PROJECT REPORT

(Project Term August-November, 2021)

CUSTOMER CHURN PREDICTION

Submitted by

Rupesh Kumar   Registration Number : 11904657

Project Group Number: NA

Course Code: INT246

Under the Guidance of

Dr. Sagar Pande

School of Computer Science and Engineering

LOVELY PROFESSIONAL UNIVERSITY

DECLARATION

We hereby declare that the project work entitled ("Customer Churn Prediction") is an authentic record of our own work carried out as requirements of Project for the award of B.Tech degree in CSE from Lovely Professional University, Phagwara, under the guidance of Dr. Sagar Pande, during August to November 2021. All the information furnished in this project report is based on our own intensive work and is genuine.

Project Group Number: NA

Name of Student : Rupesh kumar
Registration Number: 11904657

Rupesh Kumar
Date: 20th November, 2021

ACKNOWLEDGEMENT

The success and final outcome of learning **Machine Learning** required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my course and few of the projects. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank **LPU**, for providing me an opportunity to do the course and project work and giving me all support and guidance, which make me complete the course duly. I am extremely thankful to the course advisor **Dr. Sagar Pande** sir.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of **LPU** which helped us in successfully completing my course and project work.

Name: Rupesh Kumar

Reg No: 11904657

Date: 20th November, 2021

Signature:

Rupesh Kumar

CERTIFICATE

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfillment of the conditions for the award of B.Tech degree in CSE from Lovely Professional University, Phagwara.

Signature and Name of the Mentor

Designation

School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab.

TABLE OF CONTENTS

❖ CHAPTER– 1 (Descriptive Statistics)

➢ <u>Probability and Statistics for Data Science:</u>

Probability and Statistics form the basis of Data Science. The probability theory is very much helpful for making the prediction. Estimates and predictions form an important part of Data science. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability. And all of probability and statistics is dependent on Data.

Data:

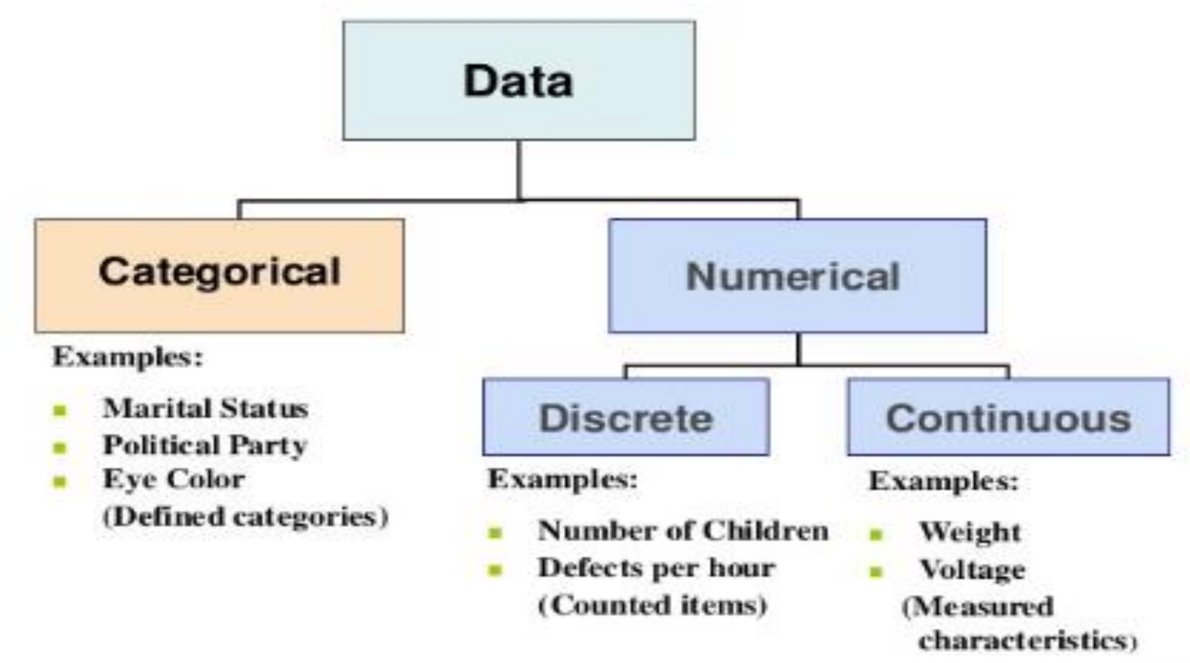
Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

Data — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process.

Why does Data Matter?

- Helps in understanding more about the data by identifying relationships that may exist between 2 variables.
- Helps in predicting the future or forecast based on the previous trend of data.
- Helps in determining patterns that may exist between data.
- Helps in detecting fraud by uncovering anomalies in the data.

Data matters a lot nowadays as we can infer important information from it. Now let's delve into how data is categorized. Data can be of 2 types categorical and numerical data. For Example in a bank, we have regions, occupation class, gender which follow categorical data as the data is within a fixed certain value and balance, credit score, age, tenure months follow numerical continuous distribution as data can follow an unlimited range of values.

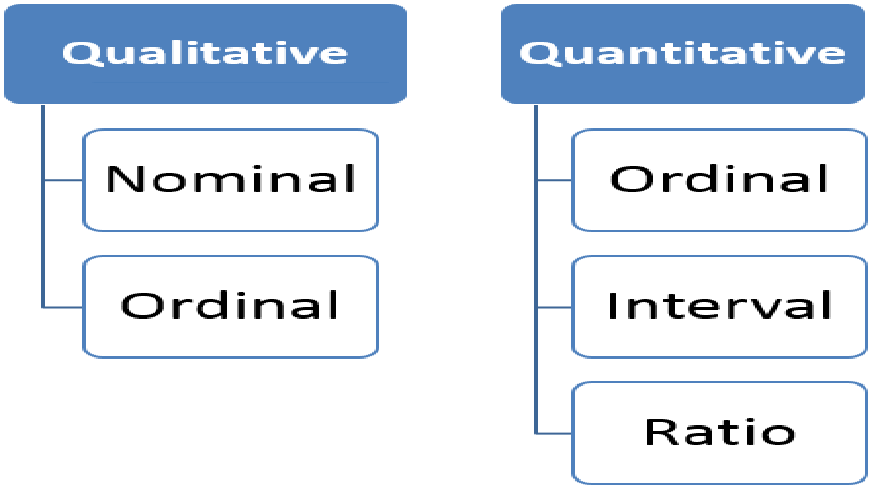


Note: Categorical Data can be visualized by Bar Plot, Pie Chart, Pareto Chart. Numerical Data can be visualized by Histogram, Line Plot, Scatter Plot.

Descriptive Statistics:

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. It helps us in knowing our data better. It is used to describe the characteristics of data.

Measurement level of Data:



The qualitative and quantitative data is very much similar to the above categorical and numerical data.

Nominal: Data at this level is categorized using names, labels or qualities. eg: Brand Name, ZipCode, Gender.

Ordinal: Data at this level can be arranged in order or ranked and can be compared. eg: Grades, Star Reviews, Position in Race, Date.

Interval: Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth.

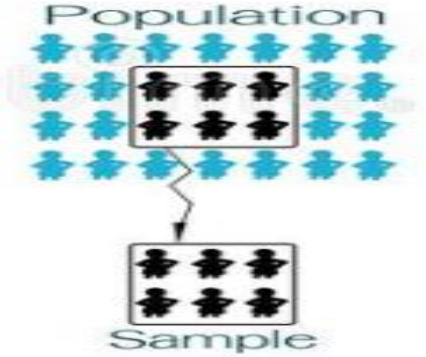
Ratio: Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points. eg: Height, Age, Weight.

Population or Sample Data:

Before performing any analysis of data, we should determine if the data we’re dealing with is population or sample.

Population: Collection of all items (N) and it includes each and every unit of our study. It is hard to define and the measure of characteristic such as mean, mode is called parameter.

Sample: Subset of the population (n) and it includes only a handful units of the population. It is selected at random and the measure of the characteristic is called as statistics.



For Example, say you want to know the mean income of the subscribers to a movie subscription service(parameter). We draw a random sample of 1000 subscribers and determine that their mean income(x̄) is $34,500 (statistic). We conclude that the population mean income (μ) is likely to be close to $34,500 as well.

Now before looking at distributions of data. Let’s take a look at measures of data.

Measures of Central Tendency:

The measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

Mean: The mean is equal to the sum of all the values in the data set divided by the number of values in the data set i.e the calculated average. It susceptible to outliers when unusual values are added it gets skewed i.e deviates from the typical central value.

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

Median: The median is the middle value for a dataset that has been arranged in order of magnitude. Median is a better alternative to mean as it is less affected by outliers and skewness of the data. The median value is much closer than the typical central value.

If the total number of values is odd then

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

If the total number of values is even then

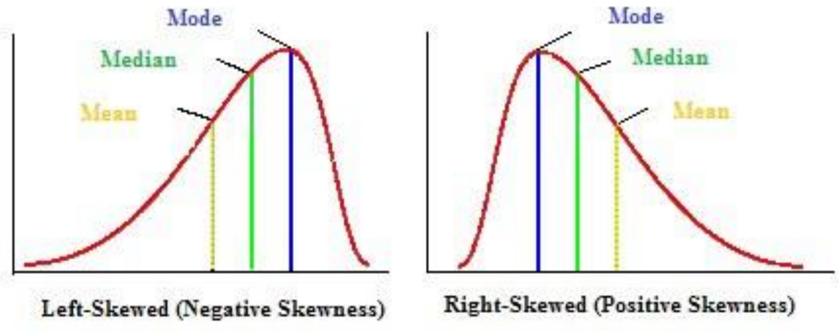
$$\text{Median} = \left(\frac{(\frac{n}{2})^{th}\,term + (\frac{n}{2}+1)^{th}\,term}{2}\right)^{\text{th}}\ term$$

Mode: The mode is the most commonly occurring value in the dataset. The mode can, therefore sometimes consider the mode as being the most popular option.

