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?

The boxplot of various categorical variables was created using seaborn and the following conclusions were drawn,

• The features 'season' and 'mnth' are similarly distributed which is to be expected.

• The demand has clearly increased from **2018 to 2019**.

• There is very little variation in the demand between different days of the week and the same is true whether it is a working day or not.

• There is clearly less demand on holidays.

• The demand is most during clear weather and declines slightly on cloudy days with a sharp drop on rainy days.

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

It is essential to use the command when creating dummy variables since it can lead to issues of multicollinearity between the variables if it's not done.

A feature that has N levels only requires N-1 dummy variables to perfectly capture the data as a zero value for all the N-1 dummy variables would automatically indicate that the dropped level takes the value 1.

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

The features ‘temp’(temperature) and ‘atemp’(temperature felt) exhibit a strong correlation to the target variable.

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

Residual analyses was performed to the model to determine the following,

• The Mean of the residuals was centered zero and the distribution was normal.

• The plot of the residuals versus the predicted values showed that the errors were randomly distributed with no clear pattern suggesting 'Homoscedasticity'.

Thus, the model was validated for the assumptions of linear regression.

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

The top three features that affect demand the most are,

• Temperature

• Weather

• Year

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

The linear regression algorithm starts by assuming that there exists a linear relationship between the independent variables and the target variable.

The algorithm tries to find a line to fit through the data so that the predicted values are as close to the actual values.

So, for a simple linear regression model the algorithm returns two values, the intercept of the line with y-axis and the slope of the line otherwise known as the beta-coefficient.

$$Y = mX + C$$

The algorithm is trained on a data set to then predict values on a test set, the algorithm continuously tries to minimize the cost function which is the squared difference in the predicted and actual value or errors squared.

The minimization of the cost function is achieved through the gradient descent algorithm that calculates the gradient of cost function and continuously updates the variables in the opposite direction of the gradient towards the minima which is the point of least error between the predicted and actual values.

Linear regression makes the following assumptions,

- **Linearity:** The relationship between independent and dependent variables is assumed to be linear.
- **Independence:** The observations are assumed to be independent.
- **Homoscedasticity:** The variance of the residuals is constant across the independent variables.
- **Normality of Residuals:** The residuals are normally distributed with mean zero.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite distinct when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics.

The datasets in Anscombe's Quartet illustrate that datasets with the same summary statistics can have different patterns and distributions. Relying solely on summary statistics like mean, variance, and correlation can lead to misleading interpretations.

Anscombe's Quartet is a reminder of the importance of exploratory data analysis like visual inspection of the data.

3. What is Pearson's R?

Pearson's correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. The value of R ranges from -1 to 1, where:

• R =1, indicates a perfect positive correlation.
• R =0, indicates there is no relation between the variables.
• R =-1, indicates a perfect negative correlation.

The magnitude of R indicates the strength of the linear relationship, Pearson's R assumes the relationship between the variables is linear. The value of R is very much influenced by outliers and the correlation should not be assumed to imply causation.

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

• Scaling is a step performed before model building to ensure all the values of the features are in a standard range.

• When the feature variables are all on the same scale their impact on the target variable can be easily inferred. Also, algorithms like gradient descent that are used to find the regression line, converge faster if the feature vectors are all on the same scale.

• Normalized scaling also known as the Min-Max scaling scales the features to a range between 0 and 1 with the minimum value at 0 and the maximum value at 1. it is given by,

$$X = (X - Xmin)/ (Xmax - Xmin)$$

• Standardized scaling ensures that the data has a mean of 0 and the standard deviation is 1. It is given by,

$$X = (X - mean(x)) / std(X)$$

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

The formula for VIF is given as,

$$VIF = 1/(1-R^2)$$

Here if the R square value for a particular variable is 1 then the VIF value approaches infinity. This means that the feature is perfectly predicted by another feature or a combination of other features.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used in statistics to assess if a dataset follows a particular theoretical distribution. It is particularly useful for checking the normality assumption of a dataset.

If the points in the plot approximately lie on a straight line, it indicates that the data is well-modeled by the chosen distribution.

Importance in linear regression,

• Q-Q plots are commonly used to check whether the residuals in linear regression follow a normal distribution.

• Q-Q plots can help identify outliers in the dataset. Outliers may appear as points deviating from the straight line in the plot.