MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Minor Project Report

on

Using Big Data Analytics to Design a Smart Market Basket- V Mart

A project report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

in

ENGINEERING MATHEMATICS & COMPUTING

Submitted by:

Akshay Kumar Koshta (0901MC201004)

Anshika Gupta (0901MC201012)

Deeksha Pathak (0901MC201018)

Dev Agrawal (0901MC201019)

Divyanshi Singh Parmar (0901MC201021)

Faculty Mentor:
Santosh Kumar Bharadwaj, Assistant Professor

Submitted to:

Department of Engineering Mathematics & Computing

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR - 474005

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Certificate

This is certified that **Akshay Kumar Koshta (0901MC201004), Anshika Gupta (0901MC201012), Deeksha Pathak (0901MC201018), Dev Agrawal (0901MC201019), Divyanshi Singh Parmar (0901MC201021)** have submitted the project report titled **BIG DATA** under the mentorship of **Dr. Vikas Shinde, Dr. Santosh Kumar Bharadwaj** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Engineering Mathematics & Computing** from Madhav Institute of Technology and Science, Gwalior.

Dr. Santosh K. Bharadwaj
Assistant Professor
Engineering Mathematics & Computing

Dr. Vikas Shinde
Professor and Head,
Department of
MAC

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Declaration

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Engineering Mathematics & Computing at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Dr. Santosh Kumar Bharadwaj**, Assistant Professor Mathematics and Computing**.**

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Date:

Place: Gwalior

Akshay Kumar Koshta (0901MC201004)

Anshika Gupta (0901MC201012)

Deeksha Pathak (0901MC201018)

Dev Agrawal (0901MC201019)

Divyanshi Singh Parmar (0901MC201021)

BTech III Year

Engineering Mathematics & Computing

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Acknowledgement

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Engineering Mathematics & Computing,** for allowing me to explore this project. I humbly thank **Dr. Vikas Shinde** , Professor and Head, Department of Engineering Mathematics & Computing , for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Santosh Kumar Bharadwaj**, Assistant Professor, Mathematics and Computing**,** for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Akshay Kumar Koshta (0901MC201004)

Anshika Gupta (0901MC201012)

Deeksha Pathak (0901MC201018)

Dev Agrawal (0901MC201019)

Divyanshi Singh Parmar (0901MC201021)

BTech III Year,
Engineering Mathematics & Computing

Table of Contents

Abstract

As the name implies, Big Data literally means large collections of data sets containing abundant information. Big data has the potential to revolutionize the art of management to take appropriate decision on time. Extremely large data sets that may be analysed computationally to reveal patterns, trends, and association from unstructured data into structured ones to find a solution for a business is the key factor in today's market. Despite the high operational and strategic impacts, there is a scarcity of empirical research to assess the business value of big data. Big Data Analytics is increasingly becoming a trending practice that many organizations are adopting with the purpose of constructing valuable information from Big Data. This paper provides in depth analysis of Big data its challenges and its future scope where it is leading too and Big Data Analytics methods used by different organizations that helps their business to make a strong investment decision. Paper also covers different big data tools used with its salient features. Future research directions in this field are wide opened; but this paper has tried to facilitate the exploration of the domain and the development of optimal techniques to address Big Data.

List of Attributes

ATTRIBUTE	DESCRIPTION
1. Store_id	Different stores (amazon, big basket, flipkart) id's are mentioned
2. Item_id	Product unique id, having different store id's
3. Item_visibility	The percentage of total display area of all products in a store allocated to the particular product.
4. Item_type	The categories to which the product belongs.
5. Item_name	Name of the Product.
6. Item_MRP	Price of the product per unit.
7. Discount	Percentage of discount on the product of a particular outlet (& store).
8. Item_outlet_sales	Number of units of the product sold in the outlet.
9. Final_MRP	(Item_MRP - Discount) i.e., the final price after offering discount.
10. Percentage_searched	Searching rate of a particuar product on the store site.
11. Rating	Customer ratings for the product out of 5 in an outlet.
12. Off price	Item_MRP*(100-DISCOUNT)....whole divide by 100

Vision

The main purpose to create this project is to analyze the big data, and to visualize it using certain tools. The ideology behind this project is to collect all the raw data and to transform it into some useful information.

This project is all about a platform where we provide recommendations to the customers, to help them select the best quality products, out of all the available ones. It also contains a feature of feedback system to ensure customer’s satisfaction.

This study aims to shoring up the customer relationship by designing and analyzing an intelligent system. It utilizes big data analytics to study the market basket analysis for one of the top retail series called "V Mart".

To accomplish this, **power bi** tool will be used to convert the raw data into a readable and visualized format

Big Data

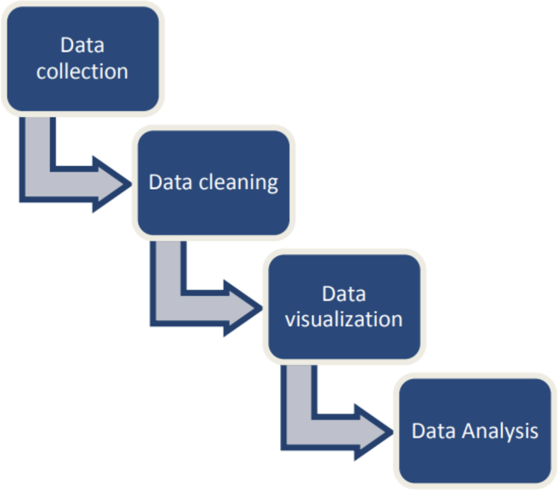
Introduction to Big Data

According to Gartner, the definition of Big Data – "Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." This definition clearly answers the "What is Big Data?" question – Big Data refers to complex and large data sets that have to be processed and analysed to uncover valuable information that can benefit businesses and organizations. However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analysed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyse the data.

