INTRODUCTION

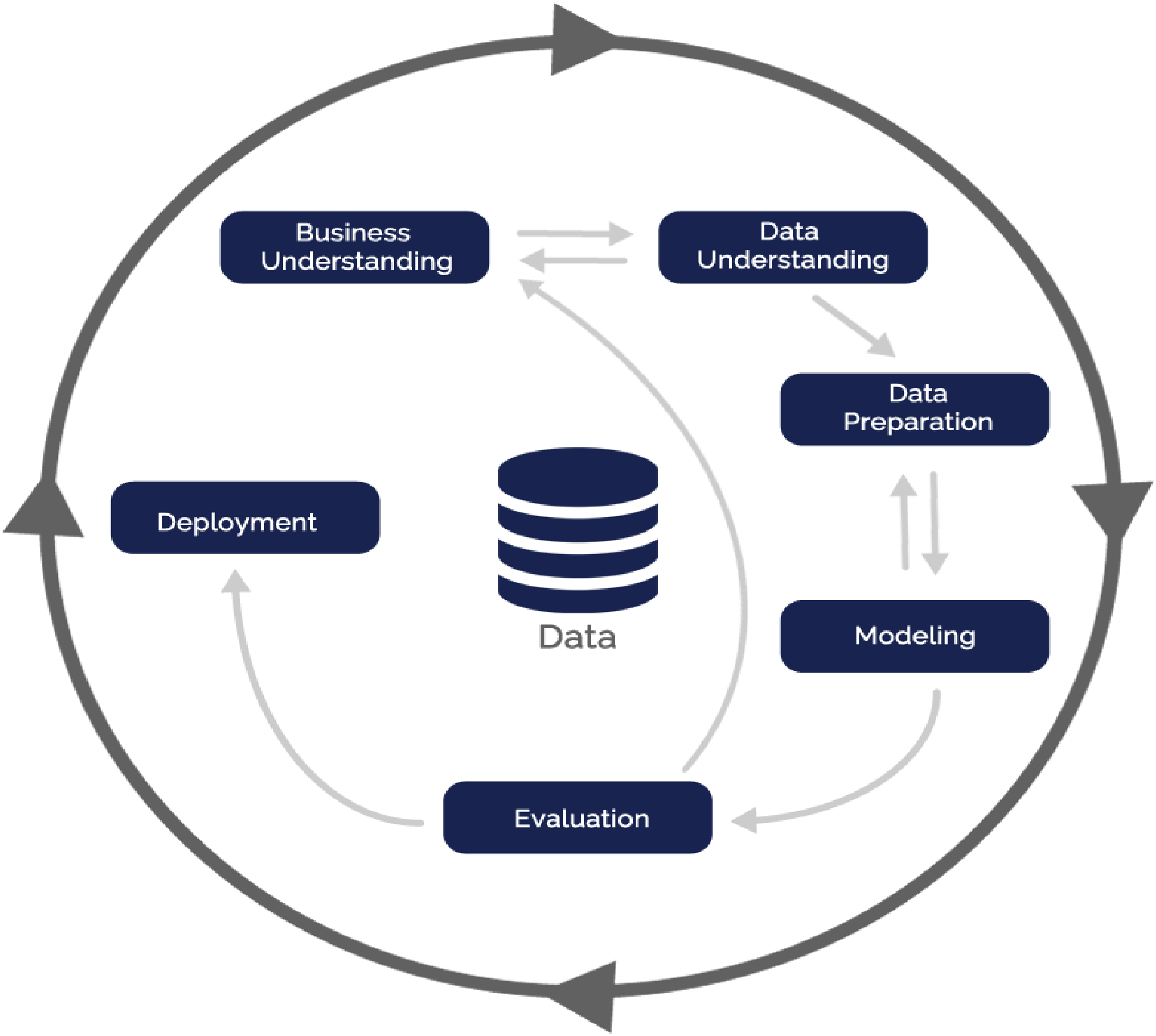
Using data to make informed decisions is the best decision. Every day the amount of data available grows exponentially .as a result, effective interpretation is more important than ever. Data analytics is quickly becoming one of the world's most exciting and rewarding career paths. Business analytics skills will most likely be in higher demand over the next decade than any other career (10.9% vs 5.2%). career of labor statistics).

Companies all over the world require qualified data analysts to solve problems and assist them in making the best business decisions possible. Currently ,59% of companies intend to add even more positions requiring data analysis skills (source: SHRM).

In this article, I provide a summary of my investigation into and analysis of funding for Indian start-ups from 2018 to 2021. I have been assigned the responsibility of studying the Indian start-up ecosystem and suggesting a suitable plan of action for our team's development. To accomplish this task, I will create a distinctive narrative based on the 2018 to 2021 Indian start-up datasets. This will involve formulating and evaluating a hypothesis, preparing research questions, conducting analysis, and presenting insights through appropriate visualizations.

The goal was to use the data to explore the ecosystem, uncover insights, and recommend the best course of action to my fictional team that is attempting to enter the Indian start-up ecosystem. I was also prepared to use the opportunity to develop and harness data analytic skills. By the time I am done with my investigations into the Indian start-ups from 2018 to 2021, I hope to have made smart and strategic decisions which is data driven. I hope to have asked the right questions, summarized data, connected business objectives to data analysis, identified and cleaned the data, created visualizations, and above all I hope to have told a data driven story.

Every data science or data analytics project follows a certain kind of data science process. Scrum, Kanban, and Agile are all methods that data science teams adopt to complete their projects but, in this project, I will be working with CRISP-DM.



