SECTION-1: *Assignment-based Subjective Questions*

1. From my analysis of the categorical variables from the dataset what I could infer is – '**season**' and '**weathersit**' plays a significant role in determining the effect on the dependent variable which is '**cnt**' in our case. For example, based on the correlation heatmap generated for the final model dataset it is evident that Spring & Summer has significance when it comes to determining the effect on count of shared bikes. On the similar lines, weather (cloudy/ rainy) has an impact as well on the count of shared bikes

2. When we create dummy variables from the categorical variables we use "**drop_first = True**" to avoid multicollinearity that happens when 2 or more features are highly correlated which can lead to unstable estimations of regression coefficients & thereby making it hard for the interpretation of the importance of feature variables

3. Based on the pair-plot drawn for the numerical variables, the variable 'registered' has the highest correlation with the target variable 'cnt'

4. After building the linear regression model, following techniques had been used to validate the assumptions on the training dataset:
   - Did look at the residual curve which is normally distributed
   - The residual curve is more or less centered around 0 (though not exactly 0 but close)
   - Did compare the r_squared value for both training dataset & test dataset & have compared (both are equal in this case)

5. Based on the final model, the top 3 features contributing towards explaining the demand of the shared bikes – 'registered', 'casual' and 'spring'

SECTION-2: *General Subjective Questions*

1. The aim of the linear regression algorithm is to find the best fitting linear relationship between one or more features (independent variables) and a dependent variable (target) by estimating the coefficients of the linear equation. The intent is to make the predicted value from the linear model as close as the actual values of the target variable thereby reducing the error component and increase the accuracy in predictions from the model. We actually use a training dataset to make the model learn from the data and we use a test dataset to check how accurate the model is in terms of predictions from the unseen data. Certain statistical components like R_squared, p-value, VIF (variance inflation factor) is used to measure how significant a feature is in determining the outcome of the target and thereby helps us in deciding the relevant features that drives the outcome of a model and predicting the value of the target variable

2. Anscombe's quartet refers to a set of four small datasets that have nearly identical statistical properties like mean, variance, correlation, and linear regression line but when visualized, they reveal significant differences in their distributions and relationships. This helps statisticians to get rid of relying only on the summary statistics and hence calls for closer look at the graphical exploration of the data. Usually, each dataset contains 11 data points and when graphed they show diverse patterns. In short, the scatter plots for these 4 datasets look very distinct and hence data visualizations are required to understand the relationships at more granular level.

3. Pearson's correlation coefficient denoted as "r" or "Pearson's r," is a measure in statistics that quantifies the strength and direction of the linear relationship between two continuous variables. It’s value ranges from -1 to 1, where:

- 1 indicates a perfect positive linear correlation: As one variable increases, the other variable also increases proportionally
- 0 indicates no linear correlation: There is no consistent linear relationship between the two variables
- -1 indicates a perfect negative linear correlation: As one variable increases, the other variable decreases proportionally

4. Feature scaling is the preprocessing technique in machine learning to ensure that all feature variables in the dataset are on a similar scale. Here, features get transformed so that they have a consistent range which can benefit certain ML algorithms. Feature scaling especially helps in models that are sensitive to the scale of the input. In case of normalized or Min-Max scaling the transformed data ranges between 0 and 1. In case of standardized scaling (z-score scaling) each feature is transformed in a way that it has mean of 0 with standard deviation of 1.
Feature scaling helps prevent features with larger magnitudes from dominating the learning in models & can help in increasing the accuracy for predicting on unseen data

5. VIF becomes infinite when we encounter a situation in which the variance in the independent variables is extremely high due to strong correlation with other variables in the model hence it becomes difficult to identify which variables is the primary driver to determine the outcome of the target variable which can lead to several issues in statistical analysis

6. A Q-Q plot or Quartile-Quartile plot is a graphical tool in statistics to assess the similarity between the normal distribution and the observed distribution of a dataset. It’s a process to visually determine if the data follows a particular distribution or not. The importance of Q-Q plot is the fact that it helps statisticians and data analysts make better informed choices in the analysis process. It also helps in Normality checks within a dataset which is essential in many statistical techniques like hypothesis testing and linear regression