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:

I) A coefficient value of year (0.233) indicates that a unit increase in

year variable increases the bike hire numbers.

ii) A coefficient value of season_4(0.129) indicates that w.r.t season_1,

a unit increase in season_4 variable increases the bike hire numbers.

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

Answer:

I) We should drop the first column, because it is completely defined by other

variables, if all other variables are zero then it quite obvious it is first one

so, it does not add any unique information to our model.

3. Looking at the pair_plot among the numerical variables, which one has the highest

correlation with the target variable?

Answer: Temperature (0.049) with positively correlated.

4. How did you validate the assumptions of linear regression after building the

model on the training set?

Answer:

I) Plotted a displot with the residuals (Actual y value - Predicted y value) and verified whether the error terms are uniformly distributed and the mean is at zero.

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

Answer:

1) Temperature (0.549)
2) Year (0.233)
3) Season_4(0.129)

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is used to predict the value of dependent variable with the independent variables. In simple linear regression the algorithm fits a straight line where in case of multiple linear regression it fits a surface that minimize the difference between the predicted and actual output values. To find the best line or surface gradient descent algorithm is used to reduce the cost function (Calculation of error between predicted value and actual value)

2. Explain the Anscombe's quartet in detail.

Different datasets with exactly same summary statistics such as mean, variance, corelation coefficient and line of best fit does not mean that they look exactly similar when visualised. The effect of outliers and curvature might drastically throw off our summary statistics and this is called Anscombe's quartet and demonstrates how important it is to always plot your data rather than relying on summary statistics alone.

3. What is Pearson's R?

Pearson's R measures the strength of the linear relationship between two variables and it is always between -1 to 1. R would be 1 if there is a perfect positive linear relationship (i.e., Y increases exactly with increase in X) and R would be -1 if there is a perfect negative liner relationship (i.e., Y decreases exactly with increase in X).

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

Scaling is the process that brings all the variables (dependant and independent) to the same scale. We perform scaling for ease interpretation and faster convergence for gradient descent method. Standardized scaling basically brings all the data into a standard normal distribution with mean zero and standard deviation one whereas normalized scaling or minmax scaling brings all the data in the range of 0 and 1

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

VIF shows the correlation between the independent variables and the VIF formula is 1/(1-r^2) where r^2 is a correlation between the variables. If the variable is perfectly correlated then the r^2 value would be 1 and the VIF formula would become infinite as the denominator will be zero.

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

Q-Q plot (Quantile-Quantile) plot helps us to identify how the data in the dataset are distributed. A Q-Q plt has two axes like a scatter plot in which we plot distribution quantile (i.e., uniform, normal etc) vs Data quantile. If the Q-Q plot show linear relationship then the distribution with which we the Q-Q is same as that of the data is distributed else it is not.