Start-Up Analysis

Final Report

Venkata Krishnan Ravichandran

1. Introduction and Research Questions

Startups play a crucial role in driving innovation and economic growth and have the potential to create significant value for investors and society as a whole. However, launching and scaling a startup is a challenging and risky undertaking, and many startups fail in their early stages. Understanding the factors that contribute to startup success or failure is therefore of great interest to entrepreneurs, investors, and researchers.

The Big Startup Success/Fail Dataset is a comprehensive and informative dataset that provides a wealth of data on startups that have either succeeded or failed in the past. The dataset was obtained from Kaggle, a platform for sharing and discovering datasets and other resources related to data science and machine learning.

This project aims to analyze the Big Startup Success/Fail Dataset to gain insights into the factors that contribute to startup success or failure. Specifically, we aim to answer the following research questions:

1. What are the key factors that contribute to startup success or failure, such as funding rounds, founding year, industry, location, and other key metrics?
2. How does the impact of these factors vary across different industries and locations?

To answer these questions, we will use a variety of statistical and machine learning techniques, including data cleaning, exploratory data analysis, regression analysis, and classification algorithms.

The motivation for this project is to provide valuable insights for entrepreneurs, investors, and policymakers seeking to improve the success rate of startups. By analyzing this dataset, we can identify patterns and trends that can inform strategic decision-making, such as identifying promising industries or geographic regions for startups or understanding the importance of certain factors such as funding and founding year.

In conclusion, this project seeks to leverage the power of analytics to study a comprehensive and informative dataset on startup success and failure. By answering our research questions, we aim to provide valuable insights for entrepreneurs, investors, and policymakers seeking to improve the success rate of startups and drive innovation and economic growth.

2. Data Collection

The data was mined and engineered from various sources and file formats. They were then extracted and transformed to the format of our requirement. Transformation of data included Data Cleaning. Steps like Removing Outliers, Identifying Missing Values, dropping missing values, imputing missing values and Interpolation were followed to ensure Data Integrity. The data dictionary provided a comprehensive guide that outlined the metadata including data types, data constraints, relationships between data elements, and other relevant information. Our data dictionary contains information like Variable and Description which provides us with information about the predictors available and their nature.

Data Sample:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	permalink	name	homepage_u	category_list	funding_tota	status	country_cod	state_code	region	city	funding_rou	founded_at	first_funding	last_funding_at
2	/organizatior	#fame	http://livfam	Media	10000000	operating	IND	16	Mumbai	Mumbai	1		1/5/15	1/5/15
3	/organizatior	:Qounter	http://www.	Application F	700000	operating	USA	DE	DE - Other	Delaware Cit	2	9/4/14	3/1/14	10/14/14
4	/organizatior	(THE) ONE of	http://oneof	Apps|Games	3406878	operating					1		1/30/14	1/30/14
5	/organizatior	0-6.com	http://www.	Curated Web	2000000	operating	CHN	22	Beijing	Beijing	1	1/1/07	3/19/08	3/19/08
6	/organizatior	004 Technol	http://004gr	Software	-	operating	USA	IL	Springfield, Il	Champaign	1	1/1/10	7/24/14	7/24/14
7	/organizatior	01Games Te	http://www.	Games	41250	operating	HKG		Hong Kong	Hong Kong	1		7/1/14	7/1/14
8	/organizatior	Ondine Biom	http://ondin	Biotechnolog	762851	operating	CAN	BC	Vancouver	Vancouver	2	1/1/97	9/11/09	12/21/09
9	/organizatior	H2O.ai	http://h2o.a	Analytics	33600000	operating	USA	CA	SF Bay Area	Mountain Vi	4	1/1/11	1/3/13	11/9/15
10	/organizatior	One Inc.	http://whati	Mobile	1150050	operating	USA	CA	SF Bay Area	San Francisc	3	8/1/11	7/20/11	2/5/14
11	/organizatior	1,2,3 Listo	http://www.	E-Commerce	40000	operating	CHL	12	Santiago	Las Condes	1	1/1/12	2/18/13	2/18/13
12	/organizatior	1-4 All		Entertainmer	-	operating	USA	NC	NC - Other	Connellys Sp	1		4/21/13	4/21/13
13	/organizatior	1.618 Techn	http://www.	Networking|	-	operating	USA	FL	Orlando	Orlando	1	12/7/13	1/22/14	1/22/14
14	/organizatior	1-800-DENTI	http://www.	Health and W	-	operating	USA	CA	Los Angeles	Los Angeles	1	1/1/86	8/19/10	8/19/10
15	/organizatior	1-800-DOCT	http://1800	Health and W	1750000	operating	USA	NJ	Newark	Iselin	1	1/1/84	3/2/11	3/2/11
16	/organizatior	1-800-Public	http://www.	Internet Mar	6000000	operating	USA	NY	New York Cit	New York	1	10/24/13	2/1/15	2/1/15
17	/organizatior	1 Mainstrear	http://www.	Apps|Cable|	5000000	acquired	USA	CA	SF Bay Area	Cupertino	1	3/1/12	3/17/15	3/17/15
18	/organizatior	1 of 99		Entertainmer	100000	operating	USA	CA	SF Bay Area	Mountain Vi	1	12/1/14	12/13/14	12/13/14
19	/organizatior	10-20 Media	http://www.	E-Commerce	2050000	operating	USA	MD	Baltimore	Woodbine	4	1/1/01	6/18/09	12/28/11
20	/organizatior	10 Minutes \	http://10mir	Education	4400000	operating	GBR	H9	London	London	2	1/1/13	1/1/13	10/9/14
21	/organizatior	1000 Corks	http://1000	Search	40000	operating	USA	OR	Portland, Ore	Lake Oswego	1	1/1/08	8/23/11	8/23/11
22	/organizatior	1000 Market	http://www.	Art|E-Comm	500000	acquired	USA	WA	Seattle	Seattle	1	1/1/09	5/15/09	5/15/09
23	/organizatior	Beijing 1000	http://www.	Mobile	43923865	operating					1	1/1/08	4/1/10	4/1/10
24	/organizatior	1000Lookz	http://1000	Beauty	-	operating	IND	25	Chennai	Chennai	1	1/1/08	7/22/13	7/22/13
25	/organizatior	1000memor	http://1000	Curated Web	2535000	acquired	USA	CA	SF Bay Area	San Francisc	2	7/1/10	1/1/10	2/16/11

Data Dictionary:

Variable	**Description**
permalink	Link to Organization
name	Name of the Startup
homepage_url	Website URL of the Startup
category_list	Field of the company
funding_total_usd	Total funding compensation received
status	Operational status

country_code	The country of Origin
state_code	The state at which the company is located
region	Region of company location
city	City of company location
funding_rounds	The number of rounds company went for funding
founded_at	When the company was established
first_funding_at	Date at which the company was first funded
last_funding_at	The last time the company was funded

3. Data Preprocessing

The dataset, which was obtained from various sources, contains 66368 rows and 14 columns. Real-world datasets are often imperfect, containing missing values, incorrect data types, unreadable characters, and unexpected values. Poor data quality can manifest in various forms, such as empty cells, data in the wrong format, outliers, and duplicates. Empty cells refer to data points that are not present in the dataset, typically represented as NaN (Not a Number) in Pandas. These missing values can be handled by dropping them, filling them with a constant value, or interpolating them to estimate the missing value. All the NaN and NULL values have been removed.

Data Summary:

