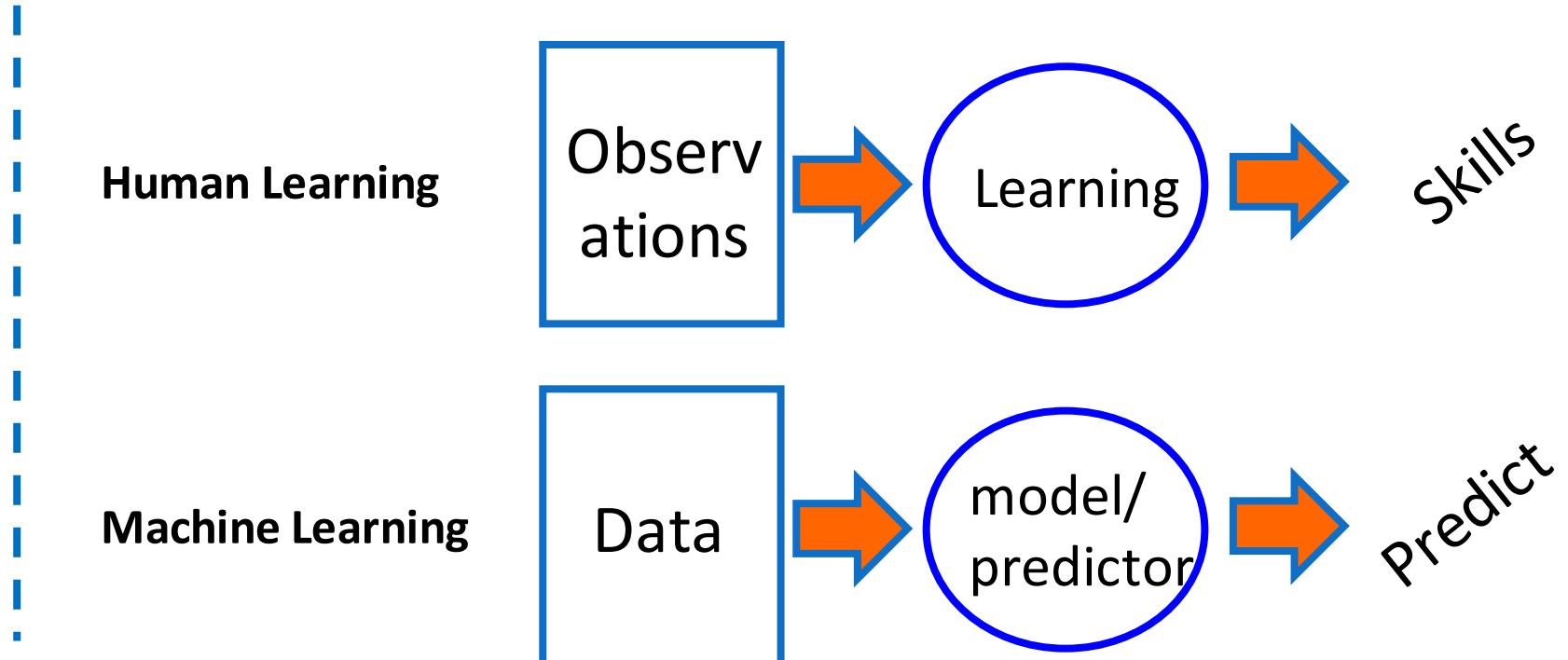


MACHINE LEARNING

Machine learning is programming computers to optimize a performance criterion using example data or past experience. Machine learning can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.

What



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Why

- Organizations/governments are collecting a lot of data
- Information from data is being used to take key business/ political decisions
- At lower levels in organization, data is used for MIS reporting
- At higher levels data based prescriptive and predictive models are being built
- Machine learning** is the most popular technique of creating these predictive and prescriptive model



MACHINE LEARNING

Machine learning is closely associated with Statistics, AI and Data mining

ML vs others

Machine Learning Vs. Statistics

- Traditional Statistics focuses on provable results with mathematical assumptions, and care less about computation
- “Statistics: A useful tool for Machine Learning”

Machine Learning Vs. Artificial Intelligence

- “Machine Learning is one possible route to realize AI”

Machine Learning Vs. Data Mining

- Traditional DM focuses on provable results with math assumptions along with efficient computation in large dataset
- “Difficult to distinguish ML and DM in reality”



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Example



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

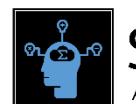
Example



machine l

- machine **learning**
- machine **liker**
- machine **learning course**
- machine **language**
- machine **liker apk**
- machine **learning projects**
- machine **learning tutorial**
- machine **likes**

Gmail Images Sign in



Start-Tech
ACADEMY

Start-Tech Academy

MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Use cases
Banking /
Telecom / Retail

- Identify:
 - Prospective customers
 - Dissatisfied customers
 - Good customers
 - Bad payers
- Obtain:
 - More effective advertising
 - Less credit risk
 - Fewer fraud
 - Decreased churn rate

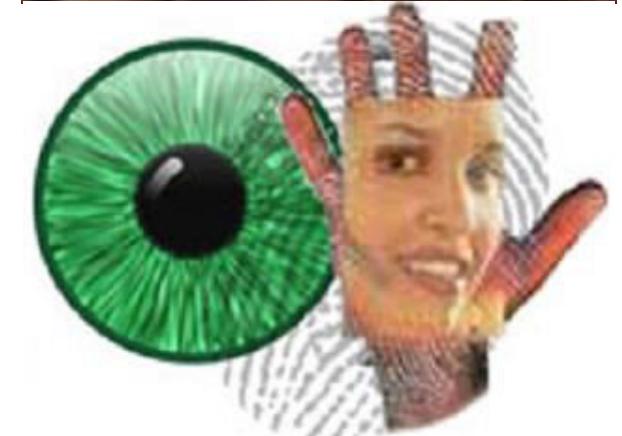


MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Use cases Biomedical / Biometrics

- **Medicine:**
 - Screening
 - Diagnosis and prognosis
 - Drug discovery
- **Security:**
 - Face recognition
 - Signature / fingerprint / iris verification
 - DNA fingerprinting



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Use cases Computer / Internet

- Computer interfaces:
 - Troubleshooting wizards
 - Handwriting and speech
 - Chat bots
- Internet
 - Hit ranking
 - Spam filtering
 - Text categorization
 - Text translation
 - Recommendation



MACHINE LEARNING

Machine learning based on statistics is basically attempting to find the relationship between input and output variables.

Example

For example, a real estate agent who wants to price a particular property will have:

Output variable: Price of property (Y)

Input variables: Area covered (X1), Number of bedrooms (X2), proximity to a landmark (X3), proximity to market (X4), recent sale price of a neighborhood property (X5) and so on

The real estate wants to find out

$$Y = f(X_1, X_2, X_3, X_4, X_5\dots)$$

So that whenever s/he gives a value of the input variables to this function, s/he can get the price of the property.



WHY ESTIMATE $f(x)$

$f(x)$ defines the relationship between dependent and independent variables.

Types

There are two major reasons to estimate $f(x)$:

1. **Prediction** – When the values of input variables is available and output variable is to be predicted. We are only interested in the value of Y, not in the relationship of Y with other variables
2. **Inference** – When the relationship between input and output variable is important. We want to establish how output variable varies with change in each predictor variable



WHY ESTIMATE $f(x)$

$f(x)$ defines the relationship between dependent and independent variables.

Choice of Model

Choice of model for estimating will depend on whether we want to **predict** or **infer**.

- For Prediction, **accuracy** of predicted function is the most important
- For Inference, **interpretability** of predicted function is most important

For example, **linear regression** is simple to interpret but may not give very accurate predicted values of Y

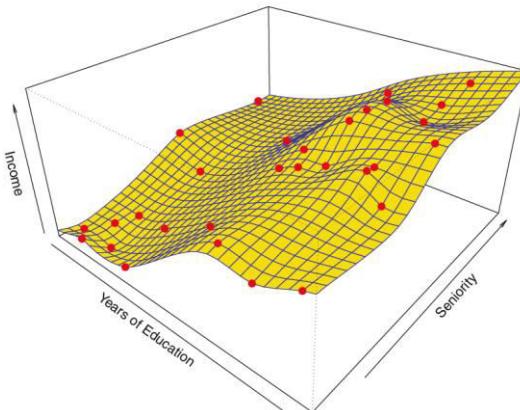
Whereas highly **non-linear models** may be predicting very accurately but the relationship may be very difficult to interpret



HOW TO ESTIMATE $F(x)$

Next, we need to specify the type of **learning method**.

Parametric vs Non parametric



In **Parametric approach**, we assume the functional form of the relationship between predictor and predicted variable

For example, we may assume linear relationship between house price with other variables

$$\text{Price (Y)} = a_0 + a_1*x_1 + a_2*x_2 + a_3*x_3 \dots a_n*x_n$$

Then we will use the training data to estimate the values of $a_0, a_1, a_2, a_3 \dots a_n$

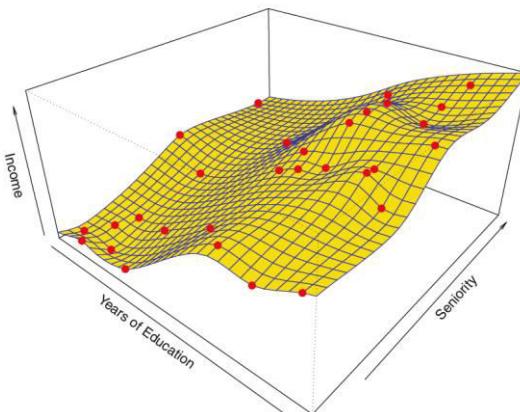
In **non-parametric approach**, we do not assume any functional form for the relationship. Instead f is estimated by getting as close to the training points. For example, in the image shown, for three variables, a three dimensional spleen is created which is as close to the points and has a smooth surface



HOW TO ESTIMATE $F(x)$

Parametric vs Non parametric

Parametric vs Non parametric



Parametric approach

- Usually more interpretable
- May not be as accurate
- Preferable if inference is the reason estimating $f(x)$

Non-parametric approach, w

- Less interpretable
- Potentially more accurate
- Needs large amount of data to train
- Preferable if prediction is the priority



TYPES OF LEARNING

Supervised Vs Unsupervised

Supervised vs Unsupervised learning

Supervised Learning:

- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.
- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

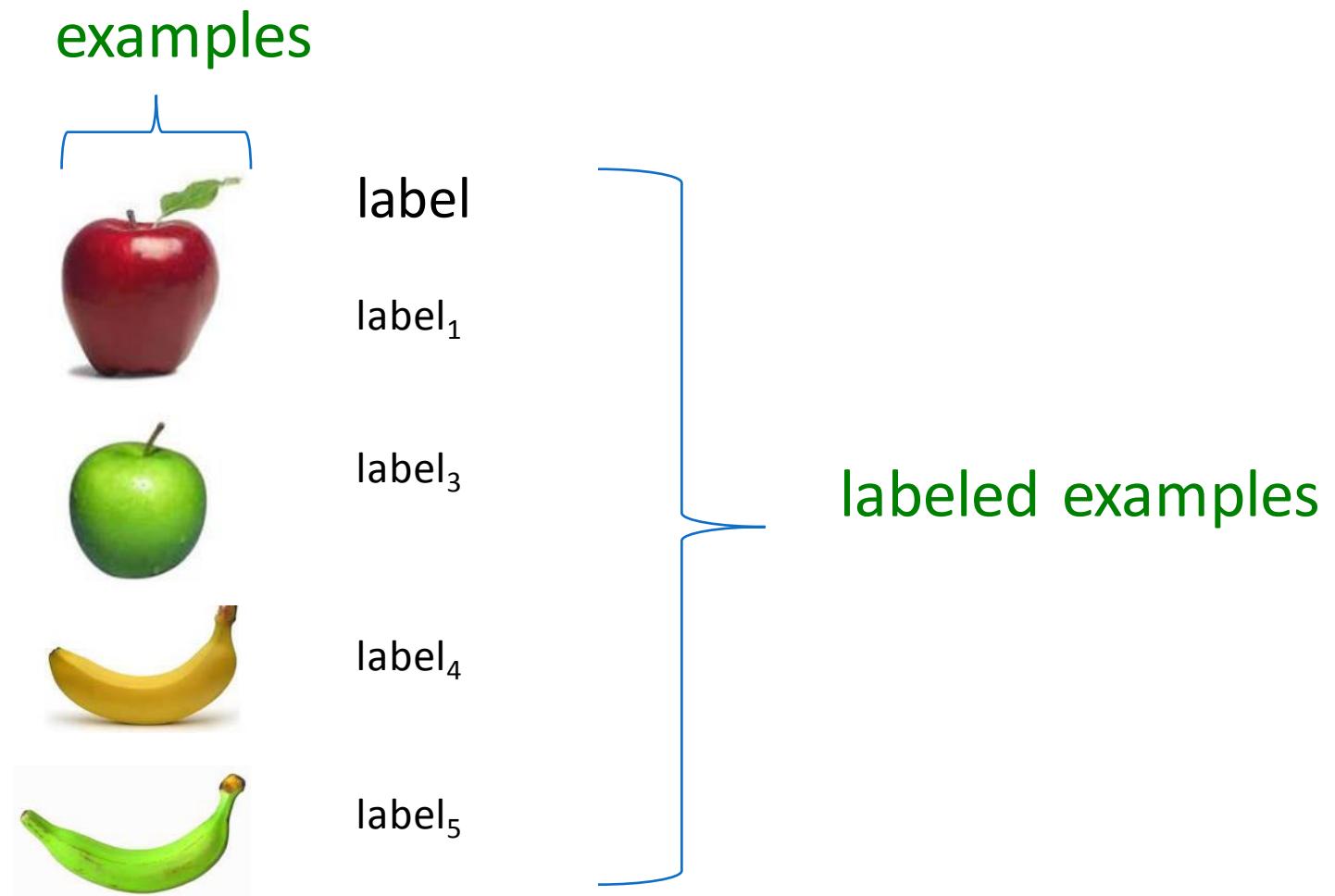
Unsupervised Learning:

- Unsupervised learning is where you only have input data (X) and no corresponding output variables.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.



Supervised Learning: Example

Supervised Learning Example



Supervised Learning: Example

Supervised Learning Example

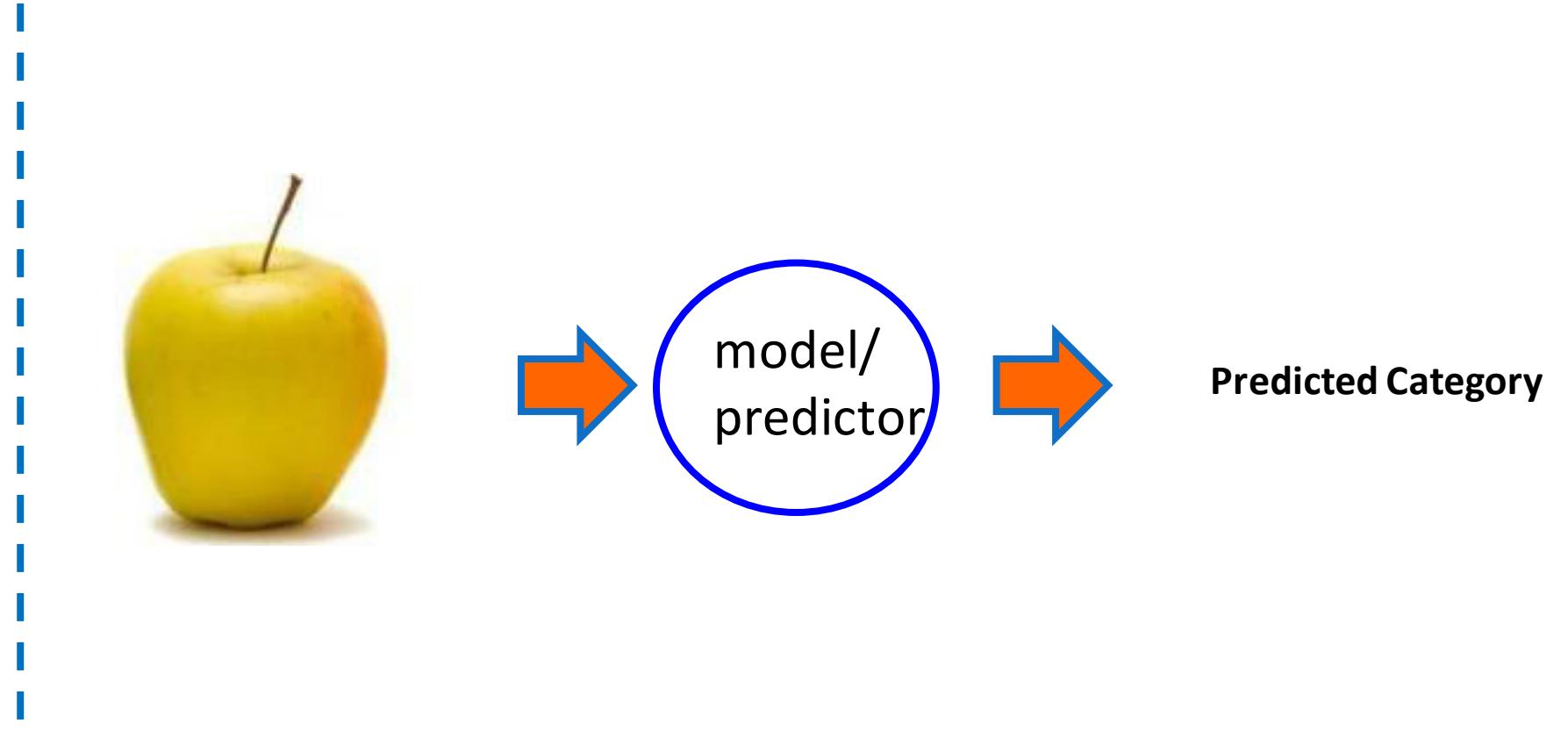
	Category	Weight	
	Apple	100 gm	
	Apple	80 gm	
	Banana	40 gm	
	Banana	60 gm	

model/
predictor



Supervised Learning: classification

**Supervised
Learning
Example
(classification)**



Supervised Learning (classification)

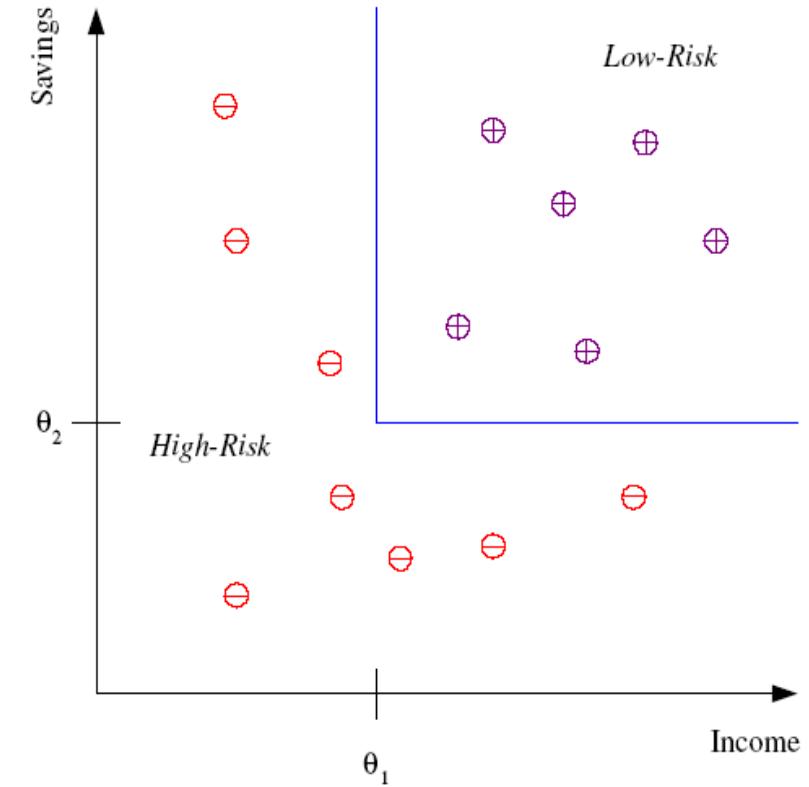
Supervised Learning (classification)

Classification:

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their income and savings
- Model - **Discriminant**
IF $income > \theta_1$ AND $savings > \theta_2$
THEN **low-risk** ELSE **high-risk**

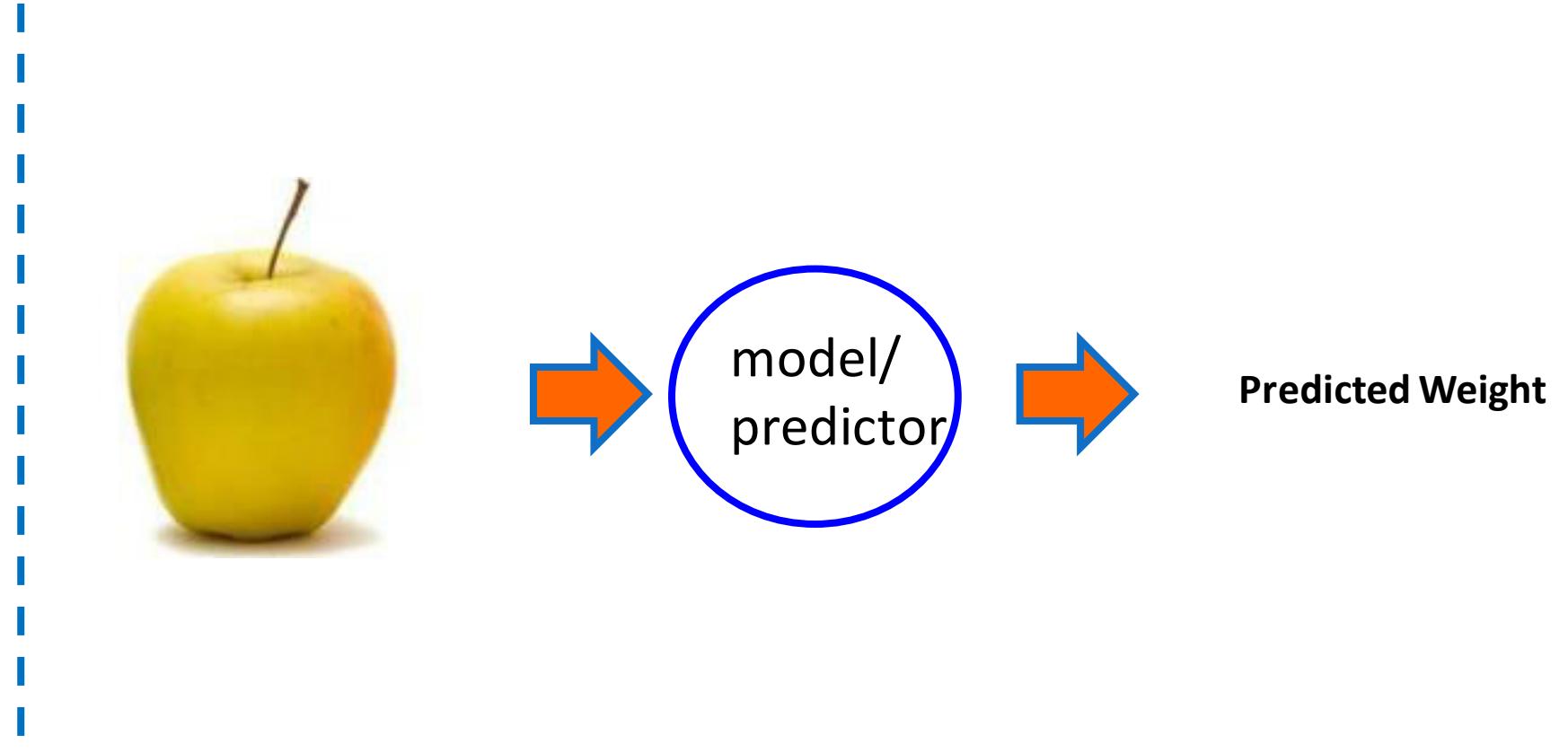
Applications :

- Pattern recognition
- Face recognition
- Character recognition
- Medical diagnosis
- Web Advertising



Supervised Learning: Regression

Supervised
Learning
Example
(Regression)



Supervised Learning (Regression)

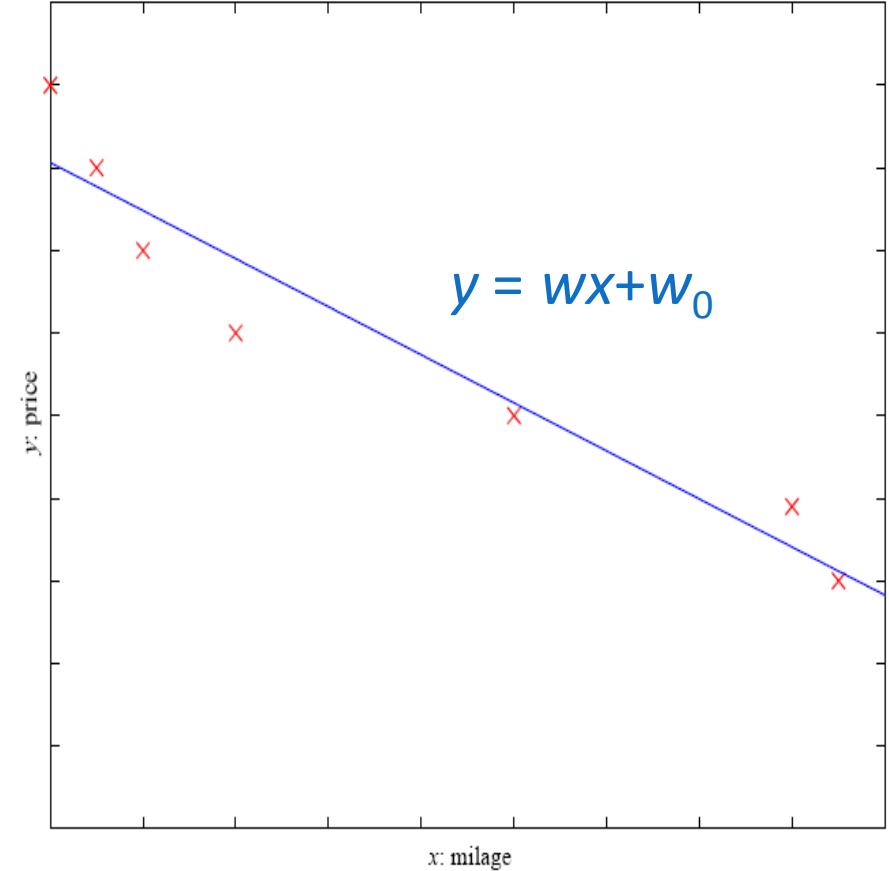
Supervised Learning (Regression)

Regression:

- Example: Price of a used car
- x : car attributes
(e.g. mileage)
- y : price
- Model – **Linear Regression**
 $y = wx + w_0$

Applications :

- Weather forecast
- Sales forecasting
- Advertising budget allocation
- Product pricing



Supervised Learning Algorithms

Supervised
Learning
Algorithms



A circular graphic containing various supervised learning algorithm names, including ANN, Neural Networks, Naive Bayes, Ensemble, Discriminant, Regression, SVM, Analysis, Random Forest, CHAID, Logistic, and Tree.

ANN
Neural Networks
Naive Bayes
Ensemble
Discriminant
Regression
SVM
Analysis
Random Forest
CHAID
Logistic
Tree

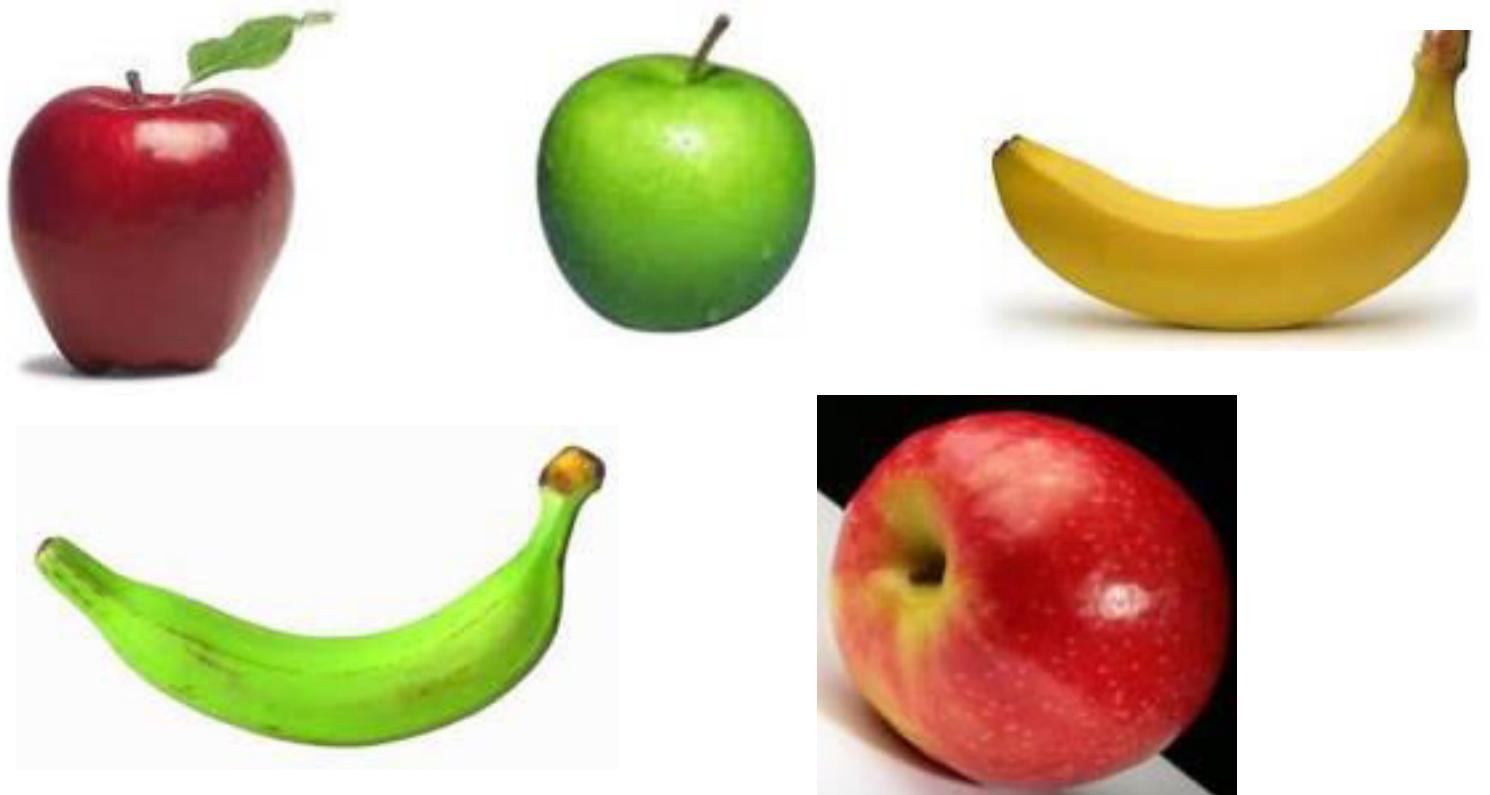


Start-Tech
ACADEMY

Start-Tech Academy

Unsupervised Learning: Example

**Unsupervised
Learning
Example**



Unsupervised learning: given data, i.e. examples, but no labels



**Start-Tech
ACADEMY**

Start-Tech Academy

Unsupervised Learning Algorithms

Unsupervised Learning Algorithms

Unsupervised Learning - Algorithms:

- Clustering
 - K means
 - Hierarchical clustering
- Hidden Markov Models (HMM)
- Dimension Reduction (Factor Analysis, PCA)
- Feature Extraction methods
- Self-organizing Maps (Neutral Nets)



Machine Learning Model

Steps

Steps in Building ML Model

1. Problem formulation
2. Data Tidying
3. Pre-Processing
4. Train-Test Split
5. Model Building
6. Validation and Model Accuracy
7. Prediction



Machine Learning Model

1. Problem formulation

- Convert your business problem into a Statistical problem
- Clearly define the dependent and independent variable
- Identify whether you want to predict or infer



Machine Learning Model

2. Data Tidying

- Transform collected data into a useable data table format
- Example

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2

Machine Learning Model

3. Data Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction



Machine Learning Model

4. Test - Train Split

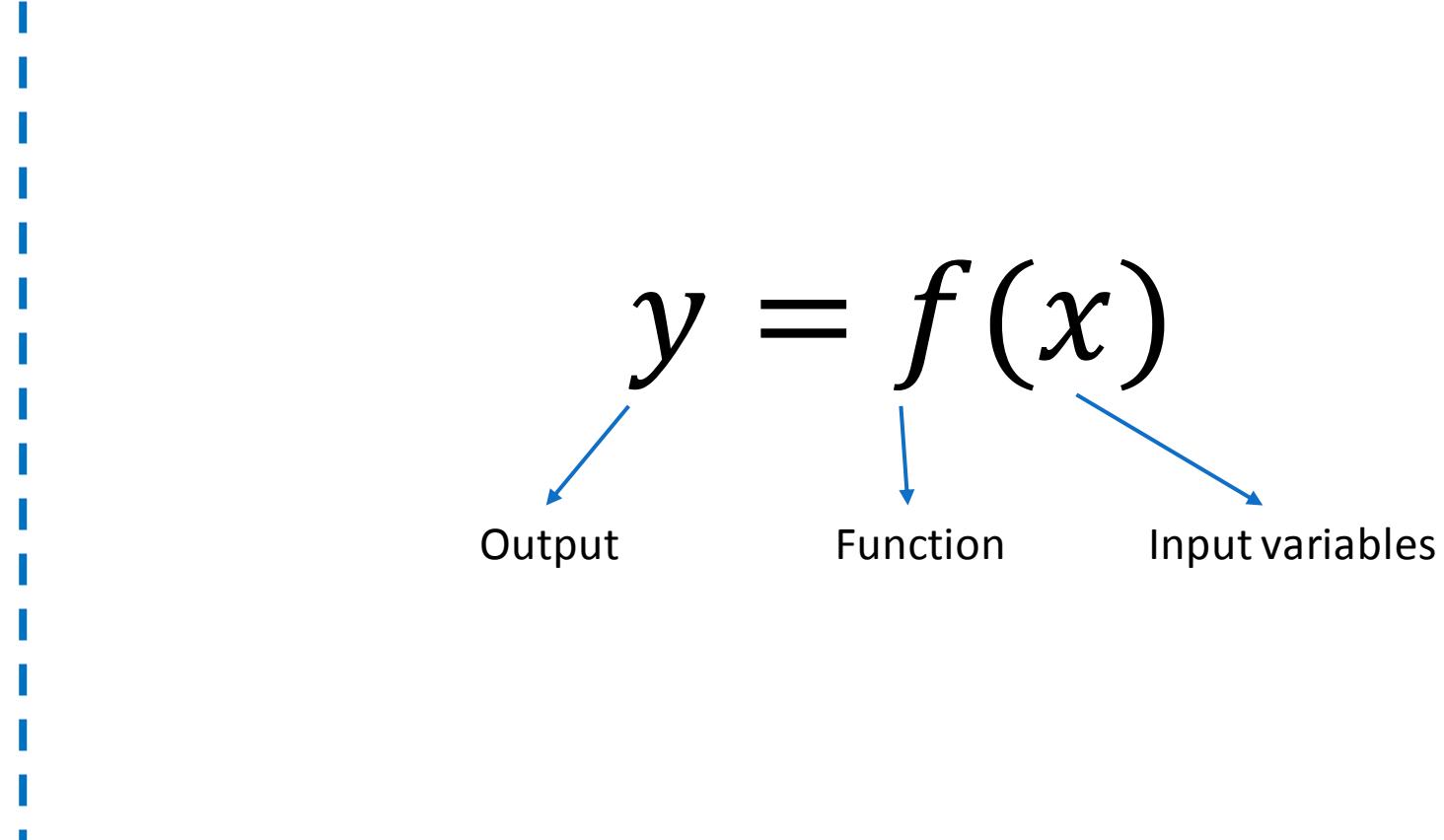
Training data is the information used to train an algorithm. The training data includes both input data and the corresponding expected output. Based on this data, the algorithm can learn the relationship between input and output variables.

Testing data includes only input data, not the corresponding expected output. The testing data is used to assess the accuracy of model created or the predictor function created using the training data.

