Data Preparation

Data-driven Healthcare - Module 2

Lecturer: Enayat Rajabi

Data Analytics and Service Innovation based on Artificial Intelligence (MAISTR) Programme

Spring 2022

Agenda

- Introduction to Healthcare Analytics
- Data Understanding
- Data Visualization
- Data Preparation
- Data Science with Python

Introduction to Healthcare Analytics

Healthcare Analytics

- In healthcare, professionals have access to vast amounts of data these days in the form of staff records, electronic patient records, clinical findings, diagnoses, prescription drugs, medical imaging procedures, mobile health, etc.
- Managing data to understand and analyze it to make well-informed decisions properly is a challenge for managers and healthcare professionals.
- More data analytics tools, introduced by large companies such as IBM and smaller companies such as Tableau and Qlik, are becoming more powerful, affordable, and easier to use.

Analytics

- Analytics involves using data, analysis and modelling to arrive at a solution to a problem or identify new opportunities via the decision-making progress.
- Data analytics can answer questions such as (1) what has happened in the past and why, referred to as descriptive analytics; (2) what could happen in the future and with what certainty, referred to as predictive analytics; and (3) what actions can be taken now to control events in the future.
- In the healthcare field, analytics can answer questions such as:
  - Is there a cancer present in this X-ray image?
  - How many nurses do we need during the upcoming holiday season, given the patient admission pattern we had last year and the number of patients with flu that we admitted the previous month?
  - How can we optimize the emergency department processes to reduce wait times?

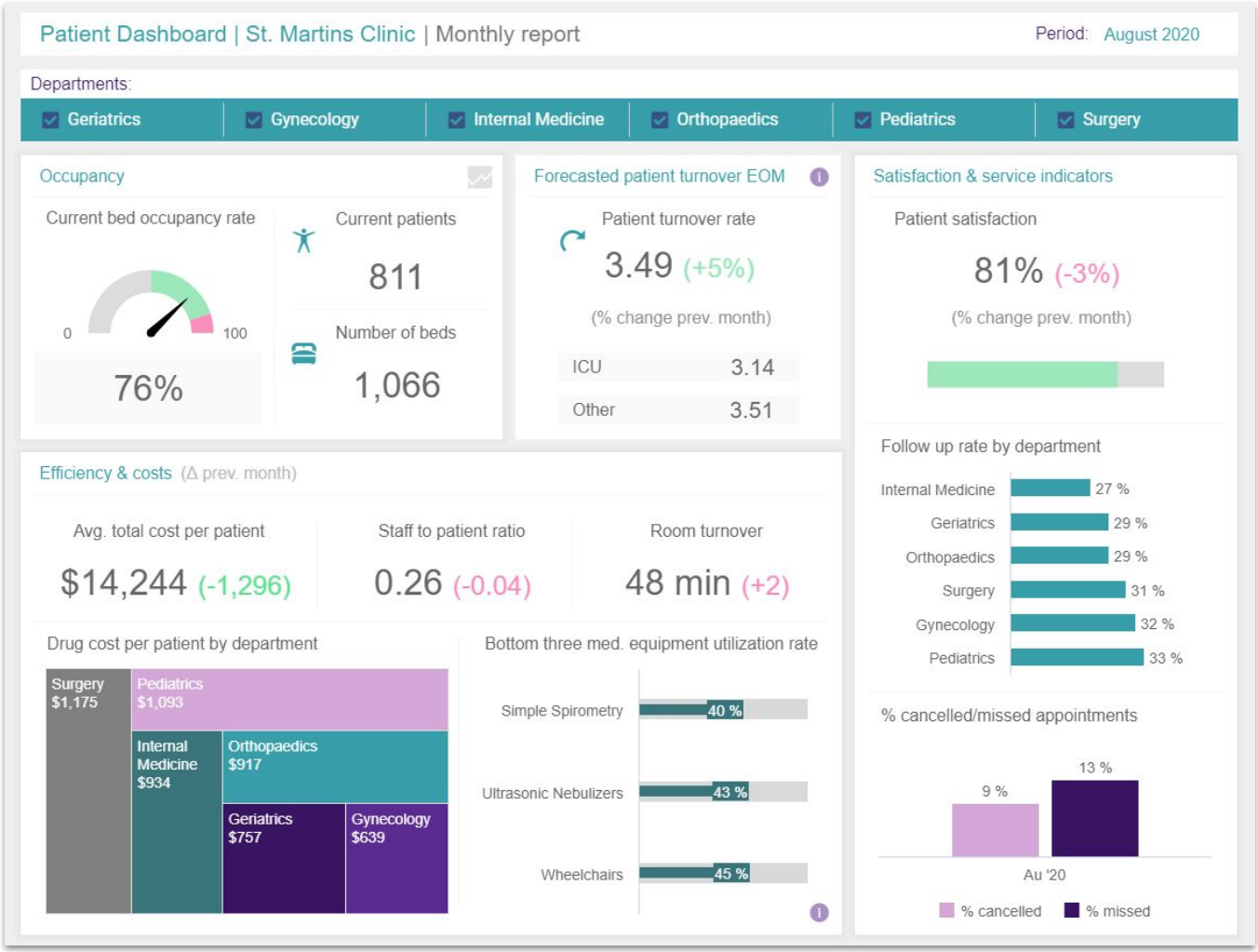
Decision Making in Healthcare

- From an analytics perspective, one can look at healthcare as a domain for decision-making:
  - A nurse or a doctor collects data about a patient (e.g., temperature, blood pressure), reviews an echocardiogram (ECG) screen, and then assesses the situation (i.e., processes the data) and decides on the next step to move the patient forward towards healing.
  - A radiologist who accesses a digital image (e.g., X-ray, ultrasound, uses the digital image processing tools available on her/his diagnostic workstation to make a diagnosis and reports the presence or absence of disease.
  - A committee might access admission data, operating room (OR) data, intensive care unit (ICU) data, financial data, or human resource data and use software to prescribe a reorganization of schedules for optimization.

Why Healthcare Analytics?

- Healthcare managers need to make sense of the data available using analytics and graphical dashboards to improve quality and performance.
- Hospitals may build many types of metrics to measure their performance and quality of care, for example, the hospital readmission rate within 30 days of discharge, the emergency department wait time, bed occupancy, the length of stay in the hospital, and the number of adverse drug events.
- An indicator allows managers to detect the current performance state and how far it is from a set target.
- Indicators can be consolidated on a screen using visualization tools such as figures, charts, colours, or numbers. These indicators displayed in a simple and easy-to-understand way are called a dashboard; dashboards display a snapshot of the "health" of an organization (e.g., a hospital).

A Healthcare Analytics Dashboard



Data Mining Process

The Cross-Industry Standard Process Model for Data Mining (CRISP-DM) describes the data mining process in six steps:

1- Business Understanding

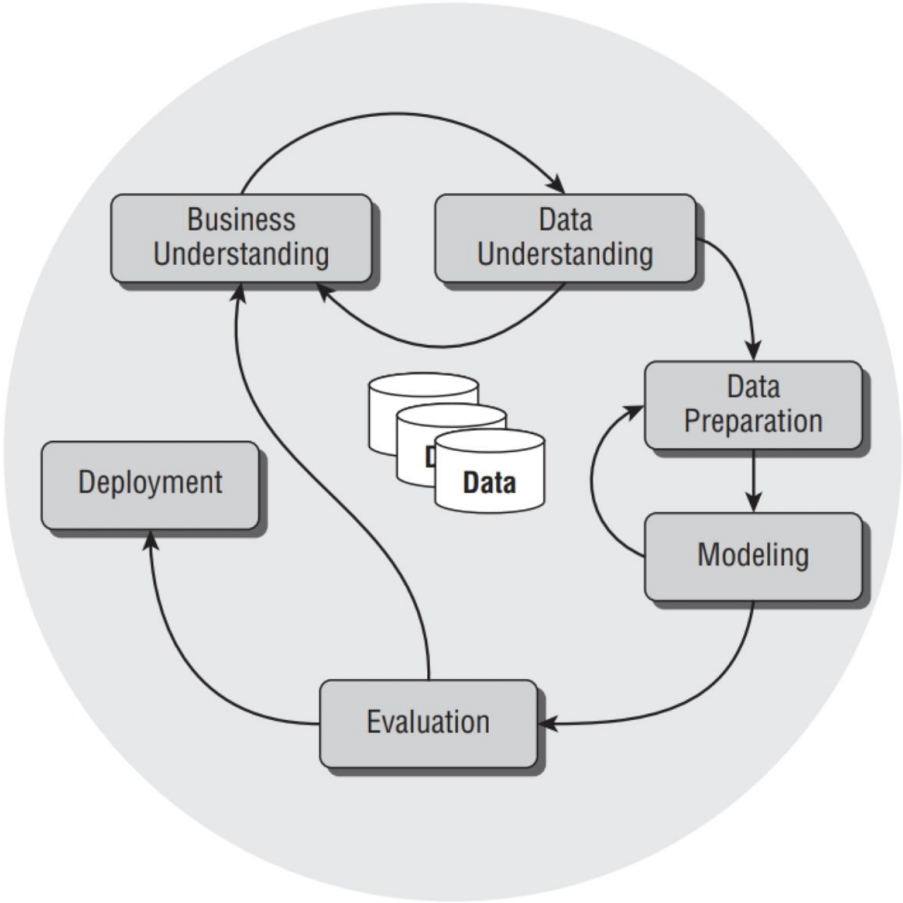
2- Data Understanding

3- Data Preparation

4- Modelling

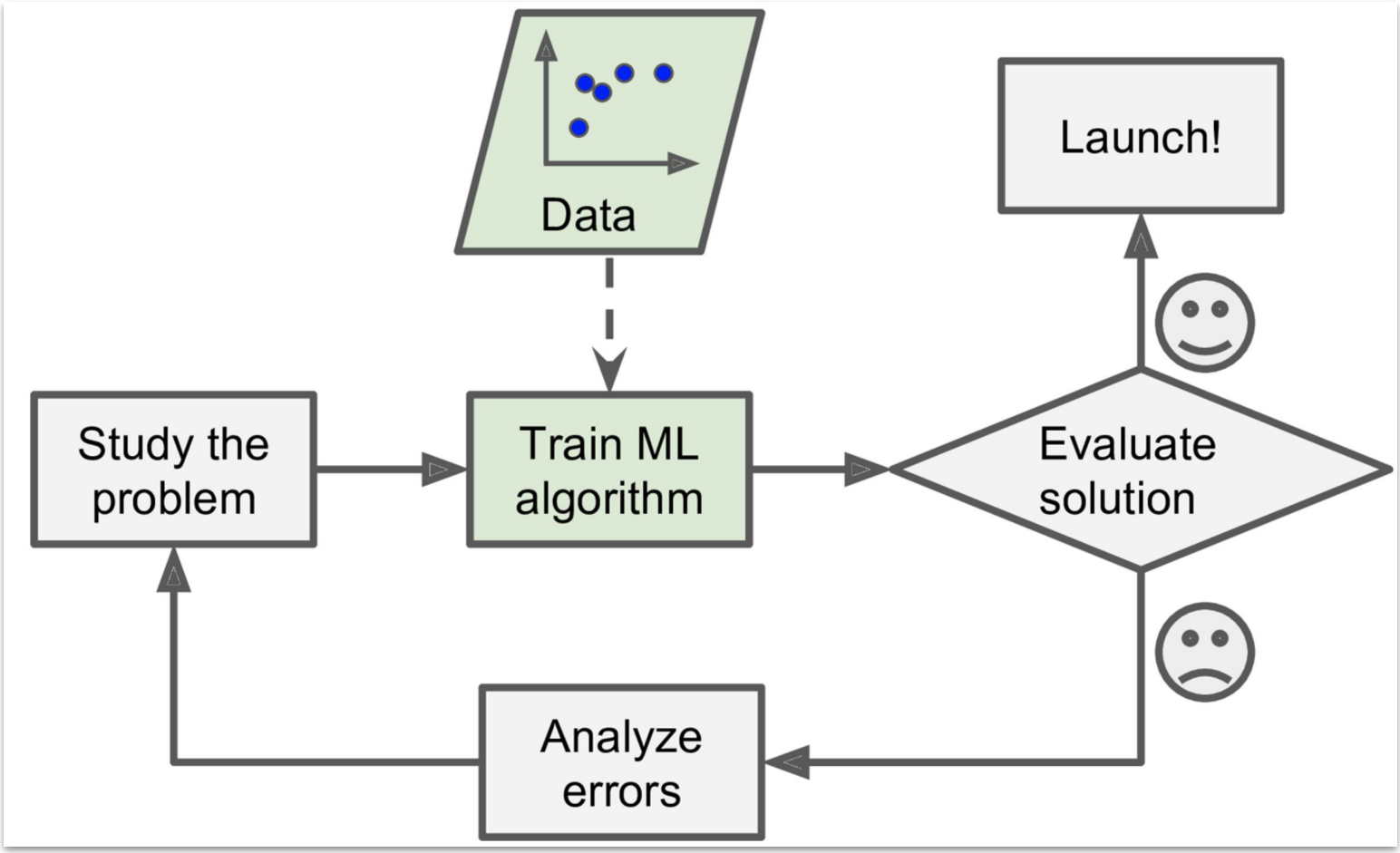
5- Evaluation

6- Deployment



The CRISP-DM process model

Typical Machine Learning Process



https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch01.html

Data Understanding

Data Understanding

- Data Understanding, as the first analytical step in data analytics, has the following purposes:
  - Examining key summary characteristics about the data to be used for modelling (records, variables, target variables)
  - Beginning enumerate problems with the data, including inaccurate or invalid values, missing values, unexpected distributions, and outliers.
  - Visualizing data to gain further insights into the characteristics of the data, especially those masked by summary statistics.

