Exploratory Data Analysis of Indian Startup Funding

CH. VASUNDHARA, EEPENAGANDLA SRIKANTH
Assistant Professor, Department Of MCA, MCA Final Semester,
Master of Computer Applications,
Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India.

Abstract:

The Indian startup ecosystem has witnessed rapid growth over the past decade, making it crucial to understand funding patterns and investment behaviour. This project aims to analyse and visualize startup funding trends in India through Exploratory Data Analysis (EDA). A comprehensive dataset was cleaned and pre-processed to correct inconsistencies in dates, city names, investment types, and funding amounts. Missing and undisclosed values were handled appropriately to ensure data quality. The analysis uncovers key insights such as the rise and fall of funding activity over time, the dominance of cities like Bangalore and Mumbai, and the popularity of sectors like E-Commerce, Technology, and Healthcare. It identifies top investors and explores the most common funding types, including Seed and Private Equity. A geographic visualization highlights major startup hubs across India. Interactive charts and maps were built using Streamlit, matplotlib, seaborn, plotly, and folium, making the dashboard user-friendly and insightful. This EDA serves as a powerful tool for stakeholders to explore the dynamics of startup funding, guiding data-driven decisions for entrepreneurs, investors, and policymakers.

Index Terms: Exploratory Data Analysis (EDA), Data Cleaning, Data Preprocessing, Missing Value Imputation, Feature Engineering, Date Normalization, Streamlit, Plotly

1.Introduction:

This project presents an interactive data visualization dashboard built using Stream lit to analyse startup funding trends in India. The data is sourced from a CSV file containing startup investment details such as funding amount, city, industry, investors, and more. Comprehensive data cleaning steps were applied to correct date formats, handle missing values, and normalize inconsistent entries. The dashboard enables users to explore funding patterns across time, cities, industries, and investment types. Users can view monthly funding trends through line charts to identify fluctuations and growth periods. Bar plots highlight the top 10 most common investment types and the industries receiving the most deals. The Cities section identifies startup hubs and visualizes them on an interactive Folium map ofIndia. An Investors section showcases the top backers based on the number of deals. A Word Cloud illustrates the popularity of various funding types, providing quick insights into funding diversity. The Overview tab allows users to explore raw data and understand the structure of the dataset. Data visualizations use libraries like Matplotlib, Seaborn, and Plotly for rich graphical representation. The code is modular, leveraging caching (@st.cache_data) for performance optimization in data loading. It includes UI enhancements such as sidebars and tabs for easy navigation through different analysis sections. This project is valuable for entrepreneurs, investors, and researchers interested in India's startup ecosystem. It demonstrates how Python, data analysis, and Streamlit can come together to build practical business intelligence tools.

1.1. Existing system

The existing system is a **Streamlit-based interactive dashboard** designed for analyzing Indian startup funding trends using a cleaned dataset (startup_funding.csv). The backend performs extensive data preprocessing including fixing inconsistent dates, standardizing column names, and cleaning monetary values by replacing missing or non-numeric entries with meaningful estimates. The dashboard features multiple tabs such as

Overview, Trends, Funding Types, Industries, Cities, Investors, and Word Cloud, offering insights into investment patterns. Users can explore temporal funding trends, city-wise startup distributions, and top investors. Visualizations are created using **Matplotlib, Seaborn, Plotly**, and **Folium maps** for geographic insights. Key statistics like **top funding types**, **industry-wise trends**, and **city funding rankings** are prominently featured. The data is cached for performance, and interactive plots allow users to drill into specific metrics. Overall, the system aims to simplify complex startup funding data through a user-friendly interface and rich visual analytics.

1.1.1. Challenges

Data Quality Issues:

❖ The dataset has many inconsistencies such as malformed dates, 'undisclosed' funding values, and noisy strings, which require extensive preprocessing to ensure accurate analysis.

Missing Values Handling

❖ Several funding amounts and city names are missing or inconsistent, demanding careful imputation strategies (e.g., replacing missing funding with mean) to avoid misleading results.

Dynamic Visual Mapping:

❖ Integrating Folium maps with accurate city coordinates and startup counts introduces complexity, especially when dealing with cities having similar names or ambiguous locations.

Categorical Normalization:

❖ There’s a need to unify similar investment types and city names (e.g., "Seed\nFunding" vs "Seed Funding") for meaningful aggregations and visual clarity

Streamlit Limitations

❖ While Streamlit offers simplicity, creating multi-tab dashboards with complex visualizations (like maps and word clouds) requires careful layout handling to maintain performance and user experience

1.2.Proposed system:

The proposed system is a data-driven web application that performs Exploratory Data Analysis (EDA) on Indian startup funding data. It starts by loading and cleaning the dataset to correct formats, handle missing values, and normalize entries for consistency. Using Python libraries like Pandas, Matplotlib, Seaborn, and Plotly, the system performs descriptive and visual analysis to uncover patterns and trends. A Streamlit-based interactive dashboard is then used to present insights across different views such as industries, cities, investors, and funding types. This system enables users to explore funding dynamics and gain actionable insights into the startup ecosystem.

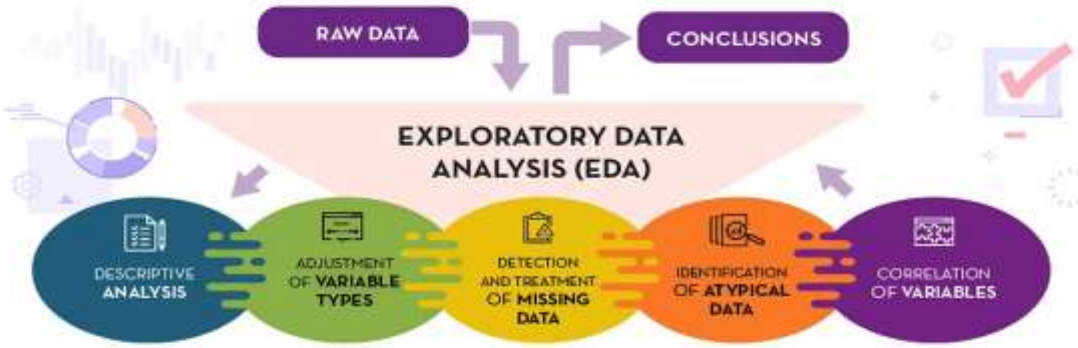

RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

Fig: 1 Proposed Diagram

1.1.1 Advantages:

Interactive Visualization

- ❖ Streamlit and Plotly allow users to explore funding trends, top investors, and city hotspots interactively, improving decision-making.

Data Cleaning Automation

- ❖ The project includes robust preprocessing steps to clean inconsistent date formats, undisclosed amounts, and city/investor names.

Multiple Analytical Views

- ❖ It provides different perspectives like trends over time, industry-wise funding, and investment types, which help uncover hidden patterns.

Geo-Mapping Capability

- ❖ Folium integration enables visual mapping of startup density across Indian cities, making location-based insights more intuitive.

User-Friendly Interface

- ❖ The sidebar navigation and modular layout make it easy for non-technical users to explore the data and insights.

2.1 Architecture:

The architecture begins with **raw data ingestion** (startup_funding.csv) and flows through a **data cleaning pipeline** that handles missing values, standardizes inconsistent entries, and parses dates and numerical fields. Using **Pandas and NumPy**, the data is then explored and transformed to extract key characteristics. **Visualizations** are generated with **Matplotlib, Seaborn, Plotly, and Folium** to understand trends, outliers, relationships, and funding geography. The processed insights are presented through an interactive **Streamlit frontend** with sidebar-based navigation, allowing users to analyse trends by time, investment types, industries, cities, and investors. The architecture concludes with visual communication of findings using graphs, word clouds, and maps to derive meaningful conclusions from the data.

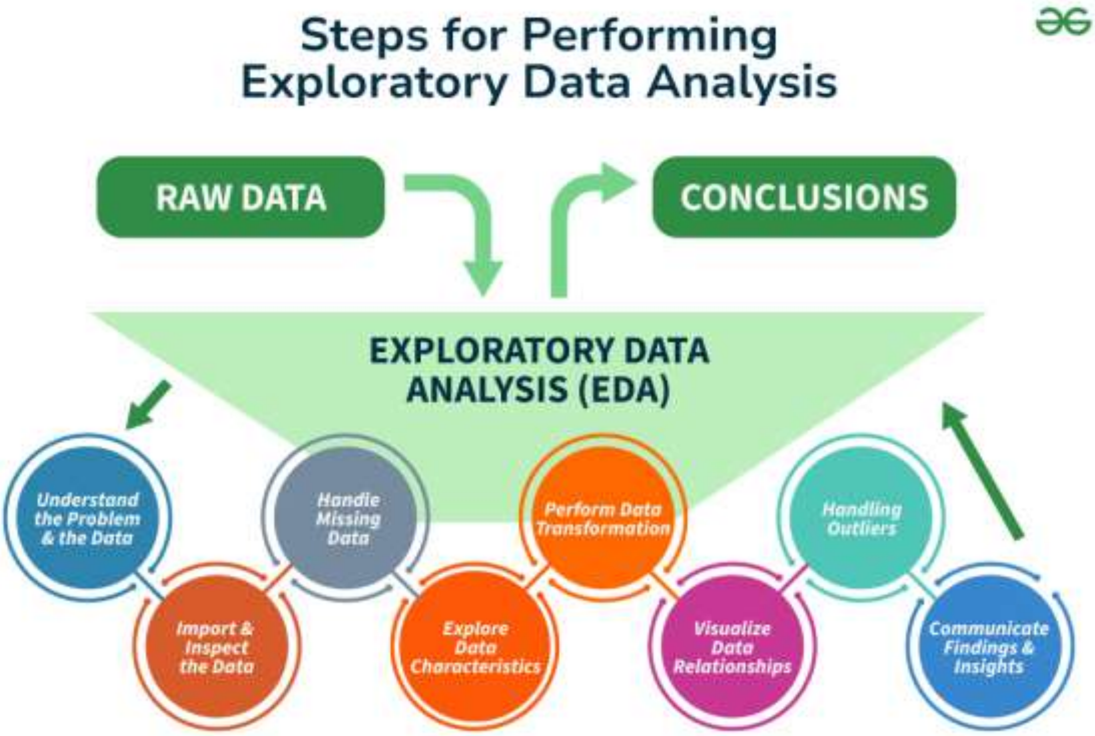

Steps for Performing Exploratory Data Analysis
RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
Understand the Problem & the Data
Import & Inspect the Data
Handle Missing Data
Explore Data Characteristics
Perform Data Transformation
Visualize Data Relationships
Handling Outliers
Communicate Findings & Insights

Fig:2 Architecture

2.2 Algorithm:

The algorithm starts by importing the necessary Python libraries for data analysis, visualization, and UI rendering using Streamlit. It then loads the startup_funding.csv dataset and performs extensive data cleaning, including fixing inconsistent date formats, normalizing text fields (like city names and investment types), and handling missing or malformed funding values. After preprocessing, new time-based features (year and month) are extracted for trend analysis. The cleaned data is visualized in a Streamlit dashboard with multiple tabs such as Overview, Trends, Funding Types, Industries, Cities, Investors, and Word Cloud. Each tab presents insightful visuals: line plots for trends, bar charts for top industries and cities, and word clouds for funding types. A Folium map is included to show geographic startup hotspots. Aggregated statistics like total and average funding over time are calculated using groupby operations. The app dynamically updates based on sidebar selections, allowing interactive exploration. This project offers a comprehensive visual analysis of Indian startup funding patterns over time.

2.3 Techniques:

In this project several techniques were applied for Exploratory Data Analysis (EDA) of Indian startup funding data. Initially, data cleaning techniques were implemented to correct inconsistent date formats, normalize city names (e.g., 'Bengaluru' to 'Bangalore'), and handle missing or malformed funding amounts by replacing them with mean values.

Data transformation techniques were used to extract the year and month from dates and convert funding values to numeric for analysis. Then, **grouping and aggregation** were applied to examine total, average, and count of investments over time and across sectors, using pandas' groupby() and agg() functions. **Visualization techniques** included bar charts, line plots, scatter plots, and word clouds created using matplotlib, seaborn, plotly, and WordCloud, enabling clear insights into funding trends, top investors, and city-wise distributions. **Geospatial mapping** was achieved using folium, showing startup hotspots across major Indian cities with interactive maps.

Additionally, **Streamlit** was used to build an interactive dashboard with tabbed sections for overview, trends, funding types, industries, cities, investors, and word clouds, facilitating dynamic exploration of insights. These combined techniques helped transform raw funding data into a comprehensive and visually engaging analytics platform.

2.4 Tools:

The project uses several powerful tools to perform Exploratory Data Analysis (EDA) and build an interactive dashboard for Indian startup funding insights. Streamlit is used as the core framework to develop the frontend web dashboard, offering sidebar navigation, tabs, and data visualizations. Pandas is the primary data manipulation tool used for cleaning, transforming, and summarizing the startup funding dataset. NumPy supports numerical operations, especially when handling missing values and replacing invalid entries. For visual analysis, the project employs Matplotlib and Seaborn to plot trends, funding types, and industry insights. Plotly Express and Graph Objects enhance interactivity in charts such as scatter plots and time series. Folium and streamlit-folium are integrated to render geographical maps showing startup density across Indian cities. Additionally, the WordCloud library provides a visual representation of investment types by frequency. Warnings are suppressed for clean output using Python's warnings module. The dataset is cached using st.cache_data() for performance optimization. These tools collectively enable comprehensive exploratory data analysis, visual storytelling, and geographic insights in a streamlined user interface.

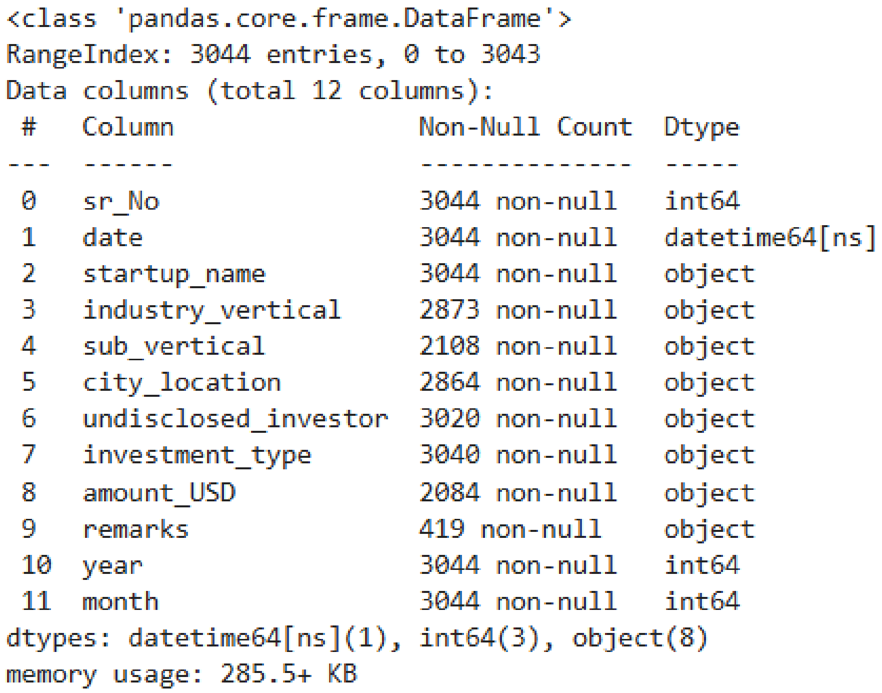
2.5 Methods:

The project follows a systematic methodology to Indian Startup Funding Analysis include several key data analysis and visualization techniques. First, data cleaning and preprocessing was applied to fix inconsistent dates, handle missing values, and normalize columns like city names and investment types. Next, feature extraction was performed by deriving new columns like year and month from the date field. For data visualization, multiple libraries such as Matplotlib, Seaborn, Plotly, and Folium were used to plot trends, top cities, industries, and investors. Additionally, grouping and aggregation techniques using Pandas were utilized to compute metrics like total funding and average investments. The WordCloud method was used for visually representing investment types. The Streamlit framework provided the interactive web interface with tabs for navigation, displaying data tables, charts, and maps. Each section—Overview, Trends, Funding Types, Cities, and Industries—uses specific plotting methods suited for that analysis. Finally, geospatial visualization with Folium plotted the distribution of startups across Indian cities, enhancing the interpretability of location-based insights.

