Top Data Analyst Interview Questions

ZaranTech
knowledge is Everything

Contents

Data Analyst Interview Questions for Freshers

Data Analyst Interview Questions for Experienced

Data Analyst Interview Questions for Experienced

(.....Continued)

19. Write characteristics of a good data model.
20. Write disadvantages of Data analysis.
21. Explain Collaborative Filtering.
22. What do you mean by Time Series Analysis? Where is it used?
23. What do you mean by clustering algorithms? Write different properties of clustering algorithms?
24. What is a Pivot table? Write its usage.
25. What do you mean by univariate, bivariate, and multivariate analysis?
26. Name some popular tools used in big data.
27. Explain Hierarchical clustering.
28. What do you mean by logistic regression?
29. What do you mean by the K-means algorithm?
30. Write the difference between variance and covariance.
31. What are the advantages of using version control?
32. Explain N-gram
33. Mention some of the statistical techniques that are used by Data analysts.
34. What's the difference between a data lake and a data warehouse?

Let's get Started

What is Data Analysis?

Data analysis is basically a process of analyzing, modeling, and interpreting data to draw insights or conclusions. With the insights gained, informed decisions can be made. It is used by every industry, which is why data analysts are in high demand. A Data Analyst's sole responsibility is to play around with large amounts of data and search for hidden insights. By interpreting a wide range of data, data analysts assist organizations in understanding the business's current state.

Data Analyst Interview Questions for Freshers

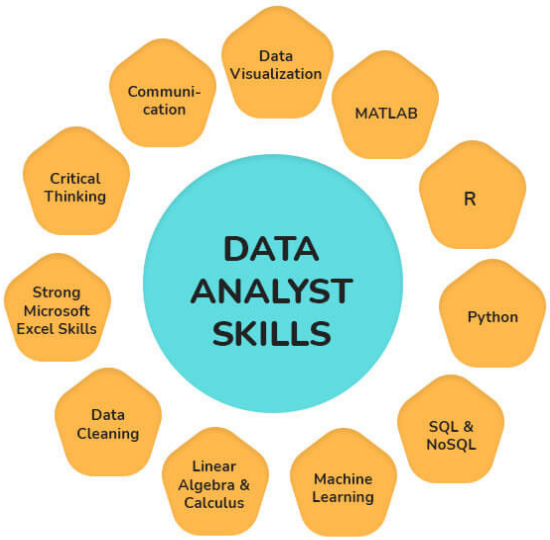
1. What are the responsibilities of a Data Analyst?

Some of the responsibilities of a data analyst include:

- Collects and analyzes data using statistical techniques and reports the results accordingly.
- Interpret and analyze trends or patterns in complex data sets.
- Establishing business needs together with business teams or management t eam s.
- Find opportunities for improvement in existing processes or areas.
- Data set commissioning and decommissioning.
- Follow guidelines when processing confidential data or information.
- Examine the changes and updates that have been made to the source production systems.
- Provide end-users with training on new reports and dashboards.
- Assist in the data storage structure, data mining, and data cleansing.

2. Write some key skills usually required for a data analyst.

Some of the key skills required for a data analyst include:



- Knowledge of reporting packages (Business Objects), coding languages (e.g., XML, JavaScript, ETL), and databases (SQL, SQLite, etc.) is a must.
- Ability to analyze, organize, collect, and disseminate big data accurately and efficiently.
- The ability to design databases, construct data models, perform data mining, and segment data.
- Good understanding of statistical packages for analyzing large datasets (SAS, SPSS, Microso Excel, etc.).
- Effective Problem-Solving, Teamwork, and Written and Verbal Communication Skills.
- Excellent at writing queries, reports, and presentations.
- Understanding of data visualization so ware including Tableau and Qlik.
- The ability to create and apply the most accurate algorithms to datasets for finding solutions.

3. What is the data analysis process?

Data analysis generally refers to the process of assembling, cleaning, interpreting, transforming, and modeling data to gain insights or conclusions and generate reports to help businesses become more profitable. The following diagram illustrates the various steps involved in the process:

- **Collect Data:** The data is collected from a variety of sources and is then stored to be cleaned and prepared. This step involves removing all missing values and outliers.
- **Analyse Data:** As soon as the data is prepared, the next step is to analyze it. Improvements are made by running a model repeatedly. Following that, the model is validated to ensure that it is meeting the requirements.
- **Create Reports:** In the end, the model is implemented, and reports are generated as well as distributed to stakeholders.

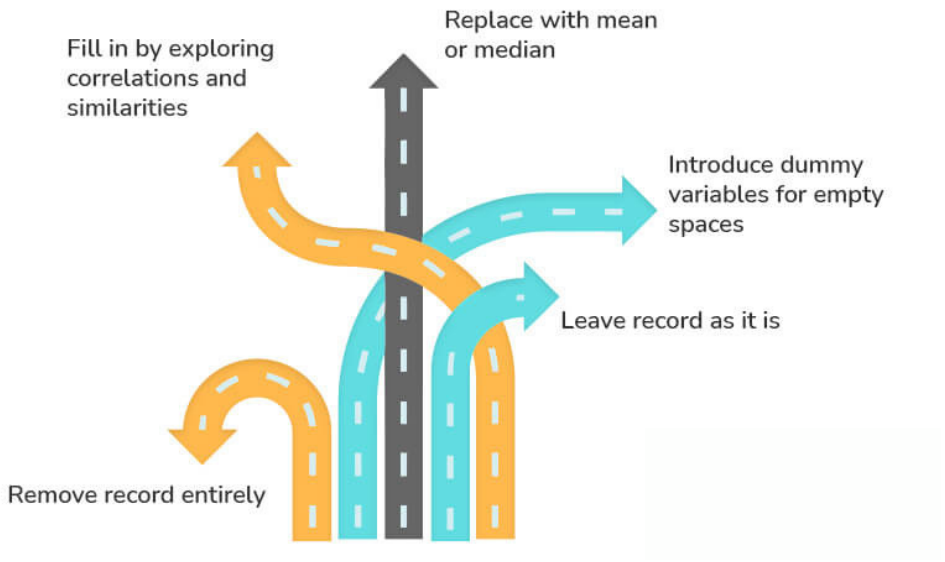
4. What are the different challenges one faces during data analysis?

While analyzing data, a Data Analyst can encounter the following issues:

- Duplicate entries and spelling errors. Data quality can be hampered and reduced by these errors.
- The representation of data obtained from multiple sources may differ. It may cause a delay in the analysis process if the collected data are combined a er being cleaned and organized.
- Another major challenge in data analysis is incomplete data. This would invariably lead to errors or faulty results.
- You would have to spend a lot of time cleaning the data if you are extracting data from a poor source.
- Business stakeholders' unrealistic timelines and expectations
- Data blending/ integration from multiple sources is a challenge, particularly if there are no consistent parameters and conventions
- Insufficient data architecture and tools to achieve the analytics goals on time.

5. Explain data cleansing.

Data cleaning, also known as data cleansing or data scrubbing or wrangling, is basically a process of identifying and then modifying, replacing, or deleting the incorrect, incomplete, inaccurate, irrelevant, or missing portions of the data as the need arises. This fundamental element of data science ensures data is correct, consistent, and usable.



6. What are the tools useful for data analysis?

Some of the tools useful for data analysis include:

- RapidMiner
- KNIME
- Google Search Operators
- Google Fusion Tables
- Solver
- NodeXL
- OpenRefine
- Wolfram Alpha
- io
- Tableau, etc.

7. Write the difference between data mining and data profiling.

Data mining Process: It generally involves analyzing data to find relations that were not previously discovered. In this case, the emphasis is on finding unusual records, detecting dependencies, and analyzing clusters. It also involves analyzing large datasets to determine trends and patterns in them.

Data Profiling Process: It generally involves analyzing that data's individual attributes. In this case, the emphasis is on providing useful information on data attributes such as data type, frequency, etc. Additionally, it also facilitates the discovery and evaluation of enterprise metadata.

Data Mining	Data Profiling
It involves analyzing a pre-built database to identify patterns.	It involves analyses of raw data from existing datasets.
It also analyzes existing databases and large datasets to convert raw data into useful information.	In this, statistical or informative summaries of the data are collected.
It usually involves finding hidden patterns and seeking out new, useful, and non-trivial data to generate useful information.	It usually involves the evaluation of data sets to ensure consistency, uniqueness, and logic.
Data mining is incapable of identifying inaccurate or incorrect data values.	In data profiling, erroneous data is identified during the initial stage of analysis.
Classification, regression, clustering, summarization, estimation, and description are some primary data mining tasks that are needed to be performed.	This process involves using discoveries and analytical methods to gather statistics or summaries about the dat a.

