Assignment-based Subjective Questions

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

There are no category variables in the given data hence no dummy variables are derived

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

With respect to dummy variables, lets say if we have n variables, whatever is conveying is also achived by n-1 variables, so the drop_first = true helps in reducing the columns

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

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

This can be validated by depicting the distributed plot where the plot is distributed normally and centric to zero then we can infer it as it is honoring the assumptions

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

Registered, casual and weatherist

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm depicts the linear association between various variables involved.This can be achieved by follwing steps like 1.Preparing the data for modeling

1. Encoding by converting binary vars to 1/0 and Category variables to dummy variables
2. Splitting into train and test
3. Rescaling the variables
4. Train the model 3.Residual analysis
5. Prediction and evaluation on test set

2. Explain the Anscombe's quartet in detail.

It is a set of four datasets that challenges our reliance on summary statistics and underscores the importance of data visualization.More over all the four datasets share identical summary statistics: Same means for both x and y. Same variances for both x and y. Same correlation coefficients. Same linear regression lines.

3.What is Pearson’s R?

Pearson's correlation coefficient which measures the quantative Linear relationship b/w two quantative variables. It ranges between -1 to 1 /n A positive R(0<R<=1): indicates if one varibale indicates the other one will also increases /n A negetive R(-1<=R<0): indicates if one varibale indicates the other one will decrease /n R close to zer0 indicates there is no linear relationship between the variables

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

scaling It ensures that the input variables are on a similar scale which helps for a better model perfromance Why: Lets say we have independent variables like Area and Bedrooms. Ideally the area value are far larger than rooms count.When we calculate the coefficients those will have larger difference and it make the inferences complicate Normalization scaling: which also known as Min-Max scaling. Here the data points are calculated such that it is between 0 & 1 Standardized scaling: Also known as Z-score Normalization. Here the data points are calculated WRT Mean and standard deviation

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

It represents the multicollinearity. Why:when one predictor variable is a linear combination of other predictors. For example, if you include both height in inches and height in feet as predictors, they are perfectly correlated.

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

Quantile-Quantile plot is used to infer whether data follows a normal distribution In linear regression, we often encounter separate training and test data sets. A Q-Q plot can confirm whether both data sets are drawn from populations with the same distributions.

In [ ]: