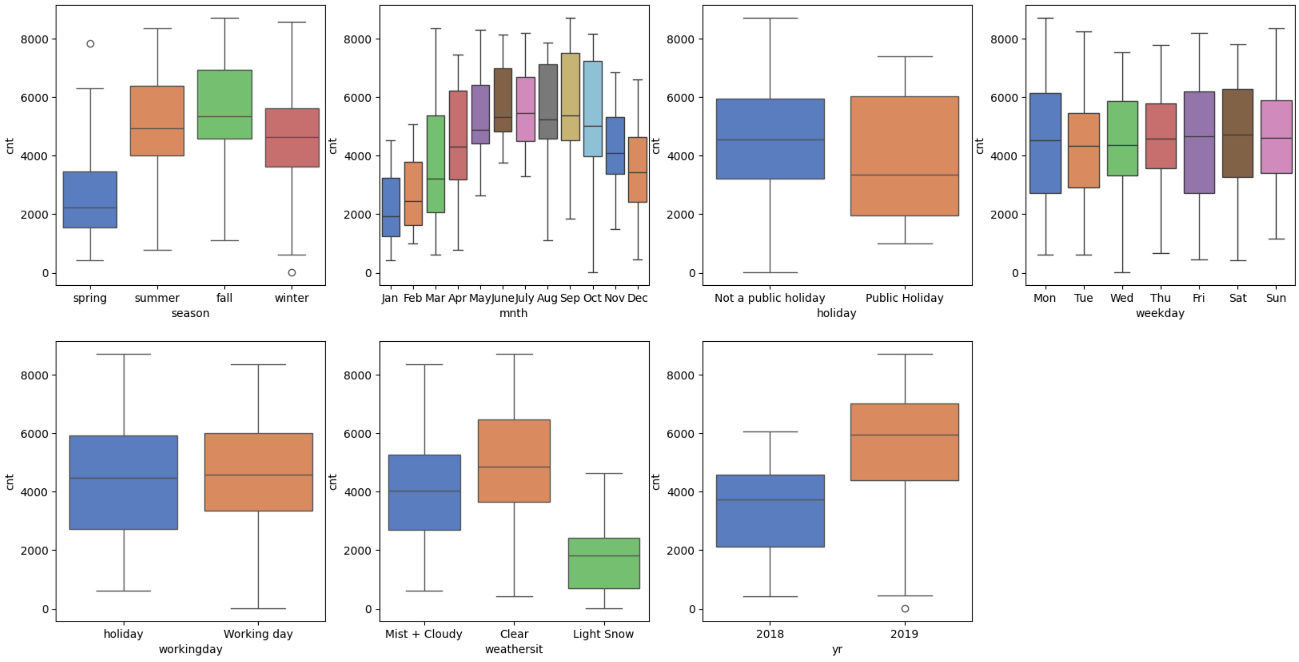
Linear Regression Assignment Subjective Questions

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?**



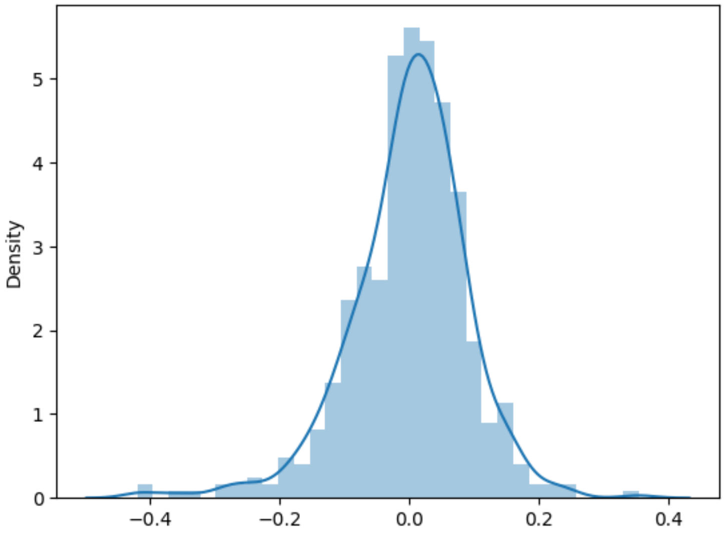
- People are most likely to rent bikes in fall, with summer next highest
- The number of people renting bikes in the months of June, July, August and September is high
- The number of bikes rented is higher on non-public holidays.
- The day of the week does not make a difference in the number of bikes being rented
- Working or non-working day does not impact the number of bikes being rented
- People are least likely to rent bikes when the weather is snowy. Clear weather has the highest number of bikes rented.
- There is an increase in the usage of rental bikes from the year 2018 to 2019.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   - The number of variables required to represent ‘m’ levels of a categorical variable is m-1.

