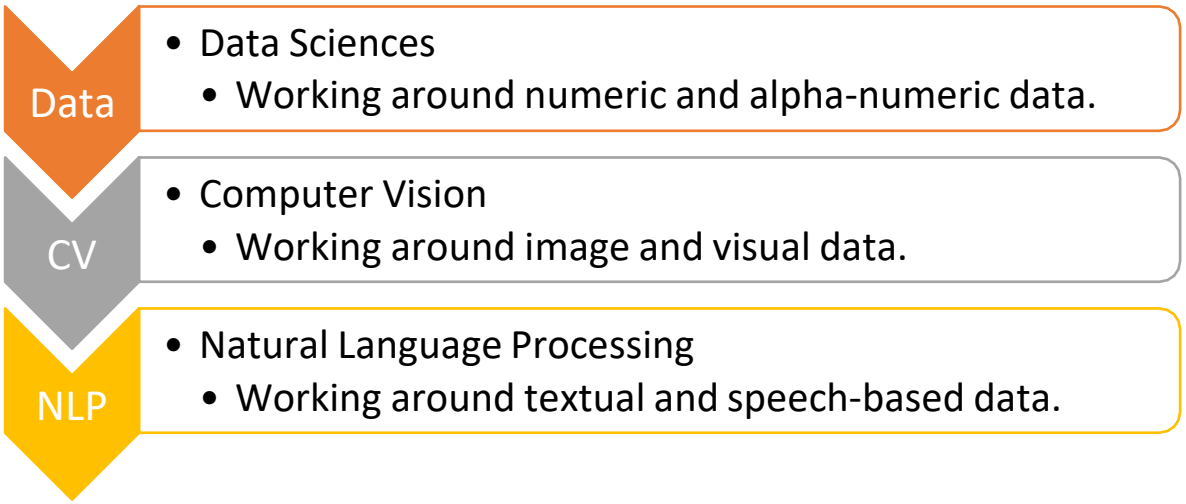
Data Sciences

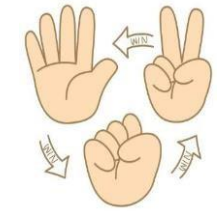
Introduction

As we have discussed earlier in class 9, Artificial Intelligence is a technology which completely depends on data. It is the data which is fed into the machine which makes it intelligent. And depending upon the type of data we have; AI can be classified into three broad domains:



Each domain has its own type of data which gets fed into the machine and hence has its own way of working around it. Talking about Data Sciences, it is a concept to unify statistics, data analysis, machinelearning and their related methods in order to understand and analyse actual phenomena with data. It employs techniques and theories drawn from many fields within the context of Mathematics, Statistics, Computer Science, and Information Science.

Now before we get into the concepts of Data Sciences, let us experience this domain with the help ofthe following game:



*** Rock, Paper & Scissors:**

Go to this link and try to play the game of Rock, Paper Scissors against an AI model. The challenge hereis to win 20 games against AI before AI wins them against you.

Did you manage to win?

What was the strategy that you applied to win this game against the AI machine?

_____
_____
_____
_____

Was it different playing Rock, Paper & Scissors with an AI machine as compared to a human?

_____
_____
_____
_____

What approach was the machine following while playing against you?

_____
_____
_____
_____

Applications of Data Sciences

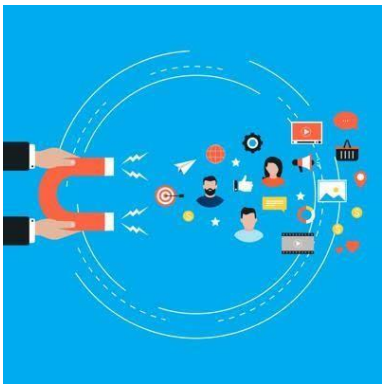
Data Science is not a new field. Data Sciences majorly work around analysing the data and when it comes to AI, the analysis helps in making the machine intelligent enough to perform tasks by itself. There exist various applications of Data Science in today's world. Some of them are:

Fraud and Risk Detection*: The earliest applications of data science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them from losses.

Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyse the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

Genetics & Genomics*: Data Science applications also enab ment personalization through research in genetics and genomics. Th pact of the DNA on our health and find individual biologica etics, diseases, and drug response. Data science techniques allow in data with genomic data in disease research, which provides a deep ssues in reactions to particular drugs and diseases. As soon as we ome data, we will achieve a deeper understanding of the human risk prediction will be a major step towards more individual care.

Internet Search*: When we talk about search engines, we think 'Google'. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL, and so on. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in the fraction of a second. Considering the fact that Google processes more than 20 petabytes of data every day, had there been no data science, Google wouldn'thave been the 'Google' we know today.



Targeted Advertising*: If you thought Search would have been the biggest of all data science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banilrs on various websites to the digital billboards at the airports
– almost all of them are decided by using data science algorithms. This is the reason why digital ads have been able to get a much higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user's past behaviour.

Website Recommendations:* Aren't we all used to the suggestions about similar products on Amazon? They not onlyhelp us find relevant products from billions of productsavailable with them but also add a lot to the user experience. A lot of companies have fervidly used this engine to promotetheir products in accordance with the user's interest andrelevance of information. Internet giants like Amazon, Twitter,Google Play, Netflix, LinkedIn, IMDB and many more use thissystem to improve the user experience. The recommendationsare made based on previous search results for a user.

Airline Route Planning*: The Airline Industry across the world is known to bear heavy losses. Except for a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air-fuel prices and the need to offer heavy discounts to customers, the situation has got worse. It wasn't long before airline companies started using
Data Science to identify the strategic areas of improvements. Now, while using Data Science, the airline companies can:

- Predict flight delay
- Decide which class of airplanes to buy
- Whether to directly land at the destination or take a halt in between (For example, A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
- Effectively drive customer loyalty programs

Getting Started

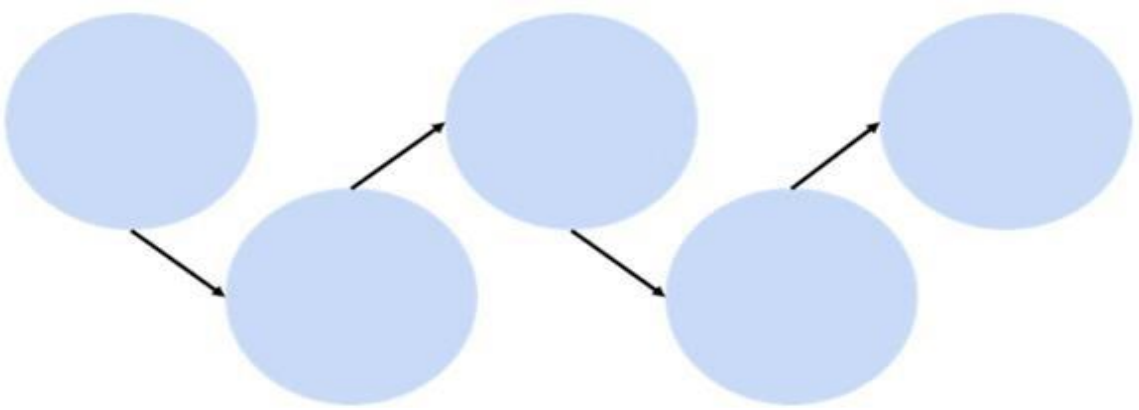
Data Sciences is a combination of Python and Mathematical concepts like Statistics, Data Analysis, probability, etc. Concepts of Data Science can be used in developing applications around AI as it gives a strong base for data analysis in Python.