Big Data Analytics

This section introduces the big data analytics mechanism used to deal with the market basket analysis problem. The data analytic mechanism used in this study involves the following four main stages



Data collection

The invoices contain the items that has be bought by a certain customer. As depicted in Table 1, the dataset contains a unique ID for each invoice (Invoice No); each invoice includes a basket that contains certain items (Stock Code). In addition, the dataset contains a unique ID for each customer (Customer) -Only the customers who are VIP are taken into consideration since they always buy from V Mart -The Customer identification number is his/her unique phone number; it is saved in the ERP system. But for the confidentiality purposes, the ID is coded

Data cleaning

The numbers are errors in data entry or they are some bugs in the system; thus, they need to be eliminated from the dataset. In addition, some data points are not available Nan; those are also eliminated. -If an invoice includes an item that is (are) repeated more than once, only just one record will be kept. In this study, the similarity or correlation between items is considered regardless of their quantity.

Data visualization

Data visualization is a critical step in the data science process, helping teams and individuals convey data more effectively to colleagues and decision makers. Teams that manage reporting systems typically leverage defined template views to monitor performance. However, data visualization isn't limited to performance dashboards. For example, while text mining an analyst may use a word cloud to to capture key concepts, trends, and hidden relationships within this unstructured data. Alternatively, they may utilize a graph structure to illustrate relationships between entities in a knowledge graph. There are a number of ways to represent different types of data, and it's important to remember that it is a skillset that should extend beyond your core analytics team. Dashboards include common visualization techniques, such as:

Tables: This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way, but they can also overwhelm users that are simply looking for high-level trends.

Pie charts and stacked bar charts: These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.

Line charts and area charts: These visuals show change in one or more quantities by plotting a series of data points over time and are frequently used within predictive analytics.

Histograms: This graph plots a distribution of numbers using a bar chart (with no spaces between the bars), representing the quantity of data that falls within a particular range. This visual makes it easy for an end user to identify outliers within a given dataset.

Scatter plots: These visuals are beneficial in reveling the relationship between two variables, and they are commonly used within regression data analysis.

Heat maps: These graphical representation displays are helpful in visualizing behavioral data by location. This can be a location on a map, or even a webpage.

Tree maps, which display hierarchical data as a set of nested shapes, typically rectangles. Tree maps are great for comparing the proportions between categories via their area size.

Data Analysis

Data analysis is the process of examining, cleansing, transforming, and modelling data with the objective of extracting useful information for decision-making. It is often used in different domains, such as business, science, and the humanities.

The History of Big Data

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centres and the development of the relational database. Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyse big data sets) was developed that same year. NoSQL also began to gain popularity during this time. The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it. With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data. While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

Benefits of Big Data and Data Analytics

- Big data makes it possible for you to gain more complete answers because you have more information.

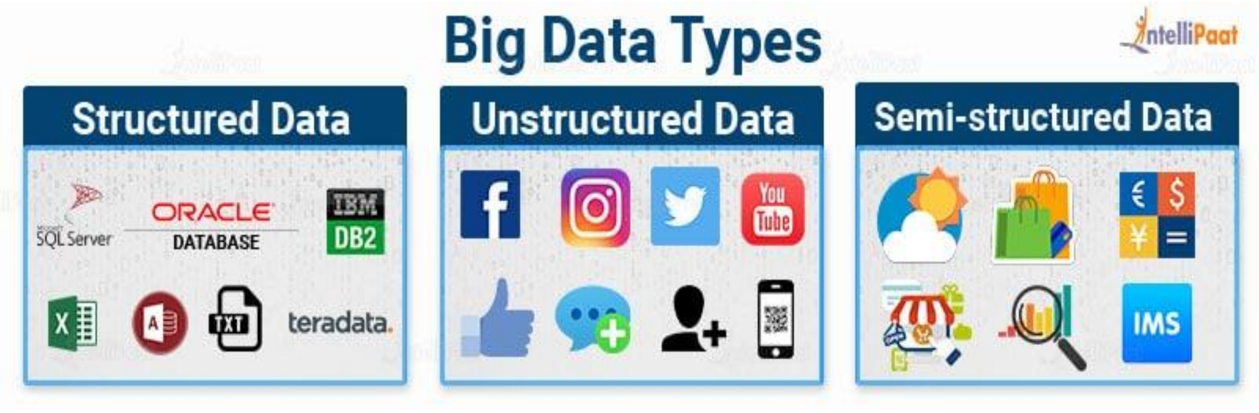
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems. Types of Big

Data Now that we are on track with what is big data, let’s have a look at the types of big data:

a) **Structured**: Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

b) **Unstructured**: Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyse unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

c) **Semi-structured**: Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus, we come to the end of types of data.



Characteristics of Big Data

Characteristics of Big Data Back in 2001, Gartner analyst Doug Laney listed the 3 V's of Big Data – Variety, Velocity, and Volume. Let's discuss the characteristics of big data. These characteristics, isolated, are enough to know what big data is. Let's look at them in depth:

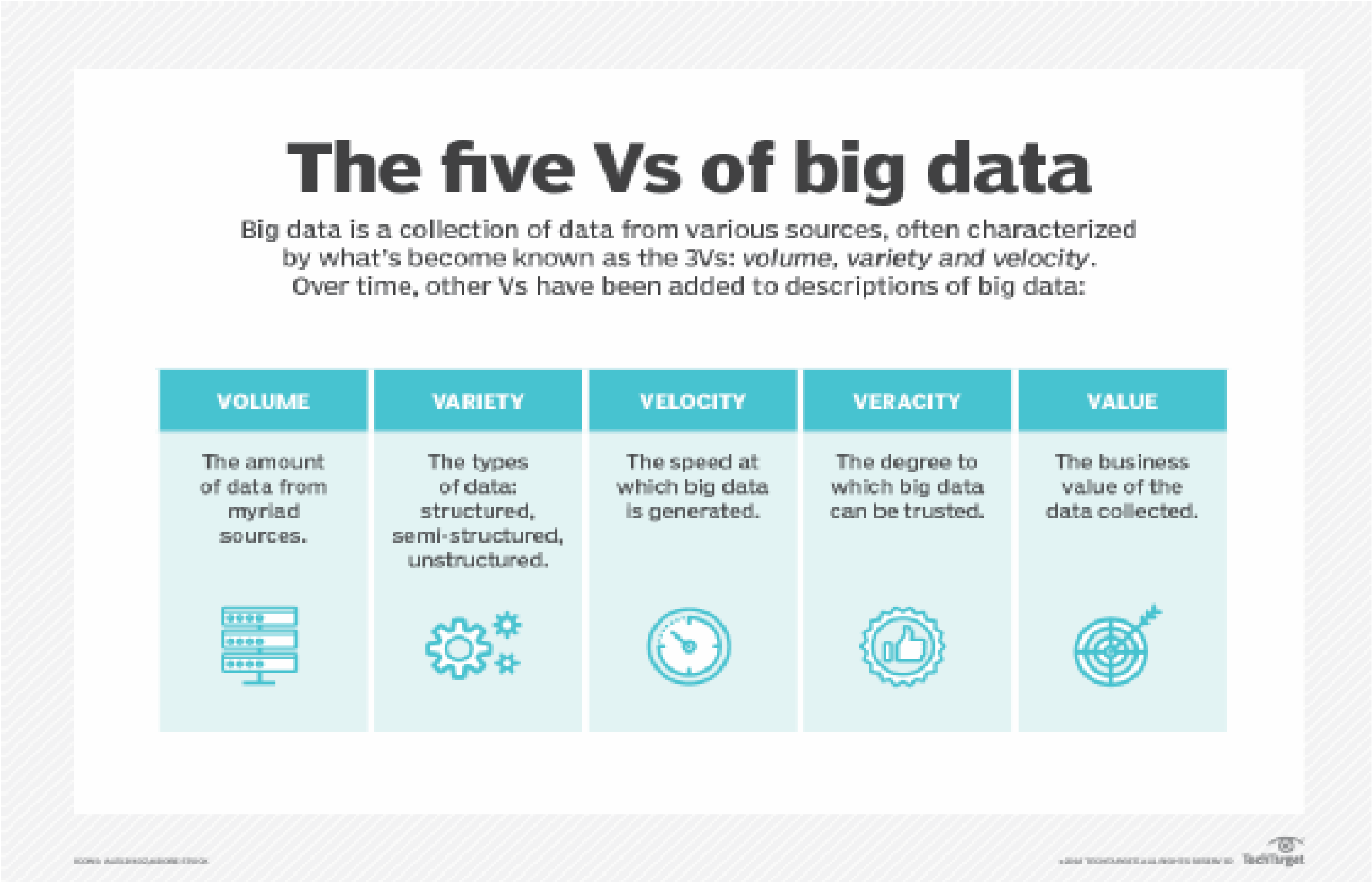
a) **Variety:** Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data. BIG DATA ANALYTICS 3

b) **Velocity**: Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

c) **Volume:** Volume is one of the characteristics of big data. We already know that Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses. Thus, comes to the end of characteristics of big data.

d) **Veracity:** It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control. Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources. Example: Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

e) **Value**: After having the 4 V’s into account there comes one more V which stands for Value. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful. Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V’s.



The five Vs of big data

Big data is a collection of data from various sources, often characterized by what’s become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.

TechTarget

Importance of Big Data

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyse it to find answers which will enable:

1. **Cost Savings**: Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.

2. **Time Reductions:** The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analysing data immediately and make quick decisions based on the learning.

3. **Understand the market conditions:** By analysing big data you can get a better understanding of current market conditions. For example, by analysing customers' purchasing behaviours, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

4. **Control online reputation**: Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.

5. **Using Big Data Analytics to Boost Customer Acquisition and Retention**: The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business

cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behaviour is important to trigger loyalty.

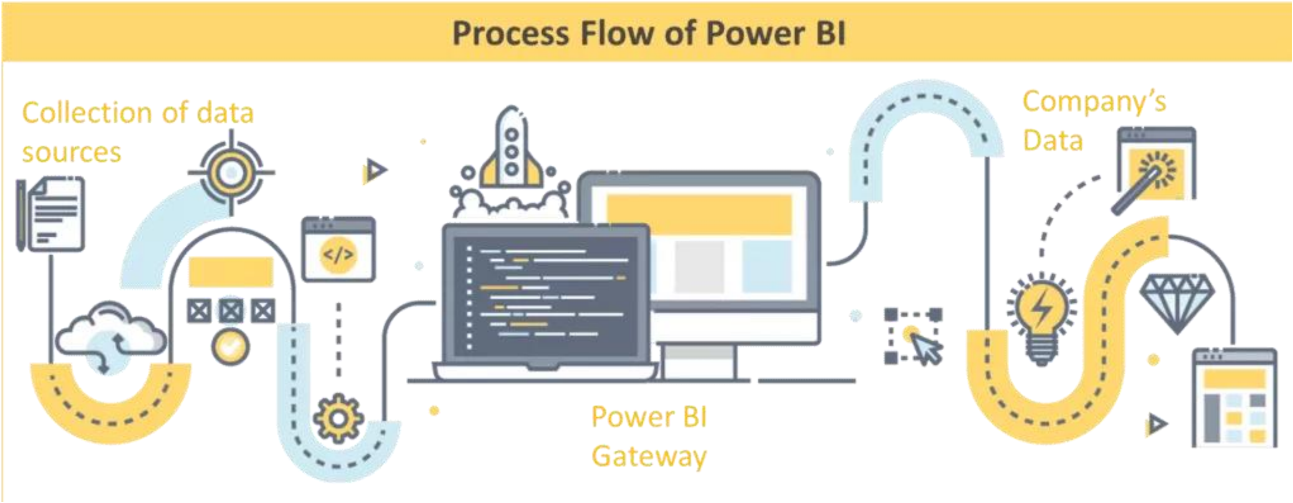
6. **Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights**: Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company’s product line and of course ensuring that the marketing campaigns are powerful.

7. **Big Data Analytics as a Driver of Innovations and Product Development**: Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

Power Bi

Introduction to Power Bi

Microsoft Power BI is a Business Intelligence service that enables you to create visually rich and interactive dashboards and reports based on the raw business data acquired from various sources. Power BI is a cloud-based user-friendly business intelligence platform that helps organisations and users to collate, manage and analyse data from a variety of sources. Power BI is used strategically to convert raw data into intelligible insights using visual charts and dashboards. This allows users to generate convincing reports and share the present state of the business. Business users utilize these services to collect data and generate BI reports. These three components are all meant to assist in building, exchanging, and leveraging business insights in the most efficient way possible for any business.

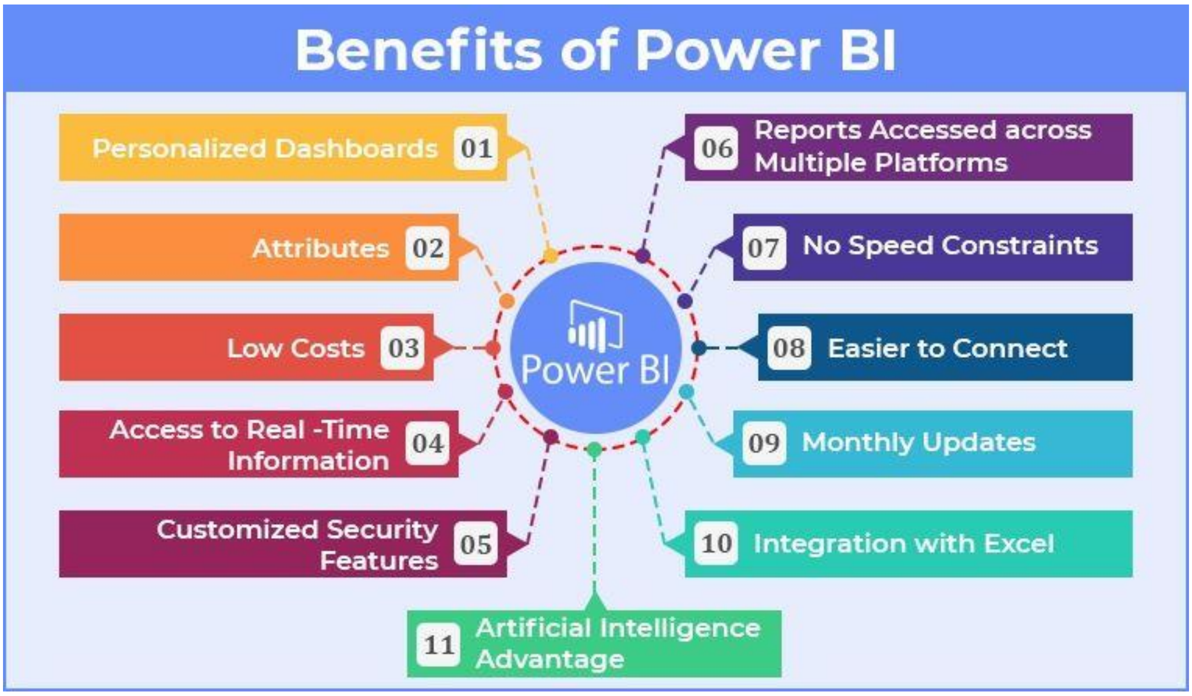


Benefits of Using Power BI

With Power BI, businesses can scrutinise and visualise their data from across the organisation. This gives them greater insights into their operations and performance which allows for informed decisions based on real data.

