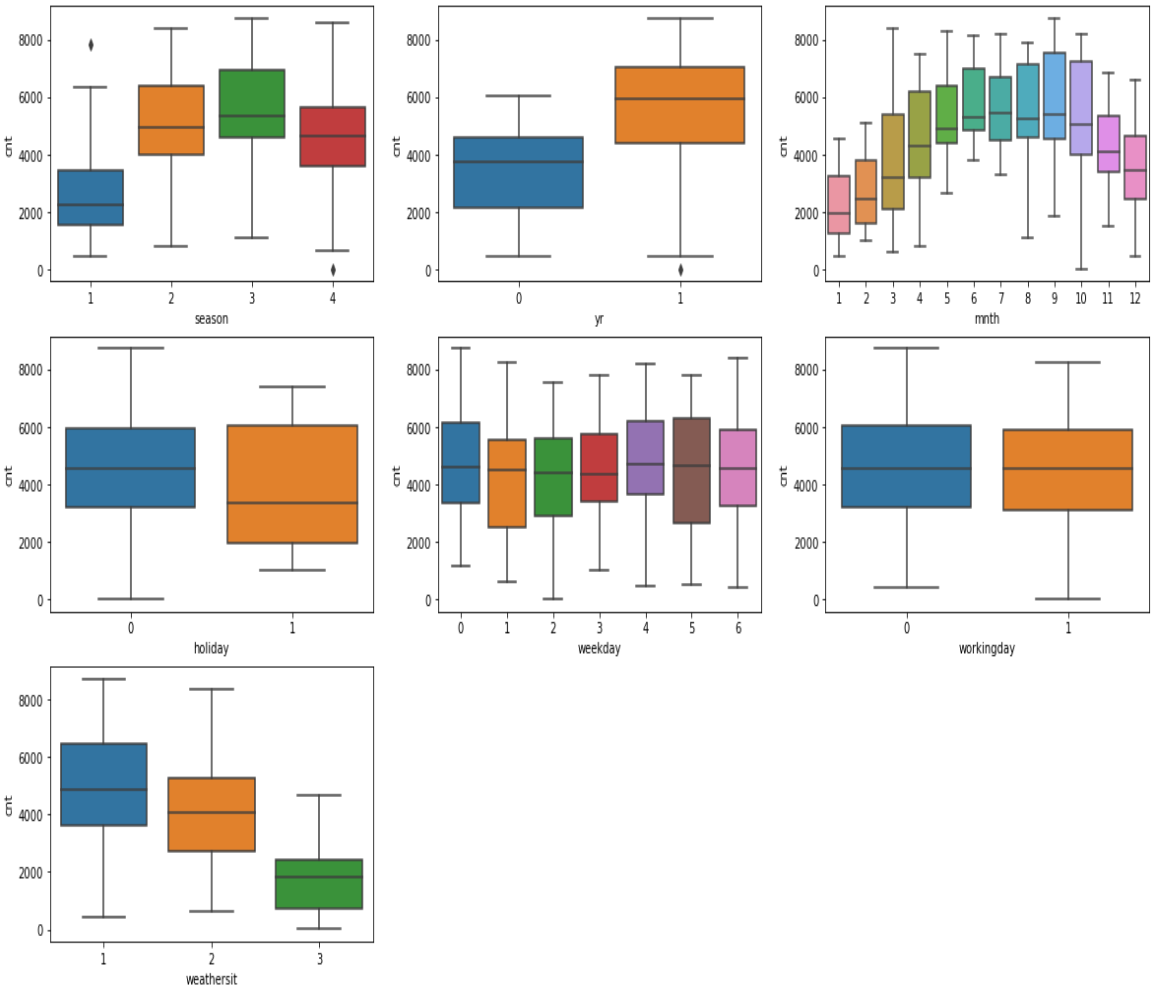
Assignment-based Subjective Questions

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

From the dataset, we mainly get 7 categorical variables: season, yr, mnth, holiday, weekday, workingday and weathersit.



As per above plot, we can infer:

season: Demand is higher for season 3(fall) and least for season 1(spring). Season 2(summer) have higher demand than season 4(winter).

yr: Demand is higher for yr value 1 i.e., 2019 than yr value 0 i.e., 2018.

mnth: Demand(cnt) vs month follow same trend as in season. It increases from 1 to 6, becomes saturated for 7,8 and 9 and starts decreasing from 10th month.

holiday: Demand is higher when holiday is 0 i.e., when it is not a holiday demand is more in comparison to when it is holiday.

weekday: If we look at the medians of each day, it seems approximately equal. Whereas if looking at upper bounds, it can be said that for weekday 0(Monday) demand is higher than the other days.

workingday: Median value for both category almost lies at same level.

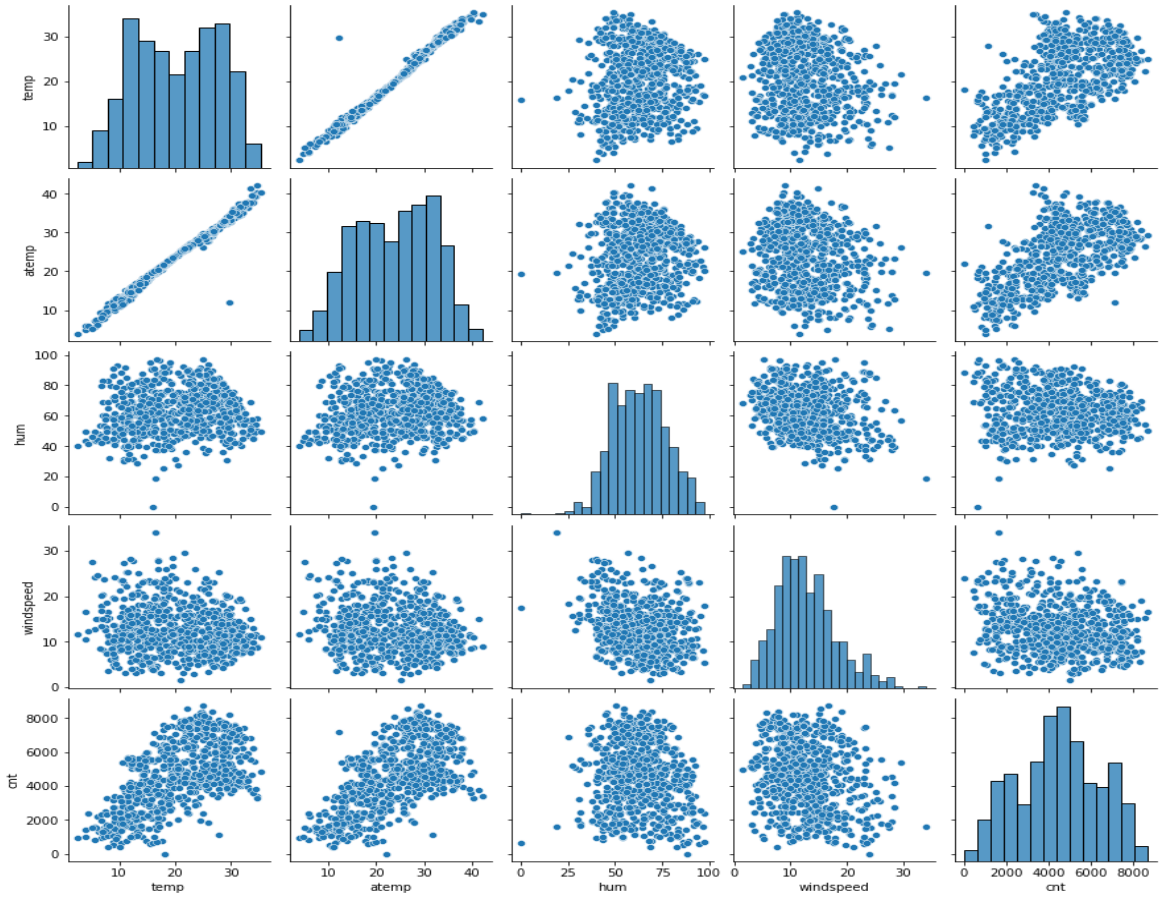
weathersit: As per the data dictionary, there are 4 categories, but as per the graph between weathersit and cnt, there are only 3 categories (1,2,3). Category 4 is missing, which represents demand for category 4 is 0. Demand for category 1 is highest and is decreasing as category moves from 1 to 4.

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

It is important to use **drop_first=True** during dummy variable creation as it helps in **reducing no of columns** created. It helps in reducing correlation between the variables, and eases the process of building the model.

For e.g., a column of months contains all values starting from January to December. Total different values these columns have is 12. If we create dummy variable for this column without drop_first=True, then it will create 12 different columns, whereas if we create dummy variable with drop_first=True, it will only create 11 columns. For 11 columns, lets assume it drops column with month January. So, if no other column is having value as 1, it automatically means that it is case of January. While creating model, we have to deal with 1 less column and it will reduce correlation also.

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

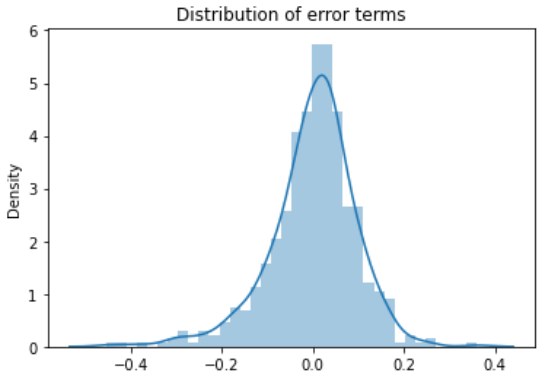


From the above plot, it shows **temp** variable have highest correlation with the target variable. Although temp and atemp have high correlation with each other, because of that, temp and atemp both are having high correlation with the target(cnt) variable.

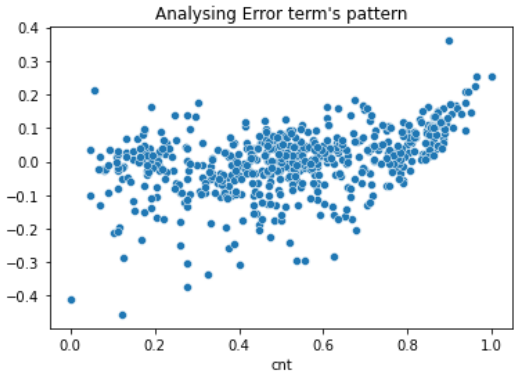
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

After building Linear Regression Model, assumptions are validated by

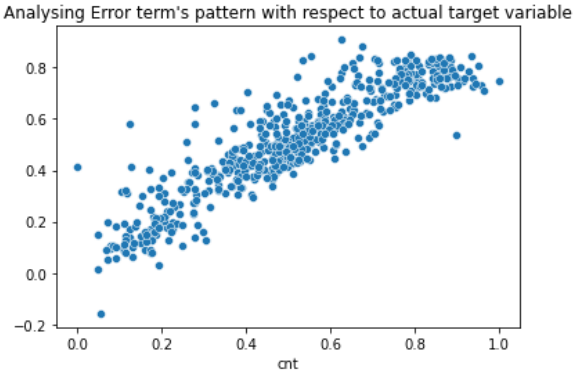
- Plotting distribution of error terms of training data. It shows error terms are normally distributed around 0.



- Plotting error terms versus target variable.



- Plotting predicted target values versus actual target values.



From the above 2 plots, no pattern is observed. Hence assumptions of Linear Regression are validated successfully.

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

As per the final model, we get below equation:

cnt = 0.411 * const + 0.49 * atemp + 0.224 * yr + 0.084 * season_winter - 0.067 * mnth_December - 0.072 * mnth_July - 0.081 * mnth_November - 0.139 * season_spring - 0.149 * windspeed - 0.181 * weathersit_Light Snow/Rain - 0.256 * hum

From this equation, we can say top 3 features are **atemp**, **yr** and **hum**.

General Subjective Questions

6. Explain the linear regression algorithm in detail.

It is a type of Supervised Machine Learning Algorithm used for Predictive Analysis. Linear regression is a statistical regression approach for predicting the relationship between variables. Linear regression, as the name implies, depicts a linear relation between the independent variable and the dependent variable. When there is only one independent variable, it is known as Simple Linear Regression, and when number of independent variables are more than one, it is known as Multiple Linear Regression. A sloped straight line describing the relationship between the variables is produced by the linear regression model.

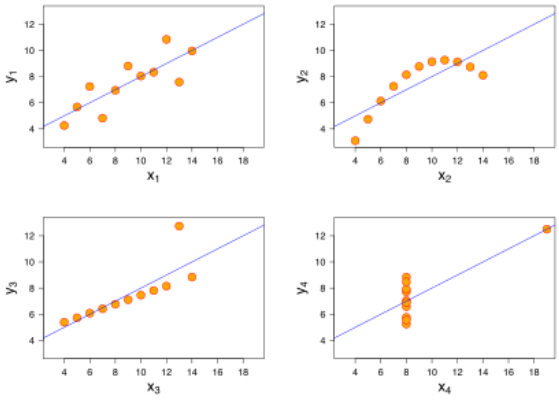
Equation for Linear Regression:

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

Y: Dependent/Target variable.
$X_1 \ldots X_p$: Independent/Predictor Variables.
β_0: Y intercept (value of Y when X = 0).
$\beta_1 \ldots \beta_p$: These are coefficients of different X variables (X_1 to X_p).
ε: Error term

7. Explain the Anscombe's quartet in detail.

According to Wikipedia, Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



Anscombe's quartet							
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of *x*	9
Sample variance of *x*: s^2	11
Mean of *y*	7.50
Sample variance of *y*: s^2	4.125
Correlation between *x* and *y*	0.816
Linear regression line	$y = 3.00 + 0.500\ x$

8. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r is a measure of linear correlation between two sets of data.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

Equation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r: correlation coefficient
x_i: values of the x-variable in a sample $\bar{x}$Type equation here.= mean of the values of the x-variable y_i: values of the y-variable in a sample $\bar{y}$= mean of the values of the y-variable

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

It is one of the most important data pre-processing step that is used on independent variables in order to normalise the data within a given range. It helps in increasing speed of calculations in the algorithms.

Maximum times, the collected data set contains features with a wide range of magnitudes, units, and ranges. If scaling is not done, the machine learning algorithm will only consider magnitude rather than units, resulting in inaccurate modelling. To handle this situation, scaling is performed on the collected dataset.

Normalization Scaling: It is also known as min-max scaling. It scales all data in range of 0 and 1.

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

Standardized Scaling: It replaces all the values by their Z-score and brings it to standard normal distribution which has mean = =0 and standard deviation = 1.

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

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

Within a multiple regression, the Variance Inflation Factor (VIF) is a measure of collinearity across predictor variables.

VIF equals to infinite, it shows perfect correlation. This demonstrates that two independent variables have a perfect correlation. We get R square = 1 in the event of perfect correlation, which leads to 1/ (1-R square) equals to infinite. To overcome this problem, we need to remove one of the variables that is producing the perfect multicollinearity.

An infinite VIF value suggests that a linear combination of other variables may exactly express the related variable.

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

A plot of the quantiles of two distributions against each other, or a plot based on quantile estimations, is known as a Q–Q plot. To compare the two distributions, the pattern of points in the figure is used.

When viewed from left to right, the points plotted in a Q–Q plot are always non-decreasing. The Q–Q plot follows the 45°line y = x if the two distributions being compared are identical. The Q–Q plot will follow some line, but not necessarily the line y = x, if the two distributions agree after linearly

transforming the data in one of the distributions. The distribution represented on the horizontal axis is more dispersed than the distribution plotted on the vertical axis if the general trend of the Q–Q plot is flatter than the line y = x. In contrast, if the general trend of the Q–Q plot is steeper than the line y = x, the vertically projected distribution is more dispersed than the horizontally displayed distribution. The "S" form of Q–Q plots indicates that one of the distributions is more skewed.