Revisiting AI Project Cycle

But, before we get deeper into data analysis, let us recall how Data Sciences can be leveraged to solve some of the pressing problems around us. For this, let us understand the AI project cycle framework around Data Sciences with the help of an example.

Do you remember the AI Project Cycle? Fill in

all the stages of the cycle here:



*The Scenario**

Humans are social animals. We tend to organise and/or participate in various kinds of social gatherings all the time. We love eating out with friends and family because of which we can find restaurants almost everywhere and out of these, many of the restaurants arrange for buffets to offer a variety of food items to their customers. Be it small shops or big outlets, every restaurant prepares food in bulk as they expect

a good crowd to come and enjoy their food. But in most cases, after the day ends, a lot of food is left which becomes unusable for the restaurant as they do not wish to serve stale food to their customers the next day. So, every day, they prepare food in large quantities keeping in mind the probable number of customers walking into their outlet. But if the expectations are not met, a good amount of food gets wasted which eventually becomes a loss for the restaurant as they either have to dump it or give it to hungry people for free. And if this daily loss is taken into account for a year, it becomes quite a big amount.

Problem Scoping

Now that we have understood the scenario well, let us take a deeper look into the problem to find out more about various factors around it. Let us fill up the 4Ws problem canvas to find out.

Who Canvas – Who is having the problem?

Who are the stakeholders?	○ Restaurants offering buffets ○ Restaurant Chefs
What do we know about them?	○ Restaurants cook food in bulk every day for their buffets to meet their customer needs. ○ They estimate the number of customers that would walk into their restaurant every day.

What Canvas – What is the nature of their problem?

What is the problem?	○ Quite a large amount of food is leftover everyday unconsumed at the restaurant which is either thrown away or given for free to needy people. ○ Restaurants have to bear everyday losses for the unconsumed food.
How do you know it is a problem?	o Restaurant Surveys have shown that restaurants face this problem of food waste.

What is the context/situation in which the stakeholders experience this problem?	○ Restaurants which serve buffet food ○ At the end of the day, when no further food consumption is possible

Where Canvas – Where does the problem arise?

Why? – Why do you think it is a problem worth solving?

What would be of key value to the stakeholders?	o If the restaurant has a proper estimate of the quantity of food to be prepared every day, the food waste can be reduced.
How would it improve their situation?	○ Less or no food would be left unconsumed. ○ Losses due to unconsumed food would reduce considerably.

Now that we have noted down all the factors around our problem, let us fill up the problem statement template.

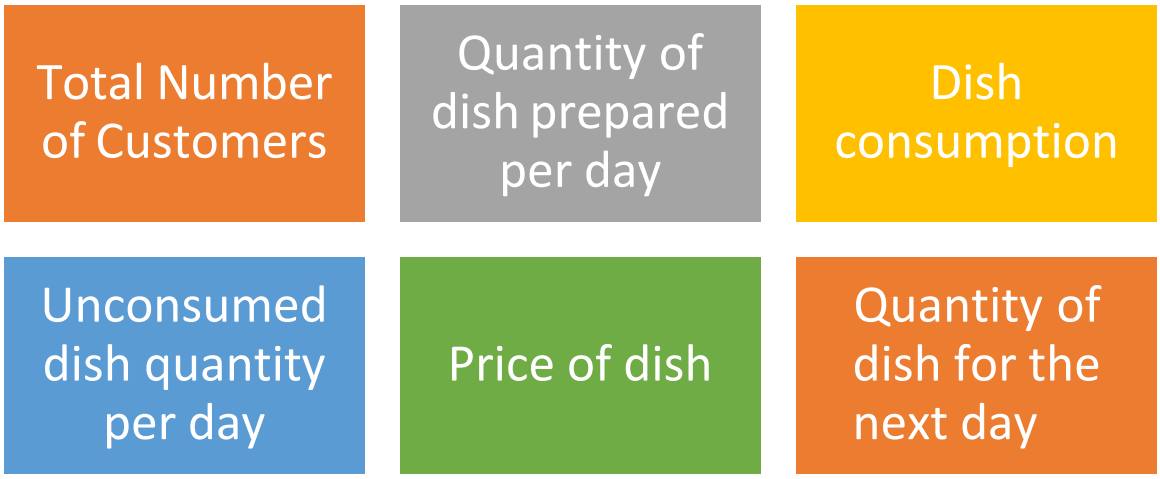
Our	**Restaurant Owners**	**Who?**
Have a problem of	Losses due to food wastage	**What?**
While	The food is left unconsumed due to improper estimation	**Where?**
ideal solution would	Be to be able to predict the amount of food to be prepared for every day consumption	**Why**

The Problem statement template leads us towards the goal of our project which can now be statedas:

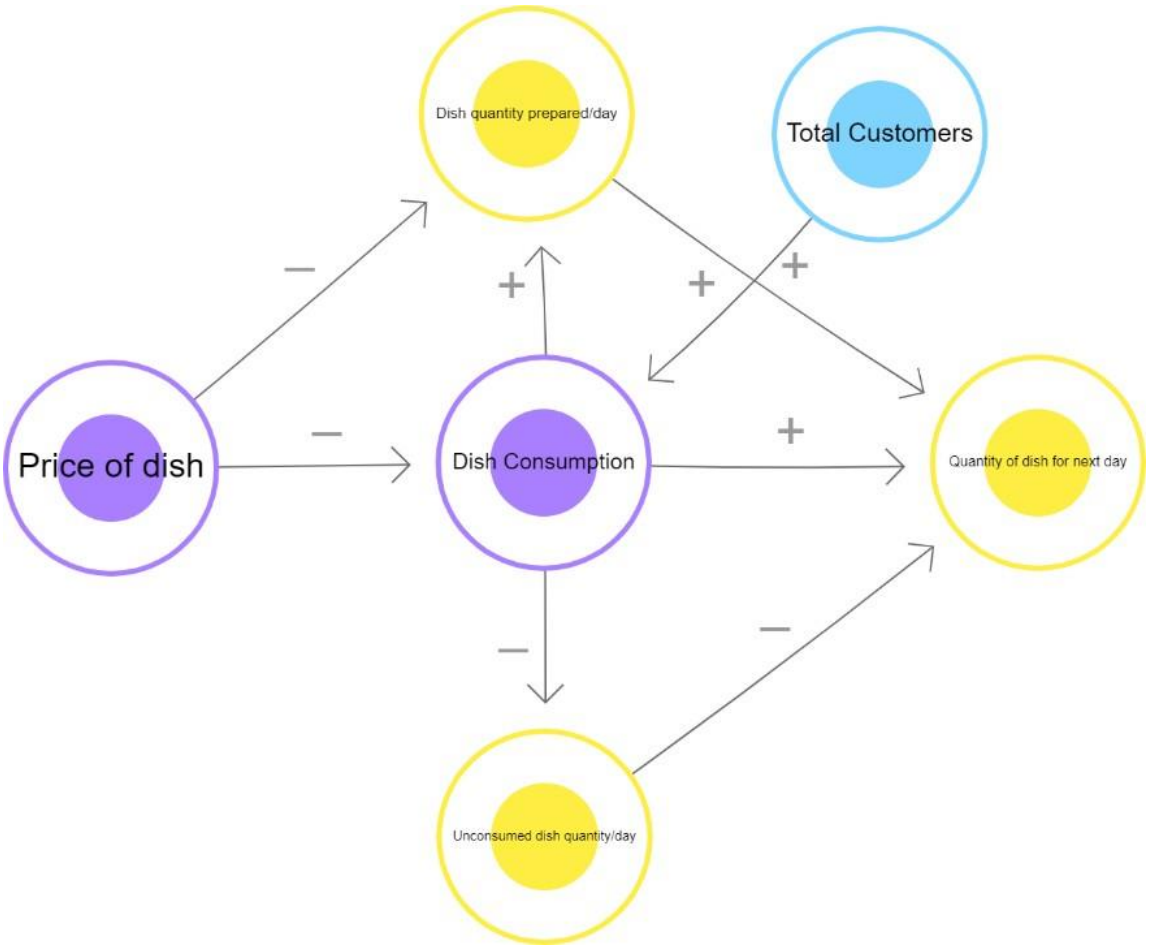
"To be able to predict the quantity of food dishes to be prepared for everyday consumption in restaurant buffets."