For Example, In a dataset containing {13,35,54,54,55,56,57,67,85,89,96} values. Mean is 60.09. Median is 56. Mode is 54.

Measures of Asymmetry:

Skewness: Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right. Skewness indicates whether the data is concentrated on one side.

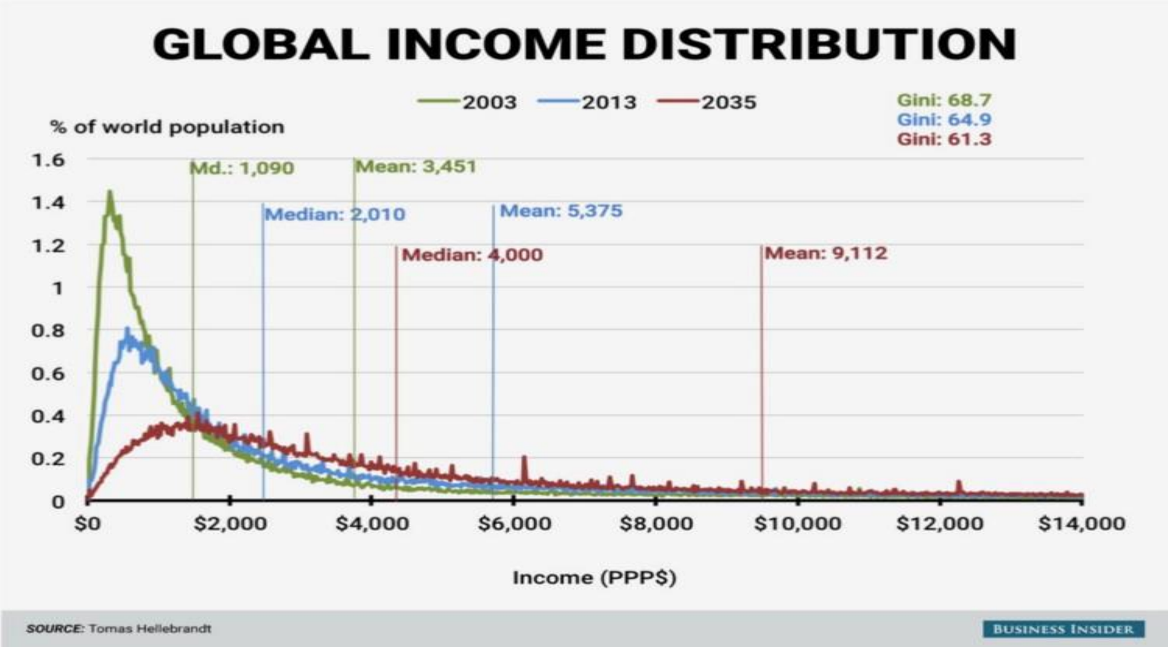

Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Positive Skewness: Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.

Negative Skewness: Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.

Skewness is important as it tells us about where the data is distributed.

For eg: Global Income Distribution in 2003 is highly right-skewed.We can see the mean $3,451 in 2003(green) is greater than the median $1,090. It suggests that the global income is not evenly distributed. Most individuals incomes are less than $2,000 and less number of people with income above $14,000, so the skewness. But it seems in 2035 according to the forecast income inequality will decrease over time.



Measures of Variability(Dispersion):

The measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observation fully. The measure of dispersion helps us to study the variability of the items i.e the spread of data.

Remember: Population Data has N data points and Sample Data has (n-1) data points. (n-1) is called Bessel's Correction and it is used to reduce bias.

Range: The difference between the largest and the smallest value of a data, is termed as the range of the distribution. Range does not consider all the values of a series, i.e. it takes only the extreme items and middle items are not considered significant. eg: For {13,33,45,67,70} the range is 57 i.e(70–13).

Variance: Variance measures how far is the sum of squared distances from each point to the mean i.e the dispersion around the mean.

Variance is the average of all squared deviations.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} \text{ for populations}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ for samples}$$

Note: The units of values and variance is not equal so we use another variability measure.

Standard Deviation: As Variance suffers from unit difference so standard deviation is used. The square root of the variance is the standard deviation. It tells about the concentration of the data around the mean of the data set.

Population standard deviation: σ

= square root of the population variance

$$\sigma = \sqrt{\sigma^2}$$

Sample standard deviation: s

= square root of the sample variance, so that

$$s = \sqrt{s^2}$$

For eg: {3,5,6,9,10} are the values in a dataset.

$$\text{Mean} = \frac{3+5+6+9+10}{5} = 6.6$$

$$\text{Variance} = \frac{(3-6.6)^2 + (5-6.6)^2 + (6-6.6)^2 + (9-6.6)^2 + (10-6.6)^2}{5}$$

$$= \frac{12.96 + 2.56 + \ 0.36 + 5.76 + 11.56}{5} = \frac{33.2}{5} = 6.64$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{6.64} = 2.576$$

Coefficient of Variation(CV): It is also called as the relative standard deviation. It is the ratio of standard deviation to the mean of the dataset.

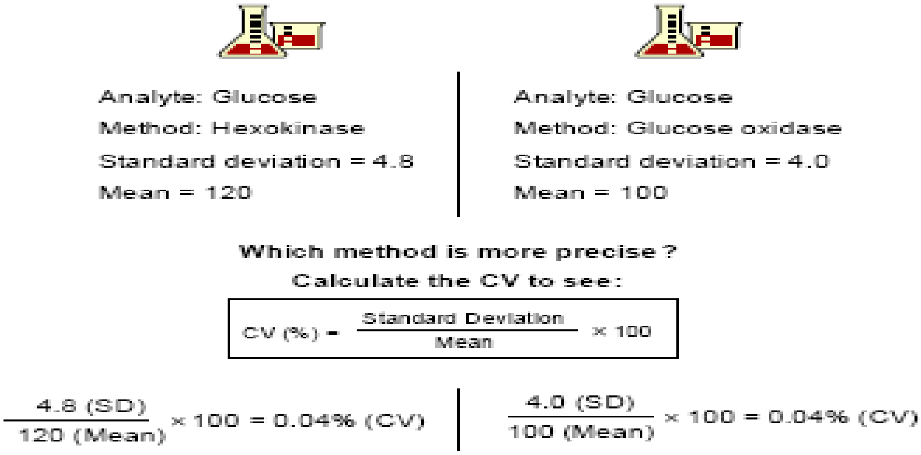
CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

Standard deviation is the variability of a single dataset. Whereas the coefficient of variance can be used for comparing 2 datasets.



Analyte: Glucose
Method: Hexokinase
Standard deviation = 4.8
Mean = 120

Analyte: Glucose
Method: Glucose oxidase
Standard deviation = 4.0
Mean = 100

Which method is more precise?
Calculate the CV to see:

$$\boxed{\text{CV (\%)} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100}$$

$$\frac{4.8\ \text{(SD)}}{120\ \text{(Mean)}} \times 100 = 0.04\%\ \text{(CV)}$$

$$\frac{4.0\ \text{(SD)}}{100\ \text{(Mean)}} \times 100 = 0.04\%\ \text{(CV)}$$

From the above example, we can see that the CV is the same. Both methods are precise. So it is perfect for comparisons.

Measures of Quartiles:

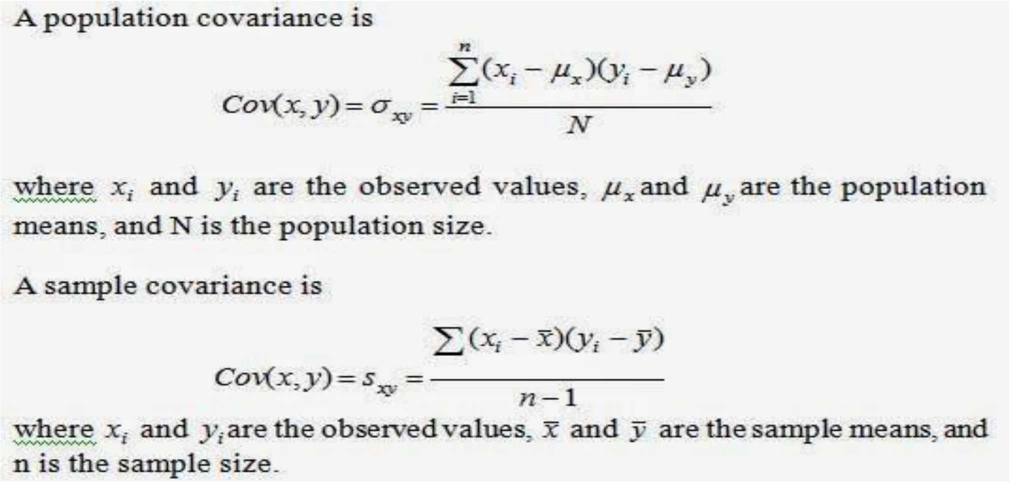
Quartiles are better at understanding as every data point considered.

Check my previous post — In the Boxplot Section, I have elaborated on Quartiles.

Measures of Relationship:

Measures of relationship are used to find the comparison between 2 variables.

Covariance: Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.



A population covariance is

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

where x_i and y_i are the observed values, μ_x and μ_y are the population means, and N is the population size.

