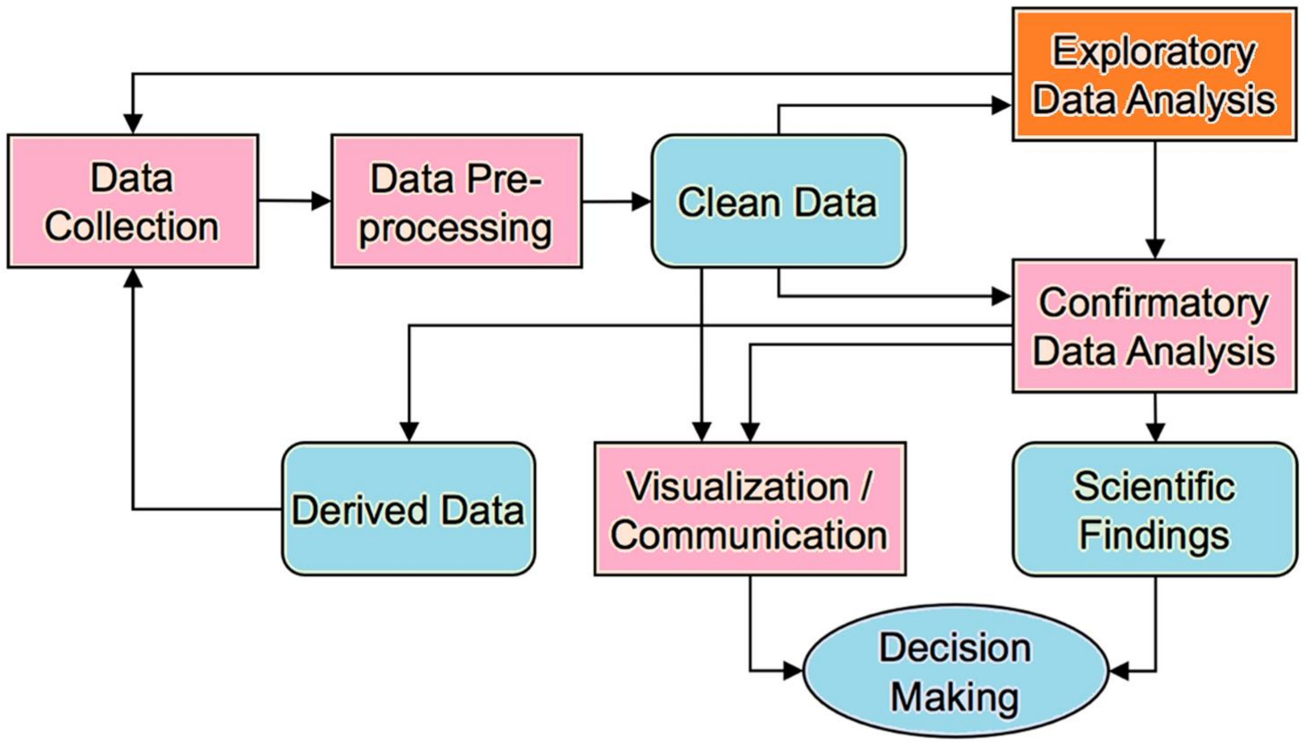
Exploratory Data Analysis

Sharad B Nalawade



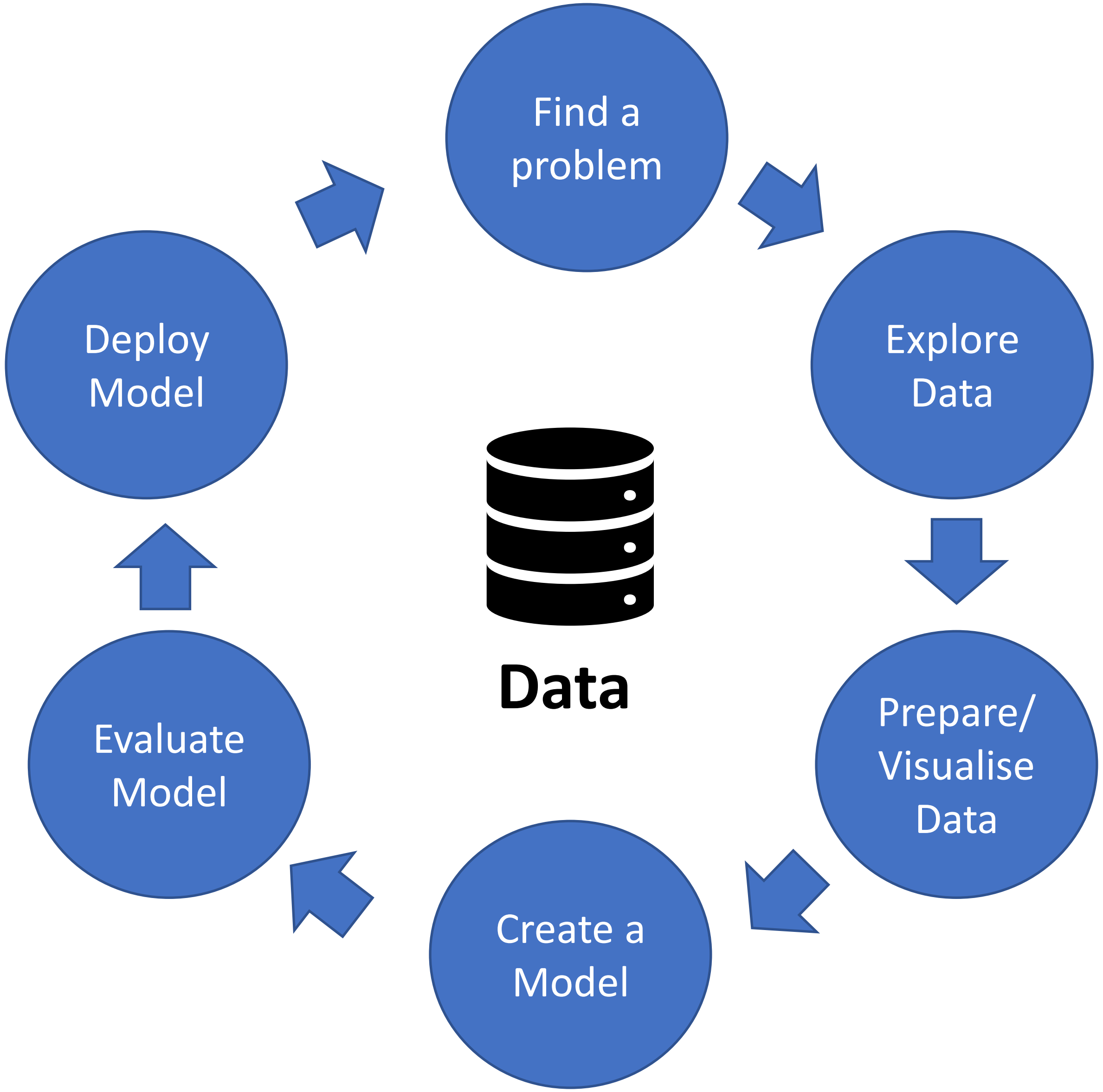
Trainer Profile: Sharad Nalawade

- 35 years in the IT industry
- Founder – ***Ultitude Consulting***, India
- President – ***Ambharii Labs,*** USA
- Stints at ***Cisco, Accenture, Symantec, Wipro, C-DoT and CMC Ltd.***
- Consultant on emerging technologies – Data Science, Machine Learning, Cloud Computing, RPA, Digital Transformation, Solution Design and Enterprise Architecture
- Author of three books on Science and Technologies
- Visiting Faculty at IIT-Jodhpur, IIM-Rohtak, IIM-Indore, IIM-Lucknow

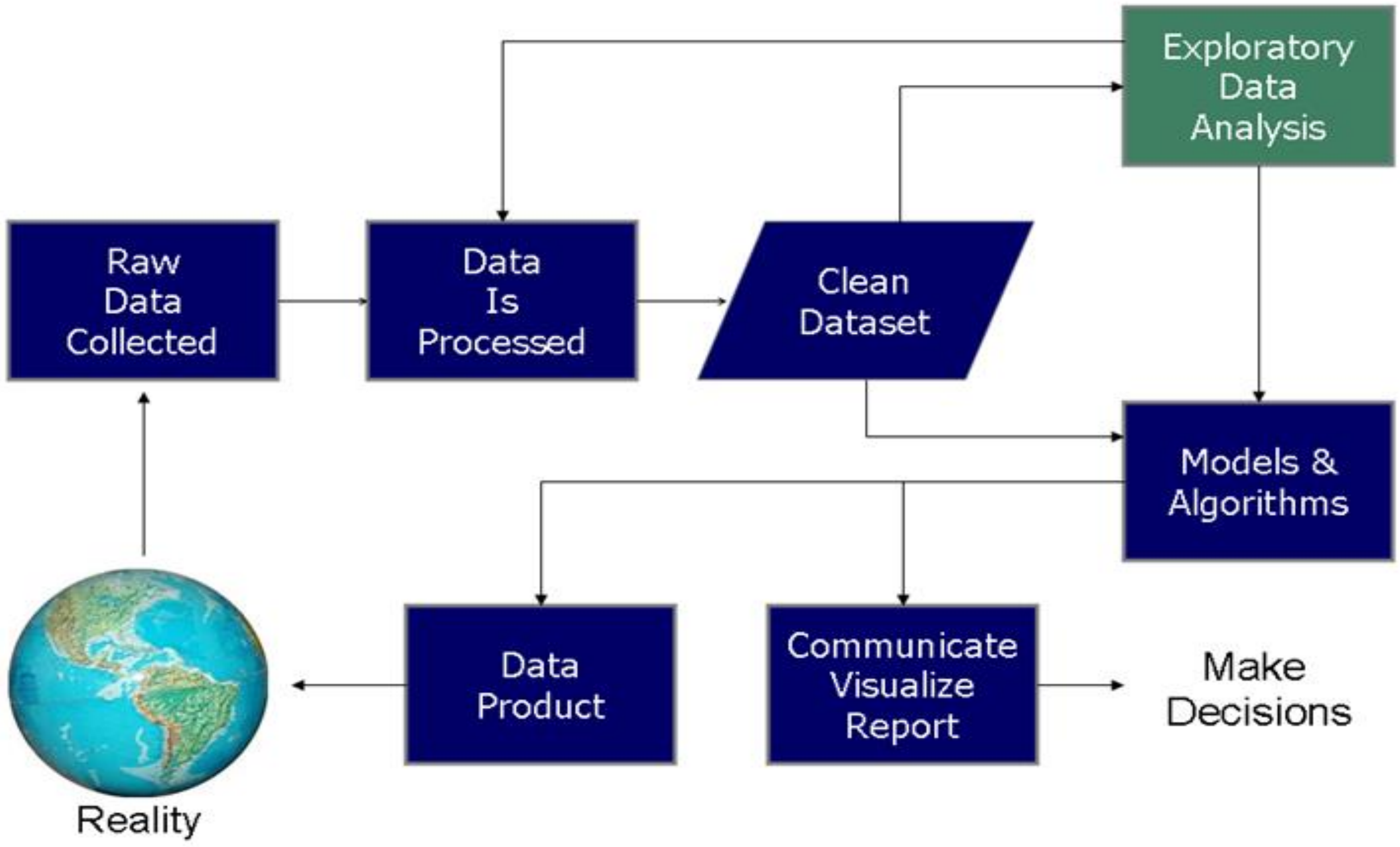
What is Exploratory Data Analysis?

In statistics, **exploratory data analysis** (EDA) is an approach of analysing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

Typical Data Science Process



Exploratory Data Analysis Process



Exploratory Data Analysis - Demo

Data set:

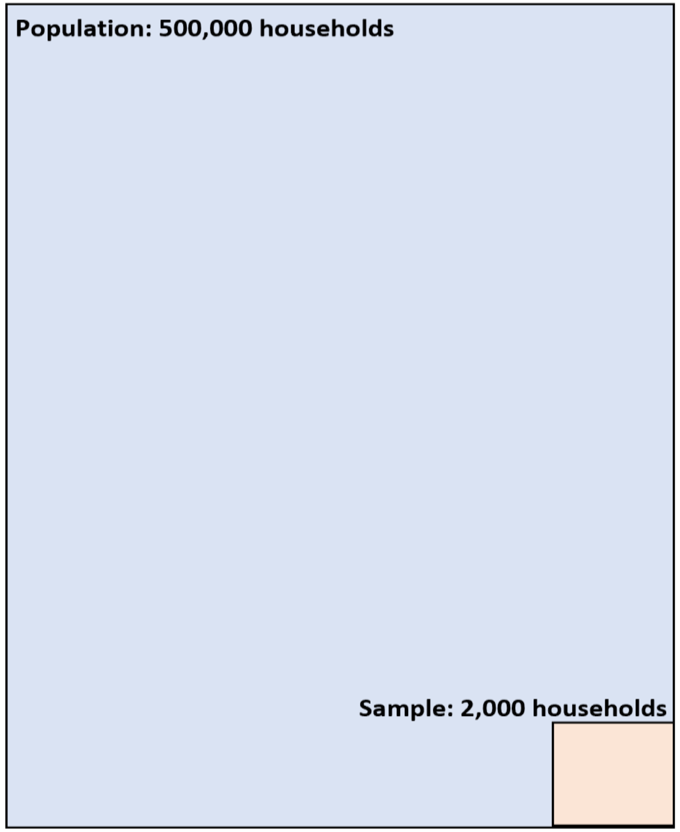
https://www.kaggle.com/josemmguerrero/analysis-and-visualization

Microsoft Excel
ma Separated Valu

Supermaket Sales Data Analytics Demo in Jupyter Notebook

Collecting Data

Population: Every possible individual element that we are interested in measuring.
Sample: A portion of the population.



