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?

Ans:

a) Bikes demand is at its highest during the fall but undergoes the most significant decline during spring, reaching its lowest point.

b) In 2019, there was an increase in bikes demand compared to 2018.

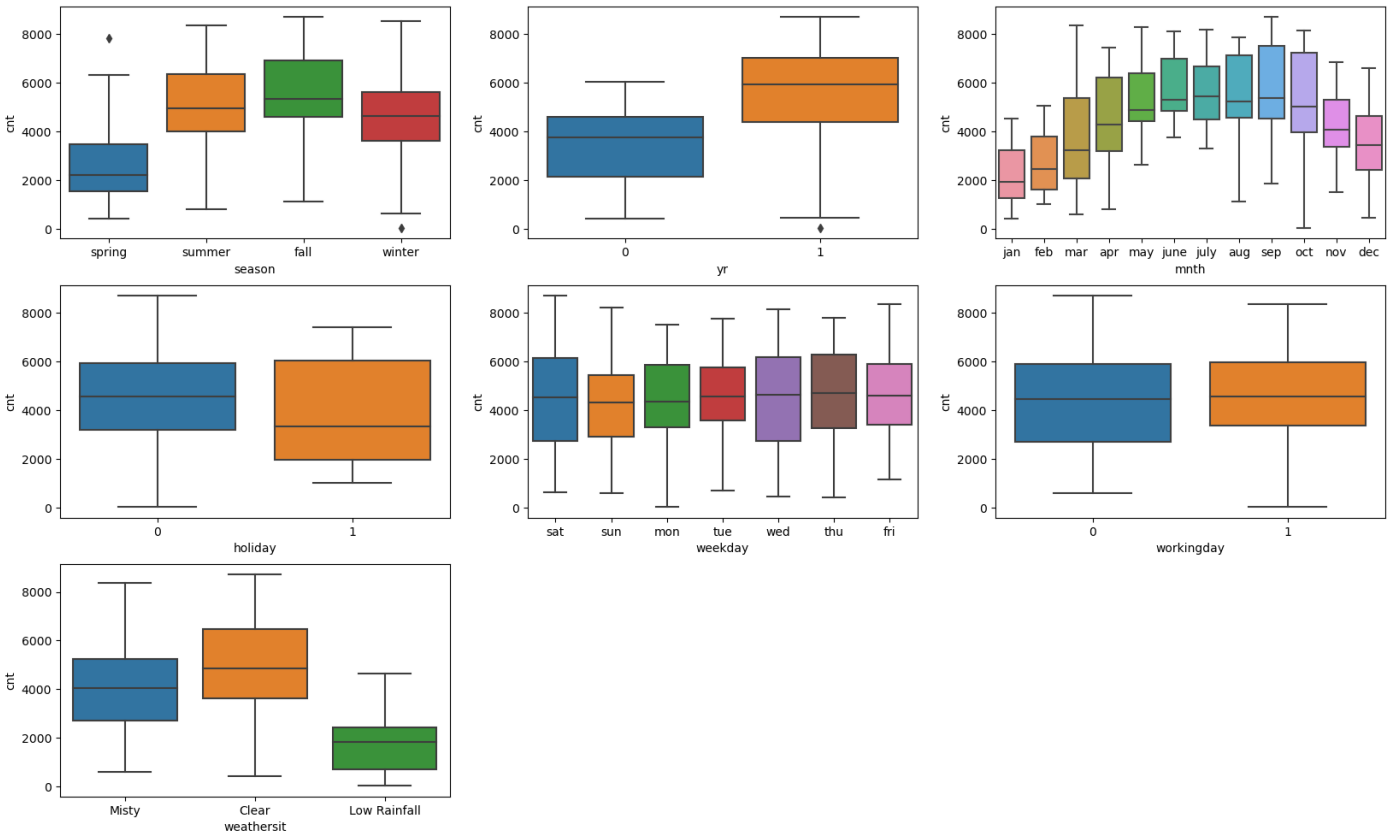
c) Bikes are in high demand during the months spanning from May to October.

d) Clear or slightly cloudy weather conditions result in higher bicycle demand, whereas light rain or snow leads to lower demand.

e) During non-holidays, the demand is higher when compare of holidays.

f) Bicycle demand remains consistent throughout the weekdays.

g) Whether it's a working day or not, bicycle demand remains unchanged.



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

Ans:

Using `drop_first=True` during dummy variable creation is important for several reasons:

1. **Multicollinearity:** Including all dummy variables can lead to multicollinearity, where variables are highly correlated, potentially causing unstable coefficient estimates.

2. **Interpretability:** It simplifies interpretation by providing a clear baseline category for comparison, making it easier to understand the effects of different categories.

3. **Dimensionality Reduction:** By excluding one category, it reduces dimensionality, which can enhance model stability and reduce computational complexity. This approach is particularly useful in large datasets.

4. **Achieving k-1 Dummy Variables:** It ensures that you have exactly k-1 dummy variables for a categorical variable with k categories, eliminating redundancy and avoiding perfect multicollinearity. This is essential for accurate modeling.

In summary, `drop_first=True` helps prevent multicollinearity, improves interpretability, reduces dimensionality, and ensures that you have the right number of dummy variables to effectively represent categorical data in various statistical and machine learning models.

Creating the Dummy Variables

In [19]:
```
# Get the dummy variables for the feature 'season', 'weathersit', 'mnth' and 'weekday'

d1 = pd.get_dummies(bike['season'], drop_first=True) ## for season feature
bike = pd.concat([bike, d1], axis = 1)  # Add the results to the original bike dataframe

d2 = pd.get_dummies(bike['weathersit'],drop_first=True) ## for weathersit feature
bike = pd.concat([bike, d2], axis = 1)  # Add the results to the original bike dataframe

d3 = pd.get_dummies(bike['mnth'], drop_first=True) ## for season feature
bike = pd.concat([bike, d3], axis = 1)  # Add the results to the original bike dataframe

d4 = pd.get_dummies(bike['weekday'],drop_first=True) ## for weathersit feature
bike = pd.concat([bike, d4], axis = 1)  # Add the results to the original bike dataframe
```

In [20]:
```
d2.head()
```

Out[20]:

	Low Rainfall	Misty
0	0	1
1	0	1
2	0	0
3	0	0
4	0	0

- The data set has only 3 values for `weathersit` variable, so creating dummies has only created 2 sets of columns.

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

Ans:

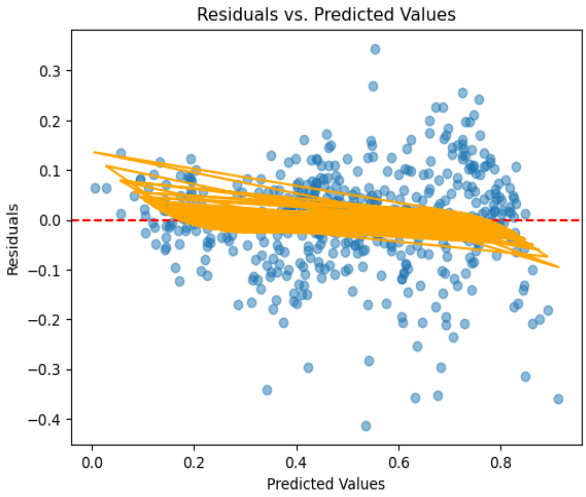
Both “atemp” and “temp” exhibit a strong positive correlation of 0.63 with the target variable, which happens to be the highest among all numerical variables in the dataset. This suggests that changes in both “atemp” (feeling temperature) and “temp” (actual temperature) are associated with similar changes in the target variable.

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

Ans:

To validate the assumptions of Linear Regression after building the model on the training set, I employed the following key steps:

1. **Residual Analysis:**



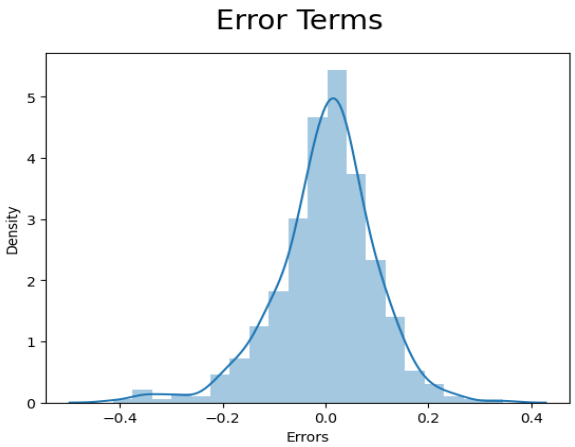
- Calculated the residuals by subtracting the predicted values from the actual target values in the training set.

- Plotted the residuals against the predicted values to identify any patterns, ensuring there were no clear systematic deviations.

