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

The categorical variable in the dataset were season, year, holiday, weekday, working day, and weathersit and month. These were visualized using a boxplot. These variables had the following effect on our dependant variable.

♣ Season - The boxplot showed that spring season had least value of count whereas fall had maximum value of count. Summer and winter had intermediate value of count.

♣ Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was’ Clear, Partly Cloudy.

♣ Year - The number of rentals in 2019 was more than 2018

♣ Holiday - Rentals reduced during holiday.

♣ Month – In September there was highest no of rentals while in December it was least.

♣ Weekday - The count of rentals is almost uniform throughout the week

♣ Workingday – The median count of users is almost constant throughout the week.

1. **Why is it important to use drop\_first=True during dummy variable creation?**

It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. It helps in reducing dimensionality.

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

Looking at pairplot it can be seen that , “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

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

The following tests were done to validate the assumptions of linear regression:

1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.

2. Secondly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.

3. Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

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

1. temp - coefficient : 0.4581

2. yr - coefficient : 0.2343

3. weathersit\_Light Snow & Rain - coefficient : -0.2878