Introduction to Business Analytics and Data Visualization

What is Business Analytics?

Business analytics is the use of:

- Data,
- Information technology,
- Statistical analysis,
- Quantitative methods, and
- Mathematical or computer-based models

To help managers gain improved insight about their business operations and make better, fact-based decisions.

Importance of analytics

Data, facts, and analysis are powerful aids to decision making and that the decisions made on them are better than those made through intuition or gut instinct (Davenport, Harris & Morrison 2010).

- What makes decision making complicated today is the overwhelming amount of available data and information (Evans 2013).

Scope of Business Analytics

Business analytics begins with the collection, organisation, and manipulation of data and is supported by three major components:

- Descriptive analytics
- Predictive analytics
- Prescriptive analytics

Descriptive analytics (understand what happened)

Uses data to understand past and present performance and make informed decisions.

- Most commonly used and well understood type of analytics.

Uses fundamental tools and methods of data analysis focusing on:

- Descriptive statistical measures and data visualisation
- Probability distributions / confidence interval
- Hypothesis test

Predictive Analytics (predict what will happen)

Analyses past performance in an effort to predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time.

Techniques include:

- Regression and forecasting
- Data Mining and Machine Learning
- Inferential statistics

Prescriptive Analytics (what should I do?)

Uses optimization to identify the best alternative to minimize or maximize some objective.

Techniques include the use of mathematical models:

- Decision analysis
- Optimization
- Simulation

Data for business Analytics

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

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

Data Sources

Primary sources: internal company records and business transactions, automated data-capturing equipment, customer market surveys

Secondary sources: Government and commercial data sources, custom research providers, online research

So what is Big Data?
Big data refer to massive amounts of business data from a wide variety of sources, much of which is available in real time, and much of which is uncertain or unpredictable.

- Volume: amount of data we produce [2.5 exabyte a day and doubling every 40 months]
- Variety: big data takes the form of messages, updates and images posted to social networks, readings from sensors, GPS signals from cell phones and more
- Velocity: means both how fast data is being produced and how fast the data must be processed to meet the demand

Structured and Unstructured Data

		Data Type	
		Structured	***Unstructured***
Data Source	*External*	Public Data Postcode data House Hold data Credit Scoring Market Research data	Social Media (Twitter, Facebook, Instagram) Blogs External sensor data
	Internal	CRM data Sales data Transaction data Invoice data Usage data Campaign data	Customer contact data (mail, email, text message, call centre, shop, website) Sensor data Mobile data

Structured: data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis.

Unstructured: Info that either does not have pre-defined data model or is not organized in a pre-defined manner.

Datasets, entities, variables, and records

- The data collected in a particular study are referred to as a dataset.
- The people, places or things for which we store and maintain information are called entities. Ie. Employees, teachers, male, female
- A variable (or attribute) is a characteristic of interest for the entities ie. Sex/gender/age
- The set of measurements collected for a particular entity is called a record (or observation).

Types of data
Categorical Data (quantitative)

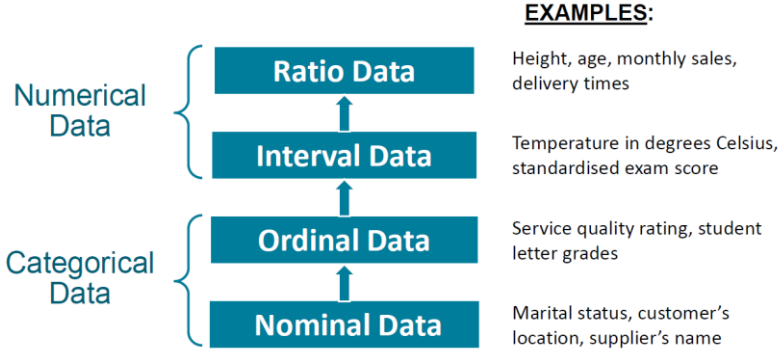
- Labels or names used to identify an attribute of each entity.
- Often referred to as qualitative data
- Can be recorded in either numeric or nonnumeric format
- Appropriate statistical analyses are rather limited
- Usually counted or expressed as a proportion or a percentage

Numerical Data (qualitative)
Numerical data indicate how many or how much and often referred to as quantitative data. Ordinary arithmetic operations are meaningful for numerical data.

- **Discrete:** if measuring how many (occupy a particular value ie. Visits)
- **Continuous:** if measuring how much (any value ie. height)

Numerical data can be converted to categorical data
e.g., salary can be converted into 'low', 'middle' and 'high'. But cannot convert 'high' salary back into a specific salary figure.

Scales of Measurement



Nominal Data
Data are labels or names used to identify an attribute to the entity.
A non-numeric label or numeric code may be used.

- cannot be sorted/ordered and measured (doesn't provide quantitative data) (one is not better than another)
- Customers can be classified by their geographical location (New South Wales, Victoria, Western Australia, South Australia, Tasmania, ACT and NT).

Ordinal Data
The data have the properties of nominal data and the categories have a meaningful rank.

A nonnumeric label or numeric code may be used.

- Rating customer service as "Poor", "average", "good", "very good" or excellent

Ordinal data have no fixed unit of measurement. Thus cannot make meaningful statements about the difference between categories.

- Cannot say that the difference between "excellent" and "very good" is the same as between "good" and "average".

Interval data
The data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.

Interval data are always numeric and have no true zero.

- Both Fahrenheit and Celsius scales represent specific measure of distance – degrees of temperature but have no true zero.

Thus cannot make meaningful ratios

- We cannot say that 50 degrees is twice as hot as 25 degrees
- defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another. Always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.

Ratio data
The data have all the properties of interval data and the ratio of two values is meaningful.

This scale must contain a zero value that indicates that nothing exists for the variable at the zero point.

- Defined as quantitative data, having the same properties as interval data with an equal and definitive ratio between each data and absolute "zero" being treated as a point of origin. CAN BE NO NEGATIVE NUMERICAL VALUE IN RATIO DATA.

Data Reliability and Validity
Reliability: data are accurate and consistent
Validity: data correctly measures what it is supposed to measure

- A tire pressure gage that consistently reads several pounds of pressure below the true value is not reliable, although it is valid because it does measure tire pressure.

Models in business analytics

- A **model** is a representation of a system, idea, or object

Models capture the most important features a problem and present them in a form that is easy to understand. It is often less expensive to analyse decision problems using a model.

Most importantly models allows us to gain insight and understanding about the decision problem at hand

Analytics as a problem-solving approach

1. Recognize problem
2. Define problem
3. Structure problem
4. Analyse problem.

And finally "good communication is vital in all steps"

Descriptive Analytics

Role of data visualisation

Data visualisation in Business Analytics serves two main purposes

1. Assist in analysis (as an exploratory tool)
2. Assist in communication of insight (as a way of telling a story)

Dashboard

- A dashboard is a visual representation of a set of key business measures. It is derived from the analogy of an automobiles control panel.
- Dashboards provide important summaries of key business information to help manage a business process or function.

Common visualisation

- Bar charts: comparing data across categories
- Pie charts: observe a portion of the total (a particular category)
- Line charts: observe trend over time
- Box/Dot plots: view distribution and unusual data elements (grouping
- Multiple box plot: comparison to find the difference between groups
- Scatter plots: investigate relationship between variables

Histogram

A graphical depiction of a frequency distribution for numeracy for numerical data in the form of a column chart is called a histogram.

Histograms for numerical data

For numerical data that have many different discrete values with little repetition or are continuous, a frequency distribution requires that we define by specifying

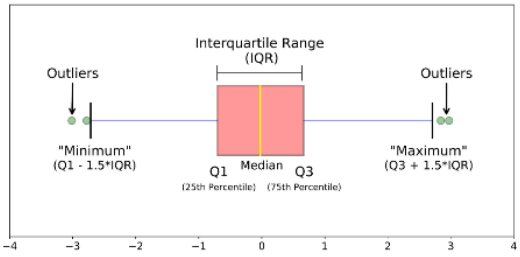
- The number of groups,
- The width of each group, and
- The upper and lower limits of each group

Choose between 5 to 15 groups, and the range of each should be equal.

