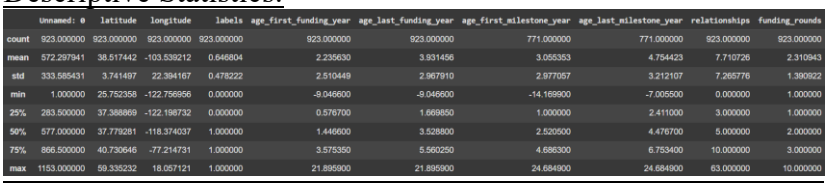
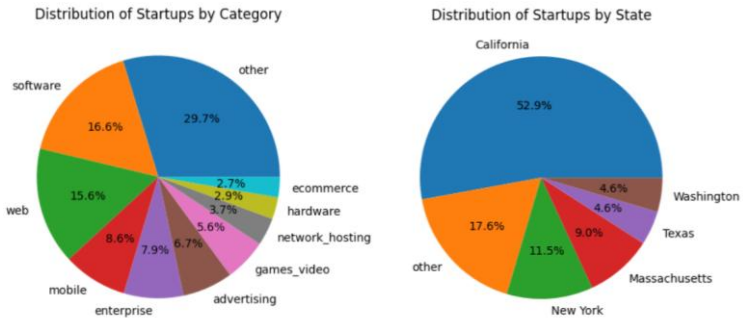
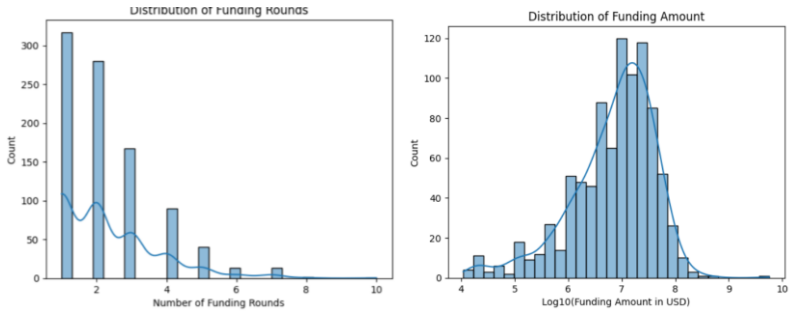
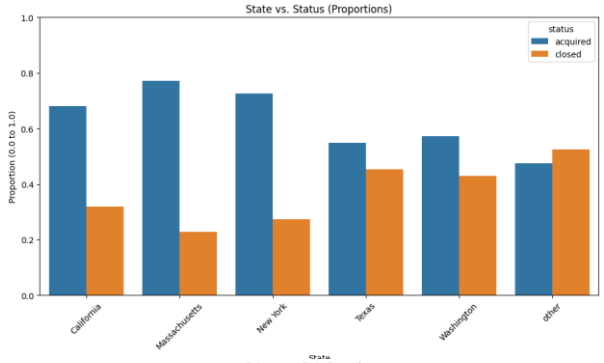
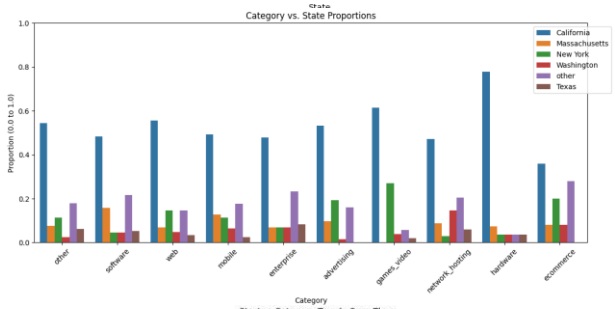
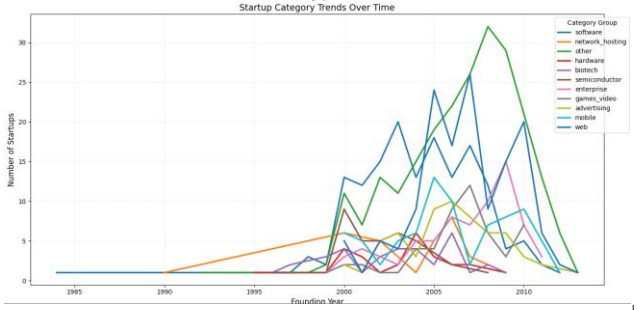
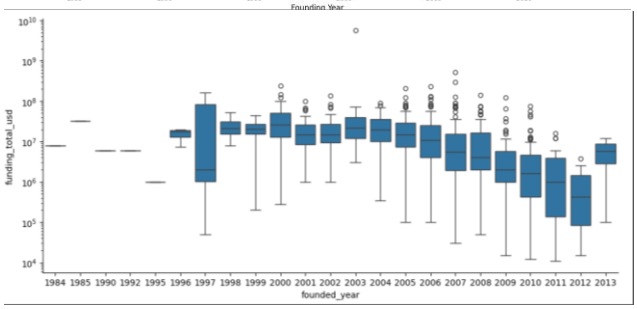
Data Collection and Preprocessing Phase

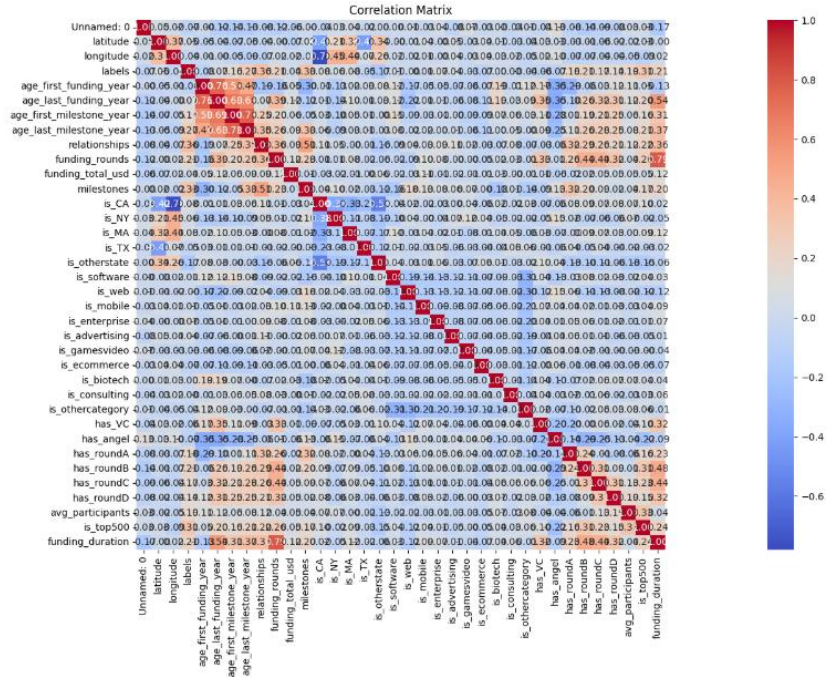
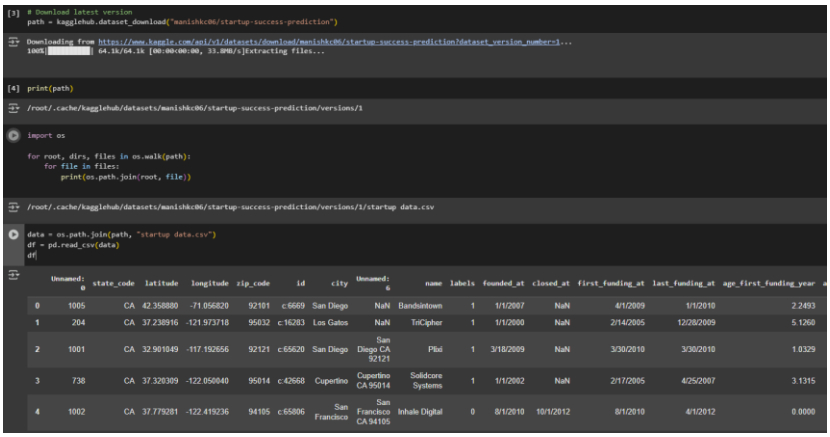
Date	18 June 2025
Team ID	SWTID1749880888
Project Title	Prosperity Prognosticator: Machine Learning for Startup Success Prediction
Maximum Marks	6 Marks

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	Dimension: 923 rows x 49 columns Descriptive Statistics: 
Univariate Analysis	

	
Bivariate Analysis	   

Multivariate Analysis	
Data Preprocessing Code Screenshots	
Loading Data	
Dropping Irrelevant Columns	

Feature Engineering

```
# Select only the 7 features that will be displayed
selected_features = [
    'funding_rounds',
    'milestones',
    'relationships',
    'is_top500',
    'funding_total_usd',
    'has_roundB',
    'avg_participants',
    'labels'
]

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

	funding_rounds	milestones	relationships	is_top500	funding_total_usd	has_roundB	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'
```