A sample covariance is

$$Cov(x,y) = s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where x_i and y_i are the observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

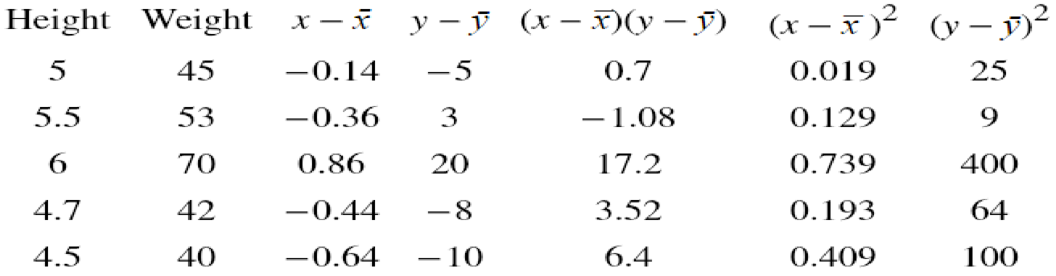
Covariance does not give effective information about the relation between 2 variables as it is not normalized.

Correlation: Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.

$$Correlation = \rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.

For Example,



Height	Weight	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
5	45	-0.14	-5	0.7	0.019	25
5.5	53	-0.36	3	-1.08	0.129	9
6	70	0.86	20	17.2	0.739	400
4.7	42	-0.44	-8	3.52	0.193	64
4.5	40	-0.64	-10	6.4	0.409	100

Sum(Height) = 25.7 Mean(Height) = 5.14

Sum(Weight) = 250 Mean(Weight) = 50

$$\sum(x-\bar{x})(y-\bar{y}) = 26.74$$

$$\sum(x-\bar{x})^2 = 1.489$$

$$\sum(y-\bar{y})^2 = 598$$

$$\text{Correlation} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{26.74}{\sqrt{1.489}\sqrt{598}} = \frac{26.54}{1.220 * 24.454} = 0.889$$

Correlation 0.889 tells us Height and Weight has a positive correlation. It is obvious that as the height of a person increases weight too increases.

Note: Correlation does not imply causation, Spurious Correlation for some strange correlations.

❖ CHAPTER- 2 (Inferential Statistics)

Descriptive statistics describe the important characteristics of data by using mean, median, mode, variance etc. It summarises the data through numbers and graphs.

In Inferential statistics, we make an inference from a sample about the population. The main aim of inferential statistics is to draw some conclusions from the sample and generalise them for the population data. E.g. we have to find the average salary of a data analyst across India. There are two options.

1. The first option is to consider the data of data analysts across India and ask them their salaries and take an average.
2. The second option is to take a sample of data analysts from the major IT cities in India and take their average and consider that for across India.

The first option is not possible as it is very difficult to collect all the data of data analysts across India. It is time-consuming as well as costly. So, to overcome this issue, we will look into the second option to collect a small sample of salaries of data analysts and take their average as India average. This is the inferential statistics where we make an inference from a sample about the population.

In inferential statistics, we will discuss probability, distributions, and hypothesis testing.

✓ Importance of Inferential Statistics

- Making conclusions from a sample about the population
- To conclude if a sample selected is statistically significant to the whole population or not
- Comparing two models to find which one is more statistically significant as compared to the other.
- In feature selection, whether adding or removing a variable helps in improving the model or not.

➢ Probability:

It is a measure of the chance of occurrence of a phenomenon. We will now discuss some terms which are very important in probability:

- **Random Experiment:** Random experiment or statistical experiment is an experiment in which all the possible outcomes of the experiments are already known. The experiment can be repeated numerous times under identical or similar conditions.
- **Sample space:** Sample space of a random experiment is the collection or set of all the possible outcomes of a random experiment.
- **Event:** A subset of sample space is called an event.
- **Trial:** Trial refers to a special type of experiment in which we have two types of possible outcomes: success or failure with varying Success probability.
- **Random Variable:** A variable whose value is subject to variations due to randomness is called a random variable. A random variable is of two types: Discrete and Continuous variable. In a mathematical way, we can say that a real-valued function X: S -> R is called a random variable where S is probability space and R is a set of real numbers.

Conditional Probability:

Conditional probability is the probability of a particular event Y, given a certain condition which has already occurred , i.e., X. Then conditional probability, P(Y|X) is defined as,

$P(Y|X) = N(X \cap Y) / N(X)$; provide $N(X) > 0$

N(X): – Total cases favourable to the event X

N(X∩Y): – Total favourable simultaneous

Or, we can write as:

$$P(Y|X) = P(X \cap Y) / P(X); \; P(X) > 0$$

Probability Distribution and Distribution function:

The mathematical function describing the randomness of a random variable is called probability distribution. It is a depiction of all possible outcomes of a random variable and their associated probabilities.

For a random variable X, CDF (Cumulative Distribution function) is defined as:

$$\mathbf{F(x) = P\ \{s\ \varepsilon\ S;\ X(s) \leq x\}}$$

Or,

$$\mathbf{F(x) = P\ \{X \leq x\}}$$

E.g. $P(X > 7) = 1 - P(X \leq 7)$

$= 1 - \{P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7)\}$

Sampling Distribution:

Probability distribution of statistics of a large number of samples selected from the population is called sampling distribution. When we increase the size of sample, sample mean becomes more normally distributed around population mean. The variability of the sample decreases as we increase sample size.

Central Limit Theorem:

CLT tells that when we increase the sample size, the distribution of sample means becomes normally distributed as the sample, whatever be the population distribution shape. This theorem is particularly true when we have a sample of size greater than 30.

The conclusion is that if we take a greater number of samples and particularly of large sizes, the distribution of sample means in a graph will look like to follow the normal distribution.

In the above graph we can see that when we increase the value of n i.e. sample size, it is approaching the shape of normal distribution.

Confidence Interval:

Confidence Interval is an interval of reasonable values for our parameters. Confidence intervals are used to give an interval estimation for our parameter of interest.

The margin of error is found by multiplying the standard error of the mean and the z-score.

Margin of error = $(z. \sigma)/ \sqrt{n}$

And Confidence interval is defined as:

Confidence interval having a value of 95% indicates that we are 95% sure that the actual mean is within our confidence interval.

Hypothesis Testing:

Hypothesis Testing is a part of statistics in which we make assumptions about the population parameter. So, hypothesis testing mentions a proper procedure by analysing a random sample of the population to accept or reject the assumption.

Type of Hypothesis:

A hypothesis is of two types:

1. Null hypothesis: Null hypothesis is a type of hypothesis in which we assume that the sample observations are purely by chance. It is denoted by H0.

2. Alternate hypothesis: The alternate hypothesis is a hypothesis in which we assume that sample observations are not by chance. They are affected by some non-random situation. An alternate hypothesis is denoted by H1 or Ha.

Steps of Hypothesis Testing:

The process to determine whether to reject a null hypothesis or to fail to reject the null hypothesis, based on sample data is called hypothesis testing. It consists of four steps:

1. Define the null and alternate hypothesis
2. Define an analysis plan to find how to use sample data to estimate the null hypothesis
3. Do some analysis on the sample data to create a single number called **'test statistic'**
4. Understand the result by applying the decision rule to check whether the Null hypothesis is true or not

If the value of t-stat is less than the significance level we will reject the null hypothesis, otherwise, we will fail to reject the null hypothesis.

Technically, we never accept the null hypothesis, we say that either we fail to reject or we reject the null hypothesis.

Terms in Hypothesis testing:

Significance level:

The significance level is defined as the probability of the case when we reject the null hypothesis but in actual it is true. E.g., a 0.05 significance level indicates that there is 5% risk in assuming that there is some difference when in actual there is no difference. It is denoted by alpha (α).

The above figure shows that the two shaded regions are equidistant from the null hypothesis, each having a probability of 0.025 and a total of 0.05 which is our significance level. The shaded region in case of a two-tailed test is called critical region.

P-value:

The p-value is defined as the probability of seeing a t-statistic as extreme as the calculated value if the null hypothesis value is true. Low enough p-value is ground for rejecting the null hypothesis. We reject the null hypothesis if the p-value is less than the significance level.

Errors in hypothesis testing:

We have explained what is hypothesis testing and the steps to do the testing. Now during performing the hypothesis testing, there might be some errors.

We classify these errors in two categories.

1. Type-1 error: Type 1 error is the case when we reject the null hypothesis but in actual it was true. The probability of having a Type-1 error is called significance level alpha(α).
2. Type-2 error: Type 2 error is the case when we fail to reject the null hypothesis but actually it is false. The probability of having a type-2 error is called beta(β).

Therefore,

α= P (Null hypothesis rejected | Null hypothesis is true)

β= P (Null hypothesis accepted | Null hypothesis is false)

Power of test is defined as

P= 1- Type-2 error

= 1 – β

Lesser the type-2 error more the power of the hypothesis test.

Decision –> / **Actual**	Reject the null hypothesis	Fail to reject the null hypothesis
Null Hypothesis is True	Type-1 Error	Decision is correct
Alternate hypothesis is true	Decision is correct	Type-2 Error

Z-test:

A Z-test is mainly used when the data is normally distributed. We find the Z-statistic of the sample means and calculate the z-score. Z-score is given by the formula,

$$\textbf{Z-score} = (x - \mu) / \sigma$$

Z-test is mainly used when the population mean and standard deviation are given.

T-test:

The t-test is similar to z-test. The only difference is that it is used when we have sample standard deviation but don’t have population standard, or have a small sample size (n<30).

Different types of T-test:

One Sample T-test:

The one-sample t-test compares the mean of sample data to a known value like if we have to compare the mean of sample data to the population mean we use the One-Sample T-test.

We can run a one-sample T-test when we do not have the population S.D. or we have a sample of size less than 30.

Two sample T-test:

We use a two-sample T-test when we want to evaluate whether the mean of the two samples is different or not. In two-sample T-test we have another two categories:

- Independent Sample T-test: Independent sample means that the two different samples should be selected from two completely different populations. In other words, we can say that one population should not be dependent on the other population.
- Paired T-test: If our samples are connected in some way, we have to use paired t-test. Here connecting means thatThe samples are connected as we are collecting data from the same group two times e.g. blood test of patients of a hospital before and after medication.

