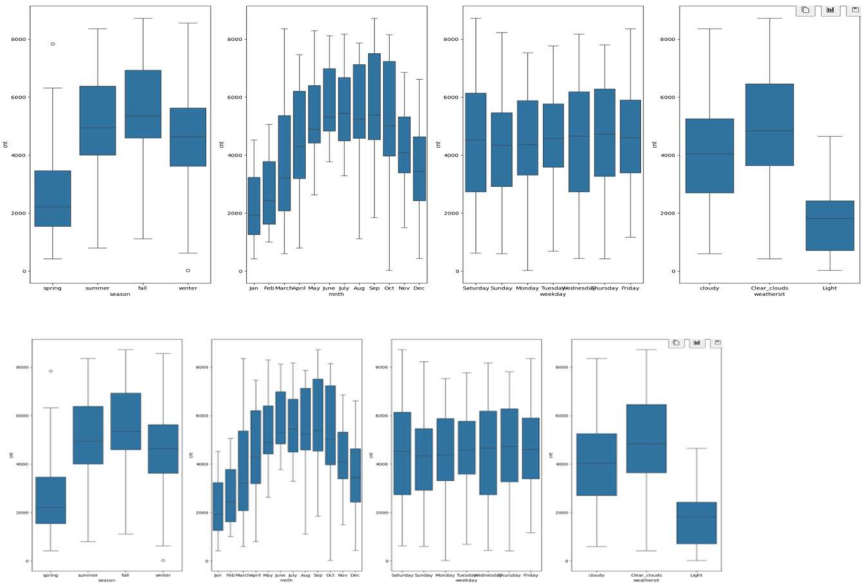
Assignment-based Subjective Questions

1. ***From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?***

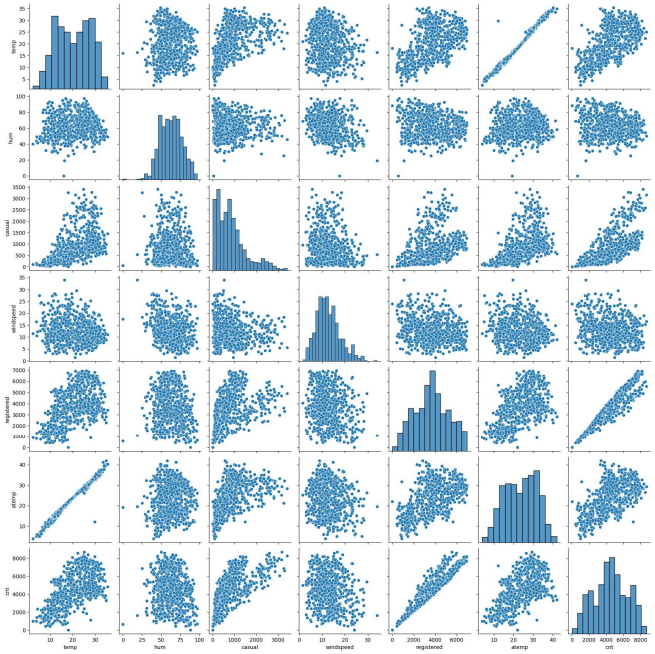


- Season and Month: These variables have a significant impact on bike rental counts, with higher rentals during warmer months (spring and summer) and lower rentals in colder months (autumn and winter). This indicates a strong seasonal pattern in bike rentals.
- Day of the Week: This variable has minimal impact on bike rental counts, with consistent rentals throughout the week, suggesting that bike usage is stable regardless of the day.
- Weather: Weather conditions significantly affect bike rental counts. Clear or cloudy weather conditions are associated with higher rentals, while light rain leads to a noticeable decrease in rentals.

Overall, bike rentals are heavily influenced by seasonal changes and weather conditions, with higher rentals during favorable weather and warmer seasons. The day of the week appears to have a negligible effect on rental counts.

2. ***Why is it important to use drop_first=True during dummy variable creation?***
   By setting drop_first=True, you drop the first dummy variable and create k−1 dummy variables. This reduces redundancy and helps to avoid multicollinearity, ensuring the model matrix is full rank and less complex.

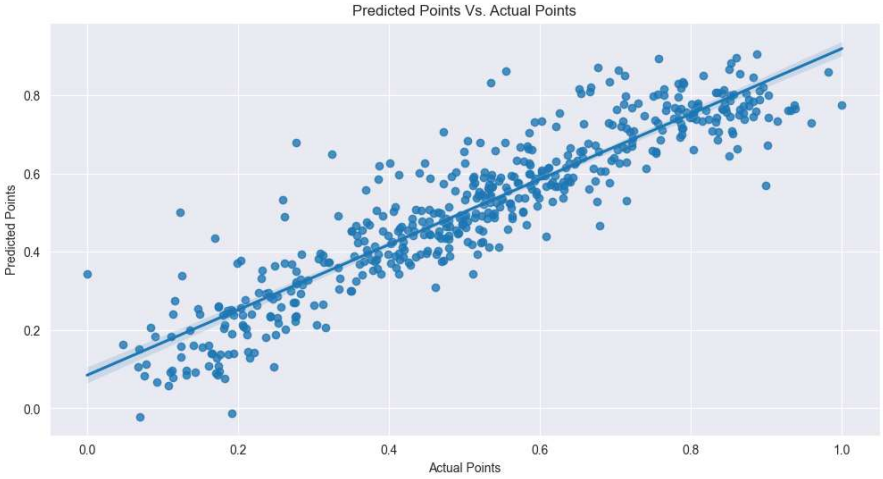
3. ***Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?***



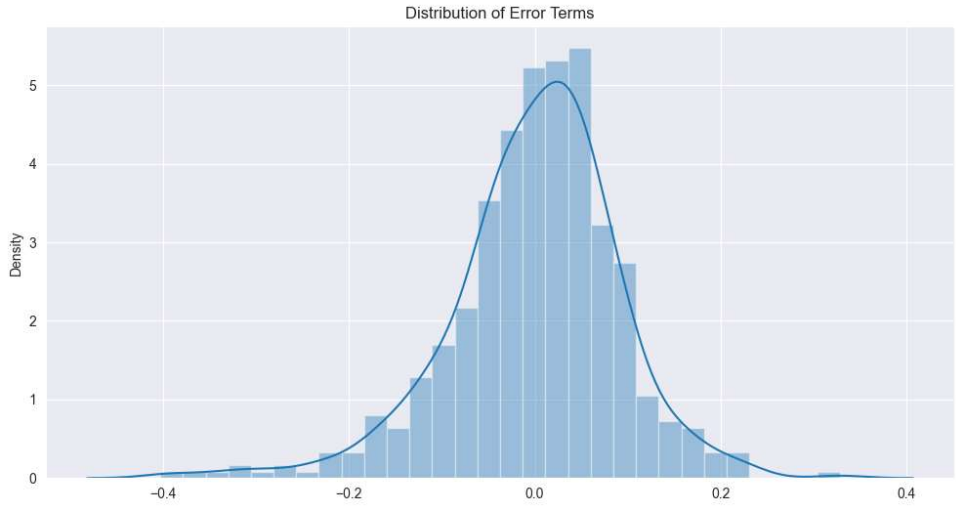
Temp and atemp has the highest correlation
Casual and registered here can be ignored as its derived from target variable

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

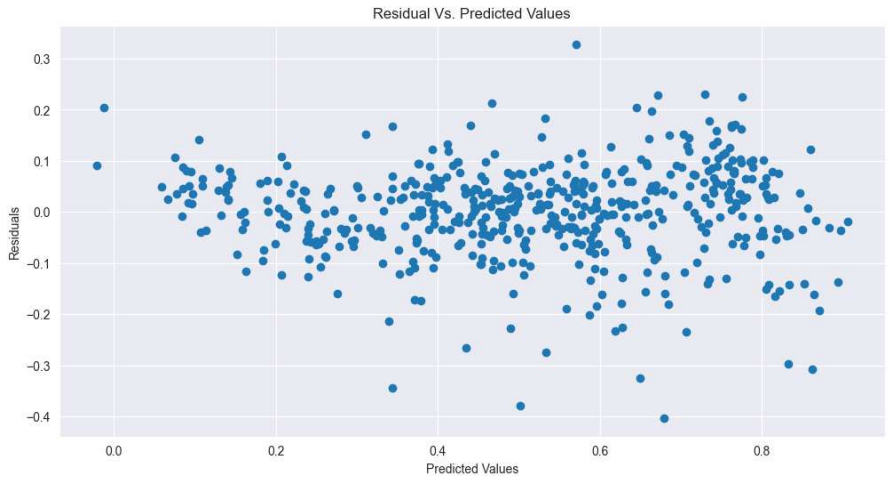
1. **Linear relationship between independent and dependent variables** – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