Data Acquisition

After finalising the goal of our project, let us now move towards looking at various data features which affect the problem in some way or the other. Since any AI-based project requires data for testing and training, we need to understand what kind of data is to be collected to work towards the goal. In our scenario, various factors that would affect the quantity of food to be prepared for the next day consumption in buffets would be:



Now let us understand how these factors are related to our problem statement. For this, we can use the System Maps tool to figure out the relationship of elements with the project's goal. Here is the System map for our problem statement.



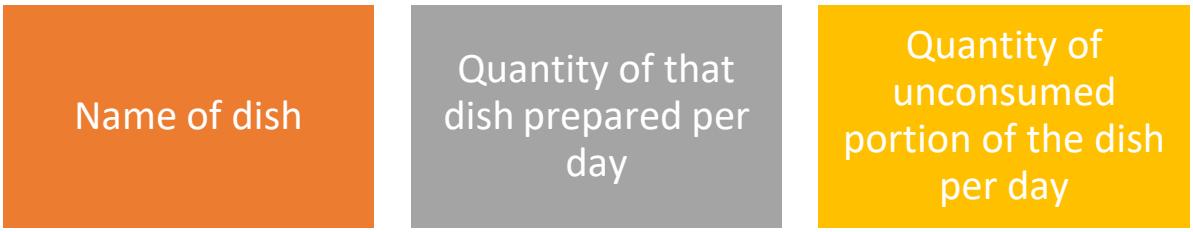
In this system map, you can see how the relationship of each element is defined with the goal of our project. Recall that the positive arrows determine a direct relationship of elements while the negative ones show an inverse relationship of elements.

After looking at the factors affecting our problem statement, now it's time to take a look at the data which is to be acquired for the goal. For this problem, a dataset covering all the elements mentioned above is made for each dish prepared by the restaurant over a period of 30 days. This data is collected offline in the form of a regular survey since this is a personalised dataset created just for one restaurant's needs.

Specifically, the data collected comes under the following categories: Name of the dish, Price of the dish, Quantity of dish produced per day, Quantity of dish left unconsumed per day, Total number of customers per day, Fixed customers per day, etc.

Data Exploration

After creating the database, we now need to look at the data collected and understand what is required out of it. In this case, since the goal of our project is to be able to predict the quantity of food to be prepared for the next day, we need to have the following data:



Thus, we extract the required information from the curated dataset and clean it up in such a way that

there exist no errors or missing elements in it.

Modelling

Once the dataset is ready, we train our model on it. In this case, a regression model is chosen in whichthe dataset is fed as a dataframe and is trained accordingly. Regression is a Supervised Learning modelwhich takes in continuous values of data over a period of time. Since in our case the data which we have is a continuous data of 30 days, we can use the regression model so that it predicts the next values to it in a similar manilr. In this case, the dataset of 30 days is divided in a ratio of 2:1 for training and testing respectively. In this case, the model is first trained on the 20-day data and then gets evaluated for the rest of the 10 days.

Evaluation

Once the model has been trained on the training dataset of 20 days, it is now time to see if the modelis working properly or not. Let us see how the model works and how is it tested.

Step 1: The trained model is fed data regards the name of the dish and the quantity produced for the same.

Step 2: It is then fed data regards the quantity of food left unconsumed for the same dish on previous occasions.

Step 3: The model then works upon the entries according to the training it got at the modelling stage.
Step 4: The Model predicts the quantity of food to be prepared for the next day.

Step 5: The prediction is compared to the testing dataset value. From the testing dataset, ideally, we can say that the quantity of food to be produced for next day's consumption should be the total quantity minus the unconsumed quantity.

Step 6: The model is tested for 10 testing datasets kept aside while training.

Step 7: Prediction values of testing dataset is compared to the actual values.

Step 8: If the prediction value is same or almost similar to the actual values, the model is said to be accurate. Otherwise, either the model selection is changed or the model is trained on more data for better accuracy.

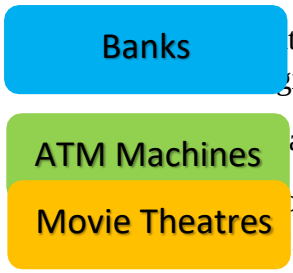
Once the model is able to achieve optimum efficiency, it is ready to be deployed in the restaurant for real-time usage.

Data Collection

Data collection is nothing new which has come up in our lives. It has been in our society since ages. Even when people did not have fair knowledge of calculations, records were still maintained in some way or the other to keep an account of relevant things. Data collection is an exercise which does not require even a tiny bit of technological knowledge. But when it comes to analysing the data, it becomes a tedious process for humans as it is all about numbers and alpha-numerical data. That is where Data Science comes into the picture. It not only gives us a clearer idea around the dataset, butalso adds value to it by providing deeper and clearer analyses around it. And as AI gets incorporated in the process, predictions and suggestions by the machine become possible on the same.

Now that we have gone through an example of a Data Science based project, we have a bit of clarity regarding the type of data that can be used to develop a Data Science related project. For the data domain-based projects, majorly the type of data used is in numerical or alpha-numerical format and such datasets are curated in the form of tables. Such databases are very commonly found in any institution for record maintenance and other purposes. Some examples of datasets which you must

already be aware of are:

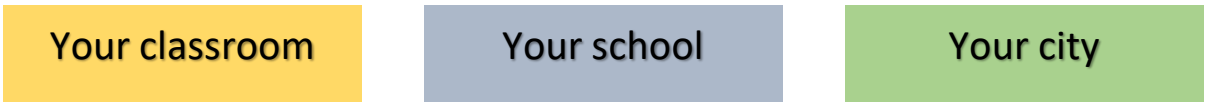


tabases of loans issued, account holder, locker owners, employee gistrations, bank visitors, etc.

age details per day, cash denominations transaction details, visitordetails, etc.

ovie details, tickets sold offline, tickets sold online, refreshmentpurchases, .

Now look around you and find out what are the different types of databases which are maintained inthe places mentioned below. Try surveying people who are responsible for the designated places to get a better idea.



As you can see, all the type of data which has been mentioned above is in the form of tables. Tables which contain numeric or alpha-numeric data. But this leads to a very critical dilemma: are these datasets accessible to all? Should these databases be accessible to all? What are the various sources of data from which we can gather such databases? Let's find out!

Sources of Data

There exist various sources of data from where we can collect any type of data required and the data collection process can be categorised in two ways: Offline and Online.

Offline Data Collection	Online Data Collection
Sensors	Open-sourced Government Portals
Surveys	Reliable Websites (Kaggle)
Interviews	World Organisations' open-sourced statistical websites
Observations	

While accessing data from any of the data sources, following points should be kept in mind:

1. Data which is available for public usage only should be taken up.
2. Personal datasets should only be used with the consent of the owner.
3. One should never breach someone's privacy to collect data.
4. Data should only be taken form reliable sources as the data collected from random sourcescan be wrong or unusable.
5. Reliable sources of data ensure the authenticity of data which helps in proper training of theAI model.

Types of Data

For Data Science, usually the data is collected in the form of tables. These tabular datasets can be stored in different formats. Some of the commonly used formats are:

1. CSV: CSV stands for comma separated values. It is a simple file format used to store tabular data. Each line of this file is a data record and reach record consists of one or more fields whichare

separated by commas. Since the values of records are separated by a comma, hence they are known as CSV files.

2. Spreadsheet: A Spreadsheet is a piece of paper or a computer program which is used for accounting and recording data using rows and columns into which information can be entered. Microsoft excel is a program which helps in creating spreadsheets.
3. SQL: SQL is a programming language also known as Structured Query Language. It is a domain-specific language used in programming and is designed for managing data held in different kinds of DBMS (Database Management System) It is particularly useful in handling structured data.

A lot of other formats of databases also exist, you can explore them online!

Data Access

After collecting the data, to be able to use it for programming purposes, we should know how to accessthe same in a Python code. To make our lives easier, there exist various Python packages which help us in accessing structured data (in tabular form) inside the code. Let us take a look at some of this package.

NumPy

NumPy, which stands for Numerical Python, is the fundamental package for Mathematical and logical operations on arrays in Python. It is a commonly used package when it comes to working around numbers. NumPy gives a wide range of arithmetic operations around numbers giving us an easier approach in working with them. NumPy also works with arrays, which is nothing but a homogenous collection of Data.

An array is nothing but a set of multiple values which are of same datatype. They can be numbers, characters, booleans, etc. but only one datatype can be accessed through an array. In NumPy, the arrays used are known as ND-arrays (N-Dimensional Arrays) as NumPy comes with a feature of creating n-dimensional arrays in Python.

An array can easily be compared to a list. Let us take a look at how they are different:

NumPy Arrays	Lists
1. Homogenous collection of Data. 2. Can contain only one type of data, hence not flexible with datatypes. 3. Cannot be directly initialized. Can be operatedwith Numpy package only. 4. Direct numerical operations can be done. For example, dividing the whole array by 3 divides every element by 3. 5. Widely used for arithmetic operations. 6. Arrays take less memory space. 7. Functions like concatenation, appending, reshaping, etc are not trivially possible with arrays. 8. Example: To create a numpy array 'A': **import numpy** **A=numpy.array([1,2,3,4,5,6,7,8,9,0])**	1. Heterogenous collection of Data. 2. Can contain multiple types of data, hence flexible with datatypes. 3. Can be directly initialized as it is a partof Python syntax. 4. Direct numerical operations are not possible. For example, dividing the whole list by 3 cannot divide every element by 3. 5. Widely used for data management. 6. Lists acquire more memory space. 7. Functions like concatenation, appending, reshaping, etc are trivially possible with lists. 8. Example: To create a list: **A = [1,2,3,4,5,6,7,8,9,0]**

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. The name is derived from the term "panel data", an econometrics term for data sets that

include observations over multiple time periods for the same individuals.

Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a Pandas data structure

The two primary data structures of Pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. Pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Here are just a few of the things that pandas do well:

- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations
- Intelligent label-based slicing, fancy indexing, and subsetting of large data sets
- Intuitive merging and joining data sets
- Flexible reshaping and pivoting of data sets

*Matplotlib**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some types of graphs that we can make with this package are listed below:

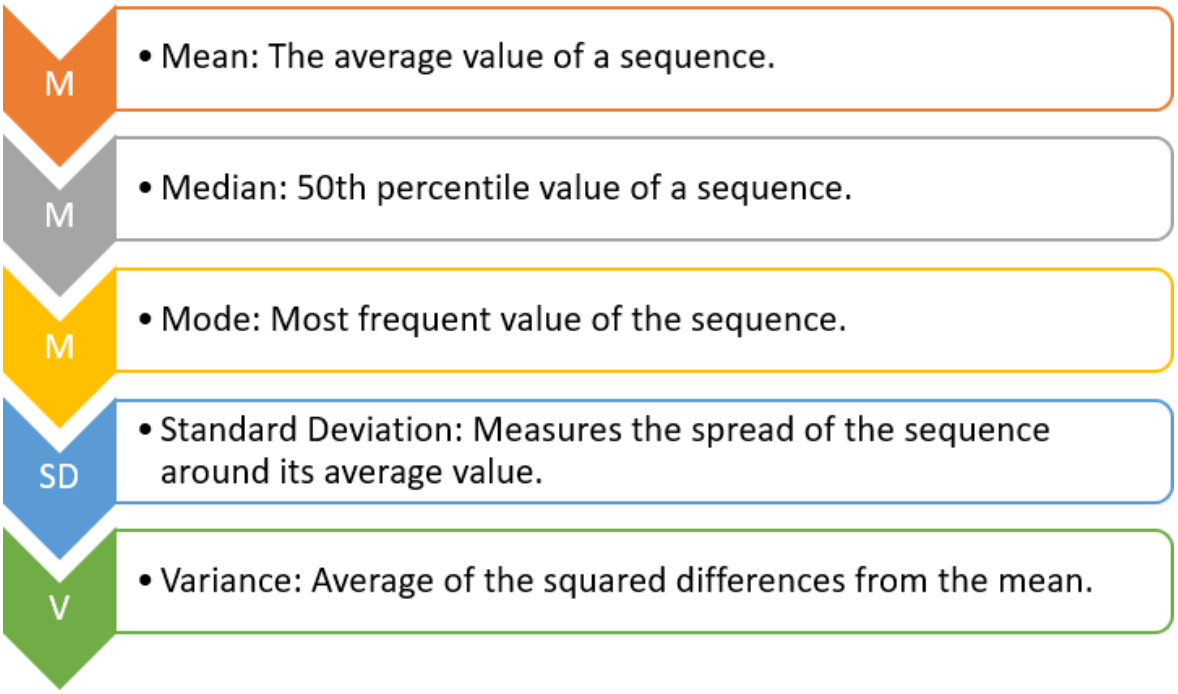
Not just plotting, but you can also modify your plots the way you wish. You can stylise them and make them more descriptive and communicable.

