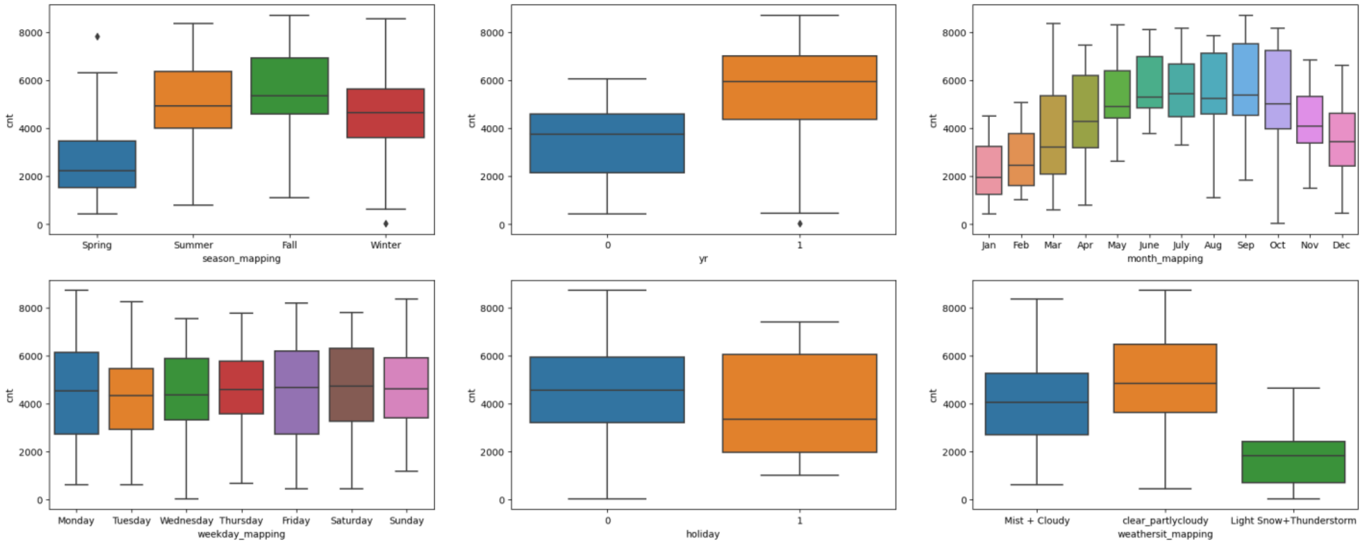
Assignment-based Subjective Questions

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

Answer:

Following figures show the boxplot of dependent variable (cnt) vs different categorical variables- Season, yr, mnth, weekday, holiday and weathersit.



- Higher median values for summer and fall seasons indicate higher number of users during these seasons, as the weather is more favourable.
- Increase in no of users in 2019, as the COVID restrictions would have eased.
- Increase in the no. of users from the months January to September and then it starts decreasing, probably due to winters kicking in.
- Higher no. of users is present on a holiday/weekend. In above graph "0" indicates holiday/weekend.
- More no. of users is observed if the weather is "clear or partly cloudy", compared to when it is misty and cloudy.

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

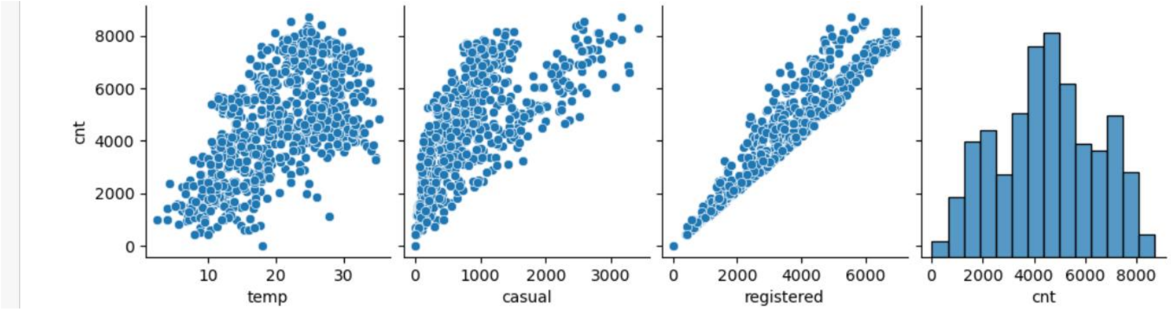
Answer:

The ‘drop_first = True’ is used while creating dummy variables to drop the base/reference category. Including a dummy variable for every category of a categorical variable can lead to multicollinearity, this can be avoided by using ‘drop_first=True’.

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

   Answer:

   Here the variable “registered” has the highest correlation with the target variable. However, the target variable is the sum of casual and registered users, hence it would make sense to exclude them from further analysis. Additional to “casual” and “registered”, the variables “temp” have the highest correlation with the target variable.

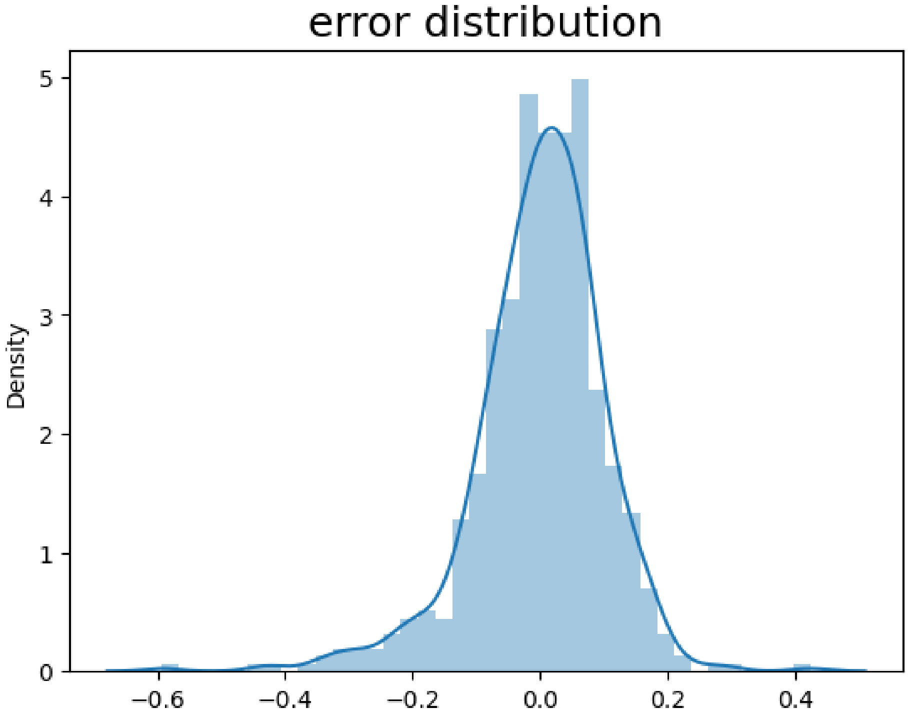


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

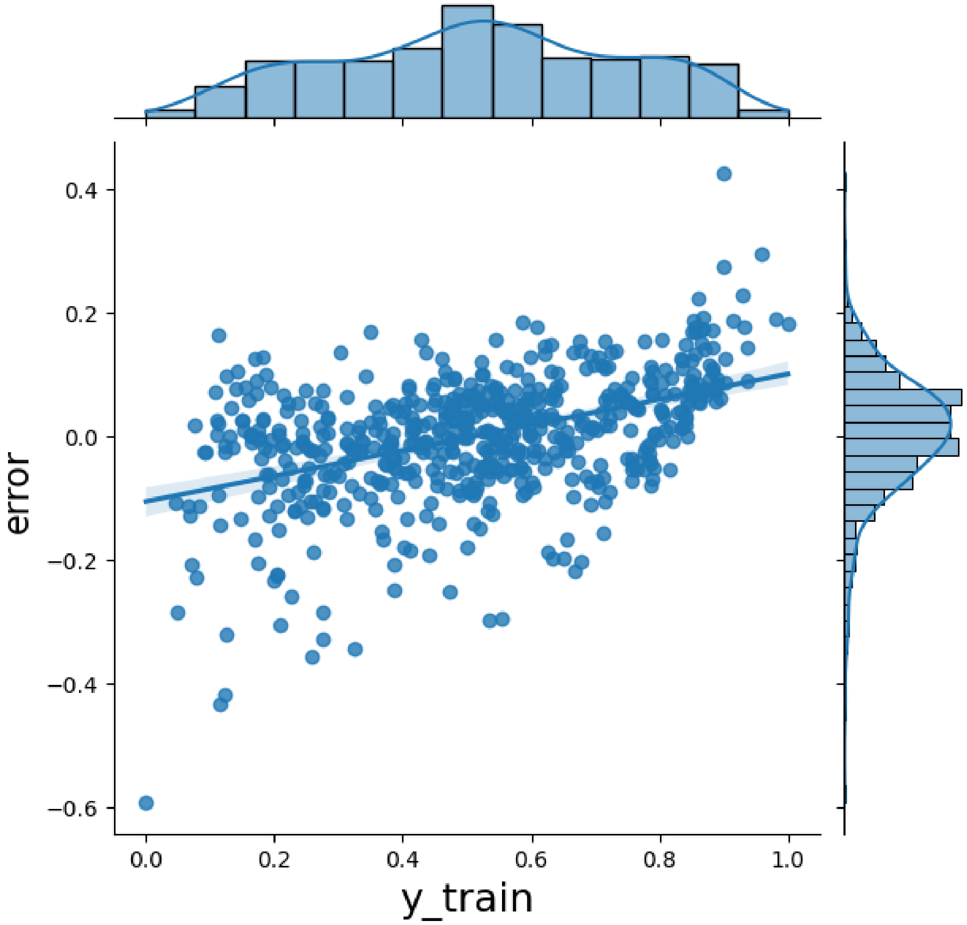
Answer:

It is observed that there is some linear relationship between the target variable and other variables such as “temp” and “atemp”, hence linear regression model is selected.

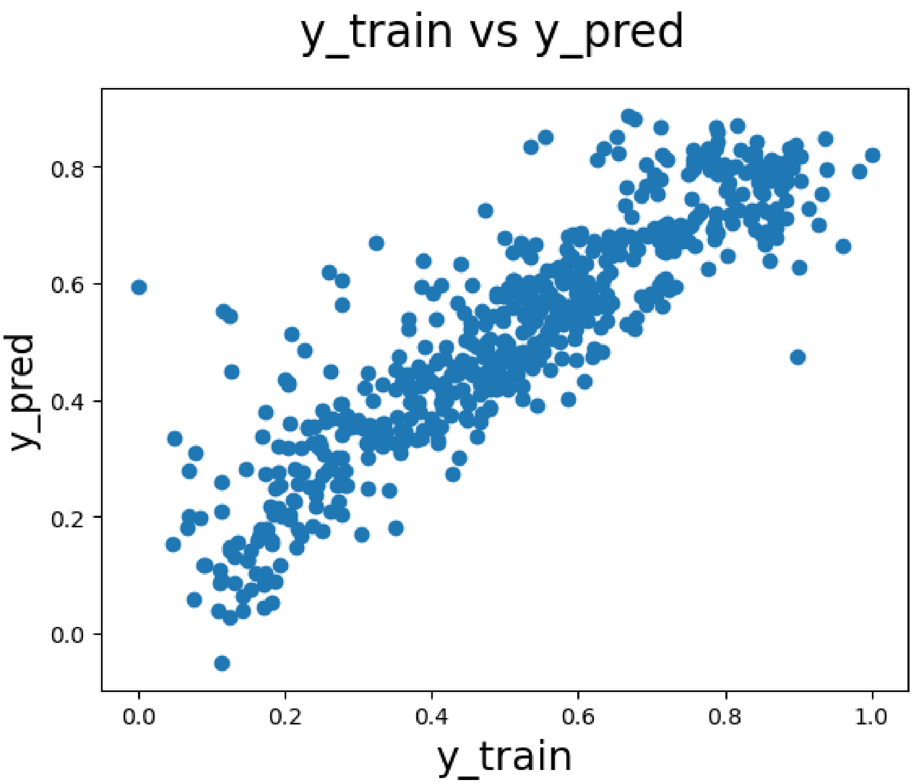
Assumption 1: Normal distribution of error terms are observed for the training dataset. Following figure shows the distribution of error terms-



Assumption 2: The error terms are independent. They are not following any pattern. The error distribution does not seem independent. Following figure demonstrates this distribution.



<u>Assumption 3:</u> There is constant variance in the distribution of error terms (homoscedasticity).



