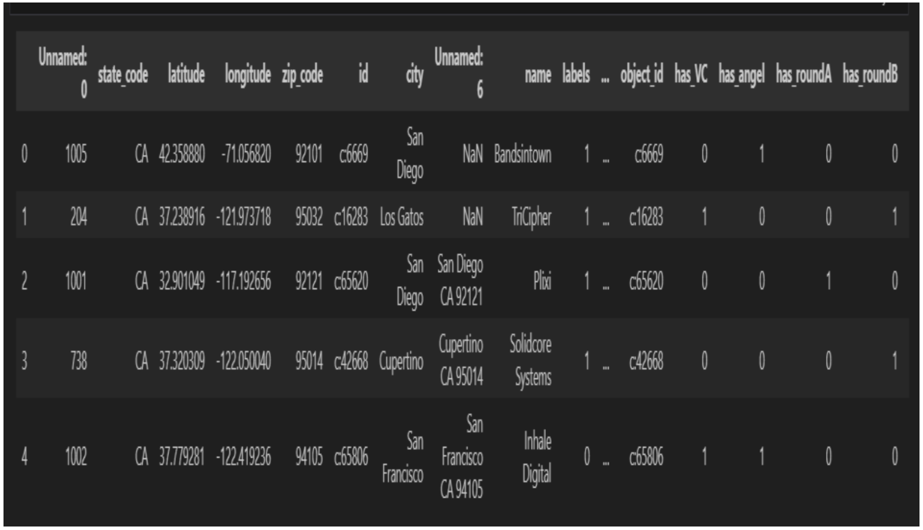
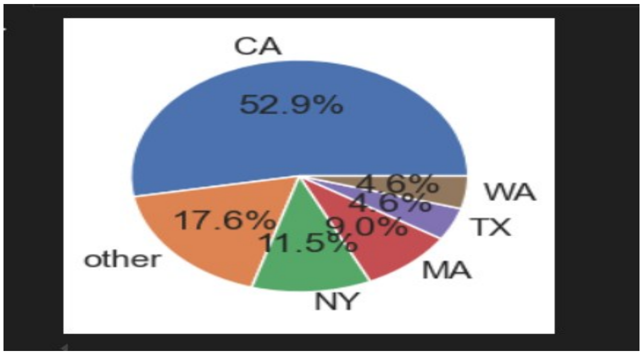
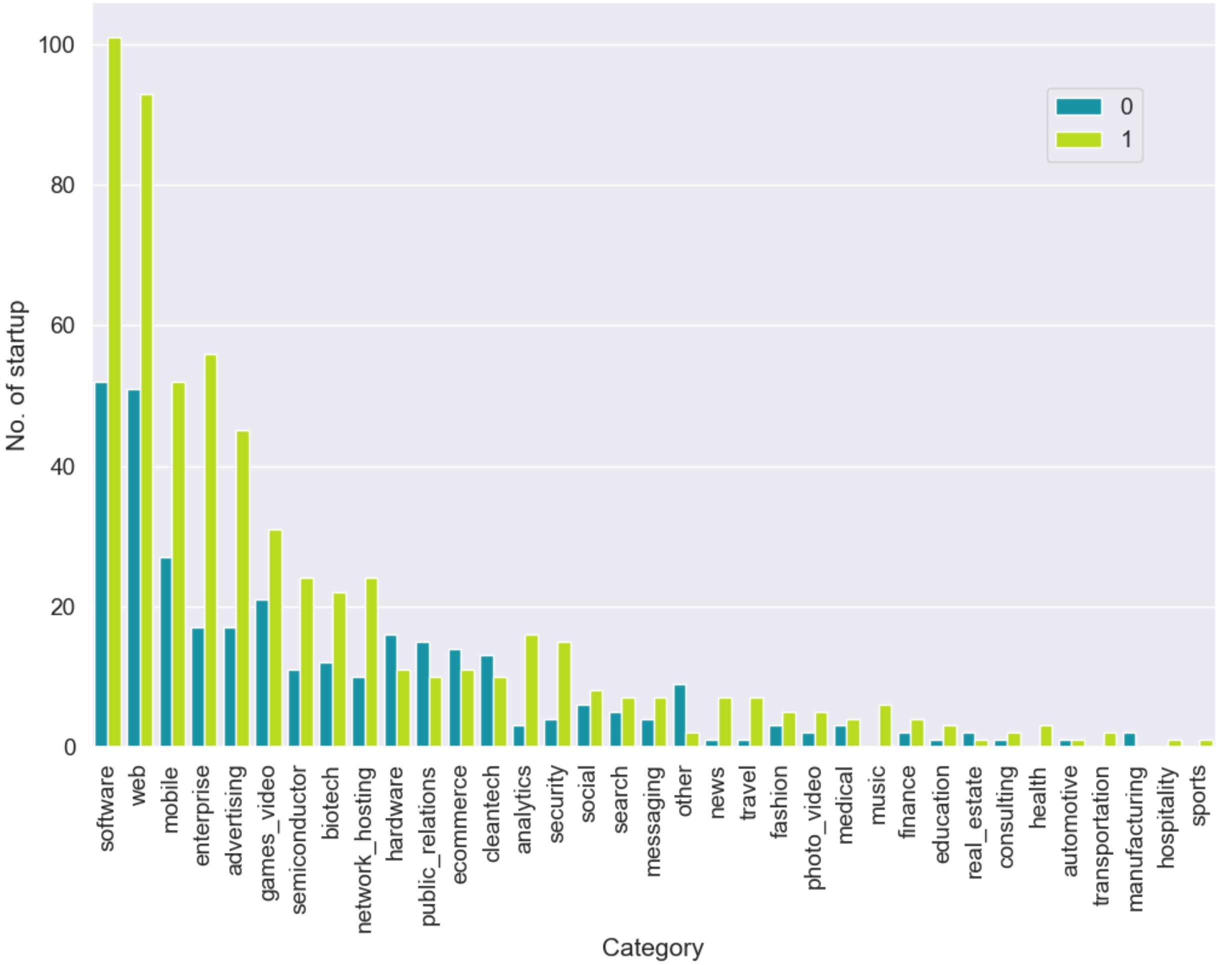
# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 7 June 2024 |
| Team ID | team-739778 |
| 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.

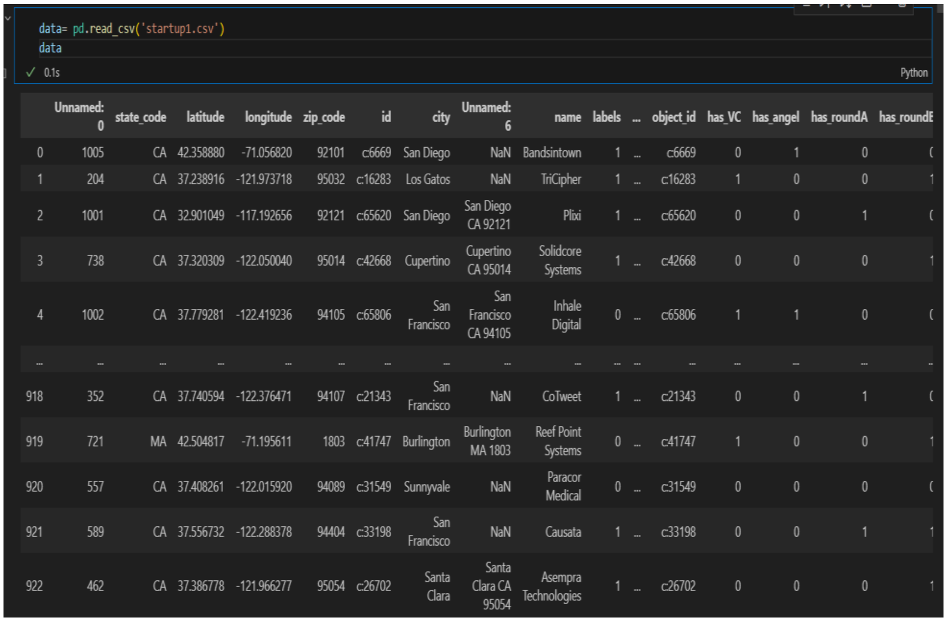
<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimension:<br>923 rows × 49 columns<br>Descriptive statistics:<br>
<table>
<tr><th></th><th>Unnamed: 0</th><th>state_code</th><th>latitude</th><th>longitude</th><th>zip_code</th><th>id</th><th>city</th><th>Unnamed: 6</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>
<tr><td>0</td><td>1005</td><td>CA</td><td>42.358880</td><td>-71.056820</td><td>92101</td><td>c6669</td><td>San Diego</td><td>NaN</td><td>Bandsintown</td><td>1</td><td>...</td><td>c6669</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>37.238916</td><td>-121.973718</td><td>95032</td><td>c16283</td><td>Los Gatos</td><td>NaN</td><td>TriCipher</td><td>1</td><td>...</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>32.901049</td><td>-117.192656</td><td>92121</td><td>c65620</td><td>San Diego</td><td>San Diego CA 92121</td><td>Plixi</td><td>1</td><td>...</td><td>c65620</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>37.320309</td><td>-122.050040</td><td>95014</td><td>c42668</td><td>Cupertino</td><td>Cupertino CA 95014</td><td>Solidcore Systems</td><td>1</td><td>...</td><td>c42668</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>-122.419236</td><td>94105</td><td>c65806</td><td>San Francisco</td><td>San Francisco CA 94105</td><td>Inhale Digital</td><td>0</td><td>...</td><td>c65806</td><td>1</td><td>1</td><td>0</td><td>0</td></tr>
</table>
</td></tr>
<tr><td>Univariate Analysis</td><td></td></tr>
</table>

