# SUMMARY: LIFECYCLE OF DATA ANALYSIS

## SESSION OVERVIEW:

By the end of this session, students will be able to:

- Understand the **lifecycle of data analysis** and its importance with the help of real world examples.
- Understand the **data cleaning methods**.
- Differentiate between **tools** used in each stage of the **data analysis lifecycle**.

## KEY TOPICS AND EXAMPLES:

- **Understand the lifecycle of data analysis:**

### a. Objective:

To begin, precisely identify the issue that has to be resolved. With data analysis, what precise question or goal are you trying to answer? Make sure the problem statement is clear and applicable.
This involves understanding the context, identifying stakeholders, and defining specific goals or questions you seek to answer with the analysis.
Also ensure which timeframe is asked in the analysis. Data alignment with the timeframe asked in the analysis is very important.

***EXAMPLE:***

Let’s suppose there is a given problem where we need to figure out the general areas where the higher income groups and luxury houses reside in the state of Goa.

### b. Getting relevant data:

Once the problem is defined, gathering data becomes the next important step which will help to address the problem. The data collected must be a representation of the problem.

***EXAMPLE:***

In this step, we need to gather relevant data w.r.t the problem statement mentioned in the previous point. The variables of the data could be the areas of Goa which might include the name of the area with longitude and latitude, the average income of the citizens in that particular area, the average house price they are residing on, etc.

### c. Understanding the data:

This step helps us understand the format of data that has been collected.
Determination of data types of every column in the data set. If there is any column which is not very important for the analysis concern.
For this we ensure that there are sufficient samples per attribute of the objective solution. We need to ensure sufficient representation per attribute in the data. For

example, we need to ensure that all key areas in Goa are covered in the data to ensure completeness of the analysis. If key areas are missing, it should be highlighted while communicating the analysis.

### d. Data cleaning and exploratory data analysis:

Investigate the gathered data to learn about its relationships, structure, and quality. Examining summary data, displaying distributions, spotting outliers, and seeing patterns or trends are all included in this. Cleanse and preprocess the data to guarantee its accuracy and dependability for analysis by dealing with missing information, eliminating duplicates, and fixing mistakes.

#### *a. Merging tables:*

Data might be in different tables or files and become difficult for the analyst to perform the analysis in different files. Thus, it becomes important to bring all the relatable data at one place. Thus,  Data analysts will perform merging. Files should have at least one common column by which they can be merged.

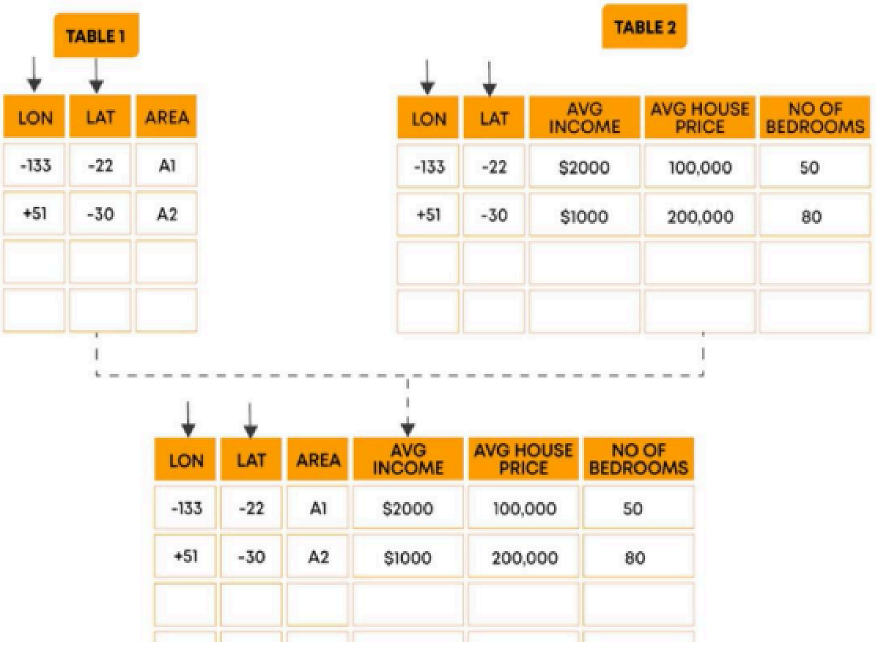


TABLE 1

| LON | LAT | AREA |
|---|---|---|
| -133 | -22 | A1 |
| +51 | -30 | A2 |
| | | |
| | | |

TABLE 2

| LON | LAT | AVG INCOME | AVG HOUSE PRICE | NO OF BEDROOMS |
|---|---|---|---|---|
| -133 | -22 | $2000 | 100,000 | 50 |
| +51 | -30 | $1000 | 200,000 | 80 |
| | | | | |
| | | | | |

| LON | LAT | AREA | AVG INCOME | AVG HOUSE PRICE | NO OF BEDROOMS |
|---|---|---|---|---|---|
| -133 | -22 | A1 | $2000 | 100,000 | 50 |
| +51 | -30 | A2 | $1000 | 200,000 | 80 |
| | | | | | |

*Figure 1 The content of Table1 is merged with Table2. LON & LAT are the common features.*

**EXAMPLE:**

In the above diagram, there are two tables.
Table1- Details of areas.
Table2- Details of the Income and houses.
Representation of the merging of two tables with the help of two common columns/variables in both the tables.

**How to merge two tables without having any common attributes:**

- We can merge two tables without having any common attribute with the help of power query (will be discussed later), VBA macros.

***b. Handling data types:***

Every column has a definite data type. All the variables should be in an expected format. For example, all the date columns, like birth date, transaction date, etc., need to be in date format.

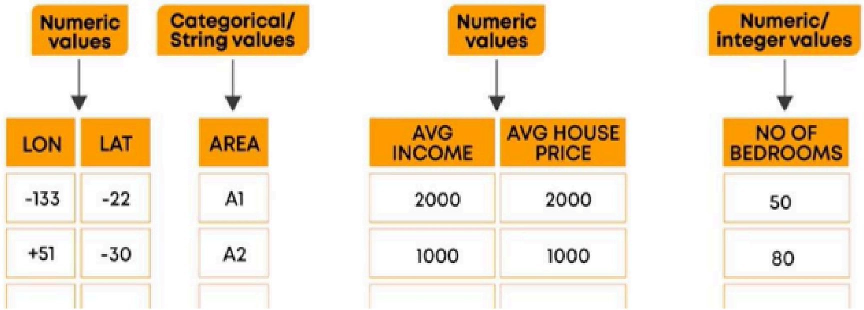


*Figure 2 Different Datatypes (numeric, string) in a Dataset*

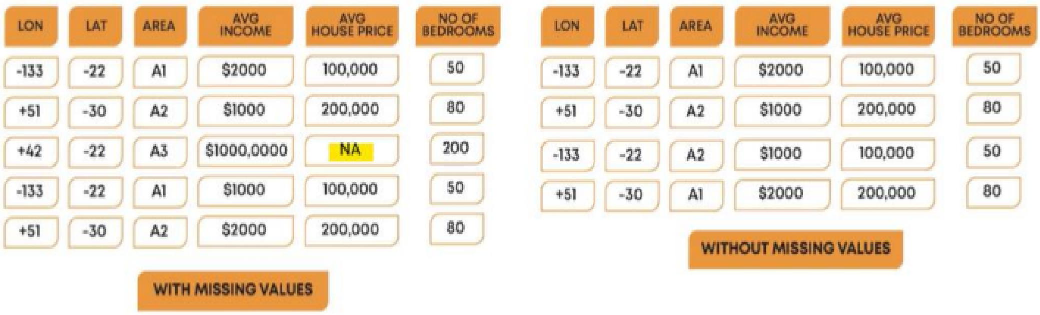
**c.** ***Handling missing values:***

In the dataset there are chances some data is not available for any attribute. That data is known as missing values. You can ignore those values if that particular attribute with missing values is not required for your data analysis. Still, if that specific column is necessary for the analysis, you must handle those missing values.

There are various method lines to handle these missing values. (Explain briefly to introduce the concepts, as these topics will be explained elaborately in the upcoming sessions).

- Deleting those rows (if missing value rows are too few compared to the whole dataset).
- Putting zero or replacing the value using the statistics method (Mean/Median/Mode, moving averages etc., all these methods depend on the dataset and the need for the attribute in the analysis).

