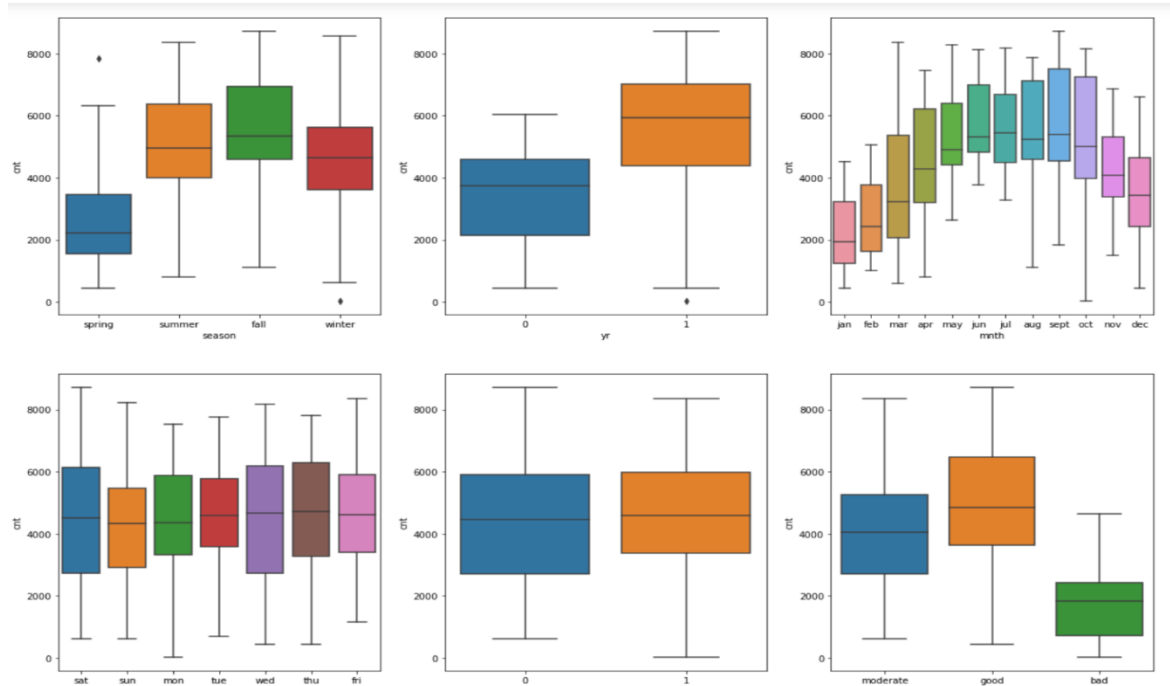


## Assignment-Based Subjective Questions

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

**ANSWER:** There are a couple of categorical variables namely season, mnth, yr, weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same



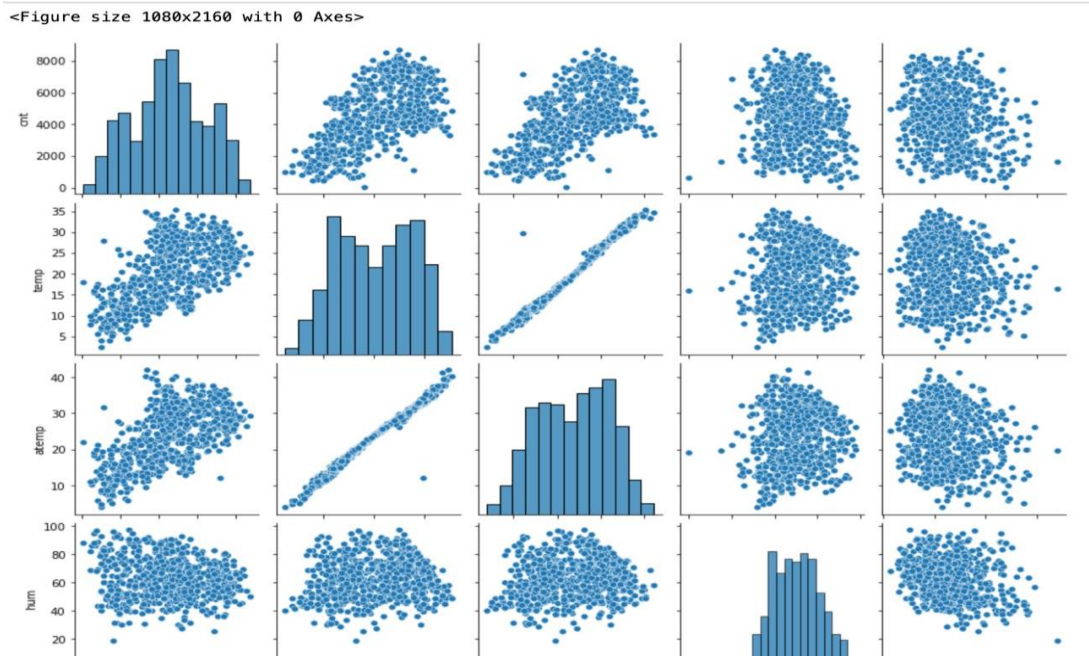
These variables are visualized using bar plot and Box plot both.

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

**Answer:** The argument `drop_first=True` is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we have a categorical variable with  $n$ -levels, then we need to use  $n-1$  columns to represent the dummy variables.

Consider a Categorical column with 3 types of values, we want to create dummy variable for that column. If one variable is neither furnished nor semi\_furnished, then it is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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



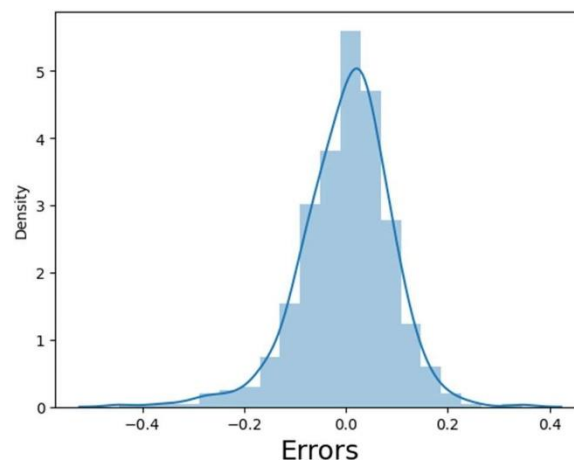
The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

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

We have done following tests to validate assumptions of Linear Regression:

1. There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. (ref. see above question's pairplot)
2. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.

Error Terms



3. linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another.

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

The top 3 significant features are:

1. temp - coefficient : 0.437655
2. yr - coefficient : 0.234287
3. weathersit\_Light Snow & Rain - coefficient : -0.292892

## **General Subjective Questions**

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

Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for  $a_0$  and  $a_1$  to find the best fit line and the best fit line should have the least error.

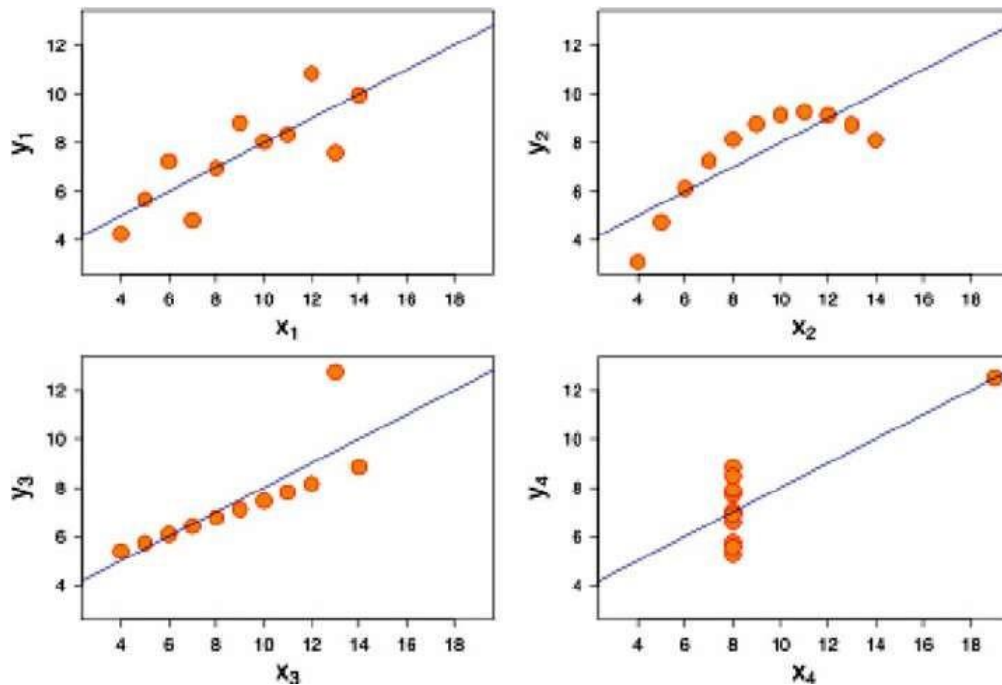
In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for  $a_0$  and  $a_1$ , which provides the best fit line for the data points.

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

- a. The first scatter plot (top left) appears to be a simple linear relationship.
- b. The second graph (top right) is not distributed normally; while there is a relation between them it is not linear.
- c. In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

d. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



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

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's  $r$ , the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

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

**Feature scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the low values, irrespective of the units of the values.

Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be

necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

**5. You might have observed that sometimes the value of VIF is infinite.**

**Why does this happen?**

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then  $VIF = \text{infinity}$ . It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

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