2. Error terms are normally distributed: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



3. Error terms are independent of each other – We can see 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



5. ***Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?***
   In my model the top 3 features contributing to demand of bike are
   1. Temperature
   2. Year.
   3. Weathersit like Raining, Humidity, Windspeed and Cloudy affects.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Linear regression is a fundamental and widely used algorithm in statistics and machine learning for predicting a continuous target variable based on one or more predictor variables.

The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

There are 2 types of linear regression algorithms

- Simple Linear Regression – Single independent variable is used.
  - $Y = \beta 0 + \beta 1 X$ is the line equation used for SLR.
- Multiple Linear Regression – Multiple independent variables are used.
  - $Y = \beta 0 + \beta 1 X 1 + \cdots + \beta p X p + \in$ is the line equation for MLR.
  - $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient.$

Cost functions – The cost functions helps to identify the best possible values for the $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable.

There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.

- Sum of squared function is used as a cost function to identify the best fit line.

The cost functions are usually represented as

- The straight-line equation is $Y = \beta 0 + \beta 1 X$
- The prediction line equation would be $Ypred = \beta 0 + \beta 1 xi$ and the actual Y is a Yi.
- $Now\ the\ cost\ function\ will\ be\ J(\beta 1, \beta 0) = \Sigma(yi - \beta 1 xi - \beta 0)2$

- The unconstrained minimization are solved using 2 methods
  - Closed form
  - Gradient descent

• While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.

- $ei = yi - ypred$ is provides the error for each of the data point.
- OLS is used to minimize the total e2 which is called as Residual sum of squares.
- RSS = $= \Sigma(yi - ypred)2ni = 1$

• Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. ***Explain the Anscombe's quartet in detail.***

Anscombe's Quartet is a set of four distinct datasets that have nearly identical simple descriptive statistics but very different distributions and graphical properties. The quartet was constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistics alone can be misleading.

Visual Inspection:

- Anscombe's Quartet emphasizes the necessity of visualizing data through scatter plots before performing statistical analyses. Graphical representations can reveal underlying patterns, anomalies, and relationships that summary statistics might miss.

Misleading Statistics:

- The quartet illustrates how datasets with identical statistical properties can have very different distributions and relationships. This underscores the potential for summary statistics to be misleading if used in isolation.

Model Appropriateness:

- The datasets highlight the importance of choosing the right model for data**Impact of Outliers**:
- The presence and influence of outliers can distort statistical measures and models.

3. ***What is Pearson's R?***

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

• -1 coefficient indicates strong inversely proportional relationship.
• 0 coefficient indicates no relationship.
• 1 coefficient indicates strong proportional relationship.

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

Where:
N = the number of pairs of scores

Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx^2 = the sum of squared x scores
Σy^2 = the sum of squared y scores

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

The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
Most of the times 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 into 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. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

- $MinMaxScaling$

  $x=(x-\min(x))/(\max(x)-\min(x))$

Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

- $Standardization$:

  $x=(x-mean(x))/(sd(x))$

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

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

Q-Q plots are graphical tools used to assess whether two datasets come from a common distribution. They can be used to compare an empirical dataset to a theoretical distribution (e.g., normal, exponential, or uniform). In linear regression, Q-Q plots help determine if the training and test datasets are from populations with the same distribution. They can also check if a dataset follows a normal distribution, with the following interpretations:

Interpretations

- **Similar Distribution**: If all data points lie around a straight line at a 45-degree angle from the x-axis, the datasets have a similar distribution.
- **Y-values < X-values**: If the quantiles of y-values are lower than those of x-values, it indicates a potential discrepancy between the datasets.
- **X-values < Y-values**: If the quantiles of x-values are lower than those of y-values, it also indicates a potential discrepancy between the datasets.
- **Different Distributions**: If the data points lie away from the straight line, the datasets have different distributions.

Advantages

- **Comprehensive Distribution Analysis**: A single Q-Q plot can reveal various distributional aspects such as location, scale shifts, symmetry changes, and outliers.

- **Sample Size Representation**: The plot allows for the inclusion of sample size information.