Variable Type

- Variables can be numeric, strings, and dates:
  - Dates are particularly cumbersome in software because so many different formats can be used.
  - Continuous variables are numbers that range from negative infinity to positive infinity. These numbers can be integers or real values, sometimes called ints and doubles, respectively (income, profit/loss, invoice amount).
  - Categorical variables have a limited number of values to label a variable rather than measure it.

Statistics Summary

- The simplest way to gain insight into variables is to assess them one at a time by calculating summary statistics, including the mean, standard deviation, skewness, and kurtosis.

- This figure shows a summary of statistics of a dataset in Python

→

	age	bmi	children	expenses
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.665471	1.094918	13270.422414
std	14.049960	6.098382	1.205493	12110.011240
min	18.000000	16.000000	0.000000	1121.870000
25%	27.000000	26.300000	0.000000	4740.287500
50%	39.000000	30.400000	1.000000	9382.030000
75%	51.000000	34.700000	2.000000	16639.915000
max	64.000000	53.100000	5.000000	63770.430000

Measures for Numeric Data

- **Measures of centre:**
  - Mean, Median, Mode
- **Measures of variation:**
  - Range, Variance
  - Quartile, Percentile, Interquartile range
  - Standard deviation
- **Shape**
  - Skew, Symmetric, Kurtosis

Mean

- The mean of a distribution is its average, simply the sum of all values for the variable divided by the count of how many values the variable has. It is sometimes represented by the Greek symbol mu, μ.
- The mean value is often understood to represent the middle of the distribution or a typical value. This is true when variables match a normal or uniform distribution, but often this is not the case.
- Example: (12, 14, 19) → Mean= (12+14+19)/3=15

Median

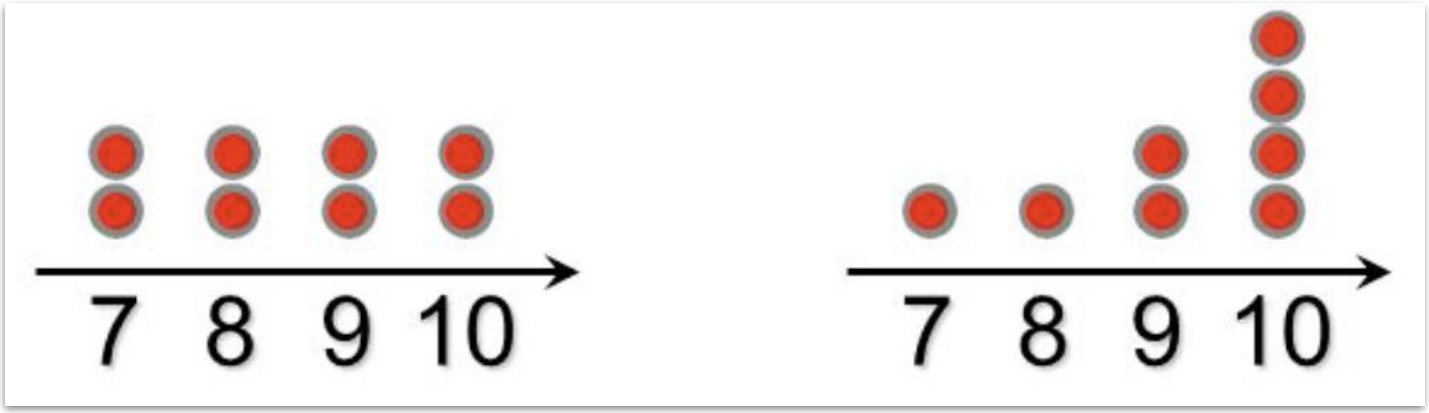
- Middle value in ordered sequence
  - If Odd n, middle value of sequence
  - If Even n, average of two middle values
- Position of median in the sequence
  - Positioning point = (n+1)/2
- Examples:
  - (12, 14, 16, 17, 19) → Median= a[(n+1)/2]=a[3]=16
  - (12, 14, 16, 17, 19, 20) → Median= (a[3]+a[4])/2=16.5

Mode

- Mode is the value that occurs most often.
- A variable maybe has no mode or several modes.
- Examples:
  - (12, 14, 16, 17, 14, 20) → Mode= 14
  - (12, 14, 16, 17, 19, 20) → No Mode
  - (12, 14, 12, 17, 14, 20) → Modes= 12, 14

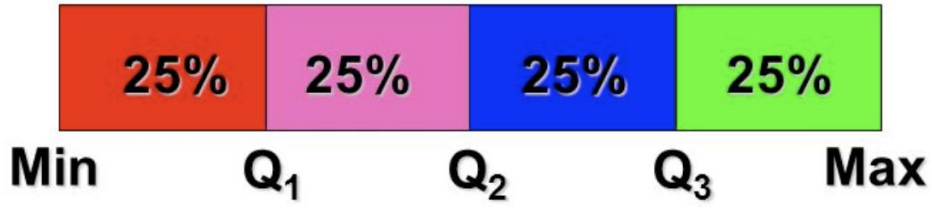
Measure of Variance: Range

- Range is a measure of spread and variation.
- Range is difference between largest and smallest observations.
  - Range = Largest (X_i) - Smallest(X_i)
- Range ignores how data are distributed.



Measure of Variance: Quartile

- The spread ordered dataset into four quarters, which is called quartiles. Quartiles require the data first be sorted.



- Example:
  - Raw data: 10, 4, 8, 19, 22, 12, 9, 33
  - Ordered: 4, 8, 9, 10, 12, 19, 22, 33
  - Min: 4, Q_1: 8.75, Q_2: 11, Q_3: 19.75, Max: 33

Measure of Variance (cont.)

- The consensus of predictive analytics software is to use quartile, quintile, decile, and percentile.

QUANTILE LABEL	NUMBER OF BINS	PERCENT OF POPULATION
Quartile	4	25
Quintile	5	20
Decile	10	10
Demi-decile Vingtile Twentile	20	5
Percentile	100	1

Measure of Variance: Percentile

- There are 99 percentiles, denoted as $P_1, P_2, \ldots, P_{99}$.
- The K^{th} percentile is the value in which k% of all observations are below that value. For example, $Q_1=P_{25}$, $Q_2=P_{50}$, $Q_3=P_{75}$.
- Percentile of value X is the number of values less than x over the total number of observations multiplied by 100.

PERCENTILE	METRIC
0th	Minimum
25th	1st quartile
50th	2nd quartile; the median
75th	3rd quartile
100th	Maximum

Percentile Example

Original Data

Best Actresses

22	37	28	63	32	26	31	27	27	28
30	26	29	24	38	25	29	41	30	35
35	33	29	38	54	24	25	46	41	28
40	39	29	27	31	38	29	25	35	60
43	35	34	34	27	37	42	41	36	32
41	33	31	74	33	50	38	61	21	41
26	80	42	29	33	35	45	49	39	34
26	25	33	35	35	28				

Sorted Data

Sorted Ages of 76 Best Actresses

21	22	24	24	25	25	25	25	26	26
26	26	27	27	27	27	28	28	28	28
29	29	29	29	29	29	30	30	31	31
31	32	32	33	33	33	33	33	34	34
34	35	35	35	35	35	35	35	36	37
37	38	38	38	38	39	39	40	41	41
41	41	41	42	42	43	45	46	49	50
54	60	61	63	74	80				

$$\text{Percentile of value } 30 = \frac{\text{number of values less than } 30}{\text{total number of values}} \times 100$$

$$= \frac{26}{76} \times 100 = 34\%$$

Interpretation: The age of 30 years is the 34th percentile, that is, $P_{34} = 30$

Measure of Variance: IQR

- The analogous measure to standard deviation using quartiles is the Inter-Quartile Range (IQR) which is the difference between the 3rd quartile value and the 1st quartile value, or in other words, the range between the 25th and 75th percentiles.
- A rule of thumb: values more than 1.5 times the IQR from the upper or lower quartiles (lower/upper limit) are considered outliers.

Example →

```
ages= [12, 34, 43, 45, 56, 67, 69, 70, 71, 71, 76, 100, 230]
ages_df=pd.DataFrame(ages,columns=['age'])

q75, q25 = np.percentile(ages, [75 ,25])
iqr = q75 - q25
ages_df[ages_df['age']>(1.5*iqr+q75)]
```

	age
12	230

Standard Deviation

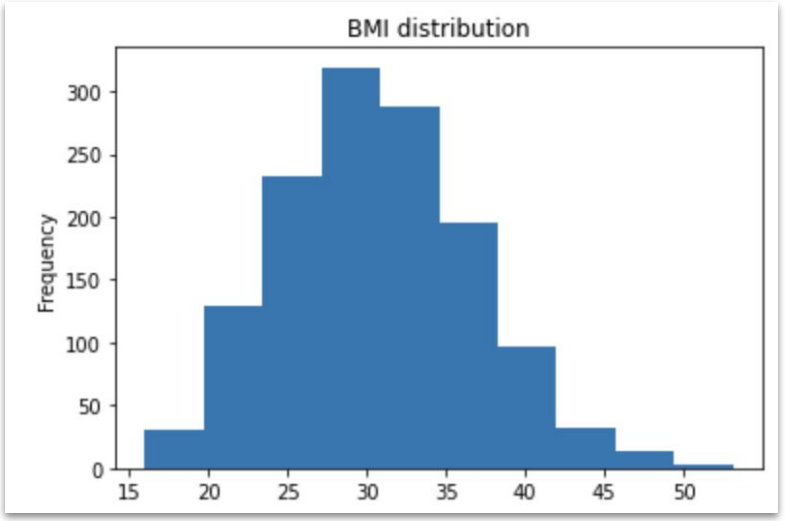
- Standard deviation measures the distribution spread; a larger standard deviation means the distribution of values for the variable has a greater range.
- Most often, in predictive analytics, the standard deviation is considered in the context of normal distributions. The Greek symbol sigma, σ, is often used to denote the standard deviation.
- The deviations display the spread of the values Xi about their mean X. Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean.

Standard Deviation (cont.)

- The variance is the average squared deviation.
- s^2 and s will be large if the observations are widely spread about their mean, and small if they are all close to the mean.
- 's' measures spread about the mean and should be used only when chosen as the center measure.
- s = 0 only when there is no spread and all observations have the same value. Otherwise, s > 0. As the observations become more spread out about their mean, s gets larger.
- 's' is not resistant. A few outliers can make it very large.

Data Shape: Normal Distribution

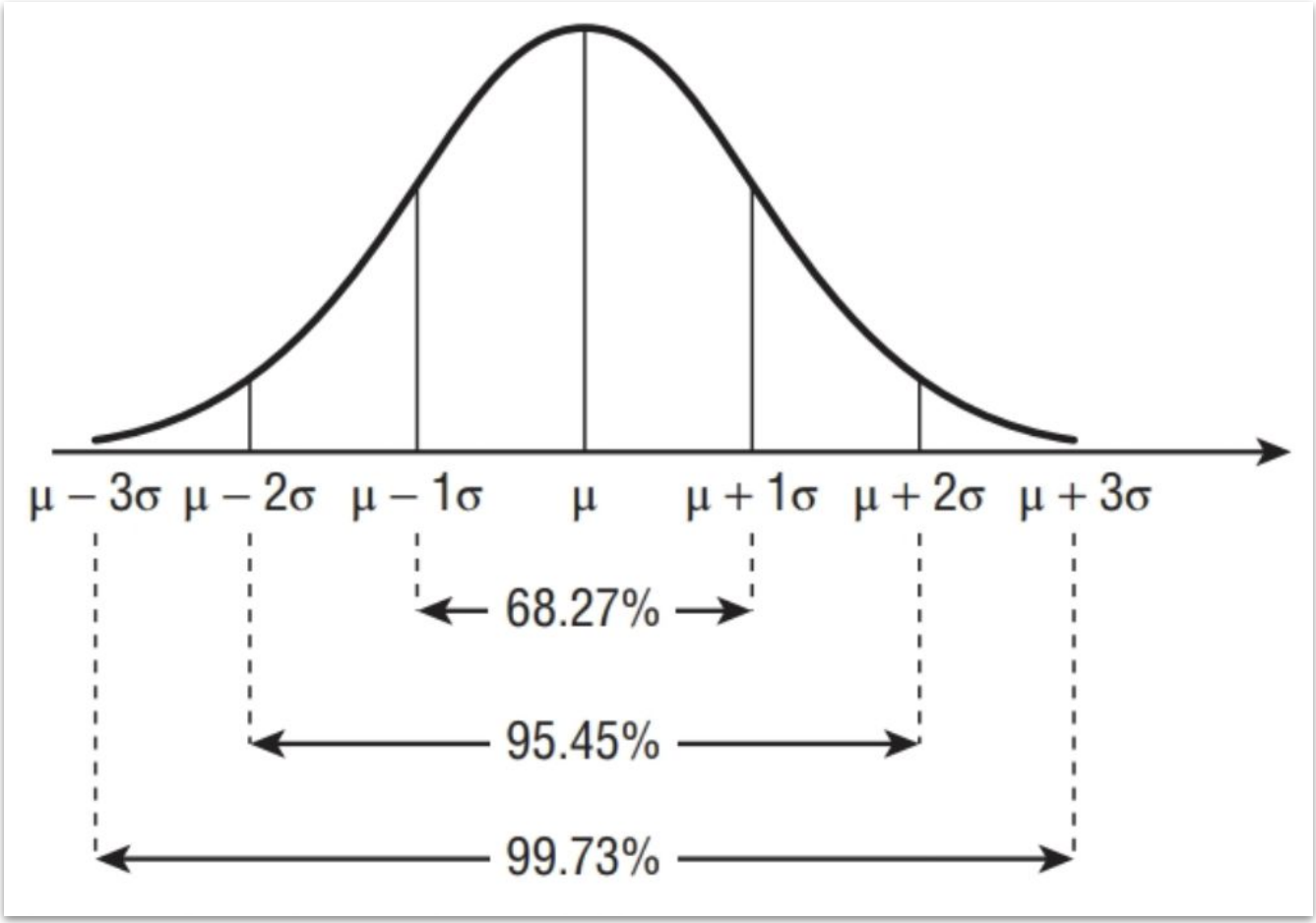
- It is also called bell curve or Gaussian distribution.
- Many algorithms assume normal distributions either explicitly or implicitly.
- Some analysts expend considerable effort to find transformations that change the variables so that they become normally distributed.



Normal Distribution

- The normal distribution is symmetric.
- The mean value is the most likely value to occur in the distribution.
- The mean, median, and mode are all the same value.
- Approximately 68 percent of the data will fall between the mean and +/–1 standard deviation from the mean.
- Approximately 95 percent of the data will fall between the mean and +/–2 standard deviations from the mean.
- Approximately 99.7 percent of the data will fall between the mean and +/– 3 standard deviations from the mean.

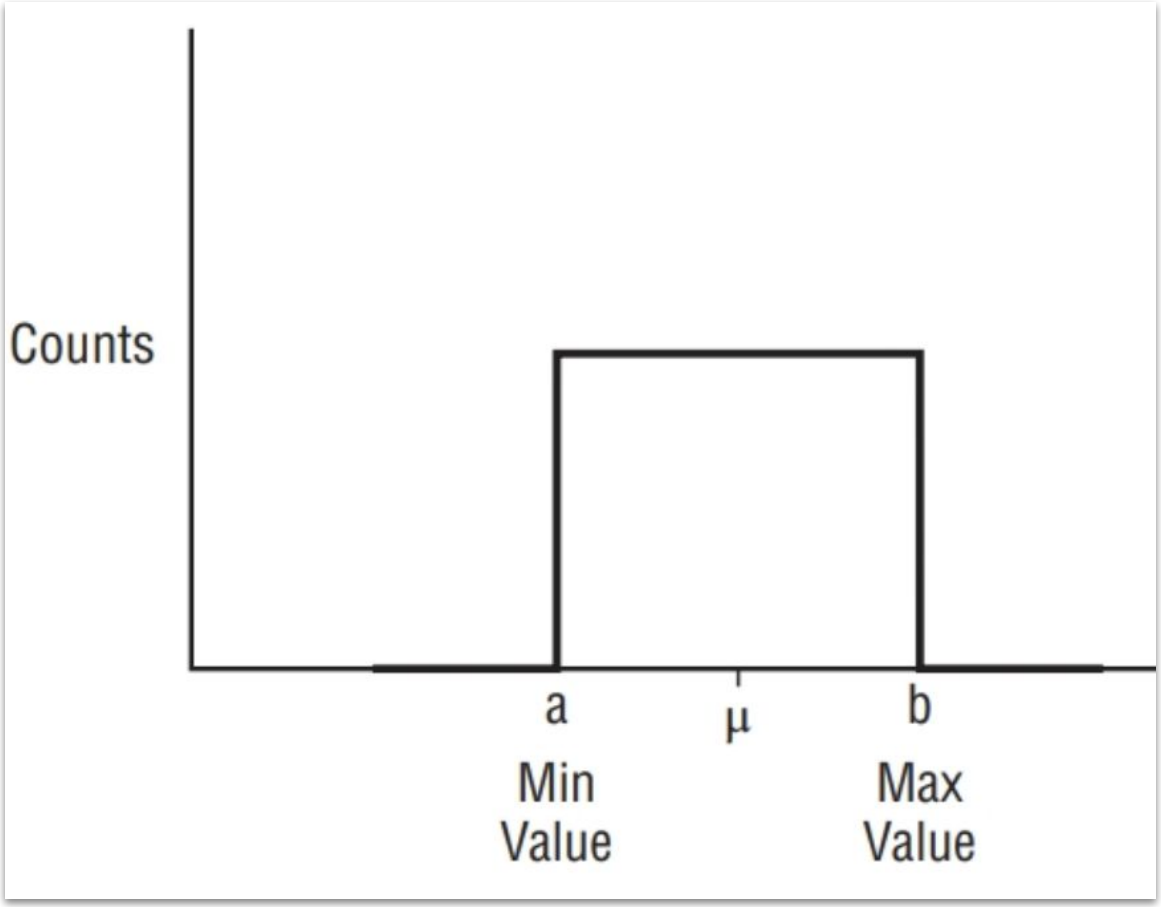
Normal Distribution (cont.)



Uniform Distribution

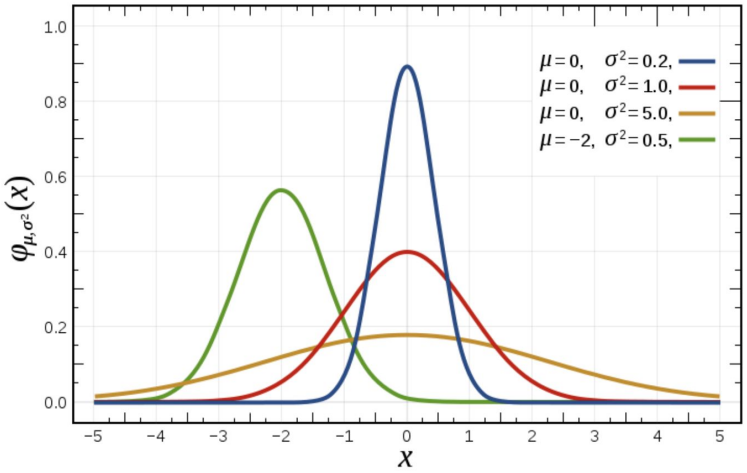
- The uniform distribution assumes that a variable has a fixed, finite range.
- The distribution is symmetric about the mean.
- The distribution is finite, with a maximum and minimum value.
- The mean and midpoint of the distribution are the same value.
- Random number generators create uniform random distributions, most often in the range 0 to 1 as their default.
- Standard deviations can be computed but are rarely used in predictive modelling algorithms.

Uniform Distribution (cont.)



Basic Statistics for Distribution Analysis

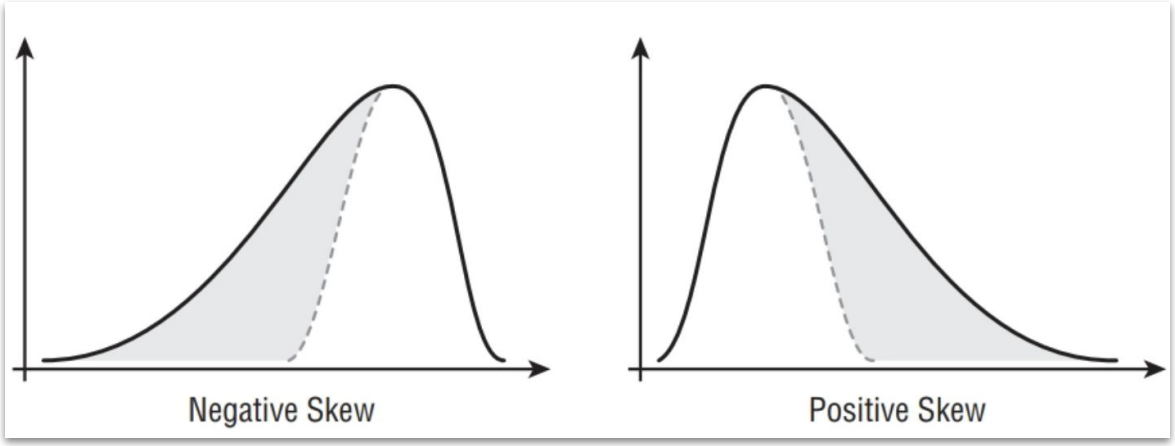
- Mean, variance, skewness, and kurtosis are important quantities in statistics.
- Some of the calculations involve sums of squares, which for large values may lead to overflow.
- To avoid loss of precision, we have to realize that variance is invariant under shift by a certain constant number.



picture from here

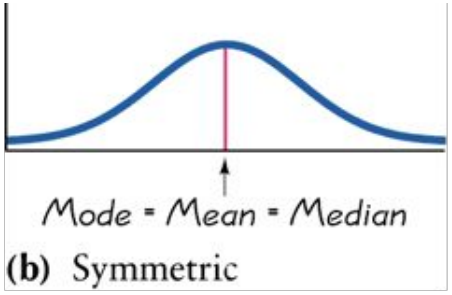
Data Shape: Skewness

- Skewness measures how balanced the distribution is. A normal distribution has a skewness value of 0.
- Positive skew: skew values greater than zero—indicates that the distribution has a tail to the right of the main body of the distribution.
- Negative skew indicates the converse: The distribution has a tail to the left of the main body.

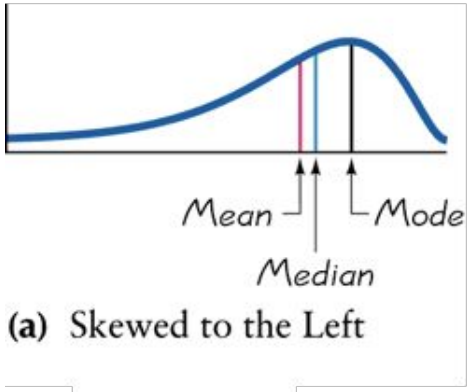


Skewness (cont.)

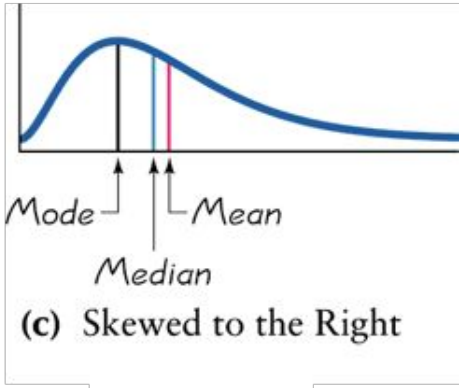
- Some algorithms assume variables have normal distributions, and significant deviations from those assumptions affect the model accuracy or at least the model interpretation.
- In positive skew, the mean is larger than the median. Also, the mode is less than the median, which is also less than the mean.



(b) Symmetric



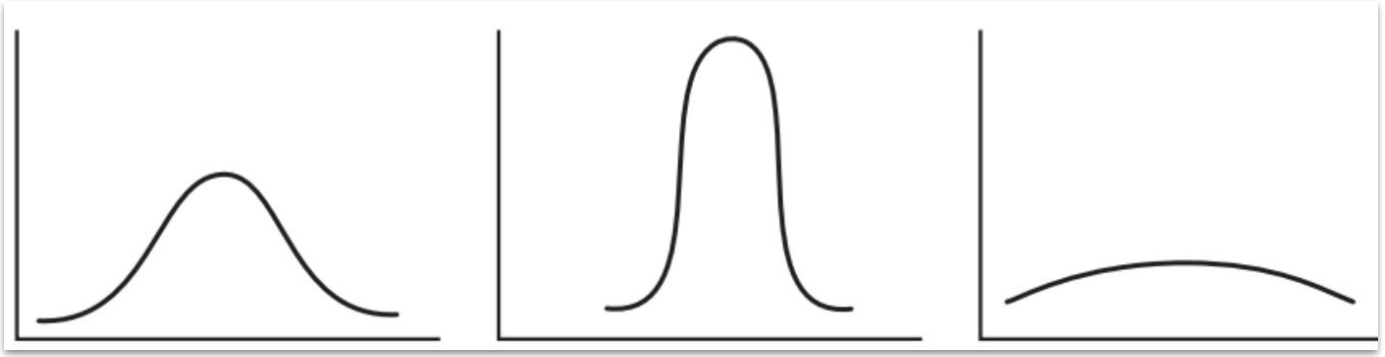
(a) Skewed to the Left



(c) Skewed to the Right

Data Shape: Kurtosis

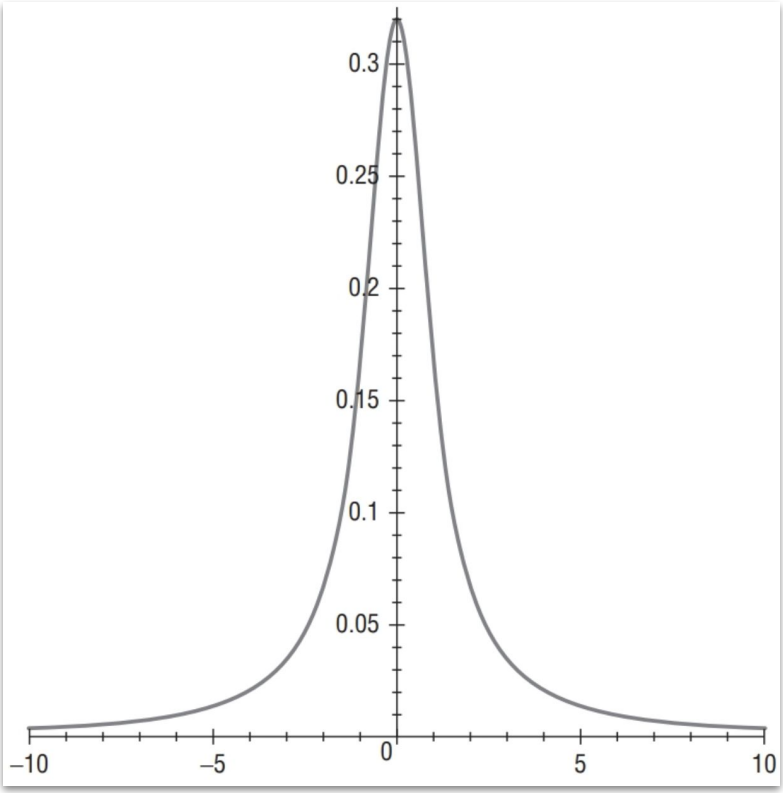
- Kurtosis measures how much thinner or fatter the distribution is compared to the normal distributions.
- Normal distribution, Skinnier-than-normal distribution (leptokurtic); and fatter-than-normal distribution (called platykurtic).



- In a normal distribution, kurtosis is equal to 3. While in leptokurtic distribution, it is less than 3, and a platykurtic distribution is greater than 3.

Kurtosis (cont.)

- The skew will be 0 (the tails are symmetric).
- The kurtosis is very large (platykurtic), because of the large tails.
- The variable's standard deviation will not represent the true spread in the data for algorithms that use standard deviation or variance in the model.



Categorical Variables

- Categorical or nominal variables are assessed by counting the number of occurrences for every level.
- There are several issues typically addressed with frequency counts.
  - Do the values make sense? Are there values that are unexpected, erroneous, or indicators of missing or unknown values?
  - Are there missing values? How many? How are they coded?
  - How many levels are there? Is there only one value? Are there more than 50 or 100 levels?
  - What is the mode of the levels?

Hypothesis Testing, Alpha Levels, Type-I and Type-II Errors

- Statistics are often used to test theories or predictions, such as that smoking is associated with lung cancer.
- In general, this is done by inference testing, which is drawing conclusions about a population of interest based on findings from a sample obtained from that population.
- The specific claim or statement that we wish to test is called a research hypothesis.
- Example: There is a link between smoking and lung cancer.
  - Independent variable (the first variable): smoking,
  - Dependent variable (the second variable): lung cancer status (since we hypothesize that its values depend on smoking).

Hypothesis Testing

- The claim that there is no link between smoking and lung cancer is called the null hypothesis and is denoted as H0. The alternative hypothesis or research hypothesis is denoted by H1.
- When testing the hypothesis (also referred to as statistical inference or significance testing), we assume that the null hypothesis is true, and try to refute it.
- If the null hypothesis is rejected after statistical analysis (for example using a t-test or correlation covered later in this chapter), then we can draw a conclusion that the association between lung cancer and smoking is significant.

Hypothesis Testing (cont.)

- When we statistically test a hypothesis, we can accept a certain level of significance known as α (alpha).
- When we say that a finding is "statistically significant,", it means that the finding is unlikely to have occurred by chance and that the level of significance is the maximum chance that we are willing to accept.
- A very common threshold for the level of significance is 0.05 or 5%, with 0.01 or 1% considered marginally significant.

Type-I and Type-II Errors

- Two types of errors may result from hypothesis testing: Type-I and Type-II errors.
- Type-I error occurs when we reject the null hypothesis (for example, we conclude that there is a significant association between two variables) while in fact the null hypothesis is true (there is no significant association or difference between the variables).
- Type-II error occurs when we do not reject the null hypothesis when in fact it is false.
- If a type-I error is costly, meaning your belief that your theory is correct when it is not could be problematic, then you should choose a low value for it to avoid that error.

Type-I and Type-II Examples

- For example, the blood pressures of a group that took a new drug are significantly lower than those of a group who took placebo would be considered "costly" (risky for patients), and a low α should be adopted.
- A common value used for α in this case is 0.01 or 1%. If a type-II error is costly, then you should choose a higher value for α to avoid that error, such as 0.1 or 10%.

Statistical Significance and P-Values

- To assess the level of significance of our statistical test (t-test, chi-square, correlation, etc.), we depend on an outcome called the p-value.
- A p-value is generated by default with different statistical tests.
- We form a decision rule for our hypothesis testing depending on the p-value; if the p-value is less than our selected level of significance α, then we cannot accept (i.e., we reject) the null hypothesis, and we must conclude that our alternative hypothesis is true.
- The lower the p-value is, the greater the significance of our finding is.

Correlation

- A correlation is a test used when both independent and dependent variables have continuous values.
- In its simplest terms, a linear correlation represents the degree to which a straight line describes the relationship between two variables, such as height and weight.
- A degree or coefficient of correlation ranges from +1 (strong positive correlation or relationship between the two variables) to −1 (strong negative correlation). Values close to zero indicate a weak relationship between the two variables.

Chi-Square Correlation

- Chi-square is a test used when both independent and dependent variables have categorical values.
- Chi-square is used to evaluate if there are significant associations between a given exposure (independent variable) and outcome (dependent variable).
- Commonly, a 2x2 table is used to present categorical data where, for example, a column represents exposure or not to a chemical (yes/no) and a row represents a disease or health outcome (yes/no).

Test Differences

- t-test is most commonly used to test the difference between the two groups' means of the dependent variable (i.e., outcome variable).
- t-test is used when one of the variables of interest is continuous (systolic blood pressure) and the other is dichotomous, i.e., nominal with only two values (taking the drug or not).

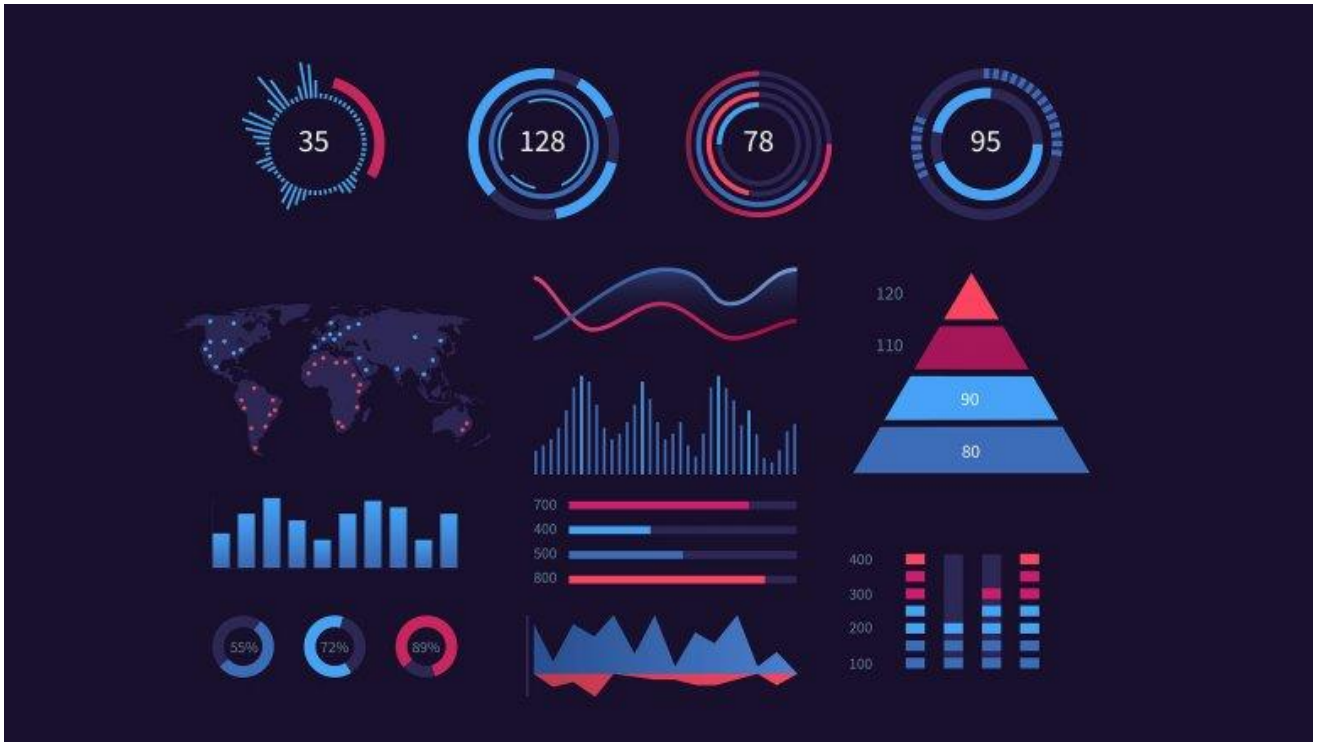
Test Differences: Example

- A new drug is administered to group A of individuals, while group B receives a placebo. The null hypothesis would be no difference in the mean systolic blood pressures of the experimental group A and the placebo group B.
- We select an α of 5%, for example, and test the data, which consists of the systolic blood pressures of the individuals.
- Suppose the p-value is less than the selected α of 0.05. In that case, we reject the null hypothesis and conclude that there is a difference in the mean between the two groups and that the new drug was effective (with a 5% level of significance or risk of type-I error).
- If we test, however, an outcome or dependent variable for the same sample at two different times, or if we match pairs of unrelated individuals (for example, having closely matched behavioural or physiological characteristics that are relevant to the outcome variable), then we call it a dependent or paired-samples t-test.

ANOVA

- Analysis of Variance, or ANOVA, is similar to the t-test, but is used when we want to compare more than two groups at a time.
- ANOVA is used when one of the variables of interest is continuous and the other is nominal with more than two values.
- One-way ANOVA examines the effect of one independent variable with comparison to three or more groups, called between subjects ANOVA, or the same group of subjects at different points of time, called repeated measures ANOVA.
- For example, to test the effect of a new anti-depression drug, the depression levels of a group of patients are measured before and at several points during the treatment.

Data Visualization



Anscombe's Quartet: Statistics can be deceiving

- Example: Anscombe's Quartet, created in 1973 by the statistician Francis Anscombe.

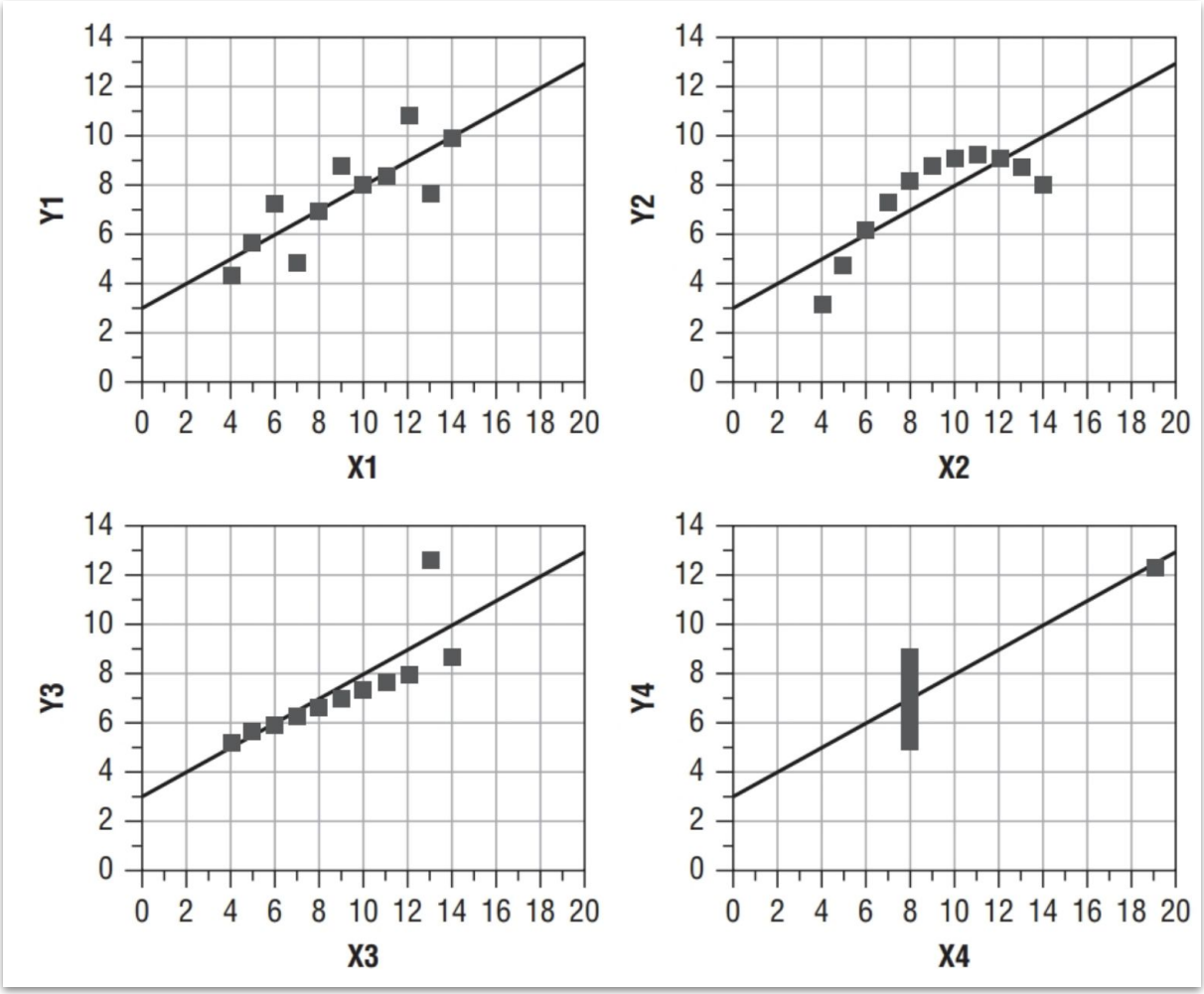
SET 1		SET 2		SET 3		SET 4	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Anscombe's Quartet (cont.)

- Interestingly, each of the four pairs of data points, has the following properties:

MEASURE	VALUE
Mean X	9.0
Variance X	11.0
Mean Y	7.50
Variance Y	4.12 to 4.13
Correlation X vs. Y	0.816 to 0.817
Regression Equation	Y = 3.0 + 0.500*X
R^2 for fit	0.666 to 0.667

Anscombe's Quartet (cont.)



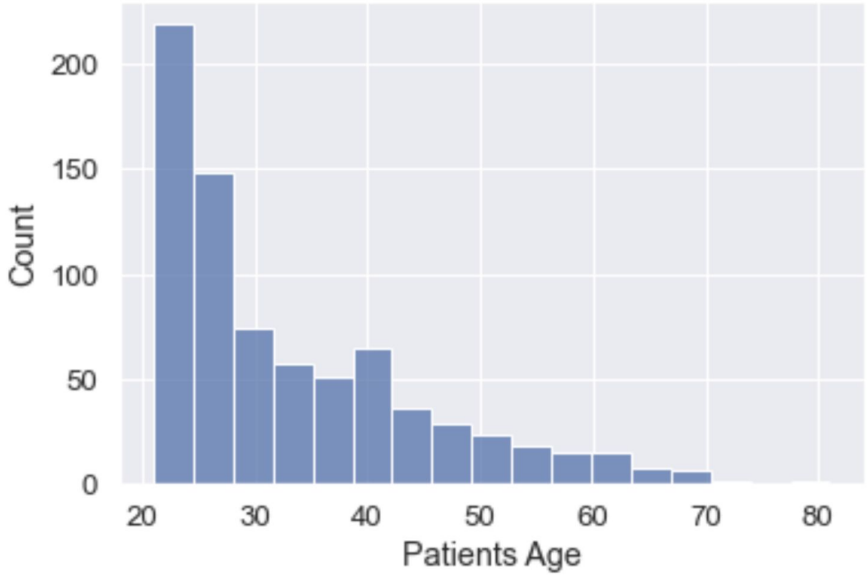
Data Visualization

- Data visualization is a graphical representation of the data for different purposes, for example:
  - seeing distances between some data
  - revealing if the data is normal, uniform, or neither
  - identifying high skew or excess kurtosis

- Visualization is done primarily in two stages of a modelling project:
  - in the beginning of a project during the Data Understanding stage
  - after building the models to understand why the most important variables in the models were effective in predicting the target variable

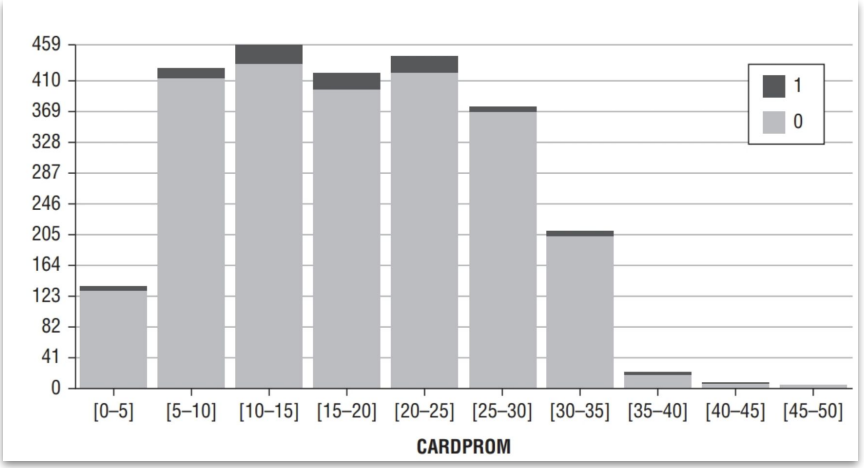
Histograms

- Histograms operate on a single variable and show an estimate of the shape of the distribution by creating bins of the data and counting how many records of the variable fall within the bin boundaries.
- The x-axis typically contains equal-width bins of the variable, and the y-axis contains the count or percentage of count of records in the bins.



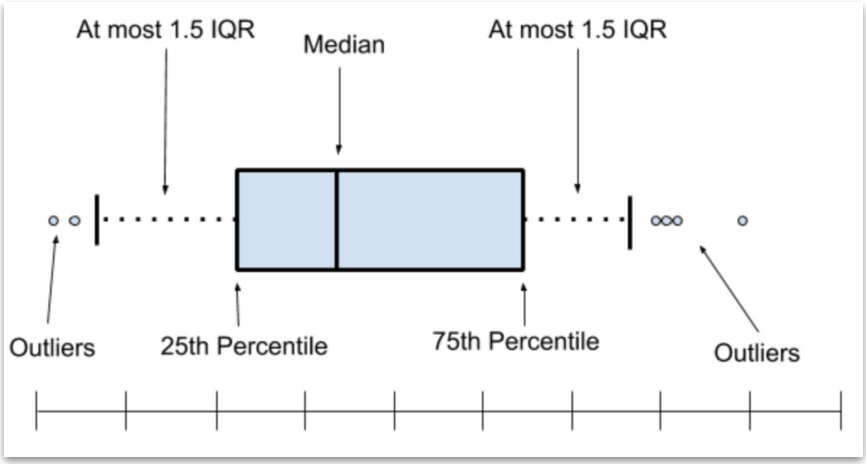
Histograms (cont.)

- Histograms can be customized by:
  - the number of bins: identifying data shape (some software allow changing the number of bins and visualizing the graph dynamically)
  - change the y-axis scale: when a few bins have large bin counts—spikes in the histogram
  - overlay a second categorical variable

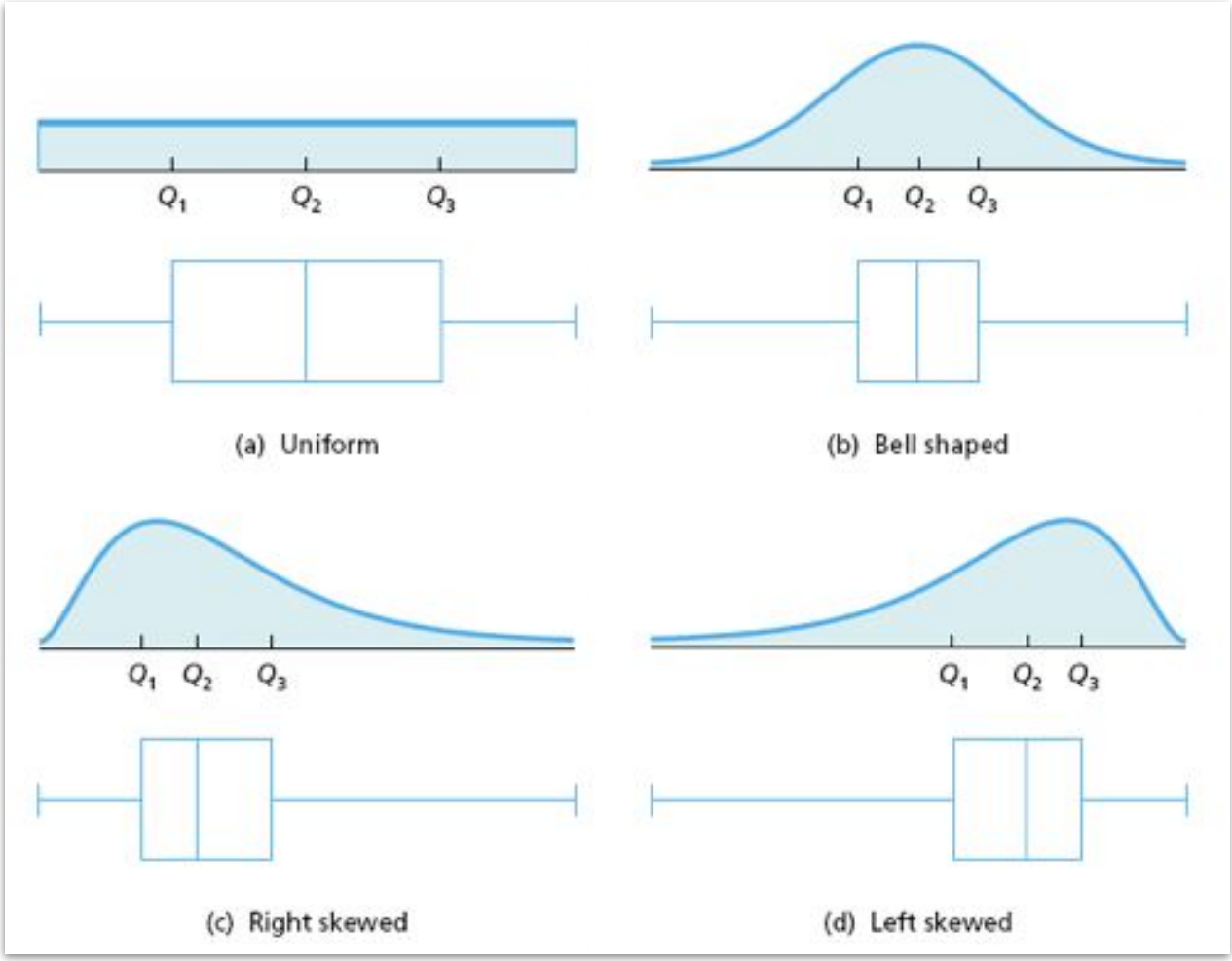


Box plot

- The box plot is a graphical representation of the quartile statistics of a variable and is an excellent method to gain quick insight into the characteristics of numeric data.
- In the box plot, the "box" is the IQR; the median is typically represented by a line cutting the IQR box in half.
- An upper quartile outlier boundary is +1.5 x IQR from the 3rd quartile and the lower quartile outlier boundary is –1.5 x IQR from the 1st quartile.



Boxplot and Distribution Shape



Correlations

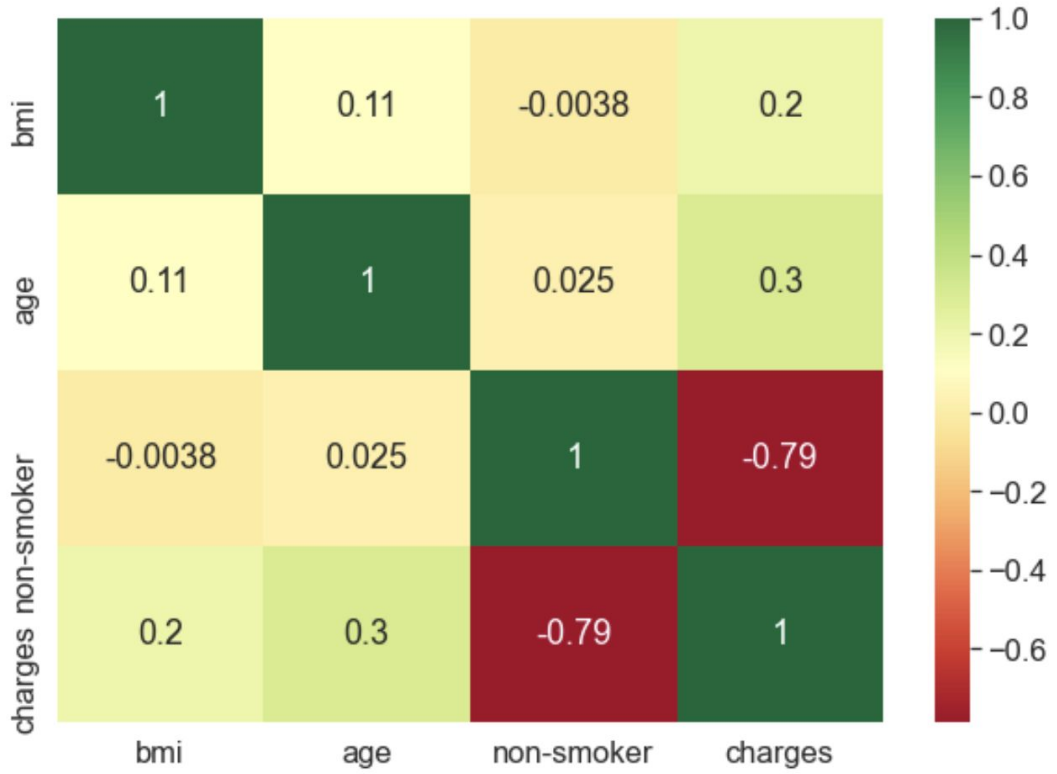
- Correlations measure the numerical relationship of one variable to another, a useful way to identify variables in a modeling data set that are related to each other.
- "Correlation does not imply causation". Just because two variables are numerically related to one another doesn't mean they have any informational relationship.
- Finding pairs of variables with high correlation indicates there is redundancy in the data, which is not good, as:
  - redundant variables don't provide new, additional information to predict a target variable.
  - some algorithms can be harmed numerically by including high correlated variables.

Correlations (cont.)

- Perfect negative correlation indicates the slope is negative and constant. A correlation value of 0 indicates there is no relationship between the two variables; they appear to be random with respect to one another.
- Correlations are affected severely by outliers and skewness. Even a few outlier value pairs can change the correlation between them from something small, like 0.2, to something large, like 0.9. Visualization is a good idea to verify that high correlation is real.

Correlations (cont.)

- We can use heat maps to visualize the correlations between variables in a dataset.



Crosstabs

- Crosstabs, short for cross-tabulations, are counts of the intersection between two variables, typically categorical variables.
- Crosstabs are a powerful and easy-to-use tool provided by many data analytics tools to understand your data in a visual form.
- It is particularly useful for conducting timely analysis and quality assurance testing on record distributions.

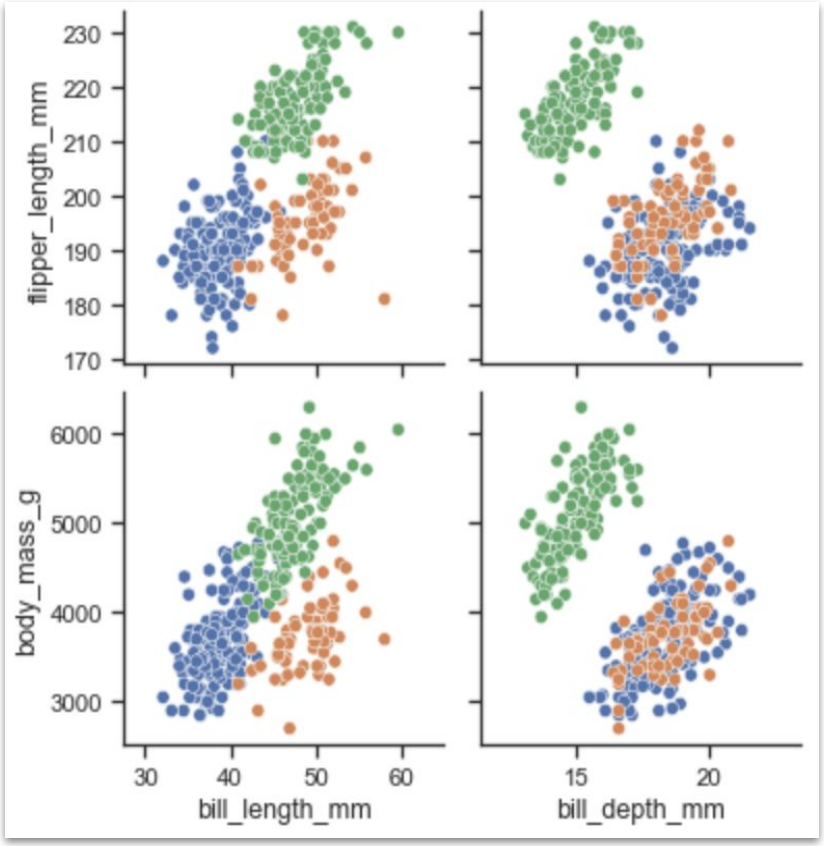
Scatter plot

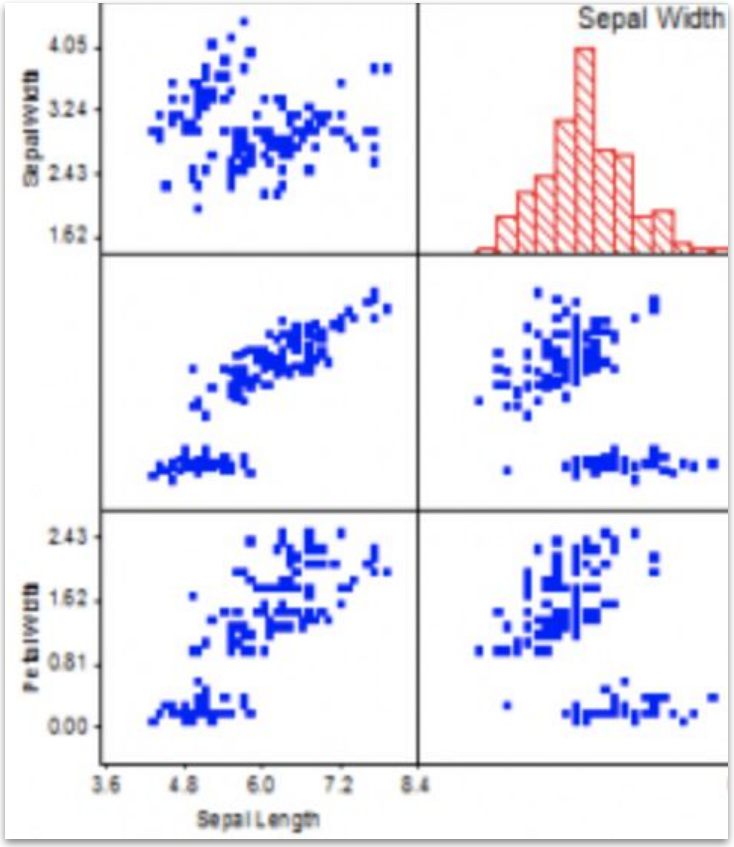
- Scatter plots are the most commonly used two-dimensional visualization method.
- Scatter plots are effective ways to show visually how correlated pairs of variables are: The more cigar-shaped the distribution and the narrower that cigar is, the stronger the correlation.
- Too many plotted data points make the visualization difficult to see.
- When there is severe skew or severe outliers in the data, the data is sparse in the distribution's tail. We can scale the data before plotting it or change the scale of the plot axes.

Scatter plot Matrices

- Just like a correlation matrix shows all pairwise correlations in the data, the scatter plot matrix shows all pairwise scatter plots for variables included in the analysis.
- They provide a good way to show several relationships between pairs of variables in one plot.
- Scatter plots take significant amounts of space to display.
- They require significant resources to draw, so typically, the number of variables included in the scatter plot is kept to less than 10.

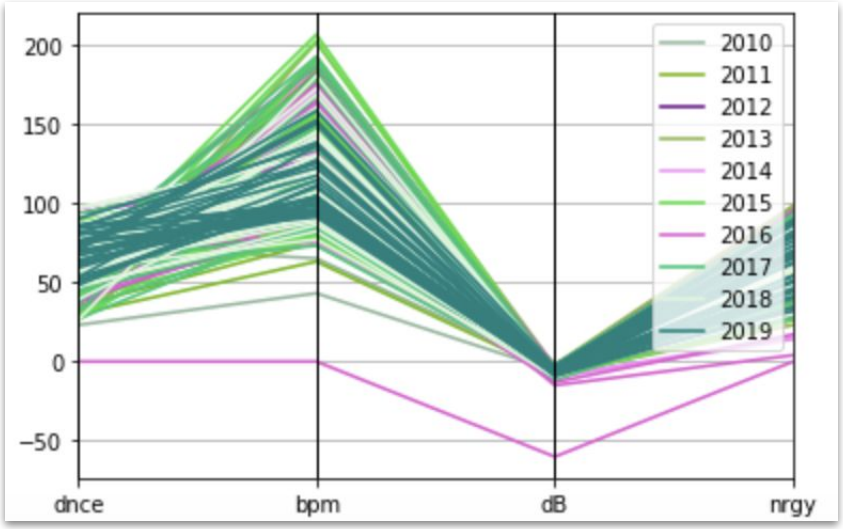
Scatter plot Matrices (cont.)





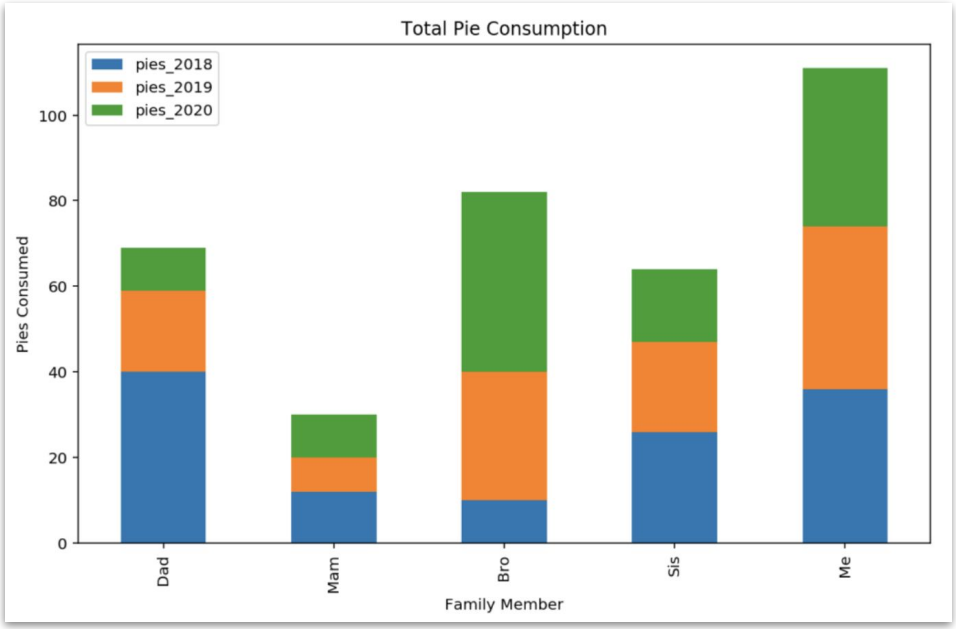
Parallel Coordinates

- Plotting many variables in one plot
- Each variable is represented in a column with its range independently scaled to its minimum to maximum range.
- The points on the vertical axes are the values of that variable. The lines connect data points for a single record, so in principle, you could trace an entire record from left to right in a parallel coordinate plot.



