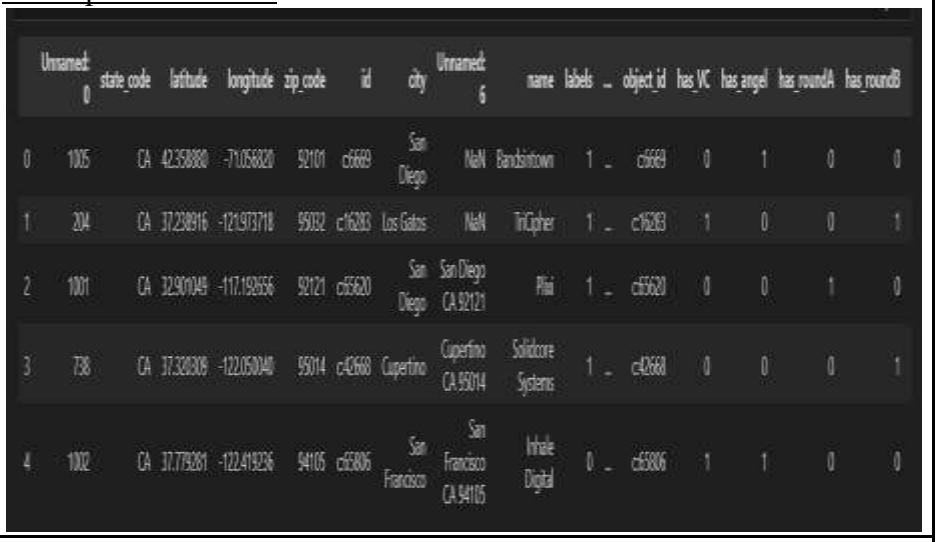
Data Collection and Preprocessing Phase

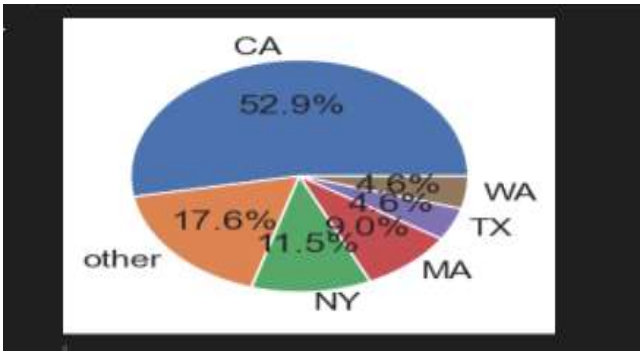
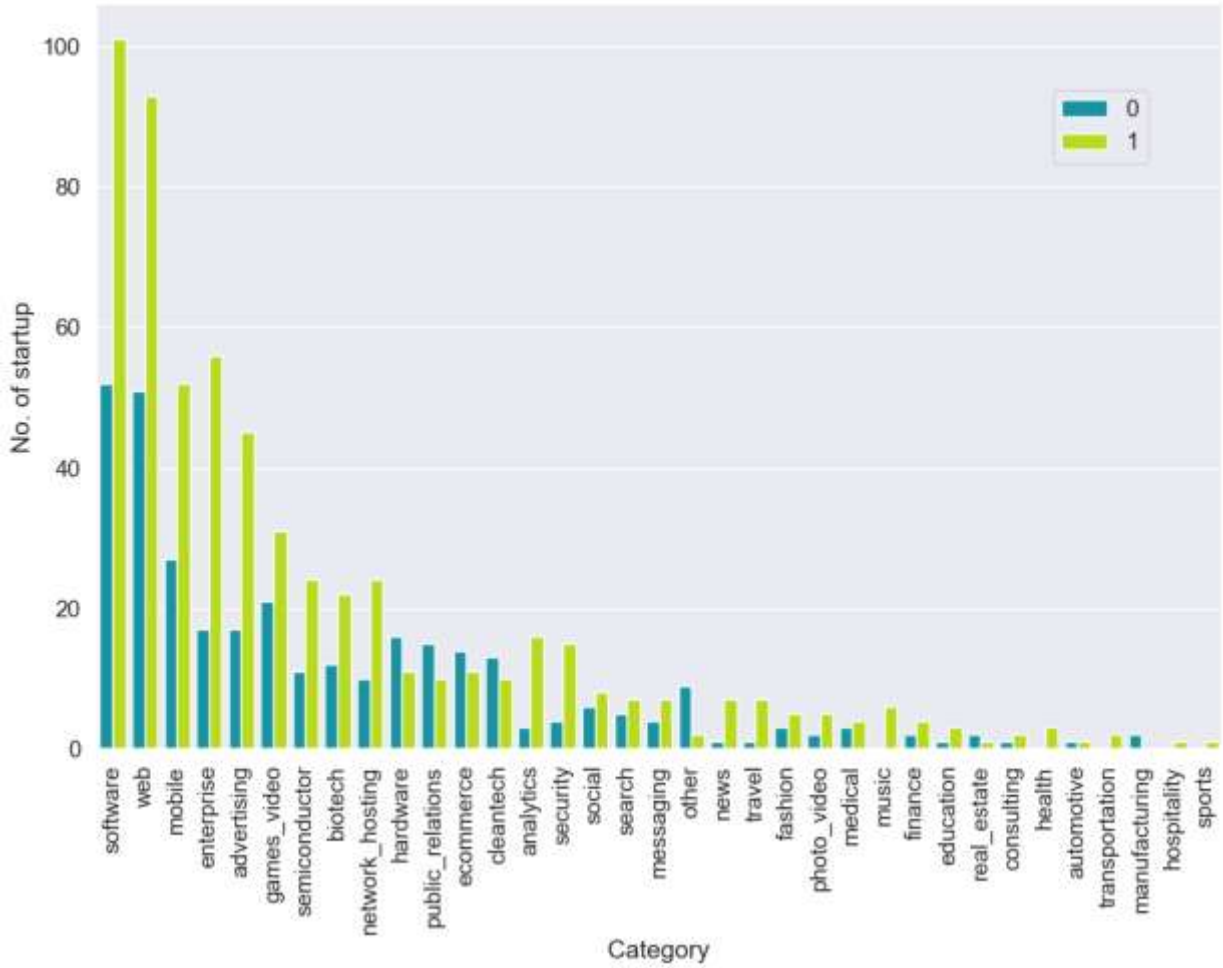
Date	7 June 2024
Team ID	739865
Project Title	prosperity Prognosticator : Machine Learning for Startup Success Prediction
Maximum Marks	6 Marks

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	<u>Dimension:</u> 923 rows × 49 columns <u>Descriptive statistics:</u> 
Univariate Analysis	

	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

	
Bivariate Analysis	

Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data

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

✓ 0.1s

	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	Bandcintown	1	...	c5669	0	1	0	0
1	294	CA	37.239916	-121.973718	95032	c16283	Los Gatos	NaN	TriCipher	1	...	c16283	1	0	0	1
2	1001	CA	32.901049	-117.192656	92121	cik5620	San Diego	San Diego CA 92121	Plus	1	...	cik5620	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	cik5806	San Francisco	San Francisco CA 94105	Inhale Digital	0	...	cik5806	1	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
918	352	CA	37.240594	-122.376471	94107	c21343	San Francisco	NaN	CoTwoot	1	...	c21343	0	0	1	0
919	721	MA	42.504817	-71.195611	1803	cik1747	Burlington	Burlington MA 1803	Reef Point Systems	0	...	cik1747	1	0	0	1
920	557	CA	37.408261	-122.015830	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	Avempra Technologies	1	...	c26702	0	0	0	1

Handling Missing Data

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

✓ 0.0s

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

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

✓ 0.0s

```
Total Missing Values Kolom "Unnamed: 6":  0
```

```python
#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)
```

✓ 0.0s

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

Data Transformation	
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-