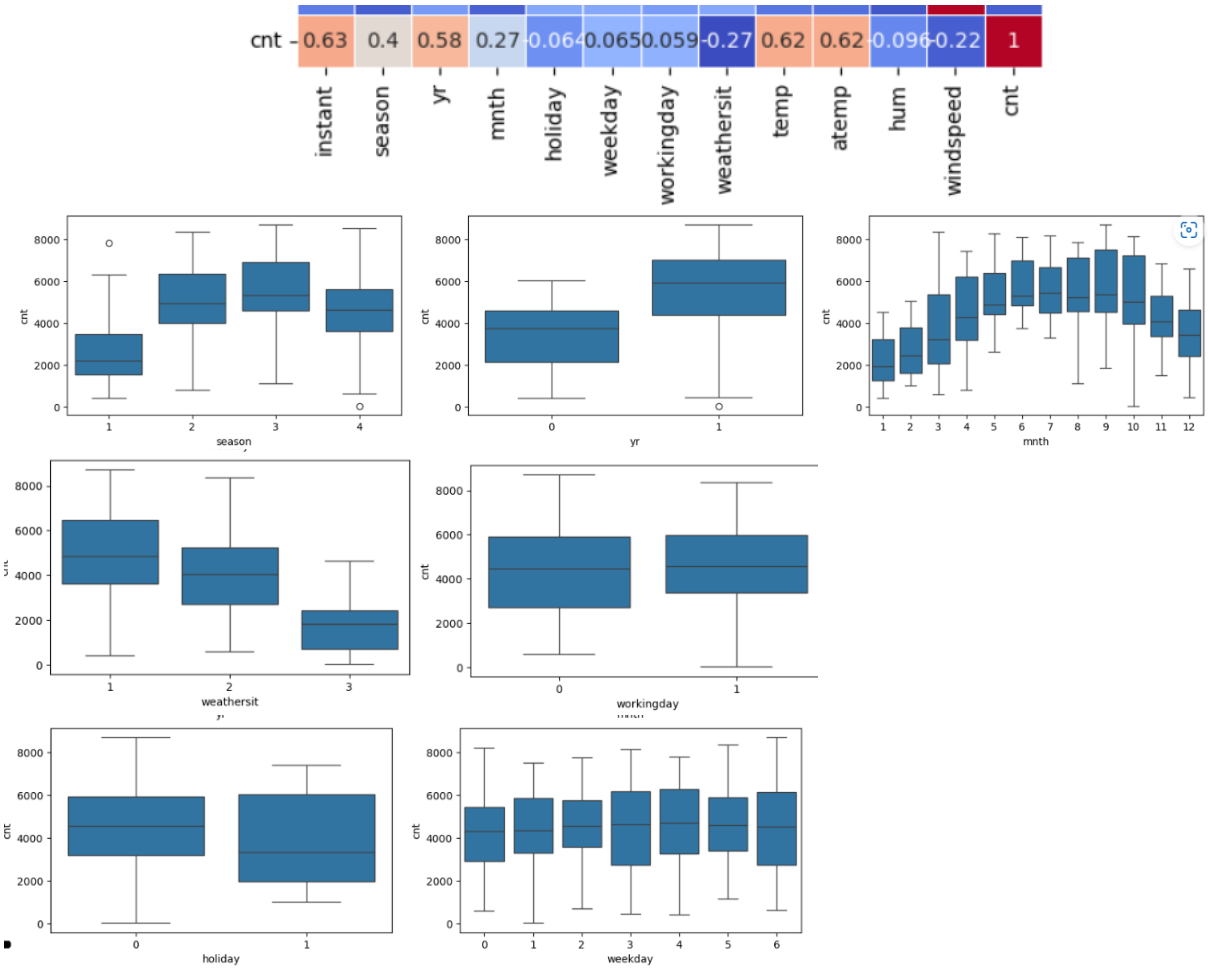
MLR Assignment – Ekansh Chaturvedi

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 mks)
    a. Considering the 'cnt' – the count of users that take a rides have **correlation with year** and season (temperature being numerical).



As an outcome of initial Bi-Variate Analysis:

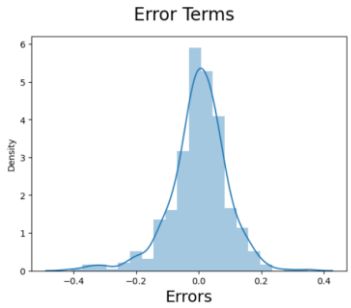
- Data represents usage of Bikes during SPRING Season is ~40-50% lower than other season.
- Also , referring weather situation, we find as logically expected that most riders favour a clear weather or moderate Cloudy while rides gets reduced during Thunderstorm, heavy Snow weather.
- Rides in general shows positive inclination from 2018 to 2019 by 50% increase in rides.
- While there No mentionable variance due to holiday, weekday or Working day impacting number of rides.

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

    a. Its important to avoid dummy variable trap, by removing redundant information avoiding it strengthens model stability and performance , also ensures model coefficients are interpretable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mks)
    a. Temperature. (both actual and Feel)

4. How did you validate the assumptions of Linear Regression after building the model on the training set.
    a. Yes , Plot the residuals against the predicted values to assess homoscedasticity.



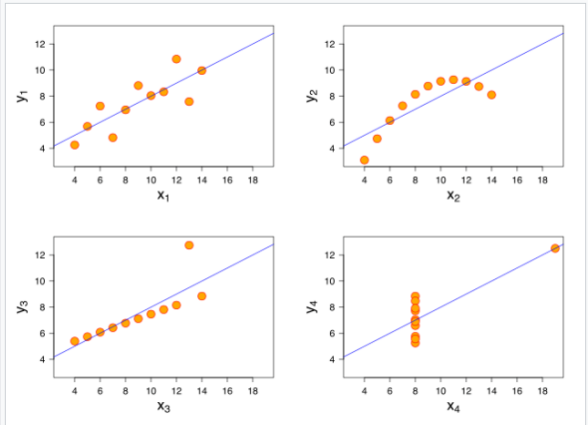
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. Year (yr):
        i. The coefficient for the yr variable is 0.2319.
        ii. This positive coefficient suggests that the demand for bike rentals increases with each passing year.
    b. Temperature (temp):
        i. The coefficient for the temp variable is 0.4934.
        ii. A positive coefficient indicates that higher temperatures lead to increased bike rentals.
    c. Season (summer):
        i. The coefficient for the summer variable is 0.1176.
        ii. This positive coefficient suggests that summer months contribute to higher bike rental demand.

General Subjective Question

1. Explain the linear regression algorithm in detail?
    a. Linear regression is a statistical method used to model the relationship between a dependent variable (response) and one or more independent variables (predictors).
    b. It assumes a linear relationship between the variables and estimates coefficients to create a linear equation (e.g., (y = mx + b)).
    c. A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized (i.e. the sum of squared differences between actual and predicted values) by estimating coefficients using OLS.
    d. Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable
    e. Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.
    f. Assumptions:
        i. **Linearity:** Assumes a linear relationship between predictors and response.
        ii. **Independence:** Residuals should be independent.
        iii. **Homoscedasticity:** Residuals should have constant variance.
        iv. **Normality:** Residuals follow a normal distribution.

2. Explain the Anscombe's quartet in detail.
    a. Anscombe's quartet consists of four datasets with nearly identical summary statistics (mean, variance, correlation) but different scatter plots.
    b. It highlights the importance of visualizing data and not relying solely on summary statistics.
    c. Despite their similar statistics, the datasets look distinct when plotted on scatter plots.
    d. These visual differences highlight the importance of data visualization over relying solely on summary statistics.

e. Regression algorithms can be misled by these variations, emphasizing the need to explore data visually before modeling or analysis.



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

1. Data Set 1: Fits a linear regression model well.
2. Data Set 2: Non-linear data, unsuitable for linear regression. While a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
3. Data Set 3: Contains outliers, challenging for linear regression.
4. Data Set 4: Also has outliers, making linear regression inadequate.

3. What is Pearson's R?
   a. Pearson's correlation coefficient measures the linear relationship between two continuous variables.
   b. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).
   c. **-1:** Perfect negative correlation (as one variable increases, the other decreases linearly).
      0: No linear correlation (variables are independent).
      1: Perfect positive correlation (both variables increase linearly).

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   a. Scaling ensures that all features contribute equally to the model.
   b. It improves convergence for gradient-based algorithms (like gradient descent).
   c.
      Normalized Scaling (Min-Max Scaling):

      i. Scales features to a range of [0, 1].
      ii. Formula:
         1. xnormalized=max(x)−min(x)x−min(x)

      Useful when features have similar distributions and no outliers.
   d. Standardized Scaling (Z-Score Scaling):

      i. Transforms features to have mean 0 and standard deviation 1.
      ii. Formula:
         1. Xstandardized=x−μ/σ
      iii. Robust to outliers and works well with any distribution.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   a. Infinite VIF occurs when perfect multicollinearity exists (one predictor is a linear combination of others).
   b. This makes it impossible to estimate the effect of individual predictors.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   a. A Q-Q plot compares the quantiles of a sample distribution to those of a theoretical distribution (usually normal).

b. It helps assess whether the residuals follow a normal distribution.
c. Deviations from the diagonal line indicate non-normality.