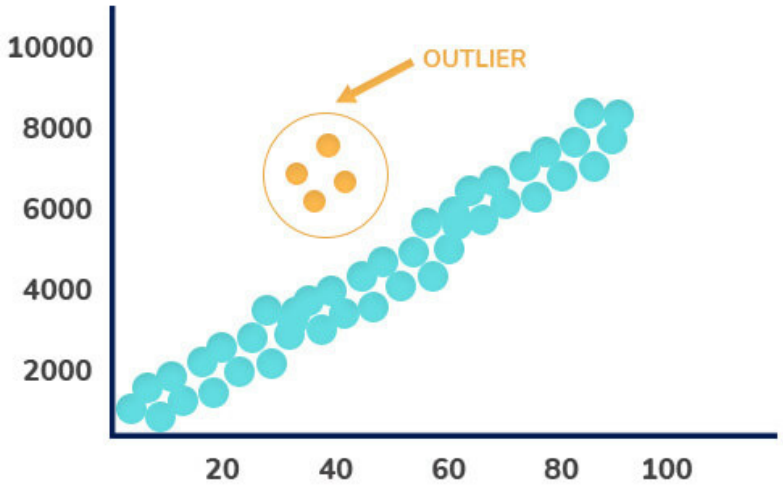
8. Which validation methods are employed by data analysts?

In the process of data validation, it is important to determine the accuracy of the information as well as the quality of the source. Datasets can be validated in many ways. Methods of data validation commonly used by Data Analysts include:

- **Field Level Validation**: This method validates data as and when it is entered into the field. The errors can be corrected as you go.
- **Form Level Validation**: This type of validation is performed a er the user submits the form. A data entry form is checked at once, every field is validated, and highlights the errors (if present) so that the user can fix them.
- **Data Saving Validation**: This technique validates data when a file or database record is saved. The process is commonly employed when several data entry forms must be validated.
- **Search Criteria Validation**: It effectively validates the user's search criteria in order to provide the user with accurate and related results. Its main purpose is to ensure that the search results returned by a user's query are highly relevant.

9. Explain Outlier.

In a dataset, Outliers are values that differ significantly from the mean of characteristic features of a dataset. With the help of an outlier, we can determine either variability in the measurement or an experimental error. There are two kinds of outliers i.e., Univariate and Multivariate. The graph depicted below shows there are four outliers in the dataset.



10. What are the ways to detect outliers? Explain different ways to deal with it.

Outliers are detected using two methods:

- **Box Plot Method**: According to this method, the value is considered an outlier if it exceeds or falls below 1.5*IQR (interquartile range), that is, if it lies above the top quartile (Q3) or below the bottom quartile (Q1).
- **Standard Deviation Method**: According to this method, an outlier is defined as a value that is greater or lower than the mean ± (3*standard deviation).

11. Write difference between data analysis and data mining.

Data Analysis: It generally involves extracting, cleansing, transforming, modeling, and visualizing data in order to obtain useful and important information that may contribute towards determining conclusions and deciding what to do next. Analyzing data has been in use since the 1960s.
Data Mining: In data mining, also known as knowledge discovery in the database, huge quantities of knowledge are explored and analyzed to find patterns and rules. Since the 1990s, it has been a buzzword.


DATA MINING
VS
DATA ANALYSIS

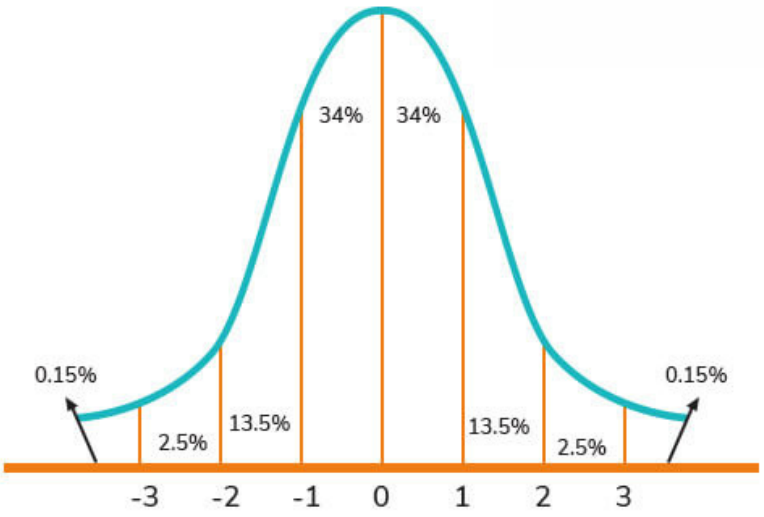
Data Analysis	Data Mining
Analyzing data provides insight or tests hypotheses.	A hidden pattern is identified and discovered in large datasets.
It consists of collecting, preparing, and modeling data in order to extract meaning or insights.	This is considered as one of the activities in Data Analysis.
Data-driven decisions can be taken using this way.	Data usability is the main objective.
Data visualization is certainly required.	Visualization is generally not necessary.
It is an interdisciplinary field that requires knowledge of computer science, statistics, mathematics, and machine learning.	Databases, machine learning, and statistics are usually combined in this field.
Here the dataset can be large, medium, or small, and it can be structured, semi-structured, and unstructured.	In this case, datasets are typically large and structured.

12. Explain the KNN imputation method.

A KNN (K-nearest neighbor) model is usually considered one of the most common techniques for imputation. It allows a point in multidimensional space to be matched with its closest k neighbors. By using the distance function, two attribute values are compared. Using this approach, the closest attribute values to the missing values are used to impute these missing values.

13. Explain Normal Distribution.

Known as the bell curve or the Gauss distribution, the Normal Distribution plays a key role in statistics and is the basis of Machine Learning. It generally defines and measures how the values of a variable differ in their means and standard deviations, that is, how their values are distributed.



The above image illustrates how data usually tend to be distributed around a central value with no bias on either side. In addition, the random variables are distributed according to symmetrical bell-shaped curves.

14. What do you mean by data visualization?

The term data visualization refers to a graphical representation of information and data. Data visualization tools enable users to easily see and understand trends, outliers, and patterns in data through the use of visual elements like charts, graphs, and maps. Data can be viewed and analyzed in a smarter way, and it can be converted into diagrams and charts with the use of this technology.

15. How does data visualization help you?

Data visualization has grown rapidly in popularity due to its ease of viewing and understanding complex data in the form of charts and graphs. In addition to providing data in a format that is easier to understand, it highlights trends and outliers. The best visualizations illuminate meaningful information while removing noise from data.

16. Mention some of the python libraries used in data analysis.

Several Python libraries that can be used on data analysis include:

- Num Py
- Bokeh
- Matplotlib
- Pandas
- SciPy
- SciKit, etc.

17. Explain a hash table.

Hash tables are usually defined as data structures that store data in an associative manner. In this, data is generally stored in array format, which allows each data value to have a unique index value. Using the hash technique, a hash table generates an index into an array of slots from which we can retrieve the desired value.

18. What do you mean by collisions in a hash table? Explain the ways to avoid it.

Hash table collisions are typically caused when two keys have the same index. Collisions, thus, result in a problem because two elements cannot share the same slot in an array. The following methods can be used to avoid such hash collisions:

- **Separate chaining technique:** This method involves storing numerous items hashing to a common slot using the data structure.
- **Open addressing technique:** This technique locates unfilled slots and stores the item in the first unfilled slot it finds.

Data Analyst Interview Questions for Experienced

19. Write characteristics of a good data model.

An effective data model must possess the following characteristics in order to be considered good and developed:

- Provides predictability performance, so the outcomes can be estimated as precisely as possible or almost as accurately as possible.
- As business demands change, it should be adaptable and responsive to accommodate those changes as needed.
- The model should scale proportionally to the change in data.
- Clients/customers should be able to reap tangible and profitable benefits from it.

20. Write disadvantages of Data analysis.

The following are some disadvantages of data analysis:

- Data Analytics may put customer privacy at risk and result in compromising transactions, purchases, and subscriptions.
- Tools can be complex and require previous training.
- Choosing the right analytics tool every time requires a lot of skills and expertise.
- It is possible to misuse the information obtained with data analytics by targeting people with certain political beliefs or ethnicities.

21. Explain Collaborative Filtering.

Based on user behavioral data, collaborative filtering (CF) creates a recommendation system. By analyzing data from other users and their interactions with the system, it filters out information. This method assumes that people who agree in their evaluation of particular items will likely agree again in the future. Collaborative filtering has three major components: users- items- interests.

Example:
Collaborative filtering can be seen, for instance, on online shopping sites when you see phrases such as "recommended for you".

22. What do you mean by Time Series Analysis? Where is it used?

In the field of Time Series Analysis (TSA), a sequence of data points is analyzed over an interval of time. Instead of just recording the data points intermittently or randomly, analysts record data points at regular intervals over a period of time in the TSA. It can be done in two different ways: in the frequency and time domains. As TSA has a broad scope of application, it can be used in a variety of fields. TSA plays a vital role in the following places:

- Statistics
- Signal processing
- Econometrics
- Weather forecasting
- Earthquake prediction
- Astronomy
- Applied science

23. What do you mean by clustering algorithms? Write different properties of clustering algorithms?

Clustering is the process of categorizing data into groups and clusters. In a dataset, it identifies similar data groups. It is the technique of grouping a set of objects so that the objects within the same cluster are similar to one another rather than to those located in other clusters. When implemented, the clustering algorithm possesses the following properties:

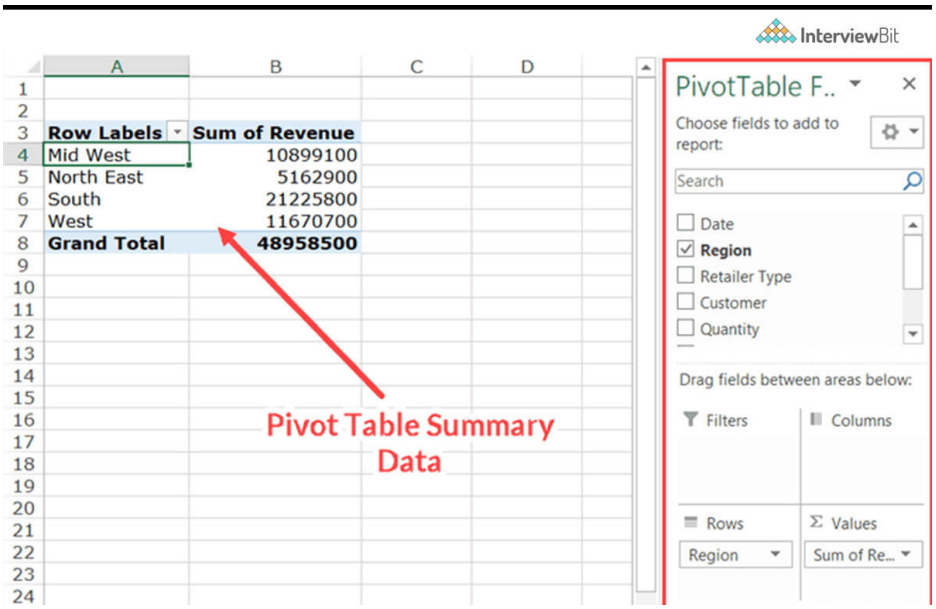
- Flat or hierarchical
- Hard or So
- Iterative
- Disjunctive

24. What is a Pivot table? Write its usage.

One of the basic tools for data analysis is the Pivot Table. With this feature, you can quickly summarize large datasets in Microso Excel. Using it, we can turn columns into rows and rows into columns. Furthermore, it permits grouping by any field (column) and applying advanced calculations to them. It is an extremely easy-to-use program since you just drag and drop rows/columns headers to build a report. Pivot tables consist of four different sections:

Value Area: This is where values are reported.

Row Area: The row areas are the headings to the le of the values.
Column Area: The headings above the values area make up the column area.
Filter Area: Using this filter you may drill down in the data set.



25. What do you mean by univariate, bivariate, and multivariate analysis?

- **Univariate Analysis:** The word uni means only one and variate means variable, so a univariate analysis has only one dependable variable. Among the three analyses, this is the simplest as the variables involved are only one.
  Example: A simple example of univariate data could be height as shown below:

Heights (in cm)	164	167.3	170	174.2	178	180	186

- **Bivariate Analysis:** The word Bi means two and variate mean variables, so a bivariate analysis has two variables. It examines the causes of the two variables and the relationship between them. It is possible that these variables are dependent on or independent of each other.
  Example: A simple example of bivariate data could be temperature and ice cream sales in the summer season.

