Basic Data Analyst Interview Questions For Freshers

1. What are the key differences between Data Analysis and Data Mining?

Data analysis involves the process of cleaning, organizing, and using data to produce meaningful insights. Data mining is used to search for hidden patterns in the data.

Data analysis produces results that are far more comprehensible by a variety of audiences than the results from data mining.

2. What is Data Validation?

Data validation, as the name suggests, is the process that involves determining the accuracy of data and the quality of the source as well. There are many processes in data validation but the main ones are data screening and data verification.

- Data screening: Making use of a variety of models to ensure that the data is accurate and no redundancies are present.
- Data verification: If there is a redundancy, it is evaluated based on multiple steps and then a call is taken to ensure the presence of the data item.

3. What is Data Analysis, in brief?

Data analysis is the structured procedure that involves working with data by performing activities such as ingestion, cleaning, transforming, and assessing it to provide insights, which can be used to drive revenue.

Data is collected, to begin with, from varied sources. Since the data is a raw entity, it has to be cleaned and processed to fill out missing values and to remove any entity that is out of the scope of usage.

After preprocessing the data, it can be analyzed with the help of models, which use the data to perform some analysis on it.

The last step involves reporting and ensuring that the data output is converted to a format that can also cater to a non-technical audience, alongside the analysts.

4. How to know if a data model is performing well or not?

This question is subjective, but certain simple assessment points can be used to assess the accuracy of a data model. They are as follows:

- A well-designed model should offer good predictability. This correlates to the ability to be easily able to predict future insights when needed.
- A rounded model adapts easily to any change made to the data or the pipeline if need be.
- The model should have the ability to cope in case there is an immediate requirement to large-scale the data.
- The model's working should be easy and it should be easily understood among clients to help them derive the required results.

5. Explain Data Cleaning in brief.

Data Cleaning is also called Data Wrangling. As the name suggests, it is a structured way of finding erroneous content in data and safely removing them to ensure that the data is of the utmost quality. Here are some of the ways in data cleaning:

- Removing a data block entirely
- Finding ways to fill black data in, without causing redundancies
- Replacing data with its mean or median values
- Making use of placeholders for empty spaces

6. What are some of the problems that a working Data Analyst might encounter?

There can be many issues that a Data Analyst might face when working with data. Here are some of them:

- The accuracy of the model in development will be low if there are multiple entries of the same entity and errors concerning spellings and incorrect data.
- If the source the data being ingested from is not a verified source, then the data might require a lot of cleaning and preprocess before beginning the analysis.
- The same goes for when extracting data from multiple sources and merging them for use.
- The analysis will take a backstep if the data obtained is incomplete or inaccurate.

7. What is Data Profiling?

Data profiling is a methodology that involves analyzing all entities present in data to a greater depth. The goal here is to provide highly accurate information based on the data and its attributes such as the datatype, frequency of occurrence, and more.

8. What are the scenarios that could cause a model to be retrained?

Data is never a stagnant entity. If there is an expansion of business, this could cause sudden opportunities that call for a change in the data. Furthermore, assessing the model to check its standing can help the analyst analyze whether the model is to be retrained or not.

However, the general rule of thumb is to ensure that the models are retrained when there is a change in the business protocols and offerings.

9. What are the prerequisites to become a Data Analyst?

There are many skills that a budding data analyst needs. Here are some of them:

- Being well-versed in programming languages such as XML, JavaScript, and ETL frameworks
- Proficient in databases such as SQL, MongoDB, and more
- Ability to effectively collect and analyze data
- Knowledge of database designing and data mining
- Having the ability/experience of working with large datasets

10. What are the top tools used to perform Data Analysis?

There is a wide spectrum of tools that can be used in the field of data analysis. Here are some of the popular ones:

- Google Search Operators
- RapidMiner
- Tableau
- KNIME
- OpenRefine

11. What is an outlier?

An outlier is a value in a dataset that is considered to be away from the mean of the characteristic feature of the dataset. There are two types of outliers: univariate and multivariate.

12. How can we deal with problems that arise when the data flows in from a variety of sources?

There are many ways to go about dealing with multi-source problems. However, these are done primarily to solve the problems of:

- Identifying the presence of similar/same records and merging them into a single record
- Re-structuring the schema to ensure there is good schema integration

13. What are some of the popular tools used in Big Data?

There are multiple tools that are used to handle Big Data. Some of the most popular ones are as follows:

- Hadoop
- Spark
- Scala
- Hive
- Flume
- Mahout

14. What is the use of a Pivot table?

Pivot tables are one of the key features of Excel. They allow a user to view and summarize the entirety of large datasets simply. Most of the operations with Pivot tables involve drag-and-drop operations that aid in the quick creation of reports.

15. Explain the KNN imputation method, in brief.

KNN is the method that requires the selection of several nearest neighbors and a distance metric at the same time. It can predict both discrete and continuous attributes of a dataset.

A distance function is used here to find the similarity of two or more attributes, which will help in further analysis.

16. What are the top Apache frameworks used in a distributed computing environment?

MapReduce and Hadoop are considered to be the top Apache frameworks when the situation calls for working with a huge dataset in a distributed working environment.

17. What is Hierarchical Clustering?

Hierarchical clustering, or hierarchical cluster analysis, is an algorithm that groups similar objects into common groups called clusters. The goal is to create a set of clusters, where each cluster is different from the other and, individually, they contain similar entities.

18. What are the steps involved when working on a data analysis project?

Many steps are involved when working end-to-end on a data analysis project. Some of the important steps are as mentioned below:

- Problem statement
- Data cleaning/preprocessing
- Data exploration
- Modeling
- Data validation
- Implementation
- Verification

19. Can you name some of the statistical methodologies used by data analysts?

Many statistical techniques are very useful when performing data analysis. Here are some of the important ones:

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