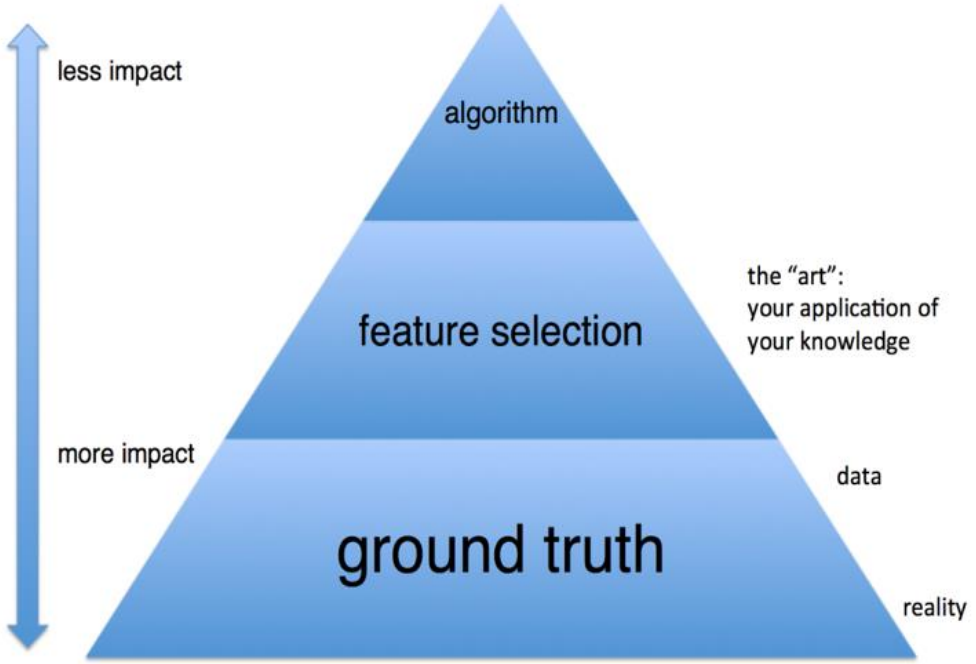
特徵工程

講者：Isaac

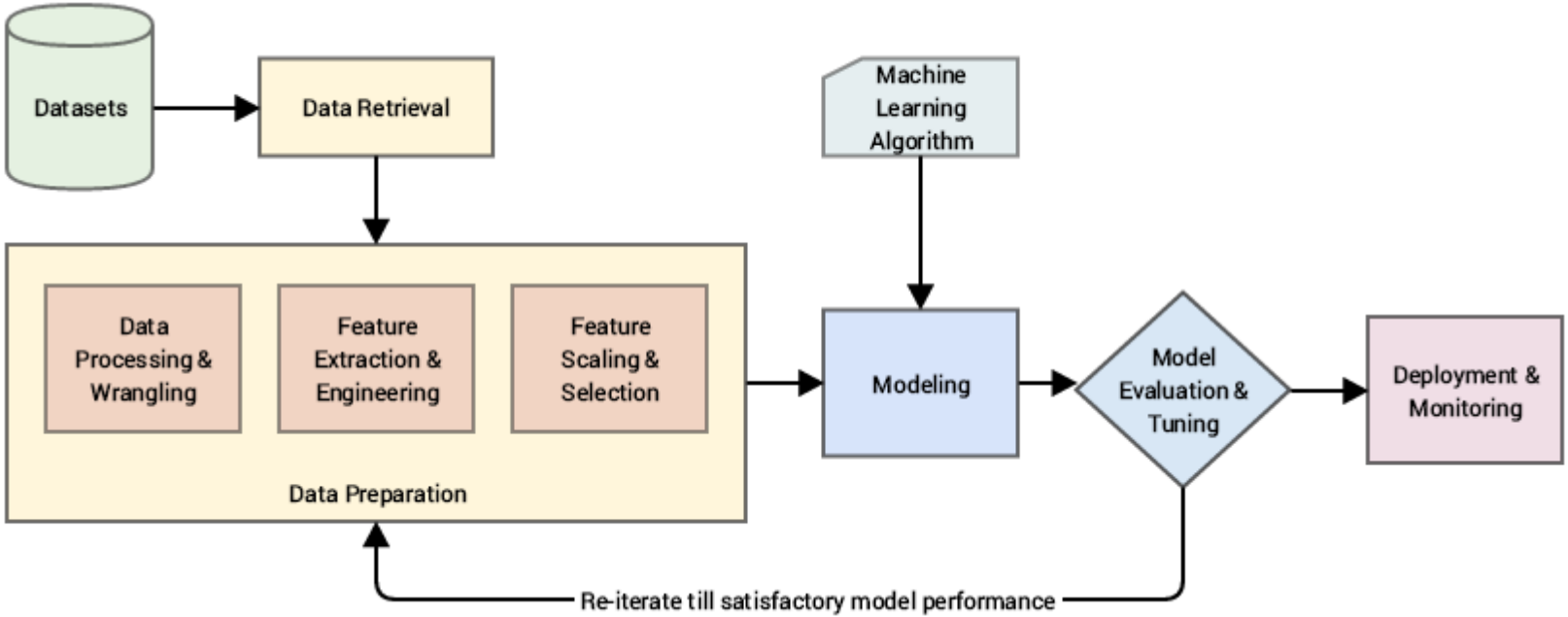
Outline

- Data Exploration
- Feature Cleaning
- Feature Engineering
- Feature Selection

Data Exploration


less impact
more impact
algorithm
feature selection
ground truth
the "art":
your application of
your knowledge
data
reality

ML Flow


Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

Types of Variable

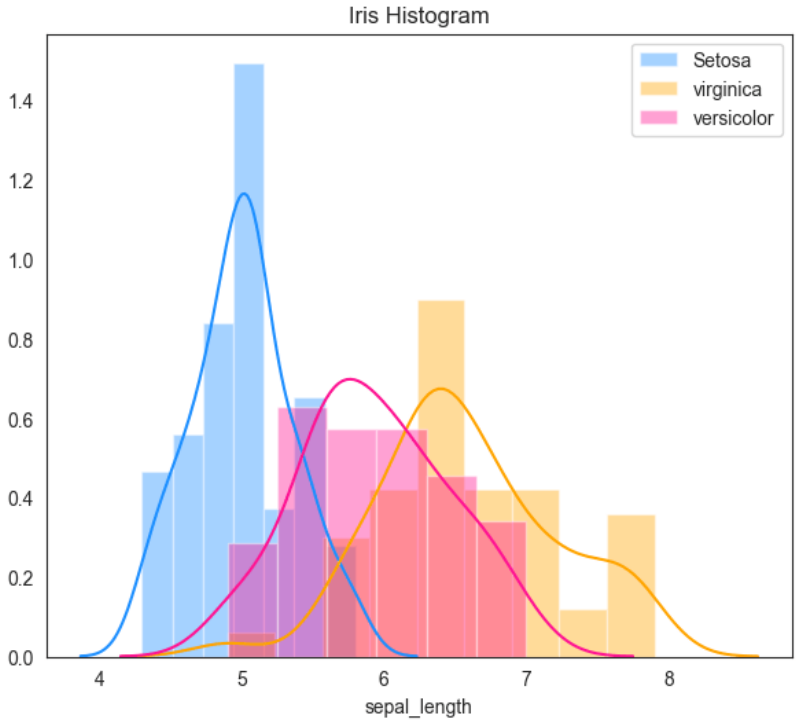
Type	Sub-type	Definition	Example
Categorical	Nominal	Variables with values selected from a group of categories, while not having any kind of natural order.	Gender, car types
	Ordinal	A categorical variable whose categories can be meaningfully ordered.	Grade of an exam
Numerical	Discrete	Variables whose values are either finite or countably infinite.	Number of children in a family
	Continuous	Variable which can take on infinitely many, uncountable values.	House prices, time passed

Univariate Analysis

- Descriptive statistics on one single variable

Variable	**What to look for**
Categorical	**Shape**: Histogram/ Frequency table...
Numerical	**Central Tendency**: Mean/ Median/ Mode **Dispersion**: Min/ Max/ Range/ Quantile/ IQR/ MAD/ Variance/ Standard Deviation/ **Shape**: Skewness/ Histogram/ Boxplot...

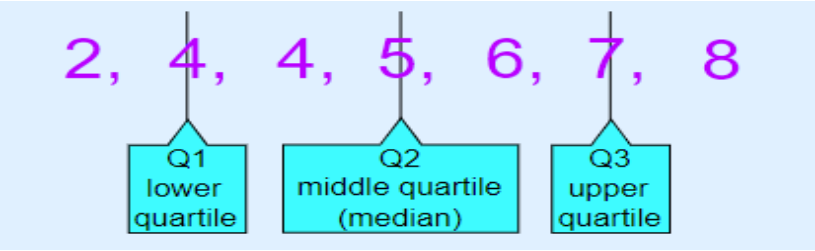
Histogram

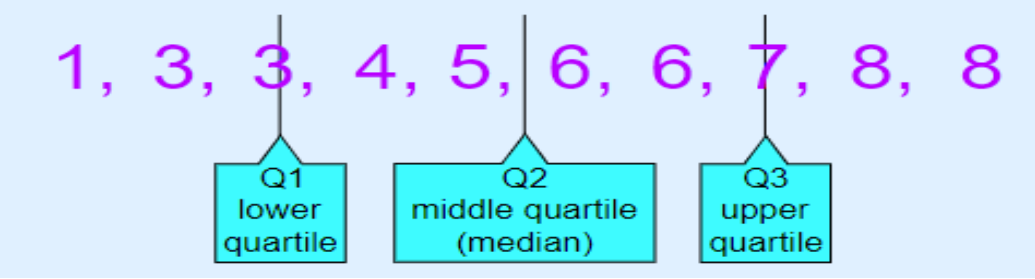

Iris Histogram
Setosa
virginica
versicolor
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
4
5
6
7
8
sepal_length

Frequency table

Score	Frequency
6	2
7	3
8	7
9	7
10	1

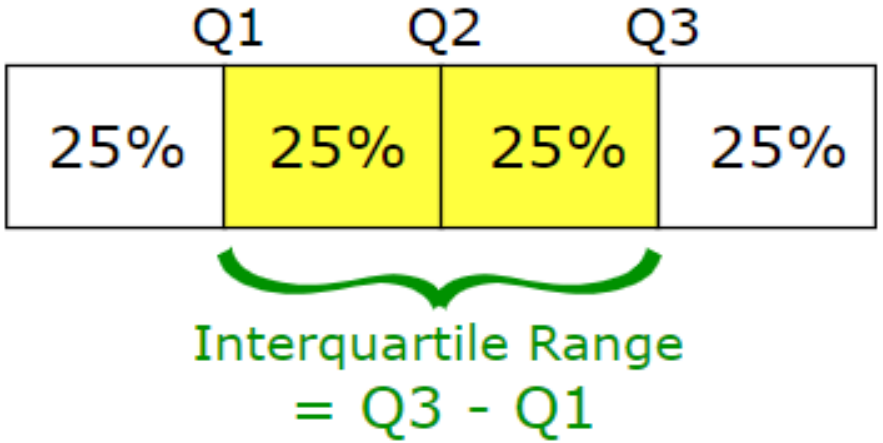
median, Q1, Q3, IQR





median, Q1, Q3, IQR

- Interquartile Range(IQR) = Q1 – Q3



median absolute deviation (MAD)

(1, 1, 2, 2, 4, 6, 9) Mean = 2

absolute deviations

(1, 1, 0, 0, 2, 4, 7)

sorting

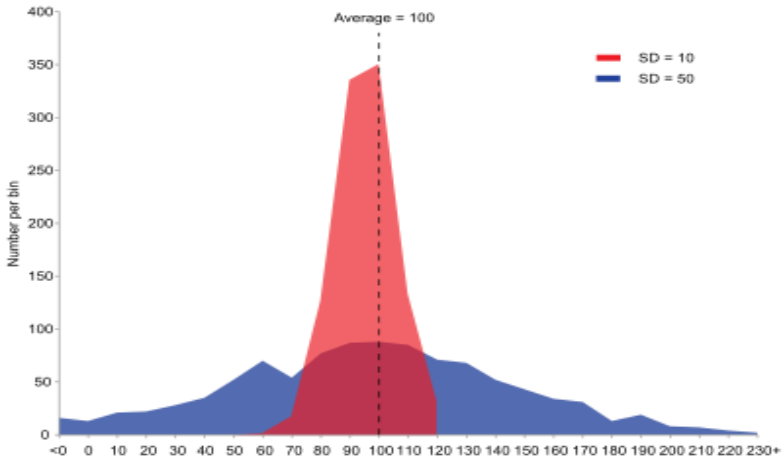
(0, 0, 1, **1**, 2, 4, 7))

standard deviation, variance

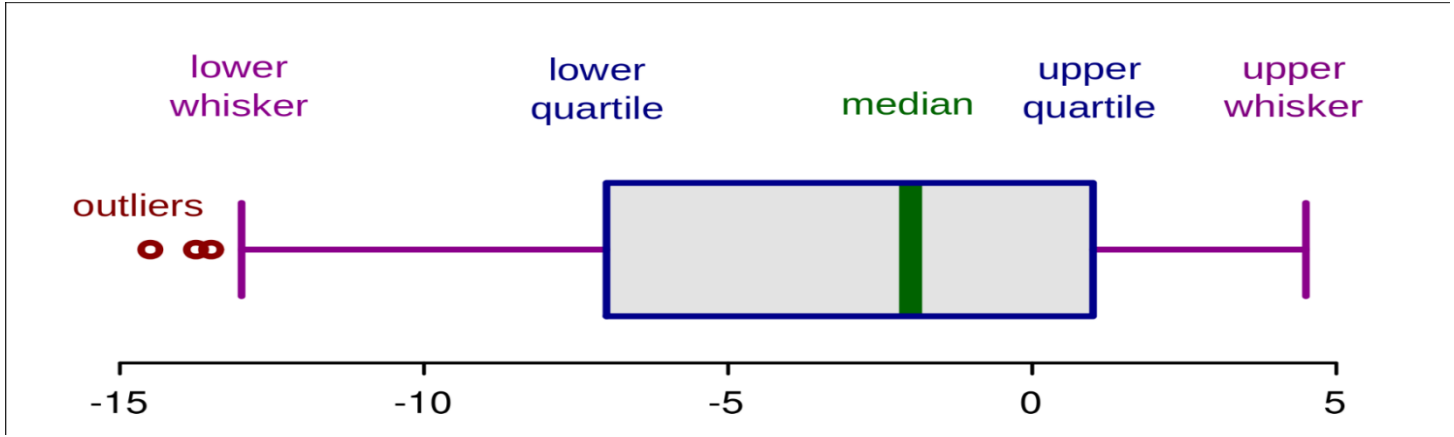
$$variance = \sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

$$standard\ deviation = \sigma = \sqrt{\frac{\sum(x_i - \mu)}{n}}$$

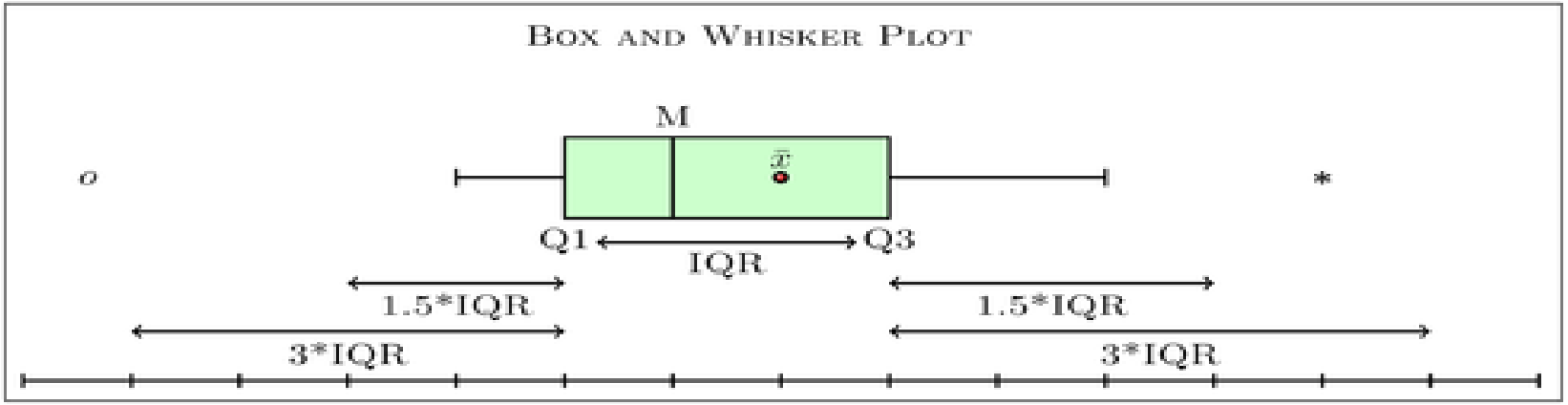
$$\mu : mean$$



Boxplot

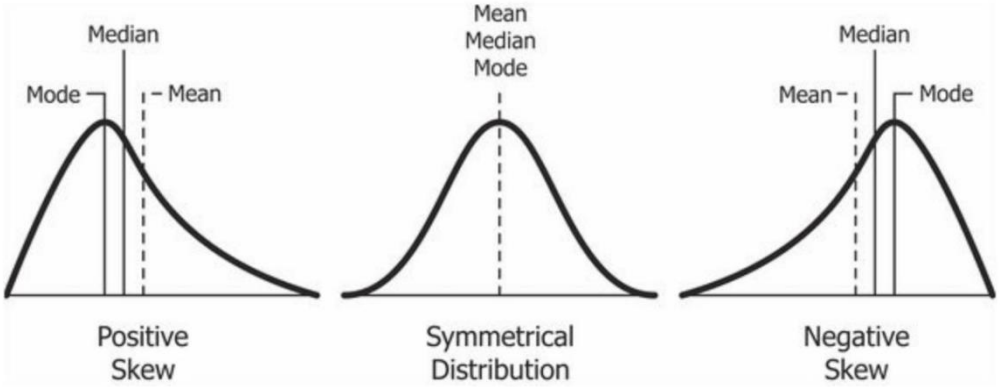

lower whisker
lower quartile
median
upper quartile
upper whisker
outliers
-15
-10
-5
0
5

Boxplot



Skewness

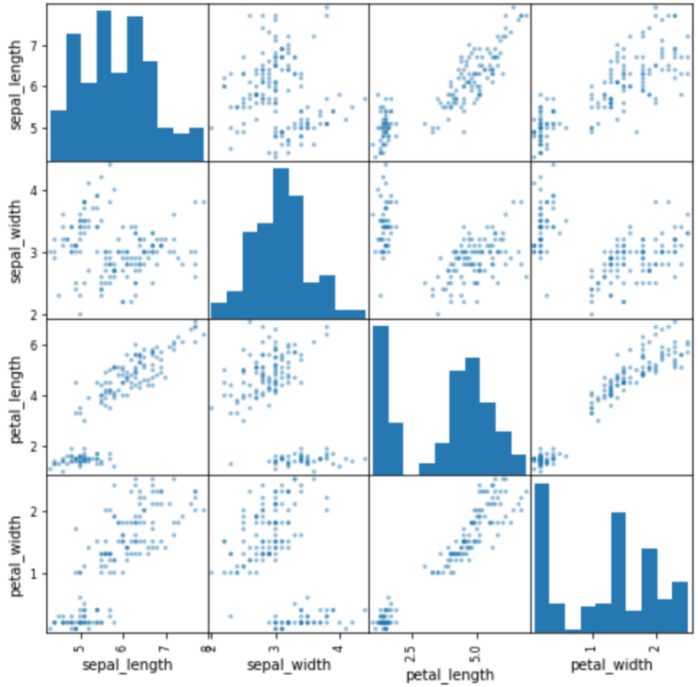
Skewness: $$S_k = \frac{\sum_{i=1}^{T} (x_i - \bar{x})^3}{\sigma^3}$$



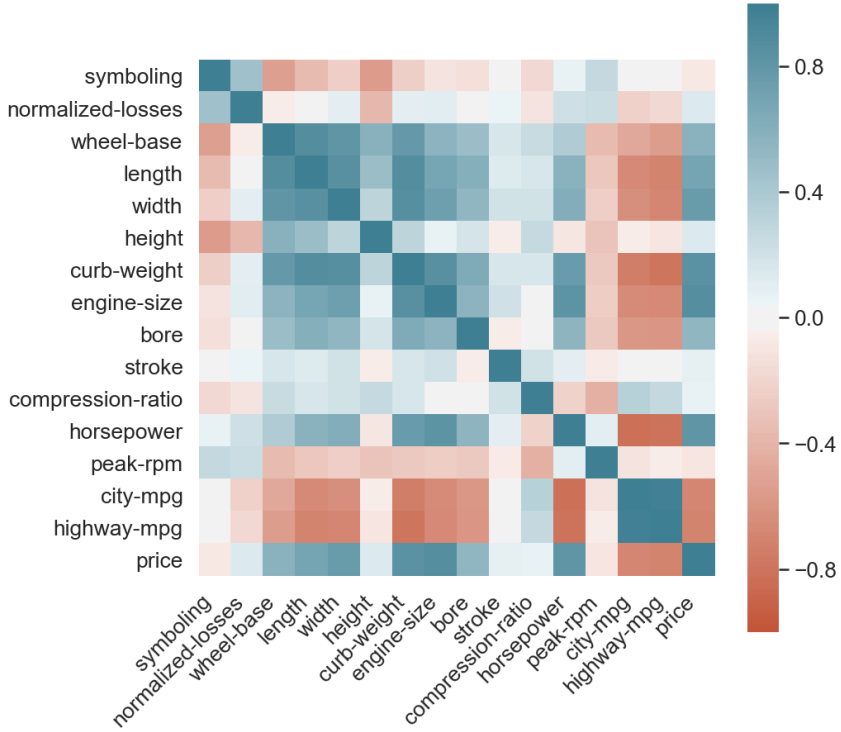
Bi-variate Analysis

- Descriptive statistics between two or more variables.
  - Usually use Scatter Plot, Correlation Plot

Scatter Plot



Correlation Plot



Feature Cleaning

Missing Values

- **Missing Values**
  - no value is stored in a certain observation within a variable
- **Why Missing Data Matters**
  - certain algorithms cannot work when missing value are present
  - even for algorithm that handle missing data, without treatment the model can lead to inaccurate conclusion

Types of Missing Values

- **Missing Completely at Random** (MCAR)
  - there is absolutely no relationship between the data missing and any other values within the dataset
- **Missing at Random**(MAR)
  - men are more likely to disclose their weight than women
- **Missing Not At Random - Depends on Unobserved Predictors**
  - if a particular treatment causes discomfort, a patient is more likely to drop out of the study (and 'discomfort' is not measured)
- **Missing Not At Random - Depends on Missing Value Itself**
  - Missingness depends on the (potentially missing) variable itself. E.g., people with higher earnings are less likely to reveal them.

How to Assume a Missing Mechanism

- By **business understanding**
  - In many situations we can assume the mechanism by probing into the business logic behind that variable.
- By **statistical test**
  - Divide the dataset into ones with/without missing and perform t-test to see if there's significant differences. If there is, we can assume that missing is not completed at random.

How to Handle Missing Data

Method	Definition	Pros	Cons
Listwise Deletion	excluding all cases (listwise) that have missing values	preserve distribution if MCAR	1. may discard too much data and hurt the model 2. may yield biased estimates if not MCAR (as we keep a special subsample from the population)
Mean/Median/Mode Imputation	replacing the NA by mean/median/most frequent values (for categorical feature) of that variable	good practice if MCAR	1. distort distribution 2. distort relationship with other variables
End of distribution Imputation	replacing the NA by values that are at the far end of the distribution of that variable, calculated by mean + 3*std	Captures the importance of missingness if there is one	1. distort distribution 2. may be considered outlier if NA is few or mask true outlier if NA is many. 3. if missingness is not important this may mask the predictive power of the original variable

How to Handle Missing Data

Method	Definition	Pros	Cons
Random Imputation	replacing the NA by taking a random value from the pool of available observations of that variable	preserve distribution if MCAR	not recommended in business settings for its randomness (different result for same input)
Arbitrary Value Imputation	replacing the NA by arbitrary values	Captures the importance of missingness if there is one	1. distort distribution 2. typical used value: -9999/9999. But be aware it may be regarded as outliers.
Add a variable to denote NA	creating an additional variable indicating whether the data was missing for that observation	Captures the importance of missingness if there is one	expand feature space

Outliers

- An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism
- **Outliers, depending on the context, either deserve special attention or should be completely ignored**
  - an unusual transaction on a credit card is usually a sign of fraudulent activity → should pay more attention
  - a height of 1600cm of a person is very likely due to measurement error → should filter out

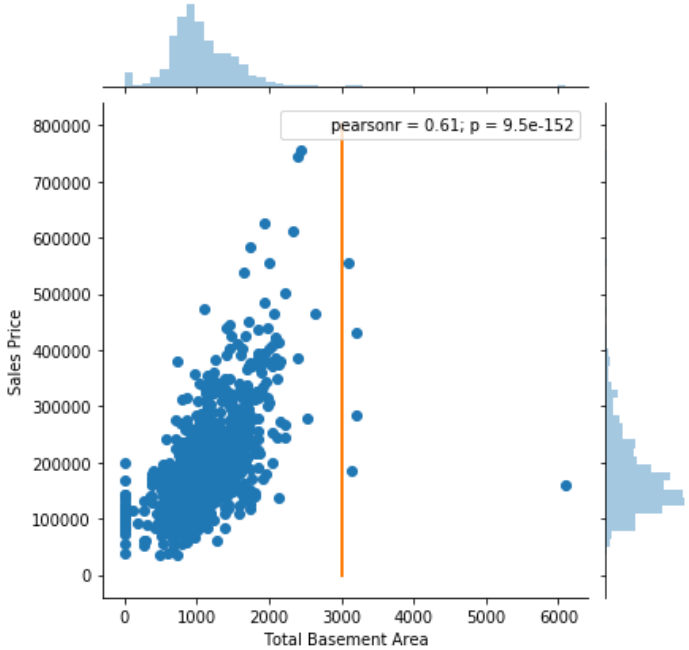
Why Outlier Matters

- make algorithm not work properly
- introduce noises to dataset
- make samples less representative
- Algorithm that is sensitive to outlier
  - Adaboost
- Algorithm that is not sensitive to outlier
  - decision trees, DBSCAN

Outlier Detection

Method	Definition	Pros	Cons
Detect by arbitrary boundary	identify outliers based on arbitrary boundaries	flexiable	require business understanding
Mean & Standard Deviation method	outlier detection by Mean & Standard Deviation Method	good for variable with Gaussian distribution (68-95-99 rule)	sensitive to extreme value itself (as the outlier increase the sd)
IQR method	outlier detection by Interquartile Ranges Rule	robust than Mean & SD method as it use quantile & IQR. Resilient to extremes.	can be too aggressive
MAD method	outlier detection by Median and Median Absolute Deviation Method	robust than Mean & SD method. Resilient to extremes.	can be too aggressive

Detect by arbitrary boundary



How to Handle Outliers

Method	Definition	Pros	Cons
Mean/Median/Mode Imputation	replacing the outlier by mean/median/most frequent values of that variable	preserve distribution	lose information of outlier if there is one
Imputation with arbitrary value	impute outliers with arbitrary value.	flexiable	hard to decide the value
Windsorization	top-coding & bottom coding (capping the maximum of a distribution at an arbitrarily set value, vice versa).	prevent model over-fitting	distort distribution
Discard outliers	drop all the observations that are outliers	/	lose information of outlier if there is one

Feature Engineering

Feature Scaling

- Feature scaling is a method used to standardize the range of independent variables or features of data
  - also known as data normalization

Why Feature Scaling Matters

- if range of inputs varies, in some algorithms, object functions will not work properly
- Gradient descent converges much faster with feature scaling done
- Algorithms that involve distance calculation are also affected by the magnitude of the feature.

How to Handle Feature Scaling

Method	Definition	Pros	Cons
Normalization - Standardization (Z-score scaling)	removes the mean and scales the data to unit variance. z = (X - X.mean) / std	feature is rescaled to have a standard normal distribution that centered around 0 with SD of 1	compress the observations in the narrow range if the variable is skewed or has outliers, thus impair the predictive power.
Min-Max scaling	transforms features by scaling each feature to a given range. Default to [0,1]. X_scaled = (X - X.min / (X.max - X.min)	/	compress the observations in the narrow range if the variable is skewed or has outliers, thus impair the predictive power.
Robust scaling	removes the median and scales the data according to the quantile range (defaults to IQR) X_scaled = (X - X.median) / IQR	better at preserving the spread of the variable after transformation for skewed variables	/

Z-score V.S. Max-min V.S. Rubost Scaler

1	Original	Standardization	Max-Min Scaler	Rubost Scaler
2	6.9314183	-0.2244971	0.0000003	0.8283487
3	2.6674115	-0.2244979	0.0000001	0.0690181
4	7.7248183	-0.2244970	0.0000003	0.9696367
5	5.7388433	-0.2244973	0.0000002	0.6159760
6	0.8965615	-0.2244982	0.0000000	-0.2463333
7	4.5147618	-0.2244975	0.0000002	0.3979926
8	2.9934144	-0.2244978	0.0000001	0.1270724
9	4.8708377	-0.2244975	0.0000002	0.4614023
10	4.2797819	-0.2244976	0.0000002	0.3561476
11	1.0085616	-0.2244982	0.0000000	-0.2263885
12	5.5166580	-0.2244974	0.0000002	0.5764094
13	1.1171326	-0.2244981	0.0000000	-0.2070542
14	0.4069897	-0.2244983	0.0000000	-0.3335159
15	5.0536949	-0.2244975	0.0000002	0.4939654
16	8.4068370	-0.2244969	0.0000003	1.0910900
17	8.9588050	-0.2244968	0.0000003	1.1893840
18	0.9543401	-0.2244982	0.0000000	-0.2360442
19	94750.5292279	-0.2079018	0.0037104	16872.6857158
20	2051.2433203	-0.2241390	0.0000803	364.8776314
21	25536631.9371928	4.2485000	1.0000000	4547540.7645023

Discretize

- Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that spans the range of the variable's values.

Why Discretize Matters

- help to improve model performance by grouping of similar attributes with similar predictive strengths
- bring into non-linearity and thus improve fitting power of the model
- enhance interpretability with grouped values
- minimize the impact of extreme values/seldom reversal patterns

How to Handle Discretization

Method	Definition	Pros	Cons
Equal width binning	divides the scope of possible values into N bins of the same width	/	sensitive to skewed distribution
Equal frequency binning	divides the scope of possible values of the variable into N bins, where each bin carries the same amount of observations	may help boost the algorithm's performance	this arbitrary binning may disrupt the relationship with the target
K-means binning	using k-means to partition values into clusters	/	needs hyper-parameter tuning
Discretization using decision trees	using a decision tree to identify the optimal splitting points that would determine the bins	observations within each bin are more similar to themselves than to those of other bins	1. may cause over-fitting 2. may not get a good performing tree
ChiMerge	supervised hierarchical bottom-up (merge) method that locally exploits the chi-square criterion to decide whether two adjacent intervals are similar enough to be merged	robust and make use of a priori knowledge	cannot handle unlabeled data

Feature Encoding

- transform strings of categorical variables into numbers so that algorithms can handle those values

How to Handle Feature Encoding

Method	Definition	Pros	Cons
One-hot encoding	replace the categorical variable by different boolean variables (0/1) to indicate whether or not certain label is true for that observation	keep all information of that variable	1. expand feature space dramatically if too many labels in that variable 2. does not add additional value to make the variable more predictive
Ordinal-encoding	replace the labels by some ordinal number if ordinal is meaningful	straightforward	does not add additional value to make the variable more predictive

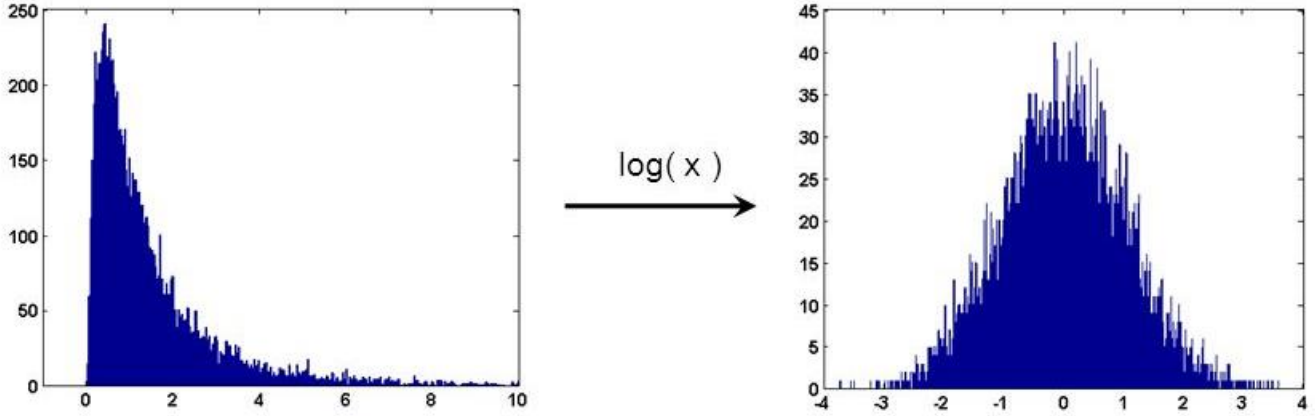
How to Handle Feature Encoding

Method	Definition	Pros	Cons
Count/frequency encoding	replace each label of the categorical variable by the count/frequency within that category	/	1. may yield same encoding for two different labels (if they appear same times) and lose valuable info. 2. may not add predictive power
Mean encoding	replace the label by the mean of the target for that label. (the target must be 0/1 valued or continuous)	1. Capture information within the label, therefore rendering more predictive features 2. Create a monotonic relationship between the variable and the target 3. Do not expand the feature space	Prone to cause over-fitting

Feature Transformation

Method	Definition
Logarithmic transformation	log(x+1). We use (x+1) instead of x to avoid value of 0
Reciprocal transformation	1/x. Warning that x should not be 0.
Square root transformation	x**(1/2)
Exponential transformation	X**(m)
Box-cox transformation	(X**λ-1)/λ
Quantile transformation	transform features using quantiles information

Logarithmic transformation



Feature Selection

Feature Selection

- Feature Selection is the process of selecting a subset of relevant features for use in machine learning model building.

Filter Method

- select features based on a performance measure regardless of the ML algorithm later employed

Filter Method

Method	Definition
Variance	removing features that show the same value for the majority/all of the observations (constant/quasi-constant features)
Correlation	remove features that are highly correlated with each other

Wrapper Method

- use a search strategy to search through the space of possible feature subsets
  - evaluate each subset by the quality of the performance on a ML algorithm

Wrapper Method

- use ML models to score the feature subset
- train a new model on each subset
- very computationally expensive
- usually provide the best performing subset for a give ML algorithm, but probably not for another
- need an arbitrary defined stopping criteria

Wrapper Method

Method	Definition
Forward Selection	evaluates all possible combinations of the selected feature and a second feature, and selects the pair that produce the best performing algorithm based on the same pre-set criteria
Backward Elimination	starts by fitting a model using all features and then remove the one that produces the highest performing algorithm

Embedded Method

- Use learning algorithm to perform feature selection and classification at same time
  - Lasso, Random forest, ……