```
# This jupyter notebook is prepared by David Winfield

from google.colab import files
uploaded = files.upload()
```

Browse... startup_info_.csv

startup_info_.csv(application/vnd.ms-excel) - 168764 bytes, last modified: n/a - 100% done

```
Saving startup_info_.csv to startup_info_.csv
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb


df = pd.read_csv("startup_info_.csv")

rows, columns = df.shape

print("The dataset contains {} rows and {} columns".format(rows, columns))
```

```
The dataset contains 923 rows and 28 columns
```

```
df.describe()
```

	Unnamed: 0	latitude	longitude	labels	age_first_funding_year	age_la
count	923.000000	923.000000	923.000000	923.000000	923.000000	
mean	572.297941	38.517442	-103.539212	0.646804	2.235630	
std	333.585431	3.741497	22.394167	0.478222	2.510449	
min	1.000000	25.752358	-122.756956	0.000000	-9.046600	
25%	283.500000	37.388869	-122.198732	0.000000	0.576700	
50%	577.000000	37.779281	-118.374037	1.000000	1.446600	
75%	866.500000	40.730646	-77.214731	1.000000	3.575350	
max	1153.000000	59.335232	18.057121	1.000000	21.895900	

```
columns = df.columns

print("Attribute columns:", columns)
```

✓ 0s completed at 5:21 PM

```
       'city', 'Unnamed: 6', 'name', 'labels', 'founded_at', 'closed_at',
       'first_funding_at', 'last_funding_at', 'age_first_funding_year',
       'age_last_funding_year', 'age_first_milestone_year',
       'age_last_milestone_year', 'relationships', 'funding_rounds',
       'funding_total_usd', 'milestones', 'state_code.1', 'category_code',
       'object_id', 'avg_participants', 'is_top500', 'status'],
      dtype='object')
```

```
df.drop(columns=["Unnamed: 0", "Unnamed: 6", "state_code.1", "object_id"], inplace=True)
```

