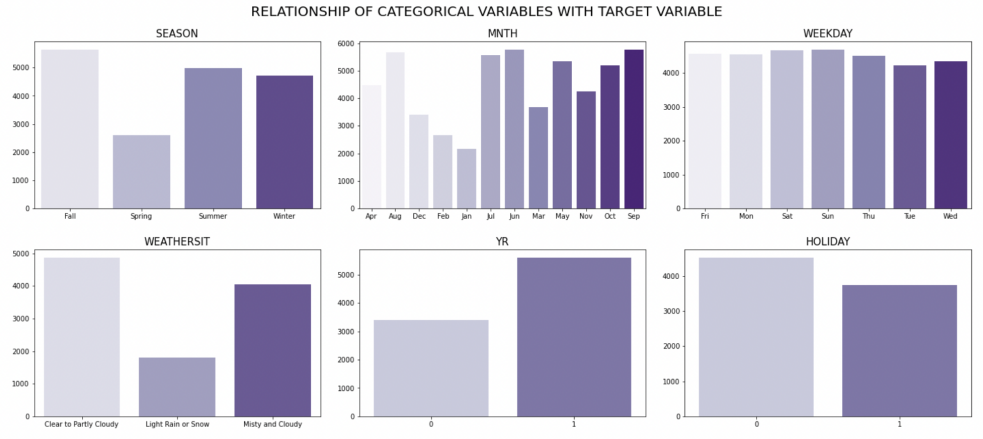
Assignment-based Subjective Questions

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



• The peak demand for bikes is evident during the Fall season, followed by Summer and Winter. In contrast, Spring experiences a notable decline in demand.

• Analysis of monthly trends reveals that August, June, and September exhibit the highest demand, with July, May, and October closely following suit. This establishes a robust six-month period, spanning from May to October, characterized by consistently high demand. The impact of months appears to have a significant influence on business dynamics.

• The demand for bikes demonstrates remarkable traction on Fridays, Saturdays, Sundays, and Thursdays. This trend suggests that bike usage is diversified for both work-related and leisure purposes.

• Unsurprisingly, days with clear weather conditions witness exceptionally high demand for bikes, indicating a positive correlation between clear days and increased usage.

• Notably, the demand for bikes on holidays appears to be slightly lower compared to weekdays, emphasizing a nuanced pattern in bike usage based on the day of the week.

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

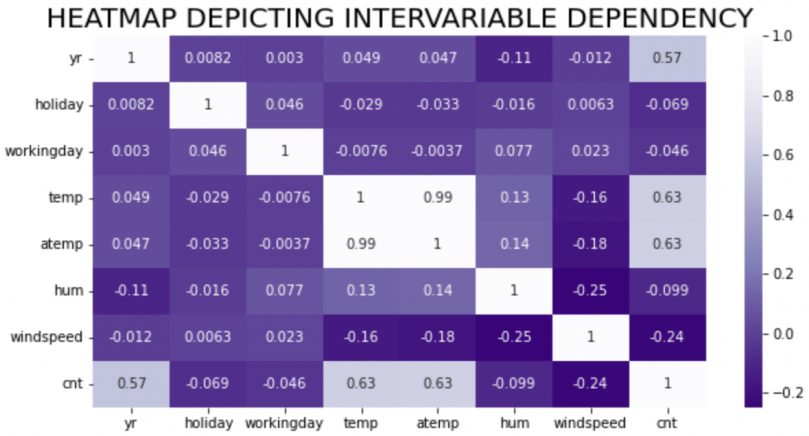
Using drop_first=True during dummy variable creation is important to prevent multicollinearity issues in regression models. When we create dummy variables for categorical features, we essentially convert categorical data into the numerical format, introducing a binary representation for each category.

For a categorical variable with n categories, creating n dummy variables would result in multicollinearity because the information in one category can be perfectly predicted from the others. This creates redundancy and can lead to problems in regression analysis.

By setting drop_first=True, we avoid this multicollinearity issue by excluding one of the dummy variables. For a categorical variable with n categories, we create only n−1 dummy variables. This ensures that each dummy variable is independent and provides unique information.

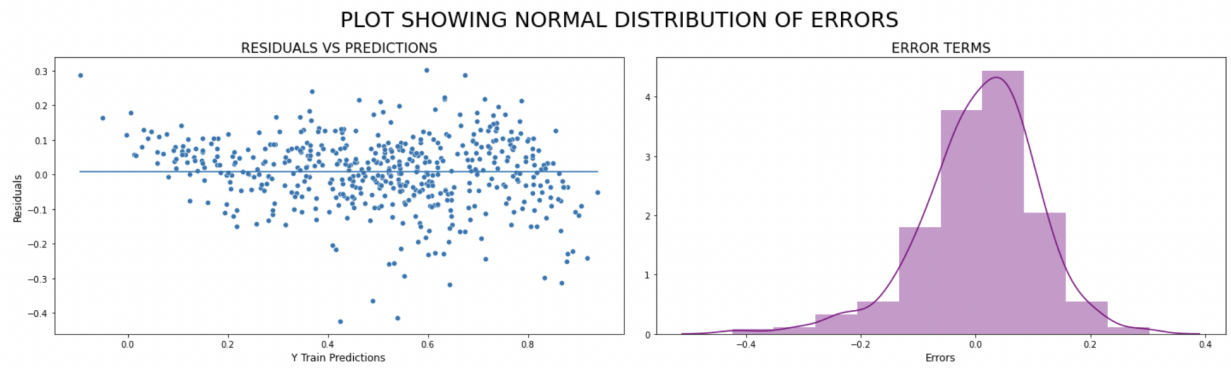
In the context of linear regression, multicollinearity can cause problems such as inflated standard errors, affecting the reliability of coefficient estimates. By dropping one dummy variable, we make the model more stable and improve its interpretability. It's a common practice to set drop_first=True to address multicollinearity when working with dummy variables in regression analysis.

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

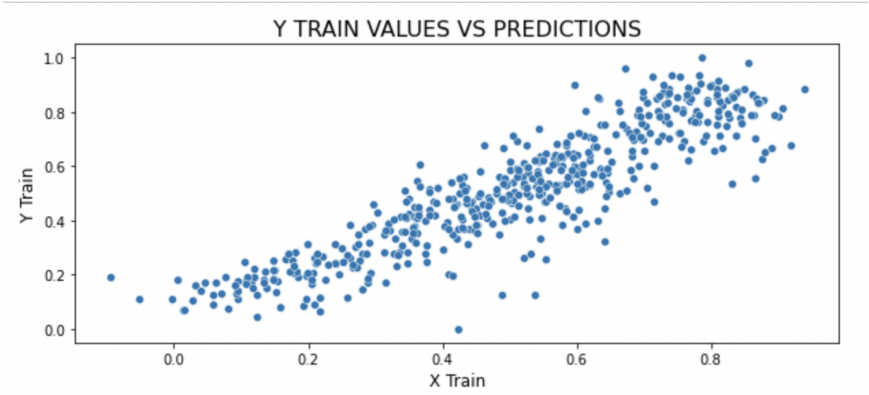


The numerical variables 'atemp' and 'temp' exhibit the strongest correlation with the target variable (cnt), both registering a substantial correlation coefficient of 0.63.

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



Mean of the residuals is extremely close to 0. For smaller values of predictions, the residuals are found to be on the higher side and focused above the mean value of 0. This validates our assumption of normal distribution errors around 0.



The plot shows an almost constant variance of predictions and thus the errors validating the assumptions of homoscedasticity.
There are no correlations between error terms. Hence, one more assumption of linear regression has been proved.

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

The most important factor affecting demand is temperature. With a coefficient of 0.73126, for every change in temperature of 1 degree, demand increases by a factor of 0.73126 (temperature x 0.73126).

The second most important factor is Light Rain or Snow with a coefficient of -0.27750. Hence, if a particular day has light rains, it is expected to reduce the demand by 27.7%.

The third most important factor is year with a coefficient value of 0.24236. Based on the historical data, given all internal and external factors remain unchanged, the company is expected to see annual growth over last year at around 24%.

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning algorithm used for predicting a continuous outcome variable (also known as the dependent variable) based on one or more predictor variables (independent variables). The fundamental idea behind linear regression is to model the relationship between the independent variables and the dependent variable as a linear equation.

The basic form of a linear regression equation for a single independent variable is: $y=mx+b$, where:

- y is the dependent variable (the variable to be predicted),
- x is the independent variable,
- m is the slope of the line (coefficient),
- b is the y-intercept.

The goal of linear regression is to find the best-fitting line that minimizes the sum of squared differences between the predicted and actual values.

The model is trained by minimizing a cost function, typically the sum of squared differences between predicted and actual values. The Ordinary Least Squares (OLS) method is commonly used for this purpose.

The coefficients ($b0,b1,...,bn$) are estimated during training to minimize the cost function.

During the training phase, the algorithm iteratively adjusts the coefficients to minimize the cost function using optimization techniques like gradient descent.

