Why Do Some Restaurants Survive?
Insights from Yelp and Neighborhood Data in Philadelphia

Chris Low — Data Visualization, Project Check-In (Week 6)

Research Question

This project asks: *What factors explain why some Philadelphia restaurants stayed open while others closed?*

Restaurant survival isn't just about business success. It also reflects the strength of local communities. I'm using Yelp data together with neighborhood income information to see how factors like location, wealth, and customer engagement relate to whether restaurants remain open. At first, I thought wealthier ZIP codes would show higher survival rates, but the data tells a more mixed story. Some middle- and lower-income areas actually have more restaurants that stayed open, possibly because they depend more on local customers or have lower operating costs. This suggests that community engagement, things like frequent reviews, check-ins, and tips, might matter more than income alone.

Dataset

I combined three open datasets to address this question:

1. **Yelp Open Dataset** (business, check-in, and tip files): includes each restaurant's name, category, rating, review count, check-ins, compliments, and open/closed status. Available at Yelp Open Dataset.

2. **Census Income by ZIP** (RowZero CSV): reports median and mean household income for each Philadelphia ZIP code. Available at Row Zero.

3. **Philadelphia ZIP Code Boundaries** (OpenDataPhilly GeoJSON): provides polygon shapes for mapping neighborhood-level data. Available at Open Data Philly.

From Yelp, the main variables are the restaurant `name`, `business_id`, `categories`, `latitude`, `longitude`, average `star rating`, `review_count`, `checkin_count`, `tip_count`, average `compliments`, and the binary `is_open` field that indicates whether the restaurant is still operating. From the Census income data, the key variables are `median_income`, `mean_income`, and total `population` by ZIP code. The GeoJSON adds a `geometry` column used to map each ZIP code's boundaries. Together, these variables capture both local economic context and customer engagement, making the dataset well suited to study what predicts restaurant survival across neighborhoods.

Each restaurant record is merged with its ZIP-level income information. Aggregating to the ZIP level gives us clear visualization of spatial trends in restaurant survival and wealth.

Preliminary Findings and Visualization Value

The initial visualizations demonstrate the usefulness of this dataset (shown next page):

- A choropleth of *Share of Restaurants Still Open by ZIP* reveals clear neighborhood differences in survival rates.

- A choropleth of *Median Household Income by ZIP* highlights economic variation across the city.

- A scatterplot of *Income vs. Restaurant Survival* shows that wealthier ZIPs do not always have more surviving restaurants; several lower-income areas display high resilience.

These results suggest that restaurant survival depends on both economic context and customer engagement. In the next stage of the project, I plan to explore which specific factors are most closely linked to restaurant survival. I'll start by examining basic patterns in the Yelp data, like how ratings, review counts, and check-ins differ between restaurants that are open and those that are closed. Then I'll build a simple predictive model, likely a logistic regression or random forest, to see which features best explain whether a restaurant stays open. This will help test whether engagement metrics, reputation, or neighborhood income are the strongest signals of resilience. The results should give a clearer picture of what drives long-term survival in Philadelphia's restaurant scene.
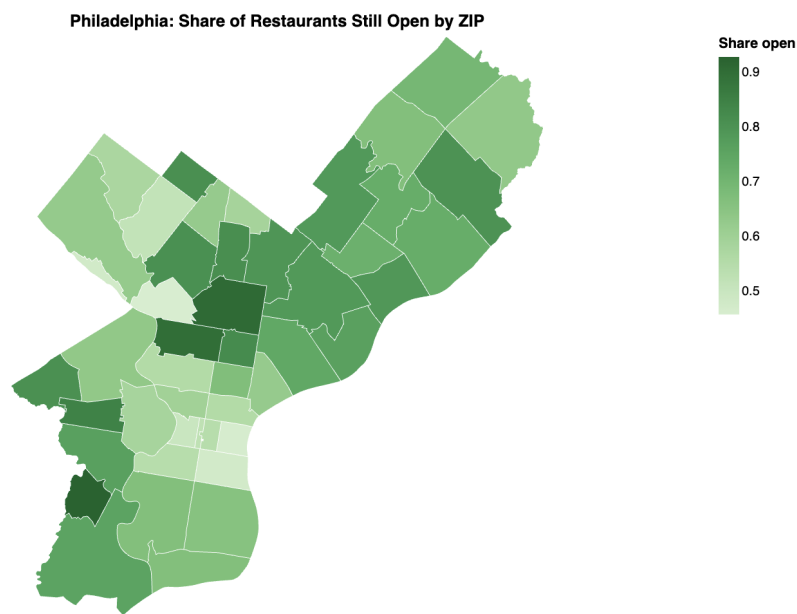

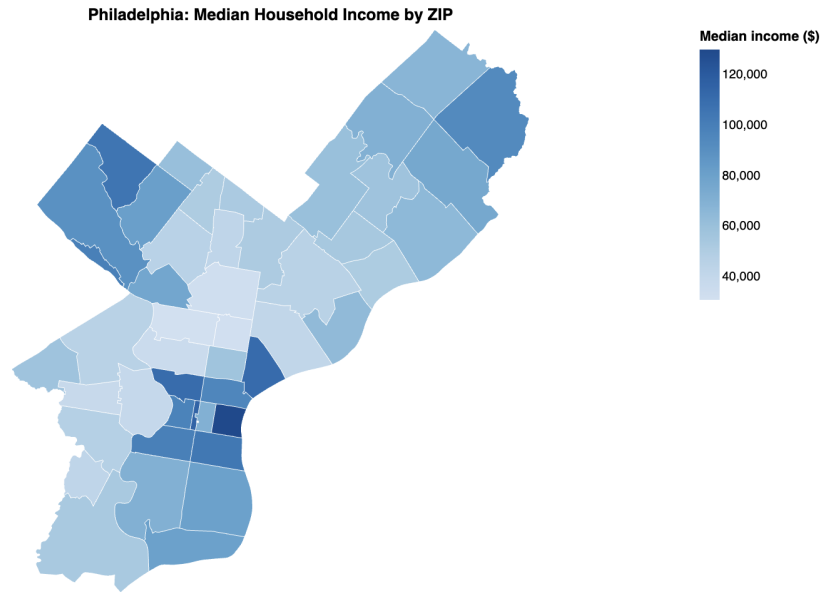Figure 1: Share of Restaurants Still Open by ZIP
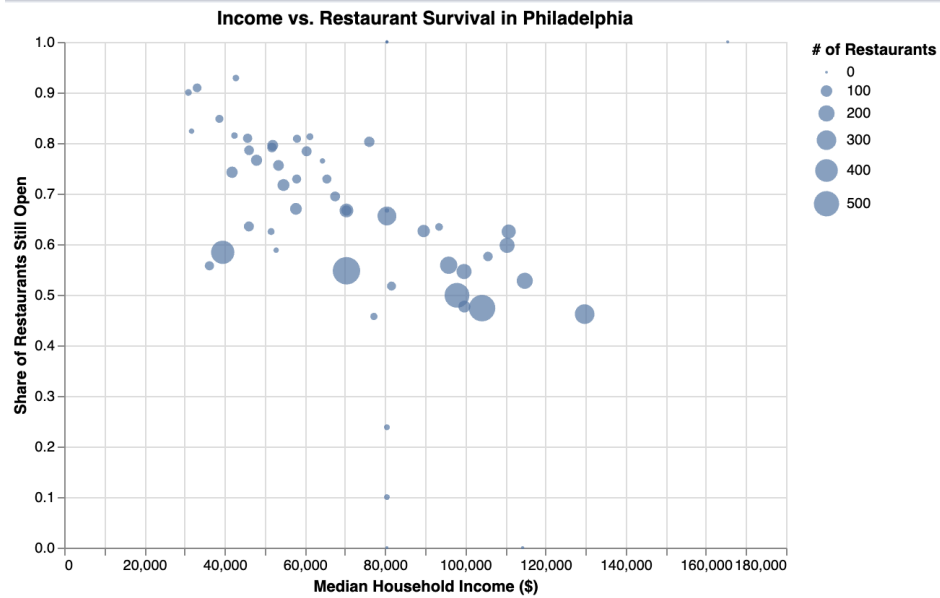
Figure 2: Median Household Income by ZIP

Figure 3: Scatterplot of Income vs Restaurant Survival