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. We planned to use:

- **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 aimed to store this data in a unified SQLite database and use it to analyze trends and correlations across different ZIP codes in Michigan.

Goals Achieved

We successfully worked with:

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

We built a normalized database with four tables: income, education, restaurants, and a test set education_data_48103.

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 which includes:

- Total reviews per ZIP

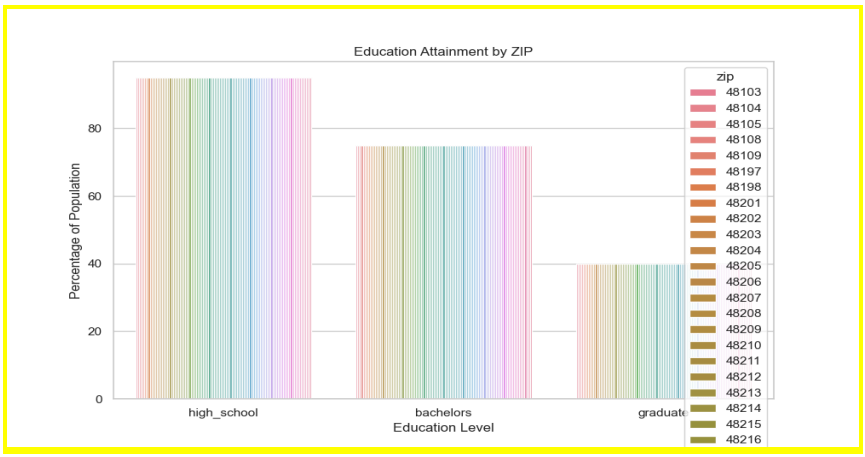
- Restaurant count

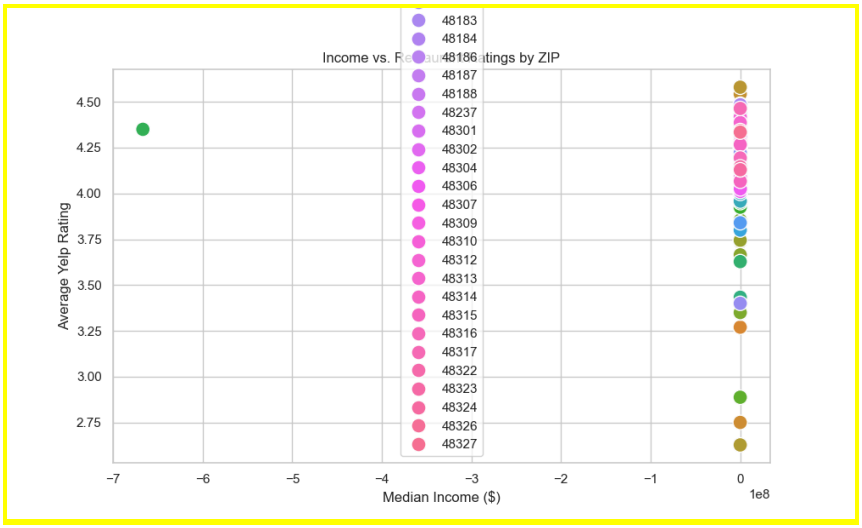
- Average reviews per restaurant

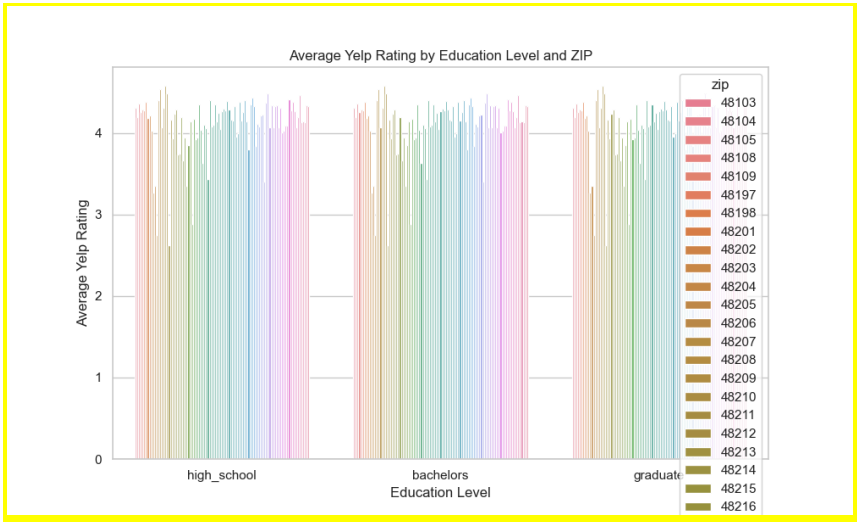
- Income bracket-based analysis

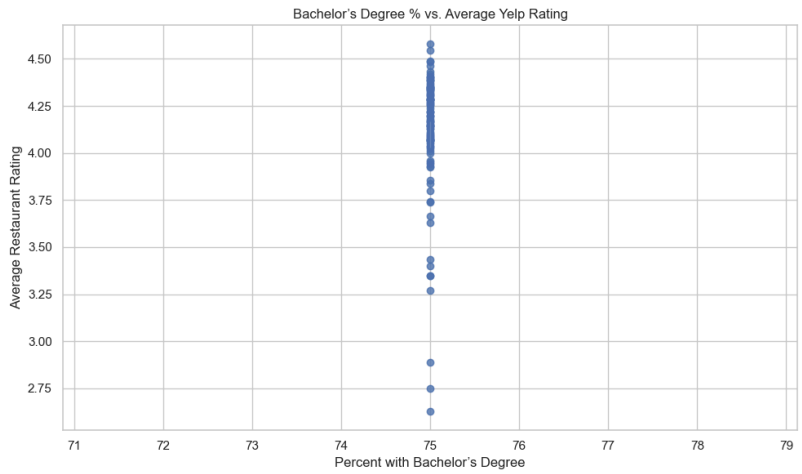
```
Average Yelp Reviews per Restaurant by Income Bracket:

income_bracket  avg_reviews_per_restaurant zip_count
          100k+                    189.95         20
         50–75k                    145.33         28
        75–100k                    196.67         16
           <50k                    135.39         35
```

Visualizations


Education Attainment by ZIP
Percentage of Population
0
20
40
60
80
high_school
bachelors
graduate
Education Level
zip
48103
48104
48105
48108
48109
48197
48198
48201
48202
48203
48204
48205
48206
48207
48208
48209
48210
48211
48212
48213
48214
48215
48216


48183
48184
48186
48187
48188
48237
48301
48302
48304
48306
48307
48309
48310
48312
48313
48314
48315
48316
48317
48322
48323
48324
48326
48327
Income vs. Restaurant Ratings by ZIP
Average Yelp Rating
2.75
3.00
3.25
3.50
3.75
4.00
4.25
4.50
−7
−6
−5
−4
−3
−2
−1
0
1e8
Median Income ($)


Average Yelp Rating by Education Level and ZIP
Average Yelp Rating
0
1
2
3
4
high_school
bachelors
graduate
Education Level
zip
48103
48104
48105
48108
48109
48197
48198
48201
48202
48203
48204
48205
48206
48207
48208
48209
48210
48211
48212
48213
48214
48215
48216



Instructions for Running Code

1. Open Terminal or Anaconda Prompt

   Navigate to the project folder:
   `cd ~/Desktop/SI206/FINALPROJECT`

2. Run the data gathering scripts in batches (Yelp, Census, Education):

   ```
   python census.py

   python education.py

   python yelp.py
   ```

3. (Re-run each 4–5 times for 100+ records)
   To run calculations and create visualizations:
   `python analyze.py`

Function Documentation

File	Function Name	Description
census.py	fetch_income_data()	Fetches income data for ZIPs using the Census API and stores it in SQLite.

education.py	scrape_education_stats()	Simulated scraping of education data and inserts it into the database.
yelp.py	fetch_yelp_data()	Pulls restaurant info from Yelp API and saves 25 businesses per ZIP.
analyze.py	pd.read_sql_query()	Uses SQL joins to combine data and generate calculated DataFrames.

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	GitHub Docs + 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