Chi-square test:

Chi-square test is used in the case when we have to compare categorical data. Chi-square test is of two types. Both use chi-square statistics and distribution for different purposes.

- Goodness of fit: It determines if sample data of categorical variables matches with population or not.
- Test of Independence: It compares two categorical variables to find whether they are related with each other or not.

ANOVA (Analysis of variance):

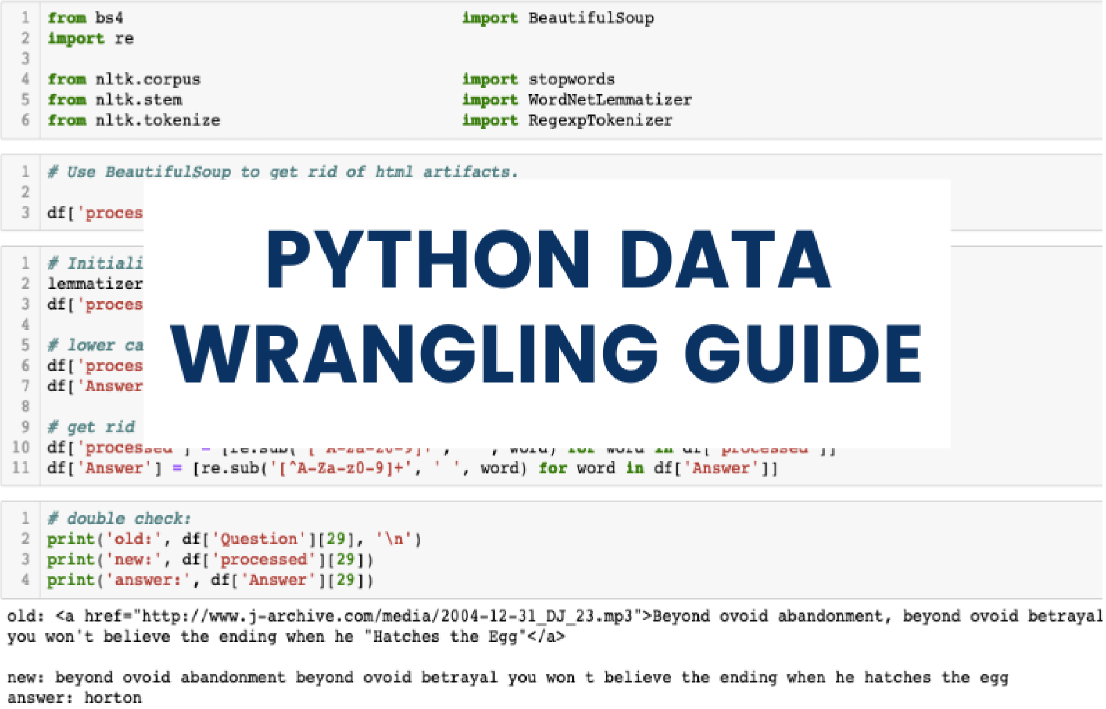
ANOVA test is a way to find out if an experiment results are significant or not. It is generally used when there are more than 2 groups and we have to test the hypothesis that the mean of multiple populations and variances of multiple populations are equal.

E.g. Students from different colleges take the same exam. We want to see if one college outperforms others.

There are two types of ANOVA test:

1. One-way ANOVA
2. Two-way ANOVA

❖ CHAPTER- 3 (Data Wrangling And Visualization Using Python Libraries)



```
from bs4                                import BeautifulSoup
import re

from nltk.corpus                        import stopwords
from nltk.stem                          import WordNetLemmatizer
from nltk.tokenize                      import RegexpTokenizer
```

```
# Use BeautifulSoup to get rid of html artifacts.

df['proces
```

```
# Initiali
lemmatizer
df['proces

# lower ca
df['proces
df['Answer

# get rid
df['processed'] = [re.sub('[^A-Za-z0-9]+', ' ', word) for word in df['processed']]
df['Answer'] = [re.sub('[^A-Za-z0-9]+', ' ', word) for word in df['Answer']]
```

```
# double check:
print('old:', df['Question'][29], '\n')
print('new:', df['processed'][29])
print('answer:', df['Answer'][29])
```

```
old: <a href="http://www.j-archive.com/media/2004-12-31_DJ_23.mp3">Beyond ovoid abandonment, beyond ovoid betrayal
you won't believe the ending when he "Hatches the Egg"</a>

new: beyond ovoid abandonment beyond ovoid betrayal you won t believe the ending when he hatches the egg
answer: horton
```

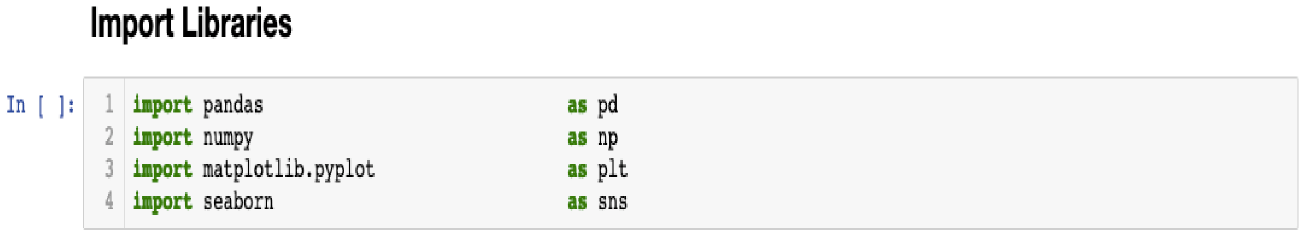
Data wrangling (otherwise known as data munging or preprocessing) is a key component of any data science project. Wrangling is a process where one transforms "raw" data for making it more suitable for analysis and it will improve the quality of your data. We will use Jeopardy questions from the Jeopardy Archive to wrangle textual data and process them for classification. Although one might not use all of the methods listed here depending on the data, there are several preprocessing methods that encapsulate the entire process:

1. **Data Exploration:** Checking for feature data types, unique values, and describing data.
2. **Null Values:** Counting null values and deciding what to do with them.
3. **Reshaping and Feature Engineering:** This step transforms raw data into a more useful format. Examples of feature engineering include one-hot encoding, aggregation, joins, and grouping.
4. **Text Processing:** BeautifulSoup and Regex (among other tools) are often used to clean and extract web scraped texts from HTML and XML documents.

Importing Libraries:

Let’s start by importing several libraries we’ll need for exploring our data and cleaning textual data:

1. **Pandas:** We will need Pandas to navigate our dataframe and check for each column’s data type, null values, and unique values.
2. **NumPy:** This package is essential for any data science project. It has a lot of mathematical functions that operate on multi-dimensional arrays and data frames.
3. **Matplotlib & Seaborn:** They are plotting and graphing libraries that we will use to visualize data in an intuitive way.

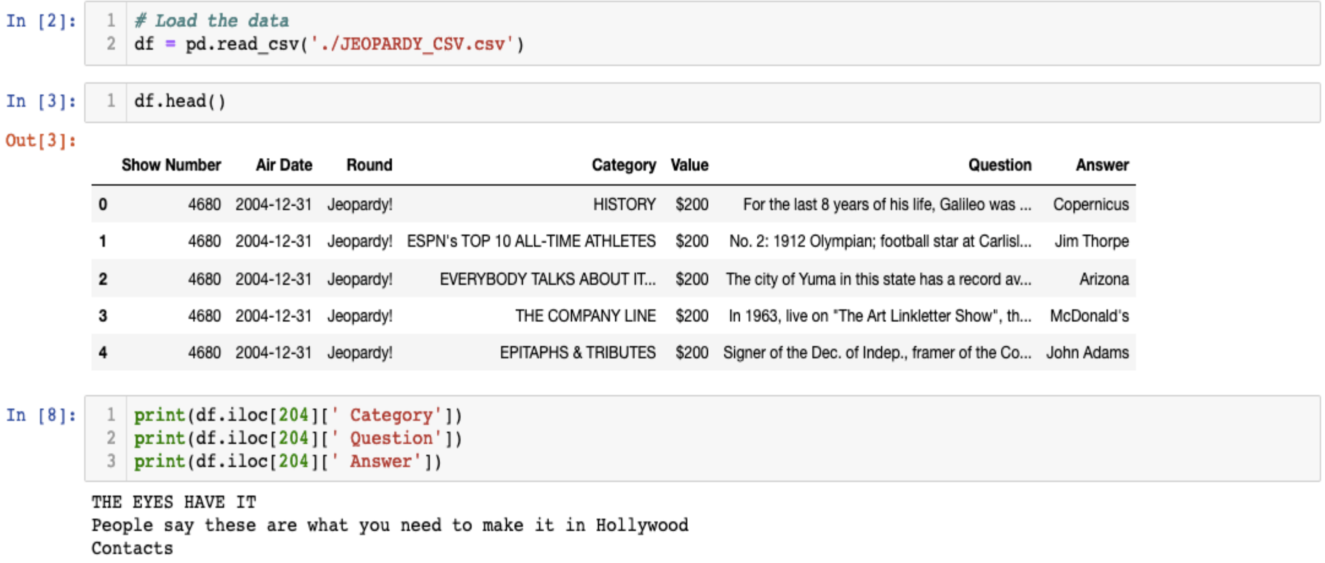


Import Libraries

```
In [ ]:
import pandas                    as pd
import numpy                     as np
import matplotlib.pyplot         as plt
import seaborn                   as sns
```

Loading the Data:

Let’s load our data, run df.head() on it, and see if we can load up a question.



```
In [2]:
# Load the data
df = pd.read_csv('./JEOPARDY_CSV.csv')
```

```
In [3]:
df.head()
```

Out[3]:

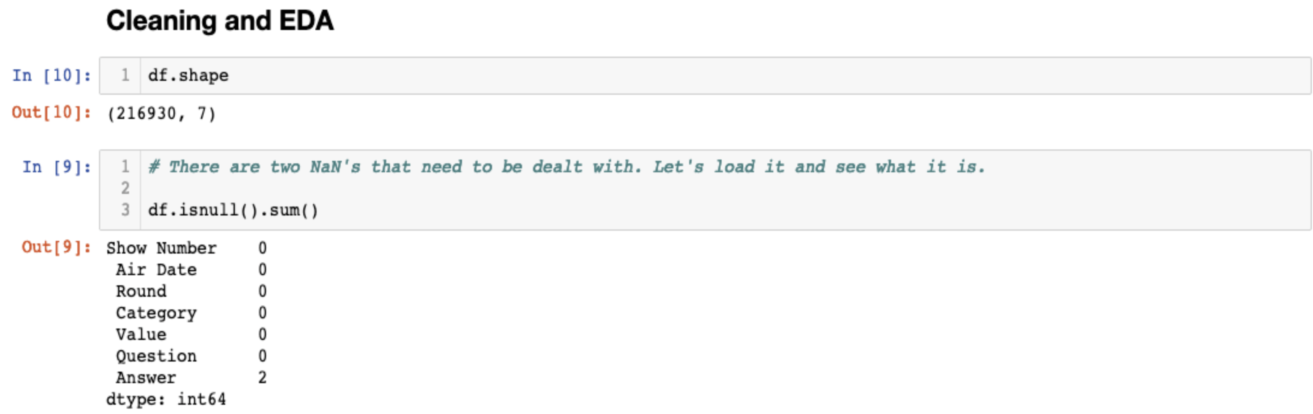
	Show Number	Air Date	Round	Category	Value	Question	Answer
0	4680	2004-12-31	Jeopardy!	HISTORY	$200	For the last 8 years of his life, Galileo was ...	Copernicus
1	4680	2004-12-31	Jeopardy!	ESPN's TOP 10 ALL-TIME ATHLETES	$200	No. 2: 1912 Olympian; football star at Carlisl...	Jim Thorpe
2	4680	2004-12-31	Jeopardy!	EVERYBODY TALKS ABOUT IT...	$200	The city of Yuma in this state has a record av...	Arizona
3	4680	2004-12-31	Jeopardy!	THE COMPANY LINE	$200	In 1963, live on "The Art Linkletter Show", th...	McDonald's
4	4680	2004-12-31	Jeopardy!	EPITAPHS & TRIBUTES	$200	Signer of the Dec. of Indep., framer of the Co...	John Adams

```
In [8]:
print(df.iloc[204][' Category'])
print(df.iloc[204][' Question'])
print(df.iloc[204][' Answer'])
```

```
THE EYES HAVE IT
People say these are what you need to make it in Hollywood
Contacts
```

Fortunately, the data looks pretty clean so far! One thing I’ve noticed right away that might cause a little inconvenience is how there’s an extra space in the column names. I will have to return to that issue and fix it. We’ll learn how to rename column variables soon. But first, let’s check the shape of our dataframe and check for null values.

Exploratory Data Analysis:



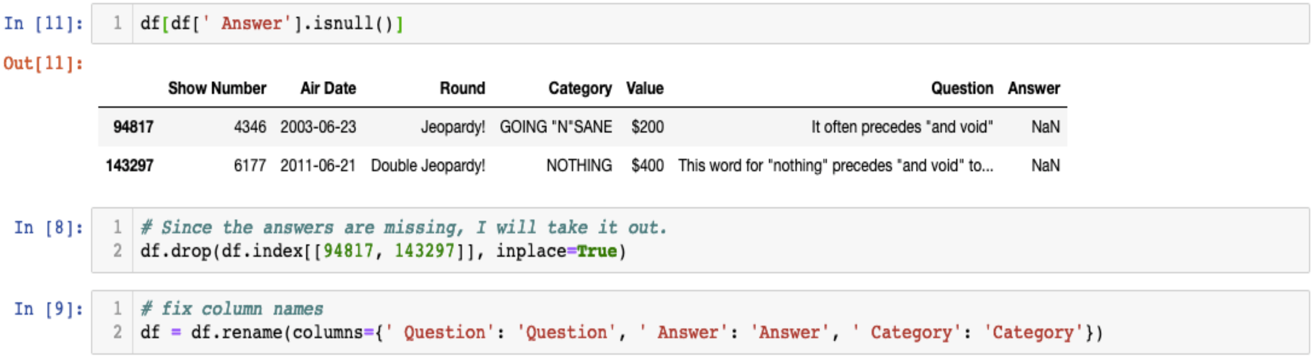
Cleaning and EDA

```
In [10]: df.shape
Out[10]: (216930, 7)
```

```
In [9]:  # There are two NaN's that need to be dealt with. Let's load it and see what it is.

         df.isnull().sum()
Out[9]:  Show Number    0
          Air Date      0
          Round         0
          Category      0
          Value         0
          Question      0
          Answer        2
         dtype: int64
```

It looks like the Answer column has two missing values. I'll load the two rows with the missing answers to see how I will deal with them.



```
In [11]: df[df[' Answer'].isnull()]
Out[11]:
```

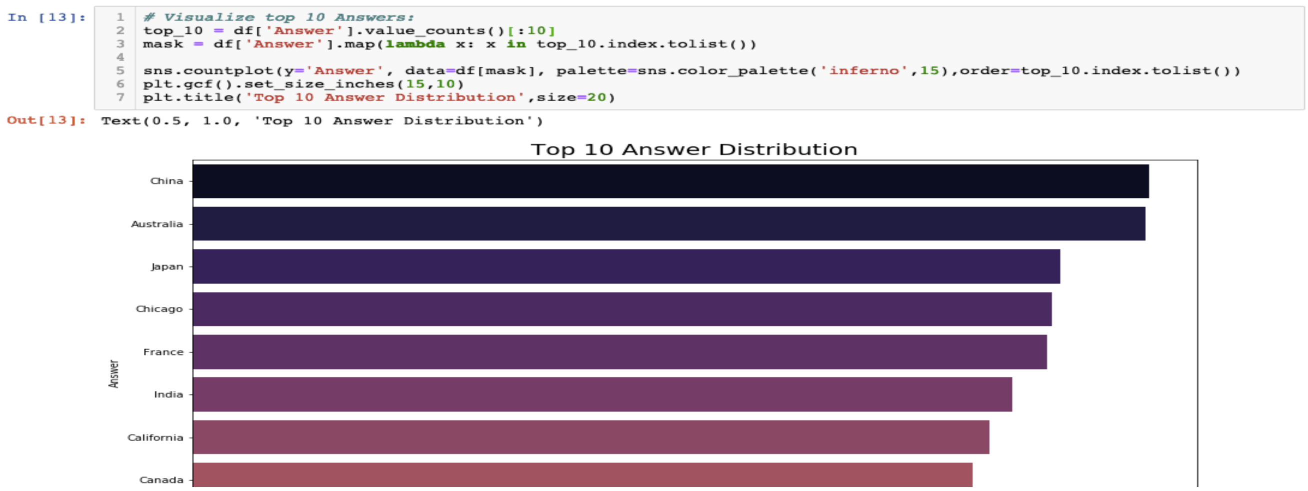
	Show Number	Air Date	Round	Category	Value	Question	Answer
94817	4346	2003-06-23	Jeopardy!	GOING "N"SANE	$200	It often precedes "and void"	NaN
143297	6177	2011-06-21	Double Jeopardy!	NOTHING	$400	This word for "nothing" precedes "and void" to...	NaN

```
In [8]:  # Since the answers are missing, I will take it out.
         df.drop(df.index[[94817, 143297]], inplace=True)
```

```
In [9]:  # fix column names
         df = df.rename(columns={' Question': 'Question', ' Answer': 'Answer', ' Category': 'Category'})
```

I've used a technique called "masking" to highlight the two missing rows that have null values. Once I can see the index number associated with the rows, I can decide what to do with them. In this case, since there are only two null values, I've decided to drop them (especially since they are text data and I don't know the answer to the questions)! Lastly, I've fixed the column names.

Now that the null values are taken care of, let's do some EDA to see what the top 10 most popular answers to Jeopardy questions are:

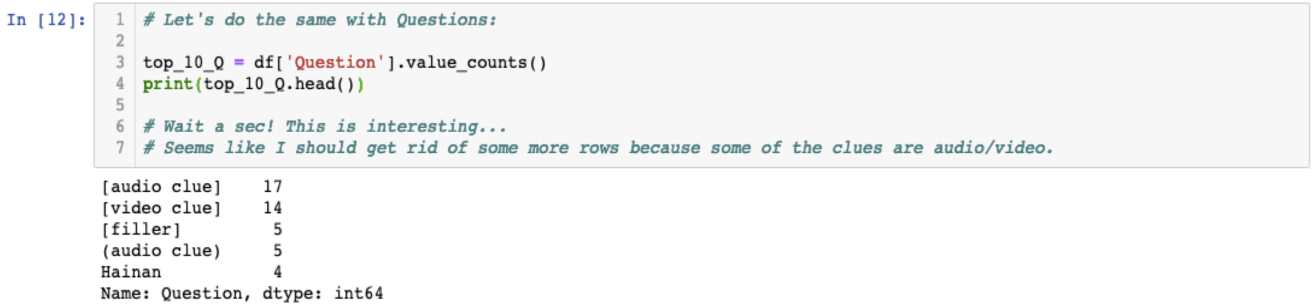


```
In [13]: # Visualize top 10 Answers:
         top_10 = df['Answer'].value_counts()[:10]
         mask = df['Answer'].map(lambda x: x in top_10.index.tolist())

         sns.countplot(y='Answer', data=df[mask], palette=sns.color_palette('inferno',15),order=top_10.index.tolist())
         plt.gcf().set_size_inches(15,10)
         plt.title('Top 10 Answer Distribution',size=20)
Out[13]: Text(0.5, 1.0, 'Top 10 Answer Distribution')
```

I’ve used a method called value_counts() to count the number of answers that are the same and stored in a variable called top_10. The ‘[:10]’ tacked on the end grabs the first 10 rows. I’ve created a mask that I will use to plug in Seaborn to create a horizontal bar chart to visualize the top 10 most common answers in Jeopardy.

