EXPERIMENT - 01

AIM: To study various data analysis and data collection tools.

THEORY:

Types of Data Analysis Tools

- **R**

R is the leading analytics tool in the industry and widely used for statistics and data modeling. It can easily manipulate your data and present in different ways. It has exceeded SAS in many ways like capacity of data, performance and outcome. R compiles and runs on a wide variety of platforms viz -UNIX, Windows and MacOS. It has 11,556 packages and allows you to browse the packages by categories. R also provides tools to automatically install all packages as per user requirement, which can also be well assembled with Big data.

Data analysis with R is done in a series of steps; programming, transforming, discovering, modelling and communicate the results.

- **PYTHON**

Python is an object-oriented scripting language which is easy to read, write, maintain and is a free open source tool. It was developed by Guido van Rossum in late 1980's which supports both functional and structured programming methods.
Python is easy to learn as it is very similar to JavaScript, Ruby, and PHP. Also, Python has very good machine learning libraries viz. Scikitlearn, Theano, Tensorflow and Keras. Another important feature of Python is that it can be assembled on any platform like SQL server, a MongoDB database or JSON. Python can also handle text data very well.

Both Python and R are developing their features and functionalities to ease the process of Data Analysis with high speed and accuracy.

- **SAS**

Sas is a programming environment and language for data manipulation and a leader in analytics, developed by the SAS Institute in 1966 and further developed in 1980's and 1990's. SAS is easily accessible, managable and can analyze data from any sources. SAS introduced a large set of products in 2011 for customer intelligence and numerous SAS modules for web, social media and marketing analytics that is widely used for profiling customers and prospects. It can also predict their behaviors, manage, and optimize communications.

- **TABLEAU**

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data in a very easily understandable format. Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards. Data analysis is very fast with Tableau tool and the visualizations created are in the form of dashboards and worksheets.

For a clear understanding, data analytics in Tableau tool can be classified into two sections.

1. **Developer Tools:** The Tableau tools that are used for development such as the creation of dashboards, charts, report generation, visualization fall into this category. The Tableau products, under this category, are the Tableau Desktop and the Tableau Public.
2. **Sharing Tools**: As the name suggests, the purpose of these Tableau products is sharing the visualizations, reports, dashboards that were created using the developer tools. Products that fall into this category are Tableau Online, Server, and Reader.

Both Excel and Tableau are data analysis tools, but each tool has its unique approach to data exploration. However, the analysis in Tableau is more potent than excel. Excel works with rows and columns in spreadsheets whereas Tableau enables in exploring excel data using its drag and drop feature.

- **APACHE SPARK**

Apache Spark is one of the most successful projects in the Apache Software Foundation and is a cluster computing framework that is **open-source and is used for real-time processing.** Being the most active Apache project at the moment, it comes with a fantastic **open-source community** and an **interface for programming**. This interface makes sure of fault tolerance and implicit data parallelism. Apache Spark keeps on releasing new releases with new features. You can also choose the various package types for Spark.

Companies such as Oracle, Hortonworks, Verizon, Visa use Apache Spark for real-time computation of data with ease of use and speed.

Features: -

1. In today's world Spark runs on Kubernetes, Apache Mesos, standalone, Hadoop, or in the cloud.
2. It provides high-level APIs in Java, Scala, Python, and R, and Spark code can be written in any of these four languages.
3. Spark's MLlib – the Machine Learning component is handy when it comes to Big Data processing.

Types of Data Collection Tools

Data collection tools refer to the devices/instruments used to collect data, such as a paper questionnaire or computer-assisted interviewing system. Case Studies, Checklists, Interviews, Observation sometimes, and Surveys or Questionnaires are all tools used to collect data.

- **INTERVIEW**

An interview is a face-to-face conversation between two individuals with the sole purpose of collecting relevant information to satisfy a research purpose. Interviews are of different types namely; Structured, Semi-structured, and unstructured with each having a slight variation from the other.

Pros

- In-depth information
- Freedom of flexibility
- Accurate data.

Cons

- Time-consuming
- Expensive to collect.

- **QUESTIONNAIRES**

This is the process of collecting data through an instrument consisting of a series of questions and prompts to receive a response from individuals it is administered to. Questionnaires are designed to collect data from a group.

For clarity, it is important to note that a questionnaire isn't a survey, rather it forms a part of it. A survey is a process of data gathering involving a variety of data collection methods, including a questionnaire.

On a questionnaire, there are three kinds of questions used. They are; fixed-alternative, scale, and open-ended. With each of the questions tailored to the nature and scope of the research.

Pros

- Can be administered in large numbers and is cost-effective.
- It can be used to compare and contrast previous research to measure change.
- Easy to visualize and analyze.
- Questionnaires offer actionable data.
- Respondent identity is protected.
- Questionnaires can cover all areas of a topic.
- Relatively inexpensive.

Cons

- Answers may be dishonest or the respondents lose interest midway.
- Questionnaires can't produce qualitative data.
- Questions might be left unanswered.
- Respondents may have a hidden agenda.

- Not all questions can be analyzed easily.

• OBSERVATION

This is a data collection method by which information on a phenomenon is gathered through observation. The nature of the observation could be accomplished either as a complete observer, an observer as a participant, a participant as an observer, or as a complete participant. This method is a key base for formulating a hypothesis.

Pros

- Easy to administer.
- There subsists a greater accuracy with results.
- It is a universally accepted practice.
- It diffuses the situation of an unwillingness of respondents to administer a report.
- It is appropriate for certain situations.

Cons

- Some phenomena aren't open to observation.
- It cannot be relied upon.
- Bias may arise.
- It is expensive to administer.
- Its validity cannot be predicted accurately.

• DOCUMENTS AND RECORDS

This method involves extracting and analysing data from existing documents. The documents can be internal to an organization (such as emails, sales reports, records of customer feedback, activity logs, purchase orders, etc.) or can be external (such as Government reports).

Pros

- Ease of data collection
- No need for searching and motivating respondents to participate.
- Allows you to track progress. Helps you understand the history behind an event and track changes over a period of time.

Cons

- Easy to administer.
- Information may be out of date or inapplicable.
- The process of evaluating documents and records can be time-consuming.
- Can be an incomplete data collection method because the researcher has less control over the results.
- Some documents may be not publicly available.

- **FOCUS GROUPS**

The opposite of quantitative research which involves numerical-based data, this data collection method focuses more on qualitative research. It falls under the primary category for data based on the feelings and opinions of the respondents. This research involves asking open-ended questions to a group of individuals usually ranging from 6-10 people, to provide feedback.

Pros

- Information obtained is usually very detailed.
- Cost-effective when compared to one-on-one interviews.
- It reflects speed and efficiency in the supply of results.

Cons

- Lacking depth in covering the nitty-gritty of a subject matter.
- Bias might still be evident.
- Requires interviewer training
- The researcher has very little control over the outcome.
- A few vocal voices can drown out the rest.
- Difficulty in assembling an all-inclusive group.

RESULT:

We became aware of various data analysis and data collection tools.