2.19-2.25 Weekly Record

Check whether `is_open=0` signifies temporary or permanent closure

1. **Yelp Dataset Documentation**

Yelp's official dataset documentation indicates that:

- `is_open = 1` → Open Business
- `is_open = 0` → Closed Business

However, it does **not specify** whether `0` means **permanently closed** or **temporarily closed**.

2. **Manual Check via Yelp Website (Combining with Google Maps)**

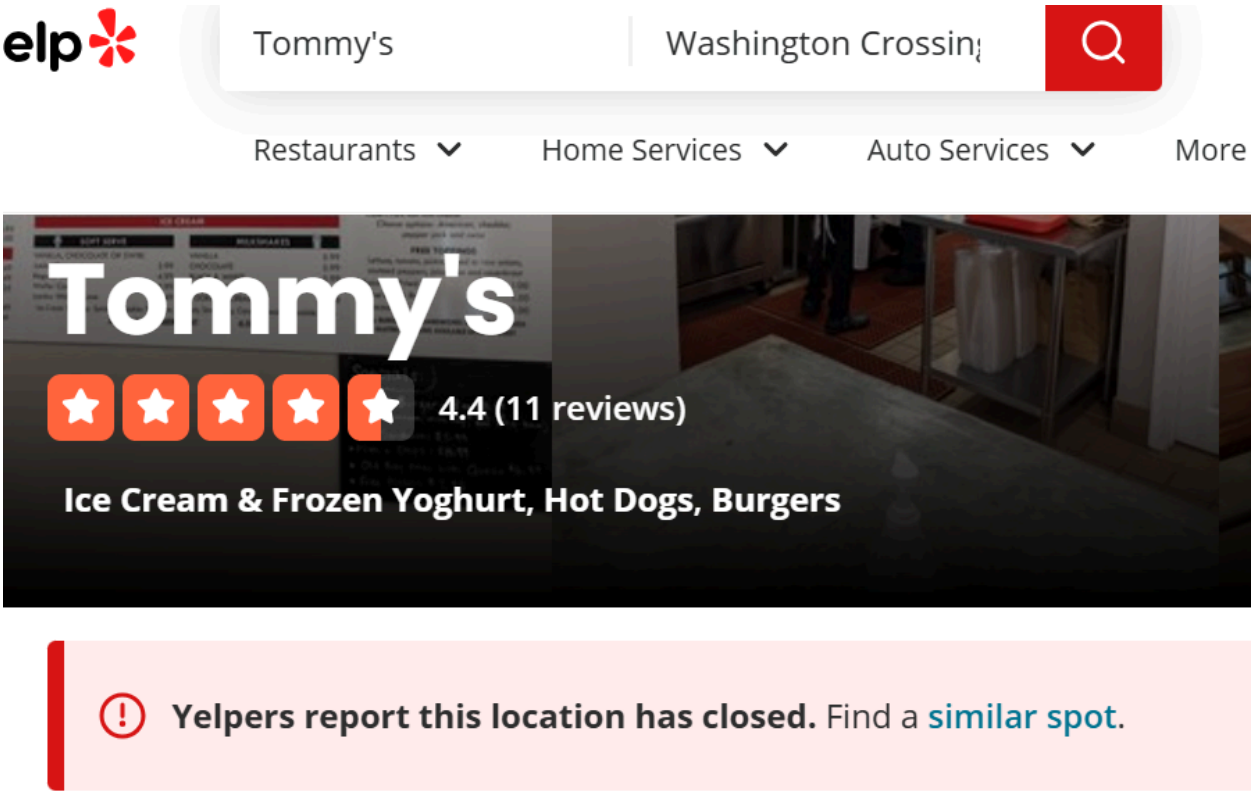
1) Randomly sample 10 `is_open=0` **businesses**

//0-no longer exist, 1-exist
Mubarak Shawarma - 225 S 45th St - Philadelphia // 0
Big Jar - 55 N 2nd St - Philadelphia //0
El Café - 31 S 19th St - Philadelphia
//1
The Jerk Pit At Chestnut Hill - 8221 Germantown Ave - Philadelphia //0
Santucci's Original Square Pizza - 4019 O St - Philadelphia
// 0, same name but not the same address
Asian Wok - 312 W Swedesford Rd - Berwyn //0
Milas Cafe - 8901 W Chester Pike - Upper Darby //0
Dunkin' - 1399 Skippack Pike - Blue Bell
// 0,chain store, but can't search the one at same location
Mazza Healthy Mediteranean - 1100 Jackson St - Philadelphia //0
Cafe Huong Lan - 1037 S 8th St - Philadelphia //0

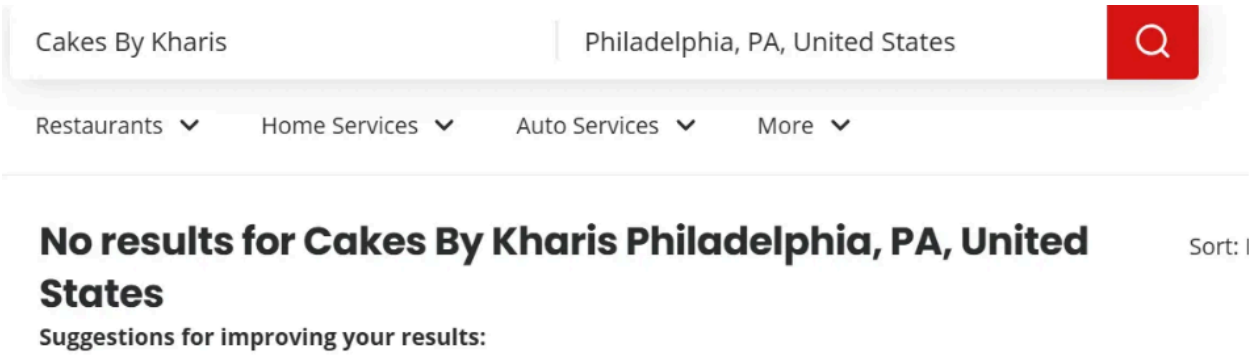
2) Randomly sample 10 `is_open=1` **businesses**

Wings To Go - 413 W Moreland Rd - Willow Grove //1
Bravo's Pizza and Family Restaurant - 212 W Walnut St - Perkasie //1

Smoked and Chopped - 3300 Fairmount Ave - Philadelphia //1
Jersey Mike's Subs - 4347 W Swamp Rd - Doylestown
//1, chain store, and find the one at same location
Pholosophy - 226 Haverford Ave - Narberth //1
Odd Logic Brewing Company - 500 Bristol Pike - Bristol //1
Jasper's Backyard - 101 E 7th Ave - Conshohocken //1
Enzo's Pizzeria - 1862 W Maple Ave - Langhorne //1
Vagrant Coffee - 4435 Baltimore Ave - Philadelphia //1
Tommy's - 1118 Taylorsville Rd - Washington Crossin // 0, find it, but "Closed" reported



This is different from the page when I searched `is_open=0` business



Maybe this is another evidence

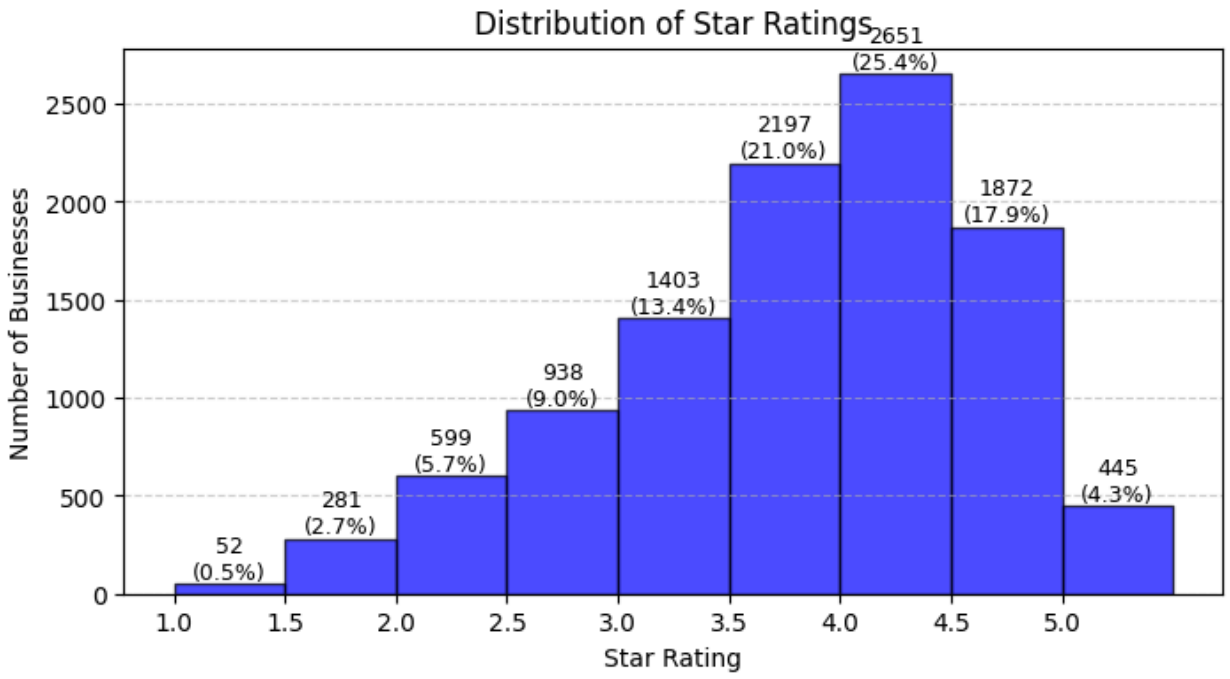
Try different thresholds for review counts and ratings, and note how many businesses are filtered out.

```
for threshold in [1, 5, 10, 15, 20, 50]:
    kept = sum(1 for x in review_counts if x >= threshold)
    print(f"Threshold {threshold}: {kept}/{len(review_counts)} businesses kept")
```

```
Threshold 1: 10438/10438 businesses kept
Threshold 5: 10438/10438 businesses kept
Threshold 10: 8637/10438 businesses kept
Threshold 15: 7464/10438 businesses kept
Threshold 20: 6639/10438 businesses kept
Threshold 50: 3987/10438 businesses kept
```