For thoroughness' sake, let’s repeat the same process to see what the top 10 most popular questions are. I assume Alex wouldn’t reuse questions so I’m expecting everything to be one.

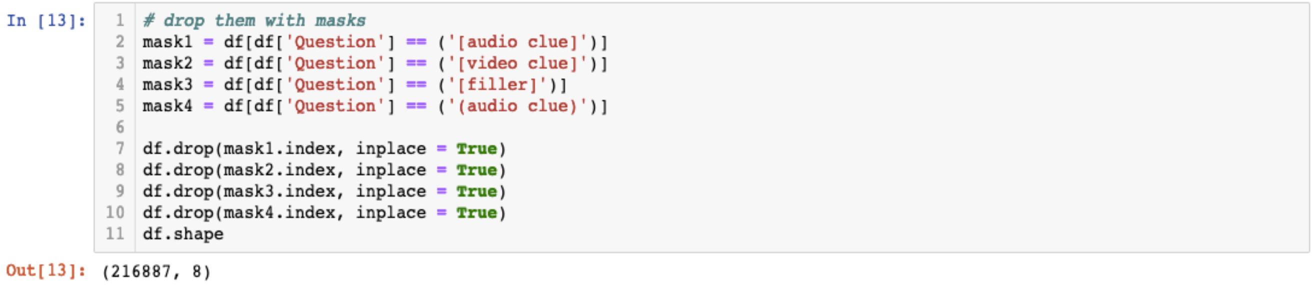


```
In [12]:
# Let's do the same with Questions:

top_10_Q = df['Question'].value_counts()
print(top_10_Q.head())

# Wait a sec! This is interesting...
# Seems like I should get rid of some more rows because some of the clues are audio/video.
```

```
[audio clue]    17
[video clue]    14
[filler]         5
(audio clue)     5
Hainan           4
Name: Question, dtype: int64
```



```
In [13]:
# drop them with masks
mask1 = df[df['Question'] == ('[audio clue]')]
mask2 = df[df['Question'] == ('[video clue]')]
mask3 = df[df['Question'] == ('[filler]')]
mask4 = df[df['Question'] == ('(audio clue)')]

df.drop(mask1.index, inplace = True)
df.drop(mask2.index, inplace = True)
df.drop(mask3.index, inplace = True)
df.drop(mask4.index, inplace = True)
df.shape
```

```
Out[13]: (216887, 8)
```

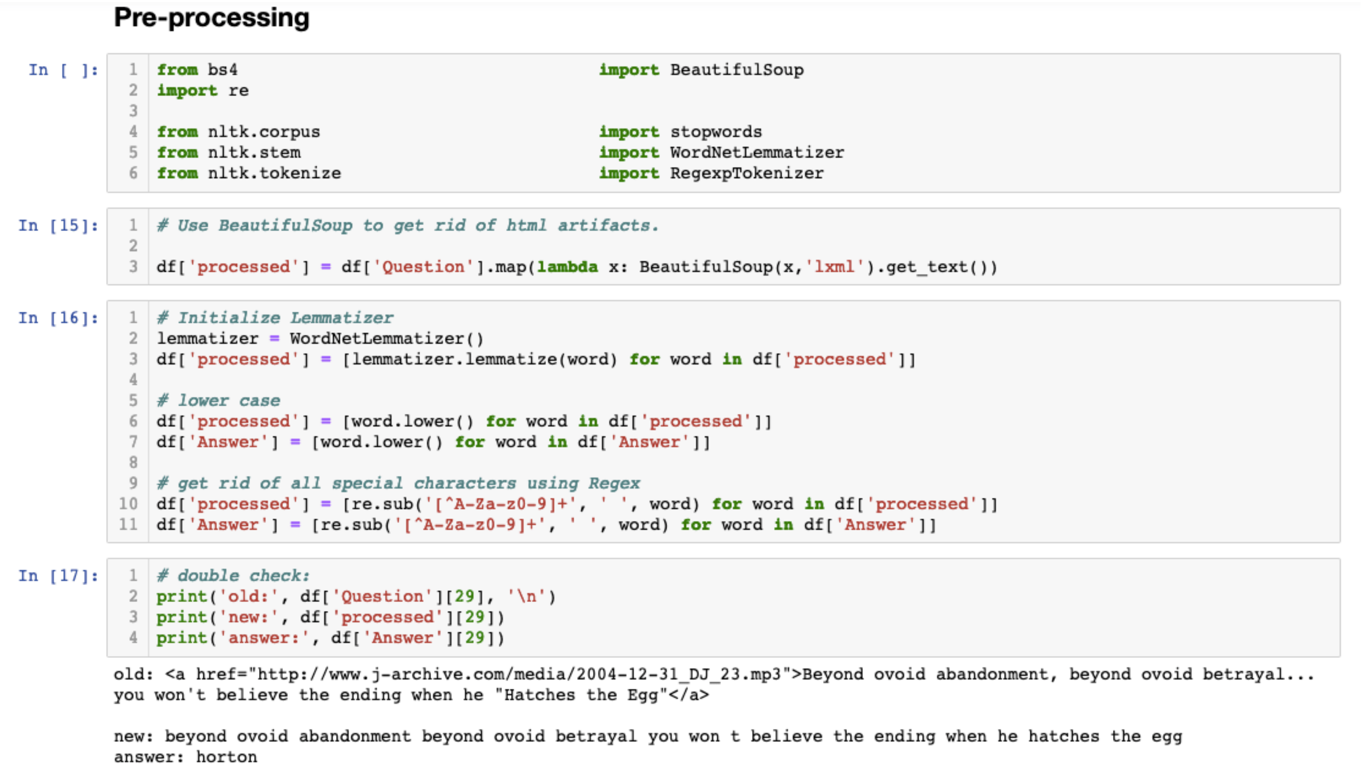
Aha! I’m glad I checked the questions and challenged my assumptions because there are quite a lot of audio and video clues! Since I will ultimately be using this data for classification and clustering, I should delete them from the dataset. I’ve gone ahead and listed them one by one instead of creating a loop for clarification's sake. This would’ve created some problems for our model if we weren’t paying attention!

Text Preprocessing:

While I was looking around the dataset, I realized some of the values had URLs and HTML tags embedded within texts. We will need to clean that up before we can analyze the data. We need to import a library called BeautifulSoup (BS4) to accomplish this. BS4 extracts text data from HTML and XML documents and it is a widely used package when dealing with scraped data.

Besides getting rid of HTML artifacts, we will need to get rid of punctuation marks in our texts, lowercase everything, and lemmatize words. For getting rid of punctuations, I’ll use regular expression (or regex as the cool kids call it) and for lower-casing everything, I’ll use Python’s built-in method.

Before we start munging our text data, I would like to highlight an important best practice. Instead of munging our original data, it is better to copy the data we’ll be using and create a new column from it. In this case, I’ve created a feature called “processed”.



Pre-processing

In [ ]:
```
from bs4                                 import BeautifulSoup
import re

from nltk.corpus                         import stopwords
from nltk.stem                           import WordNetLemmatizer
from nltk.tokenize                       import RegexpTokenizer
```

In [15]:
```
# Use BeautifulSoup to get rid of html artifacts.

df['processed'] = df['Question'].map(lambda x: BeautifulSoup(x,'lxml').get_text())
```

In [16]:
```
# Initialize Lemmatizer
lemmatizer = WordNetLemmatizer()
df['processed'] = [lemmatizer.lemmatize(word) for word in df['processed']]

# lower case
df['processed'] = [word.lower() for word in df['processed']]
df['Answer'] = [word.lower() for word in df['Answer']]

# get rid of all special characters using Regex
df['processed'] = [re.sub('[^A-Za-z0-9]+', ' ', word) for word in df['processed']]
df['Answer'] = [re.sub('[^A-Za-z0-9]+', ' ', word) for word in df['Answer']]
```

In [17]:
```
# double check:
print('old:', df['Question'][29], '\n')
print('new:', df['processed'][29])
print('answer:', df['Answer'][29])
```

```
old: <a href="http://www.j-archive.com/media/2004-12-31_DJ_23.mp3">Beyond ovoid abandonment, beyond ovoid betrayal...
you won't believe the ending when he "Hatches the Egg"</a>

new: beyond ovoid abandonment beyond ovoid betrayal you won t believe the ending when he hatches the egg
answer: horton
```

I’ve printed out the original and processed versions.

EDA is an important process even when you’re working with textual data and it helped us catch some interesting quirks we otherwise would’ve missed.

❖ CHAPTER- 4 (Introduction to machine learning and deep learning)

Machine Learning:

Machine learning is a field of study which allows machines(computers) to learn from data or experience and make a prediction based on the experience.

It enables the computers or the machines to make data-driven decisions rather than being explicitly programmed for carrying out a certain task. These programs or algorithms are designed in a way that they learn and improve over time when are exposed to new data.

Types of machine learning:

Machine learning can be broadly divided into 3 subcategories:

1. Supervised learning: In supervised learning, we have a labeled data containing input X and a label Y. In supervised learning, our task is to find the mapping between the input variable(X) called the independent variable and output variable(Y) called the dependent variable. Supervised learning can further be divided into two types of tasks:

1. **Regression** — Regression problem is when the output variable is continuous and real value. For example price, weight, etc.
2. **Classification** — Classification is a problem when the output variable is a category, such as “red” or “blue” or “disease” or “no disease”.

