Data Scientist (Analyst) interview questions

This **Data Scientist (Analyst)** interview profile offers a sample of suitable interview questions for data analysts. Feel free to modify these interview questions for candidates to fit the specific requirements and needs of your data science team.

Data Scientist Analysis Interview Questions

Successful data scientists, managers and analysts excel at deriving actionable insights from the data that an organization generates. They have a good sense of what data they need to collect and have a solid process for carrying out effective data analyses and building predictive models.

The **data scientist role that is focused on data analysis** requires candidates with a very strong foundation in topics such as statistics, operations research and machine learning as well database skills such as SQL in order to retrieve, clean and process data from a variety of sources. Several pathways can lead to this role so candidates could approach the data analysis interview questions from a mathematics or statistics background although many will come from computer science or engineering.

This type of data scientist will often program in a scripting language such as R, Python or MATLAB and the role will typically not place emphasis on the programming languages, practices and general software engineering skills necessary for working with production quality software. There may be a need to modify some questions to more quantitative, statistical analysis interview questions. This type of role often incorporates the need to present the findings of an analysis. Consequently, information visualization skills, such as knowledge of Tableau or D3.js, as well as being a good communicator can be highly valuable.

Operational questions

Data Analytics Interview Questions

- Describe the steps you follow when designing a data-driven model to tackle a business problem. An example might be to automatically classify customer support emails by topic or sentiment. Another might be to predict a company's employee churn.
- Describe different pre-processing steps that you might carry out on data before using them to train a model and state under what conditions they might be applied.
- What models would you characterize as simple models and which ones as complex? What are the relative strengths and weaknesses of choosing a more complex model over a

simpler one?
- In what ways can models be combined to form model ensembles and what are some advantages of doing this?
- What is dimensionality reduction? What are some ways to perform this? When and why might we want to do this?

Role-specific questions

(Basic ideas in statistics, probability and machine learning)

- What is a confidence interval and why is it useful?
- What is the difference between statistical independence and correlation?
- What is conditional probability? What is Bayes' Theorem? Why is it useful in practice?
- Suppose we are training a model using a particular optimization procedure such as stochastic gradient descent. How do we know if we are converging to a solution? If a training procedure converges will it always result in the best possible solution?
- How do we know if we have collected enough data to train a model?
- Explain why we have training, test and validation data sets and how they are used effectively?
- What is clustering? Give an example algorithm that performs clustering. How can we know whether we obtained decent clusters? How might we estimate a good number of clusters to use with our data?
- We often say that correlation does not imply causation. What does this mean?
- What is the difference between unsupervised and supervised learning?
- What is the difference between regression and classification?
- What do we mean when we talk about the bias-variance tradeoff in statistical models?
- What is over-fitting? How is this related to the bias-variance trade-off? What is regularization? Give some examples of regularization in models.
- Suppose we want to train a binary classifier and one class is very rare. Give an example of such a problem. How should we train this model? What metrics should we use to measure performance?
- How many unique subsets of n different objects can we make?
- How would you build a data-driven recommender system? What are the limitations of this approach?

(Tools, visualization and presentation)

- In which environment(s) do you usually run your analyses?
- Describe your experience in working with data from databases. Are you familiar with SQL?
- What visualization tools (Tableau, D3.js, R and so on) have you used?
- Do you have a presentation you can show us, such as on SlideShare?
- Do you have experience presenting reports and findings directly to senior management in

your previous roles?
- Are you comfortable speaking in public? Have you ever presented a technical topic to a large audience?

For more data science questions that emphasize programming skills and deploying models in the real world, check out the interview questions for the data scientist (coding) role.