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?
(3 marks)

Answer:

The analysis of categorical columns through box plots and bar plots reveals several noteworthy trends:

Seasonal Booking Trends:
More bookings were attracted during the fall season, and there was a significant increase in booking counts across all seasons from 2018 to 2019.

Monthly Booking Patterns:
The majority of bookings occurred in May, June, July, August, September, and October. The booking trend showed a rise from the beginning of the year to the middle, followed by a decrease towards the year-end.

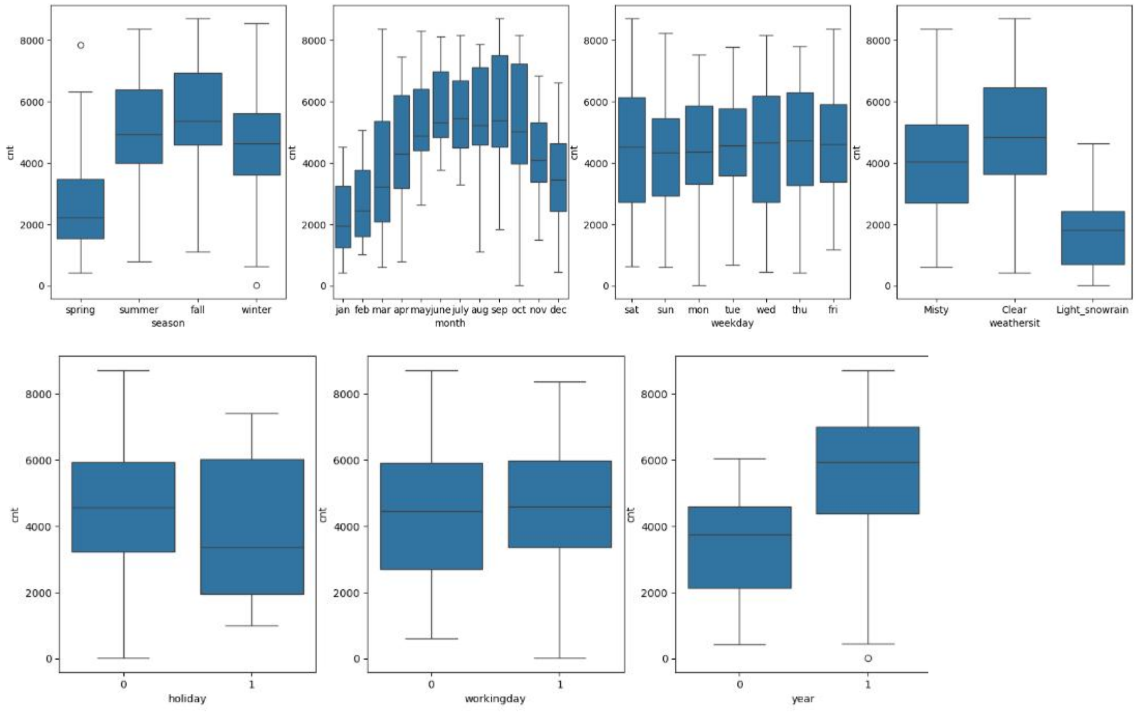
Weather Impact:
Clear weather was associated with a higher number of bookings, aligning with expectations.

Day-wise Booking Patterns:
Thursday, Friday, Saturday, and Sunday exhibited higher booking counts compared to the weekdays.

Holiday Influence:
Bookings were observed to be lower on non-holidays, which is logical as people may prefer spending time at home with family during non-holiday periods.

Working Day vs. Non-Working Day:
Booking counts appeared to be relatively equal on both working and non-working days.

Yearly Progress:
The year 2019 witnessed a greater number of bookings compared to the previous year, indicating positive business growth.



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

Answer:

Using drop_first=True is crucial in dummy variable creation as it effectively reduces the number of generated columns, specifically by omitting one dummy variable to mitigate correlations among them. This parameter is valuable in scenarios where a categorical column with k levels is transformed into dummy variables. By excluding the first level (k-1 dummies out of k), it helps prevent multicollinearity issues. For instance, if there are three types of values in a categorical column, dropping the first dummy variable suffices to identify the third type, given the absence of the other two.

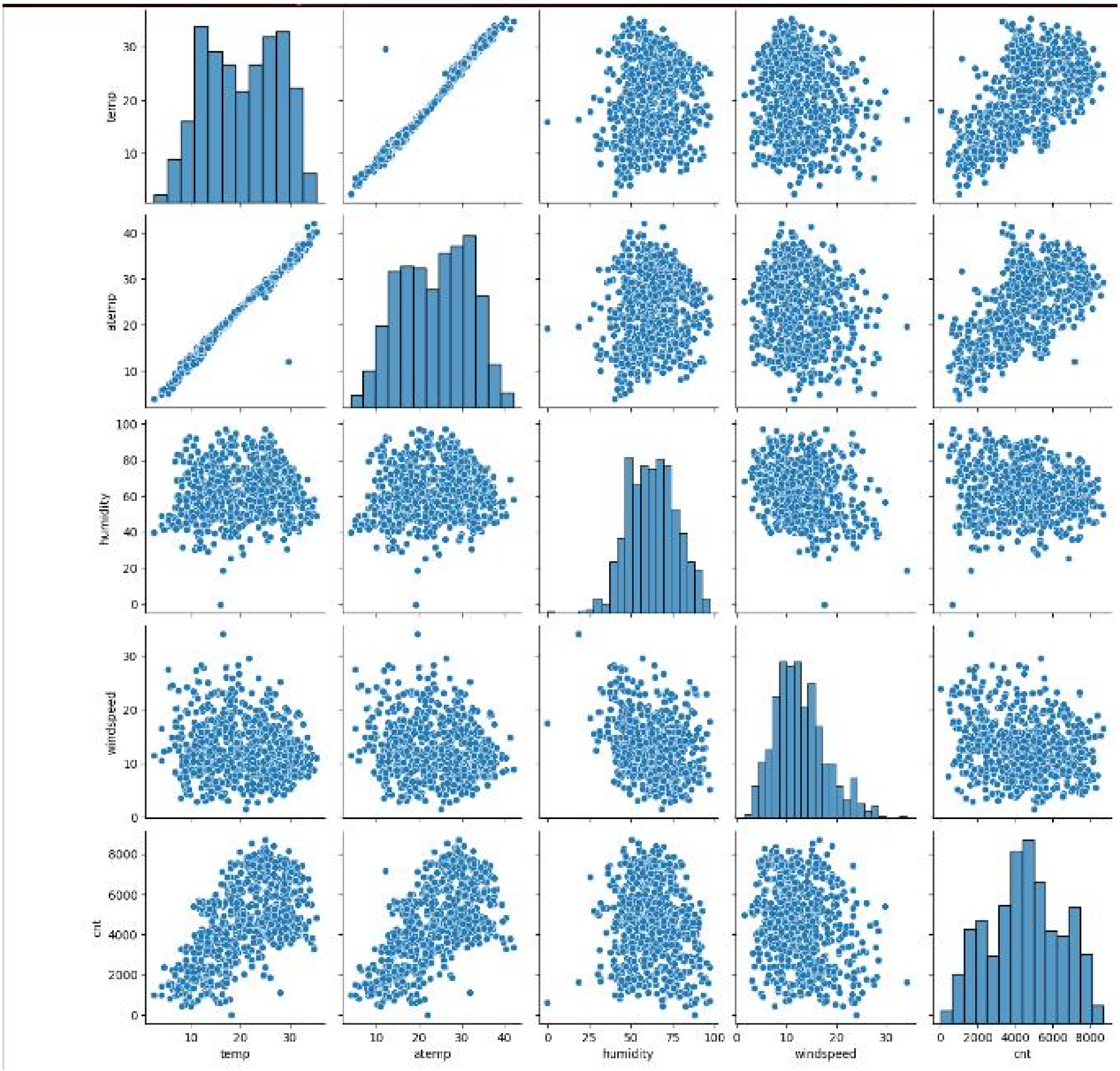
The syntax for this is

drop_first: bool, default False,
indicating whether to exclude the first level during dummy variable creation.

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

Answer:

The variable 'temp' exhibits the highest correlation with the target variable.



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
(3 marks)
Answer:
The evaluation of Linear Regression Model's compliance with the following five presumptions:

Error word normality:
A normal distribution should be followed by error terms.

Verify multicollinearity:
The variables shouldn't significantly multicollinear.

Validation of linear relationships:
There should be a clear linear relationship between the variables.

Homoscedasticity:
There should be no pattern in residual values.

Residuals' independence:
The residuals shouldn't exhibit any auto-correlation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
(2 marks)
Answer:
The following are the top 3 characteristics that greatly contribute to the explanation of the shared bike demand:

1. temp
2. winter
3. sept.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
(4 marks)
Answer:
Linear regression is a statistical model that examines the linear relationship between a dependent variable and a given set of independent variables. In this context, a linear relationship implies that as the value of one or more independent variables changes (increases or decreases), the dependent variable's value will also change accordingly (increase or decrease).
Mathematically, this relationship is represented by the equation:

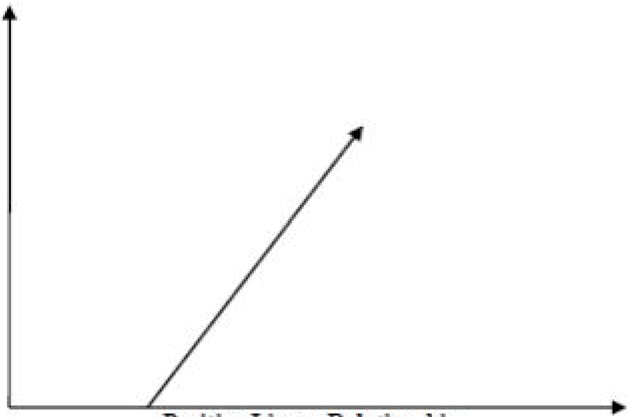
Y=mX+c

Here,

Y is the dependent variable we aim to predict.
X is the independent variable used for predictions.
m is the slope of the regression line, representing the effect X has on Y.
c is a constant known as the Y-intercept. If X=0, Y would be equal to c.

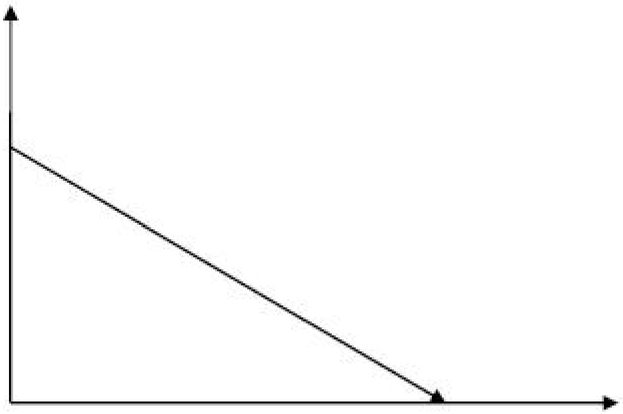
The linear relationship can be either positive or negative:

Positive Linear Relationship: Both independent and dependent variables increase.



Positive Linear Relationship

Negative Linear Relationship: Independent variable increases while the dependent variable decreases.



Negative Linear Relationship

Linear regression comes in two forms:

Simple Linear Regression
Multiple Linear Regression

There are several assumptions made by the Linear Regression model about the dataset, including:

Multi-collinearity: Assumes little or no multi-collinearity in the data, meaning that independent variables or features should have minimal dependency among them.

Auto-correlation: Assumes little or no auto-correlation in the data, meaning that there is minimal dependency between residual errors.

Relationship between variables: Assumes that the relationship between response and feature variables must be linear.

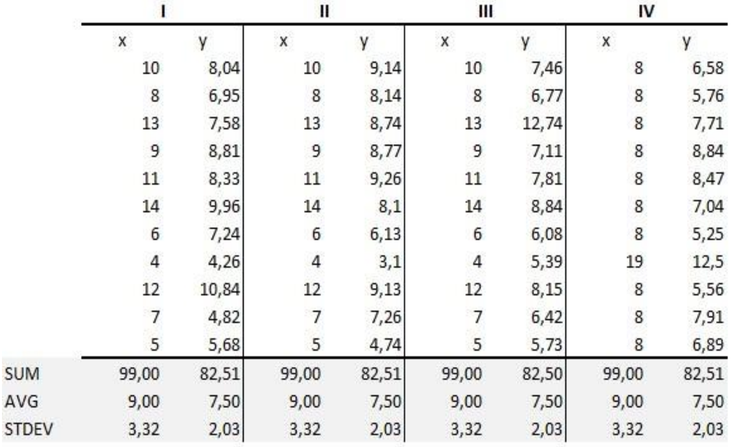
Normality of error terms: Error terms should be normally distributed.

Homoscedasticity: There should be no visible pattern in residual values.

2. Explain the Anscombe’s quartet in detail.
(3 marks)

Answer:

Anscombe’s Quartet, developed by the statistician Francis Anscombe, consists of four datasets, each containing eleven (x, y) pairs. What makes these datasets remarkable is that they share identical descriptive statistics. However, the real distinction, and I must stress, a significant one, becomes apparent when these datasets are graphed. Despite having similar summary statistics, each graph tells a different story.

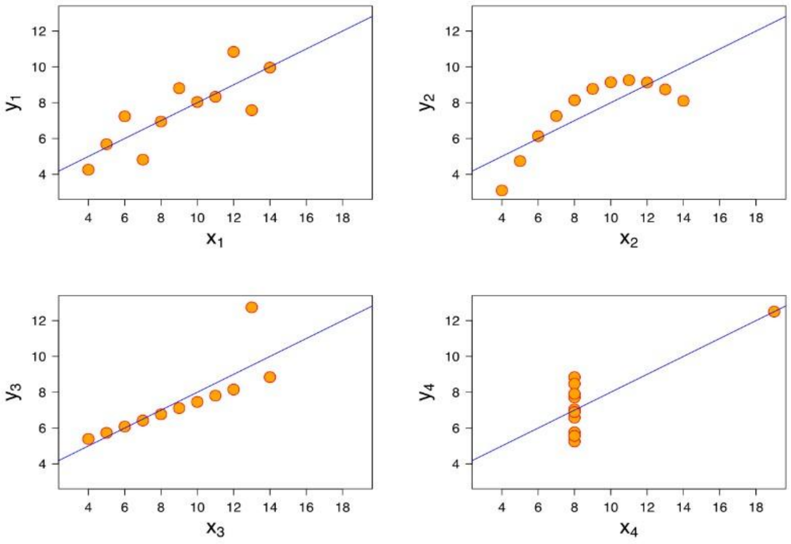


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics indicate that the means and variances for both x and y are the same across the groups:

- The mean of x is 9, and the mean of y is 7.50 for each dataset.
- Likewise, the variance of x is 11, and the variance of y is 4.13 for each dataset.
- The correlation coefficient (indicating the strength of the relationship between two variables) between x and y is 0.816 for each dataset.

When plotted on an x/y coordinate plane, these datasets exhibit the same regression lines. However, each dataset narrates a distinct story:



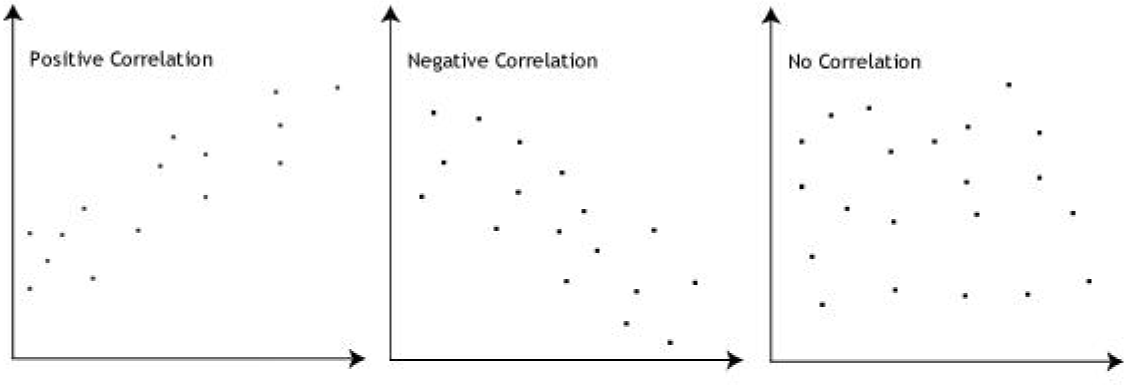
Dataset I portrays clean and well-fitting linear models.
Dataset II does not follow a normal distribution.
Dataset III, the distribution is linear, but the calculated regression is affected by an outlier.
Dataset IV demonstrates that a single outlier can generate a high correlation coefficient.

This quartet underscores the vital role of visualization in data analysis. Examining the data visually reveals the structure and provides a clear understanding of the dataset.

3. What is Pearson’s R?
(3 marks)
Answer:
A numerical representation of the degree of the linear relationship between variables is provided by Pearson's r. The correlation coefficient is positive when there is a tendency for the variables to rise and fall together. On the other hand, the correlation coefficient is negative if the variables have a tendency to move in the opposite directions, with low values of one variable correlating to high values of the other.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

"r," or the Pearson correlation coefficient, is a number between +1 and -1. There is no correlation between the two variables when the value is 0. A positive correlation is denoted by a value larger than 0, which means that as one variable rises, the other also does. Conversely, a value less than 0 denotes a negative correlation, meaning that when one variable rises, the other falls. The graphic below provides an illustration of this relationship:

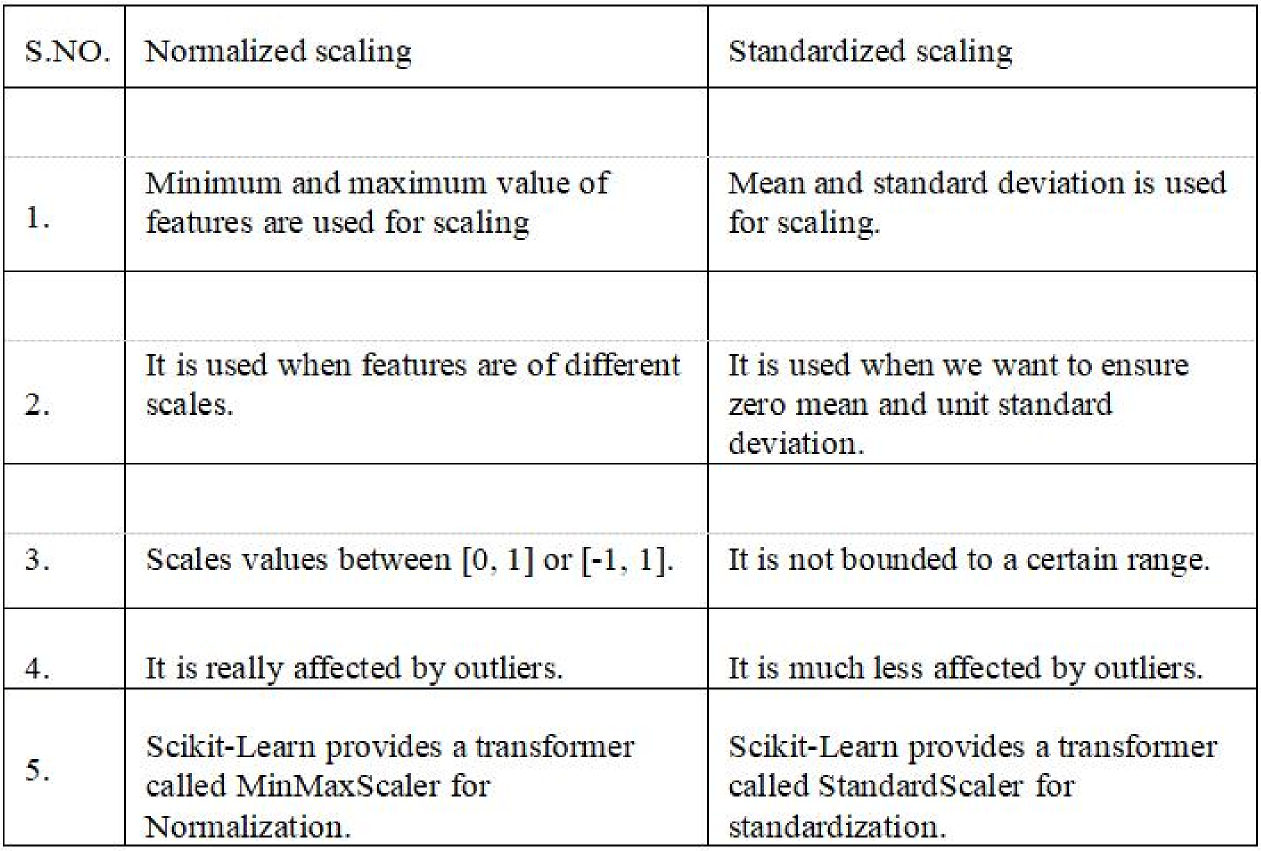


4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
(3 marks)
Answer:

A data pre-processing method called feature scaling is used to bring the independent features of a dataset into a uniform range. This technique is used to handle differences in the amounts, values, or units that are contained in the data. If feature scaling is not done, machine learning algorithms could give more weight to larger values and less weight to smaller ones, regardless of the units of measurement used.

Take an algorithm that does not do feature scaling as an example. It might erroneously interpret 3000 meters as being more than 5 kilometers. Actually, this is not correct. By bringing all values to the same magnitudes, feature scaling is used to reduce such disparities. This guarantees that the algorithm accurately understands the input and produces more precise predictions.

Difference between normalized scaling and standardized scaling



S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

The Variance Inflation Factor (VIF) is infinite when there is perfect correlation between the variables. A high VIF number denotes strong correlation between the variables; if the VIF is 4, for instance, it implies that multicollinearity has inflated the model coefficient's variance by a factor of 4.

A condition of perfect correlation between two independent variables is shown when the VIF approaches infinity. When this occurs, R-squared (R2) = 1, which makes 1 / (1 - R2) endless. In order to resolve this problem, one of the factors generating this perfect multicollinearity needs to be removed from the dataset.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

A graphical method for determining whether two data sets come from populations with a common distribution is the quantile-quantile (q-q) plot.

Application of Q-Q Plot:

The quantiles of the first and second data sets are compared using a Q-Q plot. The proportion or percentage of data points below a specified value is represented by quantiles. For example, the data point where 30% of the data falls below and 70% falls above the 0.3 quantile is represented by that value. The plot includes a reference line that is 45 degrees. Should the distributions of the two data sets be similar, then the points ought to be about in line with this reference line. There is more proof that the two data sets come from separate distributions when there is a deviation from the reference line.

Significance of Q-Q Plot:

It's critical to evaluate the assumption of a common distribution when there are two available data samples. Verifying a common distribution enables the pooling of both data sets for the purpose of estimating scale and common location characteristics. In the event that there are differences between the samples, the q-q plot sheds light on the nature of those differences more insightfully than analytical techniques like chi-square and Kolmogorov-Smirnov 2-sample tests do. Should the distributions of the two data sets be similar, then the points ought to be about in line with this reference line. There is more proof that the two data sets come from separate distributions when there is a deviation from the reference line.