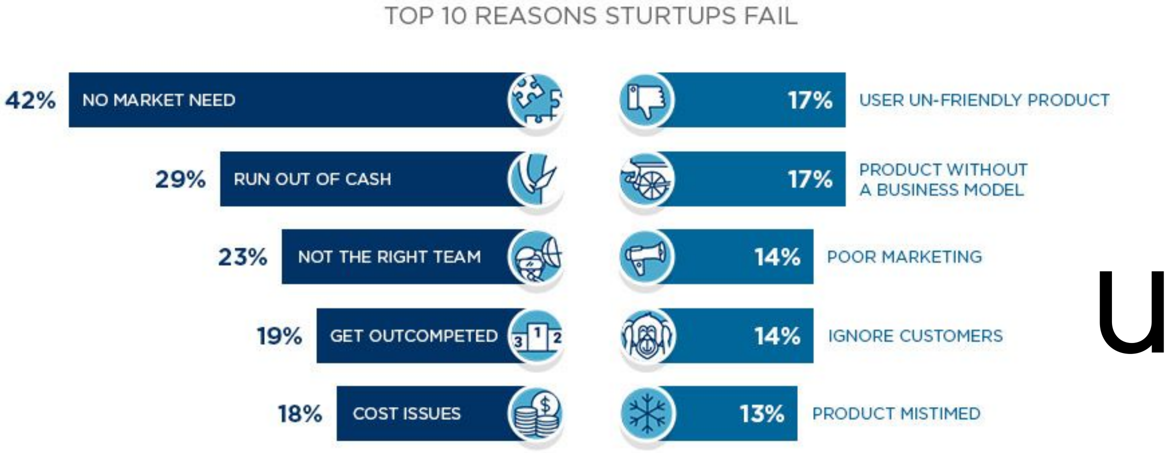
Funding a startup!

Using VC Investments Data to predict the survival of startups

Patchara Lertmanatchai | General Assembly : Data Science Immersive | Capstone Project

Introduction

Problem statement



About 1% of startups evolve into a Unicorn

Uber docker airbnb slack

The main goal of this project is to

1. Find the insights and some patterns about the condition of startup founded in the past then
2. Predict whether a startup’s current status is currently in **operating** status, **acquired** status or **closed** status by using various features variables in investment venture capital data set.

Overview of the Dataset

- Crunchbase (2014 Snapshot) provided by **True Incube** - Corporate Venture of True Corporation with totally **54,249** records and **39** feature variables in this dataset
    - Founded Year
    - Round Funding
    - Current Status
    - Age of the startup
    - Location based
    - Market
    - etc.

Issues and Bias

- **Survivorship bias** with almost 70% of data on operating start-ups.
- **Regional bias** with 46.3% of data from USA start-ups among 36% of which was from CA.
- **Market bias** with more information on Software and Biotechnology start-ups.
- Inconsistent data dated before the 1980s.

Overall Approach

- **Problem Statement**
- **Gathering Data and Data Preprocessing**
  Important to really understand the data we are working with
- **Exploratory Data Analysis (EDA)**
- **Data Preprocessing**
- **Feature Engineering**
- **Model Building**
- **Model Evaluation and Interpretation**

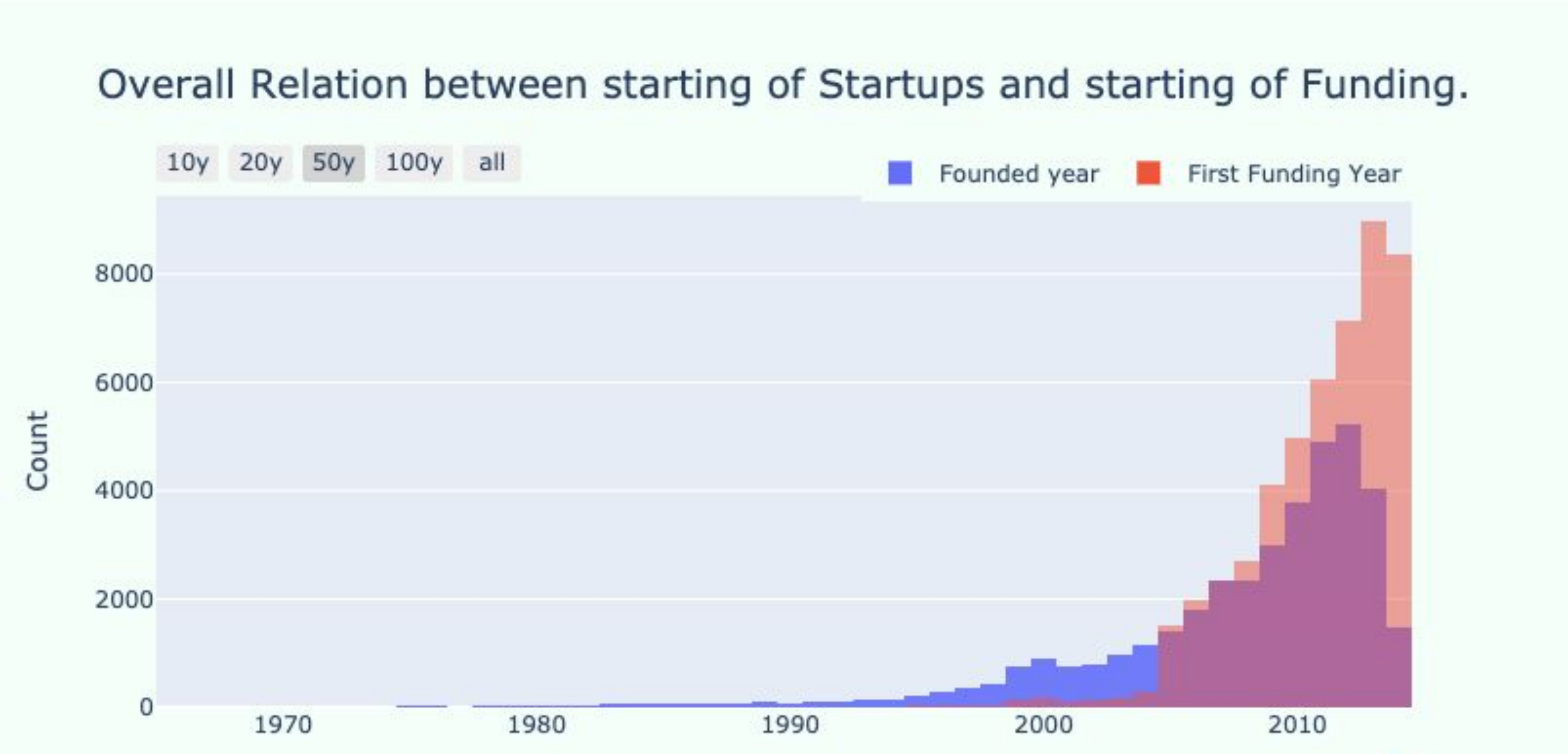
Data Preprocessing

- Stripping white spaces
- Dropping Rows and Columns
- Checking and deleting rows where all the values are NaN (missing)
- Checking for duplicate records
- Dropping rows where target variable ‘status’ is NaN
- Removing the rows for which we don’t have total funding and the breakup of the funding is null as well
- Rows with high NaN percentage
- Removing Columns that don’t contribute to model building
- Removing redundant features

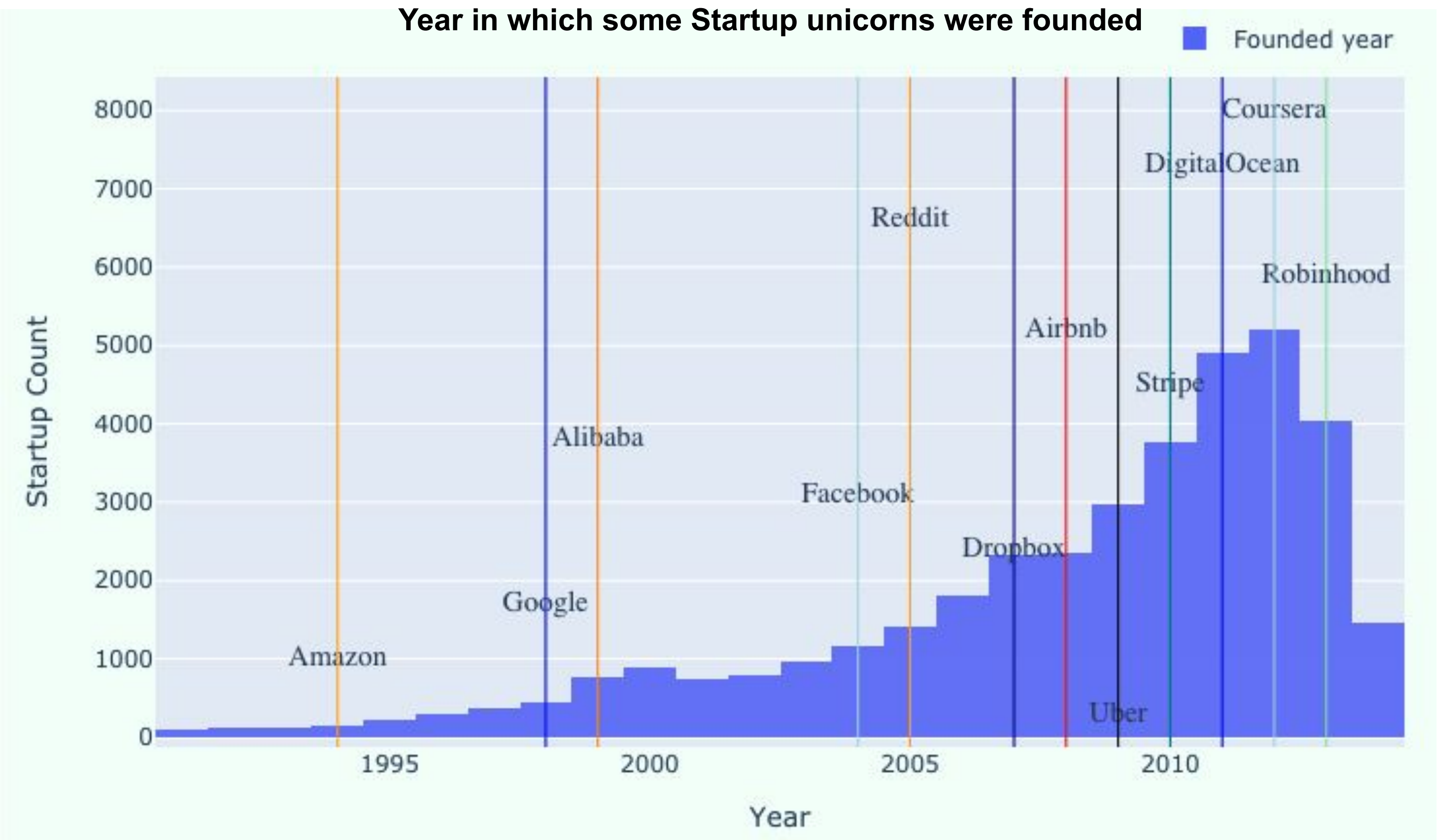
Final Data cleaning - 39,802 rows & 34 feature variables

Exploratory Data Analysis

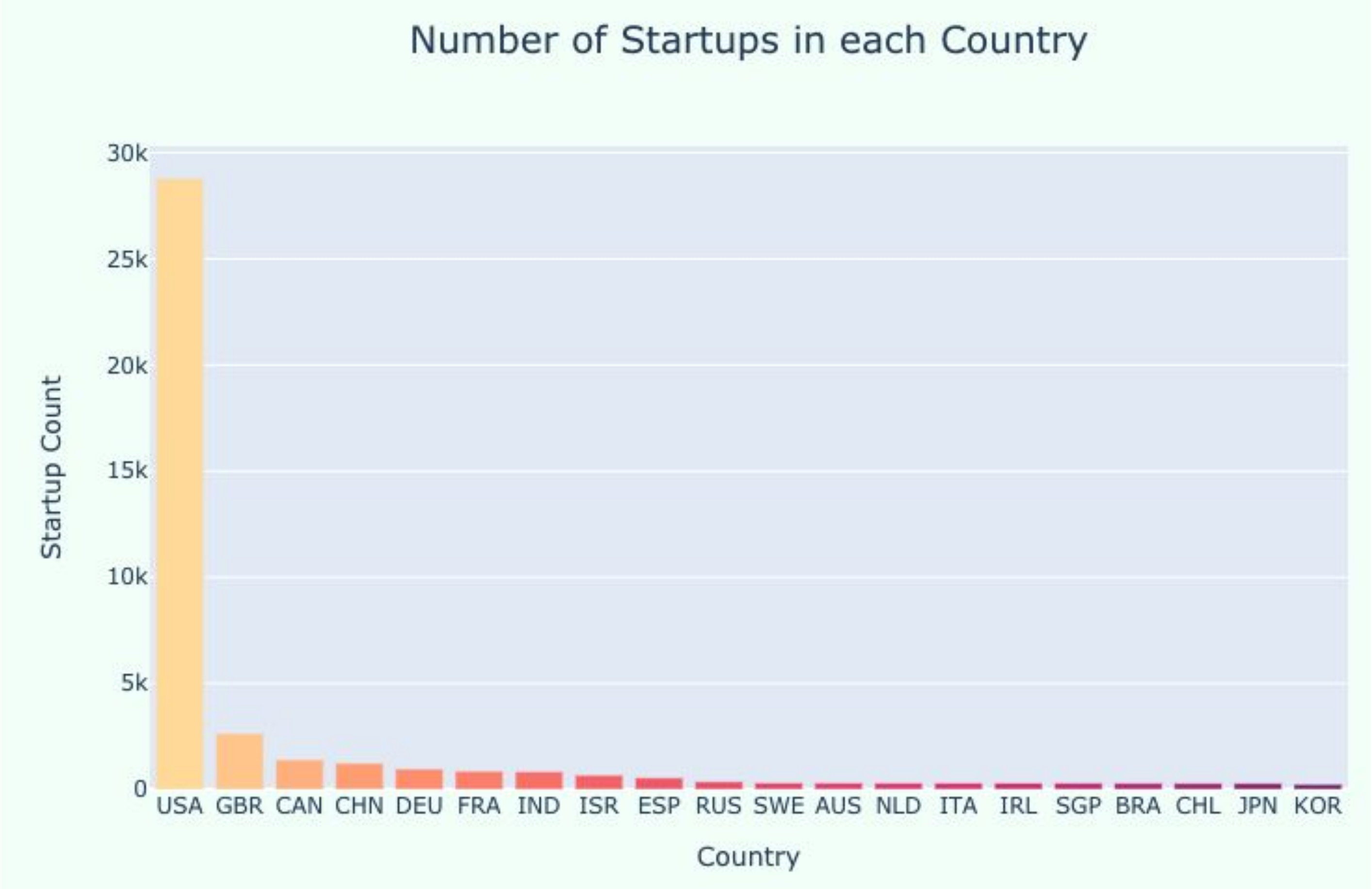
Number of Startups founded and funded each year



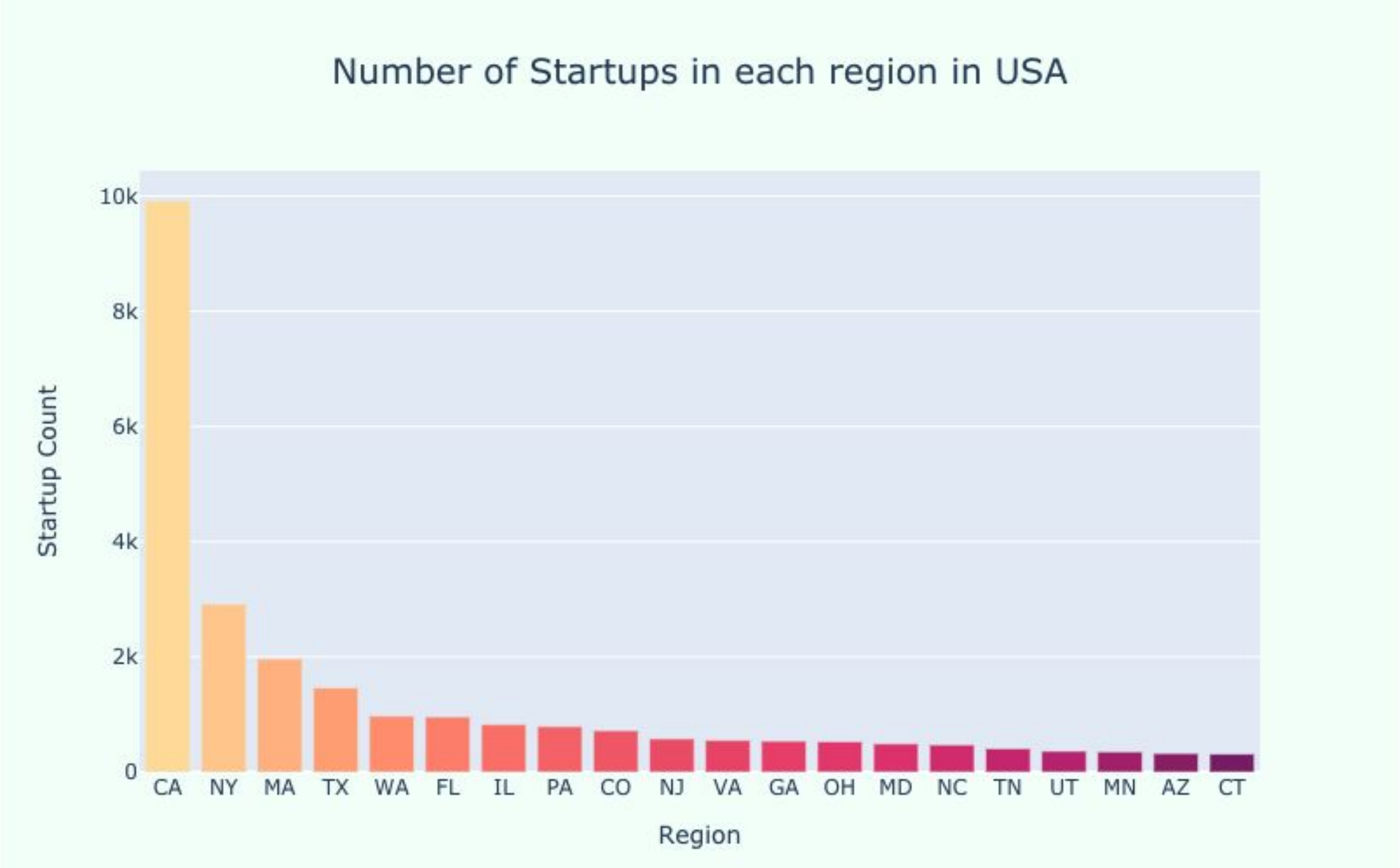
Founding year of the Unicorns



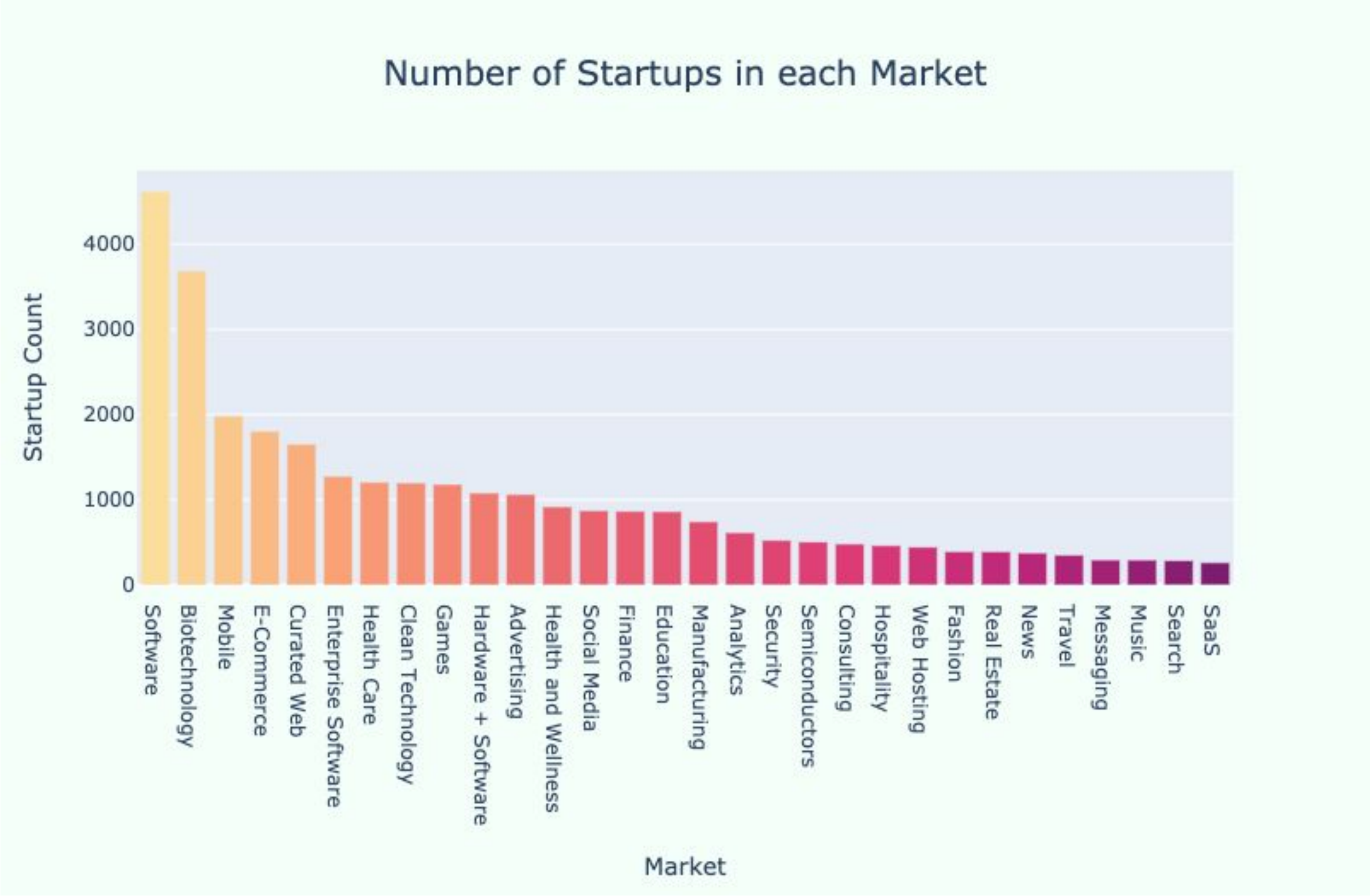
Distribution of startups in various countries



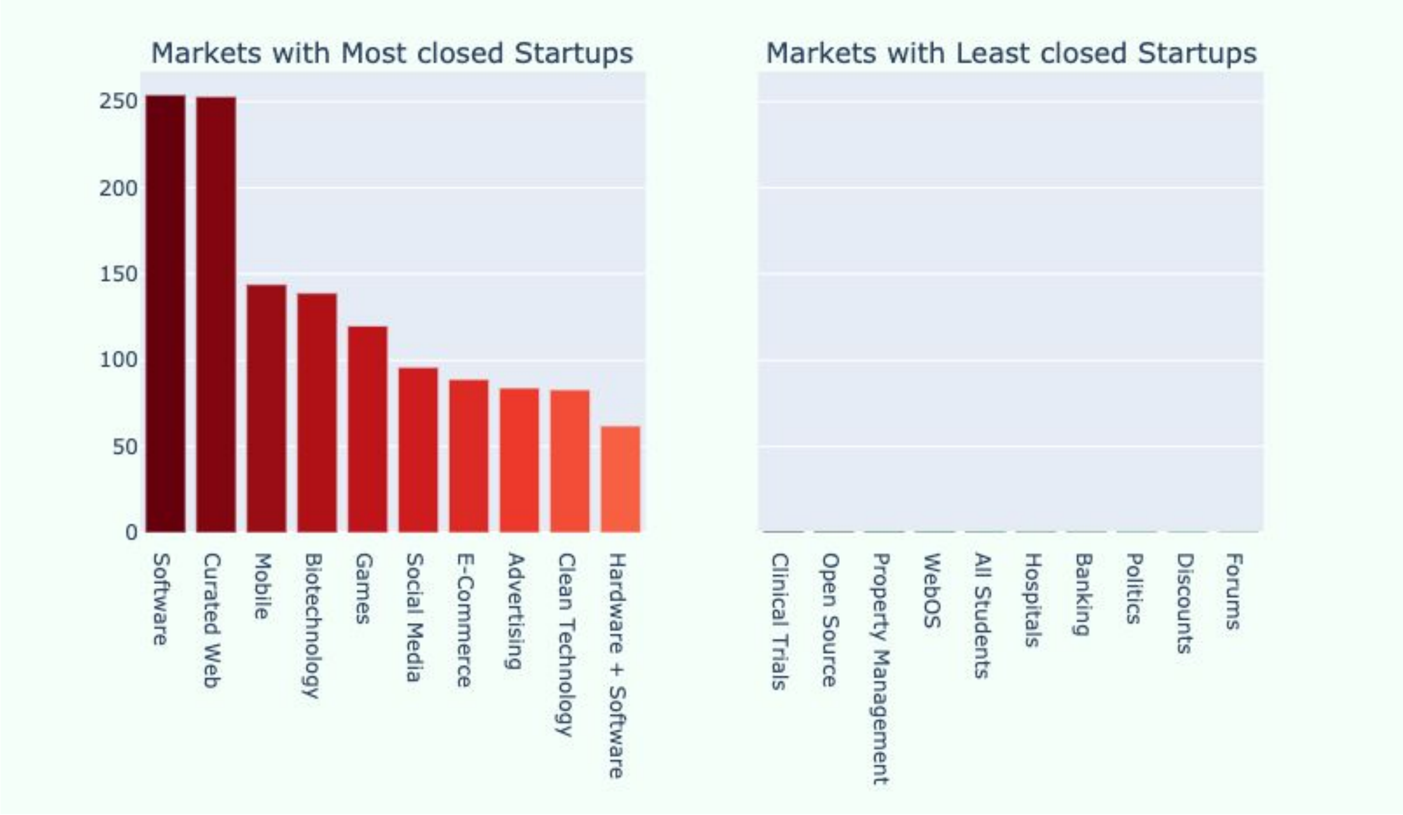
Distribution of startups among various region in USA



