**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?**

* Month wise demand of share bikes stands minimum in the month of February
* Median value of demand of shared bikes is maximum in fall season followed by summer, winter and spring.
* Peak demand of shared bikes is in the month of September.
* Year 2019 has the maximum amount of demand in shared bikes compared to 2018
* Median value of demand on a non-holiday day is more compared to a holiday.
* Average and peak count of demand on clear cloud day is maximum followed by mist cloud and light snow.

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

Using drop\_first=True during dummy variable creation is important to avoid multicollinearity in linear regression models. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, which can lead to several issues:

* Model Interpretation: Multicollinearity makes it challenging to interpret the individual impact of each predictor variable on the target variable. It becomes difficult to discern which predictor is truly contributing to the model's predictions.
* Model Coefficients: In the presence of multicollinearity, the coefficients of the correlated variables can become unstable and sensitive to small changes in the data. This instability can make it challenging to trust the results of the regression analysis.
* Inflated Standard Errors: Multicollinearity tends to inflate the standard errors of the coefficients. Larger standard errors lead to wider confidence intervals, making it harder to determine if a coefficient is statistically significant.
* Reduced Predictive Power: Multicollinearity can reduce the predictive power of the model because it introduces noise and instability into the relationship between predictors and the target variable.
* By setting drop\_first=True during dummy variable creation, we are essentially omitting one of the categories (levels) of the categorical variable as a reference category. This reference category is represented by zeros in the dummy variables for the other categories. Omitting one category helps mitigate multicollinearity because it ensures that the dummy variables for the remaining categories are not perfectly correlated.

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

* After dropping variable registered and casual, highest correlation with respect to Y is atemp or temp variable. Temp and atemp is also mutually correlated with each other, hence any one variable either atemp or temp can be considered.

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

* P-value of each variable must be less than 0.05
* VIF of the independent variable must be less than 5
* R-Square and adjusted R-square of the model must be as high as possible considering the above criteria.
* Verifying or variable selection through recursive feature elimination and re-iteration.
* Error terms must be normally distributed.

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

Below are the top 3 features explaining the demand for shared bikes:-

* + 1. Ratio of Casual to Registered (derived metric)
    2. Fall
    3. Year