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: The effect of categorical variables 'season', weathersit', 'weekday', and 'month' on Target 'count' is visualized and the following conclusions are drawn:

   i) Season:
      (1) 'spring' shows minimum number of median counts among all seasons, but outliers can be observed.
   ii) Weather:
      (1) 'light_snow' shows significantly smaller number of counts among all weathers.
      (2) 'heavy_snow' has 0 counts indicating that services are unavailable during this weather condition.
   iii) Weekday:
      (1) The number of median counts is generally same for every day.
   iv) Month:
      (1) The beginning and ending months of year shows smaller number of counts.
      (2) The number of counts in the mid of the year is (June, July and August) significantly high.

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

   Ans: A dummy variable is a numeric variable that represents categorical data, such as season, days of week, gender, race, marital status, educational qualification, etc. Dummy variables are dichotomous and quantitative variables, technically i.e., their range of values is small and they can take on only two quantitative values. Regression results are easiest to interpret when dummy variables are limited to two specific values, 1 or 0. Where, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

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

   Ans: 'real_temp' has the highest correlation with target 'count'. The graph suggests that 'count' increases with increase in 'real_temp' in a linear fashion.
   Also, the pairplot below descripts the same.

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

   Ans. By observing the distribution of residuals: The mean of residuals should follow a normal distribution with mean equal to zero or close to zero. This is done in order to check whether the selected line is actually the line of best fit or not.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans. Top 3 are
   1. 'real_temp': positive relationship
   2. 'light_snow' weather: negative relationship
   3. year: positive relationship

General Subjective Questions

1. **Explain the linear regression algorithm in detail**

   Ans: Linear Regression is a machine learning technique based on supervised learning, that allows us to associate one or more predictor variables with a dependent variable. All machine learning models try to approximate f(x) [the function that accurately describes the relationship between the independent (predictor) and dependent variables], in Linear Regression, it is assumed that f(x) is linear.

   A linear algebraic function y is given by y = mx + b, where y is the dependent variable, m is the slope, or derivative, and b, is the intercept, or the value of y when x, the predictor variable is equal to 0.

   The goal of linear regression is to obtain a line that best fits the data. In other words, the best suited values for m and b, for which total prediction error (all data points) is as small as possible.

2. **Explain the Anscombe's quartet in detail.**
   Ans: Anscombe generated a quartet of made-up data in the early 1970's. Each dataset has 11 {x, y} pairs of numbers. The means of the x values are almost identical for all four sets and the means of the y values are also almost identical. The summary statistics of the data is observed to be as follows:
   • The average x value is 9 for each dataset
   • The average y value is 7.50 for each dataset
   • The variance for x is 11 and the variance for y is 4.12
   • The correlation between x and y is 0.816 for each dataset
   • A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

3. **What is Pearson's R?**
   Ans: Correlation is a technique for investigating the relationship between two quantitative, continuous variables. Pearson's correlation coefficient (R) or the Pearson's R is a measure of the strength of the association between the two variables.
   It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0

indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Ans: Scaling is a technique to standardize the predictor variables (independent features) present in the data to a fixed range.
   It is performed during the data pre-processing to handle highly varying magnitudes or units. If feature scaling is not done, then a machine learning model tends to weigh greater values, higher and smaller values as the lower values, regardless of the unit of the values.
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Ans: In regression analysis, variance inflation factor (VIF) is a measure or multicollinearity. Multicollinearity is referred to as the correlation between predictors (independent variables) in a regression model, whose presence can significantly affect the results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. Using the R-squared (R2) value, VIF can be calculated by the formula:
   It ranges from 1 and upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) of a regression coefficient is inflated due to multicollinearity in the model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Ans: Q-Q (quantile-quantile) plots are an essential tool to graphically analyse and compare the probability distributions of two datasets by plotting their quantiles against each other. If the two distributions are exactly equal then the points on the Q-Q plot lie perfectly on a straight line (y = x).
   Q-Q plots are helpful in determining the type of distribution for a random variable whether it is a Gaussian Distribution, Uniform Distribution or Exponential Distribution, etc. Also, looking at Q-Q plots can give insights about the skewness and the measure of tailedness of the distribution. If the bottom end of the Q-Q plot deviates from the straight line and the upper end does not, it can be concluded that the distribution is left-skewed (negatively skewed). Similarly, if the upper end deviates from the straight line but not the lower then it is right-skewed (positively skewed).