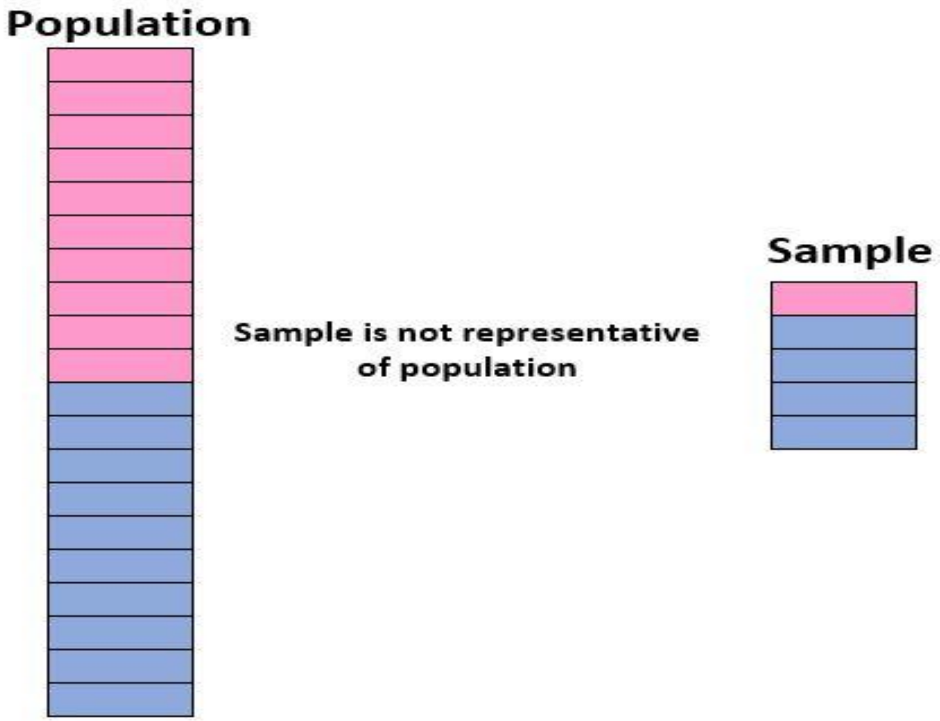
1. It is too time-consuming to collect data on an entire population.

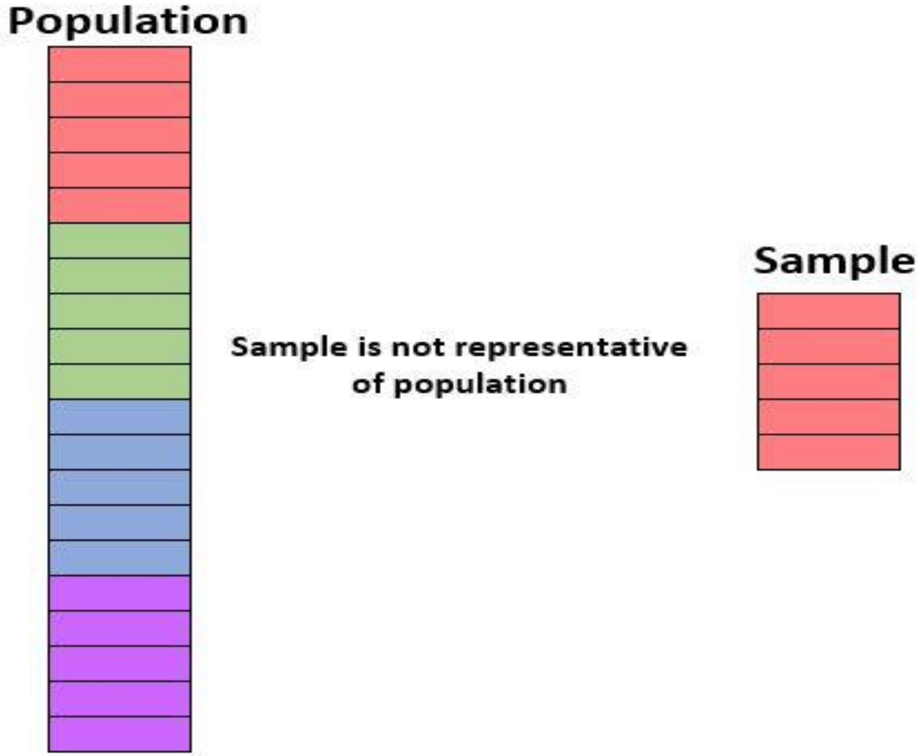
2. It is too costly to collect data on an entire population.

3. It is unfeasible to collect data on an entire population.

Collecting Data

The **Sample** should be representative of the population.



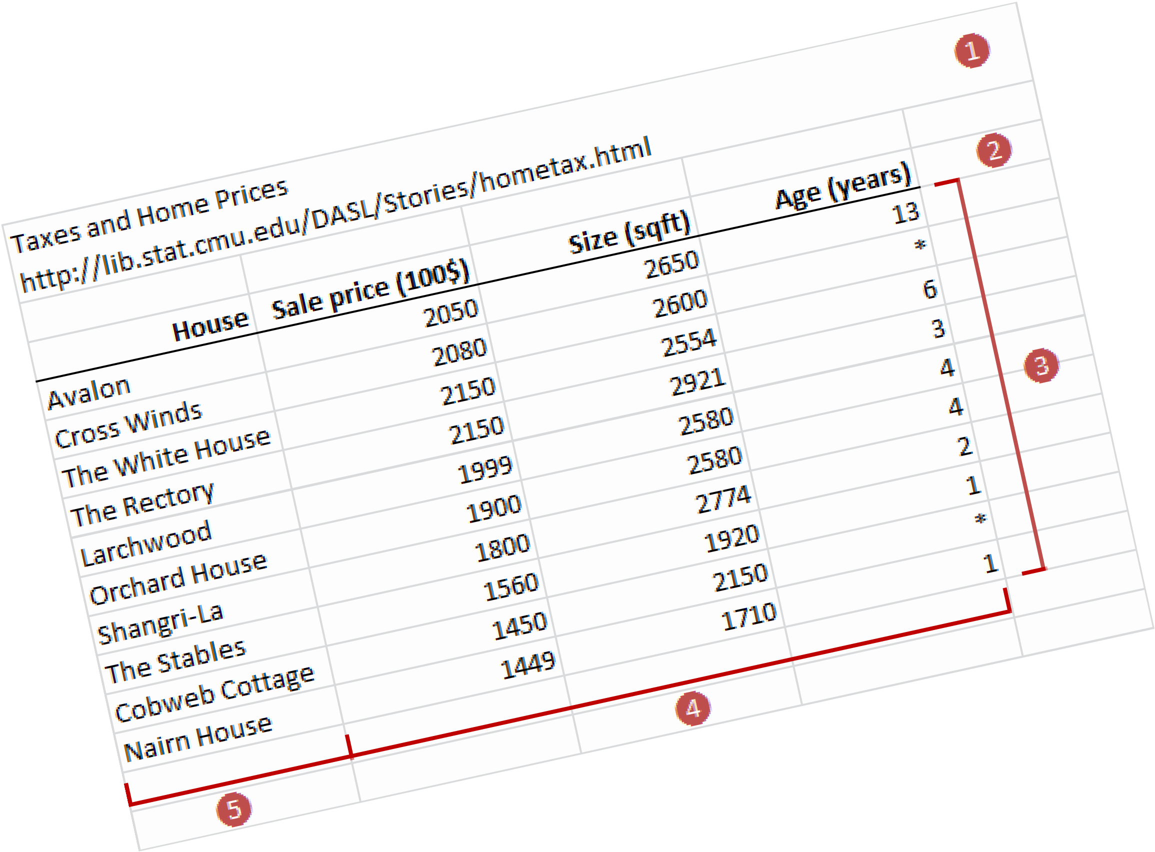


If the overall student population is composed of 50% girls and 50% boys, our sample would not be representative if it included 90% boys and only 10% girls.

When sampling sales data about four products, if we create sample data consisting of only one product sales, then it is not a true representation of the population.

What’s a Dataset?

A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.



Taxes and Home Prices
http://lib.stat.cmu.edu/DASL/Stories/hometax.html

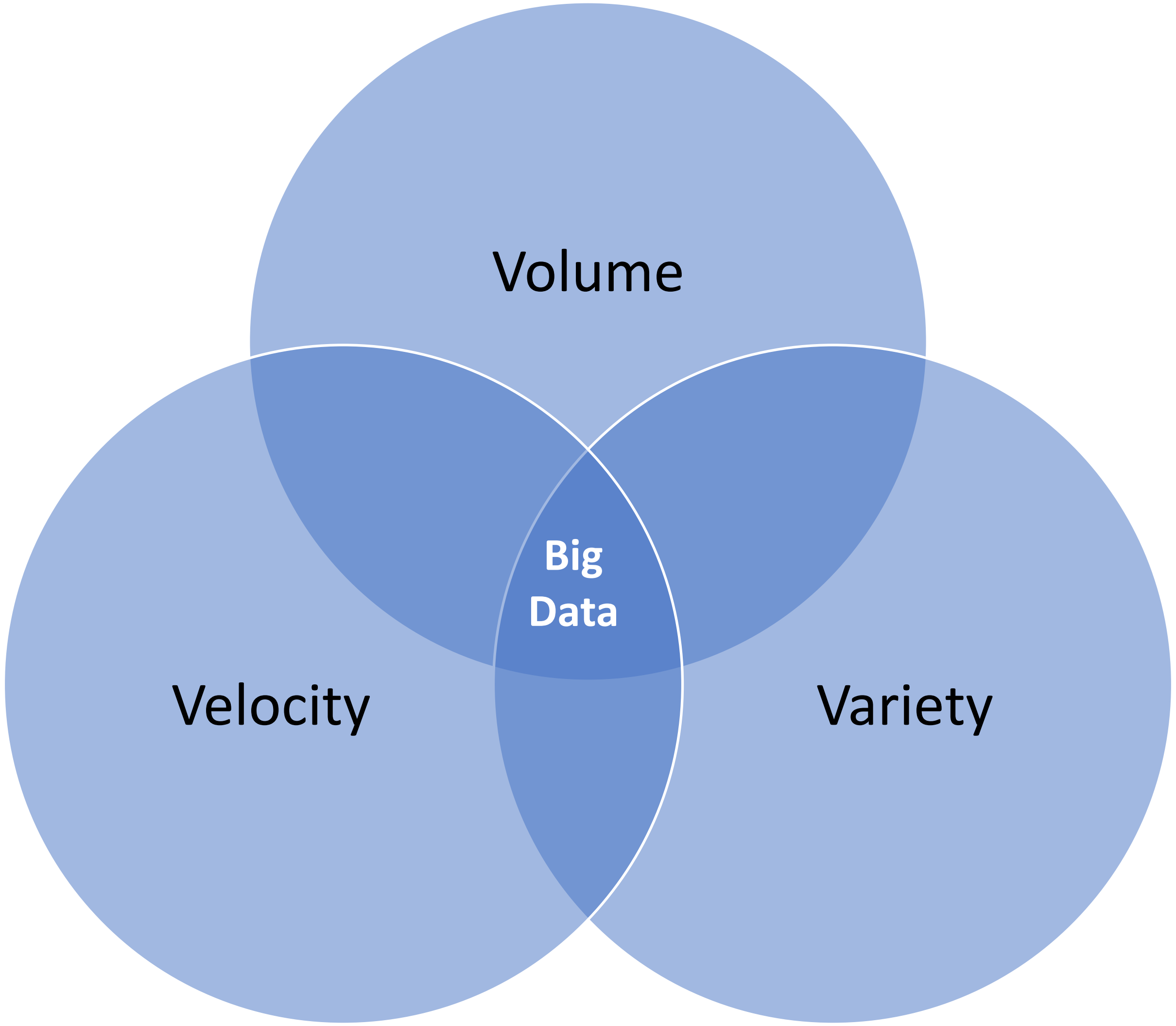
House	Sale price (100$)	Size (sqft)	Age (years)
Avalon	2050	2650	13
Cross Winds	2080	2600	*
The White House	2150	2554	6
The Rectory	2150	2921	3
Larchwood	1999	2580	4
Orchard House	1900	2580	4
Shangri-La	1800	2774	2
The Stables	1560	1920	1
Cobweb Cottage	1450	2150	*
Nairn House	1449	1710	1

Dataset Types

- Most datasets are in the table or matrix formats
- Some datasets are graph-based (nodes and edges)
- Datasets can be in the CSV or Excel or Documents (JSON) formats
- Datasets can also be sequential – *Genetic sequence*
- Datasets can be time series – *Hourly or daily stock values*
- It can be demographic – *Geographical spread of customers*

Can you think for different types of datasets from your domain?

What is Big Data?



Datasets online

https://www.kaggle.com/datasets

https://archive.ics.uci.edu/ml/datasets.php

https://data.world/datasets/data

https://data.fivethirtyeight.com/

Data Exploration and Preparation

Steps to understand, clean and prepare your data for building your predictive model:

- Variable Identification or Feature Engineering
- Univariate Analysis, Time Series Analysis
- Bi-variate Analysis
- Missing values treatment
- Outlier treatment
- Variable transformation (String to Float/Integer)
- Variable creation
- Dataset Homogeneity
- Normalization and Standardization

Data Exploration and Preparation

- Data plays a big role in Data Analytics
- Quality of the input data determines the accuracy of results
- Data exploration, data cleaning and data preparation takes 70% of the effort

	A	B	C	D
1		Column 1	Column 2	Column 3
2	Row 1	2.2	2.3	1
3	Row 2	2.3	2.6	0
4	Row 3	2.1	2	1

Dataset Types

Univariate:

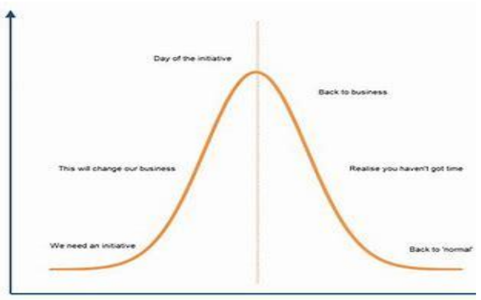
- ➢ Single variable, no relationship
- ➢ Used for time series analysis

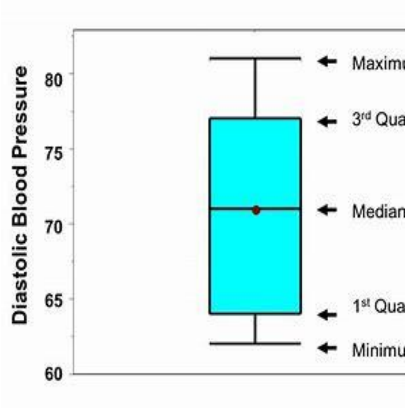
Ex: List of heights or list of stocks

Heights (in cm)	164	167.3	170	174.2	178	180

Describing univariate data with charts:

- •Frequency Distribution Tables
- •Bar Charts
- •Histograms
- •Pie Charts
- •Time series





Dataset Types

Multivariate:

- One or more variables determine output
- Ex: Height versus Weight
- Ex: Tech support dataset

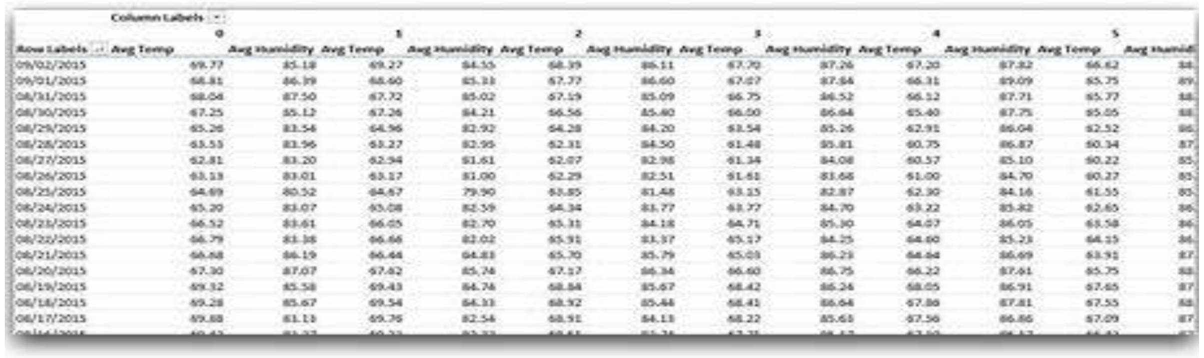
PROBLEM_TYPE	no_of_cases	Avg_pending_calls	Avg_resol_time	recurrence_freq	Replace_percent	In_warranty_percent	Post_warranty_percent
Temperature control not working	170	1.3	32	0.04	0	75	25
power chord does not tightly fit	12	2	150	0.01	0.5	5	95
Fan swing not working	5	1	35	0.02	0.2	90	10
Main switch does not on	3	2	8	0.01	0.7	5	95
Forgot mobile app password	45	2.3	54	0.15	0	99	1
AC stops abruptly	47	3.1	132	0.3	0.03	100	0
App hangs while entring commands	12	4	154	0.02	0.05	100	0
App takes a long time to initialize	165	1.2	32	0.03	0	95	5
Unable to connect the App to the device	321	1	5	0.21	0	60	40
Reinstalling the app hangs the phone	22	3.3	140	0.14	0.01	95	5
Mobile screen blanks on use of phone	23	4.3	143	0.21	0.06	100	0
Cluttered with popup messages	230	1.3	23	0.02	0	10	90
AC controller sofware missing	193	1.2	33	0.03	0	100	0

Can you develop a multivariate dataset from your domain?

Dataset Characteristics

Dimensionality:

- Measure of number of attributes
- Too many attributes is a *Curse of Dimensionality*



	Column Labels												
	0		1		2		3		4		5		
Row Labels	Avg Temp	Avg Humidity	Avg Temp	Avg Humidity	Avg Temp	Avg Humidity	Avg Temp	Avg Humidity	Avg Temp	Avg Humidity	Avg Temp	Avg Humidi	
09/02/2015	69.77	85.18	69.27	84.55	68.39	86.11	67.70	87.26	67.20	87.82	66.62	88	
09/01/2015	68.81	86.39	68.60	85.33	67.77	86.60	67.07	87.84	66.31	89.09	65.75	89	
08/31/2015	68.04	87.50	67.72	85.02	67.19	85.09	66.75	86.52	66.12	87.71	65.77	88	
08/30/2015	67.25	85.12	67.26	84.21	66.56	85.40	66.50	86.64	65.40	87.75	65.05	88	
08/29/2015	65.26	83.54	64.96	82.92	64.28	84.20	63.54	85.26	62.91	86.04	62.52	86	
08/28/2015	63.53	83.96	63.27	82.95	62.31	84.50	61.48	85.81	60.75	86.87	60.34	87	
08/27/2015	62.81	83.20	62.94	81.61	62.07	82.98	61.34	84.08	60.57	85.10	60.22	85	
08/26/2015	63.13	83.01	63.17	81.00	62.29	82.51	61.61	83.68	61.00	84.70	60.27	85	
08/25/2015	64.69	80.52	64.67	79.90	63.85	81.48	63.15	82.87	62.30	84.16	61.55	85	
08/24/2015	65.20	83.07	65.08	82.39	64.34	83.77	63.77	84.70	63.22	85.82	62.65	86	
08/23/2015	66.52	83.61	66.05	82.70	65.31	84.18	64.71	85.30	64.07	86.05	63.58	86	
08/22/2015	66.79	83.38	66.66	82.02	65.91	83.37	65.17	84.25	64.60	85.23	64.15	86	
08/21/2015	66.68	86.19	66.44	84.83	65.70	85.79	65.03	86.23	64.64	86.69	63.91	87	
08/20/2015	67.30	87.07	67.62	85.74	67.17	86.34	66.60	86.75	66.22	87.61	65.75	88	
08/19/2015	69.32	85.58	69.43	84.74	68.84	85.67	68.42	86.24	68.05	86.91	67.65	87	
08/18/2015	69.28	85.67	69.54	84.33	68.92	85.44	68.41	86.64	67.86	87.81	67.55	88	
08/17/2015	69.88	83.13	69.76	82.54	68.91	84.13	68.22	85.63	67.56	86.86	67.09	87	

The **curse of dimensionality** refers to the phenomena that occur when classifying, organizing, and analysing high dimensional data that does not occur in low dimensional spaces, specifically the issue of data sparsity and “closeness” of data.

Dataset Characteristics

- **Resolution:**
  - If the resolution is too fine, a pattern may not be visible or may be buried in noise; if the resolution is too coarse, the pattern may disappear.
  - *Ex: Atmospheric pressure data on hourly scale is different than on scale of months*

Steps in Data Exploration and Preparation

Variable Identification

- First, identify **Predictor** (Input) and **Target** (output) variables.

- Next, identify the type and category of the variables.

Example:

Predictors **Target**

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

Steps in Data Exploration and Preparation

Uni-variate Analysis

- Uni-variate analysis is used to highlight the missing or outlier values in the given dataset

Continuous Variables Analysis

- In case of continuous variables, we need to understand the central tendency and spread of the variable.
- We use measures like ***mean, median, min, max, mode, variance, standard deviation***, etc to explore data.

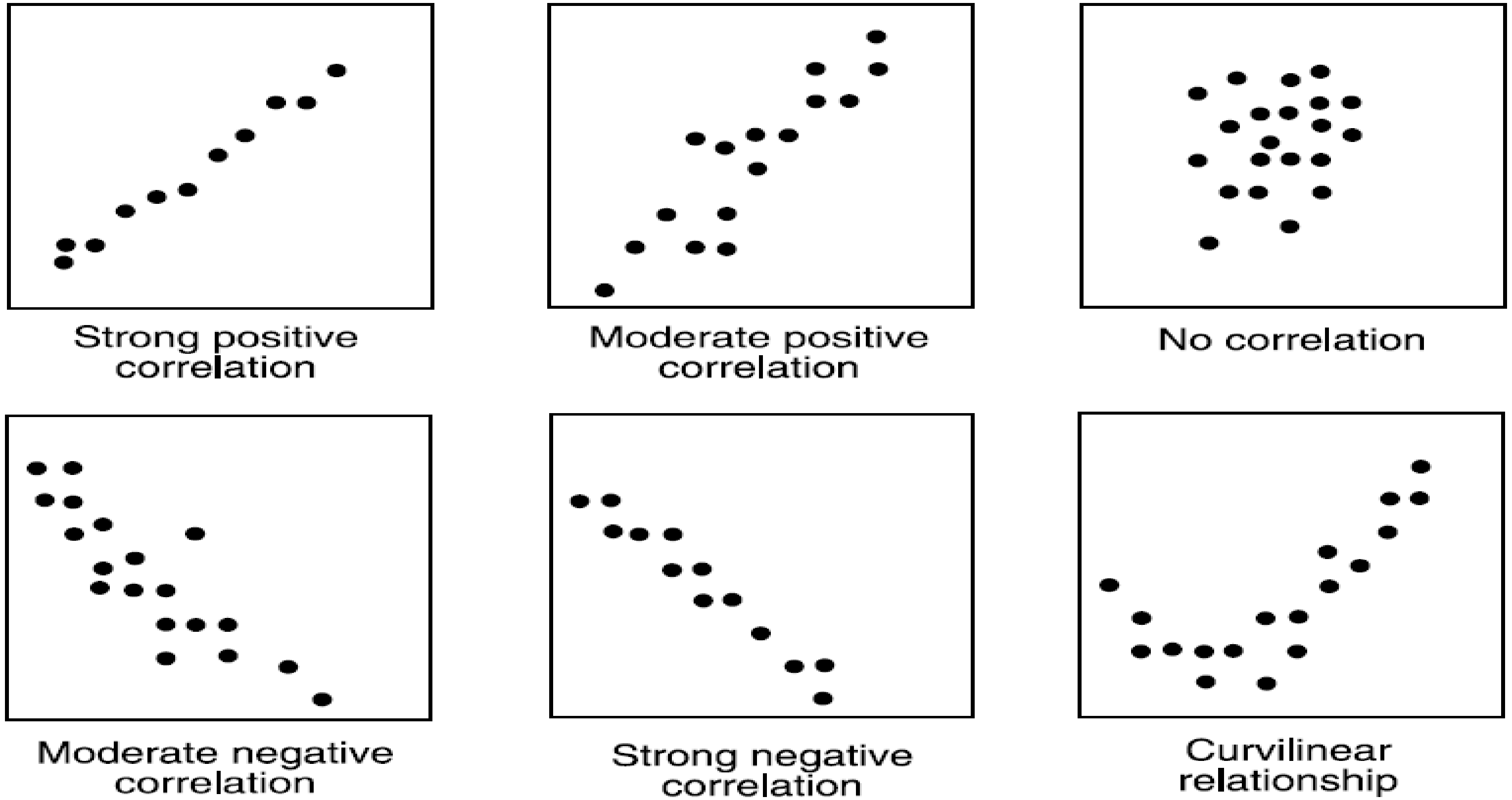
Categorical Variables

- In case of categorical variables, we need to understand the ***frequency*** and ***% of the count***.

Steps in Data Exploration and Preparation

Bi-variate Analysis

Bi-variate analysis is used to find relationship between two variables. We look for ***association*** or ***disassociation*** between variables. We use scatter-plots.



Steps in Data Exploration and Preparation

Missing Value Treatment

Missing value will lead to bias in our analysis of the data

Reasons for missing values

- ✓Data collection errors
- ✓Data extraction errors
- ✓Random entries
- ✓Missing observations

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Steps in Data Exploration and Preparation

Treating the missing values

In ***list wise deletion***, we delete observations where any of the variable is missing.

Gender	Manpower	Sales
M	25	343
~~F~~		~~280~~
M	33	332
~~M~~		~~272~~
~~F~~	~~25~~	
M	29	326
	~~26~~	~~259~~
M	32	297

Steps in Data Exploration and Preparation

In ***pair wise deletion***, we perform analysis with all cases in which the variables of interest are present.

If you delete pairwise then you'll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult.

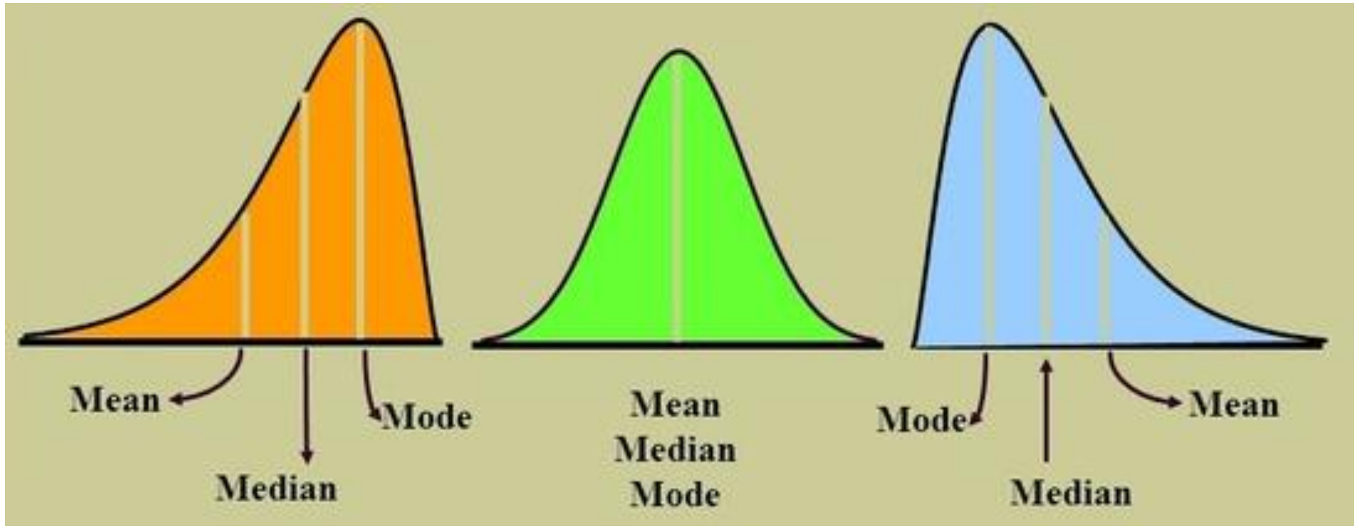
SN	Age	Income	Political Affiliation
1	18	10000	A
2	NA	25000	B
3	50	45000	C
4	22	NA	C
5	45	45000	A
6	70	20000	NA
7	NA	40000	B

Without deleting any rows, you can still do analysis of age versus political affiliation and income versus political affiliation.

