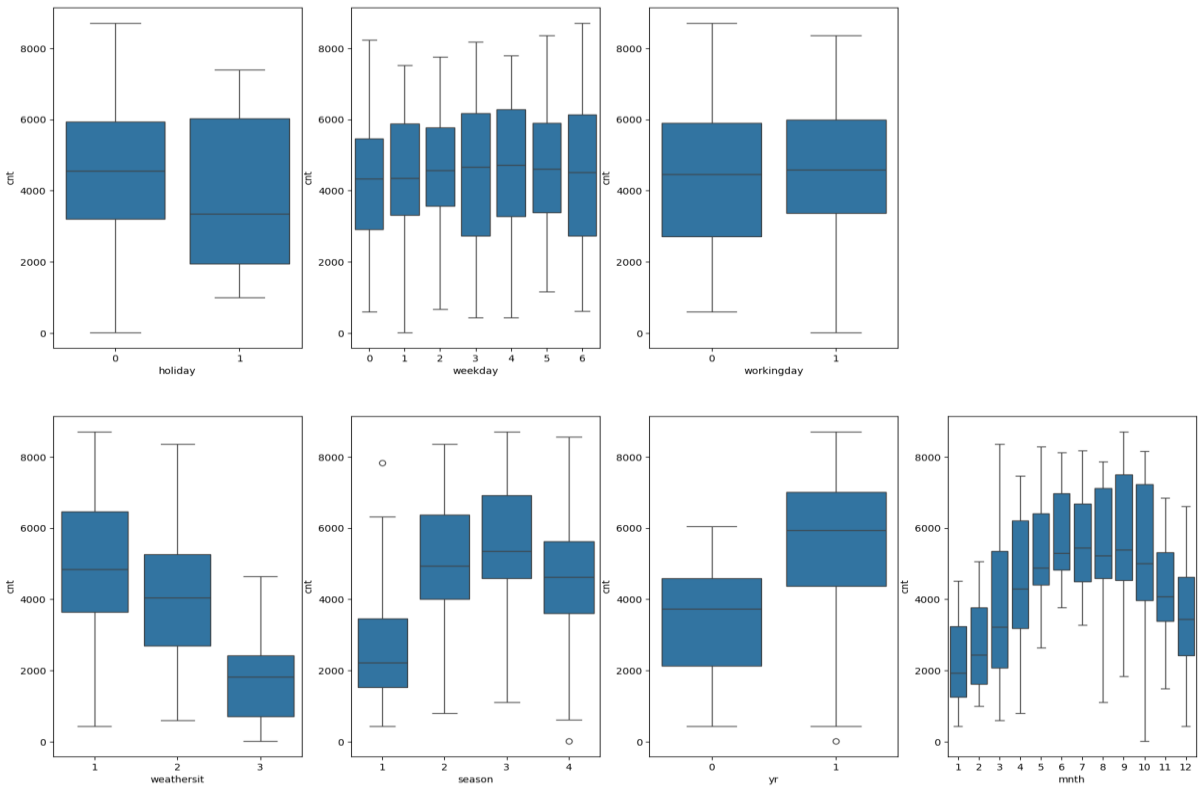
Assignment-based Subjective Questions

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

Categorical variables are Season, Yr, weekday, holiday, workingday, weathersit and mnth



1. The graph effectively illustrates the qualitative distribution of the data. When combined with the model's identified key predictors, these graphs enhance our confidence in the model's predictions. For the variable "season," it is evident that Category 3 (Fall) has the highest median, indicating increased demand during this season, while Category 1 (Spring) shows the lowest demand.

2. In comparison, 2019 saw a higher user count than 2018. Rental counts are fairly consistent throughout the week. There is a noticeable drop in rentals during heavy rain or snow, suggesting these weather conditions are particularly adverse. The highest rental counts occur during Clear and Partly Cloudy weather conditions.

3. Rentals peaked in September, while December saw a decrease, likely due to typical substantial snowfall during that month. Additionally, user counts are lower during holidays.

4. The workingday boxplot shows that most bookings occur between 4000 and 6000, with the median user count remaining relatively stable throughout the week. There is little difference in bookings whether it is a working day or not.

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

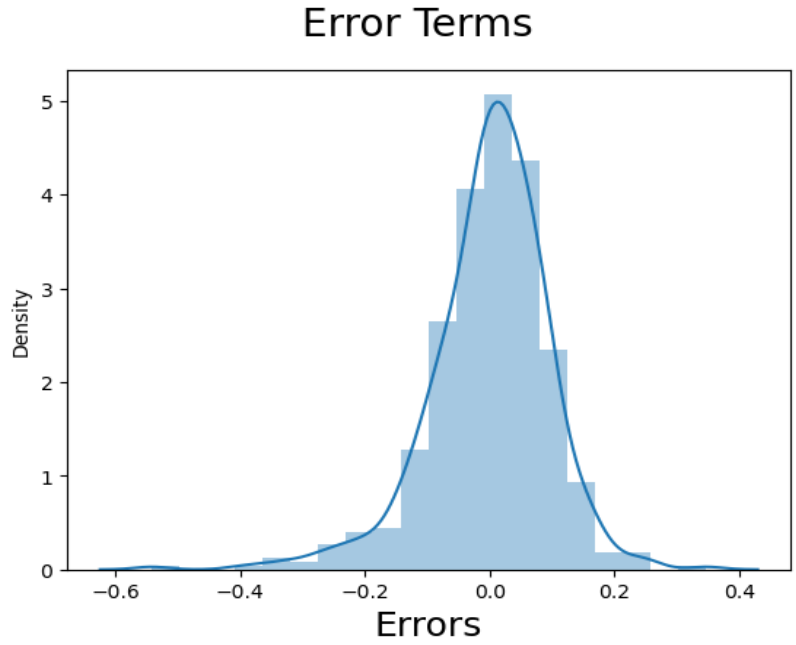
Using drop_first=True is crucial because it prevents the creation of an extra column during dummy variable encoding, thereby reducing the correlations among the dummy variables. For a categorical variable with n levels, this approach requires using only n−1 columns to represent the dummy variables.

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

Temp and atemp variables are highly correlated with the target variable cnt.

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

1. First we will always validate for there should be linear relationship between independent and dependant variables. To check we can use pair plot which we have done
2. Secondly Residuals follows always follows normal distribution and mean always zero. We validated this assumption by plotting displot of residuals.



3. Linear regression assumes minimal or no multicollinearity among the data. Multicollinearity arises when the independent variables are highly correlated with each other. To assess the extent of this correlation, we calculated the Variance Inflation Factor (VIF), which measures how strongly the feature variables in the new model are related to one another.

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

3 Significant features are 1. Temperature (0.493719)

2. weathersit : Pleasant (0.097303)
3. yr (0.237519)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a foundational algorithm in machine learning, falling under the category of supervised learning. It performs regression tasks, where it predicts a dependent (target) value based on independent variables. This technique is primarily used to explore relationships between variables and for forecasting. Various regression models differ in the type of relationship they assume between dependent and independent variables, as well as the number of independent variables they incorporate.

Linear regression is used to predict the value of a dependent variable (y) based on a given independent variable (x). This technique identifies a linear relationship between the input (x) and the output (y), which is why it is called Linear Regression.

Linear Regression may further divided into

1. **Simple Linear Regression**
2. **Multiple Linear Regression**

The mathematical equation can be given as:
$Y = \beta0 + \beta1*x$
Where

- Y is the response or the target variable
- x is the independent feature
- β1 is the coefficient of x
- β0 is the intercept

β0 and β1 are the model coefficients (or weights). To create a model, we must learn the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the target variable such as Sales!
NOTE: The main aim of the regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line.

Assumptions of Simple Linear Regression

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality

Multiple Linear Regression:

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable

Formula : y = A+B1x1+B2x2+B3x3+B4x4

Assumptions of Simple Linear Regression

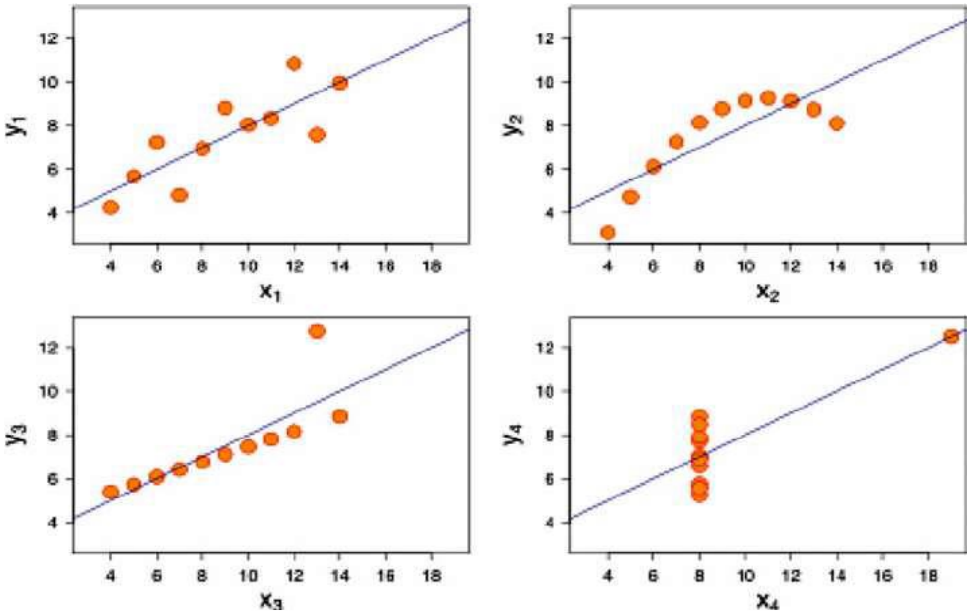
1. No multicollinearity
2. Additivity

3. Feature selection
4. Over fitting

2.Explain the Anscombe's quartet in detail.

Anscombe's Quartet, created by statistician Francis Anscombe, consists of four datasets with nearly identical statistical properties but vastly different distributions and appearances when plotted. This quartet was designed to highlight the importance of visualizing data before analysis and to demonstrate the impact of outliers and other influential observations on statistical measures.

- **The first scatter plot (top left)**: Displays a straightforward linear relationship.
- **The second plot (top right)**: Shows a non-normal distribution; the relationship is evident but not linear.
- **The third plot (bottom left)**: Exhibits a linear distribution, but the presence of an outlier affects the regression line. This outlier significantly influences the results, reducing the correlation coefficient from 1 to 0.816.
- Finally, the fourth plot (bottom right) illustrates a case where a single high-leverage point can produce a high correlation coefficient, despite the other data points showing no apparent relationship between the variables.



3.What is Pearson's R?

Pearson's r, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the association between them. Here's a detailed explanation:

Definition

Pearson's r is a statistical coefficient that ranges from -1 to 1:

- **r=1**: Perfect positive linear relationship.
- **r=−1**: Perfect negative linear relationship.
- **r=0**: No linear relationship.

Formula

The Pearson correlation coefficient is calculated using the formula:

Formula

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

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Interpretation

- **Positive rrr**: Indicates a positive linear relationship, meaning that as one variable increases, the other also tends to increase.
- **Negative rrr**: Indicates a negative linear relationship, meaning that as one variable increases, the other tends to decrease.

Assumptions

Pearson's r relies on several assumptions:

1. **Linearity**: The relationship between the variables should be linear.
2. **Homogeneity of Variance**: The variance of the variables should be roughly equal across the range of the data.
3. **Normality**: The variables should be approximately normally distributed, though Pearson's rrr is fairly robust to deviations from normality with large sample sizes.

Usage

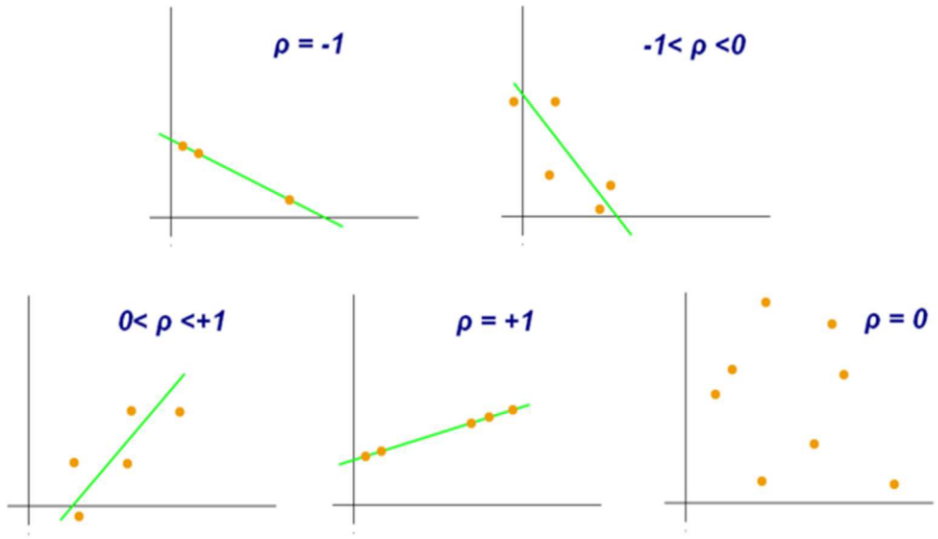
Pearson's rrr is used in various fields to:

- Determine the strength and direction of linear relationships between variables.
- Test hypotheses about the correlation between variables.
- Inform regression analysis and other statistical models.

Limitations

- **Non-Linear Relationships**: Pearson's rrr is not suitable for detecting non-linear relationships.
- **Outliers**: Outliers can significantly affect the value of Pearson's rrr and may give a misleading impression of the relationship.

As illustrated in the graph below, r=1r = 1r=1 indicates a perfect positive linear relationship, r=−1r = -1r=−1 signifies a perfect negative linear relationship, and r=0r = 0r=0 denotes no linear association between the variables.



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

Scaling is a process in data preprocessing that involves adjusting the range of feature values to make them comparable. It transforms the data to a common scale without distorting differences in the range of values.

Why is Scaling Performed?

Scaling is performed for several reasons:

1. **To Ensure Uniformity**: Features measured on different scales (e.g., height in centimeters and weight in kilograms) can affect the performance of algorithms, particularly those that use distance calculations, like k-nearest neighbors and gradient descent.
2. **To Improve Convergence**: Algorithms that rely on optimization, such as gradient descent, often converge faster and more reliably when features are scaled to a similar range.
3. **To Avoid Bias**: Scaling prevents features with larger numerical ranges from disproportionately influencing the model.

Difference Between Normalized Scaling and Standardized Scaling

1. Normalized Scaling:

- **Definition**: Also known as Min-Max scaling, it transforms the data to fit within a specified range, typically [0, 1].
- **Formula**: Xnorm=X−Xmin/Xmax−Xmin
- **Use Case**: Useful when features have different units or ranges and when you need to bound the feature values within a specific range.
- **Characteristics**: Preserves the relationships between the original values but can be sensitive to outliers, which can skew the range.

2. Standardized Scaling:

- **Definition**: Also known as Z-score normalization, it centers the data around the mean and scales it according to the standard deviation.
- **Formula**: Xstandard = (X- μ)/ σ

where μ is the mean of the feature, and σ is the standard deviation.

- **Use Case**: Commonly used when the data is normally distributed or when you want to standardize the features for algorithms that assume normally distributed data, like linear regression.
- **Characteristics**: Converts the feature values into a distribution with a mean of 0 and a standard deviation of 1. It is less affected by outliers compared to normalization.

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

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF value can become infinite in the following scenarios:

Reasons for Infinite VIF

1. **Perfect Multicollinearity**:
    - **Definition**: This occurs when one predictor variable is a perfect linear combination of one or more other predictor variables.
    - **Example**: If variable X1 can be exactly expressed as a linear combination of X2 and X3 (e.g., $X1=2X2-X3$), the matrix of predictors becomes singular, and the determinant of the matrix is zero. This leads to an infinite VIF because the calculation involves dividing by zero.
2. **Singular Matrix**
    - **Definition**: When the matrix of predictors (design matrix) is singular or nearly singular, it means the matrix does not have full rank. This happens if there is exact linear dependence among the predictors.
    - **Example**: If you include a column of ones in your design matrix for an intercept term and another column that is a perfect multiple of it, the matrix becomes singular, leading to an infinite VIF for those predictors.
    - **Numerical Precision Issues:**
    - Definition: In some cases, numerical precision issues can cause problems in the computation, especially when working with very large or very small numbers.
    - Example: Floating-point precision limits can sometimes result in very high values for VIF, which may be treated as infinite in practical calculations.

Consequences of Infinite VIF

- **Model Instability**: Infinite VIF indicates extreme multicollinearity, which can cause instability in the regression coefficients and make the model's results unreliable.
- **Inaccurate Coefficients**: High multicollinearity can lead to large standard errors for the regression coefficients, making it difficult to assess the individual impact of predictors.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the dataset with the quantiles of a theoretical distribution.

I**mportance of a Q-Q Plot in Linear Regression**

In the context of linear regression, a Q-Q plot is primarily used to check the assumption of normality of residuals. Here's why it's important:

1. **Normality of Residuals**:
    - **Assumption**: Linear regression assumes that the residuals (the differences between observed and predicted values) are normally distributed. This is crucial for valid hypothesis testing and confidence intervals for the regression coefficients.
    - **Q-Q Plot Use**: A Q-Q plot of the residuals can help visually assess whether this normality assumption holds. If the residuals follow a normal distribution, they should plot approximately along the reference line.
2. **Detecting Non-Normality:**
    - **Implications**: If the residuals deviate significantly from the reference line, it may indicate that the normality assumption is violated. This can affect the validity of p-values, confidence intervals, and predictions made by the regression model.
    - **Action**: If non-normality is detected, it may be necessary to apply transformations to the dependent variable or use robust regression techniques that do not rely on the normality assumption.
3. **Model Diagnostics:**
    - **Good Fit:** Besides normality, the Q-Q plot can also help in diagnosing other model issues if residuals show systematic patterns or deviations, such as heteroscedasticity or outliers.
    - **Refinement:** Identifying deviations allows for model refinement, such as adding or removing variables, applying transformations, or using alternative statistical methods that better handle non-normality.