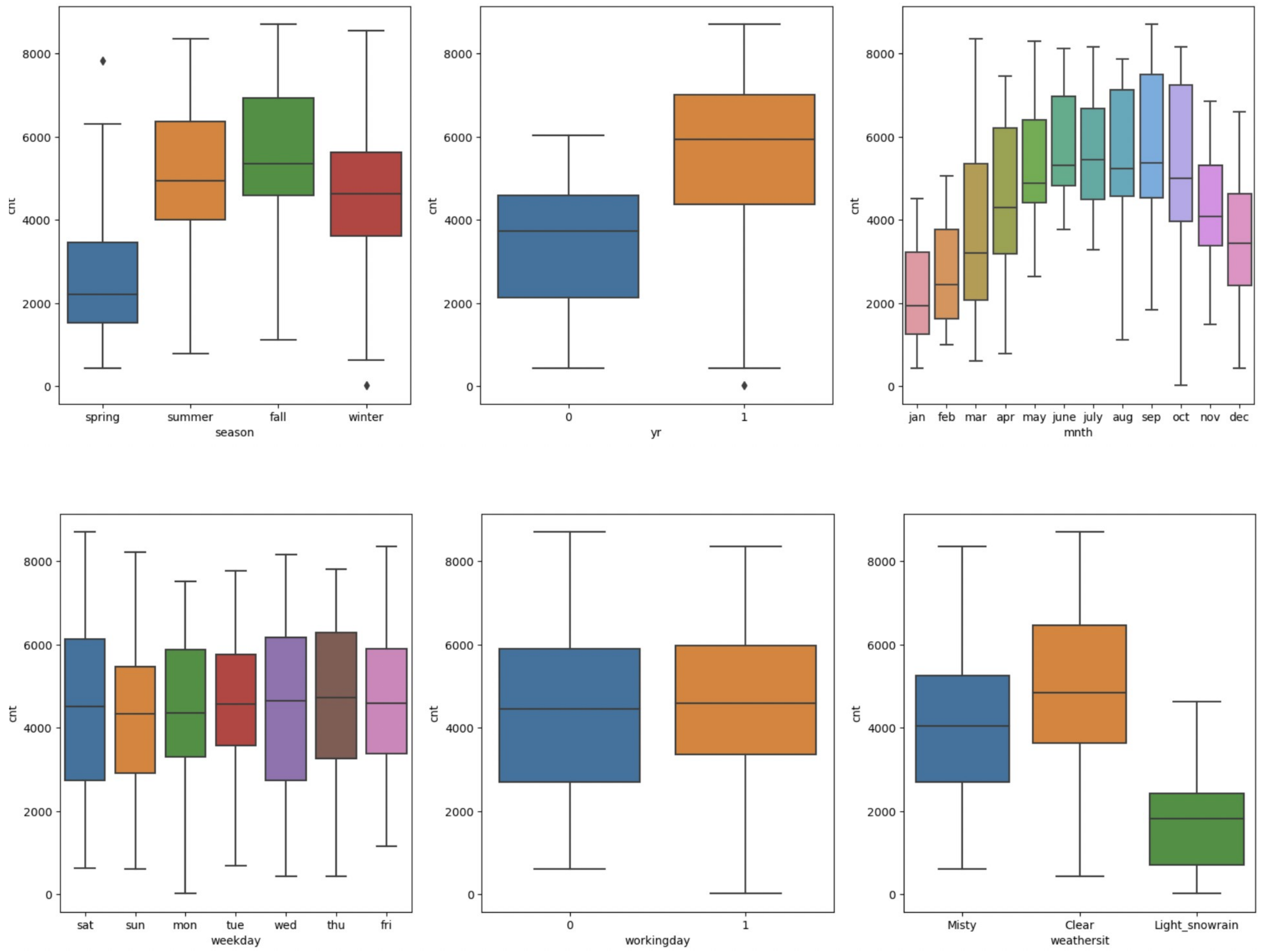
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

**Ans :**

There are a multiple categorical variables like season,yr,mnth,weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable ‘cnt’.