TEMPERATURE (IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

- **Multivariate Analysis:** In situations where more than two variables are to be analyzed simultaneously, multivariate analysis is necessary. It is similar to bivariate analysis, except that there are more variables involved.

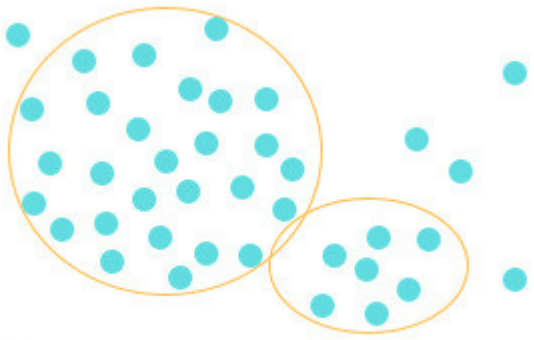
26. Name some popular tools used in big data.

In order to handle Big Data, multiple tools are used. There are a few popular ones as follows:

- Hadoop
- Spark
- Scala
- Hive
- Flume
- Mahout, etc.

27. Explain Hierarchical clustering.

This algorithm group objects into clusters based on similarities, and it is also called hierarchical cluster analysis. When hierarchical clustering is performed, we obtain a set of clusters that differ from each other.



This clustering technique can be divided into two types:

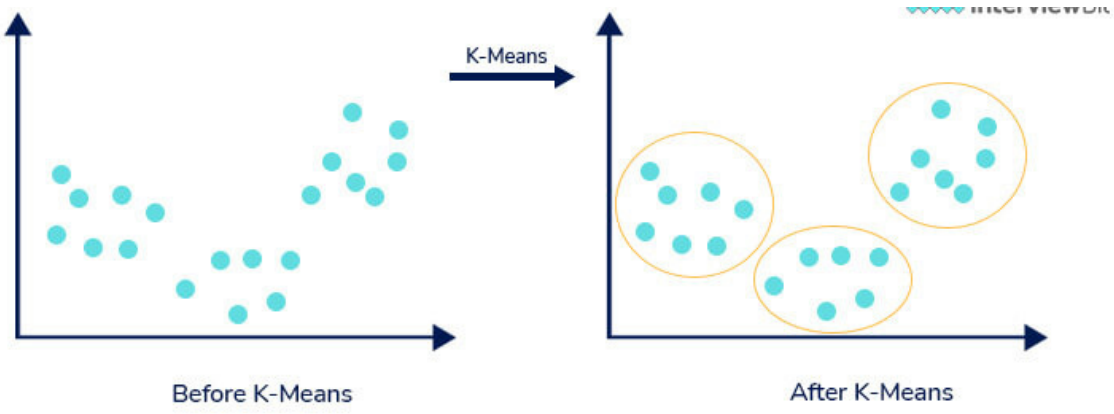
- Agglomerative Clustering (which uses bottom-up strategy to decompose clusters)
- Divisive Clustering (which uses a top-down strategy to decompose clusters)

28. What do you mean by logistic regression?

Logistic Regression is basically a mathematical model that can be used to study datasets with one or more independent variables that determine a particular outcome. By studying the relationship between multiple independent variables, the model predicts a dependent data variable.

29. What do you mean by the K-means algorithm?

One of the most famous partitioning methods is K-mean. With this unsupervised learning algorithm, the unlabeled data is grouped in clusters. Here, 'k' indicates the number of clusters. It tries to keep each cluster separated from the other. Since it is an unsupervised model, there will be no labels for the clusters to work with.



30. Write the difference between variance and covariance.

Variance: In statistics, variance is defined as the deviation of a data set from its mean value or average value. When the variances are greater, the numbers in the data set are farther from the mean. When the variances are smaller, the numbers are nearer the mean. Variance is calculated as follows:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Here, X represents an individual data point, U represents the average of multiple data points, and N represents the total number of data points.