2. **Homoscedasticity (Constant Variance):**

- Examined scatter plots of residuals against predicted values or independent variables.

- Checked for consistent spread of residuals across the range of predicted values, confirming the assumption of constant variance.



3. **Normality of Residuals:**

- Created a histogram and Q-Q plot of the residuals to assess their distribution.

- Ensured that the residuals approximately followed a normal distribution, validating the assumption of normally distributed errors.

In summary, I used residual analysis, assessed the normality of residuals, and examined homoscedasticity to validate the key assumptions of Linear Regression after building the model on the training set. These steps are essential to ensure the reliability and appropriateness of the regression model for making predictions.

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

Ans:

The top three features contributing significantly to explaining the demand for shared bikes in the final linear regression model, based on their p-values, are:

1. **‘yr’** (year) with a p-value of approximately 1.97e-96.
2. **‘atemp’** (feeling temperature) with a p-value of approximately 5.75e-46.
3. **‘spring’** (season: spring) with a p-value of approximately 4.73e-29.

These features have extremely low p-values, indicating high statistical significance and strong contributions to the model's explanatory power.

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a fundamental statistical and machine learning algorithm used for modelling the relationship between a dependent variable and one or more independent variables. It assumes that the relationship between these variables is linear, meaning that a change in the independent variable(s) corresponds to a constant change in the dependent variable.

1. Assumptions:

Linear regression relies on several key assumptions:

- Linearity: The relationship between the dependent variable (y) and the independent variable(s) (X) is linear.

- Independence: Observations are independent of each other.

- Homoscedasticity: The variance of the residuals (the differences between observed and predicted values) is constant across all levels of the independent variable(s).

- Normality: The residuals are normally distributed.

- No or Little Multicollinearity: Independent variables are not highly correlated.

2. Simple Linear Regression:

- In simple linear regression, there is one independent variable (X) and one dependent variable (y).

- The model assumes a linear relationship between y and X, which can be represented as:

$y = \beta_0 + \beta_1 X + \varepsilon$

- y: Dependent variable

- X: Independent variable

- β_0: Intercept (the value of y when X is 0)

- β_1: Coefficient of X (the slope of the line)

- ε : Error term (captures the unexplained variability)

3. Multiple Linear Regression:

- In multiple linear regression, there are multiple independent variables (X_1, X_2, ..., X_p) and one dependent variable (y).

- The model extends the simple linear regression equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

- β_0: Intercept

- β_1, β_2, ..., β_p: Coefficients of the independent variables

- ε : Error term

4. Fitting the Model:

- The goal is to estimate the coefficients (β_0, β_1, β_2, ..., β_p) that minimize the sum of squared residuals.

- This is typically done using a method like Ordinary Least Squares (OLS) for simple linear regression or gradient descent for multiple linear regression.

5. Assessing Model Fit:

- Various statistics and plots are used to assess how well the model fits the data, including the R-squared value (a measure of how much variance in y is explained by the model) and residual plots.

6. Making Predictions:

- Once the model is fitted, it can be used to make predictions. Given new values of the independent variables, the model can predict the corresponding value of the dependent variable.

7. Interpretation:

- The coefficients (β_1, β_2, ..., β_p) represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.

8. Assumptions Checking:

- It's important to check whether the assumptions of linear regression are met. Violations can lead to unreliable results.

2. Explain the Anscombe’s quartet in detail.

Ans:

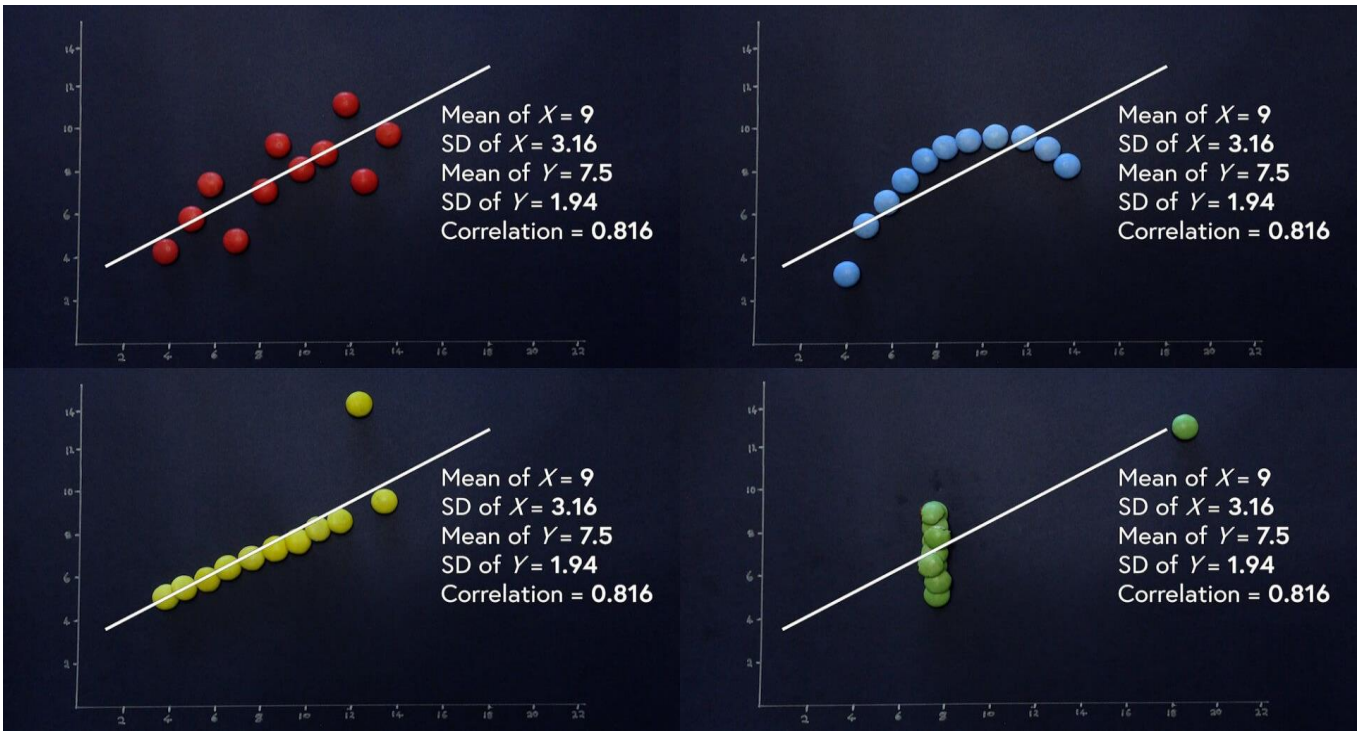
Anscombe's quartet is a set of four datasets that was created by the statistician Francis Anscombe in 1973. These datasets have identical simple descriptive statistics (mean, variance, correlation, and linear regression) but are profoundly different when graphically represented. Anscombe's quartet is often used to illustrate the importance of data visualization and the limitations of relying solely on summary statistics. Here's a detailed explanation of Anscombe's quartet:

1. Dataset Creation:

- Anscombe created four datasets, each containing 11 data points.
- Each dataset includes two variables: X and Y.

2. Descriptive Statistics:

- Remarkably, when you calculate the basic descriptive statistics for each dataset (mean, variance, correlation, and linear regression parameters), they are nearly identical. This includes:
  - Mean of X: Approximately 9.0
  - Mean of Y: Approximately 7.5
  - Variance of X: Approximately 11.0
  - Variance of Y: Approximately 4.12
  - Correlation between X and Y: Approximately 0.816
  - Linear regression equation: Y = 3.00 + 0.5X



3. Graphical Representation:

- Despite having nearly identical summary statistics, when you create scatterplots of these datasets, they look strikingly different.

- Each dataset has its unique distribution, shape, and pattern of points when plotted.

4. Understanding the Quartet:

- Dataset I: Forms a clear linear relationship.

- Dataset II: Linear relationship with an outlier.

- Dataset III: Non-linear relationship.

- Dataset IV: Linear relationship with one influential outlier.

5. Implications:

- Anscombe's quartet demonstrates the importance of data visualization in understanding data. Summary statistics alone do not provide a complete picture.

- It highlights the danger of making assumptions based solely on summary statistics.

- It underscores the value of exploring data through graphs to uncover patterns, outliers, and relationships.

6. Practical Use:

- In data analysis and statistics, Anscombe's quartet is often used to emphasize the importance of graphical exploration and to caution against over-reliance on summary statistics.

- It serves as a reminder that datasets can exhibit vastly different characteristics even when their basic statistics appear similar.

3. What is Pearson’s R?

Ans:

Pearson's correlation coefficient, often denoted as "r" or Pearson's "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses how well the relationship between these variables can be described by a straight line. Pearson's R is widely used in statistics to analyse the degree and direction of correlation between two variables.

Key characteristics of Pearson's R:

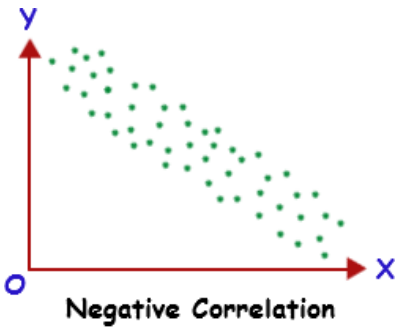
1. Range: The value of Pearson's R ranges from -1 to 1.

- An R value of -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases linearly).

- An R value of 1 indicates a perfect positive linear relationship (both variables increase linearly together).

- An R value of 0 suggests no linear relationship; the variables are not correlated.

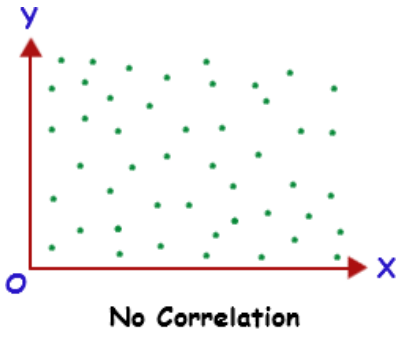
2. Direction: The sign of the correlation coefficient indicates the direction of the linear relationship:



- Positive R: Indicates a positive linear relationship (as one variable increases, the other tends to increase).

- Negative R: Indicates a negative linear relationship (as one variable increases, the other tends to decrease).

3. Strength: The absolute value of R (i.e., ignoring the sign) measures the strength of the linear relationship. A higher absolute R value indicates a stronger linear association between the variables.



4. Assumption: Pearson's correlation coefficient assumes that the relationship between the variables is linear and that the data follows a bivariate normal distribution.

5. Calculation: Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula is as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,

r = Pearson correlation coefficient

x = Values in the first set of data

y = Values in the second set of data

n = Total number of values.

Pearson's R is commonly used in various fields, including statistics, social sciences, economics, and data analysis, to assess the strength and direction of relationships between variables. It provides valuable insights into how two variables are related and is a fundamental tool for correlation analysis.

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

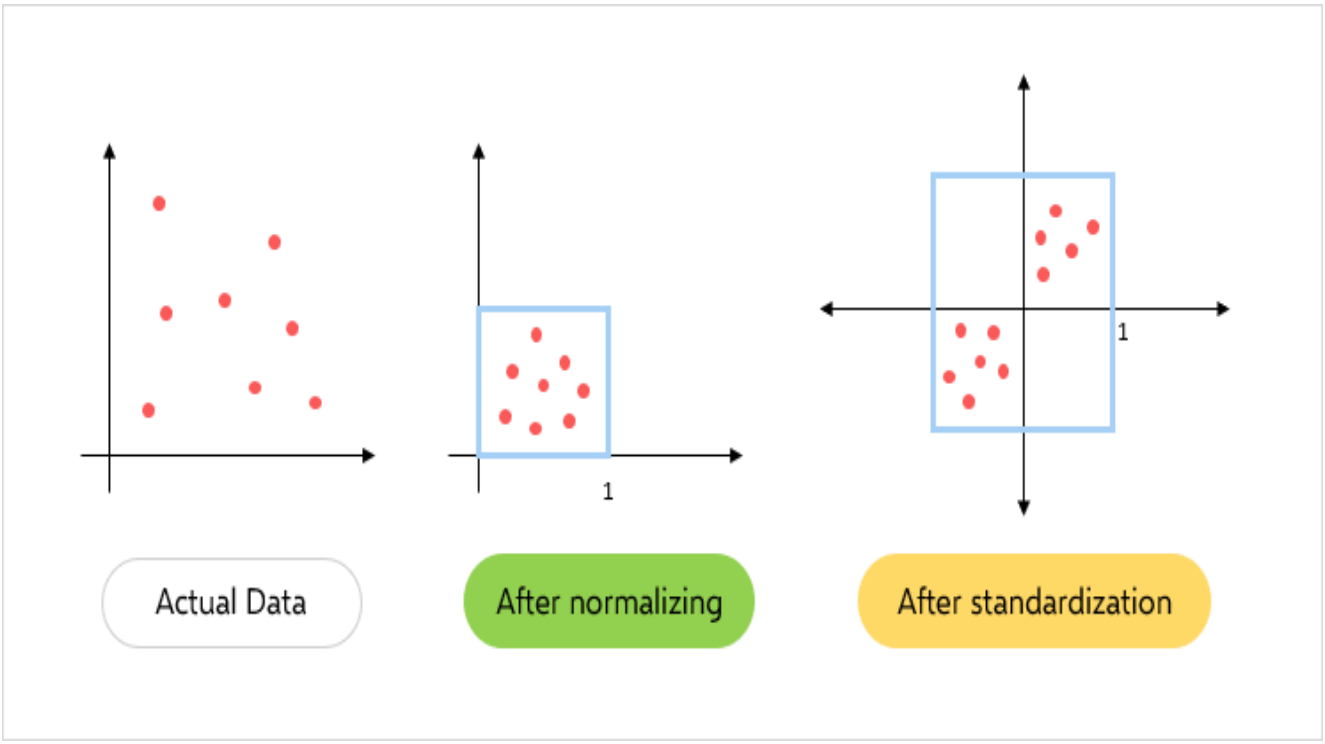
Ans:

Scaling in the context of data pre-processing refers to the process of transforming the values of variables to a standard range or distribution. It is performed to ensure that the variables have similar scales or units of measurement, making them more suitable for various data analysis techniques. Scaling helps in comparing and interpreting variables more effectively. There are two common scaling techniques: normalized scaling and standardized scaling.

Scaling is performed for several reasons:

1. Comparability: Variables with different units or scales can't be directly compared or combined. Scaling makes them comparable, enabling meaningful analysis.

2. Algorithm Performance: Many machine learning algorithms are sensitive to the scale of input variables. Scaling can improve the performance and convergence of algorithms like gradient descent, k-means clustering, and principal component analysis.



3. Interpretability: Scaling makes it easier to interpret coefficients in regression models. Without scaling, coefficients may represent different units, making their interpretation challenging.

4. Visualization: Scaling helps in creating meaningful visualizations. Plotting variables with similar scales allows for clearer insights.

Normalized Scaling vs. Standardized Scaling:

1. Normalized Scaling:

- Also known as "Min-Max scaling."

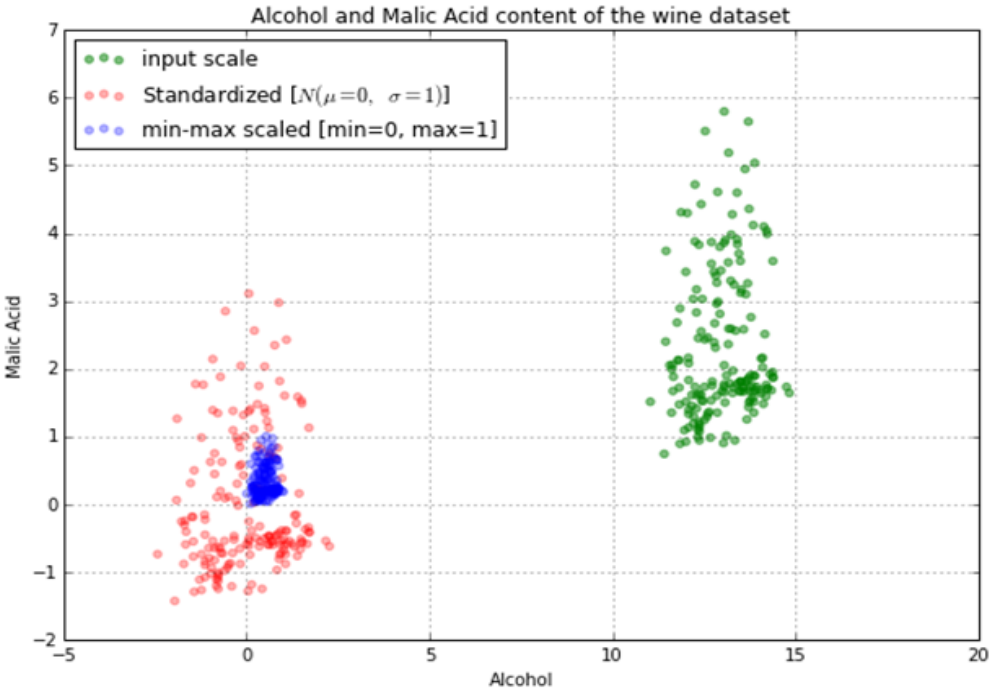
- Scales the data to a specific range, usually between 0 and 1.

