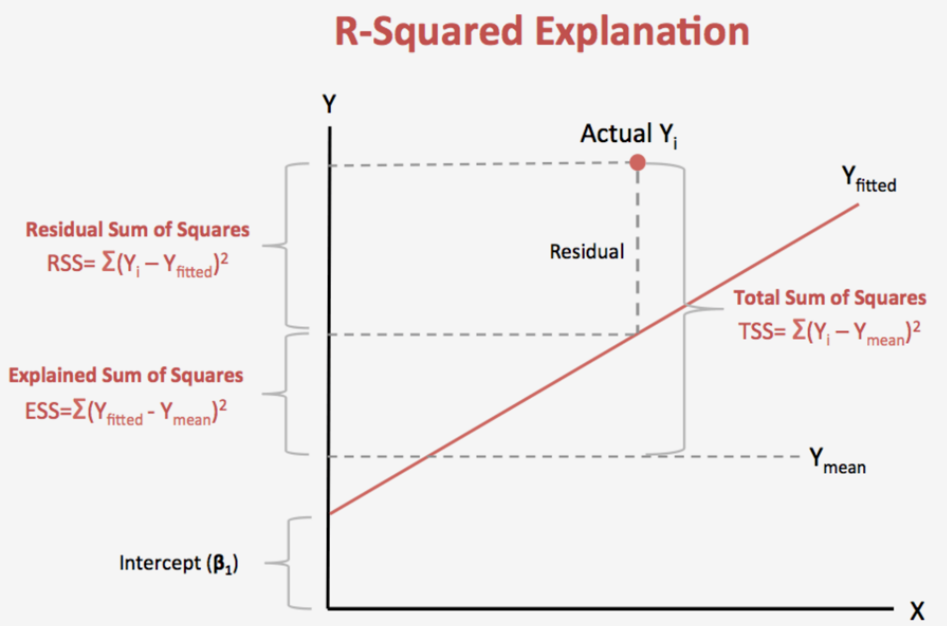
Split Data (Train & Test)– K-Fold



08

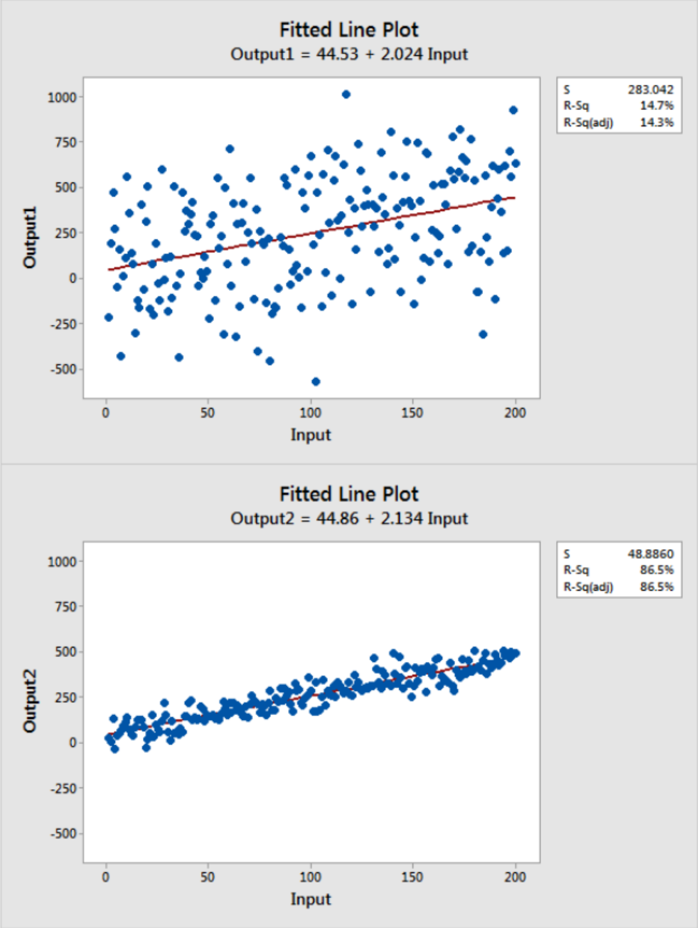
Performance Metrics

R^2 And Adjusted R^2 Calculation



$$R^2_{adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

$$R_{Sq} = 1 - \frac{RSS}{TSS}$$



https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

$\mathbf{R^2}$ And Adjusted $\mathbf{R^2}$ Calculation

The Formula for R-Squared Is

$$\mathrm{R}^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Adjusted $\boldsymbol{R^2}$: $R^2 - (1 - R^2)\frac{p}{n-p-1}$

https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/ https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

Regression Models Metrics

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/

https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

Örnekler

https://www.kaggle.com/mihirhalai/sydney-house-prices

https://www.kaggle.com/hellbuoy/car-price-prediction

https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/ https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values

Garanti BBVA
Technology

Teşekkürler

09

Relationship Between Variables

Chi Square

How homogeneous the relationship is between 2 categorical variables

To determine if there is a significant difference between the expected and observed requencies in one or more categorical variables.

$$x^2 = \sum \frac{(Obs\ Freq\ - Exp\ Freq)^2}{Exp\ Freq}$$

We establish a **critical probability** (alpha=0.05) and **we compare the probability associated with our chi^2 in the Chi Square distribution**, for (n - 1)*(m - 1) degrees of freedom.

- if P(alpha) >= P(chi), there is no significant difference
- if P(alpha) < P(chi), there is a significant difference

Age Group/Sex	F	M	Total
Young	2	3	5
Adult	2	1	3
Old	3	0	3
Total	7	4	11

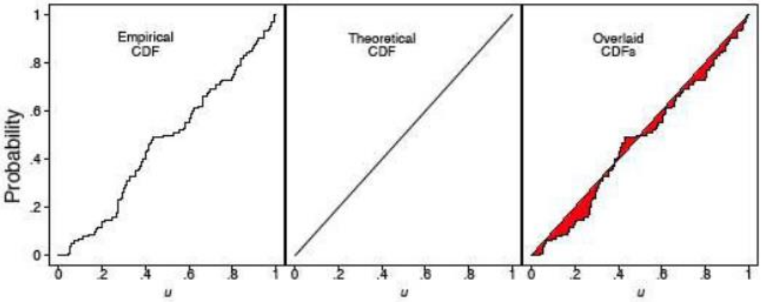
Age Group/Sex	F	M	Total
Young	3,18	1,81	5
Adult	1,90	1,09	3
Old	1,90	1,09	3
Total	7	4	11

