MACHINE LEARNING

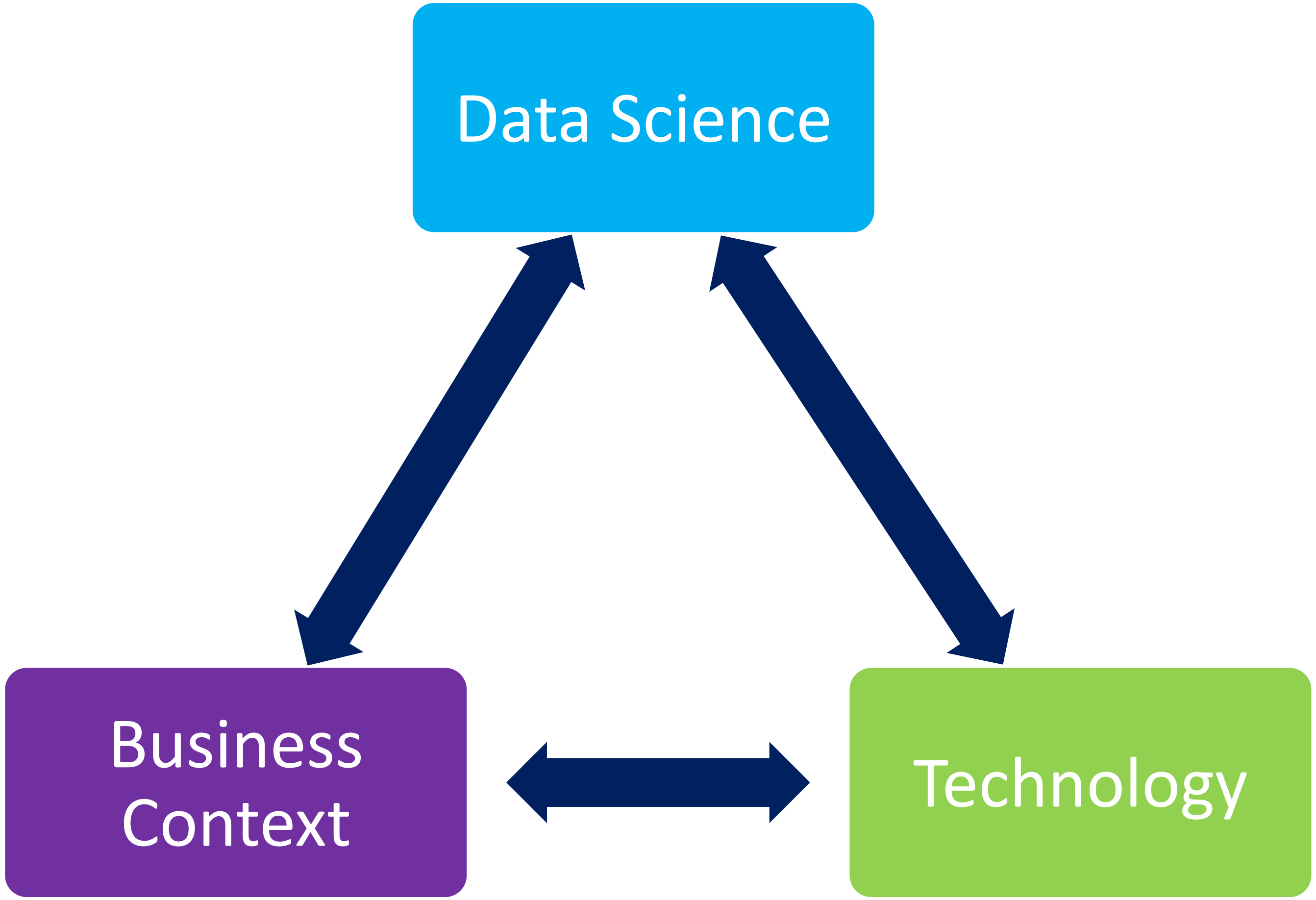
“In God we trust; all others must bring Data”

-Edwards Deming

BUSINESS ANALYTICS

DECISION MAKING

Statistical models & Machine Learning Algorithms



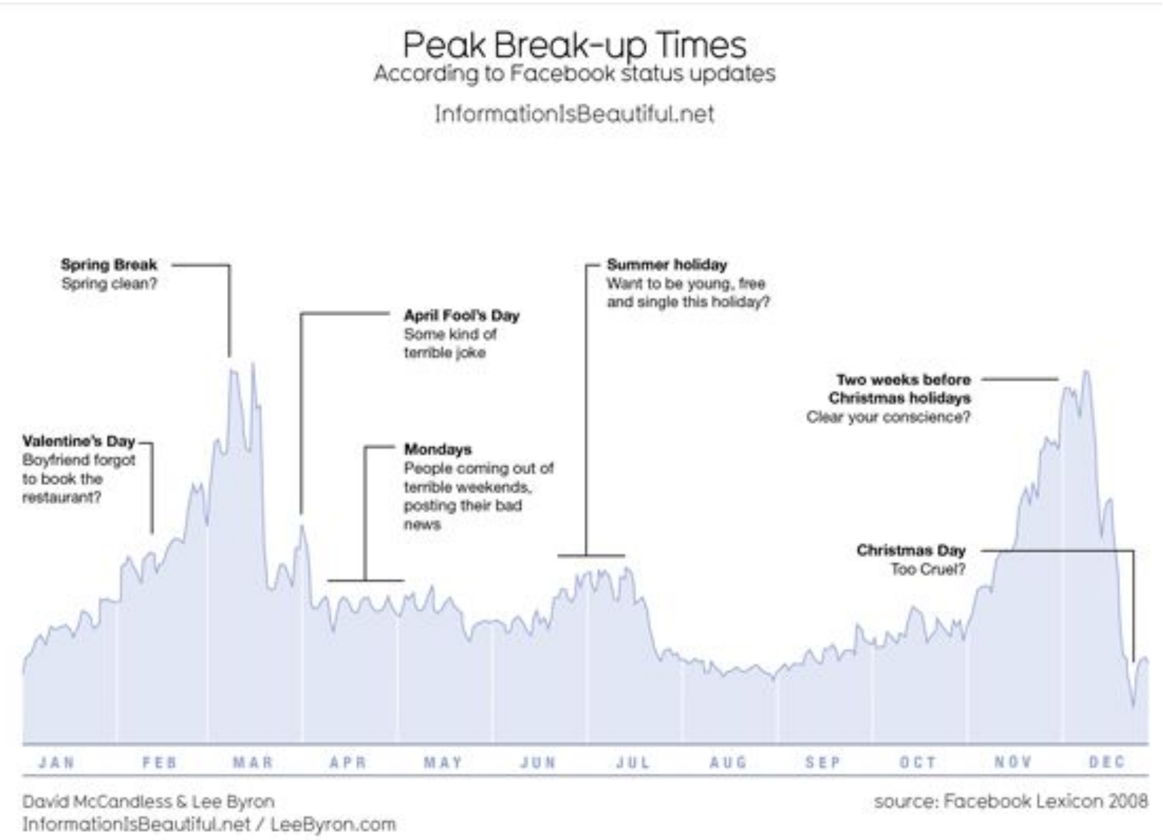
Data Science

Business Context

Technology

Problems, Opportunities, Decision Scenarios

Data collection, storage, retrieval Software tools



Peak Break-up Times
According to Facebook status updates
InformationIsBeautiful.net

Spring Break
Spring clean?

