SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 4: Data Cleaning and Normalization

Data Cleaning

- Cleaning raw input data is often the most important, and time consuming, part as a data scientist.
- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

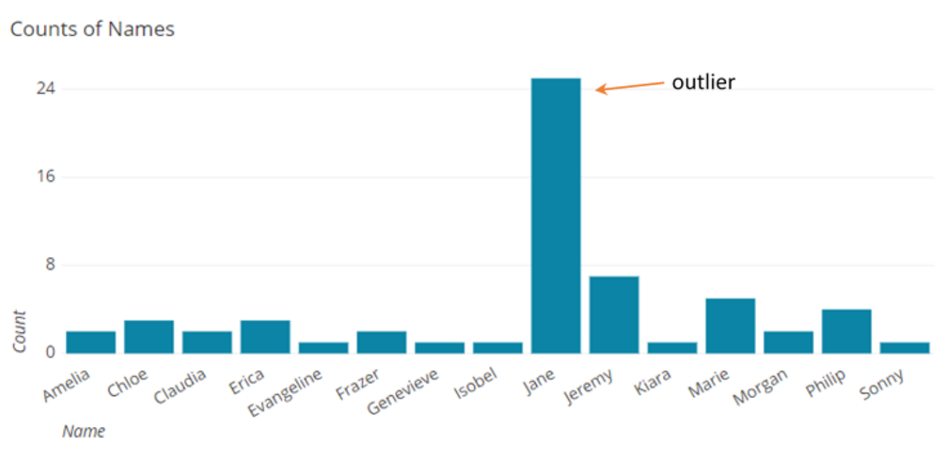
Why data cleaning is important?

- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.
- If data cleaning process is not done, then it's going to skew the results, and it will ultimately end up in the wrong decisions.
- If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

Data Cleaning

There are a lot of different kinds of problems and data that you need to watch out for:

- Outliers,
- Missing Data
- Erroneous data
- Irrelevant data
- Inconsistent data
- Formatting



Outliers

- Outliers are those data points that are really far from the rest of your data points. In other words an outlier is a value or data point that differs substantially from the rest of the data.
- An outlier is an extremely high or extremely low data point relative to the rest of the data points in a dataset.
- They can show up due to errors in data entry or measurement, or just because there's variation in the the population you're looking at and you happened to see one of the more unusual values.

Data Cleaning

Missing Data
A missing value can signify a number of different things. Perhaps the field was not applicable, the event did not happen, or the data was not available.

Erroneous data – Wrong data, invalid data that the program cannot process and should not accept.

Irrelevant data - Irrelevant data are those that are not actually needed, and don't fit under the context of the problem we're trying to solve.

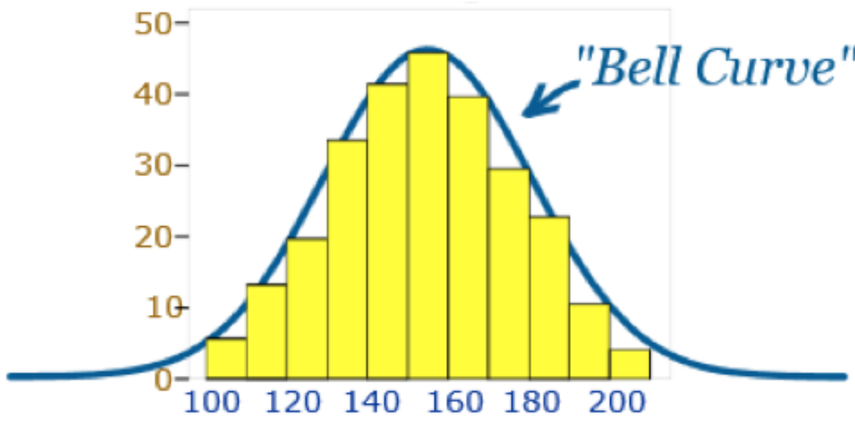
Inconsistent data - Data inconsistency is a situation where there are multiple tables within a database that deal with the same data but may receive it from different inputs.

Formatting - Data can be inconsistently formatted. Take the example of dates: in the US we always do month, day, year (MM/DD/YY), but in other countries they might do day, month, year (DD/MM/YY).

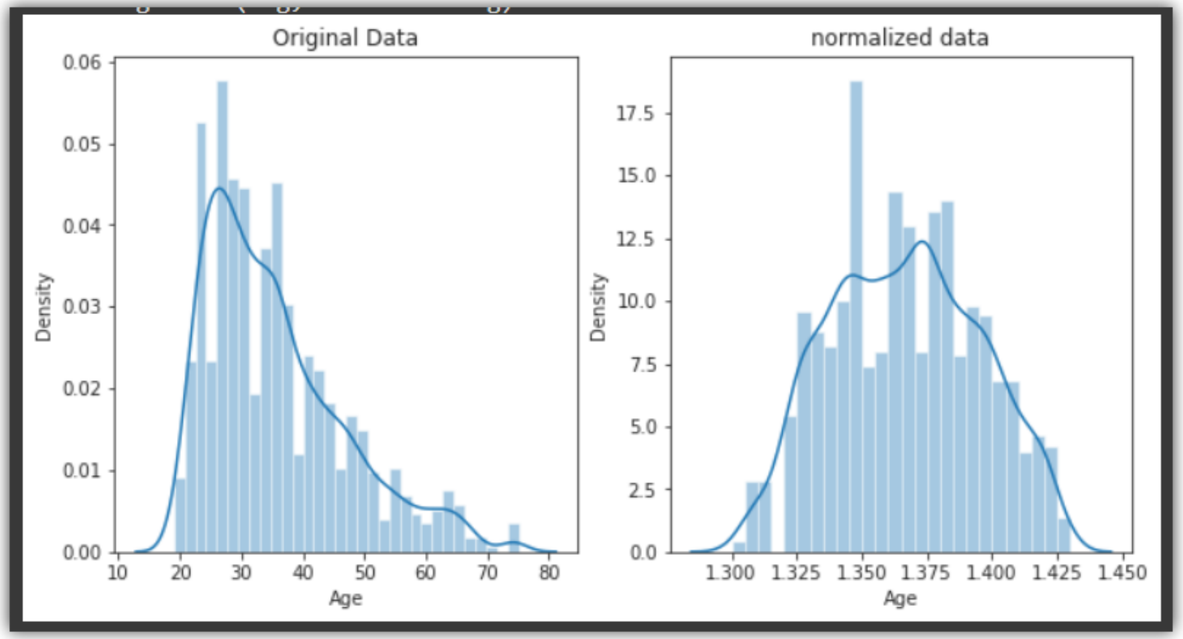
Data Normalization

- Normalization is a data preparation technique that is frequently used in machine learning.
- The process of transforming the columns in a dataset to the same scale is referred to as normalization.
- Every dataset does not need to be normalized for machine learning.
- It is only required when the ranges of characteristics are different.
- Normalization is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0.
- Normalization helps to improve the performance as well as the accuracy of your model better.
- It will not affect regression model that much but it needed for linear discriminant analysis (LDA) and Gaussian naive Bayes.
- It is useful when the feature distribution of data does not follow a Gaussian (bell curve) distribution.

Data Normalization



person_name	Salary	Year_of_ experience
Aman	100000	10
Abhinav	78000	7
Ashutosh	32000	5
Dishi	55000	6
Abhishek	92000	8
Avantika	120000	15
Ayushi	65750	7



Data Normalization

Normalization techniques in Machine Learning

- Although there are so many feature normalization techniques in Machine Learning, few of them are most frequently used. These are as follows:

- **Min-Max Scaling**: This technique is also referred to as scaling. The Min-Max scaling method helps the dataset to shift and rescale the values of their attributes, so they end up ranging between 0 and 1.

- **Standardization scaling**: Standardization scaling is also known as **Z-score normalization**, in which values are centered around the mean with a unit standard deviation, which means the attribute becomes zero and the resultant distribution has a unit standard deviation.

- This technique is helpful for various machine learning algorithms that use distance measures such as KNN, K-means clustering, and Principal component analysis

Detecting Outliers

Reasons for outliers in data

- Errors during data entry or a faulty measuring device (a faulty sensor may result in extreme readings).
- Natural occurrence

Box plots

- Box plots are a visual method to identify outliers. Box plots is one of the many ways to visualize data distribution.
- Box plot plots the q1 (25th percentile), q2 (50th percentile or median) and q3 (75th percentile) of the data along with (q1–1.5*(q3-q1)) and (q3+1.5*(q3-q1)).
- Outliers, if any, are plotted as points above and below the plot.

IQR method

- IQR method is used by box plot to highlight outliers. IQR stands for interquartile range, which is the difference between q3 (75th percentile) and q1 (25th percentile).

- The IQR method computes lower bound and upper bound to identify outliers.
- Lower Bound = q1–1.5*IQR
- Upper Bound = q3+1.5*IQR

Detecting Outliers

Z-score method

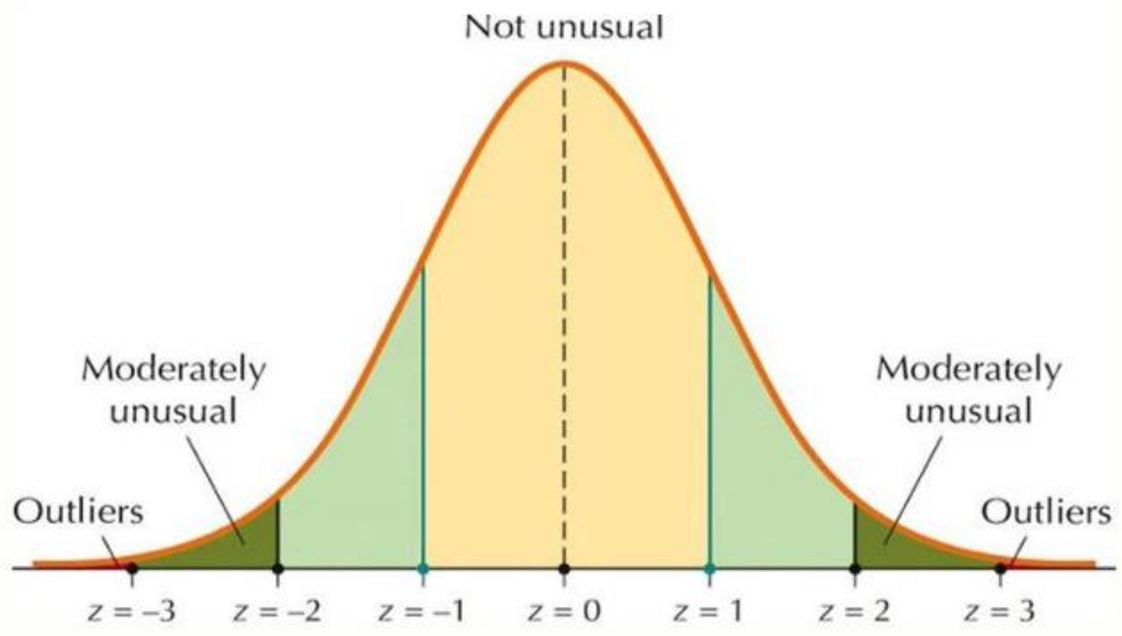
- Z-score method is another method for detecting outliers. This method is generally used when a variable' distribution looks close to Gaussian.

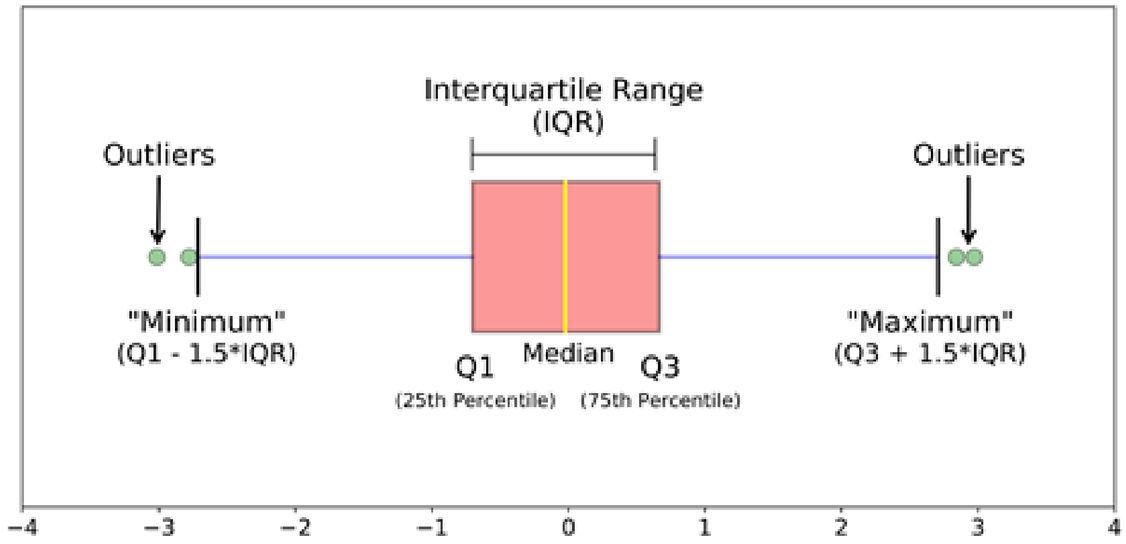
- Z-score is the number of standard deviations a value of a variable is away from the variable' mean.
  Z-Score = (X-mean) / Standard deviation

- when the values of a variable are converted to Z-scores, then the distribution of the variable is called standard normal distribution with mean=0 and standard deviation=1.

- The Z-score method requires a cut-off specified by the user, to identify outliers. The widely used lower end cut-off is -3 and the upper end cut-off is +3.

Detecting Outliers



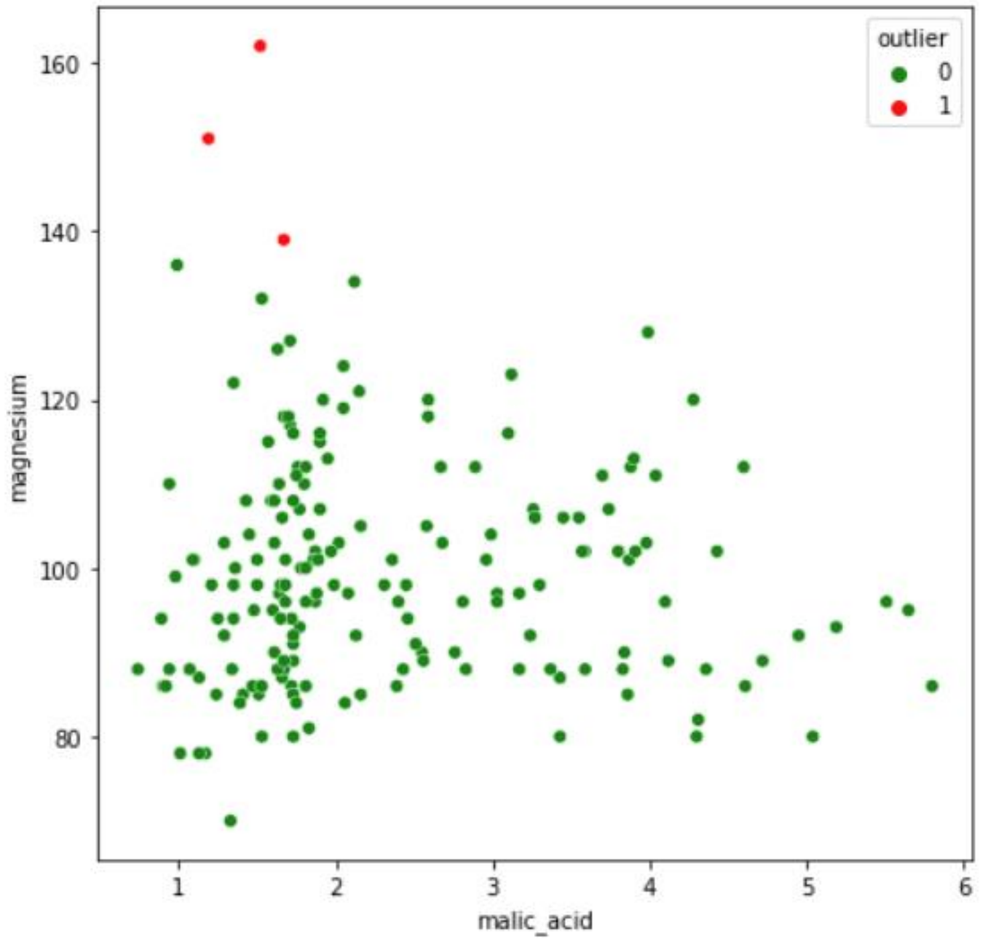


IQR to detect outliers

Detecting Outliers

Distance from the mean' method (Multivariate method)

- Unlike the previous methods, this method considers multiple variables in a data set to detect outliers.
- This method calculates the Euclidean distance of the data points from their mean and converts the distances into absolute z-scores.
- Any z-score greater than the pre-specified cut-off is considered to be an outlier.



THANK YOU