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?

Ans: Here are some insights from the analysis of categorical variables and their impact on the dependent variable:

- The months of June to September have higher rental counts.
- Bike rentals increased in 2019 compared to 2018, possibly due to increased environmental awareness.
- Bike rentals are more frequent on non-holiday days than on holidays, suggesting that people prefer staying at home during holidays.
- The fall season has the highest median bike rentals, followed by summer, likely due to the favorable weather conditions in these seasons.
- There is no significant difference between the medians of weekdays and working days.
- Clear weather conditions have the highest median rental counts, while light rain has the lowest. The weather situation Thunderstorm ('Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog') records zero bike rentals.

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

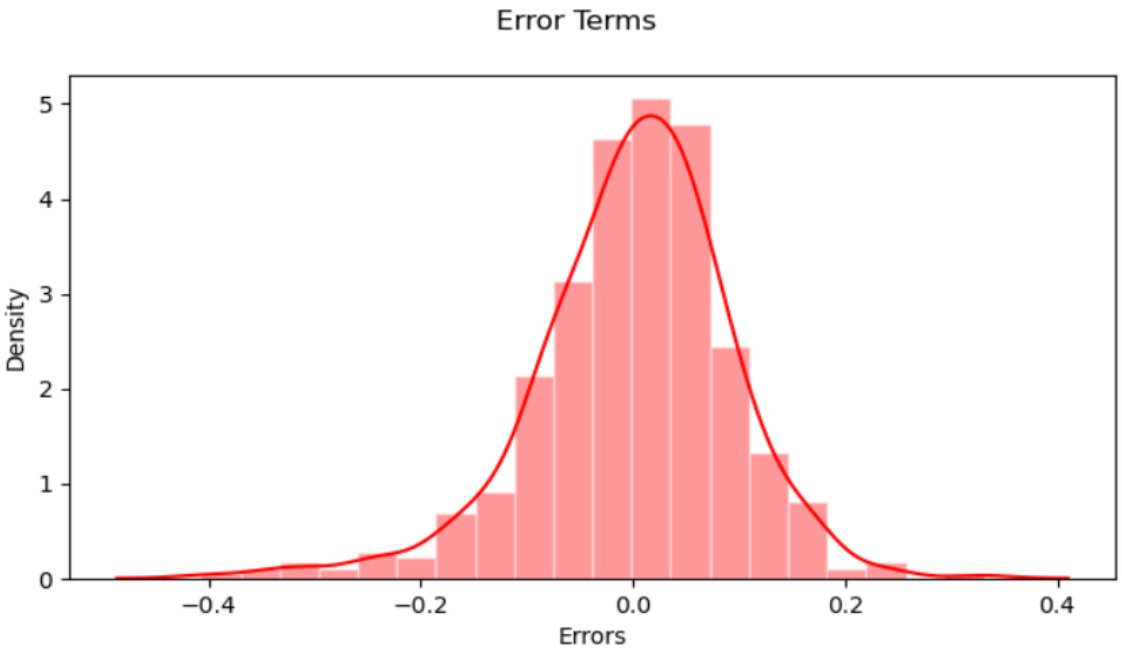
Ans: Using `**drop_first=True**` in dummy variable creation helps prevent multicollinearity, as it omits one level of each categorical variable to avoid perfect multicollinearity, leading to more stable and interpretable models.

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

Ans: The variables 'atemp' and 'temp' exhibit the strongest correlation (0.65) with the target variable 'cnt'.

4) How did you validate the assumptions of linear regression after building the model on the training set?

Ans: We assess the assumptions of linear regression by creating a distribution plot of the residuals and examining whether they follow a normal distribution and have a mean of 0. The plot below confirms a normal distribution with a mean of 0.



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

Ans: Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. Temperature (temp) with a coefficient of 0.5480, indicating a positive and substantial impact on bike demand
2. Year 2019 (Year_2019) with a coefficient of 0.2329, suggesting a positive influence on bike demand.
3. Winter Season (Season_winter) with a coefficient of 0.1293, indicating a positive effect on bike demand.

General Subjective Questions

1) Explain the linear regression algorithm in detail ?

Ans: Linear regression is a fundamental statistical and machine learning algorithm used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables). Its primary purpose is to model the linear relationship between the independent variables and the dependent variable.

Linear regression can be classified into two types, simple linear Regression and multiple linear Regression

Steps to follow while using Algorithm

- Data Collection: Gather a dataset with the target variable and one or more feature variables.
- Data Preprocessing: Clean and preprocess the data by handling missing values, outliers, and normalizing the data.
- Feature Selection: Choose relevant predictor variables that have a strong correlation with the target variable.
- Model Training: Fit a linear regression model to the training data, which estimates the coefficients (weights) for each feature.
- Prediction: Use the trained model to make predictions on new or test data.
- Model Evaluation: Assess the model's performance using evaluation metrics such as mean squared error (MSE), R-squared, or root mean squared error (RMSE).

Key Components:

- Linear Equation: Linear regression assumes that the relationship between the target variable (Y) and the predictors (X) can be expressed as a linear equation: $Y = \beta0 + \beta1X1 + \beta2X2 +... + \beta nXn$.

- Coefficients (β): These are the weights assigned to each predictor variable, indicating their impact on the target variable. The goal is to learn the best coefficients during model training.

- Cost Function: The model minimizes a cost function (e.g., MSE) to find the optimal coefficients that provide the best fit to the data.

- Least Squares Method: Linear regression often uses the least squares method to minimize the sum of the squared differences between predicted and actual values.

2) Explain Anscombe's quartet in detail ?

Ans: Anscombe's Quartet is a set of four small datasets that have nearly identical simple descriptive statistics (mean, variance, and correlation) but differ significantly when graphed. These datasets were created by Francis Anscombe in 1973 to demonstrate the importance of data visualization in statistical analysis.

