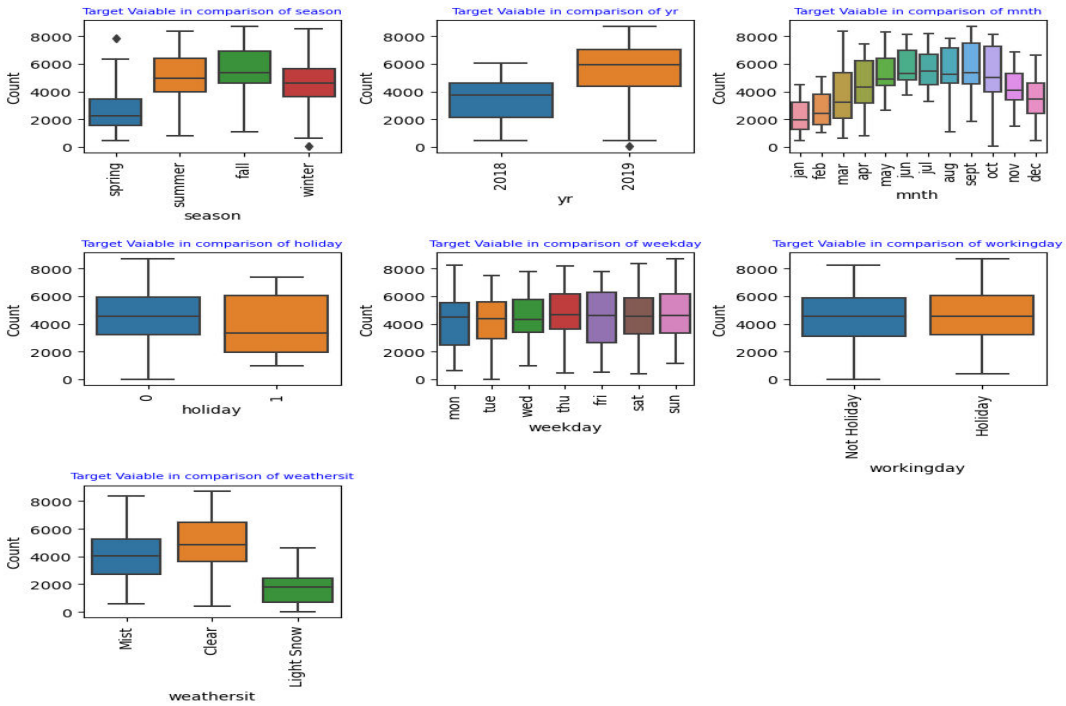
Bike Sharing Assignment

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?

Answer: Here are the rephrased inferences about the effect of categorical variables on the dependent variable:

- There is an increase in the demand for shared bikes during the fall season.
- The demand for shared bikes has been rising year over year. As the demand of shared bike is more in 2019 in comparison of 2018
- The demand for shared bikes is highest in September.
- Clear weather conditions are associated with a very high demand for shared bikes.
- The demand for shared bikes is lower on holidays because may be people like to stay in homes on holidays.
- The demand of shred bike is almost equal for working day and non-working day.
- When the weather is clear, the demand for shared bikes is very high.



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

Answer: It is very important to use drop_first = true during dummy variable creation because :

1. **Multicollinearity :** Multicollinearity happens when one or more variable are highly correlated with each other . If all categories of a independent variable are included as dummy variables, one of these categories can be perfectly predicted from the others, leading to multicollinearity. To avoid this we use drop_first = True .
2. **Redundancy:** Including all the dummy variables in the dataset cause the redundancy dropping one variables removes this redundancy, making the model more efficient without losing any information.
3. **Simple Interpretation:** Using on category as a reference, coefficient of remaining dummy variables is different from each category. And low number of columns also reduces the complexity.

