Subjective Questions Solution

Assignment based subjective questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans) In my analysis the dependent variable is ‘cnt’ and categorical variables are 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit'. The effect of categorical variables on ‘cnt’ is following:

- Among all seasons, in fall season 3, the demand of bikes is highest.
- Demand of bike in 2019 is higher than 2018.
- Over a year, in September month, the demand of bikes is highest.
- Holiday, weekday and working day do not impact the demand of bikes.
- In Clear, few clouds, Partly cloudy, Partly cloudy weather conditions the demand for bikes increases.

Q2) Why is it important to use **drop_first=True** during dummy variable creation?

Ans) When creating dummy variables from categorical features, especially in the context of machine learning models, using drop_first=True is important to avoid multicollinearity issues and to maintain model interpretability.

Here's why it's important:

- **Avoiding Multicollinearity:** When you create dummy variables for categorical features, you essentially create a binary indicator for each category. Let's say you have a feature with three categories: A, B, and C. If you create dummy variables without dropping the first category, you'll end up with two dummy variables: A, B, and C. However, you only need two variables to represent these three categories. By dropping one category (usually the first), you avoid perfect multicollinearity among the dummy variables. The dropped category becomes the reference category, and the model will implicitly understand that if both B and C are 0, then the observation must be in category A.

- **Interpretability:** When interpreting the coefficients of the model, having a reference category allows for easier interpretation. With drop_first=True, the coefficients of the remaining dummy variables represent the change in the dependent variable relative to the dropped category. This makes the interpretation more straightforward. For example, if you have dummy variables for gender (male, female), and you drop 'male', the coefficient for 'female' will represent the difference in the dependent variable between females and the reference category (males).

- **Reducing Dimensionality:** Dropping one category reduces the dimensionality of the dataset. This can be helpful, especially when dealing with high-dimensional data, as it reduces the computational burden and can prevent overfitting.

In summary, using drop_first=True during dummy variable creation helps in avoiding multicollinearity, enhances interpretability of the model coefficients, and reduces dimensionality, making it a common practice in data preprocessing for machine learning tasks.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

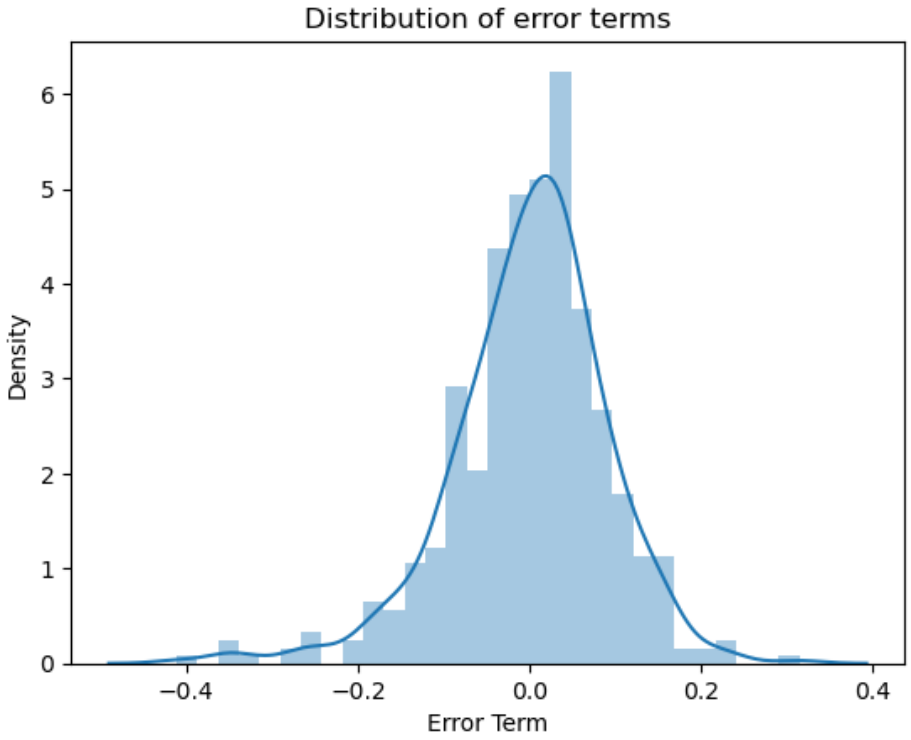
Ans) 'temp' has the highest correlation of 0.63 with the target variable 'cnt'.

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

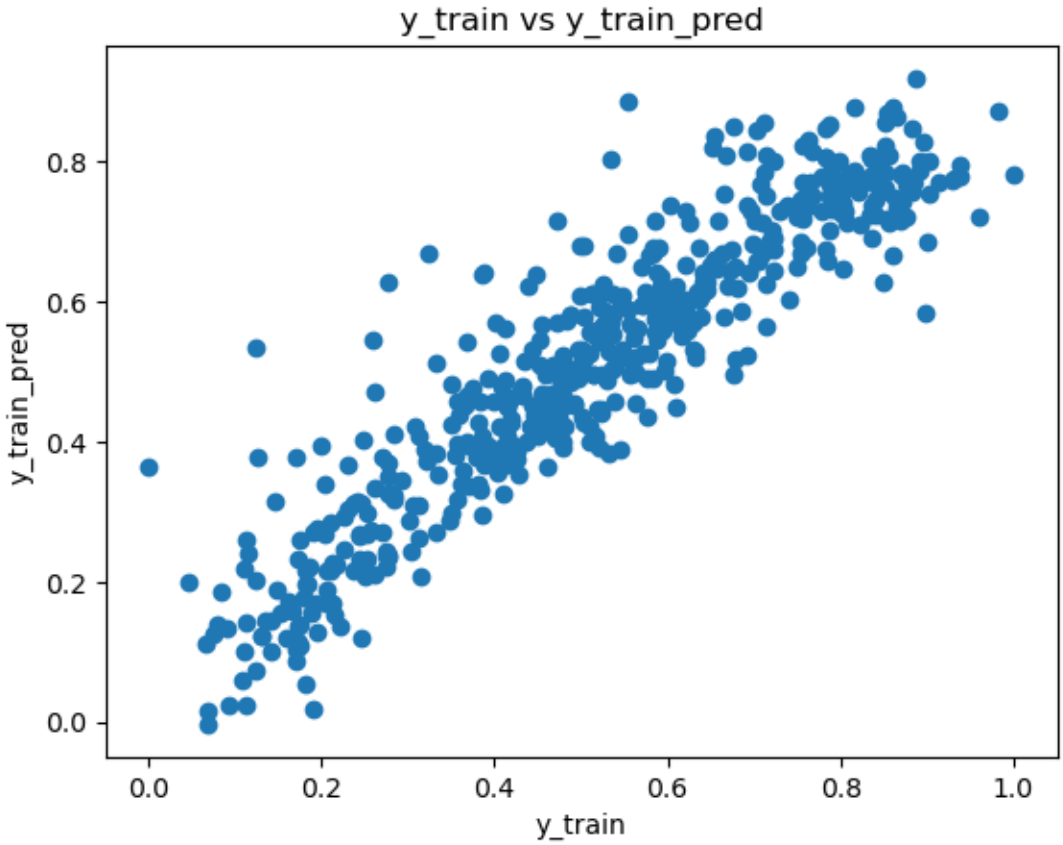
Ans) There are following assumptions on the linear regression, and we are validating them in the following manner:

- ***Both independent variables (X) and dependent variables (y) should be linearly dependent:*** The relationship between X and y is determined by the equation
  $\text{demand(cnt)} = 0.075 + (0.55 * \text{temp}) + (0.056 * \text{workingday}) + (-0.155 * \text{windspeed}) + (0.089 * \text{season2}) + (0.131 * \text{season4}) + (-0.08 * \text{weathersit2}) + (0.233 * \text{yr}) + (0.068 * \text{weekday6}) + (0.097 * \text{mnth9}) + (-0.287 * \text{weather3})$
  Above equation clearly shows a linear relationship between X and y.

- ***Error terms should be normally distributed along with 0 mean:*** The following plot describes the normal distribution nature of the error terms with 0 mean.



- ***Error terms should have a constant variance(Homoscedasticity):*** Below plot shows that the error term has a finite variance.



Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans) Top 3 features contributing significantly towards the demand of the shared bikes are:

- **temp** has coefficient of 0.55
- **yr** has coefficient of 0.231
- **season_4** (winter) has coefficient of 0.131

General Subjective Questions

Q1) Explain the linear regression algorithm in detail.

Ans) Linear regression is a fundamental statistical and machine learning technique used for modeling the relationship between a dependent variable (often denoted as y) and one or more independent variables (often denoted as x1,x2,...,xn). The goal of linear regression is to find the best-fitting straight line (or

hyperplane in higher dimensions) that describes the relationship between the independent variables and the dependent variable. Here's a detailed explanation of the linear regression algorithm:

- **Assumption:** Linear regression assumes that there is a linear relationship between the independent variables and the dependent variable. This means that changes in the dependent variable are proportional to changes in the independent variables.

- **Model Representation:** The linear regression model can be represented mathematically as:
  $y=\beta 0+\beta 1 \cdot x1+\beta 2 \cdot x2+...+\beta n \cdot xn+\epsilon$
  o y is the dependent variable.
  o x1,x2,...,xn are the independent variables.
  o β0 is the intercept (the value of y when all independent variables are zero).
  o β1,β2,...,βn are the coefficients (also known as slopes) of the independent variables, representing the change in y for a one-unit change in the corresponding independent variable.
  o ϵ is the error term, representing the difference between the observed and predicted values of y. It captures all other factors that influence y but are not included in the model.

- **Objective Function**: The objective of linear regression is to minimize the sum of squared differences between the observed values of the dependent variable and the values predicted by the linear model. This objective function is often denoted as the "least squares" criterion:

  $$\text{Minimize} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  Where *yi* are the observed values of the dependent variable, and *y^i* are the predicted values obtained from the linear regression model.

- **Parameter Estimation**: The coefficients β0,β1,...,βn are estimated using various techniques such as Ordinary Least Squares (OLS), gradient descent, or analytical solutions. OLS is a commonly used method that minimizes the sum of squared differences between the observed and predicted values.

- **Model Evaluation**: Once the coefficients are estimated, the model's performance is evaluated using various metrics such as R^2 (coefficient of determination), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), etc. These metrics assess how well the model fits the data and the predictive accuracy of the model.

- **Assumptions Checking**: It's essential to check the assumptions of linear regression, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. Violations of these assumptions can affect the validity and reliability of the regression model.

Q2) Explain the Anscombe's quartet in detail.

Ans) Anscombe's quartet is a collection of four datasets that have nearly identical descriptive statistics but differ significantly when plotted visually. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and to demonstrate the limitations of relying solely on summary statistics.

Here's a detailed explanation of Anscombe's quartet:

- **The Datasets:** Anscombe's quartet consists of four datasets, each containing 11 data points. Despite their identical summary statistics (mean, variance, correlation, and linear regression coefficients), the datasets have distinct patterns when graphed:

  Dataset I: A simple linear relationship.
  Dataset II: Non-linear relationship, but still fits a linear regression model.
  Dataset III: Outlier significantly influences the regression line.
  Dataset IV: Appears to have no correlation when, in fact, it has a strong one, with an outlier influencing the relationship.

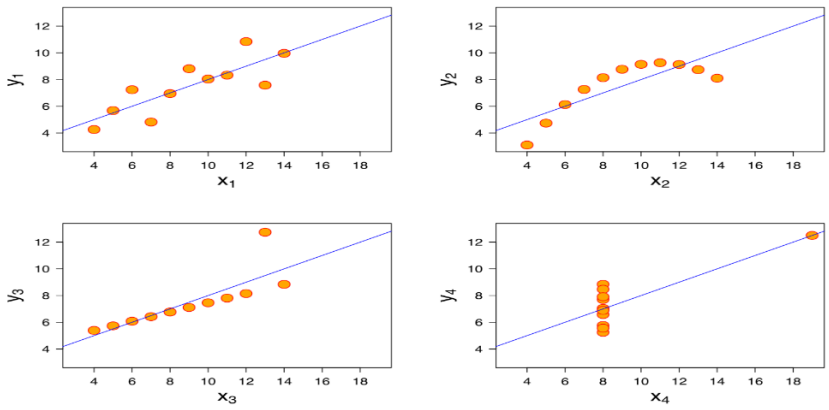
- **Visual Representation:** When plotted, the datasets reveal striking differences, despite having similar statistical properties. For example, Dataset I forms a clear linear relationship when graphed, while Dataset II shows a non-linear relationship. Dataset III has a strong linear relationship except for one outlier point, and Dataset IV appears to have no discernible pattern.

- **Implications:** Anscombe's quartet highlights the limitations of relying solely on summary statistics. While summary statistics can provide valuable insights into the characteristics of a dataset, they may not capture the full complexity or nuances present in the data. Visual exploration through graphs and plots is crucial for understanding the underlying patterns and relationships within the data.

- **Educational Tool:** Anscombe's quartet is often used as an educational tool in statistics and data science to emphasize the importance of data visualization. It underscores the need for exploratory data analysis (EDA) and the potential pitfalls of drawing conclusions based solely on summary statistics without visual inspection of the data.

- **Statistical Concepts Illustrated:** Anscombe's quartet illustrates several key statistical concepts, including:

  - The importance of graphing data to visually inspect patterns and relationships.
  - The limitations of summary statistics in describing the entire dataset.
  - The influence of outliers on statistical analysis and interpretation.
  - The dangers of assuming linear relationships without visual confirmation.
  - In summary, Anscombe's quartet serves as a powerful reminder of the value of data visualization in exploratory data analysis and the potential discrepancies between summary statistics and visual observations.



Q3) What is Pearson's R?

Ans) Pearson's correlation coefficient, often denoted as r, or Pearson's r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, who developed the coefficient.

