Emily Fanning, El Auria Atienza, Alex Hafley
SI 206 Final Report

Github repository:

https://github.com/elatien/Final-Project

Project Goals

The goal of this project was to explore how socioeconomic indicators relate to restaurant quality and visibility on Yelp.

- **Yelp API**: to collect restaurant names, ratings, and review counts
- **Census API**: to gather median income per ZIP code
- **City-Data.com (scraped with BeautifulSoup)**: to get educational attainment data per ZIP

We wanted to store this data in a SQLite database and use it to analyze trends and correlations across different ZIP codes in Michigan.

Goals Achieved

- **Yelp API**: collected 100 restaurants across multiple ZIPs
- **Census API**: got median household income for 100 ZIPs
- **Web scraping via BeautifulSoup**: simulated education data per ZIP

We built a database with 3 tables: income, education, restaurants.

We calculated and visualized relationships between education, income, and restaurant ratings/reviews across Michigan ZIP codes.

Problems Faced

- We initially received errors due to mismatched column names in SQL joins.

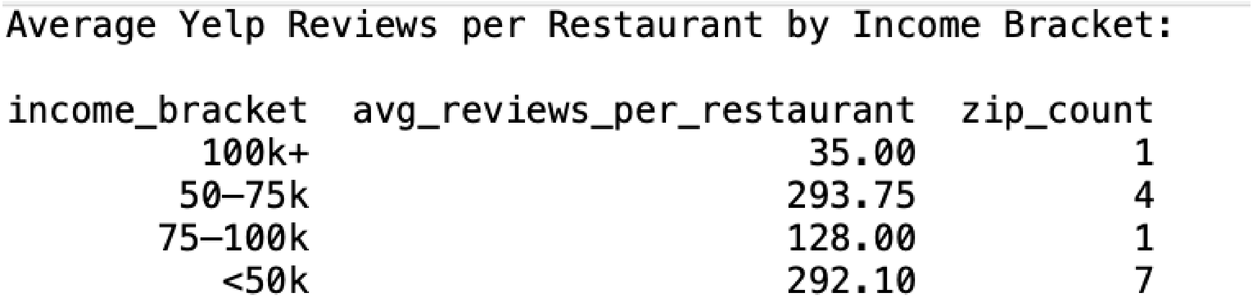
- Yelp API limited us to 25 results per request; we solved this by batching ZIPs and rerunning the script multiple times.

- We had to break data collection into chunks to meet the requirement of 25 inserts per run.

- GitHub push issues occurred due to mismatched remote names and pre-initialized repos.

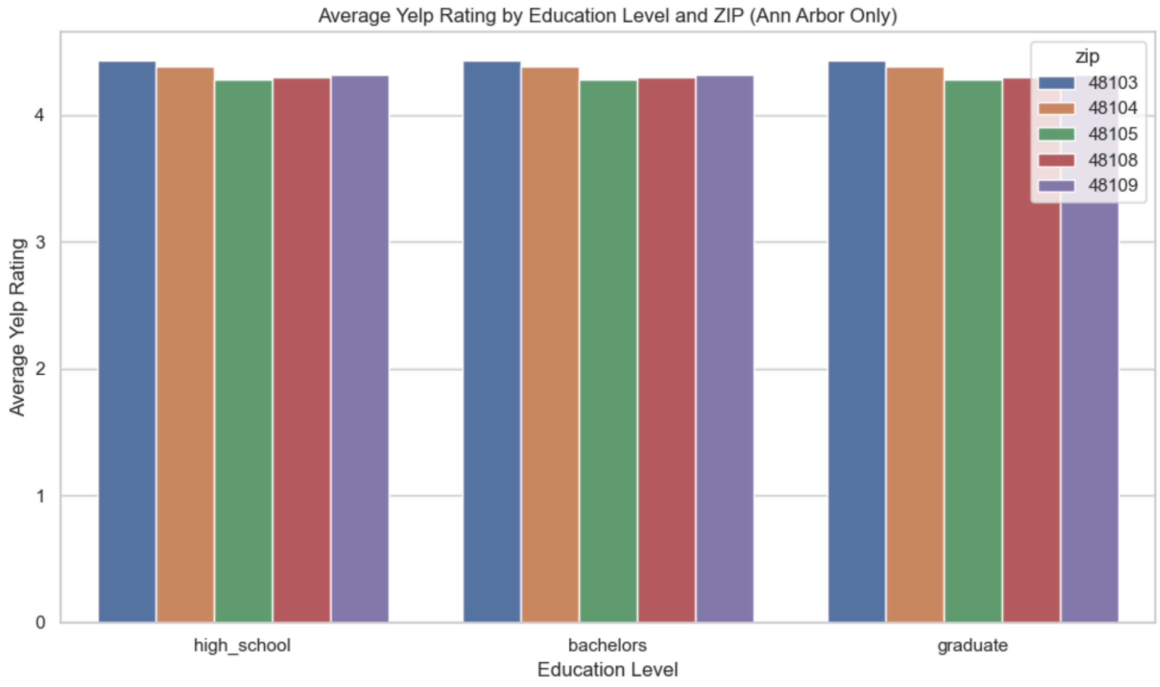
Database Calculations

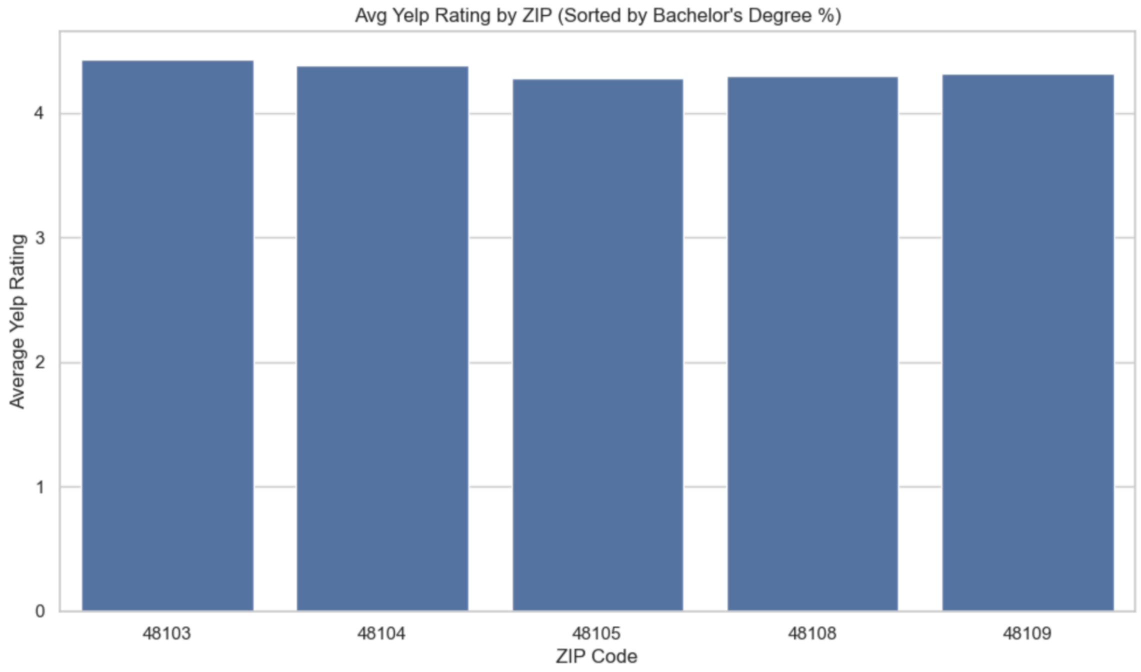
See attached screenshot of calculated_data.txt.



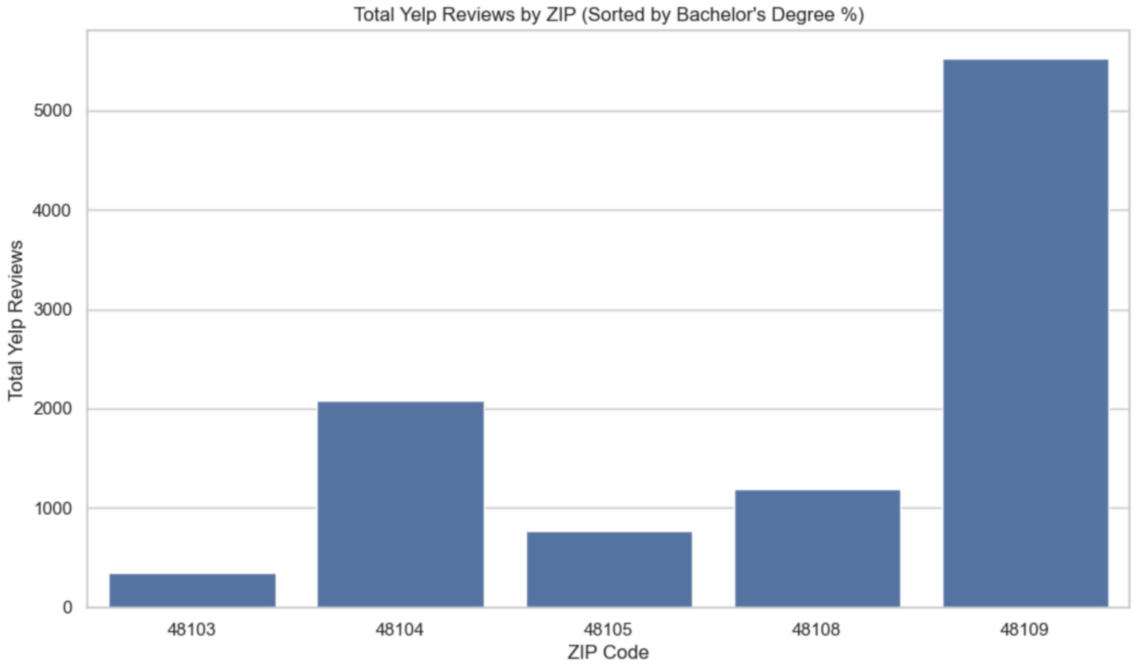
```
Average Yelp Reviews per Restaurant by Income Bracket:

income_bracket  avg_reviews_per_restaurant  zip_count
         100k+                       35.00          1
        50–75k                      293.75          4
       75–100k                      128.00          1
          <50k                      292.10          7
```

