Unit I

CHAPTER 1

Introduction to Data Science

Syllabus Topic : What Is Data?

1.1 What Is Data?

Q. 1.1.1 Explain data, information and knowledge triangle. (Ref. Sec. 1.1) (5 Marks)

- Data is defined as raw facts and figures collected together and stored in database. Data can be in structured, semi structured and unstructured format. Data are records which are collected by various ways, large number of resources generates data and this data is in different formats.
- For example, if number of males and females are counted in specific location then the values represented as no of males and females is data as it is a fact.
- Data can be processed to get information. Processing can include combining data from various sources, collecting data, converting or transforming data into a specific format, summarizing, modelling etc.
- Data can be measured, collected, presented, and analyzed by using various tools. Once the data is collected from various sources it is in the raw format and hence it is also known as raw data.
- Raw data then has to undergo through the important phase of cleaning. Data cleaning method is used to remove the garbage and unwanted values.
- Raw data collected from various source will not always provide the proper values and hence data cleaning and then data evaluating phase after collection provides assurance about the genuineness of the data. Data is usually evaluated by comparing it with some standard values or by validating it through the experts.
- Fig. 1.1.1 shows the knowledge data and information pyramid.

Fig. 1.1.1 : Data, information, knowledge triangle

➜ **1. Data**

Raw Facts and Figures.

➜ **2. Information**

Processed Data provides Information.

➜ **3. Knowledge**

Mastering the use of information in particular fashion provides knowledge.

➜ **4. Wisdom**

Application of the knowledge is known as wisdom.

Syllabus Topic : Different Kinds of Data

1.2 Different Kinds of Data

Data at its highest classification levels is of two types :

1. Quantitative data
2. Qualitative data

➜ **1. Quantitative data**

It is represented using numbers or anything through which someone can measure various dimensions such as height, weight, width, length, temperature, humidity, cost etc.

Quantitative data are further classified as discrete data and continuous data.

(a) Discrete data
(b) Continuous data

➜ **(a) Discrete data**

Data which can be counted completely is discrete data it is mostly the integer values for example number of children in family, no of players in cricket team etc, are the discrete values.

➜ **(b) Continuous data**

Continuous data is divided into the finer levels and they are usually floating point values. Example of continuous data can be height, weight, length etc.

➜ **2. Qualitative data**

It provides the characteristics and descriptors which cannot be easily measured. Qualitative data can be observed subjectively. For example smell, taste, textures, color etc.

Qualitative data are further classified as binomial, nominal and ordinal data.

(a) Binomial data
(b) Nominal data
(c) Ordinal data

➜ **(a) Binomial data**

Binomial means two values which are similar to binary data, two values can be true or false, yes or no, except or reject, right or wrong etc.

➜ **(b) Nominal data**

It is also known as unordered data here every individual element will not have a kind of ranking but it will have some categories. For example let us say there are 10 items which are having different colours so they can be categories according to color if the next value comes then it is easily categories.

➜ **(c) Ordinal data**

Ordinal data is also known as ordered data here every element have some kind of order for example short, medium, tall can be three categories for height and now if look at their names they follow some order.

Syllabus Topic : Introduction to High Level Programming Language + Integrated Development Environment (IDE)

1.3 Introduction to High Level Programming Language + Integrated Development Environment (IDE)

Q. 1.3.1 Explain high level language. (Ref. Sec. 1.3) **(5 Marks)**

1.3.1 High Level Programming Language

Programming languages provides a set of grammar to a computer program or computing devices. These languages are used to make certain application to do something or in other words the programming languages are used to give commands and instruction to the computing device.

