A

MINOR PROJECT REPORT

on

STARTUP INSIGHTS

BE(AI&DS)-VI Sem

By

Sardar Rechel Blessy (160120771014)

Supraja Balerao (160120771018)

Kushal Rathi (160120771034)

Under the guidance of

Dr. T Prathima

Associate Professor

AIDS Department

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY (A)

(Affiliated to Osmania University; Accredited by NBA(AICTE) and NAAC(UGC), ISO Certified 9001:2015)

KOKAPET(V),GANDIPET(M),RR District HYDERABAD - 75

Website: www.cbit.ac.in

2022-2023

This is to certify that the project work entitled "**STARTUP INSIGHTS**" submitted to CHAITANYA BHARATHI INSTITUTE OFTECHNOLOGY, in partial fulfillment of the requirements for the completion of Minor Project-II of VI Semester B.E. in Artificial Intelligence and Data Science, during the Academic Year 2022-2023, is a record of original work done by **Sardar Rechel Blessy(160120771014), Supraja Balerao(160120771018) and Kushal Rathi(160120771034)** during the period of study in the Department of AI&DS, CBIT, HYDERABAD, under our guidance.

Project Guide	**Head of the Department**
Dr. T Prathima	**Dr. K Ramana**
Associate Professor, Dept. of IT,	**Professor, Dept. of AI&DS,**
CBIT, Hyderabad.	**CBIT, Hyderabad.**

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to **Dr. T Prathima**, our project guide, for her invaluable guidance and constant support, along with her capable instruction and persistent encouragement.

We are grateful to our Head of Department, **Dr. K Ramana**, for his steady support and for the provision of every resource required for the completion of this project.

We would like to take this opportunity to thank our Principal, **Dr. P Ravinder Reddy**, as well as the Management of the Institute, for having designed an excellent learning atmosphere.

Our thanks are due to all members of the staff and our lab assistants for providing us with the help required to carry out the groundwork of this project.

ABSTRACT

The objective of this research is to analyze and comprehend the trends in the startup culture of India. Startup insights refer to the knowledge and understanding gleaned from the analysis and visualization of various startup venture information.

Startup insights can assist entrepreneurs, investors, and other stakeholders in the startup ecosystem to better understand the factors that come up while building a new business, and to make informed decisions regarding strategy, resource allocation, and risk management.

The research involved collection and analysis of data on numerous startup companies. Descriptive statistics were used to explain trends and patterns in the data such as, which sectors are attracting the most startups, which locations are seeing the most startup activity, yearly India has seen how many numbers of startups that have been established etc.

The results show that there is a positive relationship between the features and only few startups have entered the stock market. Moreover, majority startups belong to service-based model and series A. Lastly, it highlights huge spatial and sectoral investment concentration. Overall, startup insights are a valuable source of knowledge that can help entrepreneurs to build successful and sustainable businesses.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1.1 MOTIVATION

A startup is a newly established business venture, typically characterized by innovation and high growth potential. Startups are often founded by entrepreneurs who have a vision for a product or service that they believe will solve a problem or meet a need in the market. They often operate with limited resources and rely on creativity, ingenuity, and a willingness to take risks to succeed.

Successful startups can grow rapidly and become major players in their industries, often disrupting traditional business models along the way. However, startups also face many challenges, including raising capital, developing a viable business model, building a customer base, and competing with established companies.

In recent years, the startup ecosystem has grown significantly, with many resources available to support entrepreneurs, such as incubators, accelerators, and venture capital funding. Overall, startups play a vital role in driving innovation and growth in the economy.

1.2 OBJECTIVE OF THE PROJECT

The objective behind startup insights is to gain a deeper understanding of the startup ecosystem, including emerging trends, challenges, and opportunities. This knowledge can be used to inform decision-making, develop effective strategies, and drive innovation and growth. By analyzing this, stakeholders can gain insights into the needs and preferences of customers, the competitive landscape, and potential barriers to entry.

The primary goal of the startup insights is to help entrepreneurs, investors, and other stakeholders make informed decisions about business opportunities and investments. This includes identifying emerging trends and opportunities, assessing market demand, and developing strategies to address challenges and stay ahead of the competition. Overall, the objective behind startup insights is to foster a deeper understanding of the startup ecosystem and leverage that knowledge to drive innovation, growth, and economic development.

LITERATURE SURVEY

There is a limited amount of research analyzing startups in India, with most of the literature being descriptive in nature. However, there have been some attempts to study the Indian startup ecosystem and identify trends. In this literature survey, we will review some of the relevant studies that describe these trends.

Piyush Anand Verma and Vikas Singhal (2018) [1] observed that the events like Demonetization, Surgical Strike and Digital India Initiative affected the funding by investors into the companies. Also, the cities and industry verticals play an important part in acquiring funding from investors, and foreign investors invest a lot into Indian startups.

Nopadol Rompho (2008) [2] recommends that startups should focus on measuring and monitoring the right metrics for their particular stage of development, instead of trying to excel in all areas. This approach helps them maintain focus and avoid spreading themselves too thin, which can ultimately lead to failure.

According to Giurca Vasilescu (2009) [3] suggestion, developing companies primarily rely on investors to provide funding. These investors offer financial assistance to the company from the outset until it is ready to enter the capital market. Additionally, investors offer managerial support to help companies survive and succeed in the competitive market.

Kala Seetharam Sridhar and Fakih Amrin Kamaluddin (2021) [4] found that investment is heavily concentrated in platform and aggregator startups which comprise 30% of the sample and attract more than two-thirds of the total investment. Investment is also concentrated in the e-commerce sector.

Tracxn (Velayanikal, 2015) [5], a venture capital analytics firm discovered in their study that a total amount of $6.4 billion funding was given to various startups in the first nine months of 2015, out of which $3.4 billion was invested in online marketplace in India.

In their research on the Indian startup ecosystem between 2015-16 and 2018-19, Narayan et al. (2019) [6] used annual funding reports from Trak.in to identify investment trends. They discovered that there was no significant correlation between the level of development of a startup and the stage of funding it received.

In their study on angel investment trends from 2011 to 2015, Rao and Kumar (2016) [7] noted a significant increase in angel investment during that time period. They also observed that angel investors play a crucial role in funding early-stage startups, in contrast to venture capitalists. Furthermore, they found that angel investors invest in a diverse range of sectors.

According to Parul Grover (2019) [8], The growth of the Indian startup community can benefit the overall economy by driving growth and development. To establish ecosystems that enable the growth and sustenance of entrepreneurship, collaboration between established business communities and new firms is necessary. This collaboration can accelerate the identification of startups and improve the idea generation procedure. Venture capitalists in India can also provide mentorship, guidance, business insights, and capital funding to accelerate the growth of startups.

DATASET ACQUISITION

3.1 Existing dataset

The dataset "Indian Startups funding (in 2021)" was acquired from kaggle.com [11]. This dataset contains various details about startups that have been established in the past 40 years. It consists of 1195 rows and 10 features. Additional rows were obtained from "Indian Startups-Funding and Investors data" from the Startup Talky [12] website. Startup-Talky is a top startup media platform which contains the same details about many other startups. The features in the existing dataset are given in Table. 3.1.

Table. 3.1: Existing Dataset Features

Feature	**Description**	**Domain**
Company/Brand	Name of the company	Character
Founded	Year when the company was established	Numerical
Headquarters	Location of headquarter	Character
Sector	Domain to which the company belongs	Character
Description	General details about what the company does	Character
Founder's	People who founded the company	Character
Investor's	People who invested in the company	Character
Amount	Total valuation of the company	Character
Funding Stage	Stage of funding of the company	Character
Funding Month	Month of funding of the company	Numerical

3.2 New Dataset:

The additionally obtained rows were merged into the existing dataset. Five additional features were initially considered to be included in it. The five additional features are employee count, IPO, product, or service based, paid up capital and authorization. After some research, paid up capital and authorization features were eliminated because these details are undisclosed by the companies. Finally, the three additional features which were included in the dataset are given in Table. 3.2.

Table. 3.2: Newly Added Features

Feature	Description	Domain
Employee Count	Number of employees in the company	Character
IPO	whether the company is listed or not in the stock market	Character
Type	whether the company is product based or service based	Character

The resources used for retrieving the values of above three additional features are LinkedIn [13] and Crunchbase [14] for employee count, BSE [15] and NSE [16] for IPO, and Crunchbase [14], CorporateInformation.com [17], and The Company Check [18] websites to check whether a company is product based or service based.

METHODOLOGY

4.1 Methodology for Data Analysis and Visualization:

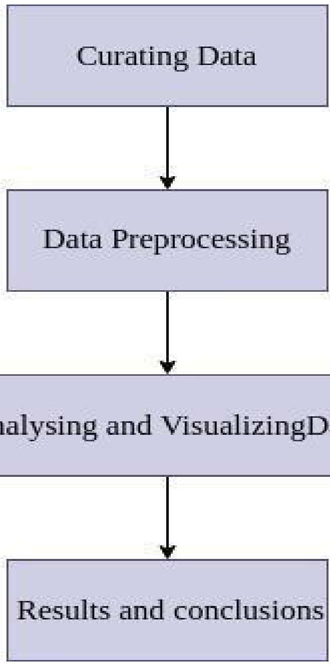

Fig 4.1: Methodology for Data Analysis and Visualization

4.2 Classification based on Type of Company:

An organization that manufactures and offers specialized products to meet the unique needs of a large customer base is known as a product-based firm. These businesses' main goals are to develop a loyal client base and introduce or supply their items to customers.

A business that focuses on providing services to other businesses or its clients, customers, or clientele is known as a service-based business. The company's service is a non-tangible good that can be delivered through facilities, amenities, and expertise.

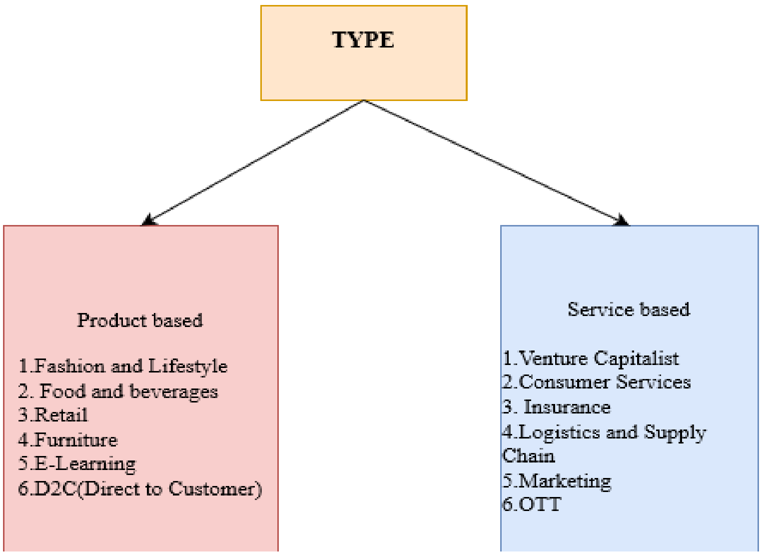

Fig. 4.2: Classification based on Type of Company

4.3 ANALYSIS TASKS

Analysis tasks are important in startup visualization projects as they help to identify patterns, trends, and insights from the data, which can be presented visually through charts, graphs, and other visual representations. In this research various tasks were performed that help in gaining insights into startups in India. They are:

1. Which city has produced more startups- Visualization using Bar chart and Pie Chart
2. Which sector has given more startups- Visualization using Pie Chart
3. Which startups have successfully entered stock market- Visualization using Bar chart and Pie Chart
4. How many are product based or service-based startups – Visualization using Bar chart (Done city wise also)
5. Relation between Sector and Type (product or service-based startup)- Machine learning algorithm (Naïve Bayes algorithm, Random Forest and Decision Tree) and Statistics (CramerV correlation coefficient)
6. Year wise India has produced how many startups- Visualization with Line chart
7. City wise top 5 sectors – Visualization with Stacked Bar chart
8. Analysis based on stage funding – Visualization with Pie Chart
9. Which startups are in what month of their funding – Visualization with Bar chart
10. Employees count working in most startups – Visualization with Bar chart
11. Top Investors in the startups – Text segmentation and creating a new csv file, Visualization with the help of a Histogram
12. Investments made more in which type of company – Visualization with Bar chart
13. Valuation Distribution of Top sectors – Visualization with Bar Chart

Overall, analysis tasks are critical in startup visualization projects because they provide the insights that can be visualized to help stakeholders understand the data and make informed decisions. By using visualizations to communicate insights, stakeholders can more easily understand the information and make better decisions based on the data.