These packages help us in accessing the datasets we have and also in exploring them to develop a better understanding of them

* Images shown here are the property of individual organizations and are used here for reference purpose only.

Basic Statistics with Python

We have already understood that Data Sciences works around analysing data and performing tasks around it. For analysing the numeric & alpha-numeric data used for this domain, mathematics comesto our rescue. Basic statistical methods used in mathematics come quite hAmanin Python too for analysing and working around such datasets. Statistical tools widely used in Python are:



Do you remember using these formulas in your class? Let us recall all of them here:

1. What is Mean? How is it calculated?

2. What is Median? How is it calculated?

3. What is Mode? How is it calculated?

4. What is Standard Deviation? How is it calculated?

5. What is Variance? How is it calculated?

Advantage of using Python packages is that we do not need to make our own formula or equation to find out the results. There exist a lot of pre-defined functions with packages like NumPy which reduces this trouble for us. All we need to do is write that function and pass on the data to it. It's that simple!

Let us take a look at various Python syntaxes that can help us with the statistical work in data analysis. Head to the Jupyter Notebook of Basic statistics with Python and start exploring! You may find the Jupyter notebook here:

Data Visualisation

While collecting data, it is possible that the data might come with some errors. Let us first take a look at the types of issues we can face with data:

1. Erroneous Data: There are two ways in which the data can be erroneous:

- Incorrect values: The values in the dataset (at random places) are incorrect. For example, in the column of phone number, there is a decimal value or in the marks column, there is a name mentioned, etc. These are incorrect values that do not resemble the kind of data expected in that position.
- Invalid or Null values: At some places, the values get corrupted and hence they become invalid. Many times you will find NaN values in the dataset. These are null values which do not hold any meaning and are not processible. That is why, these values (as and when encountered) are removed from the database.

2. Missing Data: In some datasets, some cells remain empty. The values of these cells are missing and hence the cells remain empty. Missing data cannot be interpreted as an error as the values here are not erroneous or might not be missing because of any error.

3. Outliers: Data which does not fall in the range of a certain element are referred to as outliers. To understand this better, let us take an example of marks of students in a class. Let us assume that a student was absent for exams and hence has got 0 marks in it. If his marks are taken into account, the whole class's average would go down. To prevent this, the average is taken for the range of marks from highest to lowest keeping this particular result separate. This makes sure that the average marks of the class are true according to the data.

Machines work efficiently on numbers, humans need visual aid to understand and comprehend the information passed. Hence, data visualization is used to interpret the data collected and identify patterns and trends out of it.

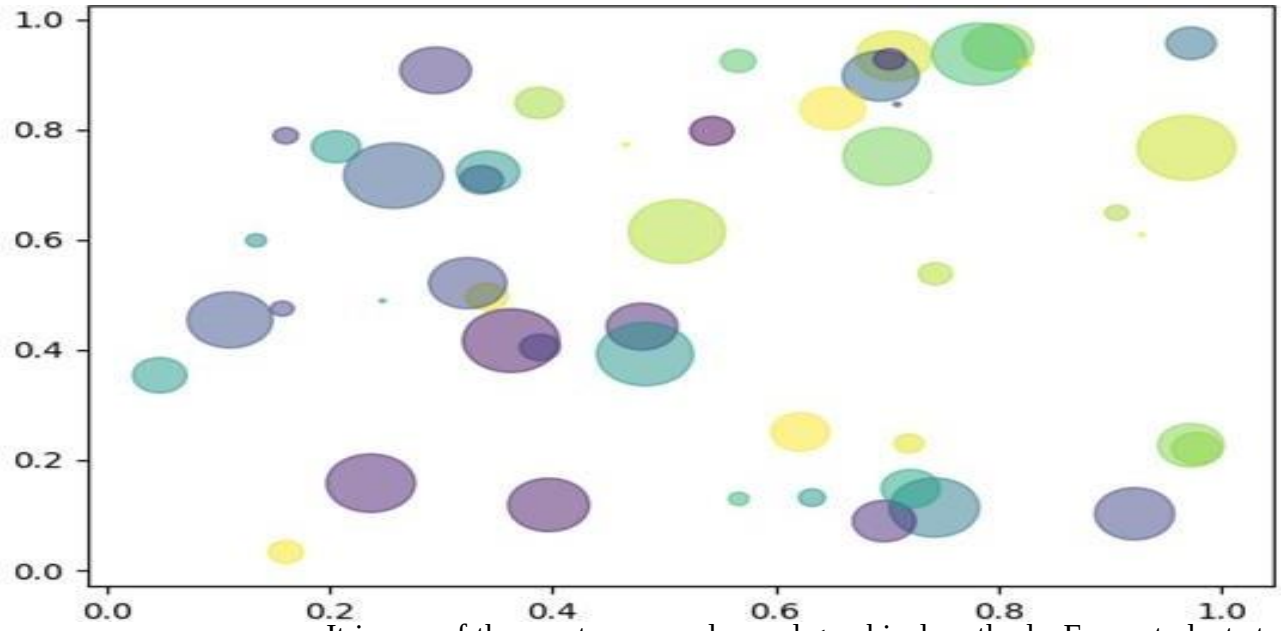
In Python, Matplotlib package helps in visualising the data and making some sense out of it. As we have

already discussed before, with the help of this package, we can plot various kinds of graphs. Let us discuss some of them here:

Scatter Plot

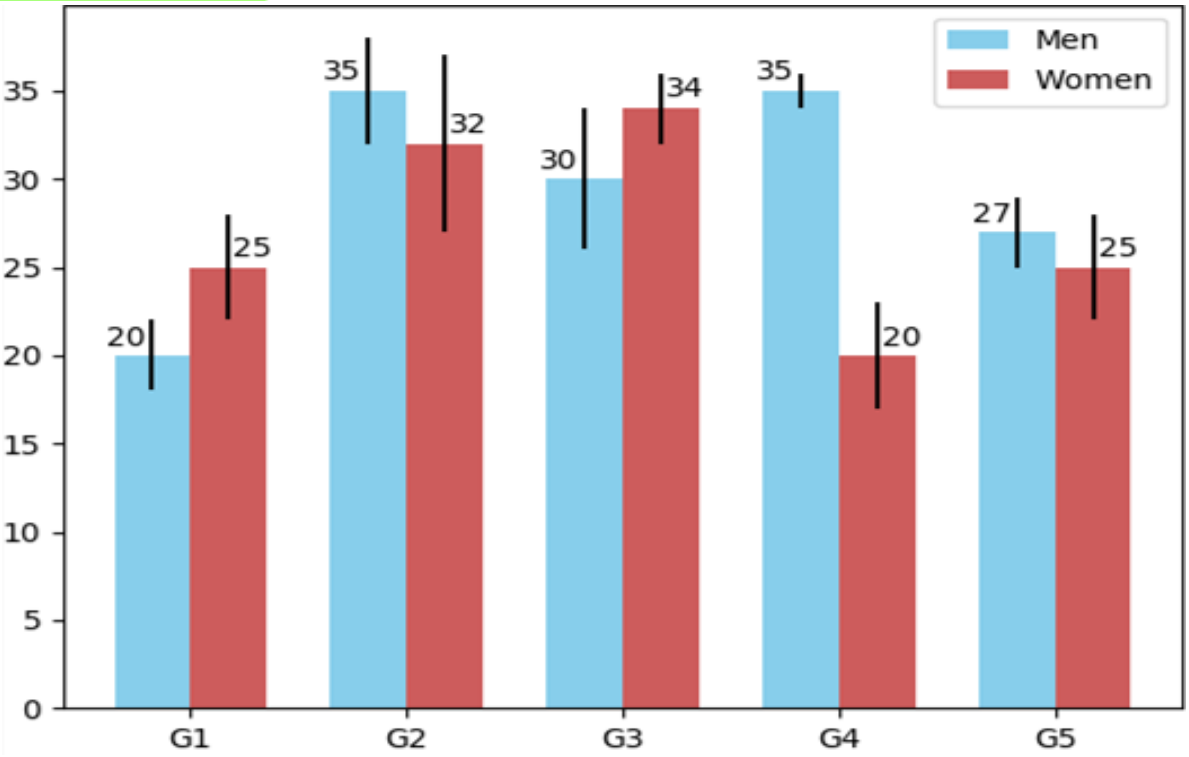
Scatter plots are used to plot discontinuous data; that is, the data which does not have any continuity in flow is termed as discontinuous. There exist gaps in data which introduce discontinuity. A 2D scatter plot can display information maximum upto 4 parameters.

In this scatter plot, 2 axes (X and Y) are two different parameters. The colour of circles and the size both represent 2 different parameters. Thus, just through one coordinate on the graph, one can visualise 4 different parameters all at once.



Bar Chart

It is one of the most commonly used graphical methods. From students to scientists, everyone uses bar charts in some way or the other. It is a very easy to draw yet informative graphical representation. Various versions of bar chart exist like single bar chart, double bar chart, etc.

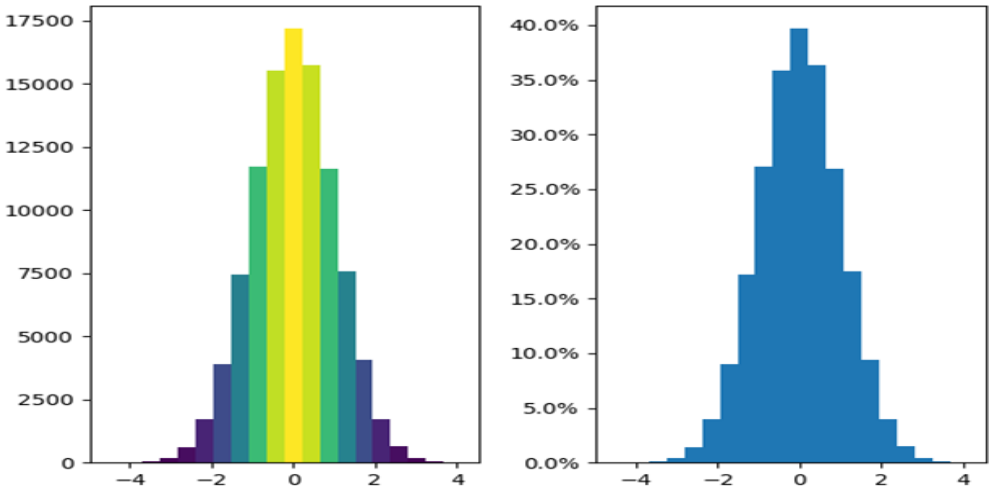


This is an example of a double bar chart. The 2 axes depict two different parameters while bars of different colours work with different entities (in this case it is women and men). Bar chart also workson discontinuous data and is made at uniform intervals.

Histogram

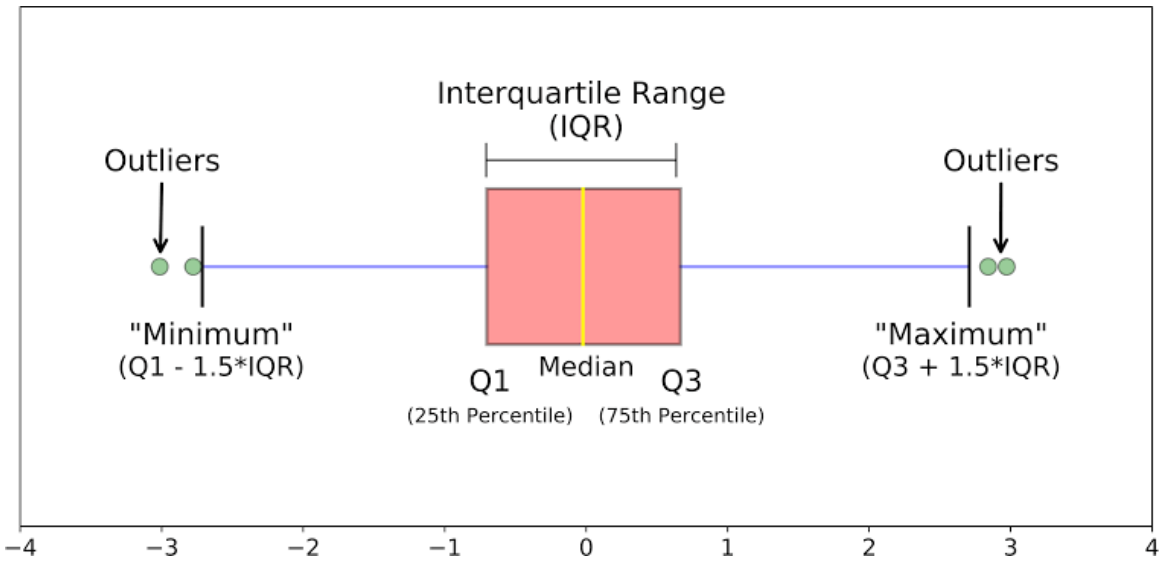
Histograms are the accurate representation of a continuous data. When it comes to plotting the variation in just one entity of a period of time, histograms come into the picture. It represents the frequencyof the variable at different points of time with the help of the bins.

In the given example, the histogram is showing the variation in frequency of the entity plotted with the help of XY plane. Here, at the left, the frequency of the element has been plotted and it is a frequency map for the same. The colours show the transition from low to high and vice versa. Whereason the right, a continuous dataset has been plotted which might not be talking about the frequency of occurrence of the element.



Box Plots

When the data is split according to its percentile throughout the range, box plots come in haman. Box plots also known as box and whiskers plot conveniently display the distribution of data throughoutthe range with the help of 4 quartiles.



