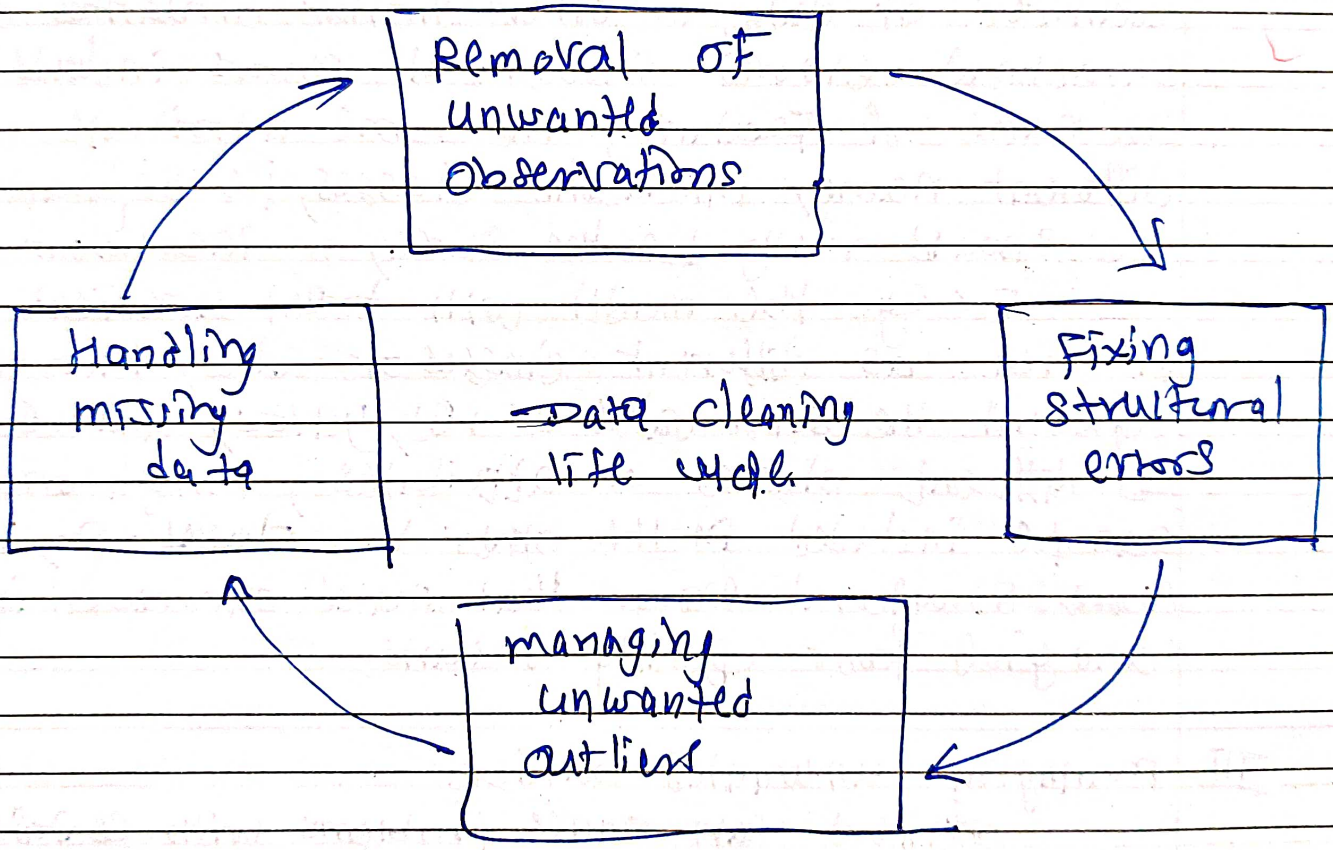
UNIT II Data cleaning

Data cleaning :-

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, dupliate or incomplete data within a dataset.

When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.



I Removal of unwanted observations

This includes deleting duplicate/ redundant or irrelevant values from your dataset. Duplicate observations most frequently arise during data collection and irrelevant observations are those that don't actually fit the specific problem that you are trying to solve.

