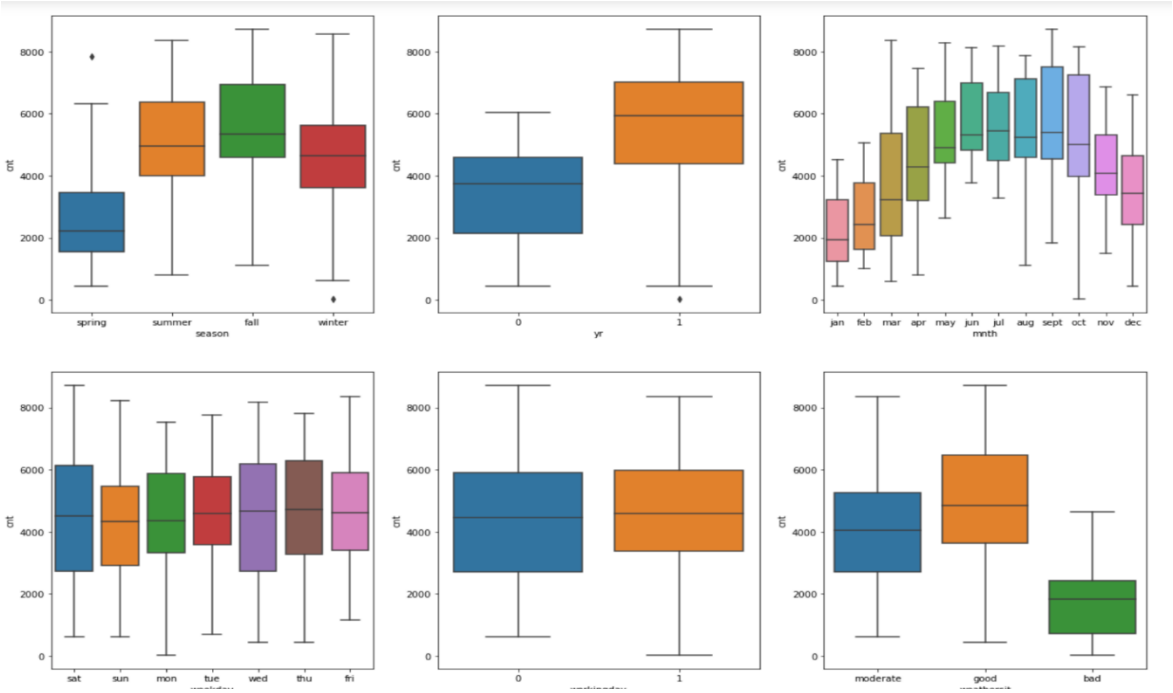
Assignment-Based Subjective Questions

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



Answer:

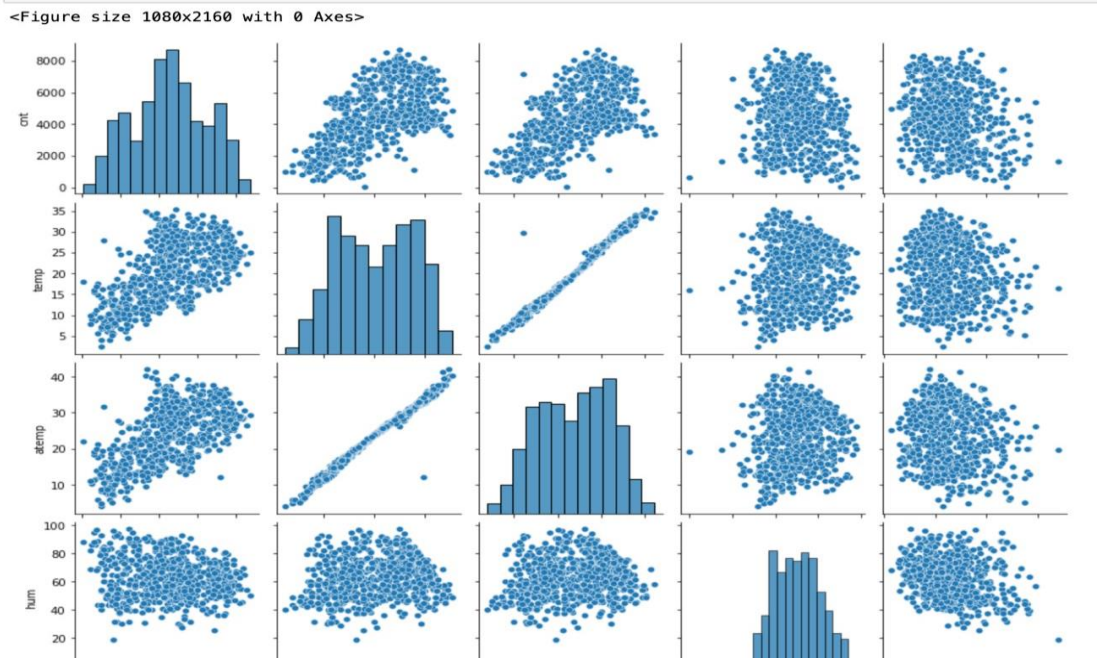
The dataset contains categorical variables, including season, year, holiday, weekday, workingday, weathersit, and month. These variables were visualized using a boxplot (refer to the attached figure). The impact of these variables on our dependent variable can be summarized as follows:

