Assignment-based Subjective Questions

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

- Season plays important role and its directly reflecting on the temperature parameter.

- 2019 have huge increase in demand compared to 2018.

- 70% of sale is on holiday and fall weather.

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

- When creating dummy variables, it is important to use drop_first=True to avoid the dummy variable trap. The dummy variable trap occurs when there is perfect multicollinearity between two or more variables, which means that the presence of one variable can be perfectly predicted by the presence or absence of the other variables.

- By using drop_first=True, we can avoid this issue by dropping the first dummy variable for each categorical variable. This ensures that there is no perfect multicollinearity between the variables, and the resulting regression model will be more stable and accurate.

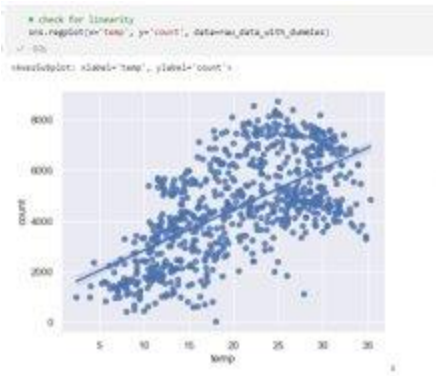
How did you validate the assumptions of Linear Regression after building the model on the training set ?

- Linearity: The relationship between the independent and dependent variables should be linear. This means that the slope of the regression line should be constant for all values of the independent variable.

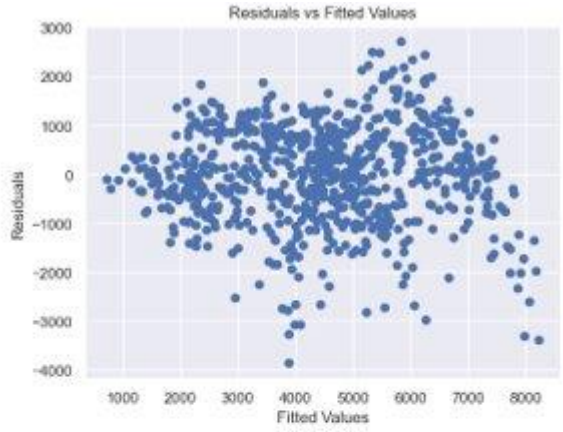
- Independence: The observations should be independent of each other. This means that the value of one observation should not depend on the value of any other observation.

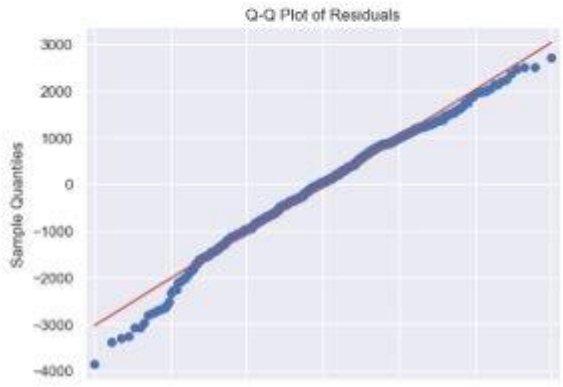
- Homoscedasticity: The variance of the residuals (the difference between the predicted and actual values) should be constant across all levels of the independent variable. This is called homoscedasticity.

- Normality: The residuals should be normally distributed. This means that the residuals should be symmetrically distributed around zero and follow a bell-shaped curve.
- No multicollinearity: The independent variables should not be highly correlated with each other. This is called multicollinearity and can lead to unstable estimates of the regression coefficients.



	Features	VIF
0	temp	3.93
1	year	1.92
2	windspeed	3.72





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

- Temp
- Year
- Windspeed

Explain the linear regression algorithm in detail ?

- Linear regression is a statistical method used to model the relationship between one or more independent variables and a dependent variable. The goal of linear regression is to find the best-fitting linear equation that describes the relationship between the independent variables and the dependent variable.

- In simple linear regression, there is only one independent variable, and the goal is to find the line of best fit that describes the

relationship between the independent variable and the dependent variable. The equation for a simple linear regression model is given by:

$y = \beta0 + \beta1 * x + \varepsilon$

- where:
  - y is the dependent variable
  - x is the independent variable
  - β0 is the intercept or constant term
  - β1 is the slope coefficient
  - ε is the error term or residual

- The slope coefficient (β1) represents the change in the dependent variable (y) for every one-unit change in the independent variable (x). The intercept term (β0) represents the value of the dependent variable (y) when the independent variable (x) is equal to 0.