column	#_Null_values	%NULL values	#of unique values	Sample data
permalink	0	0.00%	66368	['/organization/-fame' '/organization/-qounter' '/organization/-the-one-of-them-inc-' '/organization/0-6-com' '/organization/004-technologies' '/organization/01games-technology' '/organization/0ndine-biomedical-inc' '/organization/0xdata' '/organization/1' '/organization/1-2-3-listo']
name	1	0.00%	66102	['#fame' ':Qounter' '(THE) ONE of THEM,Inc.' '0-6.com' '004 Technologies' '01Games Technology' 'Ondine Biomedical Inc.' 'H2O.ai' 'One Inc.' '1,2,3 Listo']
homepage_url	5058	7.62%	61191	['http://livfame.com' 'http://www.qounter.com' 'http://oneofthem.jp' 'http://www.0-6.com' 'http://004gmbh.de/en/004-interact' 'http://www.01games.hk/' 'http://ondinebio.com' 'http://h2o.ai/' 'http://whatis1.com' 'http://www.123listo.com']
category_list	3148	4.74%	27296	['Media' 'Application Platforms|Real Time|Social Network Media' 'Apps|Games|Mobile' 'Curated Web' 'Software' 'Games' 'Biotechnology' 'Analytics' 'Mobile' 'E-Commerce']
funding_total_usd	0	0.00%	18896	['10000000' '700000' '3406878' '2000000' '-' '41250' '762851' '33600000' '1150050' '40000']
status	0	0.00%	4	['operating' 'acquired' 'closed' 'ipo']
country_code	6958	10.48%	137	['IND' 'USA' nan 'CHN' 'HKG' 'CAN' 'CHL' 'GBR' 'FRA' 'AUS']
state_code	8547	12.88%	311	['16' 'DE' nan '22' 'IL' 'BC' 'CA' '12' 'NC' 'FL']
region	8030	12.10%	1092	['Mumbai' 'DE - Other' nan 'Beijing' 'Springfield, Illinois' 'Hong Kong' 'Vancouver' 'SF Bay Area' 'Santiago' 'NC - Other']
city	8028	12.10%	5111	['Mumbai' 'Delaware City' nan 'Beijing' 'Champaign' 'Hong Kong' 'Vancouver' 'Mountain View' 'San Francisco' 'Las Condes']
funding_rounds	0	0.00%	19	[1 2 4 3 9 5 6 7 8 10]
founded_at	15221	22.93%	3978	[nan '2014-09-04' '2007-01-01' '2010-01-01' '1997-01-01' '2011-01-01' '2011-08-01' '2012-01-01' '2013-12-07' '1986-01-01']
first_funding_at	24	0.04%	4817	['2015-01-05' '2014-03-01' '2014-01-30' '2008-03-19' '2014-07-24' '2014-07-01' '2009-09-11' '2013-01-03' '2011-07-20' '2013-02-18']
last_funding_at	0	0.00%	4518	['2015-01-05' '2014-10-14' '2014-01-30' '2008-03-19' '2014-07-24' '2014-07-01' '2009-12-21' '2015-11-09' '2014-02-05' '2013-02-18']

```
df['funding_total_usd'] = df['funding_total_usd'].replace('-',np.nan)
```

```
df
```

	permalink	name	homepage_url	category_list	funding_total_usd	status	country_code	state_code	region
0	/organization/-fame	#fame	http://livfame.com	Media	10000000	operating	IND	16	Mumbai
1	/organization/-qounter	:Qounter	http://www.qounter.com	Application Platforms|Real Time|Social Network...	700000	operating	USA	DE	DE - Other
2	/organization/-the-one-of-them-inc-	(THE) ONE of THEM,Inc.	http://oneofthem.jp	Apps|Games|Mobile	3406878	operating	NaN	NaN	NaN
3	/organization/0-6-com	0-6.com	http://www.0-6.com	Curated Web	2000000	operating	CHN	22	Beijing
4	/organization/004-technologies	004 Technologies	http://004gmbh.de/en/004-interact	Software	NaN	operating	USA	IL	Springfield, Illinois
...	...	...	...	...	...	...	...	...	...
66363	/organization/zznode-science-and-technology-co...	ZZNode Science and Technology	http://www.zznode.com	Enterprise Software	1587301	operating	CHN	22	Beijing
66364	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	http://www.zzzzapp.com	Advertising|Mobile|Web Development|Wireless	114304	operating	HRV	15	Split
66365	/organization/Áeron	ÁERON	http://www.aeron.hu/	NaN	NaN	operating	NaN	NaN	NaN
66366	/organization/Ôasys-2	Ôasys	http://www.oasys.io/	Consumer Electronics|Internet of Things|Teleco...	18192	operating	USA	CA	SF Bay Area
66367	/organization/İnovatiff-reklam-ve-tanıtım-hizm...	İnovatiff Reklam ve Tanıtım Hizmetleri Tic	http://inovatiff.com	Consumer Goods|E-Commerce|Internet	14851	operating	NaN	NaN	NaN

66368 rows × 14 columns

Above, we replaced all the place holders with NULL values.

```
df[['category_list','country_code','state_code','region','city']] = \
df[['category_list','country_code','state_code','region','city']].fillna('others')

df['funding_total_usd'] = df['funding_total_usd'].fillna(0)

df = df.drop(columns = 'homepage_url', axis = 1)
df.dropna(inplace = True)
df
```

	permalink	name	category_list	funding_total_usd	status	country_code	state_code	region	city	funding_
1	/organization/-qounter	:Qounter	Application Platforms|Real Time|Social Network...	700000	operating	USA	DE	DE - Other	Delaware City	
3	/organization/0-6-com	0-6.com	Curated Web	2000000	operating	CHN	22	Beijing	Beijing	
4	/organization/004-technologies	004 Technologies	Software	0	operating	USA	IL	Springfield, Illinois	Champaign	
6	/organization/0ndine-biomedical-inc	Ondine Biomedical Inc.	Biotechnology	762851	operating	CAN	BC	Vancouver	Vancouver	
7	/organization/0xdata	H2O.ai	Analytics	33600000	operating	USA	CA	SF Bay Area	Mountain View	
...	...	...	...	...	...	...	...	...	...	
66361	/organization/zytoprotec	Zytoprotec	Biotechnology	2686600	operating	AUT	3	Vienna	Gerasdorf Bei Wien	
66362	/organization/zzish	Zzish	Analytics|Android|Developer APIs|Education|Gam...	1120000	operating	GBR	H9	London	London	
66364	/organization/zzzzapp-com	Zzzzapp Wireless ltd.	Advertising|Mobile|Web Development|Wireless	114304	operating	HRV	15	Split	Split	
66365	/organization/Áeron	ÁERON	others	0	operating	others	others	others	others	
66366	/organization/Ôasys-2	Ôasys	Consumer Electronics|Internet of Things|Teleco...	18192	operating	USA	CA	SF Bay Area	San Francisco	

51125 rows × 13 columns

In the above screenshot, we replaced all the null values with zero or considered as other categories.

```
# Changing the data types
df['founded_at'] = pd.to_datetime(df['founded_at'], errors='coerce')
df['first_funding_at'] = pd.to_datetime(df['first_funding_at'], errors='coerce')
df['last_funding_at'] = pd.to_datetime(df['last_funding_at'], errors='coerce')

# Getting the not null records
df = df.loc[~df['founded_at'].isnull()]
df = df.loc[~df['first_funding_at'].isnull()]
df = df.loc[~df['last_funding_at'].isnull()]

# Converting object data types into the necessary once.
df['funding_total_usd'] = df['funding_total_usd'].astype(float)

obj_col = df.dtypes[df.dtypes == 'object'].reset_index()['index'].tolist()

for i in obj_col:
    df[i] = df[i].astype(str)

df['city'] = df['city'].astype('str')
```

Changing the data types accordingly.

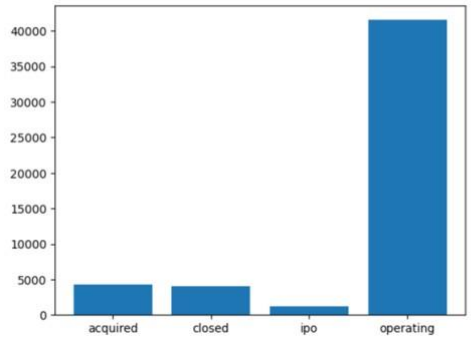
Number of Rows after cleaning: **51119**
Number of columns after cleaning: **13**

4. Data Analysis

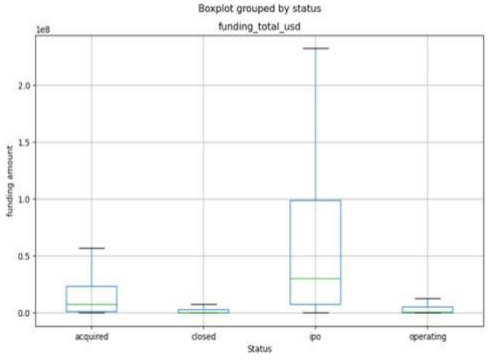
The final dataset has over 13 columns and 50 thousand + rows. The bar graph shows us that there are still 40,000 companies still under operation of all the companies listed. The boxplot gives us a representation of what the status of the startups are currently and we can see that most of them converted into IPOs. The correlation heatmap gives us the highest correlation between two variables. The tableau dashboards give us a clear representation that each region specialises in different domains of startups and the possibility of success depends majorly on the location. Hence, geospatial plotting would be ideal for the use case.

```
t = df.groupby('status', as_index = False)['name'].count()
plt.bar(t['status'], t['name'])
```

```
<BarContainer object of 4 artists>
```

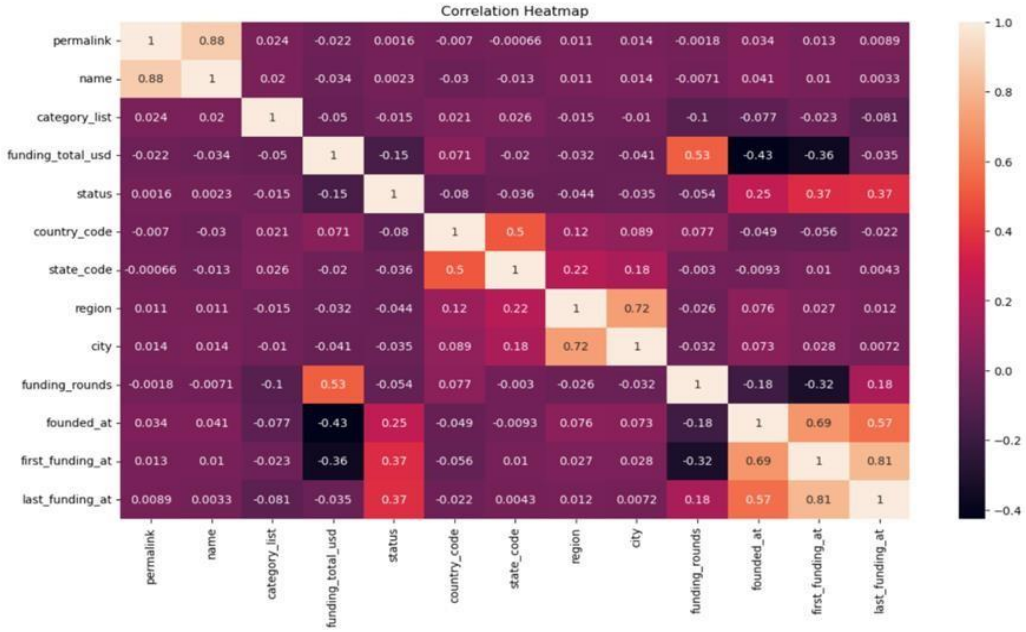



```
ax = df.boxplot(column='funding_total_usd', by='status', figsize=(10,6), showfliers=False)
ax.set_xlabel('Status')
ax.set_ylabel('funding amount')
plt.show()
```




```
import seaborn as sns
fig, ax1 = plt.subplots(figsize=(15, 8))
sns.heatmap(temp.corr(), annot = True)
plt.title("Correlation Heatmap")
```

```
Text(0.5, 1.0, 'Correlation Heatmap')
```



5. Limitations and Future Work

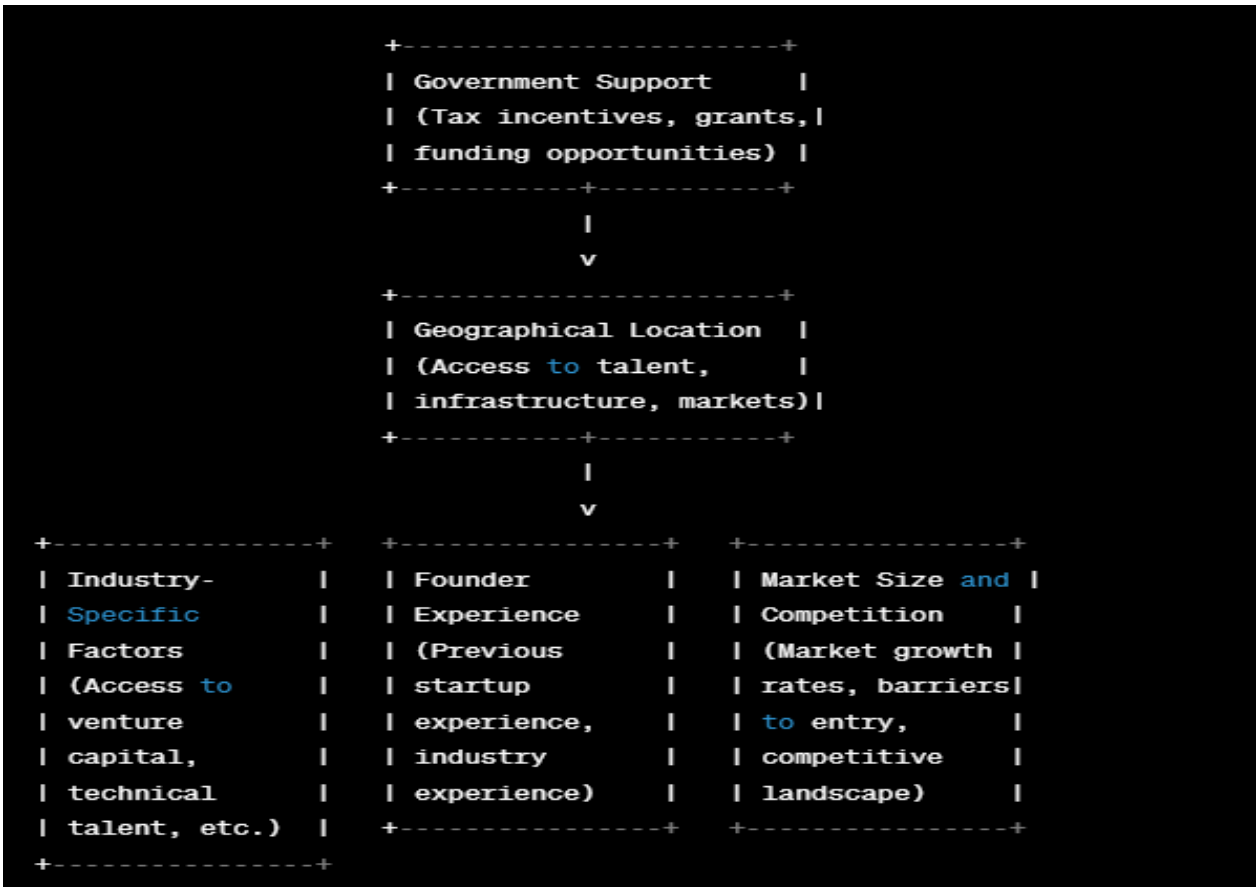
Limitations:

Data availability: One of the major limitations in analyzing startup companies over funding, category, and geography is the availability of accurate and up-to-date data. While there are various sources of information such as Crunchbase, AngelList, and Pitchbook, these databases can have incomplete, inconsistent, or outdated information.

Data quality: Even when data is available, the quality of the data can vary greatly, and it can be difficult to verify the accuracy of the information. This can lead to incorrect conclusions and inaccurate analysis.

Limited scope: The analysis of startup companies over funding, category, and geography is limited to the data that is available. It may not capture important factors such as the quality of the management team, the strength of the product, or the competitive landscape.

Future Work:



Analyzing additional attributes like government support and geographical location can provide valuable insights for startups. Here are some potential areas of focus for future analysis:

Government support: Examining the level of support that startups receive from their government can provide insight into the overall business climate and regulatory environment. This might include looking at factors such as tax incentives, grants, and funding opportunities specifically targeted at startups.

Geographical location: The location of a startup can have a significant impact on its success. Factors such as access to talent, infrastructure, and markets can all be affected by the company's location. In addition, different regions may have different regulatory environments or cultural attitudes towards entrepreneurship that can affect a startup's chances of success.

Industry-specific factors: Depending on the industry in which a startup operates, there may be additional factors that are particularly relevant to its success. For example, in the tech industry, access to venture capital funding and the availability of technical talent may be particularly important.

Founder experience: The experience and background of the startup's founders can also be an important factor in its success. For example, founders with previous startup experience or experience in the industry in which they are launching their business may be better equipped to navigate the challenges of starting a new venture.

Market size and competition: Analyzing the size of the market and the level of competition can help to identify potential opportunities and challenges for startups. This might include examining factors such as market growth rates, barriers to entry, and the competitive landscape.

By considering these additional attributes, it may be possible to gain a more comprehensive understanding of the factors that contribute to startup success or failure. However, it's important to note that there is no one-size-fits-all approach to analyzing startups, and different startups may face unique challenges and opportunities based on their individual circumstances.