**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:**

Based on the analysis of the categorical variables from the dataset, the following inferences can be made about their effect on the dependent variable:

* The **year** appears to have a significant impact on bike demand. There was a noticeable increase in bookings from 2018 to 2019, indicating a positive trend in business growth.
* The **season** also influences bike demand. The fall season, in particular, seems to attract more bookings. Furthermore, there is a significant increase in bookings from the start of the year until mid-year (May to October), after which the trend starts to decrease.
* **Weather conditions** play a crucial role in bike demand. Clear weather conditions tend to attract more bookings, which is an expected observation.
* The **day of the week** has an effect on bike demand. Thursdays, Fridays, Saturdays, and Sundays have a higher number of bookings compared to the start of the week.
* **Holidays** also influence bike demand. There are fewer bookings on holidays, which could be attributed to people preferring to spend time at home with family.
* The demand for bikes does not seem to differ significantly between working days and non-working days.
* Other factors such as **temperature**, **wind speed**, and specific **months** (December, January, November, September) and **weather types** (Snowy, Misty) also appear to influence bike demand.

These observations provide valuable insights into the factors that influence bike demand and can be used to make informed business decisions and strategies.

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

**Answer:**

The drop\_first=True parameter in dummy variable creation is used to avoid the issue of multicollinearity, which can occur when one variable can be perfectly predicted by another.

In the context of dummy variables, if we have n categories for a feature and we create n dummy variables for them, we actually introduce perfect multicollinearity. This is because the value of one dummy variable can be perfectly predicted from the others. For example, if we have a feature ‘color’ with categories ‘red’, ‘blue’, and ‘green’, and we create three dummy variables for them, knowing the values of two will automatically tell us the value of the third.

This perfect multicollinearity is problematic because it can:

* Make the coefficients/weights unstable and their interpretation misleading.
* Lead to redundancy as we are essentially adding an extra variable that doesn’t provide new information.

By using drop\_first=True, we drop the first dummy variable, reducing the n dummy variables to n-1, which eliminates this issue of multicollinearity. The dropped category’s information is not lost but is rather used as the reference category against which the others are compared. For instance, in the ‘color’ example, if we drop ‘red’, a ‘0’ in both ‘blue’ and ‘green’ would indicate ‘red’.

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

**Answer:**

‘temp’ variable has the highest correlation with the target variable.

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

**Answer:**

After building the linear regression model on the training set, I validated the assumptions of the model in the following ways:

1. **Normality of Residuals**: I checked whether the residuals (error terms) are normally distributed. This is a fundamental assumption of linear regression and can be validated using methods like Q-Q plots or statistical tests like the Shapiro-Wilk test.
2. **Absence of Multicollinearity**: I verified that there is no significant multicollinearity among the predictor variables. Multicollinearity can inflate the variance of the regression coefficients, making them unstable. Tools like the Variance Inflation Factor (VIF) can help detect multicollinearity.
3. **Linearity**: I confirmed that there is a linear relationship between the independent and dependent variables. This can be visually checked using scatter plots.
4. **Homoscedasticity**: I ensured that the residuals have constant variance at every level of the predictor variables. This is known as homoscedasticity. If the plot of residuals versus predicted values shows a funnel shape, it indicates heteroscedasticity, violating this assumption.
5. **Independence of Residuals**: I checked for autocorrelation in the residuals, which means that the residuals are not independent of each other. The Durbin-Watson test is commonly used to detect autocorrelation.
6. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

**Answer:**

According to the final model, the top 3 features that contribute significantly towards explaining the demand of shared bikes are:

* **temp**
* **weathersit\_Light Snow/Rain**
* **year**

**General Subjective Questions:**