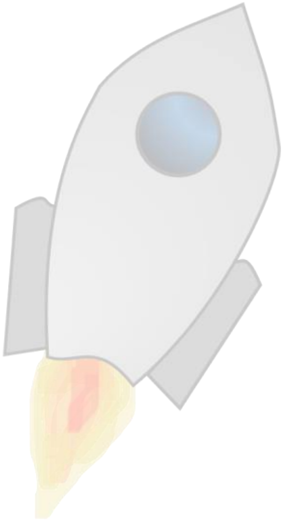
Understanding Global AI Startup Market

Radhika Patil

(SUID: radhikap)

CS 102: Working with Data – Tools and Techniques (spring 2020)

5/18/2020



Why study AI startup market

Artificial intelligence and machine learning have become the booming buzz words for todays tech market. Over last five years, technological development has rocketed by simply adding the AI technology to everyday life activities. New opportunities are being created for automating tasks that would otherwise need human involvement. Some of the most bizarre ones are systems that can answer phone calls like a human.

With this growing interest also come opportunities for entrepreneurs to start new ventures and we have seen a sharp rise in the number of AI and ML startups over the past decade. Venture capitalists are constantly on lookout for the next Google, Facebook or Amazon, the startup that would disrupt the future technological progress.

Given this, it is very interesting to study and understand the market scenario of AI startups, booming hubs, historical progress and funding situations faced when you start a new venture.

Goal

The goal of this exercise is to understand the global AI startup market, study trends and identify key insights and patterns using data processing and visualization techniques.

Dataset

For this purpose I use the crunchbase extracted dataset in the git repository https://github.com/notpeter/crunchbase-data

The dataset can be downloaded from: https://github.com/notpeter/crunchbase-data/archive/master.zip

Questions

Using the above dataset, I aim to answer following questions.

1. Which countries have most startups? Rank the countries in order of number of startups
2. For startups in USA, rank cities/states in order of number of startups. Which city has most startups? What percentage of startups are in top five cities?
3. Where are the highest funded startups located?
4. How many rounds are required for a startup to acquire funding on an average?
5. Which is the most popular sector for startups?
6. What percentage of startups are acquired? Closed? Operating?
7. How did emergence of startups change with time?
8. How much average funding is provided for startups? Minimum? Maximum?
9. What is the average funding for each sector?
10. What are the regional/country variations in average funding?
11. What are the top five highest funded startups? What is their current status?
12. Can I predict what funding a new startup could possibly acquire?

Analysis

For analyzing the data, I used a combination of following tools

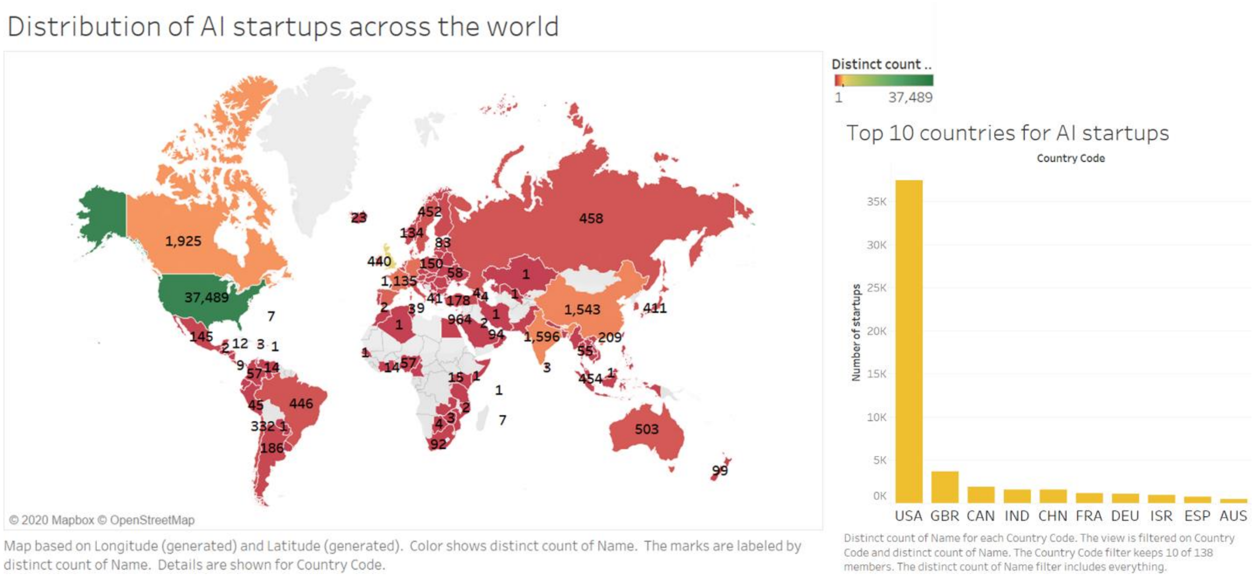
- Python – Jupyter notebook, pandas and matplotlib for data processing, analysis and partly visualization
- Tableau – Visualizations
- Excel spreadsheets – Data pre-processing and visualizations

(Full list of files used is provided at the end of the document)

I started with going through the raw data files in the dataset folder. I used the company.csv as my primary dataset. I removed first column from the file and added another column parsing the year of first funding in it using excel spreadsheets. I also used filters to find unusual characters in the file and deleted them (like hyphens in blank cells) as they could not be read in pandas dataframe. I saved this as companies_v2.csv. Then I loaded companies_v2.csv and acquisitions.csv in ipython jupyter notebook and processed and queried the data using pandas. I implemented a preliminary KNN algorithm in the ipython notebook to predict funding values for new startups. I used matplotlib to plot one of the pie charts. At query steps I saved the data for relooking. I used the original companies.csv to load in tableau and created most of the visualizations. I loaded the saved csv files from ipython notebook in excel and created visualizations like pie chars and some bar charts. The visualizations and results are elaborated below.

Geographical distribution

The companies dataset contains 66368 different startups. Visualizing their locations gives us a picture of the booming AI startup market hubs. I use tableau for creating these visualizations by counting the number of companies in each case. Figure 1 shows the data plotted on a geographic map. We see that 56% of all the startups listed are located in United States. And about 70% of the listed startups come in one of top 5 countries, USA, Britain, Canada, India or China.



Distribution of AI startups across the world

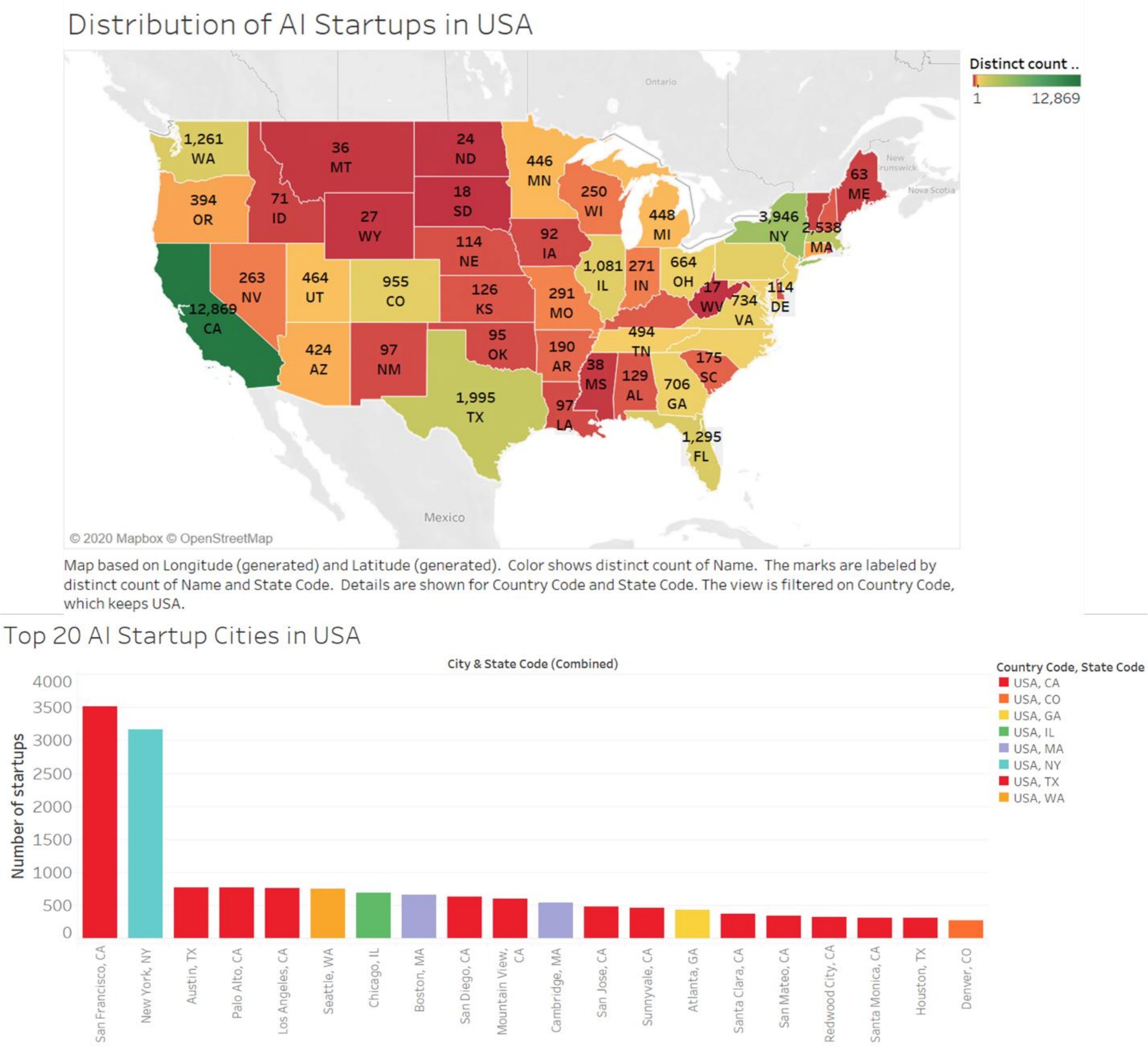
Map based on Longitude (generated) and Latitude (generated). Color shows distinct count of Name. The marks are labeled by distinct count of Name. Details are shown for Country Code.

Top 10 countries for AI startups

Distinct count of Name for each Country Code. The view is filtered on Country Code and distinct count of Name. The Country Code filter keeps 10 of 138 members. The distinct count of Name filter includes everything.

Figure 1. Global distribution of startups in the dataset. USA ranks highest in number of startups followed by Britain, Canada, India, China, France and Germany. The stark contrast of number of startups can also be seen in the bar chart

A closer look at United States is shown in Figure 2. We see similar to global scenario, California ranks highest in number of startups followed by New York, Massachusetts, Texas, Florida and Washington. These six states hold 64% of the startups. A bar chart showing city-wise distribution shows that 4 out of the top 5 and 13 out of top 20 cities are located in California. Also, around 26% of startups in USA are located in top 5 cities and 32% in top 10 cities.



Map based on Longitude (generated) and Latitude (generated). Color shows distinct count of Name. The marks are labeled by distinct count of Name and State Code. Details are shown for Country Code and State Code. The view is filtered on Country Code, which keeps USA.

Distinct count of Name for each City & State Code (Combined). Color shows details about Country Code and State Code. The view is filtered on Country Code, Inclusions (City,Country Code,State Code) and Exclusions (City,Country Code,State Code). The Country Code filter keeps USA. The Inclusions (City,Country Code,State Code) filter keeps 25 members. The Exclusions (City,Country Code,State Code) filter keeps 5,908 members.

Figure 2. USA distribution of startups in the dataset. Additionally, there are 18 startups in Alaska, not shown in map. California ranks highest, followed by New York, Massachusetts and Texas. But In a city wise distribution, cities like Austin and Seattle take higher places.

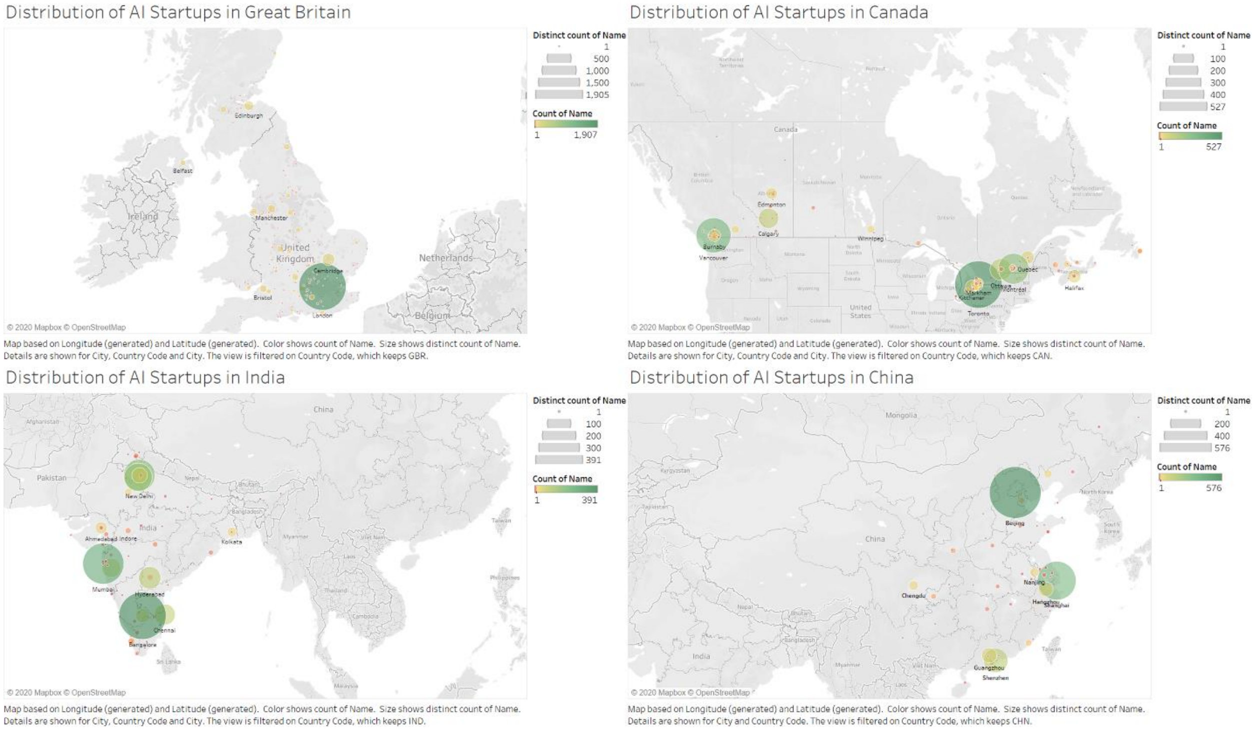
A similar analysis for other four of the five countries is shown in Figure 3. We see that startup concentrations are predominantly in the big metropolitan cities in all of them (not necessarily in sorted order) –

Great Britain – London, Cambridge Manchester, Edinburgh etc.

Canada – Toronto, Vancouver, Quebec, Ottawa etc.

India – Bangalore, Mumbai, Delhi, Hyderabad, Chennai etc.

China – Beijing, Shanghai, Shenzhen, Guangzhou etc.



Distribution of AI Startups in Great Britain

Map based on Longitude (generated) and Latitude (generated). Color shows count of Name. Size shows distinct count of Name. Details are shown for City, Country Code and City. The view is filtered on Country Code, which keeps GBR.

Distribution of AI Startups in Canada

Map based on Longitude (generated) and Latitude (generated). Color shows count of Name. Size shows distinct count of Name. Details are shown for City, Country Code and City. The view is filtered on Country Code, which keeps CAN.

Distribution of AI Startups in India

Map based on Longitude (generated) and Latitude (generated). Color shows count of Name. Size shows distinct count of Name. Details are shown for City, Country Code and City. The view is filtered on Country Code, which keeps IND.

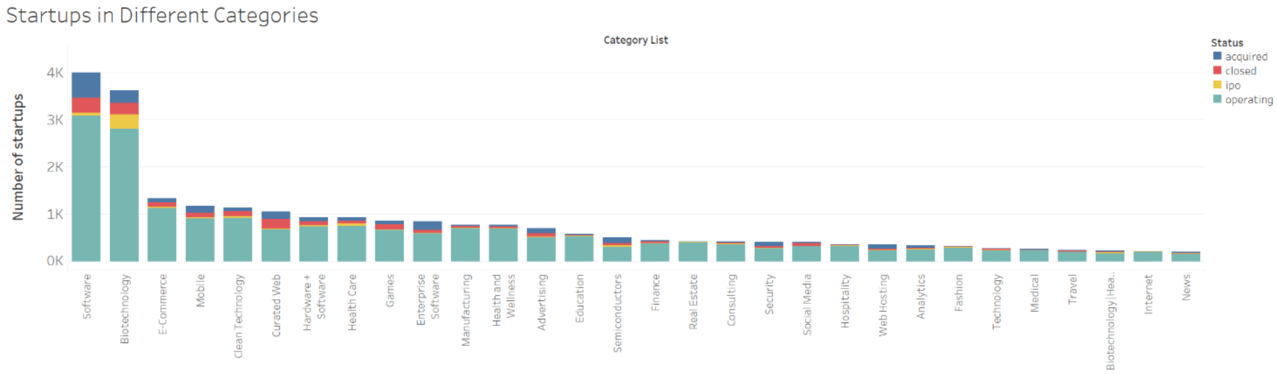
Distribution of AI Startups in China

Map based on Longitude (generated) and Latitude (generated). Color shows count of Name. Size shows distinct count of Name. Details are shown for City and Country Code. The view is filtered on Country Code, which keeps CHN.

Figure 3. Distribution of startup in Great Britain, Canada, India and China. Majority of companies are concentrated in the metropolitan cities, London, Toronto, Bangalore, Beijing etc.

Status/Categorical distribution

Categorically segregated data points out that Software sector is the most popular for AI startups. However, a lot of startups characterized in other sectors could be working on softwares related to that category. So, this does not provide accurate information on categorical distribution.



Distinct count of Name for each Category List. Color shows details about Status. The data is filtered on Action (Country Code,State Code), which keeps 918 members. The view is filtered on Category List, which keeps 30 members.

Figure 3. Top 30 popular categories for startups. Topmost is Software followed by Biotechnology, E-commerce and Mobile. The categories are also split between the status or operating condition of the companies.

To understand how many startups end up successful or unsuccessful, I look at classification according to their operating status. Table 1 shows the data and Figure 4 shows pie chart for it. We see that about 80% of the startups are still operating, but only 2.3% have been able to go public and release an IPO.

Table 1. Number of startups grouped by their operating status.

Status	Number of startups
acquired	5549
closed	6237
IPO	1547
operating	53034

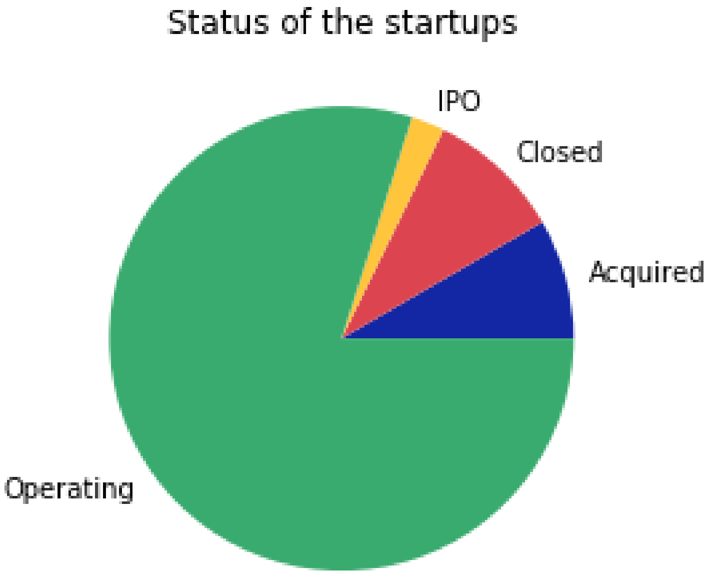


Figure 4. Percentage of startups with operating status. 80% are in operation, 2.3% released an IPO, 9.4% closed down and 8.3% were acquired by other ventures.

For this section I use a combination of Tableau and python. The data analysis and plotting for pie chart is done in python

Funding situations

The most fundamental challenge for any startup is to acquire funding. Here I try to figure out aspects of the data that might give us insights into funding situation and possibilities. I mainly use python pandas for data analysis and save analyzed data to csv files. Using excel I plot the data in these saved results.

To begin with lets start with how much average funding is provided for a startup. It is about 18.5 million USD. But the range is quite wide. The minimum is 1 USD and maximum is 30 billion USD.

Average funding (USD): 18478604.03735475

Minimum funding (USD) : 1.0

Company (Min funding)	Location	First funding on
Sentic Technologies Inc	Chicago, USA	11/12/2015
Step to the future	Moscow, RUS	3/7/2015

Maximum funding (USD): 30079503000.0

Company (Max funding)	Location	First funding on
Verizon Communications	New York, USA	1/26/2010

Table 2 and Figure 5 Show how startups funding rounds relate to operations of the startup. It makes sense that the startups that released IPO went through more funding rounds than those that closed as they were potentially more attractive to investors potentially showing more chances of success.

Table 2. Average how many funding rounds conducted for startups in each status category.

Status	Average number of funding rounds
acquired	2.116778
closed	1.399006
IPO	2.796380
operating	1.700513

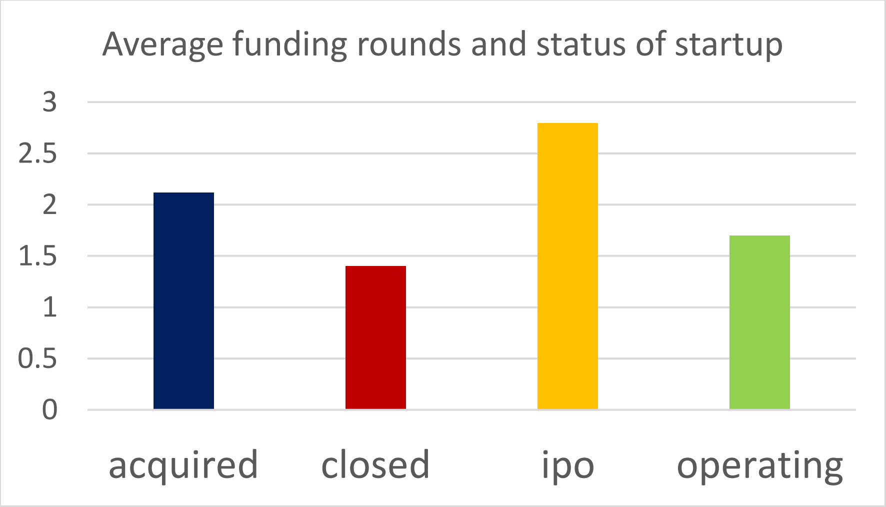


Figure 5. Classifying by operating status of the company, how many funding rounds were required to acquire funding

Another interesting aspect is to check which sectors get more funding rounds. Table 3 shows 9 categories that got on average 14 or more rounds of funding. Figure 6 shows Sectors more than 10 average rounds of funding

Table 3. Categories with on average 14+ funding rounds

Category	Average funding rounds
Communications Hardware|Enterprise Software|VoIP	17
Apps|Mobile|Mobility	16
Big Data Analytics|Clean Technology|Energy IT|Energy Management|Internet of Things	16
Biotechnology|Health and Wellness|Medical Devices|Technology	16
Banking|Finance|FinTech|Medical|Payments|Retail|Software	16
Banking|Cloud Computing|Finance|FinTech|Health Care|Technology	15
Farming|Logistics|Manufacturing|Service Providers|Supply Chain Management	15
Clean Technology|Construction	14

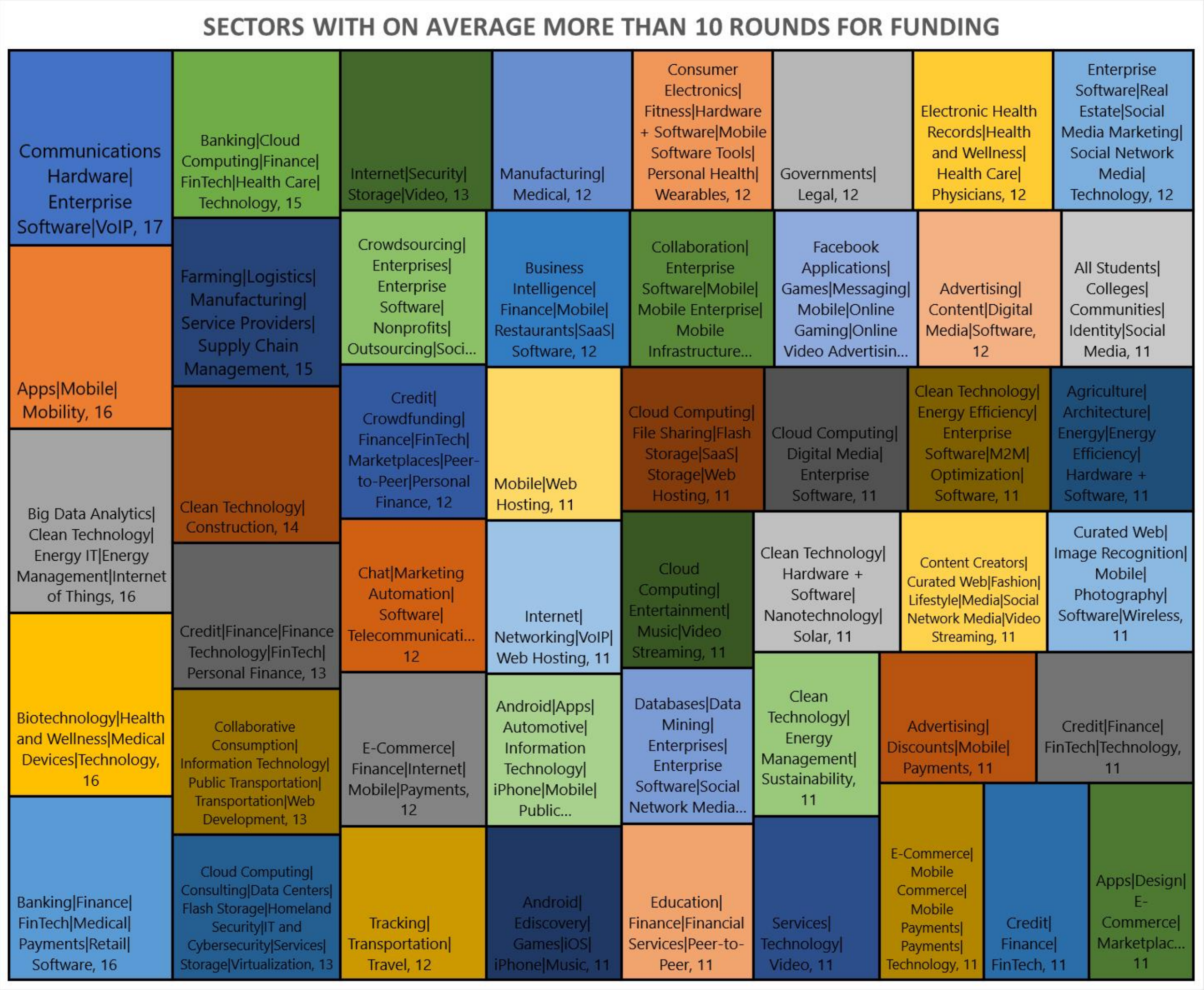
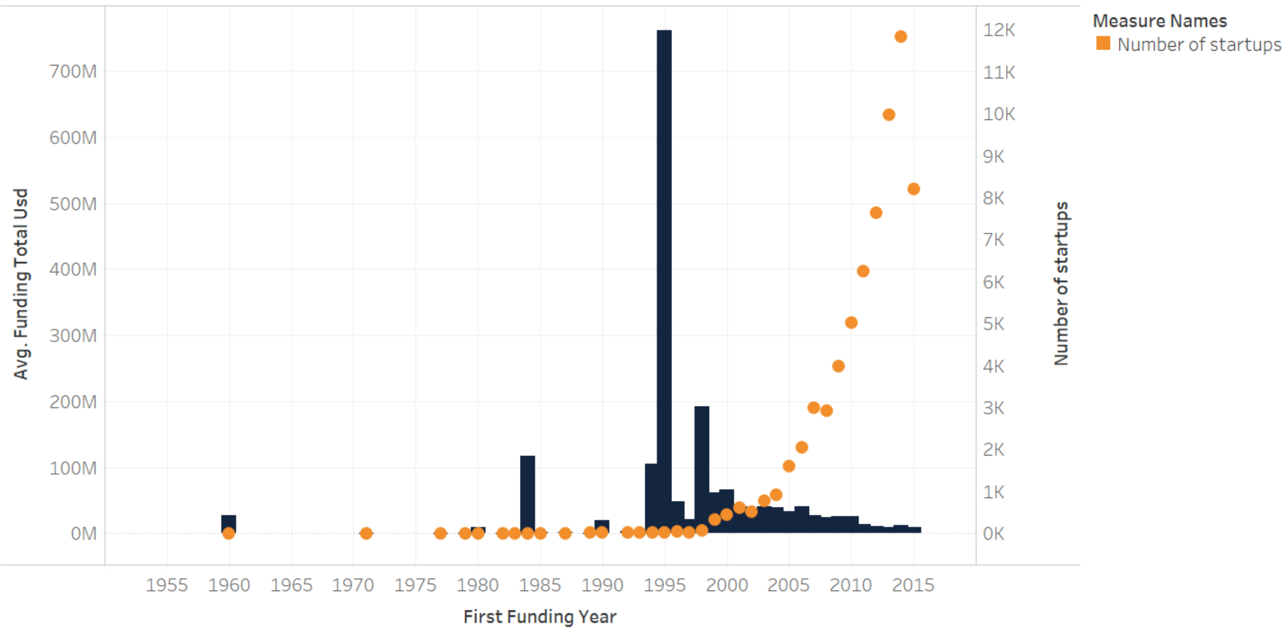


Figure 6. Category sectors that went through more than ten funding rounds on average

If we look through years, the number of startups has gone exponentially up. This is clearly evident in Figure 7 with orange circles showing increasing number of startups. This is plotted alongside the average funding per startup for each year. This in some sense tells me that around 1995 people started giving more importance to startup facilitation as we see that the average funding peaks around this time and the number of startups also starts increasing. The funding has not in reality decreased post 1995, but since the number of startups has increased to a large extent, the average funding has dropped as compared to earlier times.

Global Average Total Funding - yearly distribution



The plots of average of Funding Total Usd and Number of startups for First Funding Year. For pane Distinct count of Name: Color shows details about Number of startups. The view is filtered on average of Funding Total Usd, which keeps non-Null values only.

Figure 7. Number of startups and average funding provided along the years

Other interesting aspects to look are the top 20 highest funded startups of all time and of latest period in the dataset. These are shown in Figures 8 and 9 below. As expected from previous analysis, the highest funded startup of all time is Verizon. It is interesting to see that the second larges of all times is Freescale Semiconductors. I did some Googling to find, as per Wikipedia, "Freescale Semiconductor, Inc. was an American multinational corporation founded in 2004 headquartered in Austin, Texas, with design, research and development, manufacturing and sales operations in more than 75 locations in 19 countries" and seems to be acquired by NXP Semiconductors in 2015. Uber also ranks high, fourth, in the funding list. It is also interesting to note that Suning, the largest non-government retailer in China (as per Wikipedia) received highest funding as of 2015 (latest in this data).

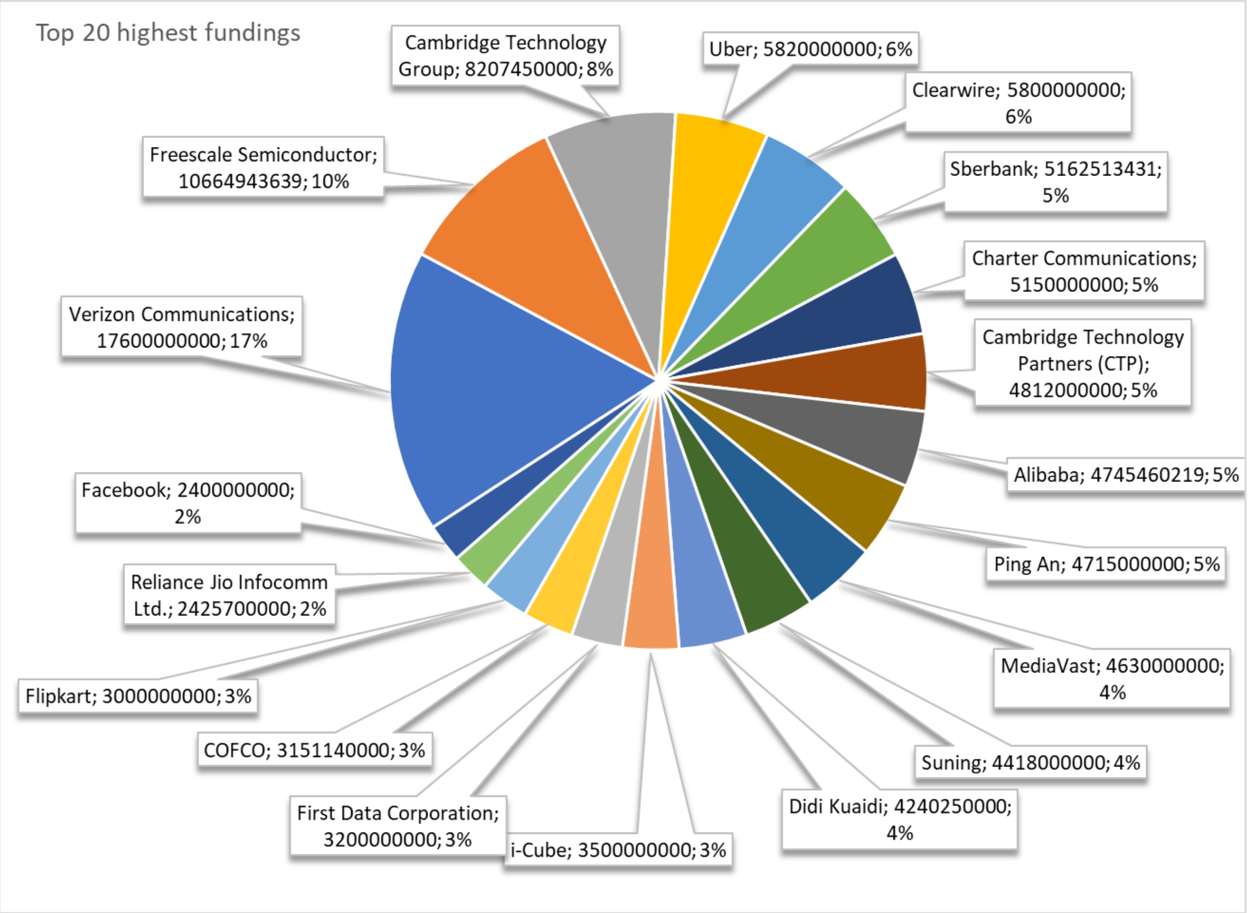


Figure 8. Top 20 highest funded startups on a pie chart. 100% represents total funding for the top 20 which is 13% of the total overall funding for all startups over all years in the dataset

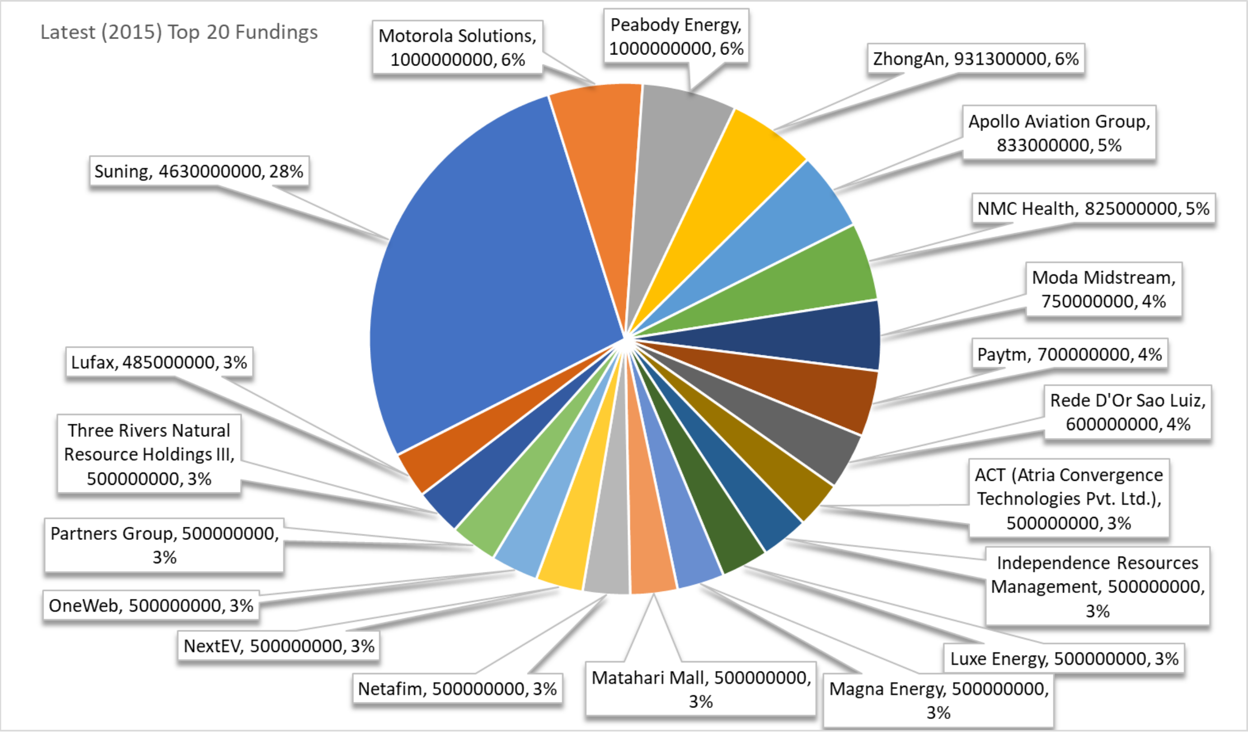


Figure 9. Top 20 highest funded newest startups (as of 2015) on a pie chart. 100% represents total funding for the new top 20 which is 12.5% of the total overall funding for all startups over all years in the dataset

Other interesting aspects to look at is the regional variation in funding. Figure 10 shows regions with average funding more than 100 million USD. Figure 11 shows the distribution (mean) and standard deviation of funding across the globe,

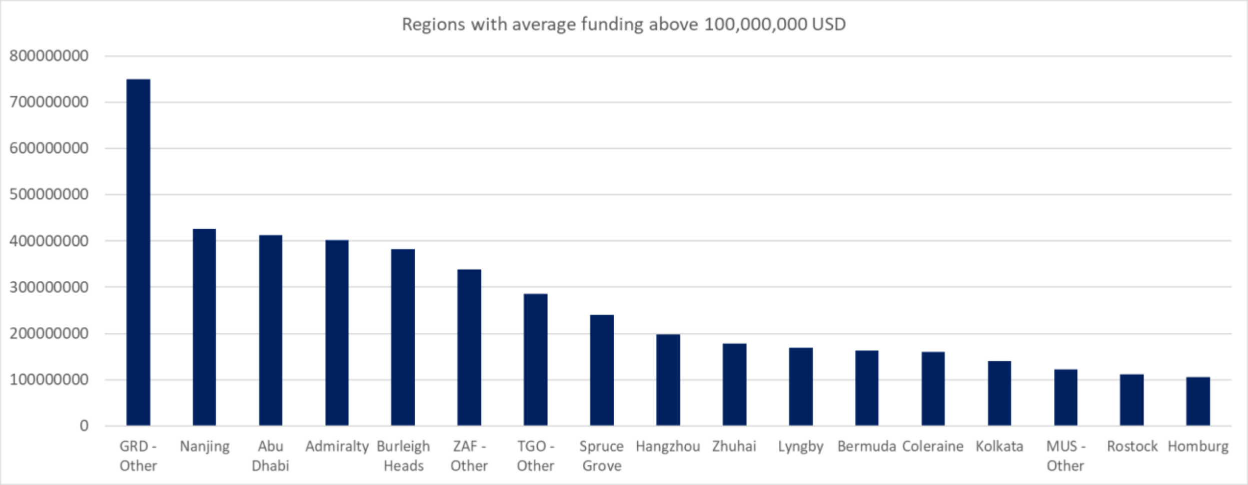
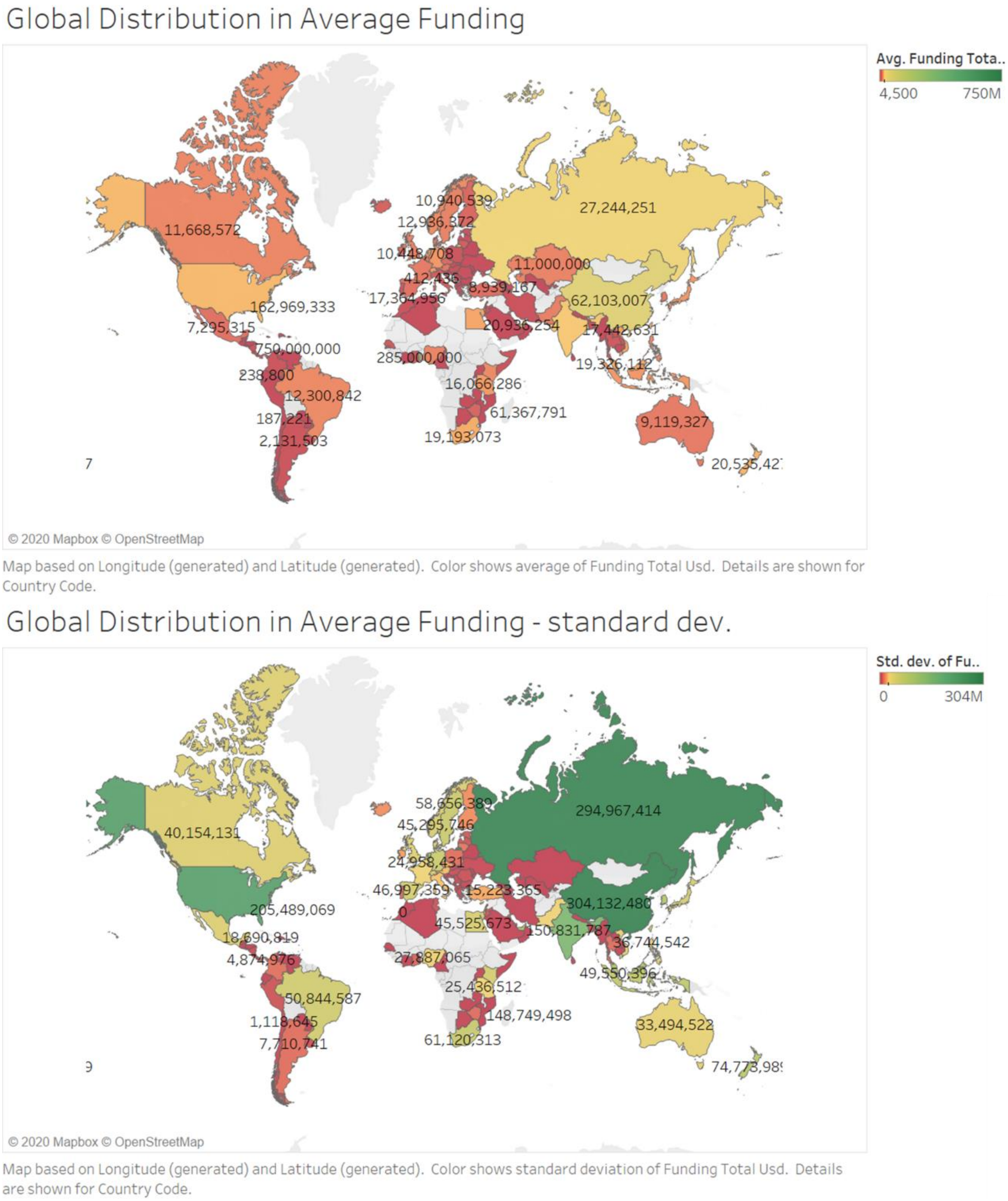


Figure 10. Regions with average funding over 100,000,000 USD.


Global Distribution in Average Funding

Map based on Longitude (generated) and Latitude (generated). Color shows average of Funding Total Usd. Details are shown for Country Code.

Global Distribution in Average Funding - standard dev.

Map based on Longitude (generated) and Latitude (generated). Color shows standard deviation of Funding Total Usd. Details are shown for Country Code.

Figure 11. Average (mean) and standard deviation of funding across the world. Large deviations in the funding values.

Acquisitions

From the acquisitions dataset I found that the average amount paid for acquisitions is around 876 million USD

```
Average price amount paid for acquisition (USD)   876532054.9628891
```

And the highest price paid is 160 billion USD.

```
Highest price amount paid for acquisition (USD)   160000000000.0
```

This was in 2015 for

Company	Category	Country	State	City
Allergan	Biotechnology|Medical|Pharmaceuticals	USA	CA	Irvine

Predicting Funding – K Nearest Neighbors

Method

Out of curiosity, I wondered if there was any way I could predict how much funding a startup could get if I knew something about it. As an exercise, I have implemented a simple K nearest neighbors algorithm in the jupyter notebook. Being the way it is, it's a lousy algorithms, slow on large data. Therefore I only choose a section of the data for finding the nearest neighbors. The idea was to say I know a company's category, location (meaning country and city) , and the year we want to predict the funding for. Can we predict the funding based on other similar company patterns.

I chose following parameters to calculate the distances.

As an example, following is a company in test data –

```
['Software', 'USA', 'Cleveland', 2012, 250000.0]
```

Category	Country code	City	First Funding Year	Total Funding (USD)	Distance
Software	USA	Cleveland	2012	25000	(To calculate)

Then I calculate distance of this company from other companies in the training data.

Distances are calculated as follows –

String comparisons (Category and City) are done using fuzzywuzzy library in python which compares strings by their levenshtein distance. For example, 'Databases for systems' and 'Database systems' are closely related categories. Country code if matched perfectly (==) is 0 distance, otherwise 1. Distance between years is the difference between them. All values are kept positive.

Below is the code snippet for the different distances.

```
d1 = 100 - fuzz.partial_ratio(company1[0],company2[0])
d2 = int(company1[1] != company2[1])
d3 = 100 - fuzz.partial_ratio(company1[2],company2[2])
d4 = abs(company1[3] - company2[3])
```

I tried 5 different types of distance weightings for the overall distance, with intention to tweak the parameters finding which one may be more important. (But it is difficult to predict that)

1. $distance = d1 + d2 + d3 + d4$
2. $distance = \sqrt{d1^2 + d2^2 + d3^2 + d4^2}$
3. $distance = d1 + 100\ d2 + d3 + 50\ d4$
4. $distance = \sqrt{d1^2 + 100\ d2^2 + d3^2 + 50\ d4^2}$
5. $distance = 5\ d1 + 100\ d2 + 8\ d3 + 10\ d4$

Then I find the K nearest neighbors with whichever distance I have chosen, and take a simple mean of their funding values as a predicted funding value. I calculate the error as a percentage of difference between actual funding received and predicted value. The cumulative error is a plain sum of the errors in the test set.

Results

Some results make sense, for example, using K = 3, and Distance #1

Test Company 4 : ['Biotechnology', 'USA', 'Bedford', 2013, 1600000.0]

Gives 3 nearest neighbors

	category	country	city	year	funding	distance
1724	Biotechnology	USA	Medford	2007	800000.0	15.231546
1668	Biotechnology	USA	Bend	2015	1000000.0	25.079872
1713	Biotechnology|Health Diagnostics	USA	Branford	2013	1277500.0	43.000000

Predicted Funding: 1025833.3333333334 & Actual Funding: 1600000.0

Error: 35.885416666666664 %

But for others, its way off.

Test Company 3 : ['Media', 'AUS', 'Melbourne', 2015, 600000.0]

Gives 3 nearest neighbors

	category	country	city	year	funding	distance
818	Advertising|Media|Technology	AUS	South Melbourne	2007	31634312.0	8
1641	Aerospace|Energy|Manufacturing|Medical Devices...	USA	Melbourne	2009	202125.0	27
838	Social Media	USA	Lilburn	2007	2300000.0	38

Predicted Funding: 11378812.333333334 & Actual Funding: 600000.0

Error: 1796.4687222222226 %

Consider the following test company.

Test Company 8 : ['Games', 'GBR', 'Reading', 2009, 3000000.0]

With K = 3 and varying distance functions, I look at the 3 nearest neighbors output.

Distance #1, #2 and #4

	category	country	city	year	funding	distance
1684	Games	CHN	Beijing	2009	10000000.0	44
253	Games	CHN	Beijing	2010	1136548.0	45
212	Games	CHN	Beijing	2008	1500000.0	45

Predicted Funding: 4212182.666666667 & Actual Funding: 3000000.0

Error: 40.4060888888889 %

Distance #3

	category	country	city	year	funding	distance
228	Accounting|Business Development|Finance|Softwa...	GBR	Edinburgh	2009	1300000.0	103
624	Analytics|Big Data|Databases|Software|Storage	GBR	London	2009	10563501.0	120
1396	Software	GBR	Warwick	2009	643000.0	127

Predicted Funding: 4168833.6666666665 & Actual Funding: 3000000.0

Error: 38.96112222222222 %

Distance #3 give more weightage to country location an year distance rather than category and city.

Distance #5

	category	country	city	year	funding	distance
1499	Aerospace|Drones|Internet of Things	GBR	Reading	2015	380000.0	360
1684	Games	CHN	Beijing	2009	10000000.0	444
212	Games	CHN	Beijing	2008	1500000.0	454

```
Predicted Funding:  3960000.0  & Actual Funding:  3000000.0
```

```
Error:  32.0 %
```

Distance #5 gives a mix of #1 and #3 results due to altered weights in the distance.

Therefore, an improved distance function is required for better modelling of the data. Also, the input parameters may not be a sufficient representation and additional parameters need to be added to the company representation. Lastly, a fast and more robust code would facilitate comparison over larger training data for improved prediction.

Conclusion

The startup market is dominated by United States where more than 50% of the startups have been located. Majority of startups are located in metropolitan areas. Funding is volatile in startup market. The startup market really started to pick up around 1995-2000 and the number of new startups increased rapidly each year.

Description of Files used

Following analysis files are used with this document

1. companies_v2.csv – Preprocessed dataset to remove/modify some columns and apply basic queries like total length of dataset
2. companies.csv and acquisitions.csv – Used as is from the dataset
3. Project1_StartupMarketForAI.iIpynb - Jupyter notebook for data processing, analysis and partly visualization
    a. Uses pandas to load and query the data (in files 1 and 2). Saves results data into separate csv files for postprocessing visualizations
    b. Matplotlib to visualize some of the data
    c. K Nearest Neighbors code and analysis
4. Two tableau workbooks – To visualize (mainly geographic and bar charts) data. One workbook was overloading so I split work in two.
    a. StartupCruchbaseData_companies.tbw
    b. StartupCruchbaseData_companies_v2.tbw
5. Eleven csv files generated from the ipython notebook in part 2. Used in excel spreadsheets to create pie charts, tree charts and bar charts, and some post processing calculations.
    a. average_funding_per_year.csv
    b. avg_funding_status.csv
    c. category_rounds.csv
    d. Country_funding.csv
    e. countrywise_count.csv
    f. latest_top20funding.csv
    g. region_funding.csv
    h. status_count.csv
    i. top20funding.csv
    j. yearwise_count.csv
    k. yearwise_funding.csv