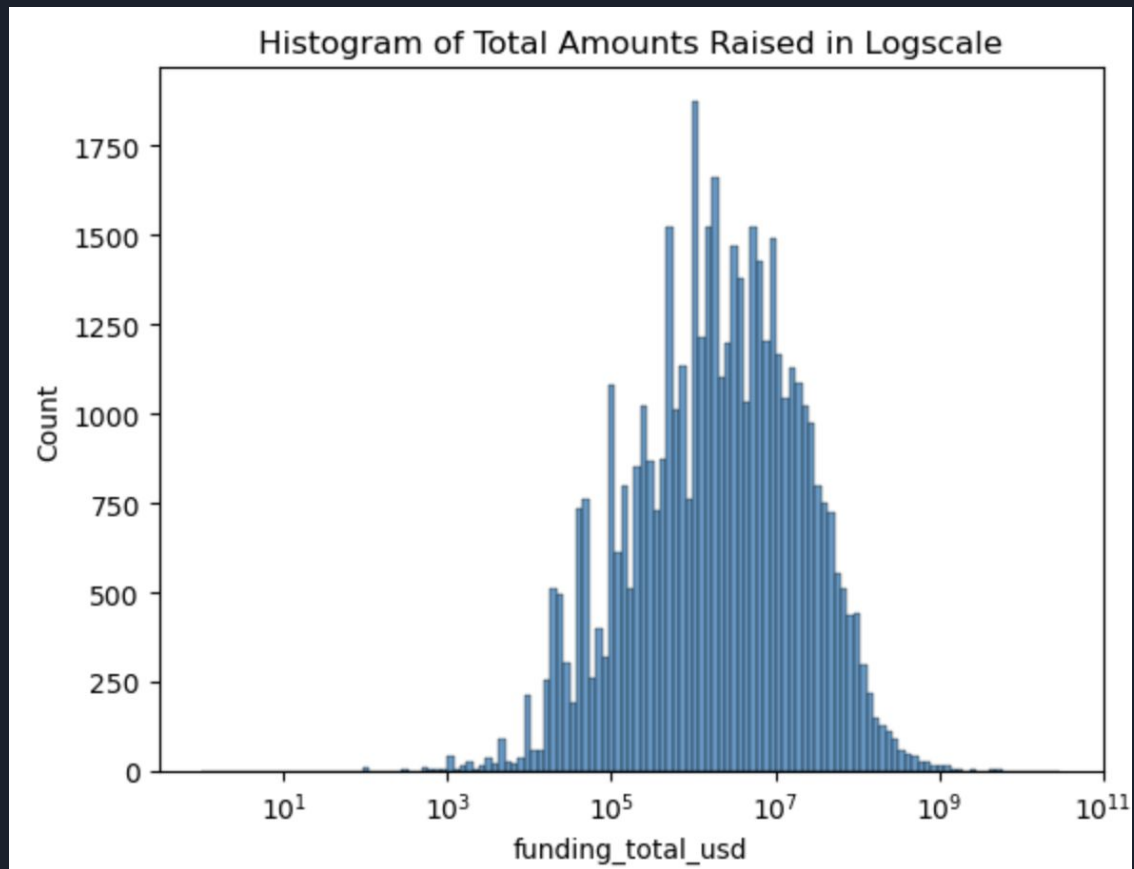# Predicting Startup Success

Recap

- The goal is to predict the success of any given startup company.

- The challenge is finding adequate data to base this model on - private companies don't have to publish or submit any information publicly.

- So far I've used data (most likely sourced from Crunchbase) which includes information such as category tags (i.e. the industry), location, important dates, funding rounds, and total amount raised by the companies.

- The plan is to analyze companies' descriptions (and logos if possible) as well to see if they produce more accurate or robust models.

- Knowing which startups are better positioned for success could help investors, as well as potential founders.
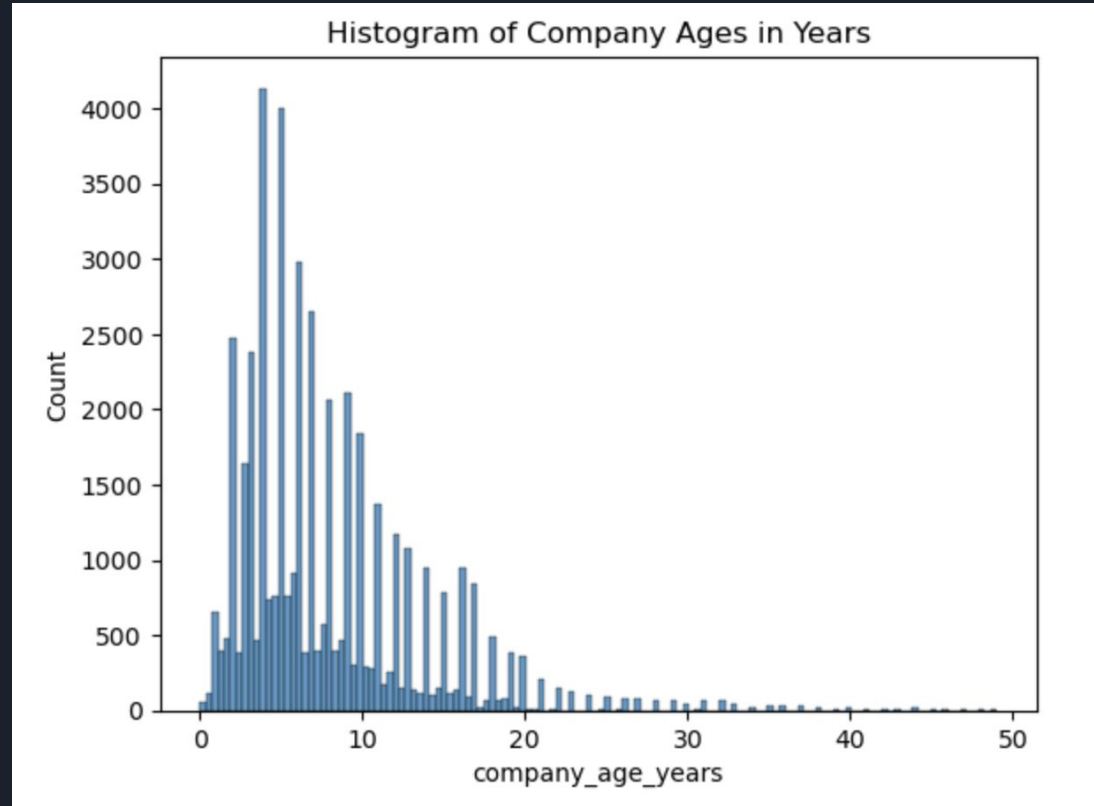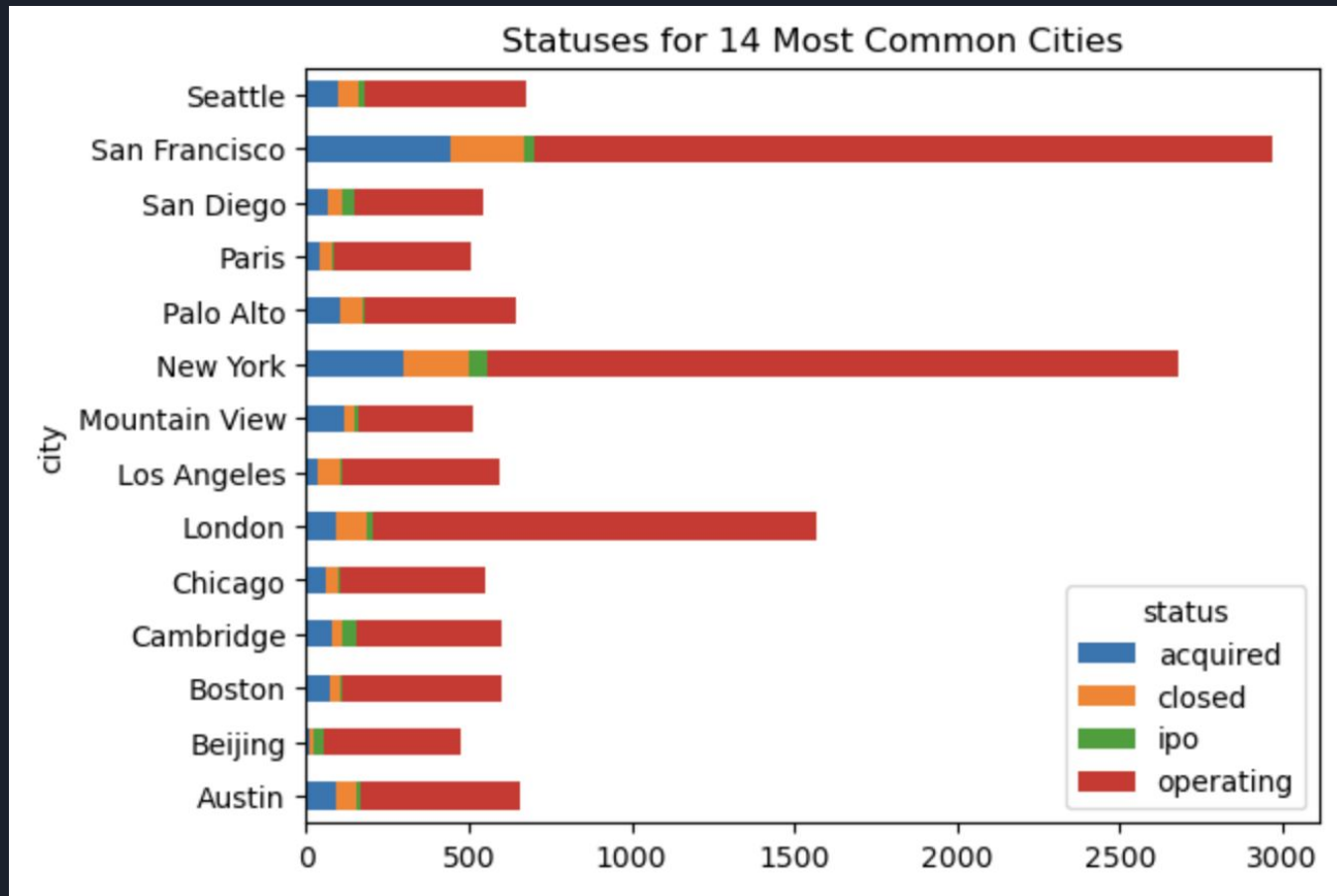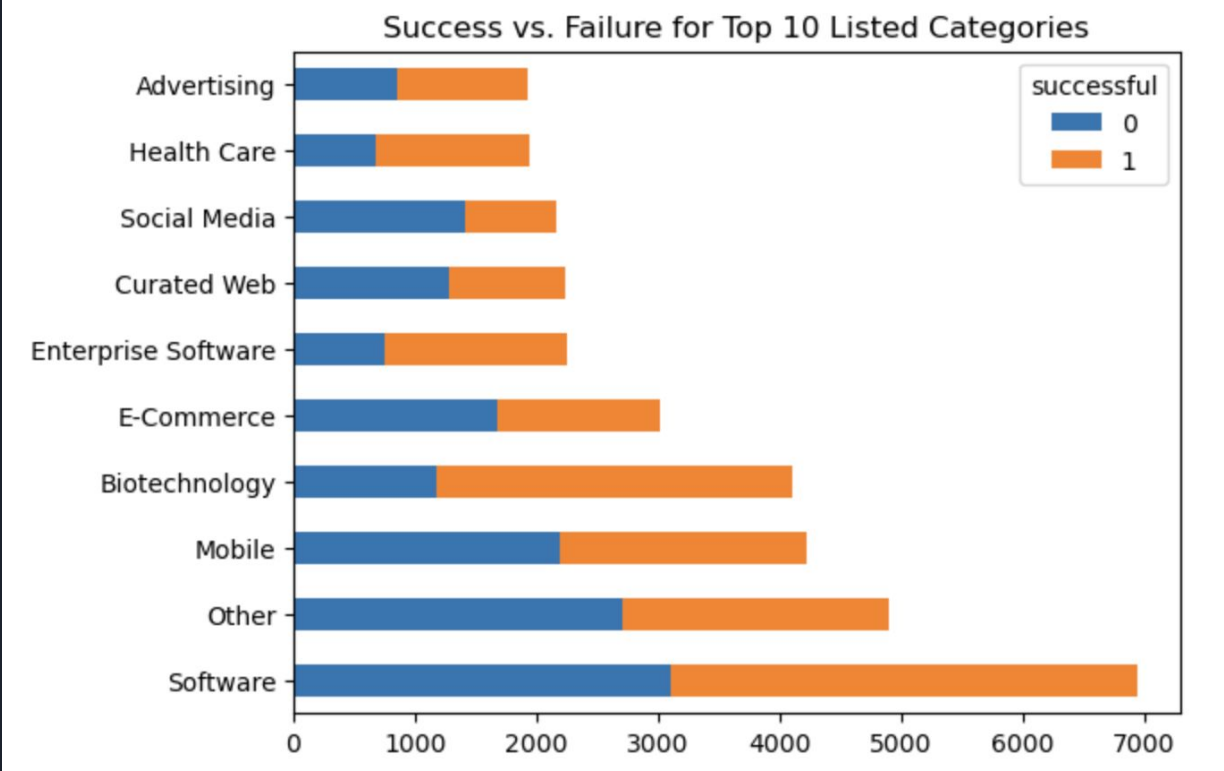
Dataset Overview