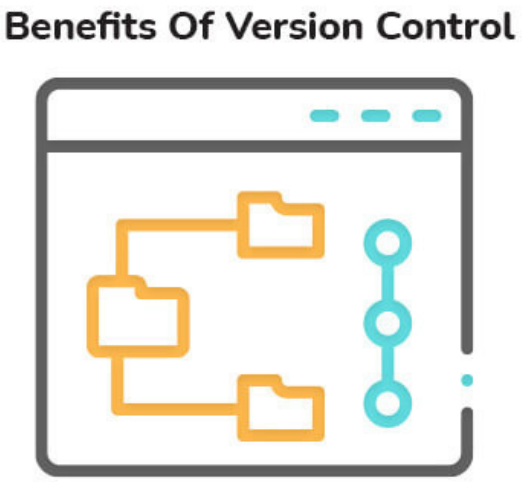
Covariance: Covariance is another common concept in statistics, like variance. In statistics, covariance is a measure of how two random variables change when compared with each other. Covariance is calculated as follows:

$$COV(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Here, X represents the independent variable, Y represents the dependent variable, x-bar represents the mean of the X, y-bar represents the mean of the Y, and N represents the total number of data points in the sample.

31. What are the advantages of using version control?

Also known as source control, version control is the mechanism for configuring so ware. Records, files, datasets, or documents can be managed with this. Version control has the following advantages:



- Analysis of the deletions, editing, and creation of datasets since the original copy can be done with version control.
- So ware development becomes clearer with this method.
- It helps distinguish different versions of the document from one another. Thus, the latest version can be easily identified.
- There's a complete history of project files maintained by it which comes in handy if ever there's a failure of the central server.
- Securely storing and maintaining multiple versions and variants of code files is easy with this tool.
- Using it, you can view the changes made to different files.

32. Explain N-gram

N-gram, known as the probabilistic language model, is defined as a connected sequence of n items in a given text or speech. It is basically composed of adjacent words or letters of length n that were present in the source text. In simple words, it is a way to predict the next item in a sequence, as in (n-1).

33. Mention some of the statistical techniques that are used by Data analysts.

Performing data analysis requires the use of many different statistical techniques. Some important ones are as follows:

- Markov process
- Cluster analysis
- Imputation techniques
- Bayesian methodologies
- Rank statistics

34. What's the difference between a data lake and a data warehouse?

The storage of data is a big deal. Companies that use big data have been in the news a lot lately, as they try to maximize its potential. Data storage is usually handled by traditional databases for the layperson. For storing, managing, and analyzing big data, companies use data warehouses and data lakes.

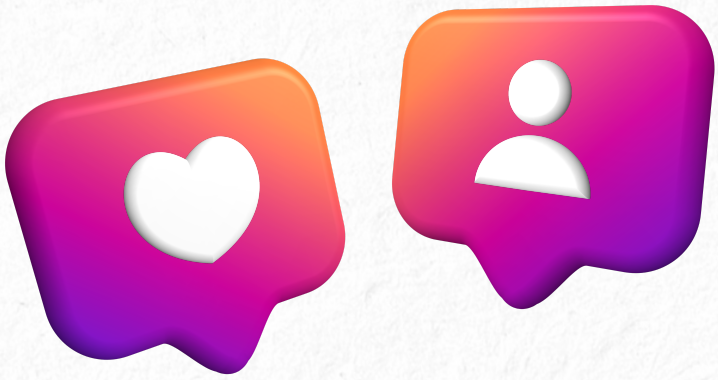
Data Warehouse: This is considered an ideal place to store all the data you gather from many sources. A data warehouse is a centralized repository of data where data from operational systems and other sources are stored. It is a standard tool for integrating data across the team- or department-silos in mid-and large-sized companies. It collects and manages data from varied sources to provide meaningful business insights. Data warehouses can be of the following types:

- **Enterprise data warehouse (EDW)**: Provides decision support for the entire organization.
- **Operational Data Store (ODS)**: Has functionality such as reporting sales data or employee data.

Data Lake: Data lakes are basically a large storage device that stores raw data in their original format until they are needed. with its large amount of data, analytical performance and native integration are improved. It exploits data warehouses' biggest weakness: their incapacity to be flexible. In this, neither planning nor knowledge of data analysis is required; the analysis is assumed to happen later, on-dem and.

Conclusion:

The purpose of Data Analysis is to transform data to discover valuable information that can be used for making decisions. The use of data analytics is crucial in many industries for various purposes, hence, the demand for Data Analysts is therefore high around the world. Therefore, we have listed the top data analyst interview questions & answers you should know to succeed in your interview. From data cleaning to data validation to SAS, these questions cover all the essential information related to the data analyst role.



@datascience-trainer

ZaranTech
knowledge is Everything