Visualizations

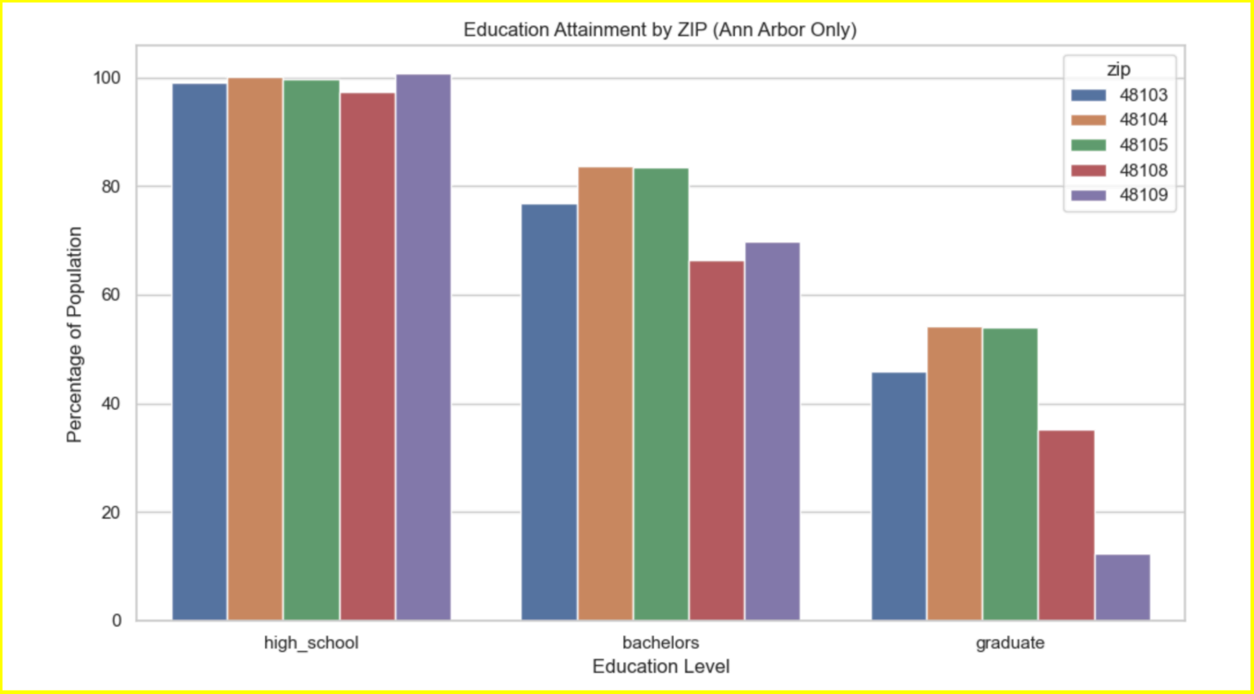


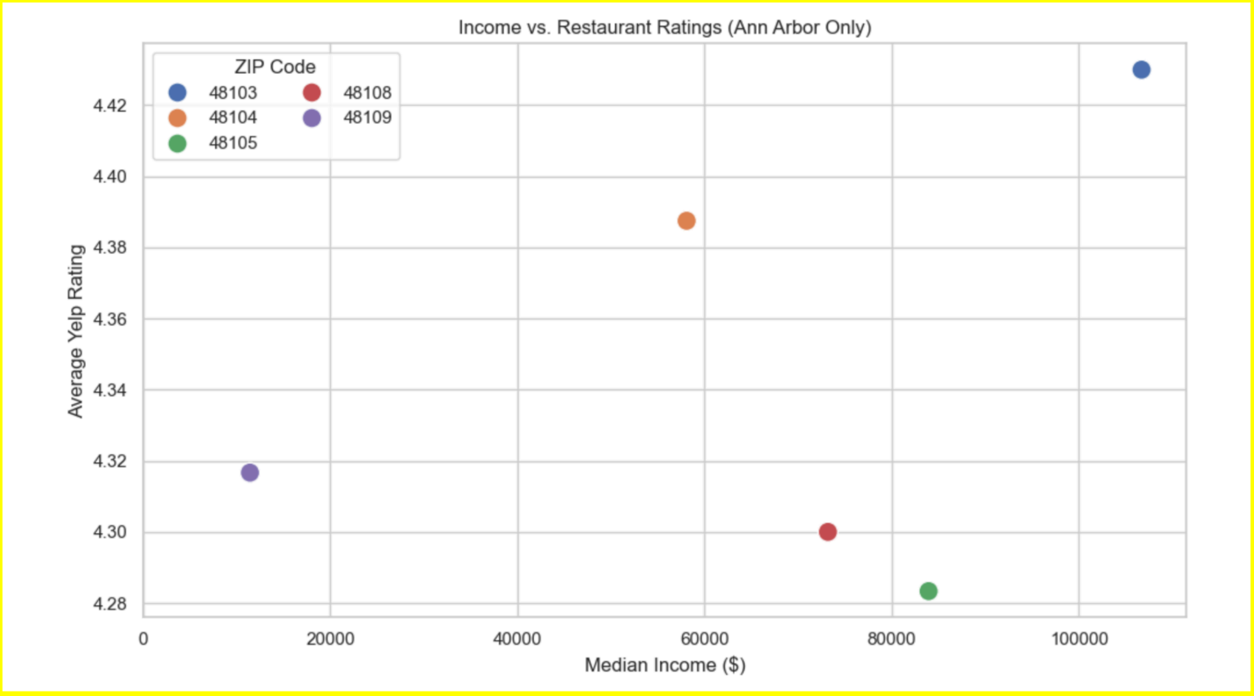


Avg Yelp Rating by ZIP (Sorted by Bachelor's Degree %)



Total Yelp Reviews by ZIP (Sorted by Bachelor's Degree %)





Instructions for Running Code

1. Open Terminal or Anaconda Prompt

Navigate to the project folder:

```
cd ~/Desktop/SI206/FINALPROJECT 5
```

2. Run the data gathering scripts in batches:

```
db_setup.py (to create the tables)

python census.py

python education.py

python yelp.py
```

3. (Re-run each 4 times for 100 records)

To run calculation and create visualizations:

```
python analyze.py
```

Function Documentation

File	Function Name	Description
db_setup.py	initialize_db()	Creates a new SQLite database (FINALPROJECTDB.db) and initializes the tables: income, education, and restaurants, with appropriate schema and foreign keys. Deletes any existing database file before creating a fresh one.
census.py	fetch_income_data()	Fetches median income data from the Census API for Michigan ZIP codes. Inserts up to 25 new records per run into the income table.
education.py	scrape_education_stats()	Scrapes education data from City-Data.com for ZIP codes (25 at a time), extracting high school, bachelor's, and graduate degree attainment percentages. Inserts into the education table.

