Assignment-based Subjective

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

Categorical variable's: "month" & "season" are having HIGH corelation.
Categorical variable's: "weekday" & "workingday" using HUE as "holiday"

- We can infer that that on a week day if its a holiday.
- especially on MONDAY:1 and FRIDAY:6 the IQR on count of Vehicles are more.

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

- When categorical variables are n-levels . Then we need to create n-1 variables to represent the variable using dummy variables.
- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

- TEMP and ATEMP has the highest corelation with the target variable "cnt".

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

- Check the RESIDUAL error terms are NORMALLY distributed for NORMALITY.
- Check the CONSTANT VARIANCE between feature's and the redisuals for HOMOSCADCITY. We should not be able to infer anything.
- Absence of multi collinearity by eliminating the features having high VIF values.Check the final data for VIF score to be seen with all features having <5..

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

- Based on the T and P_Value: Season/weathersit/temp these are the three features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm explains how independendant the factors/variabes are and how well they are reletaed.
If the they are linearly related means . They have relation in predicting the outut variable.

Linear regression algos are used in PREDICTIVE/PRESCRIPTIVE analysis.
There are two types of Regression models:

1. Simple linear regression model and
2. Multiple Linear regression model.

Using SLR model explains how good (Godness of line) we can predict. Means the variance on the target/output variable.
Here it will be single Independent variable.

We can clear all if there are any violations in the REDISUAL analysis.

1. Normalilty
2. Homoscadicity
3. Absence of Multicolinrearity

We first understand the data.
Create train and test data set as 70% and 30%
Build/train the linear model using training model.
Predict or evaluate the model usingt TEST data.

Under MLR its about overfit/using VIF removing the redundant variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a 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.

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.

3. What is Pearson's R?

Correlation measures the strength of association between two variables as well as the direction. There are mainly **three types** of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

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

Any data before passing as an parameter it should undergo a process of Standardising or Normalizing.

Standardizing :

Its used most widely in most of the models.

Its spreads data points between -3 to 4 or like that

Normalising/Scaling:

Its used less like in IMAGE PROCESSING for PIXEL intensities to be fix in 0 and 1

Scaling removes outliers yb default as its gett8ng pkaced between 0 to 1.

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

If there is a perfect corelation between independent variables then the VIF = INFINITY. If multiple or more number of variables are corelated then the VIF will be more.
VIF of more than 5 is considered to be redundant and should be dropped.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.