4.4 PRE-PROCESSING TASKS

Data preprocessing is important for visualization projects because it helps to prepare the data in a way that is suitable for visual analysis. Data from multiple sources may need to be integrated to create a comprehensive visualization. Preprocessing can help to integrate such data, ensuring that the visualizations are based on a complete and accurate picture of the data. Raw data may contain errors, missing values, or inconsistencies that can affect the accuracy and validity of the visualization. Preprocessing helps to identify and clean such errors, ensuring that the visualizations are based on accurate and reliable data.

1.Merging two datasets

A key data manipulation strategy in research is merging two or more datasets, and R has various functions to carry out this operation. When combining various data sources with diverse but relevant information, merging datasets is especially helpful. Researchers can study and understand the data more thoroughly using this technique, leading to more reliable and accurate results.

By appending the rows of one dataset to the other, the rbind() method can be used to join datasets with the same variables. The datasets must have the same variables with the same column names in order to combine them using rbind().rbind() is a simple and quick method for attaching new observations to an existing dataset or for merging two datasets with the same variables, which is one benefit of using rbind() to combine datasets.

```
# Merging two datasets
# Reading the existing datasets from Excel file
library(readxl)
existing_data <- read_excel("D:\\dsr.xlsx")
# Printing the class of each column
sapply(existing_data,class)
# Reading the new dataset to be added from Excel file
new_data <- read_excel("D:\\dsr2.xlsx")
# Printing the class of each column
sapply(new_data,class)
```

2. Replacing values and converting categorical values to logical

Replacing values and converting categorical values to logical values is a common preprocessing task that involves replacing values in a dataset with appropriate values and converting categorical variables to binary variables that can be used in statistical analysis.

Any values in the data$IPO column that contain the string "YES" are replaced with the string "Yes” .This task exemplifies data cleaning, which ensures that data is consistent and accurate.

The ifelse() function is used to convert the values in the data$IPO column to binary values (0 or 1).Any value in the data$IPO column that equals "No" is replaced with 0, and any other value is replaced with 1.This task is an example of data transformation, which is the process of converting data from one form to another in order to facilitate analysis.

```
# Replacing "YES" with "Yes"
data$IPO=gsub("YES","Yes",data$IPO)
# Replacing "No"/"Yes" with 0/1 in IPO column
data$IPO=ifelse(data$IPO=="No",0,1)
# Counting the number of occurrences of each value in IPO column
table(data$IPO)
```

3. Renaming column name

Renaming column names is a common data analysis preprocessing task that involves changing the names of columns in a dataset.This task is critical for a number of reasons, including improving data readability and interpretability and ensuring that column names are consistent across multiple datasets.

```
# Renaming the "Headquarters" column to "cities"

colnames(data)[colnames(data) == "Headquarters"] = "cities"

# Printing the class of each column in merged dataset

sapply(data,class)

# Printing unique values in cities column

unique(data$cities)
```

4. Replacing values

Replacing values entails replacing specific values in a column with other values.In this case, the values "Product" and "Service" in the "Type" column are replaced with "product" and "service," respectively.

```
# Changing all instances of 'Product' to 'product'

data$Type=gsub("Product","product",data$Type)

# Changing all instances of 'Service' to 'service'

data$Type=gsub("Service","service",data$Type)

# Displaying the first 10 rows of the 'Type' column

head(data$Type,10)
```

5. Filling empty cells with NA, changing datatype and removal of comma

This entails performing several preprocessing tasks on the "Amount(in dollars)" column.First, any empty cells in the column are replaced with NA (missing values).Second, any cells with the value "Undisclosed" are replaced with NA.Finally, any commas in the values are removed with the gsub function.Finally, the datatype of the column is changed to numeric using the as.numeric function.

```
# Replacing empty cells and cells with 'Undisclosed' value with NA

library(dplyr)

data$`Amount(in dollars)` <- na_if(data$`Amount(in dollars)`, '')
```

data$`Amount(in dollars)` <- na_if(data$`Amount(in dollars)`, 'Undisclosed')

 # Removing comma separator from Amount(in dollars)

data$`Amount(in dollars)`=gsub(",","",data$`Amount(in dollars)`)

 # Changing the datatype of the 'Amount(in dollars)' column to numeric

data$`Amount(in dollars)`=as.numeric(data$`Amount(in dollars)`)

summary(data$`Amount(in dollars)`)

 # Checking how many NA values are there in the column

sum(is.na(data$`Amount(in dollars)`))

 # Displaying the first 20 rows of the 'Amount(in dollars)' column

head(data$`Amount(in dollars)`,20)

6. Assigning variables to data ranges

It entails assigning categorical labels to numerical variables based on specific data ranges.In the project, the "Employee Count" column is converted into a categorical variable by dividing the employee counts into specific ranges and labeling each range.The factor function is used to define the ranges and labels.The levels parameter defines the numerical ranges, while the labels parameter assigns a label to each range.The range "(2-10)" is labeled "A," while the range "(11-50)" is labeled "B," and so on.

 # Printing unique values in Employee Count column

unique(data$`Employee Count`)

 # Counting the number of occurrences of each value in Employee Count column

table(data$`Employee Count`)

 # Converting Employee Count column to a factor type with ordered levels

data$ `Employee Count ‘<- factor(data$ `Employee Count`,levels = c("(2-10)","(11-50)","(51-100)","(51-200)","(101-250)","(201-500)","(251-500)","(501-1000)","(1001-5000)","(5001-10000)","(10001+)"),labels=c("A","B","C","D","E","F","G","H","I","J","K"))

 # Counting the number of occurrences of each value in Employee Count column

table(data$`Employee Count`)

This transformation can provide a more intuitive way to group the data based on the ranges defined by the labels, which can help to simplify and make the data easier to analyse.It can also help to reduce data noise and highlight important patterns or trends that may not be immediately apparent in the original numerical data.

7. Filling empty cells with NA and replacing values

Involves cleaning and transforming the values in the data's "Stage" column.The first step is to replace any empty cells with the missing data value "NA."To address missing data or inconsistent entries, the next step is to replace certain values with the "NA" value, such as "Unknown," "undisclosed," and "Undisclosed."The following set of operations consists of replacing specific values with more standardized or simplified values.To create consistency in the labels, variations of "Pre-series" and "seed" are replaced with just "Seed."Similarly, "Debt Financing" is dropped in favor of "Debt."

Other operations include standardizing label variations, such as replacing "Seed B" with "Seed" and "Early Seed" with "Seed."Finally, any remaining distinct values in the "Stage" column are shown to ensure consistency and accuracy.

Overall, these preprocessing steps are critical for cleaning and standardizing data, which can aid in the consistency and accuracy of downstream analysis and modeling tasks.

```
# Checking the unique values in the 'Stage' column
unique(data$Stage)
# Replacing empty cells and unwanted values with NA
library(dplyr)
data$Stage <- na_if(data$Stage, '')
data$Stage <- na_if(data$Stage, 'Unknown')
data$Stage <- na_if(data$Stage, 'Blue Ashva Capital, Supack Industries')
data$Stage <- na_if(data$Stage, 'undisclosed')
data$Stage <- na_if(data$Stage, 'Undisclosed')
data$Stage=gsub(c("Pre-series|seed"),"Seed",data$Stage)
data$Stage=gsub(c("Pre-seed|pre-Seed|PreSeed|Pre-Seed|Pre Seed"),"Pre seed",data$Stage)
data$Stage=gsub(c("Debt Financing"),"Debt",data$Stage)
data$Stage=gsub(c("Seed B"),"Seed",data$Stage)
data$Stage=gsub(c("Early Seed"),"Seed",data$Stage)
data$Stage=gsub(c("series A"),"Series A",data$Stage)
unique(data$Stage)
```

RESULTS & COMPARISON

5.1 Existing Tasks and Results Comparison

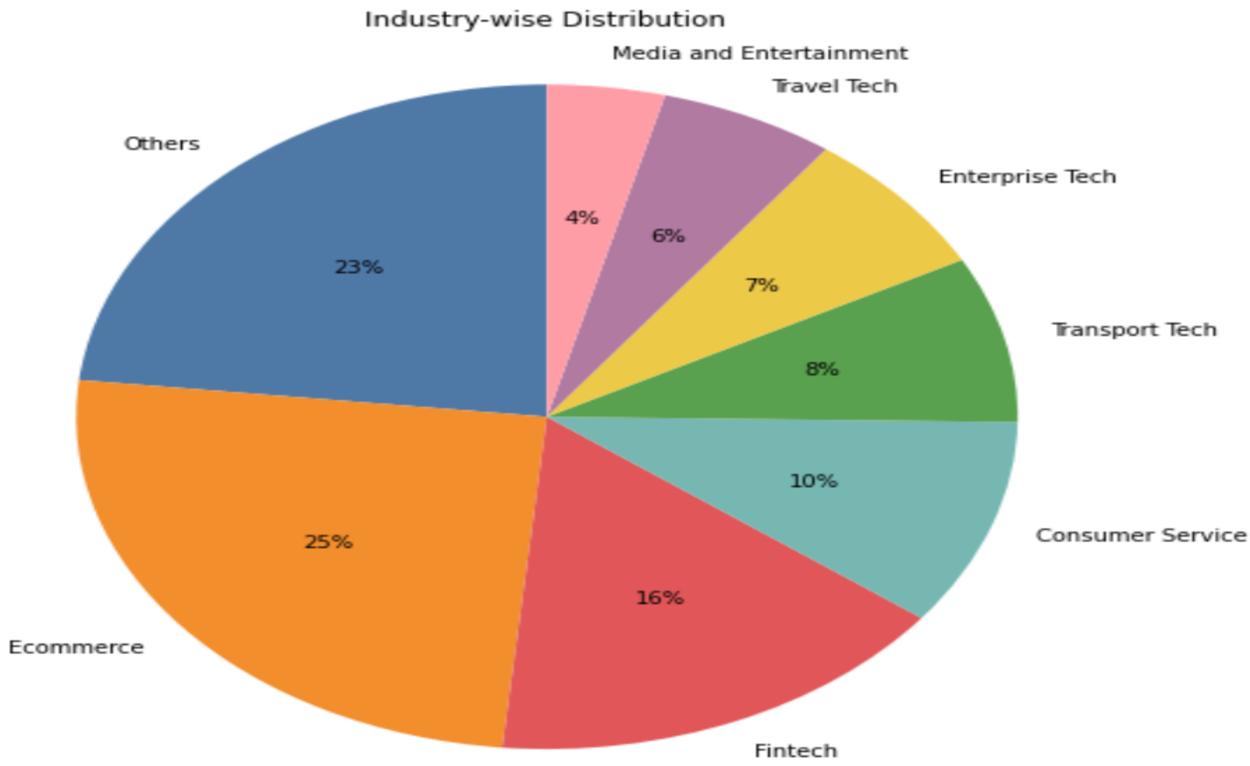

Fig. 5.1[4]: Sectoral Distribution of Investment

Based on previous results in Fig. 5.1[4] it is observed that Ecommerce Sector has got the highest sectoral distribution of investment, contributing to 25% of total investment, the Fintech sector has got 16 percent of investments.

