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:

Seasons:

- Spring and Summer has more demand than rest of the seasons

Yr:

- Demand got increased from 2018 to 2019

Weathersit

- Demand is more when the climate is clear and without raining

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

Answer:

Dummies will be generated for all values in respective categorical column. Since a dummy value for one categorical value will be redundant, we'll be using drop_first = True to drop the first column.

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

Answer:

With atemp column

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

Answer:

Error terms should be normally distributed. calculating residual by subtracting y_train and y_train_pred and plotting histogram to check the normal distribution

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:

Yr – From year to year demand is getting increased,

Workingday – In a working day, the demand is getting increased by 0.563

Sunday – On Sundays, demand is getting increased by 0.0665

General Subjective Questions

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

Answer:

- Reading and understanding the data
    - Cleaning the data
    - Visualising
- Performing Linear regression
    - Creating predictor and target variables
    - Create Train and test sets
    - Training model on train set
    - Fitting the model
    - Getting the summary
        - Identify different parameters such as R-squared, F-statistic, Prob(F-statistic) to find the overall model fit is significant or not and variance in the data

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

Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built

It tells us about the importance of visualising the data before applying various algorithms to build model

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

Answer:
Pearson's Correlation coefficient is represented as 'r', it measures how strong is the linear association between two continuous variables. Value ranges from '-1' to '+1'. Where value greater than 0 shows positive relation and less than 0 shows negative relation. 0 indicates no relation between variables

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

Answer:

Scaling is making other variables to keep in comparable values . If the scaling isn't done some of the coefficients obtained might be large or small compared to others.

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. Will be useful when we've outliers in data

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

Answer:

VIF value used to find if the variables are correlated with each other. The higher the value, the more these variables are correlated to each other

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

Answer:

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc