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?**

   After plotting bar plots and seeing the value_counts for every categorical variable, almost all the categorical variables can be a good predictor for the dependent variable. They all have a very good percentage of the dataset which makes the cnt variable(dependent variable) change as these features change. This is the inference that can be made. They all have good trends. Hence these categorical variables can be chosen in the model building.

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

   If we don't use drop_first=True then the first column of the dummy variables table won't be dropped. That would bring collinearity in the model and would make the model insignificant. Hence drop_first=True is important.

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

   Looking at the pairplot among the numerical variables, we see that the highest correlation with the target variable is temp. We would see that both temp and atemp both have a similar scatter plot. Since the data points in atemp are more spread as compared to temp, the correlation in temp is more than atemp.

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

   I plotted the dist plot and the scatter plot of the residuals which is the difference in the actual values and the predicted values. After looking at the dist plot:

   - *It can be inferred that the distribution of the residuals (Error terms) is normal around zero. Hence error terms are normally distributed.*

   After looking at the scatter plot:

   - *There was no visible pattern in the error terms which says that the error terms are independent of each other.*
   - *We can conclude that the error terms have constant variance. Hence homoscedasticity is present.*

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 final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are determined by the magnitude of the coefficients and they are :

   1. *Temp*
   2. *Light Snow*
   3. *Yr*

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression algorithm shows a linear relationship between a dependent(y) variable and one or more independent variables(x). Linear regression helps build a model which helps us to find out how the value of the dependent variable is changing according to the value of the independent variables.

   It helps us to do the following:

   1. *Finding out the effect of independent variables (x) on Target/ dependent variable (y)*
   2. *Finding out the change in the target variable with respect to one or more input variable.*
   3. *To find out upcoming or the ongoing trends.*

   The equation for linear regression algorithm is:

   $$y=b_0+b_1 x + \text{random error}$$

   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.**

   Anscombe's Quartet comprises of 4 data set that have nearly identical and simple descriptive analysis yet have 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?

- *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 scaling and standardized scaling?

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?**

   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.**

   Q-Q plot stands for Quantile-Quantile plot which basically plots the quantiles of first data set with the quantiles of the second data set.

   The uses and benefits of this plot are:

   - *It can detect outliers.*
   - *Change in scale, symmetry can be detected.*
   - *Can be used with any sample of data.*

   The importance of Q-Q plots is that they are used to assess whether a variable is normal or not. We can use Q-Q plots to check our data against any distribution, not just the normal distribution. Since the methods we apply are mostly based on normality assumptions, it is important to check the normality of the sample data. This is where Q-Q plots come into the picture.