

Summary of Data Mining and Analytics I

Overview of Data Mining

Objective: Decision assistance

Definition: Combination of artificial intelligence and statistical analysis to discover information that is “hidden in the data”

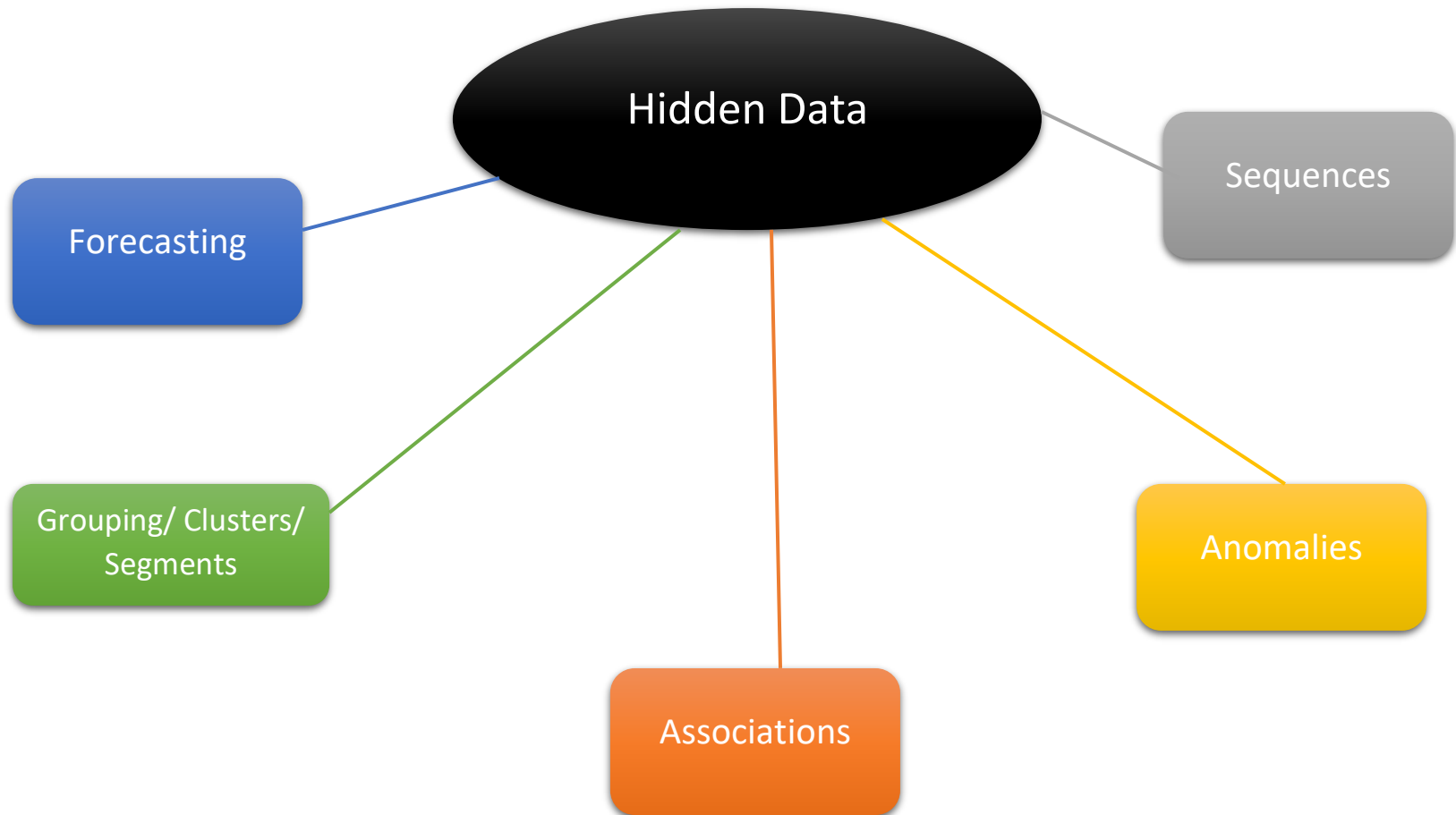
Types of Data Mining		
Basis for Comparison	Descriptive/Exploratory Mining	Predictive Mining
Basic	Brings out information that is present but hidden in a mass of data. Finds clusters/groups/associations between different products	Extrapolates new information based on present information. New information is qualitative (classification/scoring) or quantitative (regression).
Requirements	Data aggregation and data mining	Statistics and forecasting methods
Preciseness	Provides accurate data	Produce results that does not ensure accuracy.
Approach	Reactive	Proactive
Practical Analysis Methods	Standard reporting, query/drill down and ad-hoc reporting.	Predictive modelling, forecasting, simulation and alerts.
Examples	<i>Adverse events of a drug were explored by clustering the therapeutic classes; A data analyst receives detailed customer purchasing data and finds associations of any type among customers.</i>	<i>An automobile company scored customers for likelihood to return to buy a new model within the next 6 months; A credit card company offered a valued customer product for their card holder based on past card usage to determine the risk pattern; Road traffic was forecasted hourly</i>

Data Mining Aims



Important Terms

Univariate	Explores the statistics and details of 1 variable (e.g. <i>mean, median, mode, standard deviation, outliers, etc.</i>). Useful for summarizing data and finding patterns.
-------------------	--



Commercial Data Types

Transactional

Data describing an event.

e.g. orders, payments, deliveries.

Always has a time dimension and a numerical value.

Product

Data that is describing a product.

e.g. shoes, cars.

Customer

Data that describes a customer.

e.g. customer ID, first and last name.

Geodemographic

Data about a population in an area.

Technical

Data that gives a status report on something.

e.g. date of death, official titles, payer status.



Customer			Contact		
FirstName	LastName	CustID	ContactID	ContactInformation	ContactType
Steve	Stevens	101	101	555-2803	Work
Mary	Delman	102	101	555-8957	Cell
Skip	Stevens	103	102	555-8910	Work
Drew	Lakeman	104	104	555-8949	Work
Eva	Plummer	105	103	555-9850	Work
			101	555-8950	Home
			105	Plummer@akcomm.com	Email
			101	Stevens@akcomm.com	Email
			101	555-5787	Fax
			103	Stevens@akcomm.com	Email
			105	555-5675	Work
			102	Delman@akcomm.com	Email

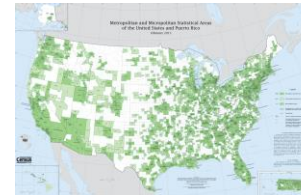
One to Many Relationship

Parent Table

Primary Key

Foreign Key

Child Table



Customer Data Types

Relational

Customer reactions to marketing.



Attitudinal

Customer satisfaction/loyalty.



Psychographic

Customer personality.



Lifetime

How long one has been a customer.