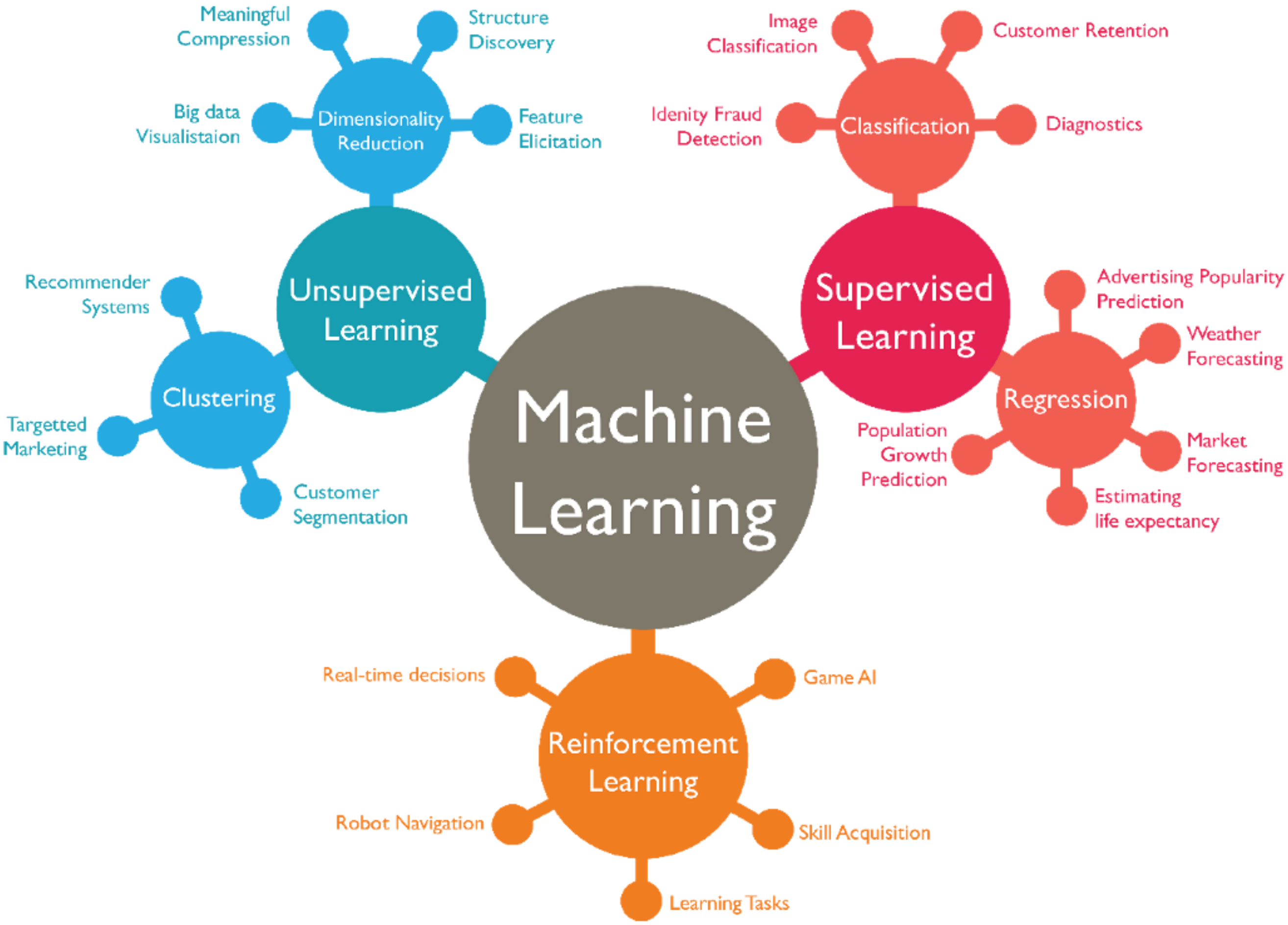
2. Unsupervised learning:

Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal of unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Unsupervised learning can further be divided into two types of tasks:

1. **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
2. **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

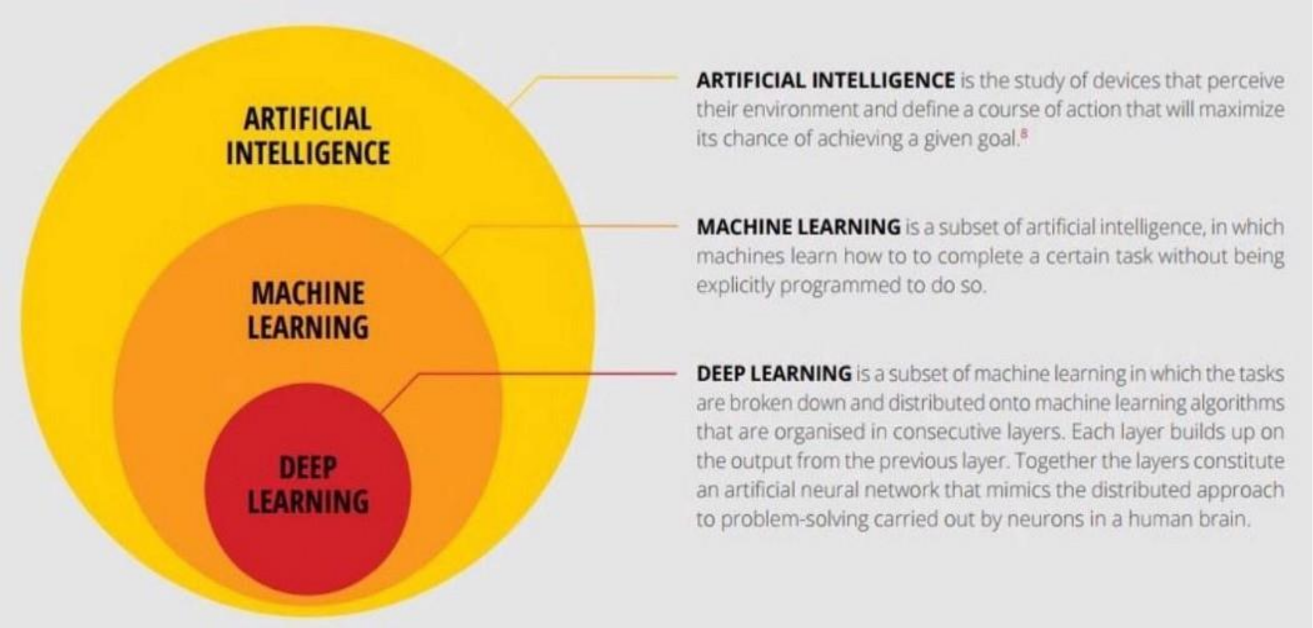
3. Reinforcement learning:

Reinforcement learning, in the context of artificial intelligence, is a type of dynamic programming that trains algorithms using a system of reward and punishment. A reinforcement learning algorithm, or agent, learns by interacting with its environment. The agent receives rewards by performing correctly and penalties for performing incorrectly. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty.



Machine learning categories with few applications

- **Deep Learning:** Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.



ARTIFICIAL INTELLIGENCE is the study of devices that perceive their environment and define a course of action that will maximize its chance of achieving a given goal.[8]

MACHINE LEARNING is a subset of artificial intelligence, in which machines learn how to to complete a certain task without being explicitly programmed to do so.

DEEP LEARNING is a subset of machine learning in which the tasks are broken down and distributed onto machine learning algorithms that are organised in consecutive layers. Each layer builds up on the output from the previous layer. Together the layers constitute an artificial neural network that mimics the distributed approach to problem-solving carried out by neurons in a human brain.

Applications of deep learning:

Applications of deep learning are vast and many of great technologies now use deep learning to improve the task. Some of the examples are:

1. Self-driving cars
2. Voice search and virtual assistants
3. Machine translation
4. Image caption generation
5. Colorization of Black and White Images
6. Game playing ai(Open Ai dota bot, google brain alpha go).
7. Real-time object recognition in the image (Google lens).

ML (Classification vs Regression):

Classification and Regression are two major prediction problems that are usually dealt with in Data mining and machine learning.

Classification is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in input and then the labels are predicted for the data.

The derived mapping function could be demonstrated in the form of "IF-THEN" rules. The classification process deal with the problems where the data can be divided into binary or multiple discrete labels.

Let's take an example, suppose we want to predict the possibility of the winning of a match by Team A on the basis of some parameters recorded earlier. Then there would be two labels Yes and No.

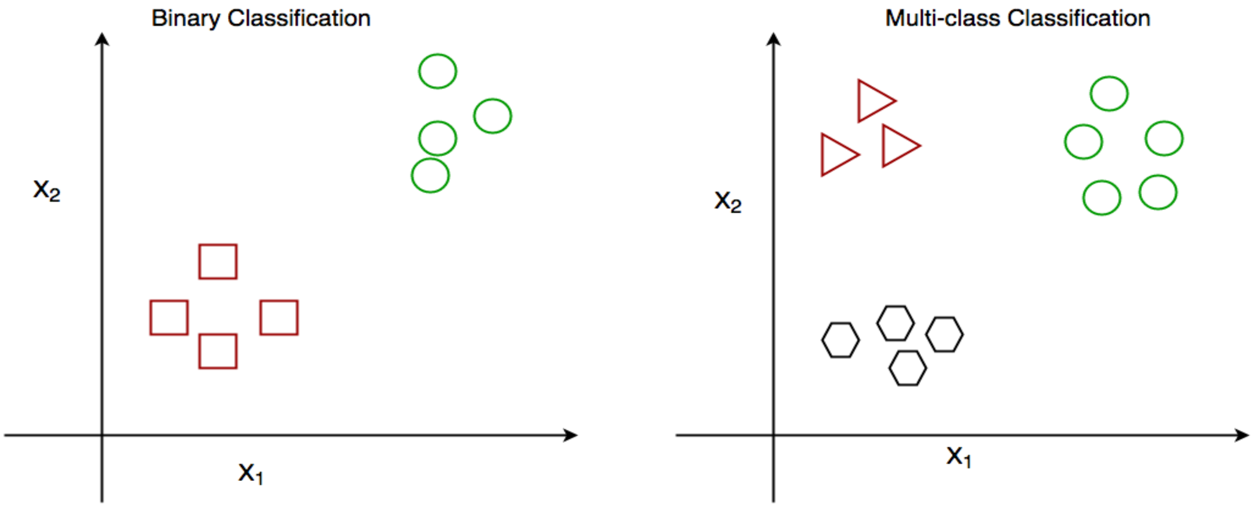


Fig: Binary Classification and Multiclass Classification

Regression is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions.

