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)

Answer: From exploratory data analysis performed, we have arrived at the below inferences regarding the effect of categorical variables on the dependent variable ('cnt' – Rental count).

1. yr - 2019 had much higher rental count than 2018, so there is a year-on-year increase in demand.
2. season – Fall season registered the highest rental count. We can observe that the count increases as the season cycles from spring to summer, then summer to fall. It peaks in fall, and reduces in winter. The same trend is observed in both the years.
3. holiday - Median rental count is lower on holidays. It may be because people may go for family outings during holidays, for which rental bikes may not be an ideal mode of transport.
4. workingday - Working day/weekend does not seem to have much effect on rental count.
5. weekday - Median rental count is more or less the same across all days of the week.
6. mnth - Rental count keeps on increasing as we move from January to June. The count somewhat plateaus in the months from June to September, and then start reducing as we move from October to December. In 2018, the count was highest in June, whereas in 2019, it was the highest in September.
7. weathersit – Adverse weather situation has a significant negative impact on rental count. Weather situations involving snow, rain, thunderstorms see the count dipping a lot. This is quite obvious and expected, as a bike does not offer any form of weather protection unlike a car.

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

Answer: While creating dummy variables for categorical variables, if there are 'n' categories in the variable, we need to create 'n-1' dummy variables. If all 'n-1' variables have value 0, then it represents the nth variable. For example, in this assignment, we have 4 seasons. So, we have to create 4-1=3 dummy variables. Why not 4? Because, if all 3 season dummy variables have value as 0, it automatically represents the fourth season, as dummy variables can have either 0 or 1 only as values.

Pandas has the get_dummies function which can automatically create dummy variables for the specified column. And 'drop_first=True' is a parameter that is used in this get_dummies function to ensure that only n-1 variables are created, by leaving out the first column.

Example code: pd.get_dummies(df['season'], drop_first = True).

This code will drop the first value from the column 'season' as we have specified drop_first=True.

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

Answer: From the pair plot, it can be observed that variables 'temp' and 'atemp' both have the highest correlation (equally high positive correlation coefficient of 0.63) with the target variable 'cnt'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: The individual assumptions have been validated as follows:

(a) Error terms are normally distributed – The residuals obtained were plotted using a distplot to check and validate that they are normally distributed with mean as zero. A perfect normal distribution curve was observed.

(b) Linear Relationship between X & Y – Linearity exists between the variables as we have obtained positive/negative coefficient values in the final fitted line equation.

(c) Error terms are independent of each other – We have created a scatter plot of the error terms, and are not able to identify any pattern followed. Hence, we can confirm that the error terms are independent of each other.

(d) Error terms have constant variance (homoscedasticity) – We have created a scatter plot of the residuals against the dependent variable. No visible pattern was observed, which confirms that the error terms are homoscedastic.

(e) Multicollinearity does not exist in the model – We have validated this assumption by checking the VIF values of the predictor variables in the final model. All the VIF values in our model are less than 5, so this assumption stands valid.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: As per the final model, the top-3 features affecting the target variable 'cnt' are as listed below:

1. temp, which has coefficient value of 0.407
2. light snow (weather situation), which has coefficient value of -0.288
3. yr (year), which has coefficient value of 0.235