Linear Regression - Bike Sharing Assignment

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1. Demand is high during the Fall season.
2. Demands of Bikes has increased from 2018 to 2019
3. July to Sept is a high demand duration
4. There is no major difference between Working day and holiday

Q2. Why is it important to use drop_first=True during dummy variable creation?
This flag helps to reduce an extra column creation

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
temp & atemp has similar correlation with respect to target variable

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?
After building a linear regression model on the training set, it is important to validate the assumptions of the model to ensure that it is reliable and accurate.

Linearity: Check if there is a linear relationship between the independent variables and the dependent variable.

Homoscedasticity: Check if the variance of the errors is constant across all levels of the independent variables.

Independence: Check if the errors are independent of each other. This can be done by examining the autocorrelation function (ACF) plot of the residuals.

Normality: Check if the errors are normally distributed. This can be done by creating a histogram of the residuals and checking if it resembles a normal distribution.

Multicollinearity: Check if there is high correlation between the independent variables. This can be done by creating a correlation matrix of the independent variables and checking for high correlation coefficients.

If any of these assumptions have been violated, the model may not be reliable, and further investigation or modifications may be necessary.

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

1. If temp increases, the bike demand increases.
2. In fall season, the demand of bike increases

General Subjective Questions

Q1. Explain the linear regression algorithm in detail

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables. It is called "linear" regression because the relationship between the variables is assumed to be linear.

Here are the steps involved in the linear regression algorithm:

Data Collection: This involves collecting data on the dependent variable

Data Preprocessing: This involves cleaning the data, handling missing values, and removing outliers.

Model Building: In linear regression, the model is a mathematical equation that describes the relationship between the dependent variable and the independent variable(s). The equation takes the form of y = mx + c, where y is the dependent variable, x is the independent variable, m is the slope of the line, and c is the intercept.

Training the Model: The next step is to train the model using the data. This involves finding the values of m and c that minimize the difference between the predicted values and the actual values of the dependent variable. This is done using a method called "least squares regression".

Model Evaluation: After the model has been trained, it needs to be evaluated. This involves testing the model on a separate set of data (the test set) to see how well it performs. The most common metric used to evaluate a linear regression model is the R-squared value, which measures how well the model fits the data.

Model Prediction: Once the model has been evaluated and found to be satisfactory, it can be used to make predictions on new data. This involves plugging in the values of the independent variable(s) into the equation and calculating the predicted value of the dependent variable.

Q2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets to demonstrate the importance of visualizing data and checking assumptions in statistical analysis. Despite having vastly different patterns, each dataset shares the same basic statistical properties, including the same mean, variance, and correlation coefficient.

The purpose of Anscombe's quartet is to illustrate that different datasets can have the same statistical properties, yet have vastly different visual patterns. This highlights the importance of visualizing data and checking assumptions in statistical analysis before drawing conclusions from the data.

Q3. What is Pearson’s R?

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

Scaling is a process of transforming the numerical features in a dataset to a common scale. This is done by applying a mathematical function to the features that changes their values without altering their distribution or relationship with other features.

Scaling is performed for various reasons.

To improve the performance of machine learning algorithms that use distance-based measures, such as k-nearest neighbors and clustering algorithms, as these algorithms are sensitive to the scale of the features.

To improve the convergence of optimization algorithms, such as gradient descent, which can be slowed down by features with large ranges.

To reduce the impact of outliers, as the values of these outliers can be relatively large compared to other data points.

There are two common methods of scaling: normalized scaling and standardized scaling.

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

The VIF (Variance Inflation Factor) is a measure of how much the variance of the estimated regression coefficient is increased due to multicollinearity among the predictor variables. VIF values greater than 1 indicate that the variance of the coefficient estimates is increased due to multicollinearity, with higher VIF values indicating higher levels of multicollinearity.

If two predator variables have the same correlation, then VIF is infinite.
I.e. 1/1-R = 1/0 = infinite

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