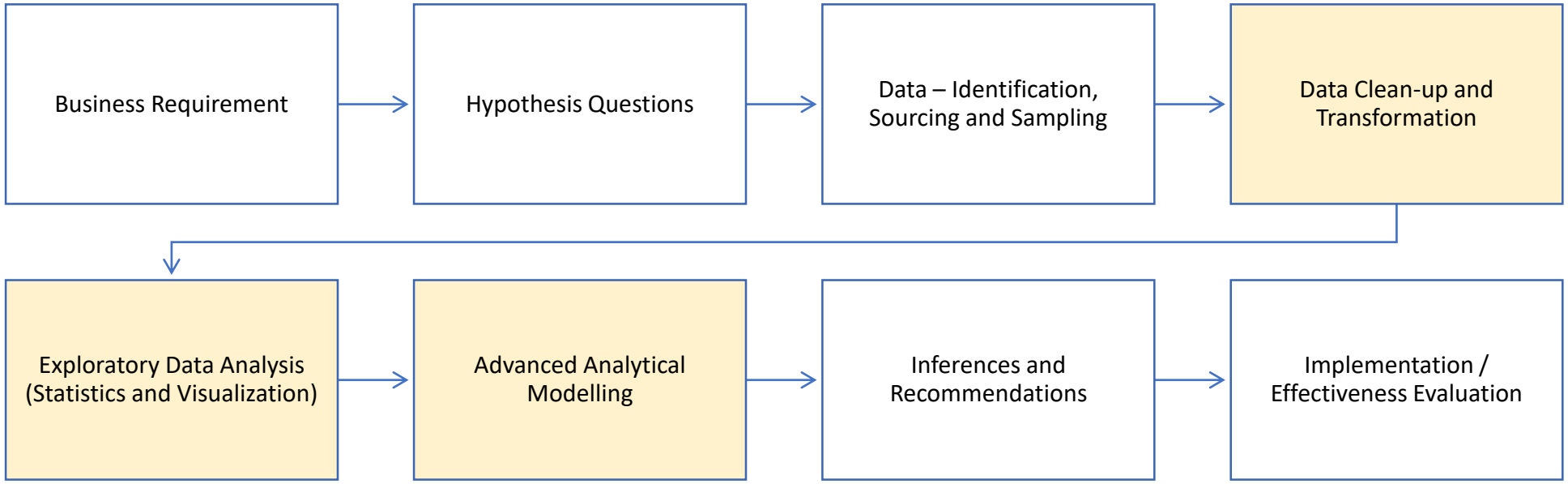
Data Analytics Methodology

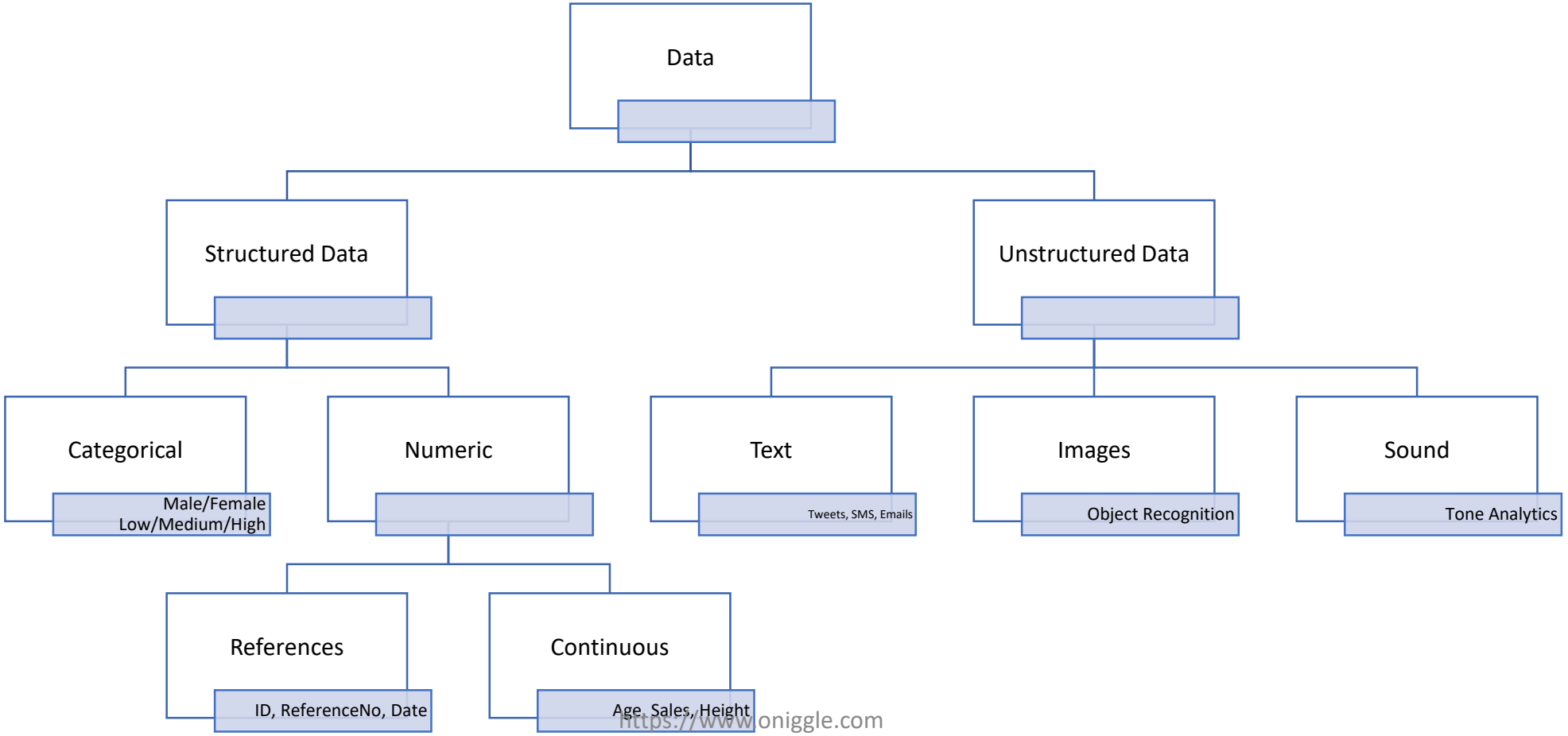


Analytical Modelling Techniques

Almost all the business problems are handled within a combination of these techniques

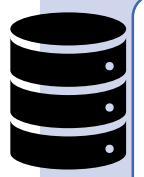
Business Problem	Analytical Technique	Algorithms
When you are trying to solve a business problem such as costs, revenue, margins, mileage of a car, viscosity of tires where the **target variable is numeric** in nature	Regression	Ordinary Least Square, Ridge, Lasso, Decision Tree, Neural Network, Random Forest …etc.
When you trying to solve a problem such as potential for loan default, risk level of an insured, presence or absence of disease where the **target variable is categorical** variable or choices instead of numeric)	Classification	Logit, Decision Tree, Random Forest, Neural Network, XGBOOST, LIGHT GBM, SVM …etc.
When you want to perform data mining to understand natural grouping of customers (like market/customer segmentation) or association of items **without a target variable**	Clustering, MBA, PCA	K-Means, Hierarchical…etc Apriori, FP-Growth

Data Types


Data
Structured Data
Unstructured Data
