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 :

1. Demand for bikes has increased from 2018 to 2019. 100 % hike against 2018.
2. Month plays very important role as Jan- Jun there is very high demand for the bikes.
3. Demand for bikes is high during the fall season and next will be summer, winter, and spring seasons.
4. Demand for bikes increase with weather.
5. Demand on working days is slightly higher. However it is not that much correlated.

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

Answer:

If you use drop_first=TRUE than it will create N-1 columns which will create and we can encode categorical variable with less dimensions.

n dummy variable creation if we will not use drop_first then it will create all separate columns for all levels of categorical variables but when we set it as True then N-1 columns will create where N is total level of categorical variable. Thus, we can encode categorical variables with less dimensions.

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

Answer:

Temp and atemp has equally high correlation with target variable cnt

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

Answer :

1. check linearity in between dependent and independent variables
2. Calculate residuals (y_train-y_train_pred
3. Check variance of residual

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

Answer : weathersit, year and season