Startups in India

Springboard Data Science Career track

Capstone 2

Patrick Long

3/29/2022

Startups in India

Overview

The purpose of this study is to gather past data on start-ups in India to better understand how, and where future companies will be funded.

I will use these factors to determine where and how start-ups are being funded:

-Amount funded

-Where the funding is coming from

-Which state the funding is going to

-Which industry is being funded

-When the start-ups are being funded

After having an initial look at the data, I formulated a hypothesis:

Most funding will go towards the most populated cities in India, Mumbai and Delhi. Those industries will be service based industries, based on current market trends in India.

Data Wrangling

Datasets

I will leverage data from multiple sources. The data will come from Kaggle, RBI, and Statisticstimes.com.

https://www.kaggle.com/paree24/india-gdp-growth-world-bank-1961-to-2017

https://rbi.org.in/Scripts/PublicationsView.aspx?id=20006

https://www.statisticstimes.com/economy/comparing-indian-states-and-countries-by-gdp.php

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

The data comes in CSV form and will need to be cleaned and merged.

The first step of this project is the Data Wrangling section, found here:

https://github.com/ptlong11/capstone2/blob/main/Capstone%20Two%20-%20Data%20Wrangling%20(2).ipynb

The approach I took to data Wrangling was to first clean the data sources. That means deleting any arbitrary columns, finding the null values, and replacing or removing them.

After each set of data was clean, I merged them all on the 'State' column to end up with one data frame that had 'State' as the de facto index. This allowed me to work within that data frame comparing the funding amounts and types by various states. This approach helped me figure out my initial hypothesis.

The Kaggle data had GDP growth in India from 1962 to 2017. I used the 2017 year, to be as up to date as possible. The RBI data had GDP growth per state, which was broken up by year. I kept it to 2017 to be the same with the Kaggle data. There were no missing data in either set for 2017.

The next Kaggle data set is crucial to the project, as it contains the startup funding. The data has great integrity and only had a few missing values, which were removed during the data wrangling section.

Exploratory Data Analysis, Pre-Processing, and Modeling

After the data collecting, cleaning, and merging, the next step is to dive into the data.

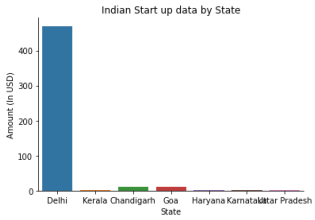
In this section I have merged the data sets on the “State” column.

State and GDP comparison. Found here:

	Rank	State	GDP
0	1	Maharashtra	399.921
1	2	Tamil Nadu	247.629
2	3	Uttar Pradesh	240.726
3	4	Gujarat	228.290
4	5	Karnataka	226.806

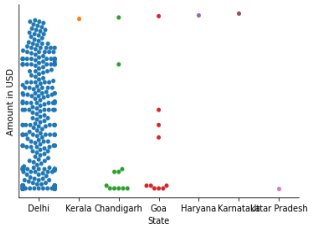
We can see the top GDP is Maharashtra followed by Tamil Nadu.

I did so by comparing first the Amount of overall funding and each Indian state. I wanted to see where the money was going:



We can see from this chart that most, if not, nearly all the startup funding is going to the state of Delhi.

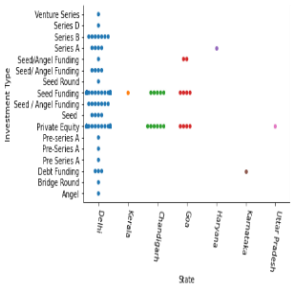
I ran another type of chart, where we can see individual dots, each representing an individual funding source:



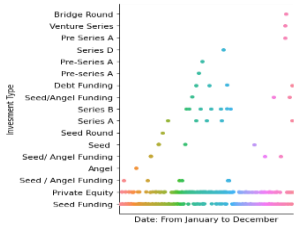
I think this chart gives us a better idea of where the money is going. Yes, most of the funding is going into Delhi, but we can see more clearly that there is still funding going into different states. Most notably Chandigarh and Goa.

Now, that we know where the funding is going. I want to see where the money is coming from. The data of funding is broken down into types of funding. To name a few; “Venture series” “Series A, B, C, D” “Seed.” These are examples of classic startup fundings. For a more indepth explanation of startup funding: https://www.investopedia.com/articles/personal-finance/102015/series-b-c-funding-what-it-all-means-and-how-it-works.asp

Startups in India

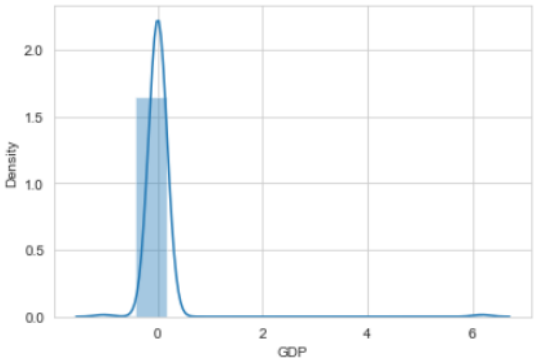


This graph gives a great look at the types of funding coming in. There are mostly Seed fundings and Private Equity rounds.



This graph gives us an example of when during the year the funding is coming in. There doesn't appear to be any 'fudning season' – funding is spread throughout the year.

I ran a radom Forest Regression analysis on the GDP (State) and the Amount of funding coming into each state. This will help me create a regression analysis.



This graph is showing that again, most funding is coming into the state of Delhi. However, it is slightly offset. That means that trends are still inline with most funding going to Delhi, but there is still enough funding outside of Delhi to be significant.

Findings and recommendations

Ultimately, trying to find and predict the best state in India to invest in is hard. Because so much of the start-up industry is in one state, Delhi, that is the de facto place to invest.

My initial hypothesis that most money in the future will be going to Delhi, I also predicted that it's not the only state to invest in. Though, due to the limits of the data, I was not able to come up with the exact industries that the funding was going into.

I still maintain that there will be more growth outside of Delhi. However, for the time being any new investor in Indian Start-ups would be remiss to not start their due diligence in Delhi.