1 Redundant observation alter the efficiency by a great extent as the data repeats and may add towards the correct side or towards the incorrect side, thereby producing unfaithful results.

2 Irrelevant observations are any type of data that is of no use to us and can be removed directly.

II Fixing structural errors

The errors that arise during measurement, transfer of data, or other similar situations are called structural errors. Structural errors include typos in the name of features, the same attribute with a different name, mislabelled classes, i.e. separate classes that should really be the same, or inconsistent capitalization.

For eg. the model will treat America and America as different classes or values, though they represent the same value or red, yellow, and red-yellow as different classes or attributes, though one class can be included in the other two classes. So, these are some structural errors that make our model inefficient and give poor quality results.

III managing unwanted outliers

outliers can cause problems with certain types of models. for eg. linear regression models are less robust to outliers than decision tree models. We should not remove outliers until we have a legitimate reason to remove them. Sometimes removing the improves performance, sometimes not. So one must have a good reason to remove the outlier, such as suspicious measurements that are unlikely to be part of real data.

IV Handling missing data:-

Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important. The two most common ways to deal with missing data are:

1 Dropping observations with missing values:

- The fact that the value was missing may be informative in itself.
- In the real world, you often need to make predictions on new data even if some of the features are missing.

2 Imputing the missing values from past observations

- "missingness" is almost always informative in itself. and you should tell your algorithm if a value was missing.
- Even if you build a model to impute your values, you are not adding any real information. you are just reinforcing the pattern already provided by other features.

Data cleaning Tools

1 OpenRefine
2 Trifacta Wrangler
3 Winpure clean or match
4 TIBCO clarity
5 melissa clean suite
6 IBM Infosphere quality stage
7 Data ladder

Why data cleaning is necessary?

There are many reasons why data cleaning is essential. list is below:

1 **Efficency:-**

Having clean data can help you in performing your analysis a lot faster. you had save a considerable amount of time by doing this task beforehand. when you clean your data before using it you had be able to avoid multiple errors. if you use data containing false values, your results wont be accurate.

2 **Error margin:-**

when you don't use accurate data for analysis you will surely make mistakes. Data cleansing helps you in that regards full stop it is a widespread practice. and you should learn the methods used to clean data.

3 **Determining data quality:**

The validity of your data is the degree to which it follows the rules of your particular requirements.

Validity errors take place when the if method is n't properly inspected. There are multiple kinds of constraints your data has to conform to for being valid.

eg. Range
Datatype
compulsory constraints
Cross field examination
unique requirements
set membership restrictions
regular pattern

4 **Accuracy:-**

you will have to focus on establishing its accuracy. even though the data is valid, it doesn't

it doesn't mean the data is accurate. and determining accuracy helps you to figure out if the data you entered was accurate or not.

5 Completeness :-

It's nearly impossible to have all the information you need. completeness is the degree to which you know all the required values. completeness is a little more challenging to achive than accuracy or validity.

6 Consistency :-

You can measure consistency by comparing 2 similar systems. you can check the data values within the same dataset to see if they are consistent or not. consistency can be relational.

7 Uniformity :-

You should ensure that all the values you have entered in your dataset are in the same unit.

Data consistency :—

Measuring data consistency can tell a researcher how valuable and useful their data. The term data consistency can be confusing. There are 3 versions of it. When the term is applied to databases. It describes data consistency within the database. When used with computing strategies, data consistency is focused on the use of data caches. The 3rd version of data consistency is used with data Analytics.

Data consistency deals with format transformations duplicated data. and missing information.

A lack of data consistency significantly increases the chances data with the system is not uniform, which would result in missing or partial data. There are normally 3 kinds of data consistency:

1 Point to time consistency :—

This focuses on ensuring all data within the system is uniform at a specific moment in time. This process prevents a loss of data if the system crashes or there are other problems in n/w. it operates by referencing bits of data in the system by way of time stamps and other consistency markers. This allows the systems to restore itself to a specific point in time.

2 Transaction consistency :—

It is used to detect incomplete transactions and roll back the data if an incomplete transcation is found.

3 Application consistency :—

It works with the transcation consistency that exists between programs. If a banking program is communicating with a tax program, application consistency promotes uniform formats between the two...

