AWSMOSIS
Transition Redefined...

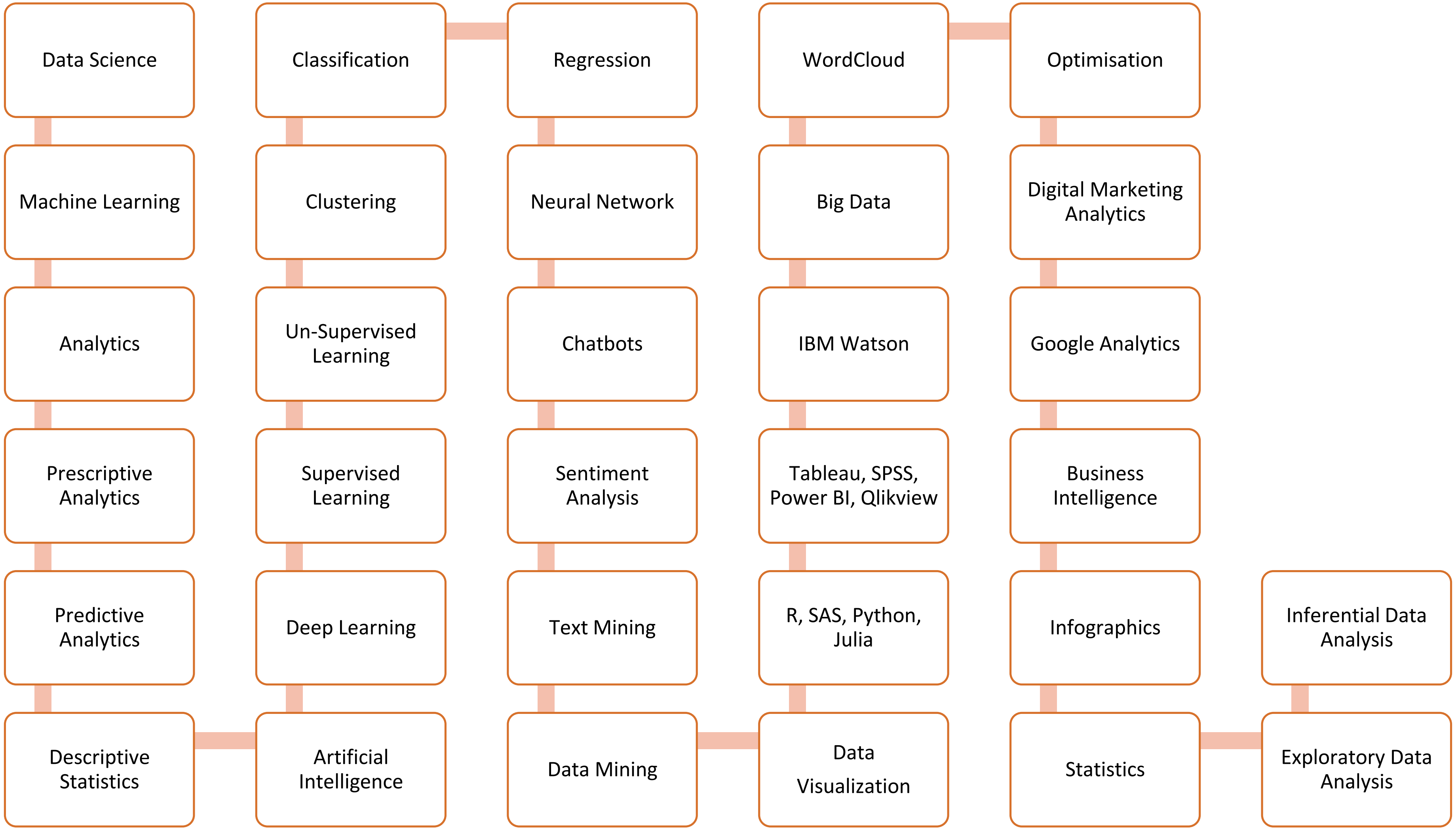
Data Analytics

Awsmosis Learning & People Solutions Pvt Ltd (A.L.P.S.)

Useful References

- www.Kaggle.com (Analytics Competition site, which gives you ideas on how companies are leveraging data analytics to solve business problems)
- http://www.informationisbeautiful.net/ (Visualization)
- http://flowingdata.com/ (Visualization)
- https://github.com/d3/d3/wiki/Gallery (Visualization)
- https://appsource.microsoft.com/en-us/marketplace/apps?product=power-bi-visuals&page=1&src=office&corrid=4b40ed50-9a49-424b-89b8-5438c04a1707&omexanonuid=efea02b2-6e57-4066-a723-998c1c3484a8

Terminologies in the Data Science World



Data Science

Machine Learning

Analytics

Prescriptive Analytics

Predictive Analytics

Descriptive Statistics

Artificial Intelligence

Deep Learning

Supervised Learning

Un-Supervised Learning

Clustering

Classification

Regression

Neural Network

Chatbots

Sentiment Analysis

Text Mining

Data Mining

Data Visualization

R, SAS, Python, Julia

Tableau, SPSS, Power BI, Qlikview

IBM Watson

Big Data

WordCloud

Optimisation

Digital Marketing Analytics

Google Analytics

Business Intelligence

Infographics

Statistics

Exploratory Data Analysis

Inferential Data Analysis

Car Industry - Examples of business requirements

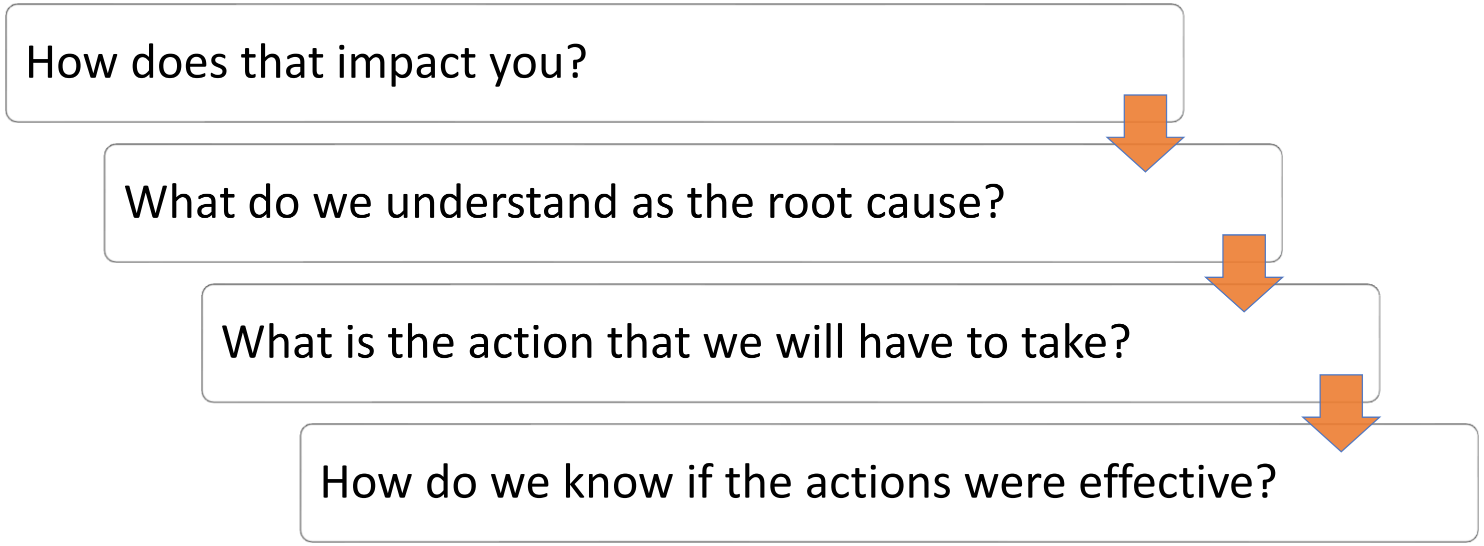
- To improve car sales you want to understand your customer better and personalize your incentives
- You have made various marketing investments and you would like to understand how effective they were in boosting sales
- You would like to understand about customers talking about you on social media
- You want to improve customer loyalty by providing better customer experience
- You have to improve our predictions for requirements of Aftermarket (spare and services)
- You would like to build a driverless car
- You would like to understand how customers use your cars (local, city, long distance)

These are all nice business requirements, but …

… do we even know if we have a problem?

Example of a business problem statement

We have a problem of customer churn!