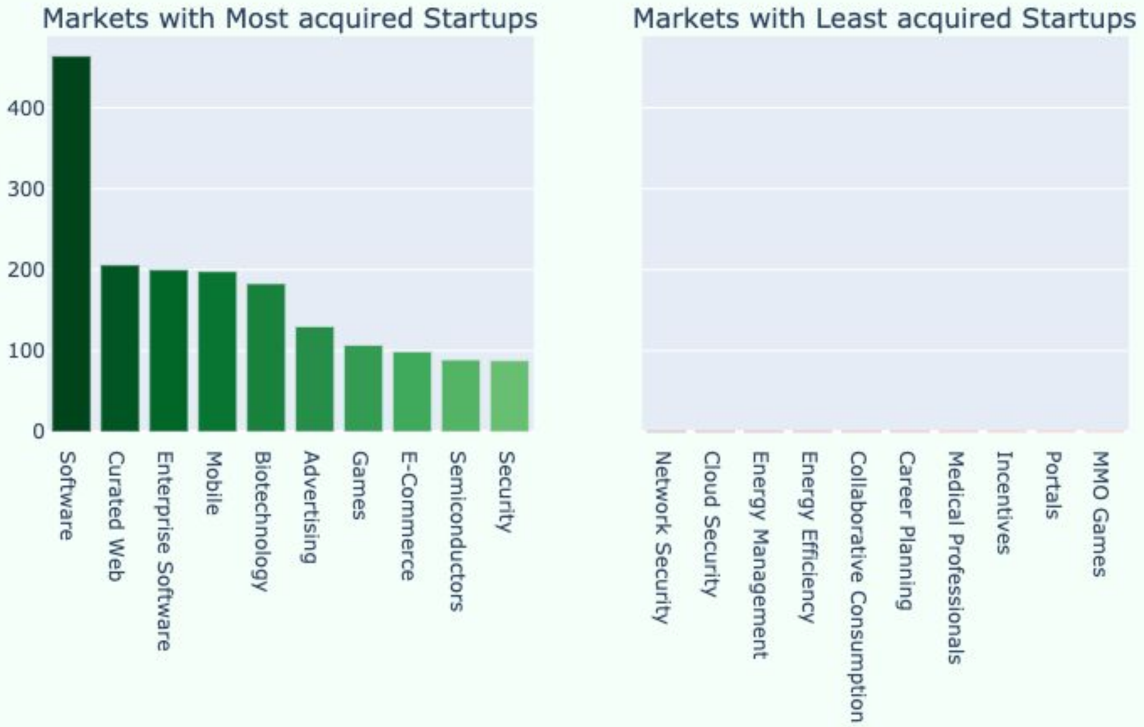
Comparing Markets



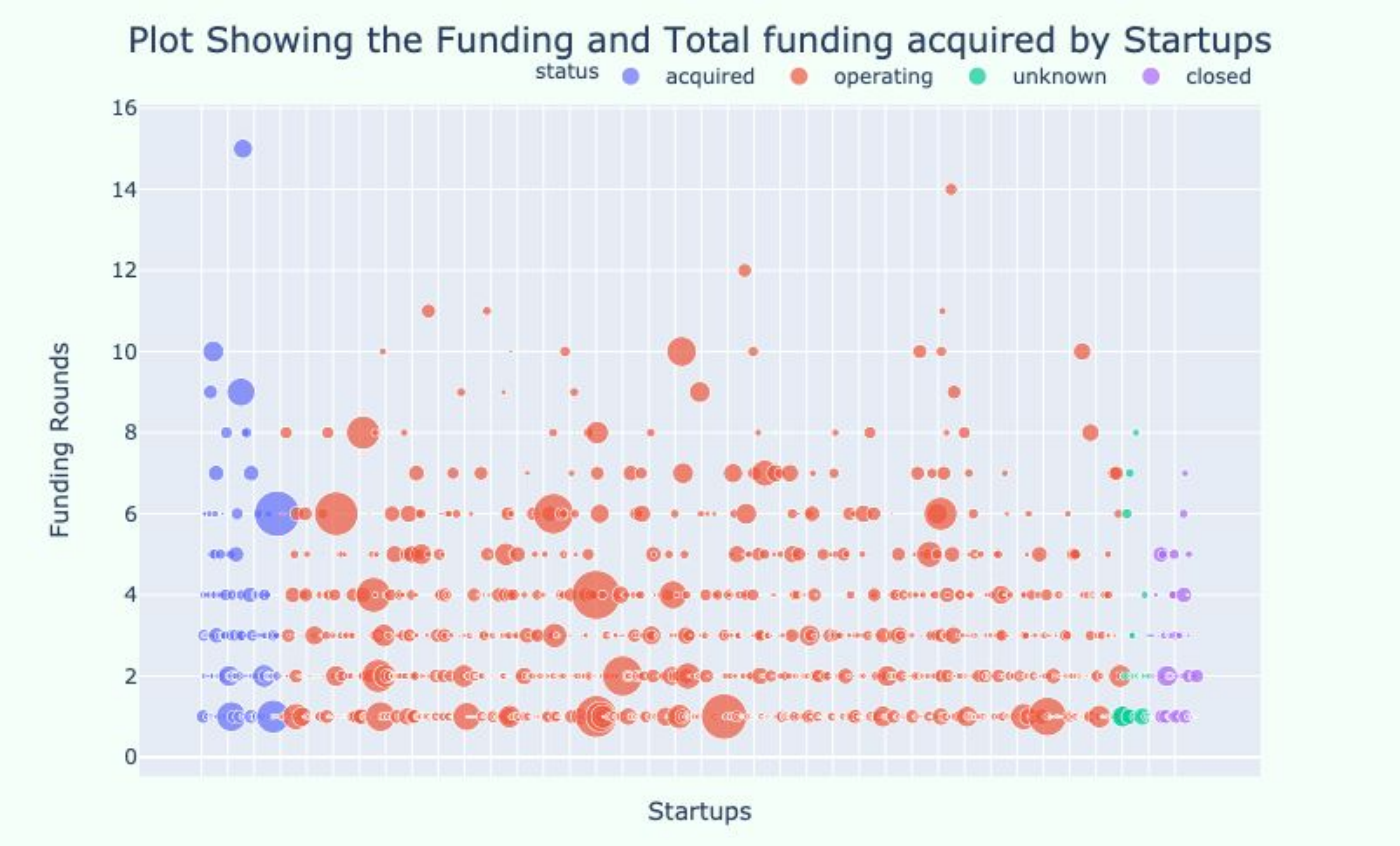
Market with Most and Least closed startups



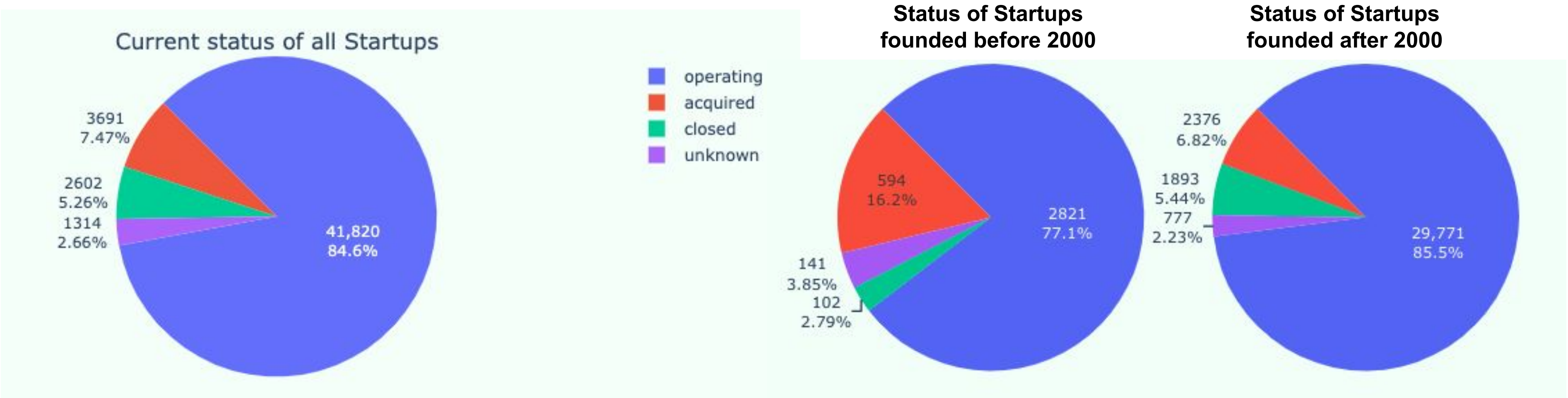
Market with Most and Least Acquired startups



Relationship between number of funding and Amount of funding in each round



Comparing Startups based on their status

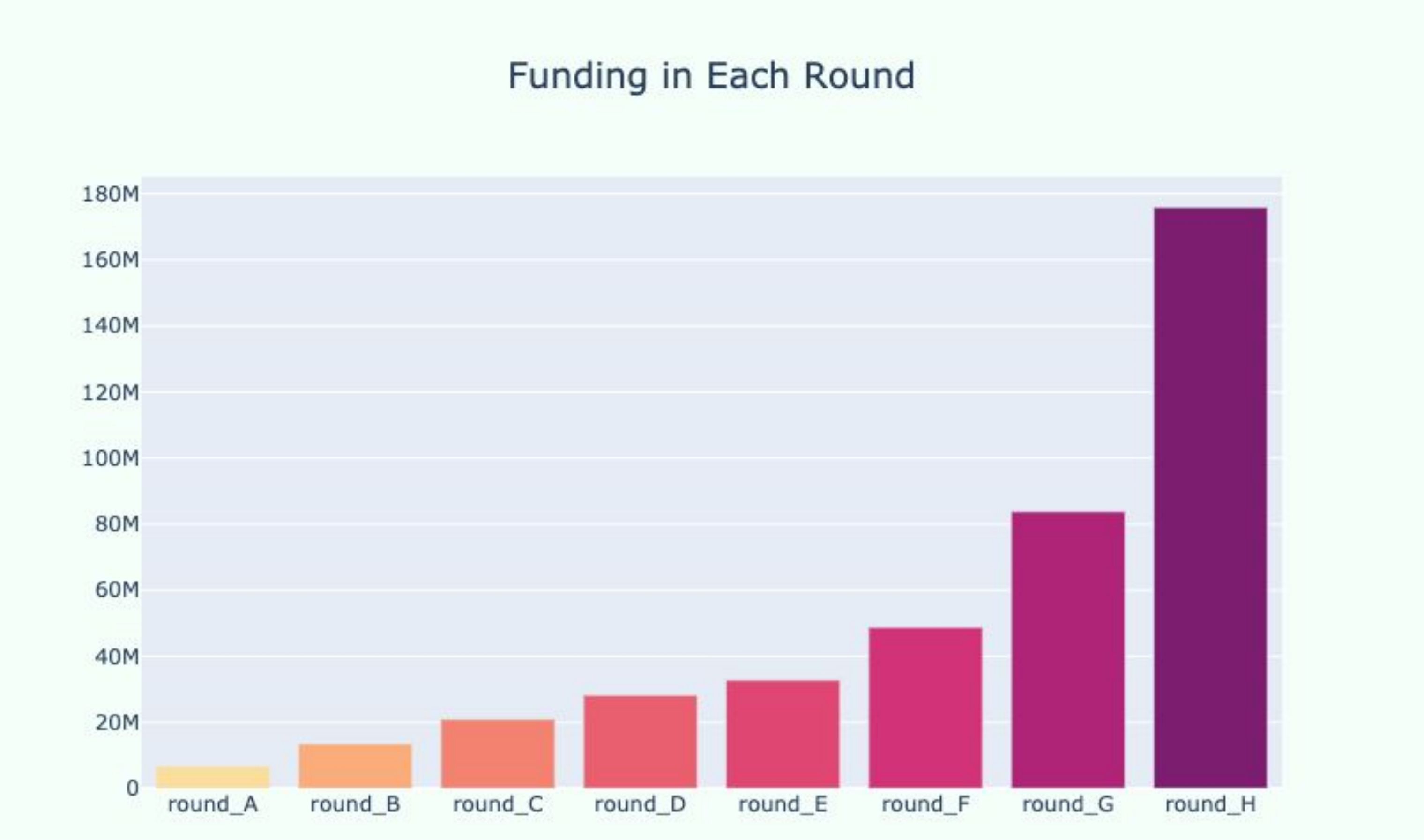


Current status of all Startups

Status of Startups founded before 2000

Status of Startups founded after 2000

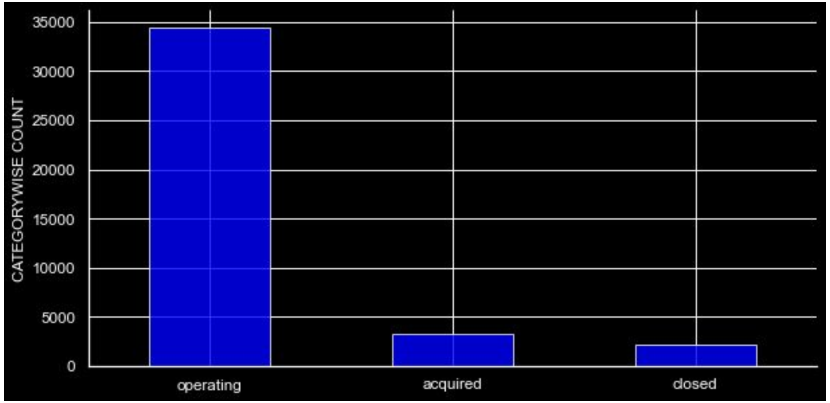
Money Invested in each round

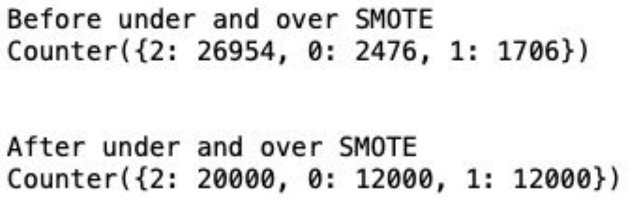


Model Building

- **Highly imbalanced dataset**

SMOTE: Synthetic Minority Over-sampling





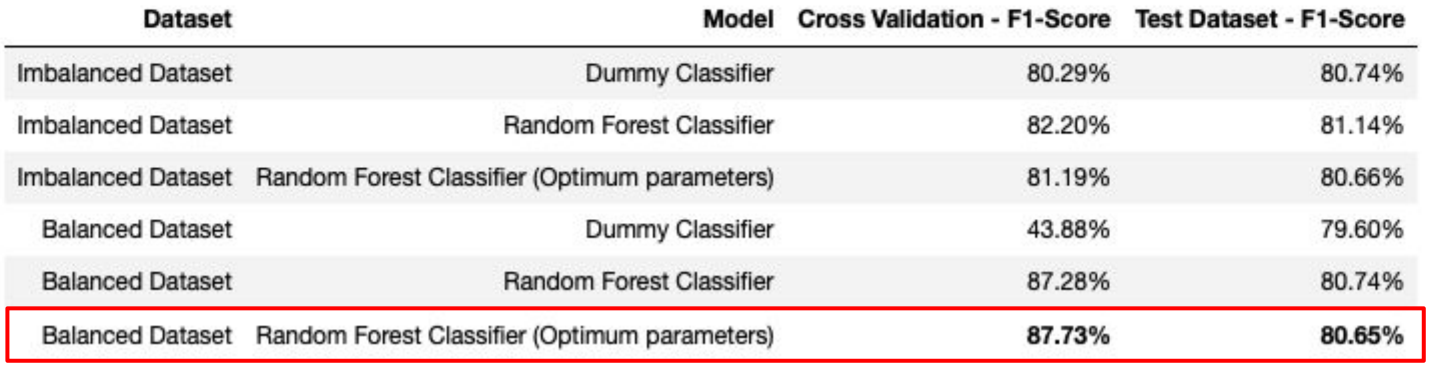
```
Before under and over SMOTE
Counter({2: 26954, 0: 2476, 1: 1706})


After under and over SMOTE
Counter({2: 20000, 0: 12000, 1: 12000})
```

- **Random Forest Classifier** model with the **Current Status** of startups as the **‘target variable’**
  - Compare with **“Dummy Classifier”** which is a baseline model that not take the relationship between the feature variables and target variables into account for predicting the target class

Model Evaluation

Model Evaluation helps us to

- Quantify the performance of a model
- Choose the best model among the models that we had built



Dataset	Model	Cross Validation - F1-Score	Test Dataset - F1-Score
Imbalanced Dataset	Dummy Classifier	80.29%	80.74%
Imbalanced Dataset	Random Forest Classifier	82.20%	81.14%
Imbalanced Dataset	Random Forest Classifier (Optimum parameters)	81.19%	80.66%
Balanced Dataset	Dummy Classifier	43.88%	79.60%
Balanced Dataset	Random Forest Classifier	87.28%	80.74%
Balanced Dataset	Random Forest Classifier (Optimum parameters)	**87.73%**	**80.65%**

Conclusion

- Since 2005, the startup scene has changed forever since there is a sensational increase of above 420%
- After some time came Facebook which in a way changed the culture and opened the door for other startups
- California (CA) has the most number of startups founded and can be seen as the startup hub of the country
- Software Market is the number market with no surprise since the technology advancement of Cloud computing, Machine Learning and AI capabilities, many new startups are trying to use these techniques to launch a new product
  - Not only in most startups founded but also for startup closed and acquired from Big Tech company
- For Venture capital, in order to invest in startup company, can’t use only financial data or numerical data
  - Management team / CEO / Culture → Data science is about Art and Science so need to be aware using technology to help making a decision

“What” and “Why” can be more interesting than “How”

- What are you trying to predict? → The current status of Startups
- Why are you doing it?
  - Help startups & VCs to make pre-decision before going to invest in Startups
- Who cares? (stakeholders) → New startups | Venture capital (VCs)
- What are some predictions your model has made?
  - Using various feature variables to predict
    - Location
    - Funding round
    -
- How can the stakeholders use it to make data-driven decisions?
  - Top industry that VC should invest in startups
  - Trending Industry that new startups should jump into the market
- Are there limitations or risks?
  - Bias Data
  - Confidential
  - Use Arts and Science

EDA and Interpret your charts - gives you practice to describe your dataset

- All your interviewers are not aware of your dataset, you are the expert in your dataset
- Introduce
  - What are the kind of features you are using
  - What do they represent in the real world,
  - over what time period was the data collected
    - Perhaps some examples from the dataset to describe the different scenarios etc.