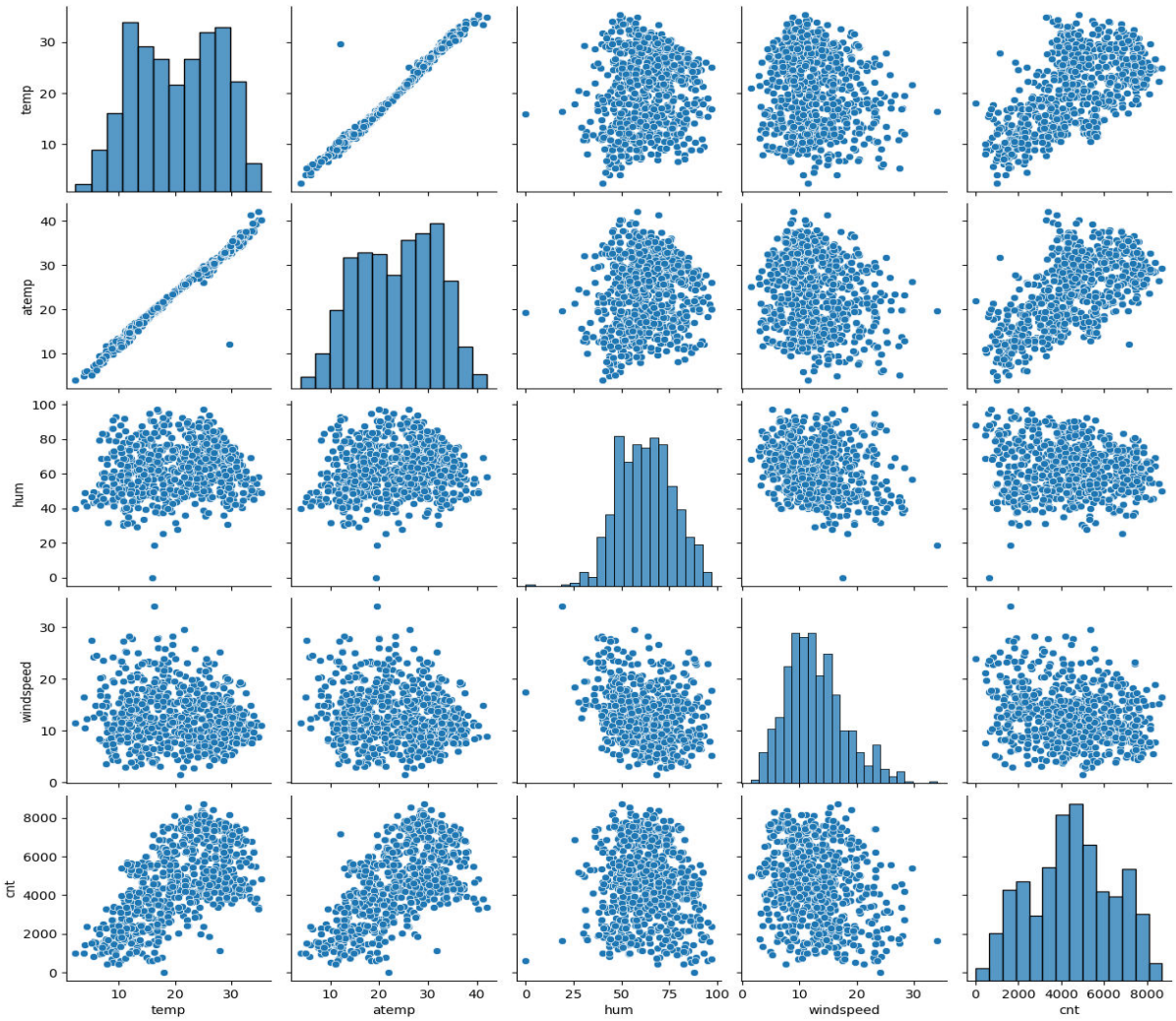
Eg: Suppose we have categorical variable " **red"**,"**Green"**," **blue"** after using drop_first=True one category (e.g., `Red`) is omitted, and the dummy variables for `Green` and `Blue` are created

- `Green`: `[0, 1]` • `Blue`: `[0, 0]`

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

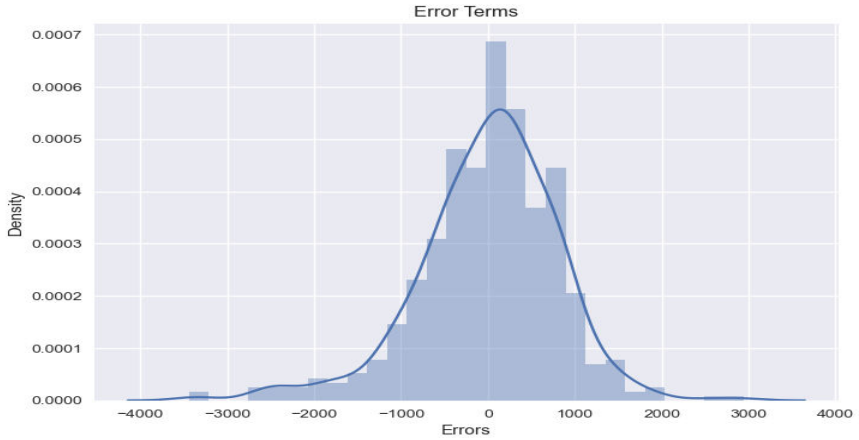
Answer: Looking at the pair-plot among the numerical variables, the numerical variables which has the highest correlation is :

- **temp**
- **atemp**

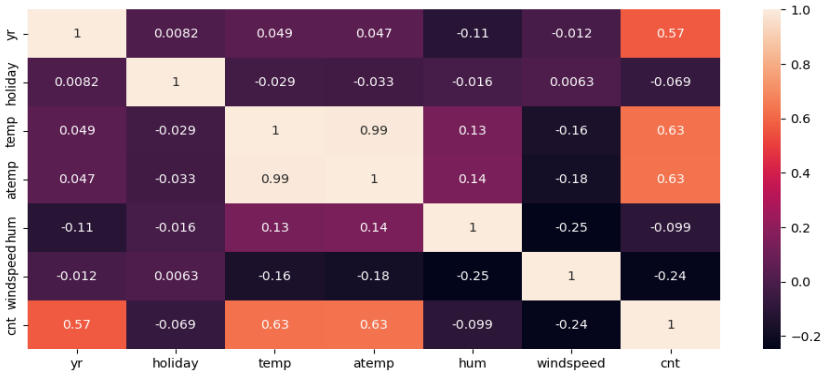


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: **1. Error terms are normally distributed with mean zero.**

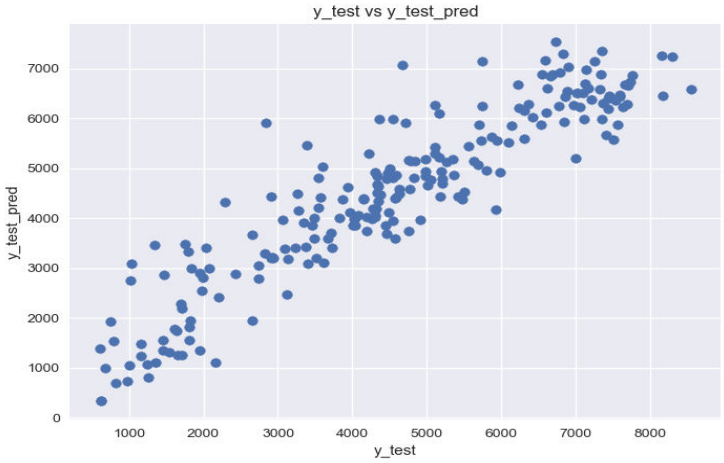


2. No Multicollinearity



if we see in above graph temp atemp and temp are highly positively correlated so we drop one of the variable.

3. Homoscedasticity



4. Error terms are independent of each other

5. **Linearity Check:** Use scatter plots of observed vs. predicted values or partial regression plots to assess linearity.

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

Answer :

- **Temperature (`temp`)**

  Coefficient: 4086.7580

  Explanation: Temperature has a positive and substantial impact on bike demand, with a large coefficient indicating that an increase in temperature leads to a significant increase in demand.

- **Year 2019 (`yr_2019`)**

  Coefficient: 1967.4799

  Explanation: The year 2019 has a strong positive effect on bike demand compared to the baseline year, suggesting an overall increase in demand in 2019.

- **Weather Situation - Light Snow (`weathersit_Light Snow`)**

  Coefficient: -2193.9944

  Explanation: The presence of light snow dramatically decreases bike demand, as indicated by the large negative coefficient. This suggests that snow significantly deters people from using shared bikes.