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

**Ans :**

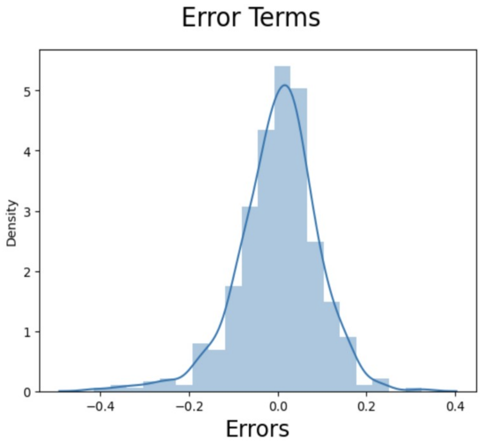
We have learned that categorical variable with ‘n’ levels, we need to create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

Eg: Here we have 4 levels for season ,first one will be dropped and 3 new columns will be created.

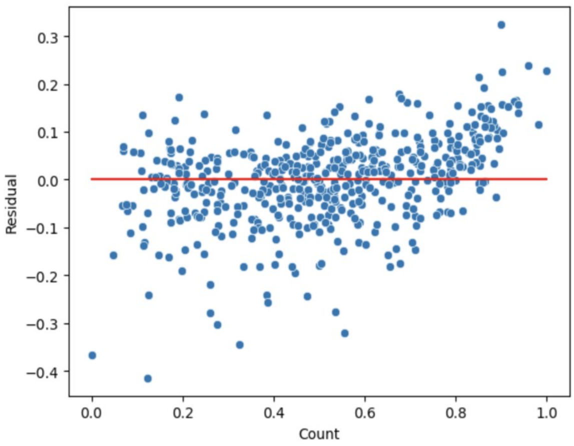
**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans:
The ‘**temp**’ and ‘**atemp**’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.

**4. How did you validate the assumptions of Linear Regression after building the model on the training set?**
Ans:
Linear Regression models are validated based on Linearity, No auto-correlation,Normality of error, Homoscedasticity, Multicollinearity.



Error terms are following normal distribution



**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
Ans:
Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season.

**1.Explain the linear regression algorithm in detail.**
Ans :
Linear regression is a form of predictive modeling technique which tells us the relationship between the dependent and independent variables .
It shows the linear relationship i.e how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, it is called **multiple linear regression**. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for b0 and b1 to find the best fit line and the best fit line should have the least error. RFE (Recursive feature elimination)or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for b0 and b1, which provides the best fit line for the data points.

**2. Explain the Anscombe’s quartet in detail.**
Ans :
Anscombe’s quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

**3. What is Pearson’s R?**
Ans:
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.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
Ans :
Scaling is transforming the data so that it fits within a specific scale and it helps in speeding of calculation in algorithm .Also if scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

| **Normalizing Scaling** | **Standardize Scaling** |
|---|---|
| Scales values between (0,1) or (-1,1) | Not bounded in a certain range. |
| Minimum and maximum value of features being used | Mean and standard deviation is used for scaling |
| It is used when features are of different scales | Used to ensure zero mean and unit standard deviation |
| Used when we don’t know about the | Used when distribution is normal. |

|  |  |
|---|---|
| distribution. |  |

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
**Ans**:
VIF helps to explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
-A VIF value of greater than 10 is considered high,
-A VIF greater than 5 should also not be ignored and inspected appropriately.

A very high VIF value shows a perfect correlation between two independent variables. In case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
Ans:
Q–Q plot is a probability plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
It is a graphical tool to help us check if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear.
The linearity assumption can best be tested with scatter plots.
**Importance of QQ Plot in Linear Regression :**
In Linear Regression when we have a train and test dataset then we can create Q-Q plot which we can confirm that both the data train and test data set are from the population with the same distribution or not.
Advantages:
● It can be used with sample size also
● Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
● If both datasets came from population with common distribution
● If both datasets have common location and common scale
● If both datasets have similar type of distribution shape
● If both datasets have tail behavior