**Springboard Foundations of Data Science**

**Capstone Project "Investigating Returns on Seed-Stage Capital in the US"**
Milestone Report
August 2, 2017

*Background*

Angel Investing is the business of providing capital to early-stage growth companies. The capital is typically the first outside money the company takes; they offer in exchange some portion of their company's equity. The company uses investor's money to fund the growth of their business. As the company grows and increases proportionately in value, additional rounds of funding may be raised, fueling additional growth. Ultimately, investors (whether Seed- or Later-stage) will crystallize their investment and generate returns via an exit. An exit results from the sale of the company to a strategic (larger, corporate) buyer or via IPO. Companies which do not proceed to exit will typically either cease operations/close, in which case investor's money is lost and return is zero, or continue to operate, where a non-zero value is achieved but the investment remains illiquid and real value unrealized.

Unlike other asset classes, there is a dearth of information around outcomes (% success/failure) and returns ($ invested, $ returned) in this space. Information that is available is largely self-reported creating a potential source of bias/positive skew. Two companies (Pitchbook and Crunchbase) collect and make available (fee basis) proprietary subsets of data. Crunchbase in 2014 published for the public a sample set of their then-current (2014) data.

*Introduction*

My study seeks to use this Crunchbase data set to investigate returns in this space. The data set includes four separate tables which describe: funding rounds, companies, acquisitions and people (investors). For my purposes, acquisitions (specifically the "acquisition price" field which describes the value of the exit, which becomes the numerator in the returns equation) and funding rounds (specifically total funding raised, returns denominator) will be the most useful.

*Core Analysis & Objectives*

These are the questions that would be most useful to intended audience and which preliminary EDA suggests will be possible to resolve. Rely on core data-analytics skills. Italics indicate either methodology, or items to needing follow-up.

1. Quantifying data scarcity: of the companies listed in this set, what proportion are missing data critical to understanding outcomes (status field from companies database) and among those how many are missing data critical to understanding returns (acq'n price/total funding). *There may also be the opportunity to put the dataset itself in the context of broader industry activity. Meaning: how much of the activity in this space (as recorded by economics bureau or elsewhere) is being captured?*
2. Quantifying proportion of positive (exit) and negative (company folds) outcomes.

3. What do (dollar on dollar) returns look like for those companies that are fully described (outcome, funding, price) in this set? Is this a sufficiently large sample from which to draw conclusions about the population?
4. Generally accepted wisdom in the industry suggests that 20 – 25% annual IRR is typical for angel portfolios. Are these figures reasonably in line with the data in our set? *Methodology: Populate missing prices fields for "acquired" companies, in a manner that achieves the 20 – 25%. Is the set of companies in the "inferred pricing" bucket statistically similar to those in the "actual pricing" bucket? Can one assume these were drawn from the same population?*
5. Resolved outcomes (exit/closure) are by their nature bimodal. What does this mean for portfolio construction? As an investor adds more companies to their angel portfolio, results (IRR) will approach industry averages. What size (# of investments) of portfolio reduces the probability of a left-tail outcome (low or zero return) to acceptable levels?

*Blue-sky Opportunities & Outcomes*

This is the really cool stuff that uses more advanced tools and programming knowledge (so there will be some more skill-building there) and may be just beyond my individual capabilities. Report will be good without any of these and fantastic will all of them..

1. What does k-means clustering tell us about this data?
2. Likewise, are there interesting conclusions to be drawn by examining the data within a graph database context (neo4j)?
3. Build a Sankey Diagram to depict the progression of companies through various stages of funding and ultimately to exit or closure, over time.
4. Machine Learning Stuff

*Cleaning & Limitations*

There are some formatting issues with the raw data: specifically dates and currency fields.

Several of the fields are sparsely populated. After screening for rows (companies) fully populated across the fields of interest with the right characteristics (US based, first funding between 2007 – 2014, first funding of the "seed" or "angel" variety in quantities greater than $100k) our data set is only 100 rows in size.

Some of the data is just wrong, and while it's not practical to verify line-by-line, those cases which appear as clear outliers in the EDA are worthy of verification and correction where necessary.

Because this data is self-reported, biases are inherent. We lack to tools to measure and correct for these biases.

*Audience/Final Report Format*

This study is intended for use by participants in the angel ecosystem. A familiarity with industry-specific language (funding rounds, exits, IRR, portfolio construction) although a great deal of room exists to present numbers that will be novel and interesting to this same audience. The report will seek to educate & inform.