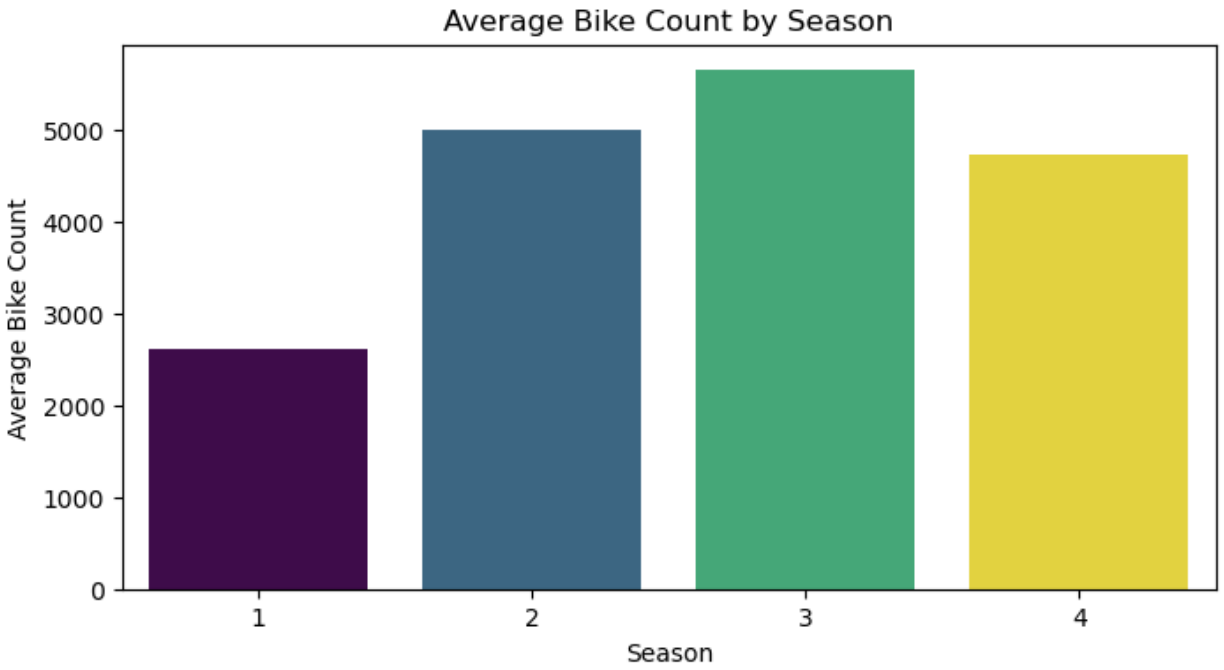
Linear Regression Subjective Questions

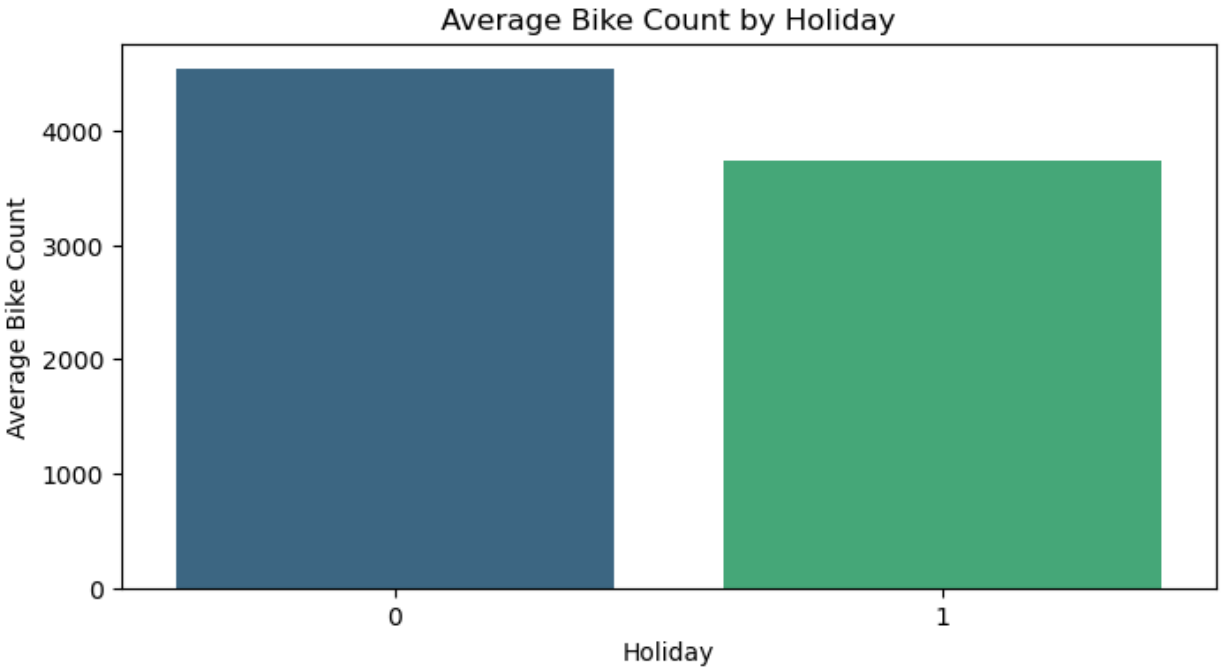
thang.truong

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



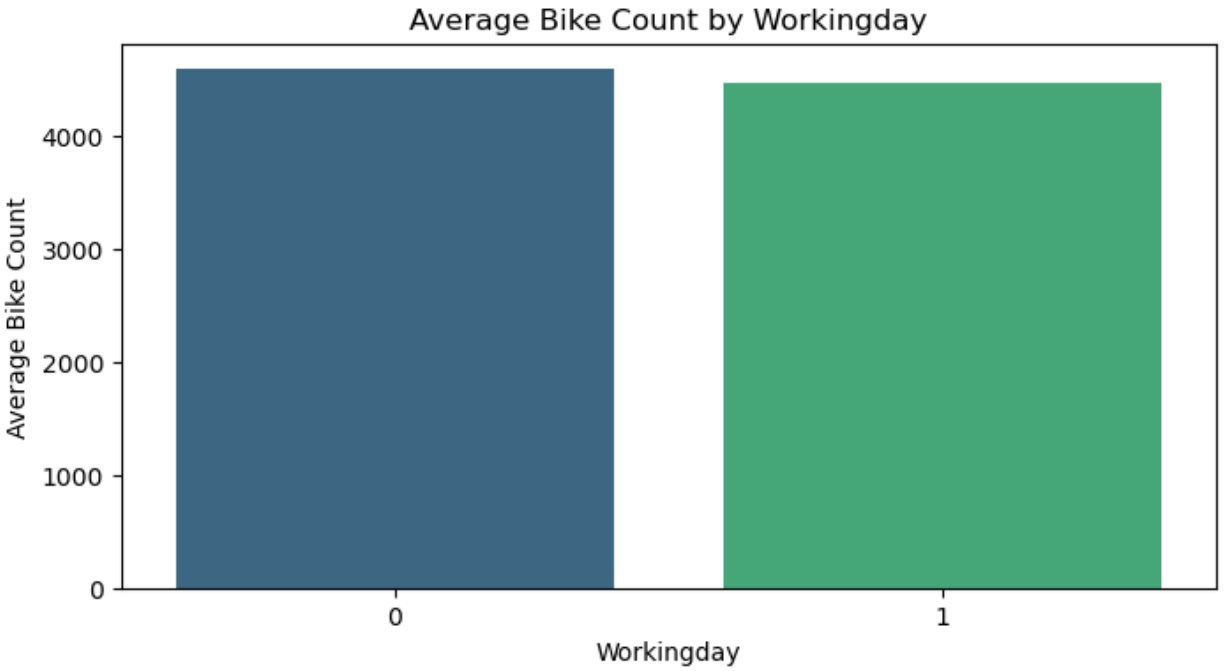
When we look at the "Average Bike Count by Season" plot, we can observe that:

- **Code 1 (Spring):** The average bike count is lower, possibly because the weather is still quite variable in early spring.
- **Code 2 (Summer):** The bike rentals increase, indicating that summer is a favorable time for biking.
