

## Data Collection and Preprocessing Phase

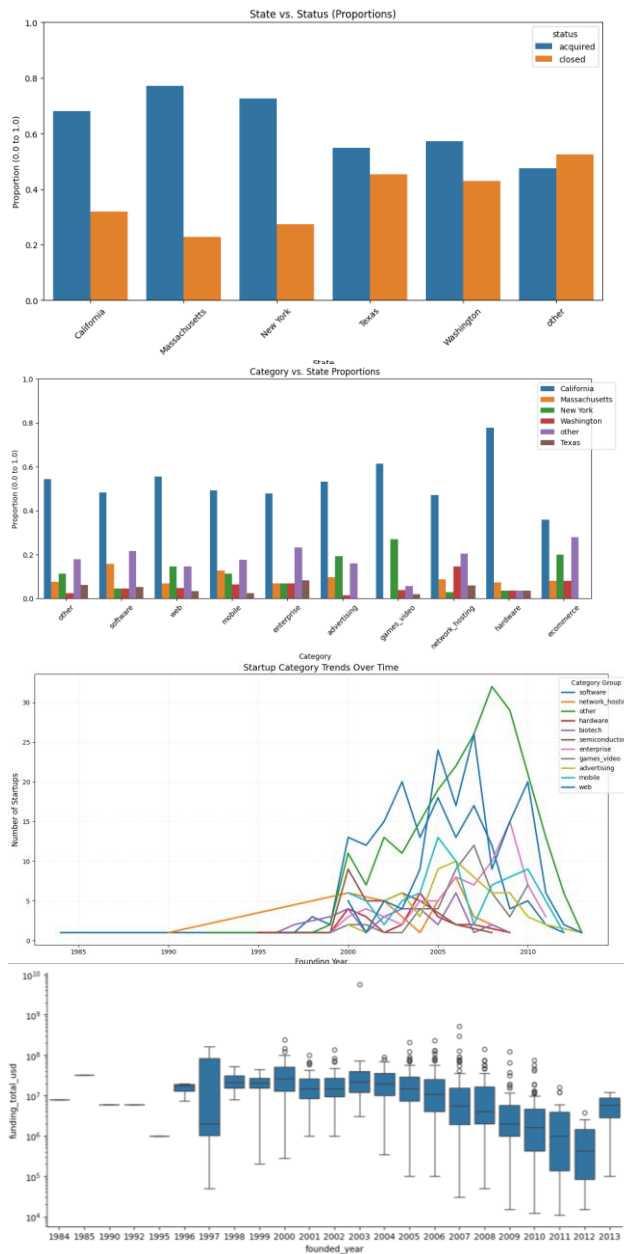
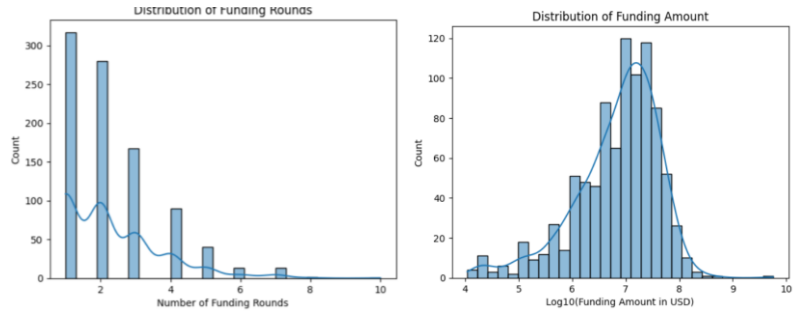
Date	18 June 2025
Team ID	SWTID1749880888
Project Title	Prosperity Prognosticator: Machine Learning for Startup Success Prediction
Maximum Marks	6 Marks

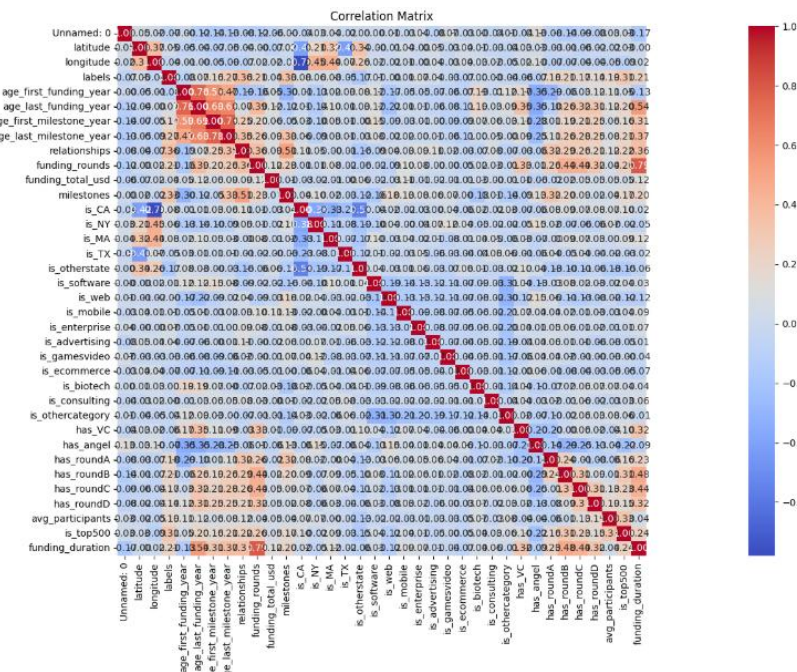
## Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																																			
Data Overview	<p><u>Dimension:</u> 923 rows x 49 columns</p> <p><u>Descriptive Statistics:</u></p> <table><tr><th></th><th>Unnamed: 0</th><th>latitude</th><th>longitude</th><th>labels</th><th>age_first_funding_year</th><th>age_last_funding_year</th><th>age_first_milestone_year</th><th>age_last_milestone_year</th><th>relationships</th><th>funding_rounds</th></tr><tr><td>count</td><td>923.000000</td><td>923.000000</td><td>923.000000</td><td>923.000000</td><td>923.000000</td><td>923.000000</td><td>771.000000</td><td>771.000000</td><td>923.000000</td><td>923.000000</td></tr><tr><td>mean</td><td>572.287941</td><td>38.517442</td><td>-103.536212</td><td>0.646804</td><td>2.235630</td><td>3.931456</td><td>3.056353</td><td>4.754423</td><td>7.710726</td><td>2.310943</td></tr><tr><td>std</td><td>333.585431</td><td>3.741497</td><td>22.394107</td><td>0.478222</td><td>2.510449</td><td>2.967910</td><td>2.877057</td><td>3.212107</td><td>7.265776</td><td>1.390922</td></tr><tr><td>min</td><td>1.000000</td><td>25.752358</td><td>-122.756956</td><td>0.000000</td><td>-9.046600</td><td>-9.046600</td><td>-14.169900</td><td>-7.005500</td><td>0.000000</td><td>1.000000</td></tr><tr><td>25%</td><td>283.500000</td><td>37.386889</td><td>-122.198732</td><td>0.000000</td><td>0.576700</td><td>1.669850</td><td>1.000000</td><td>2.411000</td><td>3.000000</td><td>1.000000</td></tr><tr><td>50%</td><td>577.000000</td><td>37.779281</td><td>-118.374037</td><td>1.000000</td><td>1.446600</td><td>3.528800</td><td>2.520500</td><td>4.476700</td><td>5.000000</td><td>2.000000</td></tr><tr><td>75%</td><td>866.500000</td><td>40.730646</td><td>-77.214731</td><td>1.000000</td><td>3.575350</td><td>5.560250</td><td>4.686300</td><td>6.753400</td><td>10.000000</td><td>3.000000</td></tr><tr><td>max</td><td>1153.000000</td><td>59.335232</td><td>18.057121</td><td>1.000000</td><td>21.895900</td><td>21.895900</td><td>24.684900</td><td>24.684900</td><td>63.000000</td><td>10.000000</td></tr></table>		Unnamed: 0	latitude	longitude	labels	age_first_funding_year	age_last_funding_year	age_first_milestone_year	age_last_milestone_year	relationships	funding_rounds	count	923.000000	923.000000	923.000000	923.000000	923.000000	923.000000	771.000000	771.000000	923.000000	923.000000	mean	572.287941	38.517442	-103.536212	0.646804	2.235630	3.931456	3.056353	4.754423	7.710726	2.310943	std	333.585431	3.741497	22.394107	0.478222	2.510449	2.967910	2.877057	3.212107	7.265776	1.390922	min	1.000000	25.752358	-122.756956	0.000000	-9.046600	-9.046600	-14.169900	-7.005500	0.000000	1.000000	25%	283.500000	37.386889	-122.198732	0.000000	0.576700	1.669850	1.000000	2.411000	3.000000	1.000000	50%	577.000000	37.779281	-118.374037	1.000000	1.446600	3.528800	2.520500	4.476700	5.000000	2.000000	75%	866.500000	40.730646	-77.214731	1.000000	3.575350	5.560250	4.686300	6.753400	10.000000	3.000000	max	1153.000000	59.335232	18.057121	1.000000	21.895900	21.895900	24.684900	24.684900	63.000000	10.000000
		Unnamed: 0	latitude	longitude	labels	age_first_funding_year	age_last_funding_year	age_first_milestone_year	age_last_milestone_year	relationships	funding_rounds																																																																																									
count	923.000000	923.000000	923.000000	923.000000	923.000000	923.000000	771.000000	771.000000	923.000000	923.000000																																																																																										
mean	572.287941	38.517442	-103.536212	0.646804	2.235630	3.931456	3.056353	4.754423	7.710726	2.310943																																																																																										
std	333.585431	3.741497	22.394107	0.478222	2.510449	2.967910	2.877057	3.212107	7.265776	1.390922																																																																																										
min	1.000000	25.752358	-122.756956	0.000000	-9.046600	-9.046600	-14.169900	-7.005500	0.000000	1.000000																																																																																										
25%	283.500000	37.386889	-122.198732	0.000000	0.576700	1.669850	1.000000	2.411000	3.000000	1.000000																																																																																										
50%	577.000000	37.779281	-118.374037	1.000000	1.446600	3.528800	2.520500	4.476700	5.000000	2.000000																																																																																										
75%	866.500000	40.730646	-77.214731	1.000000	3.575350	5.560250	4.686300	6.753400	10.000000	3.000000																																																																																										
max	1153.000000	59.335232	18.057121	1.000000	21.895900	21.895900	24.684900	24.684900	63.000000	10.000000																																																																																										
Univariate Analysis	<div><div><p>Distribution of Startups by Category</p><table><thead><tr><th>Category</th><th>Percentage</th></tr></thead><tbody><tr><td>other</td><td>29.7%</td></tr><tr><td>software</td><td>16.6%</td></tr><tr><td>web</td><td>15.6%</td></tr><tr><td>mobile</td><td>8.6%</td></tr><tr><td>enterprise</td><td>7.9%</td></tr><tr><td>advertising</td><td>6.7%</td></tr><tr><td>games_video</td><td>5.6%</td></tr><tr><td>network_hosting</td><td>3.7%</td></tr><tr><td>hardware</td><td>2.9%</td></tr><tr><td>ecommerce</td><td>2.7%</td></tr></tbody></table></div><div><p>Distribution of Startups by State</p><table><thead><tr><th>State</th><th>Percentage</th></tr></thead><tbody><tr><td>California</td><td>52.9%</td></tr><tr><td>other</td><td>17.6%</td></tr><tr><td>New York</td><td>11.5%</td></tr><tr><td>Massachusetts</td><td>9.0%</td></tr><tr><td>Texas</td><td>4.6%</td></tr><tr><td>Washington</td><td>4.6%</td></tr></tbody></table></div></div>	Category	Percentage	other	29.7%	software	16.6%	web	15.6%	mobile	8.6%	enterprise	7.9%	advertising	6.7%	games_video	5.6%	network_hosting	3.7%	hardware	2.9%	ecommerce	2.7%	State	Percentage	California	52.9%	other	17.6%	New York	11.5%	Massachusetts	9.0%	Texas	4.6%	Washington	4.6%																																																															
Category	Percentage																																																																																																			
other	29.7%																																																																																																			
software	16.6%																																																																																																			
web	15.6%																																																																																																			
mobile	8.6%																																																																																																			
enterprise	7.9%																																																																																																			
advertising	6.7%																																																																																																			
games_video	5.6%																																																																																																			
network_hosting	3.7%																																																																																																			
hardware	2.9%																																																																																																			
ecommerce	2.7%																																																																																																			
State	Percentage																																																																																																			
California	52.9%																																																																																																			
other	17.6%																																																																																																			
New York	11.5%																																																																																																			
Massachusetts	9.0%																																																																																																			
Texas	4.6%																																																																																																			
Washington	4.6%																																																																																																			

## Bivariate Analysis





```

[3] # Download latest version
path = kagglehub.dataset_download("manishkb6/startup-success-prediction")

[4] Downloading from https://www.kaggle.com/manishkb6/datasets/download/manishkb6/startup-success-prediction/dataset_version_number1...
64.12/64.1K [00:00<00, 33.0KB/s] extracting files...

[4] print(path)

~/root/.cache/kagglehub/datasets/manishkb6/startup-success-prediction/versions/1

[5] Import os

for root, dirs, files in os.walk(path):
    for file in files:
        print(os.path.join(root, file))

~/root/.cache/kagglehub/datasets/manishkb6/startup-success-prediction/versions/1/startup_data.csv

data = os.path.join(path, "startup_data.csv")
df = pd.read_csv(data)
df


```

	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	thousand: 4	name	labels	founded_at	closed_at	first_funding_at	last_funding_at	age_first_funding_year
0	1005	CA	42.358800	-71.056820	92101	c-6869	San Diego	Nah	Bandwidth	1	1/1/2007	Nah	4/1/2009	1/1/2010	2.2493
1	204	CA	37.238916	-121.973718	95632	c-16283	Los Gatos	Nah	TicCipher	1	1/1/2008	Nah	2/14/2005	12/28/2009	5.1266
2	101	CA	32.961049	-117.132656	92121	c-65620	San Diego	CA 92121	Pblid	1	3/18/2009	Nah	3/30/2010	3/30/2010	1.0329
3	738	CA	37.328309	-122.050040	95014	C-42668	Cupertino	Cupertino CA 95014	Solidcore Systems	1	1/1/2002	Nah	2/17/2005	4/25/2007	3.1315
4	1002	CA	37.773081	-122.419236	94105	c-65086	San Francisco	CA 94105	Inlabs Digital	0	8/1/2010	10/1/2012	8/1/2010	4/1/2012	8.0000

```

[32] age["age_first_funding_year","age_last_funding_year","age_first_milestone_year","age_last_milestone_year"]
    for a in range(len(age)):
        print("Is there any negative value in '{}' column : {}".format(age[a],df[age[a]][0].any()))

[33] Is there any negative value in 'age first funding year' column : True
Is there any negative value in 'age last funding year' column : True
Is there any negative value in 'age first milestone year' column : True
Is there any negative value in 'age last milestone year' column : True

[34] df=df.drop(df[df.age_first_funding_year<0].index)
df=df.drop(df[df.age_last_funding_year<0].index)
df=df.drop(df[df.age_first_milestone_year<0].index)
df=df.drop(df[df.age_last_milestone_year<0].index)

[34] for a in range(len(age)):
    print("Is there any negative value in '{}' column : {}".format(age[a],df[age[a]][0].any()))

[35] Is there any negative value in 'age first funding year' column : False
Is there any negative value in 'age last funding year' column : False
Is there any negative value in 'age first milestone year' column : False
Is there any negative value in 'age last milestone year' column : False

[36] columns_to_drop = [
    'status', 'closed_at', 'id', 'object_id', 'name',
    'city', 'zip_code', 'Unnamed: 0', 'Unnamed: 6'
]

df.drop(columns=columns_to_drop, inplace=True, errors='ignore')

```

## Feature Engineering

```
# select only the 7 features that will be displayed
```

```
selected_features = [
    'funding_rounds',
    'milestones',
    'relationships',
    'is_top500',
    'funding_total_usd',
    'has_round8',
    'avg_participants',
    'labels'
]
```

```
df_selected = df[selected_features].copy()
display(df_selected.head())
```

	funding_rounds	milestones	relationships	is_top500	funding_total_usd	has_round8	avg_participants	labels
0	3	3	3	0	375000	0	1.0000	1
1	4	1	9	1	40100000	1	4.7500	1
2	1	2	5	1	2600000	0	4.0000	1
3	3	1	5	1	40000000	1	3.3333	1
4	2	1	2	1	1300000	0	1.0000	0

## Save Processed Data

```
# Save the DataFrame with selected features to a new CSV file
```

```
output_filename = 'selected_features_data.csv'
df_selected.to_csv(output_filename, index=False)
```

```
print(f"\nDataFrame with selected features saved to '{output_filename}'")
```

```
DataFrame with selected features saved to 'selected_features_data.csv'
```