Startup Database and Recommendation Engine

K Iwasaki
kaiwasaki@berkeley.edu

Table of Contents

Problem: Finding Good Startup is Hard

For job seekers, finding a startup that matches their interests is hard because:

- Overwhelming information available online
- So many information sources to check
- Need to synthesize information
- Need to check information frequently
- So many texts to read
- Job seekers are highly biased

Frustration, Time waste, Not finding company that matches your interests

Project Overview: Finding the Best Startup For You

Create end-to-end solution from data collection, to database generation, to generation of recommendation for startups that matches your interests.

Data Collection/Preprocessing

10 articles x 100 pages

TechCrunch

~300 searches

Google

~ 300 company profiles

LinkedIn

Bloomberg

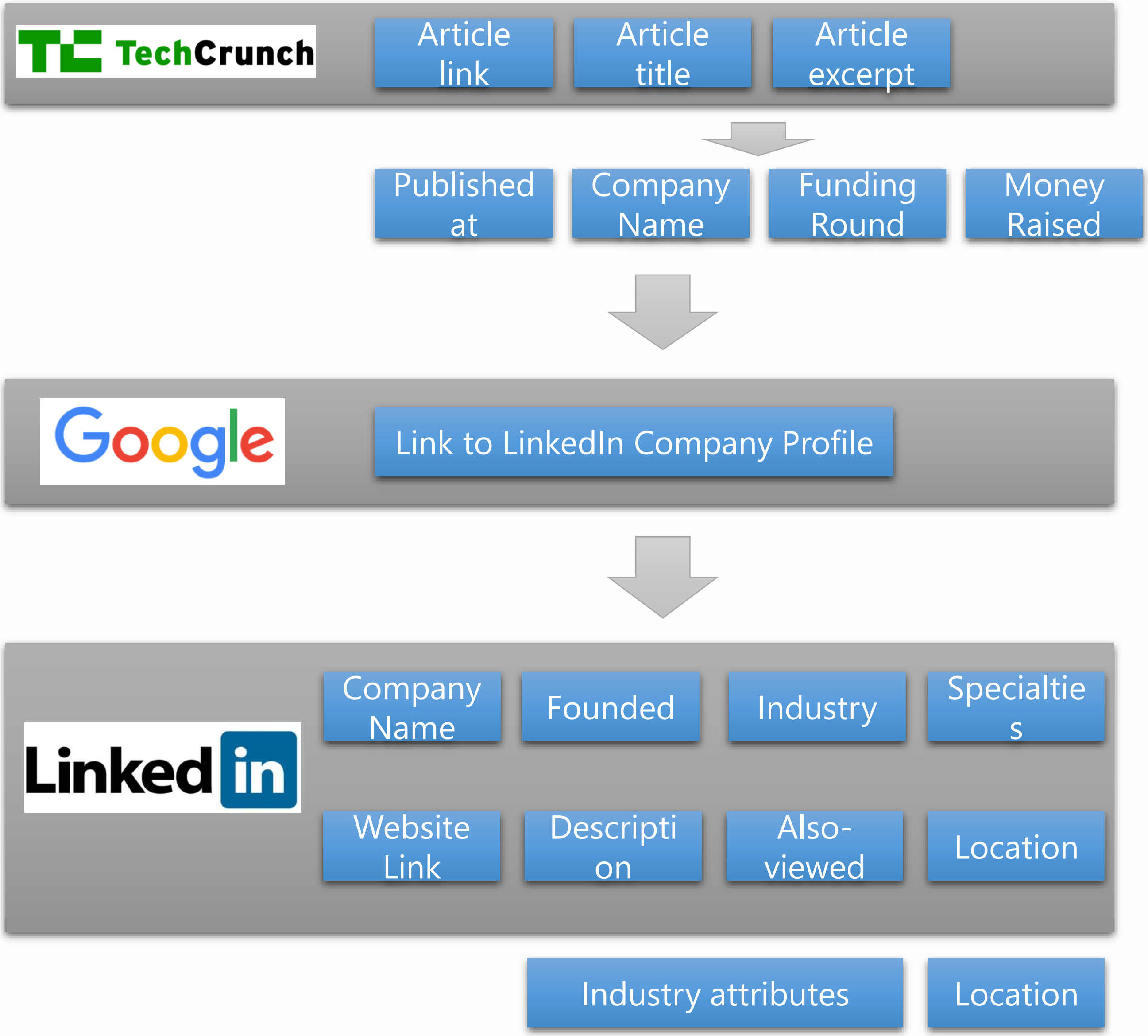
Startup Database

Company name
Company size
Money raised
Industry
Description
HQ Location
…

Recommendation Engine

Data Exploration and Visualization

Data Collection and Preprocessing Scheme



Scrape the articles about Series C fundraising from TechCrunch (article.csv)

Preprocessing1. Extract
Preprocesisng2. Extract company names
Preprocessing3. Extract funding_round and money_raised

Scrape the website links to LinkedIn Company Profiles from Google Search (linkedin_link_list.csv)

Preprocessing4. Merge the two CSV files
Preprocessing5. Validate company names

Scrape the company profile for each company from LinkedIn (linkedin_profiles.csv)

Preprocessing6. Merge the two CSV files
Preprocessing7. Extract locations
Preprocessing8. Assign industry attributes

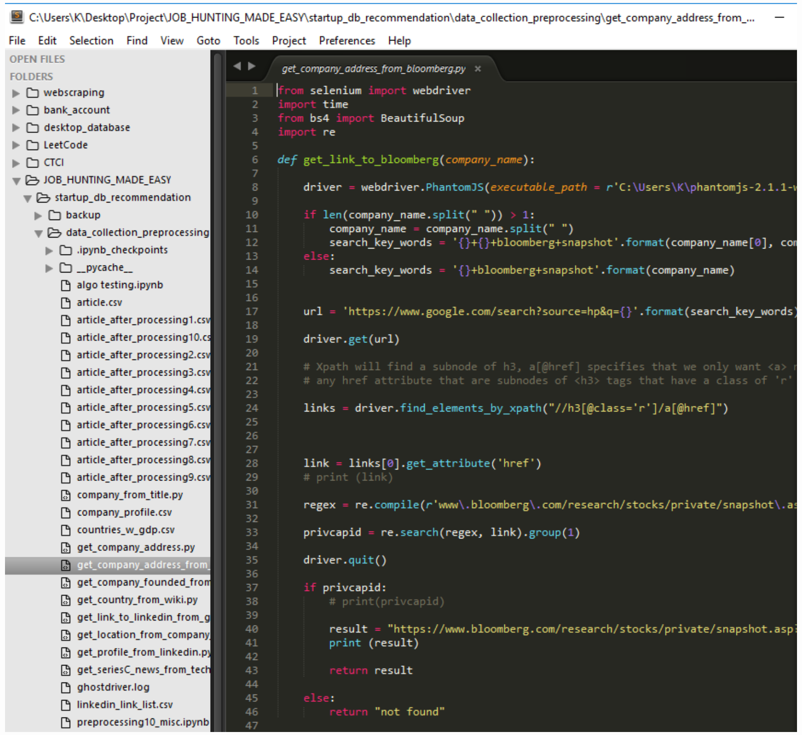
Data Collection and Preprocessing Approach

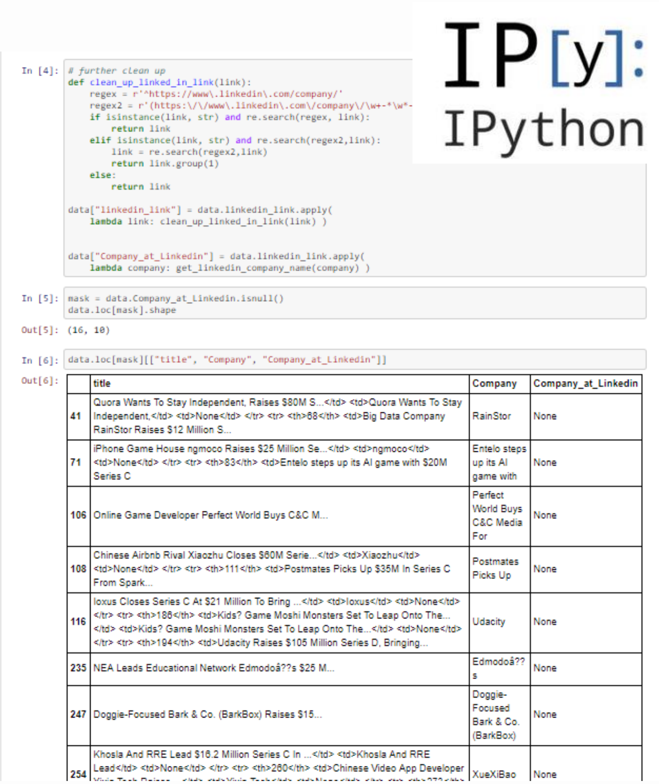
Think what information we want for the database and for the recommendation engine

Write codes and extract information from the target source

Error analysis: Confirm if we get what we want and verify missing data and why

Company name
Company size
Money raised
Industry
Description
HQ Location
…

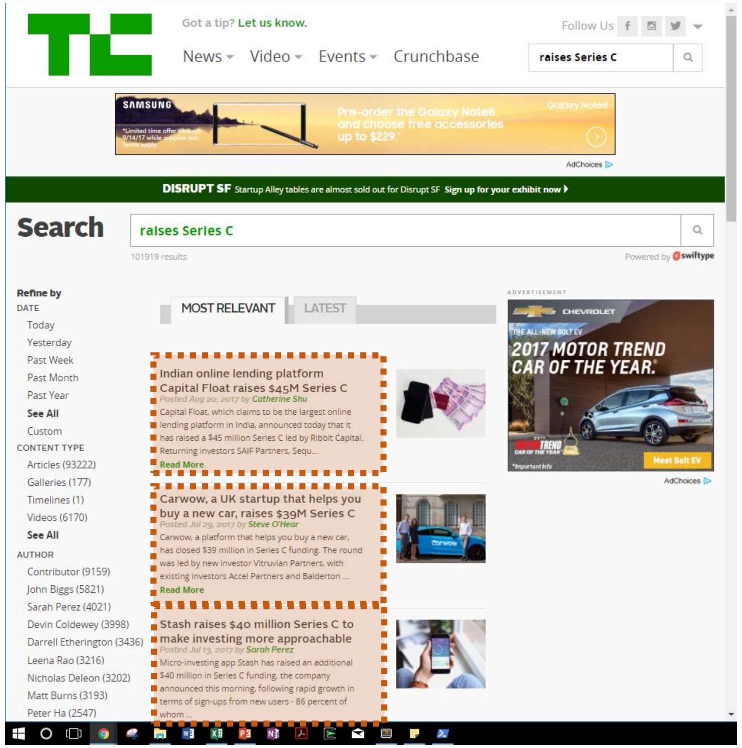




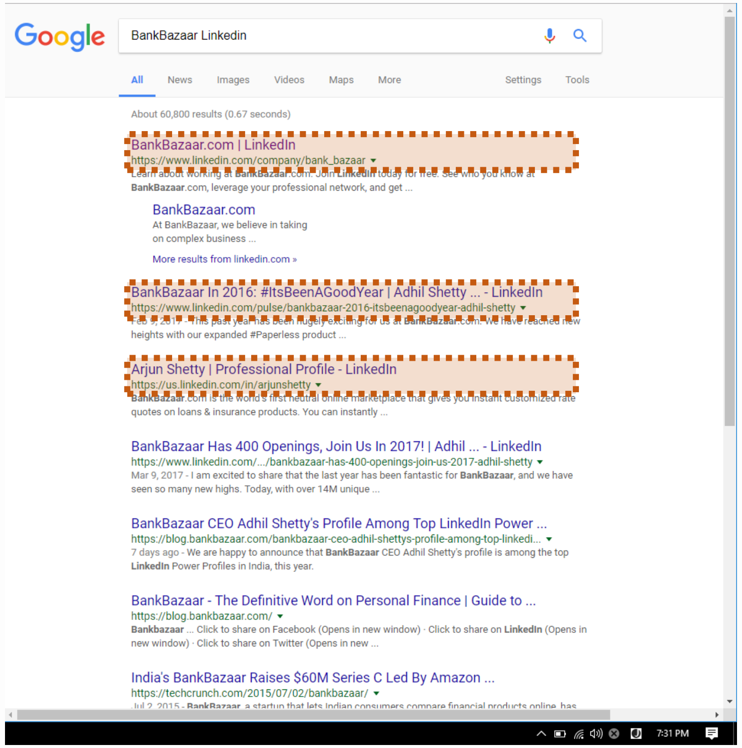
Data Scraping

Tasks: Use Selenium and BeautifulSoup to scrape information the target websites.

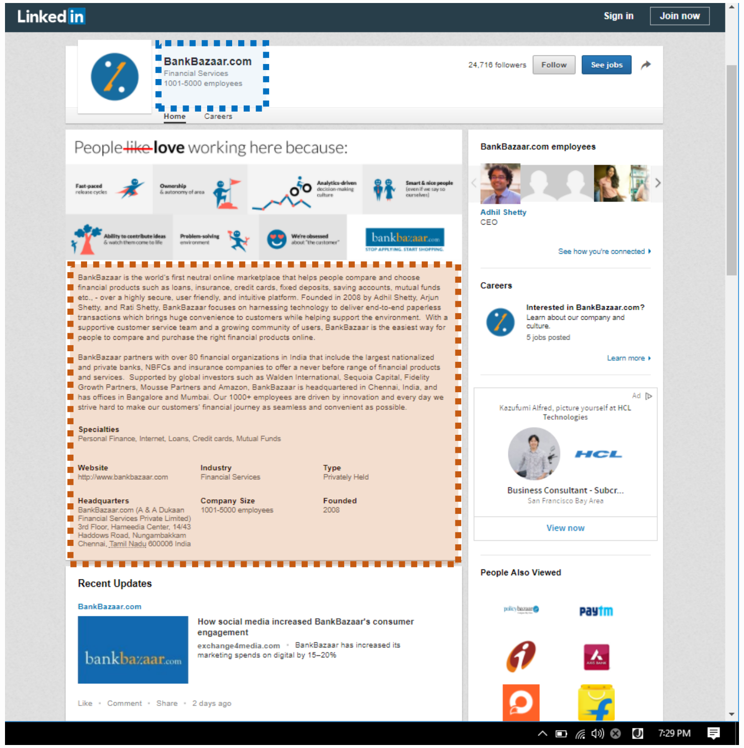
get_seriesC_news_from_techcrunch.py
Input: key words “raises Series C”
Output: articles in csv file



get_link_to_linkedin_from_google.py
Input: company name
Output: a link to company profile at LinkedIn



get_profile_from_linkedin.py
Input: a link to company profile
Output: company profiles in csv file



Data Extraction: Company Name from Article Title

Once we collect the articles, next step is extract company names from the article titles.
Challenge: a company name is irregular: it can be one word, two words, or more. It often is a mix of verb, noun, or others. Below are typical patterns that a company name shows up in an article title.

"Stash raises $40 million Series C to make investing more approachable"

"Data Storage Company Scale Computing Raises $17 Million Series C"

"Pivotal confirms Series C round is actually over $650 million"

"After bump in the road, Movinga raises $17M Series C"

"Carwow, a UK startup that helps you buy a new car, raises $39M Series "

"Confirmed: London fintech Curve raises $10M Series A"

Company names, Key verbs, decorative words

Algorithm for Company Name Extraction

Solution: algorithm to extract a company name, leveraging sentence structures of the articles that are scraped from TechCrunch. Also double-check the company name when googling it later to look for a link for a company profile page at LinkedIn. Check ***company_from_title.py for the codes***

Step1:

- Split the sentence by a key verb and keep the head
- Remove ", word word … ,"
  - If one or two words remained=> done Else: => Step2

Step2:

- Split the sentence by a key noun and keep the tail
  - If one or two words remained=> done Else: => Step2

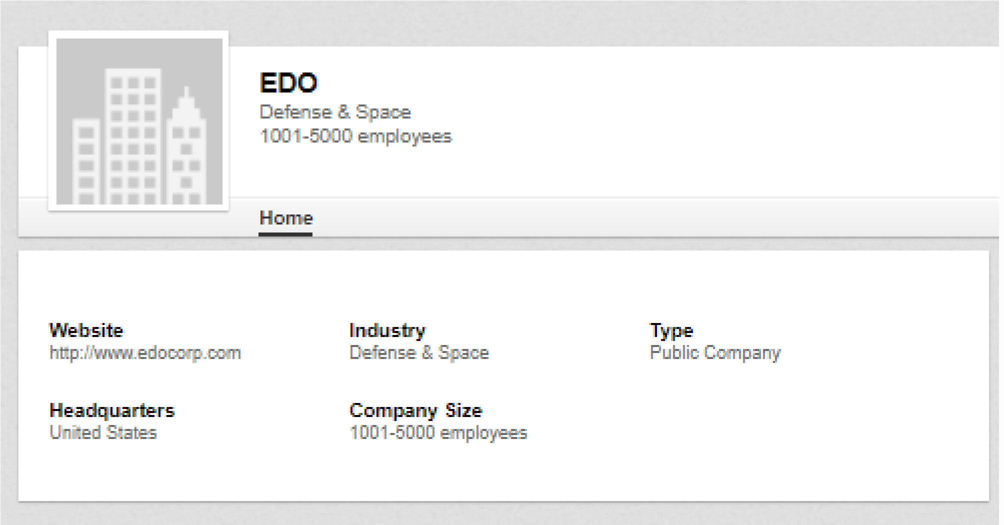
Step3:

- Split the sentence by "$" and keep the head
- Split the sentence by "Series" and keep the head

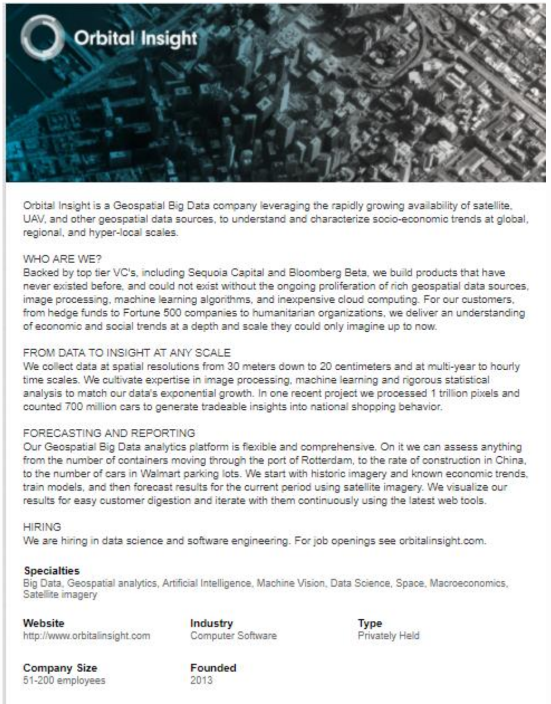
Data Extraction: Company Address

Company address is import input for the recommendation engine because many of us care where we work at.