- **Code 3 (Autumn):** The bike rentals reach the highest level, which could be due to the mild and pleasant weather in autumn.
- **Code 4 (Winter):** The bike rentals decrease, partly due to the cold and potentially adverse weather conditions.



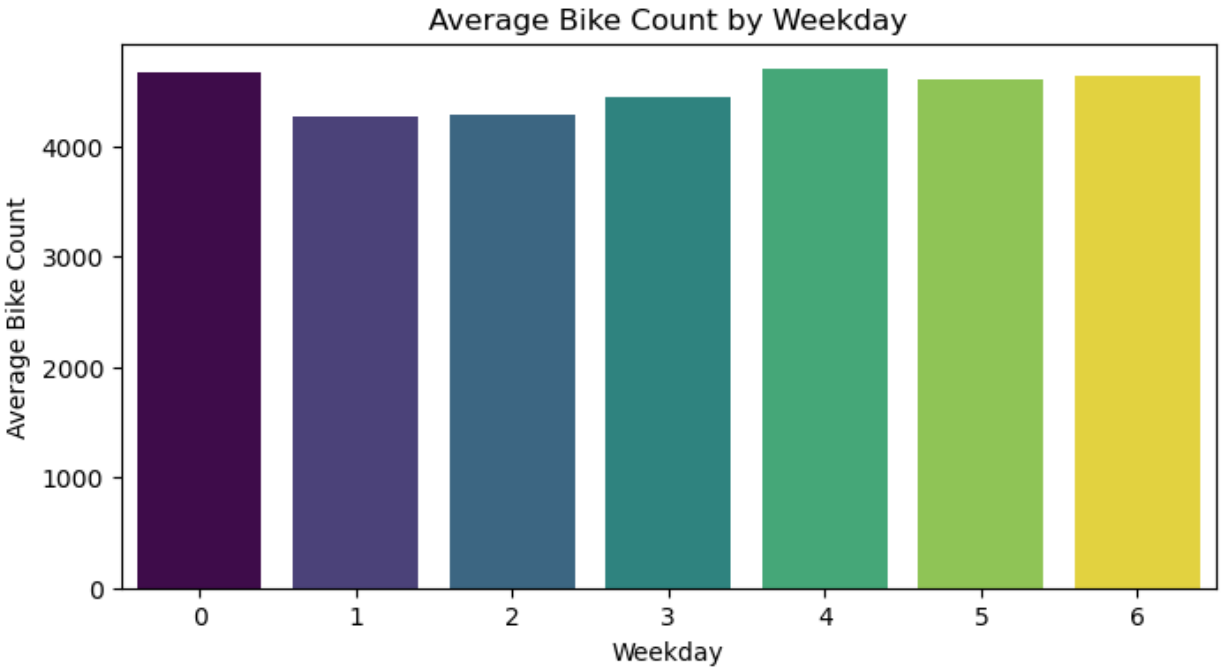
The bar chart titled **"Average Bike Count by Holiday"** shows a visual comparison between the average bike rentals on days marked as non-holidays (Holiday = 0) and holidays (Holiday = 1). Here's what we can infer:

- **Non-Holiday (0):** The bar for non-holidays is around 4500 bike rentals on average. This suggests that on common, non-holiday days—which usually correspond to regular commuting days—the demand for bike-sharing is relatively high.
- **Holiday (1):** In contrast, the holiday bar is noticeably lower, around 3500 bike rentals on average. This indicates that bike usage tends to drop on holidays, possibly because factors such as reduced commuting traffic and alternative leisure activities are more prevalent on these days.



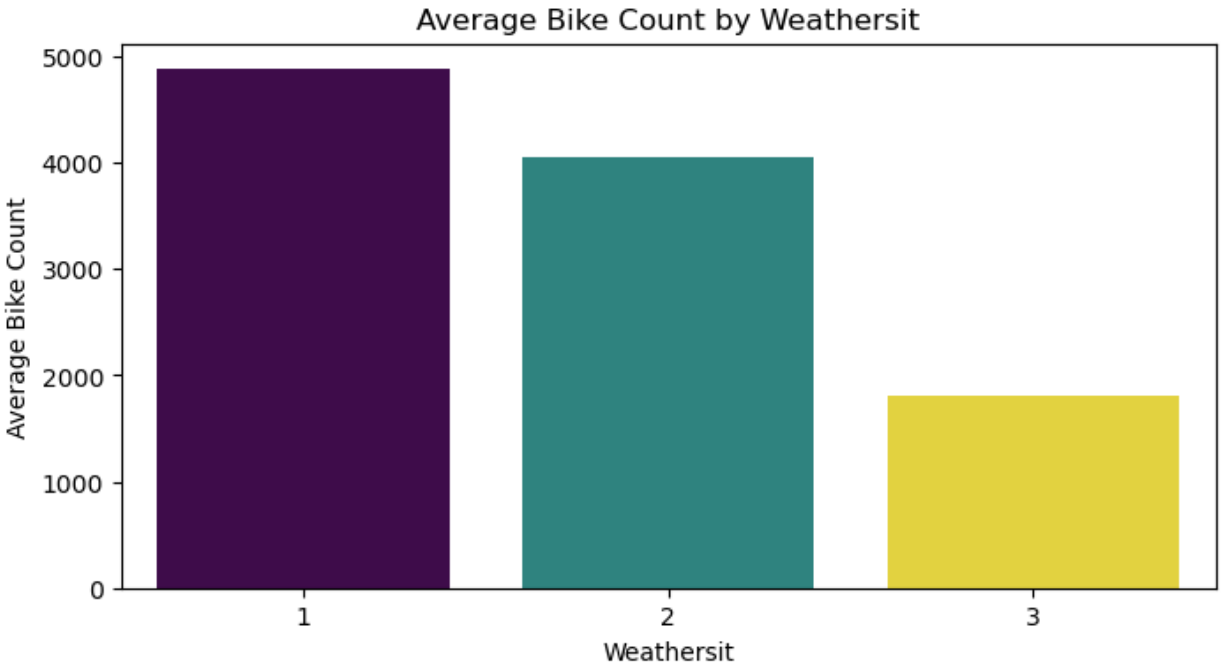
Based on the chart titled **"Average Bike Count by Workingday"**, we can observe the following:

- **Workingday = 0 (Non-working Day):** The chart shows that the average number of bike rentals for this group is higher compared to working days. This suggests that on non-working days—possibly weekends or holidays—users tend to use bikes more frequently, possibly because they have extra time for recreational activities or enjoy spending time outdoors.
- **Workingday = 1 (Working Day):** Conversely, the average number of bike rentals on working days is lower. This may be because people who work tend to prefer quicker and more convenient modes of transportation, or due to time constraints in their busy schedules.



The bar chart titled **"Average Bike Count by Weekday"** shows how bike rental activity varies across the seven days of the week, with weekdays labeled from 0 to 6. Here's what we can infer from the chart:

- **Day 0 and Day 4 (≈4500 rentals):** These days exhibit the highest average bike counts. This suggests that on these days—whatever their exact correspondence might be (for instance, if 0 represents Sunday or Monday, and 4 represents Thursday or Friday)—bike usage is at its peak. It could be due to either high commuter demand or popular leisure rides, depending on the local context and how the days are mapped.
- **Day 1 and Day 2 (≈4000 rentals):** These days show the lowest averages, indicating a dip in bike-sharing activity. This drop might point to fewer commuting trips on these days, or it could simply be a result of other factors such as less favorable conditions, lower leisure activity, or varying work schedules.
- **Intermediate Days (Day 3, Day 5, Day 6 – 4200 to 4300 rentals):** The middle-range values for these days suggest that bike rentals are moderate—neither at a peak nor at a trough—which might reflect transitional usage patterns within the week.



