**Data Analyst**

Questions for Interview

Q1. What do data analysts do?

Or

* What is the process of data analysis?
* What steps do you take to solve a business problem?
* What is your process when you start a new project?

Q2. What is your process for cleaning data?

Consider mentioning how you handle:

* Missing data
* Duplicate data
* Data from different sources
* Structural errors
* Outliers

One can also ask:

* How do you deal with messy data?
* What is data cleaning?

Q3. What data analytics software are you familiar with?

* What data software have you used in the past?
* What data analytics software are you trained in?

Q4. Questions related to Excel:

* What is a VLOOKUP, and what are its limitations?
* What is a pivot table, and how do you make one?
* How do you find and remove duplicate data?
* What are INDEX and MATCH functions, and how do they work together?
* What’s the difference between a function and a formula?

Q5. Explain the terms:

* Normal Distribution
* Data Mining
* Data Wrangling
* Clustering
* Statistical Model
* Machine Learning

Q6. What is the difference between:

* Data mining vs. data profiling
* Quantitative vs. qualitative data
* Variance vs. covariance
* Univariate vs. bivariate vs. multivariate analysis
* Clustered vs. non-clustered index
* 1-sample T-test vs. 2-sample T-test in SQL
* Joining vs. blending in Tableau

Q7. What are the various steps involved in any analytics project?

Q8. What are the common problems that data analysts encounter during analysis?

Q9. Name the methods used for detecting outliers. What do they do?

Q10. Define “Collaborative Filtering”.

**Statistical Questions**

Q1. What is the significance of Exploratory Data Analysis (EDA)?

Q2. Explain descriptive, predictive, and prescriptive analytics.

Q3. What are the different types of sampling techniques used by data analysts?

Q4. Describe univariate, bivariate, and multivariate analysis.

Q5. How can you handle missing values in a dataset?

Q6. Explain the term Normal Distribution.

Q7. What is Time Series analysis?

Q8. How is Overfitting different from Underfitting?

Q9. How do you treat outliers in a dataset?

Q10. What are the different types of Hypothesis testing?

Q11. Explain the Type I and Type II errors in Statistics?

Q12. Name the statistical methods that are highly beneficial for data analysts?

Q13. What does it mean when the p-values are high and low?

Q14. Define and explain selection bias?

Q15. What is the null hypothesis?

Q16. When do we use T-Test?

Q17. How is normal distribution different from poisson distribution?

**EXCEL Questions**

Q1. Explain how VLOOKUP works in Excel?

Q2. What function would you use to get the current date and time in Excel?

Q3. How does the AND() function work in Excel?

Q4. Can you provide a dynamic range in “Data Source” for a Pivot table?

Q5. How to delete, insert, rename a column?

**SQL Questions**

Q1. How do you subset or filter data in SQL?

Q2. What is the difference between a WHERE clause and a HAVING clause in SQL?

Q3. How are Union, Intersect, and Except used in SQL?

Q4. What is a Subquery in SQL?

Q5. Using the product\_price table, write an SQL query to find the record with the fourth-highest market price.

Q6. How do you write a [stored procedure in SQL](https://www.simplilearn.com/tutorials/sql-tutorial/stored-procedure-in-sql)?

Q7. What does the Cursor function do?

Q8. Are values and cursors different from each other? Why?

**Tableau Questions**

Q1. How is joining different from blending in Tableau?

Q2. What do you understand about LOD in Tableau?

Q3. What are the different connection types in Tableau Software?

Q4. What are the different joins that Tableau provides?

Q5. What is the difference between Treemaps and Heat Maps in Tableau?

**Python Questions**

Q1. What is the correct syntax for the reshape() function in NumPy?

Q2. What are the different ways to create a data frame in Pandas?

Q3. How will you select the Department and Age columns from an Employee data frame?

Q4. Suppose there is an array that has values [0,1,2,3,4,5,6,7,8,9]. How will you display the following values from the array - [1,3,5,7,9]?

Q5. How can you add a column to a Pandas Dataframe?

Q6. How will you print four random integers between 1 and 15 using NumPy?

Q7. Using the below Pandas data frame, find the company with the highest average sales. Derive the summary statistics for the sales column and transpose the statistics.

Q8. Explain the difference between R-Squared and Adjusted R-Squared.

Q9. How do you save filename in Python?

Q10. What is the difference between List, Tuple, Sets and Array?

Q11. How to import Numpy? Pandas? Seaborn?

Q12. How to see the top columns and rows of a data set in pandas?

Q13. What is Numpy used for?

Q14. How to see the last few rows and columns of the data set in pandas?

Q15. What command can you write to change the value of a parameter?

Q16. What does the Join function do?

**Machine Learning Questions**

Q1. What is “Clustering?” Name the properties of clustering algorithms.

Q2. What is the K-mean Algorithm?

Q3. Define “Collaborative Filtering”.

Q4. Name the statistical methods that are highly beneficial for data analysts?

Q5. What is Time Series Analysis?

Q6. What are the differences between supervised and unsupervised learning?

Q7. How is logistic regression done?

Q8. List down the conditions for Overfitting and Underfitting.

Q9. Differentiate between the long and wide format data.

Q10. What are Eigenvectors and Eigenvalues?

Q11. What is the Confusion Matrix?

Q12. What is logistic regression? State an example where you have recently used logistic regression.

Q13. What is Linear Regression? What are some of the major drawbacks of the linear model?

Q14. What is a random forest? Explain it’s working.

Q15. What is deep learning? What is the difference between deep learning and machine learning?

Q16. What is a Gradient and Gradient Descent?

Q17. How are the time series problems different from other regression problems?

Q18. What are RMSE and MSE in a linear regression model?

Q19. What are Support Vectors in SVM (Support Vector Machine)?

Q20. What does Neural Network

Q21. What is the percentage split of data before we forecast them?

Q22. What are the different methods of studying the data for forecasting them?

Q23. How does the decision tree work?

Q24. What is the difference between Decision Tree and Random Forest?

Q25. What are the different Kernel functions?