Data Mining: On what kinds of Data?

- Database-oriented data sets and applications.
  - Relational database, data warehouse, transactional database.
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - structure data, graphs, social networks and linked data.
  - object-relational databases.
  - Heterogeneous databases and legacy databases.
  - spatial data and spatio-temporal data.
  - Multimedia database
  - Text databases
  - The World-Wide Web.

A Database system also called a database management system consist of a collection of interrelated data known as data base. A set of software programs to manage and access data. A software program mechanism database structure & data storage for specifying and managing concurrent, shared or distributed data access.

for ensuring consistency & security of the information store despite system crashes or attempts of unauthorized access.

- Relational data can be accessed by database queries written in a relational query language or with the assistance of data

Data Warehouse-

A data ware house is a repository of info collected from multiple sources, stored under a unified schema and usually residing at a single site.

- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.

- It is usually a modeled by a multidimensional data structure called a cube. In which each dimension corresponds to a attribute or set of attribute in a schema. And each cell stores the value of some aggregate measure such as count or sum.

- A data cube provides a multidensional view of data and allows the a precomputation and fast access of summarized data.

Transactional data -

Each record in a transactional database captures a transaction such as a customer's purchase, a flight booking and users click on a webpage.

- A transaction typically includes a unique transaction identity number and a list of the items making up the transaction such as the items purchase in the transaction.

- A transactional database may have additional tables which cont other information related to the transaction such as item descrip info about the sales person, the branch or so on.

Predictive Analytics - Uses historical data, statistical methods and ML techniques to predict future outcomes.

- While data mining focuses on discovering patterns, predictive analytics leverages those patterns to make informed decisions.

The techniques are -

i) Predict continuous values - i.e house prices stock market trends.
ii) classification, ~~regress~~
iii) ~~Cor~~ Regression
~~iv) Predictive analytics. Time series forecasting.~~

Time Series forecasting Predicts sequencial data, i.e weather forecasting and demand prediction.

Predictive analytics - Eg: KEI industries limited, TVS motor company, e-bussines firms like flipkart, paytm, zomato, make my trip, ola are using predictive analytic function.

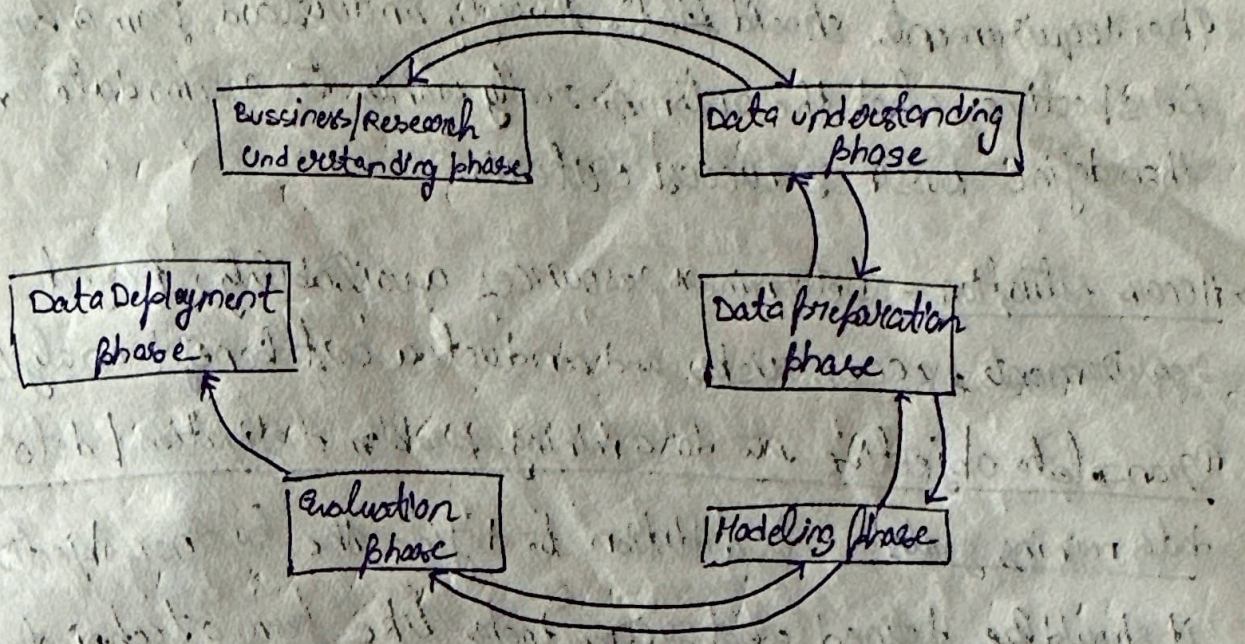
Wanted: Data Miners

- We are inundated with data in most fields, but...
- They are not trained human analyst available who are skilled to convert the data into knowledge.

Factors -

- Explosive growth in data collection, as in supermarket scanners.
- Storing the data in data warehouses.
- Increased access to data from web navigation and intranets.
- Competitive pressure to increase market share in globalized economy.
- Growth of computation power and storage capacity.

Cross Industry Standard Process: CRISP-DM



- CRISP-DM developed in 1996.
- Fits data mining into the general problem-solving strategy of bussiness/research unit.
- Contributors include Daimler Chrysler, SPSS and NCR.
- Industry, tool and application neutral.
- Non-proprietary and freely available.
- Data mining projects follow iterative, adaptive life cycle consisting of 6 phases.
- Phase sequences are adaptive.

Bussiness Understanding - has 4 phases.

- Any good project starts with a deep understanding of the customer need.
- Data mining projects are no exception and CRISP-DM recognizes this.
- The bussiness understanding phase focuses on understanding the objectives and requirements of the project.
- Aside from the 3rd class the 3 other tasks in these phase are foundational project management activities that are universal to most projects.

Phases-

1) Define project requirements & objectives - The requirements should first, thrugly understood from a bussiness perspective, what the customer really wants to accomodate and then define bussiness success criteria.

2) Access situation - Determine resources availability, project requirements, access risks and conduct a cost benefit analysis.

3) Translate objective into data mining problem defination / determine data mining goals - In addition to define the bussiness objectives, it should be defined what success looks like from a technical data mining strategies.

4) Prepare preminary strategy to meet objectives / produce a project plan - In this stage select technologies and tools and prepare a preminary strategies for achieving this objectives.

2) Data Understanding phase -
Again to the foundation of bussiness understanding drive the focus to identify and collect the dataset that can help to accomplish the project goals. This phase also has 4 tasks -

i) Collect Initial Data - Acquire the necessary data, load it into your analysis tool.

ii) Describe data - Examine the data and document its surface properties like data format, no. of records or field identities.

iii) Explore data - They differ into the data, query it, visualize it and identify relationships among the data.

iv) Verify data quality - How big or dirty is the data, document any quality issues.

• Select cases and variables appropriate for analysis.

Data Preparation phase -

This phase referred as data munging. Prepares the final dataset for modelling.

5 phases are -

i) Select data - Determine which datasets will be used and documents reasons for inclusion or exclusion.

ii) Clean data - Generally this the lengthiest task. Without it you will likely fall victim to garbage in, garbage out.

• A common practice during this task is to correct input, or remove an remote erroneous value.

iii) Construct data - Derive new attributes that will be helpful, i.e derive somebody's age from its date of birth.

iv) Integrate data - Create new datasets, combining data from multiple sources.

v) format data - Reformat data are necessary.

Ex - You might convert string values that store numbers to numeric values so that you can perform mathematical operations.

~~vi) flo~~

Data modelling phase - This is the most exciting phase and shortest phase. Here, you will like build and access various models based on several different modelling techniques.

- This phase has 4 tasks -

i) Select modelling technique - Determine which algorithm should try.

Eg, - Regression, neural network.

ii) Generate test design - Pending your modelling approach, they might need to split the data into training, testing and validation sets.

iii) Build model - several different techniques may be applied for same data mining project.

iv) Assess model - May require grouping back to data preparation phase in order to bring the form of the data into line with specific requirements of a particular data mining technique.

- Generally multiple models are competing against eachother and the data scientist needs to interpret the model, based on domain knowledge, the predefined success criteria and the test design. Although crisp DM guide ~~to iterate~~ iterate model building and assessment until you strongly believe that you have found the best model.
- In practice, teams should continue iterating until they found a good model. Proceed through the crisp DM life ~~cycle~~ cycle then further improve the model in ~~the~~ future iterations.

9/9/25

Evaluation phase - Whereas the assess ~~model~~ model task ~~of the modelling~~ of the modelling phase focuses on technical model assessment, the evaluation phase looks more broadly at which model best meets the bussiness and what to do next.

This phase has 3 tasks -

~~i) Evaluate results~~

i) Evaluate Results - Do the models meet the bussiness criteria, which one should be we approve for the ~~approve~~ bussiness.

ii) Review Process - Review the work accomplished was anything overlooked where all steps properly executed, & summarize findings and correct anything if needed.

iii) Determine next steps - Based on the previous 2 tasks, determine whether to proceed to deployment iterate further or proceed to new project.

- In this phase models are evaluated for quality and effectiveness before deployment.
- In this phase also determine whether the model in fact achieves the objectives set for it in phase 1.
- Here some important factors of the business or research problems has not been sufficiently accounted for.
- Finally come to a decision regarding the use of data mining results.

Deployment phase - Depending on the requirements the deployment phase can be as simple as generating a report or as complex as implementing a repeatative data mining process across the internet.

- A model is not particularly useful unless the customer can access its results.
- The complexity of these phase varies widely.
- This final phase has 4 tasks -

i) Plan Deployment - Develop and document a plan for deploying a model.

ii) Plan monitoring and maintainence - Develop a thorough monitoring and maintainence plan to avoid issues during the operation phase or post project phase of a model, produce final report the project team documents a summary of the project which might include a final presentation of data mining results.

iii) Review project - Conduct a project retrospective about what went well, what could have been better and how to improve in the future.

- Organisations work might not end there as a project framework CRISP-DM doesn't outline what to do after the project.

11. But if the model is going to production be sure to maintain the model in production. Constant monitoring and occational model tuning is often required.

10/9

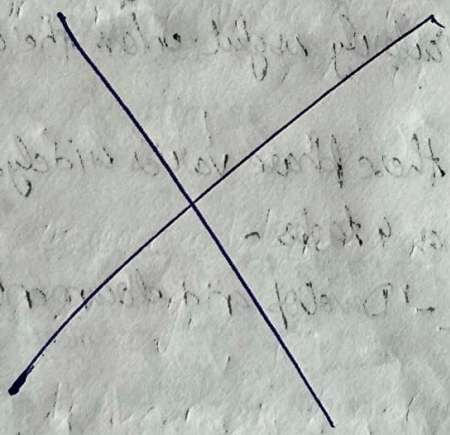
Fallacies of Data Mining -

i) There are data mining tools that we can use on our data repositories and find answers to our problems.

ii) Data mining is autonomous requiring little or low human oversight.

iii) The return rates vary depending on the startup cost, personal cost, data warehousing preparation cost.

iv) Data mining software packages are very intuitive and easy to use.



Reality -

i) There are no automated data mining tools which will mechanically solve all the problems while you wait rather data mining is a process.

CRISP-DM is one method for fitting the data mining process into the overall business in research plan of action.

ii) Data mining is not magic without skilled human supervision, blind use of data mining software will only provide you the wrong answer or a to the wrong question applied to wrong type of data. Further the wrong analysis is worse than no analysis b/o it leads to Bobby Recognization that may turn out to expensive failure.

Even after model is deployed, introduction of new data often requires updating the model. A data mining pays for itself quite quickly.

iii) Easy of use varies across project.

iv) Data mining softwares can't solve all the problems ~~while we sit Back~~. The algorithms may require specific data formats which may require substantial preprocessing.

v) Data analyst must combine domain knowledge with an analytical mind and familiarity with the overall business or research model.

— There are 3 other common falacies —

v) The data mining will identify the causes of business or research problems.

Reality

vi) The knowledge discovery process will help you to uncover patterns or behaviours. Again it is upto the humans to identify the cause.

Falacy

vi) Data mining will automatically clean up our messy database

Reality

vii) As a preliminary phase of data mining process, data preparation often deals with data that has not been examine or used in years. Therefore, organizations beginning a new data mining operation will often be confronted with the problem of data that has been lying around for years.

Falacy

viii) Data mining always provides positive results. There is no guarantee of positive results while mining data for actionable language. When data mining software used properly by people who understands the models involved the data requirements and overall project objectives, data mining can provide actionable and highly profitable results.

Predictive Analysis (PA) -

Predictive analysis is an ~~powerful tool for~~ advanced analytical discipline that leverages historical data, statistical algorithms and ML techniques to predict the likelihood of future outcomes.

- It grows beyond descriptive analytics by not only explaining what has happened but also ~~predicting~~ forecasting what is likely to happen.
- It's a powerful tool for future oriented decision making built upon the foundational data mining.
- While data mining uncovers hidden structures, predictive analytics translates these insights into actionable forecast making it crucial for industries, healthcare, finance, retail and more.
- Predictive analysis provides organisations with foresight to include decision making, optimized operations and mitigate risks.

Examples -

1) Retail
   - Amazon ~~analyzing~~ recommendation system suggest products by analyzing customers purchase history.
   - Supermarket use predictive analysis model to forecast seasonal demand and optimize stock labels.
   - Retailers predicts customers lifetime value to identify which customers and worth targeted promotions.

2) Healthcare - Hospital predict special recommendation using electronic ~~EHR~~ health records (EHR).
   - Predictive model identify high risk patient by chronic disease.

~~Pharma~~
   - Pharmaceutical use predictive analysis to predict drug trial outcome ~~outcomes~~

and success rates.

3) Banking and finance -

Credit care scoring - Banks use this model to access the probability of loan default.

Fraud Detection system - analyse spending patterns to flag unusual transaction.

Predictive model estimates future stock prices and guide investment strategies.

4) Telecommunication - Telecom service providers

Customer churn rate - Service providers predicts customer churns.

- Predictive models helps in network optimisation and maintainence.
- Telecom companies personalised offers based on predicted usage patterns.

5) Manufacture - Predictive maintainence anticipate machine failures, reducing downtown.

Models optimised supply chain operations by forecasting raw material demand.

6) Sports - PA forecasts players performance and injury likelihood.

- Teams use models to strategise game outcomes.
- Sports companies spreading ticket sales and optimize marketing campaign.

Application predictive analytics

Business decision making - Demand forecasting, dynamic pricing, Inventory optimigation.

Risk management - Predict. Credit risk analysis, Insurance

claims prediction, fraud prediction.

Customer relationship management - Personalized campaigns, churn prediction.

~~Logistic management~~ -

Healthcare - Early diagnosis
- Patient ~~mon~~ monitoring
- Personalized ~~treatment~~ recommendation.

Public sector - Crime prevention, ~~traffic~~ traffic management, disaster preparedness. ~~education predicting student dropouts, personalised learning, performance forecasting.~~

education - Predicting students dropouts, ~~personalised~~ personalised learning, performance forecasting, education recommendation learning.

