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: Following were the analysis:

- The demand is high between June to October
- Demand is more during fall season
- Clear weather also has shown increase in demand
- More demand on working day
- More demand when it's not a holiday

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

Ans: By using this parameter, we are leaving out one category as the rest of the categories will ensure all the scenarios are covered. This way it helps to reduce the multicollinearity effect.

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

Ans: Following are the correlation details:

- atemp and temp are highly correlated
- temp and atemp has high correlation with the target variable cnt too

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

Ans: Based on the R-Squared value on the training dataset, we were able to confirm that the model was able to explain 82% of the demand using the features we selected.

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

Ans: Following features contributed the most:

- atemp --> 0.373647
- yr --> 0.237669
- Weathersit/light_rain --> -0.294714

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- To model the relationship between a dependent variable and one or more independent variables, linear regression is used. It seeks to identify the straight line that fits the data the best.
- The Ordinary Least Squares(OLS) method is used in the procedure to determine the coefficients i.e. slope and intercept of the linear equation.
- By minimizing the sum of the squared differences between the predicted and actual values, the method calculates the coefficients and builds the model. It then makes predictions based on unseen data using these coefficients.
- Linear regression assumes a linear relationship between the variables and there are certain assumptions made in order to apply and evaluate a linear regression model.
- It is a useful tool for data analysis and decision-making because it shows how changes in the independent factors affect the dependent variable.

2. Explain the Anscombe's quartet in detail.

Ans:

- Anscombe's quartet explains why it is important to plot data before analyzing it and building your model.
- For x and y, and when we calculate things like the mean, variance etc. they all have almost the same values. But when we draw their numbers on a graph, you can see that their shapes and patterns are completely different.
- So, when we analyze data or build models, it's important to plot the data on a graph first. It helps us see the true scenario.

3. What is Pearson's R?

Ans:

- Pearson's R or Pearson correlation coefficient, is a way to measure how closely two sets of numbers are related to each other.
- It is a statistical measure that explains the linear correlation between two variables.
- It calculates the correlation by comparing the sum of the product of differences between corresponding values to the product of the square roots of the sums of squared differences.

- If the number is close to +1, it means they are strongly connected, if it's close to -1, they are strongly connected but in the opposite direction, and if it's close to 0, they are not really connected.

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

Ans:

- The process of scaling involves converting numerical properties to a stable range.
- If we don't have a stable range, then some of the coefficients obtained by the regression model might be very large or very small as compared to the other coefficients. At the time of model evaluation, this creates issues.
- By rescaling the features to a range from 0 to 1, normalized scaling preserves the relative relationships between the values. When there are outliers, it is helpful.
- On the other hand, standardized scaling changes the features so that they have a mean of 0 and a standard deviation of 1. When the data follows a Gaussian distribution, this preserves the original distribution's form.

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

Ans: An infinite value of VIF for a given independent variable indicates that it can be predicted by other variables in the model with extreme high accuracy. This means there is perfect multicollinearity due to redundant or highly correlated variables.

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

Ans:

- Quantile-Quantile or Q-Q plot, is a tool to help us assess if a set of data came from the same distribution.
- In linear regression, if we have separate training and test datasets, we can use a Q-Q plot to determine whether both datasets come from populations with similar distributions.
- Q-Q plots are particularly useful for evaluating the normality of residuals i.e. the difference between actual values and predicted values.