By OnlineInterviewQuestions.com

Data Analyst Interview Questions and Answers

Data Analysis is the art of collecting and analyzing data so that the company can use the same to perfect their marketing, insurance, political, and other business practices. These data analysts are highly trained professionals and they perform the analysis by using various mathematical calculations and further determine how the data samples might best be applied to increase the profit of the business. One of the critical roles in the evaluation of risk.

As most companies are always looking to expand their businesses, or at least improve their business practices, data analysis is an essential and profitable field. Data Analyst seeks to understand the origin of data and any possible distortions through the use of technology. If one can identify trends and patterns of information and also has excellent computer skills, then they can find their niche as a Data Analyst.

In this role, a person is asked to use their technical expertise to extrapolate data by using advanced computerized models. The job is professional and heavily influenced by mathematics and advanced algorithms. One can be a Data cleaner, rooting out errors in data or can be employed on Initial Analysis, whereby the assessment of the quality of data is done. As the Main Data Analyst, one is asked to look at the meaning of data, and if they work on the Final Analysis.

Practice Best Data Analyst Interview Questions and Answers

Practice Best **Data Analyst Interview Questions and Answers** for the best preparation of the Data Analyst Interview. These **Data Analyst Interview Questions** are very popular and asked various times in Data Analyst Interviews. So, practice these questions to check your final preparation for your interview. apart from this, you can also download the **data analyst interview questions pdf** completely free.

Q1. How can we differentiate between Data Mining and Data Analysis?

Here are a few considerable differences:

- **Data Mining**: Data mining does not require any hypothesis and depends on clean and well-documented data. Results of data mining are not always easy to interpret. Its algorithms automatically develop equations.
- **Data Analysis**: Whereas, Data analysis begins with a question or an assumption. Data analysis involves data cleaning. The work of the analysts is to interpret the results and convey the same to the stakeholders. Data analysts have to develop their equations based on the hypothesis.

Q2. How do we conduct a data analysis?

Data analysis deals with collecting, inspecting, cleaning, transforming and modeling data to glean valuable

insights and support better decision-making in an organization along with the motive to increase the profit. The various steps involved in the data analysis process can be a sequence in the following manner,

- **Data Exploration**: After identifying the business problem, a data analyst has to go through the data provided by the client to analyze the cause of the problem.
- **Data Preparation**: This is the most crucial step of this process wherein any data anomalies (like missing values or detecting outliers with the data have to be modeled in the right direction.
- **Data Modeling**: This step begins once the data has been prepared. Modeling is an iterative process wherein a model is repeatedly run for improvements. This ensures that the best possible result is found for a given problem.
- **Validation**: In this step, validation of the data is done between the one provided by the client and the model developed by the data analyst. The aim is to find out if the developed model will meet the business requirements or not.
- **Implementation of the Model and Tracking**: This is the final step, where the model is implemented in production and is tested for accuracy and efficiency purpose.

Q3. Define data cleansing? What best practices do you follow during data cleansing?

Data cleaning is a crucial step in the analysis process where data is inspected to find anomalies, reduce repetitive data, and eliminate incorrect information. It does not involve deleting any existing information from the database; it focuses on enhancing the quality of data so that it can be used for analysis further.
Some of the best practices for data cleansing includes:

- Development of quality data plans to identify where maximum data quality errors occur, to assess the root cause and design the plan accordingly.
- To follow a standard process of verifying the critical data before the creation of a database.
- To identify any duplicates and validate the accuracy of the data to save time during analysis.
- Tracking all the cleaning operations performed on the data is essential to repeat or remove any operations as necessary.

Q4. List the steps in an analytics project?

Steps included in an analytics project are;

- Problem definition
- Data exploration
- Data preparation
- Modeling
- Validation of data
- Implementation and tracking.

Q5. How will you define logistic regression?

Logistic regression is a statistical method that analyze a dataset, in which there are one or more independent variables and it determine the outcome. It is measured with a dichotomous variable. The objective of logistic regression is to determine the suitable fitting model to describe the relationship between the dichotomous

characteristic of interest and a set of independent variables. Logistic regression generates the coefficients of a formula to predict a logistic transformation of the probability of a presence of the characteristic of interest.

Q6. List down some of the best tools that can be useful for data-analysis?

Some of the best tools that can be useful for data-analysis are:

- Tableau
- Rapid Miner
- Open Refine
- KNIME
- Google Search Operators
- Solver
- Node XL
- IO
- Wolfram Alpha's
- Google Fusion tables

Q7. What are some of the most common problems faced by a data analyst?

Some of the common problems faced by data analyst are

- Common misspelling
- Duplication in entries
- Missing out the values
- Illegal values
- Varying value representations
- Identifying overlapping data

Q8. What missing patterns a data analyst observes?

The missing patterns that are generally observed are

- Missing completely at random
- Missing at random
- Missing that depends on the missing value itself
- Missing that depends on an unobserved input variable

Q9. Define KNN imputation method?

In KNN imputation, the missing attribute values are imputed by using the value of the attribute that is most similar to the quality whose values are missing. By using a distance function, the similarity of two attributes is determined.

Q10. List the data validations methods used by data analysts?

Usually, methods used by a data analyst for data validation are:

- Data screening
- Data verification

Q11. How to deal with multi-source problems?

To deal with multi-source problems one should:

- Get involves in a restructuring of schemas, to accomplish schema integration.
- And, Identify similar records and merge them into a single document containing all relevant attributes without redundancy.

Q12. Define Outlier?

It is a commonly used term by analysts, referred for a value that appears far away and diverges from an overall pattern in a sample. Outliers can be classified into two types;

- Univariate
- Multivariate

Q13. Define collaborative filtering?

Collaborative filtering is a simple algorithm to create a recommendation system based on user behavioral data. The most critical components of collaborative filtering are users- items- interest. One of the examples of collaborative filtering is when you see a statement like “recommended for you” on online shopping sites that pop out based on your browsing history.

Q14. What will you do if a data is suspected or missing?

In case of suspected or missing data following steps should be taken;

- Preparation of a validation report that gives information on all suspected data. Information like validation criteria that it failed and the date and time of occurrence should be taken care of.
- Experience personnel should examine the suspicious data to determine their acceptability.
- Invalid data should be assigned and replaced with a validation code.
- To work on missing data best use of analysis strategy like deletion method, single imputation methods, model-based methods, etc. should be followed up.

Q15. How to create a classification to recognize an essential customer trend in an unorganized data?

- Initially there is a need to consult with the stakeholders of the business to understand the objective of

classifying the data. Then, pull new data samples and modifying the model accordingly and evaluating it for accuracy. For this, a necessary process of mapping the data, creating an algorithm, mining the data and visualization is done. However, one can accomplish this in multiple segments by considering the feedback from stakeholders to ensure that model can produce actionable results.

- A model does not hold any value if it cannot produce actionable results, an experienced data analyst will have a different strategy based on the type of data being analyzed.

Q16. How can you define interquartile range as a data analyst?

The measure of the dispersion of data that is shown in a box plot referred to as the interquartile range. It is the difference between the upper and the lower quartile.

Q17. What criteria can define a good data model?

To say a **model is good**, the following points need to be considered.

- The developed model should have predictable performance.
- It should be adaptable easily to any changes as per business requirements.
- It should be scalable to any data change.
- A model should be efficiently consumed for actionable results.

Q18. Define the essential steps required for data validation process?

Data Validation is performed in 2 different steps:
Data Screening: In this step various algorithms are used to screen the entire data to find any erroneous or questionable values.

Data Verification: In this step each suspect value is evaluated on a case by case basis, and a decision is made if the values have to be accepted as valid, rejected as invalid or if they have to be replaced with some redundant values.

Q19. What steps can be used to work on a QA if a predictive model is developed for forecasting?

Here is a way to handle the QA process efficiently:

- Firstly, partition the data into three different sets Training, Testing and Validation.
- Secondly, show the results of the validation set to the business owner by eliminating biases from the first two sets. The input from the business owner or the client will give an idea of whether the model predicts customer churn with accuracy and provides desired results or not.
- Data analysts require inputs from the business owners and a collaborative environment to operationalize analytics. To create and deploy predictive models in production there should be an effective, efficient and repeatable process. Without taking feedback from the business owner, the model will be a one-and-done model.

Q20. How many times the retrain of a data model is required?

There is a need to refresh or retrain a model when the company enters a new market, consummate an acquisition or is facing emerging competition. As a data analyst, one should retrain the model as quickly as possible to adjust with the changing behavior of customers or change in market conditions. A good data analyst is the one who understands how changing business dynamics will affect the efficiency of a predictive model.

Q21. What is Data Mining?

Data Mining is the process to extract the data or required information from the databases. the data extracted can be used for finding solutions or in the implementation of data.