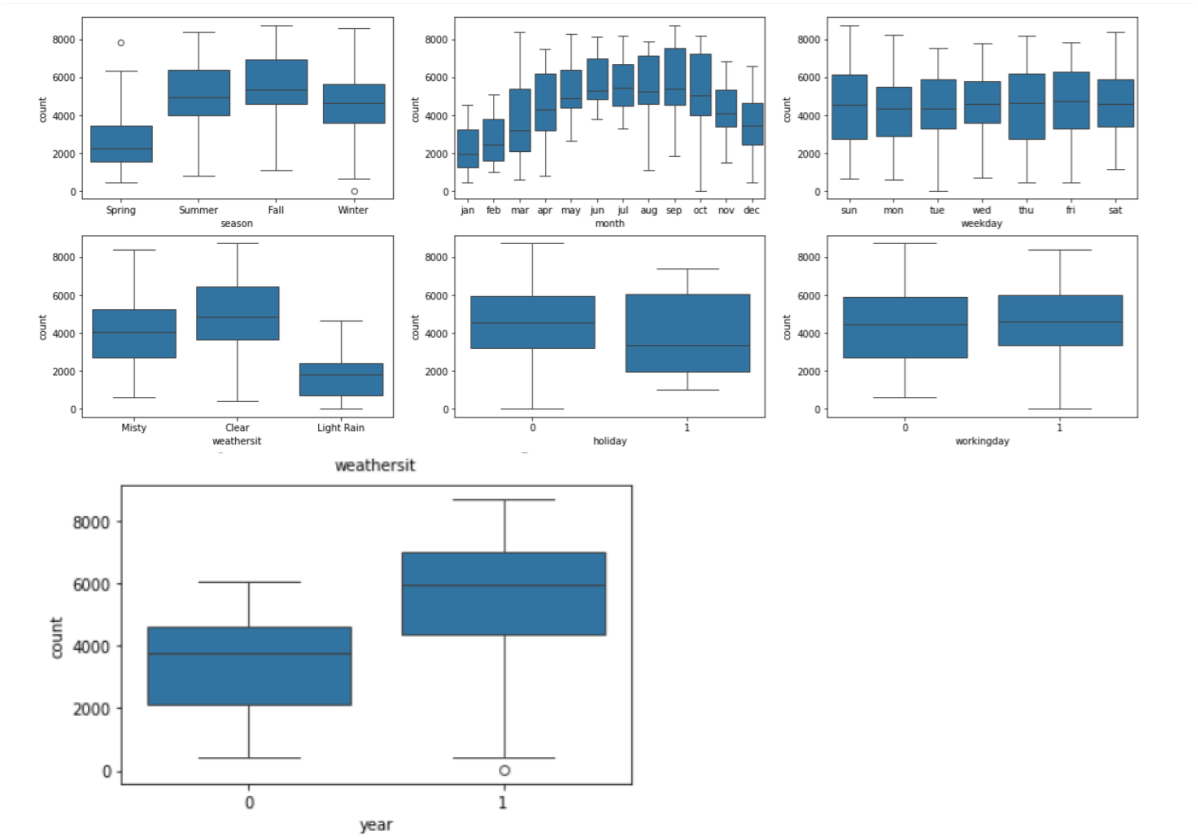
Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Visualising Categorical columns to see the coorelation of predictor variable with the target variable using Box plots were done



From the visualizations, we can derive several insights:

- Seasonal Trends: The demand for rental bikes peaks in the fall season.

- Monthly Trends: The demand for rental bikes rises steadily each month until June, with September experiencing the highest demand before it starts to decline.

- Weekdays and Working Days: There isn't much variation in demand during weekdays and working days.

- Weather Conditions: Clear weather conditions see the highest demand for bike rentals.

- Holidays: The demand for bike rentals decreases on holidays.

- Yearly Growth: There is noticeable growth in demand for the following year.

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

Answer: Using drop_first=True during dummy variable creation is important to avoid multicollinearity. By dropping the first category, we prevent the dummy variables from being

perfectly collinear, which can cause issues in the regression model. This ensures that the model remains stable and interpretable.

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

Answer: From the pair-plot analysis, it is observed that the variable temp (temperature) has the highest correlation with the target variable cnt (count of bike rentals). This indicates that temperature is a strong predictor of bike rental demand.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: To validate the assumptions of Linear Regression, the following steps were taken:

1. **Linearity**: Checked by plotting the residuals versus the fitted values to ensure there is no pattern.
2. **Normality**: Verified using a Q-Q plot to ensure the residuals are normally distributed.
3. **Homoscedasticity**: Ensured by plotting the residuals versus the fitted values to check for constant variance.
4. **Multicollinearity**: Assessed using Variance Inflation Factor (VIF) to ensure no high correlation among the independent variables.

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

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

1. **Temperature (temp)**
2. **Year**
3. **Light Rain**

These features have the highest coefficients and are statistically significant in predicting bike rental demand.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail.

Answer: Answer:

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (often called the target or outcome variable) and one or more independent variables (also known as predictors or features). The goal of linear regression is to find the best-

fitting line that describes how the dependent variable changes as the independent variables change.

The Linear Regression Equation

The equation of a simple linear regression model with one independent variable is:

[y = \beta_0 + \beta_1 x_1]

For multiple linear regression, where there are multiple independent variables, the equation is extended to:

[y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n]

Here:

- (y) is the dependent variable.
- (\beta_0) is the intercept (the value of (y) when all (x) values are zero).
- (\beta_1, \beta_2, ..., \beta_n) are the coefficients (slopes) for the independent variables (x_1, x_2, ..., x_n).

Estimating the Coefficients

The coefficients ((\beta)) are estimated using the **least squares method**, which minimizes the sum of the squared differences between the observed values and the values predicted by the model. This is known as minimizing the **residual sum of squares (RSS)**:

[RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2]

Where:

- (y_i) is the observed value.
- (\hat{y}_i) is the predicted value from the model.

Steps in Linear Regression

1. **Data Collection**: Gather the data that includes the dependent variable and the independent variables.
2. **Data Preprocessing**: Clean the data, handle missing values, and perform any necessary transformations.
3. **Model Building**: Use the training data to estimate the coefficients ((\beta)) by fitting the linear regression model.
4. **Model Evaluation**: Assess the model's performance using metrics such as R-squared, adjusted R-squared, and residual analysis.
5. **Prediction**: Use the fitted model to make predictions on new data.

Assumptions of Linear Regression

For linear regression to provide reliable results, several assumptions must be met:

- **Linearity**: The relationship between the dependent and independent variables is linear.

- **Independence**: The residuals (errors) are independent.
- **Homoscedasticity**: The residuals have constant variance at every level of the independent variables.
- **Normality**: The residuals are normally distributed.

Model Evaluation Metrics

- **R-squared (R^2)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-squared**: Adjusts the R-squared value based on the number of predictors in the model.
- **Mean Squared Error (MSE)**: The average of the squared differences between the observed and predicted values.

Example

Suppose we have a dataset with the dependent variable cnt (bike rentals) and independent variables such as temp (temperature), humidity, and windspeed. The linear regression model might look like this:

[cnt = \beta_0 + \beta_1 \cdot temp + \beta_2 \cdot humidity + \beta_3 \cdot windspeed]

By fitting this model to the data, we estimate the coefficients ((\beta)) that minimize the RSS, allowing us to predict bike rentals based on temperature, humidity, and windspeed.

Question 7. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a collection of four datasets that were constructed by the statistician Francis Anscombe in 1973. The purpose of these datasets is to illustrate the importance of graphing data before analyzing it and to demonstrate how different datasets can have identical or nearly identical simple descriptive statistics, yet reveal very different patterns when visualized.

The Four Datasets

Each of the four datasets in Anscombe's quartet consists of eleven (x, y) points. Despite having nearly identical statistical properties, such as the mean, variance, correlation, and linear regression line, the datasets exhibit distinct distributions and relationships when plotted.

Statistical Properties

For all four datasets:

The mean of x is 9.

The mean of y is approximately 7.5.

The variance of x is 11.

The variance of y is approximately 4.12.

The correlation between x and y is approximately 0.816.

The linear regression line is (y = 3 + 0.5x).

Visual Differences

When plotted, the datasets reveal the following distinct patterns:

Dataset I: This dataset appears to follow a linear relationship, and the points are scattered around the regression line with a typical spread.

Dataset II: This dataset forms a clear curve, indicating a non-linear relationship between x and y. The linear regression line does not fit the data well.

