**Q1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

A- From my analysis of categorical variables, we can infer following:

- We have more cnt in clear whether as compared to otherwise.
- Day of the week doesn't seem to have much effect on cnt
- We have relatively more cnt in month May - Oct as compared with otherwise.
- We have relatively more cnt in year 2019 as compared to 2018.
- We have relatively equal cnt on winter season as compared to otherwise.
- We have more cnt in summer season as compared to otherwise.
- We have less cnt in spring season as compared to otherwise.
- We have relatively equal cnt on working days and non-working day which seems counter intuitive
- We have more cnt on non-holiday days as compared to holiday days.

**Q2 - Why is it important to use drop_first=True during dummy variable creation?**

A- Using drop_first=True when creating dummy variables for a multi-variable linear regression is important to avoid the issue of multicollinearity.

Multicollinearity occurs when independent variables in a regression model are highly correlated, which can distort the results and make them unreliable.

Example : you have a categorical variable 'Season' with four categories: Spring, Summer, Fall, and Winter. When you create dummy variables, you get four new columns: one for each season. However, knowing the values of three (say, Spring, Summer, Fall) inherently tells you the value of the fourth (Winter). If Spring, Summer, and Fall are all 0, Winter must be 1.

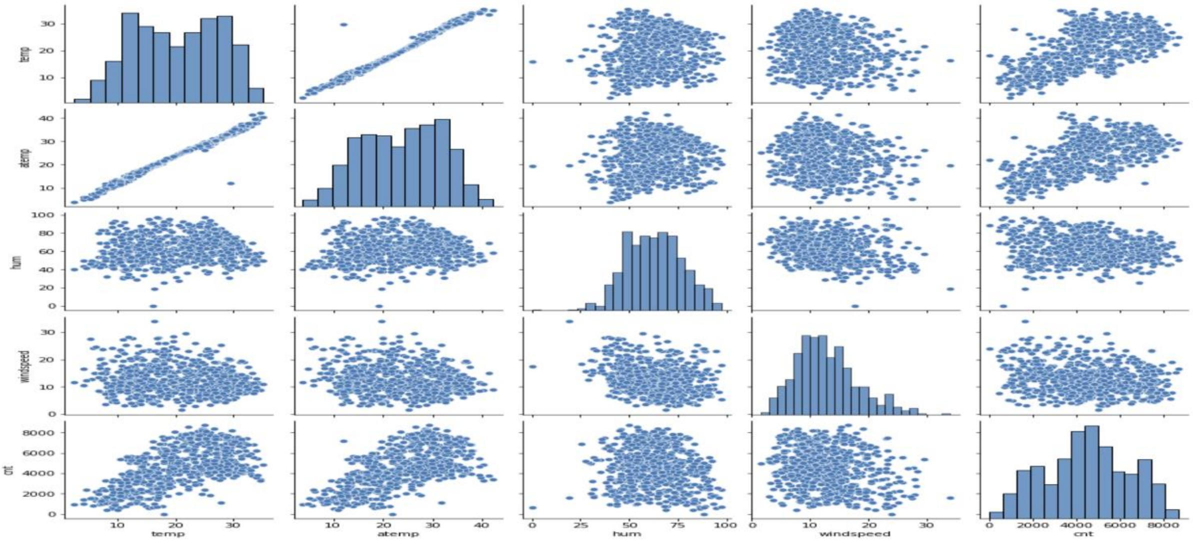
By doing drop_first=True, you drop the first dummy variable (let's say, Spring). Now, if Summer, Fall, and Winter are all 0, it implies that it's Spring. This approach reduces multicollinearity and still retains all the information about the seasons in your model. This simplification helps in making the model more stable and interpretable.

**Q3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

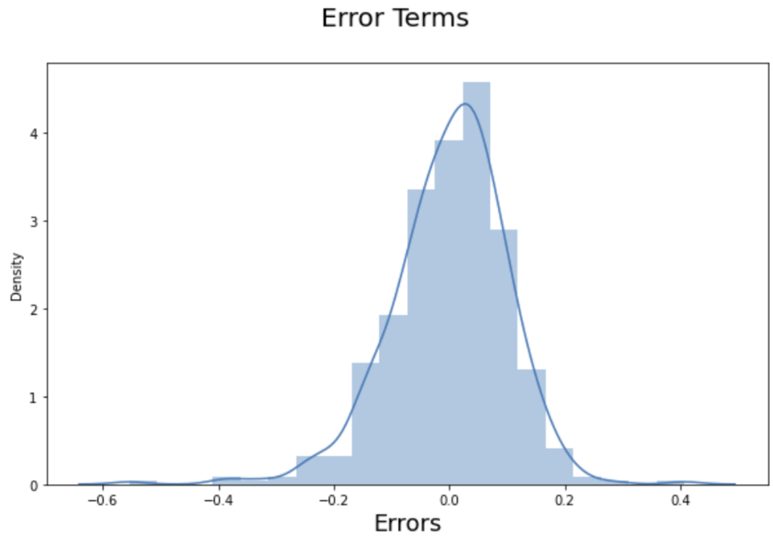
A- Both “Temp” and “atemp” have 0.63 correlation coeff which is highest amongst numerical variables.

**Q4 - How did you validate the assumptions of Linear Regression after building the model on the training set?**

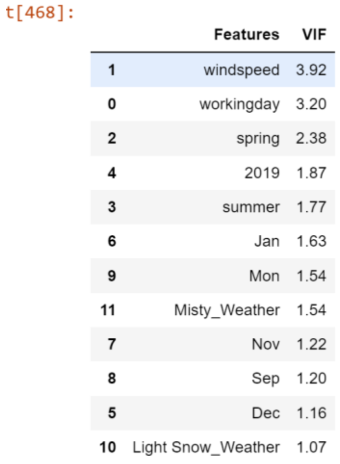
We ploted a scatter plot to check if the Independent and dependent variables are linearly correlated .
For which we use pair plot command in jupyter notebook



We checked if the error terms are normally distributed by creating a hist plot as below :



We checked Error terms are independent of each of other through VIF and kept VIF value below 5 for each variable :


t[468]:

|  | Features | VIF |
|---|---|---|
| 1 | windspeed | 3.92 |
| 0 | workingday | 3.20 |
| 2 | spring | 2.38 |
| 4 | 2019 | 1.87 |
| 3 | summer | 1.77 |
| 6 | Jan | 1.63 |
| 9 | Mon | 1.54 |
| 11 | Misty_Weather | 1.54 |
| 7 | Nov | 1.22 |
| 8 | Sep | 1.20 |
| 5 | Dec | 1.16 |
| 10 | Light Snow_Weather | 1.07 |

**Q5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

A- Top 3 factors are
1) Weather (-0.3207 * Light Snow_Weather )
2) Season (-0.2381*spring)
3) Windspeed (-0.1926 )

## General Subjective Questions

**Q1 Explain the linear regression algorithm in detail.**

A- Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The basic idea is to find the best-fitting straight line through the data points.

Formula:

The formula for a simple linear regression (one dependent and one independent variable) is:
Y=β0+β1X+ϵ
Here:

- Y is the dependent variable you're trying to predict.
- X is the independent variable you're using for the prediction.
- β0 is the y-intercept, the predicted value of Y when X is 0.
- β1 is the slope of the line, representing the change in Y for a one-unit change in X.
- ϵ is the error term, the part of Y the line can't explain.

Finding the Best Line
The "best-fitting" line is the one where the sum of the squares of the differences between the observed values (actual data points) and the predicted values (points on the line) is minimized. This method is known as Ordinary Least Squares.

Example

Imagine you're studying the relationship between hours studied (independent variable X) and test scores (dependent variable Y). Your linear regression model will try to find a line that best predicts test scores based on hours studied.

Multi-variable Linear Regression
When there are multiple independent variables, the formula expands to:
$Y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$

Each $\beta$ represents the relationship between Y and a different independent variable Xi .

In real-world scenarios, linear regression can be implemented using statistical software or programming languages like Python, R, etc. They provide functions or libraries (like scikit-learn in Python) to fit a linear model to your data and make predictions.

**Assumptions**
Linear regression assumes:

- Linear relationship between the dependent and independent variables.
- Independence of the errors.
- Homoscedasticity (constant variance of the errors).
- Normal distribution of the errors.

**Q2 Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a fascinating example in the field of statistics that demonstrates the importance of visualizing data before analyzing it. It comprises four different datasets (hence the name "quartet") that have nearly identical simple statistical properties, yet appear very different when graphed.

The Quartet
Each dataset in Anscombe's quartet consists of eleven points and was constructed in 1973 by the statistician Francis Anscombe. The four datasets are usually labeled I, II, III, and IV.

Key Statistical Properties
The intriguing part about these datasets is that they have nearly identical:

Mean and variance for both the X and Y variables.
Correlation between the X and Y variables.
Linear regression line (same slope and y-intercept) when performing a simple linear regression of Y on X.

The Twist
Despite these similarities, when you graph these datasets, they look very different. Here's a brief description of each:

Dataset I: Follows a simple linear relationship (the kind you would expect when assuming a linear model).
Dataset II: Forms a perfect curve (quadratic relationship), but not a straight line.
Dataset III: Appears as a tight linear relationship, but with one significant outlier.
Dataset IV: Has a distinct pattern where X values are mostly constant (except for one outlier), which challenges the assumptions of regression analysis.

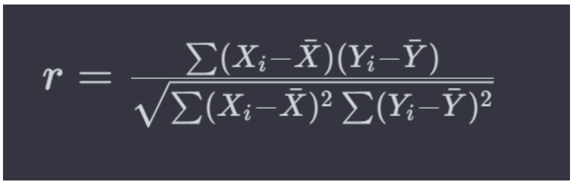
The Lesson
The key lesson from Anscombe's quartet is that relying solely on summary statistics (like mean, variance, and correlation) can be misleading. It highlights the importance of graphically inspecting the data before using statistical tools. No matter how similar datasets may appear statistically, they can behave very differently.

Anscombe’s quartet is often used to advocate for the necessity of data visualization and a reminder that statistical analysis should not be the only tool used when exploring and interpreting data. It’s a classic example to demonstrate why the assumption “if the statistics are the same, then the data is the same” is flawed.

**Q3 What is Pearson’s R?**

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It's a widely used statistical tool to assess how two variables are related in a dataset.

How It's Calculated
The formula for Pearson's correlation coefficient is:



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

1. Here:
   - *Xi* and *Yi* are individual sample points.
   - ¯*X*¯ and ¯*Y*¯ are the means of the *X* and *Y* datasets.
   - The summations run over all data points.

**Understanding the Value of R**

- The value of Pearson’s R lies between -1 and +1.
- 1*R*=1 indicates a perfect positive linear relationship: as one variable increases, the other variable increases at a constant rate.
- −1*R*=−1 indicates a perfect negative linear relationship: as one variable increases, the other decreases at a constant rate.
- *R*=0 suggests no linear relationship between the variables.
- Values close to +1 or -1 suggest a strong linear relationship, while values near 0 indicate a weak relationship.

**Example**

Imagine you have data on hours studied and test scores for a group of students. If Pearson's R for these two variables is close to +1, it suggests that more hours studied is associated with higher test scores, indicating a strong positive linear relationship.

**Context and Cautions**

- Pearson's R only measures linear relationships. Non-linear relationships can exist even if Pearson's R is close to 0.
- A high or low R value doesn't imply causation. It simply indicates the degree of linear association between two variables.

- Outliers can significantly affect the value of R, potentially giving a misleading impression of the relationship.

Pearson's correlation coefficient is a fundamental tool in statistics for quantifying linear relationships and is widely used across various fields, from psychology to finance. However, its interpretation always needs to be contextualized and approached with an understanding of its limitations.

**Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A- Scaling in the context of multivariate linear regression is a method of transforming the range of values of your variables (features) so that they can be compared on common grounds. It's particularly important in models that rely on the distance between data points, like linear regression with multiple variables, because variables on larger scales can unduly influence the model.

**Why Perform Scaling?**

1. Consistent Scale: Variables might be measured in different units (e.g., dollars, kilometers, kilograms). Scaling ensures these different units don't distort the model.
2. Numerical Stability: Some algorithms converge faster when features are on a similar scale.
3. Importance Balancing: Prevents features with larger scales from dominating those in smaller scales in the model.

**Types of Scaling**

There are two common types of scaling: Normalization and Standardization.

Normalized Scaling (Min-Max Scaling)

Normalization rescales the features to a range of [0, 1] or [-1, 1]. The formula is:

$X\mathrm{norm}= (X-X\mathrm{min}) / (X\mathrm{max}-X\mathrm{min})$

Here, $X$ is the original value, $X\mathrm{min}$ is the minimum value in the feature, and $X\mathrm{max}$ is the maximum value.

Pros:

- Bounding the values can be useful for algorithms that require input values within a specific range.
- It preserves the original distribution without distorting the differences in the ranges of values.

Cons:

- It's sensitive to outliers. An extreme value can skew the scale.

Standardized Scaling (Z-score Normalization)

Standardization rescales features so they have a mean of 0 and a standard deviation of 1. The formula is:

$X\mathrm{std}=X-\bar{X}/\sigma$

*Where* $\bar{X}$ *is the mean and* $\sigma$ *is the standard deviation.*

Pros:

- Less affected by outliers.
- Suitable for techniques which assume data is normally distributed.

Cons:

- Doesn't bound values to a specific range, which might be necessary for some algorithms.

Choosing the Right Method

- Normalization is typically used when the data does not follow a Gaussian distribution or when the scale of the input feature is irrelevant (like image data).
- Standardization is generally preferred for methods that assume the input features are normally distributed, or when the algorithm involves distance calculations (like K-Means, or when using L2 regularization).

In summary, both normalization and standardization are forms of feature scaling that are important in the context of multivariate linear regression and other machine learning algorithms. The choice between them depends on the specific algorithm and the nature of the data.