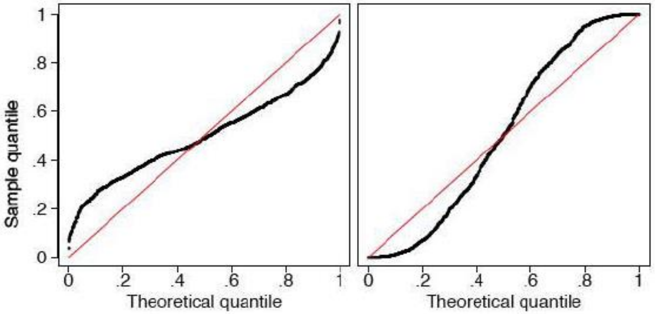
QQ - Plot

Compare distribution of data with known values

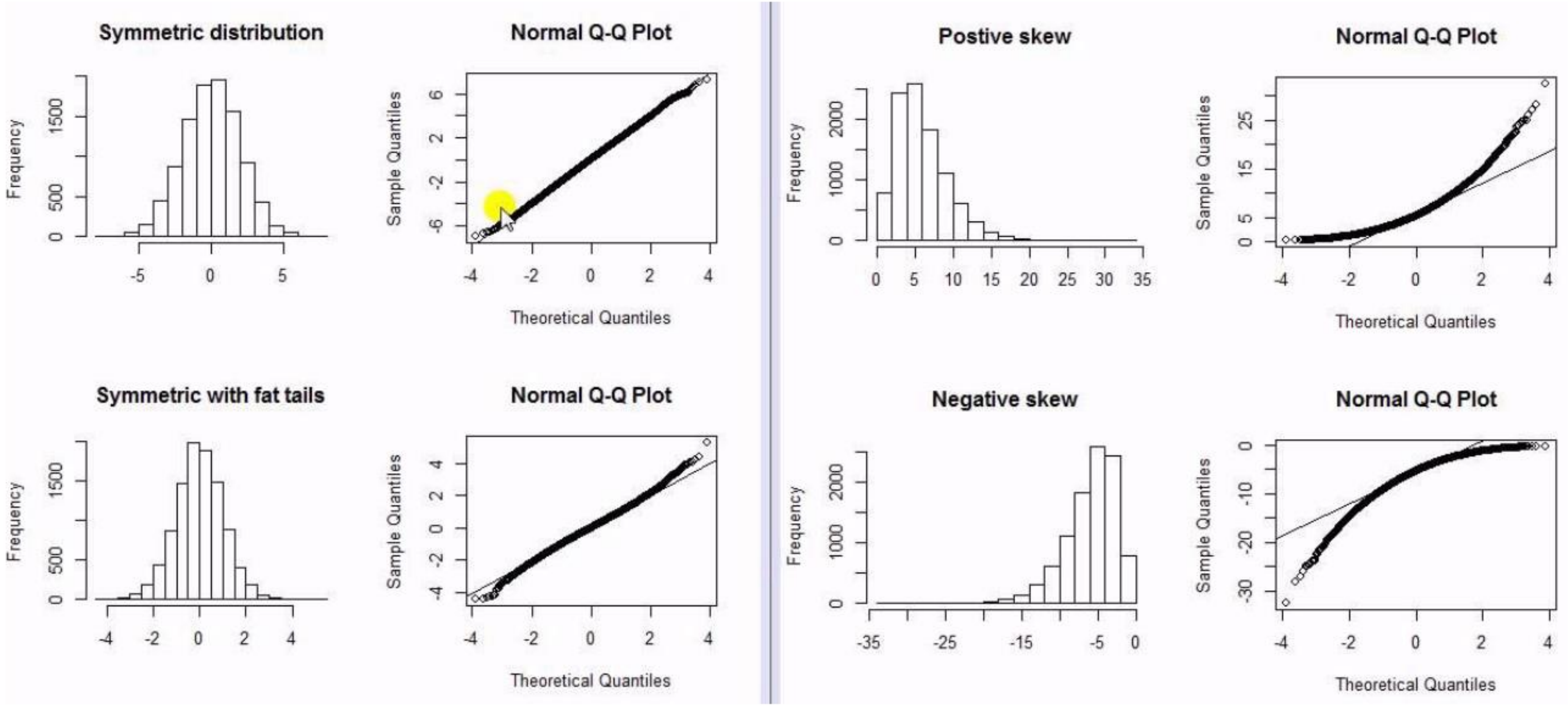
- Are there 2 datasets with common distributions?
- Do the distributions of 2 sets have the same form?
- Do they behave the same at extreme values?
- Is their distribution normal (or like another distribution type)?

