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?

- **Season**: Most of the bike bookings are done in Season 2 (Summer) and Season 3 (Fall).
- **Month**: Most of the bike bookings are done on Month 6 (June) ,7 (July),8 (August) and 9 (September)
- **Weekday**: Most of the bike bookings are done on all the days. No variation in the data
- **Weathersit:** Most of the bike bookings are done at weathersit1 (Clear, Few clouds, partly cloudy, partly cloudy) and Weathersit2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)

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

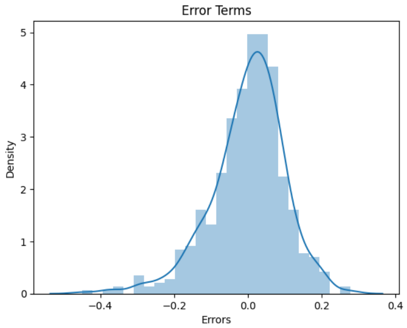
Using drop_first= True during dummy variable creation is important to address the multicollinearity issue. So Instead of K variables, K-1 dummy variables are created so that it breaks the correlation between the dummy variables and its easy to interpret the data.

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

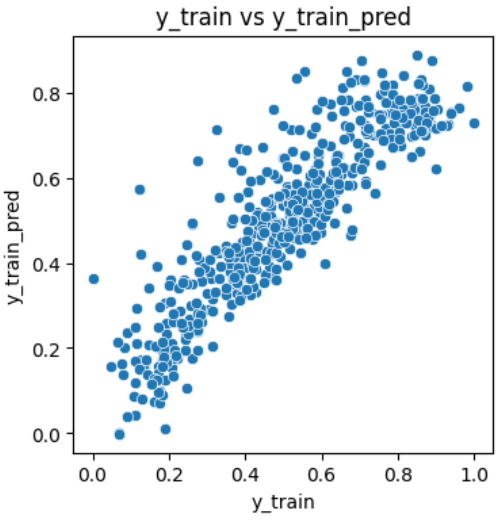
From the pair-plot among the numerical variables, we can understand 'temp' and 'atemp' is highly correlated with 'cnt' (Target Variable)

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

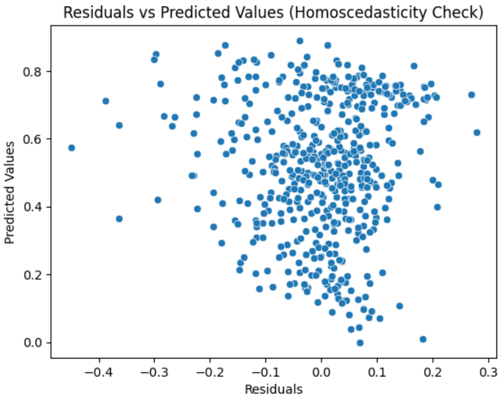
- **Normality of residuals:** In the Residual Analysis, I checked whether the errors or residuals follow normal distribution or not. Validated this phenomenon by plotting the errors in a distribution plot.



- **Linearity Check:** The relationship between target variable and independent variable should be Linear. Validated this phenomenon by plotting the target variable (y_train) and predictors (y_train_pred)



- **Homoscedasticity Check**: plot of residuals against predicted values can help identify heteroscedasticity, where the spread of residuals changes with the predicted values.



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

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of '0.571918' indicated that a unit increase in temp variable increases the bike hire numbers by 0.571918 units.
- **Weather Situation 3 (weathersit_3)** - A coefficient value of '-0.250231' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by -0.250231 units.
- **Year (yr)** - A coefficient value of '0.2308' indicated that a unit increase in yr variable increases the bike hire numbers by 0.233690 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   - Linear regression is a process of estimating relationships between the dependent variable and one or more independent variables.
   - Simple Linear regression involves only one independent variable, but Multiple Linear Regression involves one or more independent variables.
   - The simple linear regression model is represented by the equation: $y=\beta0+\beta1\cdot x+\varepsilon$
   - The Multiple linear regression model is represented by the equation: $y=\beta0+\beta1\cdot x1+\beta2\cdot x2+...+\beta n\cdot xn+\varepsilon$
     - Y is the dependent variable
     - β0 is the intercept.
     - β1, β2, …., βn are the coefficients.
     - ε is the error term.
   - The coefficients β0 and β1 are estimated using methods like the least squares method.
   - **Assumptions of Linear Regression:**
     - The relationship between the independent and dependent variables is linear.
     - Residuals are independent of each other.
     - Residuals have constant variance across all levels of the independent variables.
     - Residuals are normally distributed.
     - The independent variables are not perfectly correlated.

2. Explain the Anscombe’s quartet in detail.
    - Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they exhibit significant differences when graphically visualized. This provides the importance of visualizing the data and not relying on statistics alone.
    - The four datasets in Anscombe's quartet are:
        - Dataset I: Linear relationship with outliers
            - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
            - y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
        - Dataset II: Non-linear relationship with outliers
            - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
            - y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
        - Dataset III: Linear relationship with an influential point
            - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
            - y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
        - Dataset IV: No apparent relationship
            - x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
            - y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
    - Key observations:
        - Though they are having similar means, variances, and correlation coefficients, these datasets show unique patterns when visualized.
        - This shows the importance of visualizing the data and not just relying only on the statistics of the dataset.
        - Anscombe's quartet shows that data visualization more crucial for understanding data distribution and relationships.
3. What is Pearson’s R?
    - Pearson’s R is nothing but what we called it as Pearson’s correlation coefficient. So, this is a measure of the strength and direction of linear relationship between two variables. It ranges from -1 to 1.
    - If r=1, it indicates a positive linear relationship between the variables
    - If r=0, it indicates there is no linear relationship between the variables
    - If r=-1, it indicates there a negative linear relationship between the variables

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    - Scaling is the process of transforming numerical variables to a specific range or distribution.
    - Scaling is performed to ensure that all features in the data are on a similar scale.
    - Normalized scaling (Min-Max) transforms data to a specific range, typically 0 to 1 whereas Standardized scaling (Z-score) transforms data to have a mean of 0 and standard deviation of 1.
    - Normalized scaling (Min-Max scaling) preserving the original distribution's shape but sensitive to outliers whereas Standardized scaling (Z-score normalization) is less sensitive to outliers but altering the distribution's shape.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    - Variable Inflation Factor (VIF) measures how well a predictor variable can be predicted using all other predictor variables.
    - If a predictor variable is highly correlated then the correlation coefficient R will become 1, If R=1 in the VIF formula given below
        - VIF=1/1-R^2
    - Then VIF becomes 1/0 which is infinite.
    - So VIF becomes infinite when there is multicollinearity between independent variables.
    - To handle this, we need to remove one of the highly correlated variables.

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) visually compares observed residuals to the expected quantiles of a normal distribution.
    - In linear regression, it helps assess the normality assumption of residuals, detecting departures and ensuring the reliability of statistical inferences, model validity, and outlier identification.