The first milestone of this project will be to select the data you want to work with. You will need to select 3 different data sources that have different file types of information – and the data will need to have a relationship between them. If one doesn’t exist, you will have to create one. It is likely you will need to manipulate the data to create a relationship. Finding the data, you want to work with for this project, will likely be the hardest part of the project. You must have one of each of the following types of datasets – and you need a minimum of 1000 rows across all datasets. Each dataset should have a minimum of 10 columns/variables.

* CSV/Excel/PDF or another flat file source.
* Website you want to pull data from--you will want to identify a website that has data stored in a table, similar to the screenshot below.
* API you will pull data from.

Startup Venture Funding

* Raise capital through rounds of external funding
  + These funding rounds provide investors the opportunity to invest cash in a growing company in exchange for equity, or partial ownership of that company
* Outside investment
* Other types of funding rounds:
  + “Seed” funding or angel investor funding at the outset
    - Can be followed by Series A, B and C funding rounds as well as additional efforts to earn capital as well, if appropriate
* **What does a round mean for the prospects and direction of a company?**
* **How do the rounds of funding/early investment indicate the success of a startup … or even can they?**

How Funding Works

* Common for a company to begin with a seed round and continue with A,B and then C funding rounds
* Other side 🡪 potential investors
  + Need to gain something back from the investment 🡪 partial ownership (equity)
* Before any round of funding begins, analysts undertake a valuation of the company in question
  + Management, proven track record, market size and risk
* One of the key distinctions between funding rounds has to do with the valuation with the business, as well as its maturity level and growth prospects

Pre-Seed Funding

* Company’s founders are first getting their operations off the ground
* Founders, friends, family, etc.
* Investors have not started yet usually

Seed Funding

* First official equity funding stage
* First official money that a business venture or enterprise raises
* Some companies never extend beyond this round
* One of the most common types of investors participating in seed funding is a so-called “angel investor”
  + Tend to appreciate riskier ventures 🡪 expect an equity stake

Series A Funding

* Company may opt for Series A funding in order to further optimize its user base and product offerings
* How will the company monetize its business?
* Companies are looking for great ideas and strong strategies
  + Common for firms going through Series A funding rounds to be valued at up to $23 million

Series B Funding

* Taking businesses past the development stage
* Expanding market reach
* Valuations tend to reflect the companies’ well-establishment
* Often led by many of the same characters as the earlier round, including a key anchor investor that helps to draw in other investors
* Difference with Series B 🡪 addition of a new wave of other venture capital firms that specialize in later-stage investing

Series C Funding

* Businesses are already quite successful
  + Companies look for additional funding in order to help them develop new products, expand into new markets, or even to acquire other companies
* Operation gets less risky
* Company will end its external equity funding with Series C
* Anticipation of an IPO

Evaluating entrepreneurial prospects

1. AirBnB
   1. Most popular time of the year for renting
   2. Location popularity
   3. Host popularity/review
   4. What is the demographic of the people renting AirBnBs?
      1. Could look based on price/size of place
   5. Could combine with a crime dataset for the neighborhoods to look at crime rates vs. rent rates
      1. <https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system>
   6. <https://calendarific.com/holidays/2021/US> --> use holidays table to determine which holidays are popular for renting
   7. Geographic API for getting reviews of the area of the AirBnB
      1. <https://www.programmableweb.com/api/foursquare>
      2. <https://www.programmableweb.com/api/geonames>
2. College Scorecard
   1. Dataset: <https://www.kaggle.com/wsj/college-salaries/version/1?select=salaries-by-region.csv>
   2. Website: <https://oedb.org/rankings/acceptance-rate/>

Table, calendar

Description automatically generated with medium confidence

* 1. API: <https://collegescorecard.ed.gov/data/documentation/>

First Milestone:

* 3 data sources, along with a description of each one
  + CSV file: <https://www.kaggle.com/wsj/college-salaries/version/1?select=salaries-by-region.csv>
    - Description: This CSV file was obtained from the Wall Street Journal and is based on data from Payscale, Inc. It represents universities by region and the post-graduate earnings for the students at these various universities.
  + Website (table starts at the bottom of the page): <https://oedb.org/rankings/acceptance-rate/>
    - The tabular layout in this website represents various numerical factors for universities in the United States. These factors include: student-to-faculty ratio, graduation rate, retention rate, acceptance rate, enrollment rate, school aid rate and default rate.
    - API: : <https://collegescorecard.ed.gov/data/documentation/>
      * College Scorecard provides data at the institution-level and by field of study. It is a resource for prospective students to utilize for searching many degrees of information on colleges/universities and assessing their fit.
* The relationships between them, or the relationship you will make between them
  + All of these data sources have a relationship based on university/institution name!
  + CSV: School Name
  + Website: School Name
  + API: Institution Name
* What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words

The first thing I plan on doing for meeting the milestones is to merge the two ‘salaries’ CSV files found at the Kaggle dataset link, in order to increase the amount of fields/information for the data source. I will join them all on School Name, and it will help the data also be in a more readable format.

In terms of further cleaning and formatting the data sources, I will be checking for missing and null values. Depending on the number of these values and their business relevance, I will either remove them from the sources or choose one of the methods for filling in the missing data, such as with the overall mean. In the College Scorecard API, there is some missing information due to the privacy for the university being upheld. These values will have to be accounted for, and some columns may have to removed depending on the number of records remaining.

Furthermore, in terms of data cleansing, I will be replacing the headers, mostly for the College Scorecard API, to clarify the actual definition and meaning for the fields since some are abbreviated and not clear in representation. I will use the metadata to better define the fields.

My interpretation of what the data means is that the university that one attends can have an impact on their career salary post-graduation. One’s education is usually considered when applying for jobs, and the prestigiousness or reputation of that education could determine how much one makes. I’d also like to assess the reputation of the universities in terms of retention rates, graduation rates, etc. Even if a university is highly ranked, it may still not have an impact on the students’ post-graduation career if the university cannot retain or get students to their graduation days (although other factors must be considered for this as well).

* Data dictionaries:
  + API: <https://collegescorecard.ed.gov/data/documentation/>
  + Website variables:
    - Student-to-faculty ratio: **the number of students who attend a school or university divided by the number of teachers in the institution**
    - Graduation Rate: this rate indicates how many students finish their degrees in a timely manner upon enrolling
    - Retention Rate: the percentage of a school’s first-time, first-year undergraduate students who continue at that school the next year
    - Acceptance Rate: the percentage of applicants who are admitted
    - Enrollment Rate: expressed as net enrollment rates, which are calculated by dividing the number of students of a particular age group enrolled in all levels of education by the size of the population of that age group
    - School Aid Rate:
    - Default Rate: the percentage of college attendees that default on their student loans
  + CSV variables:
    - School Name
    - School Type
    - Region
    - Starting Median Salary
    - Mid-Career Median Salary
    - Mid-Career 10th Percentile Salary
    - Mid-Career 25th Percentile Salary
    - Mid-Career 75th Percentile Salary
    - Mid-Career 90th Percentile Salary