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

1. Season
2. year
3. Month
4. holiday
5. weekday
6. workingday
7. weathersit

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

Using drop_first=True during dummy variable creation helps to improve the performance of regression model to avoid multicollinearity issues.

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. When dummy variables are created without dropping one of the levels, the dummy variables become perfectly collinear.

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

The 'Temp' and 'aprarent_temp' have strong correlation of value 0.63.

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

1. Residual Analysis
2. Linearity
3. Multicollinearity
4. Distribution of error terms
5. No corelation in the residuals (Durbin - watson value is 2.08)

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

a) The temp with coefficient of 0.5499
b) weathersit_Light Snow/Rain (weathersit = 3) with coefficient of -0.2880
c) The 'year' with a coefficient of 0.2331

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables and the dependent variable to find the best-fitting straight line that minimizes the difference between the observed and predicted values.

There are two types of Linear Regressions:

Simple Linear Regression
This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.

Multiple Linear Regression
This involves more than one independent variable and one dependent variable.

The aim of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.
In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

What is Best Fit Line?
The error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

Formula: $y = mx + c$

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous example in statistics that illustrates the importance of graphically exploring data before drawing conclusions based on summary statistics. It consists of four datasets, each containing eleven (x, y) points, which have nearly identical simple descriptive statistics but appear very different when graphed.

Dataset I: This dataset consists of points that approximately follow a linear relationship. The relationship between the x and y variables is characterized by a strong linear trend. The summary statistics, such as the mean, variance, and correlation coefficient, accurately reflect this linear relationship.

Dataset II: Similar to Dataset I, Dataset II also displays a linear relationship between x and y. However, there's an outlier that significantly influences the regression line and correlation coefficient. Despite the outlier, the mean, variance, and correlation coefficient may still suggest a linear relationship, potentially leading to an oversimplified interpretation of the data.

Dataset III: This dataset illustrates the importance of considering the full distribution of data points. While the x variable remains constant for all data points, the y variable varies widely. Despite having the same mean, variance, and correlation coefficient as Datasets I and II, the data points are not well-described by a linear model. Instead, they follow a non-linear relationship.

Dataset IV: Dataset IV consists of an apparent relationship between x and y that is heavily influenced by a single data point. Without this point, the relationship between x and y would be entirely different. This dataset emphasizes the impact outliers can have on summary statistics and regression analysis.

To Summarize Anscombe's quartet is relying on summary statistics can be misleading. Graphical exploration of data, through techniques like scatter plots, can provide valuable insights into the underlying structure of the data and help identify patterns and relationships.

3. What is Pearson’s R?

Pearson’s correlation also known as Pearson’s R. Correlation coefficients are used to check how strong a relationship (statistical measure that quantifies the strength and direction) between two continuous variables . This is the most popular one and is commonly used for linear regression. The Pearson’s correlation coefficient is denoted with the “R”. It returns a value between 1 and -1.

-1 indicates a strong negative relationship (If one variable increases and other variable also increases proportionally)
1 indicates strong positive relationship (If one variable increases the other variable decreases proportionally)
0 indicates "No linear relationship" at all

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

Scaling is one of the most important preprocessing technique used in data analysis and machine learning. It is used to normalize the range of variables within a dataset to a similar scale.
There are some feature scaling techniques such as Normalization and Standardization.

Normalized scaling: It is also known as min-max scaling, the values of features are transformed to a scale between 0 and 1.

Standardized scaling: This is also known as z-score scaling, the values of features are transformed to have a mean of 0 and a standard deviation of 1.

Normalized scaling transforms features to a fixed range (typically between 0 and 1) while preserving their original distribution. On the other hand standardized scaling standardizes features to have a mean of 0 and a standard deviation of 1, resulting in a standard normal distribution. Based on the requirement the scaling methods depends on the specific requirements of the machine learning algorithm and the characteristics of the dataset

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

The Variance Inflation Factor (VIF) is a measure used in regression to assess the severity of multicollinearity among predictor variables. A high VIF value indicates that the variance of the estimated regression coefficients is inflated, making it difficult to assess the individual effects of each predictor variable accurately.

Here are some common scenarios that can lead to perfect multicollinearity and, consequently, infinite VIF values:

Linear dependencies: If one independent variable is a constant multiple of another (e.g., X2 = 2 * X1), perfect multicollinearity occurs, leading to an infinite VIF for one of the variables.

Data coding issues: In categorical variables with dummy coding, perfect multicollinearity can occur if one category can be perfectly predicted from the other categories.

Overfitting: Including variables that are linear combinations of other variables already in the model can lead to perfect multicollinearity.

Measurement error: If two variables are measured with significant error and the errors are correlated, it can lead to spurious correlations and, consequently, multicollinearity.

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

The quantile-quantile(q-q plot) plot is a graphical method/tool used to determine whether a dataset follows a particular probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly used for assessing whether a dataset is normally distributed (mean of 0 and standard deviation of 1) or certain if it follows some other known distribution. They are widely used in statistics, data analysis, and quality control.

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the observed data to the quantiles of a theoretical distribution, typically a standard normal distribution (mean of 0 and standard deviation of 1).

Normal Distribution:
The normal distribution (aka Gaussian distribution Bell curve) is a continuous probability distribution representing distribution obtained from the randomly generated real values.