Applying Statistical Analysis to the Data

Mean, Median & Mode

- **Mean** is the average of a set of numbers, indicating the central tendency
- **Median** is the 50^{th} percentile, the middle number that divides the set into two equal parts
- **Mode** is the most frequent number in a set



Unlike mean, the median is not affected by an outlier in the set.

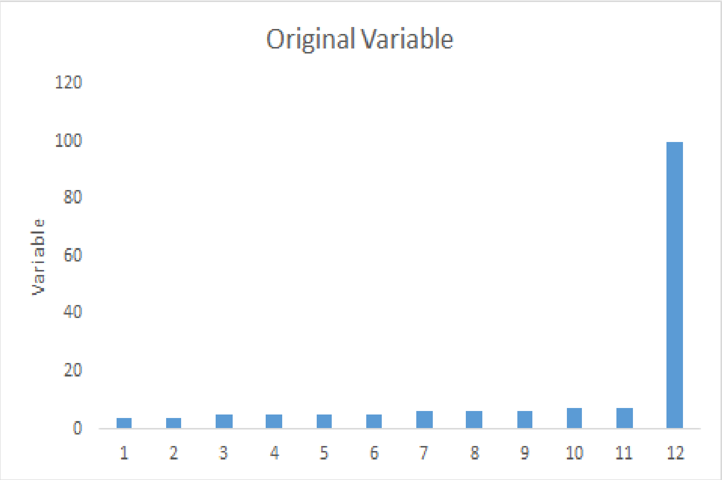
Steps in Data Exploration and Preparation

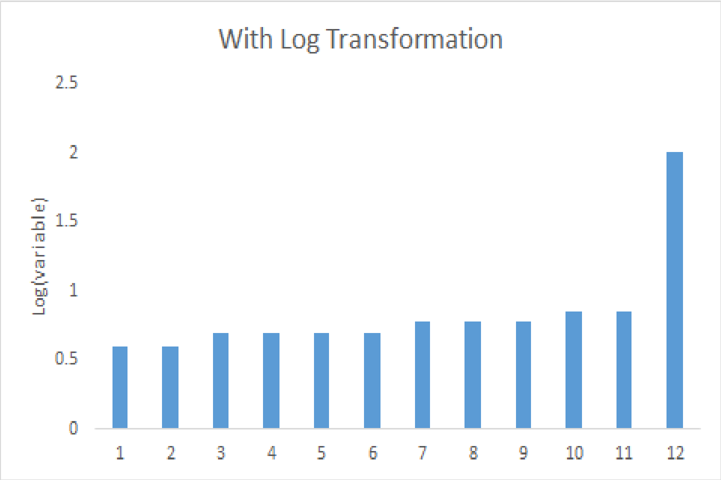
Outlier Detection and Treatment

Example:

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Treatment: ***Variable Transformation*** **or** ***Binning***





Basic Math/Statistics Functions

Mean

In probability and statistics, population **mean** or expected value are used synonymously to refer to one measure of the central tendency.

$$\overline{X} = \frac{\Sigma X}{n}$$

For example, the mean of 1, 2, 3, 4, and 5 is (15 ÷ 5) = 3.

✓It is a measure of the mid-point (around which all other values cluster) of a set of values

✓Prone to distortion by the presence of extreme values and may require use of a measure of distortion (such as mean deviation or standard deviation).

Basic Math/Statistics Functions

Median The **median** is the value separating the higher half of a dataset

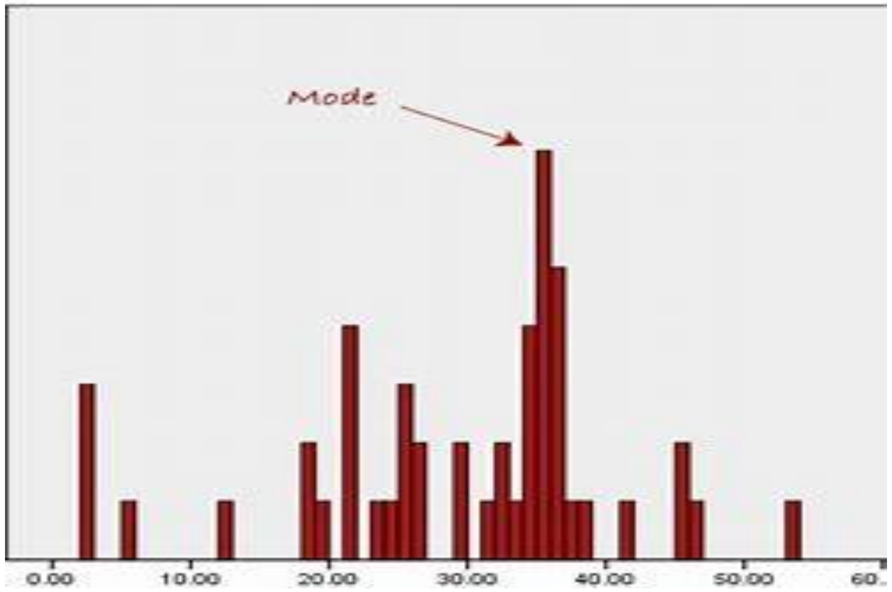
Example: The median value of the dataset 1, 2, 2, 3, 4, 7, 9 is **3**.

- To calculate Median, you have to first sort the dataset.
- If the number of values is odd, then the middle number is the median
- If the no. of values is even, the you take the mean of the middle two values

Basic Math/Statistics Functions

Mode The **mode** of a set of data values is the value that appears most often

Example: The mode in the dataset 1, 2, 2, 3, 4, 7, 9 is **2**.

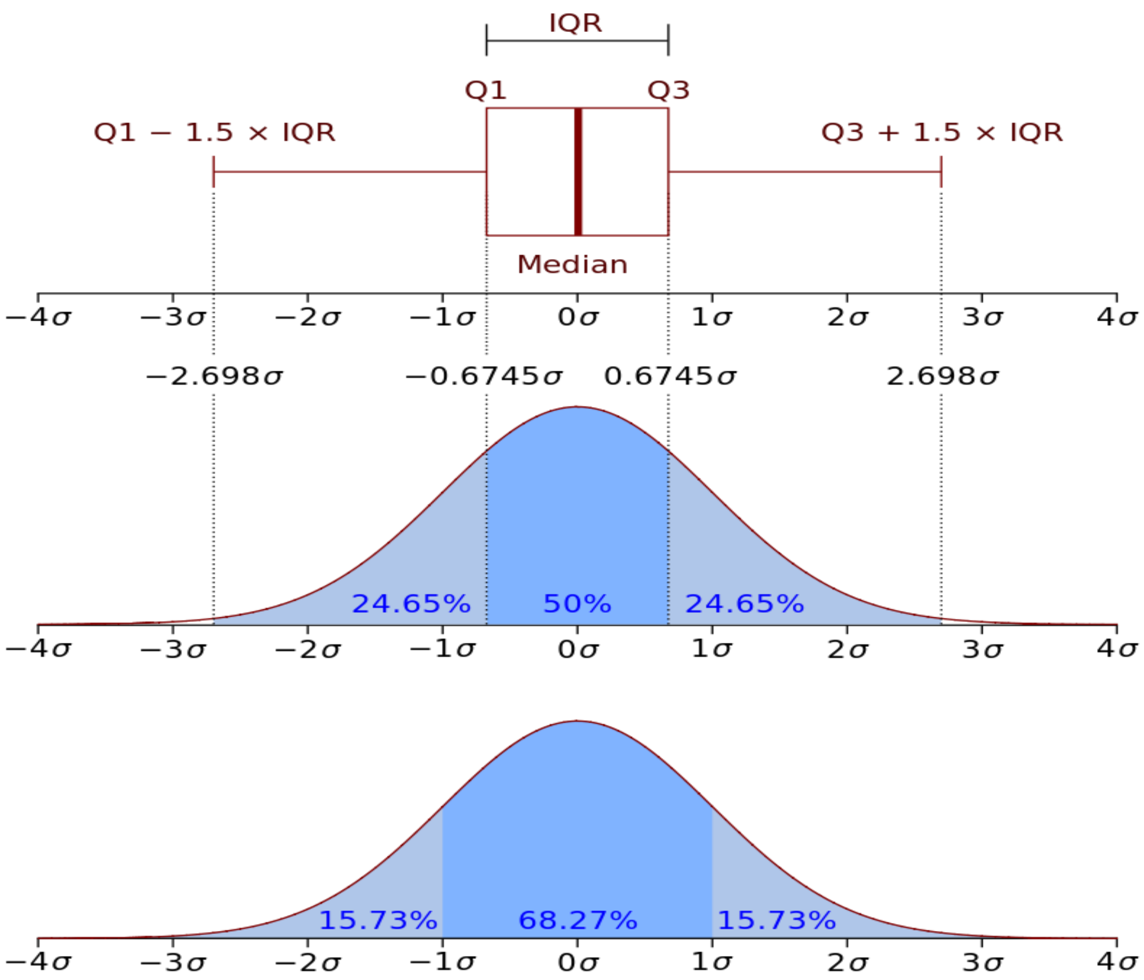


Quartile

A **quartile** is a number that divides the number of data points into four more-or-less equal parts, or quarters.

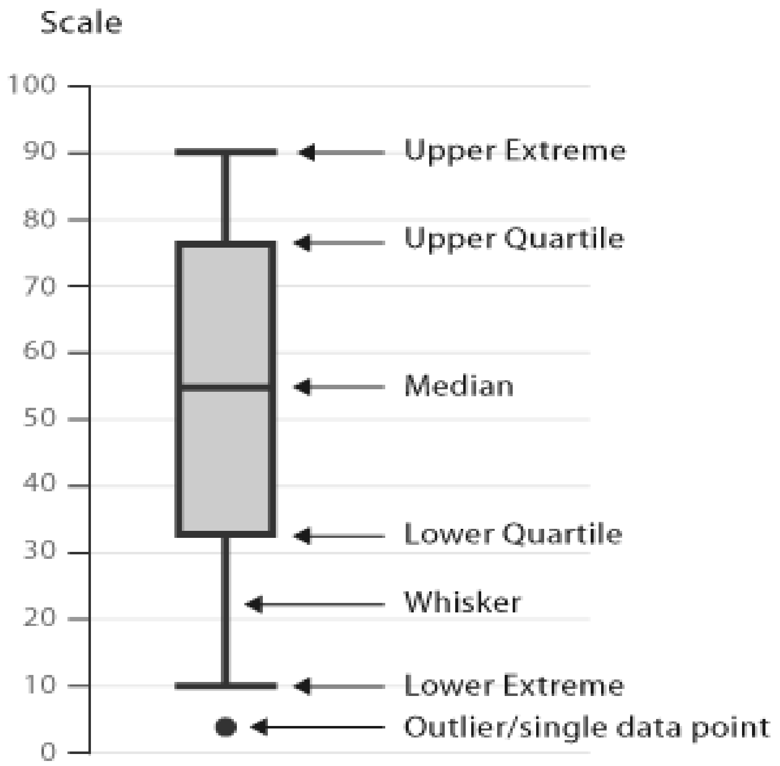
Symbol	Names	Definition
Q_1	first quartile lower quartile 25th percentile	splits off the lowest 25% of data from the highest 75%
Q_2	second quartile median 50th percentile	cuts data set in half
Q_3	third quartile upper quartile 75th percentile	splits off the highest 25% of data from the lowest 75%

Quartile



IQR = Inter-quartile Range is the middle 50% of the data

Box-Whisker Plot



Quartile Example

Basic Math/Statistics Functions

The variance measures how far each number in the set is from the mean.

Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

X: individual data point

μ: mean of data points

N: total # of data points

Basic Math/Statistics Functions

- **In stock market, variance is a measure of stock's volatility.**

- **<u>Example:</u> Let's say, the returns for a stock are 10% in year 1, 20% in year 2 and -15% in year 3.**

- **The average of these three returns is 5%.**

- **The differences between each return and the average are 5%, 15%, and -20% for each consecutive year.**

- **Squaring these deviations yields 25%, 225% and 400%, respectively; summing these squared deviations gives 650%.**

- **Dividing the sum of 650% by the number of returns in the data set (3 in this case) yields the variance of 216.67%.**

- **Taking the square root of the variance yields the standard deviation of 14.72% for the returns.**

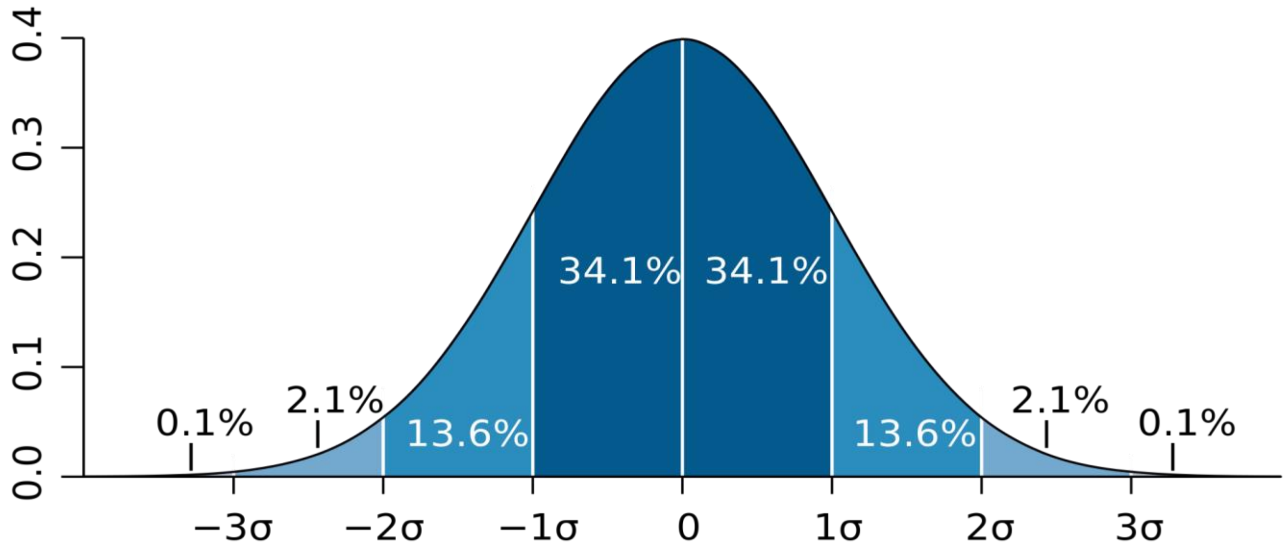
Basic Math/Statistics Functions

Standard Deviation

Standard deviation is used to measure spread or dispersion around the mean of a data set.

A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Bell curve - 68–95–99.7 rule



Each band has a width of 1 standard deviation

Basic Math/Statistics Functions

Standard Deviation

Standard deviation is sensitive to outliers. A single outlier can raise the standard deviation and in turn, distort the picture of spread.

Standard deviation is never negative.

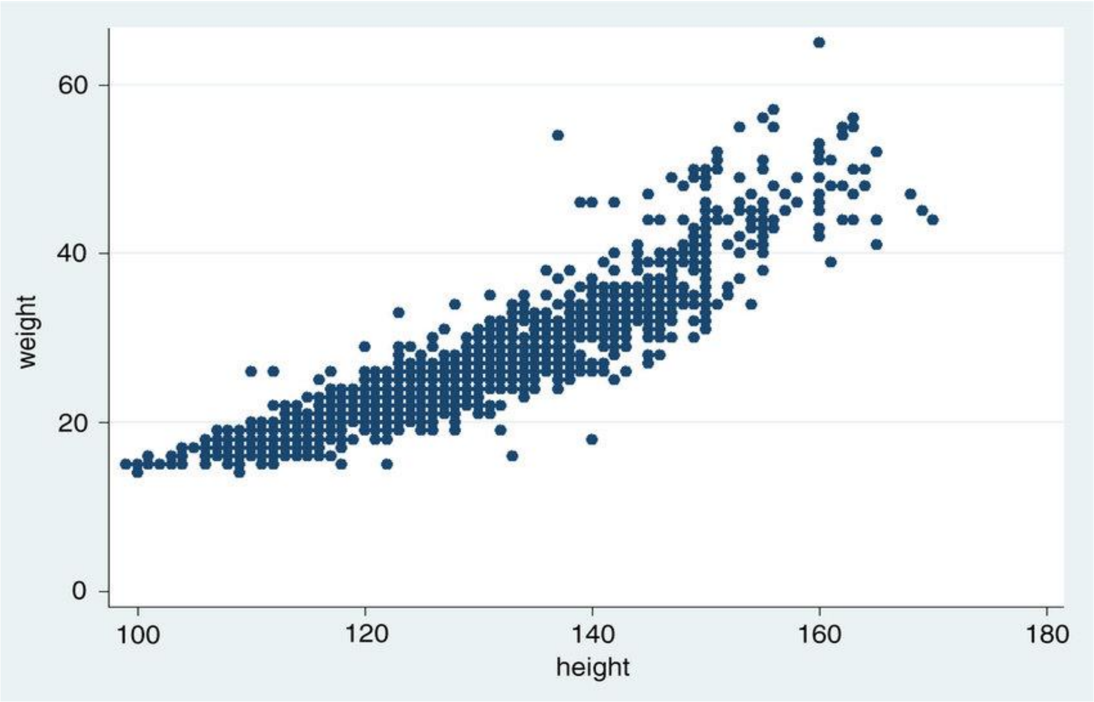
If all values of a data set are the same, the standard deviation is zero (because each value is equal to the mean).

Examples of datasets where you can apply Standard Deviation: *Population Age, Population Heights, Employee Salary, Students score, Shoe size, Tossing a coin, etc.*

Basic Math/Statistics Functions

Covariance

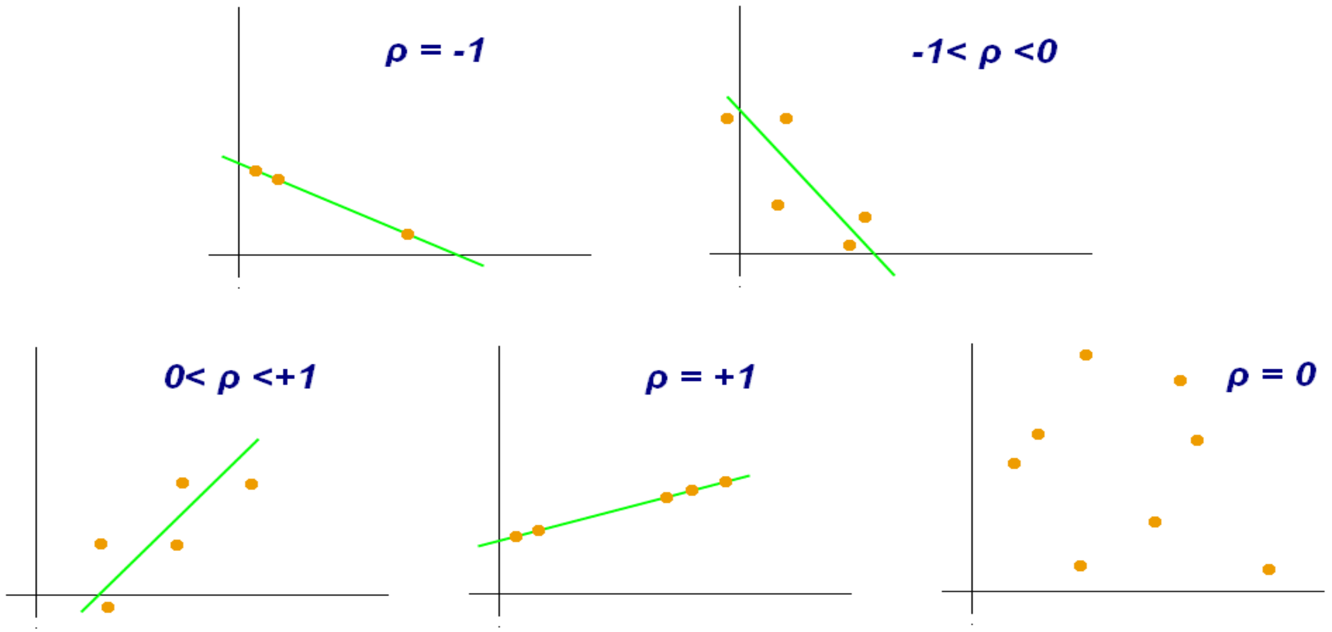
A covariance refers to the measure of how two random variables will change together and is used to calculate the correlation between variables.



Basic Math/Statistics Functions

Covariance

It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



We can use PEARSON function within MS Excel to calculate co-variance.

Basic Math/Statistics Functions

RMS The ***Root Mean Square*** is a technique to measure the error in a model.

Example: Find the RMS of -2, 5, -8, 9, -4

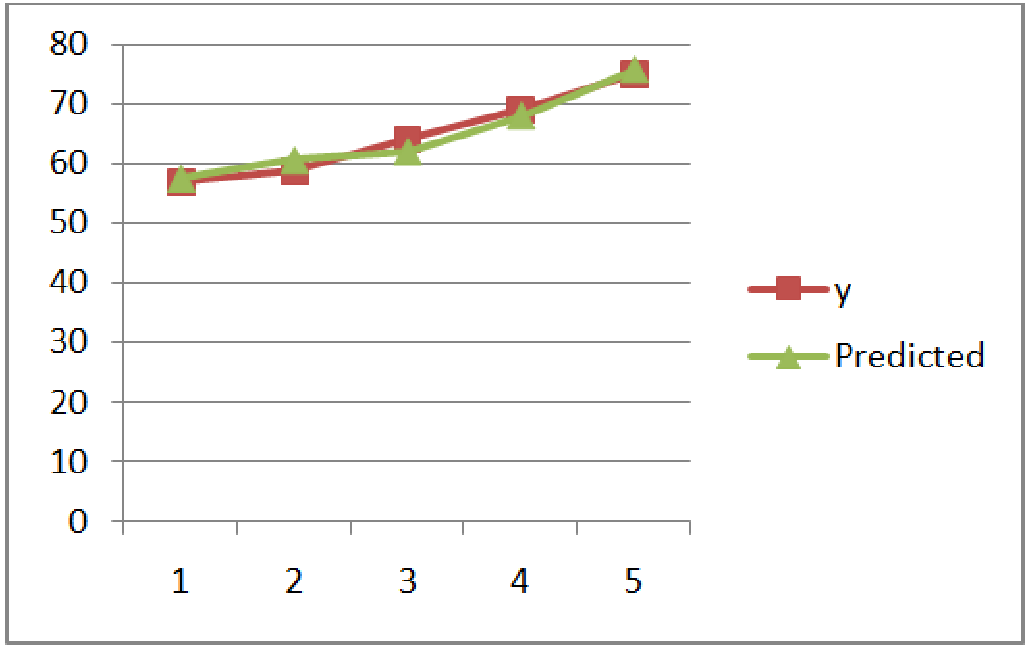
TO CALCULATE RMS:

1. SQUARE all the values
2. Take the average of the squares
3. Take the square root of the average

For example, the RMS of -2, 5, -8, 9, -4 is 6.16

RMS Example

X (Height)	Y (Weight) Actual	Predicted	Difference
165	57	**57.2**	**0.2**
167	59	**58**	**-1**
168	64	**64.6**	**0.6**
172	69	**69.8**	**0.8**
177	75	**74.4**	**-0.6**



RMS Calculation:

Differences	Squares	Average	Square Root	RMS
0.2	0.04	0.48	0.69282032	**0.69282**
-1	1			
0.6	0.36			
0.8	0.64			
-0.6	0.36			

Pre-processing your Data

Different data science projects need different structure of the dataset. It's a good idea to transform data to fit your needs. *Some key pre-processing techniques:*

Normalization – Rescaling the attributes to a common scale between 0 and 1. Useful for Regression, Classification, k-NN and Neural Networks

Standardization – Useful for Gaussian datasets, where mean will be set to 0 and Standard Deviation to 1.
Useful for Linear & Logistic Regressions using normally distributed data.

Pre-processing your Data

Let’s take examples of Normalization and Standardization

Name	Salary	Experience	Position Level
Hari	100000	10	2
Giri	78000	7	4
Siri	32000	5	8
Lata	55000	6	7
Geeta	92000	8	3
Sita	120000	15	1
Rafi	65750	7	5

Normalization and Standardization

Normalization refers to **rescaling** an input variable to the range between 0 and 1. Normalization requires that you know the minimum and maximum values for each attribute.

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

If you don't normalize the dataset, there is high chance of bias in predictive analytics.

Normalization Assignment in Excel

Normalization and Standardization

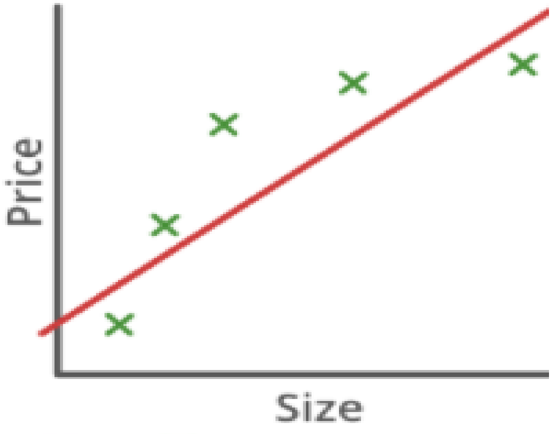
Standardisation: Standardization is a rescaling technique that refers to **centering** the distribution of the data on the value 0 and the standard deviation to the value 1.

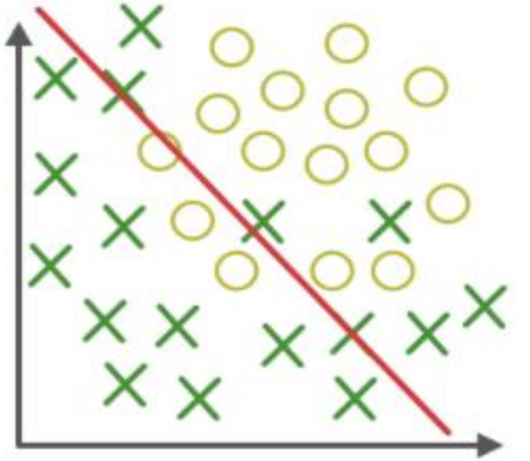
Together, the **mean** and **the standard deviation** can be used to summarize a normal distribution, also called the Gaussian distribution or Bell curve.

Bias, Variance, Underfitting & Overfitting

Bias-Variance-Underfitting-Overfitting

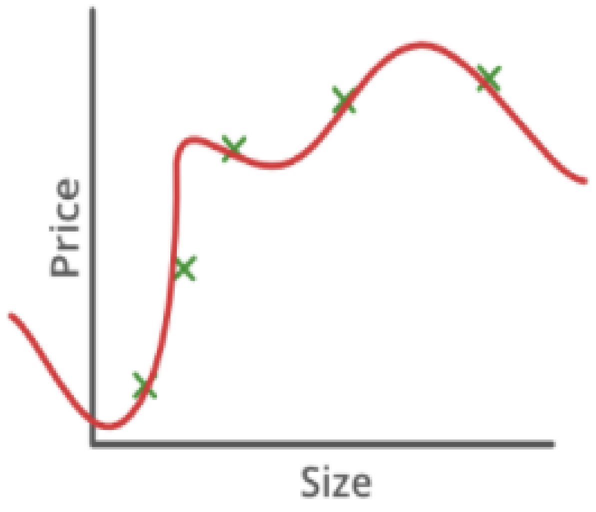
- **Data bias** in machine learning is a type of error in which certain elements of a dataset are more heavily weighted and/or represented than others.
- A biased dataset does not accurately represent a model's use case, resulting in skewed outcomes, low accuracy levels, and analytical errors.
- High bias can cause algorithm to miss crucial relationships between features and target outputs.
- Due to high bias, the algorithm **underfits** the data.
- Reasons for high bias and underfitting are many:
  - Wrong feature selection or insufficient features
  - Insufficient data
  - **Sample bias** – *Ex: Face recognition algo trained primarily on white men*
  - **Exclusion bias** – *Ex: Delete location details for customers who matter most*
  - **Recall bias** – Make feature labelling mistakes in the dataset
  - **Association bias** – Doctors are always men and nurses are always women

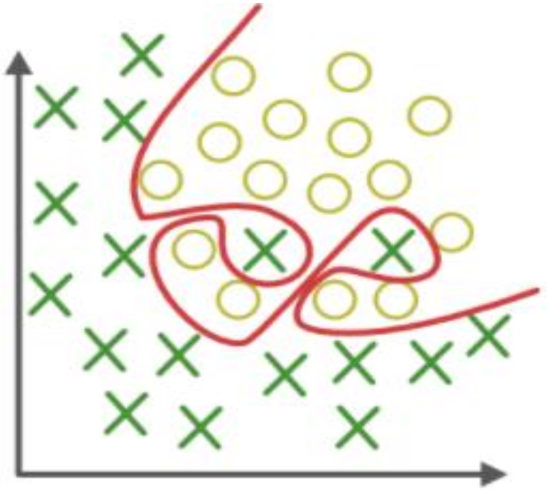




Bias-Variance-Underfitting-Overfitting

- **Variance** is an error from sensitivity to small fluctuations in the training dataset.
- Variance occurs when a model is trained too many times on the same dataset.
- High variance will occur from modelling even the noise in the dataset.
- Variance leads to **overfitting**.
- Reasons for overfitting are:
  - Training algorithm with too much data
  - Learning models get too much freedom to learn from data leading to unrealistic models.
  - The model is too complex





Bias – Variance Tradeoff

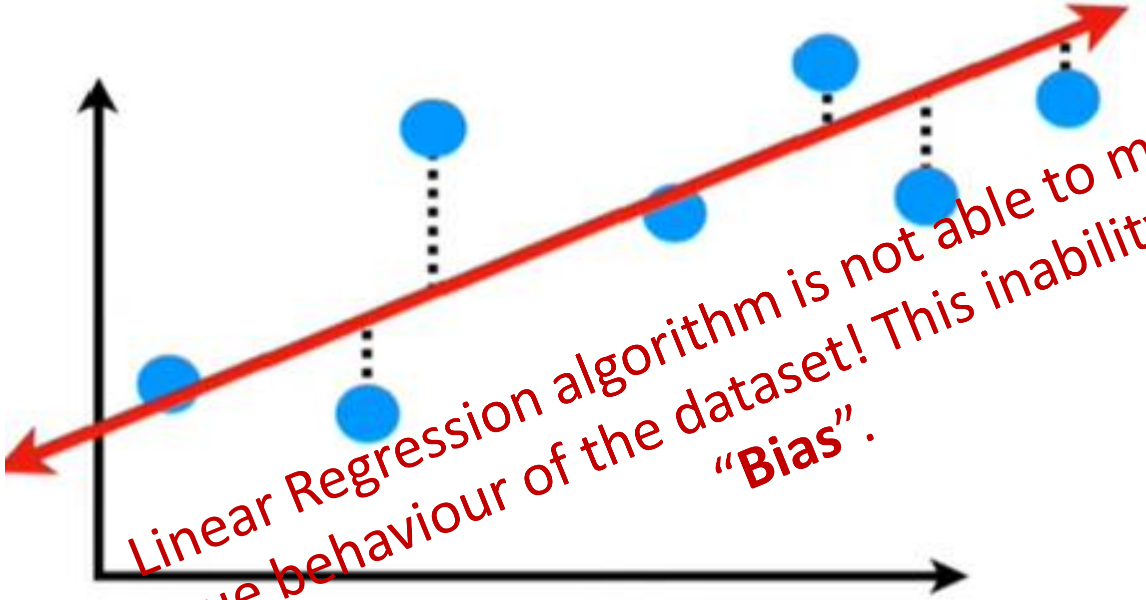
Lets start with an example of a dataset – *Height* v/s *Weight*

- Let's divide the dataset into – training and test data
- Let's say, **blue is train data** and **green is test data**
- Let's use a **Linear Regression** model to train & test with Train data only

Height (cm)	Weight (kg)
174	96
189	87
149	61
189	87
149	61
189	104
147	92
154	111
174	90
195	81
155	51
191	79
155	51
191	79
140	129
144	102
172	139
157	110
185	118

Train data

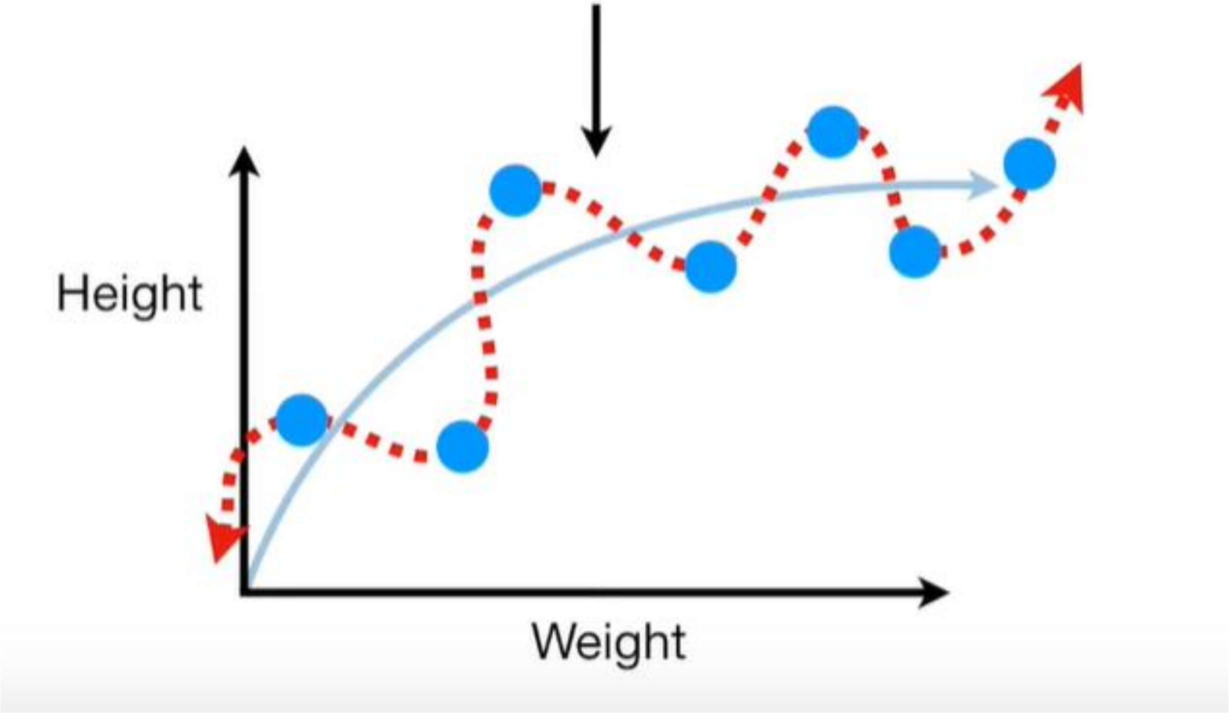
Test data



Linear Regression algorithm is not able to model the true behaviour of the dataset! This inability is called "**Bias**".

Bias and Variance

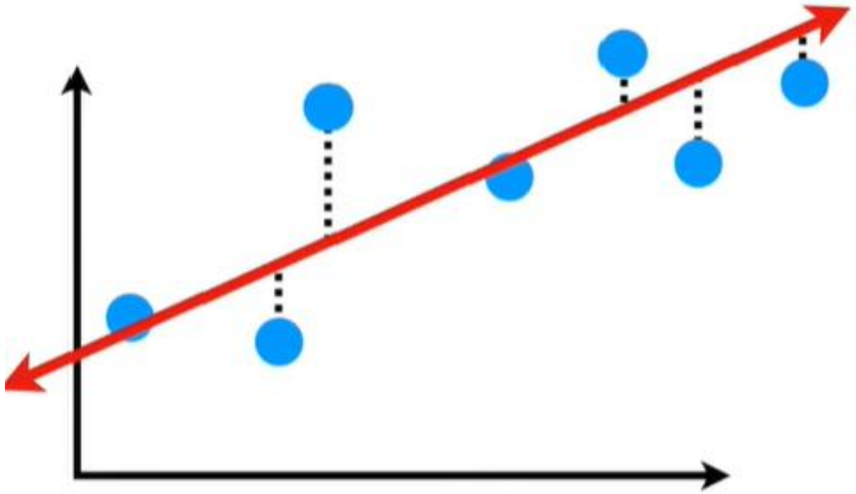
Let’s say, we have a invented a fantastic ML algorithm that accurately models the behaviour of a given dataset



Bias and Variance

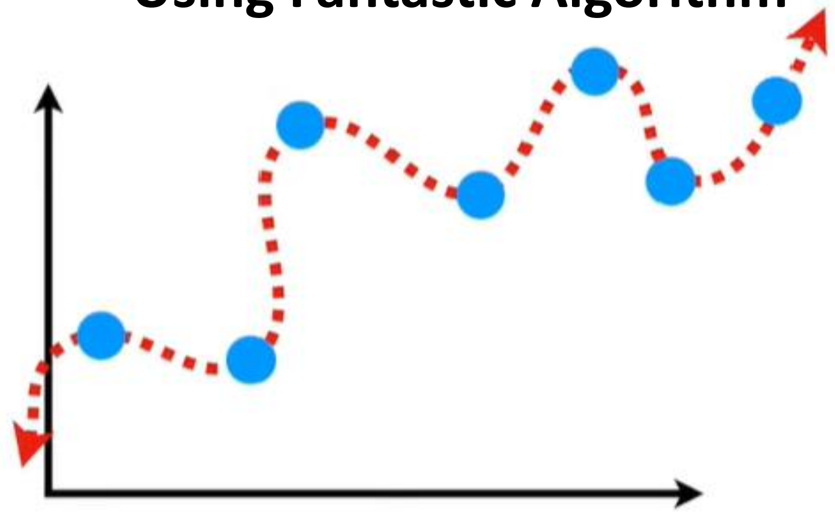
So, which algorithm would you prefer?

Using Linear Algorithm



- This one has a bias
- RMS Error is high

Using Fantastic Algorithm

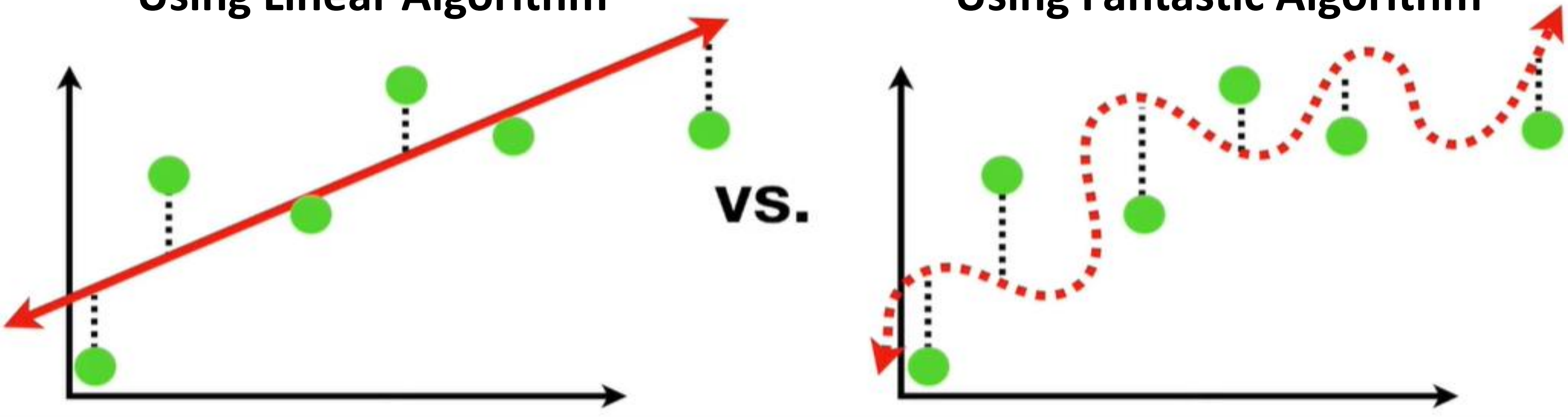


- This one has no bias
- RMS Error is Zero
- This one is a clear winner!

Bias and Variance

Let’s now test our algorithms with test data

Using Linear Algorithm **Using Fantastic Algorithm**

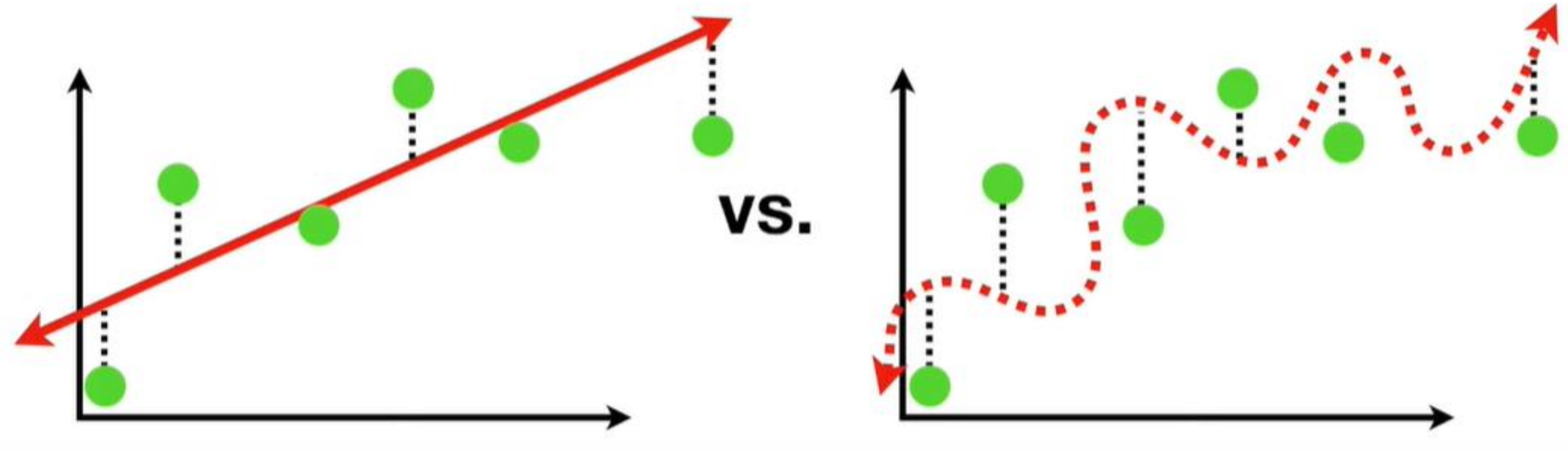


- This one fits the test data better!
- Model is consistent between train and test data
- Linear Regression Algorithm wins!

- This one does not fit the test data at all!
- Even though it did a great job with training dataset, it did a terrible job with the testing dataset.

Bias and Variance

Using Linear Algorithm | **Using Fantastic Algorithm**



Using Linear Algorithm

- ✓ This one has **high bias** because it cannot capture the true relationship between weight & height.
- ✓ It has **low variance** because RMS error is similar between datasets.
- ✓ **Underfitting** the curve

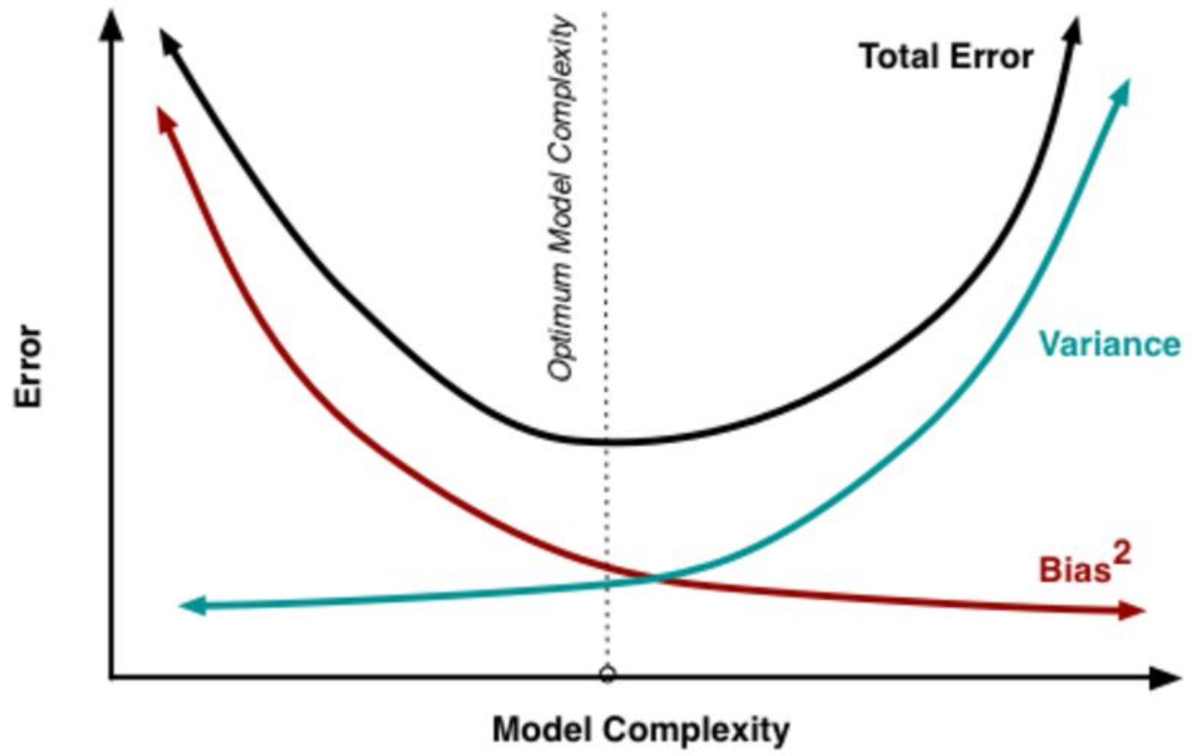
Using Fantastic Algorithm

- ✓ This one has **low bias** because it truly models the behaviour of weight and height.
- ✓ But it has **high variability** from one dataset to another.
- ✓ **Overfitting** the curve

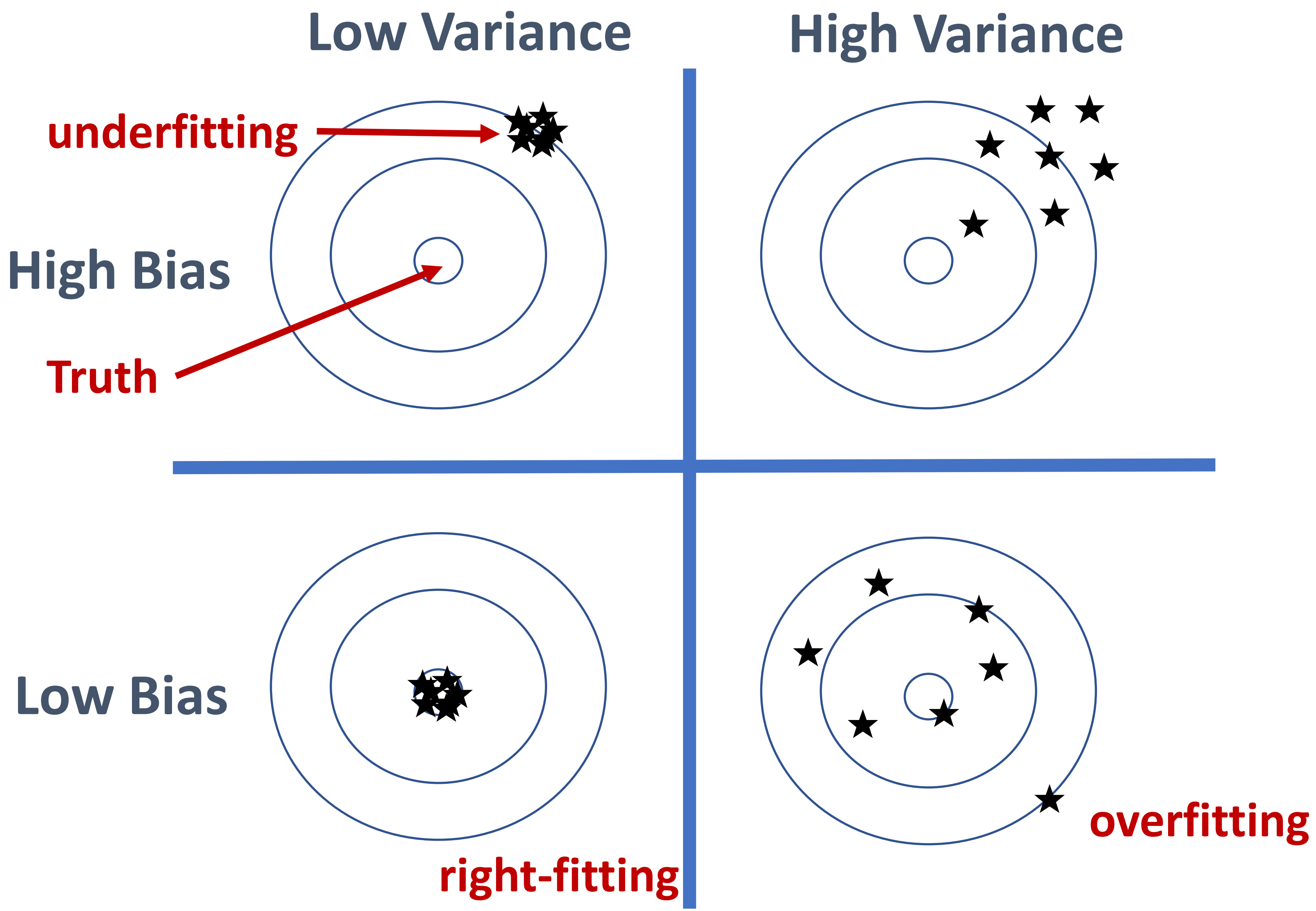
Bias – Variance *Trade-off*

- There is no escaping the relationship between bias and variance in machine learning.

- **Increasing the bias will decrease the variance.**
- **Increasing the variance will decrease the bias.**



Bias – Variance *Trade-off*

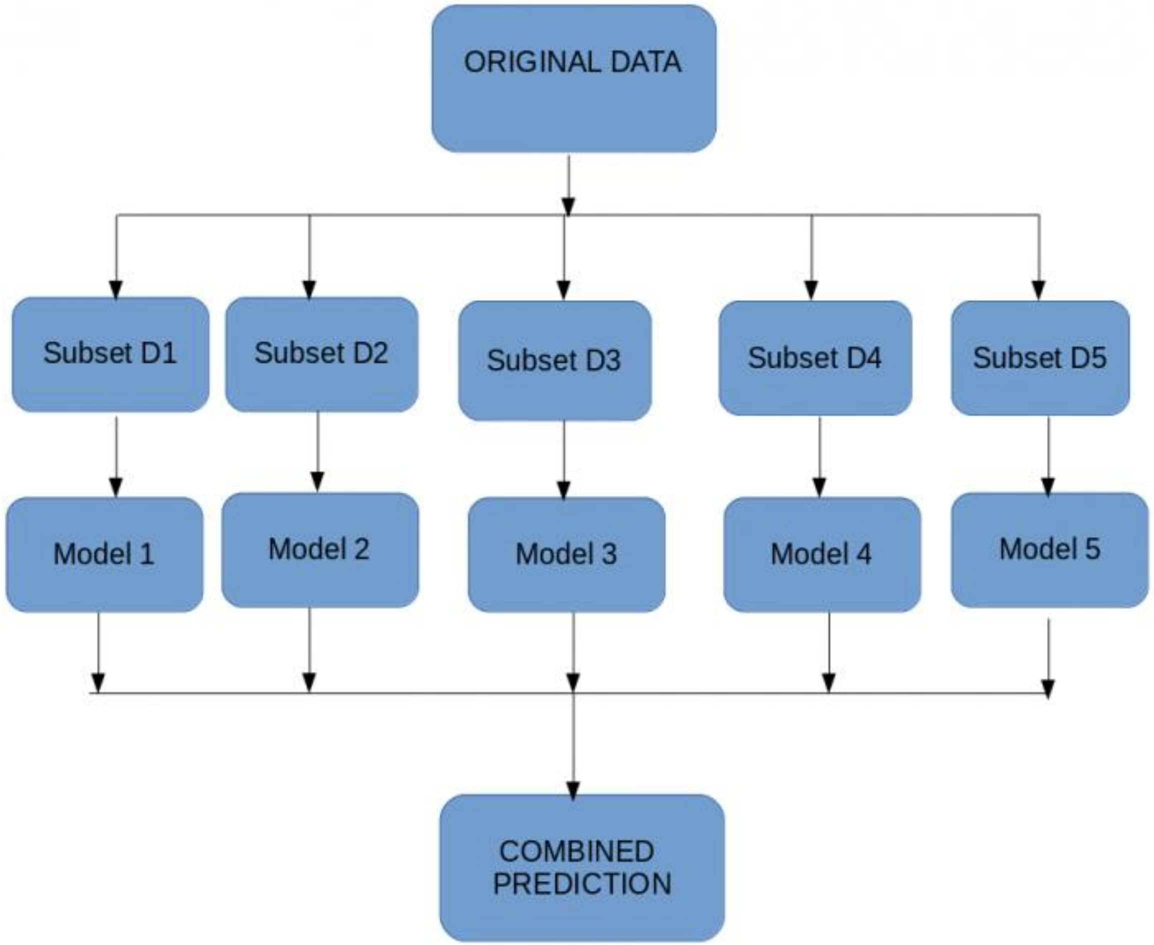


How to reduce Bias?

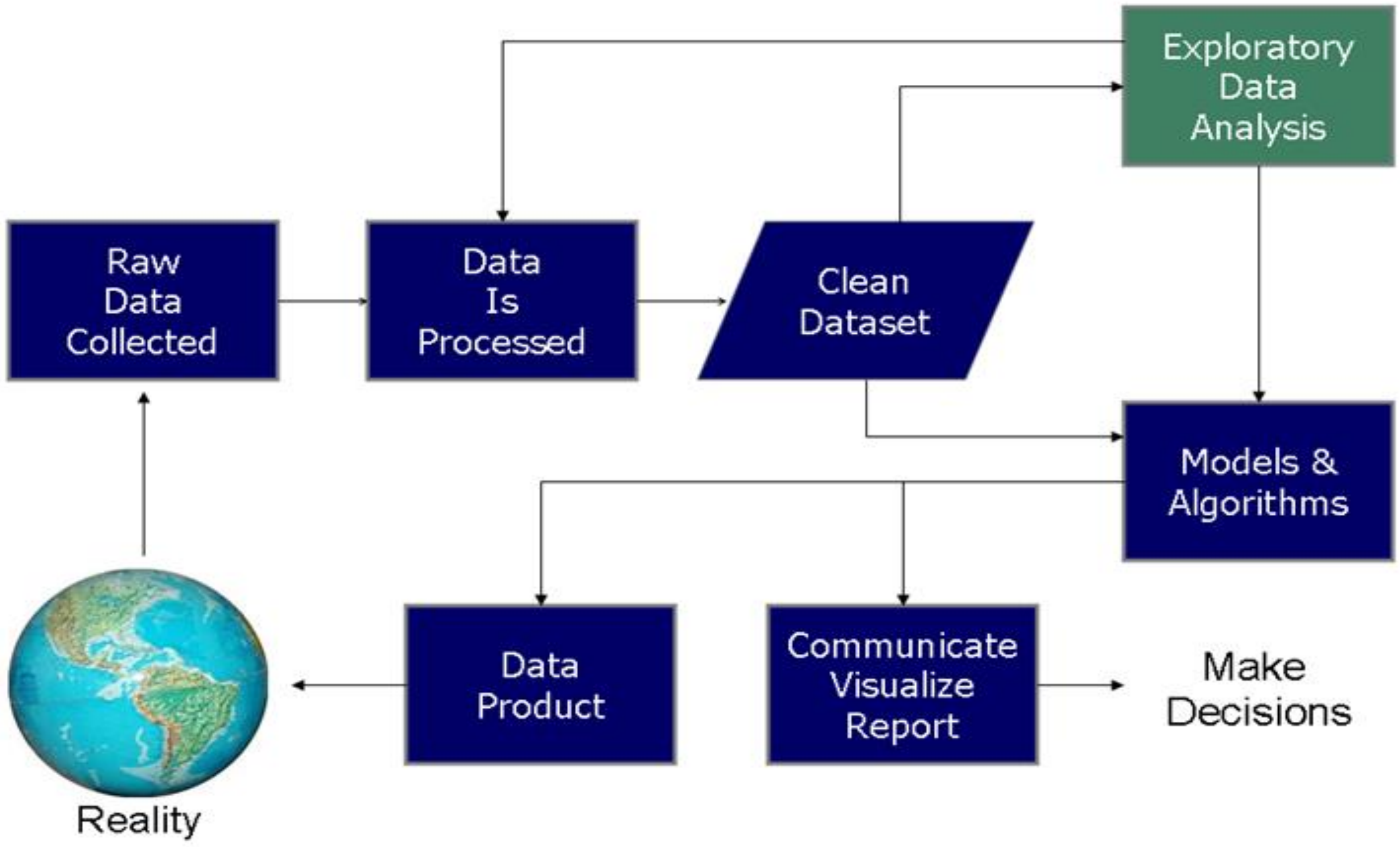
- Ensure that the data collected has diverse factors such as population, markets, demographics
- Ensure that the data truly represents the underlying pattern & trend
- Focus on fairness, inclusivity, transparency and reliability
- Ensure proper features engineering is done – include all influencing features
- Data size should be large
- Select the right model suitable for that dataset

How to reduce variance?

- Train using multiple models
- Reduce noise by eliminating unwanted features
- Use Bagging (Bootstrap aggregation) to build better datasets



Exploratory Data Analysis Process



Know your customer through their data

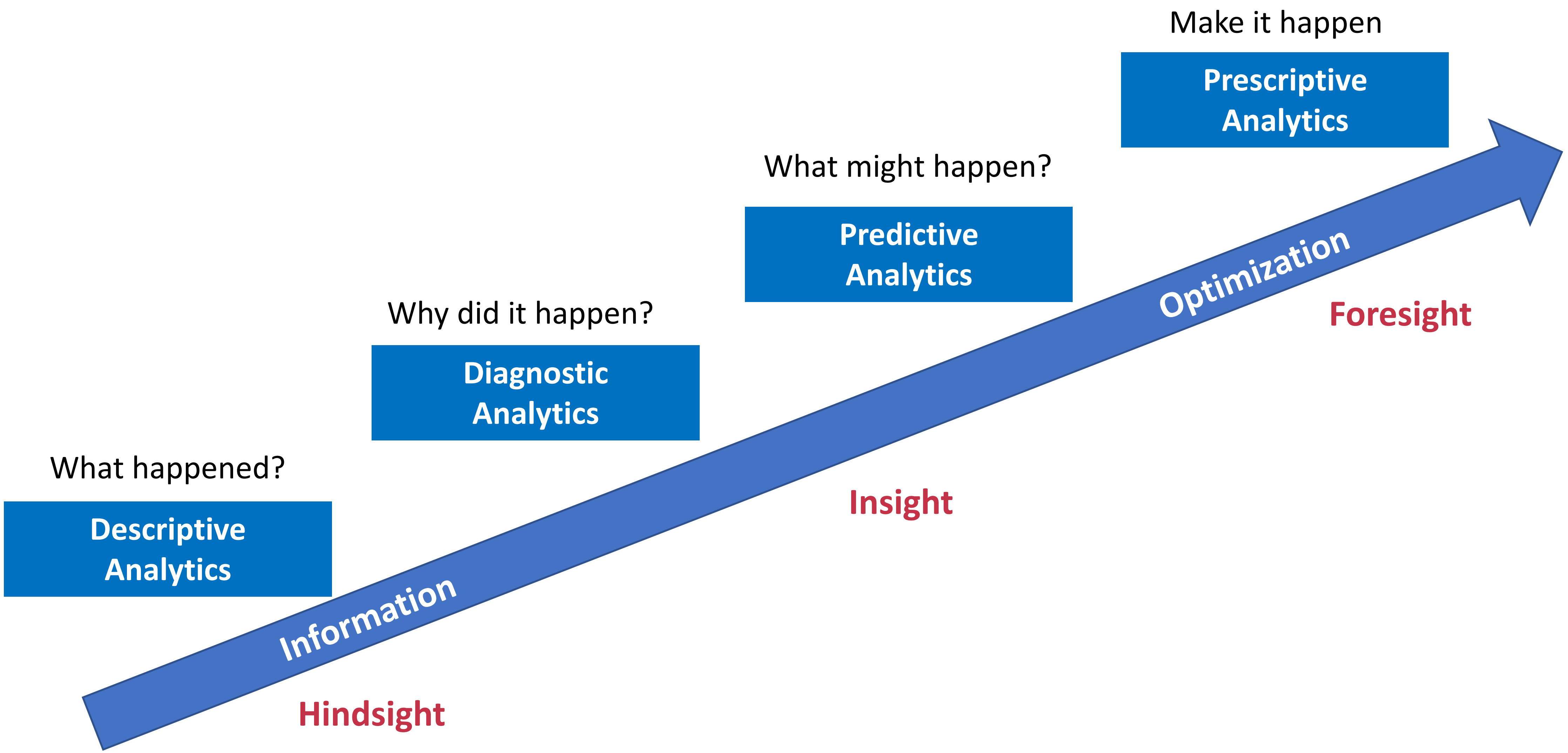
Know your audience

Collect relevant datasets

Choose the right analysis

Choose the right chart

Types of Exploratory Data Analytics



Role based Exploratory Data Analysis

SN	Report	KPIs
1	**Management Report**	• Quarterly Revenue • # of customers added • Avg revenue per customer • Customer Acquisition Cost
2	**CMO Report**	• # Leads generated (MQL & SQL) • Conversion Rate • Avg Ad spend per customer
3	**CFO Report**	• ………………
4	**Manufacturing Production Report**	• Quantity/Volume produced per a time period • Product backlogs • # Defects • Manufacturing Cost 6
5	**CTO Report**	
6	**Talent Management Report**	• Demand versus Supply • Talent attrition rate • Talent Rating
7	**Hospital Admin Report**	---------------

Benefits of Exploratory Data Analysis

- Enables C-level decision making betters and faster
- Provide invaluable and highly insightful views of the business
- Enables creation and sharing of reports across organization
- Promotes disciplined approach to reporting

Use cases for Exploratory Data Analysis

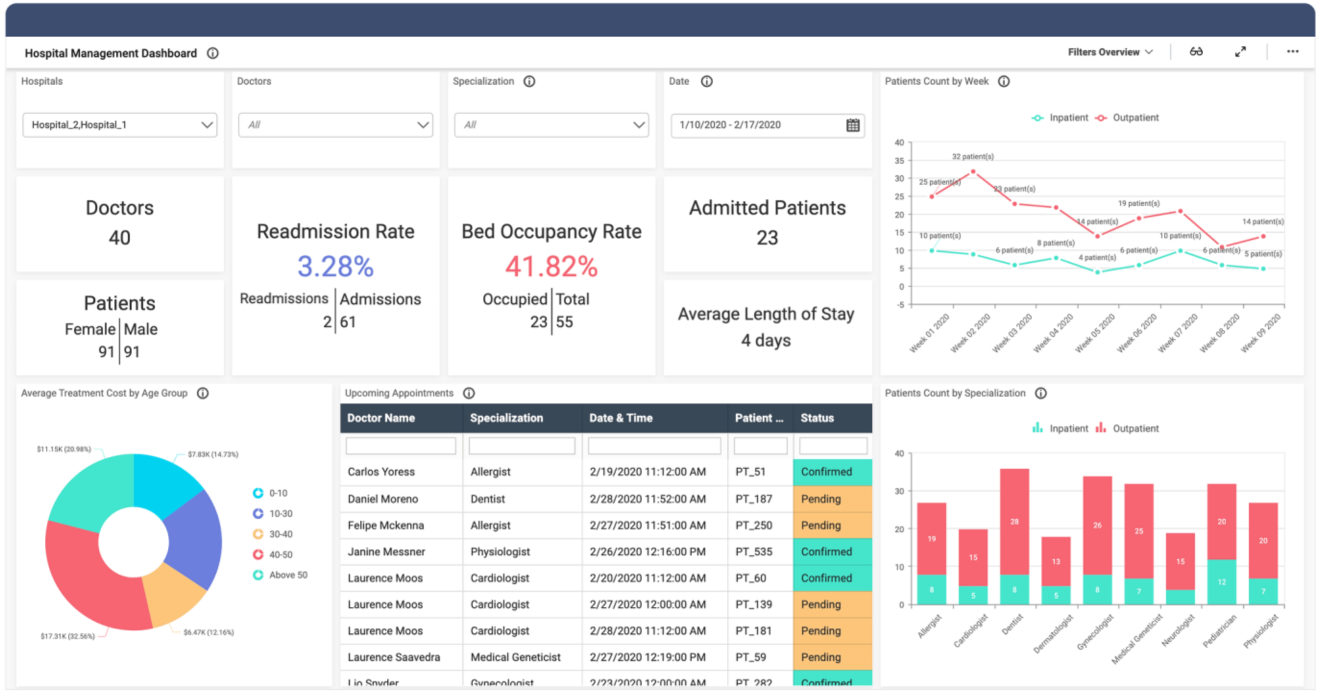
- Customer Profitability Report
- HR Performance Report
- Procurement Analysis Report
- Retail Business Analysis Report
- Sales and Marketing Report
- Supplier Quality Analysis Report
- IT Spend Analytics Report

Interactive Dashboards used in EDA

- A single window to critical performance parameters, be it business or operations

- It helps to monitor and track changes as they happen

- A great tool for date driven decision making



Interactive Dashboards Demos

Marketing Lead Trends Dashboard

https://www.inetsoft.com/sree/app/viewer/view/global/gallery/Marketing%20Leads

Call Centre Team Performance Dashboard

https://www.inetsoft.com/sree/app/viewer/view/global/gallery/Call%20Center%20Analysis

Call Centre Monitoring Dashboard

https://www.inetsoft.com/sree/app/viewer/view/global/gallery/Call%20Center%20Monitoring

Executive Sales Summary Dashboard

https://www.inetsoft.com/evaluate/bi_visualization_gallery/dashboard.jsp?dbIdx=8

Global Visual Dashboard

Global Visual Analysis (inetsoft.com)

Source: Inetsoft.com

Top Enterprise Report Building Tools/Languages

EDA Demos

Setting up Python Environment

Setting up Python Environment

- **Anaconda** is a open-source Data science platform
- Includes poplar Python packages like **NumPy** and **Pandas**
- **Jupyter Notebook** – Open source web based interface
- **JupyterLab** – Full-pledged IDE for Python
- Both support Interactive interface with code, narrative text and visualization
- You can also use **Google Colab** online

    https://colab.research.google.com/#scrollTo=P-H6Lw1vyNNd

Python EDA Examples Demos in Jupyter Notebook

EDA Demos using Power BI Desktop

Questions?