April Fool's Day
Some kind of terrible joke

Summer holiday
Want to be young, free and single this holiday?

Two weeks before Christmas holidays
Clear your conscience?

Valentine's Day
Boyfriend forgot to book the restaurant?

Mondays
People coming out of terrible weekends, posting their bad news

Christmas Day
Too Cruel?

David McCandless & Lee Byron
InformationIsBeautiful.net / LeeByron.com

source: Facebook Lexicon 2008

- There will be more traffic to online dating sites during December
- There will be greater demand for relationship counsellors and lawyers
- There will be greater demand for housing
- People would like to forget the past, so they might change the brand of beer they drink

Few examples of insights obtained using descriptive analytics reported in literature

- Most shoppers turn towards the right side when they enter a retail store
- Married men who kiss their wife before going to work live longer, earn more and get into less number of accidents as compared to who do not
- Divorce spike in the month of December and January
- Men are more reluctant to use coupons as compared to women.
- Strawberry pop-tarts sell 7 times more during hurricane compared to regular period (Wal Mart).

Definition – TOM MITCHELL

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”

Example: playing checkers.

E = the experience of playing many games of checkers

T = the task of playing checkers.

P = the probability that the program will win the next game.

CLASSIFICATION OF MACHINE LEARNING ALGORITHMS

- **SUPERVISED LEARNING** – both input and output of the training set known
- **UNSUPERVISED LEARNING** – only input variable is known but not the outcome
- **REINFORCEMENT LEARNING** – both input and output variables are uncertain

SUPERVISED LEARNING

- REGRESSION

  predicting results in a continuous output

- CLASSIFICATION

  predicting results in a discrete output

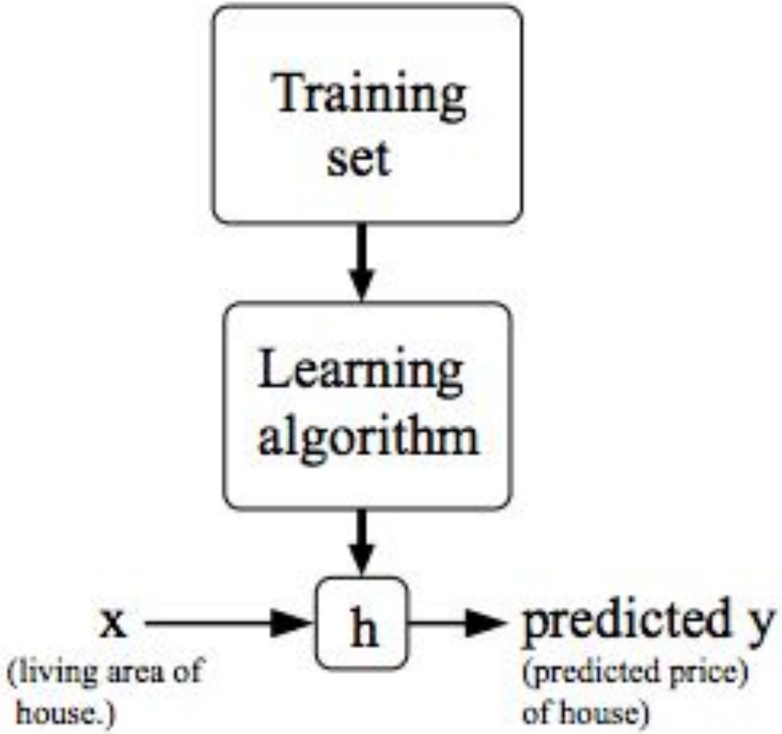
UNSUPERVISED LEARNING

Unsupervised learning allows us to approach problems with little or no idea.