- There should not be any overlap between the two.
- Usually, 70-80% of the available data is used as training data and 20-30% as testing data

Machine Learning Model

5. Model Training



Machine Learning Model

6. Performance Metrics and Validation

In Sample error

- Error resulted from applying your prediction algorithm to the dataset you built it with

Out of Sample error

- Error resulted from applying your prediction algorithm to a new data set



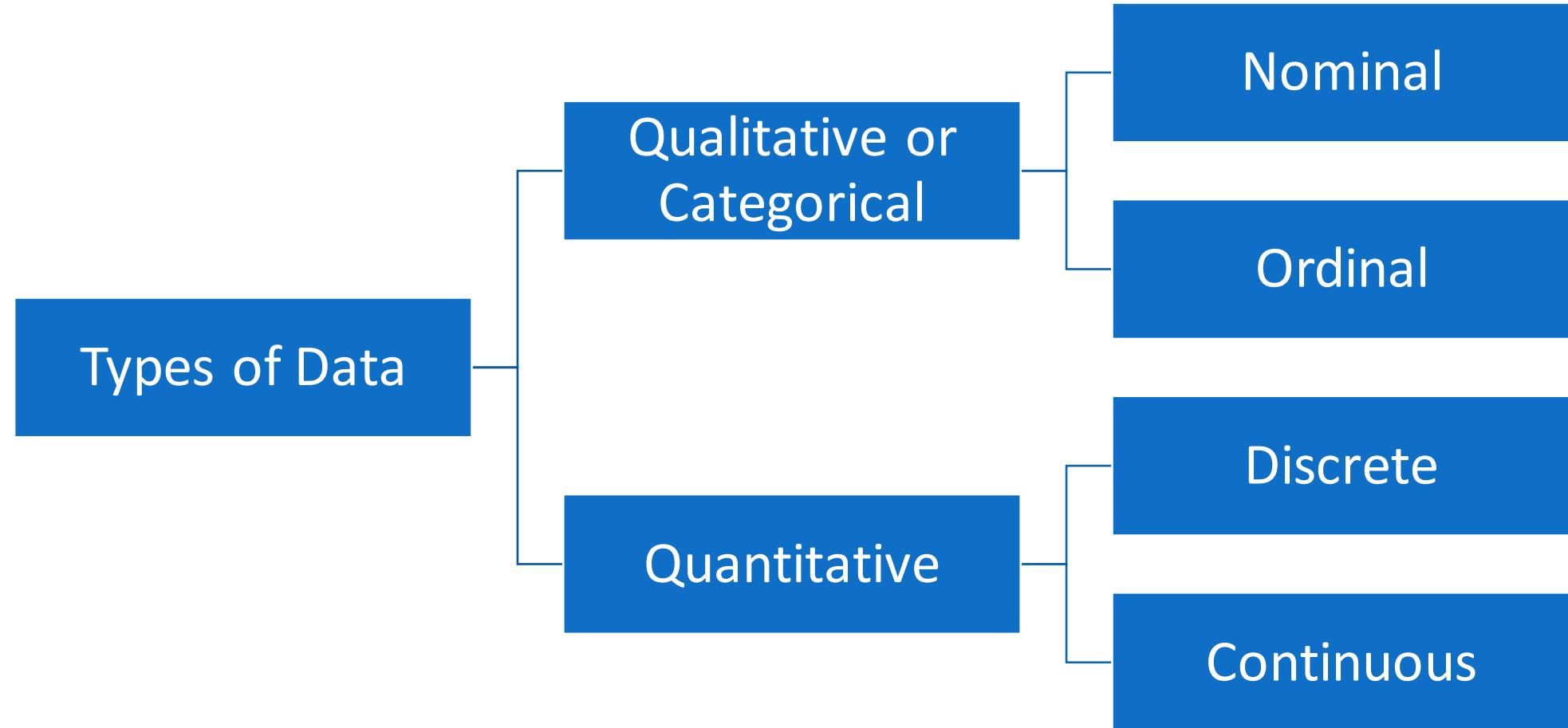
Machine Learning Model

7. Prediction

- Setup a pipeline to use your model in real life scenario
- Improve by monitoring your model over time
- Try to automate



TYPES OF DATA



QUALITATIVE DATA

A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a qualitative or categorical variable. The data collected on such a variable are called **qualitative** data.

Types

There are **two types** of Qualitative variables:

1. **Nominal Variables** The values are not ordered. Example: Nationality, Gender etc.
2. **Ordinal Variables** - The values are ordered or ranked. Example: Satisfaction score (Not satisfied, Satisfied, Delighted), Spiciness of food (Less spicy, mild & Hot 



QUANTITATIVE DATA

A Variable that can be measured numerically is called a quantitative variable.
The data collected on a quantitative variable are called quantitative data.

TYPES

There are **two types** of Quantitative variables:

1. **Discrete Variables** - A variable whose values are countable is called a discrete variable. In other words, a discrete variable can assume only certain values with no intermediate values. Example: Number of heads in 10 tosses etc.
2. **Continuous Variables** - A variable that can assume any numerical value over a certain interval or intervals is called a continuous variable. Example: Height of person etc.



TYPES OF STATISTICS

Types of Statistics

Descriptive

Inferential

DESCRIPTIVE STATISTICS

Descriptive statistics consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.

Tools

- Measures of Central Tendency
 - Mean
 - Median
 - Mode
- Measures of Dispersion
 - Range
 - Standard Deviation
- Frequency Distributions
- Histograms



INFERRENTIAL STATISTICS

Inferential statistics consists of methods that use sample results to help make decisions or predictions about a population.

Tools

- Hypothesis testing
- ANOVA
- Chi-Squared Tests
- Regression



Frequency distribution

A **frequency distribution for qualitative data** lists all categories and the number of elements that belong to each of the categories.

EXAMPLE

Branch/ Specialization	Number of students
Computer Science	100
Mechanical Engineering	80
Electrical Engineering	130
Biotechnology	60
Mathematics	50
Total	420

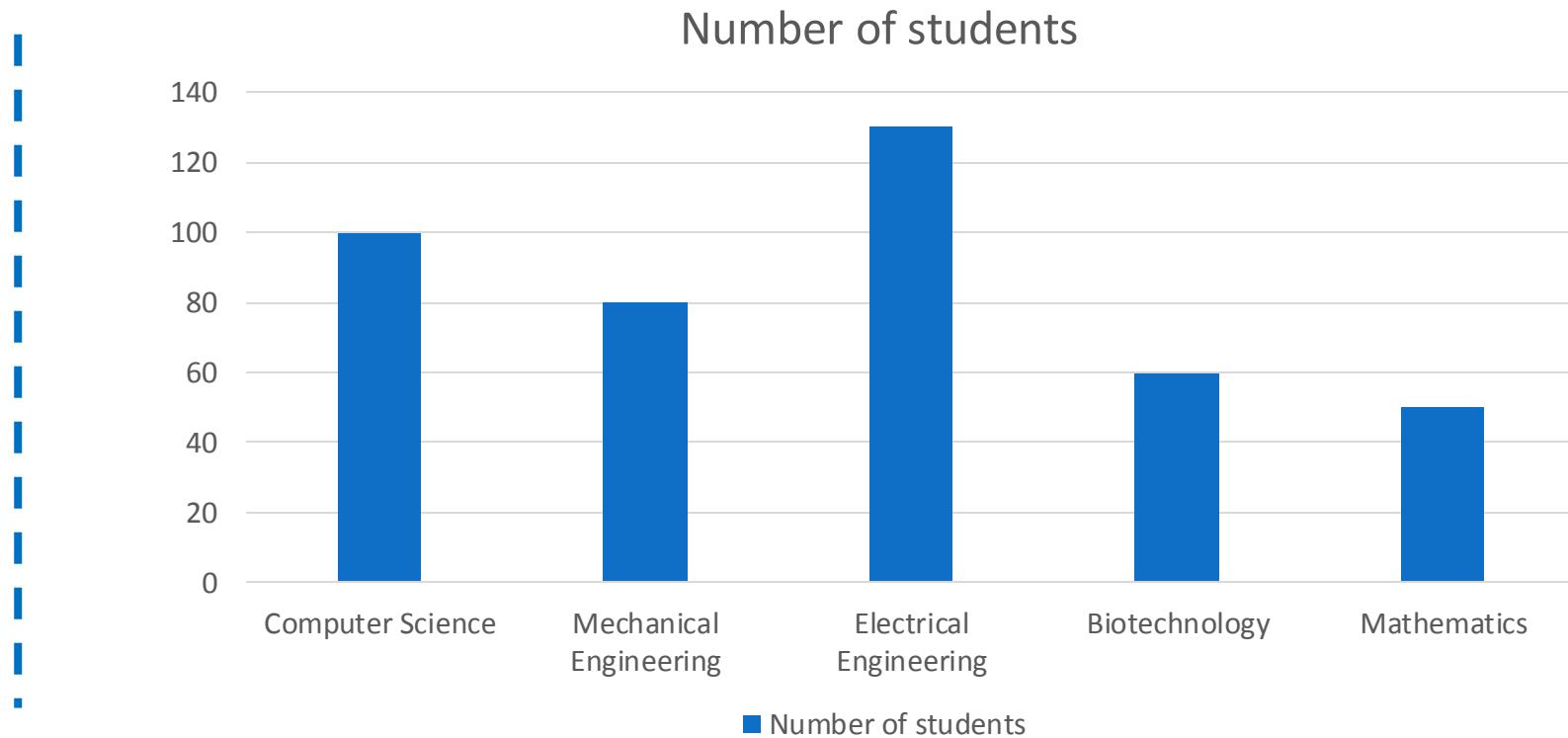
$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

Therefore, relative frequency of Biotechnology is equal to $60/420 = 0.142 \sim 14.2\%$

BAR GRAPH

A **BAR graph** is a graph made of bars whose heights represent the frequencies of respective categories is called a bar graph

TYPES

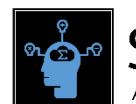


Frequency distribution

A **frequency distribution** for **quantitative** data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

EXAMPLE

Science Marks	Number of students
0-35	5
36-55	21
56-70	12
71-85	18
86-100	9
Total	65



Frequency distribution

A **frequency distribution** for **quantitative** data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

Process

1. Select the number of classes, usually between 5 and 20.
2. Calculate the class width.
Class width \approx (maximum data value) - (minimum data value) / number of classes
Round this result to get a convenient number.
3. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.
4. Using the first lower class limit and the class width, list the other lower class limits.
5. List the lower class limits in a vertical column and then determine and enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.



Frequency distribution

A **frequency distribution** for **quantitative** data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

For the list of numbers below, create the frequency distribution

10	14	26	25	30	34	14	33	33
13	21	25	29	28	7	31	31	30
25	33	31	13	28	33			

1. Select the number of classes, we will select 5.

2. Calculate the class width

$$\text{Class width} \approx (34) - (7) / 5 = 27/5 = 5.4 \rightarrow 5$$

EXAMPLE

Groups	Tally	Frequency
7 – 12		2
13 – 18		4
19 – 24		1
25 – 30	\	9
31 – 36	\	8

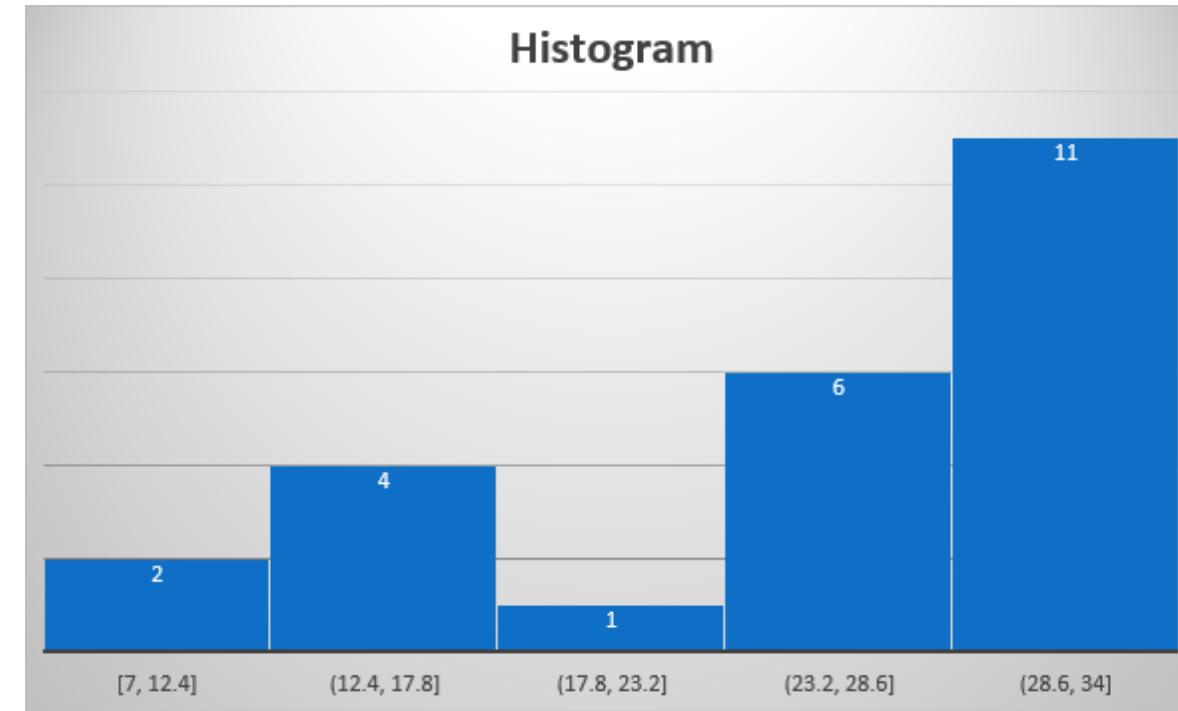


Histograms

Histogram is a graph consisting of bars of equal width drawn adjacent to each other.

The **horizontal scale** represents classes of quantitative data values and the **vertical scale** represents frequencies.

Example



The heights of the bars correspond to the frequency values.



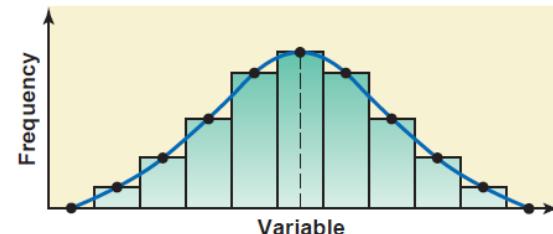
Histogram Shapes

Histogram is a graph consisting of bars of equal width drawn adjacent to each other.

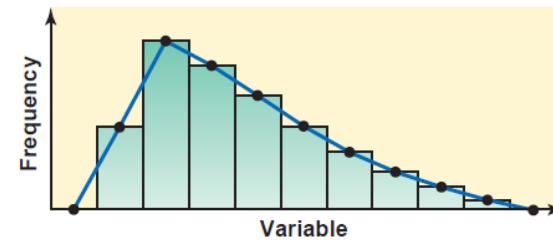
The **horizontal scale** represents classes of quantitative data values and the **vertical scale** represents frequencies.

Example

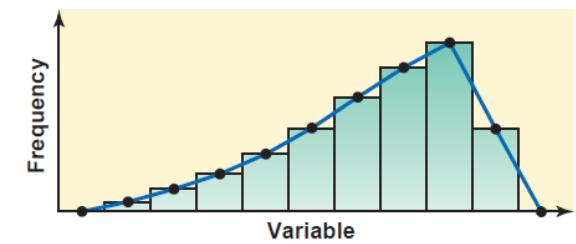
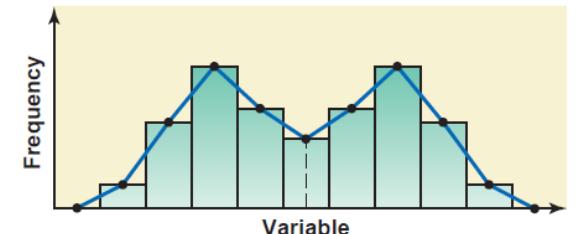
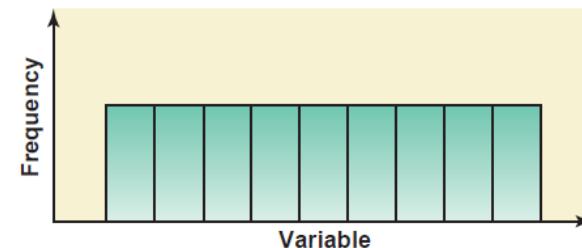
1. Symmetric



2. Skewed (Not symmetric)



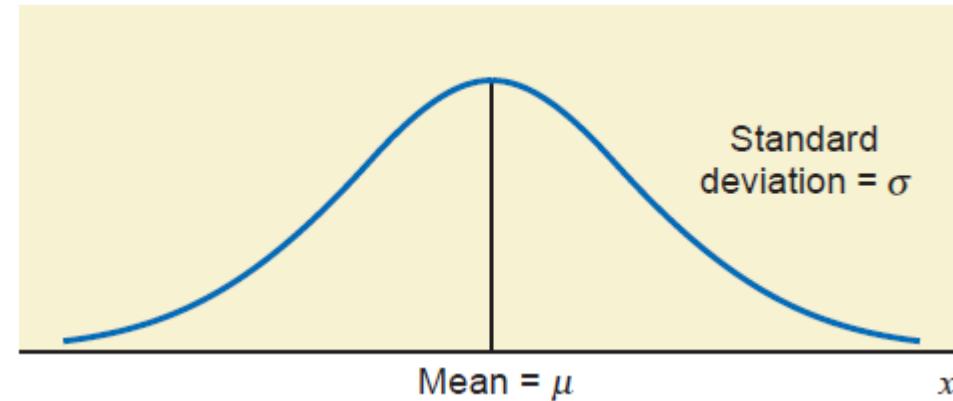
3. Uniform or rectangular



Normal Distribution

If a continuous random variable has a distribution with a graph that is symmetric and bell-shaped, and it can be described by the equation given below, we say that it has a normal distribution

$$y = \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}}$$



Properties

1. The total area under the curve is 1.0.
2. The curve is symmetric about the mean.
3. The two tails of the curve extend indefinitely.



Frequency distribution

A **frequency distribution for qualitative data** lists all categories and the number of elements that belong to each of the categories.

