Data Analysis Question & Answers

1. Explain the typical data analysis process.

Data analysis deals with collecting, inspecting, cleaning, transforming and modeling data to glean valuable insights and support better decision making in an organization. The various steps involved in the data analysis process include –

Data Exploration –

Having identified the business problem, a data analyst has to go through the data provided by the client to analyze the root cause of the problem.

Data Preparation

This is the most crucial step of the data analysis process wherein any data anomalies (like missing values or detecting outliers) with the data have to be modeled in the right direction.

Data Modelling

The modeling step begins once the data has been prepared. Modeling is an iterative process wherein the model is run repeatedly for improvements. Data modeling ensures that the best possible result is found for a given business problem.

Validation

In this step, the model provided by the client and the model developed by the data analyst are validated against each other to find out if the developed model will meet the business requirements.

Implementation of the Model and Tracking

This is the final step of the data analysis process wherein the model is implemented in production and is tested for accuracy and efficiency.

2. What is the difference between Data Mining and Data Profiling?

Data Profiling also referred to as Data Archeology is the process of assessing the data values in a given dataset for uniqueness, consistency, and logic. Data profiling cannot identify any incorrect or inaccurate data but can detect only business rules violations or anomalies. The main purpose of data profiling is to find out if the existing data can be used for various other purposes.

Data Mining refers to the analysis of datasets to find relationships that have not been discovered earlier. It focusses on sequenced discoveries or identifying dependencies, bulk analysis, finding various types of attributes, etc.

3. How often should you retrain a data model?

A good data analyst is the one who understands how changing business dynamics will affect the efficiency of a predictive model. You must be a valuable consultant who can use analytical skills and business acumen to find the root cause of business problems.

The best way to answer this question would be to say that you would work with the client to define a time period in advance. However, I would refresh or retrain a model when the company enters a new market, consummate an acquisition or is facing emerging competition. As a data analyst, I would retrain the model as quickly as possible to adjust with the changing behavior of customers or change in market conditions.

4. What is data cleansing? Mention few best practices that you have followed while data cleansing.

From a given dataset for analysis, it is extremely important to sort the information required for data analysis. Data cleaning is a crucial step in the analysis process wherein data is inspected to find any anomalies, remove repetitive data, eliminate any incorrect information, etc. Data cleansing does not involve deleting any existing information from the database, it just enhances the quality of data so that it can be used for analysis.

Some of the best practices for data cleansing include –

- Developing a data quality plan to identify where maximum data quality errors occur so that you can assess the root cause and design the plan according to that.
- Follow a standard process of verifying the important data before it is entered into the database.
- Identify any duplicates and validate the accuracy of the data as this will save lot of time during analysis.
- Tracking all the cleaning operations performed on the data is very important so that you repeat or remove any operations as necessary.

5. How will you handle the QA process when developing a predictive model to forecast customer churn?

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 just be a one-and-done model.

The best way to answer this question would be to say that you would first partition the data into 3 different sets Training, Testing and Validation. You would then show the results of the validation set to the business owner by eliminating biases from the first 2 sets. The input from the business owner or the client will give you an idea on whether you model predicts customer churn with accuracy and provides desired results.

6. Mention some common problems that data analysts encounter during analysis.

- Having a poor formatted data file. For instance, having CSV data with un-escaped newlines and commas in columns.
- Having inconsistent and incomplete data can be frustrating.
- Common Misspelling and Duplicate entries are a common data quality problem that most of the data analysts face.
- Having different value representations and misclassified data.

7. What are the important steps in the 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. Such values need to be examined and should be handled.

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

8. How will you create a classification to identify key customer trends in unstructured data?

A model does not hold any value if it cannot produce actionable results, an experienced data analyst will have a varying strategy based on the type of data being analyzed. For example, if a customer complains was retweeted then should that data be included or not. Also, any sensitive data of the customer needs to be protected, so it is also advisable to consult with the stakeholder to ensure that you are following all the compliance regulations of the organization and disclosure laws if any.

You can answer this question by stating that you would first consult with the stakeholder of the business to understand the objective of classifying this data. Then, you would use an iterative process by pulling new data samples and modifying the model accordingly and evaluating it for accuracy. You can mention that you would follow a basic process of mapping the data, creating an algorithm, mining the data, visualizing it and so on. However, you would accomplish this in multiple segments by considering the feedback from stakeholders to ensure that you develop an enriching model that can produce actionable results.

9. What are the criteria to say whether a developed data model is good or not?

- The developed model should have predictable performance.
- A good data model can adapt easily to any changes in business requirements.
- Any major data changes in a good data model should be scalable.
- A good data model is one that can be easily consumed for actionable results.

10. According to you what are the qualities/skills that a data analyst must possess to be successful in this position?

Problem Solving and Analytical thinking are the two important skills to be successful as a data analyst. One needs to skilled ar formatting data so that the gleaned information is available in an easy-to-read manner. Not to forget technical proficiency is of significant importance. You can also talk about other skills that the interviewer expects in an ideal candidate for the job position based on the given job description.