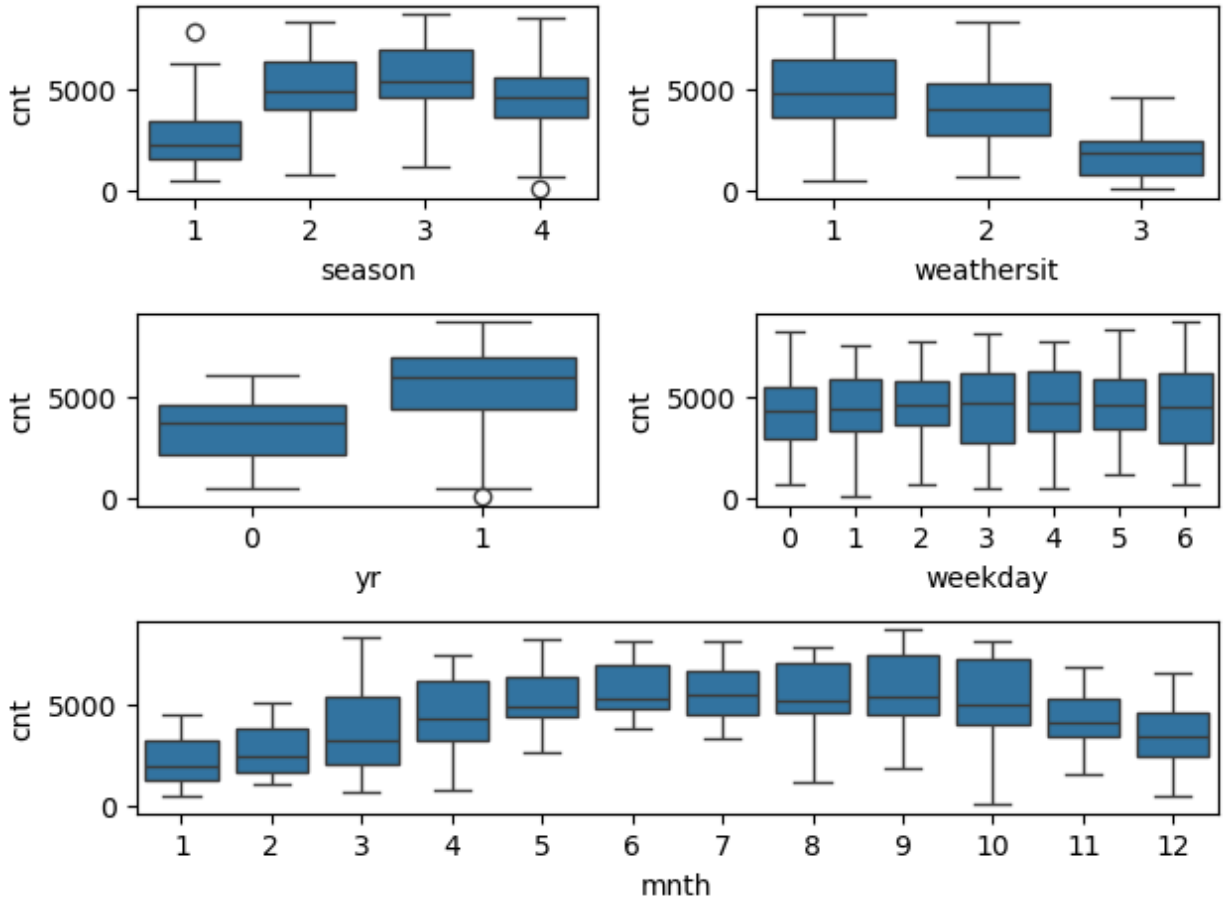
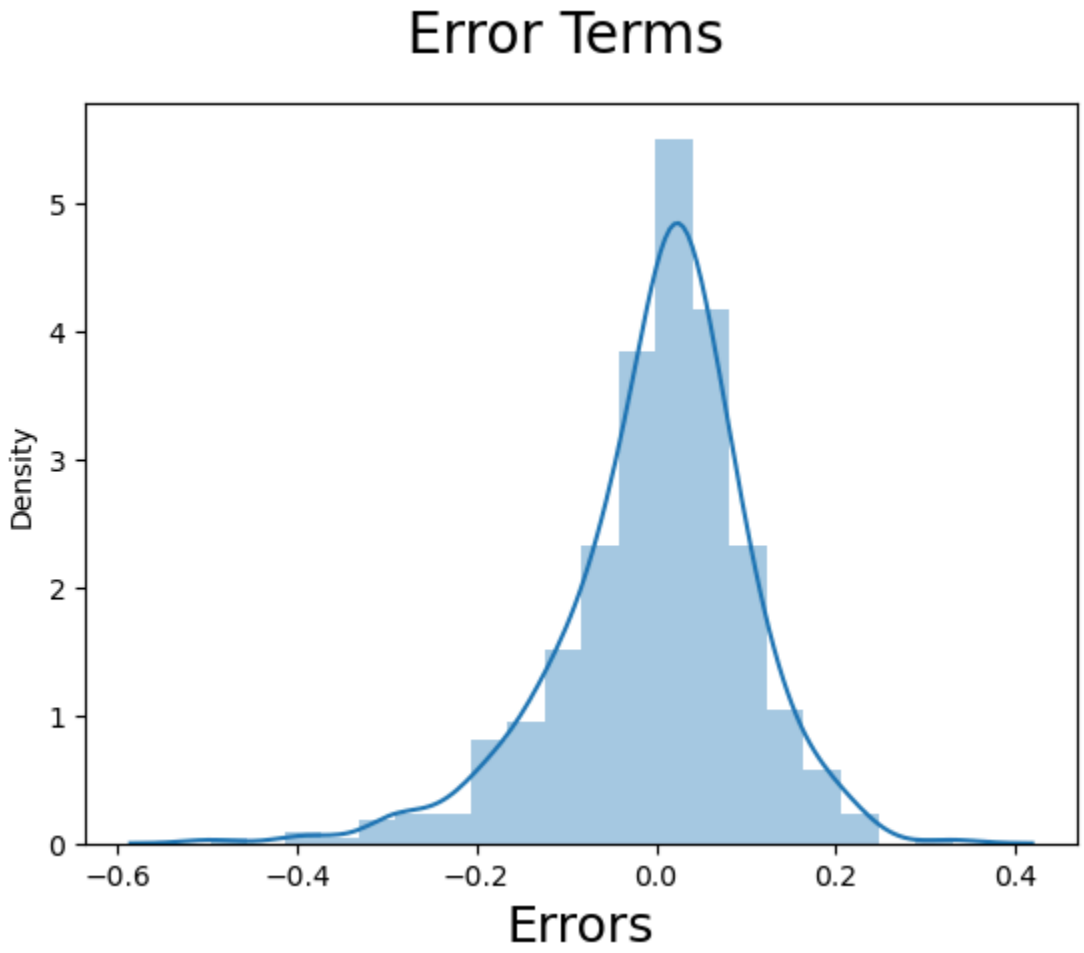
Assignment-based Subjective Questions

1. The following can be inferred from each of the categorical variables:
    a. season: summer and fall have the most riders and thus affect the dependent variable the most
    b. weathersit: most riders ride during clear, partly cloudy weather. On the other hand, the least riders are seen during light snow or rainy conditions.
    c. yr: most riders were seen during the year 2019.
    d. weekday: there isn't much of a difference between the median ridership during the days of the week.
    e. mnth: the summer months (June, July, August) have the highest median ridership, whereas the winter months (Dec, Jan, Feb) have the least.

    Overall, there is a significant effect of the categorical variables on the dependent variable.



2. It is important to use **drop_first=True** because we always need (n-1) dummy variables for n values of categorical variables. This helps in reducing the correlation between the dummy variables.
3. **temp** and **atemp** have the highest correlation with the target variable.
4. The following assumptions were validated:
    a. Homoscedasticity: The residuals are normally distributed according to the plot:



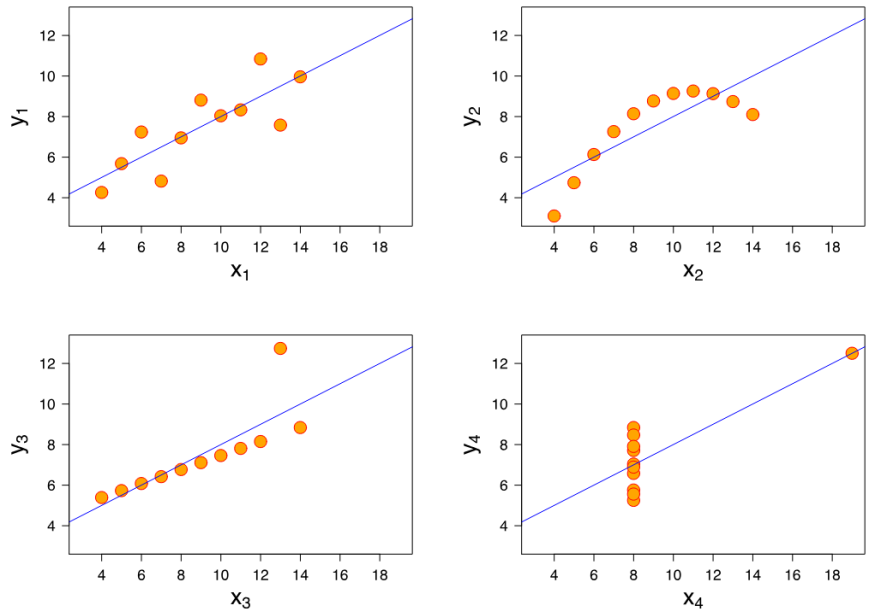
    b. No multi-collinearity: The VIFs were calculated according to the table below. As you can see, none of them have a VIF value more than 10.