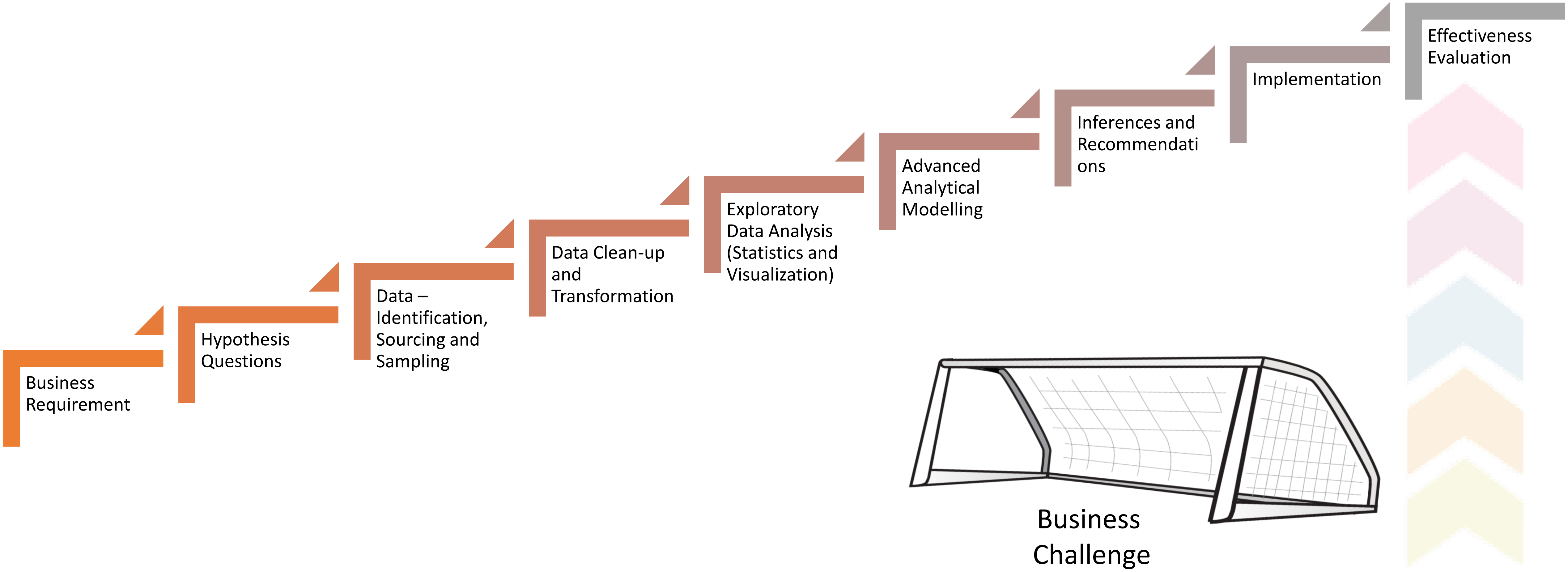
How does that impact you?

What do we understand as the root cause?

What is the action that we will have to take?

How do we know if the actions were effective?

Data Analytics Methodology



Business Requirement

Hypothesis Questions

Data – Identification, Sourcing and Sampling

Data Clean-up and Transformation

Exploratory Data Analysis (Statistics and Visualization)

Advanced Analytical Modelling

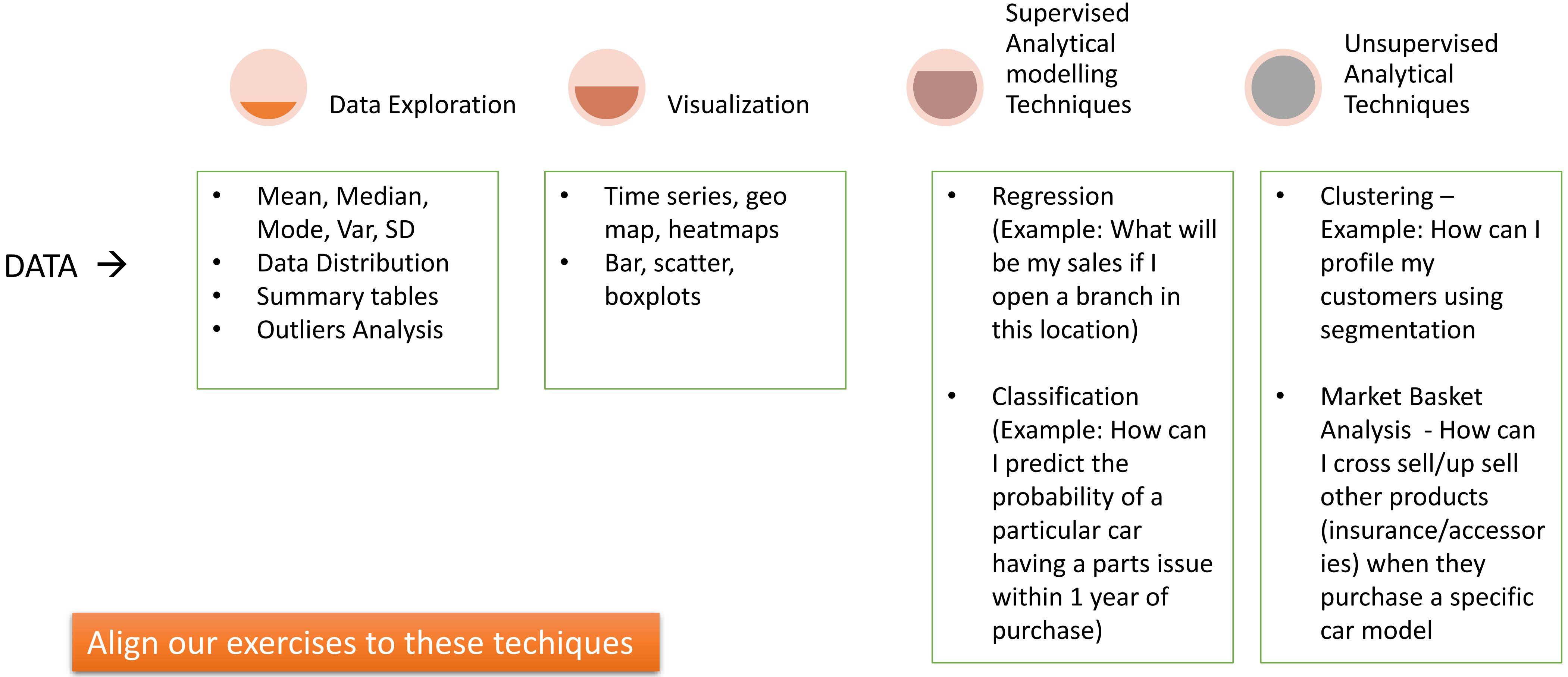
Inferences and Recommendations

Implementation

Effectiveness Evaluation

Business Challenge

Data Analytics Stages – Deeper Look



DATA →

Data Exploration

- Mean, Median, Mode, Var, SD
- Data Distribution
- Summary tables
- Outliers Analysis

Visualization

- Time series, geo map, heatmaps
- Bar, scatter, boxplots

Supervised Analytical modelling Techniques

- Regression (Example: What will be my sales if I open a branch in this location)
- Classification (Example: How can I predict the probability of a particular car having a parts issue within 1 year of purchase)

When you have a specific question on an interest variable

Unsupervised Analytical Techniques

- Clustering – Example: How can I profile my customers using segmentation
- Market Basket Analysis - How can I cross sell/up sell other products (insurance/accessories) when they purchase a specific car model

When you are fishing for information

Align our exercises to these techiques

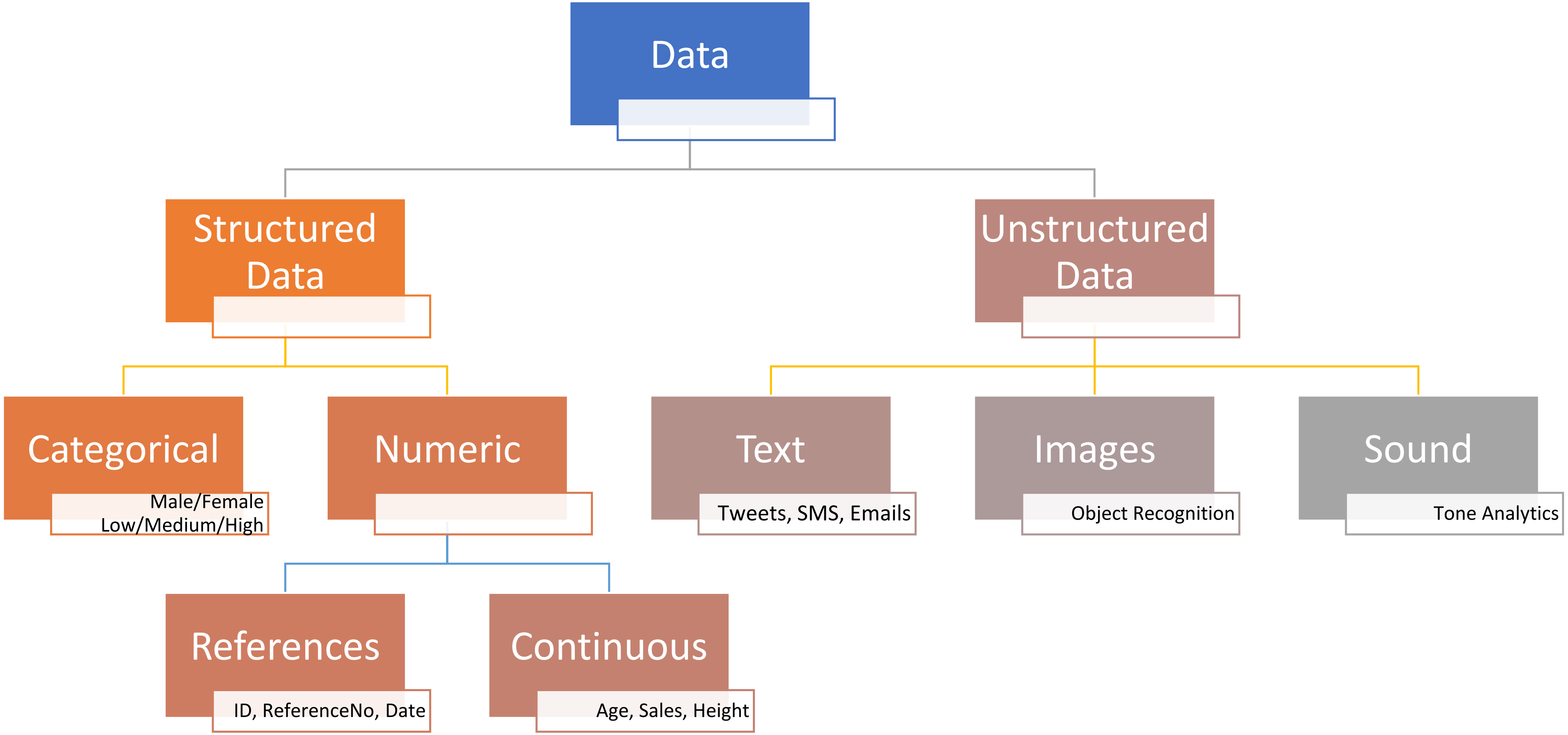
Inferences - Keep the explanation simple

Business Requirement	Hypothesis	Data Sourcing & Sampling / Clean-up	Data Exploration / Visualization	Analytical Modelling	Recommendations
Target Variable(s)	Input Variables V1 V2 V3 .. .. Vx. Creativity	V1 V2 V4 V7 V13 V15	V1 V2 V4 V7 V13 V15	V1 V2 V7 V15 V21 V24	V1 V2 V15 .. .. .. Simplicity
	Imagine every possible contributing variables using business intuition and structured design thinking process	Practicality of data collection process eliminates some insights	Hotspots are identified (probable causes)	Variables not correlated with Target Variables are eliminated	80-20 Models prioritise important variables

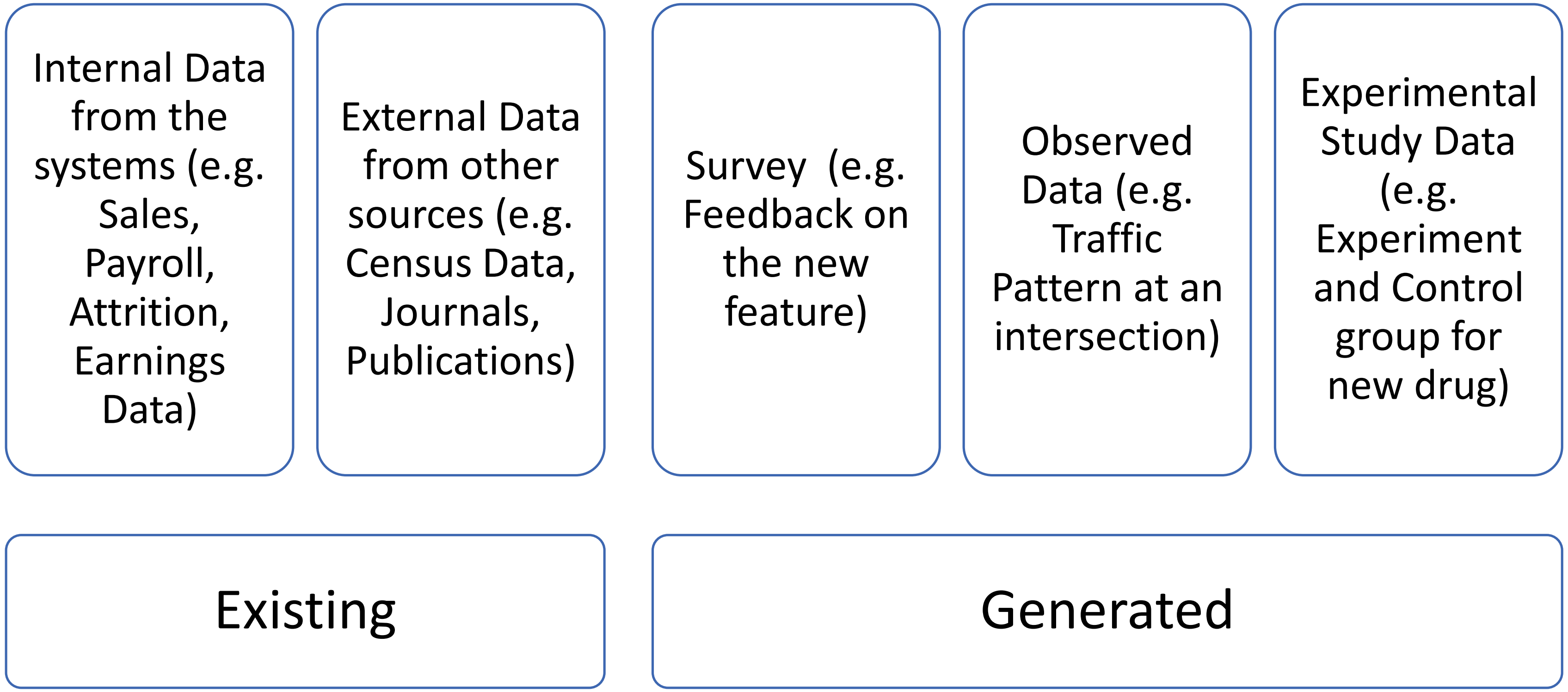
Guide to Data Preparation

- Data Types
- Data Sourcing
- Data Sampling
- Data Clean up

Type of Data


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



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

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

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)

Existing

Generated

Sampling

Why should we perform Data Sampling?

Because …

- Getting data for full population is simply not possible
- It is very expensive
- It is time consuming
- It takes a lot of effort

Sampling Types

What is the average age of male who participated in the 10K Run in Mumbai marathon in 2018

44	155	123	49	148	73	157	195	98	54
48	112	6	171	116	185	175	181	136	135
53	163	120	194	123	192	119	113	48	106
62	11	80	140	10	137	90	9	22	183

SIMPLE RANDOM
Data is selected completely in Random from original population.

Sampling Types - continued

What is the average age of male who participated in the 10K Run in Mumbai marathon in 2018

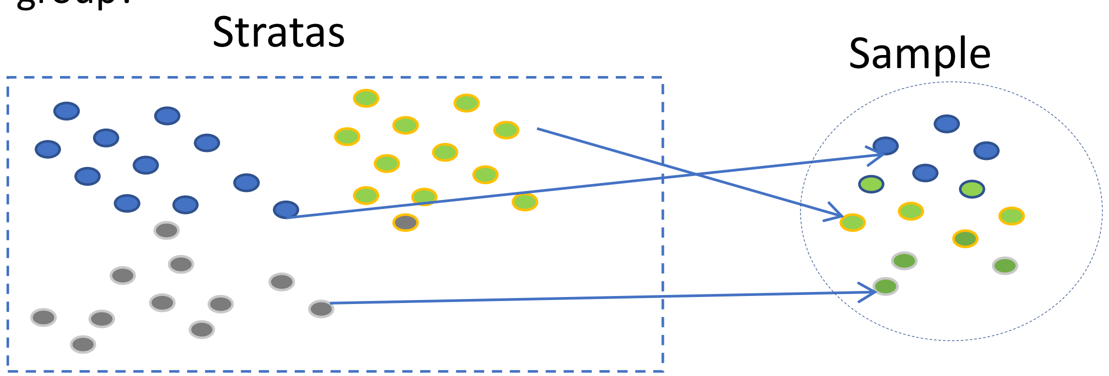
Randomly Chosen column

6	6	9	9	10	10	10	11
22	22	26	44	48	48	48	49
53	54	54	62	73	73	80	80
90	90	98	98	106	106	112	113
113	116	116	116	116	119	119	120
120	123	123	123	123	135	135	136
136	137	137	137	140	140	148	148
155	157	157	163	171	171	171	175
175	181	181	183	183	185	185	185
185	192	192	192	194	194	195	195

SYSTEMATIC RANDOM
Data is selected in Random from original population within a systematic process.

Sampling Types - continued

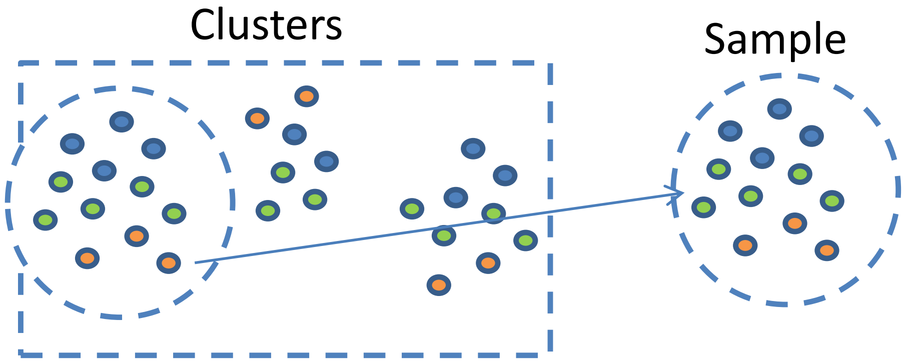
If you are selling a product, how do you ensure that you have marketed to a variety of folks in different age group?



STRATIFIED RANDOM
A heterogeneous data sample is created from homogeneous groups

Sampling Types – cont.

If you are running for election, how do you ensure that you have campaigned to a good sample of cities, villages and towns?



CLUSTER RANDOM
A heterogeneous data sample is created from heterogeneous groups

Sampling Types – cont.

You are opening a new shop in a new geography, but you have no idea about local preferences

Expert Opinion
Take opinion of a few experts to determine your hypothesis.

You do not have time to do an expensive survey for a product. You decide to partner with a local grocer

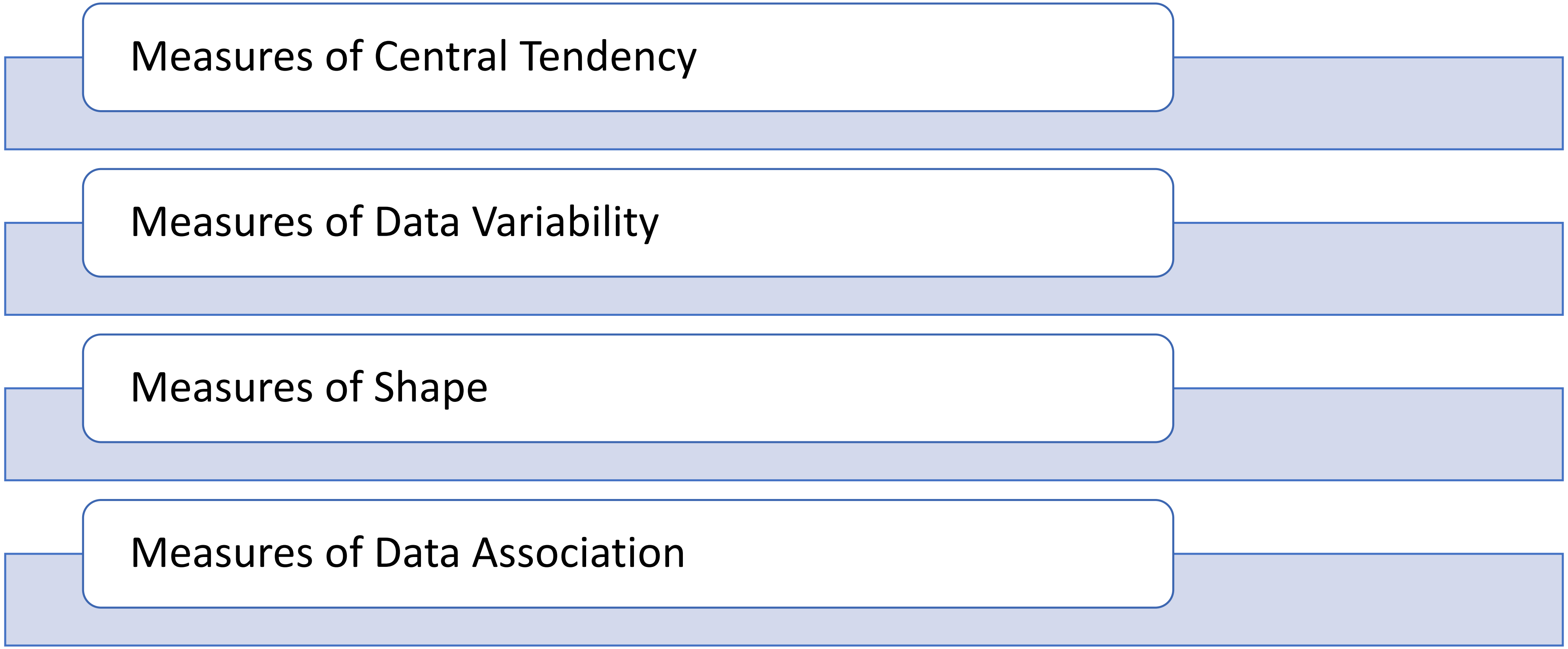
Convenience Sampling
You talk to only those customers who come into the store.

Missing Data Treatment

How do we treat missing values?

- Replace with zero
- Replace with Mean or Median
- Delete entire Observation
- Replace with alternate source
- Business logic based data update
- Manual correction from source

Data Awareness



Measures of Central Tendency

Measures of Data Variability

Measures of Shape

Measures of Data Association

Central Region - Mean

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

Central Region - Median

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

Central Region - Mode

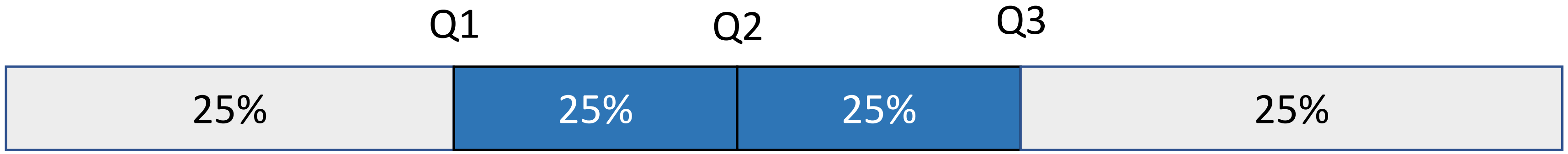
Term	Definition	Example
Mode	Number that occurs in maximum frequency in a given sequence of numbers	X = 1,2,2,3,2,4,5,4,4,4,4,5 Mode (x) = 4

Central Region - Percentiles

Term	Definition	Example

Percentiles	Divide the data into 100 parts and “Percentile” indicates what % of data observations are below that value
	
	i = (P / 100) * n
	i = Percentile Location
	P = Percentile
	n = sample size
		Revenue for Car Dealers (in crores) are as follows
		7.2, 1.2, 4.8, 1.5, 1.8, 2.1, 4.3, 6.2, 4.3, 1.7, 1.9, 2.2, 4.3
		
		Which revenue represents 30% percentile?
		
		Arrange the data in ascending order
		1.2 1.5 1.7 1.8 1.9 2.1 2.2 4.3 4.3 4.3 4.8 6.2 7.2
		
		i = 30 / 100 * 13 = 3.9
		The 4th element represents the revenue at 30th percentile which is 1.8
		
		1.2 1.5 1.7 **1.8** 1.9 2.1 2.2 4.3 4.3 4.3 4.8 6.2 7.2

Central Region - Quartile



Q1 Q2 Q3

25% 25% 25% 25%

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

Measure of Variability - Variance

Term	Definition
Variance	Variance is a measure of spread of the values Variance = Σ (Xi – Mean)^2 / No. of Values (where i = 1 to No. of Values)

X	Mean	Diff	Squared Difference
60	86.33	-26.33	693.2689
70	86.33	-16.33	266.6689
83	86.33	-3.33	11.0889
92	86.33	5.67	32.1489
101	86.33	14.67	215.2089
112	86.33	25.67	658.9489

Total Sum Squared	1877.333
Variance	312.8889
SD	17.68867

Measure of Variability – Standard Deviation

Term	Definition
Standard Deviation	SD = √ Variance

X	Mean	Diff	Squared Difference
60	86.33	-26.33	693.2689
70	86.33	-16.33	266.6689
83	86.33	-3.33	11.0889
92	86.33	5.67	32.1489
101	86.33	14.67	215.2089
112	86.33	25.67	658.9489
		Total Sum Squared	1877.333
		Variance	312.8889
		SD	17.68867

Population versus Statistic

- Population is entire data set and Parameter is a variable of the population (like age, weight)
- Sample is a subset of Population and Statistic is a variable of the Sample (like age, weight)

Population	Sample
Population size = N Mean $\mu = \Sigma x_i / N$	Sample Size = n Mean $x = \Sigma x_i / n$
Variance $\sigma^2 = (\Sigma(x_i-\mu)^2 / N)$	Variance $\sigma^2 = \Sigma(x_i-\mu)^2 /$ **(n-1)**
Standard Deviation = σ = √ $(\Sigma(x_i-\mu)^2 / N)$	Standard Deviation σ = √ $(\Sigma(x_i-\mu)^2 /$ **n-1**$)$
	Standard Error of the sample mean = σ / √n

Quick Question: What should be a minimum sample size as a general rule of thumb?

Measure of Variability – Coefficient of Variation

Term	Definition
CV	CV = $\sigma \div \mu$ * 100 Standard Deviation / Mean * 100

X	Mean	Diff	Squared Difference
60	86.33	-26.33	693.2689
70	86.33	-16.33	266.6689
83	86.33	-3.33	11.0889
92	86.33	5.67	32.1489
101	86.33	14.67	215.2089
112	86.33	25.67	658.9489

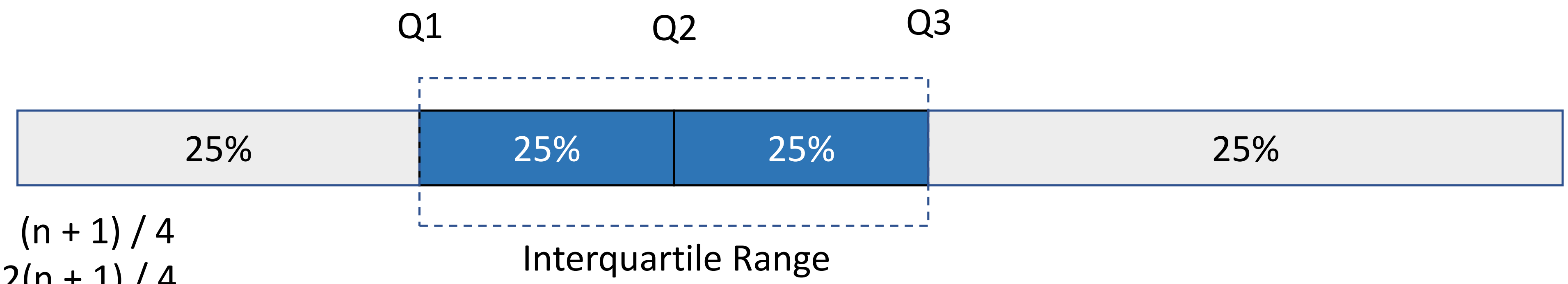
Total Sum Squared	1877.333
Variance	312.8889
SD	17.68867
CV	20.47

Outliers in Data

What is an outlier?