Comparision : Data mining vs Predictive Analysis

Aspect	Datamining	Predictive Analytics
Defn:	Process of discovering patterns, correlation and useful info from large data sets.	Process of using statistical ML. AI models to predict future outcomes based on past data.
Focus:	What's hidden in the data (finding unknown patterns and relationships).	What will happen next. (forecasting future events).
Approch:	explanatory - knowledge discovery.	Action oriented decision making based on forecasts.
Techniques:	clustering classifications, association rule mining anomally detection.	Regression analysis, Time series forecasting decision trees, neural nw, ensemble.

Outcome:	Insights, rules, patterns (eg: customers who buy bread often buy butter)	Predictions and probabilities (eg: customer X has an 80% chance of buying butter tomorrow)
Scope:	Broad or (can be descriptive, diagnostics or predictive)	Narrow: specifically focused on predictions and forecasting

15/9

Clustering - is a technique used to group similar data points together based on their features and characteristics.

- It can also be refer to as a process of grouping a set of objects so that objects in the same group are more similar to ~~objects~~ each other than those in other groups. It's an unsupervised learning technique that aims to identify similarities and patterns in a data set.
- Clustering algorithms typically require defining the no. of clusters, similarity measures and clustering methods. These algorithms aim to group data points together any way that maximizes similarity within the groups and minimizes similarities b/w similar groups.

Cluster - group of data points with similar characteristics or features is called cluster.

Summarization - is the process of condensing large datasets into a certain more understandable format that highlights key patterns, trends and relationships.

- It's the process of generating a compact and informative representation of a large dataset.
- It's a foundational step in data preprocessing that makes data more ~~manageable~~ manageable and understandable for analysis.
- The goal is to reveal underlying patterns and trends efficiently without loosing critical information.

Types of data summarization -

i) Descriptive statistics - uses statistical measures to describe the main features of numerical data.

Eg: Measures of central tendency → mean, median, mode.
Measures of dispersion → standard deviation, range, variance.

ii) Aggregation - Involves combining data into simpler summaries, i.e sum, average, group count.

iii) Dimensionality reduction - Reduces no. of variables in a dataset while preserving essential information.

iv) Clustering - Grouping of similar data points into clusters.

v) Sampling - selects a representative subset of the data.

vi) Take summarization - Uses Natural language processing to condense large text documents into certain coherent summaries.

vii) Extractive summarization - selects most important sentences and phrases of the original text.

viii) Abstractive summarization - Generate new sentences to capture essence of original text.

ix) Data visualization - It represents summaries graphically, making patterns and trends more intuitive and easily identifiable.
Eg - Histogram, bar chart, barcharts.

Association - Association rule mining finds interesting association and relationships among large datasets of data atoms. This rule shows how frequently an item set occurs in a transaction.

Eg: Market basket analysis. (Bread with butter)

Association rule - It is a data mining technique that identifies association b/w items frequently purchased together.

- Association rule mining aims to find if there patterns often ~~express~~ expressed as association rules which indicates how frequently certain items occur together.

16/9

Chapter 2 -

Descriptive statistics - Refers to method for summarizing and organizing information in a dataset.

- The entities for which info is collected are called the **elements**.
- Elements are also called ~~cases~~ cases or **subjects**.

Variable - is a characteristic of an element which takes an different values ~~of elements~~ for different element.

- Other names of variables → attributes, features, dimensions.

Applicant, Marital Status, Mortgage, Income, Rank,

The set of variable values for a particular element is an **observation**.

- Other names of observation are → records, samples, objects, instances, subjects.

Applicant	Marital Status	Mortgage	Income (£)	Rank	Year	Risk
1	single	Y	38,000	2	2009	Good
2	Married	Y	32,000	7	2010	Good
3	Other	N	25,000	9	2011	Good
4	Other	N	36,000	3	2009	Good
5	other	Y	24,000	4	2010	Good
6	other	N	25,100	10	2008	Bad

Variables can be 2 types

i) Qualitative

ii) Quantitative (Numerical variables)

i) **Qualitative** - enables the elements to be classified or categorised acc to some characteristics.

- Are also called as **categorical variable**.

Eg: Rank, Marital status

ii) Quantitative variables - An arithmetic operations can be performed on it. Also called as Numerical values.

Eg: Income, Year etc.

Data - can be classified in 4 ~~types~~ levels of measurement

i) ~~Nominal~~ Nominal
ii) Ordinal
iii) Interval
iv) Ratio.

Ordinal, Interval, Ratio → Qualitative

Nominal - means relating to name. The value of a nominal attributes are symbols, names, things. Each value represents some kind of category or state and so nominal attributes are also referred to as categorical.

- The values don't have any meaningful order. It's also known as ~~enum~~ enumeration.

Eg: Ranks, status ~~etc~~, branch etc

- It is possible to represent ~~nominal~~ nominal attributes by numbers.
- However no.s are not intended to be used quantitatively i.e mathematical operations and values of ~~nominal~~ nominal attribute are not meaningful.
- Mean or median doesn't work on ~~nominal~~ nominal attribute.
- The most occuring value can be considered as mode of particular attribute.

Binary attribute - Male-female, test for HIV.

↓ symmetric ↓ Asymetric

Ordinal attribute - with possible value that have a meaningful order or ranking among them. But the magnitude b/w successive values is not ~~known.~~ known.

size of popcorn small, medium, large.

- Ordinal attribute are useful registering subjective assesment that can't be measured objectively.
- Are often used in surveys or ratings.

Numeric Attribute - is quantitative, i.e it is a measurable quantity. represented in integers, real values.

- Can be interval scaled or ratio scaled.

- Interval scaled attributes - are measured on a scale of equal size units. The values of interval scaled attributes have order. can be +ve, 0 or -ve.
  - In addition to provide ranking of values, such attribute allows us to compare and quantify the difference between values.
  Eg: Temperature attribute is an interval scale, calender dates.
  - Mean value can be computed in addition to median and mode of central tendency.

- Ratio scale attributes - is a numeric attribute with an inherent zero point, i.e a measurement in ratio scale we can speak of a value as being a multiple or ratio of another value.
  - In addition the values are ordered and we can also compute the difference between values as well as mean, median, mode.
  Eg: Experience, age etc.

Discrete vs Continuous attribute -

i) A discrete attribute has a finite or countable infinite set of values which may or may not be represented as integers.
   Eg: Hair colour, medical tests, drink size each have a finite no. of values and so are discrete.

An attribute is constantly infinite if the set of possible values is infinite. But the values can be put in one to one correspondence with natural numbers.

Eg: Customer id is constantly infinite i.e the no. of customers can grow up to infinite but in reality the actual set of values is countable.

ii) Even if a attribute is not discrete it is continious, i.e in practice real values are represented using a finite no. of digits.

iii) Continious attributes are typically represented as floating point variables.

Basic statistical description of data -

Predictor variables	Response variables
i) Independent variables	i) Dependent variables.
ii) A predictor variable is a variable whose value is used to help predict the value of the response variable.	ii) A response variable is a variable of interest whose value is presumably determined at least in part by the set of predictor variables
iii) The predictor variable in Table are all variable except risk.	iii) The response variable in Table is risk.

- It is used to identify properties of the data and highlight what value should be treated as noise or ~~out ware~~ outlayers.

Measure of central tendency - ~~Mean~~

* **Mean** - The most common and effective numeric measure for center of a set of data is the arithmetic mean.

Let, $x_1, x_2, \dots x_n$ be a set of N values for observations such as for some numeric attribute x like ~~salary~~ salary, the mean of set of values is

i) $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \dots x_n}{N}$

Eg: {2, 4, 6, 8, 10}, Mean = $\frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$

OR

$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1x_1 + w_2x_2 + \dots w_Nx_N}{w_1 + w_2 + \dots w_N}$ (Weighted avg)

- Sometimes each value "x_i" in a set may be associated with weight w_i for $i = 1$ to N.