General Subjective Questions

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

   Answer**:**

   Linear regression is a type of supervised machine learning algoithm. It is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It's widely used for prediction purpose.

Types of Linear Regression

1 **Simple Linear Regression:** Involves one dependent variables and one independent variable.

2. **Multiple Linear Regression:** Involves one dependent variables and two or more than two independent variable.

The **basic idea of linear regression is** to find a linear relationship between the dependent variable y and one or more independent variables and **creating the best fit line**.

Simple Linear Regression Model:

y_predicted = a + bx

here :

a is the intercept(value of y when x =0)

x is Independent variable (predictor)

b is slope of the line ((change in y for a unit change in x)

Multiple linear regression model:

y_predicted = a + b1 x1 + b2 x2 + … + bk xk

here :

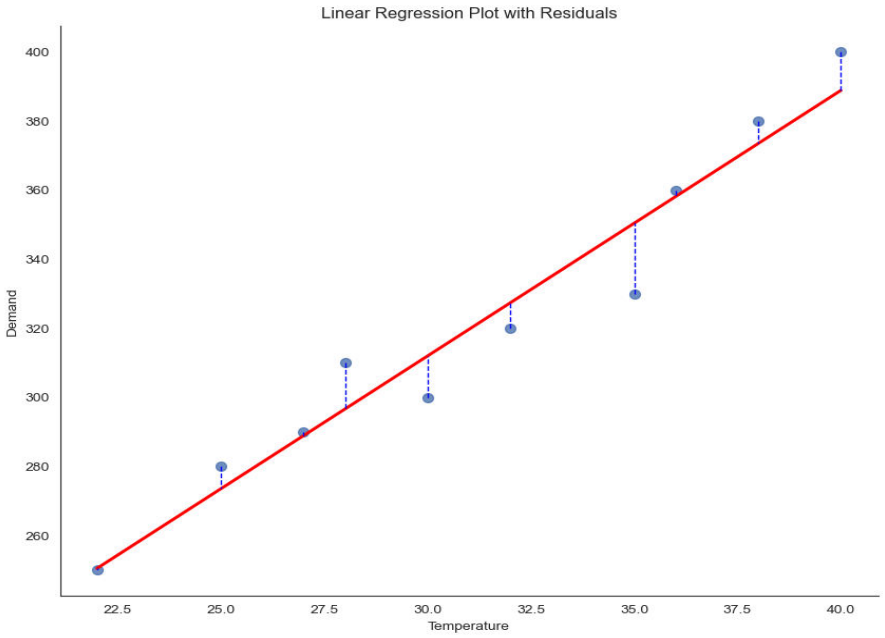
a is the intercept(value of y when x =0)

x1,x2,..., xk is Independent variable (predictor)

b is slope of the line ((change in y for a unit change in x)

Assumption of Linear Regression:

1. **Linearity**: Relation between dependent and independent variable should be linear.
2. **Independence**: observations should be independent of each other.
3. **Homoscedasticity**: The residuals (y-y_predicted) have constant variance at all levels of the independent variable(s).
4. **Normality**: The residuals of the model are normally distributed
5. **No Multicollinearity**



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

Answer:

Anscombe's Quartet is a classic example used to **illustrate the significance of data visualization** in revealing insights that **might be missed through statistical analysis alone**.

The **Quartet** consists of **four datasets**, each containing **eleven (x, y) points**. Despite **having identical descriptive statistics** (such as mean, variance, and standard deviation), each dataset **displays distinct characteristics when graphed**. This highlights the importance of visualizing data to uncover different patterns and behaviors that may not be apparent through statistical measures alone.

Components of Anscombe's Quartet

Mean of X values: The average value of the X variable is the same for all four datasets.
Mean of Y values: The average value of the Y variable is the same for all four datasets.
Variance of X values: The measure of how spread out the X values are from the mean is the same.
Variance of Y values: The measure of how spread out the Y values are from the mean is the same.
Correlation between X and Y: The strength and direction of the relationship between X and Y are the same.
Linear Regression Line: The best-fit line (least-squares line) for predicting Y from X is the same in all datasets

The Four Datasets

Dataset 1:

This dataset appears to follow a linear relationship. When plotted, it shows a scatter plot where the points fall closely around a straight line.

Dataset 2:

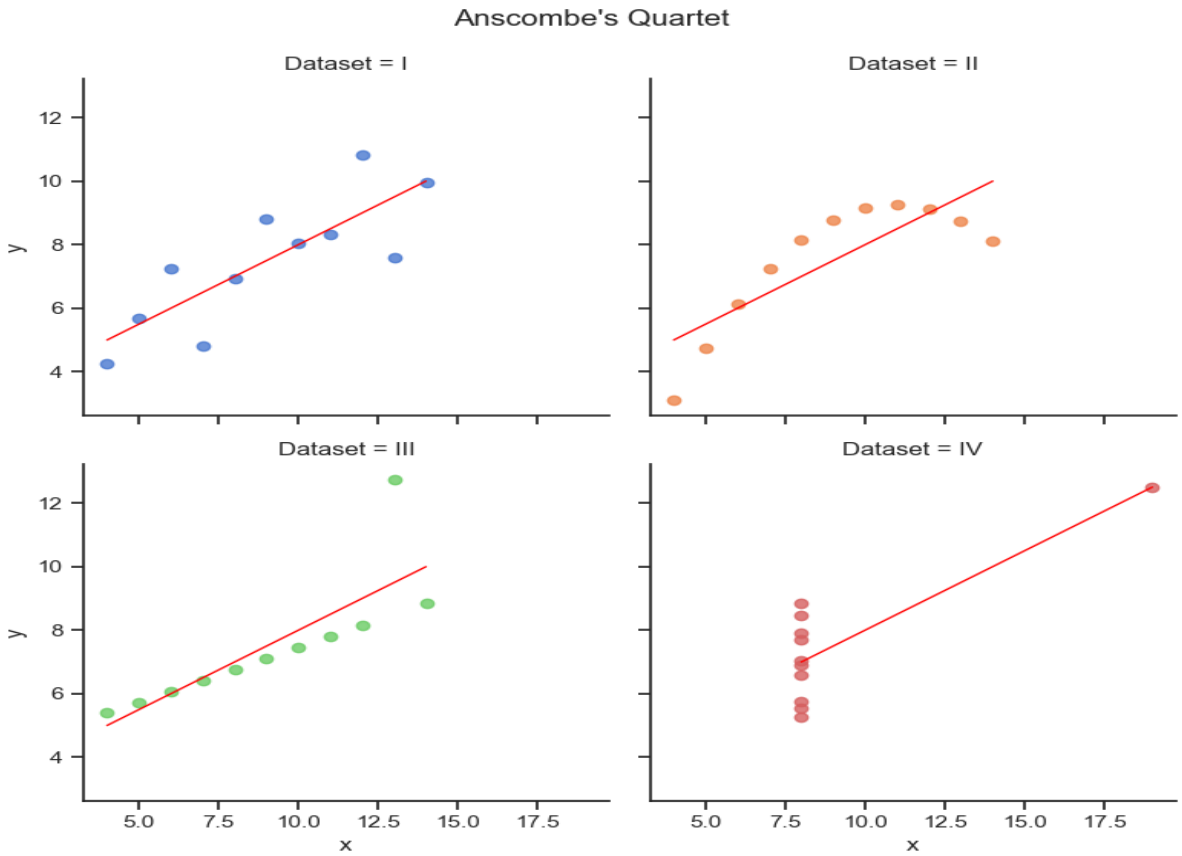
This dataset forms a curve. Even though the summary statistics are similar, the scatter plot reveals a quadratic relationship.

Dataset 3:

In this dataset, most points lie on a straight line, but there is one outlier that can heavily influence the correlation and regression line.

Dataset 4:

Here, most of the X values are the same except one, making a vertical line. The Y values are quite spread out, with one extreme outlier that affects the overall statistics.



Importance and Implications

Anscombe's Quartet shows that datasets with identical statistical properties can have vastly different distributions. This underscores the importance of visualizing data to fully understand its structure and relationships.

The Quartet emphasizes that relying solely on summary statistics can be misleading. Visualizations, such as scatter plots, reveal patterns, outliers, and relationships that statistics alone may not capture.

The Quartet highlights the necessity of validating the assumptions of statistical models. For instance, assuming linearity when the data is not linear can lead to incorrect conclusions.

Conclusion

- Before drawing conclusions, plot the data to see its shape and trends.
- Summary Statistics Aren't Enough
- Mean, variance, and correlation can't capture the entire story.
- Visualization helps in identifying anomalies and true relationships

3. What is Pearson's R?

Answer: **Pearson's R**, also known as the **Pearson correlation coefficient**,
It is a statistical measure that **computes the strength and direction of the linear relationship between two continuous variables**.
It is commonly used to assess how closely the data points in two datasets relate to each other.

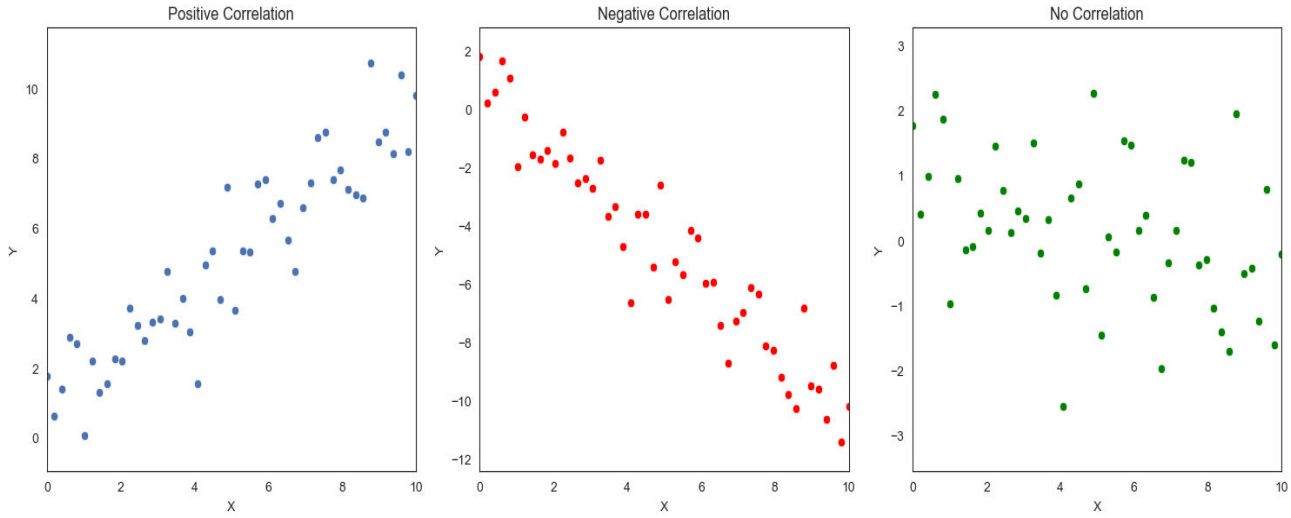
it is **Mathematically represented** as :

$$r = COV(X,Y) \ / \ \sigma X \, \sigma Y$$

Cov(X,Y) is the covariance between variables
σX σY are the standard deviations of X and Y

Pearson correlation coefficient **ranges** between **0 to -1**

1 means perfect **positive correlation**
-1 means prefect **negative correlation**
0 means no relation
Values close to 1 indicate **strong positive correlation.**
Values close to -1 indicate **strong negative correlation.**
Values close to 0 indicate **weak or no correlation.**



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

Answer:

Scaling: Scaling is a technique used to transform the numerical independent features so they fall in a specific range or have specific properties. So each features contributes equally.

Why we perform scaling:

-**Standardizing** all **numerical column** before model training is very important step, particularly when working on machine learning model because it ensures all features contribute equally, when features are measured in different units.

- **Standardizing** can **make data less sensitive to outliers** by centering data around zero, which can help in identifying and mitigating their impact.

- **Standardizing improves** the **Performance of the model** and **convergence speed.**

Difference between normalized scaling and standardized scaling:

Range:

Normalized Scaling : Scales ranges usually between 0 to 1.

Standardized scaling: Range is not fixed depend on the data**.**

Sensitivity to Outliers:

Normalized Scaling: Sensitive to outliers because the minimum and maximum values are directly used in scaling.

Standardized scaling: Less sensitive to outliers compared to normalization because the scaling is based on mean and standard deviation, which are less affected by extreme values.

Data Distribution:

Normalized Scaling: Does not assume any particular distribution of the data. It just scales the feature to fit within the specified range.

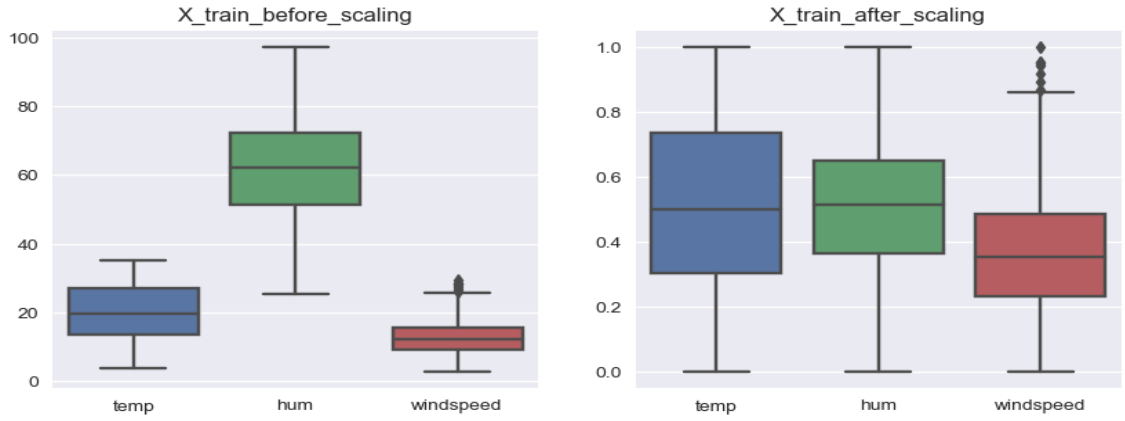
Standardized scaling: Assumes that the data can be approximately normally distributed. It centers the data and scales based on variance.

Formula:

Normalized Scaling: $x - \min(x) / \max(x) - \min(x)$

Standardized scaling: $X - \mu / \sigma$

Let see the comparison between before and after scaling of the feature:



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

Answer:

Variance Inflation Factor (VIF) :

vif measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

Reason of having the value of VIF Infinite:

1. **Perfect multicollinearity** It occurs when one independent variable is a perfect linear function of one or more other independent variables. In this case, the independent variables are perfectly correlated.

2. **If the matrix of predictors (design matrix) is singular or nearly singular** if there is exact or near exact linear dependency among the independent variables. this happens because when we have too many independent variables relative to the number of observations.

3. **Incorrectly entered data** or duplicate entries might create scenarios where predictors are perfectly correlated.

4. **Including redundant variables or dummy variables** without proper reference categories can cause high multicollinearity, potentially leading to infinite VIFs

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

ANSWER:

QQ plot is a **graphichal method** for comparing **two probabilty distributiions** by plotting their **quantiles each other.**

Use and Importance of Q-Q Plot in Linear Regression

Checking Normality of Residuals: It helps identify deviations from normality, such as skewness or kurtosis, in the residuals of a regression model.

Model Adequacy: If the residuals are not normally distributed, it might indicate that the model is not adequately capturing the underlying relationship or that there are outliers or influential points affecting the results. This could suggest the need for model adjustments or the use of different statistical techniques.

Identifying Outliers and Influential Points: Points that deviate significantly from the expected line in a Q-Q plot might be outliers or influential data points that have an unusual effect on the regression results.

Influence of Data Points: Identifying such points can help in deciding whether they should be removed.

Transformations:If the Q-Q plot indicates non-normality, transformations (e.g., log, square root) can be applied to achieve normality.

Conclusion

The Q-Q plot is an essential diagnostic tool in linear regression, providing visual evidence of the normality of residuals. By using Q-Q plots, analysts can ensure that regression assumptions hold, leading to more accurate and reliable models.