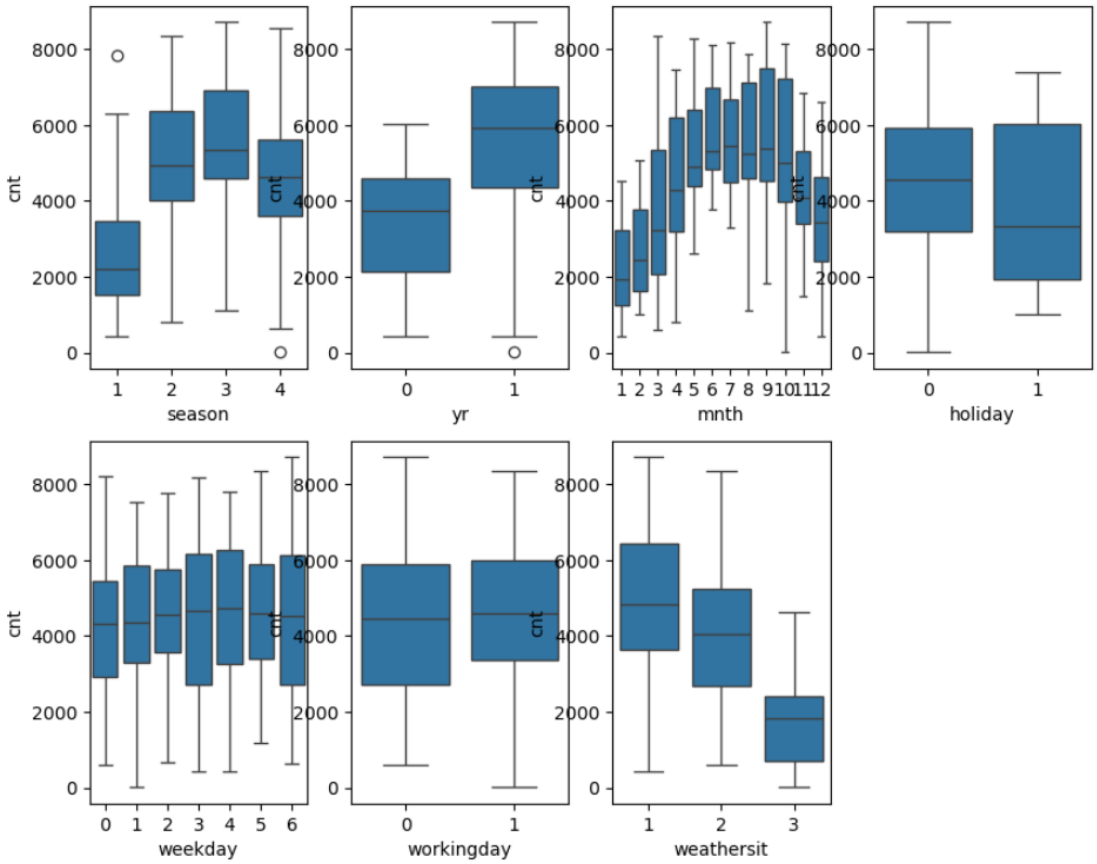
Assignment-based Subjective Questions

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

List of categorical variables: season, yr, mnth, holiday, weekday, workingday, weathersit

Visualizing their impact on 'cnt' using box plot:



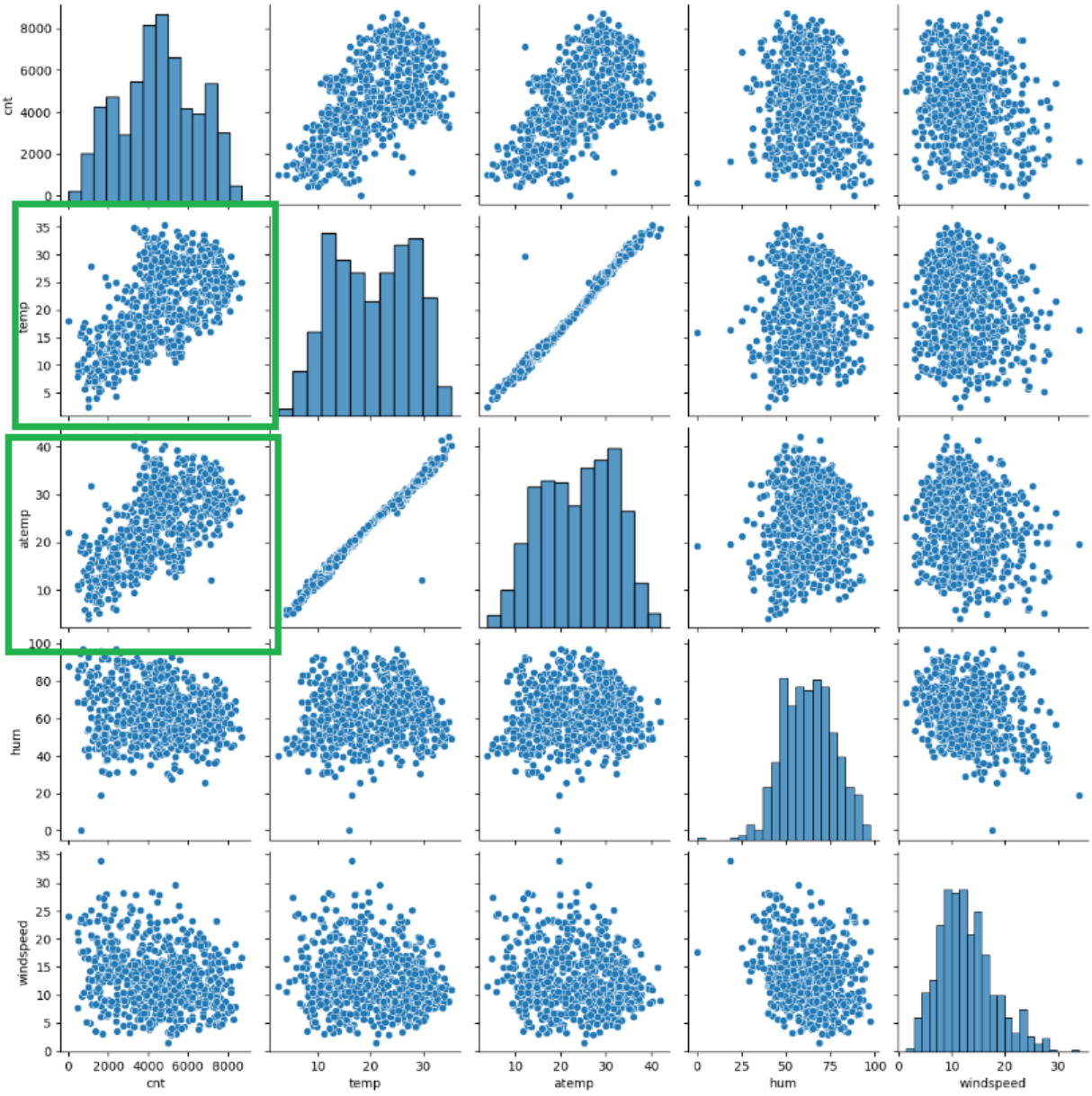
All the categorical variables mentioned above except, weekday and Workingday have a significant effect on the bike demand. i.e., 'cnt' variable

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

In multiple linear regression, using drop_first=True when creating dummy variables is important to avoid multicollinearity. Multicollinearity occurs when independent variables are highly correlated, which can distort the regression coefficients and make the model unstable. By dropping the first dummy variable, we avoid this issue by ensuring that the dummy variables are independent, leading to more reliable and interpretable regression results.

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

'Temp' and 'atemp' seems to have the highest correlation with target variable. Later we identified both temp and atemp are dependent on each other and we dropped atemp. So we can consider temp to be the highest correlating variable.



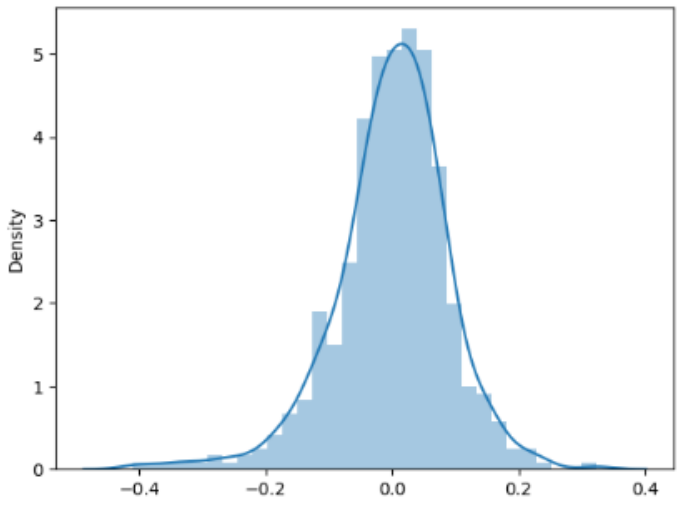
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

We could confirm,

- Error terms are normally distributed with mean zero

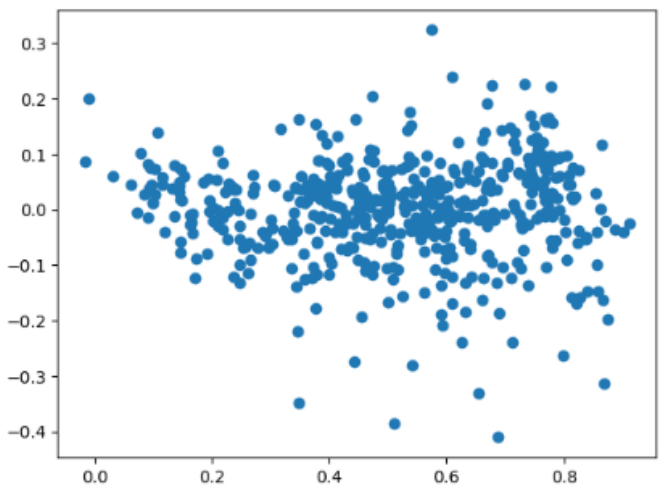
```
plt.figure()
res=y_train-y_train_pred
sns.distplot(res)
```

```
65]: <Axes: ylabel='Density'>
```



Error terms are independent of each other.

```
[77]: plt.scatter(y_train_pred,res)
plt.show()
```



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

Below are the coef value of final features. Below highlighted ones would be the top 3 contributing features.

const	0.223600
yr	0.228443
temp	0.598613
hum	-0.175693
windspeed	-0.189754
season_summer	0.083390
season_winter	0.134873
mnth_July	-0.043632
mnth_September	0.091732
weekday_Sunday	-0.041797
weathersit_Light Snow	-0.234148
weathersit_Mist & Cloudy	-0.051438

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

Types of Linear Regression
There are two main types of linear regression:

Simple Linear Regression
This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$y=\beta 0+\beta 1X$
where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\ldots\beta nX$

where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

The primary focus is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

This demonstrates 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.

Purpose of Anscombe's Quartet

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3.What is Pearson's R?

(Pearson's Correlation Coefficient)

Pearson's R is a measure of the linear relationship between two continuous variables.

Range:

- **1**: Perfect positive correlation (both variables increase together).
- **-1**: Perfect negative correlation (one variable increases while the other decreases).
- **0**: No linear correlation.

Interpretation:

- **Positive R**: Indicates a positive relationship (as one variable increases, so does the other).
- **Negative R**: Indicates a negative relationship (as one variable increases, the other decreases).

Pearson's R helps in quantifying the strength and direction of a linear relationship between two variables.

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

Scaling

Definition: Scaling is the process of adjusting the range of features in your data to ensure they are on a similar scale.

Purpose:

- **Improves Model Performance**: Many machine learning algorithms perform better or converge faster when features are on a similar scale.
- **Prevents Bias**: Prevents features with larger ranges from dominating the model.

Types of Scaling

1. **Normalized Scaling**:
    - **Definition**: Rescales the feature values to a range of [0, 1].
    - **Formula**: Normalized value=(X−Xmin)/(Xmax−Xmin)
    - **Use Case**: Useful when the data does not follow a Gaussian distribution or when you want features in a specific range.
2. **Standardized Scaling**:
    - **Definition**: Rescales the feature values to have a mean of 0 and a standard deviation of 1.
    - **Formula**: Standardized value=(X−μ)/σ
    - **Use Case**: Useful when the data follows a Gaussian distribution and when features are required to have a standard normal distribution (mean = 0, standard deviation = 1).

In summary, scaling ensures that different features contribute equally to the model's performance by adjusting their ranges. Normalized scaling puts data in a [0, 1] range, while standardized scaling adjusts data to have a mean of 0 and a standard deviation of 1.

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

This occurs when one predictor variable is an exact linear combination of one or more other predictor variables. In other words, there's a perfect correlation between some predictor variables.

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

A Q-Q (Quantile-Quantile) plot assesses if data follows a theoretical distribution, like normal distribution.

Use:

- **Assess Normality**: Checks if residuals are normally distributed.
- **Detect Outliers**: Identifies outliers affecting model reliability.
- **Validate Assumptions**: Ensures model assumptions (like normality) are met for accurate predictions.