- Season: The boxplot indicates that the spring season has the lowest cnt (count), while the fall season has the highest cnt. Summer and winter exhibit intermediate cnt values.
- Weathersit: There are no users recorded during heavy rain or snow, indicating extremely unfavorable weather conditions. The highest count is observed when the weathersit is 'Clear, Partly Cloudy.'
- Year (Yr): Rentals in 2019 surpass those in 2018.
- Holiday: Rental counts decrease during holidays.
- Month (Mnth): September records the highest number of rentals, whereas December has the lowest. This observation aligns with the weathersit data, as December typically experiences heavy snowfall, which could explain the decrease in rentals.
- Weekday: The rental counts remain relatively consistent throughout the week.
- Workingday: The median count of users remains nearly constant throughout the week.

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

Answer: Setting **drop_first=True** during dummy variable creation in regression analysis is crucial for preventing multicollinearity. It omits one category reference to avoid redundancy and ensures that the model remains numerically stable. This step enhances model interpretability, preventing the "dummy variable trap" where two dummy variables become perfectly correlated. Without this option, regression models may produce unreliable results, leading to incorrect interpretations and potentially biased coefficient estimates. Thus, using **drop_first=True** is a standard practice to create robust and meaningful regression models.

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



Among all variables in the dataset, 'temp' and 'atemp' exhibit the strongest correlation with the target variable, 'cnt.

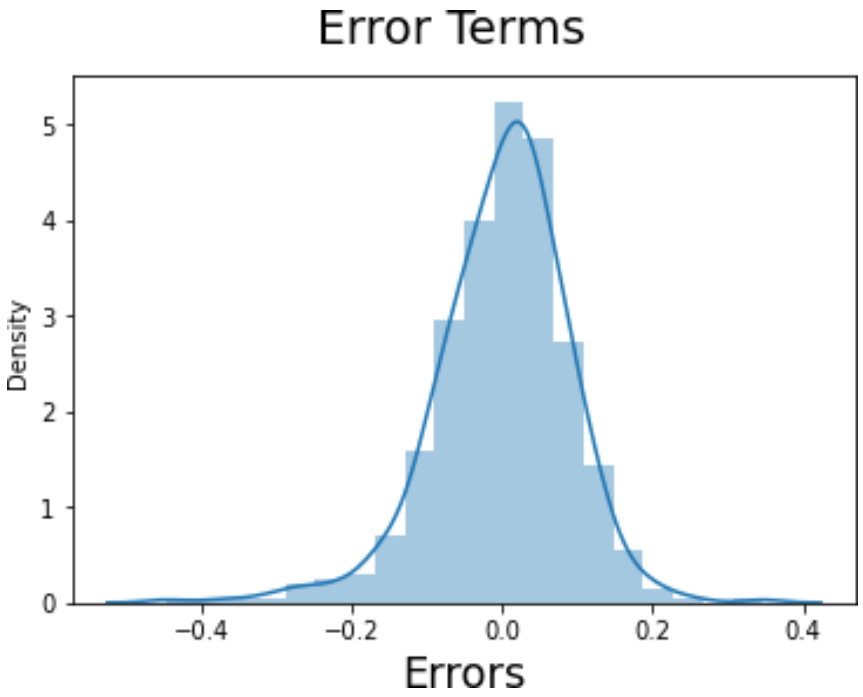
4. **How did you validate the assumptions of Linear Regression after buildingthe model on the training set?**

To validate the assumptions of linear regression, we employed a series of tests:

1. **Linearity Check:** Linear regression presupposes a linear relationship between independent and dependent variables. To assess this, we conducted a pairplot analysis of numeric variables. Refer to the notebook for a detailed exploration.

2. **Residual Normality:** The distribution of residuals must adhere to normality, with a mean centered around 0. We verified this assumption by plotting a distplot of residuals. The visualization demonstrated that residuals align closely with a mean of 0.

3. **Multicollinearity Examination:** Linear regression assumes minimal multicollinearity among independent variables. Multicollinearity occurs when predictors exhibit high correlation. To quantify this, we calculated the Variance Inflation Factor (VIF). For comprehensive insights, please consult the notebook. These assessments ensure the reliability of our linear regression model.



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

The most influential factors in explaining shared bike demand are temperature, year, and season. Temperature has a direct impact on rider comfort, while year and season capture broader trends and seasonal variations affecting bike usage patterns.

General Subjective Questions

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

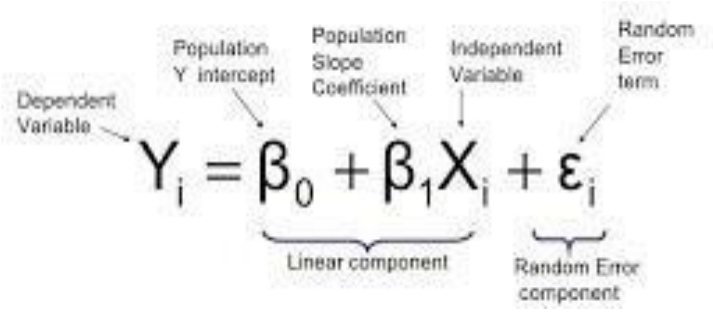
- Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis.Regression is the most commonly used predictive analysis model.

- 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 thepredictor(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 isof continuous data type and Predictors or independent variables could be of any data type likecontinuous, 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 thecoefficient 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 onlyone independent variable.

The equation for SLR will be:



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

The equation for MLR will be:

$$observed\ data \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

$$predicted\ data \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

$$error \rightarrow \quad \varepsilon = y - y'$$

B1 = coefficient for X1variable

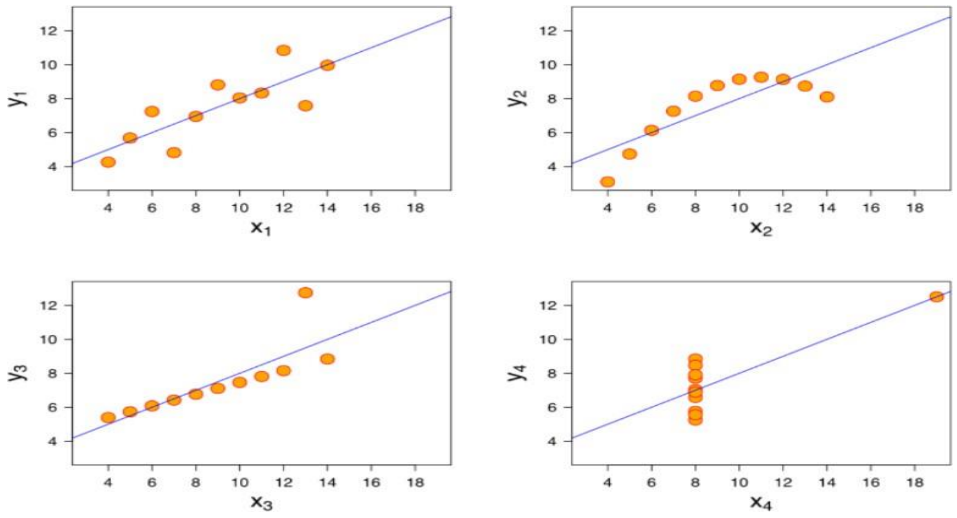
B2 = coefficient for X2 variable

B3 = coefficient for X3 variable and s on…

B0 is the intercept(constant term)