Comparisons are done with quantile to quantile



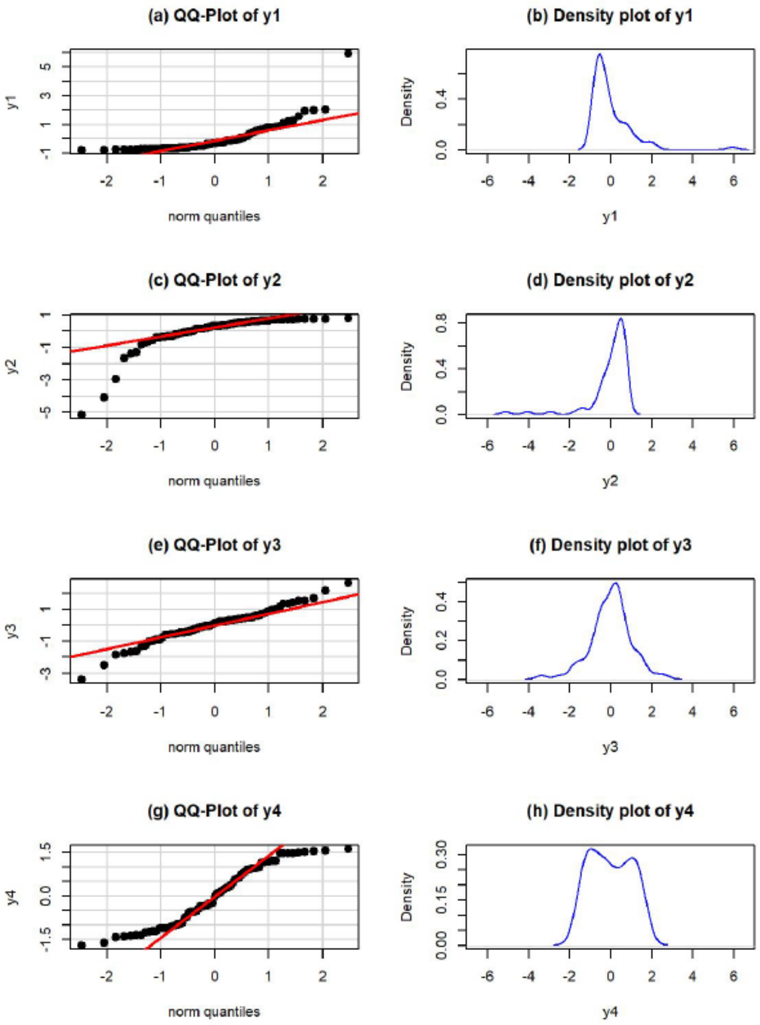


QQ - Plot



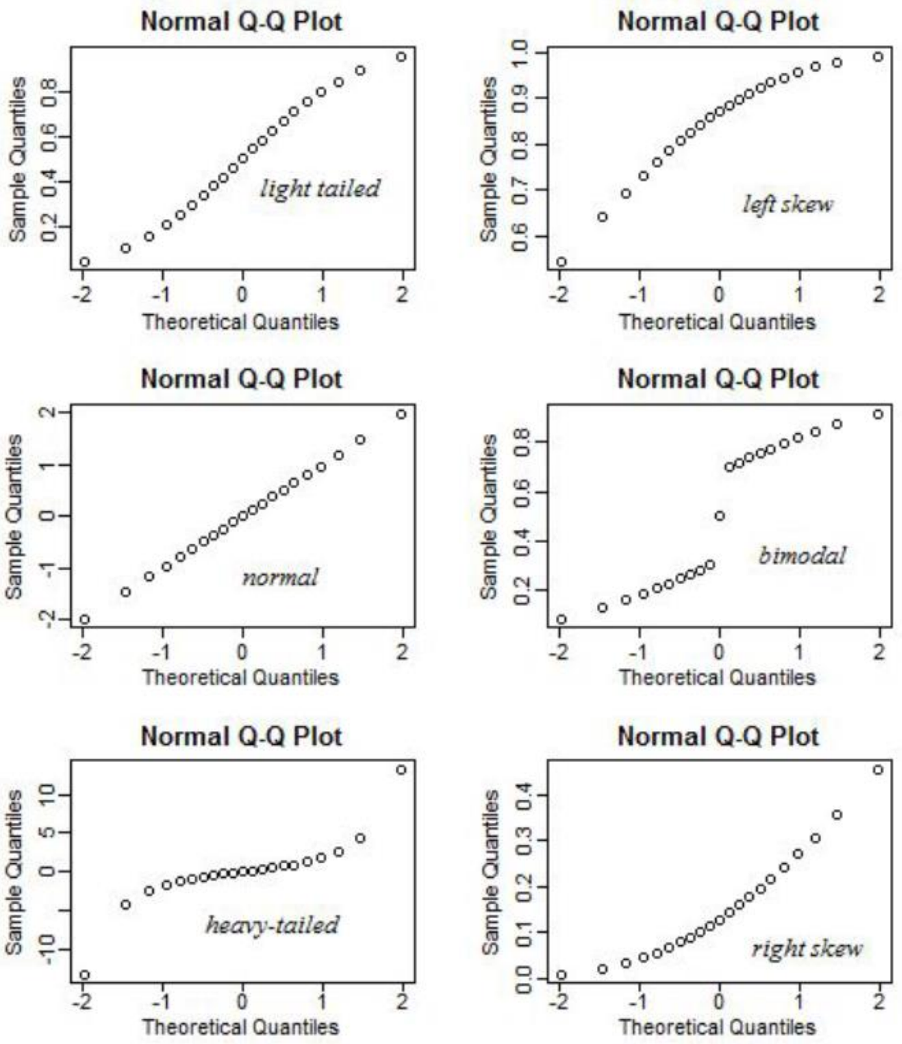
Symmetric distribution
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Postive skew
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Symmetric with fat tails
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
Negative skew
Frequency
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

QQ - Plot



(a) QQ-Plot of y1
(b) Density plot of y1
(c) QQ-Plot of y2
(d) Density plot of y2
(e) QQ-Plot of y3
(f) Density plot of y3
(g) QQ-Plot of y4
(h) Density plot of y4
norm quantiles
Density
y1
y2
y3
y4