- In multiple linear regression, there are two or more independent variables, and the goal is to find the best-fitting linear equation that describes the relationship between the independent variables and the dependent variable. The equation for a multiple linear regression model is given by:

  $y = \beta0 + \beta1 * x1 + \beta2 * x2 + ... + \beta k * xk + \varepsilon$

- where:
  - y is the dependent variable
  - x1, x2, ..., xk are the independent variables
  - β0 is the intercept or constant term
  - β1, β2, ..., βk are the slope coefficients
  - ε is the error term or residual

- The slope coefficients (β1, β2, ..., βk) represent the change in the dependent variable (y) for every one-unit change in the corresponding independent variable (x1, x2, ..., xk). The intercept term (β0) represents the value of the dependent variable (y) when all the independent variables are equal to 0.

- The performance of a linear regression model is typically evaluated using metrics such as mean squared error (MSE), root mean squared error (RMSE), R-squared (R^2), and mean absolute

error (MAE). These metrics help to assess how well the model fits the data and make predictions. Linear regression is widely used in various fields such as finance, economics, engineering, and social sciences, among others.

Explain the Anscombe's quartet in detail.

- Despite having different descriptive statistics, the four datasets in the quartet all produce the same linear regression model: y = 3 + 0.5x. This demonstrates the limitations of relying solely on summary statistics and the importance of visualizing data to understand the underlying relationships.

- Here are the details of the four datasets in Anscombe's quartet:
  - Dataset I: This dataset has a linear relationship between x and y, with no apparent outliers. The summary statistics for x and y, such as mean, variance, and correlation coefficient, are all close to the values in the linear regression model.

  - Dataset II: This dataset also has a linear relationship between x and y, but with one outlier point that has a much higher y value than the other points. The linear regression model still fits the data well, but the outlier point has a large influence on the regression line.

  - Dataset III: This dataset has a non-linear relationship between x and y, with one outlier point that has a much lower y value than the other points. The linear regression model is not a good fit for this dataset, as it assumes a linear relationship between x and y.

  - Dataset IV: This dataset has a perfect quadratic relationship between x and y, with no outliers. The linear regression model is again not a good fit for this dataset, as it assumes a linear relationship between x and y.

What is Pearson's R?

- Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and takes values between -1 and 1.

- A value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable 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 between the two variables.

- The Pearson correlation coefficient is widely used in fields such as statistics, economics, psychology, and engineering to measure the degree of association between two variables. It assumes that the relationship between the variables is linear and that the data are normally distributed.

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

- Scaling is the process of transforming numerical data into a specific range to make it easier to compare and analyze. It involves changing the range of values for a variable, so they fall within a specific range or scale.

- Scaling is performed to bring all the features of the data to a common scale so that no feature has undue influence on the model. It also helps in improving the performance of the model, as it can help in reducing the impact of outliers and improve convergence of the optimization algorithms.

- There are two commonly used methods for scaling data: normalized scaling and standardized scaling.

- Normalized scaling involves scaling the data so that it falls within a specific range, usually between 0 and 1. This is done by subtracting the minimum value from each observation and dividing by the range. This method is useful when the minimum and maximum values of the variable are known and the data is uniformly distributed.

- Standardized scaling involves transforming the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean value from each observation and dividing by the standard deviation. This method is useful when the mean and standard deviation of the variable are known and the data is normally distributed.

- The main difference between normalized scaling and standardized scaling is the range of values. Normalized scaling limits the range of values to between 0 and 1, while standardized scaling retains the original range of values but transforms them to have a mean of 0 and standard deviation of 1.

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

- Sometimes, the value of VIF can be infinite. This happens when one or more of the independent variables in the model are perfectly correlated with each other. When two or more variables are perfectly correlated, it means that they have a correlation coefficient of 1 or -1. In this case, the VIF value for one of the variables will be infinite because the denominator in the VIF equation becomes zero.

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

- A Q-Q plot, or quantile-quantile plot, is a graphical technique used to assess whether a set of data follows a particular distribution, such as a normal distribution. The Q-Q plot compares the distribution of the observed data to the expected distribution, which can be based on a theoretical distribution or on the distribution of a reference dataset. The Q-Q plot is created by plotting the quantiles of the observed data against the quantiles of the expected distribution.

- In linear regression, Q-Q plots can be used to check the normality assumption of the errors, which is one of the key assumptions of linear regression. The normality assumption requires that the errors follow a normal distribution, with a mean of zero and a constant variance. If the errors are not normally distributed, it can lead to biased estimates of the coefficients, and the results of the analysis may not be reliable.