Train In Data

Variable Types

Objectives

- Understand the different types of variables
- Identify different types variables
- Examples of the different variables in a dataset

What is a Variable?

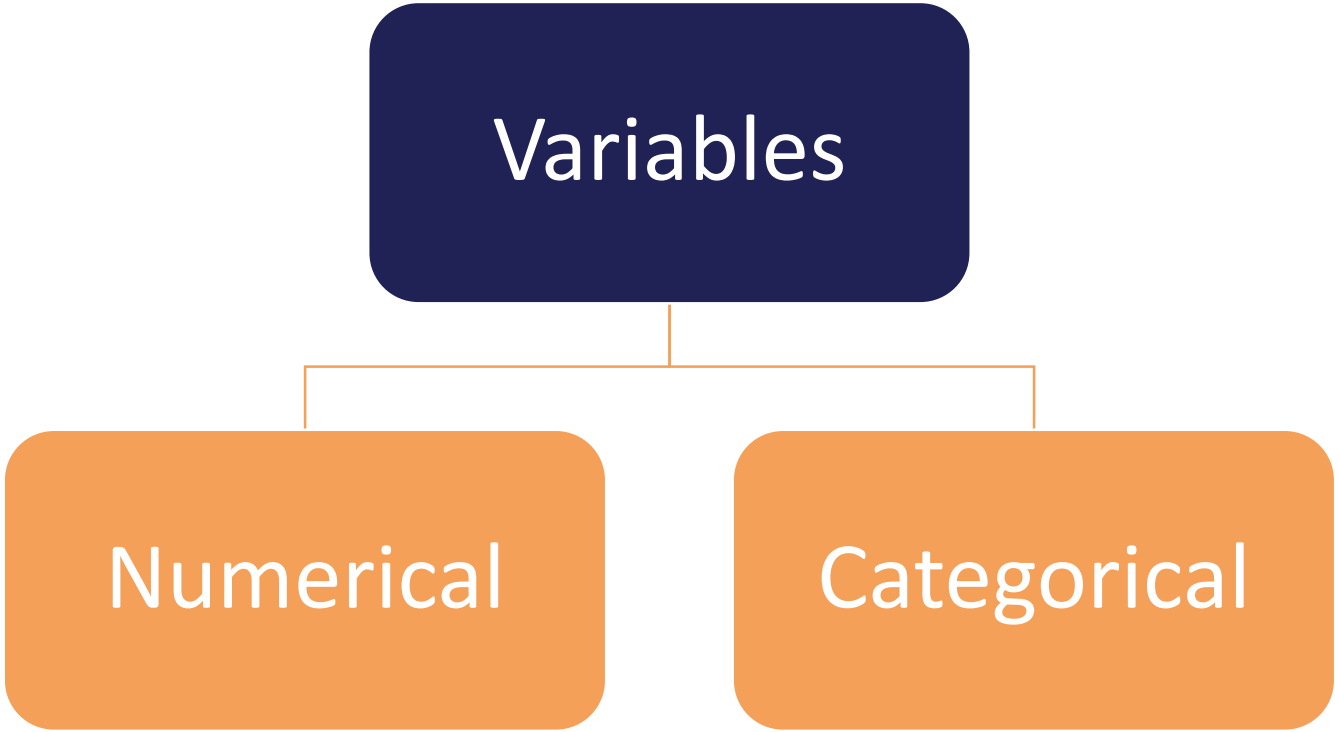
Variable

A variable is any characteristic, number, or quantity that can be measured or counted.

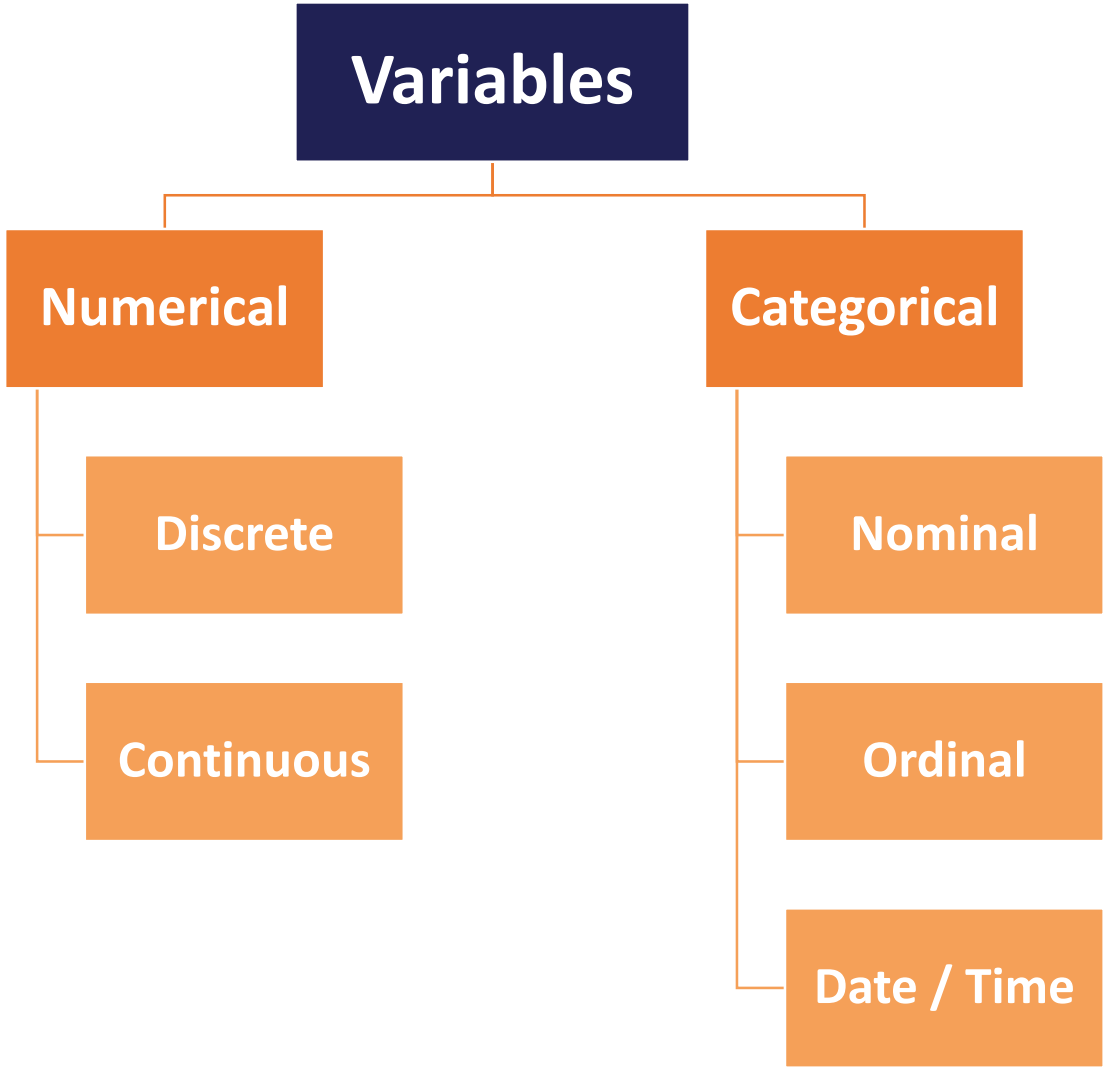
Variable Examples

- Age (21, 35, 62, ...)
- Gender (male, female)
- Income (GBP 20000, GBP 35000, GBP 45000, ...)
- House price (GBP 350000, GBP 570000, ...)
- Country of birth (China, Russia, Costa Rica, ...)
- Eye colour (brown, green, blue, ...)
- Vehicle make (Ford, Volkswagen, ...)

Variable Types



Variable Types

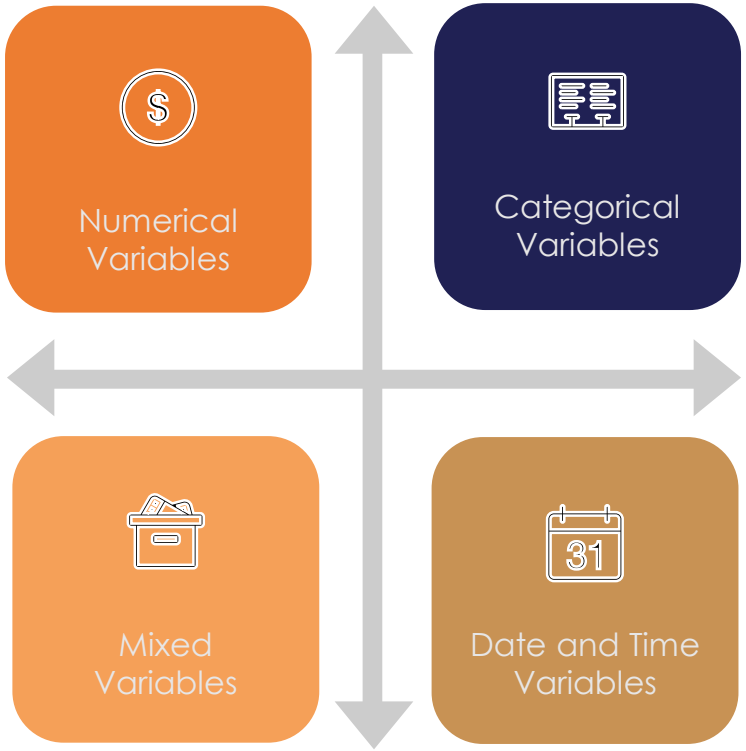


In the next lectures…

Numerical Variables

Discrete and

Continuous



Categorical Variables

Strings

Mixed Variables

With strings and

numbers

Date and Time

Time, date or both

Train In Data

Variable Characteristics

Objectives

Understand the characteristics of the variables which need to be addressed before building machine learning models. Specifically:

- Identify variable characteristics
- Understand how they impact machine learning models
- Examples of variables characteristics in real datasets

Final Summary

- Final article summarizing how the different variable characteristics affect the different machine learning models at the end of the section.

- Additional reading resources.

Variable Characteristics

Understand what are the different things we need to look out for when analyzing the variables in our datasets

Variable Characteristics

LINEAR MODEL ASSUMPTIONS 03
Do the variables fulfill those assumptions?

DISTRIBUTIONS 04
Normal, skewed and others.

CATEGORICAL VARIABLES 02
Variables contain strings instead of numbers

OUTLIERS 05
Unusual or unexpected values

MISSING DATA 01
Lack of information for some observations within a variable.

FEATURE MAGNITUDE 06
Scale of the different features

Train In Data

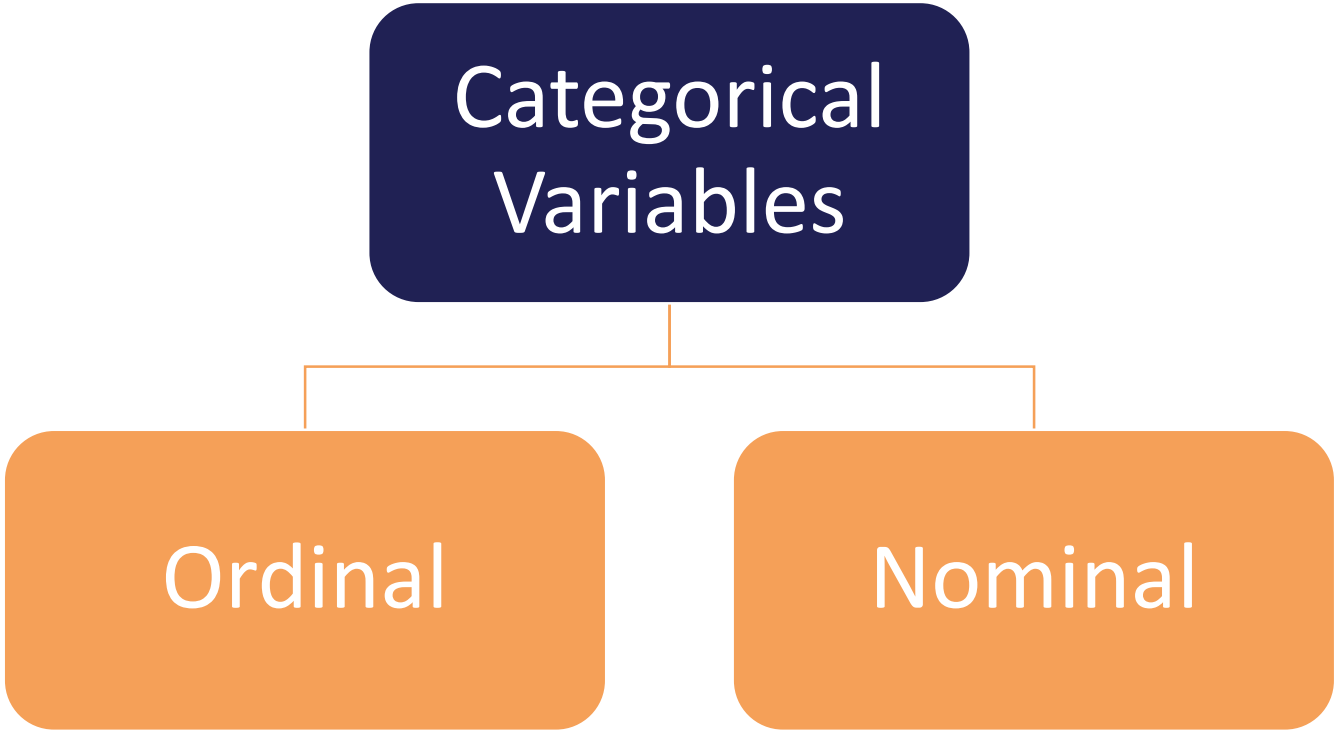
Categorical Variables

Categorical Variables

The values of a categorical variable are selected from a group of **categories**, also called **labels**. Examples:

- Marital status (married, single, …)
- Intended use of loan (debt-consolidation, car purchase, ...)
- Mobile network provider (Vodafone, Orange, ...)
- Gender (male, female)

Variable Types



Ordinal Variables

Categorical variables in which categories can be meaningfully ordered are called ordinal. Examples:

- Student's grade in an exam (A, B, C or Fail)
- Days of the week (Monday = 1 and Sunday = 7)
- Educational level, with the categories: Elementary school, High school, College graduate and PhD ranked from 1 to 4

Nominal Variables

Show no intrinsic order of the labels. Examples:

- Country of birth (Argentina, England, Germany)
- Postcode
- Vehicle make (Citroen, Peugeot, ...)

