Data Analysis

First of all, try to understand that what is data analysis. Wikipedia says “Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making”. Okay but let’s understand in simpler words. We get the data from somewhere and it may be some dirty data and we first clean it and then we try to extract the meaningful things from the data by means of understanding it… The process of doing this is known as “Data Analysis”.



# 

# But why do we need the data analysis?

Yes, we need to analyze the to understand how things going… Let’s understand it with an easy example... Do you remember when our mom dads leave all the works on the last Sunday of the month and calculate how much they spent that month and then worried about them again failing to save more money this month too? So why do they calculate? Yes, they calculate to understand how and where the money goes… To fix the unwanted money flow they need to know all the money spend data and they also priorities their works accounts to the needs.

I believe do you have ever seen some graphs that are representing the sales of the company. So why that graph was created? Yes, you got that to understand the sales of the company over time. So we can see how well the company is performing so users can choose that company for their benefits …



Yes, I know there is a silly example let’s get into more real-life problems.

There was a middle size company and all their employees are working hard for the success of the company. But sales started decreasing quarter by quarter. That was very painful for Company CEO so the company CEO hired a Data Analyst to understand how his company sales is going doing day by day...

So analyst collected all the data of all company employee like check-in time, check-out time, No of leaves taken, what he searches in office pc, how many dependent in his family, what is the age of the employee and many other things…

After collecting the data he cleaned the data and then he start analyzing it… analyzing he created the report and submitted it to the CEO. When CEO reads the report he understood that why his company sales were going down…

The reports say that

* His company has most of employee are aged between 20 to 30
* Most of the employees have 1 or zero dependent
* A young age people are likely to play games in office ours
* Young age people take leaves a lot

After checking the report he understood that Young aged people care less about his/their job than the dependent ones so they don’t work properly play video games in the office and don’t much dedicate to their work…

The analyst also created a report that explains the effectiveness of a group of employe and then CEO remoted that group of employees and hires the employee that has more dedication towards the work… and surprisingly company growth was starting going up in few months…

So we understood that why data need to be analyzed…

# So what is the process to perform data analysis?

To perform the data analysis we need to understand the steps of data analysis...

So here are the 5 baby steps to perform Data analysis: -

1. Need of analysis
2. Collecting the data
3. Clean the Data
4. Analyze the data
5. Interpret and apply the result



# Need of analysis

First of all, we need to understand that why we need to do the analysis because understanding the problem is the first thing to find the solution to the problem. So we need to first understand that why we need the analysis. Once you got the idea of why you need the analysis you can prepare a roadmap for what needs to be done to achieve the solution.

Let’s see few examples of why we need the analysis

1. To find the effective employee in the company to increase the sales
2. To find the best price of the old car
3. To find the best path to reach the destination
4. To find the best mixture of concrete to build an earthquake-proof building

So understanding the need for analysis will push us to the right path that needs to be followed to get to the conclusion. Understanding the need for analysis we have to be a good communicator and we need to communicate with the team that needs a solution you also have to understand what sort of things are needed in the solution.

Good communication skills matter a lot to understand the problem.



I’m sharing few methods that need to be a good communicator

1. Learn to listen.

2. Pay attention to body language.

3. Observe how others communicate in different contexts.

4. Don't be afraid of a bit of silence.

5. Use action verbs and confident language.

6. Ask questions.

7. Find common ground, even in an argument.

8. Be prepared and know what you’re talking about.

9. If you want to subtly change the subject, find a verbal bridge.

10. Find the best way to frame your story.

11. Be relax

12. Practice

So that’s all about understanding the problem.

# Collecting the data

Yes, the next step is collecting the data But where to go the collect the data. Don’t worry while understanding the need for the analysis you also get an idea about where data comes from. It may come from any API or any website any article or from any database or from many places because data is everywhere… 

Basically, data can be divided into 2 types

1. Structured data
2. Un-structured data

Structured data are those data which is structured and well-formatted so we can use and consume the data very easily but on the other hand, unstructured data is very messy and needs a lot of time to extract the actual data and put into the structured form… because before doing anything we need the structured data that needs to be processed easily

We can also divide the data on the basis of form which could be

1. Audio
2. Video
3. Text
4. Image

The concerned team can also provide you with any database, API, or text data to process.

But the day could be

1. Internal Data
2. External Data

# 

Internal Data: This type of data is collected within the organization or third party organization to understand the problem This may include

1. CRM
2. ERP
3. Marking tool
4. Apps

External Data: External Data are those which are available on the internet and you may need to fetch the data if they already didn’t collect it...

This type of data includes

1. World Health data
2. Google trends
3. Twitter Tweets and trends
4. Reviews on shopping sites
5. Stock exchange data

After collecting the data you have to get into the cleaning of the Data

For collecting the data python provides tons of ways you can fetch the data from every possible way in python.

For API we can use python request

For web scraping, we can use Beautifulsuop or Selinum

For getting data from a database like MySQL use the MySQL package

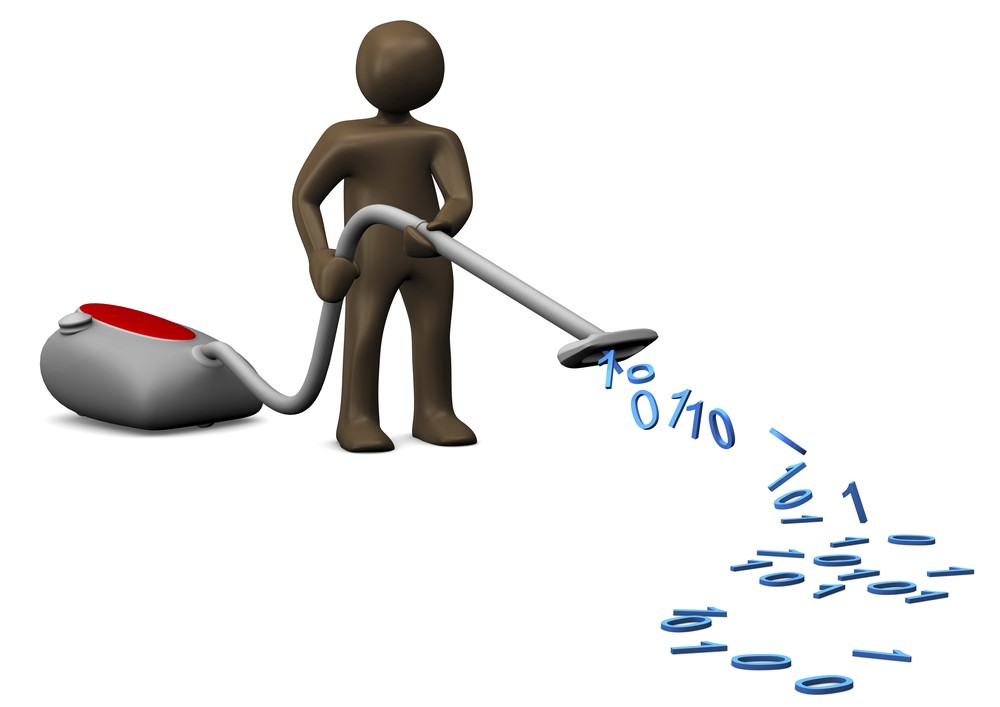
From getting data from formatted text file CSV use pandas

For getting unstructured data from the text file you have to use manual ways depending on the details.

# Clean the Data

Cleaning the data is one of the most challenging and time-consuming parts of the process which needs lots of time and effort to clean Data. According to ldodds.com, 80% of the time goes into cleaning the data.  
But why do you need the clean the data though? Because the data is collected by and from human and human makes mistakes and that mistake needs to be removed to get the insight from the data. Let’s understand by an example.

In your school, you got a job to collect the name, phone no, hobby, favorite sports, favorite actor name, and marks in the last exam from all the students of your school. You have collected all the in a sheet but at the end of the submission, you realized that you have missed some data like a hobby and favorite actor of some students or they refused to share with you. Also, you noticed you have entered more than 100 marks for some students which is technically impossible So that data needs to be handled to get the insight.



You must have to process

1. Null values
2. Duplicate values
3. Wrong data (Like the length of cars can’t 0 marks can’t be more than 100)

For cleaning the data in python if you are using pandas to store data then pandas have lots of functions to clean the data and restructure the data if required. You can also use the manual way or use other tools like OpenRefine and Drake

# Analyze the data

Analyzing the data is the process of finding the pattern in the data to extract meaningful information from the data. We also use visualization tools to analyze the data and to understand it and in the end, we generate Dashboard, Graphs, or scoreboards to represent the insight to the user. To analyze the data you can use

* Cluster analysis
* Anomaly deduction
* Associate rule

In clustering, we group similar data into different groups so it will be easy to understand the data for example HR group the employee in two or more parts depending on the dedication of his/her work…



There are lots of methods are there which can be used to analyze the data that may include

* Quantitative data analysis
* Qualitative data analysis
* Statistical analysis
* Textual analysis
* Descriptive analysis
* Predictive analysis
* Prescriptive analysis
* Diagnostic analysis

We also use algorithms and models to understand the data and to find the patterns in that.

For analyzing the data Python provides lots of function and library which is very flexible to use depending on requirement which includes

* Matplotlib
* Seaborn
* Plotly
* Geoplotlib
* ggplot

#### Matplotlib

It is the first python data visualization library and it’s the core of many libraries and that are built on top of matplotlib. It’s extensively used in python because it’s the oldest one. Pandas also used the matplotlib lib as the default backend.

Seaborn

#### Seaborn

Seaborn is built on the top of matplotlib and is closely integrated with a pandas data structure to prived easy implimentation with pandas The main difference between matplotlib and seaborn is seaborn creates beautiful charts in few lines of code Because seaborn uses default style and color palettes.

#### Plotly

Plotly is an online graph tool that also provides offline interactive graph but when you go online you have to pay some sorts of money. We can create advanced graphs in Plotly very easily

#### Geoplotlib

It is used for creating maps and plotting geographical data. Anyone can use this library to create a variety of maps like heatmaps don’t density map and choropleths graph easily It is used for maps because other libraries don’t support maps

#### ggplot

ggplot is the python version of the famous ggplot2 of R and Gramer of Graphics language. ggplot is a declarative style library which tightly coupled with pandas So better to use pandas when you want to go with ggplot

# Interpret and apply the result

This is the last step of the cycle which depends on the Organisation and their teams after an analyst submits his report, insight, and parts they find the data to the organization then the organization finally understand the problem and try to fix this by taking action, for example, an analyzed submitted the employee efficiency and dedication report based on their habits then the organization will thing about it what organization needs to do with the good or bad employee should they need to fire them or should organization warn them and again analyze the result or they should give bounces to the good employe that work harder and work to grow the organsiaton.



So here we are at the end of the article where we understood About the Daya analysis and its process and how we can perform it.