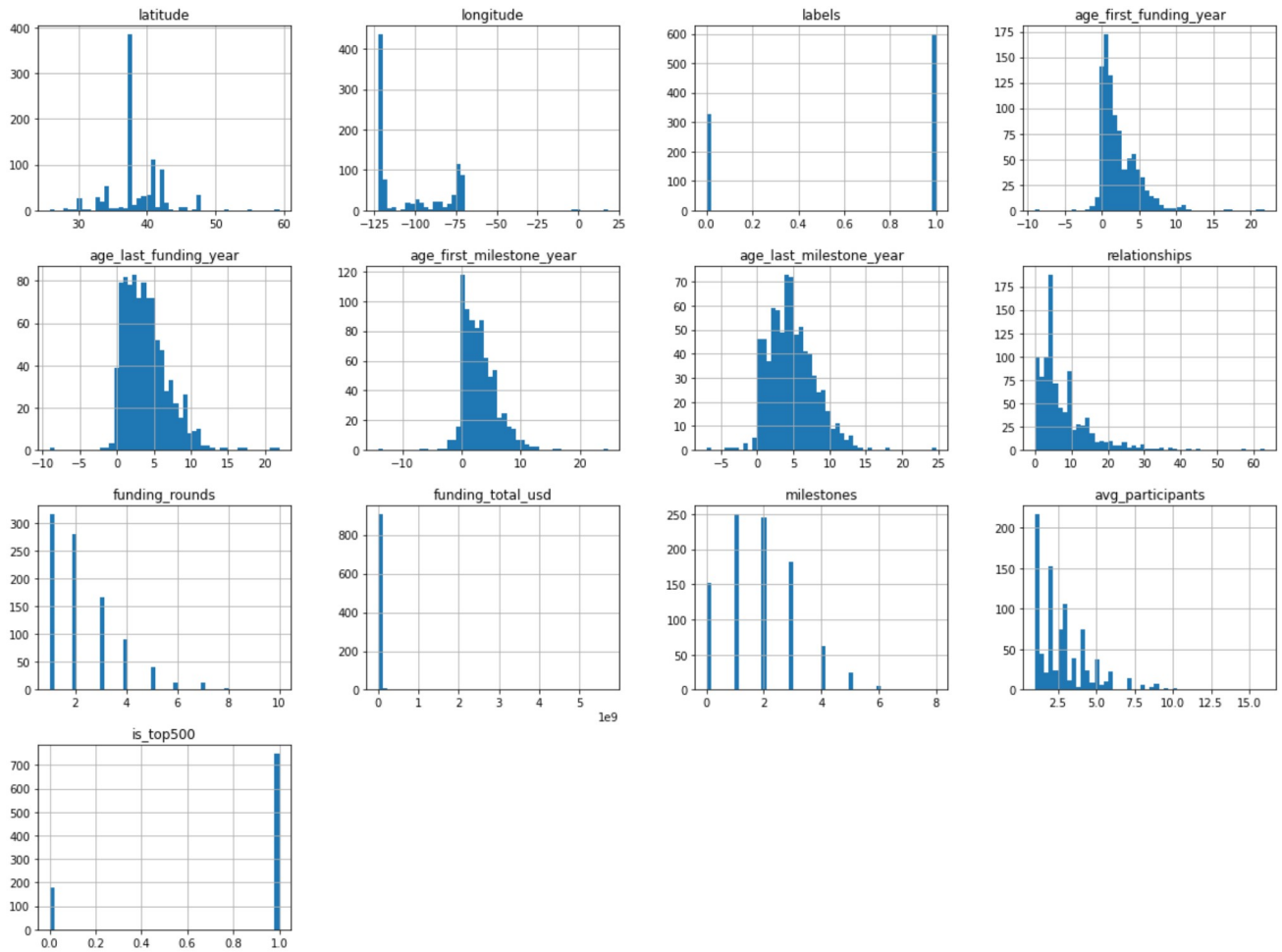
```
numeric_df = df._get_numeric_data()
```

```
print(numeric_df)
```

```
      latitude   longitude  labels  age_first_funding_year  \
0    42.358880  -71.056820       1                  2.2493
1    37.238916 -121.973718       1                  5.1260
2    32.901049 -117.192656       1                  1.0329
3    37.320309 -122.050040       1                  3.1315
4    37.779281 -122.419236       0                  0.0000
..         ...         ...     ...                     ...
918  37.740594 -122.376471       1                  0.5178
919  42.504817  -71.195611       0                  7.2521
920  37.408261 -122.015920       0                  8.4959
921  37.556732 -122.288378       1                  0.7589
922  37.386778 -121.966277       1                  3.1205

     age_last_funding_year  age_first_milestone_year  age_last_milestone_year  \
0                   3.0027                    4.6685                   6.7041
1                   9.9973                    7.0055                   7.0055
2                   1.0329                    1.4575                   2.2055
3                   5.3151                    6.0027                   6.0027
4                   1.6685                    0.0384                   0.0384
..                     ...                       ...                      ...
918                 0.5178                    0.5808                   4.5260
919                 9.2274                    6.0027                   6.0027
920                 8.4959                    9.0055                   9.0055
921                 2.8329                    0.7589                   3.8356
922                 3.1205                    4.0027                   4.0027

     relationships  funding_rounds  funding_total_usd  milestones  \
0                3               3             375000           3
1                9               4           40100000           1
2                5               1            2600000           2
3                5               3           40000000           1
4                2               2            1300000           1
..             ...             ...                ...         ...
918              9               1            1100000           2
919              1               3           52000000           1
920              5               1           44000000           1
921             12               2           15500000           2
922              4               1           20000000           1
```

```
922                4                 1            20000000              1

     avg_participants  is_top500
0              1.0000          0
1              4.7500          1
2              4.0000          1
3              3.3333          1
4              1.0000          1
..                ...        ...
918            6.0000          1
919            2.6667          1
920            8.0000          1
921            1.0000          1
922            3.0000          1

[923 rows x 13 columns]
```

```
numeric_df.hist(bins=50, figsize=(20, 15))

plt.show()

skew = numeric_df.skew()

print("Skew of each numeric column:")
print(skew)
```



```
Skew of each numeric column:
latitude                     0.309298
longitude                    0.873708
labels                      -0.615290
age_first_funding_year       2.104001
age_last_funding_year        1.092075
age_first_milestone_year     0.944468
age_last_milestone_year      0.711934
relationships                2.329961
funding_rounds               1.356917
funding_total_usd           29.152461
milestones                   0.577378
avg_participants             1.767554
is_top500                   -1.577343
dtype: float64
```

```
categorical_df = df.select_dtypes(include='object')

print(categorical_df)
```

```
     state_code zip_code       id           city                  name  \
0            CA    92101   c:6669      San Diego           Bandsintown
1            CA    95032  c:16283      Los Gatos             TriCipher
2            CA    92121  c:65620      San Diego                 Plixi
3            CA    95014  c:42668      Cupertino     Solidcore Systems
4            CA    94105  c:65806  San Francisco        Inhale Digital
..          ...      ...      ...            ...                   ...
918          CA    94107  c:21343  San Francisco               CoTweet
919          MA     1803  c:41747     Burlington    Reef Point Systems
920          CA    94089  c:31549      Sunnyvale       Paracor Medical
921          CA    94404  c:33198  San Francisco               Causata
922          CA    95054  c:26702    Santa Clara  Asempra Technologies
```

```
    founded_at  closed_at first_funding_at last_funding_at category_code  \
0     1/1/2007        NaN         4/1/2009        1/1/2010         music
1     1/1/2000        NaN        2/14/2005      12/28/2009    enterprise
2    3/18/2009        NaN        3/30/2010       3/30/2010           web
3     1/1/2002        NaN        2/17/2005       4/25/2007      software
4     8/1/2010  10/1/2012         8/1/2010        4/1/2012   games_video
..         ...        ...              ...             ...           ...
918   1/1/2009        NaN         7/9/2009        7/9/2009   advertising
919   1/1/1998  6/25/2008         4/1/2005       3/23/2007      security
920   1/1/1999  6/17/2012        6/29/2007       6/29/2007       biotech
921   1/1/2009        NaN        10/5/2009       11/1/2011      software
922   1/1/2003        NaN        2/13/2006       2/13/2006      security

       status
0    acquired
1    acquired
2    acquired
3    acquired
4      closed
..        ...
918  acquired
919    closed
920    closed
921  acquired
922  acquired

[923 rows x 11 columns]
```

```
missing_values = df.isna().sum().sort_values(ascending=False)

print(missing_values)
```