EXAMPLE

Branch/ Specialization	Number of students
Computer Science	100
Mechanical Engineering	80
Electrical Engineering	130
Biotechnology	60
Mathematics	50
Total	420

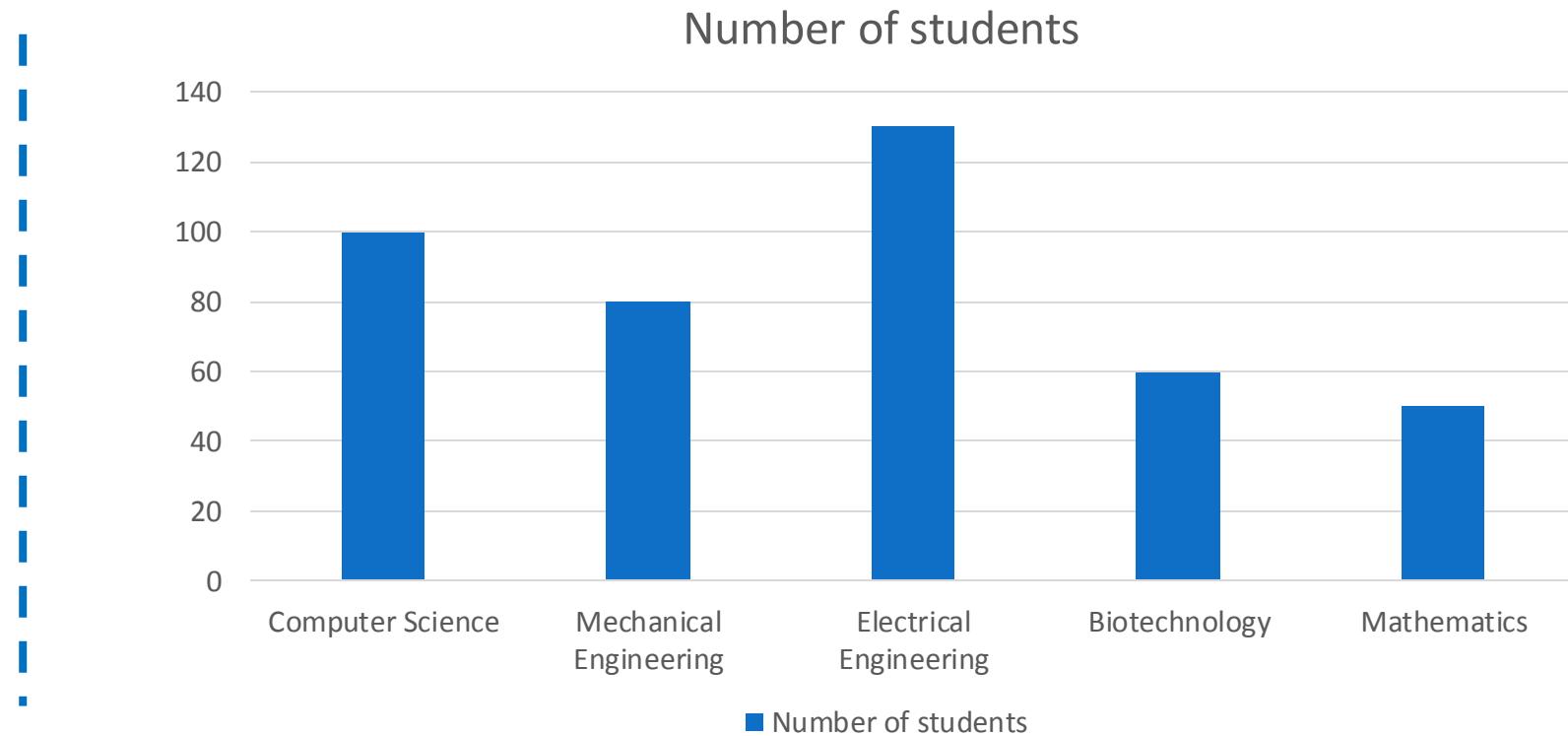
$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

Therefore, relative frequency of Biotechnology is equal to $60/420 = 0.142 \sim 14.2\%$

BAR GRAPH

A **BAR graph** is a graph made of bars whose heights represent the frequencies of respective categories is called a bar graph

TYPES

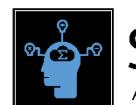


Frequency distribution

A **frequency distribution** for **quantitative** data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

EXAMPLE

Science Marks	Number of students
0-35	5
36-55	21
56-70	12
71-85	18
86-100	9
Total	65



Frequency distribution

A **frequency distribution** for **quantitative** data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

Process

1. Select the number of classes, usually between 5 and 20.
2. Calculate the class width.
Class width \approx (maximum data value) - (minimum data value) / number of classes
Round this result to get a convenient number.
3. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.
4. Using the first lower class limit and the class width, list the other lower class limits.
5. List the lower class limits in a vertical column and then determine and enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.



Frequency distribution

A **frequency distribution** for **quantitative** data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

For the list of numbers below, create the frequency distribution

10	14	26	25	30	34	14	33	33
13	21	25	29	28	7	31	31	30
25	33	31	13	28	33			

1. Select the number of classes, we will select 5.

2. Calculate the class width

$$\text{Class width} \approx (34) - (7) / 5 = 27/5 = 5.4 \rightarrow 5$$

EXAMPLE

Groups	Tally	Frequency
7 – 12		2
13 – 18		4
19 – 24		1
25 – 30	\	9
31 – 36	\	8

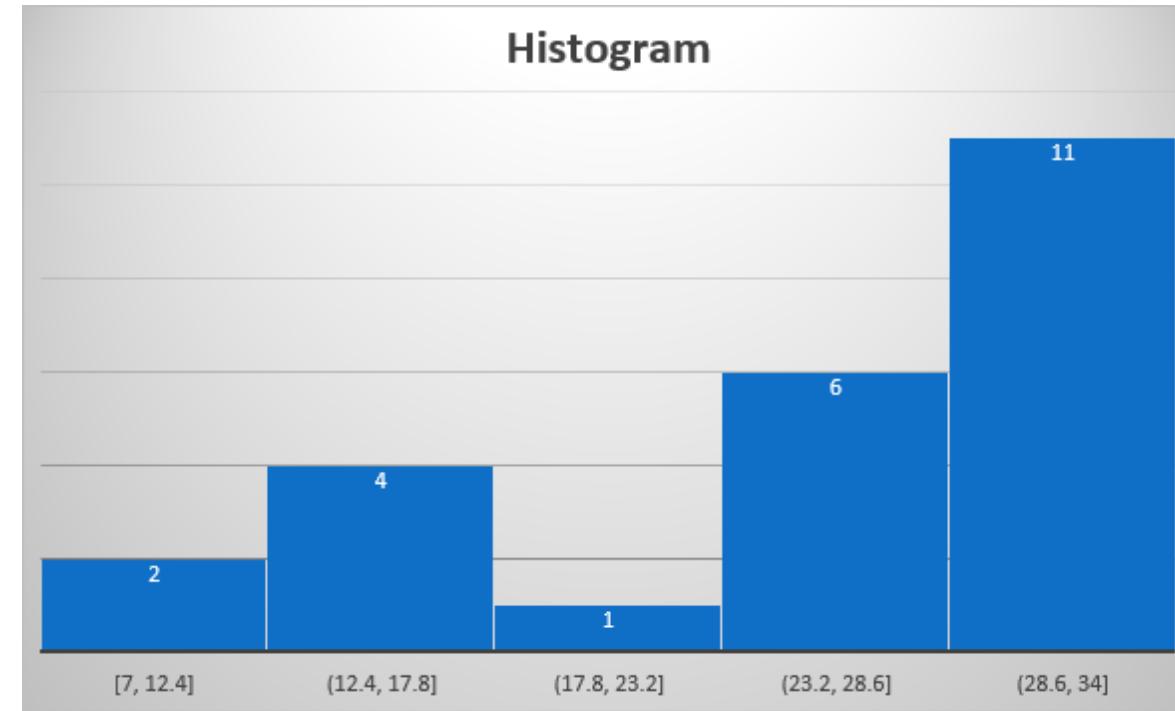


Histograms

Histogram is a graph consisting of bars of equal width drawn adjacent to each other.

The **horizontal scale** represents classes of quantitative data values and the **vertical scale** represents frequencies.

Example



The heights of the bars correspond to the frequency values.



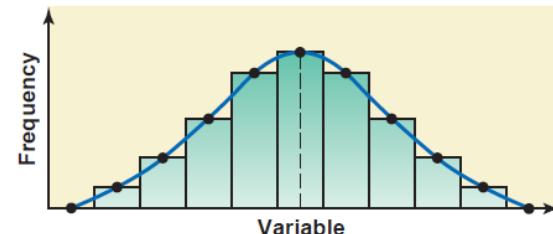
Histogram Shapes

Histogram is a graph consisting of bars of equal width drawn adjacent to each other.

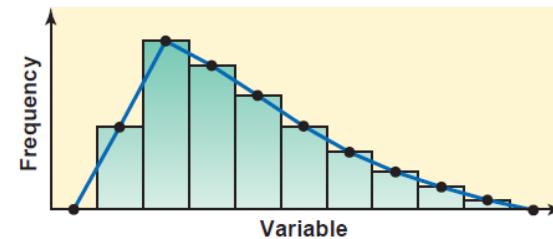
The **horizontal scale** represents classes of quantitative data values and the **vertical scale** represents frequencies.

Example

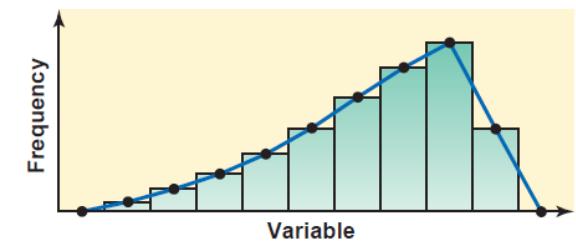
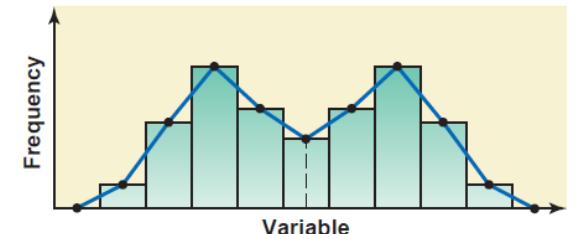
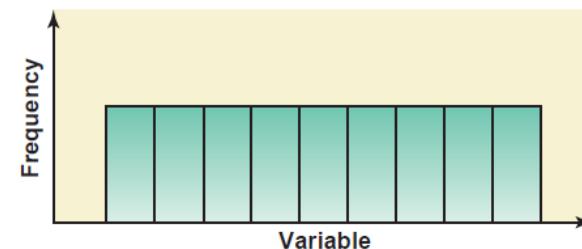
1. Symmetric



2. Skewed (Not symmetric)



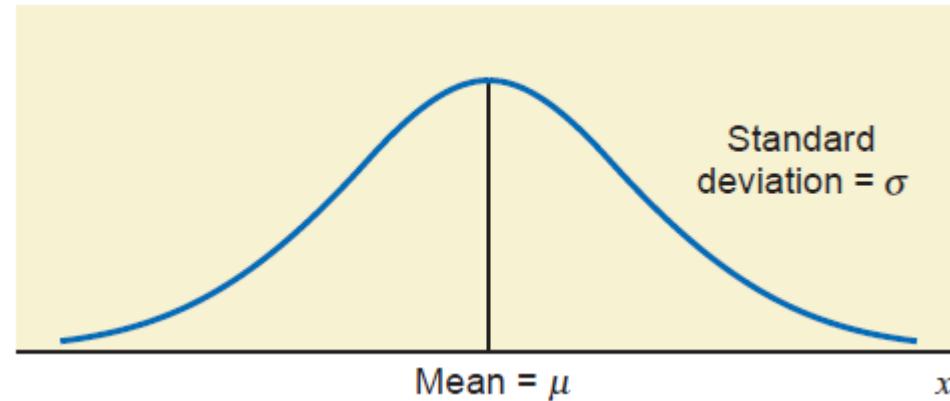
3. Uniform or rectangular



Normal Distribution

If a continuous random variable has a distribution with a graph that is symmetric and bell-shaped, and it can be described by the equation given below, we say that it has a normal distribution

$$y = \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sigma\sqrt{2\pi}}$$



Properties

1. The total area under the curve is 1.0.
2. The curve is symmetric about the mean.
3. The two tails of the curve extend indefinitely.



Measures of Center

A measure of center is a value at the center or middle of a data set.

TYPES

- Mean
- Median
- Mode
- Mid-range



MEAN

The **Mean** is obtained by dividing the sum of all values by the number of values in the data set

Formula

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

Mean for population data: $\mu = \frac{\sum x}{N}$

Mean for sample data: $\bar{x} = \frac{\sum x}{n}$

Where Σx is the sum of all values, N is the population size, n is the sample size, μ is the population mean, and \bar{x} is the sample mean.



MEAN

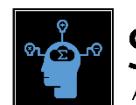
The **Mean** is obtained by dividing the sum of all values by the number of values in the data set

Example

What is the average or the mean height of the following students:

Student	Height
A	168
B	182
C	170
D	155
E	159

$$\text{Mean} = (168+182+170+155+159)/ 5 = 824/5 = 164.8$$



MEDIAN

The **Median** is the value of the middle term in a data set that has been ranked in increasing order.

Steps

1. Rank the data set in increasing order.
2. Find the middle term. The value of this term is the median.

Note that if the number of observations in a data set is odd, then the median is given by the value of the middle term in the ranked data. However, if the number of observations is even, then the median is given by the average of the values of the two middle terms.



MEDIAN

The **Median** is the value of the middle term in a data set that has been ranked in increasing order.

Example

What is the median height of the following students:

Student	Height
A	168
B	182
C	170
D	155
E	159

1. Order in ascending order -> 155,159,168,179,182
2. Select the middle value -> 168



MODE

The **Mode** of a data set is the value that occurs with the greatest frequency.

Note

- When **two data values** occur with the same greatest frequency, each one is a mode and the data set is **bimodal**
- When **more than two data values** occur with the same greatest frequency, each is a mode and the data set is said to be **multimodal**
- When **no data value** is repeated, we say that there is **no mode**



MODE

The **Mode** of a data set is the value that occurs with the greatest frequency.

Example

What is the Mode height of the following students:

Student	Height
A	168
B	182
C	168
D	155
E	159

168 appears two times while rest of the values appear only once.
Therefore, 168 is the Mode height

MIDRANGE

The **Midrange** is the value midway between the maximum and minimum values in the original data set.

Formula

$$\text{midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$$



MIDRANGE

The **Midrange** is the value midway between the maximum and minimum values in the original data set.

Example

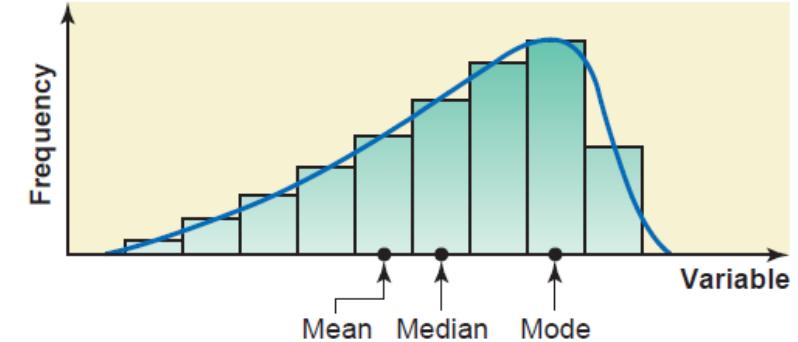
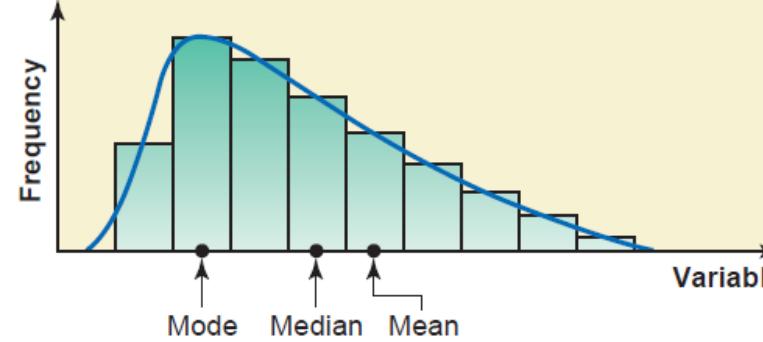
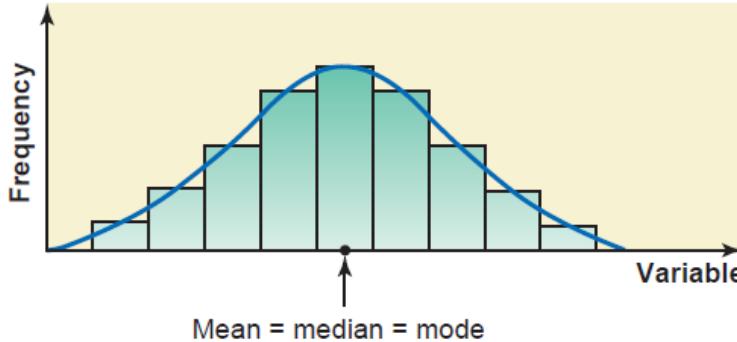
What is the Midrange of the following students:

Student	Height
A	168
B	182
C	168
D	155
E	159

$$\text{Midrange} = (155 + 182)/2 = 168.5$$



COMPARISON OF CENTERS



- For a symmetric distribution curve with one peak, the values of the mean, median, and mode are identical
- For a distribution curve skewed to the right, the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two.
- Mean is not always the best measure of central tendency because it is heavily influenced by outliers. Median is preferred over mean for this reason
- One advantage of the mode is that it can be calculated for both kinds of data—quantitative and qualitative—whereas the mean and median can be calculated for only quantitative data.



MEASURES OF DISPERSION

The measures that help us learn about the spread of a data set are called the measures of dispersion.

Steps

There are three measures of dispersion:

- Range
- Standard Deviation
- Variance



RANGE

Range is the difference between the largest and the smallest values in a data set

Example

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

For the ages of people attending a party below, what is the range?

10	14	26	25	30	34	14	33	33
13	21	25	29	28	7	31	31	30
25	33	31	13	28	33			

$$\text{Range} = 34 - 7 = 27$$

Range is influenced by outliers, therefore may not be very useful.



VARIANCE & STANDARD DEVIATION

Variance is the average of the squared differences from the Mean.

Standard Deviation is the square root of variance.

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Population standard deviation

Sample standard deviation

Formula

Where Σ is the sum of, N is the population size, n is the sample size, μ is the population mean, and \bar{x} is the sample mean.

Larger sigma value means that the data is more widely spread.



VARIANCE & STANDARD DEVIATION

Variance is the average of the squared differences from the Mean.

Standard Deviation is the square root of variance.

Example

For the ages of people attending a party below, what is the standard deviation?

