PYTHON FOR DATA ANALYSIS

[1]Deepa R, [2]Dr. Ravikumar G K, [3]Kavitha H M, [4]Divya B M

[1]Student, [2]Professor, [3]Research Scholar, [4]Assistant Professor
[1]Computer Science and Engineering,
[1]Adichunchanagiri University, B G Nagara, Mandya, Karnataka, India

Abstract : Increased volume of data and demand for analysis of data leads to the rise of one of the biggest problems of Big Data called Big Data analysis. In general, it refers to the process of collecting large and complex datasets which are tough to process using traditional database management tools or data processing applications. These are available in structured, semi-structured, and unstructured format in petabytes and beyond that needs to be analyzed. Considering this need, various data analytical tools and services have evolved as a means to solve this problem. Data analysis plays a vital role in decision making. Data is accessed and collected from different online and offline sources. Analyzing the data after collection and mining data from huge data repository and plotting the graph based on prediction on particular issue using the data collected with the help of libraries provided in python. Therefore trending area of research and development is big data analysis.

IndexTerms **– Data Analysis, Python, Python Libraries**

I. INTRODUCTION

There are various forms of data (i.e., audio, video, text, image, etc) both structured and unstructured which grows rapidly in day-to-day life. So analysis of this big data is a challenging task. Hence the data is collected from various forms. Once the data is collected the next important step is data analysis or statistical analysis. We use some statistical tools and statistical techniques in statistical data analysis. We use SPSS statistical software for outputs of statistical analysis for descriptive statistics, graphs and different tests. The variables given in the data set are summarised by descriptive statistics. We also use some inferential statistics to check some claims about the car data. We have to check the relationship between the given variables. Then do some graphical analysis for given variables. For checking our claims about the given data we have to do some inferential statistics or hypothesis testing. After statistical analysis, we will do some conclusions regarding the available data.

II.DATA ANALYSIS

Data Analysis is a step by step procedure of collecting, transforming, cleaning, and modelling data with the aim of discovering the required information and to discover needful information for business decision-making. While taking any decision in our day-to-day life is by thinking about what happened in past or what will happen in future by choosing that specific decision. In short, analyzing our past or future and make decisions based on it is Data analysis. The obtained results are communicated, suggesting conclusions, and supporting decision-making. Data visualization is used to represent the data to ease the discovering the useful patterns in the data. The terms Data Analysis and Data Modelling mean the same.

Why Data Analysis?

To bald the business even to grow in your life, sometimes all you need to do is Analysis. If there is no growth in business, then you have to look backward and attorn your mistakes and make a necessary plan again without repeating those previous mistakes. Eventhough the business is growing, then you have to look forward to make the business to grow more towards success. The main goal is to analyze your business data and business processes.

DATA ANALYSIS PROCESS

Data Analysis Process consists of the following phases that are iterative in nature −

- Data Requirements Specification
- Data Collection
- Data Processing
- Data Cleaning
- Data Analysis
- Communication

Data Requirements Specification

Initially we have to think about why it is important to do this data analysis? Most importantly we need to find out the aim of doing the Analysis, which type of data analysis is to done. In this stage, we will decree what to analyze and how it is measured, we need to interpret why we are investigating and what precautions are to be taken to do this Analysis. Based on the requirements and specifications of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Variants of cars). Specific variables regarding cars (e.g., make and model) may be specified and obtained. Data may be categorical or numerical.

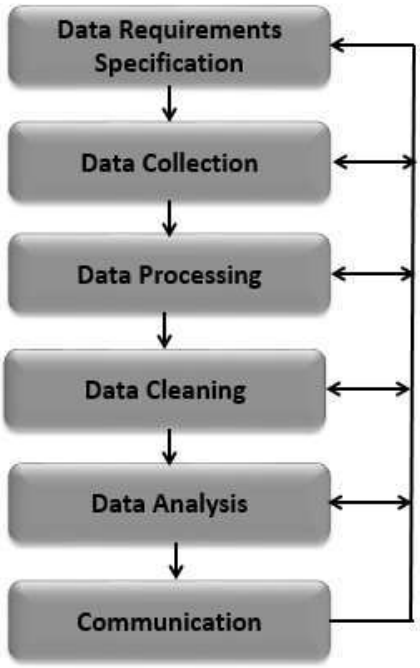

Figure a: Phases of Data Analysis Process

Data Collection

Data Collection is the process of gathering relevant information on specific variables identified as data requirements. The focus is on securing accurate and honest collection of data. The decisions are validated based on the data gathered by Data Collection .It gives both a target to improve and baseline to measure. Data is gathered from various sources ranging from organizational databases to the information in web pages. The data thus obtained, and it may be unstructured and may contain irrelevant information. Hence, the collected data is required to be processed and cleaned.

Data Processing

Before analysis, the data collected must be processed or organized. This includes the task of structuring the data as required for the relevant Analysis Tools. For example, the data is stored in a table in the form of rows and columns within a Spreadsheet or Statistical Application. A Data Model might have to be created.

Data Cleaning

Whatever data is collected now may not be useful or irrelevant to your aim of Analysis, hence it should be cleaned. There might be duplicate data in collected records of data, white spaces or errors. Preventing and correcting these errors is referred as Data Cleaning. There are different types of Data Cleaning depending on the type of data. The data should be clean and error free. This phase must be done before Analysis because the expected output of the analysis is achieved based on data cleaning.

Data Analysis

After data is collected, cleaned, and processed, it is ready for Analysis. Since we manipulate data, we may find that we have the exact information needed, or we might need to gather more data. We can use data analysis tools and software which will helps to understand, interpret, and derive conclusions based on the requirements during this phase. Based on the requirements, there are various data analysis techniques available to understand, interpret, and derive conclusions. To examine the data in graphical format, to obtain added insight regarding the messages within the data, Data Visualization may also be used. To identify the relations among the data variables, Statistical Data Models such as Correlation, Regression Analysis can be used. To simplify analysis and communicate results these models that are descriptive of the data are helpful. The process is iterative since it might require additional Data Cleaning or additional Data Collection.

Communication

Once the data is analysed, it's finally time to moralise your results. To support the decisions and further action of the users, the results of the data analysis are to be reported in a format as required by the users. Data visualization techniques, such as tables and charts can be adopted by data analysts, which help in delivering the message clearly and efficiently to the users. There might be additional analysis based on the feedback from the users.

DATA ANALYSIS TOOLS

Using Data analysis tools, it easier for data anlysts to process and manipulate data, analyze the relationships and correlations between data sets, and it also helps to identify patterns and trends for interpretation. Here is a complete list of tools.

Figure b: Data Analysis Tools

III. PYTHON

Python is an interpreted, high-level, general-purpose programming language created and released in 1991 by Guido van Rossum and initially it's design philosophy emphasizes code readability with its notable use of significant whitespace Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

4.2 Why Python for Data Analysis?

For many people, the Python language is easy to adopt. Since1991, along with Perl, Ruby, and other programming languages Python has become one of the most popular dynamic programming language. For building websites, Python and Ruby have become especially popular in recent years using their numerous web frameworks, like Rails (Ruby) and Django (Python). Such languages can be used to write quick-and-dirty small programs, or scripts, they are often referred as scripting languages. They cannot be used for building mission-critical software. Among interpreted languages Python is distinguished by its large and active scientific computing community. Since the early 2000s, espousal of Python for scientific computing in both academic research and industry applications has increased significantly.

Python will inevitably draw comparisons with the many other domain-specific open source and commercial programming languages for data analysis and interactive, exploratory computing and data visualization and tools such as R, MATLAB, SAS, Stata, and others are widely used. Recently Python's improved library support (primarily pandas) has made a robust alternative for data manipulation tasks. Python's strength is an excellent choice as a single language for building data-centric applications as rolled into one with its strength in general purpose programming.

4.3 Essential Python Libraries

Following are the overview of each library of python for those who are less familiar with the scientific Python ecosystem and the libraries used.

4.3.1 NumPy

NumPy, short form of Numerical Python, is the initial basic package for scientific computing in Python. For storing and manipulating data for numerical data other than built-in Python data structures, NumPy arrays are a much more efficient. The data stored in a NumPy array can be operated without duplicating any data and also libraries written in a lower-level language, such as C or FORTRAN.

4.3.2 Pandas

Pandas enable rich data structures and functions designed to work with structured data fast, easy, and expressive. It is one of the critical ingredients providing Python to be a powerful and productive data analysis environment. Data Frame, a two dimensional tabular, column oriented data structure with both row and column labels are the primary object in pandas that will be used. Pandas feature rich, high-performance time series functionality and tools well-suited for working with financial data. Pandas are initially designed as an ideal tool for financial data analysis applications.

4.3.3 Matplotlib

For Python programming language Matplotlib is a plotting library and its numerical mathematics extension is NumPy. It is the most accepted Python library for producing plots and other 2D data visualizations. For publication, it is suited well for creating plots. It provides a comfortable interactive environment for plotting and exploring data since it integrates well with IPython. The plots are interactive, the section of plot can be zoomed in and out, and pan around the plot using the toolbar in the plot window. Using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+, It afford an object-oriented API for engrafting plots into applications.

4.3.4IPython

IPython is the component in the standard scientific Python toolset that binds everything together. It gives a robust and productive environment for interactive and exploratory computing. It is an upgraded Python shell designed to speed up the writing, testing, and debugging of Python code. It is specifically effectual for interactively working with data and visualizing data with matplotlib. IPython is usually pertained with the bulk of Python work, including running, debugging, and testing code

4.3.5 SciPy

The collection of packages addressing a numerous standard problem domains in scientific computing is SciPy. Combination of NumPy and SciPy form a reasonably complete computational replacement for much of MATLAB along with few of its add-on toolboxes.

IV. CONCLUSION

Because of increase in the amount of data in current environment, it becomes difficult to handle the data, and analysis of the data sets. Although there are many sources of data that are currently fueling the rapid growth in data volume, Massive data analysis creates new challenges at the interface between humans and computers. Here we collected the data through the Google forms , mined the data collected and given it as input , after some operations on data sets using particular libraries in python, we visualize the output by plotting the graph based on given input.

V. ACKNOWLEDGMENT

I would like to express my sincere gratitude towards my guide Dr. RaviKumar G K, Professor, Head R & D, Dept. of CSE, B.G.S Institute of Technology, B.G Nagar, for the help, guidance and advice in development of this methodology. I would like to express my sincere gratitude towards Mrs. Divya B M, Asst. Professor, BGSIT, for the support and guidance.

References

1 **Spatial Panel Data Analysis** J.Paul Elhorst Faculty of Economics and Business, University of Groningen, the Netherlands Elhorst, J.P. (2017) Spatial Panel Data Analysis. In: Shekhar S., Xiong H., Zhou X. (Eds.) Encyclopedia of GIS, 2nd edition, pp. 2050-2058. Springer International Publishing, Cham, Switzerland.

2 **Compositional data analysis of household waste recycling centres in Denmark** Edjabou, Maklawe Essonanawe; Martín-Fernández, J. A.; Boldrin, Alessio; Astrup, Thomas Fruergaard

3 **Python for Data Analysis**
Data Wrangling with Pandas, NumPy, and IPython By William McKinney