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?

Ans:

Week Day – Demand almost same on most week days.

Working Day - Demand slightly more on working days than weekend / holiday

Holiday - Lesser demand on holidays

Month - Demand most in June - August (maxm in June and August)

Year – Demand more in 2019 than 2018

Season – Demand maximum in fall / autumn

Weather situation – Demand maximum during clear weather

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

Ans

For n categories in a categorical variable, n-1 dummy variables need to be created as the information in n variables is adequately in n-1 variables.

“drop_first = true” is basically to drop the original column of the categorical variable that has now been broken down into dummy variables

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

Ans

Ttemp has the highest correlation.

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

Ans:

Linear relationship between X & Y - Validated via scatter plot

The error terms are normally distributed – Validated via residual analysis

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

Ans : Temperature(temp) , Year(yr) , season_4 (in decreasing order of coefficients) contribute significantly towards demand of shared bikes

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   There are two main types of linear regression:

   Simple Linear Regression

   This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

   $y=\beta 0+\beta 1X$ $y=\beta 0+\beta 1X$

   where:

   - Y is the dependent variable
   - X is the independent variable
   - β0 is the intercept
   - β1 is the slope

   Multiple Linear Regression

   This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

   $y=\beta 0+\beta 1X1+\beta 2X2+\ldots\ldots\ldots\beta nXn$ $y=\beta 0+\beta 1X1+\beta 2X2+\ldots\ldots\ldots\beta nXn$

   where:

   - Y is the dependent variable
   - X1, X2, …, Xn are the independent variables
   - β0 is the intercept
   - β1, β2, …, βn are the slopes

   The goal 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.

2. **Explain the Anscombe's quartet in detail.**

Ans:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

3. What is Pearson's R?

Ans :

Pearson's correlation coefficient(R) has a value between -1 and +1 .

It refers to the degree of correlation between 2 variables.

The closer the value is to +/- 1, the higher the correlation.

+ve value means both variables increase/decrease in the same direction while -ve value means they increase/decrease in the opposite direction.

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

Ans:

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods .

You can scale the features using two very popular methods:

Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

X= x-mean(x)/SD(x)

MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

X= x- min(x) / max(x) – min(x)

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

Ans : This means that the variable is 100% correlated with all other variables , VIF = 1/1-Rsq , so whenever the Rsq value of variable is 1 , its VIF will be Inf

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

Ans : The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.