Heterogeneous data

Heterogeneous data are any data with high variability of data types and formats. They are possibly ambiguous and low quality due to missing values, high data redundancy, and untrustworthiness. It is difficult to integrate heterogeneous data to meet the business information demands.

Homogenous data :-

Homogenous data structures are ones that can only store a single type of data. (numeric, integer, character etc.).

Missing data :-

Missing data or missing values is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data.

missing data are typically grouped into 3 categories.

1 Missing completely at random (MCAR)

when data are MCAR, the fact that the data are missing is independent of the observed and unobserved data.

2 missing at random (MAR)

3 missing not at random (MNAR)

How do you deal with missing data?

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data:-

The imputation method develops reasonable guesses for missing data. it's most useful when the percentage of missing data is low.

What are the causes of missing data.?

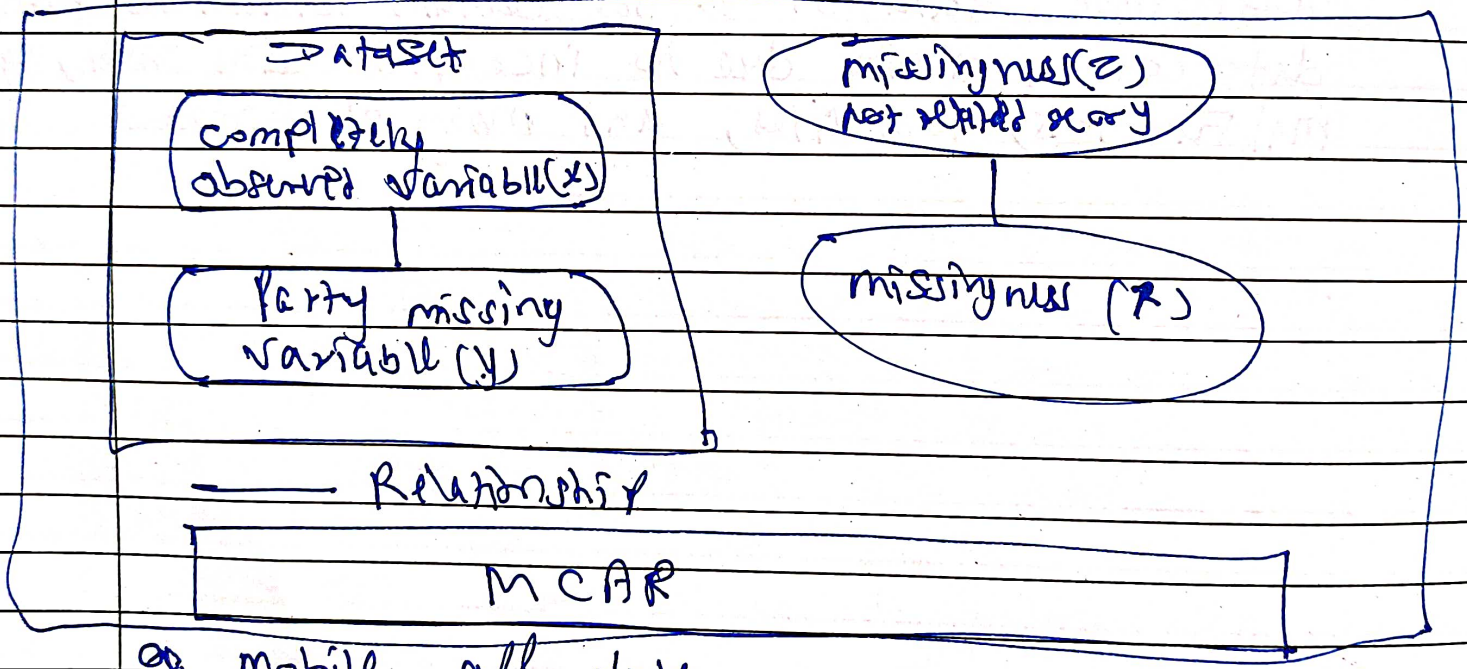
missing data or missing values, occur when you don't have data stored for certain variables or participants. data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons.

