Katy Waterman
STAT 3870 - P3
9/30/25

Part 1: Your Scenario (20 points)

Main Objective: Collect U.S. education enrollment data to analyze grade-level patterns across multiple states over time. The goal is to compare enrollment trends in different regions and identify shifts in grade distribution between 2017–2020.
Data Sources: Urban Institute Education Data Portal
Data Types: Year, Grade Level, State, Enrollment Counts, School Level
Geographic Scope: Three states for comparison: California (FIPS 6), New York (FIPS 36), and Texas (FIPS 48)
Time Range: Academic years 2017-2020

Part 2: Learning about API (15 points)

Before this project, I had very limited knowledge about APIs. Working through the exercises really helped me understand how APIs work in practice, for example how you can make a request and get structured data back instantly. Using R, I learned how to send GET requests, deal with JSON data, and write the results into files I could reuse. It felt a bit intimidating at first, but once I got the hang of the code and saw the data coming in, it was really interesting. I also learned how important it is to handle errors and be respectful of API limits so I don't accidentally overload someone's server. Overall, it was a super practical intro to working with live data, and I can definitely see myself using APIs more in this project.

Part 3: Setting Up Free API Access (10 points)

For this project, I used the Urban Institute Education Data API, which is fully open and does not require any authentication or API key. This made setup straightforward and accessible for first-time users like me.

However, if I were using an API that did require authentication, such as the OpenWeatherMap API or NewsAPI, I would follow proper security best practices to keep my key safe.

Part 4: Build Your AI Data Collection Agent (35 points)
Screenshots of your agent running

```
Fetching year: 2018 grade: 3 state (FIPS): 36
No encoding supplied: defaulting to UTF-8.
Fetching year: 2018 grade: 3 state (FIPS): 48
No encoding supplied: defaulting to UTF-8.
Fetching year: 2018 grade: 8 state (FIPS): 6
No encoding supplied: defaulting to UTF-8.
Fetching year: 2018 grade: 8 state (FIPS): 36
```

```
Collection Summary:
> cat("Total records:", nrow(edu_data), "\n")
Total records: 101433
> cat("Success rate:", round(collection_stats$successful_requests / collec
tion_stats$total_requests, 2), "\n")
Success rate: 1
> cat("Quality score:", quality_report$summary$overall_quality_score,
"\n")
Quality score: 1
```

Part 5:Documentation (20 pts)

Quality assessment report and Collection Summary:

Total number of records: 101,433
Collection success rate: 100% (no failed API calls)
Quality Score / Completeness:

- All major fields (year, ncessch, grade, sex, race, enrollment) show >99.9% completeness
- No negative or obviously invalid values observed

quality_report	list [2]	List of length 2
summary	list [3]	List of length 3
total_records	integer [1]	101433
collection_success...	double [1]	1
overall_quality_score	double [1]	1
completeness	double [9]	1 1 1 1 1 1 ...
year	double [1]	1
ncessch	double [1]	1
ncessch_num	double [1]	1
grade	double [1]	1
race	double [1]	1
sex	double [1]	1
enrollment	double [1]	0.9993296
fips	double [1]	1
leaid	double [1]	1

Collection Summary:
Total data points collected: 101,433

Years Covered: 2017–2020
States Covered: California, New York, Texas
Grades Covered: 3 and 8
Success Rate: 100% - All 24 API calls returned data
Quality Metrics:

- Completeness score per variable: 99.9% or higher
- Data format consistent across all years


```
{
  "summary": {
    "total_records": [101433],
    "completeness": [1, 1, 1, 1, 1, 1, 0.9993, 1, 1]
  }
}
```


Challenges:

- API queries for large datasets can be slow
- Requires delays to avoid overloading the server

Recommendations:

- Expand to additional states and grades
- Combine with IPEDS datasets for higher education insights