**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)**

June, July has higher count of bikes are rented

Summer and fall has higher count of bikes

Partly cloudy has higher count of bikes

**2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)**

drop\_first=True is used, to reduce the additional columns created during dummy variable creation. It reduces the correlations created by dummy variables. For a categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

1. count is positively correlated to casual and registred features
2. Relation between cnt to independent variables: count and temp are correlated
3. relation between independent variabels: count is negatively correlated with windspeed, so we can say during windy days demand is somewhat reducing.
4. atemp and temp are positively correlated Hence these are exhibiting multi collinearity. Have to remove these terms

**4. How did you validate the assumptions of Linear Regression after building the model on the**

**training set? (3 marks)**

1. Absence of multicollinearity
2. Homoscedasticity
3. Absence of linear relationship

**5. Based on the final model, which are the top 3 features contributing significantly towards**

**explaining the demand of the shared bikes? (2 marks)**

Humidity, wind speed and holiday