- Formula for Min-Max scaling:

X_normalized = (X - X_min) / (X_max - X_min)

- Advantages: Maintains the original distribution shape, useful when the original data's distribution needs to be preserved.

- Disadvantages: Susceptible to outliers, may not handle extreme values well.



2. Standardized Scaling:

- Also known as "Z-score scaling" or "Mean-Std scaling."

- Centers the data around mean (0) and scales it to have a standard deviation of 1.

- Formula for Standardized scaling:

X_standardized = (X - μ) / σ

- μ **(mu):** Mean of the variable

- σ **(sigma):** Standard deviation of the variable

- Advantages: Handles outliers better, useful when the distribution's shape is not critical, and for algorithms sensitive to scale.

- Disadvantages: Alters the original distribution shape.

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

Ans:

The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple regression analysis. It quantifies how much the variance of an estimated regression coefficient is increased due to multicollinearity among the predictor variables. A VIF value of infinity, or "infinite VIF," typically occurs when there is a perfect linear relationship (perfect multicollinearity) among the predictor variables.

Here's why infinite VIF can happen and what it indicates:

1. Perfect Multicollinearity:

- Infinite VIF arises when two or more predictor variables in the regression model are perfectly correlated or can be expressed as exact linear combinations of each other.

- For example, if you have two predictor variables X1 and X2, and X2 is equal to 2*X1 for every observation, then there is a perfect linear relationship between them.

2. Mathematical Consequences:

- When there's perfect multicollinearity, the ordinary least squares (OLS) regression algorithm cannot estimate unique coefficients for the correlated variables.

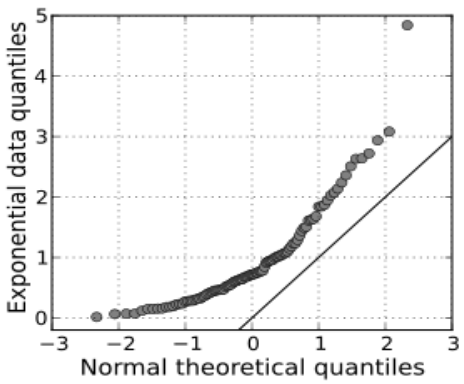
- As a result, the estimated variance of the coefficients becomes infinite, leading to an infinite VIF.

3. Impact on the Model:

- Infinite VIF is a severe issue because it indicates that the model's coefficients are not identifiable, and the model cannot be effectively estimated or interpreted.

- This situation makes it impossible to assess the individual effects of the correlated variables on the dependent variable.

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



Ans:

A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It is a powerful visualization technique that helps you compare the distribution of your data to an expected or

idealized distribution. Q-Q plots are widely used in various statistical analyses, including linear regression, for several purposes:

Use and Importance of Q-Q Plot in Linear Regression:

1. Checking Normality Assumption:

- In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) follow a normal distribution. This assumption is crucial for the validity of statistical inference and hypothesis testing.

- Q-Q plots are used to visually assess whether the residuals approximate a normal distribution. If the points on the Q-Q plot closely follow a straight line, it suggests that the residuals are normally distributed. Deviations from the line indicate non-normality.

2. Identifying Departures from Normality:

- Q-Q plots can reveal specific departures from normality. For example, if the points in the Q-Q plot curve upwards or downwards, it may indicate skewness in the data.

- If the Q-Q plot exhibits S-shaped patterns or strong deviations from the straight line in the tails, it may suggest heavy tails or kurtosis in the distribution.

3. Residual Transformation:

- If the Q-Q plot reveals non-normality in the residuals, you may need to consider transforming the residuals (e.g., log transformation) to make them more closely resemble a normal distribution.

- Transformations can improve the validity of statistical tests and confidence intervals.

4. Model Assessment:

- Q-Q plots can be used not only for checking normality of residuals but also for other diagnostic purposes in linear regression.

- They can help identify outliers or influential data points that may have a substantial impact on the model.

5. Model Assumptions:

- By using Q-Q plots, you can gain insights into whether linear regression assumptions, such as linearity and constant variance of residuals, are met.

- While Q-Q plots primarily focus on normality, they indirectly provide information about other assumptions as well.