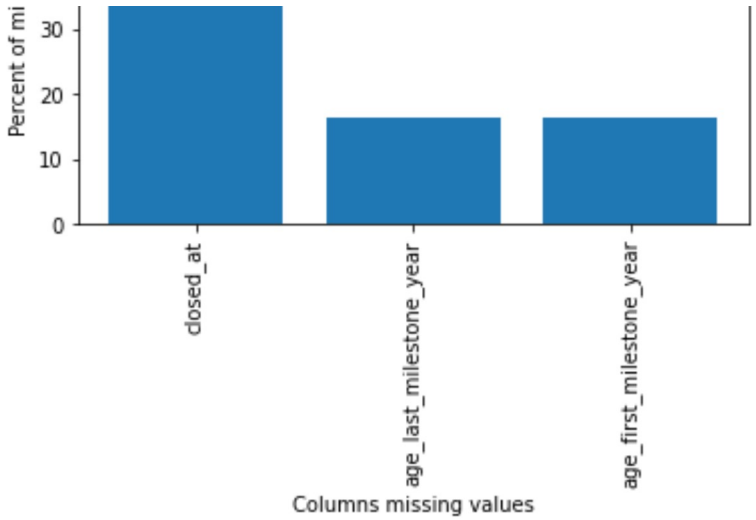
```
closed_at                  588
age_last_milestone_year    152
age_first_milestone_year   152
state_code                   0
age_last_funding_year        0
is_top500                    0
avg_participants             0
category_code                0
milestones                   0
funding_total_usd            0
funding_rounds               0
relationships                0
age_first_funding_year       0
latitude                     0
last_funding_at              0
first_funding_at             0
founded_at                   0
labels                       0
name                         0
city                         0
id                           0
zip_code                     0
```

```
longitude                   0
status                      0
dtype: int64
```

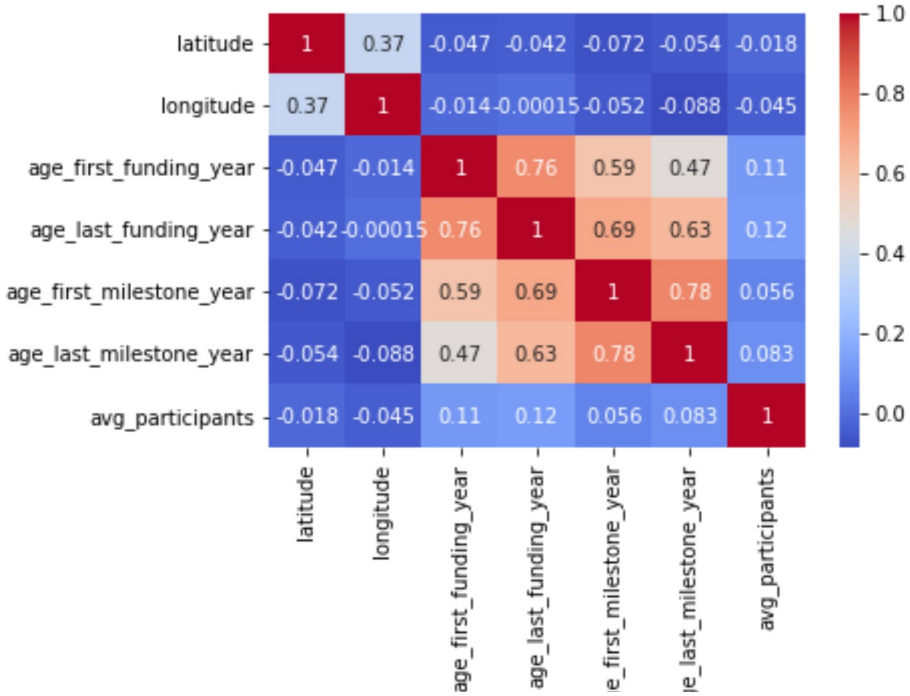
```
missing_values_percentage = (df.isna().mean() * 100).sort_values(ascending=False)

print(missing_values_percentage)
```

```
closed_at                  63.705309
age_last_milestone_year    16.468039
age_first_milestone_year   16.468039
state_code                  0.000000
age_last_funding_year       0.000000
is_top500                   0.000000
avg_participants            0.000000
category_code               0.000000
milestones                  0.000000
funding_total_usd           0.000000
funding_rounds              0.000000
relationships               0.000000
age_first_funding_year      0.000000
latitude                    0.000000
last_funding_at             0.000000
first_funding_at            0.000000
founded_at                  0.000000
labels                      0.000000
name                        0.000000
city                        0.000000
id                          0.000000
zip_code                    0.000000
longitude                   0.000000
status                      0.000000
dtype: float64
```

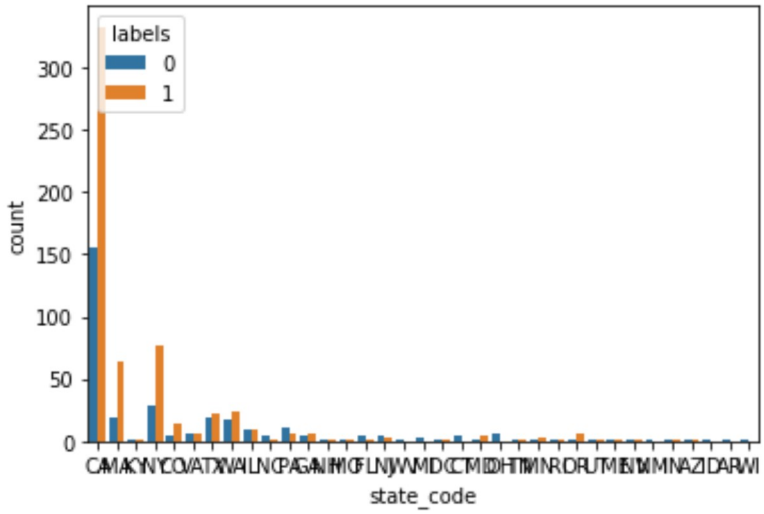
```
columns_with_missing_values = missing_values_percentage[missing_values_percentage > 0].inde
missing_values_percentage = missing_values_percentage[missing_values_percentage > 0]

# Plot the columns with missing values and their percentage count
plt.bar(columns_with_missing_values, missing_values_percentage)
plt.xlabel('Columns missing values')
plt.ylabel('Percent of missing values')
plt.title('Columns missing values and their respective percentage count')
plt.xticks(rotation=90)
plt.show()
```



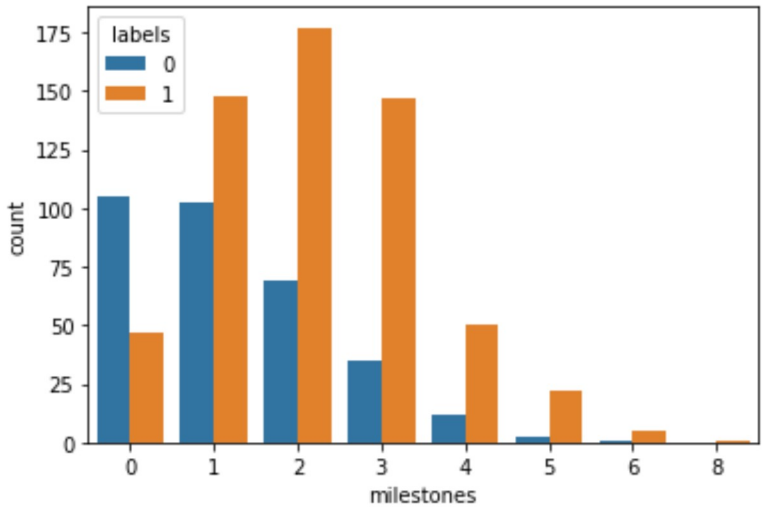

```
from sklearn.preprocessing import LabelEncoder

df_encoded = df.copy()

le = LabelEncoder()

df_encoded['status'] = le.fit_transform(df_encoded['status'])


df_numeric = df.select_dtypes(include=['float64'])

corr = df_numeric.corr()

sb.heatmap(corr, annot=True, cmap='coolwarm')

plt.show()
```



```
sb.countplot(x='state_code', hue='labels', data=df)

plt.show()

# CA produced the majority of successful startups
```




```
sb.countplot(x='milestones', hue='labels', data=df)

plt.show()
# milestone 2 made the statistically highest number of successful startups
```

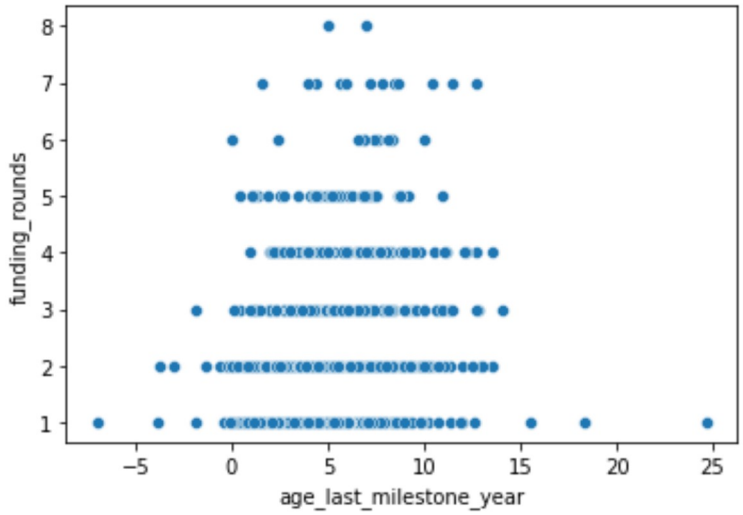



```
df.drop_duplicates(inplace=True)

print("New shape:", df.shape)
```

```
New shape: (923, 24)
```

```
sb.scatterplot(x='age_last_milestone_year', y='funding_rounds', data=df)
# age of last milestone year seems to have higher correlation with number of funding rounds
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5fbafec490>
```




```
fig, ax = plt.subplots(figsize=(12,8))

numeric_features = df_numeric.select_dtypes(include=[np.number])

for i, col in enumerate(numeric_features.columns):
    plt.subplot(3, 4, i+1)
    plt.boxplot(numeric_features[col])
    plt.title(col)

plt.tight_layout()
plt.show()
```