Performing Data Processing and Visualization on RawData Containing State-Wise, Sector Wise Distribution of Startups in India from 2016-2022

Presented By: Gajanan Satote

Index:

SR NO	TITLE	PAGE
1	ABSTRACT	3
2	INTRODUCTION	4
3	THE DATA	4
4	THE VISUALIZATION	4
5	SOFTWARE AND PACKAGES USED	4
6	READING AND FILTERING DATA	5
7	SECTOR/INDUSTRY VS STARTUP DATA	7
8	STATE VS NUMBER OF STARTUPS	9
9	INDIA WIDE DENSITY OF STARTUPS	11
10	CO RELATION OF YEAR AND NUMBER OF STARTUPS.	14
11	3D SCATTER PLOTS	15
12	SECONDARY DATA PROCESSING	16
13	WORDCLOUD OF BIGGEST INVESTORS	17
14	CITYWISE DISTRIBUTION OF UNICORNS	18
15	CONCLUSIONS AND INFERENCES	19
16	REFERENCES	19

ABSTRACT:

Narendra Modi's startup campaign in 2015 has given a boost to number of startups in India. Thus, as a motivation for the project, the startup data available at the government website is taken into processing. Analysis cleaning, filtering of raw data and visualization is conducted in the report to generate meaningful outcome. Software Packages in Python, SQLite, DB Browser are used to query, filter and process data. Finally, conclusions are drawn on the basis of results obtained in the report.

INTRODUCTION:

Narendra Modi launched the much-needed program "Start-up India" in 2015. India is a country known for its skilled populace. The kids, however, do not have many chances to realize their aspirations. The youth can thus use this campaign as a terrific springboard to achieve their objectives. On Independence Day, our prime minister made the announcement. This was introduced on January 16 in order to assist the youth. Since then, data of number of startups registered in India Under various categories is huge and is maintained by government.

In this report raw data is from data.gov.in is acquired and processed to conclude the data and represented in a meaningful way.

THE DATA:

The data contains two parts:

1. Primary Data containing sate-wise list of startups in various categories
2. Secondary Data containing list of Startup turned Unicorns in the period

Both the Data is present in csv i.e. (Comma Separated Values) format, which is read into a processing software, here python, and added to a particular database here SQLite and then processed to be able to visualize and get some Output.

THE VISUALIZATION:

The downloaded data will be filtered and processed in python and visualized using different charts and graphs to draw usable conclusions of the data.

SOFTWARES and PACKAGES USED:

The project is conducted on Python software, and following libraries are used:

- Pandas
- Numpy
- Plotnine
- Plotly
- GeoPandas
- WordCloud
- Matplotlib

READING and FILTERING DATA:

Reading Primary Data:

The data named gov-data-final is imported and checked for its values in panda data frame. It shows the values as follows:

	S No.	Year	State	Industry	Count	Last Update
0	1	2022	Andaman and Nicobar Islands	Agriculture	1	2/10/2022 4:00
1	2	2022	Andaman and Nicobar Islands	AR VR (Augmented + Virtual Reality)	1	2/10/2022 4:00
2	3	2022	Andaman and Nicobar Islands	Construction	1	2/10/2022 4:00
3	4	2022	Andaman and Nicobar Islands	Internet of Things	1	2/10/2022 4:00
4	5	2022	Andaman and Nicobar Islands	Marketing	1	2/10/2022 4:00

Out of these the Last update column is of no use for the scope of the project, hence it dropped from the dataframe and the updated dataframe looks like:

	S No.	Year	State	Industry	Count
0	1	2022	Andaman and Nicobar Islands	Agriculture	1
1	2	2022	Andaman and Nicobar Islands	AR VR (Augmented + Virtual Reality)	1
2	3	2022	Andaman and Nicobar Islands	Construction	1
3	4	2022	Andaman and Nicobar Islands	Internet of Things	1
4	5	2022	Andaman and Nicobar Islands	Marketing	1

Storing Primary Data to SQL:

Python was used to Query the data and store it in SQL using SQLite in the database named: StartUp_Primary.db

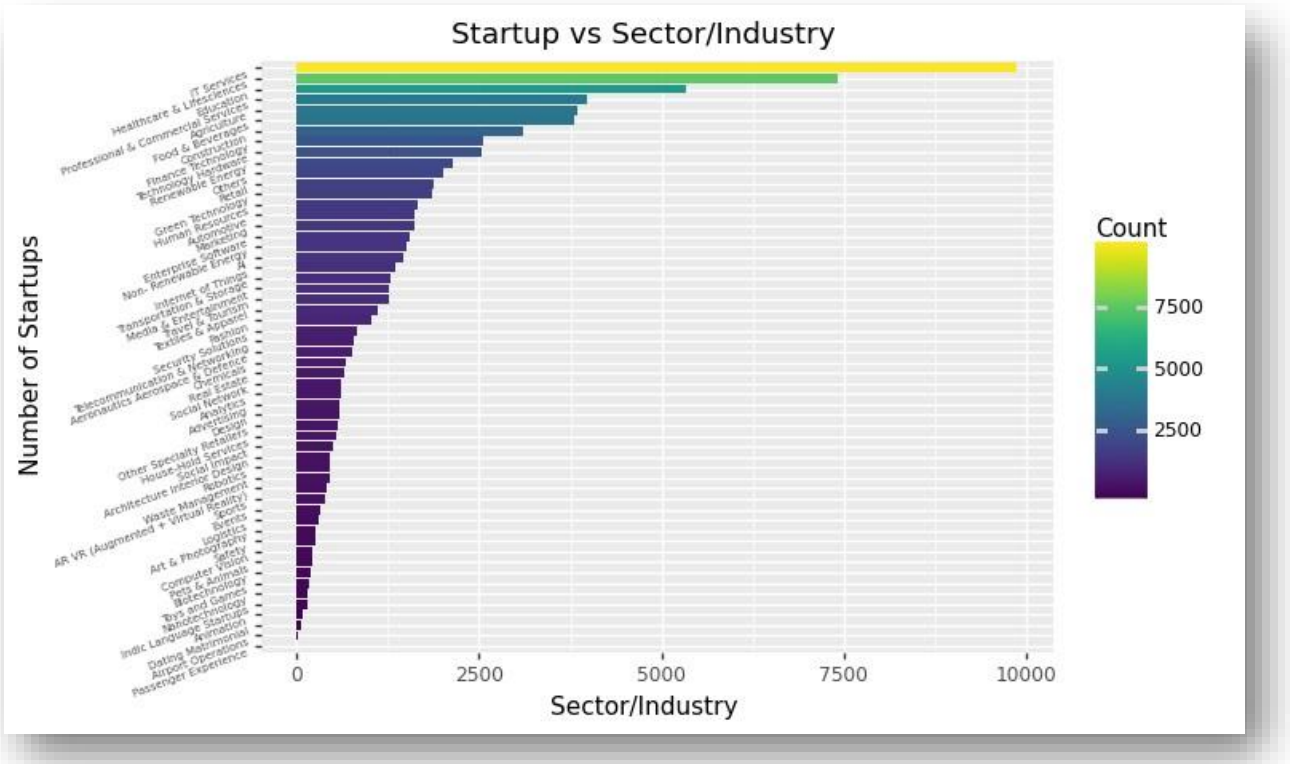
S_NO	YEAR	STATE	INDUSTRY	COUNT
Filter	Filter	Filter	Filter	Filter
1	2022	Andaman and Nicobar Islands	Agriculture	1
2	2022	Andaman and Nicobar Islands	AR VR (Augmented + Virtual Reality)	1
3	2022	Andaman and Nicobar Islands	Construction	1
4	2022	Andaman and Nicobar Islands	Internet of Things	1
5	2022	Andaman and Nicobar Islands	Marketing	1
6	2022	Andaman and Nicobar Islands	Other Specialty Retailers	1
7	2022	Andaman and Nicobar Islands	Transportation & Storage	1
8	2022	Andaman and Nicobar Islands	Travel & Tourism	2
9	2022	Andhra Pradesh	Advertising	2
10	2022	Andhra Pradesh	Aeronautics Aerospace & Defence	5
11	2022	Andhra Pradesh	Agriculture	21
12	2022	Andhra Pradesh	AI	3

SECTOR/INDUSTRY vs STARTUP DATA:

It is logical to check which sector has the greatest number of startups year wise. To do so a new dataframe is created which stores the count of startups, industry/ sector wise. This is done using the groupby function in pandas python, where the final modified dataframe looks like:

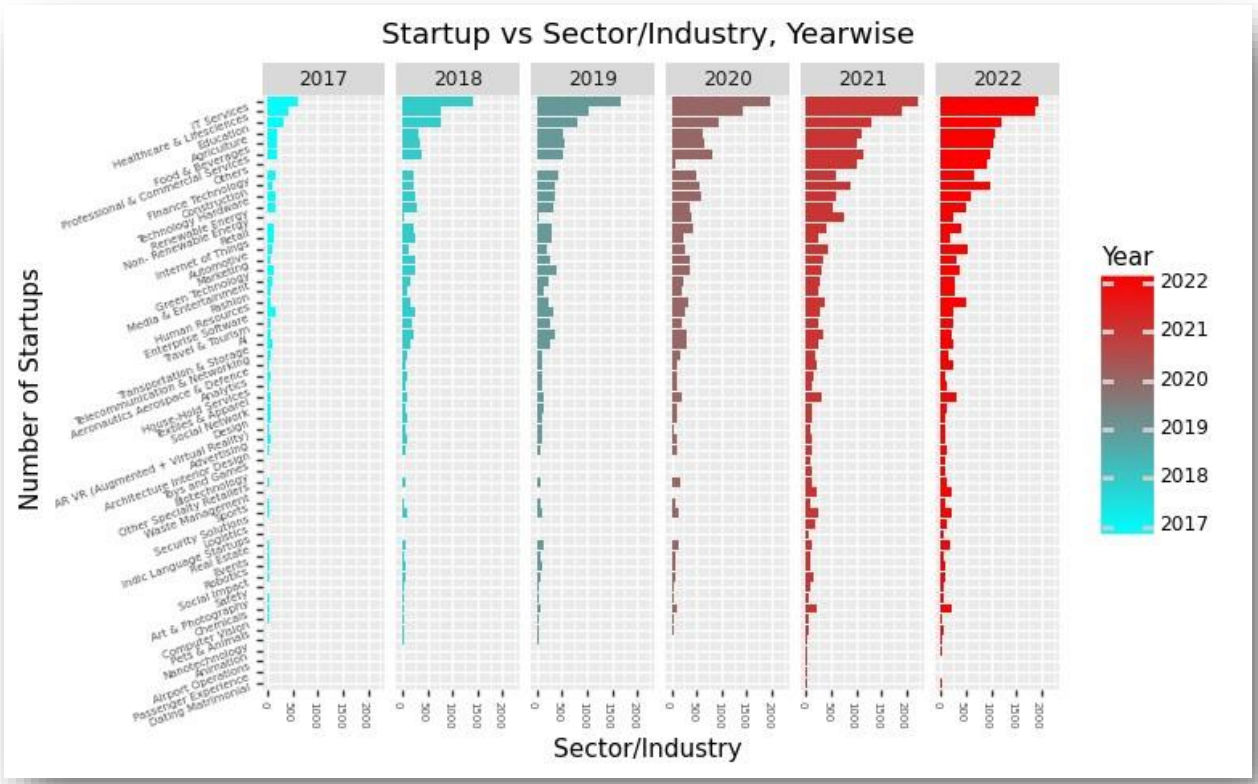
	Industry	Count
0	AI	1473
1	AR VR (Augmented + Virtual Reality)	420
2	Advertising	594
3	Aeronautics Aerospace & Defence	763
4	Agriculture	3848
5	Airport Operations	10
6	Analytics	600
7	Animation	81
8	Architecture Interior Design	454
9	Art & Photography	262

The Count vs the Sector data then is visualized as follows:



It is clear that IT services have the greatest number of Startups followed by Healthcare, Education, Professional/Commercial Services and Agriculture to sum up the Top 5 sectors.

Year wise trend of the data is also visualized to get a better idea how have the startup sectors progressed year wise.



The data shows since the announcement of Startup India Campaign the number of startups have significantly increased even in times such as COVID, and these numbers seem continue to grow.

STATE vs NUMBER of STARTUPS:

To visualize this data a new dataframe was created that stores the total count of startups state wise.

	State	Count
0	Maharashtra	14667
1	Karnataka	9415
2	Delhi	9158
3	Uttar Pradesh	7320
4	Gujarat	5516
5	Tamil Nadu	4357
6	Haryana	4287
7	Telangana	4193
8	Kerala	3544
9	West Bengal	2581
10	Rajasthan	2566
11	Madhya Pradesh	2354
12	Odisha	1444

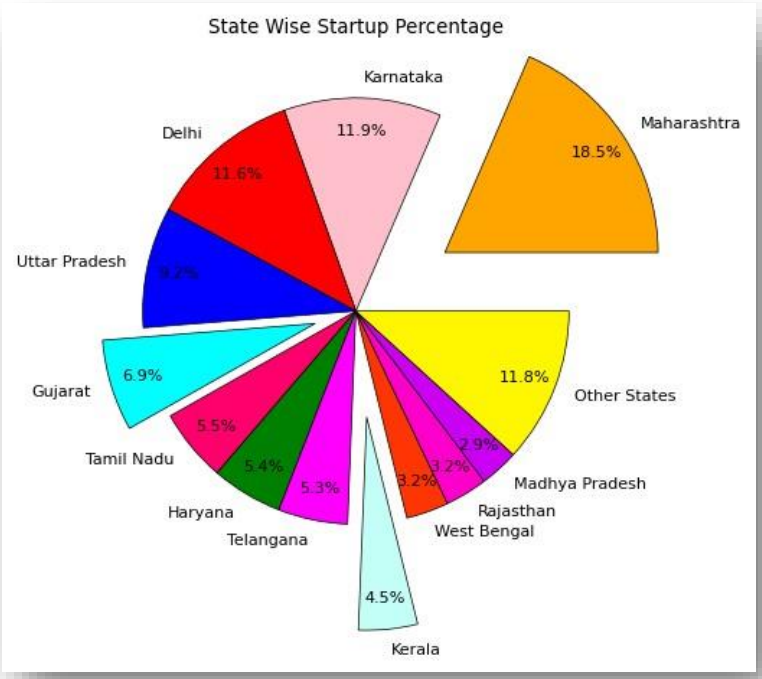
To filter some data, we add a percentage column to the data which makes the output as follows:

	State	Count	Percentage
0	Maharashtra	14667	18.49
1	Karnataka	9415	11.87
2	Delhi	9158	11.55
3	Uttar Pradesh	7320	9.23
4	Gujarat	5516	6.95
5	Tamil Nadu	4357	5.49
6	Haryana	4287	5.4
7	Telangana	4193	5.29
8	Kerala	3544	4.47
9	West Bengal	2581	3.25
10	Rajasthan	2566	3.24
11	Madhya Pradesh	2354	2.97
12	Odisha	1444	1.82

The data fewer than 2% weightage is categorized into a new row which denotes Other States. This can be viewed as follows:

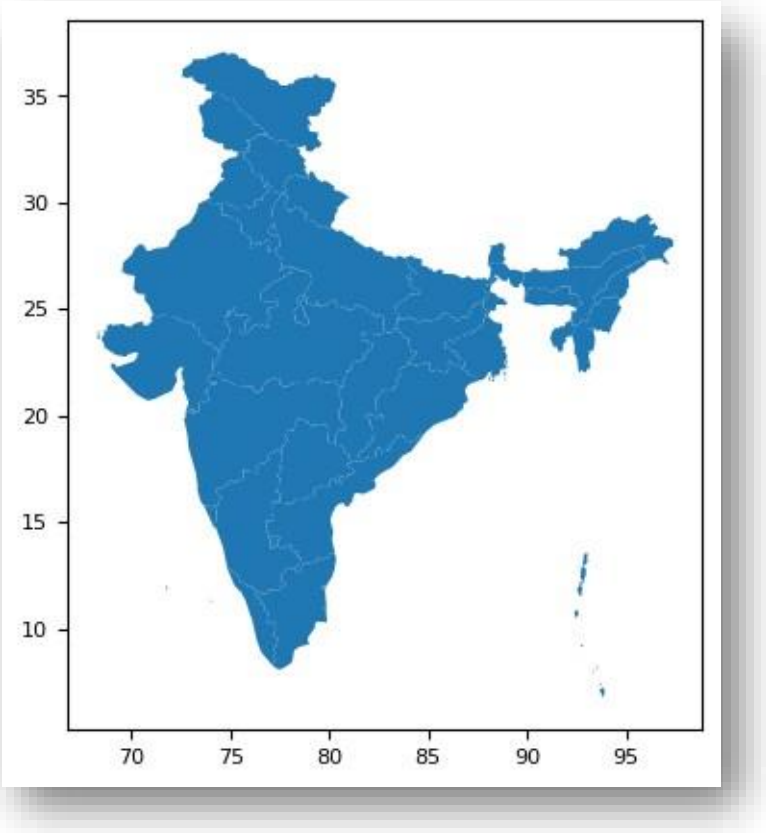
	State	Count	Percentage
0	Maharashtra	14667	18.49
1	Karnataka	9415	11.87
2	Delhi	9158	11.55
3	Uttar Pradesh	7320	9.23
4	Gujarat	5516	6.95
5	Tamil Nadu	4357	5.49
6	Haryana	4287	5.4
7	Telangana	4193	5.29
8	Kerala	3544	4.47
9	West Bengal	2581	3.25
10	Rajasthan	2566	3.24
11	Madhya Pradesh	2354	2.97
12	Other States	9361	11.81

This helps in visualizing State-wise data into a better Pie Chart:



INDIA WIDE DENSITY OF STARTUPS:

An effort was made to display the distribution of Count in form of Color Density on the map of India. For this a shape file of India is imported into python using geopanda library, a basic boundary plot of map of India in python is as shown:



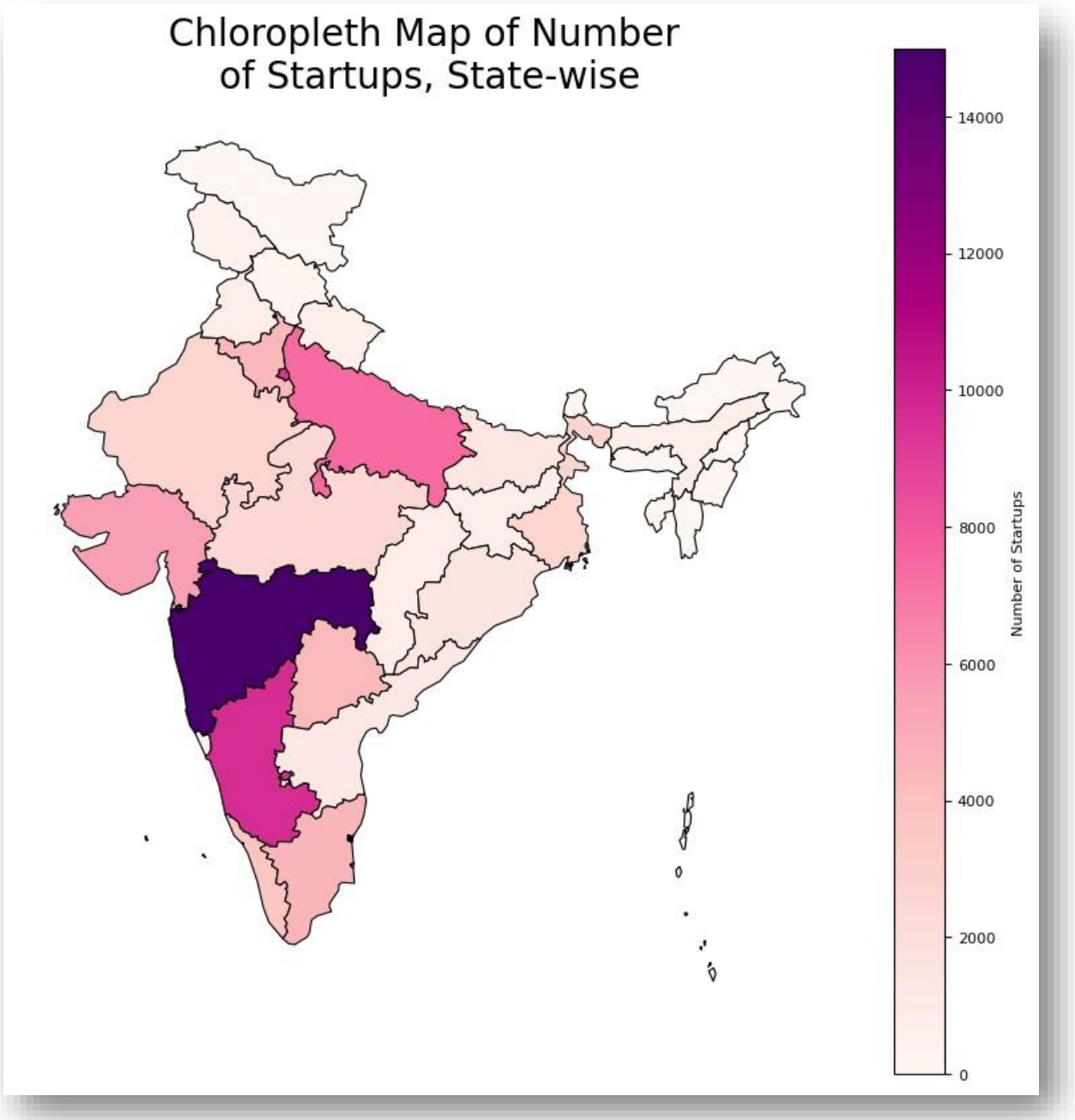
This is executed using the geometry column in the shape file data frame which looks like:

	id	st_nm	geometry
0	None	Andaman and Nicobar Islands	MULTIPOLYGON (((93.84831 7.24028, 93.92705 7.0...
1	None	Arunachal Pradesh	POLYGON ((95.23643 26.68105, 95.19594 27.03612...
2	None	Assam	POLYGON ((95.19594 27.03612, 95.08795 26.94578...
3	None	Bihar	POLYGON ((88.11357 26.54028, 88.28006 26.37640...
4	None	Chandigarh	POLYGON ((76.84208 30.76124, 76.83758 30.72552...

This shape file is merged with our dataframe which contains the column Count and thus can be used to plot final choropleth map of India.

st_nm	id	geometry	Count	Percentage
Andaman and Nicobar Islands	0	MULTIPOLYGON (((93.84831 7.24028, 93.92705 7.0...	38.0	0.05
Arunachal Pradesh	0	POLYGON ((95.23643 26.68105, 95.19594 27.03612...	13.0	0.02
Assam	0	POLYGON ((95.19594 27.03612, 95.08795 26.94578...	663.0	0.84
Bihar	0	POLYGON ((88.11357 26.54028, 88.28006 26.37640...	1376.0	1.73
Chandigarh	0	POLYGON ((76.84208 30.76124, 76.83758 30.72552...	274.0	0.35
Chhattisgarh	0	POLYGON ((83.94694 23.62196, 83.95594 23.62406...	827.0	1.04
Dadra and Nagar Haveli	0	POLYGON ((73.20640 20.12165, 73.20865 20.10695...	0.0	0.00
Daman and Diu	0	POLYGON ((72.80144 20.37378, 72.84418 20.47463...	0.0	0.00
Goa	0	POLYGON ((74.11982 15.65278, 74.24806 15.65698...	327.0	0.41
Gujarat	0	MULTIPOLYGON (((68.35808 23.80475, 68.41658 23...	5516.0	6.95

The final visualization can be observed as follows:

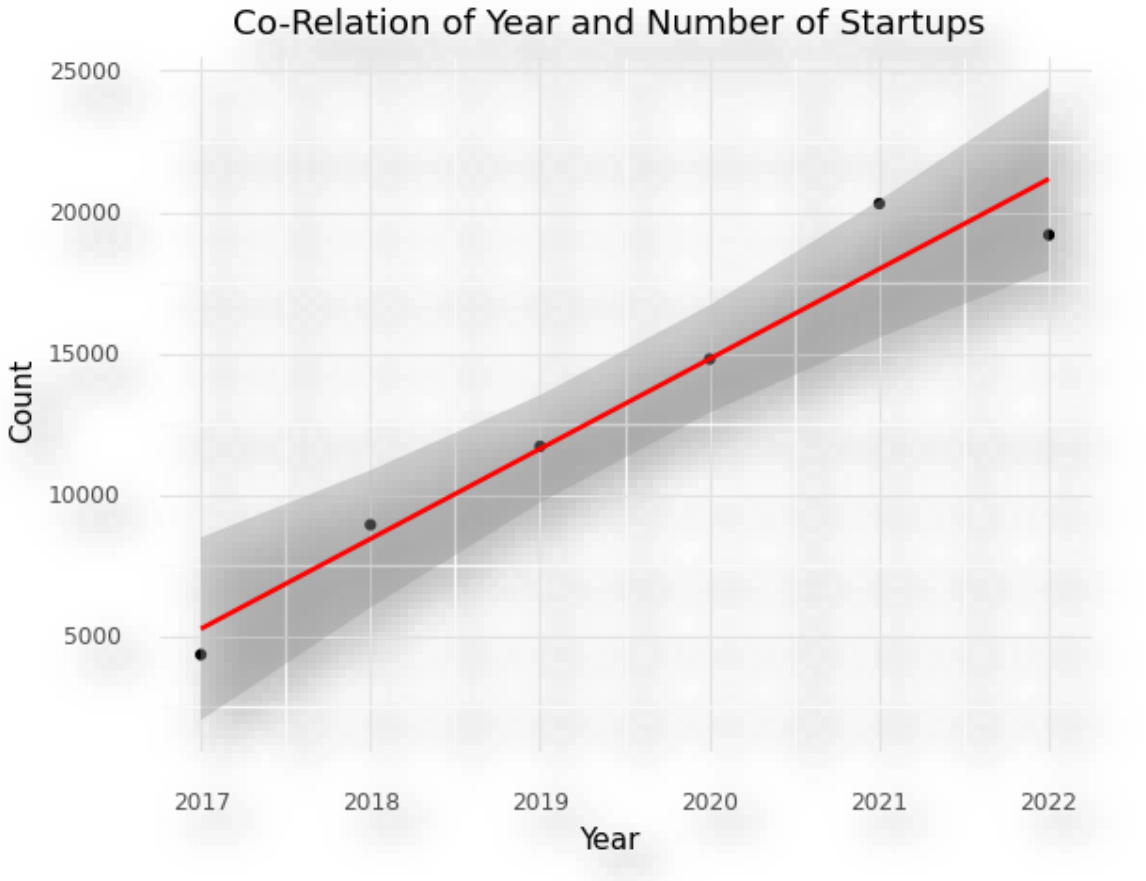

Chloropleth Map of Number of Startups, State-wise
14000
12000
10000
8000
6000
4000
2000
0
Number of Startups

CO RELATION with YEAR and NUMBER of STARTUPS:

To find the trend of number of startups vs year wise growth a new dataframe is formed as follows:

	Year	Count
0	2017	4352
1	2018	8944
2	2019	11718
3	2020	14806
4	2021	20303
5	2022	19196

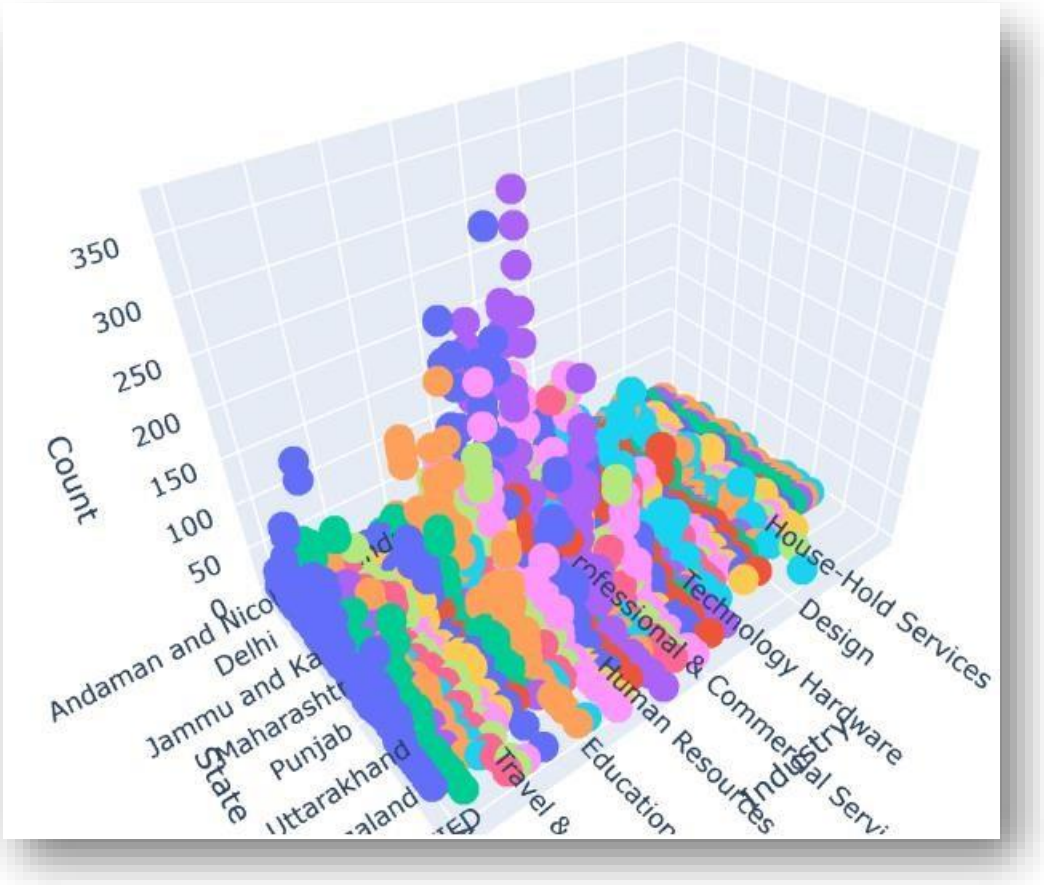
This is then used to plot the number of startups occurring yearly as follows:



It was seen almost perfect positive correlation between year and startup count indicating the number of Startups will definitely increase year wise.

3D SCATTER PLOT:

An attempt to visualize all three data i.e. State, Industry, Count is carried out with the Plotly interactive library whose output is as follows:



This interactive visualization targets to give each count based on Industry and State which is better viewed in an interactive interface.

SECONDARY DATA PROCESSING:

Any startup that receives a valuation of $1 billion is referred to as a unicorn in the venture capital sector. Being a unicorn is not easy, and each one today has a unique history and a unique set of characteristics that have benefited it. Such list of 102 Startup data is imported in python to process the conversion rates of the startups. The dataframe looks as follows:

	No	Company	Sector	Entry Valuation^^ ($B)	Valuation ($B)#	Entry	Location	Select Investors
0	102.0	Molbio Diagnostics	Healthtech - Diagnostics	1.53	1.53	Sep-22	Goa	Temasek, Motilal Oswal
1	101.0	Shiprocket	Aggregator - Logistics Services	1.23	1.23	Aug-22	Delhi	Lightrock India, Info Edge, Tribe Capital, Tem...
2	100.0	OneCard	Fintech - Credit Cards	1.30	1.30	Jul-22	Pune	QED Investors, Matrix Partners India, Sequoia ...
3	99.0	Leadsquared	SaaS - CRM	1.00	1.00	Jun-22	Bangalore	Stakeboat Capital, Gaja Capital, WestBridge
4	98.0	Purplle	E-Commerce - Personal Care & Cosmetics	1.10	1.10	Jun-22	Mumbai	JSW Ventures, IvyCap Ventures, Blume Ventures,...
...	...	...	...	...	...	...	...	...
97	5.0	PayTM^	Fintech - Payments & Wallet	1.70	16.00	Feb-15	Noida	Saama Capital, Elevation Capital, Alibaba, Ber...
98	4.0	Snapdeal*	E-Commerce	1.80	2.40	Oct-14	Delhi	Kalaari Capital, Nexus Ventures, Bessemer, Sof...
99	3.0	Mu Sigma	SaaS - Analytics	1.00	1.50	Feb-13	Bangalore	Accel, Sequoia Capital, General Atlantic
100	2.0	Flipkart^	E-Commerce	1.00	37.60	Feb-12	Bangalore	Accel, Tiger Global, Naspers, SoftBank, Tencent
101	1.0	InMobi	Adtech - Mobile Ads	1.00	1.00	Sep-11	Bangalore	KPCB, Sherpalo Ventures, SoftBank

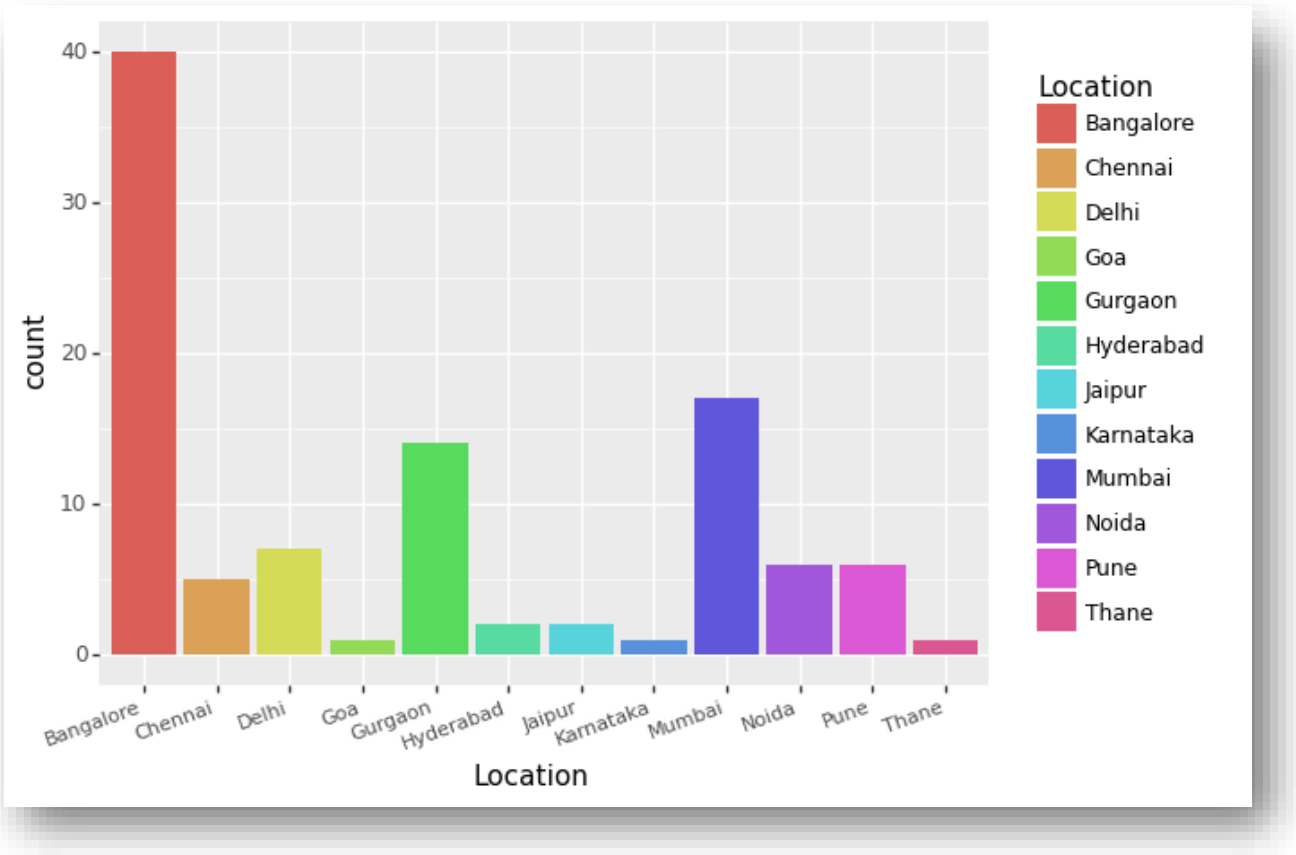
WORDCLOUD of BIGGEST INVESTORS:

The biggest supporters of any startup to become a unicorn is the Investors, among many names in the data frame some of the biggest recurring investors and all others are visualized as below:

As clearly seen Venture, Tiger Global are among the top investors in Unicorn firms, so next time you know where to knock if you need some funding for your startup.

CITY WISE DISTRIBUTION of UNICORNS:

A plot of Unicorns across the country was plot against their count. This was achieved by plotting Location vs its count in the dataframe.
The output is as displayed:



This brings us to the end of the visualization part of the data, as enough data is present to draw some conclusions out of the data.

CONCLUSIONS and INFERENCES:

The following conclusions are drawn from the above project:

- 'IT' sector is the most favorite for startups (9886) followed by Healthcare (7423)
- The year 2021 has most number of startups as of yet
- Maharashtra State has the greatest number of startups (18.5%) out of all the total startups, followed by Karnataka (11.9%) and Delhi (11.5%)
- The number of startups show a clear increase over the years which will in turn create more job opportunities for the people.
- Though % of Startups is more in Maharashtra, the number of unicorns converted are more in state of Karnataka(40) as compared to Maharahstra (26).

REFERENCES:

Primary Data: data.gov.in

DPIIT: Department for Promotion of Industry and Internal Trade

https://www.startupindia.gov.in/content/sih/en/search.html?roles=Startup&page=0

https://www.startupindia.gov.in

Secondary Data:

https://www.ventureintelligence.com/Indian-Unicorn-Tracker.php