Based on provided definitions, the weather situation variable (weathersit) is encoded as follows:

- **1: Clear, Few clouds, Partly cloudy** Conditions indicating clear weather with minimal cloud cover.
- **2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist** Conditions that include various combinations of mist and cloudiness.
- **3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** Conditions that involve light snowfall or light rain, often accompanied by thunderstorms and scattered clouds.
- **4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog** Conditions depicting severe weather such as heavy rain with ice pellets, thunderstorms, additional mist, or combinations of snow and fog.

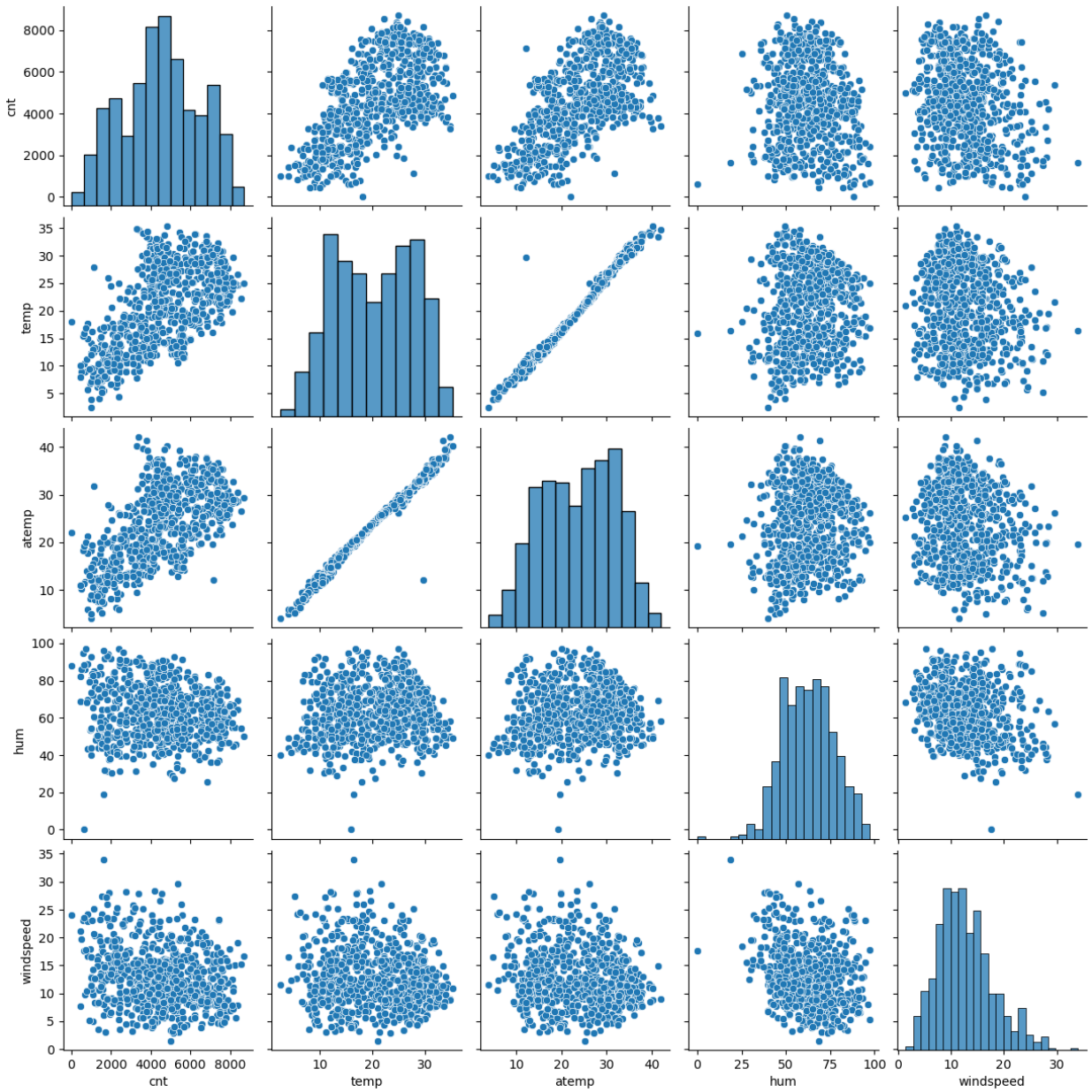
Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True when creating dummy variables is important for two primary reasons:

- **Avoiding Multicollinearity (Dummy Variable Trap):** When you create dummy variables for a categorical feature with k categories without dropping one, the resulting dummy variables are linearly dependent (they add up to one). This perfect collinearity, known as the dummy variable trap, can cause issues in regression models where the design matrix becomes singular. By setting drop_first=True, you remove one dummy variable, thereby eliminating redundancy.
- **Establishing a Baseline for Comparison:** Dropping the first category establishes a baseline (reference group) against which the effects of the remaining categories are measured. This simplifies the interpretation of the regression coefficients as each coefficient now represents the effect relative to the omitted category.

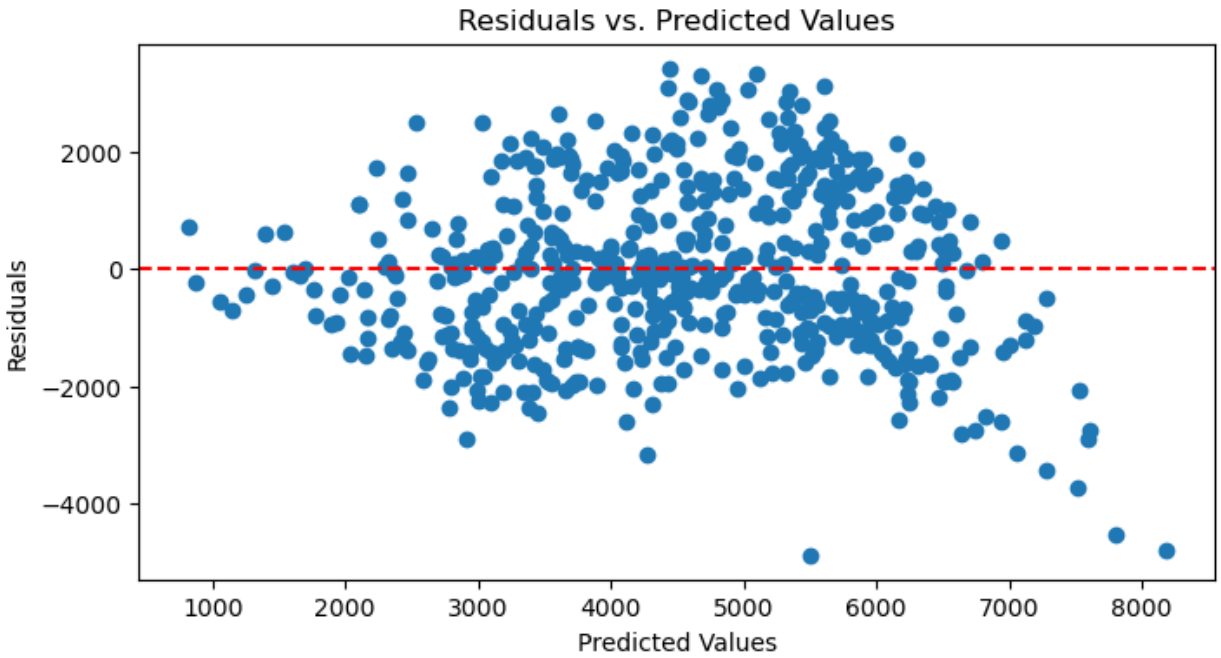
For instance, in the bike-sharing dataset, if you create dummy variables for the season variable without dropping one dummy, you may end up with perfect multicollinearity. Using drop_first=True ensures model remains stable and the resulting coefficients are interpretable.

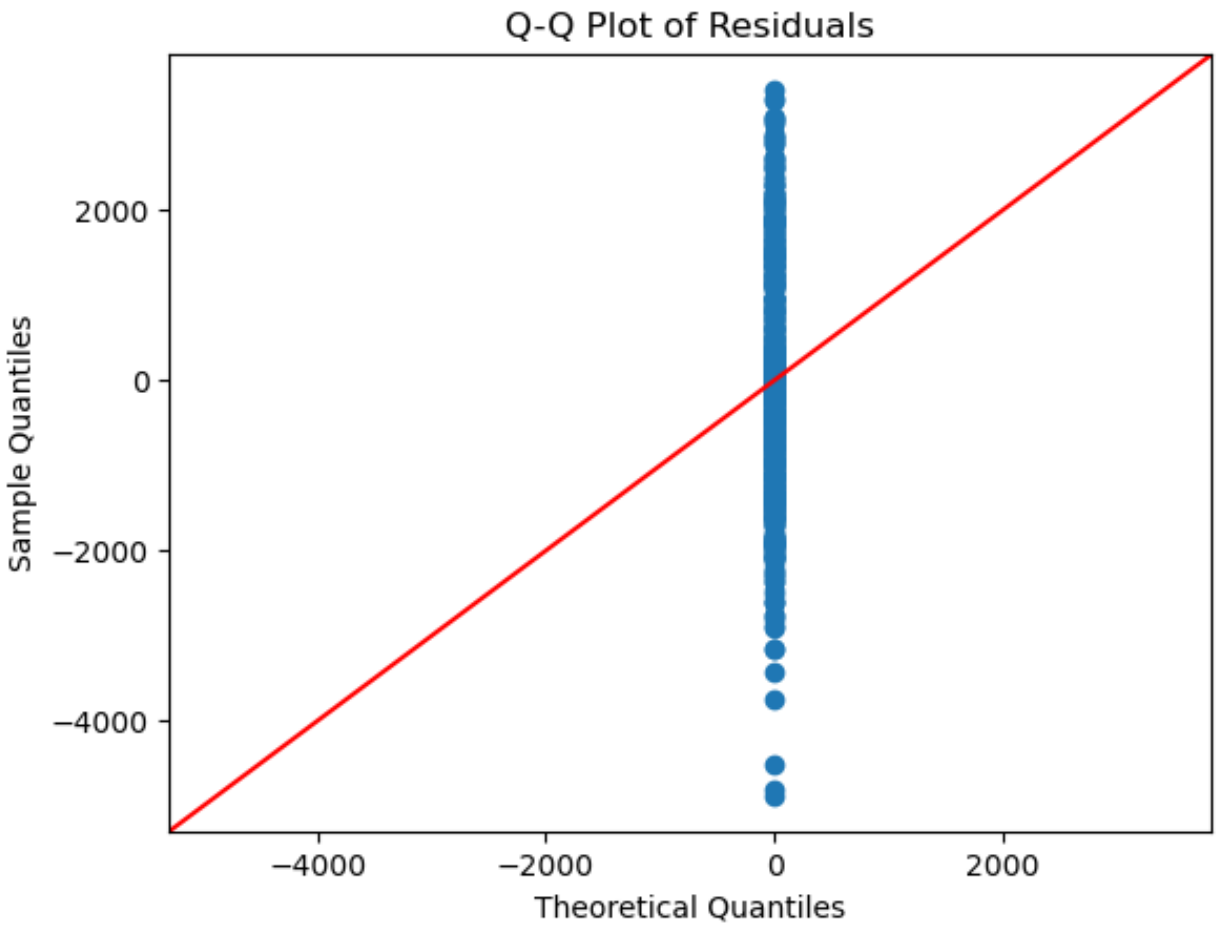
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



Based on the pair-plot among the numerical variables, it appears that the **temperature variable** (temp) has the highest correlation with the **target variable** (cnt).

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





1. **Linearity:** Check scatter plots of predictors vs. target & residuals vs. fitted values to ensure no systematic pattern.
2. **Homoscedasticity:** Inspect residuals vs. predicted plot and apply the Breusch-Pagan test for constant variance.
3. **Normality:** Use Q-Q plots and histograms of residuals (optionally Shapiro-Wilk) to verify a normal distribution.
4. **Independence:** Compute the Durbin-Watson statistic to check for uncorrelated residuals.
5. **Multicollinearity:** Evaluate VIF for each predictor to detect high correlations and ensure stable coefficient estimates.

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

Based on the final linear regression model, the top three features significantly contributing to explaining the demand for shared bikes are:

- **Temperature (temp):** This variable shows the strongest positive impact as higher temperatures generally lead to increased bike usage.
- **Feels-like Temperature (atemp):** Despite its similarity to temp, it captures the perceived temperature effect, which further drives the number of rentals.
- **Humidity (hum):** This feature is significant in a negative manner—indicating that higher humidity tends to reduce the demand for bike sharing.

Explain the linear regression algorithm in detail.

1. Model Structure

At its core, linear regression assumes that the target variable y can be expressed as a linear combination of the predictor variables x 1 , x 2 , … , x p plus an error term ε .

Mathematically, the model is written as:

$$y = \beta\ 0 + \beta\ 1\ x\ 1 + \beta\ 2\ x\ 2 + \ldots + \beta\ p\ x\ p + \varepsilon$$

- β 0 : The intercept represents the estimated average value of y when all x values are zero.
- β 1 , β 2 , … , β p : Coefficients that quantify the change in y for a unit change in the respective predictor variable.
- ε : The error term, capturing the variation in y that cannot be explained by the predictors.

For the bike-sharing dataset, for instance, the predictors might include weather conditions or time-related features, and the target could be the demand (i.e., bike counts, labeled as cnt).

2. Objective and Loss Function

The goal of linear regression is to estimate the coefficients β such that the predicted values y ^ are as close as possible to the true values y . This is typically achieved by minimizing the **Residual Sum of Squares (RSS):**

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}i)^2 = \sum i = 1^n (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

Minimizing the RSS ensures that the overall error in prediction is as small as possible.

3. Coefficient Estimation

Ordinary Least Squares (OLS):

- The most common method for estimating the coefficients is the **Ordinary Least Squares (OLS)**. With a data matrix XX (including a column of ones for the intercept) and a target vector yy, the OLS solution is given by:

$$\hat{\beta} = (XTX)-1X^T y$$

This closed-form solution finds the set of coefficients that minimize the RSS. It is computationally efficient and works well when the number of predictors is not extremely high.

Gradient Descent:

For very large datasets or when the closed-form solution is computationally expensive (e.g., due to a high number of features), **gradient descent** is used. This iterative optimization method updates the coefficients step-by-step according to the gradient of the loss function:

$$\beta_j(t+1) = \beta_j(t) - \alpha \frac{\partial \text{RSS}}{\partial \beta_j}$$

where α is the learning rate. This process continues until convergence, i.e., when changes in RSS become negligibly small.

4. Assumptions of Linear Regression

For the linear regression model to yield valid inferences and predictions, several key assumptions must be met:

- **Linearity:** The relationship between the predictors and the target is assumed to be linear. This implies that adding the contributions of the predictors accurately represents the effect on yy.
- **Independence of Errors:** The residuals (errors) are assumed to be independent. This is critical in ensuring that observations (especially in time series data) do not influence each other unduly.
- **Homoscedasticity:** The variance of the errors should remain constant across all levels of the independent variables. Non-constant variance (heteroscedasticity) can lead to inefficient estimates and affect hypothesis tests.
- **Normality of Errors:** In many applications—especially when constructing confidence intervals or conducting significance tests—the residuals are assumed to be Gaussian (normally distributed).
- **No Multicollinearity:** The predictors should not be too highly correlated with each other. High multicollinearity can make it challenging to estimate individual coefficients accurately, resulting in large standard errors.
- In practice, after fitting a model (for example, on the bike-sharing dataset), diagnostic tools such as residual plots, Q-Q plots, the Breusch-Pagan test (for homoscedasticity), Durbin-Watson test (for independence), and Variance Inflation Factor (VIF) calculations are used to validate these assumptions.