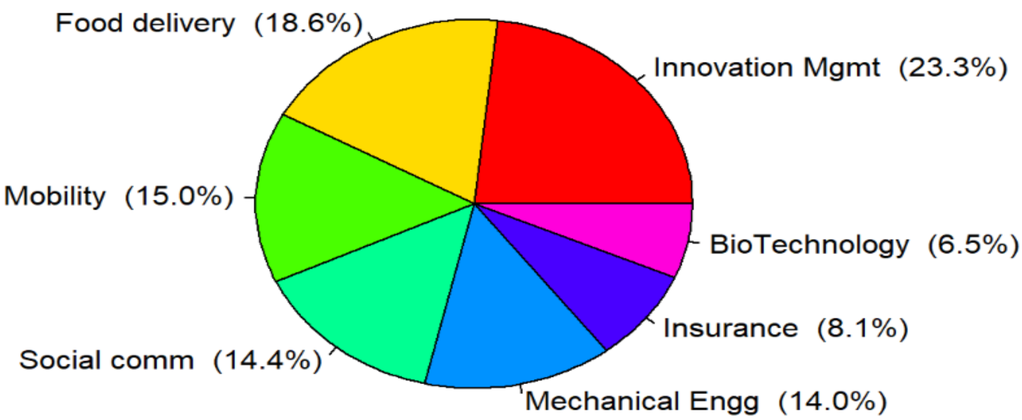

Fig. 5.2: Sectoral Distribution of Investment

From the analysis of various sectors in Fig. 5.2, it is seen that Innovation management has the Highest Valuation among all other sectors (the Innovation Management sector in startups involves the use of strategic and structured approaches to promote innovation and drive growth within the company. This may include the development of processes and tools for ideation, prototyping, testing, and commercialization of new products, services, or business models.) which is followed by Food and delivery which has seen a lot of potential growth in the past years

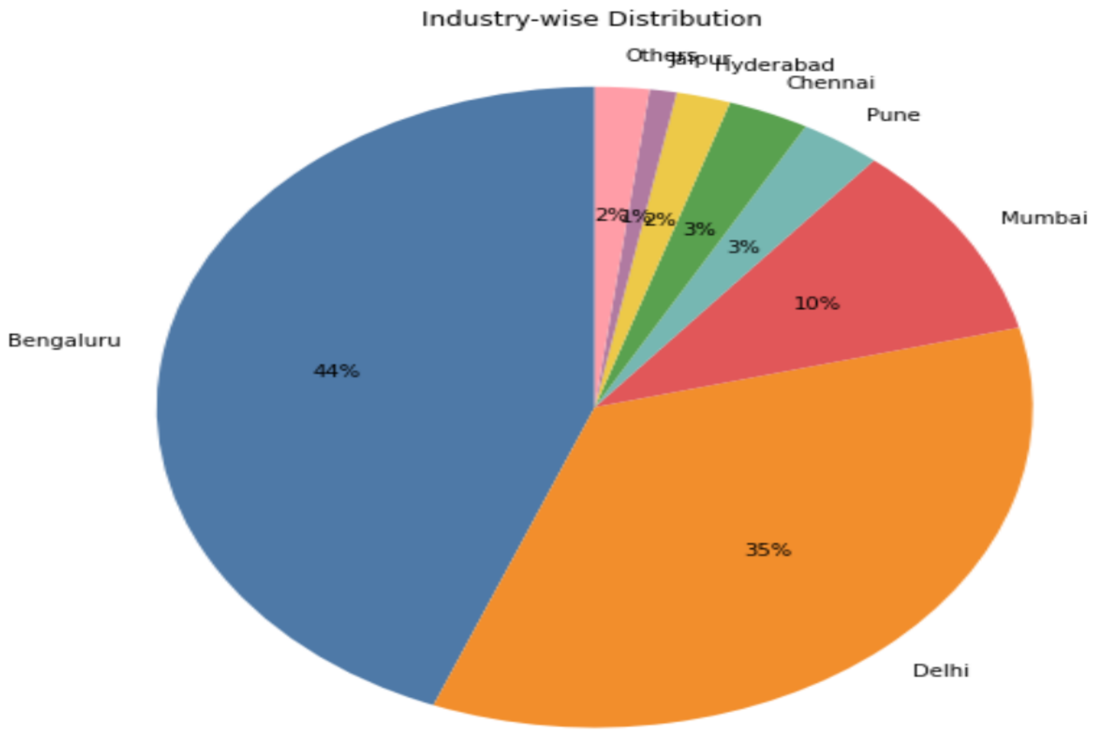

Fig. 5.3[4]: Spatial Distribution of Investment

Coming to the Sectoral Distribution of Investments in Fig. 5.3[4], Bangalore has got highest investments in startups accounting to 44% followed by the capital of India, Delhi which has got 35% of total investment received and other metropolitan cities follow by gathering investments.

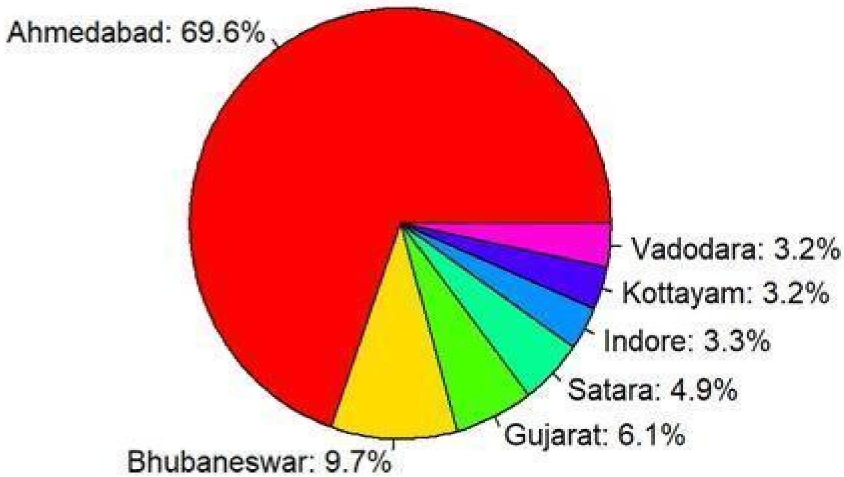

Fig. 5.4: Spatial Distribution of Investment in 2022

Coming to the results in Fig. 5.4, the past 2 years has seen more startups gathering investments in Western part of India with Ahmedabad having got 69.6 percent of total investments, followed by Bhubaneswar getting 9.7 percent of investments.

Startup funding stages are typically divided into rounds, with each round representing a significant milestone in the company's growth and development. The seed round is the first stage, in which a startup raises funds from family, friends, and angel investors to fund its early development. The Series A round is the next stage, in which a startup receives investment from venture capitalists in order to grow the business and expand its customer base. Series B, C, and beyond are the subsequent rounds, with each round providing larger amounts of funding to support further growth and expansion. Later rounds of funding may include additional investors such as private equity firms or hedge funds, as well as a secondary market for existing shareholders to sell their shares.

Typically, each round of funding involves negotiations between the startup and the investors about the company's valuation and the terms of the investment, such as the amount of equity offered and any special rights or preferences granted to the investors. The size and timing of each round are determined by a number of factors, including the startup's business model, market potential, and growth trajectory. Startups in certain industries, such as technology or biotech, may require more capital to fund R&D, whereas startups in consumer-focused industries may require less capital to reach profitability.

In general, the stages of funding in startups reflect the development of the business as it moves from a concept to a successful enterprise, and each round of funding signifies a crucial step toward realizing the goals and objectives of the company.

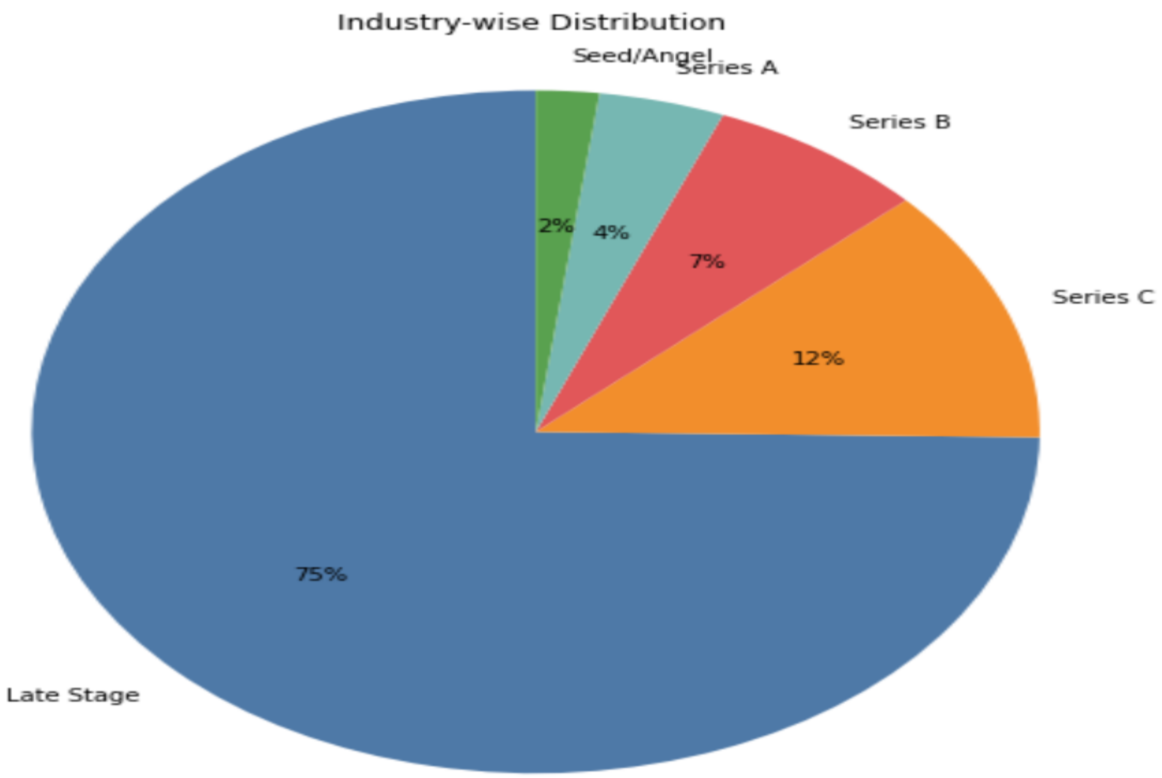

Fig. 5.5[4]: Stage-wise Distribution of Investment

5.2 Results and Visualizations

How many startups in India are product or service based?

 # To showcase the frequency of startups based on their product or service type in India and the top cities with the highest number of startups - Bangalore, Mumbai, Gurugram, and New Delhi. The first bar chart displays the frequency of startups based on their product or service type in India. The subsequent bar charts depict the same analysis for each of the top cities individually.

#The input is a table of frequency (numeric) of Type column (character). It is the respective count of product and service (for India and a particular city).

 # Plotting bar chart for all rows

freq=table(data$Type) # Counting the frequency of startups based on Type column

freq

barplot(freq, main = "How many startups in India are product or service based", xlab = "product or service", ylab = "Frequency", col = rainbow(length(freq))) # Plotting the bar chart for the frequency of startups based on product or service

df=data.frame(data) # Creating a data frame from the original data

Plotting bar chart for Bangalore

B=df[df$cities=="Bangalore",] # Filtering the data for startups in Bangalore

freq1=table(B$Type) # Counting the frequency of startups based on product or service for Bangalore

freq1

barplot.default(freq1,main = "How many startups in Bangalore are product or service based",xlab = "product or service based",ylab ="Frequency", col = rainbow(length(freq1))) # Plotting the bar chart for the frequency of startups based on product or service for Bangalore

##Plotting bar chart for Mumbai

M=df[df$cities=="Mumbai",] # Filtering the data for startups in Mumbai

freq1=table(M$Type) # Counting the frequency of startups based on product or service for Mumbai

freq1

barplot.default(freq1,main = "How many startups in Mumbai are product or service based",xlab = "product or service based",ylab ="Frequency", col = rainbow(length(freq1))) # Plotting the bar chart for the frequency of startups based on product or service for Mumbai

##Plotting bar chart for Gurugram

G=df[df$cities=="Gurugram",] # Filtering the data for startups in Gurugram

freq1=table(G$Type) # Counting the frequency of startups based on product or service for Gurugram *freq1*

barplot.default(freq1,main = "How many startups in Gurugram are product or service based",xlab = "product or service based",ylab ="Frequency", col = rainbow(length(freq1))) # Plotting the bar chart for the frequency of startups based on product or service for Gurugram

##Plotting bar chart for New Delhi

D=df[df$cities=="New Delhi",] # Filtering the data for startups in New Delhi

freq1=table(D$Type) # Counting the frequency of startups based on product or service for New Delhi

freq1

barplot.default(freq1,main = "How many startups in New Delhi are product or service based",xlab = "product or service based",ylab ="Frequency", col = rainbow(length(freq1))) # Plotting the bar chart for the frequency of startups based on product or service for New Delhi

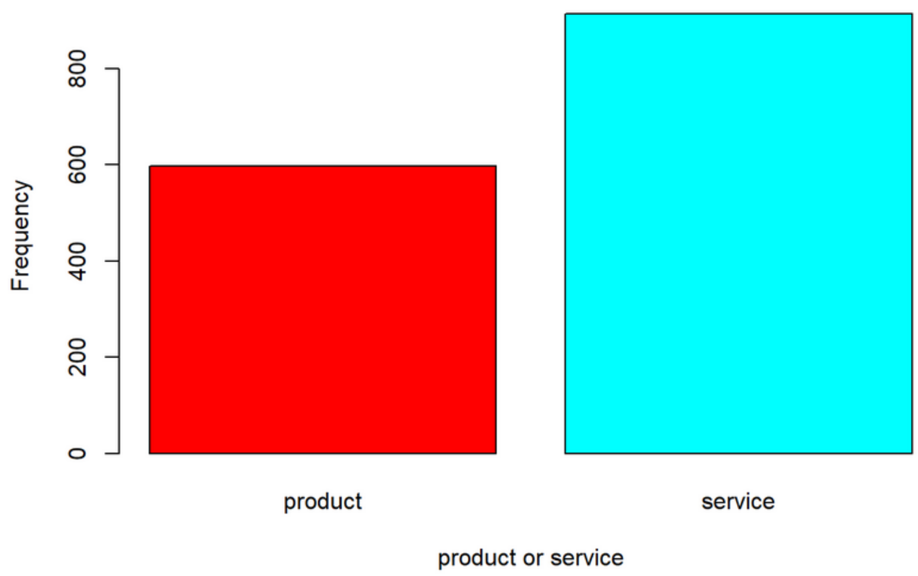

Fig. 5.6: Number of Product and Service based startups

Majority of startups in India have opted for the Service type of Business model with about 896 startups in Service type and about 594 startups have opted to work in the Product type.

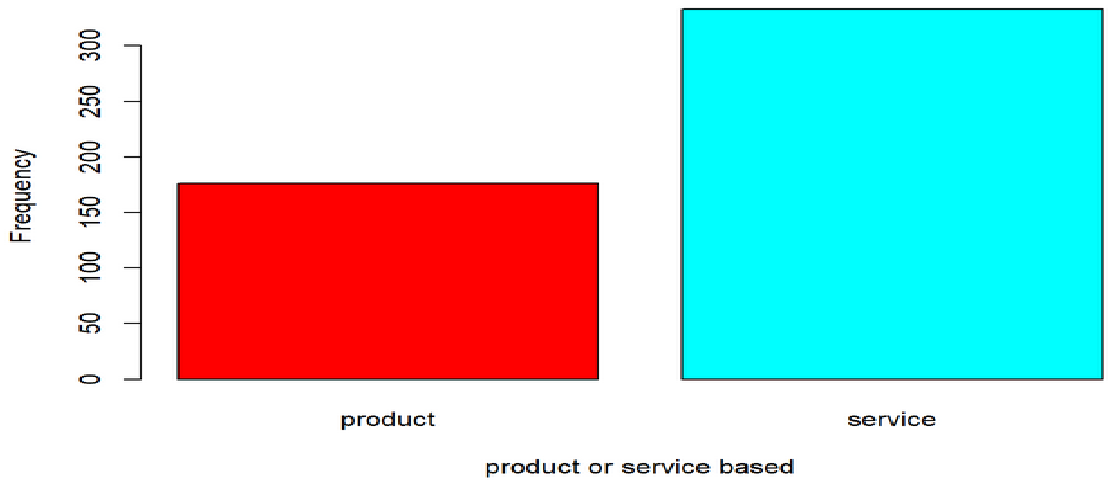

Fig. 5.7: Number of Product and service based startups in Bangalore

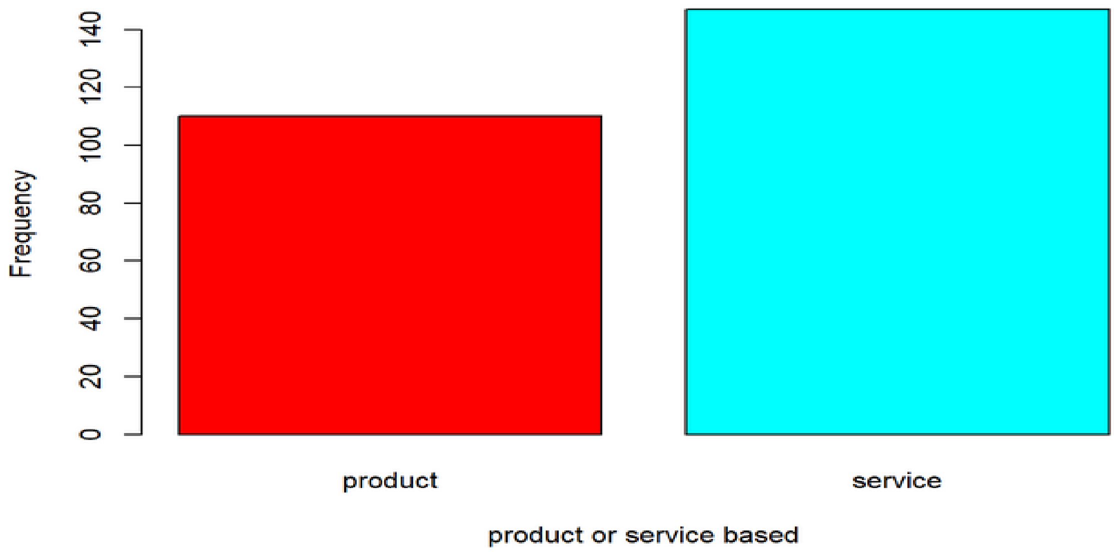

Fig. 5.8: Number of Product and service based startups in Mumbai

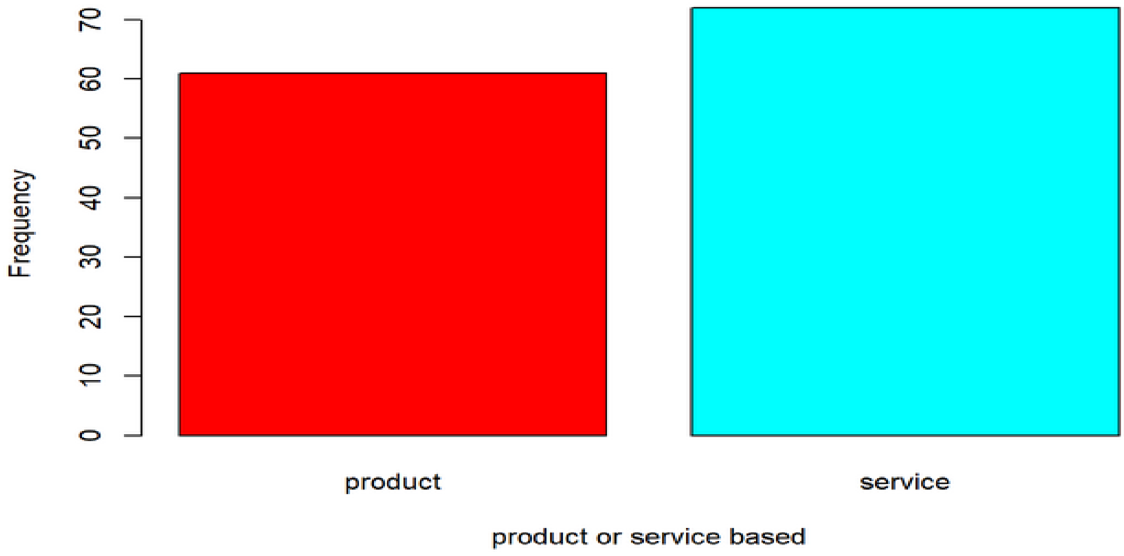

Fig. 5.9: Number of Product and service based startups in Gurugram

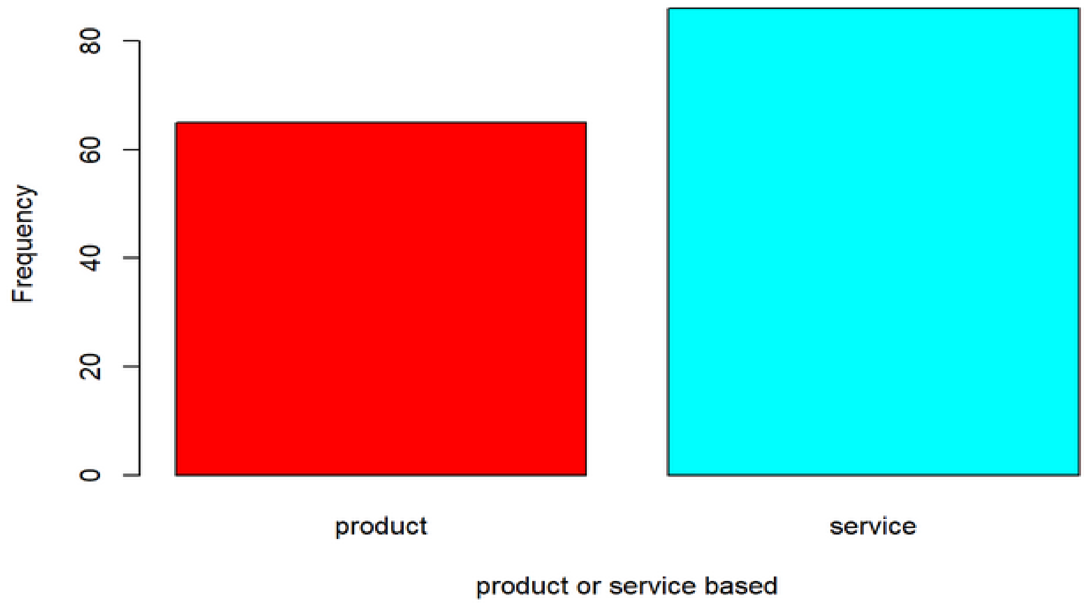

Fig. 5.10: Number of Product and service based startups in New Delhi

Number of startups over the years:

 # To visually represent the trend of number of startups founded over the years. The plot shows the trend of the number of startups founded over the past 20 years, with the blue line representing the trend.