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

<u>Answer:</u>

Top 3 contributing features are:
Temperature (temp): Temperature has significant effect on the business.
Year (yr): Year on year growth can be seen during analysis.
Season(season): Fall & Summer season has highest demand of shared bikes.

Final equation:

cnt = 0.2012 + 0.2392(yr) - 0.094(holiday) + 0.3188(temp) - 0.1469(Spring) - 0.0234(Summer) - 0.0319(Dec) - 0.0306(Feb) - 0.0584(Jan) + 0.0387(June) + 0.0455(May) + 0.0707(Sep) + 0.0141(Saturday) + 0.0324(Sunday) + 0.0015(Wednesday) + 0.0905(clear_partlycloudy)

General Subjective Questions

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

Answer: Linear regression algorithm is a supervised machine learning algorithm that finds the best straight line fitting the data. Linear regression is primarily used for understanding the impact of historical data on the target variable. Here the linear regression algorithm can be applied only for continuous data, with the assumption that there is some linear relationship between the target variable and other independent variables. The accuracy of the best fit line is obtained through the least squares method.

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.

$$\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$$
$$\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$$

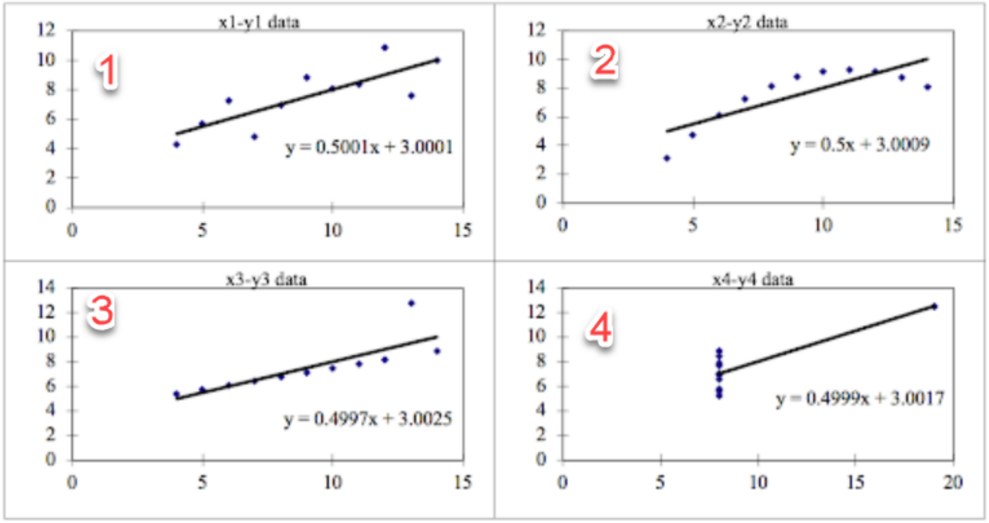
Assumptions for linear regression model:

- The errors or residuals have a normal distribution.
- The distribution of errors has constant variance (no correlation).
- The errors are independent and do not follow any pattern.

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

Answer:

Anscombe's quartet consists of 4 different plots which are similar in terms of descriptive statistics (mean and variance), however their distributions are very different from each other. Anscombe's quartet illustrates the importance of first plotting the features before analysis or building a model. Following figure shows the linear regression model fitted for these 4 datasets:



Dataset 1: Here the linear regression model is a good fit.

Dataset 2: The linear regression model cannot capture non-linear distribution.

Dataset 3: Linear regression model is sensitive to the outlier, resulting in an incorrect result.

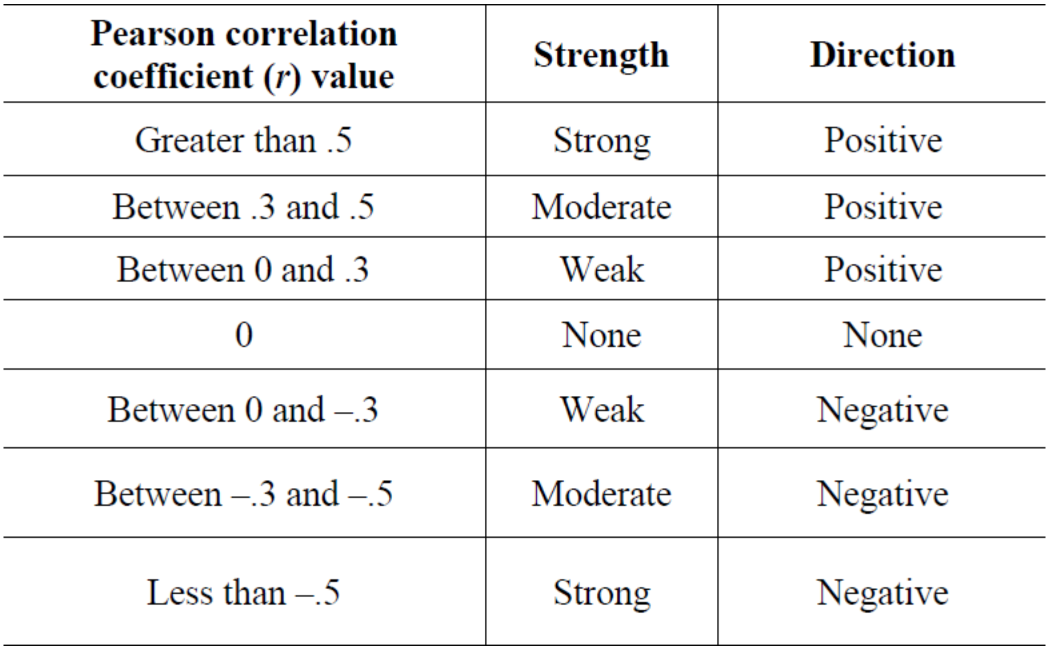
Dataset 4: Here again the sensitivity towards outliers of the linear regression model is demonstrated.

To summarize, anscombe's quartet depict how easy it is to build an incorrect regression model without due diligence of first visualizing the relationships.

3. What is Pearson's R?

Answer:

Pearson's correlation coefficient is a parameter that is used to measure the strength and direction of a linear correlation, between two variables. The value of the coefficient lies between -1 and 1. A value of "-1" indicates a negative correlation, "0" indicates no correlation and "1" indicates a strong linear correlation. Following table provides the general rules of thumb followed while interpreting the pearson's correlation coefficient:



Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The pearson’s correlation coefficient can be primarily used when all of the following points are true:

- There is linear relationship between the quantitative variables.
- There are no outliers.
- The errors are normally distributed.

The pearson’s coefficient also provides a measure of how well the observations are distributed around the best-fit line. A value of -1 or 1 indicates that all the observations are distributed on the line.

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

Answer:

What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.

Why - 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, 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.

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

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

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

Answer:

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

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

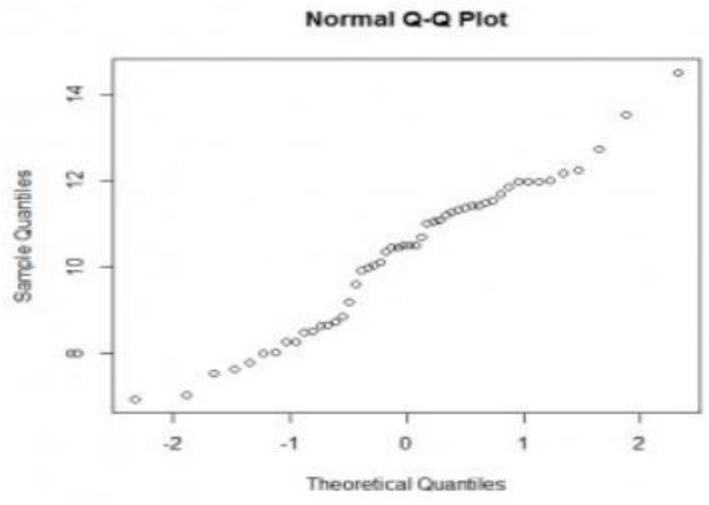
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Answer:

Quantile-quantile plot is used to determine whether two samples of data belong to the same population. The quantiles of the first dataset are plotted against the quantiles of the second dataset, if the two samples belong to the same population, then the points will lie along the same line. Uses of a Q-Q plot:

- Identify whether two samples belong to the same population.
- Determine the distribution of the sample (normal, uniform etc.)

An example of a Q-Q plot is shown in below figure:


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles