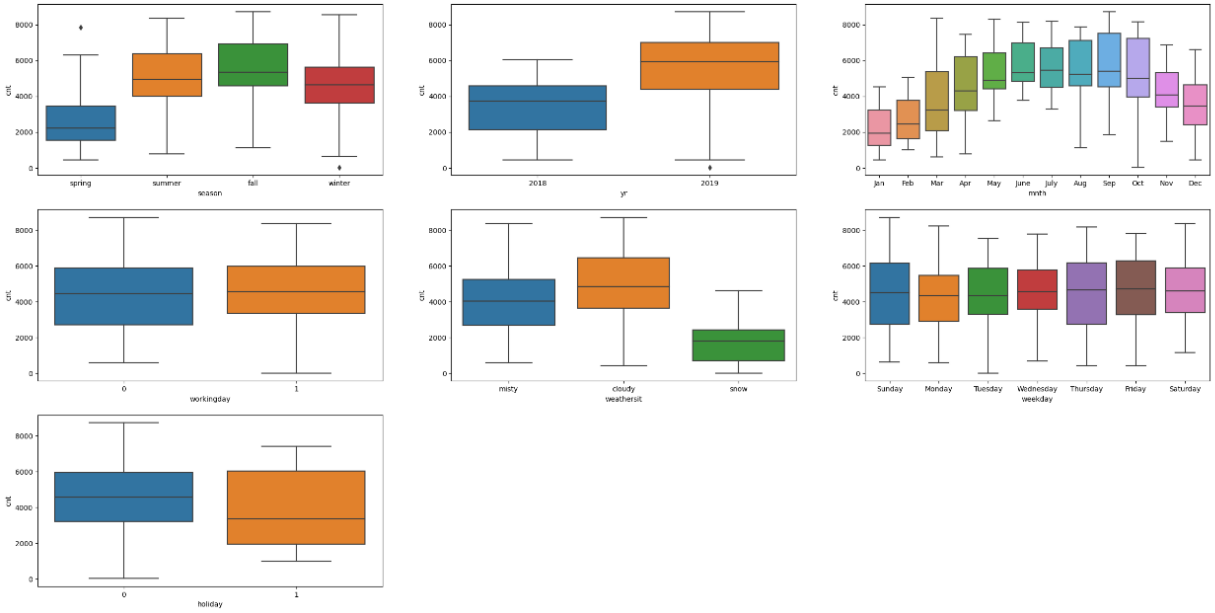
Linear Regression Assignment

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? (3 marks)

Ans: Insights from categorical variable:

Categorical variables include season, year, month, holiday, weekday,weathersit. Target variable is cnt.



- Season - We can notice a positive trend in the number of customers in 3 - Fall, 2 - Summer, and 4 - Winter seasons
- Year - The overall business shows an increasing trend in their user base year on year from 2018 to 2019
- Month - Similar to the season trend, there is a positive trend in the related months of summer, fall and winter.
- Holiday : On holidays, the users show a wider spread in the counts
- Weekday : No major difference in Weekdays or weekends
- Weathersit : Clearer weathers show a positive trend in the number of bike users

  - - 1: Clear, Few clouds, Partly cloudy, Partly cloudy → clo udy

  - - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few cloud s, Mist → misty

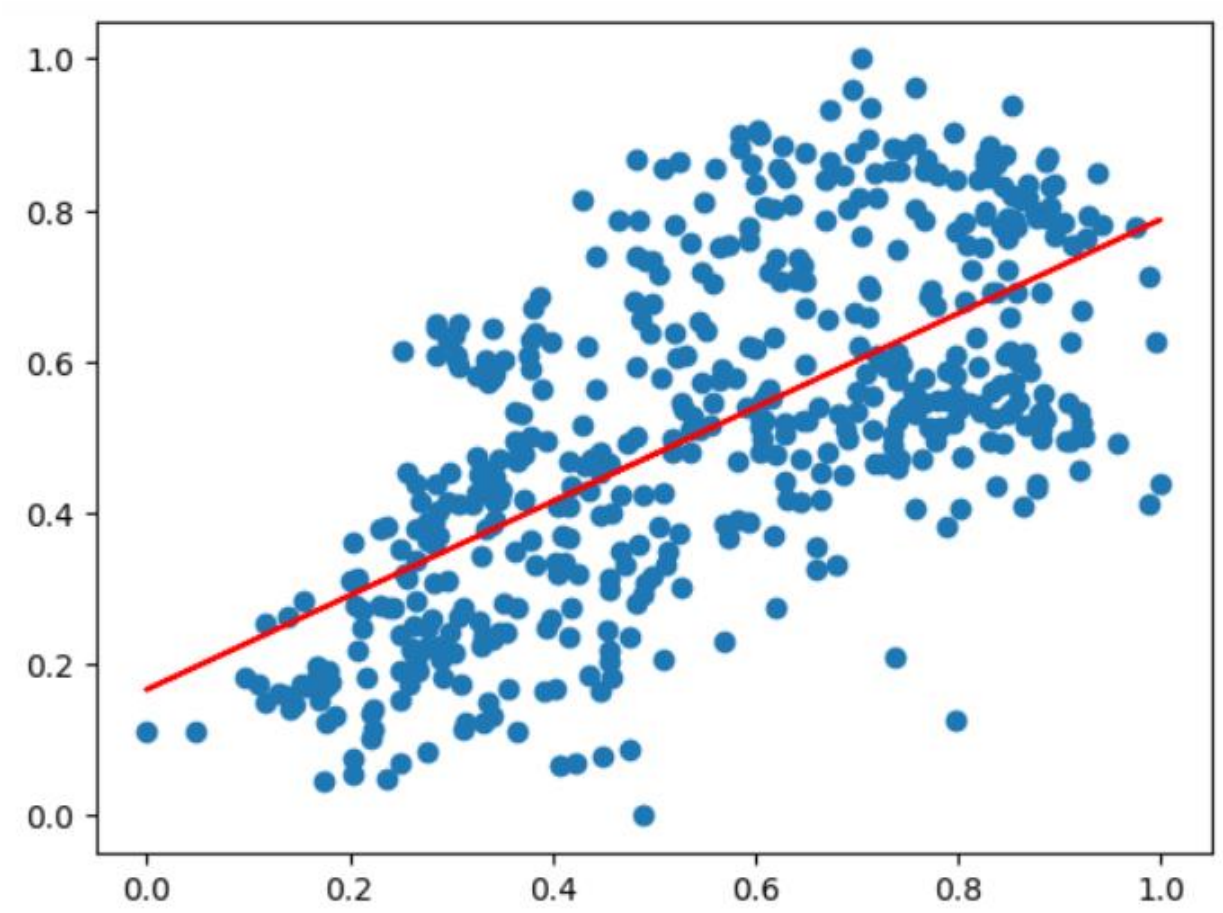
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clo uds, Light Rain + Scattered clouds →snow

    Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog → Rainy

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans: Helps to drop the first column while creating dummy variables where no dummy variables becomes n-1 and also helps to eliminate multicollinearity between dummy variables

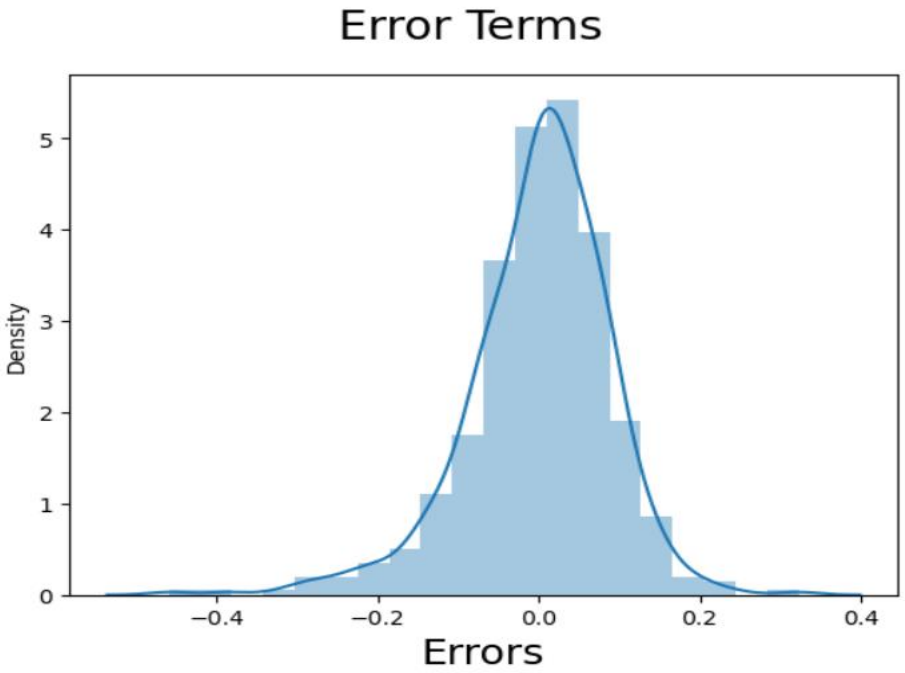
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: atemp and temp are highly correlated with the target variable cnt with value of 0.63

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans: 1. Assumptions are validated by creating a scatter plot between x_train and y_train. Creating a fit line to check the linear relationship between dependent variable and feature variable



2. Error terms are independent of each other – above there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other

3. Error terms are normally distributed: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Ans: The top 3 features for contributing towards demand are temperature,season,yr

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   Ans:

   a. Linear regression is a type of supervised machine learning algorithm to find the best linear relationship between the independent variables and dependent variables.

   b. Linear regression of two types – simple linear regression and multiple linear regression simple linear regression – explains the relationship between a dependant variable and one independent variable. $Y = \beta_0 + \beta_1 X$ is the standard equation of the regression line. β_0 represents slope and β_1 represents intercept

   c. The strength of the linear regression is determined by R^2 and residual standard error

   d. Multiple Linear regression explains the relationship between one dependent variable and several independent variables

   $$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

2. Explain the Anscombe's quartet in detail. (3 marks)

   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. It

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? (3 marks)

   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.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   a. Scaling is used for the preparation of data in building a machine learning model to standardize the independent feature variables to a fixed range
   b. The feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also, the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. On scaling. it affects the coefficients. It does not affect prediction and precision of prediction.
   c. Normalization or min-max scaling changes all the data points between the range of 0 and 1. Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

The value of VIF is infinite when there is a perfect correlation between the 2 independent variables. Formula: $VIF = \frac{1}{1-R^2}$. If R^2 is one then VIF is infinite. It indicates there is a problem in multi-collinearity ad one of these variables needs to be dropped to define the model

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

Q-Q plots are also known as Quantile-Quantile plots. It plots the quantiles of a sample distribution against quantiles of a theoretical distribution. It helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. It helps to check if two populations are of the same distribution.