<table>
<tr><td></td><td>
</td></tr>
<tr><td>Bivariate Analysis</td><td>
</td></tr>
</table>

| | |
|---|---|
| Outliers and Anomalies | - |

## Data Preprocessing Code Screenshots

Loading Data



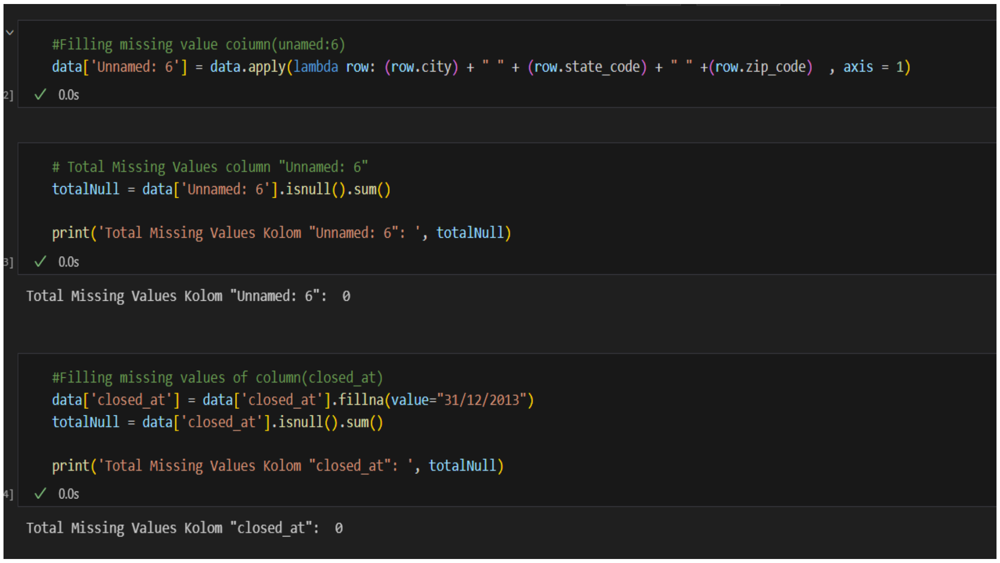
```
data= pd.read_csv('startup1.csv')
data
```

✓ 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_roundE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1005 | CA | 42.358880 | -71.056820 | 92101 | c:6669 | San Diego | NaN | Bandsintown | 1 | ... | c:6669 | 0 | 1 | 0 | ( |
| 1 | 204 | CA | 37.238916 | -121.973718 | 95032 | c:16283 | Los Gatos | NaN | TriCipher | 1 | ... | c:16283 | 1 | 0 | 0 | 1 |
| 2 | 1001 | CA | 32.901049 | -117.192656 | 92121 | c:65620 | San Diego | San Diego CA 92121 | Plixi | 1 | ... | c:65620 | 0 | 0 | 1 | ( |
| 3 | 738 | CA | 37.320309 | -122.050040 | 95014 | c:42668 | Cupertino | Cupertino CA 95014 | Solidcore Systems | 1 | ... | c:42668 | 0 | 0 | 0 | 1 |
| 4 | 1002 | CA | 37.779281 | -122.419236 | 94105 | c:65806 | San Francisco | San Francisco CA 94105 | Inhale Digital | 0 | ... | c:65806 | 1 | 1 | 0 | ( |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 918 | 352 | CA | 37.740594 | -122.376471 | 94107 | c:21343 | San Francisco | NaN | CoTweet | 1 | ... | c:21343 | 0 | 0 | 1 | ( |
| 919 | 721 | MA | 42.504817 | -71.195611 | 1803 | c:41747 | Burlington | Burlington MA 1803 | Reef Point Systems | 0 | ... | c:41747 | 1 | 0 | 0 | 1 |
| 920 | 557 | CA | 37.408261 | -122.015920 | 94089 | c:31549 | Sunnyvale | NaN | Paracor Medical | 0 | ... | c:31549 | 0 | 0 | 0 | ( |
| 921 | 589 | CA | 37.556732 | -122.288378 | 94404 | c:33198 | San Francisco | NaN | Causata | 1 | ... | c:33198 | 0 | 0 | 1 | 1 |
| 922 | 462 | CA | 37.386778 | -121.966277 | 95054 | c:26702 | Santa Clara | Santa Clara CA 95054 | Asempra Technologies | 1 | ... | c:26702 | 0 | 0 | 0 | 1 |

Handling Missing Data



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

✓ 0.0s

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

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

✓ 0.0s

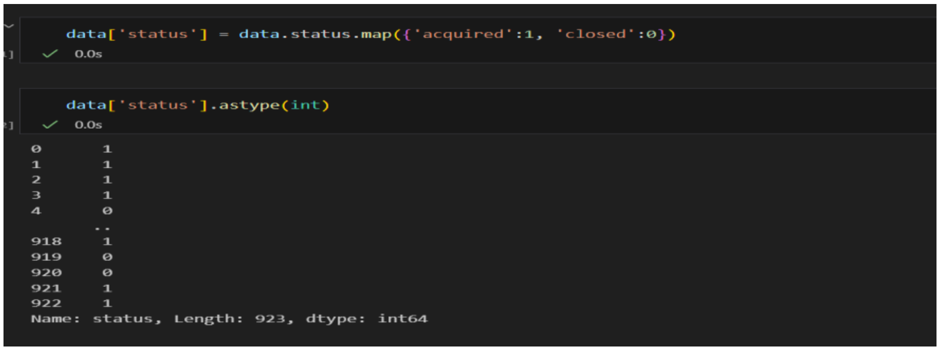
```
Total Missing Values Kolom "Unnamed: 6":  0
```

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