Business Analytics (MIS171) Summary Notes

Topic 1 Revision Notes

Business Analytics:

Definition:
-"Process of transforming data into actions through analysis and insights in the context of organisational decision making and problem solving"
-It is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain an improved insight about their business operations and make better, fact-based decisions
-Supported by various tools such as Microsoft excel, and other software packages

Importance of Analytics:
-Data, facts and analysis aid decision making, and that the decisions made on them are better than those made through gut instinct
-Decision making today is even more complicated, due to overwhelming data and information
-There is a strong relationship of use of analytics and profitability and revenye

Evolution of Business Analytics:
-Modern evolution of analytics began with the introduction of computers, as they provided the ability to store and analyze data easily.

Three major components of business analytics:

1. Descriptive Analysis (WANT TO KNOW ABOUT PAST)	-Most commonly used and most well understood type of analytics -Use data to understand past and present performance to make important decisions -Summarizes data into meaningful charts and reports
2. Predictive Analysis (WANT TO KNOW ABOUT FUTURE)	-Analyzes past performance in an effort to predict the future by examining historical data, detecting patterns or relationships in these data -Techniques include: regression and forecasting
3. Prescriptive Analysis (MAKING DECISIONS-OPTIMIZATION)	-Uses optimization to identify the best alternative to minimize or maximize some objective -Addresses questions such as: •How much should we produce to maximize profit? •What is the best way of shipping goods from our factory to minimize costs?

What is Statistics?

Statistics definition:

-"Statistics relates to the collection, analysis, interpretation, and presentation of data"

-Statistical methods are used to:

•Summarize a collection of data

•Draw inferences about an entire population

•Make predictions or forecasts

-Statistics is also the study of variation in data

-Descriptive VS. Inferential statistics:

1. Descriptive statistics:	-Are tabular, graphical, and numerical measures used to summarize data
2. Inferential statistics:	-The process of using data obtained from a sample to make estimates and test claims about the characteristics of a population

Variables:

-Characteristics of items or individuals

-EG. Gender, field of study, money in wallet, time spent in shower each day

-It is essential that all variables have an operational definition: which is defines how a variable is to be measured, otherwise confusion can occur.

Data:

-Observed characteristics of items of individuals.

Populations:

-A collection of all members of a group being investigated

-Two factors need to be specified when defining a population:

•1. The entity (EG. People or motor vehicles)

•2. The boundary

Sample:

-The portion of the population selected for analysis

-EG. Ten full time students selected for a focus group

Parameter:

-A numerical measure of some population characteristic

-EG. The average amount spent by all customers at the local shopping centre last weekend

Statistic:
-A numerical measure that describes a characteristic of a sample
-EG. The average amount spent by the 30 customers completing the market research survey

Data sources:

Four important sources of data:
-Data distributed by an organisation or an individual
-A designed experience
-A survey
-An observational study (such as a focus group)

Primary and Secondary sources:

Primary sources:	-When the data collector is the one using the data for analysis -EG. Internal company records, business transactions, customer market surveys
Secondary sources:	-When another organisation or individual has collected the data that is used for analysis by an organisation or individual -EG. Government and commercial sources, online research

Types of Data:

***BIG DATA* (Data deluge):**
-Many companies have massive amounts of data at their disposal
-This data deluge is a result of:
•Automatic data collection
•Electronic instrumentation
•Online transactional processing
-There is growing recognition of the untapped value in these data bases
-Data is produced in great volumes, in a variety of forms, and is produced very quickly=BIG DATA

1. Categorical data (Qualitative data):
-Labels or names used to identify attributes of each entity
-Can be recorded in either numeric or nonnumeric formats
-EG. 'Yes or no', 'male or female' answers
-Usually counted or expressed as a portion or a percentage

2. Numerical data (Quantitative data):
-Take numbers as their observed responses
-Numerical data can be converted to categorical data. EG Salary can be converted into