I used the cross industry standard process for data mining (CRISP-DM) model as the base for my data science process.it has six sequential phases:

1. Business understanding - what does the business need?
2. Data understanding – what data do we have/ need? Is it clean?
3. Data preparation – how do I organize the data for modelling?
4. Modelling- what modelling techniques should I apply?
5. Evaluation – which model best meets the business objectives?

6. Deployment- how do stakeholders access the results?

THE DATA

The data for each year of funding was in a separate CSV file, meaning they had to be put together at a point. The shared columns in the files were as follows (column name: description):

- Company/Brand: name of the company/start-up
- Founded: the year the start-up was founded
- Sector: sector of operation
- About Company/What it does: description of the company
- Founders: founders of the company
- Investor: investors in the deal
- Amount ($): raised funds
- Stage: round of funding
- Headquarters: headquarters of the company

ASK STAGE

At this stage, we bring the objective into view and put down the questions that we intend to answer at the end of the analysis process. The first phrase of the data analysis process is asking the right questions.

Here, with the overarching goal of making a recommendation to the team to assist their goal of entering the Indian startup ecosystem, the whole picture was considered to make sure we get the situation right. The following hypothesis was stated and questions were asked to guide the analyses.

HYPOTHESIS

Null hypothesis: Fintech is the most lucrative sector receiving the most significant funding in the Indian start-up ecosystem.

Alternative hypothesis: Fintech is not the most lucrative sector as it is not receiving the most significant amount of funding in the Indian start-up ecosystem.

RESEARCH QUESTIONS

1. Which sector received the most funding from 2018 to 2021?
2. How much funding has the fintech sector received in the Indian start-up ecosystem over the same period, compared to other sectors?
3. Have there been any significant fintech funding deals in the Indian start-up ecosystem from 2018 to 2021?
4. What are the top 10 sectors that had the most funding over the same period?
5. How many fintech start-up companies received funding from 2018 to 2021 compared to other companies in the top 10 sectors?
6. How does the average amount of funding received by fintech start-ups compare to the average amount of funding received by start-

ups in other sectors in the Indian start-up ecosystem over the same period?

7. Have the investment trends in the Indian start-up ecosystem shown a preference for fintech companies over other sectors from 2018 to 2021?

DATA PREPARATION AND PROCESSING

Here I organize the data to make it fit for analysis. Cleanliness and consistency of data are the objectives here.

I plan to analyze data from 2018 to 2021 to address the research questions, so it is necessary to merge the datasets. So, I took a quick glance at the data in excel and found out that, the column names in the 2018 datasets do not match those in 2019, 2020, and 2021 datasets, and the 2018 dataset also has fewer columns. Additionally, some of the values in the "Amount" column of the 2018 dataset are in rupees instead of dollars like the rest of the datasets.

To resolve these inconsistencies before combining the datasets, I will first adjust the column names of the 2018 dataset to align with the other datasets. After merging the datasets, any missing columns will be filled with null values. Finally, I will convert the rupee values in the "Amount" column to dollars.

LOADING PACKAGES

To start with, the basic packages for analysis were loaded into my jupyter notebook. These packages were:

Pandas: for data cleaning and manipulation

NumPy: for data cleaning and manipulation

Glob: a module that has several functions, that can help in listing files under a specified folder.

Matplotlib: visualization tool

NOTES FROM PREVIEWING THE DATA FRAMES

The individual datasets were then loaded as Pandas Data Frames

The 2018 data

1. The Data Frame has 526 rows and 6 columns.
2. Dashes were used in the amounts column for deals whose values were not known.
3. The amounts in the 2018 Data Frame are a mix of Indian Rupees (INR) and US Dollars (USD), meaning they have to be converted into the same currency.
4. The industry and location columns have multiple information. A decision is to be made between selecting the first value before the separator (,) as the main value.

The 2019 data

1. The Data Frame has 90 rows and 9 columns.

2. The data type of the “Founded” column is set to float64. It should be set to a string for uniformity.
3. The headquarter column has multiple information. A decision is to be made between selecting the first value before the separator (,) as the main value or representing that column with a word cloud.

The 2020 data

1. There is an extra column called “Unnamed:9”, giving it a total of 10 columns. It should be dropped to ensure complete alignment with the other Data Frames for ease of concatenation.

The 2021 data

1. The data type of the “Founded” column is set to float64. It should be set to a string for uniformity.

GENERAL NOTES

- The columns in 2018 are different from those of 2019–2021, meaning they have to be renamed before merging for consistency’s sake.
- The currency signs and commas have to be removed from each amount column form all datasets.
- All the columns with amounts have to be set to float.

- All the years of funding and the years founded should be converted to strings.
- The respective years of funding have to be attached to each Data Frame before combining.
- Upon examining the data frame, I discovered that some of the columns, such as "Amount ($)" and "Founded" columns contain values that should be numerical but are currently strings (objects). I will need to convert the datatypes of the values in these columns to numerical (float and/or integer).

ASSUMPTIONS

1. The average Indian Rupee (INR) to US Dollar (USD) rate for the relevant year will be used for currency conversions.
2. The first values of industry and location in the 2018 data are the primary sector and headquarters respectively.
3. Amounts without currency symbols in the 2018 dataset are in USD.
4. Imputations will not be made for undisclosed and/or unavailable (missing) amounts due to the uncertainties, risks of misstatements and possible misleading effects on the analyses.

DATA CLEANING

the major activities performed on the Data Frames with respect to data cleaning are explain below.

The detailed functions will be found in the jupyter notebook, a link to which will be attached at the end of the article.

1. I separated the values in the location and industry columns using a comma as the delimiter and selected the first value in the split column as the primary sector.

