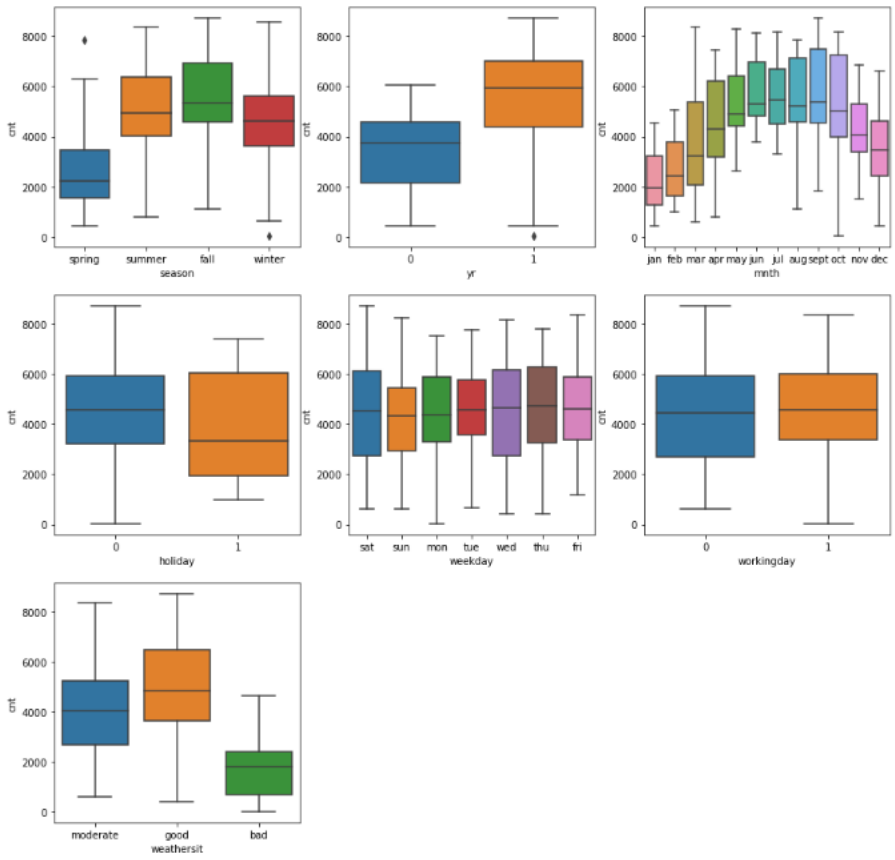
# Assignment-based Subjective Questions

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




The categorical variable used in the dataset: season , yr(year) , holiday, weekday ,workingday, and weathersit(weather situation) and mnth(month) . These were visualized using a boxplot.

These variables had the following effect on our dependant variable: -

- Season - For the variable season, we can clearly see that the category 3: Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring.
- Yr - The year 2019 had a higher count of users as compared to the year 2018.
- Holiday - rentals reduced during holiday.
- Weekday - The bike demand is almost constant throughout the week.
- Workingday – From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is quite adverse. Highest count was seen when the weather situation was Clear, Partly Cloudy.
- Mnth - The number of rentals peaked in September, whereas they peaked in December. This observation is consistent with the observations made regarding the weather. As a result of the typical substantial snowfall in December, rentals may have declined

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

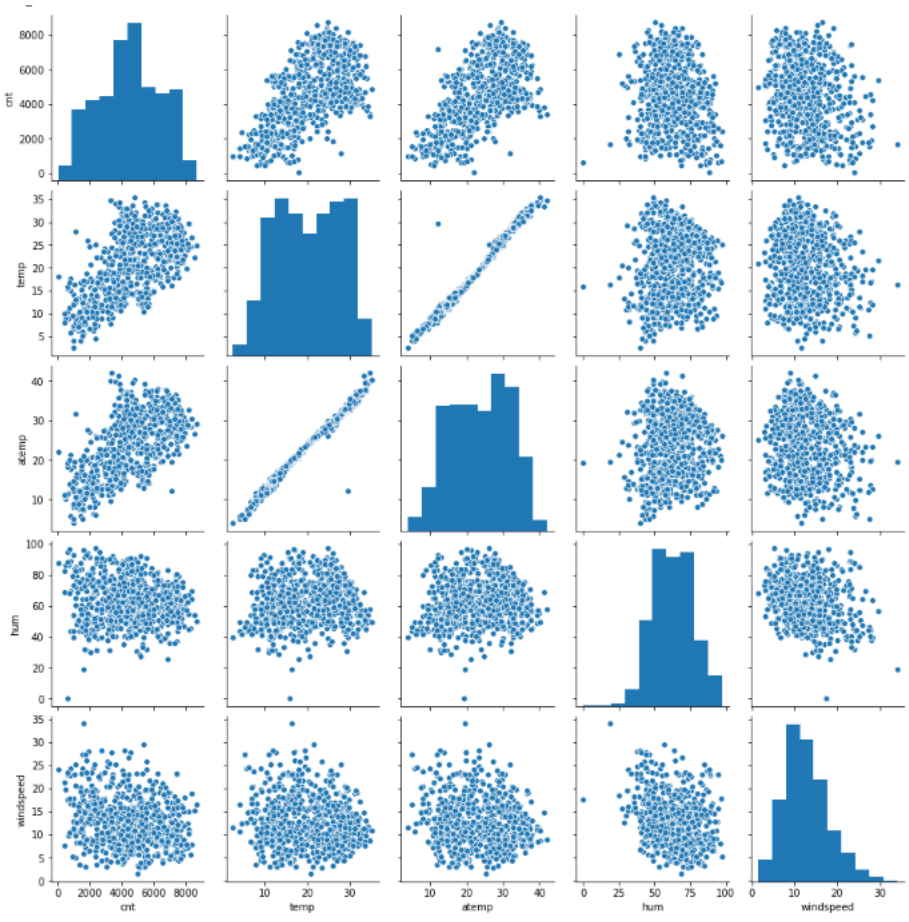
drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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.

| Value | Indicator Variable | |
|---|---|---|
| Furnishing Status | furnished | semi-furnished |
| furnished | 1 | 0 |
| semi-furnished | 0 | 1 |
| unfurnished | 0 | 0 |

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

"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).

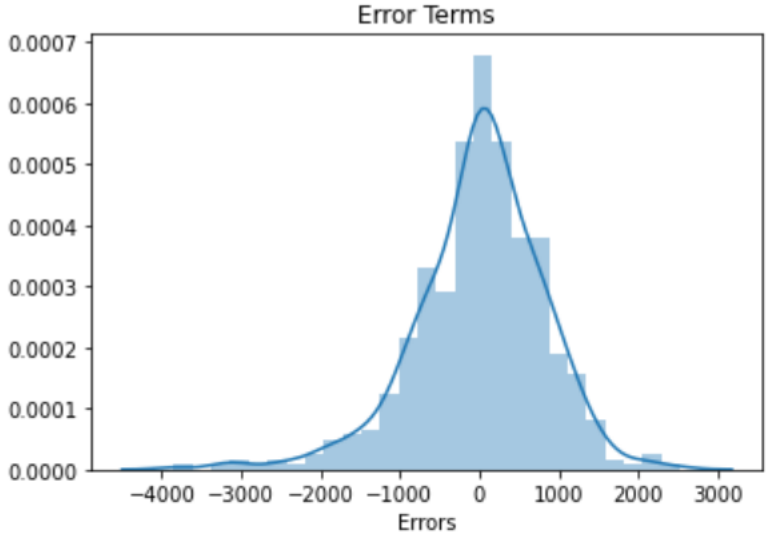


4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

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

a. 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)

b. 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.




c. 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 may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

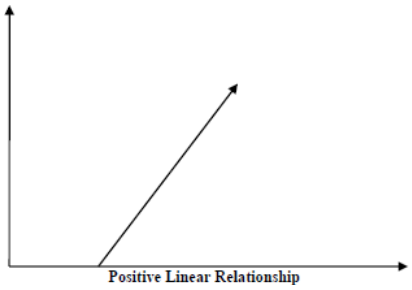
X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y c is a

constant, known as the Y-intercept. If X = 0, Y would be equal to c.

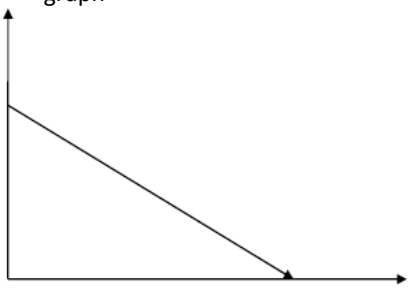
Furthermore, the linear relationship can be positive or negative in nature as explained below–

- Positive Linear Relationship:
  - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

- Negative Linear relationship:
  - A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

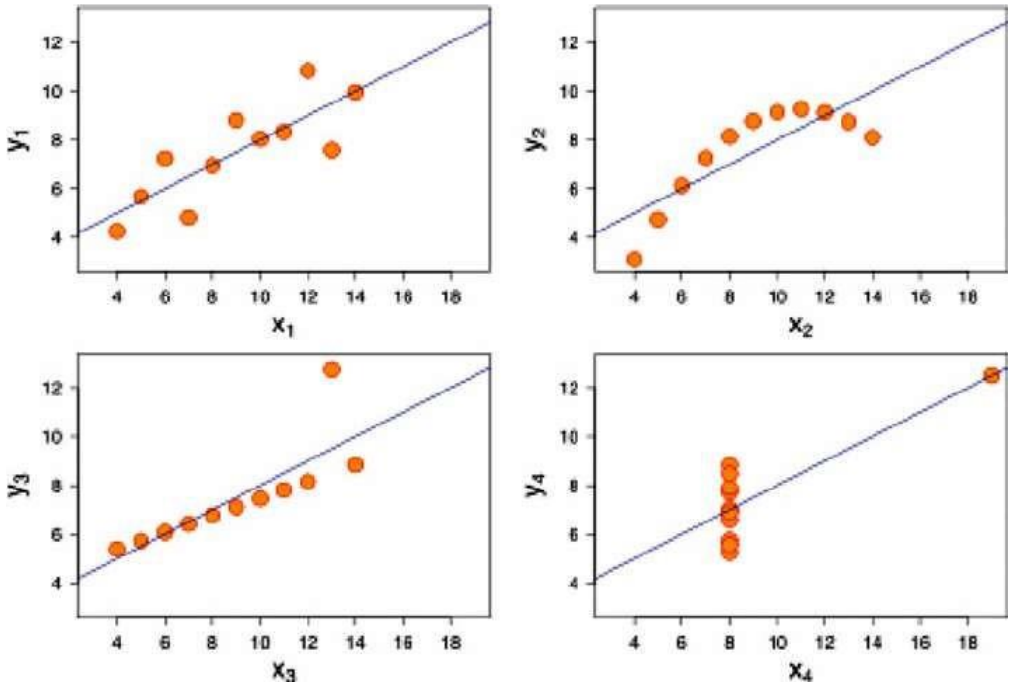
- ✓ Multi-collinearity –
    - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- ✓ Auto-correlation –
    - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- ✓ Relationship between variables –
    - Linear regression model assumes that the relationship between response and feature variables must be linear.
- ✓ Normality of error terms –
    - Error terms should be normally distributed
- Homoscedasticity –
    - There should be no visible pattern in residual values.

## 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.

- 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 is 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?

Pearson's r is a numerical summary of the strength of the linear association between the variables. It 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?"

Formula

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

$r$ = correlation coefficient

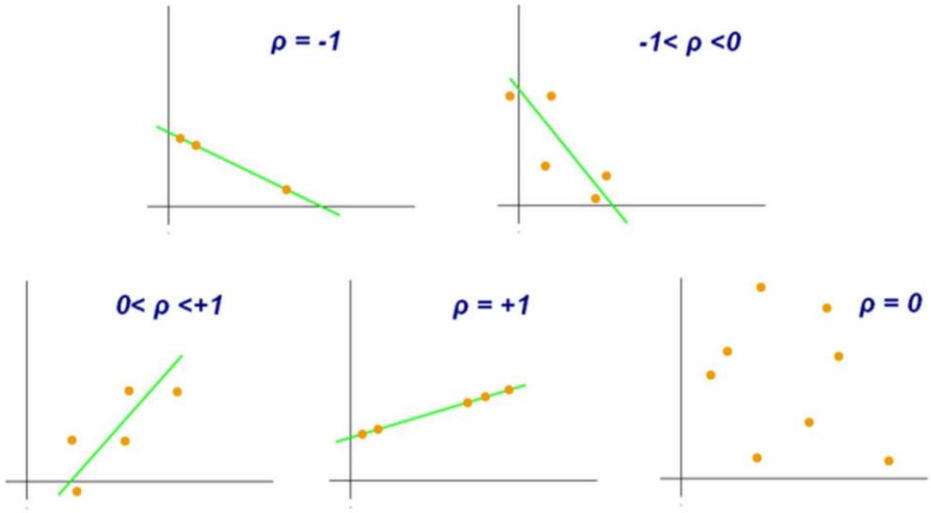
$x_i$ = values of the x-variable in a sample

$\bar{x}$ = 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

As can be seen from the graph below, 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




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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Example**:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

| S.NO. | Normalized scaling | Standardized scaling |
|---|---|---|
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for 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 = 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.

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

Where R-1 is the R-square value of that indedent 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 it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

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 behaviour?