Contents

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)

The categorical variables that are present in the dataset are:

i. Season
ii. Weekday
iii. Weather situation
iv. Year
v. Month
vi. Holiday

vii. Working Day

Please find below the effect on each of these variables on the target variable i.e., Count.

Categorical Variables	Effect on dependent variable
Season	During Summer and Fall seasons the average users of the lending bikes are more
Weekday	Weekday doesn't seem to be having an impact on the target variable
Weather Situation	Clear sky attracts more users than Mist. Whenever it snows, the usage or the count is minimal
Year	Year 2019 has a greater number of lending bike cases compared to Year 2018
Month	Months June, July, August and September or the season of summer is the peak season of lending bikes
Holiday	If it is a holiday, count is lesser compared to a non-holiday
Working Day	There is more usage of lending bikes on a weekday compared to a non-working day

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

When we create dummy variables for a categorical variable, there will be as many as the unique values of the categorical variable will be offered to be added.

For, eg., Season had 4 unique values. Hence 4 dummy variables were suggested.

Each unique combination of these variables will correspond to the unique value of the original categorical variable. In this case season. Even if we ignore the first dummy variable, we can still maintain the unique combinations that can represent each unique values of the categorical variable.

i.e., all the other dummy variables except first will have zero to represent the first value of the categorical variable.

Hence the first variable becomes redundant and by dropping that we can avoid adding one unnecessary variable to the training model. This will further increase the processing power of the processor.

It is important in order to avoid multi collinearity between the dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

General Subjective Questions