2. Explain the Anscombe's quartet in detail?

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance ofplotting the graphs before analyzing and modeling, and the effect of other observations on statistical properties. There are these four data set plots which have nearly the same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



- 1st data set fits linear regression model as it seems to be linear relationship between X and y
- 2nd data set does not show a linear relationship between X and Y , which meansit does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can't be handledby a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff. Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning mode

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.

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

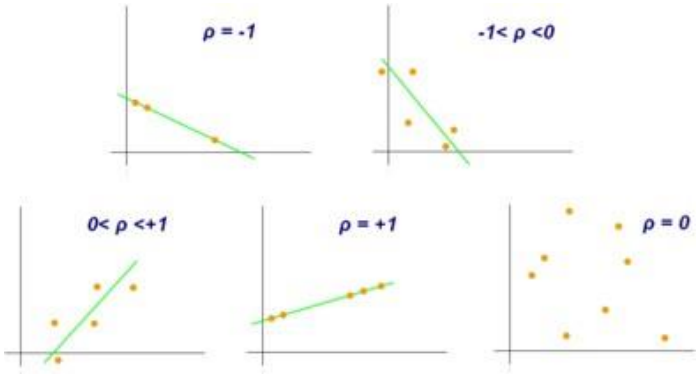
$\bar{y}$ = mean of the values of the y-variable

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 differencebetween normalized scaling and standardized scaling?

Feature scaling is a method used to normalize or standardize the range of independent variablesor 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 notfollow 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.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussiandistribution. 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. Whydoes this happen?

VIF - Variance Inflation Factor

The VIF 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 thefeature 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 independent variable which we want to check how well thisindependent variable is explained well by other independent variables. If that independent variablecan 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 thatthe 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.

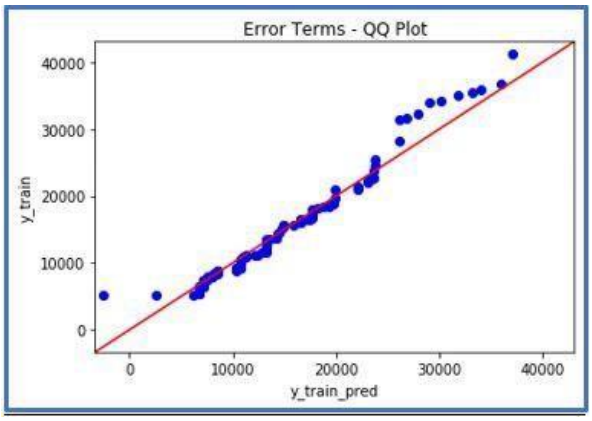
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot inlinear regression.

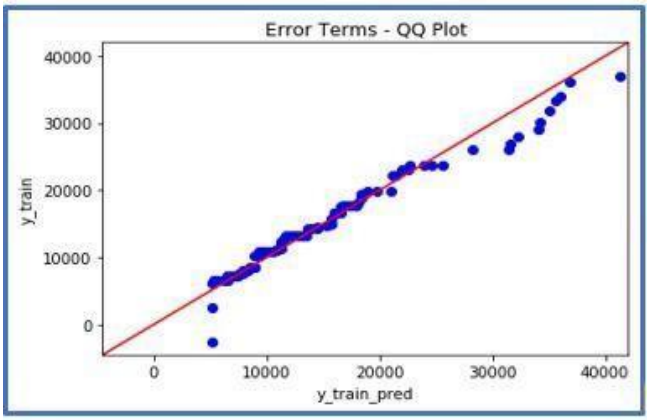
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Itis used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, 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?

Below are the possible interpretations for two data sets using a Q-Q plot:

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degreefrom x -axis
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.





3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.