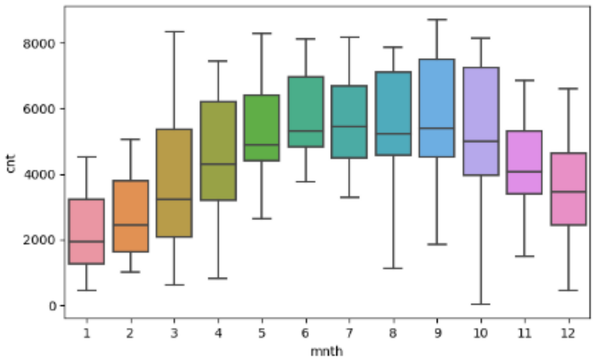
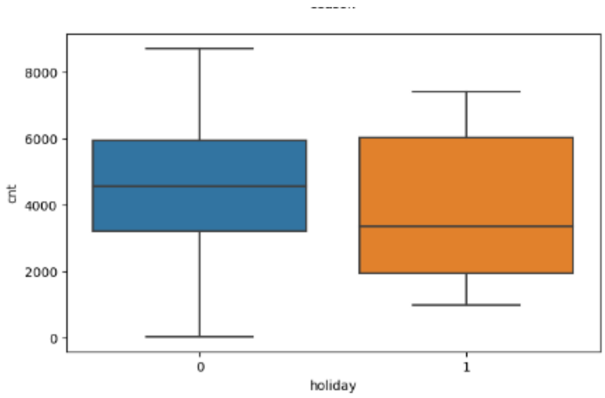
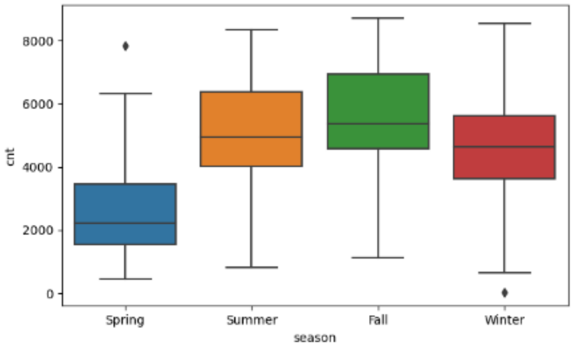
Assignment-based Subjective Questions

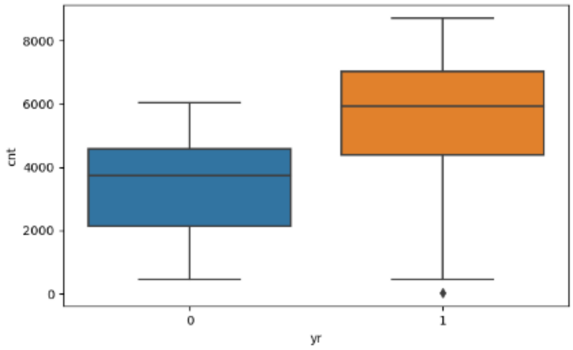
1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

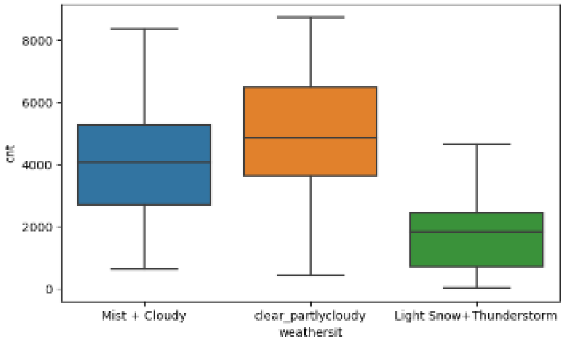
Answer: Following figures show the boxplot of dependent variable (cnt)vs different categorical variables- mnth, holiday, season, yr and weathersit











- Higher median values for summer and fall seasons indicate higher number of users during these seasons, as the weather might be more favourable
- There is a significant increase in the no. of users for the year 2019, as the COVID restrictions would have eased
- There is a gradual increase in the no. of users from the months January to July and then it starts decreasing, this would probably correspond to the seasons
- Higher no. of users are present on a holiday/weekend. It is assumed that "0" indicates holiday/weekend

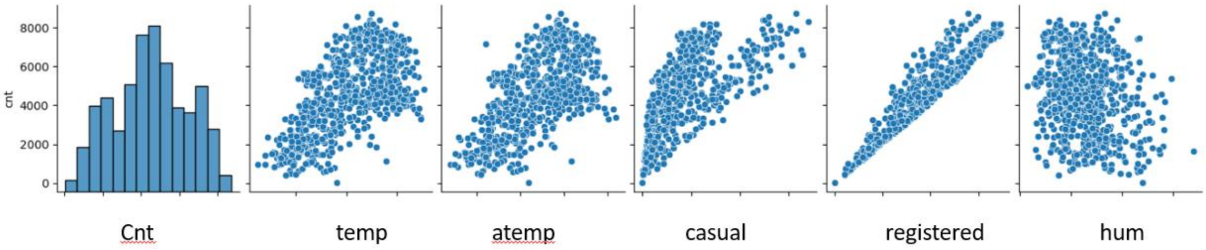
- Working day or a weekday do not have any significant impact on the no. of users as there is not much change in the median values
- More no. of users (5000 approx.) are observed if the weather is "clear or partly cloudy", compared to when it is misty and cloudy (4000 approx. users)

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

Answer: It is important to reduce the no. of columns, as this helps in reducing the complexity of the model and therefore becomes easier to interpret the variables that are having a significant impact on the dependent variable. 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: Following is the pair-plot of the dependent variable vs other numeric variables

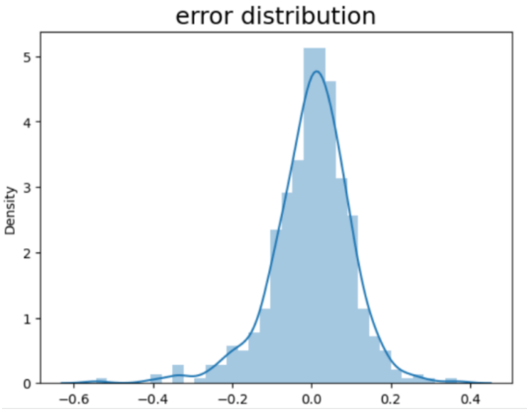


Here the variable "registered" has the highest correlation with the target variable. However, the target variable is actually the sum of casual and registered users, hence it would make sense to exclude them from further analysis. Additional to "casual" and "registered", the variables "temp" and "atemp"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

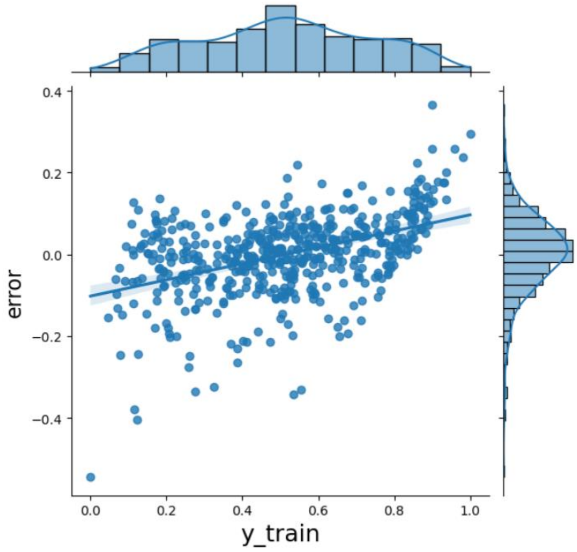
Answer: Before assuming the linear regression model, it is observed that there is some linear relationship between the target variable and other variables such as "temp" and "atemp", hence linear regression model is selected.

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

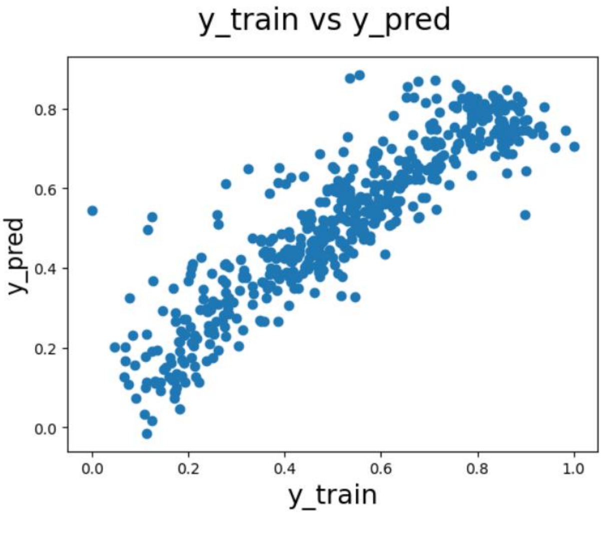


Assumption 2) The error terms are independent i.e they are not following a pattern.

The error distribution does not seem independent. A higher order model might be necessary to ascertain better information for this problem. Following figure demonstrates this distribution-



Assumption 3) There is constant variance in the distribution of error terms (homoscedastic), following figure shows that



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

Answer: Following is the equation:

cnt = 0.1338 + 0.237*(yr) - 0.0826*(holiday) + 0.46*(temp) - 0.17*(windspeed) - 0.077*(Spring) + 0.042*(Summer) + 0.0705*(Winter) + 0.091*(clear_partlycloudy)

top 3 features- “temp”, “yr” and “clear_partlycloudy”

temp: for a unit increase in the ambient temperature, the target variable increases by 0.46 times

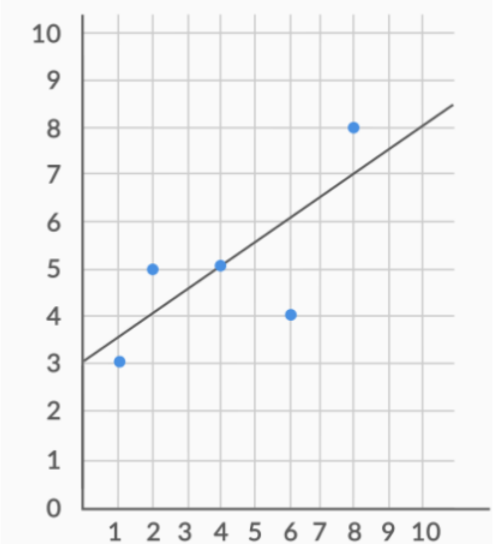
yr: for a unit increase in the feature “yr”, the target variable increases by 0.237 times

clear_partlycloudy: for a unit increase in this variable, the target variable increases by 0.091 times

General subjective questions

6) 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. Following figure shows a straight line fit for a scatter plot-



Following is the equation of a simple linear regression model-

Formula

$$y = \alpha + \beta x$$

β = slope

α = y-intercept

y = y- coordinate

x = x-coordinate

With multiple independent variables, the linear regression equation is a shown below-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

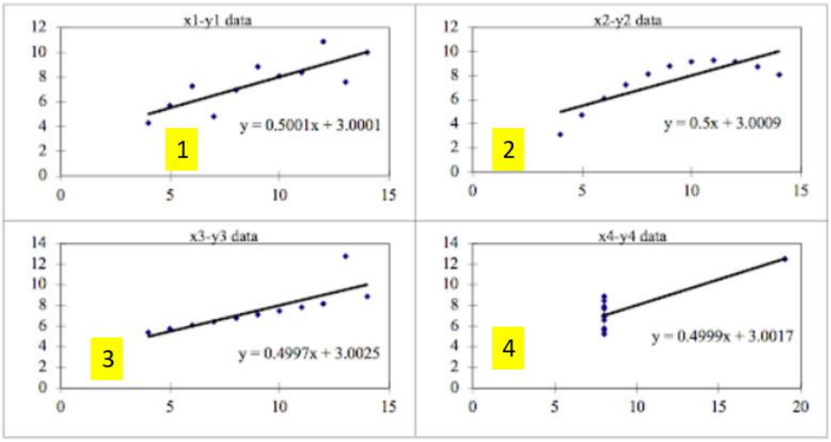
The coefficient of the variable provides information of impact of that variable on the target variable, if all other variables remain constant.

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

7) Explain the Anscombe's quartet in detail

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.

Reference for the figure: https://builtin.com/data-science/anscombes-quartet

8) 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.

9) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

a. What is scaling? Why is scaling performed?
Scaling of features is an import data pre-processing step, done to ensure that all the values are within the same range of magnitude of 0 to 1. If scaling is not done then it results in incorrect values of coefficients of the respective variables, as a result the model would end up assigning incorrect weightage to some variables (owing to their magnitudes). Hence it is important to normalize the magnitudes of all the continuous variables (dummy variables are excluded).

b. What is the difference between normalized scaling and standardized scaling?
Standard scaling: the feature is scaled by subtracting the mean from all the data points and dividing the resultant values by the standard deviation of the data. Following is the formula-

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Normalized scaling (Min-max scaling): Here the data point is subtracted with the minimum value from the datapoint and result is divided by the difference between the maximum and the minimum value. Following is the formula-

$$X_{\text{scaled}} = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Normalized scaling results in the feature values scaled to magnitudes between 0 to 1, while standardized scaling scales the feature in which the datapoints are having a mean of zero and standard deviation of 1

Standardized scaling is useful when the distribution of the feature values is not known or when there are outliers in the data

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

Answer: VIF is an index that provides a measure of how much the variance of a regression coefficient increases or explained due to multi-collinearity. If there is perfect correlation between the features, then it results in infinite value for VIF. A large value of VIF indicates that there is a correlation between the variables.
VIF can be calculated from the following formula-

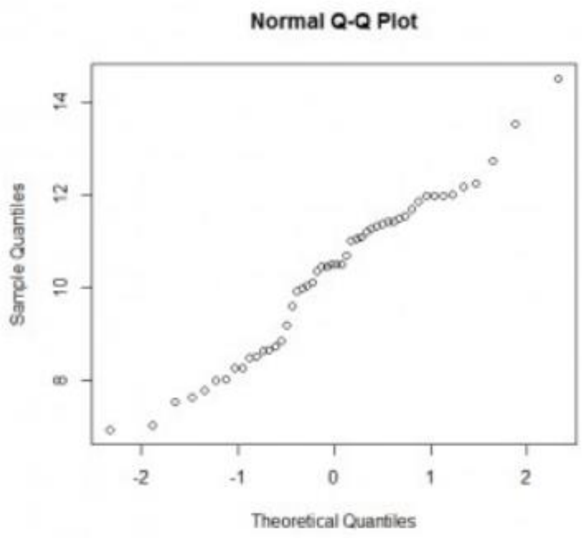
$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

11) 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 is 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