

## Data Collection and Preprocessing Phase

Date	13 February 2026
Team ID	LTVIP2026TMIDS50820
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

	<p><u>Dimension:</u>        923 rows × 49 columns</p> <p><u>Descriptive statistics:</u></p>																																																																																																
Data Overview	<table border="1"> <thead> <tr> <th>Unnamed: 0</th> <th>state_code</th> <th>latitude</th> <th>longitude</th> <th>zip_code</th> <th>id</th> <th>city</th> <th>Unnamed: 6</th> <th>name</th> <th>labels</th> <th>...</th> <th>object_id</th> <th>has VC</th> <th>has angel</th> <th>has roundA</th> <th>has roundB</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>1005</td> <td>CA</td> <td>42.358880</td> <td>-71.056820</td> <td>92101</td> <td>c5669</td> <td>San Diego</td> <td>NaN</td> <td>Bandsdown</td> <td>1</td> <td>c5669</td> <td>0</td> <td>1</td> <td>0</td> <td>0</td> </tr> <tr> <td>1</td> <td>204</td> <td>CA</td> <td>37.238816</td> <td>-121.937118</td> <td>95032</td> <td>c16283</td> <td>Los Gatos</td> <td>NaN</td> <td>Tröger</td> <td>1</td> <td>c16283</td> <td>1</td> <td>0</td> <td>0</td> <td>1</td> </tr> <tr> <td>2</td> <td>1001</td> <td>CA</td> <td>32.90049</td> <td>-117.19266</td> <td>92121</td> <td>c5580</td> <td>San Diego</td> <td>CA32121</td> <td>Psi</td> <td>1</td> <td>c5580</td> <td>0</td> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>3</td> <td>738</td> <td>CA</td> <td>37.320308</td> <td>-122.05040</td> <td>95014</td> <td>c42688</td> <td>Cupertino</td> <td>CA35014</td> <td>Solidcore Systems</td> <td>1</td> <td>c42688</td> <td>0</td> <td>0</td> <td>0</td> <td>1</td> </tr> <tr> <td>4</td> <td>1002</td> <td>CA</td> <td>37.779281</td> <td>-122.41926</td> <td>94105</td> <td>c55806</td> <td>San Francisco</td> <td>CA34105</td> <td>Inhale Digital</td> <td>0</td> <td>c55806</td> <td>1</td> <td>1</td> <td>0</td> <td>0</td> </tr> </tbody> </table>	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	c5669	San Diego	NaN	Bandsdown	1	c5669	0	1	0	0	1	204	CA	37.238816	-121.937118	95032	c16283	Los Gatos	NaN	Tröger	1	c16283	1	0	0	1	2	1001	CA	32.90049	-117.19266	92121	c5580	San Diego	CA32121	Psi	1	c5580	0	0	1	0	3	738	CA	37.320308	-122.05040	95014	c42688	Cupertino	CA35014	Solidcore Systems	1	c42688	0	0	0	1	4	1002	CA	37.779281	-122.41926	94105	c55806	San Francisco	CA34105	Inhale Digital	0	c55806	1	1	0	0
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	c5669	San Diego	NaN	Bandsdown	1	c5669	0	1	0	0																																																																																		
1	204	CA	37.238816	-121.937118	95032	c16283	Los Gatos	NaN	Tröger	1	c16283	1	0	0	1																																																																																		
2	1001	CA	32.90049	-117.19266	92121	c5580	San Diego	CA32121	Psi	1	c5580	0	0	1	0																																																																																		
3	738	CA	37.320308	-122.05040	95014	c42688	Cupertino	CA35014	Solidcore Systems	1	c42688	0	0	0	1																																																																																		
4	1002	CA	37.779281	-122.41926	94105	c55806	San Francisco	CA34105	Inhale Digital	0	c55806	1	1	0	0																																																																																		
Univariate Analysis																																																																																																	

	<table border="1"> <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>4.0%</td> </tr> </tbody> </table>	State	Percentage	CA	52.9%	other	17.6%	NY	11.5%	MA	9.0%	TX	4.6%	WA	4.0%																																																																																											
State	Percentage																																																																																																									
CA	52.9%																																																																																																									
other	17.6%																																																																																																									
NY	11.5%																																																																																																									
MA	9.0%																																																																																																									
TX	4.6%																																																																																																									
WA	4.0%																																																																																																									
Bivariate Analysis	<table border="1"> <thead> <tr> <th>Category</th> <th>No. of startup (Category 0)</th> <th>No. of startup (Category 1)</th> </tr> </thead> <tbody> <tr> <td>software</td> <td>50</td> <td>100</td> </tr> <tr> <td>web</td> <td>50</td> <td>90</td> </tr> <tr> <td>mobile</td> <td>28</td> <td>52</td> </tr> <tr> <td>enterprise</td> <td>16</td> <td>55</td> </tr> <tr> <td>advertising</td> <td>16</td> <td>45</td> </tr> <tr> <td>games_video</td> <td>21</td> <td>30</td> </tr> <tr> <td>semiconductor</td> <td>12</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>15</td> <td>12</td> </tr> <tr> <td>public_relations</td> <td>15</td> <td>10</td> </tr> <tr> <td>e-commerce</td> <td>13</td> <td>10</td> </tr> <tr> <td>cleantech</td> <td>13</td> <td>10</td> </tr> <tr> <td>analytics</td> <td>3</td> <td>15</td> </tr> <tr> <td>security</td> <td>5</td> <td>15</td> </tr> <tr> <td>social</td> <td>7</td> <td>8</td> </tr> <tr> <td>search</td> <td>5</td> <td>7</td> </tr> <tr> <td>messaging</td> <td>5</td> <td>7</td> </tr> <tr> <td>other</td> <td>10</td> <td>2</td> </tr> <tr> <td>news</td> <td>2</td> <td>6</td> </tr> <tr> <td>travel</td> <td>2</td> <td>6</td> </tr> <tr> <td>fashion</td> <td>3</td> <td>4</td> </tr> <tr> <td>photo_video</td> <td>3</td> <td>4</td> </tr> <tr> <td>medical</td> <td>3</td> <td>4</td> </tr> <tr> <td>music</td> <td>2</td> <td>5</td> </tr> <tr> <td>finance</td> <td>2</td> <td>3</td> </tr> <tr> <td>education</td> <td>2</td> <td>2</td> </tr> <tr> <td>real_estate</td> <td>2</td> <td>2</td> </tr> <tr> <td>consulting</td> <td>2</td> <td>2</td> </tr> <tr> <td>health</td> <td>2</td> <td>2</td> </tr> <tr> <td>automotive</td> <td>1</td> <td>2</td> </tr> <tr> <td>transportation</td> <td>1</td> <td>2</td> </tr> <tr> <td>manufacturing</td> <td>1</td> <td>2</td> </tr> <tr> <td>sports</td> <td>1</td> <td>2</td> </tr> </tbody> </table>	Category	No. of startup (Category 0)	No. of startup (Category 1)	software	50	100	web	50	90	mobile	28	52	enterprise	16	55	advertising	16	45	games_video	21	30	semiconductor	12	24	biotech	12	22	network_hosting	10	24	hardware	15	12	public_relations	15	10	e-commerce	13	10	cleantech	13	10	analytics	3	15	security	5	15	social	7	8	search	5	7	messaging	5	7	other	10	2	news	2	6	travel	2	6	fashion	3	4	photo_video	3	4	medical	3	4	music	2	5	finance	2	3	education	2	2	real_estate	2	2	consulting	2	2	health	2	2	automotive	1	2	transportation	1	2	manufacturing	1	2	sports	1	2
Category	No. of startup (Category 0)	No. of startup (Category 1)																																																																																																								
software	50	100																																																																																																								
web	50	90																																																																																																								
mobile	28	52																																																																																																								
enterprise	16	55																																																																																																								
advertising	16	45																																																																																																								
games_video	21	30																																																																																																								
semiconductor	12	24																																																																																																								
biotech	12	22																																																																																																								
network_hosting	10	24																																																																																																								
hardware	15	12																																																																																																								
public_relations	15	10																																																																																																								
e-commerce	13	10																																																																																																								
cleantech	13	10																																																																																																								
analytics	3	15																																																																																																								
security	5	15																																																																																																								
social	7	8																																																																																																								
search	5	7																																																																																																								
messaging	5	7																																																																																																								
other	10	2																																																																																																								
news	2	6																																																																																																								
travel	2	6																																																																																																								
fashion	3	4																																																																																																								
photo_video	3	4																																																																																																								
medical	3	4																																																																																																								
music	2	5																																																																																																								
finance	2	3																																																																																																								
education	2	2																																																																																																								
real_estate	2	2																																																																																																								
consulting	2	2																																																																																																								
health	2	2																																																																																																								
automotive	1	2																																																																																																								
transportation	1	2																																																																																																								
manufacturing	1	2																																																																																																								
sports	1	2																																																																																																								
Outliers and Anomalies	-																																																																																																									

## Data Preprocessing Code Screenshots

	<pre> data = pd.read_csv('startups.csv') data     </pre> <p>Unnamed: 0 state_code latitude longitude zip_code id city Unnamed: 6 name label ... object_id has_VC has_angel has_rainfall has_rainmonth has_rainyear</p> <table border="1"> <tbody> <tr><td>0</td><td>105</td><td>CA</td><td>32.350880</td><td>-117.056020</td><td>92107</td><td>c5668</td><td>San Diego</td><td>NAN</td><td>Banditron</td><td>1</td><td>-</td><td>c5668</td><td>0</td><td>1</td><td>0</td><td>1</td></tr> <tr><td>1</td><td>294</td><td>CA</td><td>32.289111</td><td>-121.973218</td><td>95024</td><td>c5683</td><td>Los Gatos</td><td>NAN</td><td>TelCafe</td><td>1</td><td>-</td><td>c5683</td><td>1</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>2</td><td>1001</td><td>CA</td><td>32.901041</td><td>-117.193536</td><td>92121</td><td>c5628</td><td>San Diego</td><td>San Diego (CA 32121)</td><td>Rai</td><td>1</td><td>-</td><td>c5628</td><td>0</td><td>0</td><td>1</td><td>1</td></tr> <tr><td>3</td><td>738</td><td>CA</td><td>33.203891</td><td>-122.036640</td><td>95014</td><td>c4698</td><td>Cupertino</td><td>Operon CA 35014</td><td>Solidcore Systems</td><td>1</td><td>-</td><td>c4698</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>4</td><td>1002</td><td>CA</td><td>37.770081</td><td>-122.419236</td><td>94105</td><td>c5508</td><td>San Francisco</td><td>San Francisco CA 35115</td><td>Wise Digital</td><td>0</td><td>-</td><td>c5508</td><td>1</td><td>1</td><td>0</td><td>1</td></tr> <tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr> <tr><td>98</td><td>332</td><td>CA</td><td>33.740584</td><td>-122.376471</td><td>94102</td><td>c2548</td><td>San Francisco</td><td>Nan</td><td>Collected</td><td>1</td><td>-</td><td>c2548</td><td>0</td><td>0</td><td>1</td><td>1</td></tr> <tr><td>99</td><td>701</td><td>MA</td><td>42.359697</td><td>-71.059811</td><td>1601</td><td>c4747</td><td>Burlington</td><td>Burlington MA 1601</td><td>Red Point Systems</td><td>0</td><td>-</td><td>c4747</td><td>1</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>100</td><td>337</td><td>CA</td><td>37.848881</td><td>-122.015320</td><td>94088</td><td>c31548</td><td>Sunnyvale</td><td>NAN</td><td>Princo Medical</td><td>0</td><td>-</td><td>c31548</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>101</td><td>594</td><td>CA</td><td>37.556772</td><td>-122.298778</td><td>94041</td><td>c23198</td><td>San Francisco</td><td>NAN</td><td>Gazooz</td><td>1</td><td>-</td><td>c23198</td><td>0</td><td>0</td><td>1</td><td>1</td></tr> <tr><td>102</td><td>462</td><td>CA</td><td>37.386778</td><td>-122.366277</td><td>95054</td><td>c26702</td><td>Santa Clara</td><td>Santa Clara CA 35054</td><td>Serial Technologies</td><td>1</td><td>-</td><td>c26702</td><td>0</td><td>0</td><td>0</td><td>1</td></tr> </tbody> </table>	0	105	CA	32.350880	-117.056020	92107	c5668	San Diego	NAN	Banditron	1	-	c5668	0	1	0	1	1	294	CA	32.289111	-121.973218	95024	c5683	Los Gatos	NAN	TelCafe	1	-	c5683	1	0	0	1	2	1001	CA	32.901041	-117.193536	92121	c5628	San Diego	San Diego (CA 32121)	Rai	1	-	c5628	0	0	1	1	3	738	CA	33.203891	-122.036640	95014	c4698	Cupertino	Operon CA 35014	Solidcore Systems	1	-	c4698	0	0	0	1	4	1002	CA	37.770081	-122.419236	94105	c5508	San Francisco	San Francisco CA 35115	Wise Digital	0	-	c5508	1	1	0	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	98	332	CA	33.740584	-122.376471	94102	c2548	San Francisco	Nan	Collected	1	-	c2548	0	0	1	1	99	701	MA	42.359697	-71.059811	1601	c4747	Burlington	Burlington MA 1601	Red Point Systems	0	-	c4747	1	0	0	1	100	337	CA	37.848881	-122.015320	94088	c31548	Sunnyvale	NAN	Princo Medical	0	-	c31548	0	0	0	1	101	594	CA	37.556772	-122.298778	94041	c23198	San Francisco	NAN	Gazooz	1	-	c23198	0	0	1	1	102	462	CA	37.386778	-122.366277	95054	c26702	Santa Clara	Santa Clara CA 35054	Serial Technologies	1	-	c26702	0	0	0	1
0	105	CA	32.350880	-117.056020	92107	c5668	San Diego	NAN	Banditron	1	-	c5668	0	1	0	1																																																																																																																																																																												
1	294	CA	32.289111	-121.973218	95024	c5683	Los Gatos	NAN	TelCafe	1	-	c5683	1	0	0	1																																																																																																																																																																												
2	1001	CA	32.901041	-117.193536	92121	c5628	San Diego	San Diego (CA 32121)	Rai	1	-	c5628	0	0	1	1																																																																																																																																																																												
3	738	CA	33.203891	-122.036640	95014	c4698	Cupertino	Operon CA 35014	Solidcore Systems	1	-	c4698	0	0	0	1																																																																																																																																																																												
4	1002	CA	37.770081	-122.419236	94105	c5508	San Francisco	San Francisco CA 35115	Wise Digital	0	-	c5508	1	1	0	1																																																																																																																																																																												
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...																																																																																																																																																																												
98	332	CA	33.740584	-122.376471	94102	c2548	San Francisco	Nan	Collected	1	-	c2548	0	0	1	1																																																																																																																																																																												
99	701	MA	42.359697	-71.059811	1601	c4747	Burlington	Burlington MA 1601	Red Point Systems	0	-	c4747	1	0	0	1																																																																																																																																																																												
100	337	CA	37.848881	-122.015320	94088	c31548	Sunnyvale	NAN	Princo Medical	0	-	c31548	0	0	0	1																																																																																																																																																																												
101	594	CA	37.556772	-122.298778	94041	c23198	San Francisco	NAN	Gazooz	1	-	c23198	0	0	1	1																																																																																																																																																																												
102	462	CA	37.386778	-122.366277	95054	c26702	Santa Clara	Santa Clara CA 35054	Serial Technologies	1	-	c26702	0	0	0	1																																																																																																																																																																												
Loading Data																																																																																																																																																																																												
Handling Missing Data	<pre> # Filling missing value column("unnamed:6") data["unnamed: 6"] = data.apply(lambda row: (row.city) + " " + (row.state_code) + " " +(row.zip_code) , axis = 1)     </pre> <p># Total Missing Values column "unnamed: 6"  <code>totalNull = data["unnamed: 6"].isnull().sum()</code>  <code>print('Total Missing Values Kolom "unnamed: 6": ', totalNull)</code></p> <p>Total Missing Values Kolom "unnamed: 6": 0</p> <pre> # 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)     </pre> <p>Total Missing Values Kolom "closed_at": 0</p>																																																																																																																																																																																											

Data Transformation	<pre> data["status"] = data.status.map({'acquired':1, 'closed':0})  data["status"].astype(int)  0    1 1    1 2    1 3    1 4    0       ... 938    1 939    0 940    0 941    1 942    1 Name: status, length: 943, dtype: int64 </pre>
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-