☛ **Programming languages are of mainly three types**

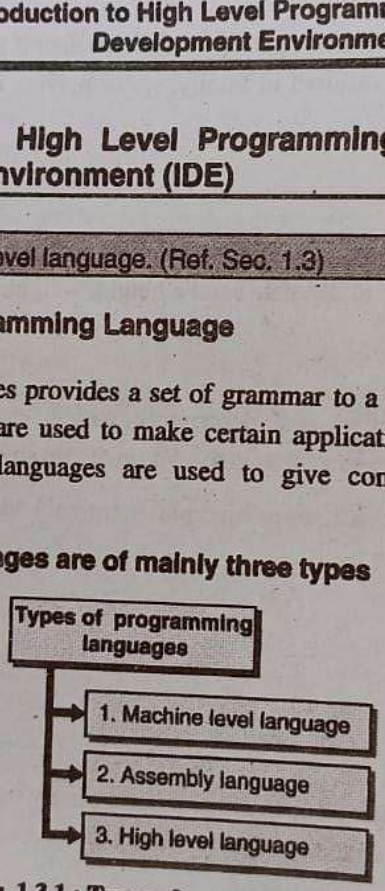

Fig. 1.3.1 : Types of programming languages

➜ **1. Machine level language**

Machine level language also known as binary or low level language consists of "0" and "1". Machine languages are mainly used by the processors in order to execute certain command. It is easy for machine to understand it but difficult for human beings to read, interpret and remember it.

2. Assembly language

Assembly language is an intermediate language in which the instructions are written in the assembly language. Assembly language can be easily understood by the IC's or chips. For example chips used in microwave, washing machine etc are given instruction in the form of assembly language. Tools like MASM, TASM are used as the editors for writing the assembly language programs. An example of the instruction in assembly language.

MOV AX, 04H

The above instruction is for 8086 microprocessor and it means that in AX register move the value 04 H, where H means in hexadecimal format. Assembly language is simpler to understand to human with respect to the binary language.

3. High level language

- It is a simplest language out of all three and it is completely human readable. It uses formats similar to english language and easy to understand.
- Every instruction in the high level language is converted to machine level language by using interpreter or compilers. There are certain advantages and disadvantages of high level language which are as follow :

Advantages of high level language

1. High level languages are user friendly.
2. They are easy to understand.
3. They are easy to learn.
4. They are problem oriented and not machine oriented.
5. They are easy to maintain.
6. Programs written in high level language can be easily converted into machine language.

Disadvantages of level language

1. They have to be converted into machine level language and hence an additional task which may need specific tool.
2. The object code generated by a translator might be inefficient when compared to equivalent assembly language program.

☞ **Types of high level languages**

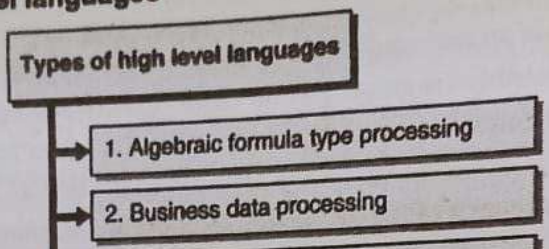

Fig. 1.3.2 : Types of high level language

➜ **1. Algebraic formula type processing**

These languages are related to mathematical and statistical problem solving techniques, for example FORTRAN, BASIC, ALGOL etc.

➜ **2. Business data processing**

These languages are used for processing the business data, for example COBOL and RPG.

➜ **3. String and list processing**

These languages used for string manipulation, pattern matching, insertion and deletion of characters and strings, for example LISP, Prolog etc.

➜ **4. Object oriented programming language**

In these types of languages the program is divided into objects, for Example C++, Java etc.

➜ **5. Visual programming language**

These languages are used for development of Windows-based applications, for example visual basic, visual java etc.

1.3.2 Integrated Development Environment (IDE)

Q. 1.3.2 Explain IDE. (Ref. Sec. 1.3.2) **(5 Marks)**

- Integrated Development Environment (IDE) is a package consisting various software's or tools which are necessary to build a complete application.IDE may include various tools or software's like text editor, compiler, linker, debugger, automation tools etc.
- Before IDE to develop application different software were used all together differently an example can be writing java code in notepad and then running it using jdk.
- Instead now a day's one can use Netbeans or Eclipsce as an IDE which has its own editor, compiler, debugger etc.

Various benefits of IDE

The main aim of IDE is to provide an environment which consists of different things together which will reduce space, time etc. Following are few benefits provided by IDE.

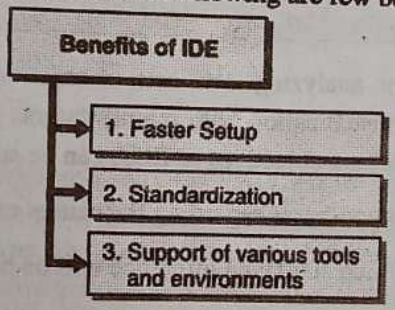

Fig. 1.3.3 : Benefit of IDE

1. Faster setup

In earlier approach when IDE was not there the required software's where installed one after another which used to take lot of memory and time as installation of each module or software use to take certain time also installing one after another will lead to delay and once installed if any error comes the process has to be repeated again which leads to wastage of time, whereas IDE provides one application file which will install all the needed software's or tools together. IDE provides option to select required component of particular software and by selecting and deselecting it memory management can be done properly.

2. Standardization

Using IDE provides similar type of environment and steps for similar type of application which will create a standard format through which the involvement of the errors can be reduced.

3. Support of various tools and environments

IDE supports many languages and tools, environment due to which for developing certain application an developer will not have to search for various other software's apart from it application created for one platform can be used in other platform too as IDE also provides platform independence.

Syllabus Topic : Exploratory Data Analysis (EDA) + Data Visualization

1.4 Exploratory Data Analysis (EDA) + Data Visualization

Q. 1.4.1 Explain EDA and data visualization. (Ref. Sec. 1.4) (5 Marks)

It is an approach used for analyzing data sets in-order to summarize their important characteristics. EDA and data visualization both mainly targets towards representation of the data in the graphical format. Various objectives of EDA can be summarized as follow :

1. To provide suggestion for hypotheses regarding the causes of observed phenomena.
2. To asses assumptions on which statistical inference will be based.
3. To support the selection of proper statistical tools and techniques.
4. To provide a base for data collection in future by considering surveys and experiments.

Various Tools which can be used by EDA are :

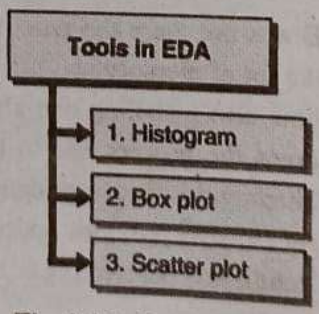

Fig. 1.4.1 : Tools in EDA

1. Histogram

It is a representation which provides the distribution of numerical data. It is an estimate of a probability distribution function.

Example of histogram is as follow :

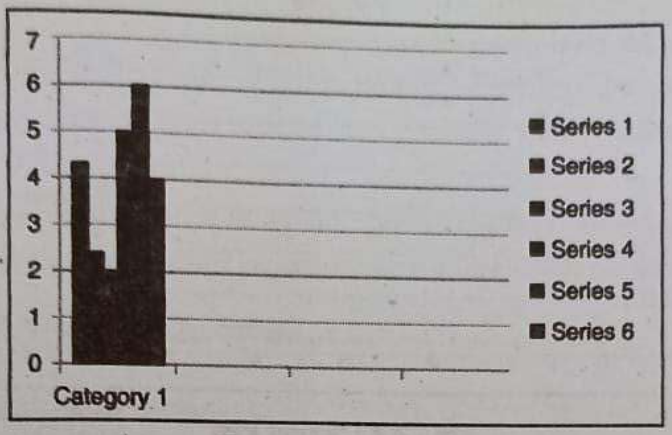

Fig. 1.4.2 : Histogram

2. Box plot

Box plot is a graphical representation method used to depict groups of numerical data through their quartiles. Box plots have lines extending vertically form the boxes indicating variability outside the upper and lower quartiles, these are called as whiskers and hence box plots are also many times called as box-and whisker plot.

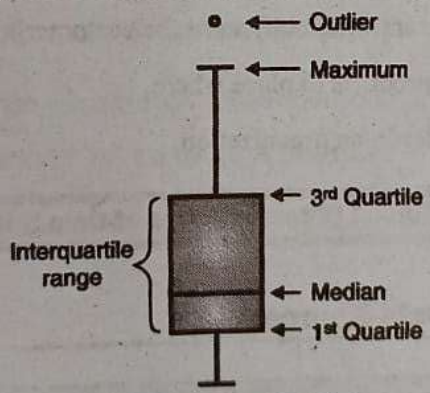

Fig. 1.4.3 : Box Plot

3. Scatter plot

It is a type of plot used to display values for mainly two variables of data set. The data in the scatter plot are represented as collection points, every point represent the position and the value related to that variable.

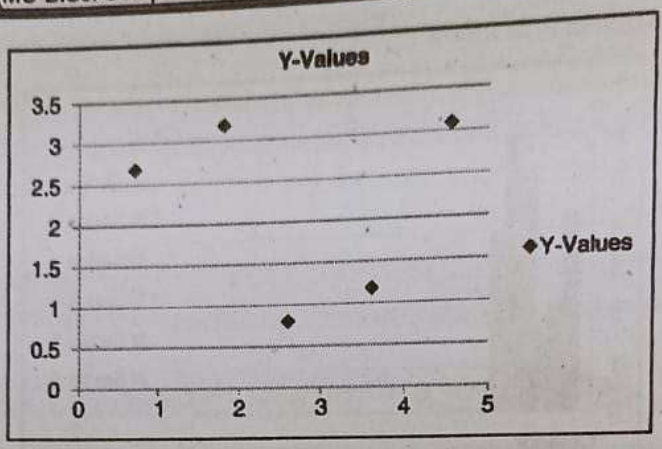

Fig. 1.4.4 : Scatter Plot

Data visualization is a process for displaying data on charts and bars. It provides the visual reporting to the users. Data visualization tools can be used as Business Intelligence (BI) reporting tool.

☛ **Data Visualization provides following benefits**

1. Helps to identify the areas that need attention or improvement.
2. Provides clarification of factors which influence the customer's behaviour.
3. Helps to understand which products to place where.
4. It predicts the volume of sales in an organization.

Syllabus Topic : Different Types of Data Sources

1.5 Different Types of Data Sources

Q. 1.5.1 List and explain various data type sources. (Ref. Sec. 1.5) **(5 Marks)**

Data sources are the resources through which the data is made available for analysis or any other use. Data available from various sources can be in structured, semi structured or unstructured format. There are various data sources which are available which includes,

1. **File systems :** File systems are for example CSV, excel files, etc.
2. **Relational Systems :** They include various databases such as Oracle, Sql Server, DB2, etc.

Unit I

CHAPTER 2

Data Management

2.1 Introduction

Q. 2.1.1 Explain data management. (Ref. Sec. 2.1) **(5 Marks)**

- It is an administrative process which includes acquiring, validating, storing protecting and processing of the data in order to ensure the ease of access, reliability and timeliness of the data. Data management process provides with wide variety of solutions to the problems faced by people for handling data.
- Data management deals with acquisition and preparation of the data for various purposes in the organization. Data management also deals with creation of back up, data storage, providing authentication, data cleaning, data extraction and analysis.
- In an organization data management should be done perfectly as it leads to optimization of various processes in the organization. Some of the best practices that an organization should have for proper data management are :

1. There should be simple access to past and present data.
2. Store data in proper format which is flexible so that any point of time it can be extracted in a particular format.
3. Scrub the data in order to provide a good quality in the existing business processes.

Managing of the data is considered to be first step towards handling large volume of data in any format produced on daily basis. Data management practices helps in data analytics.

Syllabu

2.2 Data Collection

Q. 2.2.1 Explain data collection

- Data collection is a process in v resources generate data in vari like interviews, questionnaires case studies and documents a

Data collection methods a

A. Primary data collection
B. Secondary data collecti

A. Primary data collecti

In this method the data is interviews and questionnai it's unique or undisturbed. are :

Fig

1. Questionnaires

This methods is widel be generated before da

Syllabus Topic : Data Collection

2.2 Data Collection

Q. 2.2.1 Explain data collection methods. (Ref. Sec. 2.2) (10 Marks)

- Data collection is a process in which data is gathered related to particular activity. Various resources generate data in various formats and to collect this data various means are used like interviews, questionnaires and surveys, observations, focus groups, Oral history and case studies and documents and records etc.,

Data collection methods are mainly divided into two types

A. Primary data collection methods

B. Secondary data collection methods

A. Primary data collection methods

In this method the data is collected by the observer itself using certain methods like interviews and questionnaires. The data collected in this method is first handed data and it's unique or undisturbed. Various methods included in primary data collection methods are :

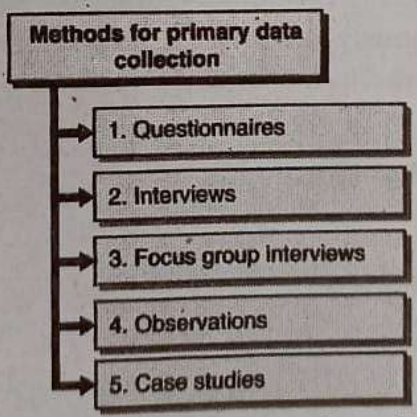

Fig. 2.2.1 : Methods for primary data collection

1. Questionnaires

This methods is widely used for collecting data, correct set of the questionnaires need to be generated before data collection in this method.

