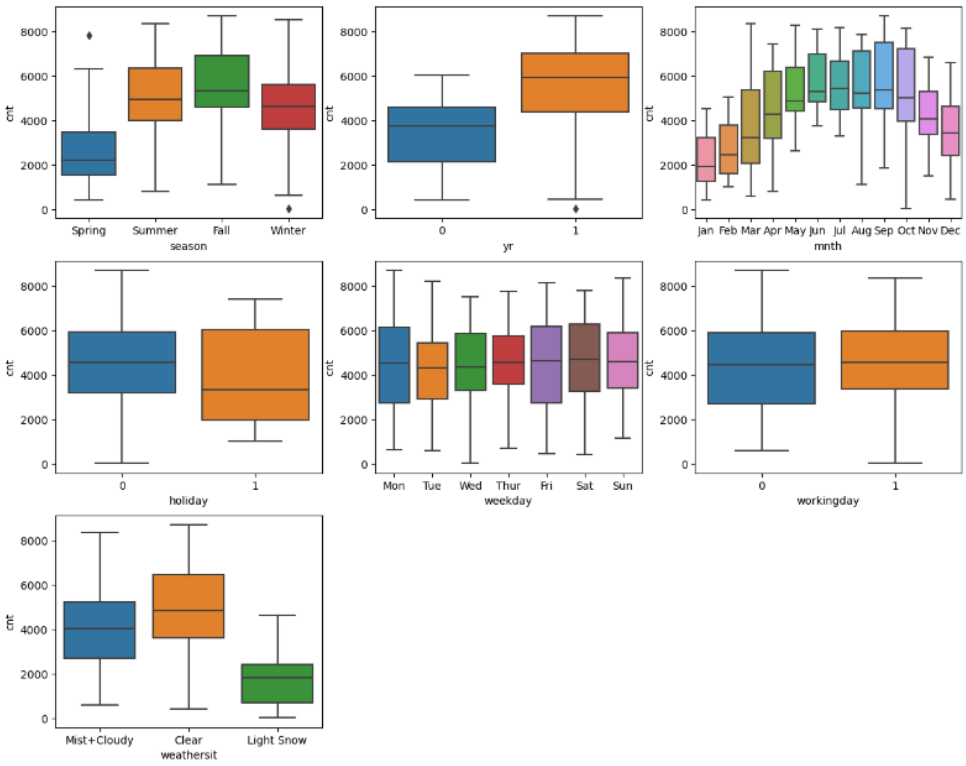
Assignment-based Subjective Questions

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

   There are total seven categorical variables in the given dataset, which are Season, Year, Month, Holiday, Weekday, Workingday and Weathersit



Observations

- **Seasons:** Demand for shared bikes varies significantly across different seasons, with Fall exhibiting the highest demand, followed by Summer, Winter, and Spring with relatively lower demand.
- **Year:** In 2019, the boxplot distribution indicates a higher demand compared to 2018, suggesting a potential yearly increase in bike sharing demand.
- **Month:** There is a consistent upward trend in demand from January to June, a dip in July, followed by an increase until September. While there is a slight drop in demand in October compared to September, demand remains relatively stable. However, a noticeable reduction in demand is observed in November and December.
- **Holiday**: On holidays, the distribution of demand is lower compared to non-holidays.
- **Weekday:** Across the days of the week, there is no significant change in demand.
- **Workingday:** The distribution of demand is higher on working days compared to non-working days.
- **Weathersit:** The demand for shared bikes is highest on clear weather days, followed by cloudy days. However, there is a significant reduction in demand during snowy weather conditions.

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

   Machine learning algorithms don't directly process categorical variables. To address this limitation, it becomes necessary to transform categorical columns into distinct columns containing binary values (0s and 1s). This process is commonly referred to as creating dummy columns using pandas, where each unique category becomes a separate column with binary indicators.

The pd.get_dummies() function is utilized to convert categorical columns into dummy columns, representing values as either 0s or 1s. Given below is the example

Unknow	Male	Female	Others
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

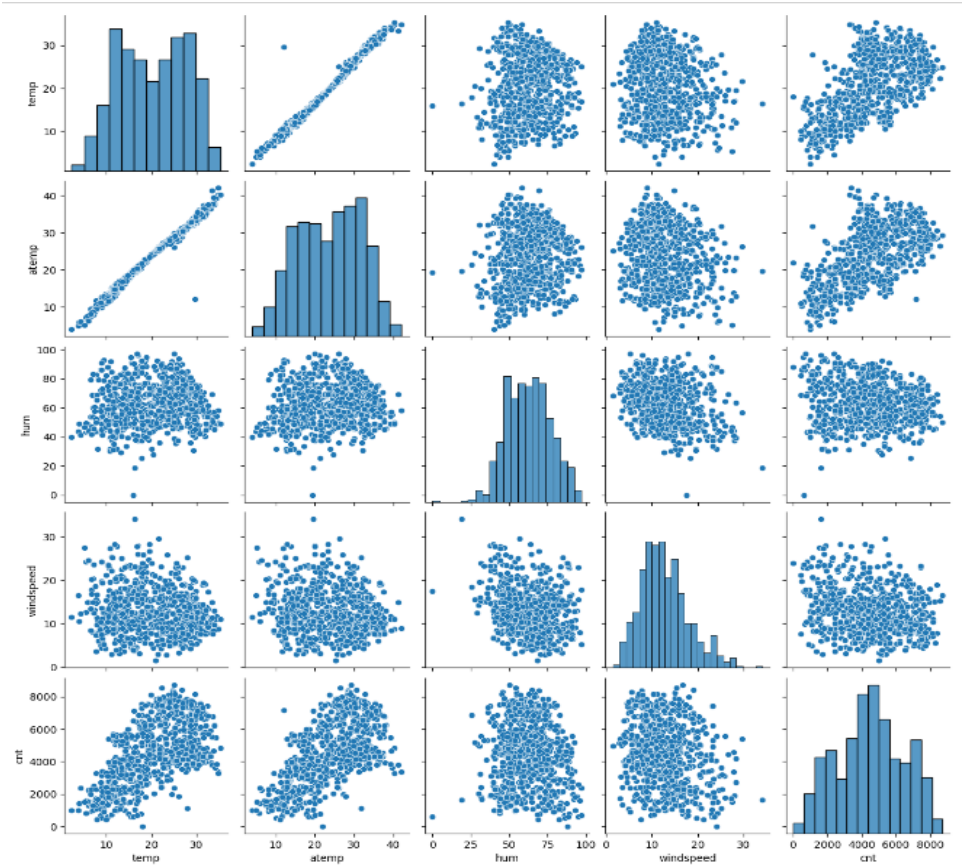
Dropping the first categorical variable "Unknown" is possible because if every other dummy column is 0, then this means your first value would have been 1.

Significance of drop_first = True

- Including all dummy variables without dropping one may result in perfect multicollinearity, as the sum of all dummy variables in a given category will consistently be constant (e.g., 1 + 0 + 0 = 1). Multicollinearity poses challenges in regression models, leading to unstable coefficients and inflated standard errors.
- In the case of iterative models, achieving convergence might be problematic. The act of dropping the first dummy variable simplifies the model by eliminating redundancy.
- This practice facilitates a more straightforward interpretation of coefficients since each coefficient now signifies the effect of a specific category relative to the omitted category.
- It contributes to computational efficiency, especially when dealing with large datasets.

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

Examining the pair plot below reveals a highest correlation between the target variable, "cnt," and the temp and atemp variables.

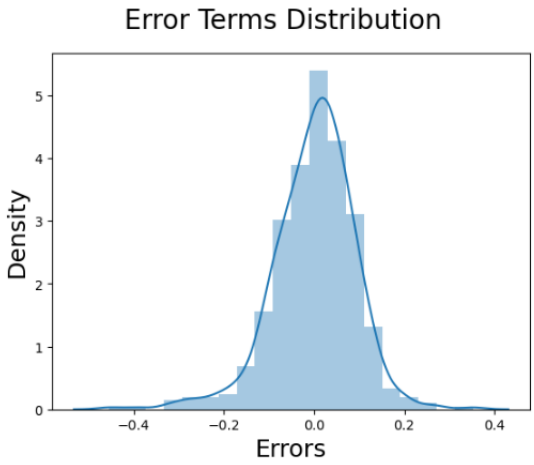


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

When constructing a linear regression model, certain assumptions are made. Let's examine whether these assumptions hold true after building the model on the training set.

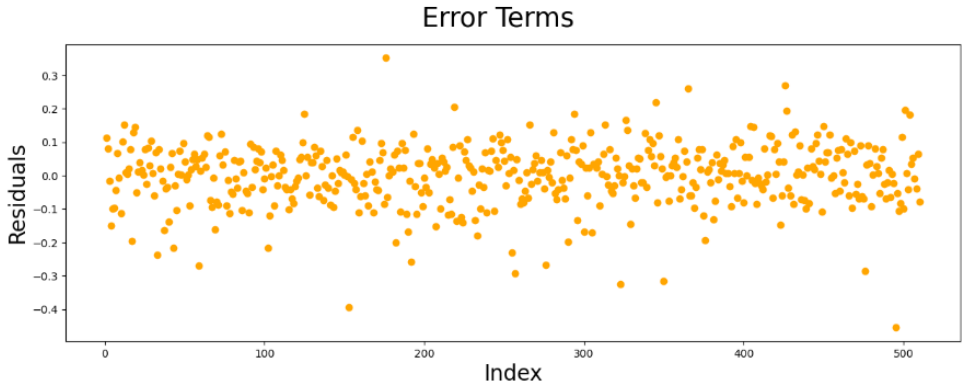
Validation 1

The image below confirm that the error terms exhibit a normal distribution with a mean of zero.



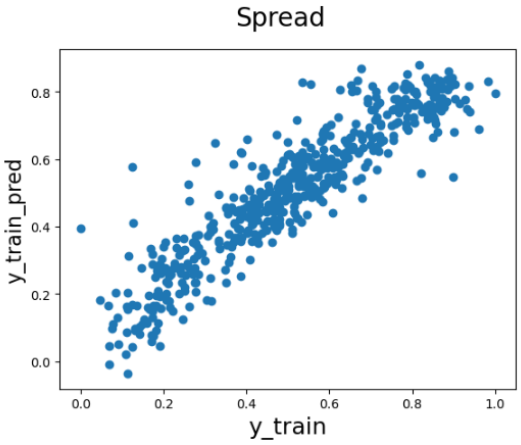
Validation 2:

The error terms demonstrate independence from one another. In the image below, no visbile patterns are evident.



Validation 3:

The image below provides confirmation that the error terms exhibit a constant variance.



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

Final linear regression equation

cnt = 0.2931 + (0.4117 * atemp) + (0.2357 * yr) + (0.058 * Winter) + (0.0557 * Sep) - (0.0501 * Nov) - (0.0531 * Dec) - (0.0562 * Jan) - (0.0598 * Jul) - (0.0881 * Mist+cloudy) - (0.0881 * holiday) - (0.1096 * Spring) - (0.1418 * Windspeed) - (0.2912 * Light Snow)

The three primary features that play a substantial role in explaining the demand for shared bikes are as follows:

- "atemp" with the coefficient magnitude of 0.4117
- "Light Snow" with the coefficient magnitude of 0.2912
- "year" with the coefficient magnitude of 0.2357

General Subjective Questions

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

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.

Assumptions:

- Linearity: The independent and dependent variables have a linear relationship with one another.
- Independence: The observations in the dataset are independent of each other.
- Homoscedasticity: Across all levels of the independent variables, the variance of the errors is constant..
- Normality: The residuals should be normally distributed. This means that the residuals should follow a bell-shaped curve.

There are two main types of linear regression:

Simple Linear Regression: This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$Y = \beta_0 + \beta_1 X$

Multiple Linear Regression: This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + \beta_n X_n$

Where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the coefficients

Best Fit lIne

The primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values is minimum. The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The

slope of the line indicates how much the dependent variable changes for a unit change in the independent variables.

Cost Function

The cost function is instrumental in determining optimal values for coefficients such as β0, β1, β2, etc., to achieve the best fit line for the given data points. This best fit line is attained by minimizing the cost function

The cost function or the loss function is nothing but the difference between the predicted value Y and the true value Y. It is the Residual Sum of squares (RSS) between the predicted value and the true value. The cost function can be written as:

Cost Function = $\sum_{n}^{i} (Y_{pred} - Y_{act})^2$

Gradient Descent for Linear Regression

A linear regression model can be trained using the optimization algorithm gradient descent by iteratively modifying the model's parameters to reduce the Residual Sum of squares (RSS) of the model on a training dataset.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that exhibit nearly identical basic statistical characteristics, yet they possess distinctive features that can mislead a regression model when visualized individually. Originating in 1973, statistician Francis Anscombe created this quartet to emphasize the significance of graphing data as a preliminary step before analysis and model development. The quartet underscores the potential pitfalls of relying solely on simple descriptive statistics without considering the nuanced patterns revealed through visualization.

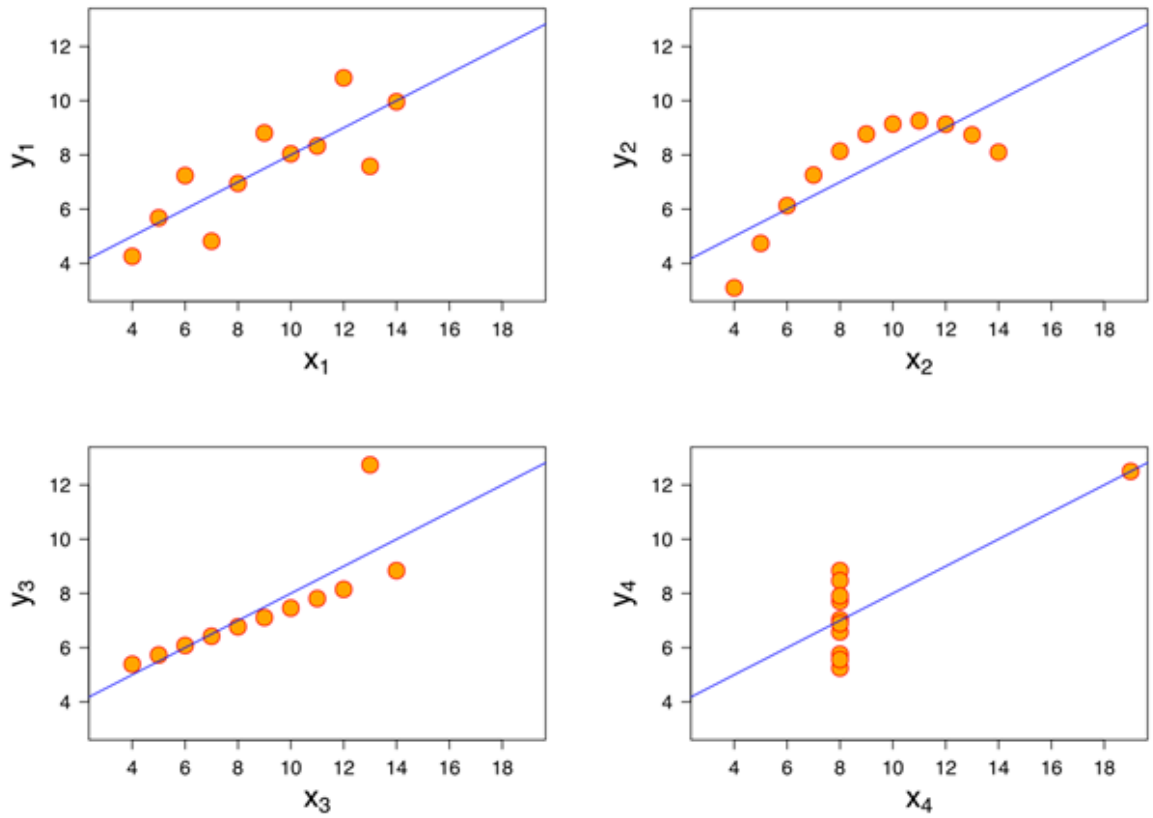
Purpose of Anscombe's Quartet in Data Visualization

Anscombe's quartet emphasizes the need to visualize data before employing algorithms, highlighting the importance of plotting data to identify anomalies. It underscores that linear regression suits linear relationships but is unsuitable for other dataset types.

