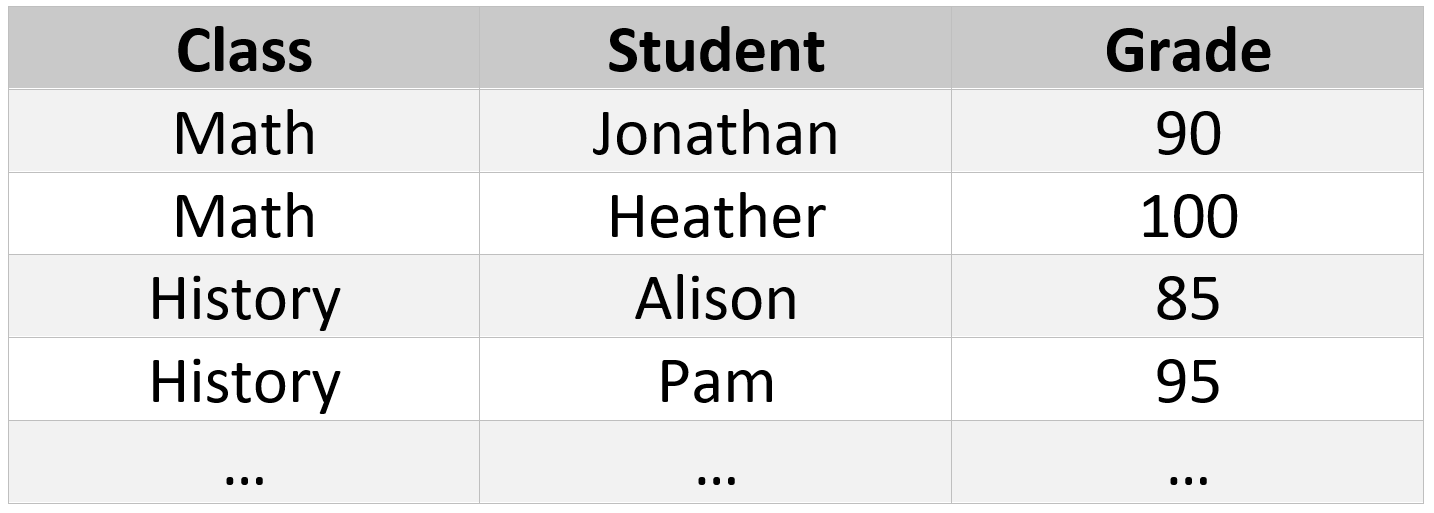
1. *Write a SQL query to create a table that shows, for each class, the value of the highest grade in the class (not the Student).*



**Answer** :

SQL Query to create the above table in Oracle Database :

Create table class\_data

( class varchar2(10),

student varchar2(10),

grade number ) ;

**SQL query to create a table as per the requirement :**

Create table class\_high\_grade as

select class, max(grade) grade

from class\_data

group by class;

++++++++++++++++++++++++++++++

1. *In pseudo-code or whatever language you would like: write a program that prints the numbers from 1 to 100. But for multiples of three print “Fizz” instead of the number and for the multiples of five print “Buzz”. For numbers which are multiples of both three and five print “FizzBuzz”.*

**Answer** :

Since I am familiar with Python, I am using Python to write the code for the above question :

**Code in Python :**

print('\n'.join(['FizzBuzz' if x % 15 == 0  
else 'Fizz' if x % 3 == 0  
else 'Buzz' if x % 5 == 0  
else str(x)  
for x in range(1,101)]))

Screen shot from my jupyter notebook for practical explanation :

A screenshot of a social media post

Description automatically generated

++++++++++++++++++++++++++++++

1. *A company selling a competitor to Microsoft Office is testing their marketing by sending out two different sets of emails. One set contains business related content, and one contains consumer related content. The bottom two graphs have the same data as the top two, only bucketed by the type of customer (based on how much he has spent with the company the year before the emails were sent). Can you see any mistake performed on the campaign? Can the campaign be optimized?*

A screenshot of a cell phone

Description automatically generated

**Answer** :

Inference from the Top two Graphs :

Though the Buisness Marketing Emails are delivered more than Consumer Marketing Emails, it’s the Consumer Marketing Emails that has more Click-through rate.

Inference from the Bottom two Graphs :

If the mails are viewed based on Customer type ( Low, Normal, High ) its noticed that Business mails have more Click through rate. Also it can be noticed that higher percentage of Business mails have been sent to Low-End Customers who have lower click through rates.

How the Campaign be Optimized ?

As per the above inference from the Graphs, we can safely assume that High End Customers generally have higher percentage of Click through rates ( for both Business and Consumer mails ) compared to both Normal and Low end customers.

Hence its more effective to deliver Business/Consumer marketing mails to High end Customers. Relatively we can increase the mails to Normal Customers as well as they also have better Click through rates when compared with low end customers.

++++++++++++++++++++++++++++++

**Statistics Questions :**

* *What is the Central Limit Theorem and why is it important?*

**Answer** :

Although there is no one complete central limit definition, the one commonly found in most statistics text books is

“ The Central Limit Theorem (CLT) is a statistical theory states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.”

In simple terms , with large samples sizes the sample means are normally distributed.

Importance :

By Central Limit Theorem, we know

"The distribution of means across repeated samples will be normal with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of n.

Since we know exactly what the distribution of means will look like for a given population, we can take the mean from a single sample and compare it to the sampling distribution to assess the likelihood that our sample comes from the same population. In other words, we can test the hypothesis that our sample represents a population distinct from the known population.

+++++++++++++++++++

* *What is sampling? How many sampling methods do you know?*

**Answer** :

To understand Sampling, we need to understand the term “Population”

“Population” is the collection of the elements which has some or the other characteristicsin common. Number of elements in the population is the size of the population.

Sample is the subset of the population. The process of selecting a sample is known as “Sampling”. Number of elements in the sample is the sample size.

There are lot of sampling techniques which can be largely grouped into two categories :

Probability Sampling

Non- Probability Sampling

The difference lies between the above two is whether the sample selection is based on randomization or not. With randomization, every element gets equal chance to be picked up and to be part of sample for study.

I have predominantly worked on Probability Sampling in my coding like :

Simple Random Sampling : Every element has an equal change of getting selected. It is used when we don’t have any kind of prior information about the target population.

For Eg : Random selection of 10 employees from a company 100. Each employee has the probability of being selected as 1/100

Startified Sampling : ( I used it often during Train-Test split of the dataset )

This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous and heterogeneous among the other subgroups formed

There is also another technique known as Cluster sampling. I do have theoretical knowledge but have not worked on it practically.

With regards to Non-Probability Sampling, we have Convenience Sampling, Purposive Sampling. ( Again just the theoretical knowledge only )

Convenience Sampling is the selection based on availability.

Purposive Sampling is the selection based on the intention (or) purpose of study.\

NOTE : I have very often used SMOTE ( Synthetic Minority Over Sampling Technique -

Finding K-nearest neighbours of minority variable and build on it ) for Un-balanced dataset

+++++++++++++++++++

* *How do you handle outliers or data points that skew data?*

**Answer** :

Outlier is a data point that is distant from other similar points.

In statistics, an outlier is an observation point that is distant from other observations

General rule :

Low Outliers are below : Q1 - 1.5 \* IQR

High Outliers are above : Q3 + 1.5 \* IQR

Types of Outliers :

1. Univariate Outlier : Data point that consists of an extreme value on one variable
2. Multivariate Outlier : Multivariate outlier is a combination of unusual scores on at least two variables

Ways to deal outliers :

1. Deleting observations: Delete outlier values if its due to data entry error, data processing error.
2. Transforming and binning values : Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variable
3. Imputing : We can also impute outliers. Use Mean, Median, Mode imputation methods. Before imputing values, we need to analyze the data
4. Treat Outliers separately : If there are significant no of outliers, we should treat them separately. One of the approach is to treat both groups as two different groups and build individual model for both groups and then combine the output.

+++++++++++++++++++++++

* *What is bias, how it can affect your model?*

**Answer** :

Theorical definition of bias : Bias in Machine learning is defined as the phenomena of observing results that are systematically prejudiced due to faulty assumptions.

At conceptual level, let’s try to understand what is bias and variance and its impact on the model with an example :

We want to predict the average marks of all the students of Grade 6 in a school – “Google”. Let’s pretend to be a novice with no knowledge about the school – “Google”. Now we make a simplistic assumption that average mark of all the students in school “Google” to be similar to the average mark of all the students in Grade 6. With this assumption, our prediction will not be accurate and will give more generic results. This excess generalization is nothing but High bias known as underfitting.

