Q1. At least 3 Main Assumptions of Linear Regression and explain ?

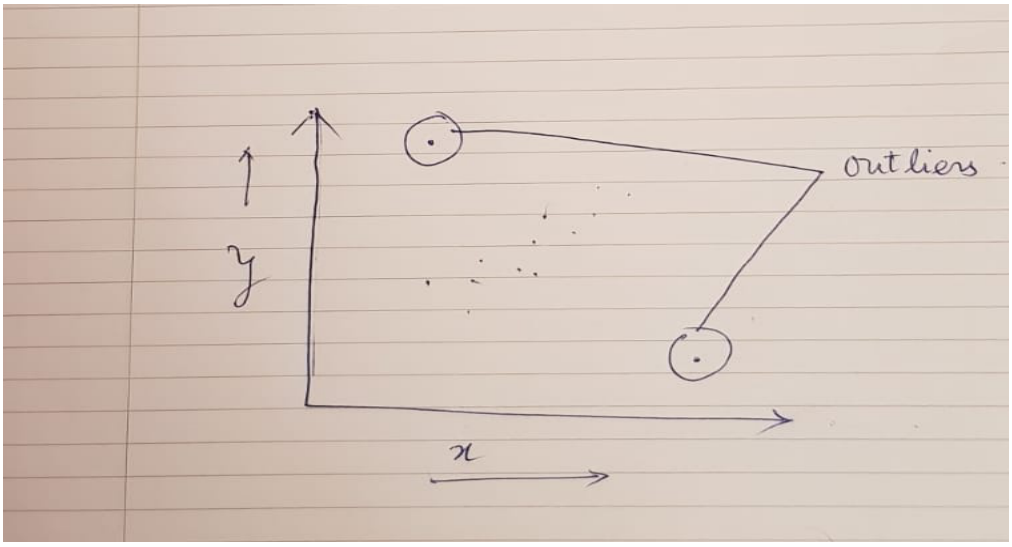
Ans 1: 3 main assumptions of Linear Regression explained below:

1. Outliers:

In a variable if there are observations or data which have values abnormal or extreme relative to all the other observation or values in that variable then these observations are called outliers.
For example: annual income contains 10 observation which range from 5-10 Lacs and if you add a new observation/data point 75 Lacs then this value is an outlier in the variable annual income.

In Linear regression the outliers can have a large influence on the line of the best fit. So, if there are independent variables which have one or more outliers and hence the best fit line will try to take these values in consideration which can adversaly affect the results of Linear Regression. Hence with dataset having outliers it becomes very difficult to predict the value of independent variable.

To check for outlier's scatter plot can be used as outliers will be positioned away from the normal data values which can be easily seen in the plot.
Below figure shows 2 outliers using a scatter plot:



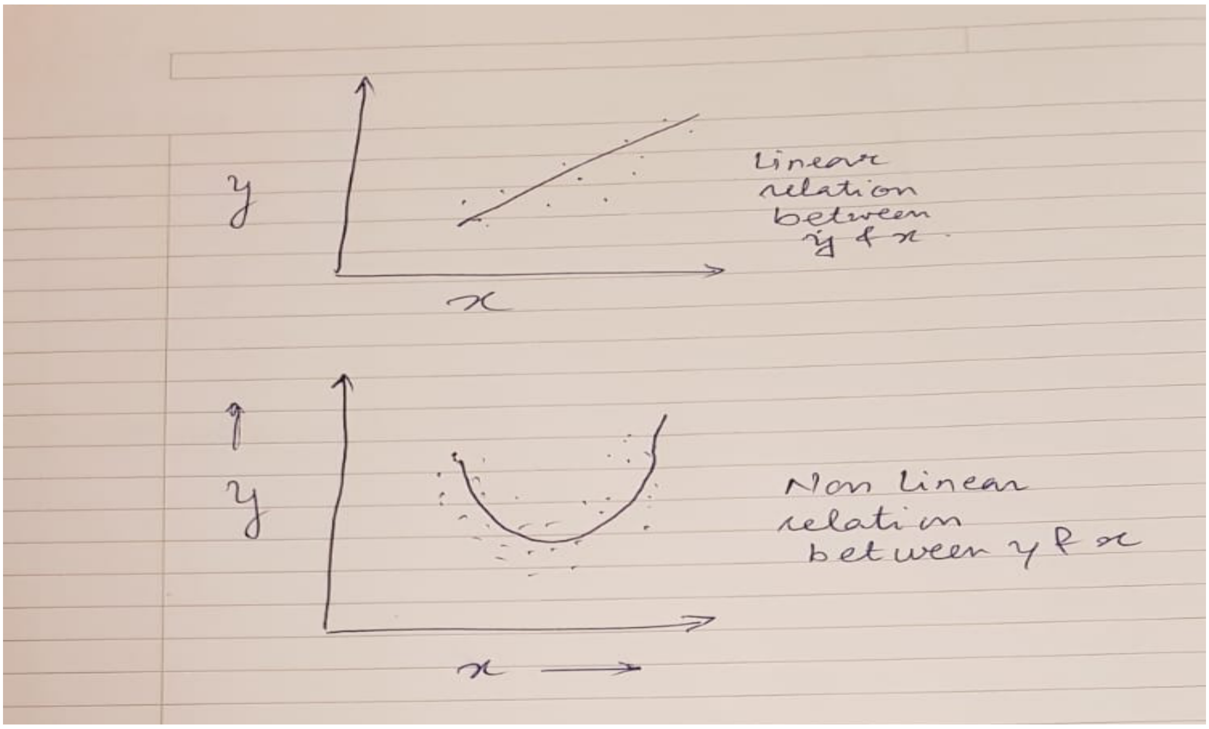
2. Linearity:

In regression model it is assumed that there is a linear relationship between the predictor or independent variables and the dependent variable.
This assumption is essential in order to predict the dependent variable for a specific value of the predictor or independent variable.
This relationship should satisfy the below equation:

Y = c + m1 * x1 + m2 *x2 …… mn * xn

Here n is the total number of independent variable, Y is the dependent variable and x1, x2 … xn are independent variables.

To check if linearity exists between say a dependent variable and independent variable draw a scatter plot between them and try to fit a straight line between the data points if you successfully able to fit a straight line then there is a linear relation between those 2 variables otherwise the relationship between the 2 variables is not Linear.
The below figure shows linear and non- relation between dependent and independent variable x and y


y
x
Linear relation between y & x.
y
x
Non linear relation between y & x

3. Multicollinearity:

Linear Regression assumes that there is no multicollinearity in the data. Multicollinearity occurs when there is high correlation between One or more Independent variables in the given data. In case if there is multicollinearity in the data set then it becomes difficult to estimate the exact relationship between the dependent variable and the independent /predictor variables.
So, for example if there are 2 independent variables x1 and x2 and y is the dependent variable then below will be the equation:

Y = C + a1 * x1 + a2 * x2

Where C is the intercept and a1 and a2 are the coefficients. Now if multicollinearity exists in the data i.e. x1 and x2 have a high correlation then it will also affect the coefficients a1 and a2 and hence will affect the prediction.
If there is a very high collinearity between 2 independent variables, then at times it would be wise to drop one of them (also keeping other evaluation matrix in mind) and we can still predict the Y.
For example, in the given Assignment -1 the Car mileage in city and Car mileage in Highways are highly correlated as the mileage is highly dependent on the internals of the car such as engine size, efficiency etc. and hence we see their correlation to be 0.99 so in this case we can either drop one of the variables and include the other in the Linear regression or we can also take an average of the 2 values.
To check for multicollinearity, we can use:

1. Scatter plot, Correlation matrix and heat map among the independent variables
2. VIF (Variance Inflation Factor) values less than 4 indicate no multicollinearity VIF values of 10 or more indicate high multicollinearity.

Q2 – Explain at least 3 model evaluation metrices?
Ans 2- Below are 3 model evaluation metrices explained along with Advantage and disadvantage:

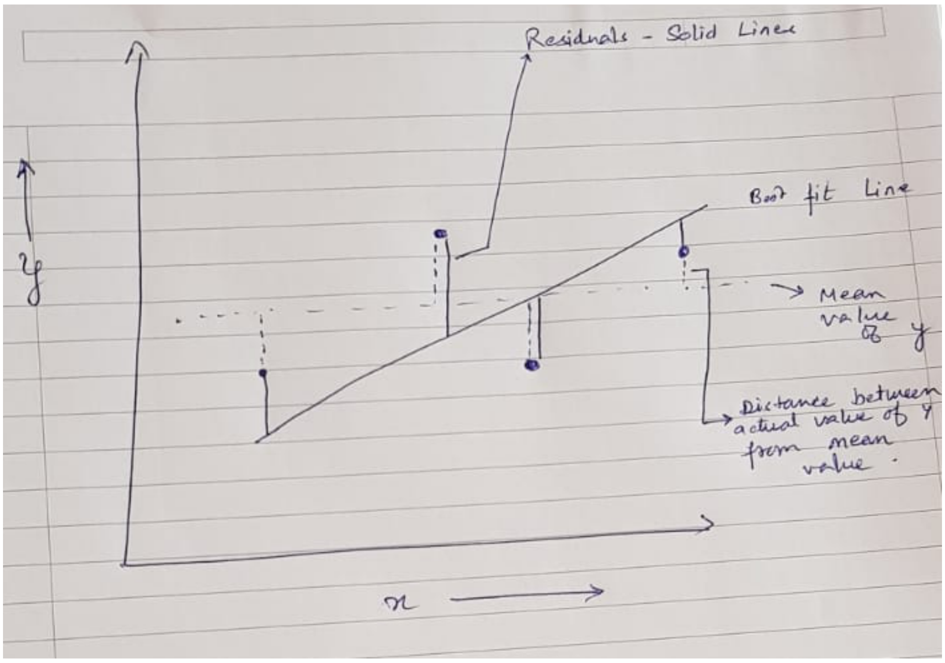
A.)R Squared: This R Squared metric is used to measure how close the actual data is to fitted regression line. This also known as coefficient of determination. So, R-squared is 1- fraction of unexplained variation divided by total variation in the predicted variable.

R Squared value lies between 0 and 1 with 0 indicating that the model explains none of the variance in the predicted variable Whereas 1 value means that the model explains all the variance in the predicted variable and the predicted values predicted by the model are equal to the actual values.

1. The final model has R Squared value of 0.889 which means the final model is able to explain 88.9% of the variance in the dependent value "price" of the car.
   Since the model has accounted for 88.9% of the variance so the predicted values should be close to the actual value

   R Squared = 1 – (RSS / TSS)

   Here RSS (Residual sum of squares) is the sum of square of residuals (distance between actual data point and the best fit line) whereas TSS (Total sum of squares) is the sum of square of distance of each data point from the mean of all the data points.



In general, better R Squared value means a better model.

Disadvantage: For every variable added in the model the R Squared value would increase.
Even when you add a random variable to model there will still be a slight increase even though adding the random variable does not add any value to the model.

B.) Adj. R Squared:

Adj. R Squared like R Squared measures the proportion of variation in the dependent variable which is explained by the model using the independent variables. But Adj. R Squared adjusts the value also based in the number of independent variable and the value it is adding to the model.
So, if you add a new variable which is random and does not add value to the model then the Adj. R Squared will penalize this model and hence causing the value of Adj. R Squared to be reduced then the previous model.

1.) In the final model lm_12 Adj. R Squared 0.884 where as that of R Squared 0.889.

The difference in both the value is because Adj. R Squared statistically adjusts the value based on the number of independent variables in the model and their significance whereas R Squared does not take significance of added independent variables in the model.
In general, the more non-significant variables you add into the model the gap between the R Squared and Adj. R Squared value increases.

Adj. R Squared = 1 – [(1 – R Squared) (n-1) / n-k-1]

Here n is the number of items in the data set and k is the number of independent variables in the data set whereas k is the number of independent variables in the model.
So, if you add a variable to the model which does not significantly add to the model that is the R Squared value remains almost similar then since k increases and causes decrease the value of denominator hence penalizing the model.

2.) Advantage of Adj. R Squared: The Adj. R Squared also statistically includes the significance of the added independent variables in the model and hence compared to R Squared it penalizes a model if you add random variables to a model or variables which do not add much value /significance to the model.

C.) VIF (Variance Inflation Factor):

VIF measure the collinearity among the independent variables so VIF metric provides a measure of correlation of 1 variable with multiple other independent variable.

1. In the final model lm_12 the VIF values of all the added variable is below 10.

If x1, x2..xn are the predictor variables in the model then VIF value of x1 is calculated by building multiple linear regression model with x1 as the target variable and all other independent variables (x2,x3.. xn) as the predictor and we calculate the R Squared value of this model R1 then VIF of this model can be calculated as

VIF(x1) = 1 / 1 – R Squared value of model R1.

So, in our final model VIF of al variable is less then 10 meaning R Squared value would be less than 0.9 because with 0.9 R Square value VIF value will be 1/0.1= 10 implying that it will be hard to predict any of the independent variables selected in the model with help the other independent variables included in the model.
Ideally the VIF value should be less than 3 but since in our model removing the variables with high VIF was degrading the model so decided to keep them.

2. Advantage of VIF : It is a simple and very useful metric which is used to measure collinearity or correlation among multiple variables and it is specially very useful in eliminating non-significant variables in a model using VIF and p value of the model in combination.