**SECTION 1 – INTRO:**

My goal for this project was to analyze the venture capital investment space with respect to funding data, and see what trends there were to discern. By analyzing the funding data, locations, involved individuals, I figured I could develop an additional set of insights for a hypothetical VC firm to add to their current investment criteria. **Therefore, the goal of this project is to create a set of insights that serve as additional criteria for an investor to use when evaluating startup VC investments, not as an investment thesis in and of itself.**

**SECTION 2 – BACKGROUND:**

The data I selected was from Crunchbase, via a Kaggle dataset, and contained a variety of funding information regarding startups between the years of 1980-2013. There were 11 separate tables:

* **Degrees** –contained degrees of some of the successful people associated with startups in the list
* **Funding Rounds** – contained detailed information on all individual funding rounds for companies in the dataset, including, the amount raised, currency, date, round, etc. This was used heavily in my analysis
* **Funds** – Included information on the individual funds that invested in companies in the datset
* **Investments** – information on all investments into companies on the list, including object id for both the investor and the company they invested in
* **IPOs** – pricing & date information on all companies within the dataset that pursued an IPO
* **Milestones** – didn’t use, but contains information on different major milestons the companies reached
* **Objects** – this was the primary table, which provided an ID number for all companies, investors, and individuals that was then used in all other tables, and tied that ID number to the name. Used in almost every query I ran
* **Offices** – This showed the location of headquarters for all companies in the list
* **People** - this was a list of important figures (executives, investors, advisors) of companies in the dataset. This was used to determine how many successful startups each of these individuals was related to.
* **Relationships** – this tied the id for the people in ‘people’ to the id for the company they were involved with

**Questions I asked to help build my analysis:**

* Which companies are the most successful? How can we define that?
* Which industries receive the most funding? Which industries receive the most funding among successful companies?
* Which industries have the highest percentage of successful companies?
* Where are the majority of successful startups located?
* How much are successful startups raising? Is it less or more than the overall average?
* How fast are successful startups raising rounds? Does it differ from the average?
* How fast do successful companies move from founding to IPO? How about founding to acquisition? How has that changed over the years?
* Who are the most successful operators?
* Who has had the most success investing?

**A note on ‘successful’ companies:**

Because I needed to define success for the sake of this analysis, I decided to focus on companies that had either pursued an IPO, or had been acquired by a larger organization. Companies that met this definition of successful companies accounted for 5.5% of the companies in the dataset. Much of the analysis in my project is based on comparing trends among this 5.5% with the same trend among all companies in the dataset.

**SECTION 3 – CLEANING & ANALYSIS**

**Tools used in this project:**

* **Python** -
* **Python** Pandas Library – to load CSV files into Python
* **Python** Profiling Report library – to conduct EDA and understand the contents of the dataset
* **Python SQL Alchemy Library** – to create a SQL database to analyze
* **PostgreSQL using PGAdmin 4** – to run queries to analyze the data
* **Tableau Desktop** – to build visualizations for the final presentation
* **Excel** – to compile and clean results from SQL queries

**Cleaning & Analytical Process:**

* Downloaded 11 CSV files from Kaggle dataset
* Uploaded CSV files into Jupyter Lab python file using Pandas library
  + Created dataframes
  + Created profiling reports for EDA
    - Read through all reports, evaluated null strategy and columns to avoid due to missing data
    - If zeros or nulls < 12% of data in column, leverage where xyz is not null or where xyz != 0 to calculate summary statistics and analyze data within PostgreSQL DB
* PostgreSQL Database Creation w/ PGAdmin
  + Created from Jupyter lab with SQL Alchemy library
  + Ran script for all 11 CSV files to create 11 tables within Capstone\_VCFundingData database
* Queries created (PostgreSQL in PGAdmin 4)
  + List of all companies that meet the definition of success
  + How much each successful startup raised in total (ordered by USD descending)
  + Average total funding raised per successful startup ($35M)
  + Average size of each funding round
    - For successful vs all companies
  + Query to find the average time from initial funding to IPO (2.89 years)
    - Average time from initial funding to acquisition
  + Location of offices of successful companies
  + Average time between seed round and series A of successful companies
    - Separate query for all companies
  + Average time between series A and series B of successful companies
    - Separate query for all companies
  + Average time between series B & C of successful companies
    - Separate query for all companies
  + Average number of funding rounds for successful companies
    - Average number of funding rounds for all companies
  + Group by of number of successful companies per category code
    - Total startups per category code
  + Total funding raised per category for successful vs total startups
  + Total funding per category for successful companies where year last funded was 2012 or 2013
    - To then compare the top IPOs of 2013/2014 and see what industry they fell into
    - Take this list and compile top IPOs from the year 2014
    - Find max date
  + Time from founded\_at to ipo
    - Time to IPO for companies founded between 1998-2013
    - Time to IPO for companies founded before 1998
  + Time from founded\_at to acquisition
    - Time to acquisition for companies founded between 1998-2013
    - Time to acquisition for companies founded before 1998
  + Count of successful startups people are tied to
    - Who to approach for co-investment, information, jobs, etc
  + How many successful companies each investor invested in
    - How many total companies each investor invested in
* Tableau Visualizations
  + Map of office locations for successful companies
  + Total funding deployed per industry for all companies and just for successful companies
  + Average size of funding round for successful vs all companies
  + Word cloud – count of successful startups individuals have been involved with
  + Count of successful companies by industry
* Excel Work
  + Built sheet with KPIs for funding round timing for ‘successful’ companies and all companies
  + Sheet with time from founding to success for both IPOs & Acquisitions
    - Broken out by total, 1980-1998, and 1998-2013

**SECTION 4 – RESULTS:**

* **Location**
  + Unsurprisingly, the majority of startups are headquartered in California, although there is a significant minority in NY, MA, TX, WA. For a VC firm looking to establish themselves, they would likely need to establisha presence in SF/Bay Area, NYC, Boston, and Austin.
* **Industry**
  + There is very little consensus here about which industries perform best. Industry specialization is best left to subject matter experts to investigate further.
* **Funding Amount**
  + $35M total funding seems to be the sweet spot. While there are likely industry-specific variations that require follow-on analysis, companies on track to raise much more than this likely have not found a clear path to profitability quickly enough, and companies that raise much left will likely not be able to scale at the pace of their competition.
* **Funding Timing**
  + Successful companies moved faster through funding rounds than the total average. A good strategy to leverage this would be targeting companies for series B or C investments that moved quicker than the average time from Seed -> A.
* **Overall Timing**
  + Startups appear to be reaching maturity quicker than ever. Speed is king, and companies demonstrating success are reaching IPO/Acquisition quicker than ever. To maximize the exit potential, focus on companies on track to IPO/be acquired within the first 7 years of existence.
* **People**
  + I have developed a comprehensive, yet non-exhaustive, list that details individuals with ties to successful startups. For the sake of this project, this functions as a good list of individuals to work with as advisors, or try to bring on as employees, or otherwise learn from
* **Investors**
  + The top VC firms appear to have the highest counts of successful exits. Further research is needed to determine if this is due to the sheer quantity of total investments, or if they are making better investments than other firms.

**SECTION 5 – CONCLUSION:**

While there is much to be taken away from this in terms of correlation of certain KPIs to success (see above results), the main takeaway for me is all of the avenues for continued analysis. As I investigated the industry trends, I realized that all of the factors I analyzed could be re-analyzed on a per-industry basis to reveal a bit more nuance about how each industry operates. The VC investors data could be further analyzed to determine a ‘win rate’ for each VC, and rank the top VCs that way.

My main desire for additional information is to find information about profitability and tie it into the analysis. In an era where so many startups are raising vast amounts of capital without turning a profit, I would like to further investigate how profitable the top startups inn this list were, how the speed at which capital was raised differed for profitable companies, and which industries saw the most profitable companies.

All in all this was a very intriguing project to work on, and one that allowed me to practice a lot of the data anlysis skills I’ve learned in the past few months of my data analytics bootcamp. Getting to use Python to clean and process data and set up a SQL database, PostgreSQL to run queries against that database, and Tableau to visualize the data I uncovered really gave me a good feel for what it feels like to work with a full analytical tech stack.

**WORKS CITED**

Data - <https://www.kaggle.com/datasets/justinas/startup-investments>