Generally the error will be more in both training and test data in case of underfitting.

NOTE : Building an effective Machine Learning model is all about striking the right balance between Bias (Underfitting) and Variance (Overfitting). Due to lack of time, we are not discussing about Variance now.

+++++++++++++++

* *Explain ANOVA and it’s applications.*

**Answer** :

ANOVA test is a statistical test to see if there’s a difference among multiple groups. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.

Note : It’s a extension of t-test over multiple groups.

Applications :

* A manufacturer has different processes to make cricket bats. They want to know if one process is better than the other.
* Employees from different organization take the same assessment. You want to see if one organization outperforms the other.

**Python Questions :**

* *What is difference between tuple and list in python*

Answer :

Kindly refer below the screen shot taken from my jupyter notebook for the explanation :

A screenshot of a social media post

Description automatically generated

* *How can we merge to (I assume it’s a typo, it should be “two” ) list in python or what function can be used*

Answer :

Kindly refer the below screen shot taken from my Jupyter notebook for the answer :

Method 1 : Straight forward. Using ‘+” Operator

Method 2 : Using the “Extend” Function

A screenshot of a cell phone

Description automatically generated

* *What is list comprehension in python*

Answer :

List comprehensions are used for creating new lists from other iterables.

How it works ?

Contains the expressions which is executed for each element along with the for loop to iterate over each element

Basic syntax :

new\_list = [expression for\_loop\_one\_or\_more conditions]

Basically using “Comprehension” improves code readability.

Note : All comprehension can be converted into for loops, but the reverse is not always possible.

Kindly refer the screen shot of my Jupyter notebook for practical explanation :

A screenshot of a cell phone

Description automatically generated

**NOTE** : My earlier answer to “FizzBuzz” also has comprehension.

* *What is the annotation classmethods used for in python*

Answer : I hope my understanding is correct on the above :

A class method is the one which can be called only with the instance of an Object. These methods usually define the behavior of an object and modify the object’s properties or instance variables.

There are two ways to create class methods in Python:

* Using classmethod(function)
* Using @classmethod annotation

Using classmethod() looks non-Pythonic and hence, in newer Python versions, we can use the @classmethod annotation decorator for making class methods.

* *You have a list of numbers in python, how can we remove the duplicates*

Answer :

We can use the function – Set.

Pls refer the below screen shot for practical explanation :

A screenshot of a cell phone

Description automatically generated

* *What is lambda in python*

Answer :

lamda is used in python to create anonymous function, which means a function without a name. lamda function is an expression

Syntax : lamda arguments : expression

Kindly refer the screen shot of my jupyter notebook for practical explanation :

( Simple lamda function to add 10 to the given number )

A picture containing screenshot

Description automatically generated

* *How does reverse indexing work in python, how can we reverse a string in a single line*

Answer :

Kindly refer the screen shot of my jupyter notebook for practical explanation :

A screenshot of a cell phone

Description automatically generated

* *How can we reindex the data frame in python*

Answer : we can use the function “reindex”

A screenshot of a cell phone

Description automatically generated

* *How to access the nth element in python dataframe*

Answer :

There are multiple ways to access the elements in Python dataframe :

By Index Position : iloc

By Index Value : loc

By Column names :

Kindly refer the screen shot of my jupyter notebook for practical explanation :

A screenshot of a cell phone

Description automatically generated

* *How can we join 2 data frames in python*

Answer :

Using Merge Command :

Note : We can use different times of Joins in the Merge command using the clause “how”, due to lack of time I am not able to get it done practically.

A screenshot of a social media post

Description automatically generated

Concat Command :

A screenshot of a social media post

Description automatically generated

Append Command :

A screenshot of a social media post

Description automatically generated

Join Command :

A screenshot of a cell phone

Description automatically generated

* *Why are dataframes immutable in python*

Answer : This is contradicting for my understanding, as far as I know the dataframes in Python – Pandas can be altered. Sorry if my understanding is wrong ? Are we referring to Spark Dataframes?

* *What is difference between merge and concat in pandas dataframe*

Answer :

* merge() is used to combine two (or more) dataframes on the basis of values of common columns (indices can also be used, use left\_index=True and/or right\_index=True),
* concat() is used to append one (or more) dataframes one below the other (or sideways, depending on whether the axis option is set to 0 or 1).