Here, the weights reflects the significance, importance or occurrence frequency attached with their respective values. In this case we can compute the 2nd formula.

Advantages - Although mean is the single most useful quantity for describing a data set it's not always the best way of measuring the center of the data. A major problem with mean is its sensativity to outlayers.

Even a small no. of extreme values can affect the mean.

- To minimize the effect caused by extreme values, trimmed mean concept is used which is the mean obtain for chopping of values at high and low extremes.

* Median - Middle value when data is arranged in ascending or descending order.

Eg: {3, 5, 7, 9, 11} ⇒ Median = 7. → odd

{3, 5, 7, 9} ⇒ Median = (5+7)/2 = 6 → even.

- For skewed or asymmetric data a better measure of center of data is median. which is the middle value in a set of ordered data values.
- It's the value which separates the higher half of a dataset from the lower half.
- Generally it appear to numeric data, sometime it appears to ordinary data.
- It's expensive to compute when we have a large no. of observation.

$$\text{Median} = L_1 + \left(\frac{(N/2 - (\Sigma \text{freq})_l)}{\text{freq}_{\text{median}}}\right) \text{width}$$

$L_1 \rightarrow$ lower boundary of median interval

$N \rightarrow$ no. of values in entire dataset.

- sum of frequency of all intervals that are lower than the median interval
- freq median → frequency of median interval width, i.e width of median interval.

*Mode - The mode is another measure of central tendency. The mode for a set of data is the value that occur most frequently in the set therefore it can be determined for qualitative and quantitative attribute.

- It is possible for the greatest frequency to correspondence to several different values which results in more than 1 mode.
- Datasets with one, two or ~~three~~ three modes are respectively called as unimodal, bimodal and trimodal.
- In general a dataset with 2 or more modes is multimodal.
- At the other extreme if is data value occurs only once then there is no mode.
- For unimodal numeric data, that are moderately skewed the following relationship can be established.

$$\text{Mean} - \text{mode} = 3(\text{Mean} - \text{Median})$$

This implies that the mode for unimodel frequency curves that are moderately skewed can easily be approximated if the mean and median values are ~~known~~ known.

*Midrange - Avg of largest and smallest value in ~~data~~ set.

Measuring the dispersion of data - captures the degree to which values differ from each other and from the center.

- the measures of variability:
  - The Range
  - Quartiles
  - Interquartile Range
  - Five-number summary
  - Box Plot
  - Z score
  - The variance
  - The Standard Deviation

Range - The range is the difference b/w the max and min values in the dataset.

Range = max value - min value

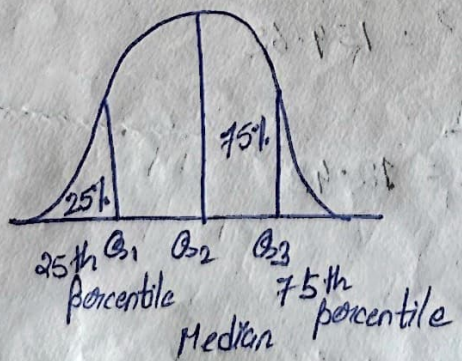
Limitations

- very sensitive to outliers (extreme values)
- Ignores distribution of values between extremes.

Quantiles - are points taken at regular intervals of a data distribution dividing it into essentially equal ~~sides~~ size consecutive sets.

Let $x_1, x_2, \ldots x_n$ be a set of observations for some numeric attribute "x".

Suppose the data for attribute x are sorted in increasing numeric order we can pick certain data points to split the data distribution into equal size consecutive sets. These data points are called Quantiles.



$Q_1 = \frac{1}{4}(n+1)$

$Q_2 = \frac{1}{2}(n+1)$

$Q_3 = \frac{3}{4}(n+1)$

Q) Suppose the data for analysis includes the attributes x, the edge values for data tuples are in increasing order. {13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}.

a) What is mean and median of data?

a) What is mode of the data? Comment on datas modality.

a) Mid range of data? ~~find~~.

a) Find 1st and 3rd Quartile of data? Find IQR

a) Give 5 nos. of data.

a) Show box plot of the data.

Sol: Mean = ~~29.69~~ $29.96 \approx 30$

Median = 25 (14th one)

Mode = 25 and 35 (It is a bi modal data)

Mid range = $\frac{70+13}{2} = 41.5$

IQR

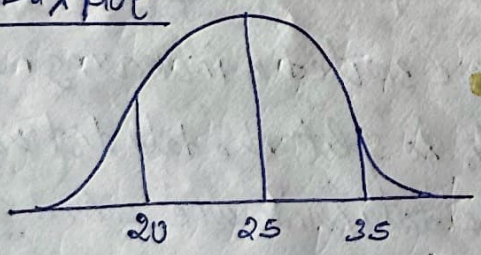
$Q_1 = \frac{1}{4}(n+1) = 1/4 \times 28 = 7 \rightarrow 20$

$Q_2 = \frac{1}{2}(n+1) = 1/2 \times 28 = 14 \rightarrow 25$

$Q_3 = \frac{3}{4}(n+1) = \frac{3}{4} \times 28 = 21 \rightarrow 35$

IQR = 35 - 20 = 15

~~Plot~~ Box plot



variance - $\sigma^2 = \frac{1}{27}\sum_{i=1}^{27} x_i - \bar{x}^2 = 154.62$

Standard deviation - $\sigma = \sqrt{154.62} = 12.43$

Variance and standard deviation -

These are majors of data dispersion. They indicate how spread out a data is. A lower standard deviation means that the data observation tends to be very close to the mean while in high standard deviation indicate that the data are spread out over a large range of values.

Variance of 'N' observations i.e $x_1, x_2, \ldots, x_n$ for a numeric attribute 'x' is $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2$

$\bar{x}$ - Mean value of observations

Standard deviation σ of the observations is the square root of the variance (σ^2).

σ = Measures spread about the mean and should be considered only when the mean is chosen as measure of center.

If $\sigma = 0$, only when there is no spread when all observations of the same value.

2419

~~Data Preprocessing~~

→ Why we need data preprocessing?
→ Preprocessing steps
→ Data cleaning

~~Data cleaning~~ -

Q) Why do we need to preprocess the data?

Sol: Real world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size and they are likely origin from multiple heterogeneous sources.

- Low quality data will lead to the low quality mining results.
- Data preprocessing is needed to clean, organize and raw data into a usable format improving its quality, consistency which leads to better model performance, more insights and more reliable decision making.
- Preprocessing handles missing ~~values~~ values, removes noise and outlayers and scales features than ensuring algorithms to use the data effectively and avoid the garbage in and garbage out problem.

Garbage in - Garbage out problem - It's the concept of flood, Baised or poor information or input produces a ~~result~~ result of similar ~~quality~~ garbage quality. It's otherwise called as rubbish in and rubbish out.

- In simple words the quality of output is directly determined by the quality of its input.

Noise - refers to errors, inaccuracies or randomness in a data that lacks significant meaning or pattern or incorrect data entry or measurement imprecision.

Outlayers - Outlayers are data points that deviate significantly from the normal pattern of the dataset but not necessarily errors, ~~potentially~~ Potentially representing valid but unusual cases that can be valuable for analysis.

• Noise are detrimental or meaningless while outlayers are significant and potentially informative data points that require separate investigation.

~~Major task in d~~

Data quality - Many factors ~~comprehensting~~ comprising ~~comprising~~ data quality including accuracy, completeness, consistency, timeliness, believeability, interpretability Mainly 3 elements define the data quality.

- accuracy
- Completeness
- Consistency

Major task in data preprocessing - are data cleaning, data integration, data reduction and data transformation.

