Enze Yan
enze@bu.edu
CS 591
Homework #6

Datasets for the Final Project:

Crunchbase

Datasets Description:

Crunchbase is the datasets that keeps a record for the startup and venture capital communities. It provides information ranging from what industries are hot (trending topics such as biotech) to the potential effects of the startup's founder experience or age. The dataset includes funding, investment, and acquisition data on over 60k startups profiles contributing by 80k people around the world.

The detailed datasets contains 4 tables (table previews are cited):

- ***crunchbase.acquisitions***: details on what companies/startups were bought by other big firms. (12167 rows)
- ***crunchbase.companies***: details on each company's funding source. (46415 rows)
- ***crunchbase.investments***: details on the investments each company received. (106874 rows)
- ***crunchbase.rounds***: details on the funding round type. (78384 rows)

Among these 4 tables, there are multiple missing fields, but I do not think this will affect the overall generalization of the dataset due to the large amount of examples (training set).

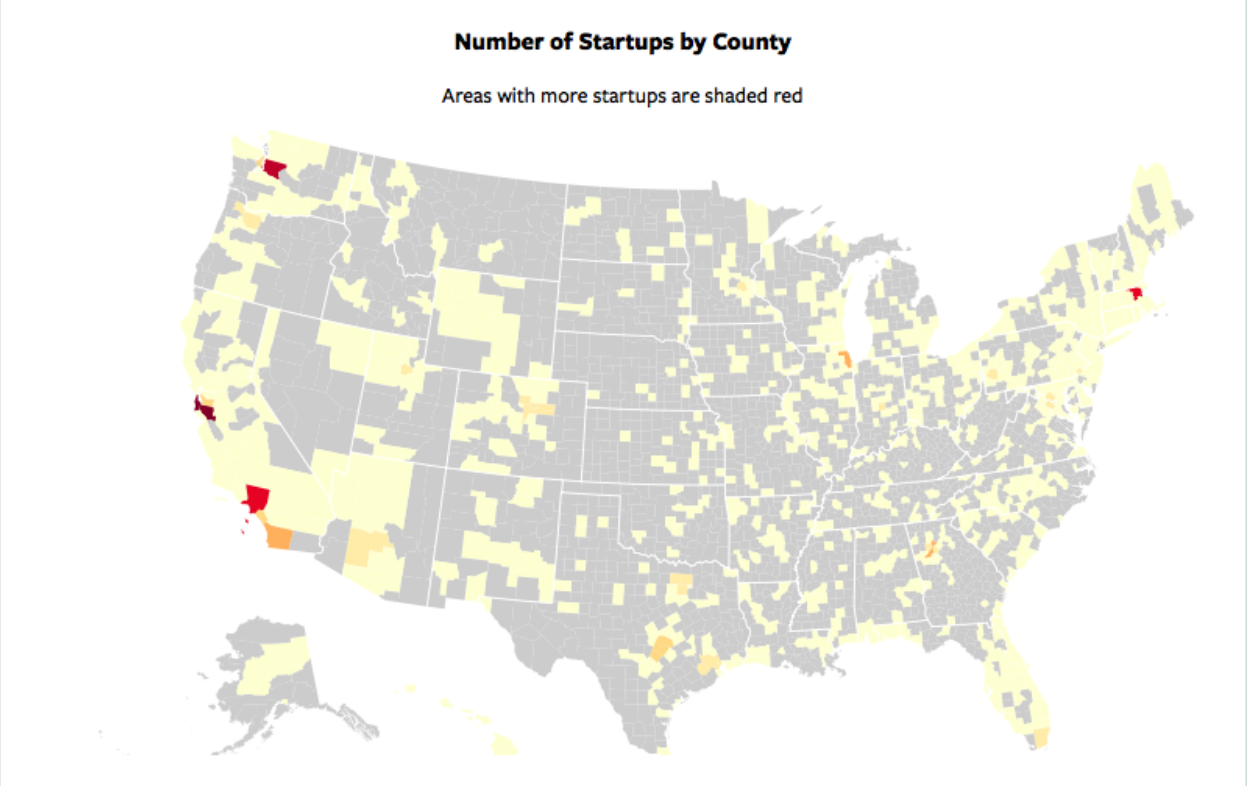
Different Hypotheses that I came up with which I could test and verify using the given datasets:

- Are there characteristics of a company—industry, location, founding year, etc.—that favor by Venture capital (VC)?
    - Using the .COMPANIES table to acquire information on industry & location, whereas .INVESTMENTS tells us more about the VC part of the stories.
- Do some VCs typically invest together, while others rarely do so?
    - Both .ROUNDS and .INVESTMENTS table provide information about VC funding type.
- Are companies raising more money earlier?
    - .COMPANIES table simply has the information on founding time versus funding size from VCs.
- Are we stuck in a bubble of a particular industry field?
- Do VCs Have an Age Bias?

- Are Experienced Founders Better?

The last three questions all requires different level of collaborations on all four given tables. For each table, I could simply use SVD or PCA to decide which characteristics are among the top importance regarding the specific hypotheses. I will be mostly using Clustering technique because it shows the grouping based on different fields. Meanwhile, Classification after training could help us better predict, say the amount of funding money a startup would receive based on its founding informations. I hope these questions I ask will indeed render me back the solution that would help those who intended to build a startup better prepare and construct the company's plan, as well as getting ready for the potential challenges in the future.

Bonus: I have done few research and found one interesting project that used this dataset to make a map showing the number of startups by the county where they are headquartered:



The above graph does indicates that location of the companies does matter when it comes to build a startup.

Sources Cited:

- https://modeanalytics.com/crunchbase/tables/acquisitions
- https://modeanalytics.com/crunchbase/tables/companies
- https://modeanalytics.com/crunchbase/tables/rounds
- https://modeanalytics.com/crunchbase/tables/investments
- http://www.bizjournals.com/portland/morning_call/2013/08/portland-is-oregons-no-3-startup.html