# DATA ANALYTICS & VISUALISATION EXPERIMENT 1

**Aim:** Getting introduced to data analytics libraries in Python and R.

**Theory:**

Data visualisation is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide accessible way to see and understand trends, outliers and patterns in any data. Additionally it provides an excellent way for employees or businesses to present data to non-technical audiences without confusion.

In the world of big data, data viz. tools and technologies are essentials of analyze massive amounts of information and make data-driven decisions.

**Advantages:**

1) Better agreement
2) Superior method
3) Simple sharing of data.
4) Precise investigation.
5) Adjustment of information
6) Geological perception.

Disadvantages:

1) Gives assesment not. exactness.
2) Absence of help
3) Inappropriate plan issue
4) One-sided results.

LIBRARIES.

1) Matplotlib:

Matplotlib is a Python vizualisation library for 2D arrays plots. It is a python library that uses the Numpy library. It includes a wide range of plots, such as scatter line, bar, histogram and others that can assist us in delving deeper into trends, behavioral patterns and correlations.

2) Seaborn:

Seaborn is a python library for creating statistical representations based on datasets. Built on top of matplotlib and is used to create various visualizations.

Built on top of pandas data structures. This library conducts the necessary modelling and aggregation internally to create insightful visuals.

It reduces the lines of codes required to produce the same result as in matplotlib.

3) Plotly:

Plotly.py is a python library viz. that is interactive, accessible, high level, declarative, and browser-based. Scientific graphs, 3D charts, statistical plots and financial charts are just a few visualizations available.

4) Pandas:

Pandas is another great library available in Python for data analysis (data manipulation, time series, integrating indexing of data, etc).

Pandas visualization is built on top of matplotlib. is a tool of Pandas library that allows us to create a visual representation of data frames. (data aligned in tabular form of rows and columns) and

series (one dimensional labelled array capable of holding data of any type) much quicker and easier way.

5) Bokeh :

Bokeh is a web based interactive modern visualisation library. It can create engaging plots and dynamic dashboards with huge or streaming data assets. The library contains many intuitive graphs that can be used to create solutions. Ideal for creating customized visuals based on specific user use cases. Visual effects can be crafted interactive to support a what if scenario model.

CONCLUSIONS -

DATA VISUALISATIONS are important for almost every career, and these libraries. (matplotlib, pandas, bokeh, seaborn, plotly) allow python to be one of the most prominent language for data visualization.

9/3/23

# DATA ANALYTICS & VISUALISATION

## EXPERIMENT 2.

**Aim:**

To implement simple linear regression on any dataset.

**Theory:**

### Linear Regression:

It is used to predict the value of a variable based on the value of another variable.

The variable you want to predict ~~to~~ is called the dependent variable & the variable you are using to predict the other variable's value which is called the 'independent variable.

### Simple Linear Regression:

Used to estimate the relationship between the two quantitative variables.

We use Simple Linear Regression, when we want to know:

1) How strong the relationship is between two variables (Eg. rainfall & soil erosion)
2) value of dependent variable at a certain

value of independent variable. (Eg. as above).

## Assumptions of Simple Linear Regression:

1) Homogenity of variance.
   Size of error in our prediction doesn't change significantly across the values of the independent variable.

2) Independence of observations.
   observations in the dataset were collected using statistically valid sampling methods.

3) Normality
   Data follows a normal distribution.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Conclusion:

Hence, we have learned & successfully implemented simple linear regression.

* * * * *

# DATA ANALYTICS & VISUALISATION

## EXPERIMENT 3

**Aim:** To implement multiple linear regression on any dataset.

**Theory:**

### Regression Models:

Regression models are used to describe relationships between variables by fitting line to the observed data.

Regression allows you to estimate how a dependent variable changes as independent variables change.

### Multiple linear regression

Used to estimate the relationship between 2 or more independent variables and one dependent variable.

We can use multiple linear regression when we want to know:

1. How strong is the relationship between two or more independent variables and one dependent variable.

Eg. rainfall & temperature affecting crop growth.

⑤ The value of dependent variable at certain values of the independent variables.

## Assumptions of Multiple linear regression:

1) Homogeneity of variance.
Size of the error in our prediction doesn't change significantly across the values of independent variable.

2) Independence of observations.
The observations in the dataset were collected using statistically valid sampling methods.

3) Normality
The data follows a normal distribution.

4) Linearity.
The line of best fit through the data points in a straight line.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

Conclusion: Thus, we have successfully learned, studied & implemented multiple linear regression.

9/3/23

# DATA ANALYTICS AND VISUALIZATIONS

## EXPERIMENT 4

**Aim:** To implement the time series ~~forecasting~~ analysing ~~algorithm~~ on a dataset using Python.

**Theory:**

1) A time series data is a series of data points or observations recorded at a different or a regular time ~~*~~ intervals. The frequency of recorded data points may be hourly, daily, weekly, monthly, quarterly or annually.

2) Time series forecasting is the process of using statistical model to predict future values of the time series based on past results.

3) Applications of time series forecasting, is used in the process of statistics, finance, business applications, etc.

4) Different components of time series are:

(i) Trend - trend shows a general direction of the time series data over a long period of time. A trend can be upward, downward or stationary.

(ii) Seasonality - exhibits a trend that repeats with respect to timing, direction & magnitude.

(iii) Stationarity - shows the mean value of the series that remains constant over the time period.

- ETS Decomposition
  used to seperate different components of a time series. ETS stands for Error, Trend and Seasonality.
- Cyclic oscillations in time series which last for more than a year are called cyclic.
- Irregularity - Fluctuations in time series data which does not correspond to trend or seasonability.

Conclusion -

Time series ~~forecasting~~ analysing has been studied and implemented on time series data.

9/3/22

CLASSMATE
Date :
Page :

DAV

# Experiment 5

AIM:

Implementation of ARIMA model.

THEORY:

ARIMA (AUTO REGRESSIVE INTEGRATED MOVING AVERAGE) is a statistical model used for time series analysis, which is analyzing data collected over time.

In simple terms, ARIMA models try to find patterns and relationships within a time series data by considering 3 important factors.

## AUTO-REGRESSION

This means that the model looks for correlation between current value of time series and its past value. It assures that the current value depends on its past value.

## Integration.

This means that the model tries to make the time series stationery, which is a situation

statistical term that means the data has a constant mean & variance overtime.

# Moving Average.

This means that the model looks for correlation between the current value of the time series and the past forecast errors.

By combining these 3 factors, ARIMA models can make forecasts of future values of a time series based on its past values and patterns.

There are 3 types of ARIMA model:

1. ARIMA (p, d, q)
2. ARMA (p, q)
3. AR (p)

## Conclusion

There are multiple models out of which we can choose the optimum one according to our requirements.

DAV

# Experiment 6

Aim:

Implementing Sentiment Analysis as a NLP technique.

Theory:

Sentiment Analysis is a natural language processing technique that involves the classification of texts into different categories based on their emotional tone such as positive, negative or neutral.

In python, there are several libraries and tools available for sentiment analysis, including NLTK, text BLOB and VADER.

```
# VADER
```

Valence Aware Dictionary and Estimation Reasoner is a rule based sentiment analysis tool that is specifically designed to analyse the sentiment of text data especially social media.

Vader works by analyzing the text and assigning sentiment scores to each word based on a predefined dictionary of sentiment scores.

The sentiment score for each word is based on its positive or negative valence as well as the ~~insen~~ intensity of the sentiment conveyed by the word.

The score for the entire text is then categorized & calculated by aggregating the scores of all the words in the text.

Vader provides sentiment scores in categories:

1. positive
2. negative
3. neutral
4. compound.

CONCLUSION

There are a number of pre-existing modules to implement sentiment analysis on any text data

CLASSMATE
Date :
Page :

# DAV Experiment 7

Aim: Data Visualisation in Python.

Theory:

1) Data visualisation is the graphical representation of data to communicate information clearly & efficiently to users.
2) In python, there are several libraries available to create data visualisation, including matplotlib, seaborn, plotly, ggplot, etc.
3) These libraries provide a range of visualisation types.

## GANTT CHART

A gantt chart is used for showing activities (tasks or events) displayed against time.

Each activity is represented by a bar, position & length of bar reflects the start & end.

## RADAR CHART.

A radar chart shows multivariate data of 3 or more quantitative variable.

It looks like a spiders web, hence is also called as spider chart.

## WORD CLOUD.

A word cloud is a visual representation of text down, which is often used to depict keyword. metadata on website, or to visualise free from test.

## NIGHTINGOLE CHART

A nightingole chart is a combition of radar chart and stacked column chart type of data visualisation.

It is useful for exploring statistical data.

## MEKKO PLOT

A marimekko chart is a graphical representation that uses statistical bar graph to visualise categorical data.

## CONCLUSION:

Data visualisations such as word cloud, nightingole chart have been implemented in python.

DAV Experiment 8

Aim: Visualizing Data in R.

Theory:

R is a popular programming language for Data visualization offering a wide range for building & customizing functions and packages to create a variety of charts and plots.

These include box charts, line charts, scatter plots, heatmaps and more using data from different sources. Such as CSV files, SQL databases and API's.

- WORD TREE

type of chart that displays the frequency of words or phrases in a hierarchal tree structure where each branch represents a subsequent word in the phrase.

- WORD CLOUD

A word cloud is a visual representation of text data where the size of each word in the cloud corresponds to its frequency of importance.

- Sun burst diagram

A sun burst diagram is a type of hierarchal chart that display data as a set of nested rings, with each ring representing a category.

- PIE CHART.

Circular chart that displays how a single variable is divided into parts with each part represented by a slice of the circle.

CONCLUSION

We have successfully learned, implemented and used R for different data visualization techniques.

* * * *