Here as we can see, the plot contains a box and two lines at its left and right are termed as whiskers. The plot has 5 different parts to it:

Quartile 1: From 0 percentile to 25th percentile – Here data lying between 0 and 25th percentile is plotted. Now, if the data is close to each other, lets say 0 to 25th percentile data has been covered in just 20-30 marks range, then the whisker would be smaller as the range is smaller. But if the range is large that is 0-30 marks range, then the whisker would also get elongated as the range is longer.

Quartile 2: From 25th Percentile to 50th percentile – 50th percentile is termed as the mean of the whole distribution and since the data falling in the range of 25th percentile to 75th percentile has minimum deviation from the mean, it is plotted inside the box.

Quartile 3: From 50th percentile to 75th percentile – This range is again plotted in the box as its deviation from the mean is less. Quartile 2 & 3 (from 25th percentile to 75th percentile) together constitute the Inter Quartile Range (IQR). Also, depending upon the range of distribution, just like whiskers, the length of box also varies if the data is less spread or more.

Quartile 4: From 75th percentile to 100th percentile – It is the whiskers plot for top 25 percentile data.

Outliers: The advantage of box plots is that they clearly show the outliers in a data distribution. Points which do not lie in the range are plotted outside the graph as dots or circles and are termed as outliers as they do not belong to the range of data. Since being out of range is not an error, that is why they are still plotted on the graph for visualisation.

Let us now move ahead and experience data visualisation using Jupyter notebook. Matplotlib library will help us in plotting all sorts of graphs while Numpy and Pandas will help us in analysing the data.

Data Sciences: Classification Model

In this section, we would be looking at one of the classification models used in Data Sciences. But before we look into the technicalities of the code, let us play a game.

Personality Prediction

Step 1: Here is a map. Take a good look at it. In this map you can see the arrows determine a quality. The qualities mentioned are:

1. Positive X-axis – People focussed: You focus more on people and try to deliver the best experience to them.
2. Negative X-axis – Task focussed: You focus more on the task which is to be accomplished and try to do your best to achieve that.
3. Positive Y-axis – Passive: You focus more on listening to people and understanding everything that they say without interruption.
4. Negative Y-axis – Active: You actively participate in the discussions and make sure that you make your point in-front of the crowd.

Think for a minute and understand which of these qualities you have in you. Now, take a chit and write your name on it. Place this chit at a point in this map which best describes you. It can be placed anywhere on the graph. Be honest about yourself and put it on the graph.

Step 2: Now that you have all put up your chits on the graph, it's time to take a quick quiz. Go to this link and finish the quiz on it individually.

On this link, you will find a personality prediction quiz. Take this quiz individually and try to answer all the questions honestly. Do not take anyone's help in it and do not discuss about it with anyone. Once the quiz is finished, remember the animal which has been predicted for you. Write it somewhere and do not show it to anyone. Keep it as your little secret.

Once everyone has gone through the quiz, go back to the board remove your chit, and draw the symbol which corresponds to your animal in place of your chit. Here are the symbols:

Lion	Otter	Golden Retriever	Beaver
			

Place these symbols at the locations where you had put up your names. Ask 4 students not to do so and tell them to keep their animals a secret. Let their name chits be on the graph so that we can predict their animals with the help of this map.

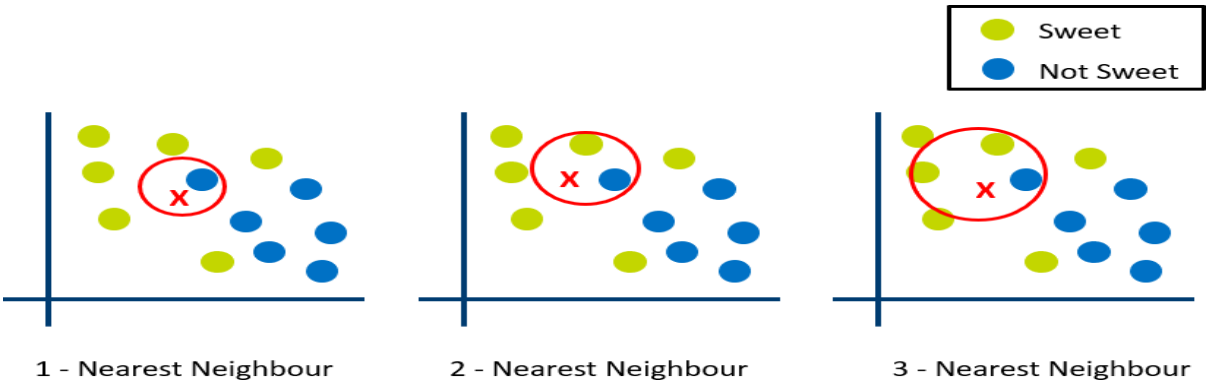
Now, we will try to use the nearest neighbour algorithm here and try to predict what can be the possible animal(s) for these 4 unknowns. Now look that these 4 chits one by one. Which animal is occurring the most in their vicinity? Do you think that if the m lion symbol is occurring the most near their chit, then there is a good probability that their animal would also be a lion? Now let us try to guess the animal for all 4 of them according to their nearest neighbours respectively. After guessing the animals, ask these 4 students if the guess is right or not.

K-Nearest Neighbour: Explained

The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other as the saying goes "*Birds of a feather flock together".* Some features of KNN are:

- The KNN prediction model relies on the surrounding points or neighbours to determine its class or group
- Utilises the properties of the majority of the nearest points to decide how to classify unknown points
- Based on the concept that similar data points should be close to each other

The personality prediction activity was a brief introduction to KNN. As you recall, in that activity, we tried to predict the animal for 4 students according to the animals which were the nearest to their points. This is how in a lay-man's language KNN works. Here, K is a variable which tells us about the number of neighbours which are taken into account during prediction. It can be any integer value starting from 1



Here, X is the value which is to be predicted. The green dots depict sweet values and the blue ones denote not sweet. Let us try it out by ourselves first. Look at the map closely and decide whether X should be sweet or not sweet? Now, let us look at each graph one by one:

1

Here, we can see that K is taken as 1 which means that we are taking only 1 nearest neighbour into consideration. The nearest value to X is a blue one hence 1-nearest neighbour algorithm predicts that the fruit is not sweet.

2

In the 2nd graph, the value of K is 2. Taking 2 nearest nodes to X into consideration, we see that one is sweet while the other one is not sweet. This makes it difficult for the machine to make any predictions based on the nearest neighbour and hence the machine is not able to give any prediction.

3

In the 3rd graph, the value of K becomes 3. Here, 3 nearest nodes to X are chosen out of which 2 are green and 1 is blue. On the basis of this, the model is able to predict that the fruit is sweet.

On the basis of this example, let us understand KNN better:

KNN tries to predict an unknown value on the basis of the known values. The model simply calculates the distance between all the known points with the unknown point (by distance we mean to say the different between two values) and takes up K number of points whose distance is minimum. And according to it, the predictions are made.

Let us understand the significance of the number of neighbours:

1. As we decrease the value of K to 1, our predictions become less stable. Just think for a minute, imagine K=1 and we have X surrounded by several greens and one blue, but the blue is the single nearest neighbour. Reasonably, we would think X is most likely green, but because K=1, KNN incorrectly predicts that it is blue.
2. Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point). Eventually, we begin to witness an increasing number of errors. It is at this point we know we have pushed the value of K too far.
3. In cases where we are taking a majority vote (e.g. picking the mode in a classification problem) among labels, we usually make K an odd number to have a tie breaker.

