Name - Anusha Kogta
Email - anukogta008@gmail.com

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?

Ans - In the bike sharing dataset, the effect of the categorical variable ‘weathersit’, 'season', 'mnth', 'weekday'on the target variable ‘cnt’. While performing EDA, I visualized the relationship between the categorical variables and the target variable and it was seen that the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being 50,000 approximately.Also, during model building the summer season the bike rental is more and less in spring season and also there is more bike rental in month of september and least in month of july and yr we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

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

Because drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Let’s say we have 3 types of values in Categorical column named as 'furnished', ' Semi-furnished','Unfirnished' and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.
Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.
Eg-

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

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

Ans - The numerical variable ‘atemp’ and 'temp' has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, we drop 'atemp' due to multicollinearity the numerical variable ‘temp’ has the highest correlation with the target variable ‘cnt’.

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

Ans - Basically Linear Relationship means that there exists a linear relationship between the dependent variable and the predictors.
However it can be verified by -
1- Pair-wise scatterplots - may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.
2- Homoscedasticity - means that the residuals have constant variance no matter the level of the dependent variable.So to verify homoscedasticity, one may look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.
3- Absence of Multicollinearity - It refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case).So to verify it the first step is to check the
a-Pairwise correlations could be the first step to identify potential relationships between various independent variables.
b- We could also check it by looking at the Variance Inflation Factors (VIF). It is calculated by (VIF = 1/1-R^2),Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large R-squared for the regression and thus a larger VIF. As a rule of thumb, VIFs scores above 5 are generally indicators of multicollinearity
4- Normality of Errors - If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.Now to verify the normality of error, an easy way is to draw the distribution of residuals against levels of the dependent variable. One can use a QQ-plot and measure the divergence of the residuals from a normal distribution. If the resulting curve is not normal (i.e. is skewed), it may highlight a problem.

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

Ans - Temp is the most significant with the largest coefficient. and also the bike rentals is more for the month of september and less month of july and bike rental are more in season summer and winter and less in season spring and also we saw the rentals reduce during holidays and also the bike rental growth are increased year wise
Hence This indicates that the bike rentals is majorly affected by **Temperature, Season and Month.**