**Q5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

A-

The Variance Inflation Factor (VIF) is a measure used to detect the presence and intensity of multicollinearity in a regression analysis. Multicollinearity occurs when two or more predictor variables (independent variables) in a model are highly correlated with each other.

Why VIF Can Be Infinite

1) The VIF for a predictor variable is calculated using the formula:
   VIF=1/1−*R*^2
2) Here, *R*^2 is the coefficient of determination of a regression model where the predictor in question is regressed against all other predictors.
3) An infinite VIF (or a very high value) typically occurs when:
   1. Perfect or Near-Perfect Collinearity**:** If *R*^2 is exactly 1, which happens in cases of perfect collinearity (i.e., the predictor can be perfectly predicted by a linear combination of other predictors), the denominator in the VIF formula becomes zero, leading to an infinite VIF. Near-perfect collinearity can push *R*^2 very close to 1, leading to extremely high VIF values.
   2. Redundant Variables**:** This can occur when you include variables that are linear combinations of other variables in your dataset. For example, if you have one variable that's the sum or difference of two others, this can lead to an infinite VIF.
   3. Highly Correlated Variables**:** Even if variables are not perfect linear combinations of each other, very high correlation can lead to high *R*^2 values, resulting in a very high VIF.

Implications

- An infinite or very high VIF indicates that the predictor is too closely related to one or more of the other predictors, which undermines the statistical significance of the variable.
- This can make the coefficients of the regression model unreliable and unstable, meaning small changes in the data could lead to large changes in the model.

**Q6 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 specific distribution, typically a normal distribution. In the context of linear regression, it's often used to check the normality of residuals.

Understanding Q-Q Plots

- In a Q-Q plot, you plot the quantiles of your data against the quantiles of a theoretical distribution (like the normal distribution).
- **Quantiles** are points in your data below which a certain proportion of your data falls. For instance, the median is a quantile where 50% of data falls below it.
- The **X-axis** of a Q-Q plot shows the theoretical quantiles. This means it doesn't depend on your data but on the distribution you're comparing your data to.
- The **Y-axis** displays your actual data quantiles.

**1 Using Q-Q Plots in Linear Regression**

- In linear regression, one of the key assumptions is that the residuals (the differences between the observed values and the values predicted by the model) are normally distributed. This is where a Q-Q plot becomes valuable.
- By creating a Q-Q plot of the regression residuals, you can visually assess if they follow a normal distribution.
- If the residuals are normally distributed, they should fall approximately along a straight line in the Q-Q plot.

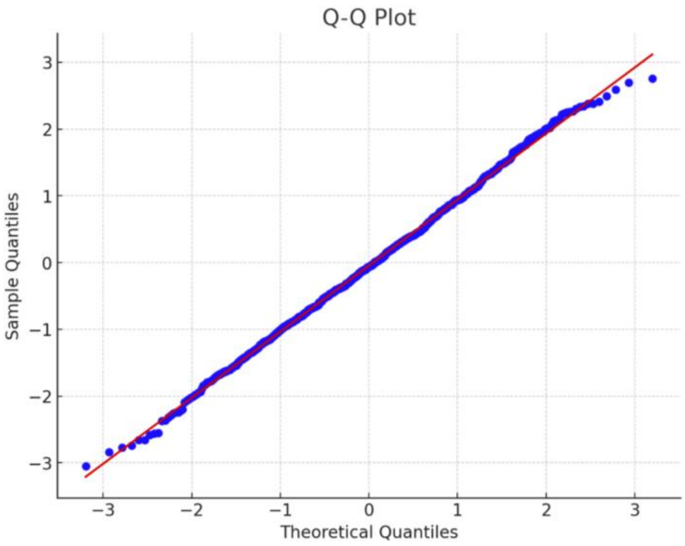
**2 Importance of Q-Q Plots in Linear Regression**

1. **Assumption Validation:** Helps in checking the normality assumption of linear regression. If the residuals are not normally distributed, the estimates of the confidence intervals and hypothesis tests may not be accurate.

2. **Model Diagnostics:** A Q-Q plot can reveal deviations from normality like skewness or kurtosis. This can guide improvements to the model, such as transformations of variables.
3. **Outlier Detection:** Outliers can also be identified in a Q-Q plot. They appear as points that fall far from the straight line.

## 3 Interpreting Q-Q Plots

- If the points lie on or close to a straight line, your data is likely normally distributed.
- If the points deviate systematically from the line, it indicates a deviation from normality. For example, a "S" shaped curve might suggest heavy tails (data has more extreme values than the normal distribution).
- In summary, the Q-Q plot is a fundamental diagnostic tool in regression analysis. It provides a visual means of assessing whether the residuals of your model are behaving as expected (normally distributed in this case), which is crucial for the reliability and validity of the linear regression model's results.


Q-Q Plot
Sample Quantiles
Theoretical Quantiles