DATA GATHERING AND CLEANING

GROUP 1: BUNQUIN, COLIGADO, DUMLAO, POMPERADA, MENDEGORIN

11 June, 2024

TOPIC OVERVIEW

01

INTRODUCTION TO THE DATA GATHERING AND CLEANING PROCESSES

02

HANDLING MISSING VALUES

03

DATA TRANSFORMATION AND STANDARDIZATION

04

OUTLIER DETECTION AND TREATMENT

05

DATA INTEGRATION AND MERGING

06

DATA VALIDATION AND VERIFICATION

07

DATA CLEANING TOOLS AND AUTOMATION

08

EXPLORATORY DATA ANALYSIS

1.) INTRODUCTION TO THE DATA GATHERING AND CLEANING PROCESSES

Data cleaning - process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset

Data cleaning, also known as data cleansing or **data preprocessing**

It is a **crucial step** in the data science pipeline to improve its quality and usability

WHY DO WE NEED TO CLEAN DATA?

For decision-making, the integrity of the conclusions drawn heavily relies on the cleanliness of the underlying data.

Without proper data cleaning, inaccuracies, outliers, missing values, and inconsistencies can compromise the validity of analytical results

Raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it.

STEPS TO PERFORM DATA CLEANLINESS

01

REMOVAL OF UNWANTED OBSERVATIONS

02

FIXING STRUCTURE ERRORS

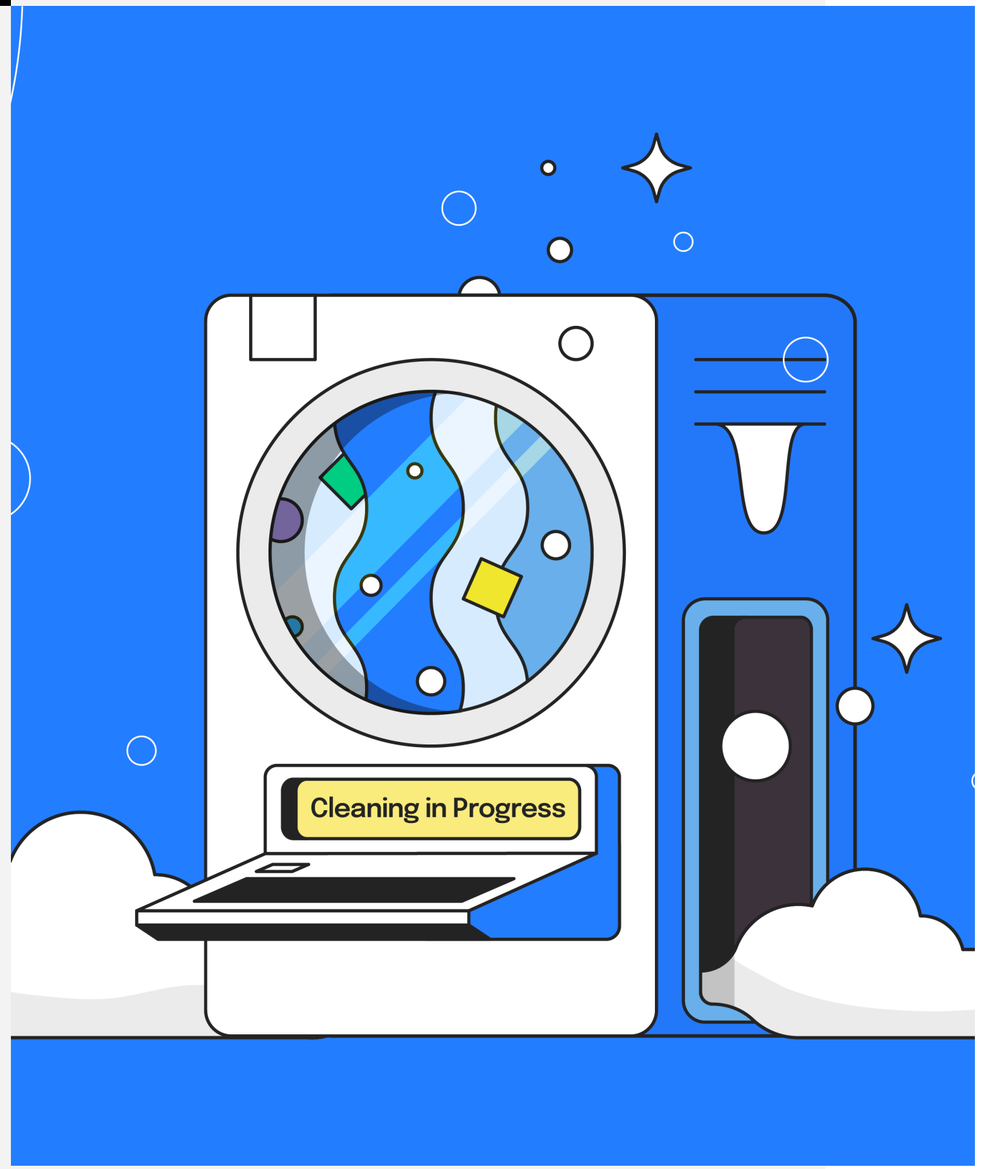
03

MANAGING UNWANTED OUTLIERS

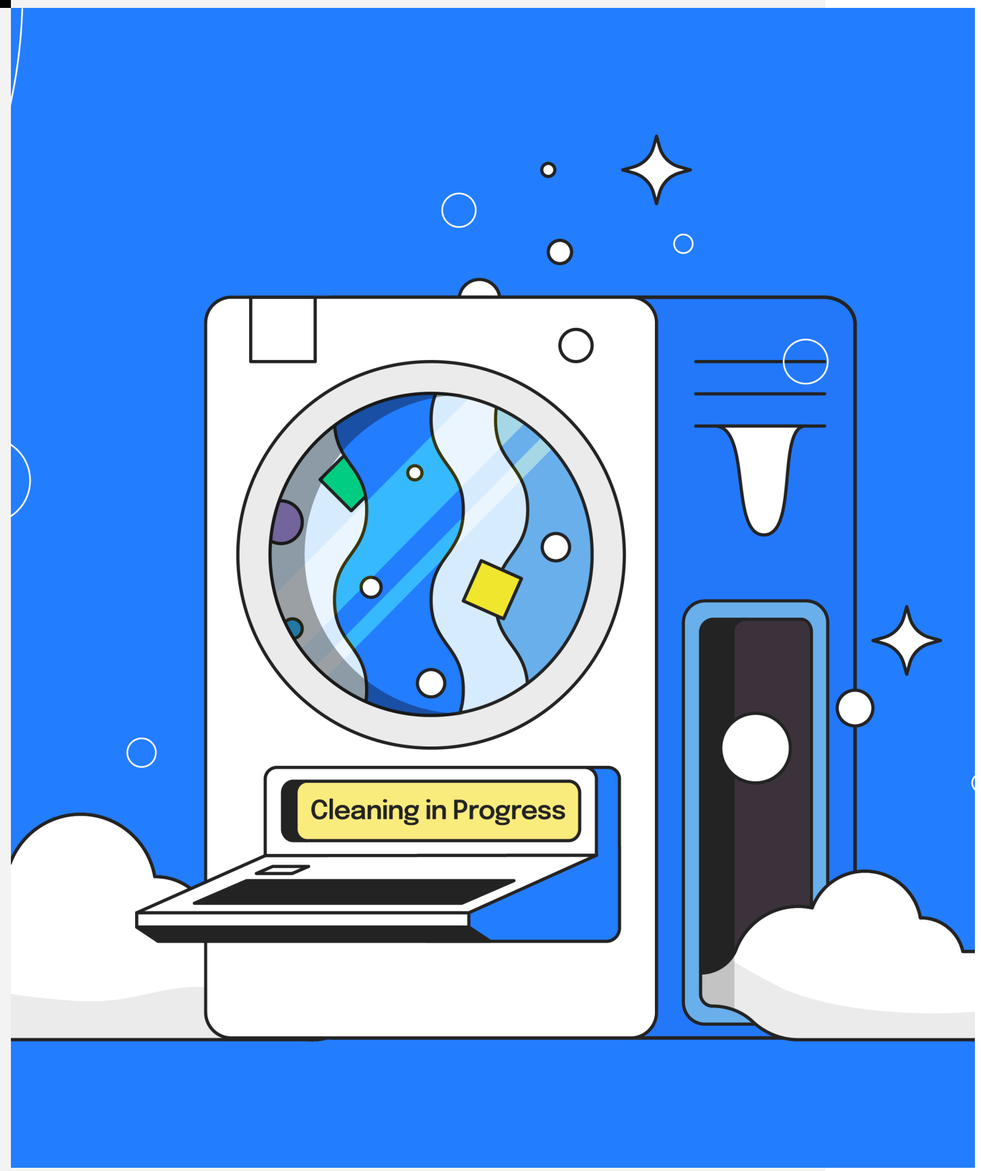
REMOVAL OF UNWANTED OBSERVATIONS

REMOVAL OF UNWANTED OBSERVATIONS

- Identify and eliminate irrelevant or redundant observations from the dataset
- Scrutinizing data entries
- improving the overall quality



FIXING STRUCTURE ERRORS



FIXING STRUCTURE ERRORS

- Inconsistencies in data formats
- Ensure uniformity in data representation
- Enhances data consistency and facilitates accurate analysis

MANAGING UNWANTED OUTLIERS

MANAGING UNWANTED OUTLIERS

- Data points significantly deviating from the norm.
- Decide whether to remove or transform outliers
- Visualize by using histogram, scatter plots, or z-score
- Removal, Transformation, Winsorization, Imputation, or Robust statistical models

2.) HANDLING MISSING VALUES

2 MAIN APPROACHES

01

DELETION

02

IMPUTATION

Before deciding which approach to employ, it helps to understand why the data is missing.

01

MISSING AT RANDOM (MAR)

The missingness is related to the observed data but not the unobserved data.

02

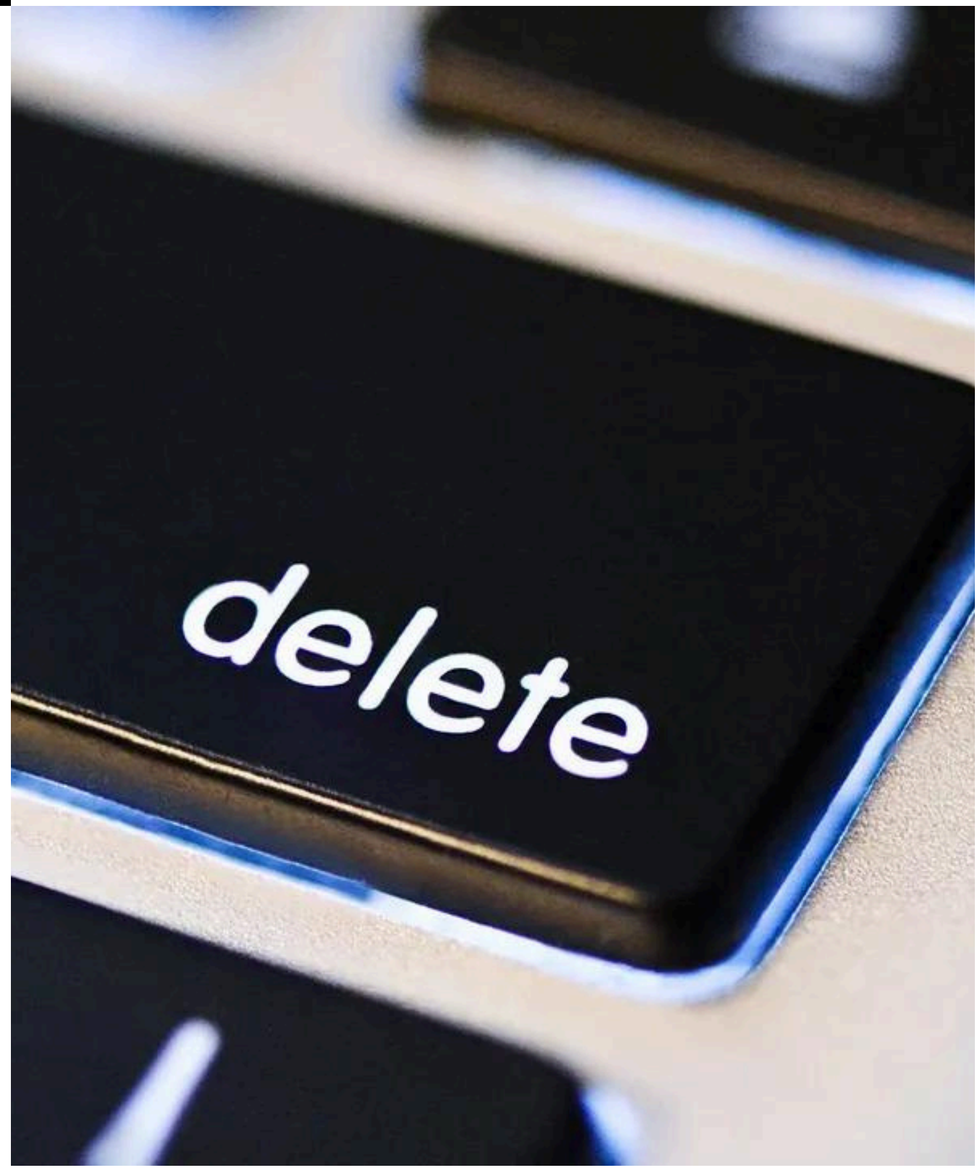
MISSING COMPLETELY AT RANDOM (MCAR)

The missingness is completely unrelated to any observed or unobserved data.

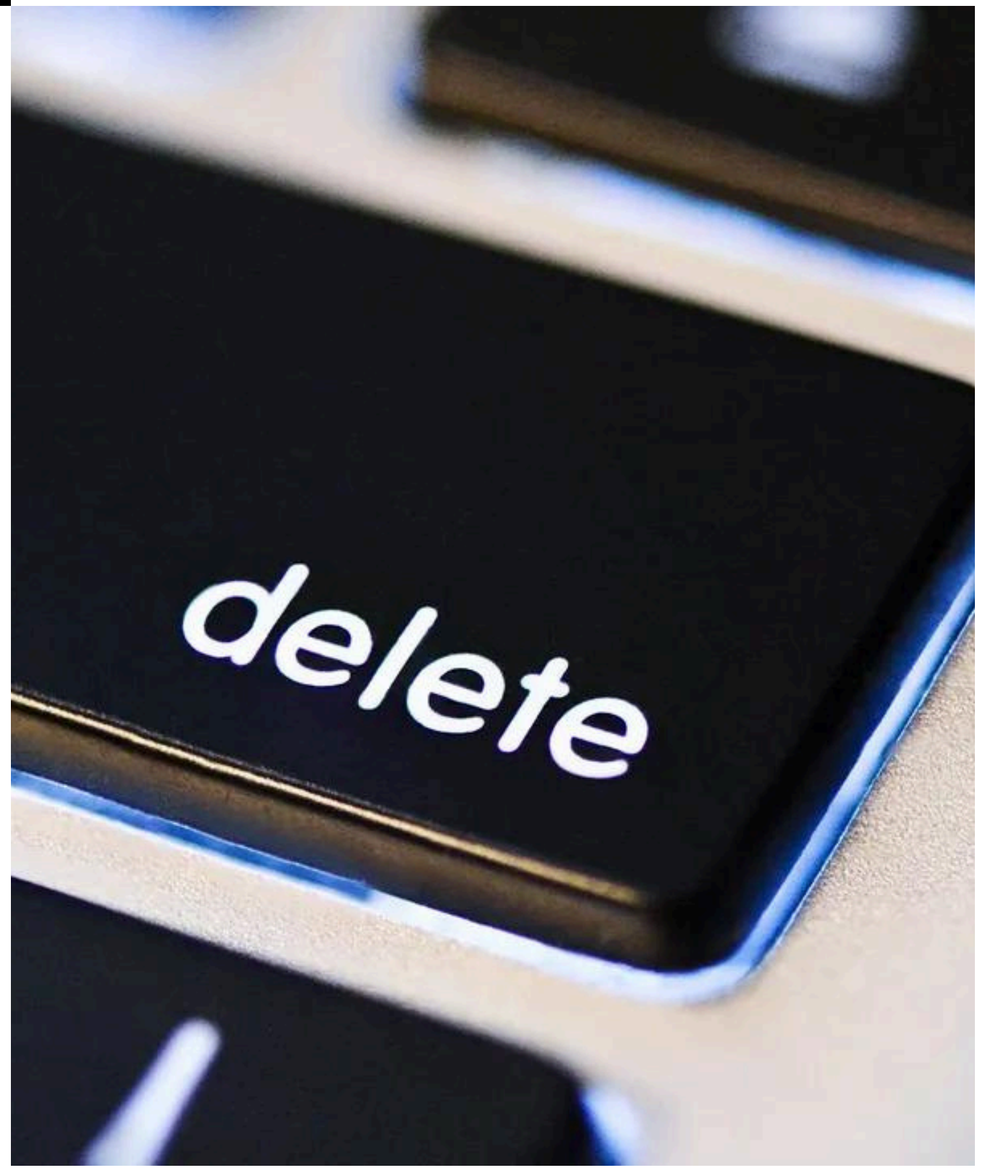
03

MISSING NOT AT RANDOM (MNAR)

The missingness is related to the unobserved data itself



DELETION



DELETION

- Removing rows or columns
- Straightforward method, but it can be problematic
- Can affect the reliability of your conclusions
- Listwise, Pairwise, Dropping Variables

IMPUTATION

IMPUTATION

- Replaces missing values with estimates
- Mean/Median/Mode Imputation
- K-Nearest Neighbors (KNN Imputation)
- Forward/Backward Fill

DATA TRANSFORMATION AND STANDARDIZATION

Data Transformation

Definition: Converting data from its original format into a format suitable for analysis.

Importance: Organizations utilize data transformation to convert data into formats that can then be used for several processes.

Data Transformation

Main Steps:

- Data Cleaning
- Data Integration
- Data Deduplication
- Data Mapping
- Data Enrichment

Data Cleaning

• Identifying and correcting errors, such as typos and missing values.

Data Integration

• Combining data from different sources into a single, coherent dataset.

Data Deduplication

• Removing duplicate entries to ensure data accuracy.

Data Mapping

- Aligning data to match a specific structure or model.

```
[15]: titanic_df.Sex.unique()

[15]: array(['male', 'female', 'F', 'M'], dtype=object)

[16]: titanic_df.Sex = titanic_df.Sex.map({'male':'M','female':'F'})

[17]: titanic_df.Sex.unique()

[17]: array(['M', 'F', nan], dtype=object)
```

Data Enrichment

- Adding additional information to enhance data quality.

```
[27]: titanic_df.head()
```

[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	person	Alone
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	male	With Family
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	female	With Family
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	female	Alone
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	female	With Family
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	male	Alone

Why Transform Data?

We transform data for:

- Improved data quality and reduced errors.
- Enhanced data organization and management.
- Increased accessibility and faster processing.

Data Standardization

Definition: Ensuring data is consistent and uniform across an organization.

Importance: Data standardization enables consistent data exchange across various systems. This means that organizations can ensure that everyone speaks the same data language by standardizing data across various systems, departments, and external partners, giving a holistic view of the company’s operations, customers, and markets.

Data Standardization

Main Steps:

- Identifying Data Sources
- Defining Data Standards
- Cleaning Data
- Performing Data Transformation
- Validating Data

Identifying Data Sources

- This allows organizations to gain insights into the data landscape and determine the scope of standardization efforts.

Defining Data Standards

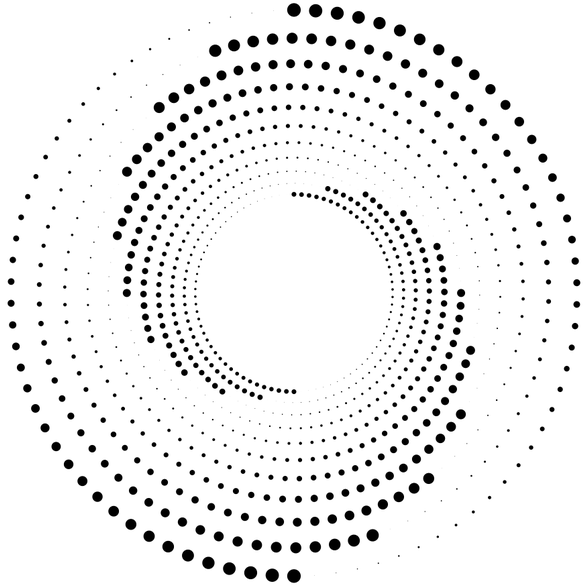
- These standards may include data formats, allowable values, validation rules, and transformation requirements.
- Defining clear standards allows organizations to ensure that data is interpreted consistently across different systems and processes.

Cleaning Data

- Data cleaning means identifying and rectifying data errors, inconsistencies, and inaccuracies. This process includes removing duplicate entries, correcting misspellings, and resolving missing or incomplete data.

Performing Data Transformation

- The next step is converting the data into a consistent format and structure to ensure that all data can be easily compared and analyzed. This includes tasks such as changing dates into a standardized format or converting units of measurement to a common standard.



Data Standardization

Validating Data

- Validate data by running tests and checks on the data, such as verifying data integrity, checking for outliers or anomalies, and validating against predefined rules or constraints.

OUTLIER DETECTION AND TREATMENT

What is an outlier?

An Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.

Why treat outliers?

Outliers are data points that significantly differ from the rest of the dataset. Detecting outliers is crucial because they can skew analyses and lead to incorrect conclusions.

Why do they occur?

Common causes of outliers are:

- Measurement errors
- Sampling errors
- Natural variability
- Data entry errors
- Experimental errors
- Sampling from multiple populations

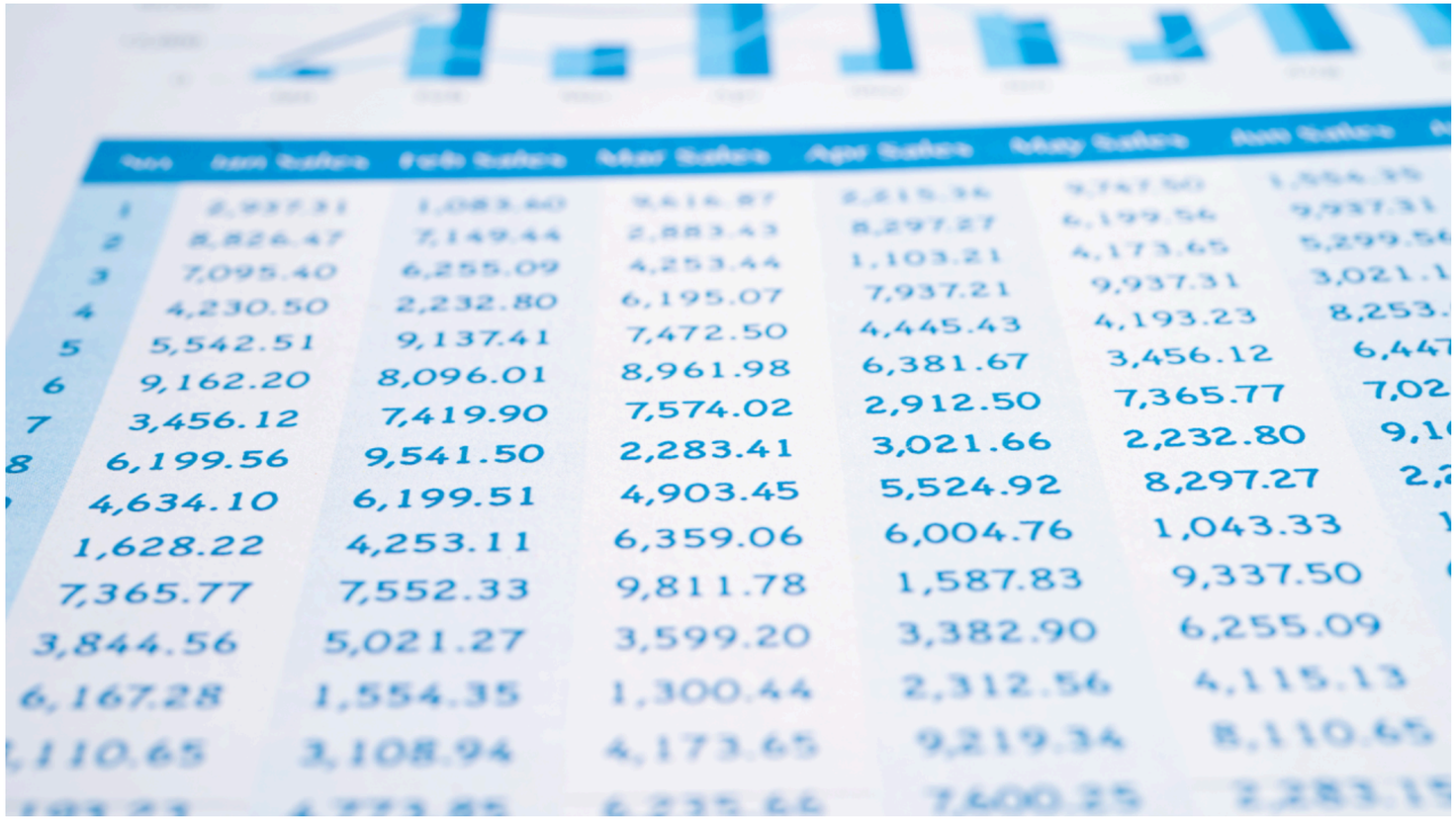
Outlier Detection

If our dataset is small, we can visually detect the outlier by just looking at the dataset.

Example:

[15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]

But what if we have a huge dataset, how do we identify the outliers then?



We need to use visualization and mathematical techniques.

These include:

- Z-Score
- Interquartile Range
- Box plots

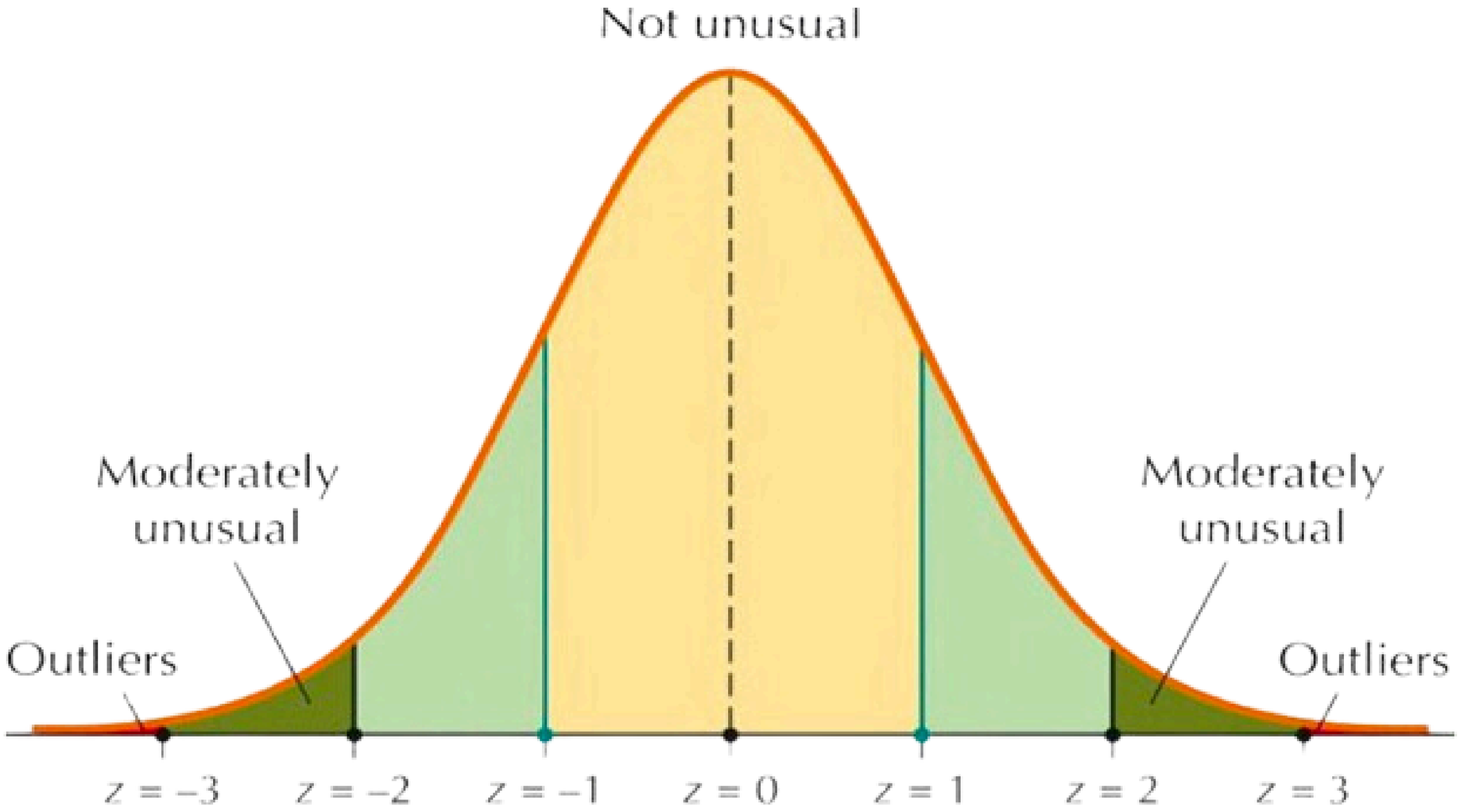
Z-Score

To compute for z-score, use the formula:

$z = (x-\mu)/\sigma$

where:

- x = a random variable
- μ = mean
- σ = standard deviation

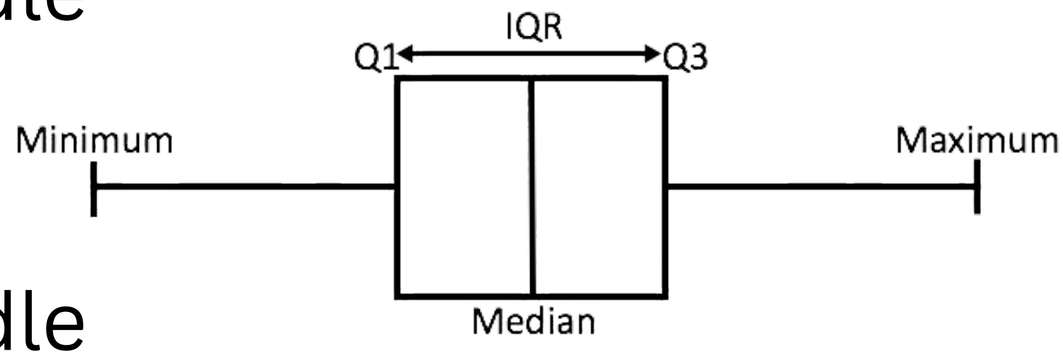


If a z-score is 3 or more, it is generally considered an **outlier** or an anomaly.

Interquartile Range

Find the:

- Q1 - 25th percentile or the middle value between the smallest number and the median
- Q3 - 75th percentile or the middle value between the median and the largest number
- IQR = the middle 50% value or Q3-Q1



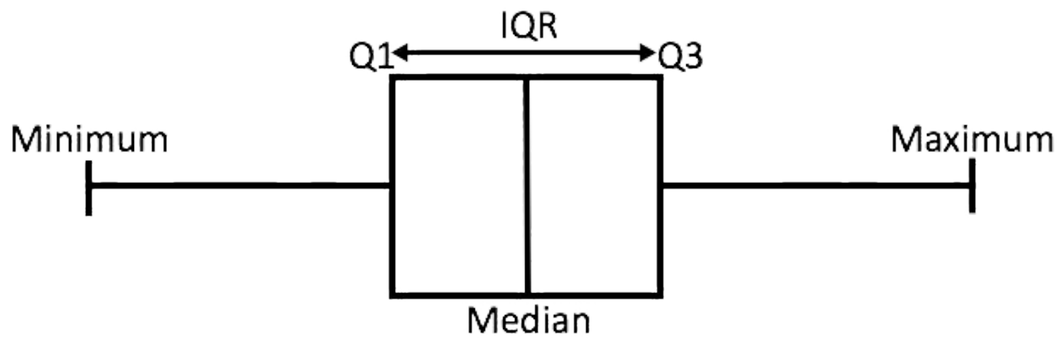
Interquartile Range

Lower bound will be:

minimum = Q1 - (IQR * 1.5)

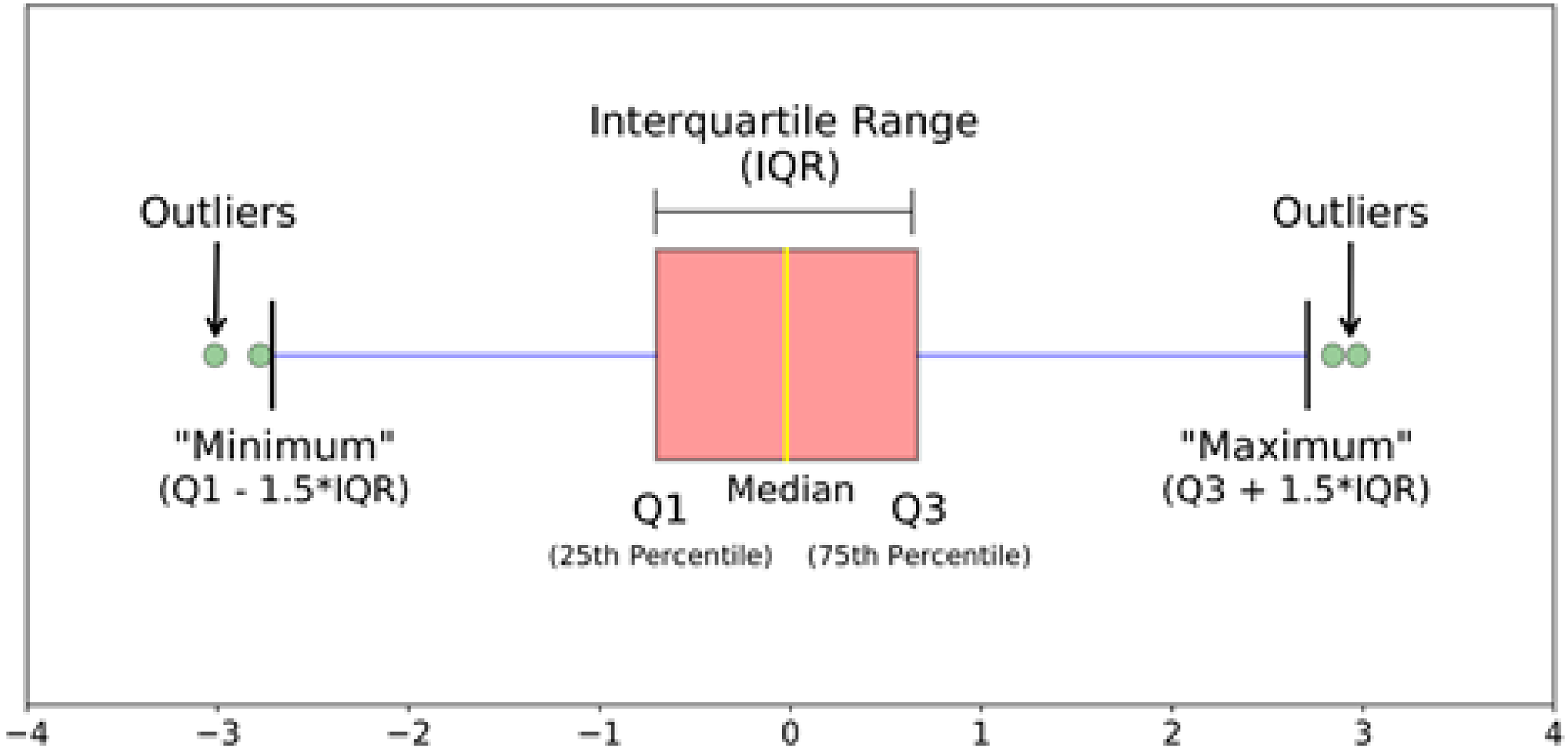
Upper bound will be:

maximum = Q3 + (IQR * 1.5)



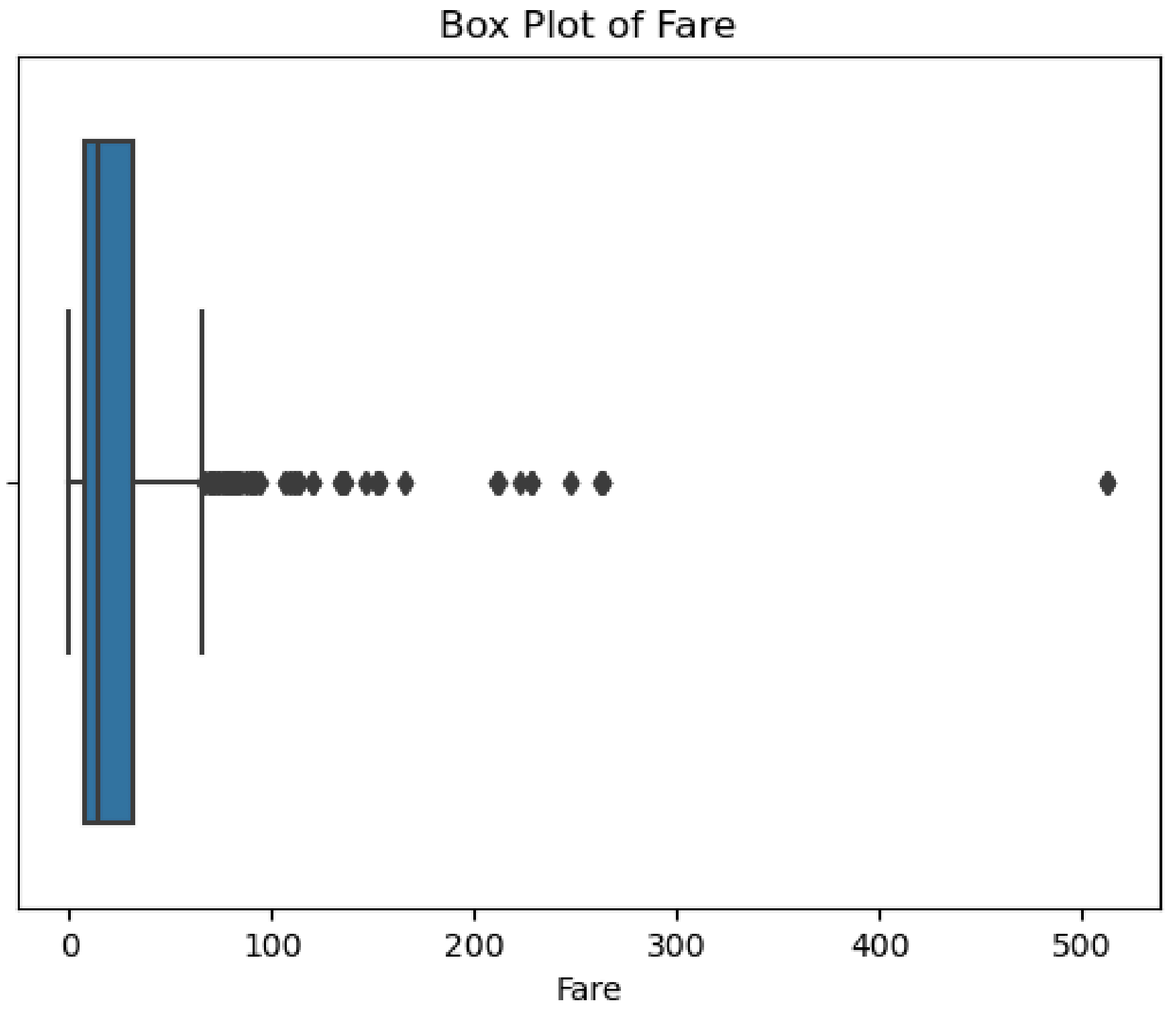
Criteria using IQR: data points that lie 1.5 times of IQR above Q3 and below Q1 are **outliers**.

Box Plots


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

Box Plots

```
sns.boxplot(x=titanic_df['Fare'])
plt.title('Box Plot of Fare')
plt.show()
```



Outlier Treatment

- It might be tempting to just remove the records where there are outliers in the data set but it’s not always the best approach.
- The outlier treatment method can vary from case to case and should be discussed with the business before finalizing the method.

There are different approaches such as:

1. replacing the outlier with the mean value, or median value
2. deleting the outlier if they are due to data entry errors

Removing them from the dataset

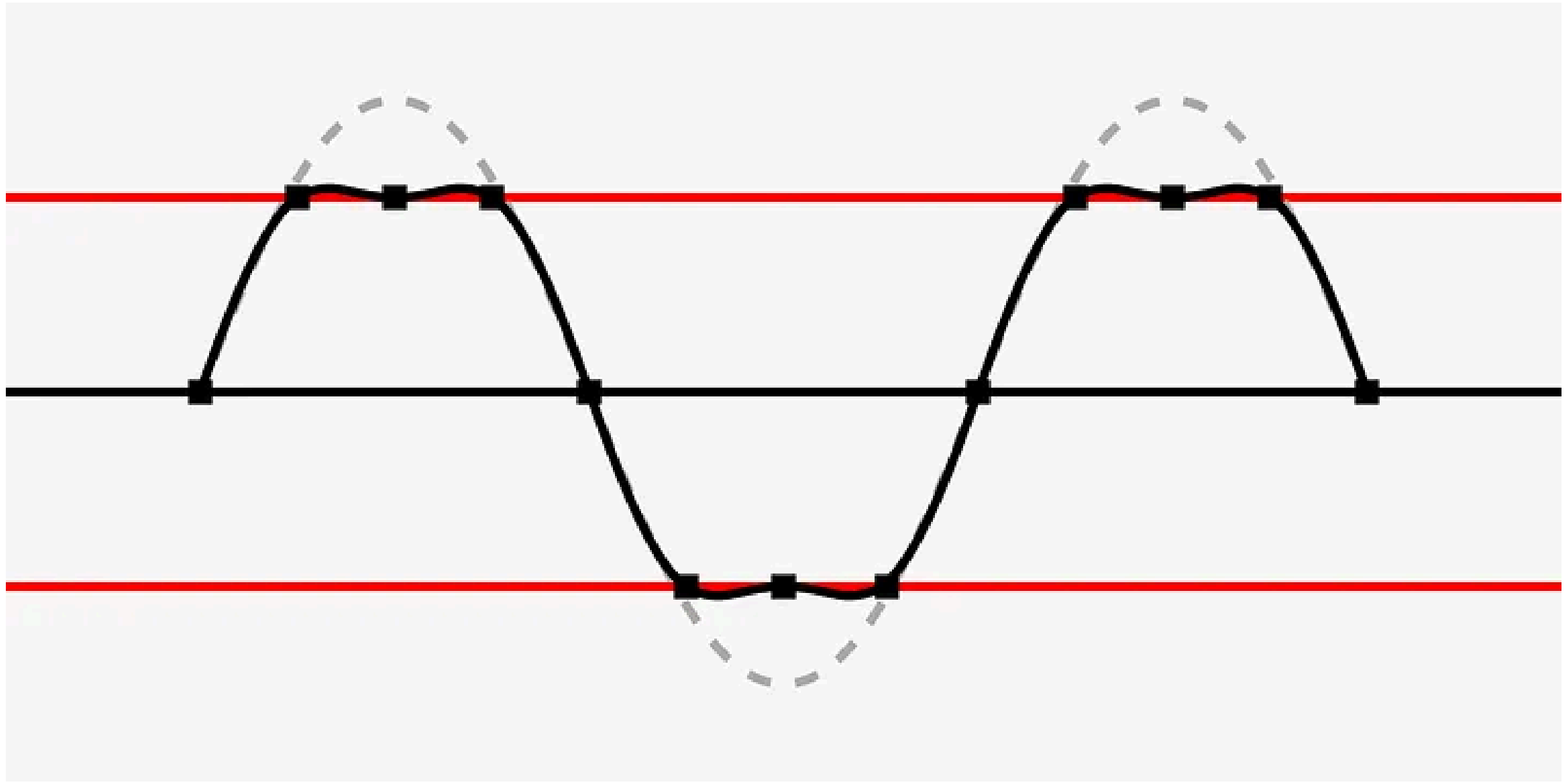
In this technique, we remove the outliers from the dataset.

Mean/Median Imputation

Replacing outliers with more appropriate values, such as the mean or median so that it would not lead to incorrect conclusions

Winsorization

Winsorization is essentially similar to imputing, but instead of imputing extreme values with mean, median, mode, min or max values, we imput those outliers with our chosen percentile.



How do we know if we have to remove the outlier?

If the outlier in question is:

- A measurement error or data entry error, correct the error if possible. If you can’t fix it, **remove** that observation because you know it’s incorrect.
- Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately **remove** the outlier.
- A natural part of the population you are studying, you should **not remove** it.

Remember

- When you decide to remove outliers, document the excluded data points and explain your reasoning. You must be able to attribute a specific cause for removing outliers. Another approach is to perform the analysis with and without these observations and discuss the differences.

DATA INTEGRATION AND MERGING

DATA INTEGRATION AND MERGING

Data Integration is **the process of combining data from different sources into a unified and coherent format** that can be used for analytical, operational, and decision-making purposes.

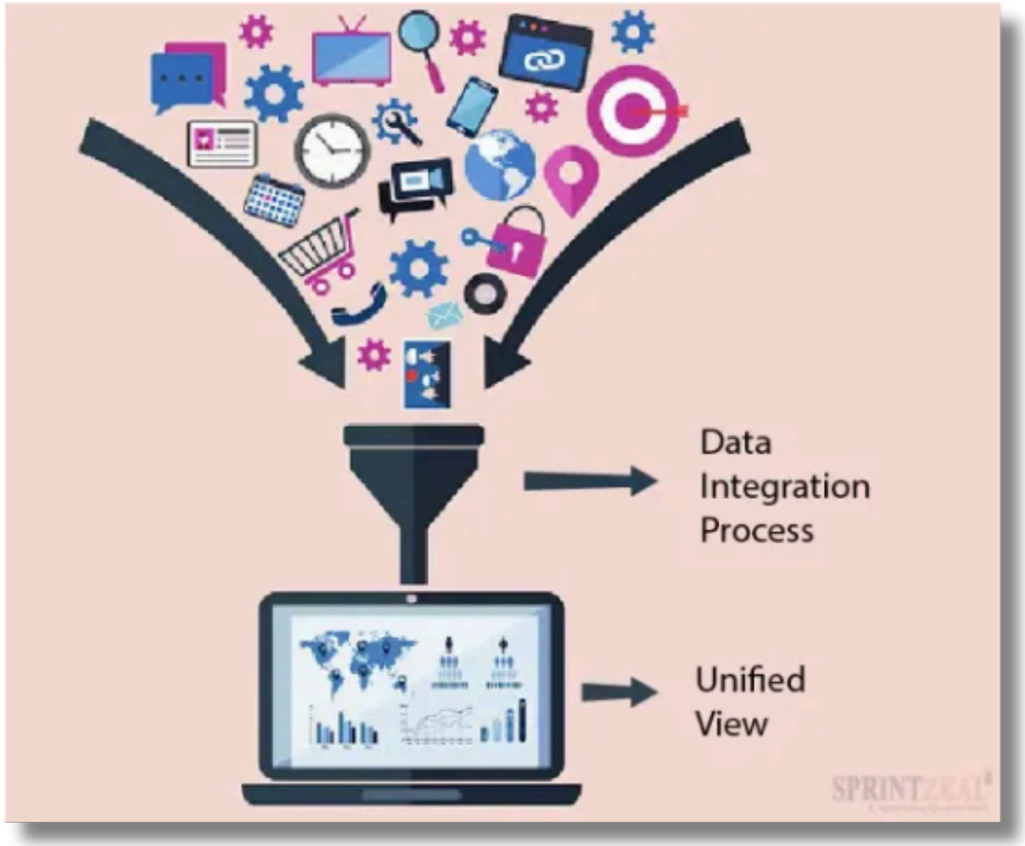


Image from: https://www.sprintzeal.com/blog/data-integration

DATA INTEGRATION AND MERGING

The process typically includes several steps:

- **Data Extraction**
  - **Gathering data from various sources** that may include databases, data warehouses, spreadsheets, APIs, and others.
- **Data Transformation**
  - The **extracted data is converted into a common format** to ensure consistency, accuracy, and compatibility. This process may involve cleaning data, handling missing values, and standardizing data formats
- **Data Loading**
  - **Storing the transformed data to a single central repository** such as a data warehouse, for further analysis and reporting.

DATA INTEGRATION AND MERGING

- **Data Synchronization**
  - Ensures that the integrated data is up to date over time and consistent across the system

- **Data Access and Analysis**
  - Once integrated, the data sets can be accessed and analyzed using various tools such as BI software, reporting tools, and analytics platforms

TYPES OF DATA INTEGRATION

- Extract, Load, Transform (ELT)
- Extract, Transform, Load (ETL)
- Real-time data integration
- Application Integration (API)
- Data Virtualization

TYPES OF DATA INTEGRATION

- **Extract, Load, Transform (ELT)**
  - Involves **extracting data** from its source, **loading it into a database** or data warehouse and then later **transforming it into a format** that suits business needs.

- **Extract, Transform, Load (ETL)**
  - Data is **transformed before loading** it into the data storage system. Meaning that the transformation happens outside the data storage system

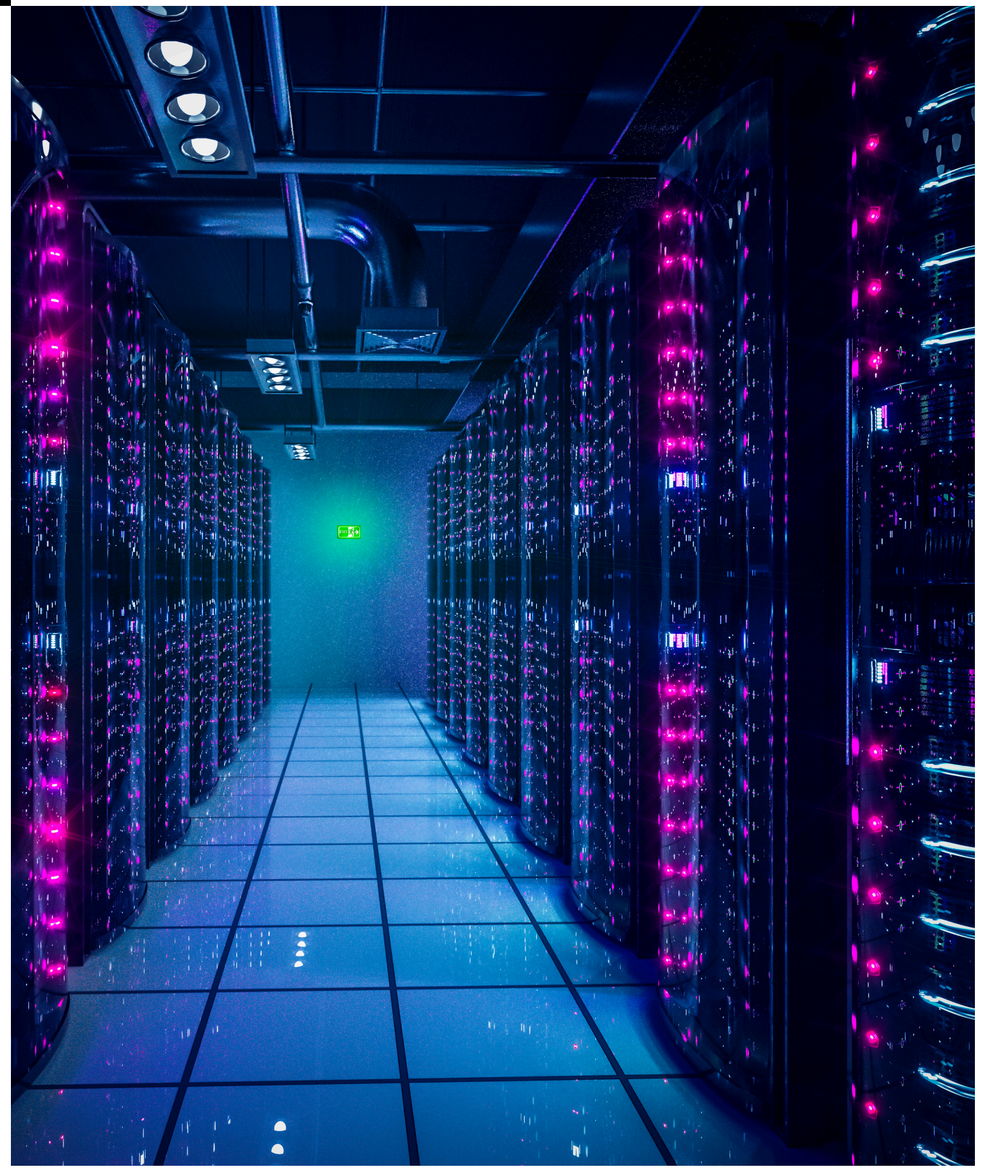
https://www.datachannel.co/blogs/what-is-etl-and-how-the-etl-process-works

TYPES OF DATA INTEGRATION

- **Real-time data integration**
  - involves capturing and processing data as it becomes available in source systems, and then **immediately integrating it into the target system.**

- **Application Integration (API)**
  - Involves **integrating data between different software applications** to ensure seamless data flow and interoperability.

- **Data Virtualization**
  - Offers a unified virtual view of data aggregated from multiple sources, sidestepping the need for physical data consolidation.



DATA INTEGRATION USE CASES

- Data warehousing
- Data lake development
- Business Intelligence and Reporting
- Processing IoT data

DATA MERGING

Data merging is a specific aspect of data integration focused on combining datasets into a single, cohesive dataset

DATA MERGING

- **Pre-Merging**
  - Profiling
  - Transformation
- **Merging**
  - Append Rows
  - Append Columns
  - Conditional Merge
- **Post-Merging**

DATA MERGING EXAMPLE

- An e-commerce company wants to integrate and merge data from different sources to gain a comprehensive view of its operations.
  - Customer Database
  - Order Database
  - Web Analytics Data

Customer Database (customer_db)

customer_id	name	email	registration_date
1	Alice Smith	alice@example.com	15/01/2020
2	Bob Johnson	bob@example.com	22/11/2019
3	Carol White	carol@example.com	03/05/2021

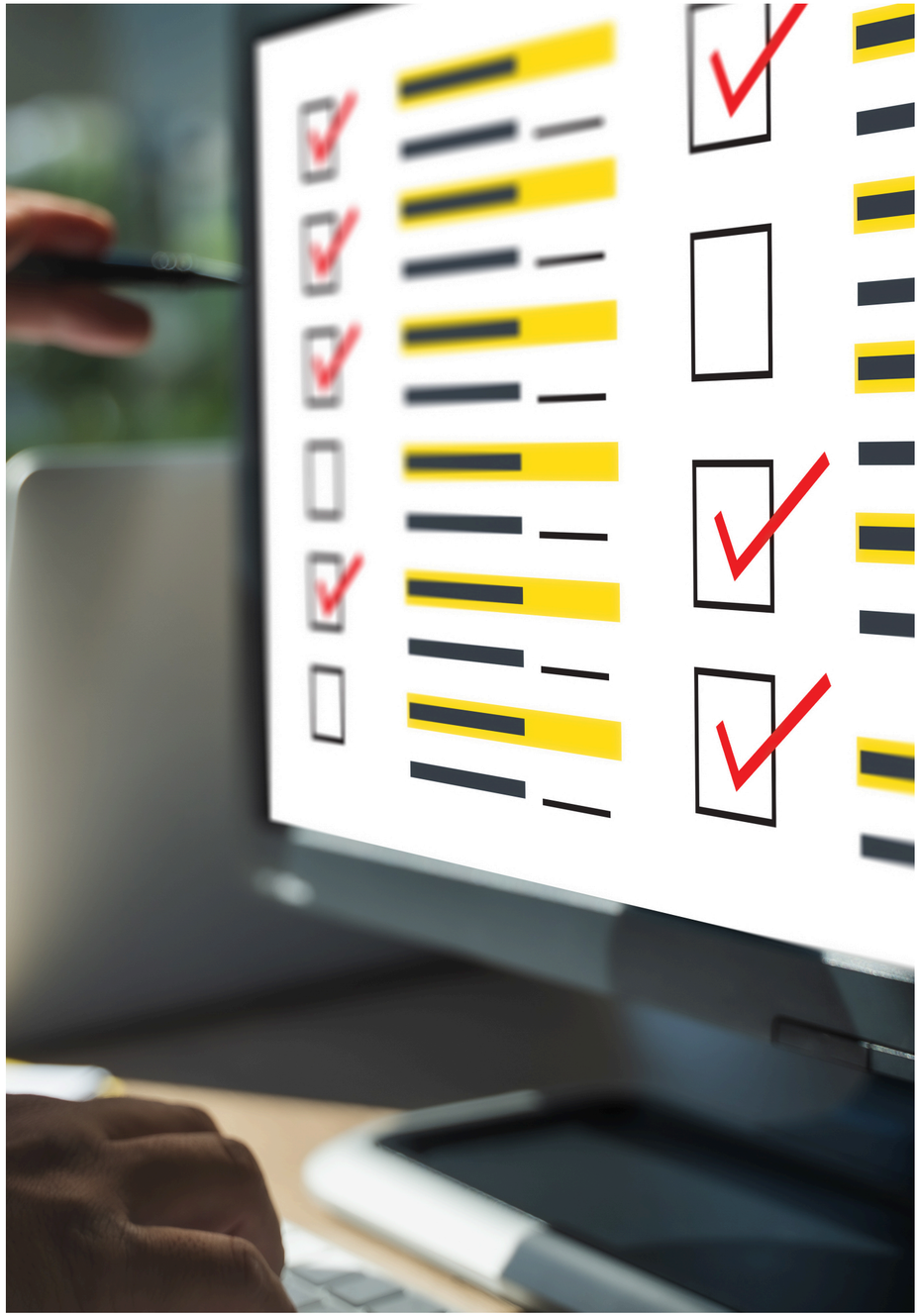
Order Database (order_db)

order_id	customer_id	order_date	amount
101	1	01/03/2022	150
102	2	02/03/2022	200
103	3	03/03/2022	300

Web Analytics Data (web_analytics)

session_id	customer_id	visit_date	page_views
1001	1	01/03/2022	5
1002	2	02/03/2022	3
1003	3	03/03/2022	7

customer_id	name	email	registration_date	order_id	order_date	amount	session_id	visit_date	page_views
1	Alice Smith	alice@example.com	15/01/2020	101	01/03/2022	150	1001	01/03/2022	5
2	Bob Johnson	bob@example.com	22/11/2019	102	02/03/2022	200	1002	02/03/2022	3
3	Carol White	carol@example.com	03/05/2021	103	03/03/2022	300	1003	03/03/2022	7



DATA VALIDATION AND VERIFICATION

DATA VALIDATION

Types of Data Validation

- Data Type Validation
- Format Validation
- Range Validation
- Consitency Validation
- Uniqueness Validation
- Code Check

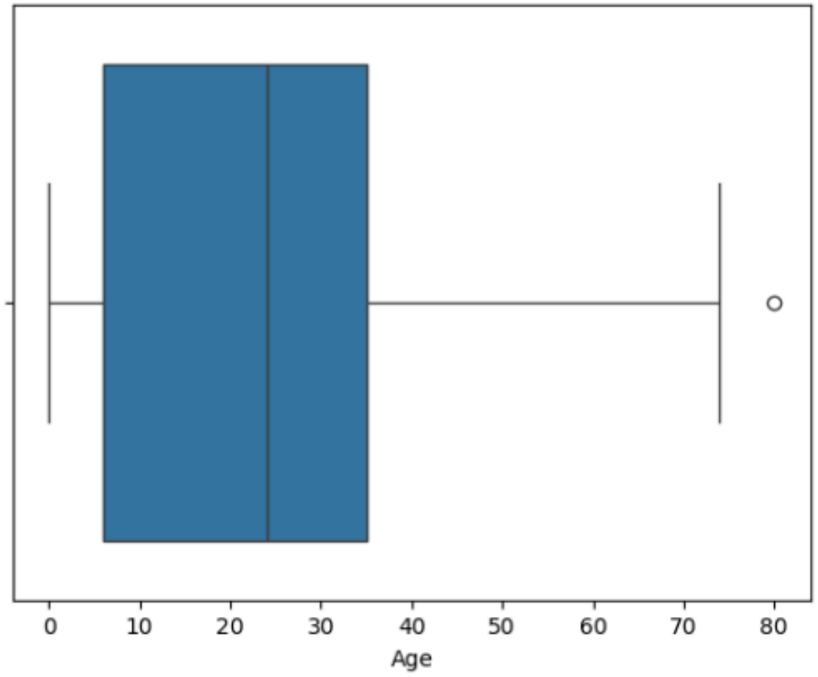
```
df.dtypes
```

```
Survived          int64
Ticket Class      int64
Name             object
Sex              object
Age             float64
SibSp             int64
Parch             int64
Ticket           object
Fare            float64
Cabin            object
Embarked         object
```

```
print("Min Age", df["Age"].min())
print("Max Age", df["Age"].max())

sns.boxplot(data=df, x="Age")
plt.show()
```

```
Min Age 0
Max Age 80
```

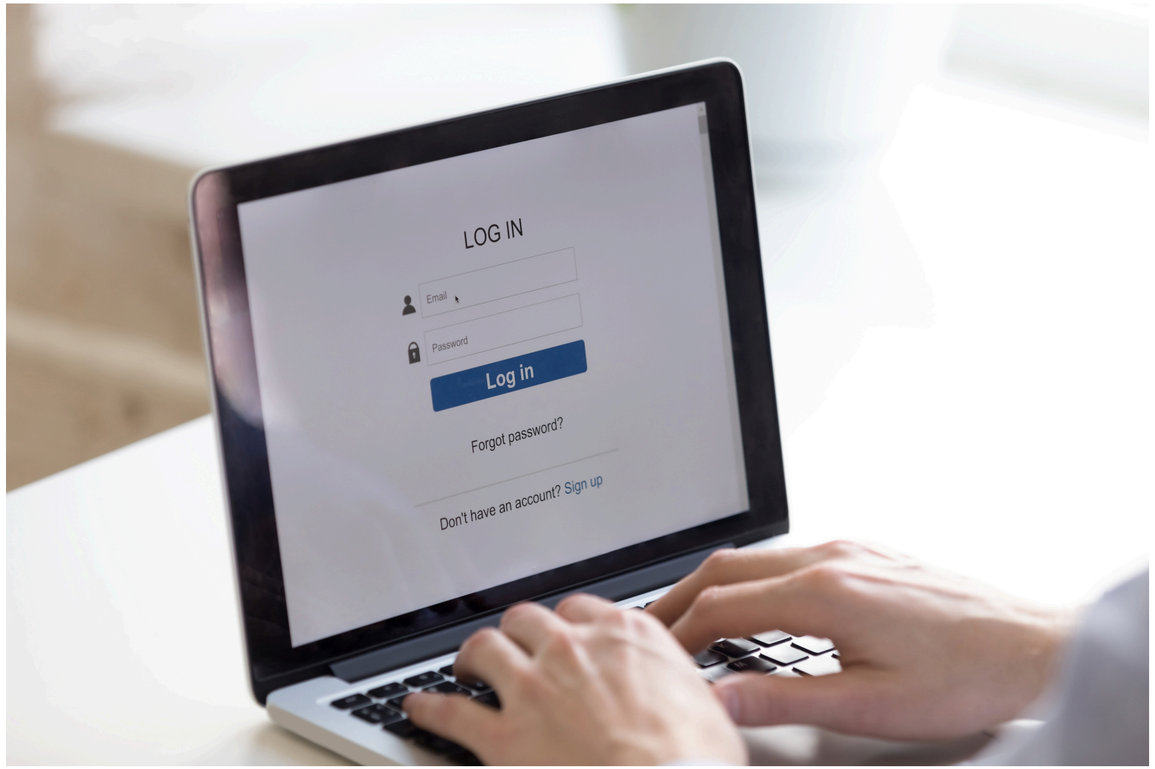



```
df.select_dtypes("number").head()
```

	Survived	Ticket Class	Age	SibSp	Parch	Fare
0	0	3	22	1	0	7.2500
1	1	1	38	1	0	71.2833
2	1	3	26	0	0	7.9250
3	1	1	35	1	0	53.1000
4	0	3	35	0	0	8.0500

DATA VERIFICATION

Data verification is the process of ensuring that data **accurately represents the real-world values** it is supposed to model.



DATA VERIFICATION

Data Verification Techniques

- **Double Entry**
  - This refers to inputting the data twice and comparing the two entries.
  - A classic example would be when **creating a new password**.
- **Proofreading Data**
  - This process calls for a thorough inspection of the data entry to make sure there are no errors and that everything is accurate.

DATA CLEANING TOOLS AND AUTOMATION

WHAT ARE DATA CLEANING TOOLS?

According to Techtarget:

“Numerous tools can be used to **automate** data cleansing tasks, including both commercial software and open source technologies.

The tools include a variety of functions for **correcting** data errors and issues, such as **adding** missing values, **replacing** null ones, **fixing** punctuation, **standardizing** fields and **combining** duplicate records.”

Reference: https://www.techtarget.com/searchdatamanagement/definition/data-scrubbing

WHY USE DATA CLEANING TOOLS?

According to Talend:

“A data cleansing tool helps provide **reliable**, **complete insights** so that you can **identify evolving customer needs** and stay on top of emerging trends. Data cleansing can produce faster response rates, generate quality leads, and improve the customer experience.”

Reference: https://www.talend.com/resources/what-is-data-cleansing

PYTHON PANDAS

Widely embraced open-source library in Python for manipulating and analyzing data.

Includes libraries like:

- **NumPy** for numerical computations
- **Matplotlib** for data visualization

Why use pandas?

- Versatile Data Structures
- Handling Missing Data
- Group-by Functionalities
- Merging and Reshaping
- IO Tools
- Reliability and Speed

Image from: https://en.wikipedia.org/wiki/Pandas_%28software%29

OPENREFINE

A powerful and versatile open-source tool tailored for managing messy data effectively.

OpenRefine Features:

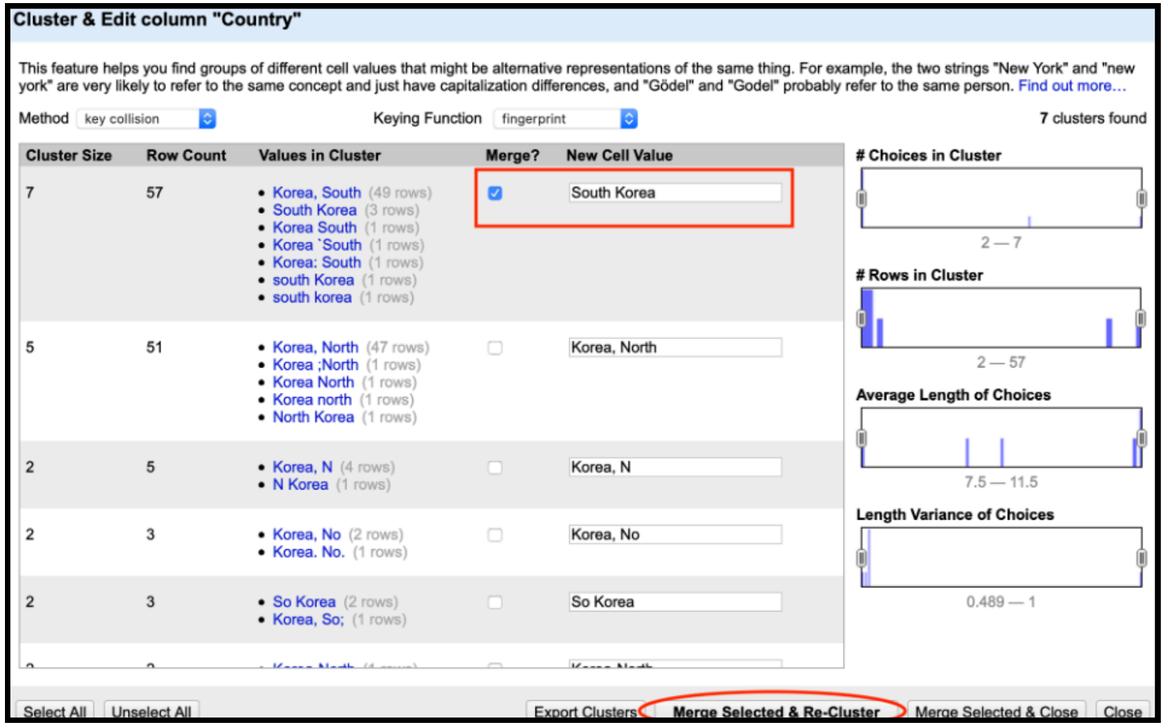
- Data Cleaning and Transformation
- Faceting Feature
- Clustering Functionality
- Data Enrichment
- Local Data Processing
- Integration with Wikidata

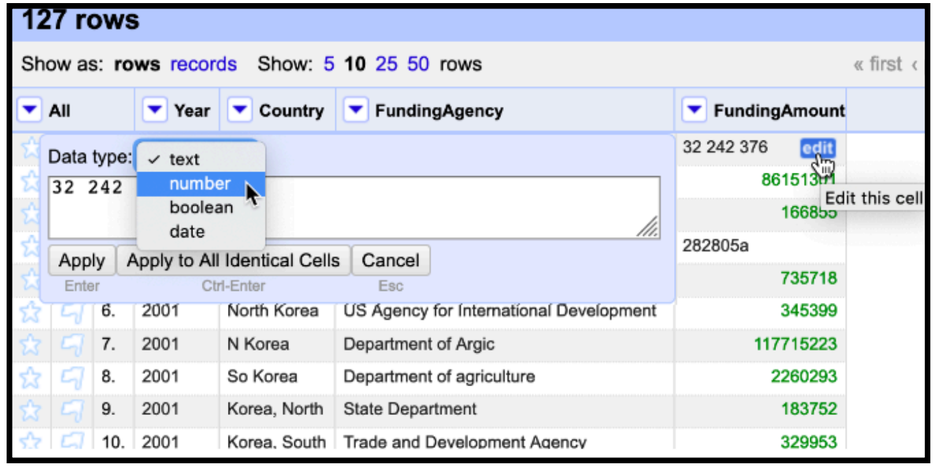
Refine use cases include:

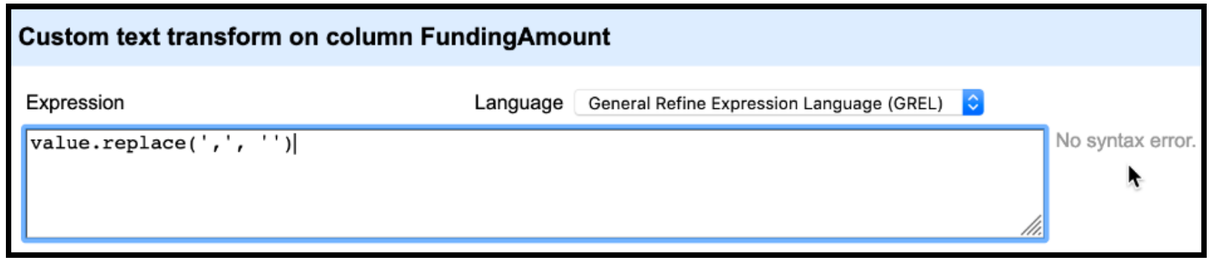
- Clean
- Transform
- Extend
- Automate

Image from: https://en.wikipedia.org/wiki/Pandas_%28software%29

OPENREFINE







Curious? View Here:

Image from: https://en.wikipedia.org/wiki/Pandas_%28software%29

ALTERYX

An analytics automation software in a low-code environment. In place of lines of codes for automation, has an entire suite of built-in tools readily available to be used for the creation of automated processes.

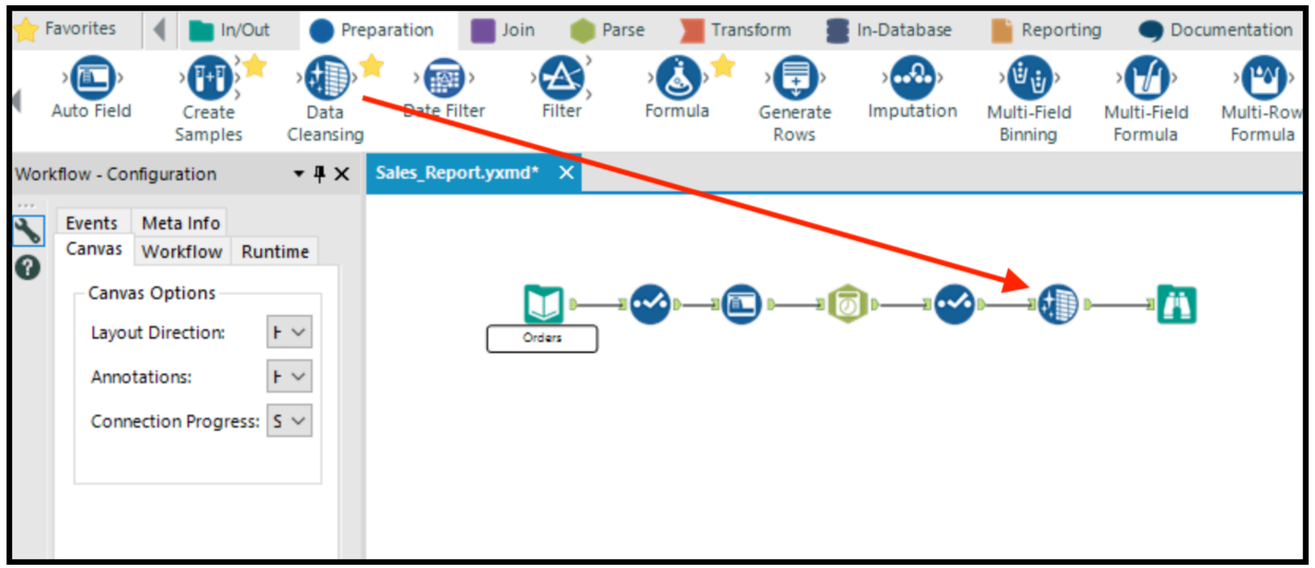
Asterix Features:

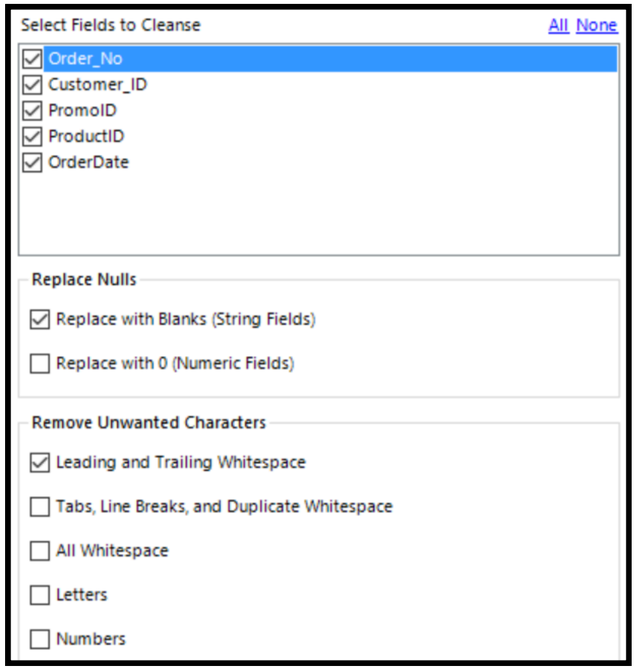
- Unique Tool
- Filter Tool
- Formula Tool
- Scatterplot Tool
- Data Cleansing Tool
- Test Tool

Image from: https://devtools.in/partners/alteryx/

ALTERYX

Order_No	Order_Date	Customer_ID	PromoID	ProductID
16601259	22/11/17	4002	[Null]	376401
19147165	01/04/17	3979	[Null]	218178
20823441	21/11/17	3382	[Null]	218178
19388155	19/11/17	3552	[Null]	355984
16842078	10/05/17	1681	[Null]	227774
16055181	26/01/17	4101	30_OFF	228722
14931234	03/01/17	3579	30_OFF	416113





Curious? View Here:

Image from: https://www.analyseup.com/learn-alteryx/alteryx-data-cleansing.html

KNIME

An open-source software that offers a user-friendly visual interface for building complex analyses, including spreadsheet automation, ETL processes, and machine learning.

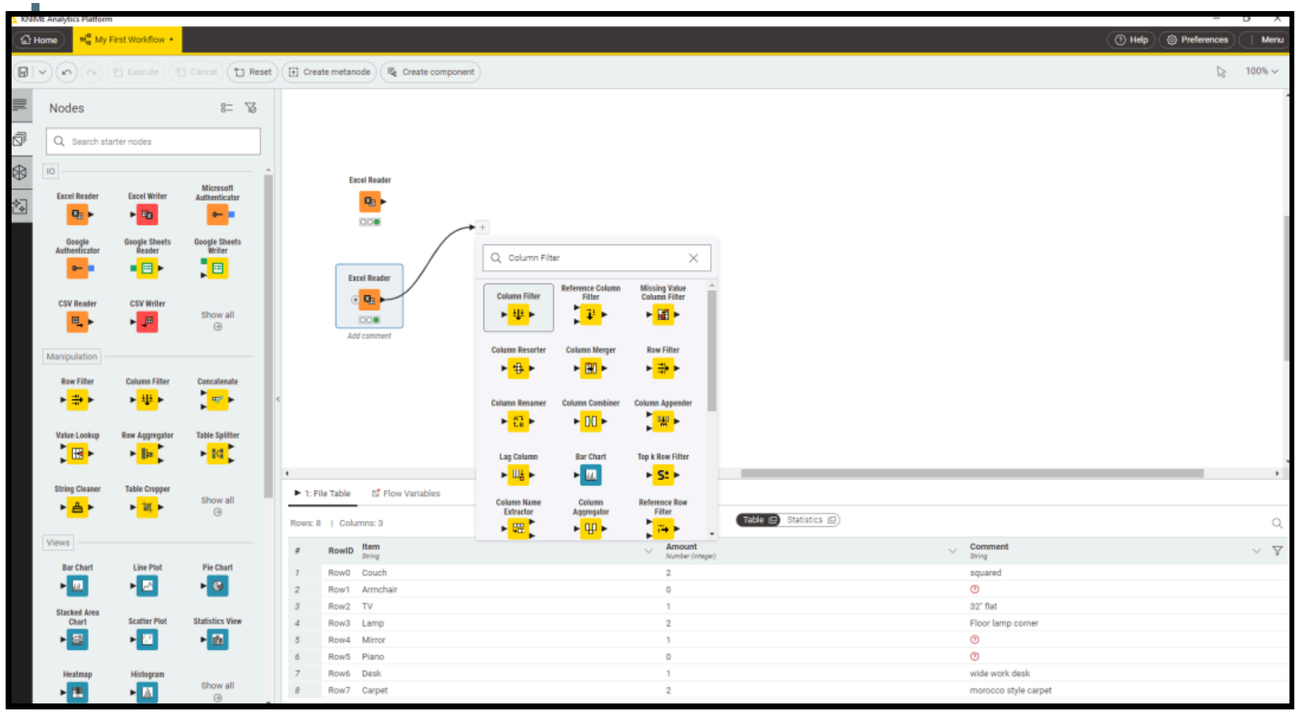
KNIME Features:

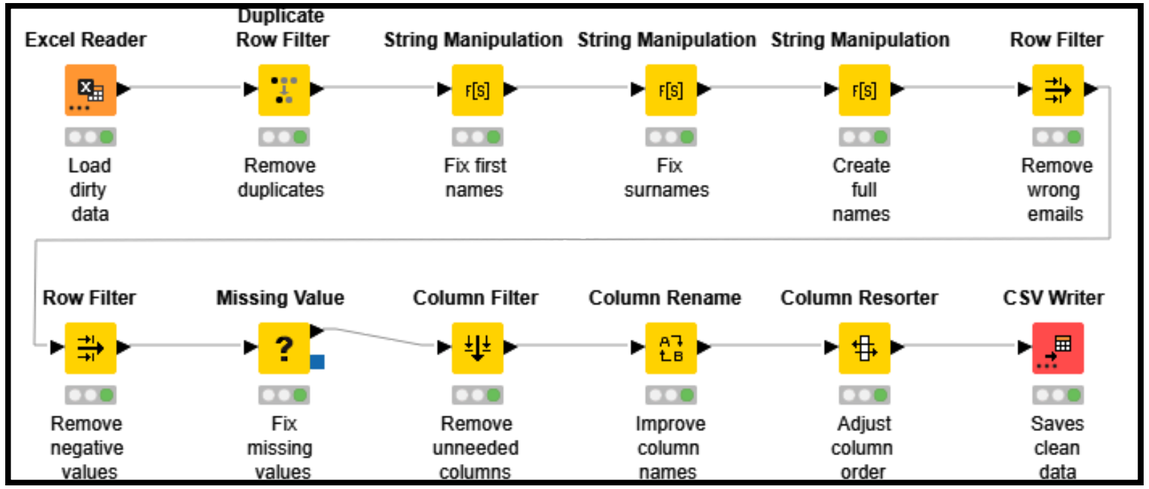
- Node-based Workflow Interface
- Data Manipulation and Transformation Nodes
- Data Enrichment and Transformation
- Duplicate Removal
- Data Validation
- Data Standardization
- Data Profiling

Open for Innovation
KNIME

Image from: https://www.knime.com/getting-started-guide

KNIME





Curious? View Here:

Image from:
https://hub.knime.com/adm/spaces/Public/Workflows/Data%20Analytics%20Made%20Easy/Chapter%202/Cleaning%20data~DLJtVEX0f9gQTCS7/current-state

https://www.knime.com/getting-started-guide

DATAIKU

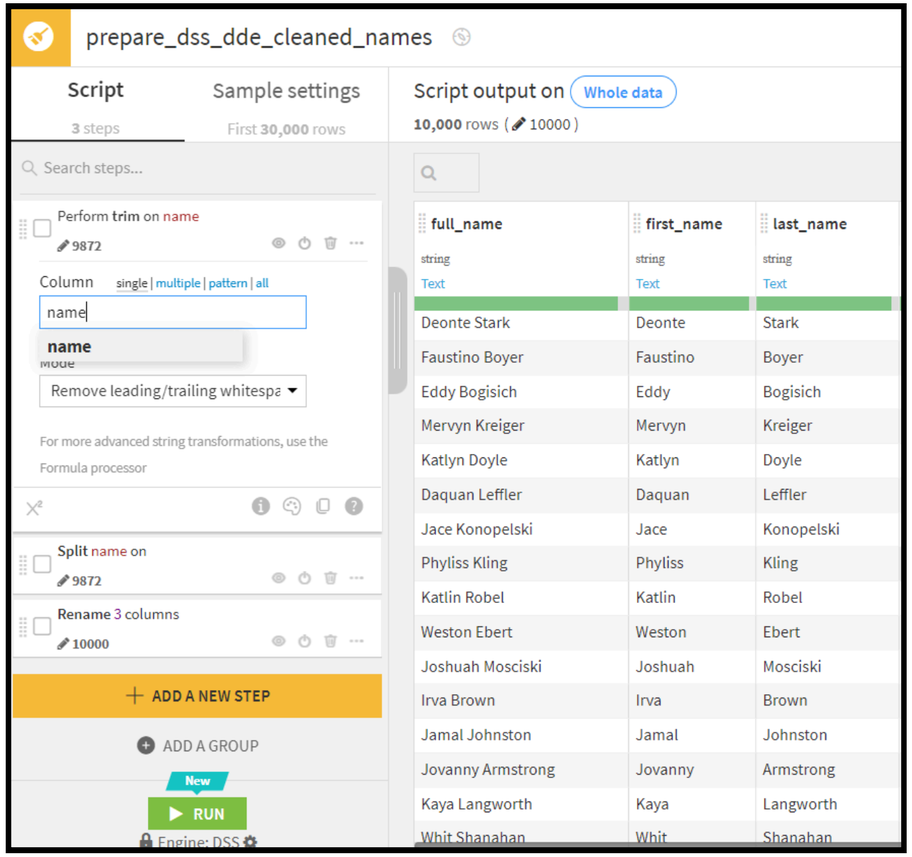
A comprehensive software platform for designing, deploying, and managing data analytics applications and predictive machine learning (ML) models.

