**07/07/2023** PYTHON -

1. What is the primary distinction between a python list and tuple?

**Answer -> Python list is Mutable and Tuple is Immutable.**

1. How can you add a new-value pair to an existing dictionary?

**Answer -> update () operator is used.**

**Eg:- dict.update({key:value})**

1. What is a regular expression in python, and what module is used for working with regular expression?

**Answer -> A regular expression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern. The Python module re is used to work with Regular Expressions. (import re)**

1. Write a function a reverse a string without using built-in reverse functions.

**Answer -> def str\_reverse(my\_string):**

**str=""**

**for i in my\_string:**

**str = i+str**

**print("Reversed string: ",str)**

**str\_reverse("My World")**

**Output -> dlroW yM**

5) Write a python function to check if a string is a pangram (contains all the letters of the alphabet).

**Answer ->**

**import string #importing string library**

**inputString = input("Enter the string : ") #Taking input from user**

**allAlphabets = 'abcdefghijklmnopqrstuvwxyz' #declaring variable and initilize with the all alphabets**

**flag = 0 #flag value 0**

**#iteration for all the characters in the allAlphabets variable**

**for char in allAlphabets:**

**#checking, Is iterated character is in the string{ In Lowercase }**

**if char not in inputString.lower():**

**flag = 1 #if yes, Flag value updated**

**if flag == 1: #checking flag value and printing the result**

**print("This is not a pangram string");**

**else:**

**print("It is a pangram string")**

**Enter the string : Pack my box with five dozen liquor jugs**

**It is a pangram string**

SQL -

1) Are multiple foreign keys possible for a table, Explain with an example?

**Answer -> A table may have multiple foreign keys, and each foreign key can have a different parent table.**

### Eg:- Customer table

* **ID---Name---Balance---*Account\_Name*---*Account\_Type* (Account\_Name,Account\_Type = Foreign Keys)**

**Account Category table (Parent Table)**

* **Account\_Type----Balance (Account\_Type = Primary Key)**

**Customer Detail table (Parent Table)**

* **Account\_Name---First\_Name----Last\_Name---Address (Account\_Name = Primary Key)**

2) What does SQL's TRUNCATE statement accomplish?

**Answer -> Truncate is used to delete all the rows from the table & free the space containing the table.**

3) What is the difference between a database and table in SQL?

**Answer -> A table is an object inside a database whereas, a database is a collection of several components like tables, indexes, stored procedures and so on.**

4) What is the purpose of the ENUM data type in SQL?

**Answer -> In SQL, the ENUM data type is a string object that allows us to limit the entries in a particular column by permitting only the String values specified for that column during the table creation.**

5) Explain the concept of index fragmentation and its impact on database performance.

**Answer -> The index fragmentation is the index performance value in percentage, which can be fetched by SQL Server. According to the index performance value, users can take the indexes in maintenance by revising the fragmentation percentage with the help of Rebuild or Reorganize operation.**

EDA -

1) What is outlier and how it be performed during EDA?

**Answer -> An outlier is something separate or different from the crowd. Outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data.**

**During EDA, Outliers can be detected using the boxplots.**

**Handling Outliers -> Different Methods**

* 1. **3R Technique (Rectify, Retain, Remove)**
  2. **Masking – False Negative**
  3. **Swamping – False Positive**
  4. **Winsorization – from feature\_engine.outliers import Winsorizer**

**Winsorizer(capping\_method=’iqr’, tail=’both’,fold=3,variables=’cols’)**

* 1. **Trimming –**

**# Detection of Outliers**

**IQR = df['col'].quantile(0.75) - df['col'].quantile(0.25) # IQR - Inter quartile range IQR = Q3-Q1**

**lower\_limit = df['col'].quantile(0.25) - (IQR \* 1.5) # Q1 - 1.5 \* IQR**

**upper\_limit = df['col'].quantile(0.75) + (IQR \* 1.5) # Q3 + 1.5 \* IQR**

**outliers\_df = np.where(df['col'] > upper\_limit,True,np.where(df['col'] < lower\_limit,True,False))**

**# if value is greater than upper limit consider it as outliers and if the value is less than lower limit consider it as outliers**

**df\_trimmed = df.loc[~(outliers\_df),] # ~ means not - it shows all false value (not outliers)**

**df.shape, df\_trimmed.shape**

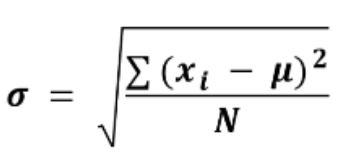
2) What is the relationship between Mean, Median and Mode for positive and negative skewed?

**Answer-> In case of a positively skewed frequency distribution, the mean is always greater than median and the median is always greater than the mode. Mean > Median > Mode.**

**In case of a negatively skewed frequency distribution, the mean is always lesser than median and the median is always lesser than the mode. Mean < Median < Mode.**

**3) What is the formula for Standard Deviation?**

**Answer -> Standard Deviation formula for Population Parameters is**

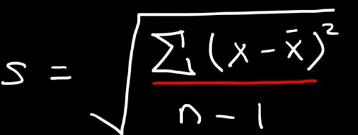
****

**xi** = every point in the dataset (observation or member of the population).

**μ**= population mean

**N** = the number of values in the population

**Standard Deviation formula for Sample Statistics is**



**xi** = every point in the dataset (observation or member of the population).

**x̄** = sample mean

**n-1** = the number of values in the sample (n) minus 1.

4) How do you interpret skewness and kurtosis in EDA?

**Answer -> Skewness :-**

**If Skewness = 0, then the data is normally distributed. If skewness is negative, then it is interpreted as left skewed and if it is positive, it is interpreted as right skewed.**

**Kurtosis:-**

**If Kurtosis = 3, then data is normally distributed or also called as Mesokurtic Distribution. If Kurtosis is negative, then it is interpreted as Platykurtic Distribution which has wide peak and thin tails and if it is positive, it is interpreted as Leptokurtic Distribution which has sharp peak and thick tails.**

5) What is the kurtosis value for standard bell curve?

**Answer -> Kurtosis value for a standard bell curve is 3 & it is called as Mesokurtic Distribution.**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**08/07/2023** PYTHON -

1) What are python keywords?

**Answer -> Python keywords are special reserved words that have specific meanings and purposes and can't be used for anything but those specific purposes.**

**Ex:- import keyword**

**keyword.kwlist**

**Few keywords are – True,False,print,class,def etc.**

2) How to handle exception in python?

**Answer->**

**When an error occurs, or exception as we call it, Python will normally stop and generate an error message. These exceptions can be handled using the try statement**

* 1. **Try block**

**The try block lets you test a block of code for errors, ex:-**

**try:  
  print(df)  
except:  
  print("An exception occurred")**

* 1. **Except block**

**The except block will handle the error**

* 1. **Else block**

**The else block execute code when there is no error**

**try:**

**print("Hello")**

**except:**

**print("Something went wrong")**

**else:**

**print("Nothing went wrong")**

* 1. **Finally block**

**The finally block lets you execute code, regardless of the result of the try- and except blocks. Ex:-**

**try:**

**print(x)**

**except:**

**print("Something went wrong")**

**finally:**

**print("The 'try except' is finished")**

3) What is a lambda function in python and where is it useful?

**Answer -> Lambda function in python is called as nameless or anonymous function.**

**Lambda functions are efficient whenever you want to create a function that will only contain simple expressions – that is, expressions that are usually a single line of a statement.**

**str1 = 'DigitMG'**

**upper = lambda string: string.upper()**

**print(upper(str1))**

**Output : DIGITMG**

4) How do you handle exceptions in python and what is the reason for using the exceptions?

**Answer->**

**When an error occurs, or exception as we call it, Python will normally stop and generate an error message. These exceptions can be handled using the try statement**

* 1. **Try block**

**The try block lets you test a block of code for errors, ex:-**

**try:  
  print(df)  
except:  
  print("An exception occurred")**

* 1. **Except block**

**The except block will handle the error**

* 1. **Else block**

**The else block execute code when there is no error**

**try:**

**print("Hello")**

**except:**

**print("Something went wrong")**

**else:**

**print("Nothing went wrong")**

* 1. **Finally block**

**The finally block lets you execute code, regardless of the result of the try- and except blocks. Ex:-**

**try:**

**print(x)**

**except:**

**print("Something went wrong")**

**finally:**

**print("The 'try except' is finished")**

5) What distinguish the python '==' and 'is' operators?

**Answer - > The ‘== ‘operator compares the value or equality of two objects, whereas the Python ‘is‘ operator checks whether two variables point to the same object in memory.**

--------------------------------------------------------------------------------------------------------------------------------------

SQL -

1) What is schema in SQL?

**Answer -> In a SQL database, a schema is a list of logical structures of data. A database user owns the schema, which has the same name as the database manager. As of SQL Server 2005, a  schema is an individual entity (container of objects) distinct from the user who constructs the object. In other words, schemas are similar to separate namespaces or containers used to handle database files. Schemas may be assigned security permissions, making them an effective method for distinguishing and defending database objects based on user access privileges. It increases the database's stability for security-related management.**

2) How many types of Normal Forms do we have in SQL? and what are they?

**Answer -> Database normalization is the process of restructuring a relational database in accordance with a series of so-called normal forms in order to reduce the data redundancy and improve data integrity.**

**There are different normal forms as in first normal form (or 1NF), second normal form (or 2NF), and third normal form (or 3NF), BCNF (Boyce-Codd Normal Form), 4NF,5NF.**

1. **1NF – The first normal form requires that a table satisfies the conditions as in rows & columns are not ordered, duplicates available, row-column intersections always have a unique value.**
2. **2NF – An entity is in a 2NF if all of its attributes depend on the whole primary key, values of different columns have dependency on other columns, table must be in 1NF & all non-key columns of the tables must depend on primary key, partial dependencies are removed and placed in separate table.**
3. **3NF – We should eliminate the fields in a table that don’t depend on the key, non-Primary key columns shouldn’t depend on the other Non-Primary key columns, there is no transitive functional dependency.**
4. **BCNF (Boyce-Codd Normal Form) – Even when a DB is in 3NF, still there would be anomalies resulted if it has more than 1 candidate key, it is also referred to as 3.5NF.**
5. **4NF – If no DB table instance contains 2 or more independent and multivalued data describing relevant entity, then it is in 4NF.**
6. **5NF – A table is in the 5NF only if it is 4NF & it can’t be decomposed into any number of smaller tables without loss of data.**

3) How can you identify missing value patterns in a SQL dataset?

**Answer -> A good way to identify missing value patterns is by replacing each column with an indicator column that is equal to 1 if the value is missing and 0 if the value is not missing.**

4) What distinguish DELETE and TRUNCATE statements from one another?

**Answer -> The TRUNCATE statement in SQL removes all data from the table and free the table's space. SQL's DELETE statement removes all data from the table but does not free the table's space.**

5) What distinguish an OUTER JOIN from an INNER JOIN?

**Answer -> The biggest difference between an INNER JOIN and an OUTER JOIN is that the inner join will keep only the information from both tables that's related to each other (in the resulting table). An Outer Join, on the other hand, will also keep information that is not related to the other table in the resulting table.**

--------------------------------------------------------------------------------------------------------------------------------------

EDA -

1) Explain the concept of correlation and which function is used to check the correlation between features?

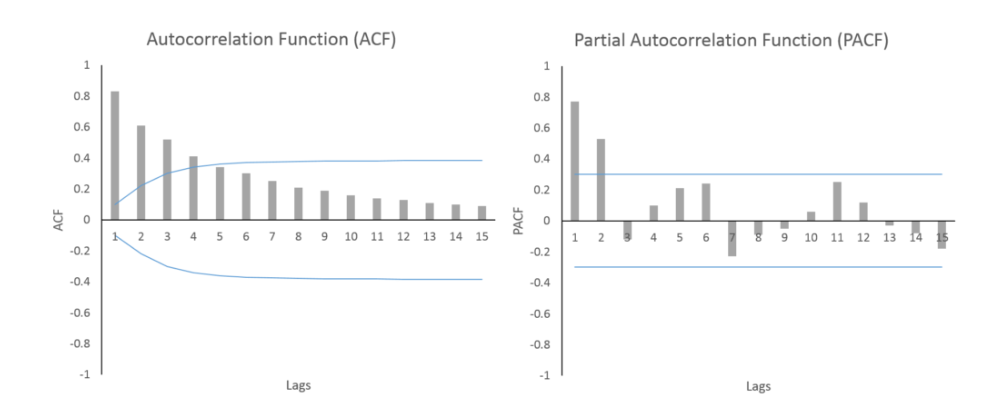
**Answer -> Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of the relationship, the correlation coefficient's value varies between +1 and -1.**

**Correlation function -> df[col1]. corr(df[col2])**

2) What is the difference between ACF and PACF plot?

**Answer -> Autocorrelation function (ACF). At lag k, this is the correlation between series values that are k intervals apart. Partial autocorrelation function (PACF). At lag k, this is the correlation between series values that are k intervals apart, accounting for the values of the intervals between.**

**The difference between ACF and PACF is the inclusion or exclusion of indirect correlations in the calculation.**

****

3) What is the formula for calculating skewness and which python function is used to get the skewness value?

**Answer -> Skewness is mostly calculated using the Fisher-Pearson Coefficient of Skewness. However, there are many more ways to calculate it such as Kelly's Measure, Bowley, and Momental. Skewness looks at the measure of skewness as the third standard moment of distribution. Skewness is also the Third Moment Business Decision in EDA.**

**Formula - > Skewness = 3(mean-median)/standard deviation.**

**Python Function -> df.col.skew()**

4) What does X- axis and Y- axis represents in a Histogram?

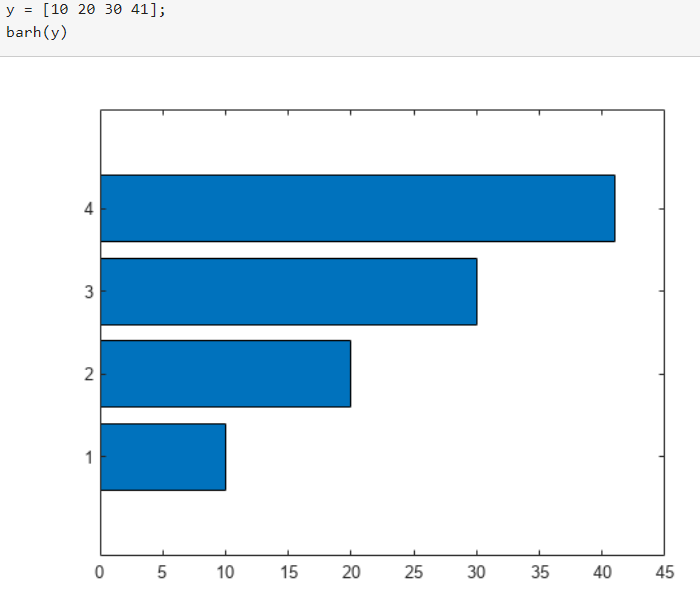
**Answer -> X-axis: The X-axis are intervals that show the scale of values which the measurements fall under.**

**Y-axis: The Y-axis shows the number of times that the values occurred within the intervals set by the X-axis.**

5) Which function is used to get horizontal bar plot?

**Answer -> barh( y ) creates a horizontal bar graph with one bar for each element in y . If y is an m-by-n matrix, then barh creates m groups of n bars. barh( x , y ) draws the bars along the vertical axis at the locations specified by x .**

**Ex: -**

****

**10/07/2023** PYTHON -

1) How can you match a string that starts with an uppercase letter followed by any number of lowercase letters using regular expressions in Python?

**Answer - Using re.search() To check if the sequence of one upper case letter followed by lower case letters we use regular expression '[A-Z]+[a-z]+$'.**

**pattern = '[A-Z]+[a-z]+$'**

**re.search(pattern,text)**

2) What is the difference between the list method append() and extend()?

**Answer - append() adds a single element to the end of the list while . extend() can add multiple individual elements to the end of the list.**

**Ex:- Append**

**# Define the list**

**>>> nums = [1, 2, 3, 4]**

**# Add the integer 5 to the end of the existing list**

**>>> nums.append(5)**

**# See the updated value of the list**

**>>> nums**

**[1, 2, 3, 4, 5]**

**# Extend**

**>>> nums = [5.6, 7.44, 6.75, 4.56, 2.3]**

**>>> new\_values = [2.3, 9.6, 4.564, 7.56]**

**# This is where the magic occurs! No more for loops**

**>>> nums.extend(new\_values)**

**# The list was updated with individual values**

**>>> nums**

**[5.6, 7.44, 6.75, 4.56, 2.3, 2.3, 9.6, 4.564, 7.56]**

3) Explain how python memory management works?

**Answer - The Python memory is primarily managed by Python private heap space. All Python objects and data structures are located in a private heap. The programmer does not have access to this private heap and interpreter takes care of this Python private heap.**

**Also, In Python the code directly gets executed without getting converted into machine language. Hence, it is slower and takes single instruction as input & the memory requirement is less.**

4) How does the string get converted to a number?

**Answer - Strings can be converted to numbers by using the int() and float() methods. If your string does not have decimal places, you'll most likely want to convert it to an integer by using the int() method.**

**Ex: - var1 = ‘abc’ # is a string**

**var1.astype(int) # code to convert the string to integer**

5) Is there a tool to help find bugs or perform static analysis?

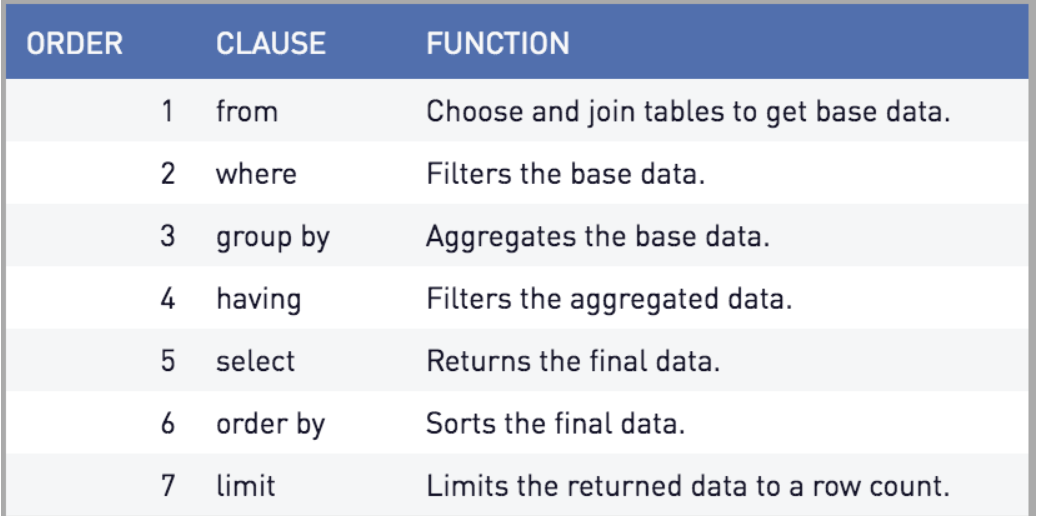
**Answer - Pychecker and Pylint are the static analysis tools that help to find bugs in python**

--------------------------------------------------------------------------------------------------------------------------------------

SQL -

1) What is the SQL server query execution sequence?

**Answer - The SQL order of execution defines the order in which the clauses of a query are evaluated.  The order of execution is as follows.**

****

2) What are the three degrees of normalization and how is normalization done in each degree?

**Answer - Normalization is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. The purpose of Normalisation in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.**

**Here is a list of Normal Forms in SQL:**

* **1NF (First Normal Form)**
* **2NF (Second Normal Form)**
* **3NF (Third Normal Form)**
* **BCNF (Boyce-Codd Normal Form)**
* **4NF (Fourth Normal Form)**
* **5NF (Fifth Normal Form)**
* **6NF (Sixth Normal Form)**

## **1NF (First Normal Form) Rules**

* **Each table cell should contain a single value.**
* **Each record needs to be unique.**

## **2NF (Second Normal Form) Rules**

* **Rule 1- Be in 1NF**
* **Rule 2- Single Column Primary Key that is not functionally dependent on any subset of candidate key relation**

## **3NF (Third Normal Form) Rules**

* **Rule 1- Be in 2NF**
* **Rule 2- Has no transitive functional dependencies**

## **BCNF (Boyce-Codd Normal Form)**

**Even when a database is in 3rd Normal Form, still there would be anomalies resulted if it has more than one Candidate Key.**

**Sometimes the BCNF is also referred as 3.5 Normal Form.**

## **4NF (Fourth Normal Form) Rules**

**If no database table instance contains two or more, independent and multivalued data describing the relevant entity, then it is in 4th Normal Form.**

## **5NF (Fifth Normal Form) Rules**

**A table is in 5th Normal Form only if it is in 4NF and it cannot be decomposed into any number of smaller tables without loss of data.**

## **6NF (Sixth Normal Form) Proposed**

**6th Normal Form is not standardized, yet however, it is being discussed by database experts for some time. Hopefully, we would have a clear & standardized definition for 6th Normal Form in the near future.**

3) What is derived column, how does it work, how it affects the performance of a database and how can it be improved?

**Answer - Derived columns let you move the processing of an expression from the target instance to the source instance. For example, you may have already defined an expression that concatenates the values of two source columns, FIRSTNAME and LASTNAME, and mapped this expression to a target column named called FULLNAME. Hence advantage is that the performance is faster.**

4) What is the difference between OLTP and OLAP?

**Answer - Online analytical processing (OLAP) and online transaction processing (OLTP) are two different data processing systems designed for different purposes. OLAP is optimized for complex data analysis and reporting, while OLTP is optimized for transactional processing and real-time updates.**

5) What are the different types of Restricted Joins?

**Answer –**

**The JOIN operations are:**

* [**INNER JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj35034.html#rrefsqlj35034)

**Specifies a join between two tables with an explicit join clause.**

* [**LEFT OUTER JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj18922.html#rrefsqlj18922)

**Specifies a join between two tables with an explicit join clause, preserving unmatched rows from the first table.**

* [**RIGHT OUTER JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj57522.html#rrefsqlj57522)

**Specifies a join between two tables with an explicit join clause, preserving unmatched rows from the second table.**

* [**CROSS JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqljcrossjoin.html#rrefsqljcrossjoin)

**Specifies a join that produces the Cartesian product of two tables. It has no explicit join clause.**

* [**NATURAL JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqljnaturaljoin.html#rrefsqljnaturaljoin)

**Specifies an inner or outer join between two tables. It has no explicit join clause. Instead, one is created implicitly using the common columns from the two tables.**

**In all cases, you can specify additional restrictions on one or both of the tables being joined in outer join clauses or in the**[**WHERE clause**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj33602.html#rrefsqlj33602)**.**

EDA -

1) Which statistical measure provides information about the spread or variability of a dataset?

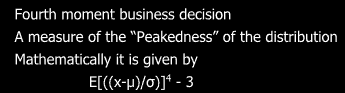
**Answer - The standard deviation is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.**

2) Which data visualisation technique is used to display the relationship between two numerical variables?

**Answer - The scatter plot is the standard way of showing the relationship between two variables. Scatter plots can also be expanded to additional variables by adding color, shape, or size to each point as indicators, as in a bubble chart.**

3) What is the formula for calculating Kurtosis and which python function is used to get the kurtosis value?

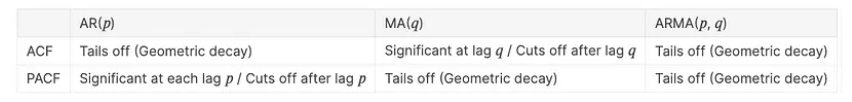
Answer – Formula for Kurtosis is as below.



The function is df.col.kurt()

4) In ACF and PACF plots, how are lower and upper bands calculated?

**Answer –**



5) Why is mean influenced by outliers, but why not median?

**Answer - A higher value or a lower value of outliers can skew the entire mean towards it. The median value represents the 50th percentile rank in an ordered data and hence it is not affected by outliers.**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**11/07/2023** PYTHON -

1) What is the purpose of the init method in python classes?

**Answer - The \_\_init\_\_ method is the Python equivalent of the C++ constructor in an object-oriented approach. The \_\_init\_\_ function is called every time an object is created from a class. The \_\_init\_\_ method lets the class initialize the object's attributes and serves no other purpose. It is only used within classes.**

2) Write a python program to check if a string is a palindrome.

**Answer - def isPalindrome(s):**

**return s == s[::-1]**

**# Driver code**

**s = "malayalam"**

**ans = isPalindrome(s)**

**if ans:**

**print("Yes")**

**else:**

**print("No")**

**Output - Yes**

3) How does python handle exceptions and what are the differences between try/except and finally blocks?

**Answer->**

**When an error occurs, or exception as we call it, Python will normally stop and generate an error message. These exceptions can be handled using the try statement**

* 1. **Try block**

**The try block lets you test a block of code for errors, ex:-**

**try:  
  print(df)  
except:  
  print("An exception occurred")**

* 1. **Except block**

**The except block will handle the error**

* 1. **Else block**

**The else block execute code when there is no error**

**try:**

**print("Hello")**

**except:**

**print("Something went wrong")**

**else:**

**print("Nothing went wrong")**

* 1. **Finally block**

**The finally block lets you execute code, regardless of the result of the try- and except blocks. Ex:-**

**try:**

**print(x)**

**except:**

**print("Something went wrong")**

**finally:**

**print("The 'try except' is finished")**

4) How can you sort a dictionary by its values in python?

**Answer – By using sorted() function.**

**my\_dict = { 'num6': 6, 'num3': 3, 'num2': 2, 'num4': 4, 'num1': 1, 'num5': 5}**

**sortedDict = sorted(my\_dict)**

**print(sortedDict)**

**# ['num1', 'num2', 'num3', 'num4', 'num5', 'num6']**

5) What is the negative indexing in lists?

**Answer - The negative indexing is the act of indexing from the end of the list with indexing starting at -1 i.e. -1 gives the last element of list, -2 gives the second last element of list and so on.**

**Ex: - list1 = [1,2,3,4]**

**print(list1[-1]) , Output = 4**

SQL -

1) What are the different types of Joins?

**Answer –**

**The JOIN operations are:**

* [**INNER JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj35034.html#rrefsqlj35034)

**Specifies a join between two tables with an explicit join clause.**

* [**LEFT OUTER JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj18922.html#rrefsqlj18922)

**Specifies a join between two tables with an explicit join clause, preserving unmatched rows from the first table.**

* [**RIGHT OUTER JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj57522.html#rrefsqlj57522)

**Specifies a join between two tables with an explicit join clause, preserving unmatched rows from the second table.**

* [**CROSS JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqljcrossjoin.html#rrefsqljcrossjoin)

**Specifies a join that produces the Cartesian product of two tables. It has no explicit join clause.**

* [**NATURAL JOIN operation**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqljnaturaljoin.html#rrefsqljnaturaljoin)

**Specifies an inner or outer join between two tables. It has no explicit join clause. Instead, one is created implicitly using the common columns from the two tables.**

**In all cases, you can specify additional restrictions on one or both of the tables being joined in outer join clauses or in the**[**WHERE clause**](https://docs.oracle.com/javadb/10.6.2.1/ref/rrefsqlj33602.html#rrefsqlj33602)**.**

2) What is a sub-query?

**Answer - A subquery is a query that is nested inside a SELECT , INSERT , UPDATE , or DELETE statement, or inside another subquery.**

3) What are the SET Operators?

**Answer - SQL set operation is used to combine two or more select statements.**

**Types of Set Operations-**

**UNION – The SQL union operation is used to combine the result of two or more SQL SELECT queries. In Union operation, the number of datatypes & columns must be the same in both tables on which UNION operation is being applied. The UNION operation eliminates the duplicate rows from its resultset.**

**UNIONALL – Combines the result of two queries and does not remove duplicate rows & does not sort by first column.**

**INTERSECT – It is used to combine two select statements. The Intersect operations returns common rows from both SELECT statements. In Intersect operation, the number of datatypes & columns must be the same. It has no duplicates & it arranges data in ascending order by default.**

**MINUS – It combine the result of two select statements. MINUS operator is used to display the rows which are present in first query but absent in second query.**

4) What are the types of views and give an example for each of them?

**Answer - There are 2 types of Views in SQL: Simple View and Complex View. Simple views can only contain a single base table. Complex views can be constructed on more than one base table.**

5) What is a RANKING function and what are the four RANKING functions?

**Answer - The RANK Function in SQL Server is a kind of Ranking Function. This function will assign the number to each row within the partition of an output. It assigns the rank to each row as one plus the previous row rank.**

**There are 4 ranking functions ROW\_NUMBER(), RANK(), DENSE\_RANK(), and NTILE() are in MS SQL. These are used to perform some ranking operation on result data set.**

EDA -

1) What can you infer from a histogram in EDA?

**Answer - A histogram helps to understand whether the distribution is unimodal, bimodal, or multimodal. Also, it tells about how widely the distribution is spread. It helps to identify outliers and high-leverage points in the dataset.**

**Primary purpose – Understand shape of data distribution**

**Secondary purpose – Identify Outliers**

2) Which data visualization technique is used to display the relationship between a categorical variable and a numerical variable?

**Answer - Bar plot is a simple plot which we can use to plot categorical variable on the x-axis and numerical variable on y-axis and explore the relationship between both variables. The blacktip on top of each bar shows the confidence Interval.**

3) What are the standard names for Positive, Negative, and normal kurtosis curves?

**Answer – Positive – Leptokurtic Distribution, Negative – Platykurtic Distribution, Normal – Mesokurtic Distribution.**

4) What is the formula for expected value?

**Answer - The basic expected value formula is the probability of an event multiplied by the number of times the event happens: (P(x) \* n). The formula changes slightly according to what kinds of events are happening.**

5) In Histogram, what is the default value for number of bins in matplotlib and seaborn?

**Answer – In Histogram, the default value for the number of bins in matplotlib and seaborn is 10.**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**12/07/2023** PYTHON -

1) What is the difference between shallow copy and deep copy in python?

**Answer - A shallow copy constructs a new compound object and then (to the extent possible) inserts references into it to the objects found in the original. A deep copy constructs a new compound object and then, recursively, inserts copies into it of the objects found in the original.**

2) What are metaclasses in python, and how are they used?

**Answer - A metaclass in Python is a class of a class that defines how a class behaves. A class is itself an instance of a metaclass. A class in Python defines how the instance of the class will behave.**

**Ex: -**

**class TestClass():**

**pass**

**my\_test\_class = TestClass()**

**print(my\_test\_class)**

**This code defines a class called TestClass using the class keyword in Python.**

**• The pass keyword is used to indicate that the class has no methods or attributes defined.**

**• Then, an instance of the TestClass class is created and assigned to the variable my\_test\_class using the parentheses () after the class name.**

**• Finally, the print() function is used to output the value of my\_test\_class, which will be a string representation of the object's memory location in memory.**

3) Explain the Global interpreter Lock in python?

**Answer - GIL or Global Interpreter Lock. It is a type of process lock used when working with multiple processes. It gives the control to only one thread. Generally, Python uses a single thread to run a single process. We get the same performance result of the single-threaded and multi-threaded processes using the GIL.**

4) What is generator in python?

**Answer - In Python, a generator is a function that returns an iterator that produces a sequence of values when iterated over. Generators are useful when we want to produce a large sequence of values, but we don't want to store all of them in memory at once.**

**Ex: -**

**def my\_generator(n):**

**# initialize counter**

**value = 0**

**# loop until counter is less than n**

**while value < n:**

**# produce the current value of the counter**

**yield value**

**# increment the counter**

**value += 1**

**# iterate over the generator object produced by my\_generator**

**for value in my\_generator(3):**

**# print each value produced by generator**

**print(value)**

**Output : 0**

**1**

**2**

5) What is the purpose of the ' global keyword ' in python?

**Answer - A global keyword is a keyword that allows a user to modify a variable outside the current scope. It is used to create global variables in Python from a non-global scope, i.e. inside a function.**

**Ex:-**

**# global variable**

**a = 15**

**b = 10**

**# function to perform addition**

**def add():**

**c = a + b**

**print(c)**

**# calling a function**

**add()**

**Output: 25**

**--------------------------------------------------------------------------------------------------------------------------------------**

SQL -

1) What is the purpose of the HAVING clause in SQL queries?

**Answer – Having clause is used to specify a search condition for a group or an aggregate. It is used in a GROUP BY Clause. If we are not using the GROUP BY clause, then we can use HAVING function like a WHERE clause.**

2) How to calculate the average, sum and count of a column in SQL?

**Answer - SQL COUNT(), AVG() and SUM() Functions**

1. **SELECT COUNT(column\_name) FROM table\_name. WHERE condition;**
2. **SELECT AVG(column\_name) FROM table\_name. WHERE condition;**
3. **SELECT SUM(column\_name) FROM table\_name. WHERE condition;**

3) Explain the difference between UNION and UNION ALL in SQL?

**Answer - UNION and UNION ALL in SQL are used to retrieve data from two or more tables. UNION returns distinct records from both the table, while UNION ALL returns all the records from both the tables.**

4) What is the difference between a subquery and join in SQL?

**Answer - Subqueries are useful for filtering rows and returning values as part of the outer query, while joins are useful for combining rows from multiple tables based on a related column. Both can be used to achieve similar goals, but the choice of which one to use depends on the specific requirements of your query.**

5) What is the purpose of the TRIGGER statement in SQL?

**Answer – A triggers is SQL is a procedural code that is automatically executed in response to certain events on a specified table. Purpose of a trigger in SQL is that, it is designed to check or change data based on a data modification or definition statement.**

**--------------------------------------------------------------------------------------------------------------------------------------**

EDA -

1) If there are more than two mode values for a dataset, What does it mean?

**Answer – If there are more than two mode values for a dataset then it is multi modal, which means the dataset has presence of clusters or outliers.**

2) What is the disadvantage of using variance?

**Answer - One drawback to variance, though, is that it gives added weight to outliers. These are the numbers far from the mean. Units get squared. Squaring these numbers can skew the data. Another pitfall of using variance is that it is not easily interpreted.**

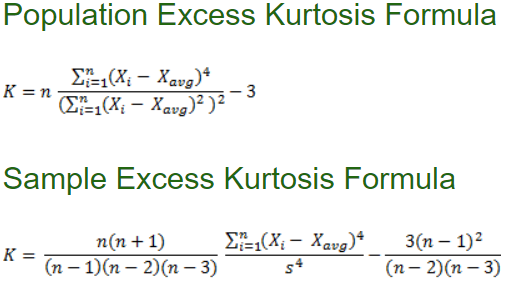
3) What is the value of vert parameter in boxplot function, in order to visualize it horizontally?

**Answer - The vert = 0 attribute creates horizontal box plot.**

4) What is the formula for excess kurtosis?

**Answer – If we subtract 3 from kurtosis, then it is called as Excess Kurtosis.**

**Formula-**

****

5) What is the primary and secondary use of scatter plot?

**Answer – Primary Purpose – Direction, Strength and linearity of variables.**

**Secondary Purpose – Identify Outliers and presence of Clusters.**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**13/07/2023** PYTHON -

1) What is the difference between a module and a package in python?

**Answer - In Python, both modules and packages organize and structure the code but serve different purposes. In simple terms, a module is a single file containing python code, whereas a package is a collection of modules that are organized in a directory hierarchy.**

2) How do you convert a string into a list?

**Answer – Ex: - def Convert(string):**

**li = list(string.split(" "))**

**return li**

**str1 = "Geeks for Geeks"**

**print(Convert(str1))**

3) What is Zip and Unzipping?

**Answer - ZIP is a common file format that's used to compress one or more files together into a single location. This reduces file size and makes it easier to transport or store. A recipient can unzip (or extract) a ZIP file after transport and use the file in the original format.**

4) What is the purpose of using backreferences in regular expressions?

**Answer - In a regular expression pattern, back-references are used to match the same content as a previously matched subexpression. Ex: - <([A-Z][A-Z0-9]\*)\b[^>]\*>.**

--------------------------------------------------------------------------------------------------------------------------------------

SQL -

1) Can you use multiple conditions in a join statement?

**Answer - SQL lets you use AND or OR clauses to join on multiple conditions and/or multiple columns.**

**Syntax : - JOIN <table name> ON <condition 1> [AND / OR] <condition 2> ... [AND / OR] <condition N>**

2) What is the difference between DATE, DATETIME, and TIMESTAMP data types?

**Answer - DATETIME: The DATETIME stores a date and time value in the form YYYY-MM-DD HH:MM:SS . It stores both the date and time. TIMESTAMP: The TIMESTAMP is similar to the DATETIME , but includes a timezone.**

3) What is the difference between INNER JOIN and OUTER JOIN in SQL?

**Answer - INNER JOIN returns the common and the matching records between the tables. OUTER JOIN returns all the records from the database tables.**

--------------------------------------------------------------------------------------------------------------------------------------

EDA -

1) Primary and Secondary purpose of Box plot?

**Answer – Primary Purpose – Identify the Outliers**

**Secondary Purpose – Understand Shape of Data Distribution.**

2) How to interpret positive and negative skewness?

**Answer - If the skewness is right-tailed (meaning the right tail is more pronounced than the left), the skew is positive. In this case, the values are more than zero. If the opposite is true and the tail is more pronounced on the left, then the skew is negative, where the values are less than zero.**

3) What is the difference between covariance and correlation?

Answer - **Both covariance and correlation measure the relationship and the dependency between two variables. Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables.**

**14/07/2023** PYTHON -

1) Write a function that takes a list of numbers as input and returns the largest number in the list?

**Answer –**

**def myMax(list1):**

**# Assume first number in list is largest**

**# initially and assign it to variable "max"**

**max = list1[0]**

**# Now traverse through the list and compare**

**# each number with "max" value. Whichever is**

**# largest assign that value to "max'.**

**for x in list1:**

**if x > max:**

**max = x**

**# after complete traversing the list**

**# return the "max" value**

**return max**

**# Driver code**

**list1 = [10, 20, 4, 45, 99]**

**print("Largest element is:", myMax(list1))**

2) How does the iter() function work in python?

**Answer - Python iter() The python iter function is used to return an iterator for the object. The iter() is used to create an object that will iterate one element at a time. The iter() takes two optional arguments as input.**

**lis1 = [1, 2, 3, 4, 5]**

**# printing type**

**print("The list is of type : " + str(type(lis1)))**

**# converting list using iter()**

**lis1 = iter(lis1)**

**# printing type**

**print("The iterator is of type : " + str(type(lis1)))**

**# using next() to print iterator values**

**print(next(lis1))**

**print(next(lis1))**

**print(next(lis1))**

**print(next(lis1))**

**print(next(lis1))**

**Output**

**The list is of type : <class 'list'>**

**The iterator is of type : <class 'list\_iterator'>**

**1**

**2**

**3**

**4**

**5**

3) What is list comprehension in python?

**Answer - List comprehension in Python is an easy and compact syntax for creating a list from a string or another list. It is a very concise way to create a new list by performing an operation on each item in the existing list. List comprehension is considerably faster than processing a list using the for loop.**

**Ex:- even\_nums = []**

**for x in range(21):**

**if x%2 == 0:**

**even\_nums.append(x)**

**print(even\_nums)**

**Output - [0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20]**

4) How do you remove duplicates from a list in python?

**Answer - l = [1, 2, 4, 2, 1, 4, 5]**

**print("Original List: ", l)**

**res = [\*set(l)]**

**print("List after removing duplicate elements: ", res)**

**Output**

**Original List: [1, 2, 4, 2, 1, 4, 5]**

**List after removing duplicate elements: [1, 2, 4, 5]**

5) Write a python program to find the factorial of a given number.

**Answer –**

**num = 7**

**# To take input from the user**

**#num = int(input("Enter a number: "))**

**factorial = 1**

**# check if the number is negative, positive or zero**

**if num < 0:**

**print("Sorry, factorial does not exist for negative numbers")**

**elif num == 0:**

**print("The factorial of 0 is 1")**

**else:**

**for i in range(1,num + 1):**

**factorial = factorial\*i**

**print("The factorial of",num,"is",factorial)**

**Output : num = 5, The factorial of 5 is 120**

--------------------------------------------------------------------------------------------------------------------------------------

SQL -

1) What is the purpose of the GROUP BY clause in SQL?

**Answer - The GROUP BY clause causes the rows of the items table to be collected into groups, each group composed of rows that have identical order\_num values (that is, the items of each order are grouped together). After the database server forms the groups, the aggregate functions COUNT and SUM are applied within each group.**

2) Explain the ACID properties in the context of database transaction and why they are important?

**Answer - In the context of transaction processing, the acronym ACID refers to the four key properties of a transaction: atomicity, consistency, isolation, and durability. All changes to data are performed as if they are a single operation. That is, all the changes are performed, or none of them are.**

**Atomicity – If an action is performed on the data, it should either be executed, completed or not executed at all. Ex: - For example, in an application that transfers funds from one account to another, the atomicity property ensures that, if a debit is made successfully from one account, the corresponding credit is made to the other account.**

**Consistency – Data is in a consistent state when a transaction starts & when it ends. For example, in an application that transfers funds from one account to another, the consistency property ensures that the total value of funds in both the accounts is the same at the start and end of each transaction.**

**Isolation – The intermediate state of a transaction is invisible to other transactions. As a result, transactions that run concurrently appear to be serialized. For example, in an application that transfers funds from one account to another, the isolation property ensures that another transaction sees the transferred funds in one account or the other, but not in both, nor in neither.**

**Durability – After a transaction successfully completes, changes to data persist & are not undone, even in the event of a system failure. For example, in an application that transfers funds from one account to another, the durability property ensures that the changes made to each account will not be reversed.**

3) Create two tables and using all join concepts?

**Answer –**

**-- Create employees table**

**CREATE TABLE #Employees(Employee\_id int, Employee\_name varchar(250), Employee\_DOB date, Department\_ID int)**

**-- Create departments table**

**CREATE TABLE #Departments(Department\_id int, Department\_Name varchar(250))**

**-- Insert values into departments table**

**INSERT INTO #Departments(Department\_id,Department\_Name)**

**VALUES(1,'Human Resources'), (2,'Development'), (3,'Sales'), (4, 'Technical Support')**

**-- Insert values into employees table**

**INSERT INTO #Employees(Employee\_id,Employee\_name, Employee\_DOB,Department\_ID)**

**VALUES (1,'Alan Smith','19890101',1),**

**(2,'Sultan Nader','19920101',1),**

**(3,'Mohd Rasheed','19990101',2),**

**(4,'Brian Wallace','19790101',3),**

**(5,'Peter Hilton','19860101',NULL)**

**Inner Join :-**

**SELECT Employee\_id,Employee\_name, Employee\_DOB, Department\_Name**

**FROM #Departments INNER JOIN #Employees**

**ON #Departments.Department\_id = #Employees.Department\_ID**

**Left Join :-**

**SELECT Employee\_id,Employee\_name, Employee\_DOB, Department\_Name**

**FROM #Employees LEFT JOIN #Departments**

**ON #Departments.Department\_id = #Employees.Department\_ID**

**Right Join :-**

**SELECT Employee\_id,Employee\_name, Employee\_DOB, Department\_Name**

**FROM #Employees RIGHT JOIN #Departments**

**ON #Departments.Department\_id = #Employees.Department\_ID**

**Full Outer Join :-**

**SELECT Employee\_id,Employee\_name, Employee\_DOB, Department\_Name**

**FROM #Employees FULL JOIN #Departments #Departments.Department\_id = #Employees.Department\_ID**

**Cross Join :-**

**SELECT Employee\_id,Employee\_name, Employee\_DOB, Department\_Name**

**FROM #Employees CROSS JOIN #Departments**

4) Explain the concept of normalization in SQL?

**Answer -** **Normalization is a database design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. The purpose of Normalisation in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.**

5) What is correlated subquery in SQL?

**Answer - A correlated subquery is a subquery that refers to a column of a table that is not in its FROM clause. The column can be in the Projection clause or in the WHERE clause. In general, correlated subqueries diminish performance.**

**Ex: -**

SELECT po\_num, ship\_date FROM orders main

WHERE 10 >

(SELECT COUNT (DISTINCT ship\_date)

FROM orders sub

WHERE sub.ship\_date < main.ship\_date)

AND ship\_date IS NOT NULL

ORDER BY ship\_date, po\_num;

**The subquery is correlated because the number that it produces depends on main.ship\_date, a value that the outer SELECT produces. Thus, the subquery must be re-executed for every row that the outer query considers.**

--------------------------------------------------------------------------------------------------------------------------------------

EDA -

1) What is the formula for calculating number of bins and bin width in Histogram?

**Answer – We can calculate the number of bins, by taking the square root of the number of data points and round up & can calculate the bin width by dividing the specification tolerance or range (USL-LSL or Max-Min value) by the # of bins.**

#### **Bin Width = (Max-Min)/sqrt(n)**

2) Which functions are used for labeling X- axis and Y- axis when using matplotlib library ?

**Answer - With Pyplot, you can use the xlabel() and ylabel() functions to set a label for the x- and y-axis.**

3) What formula are used for calculating lower limit and upper limit in Boxplot?

**Answer - Upper and lower fences cordon off outliers from the bulk of data in a set. Fences/limits are usually found with the following formulas:**

**Upper limit = Q3 + (1.5 \* IQR)**

**Lower limit = Q1 — (1.5 \* IQR).**

4) What is X-axis and Y-axis represent in ACF and PACF plots?

**Answer - ACF plot for a series. The x axis of the ACF plot indicates the lag at which the autocorrelation is computed; the y axis indicates the value of the correlation (between −1 and 1).**

5) What is the only calculation that can be performed on categorical data in EDA?

**Answer – Dummy Variable Creation –**

**One-Hot encoding – pd.get\_dummies(df, drop\_first = True)**

**from sklearn.preprocessing import OneHotEncoder**

**Label encoding - from sklearn.presprocessing import LabelEncoder**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**15/07/2023** PYTHON -

1) Write a program that will convert celsius value to fahrenheit?

**Answer - celsius = 47**

**# Converting the temperature to**

**# fahrenheit using the formula**

**fahrenheit = (celsius \* 1.8) + 32**

**# printing the result**

**print('%.2f Celsius is equivalent to: %.2f Fahrenheit'**

**% (celsius, fahrenheit))**

**Output**

**47.00 Celsius is equivalent to: 116.60 Fahrenheit**

2) When would you choose to use append() over extend() and vice versa?

**Answer - append() adds a single element to the end of the list while . extend() can add multiple individual elements to the end of the list.**

**Ex:- Append**

**# Define the list**

**>>> nums = [1, 2, 3, 4]**

**# Add the integer 5 to the end of the existing list**

**>>> nums.append(5)**

**# See the updated value of the list**

**>>> nums**

**[1, 2, 3, 4, 5]**

**# Extend**

**>>> nums = [5.6, 7.44, 6.75, 4.56, 2.3]**

**>>> new\_values = [2.3, 9.6, 4.564, 7.56]**

**# This is where the magic occurs! No more for loops**

**>>> nums.extend(new\_values)**

**# The list was updated with individual values**

**>>> nums**

**[5.6, 7.44, 6.75, 4.56, 2.3, 2.3, 9.6, 4.564, 7.56]**

3) What is the purpose of the python garbage collector, and how does it reclaim memory from objects that are no longer referenced?

**Answer - Garbage collection is to release memory when the object is no longer in use. This system destroys the unused object and reuses its memory slot for new objects. You can imagine this as a recycling system in computers. Python has an automated garbage collection.**

**--------------------------------------------------------------------------------------------------------------------------------------**

SQL -

1) Explain difference between a primary key and a unique key in SQL.

**Answer - Both keys provide a guaranteed uniqueness for a column or a set of columns in a table or relation. The main difference among them is that the primary key identifies each record in the table, and the unique key prevents duplicate entries in a column except for a NULL value.**

2) What is the purpose of the ROLLUP operator in SQL?

**Answer - ROLLUP in SQL is an extension of the**[**GROUP BY clause**](https://www.scaler.com/topics/sql/group-by-in-sql/)**, this SQL ROLLUP option allows the user to include extra rows that represent the subtotals, which are usually referred to as super-aggregate rows, along with a grand total row**.

3) How do you add or remove columns from an existing table using the ALTER statement in SQL?

Answer -

1. **ALTER TABLE table\_name ADD (Columnname\_1 datatype, Columnname\_2 datatype, … Columnname\_n datatype);**
2. **ALTER TABLE table\_name. DROP COLUMN column\_name;**
3. **ALTER TABLE table\_name. ALTER COLUMN column\_name column\_type;**

4) How do you retrieve data from a table in MySQL?

**Answer - Data can be fetched from MySQL tables by executing SQL SELECT statement through PHP function mysql\_query. You have several options to fetch data from MySQL. The most frequently used option is to use function mysql\_fetch\_array().**

**--------------------------------------------------------------------------------------------------------------------------------------**

EDA -

1) What is the primary and secondary use of Histogram?

**Answer – Primary purpose – Understand the shape of data distribution.**

**Secondary purpose – Identify Outliers.**

2) In order to visualize bar plot using Matplotlib and Seaborn, which python functions are used ?

**Answer – Matplotlib - plt.bar(x, height, width, bottom, align)**

**Seaborn - sns.barplot(x = 'time',y = 'total\_bill',data = df)**

3) What is formula for range?

**Answer – Range = Maximum Value–Minimum Value.**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**17/07/2023** PowerBI -

1) Explain the concept of calculated measures in PowerBI?

**Answer - Creating Power BI measures is often called “calculated measures,” which use DAX expressions to calculate new values from the existing table.**

**A measure is evaluated in the context of the cell evaluated in a report or in a DAX query, whereas a calculated column is computed at the row level within the table it belongs to. The context of the cell depends on user selections in the report or on the shape of the DAX query.**

2) What is the difference between calculated columns and calculated measures in PowerBIr?

**Answer - In general, calculated columns are more resource-intensive, as the calculation is performed for every row in the table, and the result is stored. Measures, on the other hand, are more flexible, as they can be calculated on the fly and updated in real-time as the underlying data changes.**

**Large Data: As mentioned earlier, calculated columns can consume a large amount of memory and slow down the performance of your data model if your data volume is large.**

**Measures, on the other hand, perform calculations on-the-fly and only aggregate the data that is necessary for the current view.**

**Reusability: Measures are reusable and can be used in multiple visualizations and reports, while calculated columns are tied to a specific table and cannot be easily reused.**

**Update Frequency: Calculated columns are only updated when the data is refreshed, while measures are updated in real-time as the underlying data changes.**

**Data Types: Calculated columns can handle a wider variety of data types, including text and date/time values, while measures are limited to numeric data types.**

**Complexity: Measures can handle more complex calculations than calculated columns, including complex aggregations and dynamic calculations based on user selections and filters.**

3) Explain the concept of data modelling in PowerBI?

**Answer - Power BI data modelling is the process of creating a relationship between common columns of multiple tables. If the column headings are the same across tables, then Power BI auto-detects the relationship between tables. Using these columns, we can merge the tables as well.**

4) How can you implement row-level security in PowerBI?

**Answer - From the ribbon, select Manage roles. From the Manage roles window, select New to create a new role. Under Roles, provide a name for the role and select enter. Under Select tables, select the table you want to apply a row-level security filter to.**

**To implement RLS, you need to create security policies, which are rules that define what data is visible to which users. The policies are usually defined using predicates, which are logical expressions that determine which rows in a table are visible to a user based on their role or permissions.**

5) How can you optimise the performance of PowerBI reports and dashboards?

**Answer –**

1. **Limit the number of visuals in dashboards and reports**

**Using too many visuals on a single report slows report performance. Limit widget visuals to eight per report page and grids to one per page. Limit tiles to 10 per dashboard. In general, limit pages 30 points, assuming each type of visual is worth a different number of points:**

* + **Cards: 1**
  + **Gauges: 2**
  + **Charts: 3**
  + **Maps: 3**
  + **Grids: 5**

1. **Remove unnecessary interactions between visuals**

**By default, all visuals on a report page can interact with one another. By disabling unnecessary interactions, you reduce the number of queries fired at the back end, which improves report performance.**

1. **Use**[**on-premises data gateway**](https://docs.microsoft.com/en-us/power-bi/connect-data/service-gateway-onprem)**instead of Personal Gateway**

**Personal Gateway imports data into Power BI. On-premises data gateway (also known as Enterprise Gateway) imports nothing, which is more efficient when working with large databases.**

1. **Use separate gateways for**[**Power BI service live connection**](https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-report-lifecycle-datasets)**and**[**scheduled data refresh**](https://docs.microsoft.com/en-us/power-bi/connect-data/refresh-scheduled-refresh)

**If you use the same gateway for a scheduled refresh and a live connection, live connection performance will slow down during the scheduled refresh.**

1. **Test custom visual performance before use**

**Custom visuals can perform poorly when handling large datasets or complex aggregations. Uncertified custom visuals are generally not tested by the Power BI team. If a custom visual performs poorly, consider replacing it with a different visual.**

1. **Limit complex measures and aggregations in data models**

**Create calculated measures instead of calculated columns. Where possible, push calculated columns and measures to the source. The closer they are to the source, the faster they are likely to perform.**

1. **Use Star schema instead of Snowflake schema when possible**

**Snowflake schema has a complex query structure, making it difficult to implement changes. Star schema is easy to read, uses fewer joins, and tends to reduce data redundancy.**

1. **Use slicers sparingly**

**Slicers are a great way of allowing users to navigate data, but they come at a performance cost. Each slicer generates two queries: one fetches the data, the other fetches selection details. Adding too many slicers drastically slows performance. To remove unnecessary slicers, use the Filter pane to evaluate which slicers are used the least.**

1. **Ensure reports and data sources are in the same region**

**With the tenant and data source in the same region, you can limit the effects of network latency. Sharing a region ensures faster data transfer and faster query execution.**

1. **Import only necessary fields and tables instead of entire data sets**

**Ensure the model is as narrow and lean as possible. Power BI works on columnar indexes, meaning longer and leaner tables perform better. When you need to import a large table, partition it and process multiple partitions in parallel.**

1. **Use templates (.PBIT files) to speed up and standardize report development instead of starting with an empty .PBIX file**

**Templates enable you to develop branded reports faster. With templates, you can save custom color palettes and themes, ensuring corporate branding is pre-applied to all pages. Templates also automatically connect to commonly used data sources and offer commonly used DAX measures.**

1. **Reduce queries**

**Reduce the number of queries sent by Power BI using the Query reduction settings. For slicers, select the “Add an Apply button to each slicer to apply changes when you’re ready” option. For filters, select “Add a single Apply button to the filter pane to apply changes at once (preview).”**

1. **Avoid bi-directional and many-to-many relationships against high cardinality columns**

**Many-to-many and bi-directional relationships navigate more pathways and check more data points. As a result, bi-directional relationships against high-cardinality columns negatively impact report performance.**

1. **Avoid using floating point data types**

**Floating point data types can result in unpredictable round-off errors and can decrease the performance of reports.**

1. **Replace the auto-generated date table with a custom date table in your model**

**Using a date table enables you to leverage the time-series function in Power BI. However, the auto-generated date table creates a date table for each date column, drastically increasing the model size. You can use a single date table to reduce model size, placing any required relationships in fact tables. When building your date table, split the date and time to improve data compression. Find out how to turn off the auto-generated date table**[**here**](https://www.youtube.com/watch?v=xu3uDEHtCrg)**and how to set and use custom date tables**[**here**](https://docs.microsoft.com/en-us/power-bi/transform-model/desktop-date-tables)**.**

1. **Set IsAvailableinMdx to false on non-attribute columns**

**Disable attribute hierarchy for measure columns and for any columns you don’t want end users to use. This reduces data size and load time.**

1. **Reduce the amount of data loaded on page load**

**Use bookmarks, drill-through pages, and tooltips to reduce the amount of data loaded on page load. This improves page load time for landing pages.**

1. **Use report backgrounds for static images**

**For static images, use report backgrounds rather than multiple visuals. This conveys the same information to the end user at a fraction of the performance cost.**

**--------------------------------------------------------------------------------------------------------------------------------------**

Data Preprocessing -

1) Converting one data type into another is called as?

**Answer - Converting one datatype into another is known as type casting or, type-conversion. For example, if you want to store a 'long' value into a simple integer then you can type cast 'long' to 'int'.**

2) What do you mean by zero or near zero variance?

**Answer - a zero variance variable is one whose values are all the same constant variable and a near-zero variance (NZV) variable is one where almost all values are constant and only a few have values that differ from that constant.**

**Usually, we ignore the columns with same entries throughout or if majority of the entries are same. Ex: - All entries of a column called Country show the same name as USA.**

3) When do we say that data has duplicates?

**Answer - Checking for duplicates**

**duplicated() method to facilitate finding duplicate data. This checks if the whole row appears elsewhere with the same values in each column. This method returns a Boolean Series , where each entry represents whether or not the row is a duplicate.**

**Getting single source of truth from various entries is one such example.**

**data.duplicated() is used to check for duplicates.**

**Or data.drop\_duplicates() to drop the duplicated values if they exist.**

4) Which python function is used for finding out the missing values?

**Answer – Imputation method is used for finding out the missing values.**

**Ex: - from sklearn.impute import SimpleImputer**

**SimpleImputer(missing\_values = np.nan,strategy=’mean’)**

**SimpleImputer(missing\_values = np.nan,strategy=’median’)**

**SimpleImputer(missing\_values = np.nan,strategy=’most\_frequent’)**

5) Define Encoding?

**Answer - Encoding is a technique of converting categorical variables into numerical values so that it could be easily fitted to a machine learning model.**

**One-hot Encoding – pd.get\_dummies(df,drop\_first=True)**

**from sklearn.preprocessing import OneHotEncoder**

**Label Encoding - from sklearn.preprocessing import LabelEncoder**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**18/07/2023** PowerBI -

1) What is PowerBI?

**Answer - Power BI is a business analytics tool developed by Microsoft. It allows users to connect, visualize, and analyze data from various sources, including cloud services and on-premises databases.**

2) Which of the following is a way to create a new visualisation in PowerBI?

**Answer – Charts**

3) In PowerBI, how do you make a calculated column?

**Answer – Using DAX Expressions.**

4) What distinguishes a measure in PowerBI from a computed column?

**Answer - a calculated column belongs to a single table, while a measure belongs to the whole data model. A calculated column is evaluated in a row context (row by row, like in an excel table), while a measure is evaluated in the filter context.**

5) How can you schedule data refresh in PowerBI service?

**Answer - The Power BI service allows users to schedule the refresh of a Power BI dataset. This can be done by configuring a refresh schedule in the dataset settings.**

--------------------------------------------------------------------------------------------------------------------------------------

Data Preprocessing -

1) Which python function is used for converting one data type into another?

**Answer – astype() function is used for converting one data type into another.**

2) When do you use transformations?

**Answer - Feature transformation is an important step in feature engineering of numeric data and it is used to handle skewed data. Variables with skewed distributions can be hard to incorporate into some modelling procedures, especially in the presence of other variables that are not skewed. In this case, applying a transformation to reduce skew will improve performance of models.**

3) Formulae for standardization, Normalization and Robust Scaler?

**Answer – Standardization = (X – Mean)/Std.Dev**

**Normalization = (X – min(X))/(max(X) – min(X))**

**Robust Scalar = (X – Median(X)) / IQR(X)**

4) Which library is imported for performing winsorization?

**Answer – from feature\_engine.outliers import Winsorizer**

**Winsorizer(capping\_method = ‘iqr’, tail = ‘both’, fold = 3, variables = ‘cols’)**

5) What is the purpose of data scaling and which module is imported for scaling the data?

**Answer – Data Scaling/Feature Scaling/Feature Shrinking will make the data scale free and unitless, hence it is used.**

**Standard Scaler = from sklearn.preprocessing import StandardScalar**

**Min-Max Scaler / Range Method = from sklearn.preprocessing import MinMaxScalar**

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**19/07/2023** PowerBI -

1) How can you create dynamic visuals and reports in PowerBI?

**Answer – We can create dynamic visuals & reports in Power BI using DAX.**

2) What is the purpose of the PowerBI DirectQuery mode, and when should it be used?

**Answer - DirectQuery lets you build visualizations over very large datasets, where it would be unfeasible to import all the data with pre-aggregation. DirectQuery reports always use current data.**

3) What is the purpose of the PowerBI Data Analysis Expressions (DAX) language, and how is it used?

**Answer - Data Analysis Expressions (DAX) is a programming language that is used throughout Microsoft Power BI for creating calculated columns, measures, and custom tables. It is a collection of functions, operators, and constants that can be used in a formula, or expression, to calculate and return one or more values.**

4) What is the difference between the SUM and SUMX function in DAX?

**Answer - The SUMX function calculates the expression for each row in the table and returns the sum of those values. In short, the SUM function is used for simple sum calculations, while the SUMX function is used for more complex calculations that require iteration over a table.**

5) Explain the concept of query folding in power Query.

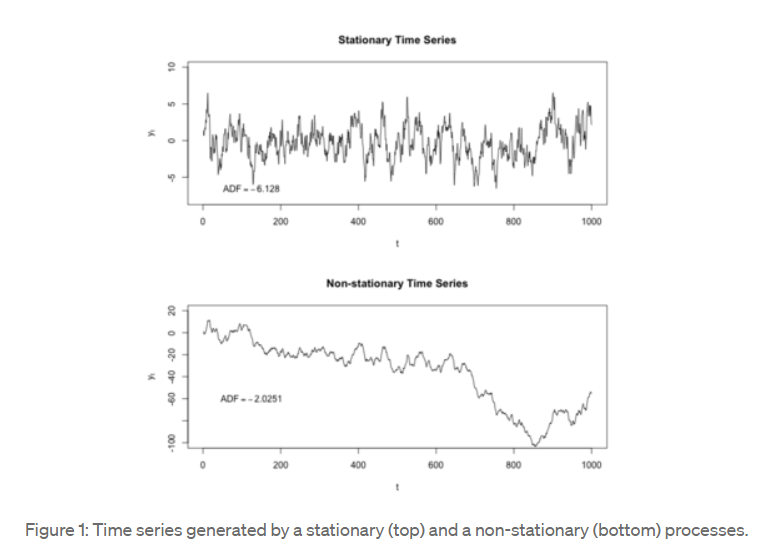
**Answer - Query folding is the ability for a Power Query to generate a single query statement to retrieve and transform source data. The Power Query mashup engine strives to achieve query folding whenever possible for reasons of efficiency.**

Data Preprocessing -

1) What is stationarity in time series data and why is it important?

**Answer - In the most intuitive sense, stationarity means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the way it changes does not itself change over time.**

**Ex: -**



2) Difference between Box Cox and Yeo Johnson transformations?

**Answer - Both Box-Cox and Yeo-Johnson transform non-normal distribution into a normal distribution. However, Box-Cox requires all samples to be positive, while Yeo-Johnson has no restrictions.**

3) How do you handle missing values in time series data and what are some techniques for imputing missing values in time series data?

**Answer – Missing values in time series data are handled by using Imputation techniques. Mean, Median ,Mode, Random, Hot Deck, Regression & KNN are some of the Imputation methods.**

4) Which functions are used for identifying and removing duplicates?

**Answer – data.duplicated() is to check for duplicates & data.drop\_duplicates() is to remove duplicates.**

5) What are the different capping methods in Winsorization?

**Answer - Different capping method are 'gaussian', 'iqr' or 'quantiles'.**

**'gaussian': the transformer will find the maximum and / or minimum values to cap the variables using the Gaussian approximation.**

**'iqr': the transformer will find the boundaries using the IQR proximity rule.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**20/07/2023** Power BI -

1) What are the different types of visuals available in Power BI?

**Answer –**

* **Area charts: Basic (Layered) and Stacked. ...**
* **Bar and column charts. ...**
* **Combo charts. ...**
* **Doughnut charts. ...**
* **Funnel charts. ...**
* **Gauge charts. ...**
* **KPIs. ...**
* **Line charts.**

2) How can you create a custom visual in Power BI?

**Answer - Importing Organizational Custom Visuals  
  
Step 1: Click on the ellipse’s icon in the Visualizations pane. Select Import from marketplace option.**

**Step 2: Select the MY ORGANIZATION section. In this section, you will find all the custom visuals developed and uploaded by the Power BI Admin from your organization.**

3) Explain the concept of drill-through in Power BI?

**Answer - You can create a drillthrough button in Power BI. This button drills through to a page with details that are filtered to a specific context. One way to drill through a report is to right-click in a visual. However, if you want the drillthrough action to be more obvious, you can create a drillthrough button instead.**

4) How can you create a report-level filter in Power BI?

**Answer - Select Edit to open the report in Editing view.**

1. **Open the Visualizations and Filters pane and the Fields pane, if they are not already open.**
2. **From the Fields pane, select the field you want to add as a new report-level filter, and drag it into the Report level filters area.**
3. **Select the values you want to filter.**

5) What is the purpose of bookmarks in Power BI?

**Answer - Bookmark is a Power BI feature that enables you to capture the current state of your report page. And, you can then refer to this captured state, as a “special” page — through the Action property on your buttons, images, shapes, etc.**

**--------------------------------------------------------------------------------------------------------------------------------------**

Data Preprocessing -

1) What is the difference between data normalisation and standardization?

**Answer - In contrast to Normalization, Standardization does not always have a bounding range; therefore, any outliers in your data won't be impacted by it. Scales for normalization fall between [0,1] and [-1,1]. Standardization has no range restrictions.**

**StandardScaler is used for scaling features of a dataset, while Normalizer is used for scaling the rows of a dataset.**

2) What is the difference between One Hot Encoding and Label Encoding?

**Answer - 1. Label Encoding: Assign each categorical value an integer value based on alphabetical order.**

**2. One Hot Encoding: Create new variables that take on values 0 and 1 to represent the original categorical values.**

3) What do you mean by discretization?

**Answer – Discretization / Binning / Grouping is a Data Preparation method of converting the continuous data to discrete data.**

4) How do you normalize text data for natural language processing (NLP)?

**Answer - Text normalization is a pre-processing step aimed at improving the quality of the text and making it suitable for machines to process. Five main steps in text normalization are case normalization, tokenization and stop word removal, Parts-of-Speech (POS) tagging, and stemming.**

5) What techniques are used for converting imbalanced data into balanced data?

**Answer – Techniques used for converting imbalanced data into balanced data are as follows –**

**Random Resampling – Oversampling & Undersampling, Bootstrap Resampling, K Fold Cross Validation, Repeated K Fold Cross Validation, Stratified K-Fold Cross Validation, LOOCV(Leave One Out Cross Validation N-Fold), SMOTE(Synthetic Minority Oversampling Technique), MSMOTE(Modified SMOTE), Cluster Based Sampling, Ensemble Techniques.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**22/07/2023** Power BI -

1) What is the difference between calculated columns and calculate tables in Power BI?

**Answer - A**[**calculated column**](https://go.christiansteven.com/bi-blog/power-bi/automate-your-company-reports-using-power-bi)**is a column of data that is added to an existing table in your model. This is done either in report view or data view using a**[**DAX**](https://go.christiansteven.com/bi-blog/what-is-dax-what-are-the-benefits-of-using-variables-in-dax)**formula to determine the data that is displayed. A calculated column can be a useful tool if you are looking to combine data from multiple columns in a table or across tables.**

**Calculated tables are created in Power Query. Calculated tables increase the size of the data model. Calculated tables cannot include calculated columns.**

2) What is the purpose of the CALCULATE function in Power BI?

**Answer - According to the Microsoft Power BI documentation, the CALCULATE function forms part of the filter function category and is defined as "evaluating an expression in a modified filter context."**

3) How are DAX and Power Query different from each other in Power BI?

**Answer – DAX uses functions to work on data that is stored in tables. Power Query (M) is used to query data sources, clean, and load data.**

4) What is Power Query in Power BI?

**Answer - Power Query is the data connectivity and data preparation technology that enables end users to seamlessly import and reshape data from within a wide range of Microsoft products, including Excel, Power BI, Analysis Services, Dataverse, and more.**

5) What are the 3 views in Power BI?

**Answer - Along the left side of Power BI Desktop are icons for the three Power BI Desktop views: Report, Data, and Model, from top to bottom.**

Data Preprocessing -

1) What does X-axis and Y-axis represent in Q-Q plot?

**Answer - The normal distribution is the base distribution and its quantiles are plotted along the x-axis as the “Theoretical Quantiles” while the sample quantiles are plotted along the y-axis as the “Sample Quantiles.”**

2) Which formulae is used to calculate lower and upper limits using gaussian capping method in box plot?

**Answer –**

**IQR = Q3-Q1**

**Lower Limit = Q1 - 1.5\*IQR**

**Upper Limit = Q3 + 1.5\*IQR**

3) What are the common techniques used for handling missing data in dataset?

**Answer – Missing Value Imputation.**

* 1. **Deletion Methods** 
     1. **Complete Case Analysis -> Case-Wise Deletion or List-Wise Deletion.**
     2. **Available Case Analysis -> Pair-Wise Deletion.**
  2. **Single Imputation Methods**
     1. **Mean Imputation**
     2. **Median Imputation**
     3. **Mode Imputation**
     4. **Random Imputation**
     5. **Hot-Deck Imputation**
     6. **Regression Imputation**
     7. **KNN Imputation**
  3. **Model Based Methods**
     1. **Maximum Likelihood (EM Algorithm)**
     2. **Multiple Imputation**

4) What is the difference between standardization and normalization in data preprocessing?

**Answer - In contrast to Normalization, Standardization does not always have a bounding range; therefore, any outliers in your data won't be impacted by it. Scales for normalization fall between [0,1] and [-1,1]. Standardization has no range restrictions.**

**StandardScaler is used for scaling features of a dataset, while Normalizer is used for scaling the rows of a dataset.**

5) How do you handle outliers in a dataset?

**Answer -** **Handling Outliers -> Different Methods**

* 1. **3R Technique (Rectify, Retain, Remove)**
  2. **Masking – False Negative**
  3. **Swamping – False Positive**
  4. **Winsorization – from feature\_engine.outliers import Winsorizer**

**Winsorizer(capping\_method=’iqr’,tail=’both’,fold=3,variables=’cols’)**

* 1. **Trimming –**

**# Detection of Outliers**

**IQR = df['col']. quantile(0.75) - df['col'].quantile(0.25) # IQR - Inter quartile range IQR = Q3-Q1**

**lower\_limit = df['col'].quantile(0.25) - (IQR \* 1.5) # Q1 - 1.5 \* IQR**

**upper\_limit = df['col'].quantile(0.75) + (IQR \* 1.5) # Q3 + 1.5 \* IQR**

**outliers\_df = np.where(df['col'] > upper\_limit,True,np.where(df['col'] < lower\_limit,True,False))**

**# if value is greater than upper limit consider it as outliers and if the value is less than lower limit consider it as outliers**

**df\_trimmed = df.loc[~(outliers\_df),] # ~ means not - it shows all false value (not outliers)**

**df.shape, df\_trimmed.shape**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**24/07/2023** Data Preprocessing -

1) What is dimensionality reduction, and what techniques can be used for it?

**Answer - Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible. In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still preserves the essence of the original data.**

**Linear patterns are handled by these techniques - > Linear Discriminant Analysis (LDA), Non-Negative Matrix Factorization (NMF), Independent Component Analysis (ICA), Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Factor Analysis.**

**Non-Linear patterns are handled by Stochastic Neighbour Embedding (SNE), T-Distributed Stochastic Neighbour Embedding(t-SNE).**

2) What is the role of data preprocessing in machine learning?

**Answer - In conclusion, preprocessing data before applying it to a machine learning algorithm is a crucial step in the ML workflow. It helps to improve the accuracy, reduce the time and resources required to train the model, prevent overfitting, and improve the interpretability of the model.**

3) How do you handle categorical variables in data preprocessing?

**Answer – Categorical variables are handled as below.**

**Encoding is a technique of converting categorical variables into numerical values so that it could be easily fitted to a machine learning model.**

**Dummy Variable Creation –**

**One-Hot encoding – Create new variables that take on values 0 and 1 to represent the original categorical values.**

**pd.get\_dummies(df, drop\_first = True)**

**from sklearn.preprocessing import OneHotEncoder**

**1. Label Encoding: Assign each categorical value an integer value based on alphabetical order.**

**from sklearn.presprocessing import LabelEncoder**

4) What are the common techniques used for handling missing values in a dataset?

**Answer – Missing Value Imputation.**

* 1. **Deletion Methods** 
     1. **Complete Case Analysis -> Case-Wise Deletion or List-Wise Deletion.**
     2. **Available Case Analysis -> Pair-Wise Deletion.**
  2. **Single Imputation Methods**
     1. **Mean Imputation**
     2. **Median Imputation**
     3. **Mode Imputation**
     4. **Random Imputation**
     5. **Hot-Deck Imputation**
     6. **Regression Imputation**
     7. **KNN Imputation**
  3. **Model Based Methods**
     1. **Maximum Likelihood (EM Algorithm)**
     2. **Multiple Imputation**

**Imputation method is used for finding out the missing values.**

**Ex: - from sklearn.impute import SimpleImputer**

**SimpleImputer(missing\_values = np.nan,strategy=’mean’)**

**SimpleImputer(missing\_values = np.nan,strategy=’median’)**

**SimpleImputer(missing\_values = np.nan,strategy=’most\_frequent’)**

5) What is feature scaling and why is it necessary?

**Answer - Data Scaling/Feature Scaling/Feature Shrinking will make the data scale free and unitless, hence it is used.**

**Standard Scaler = from sklearn.preprocessing import StandardScalar**

**Min-Max Scaler / Range Method = from sklearn.preprocessing import MinMaxScalar**

**StandardScaler is used for scaling features of a dataset, while Normalizer is used for scaling the rows of a dataset.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**26/07/2023 -** Deep Learning -

1) What Deep Learning architectures do you know for object Detection?

**Answer - Deep learning-based approaches to object detection use convolutional neural networks architectures such as RetinaNET, YOLO, CenterNet, SSD, and Region Proposals.**

2) How Region -Based CNN are used in Object Detection?

**Answer - In R-CNN, selective Search approach is applied to find region proposals for classification. selective Search can capture any possible scales and less computational complexity. 2000 regions are selected from selective search.**

3) What Image Thresholding methods do you know?

**Answer –**

* 1. **Simple Thresholding is also known as Binary thresholding. This technique sets a threshold value and compares each pixel to that particular threshold value. If the pixel value is less than or equal to the assigned threshold, then the pixel value is set to zero or to the maximum value.**
  2. **Binary Inverse thresholding – This technique is opposite of binary thresholding. In this technique, if the pixel value is greater than the assigned threshold, then the value is set to zero, otherwise it is set to maximum value.**
  3. **Truncate Thresholding – If the pixel value is greater than the threshold, then it is truncated to the assigned threshold, else It is left the same.**
  4. **Threshold to zero – If the pixel value is less than the threshold value, then it is set to zero.**
  5. **Threshold to zero Inverse - If the pixel value is greater than the threshold value, then it is set to zero.**

4) What would you do if you need to train an Image Classification network and do not have enough data?

**Answer –**

* **4 ways to handle insufficient data**
  + 1. **Model Complexity - Model complexity is nothing but building a simple model with fewer parameters. This method is less susceptible to over-fitting. Example: Naive Bayes, Linear Regression.**
    2. **Transfer Learning - Transfer Learning is used in the case of Deep Learning and Neural Networks. It uses a pre-built model, which is then tweaked on the small dataset that you have.**
    3. **Data Augmentation - Data Augmentation helps to tweak (make slight improvements) to get new images.**

**It takes the pre-existing samples and changes them in some way to  
create new samples and increase the number of training samples and  
typically used with Image data.**

**Disturb images in some way to generate new images, such as,**

* + **Scaling**
  + **Rotation**
  + **Affine Transforms**

**These image processing options are often used as pre-processing  
techniques to make image classification models built using CNN are robust**

* + 1. **Synthetic Data – Synthetic data generally refers to artificially generating samples which mimic the real-world data (it is one only if we have a good understanding of features). This may induce bias in existing data.**

5) How can you convert a dense Layer of a CNN into a Fully Convolutional Layer?

**Answer - A fully convolution network can be built by simply replacing the FC layers with their equivalent Conv layers. In the example of VGG16 we can do so by first removing the last four layers. One way to do so is to pop layers from the model. In the model stack, each popping will remove the last layer.**

**--------------------------------------------------------------------------------------------------------------------------------------**

Tableau -

1) How would you handle a large dataset in Tableau that exceeds the software's memory limitations?

**Answer –**

**Option 1:  
Use extract refresh to update hyper files on Tableau Server side instead of publishing hyper files from the outside of Tableau Server.  
  
Option 2:  
Reduce the size of the extracts (.hyper files).  
  
Option 3:  
Add the memory of Tableau Server machine.  
  
Option 4:  
Add a dedicated Data Engine (Hyper) node.**

​

2) Can you explain how Tableau's level of detail (LOD) expressions work?

**Answer - Level of Detail (LOD) expressions are used to run complex queries involving many dimensions at the data source level instead of bringing all the data to Tableau interface. A simple example is adding dimension to an already calculated aggregate value.**

## **Types of LOD**

**There are three main types of LOD expressions.**

* **FIXED LOD This expression computes values using the specified dimensions without reference to any other dimensions in the view.**
* **INCLUDE LOD This level of detail expressions compute values using the specified dimensions in addition to whatever dimensions are in the view.**
* **EXCLUDE LOD These levels of detail expressions subtract dimensions from the view level of detail.**

3) How can Power BI row-level security be implemented?

**Answer - From the ribbon, select Manage roles. From the Manage roles window, select New to create a new role. Under Roles, provide a name for the role and select enter. Under Select tables, select the table you want to apply a row-level security filter to.**

**To implement RLS, you need to create security policies, which are rules that define what data is visible to which users. The policies are usually defined using predicates, which are logical expressions that determine which rows in a table are visible to a user based on their role or permissions.**

4) Explain the concept of data blending in Tableau and when you would use it?

**Answer - Data Blending is a very powerful feature in Tableau. It is used when there is related data in multiple data sources, which you want to analyse together in a single view. It is a method for combining data that supplements a table of data from one data source with columns of data from another data source.**

5) How would you create a calculated field in Tableau, and provide an example of when it might be necessary?

**Answer – Ex : -**

## **Step 1: Create the calculated field**

1. **In a worksheet in Tableau, select Analysis > Create Calculated Field.**
2. **In the Calculation Editor that opens, give the calculated field a name.**

**In this example, the calculated field is called Profit Ratio.**

## **Step 2: Enter a formula**

1. **In the Calculation Editor, enter a formula.**

**This example uses the following formula:**

**SUM([Profit])/SUM([Sales])**

**Formulas use a combination of functions, fields, and operators. To learn more about creating formulas in Tableau, see**[**Formatting Calculations in Tableau**](https://help.tableau.com/current/pro/desktop/en-us/functions_operators.htm)**and**[**Functions in Tableau**](https://help.tableau.com/current/pro/desktop/en-us/functions.htm)**.**

1. **When finished, click OK.**

**The new calculated field is added to the Data pane. If the new field computes quantitative data, it is added to Measures. If it computes qualitative data, it is added to Dimensions.**

**You are now ready to use the calculated field in the view.**

**28/07/2023 -** Deep Learning -

1) Compare Sobel vs Canny techniques for Edge Detection in Computer Vision?

**Answer - Canny edge detection algorithm produces smoother, thinner, and cleaner images than Sobel and Prewitt filters. The input image is smoothened, Sobel filter is applied to detect the edges of the image.**

2) What is the difference between concatenate and add layers in Keras and when would you use each one?

**Answer – In Keras, Add layer adds two input tensor while concatenate appends two tensors.**

**Ex: -**

**import keras**

**import tensorflow as tf**

**import keras.backend as K**

**a = tf.constant([1,2,3])**

**b = tf.constant([4,5,6])**

**add = keras.layers.Add()**

**print(K.eval(add([a,b])))**

**#output: [5 7 9]**

**concat = keras.layers.Concatenate()**

**print(K.eval(concat([a,b])))**

**#output: array([1, 2, 3, 4, 5, 6], dtype=int32)**

3) Can autoencoders be used for feature generation? If yes, how?

**Answer - The general model of autoencoder for feature extraction.**

**The input spikes are used to train the autoencoder. Once trained, the values obtained at the code layer can be extracted as a new feature space that is used in the clustering of the spikes.**

4) How does the hidden layer in a neural network compute?

**Answer - In neural networks, a hidden layer is located between the input and output of the algorithm, in which the function applies weights to the inputs and directs them through an activation function as the output. In short, the hidden layers perform nonlinear transformations of the inputs entered into the network.**

5) What Neural Network Architecture could you use to classify videos?

**Answer - If we are able to take advantage of the temporal nature of videos, we can improve our actual video classification task results using more advanced neural networks architectures such as Long short-term memory (LSTMs) and Recurrent Neural Networks (RNNs).**

**---------------------------------------------------------------------------------------------------------------------------**

Tableau -

1) What are the main products of Tableau?

**Answer –**

**Tableau Products**

* **Tableau Desktop. Tableau Desktop allows for authoring of content for visual analytics and data exploration. ...**
* **Tableau Prep. ...**
* **Tableau Server and Online. ...**
* **User based permissions in Tableau Server and Tableau Online. ...**
* **Tableau Reader. ...**
* **Tableau Public. ...**
* **Tableau Viewer. ...**
* **Tableau Explorer.**

2) What is the purpose of a hierarchy in Tableau, and how can it be created?

**Answer - The Hierarchy in Tableau is an arrangement where the entities are presented at various levels. In common terms, Hierarchy is a system or organization that has many levels from highest to lowest; similarly, in Tableau, we can create hierarchies by bringing one dimension as a level under the principal dimension.**

3) What is the dashboard in Tableau?

**Answer - Data dashboards are a summary of different, but related data sets, presented in a way that makes the related information easier to understand. Dashboards are a type of data visualization, and often use common visualization tools such as graphs, charts, and tables.**

4) What does the image indicate if we have a double database in Tableau?

**Answer –**

| **VISUAL CUE** | **DESCRIPTION** |
| --- | --- |
|  | **The workbook is directly connected to a relational data source or file.** |
|  | **The workbook is connected to a cube (multidimensional) data source. In Tableau,**  **cube data sources are supported only in Windows.** |
|  | **The workbook is connected to an extract that still references the underlying data.** |
|  | **The workbook is connected to an Excel table, sub-table, or named range.** |
|  | **The workbook is connected to a data source that has been published to Tableau**  **Server.** |

5) What sorts of charts are available in Tableau?

**Answer – Pie, Bar, Scatter, Gantt, Histogram, Line chart, Line Graph, Treemap, Waterfall Chart, Map, Pareto, Bubble Chart & Motion Chart.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**31/07/2023** - Deep Learning -

1) What are the main components of a neural network?

**Answer – The main components of a neural network consist of an input later, a processing layer (hidden layer), and an output layer.**

2) What is backpropagation?

**Answer - Backpropagation is just a way of propagating the total loss back into the neural network to know how much of the loss every node is responsible for, and subsequently updating the weights in a way that minimizes the loss by giving the nodes with higher error rates lower weights, and vice versa.**

3) What is overfitting in deep learning? How can it be mitigated?

**Answer - It is a common pitfall in deep learning algorithms in which a model tries to fit the training data entirely and ends up memorizing the data patterns and the noise and random fluctuations.**

**These models fail to generalize and perform well in the case of unseen data scenarios, defeating the model's purpose.**

**Overfitting can be mitigated by below techniques:**

**Train with more data :**

**With the increase in the training data, the crucial features to be extracted become prominent. The model can recognize the relationship between the input attributes and the output variable. The only assumption in this method is that the data to be fed into the model should be clean; otherwise, it would worsen the problem of overfitting.**

### ****Data augmentation :****

**An alternative method to training with more data is data augmentation, which is less expensive and safer than the previous method. Data augmentation makes a sample data look slightly different every time the model processes it.**

### ****Addition of noise to the input data  :****

**Another similar option as data augmentation is adding noise to the input and output data. Adding noise to the input makes the model stable without affecting data quality and privacy while adding noise to the output makes the data more diverse. Noise addition should be done in limit so that it does not make the data incorrect or too different.**

### ****Feature selection :****

**Every model has several parameters or features depending upon the number of layers, number of neurons, etc.  The model can detect many redundant features or features determinable from other features leading to unnecessary complexity. We very well know that the more complex the model, the higher the chances of the model to overfit.**

### ****Cross-validation :****

**Cross-validation is a robust measure to prevent overfitting. The complete dataset is split into parts. In standard K-fold cross-validation, we need to partition the data into k folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining holdout fold as the test set. This method allows us to tune the hyperparameters of the neural network or machine learning model and test it using completely unseen data.**

### ****Simplify data :****

**Till now, we have come across model complexity to be one of the top reasons for overfitting. The data simplification method is used to reduce overfitting by decreasing the complexity of the model to make it simple enough that it does not overfit. Some of the procedures include pruning a decision tree, reducing the number of parameters in a neural network, and using dropout on a neutral network.**

### ****Regularization :****

**If overfitting occurs when a model is too complex, reducing the number of features makes sense. Regularization methods like Lasso, L1 can be beneficial if we do not know which features to remove from our model. Regularization applies a "penalty" to the input parameters with the larger coefficients, which subsequently limits the model's variance.**

### ****Ensembling :****

**It is a machine learning technique that combines several base models to produce one optimal predictive model. In Ensemble learning,  the predictions are aggregated to identify the most popular result. Well-known ensemble methods include bagging and boosting, which prevents overfitting as an ensemble model is made from the aggregation of multiple models.**

### ****Early stopping :****

**This method aims to pause the model's training before memorizing noise and random fluctuations from the data. There can be a risk that the model stops training too soon, leading to underfitting. One has to come to an optimum time/iterations the model should train.**

### ****Adding dropout layers :**** Large weights in a neural network signify a more complex network. Probabilistically dropping out nodes in the network is a simple and effective method to prevent overfitting. In regularization, some number of layer outputs are randomly ignored or “dropped out” to reduce the complexity of the model.

4) What are activation functions, and why are they important in neural networks?

**Answer - The activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias to it. The purpose of the activation function is to introduce non-linearity into the output of a neuron.**

5) Explain yolov8 architecture.

**Answer - YOLOv8 is the latest iteration of the YOLO object detection model, aimed at delivering improved accuracy and efficiency over previous versions. Key updates include a more optimized network architecture, a revised anchor box design, and a modified loss function for increased accuracy.**

**The architecture of YOLOv8 builds upon the previous versions of YOLO algorithms.**

**YOLOv8 utilizes a convolutional neural network that can be divided into two main parts: the backbone and the head.**

**A modified version of the**[**CSPDarknet53**](https://paperswithcode.com/method/cspdarknet53#:~:text=CSPDarknet53%20is%20a%20convolutional%20neural,through%20a%20cross%2Dstage%20hierarchy.)**architecture forms the backbone of YOLOv8. This architecture consists of 53 convolutional layers and employs cross-stage partial connections to improve information flow between the different layers.**

**The head of YOLOv8 consists of multiple convolutional layers followed by a series of fully connected layers.**

**These layers are responsible for predicting bounding boxes, objectness scores, and class probabilities for the objects detected in an image.**

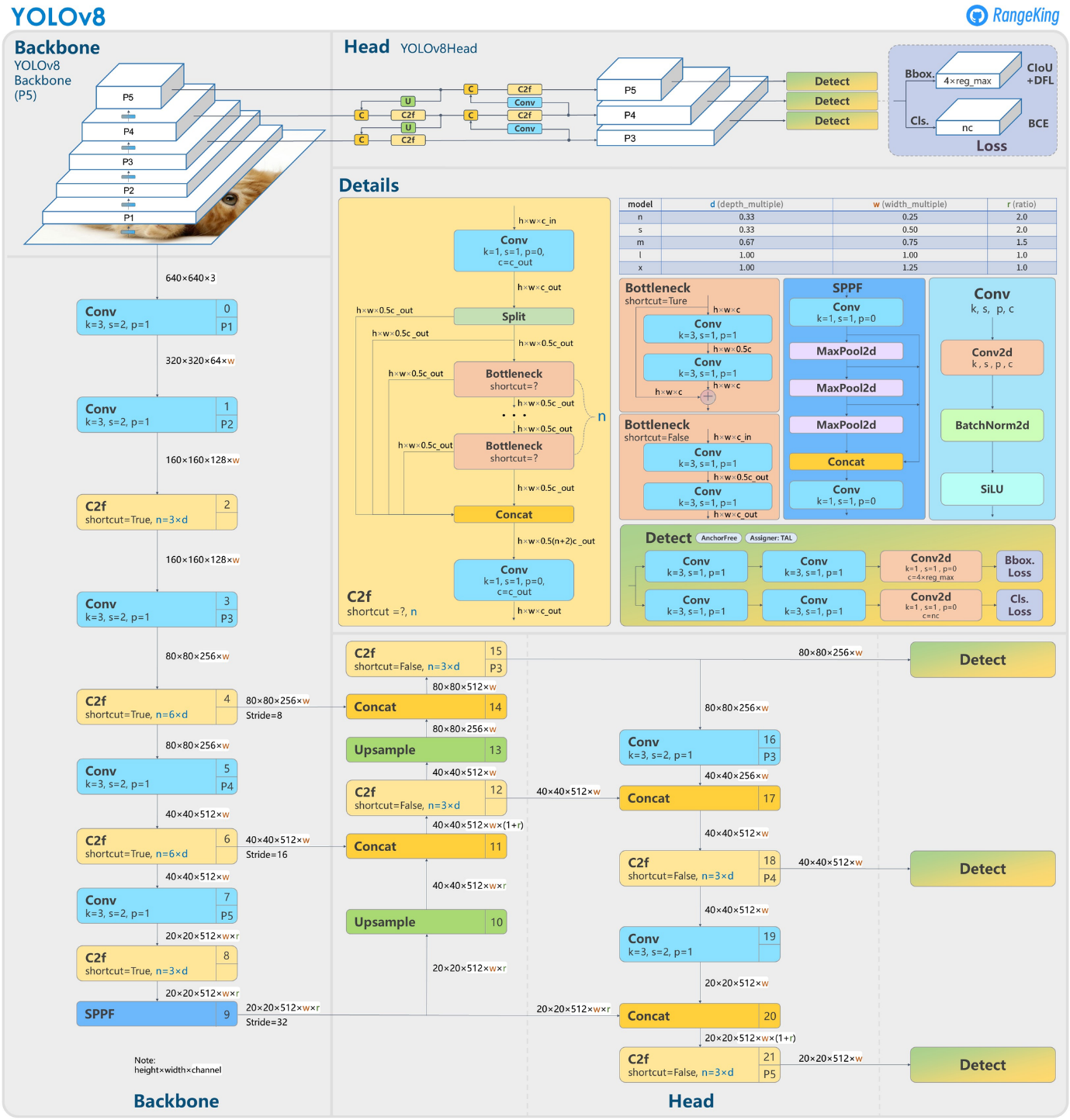
**One of the key features of YOLOv8 is the use of a self-attention mechanism in the head of the network.**

**This mechanism allows the model to focus on different parts of the image and adjust the importance of different features based on their relevance to the task.**

**Another important feature of YOLOv8 is its ability to perform multi-scaled object detection. The model utilizes a feature pyramid network to detect objects of different sizes and scales within an image.**

**This feature pyramid network consists of multiple layers that detect objects at different scales, allowing the model to detect large and small objects within an image.**

## YOLOv8 Architecture: A Deep Dive



**02/08/2023 -** Deep Learning -

1. What are auto encoders?

**Answer - An autoencoder is an unsupervised learning technique for neural networks that learns efficient data representations (encoding) by training the network to ignore signal “noise.” Autoencoders can be used for image denoising, image compression, and, in some cases, even generation of image data.**

2. What is the most used activation function in neural network and why?

**Answer - A rectified linear unit (ReLU) is an activation function that introduces the property of non-linearity to a deep learning model and solves the vanishing gradients issue. "It interprets the positive part of its argument. It is one of the most popular activation functions in deep learning. It is the most widely used activation function. Chiefly implemented in hidden layers of Neural network. Equation: - A(x) = max(0,x). It gives an output x if x is positive and 0 otherwise.**

3. Why CNN is preferred over ANN for Image Classification tasks even though it is possible to solve image classification using ANN?

**Answer - When it comes to image classification jobs, CNNs have several benefits over ANNs. They are more adapted to capture the spatial structure of pictures, which is one of their key advantages. Convolutional layers, such as edges, lines, and forms, are used by CNNs to recognize patterns in pictures.**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Artificial Neural Networks (ANNs)** | **Convolutional Neural Networks (CNNs)** |
| Architecture | Multilayer perceptron | Convolutional layers, pooling layers, and fully connected layers |
| Feature extraction | Hand-crafted or learned features | Learned features through convolutional layers |
| Spatial information | Not specifically designed to capture spatial structure | Specifically designed to capture the spatial structure of images |
| Parameter sharing | No parameter sharing | Parameter sharing through convolutional layers |
| Scalability | Less scalable due to the high number of parameters and overfitting | More scalable due to shared weights and hierarchical representations |
| Training data | Requires a large amount of training data to avoid overfitting | More efficient use of training data due to parameter sharing |
| Transfer learning | Less effective for transfer learning | Effective for transfer learning due to pre-trained models on large datasets |
| Computational efficiency | Less computationally efficient, especially for large images or deep architectures | More computationally efficient, especially for large images or deep architectures |
| Accuracy | Can achieve high accuracy on image classification tasks | Can achieve higher accuracy than ANNs on image classification tasks |

4. What would you do if you did not have enough data to train an Image Classification network?

**Answer –**

* **4 ways to handle insufficient data**
  + 1. **Model Complexity - Model complexity is nothing but building a simple model with fewer parameters. This method is less susceptible to over-fitting. Example: Naive Bayes, Linear Regression.**
    2. **Transfer Learning - Transfer Learning is used in the case of Deep Learning and Neural Networks. It uses a pre-built model, which is then tweaked on the small dataset that you have.**
    3. **Data Augmentation - Data Augmentation helps to tweak (make slight improvements) to get new images.**

**It takes the pre-existing samples and changes them in some way to  
create new samples and increase the number of training samples and  
typically used with Image data.**

**Disturb images in some way to generate new images, such as,**

* + **Scaling**
  + **Rotation**
  + **Affine Transforms**

**These image processing options are often used as pre-processing  
techniques to make image classification models built using CNN are robust**

* + 1. **Synthetic Data – Synthetic data generally refers to artificially generating samples which  
       mimic the real-world data (it is one only if we have a good understanding of features). This may induce bias in existing data.**

5. What are the Various model saving methods? Explain it one by one?

### Answer - ****Way 1: Pickle string****:

**The pickle module implements a fundamental, but powerful algorithm for serializing and de-serializing a Python object structure.**

**Pickle model provides the following functions –**

* **pickle.dump to serialize an object hierarchy, you simply use dump().**
* **pickle.load to deserialize a data stream, you call the loads() function.**

### ****Way 2:**** ****Pickled model as a file using joblib****:

**Joblib is the replacement of pickle as it is more efficient on objects that carry large numpy arrays. These functions also accept file-like object instead of filenames.**

Tableau -

1) What is the difference between a live connection and an extract in Tableau?

**Answer - An extract is a snapshot of the data, the extract will need to be refreshed to receive updates from the original data source, whether it is a local file or an on-premise database. Live connections offer the convenience of real-time updates, with any changes in the data source reflected in Tableau.**

2) What is metadata table in Tableau?

**Answer - The Tableau Metadata API discovers and indexes all of the content on your Tableau Cloud site or Tableau Server, including workbooks, data sources, flows, and metrics. Indexing is used to gather information about Tableau content, or metadata, about the schema and lineage of the content.**

3) What is the difference between a global filter and a context filter in Tableau?

**Answer - A normal filter(global filter) works independently from the other filters and always looks at the entire dataset. The context filter is higher in rank, so it goes first when it comes to filtering.**

4) What is the Tableau Desktop interface?

**Answer - The Tableau interface is the environment you will use to conduct your data analysis and create data visualizations, including all of the software's functionalities.**

5) What is the Tableau Mobile app?

**Answer - Tableau Mobile is the companion app for Tableau Cloud and Tableau Server, giving you access to your Tableau site on the go. Available for Android and iOS, Tableau Mobile lets you interact with the content on your site and discover data insights, even if you are offline.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**04/08/2023 -** Deep Learning -

1.What would you do if you did not have enough data to train an Image Classification network?

**Answer –**

* **4 ways to handle insufficient data**
  + 1. **Model Complexity - Model complexity is nothing but building a simple model with fewer parameters. This method is less susceptible to over-fitting. Example: Naive Bayes, Linear Regression.**
    2. **Transfer Learning - Transfer Learning is used in the case of Deep Learning and Neural Networks. It uses a pre-built model, which is then tweaked on the small dataset that you have.**
    3. **Data Augmentation - Data Augmentation helps to tweak (make slight improvements) to get new images.**

**It takes the pre-existing samples and changes them in some way to  
create new samples and increase the number of training samples and  
typically used with Image data.**

**Disturb images in some way to generate new images, such as,**

* + **Scaling**
  + **Rotation**
  + **Affine Transforms**

**These image processing options are often used as pre-processing  
techniques to make image classification models built using CNN are robust**

* + 1. **Synthetic Data – Synthetic data generally refers to artificially generating samples which  
       mimic the real-world data (it is one only if we have a good understanding of features). This may induce bias in existing data.**

2. Can you discuss the features and benefits of Flask for building and deploying machine learning APIs?

**Answer - Integrated Unit Testing: Flask offers an integrated unit testing feature that helps in faster debugging, robust development, and independence to do experiments.**

**Flexible and Scalable: Flask supports WSGI templates that help in flexibility and scalability in the web development process.**

**Flask helps in implementing a machine learning application in Python that can be easily plugged, extended and deployed as a web application.**

**Flask is based on two key components: WSGI toolkit and Jinja2 template engine. WSGI is a specification for web applications and Jinja2 renders web pages.**

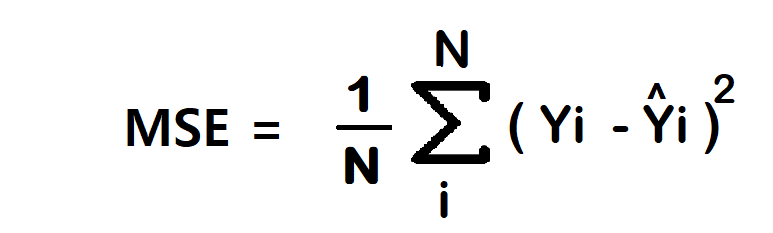
3. Explain some commonly used loss functions in deep learning?

**Answer – Loss Function in Deep Learning**

1. **Regression**
   * **MSE(Mean Squared Error)**
   * **MAE(Mean Absolute Error)**
   * **Hubber loss**
2. **Classification**
   * **Binary cross-entropy**
   * **Categorical cross-entropy**
3. **AutoEncoder**
   * **KL Divergence**
4. **GAN**
   * **Discriminator loss**
   * **Minmax GAN loss**
5. **Object detection**
   * **Focal loss**
6. **Word embeddings**
   * **Triplet loss**

### ****1. Mean Squared Error/Squared loss/ L2 loss****

**The Mean Squared Error (MSE) is the simplest and most common loss function. To calculate the MSE, you take the difference between the actual value and model prediction, square it, and average it across the whole dataset.**

****

#### **Advantage**

* **1. Easy to interpret.**
* **2. Always differential because of the square.**
* **3. Only one local minima.**

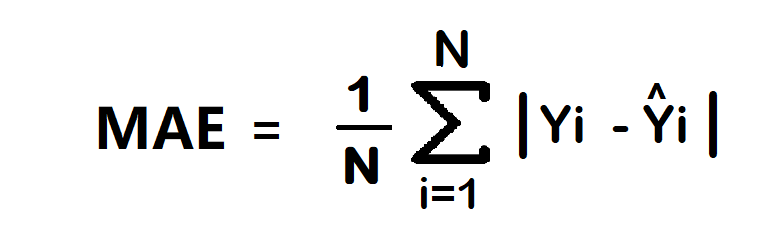
#### **Disadvantage**

* **1. Error unit in the square. because the unit in the square is not understood properly.**
* **2. Not robust to outlier**

***Note – In regression at the last neuron use linear activation function.***

### ****2. Mean Absolute Error/ L1 loss****

**The Mean Absolute Error (MAE) is also the simplest loss function. To calculate the MAE, you take the difference between the actual value and model prediction and average it across the whole dataset.**

****

#### **Advantage**

* **1. Intuitive and easy**
* **2. Error Unit Same as the output column.**
* **3. Robust to outlier**

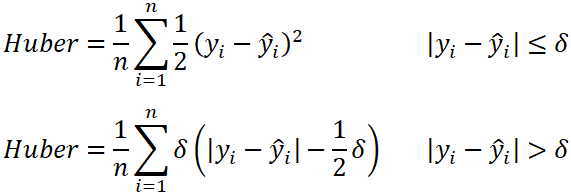
#### **Disadvantage**

* **1. Graph, not differential. we can not use gradient descent directly, then we can subgradient calculation.**

***Note – In regression at the last neuron use linear activation function.***

### ****3. Huber Loss****

**In statistics, the Huber loss is a loss function used in robust regression, that is less sensitive to outliers in data than the squared error loss.**

****

* **n – the number of data points.**
* **y – the actual value of the data point. Also known as true value.**
* **ŷ – the predicted value of the data point. This value is returned by the model.**
* **δ – defines the point where the Huber loss function transitions from a quadratic to linear.**

#### **Advantage**

* **Robust to outlier**
* **It lies between MAE and MSE.**

#### **Disadvantage**

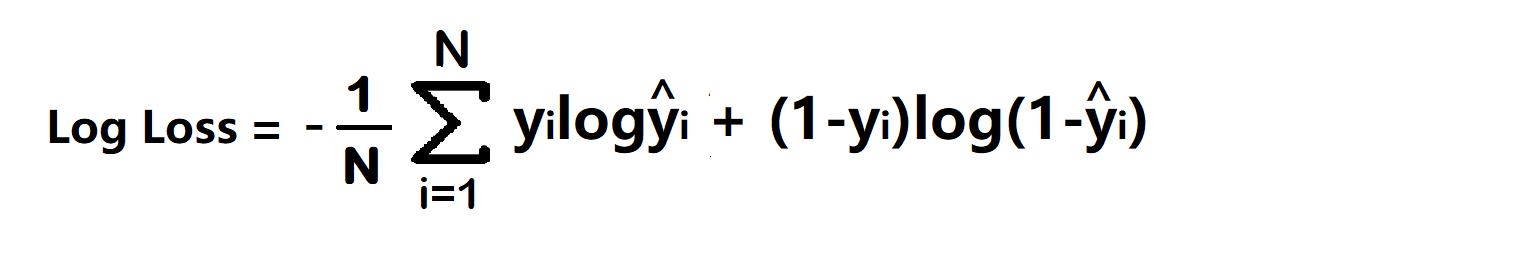
* **Its main disadvantage is the associated complexity. In order to maximize model accuracy, the hyperparameter δ will also need to be optimized which increases the training requirements.**

## **B. Classification Loss**

### ****1. Binary Cross Entropy/log loss****

**It is used in binary classification problems like two classes. example a person has covid or not or my article gets popular or not.**

**Binary cross entropy compares each of the predicted probabilities to the actual class output which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value. That means how close or far from the actual value.**

****

* **yi – actual values**
* **yihat – Neural Network prediction**

#### **Advantage –**

* **A cost function is a differential.**

#### **Disadvantage –**

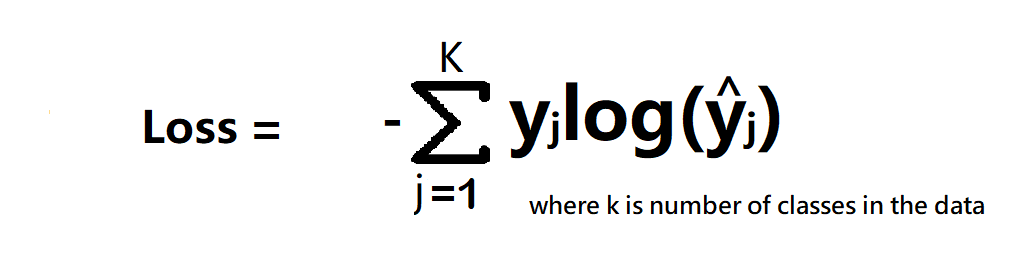
* **Multiple local minima**
* **Not intuitive**

***Note – In classification at last neuron use sigmoid activation function.***

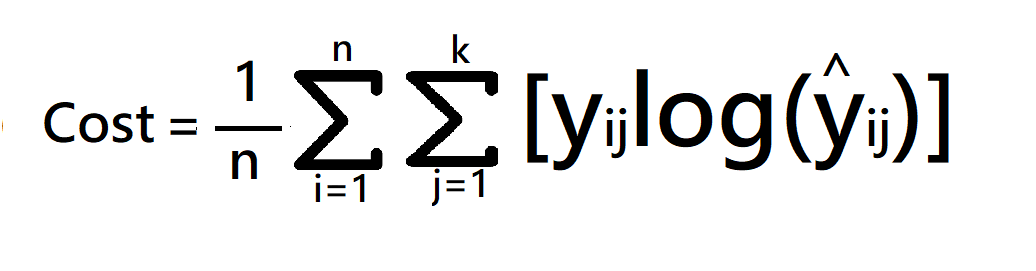
### ****2. Categorical Cross Entropy****

**Categorical Cross entropy is used for Multiclass classification and softmax regression.**

**loss function = -sum up to k(yjlagyjhat) where k is classes**

****

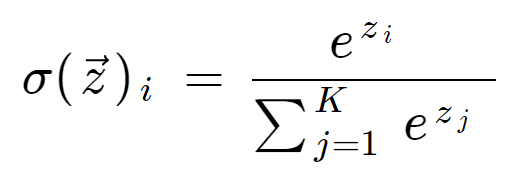
**cost function = -1/n(sum upto n(sum j to k (yijloghijhat))**

****

**where**

* **k is classes,**
* **y = actual value**
* **yhat – Neural Network prediction**

***Note – In multi-class classification at the last neuron use the softmax activation function.***

****

**if problem statement have 3 classes**

**softmax activation – f(z) = ez1/(ez1+ez2+ez3)**

4. Describe L1 and L2 Regularization. How do they differ?

**Answer - L2 regularization takes the square of the weights, so the cost of outliers present in the data increases exponentially. L1 regularization takes the absolute values of the weights, so the cost only increases linearly.**

5. How does Image Registration work?

**Answer - Image registration is an image processing technique used to align multiple scenes into a single integrated image. It helps overcome issues such as image rotation, scale, and skew that are common when overlaying images.**

**--------------------------------------------------------------------------------------------------------------------------------------**

Tableau -

1) What is the purpose of Tableau data blending, and how is it different from data joining?

**Answer - Data Blending Aggregates the data and then combines it. Joins combine the data and then aggregates it. Data Blending can combine data from different sources. Joins can combine data from the same sources only.**

2) How can you create a calculated field in Tableau, and what are some scenarios where calculated fields are useful?

**Answer – Ex : -**

## **Step 1: Create the calculated field**

1. **In a worksheet in Tableau, select Analysis > Create Calculated Field.**
2. **In the Calculation Editor that opens, give the calculated field a name.**

**In this example, the calculated field is called Profit Ratio.**

## **Step 2: Enter a formula**

1. **In the Calculation Editor, enter a formula.**

**This example uses the following formula:**

**SUM([Profit])/SUM([Sales])**

**Formulas use a combination of functions, fields, and operators. To learn more about creating formulas in Tableau, see**[**Formatting Calculations in Tableau**](https://help.tableau.com/current/pro/desktop/en-us/functions_operators.htm)**and**[**Functions in Tableau**](https://help.tableau.com/current/pro/desktop/en-us/functions.htm)**.**

1. **When finished, click OK.**

**The new calculated field is added to the Data pane. If the new field computes quantitative data, it is added to Measures. If it computes qualitative data, it is added to Dimensions.**

**You are now ready to use the calculated field in the view.**

3) How can you optimize performance in Tableau workbooks, especially when dealing with large datasets?

Answer –

#### **1. Use an extract**

**Extracting data and storing it in memory is likely to be optimal unless you need real-time / zero latency data and have the associated infrastructure to support that. For the purposes of this guide, we have assumed that your use case allows for an extract to be used.**

**If you need to have a live database connection, then avoid RAWSQL and custom SQL. Instead, use a sample extract to build the dashboards before changing to the live connection. If you don’t have an analytics database but need the live connection, then most cloud providers have turnkey solutions.**

#### **2. Limit and optimise data**

**The background for this tip is fairly straightforward: The more data Tableau has to go through, the longer it takes. There are however a few ways of minimising the amount of data used in Tableau.**

* **Use extract and data source filters**

**Most datasets contain a certain amount of data that is completely irrelevant for the analysis. This can be data older than what is in scope or sales data for products not in your department. Having this data in the extract will only slow down queries as Tableau has to look through irrelevant data every time something needs to be computed. This is where the extract and data source filters are useful. These filters make it possible to remove data from the final dataset, and in that way, limit the data Tableau has to query.**

* **Aggregate**

**Make sure the data is only at the required level of granularity. If your data is unnecessarily granular and you are ultimately going to be using a summary level, rather than the detail, you will be slowing down performance with the extra data being processed.  
Performance will be improved, and far noticeably for larger datasets, if you aggregate the data to the level required for the analysis or dashboard required. An example of a bad practice of this could be importing transaction level data to create monthly sales reports.  
Clearly, in instances like this, it is important to make sure you have carried out an effective requirements analysis with your stakeholders so that you are clear on the level of granularity that is going to be needed for your project (amongst other things).  
Creating an aggregate can be done either in a database view or by using the built-in aggregation function in Tableau – to do this, visit the extract settings menu.**

* **Hide unused fields**

**The number of columns in a dataset has a larger impact on performance than the number of rows. Therefore, removing unused columns can significantly speed up performance – especially if the dataset is very wide but only a few columns are needed. You might have found that most data sources contain more fields than what is actually required when building a Tableau dashboard. If retained, these unnecessary fields will take up space and be a part of the extract, adding to the amount of data Tableau will have to go through each time a query is performed.  
This is an easy thing to fix without having to resort to building new SQL views or creating custom SQL queries. Simply click “Hide All Unused Fields” in the dropdown menu in the data pane. This keeps the fields but excludes them from being a part of the extract that is used by Tableau – doing this has an impact on both query speed and workbook size.  
Any hidden fields can later be shown by clicking “Show Hidden Fields” in the same dropdown.**

* **Use a context filter**

**A new, temporary dataset is created when context filters are applied. This happens before other filters are applied or calculations are performed (**[**see order of operations**](https://help.tableau.com/current/pro/desktop/en-us/order_of_operations.htm)**). Applying a context filter can therefore have a great impact on performance, as it limits the amount of data that will need to be queried.  
Knowing this, it can be tempting to apply context filters to everything. However, this is not good practice as it takes time to compute and create a new dataset. While context filters do have other uses, they should only be used for performance improvements if they can reduce the amount of data to at least one tenth of the original size, as the added computation time will outweigh the performance increase in other situations.**

* **Use the correct data types - This can be multiple different issues: numbers stored as strings, dates as datetime, or something which is essentially a Boolean value stored as a string. Using incorrect data types can have an impact on performance as some comparisons are quicker than others. Typically, this is as follows:**

**Boolean > Int > Float > String > Date > Datetime**

**A quick solution here is to make sure that numbers are either integers or floats, that dates are dates, and that datetime is only used in cases when the time is also needed. Lastly, some fields can also be replaced with other data types, e.g. a field with yes/no can be made into a Boolean value.**

#### **3) Optimise Filters**

* **Reduce the number of filters**

**Only relevant and necessary filters should be used in a report. Each filter added is more work for Tableau, so irrelevant or redundant filters will only increase the time spent on querying the underlying dataset.**

* **Use the correct filters**

**Include filters are quicker than exclude filters. When Tableau encounters an exclude filter, it will have to load in the entire domain of the filtered field before starting the actual filtering. This will of course increase the time required to filter the data.  
Instead, create a new calculated field returning True/False and filter on this instead. If the filter needs to be user changeable, then use a parameter to set the True/False values in the calculated field.**

* **Include all filter values**

**Avoid using the “include only relevant values” option on filters unless necessary. While it can make the filter list nicer and more dynamic, it requires extra computational power as the list has to be recomputed every time a filter is changed.**

**4) Optimise calculations**

* **Use the correct statements**

**When testing for multiple conditions in a calculated field, using ELSEIF instead of ELSE IF will keep the evaluation in the same statement instead of moving it to a new nested statement. Nested statements will come with a performance cost and are easily avoidable.  
It is also good to make sure that the tests come in the correct order, starting with the most likely outcome. If 90% of the dataset is case A and only 10% case B, then test for case A before case B.**

* **Limit the use of nested calculations**

**Don’t reference the same calculated field in another calculation more than necessary. Nesting calculations like this adds unnecessary complexity because a new query is run every time a calculated field is referenced – I.e., if a calculated field is referenced three times in the same calculation, then running that calculation will require the calculated field to be calculated three times overall. Instead, find another way of achieving the same outcome with fewer references to the field. This can include changing multiple IF statements referencing the same field to a CASE or changing the calculation method used.**

* **Avoid COUNTD**

**Distinct counting (COUNTD) is one of the slowest aggregations in Tableau. It should be avoided if possible, as it has to be computed every time something changes – often at great computational cost. Using distinct counts is often important to the analysis, so completely avoiding any counts isn’t feasible. There are however**[**multiple alternatives to COUNTD**](https://www.tableau.com/en-gb/about/blog/2013/5/tips-tricks-alternatives-count-distinct-23253-0)**that could be worth using instead.**

#### **5. Optimise visualisations**

**Try reducing the number of marks in your visualisations. While it can be tempting to have highly detailed visualisations containing multiple different graphs of a highly granular level, it is not advisable. Adding more marks means that Tableau will need more processing power to render the visualisation. By limiting how much is shown by filtering, aggregating, or splitting the visualisation into multiple worksheets, the rendering time can be improved.**

**One final thing that can be done to improve rendering times is to limit the number of worksheets in a dashboard. Each worksheet will have to run its own queries, so the more worksheets in a single dashboard, the more queries Tableau will have to render. Instead, spread the worksheets out across multiple dashboards if possible. This will both speed up the dashboard and mitigate the information overload that comes with dashboards being crammed full of highly detailed graphs.**

4) How can you implement advanced interactivity in Tableau dashboards?

**Answer - Build an Interactive Dashboard Using Tableau**

1. **Create Graphs.**
2. **Create the First Draft of the Dashboard.**
3. **Make Your Dashboard Interactive.**
4. **Stories, New Data Sources, and More**

5) How can you integrate Tableau with other systems or applications?

Answer – Tableau can be integrated with these below applications ->

* **Amazon Web Services.**
* **Snowflake.**
* **Data Bricks.**
* **Mulesoft.**
* **Oracle.**
* **SAP.**
* **Cloudera.**
* **Google GCP.**

**Although Tableau supports direct integration with 80+ data sources, there are still a wide variety of data sources that are not directly supported by it. To help users establish connections with unsupported Tableau Data Sources seamlessly, the following functionalities were introduced:**

* [**Web Data Connector**](https://hevodata.com/learn/understanding-tableau-data-integration/#wdc)
* [**ODBC Connector**](https://hevodata.com/learn/understanding-tableau-data-integration/#odbc)
* [**Extract API**](https://hevodata.com/learn/understanding-tableau-data-integration/#extract_api)
* [**Hyper API**](https://hevodata.com/learn/understanding-tableau-data-integration/#hyper_api)

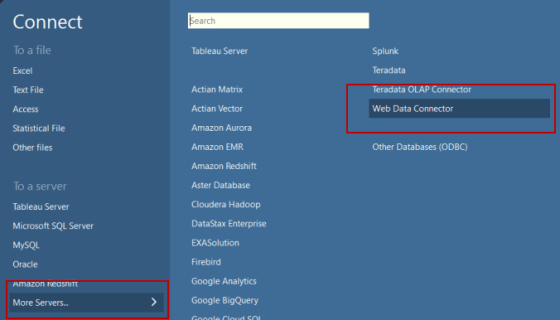
### 1) Web Data Connector

**Tableau Web Data Connector (WDC) can be defined as a set of APIs that allow developers to establish a connection between Tableau and any data on the web that is accessible over HTTP. Web Data Connector (WDC) is an HTML file that includes some JavaScript code.**

**Tableau houses an intuitive functionality that allows users to create their own Web Data Connector (WDC) or use an existing one. Users can create their own Web Data Connector (WDC) that is capable of reading data from any website publishing data in XML, JSON, or HTML format. The Web Data Connector (WDC) must be hosted on a local Web Server on your computer, on a Web Server in the user’s domain, or on any third-party Web Server.**

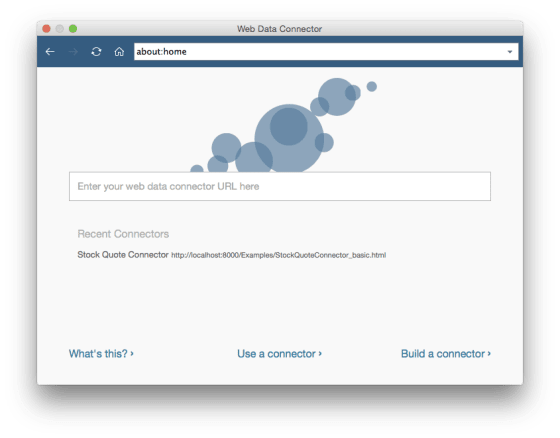
**Users can implement the following steps in order to leverage Web Data Connectors to set up unsupported Tableau Data Integrations:**

* **Step 1: Open Tableau, click on More Servers in the Connect pane, and then select Web Data Connector.**

**[](https://res.cloudinary.com/hevo/image/upload/f_auto,q_auto/v1686065763/hevo-learn-1/select_wdc.png?_i=AA)**

**Image Source:**[**https://tableau.github.io/webdataconnector/docs/wdc\_use\_in\_tableau**](https://tableau.github.io/webdataconnector/docs/wdc_use_in_tableau)

* **Step 2: Enter the URL of the WDC and press Enter.**

**[](https://res.cloudinary.com/hevo/image/upload/f_auto,q_auto/v1686065755/hevo-learn-1/url_wdc.png?_i=AA)**

**Image Source:**[**https://www.guru99.com/tableau-data-connections.html**](https://www.guru99.com/tableau-data-connections.html)

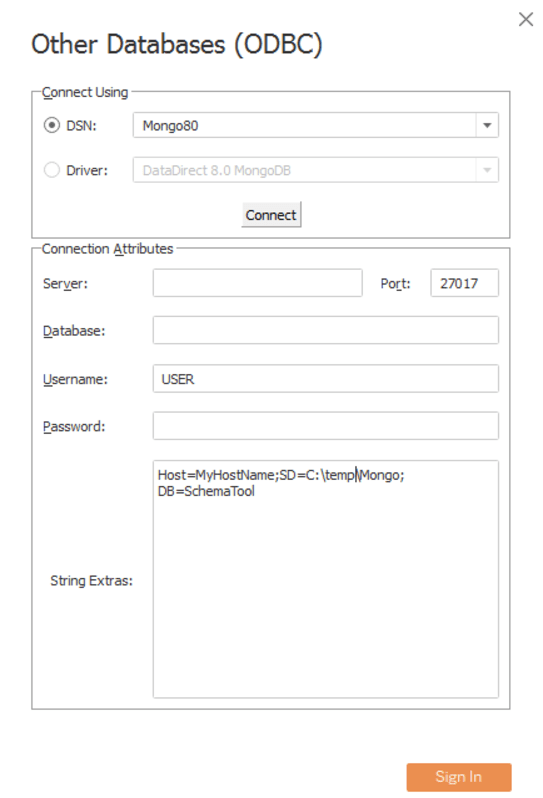
* **Step 3: If the WDC leads to a webpage, enter any required information and select Submit.**
* **Step 4: Once you’ve provided the required information, WDC will retrieve all the information as an Extract and import it into Tableau Data Connections. This data can then be analyzed as per the usual process.**

### 2) ODBC Connector

**ODBC (Open Database Connectivity) is a popular industry standard that allows a wide variety of software applications to access data stored in a database. The basis of ODBC is that software applications can make use of standard SQL queries to form a connection with a database and request data from it. The ODBC driver accepts requests in the standard syntax and converts these requests into a format that can be interpreted by the target database.**

**Hence, the ODBC driver can be seen as a translation layer capable of converting a general-purpose request into a database-specific format. Tableau houses functionality that allows users to connect with all ODBC-compliant data sources using its robust in-built ODBC Connector. This ODBC Connector can be leveraged to seamlessly integrate unsupported Tableau Data Sources by implementing the following steps:**

* **Step 1: Open Tableau, click on More Servers in the Connect pane and then select Other Databases (ODBC).**
* **Step 2: Enter the required information required to identify and establish a connection with the necessary ODBC-compliant database and click on Sign In.**

**[](https://res.cloudinary.com/hevo/image/upload/f_auto,q_auto/v1686065748/hevo-learn-1/connect_odbc.png?_i=AA)**

**Image Source:**[**https://www.progress.com/tutorials/odbc/how-to-configure-the-datadirect-odbc-mongodb-driver-with-tableau-desktop**](https://www.progress.com/tutorials/odbc/how-to-configure-the-datadirect-odbc-mongodb-driver-with-tableau-desktop)

### 3) Extract API

**The Tableau Extract API houses functionalities that allow users to perform the following operations:**

* **Create and add data to Extract files that can improve performance exponentially and provide uninterrupted offline access to all your Tableau Data Sources.**
* **Write programs capable of integrating with unsupported Tableau Data Sources, and writing data into Extract files for later use.**
* **Write programs capable of creating Extract files that contain data from multiple tables.**

**The platforms supported by Tableau Extract API are as follows:**

* **Windows 7 or later.**
* **Windows Server 2008 R2 or later.**
* **Mac OS X (10.9 and later).**
* **CentOS 7 and later.**
* **Fedora 18 and later.**
* **Ubuntu 12.04 and later.**

**The programming languages supported by Tableau Extract API are as follows:**

* **C/C++**
* **Python 2.x and 3.x**
* **Java**

**Information on how to set up Extract API depending on the programming language and platform being used can be found**[**here**](https://help.tableau.com/current/api/extract_api/en-us/Extract/extract_api.htm)**.**

### 4) Hyper API

**Tableau Hyper can be seen as a high-performance In-memory Data Engine technology that allows users to analyze large and complex datasets faster, by seamlessly evaluating all analytical queries in the transactional database.**

**Users can leverage the Tableau Hyper API to perform the following operations:**

* **Open existing Extract files.**
* **Perform various CRUD (Create, Read, Update, Delete) operations on Extract files.**
* **Create Extracts for various unsupported Tableau Data Sources.**
* **Automate various custom Extract, Transform and Load (**[**ETL**](https://hevodata.com/learn/4-best-tableau-etl-tools-for-2021/)**) operations.**
* **Use SQL queries to interact with data in Tableau Hyper Extracts.**

**The platforms supported by Tableau Hyper API are as follows:**

* **Microsoft Windows Server 2019, 2016, 2012 R2, 2012, 2008 R2**
* **Microsoft Windows 7 or newer (64-bit)**
* **Amazon Linux 2, Red Hat Enterprise Linux (RHEL) 7.3+, CentOS 7.3+, Oracle Linux 7.3+, Ubuntu 16.04 LTS, and 18.04 LTS**
* **macOS 10.13 or newer**

**The programming languages supported by Tableau Hyper API are as follows:**

* **Python (3.6 or newer)**
* **Java (Java 8 or newer)**
* **C++ (C++11 or newer)**
* **C#/.NET (.NET Standard 2.0)**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**07/08/2023 -** Deep Learning -

1. What are some popular object detection architectures apart from YOLO

**Answer –**

**1) Histogram of Oriented Gradients (HOG)**

**2) Region Based Convolution Neural Network (R-CNN)**

**3) Single Shot Detector (SSD)**

**4) RetinaNet**

**5) Faster R-CNN**

2. Difference between flask, streamlit, fastapi.

**Answer - Flask's documentation is not as comprehensive as FastAPI's but still provides ample resources for learning and using the framework. Streamlit's documentation is focused on data science use cases, with detailed examples and tutorials for creating interactive data visualizations and machine learning models.**

3. What is Matrix pooling?

**Answer - Matrix pooling (or two-dimensional array-based group testing algorithm) is performed through placing specimens in an array and then screening groups which are formed by rows and by columns.**

* **Pooling layers are used to reduce the dimensions of the feature maps. Thus, it reduces the number of parameters to learn and the amount of computation performed in the network.**
* **The pooling layer summarises the features present in a region of the feature map generated by a convolution layer. So, further operations are performed on summarised features instead of precisely positioned features generated by the convolution layer. This makes the model more robust to variations in the position of the features in the input image.**

4. Explain Confusion matrix in object classification?

**Answer - The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known.**

5. What is the use of Anchor boxes in yolo?

**Answer - “Anchor boxes are a set of predefined bounding boxes of a certain height and width. These boxes are defined to capture the scale and aspect ratio of specific object classes you want to detect and are typically chosen based on object sizes in your training datasets.”**

**--------------------------------------------------------------------------------------------------------------------------------------**

Tableau -

1) Explain the concept of data densification in Tableau?

**Answer - Data densification is basically a technique to densify the dataset or in simpler terms, increase the number of observations in the dataset. Data densification occurs when Tableau creates additional marks in the view to “compensate” for missing value; it also assigns to those marks the value it thinks it should have, based on surrounding values. Notice that these marks are not added to your data, but only on your view.**

2) What is the difference between joining and blending in Tableau?

**Answer - The main difference between the joining & blending is when the aggregation is performed. A join combines the data and then aggregates. A blend aggregates and then combines the data.**

3) What does a Gantt chart mean?

**Answer - Applies to: Tableau Desktop. Use Gantt charts to show the duration of events or activities. In a Gantt chart, each separate mark (usually a bar) shows a duration. For example, you might use a Gantt chart to display average delivery time for a range of products.**

4) What are the different kinds of Tableau filters available?

**Answer - Types of Filters in Tableau Desktop**

* **Context Filter**
* **Extract Filter**
* **Data Source Filter**
* **Dimension Filter**
* **Measure Filter**
* **Table Calculation Filter**

5) What does the Blue Mark and Blue Pill indicate in Tableau?

**Answer –**

**Blue Pill: When we highlight fields on the data pane, discrete fields will appear as a blue pill while continuous fields appear as a green pill.**

**Blue Mark: A blue check mark indicates that the data source is the primary data source in the workbook.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**09/08/2023 –** Questions

1.What is transfer learning, and how can it be beneficial in image detection tasks?

**Answer - In transfer learning, a machine exploits the knowledge gained from a previous task to improve generalization about another. For example, in training a classifier to predict whether an image contains food, you could use the knowledge it gained during training to recognize drinks.**

2.what is difference between cost function and loss function

**Answer - The loss function calculates the error per observation, whilst the cost function calculates the error over the whole dataset. Also, cost function is aggregation of loss function.**

3.Can you describe the vanishing gradient problem and how it is addressed in deep learning?

**Answer - Vanishing gradient problem is a phenomenon that occurs during the training of deep neural networks, where the gradients that are used to update the network become extremely small or "vanish" as they are backpropagated from the output layers to the earlier layers. Also, rate of change stops abruptly since value is < 1.**

4.Difference between object detection and instance segmentation?

**Answer - Object detection finds bounding boxes around objects and classifies them. Instance segmentation adds, for every detected object, a pixel mask that gives the shape of the object.**

5.Explain regularization techniques such as dropout and L2 regularization.

**Answer – Dropout - The dropout process is random, and it is performed separately for each training example. As a consequence, each training example might be trained on a different network. In the test phase they use all activations, but reduce them by a factor 'p' (to account for missing activations during training) - all neurons output are reduced by 0.5. Neuron output becomes 0 hence all connections are gone.**

**L2 regularization - L2 regularization, also known as Ridge regularization, is a machine learning technique that avoids overfitting by introducing a penalty term into the model's loss function based on the squares of the model's parameters. The goal of L2 regularization is to keep the model's parameter sizes short and prevent oversizing.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**

**11/08/2023 –** Questions

1) What is Flask and how does it differ from other web frameworks?

**Answer - Flask is a web framework, it is a Python module that lets you develop web applications easily. Flask is a lightweight Python web framework that provides useful tools and features for creating web applications in the Python Language. It gives developers flexibility and is an accessible framework for new developers because you can build a web application quickly using only a single Python file.**

**Flask: Flask is also easy to use, but requires a bit more setup than Streamlit. Flask provides a minimal core that can be extended with plug-ins, which can make it a bit more complex to get started with.**

**Django: Django can be more challenging to set up and requires a more complex development environment.**

**Streamlit is a newer framework that is designed specifically for data-driven web applications.**

2) What is Streamlit and what is it used for in the context of data science?

**Answer - Streamlit is a Python-based web application framework for visualizing data and analyzing results in a more efficient and flexible way. It is an open source library that assists data scientists and academics to develop Machine Learning (ML) visualization dashboards in a short period of time.**

**Streamlit is a free, open-source, all-python framework that enables data scientists to quickly build interactive dashboards and machine learning web apps with no front-end web development experience required.**

3) How can you create a simple interactive data visualization using Streamlit?

**Answer - Streamlit provide a bar\_chart function to plot bar charts for data visualization. Since the data frame contains 5 rows, there are bars corresponding to each row in the output. These bars contain the distribution of each column's value in that specific row and are denoted by different colours for each column.**

4) How does Streamlit handle real-time data updates and streaming?

**Answer - Streamlit is an awesome library that enables you to quickly create clean and nice-looking dashboards with a bit of code.  
It works great when you want to develop a research service that can be shared among your team and you have already existing Python code which you can reuse.**

**Example: -**

#### **Using st.session\_state and experimental\_rerun**

**import random**

**import time**

**import pandas as pd**

**import streamlit as st**

**def has\_changed() -> bool:**

**return True**

**def load\_most\_recent() -> pd.DataFrame:**

**return pd.DataFrame({"my\_col": list(range(random.randint(1, 10)))})**

**def wait\_for\_update():**

**# Simplest would be to just periodically check if something has changed**

**while True:**

**time.sleep(1)**

**if has\_changed():**

**return**

**df = load\_most\_recent()**

**st.write(df)**

**wait\_for\_update()**

**# This will trigger the streamlit server to run the script again from start**

**st.experimental\_rerun()**

**The main advantage of this approach is its simplicity and the fact that you avoid doing complex operations inside streamlit.**

#### **Using placeholder and a loop**

**import time**

**import pandas as pd**

**import streamlit as st**

**df = pd.DataFrame({"my\_col": [1, 2]})**

**# We create the placeholder once**

**placeholder = st.empty()**

**while True:**

**# It is important to exit the context of the placeholder in each step of the loop**

**with placeholder.container():**

**df = pd.concat(**

**[**

**df,**

**pd.DataFrame(**

**{"my\_col": [df["my\_col"].max() + 1]}, index=[df["my\_col"].max() + 1]**

**),**

**]**

**)**

**# placeholder object should have the same methods for displaying data as st**

**placeholder.dataframe(df)**

**time.sleep(2)**

5) How can you manage environment configuration in Flask applications?

**Answer - In order to set the environment and debug mode reliably, Flask uses environment variables. The environment is used to indicate to Flask, extensions, and other programs, like Sentry, what context Flask is running in. It is controlled with the FLASK\_ENV environment variable and defaults to production .**

## **Environment and Debug Features**

**The**[**ENV**](https://flask.palletsprojects.com/en/1.1.x/config/#ENV)**and**[**DEBUG**](https://flask.palletsprojects.com/en/1.1.x/config/#DEBUG)**config values are special because they may behave inconsistently if changed after the app has begun setting up. In order to set the environment and debug mode reliably, Flask uses environment variables.**

**The environment is used to indicate to Flask, extensions, and other programs, like Sentry, what context Flask is running in. It is controlled with the FLASK\_ENV environment variable and defaults to production.**

**Setting FLASK\_ENV to development will enable debug mode. flask run will use the interactive debugger and reloader by default in debug mode. To control this separately from the environment, use the FLASK\_DEBUG flag.**

**To switch Flask to the development environment and enable debug mode, set FLASK\_ENV:**

**$ export FLASK\_ENV=development**

**$ flask run**

**ENV**

**What environment the app is running in. Flask and extensions may enable behaviors based on the environment, such as enabling debug mode. The**[**env**](https://flask.palletsprojects.com/en/1.1.x/api/#flask.Flask.env)**attribute maps to this config key. This is set by the FLASK\_ENV environment variable and may not behave as expected if set in code.**

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**16/08/2023 – Questions**

1) What is video based counting deployment?

**Answer - A method for detecting, counting, and tracking vehicles in roundabout videos is proposed and the output is compared with both manual counting results and benchmark tracking results where the entry/exit matrix is generated using only YOLOv4 and an object tracker.**

2) What is object detection in video counting systems?

**Answer - Video Object Detection (VOD) mimics the human visual cortex. It allows machines to analyze video frame by frame and identify the objects present within them. Thus, object detection in video works similarly to AI image recognition. Such a tool aims to locate and identify objects seen on input moving images.**

3) What are some challenges in deploying video counting systems?

**Answer -**

* **Object Localisation.**
* **Multiple Aspect Ratios and Spatial Sizes.**
* **Viewpoint Variation.**
* **Occlusion.**
* **Deformation.**
* **Intra-Class Variation.**
* **Limited Data.**
* **Cluttered or textured background.**

4) How can the accuracy of a video counting system be improved?

**Answer - the system performance and accuracy can still be improved through downsizing the frames, retraining the models, increasing epoch size and exploring other models.**

5) What are the key components of a video - based counting system?

**Answer – The key components of a video-based counting systems are detector, tracker and counter. The detector identifies objects in a given frame of video and returns a list of bounding boxes, the tracker tracks the objects/images in the video & the counter, counts the number of detections done.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------18/08/2023 – Questions**

1) What is Activation Function?

**Answer - The activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias to it. The purpose of the activation function is to introduce non-linearity into the output of a neuron.**

2) What is Overfitting in deep learning and what techniques can be used to prevent it?

**Answer - Overfitting occurs when the model has a high variance, i.e., the model performs well on the training data but does not perform accurately in the evaluation set. The model memorizes the data patterns in the training dataset but fails to generalize to unseen examples.**

**You can prevent overfitting by diversifying and scaling your training data set or using some other data science strategies, like those given below.**

**Early stopping -   
Early stopping pauses the training phase before the machine learning model learns the noise in the data. However, getting the timing right is important; else the model will still not give accurate results.**

**Pruning -   
You might identify several features or parameters that impact the final prediction when you build a model. Feature selection—or pruning—identifies the most important features within the training set and eliminates irrelevant ones. For example, to predict if an image is an animal or human, you can look at various input parameters like face shape, ear position, body structure, etc. You may prioritize face shape and ignore the shape of the eyes.**

**Regularization -   
Regularization is a collection of training/optimization techniques that seek to reduce overfitting. These methods try to eliminate those factors that do not impact the prediction outcomes by grading features based on importance. For example, mathematical calculations apply a penalty value to features with minimal impact. Consider a statistical model attempting to predict the housing prices of a city in 20 years. Regularization would give a lower penalty value to features like population growth and average annual income but a higher penalty value to the average annual temperature of the city.**

**Ensembling -   
Ensembling combines predictions from several separate machine learning algorithms. Some models are called weak learners because their results are often inaccurate. Ensemble methods combine all the weak learners to get more accurate results. They use multiple models to analyze sample data and pick the most accurate outcomes. The two main ensemble methods are bagging and boosting. Boosting trains different machine learning models one after another to get the final result, while bagging trains them in parallel.**

**Data augmentation -   
Data augmentation is a machine learning technique that changes the sample data slightly every time the model processes it. You can do this by changing the input data in small ways. When done in moderation, data augmentation makes the training sets appear unique to the model and prevents the model from learning their characteristics. For example, applying transformations such as translation, flipping, and rotation to input images.**

3) What are some challenges or limitations of deep learning in real-world applications?

### Answer - Disadvantages of Deep Learning:

**While deep learning has many advantages, there are also some disadvantages to consider:**

1. **High computational cost: Training deep learning models requires significant computational resources, including powerful GPUs and large amounts of memory. This can be costly and time-consuming.**
2. **Overfitting: Overfitting occurs when a model is trained too well on the training data and performs poorly on new, unseen data. This is a common problem in deep learning, especially with large neural networks, and can be caused by a lack of data, a complex model, or a lack of regularization.**
3. **Lack of interpretability: Deep learning models, especially those with many layers, can be complex and difficult to interpret. This can make it difficult to understand how the model is making predictions and to identify any errors or biases in the model.**
4. **Dependence on data quality: Deep learning algorithms rely on the quality of the data they are trained on. If the data is noisy, incomplete, or biased, the model’s performance will be negatively affected.**
5. **Data privacy and security concerns: As deep learning models often rely on large amounts of data, there are concerns about data privacy and security. Misuse of data by malicious actors can lead to serious consequences like identity theft, financial loss and invasion of privacy.**
6. **Lack of domain expertise: Deep learning requires a good understanding of the domain and the problem you are trying to solve. If the domain expertise is lacking, it can be difficult to formulate the problem and select the appropriate algorithm.**
7. **Unforeseen consequences: Deep learning models can lead to unintended consequences, for example, a biased model can discriminate against certain groups of people, leading to ethical concerns.**
8. **Limited to the data its trained on: Deep learning models can only make predictions based on the data it has been trained on. They may not be able to generalize to new situations or contexts that were not represented in the training data.**
9. **Black box models: some deep learning models are considered as “black-box” models, as it is difficult to understand how the model is making predictions and identifying the factors that influence the predictions.**

4) How can you handle imbalanced dataset in deep learning projects?

**Answer – Ways to handle Imbalanced dataset in deep learning ->**

**1. Choose Proper Evaluation Metric**

**The accuracy of a classifier is the total number of correct predictions by the classifier divided by the total number of predictions. This may be good enough for a well-balanced class but not ideal for the imbalanced class problem. The other metrics such as precision is the measure of how accurate the classifier’s prediction of a specific class and recall is the measure of the classifier’s ability to identify a class.**

**For an imbalanced class dataset F1 score is a more appropriate metric. It is the harmonic mean of precision and recall and the expression is –**

****

**So, if the classifier predicts the minority class but the prediction is erroneous and false-positive increases, the precision metric will be low and so as F1 score. Also, if the classifier identifies the minority class poorly, i.e. more of this class wrongfully predicted as the majority class then false negatives will increase, so recall and F1 score will low. F1 score only increases if both the number and quality of prediction improves.**

**F1 score keeps the balance between precision and recall and improves the score only if the classifier identifies more of a certain class correctly.**

**2. Resampling (Oversampling and Undersampling)**

**This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling. After sampling the data we can get a balanced dataset for both majority and minority classes. So, when both classes have a similar number of records present in the dataset, we can assume that the classifier will give equal importance to both classes.**

**3. SMOTE**

**Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don’t add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use *k* nearest neighbor to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.**

### 4. BalancedBaggingClassifier

**When we try to use a usual classifier to classify an imbalanced dataset, the model favors the majority class due to its larger volume presence. A**[**BalancedBaggingClassifier**](https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedBaggingClassifier.html) **is the same as a sklearn classifier but with additional balancing. It includes an additional step to balance the training set at the time of fit for a given sampler. This classifier takes two special parameters “sampling\_strategy” and “replacement”. The sampling\_strategy decides the type of resampling required (e.g. ‘majority’ – resample only the majority class, ‘all’ – resample all classes, etc) and replacement decides whether it is going to be a sample with replacement or not.**

**5. Threshold moving**

**In the case of our classifiers, many times classifiers actually predict the probability of class membership. We assign those prediction’s probabilities to a certain class based on a threshold which is usually 0.5, i.e. if the probabilities < 0.5 it belongs to a certain class, and if not it belongs to the other class.**

**For imbalanced class problems, this default threshold may not work properly. We need to change the threshold to the optimum value so that it can efficiently separate two classes. We can use ROC Curves and Precision-Recall Curves to find the optimal threshold for the classifier. We can also use a grid search method or search within a set of values to identify the optimal value.**

5) Explain the concept of layers in a neural network and their functions.

**Answer - The Neural Network is constructed from 3 type of layers: Input layer — initial data for the neural network. Hidden layers — intermediate layer between input and output layer and place where all the computation is done. Output layer — produce the result for given inputs.**

1. **Input Layer**

**It is a layer where all the inputs are fed to the Neural Network or model.**

1. **Hidden Layers**

**Hidden Layers are the layers which are in between input and output layers which are used for processing inputs. A Neural Network can have more than one Hidden layer.**

1. **Output Layer**

**It is the Last layer of the Neural Network. It takes the information from Hidden layers after data processing. Thus, the processed data is made available at output layer.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------21/08/2023 – Questions –**

1) What is loss function in deep learning?

**Answer - The loss function calculates the error per observation in deep learning.**

2) What is backpropagation?

**Answer - Backpropagation is just a way of propagating the total loss back into the neural network to know how much of the loss every node is responsible for, and subsequently updating the weights in a way that minimizes the loss by giving the nodes with higher error rates lower weights, and vice versa.**

3) What is pooling in CNNs?

Answer - **In convolutional neural networks (CNNs), the pooling layer is a common type of layer that is typically added after convolutional layers. The pooling layer is used to reduce the spatial dimensions (i.e., the width and height) of the feature maps, while preserving the depth (i.e., the number of channels).**

4) What is dropout?

**Answer – Dropout is one of the regularization techniques to avoid Overfitting / Variance.**

**Dropout - The dropout process is random, and it is performed separately for each training example. As a consequence, each training example might be trained on a different network. In the test phase they use all activations, but reduce them by a factor 'p' (to account for missing activations during training) - all neurons output are reduced by 0.5. Neuron output becomes 0 hence all connections are gone.**

5) What is a hyperparameter in deep learning?

**Answer - Hyperparameters are parameters whose values are set before starting the model training process. Deep learning models, including convolutional neural network (CNN) and recurrent neural network (RNN) models can have anywhere from a few hyperparameters to a few hundred hyperparameters.**

**----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------**