Categorical
Male/Female
Low/Medium/High
Numeric
Text
Tweets, SMS, Emails
Images
Object Recognition
Sound
Tone Analytics
References
ID, ReferenceNo, Date
Continuous
Age, Sales, Height

Data Sources

Existing

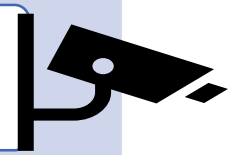


Internal Data from the systems (e.g. Sales, Payroll, Attrition, Earnings Data)

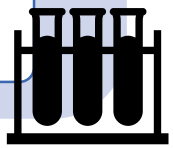
External Data from other sources (e.g. Census Data, Journals, Publications)

Generated

Survey (e.g. Feedback on the new feature)



Observed Data (e.g. Traffic Pattern at an intersection)



Experimental Study Data (e.g. Experiment and Control group for new drug)

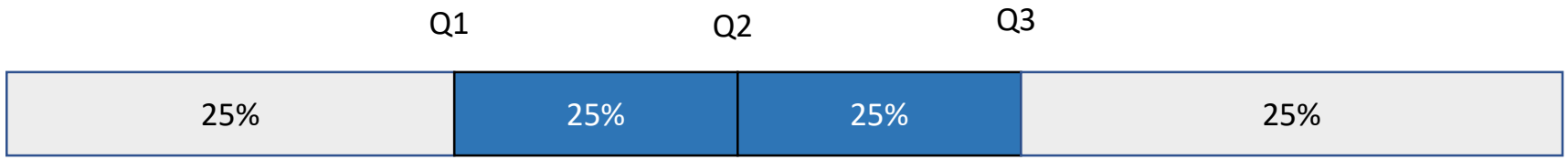
Measures of Central Tendency

Term	Definition	Example
Mean	Arithmetic Mean or Average is Σ Values / (No. of Values)	X = 8,1,2,4,6,0,7 Mean(x) = 28 / 7 = 4

Term	Definition	Example
Median	Mid point of a sequence of numbers arranged in alphabetical order (or average of mid 2 points if even)	X = 1,2,4,6,7,8,10 Median (x) = 6 X = 1,2,4,6,7,8,10, 11 Median (x) = (6 + 7) / 2 = 6.5

Term	Definition	Example
Mode	Number that occurs in maximum frequency in a given sequence of numbers	X = 1,2,2,3,2,4,5,4,4,4,4,5 Mode (x) = 4

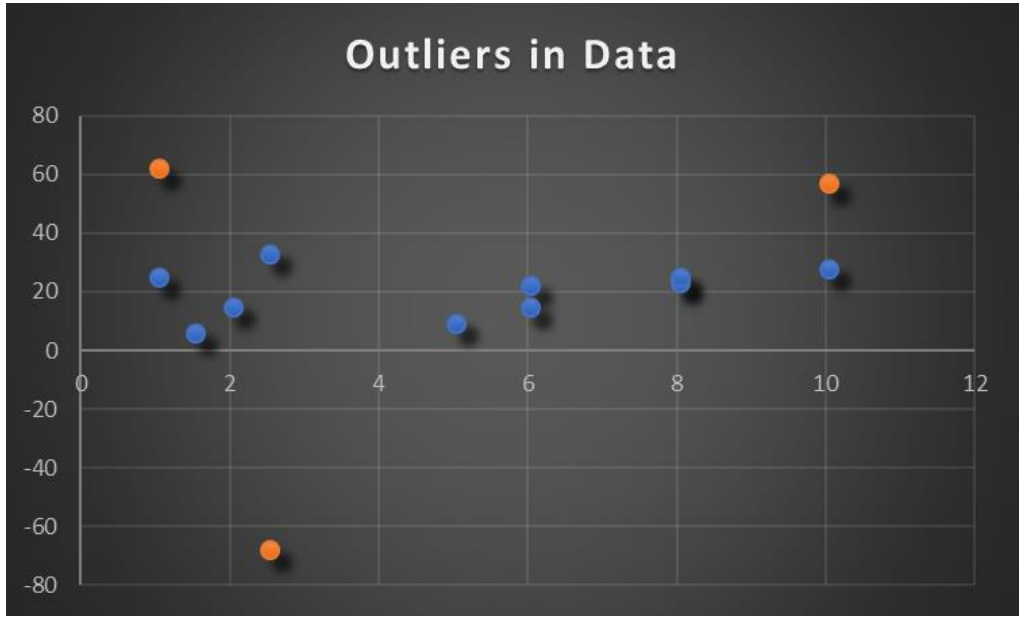
Central Region - Quartile



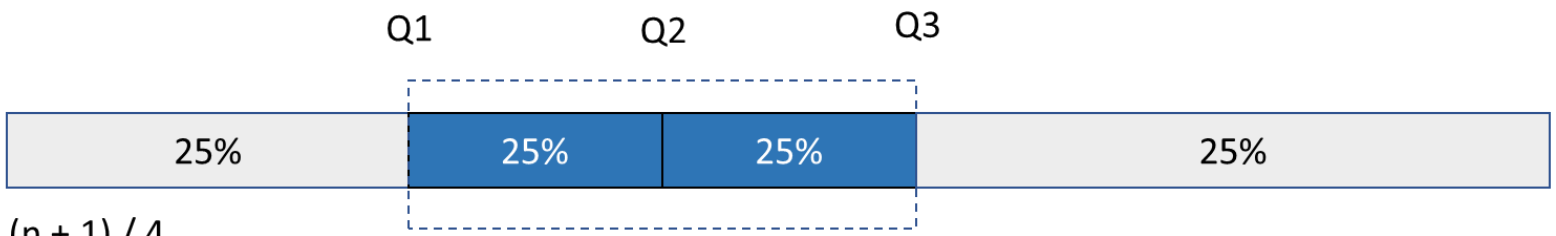
Term	Definition	Example
Quartile	Divide the group into 4 parts each with 25% of the data Q1 represents the point where you find 25% of the data below it (same as 25th percentile) Q1 = Value at (n + 1)/4 Q2 represents the point where you find 50% of the data below it (also same as median and 50th percentile) Q2 = Value at 2* (n + 1)/4 Q3 represents the point where you find 75% of the data below it (same as 75th percentile) Q2 = Value at 3* (n + 1)/4	X = 100, 105, 107, 120, 125, 135, 145, 147, 150, 152, 152,154, 156, 165, 168, 170 There are n=16 values Q1 = Value at (16+ 1)/4 = Value at 4.5 Which is average of 120 and 125 = 122.5 Q2 = Median = (147 + 150) / 2 = 148.5 Q3 = Value at 3 * (16 + 1)/4 = Value at 12.75 which is average between 154 & 156 = 155

Extreme Values - Outliers

- Outlier is an extreme value in your observations. Its an observation point that is distant from other observations in that group of data



Outliers



Q1 = (n + 1) / 4
Q2 = 2(n + 1) / 4
Q3 = 3(n + 1) / 4
n = number of observations

Example
X = 0,1,2,3,3,3,3,3,3,3,3,4,5,**5**,6,6,7,8,8,9,9,9,9,10,12,20
n = 25, Q2 position = 2 * (25 + 1) / 4 = 13th value which is 5

0,1,2,3,3,3, (Q1) ,3,3,3,3,4,5, Q2(**5**), 6,6,7,8,8,9, (Q3) 9,9,9,10,12,20