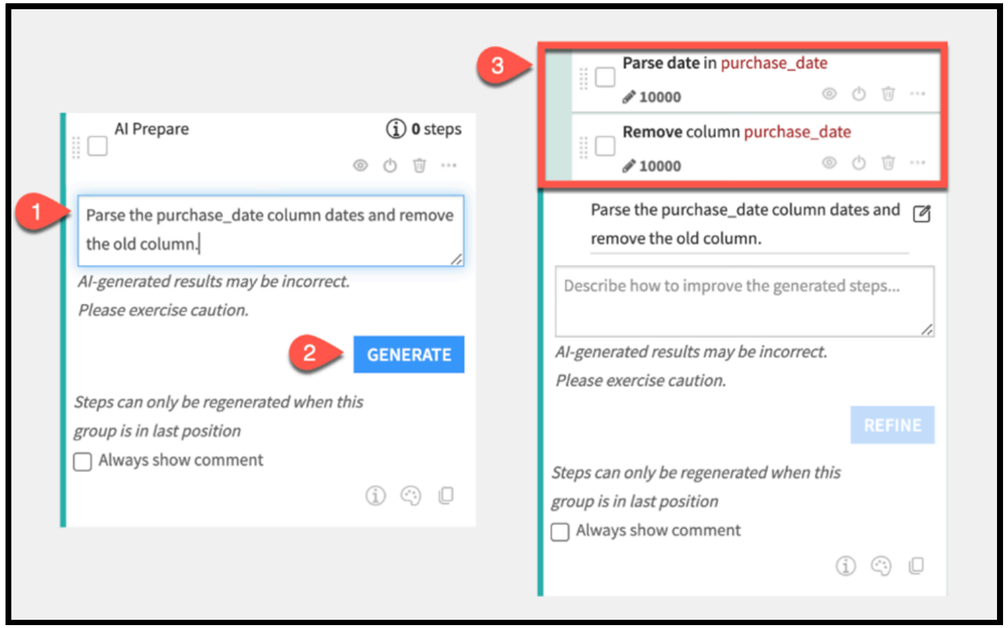
Dataiku Features:

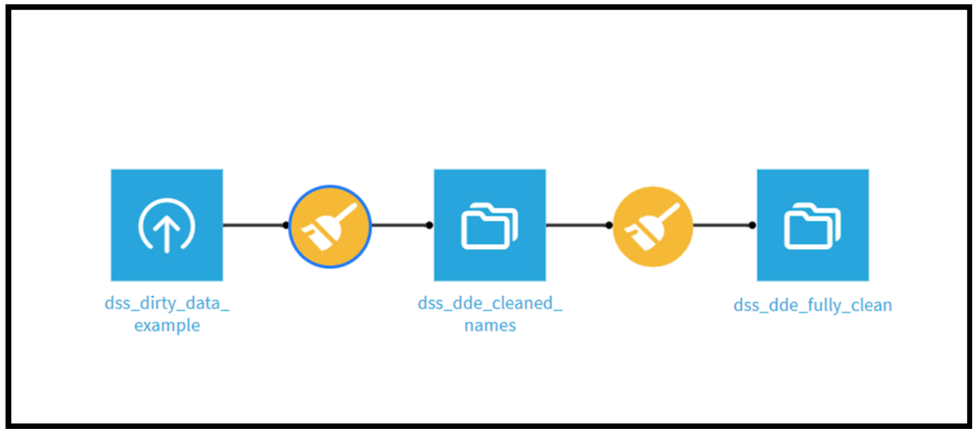
- Collaborative Environment
- Data Preparation Tools
- Model Building
- Model Validation and Testing
- Model Deployment and Monitoring

Image from: https://www.dataiku.com/learn/samples/cleaning-contacts/

DATAIKU







Curious? View Here:

Image from: https://gallery.dataiku.com/projects/DKU_CLEANING_CONTACTS/flow/

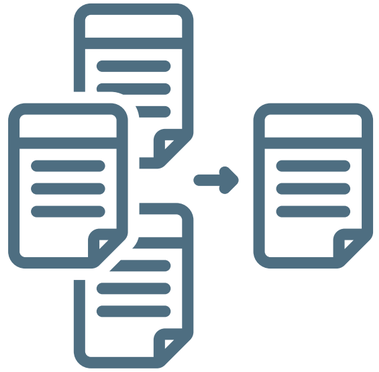
https://knowledge.dataiku.com/latest/data-preparation/prepare-recipe/concept-ai-prepare.html

EXPLORATORY DATA ANALYSIS

WHAT IS IT?

Analyze Dataset

Visualize Data



Produce Insight

WHY IS IT IMPORTANT?

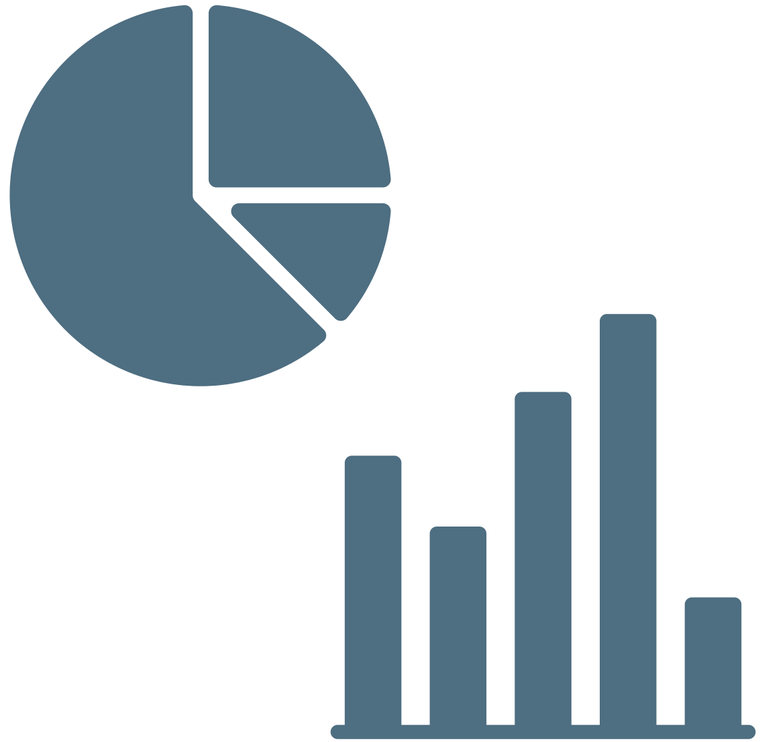
It helps in generating insights or further piques the interest and curiosity of the person

WHY IS IT IMPORTANT?

it validates the data by checking for inconsistencies, outliers, and errors

WHY IS IT IMPORTANT?

it helps in determining the correct method of presenting data



4 TYPES OF EDA

Univariate Non-graphical

- Only has single variable
- No graphical presentations
- Used for statistical measurements

Mean:	1.69
Median:	1.55
Mode:	1.5
Standard Deviation:	0.356351
Variance:	0.126986

4 TYPES OF EDA

Univariate Graphical

- Still deals with only one variable
- Uses graphical presentation to visualize data
- Helps to observe how data behaves

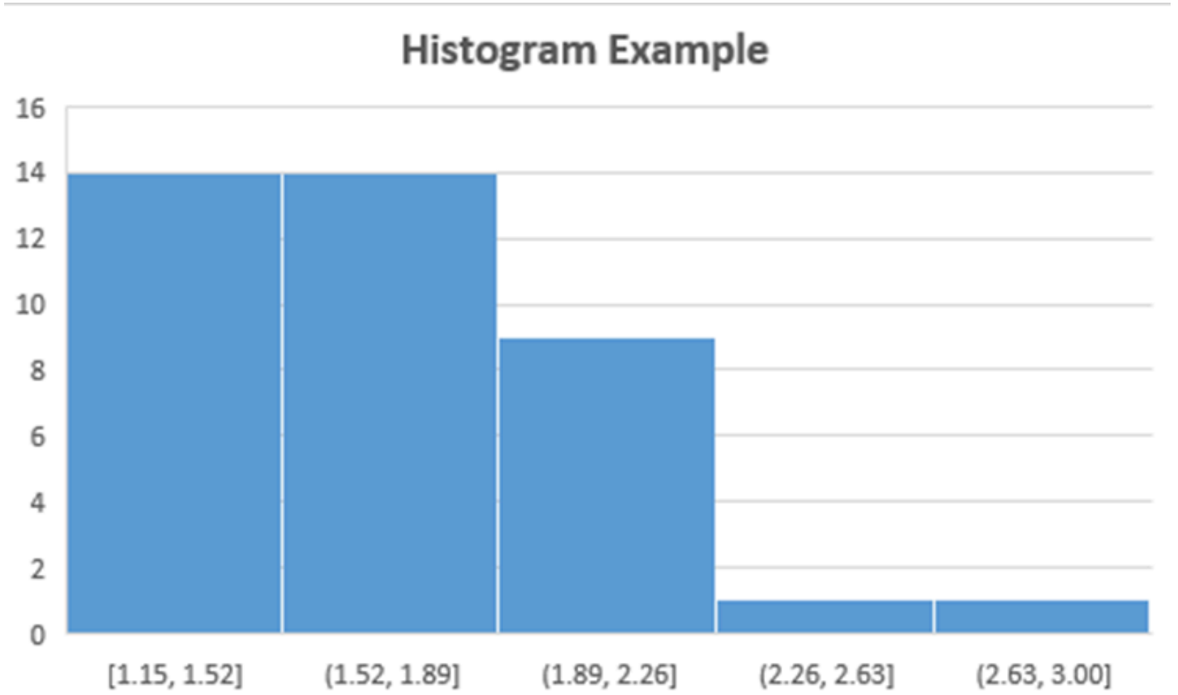
Three of the commonly used Univariate Graphics:

1. **Stem-and-Leaf Plots**
2. Histogram
3. Box Plots

Stem	Leaf																														
1	15	22	25	30	35	4	4	5	5	5	5	5	5	50	5402	55	55	55	55	55	55	65	66	6667	75	75	75	7667	9167	95	
2	0	0	0	0	0	1	19	4231																							
3	0																														

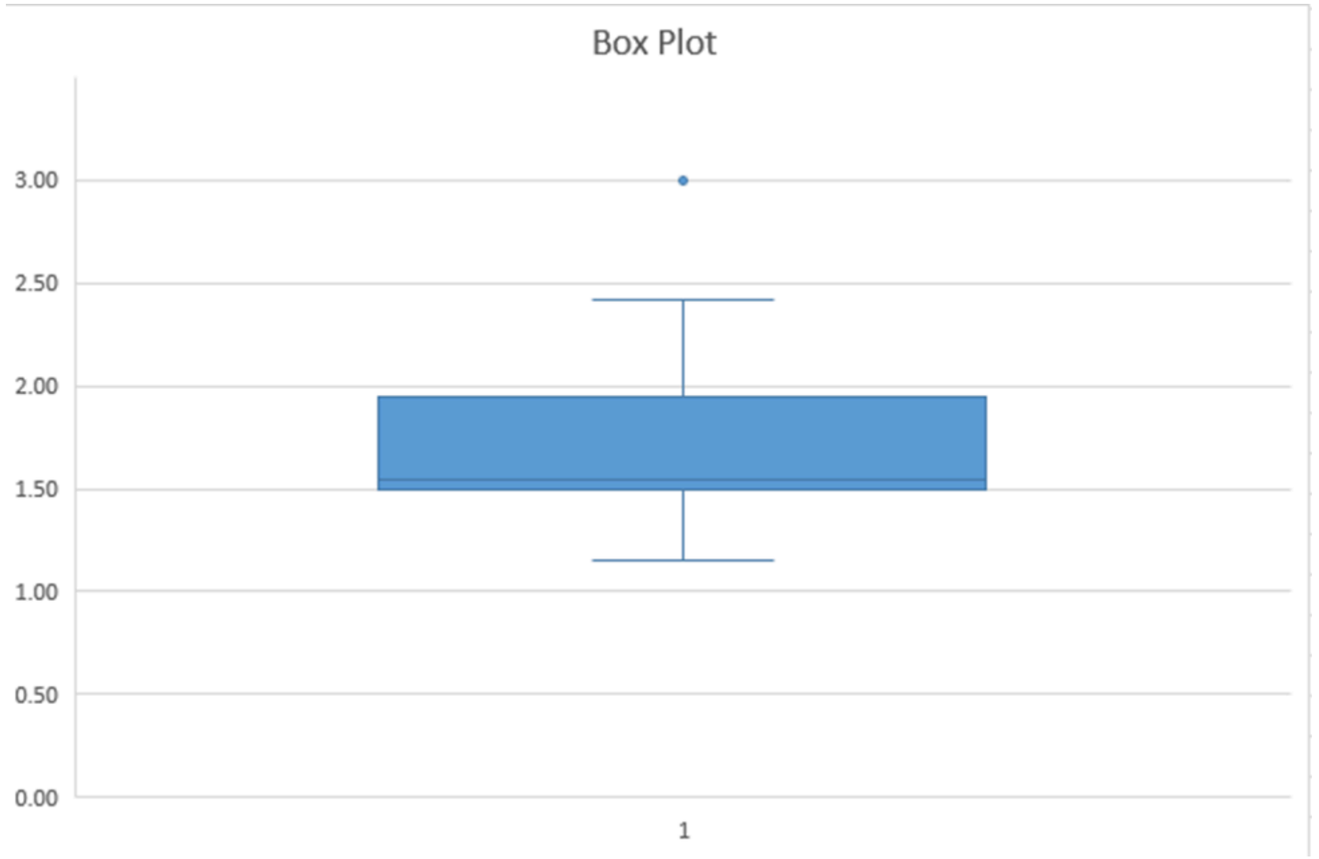
Three of the commonly used Univariate Graphics:

1. Stem-and-Leaf Plots
2. **Histogram**
3. Box Plots



Three of the commonly used Univariate Graphics:

1. Stem-and-Leaf Plots
2. Histogram
3. **Box Plots**



4 TYPES OF EDA

Multivariate Non-graphical

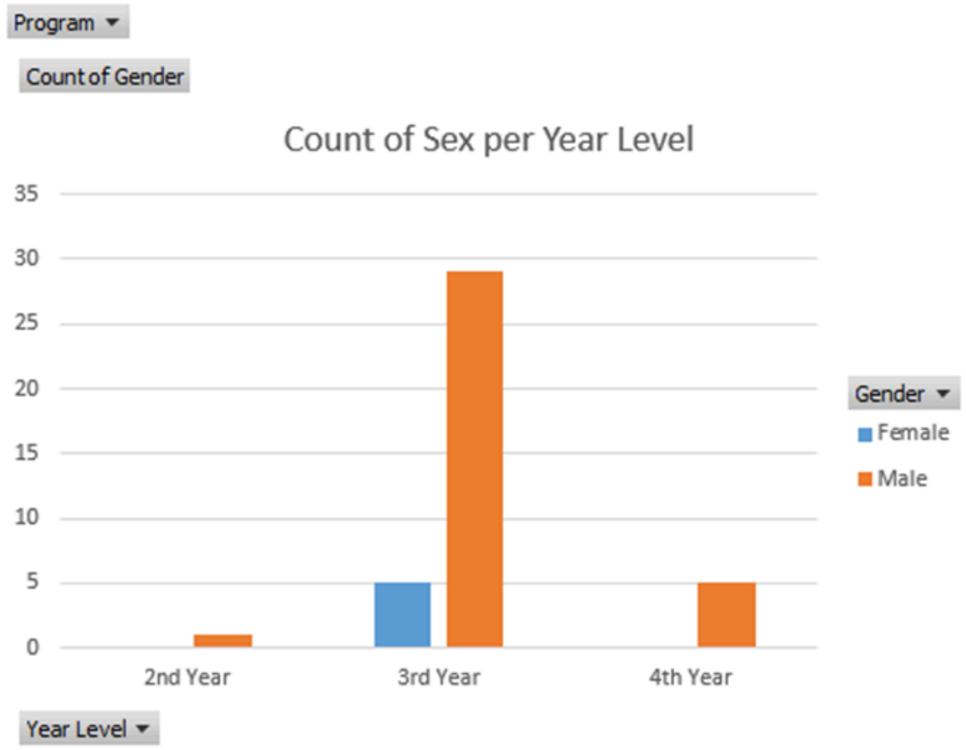
- Deals with multiple variables
- No graphical presentations
- Looks for patterns derived from multiple variables

Count of Gender	Column Labels	
Row Labels	Female	Male
2nd Year		1
3rd Year	5	29
4th Year		5

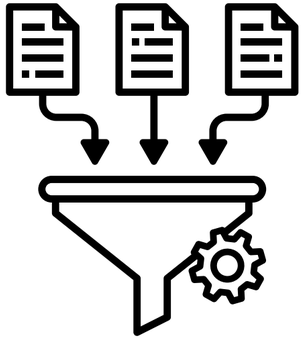
4 TYPES OF EDA

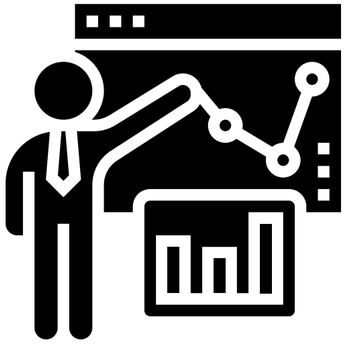
Multivariate Graphical

- Deals with multiple variables
- Uses graphical presentations to visualize data
- Displays relationship of two or more variables



PROCESS OF EDA

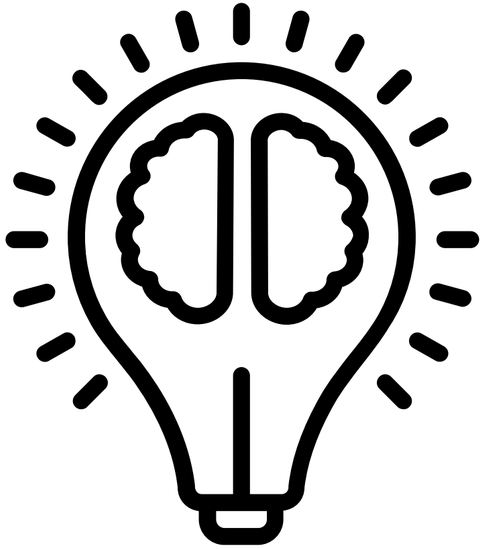




PROCESS OF EDA

Understand the Problem and the Data

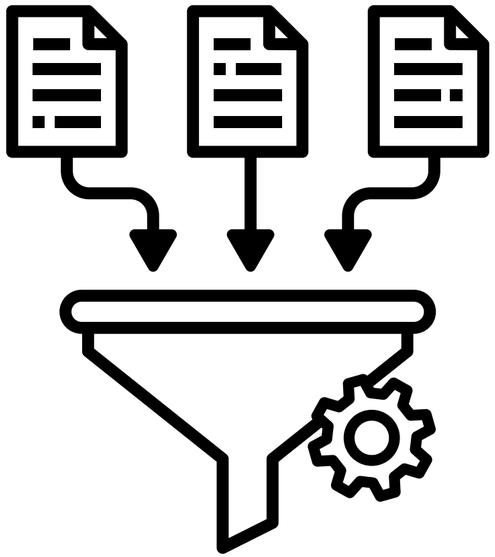
- comprehend the problem and data thoroughly
- helps data scientists in formulating guide questions
- avoids them from making incorrect assumptions



PROCESS OF EDA

Clean the Data

- Uncleaned data affects the result
- EDA helps in data cleaning
- Ensures that data is ready for exploring, summarizing, and visualizing the data



PROCESS OF EDA

Explore the Data

- Exploring data often leads to discovering meaningful insights
- It guides the scientist to understand why some phenomenon occurs in the data

PROCESS OF EDA

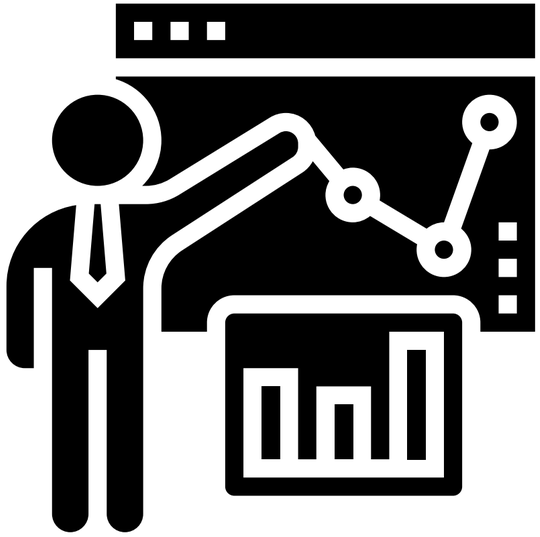
Visualize Data Relationships

- Exploring the data helps in visualizing data
- It leads the scientist to know what variables forms a relationships

PROCESS OF EDA

Communicate Findings and Insights

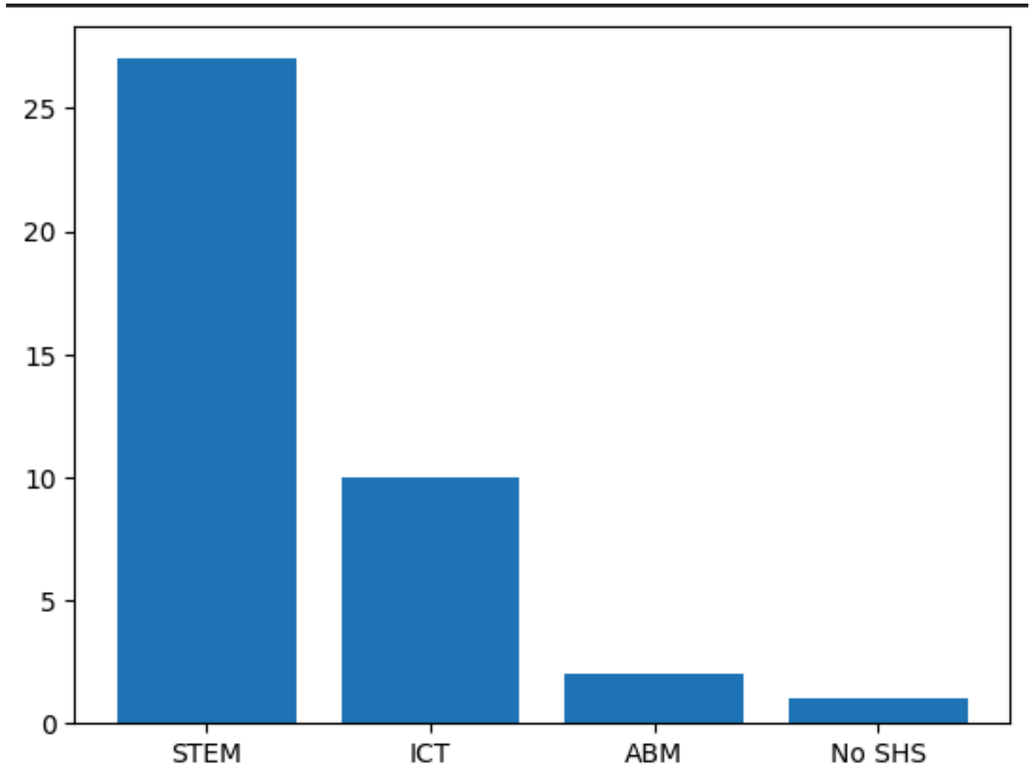
- Summarize findings, formulate insights, and discuss the result
- Helps stakeholders in making data-driven decisions



TOOLS FOR EDA

Python

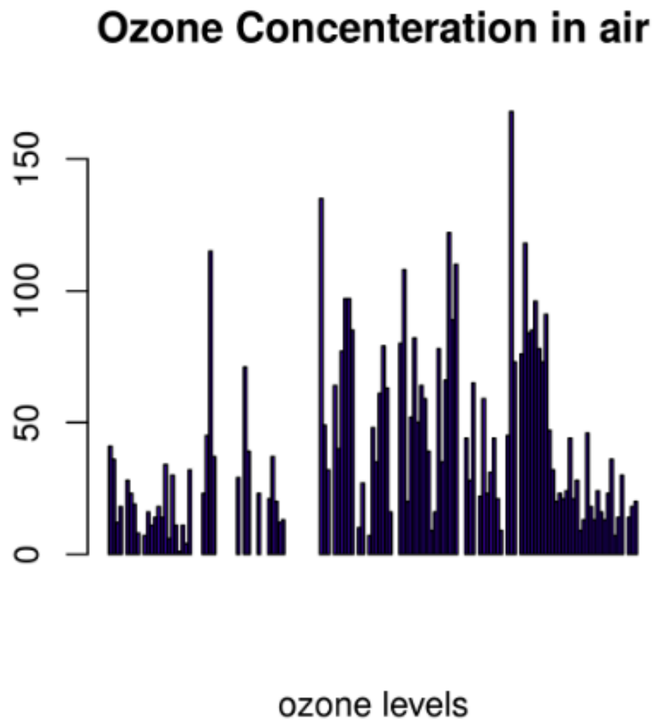
- Has multiple libraries for EDA such as Pandas, numPy, matplotlib
- Beginner-friendly



TOOLS FOR EDA

R Programming Language

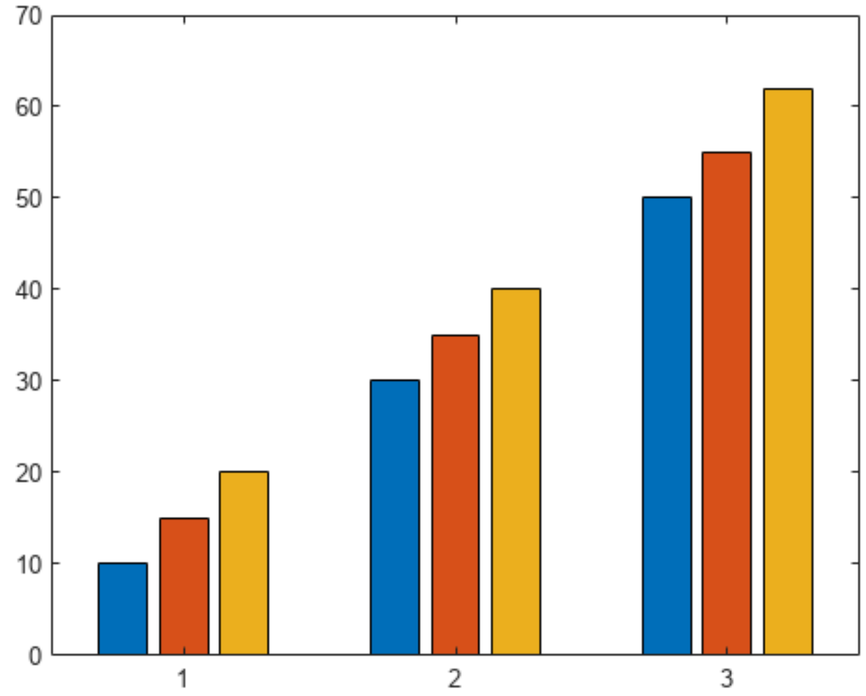
- Specifically designed for statistical analysis and data visualization
- Usually preferred for data visualizations



TOOLS FOR EDA

MATLAB

- Commonly used in the Engineering field
- Powerful mathematical computations



References:

Effective Strategies to Handle Missing Values in Data Analysis (analyticsvidhya.com)
https://www.linkedin.com/pulse/complete-guide-handling-missing-values-machine-learning-didarul-islam-1elpe
ML | Handling Missing Values - GeeksforGeeks
How to Deal with Missing Data | Master's in Data Science (mastersindatascience.org)
Missing Data Imputation using Regression (kaggle.com)
Model-based Methods for Imputation | by Sourabh Gupta | Medium
Linear Interpolation Formula with Solved Examples (byjus.com)
LINEAR INTERPOLATION definition | Cambridge English Dictionary
Data Collection and Cleaning: Key Steps in Effective Data Analysis (linkedin.com)
Data Cleaning: Definition, Benefits, And How-To | Tableau
ML | Overview of Data Cleaning - GeeksforGeeks
https://www.statisticshowto.com/probability-and-statistics/outliers/
https://www.analyticsvidhya.com/blog/2016/03/identifying-and-dealing-with-outliers-3-different-ways/
https://www.datacamp.com/community/tutorials/simple-guide-outliers-python
Top ten ways to clean your data - Microsoft Support
https://campus.datacamp.com/courses/understanding-data-science/preparation-exploration-and-visualization?ex=6 https://www.ibm.com/topics/exploratory-data-analysis
https://www.linkedin.com/advice/3/what-benefits-limits-exploratory-data-analysis-skills-data-analysis
https://www.knowledgehut.com/blog/data-science/eda-data-science

GROUP 1

THANK YOU

11 June, 2024