Data cleaning - routines work to clean the dataset filling in missing values, smoothing noisy data, identifying or removing outlayers and resolving inconsistencies.

Data integration - It involves integrating multiple databases, ~~datacubes~~ datacubes or files. In addition to data cleaning steps must be taken to avoid ~~reduni~~ redundancy during data integration.

• Addition data cleaning can be performed to detect and remove redundancies that may have resulted from data integration.

Data reduction - obtains a reduced representation of the dataset. i.e much smaller in volume. It produces the same analytical results.

• Data reduction strategies includes dimensionality reduction and numerosity reduction.

• In dimensionality reduction data encoding schemes are applied so as to obtain a reduced or compressed ~~represent~~ representation of original data.

eg: Principal component analysis, wavelet transformation.

• In numerosity reduction the data are ~~stored~~ replaced by alternative smaller representation using parametric models or non parametric models.

Data transformation - Discretization and concept hierarchy generalization are powerful tools for data mining in that they allow data mining at multiple abstraction levels. Normalization, data discretization, and concept hierarchy generation are forms of data transformation

Data cleaning -

① Missing values

a) Ignore the tuple
b) Fill in the missing value manually.
c) Use a global constant
d) Use a measure of central tendency.

Data transformation (continued) -

Data transformation operations are -

- Additional data preprocessing procedure that will contribute towards the success of.
- In summary, real world data tend to be dirty, incomplete and inconsistent data preprocessing techniques can improve quality thereby helping to improve accuracy and efficiency of a subsequent mining process.
- Data preprocessing is an important step in knowledge discovery process because quality decision must be based on quality data. Detecting data anomalies, rectifying them early and reducing data to normalized form lead to huge pay offs and decision making.

Data preprocessing -

→ Data cleaning - Real world data tend to be incomplete and inconsistent. Its duty includes filling missing values, smoothing while identifying outlayers and correct inconsistencies in data set.

• Missing values - It can be handled using following methods -

i) Ignore the tuple - This is usually done when the class label is missing. This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the % of missing values per attributes vary considerably. By ignoring the tuples, we don't make use of the remaining attribute values in a tuple. ~~Such data that it can~~.

Such data could be useful for the task in hand.

ii) Filling the missing values manually - In general this task is time consuming and may not be feasible even a large dataset with many missing values.

iii) Use a global constant to fill the missing values - ~~Replace~~ Replace all missing attributes values by the same constant such as a label like unknown or infinite. If missing values are replaced by ~~unknown~~ unknown then the mining program may mistakenly think that they form an interesting concept. Since they ~~are~~ all have values in common, that of unknown, though this method is simple it is not fullfilled.

iv) Use a measure of central tendency - For normal asymmetric data distribution, the mean can be used while skewed data distribution should use the median. ~~Use the~~.

v) Use the attribute mean and median for all attribute belonging to the same class as the given tuples.

For eg: classifying customers according to credit risk we may replace the missing values with the mean income value for customers in the same credit risk categories as that of the given tuple. If the data distribution for a given class is ~~so~~ skewed then median is a better choice.

- Use the most probable value to fill in the missing values - this may be determining with regression, using bayesian convolism or decision tree induction. It's a popular strategy in comparision to other methods, it uses the most information, from present data to predict missing values.

→ Smoothing noisy data - Noise is a random error or variance in a measured variable. Some basic statistical technique i.e box plots and scatter plots.

- Methods of data visualization can be used to identify outlayers which may represent noise.
- The noise can be removed by smoothing.

+ Binning - methods smooths a sorted data values by consulting its neighbour-hood i.e the values around it. The sorted values are distributed into a no. of buckets or bins.

- In order to smooth data, various methods can be applied which includes
  i) Smoothing by bin mean
  ii) Smoothing by bin median
  iii) Smoothing by bin boundaries
  iv) Smoothing by using regression lines
  v) Outlier analysis.

Smoothing using bin -

- Bin mean - In smoothing by bin mean, each value in a bin is replaced by mean value of the bin.

  Eg: The mean of values i.e 4, 8, 15 is 9.
  i.e, 9, 9, 9 (replaced)
  ∴ each original value is replaced by value 9.

- **Bin median** - Simply smoothing by bin median can be employed in which each bin value is replaced by the bin median.
- **Bin boundaries** - In smoothing by bin boundaries the min and max values in a given bin are identified as bin boundaries. Each bin value is then replaced by the closest boundary value. In general the larger the width, the greater effect of smoothing. So, bins may be equal width or equal depth.

Equal width or (distance) partition - It divides the range into 'N' interval of equal size. So, if 'A' and 'B' are lowest and highest values of attributes the width of interval will be $\frac{B-A}{N} = w$.
It is the most straightforward approach but outlayers may dominate presentation.

Equal depth (frequency) partition - It divides the range into 'N' intervals is containing approximately same no. of samples.

Good data scaling - Managing categorical attributes can be tricky.

Q1) Sorted data for price 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34.

Bin 1: 4, 8, 9, 15
Bin 2: 21, 21, 24, 25
Bin 3: 26, 28, 29, 34

Bin 1 Mean - 9, 9, 9, 9
Bin 2 Mean - 22·75, 22·75, 22·75, 22·75
Bin 3 Mean - 29·25, 29·25, 29·25, 29·25
Bin 1 Median - 8·5, 8·5, 8·5, 8·5

⋮

Bin 1 Boundary - 4, 4, 4, 15

⋮

Equal width Partitioning - L.V = 4
H.V = 34, Range = 34 - 4
Width of interval = $\frac{30}{3} = 10$
Bin 1 - 4, 14 ⇒ Bin 1 = 4, 8, 9
Bin 2 - 14, 24 ⇒ Bin 2 = 15, 21, 21
Bin 3 - 24, 34 ⇒ Bin 3 = 24, 25, 26, 28, 29, 34

Identifying misclassifications -

Origin

Level name	Counts
USA	1
France	1
US	156
Europe	46
Japan	51

- verify values valid and consistent
- count for USA + and France = 1?
- 2 records classified inconsistently with respect to origin of the manufacturer.
- Maintain consistency by the labelling USA to US and France to Europe.

Graphical methods for identifying outliers -

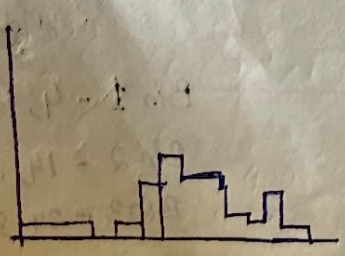
- Outliers are values that lie near extreme limits of data range.
- Identifying outlayer is imp because they may represent errors in data input also even if an outlayer is valid datapoint and not an error certain statistical methods are sensative to ~~present~~ presence the outlayers and may deliver unreliable results.
- Graphical methods for identifying ~~outlayers~~ outlayers includes histogram, box plots and scatter plots.

Box plots - outlayers are typically defined as any data points that fall below $Q_1 - 1.5\ IQR$ or above $Q_3 + 1.5\ IQR$.

IQR (Inter quartile Range) = $Q_3 - Q_1$.

Histogram - It is a graphical representation of the distribution of quantitative data displayed as a series of rectangular bars.

- It is also referred as bar charts.
- In this outlayers appear as isolated bars on either the extreme left or right side of the histogram, far from the main body of data that forms a typical distribution.



Scatterplot - It is a type of plot or mathematical diagram using cartesian coordinates to display values for typically 2 variables for a set of data.
- It uses dots to represent values for 2 different numeric variables.

Measures of central and spread - (Already done previously)

~~Data transformation~~

- Data integration - Integrating different sources of data.
  - It ~~is used~~ helps to reduce and avoid redundancies and inconsistencies in the resulting data sets which improves the accuracy and speed of subsequent data mining problem.
  - It occurs entity identification problem.
  - It occurs when we try to determine whether 2 or more records from different data sources refer to the same real world entity.

- Data transformation - It is the process of converting data from 1 format structure or value representation to another to ensure compatibility, consistency and better quality for analysis.
  - Some data mining algorithms adversely affected by differences in variable ranges.
  - Variables with greater ranges tend to have larger influence on ~~the~~ data model's result.
  - Therefore numeric fields values should be normalized.
  - Standardizes scale of effect each variables has on results.

Types -

1) Smoothing - Removes noise from data.
   Eg: Regression, Binning and

2) Aggregation - Summarizing data.
   Eg: ~~Del~~ Daily sales data aggregated to montly sales data.

3) Generalization - Replacing detailed data with higher level concepts.
   Eg: Age will be replaced with child, young, old.

4) Attribute construction - Creating new attributes from existing ones.

5) Normalization - ~~Rescaling~~ Rescaling values to a standard range i.e 0 - 1.

Min-max normalization - $X^*_{mm} = \frac{X - min(X)}{range(X)} = \frac{X - min(X)}{max(X) - min(X)}$

$midrange(X) = \frac{max(X) + min(X)}{2}$

Zscore standardization - $Zscore = \frac{X - mean(X)}{SD(X)}$

Decimal scaling - ensures that every normalized value lies Blw -1 to 1.

$X^*_{decimal} = \frac{X}{10^d}$, d represents the no. of digits in the data value with largest absolute valu

Eg: Min: $X^*_{decimal} = \frac{1613}{10^4} = 0.1613$

Max: $X^*_{decimal} = \frac{4997}{10^4} = 0.4997$

Q) Use the methods below to normalize the following group of data 200, 300, 400, 600, 1000.

i) Min-max normalization by setting min = 0, max = 1

ii) Zscore standardization

iii) Normalization by decimal scaling

Sol: i) $X^*_{mm} = \frac{X - min(X)}{range(X)} = \frac{200 - 200}{1000 - 200} = 0$

$X^*_{mm} = \frac{X - min(X)}{range(X)} = \frac{1000 - 200}{1000 - 200} = 1$

X	Normalization
200	0
300	0.125
400	~~0.375~~ 0.25
600	~~0.625~~ 0.50
1000	1

ii) Mean(X) = 500

SD(X) = 282.8

X	normalization
200	-1.06
300	-0.707
400	-0.35
600	0.35
1000	1.76

(iii)

X	Normalization
200	$200/10^3 = 0.2$
300	$300/10^3 = 0.3$
400	0.4
600	0.6
1000	0.1

Q2) Using the data for age (given in prev exercise) answer the following.

i) Use min-max normalization to transform the value 35 for age onto the range 0.0 to 1.0.

ii) Use z-score normalization to transform the value 35 for age.

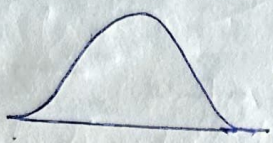
iii) Use normalization by decimal scaling to transform the value of 35 for age

Sol:

13/10

Transformations to achieve normality -

- The normal distribution is a continuous probability distribution commonly known as the bell curve, which is symmetric.
- It is centered at mean μ and has its spread determined by standard deviation σ (sigma).
- Here, $\mu = 0$ and $\sigma = 1$, known as standard normal distribution Z.



- Some data mining are statistical methods required that variables be normally distributed.
- However, real world data is not always symmetric in nature. It may be positively or negatively skewed.
- Skewed data can adversely affect the performance of data mining task leading to inaccurate predictions and biased results.

Skewness - refers to the asymmetry or lack of symmetry in a distribution of data. A data set is skewed if distribution is not symmetrical meaning that the tale of the distribution is longer in one side than the other.

- Positive skewness is right skewed. Here, the tail of the distribution extends to the right indicating that majority of the data points are concentrated on the left side with a few outliers on the right.
- Negative skewness is left skewed. Conversely, in a negatively skewed distribution the tail of the distribution extends to the left indicating that the majority of the data points are concentrated on the right side with a few outliers on the left.
- For right skewed data, mean > median, therefore skewness will be positive value for left skewed data, mean < median, therefore generating negative values for skewness.

. For perfectly ~~skewed data, mean~~ symmetrical diagram skewness = 0.

Transformation to achieve normality involve -

applying mathematical functions like logarithmic, square root or, boxcox transformations that is not normally distributed to meet the assumptions of statistical tests such as ~~pre test tests~~ t-test, anova and regression.

~~2 standardization data~~

. Z standardized data will have mean 0 and standard deviation 1 but doesn't mean that they are normally distributed ~~per~~.

. ~~Ske~~ skewness = $\frac{3(\text{Mean} - \text{median})}{\text{standard deviation}}$

. To make our data "more normally distributed", we must first make it symmetric, which means eliminating the skewness (=0).

- To eliminate skewness, we apply a transformation to the data.
  - ln(weight)
  - $\sqrt{\text{weight}}$
  - $1/\sqrt{\text{weight}}$

14/10

Logarithmic transformation -

- This is one of the most common transformation and is very effective data i.e ~~(+ve)~~ skewed.

$$X_{new} = \ln(X_{old})$$

- We can use natural logarithm or a base 10 logarithm function. It is best suited for data that follows a log ~~no~~ normal distribution or has a strong positive skewed data. For data containing 0s we can add a small constant to every value by taking the logarithm.

$$\log(X_{old} + c)$$

Square root transformation -

This is a less severe transformation than of logarithm. It is useful for data with a moderate positive skewed.

$$X_{new} = \sqrt{X_{old}}$$

- It's best for count data which often follows a Poisson distribution.

Inverse square root transformation -

It's a powerful technique used to achieve normality, especially for data with a strong positive skew.

$$x_{new} = 1/\sqrt{x_{old}}$$

Reciprocal Transformation -

This is a drastic transformation used for severely positively skewed data.

$$x_{new} = 1/x_{old}$$

Boxcox transformation - This is a familiar power transformation that finds the optimal parameters λ to make the data as normal as possible.

$$y(\lambda) = \frac{x^\lambda - 1}{\lambda}, \text{ if } \lambda \neq 0$$

$$y(\lambda) = \ln(x), \text{ if } \lambda = 0$$

- It is based for when we are unsure which transformation to use the algorithm can evaluate a range of λ values to find out the best fit.

Q1) Do logarithmic transformation, square root transformation and inverse square root transformation and reciprocal transformation for x = 15,000, 25,000, 30,000, 45,000, 50,000, 75,000, 1,50,000, & 5,00,000.

Sol: Logarithmic transformation

$x_{new} = \ln(15,000) = 9.615$

$x = (9.615, 10.126, 10.308, 10.714, 10.819, 11.225, 11.918, 13.122)$

Square root transformation

X = (122.474, 158.113, 173.205, 212.132, 223.606, 273.861, 387.298, 707.106)

Inverse square root transformation

~~X = (8164.99,~~

X = (0.008, 0.006, 0.005, 0.004, 0.0044, 0.0036, 0.002, 0.001)

Reciprocal transformation

X = (0.00006, 0.00004, 0.000023, 0.000022, 0.00002, 0.000013, 0.0000066, 0.000002)

Numerical methods for identifying outliers -

- A range of numerical methods used to identifying outlayers in a dataset from simple statistical rules to more complex ML algorithms.
- The choice of method depends largely on the data distribution and dimensionality
- Statistical methods are particularly effective for detecting outliers in univaried datasets.