10	14	26	25	30	34	14	33	33
13	21	25	29	28	7	31	31	30
25	33	31	13	28	33			

$$\text{Variance} = (10-24.875)^2 + (14-24.875)^2 + \dots / 24 = 1624.625 / 24 = 67.69$$

$$\text{Standard Deviation} = \text{Sqr root of variance} = 8.23$$



BUSINESS KNOWLEDGE

Most important part of the creating a model is to have a sound business knowledge of the problem you are trying to solve

Types of Research

1. Primary research

A. Discussions

Ask questions and gather information from the stakeholders

B. Dry Run

If possible take a dry run of problem you are trying to investigate

2. Secondary research

A. Reports and Studies

Read reports and studies by government agencies, trade associations or other businesses in your industry

B. Previous works

Go through any previous work and findings related to your problem



BUSINESS KNOWLEDGE

Examples

Cart Abandonment Analysis

Problem:

High fractions of your online customer are adding product to their cart but not purchasing it

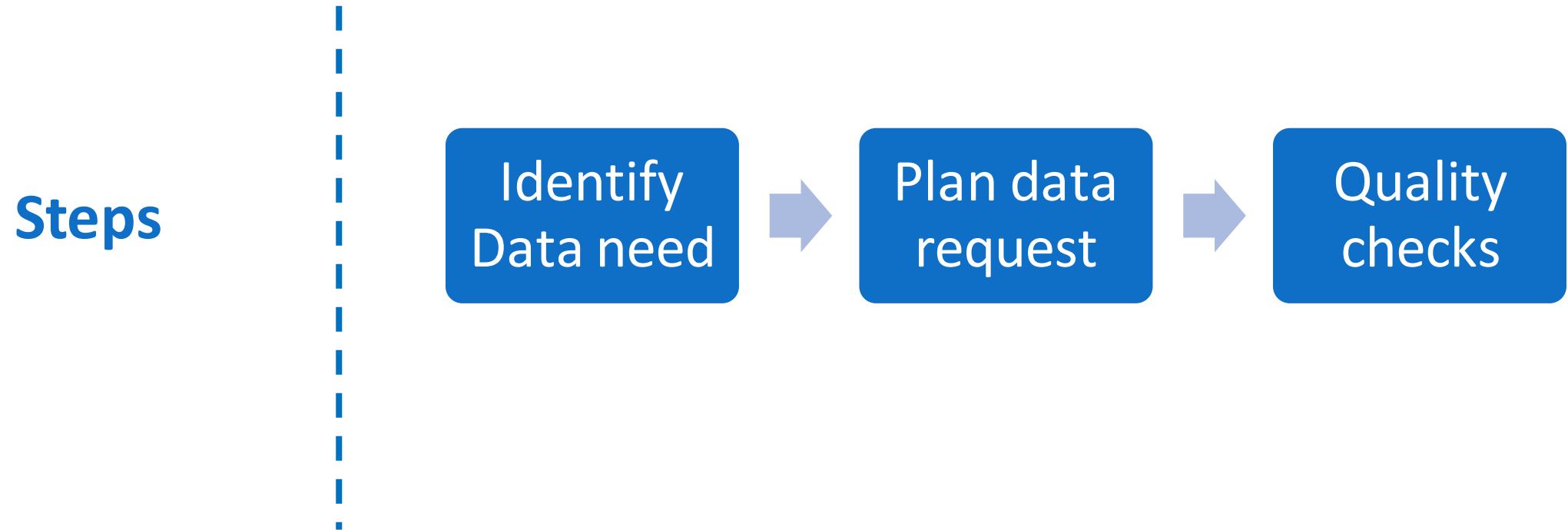
Business Knowledge that will be helpful

1. Discussions with the marketing team
2. Discussions with the product team
3. Dry run of the online purchasing process to understand customer journey
4. Research on industry reports regarding cart abandonment
5. Any previous work in your /other organization regarding cart Abandonment



Data Exploration

Next step should be to use the acquired business knowledge to search for relevant data



Steps



Data Exploration

Data Exploration

Next step should be to use the acquired business knowledge to search for relevant data

1. Internal Data

Data collected by your organization

E.g. Usage data, sales data, promotion data

2. External data

Data acquired from external data sources

E.g. Census Data, External vendor Data, Scrape data



Data Exploration

Examples

Cart Abandonment Analysis

1. Input from the marketing team –

Our 50 % comes from email marketing, 30% from organic search and rest 20% from ad word marketing

-> Gather the source website data for all customers

2. Input from the product team

We have 3 step purchase process – Cart review, Address/personal detail, Payment

-> Gather the Cart Abandonment location for all customer

3. Input from industry reports regarding cart abandonment

Customers tends to put high value item for long duration in their cart

-> Gather the data about total Cart value of all customers

4. Input from dry run

Encountered a survey link for rate website experience

-> Gather survey data for all customers



DATA DICTIONARY

Next step should be to understand the data. You should know variable definition and distribution along with table's unique identifiers and foreign keys

Data Dictionary

A Comprehensive Data Dictionary should include

1. Definition of predictors
2. Unique identifier of each table (or Primary Keys)
3. Foreign keys or matching keys between tables
<https://youtu.be/76Y6Tg1glrQ>
4. Explanation of values in case of Categorical variables



DATA DICTIONARY

Examples

Data Dictionary House Pricing Dataset

The data set contains 506 observations of house prices from different towns. Corresponding to each house price, data of 18 other variables is available on which price is suspected to depend

price	Value of the house
crime_rate	Crime rate in that neighborhood
resid_area	Proportion of residential area in the town
air_qual	Quality of air in that neighborhood
room_num	Average number of rooms in houses of that locality
age	How old is the house construction in years
dist1	Distance from employment hub 1
dist2	Distance from employment hub 2
dist3	Distance from employment hub 3
dist4	Distance from employment hub 4
teachers	Number of teachers per thousand population in the town



DATA DICTIONARY

Examples

Data Dictionary House Pricing Dataset

The data set contains 506 observations of house prices from different towns. Corresponding to each house price, data of 18 other variables is available on which price is suspected to depend

poor_prop	Proportion of poor population in the town
airport	Is there an airport in the city? (Yes/No)
n_hos_beds	Number of hospital beds per 1000 population in the town
n_hot_rooms	Number of hotel rooms per 1000 population in the town
waterbody	What type of natural fresh water source is there in the city (lake/ river/ both/ none)
rainfall	The yearly average rainfall in centimeters
bus_ter	Is there a bus terminal in the city? (Yes/No)
parks	Proportion of land assigned as parks and green areas in the town

UNIVARIATE ANALYSIS

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.

Univariate Analysis

Ways to describe patterns found in univariate data

1. Central tendency
 1. Mean
 2. Mode
 3. Median
2. Dispersion
 1. Range
 2. Variance
 3. maximum, minimum,
 4. Quartiles (including the interquartile range), and
 5. Standard deviation
3. Count / Null count



EDD (EXTENDED DATA DICTIONARY)

Example

	Age	Name	Score
count	12.000000	12	12.000000
unique	NaN	12	NaN
top	NaN	Rahul	NaN
freq	NaN	1	NaN
mean	32.500000	NaN	73.000000
std	9.209679	NaN	17.653225
min	24.000000	NaN	44.000000
25%	25.750000	NaN	64.000000
50%	29.000000	NaN	74.000000
75%	35.250000	NaN	87.500000
max	51.000000	NaN	99.000000

Missing Value Imputation

Real-world data often has missing values. Data can have missing values for a number of reasons such as observations that were not recorded and data corruption.

Missing Value Imputation

Impact

- Handling missing data is important as many machine learning algorithms do not support data with missing values.

Solution

- Remove rows with missing data from your dataset.
- Impute missing values with mean/median values in your dataset.

Note

- Use business knowledge to take separate approach for each variable
- It is advisable to impute instead of remove in case of small sample size or large proportion of observations with missing values



Missing Value Imputation

Methods

1. **Impute with ZERO**
 - Impute missing values with zero
2. **Impute with Median/Mean/Mode**
 - For numerical variables, impute missing values with Mean or Median
 - For categorical variables, impute missing values with Mode
3. **Segment based imputation**
 - Identify relevant segments
 - Calculate mean/median/mode of segments
 - Impute the missing value according to the segments
 - For example, we can say rainfall hardly varies for cities in a particular State
 - In this case, we can impute missing rainfall value of a city with the average of that state



Outlier Treatment

Outlier is a commonly used terminology by analysts and data scientists, Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Outlier Treatment

Reasons

- Data Entry Errors
- Measurement Error
- Sampling error etc

Impact

- It increases the error variance and reduces the power of statistical tests

Solution

- Detect outliers using EDD and visualization methods such as scatter plot, histogram or box plots
- Impute outliers



Outlier Treatment

Example

	Without Outlier	With Outlier
Data	6,6,6,4,4,5,5,5,5,7,7	6,6,6,4,4,5,5,5,5,7,7,300
Mean	5.45	30.0
Median	5	5.5
Mode	5	5
Standard deviation	1.04	85.03
Variance	1.08	7230.10



Outlier Treatment

Methods

1. Capping and Flooring

- Impute all the values above $3 * P99$ and below $0.3 * P1$
- Impute with values $3 * P99$ and $0.3 * P1$
- You can use any multiplier instead of 3, as per your business requirement

2. Exponential smoothing

- Extrapolate curve between P95 to P99 and cap all the values falling outside to the value generated by the curve
- Similarly, extrapolate curve between P5 and P1

3. Sigma Approach

- Identify outliers by capturing all the values falling outside $\mu \mp x\sigma$
- You can use any multiplier as x, as per your business requirement

Seasonality

Seasonality is the presence of variations that occur at specific regular intervals less than a year, such as weekly, monthly, or quarterly.

Seasonality

Reasons

- Weather,
- Vacation,
- Holidays

Examples

- Ice cream sales
- Christmas sales

Solution

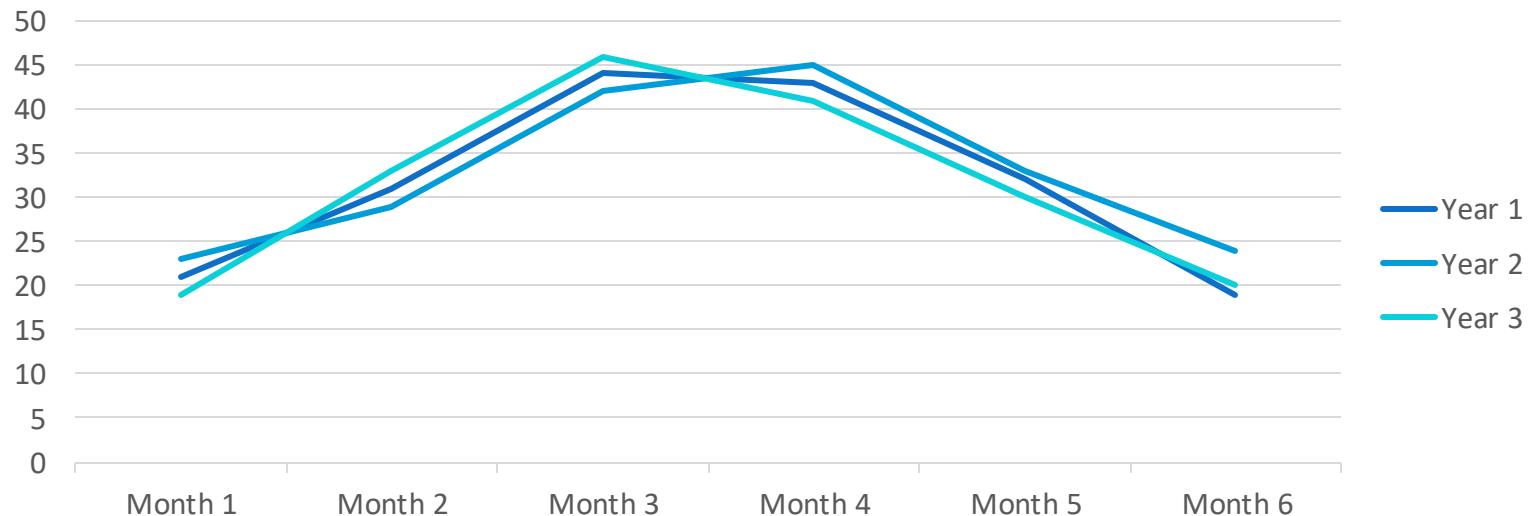
- Calculate multiplication factor for each month as $m = \mu_{Year} \div \mu_{Month}$
- Multiply each observation with its multiplication factor



Seasonality

Example

	Year 1	Year 2	Year 3
Month 1	21	23	19
Month 2	31	29	33
Month 3	44	42	46
Month 4	43	45	41
Month 5	32	33	30
Month 6	19	24	20



Seasonality

Example

	Year 1	Year 2	Year 3	Factor
Month 1	21	23	19	1.521164
Month 2	31	29	33	1.030466
Month 3	44	42	46	0.72601
Month 4	43	45	41	0.742894
Month 5	32	33	30	1.008772
Month 6	19	24	20	1.521164



	Year 1	Year 2	Year 3
Month 1	31.94	34.99	28.9
Month 2	31.94	29.88	34.01
Month 3	31.94	30.49	33.4
Month 4	31.94	33.43	30.46
Month 5	32.28	33.29	30.26
Month 6	28.90	36.51	30.42



Bivariate Analysis

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

Creating new Variables

Scatter Plot

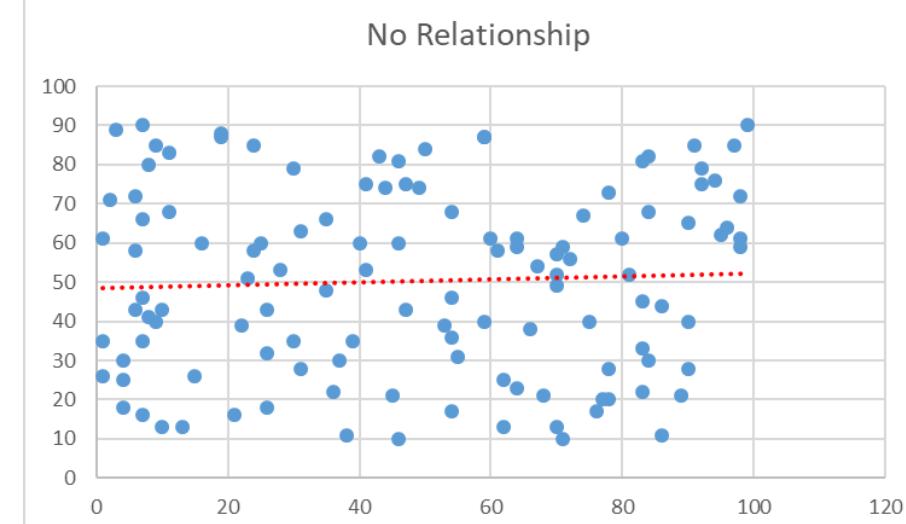
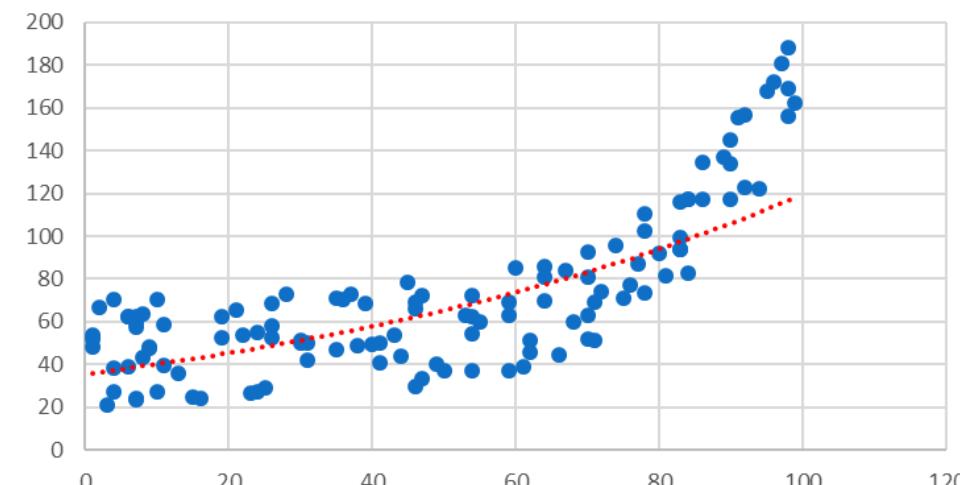
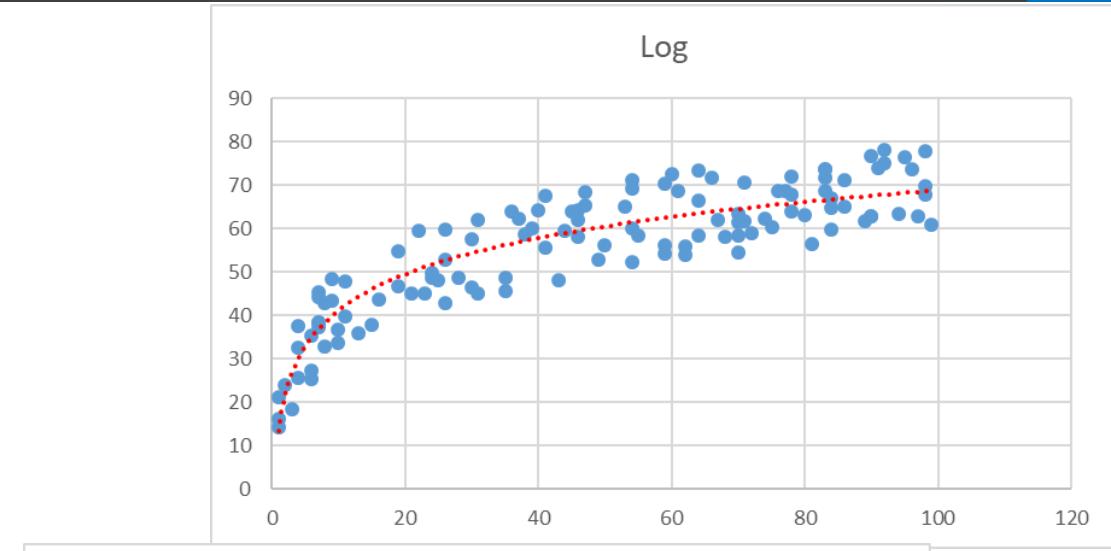
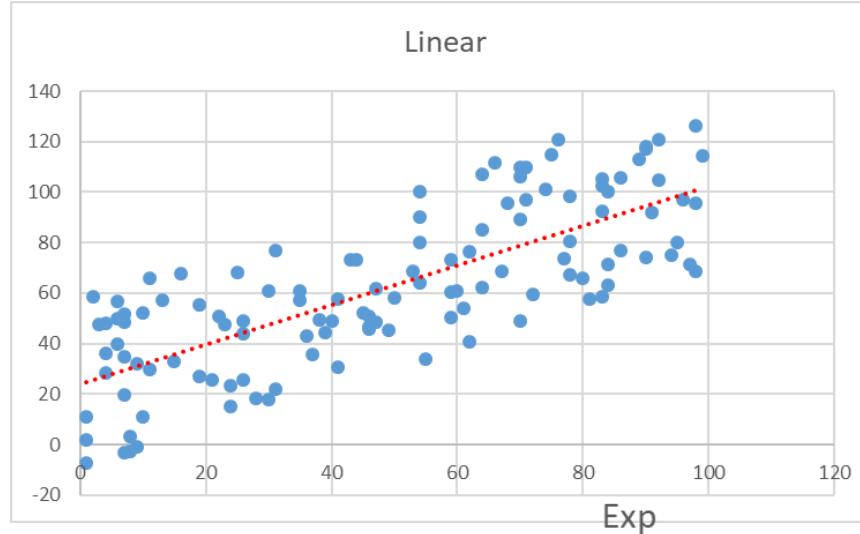
- Scatter indicates the type (linear or non-linear) and strength of the relationship between two variables
- We will use Scatter plot to transform variables

Correlation

- Linear correlation quantifies the strength of a linear relationship between two numerical variables.
- When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity.
- Correlation is used to drop Non Usable variables



Scatter plots



Variable Transformation

Transform your existing variable to extract more information out of them

Creating new Variables

Identify

- Using your business knowledge and bivariate analysis to modify variable

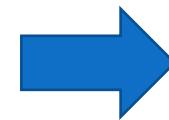
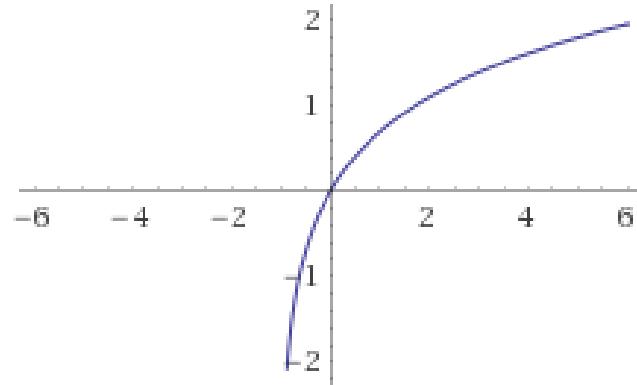
Methods

- Use Mean/Median of variables conveying similar type of information
- Create ratio variable which are more relevant to business
- Transform variable by taking log, exponential, roots etc.



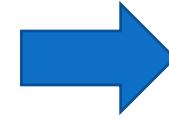
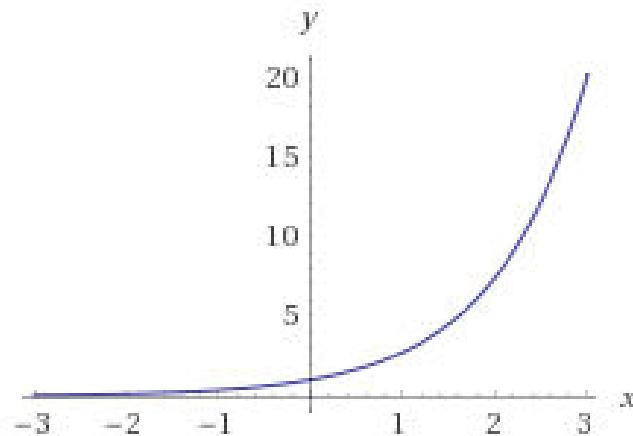
Transformation

If



Take e^x instead of x

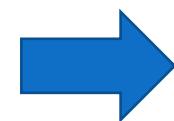
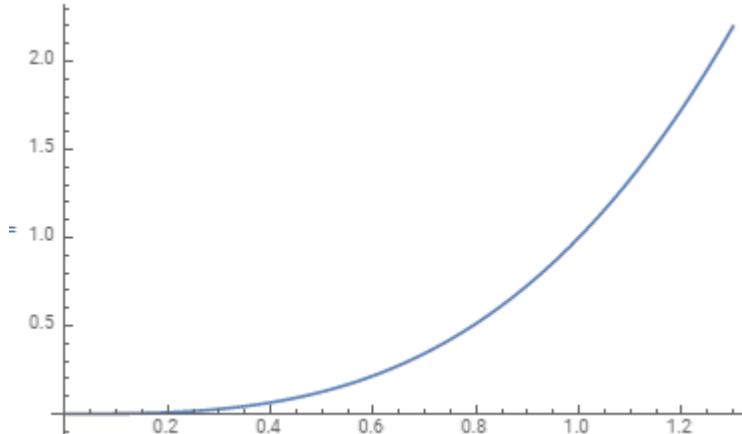
If



Take $\ln(1+x)$ instead of x

Transformation

If



Take \sqrt{x} or $\sqrt[n]{x}$ instead of x

Non Usable Variables

Identify the non usable variables to reduce the dimension of your dataset

Non Usable Variables

- | 1. **Variables with single unique value**
- | 2. **Variables with low fill rate**
- | 3. **Variables with regulatory issue**
- | 4. **Variable with no business sense**



Bivariate Analysis

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

Creating new Variables

Scatter Plot

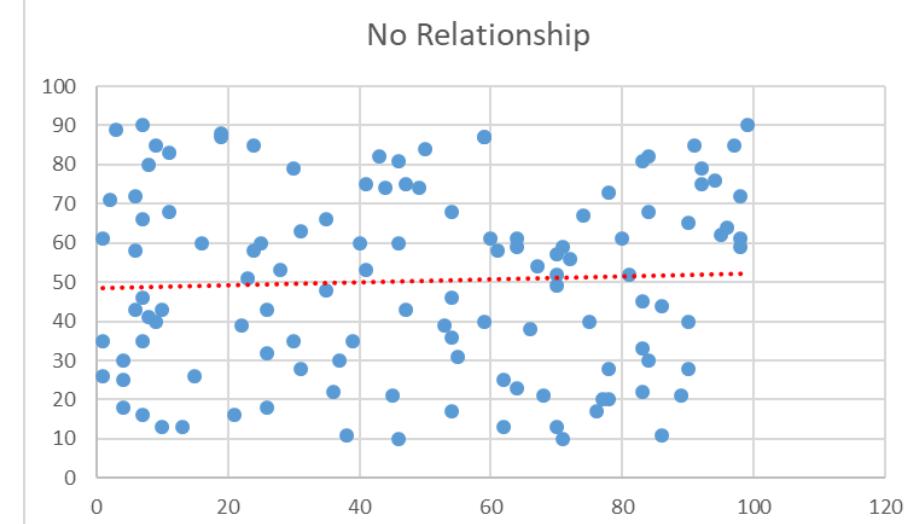
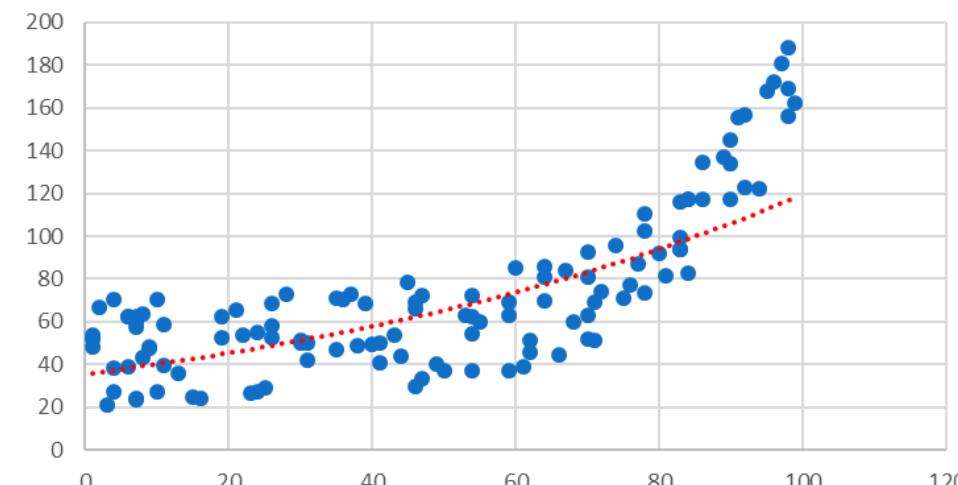
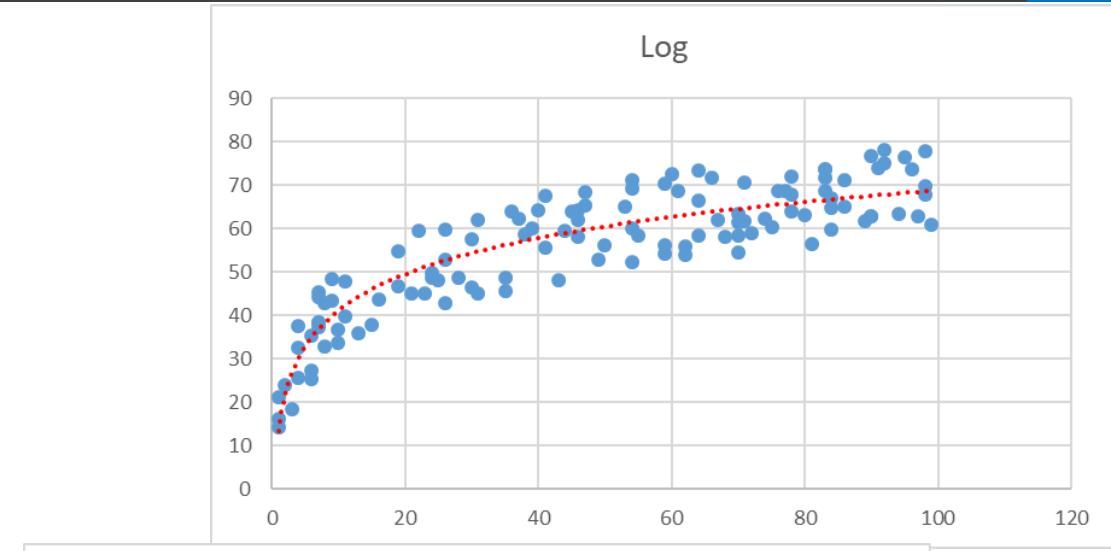
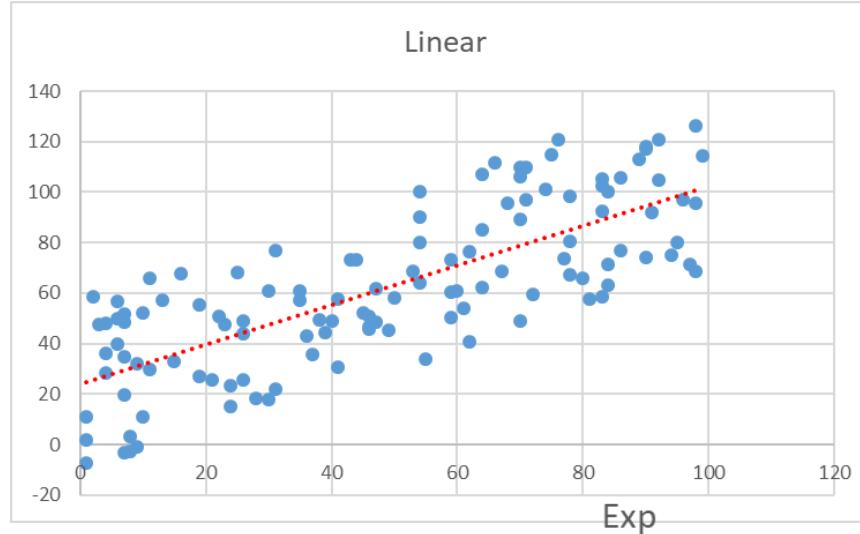
- Scatter indicates the type (linear or non-linear) and strength of the relationship between two variables
- We will use Scatter plot to transform variables

Correlation

- Linear correlation quantifies the strength of a linear relationship between two numerical variables.
- When there is no correlation between two variables, there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity.
- Correlation is used to drop Non Usable variables



Scatter plots



Variable Transformation

Transform your existing variable to extract more information out of them

Creating new Variables

Identify

- Using your business knowledge and bivariate analysis to modify variable

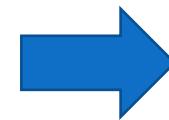
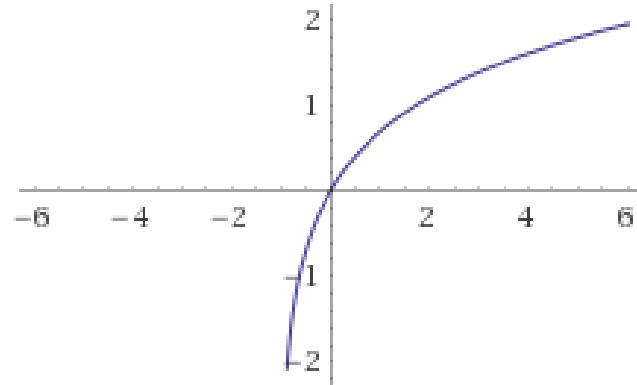
Methods

- Use Mean/Median of variables conveying similar type of information
- Create ratio variable which are more relevant to business
- Transform variable by taking log, exponential, roots etc.



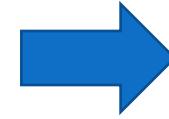
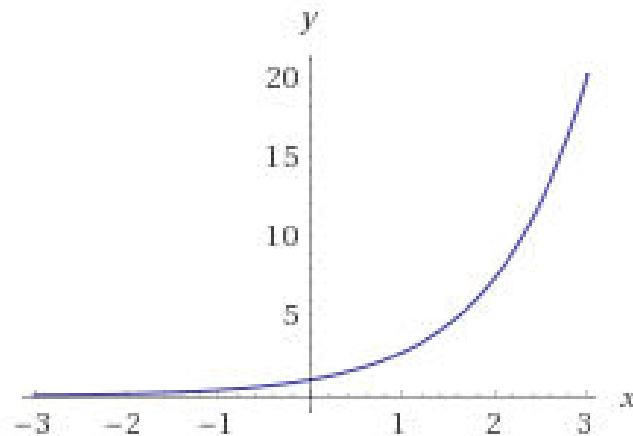
Transformation

If



Take e^x instead of x

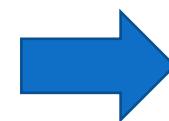
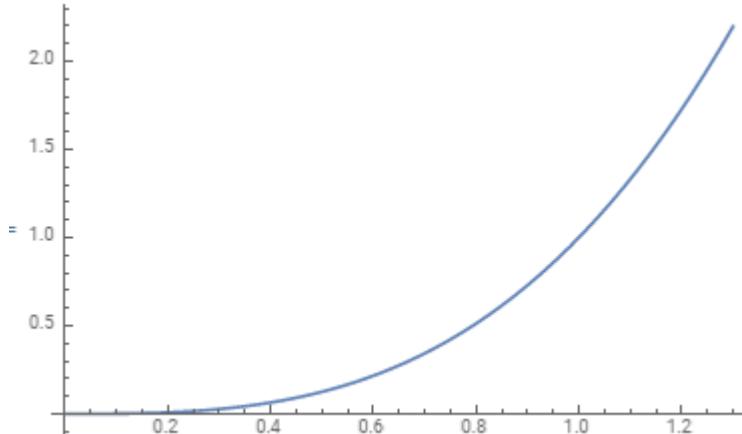
If



Take $\ln(1+x)$ instead of x

Transformation

If



Take \sqrt{x} or $\sqrt[n]{x}$ instead of x

Correlation

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

Correlation

Examples

Some examples of data that have a high correlation:

- Your caloric intake and your weight.
- The amount of time you study and your GPA.

Some examples of data that have a low correlation (or none at all):

- A dog's name and the type of dog biscuit they prefer.
- The cost of a car wash and how long it takes to buy a soda inside the station.



The Correlation Coefficient

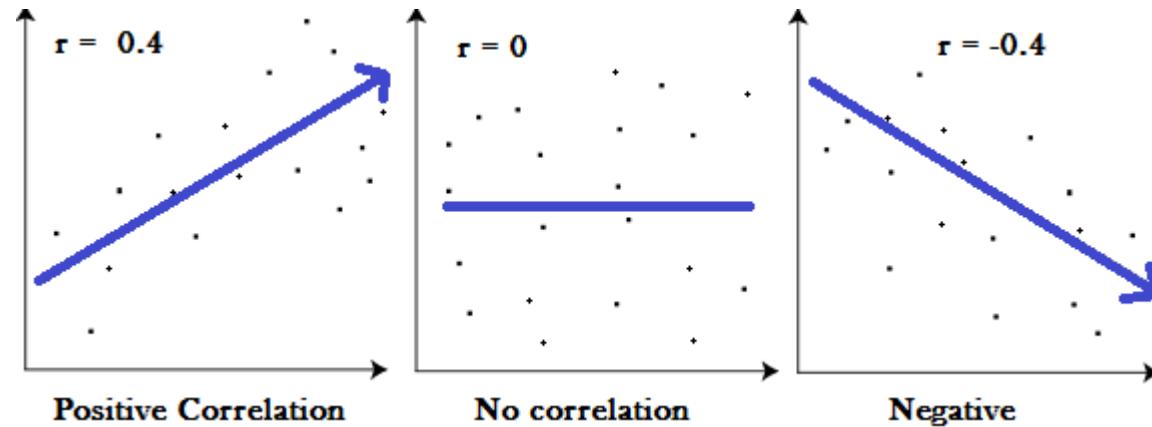
Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

Correlation Coefficient

Definition

- A correlation coefficient is a way to put a value to the relationship.
- Correlation coefficients have a value of between -1 and 1.
- A “0” means there is no relationship between the variables at all,
- While -1 or 1 means that there is a perfect negative or positive correlation

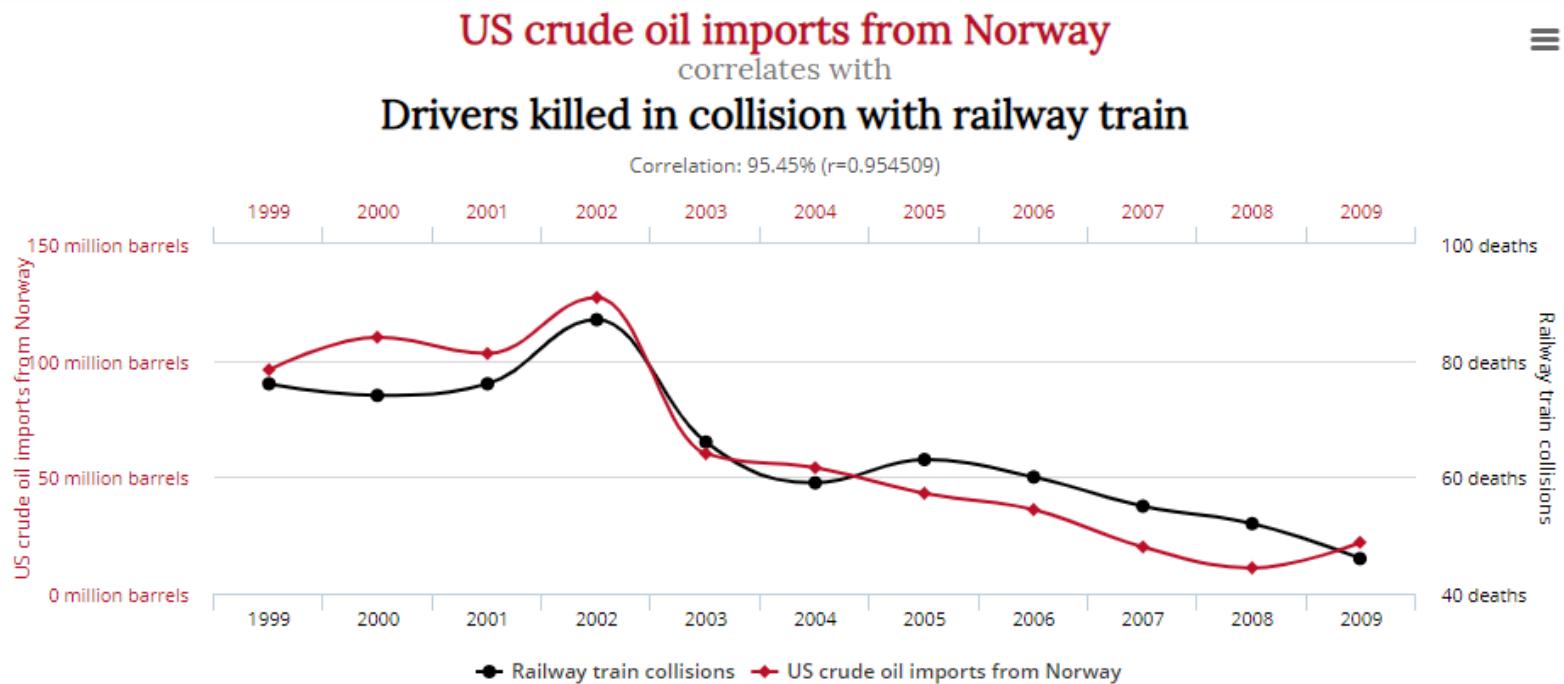
Example



Correlation vs Causation

Causation : The relation between something that happens and the thing that causes it . The first thing that happens is the cause and the second thing is the effect .