Assumptions of Linear Regression:

- Linear regression assumes a linear relationship between the independent and dependent variables.
- It assumes that the residuals (differences between predicted and actual values) are normally distributed and have constant variance.

Once trained, the linear regression model can be used to make predictions on new, unseen data. The performance of the model is often assessed using metrics like Mean Squared Error (MSE), R-squared, or other relevant regression metrics.

Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. The quartet comprises four distinct datasets, each consisting of 11 data points paired in x and y coordinates.

Here are the details of Anscombe's quartet:

Dataset I:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset II:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Dataset III:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset IV:

- x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
- y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Key Observations:

Descriptive Statistics:

- All four datasets have nearly identical simple descriptive statistics, including means, variances, and correlation coefficients.

Graphical Representation:

- When plotted, Dataset I forms a typical linear relationship.
- Dataset II exhibits a non-linear relationship.
- Dataset III shows an outlier that influences the linear regression line.

- Dataset IV demonstrates the impact of a single outlier on correlation and regression.

Importance of Graphical Exploration:

- Anscombe's quartet underscores the importance of visually exploring data to understand patterns, relationships, and potential outliers.
- Descriptive statistics alone may not reveal the nuances present in the data.

Statistical Insights:

- The quartet challenges the assumption that datasets with similar summary statistics will have similar properties.
- It highlights the limitations of relying solely on numerical summaries without visualizing the data.

In conclusion, Anscombe's quartet serves as a powerful reminder of the necessity of graphical exploration in data analysis and the potential pitfalls of relying solely on summary statistics. It emphasizes the value of visualization in uncovering hidden patterns and outliers that may significantly impact the interpretation of data.

What is Pearson's R?

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. The formula for Pearson's correlation coefficient (r) between two variables, X and Y, with n data points, is given by:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i are the individual data points.
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.

- Pearson's r ranges from -1 to 1.
- r=1 indicates a perfect positive linear relationship.

- $r=-1$ indicates a perfect negative linear relationship.
- $r=0$ indicates no linear relationship.
- The magnitude of r indicates the strength of the linear relationship. The closer $|r|$ is to 1, the stronger the linear relationship.
- Pearson's r is sensitive to outliers, and a single outlier can significantly impact the correlation coefficient.

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

Scaling is a preprocessing technique in data analysis and machine learning that involves transforming the values of variables to a specific range. The goal of scaling is to bring all variables to a comparable magnitude, which can be particularly important in algorithms sensitive to the scale of input features. The most common scaling methods include normalization and standardization.

Gradient Descent-based optimization algorithms converge faster when variables are on a similar scale. This is crucial for algorithms used in linear regression, logistic regression, and neural networks.

Scaling facilitates better interpretation of coefficients in linear models. Without scaling, the coefficients may not accurately reflect the variable's impact if the variables are on different scales.

Regularization techniques, like Ridge or Lasso regression, are sensitive to the scale of variables. Scaling ensures that regularization is applied uniformly across all features.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

- Range: Normalization scales the values of a variable to a specific range, typically between 0 and 1.
- Advantages:

        - Intuitive scaling to a specific range.
        - Preserves the relative differences between values.
    - Considerations:
        - Sensitive to outliers.

Standardized Scaling (Z-Score Scaling):

- Range: Standardization scales the values of a variable to have a mean of 0 and a standard deviation of 1.
- Advantages:
    - Less sensitive to outliers compared to normalization.
    - Preserves the shape of the distribution.
- Considerations:
    - Values are not constrained to a specific range.

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

- A predictor variable can be expressed as a perfect linear combination of other predictor variables.
- In the context of dummy variables, if one dummy variable is a perfect linear combination of others, it can lead to multicollinearity issues and result in infinite VIF.
- If a predictor variable has zero variance (i.e., all its values are the same), it can also result in an infinite VIF.

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 in statistics to assess whether a dataset follows a particular theoretical distribution. In the context of linear regression, Q-Q plots are often employed to examine whether the residuals (the differences between observed and predicted values) conform to the assumptions of a normal distribution.

- One of the key assumptions in linear regression is that the residuals are normally distributed. The Q-Q plot provides a visual check to assess the normality of the residuals.
- If the points on the Q-Q plot deviate significantly from a straight line, it suggests departures from normality. Common deviations include skewness or heavy tails.
- Q-Q plots are part of the diagnostic tools used to evaluate the goodness of fit of a linear regression model. Deviations from normality may suggest that the model assumptions are not fully met.
- The Q-Q plot is particularly useful for examining the distribution of residuals. If residuals are normally distributed, it implies that the model is well-specified and captures the underlying patterns in the data.
-