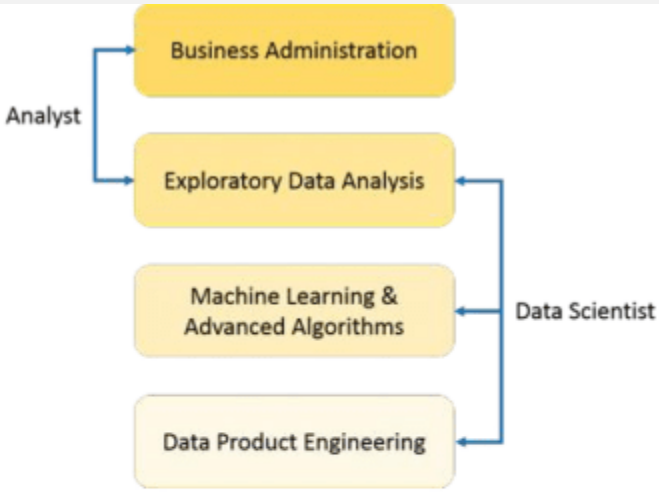
Data Science Interview Questions

What is Data Science?

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.

How is Data Science different from what statisticians have been doing for years? Also, list the differences between supervised and unsupervised learning.

The answer lies in the difference between explaining and predicting.



Supervised Learning vs Unsupervised Learning

Supervised Learning	Unsupervised Learning
1. Input data is labeled.	1. Input data is unlabeled.
2. Uses training dataset.	2. Uses the input data set.
3. Used for prediction.	3. Used for analysis.
4. Enables classification and regression.	4. Enables Classification, Density Estimation, & Dimension Reduction

By Waqas Ali Munawar

Data Science Interview Questions

What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied.

It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples.

If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

The types of selection bias include:

- **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
- **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
- **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
- **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

By Waqas Ali Munawar

Data Science Interview Questions

What is the difference between "long" and "wide" format data?

- In the **wide-format**, a subject's repeated responses will be in a single row, and each response is in a separate column.
- In the **long format**, each row is a one-time point per subject. We can recognize data in wide format by the fact that columns generally represent groups.

Name	Height	Weight
John	160	67
Christopher	182	78

Figure: *Wide Format*

Name	Attribute	Value
John	Height	160
John	Weight	67
Christopher	Height	182
Christopher	Weight	78

Figure: *Long Format*

What do you understand by the term Normal Distribution?

- Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.
- However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.
- The random variables are distributed in the form of the asymmetrical bell-shaped curve.

By Waqas Ali Munawar

Data Science Interview Questions

Properties of Normal Distribution:

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the centre
5. Asymptotic

What is the goal of A/B Testing?

- It is a statistical hypothesis testing for a randomized experiment with two variables A and B.
- The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for the business.
- It can be used to test everything from website copy to sales emails to search ads. An example of this could be identifying the click-through rate for a banner ad.

What are the differences between overfitting and underfitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A

model that has been overfitting has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

Python or R — Which one would we prefer for text analytics?

We will prefer Python because of the following reasons:

• Python would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
• R is more suitable for machine learning than just text analysis.
• Python performs faster for all types of text analytics.

Differentiate between univariate, bivariate and multivariate analysis?

- Univariate analyses are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and the analysis can be referred to as univariate analysis.

- The bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

- The multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

Data Science Interview Questions

What is Cluster Sampling?

- Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.
- For e g., A researcher wants to survey the academic performance of high school students. He can divide the entire population into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once we reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

Can you explain the difference between a Validation Set and a Test Set?

- A Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.
- On the other hand, a Test Set is used for testing or evaluating the performance of a trained machine learning model.
- In simple terms, the differences can be summarized as; training set is to fit the parameters i.e., weights and test set is to assess the performance of the model i.e., evaluating the predictive power and generalization.

By Waqas Ali Munawar

Data Science Interview Questions

Explain cross-validation

- Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will generalize to an independent data set. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

- The goal of cross-validation is to term a data set to test the model in the training phase (i.e., validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

What is Machine Learning?

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics.

What is Supervised Learning?

- Supervised learning is the machine learning task of inferring a function from labelled training data. The training data consist of a set of training examples.

- Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm and Neural Networks

- E.g., If we built a fruit classifier, the labels will be "**this is an orange, this is an apple and this is a banana"**, based on showing the classifier examples of apples, oranges and bananas.

By Waqas Ali Munawar

Data Science Interview Questions

What is Unsupervised learning?

- Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.
- Algorithms: Clustering, Anomaly Detection and Latent Variable Models.
- E.g., In the same example, a fruit clustering will categorize as **“fruits with soft skin and lots of dimples”, “fruits with shiny hard skin” and “elongated yellow fruits”.**

By Waqas Ali Munawar