Missing Data Imputation Techniques:

Missing datas are defined as values that are not available and that would be meaningful if they are observed.

I Missing completely at Random (MCAR)

When we say data are missing completely at random. We mean that the missingness has nothing to do with the observation being studied. (completely observed variable (x) and partly missing variable (y).

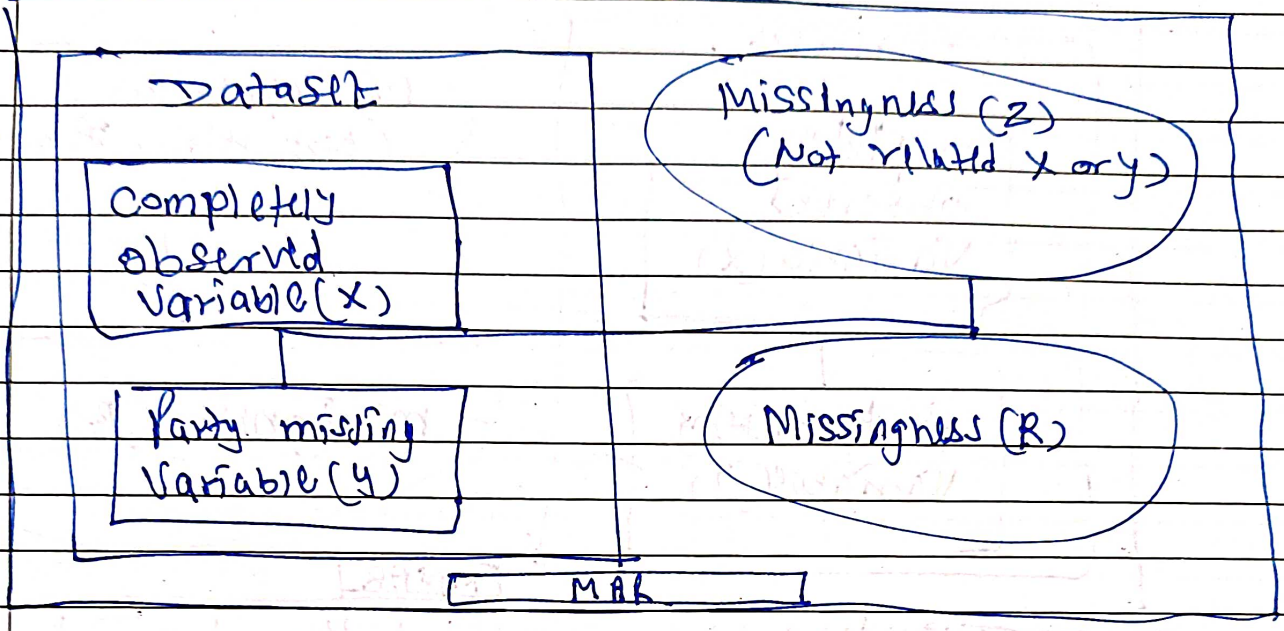


eg mobile app data.

mobile id	mobile purpose (x)	Download speed (y)	missing because of data limit fully utilized (z)
1	fast	157	
2	lite	99	
3	fast	167	
4	fast	N/A	

It might be able to predict R. from other variable.

II Missing at random (MAR)

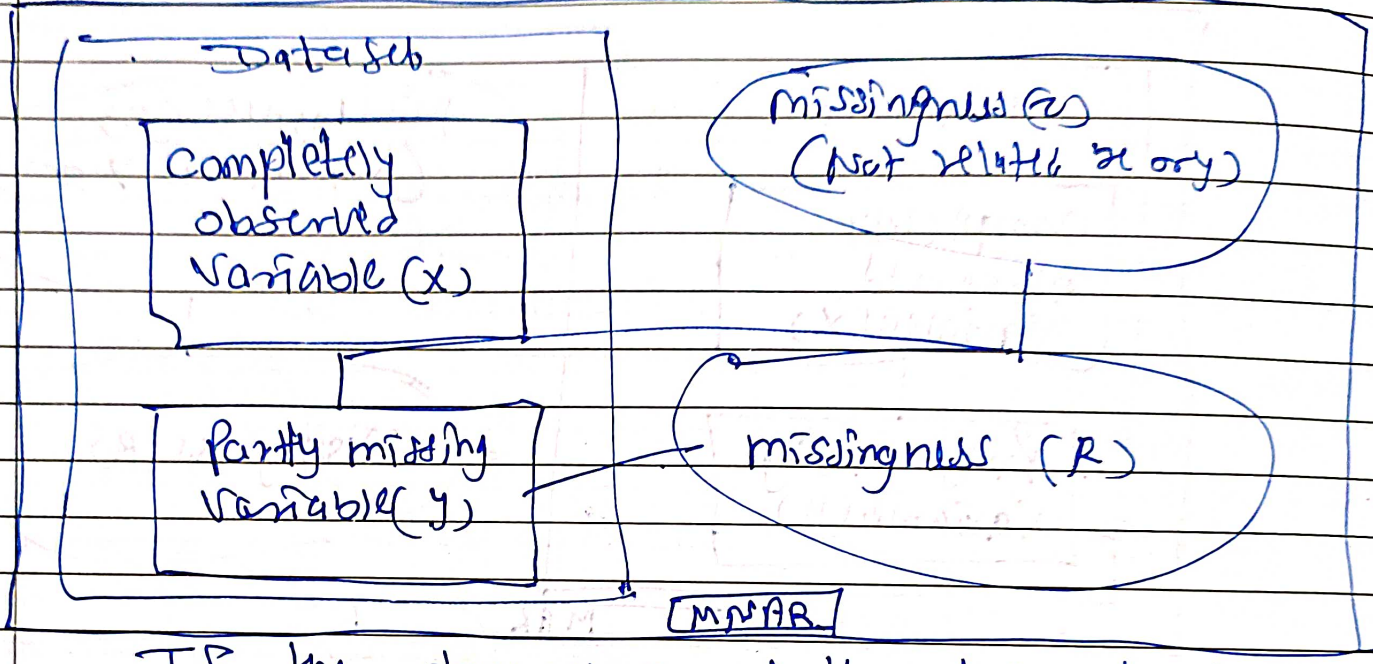


When we say data are missing at random. We mean that missing data on a partly missing variable (Y) is related to some other completely observed variables (X) in the analysis should be model but not to the values of Y itself.

Eg. if a child does not attend an examination because the child is ill.

mobile id	mobile package	download speed (X)	data unit usage (Y)	when data limit usage reached 100% missing has occurred. missing depends on other observed variable. (R, Z)
1	Fast +	157	80%	
2	Lite	99	70%	
3	Fast +	167	10%	
4	Fast +	NIL	100%	

III Missing not at Random (MNAR)

