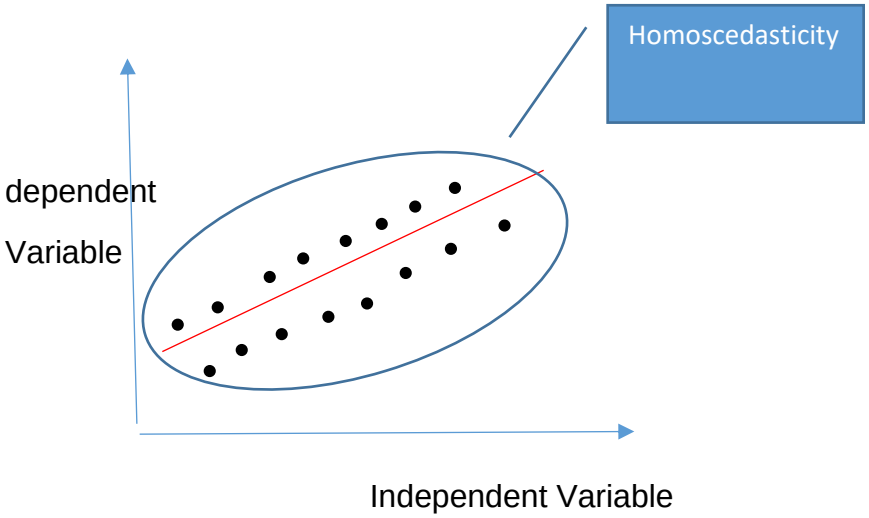
Question-1:

List down at least three main assumptions of linear regression and explain them in your own words. To explain an assumption, take an example or a specific use case to show why the assumption makes sense.

The three main assumptions of linear regression are:

Assumption 1 : Homoscedasticity

Let us take an example of the liner regression assignment itself where we are trying to predict the car price (dependent variable) using the multiple independent variable. This assumption means that the variance around the regression line (0) is the same for all values of the independent variable. However, we consider this assumption but there is hardly a dataset that fits this assumption. The general rule is If the ratio of the largest variance to the smallest variance is 1.5 or below, the data is homoscedastic. We can test homoscedastic using the Bartlett's test for homogeneity of variances



Assumption 2 : Linear Relation

Linear regression assumes that the relationship between the independent and dependent variables to be linear. If the relationship between input and output is not linear, directly fitting the model to estimate or predict the outcome based on the input will give results that indicate no relationship, or a weaker relationship than there really is. A linear relationship is one where increasing or decreasing one variable X times will

cause a corresponding increase or decrease of X times in the other variable too.
Example : For a given object, if the volume of the object is doubled, its weight will also double. This is a linear relationship. If the volume is increased 10 times, the weight will also increase by the same factor.

Linear relations are best plotted on scatter plots.

$Y = a + (\beta 1 * X1) + (\beta 2 * X2) + (\beta 3 * X3) + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\beta n * Xn)$

Assumption 3 : There is no perfect linear relationship between independent variables-No perfect multicollinearity

Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.
Multicollinearity is tested with Correlation matrix and Variance Inflation Factor (VIF).

Variance inflation factor is a metric computed for every *independent* variable that goes into a linear model. If the VIF of a variable is high, it means the information in that variable is already explained by other *independent* variables present in the given model, which means, more redundant is that variable. So, lower the VIF the better preferably below 2.

The variance inflation factor of the linear regression tells that with VIF > 10 there is an indication that multicollinearity may be present and with VIF > 100 there is certainly multicollinearity among the variables.

If multicollinearity is found in the data, the simplest way to address the issue is to remove independent variables with high VIF values.

Example of high VIF indicating multicollinearity from the linear regression assignment:

	Var	Vif
0	fueltype	inf
28	enginetype__rotor	inf
19	fuelsystem__idi	inf
1	enginelocation	inf
14	carcompany__subaru	inf
27	enginetype__ohcf	inf
26	cylindernumber__two	inf
9	compressionratio	100.830000
20	fuelsystem__mpfi	28.580000