

Data Preprocessing

Adapted from Discovering Knowledge in Data:
An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John
Wiley and Sons, Inc.,

Where & Why Do We Preprocess Data?

CRISP-DM Review



- Focusing on Data Understanding and Data Preparation Process CRISP-DM process
- For data mining purposes, database values must undergo data cleaning and data transformation
- Raw data often unprocessed, incomplete, noisy
- May contain:
 - Obsolete/redundant fields
 - Missing values
 - Outliers
 - Data in form not suitable for data mining
 - Values not consistent with common sense

Data Cleaning – Example

Can You Find Any Problems in This Tiny Data Set?

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75000	C	M	5000
1002	J2S7K7	F	—40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	F	99999	30	D	3000

Data-set of USA Customer Transactions

CustomerID field is assumed to be fine; But Zip Code, Gender?

- **Zip Code**
 - Do not assume local format
 - 90210 (U.S.) vs. J2S7K7 (Canada)
 - In a free trade era should expect some unusual values
 - Be aware of data type/conversion issues
 - Zip code 06269 stored in numeric field truncates the leading zeroes, and thus, is represented as 6269 (Zip Code for Storrs, CT)
- **Gender**
 - Value is missing for customer 1003
- **Income**
 - ?
- **Age**
 - ?
- **Marital Status**
 - ?

Handling Missing Data

- Examine *cars* dataset containing records for 261 automobiles manufactured in 1970s and 1980s
- Suppose that some fields are missing for certain records, like in figure below:

	mpg	cubicinches	hp	brand
1	14.000	350	165	US
2	31.900		71	Europe
3	17.000	302	140	US
4	15.000	400	150	
5	37.700	89	62	Japan

- **Delete Records Containing Missing Values?**
 - Dangerous, as pattern of missing values may be systematic
 - Valuable information in other fields lost
- **Three alternative methods available** – Not entirely satisfactory
 - Method 1 – Replace with a Constant
 - Method 2 – Replace with Mode or Mean
 - Method 3 – Replace with Random values
- **Data imputation methods** – Better approach
 - i.e. Imputed value based on other characteristics of the record

Handling Missing Data (*cont'd*)

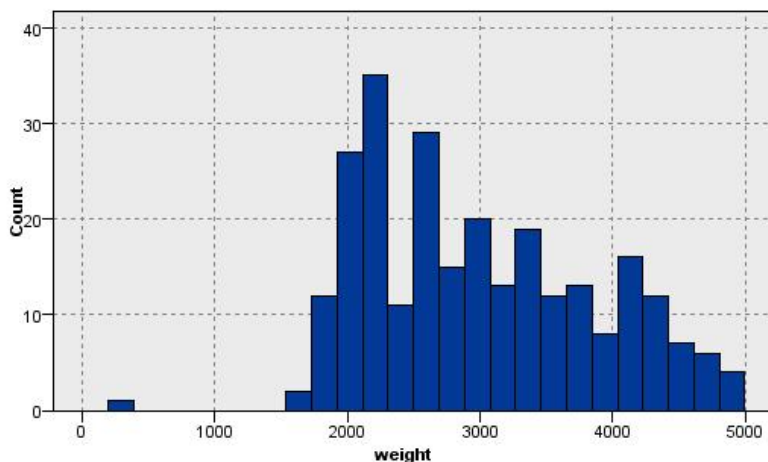
- **Data Imputation Methods**

- Imputation of Missing Data - What is the likely value, given record's other attribute values?
- Example: From two samples below, American car would be expected to have more cylinders
 - American car with 300 cubic inches and 150 horsepower
 - Japanese car with 100 cubic inches and 90 horsepower
- Requires tools like multiple regression, or classification and regression trees

Graphical Methods for Identifying Outliers

(cont'd)

- Outliers are extreme values that go against the trend of the remaining data
- Method #1 - Histogram
 - A histogram examines values of numeric fields
 - Example: Histogram shows vehicle weights for *cars* data set* (*cars2.txt*)
 - The extreme left-tail contains one outlier weighing several hundred pounds (192.5)
 - Should we doubt validity of this value? This is too light for a car.
 - Possibility: Original value was 1925 pounds. Requires further investigation.