END

QUESTIONS AND ANSWER:

MCQ:

1. Data science is the process of diverse set of data through ?

A. Organizing data
B. Processing data
C. Analysing data
D. All of the above

Answer : D

Explanation: Data science is the field which includes organizing data, processing data and analysing data to extract valuable information from data for business decision-making, strategic planning, etc. So, All of the above is correct.

2. Point out the correct statement.

A. Raw data is original source of data
B. Preprocessed data is original source of data
C. Raw data is the data obtained after processing steps
D. None of the above

Answer: A

Explanation: Raw data is original source of data is the correct answer. So, option A is correct.

3. How do we perform Bayesian classification when some features are missing?

A. We integrate the posteriors probabilities over the missing features
B. We ignore the missing features
C. We assuming the missing values as the mean of all values
D. Drop the features completely

Answer: A

Explanation: When some features are missing, while performing Bayesian classification we don't use general methods of handling missing values but we integrate the posteriors probabilities over the missing features for better predictions. So, option A is correct.

4. The modern conception of data science as an independent discipline is sometimes attributed to?

A. John McCarthy
B. Arthur Samuel
C. William S.

D. Dennis Ritchie
Answer: C
Explanation: William S. developed data science.

5. _____ graph displays information as a series of data points connected by straight line segments.

A. Bar
B. Scatter
C. Histogram
D. Line
Answer: D
Explanation: A line graph displays information as a series of data points connected by straight line segments.

6. Data fishing is sometimes referred to as

A. Data bagging
B. Data dredging
C. Data merging
D. None of the mentioned
Answer: B
Explanation: Data fishing is sometimes referred to as Data dredging so option B is correct.

7. Which is one of the significant data science skills?

A. Statistics
B. Data Visualization
C. Machine Learning
D. All of the above

8. A method used to make vector of repeated values?

A. read()
B. data()
C. rep()
D. view()
Answer: B
Explanation: data() method used to make vector of repeated values.

9. Which of the following step is performed by the data scientist after acquiring the data?

A. Data Replication
B. Data Integration
C. Data Cleansing
D. All of the Mentioned
Answer : C
Explanation: Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a database, table, or record set and it is generally performed by data scientist after acquiring the data.

10. Which of the following is good way of performing experiments in data science?

A. Measure variability
B. Generalize to the problem
C. Have Replication
D. All of the above
Answer: D
Explanation: Measure variability, generalize to the problem, have Replication all of the mentioned is good way of performing experiments in data science.

SHORT AND LONG QUESTION AND ANSWER:

11. **All of us use Smartphone's. When we install a new app, it asks us for several permissions to access our phone's data in different ways. Why do apps collect such data?**

ANS: 1. To provide customized notifications and recommendations.
2. To improve the efficiency and accuracy of the app.

12. Sirisha and Divisha want to make a model which will organize the unlabeled input data into groups based on features. Which learning model should they use and why?

ANS: Clustering model/Unsupervised learning is used to organize the unlabeled input data into groups based on features. Clustering is an unsupervised learning algorithm which can cluster unknown data according to the patterns or trends identified out of it. The patterns observed might be the ones which are known to the developer or it might even come up with some unique patterns out of it.

13**. Ajay wants to access data from various sources. Suggest him any two points that he needs to keep in mind while accessing data from any data source.**

ANS: While accessing data from any of the data sources, following points should be kept in mind:

1. Data which is available for public usage only should be taken up.
2. Personal datasets should only be used with the consent of the owner.
3. One should never breach someone's privacy to collect data.
4. Data should only be taken from reliable sources as the data collected from random sources can be wrong or unusable.
5. Reliable sources of data ensure the authenticity of data which helps in the proper training of the AI model.
6. Data should be relevant to the problem

13**. Draw the confusion matrix for the following data the number of true positive = 100 the number of true negative 47, the number of false positive = 62, the number of false negative = 290**

14. How can data science help the government in fraud detection?

ANS: Data Scientists use methods and technologies that are combined in fraud analytics to assist spot possibly fraudulent transactions. The ability to process vast amounts of data at once is the main advantage of employing Data Science and Data Analytics for fraud detection by Data Analysts.

In US government, every year there is tax evasion so mow the government uses fraud-detection protocols in the digital age. The government has improved efficiency by making multidimensional taxpayer profiles from social media data, mixed metadata, emailing analysis, electronic payment patterns and more. Based on these profiles, the government forecasts individual tax returns. It became easy to catch those individuals who tried to evade the tax. Indian Government will soon the digital KYC for the verification of each Indian citizen for checking whether they are paying taxes as per their income or are conducting any fraud using black money.

15. What is Data science? Give an example of it.

Data sciences is a domain of AI related to data systems and processes, in which the system collects numerous data, maintains data sets and derives meaning/sense out of them. The information extracted through data science can be used to make a decision about it.

OR

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

OR

Data Sciences, it is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyses actual phenomena with data.

For example: a company that has petabytes of user data may use data science to develop effective ways to store, manage, and analyze the data.

16. Where do we collect data from?

Data can be collected from various sources like –

- Surveys
- Sensors
- Observations
- Web scrapping (Internet)
- Interviews
- Documents and records.
- Oral histories

17. Why do we need to collect data?

Data to a machine is similar to food for human being to function. The world of Artificial Intelligence revolves around Data. Every company whether small or big is collecting data from as many sources as possible. Data is called the New Gold today. It is through data collection that a business or management has the quality information they need to make informed decisions from further analysis, study, and research. Data collection allows themto stay on top of trends, provide answers to problems, and analyze new insights to great effect.

18. What is data mining? Explain with example.

Data mining is the process of analyzing large data sets and extracting the useful information from it. Data mining is used by companies to turn raw data into useful information. It is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information

OR

Data mining is an automatic or semi-automatic technical process that analyses large amounts of scattered information to make sense of it and turn it into knowledge. It looks for anomalies, patterns or correlations among millions of records to predict results, as indicated by the SAS institute, a world leader in business analytics.

Example:

Price Comparison websites- They collect data about a product from different sites and thenanalyze trends out of it and show up the most appropriate results.

Data mining is also known as Knowledge Discovery in Data (KDD)

To be moved to chapter no. 3

19. What do you understand by Data Privacy?

The world of Artificial Intelligence revolves around Data. Proper and ethical handling of own data or user data is called data privacy. It is all about the rights of individuals with respect to their personal information.

Data privacy or information privacy is a branch of data security concerned with the proper handling of data – consent, notice, and regulatory obligations. More specifically, practical data privacy concerns often revolve around: Whether or how data is shared with third parties

20. Is data which is collected by various applications ethical in nature? Justify your

Yes, most of the times, the data collected by various applications is ethical in nature as the users agree to it by clicking on allow when the application asks for various permissions. They ask for our data for various facilities like - to show us personalized recommendations and advertisements and to make their app more accurate and efficient