<u>Example:</u>

- Google News
- Social network analysis
- Market segmentation

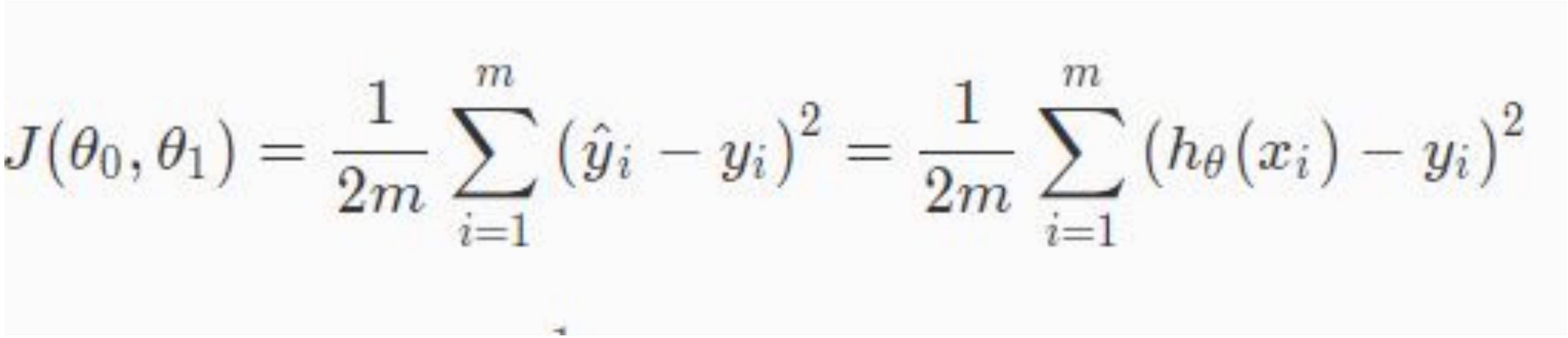
MODEL REPRESENTATION



$$h_\theta(x)=\theta_0+\theta_1 x$$

COST FUNCTION

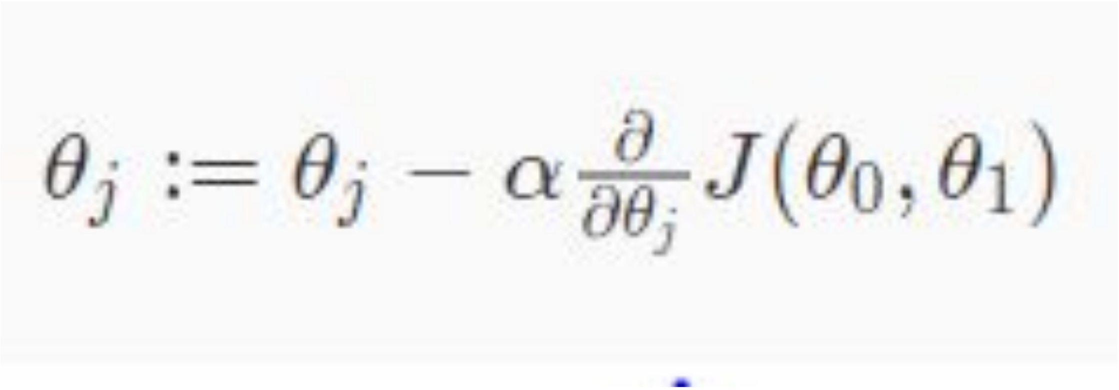
Cost function also known as “squared error function” or “mean squared error” is a measure of accuracy of our hypothesis.



$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}_i - y_i\right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i\right)^2$$

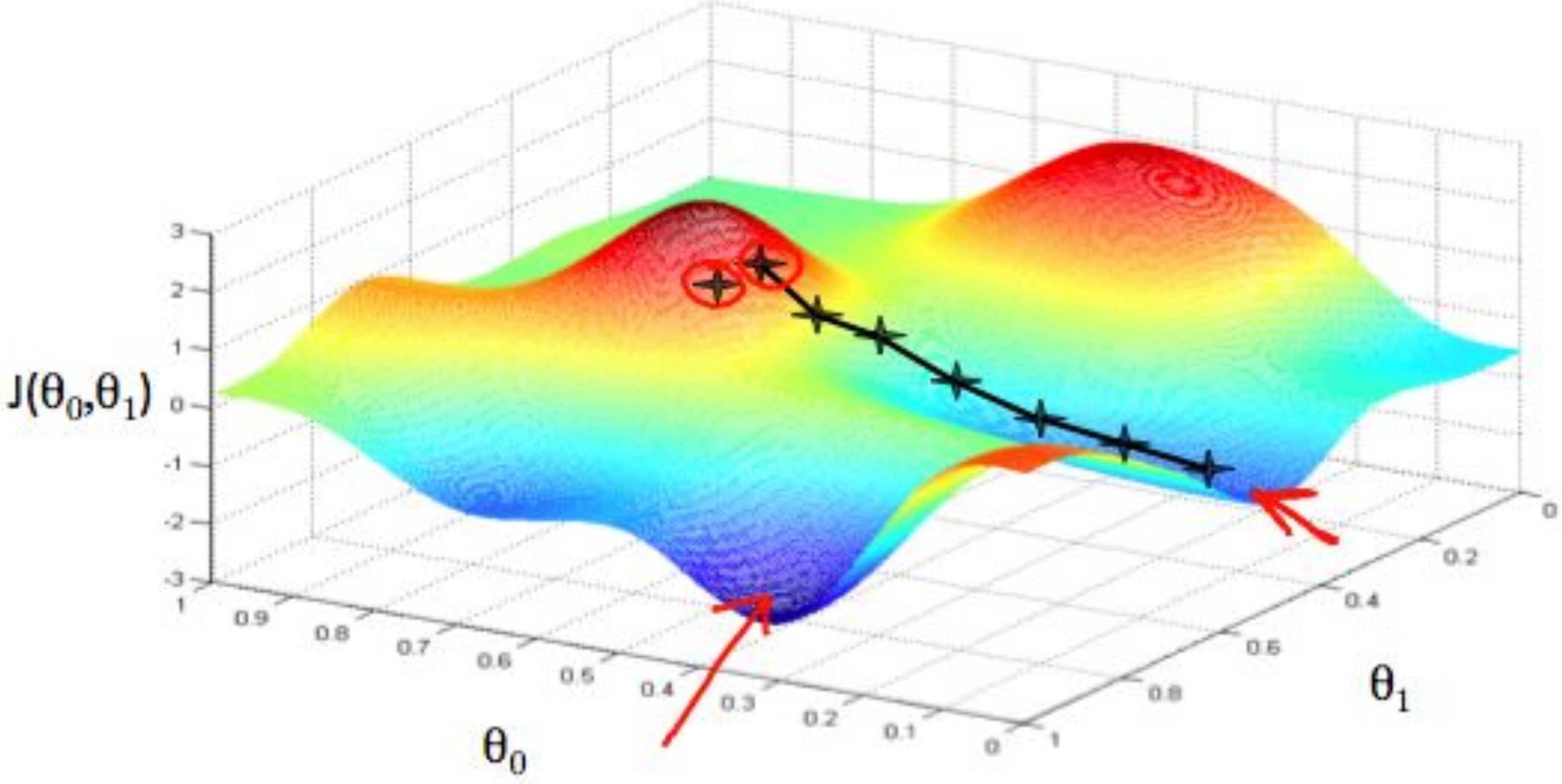
GRADIENT DESCENT

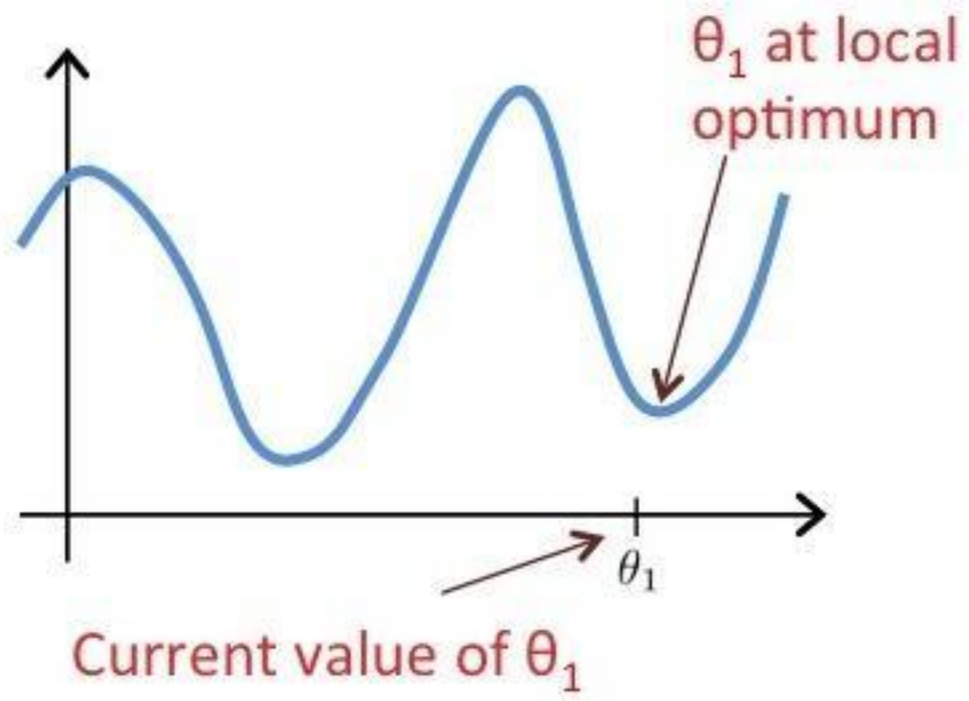
An algorithm to minimize the cost function



