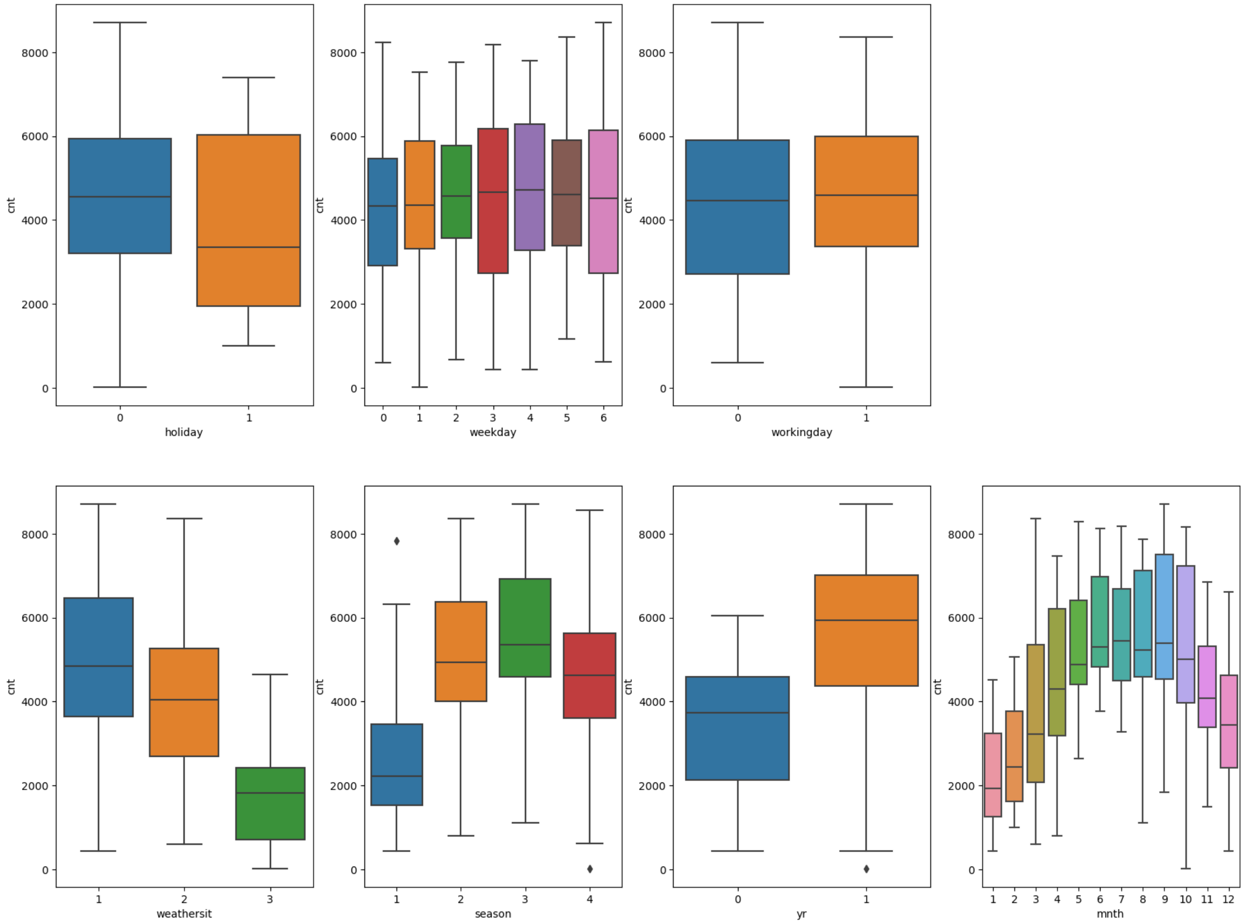
Assignment-based Subjective Questions

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 S eason, Yr, weekday, holiday, workingday, weathersit and mnth



1. The data's qualitative distribution is well-illustrated by the graph. These graphs increase our confidence in the model's predictions when paired with the important predictors that the model has discovered.It is clear that Category 3 (Fall) has the greatest median for the variable "season," indicating higher demand during this season compared to Category 1's (Spring) lowest demand.

2. By contrast, there were more users in 2019 than in 2018. The number of rentals is generally constant all during the week. When it rains or snows heavily, there is a discernible decline in rentals, indicating. These are especially bad weather conditions. The periods with the greatest rental counts are Clear and Cloudy skies with some clouds.

3. September was the greatest month for rentals, and December witnessed a decline, probably as a result of the month's usual heavy snowfall. In addition, fewer people utilize the site on vacations.

4. The workingday boxplot reveals that the majority of bookings take place between 4000 and 6000, and that the weekly median user count stays largely consistent. Bookings don't really change depending on whether it's a working day or not.

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

Drop_first=True is essential because it stops an additional column from being created during the encoding of dummy variables, which lowers the correlations between the dummy variables.

In this case, the dummy variables for a categorical variable with n levels must be represented by n−1 columns alone.

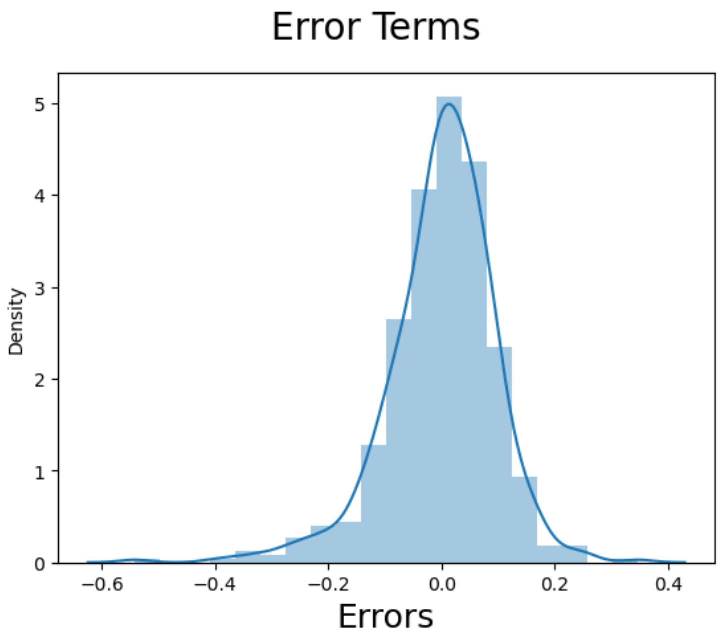
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The goal variable, cnt, has a strong correlation with the variables temp and atemp.

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

1. First, we always confirm that the independent and dependent variables have a linear relationship. We can utilize the pair plot that we completed to check.

2. Secondly, the mean is always zero and the residuals always follow a normal distribution. We used a residuals displot to verify this premise.



3. Little or no multicollinearity between the data is assumed by linear regression. When there is a strong correlation between the independent variables, multicollinearity occurs. We determined the Variance Inflation Factor (VIF), which gauges how closely the feature variables in the new model are related to one another, in order to determine the degree of this correlation.

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 fundamental algorithm in machine learning. Regression tasks are carried out by it, in which it forecasts a dependent (goal) value by using independent variables. This method is mostly applied to forecasting and the investigation of correlations between variables. The amount of independent variables and the kind of relationship between the dependent and independent variables that each regression model assumes varies.

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

The weights, or model coefficients, are denoted by β0 and β1. We need to find out these coefficients' values in order to build a model. We may use the model to forecast the target variable, like sales, once we know the values of these coefficients!

NOTE: Finding the line that best fits the data is the primary goal of the regression. The line with the lowest total prediction error (across all data points) is the best match. The distance from the points to the regression line is called the error.

2. Explain the Anscombe's quartet in detail.

The purpose of this quartet was to illustrate the significance of presenting data visually before to analysis and to show how outliers and other significant observations affect statistical measures.

- The top left scatter plot, the first one, shows a simple linear relationship.
- The second plot (upper right) displays a non-normal distribution with a clear but non-linear relationship.
- The third plot (bottom left) shows a linear distribution, although the regression line is impacted by an outlier. The results are greatly impacted by this outlier, which causes the correlation coefficient to drop from 1 to 0.816.
- Lastly, the fourth figure (bottom right) shows an instance in which there might be a single high-leverage point that generates a high correlation coefficient, even in the absence of any apparent relationship between the variables in the other data points.

3. What is Pearson's R?

The linear link between two variables is expressed as Pearson's r, or the Pearson correlation coefficient. It measures how strongly and in which a relationship exists between them.

This is a thorough 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:

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

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:

• N**on-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?

In data preparation, scaling is a procedure that modifies the feature value range to achieve comparability. It converts the data to a standard scale without warping the range of values' variations.

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 graphical tool called a Q-Q (Quantile-Quantile) plot is used to determine if a dataset adheres to a specific theoretical distribution, most frequently the normal distribution. It does this by comparing the dataset's quantiles to those of a theoretical distribution.

The Q-Q Plot's Significance in Linear Regression

A Q-Q plot is mostly used in the context of linear regression to verify the residuals' assumed normality. This is the reason it matters:

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 pvalues, 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.