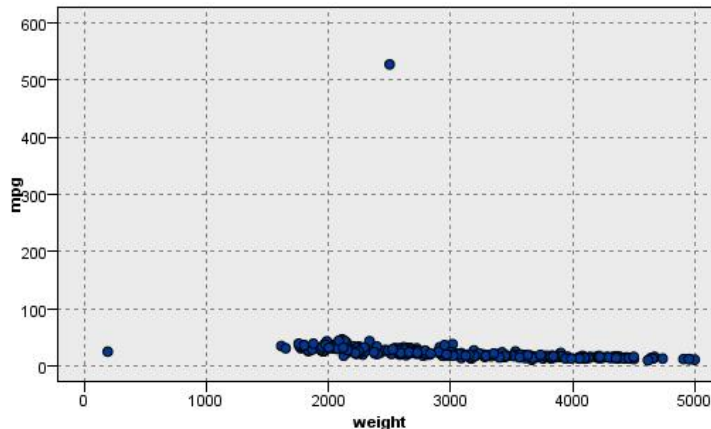


Graphical Methods for Identifying Outliers

(*cont'd*)

- Method #2 – Two-dimensional Scatter Plot

- Two-dimensional scatter plots help determine outliers in more than one variable
- Example: Scatter plot of *mpg* against *weightlbs* shows two possible outliers
 - Most data points cluster together along x-axis
 - However, one car weighs 192.5 pounds and other gets over 500 miles per gallon?
 - Important: A record may be outlier in a particular dimension, but not in the other



Measures of Center

Measures of center

Estimate where the center of a particular variable lies

- Most common *measures of center*
 - Mean, Median and Mode
 - They are a special case of *measures of location*, which indicate where a numeric variables lies (examples: percentiles and quantiles)

Measures of Spread

Measures of Spread - Introduction

- Measures of location not enough to summarize a variable
- Example: Table with P/E ratios for two portfolios (below)
 - Portfolio A – Spread with one very low and one very high value
 - Portfolio B – Tightly clustered around the center
 - P/E ratios for each portfolio is distinctly different, yet **they both** have P/E ratios with mean 10, median 11 and mode 11
- Clearly, measures of center do not provide a complete picture
- Measures of spread or measure of variability complete the picture by describing how spread the data values of each portfolio are

Stock Portfolio A	Stock Portfolio B
1	7
11	8
11	11
11	11
16	13

Measures of Spread

Measures of Spread - Introduction

- Typical measures of variability include
 - **Range** (maximum – minimum)
 - **Standard Deviation** – Sensitive to the presence of outliers (because of the squaring involved – see below)
 - **Mean Absolute Deviation** – Preferred in situations involving extreme values
 - Interquartile Range
- Sample Standard Deviation is defined by
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$
 - Interpreted as “typical” distance between a field value and the mean
 - Most field values lie within two standard deviations of the mean

Data Transformation

- Variables tend to have ranges different from each other
- In a dataset, two fields may have ranges:
 - **Annual Income:** [€15,000, €300,000]
 - **Employee Age:** [16, 70]
- Some data mining algorithms adversely affected by differences in variable ranges
 - Variables with greater ranges tend to have larger influence on data model's results
 - Therefore, numeric field values should be normalized
- Standardizing scales the effect each variable has on results
- kNearest Neighbour, Neural Networks and other algorithms that make use of distance measures benefit from normalization

Min-Max Normalization

Find Min-Max normalization for cars weighing 1613, 3384 and 4997 pounds, respectively

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

Where:

$$\min(X) = 1613$$

$$\max(X) = 4997$$

Car	Weightlbs	Formula	Result	Comments
Ultra-light vehicle	$X = 1613$	$X^* = \frac{1613 - 1613}{4997 - 1613}$	$X^* = 0$	Represents the minimum value in this variable, and has min-max normalization of zero.
Mid-range vehicle	$X = 3384$	$X = \frac{3384 - 1613}{4997 - 1613}$	$X^* = 0.5$	Weight exactly half-weight between the lightest and the heaviest vehicle, and has min-max normalization of 0.5.
Heaviest vehicle	$X = 4997$	$X = \frac{4997 - 1613}{4997 - 1613}$	$X^* = 1$	Heaviest vehicle of the dataset has min-max normalization of one.

Min-Max normalization will always have a value between 0 and 1.

Z-score Standardization

- Widely used in statistical analysis
- Takes difference between field value and field value mean
- Scales this difference by field's standard deviation

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

Find Z-score standardization for cars weighing 1613, 3384 and 4997 lbs, respectively

Summary Statistics for weightlbs

weightlbs	
Statistics	
Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

Car	Weightlbs	Formula	Result	Comments
Ultra-light vehicle	$X = 1613$	$X^* = \frac{1613 - 3005.49}{852.646}$	$X^* \approx -1.63$	Data values below the mean will have negative Z-score standardization.
Mid-range vehicle	$X = 3384$	$X = \frac{3384 - 3005.49}{852.646}$	$X^* \approx 0$	Values falling exactly on the mean will have zero (0) Z-score
Heaviest vehicle	$X = 4997$	$X = \frac{4997 - 3005.49}{852.646}$	$X^* \approx 2.34$	Data values about the mean will have a negative Z-score standardization

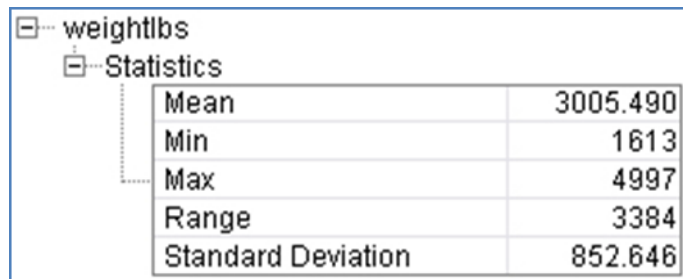
Decimal Scaling

- Ensures that normalized values lies between -1 and 1
- Defined as:

$$X^* = \frac{X}{10^d}$$

where d represents the number of digits in the data value with the largest absolute value.

- For the weight data, the largest absolute value is $|4997| = 4997$, with $d=4$ digits
- Decimal scaling for the minimum and maximum weights are:

A screenshot of a software interface showing a tree view with 'weightlbs' expanded to 'Statistics'. Below it is a table with statistical data.

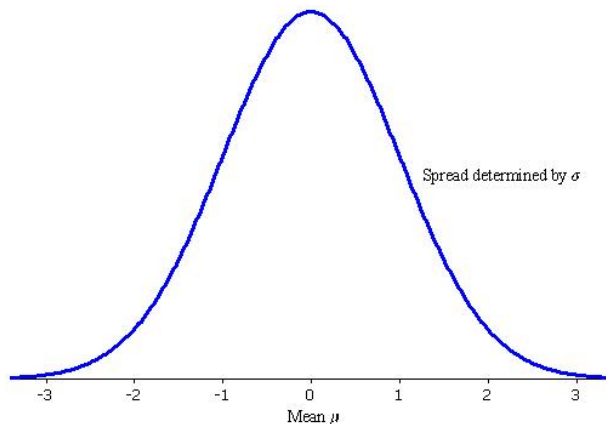
weightlbs	
Statistics	
Mean	3005.490
Min	1613
Max	4997
Range	3384
Standard Deviation	852.646

$$\text{Min: } X_{decimal}^* = \frac{1613}{10^4} = 0.1613$$

$$\text{Max: } X_{decimal}^* = \frac{4997}{10^4} = 0.4997$$

Transformations to achieve normality

- Some data mining algorithms and statistics methods require *normally distributed* variables
- Normal distribution
 - Continuous probability distribution known as the ‘bell curve’ (symmetric)
 - Centered and mean μ (myu) and spread given by σ (sigma)



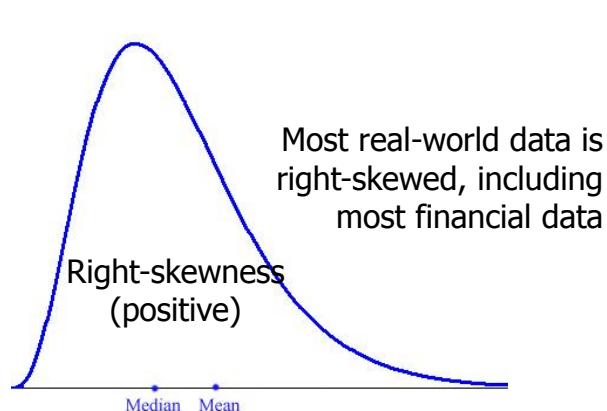
Standard normal Z-distribution
with $\mu=0$ and $\sigma=1$

Transformations to achieve normality

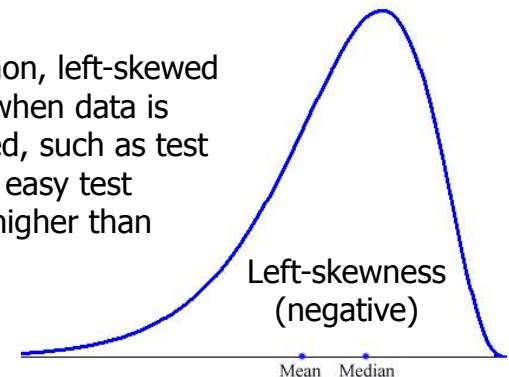
- Some data mining algorithms and statistics methods require *normally distributed* variables
- Statistics for measuring **the skewness of a distribution**:

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

- Right-skewness data – Is positive, as mean is greater than the median
- Left skewness data – Mean is smaller than the median, generating negative values
- Perfectly symmetric data – mean, median and mode are equal, so skewness is zero



Not as common, left-skewed data occurs when data is right-censored, such as test scores on an easy test (cannot get higher than 100).



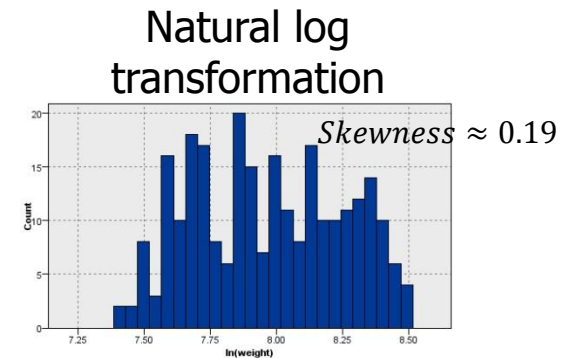
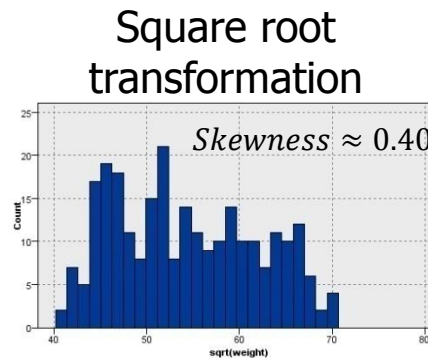
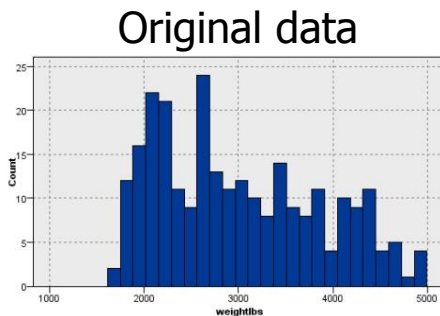
Transformations to achieve normality

- To eliminate skewness, we must apply a transformation to the data
 - This makes the data symmetric and makes it “more normally distributed”
- Common transformations are:

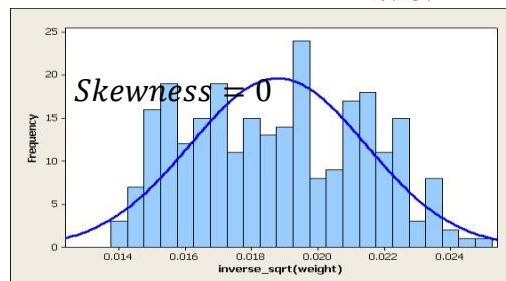
Natural Log	Square Root	Inverse Square Root
$\ln(y)$	\sqrt{y}	$\frac{1}{\sqrt{y}}$

Square root transformation somewhat reduces skewness, while ln reduces it further while inverse square root reduces further again ($y = \text{weightlbs}$)

Important: There is nothing special about the inverse square root transformation. It just worked with the skewness in the weight data



Inverse Square Root transformation

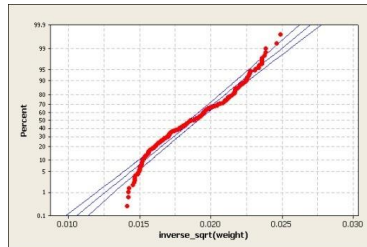


Notice that while we have achieved symmetry, we have not reached normality (the distribution does not match the normal curve)

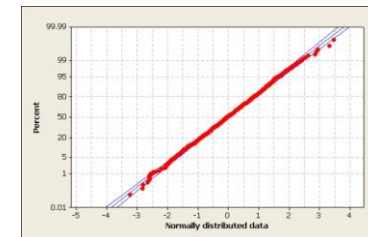
Transformations to achieve normality (*cont'd*)

- After achieving symmetry, we must also check for normality
- The Normal Probability Plot
 - Plots the quantiles for a particular distribution *against* the quantiles of the standard normal distribution
 - Similar to percentile, p^{th} quantile of a distribution is value x_p , such that $p\%$ of the distribution values are less than or equal to x_p
 - If the bulk of the points fall on a straight line, the distribution is normal; systematic deviations indicate nonnormality
- As expected, the normal probability plot for the `inverse_sqrt(weight)` indicates nonnormality

Normal probability plots



Plot for `inverse_sqrt(weight)` has systematic deviations that indicate nonnormality



Plot for normally distributed data

Numerical Methods for Identifying Outliers

- **Using Z-score Standardization to Identify Outliers**
 - Outliers are Z-score Standardization values either less than -3, or greater than 3
 - Values much beyond range [-3, 3] require further investigation to determine their validity
 - Should not automatically omit outliers from analysis
 - However, Mean and Standard Deviation are both sensitive to the presence of outliers
 - Sample Mean and sample standard deviation are both part of the formula for z-score standardization
 - If an outlier is added or deleted from the dataset, mean and standard deviation is affected
 - When selecting a method for evaluating outliers, should not use measures which are themselves sensitive to outliers

Numerical Methods for Identifying Outliers

(cont'd)

- Using Interquartile Range (IQR) to Identify Outliers
 - Robust statistical method and less sensitive to presence of outliers
 - Data divided into four quartiles, each containing 25% of data
 - First quartile (Q1) 25th percentile
 - Second quartile (Q2) 50th percentile (median)
 - Third quartile (Q3) 75th percentile
 - Fourth quartile (Q4) 100th percentile
 - IQR is measure of variability in data

Numerical Methods for Identifying Outliers *(cont'd)*

- $IQR = Q3 - Q1$ and represents spread of middle 50% of the data
- Data value defined as outlier if located:
 - $1.5 \times (IQR)$ or more below $Q1$; or
 - $1.5 \times (IQR)$ or more above $Q3$
- For example, set of test scores have 25th percentile ($Q1$) = 70, and 75th percentile ($Q3$) = 80
- 50% of test scores fall between 70 and 80 and Interquartile Range (IQR) = $80 - 70 = 10$
- Test scores are identified as outliers if:
 - Lower than $Q1 - 1.5 \times (IQR) = 70 - 1.5(10) = 55$; or
 - Higher than $Q3 + 1.5 \times (IQR) = 80 + 1.5(10) = 95$

Binning Numerical Variables

- Some algorithms require categorical predictors
- Continuous predictors are partitioned as bins or bands
 - Example: *House value* numerical variable partitioned into: *low, medium or high*
- Four common methods:

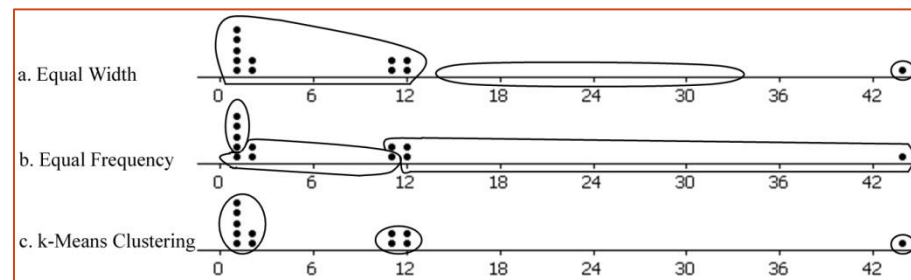
Method	Description	Notes
1. Equal width binning	Divides predictor into k categories of equal width, where k is chosen by client/analyst	Not recommended, since width of bins can be affected by presence of outliers
2. Equal frequency binning	Divides predictor into k categories, each having k/n records, where n is the total number of records	Assumes that each category is equally likely, which is not warranted
3. Binning by clustering	Uses clustering algorithm, like <i>k-means clustering</i> to automatically calculate “optimal” partitioning	Methods 3 and 4 are preferred
4. Binning based on predictive value	Methods 1 to 3 ignore the target variable; this method partitions numerical predictor based on the effect each partition has on the value of the target variable	

Binning Numerical Variables (*cont'd*)

- Example: Discretize $X = \{1,1,1,1,1,2,2,11,11,12,12,44\}$ into $k=3$ categories

Method	Low	Medium	High
a. Equal Width	$0 \leq X < 15$ Contains all values except one	$15 \leq X < 30$ Contains no data	$30 \leq X < 45$ Contains single outlier
b. Equal Frequency	First four data values $\{1,1,1,1\}$	Next four data values $\{1,2,2,11\}$	Last four data values $\{11,12,12,44\}$
c. k-means Clustering	$\{1,1,1,1,2,2\}$	$11,11,12,12$	$\{44\}$

- How is that in Equal Frequency, values $\{1,1,1,1,1\}$ are split into two categories? Equal values should belong to the same category
- As illustrated in image below, k-means clustering identifies apparently intuitive partitions



Reclassifying categorical variables

- Equivalent of binning numerical variables
- **Algorithms like Logistic Regression and C4.5 decision tree are suboptimal with too many categorical values**
- Used to reduce the number of values in a categorical field
- Example:
 - Variable *counties* {32 values} → Variable *region* {North, South, East, West}
 - Instead of 32 values, analyst/algorithm handle only 4 values
 - Alternatively, could convert *county* into *provinces*, with values {Munster, Leinster, Connacht, Ulster} or into *economic_level*, (poor_counties, rich_counties, midrange_counties)
- Data analyst should select reclassification that fits business/research problem

Removing variables that are not useful

- Some variables will not help the analysis
 - Unary variables – Take only a single value (a constant).
 - Example – In an all-girls private school, variable gender will always be female, thus not having any effect in the data mining algorithm
 - Variables which are very nearly unary – Some algorithms will treat these as unary. Analyst should consider whether removing.
 - Example - In a team with 99.9% females and 0.05% males, the variable gender is nearly unary.