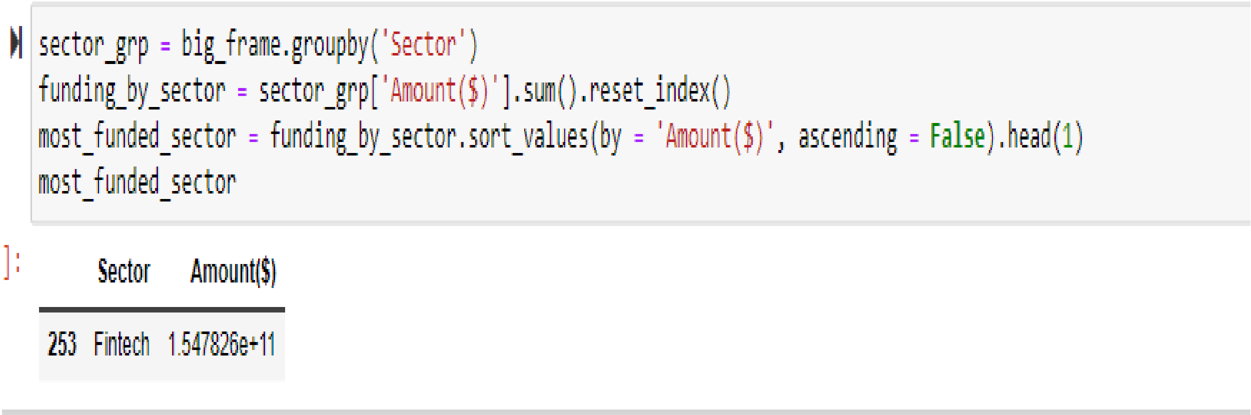
2. I applied string formatting to all columns except the amounts columns which were formatted as numeric.
3. I removed all commas and currency signs from the “Amounts” columns.
4. I replaced the “nan” values in the “Founders” column with nulls.
5. I replaced notable misplaced and/or erroneous values in the respective rows.
6. I dropped the extra unnamed in the 2020 Data Frame.
7. For each Data Frame, I appended a column indicating the year of funding

ANSWERING RESEARCH QUESTIONS

At this point, I combine the analyses and share stages of the data analysis process the coding and visualization of the merged data.

Q1. Which sector received the most funding from 2018 to 2021?

Funding in the Indian tech ecosystem has been on the rise in general. But we sought to understand which sector received the most funding.

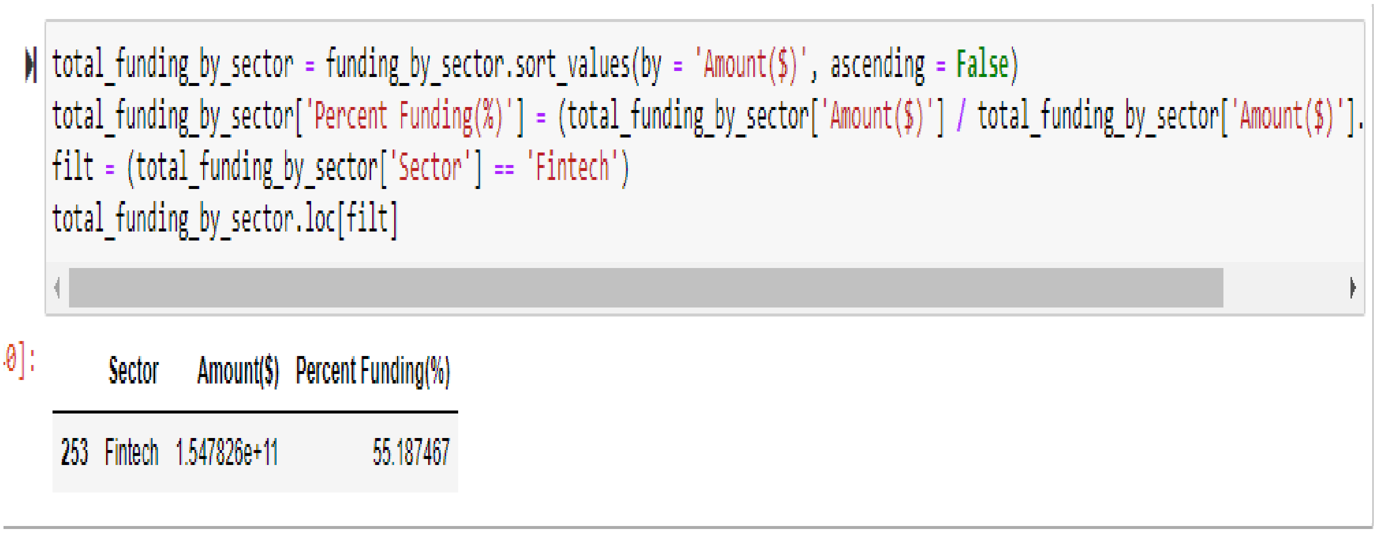


```
sector_grp = big_frame.groupby('Sector')
funding_by_sector = sector_grp['Amount($)'].sum().reset_index()
most_funded_sector = funding_by_sector.sort_values(by = 'Amount($)', ascending = False).head(1)
most_funded_sector
```

```
]:       Sector    Amount($)
    253  Fintech   1.547826e+11
```

In conclusion, I found out that fintech has the most funding from 2018 to 2021.

Q2. How much funding has the fintech sector received in the Indian start-up ecosystem over the same period, compared to other sectors?

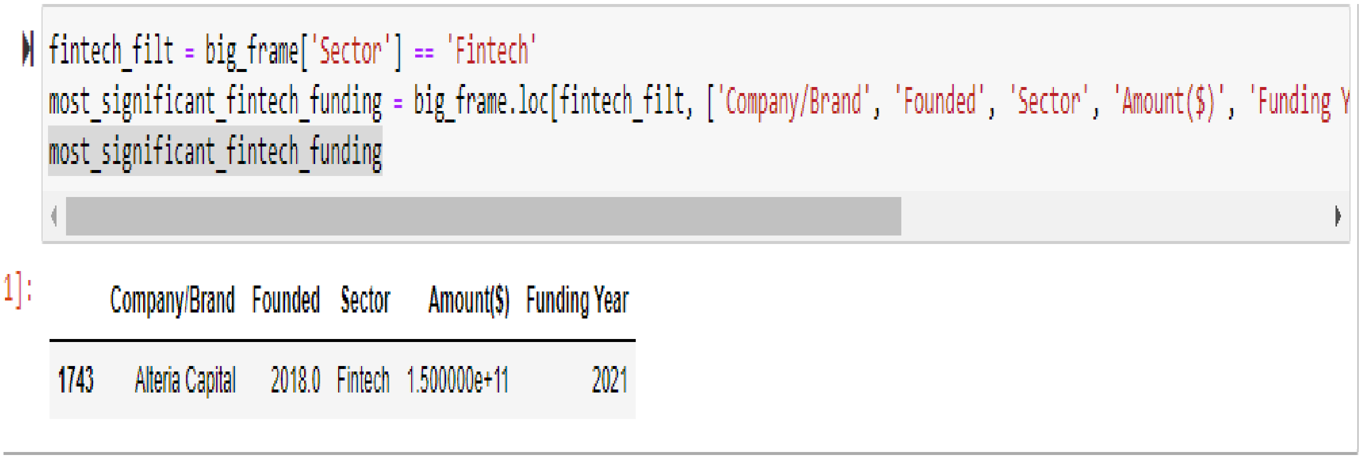


```
total_funding_by_sector = funding_by_sector.sort_values(by = 'Amount($)', ascending = False)
total_funding_by_sector['Percent Funding(%)'] = (total_funding_by_sector['Amount($)'] / total_funding_by_sector['Amount($)'].
filt = (total_funding_by_sector['Sector'] == 'Fintech')
total_funding_by_sector.loc[filt]
```

```
0]:       Sector    Amount($)     Percent Funding(%)
    253  Fintech   1.547826e+11           55.187467
```

I further wanted to explore how much money fintech which happens to be the most funded sector had received during the period of 2018 – 2021.from the picture above fintech received about 55 billion dollars.

Q3. Have there been any significant fintech funding deals in the Indian start-up ecosystem from 2018 to 2021?



```
fintech_filt = big_frame['Sector'] == 'Fintech'
most_significant_fintech_funding = big_frame.loc[fintech_filt, ['Company/Brand', 'Founded', 'Sector', 'Amount($)', 'Funding Y
most_significant_fintech_funding
```

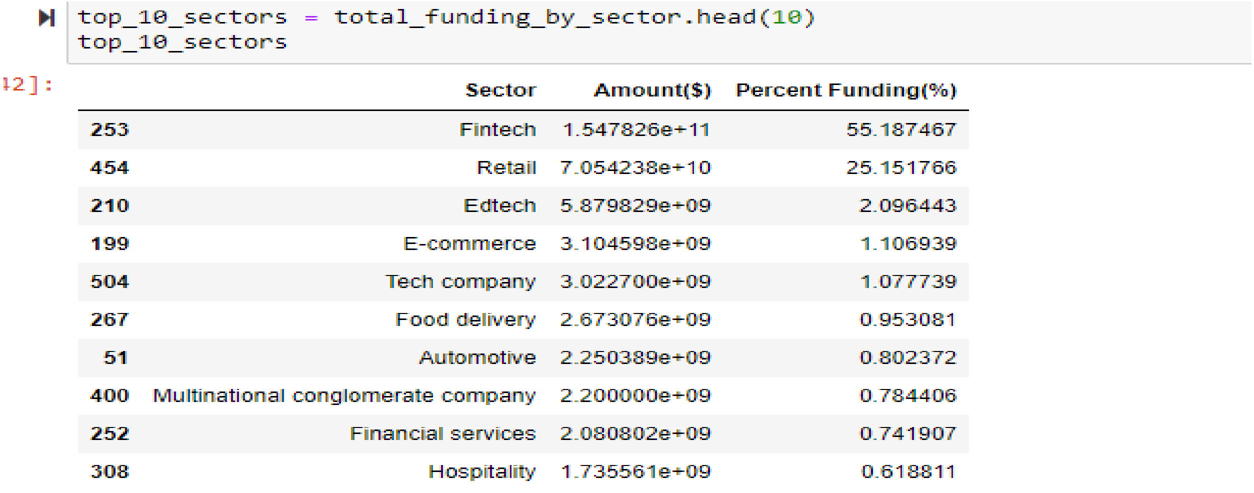
1]:

	Company/Brand	Founded	Sector	Amount($)	Funding Year
1743	Alteria Capital	2018.0	Fintech	1.500000e+11	2021

The fintech company with the most significant funding is Alteria capitals.

Q4. What are the top 10 sectors that had the most funding over the same period?

To get a better picture of what is happening in the Indian startup ecosystem, I decided to look at the top 10 sectors that had the most funding in the same period. What sectors are investors really putting their monies in.



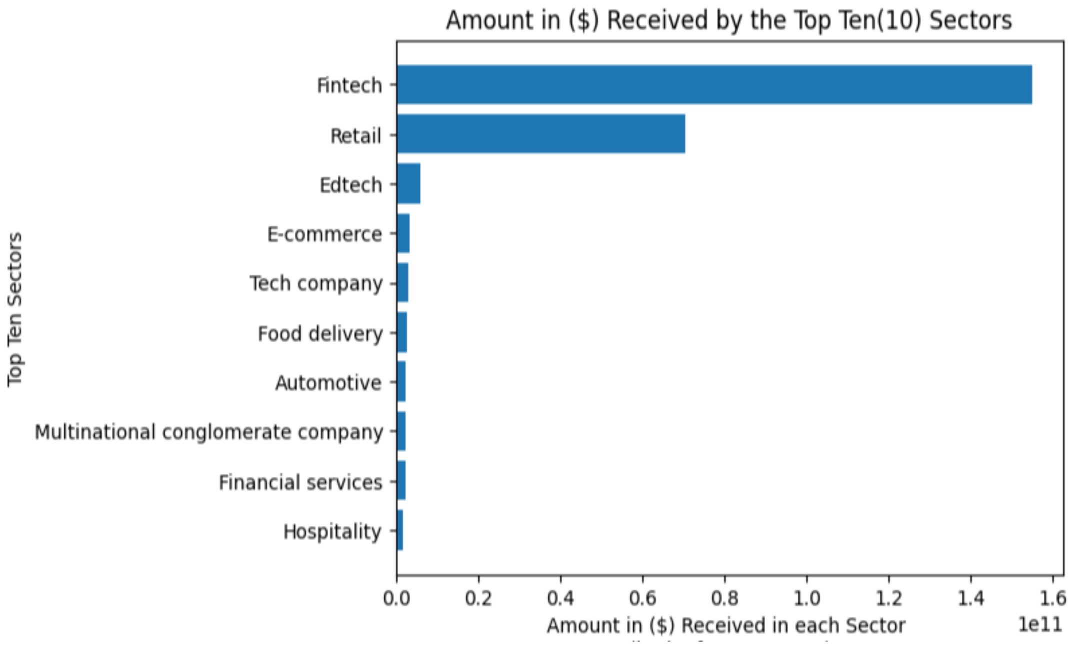
```
top_10_sectors = total_funding_by_sector.head(10)
top_10_sectors
```

12]:

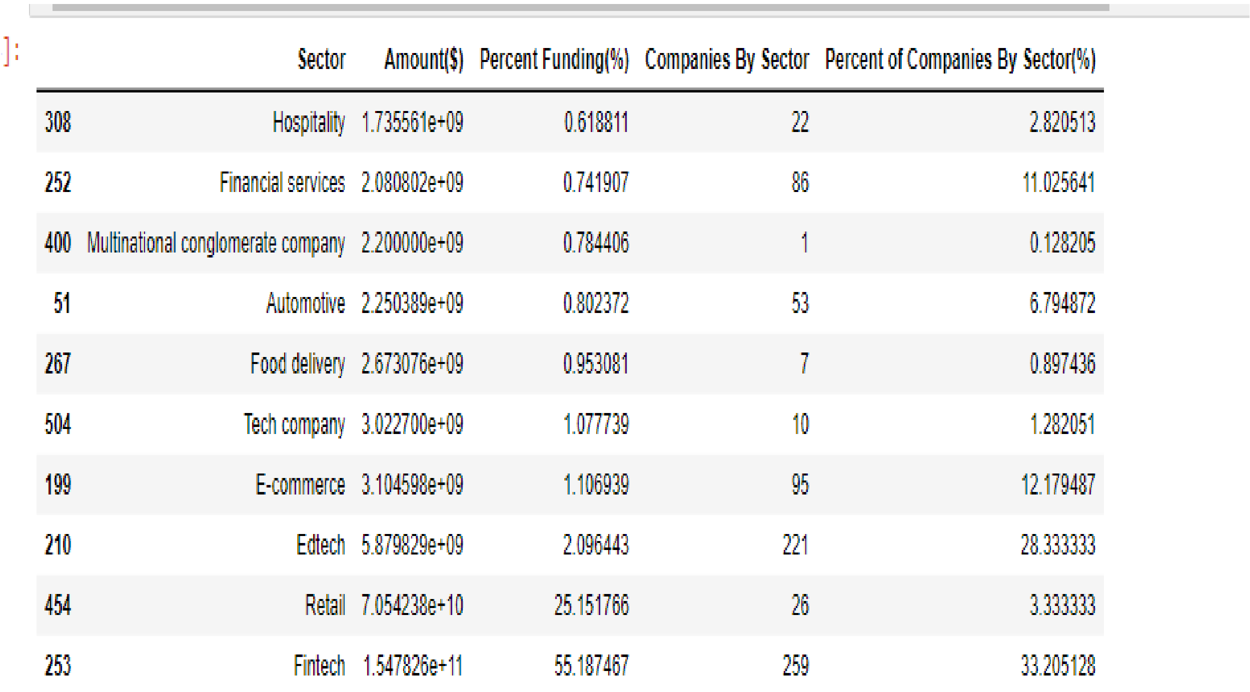
	Sector	Amount($)	Percent Funding(%)
253	Fintech	1.547826e+11	55.187467
454	Retail	7.054238e+10	25.151766
210	Edtech	5.879829e+09	2.096443
199	E-commerce	3.104598e+09	1.106939
504	Tech company	3.022700e+09	1.077739
267	Food delivery	2.673076e+09	0.953081
51	Automotive	2.250389e+09	0.802372
400	Multinational conglomerate company	2.200000e+09	0.784406
252	Financial services	2.080802e+09	0.741907
308	Hospitality	1.735561e+09	0.618811

From the image above, it is clear that fintech has the highest funding percentage wise; fintech having a lion share with about 55 percent whereas hospitality has the lowest with about 0.6 percent of total funding.

Below is a **graphical representation**

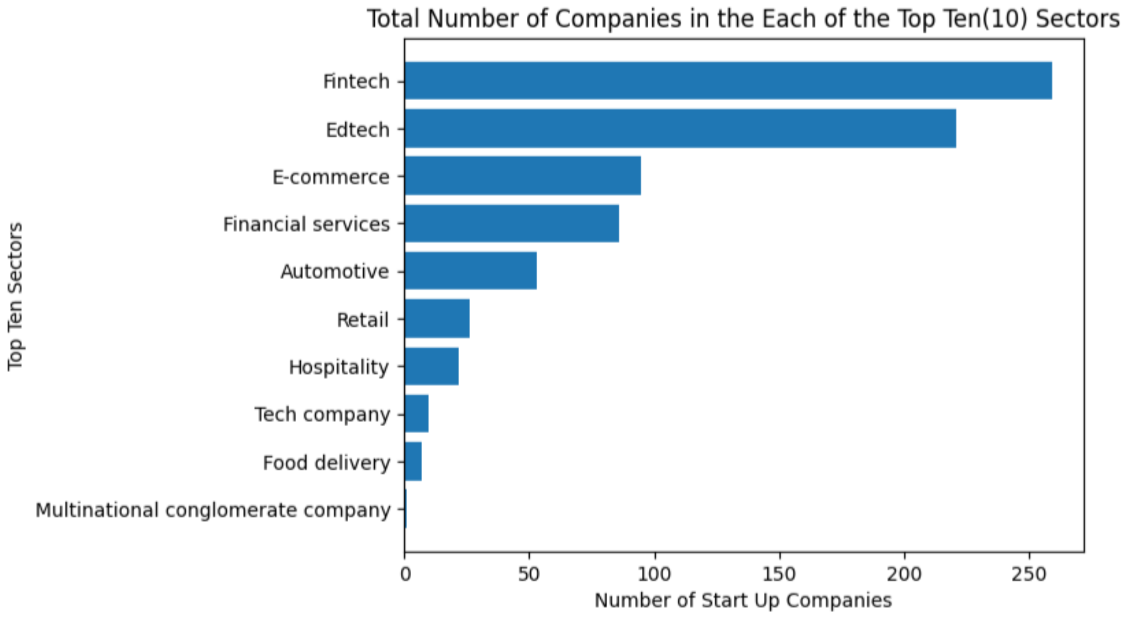


Q5. How many fintech start-up companies received funding from 2018 to 2021 compared to other companies in the top 10 sectors?

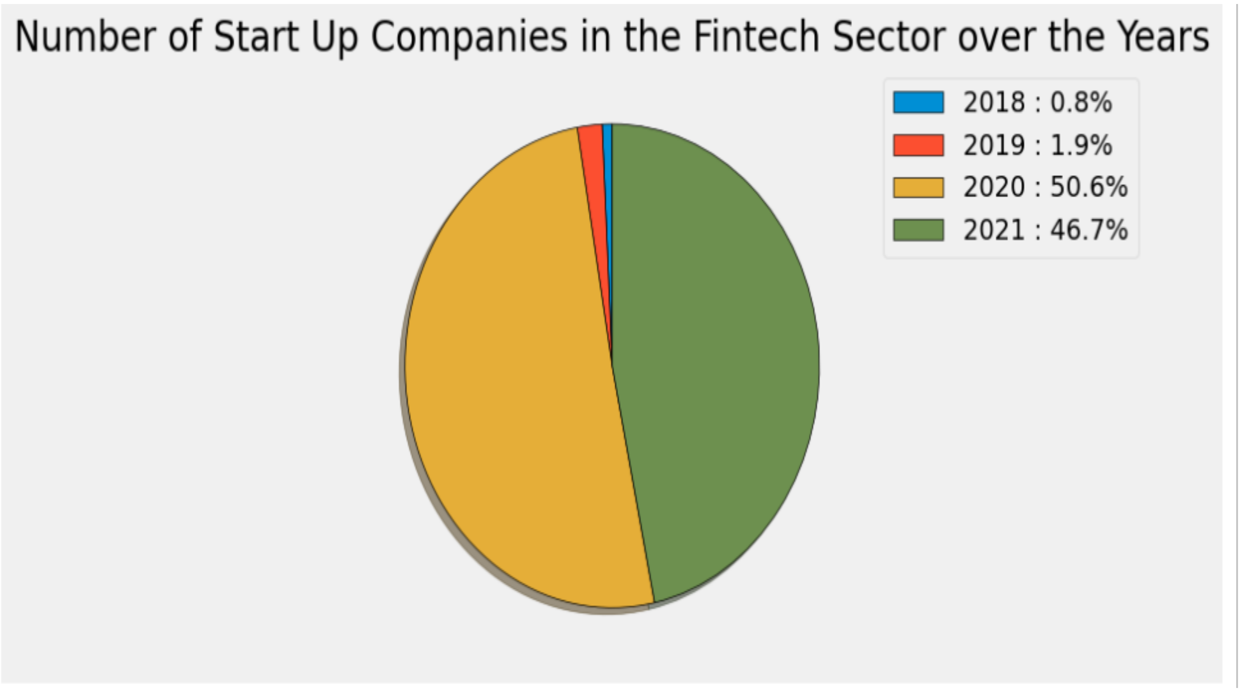


	Sector	Amount($)	Percent Funding(%)	Companies By Sector	Percent of Companies By Sector(%)
308	Hospitality	1.735561e+09	0.618811	22	2.820513
252	Financial services	2.080802e+09	0.741907	86	11.025641
400	Multinational conglomerate company	2.200000e+09	0.784406	1	0.128205
51	Automotive	2.250389e+09	0.802372	53	6.794872
267	Food delivery	2.673076e+09	0.953081	7	0.897436
504	Tech company	3.022700e+09	1.077739	10	1.282051
199	E-commerce	3.104598e+09	1.106939	95	12.179487
210	Edtech	5.879829e+09	2.096443	221	28.333333
454	Retail	7.054238e+10	25.151766	26	3.333333
253	Fintech	1.547826e+11	55.187467	259	33.205128

The total number of companies in each of the top 10 sectors are:



Q6. Number of startups companies in the fintech sector over the years?



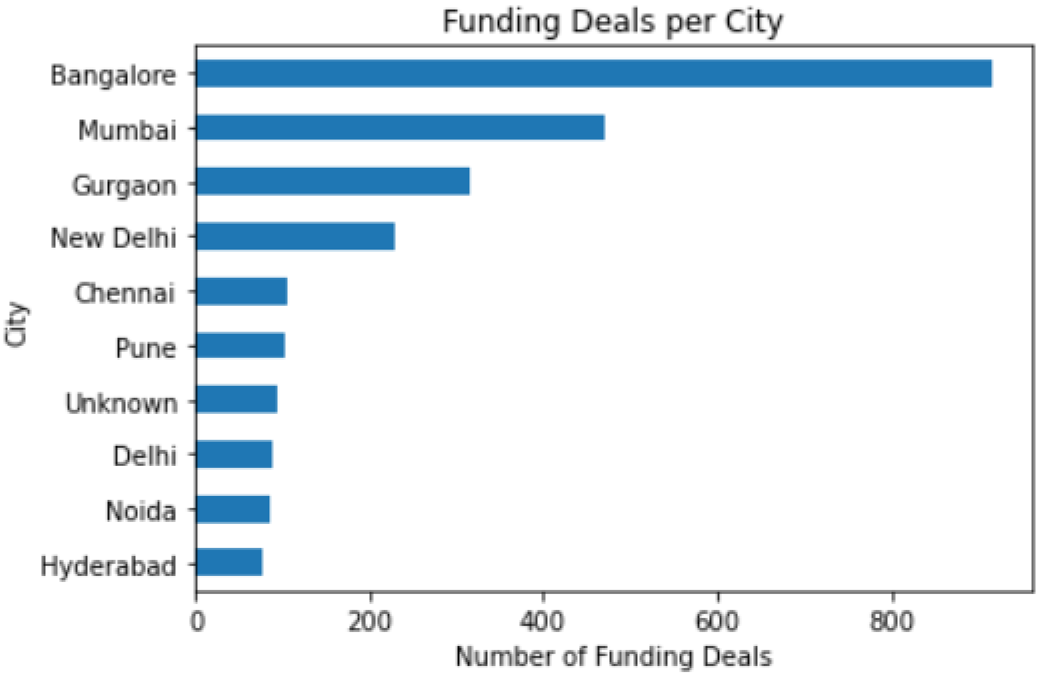
The number of startups in the fintech sector has increased sharply over the years with a slightly significant drop in the year 2021(a difference of 3.9 from 2020-2021). It is therefore safe to conclude that both the number of deals and funding received by startups in the fintech sector increased over the period of years.

ADDITIONAL ANALYSIS

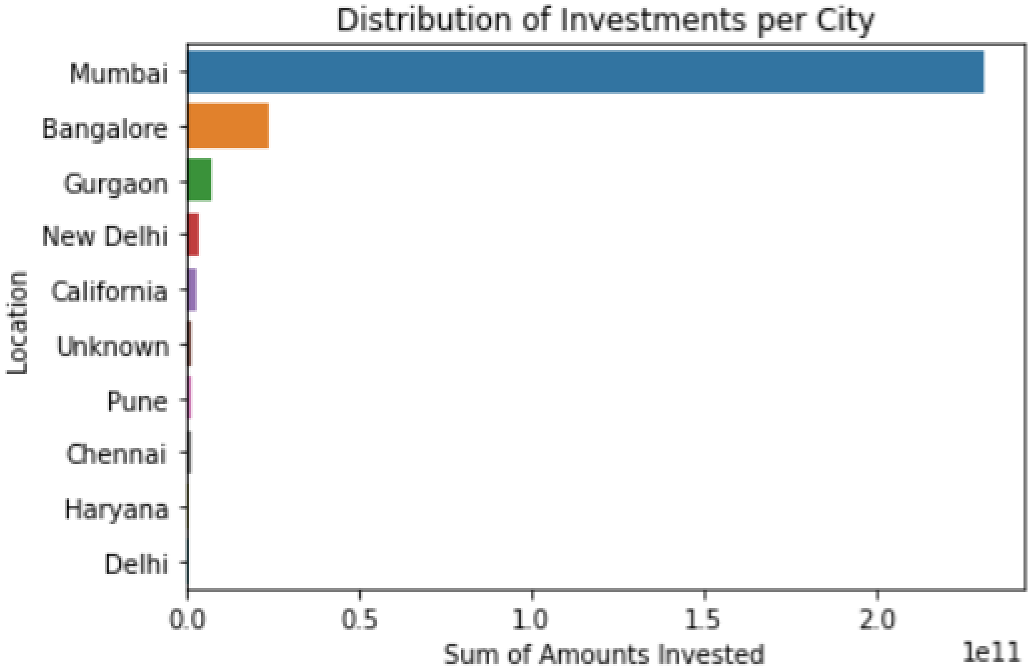
How does location affect funding to startups?

In terms of total number of deals for startups headquartered in the various locations, Bangalore (916), Mumbai (471), Gurgaon (317), and

New Delhi (230) made up the top 5. A visual representation can be seen below:



By amount of funding received by startups headquartered in the locations, Mumbai came first (USD 230.8bn), followed by Bangalore (USD 24bn), Gurgaon (USD 6.9bn), and New Delhi (USD 3.4bn) in that order. A visual representation of the gap between funding to startups headquartered in the top 10 locations is shown below:

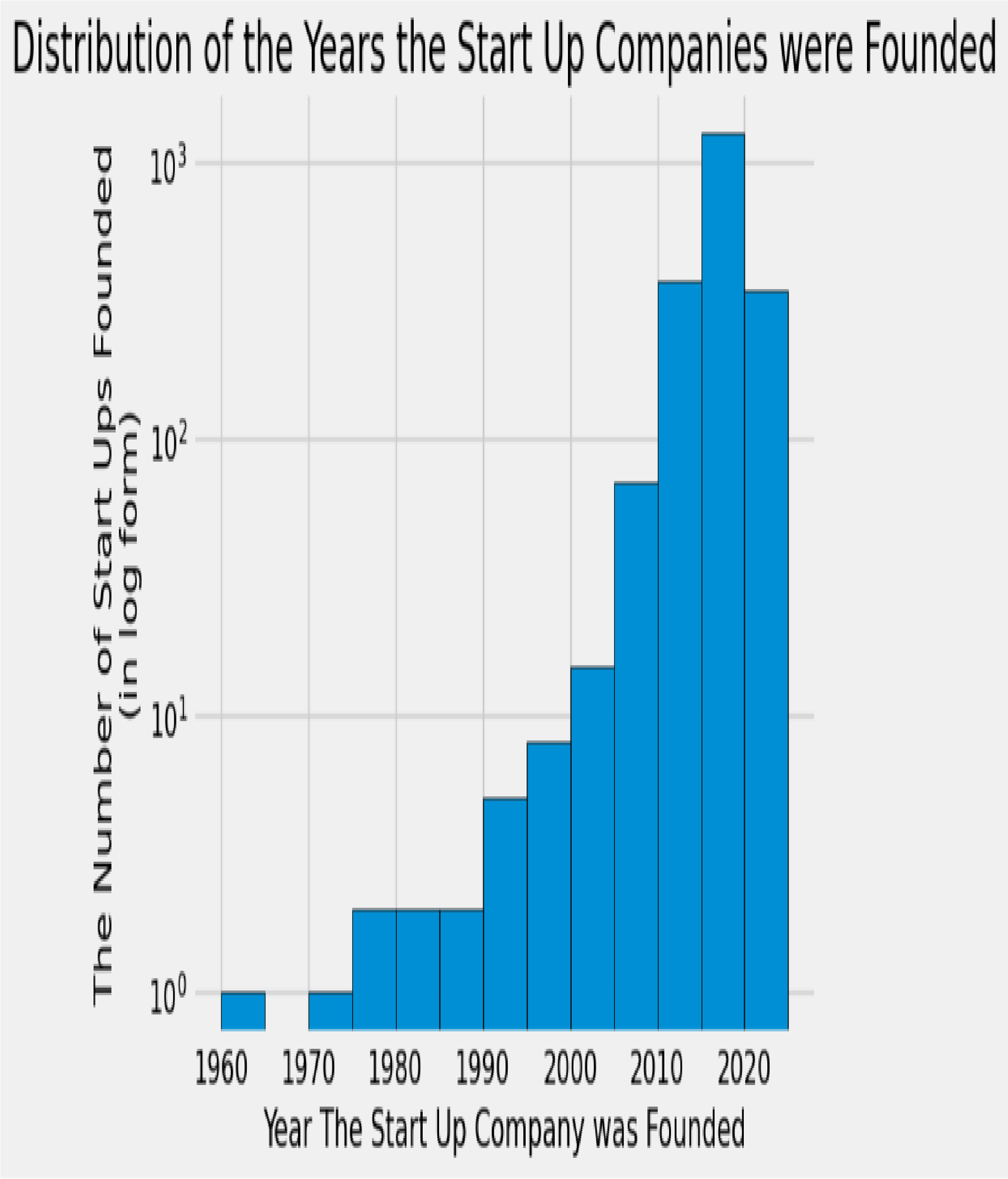


By the sheer amounts involved in deals for startups headquartered there, Mumbai (82.5%) and Bangalore (8.6%) alone took up about 91% of the total funding received by Indian startups over the period. This is a clear case of centralization of funding around specific locations.

Biju’s (10), Bharatpe (10), and Zomato (7) were involved in the greatest number of deals over the period.

Alteria Capital (USD 150bn), Reliance Retail Ventures (USD 70bn), and Snowflake (USD 3bn) made up the top 3 in funding received.

These 6 firms can thus be said to be the startups most favored by investors over the period.



CONCLUSION

In conclusion, the null hypothesis that Fintech is the most lucrative sector receiving the most significant funding in the Indian start-up ecosystem has been proven through analysis of funding data and research on the Indian start-up ecosystem datasets. We accept the null hypothesis

FINAL NOTES

Thank you so much for reading.

I will be grateful for your comments, advice, suggestions, and recommendations. Too long or too short? Too detailed or missing some details? Please let me know. You can leave a comment here or find me on Twitter (@arku_laryea).

I’m also open to collaborating on projects.

Link to the project repository on GitHub: https://github.com/arkularyea/indian-startup-funding-analysis.git

LINKEDIN : https://www.linkedin.com/pulse/exploratory-analysis-indian-startup-from-2018-2021-nii-laryea