Pearson's r can take values between -1 and 1:

• r=1: Indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.

• r=−1: Indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.

• r=0: Indicates no linear relationship between the variables.

The formula to calculate Pearson's correlation coefficient r between two variables X and Y is as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Where
• Xi and Yi are individual data points.
• X¯ and Y¯ are the means of the variables X and Y, respectively.
• n is the number of data points.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans) Scaling is the process of transforming the values of variables to a specific range or distribution. It is a common preprocessing step in machine learning and statistical analysis. Scaling ensures that all variables have similar ranges of values, which can help improve the performance and stability of various algorithms.
Scaling is performed for several reasons:

- **Algorithm Sensitivity:** Many machine learning algorithms are sensitive to the scale of input features. For example, algorithms like k-nearest neighbors (KNN), support vector machines (SVM), and gradient descent-based algorithms (e.g., linear regression, logistic regression, neural networks) can perform poorly if the input features have different scales. Scaling helps these algorithms converge faster and produce more accurate results.
- **Distance-Based Algorithms**: Algorithms that use distance measures, such as KNN and SVM, are heavily influenced by the scale of input features. Without scaling, variables with larger scales can dominate the distance calculations, leading to biased results.

- **Regularization**: Regularization techniques, such as ridge regression and lasso regression, penalize large coefficients. Scaling helps ensure that all features are penalized equally, preventing some features from dominating the regularization process.
- **Interpretability**: Scaling can improve the interpretability of coefficients in linear models by putting them on the same scale. This allows for a more direct comparison of the effects of different features on the outcome.

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

Normalized Scaling (Min-Max Scaling):

- Normalized scaling, also known as min-max scaling, transforms the values of variables to a specific range, typically between 0 and 1.
- The formula for min-max scaling is:

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- This method preserves the relative distances between data points but may not handle outliers well.

Standardized Scaling (Z-score Scaling):

- Standardized scaling, also known as z-score scaling or standardization, transforms the values of variables to have a mean of 0 and a standard deviation of 1.
- The formula for standardization is

$$x_{\text{standardized}} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

- This method centers the data around the mean and scales it according to the standard deviation, making it more robust to outliers.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans) 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 is inflated due to multicollinearity in the predictor variables. In some cases, the VIF may be calculated as infinite. This happens when there is perfect multicollinearity among the predictor variables.

Perfect multicollinearity occurs when one or more independent variables in a regression model can be perfectly predicted by a linear combination of other independent variables. In other words, there is a perfect linear relationship between two or more predictor variables. This situation makes it impossible to estimate unique regression coefficients for each predictor variable, resulting in an infinite VIF.

For example, if you have a variable that is a constant multiple of another variable (e.g., X2=2X1), then there is perfect multicollinearity between X1 and X2, and it is not possible to estimate separate coefficients for each variable. As a result, the VIF for one of these variables will be infinite.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans) A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether or not a dataset follows a particular probability distribution, typically the normal distribution. It compares the quantiles of the dataset's empirical distribution to the quantiles of a theoretical distribution (such as the normal distribution).
Here's how a Q-Q plot works:

- **Sorting the Data**: First, the data points in the dataset are sorted in ascending order.
- **Calculating Quantiles**: Quantiles of the dataset's empirical distribution are calculated. These quantiles represent the percentage of data points that fall below a given value.
- **Calculating Theoretical Quantiles**: Theoretical quantiles are calculated based on the chosen probability distribution (e.g., normal distribution).
- **Plotting**: The calculated quantiles from the dataset are plotted against the theoretical quantiles. If the dataset follows the theoretical distribution closely, the points on the Q-Q plot will fall approximately along a straight line.

The use and importance of a Q-Q plot in linear regression include:

- **Assumption Checking**: Q-Q plots are commonly used to check whether the residuals (the differences between observed and predicted values) of a linear regression model follow a normal distribution. This is an important assumption of linear regression, and violating it can affect the validity of statistical inference and prediction.
- **Detecting Departures from Normality**: A deviation from a straight line in the Q-Q plot suggests that the residuals do not follow a normal distribution. This can indicate the presence of outliers, skewness, or other departures from normality in the data. Identifying such departures helps in diagnosing potential issues with the regression model and in deciding whether corrective actions, such as data transformation or using robust regression techniques, are necessary.
- **Comparing Models**: Q-Q plots can also be used to compare the normality of residuals across different models. For example, if you have fitted several linear regression models with different sets of predictors, you can compare their Q-Q plots to assess which model's residuals better approximate a normal distribution.

Overall, Q-Q plots provide a visual and diagnostic tool for assessing the assumption of normality in linear regression models and for identifying potential issues with the model's residuals, helping to ensure the validity and reliability of regression analysis.
