Startup Success Prediction

Made by:
Malysh Igor
Tiukavkina Ekaterina

Introduction

Startups drive economic growth through innovation and job creation, but 90% fail. With exponential startup growth, investors face increasing difficulty identifying high-potential ventures early

Research goal

> Predict startup success (M&A or IPO) versus failure (shutdown) using 48+ operational, funding, and market variables to enable data-driven investment decisions

Analysis using regression models

Research objectives

1. **Data Understanding** – Explore relationships between startup success/failure and features such as funding, industry type, and geographical location.
2. **Feature Engineering** – Derive meaningful predictors from existing variables (e.g., funding intervals, milestone achievement rates).
3. **Model Development** – Build and compare classification models (e.g., logistic regression, random forest, gradient boosting) to predict startup outcomes.
4. **Model Evaluation** – Assess model performance using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.
5. **Actionable Insights** – Identify key factors influencing startup success to guide investors, founders, and policymakers.

Dataset overview

Scope

923 funded startups with 48+ features tracking their journey

Some features include:

funding_total_usd: Total capital raised across all rounds

has_roundA/B/C/D: Binary indicators for specific funding stages reached

avg_participants: Average number of investors per funding round

age_first_funding_year: Years from founding to first funding

age_last_funding_year: Years from founding to most recent funding

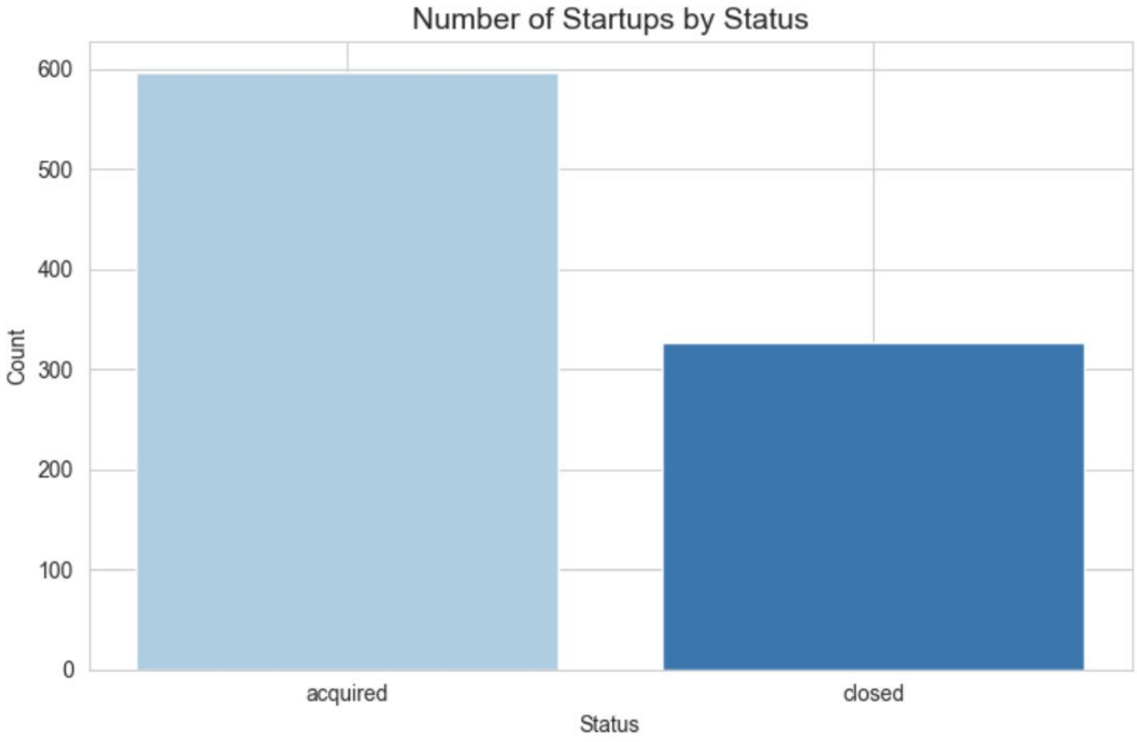
milestones: Total number of significant achievements

Target variable

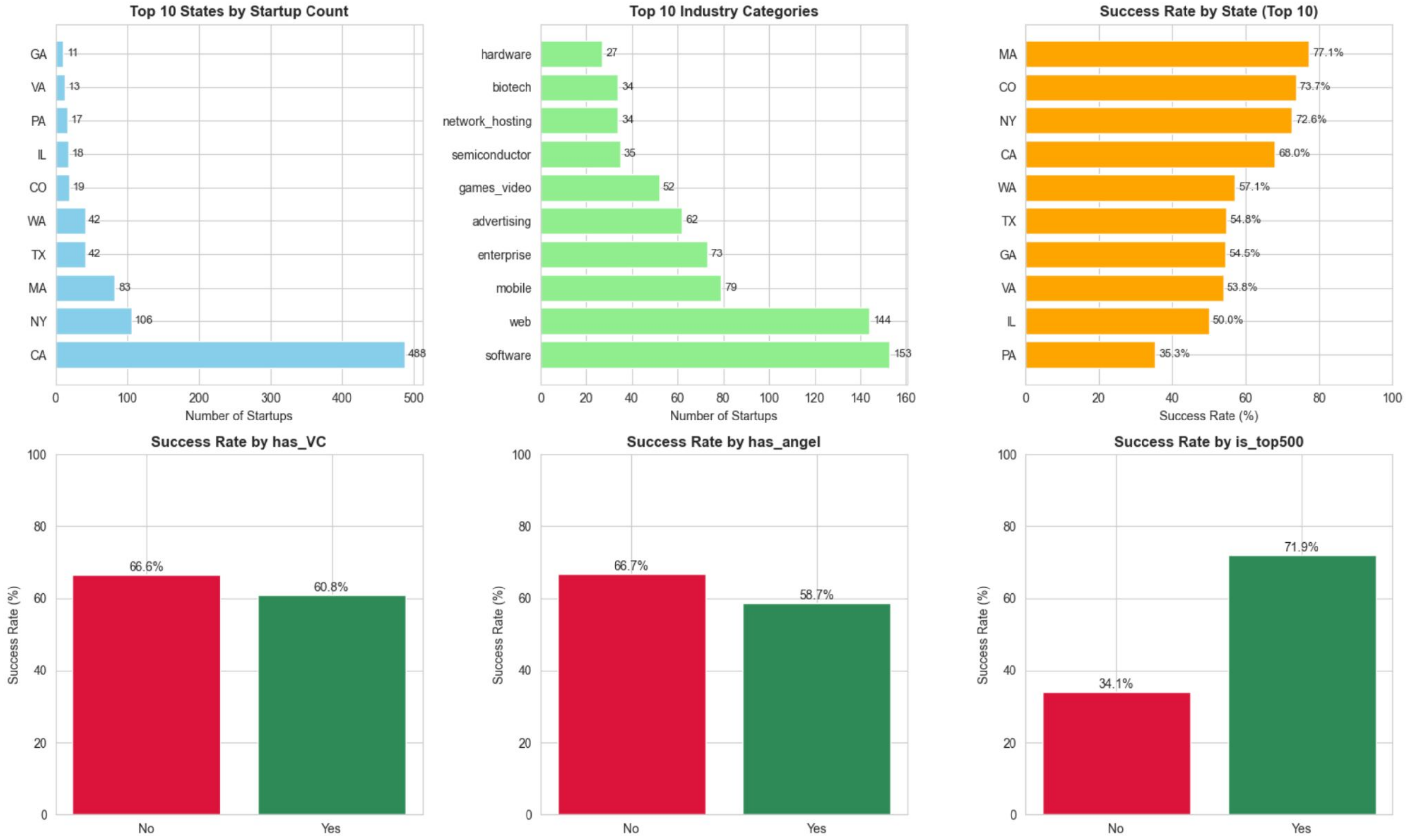
Column status is target:

> value **acquired** means success of startup (65% of dataset)

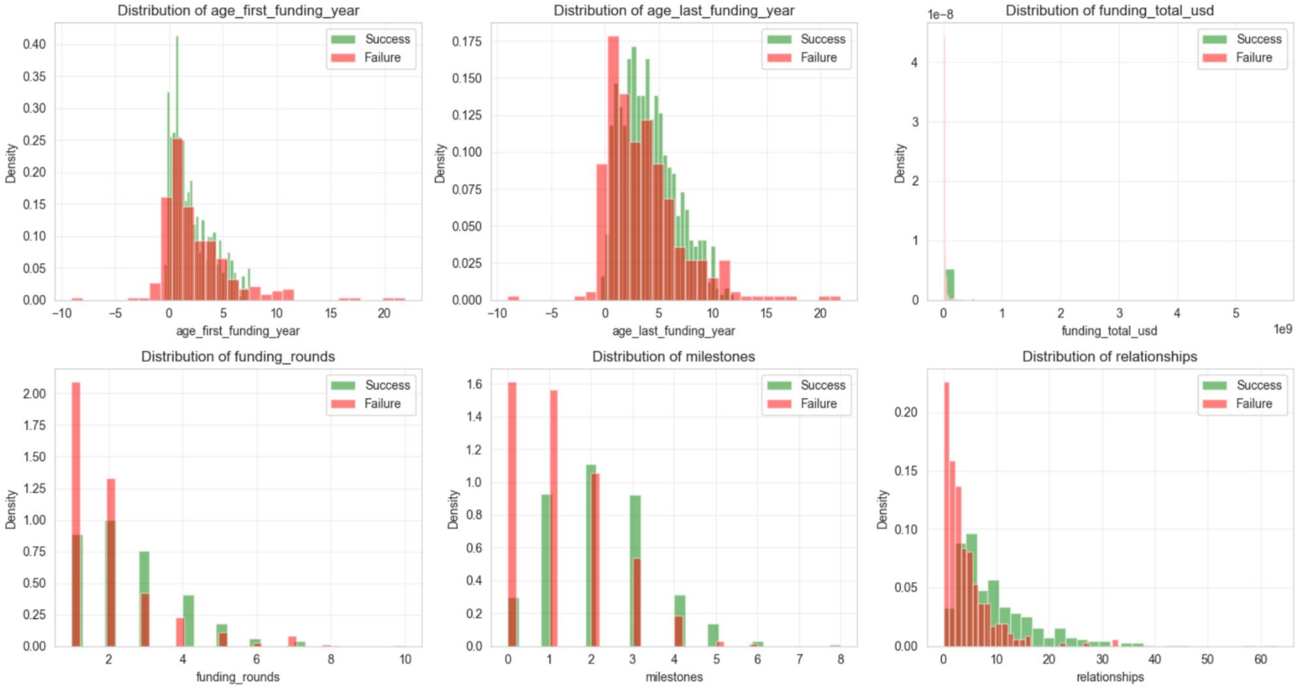
> value **closed** means failure of startup (35% of dataset)

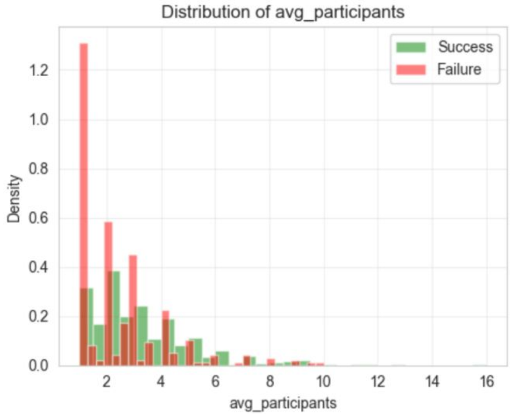


Categorical variables distribution



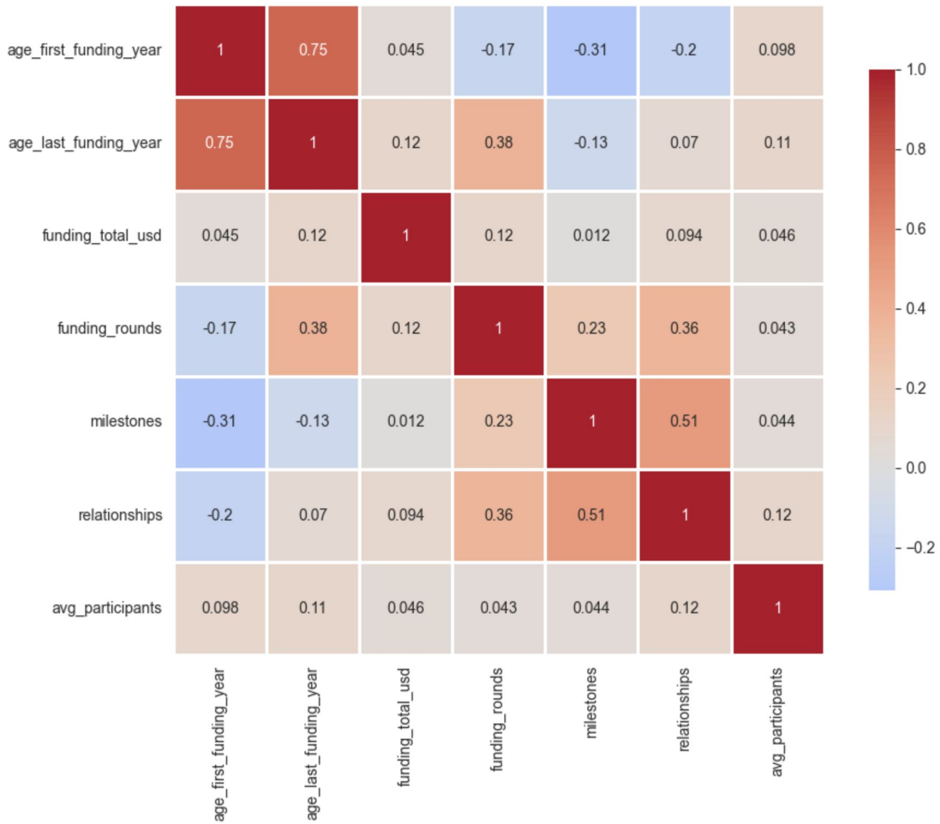
Numerical variables distribution





Multicollinearity check

- age_last_funding_year VIF = 16.09
- funding_rounds = 8.75
- age_first_funding_year = 8.40
- is_top500 = 5.74
- milestones = 3.82
- avg_participants = 3.55
- relationships = 3.30
- has_VC = 1.86
- funding_total_usd = 1.04



Logistic regression: Model

Chosen predictors:

- funding_total_usd
- has_VC
- is_top500
- milestones
- avg_participants
- age_first_funding_year

Training set size: 646

Test set size: 277

Training success rate: 64.71%

Test success rate: 64.62%

Logistic regression: Results

Pseudo R2 = 0.1652: Model explains 16.5% of variance in startup success

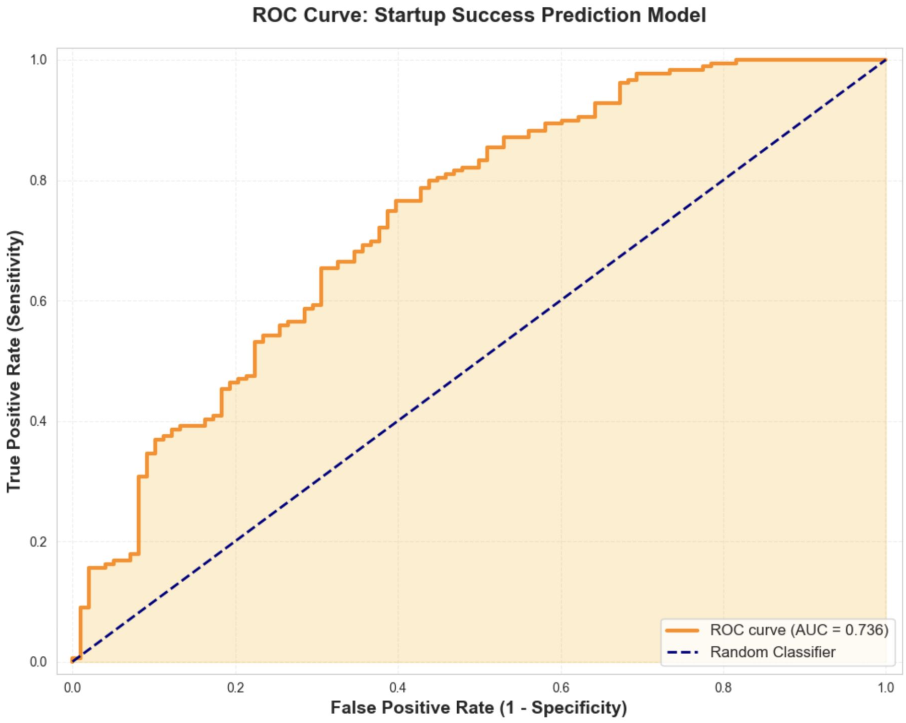
LLR p-value = 2.014e-27 (highly significant model)

The model is statistically significant but has moderate explanatory power

Accuracy = 71.5%

ROC-AUC = 0.736

Model has 73.6% chance of correctly ranking a random successful startup higher than a random failed one



Logistic regression: Significant predictors

is_top500

Coefficient: 1.2336, $p < 0.001$

Odds Ratio: $e^{1.2336} = 3.434$

Startups in top 500 lists have 3.4x higher odds of success

milestones

Coefficient: 0.5796, $p < 0.001$

Odds Ratio: $e^{0.5796} = 1.785$

Each additional milestone increases success odds by 78.5%

avg_participants

Coefficient: 0.1704, $p = 0.005$

Odds Ratio: $e^{0.1704} = 1.186$

Each additional average participant increases odds by 18.6%

General conclusions

- **Capital != success:** Funding amount doesn't predict outcomes
- **VC alone != guarantee:** Investor brand isn't enough
- **Speed != advantage:** Fast funding doesn't ensure success

What actually matters:

- **Market validation:** Top 500 status = 243% higher odds
- **Execution excellence:** Each milestone = 78.5% higher odds
- **Investor diversity:** More participants = 18.6% higher odds

Cluster analysis

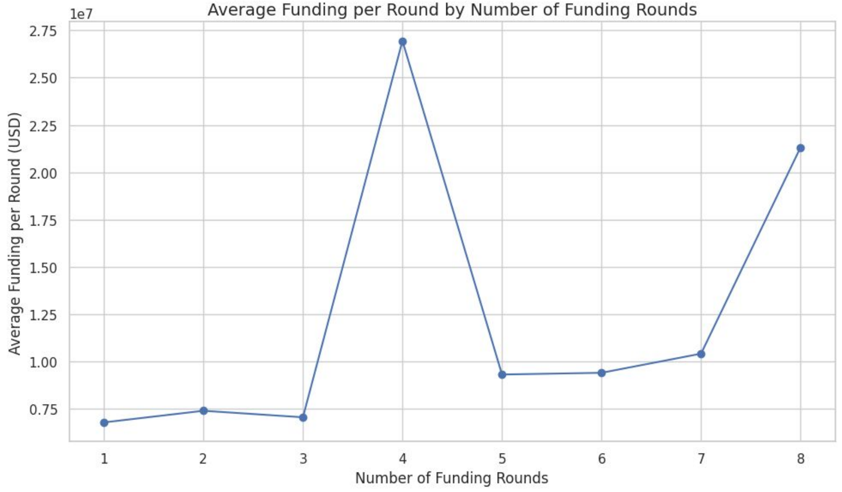
Research objectives

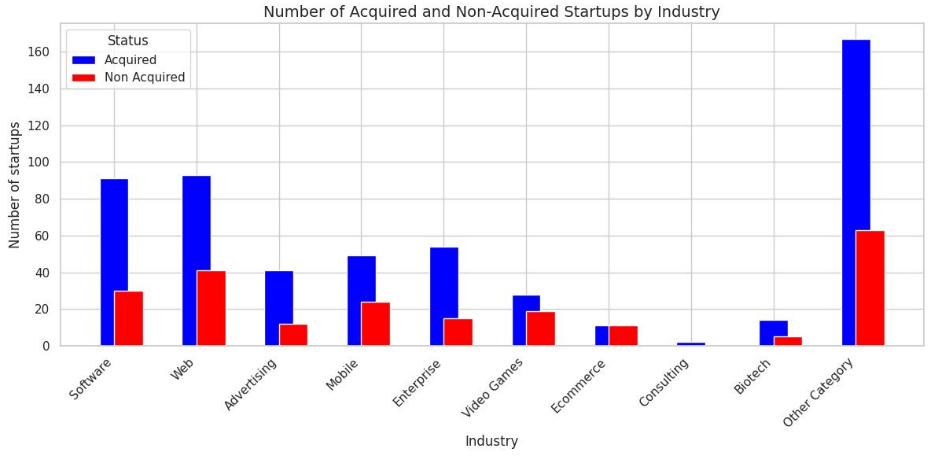
1. **Data Understanding** – Explore relationships between startup success/failure and features such as funding, industry type, and geographical location.
2. **Analyze Geographic Influence** – Determine how a startup's location impacts its success rate and access to funding.
3. **Evaluate Funding Patterns** – Assess the relationship between funding rounds, total funding raised, and investor types with startup outcomes.
4. **Examine Industry Trends** – Identify which industry categories have higher success rates.
5. **Profile Successful Startups** – Create profiles or clusters of startups to distinguish common characteristics of successful versus unsuccessful ventures.
6. **Results Evaluation** – Provide actionable insights based on cluster characteristics.

Research hypotheses

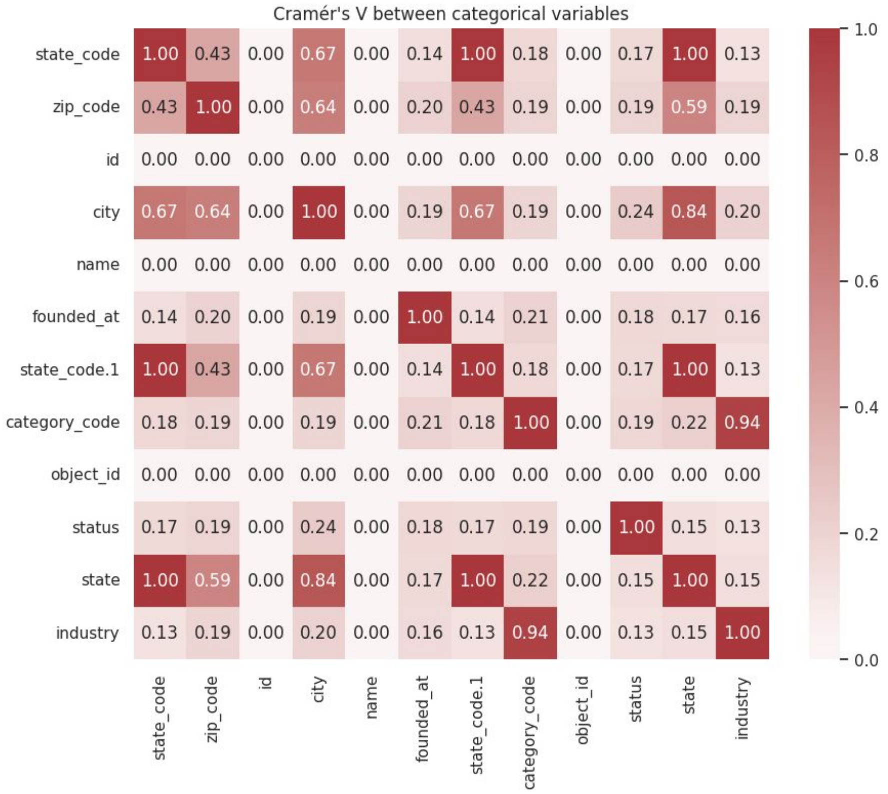
1. **Geographic Influence** – Location in California has a higher likelihood of startup success compared to other states**.**
2. **Funding Volume** – Total funds raised positively correlate with the likelihood of startup acquisition.
3. **Number of Funding Rounds**: – Startups with more funding rounds are more likely to succeed.
4. **Investor Type** – Startups with venture capital funding have a higher success rate than those with other financing types
5. **Industry Category** – Startups in the "Software" industry have more funding and higher success rates than those in other industries
6. **Number of Investors** – More investors in funding rounds are positively linked to a higher chance of startup success
7. **Top 500 Ranking** – Startups listed in the "Top 500" have a significantly higher chance of being acquired
8. **Networking Effect** – Startups with more professional connections are more likely to be acquired

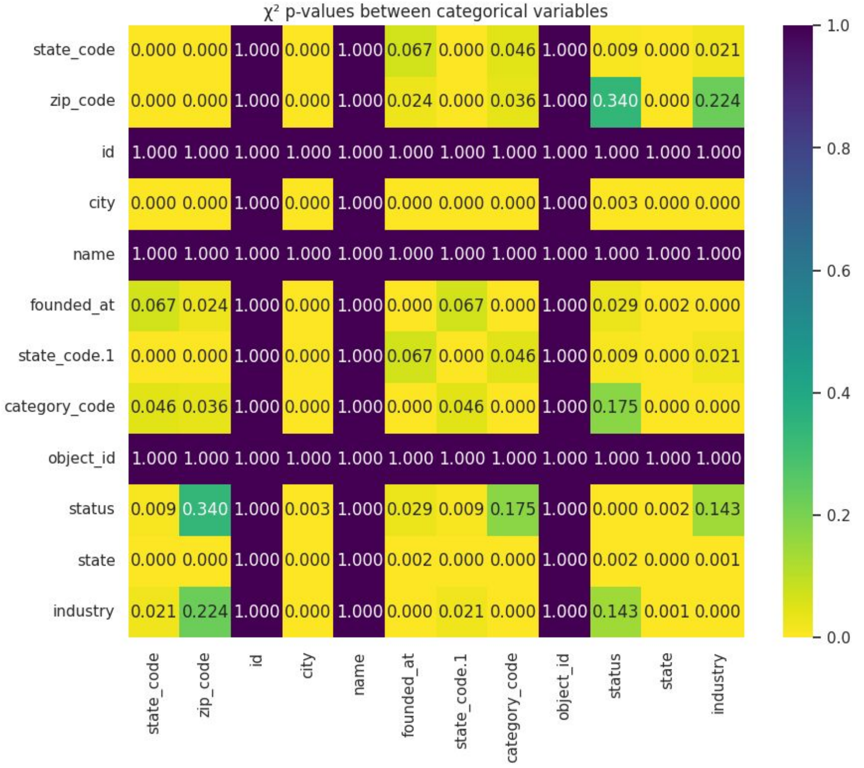
Descriptive data analysis



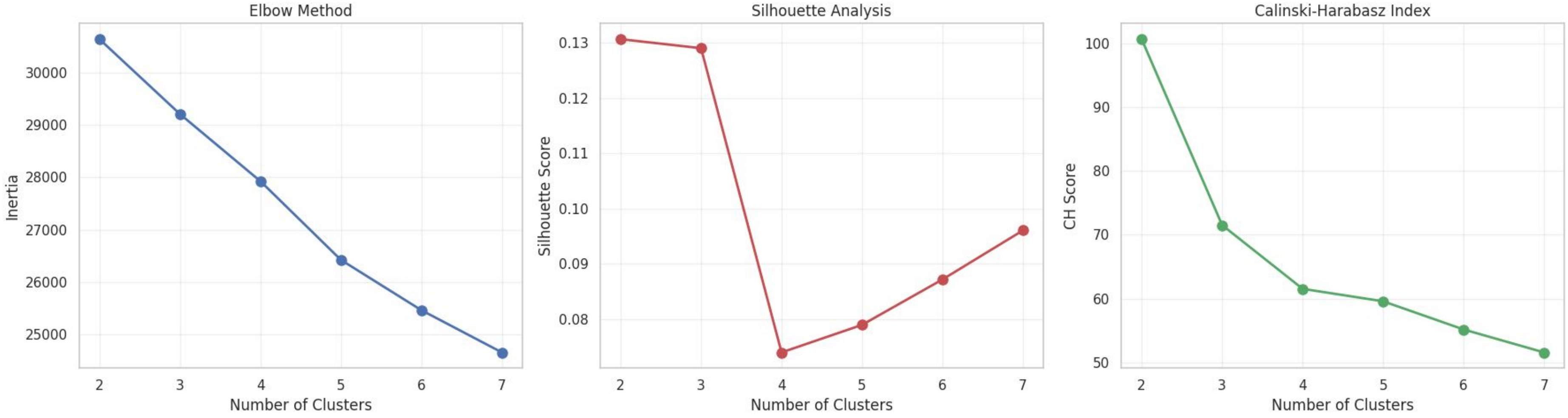


Relationships between variables



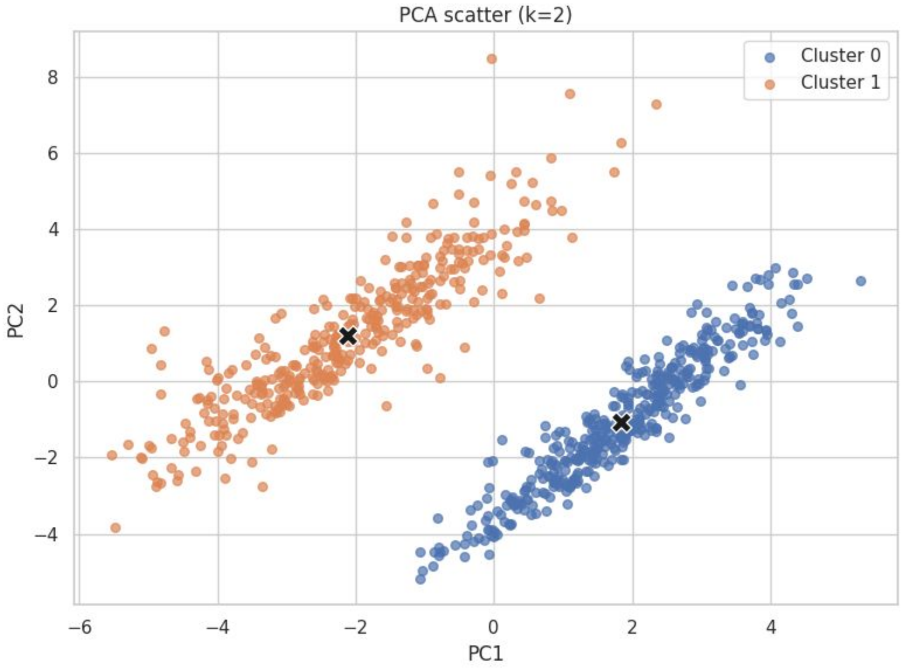


Clustering: N clusters



Optimal n clusters = 2

Clustering: Model



Cluster 0: 413 observations
Cluster 1: 357 observations

Clustering: Results

Cluster 0 "Non-California Top 500 Entities"

Representative entities:

* Jambool (CA, 1.0, Top500, $6.0M)
* Parascale (CA, 1.0, Top500, $11.4M)
* Mogad (CA, 1.0, Top500, $500k)
* RentJuice (CA, 1.0, Top500, $6.7M)
* TrustedID (CA, 1.0, Top500, $25.0M)
* threadsy (CA, 1.0, Top500, $6.3M)
* Bigfoot Networks (CA, 1.0, Top500, $20.8M)
* ViVu (CA, 1.0, Top500, $3.0M)

Key features:

- Is Ca: 1.00 (z=0.93)
- Is Ca Flag: 1.00 (z=0.93)
- State: 1.00 (z=-0.82)
- State: 3.00 (z=-0.77)
- Longitude: -120.44 (z=-0.75)
- Is Otherstate: 0.00 (z=-0.50)

Clustering: Results

Cluster 1 "California Valley Startups"

Representative entities:

* Accertify (IL, 1.0, Top500, $4.7M)
* Jumo (NY, 1.0, Top500, $3.5M)
* NSFW Corporation (NV, 1.0, Top500, $250k)
* Go Try It On (NY, 1.0, Top500, $3.8M)
* Rollstream (VA, 1.0, Top500, $7.5M)
* Summize (VA, 1.0, Top500, $750k)
* Savored (NY, 1.0, Top500, $3.8M)
* Socialthing (CO, 1.0, Top500, $415k)

Key features:

- Is Ca: -0.00 (z=-1.08)
- Is Ca Flag: -0.00 (z=-1.08)
- State: 3.41 (z=0.95)
- State: 20.02 (z=0.90)
- Longitude: -84.57 (z=0.86)
- Is Otherstate: 0.44 (z=0.58)

Conclusions: Outputs

- **Geographic Dominance**: California remains the central hub for startups, with distinct clusters (California vs. Non-CA).
- **Funding Matters**: Higher funding and more rounds correlate with startup success, especially for "Top 500" companies.
- **Industry Trends**: "Web" and "Software" industries have higher acquisition rates, reflecting market demand and scalability.

Thank you for your attention!