Let's take a similar example in regression also, where we are finding the possibility of rain in some particular regions with the help of some parameters recorded earlier. Then there is a probability associated with the rain.

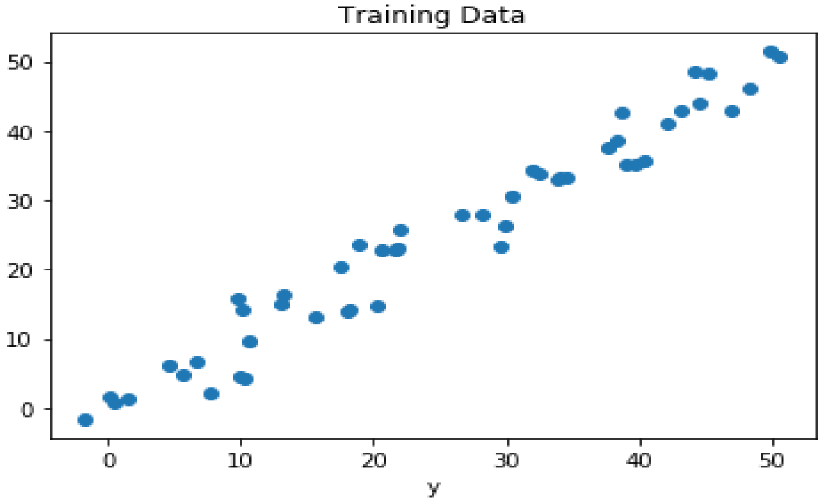


Fig: Regression of Day vs Rainfall (in mm)

ML (Underfitting and Overfitting):

Let us consider that we are designing a machine learning model. A model is said to be a good machine learning model if it generalizes any new input data from the problem domain in a proper way. This helps us to make predictions in the future data, that the data model has never seen. Now, suppose we want to check how well our machine learning model learns and generalizes to the new data. For that, we have overfitting and underfitting, which are majorly responsible for the poor performances of the machine learning algorithms.

Before diving further let's understand two important terms:
– **Bias:** Assumptions made by a model to make a function easier to learn.
– **Variance:** If you train your data on training data and obtain a very low error, upon changing the data and then training the same previous model you experience a high error, this is variance.

Underfitting:

A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. *(It's just like trying to fit undersized pants!)* Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have less data to build an accurate model and also when we try to build a linear model with fewer non-linear data. In such cases, the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

In a nutshell, **Underfitting – High bias and low variance**

Techniques to reduce underfitting:

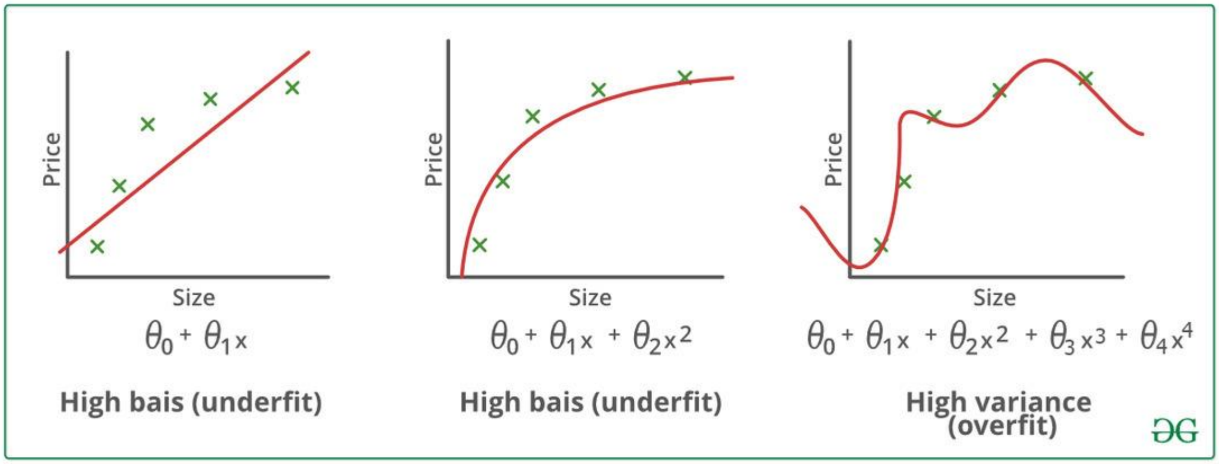
1. Increase model complexity
2. Increase the number of features, performing feature engineering
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results.

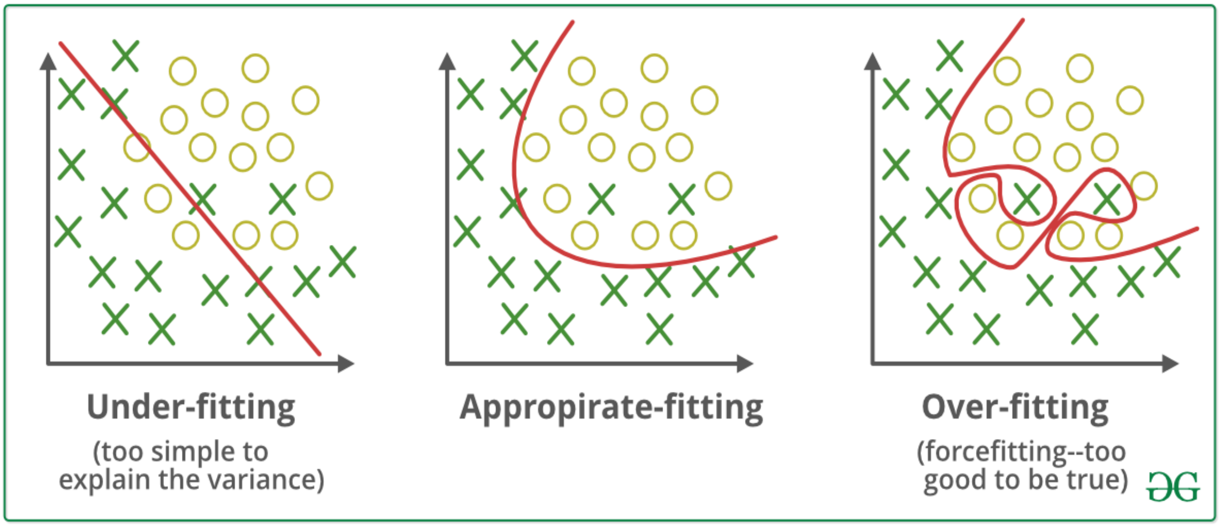
Overfitting:

A statistical model is said to be overfitted when we train it with a lot of data *(just like fitting ourselves in oversized pants!)*. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

In a nutshell, **Overfitting – High variance and low bias**

Examples:




Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Techniques to reduce overfitting:

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.

Good Fit in a Statistical Model:

Ideally, the case when the model makes the predictions with 0 error, is said to have a *good fit* on the data. This situation is achievable at a spot between overfitting and underfitting. In order to understand it we will have to look at the performance of our model with the passage of time, while it is learning from training dataset.

With the passage of time, our model will keep on learning and thus the error for the model on the training and testing data will keep on decreasing. If it will learn for too long, the model will become more prone to overfitting due to the presence of noise and less useful details. Hence the performance of our model will decrease. In order to get a good fit, we will stop at a point just before where the error starts increasing. At this point, the model is said to have good skills on training datasets as well as our unseen testing dataset.

❖ CHAPTER- 5 (CONCLUSION)

Outcomes:

- Have a good understanding of the fundamental issues and challenges of machine learning: data, model selection, model complexity, etc.

- Have an understanding of the strengths and weaknesses of many popular machine learning approaches.

- Appreciate the underlying mathematical relationships within and across Machine Learning algorithms and the paradigms of supervised and un-supervised learning.

- Be able to design and implement various machine learning algorithms in a range of real-world applications.

- Ability to integrate machine learning libraries and mathematical and statistical tools with modern technologies.

❖ REFERENCES

- ✓ PROBABILITY AND MATHEMATICAL STATISTICS by Prasanna Sahoo Department of Mathematics University of Louisville, KY 40292 USA
- ✓ FUNDAMENTALS OF MATHEMATICAL STATISTICS by S.C. GUPTA V.K. KAPOOR
- ✓ An Introduction to the Science of Statistics: From Theory to Implementation Preliminary Edition Joseph C. Watkins
- ✓ lasseschultebraucks.com/overfitting-underfitting-ml/
  - ✓ chunml.github.io/ChunML.github.io/tutorial/Underfit-Overfit/
  - ✓ geeksforgeeks.com
  - ✓ youtube.com
  - ✓ google.com
  - ✓ en.wikipedia.org/wiki/Machine_learning
  - ✓ youtube.com/playlist?list=PL2HX_yT71umDkEQdfKFnqJFxSpjMG-Vw5

---:END OF THE REPORT:---