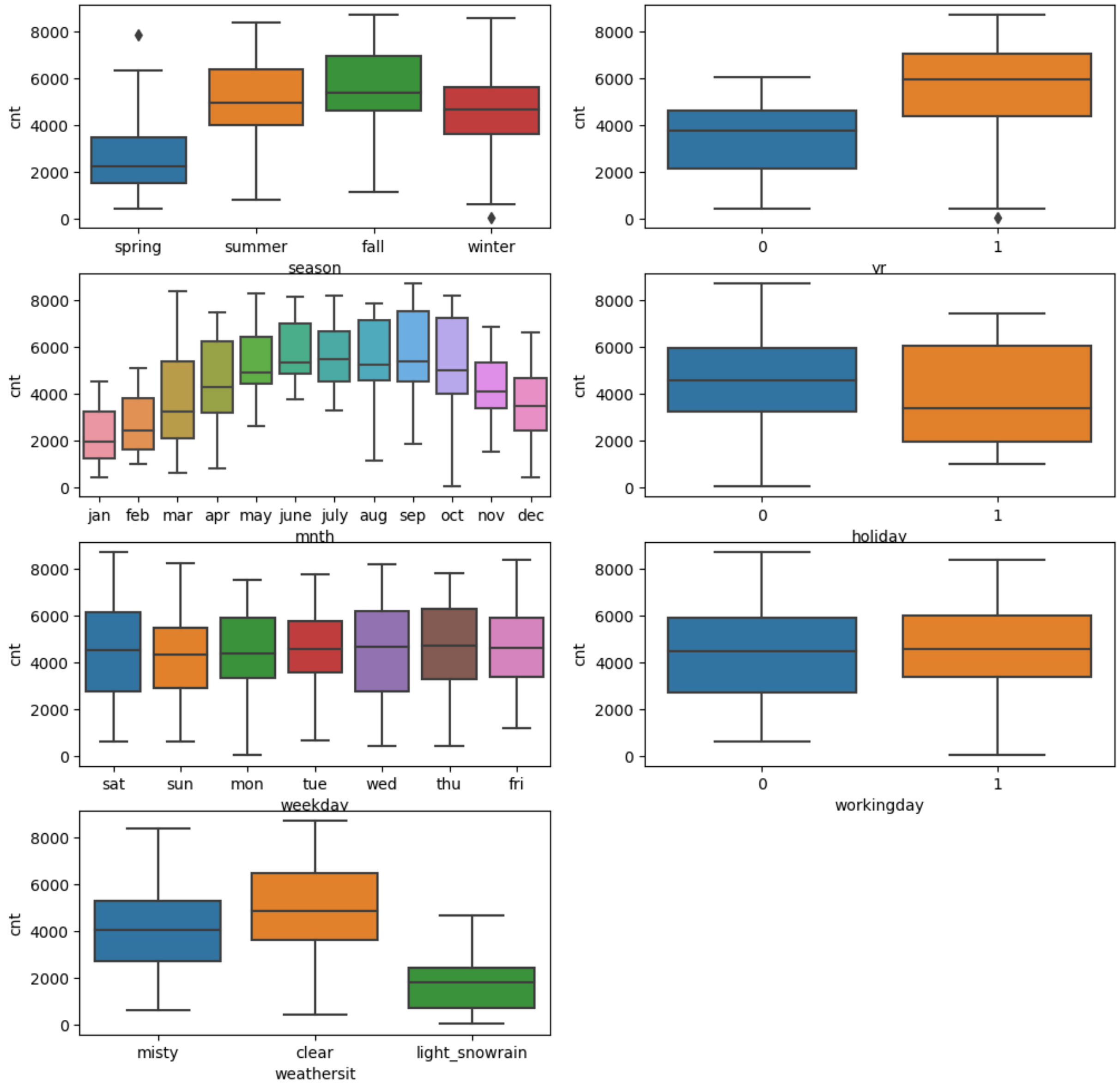
# Assignment-based Subjective Questions

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

Answer: Refer to the figure below:

- In the fall season and summer, people show a higher inclination to rent bikes.
- Bike rental rates peak in September.
- Saturday, Wednesday, and Thursday stand out as the days with the highest bike rental activity.
- Clear weather conditions contribute to most bike rentals.
- The year 2019 witnessed a higher volume of bike rentals.
- The distinction in bike rental rates between working days and non-working days is not substantially significant.



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

Answer: The primary reason for drop_first=True during dummy variable creation is important to prevent multicollinearity, ensure independence of variables.

For Example: Consider a categorical variable “season” with 4 categories: spring, summer, fall and winter. If we create dummy variables without dropping one, we might have:

Dummy_ spring, Dummy_ summer, Dummy_ fall, Dummy_ winter

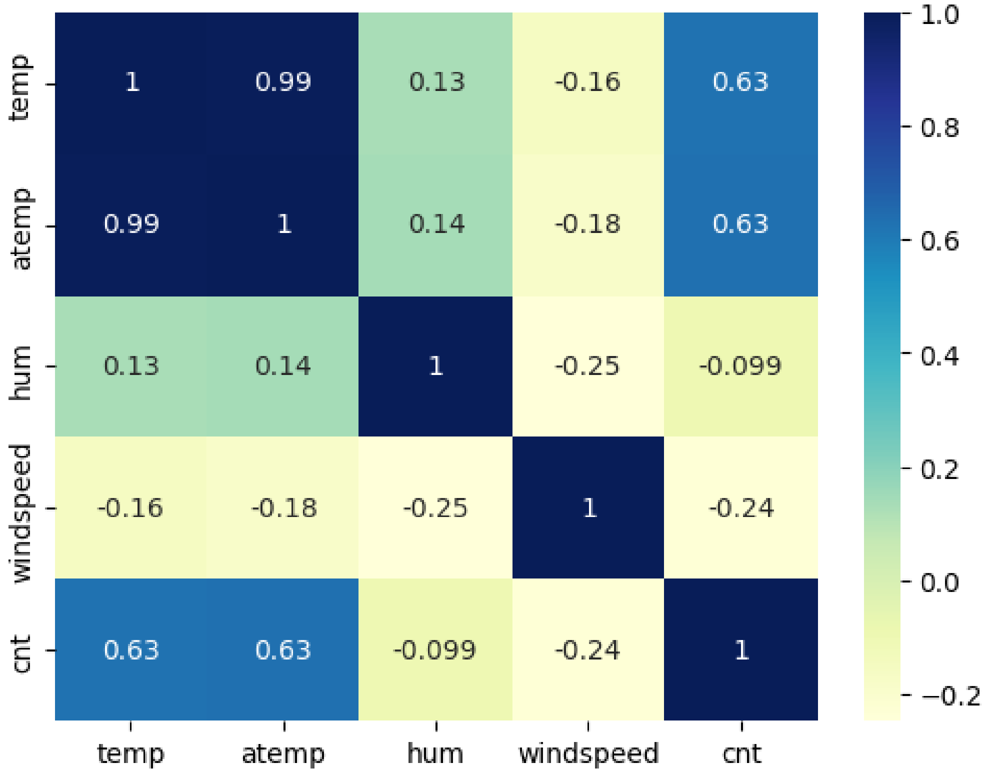
With drop_first=True, we drop one dummy variable, say Dummy_ spring, and have:

Dummy_ summer, Dummy_ fall, Dummy_ winter

This ensures that only 3 dummy variables (i.e., n-1 dummy variables) are used to represent the 4 categories preventing multicollinearity and facilitating a more interpretable and efficient model.

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

Answer: ‘Temp’ and ‘atemp’ are highly correlated with target variable ‘Cnt’. Refer to the figure below.



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

Answer:

- Examined normality of residual error using histograms to ensure they follow a normal distribution. (Figure 1 for reference)
- The predicted value has linear relationships by inspecting scatterplots of predicted values against actual values. (Figure 2 for reference)
- P-values of all variables are below the significance level (e.g., 0.05).
- Variance Inflation Factor (VIF) for each predictor variable are below 5.

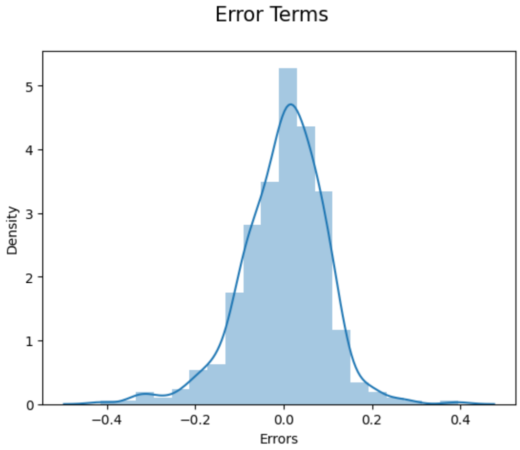


Figure 1

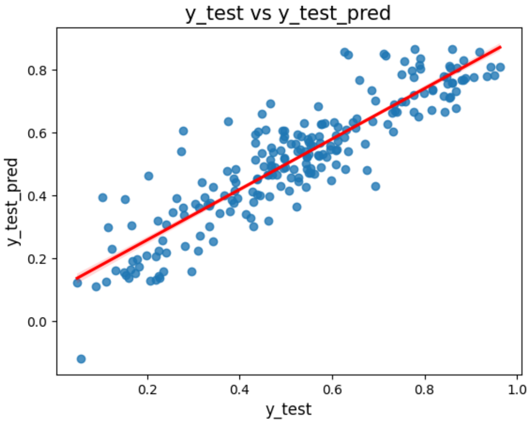


Figure 2

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

Answer: The top 3 influential features might include:

I. Year
II. Temperature
III. Weathersit (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

# General Subjective Questions

## 1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The basic idea is to find a linear equation that best fits the data points and can be used to predict the value of the dependent variable for given values of the independent variables.

The following is an example of a resulting linear regression equation:

$$Y = \beta 0 + \beta 1.X + err$$

Here,

- Dependent Variable(Y): This is the target variable which we want to predict.
- Independent Variable(X): These are used to predict the target variable.
- Equation: Y= Beta 0 + Beta1 *X + err (Beta 0 is the intercept, Beta 1 is the slope, err is the error)

Steps for Linear Regression:

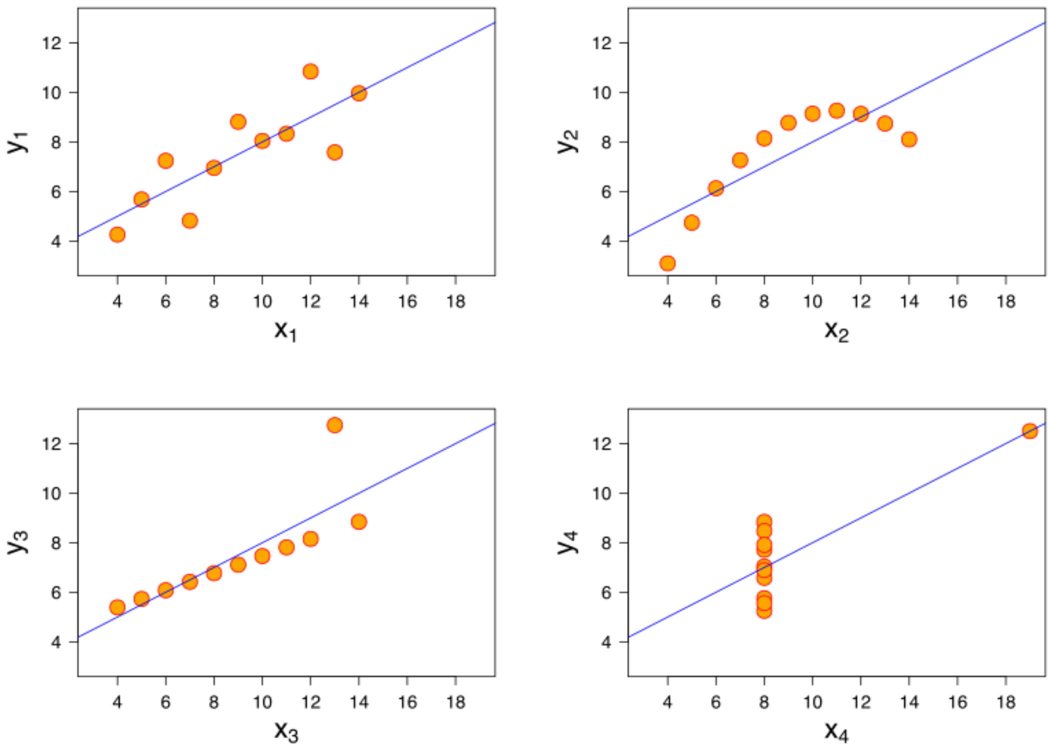
1. **Gather Data**: Collect relevant data that includes a dependent and one or more independent variables, data dictionary etc.
2. **Data Exploration (EDA)**: Explore the data to understand its distribution, identify outliers, and check for missing values.
3. **Train Test Split of Data**: Divide the dataset into training and testing sets to evaluate the model's performance on unseen data.
4. **Feature Scaling**: Standardize or normalize the features if necessary to ensure comparable scales. Min-Max scaling is one of them.
5. **Model Representation**: There are multiple approaches to build a model.
    I. Begin with one variable and keep adding the features.
    II. Begin with all the variables and keep removing the features.
    III. Use RFE to find the top 'n' features, and manually select the features to build the model.
6. **Model Training**: Train the model using the training set, adjusting the parameters to minimize the sum of squared residuals.
7. **Assumptions Checking (Residual Analysis):** Check if the error terms are normally distributed (which is in fact, one of the major assumptions of linear regression) using histogram.
8. **Model Evaluation**: Evaluate the model's performance on the testing set using metrics like
    a. P-value < 0.05
    b. Significant R^2
    c. VIF < 5
9. **Predictions**: Use the trained model to make predictions on new, unseen data.

## 2. Explain the Anscombe’s quartet in detail.

Answer: Anscombe's quartet helps show why it's essential to explore data thoroughly and not just rely on basic numbers. It points out that only looking at summary statistics might miss important details. The quartet highlights the need for using graphs and visuals to catch trends, outliers, and other important information that might not be clear from numbers alone.

Anscombe's quartet consists of four datasets, each containing eleven x-y pairs of data.

Let’s say there are 4 data sets which has statistical summary approximately similar. However, when these models are plotted on a scatter plot, each dataset produces a distinct type of plot that cannot be interpreted by any regression algorithm.



We can characterize the four datasets in Anscombe's quartet as follows:

- Dataset 1: Demonstrates a good fit for the linear regression model.
- Dataset 2: Cannot be modeled with a linear regression due to its non-linear nature.
- Dataset 3: Reveals outliers that linear regression struggles to handle.
- Dataset 4: Exhibits outliers that pose challenges for the linear regression model as well.

So, Anscombe’s quartet help us about the importance of visualizing data before applying various algorithms to build models.

## 3. What is Pearson’s R?

Answer: Pearson’s R is widely used to assess the degree of correlation or association between two sets of data points. The value of Pearson's R ranges from -1 to 1:

- Positive Correlation (0 < R < 1): Indicates a perfect positive linear relationship, meaning as one variable increases, the other variable also increases proportionally.
- No Correlation (R = 0): Suggests no linear relationship between the variables.
- Negative Correlation (-1 < R < 0): Indicates a perfect negative linear relationship, meaning as one variable increases, the other variable decreases proportionally.

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

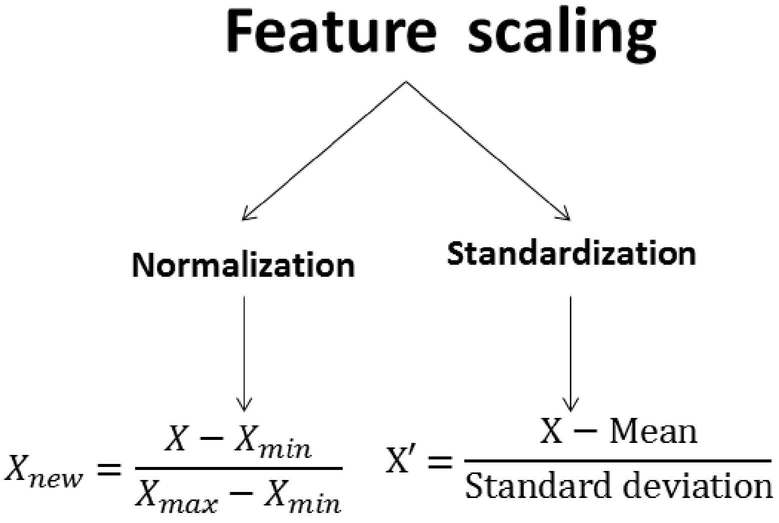
Answer: Scaling is a preprocessing step while building machine learning model that involves adjusting the range or distribution of variables to make them comparable. This is used to bring different variables to a common scale, preventing one variable from dominating the others due to differences in their magnitudes.

Scaling is performed for the following reasons:

Scaling ensures that variables with different units or magnitudes can be compared directly, and it helps algorithms converge faster and perform better.

**Normalized Scaling**: The variables are scaled in such all the values lies between 0 and 1 using max and min value of the data. This is also known as Min-Max scaling. This type of scaling used when the algorithm used is sensitive to the specific scale of the data.

**Standardized Scaling**: The variables are standardized in such a way that their mean is zero and S.D is 1. This type of scaling used when the algorithm is not scale-sensitive and perform better with normally distributed data.


Feature scaling
Normalization
Standardization
$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$
$X' = \frac{\text{X} - \text{Mean}}{\text{Standard deviation}}$

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

Answer: The VIF is defined as 1/(1-R^2). Here the value of VIF will tend to infinite when the denominator is 0. This will occur when the R value is 1 or -1. the correlation between at least two variables is perfect (correlation coefficient of ±1).

This scenario occurs when perfect multicollinearity exists among the predictor variables. Multicollinearity refers to a situation where two or more independent variables in a regression model are highly correlated, making it difficult for the model to estimate the individual effect of each variable.

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

Answer: A Q-Q plot is used to assess whether a given set of data follows a specific theoretical distribution, such as the normal distribution. It is used to check the normality assumption of the residuals/error terms.

Use and Importance in Linear Regression:

1. **Normality Assessment:** The Q-Q plot helps assess whether the distribution of residuals aligns with the normal distribution.
2. **Outlier Identification**: Outliers in the tails of the distribution can be identified through deviations from the expected quantiles.
3. **Model Validity Check**: Ensuring that the residuals are normally distributed.