# **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:  Categorical variables play a vital role in analyzing the demand of bike sharing. There are seasonal variables, month, weekday and weather sit categories which are affecting the dependent variable. After analyzing the data I can able to say that under season category summer and winter are affecting the target variable positively. And month wise we have July and September month playing significant role in affecting dependent variable. Under weather sit light snow and Mist also creating some impact in dependent variable.

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

Answer: Using drop_first during the dummy variable creation is important because it helps is in reducing one extra column and it also helps us in reducing the multicollinearity. Dropping the first category makes the interpretation easy and simple.

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

Answer:  Numerical variable “temp”(Temperature) have the highest correlation with the target variable “cnt” with 0.64.

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

Answer:  After building the Linear regression model we can validate our assumptions using the residual analysis using the residual plot which show us the pattern on how data distributed with mean zero. We can also validate using the multicollinearity. Correlation and variance inflation factor (VIF) also help us validate the assumptions of linear regression model.

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

Answer:  The top 3 features controlling the significance towards explaining the demand of the shared bikes are as follows:

I. Temp (Temperature)
II. Yr (Year)
III. Weather sit light snow
‘Temp’ and ‘Yr’ is positively affecting the demand for bike sharing. Weather sit light_snow is negatively affecting the bike sharing demand.

## General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a part of regression analysis. Linear Regression Algorithm is a machine learning algorithm based on supervised learning. In linear regression we train a model to predict the behavior of the data based on some variables. In training model we will analyze the independent variables which are affecting the target variable. For example in our bike sharing data we are determining the variables which are affecting the target variable "Count".

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a collection of 4 datasets that have identical summary statistics but exhibit different patterns when graphically visualized. The datasets showcase linear, non-linear, and outlier-influenced relationships, highlighting the need for comprehensive analysis beyond numerical summaries. It also value to the importance of using data visualization to see trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R?

Answer: Pearson's R is a measure of the linear relationship between two numerical variables, ranging from -1 to +1. It quantifies the strength and direction of the linear association, with +1 indicating a positive relationship, -1 indicating a negative relationship, and values close to 0 is little or no linear relationship.

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

Answer: Scaling is the process of transforming numerical variables to a specific range or distribution. It is performed to bring variables to a one scale and make potential issues caused by differing measures. Normalized scaling adjusts values to a range between 0 and 1, while standardized scaling transforms variables to have zero mean and unit variance. Normalization preserves the relative relationships between data points, while standardization centers the data around the mean and accounts for variability.

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

Answer: VIF can become infinite when there is perfect multicollinearity in the data. Perfect multicollinearity occurs when one or more independent variables in a regression model can be perfectly predicted by a linear combination of other independent variables. In such cases, the VIF for the variables involved becomes infinite because the estimated coefficient's variance is infinitely inflated due to the redundant information contained in the correlated variables.

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

Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular distribution, such as a normal distribution. It compares the quantiles of the observed data against the expected quantiles of a specified distribution. In linear regression, a Q-Q plot helps validate the assumption of normality for the residuals, allowing us to assess if the residuals are approximately normally distributed and if the linear regression model is appropriate for the data.