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

Identifying Outliers with Box Plots



Q1 Q2 Q3

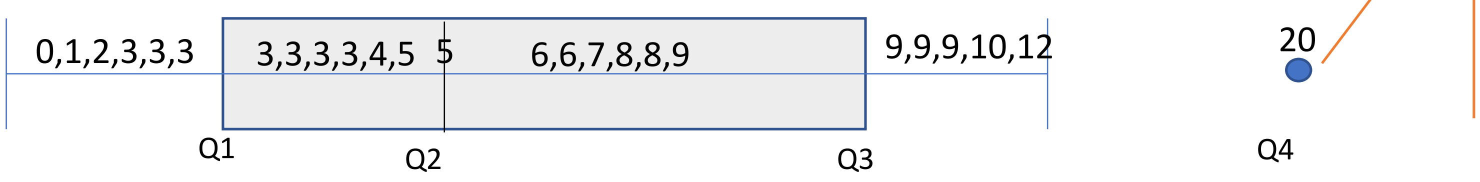
25% 25% 25% 25%

Interquartile Range

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

Example
X = 0,1,2,3,3,3,3,3,3,3,4,5,**<u>5</u>**,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(**<u>5</u>**), 6,6,7,8,8,9, (Q3) 9,9,9,10,12,20



0,1,2,3,3,3 3,3,3,3,4,5 5 6,6,7,8,8,9 9,9,9,10,12 20

Q1 Q2 Q3 Q4

Outlier is
any point > Q3 + 1.5 IQR
Or
any point < Q1-1.5*IQR

IQR = Q3-Q1 = 9-3 = 6
Q3 + 1.5 IQR = 9 + 1.5*6=18
Outlier > 18

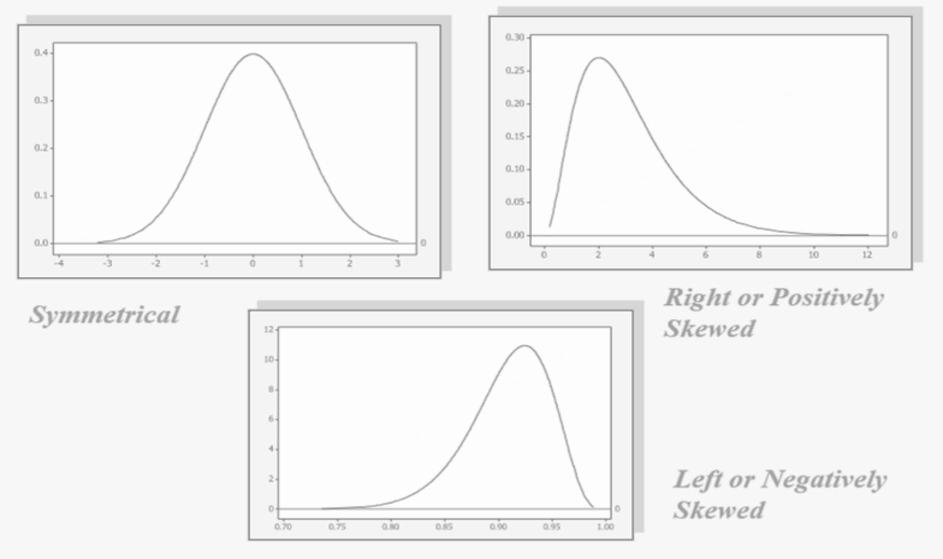
Outlier Treatment

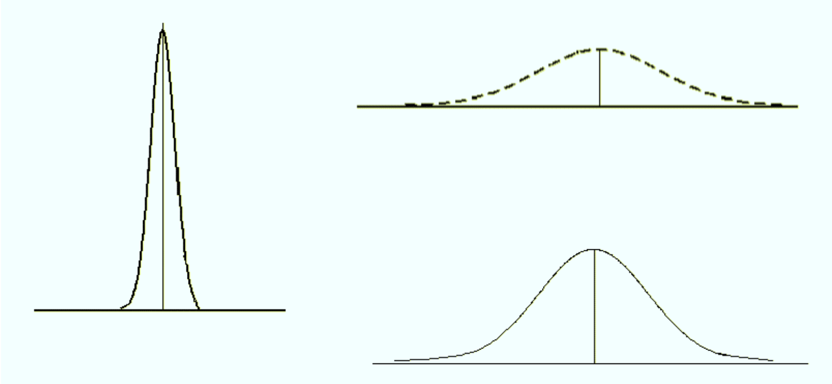
How to deal with outlier?

- Remove entire observation
- Replace with mean/median
- Manually correct errors
- Apply Log transformation or scaling
- Retain outlier as-is
- Analyse only the outliers

Measure of Shape

Skewness: Measure of absence of Symmetry

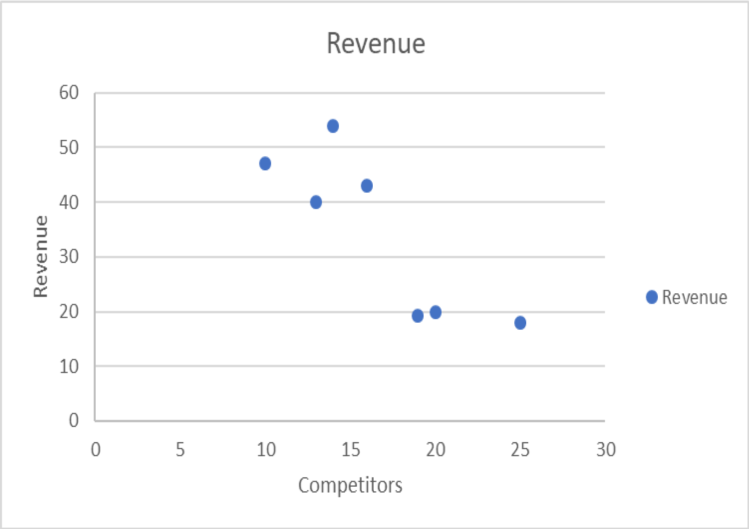


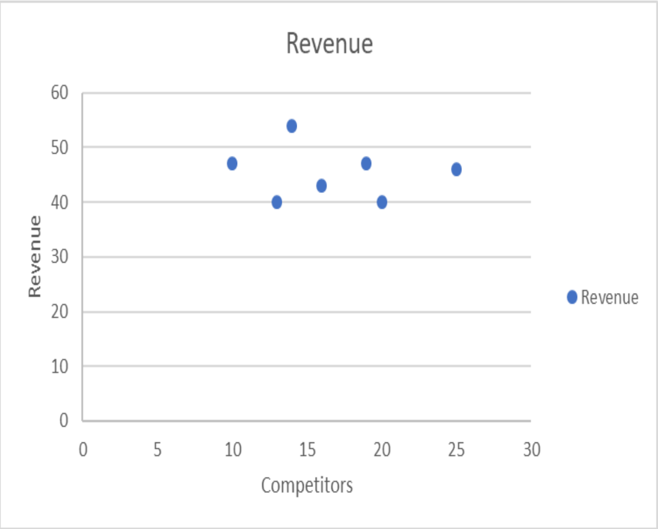


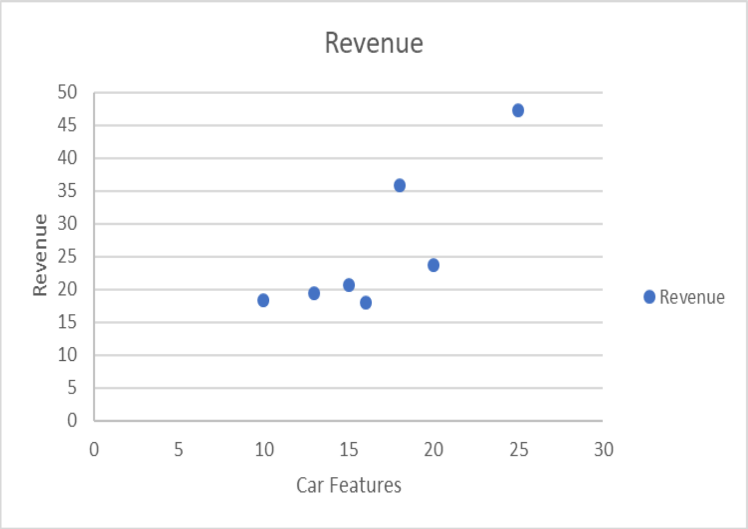
Kurtosis: Peakedness of a distribution

Measure of Association

Correlation: Measure of degree relatedness of Numeric Data
It varies between -1 to +1







Correlation = -0.84679

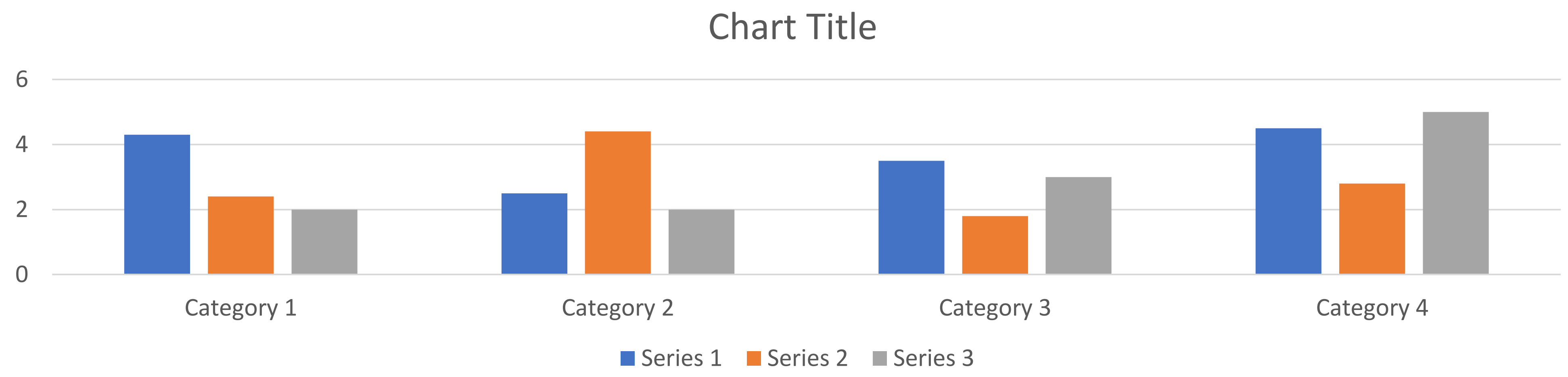
Correlation = -0.14989

Correlation = 0.828728

Key considerations in Visualization

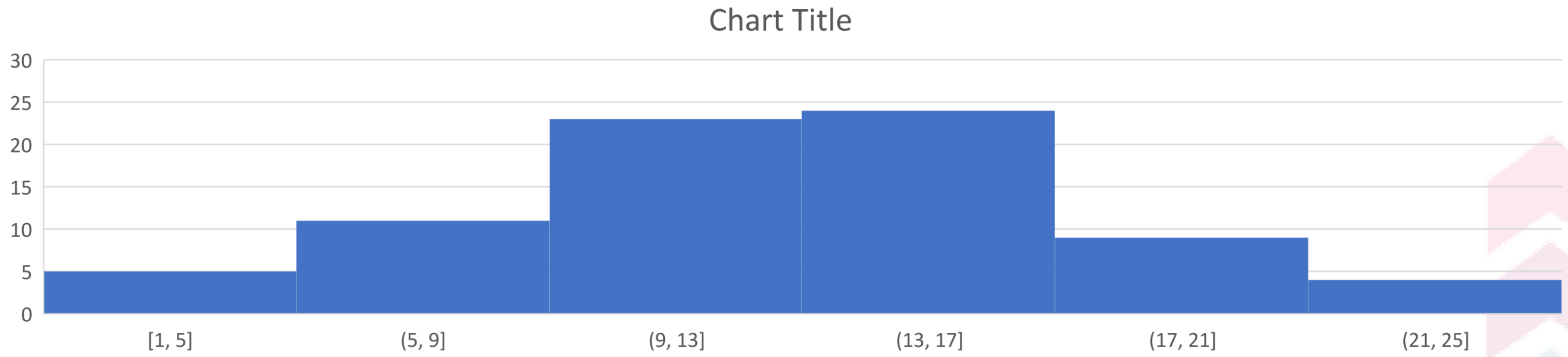
- Every aspect of the chart should tell a separate story
- Colour is not just for visual appeal, but should indicate a data characteristic
- Use the right chart for the right purpose
- Its not always readymade charts:- Use your creativity in drawing your own analytics story

When to use a bar chart



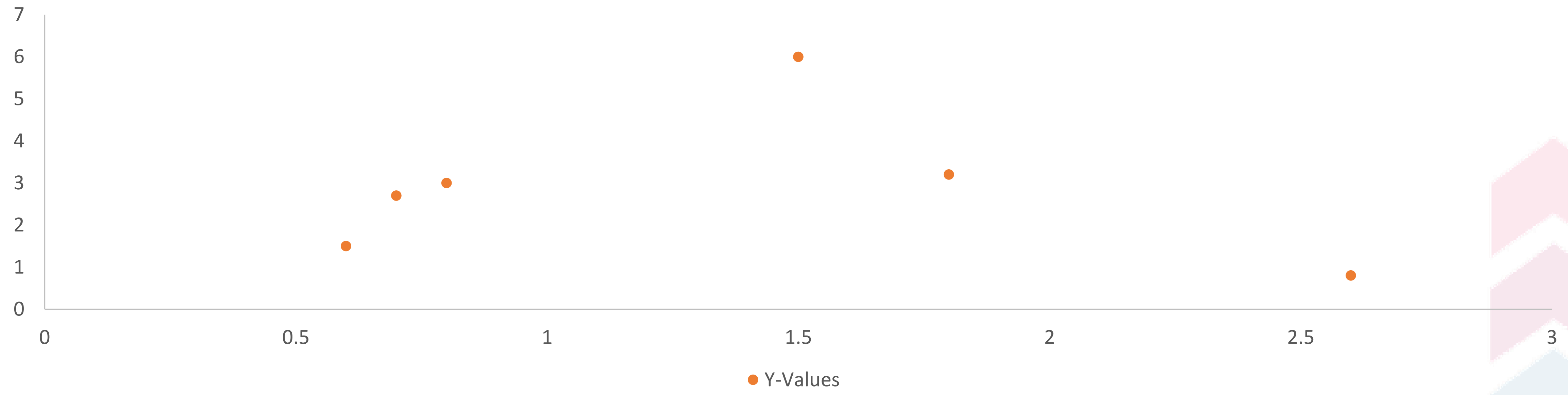
When you are comparing values of categorical variables

When to use a histogram



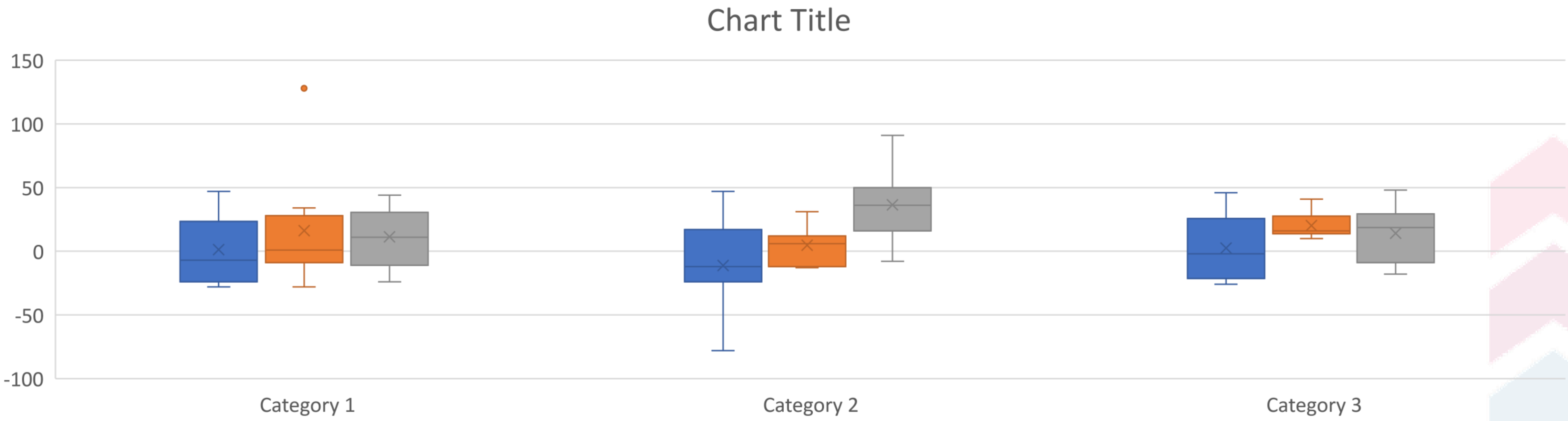
When you are looking at frequency distribution

When to use a scatter chart



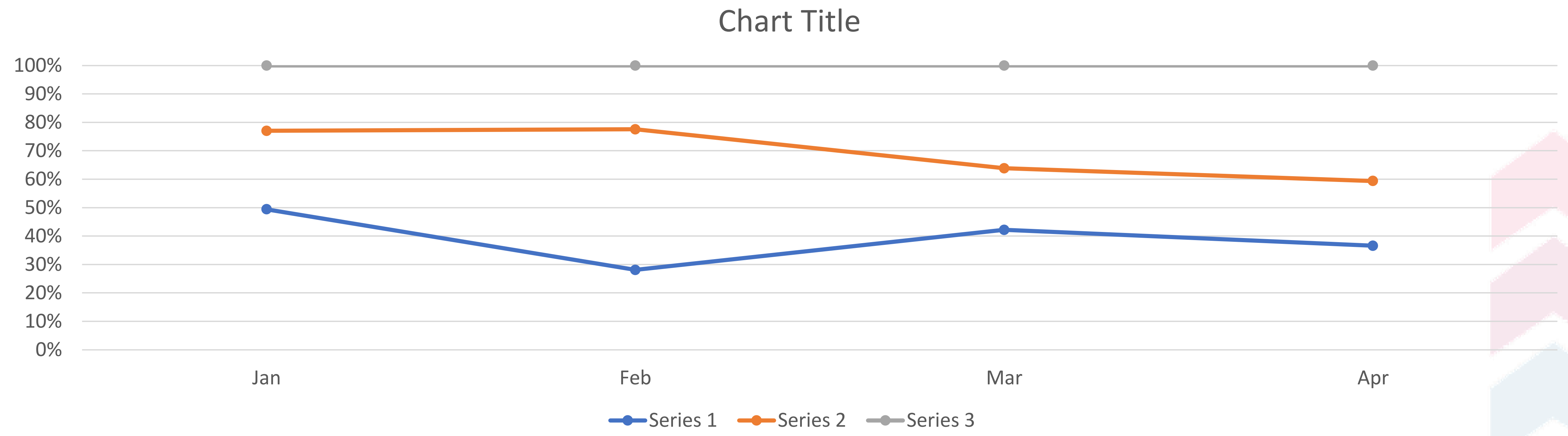
When you are identifying correlation between 2 variables

When to use a boxplot



When you are looking for outliers

When to use a Line chart



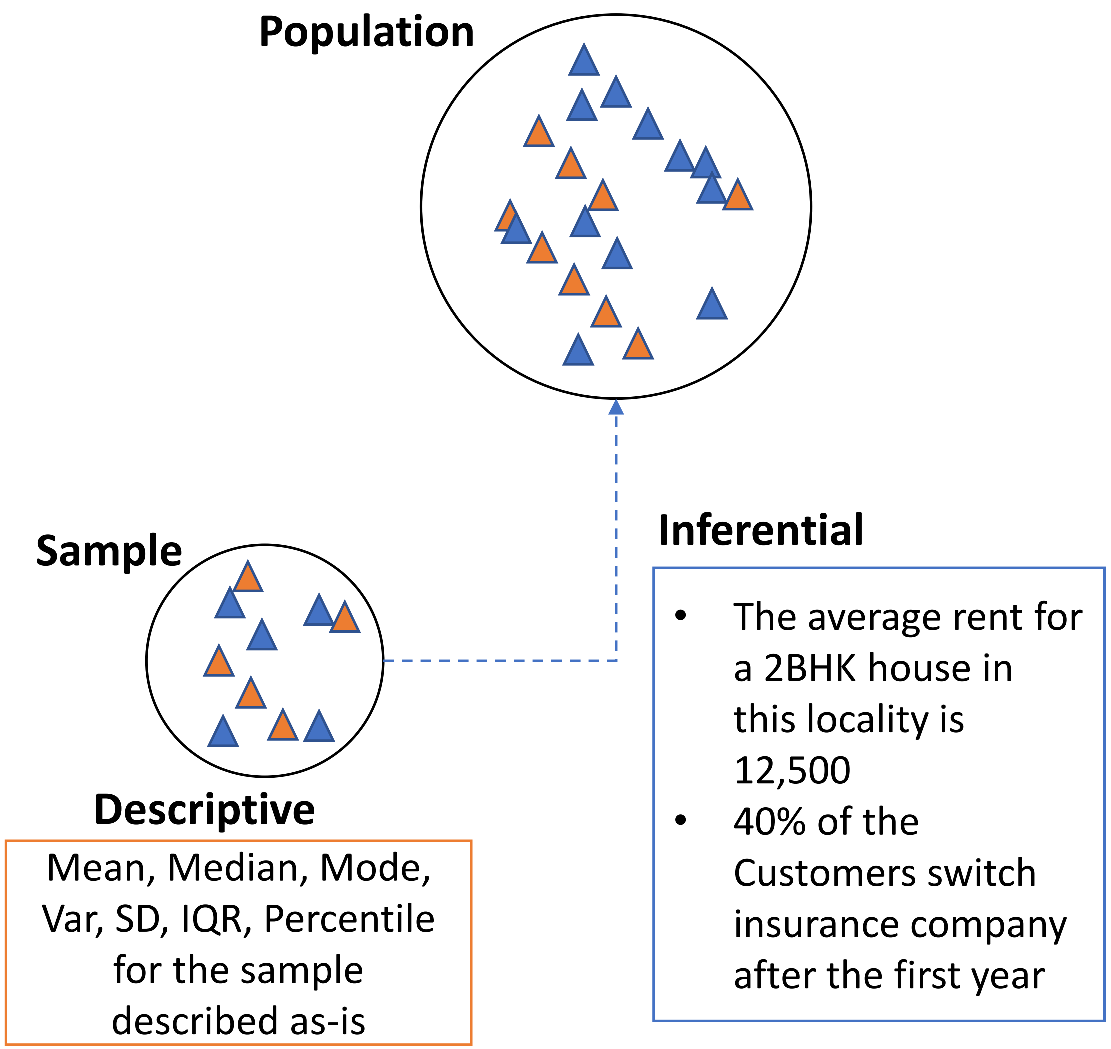
When you are comparing values as a function of Time

Inferential Statistics

- You are aware of a descriptive statistics - You have a sample and you can identify the mean, median, mode, standard deviation, variance, quartiles…etc

- Now based on this, when you draw conclusion about a population, it becomes inferential statistics

- The important question is how "confident are you?" when you make a statement about the population



Population

Sample

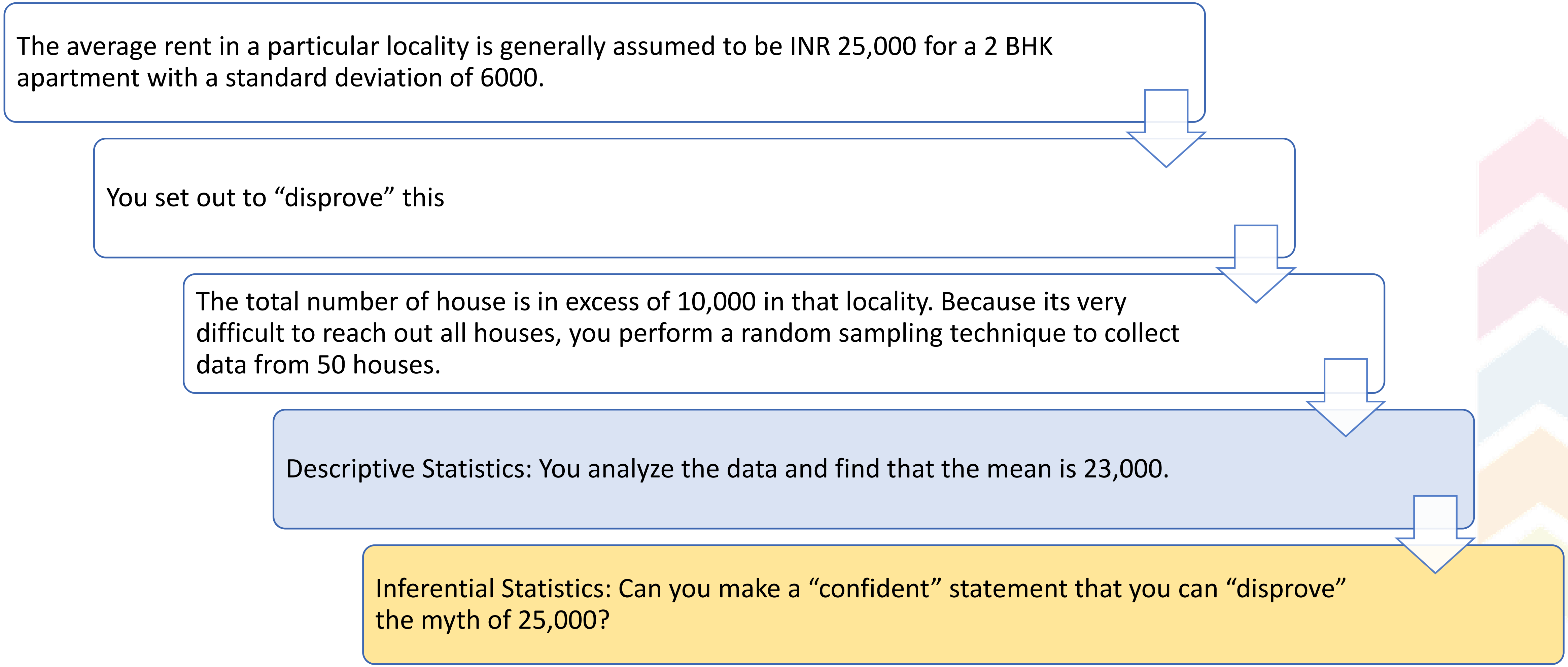
Inferential

- The average rent for a 2BHK house in this locality is 12,500
- 40% of the Customers switch insurance company after the first year

Descriptive

Mean, Median, Mode, Var, SD, IQR, Percentile for the sample described as-is

Lets take up an example to understand this



The average rent in a particular locality is generally assumed to be INR 25,000 for a 2 BHK apartment with a standard deviation of 6000.

You set out to “disprove” this

The total number of house is in excess of 10,000 in that locality. Because its very difficult to reach out all houses, you perform a random sampling technique to collect data from 50 houses.

Descriptive Statistics: You analyze the data and find that the mean is 23,000.

Inferential Statistics: Can you make a “confident” statement that you can “disprove” the myth of 25,000?

Confidence Interval

- When you are pushed to make a “confident statement” you become a bit defensive
- So, instead of stating that the rent 23000 (called as Point Estimate), you are better off saying 23000 +/- something (called as Confidence Interval)

Statistically it is written as

Point Estimate of Sample +/- Confidence Interval

CI = Z * SE = Z * (SD / $\sqrt{n}$)

Z takes the value of 1.96 @ 95% confidence level

n = sample size

CI = 1.96 * 6000 / $\sqrt{50}$ = 1663

Rent is 23000 +/- 1663

If the situation is critical and you need to provide an answer @ 99% confidence, then use a Z value of 2.58

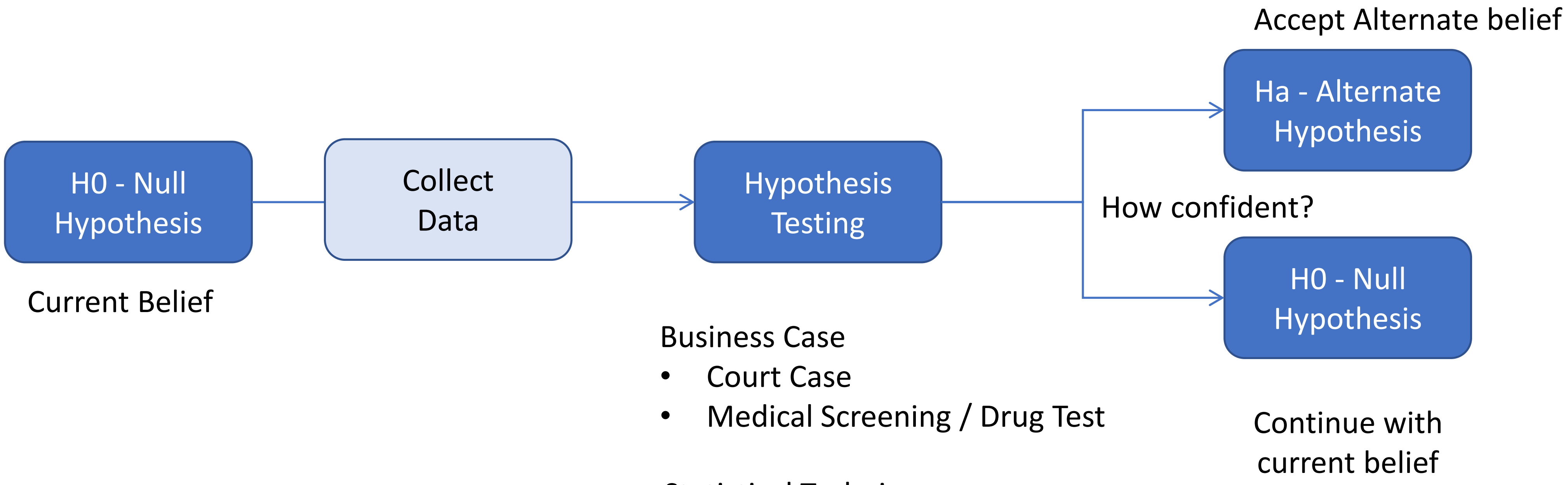
CI = 2.58 * 6000 / $\sqrt{50}$ = 2189

You could say with “95% confidence” the rent is between 21337 and 24663 and challenge the current belief 25000 (which is outside this range)

Then the range is between 20811 and 25189 which includes the current belief. You just failed to “disprove” the current belief

Hypothesis Test

What we just did can be called as a Hypothesis Test



Business Case

- Court Case
- Medical Screening / Drug Test

Statistical Techniques

- Z-distribution
- T student distribution
- Anova
- F Statistic
- Chi Square Test

When to use different statistical tests

Hypothesis Tests using distribution	When to use	Example/Use Case
Z-Distribution	• When you are comparing means between sample and population (or) 2 sets of samples and you have known population mean and Standard Deviation	• The mean monthly cell phone bill in a city is μ = 400
T-Distribution	• When you are comparing means between sample and population (or) 2 sets of samples, but you have unknown population standard deviation. • When sample size is high, the t-> Z	• Is there a difference in average dividend yield between stocks listed on the NYSE & NASDAQ?
F Distribution	• Testing hypotheses about the equality of two population Variances	• A new drug is evaluated for different dosage for different age group.
Chi-Square Tests	• X2 Test for the Difference Between Two Proportions	• Proportion of females who are left handed is equal to the proportion of males who are left handed
Anova	• The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups	• Measure if three or more different golf clubs yield different distance

Truth Table

Evaluating what may go wrong with hypothesis testing

Court Case		Hypothesis Testing based Decision	
		Do not Reject H0 (Innocent Verdict)	Reject H0 and select Ha (Guilty Verdict)
Actual Situation / Reality / Truth	Do not Reject H0 (Innocent)	✓ Okay	☹ Type 1 Error (we have convicted an innocent guy)
	Reject H0 and select Ha (Guilty)	☹ Type 2 Error (we have let go of a guilty guy)	✓ Okay

Which is costly to make Type 1 or Type 2

Regression Analysis

The *regression* equation attempts to explain the relationship between the Y and X variables through *linear* association.

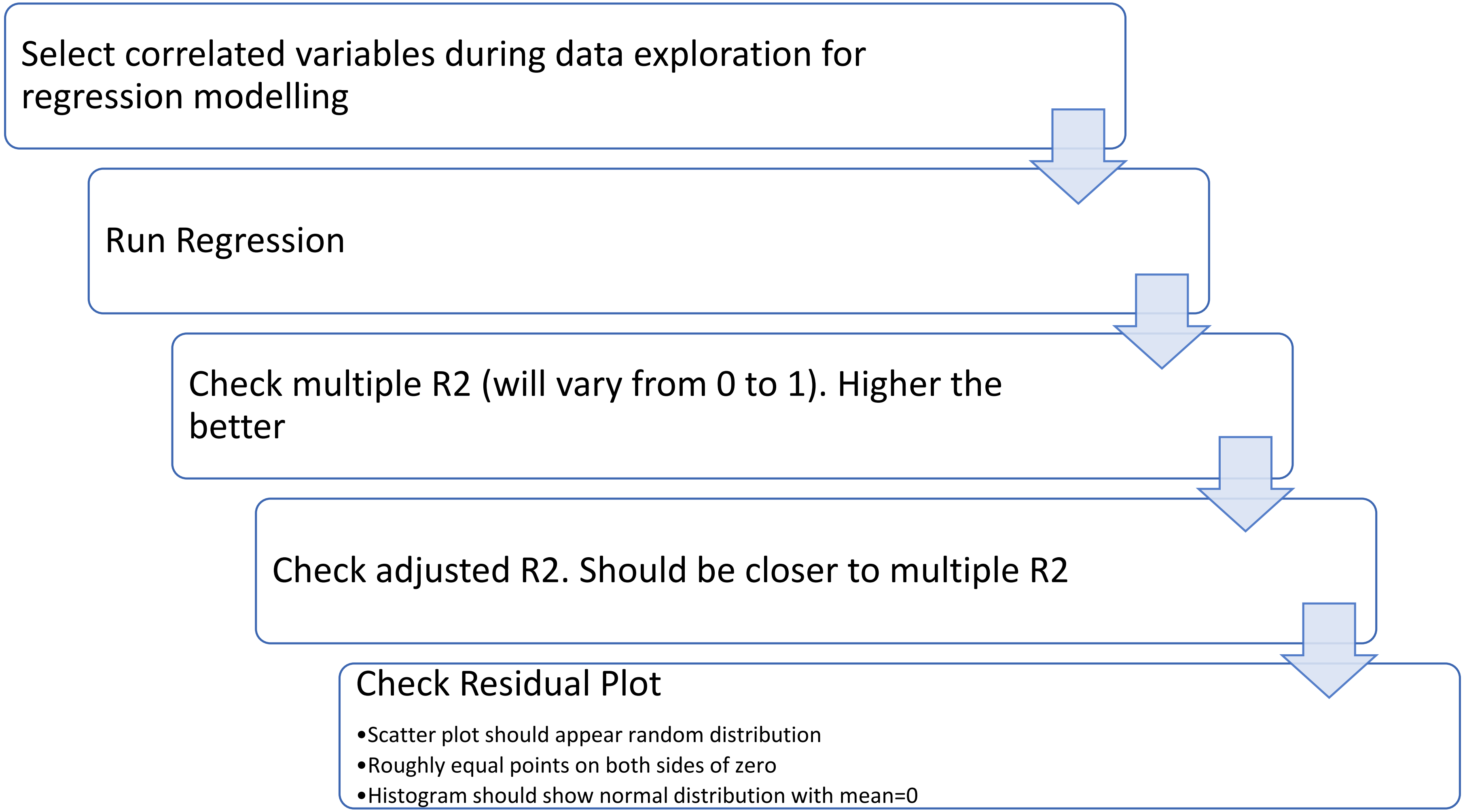
Represented as $Y = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$.

Where “b_i” represents slope or angle and “a” represents Intercept

Example

MPG of a CAR = 9.6 + 1.2*(time to reach .5KM) - 3.9*weight in tonnes

Steps in Regression Analysis



Select correlated variables during data exploration for regression modelling

Run Regression

Check multiple R2 (will vary from 0 to 1). Higher the better

Check adjusted R2. Should be closer to multiple R2

Check Residual Plot

- Scatter plot should appear random distribution
- Roughly equal points on both sides of zero
- Histogram should show normal distribution with mean=0

AWSMOSIS

Transition Redefined...

Awsmosis Learning and People Solutions Pvt. Ltd.
Address: A-10, Sapphire Apartment Condominium,
North Main Road, Koregaon Park, Pune- 411001
Phone: +91 9011067390, +91-20-26154048;
Email: info@awsmosis.in
Website: www.awsmosis.in