Channel

Channel through which contact was made.
(sponsorship, ads)

Preferred channel for contact

Preferred channel for orders

Preferred delivery channel



Sociodemographic

Personal (sex, level of education)

Family (family situation, # & ages of children, # dependents)

Occupational (income, occupation, social category, # working & retired people)

Wealth

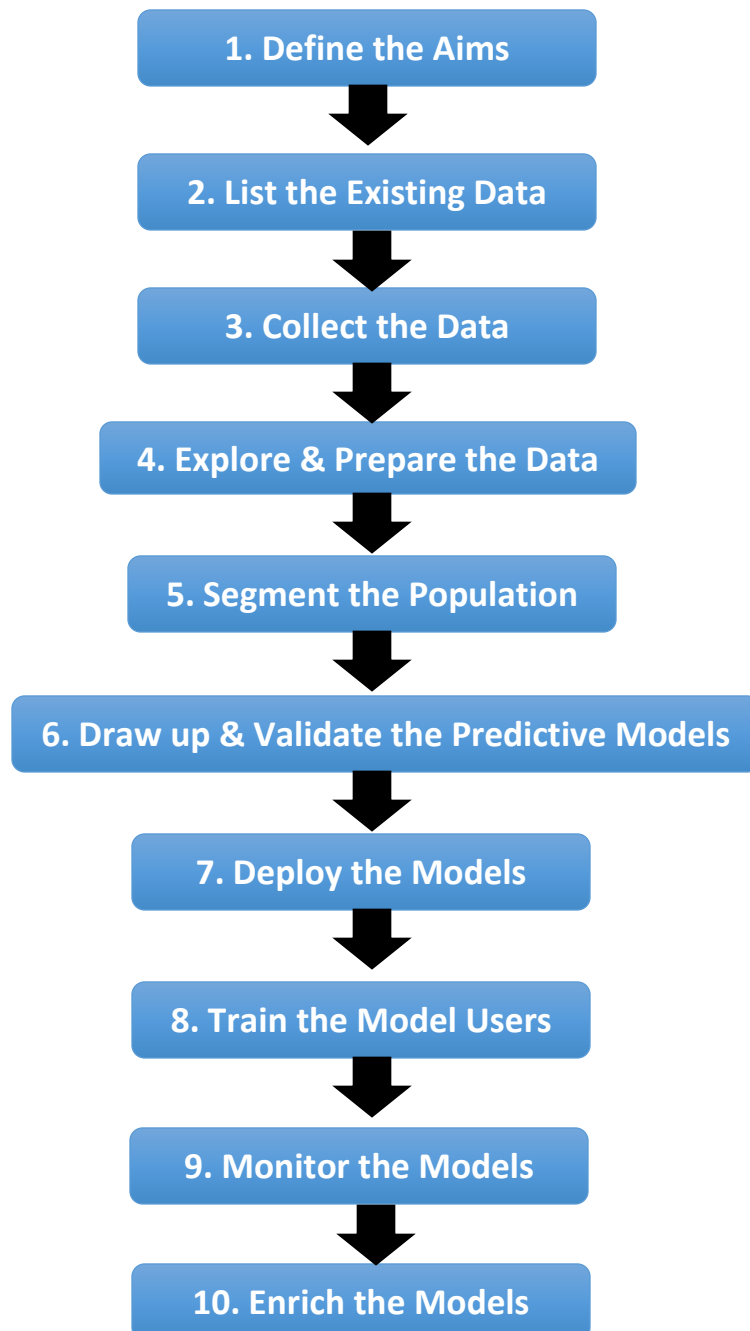
Geographical

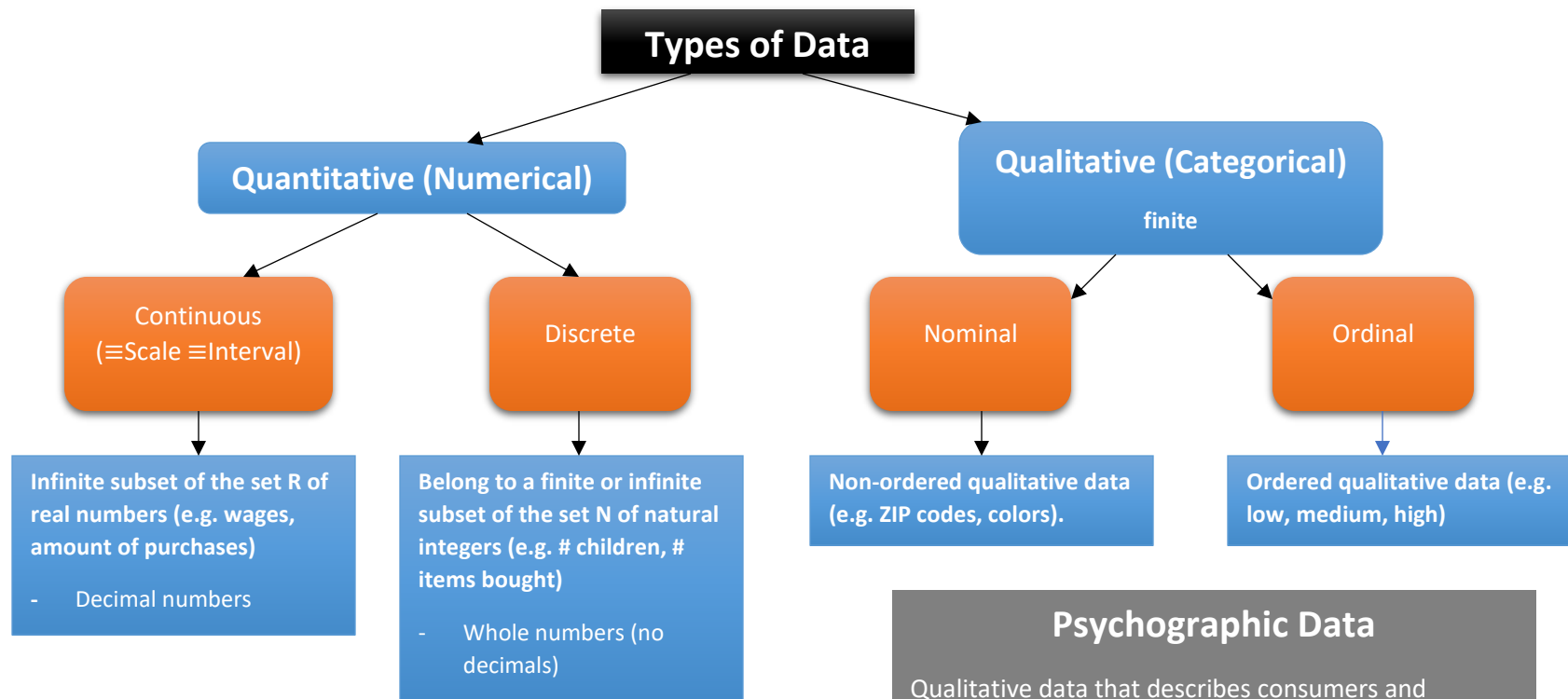
Environmental & Demographic

(competition, population, working population, unemployment rates)



Data Mining Study Development





Psychographic Data

Qualitative data that describes consumers and customers based on psychological attributes:

- Lifestyle
- Personality (shy, prudent, ambitious, outgoing, etc.)
- Values (conservative, liberal, materialistic, etc.)
- Risk aversion (trustful, mistrustful, anxious, demanding, etc.)
- Knowledge
- Focus of interest
- Opinions and behavior

Types of Values

Rare Value

Can create bias in factor analysis and other analysis, by appearing more important than they are.

What to do:

Remove

Replace with a more frequent value.

Missing Value

Gaps in the data.

What to do:

<10% data not excluded:
Remove corresponding observations

Mean substitution

Aberrant Value

Erroneous value corresponding to incorrect measurement, a calculation error, or a false declaration.

***Incorrect dates:** Unknown DOB replaced by 'round numbers', subscription dates before customer's DOB, dates of last updates in the year 2050, 29/Feb in non-leap year.*

Customers declared as 'private' when they are 'business'.

Amount input as cents when it should be in Euros.

What to do:

Delete if not too numerous and if their distribution is suitably random.

Replace with statistically imputed value.

Extreme Value

Observations in a sample so far separated in value from the remainder as to suggest they may be from a different population, or the result of an error in measurement.

What to do:

1-2% data not excluded:
Exclude outliers

Neutralize: Divide continuous values into classes

Winsorizing: Replace values of the variable beyond 99th percentile with this percentile.

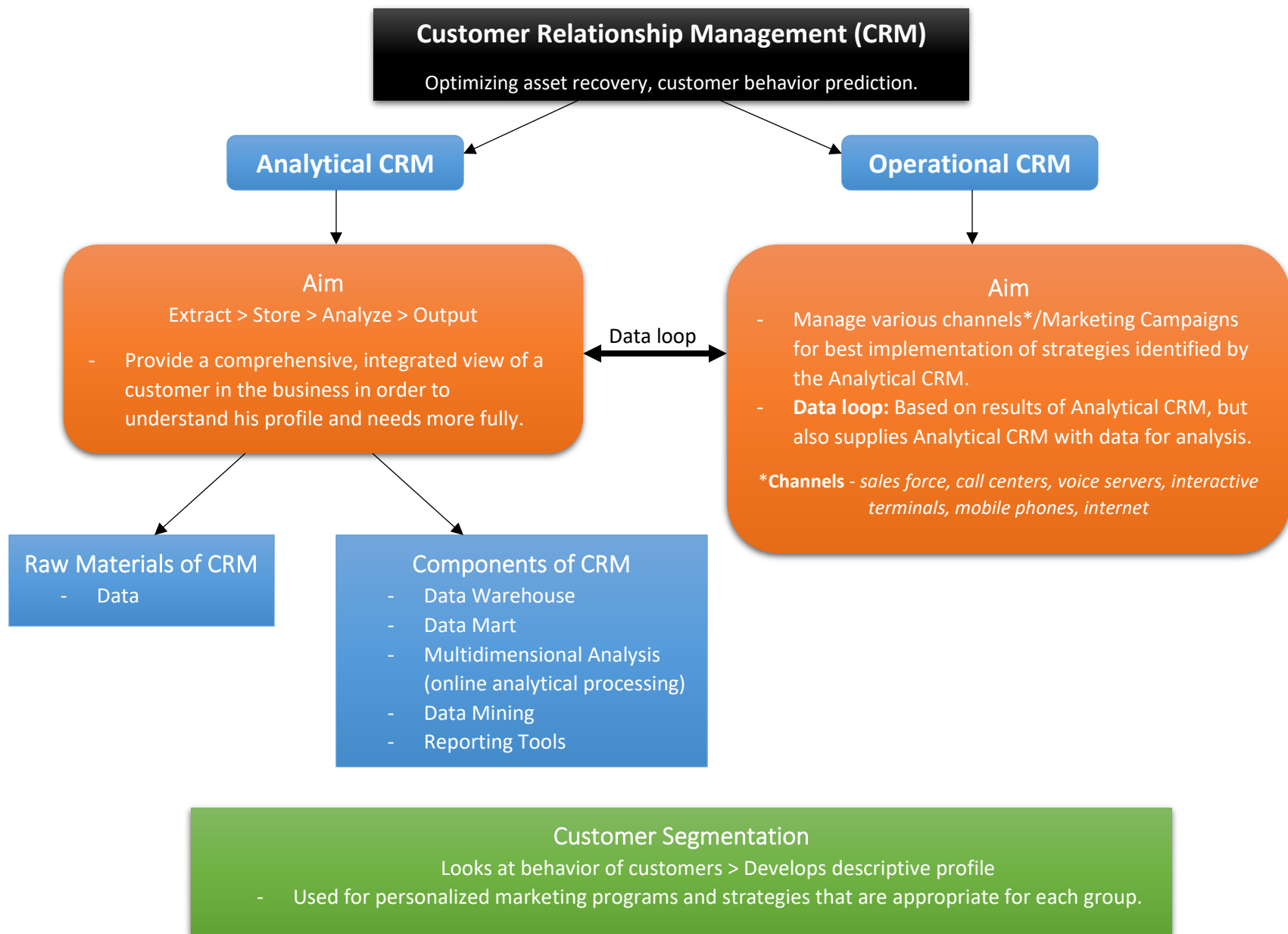
Missing Value

Name	Income	Job	name	children
Alice	8000000	trust fund retiree	Ethan	2
Bob	40000	rideshare driver		2
Charlie	1	racecar driver	Refuse to answer	1
Danielle	90000	marketing mgr	Gerald	2

Extreme Value, Aberrant Value, Rare Value

Age	Gender	Hair	Eye	Weight	Salary
14	F	Blue	Blue	143	12500
28	F	Brown	Brown	9	32150
22	M	Blue	Brown	215	34200
46	F	Brown	Orange	190	53200
75	M	Gray	Green	187	28040

Aberrant Values



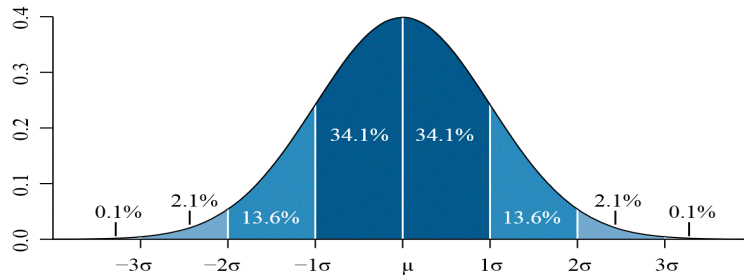
Tests Summary

Type	Test Name	When to use
Normality	Shapiro-Wilk (best)	+ Straight line, Slope = 1 Test to see if data deviates from a straight line.
	Kolmogorov-Smirnov (general)	- $p < 0.05$ or $p < 0.1$ Compares the cdf (cumulative density function) of the variable tested vs. cdf of a gaussian variable and determines the probability of observing a deviation as large or larger.
	Lilliefors	
	Anderson-Darling	
Homoscedasticity	Levene (best)	Best for non-normal distributions
	Bartlett	Best for normal distributions
	Fisher	Least robust if normality is not present
Bivariate (2 discrete variables)	Cramer's V	
	Chi-Square	
Bivariate (1 continuous, 1 discrete)	Parametric ANOVA	Requires normality & homoscedasticity
	Wilcoxon-Mann-Whitney	Non-parametric, 2 groups
	Kruskal-Wallis	Non-parametric, >2 groups

Mean comparison tests		
Form of distribution	Two samples	Three or more samples
Normality & Homoscedasticity	Student's t test	ANOVA
Normality & Heteroscedasticity	Welch's t test	Welch – ANOVA
Non-normality & Heteroscedasticity	Wilcoxon–Mann–Whitney	Kruskal–Wallis
	Median test	Median test
		Jonckheere–Terpstra test (ordered samples)

Tests

Tests for Normality



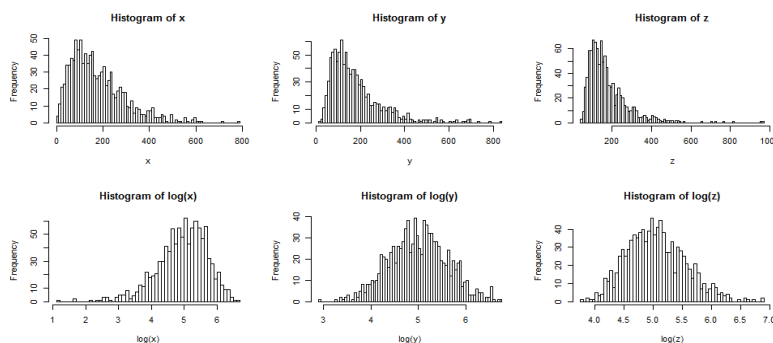
The normality of a variable can be verified by the following tests:

1. **Shapiro–Wilk test (the best)**
 - a. Normal distribution: Straight Line with **Slope = 1**
2. **Kolmogorov–Smirnov test (the most general)**
 - a. **NOT** a Normal distribution: If probability **<0.05** or **<0.10** we reject H_0
3. Lilliefors test
4. Anderson–Darling test

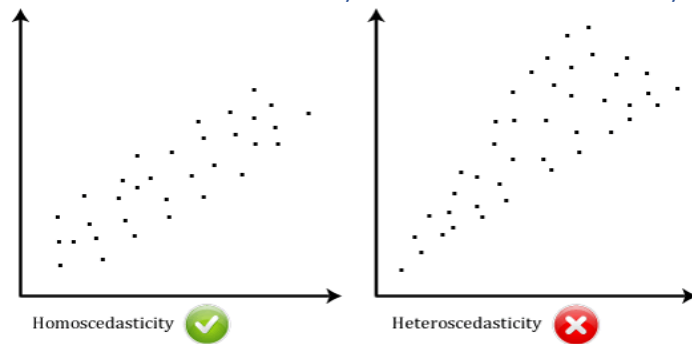
Tests for Normality				
Test	Statistic		P-Value	
Shapiro-Wilk	W	0.992264	Pr < W	0.9845
Kolmogorov-Smirnov	D	0.051999	Pr > D	> 0.1500
Cramer-von Mises	W-Sq	0.02621	Pr > W-Sq	> 0.2500
Anderson-Darling	A-Sq	0.162071	Pr > A-Sq	> 0.2500

Transformations Used to Reduce Skewness

1. **Log transformation** (has to be adjusted for 0 and negative values)
 - a. If data is skewed, log transformation will transform the data to a form where we see a more normal distribution.
2. Square (left skew)
3. Square root (right skew)
4. Cube (right skew)



Test for Homoscedasticity and Heteroscedasticity



1. **Levene Test** (best – low sensitivity to non-normality)
2. **Bartlett Test** (best if distribution is equal)
3. **Fisher Test** (least robust if normality is not present)

Test of Homogeneity of Variance			
Levene Statistic	Df1	Df2	Significance
50.448	2	396	.000

Bivariate Tests

Measuring Links between Two Discrete Variables

e.g. gender and smoking

1. **Cramer's V**
2. **Chi-Square**

Discrete



Discrete

Measuring Links between One Discrete & One Continuous Variable

e.g. dosage of a medicine and recovery time

Discrete



Continuous

1. **Parametric ANOVA Test**
 - a. Requires normality/homoscedasticity assumption
2. **Non-Parametric Approaches**
 - a. **Wilcoxon-Mann-Whitney** (2 groups)
 - b. **Kruskal-Wallis** (>2 groups)

Descriptive Learning Methods (Unsupervised)

K-means,
moving centers

Cluster Analysis

Clusters similar data points to each other that are less like those in separate clusters.

E.g. Consider a bookstore with different areas for the types of books: History, Self-Help, Romance, Mystery & Crime etc. The books in each of these clusters are more like each other than they are to other clusters.

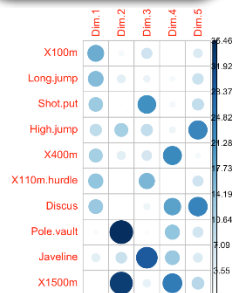


© www.khanacademy.org/a/descriptive-learning-methods

PCA

Help you identify which variables are important so you can compress the data by reducing the number of dimensions.

E.g. gender, days of week they typically shop, amount spent on average trip.



Market Basket
Analysis

Association Analysis

In a given set of records, each will contain a number of items. Association analysis allows you to determine the degree to which the items tend to be associated with one another.

E.g. people who buy hamburger buns will also likely buy mustard and hamburger meat. You can associate the items together and create rules.



Neural Clustering

Classification method - AKA Self-Organizing Map (SOM)

Unsupervised learning **neural network**

In the end of the learning, all inputs are classified as the node they are closest to. Each input that is processed changes the location of the nodes.

No variables to predict, it learns the structure of the data in order to distinguish clusters hidden within the data.

Operates like a matrix that is made up of cells (nodes), vectors, and magnitudes (weights) of each vector.

Kohonen
Network (SOM)

Factor Analysis

Multiple observed variables have similar patterns of responses because they are all associated with a latent (*i.e. not directly measured*) variable.

Involves grouping variables with similar attributes into a matrix using linear algebra techniques.

e.g. Dr. Peterson personality tests: Lots of similar questions in test, if participants answer all similarly = more likely they have that trait.

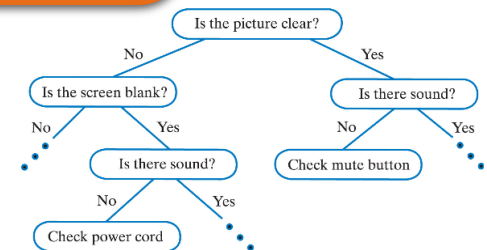
CPA, CMA, CA

Decision Trees

Both Descriptive and Predictive.

Maps possible outcomes of series of related choices.

Useful for identifying criteria to divide data into classes, weighing possible actions against one another based on their potential outcomes.



Descriptive Learning Methods (Unsupervised)

1. PCA – Principal Component Analysis

- a. Help you identify which variables are important so you can compress the data by reducing the number of dimensions.
- b. *E.g. gender, days of week they typically shop, amount spent on average trip.*

2. Cluster Analysis (K-means, moving centers)

- a. Clusters similar data points to each other that are less like those in separate clusters
- b. *E.g. Consider a bookstore with different areas for the types of books: History, Self-Help, Romance, Mystery & Crime etc. The books in each of these clusters are more like each other than they are to other clusters.*

3. Association Analysis (≡ Market Basket Analysis)

- a. In a given set of records, each will contain a number of items. Association analysis allows you to determine the degree to which the items tend to be associated with one another.

4. Neural Clustering (Kohonen Network (SOM))

5. Factor Analysis (CPA, CMA, CA)

6. Decision Trees (both Descriptive and Predictive).

1. PCA – Principal Component Analysis

2. Cluster Analysis (K-means, Moving Centers)

Clustering is used for organizing variables into smaller, homogenous groups to simplify analysis and find patterns in data.

- Variables are grouped together based on similar traits, i.e. factors like age, income, education level etc.
- Typically a highly subjective process since it is dependent on one's familiarity with the data.

Cluster Types

Cluster Types	
Exclusive Clustering	Each variable can only belong in one single cluster. It cannot belong to any other cluster.
Fuzzy Clustering	Data points are assigned a probability belonging to one or more clusters
Overlapping Clustering	Each variable can belong to more than one cluster.
Hierarchical Clustering	Iterative process that either divides or combines clusters
Probabilistic Clustering	Data is clustered using algorithms that connect variables using distances or densities. This is done via computer.

K-Means

- Clustering algorithm
- **Useful for data that have NO labels, i.e. data without defined categories or groups.**
- Allows researcher to find groups of data which are like each other and cluster them together, even if the data is not labeled.
- Clusters defined by centroids, there are k centroids in each data set. k is a value returned by the algorithm (a point considered to belong to a particular cluster if the point is closer to that cluster's centroid than any other centroid).

- Iterative method, process of assigning data points to clusters based on current centroids and then choosing next, closer centroid is done until the data set reaches convergence.

Moving Centers

- Clustering algorithm to group data points
- Akin to K-Means method as they work in the same manner.

3. Association Analysis/Rules (≡ Market Basket Analysis)

- Best if used for finding the **most frequently occurring combinations** of variables in a data set.
- Commonly used for Market Basket Analysis
- *E.g. Amazon store has “Customers who viewed this item also viewed” and “frequently bought together”.*
- Apriori is the main algorithm for detecting association rules.

4. Neural Clustering (Kohonen Network (Map))

- Classification method - AKA Self-Organizing Map (SOM)
- Unsupervised learning **neural network**
- In the end of the learning, all inputs are classified as the node they are closest to. Each input that is processed changes the location of the nodes.
- No variables to predict, it learns the structure of the data in order to distinguish clusters hidden within the data.
- Operates like a matrix that is made up of cells (nodes), vectors, and magnitudes (weights) of each vector.

Neural Network

Can be used for clustering as well as classifying data (predictive and descriptive, qualitative or quantitative dependent variable).

- Descriptive (Unsupervised) or Predictive (Unsupervised) learning
- The patterns recognized by neural networks are numerical and contained in vectors. Input data is translated from its raw form into numerical vector values.

+	Handles non-linear relations between the variables
-	Requires massive amounts of computer power.

5. Factor Analysis

- Collapses many variables into a few underlying factors.
- **Key concept: Multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable.**
- Involves grouping variables with similar attributes into a matrix using linear algebra techniques.
- In FA there are the same number of factors as there are variables.
- The eigenvalue is a measure of how much of variance of the observed variables a factor explains – any factor with an eigenvalue ≥ 1 explains more variance than a single observed variable.
- Factor analysis methods include: Principal Component Analysis (PCA), Correspondence Analysis (CA), and Multiple Correspondence Analysis (MCA).
 - **Correspondence Analysis (CA):** Intended for the analysis of qualitative variables.

- Plots variables in relation to one another.
- High frequency denotes a strong positive relationship, *i.e. two positively related categories, A&B, are opposed*, intermediate frequency denotes a weak relationship.
- Can be used to convert qualitative values into quantitative values.
- **Multiple Correspondence Analysis (MCA):** Component with > 2 variables.
 - Can be used to convert qualitative values to quantitative values.

6. Decision Trees (both Unsupervised/Descriptive & Supervised/Predictive)

- Maps possible outcomes of series of related choices.
- Useful for identifying criteria to divide data into classes, weighing possible actions against one another based on their potential outcomes.
- Detects two-way interactions between tables.

+	Easy to understand, Non-parametric, Models non-linear response of dependent variable, Can handle collinearity, missing data and outliers using isolating nodes, Simpler data preparation phase Reasonable computing times
-	Become extremely complex very quickly, Evaluates independent variables sequentially, not simultaneously, causing lack of robustness of the model and potential bias (can be overcome by resampling and using a mean, but this will cause the loss of the simplicity of the model) Large dataset required.

Predictive Learning Methods (Supervised)

Multi-Layer Perceptron

A class of feed forward artificial network consisting of 3 layers of nodes: input layer, hidden layer, output layer. MLP uses a supervised learning technique called backpropagation for training.

Radial Basis Function

Neural network.
Works with only 1 hidden layer.
Uses circle as a basis for analysis
Prediction & classification.

Support Vector Machines

Used for classification, regression, and detecting outliers in data.
SVM is used to separate classes from each other.

Naïve Bayes Method

Classifies data based on the probability that it belongs to the class given certain features.
fruit has a high probability of being classified as an apple if it is red, round, and about 3 inches in diameter.

Linear Discriminant Analysis

Maximize separability: algorithm searches for a linear combination of variables (predictors) that best separates two classes (targets)

Similar to PCA, but instead of maximizing variation, maximizes separability.

Best when assumptions of homoscedasticity and multinormality are correct.

Regression Model

ANOVA

K-Nearest Neighbors

Decision Trees

Both Descriptive and Predictive.

Maps possible outcomes of series of related choices.

Useful for identifying criteria to divide data into classes, weighing possible actions against one another based on their potential outcomes.

Predictive Learning Methods (Supervised)

1. Multi-Layer Perceptron (MLP)

Neural Network Perceptron

The perceptron of an algorithm for supervised learning of binary classifiers.

- **Binary classifiers:** functions that can decide whether an input belongs to some specific class or not.
- Uses machine learning to maximize separability of variables into categories:
 - a. Starts with random weights and sees how many it got right/wrong.
 - b. Adjusts the weights and sees if it got more right this time than last time.
 - c. Repeats until accurately separates the data into classes.
 - d. Once model is trained, the weights are applied to a different dataset.

2. Radial Basis Function (RBF)

- Neural network.
- Works with only 1 hidden layer.
- Uses circle as a basis for analysis
- Prediction & classification.

- | | |
|---|---|
| - | It may need large number of units in its hidden layer, which increases execution time of network without always yielding perfect modelling of complex structures and irregular data. |
|---|---|

3. Support Vector Machines

- Supervised learning methods used for classification, regression, and detecting outliers in data.
- Builds perceptive model that can assign new input objects into 1 category or another – Essentially SVM is used to separate classes from each other.

- | | |
|---|--|
| + | Can model non-linear phenomena,
Precision of prediction in certain cases,
Robust |
| - | Does not directly provide estimated probabilities,
Sensitive to choice of Kernel parameters,
Long computation time
Limited number of software programs that can implement it. |

4. Naive Bayes Method

- Classifies data based on the probability that it belongs to the class given certain features.
- **Can group objects based on features:** *e.g. fruit has a high probability of being classified as an apple if it is red, round, and about 3 inches in diameter.*

- | | |
|---|----------------------------------|
| - | Requires large data sets. |
|---|----------------------------------|

5. Fisher's Linear Discriminant Analysis

- **Maximize separability:** algorithm searches for a linear combination of variables (predictors) that best separates two classes (targets)
- Similar to PCA, but instead of maximizing variation, maximizes separability.

- Best when assumptions of homoscedasticity and multinormality are correct.

+	Fast calculation time, Concise models, Good at detecting global phenomena (c.f. decision trees which detect local phenomena), Does not need a lot of records.
-	Only detects linear phenomena, Can only be used for continuous variables (except for DISQUAL), Sensitive to outliers.

Disqual

- An extension of Fisher's Discriminant Analysis developed by Gilbert Saporta in 1975.
- Allows for qualitative dependent variables.

6. Regression Models

Linear regression

Assumes the following:

- Residuals** (error terms) are **normally distributed**.
- Best fitting regression line is a straight line**.
- Residuals** (error terms) have **constant variance** at every value of x .
- Residuals** (error terms) are **independent**.
- Residuals have a mean of zero** (error terms sum to 0).

7. Anova

8. K-nearest neighbors

9. Decision Trees are both predictive and descriptive

- Maps possible outcomes of series of related choices.
- Useful for identifying criteria to divide data into classes, weighing possible actions against one another based on their potential outcomes.
- Detects two-way interactions between tables.

+	Easy to understand, Non-parametric, Models non-linear response of dependent variable, Can handle collinearity, missing data and outliers using isolating nodes, Simpler data preparation phase, Reasonable computing times.
-	Become extremely complex very quickly, Evaluates independent variables sequentially, not simultaneously, causing lack of robustness of the model and potential bias (can be overcome by resampling and using a mean, but this will cause the loss of the simplicity of the model), Large dataset required.

Speed Round – Choosing a method for:

Choosing a method for:	
High precision	Discriminant Analysis Linear regression Logistic regression
Small dataset	Avoid decision trees and neural networks.
Missing values, heterogeneous data (outliers)	Association rules Decision trees Logistic regression DISQUAL ... can be used by coding the missing values as a special class.
Highly correlated independent variables	Decision Trees
Without having to prepare or homogenize data	Decision Trees
Dependent variable = qualitative	Naïve Bayes Method Linear Discriminant Analysis (+DISQUAL) Support Vector Machines Logistic Regression
Dependent variable = quantitative	Linear Regression ANOVA
Dependent variable = qualitative OR quantitative	Decision tree (1 quant, 1= qual) Neural Networks (> 1 quant, > qual) K Nearest Neighbors (1 quant or qual)

Types of Analyses

Time Series Analysis

- Used to study the consequences of measurements of variable or variables. The measurements are often made at regular time intervals.
- Explains how the past affects the future or how the two time series can “interact” with each other *i.e. Forecasts future values of the series.*
- **Has defining characteristic that order of observations is critical, because there is dependency between observations, so changing the order of the observations can change the meaning of the data.**

Stepwise Selection

- **Forward stepwise selection** starts with no variables in model initially, variables are added into the model one by one.
- **Backward stepwise selection** starts with all of the variables in the initial model, variables are removed from model one by one.
- **Combined stepwise selection** starts with forward stepwise then changes to backward stepwise, process alternates between forward and backward until no variables can be removed or added.

MOSIAC Analysis

- **Geodemographic topology based on sociodemographic data, lifestyles, behavior and preferences.**
- Used in Experian credit reporting – Comprised of financial variables like education level, size of household, occupation, income, etc.

RFM Analysis

- **RFM: Recency, Frequency, Monetary value**
- Cross-tabulates recency of last purchase in the period being studied with the frequency of purchases in that period, then examines the distribution of purchases.
- Answers the where, when, how, quantity, and what questions – analyses where the product was purchased? When was it purchased? How was it paid for? How much was purchased? What was purchased?

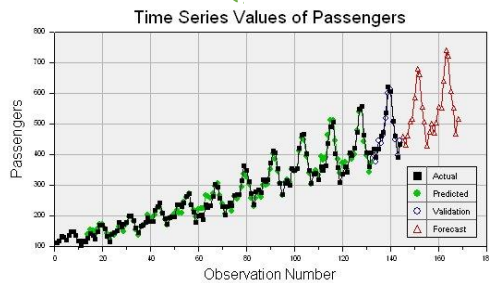
Survival Analysis

- **Sets of methods for analyzing data where the dependent variable is the period of time until an event of interest occurs.**
- **Churn and retention rates** for customers is an ideal scenario to use survival analysis.

Types of Analyses

Time Series

Used to study the consequences of measurements of variable or variables. The measurements are often made at regular time intervals.



Stepwise

Forward stepwise selection starts with no variables in model initially, variables are added into the model one by one.

Backward stepwise selection starts with all of the variables in the initial model, variables are removed from model one by one.

Combined stepwise selection starts with forward stepwise then changes to backward stepwise, process alternates between forward and backward until no variables can be removed or added.

MOSIAC

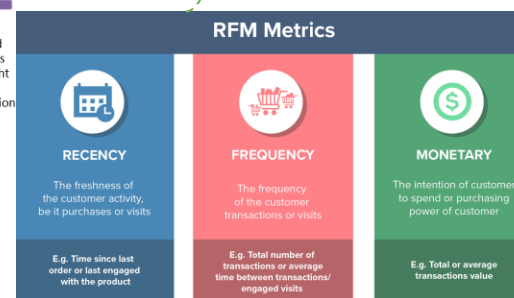
Geodemographic topology based on sociodemographic data, lifestyles, behavior and preferences.



RFM

Recency, Frequency, Monetary value.

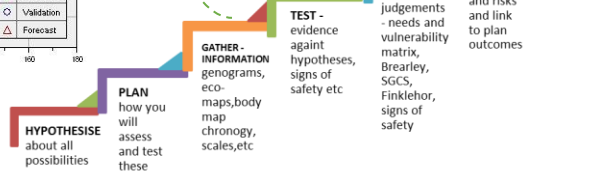
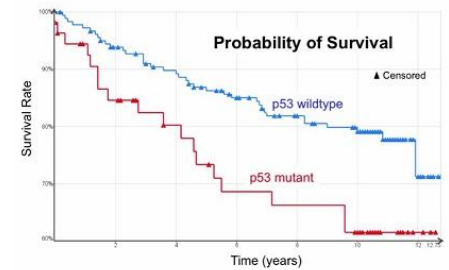
Analyses where the product was purchased? When was it purchased? How was it paid for? How much was purchased? What was purchased?



Survival

Sets of methods for analyzing data where the dependent variable is the period of time until an event of interest occurs.

Churn and retention rates for customers is an ideal scenario





Reduce Processing Times

- Work on **structured files** (SAS, SPSS, DB2, etc.) rather than flat files.
- Limit analyses to the lines and variables relevant to the current process.
- Recode the variables and make them smaller by using formats.
- Create **Booleans** such as alphanumeric variables of length 1, rather than numerical variables.
- Clearly define the length of the variables used, limiting it to the minimum possible.
- Remove intermediate files which are no longer required.
- Keep enough free space on the hard disk.
- Defragment the hard disk if necessary.
- Do not place the analyzed file or the temporary workspace on a remote network since network latency and speed will become an issue.
- **Increase the amount of RAM.**

Reduce Processing Times – SAS

- Use **KEEP** and **DROP** commands to analyze only the relevant variables.
- Use the **LENGTH** command to clearly define the length of the variables used.
- Use the **PROC DATABASES LIB = WORK KILL NOLIST** command to clear out the temp WORK directory often since it is not automatically purged until the end of the SAS session.
- Use **BY** command instead of CLASS in the MEANS procedure.
- **Create index on variables used at least 3 times** in a WHERE or BY filter.
- Use **COMPRESS = YES** command to reduce hard disk space occupied by file by removing all blank characters and spaces in data set.
- For copying tables, use **PROC COPY** or **PROC DATASETS** rather than a DATA SET step.
- Use **TAGSORT** option when sorting a large table.
- Use the **PRESORTED** option to sort the table if it has not been done already.

Check your knowledge:	
An analyst was given a table that contained unstructured data of keywords. He/she needs to identify the most frequently cited words in the documentation. Which method should the analyst use for classification?	Clustering
A data analyst wants to reduce the dimensionality of the text from a set of web pages. Which method should the data analyst apply to the dataset?	Kohonen maps
Methods in which the variables in $\epsilon [0,1]$ must be transformed.	Neural networks
Detects the two-way interactions between tables.	Decision Tree
Identifies hidden interconnected relationships	Neural Networks
Large volumes of data distilled into homogenous group	Cluster analysis
Which algorithm is both a prediction model and a classification model?	Decision Trees
Which type of analysis is used to determine the distribution of purchases during a given period of time	RFM
A retailer is looking for interesting and recurring patterns in data that will be used for targeted marketing. The retailer has given a data analyst a large list of transactions, with data on what customers purchased during each visit to the store. Which data mining method should the data analyst use in order to accomplish this task?	Association Rules
A data analyst is trying to determine how many home runs, to the exact number a player will hit based on the player's home run total from the previous year. Which method should be used?	Linear Regression
A data analyst is choosing a method to use on a dataset and needs the following capabilities for a project: 1) Capacity to process the data within a reasonable period 2) Ability to handle the possibility of incomplete and heterogeneous data that may not be numeric. Which method can the data analyst use?	DISQUAL
May need a large number of units in its hidden layer, which increases execution time of network without always yielding perfect modelling of complex structures and irregular data.	Radial Basis Function Networks.
Why might an online retailer mine the order history of its customers?	To source new products.
Which technique should be used to discover links between Age and Income?	Pearson Correlation
Which method can a data analyst use to identify the links between malignancy and nodule color variation?	Cramer's V

TYPES OF VARIABLES	
Race Number	Nominal
Place	Ordinal
Body Temp	Interval
Weight	Ratio

To do:

A data analyst is using analytical CRM to extract, store, analyze, and output relevant customer information. What is the first step within the analytical CRM phase that this analyst will be performing?

- Combining a customer's records to develop a holistic view.

Which feature of application development is unique to data mining?

- The development phase cannot be completed in the absence of data.

After performing a normality test on a dataset, results show the null hypothesis should be rejected.

Which type of test should be performed to analyze the data?

- Non-parametric test

A data analyst is looking at continuous independent variables. The data analyst discovers both hypotheses of normality and homoscedasticity are not satisfied. The variables in question consist of three or more categories.

Which test is appropriate for this non-parametric scenario?

- Kruskal-Wallis

Statistical tests applied to a dataset reveal the dataset exhibits non-normality and heteroscedasticity.

Which comparison test should be used for three or more samples?

- Kruskal-Wallis

Use the given dataset to answer the following question:

Name	Age	Gender	Income
Alice	23	F	\$40,000
Bob	55	M	\$50,000
Charlie	22	M	\$20,000
Dana	48	F	\$250,000
Ethan	34	M	\$110,000
Eric	55	M	\$5,000,000
Roberta	30	F	\$2,500
Paulette	42	F	\$90,000

Use the given dataset to answer the following question:

	Lives in urban area	Does not live in urban area	Total
Likes sushi	950	100	1050
Does not like sushi	50	1900	1950
Total	1000	2000	

Which test should be used to identify a link between the variables?

Cramer's V

Use the given table to answer the following question:

Race Number	Place	Body Temp	Weight
147	1	98.9	140
764	3	100.3	205
399	42	100.0	115
004	16	99.5	165

Match the column name to the type of data it contains.

Answer options may be used more than once or not at all. Select your answer from the pull-down list.

		YOUR ANSWER	CORRECT ANSWER
✓	Race Number	Nominal	Nominal
✓	Place	Ordinal	Ordinal
✓	Body Temp	Interval	Interval
✓	Weight	Ratio	Ratio

Which statement characterizes statistical software versus data mining software?

Statistical software: SAS/STAT is an example software package.

What are two characteristics of the R statistical software?

Free to redistribute and modify the code.

New packages are available quickly

A data analyst needs to analyze the churn rate (customer retention) and the time of possible churn of customers for a local wireless company.

Which method should be used?

Survival analysis

Which algorithm is both a prediction model and a classification model?

Decision tree

Which function is considered data preparation?

File handling

Which two functions are data preparation functions?

Transformation of variables

Analysis and imputation of missing values

43.

A data scientist is reading an extremely large SAS dataset.

Which two commands should the data scientist use to decrease processing time?

Data Table1(Compress=YES);

Data Data1(keep= total_units price customer_id);

Which two methods should a data analyst use to reduce processing time when working in SAS?

Create Booleans as alphanumeric variables of length 1.

Increase the amount of RAM.

A fast-growing international insurance company with 10 million customers wants to improve the data analytics methods it is using in SAS.

Which solution should be utilized to reduce data mining processing time?

Use structured files.

Which two changes can be made to code in SAS in order to reduce processing time?

Use BY rather than CLASS in the MEANS procedure.

Use PROC COPY or PROC DATASETS rather than a DATA SET step.

A data analyst wrote the following code:

```
MyData <- read.table(file, header = TRUE, sep = “,”)
```

Which software did the data analyst use?

R

Use the given characteristics of a data mining and statistical analysis software to answer the following question:

- 1. It is based on the same language as S-PLUS.**
- 2. It is programmable, so users can easily create a new function.**
- 3. It has a console for command, data editor command window, graphics window, and program editor window.**

4. Its source code is available.

Which software is described?

R

A retailer is looking for interesting and recurring patterns in data that will be used for targeted marketing. The retailer has given a data analyst a large list of transactions, with data on what customers purchased during each visit to the store.

Which data mining method should the data analyst use in order to accomplish this task?

Association rules

A data analyst is trying to determine how many home runs, to the exact number, a player will hit based on the player's home run total from the previous year.

Which method should be used?

Linear regression

An automobile manufacturer has obtained access to customer-related data that was previously unavailable.

Which method should the manufacturer use to perform descriptive data mining?

Parametric or semi-parametric models

A data analyst has been tasked by a pizza company to provide recommendations for three new restaurants. The best indication of success is based on the population of a surrounding area.

Which descriptive data mining method should the data analyst use to provide the recommendations?

Clustering

A marketing research team is mining customer demographic data for segmentation purposes. The team's data analyst wants to apply the hierarchical clustering method, but researchers are reluctant to use the method.

What might be the disadvantage of using the hierarchical clustering method in this scenario when restrictive assumptions about the problem to be solved are absent?

The model at level n will be decided by clustering at level $n-1$.

A data analyst has performed the following on a dataset:

Data preparation techniques

Exploratory data analysis

Identification of the dependent and independent variables

The single dependent variable is quantitative, and the single independent variable is qualitative.

Which data mining method should the data analyst be using?

Decision trees

A data analyst is choosing a method to use on a dataset and needs the following capabilities for a project:

1) Capacity to process the data within a reasonable period 2) Ability to handle the possibility of incomplete and heterogeneous data that may not be numeric
Which method can the data analyst use?

DISQUAL

Which technique should a data scientist use to predict the unknown probability of a white marble, given the known probability of red and green marbles as shown in the graphic?

Naïve Bayes

55.

A researcher has a data set containing socio-demographic data about study participants.

Use the given sample of the data to answer the following question:

Participant	Eye Color	Gender	Height	Type
J	Blue	M	5'5"	B
C		F	5'6"	B
M	Brown	M	6'0"	C
L	Blue	M	5'9"	C
N	Blue	F	5'8"	A
F	Brown	F	6'3"	F
Z	Green	M	5'8"	D
...

Knowledge about which characteristic is required in order to choose a data mining method for this scenario?

If Type is Ordinal

The following table contains an example of unstructured data of keywords, using content analytics with ranking:

urban, resource, public, animal, planning, ecological, sustainanability, residents	56
tolerance,accepatable,latent,statistic,tolerances,toleranc,metrology,statis	52
prize,award,medal,awarded,recipients,recipient,achievement,outstanding	51
hydroxide,electrolytic,eletrolysi,electrolysis,sodium,calcium,electrlyte	50
regression,guage,mile,mole,customary,are,correlation,error,correlat,var	49

A data analyst needs to identify the most frequently cited words in the documentation and classify them into groups.

Which method should the analyst use for classification?

Clustering

A data analyst wants to reduce the dimensionality of the text from a set of web pages.

Which method should the data analyst apply to the dataset?

Kohonen Maps

Use the given dataset to answer the following question:

	idname	location	income	gender	satisfaction
1	Alice	Alpharetta, GA	80,000	F	low
2	Bob	Boston, MA	110,000	M	low
3	Carol	Chicago, IL	70,000	F	low
4	David	Dallas, TX		M	low
5	Edward		500,000	M	high
6	Frank	Fort Laramie, CO	30,000	F	low

Which statement is valid in describing a method and the data it needs to perform data mining?

If no values are missing, linear regression may be performed using an “income” column

Match each data mining method to its characteristics of processing heterogeneous or incomplete data.

Answer options may be used more than once or not at all. Select your answer from the pull-down list.

		YOUR ANSWER	CORRECT ANSWER
✓	Neural networks perceptrons.	The variables in $\epsilon [0,1]$ must be transformed.	The variables in $\epsilon [0,1]$ must be transformed.
✓	Radial basis function networks.	The variables in $\epsilon [0,1]$ must be transformed.	The variables in $\epsilon [0,1]$ must be transformed.
✓	Neural networks (Kohonen).	The variables in $\epsilon [0,1]$ must be transformed.	The variables in $\epsilon [0,1]$ must be transformed.


	YOUR ANSWER	CORRECT ANSWER
Linear regression.	The variables in ϵ [0,1] must be transformed.	Numerical variables and variables without missing values.
Moving centres method and its variants.	The variables in ϵ [0,1] must be transformed.	Numerical variables and variables without missing values.

In a recent poll, the responses of the respondents were mapped with the census data available on a government portal.

Which type of predictive data mining algorithm is this?

Correspondence Analysis (CA)

Answer options may be used more than once or not at all. Select your answer from the pull-down list.

	YOUR ANSWER	CORRECT ANSWER
 Detects the two-way interactions between tables	Decision tree	Decision tree
Identifies hidden interconnected relationships	Cluster analysis	Neural networks
Large volumes of data distilled into homogeneous group	Decision tree	Cluster analysis
Marriage between lexicometry and data mining	Cluster analysis	Text mining

Which two modeling algorithms are used in data mining?

Neural networks perceptrons

Decision trees

A client requests an analysis of the reviews posted for one of the IT products that was launched last year. Data for all of the client's products was downloaded in Excel, and each word was separated. Sentiment categories were designed based on word counts. However, the sentiment model is not working on new data.

What could be the reason for this failure?

Grouping of words was not performed

A company that creates English speech recognition software would like to create automatic completion functionality for healthcare professionals.

Which two tools are required for this application?

Semantic dictionary

Syntactic analyzer

Which two sources of data are suitable for the purpose of providing a more personalized online experience?

Personal identification of user

Website cookies

A data analyst wants to determine what percentage of users that visit a web page for a new movie also view the movie stars' web pages.

What are two disadvantages of using cookies to retrieve the data in this study?

Privacy settings may block data transmission.

This method identifies the computer, but not unique users.