Handling missing values are very important as these can impact the analysis and will end up in a faulty analysis. (In several scenarios, it must be consulted with the stakeholders to consider which type of method should be used to handle missing values)



| LON | LAT | AREA | AVG INCOME | AVG HOUSE PRICE | NO OF BEDROOMS |
|---|---|---|---|---|---|
| -133 | -22 | A1 | $2000 | 100,000 | 50 |
| +51 | -30 | A2 | $1000 | 200,000 | 80 |
| +42 | -22 | A3 | $1000,0000 | NA | 200 |
| -133 | -22 | A1 | $1000 | 100,000 | 50 |
| +51 | -30 | A2 | $2000 | 200,000 | 80 |

WITH MISSING VALUES

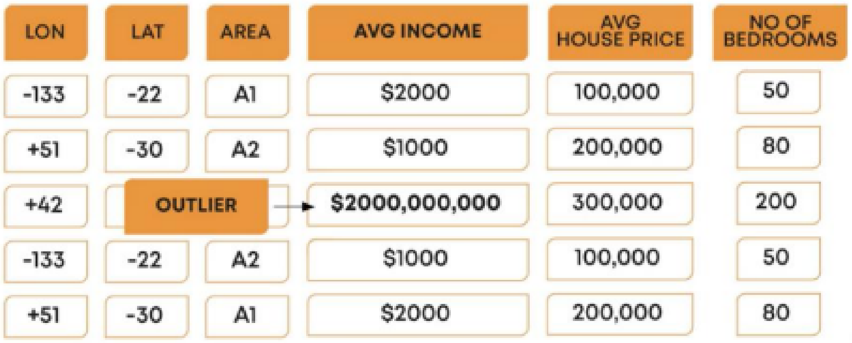
| LON | LAT | AREA | AVG INCOME | AVG HOUSE PRICE | NO OF BEDROOMS |
|---|---|---|---|---|---|
| -133 | -22 | A1 | $2000 | 100,000 | 50 |
| +51 | -30 | A2 | $1000 | 200,000 | 80 |
| -133 | -22 | A2 | $1000 | 100,000 | 50 |
| +51 | -30 | A1 | $2000 | 200,000 | 80 |

WITHOUT MISSING VALUES

*Figure 3 Tables with missing value (represented as NA) and without missing values*

d. ***Handling Outliers:***

Outliers are the data points or values which do not fall into all the other values. For example, if all values in a column are between 1 to 100 and there are two values which are 2000 and 3000, then these two will be outliers. Outliers can be an error or exceptional cases in any particular data.



| LON | LAT | AREA | AVG INCOME | AVG HOUSE PRICE | NO OF BEDROOMS |
|---|---|---|---|---|---|
| -133 | -22 | A1 | $2000 | 100,000 | 50 |
| +51 | -30 | A2 | $1000 | 200,000 | 80 |
| +42 | OUTLIER → |  | $2000,000,000 | 300,000 | 200 |
| -133 | -22 | A2 | $1000 | 100,000 | 50 |
| +51 | -30 | A1 | $2000 | 200,000 | 80 |

*Figure 4 In the Average Income column, the value $2000,000,000 is an outlier*

e. ***Ensuring data consistency*** :
Making all texts within a column consistent (Eg: ensuring all prices are in Rs or $)

f. ***Ensuring case consistency***:
Proper and consistent case should be followed for string columns

g. ***Dates*** - Ensuring all different date columns in the data are following same format

h. ***Removing unnecessary columns:***
The data might have additional confusing columns, we need to remove them to avoid confusion.

e. **Data analysis:**

- Based on the problem definition and data exploration, formulate hypotheses or research questions that you aim to test or investigate using the data. These hypotheses will guide your subsequent analysis efforts and help focus your analysis on relevant insights.

- Apply the chosen analysis techniques to the data to extract insights and answer the research questions or hypotheses formulated earlier. This may involve running statistical tests, Univariate or bivariate analysis which includes correlation analysis or performing other analytical procedures as appropriate.

### f. Visualisation:

The analysis done in the above point must be converted in the form of some visualisation which makes it easier for the stakeholders to understand, with the help of some data visualisation tool.

***EXAMPLE:***

A threshold could be set for the income above which it will be considered as a higher income category.
A threshold could be set for the prices of the house above which it will be considered as the luxury house.
A representation with the help of a map chart in a visualisation tool could be useful to showcase the areas where the higher income and luxury houses reside.

### g. Communication:

Communicate the findings of the data analysis in a clear, concise, and compelling manner to stakeholders or decision-makers. Tailor the communication to the audience, focusing on key takeaways and actionable recommendations.
It is also important to ensure complete information is passed to stakeholders. Missing information while communicating information might lead to incorrect insights delivery.
These results should be in such a way they should be understandable to the stakeholder, so it is always good to have them in pictorial form or less technical content.

## Importance of maintenance of the flow of the lifecycle of data analysis:

- **Data Consistency:** By maintaining the flow of the data analysis lifecycle, you ensure consistency in data collection, processing, and analysis methods. Consistent practices minimise errors and discrepancies in the data, leading to more reliable insights and conclusions.
- **Quality Assurance:** Regular maintenance of flow in the data helps identify and address issues with data quality, such as missing values, outliers, or inconsistencies. By continuously monitoring and cleaning the data, you improve the overall quality and reliability of the analysis results.
- **Timeliness:** Data analysis often involves handling large volumes of data collected over time. Maintaining the flow of the life cycle ensures that data is processed and analysed in a timely manner, allowing organisations to make informed decisions quickly and efficiently.
- **Adaptability to Changes**: The data landscape is constantly evolving, with new data sources, technologies, and analytical methods emerging regularly. Maintenance of the

data analysis flow enables organisations to adapt to these changes effectively, ensuring that analysis practices remain up-to-date and relevant.

- **Tools/software for Data analysis:**

1. **Data collection:**
   - **Database Management System**: MySQL, MS SQL server, Oracle Database, laptop storage (Excel).
   - **Web scraping tools**: Selenium.
2. **Data Gathering:**
   - **ETL**: Extract, Transform, Load
   - **Software**: AWS Glue, Hadoop
3. **Data processing:**
   - ETL, Python, R, SAS, Excel (Data can be gathered through channels like email, shared drives and FTP sites).
4. **Data analysis:**
   - **Statistical Analysis tools**: Excel, Python, R, SAS.
   - **Data Visualization tool**: Tableau, Power BI.
     (The statistical analysis tools mentioned above also have the visualisation capabilities but Tableau and power BI have better features to visualise.)
5. **Presentation:**
   - **Presentation software**: MS PowerPoint, Google slides, Excel.
   - **Dashboarding Tools**: Tableau, MS Power BI.

***EXAMPLE:***

Later part of the class can be focused on explanation of the lifecycle of data analysis with the help of a different scenario. As now they are aware of the steps and the theory related to it. This example will provide a greater extent of clarification on the lifecycle of data analysis.

**Objective:**

A bakery chain owner wants to optimise their product offerings to increase sales and profitability.

**Getting relevant data:**

https://docs.google.com/spreadsheets/d/19G7VII59FPI5o-vLbmH-ZmucPxLr0SpECDvuRpLA6SI/edit#gid=339746817

*The above example is a hypothetical example which has very little data, but this isn't the case with real world data.*

**Importance of adequate sample size:**

- Adequate sample size ensures that analysis results are representative of the population, minimises bias, increases statistical power, and enhances the reliability of conclusions and decisions drawn from the data.
- In essence, sample size influences the quality and validity of the entire data analysis process, from data collection and cleaning to exploratory analysis, statistical inference, model building, and decision-making.

**Data cleaning:**

a. **Table merging:**
https://docs.google.com/spreadsheets/d/19G7VII59FPI5o-vLbmH-ZmucPxLr0SpECDvuRpLA6SI/edit#gid=0

b. **Checking data type:**

https://docs.google.com/spreadsheets/d/19G7VII59FPI5o-vLbmH-ZmucPxLr0SpECDvuRpLA6SI/edit#gid=123422596

c. **Missing values:**

https://docs.google.com/spreadsheets/d/19G7VII59FPI5o-vLbmH-ZmucPxLr0SpECDvuRpLA6SI/edit#gid=1069091265

d. **Outliers:**

https://docs.google.com/spreadsheets/d/19G7VII59FPI5o-vLbmH-ZmucPxLr0SpECDvuRpLA6SI/edit#gid=132758849
Here, 100 seems like an outlier in the setting of this dataset. But, there is a possibility that one store may be more popular for tea and thus sell way more than expected.

e. **Cleaning string characters:**

https://docs.google.com/spreadsheets/d/19G7VII59FPI5o-vLbmH-ZmucPxLr0SpECDvuRpLA6SI/edit#gid=1564599194