III. METHODOLOGY

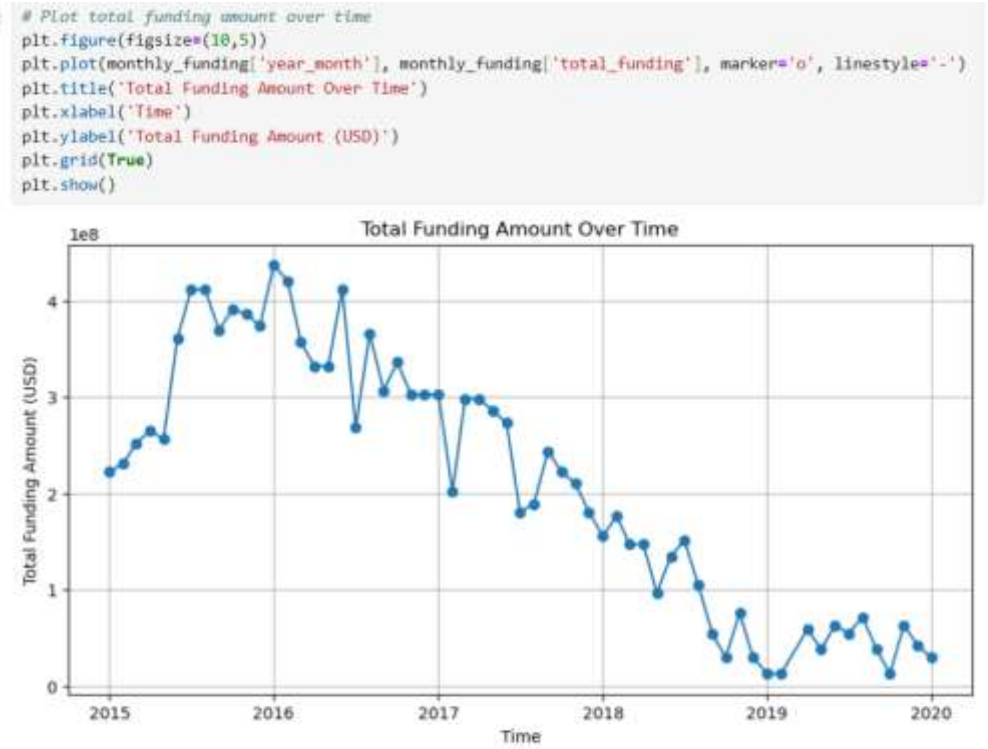
3.1 Input:

This project is an interactive data visualization dashboard built using Streamlit, aimed at analyzing startup funding trends in India. The dataset contains 3,044 records and 12 columns, including fields such as startup name, industry vertical, investment type, amount funded, and location. Extensive preprocessing was applied to clean inconsistent dates, missing values, and irregular text entries. The funding amount column was converted to numeric, with missing or zero values filled using the mean. The data spans multiple years and months, allowing temporal trend analysis. The app includes interactive plots showing funding trends, top cities, industries, and investors. Folium maps highlight city-level startup activity. A WordCloud illustrates the distribution of investment types. This tool enables a comprehensive understanding of India's evolving startup ecosystem.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3044 entries, 0 to 3043
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   sr_No                3044 non-null   int64
 1   date                 3044 non-null   datetime64[ns]
 2   startup_name         3044 non-null   object
 3   industry_vertical    2873 non-null   object
 4   sub_vertical         2108 non-null   object
 5   city_location        2864 non-null   object
 6   undisclosed_investor 3020 non-null   object
 7   investment_type      3040 non-null   object
 8   amount_USD           2084 non-null   object
 9   remarks              419 non-null    object
 10  year                 3044 non-null   int64
 11  month                3044 non-null   int64
dtypes: datetime64[ns](1), int64(3), object(8)
memory usage: 285.5+ KB
```

Fig 1: Dataset Structure and Column Summary of Startup Funding Data



```
# Plot total funding amount over time
plt.figure(figsize=(10,5))
plt.plot(monthly_funding['year_month'], monthly_funding['total_funding'], marker='o', linestyle='-')
plt.title('Total Funding Amount Over Time')
plt.xlabel('Time')
plt.ylabel('Total Funding Amount (USD)')
plt.grid(True)
plt.show()
```

Fig 2: Total Funding Trend in Indian Startups (2015–2020)

The line plot illustrates the monthly trend of total funding received by Indian startups from 2015 to early 2020. Funding volumes peaked around 2016, reflecting a period of heightened investment activity. However, a steady decline followed, with noticeable drops in 2018 and 2019. The data is sourced from a cleaned startup funding dataset and visualized using Matplotlib. This trend highlights changing investor behavior over time within the Indian startup ecosystem.

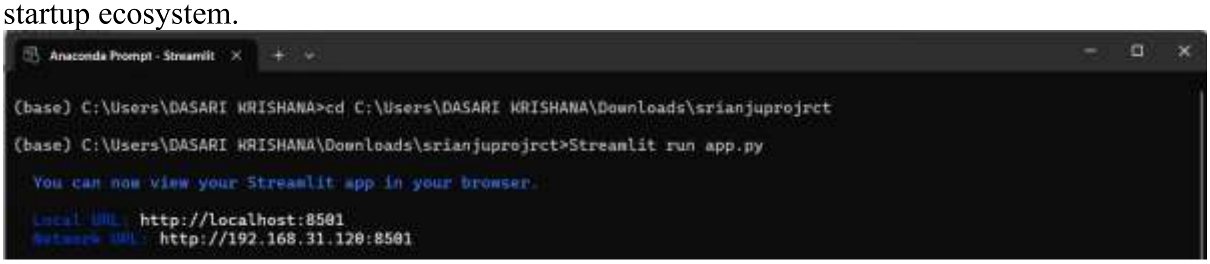
ISSN: 2583-6129
DOI: 10.55041/ISJEM04900

3.2 Method of Process:

The project follows a structured method of process to analyze Indian startup funding data and build an interactive dashboard using Streamlit. It begins with importing the necessary libraries such as Pandas, NumPy, Matplotlib, Seaborn, Plotly, and Folium. The dataset is then loaded and undergoes extensive preprocessing, including renaming columns, fixing malformed date entries, handling missing values, and standardizing categorical entries like cities, investors, and industry verticals. Numerical columns like funding amounts are cleaned by replacing non-numeric entries and imputing missing values with the mean. New features such as year and month are extracted from the date for time-based analysis. The cleaned data is used to generate multiple visualizations categorized into sections such as trends over time, funding types, industries, cities, investors, and word clouds. Streamlit's sidebar navigation enables switching between these sections. Bar plots, line charts, and a Folium map are employed for dynamic and geographical insights. The final dashboard allows users to explore top-funded sectors, investor activity, and city-based funding patterns interactively.

3.3 Output:

The Streamlit application delivers a dynamic dashboard to analyze startup funding trends in India. The data, sourced from a CSV file, undergoes significant preprocessing including date correction, amount standardization, and normalization of city names and funding types. The dashboard is divided into insightful sections—starting with an Overview showcasing sample data. The Trends tab visualizes total funding over time, revealing notable peaks in 2019 and dips during late 2018. Under Funding Types, users explore top funding categories, where Seed and Venture Capital dominate. The Industries tab highlights sectors like E-Commerce, Technology, and Healthcare as top recipients of investment. Cities analysis shows Bangalore, Mumbai, and Delhi as startup hubs, with a folium-based interactive map marking startup concentrations. The Investors section lists leading funders, including anonymous ones grouped as “Undisclosed Investors.” A compelling Word Cloud offers a creative visual of funding types. The data pipeline integrates pandas for processing, matplotlib/seaborn for plotting, plotly for interactivity, and folium for mapping—resulting in an informative and user-friendly dashboard that supports both investors and analysts in understanding India’s startup ecosystem.



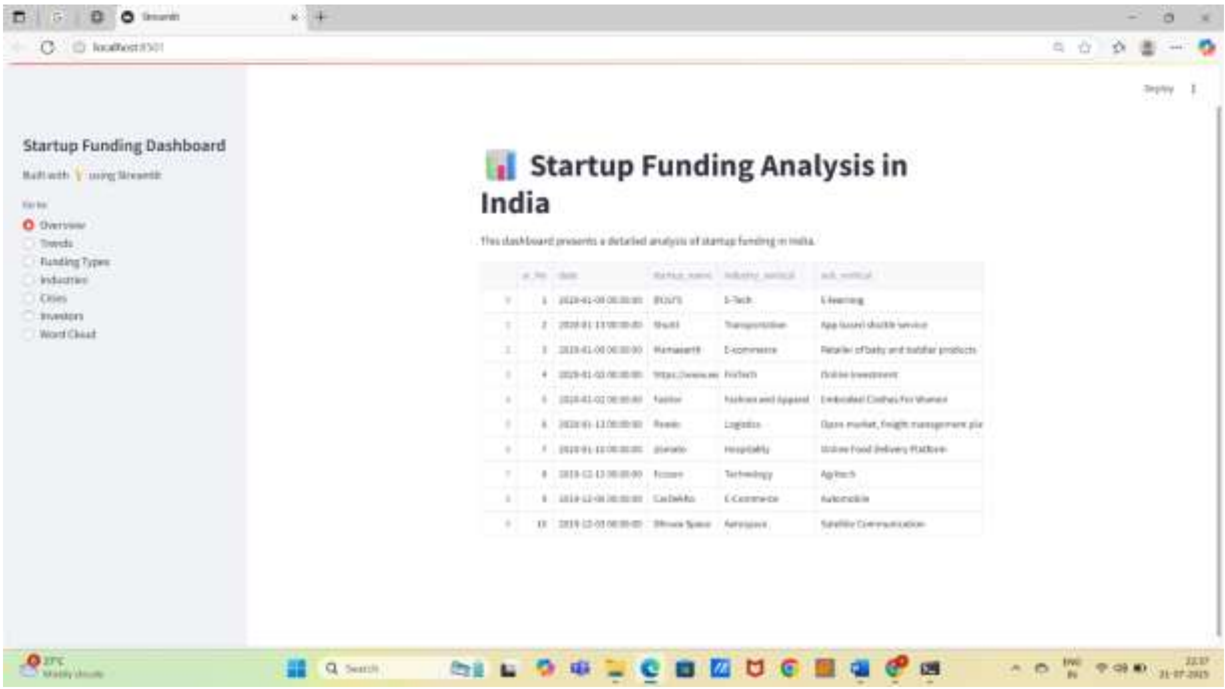
```
(base) C:\Users\DASARI KRISHANA>cd C:\Users\DASARI KRISHANA\Downloads\srianjuprojrct

(base) C:\Users\DASARI KRISHANA\Downloads\srianjuprojrct>Streamlit run app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.31.120:8501
```

Fig: Streamlit App Deployment Confirmation via Anaconda Prompt



Startup Funding Dashboard

Built with 💡 using Streamlit

Go to

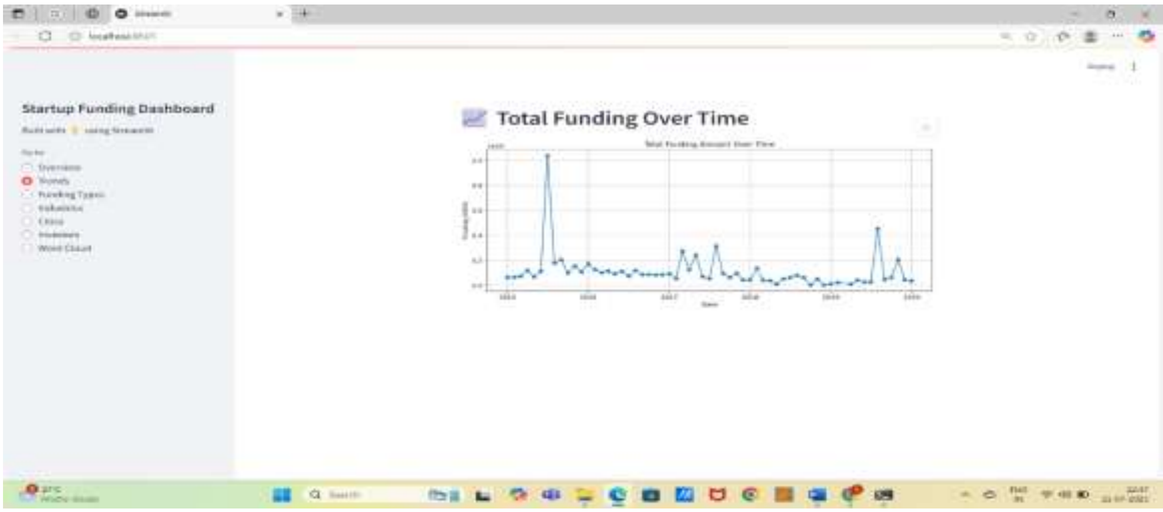
● Overview
○ Trends
○ Funding Types
○ Industries
○ Cities
○ Investors
○ Word Cloud

📊 Startup Funding Analysis in India

This dashboard presents a detailed analysis of startup funding in India.

	sr_no	date	startup_name	industry_vertical	sub_vertical
0	1	2020-01-09 00:00:00	BYJU'S	E-Tech	E-learning
1	2	2020-01-13 00:00:00	Shuttl	Transportation	App based shuttle service
2	3	2020-01-09 00:00:00	Mamaearth	E-commerce	Retailer of baby and toddler products
3	4	2020-01-02 00:00:00	https://www.we	FinTech	Online Investment
4	5	2020-01-02 00:00:00	Fashor	Fashion and Apparel	Embroidered Clothes For Women
5	6	2020-01-13 00:00:00	Pando	Logistics	Open-market, freight management pla
6	7	2020-01-10 00:00:00	Zomato	Hospitality	Online Food Delivery Platform
7	8	2019-12-12 00:00:00	Ecozen	Technology	Agritech
8	9	2019-12-06 00:00:00	CarDekho	E-Commerce	Automobile
9	10	2019-12-03 00:00:00	Dhruva Space	Aerospace	Satellite Communication

Fig: Startup Funding Analysis in India



Startup Funding Dashboard

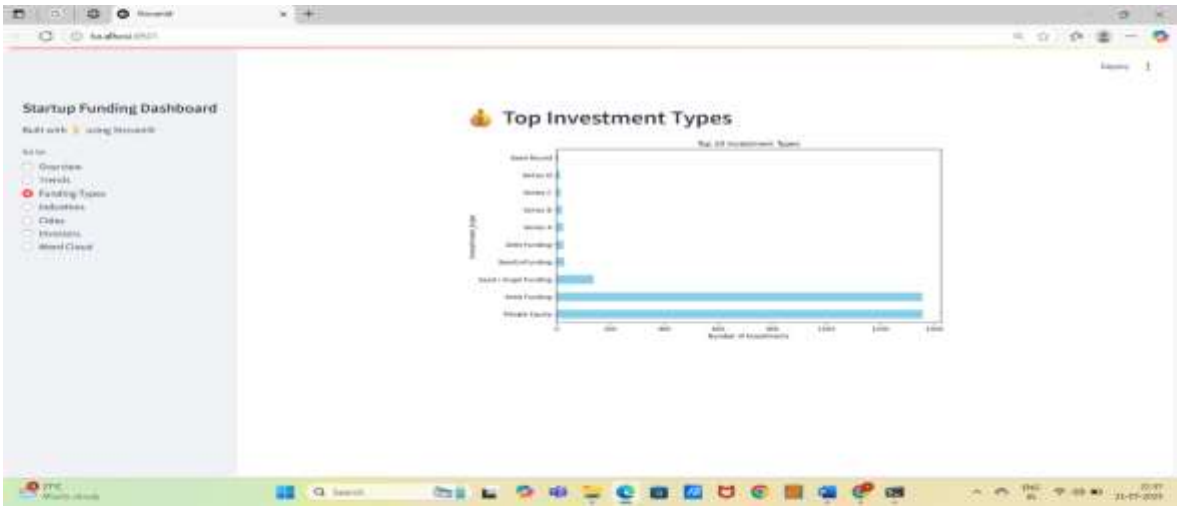
Built with 💡 using Streamlit

Go to

○ Overview
● Trends
○ Funding Types
○ Industries
○ Cities
○ Investors
○ Word Cloud

📈 Total Funding Over Time

Fig: Total Funding Over Time



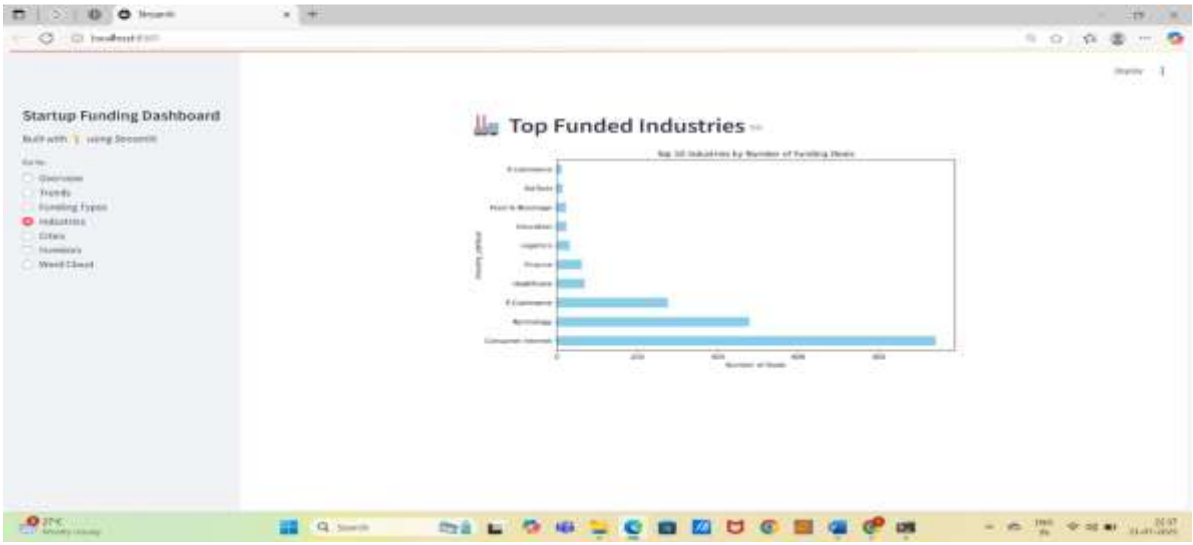
Startup Funding Dashboard

Built with using Streamlit

○ Overview
○ Trends
● Funding Types
○ Industries
○ Cities
○ Investors
○ Word Cloud

Top Investment Types

Fig: Top Investment Types



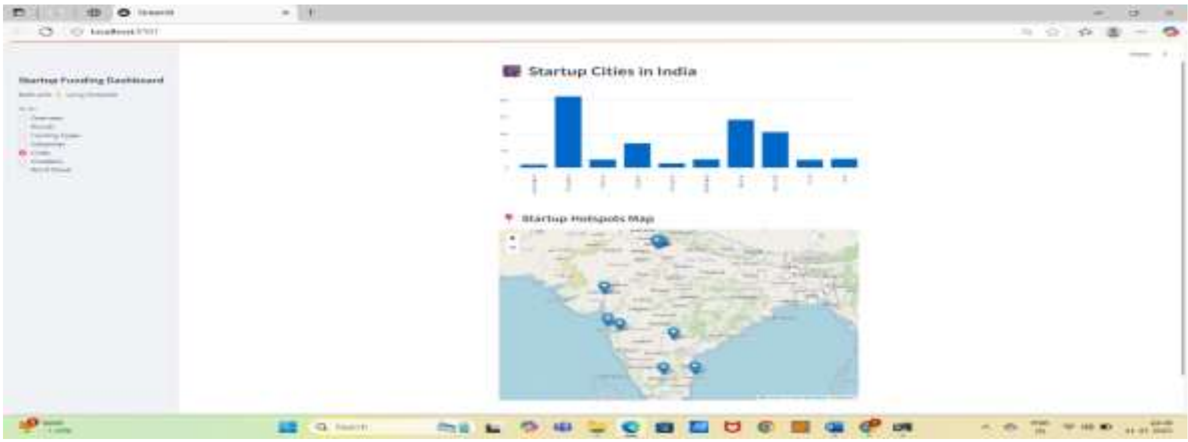
Startup Funding Dashboard

Built with using Streamlit

○ Overview
○ Trends
○ Funding Types
● Industries
○ Cities
○ Investors
○ Word Cloud

Top Funded Industries

Fig: Top Funded Industries



Startup Funding Dashboard

Startup Cities in India

Startup Hotspots Map

Fig: Startup Cities in India

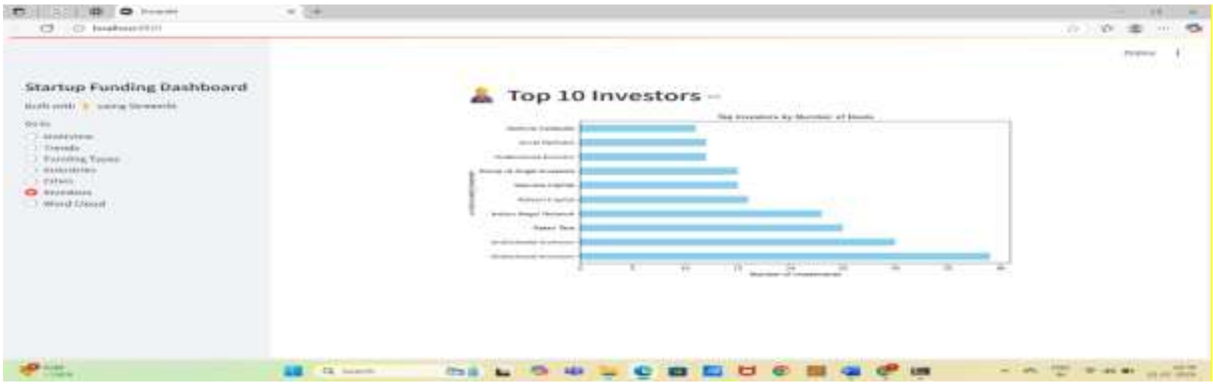


Fig: Top 10 Investors

IV. RESULTS:

The project provides a comprehensive Streamlit-based dashboard for analyzing Indian startup funding trends. It visualizes total funding over time, revealing notable fluctuations with a peak in 2019. The most common investment types include Seed Funding and Private Equity, while key industries like E-Commerce, Technology, and Consumer Internet dominate in terms of both frequency and value of investments. Major cities such as Bangalore, Mumbai, and Delhi lead in startup presence and funding volume. The dashboard also highlights top investors and presents a word cloud of funding types, offering users intuitive insights into India's startup ecosystem.

V. DISCUSSIONS:

The project focuses on creating a comprehensive and interactive dashboard for analyzing startup funding in India using Streamlit. The code systematically handles data cleaning tasks such as standardizing date formats, correcting city names, and cleaning funding amounts to ensure accurate analysis. With various visualizations like bar plots, line charts, and a word cloud, users can explore funding trends, investment types, top industries, and key investor insights. Geospatial data is also utilized to show city-wise startup concentrations on a map, enhancing user engagement. Moreover, the use of both static (Matplotlib, Seaborn) and interactive (Plotly, Folium) visualizations increases interpretability. The modular structure with sidebar navigation improves UI/UX design. Overall, the project effectively bridges data analytics with an intuitive frontend for stakeholders to explore India’s startup funding landscape

VI. CONCLUSION:

the project successfully provides an interactive dashboard using Streamlit to visualize and analyze startup funding trends in India. Through comprehensive data cleaning and processing, the analysis highlights funding patterns over time, dominant investment types, key industries, and active investor cities. Notable insights include funding peaks around 2019, Bangalore as a leading startup hub, and strong investor engagement in Seed and Venture funding. This solution offers a valuable tool for stakeholders to explore the Indian startup ecosystem effectively.

VII. FUTURE SCOPE:

The developed project provides a strong foundation for exploring startup funding patterns in India using interactive visualizations via Streamlit. In the future, the dashboard can be enhanced by integrating real-time data pipelines for up-to-date funding insights, applying machine learning models to predict funding trends, and incorporating investor and startup profiling for personalized recommendations. Expanding geographical coverage beyond India and adding filters for sector-specific deep dives will also broaden its practical utility for entrepreneurs, investors, and policymakers.

VIII. ACKNOWLEDGEMENT:

Chinthagingala. Vasundhara working as an Assistant professor in master of computer application in sanketika vidya parishad engineering college, Visakhapatnam Andhra Pradesh. With 2 years of experience in computer science and engineering (CSE), accredited by NAAC.with her area of intrest in java full stack. She is dedicated and emerging academician in the field of Computer Science, currently serving as a faculty member. With a strong foundation in technical concepts and a passion for teaching, She has begun his academic journey by actively mentoring students in practical and innovative projects. As a faculty, She has already made a significant impact by guiding student teams through their academic projects with clarity, enthusiasm, and technical proficiency. His commitment to student success and interest in research-driven teaching make him a promising contributor to academic excellence.

Eepenagandla. Srikanth is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning E. Srikanth has taken up his PG project on **Exploratory Data Analysis of Indian Startup Funding** and published the paper in connection to the project under the guidance of Chinthagingala. Vasundhara, Assistant Professor, SVPEC.

REFERENCES

1. **Streamlit – Official Documentation**

   https://docs.streamlit.io/

2. **Pandas – Python Data Analysis Library**

   https://pandas.pydata.org/docs/

3. **NumPy – Scientific Computing Library**

   https://numpy.org/doc/

4. **Matplotlib – Data Visualization in Python**

   https://matplotlib.org/stable/users/index.html

5. **Seaborn – Statistical Data Visualization**

   https://seaborn.pydata.org/

6. **Plotly – Interactive Visualization Library**

   https://plotly.com/python/

7. **Folium – Mapping in Python**

https://python-visualization.github.io/folium/

8. **WordCloud – Text Data Visualization**

https://github.com/amueller/word_cloud

9. **Jupyter Notebook**

https://jupyter.org/

10. **Visual Studio Code (VSCode)**

https://code.visualstudio.com/

11. Kaggle Dataset – Indian Startup Funding

https://www.kaggle.com/datasets/sudalairajkumar/indian-startup-funding

12. **GeeksforGeeks – Data Cleaning with Pandas**

https://www.geeksforgeeks.org/data-cleaning-using-python/

13. **GeeksforGeeks – Exploratory Data Analysis (EDA)**

https://www.geeksforgeeks.org/exploratory-data-analysis-in-python/

14. **Medium – EDA Techniques in Python**

https://medium.com/swlh/a-complete-guide-to-exploratory-data-analysis-eda-in-python-891eda4fba4e

15. **W3Schools – HTML Basics**

https://www.w3schools.com/html/

16.**W3Schools – CSS for Styling**

https://www.w3schools.com/css/

17. Bootstrap – Frontend Framework

https://getbootstrap.com/docs/5.0/getting-started/introduction/

18. **GitHub – Streamlit Projects & Dashboards**

https://github.com/streamlit/streamlit

19. **Streamlit-Folium GitHub**

https://github.com/randyzwitch/streamlit-folium

20. **DigitalOcean – Deploying Streamlit Apps**
https://www.digitalocean.com/community/tutorials/how-to-deploy-a-streamlit-application

21. **Analytics Vidhya – Data Preprocessing Techniques**
https://www.analyticsvidhya.com/blog/2021/06/data-preprocessing-techniques-you-should-know/

22. **Scikit-learn Documentation** *(if ML is added later)*
https://scikit-learn.org/stable/documentation.html

23. **Python Official Documentation**
https://docs.python.org/3/

24. **Lucidchart – UML Class Diagrams Guide**
https://www.lucidchart.com/pages/uml-class-diagram

25. **Simplilearn – ER Diagram in DBMS**
https://www.simplilearn.com/tutorials/sql-tutorial/er-diagram-in-dbms