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?** (3 marks)

Ans- As per the analysis of the categorical variables from the dataset, below are the inferences:

- Rented bikes count is more in fall and summer season
- Bikes rented in 2019 are more compared to 2018
- Bikes are rented more on Sat, Wed and Thur.
- Bikes are rented more in clear weather
- Bikes are rented more in September.

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

Ans- When creating dummy variables from categorical variables, the 'drop_first = True' parameter is used to drop one of the dummy variables representing a particular category. This is important to avoid multicollinearity, which is a situation where two or more predictor variables in a regression model are highly correlated with each other. This approach ensures that there is no perfect correlation among the dummy variable, as only one dummy variable represents each category and thus improving the reliability of the model's estimates and inference.

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

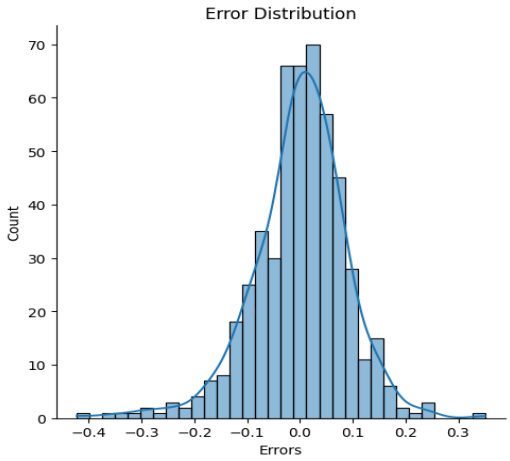
Ans- The temp variable has the highest correlation with target variable (cnt).

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

Ans- Validation on the assumptions of Linear Regression was done on basis of below:

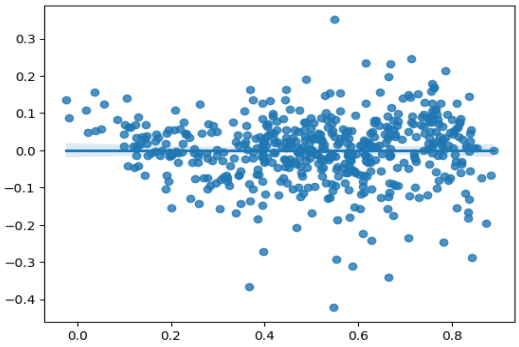
Residual analysis:

- Error distribution of residual. It is a normal distribution, and it is centered around 0.



- -Variance of Errors doesn't follow any trends

- Residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend.



- Linear relationship between the dependent variable and feature variable.

R-squared and Adjusted R-squared

VIF.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)

Ans- The top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature(temp), year(yr) and LightSnow .

General Subjective Questions

1. **Explain the linear regression algorithm in detail.** (4 marks)

Ans- Linear regression is an ML algorithm used for supervised learning. It helps in predicting the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, where the dependent variable is assumed to be a linear combination of the independent variables.

There are two types of Linear Regression- Simple Linear Regression and Multiple Linear Regression. Simple Linear Regression is used when a single independent variable is used to predict the target variable. Multiple Linear Regression is used when a multiple independent variable are used to predict the target variable.

Here are the steps involved in the linear regression algorithm:

1. Data Preparation: Gather a dataset that contains the values of the dependent variable and the independent variables. Ensure the dataset is free from missing values, outliers, and other data quality issues.
2. Variable Selection: Determine which independent variables to include in the model. This can be based on domain knowledge, exploratory data analysis, or statistical techniques like forward selection, backward elimination, or regularization methods.

3. Model Specification: Specify the form of the linear regression model by defining the relationship between the dependent variable and the selected independent variables. It takes the form:
   $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$,
   where y is the dependent variable, β_o is the intercept, β_1, β_2, ..., β_p are the coefficients for the independent variables x_1, x_2, ..., x_p, and ε represents the error term.
4. Estimation: Estimate the coefficients (β_o, β_1, β_2, ..., β_p) that minimize the sum of squared residuals (differences between observed and predicted values). This can be done using various techniques like ordinary least squares (OLS), gradient descent, or maximum likelihood estimation.
5. Model Evaluation: Assess the goodness of fit of the model. Common evaluation metrics include the coefficient of determination (R^2), which measures the proportion of variance in the dependent variable explained by the independent variables, and the standard error of the estimate (SE), which measures the average distance between the observed and predicted values.
6. Model Interpretation: Interpret the estimated coefficients to understand the relationship between the independent variables and the dependent variable. Positive coefficients indicate a positive relationship, negative coefficients indicate a negative relationship, and the magnitude of the coefficients represents the strength of the relationship.
7. Prediction: Once the model is trained and evaluated, it can be used to make predictions on new, unseen data. By plugging in the values of the independent variables into the model equation, we can obtain the predicted value of the dependent variable.
8. Model Assumptions: Linear regression relies on several assumptions, including linearity, independence of errors, constant variance of errors (homoscedasticity), normality of errors, and absence of multicollinearity. It is important to assess these assumptions and take appropriate steps if they are violated.

Overall, linear regression provides a simple yet powerful approach to model and understand the relationships between variables.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans- Anscombe's quartet refers to a set of four datasets that have nearly identical simple descriptive statistics but exhibit starkly different properties when plotted and analyzed. These datasets were originally created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and to challenge the reliance on summary statistics alone.

The four datasets in Anscombe's quartet are labeled I, II, III, and IV. Here is a detailed description of each dataset:

1. Anscombe's Dataset I:
   x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82

Dataset I is a simple linear relationship with a slight positive slope. When plotted, the data points roughly follow a straight line. The relationship between x and y is best described by a linear regression model.

2. Anscombe's Dataset II:

   x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

   y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26

Dataset II also exhibits a linear relationship, but with a slight curvilinear pattern. The points roughly follow a curve rather than a straight line. This dataset highlights the importance of considering non-linear relationships.

3. Anscombe's Dataset III:

   x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

   y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42

Dataset III represents a non-linear relationship with a clear outlier. Most of the points follow a linear pattern, except for one data point with a significantly higher y-value. This dataset emphasizes the impact of outliers on statistical analysis and regression models.

4. Anscombe's Dataset IV:

   x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8

   y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91

Dataset IV consists of a vertical line of x-values except for one extreme outlier. The relationship between x and y is nonlinear and not well-described by a simple linear regression model. This dataset illustrates the importance of examining the overall structure of the data rather than relying solely on summary statistics.

Anscombe's quartet serves as a cautionary example, highlighting that datasets with similar summary statistics can exhibit diverse patterns and relationships when visualized. It emphasizes the importance of exploratory data analysis and the limitations of relying solely on summary statistics without examining the data's underlying structure.

3. What is Pearson's R? (3 marks)

Ans- Pearson's correlation coefficient, often denoted as Pearson's R or simply R, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by the British statistician Karl Pearson and is widely used in various fields, including statistics, social sciences, and data analysis.

Pearson's R ranges from -1 to 1, where:

- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.

The calculation of Pearson's R involves the following steps:

1. Standardizing the variables: Each variable is transformed by subtracting its mean and dividing by its standard deviation. This step ensures that the variables are on a comparable scale.

2. Computing the covariance: The product of the standardized values of the two variables is calculated for each data point, and the average of these products is computed. This measure is known as the covariance.
3. Computing the standard deviations: The standard deviations of the two variables are calculated separately.
4. Calculating Pearson's R: The covariance is divided by the product of the standard deviations of the two variables to obtain the correlation coefficient, which is Pearson's R.

Pearson's correlation coefficient is widely used because it is easy to interpret and provides a measure of the linear association between variables. However, it is important to note that Pearson's R only measures the strength and direction of linear relationships and may not capture other types of relationships, such as non-linear or curvilinear associations. Additionally, Pearson's R is sensitive to outliers, and its validity can be affected by violations of assumptions, such as non-normality or heteroscedasticity.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** (3 marks)

Ans- Scaling, in the context of data analysis, refers to the process of transforming variables to a specific range or distribution. It involves adjusting the values of variables to make them comparable and facilitate meaningful comparisons between different variables or observations. Scaling is typically performed as a preprocessing step before applying various statistical or machine learning techniques.

Scaling is performed for several reasons:

1. Comparable Magnitudes: Variables may have different scales or units of measurement. Scaling brings the variables to a similar magnitude, ensuring that no single variable dominates the analysis simply because of its larger values. This allows for fair and meaningful comparisons.
2. Convergence of Algorithms: Many optimization algorithms used in machine learning and statistical models perform better when the input features are on a similar scale. Scaling helps algorithms converge faster and prevents issues where some variables have disproportionate influences on the model.
3. Interpretability: Scaling makes the variables more interpretable. It ensures that the coefficients or weights assigned to the variables can be directly compared, as they are based on the same scale.

There are two common types of scaling: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):

   Range: Transforms the variables to a specific range, typically between 0 and 1.

   Formula: The formula for normalizing a variable x is: x_normalized = (x - min(x)) / (max(x) - min(x)).

   Benefits: Normalized scaling preserves the original distribution and relative ordering of the data. It is useful when the exact range of the data is important and known.

2. Standardized Scaling (Z-score Standardization):

   Z-score: Transforms the variables to have a mean of 0 and a standard deviation of 1.

   Formula: The formula for standardizing a variable x is: x_standardized = (x - mean(x)) / standard_deviation(x).

   Benefits: Standardized scaling centers the data around the mean, ensuring a mean of 0, and scales the data based on the standard deviation. It allows for easier interpretation as the values are expressed in terms of standard deviations from the mean. Standardized scaling is useful when comparing variables with different units or distributions.

It's important to note that the choice between normalized scaling and standardized scaling depends on the specific requirements of the analysis or modeling task at hand. Both scaling methods have their own advantages and considerations, and the decision should be made based on the nature of the data and the objectives of the analysis.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?** (3 marks)

Ans- In the context of linear regression analysis, the Variance Inflation Factor (VIF) is a measure that quantifies multicollinearity, which is the correlation or high interdependency among predictor variables. VIF is used to assess the extent to which the variance of the estimated regression coefficients is inflated due to multicollinearity.

The formula for calculating the VIF of a predictor variable is as follows: VIF = 1 / (1 - R^2) where R^2 represents the coefficient of determination obtained by regressing the predictor variable against all other predictor variables. The VIF value provides insight into how much the variance of a particular predictor's coefficient is inflated due to multicollinearity.

In some cases, the VIF value can be infinite. This occurs when the coefficient of determination (R^2) for a particular predictor variable is equal to 1. A perfect correlation between a predictor variable and other predictor variables can lead to an R^2 of 1, resulting in an infinite VIF.

There are a few scenarios that can cause a predictor variable to have an R^2 of 1 and an infinite VIF:

1. Perfect Linear Relationship: The predictor variable is perfectly linearly related to one or more other predictor variables in the model. In this case, the VIF becomes infinite because the variance of the coefficient estimate cannot be determined separately from the other correlated variables.
2. Redundant Predictor: The predictor variable is a linear combination or a duplicate of another predictor variable(s) in the model. When two or more predictor variables provide the same information, it results in perfect multicollinearity and leads to an infinite VIF.

Having an infinite VIF suggests severe multicollinearity, indicating that the predictor variable is perfectly predictable from the other variables in the model. This can pose challenges in interpreting the model and estimating the effect of individual predictors. In such cases, it is necessary to address multicollinearity by identifying and resolving the high interdependency

among the predictor variables. Techniques such as removing redundant variables, transforming variables, or using dimensionality reduction methods can help mitigate multicollinearity issues.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Ans- A Q-Q (quantile-quantile) plot, also known as a quantile plot or normal probability plot, is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution, typically the normal distribution. It helps to determine if the data follows a specific distribution or if it deviates from it.

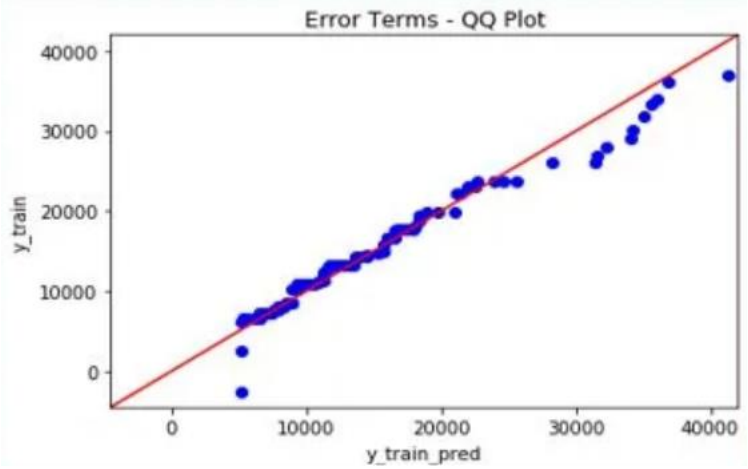
In a Q-Q plot, the observed quantiles of the data are plotted against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points in the plot should lie approximately along a straight line. Deviations from the straight line suggest departures from the assumed distribution.

In the context of linear regression, Q-Q plots are used to assess the assumption of normality for the residuals or errors. The residuals represent the differences between the observed values and the predicted values obtained from the linear regression model. The assumption of normality is important as it enables valid inference, hypothesis testing, and confidence interval estimation in linear regression.

The use and importance of Q-Q plots in linear regression are as follows:

1. Assessing Normality: By examining the Q-Q plot of the residuals, you can visually assess if the residuals follow a normal distribution. If the points on the plot closely follow the diagonal line, it suggests that the residuals are approximately normally distributed. On the other hand, deviations from the line indicate departures from normality.
2. Detecting Skewness and Outliers: Q-Q plots can reveal skewness in the distribution of residuals. If the points on the plot deviate from the straight line, it may indicate a skewed distribution. Additionally, extreme deviations from the line may indicate the presence of outliers in the residuals.
3. Model Validity: Normality of residuals is a crucial assumption for linear regression models. Violations of this assumption can lead to biased coefficient estimates, incorrect standard errors, and invalid hypothesis testing. Q-Q plots provide a graphical tool to assess the validity of the normality assumption and identify potential issues with the model.
4. Remedial Actions: If the Q-Q plot reveals deviations from normality, it indicates a need for further investigation and potential remedial actions. It may be necessary to consider transformations of variables, identify influential observations, or explore alternative modeling techniques that can handle non-normal residuals.

In summary, Q-Q plots provide a visual assessment of the distributional assumptions in linear regression, particularly the assumption of normality for residuals. They help to identify departures from normality, skewness, and outliers, which are important considerations in ensuring the validity and reliability of the regression model.


Error Terms - QQ Plot
40000
30000
20000
10000
0
y_train
0
10000
20000
30000
40000
y_train_pred