Advantages of questionnaires methods

1. It can be used in interviews or as telephonic surveys.
2. Questions can be posted or emailed.
3. Questionnaires once made can be used for large number of people.
4. They can cover large geographical area.
5. Making questionnaires involves lower cost with respect to other methods of data collection.

Disadvantages of questionnaires methods

1. Questionnaires are to be design according to the problem.
2. They should as much as simple so even layman related to that problem should be able to answer it.
3. They have low response rate.
4. Many time remainders are to be given to the responder to response to the questions.
5. It assumes that there is no literacy problem.

2. Interviews

It is a technique which is mainly used to gain an understanding of reasons and motivations of someone's attitude or behaviour. They can be one -to -one or one -to - many or in some case many- to- one also.

Advantages of Interviews areas

1. It has good response rate.
2. They provide immediate response.
3. Various recording equipments can be used.
4. Provides in depth answers related to a problem.
5. Not only provides verbal answers but also provides the details of feelings too.

Disadvantage of Interview areas

1. Lot of preparation is need like setting up the place, getting appointment of a person who will be interviewed.
2. Time constraints are there.
3. It can be expensive depending the preparation needed for certain interview.

4. Set of questions are needed.

☞ **Interviews can be classified into three types mainly**

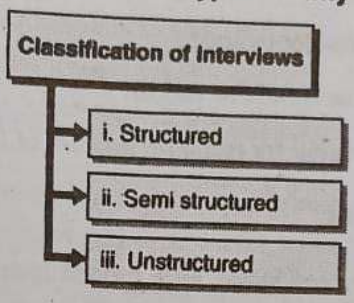

Fig. 2.2.2 : Classification of interviews

➜ **i. Structured**

It is based on a carefully worked interview schedule. It is generally done when lot of questions are not thought provoking.

➜ **ii. Semi structured**

In this type of interview the focus is to ask certain question which provides an in depth answers from the person who is been interviewed.

➜ **iii. Unstructured**

This are also know as in depth or informal interviews as there are no constrain on time and type of question. They are not specific to a particular set of problem.

☞ **Telephonic Interviews**

It is an alternate form of interview to the personal and face-to-face interview. Various Advantages and Disadvantages for this type are :

☞ **Advantages of telephonic interview**

1. They are relatively cheap as there is no need of much of prior setup.
2. They are quick with respect to personal interview.
3. They do not have location barrier which means there is no constraint on the location of interviewer and interviewee.
4. In this method call can be recorded as a proof or further use.

Disadvantages of telephonic interview

1. Questionnaires should be prepared.
2. Expressions of interviewee cannot be seen.
3. Connectivity of telephone line should be proper.
4. Interviewee should be little faster for response because of time constraint.

3. Focus group interviews

Focus Group Interviews are conducted by a trained interviewer in a non-structured and natural way with a small group of interviewee's. They are sort of discussions; the required answers for particular problem are discussed and concluded from the talk.

4. Observations

Data collection using observation methods involves recording and or tracking the behaviour of a person or various objects and events in a specific way.

Observation methods are further classified as,

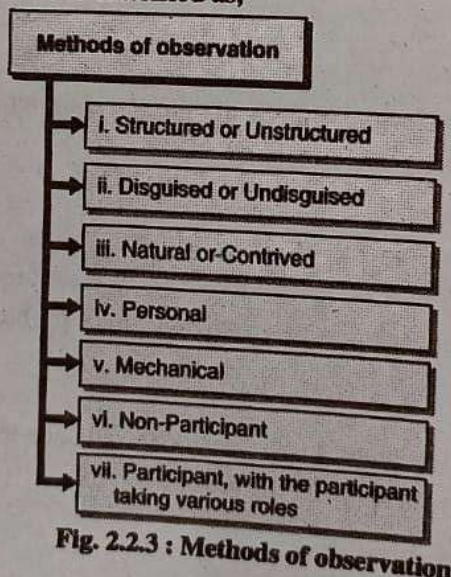

Fig. 2.2.3 : Methods of observation

i. Structured or Unstructured observations

In structured observation, the details of the observations are to be needed and various measurements metrics must be defined.

