Assignment-based Subjective Questions

1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

>> 1.People are likely to rent bikes mostly in fall season

2.Bike rentals are more in Sep

3.Bikes are mostly rented on Wed, Thu and Sat

4.Bikes are mostly rented during clear weather

5.Bikes were rented more during 2019

6.Bikes are mostly rented during holidays

2) Why is it important to use drop_first=True during dummy variable creation?

>> A variable with n levels can be represented by n-1 dummy variables. So, even if we remove the first column we can represent the data. For example, there are 3 values for furnishing status – Furnished, Sem-Furnished and Unfurnished. If represented by 1,0,0 respectively. If the value for Furnished and Sem-Furnished is 0 and 0 then we can easily make out the value for Unfurnished is 1. Hence, the 3rd value can be known from the first 2 columns and thereby, creating efficiency

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

>> Temp variable

4) How did you validate the assumptions of Linear Regression after building the model on the training set?

>> Through residual distribution. Checking the normal distribution and mean value of 0

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

>> Temp, winter and Sep

General Subjective Questions

1) Explain the linear regression algorithm in detail.

>> Linear Regression is a statistical model that analyses the relationship between a dependent variable with the given set of independent variables.

Mathematically, the relationship can be represented by

Y=mX+c

Y-Dependent variable

X – Independent Variable

m-slope which shows the effect X has on Y

c-Constant known as Y-intercept

Linear relationship can be positive or negative

Linear regression is of 2 types-

a) Simple Linear Regression
b) Multiple Linear Regression

Assumptions made about dataset by Linear Regression model-

a) Multi-Collinearity – Independent variables do not have correlation amongst them
b) Relationship between independent and dependent is linear
c) Errors should be normally distributed

2) Explain the Anscombe's quartet in detail.

>> This was developed by statistician Francis Anscombe. It comprised four datasets, each containing eleven (x,y) pairs. They share some descriptive statistics. However, when graphed they are completely different. They show the same regression lines but each dataset behaves differently

3) What is Pearson's R?

>> Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposite, the correlation coefficient will be negative. The correlation coefficient can take values from +1 to -1. A value of 0 indicates no association between the variables. A value greater than 0 shows positive association and value lesser than 0 indicates negative association.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

>> Feature scaling is a technique to standardize the independent features in a fixed range. It is done during data pre-processing to handle highly varying magnitudes or values or units. If scaling is not done then machine learning model will tend to weigh greater values higher and consider smaller values as lower, regardless of the unit.

S.No.	Normalized	Standardized

1	Min and Max values are used for scaling	Mean and standard deviation is used
2	Used when features have different scales	Used to ensure zero mean and unit standard deviation
3	Scale values between [0,1] or [-1,1]	Not bounded to a certain range
4	Affected by outliers	Not affected

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

>> If there is a perfect correlation, then VIF = Infinity. A large value of VIF indicates there is correlation between the variables. In the case of perfect correlation R2 = 1 which leads to (1/1-R2) as infinity. Hence, we will have to drop on of the variables which is causing multicollinearity

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

>> The quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution

When there are 2 data samples, it is important to know if the assumption of a common distribution is justified. The q-q plot can provide more insight into the nature of the difference than analytical methods