5. Model Evaluation and Interpretation

Once the coefficients are estimated:

- **R-squared (R^2):** This metric indicates the proportion of variability in the target variable explained by the model. A higher R2R^2 (closer to 1) means a better fit.
- **Significance Tests:** t-tests on the individual regression coefficients help determine if the predictors have a statistically significant relationship with the target variable.
- **Residual Analysis:** Examining residuals (the differences between observed and predicted values) not only helps validate model assumptions but also points out potential issues like outliers or model misspecification.
- **Interpretability:** One of linear regression's strengths is its interpretability. Each coefficient $\beta\ j$ indicates the expected change in y for a one-unit change in $x\ j$, keeping other variables constant. This clarity is particularly useful for understanding the impact of different factors—such as weather conditions or temporal factors—in the bike-sharing dataset.

Explain the Anscombe's quartet in detail.

1. Identical Summary Statistics:

- Each dataset in the quartet has the same mean of the independent variable x (usually around 9) and the same mean of the dependent variable y (approximately 7.5).
- They share nearly identical variances, correlations between x and y , and even the same linear regression line (with similar intercepts and slopes). In fact, if you compute the regression equation for each dataset, you might get something like

$$\hat{y} = 3 + 0.5x$$

1. Identical Summary Statistics:

- Each dataset in the quartet has the same mean of the independent variable x (usually around 9) and the same mean of the dependent variable y (approximately 7.5).
- They share nearly identical variances, correlations between x and y , and even the same linear regression line (with similar intercepts and slopes). In fact, if you compute the regression equation for each dataset, you might get something like

2. Diverse Data Patterns:

- **Dataset I:** Shows a typical linear relationship with data points scattered reasonably evenly around the fitted regression line, reflecting what one might expect from a "normal" linear association.
- **Dataset II:** Despite matching in summary statistics, this dataset reveals a curved pattern when plotted, indicating a non-linear relationship. The linear regression line still fits formally because the overall averages and variances are the same, but the true nature of the relationship is obscured.
- **Dataset III:** Contains a largely linear pattern with one or more influential outliers. These outliers significantly affect the regression line even though the computed summary statistics remain unchanged.
- **Dataset IV:** Almost all x values are nearly the same (except for one influential observation). This creates the illusion of a relationship when, in reality, the vast majority of the data offer little information about the linear trend. The outlier forces the regression line to appear as if there were a linear association.

3. Key Lessons and Importance:

- **Visualization Matters:** Anscombe's quartet spectacularly demonstrates that relying solely on summary statistics such as means, variances, and regression coefficients can be misleading. Visualizing data using scatter plots or other graphical tools is essential to uncover hidden patterns, outliers, or non-linear relationships.
- **Model Validation:** It serves as a cautionary tale when validating models. Even if the statistical metrics suggest a "good" fit, the underlying data structure might tell a very different story. This example is often used to encourage analysts and data scientists to perform exploratory data analysis (EDA) before settling on a model.
- **Robustness of Statistical Inference:** The quartet shows that having similar statistical summaries does not ensure that the data behave in similar ways. This reinforces the need for robust diagnostics and understanding the context of the data.

4. Historical Context:

- Created by Francis Anscombe in 1973, the quartet was originally designed to illustrate the importance of graphical display in data analysis and to caution against taking summary statistics at face value. It remains a classic educational example in statistics courses and texts.

What is Pearson's R?

Pearson's R is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Here are the key details:

- **Definition:** Pearson's R, also known as the Pearson correlation coefficient, is computed as the covariance of the two variables divided by the product of their standard deviations. In formula form, it is expressed as:

$$r = covariance(X, Y) / (\sigma_x \cdot \sigma_y)$$

or equivalently:

$$r = [\Sigma(x_i - mean(X)) \cdot (y_i - mean(Y))] / [sqrt(\Sigma(x_i - mean(X))^2) \cdot sqrt(\Sigma(y_i - mean(Y))^2)]$$

- **Range & Interpretation:** The value of r ranges from -1 to 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.

  The closer |r| is to 1, the stronger the linear relationship between the variables.

- **Usage & Considerations:** Pearson's R is widely used in data analysis and machine learning to assess how well two variables are linearly related. However, it assumes:
    - The relationship between the variables is linear.
    - The data are approximately normally distributed.
    - The observations are independent.

  It is sensitive to outliers, which can significantly affect the value of r.

In the context of the bike-sharing dataset provided in the assignment, Pearson’s R might be used to measure, for example, the correlation between temperature and bike rentals, helping to understand whether higher temperatures correspond to increased bike-sharing demand.

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

Why is Scaling Performed?