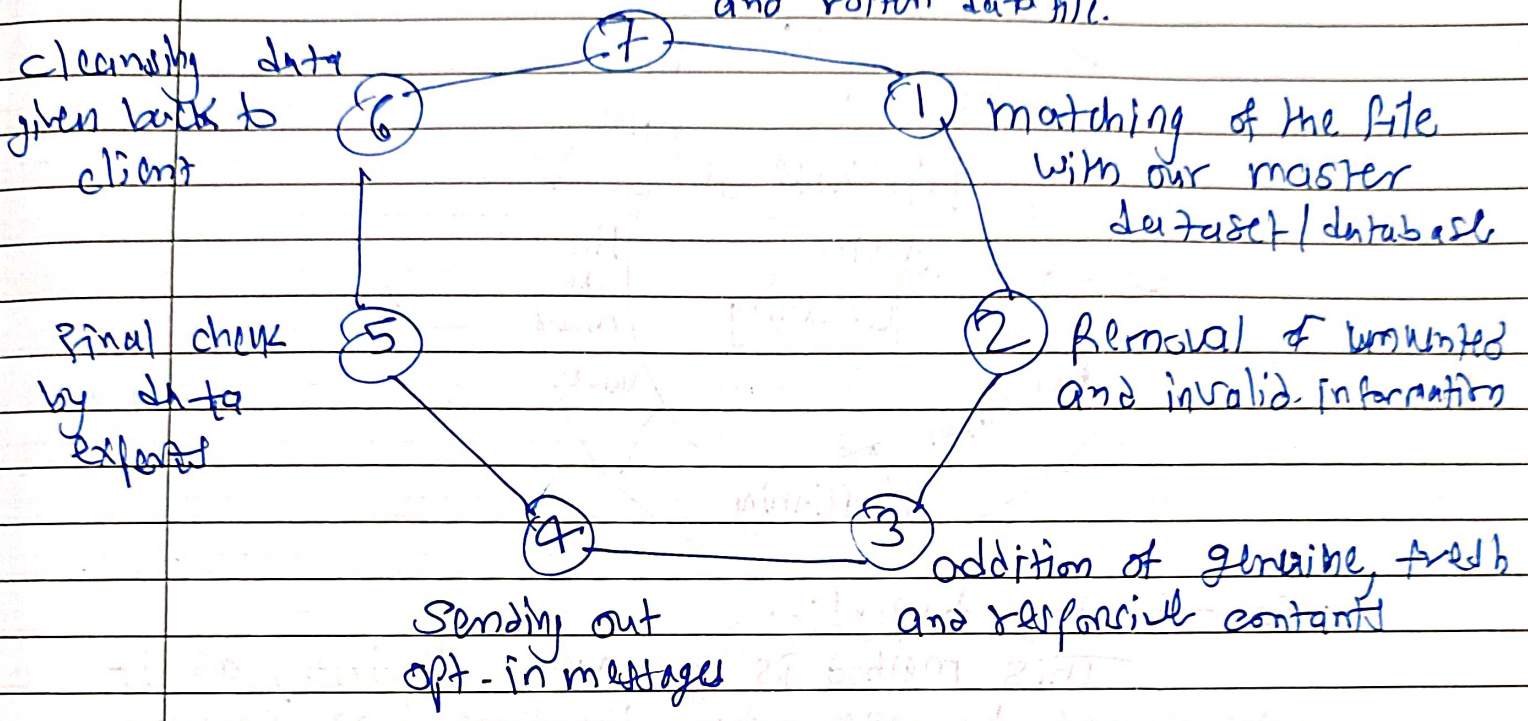


If the characters of the data do not meet those of MCAR or MAR, they fall into the category of missing not at random (MNAR). When data are missing not at random, the missingness is specifically related to what is missing.

mobile id	mobile package	download speed	data limit usage	download speed limit missing value has relation to download speed data limit usages and some other unknown value
1	Fast+	N/A	80%	
2	Lite	99	70%	
3	Fast+	167	16%	
4	Fast+	N/A	75%	

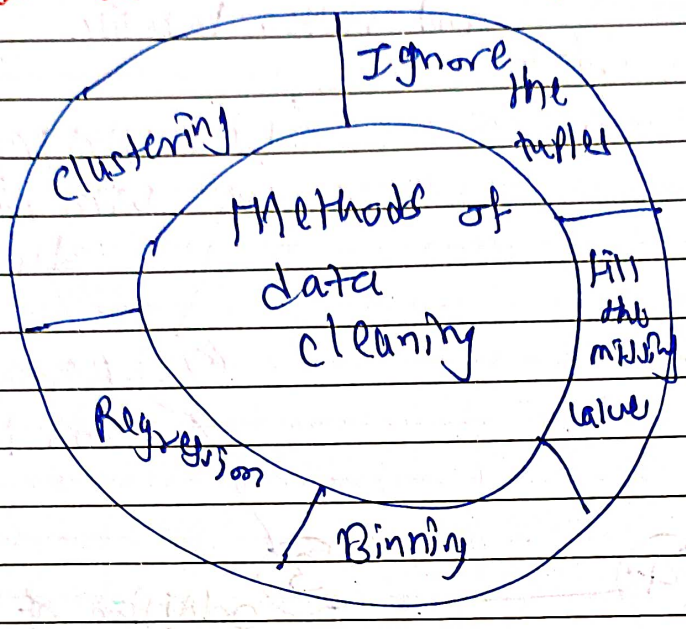
Here value is missing beyond a data limit usage range but we can not predict the value.

Data cleansing receiving incomplete and rotten data file.



Data cleansing or data cleanity is the process of detecting and correcting corrupt or inaccurate records from a record set, table or database and refers to identifying incomplete, incorrect inaccurate or irrelevant parts of the data and then replacing modifying or deleting the

Methods of data cleaning



I **Ignore the tuples:-**
This method is not very feasible, as it only comes to use when the tuple has several attributes & has missing values.

II **Fill the missing value:-**
This approach is also not very effective or feasible. It can be a time consuming method. In the approach, one has to fill in the missing value. This is usually done manually, but it can also be done by attribute mean or using the most probable value.

III **Binning Method:-** This approach is very simple to understand. The smoothing of sorted data is done using value around it. The data is then divided into several segments of equal size. After that the different methods are executed to complete the task.

IV **Regression:-** The data is made smooth with the help of using the regression function.

The regression can be linear or multiple. linear regression has only one independent variable, and multiple regressions have more than one independent variable.

V. clustering:- This method mainly operates on the group. clustering groups the data in a cluster. Then the outliers are detected with the help of clustering.

Process of data cleaning

I Monitoring the errors:-
Keep a note of suitability where the most mistakes arise. It will make it easier to determine and stabilize false or corrupt information, information is especially necessary while integrating another possible alternate with established management sys.

II Standardize the mining process:-
Standardize the point of insertion to assist and reduce the chances of duplicity.

III Validate data accuracy:-
Analyze and invest in data tools to clean the record in real time. Tools used artificial intelligence to better examine for correctness.

IV Scrub for duplicate data:-
Determine duplicates to save time when analyzing data, frequently attempted the same data can be avoided by analyzing and investing in separate data erasing tools that can analyze rough data in quantity and automate the operation.

V Research on data:-
Before this activity, our data must be standardized, validated and scrubbed for duplicates. There are many third party sources and these approved and authorized parties sources can capture information directly from our databases. They help us to clean and compile the data to ensure completeness, accuracy and reliability for business decision making.

VI Communicate with the team:- Keeping the group in the loop will assist in developing and strengthening the client and sending more targeted data to prospective customers.

Justify data cleaning affect the quality of data?

1 Invalid values:-
Some datasets have well known values. eg. gender must be only 'F' for female and 'M' male. In this case it is easy to detect wrong value.

2 Formats :- The most common issue, its possible to get values in different formats like a name written as "name, surname".

3 Attribute dependencies:—
when the value of a feature depends on the value of another feature. for eg. if we have some school data, the no. of students is related to whether the person is "teacher". If someone is not a teacher he/she can't have any student.

4 Uniqueness:- It's possible to find repeated data in features that only allow unique values. for eg. we can't have two products with the same identifier.

5 missing values:— Some features in the dataset may have blank or null values.

6 misspellings — In correctly written values.

7. misfielded values — when a feature contains the values of another.

