# Chapter 1 Introduction of Data analytics

## Key function of data Scientist

Understand the business problem->

Data Mining and Analysis Design->

Descriptive and predictive Analysis->

Devising Business strategies from the insight

## Work can do for Data Scientist

Building training sets->

Cleaning and organizing data->

collecting data sets->

mining data for patterns->

refining algorithms->

others

**Analysis VS Analytics**

Analysis refers to the process of separating a whole problem into its parts so that the parts can be critically examined at the granular level.

Analytics is a variety of methods, technologies and associated tools for creating new knowledge/insight to solve complex problems and make better and faster decisions.

**What can be learned from data?**

Data Analytics is an emerging technique that dives into a data set without prior set of hypotheses. The data derive meaningful trends or intriguing findings that were not previously seen or empirically validated. Data analytics enables quick decisions or help change policies due to trends observed.

**Data Analytics**

Accumulation of raw data captured from various sources can be used to identify fruitful patterns and relationships. Exploratory visualization uses exploratory data analytics by capturing relationships that are perhaps unknown or at least less formally formulated. Confirmatory visualization is theory driven.

**Type of Data Analytics**

(Information->analysis->decision) hindsight ->insight->foresight

Descriptive Analytics -> what’s happening? Comprehensive, accurate and live data. Effective visualization.

Diagnostic Analytics -> why happen? Ability to drill down to the root cause and to isolate all confounding information.

Predictive Analytics ->what is likely to happen? Business strategies have remained fairly consistent over time. Historical patterns being used to predict specific outcomes using algorithms. Decision are automated using algorithm and technology.

Prescriptive Analytics -> what shall I do about it? Recommended actions and strategies based on champion/ challenger testing strategy outcomes. Applying advanced analytical techniques to make specific recommendations.

Deeper look from descriptive to prescriptive.

**Data vs information**

Data is numerical or textual facts and figures that are collected through some type of measurement process.

Information is result of analysing data, that is extracting meaning from data to support evaluation and decision making.

**Data type for data Analytics**

Data sets: a collection of data. Such as marketing survey responses, a table of historical stock process and a collection of measurements of dimensions of a manufactured item.

Data base: a collection of related files containing records on people, place and things. It usually organized in a two-dimensional table where the columns correspond to each individual element of data, and rows represent records of related data elements.

Data warehouse: a collection of databases used for reporting and data analysis. Such as data mart. A departmental data warehouse that stores only relevant data.

Chapter 2 introduction to data warehouse and crisp-dm

**Data warehouse**

A logical collection of information gathers from many operational databases.

Supports business analytics activities.

Decision-making task.

Many organizations need internal and external data sources.

To aggregate information throughout an organization into a single repository for decision-making purposes.

**Database**

An organized grouping of information within a specific structure that needs to be retrieved frequently.

Most databases in use today are relational database*.*

They are designed using *many tables* which *relate* to one another in a logical fashion.

Relational databases generally contain *dozens or even hundreds of tables*, depending upon the size of the organization.

**Data mart**

Subsets of data warehouses that is highly focused and isolated for a specific population of users

Examples: Marketing data mart, inventory data mart and etc.

# Chapter 2

## Data communication (ETL)

Extraction, transformation and loading (ETL) is a process that:

Extract information from internal and external databases

Transform the information using common set of enterprise definition

Load the information into a data warehouse

## MULTI-DIMENSION INFORMATION

Database -a series of two-dimensional tables/information

Data warehouse / data mart –a multi-dimensional information that contains layers of columns and rows.

Cube –a common term to represent multi-dimension information.

|  |  |  |
| --- | --- | --- |
|  | Database | Data Warehouse |
| Purpose | For data retrieval, updating and management | For data analysis and decision making |
| Systems/ Application | OLTP (Online Transaction Processing System) | Analytical Software. i.e. data mining tools, reporting tools and OLAP (Online Analytical Processing) tools |
| Format | •Normalised  •Relational database  •Lowest level of granularity (i.e., individual transactions) | •De-normalised and integrated  •Multi-dimensional arrays or relational format  •Subject-oriented  •Granularity level depends on subject |
| Time Frame | Current / Real time | Historical |

A data set is a subset of a database or a data warehouse.

It is usually denormalized so that only one table is used.

The creation of a data set may contain several steps, including appending or combining tables from source database tables, or simplifying some data expressions.

Data sets may be made up of a representative sample of a larger set of data, or they may contain all observations relevant to a specific group.

## CRISP-DM

->CRoss-Industry Standard Process for Data Mining

**Advantage of CRISP-DM**

Non-proprietary

Application/Industry neutral

Tool neutral

Focus on business issues - As well as technical analysis

Framework for guidance

Experience base - Templates for Analysis

## CRISP-DM phase

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Business** **Understanding** | **Data** **Understanding** | **Data** **Preparation** | **Modeling** | **Evaluation** | **Deployment** |
| **Determine Business Objectives** | **Collect Initial Data** | **Select Data** | **Select Modeling** **Technique** | **Evaluate Results** | **Plan Deployment** |
| **Situation Assessment** | **Describe Data** | **Clean Data** | **Generate Test Design** | **Review Process** | **Plan Monitoring and Maintenance** |
| **Determine Data Mining Goal** | **Explore Data** | **Construct Data** | **Build Model** | **Determine Next Steps** | **Produce Final Report** |
| **Produce Project Plan** | **Verify Data Quality** | **Integrate Data** | **Assess Model** | **Review Project** |
| **Format Data** |

**Business** **Understanding**

**Determine Business Objectives** – Background, Business Objectives, Business Success, Criteria.

**Situation Assessment -** Inventory of Resources, Requirements, Assumptions, and Constraints, Risks and Contingencies, Terminology, Costs and Benefits.

**Determine Data Mining Goal -** Data Mining Goals, Data Mining Success, Criteria*.*

**Produce Project Plan -** Project Plan, Initial Assessment of Tools and Techniques*.*

**Data** **Understanding**

**Collect Initial Data -** Initial Data Collection Report*.*

**Describe Data -** Data Description Report.

**Explore Data -** Data Exploration Report*.*

**Verify Data Quality -** Data Quality Report*.*

**Data** **Preparation** (*Data Set, Data Set Description)*

**Select Data -** *Rationale for Inclusion /Exclusion*

**Clean Data -** *Data Cleaning Report*

**Construct Data -** *Derived Attributes, Generated Records*

**Integrate Data -** *Merged Data*

**Format Data -** *Reformatted Data*

**Modeling**

**Select Modeling** **Technique -** *Modeling Technique, Modeling Assumptions*

**Generate Test Design -***Test Design*

**Build Model -** *Parameter Settings, Models, Model Description*

**Assess Model -** *Model Assessment, Revised Parameter Settings*

**Evaluation**

**Evaluate Results -** *Assessment of Data, Mining Results w.r.t. , Business Success Criteria, Approved Models*

**Review Process -** *Review of Process*

**Determine Next Steps -** *List of Possible Actions, Decision*

**Deployment**

**Plan Deployment -** *Deployment Plan*

**Plan Monitoring and Maintenance -** *Monitoring and Maintenance Plan*

**Produce Final Report -** *Final Report, Final Presentation*

**Review Project –** *Experience, Documentation*

# Chapter 3 &4 Data pre-processing

## Why pre-processing?

Incomplete, noisy, inconsistent

## MAJOR TASKS IN DATA PREPROCESSING

Data cleaning - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration - Integration of multiple databases, data cubes, or files

Data transformation- Normalization and aggregation

Data reduction - Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization - Part of data reduction but with particular importance, especially for numerical data

## Why data mining?

To better understand the data: central tendency, variation and spread

**Data dispersion characteristics**

- median, max, min, quantiles, outliers, variance, etc.

**Numerical dimensions** correspond to sorted intervals

- Data dispersion: analyzed with multiple granularities of precision 🢝 Boxplot or quantile analysis on sorted intervals

Dispersion analysis on computed measures

- Folding measures into numerical dimensions

- Boxplot or quantile analysis on the transformed cube

## MEASURES OF CENTRAL TENDENCY

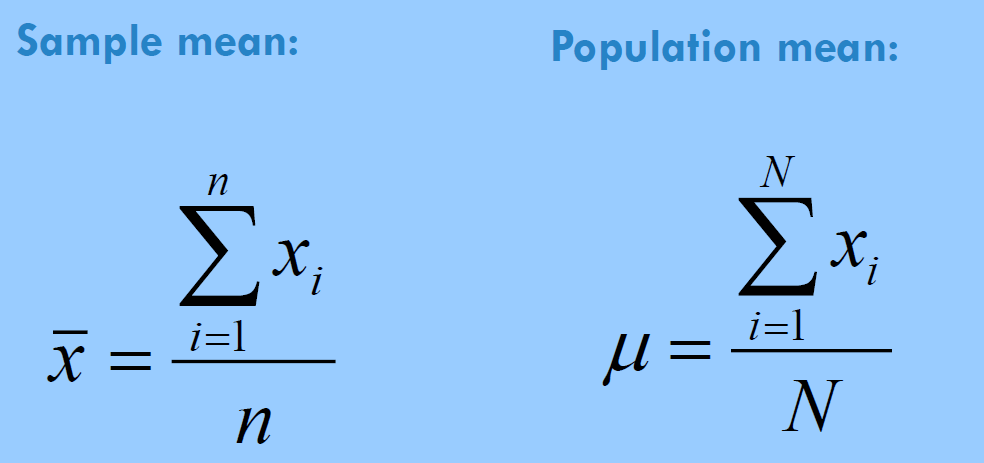
**Measures** of the location of the **middle** or the **centre** of a distribution

## Mean, median, mode

**Mean –** Most commonly used measure of central tendency **Average** of all observations

The sum of all the scores divided by the number of scores

Note: **Assuming** that each observation is **equally** significant



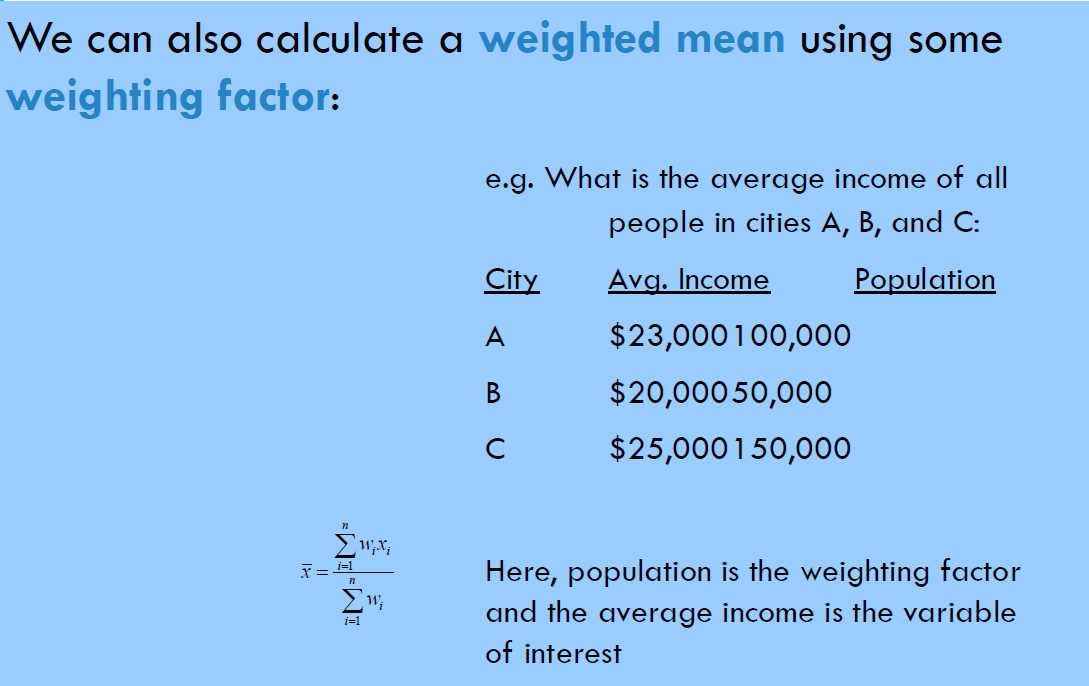
Advantage

Sensitive to any change in the value of any observation

Disadvantage

Very sensitive to outliers

**WEIGHTED MEAN**

****

**Median** – This is the value of a variable such that half of the observations are above and half are below this value i.e. this value divides the distribution into two groups of equal size. When the number of observations is **odd**, the median is simply equal to the **middle** value. When the number of observations is **even**, we take the median to be the **average** of the two values in the middle of the distribution

**Advantage**: the value is NOT affected by extreme values at the end of a distribution (which are potentially are outliers)

**Mode** - This is the **most** frequently occurring value in the distribution. This is the only measure of central tendency that can be used with **nominal data.** The mode allows the distribution's **peak** to be located quickly.

Most often, the **mean** is selected **by default.** The mean's key **advantage** is that it is sensitive to any change in the value of any observation. The mean's **disadvantage** is that it is very sensitive to **outliers.** We really must consider the **nature** of the data, the **distribution**, and our **goals** to choose properly.

## SOME CHARACTERISTICS OF DATA

Not all data is the same. There are some **limitations** as to what can and cannot be done with a data set, depending on the characteristics of the data.

## A. CONTINUOUS VS. DISCRETE DATA

**Continuous** data can include any value (i.e., real numbers)

🢝 e.g., 1, 1.43, and 3.1415926 are all acceptable values.

🢝 Geographic examples: distance, tree height, amount of precipitation, etc

**Discrete** data only consists of discrete values, and the numbers in between those values are not defined (i.e., whole or integer numbers)

🢝 e.g., 1, 2, 3.

🢝 Geographic examples: # of vegetation types,

## B. GROUPED VS. INDIVIDUAL DATA

The distinction between individual and grouped data is somewhat self-explanatory, but the issue pertains to the effects of grouping data. While a family income value is collected for each household (individual data), for the purpose of analysis it is transformed into a set of classes (e.g., <$10K, $10K-20K, > $20K)

e.g., elevation (1000m vs. < 500m, 500-1000m, 1000- 2000m, > 2000m)

In grouped data, the raw individual data is categorized into several classes, and then analyzed. The act of grouping the data, by taking the central value of  each class, as well as the frequency of the class interval, and  using those values to calculate a measure of central tendency  has the potential to introduce a significant distortion

Grouping always reduces the amount of information contained in the data

## C. SCALES OF MEASUREMENT

Data is the plural of a datum, which are generated by the recording of measurements. Measurements involves the categorization of an item (i.e.,  assigning an item to a set of types) when the measure is  qualitative or makes use of a number to give something a  quantitative measurement.

The data used in statistical analyses can divided into four types: 1. The Nominal Scale 2. The Ordinal Scale 3. The interval Scale 4. The Ratio Scale

## THE NOMINAL SCALE

Nominal scale data are data that can simply be broken down into categories, i.e., having to do with names or types. Dichotomous or binary nominal data has just two types, e.g., yes/no, female/male, is/is not, hot/cold, etc.

Multichotomous data has more than two types, e.g., vegetation types, soil types, counties, eye colour, etc

Not a scale in the sense that categories cannot be ranked or ordered (no greater/less than)

## THE ORDINAL SCALE

Ordinal scale data can be categorized AND can be placed in an order, i.e., categories that can be assigned a relative importance and can be ranked such that numerical category values have star-system restaurant rankings 5 stars > 4 stars, 4 stars > 3 stars, 5 stars > 2 stars.

BUT ordinal data still are not scalar in the sense that differences between categories do not have a quantitative meaning. i.e., a 5-star restaurant is not superior to a 4-star restaurant by the same amount as a 4-star restaurant is than a 3-star restaurant.

## THE INTERVAL SCALE

Interval scale data take the notion of ranking items in order one step further, since the distance between adjacent points on the scale are equal.

For instance, the Fahrenheit scale is an interval scale, since each degree is equal but there is no absolute zero point.

This means that although we can add and subtract degrees (100° is 10° warmer than 90°), we cannot multiply values or create ratios (100° is not twice as warm as 50°)

## THE RATIO SCALE

Similar to the interval scale, but with the addition of having a meaningful zero value, which allows us to compare values using multiplication and division operations, e.g., precipitation, weights, heights, etc

e.g., rain – We can say that 2 inches of rain is twice as much rain as 1 inch of rain because this is a ratio scale measurement

e.g., age – a 100-year-old person is indeed twice as old as a 50-year-old one

## WHICH ONE IS BETTER: MEAN, MEDIAN, OR MODE?

The **mean** is valid only for **interval** data or **ratio** data.

Consider a company that has nine employees with salaries of 35,000 a year, and their **supervisor** makes 150,000 a year.

If you want to describe the **typical** salary in the company, which statistics will you use?  I will use mode or median (35,000), because it tells what salary **most** people get.

What if you are a **recruiting officer** for the company that wants to make a **good impression** on a **prospective** employee?  The mean is (35,000\*9 + 150,000)/10 = 46,500 I would probably say: "The average salary in our company is 46,500" using mean.

## Data cleaning

Data cleaning tasks

1. Fill in missing values
2. Identify outliers and smooth out noisy data
3. Correct inconsistent data
4. Resolve redundancy caused by data integration April 25, 2022 DATA MINING: CONCEPTS AND TECHNIQUES

**HOW TO HANDLE MISSING DATA?**

Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.

Fill in the missing value manually: tedious + infeasible? Fill in it automatically with a global constant: e.g., “unknown”, a new class?!  the attribute means. Imputation: fill in the missing value using the feature mean or the most probable value.

**IMPUTING MISSING DATA**

Delete missing observations. Can lead to serious biases. If missing data is relatively small, may be okay.

COLD-DECK IMPUTATION

Fill in the data using means or other analysis of the variable to fill in the value. Measure of central tendency (mean, median, mode)

HOT-DECK IMPUTATION

Identify the most similar case to the case with a missing value and substitute the most similar case’s value for the missing case’s value.  Advantages: simplicity, maintains level of measurement, complete data at the end.  Disadvantage: can identify more than one similar case and randomly select or use average.

DISTRIBUTION-BASED IMPUTATION

Assign value based on the probability distribution of the non-missing data. Tries to capture the “observed” empirical distribution of data.

STATISTICAL IMPUTATION

Build a regressor to classify the input value. Consider the “missing” value as the “output” and the rest of the features as input Imputes the value based on other features.

PREDICTIVE IMPUTATION

Let a classifier model the underpinnings of the missingness mechanism.

## NOISY DATA

Noise: random error or variance in a measured variable

Incorrect attribute values may due to

🢝 faulty data collection instruments

🢝 data entry problems

🢝 data transmission problems

🢝 technology limitation

🢝 inconsistency in naming convention

Other data problems which require data cleaning

🢝 duplicate records

🢝 incomplete data

🢝 inconsistent data

## HOW TO HANDLE NOISY DATA?

Binning - first sort data and partition into (equal-frequency) bins. then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Regression - smooth by fitting the data into regression functions

Clustering - detect and remove outliers

Combined computer and human inspection - detect suspicious values and check by human (e.g., deal with possible outliers)

## How to handle missing value?

1. Remove the attribute
2. Create a new attribute to handle the missing values
3. Remove the data (if it is not too many)
4. Imputation of Missing Values: - Use commonly used values or average the value
5. Use supervised learning to predict the missing values and use those values

## Data Integration

Data integration:  Combines data from multiple sources into a coherent store

Schema integration: e.g., A.cust-id ≡ B.cust-# ◼ Integrate metadata from different sources

Entity identification problem:  ◼ Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

Detecting and resolving data value conflicts: ◼ For the same real-world entity, attribute values from different sources are different. ◼ Possible reasons: different representations, different scales, e.g., metric vs. British units

## Data Reduction Strategies

◼ Why data reduction?

◼ A database/data warehouse may store terabytes of data ◼ Complex data analysis/mining may take a very long time to run on the complete data set

◼ Data reduction

◼ Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

◼ Data reduction strategies

◼ Data cube aggregation:

◼ Dimensionality reduction — e.g., remove unimportant attributes

◼ Data Compression

◼ Numerosity reduction — e.g., fit data into models

◼ Discretization and concept hierarchy generation

## Data Cube Aggregation

◼ The lowest level of a data cube (base cuboid) ◼ The aggregated data for an individual entity of interest ◼ E.g., a customer in a phone calling data warehouse

◼ Multiple levels of aggregation in data cubes ◼ Further reduce the size of data to deal with ◼ Reference appropriate levels. ◼ Use the smallest representation which is enough to  solve the task

◼ Queries regarding aggregated information should be answered using data cube, when possible

## Attribute Subset Selection

◼ Feature selection (i.e., attribute subset selection):

◼ Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features

◼ reduce # of patterns in the patterns, easier to understand

◼ Heuristic methods (due to exponential # of choices):

◼ Step-wise forward selection

◼ Step-wise backward elimination

◼ Combining forward selection and backward elimination

◼ Decision-tree induction

## Heuristic Feature Selection Methods

◼ There are 2d possible sub-features of d features ◼ Several heuristic feature selection methods:

◼ Best single features under the feature independence assumption: choose by significance tests

◼ Best step-wise feature selection:

◼ The best single-feature is picked first

◼ Then next best feature condition to the first, ...

◼ Step-wise feature elimination:

◼ Repeatedly eliminate the worst feature

◼ Best combined feature selection and elimination

◼ Optimal branch and bound:

◼ Use feature elimination and backtracking

## Dimensionality Reduction: Principal Component Analysis (PCA)

◼ Given N data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) that can be best used to represent data

◼ Steps

◼ Normalize input data: Each attribute falls within the same range

◼ Compute k orthonormal (unit) vectors, i.e., principal components

◼ Each input data (vector) is a linear combination of the k principal component vectors

◼ The principal components are sorted in order of decreasing “significance” or strength

◼ Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data

◼ Works for numeric data only

◼ Used when the number of dimensions is large