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? (3 marks)

   Answer:

   Categorical variables don't have any significant effect on the dependent variable. Its because, dteday is only a categorical variable and it is unique across all the rows which is not useful for the modelling as it doesn't give a significant patter due to the uniqueness.

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

   Answer:

   It is mandatory to convert categorical in to numerical. Drop_first=True will remove the first converted variable. It helps to prevent the multicollinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Answer:

   temp and atemp variables are highly correlated with target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer:

   We have to consider some metrics to ensure that, model is significantly good. Some of the metrics are p-value and R squared.

   Follow below rules to remove the features and rebuild the model by updating the model with new features.

   - If P-value >0.05 and VIF > 5 then remove the feature
   - If P-value <0.05 and VIF < 5 then keep the feature
   - First check, If P-value is high and VIF is low then remove the feature and rebuild the model. Else If P-value is low and VIF is high then remove the feature.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   Answer:

   Humidity, Temperature, Year

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Answer:

   Linear Regression is a regression model which is useful to predict continuous variable. LR is a supervised learning method.

   LR will be used to predict the relationship between dependent and independent variables using a straight line.

   There are two types of LR.

   1. Simple LR

This model will be used to find single relationship between independent and dependent variable

Formula: Y = Beta 0 + Beta1.X

Where Beta 0 represents intercept
Beta 1 represent slope
X is input variable
Y is output variable

2. Multiple LR

This model will be used to find relationship between several independent and one dependent variables

Formula: Y = Beta 0 + Beta1.X1 + Beta2.X2 +.....+ Betan.Xn

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

It will contain four datasets and will share the same statistics (mean, variance, correlation and regression line)

Each dataset consists of 11 pairs.
1. First dataset: Will be linear
2. Second: quadratic curve
3. Third: Strong relationship but prone to outlier
4. Fourth: Has same x-values but different y-values.

Main purpose of this, to derive whether all the datasets are similar or not.

3. What is Pearson's R? (3 marks)

Answer:

It will used to measure the strength and direction of linear relationship between the continuous variables. If the relationship is not linear then Pearson's R will not be significant/accurate. Pearson's will be prone to outliers very much.

Values lies between -1 to 1
-1 : negative relationship
1 : positive relationship
0 : no linear relationship

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling will usually perform on numerical features. Scaling will convert the numerical values to in a specific range (0 to 1).

Purpose of scaling is to bring all features in a similar scale. Mean to say, features will have different scales which misleads the ML models. Due to that, scaling will be mandatory to perform the ML model better.

Difference between normalized and standardized is, Normalization will convert the values in to specific range 0 to 1. Where as Standardized, it will transforms data to have a mean of 0 and standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Answer:

   It means there is a perfect multicollinearity (highly correlated) between the predictor variables. Infinite VIF will be not significant to the model and the model will not perform well because of highly correlated features

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Answer:

   It helps to tell whether the data came from Normal, exponential or uniform distribution.
   It mainly used in Linear regression to check the residuals distribution which helps to validate the assumptions of the model.