Data Transformation

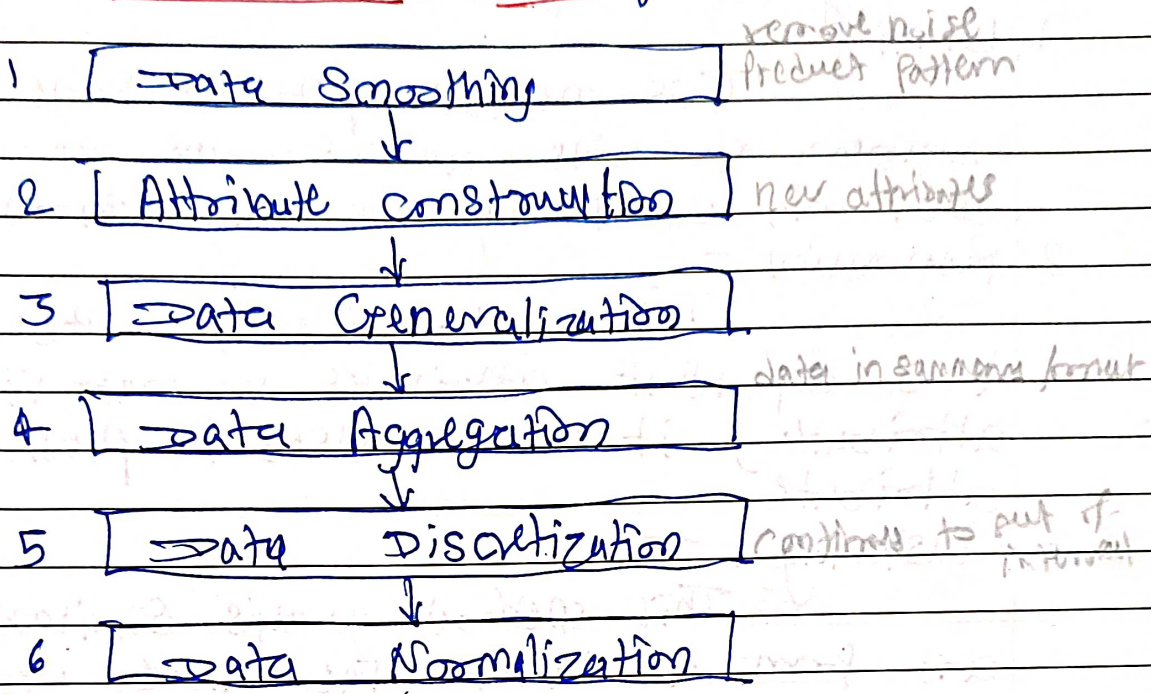
Raw data is difficult to trace or understand. Data transformation is technique used to convert the raw data into a suitable format that efficiently eases data science and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

Data transformation is an essential data preprocessing technique that must be performed on the data before data mining to provide patterns that are easier to understand.

Data transformation changes the format, structure, or values of the data and converts them into clean, usable data. Data may be transformed at two stages of the data pipeline for data analytics projects. organizations that use on-premises data warehouses generally use an ETL (extract, transform, and load) process, in which data transformation is middle step.

Most organizations use cloud based data warehouses to scale compute and storage resources with latency measured in seconds or minutes. The scalability of the cloud platform lets organizations skip preload transformations and load raw data into data warehouse, then transform it at query time.

Data Transformation Techniques.



I Data Smoothing :-

Data Smoothing is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. it helps in predicting the patterns. when collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

The concept behind the data Smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysis or trends who need to look at a lot of data which can often be difficult to digest for finding pattern that they wouldn't see otherwise.

We have seen how the noise is removed from the data using the techniques such as

binning, regression, clustering.

1. Binning –
This method splits the sorted data into the number of bins and smooths the data values in each bin considering the neighborhood values around it.
2. Regression –
This method identifies the relation among two dependent attributes so that if we have one attribute, it can be used to predict the other attribute.
3. Clustering :–
This method groups similar data values and form a cluster. The values that lie outside a cluster are known as outliers.

II Attribute Construction
In the attribute construction method, the new attributes consult the existing attributes to construct a new data set that eases data mining. New attributes are created and applied to assist the mining process from the given attributes. This simplifies the original data and makes the mining more efficient.
eg. dataset in graph.

III Data Aggregation :–
Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.

This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used.

Gathering accurate data of high quality and a large enough quantity is necessary to produce relevent results. The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations and marketing strategies.

IV Data Normalization

1. min - max Normalization
2. Z - Score normalization
3. Decimal Scaling

V Data Discretization

This is a process of converting continous data into set of data intervals. continous attribute values are substituted by small interval ~~labels~~ labels. This makes the data easier to study and analyze. If a data mining task handles a continous attribute, then its discrete value can be replaced by constant. quality attributes. This improves the efficiency of the task.

This method is also called a data reduction mechanism as it transforms a large dataset into a set of categorical data. Discretization also uses decision tree based algorithms to produce short, compact and accurate results when using discrete values.

VI Data Generalization:-

It converts low level data attributes to high level data attributes using concept hierarchy. This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data. Data generalization can be divided into two approaches.

① Data cube process approach
② attribute oriented induction approach.

Ways to handle missing data during cleaning

1. Mannual entry of missing data
2. Using attribute mean
3. Using most probable value. eg. DT, Regression, Predict
4. Using global constant
5. ignore the tuple.