***SUBJECTIVE QUESTIONS***

**Q 13)How is cluster analysis calculated?**

Cluster analysis is an exploratory analysis that tries to identify structures within the data. Clustering is an unsupervised machine learning algorithm. Since it is exploratory, it does not make any distinction between dependent and independent variables. Three important factors by which clustering can be evaluated are: (a) Clustering tendency (b) Number of clusters, **k**(c) Clustering quality

**Clustering tendency**

Before evaluating the clustering performance, making sure that data set we are working has clustering tendency and does not contain uniformly distributed points is very important. Non-uniform distribution of points in data set becomes important in clustering.

Hopkins test, a statistical test for spatial randomness of a variable, can be used to measure the probability of data points generated by uniform data distribution.

***Null Hypothesis (Ho) :****Data points are generated by uniform distribution (implying no meaningful clusters)****Alternate Hypothesis (Ha):****Data points are generated by random data points (presence of clusters)*  
If H>0.5, null hypothesis can be rejected and it is very much likely that data contains clusters. If H is more close to 0, then data set doesn’t have clustering tendency.

# ****Number of Optimal Clusters,****k

Some of the clustering algorithms like K-means, require number of clusters, k, as clustering parameter. If k is too high, each point will broadly start representing a cluster and if k is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering.

Finding right number of cluster depends upon: (a) Distribution shape (b) scale in the data set (c) clustering resolution required by user. There are two major approaches to find optimal number of clusters:  
(1) Domain knowledge  
(2) Data driven approach

**Domain knowledge**— Domain knowledge might give some prior knowledge on finding number of clusters. For example, in case of clustering iris data set, if we have the prior knowledge of species (sertosa, virginica, versicolor) , then k = 3.

**Data driven approach**— If the domain knowledge is not available, mathematical methods help in finding out right number of clusters. It consists of three methods: Empirical Method, Elbow Method and Statistical approach.

# Clustering quality

Once clustering is done, how well the clustering has performed can be quantified by a number of metrics.

There are majorly two types of measures to assess the clustering performance.

(i) Extrinsic Measures which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) Intrinsic Measures that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

**Q 14)How is cluster quality measured?**

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available.

If ground truth is available, it can be used by **extrinsic methods**, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of “cluster labels.” Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are [*unsupervised methods*](https://www.sciencedirect.com/topics/computer-science/unsupervised-method).

**Q 15)What is cluster analysis and its types?**

Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

There are four basic types of cluster analysis used in data science. These types are: Centroid Clustering, Density Clustering Distribution Clustering, and Connectivity Clustering.

## Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify.

## Density Clustering

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are.

## Distribution Clustering

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions

## Connectivity Clustering

Connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related.

**SQL**

**Q 11)What is data-warehouse?**

**A data warehouse is a large collection of business data used to help an organization make decisions.** The concept of the data warehouse has existed since the 1980s, when it was developed to help transition data from merely powering operations to fueling decision support systems that reveal [business intelligence](https://www.talend.com/resources/what-is-business-intelligence/). The large amount of data in data warehouses comes from different places such as internal applications such as marketing, sales, and finance; customer-facing apps; and external partner systems, among others.

On a technical level, a [data warehouse](https://www.stitchdata.com/resources/data-warehouse/) periodically pulls data from those apps and systems; then, the data goes through formatting and import processes to match the data already in the warehouse. The data warehouse stores this processed data so it’s ready for decision makers to access. How frequently data pulls occur, or how data is formatted, etc., will vary depending on the needs of the organization.

It’s easy to confuse a data warehouse with a [database](https://help.talend.com/reader/TKUQ4WRBbYZRnl9OyAgr5w/jlwGQOe5Yi_W62Zfpj~i9Q), since both concepts share some similarities. The primary difference, however, comes into effect when a business needs to perform analytics on a large data collection. Data warehouses are made to handle this type of task, while databases are not.

**Q 12)What is the difference between OLTP VS OLAP?**

Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.**The primary objective is data processing and not data analysis**

Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.**The primary objective is data analysis and not data processing**.

## KEY DIFFERENCE between OLTP and OLAP:

* Online Analytical Processing (OLAP) is a category of software tools that analyze data stored in a database whereas Online transaction processing (OLTP) supports transaction-oriented applications in a 3-tier architecture.
* OLAP creates a single platform for all type of business analysis needs which includes planning, budgeting, forecasting, and analysis while OLTP is useful to administer day to day transactions of an organization.
* OLAP is characterized by a large volume of data while OLTP is characterized by large numbers of short online transactions.
* In OLAP, data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database whereas OLTP uses traditional DBMS.

**Q 13)What are the various characteristics of data-warehouse?**

The key characteristics of a data warehouse are as follows:

* Some data is denormalized for simplification and to improve performance
* Large amounts of historical data are used
* Queries often retrieve large amounts of data
* Both planned and ad hoc queries are common
* The data load is controlled

In general, fast query performance with high data throughput is the key to a successful data warehouse.

