SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

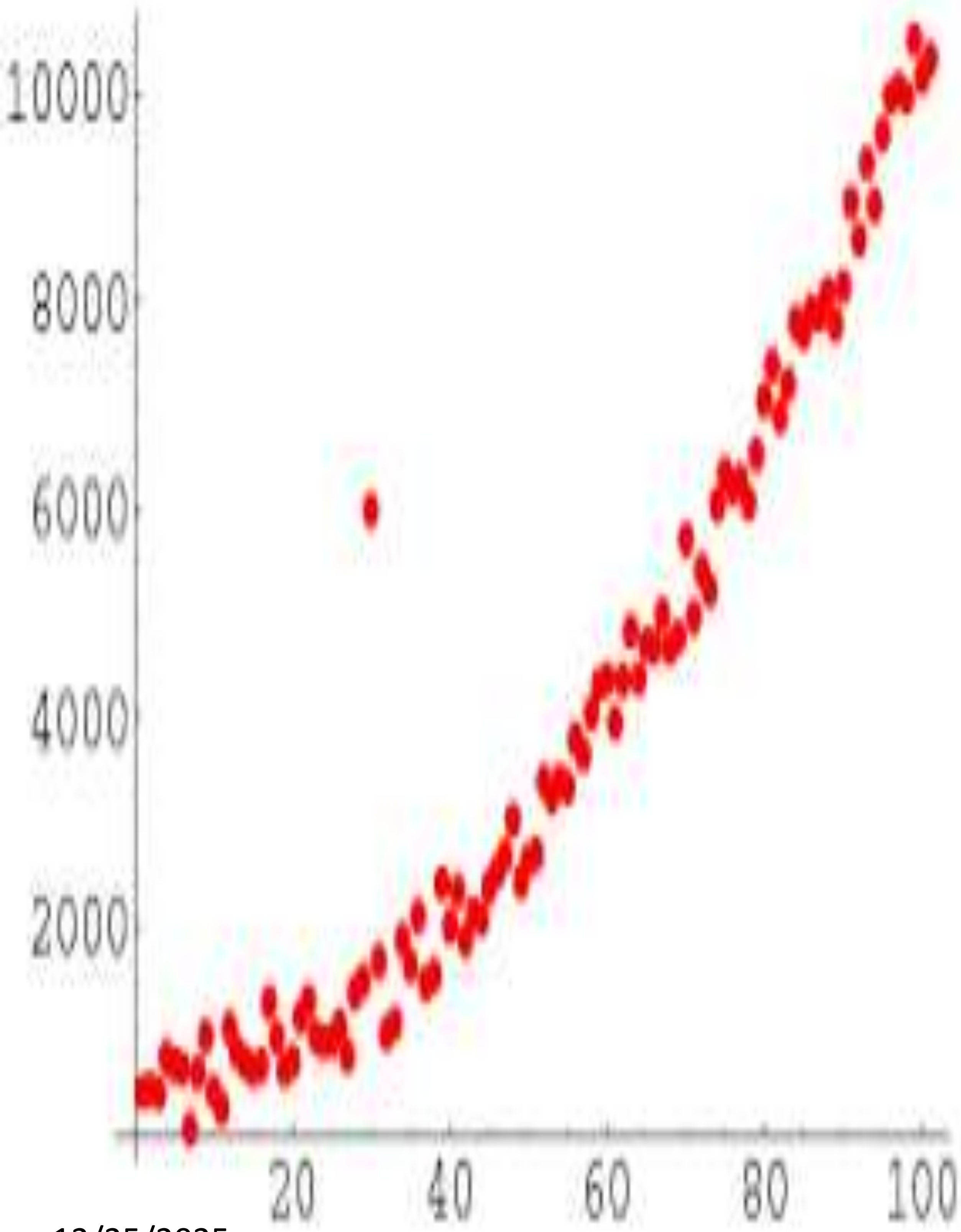
UNIT II – OUTLIERS

Outliers

EMPATHY:

- Outliers in data science are data points that differ significantly from the rest of the dataset, lying at an abnormal distance from other observations.

Made with GAMMA



What is Outlier?

- Outliers, in the context of information evaluation, are information points that deviate significantly from the observations in a dataset.
- For instance, in a dataset of monthly sales figures, if the income for one month are extensively higher than the sales for all of the different months, that high sales determine would be considered an outlier.

Why Removing Outliers is Necessary?

Impact on Analysis:

- Removing unusual values helps the analysis reflect the typical data more clearly.

Statistical Significance:

- Extreme values can change the accuracy of conclusions drawn from data.
- Removing such values, when appropriate, improves the quality of the analysis.

DEFINE:

Types of Outliers:

Univariate Outliers:

- These outliers occur when a value in one variable differs greatly from the rest of the dataset.
- For example, if you're reading the heights of adults in a sure place and most fall in the variety of 5 feet 5 inches to 6 ft, a person who measures 7 toes tall might be taken into consideration a univariate outlier.

IDEATE:

Multivariate Outliers:

- Multivariate outliers are data points that are unusual across more than one variable at the same time, showing complex relationships in the data.

- Continuing with our example, consider evaluating height and weight, and you discover a character who's especially tall and relatively heavy in comparison to the relaxation of the populace.

- This character would be taken into consideration a multivariate outlier, as their characteristics in each height and weight concurrently deviate from the normal.

Point Outliers:

- These are the points which might be far eliminated from the rest of the points.
- For instance, in a dataset of common household energy utilization, a price this is exceptionally excessive or low as compared to the relaxation is a point outlier.

Contextual Outliers:

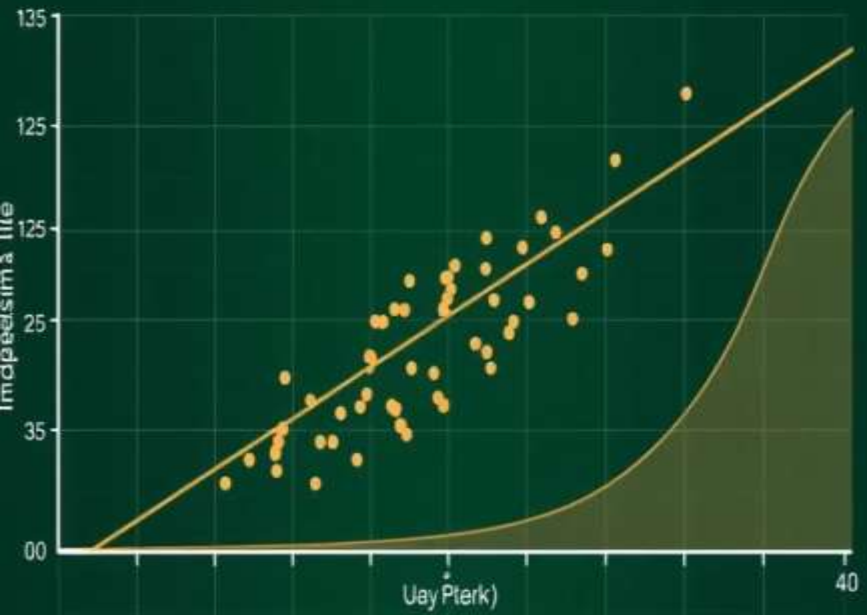
- Sometimes known as conditional outliers, these are facts factors that deviate from the normal only in a specific context or condition.
- For instance, a very low temperature might be regular in wintry weather but unusual in summer.

Collective Outliers:

- These outliers consist of a set of data factors that might not be excessive by means of themselves however are unusual as an entire.
- This type of outlier regularly shows a change in information behavior or emergent phenomena.

TESTING:

Main Causes of Outliers



Data Entry Errors: Simple human errors in entering data can create extreme values.

Measurement Error: Faulty device or experimental setup problems can cause abnormally high or low readings.

Experimental Errors: Flaws in experimental design might produce facts factors that do not represent what they're presupposed to degree.

Intentional Outliers: In some cases, data might be manipulated deliberately to produce outlier effects, often seen in fraud cases.

Data Processing Errors: During the collection and processing stages, technical glitches can introduce erroneous data.

Natural Variation: Inherent variability in the underlying data can also lead to outliers.

PROTOTYPING:

How Outliers can be Identified?

1. Outlier Identification Using Visualizations

- Visualizations offers insights into information distributions and anomalies.
- Visual tools like with scatter plots and box plots, can efficaciously spotlight information factors that deviate notably from the majority.
- In a scatter plot, outliers often seem as records factors mendacity far from the primary cluster or displaying unusual styles as compared to the relaxation.
- Box plots offer a clean depiction of the facts central tendency and spread, with outliers represented as person factors beyond the whiskers.

When Should You Remove Outliers?

- The decision to remove outliers depends on the purpose of the analysis.
- Outliers should be removed when they result from errors or unusual events and do not reflect the true nature of the data.