- The data set has locations, industries, starting and funding dates, rounds and amounts.

- The data was cleaned by handling nulls, converting column types, validating timestamps, changing categories to a list.

- Missing values were computed for dates using the average time delta.

- Dates were converted to integers representing days since funding or founding.

- Cities/countries and categories were one hot encoded.

- A success column was created based on the status and funding total columns:

  - A company is successful if it was acquired or IPO'd or is operating and has raised at least 1.7 million dollars.

  - It has failed if it closed or has not raised enough money.

EDA

Histogram of Total Amounts Raised in Logscale

Histogram of Company Ages in Years

Statuses for 14 Most Common Cities

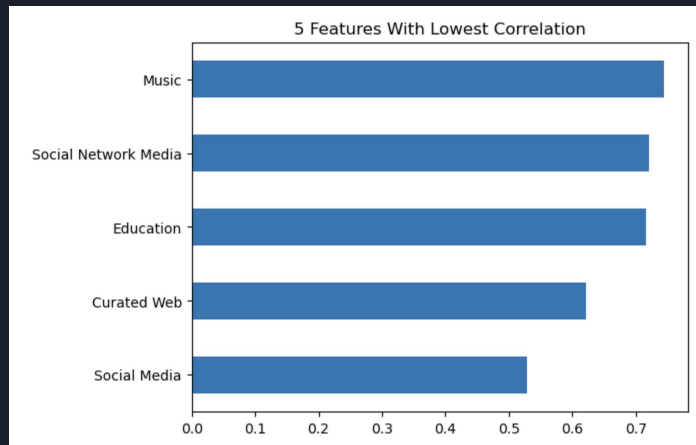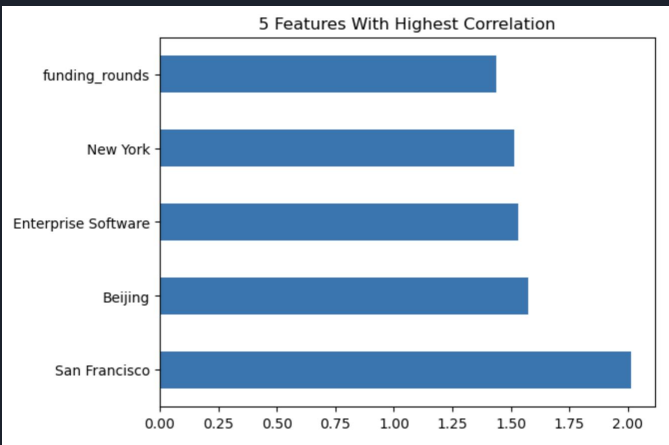Success vs. Failure for Top 10 Listed Categories

Model Evaluation

- Logistic regression model:

    - Training and testing accuracy of 72% - beating the baseline guess accuracy by 22%.

    - Precision = 76.6%, recall = 67.4%, f1 = 71.7%

5 Features With Highest Correlation

5 Features With Lowest Correlation

Next Steps

- Optimizing Hyperparameters and features

- Trying different models

- Incorporating text data of description fields

- Get a status update on the companies using URLs