Overlaying the Target Variable in Summary

- One simple but powerful extension to any data visualization method is an overlay of an additional variable, usually a categorical or target variable.
- Adding this additional dimension provides a first look-ahead to the predictive power of the input variables in the chart.



Significance Measures

BASE STATISTIC	SIGNIFICANCE MEASURE
Mean	Standard Error of Mean
Skewness	Standard Error of Skew
Kurtosis	Standard Error of Kurtosis
Crosstabs	Chi-square Statistic and p value
Correlations	Pearson correlation coefficient and p value
Binomial Test	Error at 95% Confidence Level
Anova, Difference of Means	F-statistic

Data Auditing

- A list of data audit items might include the following:
  - How many missing values are there?
  - Is there a strange minimum or maximum value? Strange mean values or large differences between mean and median?
  - Is there large skew or excess kurtosis?
  - Are there gaps in the distributions? Any high-cardinality categorical variables?
  - Are there any unusually strong relationships with the target variable, possibly indicating leakage of the target into a candidate input variable?
  - Are any variables highly correlated with each other, possibly indicating redundant variables?
  - Are there any crosstabs that show strong relationships between categorical variables, possibly indicating redundant variables?

Data Preparation

Data Preparation

- Data preparation is intended to convert data identified for modelling into a better form for predictive modelling algorithms.
- The key steps in data preparation related to the columns in the data are variable cleaning, variable selection, and feature creation.
- Data preparation steps described to the rows in the data are record selection, sampling, and feature creation.
- Data preparation is the most time-intensive step in predictive modelling (ranging between 60 and 90 percent).

Data Cleaning

- Predictive modellers should fix data problems.
- Database cleanliness is not the same as predictive modelling cleanliness.
- Predictive modeller is often the first to examine the data in such detail, thus finding data uncleaned.
- The 60–90 percent guideline is a good rule of thumb for planning the first predictive modelling project using data that has never been used for modelling before.

Variable Cleaning

- Variable cleaning refers to fixing problems with values of columns, including incorrect or miscoded values, outliers, and missing values.
- Mistakes in variable cleaning can destroy predictive power in the variables that were modified.
- Sometimes the analysis may require domain experts to interpret the values, so the analyst better understands the intent of the values saved in the data.

Categorical and Continuous Values

- For categorical variables, examine the frequency counts and identify values of a variable that are unusual, occurring very infrequently.
- For continuous values, incorrectly coded values are most often detected as outliers or unusual values, or as spikes in the distribution, where a single repeated value occurs far more often than you would otherwise expect.

Consistency in Data Formats

- The format of the variable values within a single column must be consistent throughout the column.
- Most often mismatches in variable value types occur when data from multiple sources are combined into a single table. For example, if the purchase date in a product table comes from several sources
- You cannot have some dates with the format mm/dd/yyyy and others with the format mm/dd/yy in the same column.
- Another common discrepancy in data types occurs when a categorical variable (like ZIP code) is an integer in some data tables and a string in others.

Outliers

- Outliers are unusual values that are separated from the main body of the distribution, typically as measured by standard deviations from the mean or by the IQR (Interquartile Ranges).
- Outliers can be unusual for different reasons:
  - They are just data coded wrong (e.g. age: 141).
  - They exceed a fixed number of standard deviation from the mean.
- Some outliers distort the numerical models such as linear regression, k-nearest neighbor, and k-Means clustering, as the algorithms ignore most of the data.

Identifying and Fixing Outliers

- Without data visualization, the analyst may not even know these are unusual values because the summary statistics will not immediately provide clues.
- If the outliers are correctly coded:
  - Remove the outliers from the modelling data
  - Create separate models just for outliers
  - Transform the outliers so that they are no longer outliers
  - Bin the data
  - Leave the outliers in the data without modification

Fixing Outliers (cont.)

- Removing records with outliers in the inputs could remove some associated cases, e.g., patients with a particular disease.
- If outliers are good predictors of a target variable, the decision trees will split the data into the outlier/non-outlier data subsets.
- There are techniques (e.g., transformation, scaling) to reduce the distance between the outliers and the main body of the distribution.
- Binning to identify outliers may require some manual tuning.
- In data-driven analytics, we can use only algorithms unaffected by outliers, such as decision trees. We can also use the algorithms affected by outliers to purposefully bias the models in the direction of the outliers.

Missing Values

- Missing values are problematic, typically coded in data with a null value or as an empty cell.
- Missing values might be due to simple data entry errors, or they mean nothing more than that the values are completely unknown.
- Data might be lost inadvertently through data corruption or overwriting of database tables, or it was never collected in the first place (data collection limitations or even forgetfulness).
- If data is deliberately withheld during data collection (e.g., surveys), there may be predictive information missing from the value.

Typical Missing Values

POSSIBLE REPRESENTATION OF MISSING VALUES	DESCRIPTION
Null, empty string ("")	For numeric or categorical variables
0	For numeric variables that are never equal to zero
–1	For numeric variables that are never negative
99, 999, and so on	For numeric variables that have values less than 10, 100, and so on
–99, –999, and so on	For numeric variables that can be negative
U, UU	For categorical variables, especially 1- or 2-character codes
00000, 00000-0000, XXXXX	For ZIP Codes
11/11/11	For dates
000-000-0000, 0000000000	For phone numbers

Fixing Missing Values

- Listwise deletion:
  - Removing any record with any missing values, leaving only records with fully populated values for every variable to be used in the analysis. (e.g., no geolocation in data).
- Column deletion:
  - Removing variables with any missing values at all, leaving those that are fully populated (depends on number of missing values).
- Imputation with a constant value:
  - For categorical use "U", for continuous use 0 (if means nothing) or -1.

Fixing Missing Values (cont.)

- Mean and Median imputation for continuous variables:
  - Easy, but the more values imputed with the mean, the smaller the standard deviation becomes.
- Imputing with distributions:
  - Impute randomly from a known distribution
  - It will retain the same shape as the original shape
- Random imputation from own distributions:
  - A random actual value of the variable of the non-missing values is selected. The distribution of imputed values matches the populated data. Example:

Fixing Missing Values (cont.)

- Imputing Missing Values from a Model
  - Use the other input variables that may predict this new missing value variable (target variable).
  - The training data should be large enough, and all inputs must be populated; listwise deletion is an appropriate way to remove records with any missing values.
- Imputation for Categorical Variables
  - The missing value can be imputed with a value that represents missing so that no cell contains a null any longer.
  - There may be advantages to building predictive models to predict the variable with missing values.

How Much Missing Data Is Too Much?

- 50% missing values, remove data? There is no complete answer.
- If you are trying to build a classification model with the target variable populated with 50 percent 1s and 50 percent 0s, but one of the candidate variables is 95 percent missing, that variable is very unlikely to be useful, and will usually be removed from the analysis.
- What if the target variable is populated with 99 percent 0s and 1 percent 1s? The proportion of populated input values is five times that of the target; it may still be a useful predictor.

Feature Creation

- Features are new variables to add to data that are created from one or more existing variables already in the data.
- They are sometimes called derived variables or derived attributes.
- Creating features provides more value-added to the quality of data than any other step.
- A skewed variable is typically transformed by a function to a normal distributed.

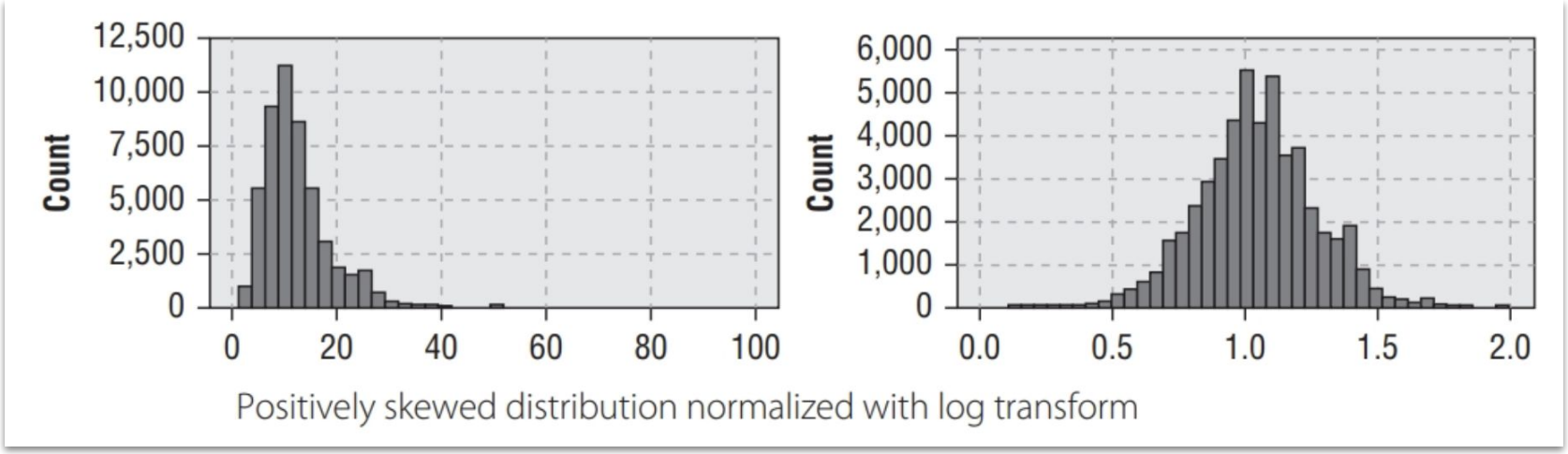
Feature Creation: Positive Skewness

- The most common corrections for positive skewness: log transform, the multiplicative inverse, and the square root transform.
- The log transform is perhaps the most often used transformation to correct for positive skew.
- The log base 10 pulls the tail in more than the natural log and is preferred sometimes for this reason.

TRANSFORM NAME	TRANSFORM EQUATION
Log transform	log(x), logn(x), log10(x)
Multiplicative inverse	$1/x$
Square root	sqrt(x)

Feature Creation: Positive Skewness (cont.)

- Take care that there are no 0 or negative values in the data, or you introduce undefined values into the transformed data.
  - For 0 values: add 1 to the original variable a common practice (log10(x + 1)
  - For negative values: add the absolute value of the minimum value of the variable (log10(|min(x)| + 1 + x) plus 1



