# 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. Season Weather Situation holiday, month, Working day, and weekday were the categorical variable in the dataset. A boxplot was used to visualise these. These variables influenced our dependent variable in the following ways:

1.**Season:** The boxplot revealed that the spring season had the lowest value of count, while the fall season had the highest value of count. Summer and winter had count values that were in the middle.

2.**Weather:** Situation: When there is heavy rain/snow, there are no users, indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was 'Clear, Partly Cloudy.

3.**Holiday:** Rentals were found to be lower during the holidays.

4.**Month:** September had the most rentals, while December had the fewest. This observation is comparable to the one made in weathersit. The weather in December is typically cold and snowy.

5.**Weekday:** Weekends saw a significant increase in book hiring compared to weekdays.

6.**Working day:** It had little effect on the dependent variable.

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

Ans. When creating dummy variables, it's essential to eliminate the first column to avoid redundancy and correlated features. Failing to do so can adversely affect certain models, particularly when dealing with low cardinality. Models that rely on iterations may struggle to converge, and assessments of variable importance may become skewed. Additionally, maintaining all dummy variables introduces multicollinearity issues among them. By discarding one column, we effectively manage and mitigate these challenges, ensuring a more stable and accurate representation in our models.

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

Ans. Upon examining the distribution of residuals through plotting, it was observed to conform to a normal distribution, centering around a mean value of 0.

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

Ans. The Following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:

Temp (0.4910) -
Year (0.2336)
Weathersit Light Snow (-0.2842)

# General Subjective Questions

1. Explain the linear regression algorithm in detail.
Ans. Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation.

**"y=mx+c"**

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression**: SLR is used when the dependent variable is predicted using only one independent variable.

2. **Multiple Linear Regression:** MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$y_i=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p$$

β1 coefficient for XI variable

β2 coefficient for X2 variable

β3- coefficient for X3 variable and so on....

β0 is the intercept (constant term)

2. What is R-squared? What is adjusted R-squared?

Ans. R-squared (R2) is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It is a scale from 0 to 1, where a value of 1 indicates that the model explains all the variability of the dependent variable, and a value of 0 indicates that the model does not explain any variability. R-squared is calculated as the ratio of the explained variance to the total variance in the data.

Adjusted R-squared, on the other hand, is a modified version of R-squared that accounts for the number of predictors in the model. While R-squared tends to increase as more variables are added to the model, adjusted R-squared penalizes the inclusion of unnecessary variables that do not contribute significantly to explaining the variance in the dependent variable. The adjusted R-squared value is always lower than the R-squared value, and it increases only if the new variable improves the model more than would be expected by chance.

In summary, R-squared measures the overall goodness of fit of the model, while adjusted R-squared adjusts for the number of predictors, providing a more accurate assessment of the model's explanatory power.

3. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a preprocessing step in data analysis and machine learning that involves transforming the numeric values of variables to a standard range or distribution. The goal is to ensure that all variables have a similar scale, preventing certain features from dominating or being neglected based solely on their original magnitude.

In short, the primary difference between normalized scaling and standardized scaling lies in the transformation range. Normalized scaling (Min-Max scaling) confines values to a specific range, often between 0 and 1, preserving relative relationships. Standardized scaling (Z-score scaling) transforms data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms assuming a Gaussian distribution. Normalized scaling is beneficial when distribution characteristics matter, while standardized scaling is robust to outliers and aligns with Gaussian assumptions. The choice depends on the specific needs and characteristics of the data.

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. The formula for VIF is: $$VIF_i = \frac{1}{1 - R_i^2}$$

Basically, if R square is 1 than VIF becomes infinite. It means that there is perfect correlation between the features.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. A Quantile-Quantile (Q-Q) plot is a scatter plot that compares the quantiles of two datasets to assess if they share the same distribution. The main objective is to visually inspect whether the data in question originated from a common source. When the datasets come from the same distribution, the Q-Q plot takes the form of a straight line, serving as a visual confirmation of their similarity. Essentially, it provides an intuitive means to check and compare the distributions of two sets of data.