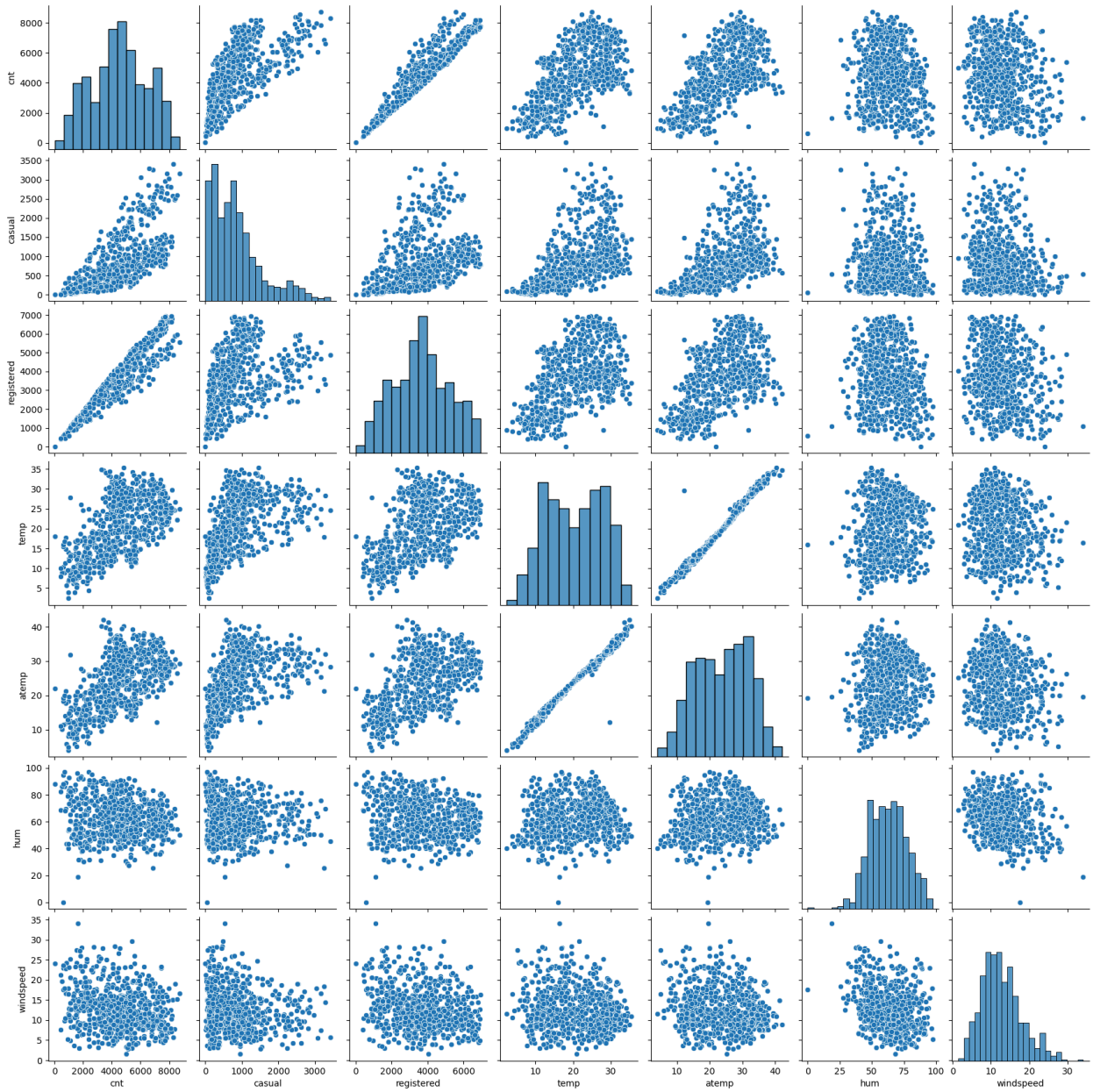
Assignment-based Subjective Questions

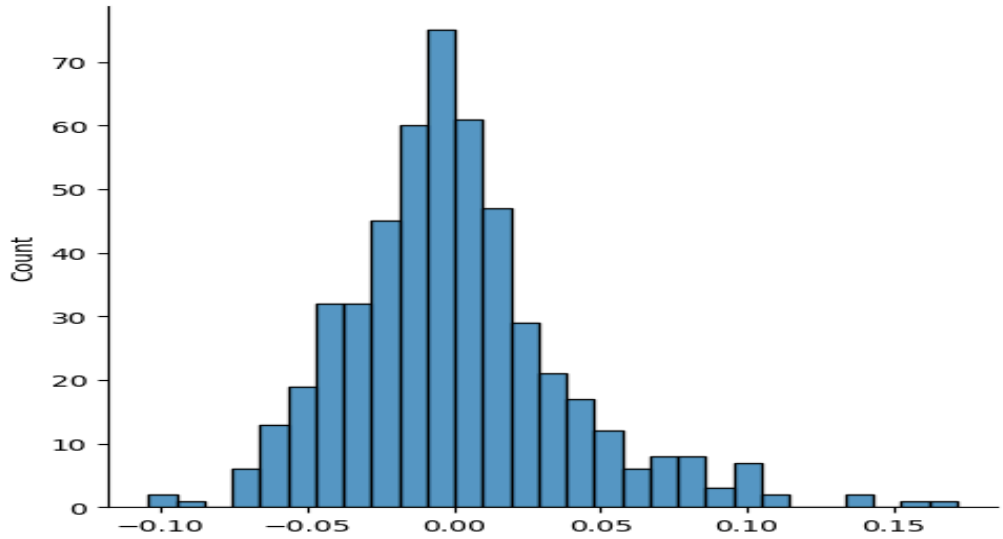
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
    - Categorical Variables
        i. season : 1: 'spring', 2: 'summer', 3: 'fall', 4:'winter'
        ii. yr : 0: '2018', 1: '2019'
        iii. mnth : 1: 'Jan', 2: 'Feb', 3: 'Mar', 4: 'April', 5: 'May', 6: 'June', 7: 'July', 8: 'Aug', 9: 'Sept', 10: 'Oct', 11: 'Nov', 12:'Dec'
        iv. weekday : 0: 'Mon', 1: 'Tue', 2: 'Wed', 3: 'Thurs', 4: 'Fri', 5: 'Sat', 6: 'Sun'
        v. weathersit : 1: 'Partly Cloudy', 2: 'Mist', 3: 'Low Rain', 4: 'Heavy Rain'
    - Each & every variable as some or other infer on our dependent variable.
    - Like when summer season is there our 'cnt' variable will increase as the coef we have is +ve 0.0238.
    - Likewise if we have season winter our 'cnt' variable decreases.
    - Same we have for other variables also that because of our categorical variable our 'cnt' variable will either decrease or increase.

2. **Why is it important to use drop_first=True during dummy variable creation?**
    - Our dummy variable formula stands for **'n-1'**
    - Also, it is important to avoid the multicollinearity issue or dummy variable trap
    - If we include all the levels of dummy variables in modeling we introduce a perfect multicollinearity among them
    - So when we **drop_first=True** it will drop the first variable which is created
    - Eg : if 21 variables are created then 20 variables will only be used.
    - The dummy variable trap occurs when you include all dummy variables representing the categories of a categorical variable in your model.
    - By dropping one dummy variable for each categorical feature, we maintain interpretability of the model coefficients.

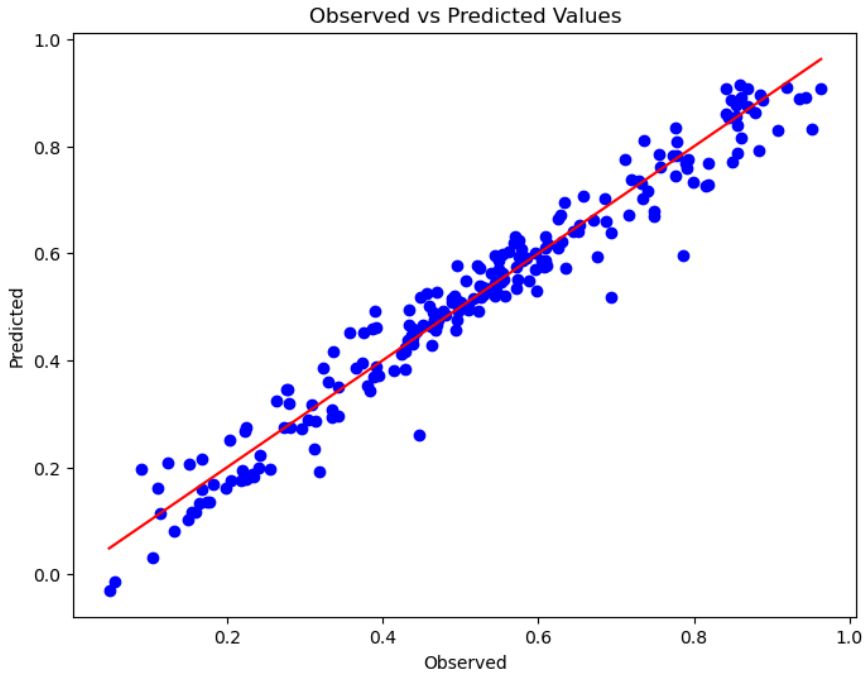
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    - Looking at our pair plot we can straight away say that our ‘registered’ & ‘casual’ variables are highly correlated with our dependent variable.



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    - As we can see that our R-squared on the training data set is 0.972 & our Adjusted R-squared is 0.970 & our r2_score on test data set is 0.952 from which we can say that our model is performing well on the unseen test data set.
    - We also have checked the VIF on our training data set & we observe that majority of the variables as VIF >5 from which we can say that there is no multicollinearity issue with our variables.
    - Also we have plotted a histogram for our residuals which has normal distribution.



    - As we can see that there is a linear relationship between our observed values & our predicted values. Our scatter plot shows that there is no violation of linearity assumption



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    - According to my assumption & modeling my top 3 features which are contributing significantly towards the demand of shared bikes are as follows
        i. workingday
            1. We can say that there is high demand for shared bikes during the working day. As people might require to travel & having shared bikes might seem more feasible than having other sources of travel.
        ii. registered
            1. We have registered users which are increasing the demand for shared bikes than the casual users.
        iii. weekday_Mon
            1. weekday_Mon also has a good impact on our dependent variable.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    a. Linear Regression is a machine learning algorithm based on supervised regression algorithm.
    b. Regression models a target prediction value based on independent variables.
    c. It is mostly used for finding out the relationship between variables and forecasting
    d. Equation of linear regression:

        $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_n x_n + \epsilon$$

    e. In Linear Regression, we predict the value by an integer number.
    f. It is based on the ordinary least square method.
    g. We calculate Root Mean Square Error(RMSE) to predict the next value.
    h. Dependent variable should be numeric and the response variable is continuous to value
    i. when we plot the training datasets, a straight line can be drawn that touches maximum plots
    j. To evaluate our model's performance quantitatively, we plot the error of each observation directly. These errors, or residuals, measure the distance between each observation and the predicted value for that observation.
    k. We make use of these residuals later when we talk about evaluating regression models
    l. The goal of linear regression is reducing this error such that we find a line/surface that 'best' fits our data

2. **Explain the Anscombe's quartet in detail.**
    a. Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics.
    b. It consists of four datasets that have nearly identical statistical properties, including means, variances, correlation coefficients, and regression lines, yet they exhibit vastly different patterns when plotted.
    c. The quartet was created by the statistician Francis Anscombe in 1973
    d. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths
    e. **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.
    f. Dataset 1:
        i. This dataset consists of a linear relationship between the variables *X* and *Y*, with some random noise added to the data points.
        ii. When plotted, it resembles a simple linear relationship, and fitting a linear regression line
    g. Dataset 2:
        i. This dataset also has a linear relationship between the variables *X* and *Y* but with an outlier that significantly affects the linear regression line
        ii. Despite the outlier, the summary statistics (mean, variance, correlation coefficient) are similar to those of Dataset I.
        iii. However, the outlier distorts the linear relationship when plotted

h. Dataset 3:
    i. This dataset also has a linear relationship between the variables *X* and *Y, specifically* a quadratic relationship.
    ii. Summary statistics still indicate a strong linear relationship, but plotting the data reveals the non-linear pattern.
    iii. A linear regression line is not appropriate for this dataset.
i. Dataset 4:
    i. This dataset consists of several groups of data points, each with the same *X* values but different *Y* values.
    ii. Summary statistics suggest a linear relationship, but plotting the data reveals distinct groups of points with different patterns.
    iii. A single linear regression line is not appropriate for this dataset.

j. Anscombe's quartet emphasizes the importance of visually inspecting data before drawing conclusions. Summary statistics alone may not capture the true nature of the data
k. A regression line may not always be the best summary of the relationship between variables, especially if the relationship is non-linear or affected by outliers

3. What is Pearson's R?

a. Pearson's *r*, commonly referred to as Pearson correlation coefficient or Pearson's *r*, is a measure of the linear correlation between two variables.
b. It quantifies the strength and direction of the linear relationship between two continuous variables.
c. It is named after Karl Pearson, who developed it in the 19th century.
d. Pearson's correlation coefficient ranges from -1 to 1
    i. $r=1$: Perfect positive correlation. As one variable increases, the other variable also increases linearly
    ii. $r=-1$: Perfect negative correlation. As one variable increases, the other variable decreases linearly
    iii. $r=0$: No linear correlation. There is no linear relationship between the two variables.
e. Pearson's *r* measures the linear relationship between variables. It assumes that the relationship between the variables can be adequately described by a straight line
f. Pearson's *r* is not affected by changes in scale (multiplication by a constant) or translation (addition of a constant) of the variables.
g. Pearson's *r* can be sensitive to outliers, particularly in small datasets.
h. The formula to compute Pearson's correlation coefficient (r) between two variables X & Y is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    a. Scaling is part of data preparation as this technique brings data points that are far from each other closer in order to increase the algorithm effectiveness and speed up the Machine Learning processing. Scaling data enables the model to learn and actually understand the problem.
    b. Why is scaling performed?
        i. Scaling is performed to get the data points that are far from each other closer to increase the algorithm effectiveness. It use to improve the model performance.
    c. Normalization:
        i. Normalization in machine learning is the process of translating data into the range [0, 1] (or any other range)

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

        ii. Xmax and Xmin are the maximum and the minimum values of the feature, respectively.
        iii. When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
        iv. When the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X' is 1
        v. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

    d. Standardization
        i. Standardization is another Feature scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.
        ii. Formula

$$X' = \frac{X - \mu}{\sigma}$$

        iii. Mu is the mean of the feature values and Sigma is the standard deviation of the feature values. In this case, the values are not restricted to a particular range.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    a. The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. It quantifies how much the variance of the estimated regression coefficients are inflated due to multicollinearity among the predictor variables. Typically, VIF values greater than 10 or 5 are considered indicative of multicollinearity.
    b. When the VIF value is calculated to be infinite, it indicates an extremely high degree of multicollinearity among the predictor variables. This situation arises when one or more predictor variables are perfectly collinear with one or more other predictor variables.
    c. **Perfect Collinearity**:
        i. One or more predictor variables can be expressed as a perfect linear combination of other predictor variables
    d. **Data Errors**:
        i. Data errors or coding mistakes can lead to perfect collinearity among predictor variables.
    e. **Dummy Variable Trap**:
        i. In regression analysis involving categorical variables encoded as dummy variables, if all levels of a categorical variable are included as dummy variables, it can lead to perfect collinearity among the dummy variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**
    a. A Q-Q (Quantile-Quantile) plot is a graphical technique used to assess whether a set of data follows a particular probability distribution, such as the normal distribution
    b. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically the standard normal distribution (mean = 0, standard deviation = 1)
    c. The Q-Q plot is a scatter plot.
    d. Observed data quantiles plotted on the horizontal axis and the theoretical quantiles (from the specified distribution) plotted on the vertical axis
    e. In linear regression, one of the key assumptions is that the residuals are normally distributed. Q-Q plots provide a visual method to check this assumption.
    f. If the residuals follow a normal distribution, the points on the Q-Q plot will approximately form a straight line.
    g. Q-Q plots allow you to identify deviations from normality.
    h. If the Q-Q plot reveals deviations from normality, it may suggest potential model improvements or transformations of variables to better meet the assumption of normality
    i. A well-fitting linear regression model should have residuals that are approximately normally distributed. Q-Q plots help evaluate the adequacy of the model fit by assessing whether the residuals conform to the expected distribution.
    j. Q-Q plots can also help identify outliers and extreme values in the residuals. Outliers may cause deviations from the expected straight line pattern in the Q-Q plot, indicating potential issues with the model or the data.