Assignment-based Subjective Questions:

Question 1: **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer 1: I found some of the categorical columns which is quite good to predict the dependent variables, the name of the categorical columns is listed below which is imp:

- Workingday
- Summer
- Fall
- Winter
- March
- April
- May
- June
- Aug
- Sept
- Oct
- Mist
- Light_Snow
- Wednesday
- Thursday
- Friday
- Saturday
- Sunday

These are some inferences:

- In 2019 the price of rental bikes is more as compared to 2018
- In season Fall and winter has reached maximum sales (Our demand is high in the season of Fall and Winter)
- In the months of July, september and october reached the max sales
- Wednesday and Friday got the maximum sales (Friday because end of the week)
- Because of the above data, weekdays have more sales compared to weekends/holidays.
- People go for outing on bikes more on when the weather is mostly Partly Cloudy

So, the business needs to focus more on Weekdays (Mostly on wednesday and Friday) in the months of July, September and october then they can increase the revenue of the company.

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

Answer 2: If we don't use this in our hyperparameter then all the categories will be converted into columns and that is not going to be a good solution or 1 extra column we will add in our dataset which is not necessary.

If we use this in our hyperparameter in **pd.get_dummies()**, then 1 column is reduced .
Example:
Let's say we have 1 column having 3 categories(yes, no, maybe) then by using **drop_first = True**, only 2 columns will be added explicitly **no and maybe** only. For yes 0,0, so no need to create this column. That's why we used **drop_first = True.**

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

Answer 3: mnth, registered and casual are the parameters which are highly correlated with the target variable in the beginning of the EDA.
Temp, yr are the parameters which are highly correlated with the target variable after EDA.

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

Answer 4: After building model, we can validate assumptions of Linear Regression:

1. Residuals are normally distributed - Make scatter plot and histograms
2. Error terms are constant variance **(Homoscedasticity) -** plotting residuals against the predicted value or again independent variable**.**
3. Check for independence by examining the residuals for any patterns or trends, and by using the Durbin-Watson test.

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

Answer 5: Top 3 Features:

- yr
- Hum
- windspeed
- temp

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer 1: Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). It is a popular technique used in machine learning and statistics for predicting the value of a continuous outcome variable based on one or more input variables.

The basic idea behind linear regression is to find the best-fitting straight line that represents the relationship between the independent and dependent variables. This line is called the regression line or the line of best fit.

The linear regression algorithm can be broken down into the following steps:

Step 1: Data Preparation
The first step in any machine learning algorithm is to prepare the data for analysis. In linear regression, this involves collecting and cleaning the data. The data should be free from any missing values, outliers, or errors.

Step 2: Model Specification
The second step is to specify the model. Linear regression models can be simple or multiple. Simple linear regression involves one independent variable, while multiple linear regression involves two or more independent variables. The model specification also involves selecting the appropriate form of the model, such as linear or quadratic.

Step 3: Model Estimation
The third step is to estimate the parameters of the model. The parameters are the coefficients that define the slope and intercept of the regression line. The most common method of estimation is the method of least squares, which involves finding the line that minimizes the sum of the squared residuals (the difference between the predicted and actual values).

Step 4: Model Evaluation
The fourth step is to evaluate the model. This involves assessing the goodness of fit of the model and testing the significance of the coefficients. The goodness of fit is measured using the R-squared statistic, which measures the proportion of variance in the dependent variable that is explained by the independent variables. The significance of the coefficients is tested using the t-test or F-test.

Step 5: Prediction
The final step is to use the model to make predictions. Once the coefficients of the model have

been estimated, new values of the independent variables can be plugged into the model to obtain predicted values of the dependent variable.

In summary, linear regression is a powerful tool for modeling the relationship between a dependent variable and one or more independent variables. By following the steps outlined above, we can use linear regression to predict the value of a continuous outcome variable based on one or more input variables.

Question 2: Explain the Anscombe's quartet in detail.

Answer 2 : Anscombe's quartet is a set of four datasets that have the same statistical properties, despite appearing very different when plotted on a graph. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data in addition to calculating summary statistics.

Each of the four datasets contains 11 (x, y) pairs of data, and they all have the same mean, variance, and correlation coefficient. However, when plotted on a graph, they appear very different.

Dataset I: This dataset shows a clear linear relationship between x and y, with a positive slope.

Dataset II: This dataset shows a non-linear relationship between x and y, with a curved shape.

Dataset III: This dataset shows a linear relationship between x and y, but with one outlier that has a large effect on the regression line.

Dataset IV: This dataset shows a perfect relationship between x and y, but only because all the y values are the same, except for one outlier that is far from the rest.

Anscombe's quartet illustrates the importance of visualizing data before drawing any conclusions. Even if summary statistics like mean, variance, and correlation coefficient are the same, the shape of the data can have a significant impact on the analysis. For example, Dataset III has a strong linear relationship between x and y, but one outlier has a large effect on the regression line. Removing that outlier may drastically change the conclusions drawn from the analysis.

In summary, Anscombe's quartet is a set of four datasets that have the same statistical properties but appear very different when plotted on a graph. It is an important reminder that summary statistics can be misleading and the visualization of data is crucial for understanding the underlying relationships.

Question 3: **What is Pearson's R?**

Answer 3: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It is a widely used statistical measure of association between two continuous variables.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It takes a value between -1 and 1, with -1 indicating a perfect negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfect positive linear relationship.

The formula for calculating Pearson's R is:

$r = (\Sigma(x - \bar{x})(y - \bar{y})) / sqrt(\Sigma(x - \bar{x})^2) * sqrt(\Sigma(y - \bar{y})^2)$

where x and y are the two variables, $\bar{x}$ and $\bar{y}$ are their respective means, and Σ is the sum of the values.

Pearson's R can be used to determine the direction and strength of the relationship between two variables. A positive value of r indicates a positive relationship, meaning that as one variable increases, the other variable also tends to increase. A negative value of r indicates a negative relationship, meaning that as one variable increases, the other variable tends to decrease.

The magnitude of r indicates the strength of the relationship, with values closer to 1 indicating a stronger relationship. However, it is important to note that Pearson's R only measures the strength of a linear relationship between two variables, and does not capture any non-linear relationships or other forms of association.

In summary, Pearson's R is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It takes a value between -1 and 1, with values closer to 1 indicating a stronger relationship. It is widely used in data analysis and can provide important insights into the relationship between variables.

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

Answer 4: Scaling is a data preprocessing technique used to transform the values of variables in a dataset to a similar scale. Scaling is performed to make sure that each variable has an equal impact on the analysis and to improve the performance of some machine learning algorithms that rely on the distance between data points.

Normalization is a scaling technique that transforms the values of a variable to fit within a specified range, typically between 0 and 1. Normalization is performed by subtracting the minimum value of the variable and dividing by the range, which is the difference between the maximum and minimum values.

Standardization is a scaling technique that transforms the values of a variable to have a mean of 0 and a standard deviation of 1. Standardization is performed by subtracting the mean value of the variable and dividing by the standard deviation.

The main difference between normalized scaling and standardized scaling is that normalized scaling changes the range of the data values, while standardized scaling changes the distribution of the data values.

Normalized scaling is useful when the range of data values varies widely, and we want to compare the relative positions of data points within the range of the variable. For example, if we have a dataset containing age and income, income will likely have a much larger range of values than age, and normalizing the income variable will ensure that it does not dominate the analysis.

Standardized scaling is useful when we want to compare the distribution of data values across variables. Standardization transforms the data values to have a mean of 0 and a standard deviation of 1, which allows us to compare the distribution of data values across variables with different units and scales.

In summary, scaling is a data preprocessing technique used to transform the values of variables in a dataset to a similar scale. Normalized scaling changes the range of the data values, while standardized scaling changes the distribution of the data values. Normalized scaling is useful when comparing the relative positions of data points, while standardized scaling is useful when comparing the distribution of data values across variables.

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

Answer 5 : VIF, or Variance Inflation Factor, is a measure of how much the variance of the estimated regression coefficient is increased due to multicollinearity in the predictor variables. A high VIF indicates that there is a strong correlation between predictor variables, which can lead to unstable and unreliable regression coefficients.

Sometimes, the value of VIF can be infinite, which means that the estimated coefficient of a predictor variable is not defined. This can happen when one or more of the predictor variables are perfectly collinear, meaning that they are perfectly correlated and can be expressed as a linear combination of each other.

When there is perfect collinearity between two or more predictor variables, the regression model cannot estimate the effect of each variable separately, as they are essentially providing the same information. This can result in an unstable model with unreliable estimates, leading to an infinite VIF value for the collinear variable.

In practice, it is uncommon to encounter perfect collinearity between predictor variables, as it would require an exact duplication of information. However, high VIF values can still indicate the presence of strong multicollinearity, which can lead to unstable and unreliable regression models.

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

Answer 6: A Q-Q (quantile-quantile) plot is a graphical technique used to compare the distribution of a sample dataset to a theoretical distribution, such as a normal distribution. The Q-Q plot compares the quantiles of the two distributions by plotting the sorted values of the sample dataset against the expected values of the corresponding quantiles of the theoretical distribution.

In linear regression, Q-Q plots are useful for assessing whether the residuals (the differences between the predicted values and the actual values of the response variable) are normally distributed. The residuals of a linear regression should ideally be normally distributed with a mean of 0, as this indicates that the model is a good fit for the data.

To create a Q-Q plot for the residuals of a linear regression, we sort the residuals from smallest to largest and plot them against the expected values of the corresponding quantiles of a normal distribution. If the residuals are normally distributed, the Q-Q plot will follow a straight line with points closely following the diagonal line.

If the Q-Q plot deviates from a straight line, it indicates that the residuals are not normally distributed. For example, if the Q-Q plot shows a curve or a significant deviation from the diagonal line, it suggests that the residuals are skewed or have heavier tails than a normal distribution. This can indicate that the linear regression model is not a good fit for the data, or that there may be issues with heteroscedasticity or outliers.

In summary, Q-Q plots are a graphical technique used to compare the distribution of a sample dataset to a theoretical distribution, such as a normal distribution. Q-Q plots are useful for assessing whether the residuals of a linear regression are normally distributed, which is an important assumption of the model. If the residuals are not normally distributed, it can indicate that the linear regression model is not a good fit for the data, or that there may be issues with heteroscedasticity or outliers.