1. **Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. What algorithms can be used to automatically select the most important features (regression, etc..)? Describe at least 3?**

Filter methods: These methods use a simple statistical test to identify features that strongly relate to the target variable. Examples include the chi-squared test, mutual information, and correlation coefficient.

Wrapper methods: These methods use the performance of a model to identify the most important features. Examples include Recursive Feature Elimination (RFE) and forward and backward feature selection.

Embedded methods: These methods are algorithms with built-in feature selection as a part of their structure. Examples include LASSO, Ridge Regression, and Random Forest Importance.

Hybrid methods combine the above-mentioned methods to find the most relevant features.

Each method has its advantages and disadvantages, and the best one to use depends on the specific problem, the data, and the model's intended use.

1. **Explain data leakage and overfitting (define each)?  
   Explain the effect of data leakage and overfitting on the performance of an ML model.**

Data leakage is when information outside of the training data is used to build the model. This can lead to a model that looks more accurate than it is because it was trained on data that will only be available when it's used in the real world. The information leakage could have been caused by a number of things, such as test data being mistakenly included in training or data that shouldn't have been seen during training but was available when making a prediction.

Overfitting is when a model has been trained too well on the training data, so it doesn't do well on new data it hasn't seen before. The model learned the noise in the training data instead of the underlying pattern. As a result, it could do better when applied to new data. Effect on performance: Leaking data can make a model look better than it is by making its performance look better than it is.

Overfitting can make a model work well with the training data but badly with new data it has never seen before. Because of this, models that have been overfitted won't work well with new data and probably won't do well in production.

Both Data leakage and Overfitting can make a model look more accurate than it is and stop it from generalizing to new data it hasn't seen before. While making a model, it's very important to be aware of these problems and take steps to avoid them.

1. **Explain what our outliers in your data?  
   Explain at least two methods to deal/treat outliers in your data?**

Outliers in data are observations that are significantly different from the rest. Data points differ greatly from the remainder of the dataset's observations. They may be the result of measuring or recording errors or signify an unexpected occurrence. Outliers can have a substantial impact on a dataset's statistical qualities as well as any subsequent modeling or analysis.

Outliers in data can be dealt with in two ways:

Data cleaning entails identifying outliers and deleting them or replacing them with a more realistic value. If an outlier is produced by a measurement error, for example, it can be deleted from the dataset. If an exceptional event caused the outlier, it could be replaced with a more realistic figure.

Data Transformation: This is applying a mathematical function to data to make it more symmetric, such as log transformation, which can lessen the impact of outliers. This strategy is especially beneficial when outliers are caused by skewed data distribution. It is also less harmful to the data than deleting outliers.

1. **What is feature scaling and why is it important to our model?  
   Explain the different between Normalization and Standardization?**

Feature scaling is a strategy for standardizing the range of independent variables or data features. Scaling features before training a model is critical because features on different scales can make it difficult for the model to converge. Normalization and standardization are the two basic strategies for feature scaling. Scaling a variable to have values between 0 and 1 is known as normalization. It is accomplished by subtracting the feature's minimum value and dividing the result by the range (i.e., max-min).

Standardization is the process of changing a variable into one with a mean of 0 and a standard deviation of 1. It is accomplished by removing the feature's mean and dividing the result by its standard deviation.

Normalization and standardization are both important in data preprocessing. Normalization comes in handy when you have a skewed distribution and want to scale a feature between 0 and 1. When translating data to a conventional normal distribution, standardization is useful. Standardization should not be utilized if your data contains outliers because it is sensitive to outliers. Outliers are more resistant to normalization.