**INDIAN START-UPS ECOSYSTEM**

This article is to share insights gained after extensive analysis had been done on the funding of the Indian startup Ecosystem data. This was to be done using the Cross-Industry Process for Data Mining (CRISP-DM) framework with much emphasis being laid exploratory data analysis.

The Indian startup ecosystem is one of the fastest growing in the world, characterized by a dynamic and rapidly evolving landscape. It comprises a diverse range of sectors which includes but is not limited to technology, e-commerce, fintech, HealthTech, edtech etc. Over the years, India has become home to numerous start-ups being honed in cities like Delhi, Bengaluru, Mumbai, and Hyderabad, and support for same have stemmed from a wide network of investors, incubators, accelerators, and government initiatives.

Some Key factors driving the growth of the Indian startup ecosystem include a large and young population, increasing internet penetration, and a supportive policy environment. Government programs such as Startup India, Digital India, and Make in India have provided a significant boost by offering incentives, funding opportunities, and simplifying regulatory processes.  
The Indian startup ecosystem continues to attract substantial domestic and international investments despite challenges in funding early in the stages of commencement and regulatory hurdles. The maturity of the market is reflected in the rise of IPOs and acquisitions, signaling a robust and promising future for startups in the country.

The project aims to analyze the funding received by startups in India from 2018 to 2021. The goal was to propose recommendations based on evidence based, and data driven insights deduced from the datasets given on a yearly basis. The analysis covered startup information such as funding amounts, investors, sectors, and more.

Data was stored across various platforms or medium, necessitating the need to gather and clean, before the extensive analysis of the data to be able to recommend as need be.

DATA UNDERSTANDING

The datasets for this project were collected from three different sources, namely;

* SQL Database server for datasets 2020 and 2021,
* Microsoft OneDrive for the 2019 dataset
* Github Repository for 2018 dataset.

For purposes of understanding all four datasets were merged at a point and the common columns identified for the analysis are listed as follows;

* **Company Brand**: Name of the company/startup
* **Founded**: year in which the companies were created
* **Headquarters**: where the companies are located in India i.e. the geographical locations of the companies
* **Sectors**: sector of service speaks to what a company does.
* **Founders**: the names of those who founded the companies
* **Investors**: these are the names of the shareholders/those who invested in the companies
* **Amount ($)**: the amount received by the companies
* **Stage**: round of funding attained by each company
* **Funding Year**: the year each company was funded (the funding year was included based on the year the data were collected)

The following are the business questions to be answered at the end of the analysis.

1. What sector has shown the highest growth in term of funding received over the past 4 years?  
   2. what geographical regions within India have emerged as the primary hubs for startup activities and investment and what factors contribute to their prominence  
   3. Are there any notable differences in funding patterns between early-stage startups and more established companies?  
   4. which sectors receive the lowest level of funding and which sector receive the highest level of funding in India and what factors contribute to this?  
   5. which investors have had more impact on start-ups over the years?  
   As part of the analytical questions, a hypothesis will be conducted on whether there is a significant difference in funding between startups in Bengaluru and other startup locations.  
   Data Preparation and Pre-processing  
   Create a virtual environment to handle the project and import all necessary libraries and packages for effectively executing this project.

# import all necessary libraries  
  
#data manipulation  
import numpy as np  
import pandas as pd   
import missingno as msno  
  
#statistical libaries  
from scipy import stats  
import statistics as stat  
import statsmodels.api as sm  
from statsmodels.formula.api import ols   
  
#data visualization libraries  
import matplotlib.pyplot as plt #both the matplotlib and seaborn will be used for visualization of data  
import seaborn as sns   
  
#database connection libraries  
import pyodbc #this library package is used to connect to database servers  
from dotenv import dotenv\_values  
  
#hide warnings  
import warnings  
warnings.filterwarnings('ignore')

1. After successfully loading the data, the following observations were noted from each year’s dataset.  
   **2018 Dataset**  
   • The dataset comprises 526 rows and 6 columns.  
   • It contains one duplicate entry.  
   • The “amount” column includes values in both dollars and rupees, along with some non-numeric characters.  
   • The “location” column features a mix of state, regional, and city information.  
   • There is a Google document link in the “Round/Series” column.  
   • The columns in 2018 are different from those of 2019–2021, meaning they have to be renamed before concatenation.  
     
   **2019 Dataset**  
   - The dataset comprises 89 rows and 9 columns.  
   - There are no duplicate entries, but it does contain null values.  
   - The “amount” column is of float datatype and includes other non-numeric characters.
2. Columns such as “Stage,” “HeadQuarters,” and “Founded” have many missing values.
3. **2020 Dataset**  
   - The dataset comprises 1055 rows and 10 columns.  
   - It contains three duplicate entries.  
   - The “amount” column is of float datatype and includes other non-numeric characters.  
   - The “Headquarters” column includes some locations outside of India.
4. **2021 Dataset**  
   • The dataset comprises 1209 rows and 9 columns.  
   • It contains 19 duplicate entries.  
   • The “amount” column is of float datatype and includes other non-numeric characters.  
   • There are instances where values have been recorded under incorrect columns.  
   All datasets, except for 2018, contain missing values. The datasets are plagued with misplaced values, duplicates, and missing entries. Overall, the datasets are messy and will require thorough cleaning.
5. To thoroughly clean the dataset, the following assumptions were made:  
   All amount values without specified currencies are assumed to be in USD.  
   -The exchange rate between USD and Indian Rupee as of 2018 is 1 USD = 83.3233 INR.

import pandas as pd  
  
# Assume df\_2018 is already defined with 'Amount($)' column  
exchange\_rate = 83.32  
  
def clean\_Amount():  
 # Copy the 'Amount($)' column  
 amount\_column = df\_2018['Amount($)'].copy()  
   
 # Remove commas from the values  
 amount\_column = amount\_column.str.replace(',', '', regex=True)  
   
 # Extract the values in rupees  
 amount\_in\_rupees = amount\_column[amount\_column.str.startswith('₹')]  
 # Remove the rupee symbol and convert to float  
 amount\_in\_rupees = amount\_in\_rupees.str.lstrip('₹').astype(float)  
 # Convert rupees to dollars  
 amount\_in\_rupees = amount\_in\_rupees / exchange\_rate  
   
 # Extract the values in dollars  
 amount\_in\_dollars = amount\_column[amount\_column.str.startswith('$')]  
 # Remove the dollar symbol and convert to float  
 amount\_in\_dollars = amount\_in\_dollars.str.lstrip('$').astype(float)  
   
 # Replace the unclean column with the clean ones  
 amount\_column.loc[amount\_in\_rupees.index] = amount\_in\_rupees  
 amount\_column.loc[amount\_in\_dollars.index] = amount\_in\_dollars  
   
 # Convert the column to numeric  
 amount\_column = pd.to\_numeric(amount\_column, errors='coerce')  
  
# Update the original dataframe  
df\_2018['Amount($)'] = amount\_column

The median was used in filling the missing values because of the outliers present in the amount column.  
Stage Column  
The data was processed by categorizing the “Stage” column into high-level funding stages, based on the stages listed on the Indian Startup Ecosystem website. A function was defined to facilitate the distribution of the stages. Additionally, the author introduced a seventh stage named “Undisclosed” to account for missing values in the “Stage” column. The Google document link mentioned during the data understanding phase was also replaced with “Undisclosed”.

# Define the sector redistribution function  
def stage\_distribution(Stage):  
 if re.search(r'Pre-seed|Pre-seed Round|PE|Private Equity|Pre seed Round|Pre seed round|Pre-Seed', Stage):  
 return 'Pre\_seed'  
 elif re.search(r'Seed fund|Seed|Seed A|Fresh funding|Early seed|Seed Funding|Seed+|Seed funding|Angel Round|Angel|Seed round|Seed Round|Seed Investment',Stage):  
 return 'Seed'  
 elif re.search(r'Seed Round & Series A|Venture - Series Unknown|Post series A|Seies A|Series A2|Series A+|Pre Series A|Pre series A1|Pre-series A1',Stage):  
 return 'Series\_A'  
 elif re.search(r'Pre series B|Series B3|Series B2|Series B+|Pre-Series B',Stage):  
 return 'Series\_B'  
 elif re.search(r'Pre series C|Mid series',Stage):  
 return 'Series\_C'  
 elif re.search(r'Series E2|Series G|Series E|Series H|Series F|Corporate Round|Series D1|Series C, D|Series F2|Series F1|Series I',Stage):  
 return 'Series\_D and Beyond'  
 elif re.search(r'Post-IPO Debt|Post-IPO Equity|Non-equity Assistance|Debt|Debt Financing|Grant|Secondary Market',Stage):  
 return 'IPO'  
 elif re.search(r'Bridge Round|Bridge',Stage):  
 return 'Bridge'  
 elif re.search(r'Edge',Stage):  
 return 'Edge'  
 else:  
 return np.nan  
 # Convert "Sector" column to string type  
df\_2020\_2021\_2019\_2018['Stage'] = df\_2020\_2021\_2019\_2018['Stage'].astype(str)  
# Apply the function to create the new column  
df\_2020\_2021\_2019\_2018['Stage'] = df\_2020\_2021\_2019\_2018['Stage'].apply(stage\_distribution)

HeadQuarter Column  
The HeadQuarter column was cleaned by creating a location dictionary to map data to the appropriate locations. Similar city names were standardized, for instance, Bangalore was mapped to Bengaluru, and Delhi to New Delhi.  
Cities outside of India were replaced with ‘Overseas’, as investors often favor well-established companies when making investment decisions.  
Regional and state information was removed.  
Sector Column  
Next, the sector column, which indicates the economic sector in which a startup operates, was cleaned. As you might have guessed, this involved using another dictionary.  
Based on research, the values in the sector column were generalized into 15 sectors commonly recognized in India. These sectors include:  
- IT & Technology  
- Financial Services  
- Healthcare & Life Sciences  
- Consumer Goods  
- Business Services  
- Media & Entertainment  
- Education  
- Manufacturing  
- Retail  
- Transportation & Logistics  
- Sports  
- Agriculture  
- Real Estate  
- Travel & Tourism  
- Energy  
- Others  
Other Columns  
The other columns were cleaned and processed as needed to ensure data consistency and accuracy.  
A new column named ‘funded\_year’ was added to all four datasets to indicate the year when the funding was received.

Furthermore, in the 2018 dataset, the column names were modified to align with the naming conventions of the other years. Any missing columns were also included for consistency across all datasets.

# add the year column to the 2018 dataframe   
df\_2018['Funding\_Year']=2018  
df\_2018['Funding\_Year'] =pd.to\_datetime(df\_2018['Funding\_Year'],format='%Y',errors ='coerce')  
  
# add investors column  
df\_2018['Investor']='Undisclosed'  
  
# add the founders column to the data frame  
df\_2018['Founders']='Undisclosed'

I assumed that all the missing values within the founded column should be replaced by the most frequent year in the column.

**Hypothesis testing**  
Firstly, a normality and homogeneity test were carried out to check if the data is normality distributed i.e. if our data is not skewed, and also with the Levene test we checked for the equality of variance in the dataset i.e. homogeneity. this needs to be done to know if we are to use a parametric test which is the t-test or a non-parametric test such as Mann-Whitney U. of course we ended up carrying out the hypothesis test with the Mann-Whitney U test due to the skewness of the data.

funding\_bengaluru = year\_of\_existence\_with\_5years[year\_of\_existence\_with\_5years['HeadQuarter'] == 'Bengaluru']['Amount($)']  
funding\_other = year\_of\_existence\_with\_5years[year\_of\_existence\_with\_5years['HeadQuarter'] != 'Bengaluru']['Amount($)']  
# Check for normality  
shapiro\_bengaluru = shapiro(funding\_bengaluru)  
shapiro\_other = shapiro(funding\_other)  
  
print(f"Shapiro test for bengaluru: Statistic={shapiro\_bengaluru.statistic}, p-value={shapiro\_bengaluru.pvalue}")  
print(f"Shapiro test for Other locations: Statistic={shapiro\_other.statistic}, p-value={shapiro\_other.pvalue}")  
  
# Check for equal variances  
levene\_test = levene(funding\_bengaluru, funding\_other)  
print(f"Levene's test: Statistic={levene\_test.statistic}, p-value={levene\_test.pvalue}")  
  
# Choose and perform the appropriate test  
if shapiro\_bengaluru.pvalue > 0.05 and shapiro\_other.pvalue > 0.05:  
 if levene\_test.pvalue > 0.05:  
 # Two-sample t-test (assuming equal variances)  
 t\_stat, p\_value = ttest\_ind(funding\_bengaluru, funding\_other, equal\_var=True)  
 else:  
 # Welch's t-test (assuming unequal variances)  
 t\_stat, p\_value = ttest\_ind(funding\_bengaluru, funding\_other, equal\_var=False)  
else:  
 # Mann-Whitney U test  
 u\_stat, p\_value = mannwhitneyu(funding\_bengaluru, funding\_other)  
  
print(f"Test result: Statistic={t\_stat if 't\_stat' in locals() else u\_stat}, p-value={p\_value}")

Based on the result above;  
For the Shapiro Wilks test which tests for the normality of the dataset, the p-value is lesser than 0.05 for both the Bengaluru location and other locations, this shows that the data is not normally distributed.  
For Levene’s test the p-value is greater than 0.05 this shows that the dataset has constant variances  
The Mann-Whitney U test was carried out instead of the t-test because the data is not normally distributed. the p-value for the test result is less than 0.05, therefore we reject the null hypothesis.  
- Conclusion:  
There is a significant difference in the funding amount between Bengaluru and other headquarters

**Answering the Business Questions**

1. **What sector has shown the highest growth in funding received over the past 4 years?**

the manufacturing sector has shown the highest growth in funding over the past four years even though it is not the most funded sector.

1. **2. what geographical region within India have emerged as the primary hubs for startup activities and investment and what factors contribute to their prominence**

Although Bengaluru is known for having the highest concentration of startups, Mumbai is receiving more funding. Based on this, I believe Mumbai is the primary hub for startups, as it attracts more investment than any other location.

1. **3. Are there any notable differences in funding patterns between early-stage startups and more established companies?**

Companies in the early stages, such as pre-seed, seed, Series A, and Series B, received relatively small amounts of funding. Similarly, more established companies at the Series C, Series D, and beyond stages also received significantly lower funds. In contrast, fully established companies at the IPO level received considerably more funding.

1. 4. **which sectors receive the lowest level of funding and which sector recieve the highest level of funding in India and what factor contribute to this?**

The financial sector attracts the most investment, whereas the sports sector receives the least funding.

1. 5. Which investor has more impact on start-ups over the years?
2. From the visualization, it is evident that ‘Tradecred’ has made the highest investment in a single sector, specifically the manufacturing sector. In contrast, ‘Tiger Global’ has diversified its investments across four different sectors: retail, financial services, education, and others. Based on this observation, we can conclude that ‘Tiger Global’ is the most impactful investor in funding for startups, due to its extensive and varied investment portfolio. it should also be noted that a substantial number of investors’ names were undisclosed, the visualization and result generated above is for the known investors

## Recommendation

1. Based on the analysis of the Indian startup ecosystem, I recommend investing in the manufacturing sector due to its highest growth rate among all sectors, despite not being the most funded. Additionally, establishing the company in Mumbai is advisable, as the city attracts more investors than any other location.

# Deployment

1. The final phase of the project is deployment — making the results accessible to stakeholders.
2. To foster a collaborative learning environment, I have made the entire project accessible through my GitHub repository. Here, you will find the complete codebase, extensive documentation, and in-depth analyses. Additionally, I have developed an interactive Power BI dashboard that showcases engaging visualizations and key trends. I encourage you to explore the GitHub repository and interact with the Power BI dashboard. Together, we can delve into this insightful journey, enabling entrepreneurs and enthusiasts to leverage data-driven strategies for a thriving Indian startup ecosystem.
3. I appreciate everyone who contributed to the success of these project.
4. Assayouti Zakariyah
5. Data Analyst/Data Science Professional