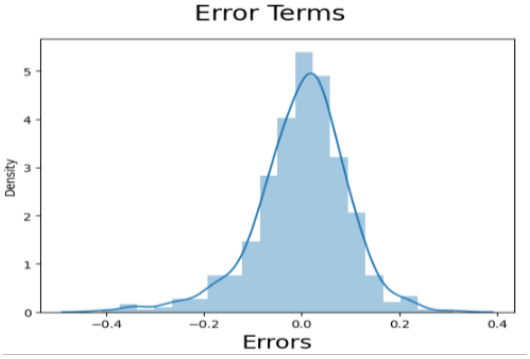
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    a. The rentals are higher in summer and fall season (season - 2, 3)
    b. The same trend is seen in months. There are more people likely to rent the bike when the season changes to summer and continues to fall.
    c. There's an obvious high trend of bike rentals in clear weather which can be indicated in wathersit value 1
    d. People tend to rent bikes more on a normal day than on a holiday.
    e. The above point can also be proven by looking and the trend in workingday analysis.

2. Why is it important to use **drop_first=True** during dummy variable creation?
    a. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
    b. Let's say we have 3 types of values in the Categorical column and we want to create a dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obviously unfurnished. So we do not need a 3rd variable to identify the unfurnished.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    a. atemp and temp variables are highly correlated with cnt target variable.
    b. Temperature affects the number of people willing to rent and ride a bike.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Error terms are normally distributed with mean zero. So I constructed the below histogram for residuals. It shows me that the residuals are normally distributed which is a property of Linear Regression hence, proving that the assumption to use linear regression is correct.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    a. temp - This predicts temperature and the number of bike rental counts is directly proportional to each. Increase in temperature will also increase the number of bike rentals.
    b. weathersit_3 - This predicts that when the weather situation is moving from light snow/light rain / thunderstorm to clear / partly cloudy, the number of bike rentals increase with it.
    c. yr - This predicts that the more years the company is in business the people tend to know about it and the bike rentals increase with each passing year.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
    b. When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.
    c. Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.
    d. The interpretability of linear regression is a notable strength.
    e. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics.
    f. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.
    g. Simple Linear Regression: This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.
    h. Multiple Linear Regression: This involves more than one independent variable and one dependent variable.
    i. Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

2. Explain the Anscombe's quartet in detail
    a. Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.
    b. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models.
    c. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).
    d. Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?
    a. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.
    b. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
    c. The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset.
    d. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Feature scaling is a vital pre-processing step in machine learning that involves transforming numerical features to a common scale.
    b. It plays a major role in ensuring accurate and efficient model training and performance.
    c. Scaling techniques aim to normalize the range, distribution, and magnitude of features, reducing potential biases and inconsistencies that may arise from variations in their values.
    d. Purpose of scaling:
        i. Enhancing Model Performance: Feature scaling can significantly enhance the performance of machine learning models. Scaling the features makes it easier for algorithms to find the optimal solution, as the different scales of the features do not influence them.
        ii. Addressing Skewed Data and Outliers: Skewed data and outliers can negatively impact the performance of machine learning models. Scaling the features can help in handling such cases. By transforming the data to a standardized range, it reduces the impact of extreme values and makes the model more robust.
        iii. Faster Convergence During Training: For gradient descent-based algorithms, feature scaling can speed up the convergence by helping the optimization algorithm reach the minima faster.

iv. Balanced Feature Influence: When features are on different scales, there is a risk that larger-scale features will dominate the model's decisions, while smaller-scale features are neglected. Feature scaling ensures that each feature has the opportunity to influence the model without being overshadowed by other features simply because of their scale.
v. Improved Algorithm Behavior: Certain machine learning algorithms, particularly those that use distance metrics like Euclidean or Manhattan distance, assume that all features are centered around zero and have variance in the same order.

e. Difference between Normalized and Standard scaling:
   i. Normalization: also known as min-max scaling, transforms the features to a range between 0 and 1. It subtracts the minimum value of the feature and divides it by the range (maximum value minus minimum value). This technique is suitable when the distribution of the data does not follow a Gaussian distribution.
   ii. Standardization transforms the features to have a mean of 0 and a standard deviation of 1. It subtracts the mean of the feature and divides it by the standard deviation. This technique is preferable when the data is normally distributed or when we don't know the distribution in advance. Standardization maintains the shape of the distribution and does not bound the features to a specific range.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   a. The VIF is equal to 1 if the regressor is uncorrelated with the other regressors, and greater than 1 in case of non-zero correlation.
   b. The greater the VIF, the higher the degree of multicollinearity.
   c. In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.
   d. There is no precise rule for deciding when a VIF is too high (O'Brien 2007), but values above 10 are often considered a strong hint that trying to reduce the multicollinearity of the regression might be worthwhile.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   a. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
   b. Also, it helps to determine if two data sets come from populations with a common distribution.
   c. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions
   d. It can be used with sample sizes also

e. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot