<u>**Multiple Linear Regression**</u>

<u>Bike Sharing Assignment</u>

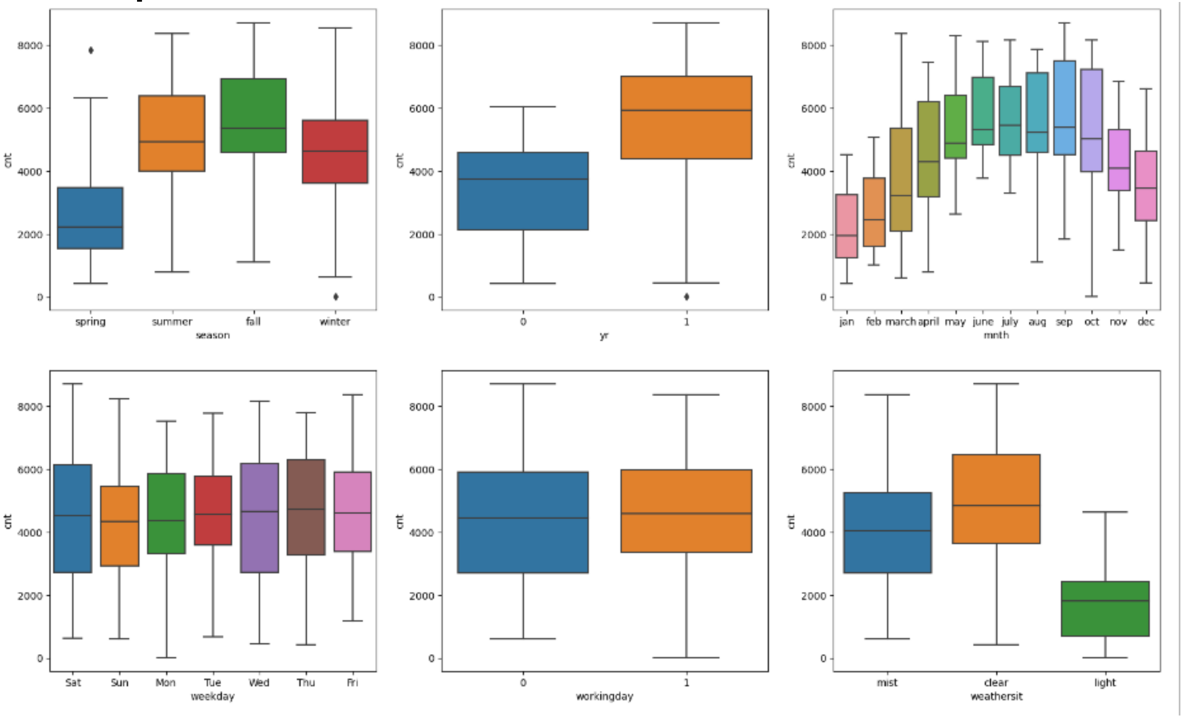
Assignment-based Subjective Questions:

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

Answer: In the given dataset, 'seasons', 'yr', 'weekday', 'holiday', 'workingday' and 'wethersit' are the categorical variables. Some of these variables have a significant effect on the dependent variable 'cnt'. Hence, correlation of these variables visualized using boxplot and inferred below points;

- More no. of bikes are rented in fall and summer **seasons**, and from May till October.
- More no. of bikes were rented in 2019(1) than in 2018(0) **years**.
- Variables like **weekday, holiday** and **workingday** don't reflect any significant pattern.
- When weather is **clear**, then more bikes are rented.

Below boxplot reflects the above inferences;



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

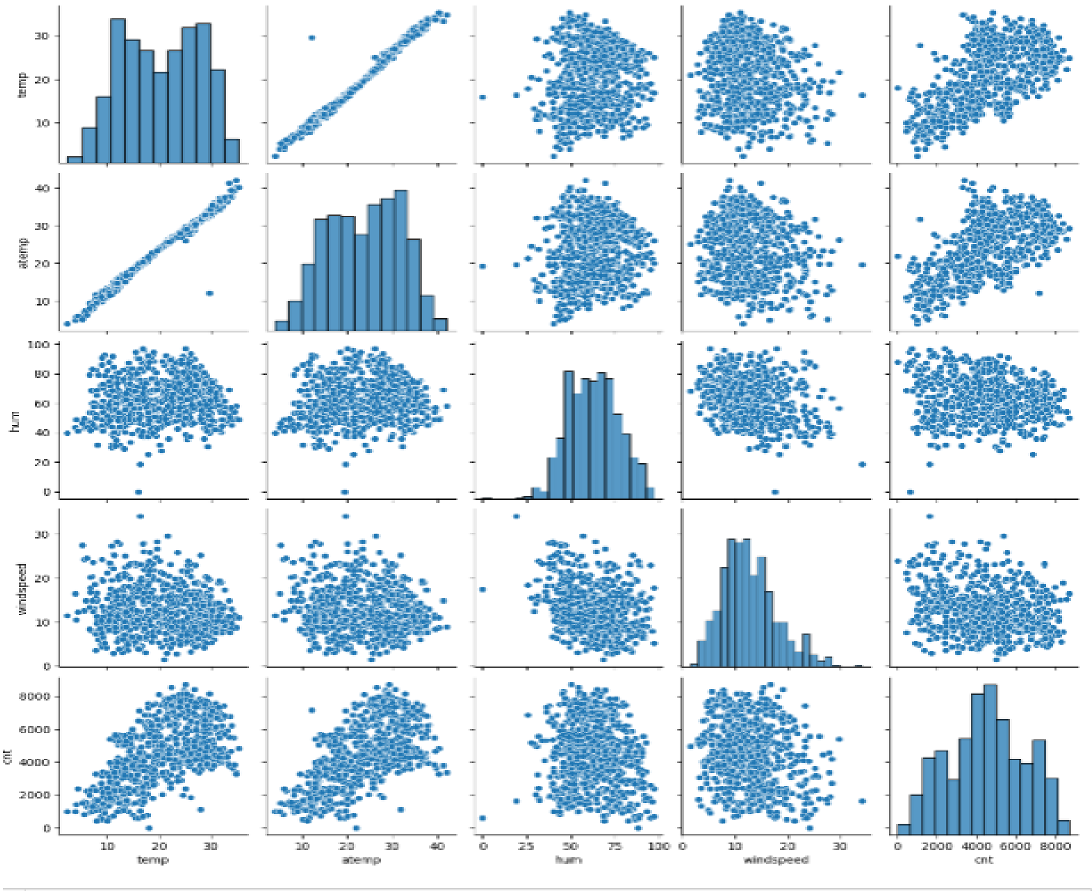
Answer: In a dataset, categorical variables should be treated with dummy variables in order to convert to numerical values, which eventually helps to determine the correlation of independent and dependent variables.

Dummy variables should be created for ‘n’ values of categorical variables is ‘n-1’ and hence drop_first = true helps to drop a first category after encoding, which prevents redundancy and multicollinearity issues.

For ex: 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obvious unfurnished. Hence, 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?**

Answer: Variables ‘temp’ & ‘atemp’ had high correlation with target variable ‘cnt’ at 63% or 0.63 contribution to the dataset.



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

Answer: Assumptions of multiple linear regression can be validated by applying concept of residual analysis on the training dataset after building and attaining the right model. There are four major assumptions in multiple linear regression are follows;

- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence**: Observations are independent of each other.
- **Normality**: For any fixed value of X, Y is normally distributed

From the above assumptions, Linearity and Normality can be visualized using **distplot** with x axis as 'Errors' and Y axis as "Density'. On the other hand, Homoscedasticity and Independence can be visualized using **scatterplot** with x axis as 'Predicted values' and y axis as 'Errors'.

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

Answer: Based on the model evaluation, below are top 3 features/variables contributing significantly towards explaining the demand of the shared bikes.

- **temp**: a 1-degree Celsius increase in temperature leads to increase in rental bikes
- **windspeed**: if the windspeed will increase by 1 unit, the number of rental bikes will decrease.
- **season_spring**: if the season will be spring, the number of rental bikes will decrease.

General Subjective Questions:

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

Answer: The linear regression algorithm is a statistical method used to understand the relationship between two or more variables. It's commonly used in predicting outcomes based on input variables. The algorithm assumes a linear relationship between the independent variables and the dependent. For example, in predicting house prices, the independent variables might be square footage, number of bedrooms, and location, while the dependent variable would be the price.

The algorithm calculates the best-fit line that minimizes the difference between the predicted values and the actual values in the dataset. This line is represented by the equation **y = mx + b**, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the intercept. The algorithm estimates the values of m and b based on the training data so that the line fits the data as closely as possible.

The linear regression algorithm involves several steps:

- Data Collection
- Data Preprocessing
- Model Building
- Model Evaluation
- Prediction

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties but vastly different graphical representations. This quartet was created by statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data before drawing conclusions based solely on statistical measures.

Each dataset in the quartet has the same mean, variance, correlation, and linear regression line. However, when plotted, they reveal distinct patterns such as linear relationships, non-linear relationships, outliers, and influential points. This demonstrates that relying solely on summary statistics can be misleading and highlights the need for visual exploration to fully understand the patterns and relationships in data.

3. What is Pearson's R?

Answer: Pearson correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1. Use the below Pearson coefficient correlation calculator to measure the strength of two variables.

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

Answer: Scaling is the process of transforming data to a common scale or range, to make comparisons easier. It's commonly used in data preprocessing before applying linear regression algorithms.

Variables with different scales can skew analyses. Scaling ensures that variables contribute equally to model building and comparisons. and also algorithms perform better and faster when features are on a similar scale. Scaling can also help in normalizing the data distribution, making it easier to interpret statistical measures.

Normalized Scaling is also known as **min-max scaling**, this method transforms data to a common range, between 0 and 1. It's calculated using the formula:

$$x' = (x - x_{min})/(x_{max} - x_{min})$$

Let's say if we have a dataset of student weight between 40 to 80 in range. After applying mix-max scaling, these weights will be transformed to a range between 0 and 1.

Standardized Scaling is also known as **z-score normalization scaling**, this method transforms data to have a mean of 0 and a standard deviation of 1. It's calculated using the formula:

$$z = \frac{x - \mu}{\sigma}$$

Let's says if we have a dataset of exam scores with different means and spreads, after standardized scaling, the scores will have a mean of 0 and a standard deviation of 1.

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

Answer: When the VIF value is infinite, it means there's a multicollinearity between variables. This happens when one or more independent variables can be exactly predicted by a combination of other independent variables. it's like having duplicate or redundant information in your data. This causes issues in linear regression analysis because it becomes impossible to figure out the contribution of each variable to predicting the outcome accurately.

If the VIF value is above 10, it's too high, and even values above 5 should be checked carefully. A high VIF indicates that two independent variables are strongly related. When they're perfectly related, we get an R-squared value of 1, which leads to an infinite VIF. To fix this, we should remove one of the variables causing this strong relationship from our dataset.

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

Answer: Quantile-Quantile (Q-Q) plot is a probability plot, is a graphical tool used to compare two probability distributions or datasets by plotting their quantiles against each other. It helps assess if a set of data matches a theoretical distribution, such as a normal, exponential, or uniform distribution

In linear regression, we use Q-Q plots to compare the distributions of the training and test datasets. This helps us determine if both datasets come from the same population with similar distributions or not..

The Q-Q plot can be applied to datasets of different sample sizes.

● This plot can detect various distributional characteristics such as shifts in location, scale, symmetry changes, and the presence of outliers.

● When used on two datasets, the Q-Q plot checks if they share a common distribution from the same population, have similar location and scale, exhibit similar distribution shapes, and demonstrate comparable tail behavior.