$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

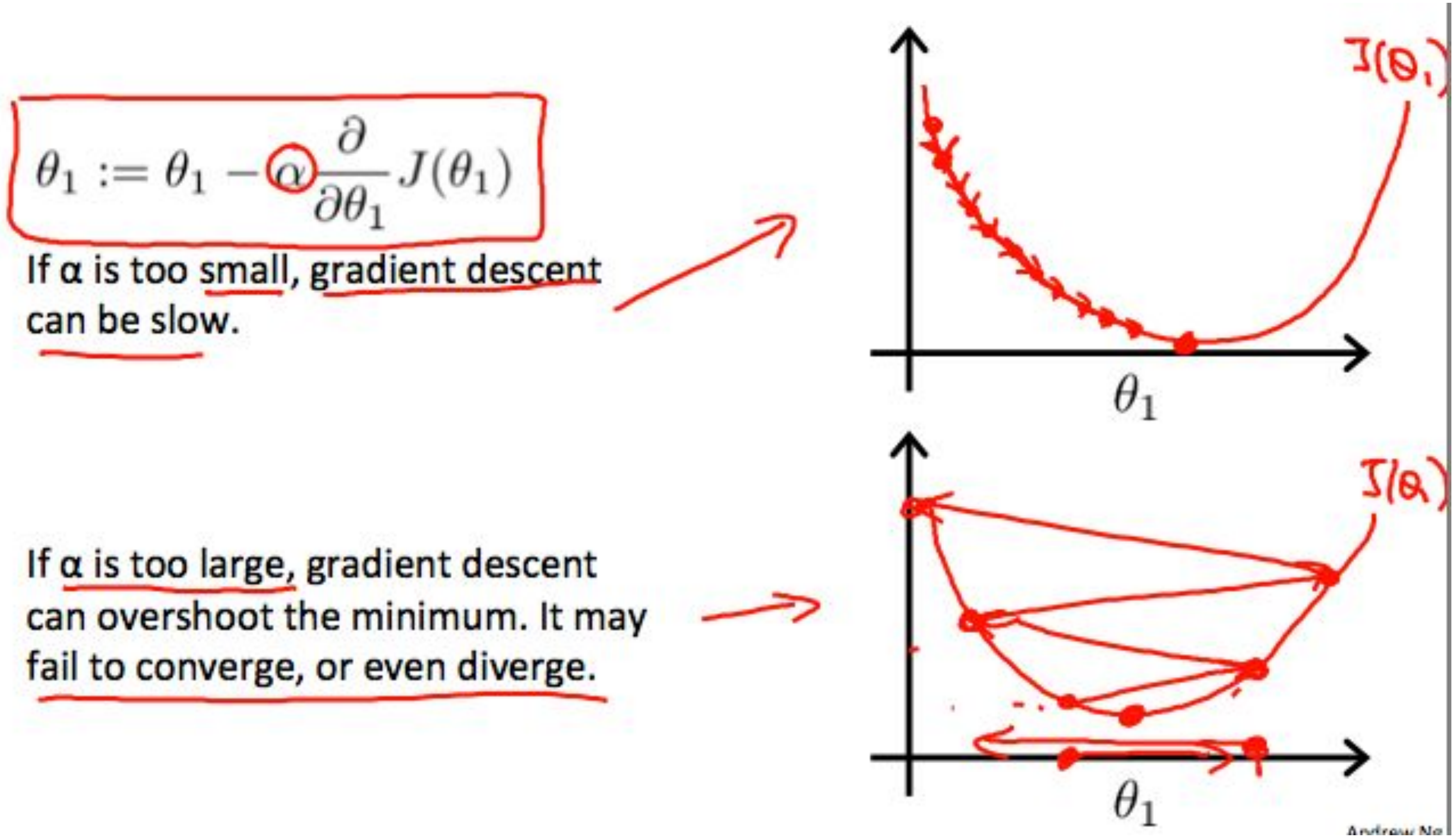
α= learning rate
j=0,1 represents the feature index number


J(θ_0,θ_1)
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ_0
0
0.2
0.4
0.6
0.8
1
θ_1



- Leave θ_1 unchanged
- Change θ_1 in random direction
- Move θ_1 in the direction of global minimum
- Decrease θ_1

Choosing learning rate α



$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If α is too small, gradient descent can be slow.

If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.

Introduction To Descriptive Analytics

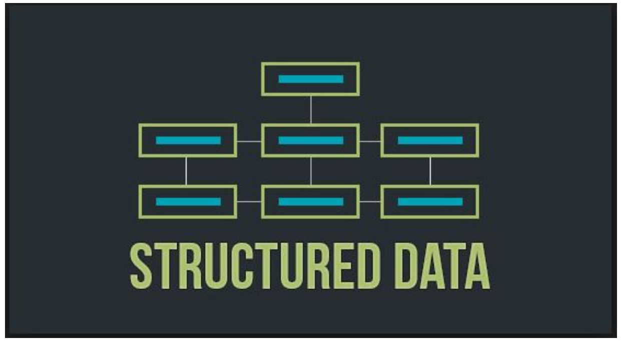
- Simple comprehension of data using data summarization

- Basic statistical measures and visualization.

- Designing effective dashboards and scorecards.

Structured and Unstructured Data

- Structured data means that the data is described in a matrix form with labelled rows and columns.



- Any data that is not originally in the matrix form with rows and columns is an unstructured data.

Structured data consisting of nominal and ratio scales

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

Data Type

Cross-Sectional Data: A data collected on many variables of interest at the same time or duration of time is called cross-sectional data.

Time Series Data: A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.

Panel Data: Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data).

TYPES OF DATA MEASUREMENT SCALES

- Nominal scale refers to variables that are basically names (qualitative data) and also known as categorical variables.
- Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- Interval scale corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade) or intelligence quotient (IQ) score are examples of interval scale
- Any variable for which the ratios can be computed and are meaningful is called ratio scale.

Population And Sample

- Population is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem.

- Sample is the subset taken from a population.

Measures Of Central Tendency

Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of

central tendency.

$$\text{Mean}=\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Mean

Symbol $\bar{X}$ is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we have the population mean which is denoted by μ (population mean).

In Table 2.1, the average salary is given by

$$\bar{X} = \frac{(270+220+240+250+180+300+240+235+425+240)\times 1000}{10} = 260000$$

Property of Mean

An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

Median (or Mid) Value

Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%.

Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position $(n + 1)/2$ when n is odd. When n is even, the median is the average value of $(n/2)^{th}$ and $(n + 2)/2^{th}$ observation after arranging the data in the increasing order.

Mode

Mode is the most frequently occurring value in the dataset

Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless.

For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female. Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database

Percentile

Percentile, decile and quartile are frequently used to identify the position of the observation in the dataset.

Percentile, denoted as P_x, is the value of the data at which x percentage of the data lie below that value

Position corresponding to $P_x \approx$ x (n+1)/100

P_x is the position in the data calculated , where n is the number of observations in the data.

Decile and Quartile

Decile corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on.

Quartile divides the data into 4 equal parts. The first quartile (Q_1) contains first 25% of the data, Q_2 contains 50% of the data and is also the median. Quartile 3 (Q_3) accounts for 75% of the data

Example (Percentile Calculation)

Time between failures of wire-cut (in hours)

2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

1. Calculate the mean, median, and mode of time between failures of wire-cuts
2. The company would like to know by what time 10% (ten percentile or P_{10}) and 90% (ninety percentile or P_{90}) of the wire-cuts will fail?
3. Calculate the values of P_{25} and P_{75}.

Measures of Variation

Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X)

Variability in the data is measured using the following measures:

- Range
- Variance
- Standard Deviation

Range and Variance

Range is the difference between maximum and minimum value of the data. It captures the data spread.

Variance is a measure of variability in the data from the mean value. Variance for population, σ^2, is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}$$

Sample Variance

In case of a sample, the Sample Variance (S^2) is calculated using

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

While calculating sample variance S2, the sum of squared deviation $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ is divided by (n-1), this is known as Bessel’s correction.

Standard Deviation

The population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \qquad S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$$

Measures of Shape – Skewness and Kurtosis

Skewness is a measure of symmetry or lack of symmetry. A dataset is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between μ and μ - kσ is same as μ and μ+ kσ, where k is some positive constant.

Pearson's moment coefficient of skewness for a dataset with *n* observations is given by

$$g_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3 / n}{\sigma^3}$$

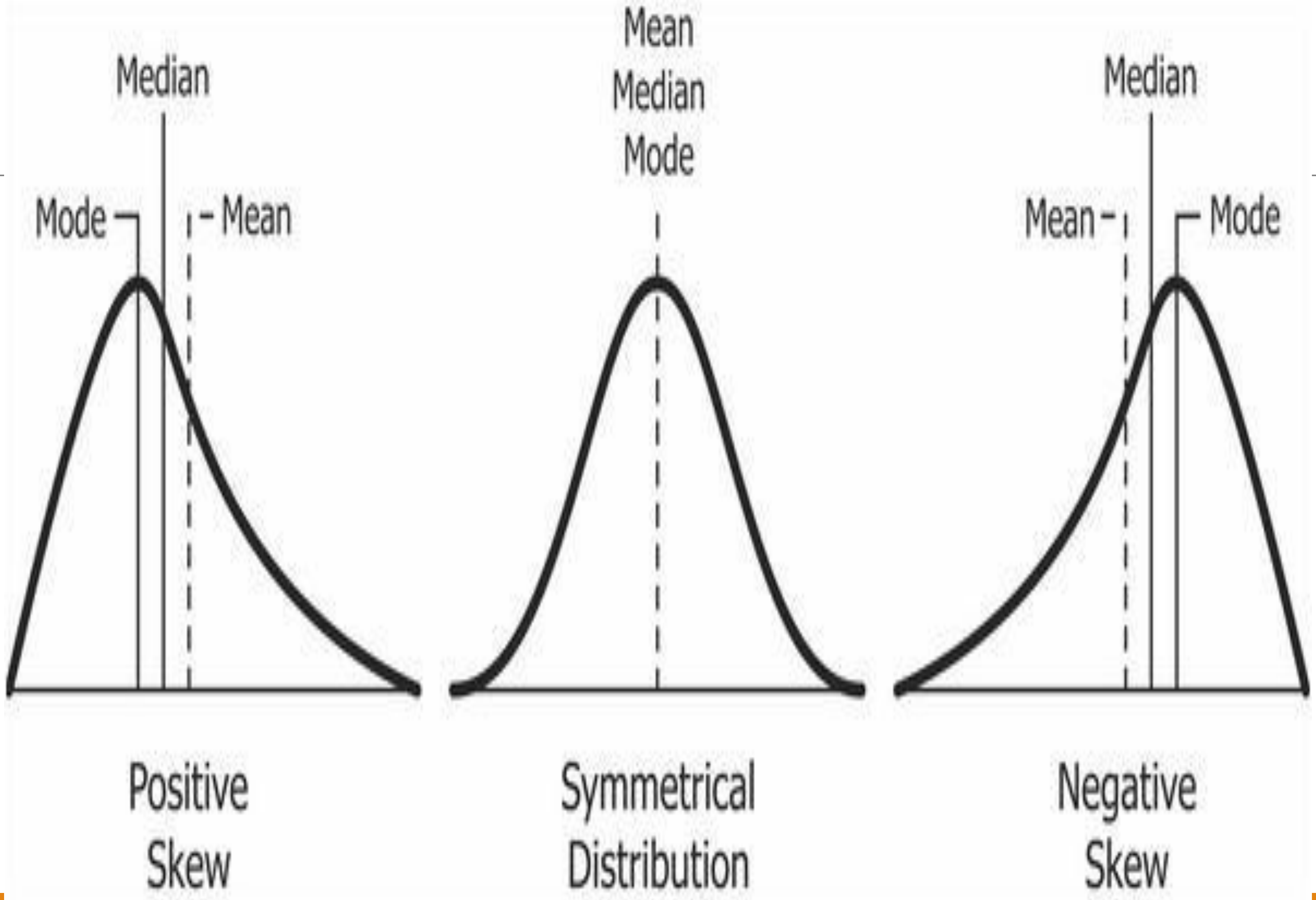
The value of g_1 will be close to 0 when the data is symmetrical. A positive value of g_1 indicates a positive skewness and a negative value indicates **negative skewness.**

Skewness

The following formula is used usually for a sample with n observations (Joanes and Gill, 1998):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

The value of $\frac{\sqrt{n(n-1)}}{n-2}$ will converge to 1 as the value of n increases.


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

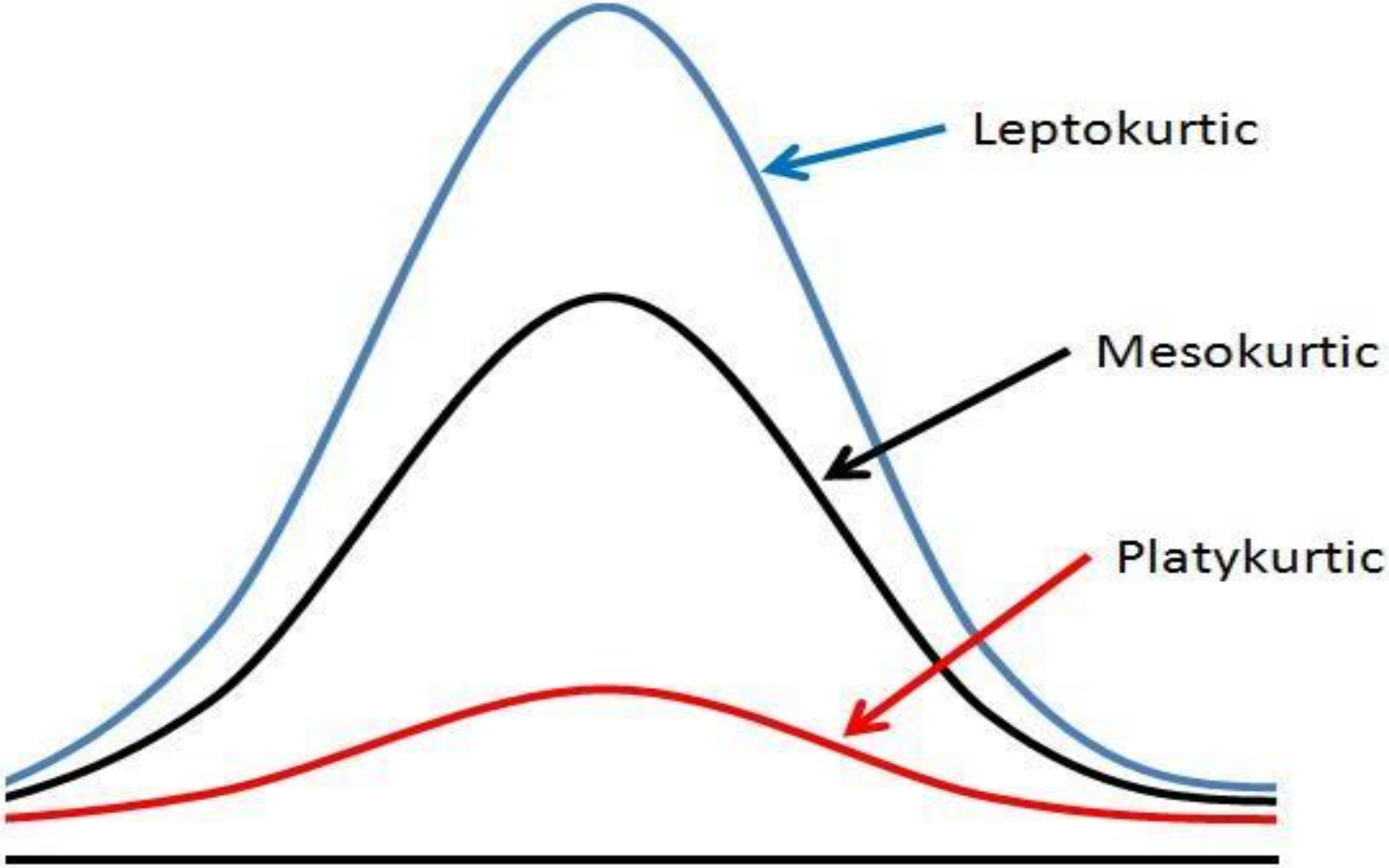
Kurtosis

Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{4}\left(X_i - \bar{X}\right)^4 / n}{\sigma^4}$$

Kurtosis value of less than 3 is called platykurtic distribution and greater than 3 is called leptokurtic distribution. The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**)

Leptokurtic, mesokurtic, and platykurtic distributions



Excess Kurtosis

The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by:

$$\text{Excess Kurtosis}= \frac{\sum_{i=1}^{4}\left(X_i-\bar{X}\right)^4/n}{\sigma^4}-3$$

Data Visualization

Data visualization is an integral part of descriptive analytics and it assists decision maker with useful insights

There are many useful charts such as histogram, bar chart, pie-chart, box-plot that would assist data scientist with visualization of the data

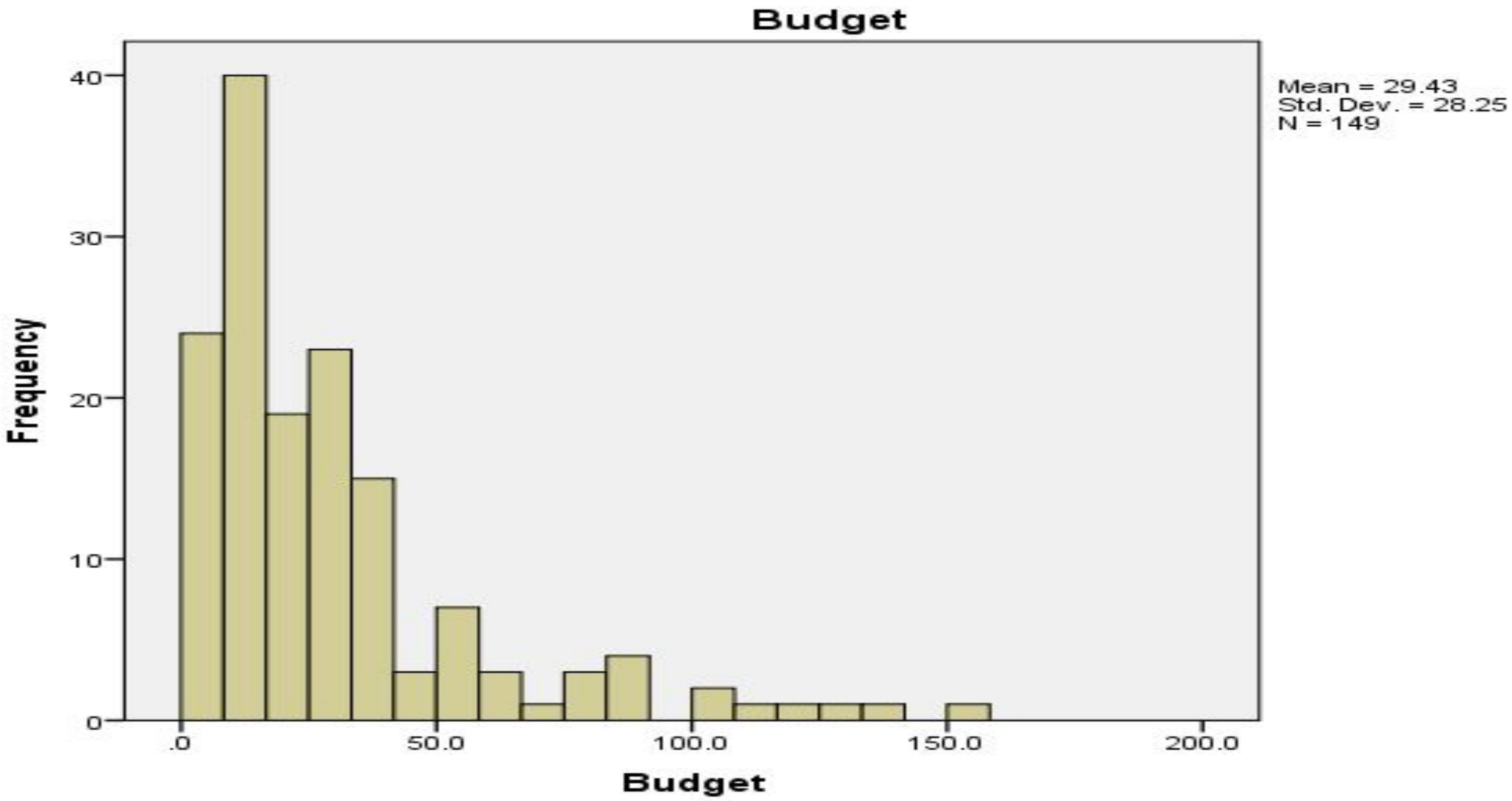
Histogram

Histogram is the visual representation of the data which can be used to assess the probability distribution (frequency distribution) of the data

Histograms are created for continuous (numerical) data.

It is a frequency distribution of data arranged in consecutive and non-overlapping intervals

Histogram of Bollywood movie budget



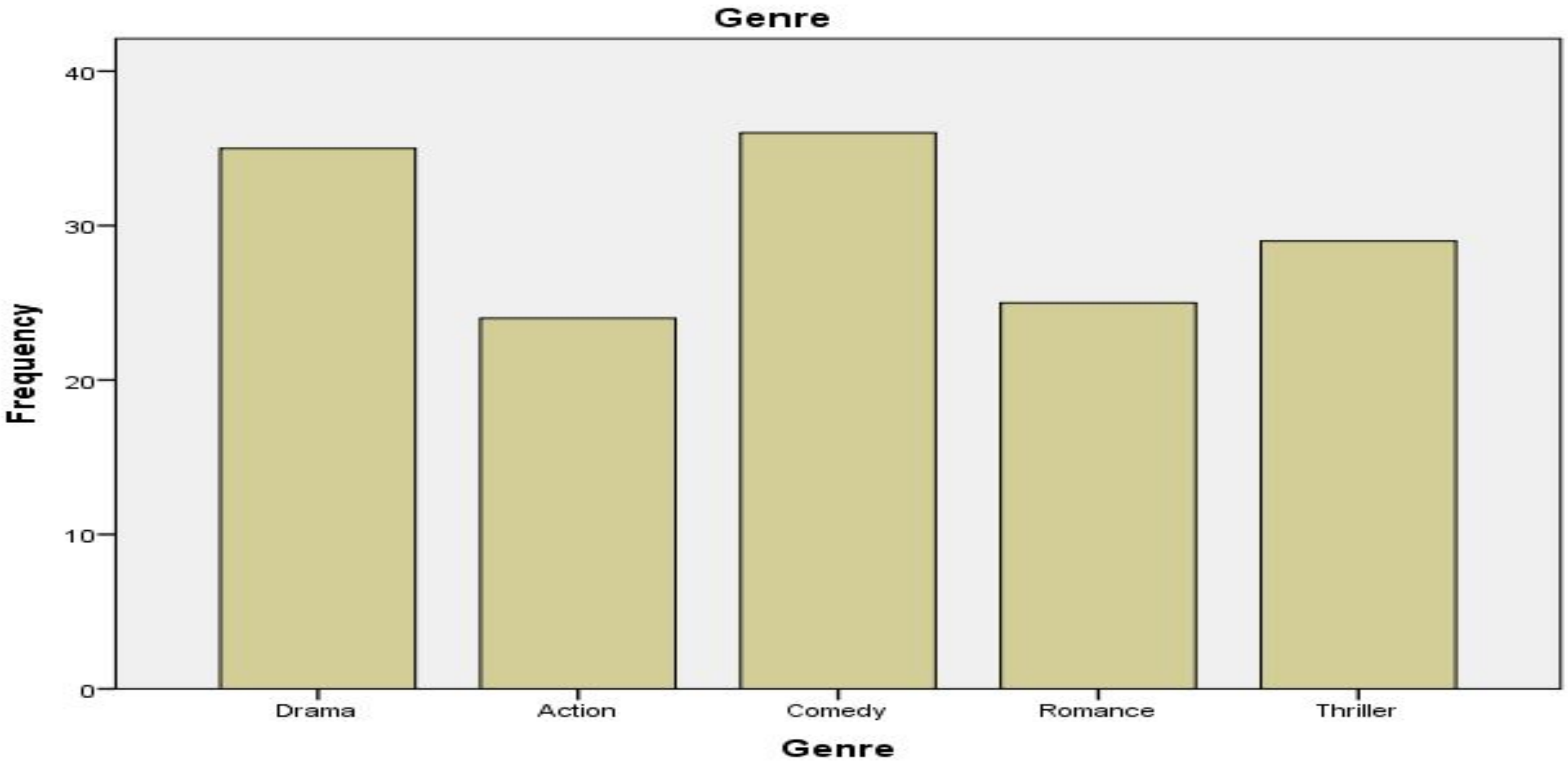
Bar Chart

Bar chart is a frequency chart for qualitative variable (or categorical variable)

Bar chart can be used to assess the most-occurring and least-occurring categories within a dataset

Histograms cannot be used when the variable is qualitative

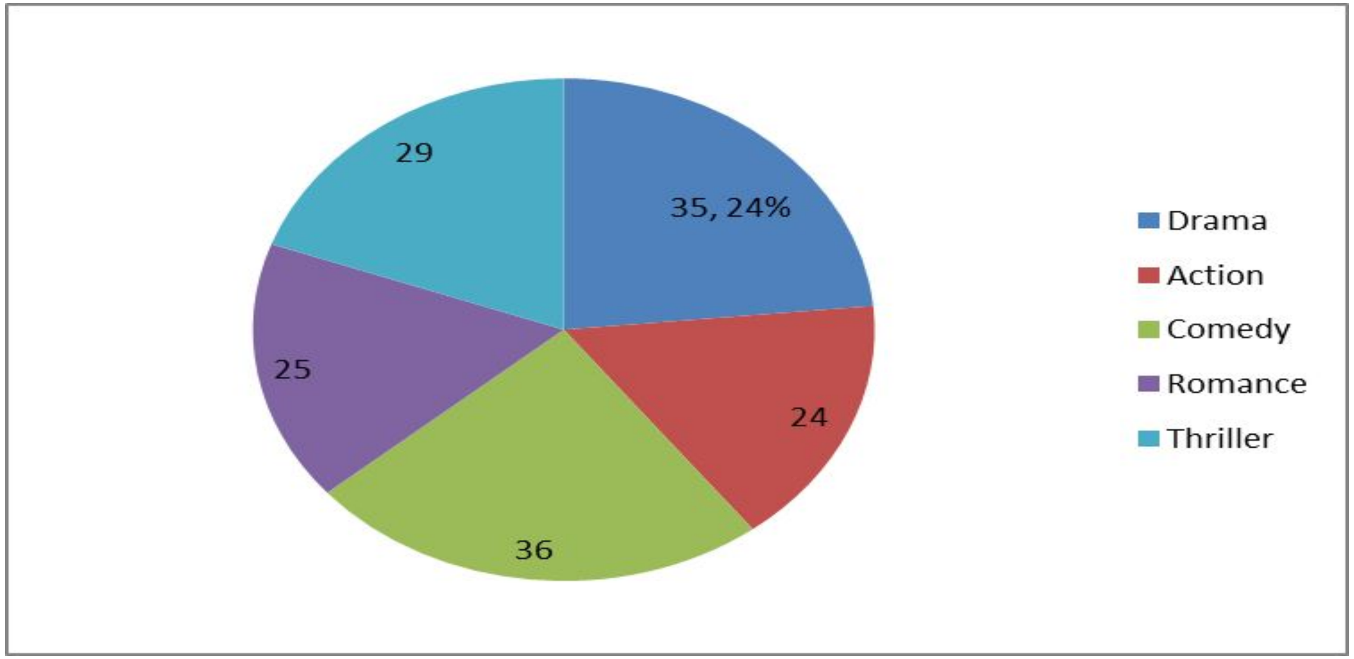
Bar chart for movie genre



Pie Chart

Pie chart is mainly used for categorical data and is a circular chart that displays the proportion of each category in the dataset

Pie chart for movie genre



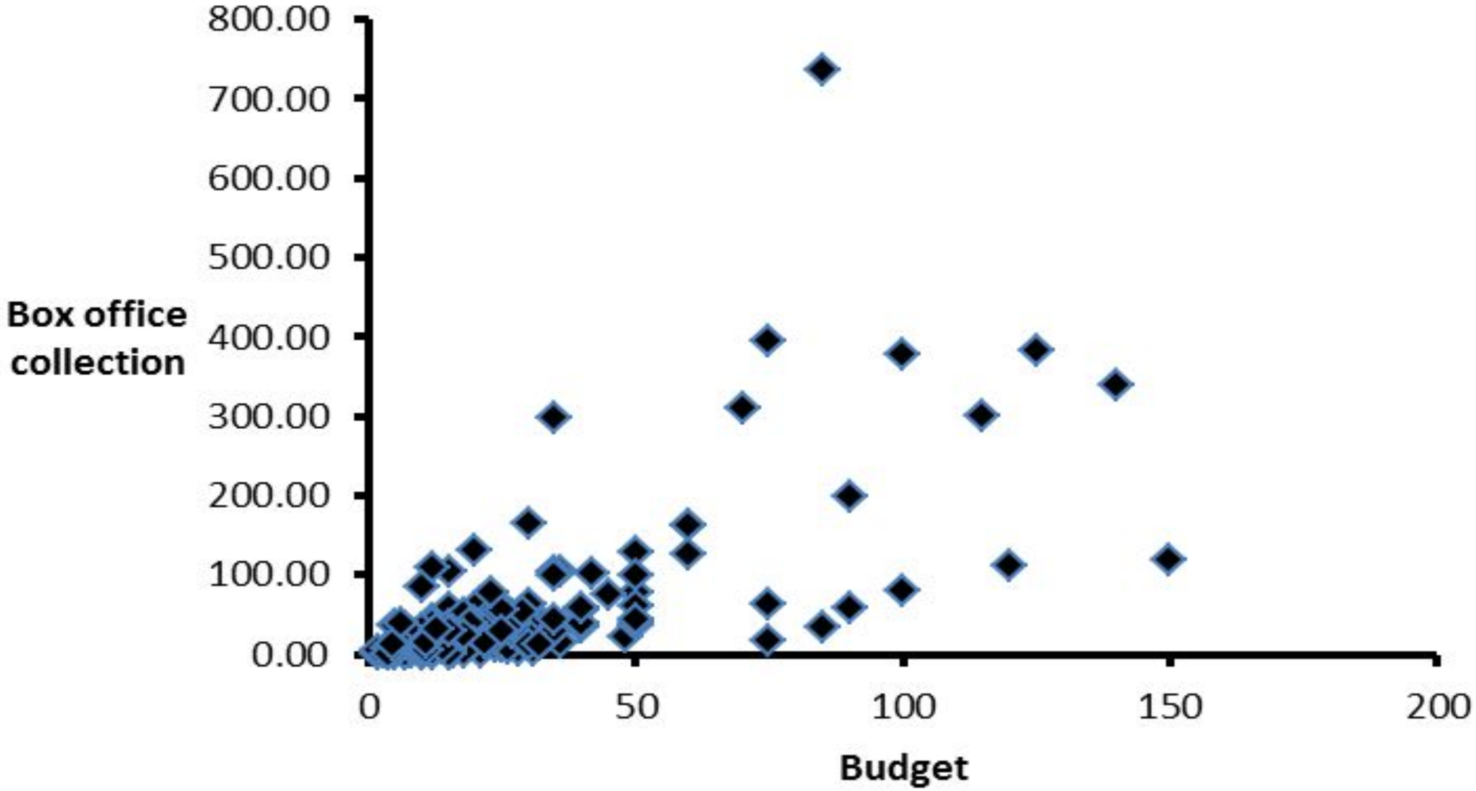
Scatter Plot

Scatter plot is a plot of two variables that will assist data scientists to understand if there is any relationship between two variables

The relationship could be linear or non-linear

scatter plot is also useful for assessing the strength of the relationship and to find if there are any outliers in the data

Scatter plot between movie budget and box office collection


Box office collection
Budget