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?

Ans- From the analysis I could find that from categorical columns mostly ‘season’,’ holiday’, ‘weekday’ and ‘month’ are highly corelated with dependent variable ‘cnt’.

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

Ans- It is important to use drop_first=True because it helps to reduce only 1 column in every categorical column while one hot encoding .

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

Ans- After looking that pair plot with numerical variable the highest correlated feature with target variable is : ‘temp’.

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

Ans- In order to validate the assumption of Linear Regression I checked with R2 score on both test and train , I also checked with MSE value for both test and train , went with RFE to find the best feature .

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

Ans- the top 3 feature that can be choosen are

A - ‘temp’ based on it we can find that in climate with higher temperature people do not try to use Bike sharing ,

B- ‘weekends’ based on weekends people generally use more than on weekdays

C- ‘season’ - based on season also people try to accept sharing bike like in spring season people want it but in summer season people do not like it .

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting, Linear Regression try to apply the Y=mx+c concept where Y is dependent variable or target or output variable and x I independent variable and c is the cosntant error needed to be there to fit properly.

2. Explain the Anscombe's quartet in detail.

Ans- it isa group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

3. What is Pearson's R?

Ans- In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

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

Ans- It is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

In regression, it is often recommended to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their mean.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Ans- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve the problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables .

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

Ans- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.