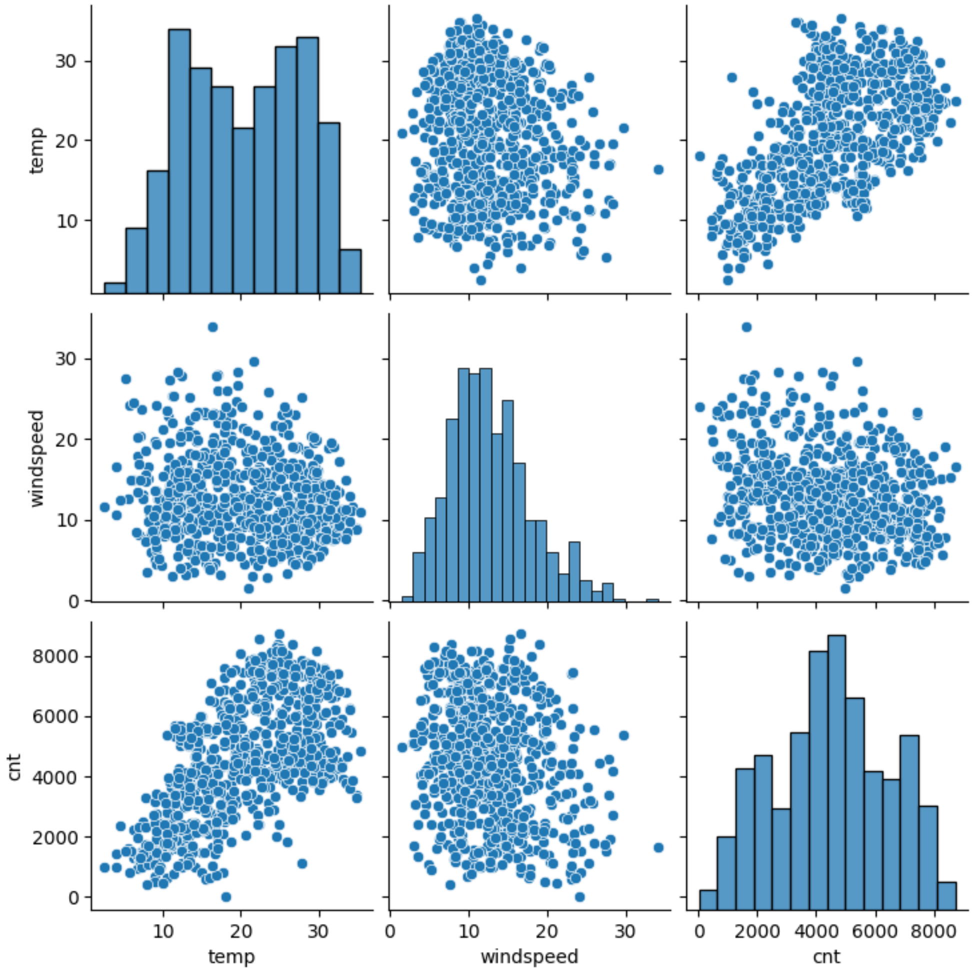
- During dummy variable creation the pandas.get_dummies() function uses ‘m’ variables to represent ‘m’ levels of categorical variables instead of using ‘m-1’ variables.
- Hence to avoid the extra column we use the drop_first=True parameter, this drops the first column of the dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - Looking at the pair plot amongst the numerical variables, ‘temp’ and ‘atemp’ numerical variables have the highest correlation with the target variable ‘cnt’.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - Assumptions of Linear regression models are validated as follows:
        - Residuals are normally distributed

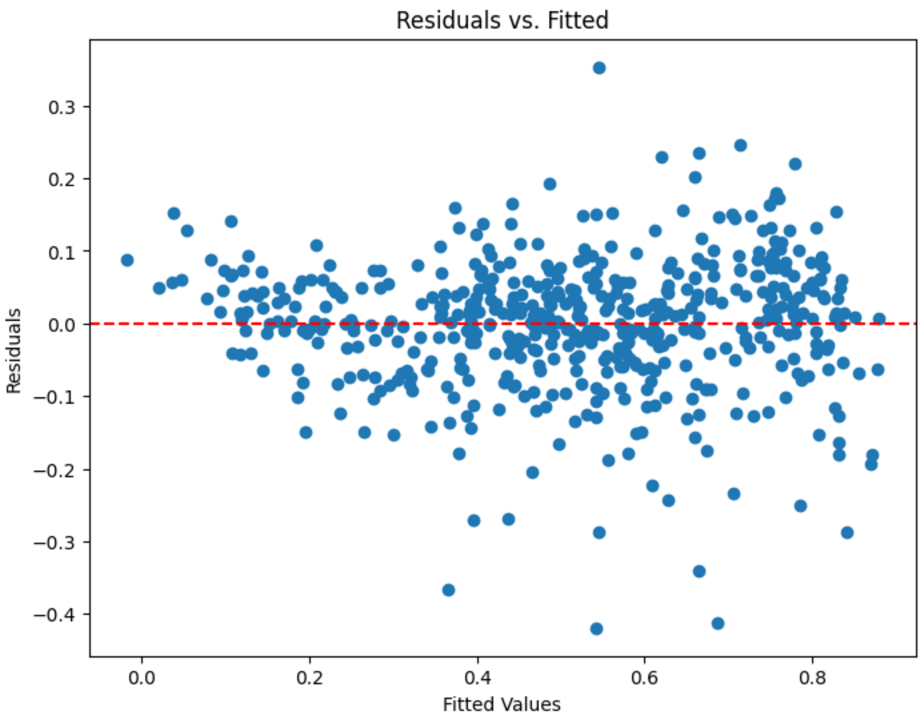


        - Multicollinearity check
            - VIF should be low, less than 5.
            - In the final model, the VIF of all but one feature is less than 5, the variable temp has VIF of 5.14, this variable was not removed from the model as it has a strong relation to the target variable, and its VIF is very close to 5.
        - Linearity is visible between the target variable and predictor variable



There is a linear relation between temp and cnt

- Error terms are independent of each other, i.e. one error term is not dependent on the previous error term
- Homeodasticity: Variance of error terms remains constant



  - There is no discernible pattern in the scatter plot hence we can conclude that the error terms are independent of each other
  - The variation of data points is not increasing significantly, hence the error terms have constant variance

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - The following top three features contribute significantly towards explaining the demand for the bikes:
     - Temperature (temp) has a coefficient of 0.4917
     - Year (yr) has a coefficient of 0.233876
     - weathersit (light snow) has a coefficient of -0.284654

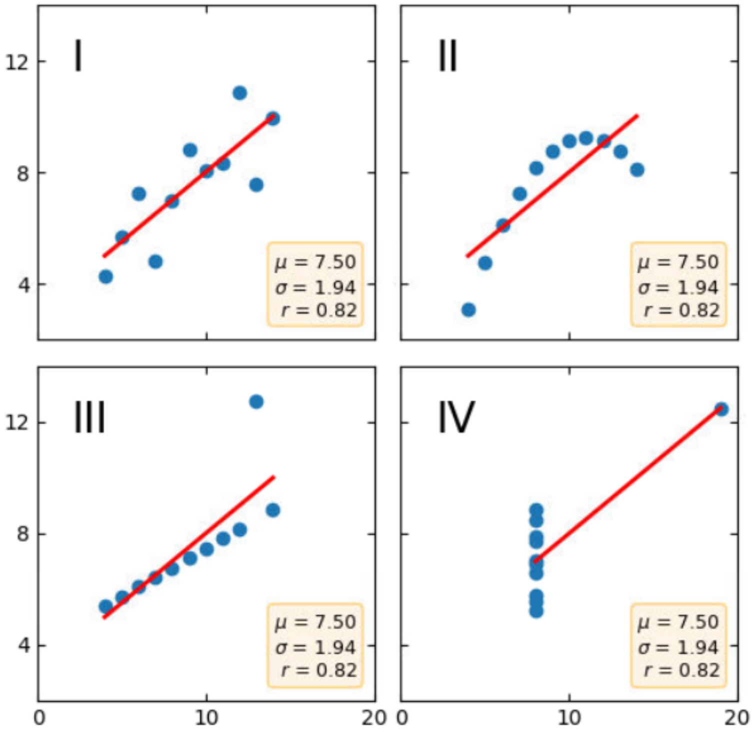
General Subjective Questions

1. **Explain the linear regression algorithm in detail**
   - **Linear regression** is a fundamental statistical method used to model the relationship between a dependent variable (also called the response variable) and one or more independent variables (predictors). The goal is to find the best-fitting linear equation that describes this relationship.
   - Simple Linear Regression:
     - Only one predictor variable

    - Equation: Y = mX + c
        - Y = target variable
        - X = predictor variable
        - m = Slope or (change in Y)/(change in X)
        - c = Value of Y when X = 0
- Multiple Linear Regression:
    - More than one predictor variable
    - Equation: Y = B0 + B1X1 + B2X2 + ... + BnXn + E
- The linear relationship can either be positive or negative.
- Positive linear relation means the dependent variable increases if the value of predictor variable increases. Slope is positive.
- Negative linear relation mans the dependent variable decreases if the predictor variable increases. Slope is negative.
- Assumptions of Linear regression
    - Linearity: The relationship between the independent and dependent variables is linear.
    - Independence: Observations are independent of each other.
    - Homoscedasticity: The residuals (errors) have constant variance at all levels of the independent variable(s).
    - Normality: The residuals should be approximately normally distributed (especially important for inference).
    - Assumptions for multiple linear regression:
        - No collinearity amongst the predictor variables.
- The objective of linear regression is to minimize the sum of squared differences between the observed values and predicted values.

2. **Explain the Anscombe's quartet in detail**
    - **Anscombe's Quartet** is a famous dataset created by statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analysing it and to show how different datasets can have the same statistical properties.
    - The quartet consists of four different datasets that have nearly identical summary statistics (mean, variance, correlation) but have very different distributions and appearances when plotted.
    - The Datasets
        - **Dataset I**:
            - A linear relationship with a positive slope.
        - **Dataset II**:
            - A non-linear relationship (a curve) with a similar correlation to Dataset I.
        - **Dataset III**:

        - A linear relationship with an outlier that significantly affects the correlation.
    - **Dataset IV**:
        - A vertical line with no correlation but with similar means and variances.



3. **What is Pearson's R**

- **Pearson's R**, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:
    - **1** indicates a perfect positive linear correlation.
    - **-1** indicates a perfect negative linear correlation.
    - **0** indicates no linear correlation.
- Pearsons R is used in linear regression to analyse the relationship between two variables

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

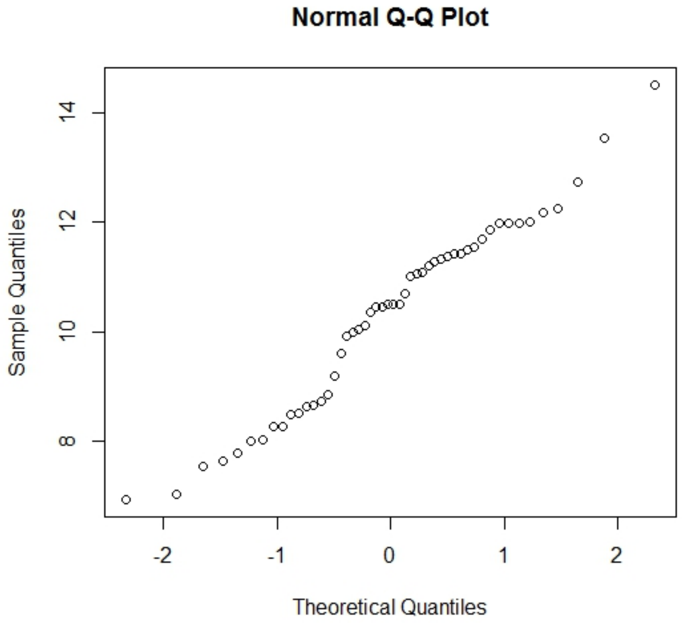
- Scaling is a technique for data preprocessing.
- Scaling is performed because:
    - Ease of interpretation. If the difference in scale of variables is very different then the coefficients in linear regression model becomes

    - harder to interpret, for instance a change in a variable measured in hundreds may have a different impact than a change in a variable measured in small units.
    - Faster convergence for gradient descent algorithms. If we have a lot of different features using different scales, say for example one variable has a scale of -1 to 1 and another variable has a scale of 1-10000, the time taken to get to convergence is very high. If the scales are comparable the convergence can be achieved faster.
- Two types of scaling:
    - Normalization
        - Normalized scaling is used when we want to bound the values in a specific range
        - It is used when the scales of the features are different
        - MinMax Scaling is a type of normalizarion
        - The values in MinMax scaling are in the range [0,1]
        - Normalized scaling is affected by outliers (can skew results)
    - Standardization
        - Standardization does not specify a fixed range of values, i.e. it is not bounded
        - It is used when the algorithms assume a normal distribution.
        - It has a mean of 0 and standard deviation of 1
        - Z-score scaling is a type of standardization
        - It is less sensitive to outliers (but still affected)

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    - VIF explains how one predictor variable is explained by all other predictor variables combined.
    - VIF becomes infinite when R-squared(j) = 1. This happens when one predictor variable is a perfect linear combination of one or more other predictor variables.
    - Example, X1 and X2 are predictor variables, if X2 = 10*X1 then VIF for X2 will be infinite.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    - A **Q-Q plot** (Quantile-Quantile plot) is a graph used to assess if a dataset follows a specified theoretical distribution, most commonly the normal distribution.

- 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. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.
- QQ Plot can be used in linear regression models to check that assumption that the residuals are normally distributed.
- Interpretation of the QQ Plot can be done as follows:
    - **Close to the Line**: If the points closely follow the diagonal line, it indicates that the residuals are approximately normally distributed.
    - **Deviations**: If the points deviate from the line, it suggests that the residuals may not be normally distributed, which could violate the assumptions of linear regression.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles