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:

Year: The demand for rental bikes have increased from 2018 to 2019.

Month: The demand for rental bikes have are increasing from Jan to July and then decreases towards the end of the year.

Holiday: There is not much difference in the upper bound for demand over the holiday vs non-holiday, however the median demand on holiday is lower than the non-holiday.

Weekdays: The median demand on all the weekdays is similar.

Workingday: There is no difference in the bike demand whether the day is working day or not.

Season: There is a high demand for rental bikes in summer and fall as compared to winter and spring.In Spring season, the demand for rental bike is the lowest.

Weather: There is a high demand for rental bikes in when the weather is clear as the demand goes on decreasing as the weather worsens.

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

Answer:

The reason for this is to avoid multicollinearity. Multicolinearity which occurs when two or more independent variables are highly correlated with each other.

When creating dummy variables, if we include all of the dummy variables created from a categorical feature, we will end up with one dummy variable being a linear combination of the others.

By dropping one of the dummy variables, we prevent multicollinearity. The dropped dummy variable can be easily guessed by using the values present in the other categories.

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

Answer:

The variable ‘temperature’ has the highest correlation with the target variable.

Note: This is after removing the non-necessary variables like casual and registered.

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

Answer:
For validating the assumptions of linear regerssion, I have performed following steps:
A. Predicted the target using the trained model and then calculated the residual as difference of actual and predicted.
B. Then used histogram to see the distribution of the residuals which showed normal distribution with mean=0
C. To test homoscadasity, plotted the residuals vs the actual values on a scatter plot. This showed that the residuals are not having any specific trend.
D. Then used VIF to check for multicolinearity within the independent variables.
E. Used scatter plot to check for the linear relation between target and independent variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
The top 3 features explaining the demand of bikes are
A. Weather of light snow and rain
B. Year
C. Winter season

General Subjective Questions
1. Explain the linear regression algorithm in detail.
Answer:
Linear regression is a machine learning algorithm used to predict a continuous output variable based on one or more input variables. The algorithm finds a linear relationship between the input variables and the output variable, which can then be used to make predictions for new data.

The goal of linear regression is to find the line that best fits the data, which is called the regression line. The line is defined by an intercept and one or more coefficients. The regression line is expressed as:
Y = mX + c
Where Y is target
X is independent variable
C is constant
M is slope

The algorithm uses a training dataset to estimate the values of the coefficients that minimize the sum of squared residuals (SSR) between the predicted and actual values of the output variable.

The coefficients can be estimated using the method of least squares, which involves finding the values that minimize the cost function.

2. Explain the Anscombe's quartet in detail.
Answer:
Anscombe's quartet is a set of four datasets that have identical statistical properties but are visually distinct to demonstrate the importance of data visualization in understanding and interpreting data.
Each dataset in the quartet consists of 11 pairs, which are plotted on a scatterplot.

3. What is Pearson's R?
Answer:
Pearson's R is Pearson correlation coefficient which indicates the extent to which two variables are linearly related. It measures the degree to which the two variables change together, and ranges between -1 and 1, with 0 indicating no correlation.
The Pearson correlation coefficient is calculated as the covariance between two variables divided by the product of their standard deviations.
A positive Pearson correlation coefficient indicates that as one variable increases, the other variable also tends to increase. A negative correlation coefficient indicates that as one variable increases, the other variable tends to decrease. A correlation coefficient of 0 indicates no linear relationship between the two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:
Scaling is the process of transforming data into a specific range. The purpose of scaling is to normalize the data and make it more suitable for use in training machine learning algorithms.

The main reasons for performing scaling are:

A. Scaling the variables can help avoid the issue of variables with larger scales dominating the algorithm.
B. Scaling can make it easier to compare variables on the same scale or interpret the importance of variables.

Normalized scaling involves scaling the data to fall within a range of 0 to 1. This is done by subtracting the minimum value of the variable from each data point and dividing by the range of the variable
Normalized scaling preserves the relative relationships between the data points, but does not preserve the mean or standard deviation of the data.

Standardized scaling involves transforming the data to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the variable from each data point and dividing by the standard deviation of the variable. Standardized scaling preserves the mean and standard deviation of the data, but does not preserve the relative relationships between the data points.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer:
The value of VIF may be infinite and it happens when one of the predictor variables in the model is a perfect linear combination of one or more of the other predictor variables. This means that one of the variables can be expressed as a linear combination of the other variables with no error, resulting in a perfect correlation and a VIF of infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer:A Q-Q plot, short for quantile-quantile plot, is a graphical technique used to compare the distribution of a sample data to a theoretical distribution, such as a normal distribution.
In a Q-Q plot, the quantiles of the sample data are plotted against the expected quantiles of the theoretical distribution. If the sample data is normally distributed, the points in the Q-Q plot should roughly fall on a straight line.
In linear regression, Q-Q plots can be used to assess the normality of the residuals, which are the differences between the observed values and the predicted values from the regression model. If the residuals are normally distributed, the Q-Q plot of the residuals

will show a straight line, while departures from the line indicate non-normality.