yelp.py	fetch_yelp_data()	Uses the Yelp Fusion API to fetch restaurant names, ratings, and review counts for each ZIP code. Limits to 25 results per ZIP and stores in the restaurants table.

Resource Log

Date	Issue Description	Location of Resource	Result
4/08/2025	Needed Census API key setup	https://api.census.gov/data/key_signup.html	Got working API access
4/09/2025	Git remote push failing due to mismatch	ChatGPT	Remote updated & pushed
4/10/2025	SQL JOIN failing due to missing columns	PRAGMA table_info() used to debug	Query corrected
4/10/2025	Git rejecting push due to diverged history	ChatGPT help: --allow-unrelated-histories	Pull resolved
4/17/2025	Fixed presentation demo bugs	ChatGPT and GSI	Debigged and fixed databases. Code runs properly

Project Updates

These are the updates we made after our demo presentation:

- We debugged our data processing pipeline to make sure that all entries are unique and capped the output to 25 records per run. This change improves both efficiency and readability for testing and final outputs.

- To make things clearer, we simplified the visuals by focusing just on Ann Arbor zip codes. We also simplified the legends and cleaned up the x-axes for all of the charts and graphs to make the visual information more accessible and user-friendly.

- We replaced placeholder education data with a proper scrape from City-Data, which improved the reliability of our dataset. The education table was also updated to show the correct information.