Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)**

   After analyzing bar plots and checking the value counts for each categorical variable, it's clear that almost all of them significantly affect the "cnt" variable (dependent variable). These features consistently show strong trends, with a substantial portion of the dataset influenced by their variations. As a result, these categorical variables appear to be valuable predictors and can be considered for inclusion in the model building.

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

   If we skip using drop_first=True, we keep the first column in the dummy variables table. This can cause collinearity issues in the model, making it less reliable. Using drop_first=True is important to avoid this problem and ensure a more meaningful and accurate model.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)**

   When we look at the pairplot of numerical variables, we notice that temperature (temp) has the strongest correlation with the target variable. The scatter plots for both temp and "feels-like" temperature (atemp) are similar, but temp shows a stronger correlation because its data points are less spread out compared to atemp. In simpler terms, temp has a clearer relationship with the target variable than atemp.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   Analyzing the dist plot and scatter plot of the residuals, we can make the following observations:

   Distribution Plot: The distribution of the residuals (error terms) is centered around zero, suggesting a normal distribution. This implies that the error terms follow a normal pattern.

   Scatter Plot:
   - No discernible pattern is visible in the scatter plot of the error terms, indicating that the errors are independent of each other.
   - The constant spread of points in the scatter plot implies that the error terms have consistent variance, confirming the presence of homoscedasticity.

   In summary, the error terms are normally distributed around zero, independent of each other, and exhibit constant variance, meeting the assumptions of a well-behaved regression model.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

According to the final model, the three most significant factors influencing the demand for shared bikes are determined by the size of their coefficients:

1. Temperature (temp)
2. Light Snow
3. Year (yr)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

The linear regression algorithm establishes a linear connection between a dependent variable (y) and one or more independent variables (x). This modeling technique enables us to understand how the value of the dependent variable changes based on the values of the independent variables.

Key purposes of linear regression include:

I. **Effect of Independent Variables:** Assessing the impact of independent variables (x) on the target or dependent variable (y).

II. **Change in the Target Variable:** Investigating how the target variable changes concerning one or more input variables.

III. **Trend Analysis:** Identifying upcoming or ongoing trends in the data.

The linear regression equation takes the form:

$$y=b0 +b1\ x + \text{random error}$$

The random errors represent everything that the model does not have into account because it would be extremely unlikely for a model to perfectly predict a variable, as it is impossible to control every possible condition that may interfere with the response variable. The errors may also include reading or measuring inaccuracies as well.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's Quartet comprises of 4 data set that have nearly identical and simple descriptive analysis yet has very different distributions and appear very different when graphed. They have quite different distributions and appear differently when plotted on scatter plots. It fools the regression model if built.

When the models particular to each of the datasets are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by the peculiarities.

Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good model.

3. What is Pearson's R? (3 marks)

Pearson's R is the test statistics that measures the statistical relationship, between two continuous variables. It gives information about the magnitude of the linear association, or correlation between 2 variables. It is denoted by r.

- It also mentions whether there is a statistically significant relationship between any 2 variables.
- It also mentions about how 2 variables are strongly related to each other.
- Pearson coefficient is sensitive to outliers.

4. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)

Scaling is the process of bringing all the features in the same standing since there might be some features whose units are very much different in magnitude than the rest of the features.

Scaling helps in making the model better. Due to difference in the units of the features, the correlation of the features takes a toll. It makes the underlying assumption that higher ranging numbers have superiority of some sort. So, these more significant number starts playing a more decisive role while training the model. It does not give accurate results. Hence making the model weak. Therefore, scaling is one of the most critical steps before creating a model.

The difference between Normalized Scaling and Standardized Scaling are as follows:

Normalized Scaling	Standardized Scaling
It scales and translates each feature individually such that it is in the given range on the training set between zero and one.	It features and scales them such that the distribution centered around 0, with a standard deviation of 1.
If data has too many outliers, then this method is not the best.	If data is not normally distributed, this is not the best method.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF is the measure of the extent of the correlation between one and other predictor variables in a model. It is used to check multi- collinearity. High values of VIF mean that there is high multicollinearity associated with the predictor variable.

An infinity value of VIF shows a perfect correlation between two predictor variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) which is infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot, short for Quantile-Quantile plot, compares the quantiles of two datasets, often used to assess the normality of a sample. The key uses and benefits of Q-Q plots include:

Outlier Detection:
Q-Q plots are effective in identifying outliers by comparing the quantiles of the observed data

with those of an expected distribution.

Detection of Scale and Symmetry Changes: Changes in scale and symmetry of the data can be detected through the visual patterns in the Q-Q plot.

Applicability to Any Sample: Q-Q plots are versatile and can be applied to any sample of data, not limited to specific distributions.

The significance of Q-Q plots lies in their ability to assess normality and compare the observed data against a reference distribution. These plots are particularly valuable in checking the normality assumption, which is essential for many statistical methods. Q-Q plots can be utilized to examine the distribution of data against various theoretical distributions, extending their applicability beyond just assessing normality. As methods often rely on normality assumptions, Q-Q plots play a crucial role in ensuring the validity of statistical analyses based on these assumptions.