- **Algorithm Performance:** Many algorithms—such as gradient descent based methods, k-nearest neighbors, or support vector machines—are sensitive to the scale of the features. If the features have widely different ranges, those with a larger scale may unduly influence the model.
- **Faster Convergence:** Scaling can lead to faster convergence during optimization because it makes the cost function more symmetrical and reduces the risk of the algorithm oscillating when updating weights.
- **Improved Interpretability and Comparability:** When all features are on a similar scale, the coefficients of models like linear regression become more interpretable, and it's easier to understand the relative impact of each feature.

Differences Between Normalized Scaling and Standardized Scaling

1. Normalized Scaling (Min-Max Scaling):

- **Process:** This approach rescales the features to a fixed range, typically [0, 1] (or sometimes [-1, 1]). The transformation is given by:

$$xnorm = x - min(x)max(x) - min(x)x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Characteristics:**
  - The transformed feature values are bounded within the specified range.
  - It is sensitive to outliers since the range max (x) − min (x) can be disproportionately affected by extreme values.
  - Commonly used when a specific range is required by an algorithm (e.g., neural networks).

2. Standardized Scaling (Z-score Normalization):

- **Process:** Standardization transforms features so that they have a mean of 0 and a standard deviation of 1. The formula is:

$$xstd = \frac{x - \mu}{\sigma}$$

where μ\mu is the mean of the feature and σ\sigma is its standard deviation.

- **Characteristics:**
  - This method centers the data and makes the variance uniform.
  - It is less sensitive to outliers compared to normalization.
  - It is preferred when the distribution of the data is

2. **Standardized Scaling (Z-score Normalization):**
   - **Process:** Standardization transforms features so that they have a mean of 0 and a standard deviation of 1. The formula is:

$$xstd = \frac{x - \mu}{\sigma}$$

     where μ is the mean of the feature and σ is its standard deviation.
   - **Characteristics:**
     - This method centers the data and makes the variance uniform.
     - It is less sensitive to outliers compared to normalization.
     - It is preferred when the distribution of the data is approximately normal, and many statistical models assume zero-centered and unit variance features.

Context: Bike-Sharing Dataset

In the context of the bike-sharing dataset, scaling might be applied to features like temperature, humidity, and windspeed. If you normalize these variables, you bring them to a common range (e.g., [0, 1]), ensuring that one feature doesn't dominate the others simply due to its scale. Alternatively, standardizing them ensures that each feature contributes equally to the analysis by setting their means to zero and variances to one—making the model more robust, especially if using regularization or algorithms sensitive to feature variances.

- In summary, **scaling** is a crucial pre-processing step to harmonize feature ranges, thereby improving model training and performance. **Normalization** confines features to a specific range, while **standardization** centers the features to have a mean of 0 and a standard deviation of 1, each serving slightly different purposes based on the characteristics of data and the requirements of chosen algorithm.

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

The Variance Inflation Factor (VIF) for a predictor is computed as:

$$VIF = \frac{1}{1 - R^2}$$

where R 2 is the coefficient of determination when the predictor is regressed on all the other predictors.

An infinite VIF occurs when R 2 = 1 . This means the predictor is perfectly collinear with one or more other predictors in model—there is an exact linear relationship. When there's perfect multicollinearity, the predictor's values can be completely determined by the other variables, causing the denominator 1 − R 2 to become zero and thus making the VIF infinite.

This situation often happens if you accidentally include redundant features. For example, it can occur when using dummy variables without dropping one category (leading to the dummy variable trap), or when certain transformations or combinations of variables result in exact linear dependencies.

In summary, an infinite VIF is a red flag that indicates severe multicollinearity, implying that one or more predictors provide duplicate information to the model.

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

A Q-Q plot—short for quantile-quantile plot—is a graphical tool used to compare the distribution of data (or, in the case of linear regression, the residuals) to a theoretical distribution, usually the normal distribution.

1. **How It Works:**
   - A Q-Q plot plots the quantiles of sample data against the quantiles of the theoretical distribution (e.g., normal distribution).
   - If the residuals (or data) follow the theoretical distribution, the points will lie approximately along a straight diagonal line.

2. **Use in Linear Regression:**
   - **Assessing Normality of Residuals:** One of the key assumptions in linear regression is that the residuals (errors) are normally distributed. By creating a Q-Q plot of the residuals, you can visually inspect whether they deviate substantially from normality.
   - **Detecting Outliers and Heavy Tails:** Deviations from the straight line in a Q-Q plot can indicate the presence of outliers, skewness, or heavy tails. This information helps diagnose potential problems with the model assumptions.
   - **Model Validation:** Ensuring that the residuals are normal is crucial for the validity of inferential statistics in regression (such as confidence intervals and hypothesis tests). If the residuals are not normal, the estimates of variability and p-values may be unreliable.

3. **Importance:**
   - **Verifying Assumptions:** A Q-Q plot is a simple yet powerful diagnostic to check the crucial assumption of normality in the error terms.
   - **Guiding Further Analysis:** If the Q-Q plot shows significant deviations from the line, it may signal that a transformation of the dependent variable or the use of a different modeling approach is needed.

In summary, the Q-Q plot is an essential tool in the diagnostic process for linear regression. It helps ensure that the residuals approximate a normal distribution, thereby reinforcing the validity of the model's statistical inference and guiding any necessary adjustments.