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- The categorical features such as season and weathersit affect the dependent variable that is cnt. As the season is summer, winter and fall there are increase in bike rentals but in season spring there are decrease in the bike rentals. In the clear and normal cloudy weather the bike rentals are more as compare to foggy or partly snow and rain weather.

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 reducing extra column created for dummy variable and hence it reduces co-relations created by the dummy variables.

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

Ans- The “registered” variable is having highest correlation with the target variable but it is not counted because target variable is derived from the registered variable. The highest correlation is having the variable named “temp”.

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

Ans- firstly I checked the statistical data in which the probability of f-statistic should be very less, p-value should be less than 0.05 of all the variables, the R^2 value should be higher than 0.70 and the adjusted R^2 value should be in range with R^2 value, multi-collinearity should be less or null between the dependent and independent variable.

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

Ans- 1. Yr 2. Temp 3. winter (in season)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans- Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of the data based on some variables. In linear regression name means that the two variables which are the x-axis and y-axis should be linearly correlated.
Mathematically the formula for the linear regression equation is $y = mx + c$
Where, m and c are given by
m(slope) = n **Σ** xy – (**Σ**x)(**Σ** y) / n**Σ**x^2-(**Σ** x^2)
c(intercept) = = n **Σ** y – b(**Σ** x) / n
x and y are two variables on the regression line.
In this, m = slope of the line
c = y-intercept of the line
x = Independent variable from dataset
y = Dependant variable from dataset

2. Explain the Anscombe's quartet in detail.

Ans- 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.
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features mut be plotted in order to see the samples that can help you identify the various anomalies present in data like outliers, diversity of the data, linear separability of the data etc. Also linear relationships can only be considered a fit for the data with linear relationships and is capable of handling any other kind of datasets.

3. What is Pearson's R?

Ans- Pearson's R is 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 opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

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

Ans- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalized Scaling	Standardized Scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard devaition
Scales values between 0,1 or -1,1	It is bounded to a certain range
It is affected by outliers	It is much less affected by outliers

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 this 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 (which show an infinite VIF as well).

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.

QQ plot is a graphical tool to help us assess if a set of data came from some theoretical distribution such as normal, exponential or uniform distribution. Also, it helps us to determine if the two data sets come from populations with a common distribution.