Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

- Category variable like season, weather, month are affecting the count of booking in bike sharing.
- In the month of September, in winter and summer season and when cloud is clear and less snow, there are more booking.

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

Answer:

The drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Also, model would be more performant.

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

Answer:

Temp vs cnt is highly correlated.

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

Answer:

We will check it through the test data to validate our assumptions.

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

Answer:

Temp, year, light snow weather

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail

Answer:

As name suggested linear regression is based on correlation of two variable x, y. Based on the we predict how two variable is linearly dependent on each other. The basic formula for the linear equation is –

Y = mx + c, where m is the slop and c is the intercept. X is the independent variable and y is the dependent variable.

Question 2: Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Question 3: What is Pearson's R?

Answer:

The Pearson's correlation coefficient varies between -1 and +1. Pearson's r is a numerical summary of the strength of the linear association between the variables.

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

Answer:

Scaling is the process to normalize the numeric variable with different range and magnitude. Different magnitude can affect the coefficient of variable hence can provide wrong result in modelling so it is required.

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

Answer:

If there is perfect correlation the VIF is infinite, since R square is 1.

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

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.