**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 pring 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**

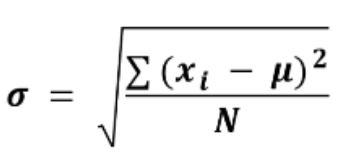
2) What is the relationship between Mean, Median and Mode for positive and negative skewed?

**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**

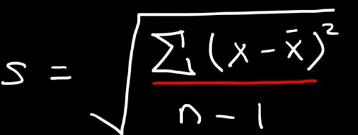
****

**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?

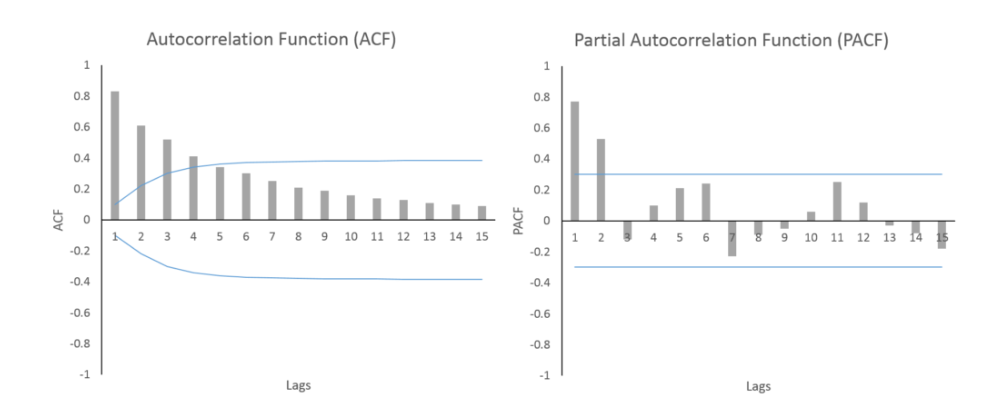
**Answer -> Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of the relationship, the correlation coefficient's value varies between +1 and -1.**

**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?

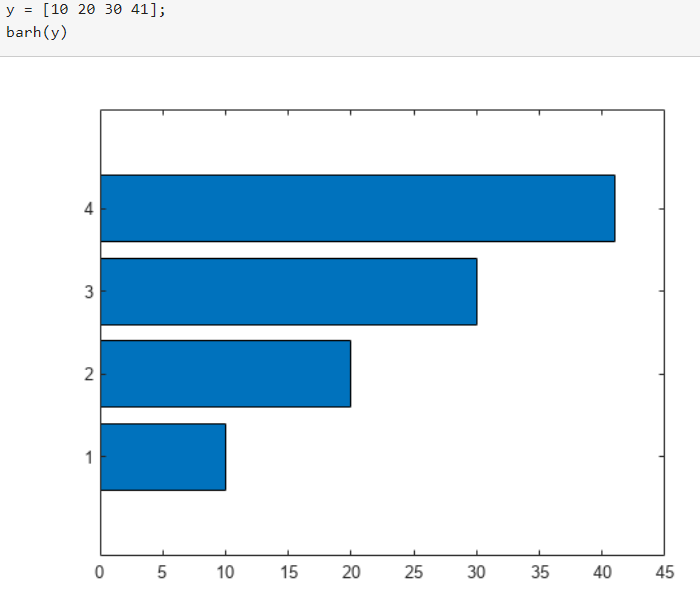
**Answer -> X-axis: The X-axis are intervals that show the scale of values which the measurements fall under.**

**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?

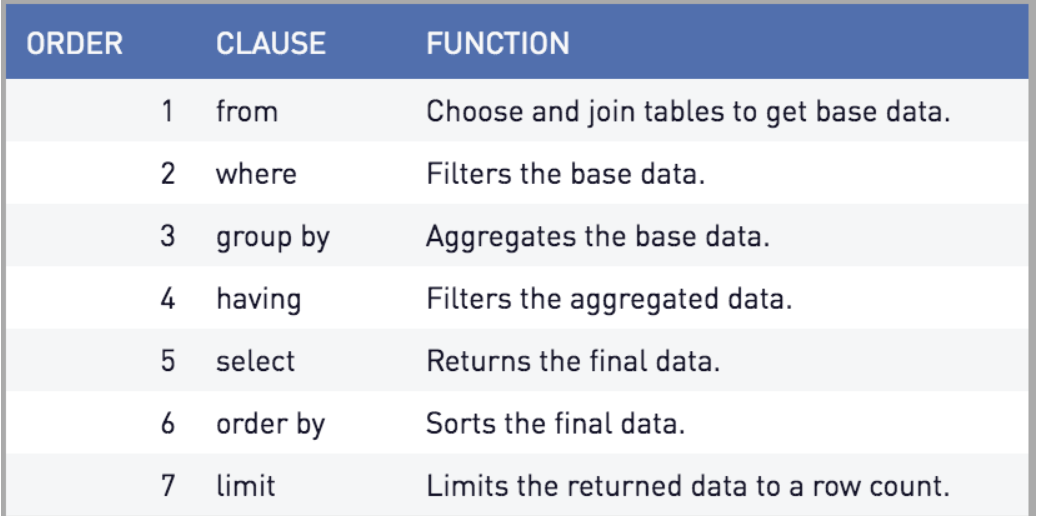
**Answer - Pychecker and Pylint are the static analysis tools that help to find bugs in python**

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

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 does not functionally dependant 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 is 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?

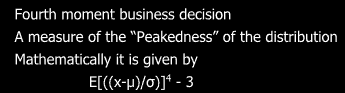
**Answer - The standard deviation is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.**

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?

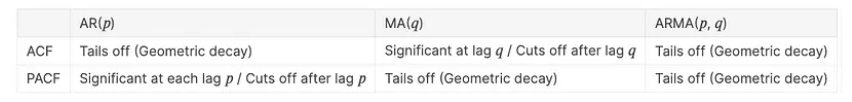
Answer – Formula for Kurtosis is as below.



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 amount 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 - What is the purpose of global keywords in python? 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.**

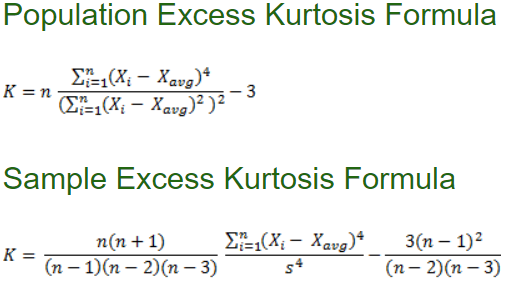
3) What is the value of vert parameter in boxplot function, in order to visualize it horizontally?

**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**

**# fehrenheit 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 PowerBI?

**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 modeling 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**

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