Challenge: Some companies don’t input their company address at LinkedIn. Some companies are based outside of US and thus their addresses have different formatting.



It only says United States for Headquarters.



It doesn’t have address at all.

Solution to extract/revise Company Address

Solution: Multi-step approach: first focus on label countries and then focus on US companies to extract zip code.

- Step1: Complete labeling by countries
  - Country list scraped from Wikipedia --- *get_country_from_wiki.py*
  - Extract country information from features collected so far
- Step2: Focus on the US companies and get zip code for them

  For missing or insufficient information
  - Google Search --- *get_comnay_address.py*
  - Company Website --- *get_location_from_company_website.py*
  - Bloomberg --- *get_company_address_from_Bloomberg.py*
- Step3: Gain state, city, geo location from the zip code for US companies

  Use two python modules to capture city and state because both of them have some missing data

Data Extraction: Industry Attributes from Description

Challenges: Industries have been arbitrarily assigned to companies. As a result, there are 49 unique industries for about 300 companies. There are three problems in order for recommendation engine to work:

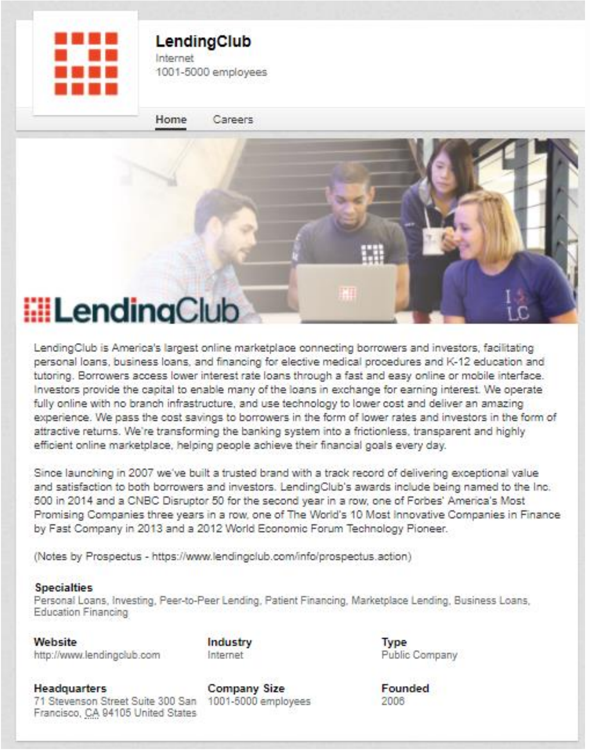
1) ***Some industries are quite similar*** thus should be merged.

2) ***Some industries have lots of companies*** such as Computer Software. They should be split into more smaller segment.

3) ***one industry is not sufficient to describe a nature of a company*** because its business is often a combination of different elements. For example, the company below is internet x financial service, instead of internet alone

```
data.Industry.unique()
```

```
array(['Financial Services', 'Information Technology and Services',
       'Human Resources', 'Computer Software',
       'Logistics and Supply Chain', 'Internet',
       'Computer & Network Security', 'Food & Beverages',
       'Marketing and Advertising', 'Medical Devices', 'E-Learning',
       'Consumer Services', 'Sports', 'Consumer Electronics',
       'Computer Hardware', 'Education Management', 'Apparel & Fashion',
       'Entertainment', 'Consumer Goods', 'Biotechnology',
       'Management Consulting', 'Real Estate', 'Fund-Raising',
       'Commercial Real Estate', 'Food Production', 'Online Media',
       'Mechanical or Industrial Engineering', 'Renewables & Environment',
       'Farming', 'Electrical/Electronic Manufacturing',
       'Leisure, Travel & Tourism', 'Sporting Goods', 'Retail',
       'Semiconductors', 'Cosmetics', 'Insurance', 'Telecommunications',
       'Health, Wellness and Fitness', 'Textiles',
       'Staffing and Recruiting', 'Nanotechnology',
       'Luxury Goods & Jewelry'], dtype=object)
```



Algorithm to Assign Industry Attributes to Each Company

Solution Part1: algorithm to simplify the industry classification by merging some industries so that minor industry labels are eliminated

Industry (Original)	Industry_consolidated (New)
["Apparel & Fashion", "Consumer Goods", "Consumer Services", "Cosmetics", "Luxury Goods & Jewelry", "Retail", "Leisure, Travel & Tourism", "Sporting Goods", "Textiles"]	Consumers Goods & Services
["Computer Software"]	Computer Software
["Computer & Network Security", "Computer Hardware"]	Computer & Network Security & Hardware
['E-Learning', 'Education Management']	Education
["Entertainment"]	Entertainment
["Marketing and Advertising"]	Marketing and Advertising
["Farming", "Food & Beverages", "Food Production", "Restaurants"]	Food Business
["Insurance", "Fund-Raising", "Financial Services"]	Financial Services
["Information Technology and Services"]	Information Technology and Services
["Internet", "Online Media"]	Internet
["Commercial Real Estate", "Real Estate"]	Real Estate
['Health, Wellness and Fitness', 'Medical Devices', "Sports"]	Healthcare_health
["Human Resources", "Staffing and Recruiting"]	Human Resources
["Telecommunications", "Renewables & Environment", "Logistics and Supply Chain"]	Infrastructure
["Semiconductors", "Nanotechnology", "Biotechnology", "Management Consulting", "Electrical/Electronic Manufacturing" "Mechanical or Industrial Engineering"]	Niche

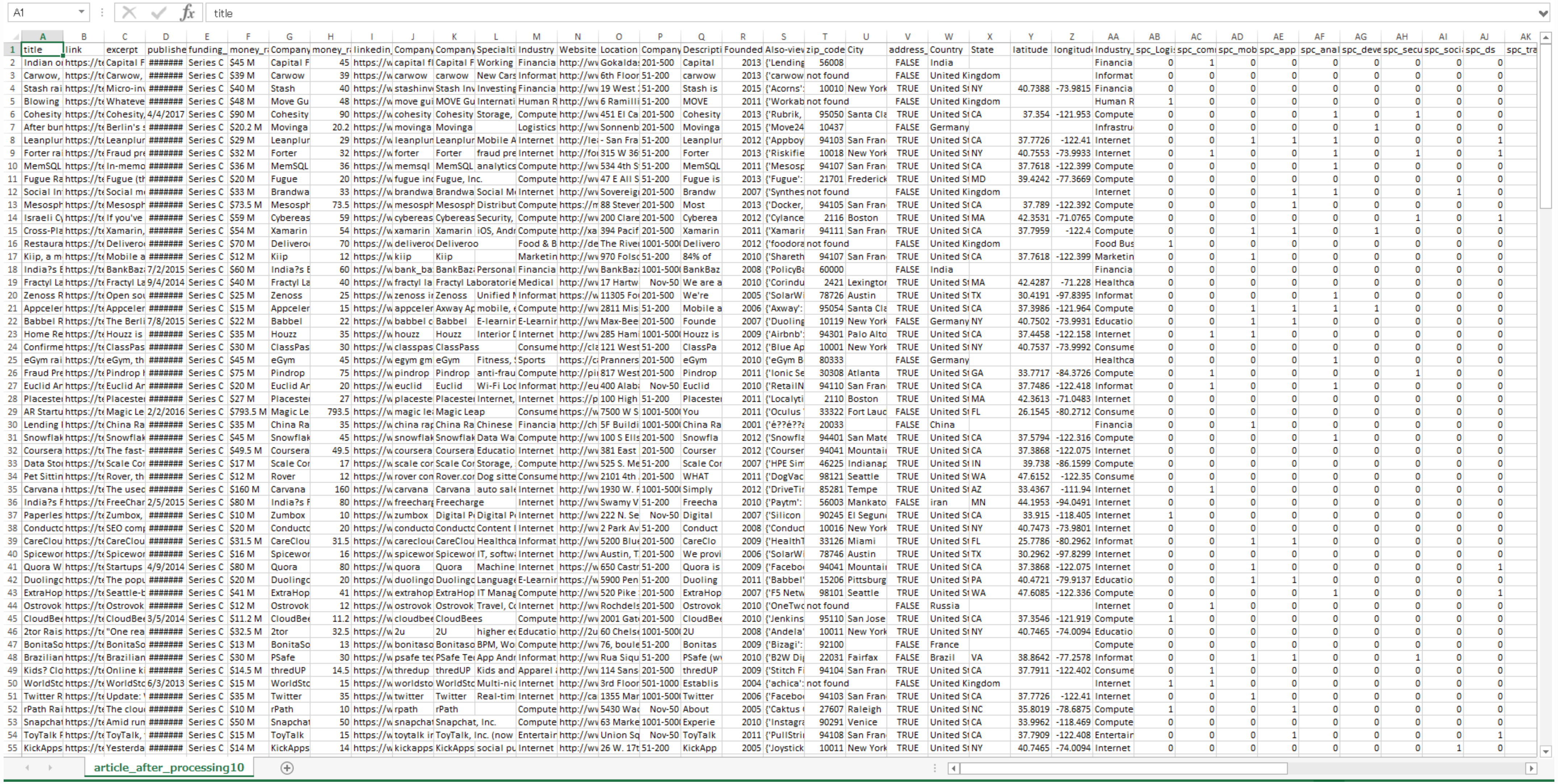
Algorithm to Assign Industry Attributes to Each Company

Solution Part2: Algorithm to add new features to represent company businesses better based on the key words in appeared in company profiles

```
key_words_dict = {
    "Food Business": ["restaurant", "farm", "greenhouse", "Gastronomie"],
    "Education": ["Online Learning", "Education", "Tutor"],
    "Financial Services": ["payment", "loan", "financ", "fundraising",
                "investing", "lending"],
    "Healthcare_health": ["healthcare", "medical", "genetic", "therapy", "disease",
                   "fitness", "wellness", "welfare","wearable", "gym"],
    "Human Resources": ["recruit", "workforce", "Human Resource"],
    "Logistics and Supply Chain": ["delivery", "drone",
                                   "transportation", "supply chain"],
    "Entertainment": ["entertainment", "game"],
    "Computer & Network Security & Hardware": ["storage","backup", "recovery",
                                               "privacy"],
    "Real Estate": ["Real Estate"],
    "Marketing and Advertising": ["marketing", "advertising", "advertisement"],

    "commerce": ["eCommerce", "Commerce", "Retail"],
    "mobile" : ["mobile"],
    "app": ["mobile app", "app\s"],
    "analysis": ["analytics", "analysis"],
    "developer": ["developer"],
    "security" : ["fraud", "detection", "protection"],
    "social": ["Social Media"],
    "ds": ["artificial intelligence", "machine learning",
          "deep learning", "big data"],
    "travel": ["Travel"],
    "booking_ticketing": ["booking", "ticket"],
    "Apparel": ["fashion", "clothing", "shoes", "Sporting Goods"],
    "cloud": ["cloud"],
    "API": ["API"],
    "device": ["device"],
    "design": ["design"],
    "enterprise": ["enterprise", "productivity", "collaboration"],
    "robotics_manufacturing": ["Manufact", "robotics", "3d"]
 }
```

Now Database is Set! --- 232 rows by 47 columns



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	title	link	excerpt	publishe	funding_	money_r	Company	money_r	linkedin_	Company	Company	Specialti	Industry	Website	Location	Company	Descripti
2	Indian o	https://te	Capital F	#######	Series C	$45 M	Capital F	45	https://w	capital fl	Capital F	Working	Financia	http://wv	Gokalda:	201-500	Capital
3	Carwow,	https://te	Carwow,	#######	Series C	$39 M	Carwow	39	https://w	carwow	carwow	New Cars	Informat	http://wv	6th Floor	51-200	carwow
4	Stash rai	https://te	Micro-inv	#######	Series C	$40 M	Stash	40	https://w	stashinv	Stash Inv	Investing	Financia	http://wv	19 West	51-200	Stash is
5	Blowing	https://te	Whateve	#######	Series C	$48 M	Move Gu	48	https://w	move gui	MOVE Gu	Internati	Human R	http://wv	6 Ramilli	51-200	MOVE
6	Cohesity	https://te	Cohesity,	4/4/2017	Series C	$90 M	Cohesity	90	https://w	cohesity	Cohesity	Storage,	Compute	http://wv	451 El Ca	201-500	Cohesity
7	After bun	https://te	Berlin's s	#######	Series C	$20.2 M	Movinga	20.2	https://w	movinga	Movinga		Logistics	http://wv	Sonnenb	201-500	Movinga
8	Leanplur	https://te	Leanplur	#######	Series C	$29 M	Leanplur	29	https://w	leanplun	Leanplur	Mobile A	Internet	http://lea	- San Fra	51-200	Leanplur
9	Forter rai	https://te	Fraud pre	#######	Series C	$32 M	Forter	32	https://w	forter	Forter	fraud pre	Internet	http://for	315 W 36	51-200	Forter
10	MemSQL	https://te	In-memo	#######	Series C	$36 M	MemSQL	36	https://w	memsql	MemSQL	analytics	Compute	http://wv	534 4th S	51-200	MemSQL
11	Fugue Ra	https://te	Fugue (th	#######	Series C	$20 M	Fugue	20	https://w	fugue inc	Fugue, Inc.		Compute	http://wv	47 E All S	51-200	Fugue is
12	Social In	https://te	Social m	#######	Series C	$33 M	Brandwa	33	https://w	brandwa	Brandwa	Social M	Internet	http://wv	Sovereig	201-500	Brandw
13	Mesosph	https://te	Mesosph	#######	Series C	$73.5 M	Mesosph	73.5	https://w	mesosph	Mesosph	Distribut	Compute	https://m	88 Stever	201-500	Most
14	Israeli Cy	https://te	If you've	#######	Series C	$59 M	Cybereas	59	https://w	cybereas	Cybereas	Security,	Compute	http://wv	200 Clare	201-500	Cyberea
15	Cross-Pla	https://te	Xamarin,	#######	Series C	$54 M	Xamarin	54	https://w	xamarin	Xamarin	iOS, Andr	Compute	http://xa	394 Pacif	201-500	Xamarin
16	Restaura	https://te	Deliveroo	#######	Series C	$70 M	Deliveroo	70	https://w	deliveroo	Deliveroo		Food & B	http://de	The River	1001-500	Delivero
17	Kiip, a m	https://te	Mobile a	#######	Series C	$12 M	Kiip	12	https://w	kiip	Kiip		Marketin	http://wv	970 Folsc	51-200	84% of
18	India?s E	https://te	BankBaz	7/2/2015	Series C	$60 M	India?s E	60	https://w	bank_ba	BankBaz	Personal	Financia	http://wv	BankBaz	1001-500	BankBaz
19	Fractyl La	https://te	Fractyl La	9/4/2014	Series C	$40 M	Fractyl La	40	https://w	fractyl la	Fractyl Laboratorie		Medical	http://wv	17 Hartw	Nov-50	We are a
20	Zenoss R	https://te	Open sou	#######	Series C	$25 M	Zenoss	25	https://w	zenoss ir	Zenoss	Unified M	Informat	https://w	11305 Fo	201-500	We're
21	Appceler	https://te	Appceler	#######	Series C	$15 M	Appceler	15	https://w	appceler	Axway Ap	mobile, e	Compute	http://wv	2811 Mis	51-200	Mobile a
22	Babbel R	https://te	The Berli	7/8/2015	Series C	$22 M	Babbel	22	https://w	babbel c	Babbel	E-learnir	E-Learnir	http://wv	Max-Bee	201-500	Founde
23	Home Re	https://te	Houzz is	#######	Series C	$35 M	Houzz	35	https://w	houzz	Houzz	Interior D	Internet	http://wv	285 Hami	1001-500	Houzz is
24	Confirme	https://te	ClassPas	#######	Series C	$30 M	ClassPas	30	https://w	classpas	ClassPass		Consume	http://cla	121 West	51-200	ClassPa
25	eGym rai	https://te	eGym, th	#######	Series C	$45 M	eGym	45	https://w	egym gm	eGym	Fitness, S	Sports	https://ca	Pranners	201-500	eGym
26	Fraud Pre	https://te	Pindrop I	#######	Series C	$75 M	Pindrop	75	https://w	pindrop	Pindrop	anti-frau	Compute	http://pi	817 West	201-500	Pindrop
27	Euclid Ar	https://te	Euclid Ar	#######	Series C	$20 M	Euclid Ar	20	https://w	euclid	Euclid	Wi-Fi Loc	Informat	http://eu	400 Alab	Nov-50	Euclid
28	Placester	https://te	Placester	#######	Series C	$27 M	Placester	27	https://w	placeste	Placester	Internet,	Internet	https://p	100 High	51-200	Placester
29	AR Startu	https://te	Magic Le	2/2/2016	Series C	$793.5 M	Magic Le	793.5	https://w	magic lea	Magic Leap		Consume	https://w	7500 W S	1001-500	You
30	Lending I	https://te	China Ra	#######	Series C	$35 M	China Ra	35	https://w	china rap	China Ra	Chinese	Financia	http://ch	5F Buildi	1001-500	China Ra
31	Snowflak	https://te	Snowflak	#######	Series C	$45 M	Snowflak	45	https://w	snowflak	Snowflak	Data Wa	Compute	http://wv	100 S Ells	201-500	Snowfla
32	Coursera	https://te	The fast-	#######	Series C	$49.5 M	Coursera	49.5	https://w	coursera	Coursera	Educatio	Internet	http://wv	381 East	201-500	Courser
33	Data Sto	https://te	Scale Cor	#######	Series C	$17 M	Scale Cor	17	https://w	scale cor	Scale Cor	Storage,	Compute	http://wv	525 S. Me	51-200	Scale Cor
34	Pet Sittin	https://te	Rover, th	#######	Series C	$12 M	Rover	12	https://w	rover con	Rover.cor	Dog sitte	Consume	http://wv	2101 4th	201-500	WHAT
35	Carvana	https://te	The usec	#######	Series C	$160 M	Carvana	160	https://w	carvana	Carvana	auto sale	Internet	http://wv	1930 W. F	1001-500	Simply
36	India?s F	https://te	FreeChar	2/5/2015	Series C	$80 M	India?s F	80	https://w	freecharg	Freecharge		Internet	http://wv	Swamy V	51-200	Freecha
37	Paperles	https://te	Zumbox,	#######	Series C	$10 M	Zumbox	10	https://w	zumbox	Digital P	Digital P	Internet	http://wv	222 N. Se	Nov-50	Digital
38	Conducto	https://te	SEO com	#######	Series C	$20 M	Conducto	20	https://w	conducto	Conducto	Content I	Internet	http://wv	2 Park Av	51-200	Conduct
39	CareClou	https://te	CareClou	#######	Series C	$31.5 M	CareClou	31.5	https://w	careclou	CareClou	Healthca	Informat	http://wv	5200 Blue	201-500	CareClo
40	Spicewor	https://te	Spicewor	#######	Series C	$16 M	Spicewor	16	https://w	spicewor	Spicewor	IT, softw	Internet	http://wv	Austin, T	201-500	We provi
41	Quora W	https://te	Startups	4/9/2014	Series C	$80 M	Quora	80	https://w	quora	Quora	Machine	Internet	https://w	650 Castr	51-200	Quora is
42	Duolingo	https://te	The popu	#######	Series C	$20 M	Duolingo	20	https://w	duolingo	Duolingo	Language	E-Learnir	https://w	5900 Pen	51-200	Duoling
43	ExtraHop	https://te	Seattle-b	#######	Series C	$41 M	ExtraHop	41	https://w	extrahop	ExtraHop	IT Manag	Compute	http://wv	520 Pike	201-500	ExtraHop
44	Ostrovok	https://te	Ostrovok	#######	Series C	$12 M	Ostrovok	12	https://w	ostrovok	Ostrovok	Travel, C	Internet	http://wv	Rochdels	201-500	Ostrovok
45	CloudBee	https://te	CloudBee	3/5/2014	Series C	$11.2 M	CloudBee	11.2	https://w	cloudbee	CloudBees		Compute	http://wv	2001 Gat	201-500	CloudBee
46	2tor Rais	https://te	"One rea	#######	Series C	$32.5 M	2tor	32.5	https://w	2u	2U	higher ed	Educatio	http://2u	60 Chelse	1001-500	2U
47	BonitaSo	https://te	BonitaSo	#######	Series C	$13 M	BonitaSo	13	https://w	bonitaso	Bonitaso	BPM, Wo	Compute	http://wv	76, boule	51-200	Bonitas
48	Brazilian	https://te	Brazilian	#######	Series C	$30 M	PSafe	30	https://w	psafe tec	PSafe Tec	App Andr	Informat	http://wv	Rua Siqu	51-200	PSafe (w
49	Kids? Clo	https://te	Online ki	#######	Series C	$14.5 M	thredUP	14.5	https://w	thredup	thredUP	Kids and	Apparel &	http://wv	114 Sans	201-500	thredUP
50	WorldSto	https://te	WorldSto	6/3/2013	Series C	$15 M	WorldSto	15	https://w	worldsto	WorldSto	Multi-nic	Internet	http://wv	3rd Floor	501-1000	Establis
51	Twitter R	https://te	Update:	#######	Series C	$35 M	Twitter	35	https://w	twitter	Twitter	Real-tim	Internet	http://ca	1355 Mar	1001-500	Twitter
52	rPath Rai	https://te	The clou	#######	Series C	$10 M	rPath	10	https://w	rpath	rPath		Compute	http://wv	5430 Wac	Nov-50	About
53	Snapchat	https://te	Amid run	#######	Series C	$50 M	Snapchat	50	https://w	snapchat	Snapchat, Inc.		Compute	http://wv	63 Marke	1001-500	Experie
54	ToyTalk F	https://te	ToyTalk,	#######	Series C	$15 M	ToyTalk	15	https://w	toytalk ir	ToyTalk, Inc. (now		Entertair	http://wv	Union Sq	Nov-50	ToyTalk
55	KickApps	https://te	Yesterda	#######	Series C	$14 M	KickApps	14	https://w	kickapps	KickApps	social pu	Internet	http://wv	26 W. 17t	51-200	KickApp