Review Count Percentiles:
10th percentile: 7.0
25th percentile: 13.0
50th percentile (median): 32.0
75th percentile: 85.0
90th percentile: 197.0

75% of businesses have reviews exceeding 13



Over 80%: above 3.0

Try different combinations: around 13 reviews + around 3 stars

```
Total open businesses: 10438
Removed businesses with review count < 10: 1801 (17.25%)
Removed businesses with rating < 2.5: 932 (8.93%)
Removed businesses that failed both conditions: 267 (2.56%)
Remaining businesses after filtering: 7972 (76.37%)
```

```
Total open businesses: 10438
Removed businesses with review count < 10: 1801 (17.25%)
Removed businesses with rating < 3: 1870 (17.92%)
Removed businesses that failed both conditions: 442 (4.23%)
Remaining businesses after filtering: 7209 (69.06%)
```

Each criterion filters out 10% to 20%

70% businesses remained

Removed:

is_open=0 + review_counts=0 + rating<2.5

↓

is_open=0 + review_counts<10 + rating<3

Utilize Categories

categories: useful for structured vectors

Extract all unique categories from `pa_filtered_dining_businesses.json` → 484 unique categories

✅ **Not All Categories Are Related to Food**

- Examples: Playgrounds, Gyms, Golf
- **Solution:** Remove **non-food-related** categories.

✅ **Some Categories Are Too General**

- `"Food"`, `"Restaurants"`, `"Specialty Food"`
- **Solution:** Remove overly broad terms and rely on **more specific categories** like `"Chinese"`, `"Italian"`.

✅ **Synonyms and Redundancies**

- Example: `"American (New)"` and `"American (Traditional)"` → **Merge to "American"**
- Example: `"Bubble Tea"` and `"Tea Rooms"` → **Consider merging or Create Hierarchy (?)**

LLM + Manual Check

Refined categories count: 484→ 208→ 167

Not yet finished, since there are still some synonyms and irrelevant categories through manual check

Create a Hierarchical or Clustered Category Representation (TODO)

// It is better to choose Clustering

- Instead of treating each category independently, group them into **higher-level clusters**:
  - **Cuisine Type**: *Mexican, Italian, Chinese, American, Indian, Thai, etc.*
  - **Service Style**: *Fast Food, Fine Dining, Buffet, Takeout, Delivery*.
  - **Dietary**: *Vegan, Gluten-Free, Halal*.

Suggestions about `attributes`

Check how attribute values are split within the dataset

Discard **attributes with low occurrence frequencies** to enhance dataset relevance. Given that there are over 7,200 businesses, attributes appearing in at least 5,000 instances can be considered more representative

Simplify overly complex attributes. For example, nested attributes like `BusinessParking` can be converted into a binary format

Next Step

Continue to investigate relationships between `categories`

Process `attributes`

Technical Approach