Box Plot

A box and whisker plot – also called a box plot – displays the five-number summary of a set of data including: the first quartile, median, third quartile, and two whiskers. The whiskers can represent:

- The minimum and maximum of all of the data (as in figure below);
- The lowest data still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile often called the (turkey boxplot);
- One standard deviation above and below the mean of the data;
- The 9th percentile and the 91st percentile;



Tufte's principles of graphical display

Graphical displays should:

- Show the data;
- Tell the truth (avoid distorting what the data has to say);
- Focus on the content (help the viewer think about the information rather than design);
- Encourage the eye to compare the data;
- Make large data sets coherent;
- Reveal data at several levels of detail;
- Closely integrate statistical and verbal descriptions.

Descriptive Analytics: Univariate Analysis (Central Tendency etc.)

Univariate Analytics

Descriptive Analysis is the elementary transformation of raw data in a way that describes the basic characteristics of the data collected.

When we investigate one variable in isolation, we this **univariate Analysis**. Univariate analysis is the simplest form of analyzing data. It doesn't deal with causes or relationships and it's major purpose is to describe; it takes data, summaries that data, and finds patterns in the data.

Populations and samples

Population – all items of interest for a particular decision or investigation

- All married drivers over 25 years old
- All subscribers to Netflix

Sample – a subset of the population

- A list of individuals who rented a comedy from Netflix in the past year

The purpose of sampling is to obtain sufficient information to draw a valid inference about a population.

Frequency Table

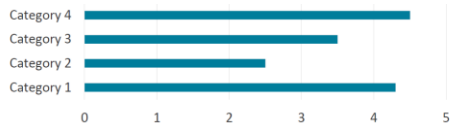
A **frequency table** (summary table) is a way of counting how often each category of the variable in question occurs. It may be enhanced by the addition of percentages that fall into each category.

Location	Number of properties	Percentage of properties
Rural	68	34.0
Town	132	66.0
Total	200	100.0

Bar Charts

A bar chart is (typically) the graphical representation of a summary table for categorical data.

The length of each bar represents the proportion, frequency, or percentage of data values in a category



- In most cases, bar graph better than frequency table

Univariate analysis of numerical data

Some ways you can describe patterns found in univariate data include:

- **Central tendency** (e.g., mean, mode, median)
- **Dispersion** (e.g., range, variance, maximum, minimum, standard deviation);
- **Shape** (e.g., skewness, kurtosis)

Measure of central tendency

A **measure of central tendency** is a single value that attempts to describe a set of data by identifying its central position.

Measures of central tendency are sometimes called measures of central location or averages. They are also classed as summary statistics.

Mean

Also called Arithmetic Mean

Population mean: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

- Gives equal weight to all values to find true average. Average of the data.

Median

The **median** specifies the middle value when the data are arranged from least to greatest (an ascending array).

- For an odd number of observations, the median is the middle of the sorted numbers.
- For an even number of observations, the median is the mean of the two middle numbers.

- Not affected by outliers/extreme values and skewed data.

Mode

The **mode** is the observation that occurs most frequently.

- The greatest frequency can occur at one (unimodal), two (bimodal), or more different values (multimodal).
- For nominal data, the mode is the only measure of central tendency that we can use.

Measures of dispersion (variation)

Dispersion refers to the degree of variation in the data; that is, the numerical spread (or compactness) of the data.

Using measures of variation in conjunction with measures of central tendency gives a more complete numerical description of the data.

Range

The **range** is the simplest and is the difference between the maximum value and the minimum value in the data set. It ignores all data points except the two extremes.

- The range is affected by outliers, and is often used only for very small data sets.

Quartiles

Quartiles are the value that divide a list of numbers into quarters:

- The **first quartile** (Q1) is the middle number between the smallest number and the median of the data set.
- The **second quartile** (Q2) is the median of the data.
- The **third quartile** (Q3) is defined as the middle value between the median and the highest value of the data set.
- Distance between first and third quartile is 50% (middle 50% of data) leave out first 25% may be outliers etc. and max 25% peaking middle 50%

Interquartile Range

The **interquartile range (IQR)**, or the mid-spread is the difference between the first and third quartiles, $Q_3 - Q_1$.

This includes only the middle 50% of the data and, therefore, is not influenced by extreme values. It is not sensitive to extreme data values.