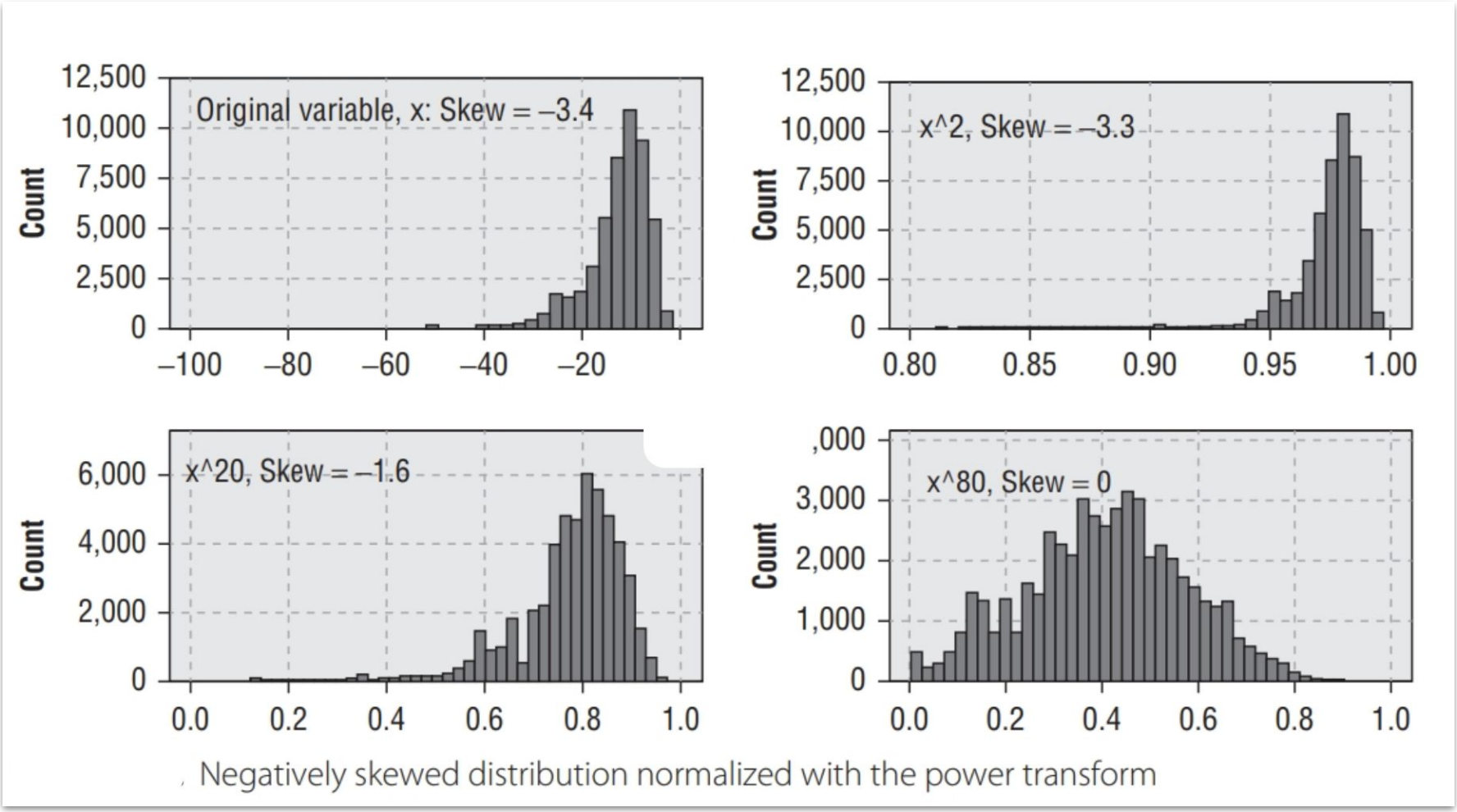
Feature Creation: Negative Skewness

- Negative skew is less common than positive skew but has the same problems with bias that positive skew has.
- Use power transform: square, cube, or raise the variable to a higher power. It is advisable to scale the variable first by its magnitude before raising it to a high power.
- The equation for transforming either a positively or negatively skewed variable is as follows:

$$\text{Transformed\_x} = \text{sgn}(x) \times \log10\,(1 + \text{abs}(x))$$

- the original sign of the variable back to the distribution
- Shift the distribution from zero
- Make the values positive

Feature Creation: Negative Skewness (cont.)



Negatively skewed distribution normalized with the power transform

Fixing Skewness: Summary

Effective transformations in converting positively or negatively skewed distributions to more balanced distributions.

PROBLEM	TRANSFORM
Positive skew	log10(1 + *x*), 1/*x*, sqrt(*x*)
Negative skew	x^n, –log10(1 + abs(*x*))
Big tails, both directions	sgn(*x*) x log10(1 + abs(*x*))

Numeric Variable Scaling

- Larger magnitudes produce larger distances. Scaling variables so that all have the same magnitudes removes this bias.
- A list of commonly used normalization methods appears below:

SCALING METHOD	FORMULA	RANGE
Magnitude scaling	$x' = x / \max\|x\|$	[-1,1]
Sigmoid	$x' = 1 / (1 + e(-x/c))$	[0,1]
Min-max normalization	$x' = (x - x_{min})/(x_{max} - x_{min})$	[0,1]
Z-score	$x' = (x - x_{mean})/x_{std}$	mostly [–3,3]
Rank binning	100*rank order / # records	[0,100]

Numeric Variable Scaling (cont.)

Z-scores are most appropriate when the data is normally distributed because the mean and standard deviation used in the scaling assume normal distributions.

Sample Z-Scored Data

AGE	AGE (MIN-MAX)	AGE (Z-SCORE
81	0.7821	1.1852
62	0.5385	−0.0208
79 close to mean	0.7564	1.0583
66	0.5897	0.2331
69	0.6282	0.4235
73	0.6795	0.6774
28	0.1026	−2.1790
50	0.3846	−0.7825
89	0.8846	1.6930
58	0.4872	−0.2747

Numeric Variable Scaling (cont.)

- Normalizing data can aid considerably in data visualization.
- Normalization of the data can help utilize more the space on the printed page available to view the data.
- For example, if a variable is log transformed to compress the effect of outliers and skew, and then is scaled to the range [0,1], and if another variable is also scaled to the range [0,1], you might find a pattern in the plot.
- Scaling and transforming data might force analysts to keep the translations in the head and show the original values of the variables on the plots (e.g., in the log transformation).

Nominal Variable Transformation

- Nominal variables present problems for numeric algorithms because they usually are not in numeric form. The most common approach to using nominal variables is to explode them into dummy variables.
- Dummies are usually coded with 0s and 1s, where the value 1 corresponds to "true" and 0 to "false."
- The recoding of categorical variables is accomplished by exploding a single column of categorical values to N columns of dummy variables.

STATE	CA	TX	FL	NY
CA	1	0	0	0
TX	0	1	0	0
FL	0	0	1	0
NY	0	0	0	1

Ordinal Variable Transformations

- Ordinal variables are problematic for predictive modelling and most algorithms do not have specific ways to incorporate ordinal variables in the algorithms.
- The modeller must usually make a choice: Should the ordinal variable be treated as continuous or categorical?
- For variables such as education:
  - Categorical: high school, bachelor's degree, and Ph.D. in one group
  - Numerical: a bachelor's degree will be considered four time as big as high school
  - Or, a thermometer scale may be more appropriate.

Ordinal Variable Transformations (cont.)

EDUCATION LEVEL	HIGH SCHOOL	SOME COLLEGE	ASSOCIATE'S DEGREE	BACHELOR'S DEGREE	MASTER'S DEGREE	PHD
High school graduate	1	0	0	0	0	0
Some college	1	1	0	0	0	0
Associate's degree	1	1	1	0	0	0
Bachelor's degree	1	1	1	1	0	0
Master's degree	1	1	1	1	1	0
PhD	1	1	1	1	1	1

Date and Time Variable Features

- Date and time variables are difficult to work with in predictive modelling and nearly always require some feature creation to be used effectively.
- Problems such as
    - date/time representation: March 1, 2013, 1-March-2013, 1-Mar-99, 3/1/99, 3/1/2013, 03/01/2013, 03012013
    - cyclical nature: days, months (Jan to Dec)

Which Version of a Variable Is Best?

- The original version of a variable and its related log10 transform will be very highly correlated.
- In summary, use an empirical approach to select the best version of a variable for predictive modelling.
- If you use each version as a candidate input variable by itself to predict the target variable, you can score each version of the variable and pick the one that is the best predictor.

Multidimensional Feature

- Two common examples of multidimensional features are interactions (created by multiplying two variables together) and ratios (created by dividing one variable by another).
- Creating ratios by domain experts helps the algorithms find patterns more clearly and simply than if they had to approximate the ratio through a series of multiplies and adds.
- Principal Component Analysis (PCA) is a method that helps to find correlated variables, variables that have significant overlap with each other, and describe them as Principal Components (PCs).

Multidimensional Feature (cont.)

- PCA can be very useful only if there is a linear projection of the data.
- Clustering algorithms also find records that fall into groups from the variable values and assign a label for each group.
- The most straightforward feature to create from a clustering model is the cluster label (dummy variable) itself. You can create a distance (a new feature) between the record and each of the cluster centers.
- Decision trees are perhaps the most commonly used algorithm for creating new features because they produce rules that are easy to understand and easy to implement.

Variable Selection Prior to Modelling

- Reduce variables through variable selection. Remove irrelevant variables, redundant variables, and variables that are unlikely to improve model accuracy.

- Stepwise Regression is a common variable selection technique that has been integrated into linear regression.

- More sophisticated techniques such as using guided random search (genetic algorithms or simulated annealing) have been added to neural networks to improve variable selection.

Removing Irrelevant and Redundant Variables

- Irrelevant variables that have either no relationship with the target variable or are improper to include in the models.
- Irrelevant variables include labels, such as ZIP code and customer ID might have some meaning, but generally can be removed. (e.g, older ID → older patient)
- Redundant variables are whose information is conveyed by one or more other variables.
- Correlation matrices identify highly correlated variables (between -0.9 to 0.9) that are good candidates for removal.

List of Techniques for Variable Selection

TECHNIQUE	COMPARISON TYPE	INCLUSION METRIC
Chi-square test	Categorical input vs. categorical target	*p* value or top *N* variables
CHAID tree stump using Chi-square test	Continuous or categorical input vs. continuous or categorical target	*p* value or top *N* variables
Association Rules confidence, 1 antecedent	Categorical input vs. categorical target	Confidence, Support
ANOVA	Continuous input vs. categorical target	*p* value, top *N* variables
Kolmogorov-Smirnov (K-S) Distance, two sample test	Continuous input vs. continuous target	K-S test critical value, top *N* variables
Linear regression forward selection (1 step)	Continuous input (or dummy) vs. continuous target	*p* value, AIC, MDL, top *N* variables
Principle Component Analysis (select top loader for each PC)	Continuous input vs. continuous target	Top *N* variables

Variable Creation and Selection Process

- The quick and automated way to assess all interactions efficiently is through the use of association rules.
- Association rules work only with categorical variables, so any continuous variables must first be binned, but association rules are fast, efficient ways to find interesting interactions.
- Variable creation and selection process
  - Clean single variables from problems such as incorrect values, missing values, and outliers.
  - Create new features from existing variables.
  - Reduce the number of candidate inputs by removing redundant and irrelevant variables and features.

Sampling Datasets

- The most common sampling strategy is splitting the data into training and testing subsets: One builds the model on the training data and then assesses the model on the held-out, testing subset.
- It is certainly possible that the testing data becomes a driver for how the algorithm builds the model. One solution to this potential problem is to have a third data set available: validation data.
- Each of these three sampled subsets of the entire data set must be large enough to be representative samples, but are separate from one another: Each record belongs to one and only one of the three subsets.

Typical Process for Building Models

1. Build a model on training data.
2. Assess the model on testing data.
3. Change settings, parameters, or inputs to the model.
4. Re-build the model on training data.
5. Assess this new model on testing data.
6. Change settings, and repeat the process.

Rules of Thumb for Determining Data Size

- We need enough records in the data to build reliable models. The actual number of records, however, is not easy to determine precisely and depends on patterns found in the data itself

- There are two primary considerations for determining sample size:
  - Number of input variables (dimensionality)
  - Sufficiently populated target variable

- These rules can also be used to reduce the size of the large data in building models. If the rules indicate that you need only 50,000 records, most of these records are not needed in the training set.

Summary of Data Preparation

- Data preparation takes considerable thought and effort to ensure the data is presented to the algorithms so they can be used effectively.
- Predictive modellers need to understand how the algorithms interpret the data to prepare the data appropriately for the algorithm.
- The data preparation stage is often revisited once problems or deficiencies are discovered while building models.
- Feature creating, in particular, is iterative as you find out which kinds of features work well for the data.

Python: A Data Science Tool

Python and Predictive Analytics

- We use Python in this lecture to:
  - prepare data for a predictive model,
  - explore data and visualize it,
  - apply machine learning model,
  - train a model and test it, and
  - evaluate a predictive model.
- Python vs. R for predictive models:
  - R is used for statistical modelling, by statisticians and data scientists with no deep programming skills.
  - Python is used by both data scientists and developers for various purposes, including developing machine learning models.

Python Popular Libraries

- **Numpy**: Math library to work with N-dimensional array
- **Scipy**: Scientific library for high performance computation
- **Matplotlib**: Popular package for working with 2D or 3D plotting
- **Pandas**: High level library for data importing, manipulation and analysis
- **Scikit-learn**: Working with machine learning algorithms, works with Numpy and Scipy
  - preprocessing package
  - feature selection package
  - functions for training datasets and model assessment

References

- El Morr, C., & Ali-Hassan, H. (2019). Analytics in Healthcare: A Practical Introduction. Springer.
- R. Sharda, D. Delen, E. Turban, J. Aronson, and T. P. Liang, Business Intelligence and Analytics: Systems for Decision Support. 2014.
- T. L. Strome, Healthcare analytics for quality and performance improvement. Hoboken, NJ: Wiley, 2013.
- L. Madsen, "The Tenets of Healthcare BI," in Healthcare Business Intelligence: A Guide to Empowering Successful Data Reporting and Analytics: Wiley, 2012.
- Abbott, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. 2014.
- Witten, I. H., Frank, Eibe, Hall, Mark A. Data Mining: Practical Machine Learning Tools and Techniques. 1884-1970.; Palestro, Christopher J.
- Kuhn, M., & Johnson, K. Applied predictive modelling (Vol. 26). New York: Springer, 2013.