Correlation
vs
Causation



Source :<http://www.tylervigen.com/spurious-correlations>



Start-Tech
ACADEMY

Start-Tech Academy

Correlation Matrix

The Correlation Matrix

Definition

- A correlation matrix is a table showing correlation coefficients between variables.
- Each cell in the table shows the correlation between two variables.
- A correlation matrix is used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

Example

	Always to vote in elections	Never to try to evade taxes	Always to obey laws	Keep watch on action of govt
Always to vote in elections	1.00	.94	.94	.94
Never to try to evade taxes	.94	1.00	.97	.95
Always to obey laws	.94	.97	1.00	.96
Keep watch on action of govt	.94	.95	.96	1.00

Application

- To summarize a large amount of data where the goal is to see patterns.
- To Identify collinearity in the data



Multicollinearity

Multicollinearity

Definition

- Multicollinearity exists whenever two or more of the predictors in a regression model are moderately or highly correlated.

Effects

- Multicollinearity results in a change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.
- Multicollinearity makes it tedious to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.

Solution

- Remove highly correlated independent variables by looking at the correlation matrix and VIF



Dummy Variable

A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels.

Dummy Variable

Why

- Regression analysis treats all independent (X) variables in the analysis as numerical.
- Nominal variables, or variables that describe a characteristic using two or more categories, are commonplace in regression research, but are not always useable in their categorical form.
- Dummy coding is a way of incorporating nominal variables into regression analysis

How

- We can make a separate column, or variable, for each category.
- This new variables can take value 0 or 1 depending on the value of the categorical variable

Dummy Variable

A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels.

Student	Favorite class	Science	Math
1	Science	1	0
2	Science	1	0
3	English	0	0
4	Math	0	1

Dummy Variable Example

Things to keep in mind

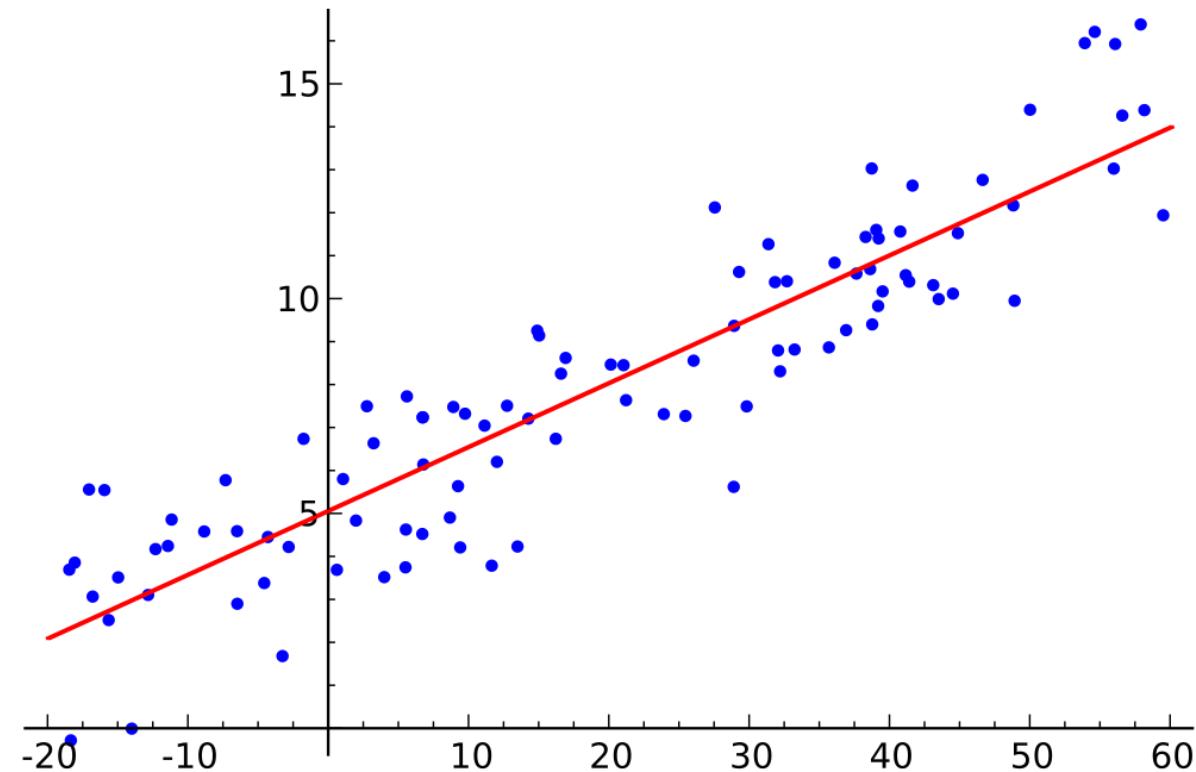
- The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels (categories) in that variable minus one.
- We cannot code variables like science = 1, math = 2, and English = 3. As, we can see that there is no such thing as an increase in favorite class – math is not higher than science, and is not lower than language either. And even if there is increase , we cannot quantify that increase



Linear Regression

linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables

Introduction



Linear Regression

price	crime_rate	resid_area	air_qual	room_num	age	dist1	dist2	dist3	dist4	teachers	poor_prop	airport	n_hos_beds	n_hot_rooms	waterbody	rainfall	bus_ter	parks
24	0.00632	32.31	0.538	6.575	65.2	4.35	3.81	4.18	4.01	24.7	4.98 YES		5.48	11.192 River		23 YES	0.04935	
21.6	0.02731	37.07	0.469	6.421	78.9	4.99	4.7	5.12	5.06	22.2	9.14 NO		7.332	12.1728 Lake		42 YES	0.04615	
34.7	0.02729	37.07	0.469	7.185	61.1	5.03	4.86	5.01	4.97	22.2	4.03 NO		7.394	101.12 None		38 YES	0.04576	
33.4	0.03237	32.18	0.458	6.998	45.8	6.21	5.93	6.16	5.96	21.3	2.94 YES		9.268	11.2672 Lake		45 YES	0.04715	
36.2	0.06905	32.18	0.458	7.147	54.2	6.16	5.86	6.37	5.86	21.3	5.33 NO		8.824	11.2896 Lake		55 YES	0.03947	
28.7	0.02985	32.18	0.458	6.43	58.7	6.22	5.8	6.23	5.99	21.3	5.21 YES		7.174	14.2296 None		53 YES	0.04591	
22.9	0.08829	37.87	0.524	6.012	66.6	5.87	5.47	5.7	5.2	24.8	12.43 YES		6.958	12.1832 River		41 YES	0.05217	
22.1	0.14455	37.87	0.524	6.172	96.1	6.04	5.85	6.25	5.66	24.8	19.15 NO		5.842	12.1768 Lake		56 YES	0.05707	
16.5	0.21124	37.87	0.524	5.631	100	6.18	5.85	6.3	6	24.8	29.93 YES		5.93	12.132 None		55 YES	0.0563	
18.9	0.17004	37.87	0.524	6.004	85.9	6.67	6.55	6.85	6.29	24.8	17.1 YES		9.478	14.1512 River		45 YES	0.05073	
15	0.22489	37.87	0.524	6.377	94.3	6.65	6.31	6.55	5.88	24.8	20.45 NO		6	11.12 Lake		29 YES	0.05778	
18.9	0.11747	37.87	0.524	6.009	82.9	6.27	5.93	6.51	6.19	24.8	13.27 NO		9.278	13.1512 Lake and Riv		23 YES	0.05524	
21.7	0.09378	37.87	0.524	5.889	39	5.76	5.14	5.58	5.33	24.8	15.71 YES		5.534	10.1736 Lake and Riv		57 YES	0.05742	

Questions

Here are a few important questions that we might seek to address:

1. Prediction Question

How accurately can I predict the price of a house , given the values of all variables

2. Inferential Question

How accurately can we estimate the effect of each of this variables on the house price



Simple Linear Regression

Simple linear regression is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y .

Introduction

Model Equation

$$Y \approx \beta_0 + \beta_1 X$$

β_0 is known as Intercept

β_1 is known as slope

Together β_0 and β_1 known as the model *coefficients or parameters*.

For House Price data

- X will represent Room_num
- Y will represent Price

$$\text{Price} \approx \beta_0 + \beta_1 \times \text{Room_num}$$

From our training data we will get $\hat{\beta}_0$ and $\hat{\beta}_1$



Simple Linear Regression

Estimating the Coefficients

- Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well
- Total number of rows (Data Point) $\Rightarrow n = 506$
- Data $\Rightarrow (x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_{506}, y_{506})$
- Lets call calculated y value as \hat{y}
$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$
$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$$
$$\hat{y}_{506} = \hat{\beta}_0 + \hat{\beta}_1 x_{506}$$
- The difference between residual the i th observed response value and the i th response value that is predicted by our linear model is known as residual
$$e_i = y_i - \hat{y}_i$$



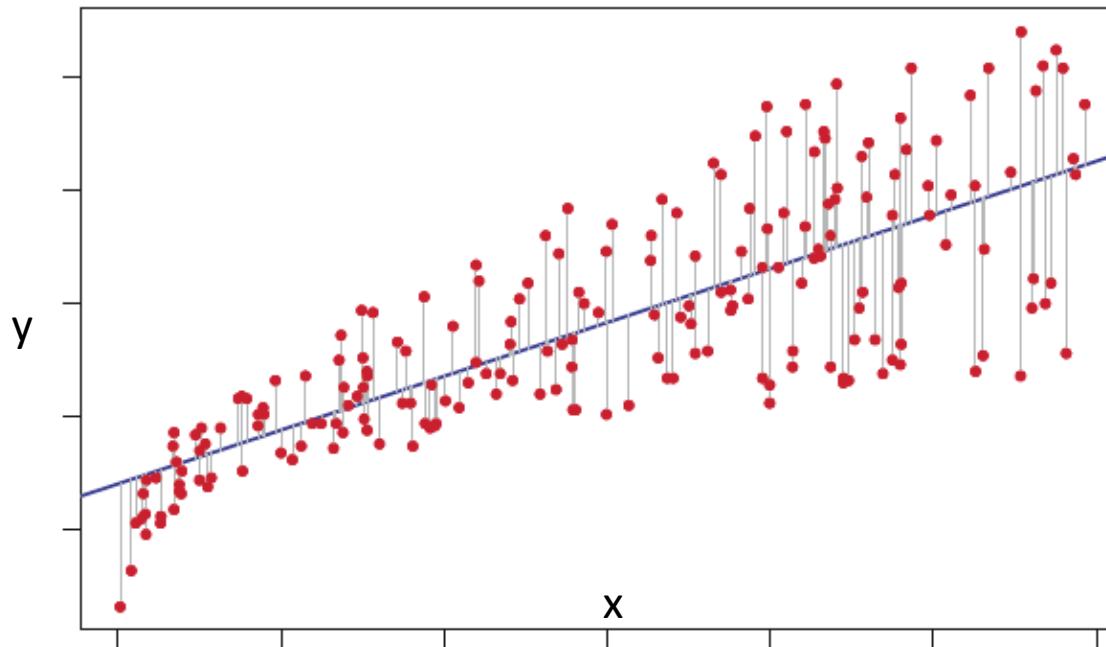
Simple Linear Regression

Residual

Residual –

The difference between residual the i th observed response value and the i th response value that is predicted by our linear model is known as residual

$$e_i = y_i - \hat{y}_i$$



Simple Linear Regression

RSS

Residual sum of squares (RSS)

$$RSS = e_1^2 + e_2^2 + \dots + e_n^2$$

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

Using some calculus, one can show that the minimizers are

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x},\end{aligned}$$



Start-Tech
ACADEMY

Start-Tech Academy

Simple Linear Regression

Model

For our Model

Residuals:

Min	1Q	Median	3Q	Max
-23.336	-2.425	0.093	2.918	39.434

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-34.6592	2.6421	-13.12	<2e-16 ***
room_num	9.0997	0.4178	21.78	<2e-16 ***

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1

Residual standard error: 6.597 on 504 degrees of freedom

Multiple R-squared: 0.4848, Adjusted R-squared: 0.4838

F-statistic: 474.3 on 1 and 504 DF, p-value: < 2.2e-16



Simple Linear Regression

we assume that the true relationship between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f , where ε is a mean-zero random error term.

Assessing the Accuracy

If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

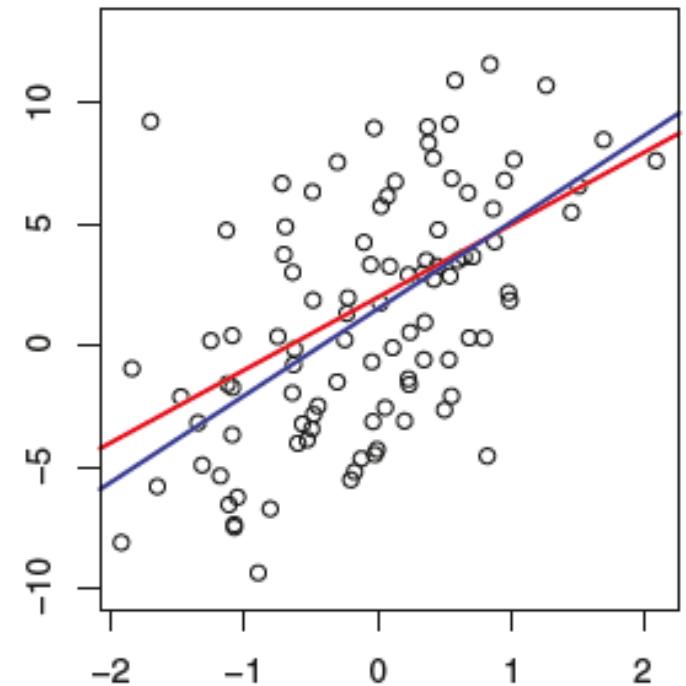
β_0 is known as Intercept

β_1 is known as slope

ε is an error term

— Population regression line

— Sample regression line



Simple Linear Regression

Standard error In Coefficients

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\sigma^2 = \text{Var}(\varepsilon)$$

σ^2 is not known, but can be estimated from the data. This estimate is known as the *residual standard error (RSE)*

$$\text{RSE} = \sqrt{\text{RSS}/(n - 2)}$$

There is approximately a 95% chance that the interval

$$[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)]$$

will contain the true value of β_1



Simple Linear Regression

Hypothesis tests

Is there any relationship between X and Y

$$Y = \beta_0 + \beta_1 X$$

- If β_1 is zero, it means there is no relationship

$H_0 : \text{There is no relationship between } X \text{ and } Y$

$H_a : \text{There is some relationship between } X \text{ and } Y$

$$H : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0,$$



Simple Linear Regression

Hypothesis tests

- To disapprove H_0 , we calculate T statistics
- We also compute the probability of observing any value equal to $|t|$ or larger
- We call this probability the *p-value*
- A small p-value means there is an association between the predictor and the response (typically less than 5% or 1 %)

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

Residuals:

Min	1Q	Median	3Q	Max
-23.336	-2.425	0.093	2.918	39.434

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-34.6592	2.6421	-13.12	<2e-16
room_num	9.0997	0.4178	21.78	<2e-16



Simple Linear Regression

Quality of Fit

RSE

The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.

Residual Standard Error

$$\text{RSE} = \sqrt{\frac{1}{n-2} \text{RSS}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

- RSE is the average amount that the response will deviate from the true regression line
- RSE is also considered as a measure of lack of fit of the model to the data

```
Residual standard error: 6.597 on 504 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.4838
F-statistic: 474.3 on 1 and 504 DF,  p-value: < 2.2e-16
```



Simple Linear Regression

Quality of Fit

R^2

The RSE provides an absolute measure of lack of fit of the model to the data.

