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

Ans: In the given dataset all categorical variables are already converted into numerical variables. We can easily scale them if they are in numerical state. Below are the variables converted from categorical into numerical.

season : season (1:spring, 2:summer, 3:fall, 4:winter)
yr : year (0: 2018, 1:2019)
mnth : month (1 to 12)
holiday : weather day is a holiday or not
weekday : day of the week
workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
weathersit :
1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

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

Ans: By dropping first column while creating dummies from the variable it will reduce redundancy from the dataset also we can achieve same results by dropping first column in it.

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

Ans: temp and atemp has highest correlation also registered and cnt variables

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

Ans: After building model I have scatter plot between y-test and y-pred, it is correlate with each other also coefficients of each final predictors are similar. And all p-values are less than 0.05 and variance inflation factor(VIF) is less than 5. So that we can decide that derived model is appropriate and will fit to given data set.

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

Ans: season, yr, weathersit are the top 3 features to explain demand of the shared bikes

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a supervised learning algorithm used for predicting continuous numeric values based on the relationship between independent variables (features) and a dependent variable (target). It assumes a linear relationship between the features and the target variable and aims to find the best-fitting line that minimizes the overall error.

There are two types of linear regression:

a. Simple linear regression
b. Multiple linear regression

Simple Linear regression:

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straightline. The straight line is plotted on the scatter plot of these two points.

The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

R^2 or Coefficient of Determination:

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

RSS(ResidualSumofSquares), TSS(Total sumofsquares)

Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

The formulation for multiple linear regression is also similar to simple linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$$Y = \beta_0 + \beta_1X1 + \beta2X2 + \beta3X3 + \beta4X4 + . . . + \beta nXn$$

Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as: 1. Model now fits a 'hyperplane' instead of a line 2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion) 3. For inference, the assumptions from Simple Linear Regression still hold ○ Zero mean ,Independent, Normally distributed error terms that have constant

There are a few new considerations that you need to make when moving to multiple linear regression, such as:

1. Adding more isn't always helpful
    a. Model may 'overfit' by becoming too complex
        i. Model fits the train set 'too well', doesn't generalize
        ii. Symptoms: high train accuracy, low test accuracy
    b. Multicollinearity
        i. Associations between predictor variables
2. Feature selection

Model Evaluation: Once the coefficients are estimated, the model's performance is evaluated using various metrics. Common evaluation metrics for linear regression include:

R-squared (R^2): Measures the proportion of variance in the target variable explained by the model. R^2 ranges from 0 to 1, where higher values indicate a better fit.

Prediction: Once the model is trained and evaluated, it can be used for making predictions on new, unseen data. Given a set of features, the model applies the learned coefficients to calculate the predicted value for the target variable.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet refers to a set of four datasets that have nearly identical statistical properties, yet display remarkably different patterns when visualized. The quartet was introduced by the statistician Francis Anscombe in 1973 to highlight the importance of visualizing data and not relying solely on summary statistics.

The datasets in Anscombe's quartet consist of pairs of x and y values. Let's examine each dataset in detail:

Dataset I:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset II:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Dataset III:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset IV:
x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Now, let's examine the statistical properties of these datasets:
Mean of x: All datasets have a mean of 9.
Variance of x: All datasets have a variance of 11.
Mean of y: Dataset I, II, and III have a mean of approximately 7.5, while Dataset IV has a mean of 7.0.
Variance of y: Dataset I, II, and III have a variance of approximately 4.12, while Dataset IV has a variance of 4.12.

Despite these statistical similarities, the patterns revealed by visualizing the datasets differ significantly:
Dataset I: Displays a relatively linear relationship between x and y.
Dataset II: Shows a non-linear pattern, where there is a slight upward curve.
Dataset III: Appears to have a strong linear relationship, except for an outlier at (13, 12.74) that influences the linear fit.

Dataset IV: Consists of a horizontal line with a single outlier at (19, 12.50), showcasing the impact of an influential point.

The main purpose of Anscombe's quartet is to emphasize that relying solely on summary statistics can be misleading. It highlights the importance of visualizing data to gain a deeper understanding of the underlying patterns, relationships, and potential outliers or influential points. By examining the quartet, one can appreciate that data exploration and visualization are crucial steps in the statistical analysis process.

3. What is Pearson's R?

Ans: Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between two continuous variables.

Pearson's R is a value that ranges between -1 and 1, where:

- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship or correlation between the variables.

The formula for calculating Pearson's R is as follows:
$R = (\Sigma((x_i - \bar{x})(y_i - \bar{y}))) / (\text{sqrt}(\Sigma(x_i - \bar{x})^2) * \text{sqrt}(\Sigma(y_i - \bar{y})^2))$
Where:
x_i and y_i are the individual values of the two variables.
$\bar{x}$ and $\bar{y}$ are the means of the two variables.
Σ denotes the summation of values.

To compute Pearson's R, the formula involves calculating the covariance between the two variables ($\Sigma((x_i - \bar{x})(y_i - \bar{y}))$) and dividing it by the product of their standard deviations ($\text{sqrt}(\Sigma(x_i - \bar{x})^2) * \text{sqrt}(\Sigma(y_i - \bar{y})^2)$).

Pearson's R has several important properties:
It is symmetric, meaning that the correlation between variable X and variable Y is the same as the correlation between variable Y and variable X.

It is affected by the scale of measurement of the variables but not by their units. Thus, it is unitless.

It only measures the linear relationship between variables and may not capture non-linear relationships.

It is sensitive to outliers, meaning that extreme values can have a significant impact on the correlation coefficient.

Pearson's R is widely used in various fields, including statistics, social sciences, finance, and machine learning. It provides a useful summary statistic to understand the **direction** and strength of the linear relationship between two continuous variables.

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

Ans: Scaling is the process of transforming the values of variables to a specific range or distribution. It is done to ensure that different variables are on a similar scale and have comparable magnitudes.

Normalized Scaling (Min-Max Scaling): It rescales the values of variables to a specific range, usually between 0 and 1. Normalized scaling preserves the relative relationships and distribution shape of the variable.

Standardized Scaling (Z-score Scaling): It transforms the values of variables to have a mean of 0 and a standard deviation of 1. Standardized scaling centers the variable around 0 and scales it based on its dispersion.

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

Ans: The phenomenon of an infinite value of VIF (Variance Inflation Factor) occurs when there is perfect multicollinearity in the data. Multicollinearity refers to a high correlation between independent variables in a regression model, which can cause issues in the model's estimation and interpretation.

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

Ans: A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It helps to evaluate whether the data follows a particular distribution and to identify deviations from that distribution.

The Q-Q plot compares the quantiles of the observed data against the quantiles of a reference distribution, typically a normal distribution.

Q-Q plots are useful tools for assessing the normality assumption of the residuals in linear regression. They help in detecting deviations from the assumed distribution, identifying outliers, and evaluating the goodness of fit of the model. By examining Q-Q plots, researchers and analysts can make informed decisions about the validity and reliability of the regression model and its assumptions.