The Battle of Neighborhoods

Minshen Li

Background

- Local preferences are important factors in determining whether a business would be successful in this area
- Preference information is not directly available
  - Looking at what local businesses thrive could help
- If similar types of merchants thrive in two neighborhoods:
  - Local preferences (for food, service, activities, etc.) may well be similar between the two areas

Problem

- What neighborhoods are similar? How do we identify an area for the business?
  - Firm wants to find an area similar to the “successful area” but also with market not saturated
- The key is thus to pin down similar neighborhoods and identify what businesses thrive in those clusters

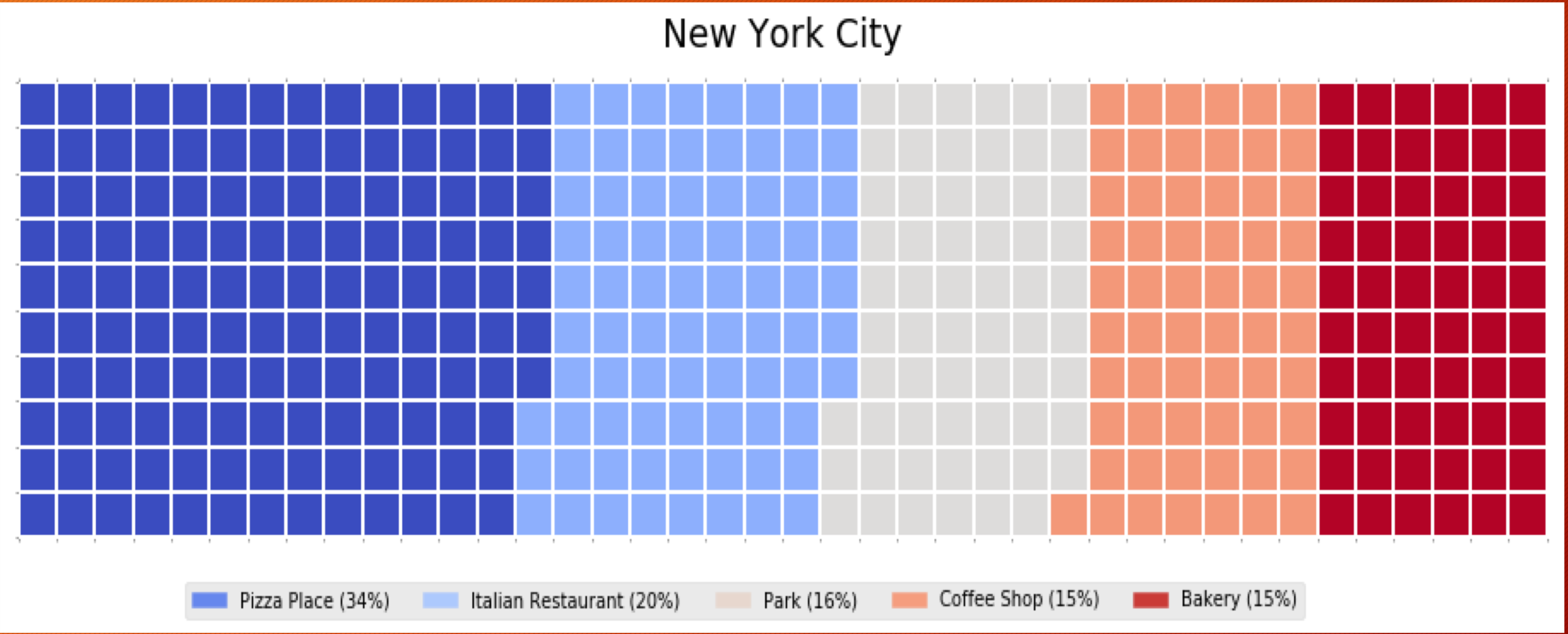
Approach

- In this case study, I analyze similarity of a total of 560 neighborhoods of New York City, Philadelphia, and Toronto:
  - Foursquare location data
    - Distribution of popular venue categories for each neighborhood
  - Cluster analysis
    - Obtain clusters of neighborhoods
  - Word clouds
    - Summarize top categories that thrive within each cluster
  - Rank similarity of clusters
    - Rank priority on the sequence of neighborhoods to be considered
  - Rank similarity of cities
    - Rank priority on the sequence of cities to be considered

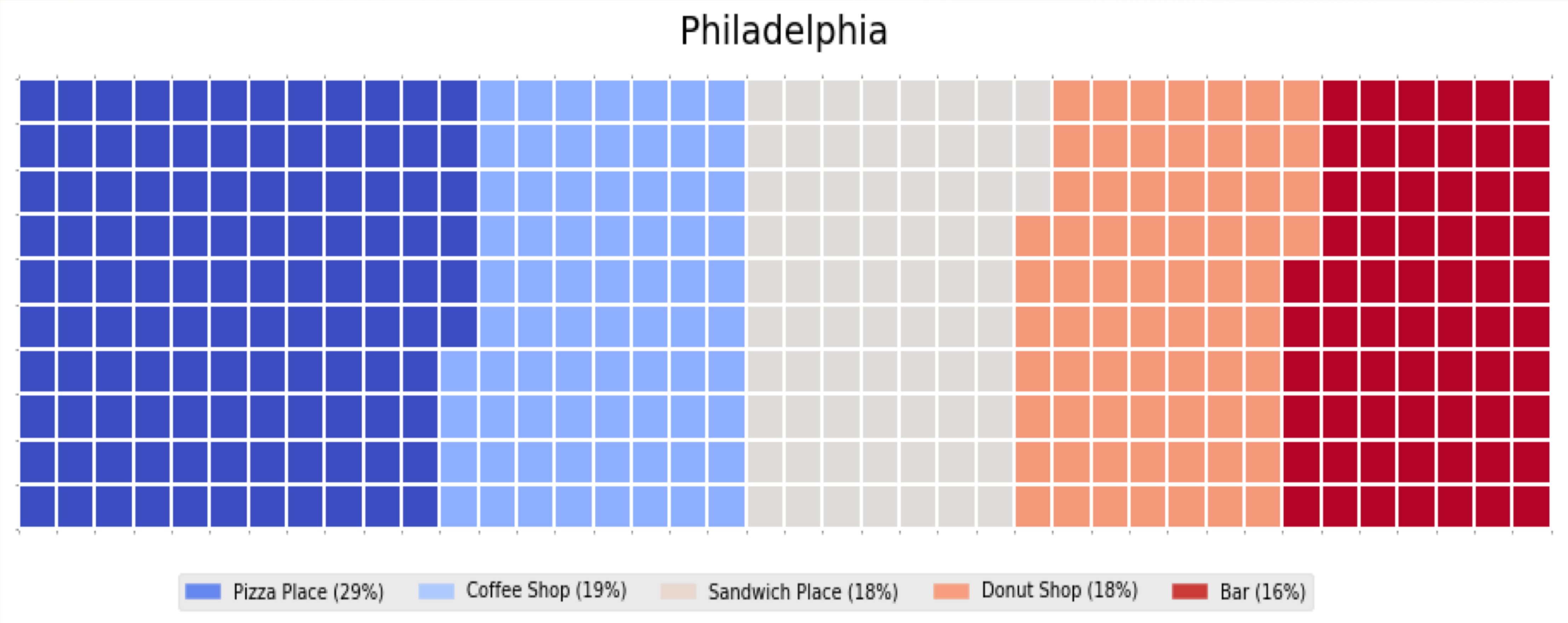
Data

- Neighborhood information: scraped from Wikipedia
  - # of neighborhoods:
    - New York City: 306
    - Philadelphia: 161
    - Toronto: 103
- Venue information: Foursquare API
  - Top 100 venues within a radius of 2km from the CenterPoint of each neighborhood.
    - # of popular venues
      - New York City: 28,304 (55.6%)
      - Philadelphia: 14,144 (27.8%)
      - Toronto: 8,446 (16.6%)
      - 479 unique popular categories

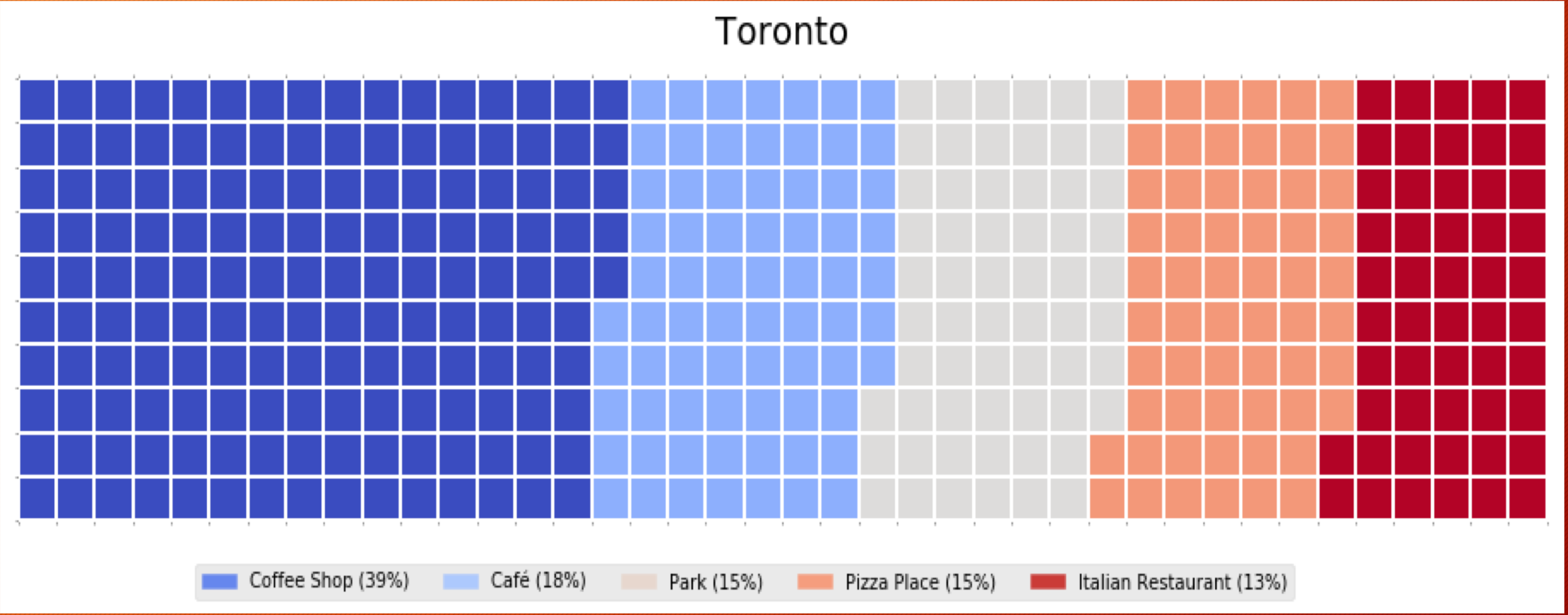
Data: Top 5 Categories in New York City



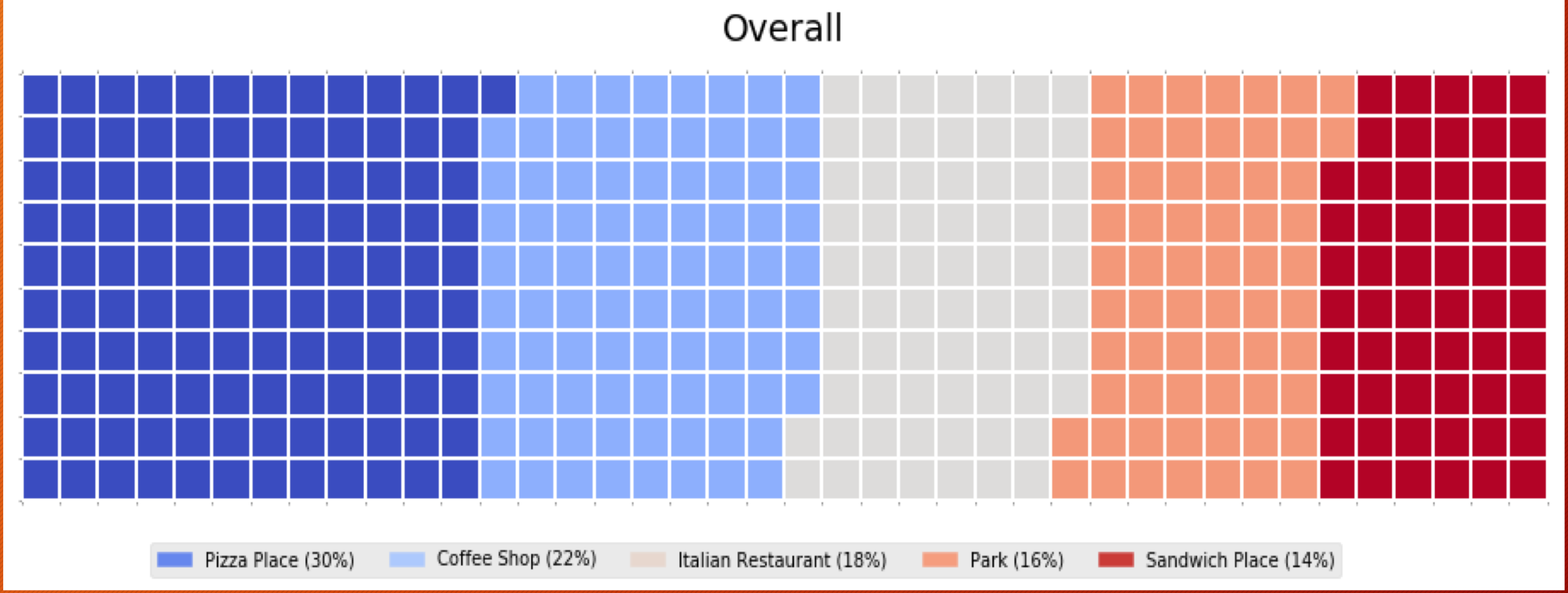
Data: Top 5 Categories in Philadelphia



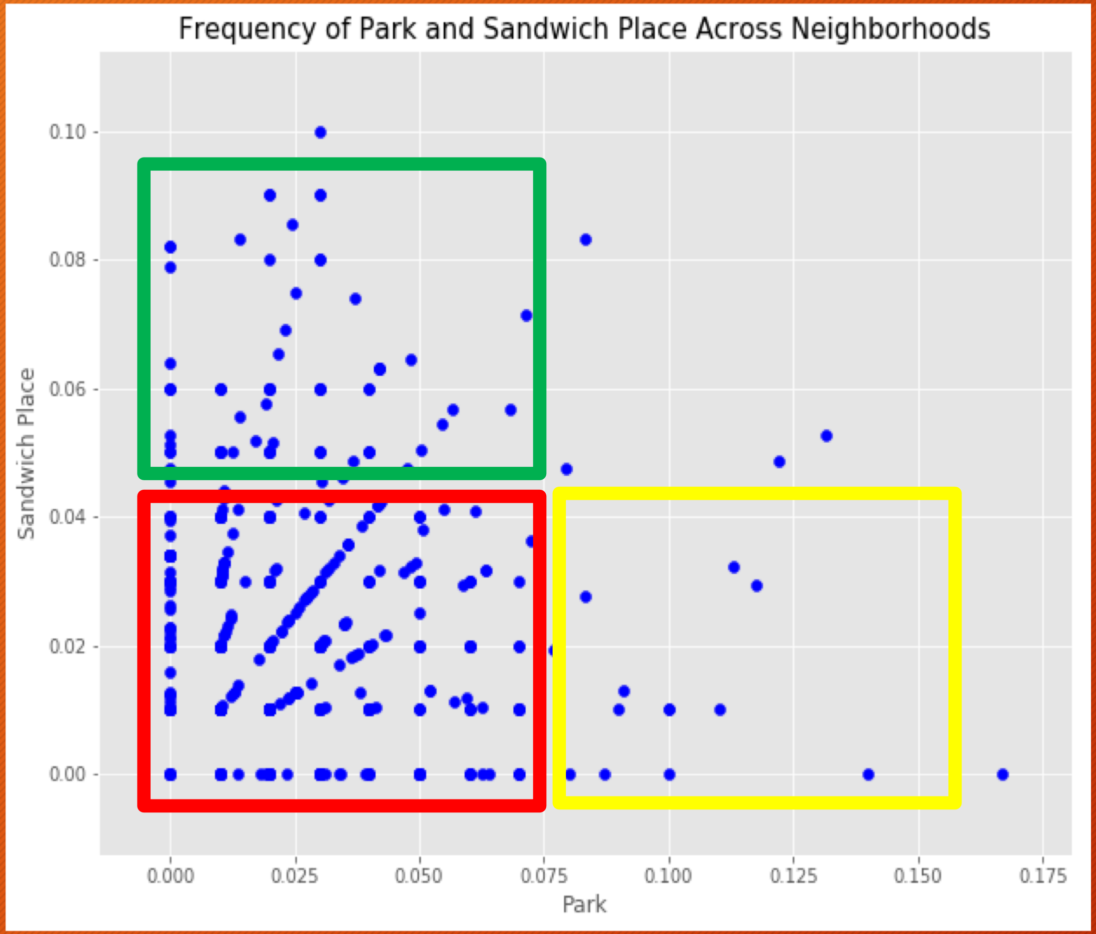
Data: Top 5 Categories in Toronto



Data: Top 5 Categories Overall



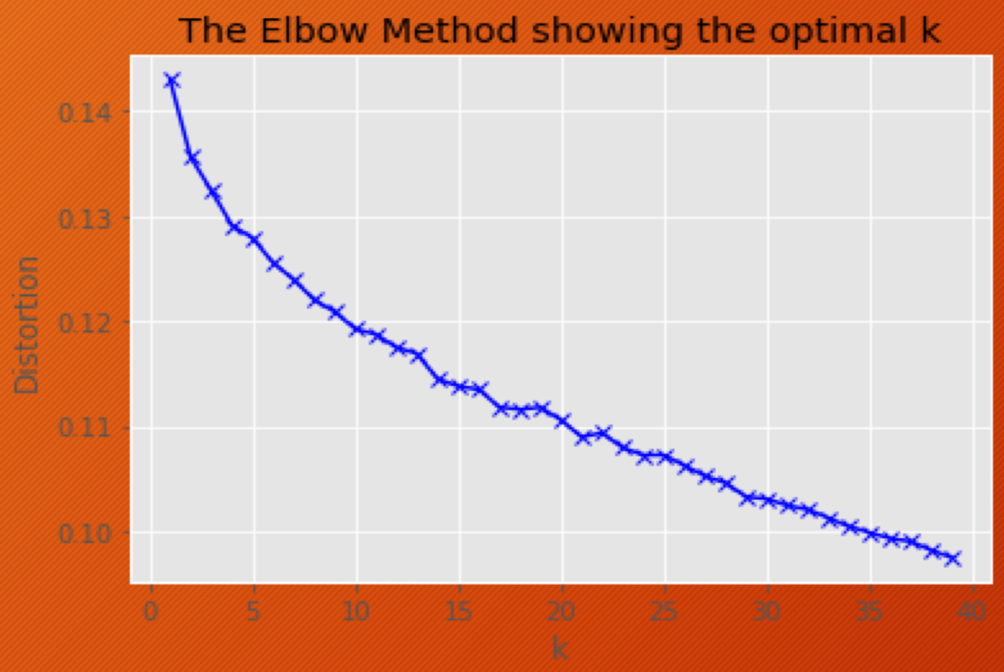
Data



- Neighborhoods are heterogeneous and clusters can be seen from the graph

Cluster Analysis

- Use K-means Cluster analysis
  - Pin down the optimal number of clusters using elbow method



- Choose k=4
  - Decreases in distortion for k>=4 are smaller than the decrease in distortion when k is increased from 3 to 4. Therefore, I go with four clusters.

Word Cloud Analysis

- Cluster 1:
  - Coffee shops, pizza places, gyms, bars, cafés, bakery shops, and Italian restaurants
- Cluster 2:
  - Pizza places, donut shops, and pharmacies
- Cluster 3:
  - Pizza places, sandwich places, and fast food restaurants
- Cluster 4:
  - Italian restaurants, pizza places, and sandwich places

- Example: To open a donut shop, check neighborhoods in Cluster 2 first

Cluster Similarity

Table 4. Euclidean Distance Among Clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0	0.126	0.113	0.083
Cluster 2		0	0.127	0.084
Cluster 3			0	0.118
Cluster 4				0

- The closest pairs of clusters include:
  - (Cluster 1, Cluster 4)
  - (Cluster 2, Cluster 4).
  - If Cluster 2 is saturated with donut shops, consider Cluster 4

City Similarity

- Obtain the distribution of clusters for each city

City Similarity

Table 5. Euclidean Distance Among Cities

	New York City	Philadelphia	Toronto
New York City	0	0.432	0.735
Philadelphia		0	0.741
Toronto			0

- Calculate Euclidean distance among cities
  - Philadelphia and New York City are more similar than to Toronto
  - Firms should consider switching between New York City and Philadelphia for business first

Summary and Future Directions

- The 570 neighborhoods of New York City, Philadelphia, Toronto can be clustered into four clusters:
  - Each cluster has hot categories
  - Firms should consider neighborhoods within the cluster first, then down the list of ranking of similarity
- Future directions:
  - Cluster analysis on the cities
  - more information, such as local employment rates, housing prices, etc., to generate more insightful cluster results
  - Better solution for finding the optimal k value