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?

The categorical variable has strong relationship with the dependent variable namely “CNT”. Below are few examples:

Categorical Variable	Relationship with dependent variable “CNT”
1. YR	2019 has more bike rentals than 2018
2. WEATHERSIT	clear to cloudy weather situation has more bike rentals among all the levels of weathersit. And misty to light rain and snow has lowest share of bike rental.
3. MNTH	From January , as the months has increased, the share of bike rental has increased till October and then again in November and December, it has dropped.
4. SEASON	Fall has the highest share in bike rental and then summer. Winter has lower and spring has lowest share of bike rentals
5. Holiday, workingday and weekday	Does not have strong relationship with dependent variable CNT.

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

Dummy variable creation is nothing but converting different levels of character variable to binary variables and thus converting them to numeric form. Out of n levels, the variable can be explained completely by creating n-1 variables. Creating n variables does not make sense and also while creating adjusted r2, having more number of variables is penalised and also will increase multi collinearity. This is why drop_first=True during dummy variable creation is important.\

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

Temp variable has the highest correlation with the target variable.

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

To validate the assumptions of Linear Regression after building the model on the training steps, below are the steps that needs to be followed:

- first perform a residual analysis of the error terms
- then make the predictions on the test set
- evaluate the model based on the predictions

Residual analysis helps to validate the assumption of the error being normally distributed which can be done by plotting a histogram.

Also the error term is homoscedastic or not is visualized by residual plot.

Pair-wise scatterplots i.e. pair-plots helped in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.

A heatmap for correlation helped in finding out multicollinearity and vif calculation helped to get rid of such variables.

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

The top 3 variables contributing significantly towards explaining the demand of the shared bikes are

- Yr
- Temp
- weathersit _misty_to_light_rain_snow

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is nothing but the linear relationship of few variables with a given variable. The objective is to estimate or predict one variable with the combination of variable which are linearly related with the dependent variable. This is regression technique which is supervised in nature and finds out the linear relationship between dependent and independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties

3. What is Pearson's R?

Pearson's R is nothing but Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is a measure of the strength of a linear association between two variables.

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

Scaling is nothing but bringing different variables in the same unit of measurement.

Scaling is performed so that while explaining the variables , interpretability is better and explanation makes more sense.

Normalised scaling is minmax scaling i.e. ((var-min)/(max-min)) and it lies between 0 and 1.

Standardised scaling is ((var-mean)/standard deviation) and though it reduces the variability, but it does not brings the value between 0 and 1. It varies beyond 0 and 1. Also, having outlier in the data affects this type of standardization.

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

If there is perfect correlation, then VIF takes the value 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.

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

QQ plot is used to see if the points lie approximately on the line. If they don't, it means, residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of QQ plots:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The qq plot can provide more insight into the nature of the difference than analytical methods.