Q1: Can you describe what a typical workday looks like for a Data Science Analyst?

A1: A typical workday can vary depending on projects and deadlines. However, a significant part of my day involves querying, cleaning, exploring, and modeling data. I also spend time meeting with stakeholders to discuss findings or to get more understanding of the business context.

Q2: Can you explain how you would handle missing data?

A2: Missing data can be handled in several ways. One could use deletion, where missing data is simply removed. However, this is unlikely if the missing data is not random. You could also use imputation, filling in the missing value based on other data. Lastly, you could use prediction models to fill in missing data.

Q3: Please describe a project where you had to use machine learning.

A3: In one project, I had to predict customer churn for a telecom company. I used a logistic regression model, random forest, and gradient boosting to model the customer churn. In the end, gradient boosting performed best in terms of accuracy and F1 score.

Q4: Can you explain what Data ETL is?

A4: ETL stands for Extract, Transform and Load. It's a crucial concept in data science and it refers to the process of extracting data from varied sources, transforming the data into a format that can be easily analyzed, and then loading it into a different database or data warehouse.

Q5: Can you describe a time when you used data visualization to explain a result to non-technical stakeholders?

A5: In one project, I had to explain to stakeholders why a new advertising campaign was not leading to increased sales. I used a time-series chart to show that while we had increased expenditure on ads, sales remained static, and their correlation was almost zero, which made it easy for the

stakeholders to grasp the concept.

Q6: What are some of the steps you would take in analyzing a new dataset?

A6: Before running any algorithms, I would start by getting an overall understanding of what information is available in the dataset. This includes knowing the total number of records, types of variables, doing a univariate analysis, checking missing values, and checking for outliers. Then, I would visualize the data and finally set some hypotheses that I would test while modeling the data.

Q7: How do you handle unstructured data?

A7: Unstructured data can be processed using various Natural Language Processing (NLP) techniques. For instance, tokenization, stemming, or lemmatization could be used to dissect unstructured text. Also, techniques like TF-IDF, Word2Vec, or BERT could be used to transform that text into numerical data that can be used for modeling.

Q8: What is the significance of a p-value?

A8: In hypothesis testing, the p-value is used to ascertain the significance of results. A smaller p-value suggests that the null hypothesis can be rejected, indicating that the findings are statistically significant.

Q9: How do you prevent overfitting in a model?

A9: Overfitting can be mitigated by various methods. This includes cross-validation, where the dataset is partitioned into a training set and validation set. This way, the model can be trained on one set of data and validated on another. Other methods include regularization, pruning, and ensembling.

Q10: Can you explain the difference between a Random Forest and Gradient Boosting?

A10: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes or mean prediction of individual trees. Gradient Boosting, on the other hand, works by building trees sequentially, where each subsequent tree aims to minimize the errors of the previous tree, thus improving the model's prediction incrementally.