Special Cases

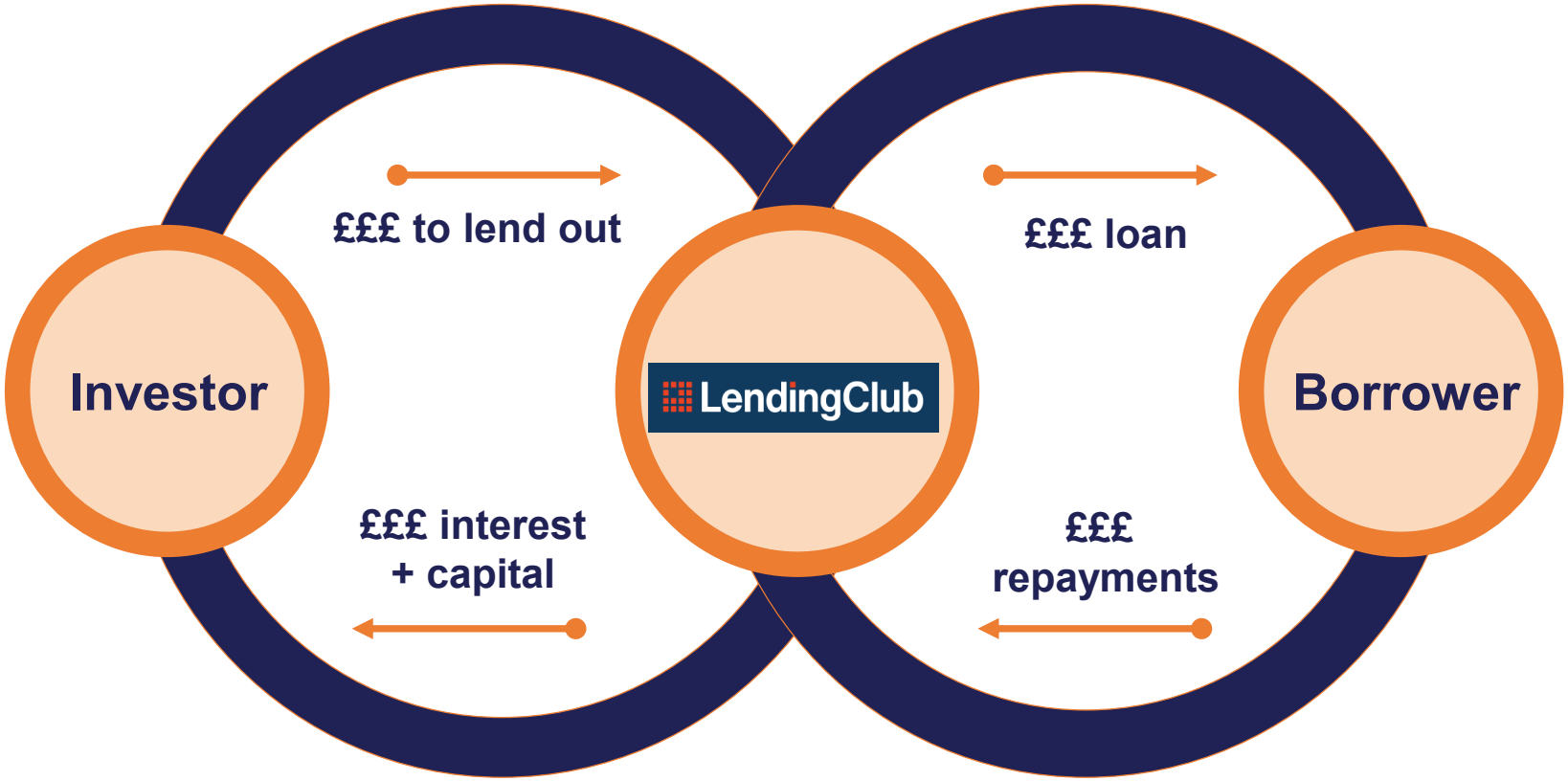
Special cases

- Categorical variables where categories are encoded as numbers (e.g. gender may be coded as 0 for males and 1 for females)
- Id variables: number that uniquely identifies an observation

Categorical Variables

Notebook demo

Peer to Peer Finance

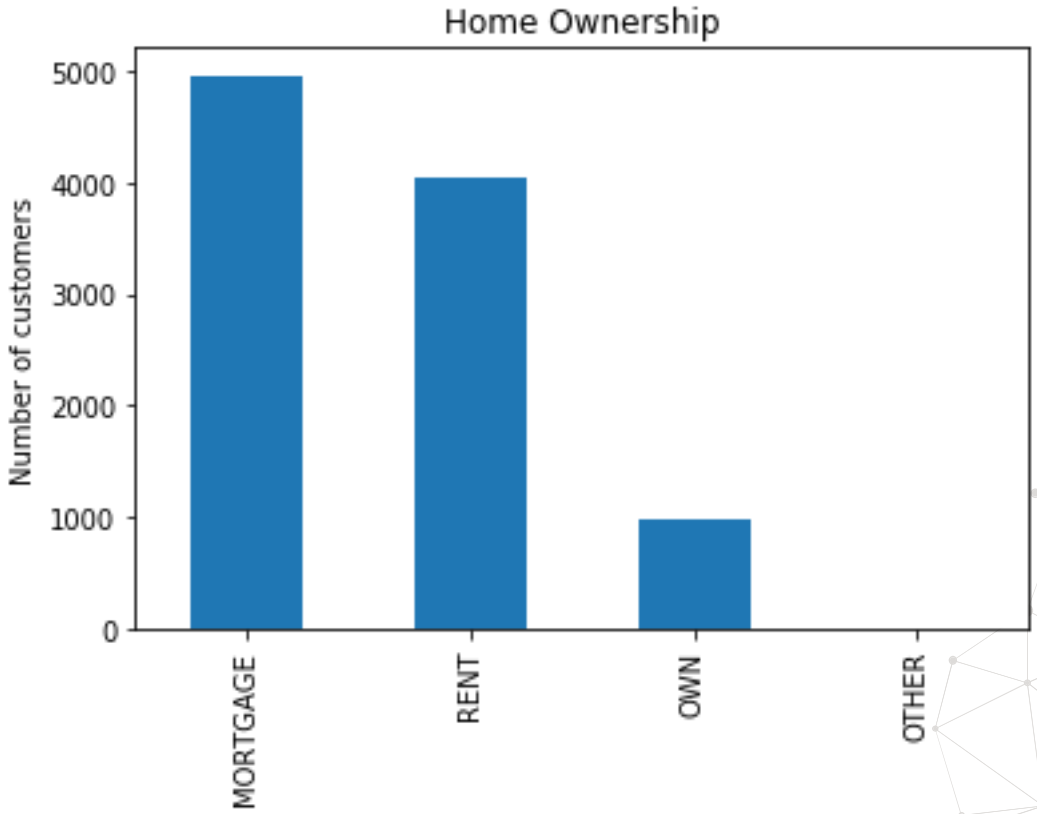


Categorical Variable Examples

- Intended use of loan - Nominal
- Home ownership - Nominal
- Loan status - Nominal

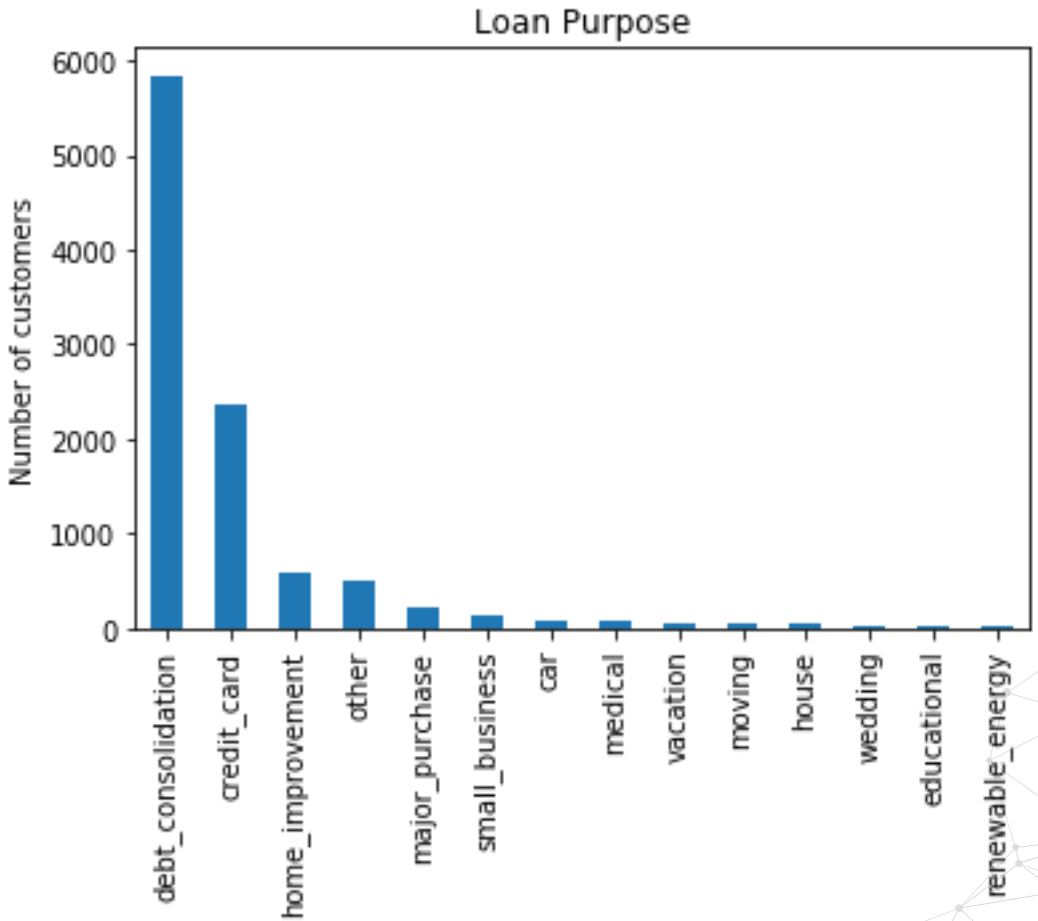
Home Ownership

- Example values: [Mortgage , Rent, Own, Other]



Loan Purpose

- Example values: [Debt consolidation, car, credit car, moving, etc.]



Train In Data

Assumptions of Linear Models

Linear Model Assumptions

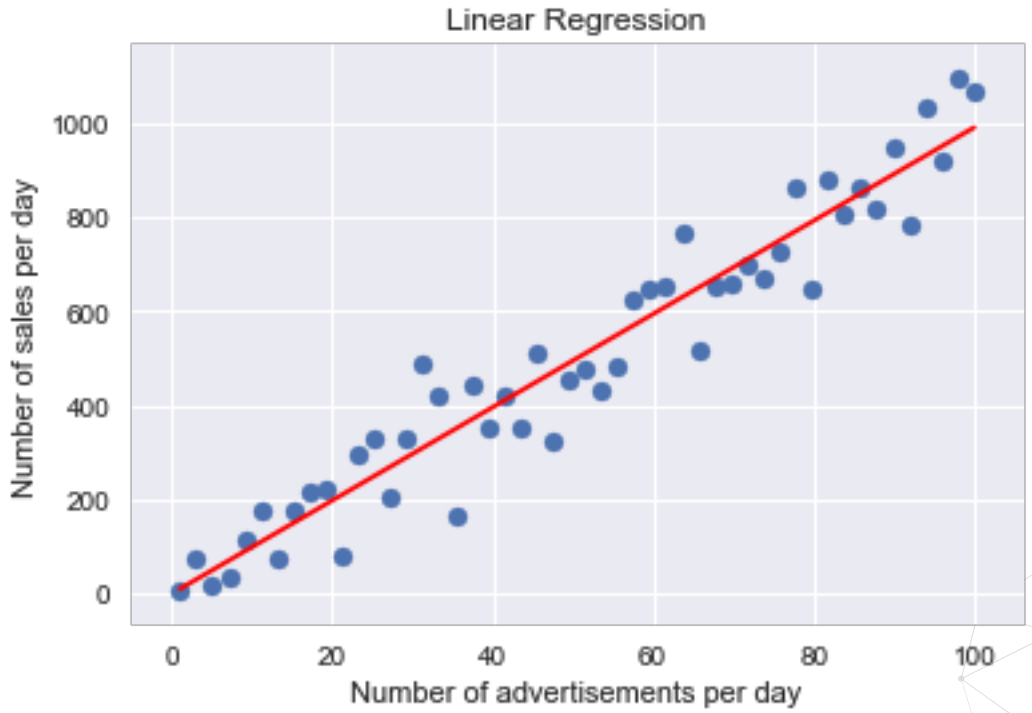
Linear models make the following assumptions about the independent variables (Xs)

- Linear relationship between the variables and the target
- Multivariate normality
- No or little co-linearity
- Homoscedasticity

Linear Relationship

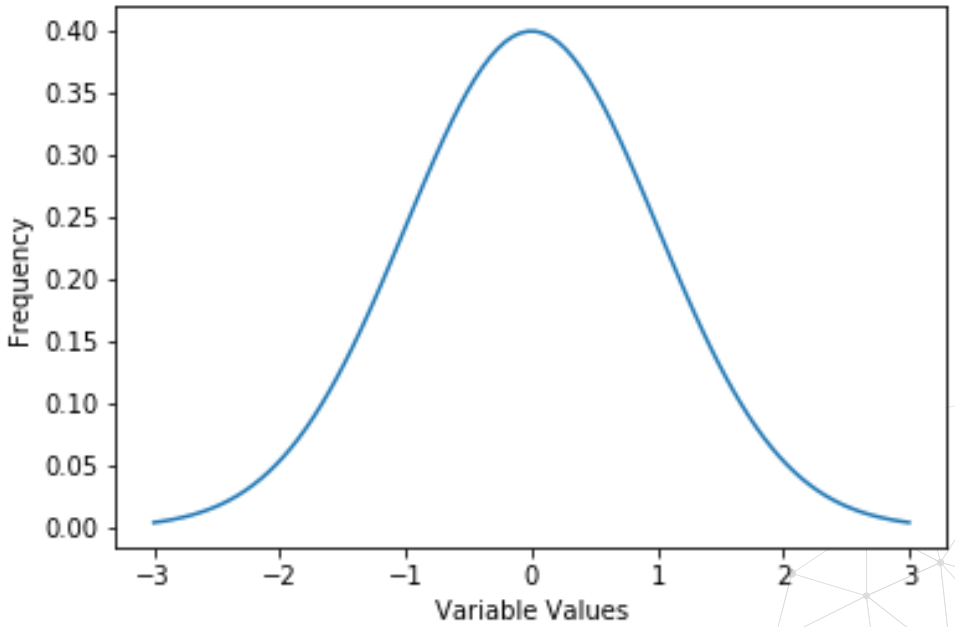
$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn$

- Linear relationship can be assessed with scatter plots
- Sometimes non-linear transformations of the variables and the target improve the linear relationship



Normality

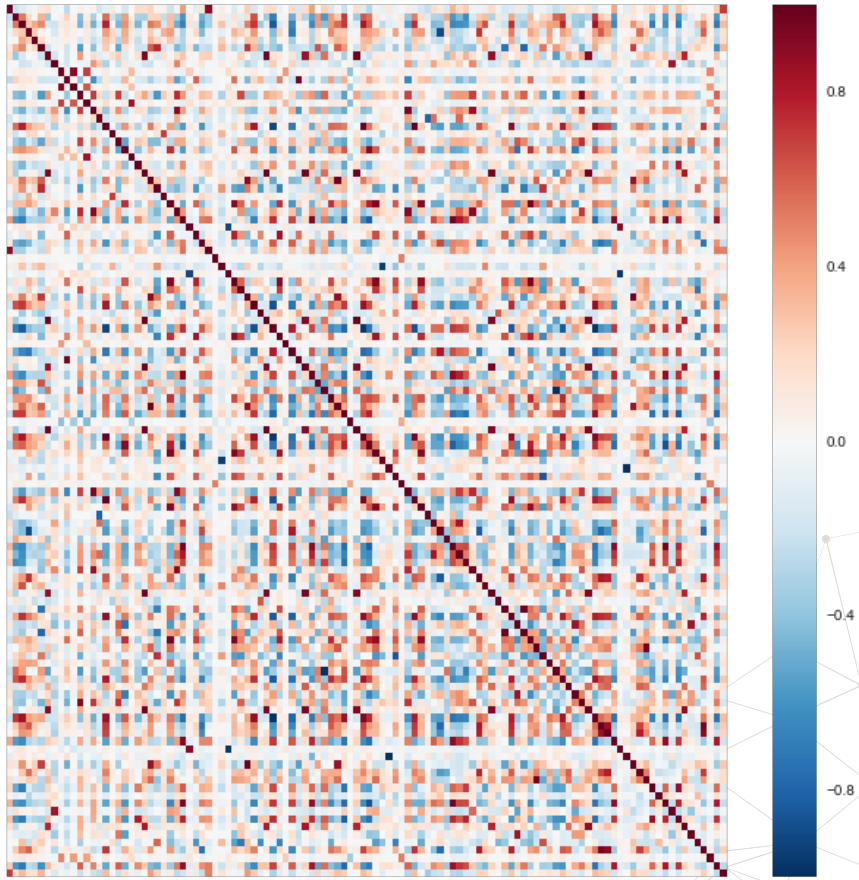
- Variables follow a Gaussian Distribution
- Normality can be assessed with histograms and Q-Q plots
- Normality can be statistically tested, for example with the Kolmogorov-Smirnov test.
- When the variable is not normally distributed a non-linear transformation (e.g., logarithm-transformation) may fix this issue.



No co-linearity

- Multicollinearity occurs when the independent variables are correlated with each other

- Multicollinearity can be assessed with a correlation matrix or the variance inflation factor (VIF)
  - Outside of the scope of this course
  - Check the course Feature Selection for Machine Learning



Homoscedasticity

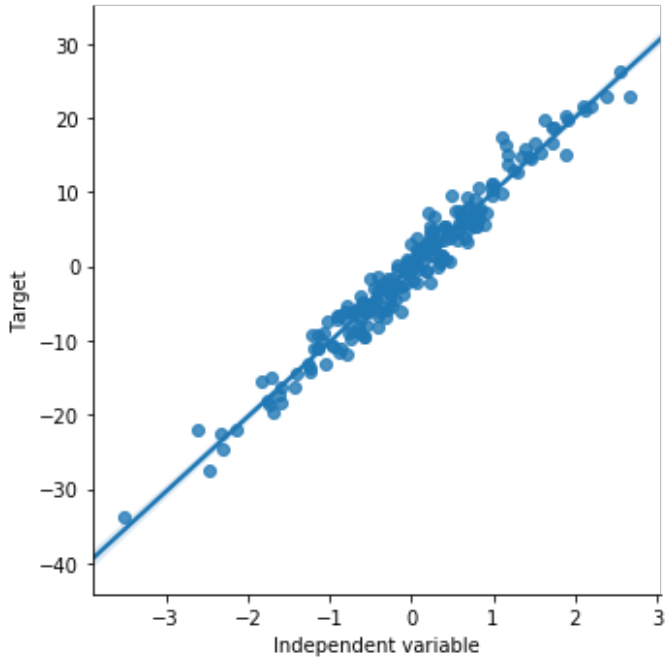
- The independent variables have the same finite variance.
- Also known as homogeneity of variance.
- There are tests and plots to determine homoscedasticity.
  - Residuals plot
  - Levene's test
  - Barlett's test
  - Goldfeld-Quandt Test
- Non-linear transformations and feature scaling can help improve homogeneity of variance

Evaluate model assumptions

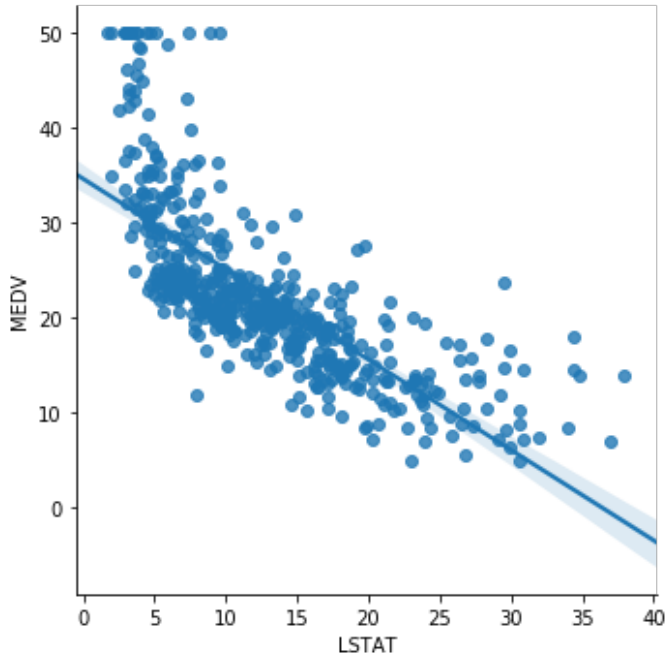
Compare model assumptions in simulated and real data

Linear Relationship – Scatter plots

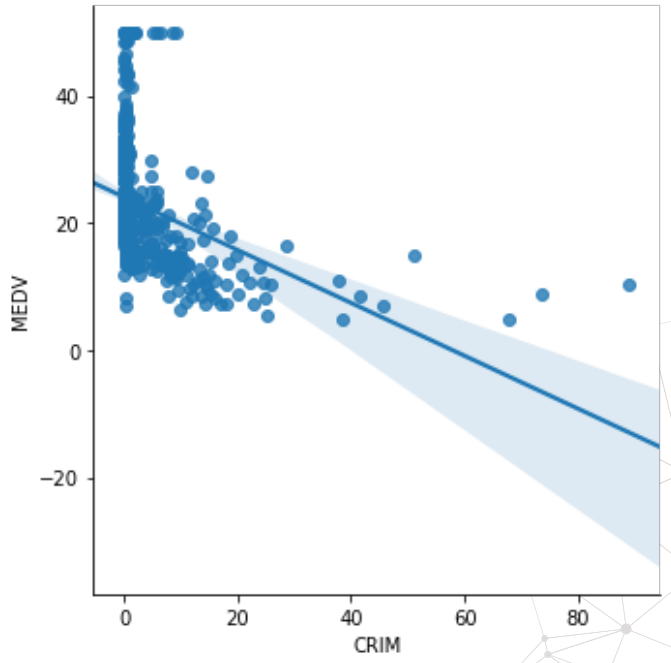
Expected – Simulated data



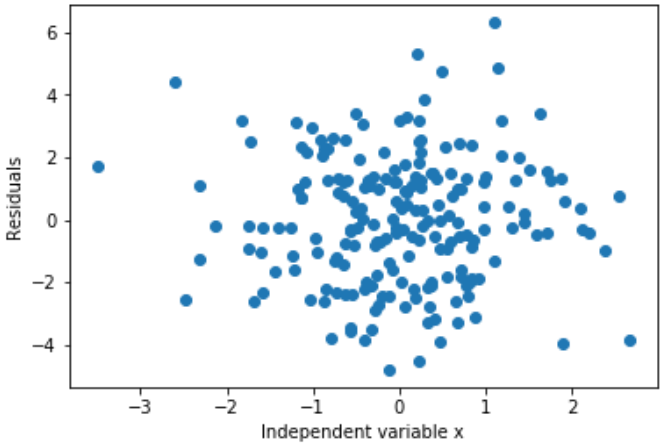
Somewhat linear relationship

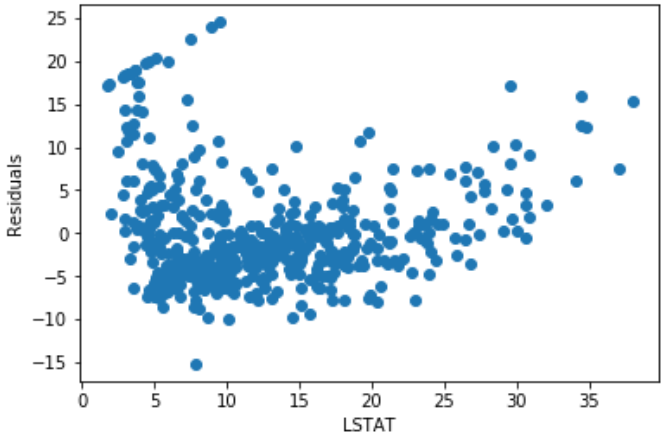


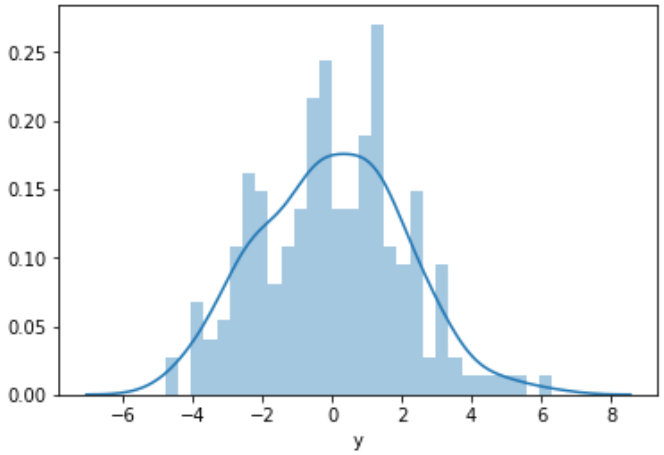
Non-linear relationship

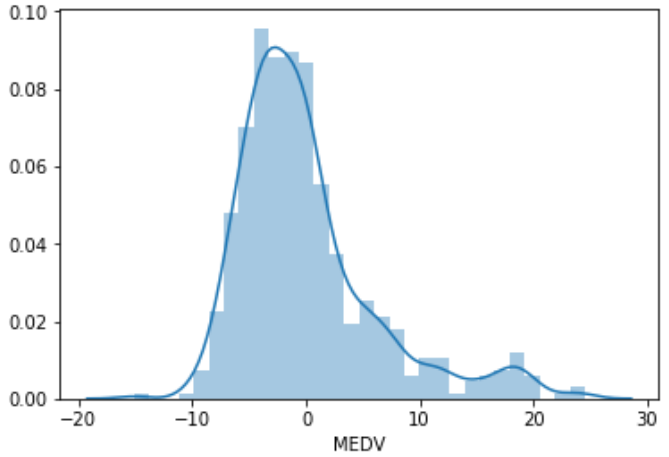


Linear Relationship – Residual plots









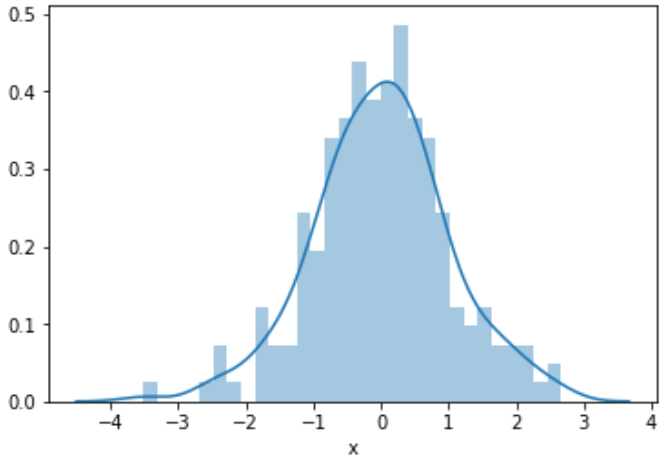
Expected – Simulated data

Somewhat linear relationship

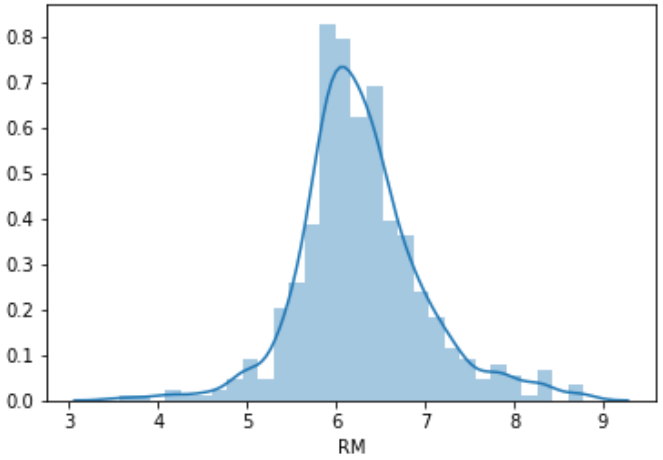
- If relationship between X and y is linear, residuals should be normally distributed and centred around 0
- Residuals are the difference between the predictions and the real value y.

Normality – Histograms

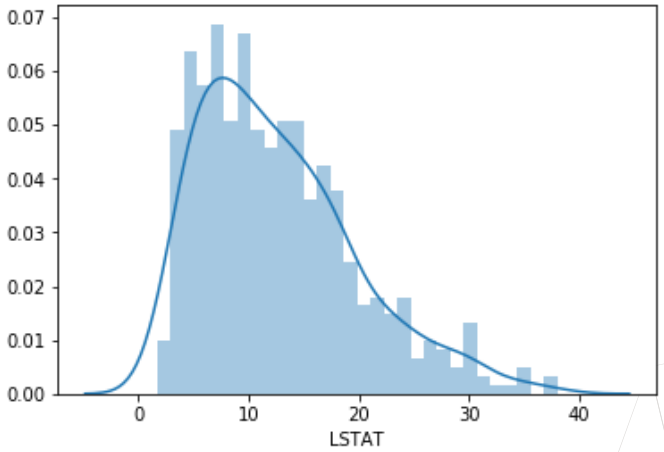
Expected – Simulated data



Somewhat linear relationship (RM)



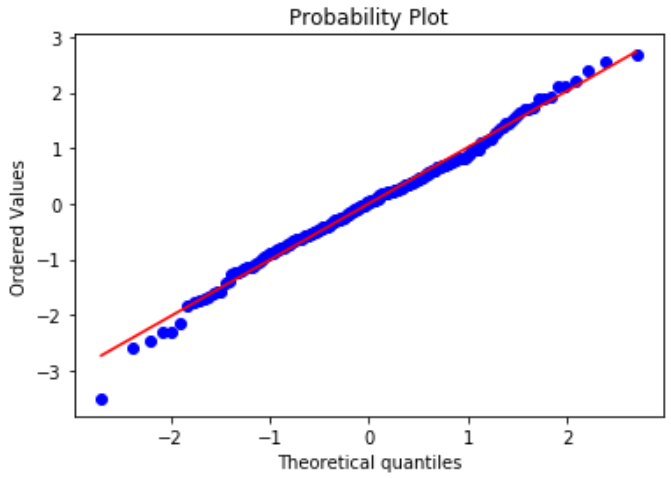
Non-linear relationship (LSTAT)



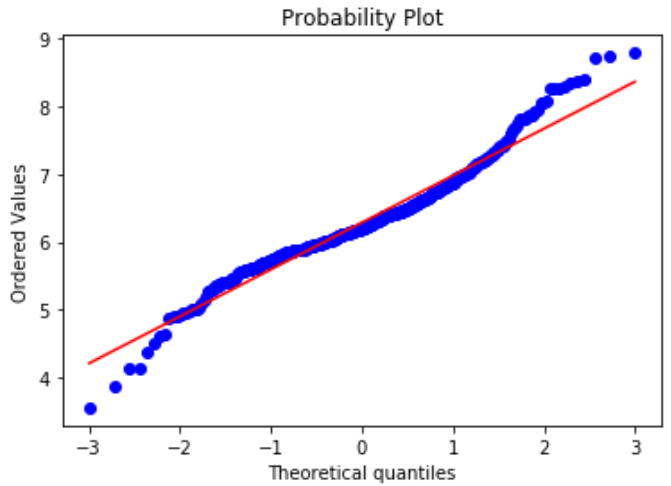
Gaussian distributions adopt a bell shape

Normality – Q-Q plots

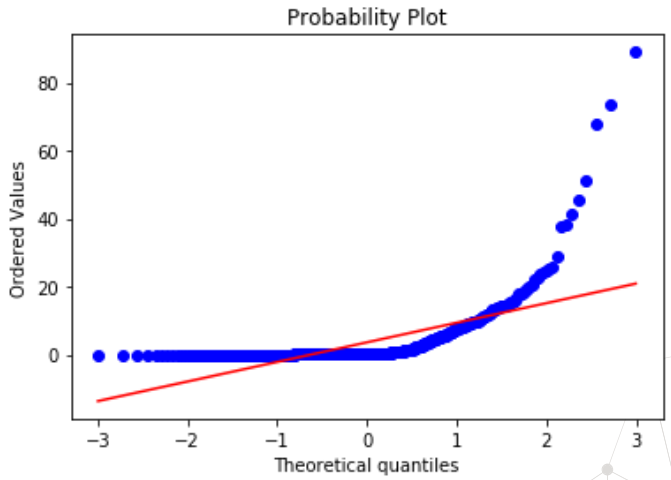
Expected – Simulated data



Somewhat linear relationship



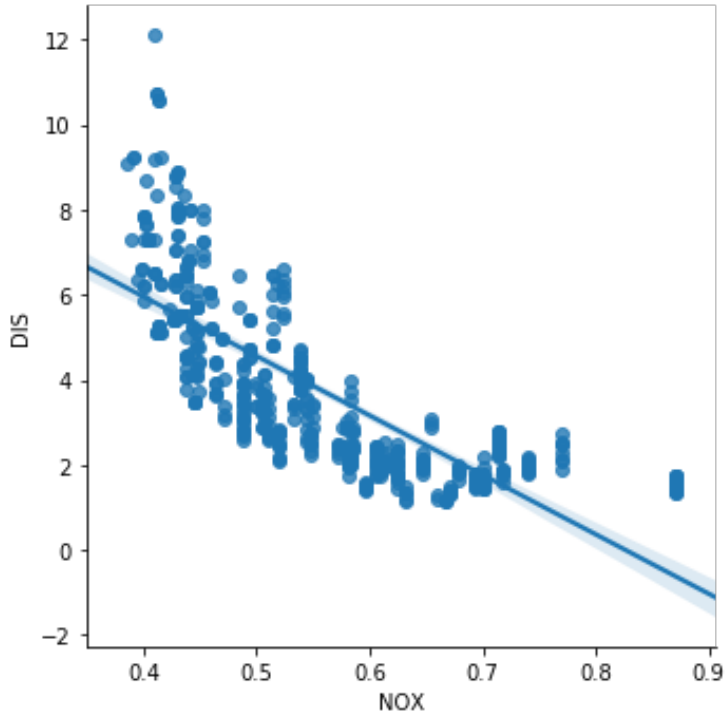
Non-linear relationship

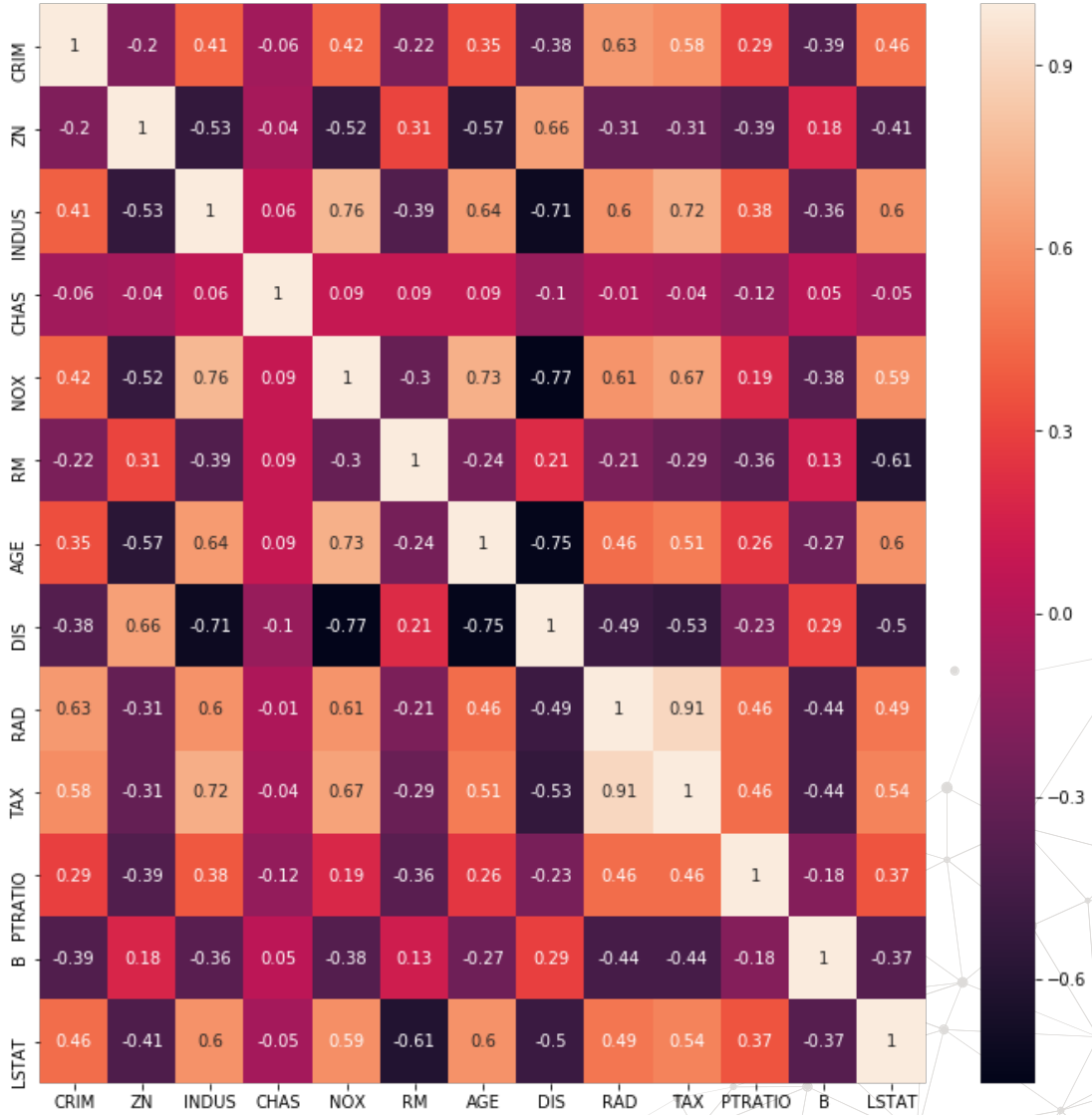


- Q-Q plots plot the variable quantiles in the y-axis and the expected quantiles of the normal distribution on the x-axis.
- If variable is normally distributed, the blue dots should fall on a 45 degree line

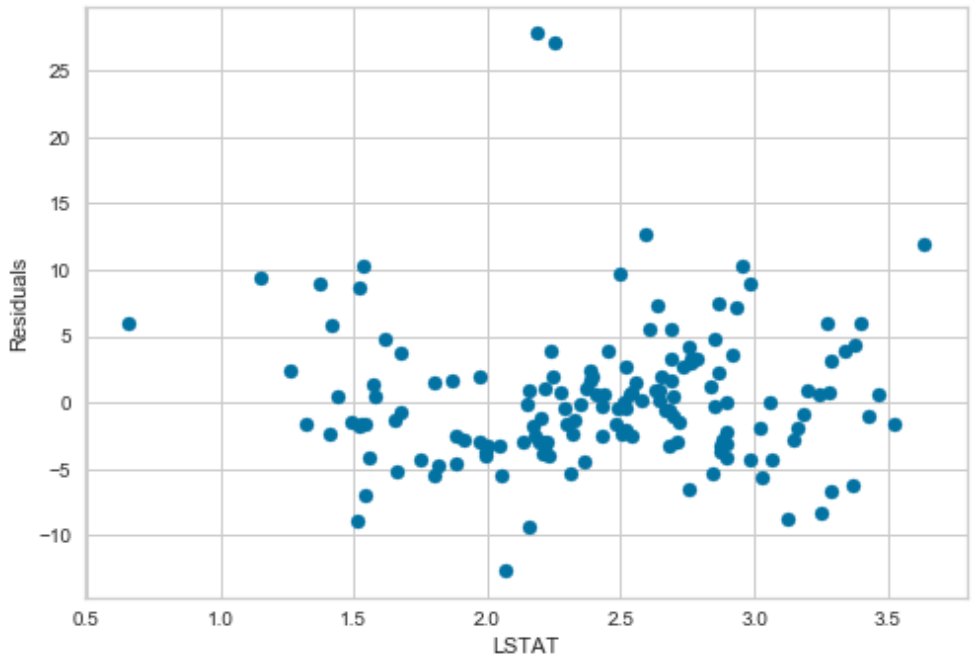
Multi Co-linearity

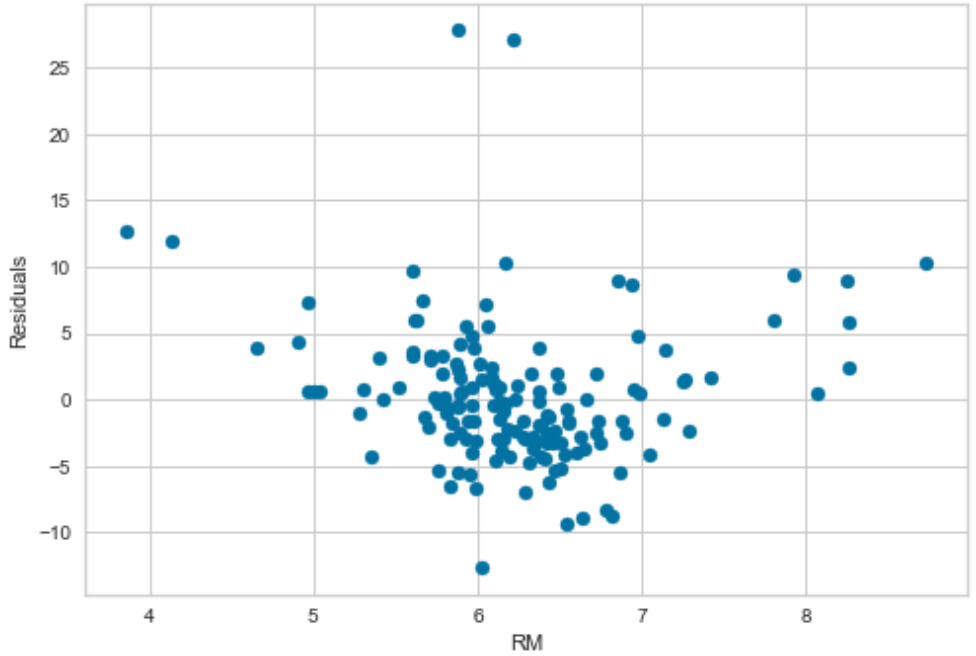
Evaluated by correlation





Homoscedasticity





Homoscedasticity: the error term (that is, the "noise" in the relationship between the independent variables X and the dependent variable Y) is the same across all the independent variables.

To identify homoscedasticity we need to plot the residuals vs each of the independent variables.

The distributions should be similar.

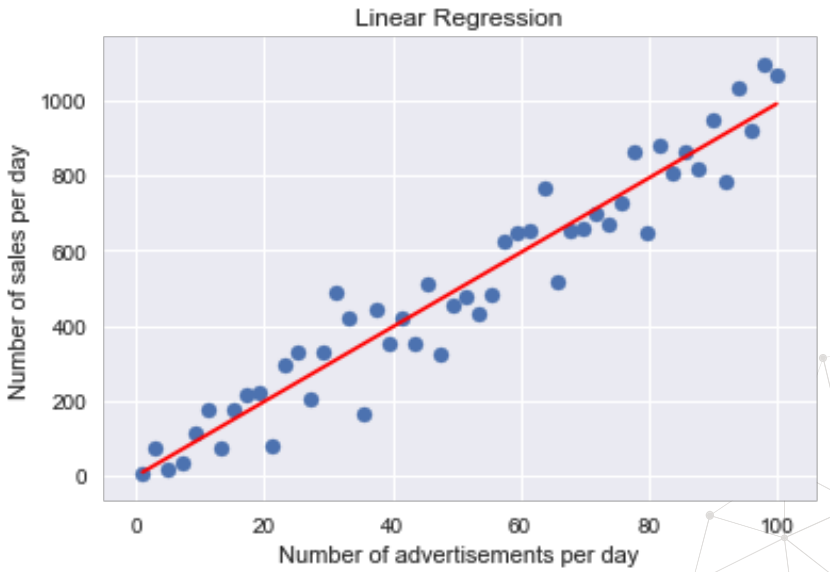
Train In Data

Feature Magnitude

Linear Models

$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn$

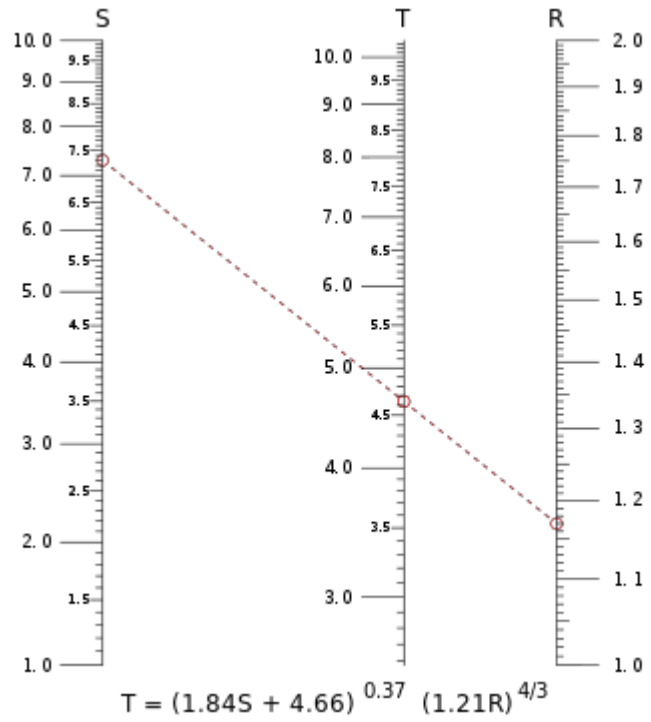
- β indicates the change in Y per unit change of X
- If X changes scale, β will change its value
- Regression coefficients depend of the magnitude of the variable
- Features with bigger magnitudes dominate over features with smaller magnitudes



Feature Magnitude matters

- The regression coefficient is directly influenced by the scale of the variable
- Variables with bigger magnitude / value range dominate over the ones with smaller magnitude / value range
- Gradient descent converges faster when features are on similar scales
- Feature scaling helps decrease the time to find support vectors for SVMs
- Euclidean distances are sensitive to feature magnitude.

Algorithms sensitive to magnitude



The machine learning models affected by the magnitude of the feature:

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-means clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

Machine learning models insensitive to feature magnitude are the ones based on Trees:

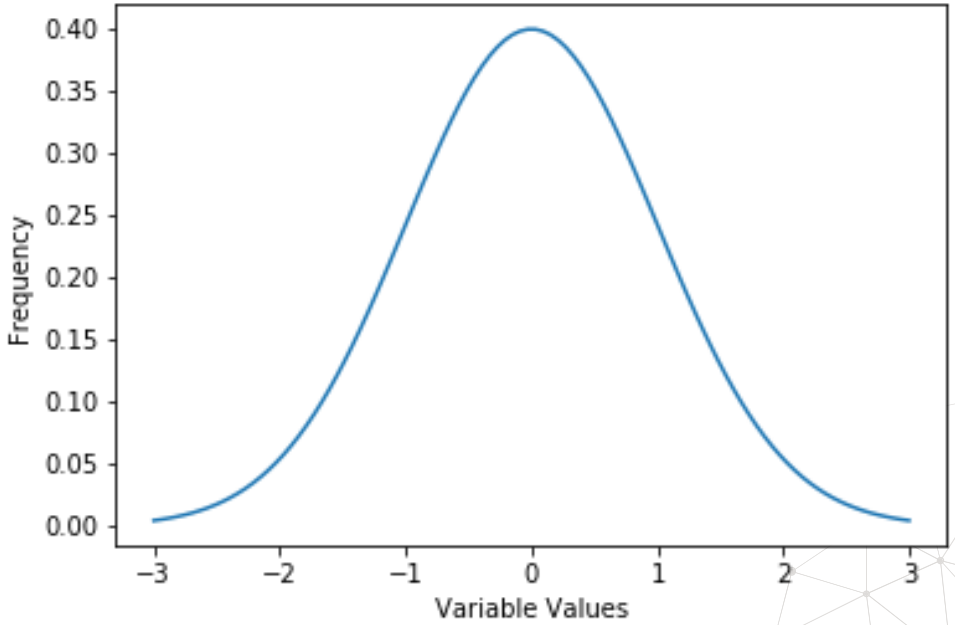
- Classification and Regression Trees
- Random Forests
- Gradient Boosted Trees

Train In Data

Variable Transformation

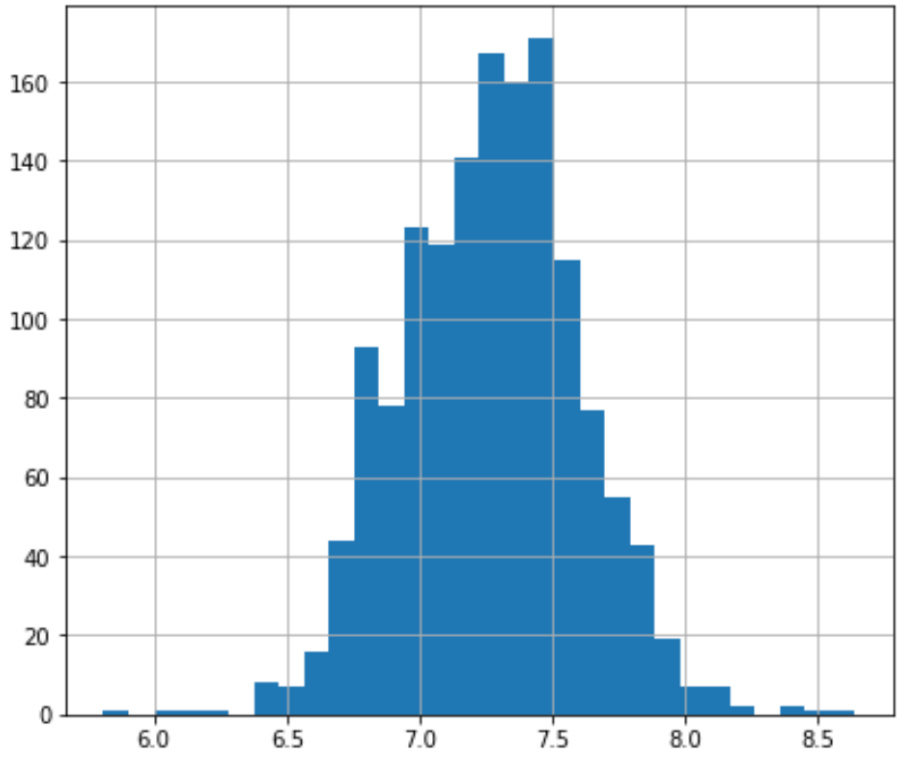
Normality in linear models

- Variables follow a Gaussian Distribution
- Normality can be assessed with histograms and Q-Q plots

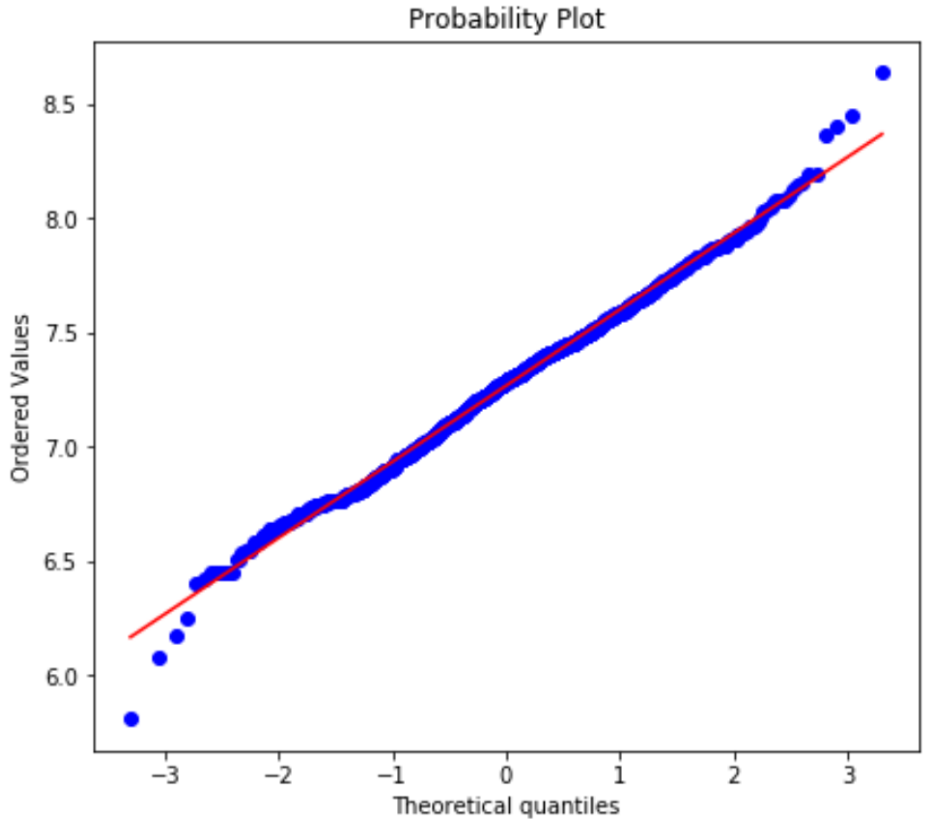


Normality assessment

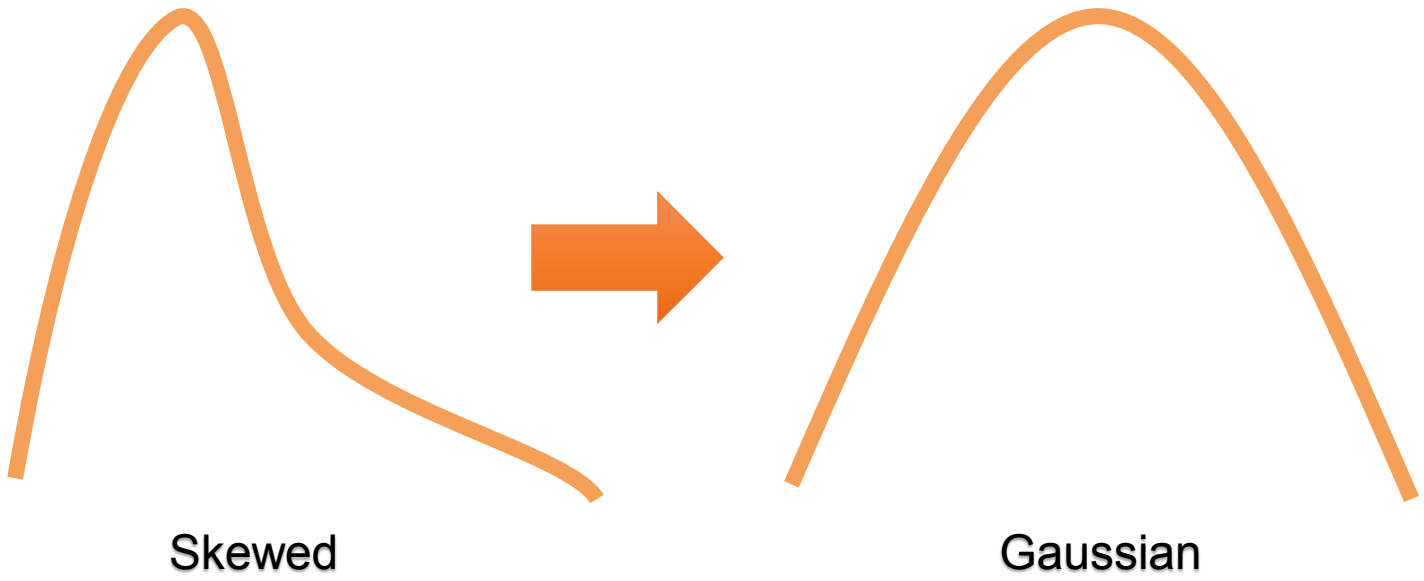
Histogram



Q-Q plot



Mathematical transformations



Variable transformation

- Logarithmic
- Exponential
- Reciprocal
- Box-Cox
- Yeo-Johnson

Mathematical transformations

Logarithmic	Reciprocal	Power / Exponential	Exponential special cases
Log(X)	1 / X	X exp(⋋)	Box-Cox (X>0)
X > 0	X != 0	sqr(X) / cube(X)	Yeo-Johnson
		Not defined for all X	

Box-Cox transformation

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x_i) & \text{if } \lambda = 0, \end{cases}$$

Yeo-Johnson transformation

$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^\lambda - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i) + 1 & \text{if } \lambda = 0, x_i \geq 0 \\ -[(-x_i + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x_i < 0, \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases}$$

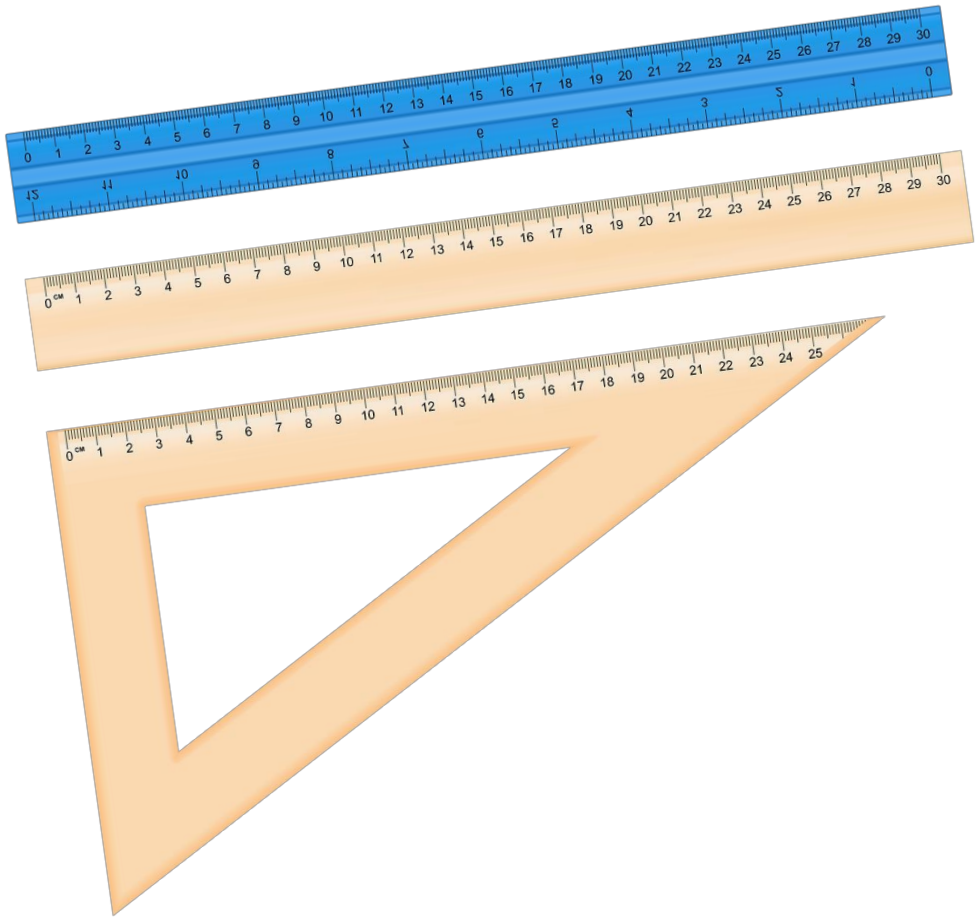
Train In Data

Feature Scaling

Feature Magnitude matters

- The regression coefficient is directly influenced by the scale of the variable
- Variables with bigger magnitude / value range dominate over the ones with smaller magnitude / value range
- Gradient descent converges faster when features are on similar scales
- Feature scaling helps decrease the time to find support vectors for SVMs
- Euclidean distances are sensitive to feature magnitude.

Algorithms sensitive to magnitude



The machine learning models affected by the magnitude of the feature:

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-means clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

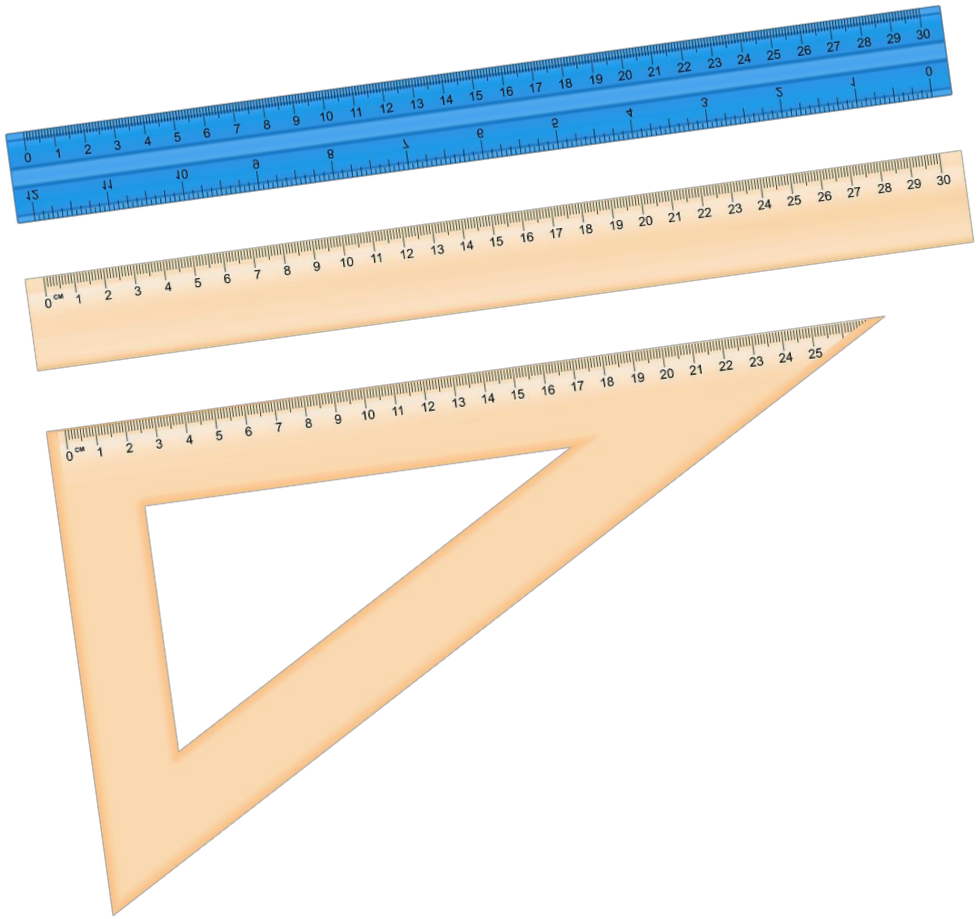
Machine learning models insensitive to feature magnitude are the ones based on Trees:

- Classification and Regression Trees
- Random Forests
- Gradient Boosted Trees

Feature Scaling

- Feature scaling refers to the methods used to normalize the range of values of independent variables.
- In other words, the methods to set the feature value range within a similar scale.
- Feature scaling is generally the last step in the data pre-processing pipeline, performed just before training the machine learning algorithms.