Each dataset in the quartet consists of 11 data points with two variables (x and y). Here's an overview of the four datasets:

1. Dataset I: Linear Relationship
   - 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

   This dataset exhibits a clear linear relationship, and a linear regression model fits it well.

2. Dataset II: Non-linear Relationship
   - x values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26

   While the x and y values are the same as in Dataset I, this dataset shows a non-linear relationship. A linear regression model is not appropriate for this data.

3. Dataset III: Linear Relationship with an Outlier
   - 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

   This dataset still maintains a linear relationship, but an outlier significantly affects the linear regression model's parameters.

4. Dataset IV: No Clear Relationship

- 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, 12.50, 5.56, 7.91

Dataset IV demonstrates that a linear regression model can provide misleading results when there is no substantial relationship between the variables.

3) What is Pearson's R?

Ans: Pearson's correlation coefficient (Pearson's R), also known as Pearson's R, is a statistic that quantifies the linear relationship (correlation) between two continuous variables. It measures both the strength and direction of the linear association between these variables. Pearson's R is a widely used method for assessing the degree to which two variables are related.

Key characteristics of Pearson's R:

1. Range: Pearson's R varies between -1 and 1.
   - A value of 1 indicates a perfect positive linear relationship.
   - A value of -1 indicates a perfect negative linear relationship.
   - A value of 0 suggests no linear relationship.

2. Interpretation:
   - If R is close to 1, it implies a strong positive linear relationship.
   - If R is close to -1, it implies a strong negative linear relationship.
   - If R is close to 0, it suggests a weak or no linear relationship.

3. Assumptions:
   - Pearson's R assumes that both variables are continuous and have a linear relationship.
   - It is sensitive to outliers.

4. Formula:
   - The formula for Pearson's R is:

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

- Where:

- r represents the correlation coefficient.
- xi denotes the values of the x-variable in a sample.
- $\bar{x}$ stands for the mean of the values of the x-variable.
- yi represents the values of the y-variable in a sample.
- $\bar{y}$ represents the mean of the values of the y-variable.

Pearson's R is valuable in various fields, including statistics, social sciences, and natural sciences, as it allows researchers to assess and quantify relationships between variables. It is a fundamental tool for bivariate correlation analysis and is often used in statistical software and data analysis tools for hypothesis testing and predictive modeling.

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

Ans : Scaling is a data preprocessing technique used in machine learning and statistics to transform the range of data values while maintaining their proportional relationships. It is performed to ensure that the features (variables) in a dataset are on a similar scale or have the same order of magnitude. Scaling is essential for many machine learning algorithms, particularly those based on distances or gradient descent, as it helps improve the performance and convergence of these algorithms. There are two common scaling techniques: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):
   - In normalized scaling, data values are transformed to a specific range, typically between 0 and 1.
   - It is performed using the following formula for each data point x:
     Xnormalized= { (X –Xmin) / (Xmax– Xmin) }

- The minimum value in the dataset is mapped to 0, and the maximum value is mapped to 1.

2. Standardized Scaling (Z-score Scaling):
- Standardized scaling, also known as Z-score scaling, transforms data into a standard normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1.
- It is performed using the following formula for each data point x:
  X standardized = (x - μ) / σ
- Here, μ is the mean of the dataset, and σ is the standard deviation.
- The result of this scaling is that the transformed data will have a mean of 0 and unit variance.

Differences:
- Normalized scaling ensures that data values are in a specific range, typically [0, 1], making it useful when the exact range is important.
- Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1. It is more appropriate when data follows a normal distribution and when comparing different datasets.
- Normalized scaling may not handle outliers well, as they can significantly affect the minimum and maximum values.
- Standardized scaling is robust to outliers and is often a better choice when data contains outliers.
- The choice between these two scaling methods depends on the characteristics of the data and the requirements of the machine learning algorithm being used.

5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans : A VIF (Variance Inflation Factor) becomes infinite when perfect multicollinearity is present in the dataset. Perfect multicollinearity occurs when one independent variable in a regression model can be exactly predicted by a linear combination of one or more other independent variables. In other words, there is an exact linear relationship between variables.

This can happen when, for example, two or more variables are linearly dependent on each other, meaning that one variable can be expressed as a constant multiple of the other(s). When calculating VIF, the formula involves dividing by 1 minus the squared correlation (R-squared) between a variable and the other variables in the

model. If the R-squared is exactly 1, the denominator becomes 0, causing the VIF to be infinite.

To avoid infinite VIF values, it's important to detect and address multicollinearity by removing or transforming variables in the regression model, ensuring that they are not perfectly correlated.

6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution, usually the normal distribution. It compares the quantiles of the dataset to the quantiles of the chosen theoretical distribution.

In the context of linear regression, Q-Q plots are important for several reasons:

- Assumption Checking: Linear regression models often assume that the residuals (the differences between actual and predicted values) are normally distributed. By plotting the residuals on a Q-Q plot, you can check whether they follow a normal distribution. Deviations from a straight line on the plot can indicate departures from the normality assumption.

- Identifying Outliers: Q-Q plots can help identify outliers or extreme observations in the dataset. Outliers can strongly affect the regression results, so it's crucial to detect them and decide whether to remove or transform these data points.

- Model Assessment: A well-behaved Q-Q plot suggests that the model's assumptions hold, which can lead to more reliable and interpretable regression results.

To create a Q-Q plot, you typically sort the residuals and plot them against the expected values of the chosen theoretical distribution (usually a normal distribution). If the points lie approximately on a straight line, it indicates that the data follows the distribution. Deviations from the line suggest deviations from the theoretical distribution, and this may require further investigation or data transformation.