- **Large volumes of data** – Power BI can take advantage of huge volumes of data that most other platforms would struggle with. Power BI has impressive compression abilities which allow to analyse and visualise data that cannot be opened in Excel. Large datasets need not be sampled and aggregated to show a brief analysis. Power BI allows for all the granular details to be accessible through drill-downs.
- **Modify and prepare data for analysis** – Data cleaning and transformations can be performed using Power BI which includes changing data formats, adding and deleting rows and columns, transposing, pivoting and unpivoting tables, creating calculated measures, columns and tables. Creating relationships between multiple tables especially when the data warehouse uses star or snowflake schema. New datasets can be added into the data model without the need for restructuring the entire data model.
- **Rich personalised dashboards** – Information dashboards can be customised to meet the exact requirements of an individual or a team in an organisation.
- **Publish reports to multiple consumers securely** – Power BI helps to set up periodic data refreshes and publish reports allowing all the users to avail the latest information. It also ensures data security, offering controls on accessibility of the reports, dashboards and data both internally and externally.

- **Supports Advanced Data Analysis** – Power BI has built-in machine learning features that can analyse data and helps users spot valuable trends and make educated predictions. Power BI has a forecasting feature based on historic data. Automated machine learning (AutoML) for dataflows enables business analysts to train, validate, and invoke Machine Learning models directly in Power BI. Power Bi can also be used for text analytics for example visualising customer comments.
- **Power BI Integrations** – The dashboards and reports can be embedded and integrated into different applications using the JavaScript SDK with the dashboard embedding API. The platform also with business management tools like Office 365, SharePoint, Dynamics 365, Spark, Hadoop, Google Analytics, SAP, Salesforce and Mail Chimp.
- **Specialised Technical Support** – Power BI provides an agile approach and analysis so that there is no requirement for specialised technical support. It supports a powerful Natural language interface with the use of graphical designer tools.



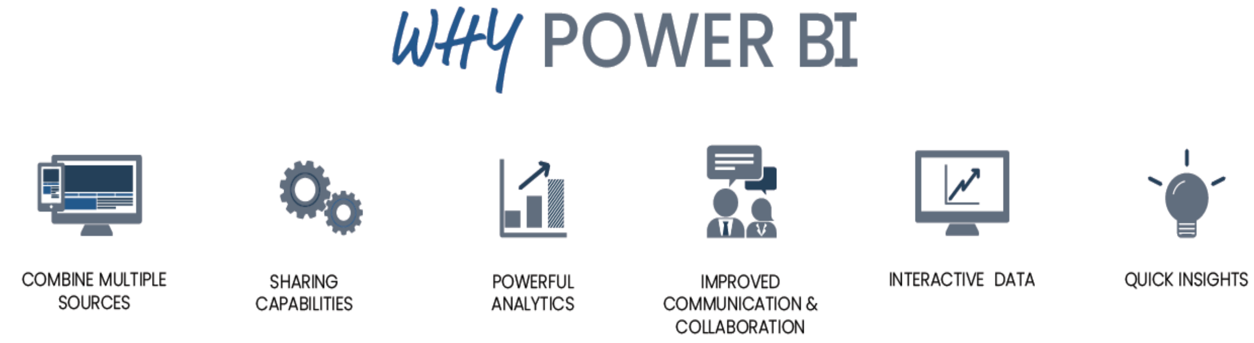
Benefits of Power BI
Personalized Dashboards 01
Attributes 02
Low Costs 03
Access to Real -Time Information 04
Customized Security Features 05
06 Reports Accessed across Multiple Platforms
07 No Speed Constraints
08 Easier to Connect
09 Monthly Updates
10 Integration with Excel
11 Artificial Intelligence Advantage
Power BI

Uses of Power Bi

There are various tools and techniques for analytics and machine learning in the fascinating and extensive realm of data science. Power BI is a high-level, all-in-one solution for data analytics in data science. Data science aids in the discovery of relevant and productive trends and insights. It involves analyzing the data and also assists us in identifying entirely new features in it. Business intelligence is sifting through data to extract meaningful organizational ideas and insights. BI enhances and strengthens the business infrastructure to get desired or projected results.

Many data sciences and analysis tasks can be automated with Power BI, eliminating the need for spreadsheets and static presentation tools. One of Power-most BI's most impressive features is its ability to create stunning visualizations. The software is packed with excellent and eye-catching visualization templates. The integration of Power BI into Data Science holds great importance for businesses. This allows for smooth and effective data visualization, which plays a vital role in an organization’s success.

With the help of Power BI, visualization in Data Science can be taken a notch further. Businesses and Data Scientists rely heavily on Power BI-aided data visualization for various projects.



Jupyter Notebook: The Classic Notebook Interface

The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.

Introduction to Jupiter Notebook

The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document. You can use Jupyter Notebooks for all sorts of data science tasks including data cleaning and transformation, numerical simulation, exploratory data analysis, data visualization, statistical modelling, machine learning, deep learning, and much more.

A Jupyter Notebook provides you with an easy-to-use, interactive data science environment that doesn't only work as an integrated development environment (IDE), but also as a presentation or educational tool. Jupyter is a way of working with Python inside a virtual "notebook" and is growing in popularity with data scientists in large part due to its flexibility. It gives you a way to combine code, images, plots, comments, etc., in alignment with the step of the "data science process." Further, it is a form of interactive computing, an environment in which users execute code, see what happens, modify, and repeat in a kind of iterative conversation between the data scientist and data. Data scientists can also use notebooks to create tutorials or interactive manuals for their software.

A Jupyter notebook has two components. First, data scientists enter programming code or text in rectangular "cells" in a front-end web page. The browser then passes the code to a back-end "kernel" which runs the code and returns the results. Many Jupyter kernels have been created, supporting dozens of programming languages. The kernels need not reside on the data scientist's computer. Notebooks can also run in the cloud such as Google's Collaboratory project. You can even run Jupyter without network access right on your own computer and perform your work locally.

Uses of Jupyter Notebooks

Jupyter is being used to do Python machine learning work. It’s a great environment with which to develop code, and also communicate results.

The name “Jupyter” was chosen to bring to mind the ideas and traditions of science and the scientific method. Additionally, the core programming languages supported by Jupyter are Julia, Python, and R. While the name Jupyter is not a direct acronym for these languages (Julia (Ju), Python (Py) and R), it does establish a firm alignment with them.

Summarizing The Pros and Cons:

We can conclude that these Notebooks are absolutely amazing in performing tasks related to visualizations, cleansing of data, and any projects related to data science or Python in general.

Pros:

- Best Platform for getting started with data science.
- Easy to share notebooks and visualizations.
- Availability of markdowns and other additional functionalities.

Cons:

- Lack of powerful features which are included in some IDE’s.

NumPy

Introduction to NumPy

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely. NumPy stands for Numerical Python

Uses of NumPy

In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy. Arrays are very frequently used in data science, where speed and resources are very important.

Difference Between NumPy

NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently. This behaviour is called locality of reference in computer science. This is the main reason why NumPy is faster than lists. Also, it is optimized to work with latest CPU architectures.

Language of NumPy

NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

Pandas

Introduction to Pandas

Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Uses of Pandas

Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

Matplotlib

Introduction to Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

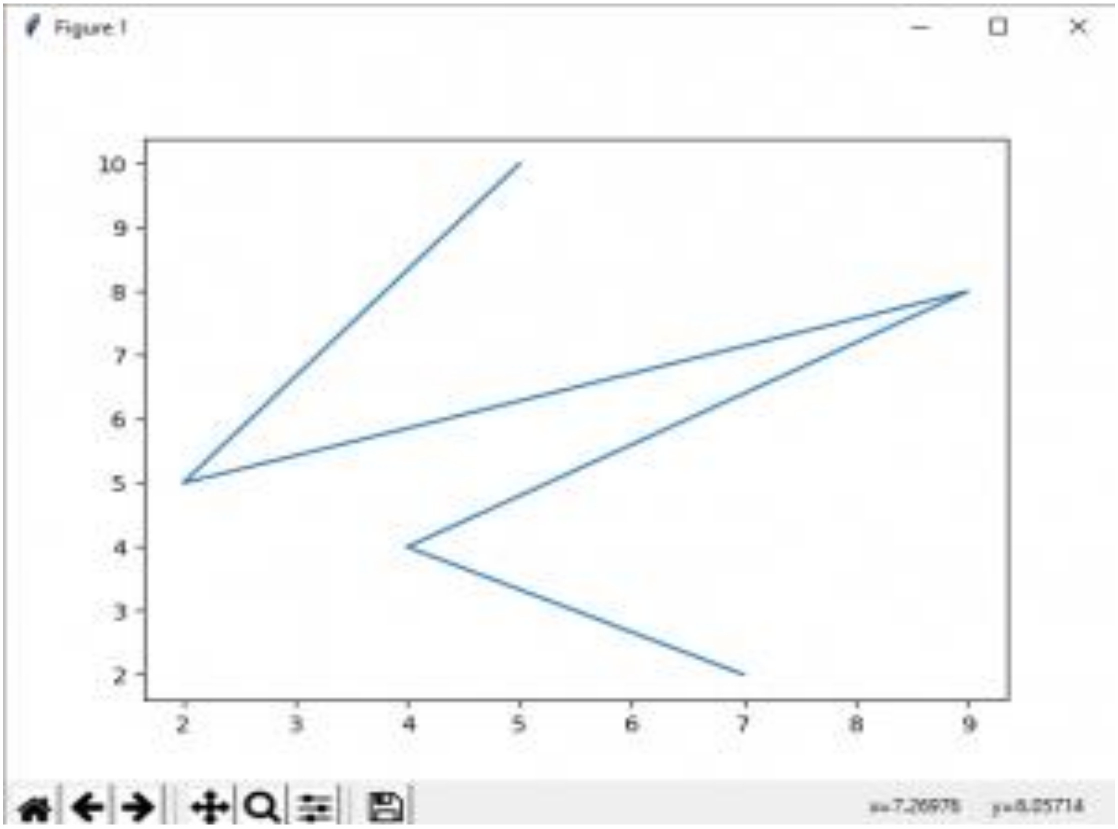
Basic plots in Matplotlib:

Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some of the sample plots are covered here.

Line Plot:

```
# importing matplotlib module
from matplotlib import pyplot as plt

# x-axis values
x = [5, 2, 9, 4, 7]

# Y-axis values
y = [10, 5, 8, 4, 2]

# Function to plot
plt.plot(x,y)

# function to show the plot
plt.show()
```

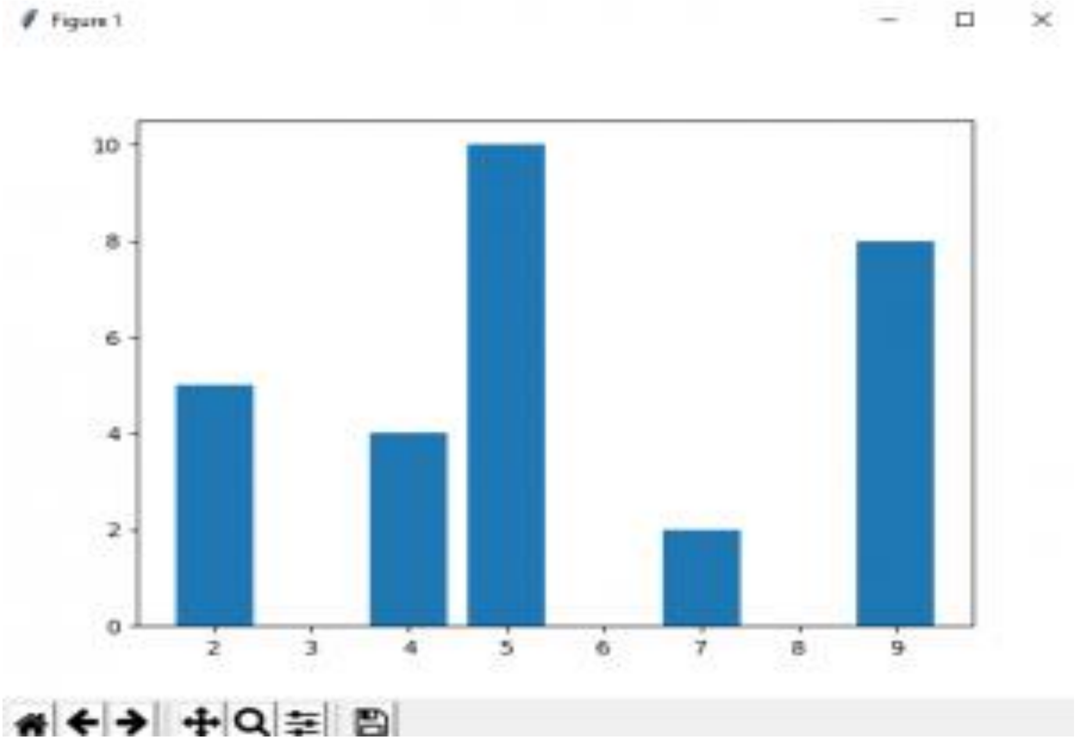
Output:



Bar Plot:

```
# importing matplotlib module
from matplotlib import pyplot as plt

# x-axis values
x = [5, 2, 9, 4, 7]

# Y-axis values
y = [10, 5, 8, 4, 2]

# Function to plot the bar
plt.bar(x,y)

# function to show the plot
plt.show()
```

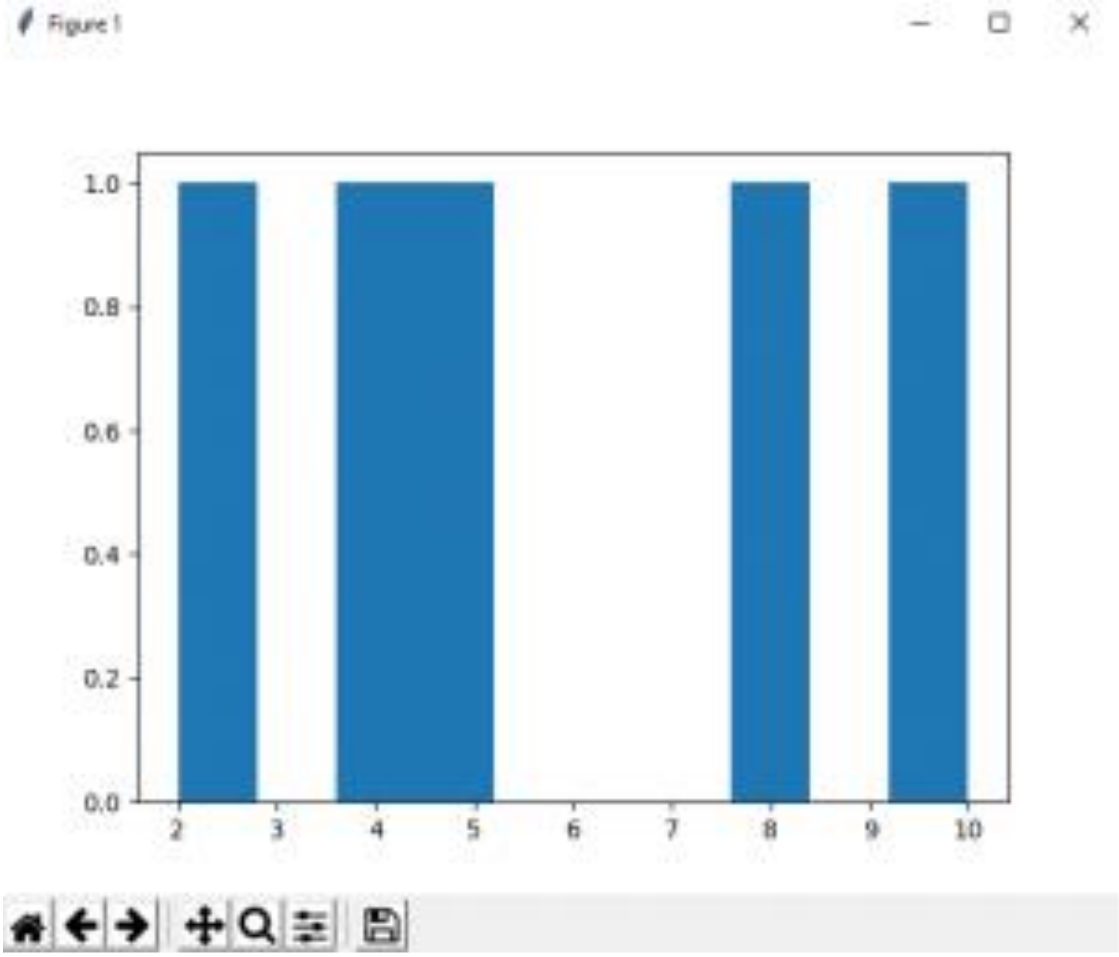
Output:



Histogram:

```
# importing matplotlib module
from matplotlib import pyplot as plt

# Y-axis values
y = [10, 5, 8, 4, 2]

# Function to plot histogram
plt.hist(y)

# Function to show the plot
plt.show()
```

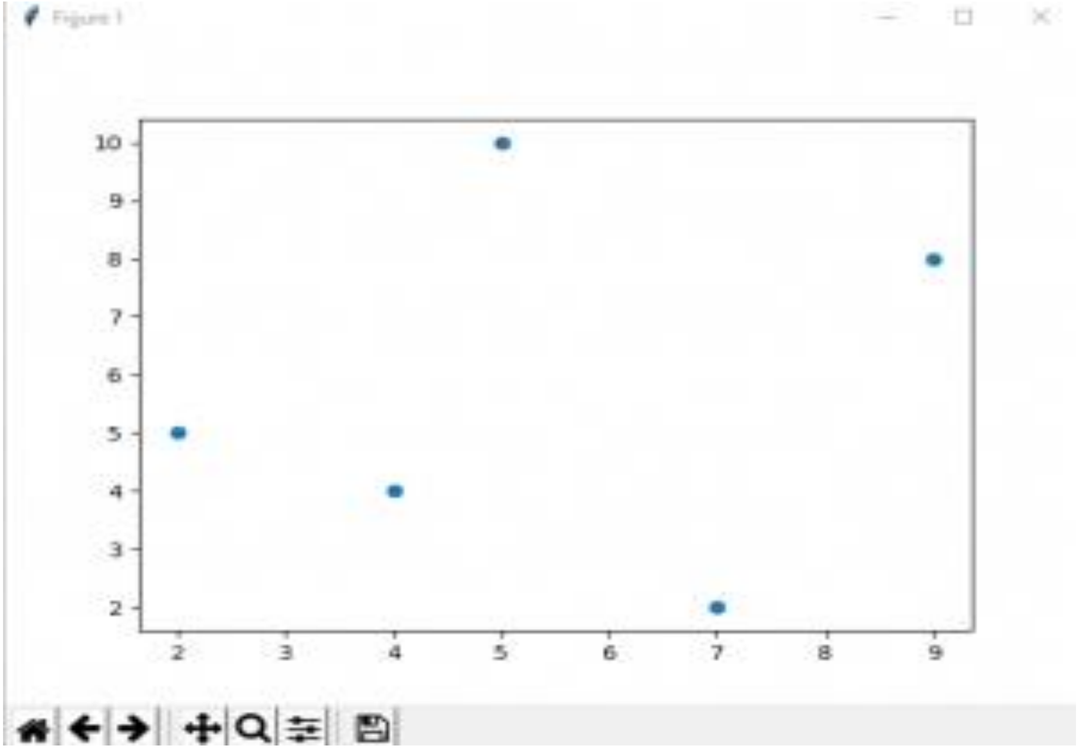
Output:



Scatter Plot:

```
# importing matplotlib module
from matplotlib import pyplot as plt
# x-axis values
x = [5, 2, 9, 4, 7]

# Y-axis values
y = [10, 5, 8, 4, 2]

# Function to plot scatter
plt.scatter(x, y)

# function to show the plot
plt.show()
```

Output:



CONCLUSION

This study presents an intelligent market basket analysis using big data analytics. It is applied in V Mart, a top retail store in Jordan. It presents recommendation for the management to adjust marketing, sales and layout design strategy to cope up with the explosion in the digitization revolution and to account for the negative effects of COVID-19 pandemic. Big data analytics utilizes cosine-similarity indexes to find correlations. Customized baskets are generated based on customers’ shopping behaviour; when calculating similarity index between customers. General baskets are generated based on correlation between stock items which are usually bought together; within same basket. Both physical and online merchandizing is adjusted; where customized marketing which targets VIP customers is recommended. This gives customers a more satisfying experience in shopping at V Mart because, they easily find items; which enlarges the market basket size and then increases sales.

In this study, the association between items is considered regardless of their quantity. Further study could be conducted to include the quantity of each item. In addition, future research should aim to include different measures related customer stay time, distance moved, storage capacity, etc. this can help improve understanding of in-store behaviour which is one important factor impacting customers’ shopping behaviour.

References

[1] S. Amaro, How the coronavirus is changing the way we shop — and what we’re buying, JUL 27 2020. https://www.cnbc.com/2020/07/27/the-future-of-retailamid-covid-19.html

[2] R. Vader, P. Martin, J. Qian, The realities of retailing in a COVID-19 world, KPMG Insights, 2020. https://home.kpmg/xx/en/home/insights/2020/03/realitie s-of-retailing-in-covid-19-world.html

[3] D. Andrews, Shoring Up Your Marketing Strategy in Turbulent Times, customer think April 10, 2020 https://customerthink.com/shoring-up-your-marketingstrategy-in-turbulent-times/

[4] Steve Harris, Safe shopping: the impact of COVID-19 on retail, https://www.orange-business.com/en/blogs/safeshopping-impact-covid-19-retail, August 14, 2020, Digital Transformation

[5] M. Alshriem, The Use of Artificial Intelligence in Traffic Violation Data Analysis, International Journal of Engineering Research and Technology. ISSN 0974-3154 Vol.13, No.4 (2020), pp. 644-652.