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)

Answer 1:

I can infer below effects on the dependent variable (count of bike bookings) based on the analysis of Categorical variables:

1. The bike booking has **increased** substantially in **year** **2019** than **2018**.
2. There are **more** bookings in **fall season** than other seasons.
3. Bookings are **increased** in **September month** by **0.0767** unit.
4. Bookings **reduce** are done in **July month** by **0.0524** unit.
5. On **holiday** there is **decrease** of **0.0980** units of bike bookings.
6. Every time the **weather** is **Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** the bike booking is **reduced** by **0.2852** units.
7. When **weather** is **Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist** there is **decrease** of **0.0816** units in the bike bookings.
8. **Winter season** there is **increase** of **0.0831** units of bike bookings.
9. **Summer season** there is **increase** of **0.0453** units of bike bookings.
10. **Spring season** there is **decrease** of **0.0669** units of bike bookings.

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

Answer 2:

1. In Linear Regression categorical variables are converted to numerical variables by dummy variable creation.
2. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. When all new dummy columns have value 0 it indicates nth level exist implicitly.
3. This is done by get_dummies() function of pandas library which uses parameter **drop_first**, when set True **will not create new dummy column for level 'n'** and is by default False**.**
4. **This will reduce correlation of one dummy column with target variable.**

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

Answer 3: 'temp' variable has the highest correlation with the target variable.

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

Answer 4:

I validate the assumptions of Linear Regression after building the model of the training set based on below factors:

1. **Normality of Error terms:** Error terms should be normally distributed.
2. **Multicollinearity:** Detecting associations between predictor variables and removing it
3. **Linearity:** There should be linear relationship between feature variable and target variable.
4. **Homoscedasticity:** Error terms should have constant variance or there should be no pattern in residual values.
5. **No auto-correlation:** Error terms should be independent of each other

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 5:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes, based on final model:

1. temp
2. year
3. weather_light (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer 1:

2. Sa
3. Sa
4. Sa
5. Sa
6. Sa
7. Sa
8. sa