* Z-score and IQR

- Z score methods for identifying outliers states that a data value is an outlier if it has a z-score that is either less than -3 or greater than 3.
- Unfortunately, the mean and standard deviation, which are both part of the formula for z-score ~~transformation~~ standardization, are both rather sensative to the presence of outliers.
- Therefore, data analyst have developed more robust statistical methods for the detection, which are less sensative to the presence of the outliers themselves.
- One method is the interquartile range (IQR):
  - 1st quartile (Q_1) is 25th percentile.
  - 2nd quartile (Q_2) is 50th ~~qua~~ percentile.
  - 3rd quartile (Q_3) is 75th percentile.
- The IQR is calculated as $IQR = Q_3 - Q_1$, and may be interpreted to

to see the spread of the middle @ 50% of the data.

- A data value is an outlier if:
  - It is located 1.5(IQR) or more below Q1, or
  - It is located 1.5(IQR) or more above Q3

15/10

Flag variables -

- A flag variable (or dummy variable or indicator variable) is a categorical variable taking only 2 values 0 and 1.
- When a categorical predictor takes $K \geq 3$ possible values, then define $K-1$ dummy variables, and use the unassigned category as a reference category.
- Eg: if a categorical predictor region has $K=4$ possible categories {N, E, S, W}, then the analyst could define the following $K-1=3$ flag variables.

Transforming categorical variables into numerical variables -

- In most instances, the data analyst should avoid transforming categorical variables to numerical variables.
- The exception is for categorical variables that are clearly ordered, such as the variable survey_response, taking values, always, usually, sometimes, never.

Binning numerical variables - (continued)

- Binning by clustering - uses a clustering algo, such as K-means clustering to automatically calculate the "optimal" partitioning.
- Binning based on predictive values - As other methods ignore the target variables; binning based on predictive value partitions the numerical predictor based on the effect each partition has on the value of the target variable.

Eg,

$X = \{1,1,1,1,1,2,2,11,11,12,44\}$

clustering $\Rightarrow$ bin 1 $= \{1,1,1,1,1,2,2\}$

bin 2 $= \{11,11,12\}$

bin 3 $= \{44\}$

Reclassifying categorical variables - is categorical equivalent of binning numerical variables.

- Often, a categorical variable will contain too many fields values to be easily analyzable.
- Data mining methods such as logistic regression and the C4.5 decision tree algo perform suboptimally when confronted with predictors containing too many field values.
- Eg: the 50 states could each be classified as the variable region, containing field values Northeast, Southeast, Northcentral, Southwest, and West.
- Alternatively, the 50 states could be reclassified as the variable economic-level with 3 field values containing the richer states, the mid-range states and the poorer states.

Adding an index field :-

- It is recommended that the data analyst create an index field, which tracks the sort order of the records in the database.
- Data mining data sets partitioned at least once.
- It is helpful to have an index field so that the original sort order may be recreated.
- Eg: using IBM/SPSS modeler, you can use the @Index function in the Derive node to create an index field.

Removing variables that are not useful -

- The data analyst may wish to remove variables that will not help the analysis, regardless of the proposed data mining task or algorithm.
  - Unary variables.
  - Variables which are very nearly unary.
- Unary variable take on only a single value, so a unary variable is not so much a variable as a constant.
- Sometimes a variable can be very nearly unary.
- Eg: suppose that 99.95% of the players in a field hockey league are female, with the remaining 0.05% male.
- While it may be useful to investigate the male players, some algorithms will tend to treat the variable as essentially unary.

Variable that should probably not be removed -

- It's a common - through questionable - practice to remove from analysis the following types of variables.
  - variables for which 90% or more of the values are missing.
  - variables of which are strongly correlated.
- Variable which contain 90% missing values present a challenge to any strategy for imputation of missing data.
- Conceivably, those who donate a lot would be inclined to report their donations, while those not donate much may be inclined to skip this survey question.
- Thus, the 10% who report are not representative of the whole.
- In this case, it may be preferable to construct a flag variable, donation-flag, since there is a pattern in the missingness which may be turn out to have predictive power.
- Inclusion of ~~correlated~~ correlated variable may be best - double count a ~~particular~~ aspect particular.

21/10

1) Exploratory Data analysis.
2) Types of data analysis
3) Common data analysis technique
4) Types of exploratory data analysis
5) Hypothesis testing vs exploratory data analysis

- Data analysis is a systematic process of collecting, cleaning, transforming and interpreting data to discover useful information, identifying patterns and support strategic decision making.
- The core process of generally follows - structured iterative approach

Steps

1. Define the objective
2. Collect data
3. Clean and organize data

4. Analyze data
5. Interpret results
6. Present results

Communicate the insides clearly to stake holders using visualizations like charts, graphs, and dash boards.

- The presentation should provide actionable recommendation.

Types of data analysis -

Depending on the goal data analysis is typically categorized into 4 main types.

1) Descreptive analytics - It answers the what happended by summarizing past data usually in the form of ~~dash board~~ dash board.

2) Diagnostic analysis - investigates historical data to understand why a certain outcomes occured.

3) Predictive analysis - uses historical data and ~~state~~ statistical methods to forecast what is likely to happen in the future ~~Ans~~ such as anticipating customers demands.

4) ~~Comm~~ Prescriptive analysis - Builds on predictive insides by ~~recom~~ recommending specific actions to take achieve a desired ~~outcome~~ outcome in analysis tech.

Common data analysis technique - analyst use a variety of techniques. to derive insights from data including explarotory data analysis, regression analysis, cluster analysis, text analysis.

22/10

Exploratory data analysis - It is used to analyse and investigate data sets and summarize their main characteristics often employing data visualization method. It finds pattern and discover how different parts of ~~data~~ data are connected.

Q) Why EDA is important ?

- EDA helps to understand the data set by how many features has what type of data each feature contain and how the data is distributed

- EDA helps to identify hidden patterns and relationship between different data points which helps us to in model building.

- EDA allows to identify errors or unusual data points that could affect our results.

- The insides gain from EDA helps us to identify most important features for building models and guide us on how to prepare them for better performance by understanding the data it helps us in choosing best modeling technique and adjusting them for better results.

Types of EDA -

There are various types of EDA based on nature of records depending on the no. of columns we are analyzing we can divide EDA into 3 types -

1. Univariate analysis - it focuses on studying one variable to understand its characteristics. It helps to describe data and find patterns using a single feature. Various common methods like histograms : Histograms are use to show data distribution. Boxplot to detect outliers and understand data spreads and bar charts for categorical data.

2. Bivariate analysis - It focuses on identifying relationship btw two variables to find connections, co-relation and dependencies. It helps to understand how two variables interact with each other. Some key techniques include scatter plots which visualize relationship btw two continous variables, co-relation, co-efficient measures how strongly the two variables are co-related cross tabulation or continous table shows the frequency distribution of two categorical variables and help to understand their relationship. In line graphs are useful for comparing two variables over time in Time series data to identify trends or patterns.

- Co-variant measures how 2 variables change together but it is paired with co-relation, co-efficient for deeper and more standardize understanding of relationship.

3. Multivariate analysis - It identify relationship blw two or more variables in a data set aims to understand how variables interact with one another which is important for statical modelling techniques. It includes techniques like pair-plots, which shows the relationship blw multiple variables at once and helps in understanding how it works, them PCA which reduces the complexity of large data set by simplifying them by keeping the most important information.

- Spatial analysis is used for geographical data by using maps and spatial plotting to understand the geographical distribution of variables.
- Time series analysis it is used for data sets that involves time based data and it involves understanding and modelling patterns and trains over time.
- Common technique includes line plots, auto co-relation analysis, moving average, and arima models.

Hypothesis Testing vs Exploratory Data -

- EDA often comes first (discovery stage) - helps analysts understand the dataset, distributions, and uncover important relationships and patterns that could indicate important areas for further investigations.
- Hypothesis testing follows (confirmation stage) - validates the patterns or suspicions suggested by EDA with statistical rigor, i.e testing assumptions with formal procedures.
- Together, they form a powerful cycle of discovery and confirmation in data mining and statistical analysis.

Aspect	Hypothesis testing	Exploratory Data Analysis
Purpose	To confirm or reject a pre-specified idea	To discover patterns, understand distributions, and generate new ideas.
Approach	Deductive, confirmatory	Inductive, discovery-oriented
When used	When clear, theory-driven questions exist	When data are unfamiliar, large or complex
Focus	Formal decision-making	Investigation of variables, distributions and relationships
Tools	Statistical tests (t-test, chi-square, anova, regression)	Graphical (Histograms, scatter plots, box plots and descriptive statistics).
Outcomes	Binary decision (reject / fail to reject H_0)	Insights, hypotheses, directions for further study.
Flexibility	~~Binary decision (~~ Rigid, structured	Flexible, Iterative

27/10

Hypothesis testing -

- Hypothesis testing is a formal statistical procedure used to evaluate weather a statement (hypothesis) about a population parameter is supported by sample data.

Key features

- Starts with an a priori hypothesis (before examining the data in details).
- Involves null hypothesis (H_0) and alternative hypothesis (H_1).
- Provides a yes/no decision (reject or fail to reject H_0).

Example - If a company says its website gets 50 visitors each day on average we use hypothesis testing to look at past visitor data and see if this claim is true or if the actual number is different.

Common EDA techniques -

- Graphical - Histograms, scatter plots, box plots, correlation heatmaps.
- Numerical - summary statistics (mean, median, variance & skewness), correlation coefficients.
- Subsets / Group analysis - Identifying clusters, trends or interesting subsets.
- EDA acts as the foundation of data analysis shaping the direction of further investigation and hypothesis testing.

Exploring categorical variables -

Objective: To identify the categorical variables influencing the minority class

- We are to test two categorial values.
  - International plan
  - Voice mail plan
- Understanding target distribution

Value	Proportion	%	Count
False		85.51	2850
True		14.49	483

28/10

only 14.49% of customer churned.

churn example -

- One of the primary reasons for performing EDA is investigate the variables,
  - examine the distribution of categorical variables,
  - look at the histograms of the numeric variables, and
  - explore the relationships among sets of variables.

International Plan

churn	No	Yes	Total
False	2664	186	2850
True	346	137	483
	3010	323	3333

churn % who not take international plan = 11.49%

churn % who take international plan = 42.41%

churn false for not international burp = 88.50%

churn false for international burp = 57.58%

~~2664~~

Row percentage

churn false

Plan (Yes) = $\frac{186}{2850}$ = 6.52%

Plan (No) = $\frac{2664}{2850}$ = 93.47%

churn true

Plan (No) = $\frac{364}{483}$ = 75.36%

Plan (Yes) = $\frac{137}{483}$ = 28.36%

29/10

- To summarize this EDA on the international plan has indicated that perhaps we should investigate what it is about our international plan i.e inducing our customers to live.

- We should expect that whatever data mining algorithms we use to predict churn the model will probably include whether or not the customer selected the international plan.

Voice mail plan -

	Voice mail plan		
churn	No	Yes	Total
~~churn~~ False	2008	842	2850 = 85.5%
~~False~~ True	403	80	483 = 14.5%
Total	2411	922	3333

Col percentage

churn false

voice plan (Yes) = $\frac{842}{922}$ = 91.3%

voice ~~plan~~ plan (No) = $\frac{2008}{2411}$ = 83.3%

churn true

voice plan (Yes) = 8.67% ≈ 8.7%

voice plan (No) = 16.7% (no plan)

Conclusion - People with no voice plan are more likely to churn out.

Without voice mail plans are churners

- EDA on voice mail plan has indicated that, perhaps we should enhance our voice mail plan still further or make it easier for customers to join it as an instrument for increasing customer loyalty.

- We should expect that whatever data mining algorithm we use to predict churn the model will probably include whether or not the customer

selected the voice mail plan.

- The confidence in this expectation is perhaps not quite as high as the international plan.
- The 2-way interactions can be explored among categorical variables with respect to churn.

Results for voice mail plan = no

Rows: Churn Columns: International Plan

	no	yes	all
False	1878	130	2008
True	302	101	403
All	2180	231	2411

Results for voice mail plan = Yes

Rows: Churn Columns: International Plan

	no	yes	all
False	786	56	842
True	44	36	80
All	830	92	922

Numeric variable

Customer service call and churn -

- Normalized histograms are useful for teasing out the relation between a numerical predictor and the target. However data analyst should always provide the companion in along with the ~~none~~ normalized histogram because the normalized histogram doesn't provide any information on frequency distribution of the variable.
- This EDA on customer service calls has indicated that:

6/11

CHAPTER-4

Dimension Reduction

- The need for dimensionality reduction arises when working with datasets that have a large no. of variables.
- High dimensional data can make analysis, visualization and modelling complex, slow and sometimes inaccurate.
- Dimensionality reduction techniques simplify data by retaining as much important information as possible.

Definition - Dimensionality reduction refers to the process of converting a dataset with many variables into one with fewer variables while still preserving most of the important information or patterns.

Need for dimensionality reduction -

i) Remove redundancy

ii) **To overcome the curse of dimensionality** - When a no. of dimension increases the data becomes sparse and it become difficult for algorithms to find patterns.

- Reducing dimensions helps make the data denser and more meaningful.
- In 100 dimensional space data points are far apart reducing algorithm performance.

iii) **To remove multicolinearity** - ~~Highly co-related features provide~~ It is a statistical phenomenon that occurs when 2 or more independent predictor variables in a regression model are highly co-related. i.e they provide overlapping or redundant information

about the independent variable.

- Multicolinearity means that some of the predictors are so closely related that it becomes difficult to tell which one actually affects the outcome.

iv) To reduce overfitting - Having too many features with limited data can cause the model to memorize noise leading to overfitting.
Ex- 1000 dimensions but only 100 samples may cause poor generalization.

v) To improve computational efficiency - High dimensional data requires more time, memory and computational power.
Eg: Face recognising data with 10,000 pixel features can be reduced to 100 features using PCA.

vi) To improve visualization and understanding - It is difficult to visualize data with more than 3 dimensions. Dimensionality reduction allows visualization in 2D or 3D space for pattern discovery.

vii) To remove noise and irrelevant features - Real world data often has irrelevant or noisy variables that can confuse models.
Eg: Color of the door in a house price dataset is irrelevant and can be removed.

viii) To enhance model performance - Reducing irrelevant features helps algorithms converge faster and perform better.
Eg: Logistic regression models perform better when trained on fewer meaningful features.

PCA

Problem:

Suppose we have variables exam 1, exam 2, that provide score for 5 students. Apply PCA to compute principal components.

Students	Exam 1	Exam 2
A	90	85
B	70	65
C	80	78
D	65	60
E	95	92

Steps of PCA -

1) Let us assume we have data sets with 'n' observations & 'p' variables then standardize the data. Each variable x_j is transfer to half mean 0 and standard deviation 1. i.e $Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$

x_{ij} → value of variable j for observation i.

$\bar{x}_j$ → Mean of variable j

s_j → Standard deviation of j.

2) Compute the variance - The covariance matrix can be determined as follows

Covariance (x_1, x_2)

$$cov(x_i, x_i) = \frac{\sum x_i^2}{n-1}$$

$$cov(y_i, y_i) = \frac{\sum y_i^2}{n-1}$$

$$cov(x_i, y_i) = \frac{\sum x_i y_i}{n-1}$$