 # The input is a table of frequencies (numeric) of unique values in Founded column (numeric).

 # Line chart showing the trend of number of startups over the years

freq2 = table(df$Founded) # count number of startups founded in each year *freq2*

plot(tail(freq2, 20), type = "l", xlab = "Year", ylab = "Count", main = "Number of startups over the years", col='blue') # plot the trend of number of startups over the years

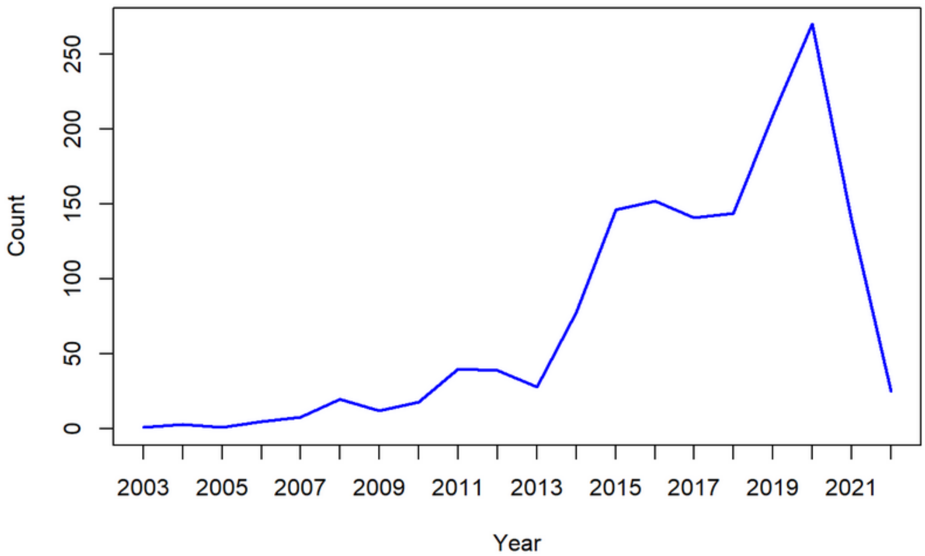

Fig. 5.11: Number of Startups over the years

Startups are critical to the economic development of a country. They drive innovation, create jobs, and stimulate growth across multiple industries. With new technologies, products, and services, startups have the potential to disrupt traditional industries, creating new markets and opportunities. Startups can contribute to the growth of other businesses and industries by attracting investment and talent. Furthermore, by providing solutions and innovations that have a positive impact, startups can help to address social and environmental challenges. Overall, startups are critical for fostering a dynamic and competitive business environment and driving a country's economic growth and development.

In Fig. 5.11 From 2003 onwards, startups were progressively developed. Since 2013, the number of startups has suddenly increased, and in 2020, the number of startups established reached its greatest point to date (about more than 250).

Number of employees in startups:

 # To visually represent the frequency of different employee count labels in the startup's dataset. The result expected will give us an understanding of the strength among startups in recent years.

 # The input is a table of frequencies (numeric) of assigned labels in Employee Count column.

 # Employee rate visualization

freq3 = table(df$Employee.Count) # count the frequency of different employee count labels *barplot.default(freq3, main = "Number of employees in startups", xlab = "Employee labels", ylab = "Frequency", col = rainbow(length(freq3))) # plot a bar chart showing the frequency of different employee count labels*

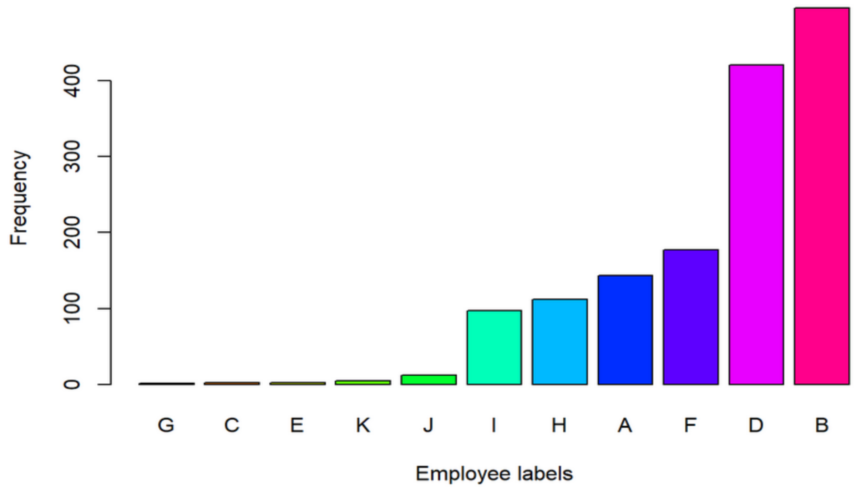

Fig. 5.12: Number of Employees in startups

Employees are essential to the success of startups because they carry out the founder's vision and create the company's goods, services, and culture. They are frequently required to put in long hours, wear many hats, and be versatile and flexible. Employees in startups are frequently given a great deal of autonomy and urged to take responsibility for their work as well as to contribute to the expansion and success of the business.

From Fig. 5.12 it is observed that about more than 400+ startups have employee count in the range of (11-50).

Top 5 cities with highest number of startups:

 # To analyze the top 5 cities with the highest number of startups in India. It provides valuable insights into the country's startup ecosystem. By visualizing the frequency and percentage of startups in top 5 cities using a bar chart and a pie chart, the analysis helps us understand which cities are leading the charge in terms of entrepreneurial activity and innovation.

 # The input is a table of frequencies (numeric) of unique values in Cities column (character).

 # Analysis of top 5 cities

freq = table(data$cities) # count the frequency of different cities where startups are located

freq = sort(freq)

freq = tail(freq, 5) # consider the top 5 cities with the highest number of startups

barplot(freq, main = "Top 5 cities with highest number of startups", xlab = "Cities", ylab = "Frequency", col = rainbow(length(freq))) # plot a bar chart showing the top 5 cities with the highest number of startups *percentages <- round(100 * prop.table(freq))*

labels <- paste(names(freq), "(", percentages, "%)", sep = "")

pie(freq, labels = labels, col = rainbow(length(freq)), main = "Top 5 cities with highest number of startups") # plot a pie chart showing the top 5 cities with the highest number of startups

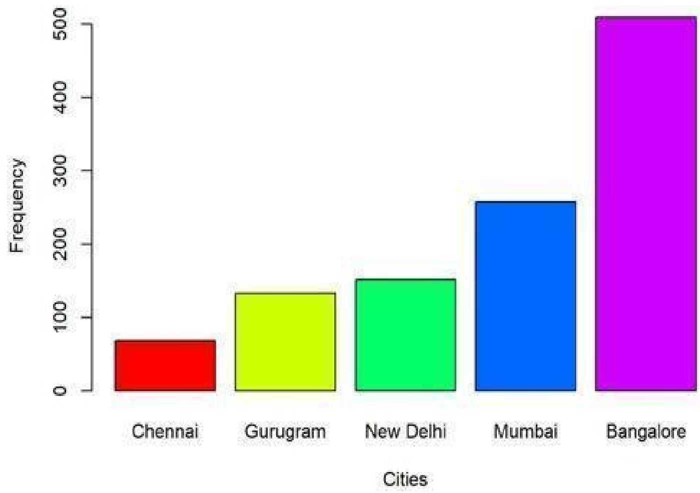

Fig. 5.13: Top 5 cities with highest number of startups

By giving access to a varied talent pool, a community of like-minded business people, and tools like incubators, accelerators, and co-working spaces, cities play a critical role in assisting businesses. Cities also provide businesses with access to resources like funding, networking opportunities, and possible clients, which can help them develop and expand. Cities can also offer a dynamic and lively environment that encourages and inspires businesspeople to follow their dreams and take chances.

It has been seen in Fig. 5.13 that Bangalore has produced the highest number of startups so far in the country, hence playing a major role in the Development of India.

Top 5 Sectors in top 5 cities:

 # To visualize the distribution of top 5 startup sectors across the top 5 cities. It can help identify which of the top sectors are more prominent in certain cities and how they compare to other cities.

 # Stacked bar chart showing overall top 5 sectors in the top 5 cities

library(ggplot2)

Get the top 5 Headquarters

top_headquarters <- head(sort(table(data$cities), decreasing = TRUE), 5) # count the frequency of different cities and get the top 5

Get the top 5 Sectors

top_sectors <- head(sort(table(data$Sector), decreasing = TRUE), 5) # count the frequency of different sectors and get the top 5

```
## Create an empty data frame to store the results
result_df <- data.frame(Cities = character(),
                        Sector = character(),
                        Percentage = numeric())
## For each of the top 5 Headquarters and top 5 Sectors, calculate the percentage
for (hq in names(top_headquarters)) {
for (sector in names(top_sectors)) {
## Calculate the percentage of startups in the current headquarters and sector combination
percentage <- 100 * sum(data$cities == hq & data$Sector == sector) / top_headquarters[hq]
 ## Add the results to the data frame
result_df <- rbind(result_df, data.frame(Cities = hq, Sector = sector, Percentage = percentage))}}
## Create a stacked bar plot using ggplot2
ggplot(result_df, aes(x = Cities, y = Percentage, fill = Sector)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 5 Sectors in Top 5 Cities", x = "Cities", y = "Percentage") +
  scale_fill_brewer(palette = "Paired") +
  theme(legend.position = "bottom")
```

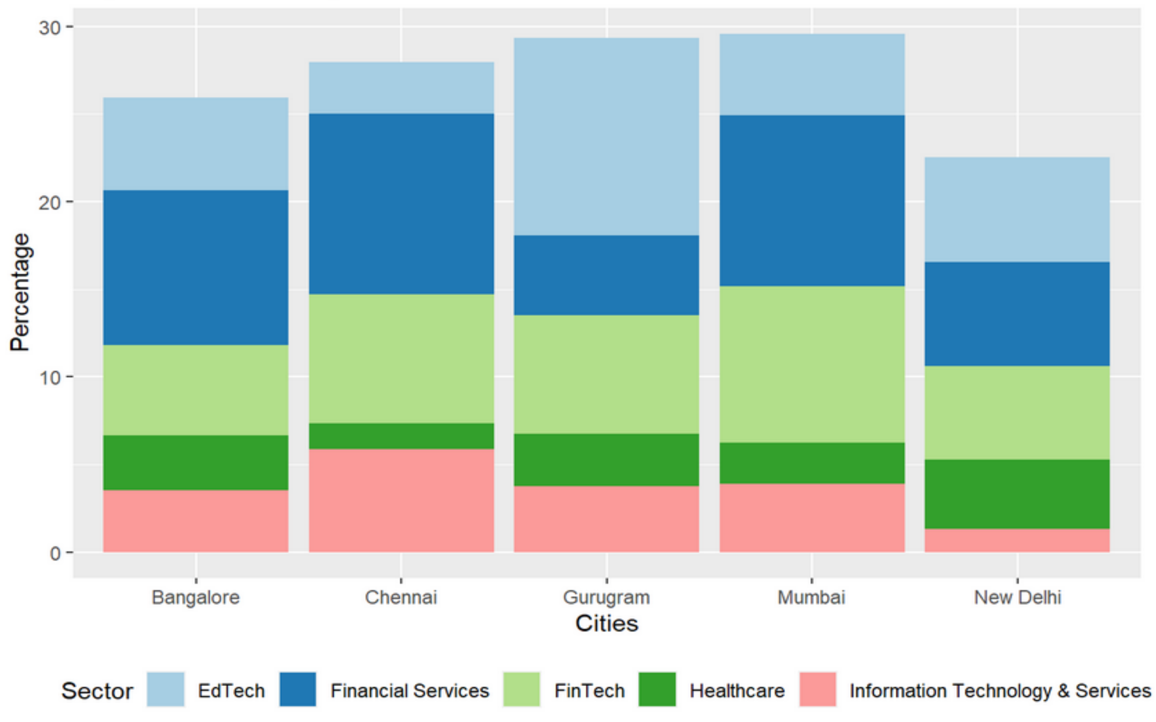

Fig. 5.14: Top 5 Sectors in Top 5 Cities

Startups can be found in a variety of industries, including technology, healthcare, finance, and education. Healthcare companies work to enhance the caliber and availability of healthcare services, whereas technology startups concentrate on creating cutting-edge software and hardware solutions. While education entrepreneurs use technology to revolutionize how people learn and teach, finance startups utilize it to challenge established financial services and offer substitutes. Startups are also widely used in other industries, such as e-commerce, renewable energy, and transportation.

From Fig. 5.14, Top 5 sectors that have emerged are EdTech, Financial Services, FinTech, Healthcare and Information Technology and Services.

Companies/Brands by month wise funding:

 # To visualize the funding months across startups. Provides insights into the seasonality of funding for startups.

 # The input is a table of frequency (numeric) of Month column (numeric).

```
# Aggregate the counts by month
counts <- table(data$Month)
# Convert the counts to a data frame
df <- data.frame(Month = names(counts), Count = as.numeric(counts))
# Plot the data in a bar chart
ggplot(df, aes(x = Month, y = Count)) +
  geom_bar(stat = "identity") +
  xlab("Month") +
  ylab("Count") +
  ggtitle("Companies/Brands by Month")
```

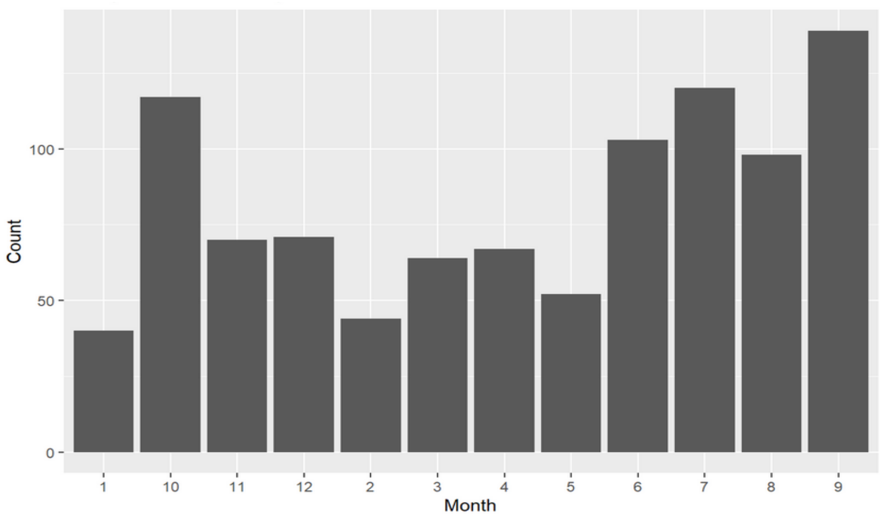

Fig. 5.15: Month Wise Funding

When examining trends and patterns across time, such as the regularity of fundraising rounds or the seasonality of product introductions, the Months can be helpful. Analysts can learn more about the timing of significant events and how they might be influenced by outside variables, such as market conditions or industry trends, by looking at the "Month" column.

From the analysis through the Data collected in Fig. 5.15, most of the startup companies were in their 9th month of funding round.

How Many startups have entered the stock market:

```
# To provide insights into the distribution of startups that have had IPOs versus those that have not. The frequency table shows the number of startups with IPOs and without IPOs.

# The input is a table of frequencies (numeric) of unique values in IPO column (numeric).

# Create a frequency table for the IPO column

freq1=table(data$IPO)

# Sort the table in ascending order

freq1=sort(freq1)

# Display the frequency table

freq1

# Create a bar plot of the top 5 frequencies, with specified properties

barplot(tail(freq1,5),

        main = "IPO",

        xlab = "Yes or No",

        ylab = "Frequency",

        col = rainbow(length(freq)))

# Create a pie chart of the top 5 frequencies, with specified properties

pie(tail(freq1,5),

    main = "IPO",

    col =c('blue','pink'))
```

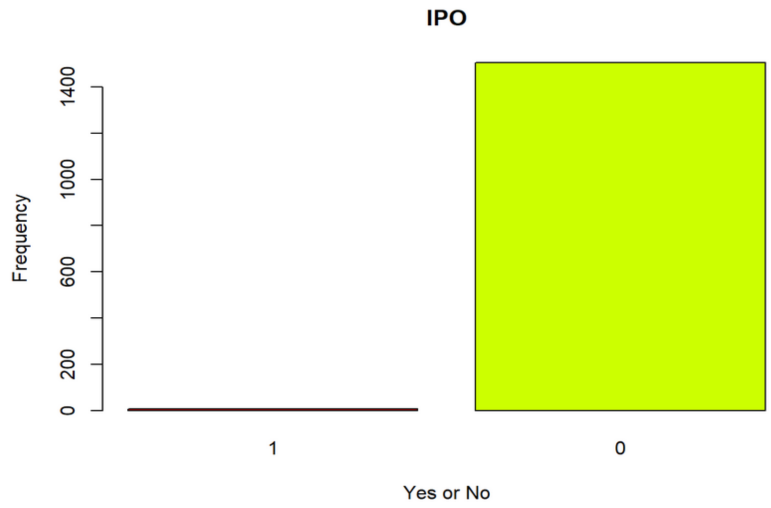

Fig. 5.16: Number of Startups entered the Stock Market (0- No; 1- Yes)

A private firm can go public through an IPO (Initial Public Offering) by selling shares of its stock to the general public. This enables the business to raise money and could boost its reputation and marketability. But IPOs also take a lot of time and money to complete, and they can make the business more exposed to shareholder demands and increased regulatory scrutiny. The individual objectives and circumstances of the startup will determine whether or not to pursue an IPO, and rigorous evaluation of the advantages and disadvantages is required.

It is observed in Fig 5.16 that only 7 Startups have successfully entered the stock market and have launched their IPO.

Who are the top 10 Investors:

```
# To identify the top 10 investors and visualize a pie chart to show their relative frequencies.

# The inputs are the investors (character) and their respective Amount (in dollars) (numeric) values.

data1=data

# Convert the Investment column to numeric

colnames(data1)[colnames(data1) == "Amount(in dollars)"] = "Investment"

data1$Investment <- as.numeric(data1$Investment)

# Create a new data frame to store the individual investments

individual_investments <- data.frame(Investor = character(), Investment = numeric())

# Loop through each row in the original data frame

for (i in 1:nrow(data1)) {

  # Get the list of investors and split them by comma
```

```
  investors <- strsplit(data1$Investors[i], ",")[[1]]
  # Calculate the investment per investor
  investment_per_investor <- data1$Investment[i]
  # Loop through each investor and add a new row to the individual_investments data frame
  for (j in 1:length(investors)) {
    individual_investments <- rbind(individual_investments,
                                    data.frame(Investor = investors[j],
                                               Investment = investment_per_investor))}}
# Write the individual investments data frame to a new CSV file
library(openxlsx)
# Write the data frame to an Excel file
write.xlsx(individual_investments, "individual_investments.xlsx", rowNames = FALSE)
d=read_excel("D:\\individual_investments.xlsx")
library(dplyr)
d$Investor <- na_if(d$Investor, '')
freq4=table(d$Investor)
freq4=sort(freq4)
freq4=tail(freq4,10)
percentages <- round(100 * prop.table(freq4))
labels <- paste(names(freq4), "(", percentages, "%)", sep = "")
pie(freq4, labels = labels, main = "top 10 investors",col =rainbow(length(freq4)))
```

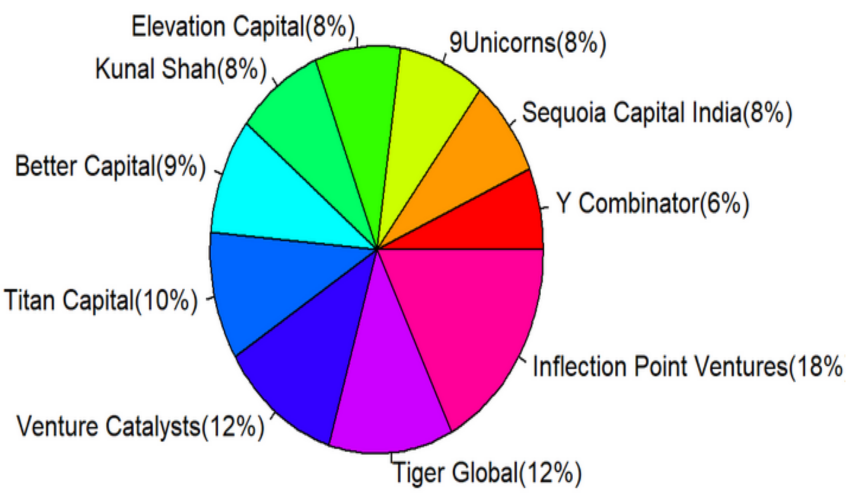

Fig. 5.17: Top 10 Investors

From friends and family to institutional investors and venture capitalists, startups can draw a variety of investors. In order to get a startup off the ground, friends and family generally provide seed money, while angel investors and venture capitalists pay bigger sums of money in exchange for equity in the company. At later phases of growth, institutional investors like hedge funds and private equity firms may also invest in startups. Investors can offer startup companies invaluable experience and direction in addition to financial support by helping them manage the difficulties of a firm's expansion.

Investors that provide investment to startups and growing businesses with strong growth potential are known as venture capitalists (VCs). Unlike traditional lenders, VCs frequently contribute equity financing in return for a portion of the company's ownership. Typically, venture capitalists (VCs) concentrate on early-stage businesses and may offer these businesses not only financial support but also advice and resources to help them develop and thrive. A target sector or location, as well as a minimum investment amount, are common investment criteria for VCs. They may also have specific expectations for the investment's exit strategy, such as an acquisition or an IPO.

With the analysis drawn from the data collected in Fig. 5.17, these are the top 10 investors in startups, with Inflection Point Ventures providing the maximum investment.

Which are the Top 5 Sectors who have the most Valuation:

 # To analyze the top 5 sectors with the highest number of startups in India. It provides valuable insights into the country's startup ecosystem. By visualizing the frequency and percentage of startups in top 5 sectors using a bar chart and a pie chart, the analysis helps us understand which sectors are leading the charge in terms of entrepreneurial activity and innovation.

 # The input is a table of frequencies (numeric) of unique values in Sector column (character).

 # Analysis of the top 5 sectors

freq2 = table(data$Sector) # count the frequency of different sectors

freq2 = sort(freq2)

freq2 = tail(freq2, 5) # consider the top 5 sectors with the highest number of startups

barplot(freq2, main = "Top 5 sectors with highest number of startups", xlab = "Sectors", ylab = "Frequency", col = rainbow(length(freq))) # plot a bar chart showing the top 5 sectors with the highest number of startups

*percentages <- round(100 * prop.table(freq2))*

labels <- paste(names(freq2), "(", percentages, "%)", sep = "")

pie(freq2, labels = labels, col = rainbow(length(freq2)), main = "Top 5 sectors with highest number of startups") # plot a pie chart showing the top 5 sectors with the highest number of startups

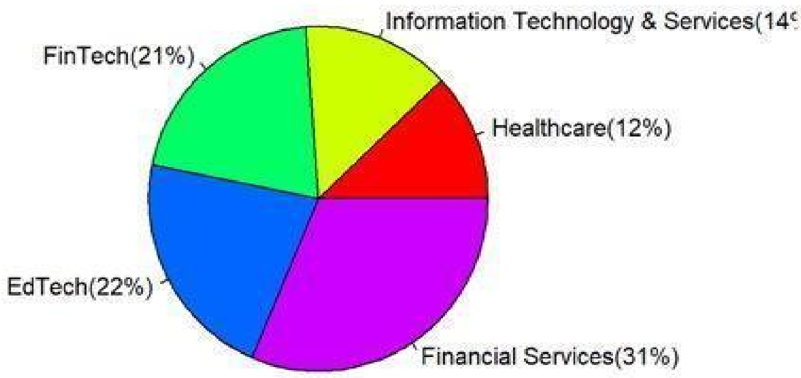

Fig. 5.18: Highest number of startups in top 5 sectors

According to Fig. 5.18, depending on the sector and stage of the business, startup sectors are valued differently. Since they have the potential for rapid growth and scalability, some industries, like technology, are known for having high valuations. Startups may be valued using the discounted cash flow method, the comparable company method, or the venture capital method. The quality of a startup's management team, its intellectual property, the size and potential of its market, its revenue and growth rate, and its competitive environment can all have an impact on its valuation. The market and the willingness of investors to invest in the company ultimately determine valuation, which can be influenced by elements like general economic conditions, industry trends, and investor perceptions of the startup's potential.

Naive Bayes Algorithm Results:

```
# To perform a classification task using a Naive Bayes model to predict the Type column of the dataset based on the Sector column. A confusion matrix is created and the correlation coefficient is calculated.
#Load necessary packages
library(e1071)
library(caTools)
library(rcompanion)
library(caret)
# Subset the data to only include columns 6 and 13, then create a new data frame
df4=data[,c(6,13)]
df4=data.frame(df4)
# Rename the second column to 'type'
colnames(df4)[2]="type"
# Convert the 'Sector' and 'type' columns to factors
df4$Sector=factor(df4$Sector)
```

```
df4$type=factor(df4$type)
# Set a random seed for reproducibility, then split the data into a training and test set
set.seed(123)
trainIndex <- createDataPartition(df4, p = 0.7, list = FALSE)
train <- df4[trainIndex, ]
test <- df4[-trainIndex, ]
# Fit a Naive Bayes model to the training data
model <- naiveBayes(type~ Sector, data = df4)
# Make predictions on the test set using the Naive Bayes model
predictions <- predict(model, newdata = test)
# Calculate the accuracy of the classifier by comparing predicted values to actual values
accuracy <- mean(predictions == test$type)
# Print the accuracy of the classifier
accuracy
# Create a confusion matrix to evaluate the performance of the classifier
confusionMatrix(predictions, test$type)
# Calculate the correlation coefficient (Cramer's V) between 'Sector' and 'type' using the rcompanion package
cramerV(df4$Sector,df4$type)
conf_mat <- confusionMatrix(predictions, test$type)
# Create data frame from confusion matrix
conf_mat_df <- as.data.frame.matrix(conf_mat$table)
# Convert row names to a variable
conf_mat_df$Reference <- rownames(conf_mat_df)
# Reshape data from wide to long format
library(tidyr)
conf_mat_long <- gather(conf_mat_df, key = "Prediction", value = "Frequency", -Reference)
# Plot stacked bar chart
library(ggplot2)
ggplot(conf_mat_long, aes(x = Reference, y = Frequency, fill = Prediction)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("#619CFF", "#FFAA5E")) +
```

labs(title = "Confusion Matrix", x = "Reference", y = "Frequency", fill = "Prediction") +

theme_minimal() +

theme(legend.position = "bottom", axis.text.x = element_text(angle = 90, vjust = 0.5))

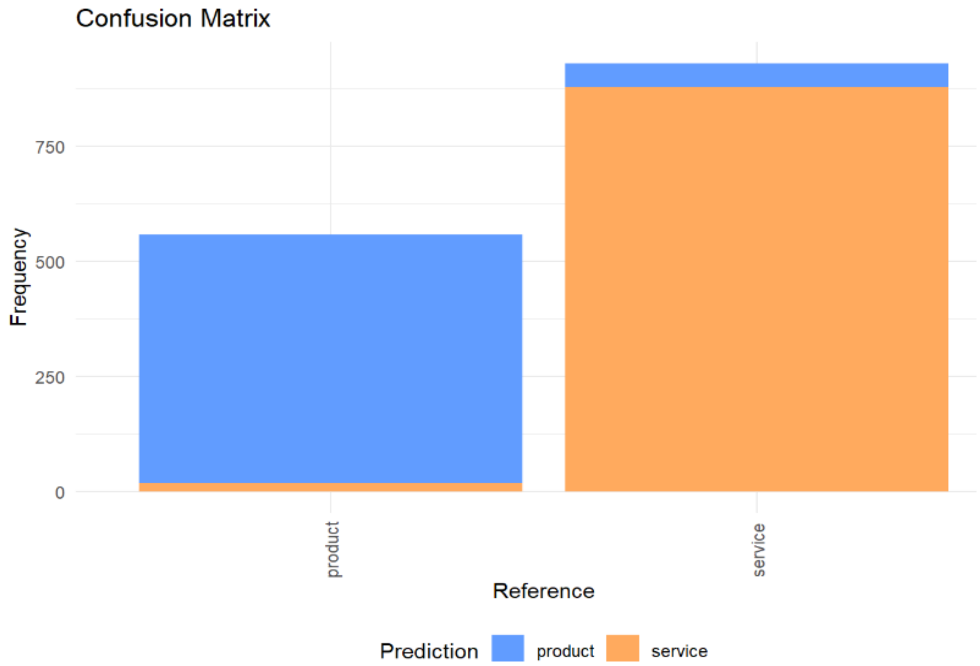

Fig. 5.19: Confusion Matrix for type feature based on Naive Bayes Algorithm

A classification algorithm called Naive Bayes is based on the Bayes theorem, which states that the likelihood of a hypothesis (in this case, a data point belonging to a particular class) given the observed evidence (the data point's features) is proportional to the likelihood of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis.

To put it another way, the algorithm determines the probability of a new data point belonging to each class given its features (in this case, x) and assigns it to the class with the highest probability.

The term "naive" refers to the assumption that the features are conditionally independent given the class. In other words, given the class, the presence or absence of one feature has no effect on the presence or absence of another. This assumption simplifies probability computation by allowing us to calculate the probabilities of each feature independently.

After computation of the algorithm, it is observed in Fig. 5.19 that very few classes in the product type company have been predicted as service and few classes (relatively smaller portion) in the service type company have been predicted as product.

Random Forest Algorithm Results:

 # Random Forest Implementation. We predict the Type (whether it's a product or service) with the help of predictors- Sector. A confusion matrix is built and plotted to get a better understanding and the accuracy of the model is printed.

```
library(dplyr)

library(randomForest)

# Convert variables to numeric where applicable

# Remove non-numeric variables

datarf <- data %>% select(-`Company Name`,-IPO, -cities, -Description, -Founders, -Investors, -Stage, -Month,-`Amount(in dollars)`,-Founded,-`Employee Count`)

datarf$Type <- as.factor(datarf$Type)

# Remove rows with missing values

datarf <- na.omit(datarf)

# Split the data into training and testing sets

set.seed(123)

trainIndex <- createDataPartition(datarf$Type, p = 0.7, list = FALSE)

trainData <- datarf[trainIndex, ]

testData <- datarf[-trainIndex, ]

# Train the random forest model

model <- randomForest(Type ~ ., data = trainData, ntree = 100, mtry = 2, type = "class")

print(model)

# Make predictions on the test data

predictions <- predict(model, testData)

# Compute the confusion matrix

conf_mat <- table(predictions, testData$Type)

conf_mat_pct <- prop.table(conf_mat, margin = 1)

# Print the confusion matrix

print(conf_mat_pct)

accuracy <- sum(diag(conf_mat)) / sum(conf_mat)

accuracy

library(ggplot2)

# Convert the confusion matrix to a data frame

conf_mat_df <- as.data.frame.matrix(conf_mat_pct)
```

```
conf_mat_df$predicted <- rownames(conf_mat_df)
conf_mat_df <- tidyr::gather(conf_mat_df, actual, value, -predicted)
# Plot the stacked bar chart
ggplot(conf_mat_df, aes(x = predicted, y = value, fill = actual)) +
  geom_col(position = "stack") +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(x = "Predicted Type", y = "Percentage", title = "Confusion Matrix")
```

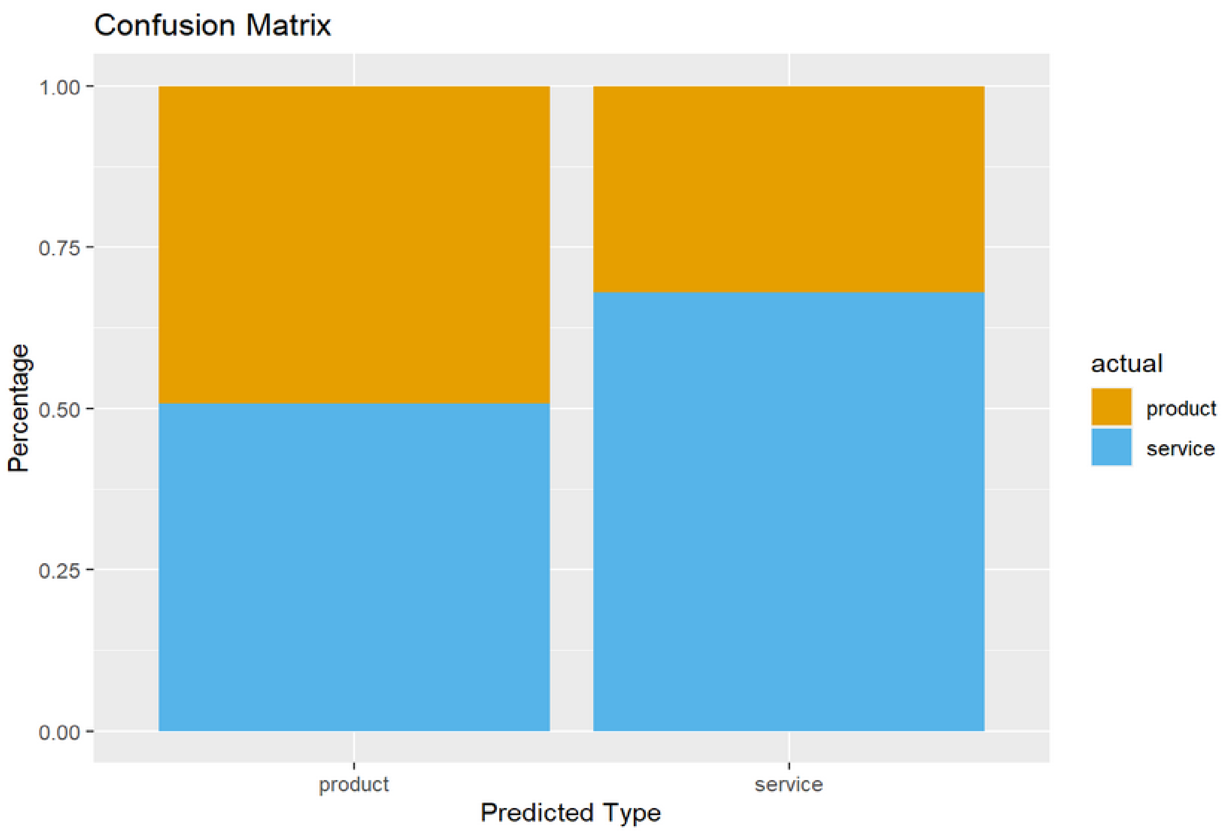

Fig. 5.20: Confusion matrix for type feature based on Random Forest Algorithm

Random Forest is a popular machine learning algorithm used for classification and regression tasks. In the context of the given problem, the random forest algorithm is used to predict the `Type` of a company based on the values of various features or columns in the data set, including the `Sector` column.

In a random forest model, a decision tree is constructed for each subset of the data, using a random sample of features. This randomness helps to reduce overfitting and improve the generalization of the model. The trees in the forest are then combined to make a final prediction for a new observation. In the case of predicting the `Type` of a company, the random forest algorithm will use the values in the `Sector` to build decision trees. Each tree will make its own prediction for the `Type` of the company, and the final prediction will be the mode of all the individual tree predictions. The model will be trained on a portion of the available data, and the remaining portion will be used to evaluate the accuracy of the model. The accuracy is calculated as the percentage of correctly predicted `Type` labels compared to the total number of labels in the test set.

On computing the Random Forest algorithm for classification in Fig. 5.20, it is observed that nearly 50% of the classes in the product type have been predicted as service and around 30% if the companies in service type have been predicted as product.

Comparing both these algorithms based on their predictions, it is inferred that Naive Bayes is the better algorithm as it has very less classes misclassified, even though there are some exceptions in the type of the company.

Decision Tree Algorithm Results:

```
# Load the dataset
df <- data
# Preprocess the data
df$Sector <- as.factor(df$Sector)
df$Type <- as.factor(df$Type)
# Split the data into train and test sets
set.seed(123)
trainIndex <- createDataPartition(df$Type, p = .8, list = FALSE)
train <- df[trainIndex,]
test <- df[-trainIndex,]
# Fit a Naive Bayes model to the training data
model <- train(Type ~ Sector, data = train, method = "rpart")
print(model)
# Make predictions on the test data
predictions <- predict(model, testData)
# Compute the confusion matrix
conf_mat <- table(predictions, testData$Type)
conf_mat_pct <- prop.table(conf_mat, margin = 1)
# Print the confusion matrix
print(conf_mat_pct)
accuracy <- sum(diag(conf_mat)) / sum(conf_mat)
accuracy
```

```
library(ggplot2)
# Convert the confusion matrix to a data frame
conf_mat_df <- as.data.frame.matrix(conf_mat_pct)
conf_mat_df$predicted <- rownames(conf_mat_df)
conf_mat_df <- tidyr::gather(conf_mat_df, actual, value, -predicted)
# Plot the stacked bar chart
ggplot(conf_mat_df, aes(x = predicted, y = value, fill = actual)) +
  geom_col(position = "stack") +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(x = "Predicted Type", y = "Percentage", title = "Confusion Matrix")
```

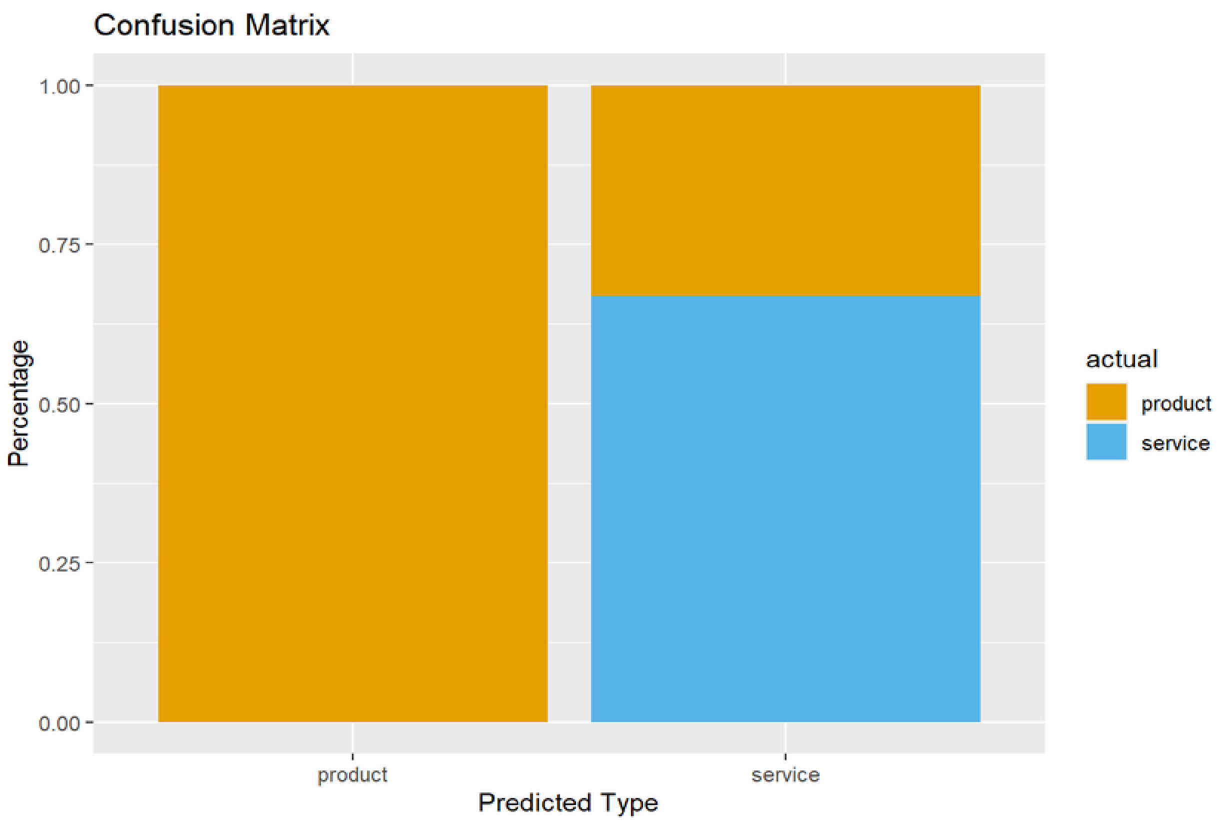

Fig. 5.21: Confusion matrix for type feature based on Decision Tree Algorithm

An example of a supervised learning algorithm is a decision tree, which is typically applied to classification issues in machine learning. It is a visual representation of every outcome that could result from a choice under specific circumstances. There are nodes, branches, and leaves in a decision tree. The nodes stand in for the decision-making points, the branches for the decision-making results, and the leaves for the classification or final decision.

The root node is the first node in the decision tree algorithm, and it is where a feature is evaluated. The data is divided into two or more branches using the feature that maximizes information gain. Up until all the data points are classified, the procedure is repeated for each additional node.

Decision tree algorithms come in two varieties: ID3 (Iterative Dichotomiser 3) and C4.5. Entropy and information gain are used by the straightforward ID3 algorithm to assess the features. C4.5 is a sophisticated algorithm that can handle categorical and continuous features and uses gain ratio rather than information gain. Decision tree classifiers have the benefit of being simple to understand and visualize. They can handle missing data as well as continuous and categorical data. If the tree is too deep or too complex, they tend to overfit the data. Pruning methods can be used on the tree to avoid overfitting.

On computing the Decision Tree algorithm for classification (in Fig 5.21) it is observed that no classes in the product type have been misclassified and around 36% of the companies in service type have been predicted as product. The Model gave an accuracy of about 68 %.

CONCLUSION & FUTURE SCOPE

The future of startups in India looks promising as the country has a growing economy, a large and young population, and a thriving startup ecosystem. The rise of digital technology and e-commerce has created new opportunities for startups in areas such as Fintech, Edtech, Healthnet, and E-commerce.With the right guidance and perfect expansion plan, many startups can grow much bigger and be sources of employment to large number of people in India.

As Data Analytics technology continues to evolve, startup insights and visualization tools can incorporate advanced techniques to provide more accurate insights and predictions. Interactive visualization dashboards can be created to provide users with a more engaging and intuitive way to explore the data. A startup insights and visualization project can be made accessible on mobile platforms. This will provide users with real-time updates and insights. While the startup ecosystem is a significant area for potential growth, there are many other industries that could benefit from insights and visualization tools.

REFERENCES

[1]Verma, Piyush Anand, and Vikas Singhal. "Data Analysis of startups investments and funding trends in India." (2018).

[2] Rompho, N. (2018), "Operational performance measures for startups", *Measuring Business Excellence*, Vol. 22 No. 1, pp. 31-41.

[3] Vasilescu, Laura. (2009). Business angels: Potential financial engines for start-ups. Ekonomska Istrazivanja. 22. 86-98.

[4]. Kamaluddin, Fakih Amrin, and Kala Seetharam Sridhar. *Indian Startup Ecosystem: Analyzing Investment Concentration and Performance of Government Programmes*. No. 514. 2021.

[5] 2015 in review: India's funding pattern in 5 graphs and what you can learn from it- https://www.techinasia.com/2015-in-review-startups-funding-in-india-5-graphs

[6]Narayan, Madhusudan, Birajit Mohanty, and Mahesh Kumar. "Growth pattern and trends in start-up funding in India." *International Journal of Innovative Technology and Exploring Engineering* 8, no. 12 (2019): 3721-3724.

[7] Rao, S. V. R., & Kumar, L. (2016). Role of Angel Investor in Indian Startup Ecosystem. *FIIB Business Review*, *5*(1), 3–14.

[8] Grover, Parul. "Sustainability Of Startups In India." *Think India Journal* 22, no. 14 (2019): 10044-10048.

[9] The Indiastartup report - http://userfilescdn.yourstory.com/ag4zzlb1-IndianStartupEcosystem_June-2018_Vol_1.pdf

[10] Kshetri, Nir, and Nir Kshetri. "Fostering startup ecosystems in India." *Asian Research Policy* 7, no. 1 (2016): 94-103.

[11] Dataset used- https://www.kaggle.com/datasets/ramjasmaurya/indian-startupsin-2021

[12] Startup Media- https://startuptalky.com/indian-startups-funding-investors-data-2022/#Indian_Startups_Funding_Data_

[13] Employee count- https://www.linkedin.com/home

[14] Startup Details- https://www.crunchbase.com/

[15] Bombay Stock Exchange – https://www.bseindia.com

[16] NationalStock Exchange – https://www.nseindia.com/

[17] IndustryInformation-
https://www.corporateinformation.com/

[18] Company Check is a registered information and data resource technology platform-
https://www.thecompanycheck.com/