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?

   In data When we look at fields like weekday, season, mnth, weathersit having values in integer. So, we have converted into categorical form (Text form). Which convey a importance or value for this integer value.
   Eg. We have values in weekday are 0-6, So we converted the into week days.
   For solution we used dummy variables. Dummy variables are used to replace column information into the numbers of columns.
   We can't use categorical variable directly in model.

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

   We create dummy variable for all unique values are present in the columns. So, if the 4 different values are in column, then we add 3 instead 4 columns.
   If 3 value combination is not present in the column the model automatically detects 4 variables.
   drop_first = True we used for the drop one column from dummy variables.

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

   'registered' has the highest correlation with the target variable count('cnt').

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   1. Train the data
   2. Compare the Predicted values and Trained values.
   3. Using Scatter plot between Predicted values and Trained values I have checked the accuracy.

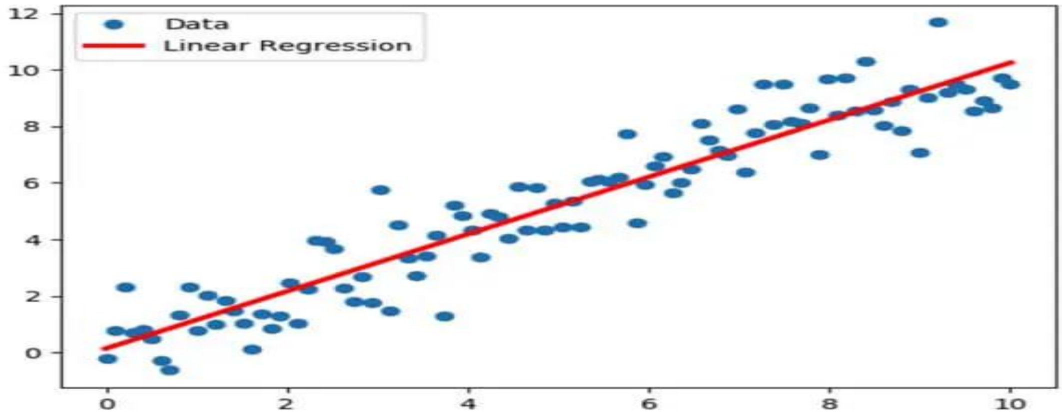
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   1. Sunday
   2. Light_Snow
   3. Holiday

   Because of this feature have low VIF

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   A linear regression model makes a prediction by simply computaing weighted sum of the input features, plus a constant.



   Find the predicted value = intercept + slop * Variable + constant

   In the case Linear regression, the linear regression model learns to find out the optimum values of theta 0 to till n values in order for correct prediction of y given X. The relationship that the algorithm learns the linear regression model.

   In simple word, A ML algorithm learns a function f such that f(x) maps to y. The algorithm learns how to model relationship between X features y target variables.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

   The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson’s R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

Scaling involves converting the numeric values to standard minimized scale
There are two types of scaling

a. Standardization
b. MinMaxScaling

Standardization:

Standardization is another technique where the values are cantered around the mean with a unit standard deviation. Theas means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
The scale values are not restricted to particular range.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

MinMaxScaling:

MinMaxScaling involves scaling the values in each column. Separately and proportionally to values between 0 and 1. These scaled values are floating point numbers corresponding to each original values.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

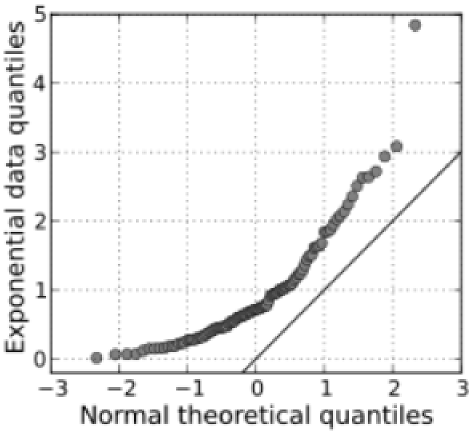
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

$$VIF = \frac{1}{1 - R_i^{\,2}}$$

We are using the formula as shown in the above figure.
For calculation we use R square value is 1 then VIF will get infinity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q-Q plot is called a **normal quantile-quantile (Q-Q) plot.** The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.