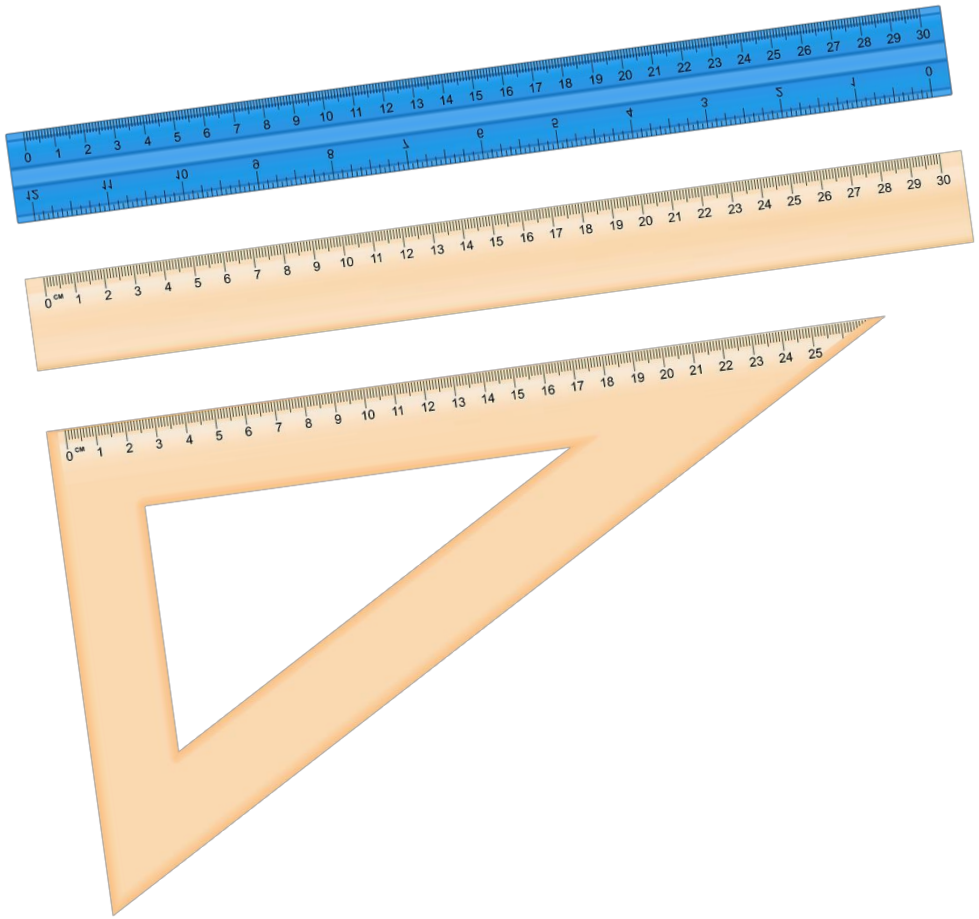
Feature scaling methods



Scaling methods

- Standardisation
- Mean normalisation
- Scaling to maximum and minimum
- Scaling to absolute maximum
- Scaling to median and quantiles
- Scaling to unit norm

Feature scaling methods



Scaling methods

- **Standardisation**
- Mean normalisation
- **Scaling to maximum and minimum**
- Scaling to absolute maximum
- Scaling to median and quantiles
- Scaling to unit norm

Train In Data

Standardisation

Standardisation

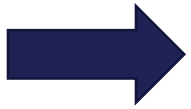
- Centres the variable at zero and sets the variance to 1.

$$\text{Z-score} = \frac{X - mean(X)}{Std(X)}$$

Standardisation: example

Price
100
90
50
40
20
100
50
60
120
40
200

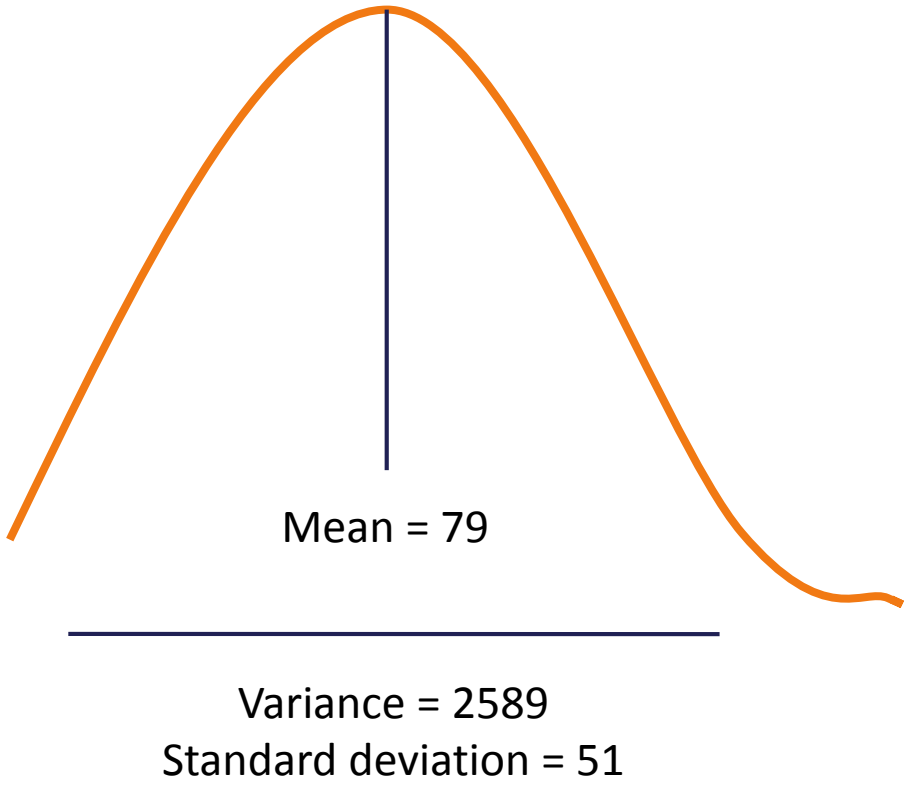
Mean = 79
Standard dev = 51



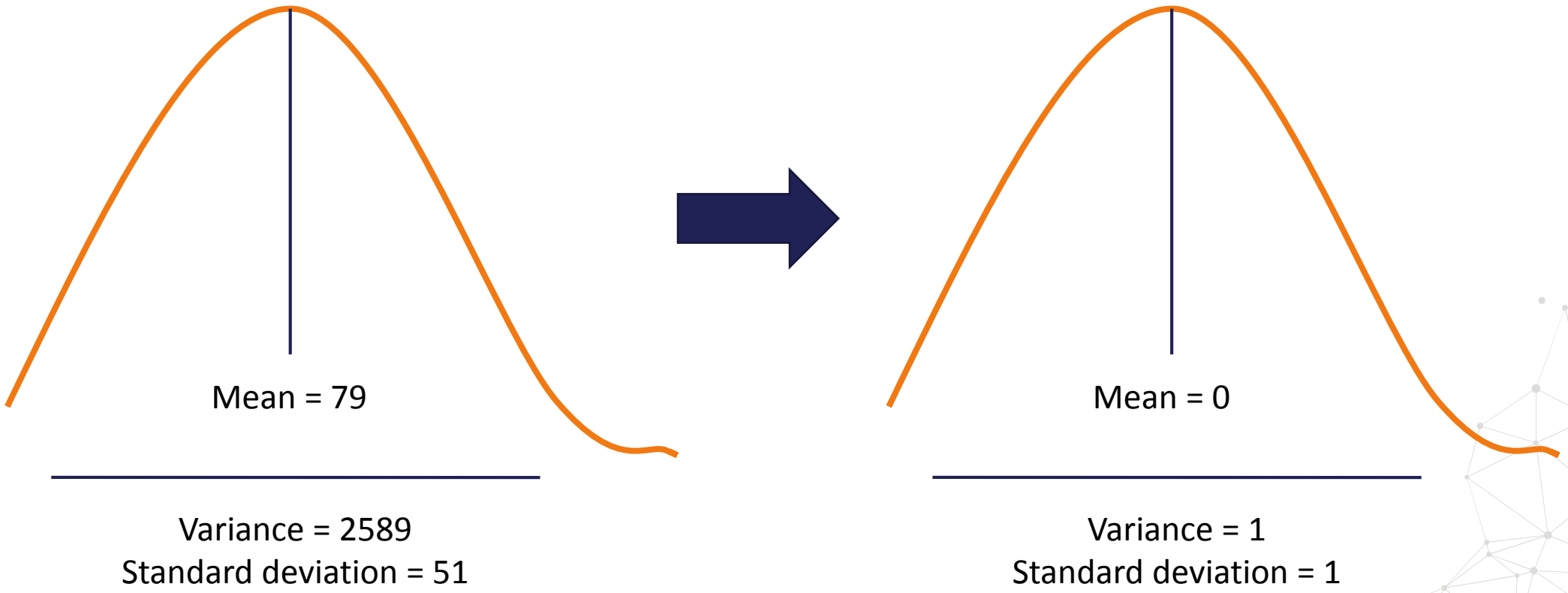
$$\frac{\text{Obs. - Mean}}{\text{Standard dev}}$$

Price
0.41
0.22
-0.57
-0.76
-1.16
0.41
-0.57
-0.37
0.80
-0.76
2.37

Standardisation: effect

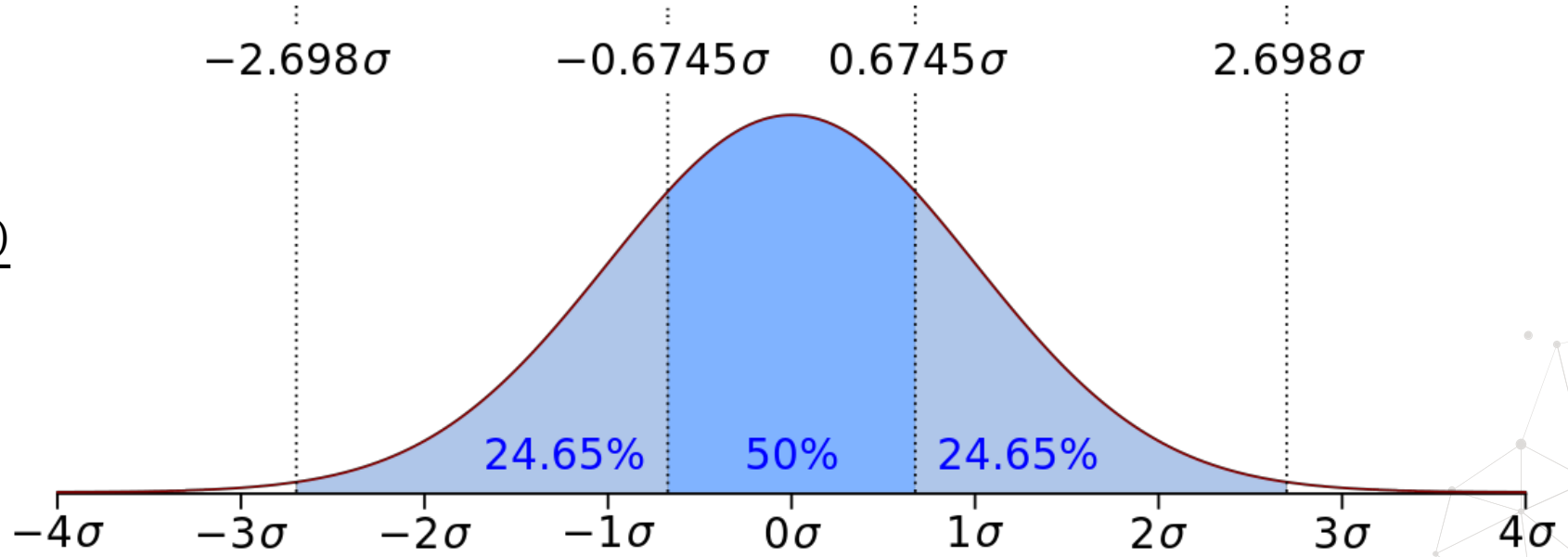


Standardisation: effect



Standardised variable: meaning

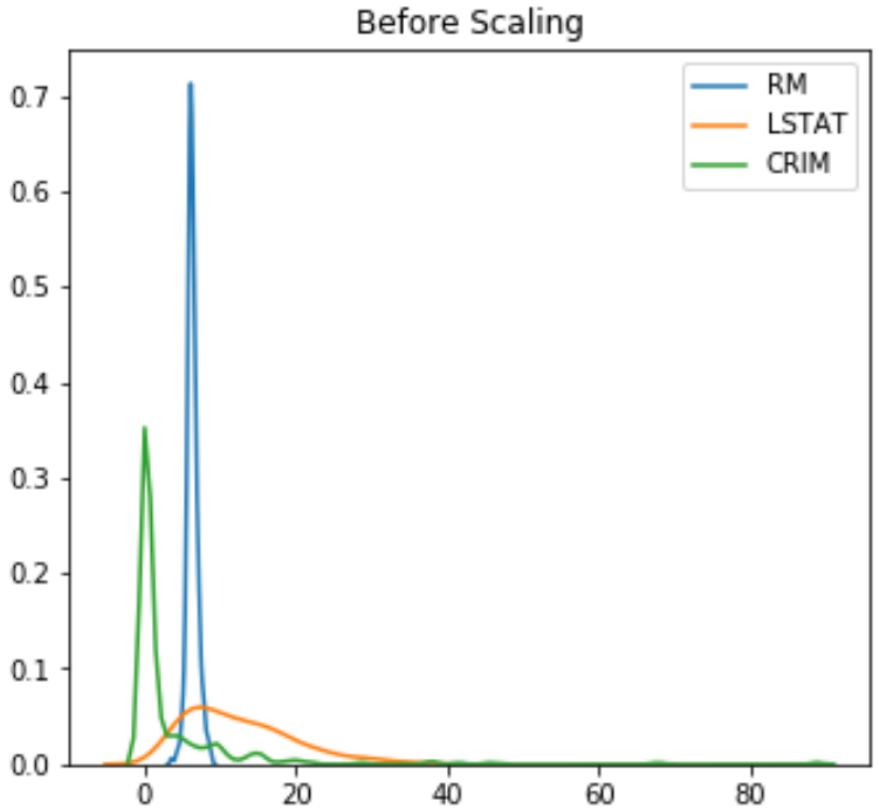
$$\text{Z-score} = \frac{X - mean(X)}{Std(X)}$$

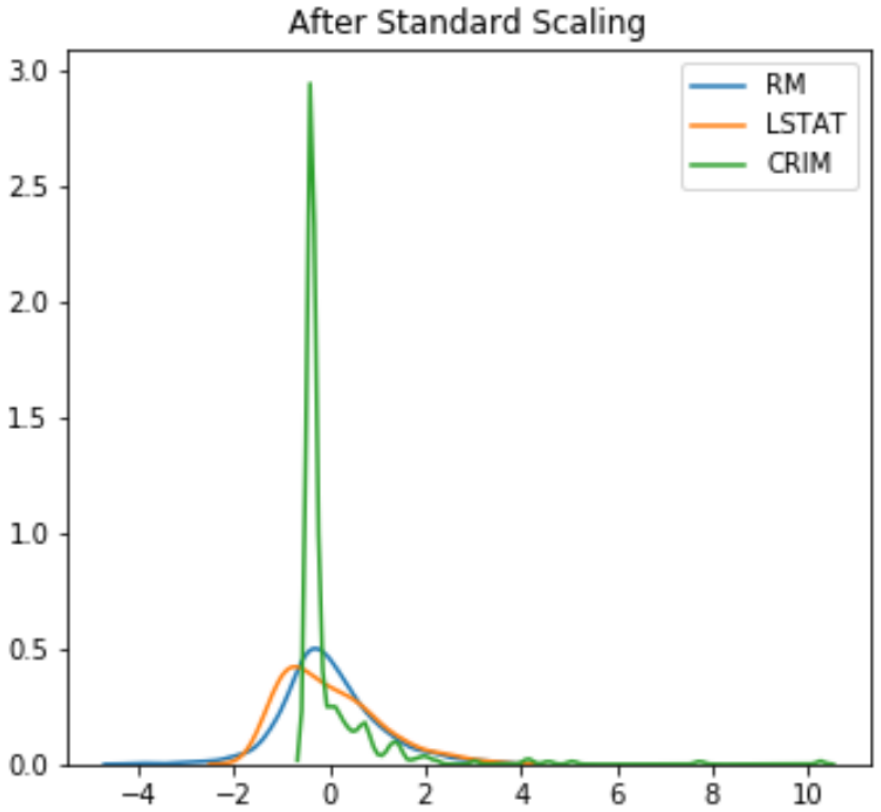


Standardisation: summary

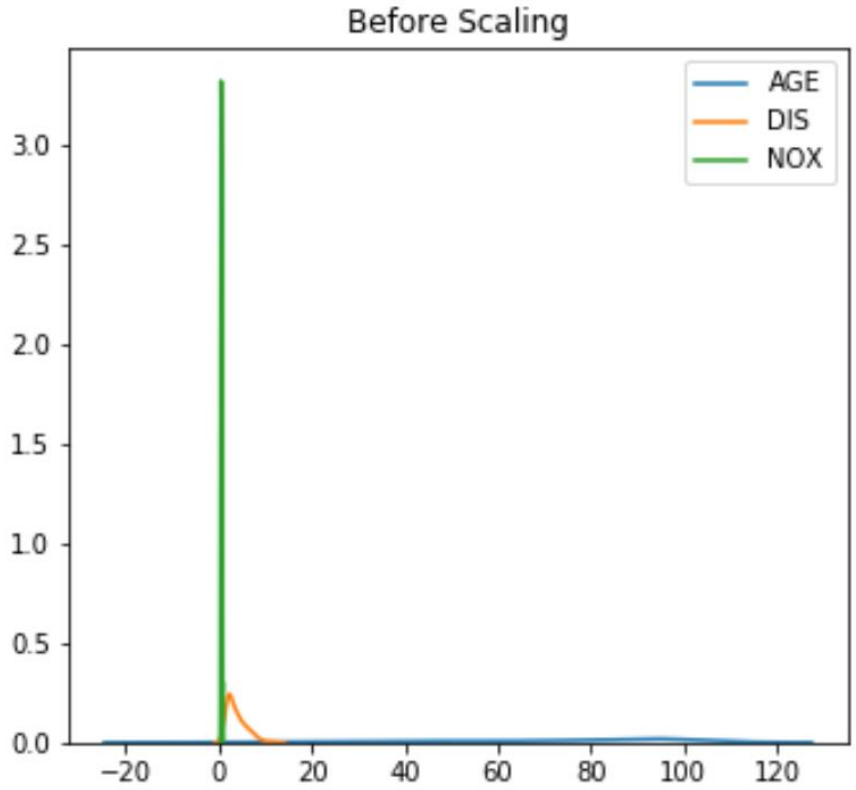
- Centres the mean at 0
- Scales the variance at 1
- Preserves the shape of the original distribution
- Minimum and maximum values vary
- Preserves outliers

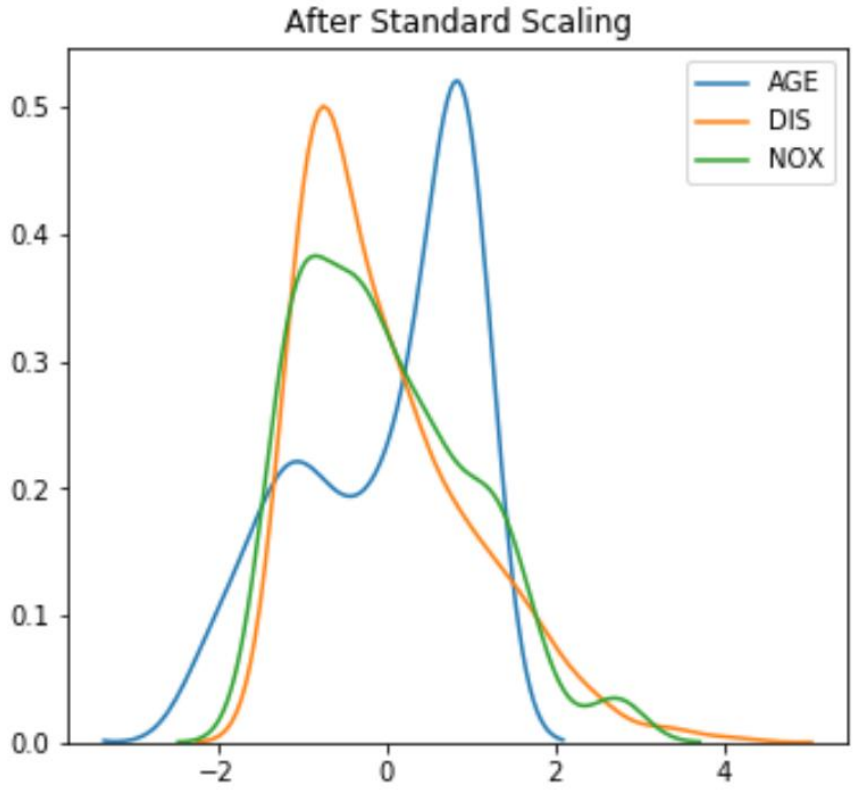
Standardisation: Notebook examples





Standardisation: Notebook examples





Train In Data

Mean Normalisation

Mean Normalisation

- Centres the variable at 0 and re-scales the variable to the value range.

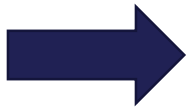
$$X\_scaled = \frac{X - mean(X)}{max(X) - \min(X)}$$

Mean normalisation: example

Price
100
90
50
40
20
100
50
60
120
40
200

Mean = 79

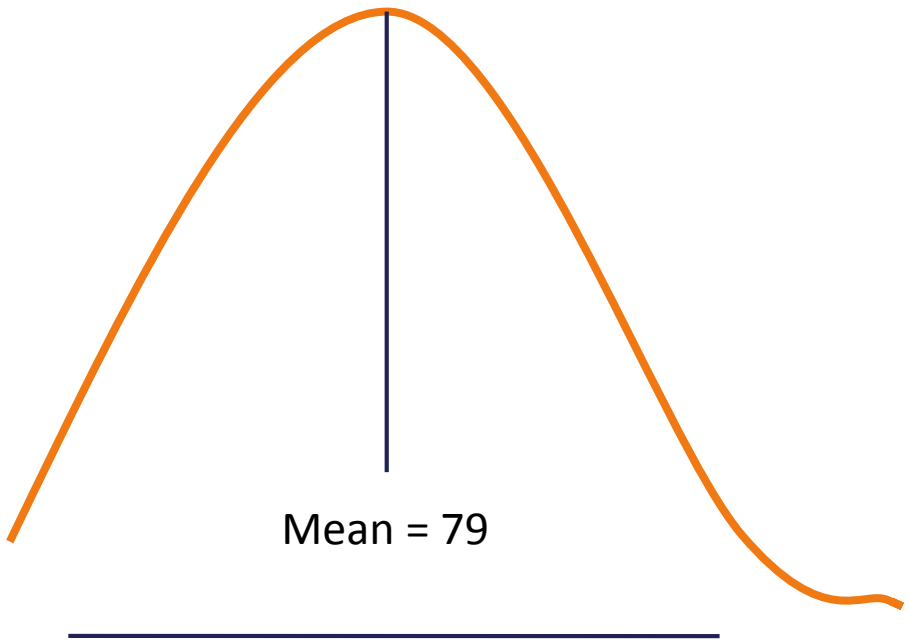
Max = 200
Min = 20
Range = 200 – 20 = 180



$$\frac{\text{Obs. - Mean}}{\text{Range}}$$

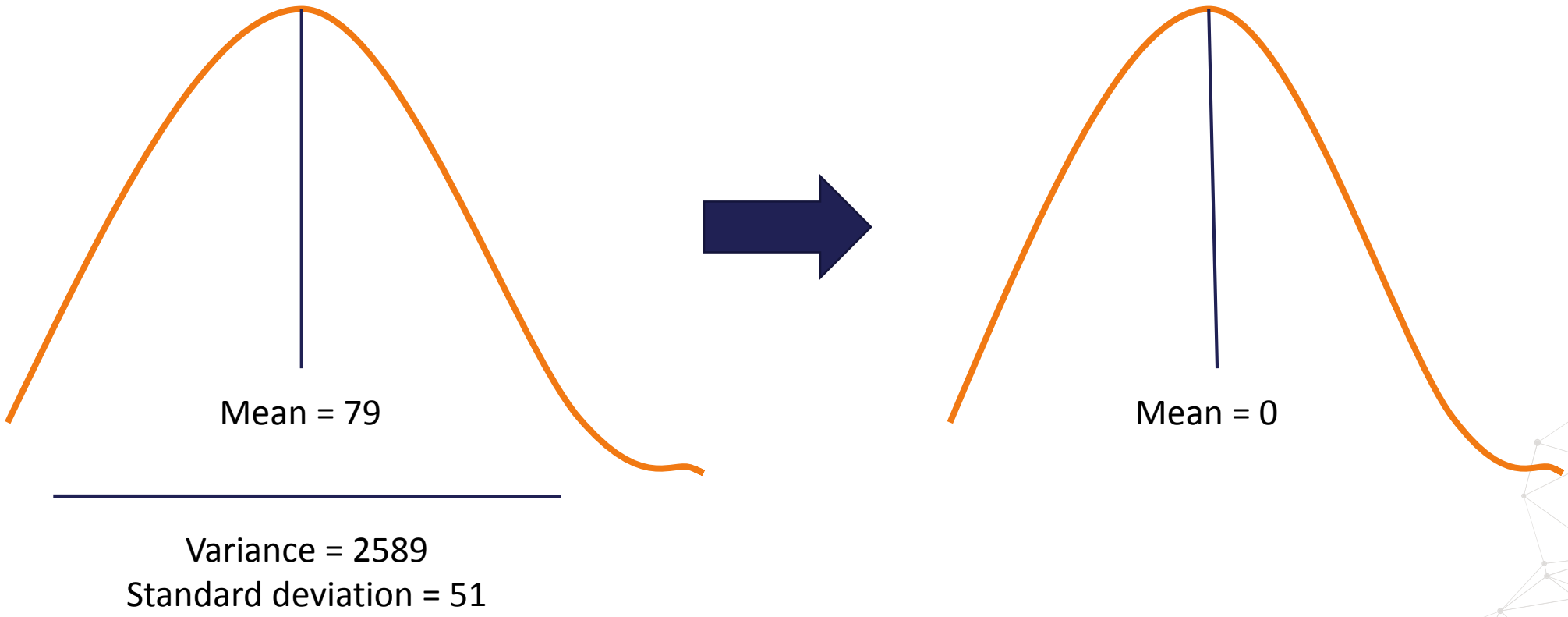
Price
0.12
0.06
-0.16
-0.22
-0.33
0.12
-0.16
-0.11
0.23
-0.22
0.67

Mean normalisation: effect



Variance = 2589
Standard deviation = 51

Mean normalisation: effect

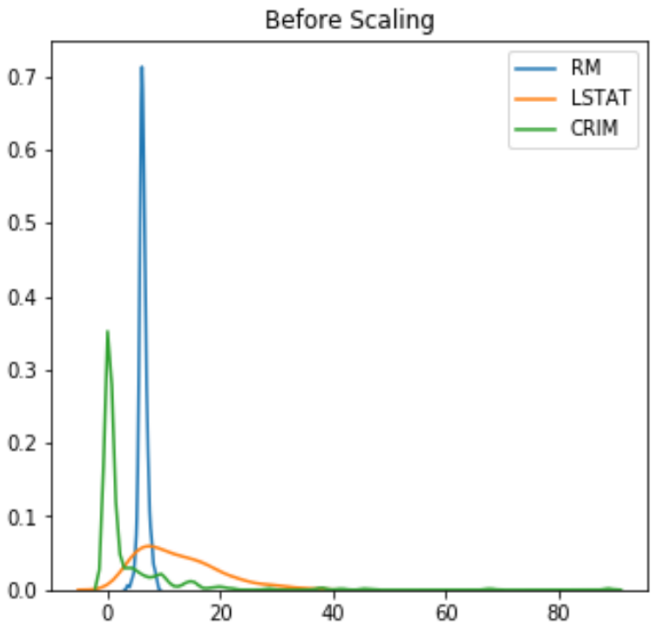


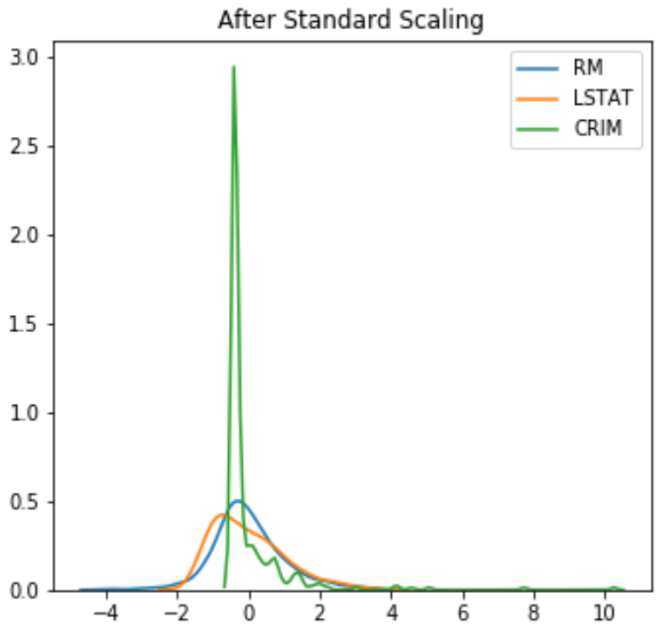
Variance = 2589
Standard deviation = 51

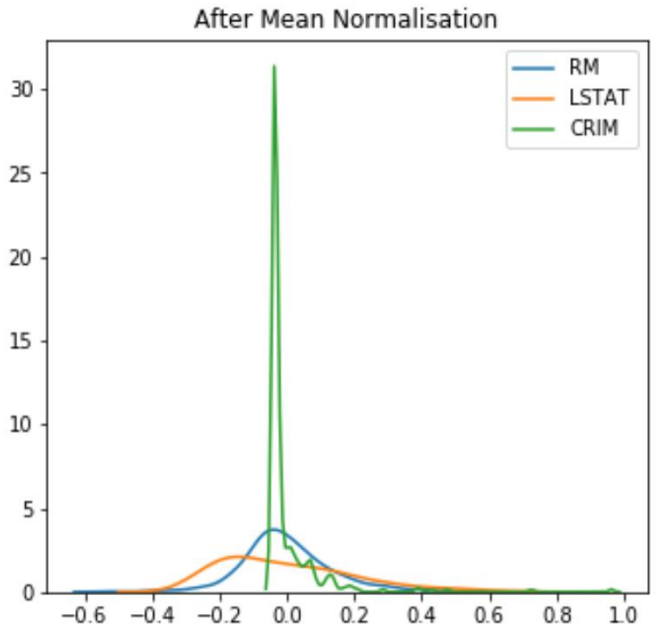
Mean normalisation: summary

- Centres the mean at 0
- Variance varies
- May alter shape of the original distribution
- Minimum and maximum values within [-1;1]
- Preserves outliers

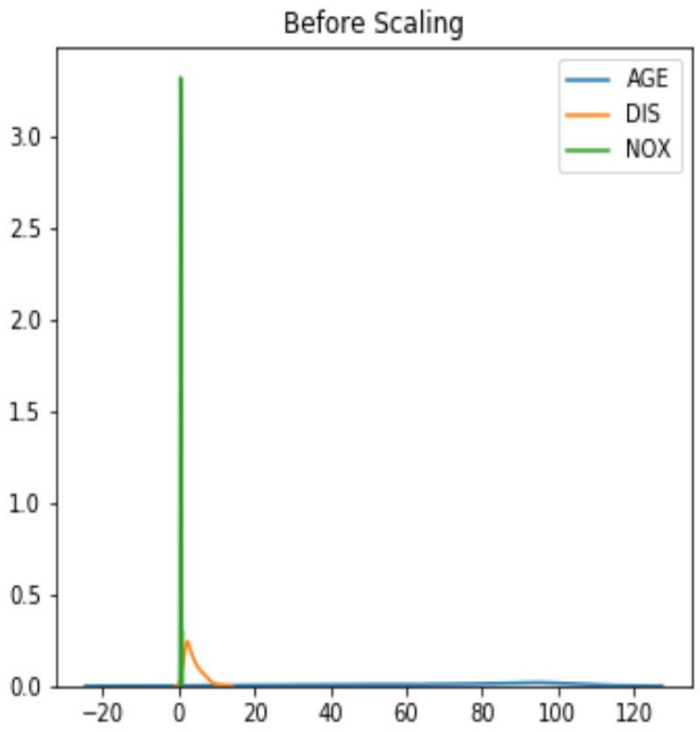
Mean normalisation: Notebook

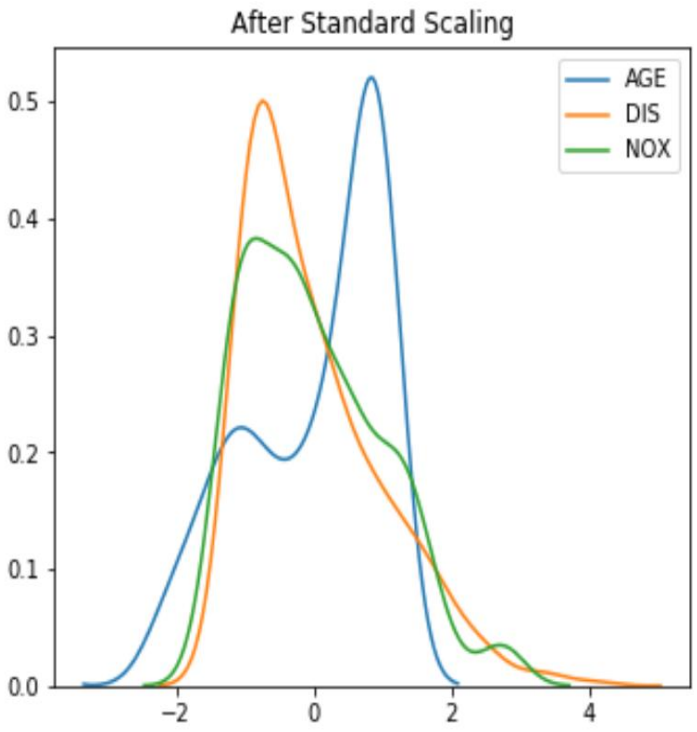


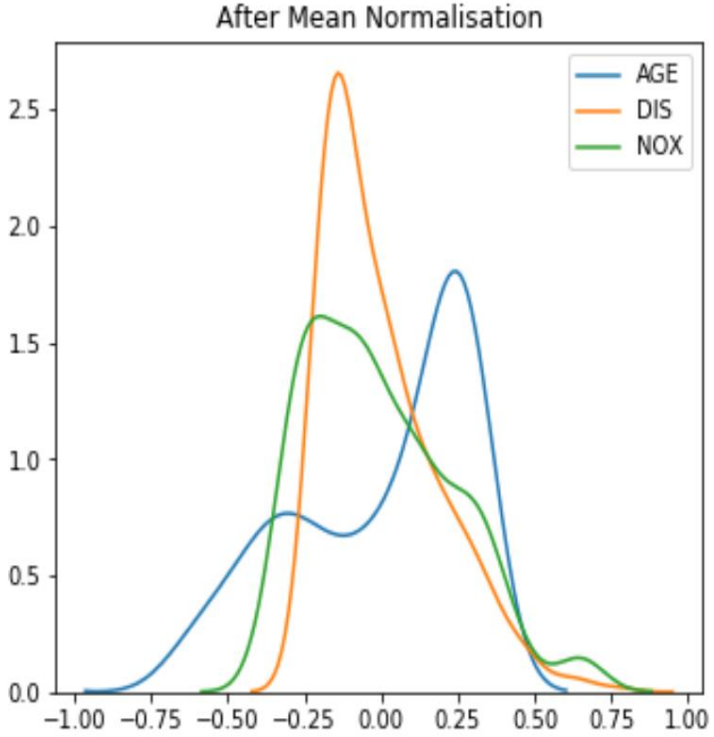




Mean normalisation: Notebook







Train In Data

Min-Max Scaling

MinMaxScaling

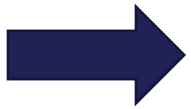
- Scales the variable between 0 and 1

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

MinMaxScaling: example

Price
100
90
50
40
20
100
50
60
120
40
200

Max = 200
Min = 20
Range = 200 – 20 = 180



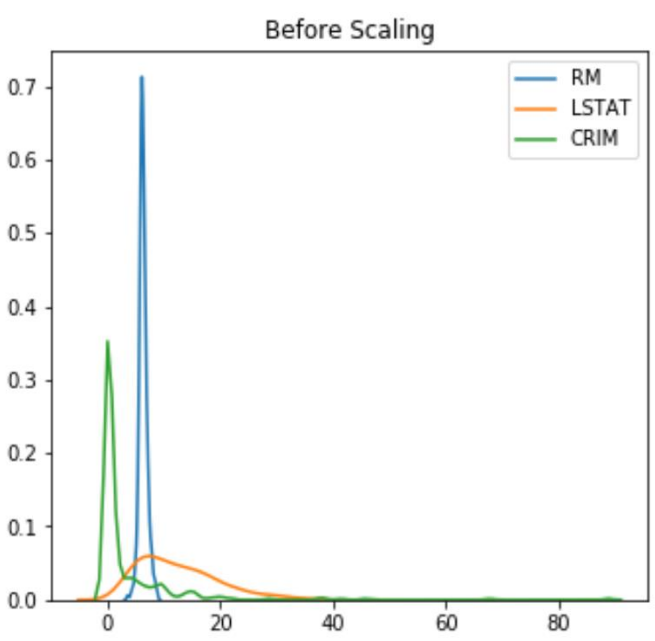
$$\frac{\text{Obs. - Min}}{\text{Range}}$$

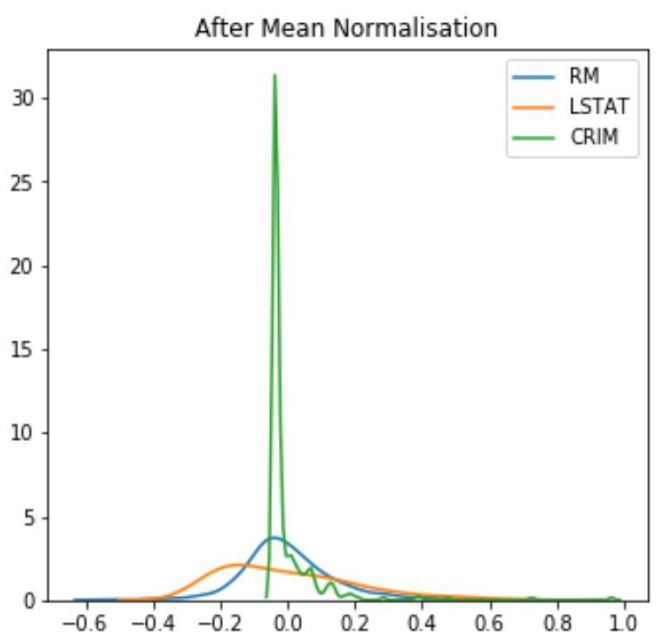
Price
0.44
0.39
0.17
0.11
0.00
0.44
0.17
0.22
0.56
0.11
1.00

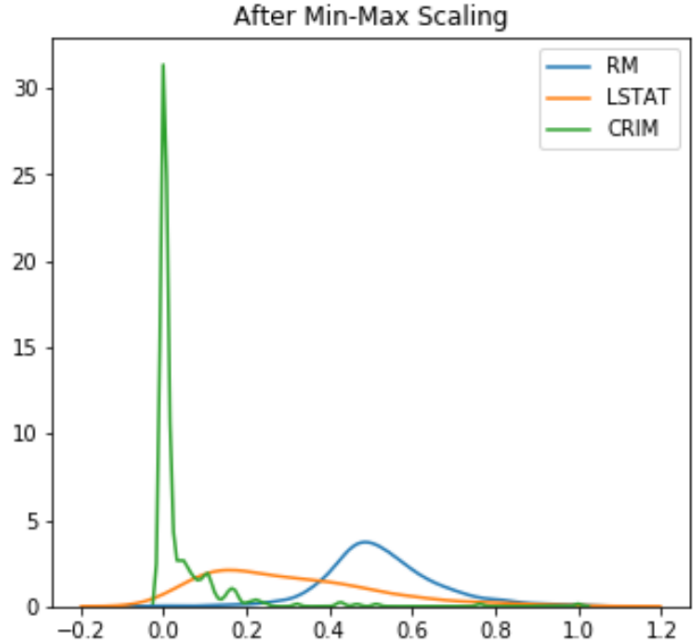
MinMaxScaling: summary

- Mean varies
- Variance varies
- May alter shape of the original distribution
- Minimum and maximum values within [0;1]
- Preserves outliers

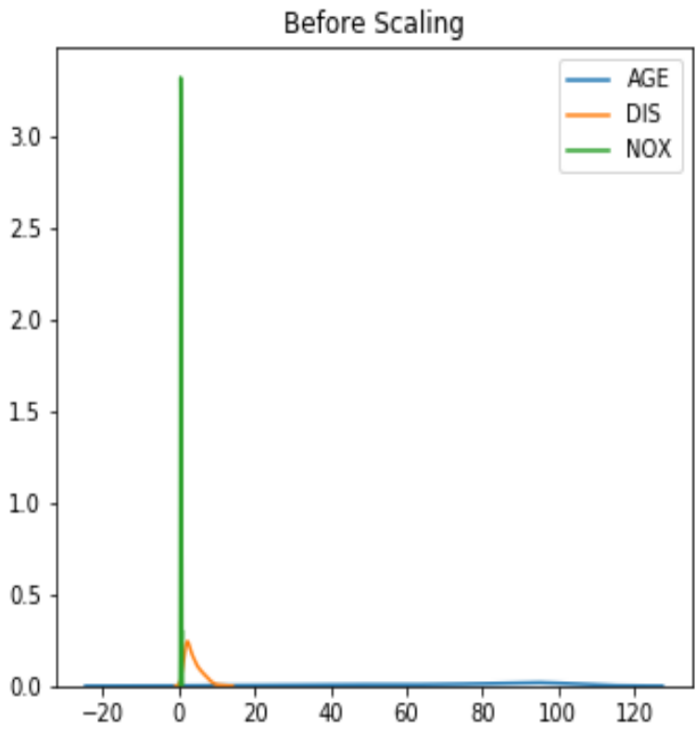
MinMaxScaling: Notebook

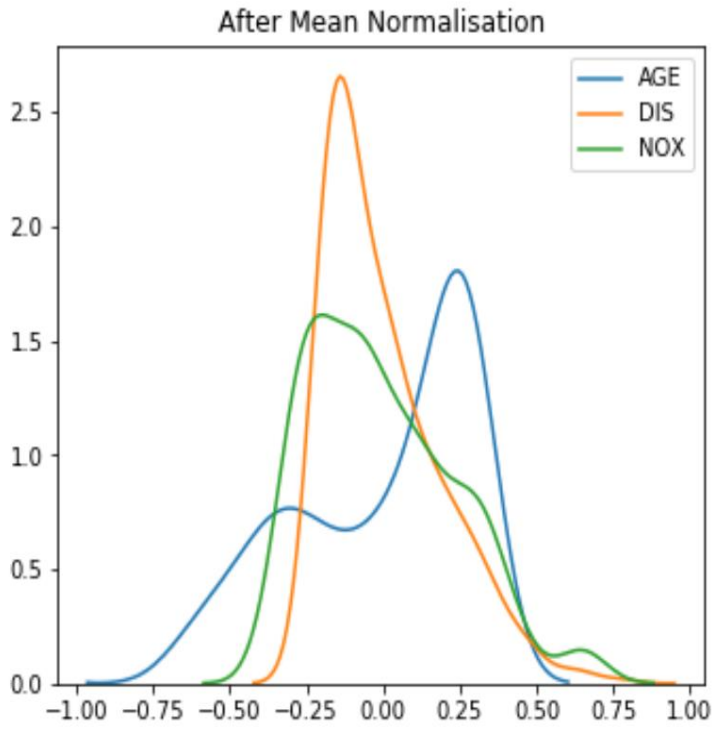


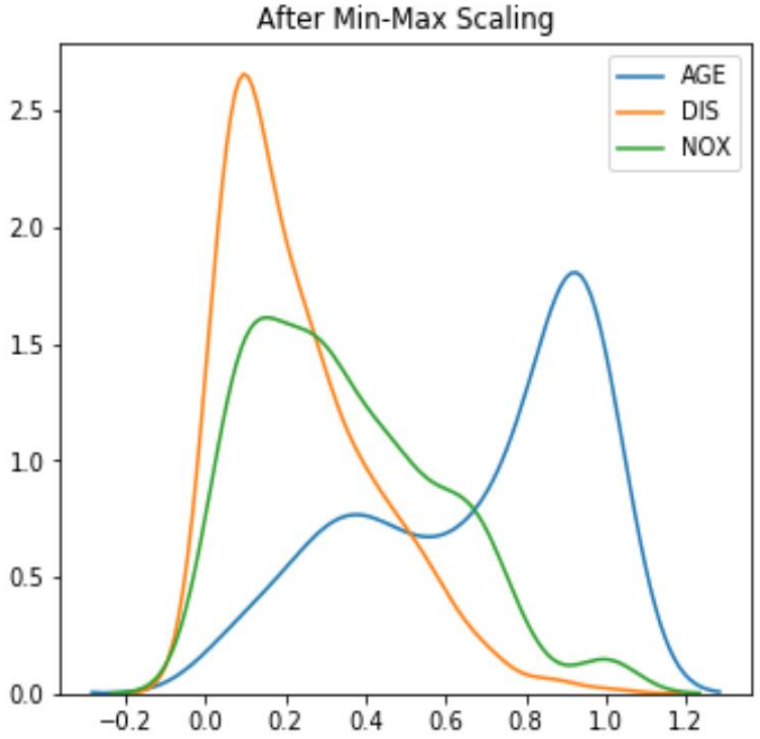




MinMaxScaling : Notebook







Train In Data

THANK YOU

www.trainindata.com