Variables that should probably not be removed

Variables with 90% or more missing values

- Consider that there may be a pattern in missingness
- Imputation becomes challenging
- Example: Variable `donation_dollars` in self-reported survey
 - Top 10% donors might report donations, while others do not – the 10% is not representative
 - Preferable to construct a flag variable, *donation_flag*, since missingness might have predictive power
 - If there is reason to believe that 10% is representative, then proceed to imputation using regression or decision tree

Variables that should probably not be removed (*cont'd*)

Strongly correlated variables

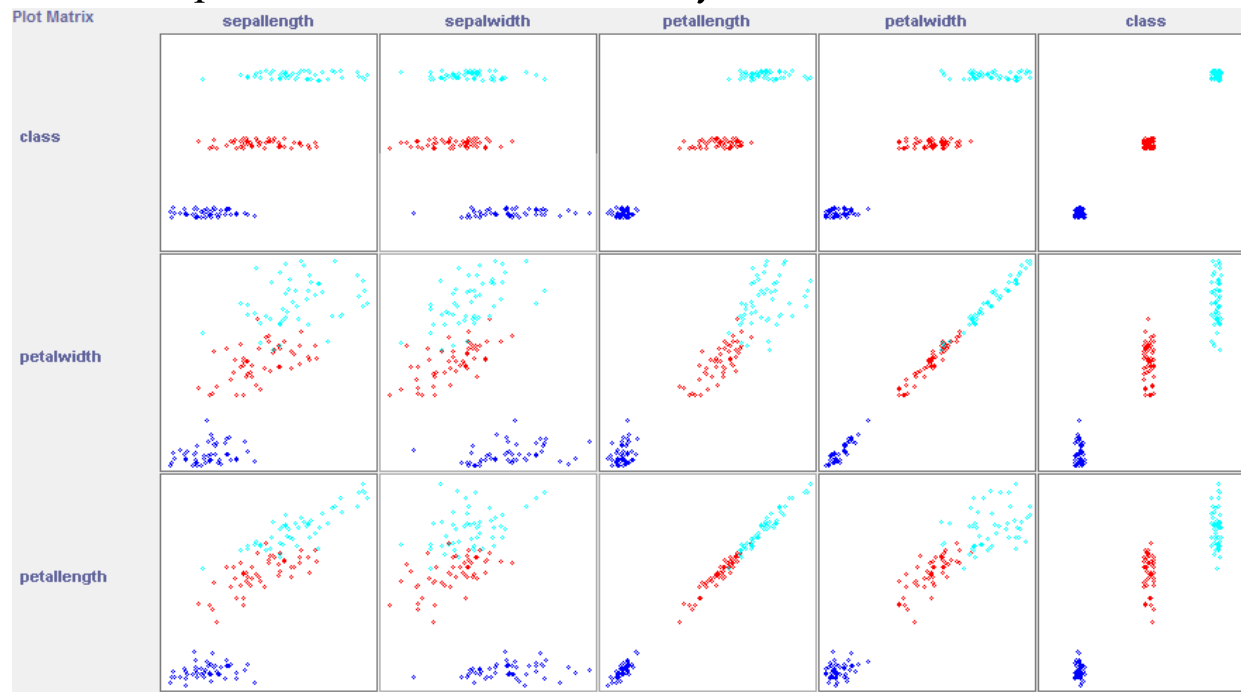
- Important information might be discarded when removing correlated variables
- Example: Variables *precipitation* and '*attendance at the beach*' are negatively correlated
 - They might double-count aspect of the analysis or cause instability in model results – prompting analyst to remove one variable
 - Should perform Principal Component analysis instead, to convert into a set of uncorrelated principal components

Removal of duplicate records

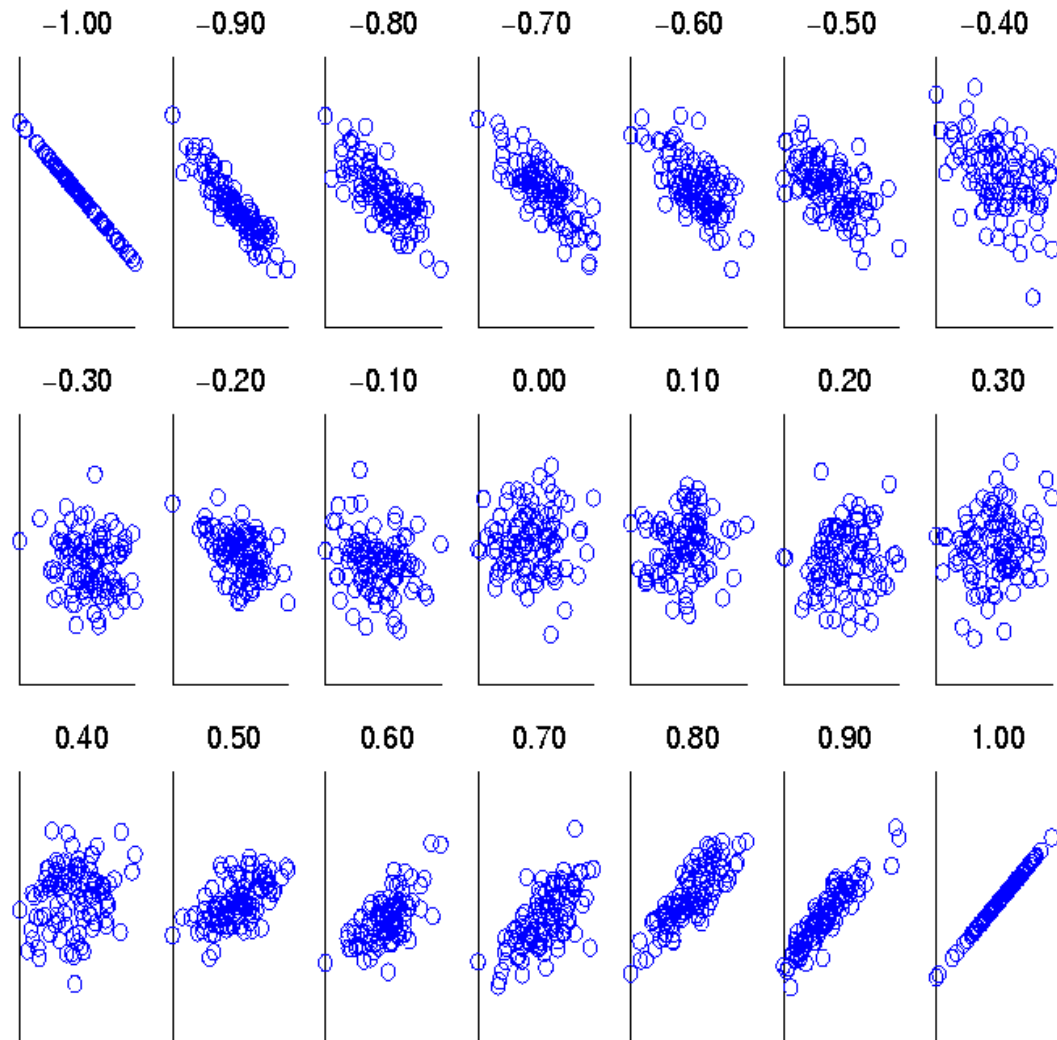
- Records might have been inadvertently copied, creating duplicates
 - Duplicate records lead to overweighting of their data values – therefore, they should be removed
- Example – If ID field is duplicated, then remove it
- But, consider genuine duplicates
 - When the number of records is higher than all possible combination of field values, there will be genuine duplicates

Dealing with Correlated Variables

- Using correlated variables in data model:
 - Should be avoided if they are perfectly correlated!
 - Incorrectly emphasizes one or more data inputs
 - May create model instability and produces unreliable results
 - Matrix plot of Iris Data below. Any correlations



Visually Evaluating Correlation



**Scatter plots
showing the
similarity from
-1 to 1.**

A word about ID fields

- ID fields have different value for each record
- Might be hurtful, with algorithm finding spurious relationships between ID field and target
- Recommendation: Filter ID fields from data mining algorithm, but do not remove them from the data, so that analyst can still differentiate the records