- In other words a proper problem definition is needed, whereas in unstructured observation the observer does not have any fixed problem definition, and hence he/she has to absorb the required details from the complete observation.

➜ **ii. Disguised or Undisguised observation**

In Disguised observation, the person who is being observed is unaware of it. Here observation is done secretly; whereas in undisguised observation the person is aware of that he/she is being observed.

➜ **iii. Natural or Contrived**

Natural observation takes place in natural environment and it is recorded and used as it is, whereas the contrived observation takes place in an artificially developed environment.

➜ **iv. Personal**

Personal Observation is similar to the natural observation.

➜ **v. Mechanical**

Mechanical Observations need various devices for recording and keeping track of the observations.

➜ **vi. Non-Participant**

It does not involve the direct or indirect involvement of the observer.

➜ **vii. Participant**

It completely depends on the involvement of the observer.

☛ **Case-Studies**

- Case studies are considered as historical approach where the measurement of what is there and how it got there.
- Case studies provide various details about the problem and also explain various scenarios. Many times in case study approach certain examples are present and hence it provides an easy way to come across certain conclusion.

☛ **Case study approach has following four steps**

1. Determine the present situation.
2. Gather background information about the past and key variables.
3. Test the hypotheses.
4. Take remedial action.

➜ **B. Secondary data collection methods**

Secondary data means that the data is available might be in some other format and it is needed to update or alter it according to the use. Secondary data are usually published in the papers or journals but there may be unpublished data also which are used by the private firms for their usage according to their needs.

Data which is published are available on :

1. Publications of government
2. Reports of various business
3. Public Records
4. Historical Documents etc.

Unpublished data can be kept somewhere in diaries, letters or on individual systems etc.

Syllabus Topic : Data Cleaning/Extraction

2.3 Data Cleaning/Extraction

Q. 2.3.1 Explain data cleaning. (Ref. Sec. 2.3) (5 Marks)

- Data cleaning is also known as data cleansing and it is process of detecting and correcting the errors or incorrect/invalid entries from the data set. Incorrect or inaccurate data can create inconsistency in the dataset, once the data is cleaned the inconsistency of the data is removed.
- Data cleaning process mainly remove typographical errors. Data cleaning process also helps for validation of the data this is done by comparing the data to be cleaned with the correct or related consistent data and then by removing unwanted data. Data should be cleaned in-order to get high quality data.
- Data cleaning also sometimes related to data purging process where the old, incomplete and duplicate data is removed. Main difference between data purging and data cleaning is data purging mainly aims to remove data for creating space for new data entries where as data cleaning is done to get the accurate data from the set of inconsistence data.
- Various tools can be used for cleaning data which includes JASP, Rattle, RapidMiner, Orange, Talend data preparation, Trifacta Wranger etc., Data cleaning includes tasks like Removing extra spaces, removing duplicates, correcting grammar, Spelling check.

Steps in data cleaning process

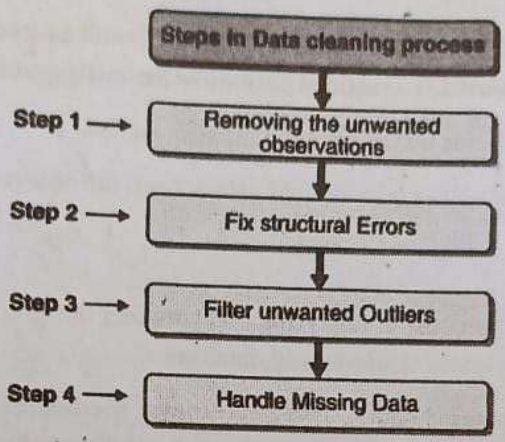

Fig. 2.3.1: Steps in data cleaning process

Step 1 : Removing the unwanted observations.

In this stage the unwanted observations are removed, unwanted observations are mainly divided into two types :

a. Duplicate observations
b. Irrelevant observations

a. Duplicate observations

Duplicate records arise in data collections stage here when the datasets are combined from multiple places there might stages when the same or similar type of the data may present.

b. Irrelevant observations

Irrelevant observations are those which do not fit in the specific problem.

Step 2 : Fix structural errors

Structural errors come during data measurement or data transfer. It can also be considered as typo errors.

Step 3 : Filter unwanted outliers

Outliers are the entries which do not belong to particular class. Outliers can lead to deviation in the output.