QQ - Plot

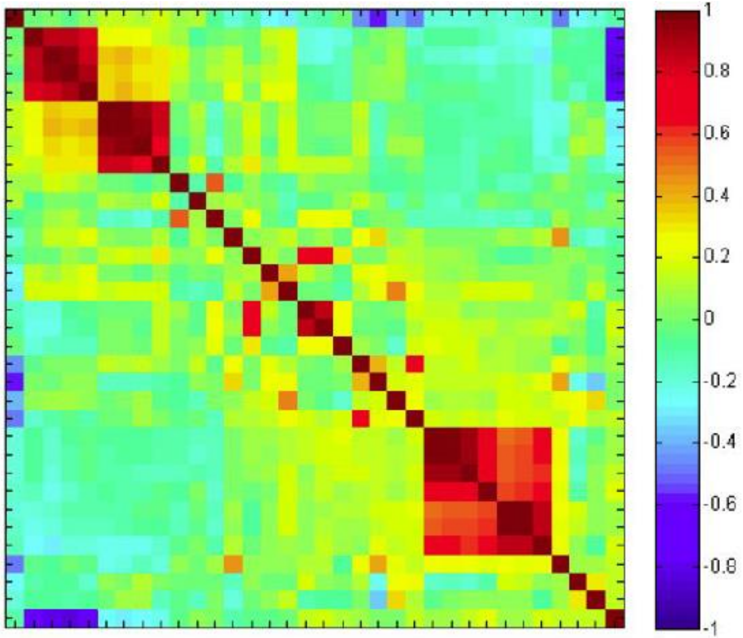

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
light tailed
left skew
normal
bimodal
heavy-tailed
right skew
-2
-1
0
1
2

Correlation Map

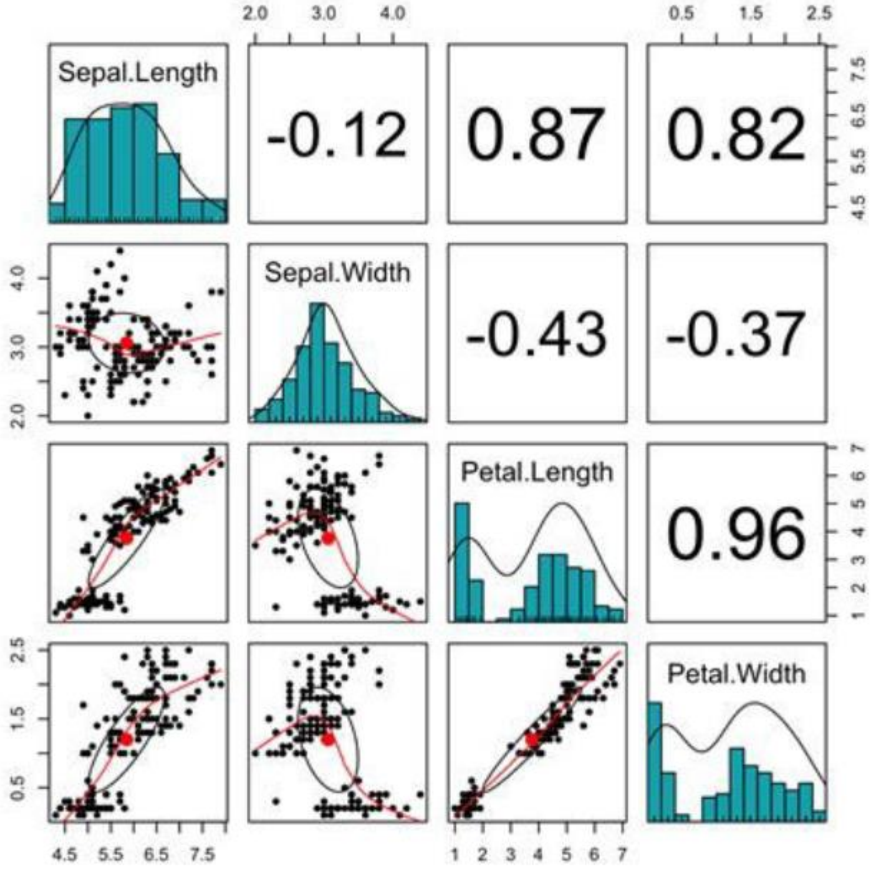
Relationship between numerical values at a glance

3 or more numerical variables

"Drawing a square matrix with as many rows as numerical values, representing the correlation of each pair with a color scale from -1 to 1"



Distributions and dispersions



Teşekkürler