	Features	VIF
2	temp	6.95
3	windspeed	5.02
1	workingday	4.20
0	yr	2.02
7	spring	1.96
6	weekday_sat	1.67
8	winter	1.45
4	mnth_jul	1.44
5	mnth_sep	1.19
9	light snow	1.08

5. The top 3 features contributing significantly towards explaining the demand of shared bikes are:
    a. temp : coeff of 0.4927
    b. year: coeff of 0.2391
    c. season (the presence of light snow): coeff of -0.2553

General Subjective Questions

1. Linear regression is a statistical technique used to model the relationship between a dependent variable (y) and one or more independent variables (predictors x1, x2, x3…). The goal of linear regression is to find the best-fitting linear relationship between the dependent and independent variables.
   The following steps are involved in linear regression:
    1. Start by assuming the relationship between the dependent and independent variables can be approximated by a linear equation. The task is to find the coefficients β_0 , β_1 etc.
    2. Estimate the coefficients using the OLS (ordinary least squares) method.
    3. Once the coefficients are estimated, predictions can be made using the linear model.
    4. Evaluate the model:
        a. R-squared value
        b. Adjusted R-squared value
        c. Mean square error
    5. Validate the model's assumptions. You need to check the linearity, independence, homoscedasticity, normality of residuals, etc. to ensure the model is valid.
2. The Anscombe's Quartet is a collection of four datasets that have nearly identical statistical properties (mean, variance, correlation, linear regression line) but are graphically distinct when plotted. The following are the properties of Anscombe's Quartets:
    a. Mean of the x-values is the same in all datasets.
    b. Mean of the y-values is the same in all datasets.
    c. Variance of the x-values is the same in all datasets.
    d. Variance of the y-values is the same in all datasets.
    e. Correlation coefficient between the x and y variables is approximately the same in all datasets. (0.816)
    f. The least-squared regression line is almost identical across the datasets.
    g. R-squared value is also nearly identical for each dataset.



The key lesson from Anscombe's Quartet is the importance of visualization of data before conducting statistical analysis. Despite having nearly identical summary statistics, the datasets reveal patterns when plotted which could lead to vastly different conclusions if these plots were not examined.

3. Pearson's r is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the letter r and ranges from -1 to 1.
    a. R = 1 : Perfect positive linear correlation.
    b. R = -1: Perfect negative linear correlation.
    c. R = 0: No linear correlation.
4. Scaling is the process of transforming the features (variables) in a dataset so that they are on a comparable scale. This is especially important when the features have different units or magnitudes, as it ensures that no single feature disproportionately influences the model.

    Scaling is performed for the following reasons:
    1. Improves model performance
    2. Ensures faster convergence
    3. Enhances interpretability
    4. Reduces sensitivity to outliers

    Difference between normalization and standardization:
    Normalization: aka min-max scaling is the process of rescaling the values of a feature to a range 0 to 1. This method gets completely rid of outliers.

Standardization: aka z-score normalization transforms the data so that it has a mean of 0 and a standard deviation of 1. This method centers the data around the mean and scales it based on standard deviation. This method preserves outliers.

5. This can happen in situations where there is perfect multicollinearity between one predictor variable and one or more of the other predictor variables. Perfect multicollinearity occurs when one predictor variable is an exact linear combination of the other predictor variables. According to the formula, here r squared would be 1, so the VIF value would be 1/1-1 which gives infinity.

6. A QQ 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.

In the context of linear regression, q-q plots are primarily used to check the assumption that the residuals are normally distributed. Normality of residuals is an important assumption because it ensures the validity of statistical tests (like t-tests) and reliability of confidence intervals. Further, q-q plots can also be used to detect outliers. Points that fall far from the reference line in a q-q plot indicate potential outliers in the data. It also helps identify Kurtosis.