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: Based on analysis we can easily see that count variable (dependent variable) is inferences by year, season, month and weather situation. We can see the increase of count from 2018 to 2019 year.

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

Ans: Drop_first=True helps to avoid the "dummy variable trap," which occurs when all categories are included in dummy variables for categorical characteristics and might cause multicollinearity. We ensure that the categories are independent of one another by eliminating the first category, which increases the stability of the model.

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

Ans: atemp/temp has highest correlation with the target variable.

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

Ans: Validate the assumption by using Q-Q plot graph, checking VIF values (should be below 5), by comparing train and test prediction and plotting them on graph for batter visualization. We should also Check for constant variance of residuals (Homoscedasticity).

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

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- temp/atemp
- year
- season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is Supervised learning algorithm that predicts a continues output (numerical) variable based on one or more input feature. Where one input feature is called as simple linear regression and more input feature is called as multiple linear regression.

Linear regression works as finding the best-fit linear line that minimize the sum of squared errors between predicted and actual values. The line is represented by an equation

$$y = mx + c$$

m = slop of line

c = intercept

x = independent variable

y = dependent variable

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four small datasets that have similar statistical properties but different distinct distribution and it look very different when presented graphically. There will be situations when based on the statistical properties it's too early to make any decision. When we present data visually using different graphs it will give better understanding. This highlights the importance of visualizing data beyond just statistical summaries.

The main purpose of Anscombe's quarter is to first visualize the dataset before describing statistical properties. Anscombe's quartet consists of four datasets, each containing eleven x-y pairs of data. Plotting each dataset appears to show a different relationship between x and y, with varied correlation strengths and variability patterns.

3. What is Pearson's R?

Ans: Pearson's R, is a statistical measure of the linear relationship between two continuous variables. The scale goes from 1 perfect positive correlation to -1 denoting perfect negative correlation where in 0 indicating no correlation

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

Ans: Scaling is a process which is performed as pre-processing during building a linear regression model which transforms input features to a common range to prevent features with large ranges from dominating the model.

If we don't perform the scaling then model prediction will be impacted by features whose values are high. Scaling technique ensures that all input features are on a similar scale, preventing larger-magnitude features from dominating

There are two types:

- Normalized scaling (Min-Max Scaler): Scales values to [0, 1] range.

- Standardized scaling (Standard Scaler): Scales values to have mean = 0 and standard deviation = 1

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

Ans: Multicollinearity between features is measured by VIF. When two or more characteristics have perfect correlation, the model becomes unstable and there is an infinite VIF. If you miss the feature variable scaling and values are very high for specific variable, you may see a higher number difference in VIF as it is calculated by $1/1\text{-}R^2$. If there's perfect multicollinearity, R^2 becomes 1, resulting in a denominator of zero. Any number divided by zero equals infinity.

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

A Quantile-Quantile plot (Q-Q plot) is a graphical tool used to compare the distribution of two datasets or a dataset to a theoretical distribution. It's commonly used in statistical analysis and linear regression to:

1. Check for normality of residuals: Verify if the residuals follow a normal distribution, which is an assumption of linear regression.

2. Identify outliers: Detect data points that deviate significantly from the expected pattern.

3. Evaluate model fit: Compare the distribution of actual values to the predicted values from a model.