Question 3

1. Conduct exploratory data analysis to identify crucial features that will be utilized in the model:

   To identify the crucial features for building a machine learning model for forecasting the approximate value of used cars, we can perform the following steps:

- Check the correlation between features and target variable (price) using a correlation matrix or heatmap.
- Analyze the distribution of the target variable (price) to identify any skewness that might need to be addressed.
- Check the distribution of the independent variables to identify any outliers, missing values, or skewness.
- Check for multicollinearity between the independent variables to avoid using redundant features in the model.
- Perform feature engineering to extract new features that could be relevant for the prediction task.

2. Please justify the selection of these features and aim to incorporate as many as possible: Based on the exploratory data analysis, some of the features that can be considered for the machine learning model are:

- brand: Some car brands are known to hold their value better than others, so including this feature could be relevant.
- model: The specific model of a car can also affect its resale value, so this feature should be included.
- vehicleType: The type of vehicle can be a significant factor in determining its value. For example, SUVs may hold their value better than sedans.
- yearOfRegistration: The age of the vehicle is a critical factor in determining its value.
- fuelType: The type of fuel can affect the value of a vehicle, as some fuel types may be more expensive to maintain than others.
- gearbox: The type of gearbox can also affect the value of a car, as automatic gearboxes are typically more expensive than manual ones.
- powerPS: The power of the vehicle's engine can also be a crucial factor in determining its value.
- kilometer: The distance the vehicle has been driven can significantly impact its value.

3. Kindly identify any potential challenges or limitations you anticipate/encounter during the feature selection process (if any): Some potential challenges or limitations that may be encountered during the feature selection process are:

- The presence of missing values or outliers in the data, which could skew the results of the analysis.
- The need to balance the number of features included in the model with the risk of overfitting, which could lead to poor performance on new data.
- The difficulty in determining which features are most relevant for the prediction task, as some features may be highly correlated with others or provide limited predictive power.

4. (Bonus) Try to propose a good model you feel would be able to best fit the features you have selected to make predictions: Based on the features selected, a regression model could be a good fit for this prediction task. Some potential models that could be used include:

- Linear regression: A simple and interpretable model that can capture linear relationships between features and the target variable.
- Random Forest regression: A more complex model that can capture nonlinear relationships between features and the target variable, as well as handle missing values and outliers in the data.
- XGBoost regression: A popular machine learning model that can handle missing values, outliers, and nonlinear relationships between features and the target variable, as well as provide feature importance scores to aid in the feature selection process.