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?

Answer: Based on observations,I can draw inference that season appears to have a significant effect on dependent variable.For e.g summer and fall season may have an positive impact on bike rentals as compared to winter and spring

Weathersit seem to influence werathersit_1 seems to influenceare likely associated with higher bike rentals compared to situations with mist, clouds, or precipitation (weathersit_2 and weathersit_3).

Year says it is better in 2019 as compared to 2018 shows popularity of bikes in increased with increase in time

The "mnth" variable, representing months from 1 to 12, may have an effect on bike rentals.

The "weekday" and "workingday" variables indicate the day of the week and whether it's a working day or not. These variables may influence bike rentals, with higher rentals on non-working days and potentially varying patterns on different weekdays.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Answer: **`drop_first=True`** during dummy variable creation is important to avoid multicollinearity and improve the interpretability of regression models.It also helps in reducing redundancy of variables which decreases the complexity of model.

Q3. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: It is been validated through the tests performed in the analysis:

1. The residuals are approximately normally distributed and centered around zero.
2. Constant variance was found so homo-scedascity is also met.
3. Almost got a straight line with predicted and test data so linear relationship assumption is also validated
4. Model is robust and multi-collinearity is also checked.

Q4. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Top three features are Temperature,Year and month.

Temperature (temp): Temperature has a positive coefficient, which means that as the temperature increases, the demand for shared bikes tends to increase.

Year (yr): Year is a binary variable representing the year (0 for 2018, 1 for 2019). A positive coefficient for the year variable suggests that there was an overall increase in bike demand from 2018 to 2019.

Month (mnth): The month variable represents the month of the year. Specific months may have a positive or negative impact on bike demand.

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a basic algorithm which is widely used supervised machine learning algorithm used for predictive modelling and helps in understanding the relationship between a dependent variable(DV) and one or more independent variables(IV). It is useful when you want to predict a continuous numeric outcome based on one or more input features present in dataset.

Terminologies used:

Dependent Variable (DV): Variable you want to predict or explain. It's also called the target variable or response variable. In linear regression, the DV is assumed to have a linear relationship with the independent variables.

Independent Variables (IV): These are the input variables or predictors. These are used to explain or predict the dependent variable. In simple linear regression, there's one independent variable, while in multiple linear regression, there are multiple independent variables present.

There are two different types of regression:

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression: This is used when there's only one independent variable. The relationship between the independent variable (X) and the dependent variable (Y) can be expressed as:

$Y=\beta 0+\beta 1X+\varepsilon$

Where X is dependent,Y is independent variable, $\beta 0$ is the intercept (the value of Y when X is 0, $\beta 1$ is the slope(change in Y for unit change in X), ε is the error term.

Multiple Linear Regression*:* Multiple Linear Regression (MLR) extends SLR to cases with more than one independent variable. The relationship is now expressed as:

$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta kXk+\varepsilon.$

- *We can evaluate the model using below methods:*
- **Mean Squared Error (MSE)**: It measures the average squared difference between predicted and actual values.
- **R-squared (R^2)**: It provides a measure of how well the independent variables explain the variation in the dependent variable.
- **Root Mean Squared Error (RMSE)**: The square root of the MSE, which gives the error in the original units of the dependent variable is RMSE value.

Q2. Explain the Anscombe’s quartet in detail

Answer : Anscombe's quartet is a famous statistical dataset consisting of four sets of data that have nearly identical simple descriptive statistics, which includes mean, variance, correlation, and linear regression . However, when visualized, these datasets looks significantly different . Anscombe's quartet was created by the statistician Anscombe to illustrate the importance of data visualization in understanding and analysing of data. It highlights the limitations of relying solely on summary statistics without visualizing the data.

Consider 4 datasets: Dataset 1:

- **x**: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7]
- **y**: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82]

When we calculate the mean of x, mean of y, variance of x, variance of y, correlation between x and y, and perform a linear regression,. The linear regression line for this dataset is a fairly good fit.

Dataset 2:

- **x**: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7]
- **y**: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26]

Dataset 2 is also quite similar as compared of summary statistics to Dataset I. However, when visualising this dataset, we'll see that it forms a parabolic shape, indicating a curved relationship between x and y.

Dataset 3:

- **x**: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7]
- **y**: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42]

Dataset 3 has a very different distribution. The relationship between x and y is not linear, but rather, it has an outlier that significantly influences the linear regression line as seen. The presence of this outlier is not evident from the summary statistics alone.

Dataset 4:

- **x**: [8, 8, 8, 8, 8, 8, 8, 19, 8, 8]
- **y**: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91]

Dataset 4 is intentionally designed to have a very high outlier. While the summary statistics might still appear similar to the other datasets, the scatterplot shows a clear outlier, which dramatically impacts the linear regression model.

Q3. What is Pearson's R?

