Q1. 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?

A1. As per the sklearn documentation to select the most important features from model - We can Use the decision tree classifier, Linear Regression.

Techniques used are -

1. Generic Univariate Select - Allows you to select features from dataset by a scoring function
2. Variance Threshold - It removes features whose variance does not meet some threshold.
3. Select K Best - The k best highest scoring features
4. Select Percentile - Based on user specified percentile

The chi2 test and r_regression method are useful to identify the important features from sklearn

The Model based are - as suggested by sklearn documentation -

1. linear_model.lasso or trees.DecisionTreesRegressor for regression
2. linear_model.LogisticRegression or svm.LinearSVC or trees.DecisionTreesClassifier for classifier

https://scikit-learn.org/stable/modules/feature_selection.html#feature-selection-using-selectfrommodel

Q2. Explain data leakage and overfitting (define each)? Explain the effect of data leakage and overfitting on the performance of an ML model

A2. Data leakage refers to a mistake that is made when sharing of information between test and training data sets occurs. When splitting a data set into test and training sets, the goal is to ensure no data is shared between these two. Data leakage occurs when the data used in the training process contains information about what the model is trying to predict, hence it leads to overfitting, hence an optimistic model which performs too well on test data also.

Q3. Explain what our outliers in your data? Explain at least two methods to deal/treat outliers in your data?

A3. Outlier is the data point which lies too far or at abnormal distance from the majority of points. Usually a box plot or distribution of data can reveal the outlier in data which fall beyond the standard deviation of 3 times. It is beyond the Q3 + 1.5 IQR.

First is to remove outlier but it result in loss of data and not a good practice. Second is to cap the dataset based on percentile like 10th as minimum and 90th as maximum to identify outlier and replace them with floor values for the 10th percentile and maximum values are replaced with 90th percentile. Third - replace the outlier with median values.

Q4. What is feature scaling and why is it important to our model? Explain the different between Normalization and Standardization?

A4. Feature scaling is a technique to standardize the independent features in the data.It is important as so some features might be too variable to ranges and hence our model will behave differently when features with invariably large range are introduced.
Normalization is a technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. Standardization - Values are centered around the mean (0) and standard deviation (1)