The statistical information for these four data sets are approximately similar.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

When these models are plotted on a scatter plot, each dataset produces a distinct plot that cannot be interpreted by any regression algorithm.



We can describe the four data sets as:

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

In simple words, it measures the strength and direction of the linear relationship between two continuous variables.

Strength in a relationship is indicated by values close to +1 or -1, reflecting a strong connection between two factors. The proximity of data points to the line strengthens this correlation, while scattered points weaken it.

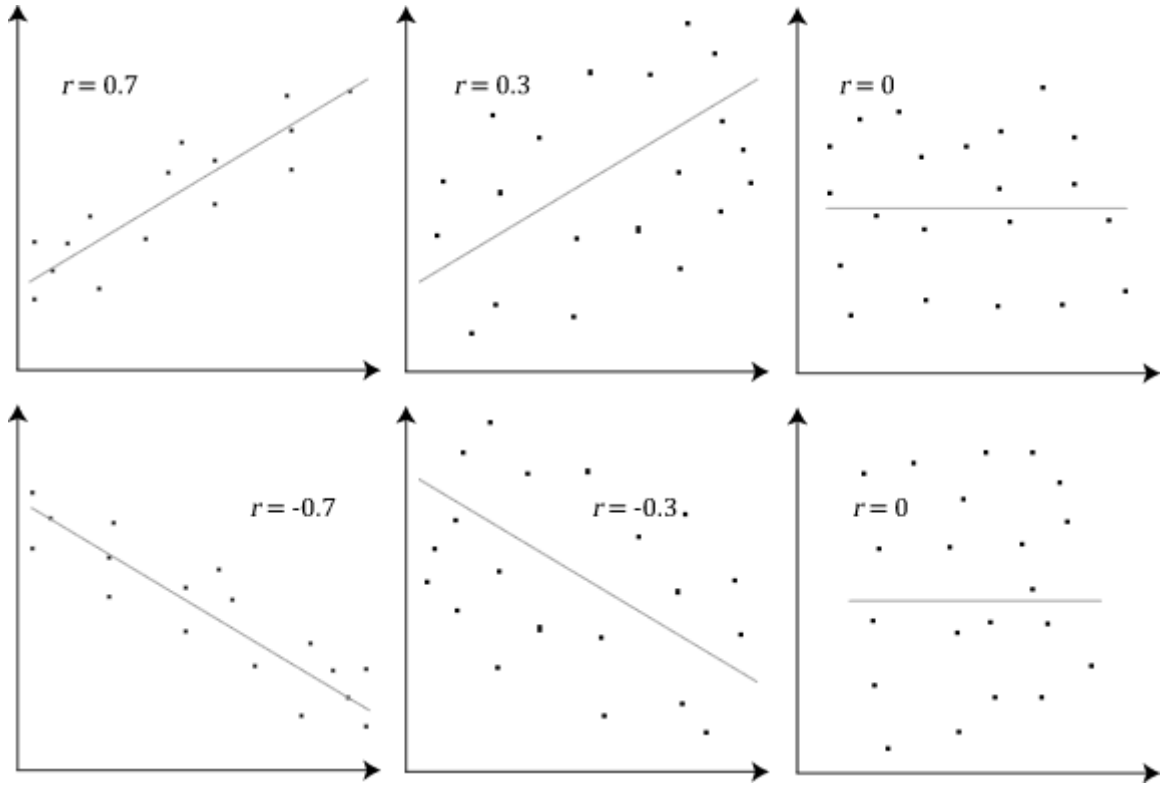
Direction refers to the slope of the line: an upward slope signifies a positive relationship, where an increase in one variable corresponds to an increase in the other, while a downward slope indicates a negative relationship, where an increase in one variable leads to a decrease in the other.

Formula

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

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

Examples of Pearson correlation coefficient



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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Methods for Scaling

- **Normalization**

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

It is really affected by outliers

- **Standardization**

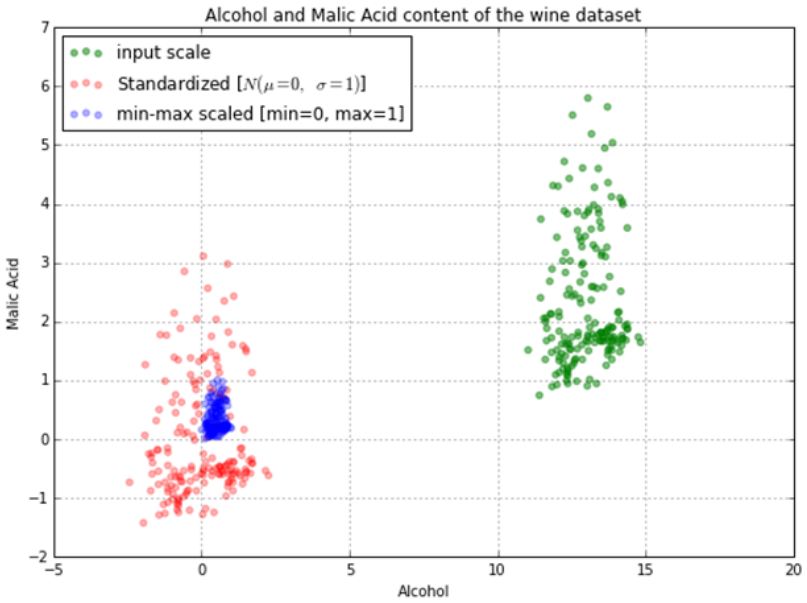
Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

It is much less affected by outliers.

Example



The impact of Standardization and Normalisation on the Wine dataset

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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a

multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

The formula for VIF is:

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

R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of R^2 means that the variable is highly correlated with the other variables.

In case where the R^2 approaches 1, resulting in an infinite Variance Inflation Factor (VIF) for a specific independent variable, it indicates that this variable can be perfectly predicted by other variables in the model.

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

The quantile-quantile or Q-Q plot is a graphical method for determining whether two samples of data came from the same population or not. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

Interpretation

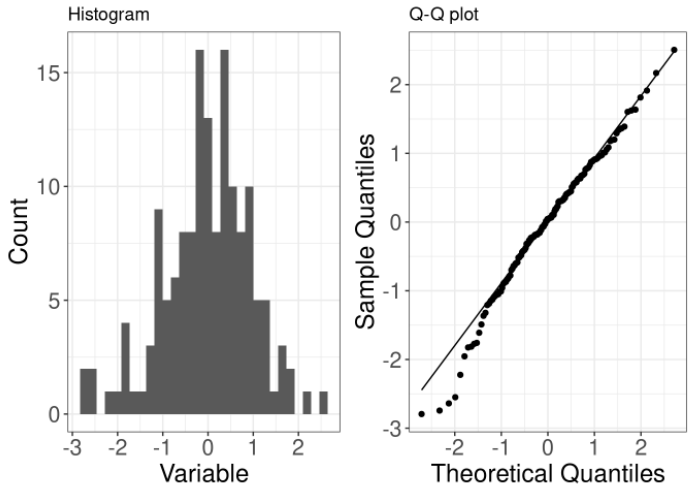
If the points lie on or close to a 45-degree line, it means that the data follow the reference distribution closely. If the points deviate from the line, it means that there are some differences between the data and the reference distribution. For example, if the points are curved, it means that the data are skewed or have heavy tails. If the points are scattered or have gaps, it means that the data have outliers.

Importance in linear regression

- Q-Q plot allows us to visually compare the distribution of residuals to a theoretical normal distribution.
- Outliers in the residuals can be detected using a Q-Q plot. Points that deviate significantly from the expected diagonal line may indicate potential outliers in the data.
- Q-Q plots can also help in assessing whether the relationship between the dependent and independent variables is truly linear. Deviations from the expected pattern in the Q-Q plot may suggest non-linearity in the relationship.
- Q-Q plots can provide insights into the homoscedasticity assumption (constant variance of residuals). If the spread of points in the Q-Q plot widens or narrows as you move along the line, it may indicate heteroscedasticity.

Normally distributed data

Below is an example of data that are drawn from a normal distribution. The normal distribution is **symmetric**, so it has no skew (the mean is equal to the median).



On a Q-Q plot normally distributed data appears as roughly a straight line (although the ends of the Q-Q plot often start to deviate from the straight line).