**Q 14)What is Star-Schema?**

­A star schema is a data warehousing architecture model where one fact table references multiple dimension tables, which, when viewed as a diagram, looks like a star with the fact table in the center and the dimension tables radiating from it. It is the simplest among the data warehousing schemas and is currently in wide use.

In the star schema, there is a single fact table. The star schema has been optimized for querying large data sets and is generally used in data marts and warehouses in order to support OLAP cubes, ad hoc queries, analytic applications and business intelligence.  
The fact tables in a star schema usually have two columns: the first is for the foreign keys pointing to the dimension tables, and the second is for the measures that contain numeric facts, hence, the name fact table.

**Q 15)What do you mean by SETL?**

Short for Set Theory as a Language (or Set Language), SETL is a [high-level programming language](https://www.webopedia.com/definitions/high-level-language/) that’s based on the mathematical theory of sets. It was developed in the early 1970’s by mathematician Professor J. Schwartz. SETL is an interpreted language with a [syntax](https://www.webopedia.com/definitions/syntax/) that is resembles [C](https://www.webopedia.com/definitions/c-language/) and in many cases similar to [Perl](https://www.webopedia.com/definitions/perl/). In SETL every statement is terminated by a semicolon. [Variable](https://www.webopedia.com/definitions/variable/) names are case-insensitive and are automatically determined by their last assignment.

**Statistics**

**Q 10)What do you understand by the term Normal Distribution?**

Normal distribution, also known as the Gaussian distribution, is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a [bell curve](https://www.investopedia.com/terms/b/bell-curve.asp). In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

In reality, most pricing distributions are not perfectly normal.

The standard normal distribution has two parameters: the mean and the [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp). For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

The normal distribution model is motivated by the [Central Limit Theorem.](https://www.investopedia.com/terms/c/central_limit_theorem.asp)

**Q 11)How do you handle missing data? What imputation techniques do you recommend?**

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

* **Missing Completely At Random (MCAR):**When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
* **Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data.
* **Not Missing At Random (NMAR):** When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Some of the other common methods are :

1)**Mean or Median Imputation**

A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.

## 2) Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

## 3) Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple [decision trees](https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods" \t "_blank) to estimate missing values and outputs OOB (out of bag) imputation error estimates.

**Q 12)What is A/B testing?**

A/B testing (also known as [split testing](https://www.optimizely.com/optimization-glossary/split-testing/) or [bucket testing](https://www.optimizely.com/optimization-glossary/bucket-testing/)) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

In an A/B test, you take a webpage or app screen and modify it to create a second version of the same page. This change can be as simple as a single headline or button, or be a complete redesign of the page. Then, half of your traffic is shown the original version of the page (known as the control) and half are shown the modified version of the page (the variation).

A/B testing allows individuals, teams, and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses, and to learn better why certain elements of their experiences impact user behavior. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.

**Q 13)Is mean imputation of missing data acceptable practice?**

Mean imputation (also called mean substitution) is popular solution to missing data, despite its drawbacks. Mainly because it’s easy, it can be really painful to lose a large part of the sample you so carefully collected, only to have little [power](https://www.theanalysisfactor.com/5-ways-to-increase-power-in-a-study/). There are many [alternatives to mean imputation](https://www.theanalysisfactor.com/missing-data-two-recommended-solutions/) that provide much more accurate estimates and standard errors, so there really is no excuse to use it.

### Mean imputation does not preserve the relationships among variables.

By imputing the mean, we are able to keep sample size up to the full sample size. This is the original logic involved in mean imputation.If all we are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias our parameter estimate. It *will* still bias our standard error

### Mean Imputation Leads to An Underestimate of Standard Errors

In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.

**Q 14)What is linear regression in statistics?**

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A [scatterplot](http://www.stat.yale.edu/Courses/1997-98/101/scatter.htm) can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the [correlation coefficient](http://www.stat.yale.edu/Courses/1997-98/101/correl.htm), which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

**Q 15)What are the various branches of statistics?**

 If we consider the branches of statistics, there are two branches in it. They are

* **Descriptive statistics**
* **Inferential statistics**

**Descriptive-statistics:**

It organizes raw data into meaningful information. Anhouse hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full. Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this particular branch “ descriptive-statistics". That is, it focuses on collecting, summarizing and presenting set of data.

**For example,**Industrial statistics, population statistics, trade statistics etc.,

**Inferential-statistics:**

It analyses sample data to draw conclusion about population. It analyses sample data to draw conclusion about population. Marketing research team of a company wants to know how far the people need a particular product manufactured by the company. There is one hundred thousand population in a particular city. It is bit difficult to go and ask all one hundred thousand people, due to time consumption and other factors. Hence, it takes a sample of 1000 people to draw conclusion for the whole population. That is making general statement from the study of particular cases or any treatment of data, which leads to prediction or inference concerning a larger group of data.

**For example,** we want to have an idea about percentage of illiterates in a country. We take a sample from a population and the proportion of illiterates in the sample. That sample with the help of probability enables us to find the proportion to the original population.

When we use inferential statistics, we start with a hypothesis, and see whether the data is consistent with that hypothesis.