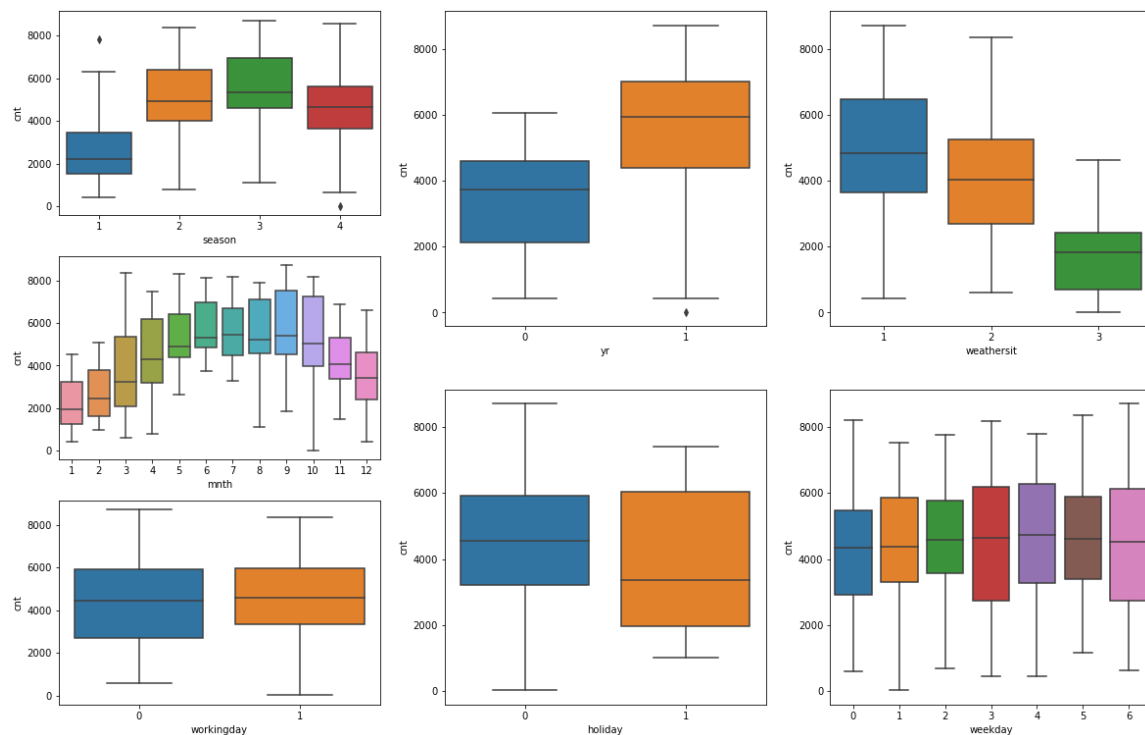


## Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



The categorical variables in the dataset are season, yr, mnth, working, weekday, weathersit, holiday. These were visualized using a boxplot and these variables has the following effect on the dependent variable (target variable - cnt)

- i) **Season** – From the boxplot, the spring season has least value of cnt whereas fall has maximum value of cnt. Summer and winter have intermediate value of cnt.
- ii) **Weathersit** - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. Highest count was seen when the weathersit is 'Clear, Few clouds, Partly cloudy, Partly cloudy'
- iii) **Holiday / Workingday** - Rentals reduces during holiday and more during working day.
- iv) **Mnth** - September has highest number of rentals and November has the least.
- v) **Yr** - The number of rentals in 2019 was more than 2018

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

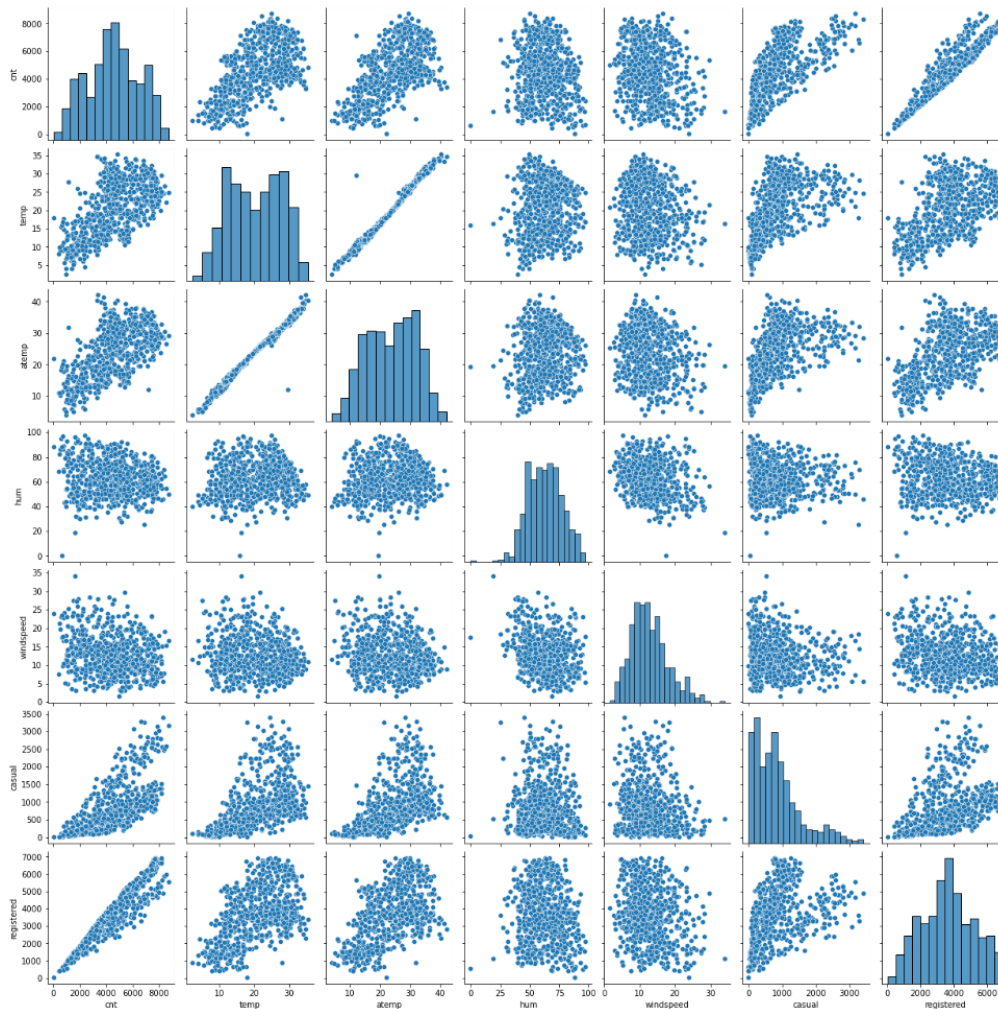
`drop_first=True`, helps in reducing the extra column created during dummy variable creation. Also, it reduces the correlations created among dummy variables.

If we don't keep `drop_first=True`, number of dummy variable column is more and that new column may affect the correlations of the models and if we have all dummy variables it leads to multicollinearity between the dummy variables.

Hence if we have categorical variable with  $n$ -levels, then we need to use  $n-1$  columns to represent the dummy variables.

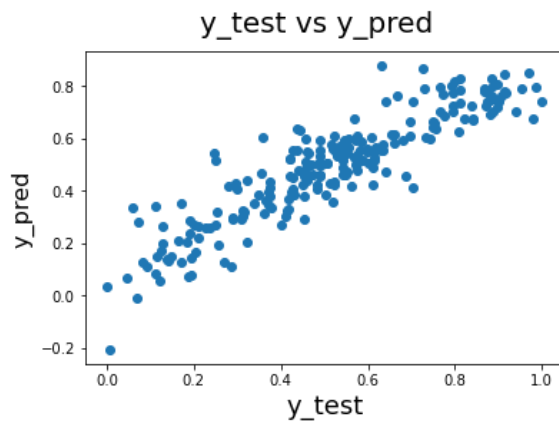
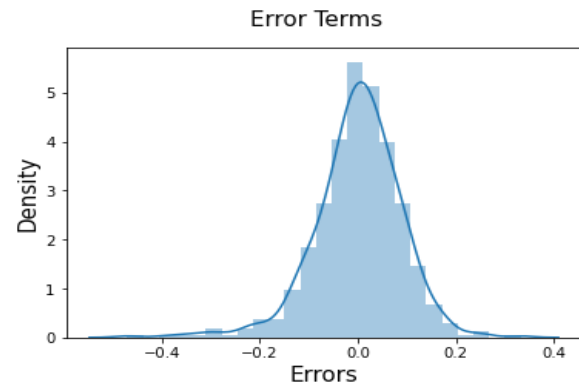
## 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Looking at the below pair plot, **temp** and **atemp** are having highest correlation with the target variable



**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Residuals distribution should follow the normal distribution and centered around 0 (mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.



The variance should not increase (or decrease) as the error values change. Also, the variance should not follow any pattern as the error terms change. It is validated in the graph.

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

The top 3 features are

S. No	Features	Co-efficient
1	temp	0.412155
2	weathersit - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds	-0.296155
3	yr	0.234608

	Variables	Coefficient value
index		
3	temp	0.412155
0	const	0.244055
1	yr	0.234608
6	winter	0.069368
9	saturday	0.066760
2	workingday	0.054321
11	jan	-0.057768
10	dec	-0.058544
13	nov	-0.059103
12	jul	-0.071054
8	mist	-0.081541
5	spring	-0.105609
4	windspeed	-0.163490
7	light_snow	-0.296155

## **General Subjective Questions**

### **1. Explain the linear regression algorithm in detail. (4 marks)**

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation " $y = mx + c$ ".

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

**1. Simple Linear Regression:** SLR is used when the dependent variable is predicted using only **one** independent variable.

**2. Multiple Linear Regression:** MLR is used when the dependent variable is predicted using multiple independent variables.

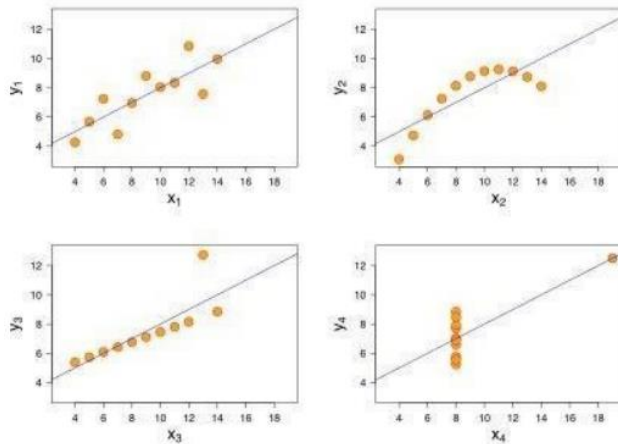
The equation for MLR will be:

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

$\beta_1$  = coefficient for  $X_1$  variable  $\beta_2$  = coefficient for  $X_2$  variable  $\beta_3$  = coefficient for  $X_3$  variable and so on...  $\beta_0$  is the intercept (constant term).

### **2. Explain the Anscombe's quartet in detail. (3 marks)**

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



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.
- 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.
- 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? (3 marks)

Pearson's  $r$  is a numerical summary of the strength of the linear association between the variables. Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data?  $r = 1$  means the data is perfectly linear with a positive slope  $r = -1$  means the data is perfectly linear with a negative slope  $r = 0$  means there is no linear association

The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

#### Example:

- Positive linear relationship: In most cases, universally, the income of a person increases as his/her age increases
- Negative linear relationship: If the vehicle increases its speed, the time taken to travel decreases, and vice versa

### 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks).

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 preprocessing 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 lower 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 Neighbors and Neural Networks. **It ranges from 0 to 1.**

$$\text{Minmaxscaler} = (x - x_{\min}) / (x_{\max} - x_{\min})$$

- 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.

$$\text{Standardisation} = (x - \mu) / \sigma$$

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity ( $VIF = 1 / (1 - R^2)$ ). If there is perfect correlation, then  $VIF = \infty$ , where  $R^2$  is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1. So,  $VIF = 1 / (1 - 1)$  which gives  $VIF = 1/0$  which results in "infinity"

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

The q-q plot is like a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution