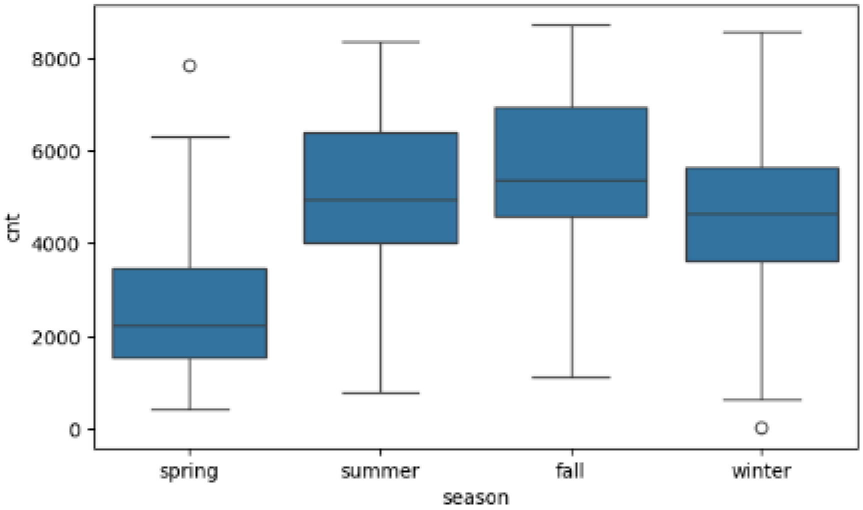
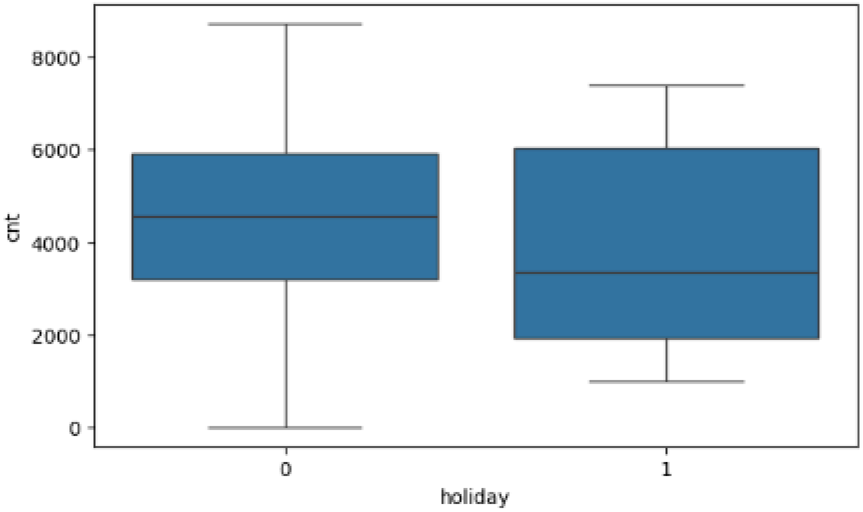
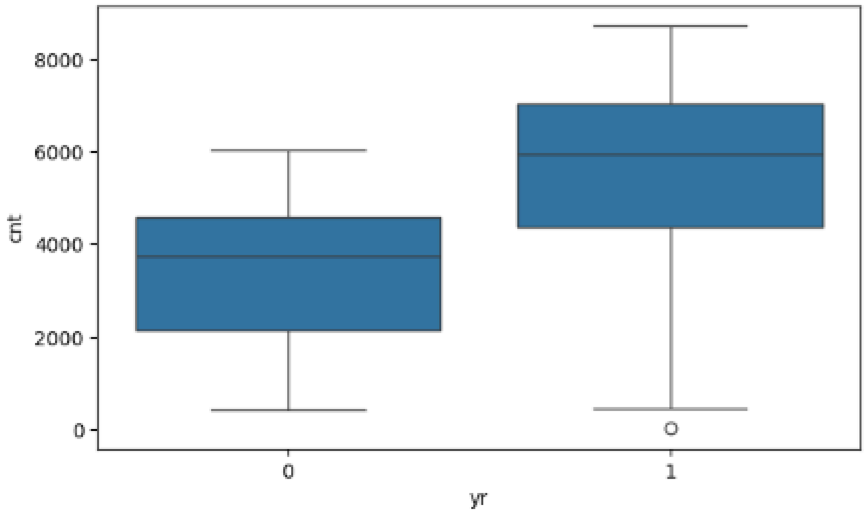
Assignment-based Subjective Questions:

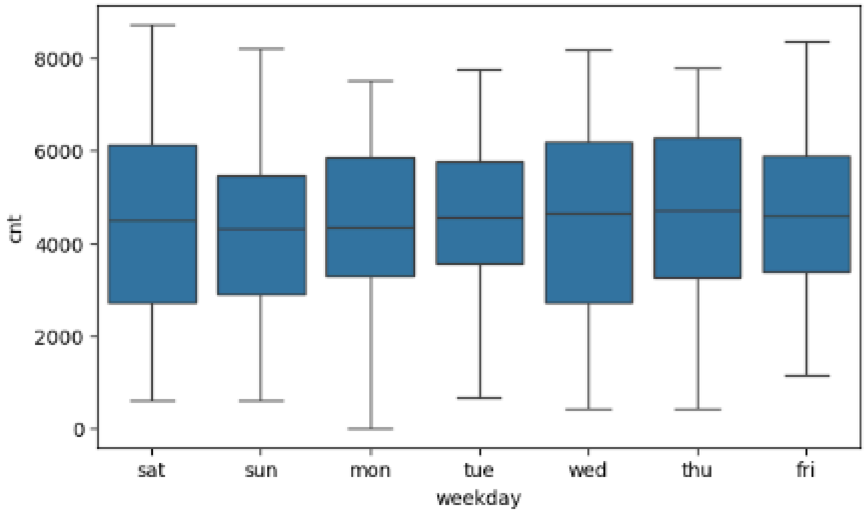
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

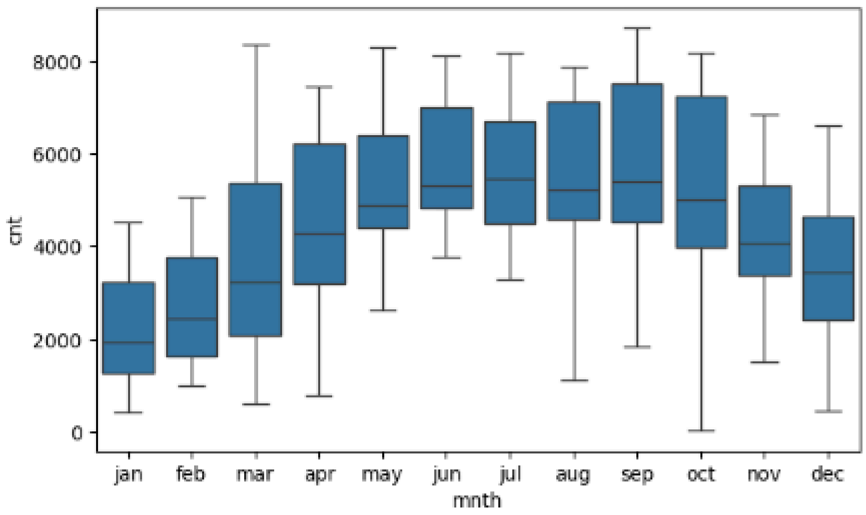
The categorical variables in the dataset are **season, yr, mnth, holiday, weekday, workingday, weathersit.** The effect of these variables on the dependent variable ‘cnt’ is shown below:

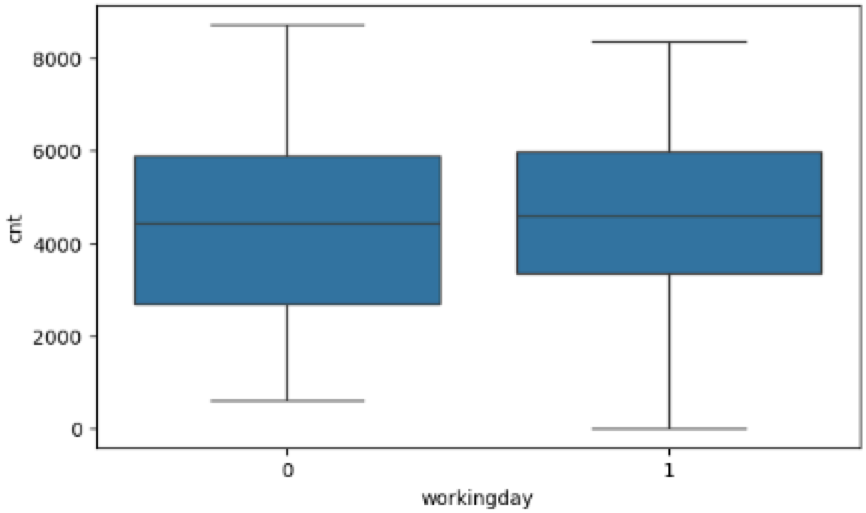


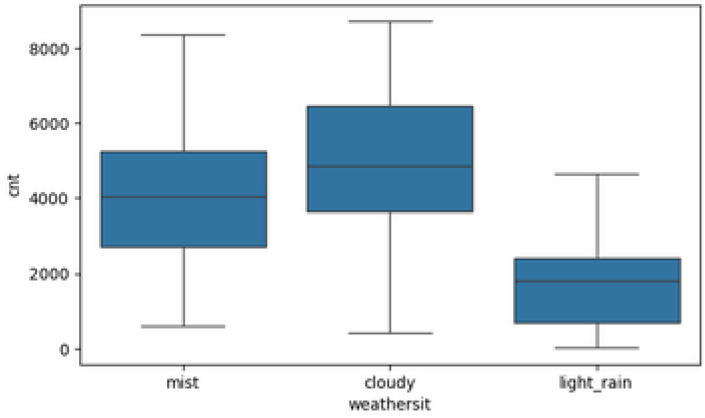



8000
6000
4000
2000
0
cnt
0
1
yr


8000
6000
4000
2000
0
cnt
sat
sun
mon
tue
wed
thu
fri
weekday







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

For a categorical variable with ‘n’ levels, we create ‘n-1’ columns. Each column indicates whether that particular level exists or not. **drop_first=True** is used so that the resultant dataframe can have ‘n-1’ columns. drop_first will drop the first column when creating dummies.

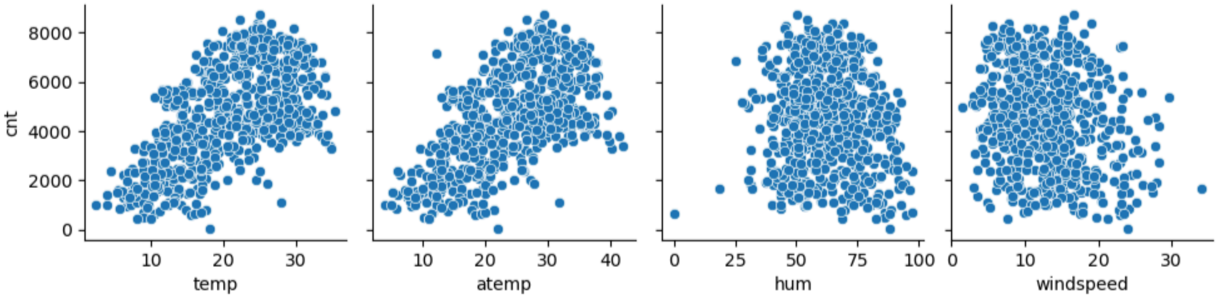
Let’s say we want to express if a house is furnished or semi-furnished or unfurnished, we can have 2 columns unfurnished, semi-furnished. In a row,

- if the values are 1,0 for these 2 columns, then the house is unfurnished
- If the values are 0,1 for these 2 columns, then the house is semi-furnished
- If the values are 0,0 for these 2 columns, then the house is furnished

We can see that the data with 3 levels can be expressed with just 2 columns, so having the 3rd column is redundant.

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

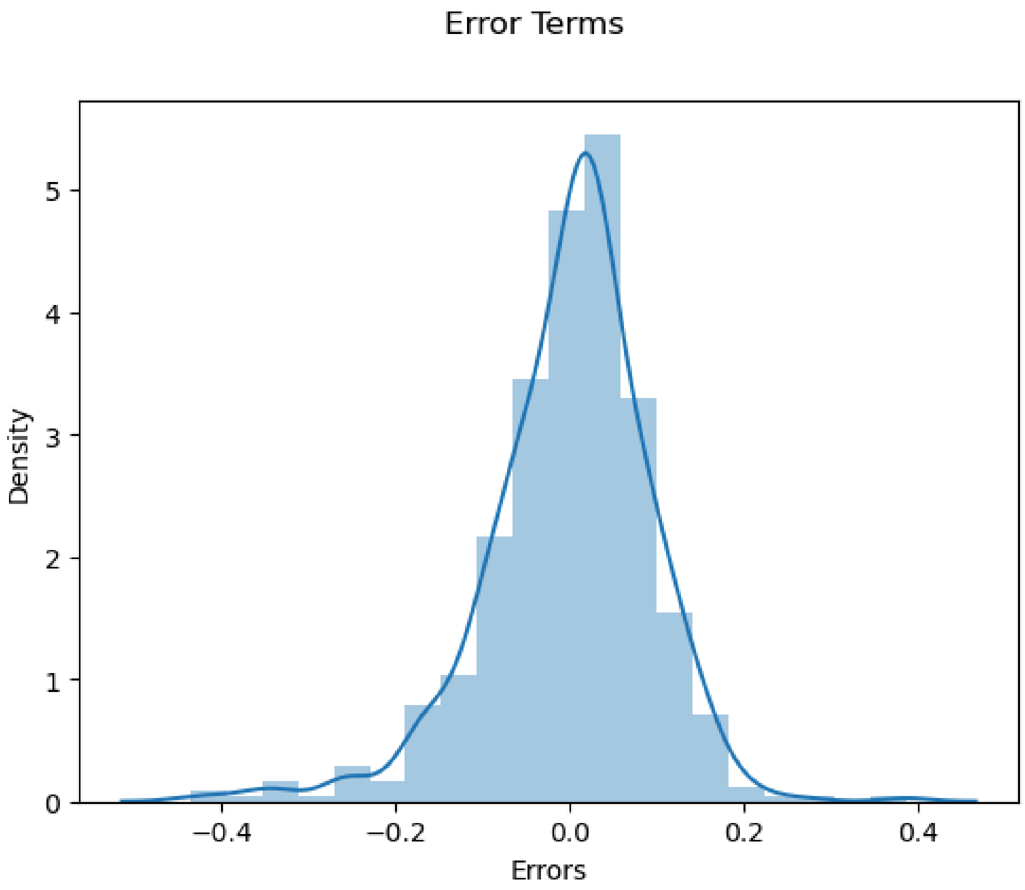
From the pair-plot among the numeric variables, the temp & atemp columns have highest correlation with the target variable.



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

The assumptions of Linear Regression can be validated by **plotting the error terms**:
sns.distplot((y_train-y_train_pred),bins=20)

- The variables x&y should have a linear relationship
- The error terms should follow a normal distribution with zero mean
- The error terms are independent of each other
- The error terms have constant variance (**homoscedasticity**)



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

- Year
- Temperature
- Season

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting line (or hyperplane in multiple dimensions) through the data points.

The goal of linear regression is to model the relationship between a dependent variable y and one or more independent variables x by fitting a linear equation to observed data. The equation for linear regression is:

$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta pxp+\epsilon y=\beta0+\beta1x1+\beta2x2+\cdots+\beta pxp+\epsilon$

Linear regression has the following assumptions:

- The variables x&y should have a linear relationship
- The error terms should follow a normal distribution with zero mean
- The error terms are independent of each other
- The error terms have constant variance (**homoscedasticity**)

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 the coefficients to find the best fit line and the best fit line should have the least error.

The model is evaluated using the value of R-squared. R-squared explains the variance in the target variable that is predicted from the independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a collection of four datasets that have nearly identical statistical properties but appear very different when graphed. Created by the statistician Francis Anscombe in 1973, this quartet is designed to illustrate the importance of graphical data analysis and demonstrate that summary statistics alone can be misleading. Here's a detailed explanation:
Purpose and Concept

Anscombe's Quartet is intended to show that relying solely on summary statistics (such as mean, variance, correlation) to understand data can be misleading. It emphasizes the importance of visualizing data to uncover underlying patterns and anomalies that summary statistics might not reveal.
The Four Datasets

Each dataset in Anscombe's Quartet consists of 11 pairs of (x, y) values. Despite their similarity in summary statistics, the datasets have very different distributions and patterns when plotted. Here are the key characteristics of each dataset:

Dataset I:

This dataset shows a linear relationship between xx and yy, with a small amount of noise.

Plot: A scatter plot will reveal a clear, straight line relationship.
Summary Statistics:
Mean of xx: 9
Mean of yy: 7.5
Correlation between xx and yy: 0.816

Dataset II:
This dataset also shows a linear relationship but with a clear outlier that significantly affects the results.
Plot: The scatter plot will show a similar trend to Dataset I but with a noticeable outlier that disrupts the line.
Summary Statistics:

Mean of xx: 9
Mean of yy: 7.5
Correlation between xx and yy: 0.816

Dataset III:

Description: This dataset demonstrates a non-linear relationship where yy is a quadratic function of xx, with the data forming a curved pattern.
Plot: The scatter plot will reveal a parabolic trend rather than a straight line.
Summary Statistics:

Mean of xx: 9
Mean of yy: 7.5
Correlation between xx and yy: 0.816

Dataset IV:

Description: This dataset features a relationship between xx and yy that is influenced by a single extreme value, which has a large effect on the overall pattern.

Plot: The scatter plot will show a trend similar to Dataset I but with a different kind of deviation due to the influence of the extreme value.

Summary Statistics:

Mean of xx: 9
Mean of yy: 7.5
Correlation between xx and yy: 0.816

3. What is Pearson's R?

Pearson's rrr, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is one of the most widely used measures of correlation and is defined as follows:

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

Assumptions:

1. **Linearity**: The relationship between the variables is linear.
2. **Homogeneity of Variance**: The spread of data points is roughly the same across all levels of the variables.
3. **Normality**: Both variables are approximately normally distributed (this is particularly important for hypothesis testing).

Applications:

- **Exploratory Data Analysis**: Helps identify and quantify the strength of relationships between variables.
- **Statistical Modeling**: Used in regression analysis to understand the degree to which predictors are related to the response variable.
- **Correlation Analysis**: Useful in fields like psychology, finance, and social sciences to investigate relationships between variables.

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

Scaling refers to the process of transforming data so that all features fit within a specified range or scale. This is an essential preprocessing step in data analysis and machine learning, as it helps improve the efficiency and accuracy of algorithms.

Data often come with features that vary widely in magnitude, units, and range. Without scaling, algorithms may disproportionately prioritize features with larger values, which can lead to biased or incorrect model performance. By standardizing or normalizing the data, you ensure that each feature contributes equally to the model, thereby enhancing the algorithm’s ability to learn from all features effectively.

Difference between Normalized Scaling and Standardized Scaling:
1. In normalized scaling minimum and maximum value of features being used whereas in Standardized scaling, mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.

3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
When some variables are able to create a perfect multiple regression on other variables (multicollinearity is perfect), the VIF value tends to infinity.

6. 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 is a type of probability plot that graphically compares two probability distributions by plotting their quantiles against each other. It helps determine if a dataset likely originates from a theoretical distribution, such as the Normal, Exponential, or Uniform distribution.

A Q-Q plot can also be used to assess whether two distributions are similar. If the distributions are similar, the Q-Q plot will generally show a linear pattern. To evaluate this linearity, scatter plots are particularly useful. Additionally, linear regression analysis often requires the assumption of multivariate normality for all variables, which can be checked using histograms or Q-Q plots.

Uses of a Q-Q Plot in Linear Regression:

1. Assessing Normality of Residuals:
2. Detecting Deviations from Normality
3. Evaluating Model Fit

Importance of a Q-Q Plot in Linear Regression:

1. Validating Assumptions
2. Identifying Outliers and Influential Points
3. Model Improvement
4. Enhancing Predictive Accuracy