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1	Founded	Also-view	zip_code	City	address_	Country	State	latitude	longitude	Industry_	spc_Logi	spc_com	spc_mob	spc_app	spc_anal	spc_deve	spc_secu	spc_soci	spc_ds	spc_tra
2	2013	{'Lending	56008		FALSE	India				Financia	0	1	0	0	0	0	0	0	0	
3	2013	{'carwow	not found		FALSE	United Kingdom				Informat	0	0	0	0	0	0	0	0	0	
4	2015	{'Acorns':	10010	New York	TRUE	United St	NY	40.7388	-73.9815	Financia	0	0	0	0	0	0	0	0	0	
5	2011	{'Workab	not found		FALSE	United Kingdom				Human R	1	0	0	0	0	0	0	0	0	
6	2013	{'Rubrik,	95050	Santa Cla	TRUE	United St	CA	37.354	-121.953	Compute	0	0	0	0	1	0	1	0	0	
7	2015	{'Move24	10437		FALSE	Germany				Infrastru	0	0	0	0	0	1	0	0	0	
8	2012	{'Appboy	94103	San Fran	TRUE	United St	CA	37.7726	-122.41	Internet	0	0	1	1	1	0	0	0	1	
9	2013	{'Riskifie	10018	New York	TRUE	United St	NY	40.7553	-73.9933	Internet	0	1	0	0	1	0	1	0	1	
10	2011	{'Mesosp	94107	San Fran	TRUE	United St	CA	37.7618	-122.399	Compute	0	0	0	0	1	0	0	0	0	
11	2013	{'Fugue':	21701	Frederick	TRUE	United St	MD	39.4242	-77.3669	Compute	0	0	0	0	0	0	0	0	0	
12	2007	{'Synthes	not found		FALSE	United Kingdom				Internet	0	0	0	1	1	0	0	1	0	
13	2013	{'Docker,	94105	San Fran	TRUE	United St	CA	37.789	-122.392	Compute	0	0	0	1	0	0	0	0	0	
14	2012	{'Cylance	2116	Boston	TRUE	United St	MA	42.3531	-71.0765	Compute	0	0	0	0	0	0	1	0	1	
15	2011	{'Xamarir	94111	San Fran	TRUE	United St	CA	37.7959	-122.4	Compute	0	0	1	1	0	1	0	0	0	
16	2012	{'foodora	not found		FALSE	United Kingdom				Food Bus	1	0	0	0	0	0	0	0	0	
17	2010	{'Shareth	94107	San Fran	TRUE	United St	CA	37.7618	-122.399	Marketin	0	0	1	0	0	0	0	0	0	
18	2008	{'PolicyBa	60000		FALSE	India				Financia	0	0	0	0	0	0	0	0	0	
19	2010	{'Corindu	2421	Lexingtor	TRUE	United St	MA	42.4287	-71.228	Healthca	0	0	0	0	0	0	0	0	0	
20	2005	{'SolarWi	78726	Austin	TRUE	United St	TX	30.4191	-97.8395	Informat	0	0	0	0	1	0	0	0	0	
21	2006	{'Axway':	95054	Santa Cla	TRUE	United St	CA	37.3986	-121.964	Compute	0	0	1	1	1	1	0	0	0	
22	2007	{'Duoling	10119	New York	FALSE	Germany	NY	40.7502	-73.9931	Educatio	0	0	1	1	0	0	0	0	0	
23	2009	{'Airbnb':	94301	Palo Alto	TRUE	United St	CA	37.4458	-122.158	Internet	0	1	1	0	0	0	0	0	0	
24	2012	{'Blue Ap	10001	New York	TRUE	United St	NY	40.7537	-73.9992	Consume	0	0	0	0	0	0	0	0	0	
25	2010	{'eGym B	80333		FALSE	Germany				Healthca	0	0	0	0	1	0	0	0	0	
26	2011	{'Ionic Se	30308	Atlanta	TRUE	United St	GA	33.7717	-84.3726	Compute	0	1	0	0	0	0	1	0	0	
27	2010	{'RetailN	94110	San Fran	TRUE	United St	CA	37.7486	-122.418	Informat	0	1	0	0	1	0	0	0	0	
28	2011	{'Localyti	2110	Boston	TRUE	United St	MA	42.3613	-71.0483	Internet	0	0	0	0	0	0	0	0	0	
29	2011	{'Oculus	33322	Fort Lauc	FALSE	United St	FL	26.1545	-80.2712	Consume	0	0	0	0	0	0	0	0	0	
30	2001	{'é??é??a	20033		FALSE	China				Financia	0	0	1	0	0	0	0	0	0	
31	2012	{'Snowfla	94401	San Mate	TRUE	United St	CA	37.5794	-122.316	Compute	0	0	0	0	1	0	0	0	0	
32	2012	{'Courser	94041	Mountain	TRUE	United St	CA	37.3868	-122.075	Internet	0	0	0	0	0	0	0	0	0	
33	2007	{'HPE Sim	46225	Indianap	TRUE	United St	IN	39.738	-86.1599	Compute	0	0	0	0	0	0	0	0	0	
34	2011	{'DogVac	98121	Seattle	TRUE	United St	WA	47.6152	-122.35	Consume	0	0	0	0	0	0	0	0	0	
35	2012	{'DriveTir	85281	Tempe	TRUE	United St	AZ	33.4367	-111.94	Internet	0	1	0	0	0	0	0	0	0	
36	2010	{'Paytm':	56003	Mankato	FALSE	iran	MN	44.1953	-94.0491	Internet	0	0	0	0	0	0	0	0	0	
37	2007	{'Silicon	90245	El Segun	TRUE	United St	CA	33.915	-118.405	Internet	1	0	0	0	0	0	0	0	0	
38	2008	{'Conduct	10016	New York	TRUE	United St	NY	40.7473	-73.9801	Internet	0	0	0	0	0	0	0	0	0	
39	2009	{'HealthT	33126	Miami	TRUE	United St	FL	25.7786	-80.2962	Informat	0	0	1	1	0	0	0	0	0	
40	2006	{'SolarWi	78746	Austin	TRUE	United St	TX	30.2962	-97.8299	Internet	0	0	0	0	0	0	0	0	0	
41	2009	{'Facebo	94041	Mountain	TRUE	United St	CA	37.3868	-122.075	Internet	0	0	1	0	0	0	0	0	1	
42	2011	{'Babbel'	15206	Pittsburg	TRUE	United St	PA	40.4721	-79.9137	Educatio	0	0	1	1	0	0	0	0	0	
43	2007	{'F5 Netw	98101	Seattle	TRUE	United St	WA	47.6085	-122.336	Compute	0	0	0	0	1	0	0	0	1	
44	2010	{'OneTwo	not found		FALSE	Russia				Internet	0	1	0	0	0	0	0	0	0	
45	2010	{'Jenkins	95110	San Jose	TRUE	United St	CA	37.3546	-121.919	Compute	1	0	0	0	0	0	0	0	0	
46	2008	{'Andela'	10011	New York	TRUE	United St	NY	40.7465	-74.0094	Educatio	0	0	0	0	0	0	0	0	0	
47	2009	{'Bizagi':	92100		FALSE	France				Compute	0	0	0	0	0	0	0	0	0	
48	2010	{'B2W Dig	22031	Fairfax	FALSE	Brazil	VA	38.8642	-77.2578	Informat	0	0	1	1	0	0	1	0	0	
49	2009	{'Stitch Fi	94104	San Fran	TRUE	United St	CA	37.7911	-122.402	Consume	0	1	0	0	0	0	0	0	0	
50	2004	{'achica':	not found		FALSE	United Kingdom				Internet	1	1	0	0	0	0	0	0	0	
51	2006	{'Facebo	94103	San Fran	TRUE	United St	CA	37.7726	-122.41	Internet	0	0	1	0	0	0	0	0	0	
52	2005	{'Caktus	27607	Raleigh	TRUE	United St	NC	35.8019	-78.6875	Compute	1	0	0	1	0	0	0	0	0	
53	2010	{'Instagra	90291	Venice	TRUE	United St	CA	33.9962	-118.469	Compute	0	0	0	0	0	0	0	0	0	
54	2011	{'PullStri	94108	San Fran	TRUE	United St	CA	37.7909	-122.408	Entertair	0	0	0	1	0	0	0	0	1	
55	2005	{'Joystick	10011	New York	TRUE	United St	NY	40.7465	-74.0094	Internet	0	0	0	0	0	0	0	1	0	

article_after_processing10

Data Exploration to Recommendation Generation

Data Exploration/ Visualization

Data exploration is to gain insights for algorithm selection, feature selection, feature transformation through following steps:

- Summary statistics, variable category, NA value detection
- Univariate analysis
- Bivariate analysis

Documentation and codes:

Recommendation Engine

Based on the inputs from the data exploration, we create the recommendation and generate recommendations in the following steps:

- Algorithm Selection
- Feature Transformation and Engineering
- Recommendation Output

Documentation and codes:

Summary Statistics, Variable Category, and NA values

Refer the document for the details: explatory_data_analysis/explatory_data_analysis

```r
summary(data[, 1:13 ])
```

```
##      published_at   funding_round money_raised_float
##  1/14/2016 :  3             : 12   Min.   : 10.00
##  7/29/2015 :  3   Series B: 48   1st Qu.: 15.00
##  11/15/2016:  2   series C:  1   Median : 25.00
##  11/3/2015 :  2   Series C:152   Mean   : 41.17
##  12/11/2013:  2   Series D: 14   3rd Qu.: 45.00
##  2/11/2008 :  2   Series E:  5   Max.   :793.50
##  (Other)   :218
##                                              CompanyName     CompanySize
##  2U                                                :  1   51-200   :101
##  3D Robotics                                       :  1   201-500  : 64
##  aCommerce - Ecommerce Solutions for Southeast Asia:  1   Nov-50   : 27
##  Affle                                             :  1   1001-5000: 21
##  App Annie                                         :  1   501-1000 : 14
##  Appear Here                                       :  1   10,001+  :  2
##  (Other)                                           :226   (Other)  :  3
##     Founded                City     address_check           Country
##  Min.   :1939   San Francisco:49   False: 59     United States :178
##  1st Qu.:2007                :46   True :173     United Kingdom: 17
##  Median :2010   New York     :27                 Germany       :  8
##  Mean   :2009   Mountain View:11                 Canada        :  4
##  3rd Qu.:2012   San Mateo    : 8                 India         :  4
##  Max.   :2017   Boston       : 6                 Singapore     :  4
##                 (Other)      :85                 (Other)       : 17
```

```r
# check columns 1:13. Columns 13: have same format.
str(data[, 1:13 ])
```

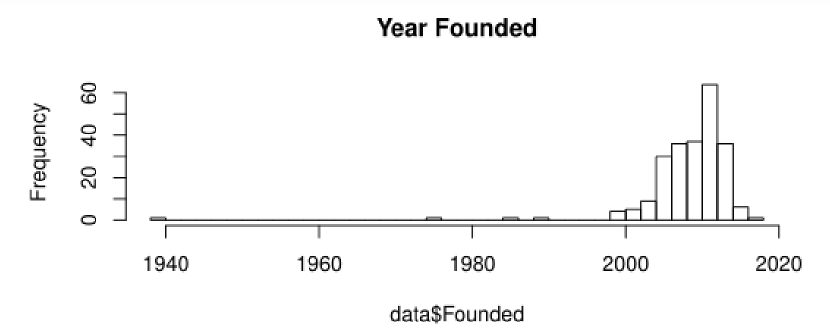
```
## 'data.frame':    232 obs. of  13 variables:
##  $ published_at                 : Factor w/ 209 levels "1/11/2010","1/14/2016",..: 189 176 160 15
##  $ funding_round                : Factor w/ 6 levels "","Series B",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ money_raised_float           : num  45 39 40 48 90 20.2 29 32 36 20 ...
##  $ CompanyName                  : Factor w/ 232 levels "2U","3D Robotics",..: 25 29 185 126 39 1
##  $ CompanySize                  : Factor w/ 8 levels "10-Jan","10,001+",..: 4 7 7 7 4 4 7 7 7 7 .
##  $ Founded                      : num  2013 2013 2015 2011 2013 ...
##  $ City                         : Factor w/ 68 levels "","Arlington",..: 1 1 36 1 55 1 52 36 52 2
##  $ address_check                : Factor w/ 2 levels "False","True": 1 1 2 1 2 1 2 2 2 2 ...
##  $ Country                      : Factor w/ 23 levels "Belgium","Brazil",..: 8 22 23 22 23 7 23 2
##  $ latitude                     : num  NA NA 40.7 NA 37.4 ...
##  $ longitude                    : num  NA NA -74 NA -122 ...
##  $ Industry_consolidated        : Factor w/ 16 levels "Computer & Network Security & Hardware",..
##  $ spc_Logistics.and.Supply.Chain: int  0 0 0 1 0 0 0 0 0 0 ...
```

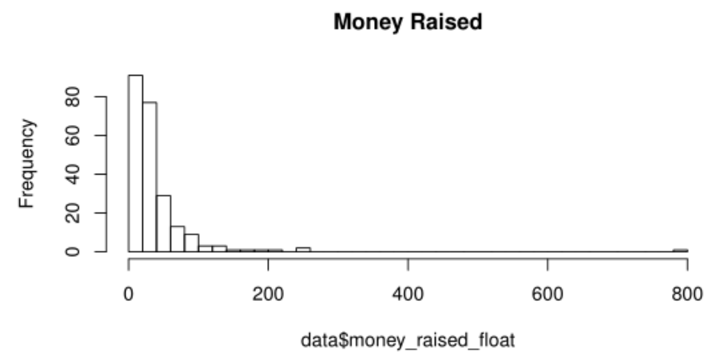
```r
# show columns with na
na = lapply(data, function(x) sum(ifelse(is.na(x) | x == "" | x == "not found", TRUE, FALSE)))
na[na > 0]
```

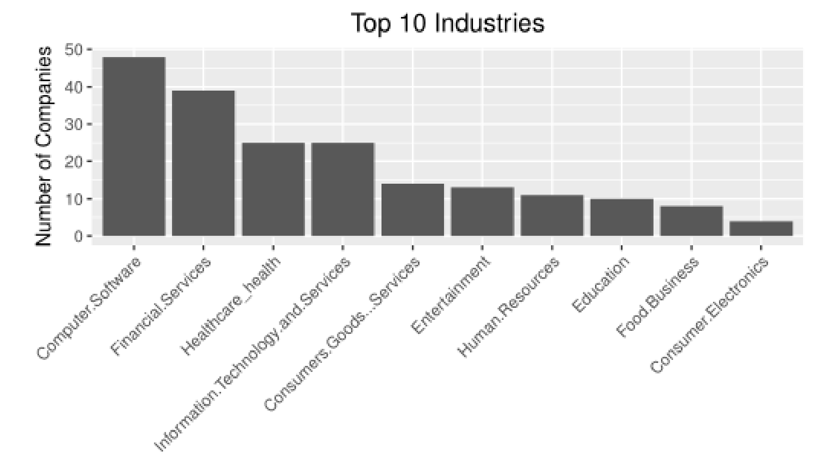
```
## $funding_round
## [1] 12
##
## $City
## [1] 46
##
## $latitude
## [1] 46
##
```

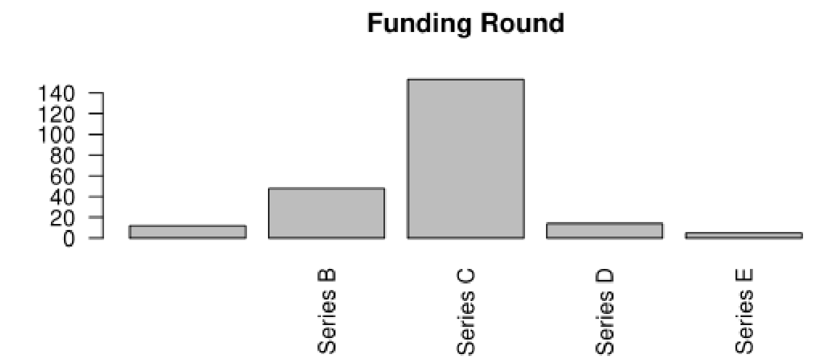
Univariate analysis

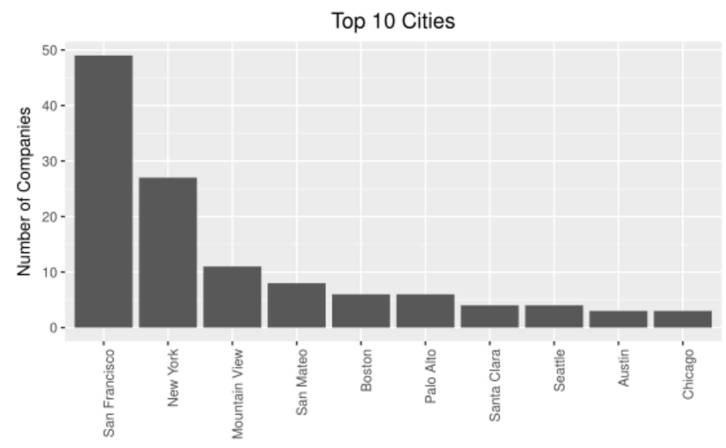
Refer the document for the details: explatory_data_analysis/explatory_data_analysis





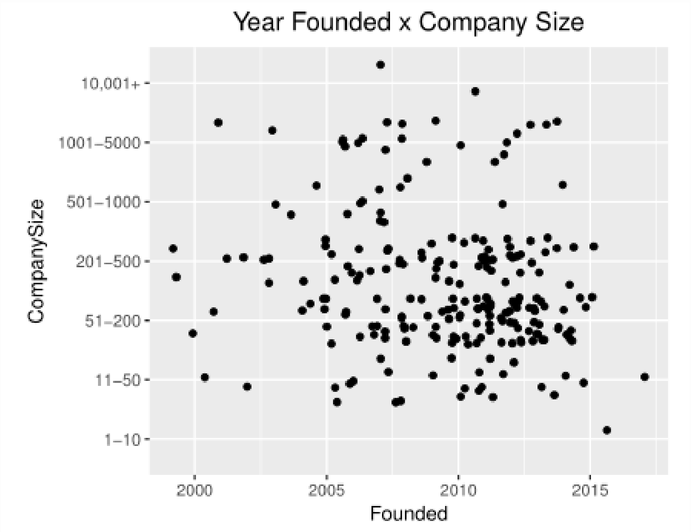


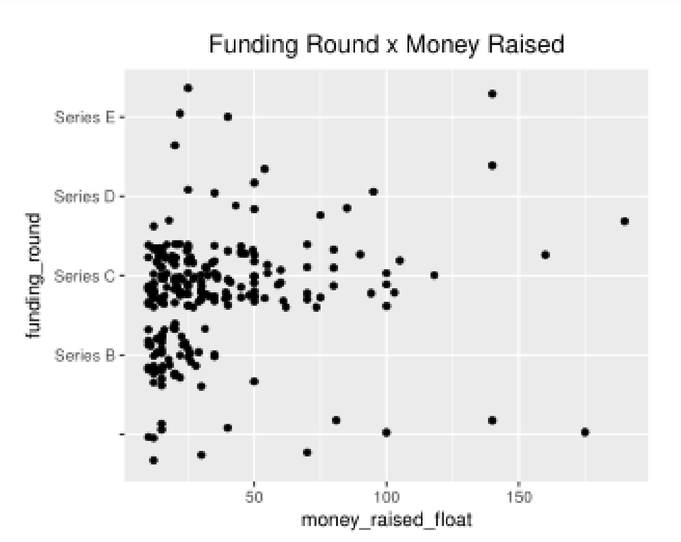


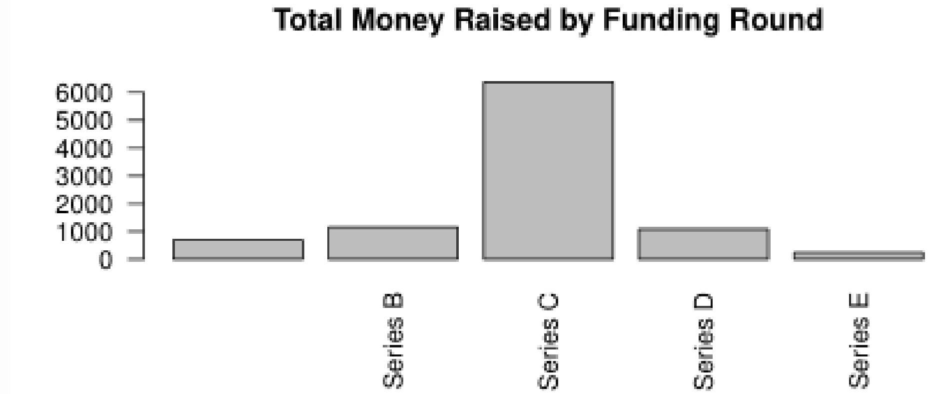


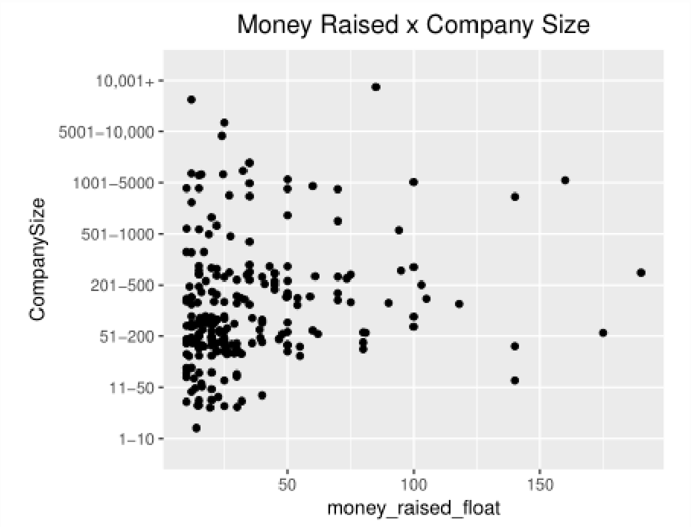
Bivariate analysis

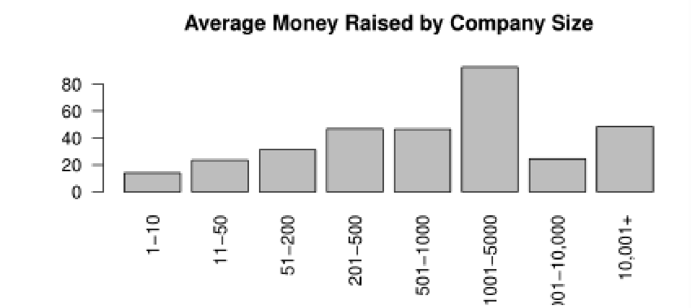
Refer the document for the details: explatory_data_analysis/explatory_data_analysis

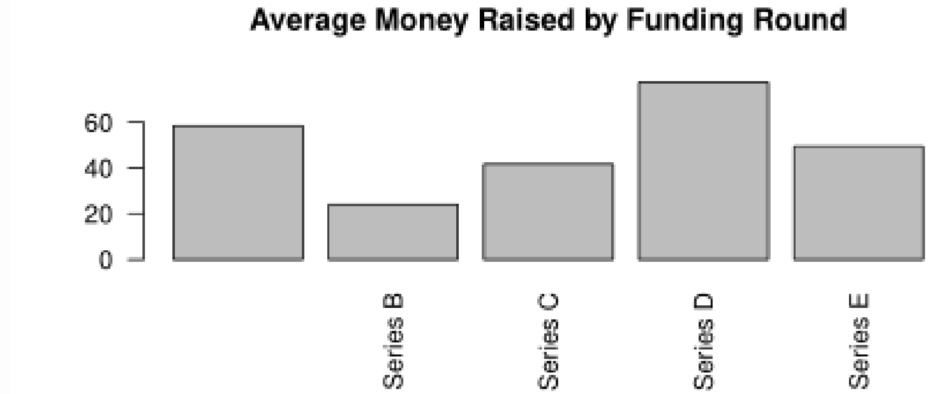






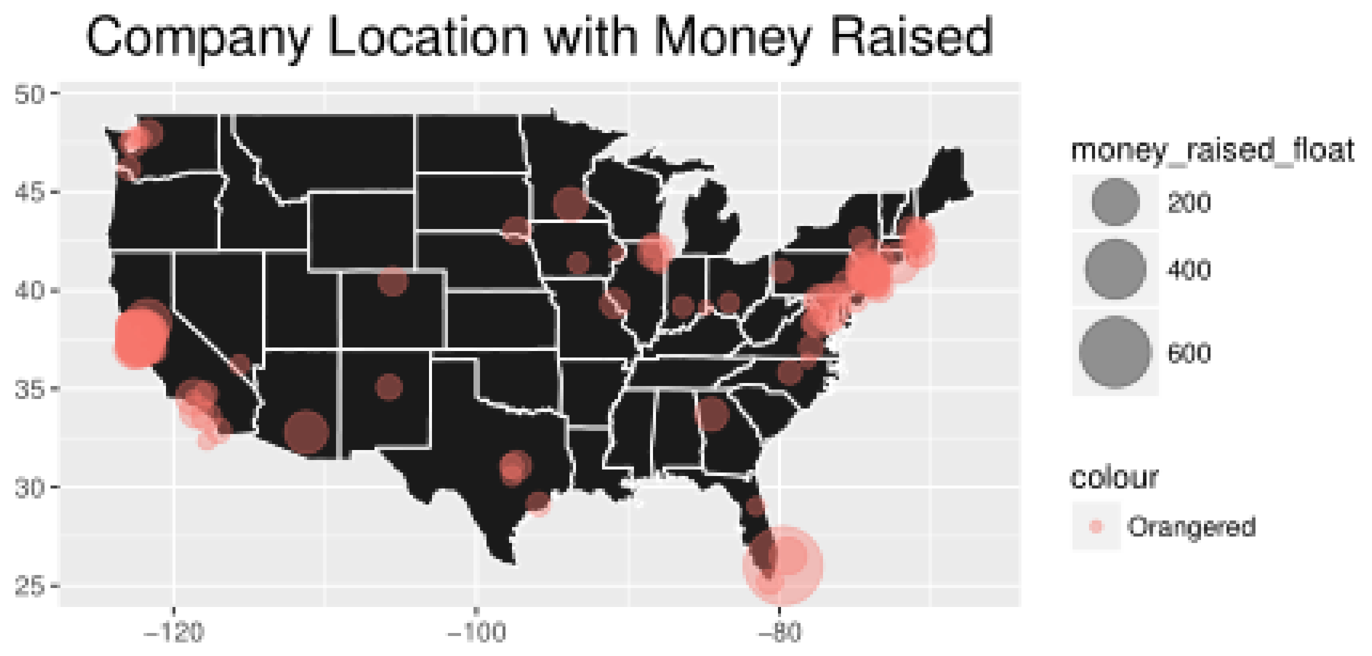


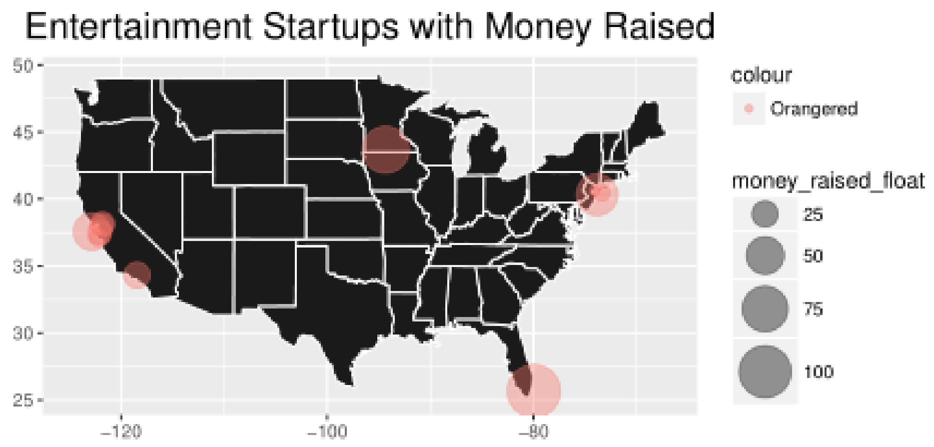


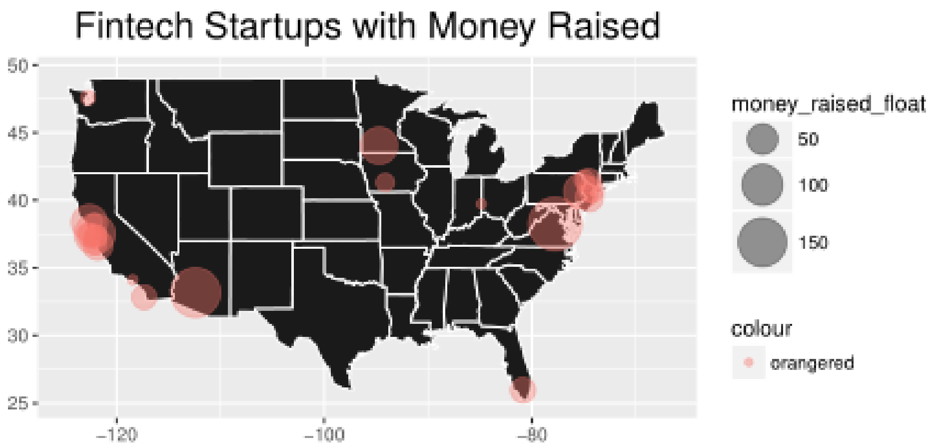


Bivariate analysis

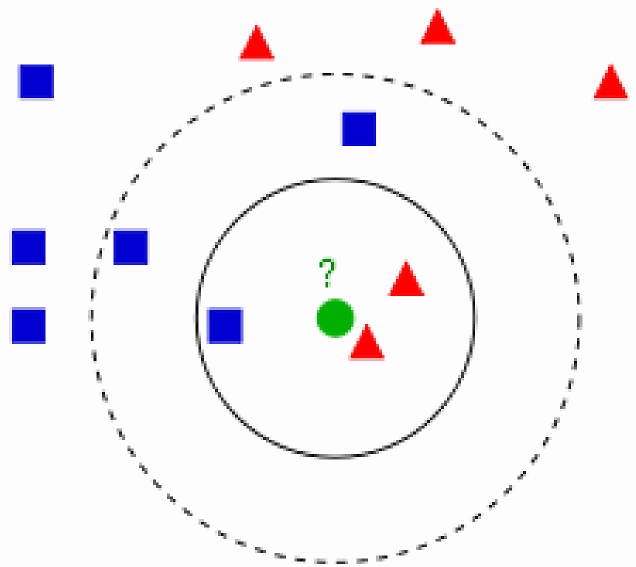
Refer the document for the details: explatory_data_analysis/explatory_data_analysis







Recommendation Engine – K-Nearest Neighbor



How the recommendation engine should work:

Given user inputs such as industry, company size, and year founded, it provides a few companies that matches the inputs.

Algorithm choice: K-nearest neighbors (KNN)

Justification for the choice: KNN works well for multi-class problems like this problem where we want to assign the user input to a label (company) as outputs out of all the different labels. It also produces several neighbors which we can use as a secondary recommendations for the user.

How KNN works: A green dot as the user input and other dots are startups in the database. KNN calculates the distance between the green dot and other dots and come up with K dots that are closest to the green dot. The shorter the distance is is, the better matches between the input and the neighbors are. These neighbors become the recommendation.

Feature Transformation and Engineering

KNN requires features to be scaled properly because KNN is distance-based algorithm and calculates a selected distance metric between the user inputs and each example of the training data.
This implies that KNN only takes a numeric variable and a dummy variable. Thus, I made transformations as followings for the features.

Features	Money Raised	Company Size	Location	Year Founded	Industry
Transformations	Log transformation to handle outliers	Binning	Convert city to geo coordinate	Removal of outliers	Key word matching
	Min max transformation	Min max transformation	Min max transformation	Min max transformation	Dummy variables

Refer the document for the details: recommendation/recommendation

Recommendation Output

```
In [51]: generate_recommendation(train, test.ix[0:0,], x_test.ix[0:0,])

---------------------------------------------------------------------------------------------------
Thank you for providing your interests! Below are the summary of your interests

Headquarters:        San Francisco
Year founded:        2015
Company size:        11-50
Industry:            Education & Internet
---------------------------------------------------------------------------------------------------
We recommend to check 'Edmodo' that matches your interests!

About the start up

Our mission is to connect all learners to the people and resources needed to achieve their full potential. We are the world's l
eading global education network that provides communication, collaboration, and coaching tools for all members of the school co
mmunity. We were founded in 2008 and currently have over 70 million members across 350,000+ schools in 150 countries.

The investors backing Edmodo are some of the best-recognized firms in the world, including Benchmark Capital, Greylock Venture
s, Index Ventures, Union Square Ventures, Learn Capital and our Chairman is Reid Hoffman, founder of LinkedIn.

So join the team that is changing how teachers and students learn - change lives, build your career and rack up the karma.

Company details

Website:             http://www.edmodo.com
Headquarters:        San Mateo, CA
Year founded:        2008
Company size:        51-200
Techcrunch article:  https://techcrunch.com/2012/07/19/nea-leads-educational-network-edmodos-25-million-series-c/

---------------------------------------------------------------------------------------------------
We also suggest checking following startups

                Company  Money_raised  Founded Company Size          City
99       Varsity Tutors            50    2,007     201-500    Saint Louis
20             Coursera            50    2,012     201-500  Mountain View
158  Boomerang Commerce            12    2,012      51-200  Mountain View
42          Engine Yard            19    2,006      51-200  San Francisco
135            Base CRM            15    2,009     201-500  Mountain View
---------------------------------------------------------------------------------------------------
```

Summery of user inputs

Summary of the top recommendation

Secondary recommendations

Refer the document for the details: recommendation/recommendation

Future Development

Data collection	Incorporate more data sources such as Glassdoor. Create data pipeline that is based on once a day batch processing from multiple data sources. Improve algorithms for various data extraction works by utilizing existing NLP packages.
Data storage	Store the data in database such as PostgreSQL for better data management and data retrieving capability.
Data preprocessing	Clean up codes and streamline the process. Incorporate better handlings.
Recommendation Engine	Store companies also-viewed for each company profile at linked in Graph DB such as Neo4j and generate startup recommendations based on the DB.
Interface	Create a Web application using Flask and develop GUI to enable users to input their preferences and to view recommendation outputs.