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:

   From the analysis below can be derived:

   a. **spring have less 'cnt' while summer, fall and winter have some more cnt and among these, the medians lie almost nearer to each other**
   b. **There is an increase on 'cnt' duing middle of year compared to start and end of the year**
   c. **all the weekdays have same median but 'Tue' have less overall 'cnt' compared to others**
   d. **when weather situation is light rains, the count is very less compared to mist and clear weather situations**

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

   While creating dummy variables we will create a column for each level of the categorical variable and give 1's and 0's as their values. We can identify the nth column value with the n-1 column values as all 0's of n-1 columns represent the nth column .

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

   'Temp' and 'atemp' columns have highest correlation with 'cnt'

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

   a. **Y checking the p-values and VIF of the features**
   b. **By taking the residuals from the predicted values of model and then created a distplot to see for normal distribution.**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer:
   a. **Positive contributions: 'yr - 2019', 'Sat' weekday, 'workingday'**
   b. **Negative contribution: 'Light' weather situation, 'spring' season, 'windspeed'**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. It uses the concept of Linear line (Y=mX+c**) concept to derive the best-fit model of the given data.**

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

This is designed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties

3. What is Pearson's R?

Answer:

the Pearson correlation coefficient (PCC), also referred to as Pearson's *r*, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.

It is essentially a normalized measurement of the covariance.
The value will always be between −1 and 1.

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

Answer:

Scaling is a measure of converting the data to common scale. SO, that the model can learn data properly.

There are many types of scaling and most popular are:

a. **Standardization: Converting the values based on means and standard deviation**

   Formulization: (x – mean(x)) / (SD(x))

b. **Min-Max scaling (normalized): Normalizing the values based on min and max values. So, the values will always be in between 0 and 1.**

   Formulation: (x – min(x)) / (max(x) - min(x))

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

Answer:

If there is a perfect correlation between two features (variables) then the VIF will be infinite.

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption