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!
    - The CSV file has a 1:1 relationship with the API, and the website also has a 1:1 relationship with the API, as they contain one record per university and the API does as well.
  + 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