Dataset III: This dataset includes an outlier that significantly affects the regression line. Without the outlier, the data points would show a linear relationship.

Dataset IV: This dataset has a vertical line of points with one outlier. The outlier heavily influences the regression line, making it appear as if there is a relationship between x and y when there is none.

Importance of Anscombe's Quartet

Anscombe's quartet emphasizes several key points in data analysis:

Graphical Analysis: Visualizing data through plots and graphs is crucial for understanding the underlying patterns and relationships. Relying solely on summary statistics can be misleading.

Outliers and Influential Points: Outliers and influential points can significantly affect statistical measures and regression models. Identifying and addressing these points is essential for accurate analysis.

Model Appropriateness: The choice of model should be based on the data's characteristics. For example, a linear model may not be appropriate for data that exhibits a non-linear relationship.

By highlighting these points, Anscombe's quartet serves as a powerful reminder of the importance of exploratory data analysis and the limitations of relying solely on summary statistics..

Question 8. What is Pearson's R?

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which two variables are

related and is widely used in statistics to understand the strength and direction of the relationship.

Understanding Pearson's R

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

- **1** indicates a perfect positive linear relationship.
- **-1** indicates a perfect negative linear relationship.
- **0** indicates no linear relationship.

Calculation of Pearson's R

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:

- (X_i) and (Y_i) are the individual sample points.
- (\bar{X}) and (\bar{Y}) are the means of the X and Y variables, respectively.

Interpretation of Pearson's R

- **Positive Correlation**: If Pearson's R is greater than 0, it indicates a positive correlation, meaning that as one variable increases, the other variable also increases.
- **Negative Correlation**: If Pearson's R is less than 0, it indicates a negative correlation, meaning that as one variable increases, the other variable decreases.
- **No Correlation**: If Pearson's R is close to 0, it indicates no linear correlation between the variables.

Example

Let's say we have two variables, X and Y, representing the number of hours studied and the scores obtained in an exam, respectively. By calculating Pearson's R, we can determine how strongly the number of hours studied is related to the exam scores.

Applications of Pearson's R

Pearson's R is used in various fields such as:

- **Psychology**: To measure the relationship between different psychological traits.
- **Economics**: To understand the relationship between economic indicators.
- **Medicine**: To study the correlation between different health parameters.

Limitations of Pearson's R

- **Linear Relationship**: Pearson's R only measures linear relationships. It does not capture non-linear relationships.
- **Outliers**: The presence of outliers can significantly affect the value of Pearson's R.

- **Assumptions**: It assumes that the variables are normally distributed and have a linear relationship.

In summary, Pearson's R is a powerful statistical tool for measuring the strength and direction of the linear relationship between two variables. It provides valuable insights into how variables are related, helping researchers and analysts make informed decisions.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is the process of transforming the features of a dataset to a common scale. It is performed to ensure that all features contribute equally to the model and to improve the convergence of gradient-based optimization algorithms. Normalized scaling transforms the data to a range of [0, 1], while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: The value of VIF can be infinite when there is perfect multicollinearity among the independent variables, meaning that one variable is a perfect linear combination of others. This causes the denominator in the VIF calculation to be zero, resulting in an infinite value. Perfect multicollinearity indicates that the model cannot uniquely estimate the coefficients of the independent variables.

Why does this happen?

- **Perfect Multicollinearity**: When one predictor variable can be perfectly predicted from the others, the matrix used in the VIF calculation becomes singular, leading to a division by zero.
- **Implications**: Perfect multicollinearity indicates that the model cannot uniquely estimate the coefficients of the independent variables. This makes it impossible to determine the individual effect of each predictor on the dependent variable, leading to unreliable and unstable estimates.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

Checking Normality: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps in visually assessing this assumption. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed.

Identifying Deviations: Deviations from the straight line in a Q-Q plot indicate departures from normality. This can help identify skewness, kurtosis, or other distributional issues in the residuals.

Model Validity: Ensuring that the residuals are normally distributed is crucial for the validity of the regression model. It affects the accuracy of confidence intervals and hypothesis tests. A Q-Q plot provides a simple and effective way to check this assumption.

In summary, a Q-Q plot is an essential diagnostic tool in linear regression for verifying the normality of residuals, which is a critical assumption for the validity of the model.