R^2

- R^2 is the proportion of variance explained
- R^2 always takes on a value between 0 and 1,
- R^2 is independent of the scale of Y.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- *TSS - total sum of squares*
- *RSS - residual sum of squares*

```
Residual standard error: 6.597 on 504 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.4838
F-statistic: 474.3 on 1 and 504 DF,  p-value: < 2.2e-16
```



Multiple Linear Regression

In Multiple linear regression more than one predictor variables are used to predict the response variable

Multiple Linear Regression

Relationship for Multiple linear regression can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon,$$

β_0 is known as Intercept

p is the number of predictors

ϵ is an error term

For our Model,

The equation is

$$\text{Price} = \beta_0 + \beta_1 \text{Crime_rate} + \beta_2 \text{poor_pop} \dots \dots \beta_{16} \text{avg_dist}$$



Multiple Linear Regression

Estimating Regression Coefficients

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.498625	5.264076	-1.235	0.2176
crime_rate	0.009710	0.348185	0.028	0.9778
resid_area	-0.040875	0.057585	-0.710	0.4782
air_qual	-15.897400	4.003793	-3.971	8.24e-05 ***
room_num	4.019017	0.426606	9.421	< 2e-16 ***
age	-0.005715	0.013606	-0.420	0.6747
teachers	1.007001	0.122098	8.247	1.50e-15 ***
poor_prop	-0.577271	0.052695	-10.955	< 2e-16 ***
airportYES	1.131516	0.454266	2.491	0.0131 *
n_hos_beds	0.329221	0.152239	2.163	0.0311 *
n_hot_rooms	0.091868	0.082174	1.118	0.2641
waterbodyLake	0.264086	0.641963	0.411	0.6810
'waterbodyLake and River'	-0.687556	0.714023	-0.963	0.3361
waterbodyRiver	-0.291319	0.546656	-0.533	0.5943
rainfall	0.016119	0.017839	0.904	0.3667
avg_dist	-1.218640	0.188933	-6.450	2.68e-10 ***

Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *
	.	.	.	1

Residual standard error: 4.925 on 490 degrees of freedom
Multiple R-squared: 0.7208, Adjusted R-squared: 0.7123
F-statistic: 84.34 on 15 and 490 DF, p-value: < 2.2e-16



Multiple Linear Regression

In Multiple linear regression more than one predictor variables are used to predict the response variable

Multiple Linear Regression

Relationship for Multiple linear regression can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon,$$

β_0 is known as Intercept

p is the number of predictors

ϵ is an error term

For our Model,

The equation is

$$\text{Price} = \beta_0 + \beta_1 \text{Crime_rate} + \beta_2 \text{poor_pop} \dots \dots \beta_{16} \text{avg_dist}$$



Multiple Linear Regression

Estimating Regression Coefficients

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.498625	5.264076	-1.235	0.2176
crime_rate	0.009710	0.348185	0.028	0.9778
resid_area	-0.040875	0.057585	-0.710	0.4782
air_qual	-15.897400	4.003793	-3.971	8.24e-05 ***
room_num	4.019017	0.426606	9.421	< 2e-16 ***
age	-0.005715	0.013606	-0.420	0.6747
teachers	1.007001	0.122098	8.247	1.50e-15 ***
poor_prop	-0.577271	0.052695	-10.955	< 2e-16 ***
airportYES	1.131516	0.454266	2.491	0.0131 *
n_hos_beds	0.329221	0.152239	2.163	0.0311 *
n_hot_rooms	0.091868	0.082174	1.118	0.2641
waterbodyLake	0.264086	0.641963	0.411	0.6810
'waterbodyLake and River'	-0.687556	0.714023	-0.963	0.3361
waterbodyRiver	-0.291319	0.546656	-0.533	0.5943
rainfall	0.016119	0.017839	0.904	0.3667
avg_dist	-1.218640	0.188933	-6.450	2.68e-10 ***

Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *
	.	.	.	1

Residual standard error: 4.925 on 490 degrees of freedom
Multiple R-squared: 0.7208, Adjusted R-squared: 0.7123
F-statistic: 84.34 on 15 and 490 DF, p-value: < 2.2e-16



Multiple Linear Regression

F statistics

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.498625  5.264076 -1.235  0.2176
crime_rate   0.009710  0.348185  0.028  0.9778
resid_area  -0.040875  0.057585 -0.710  0.4782
air_qual    -15.897400  4.003793 -3.971 8.24e-05 ***
room_num     4.019017  0.426606  9.421 < 2e-16 ***
age          -0.005715  0.013606 -0.420  0.6747
teachers     1.007001  0.122098  8.247 1.50e-15 ***
poor_prop   -0.577271  0.052695 -10.955 < 2e-16 ***
airportYES  1.131516  0.454266  2.491  0.0131 *
n_hos_beds  0.329221  0.152239  2.163  0.0311 *
n_hot_rooms 0.091868  0.082174  1.118  0.2641
waterbodyLake 0.264086  0.641963  0.411  0.6810
`waterbodyLake and River` -0.687556  0.714023 -0.963  0.3361
waterbodyRiver -0.291319  0.546656 -0.533  0.5943
rainfall      0.016119  0.017839  0.904  0.3667
avg_dist     -1.218640  0.188933 -6.450 2.68e-10 ***
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 0.1 ' ' 1

Residual standard error: 4.925 on 490 degrees of freedom
Multiple R-squared:  0.7208   Adjusted R-squared:  0.7123
F-statistic: 84.34 on 15 and 490 DF,  p-value: < 2.2e-16
```

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

H_a : at least one β_j is non-zero.

Multiple Linear Regression

F statistics

Coin is biased if I get 5 consecutive heads in 5 tosses



Probability of Head	1/2	1/2	1/2	1/2	1/2
---------------------	-----	-----	-----	-----	-----

Probability of classifying a fair coin as a biased coin = $(1/2)^5 = 0.03125$

If 100 coins are tossed 5 times each ,
What is the probability of getting all heads in at least one of the coin

$$1 - \left(1 - \frac{1}{32}\right)^{100} \approx 95\%$$



Multiple Linear Regression

F statistics

For our model

$$\beta \quad \beta \quad \beta \quad \beta \quad \beta$$

Probability of
Wrongly
classifying B as
significant

5%

5%

5%

5%

5%

If number of variables is large, there is very high chance that one of the B is wrongly classified

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)},$$



Multiple Linear Regression

Categorical Variables

airportYES	waterbodyLake	waterbodyLake and River	waterbodyRiver
1	0	0	1
0	1	0	0
0	0	0	0
1	1	0	0
0	1	0	0
1	0	0	0
1	0	0	1
0	1	0	0
1	0	0	0

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{If Airport is present} \\ \beta_0 + \epsilon_i & \text{If Airport is not present} \end{cases}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
airportYES	1.131516	0.454266	2.491	0.0131 *



Multiple Linear Regression

Categorical Variables

airportYES	waterbodyLake	waterbodyLake and River	waterbodyRiver
1	0	0	1
0	1	0	0
0	0	0	0
1	1	0	0
0	1	0	0
1	0	0	0
1	0	0	1
0	1	0	0
1	0	0	0

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
waterbodyLake	0.264086	0.641963	0.411	0.6810
'waterbodyLake and River'	-0.687556	0.714023	-0.963	0.3361
waterbodyRiver	-0.291319	0.546656	-0.533	0.5943

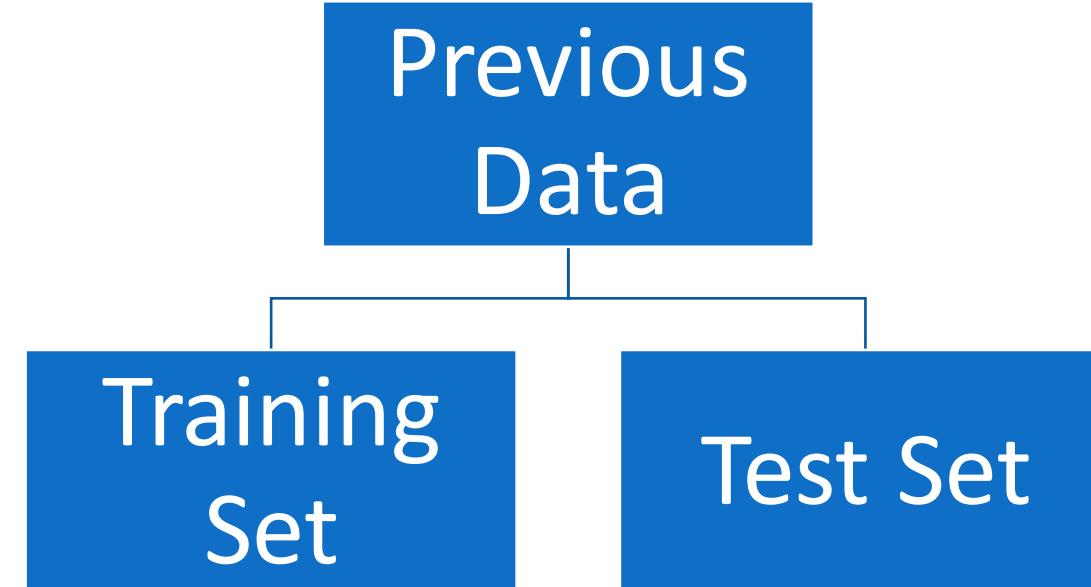


Linear Regression

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

- Training error – Performance of model on the previously **seen** data
- Test error – Performance of model on the **unseen** data

Test-Train Split



Linear Regression

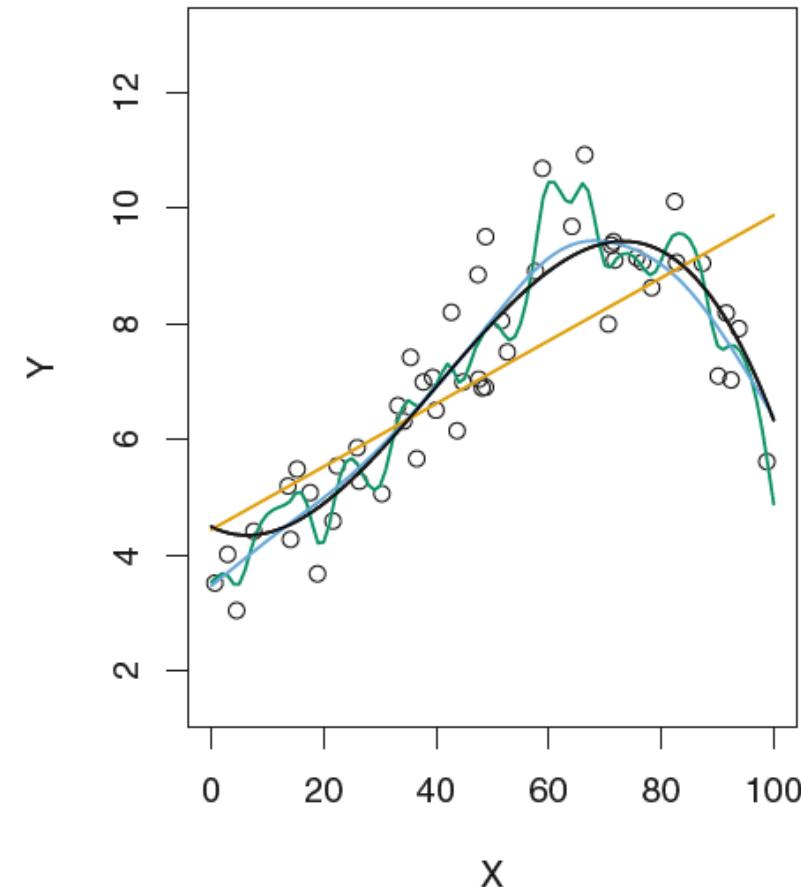
Test-Train Split

	Training Set	- $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
	Model is trained	$y = f(x)$
	Test Set	- Previously unseen data (x_0, y_0)
	Test MSE	- $\text{Ave}(\hat{f}(x_0) - y_0)^2$

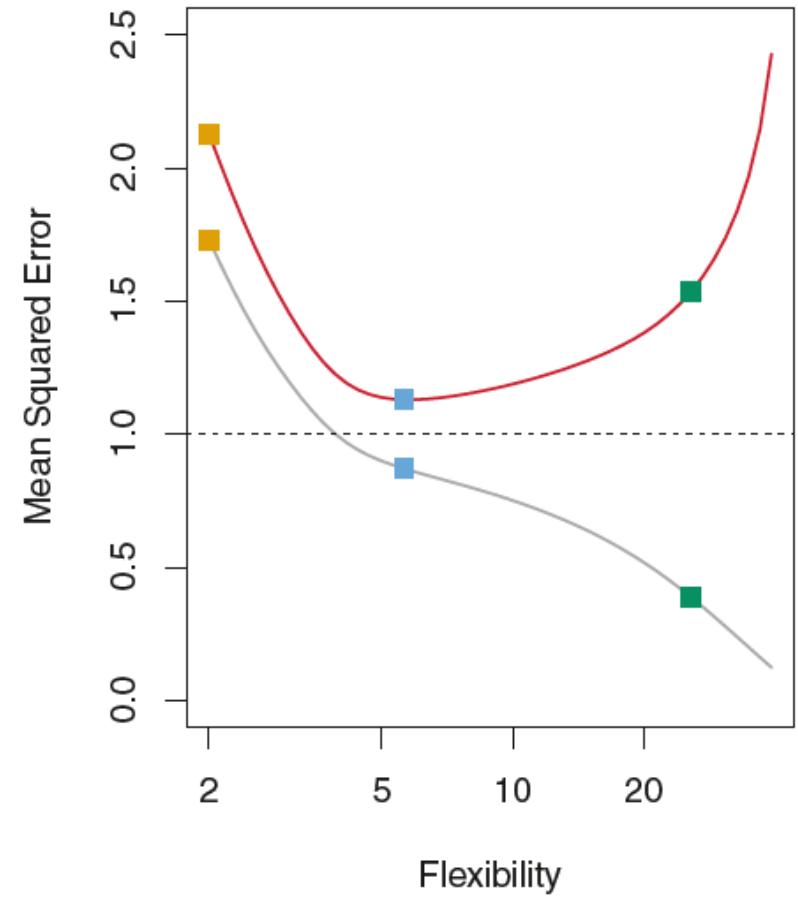


Other Linear Regression

Test-Train Split



- True Function
- Output of linear Model
- Output of more flexible model



- Test error
- Training error

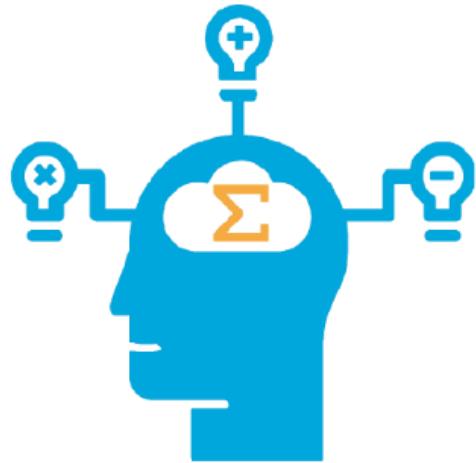


Linear Regression

Test-Train Split Techniques

- 1. Validation set approach**
 - Random division of data into two parts
 - Usual split is 80:20 (Training : Test)
 - When to use – In case of large number of observations
- 2. Leave one out cross validation**
 - Leaving one observation every time from training set
- 3. K-Fold validation**
 - Divide the data into k set
 - We will keep one testing and K-1 for training





Start-Tech Academy

Linear Regression

Bias-Variance Trade-Off

$$\text{Expected test error} = E(\text{Bias}) + E(\text{Variance}) + E(\epsilon)$$

$E(\epsilon)$

Variance of error, Irreducible

$E(\text{Variance})$

Amount by which predicted function will change if we change training dataset

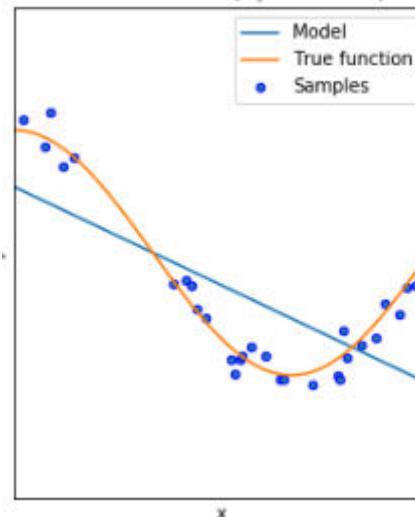
$E(\text{Bias})$

Error due to approximation of complex relationship as a simpler model such as linear model

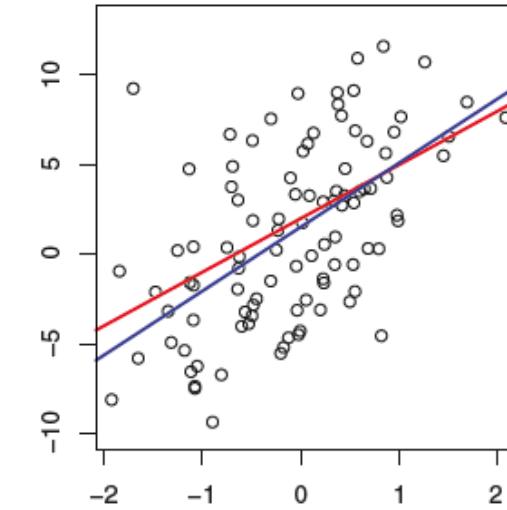


Linear Regression

Variance

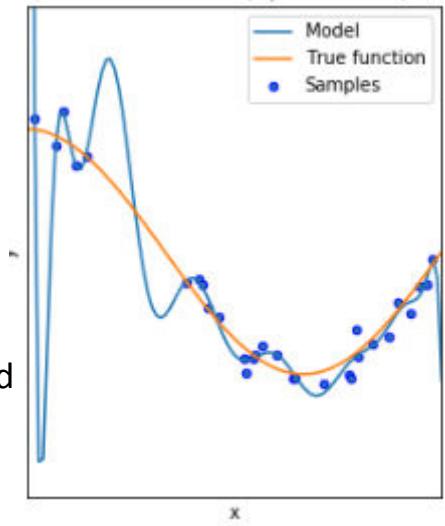


Less Flexible method



- Population regression line
- Sample regression line

More Flexible method

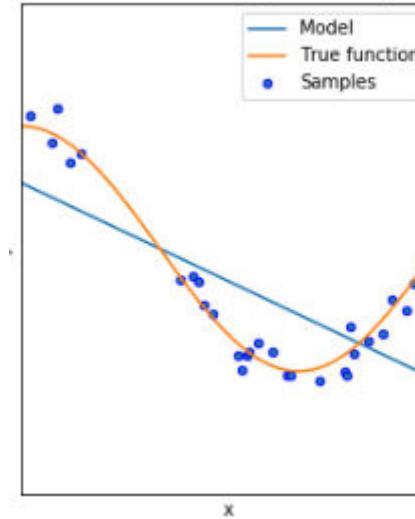


Linear Regression

Bias

$E(Bias)$

Error due to approximation of complex relationship as a simpler model such as linear model



Linear Regression

Bias-Variance Trade-Off

The Tradeoff

If we try to decrease one by changing model flexibility, other one increases

- Bias + Variance
- Bias
- Variance

