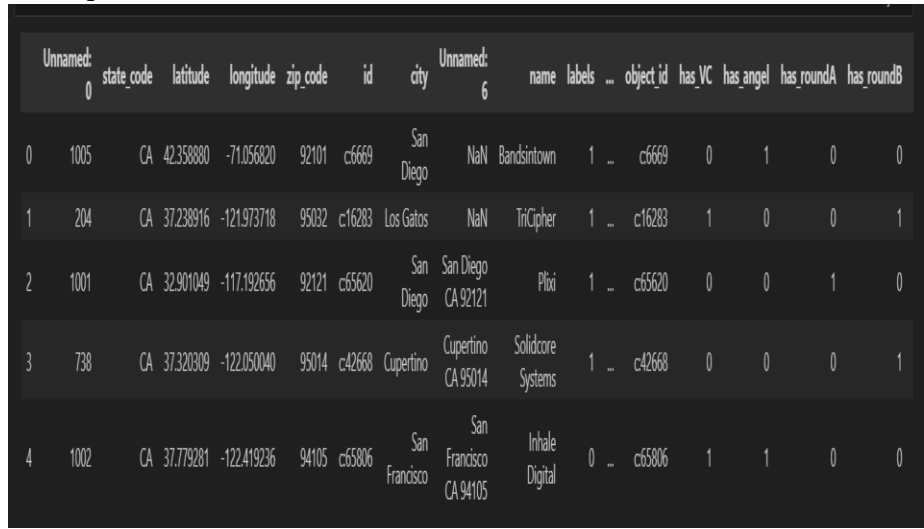


Data Collection and Preprocessing Phase

Date	7 June 2024
Team ID	team-739937
Project Title	prosperity Prognosticator : Machine Learning for Startup Success Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	<p><u>Dimension:</u> 923 rows \times 49 columns</p> <p><u>Descriptive statistics:</u></p> 
Univariate Analysis	

	<table border="1"><caption>Startup Distribution by State</caption><thead><tr><th>State</th><th>Percentage</th></tr></thead><tbody><tr><td>CA</td><td>52.9%</td></tr><tr><td>other</td><td>17.6%</td></tr><tr><td>NY</td><td>11.5%</td></tr><tr><td>MA</td><td>9.0%</td></tr><tr><td>TX</td><td>4.6%</td></tr><tr><td>WA</td><td>4.6%</td></tr></tbody></table>	State	Percentage	CA	52.9%	other	17.6%	NY	11.5%	MA	9.0%	TX	4.6%	WA	4.6%																																																																																														
State	Percentage																																																																																																												
CA	52.9%																																																																																																												
other	17.6%																																																																																																												
NY	11.5%																																																																																																												
MA	9.0%																																																																																																												
TX	4.6%																																																																																																												
WA	4.6%																																																																																																												
Bivariate Analysis	<table border="1"><caption>Startup Count by Category</caption><thead><tr><th>Category</th><th>Group 0</th><th>Group 1</th></tr></thead><tbody><tr><td>software</td><td>52</td><td>101</td></tr><tr><td>web</td><td>51</td><td>93</td></tr><tr><td>mobile</td><td>27</td><td>52</td></tr><tr><td>enterprise</td><td>17</td><td>56</td></tr><tr><td>advertising</td><td>17</td><td>45</td></tr><tr><td>games_video</td><td>21</td><td>31</td></tr><tr><td>semiconductor</td><td>11</td><td>24</td></tr><tr><td>biotech</td><td>12</td><td>22</td></tr><tr><td>network_hosting</td><td>10</td><td>24</td></tr><tr><td>hardware</td><td>16</td><td>11</td></tr><tr><td>public_relations</td><td>15</td><td>10</td></tr><tr><td>ecommerce</td><td>14</td><td>11</td></tr><tr><td>cleantech</td><td>13</td><td>10</td></tr><tr><td>analytics</td><td>3</td><td>16</td></tr><tr><td>security</td><td>4</td><td>15</td></tr><tr><td>social</td><td>6</td><td>8</td></tr><tr><td>search</td><td>5</td><td>7</td></tr><tr><td>messaging</td><td>4</td><td>7</td></tr><tr><td>other</td><td>9</td><td>2</td></tr><tr><td>news</td><td>1</td><td>7</td></tr><tr><td>travel</td><td>1</td><td>7</td></tr><tr><td>fashion</td><td>3</td><td>5</td></tr><tr><td>photo_video</td><td>2</td><td>5</td></tr><tr><td>medical</td><td>2</td><td>4</td></tr><tr><td>music</td><td>1</td><td>6</td></tr><tr><td>finance</td><td>2</td><td>4</td></tr><tr><td>education</td><td>1</td><td>3</td></tr><tr><td>real_estate</td><td>2</td><td>2</td></tr><tr><td>consulting</td><td>1</td><td>2</td></tr><tr><td>health</td><td>1</td><td>3</td></tr><tr><td>automotive</td><td>1</td><td>1</td></tr><tr><td>transportation</td><td>1</td><td>2</td></tr><tr><td>manufacturing</td><td>2</td><td>1</td></tr><tr><td>hospitality</td><td>1</td><td>1</td></tr><tr><td>sports</td><td>1</td><td>1</td></tr></tbody></table>	Category	Group 0	Group 1	software	52	101	web	51	93	mobile	27	52	enterprise	17	56	advertising	17	45	games_video	21	31	semiconductor	11	24	biotech	12	22	network_hosting	10	24	hardware	16	11	public_relations	15	10	ecommerce	14	11	cleantech	13	10	analytics	3	16	security	4	15	social	6	8	search	5	7	messaging	4	7	other	9	2	news	1	7	travel	1	7	fashion	3	5	photo_video	2	5	medical	2	4	music	1	6	finance	2	4	education	1	3	real_estate	2	2	consulting	1	2	health	1	3	automotive	1	1	transportation	1	2	manufacturing	2	1	hospitality	1	1	sports	1	1
Category	Group 0	Group 1																																																																																																											
software	52	101																																																																																																											
web	51	93																																																																																																											
mobile	27	52																																																																																																											
enterprise	17	56																																																																																																											
advertising	17	45																																																																																																											
games_video	21	31																																																																																																											
semiconductor	11	24																																																																																																											
biotech	12	22																																																																																																											
network_hosting	10	24																																																																																																											
hardware	16	11																																																																																																											
public_relations	15	10																																																																																																											
ecommerce	14	11																																																																																																											
cleantech	13	10																																																																																																											
analytics	3	16																																																																																																											
security	4	15																																																																																																											
social	6	8																																																																																																											
search	5	7																																																																																																											
messaging	4	7																																																																																																											
other	9	2																																																																																																											
news	1	7																																																																																																											
travel	1	7																																																																																																											
fashion	3	5																																																																																																											
photo_video	2	5																																																																																																											
medical	2	4																																																																																																											
music	1	6																																																																																																											
finance	2	4																																																																																																											
education	1	3																																																																																																											
real_estate	2	2																																																																																																											
consulting	1	2																																																																																																											
health	1	3																																																																																																											
automotive	1	1																																																																																																											
transportation	1	2																																																																																																											
manufacturing	2	1																																																																																																											
hospitality	1	1																																																																																																											
sports	1	1																																																																																																											
Outliers and Anomalies	-																																																																																																												

Data Preprocessing Code Screenshots

Loading Data

```
data= pd.read_csv('startup1.csv')
data
```

✓ 0.1s Python

	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	Unnamed: 6	name	labels	object id	has VC	has angel	has roundA	has roundB
0	1005	CA	42.358880	-71.056820	92101	c6669	San Diego	NaN	Bandsintown	1	c6669	0	1	0	0
1	204	CA	37.238916	-121.973718	95032	c16283	Los Gatos	NaN	TriCipher	1	c16283	1	0	0	1
2	1001	CA	32.901049	-117.192656	92121	c65620	San Diego	San Diego CA 92121	Plixi	1	c65620	0	0	1	0
3	738	CA	37.320309	-122.050040	95014	c42668	Cupertino	Cupertino CA 95014	Solidcore Systems	1	c42668	0	0	0	1
4	1002	CA	37.779281	-122.419236	94105	c65806	San Francisco	San Francisco CA 94105	Inhale Digital	0	c65806	1	1	0	0
...
918	352	CA	37.740594	-122.376471	94107	c21343	San Francisco	NaN	CoTweet	1	c21343	0	0	1	0
919	721	MA	42.504817	-71.195611	1803	c41747	Burlington	Burlington MA 1803	Reef Point Systems	0	c41747	1	0	0	1
920	557	CA	37.408261	-122.015920	94089	c31549	Sunnyvale	NaN	Paracor Medical	0	c31549	0	0	0	0
921	589	CA	37.556732	-122.288378	94404	c33198	San Francisco	NaN	Causata	1	c33198	0	0	1	1
922	462	CA	37.386778	-121.966277	95054	c26702	Santa Clara	Santa Clara CA 95054	Asempra Technologies	1	c26702	0	0	0	1

Handling Missing Data

```
#Filling missing value column(unnamed:6)
data['Unnamed: 6'] = data.apply(lambda row: (row.city) + " " + (row.state_code) + " " + (row.zip_code) , axis = 1)

# Total Missing Values column "Unnamed: 6"
totalNull = data['Unnamed: 6'].isnull().sum()

print('Total Missing Values Kolom "Unnamed: 6": ', totalNull)

Total Missing Values Kolom "Unnamed: 6": 0

#Filling missing values of column(closed_at)
data['closed_at'] = data['closed_at'].fillna(value="31/12/2013")
totalNull = data['closed_at'].isnull().sum()

print('Total Missing Values Kolom "closed_at": ', totalNull)

Total Missing Values Kolom "closed_at": 0
```

Data Transformation	<pre>data['status'] = data.status.map({'acquired':1, 'closed':0}) ✓ 0.0s data['status'].astype(int) ✓ 0.0s 0 1 1 1 2 1 3 1 4 0 .. 918 1 919 0 920 0 921 1 922 1 Name: status, Length: 923, dtype: int64</pre>
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-