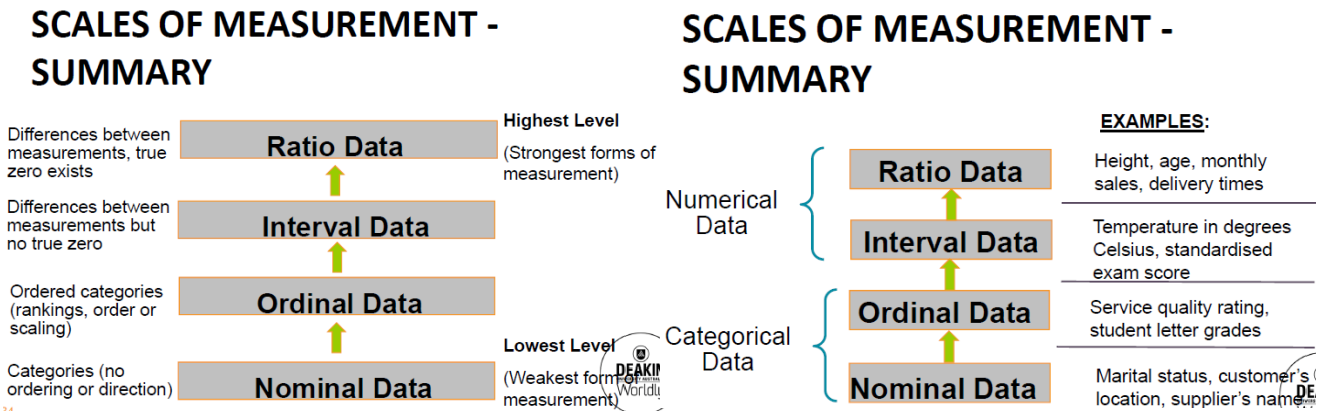
low/medium/high. However you cannot convert categorical data back to numerical data

-There are two types of numerical data:

Discrete:	-If measuring how many (Whole numbers)
Continuous:	-If measuring how much (Decimal places)

Scales of Measurement:

Categorical Measurements	
Nominal:	-A classification of categorical data that implies no ranking -EG. Favorite soft drink, gender
Ordinal:	-Scale of measurement where values are assigned by ranking -EG. Rating customers service as 'very good, good, average, or poor'
Numerical Measurements	
Interval:	-A ranking of numerical data where differences are meaningful but there is no true zero point -EG. Shoe sizes 9, 9.5, 10
Ratio:	-A ranking of numerical data where differences between measurements involve a true zero point -EG. Length, weight, age, salary measurements



Two Broad Types of Data:

Cross-sectional data:	"Relates to a group of items or individuals at a given point of time"
Time ordered (time series) data:	"Relates to a particular entity or situation at different points of time"

Topic 2 Revision Notes

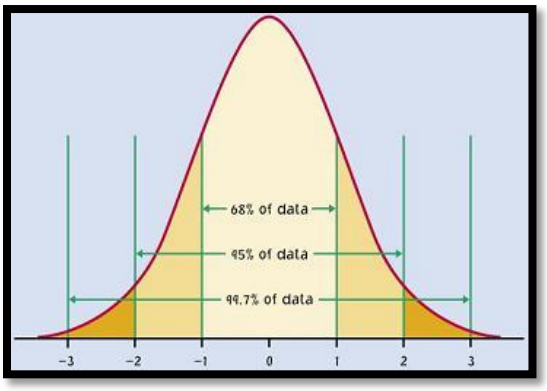
-"Provide a relative measure of the distance an observation is from the mean (in terms of standard deviations)"

$$Z = \frac{X - \bar{X}}{S}$$

-As a general rule a Z score above +3 or below -3 is considered an outlier
-EG. A Z score of 2 means that a value is 2 SDs away from the mean

The Chebyshev Rule (for any data set):
-At least 75% of the data values must be within Z=2 Standard deviations of the mean
-At least 89% of the data values must be within Z=3 Standard deviations of the mean
-At least 94% of the data values must be within Z=4 Standard deviations of the mean

The Empirical Rule (for a data set that is bell-shaped):
-Approx 68% of the data values lie within Z=1 Standard deviations of the mean
-Approx 95% of the data values lie within Z=2 Standard deviations of the mean
-Approx 99.7% of the data values lie within Z=3 Standard deviations of the mean
(ONLY WORKS FOR SYMMETRICAL DATA)



4. Measures of Variability:

WORST MEASURE OF VARIABILITY (starting at range) BEST MEASURE OF VARIABILITY (through to coefficient of variation)

1. Distance Measures:

Range:	-"Difference between largest and smallest data values" =Max – Min
Interquartile Range (IQR):	-"Difference between the third quartile and the first quartile" =Q3 – Q1 -It is the range for the middle 50% of the data

2. Average Variation:

-Measure the average scatter around the mean. That is how larger values fluctuate above it and how smaller values are distributed below it.

Variance (S^2):	-Expressed in square units $s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$
Standard deviation (S):	-"Estimate of the average deviation of individual values away from the mean" -SD is preferred over S^2 because it maintains the original unit $-S(\sigma) = \sqrt{S^2}$

3. Relative Variation:

Coefficient of Variation:	-"Indicates how large the standard deviation is in relation to the mean" $\text{CoV} = \frac{s}{x} \times 100$ -Useful for comparing variability between data sets in different units -EG. Relative to the mean, the package volume is more variable than the package weight

Summary:

-The more spread out the data: the larger the range + IQR + SD

-The more concentrated or similar the data: the smaller the range+ IQR + SD

-If the value are the same: the range + IQR + SD will be zero

-No measure of variation can ever be negative

5. Shape:

Symmetrical Data (normal distribution):