- Answer: Pearson's correlation coefficient,It is a statistic that quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most commonly used measures of correlation in statistics .Pearson's r can take values from -1 and 1, where: 1 represents a perfect positive linear correlation: As one variable increases, the other also increases, and the data points fall perfectly along a straight line and a positive slope.
- 0 means no linear correlation: The data points are scattered without following a particular trend.
- -1 means a perfect negative linear correlation, As one variable increases, the other decreases, and the data points fall perfectly along a straight line with a negative slope

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer : Scaling in the context of data pre-processing refers to the process of transforming data to a specific range or distribution. It is performed to ensure that all features (variables) have similar

scales or magnitudes. Scaling is essential because many machine learning algorithms are sensitive to the scale of the input features in the given dataset
Scaling is performed due to reasons like:
Equalizing Feature Influence: Scaling ensures that all features contribute equally to the modelling process. If one feature has a much larger scale than others, it can dominate the learning algorithm's calculations and gets better results
Speed Convergence: Many machine learning algorithms, like gradient descent-based methods, converge faster when the features are on similar scales. This helps to improve the training process and reduces the number of iterations needed.
Avoid Numerical Instabilities: Some algorithms are sensitive to the scale of data and can experience numerical instability or convergence issues if the scales are vastly different so scaling can be done.
Improves Interpretability: Scaling also helps to improve the interpretability of model coefficients in linear models. It makes it easier to understand the relative importance of different features

Normalized Scaling (Min-Max Scaling): In normalized scaling, also known as Min-Max scaling, the values of the features are transformed to fit within a specific range, usually [0, 1]. The formula for Min-Max scaling given below:

$Xnew=(X-Xmin)/(Xmax-Xmin)$
Here, X is the original value, $Xnew$ is the scaled value, $Xmin$ is the minimum value in the feature, and $Xmax$ is the maximum value in the feature.

This scaling method is appropriate when we want to preserve the original distribution of the data while ensuring that it falls within a specified range

Standardized Scaling (Z-Score Scaling): In standardized scaling, also known as Z-score scaling, the values of the features are transformed to have a mean of 0 and a standard deviation of 1. The formula for standardization given:

$Xnew=(X-Xmean)/Xstd,$ Here, X is the original value, $Xnew$ is the scaled value $Xmean$ is the mean of the feature, and $Xstd$ is the standard deviation of the feature.

This scaling method centers the data to almost zero and scaled based on the variability in the data. It is useful when you want to standardize features to have comparable scales, especially when working with algorithms that assume normally distributed data.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: This can occur when there is perfect multicollinearity in your dataset. Perfect multicollinearity happens when two or more independent variables in a regression model are perfectly correlated or linearly dependent on each other. In this scenario, one variable can be expressed as a linear combination of the others, making it impossible to calculate a unique set of coefficients.

Here are some common situations that arises and can lead to perfect multi-collinearity:

1. **Duplicate Variables:** If you have multiple columns in your dataset that contain the same information or are linear transformations of each other, it can lead to perfect multi-collinearity.
2. **One-Hot Encoding:** In regression models with categorical variables, using one-hot encoding (creating dummy variables) can introduce multi-collinearity if the categories are perfectly correlated.

3. **Over-parameterization:** Including too many variables in a regression model, especially when you have fewer observations than variables, can lead to multi-collinearity.
4. **Linear Relationships:** If two or more independent variables are linearly related (e.g., one variable is a constant multiple of another), it can lead to multi-collinearity.

To solve the issue of infinite VIF values due to perfect multi-collinearity,we should:

1. **Identify the Source:** Determine which variables are causing the multi-collinearity. This may involve examining the correlation matrix, looking for duplicate variables, or considering the nature of your data.
2. **Resolve the Issue:** Depending on the source of multi-collinearity, you may need to drop one of the correlated variables, combine them into a single variable, or use other techniques like ridge regression (which can handle multi-collinearity).
3. **Re-compute VIF:** After resolving the multi-collinearity issue, you can recalculate the VIF values for your variables to ensure that they are within an acceptable range.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: Q-Q (Quantile-Quantile) plot is a graphical tool that is used in statistics and data analysis to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles (ordered values) of the observed data against the quantiles of the theoretical distribution. The Q-Q plot provides a visual way to evaluate whether the data closely matches the expected distribution or if there are deviations from it.

The importance and use of Q-Q plots in linear regression are as follows:

1. **Normality Assumption**: Linear regression often assumes that the residuals (the differences between the observed and predicted values) are normally distributed. Checking the normality of residuals is crucial because violations of this assumption can affect the validity of statistical inferences and hypothesis tests in linear regression.
2. **Detecting Departures from Normality**: By creating a Q-Q plot of the residuals, you can visually assess whether they follow a normal distribution. If the points on the Q-Q plot deviate from the straight line, it may indicate non-normality.
3. **Identifying Outliers**: Outliers in the residuals can distort the assumptions of linear regression. Q-Q plots can help us identify outliers by examining data points that are far from the expected line in the plot.
4. **Model Adequacy**: Q-Q plots are part of a suite of diagnostic tools used to assess the adequacy of a linear regression model. Together with other diagnostic plots like residual plots and leverage plots, Q-Q plots help us determine if your model is appropriate for your data.

Q-Q plots are a valuable tool in linear regression analysis because they provide a graphical and intuitive way to assess whether your data conforms to the normality assumption and to detect deviations that could impact the validity of your regression model. By using Q-Q plots and other diagnostic tools, you can make informed decisions about the adequacy and reliability of your linear regression analysis.