
# Content-based Recommender System for the Clusters of potential interest

## Based on OSM data for San Francisco

Mariia Sundeeva. Data Science Capstone. 3/1/23

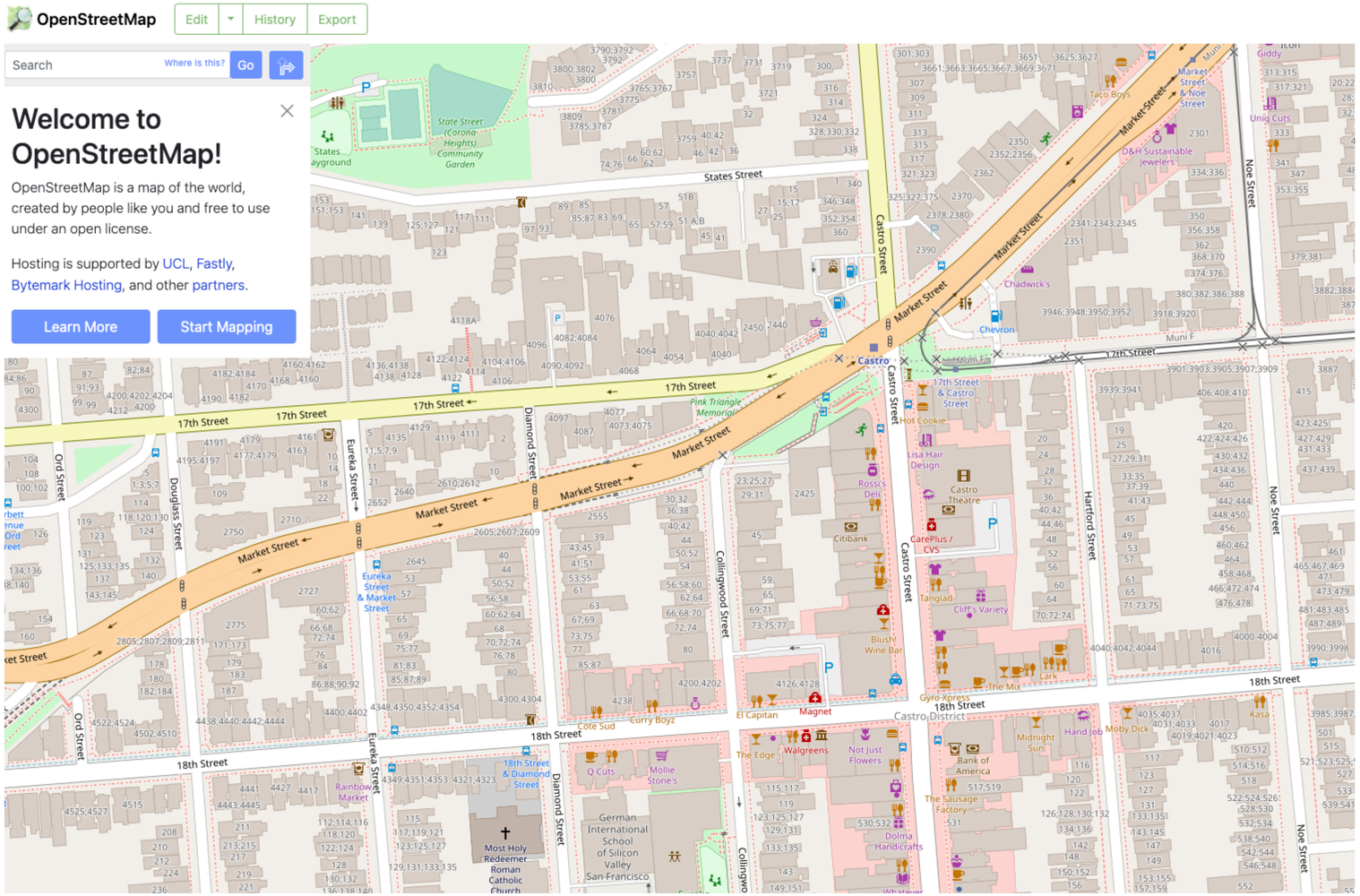
# How to measure area’s vibe?

## For people with non-mainstream preferences

- Cluster is something an average person can walk around in a comfortable pace
- Consists of at least 3 places
- Top contributors: cafes, restaurants, places of worship, fast food, bars, banks, pubs

# Data Collection

## Open Street Map via OSMnx




Tag: amenity

Geometry: points and ways

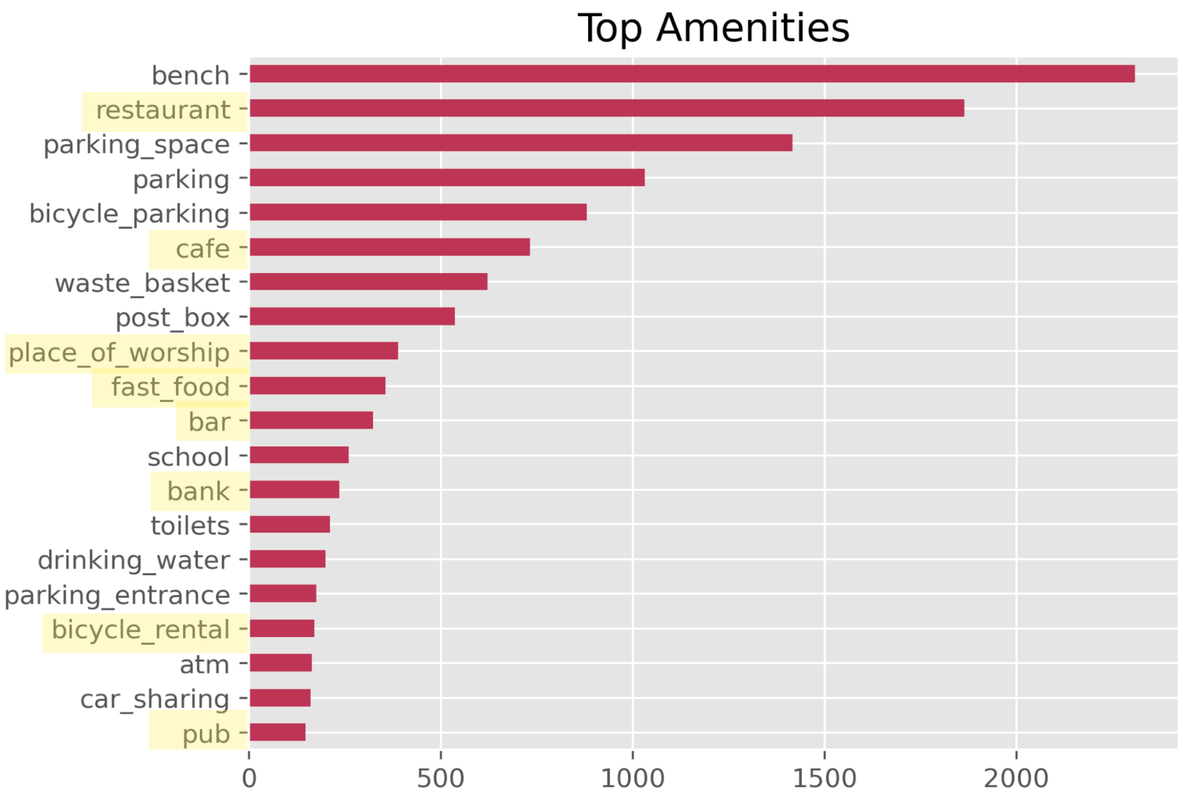
Location: San Francisco

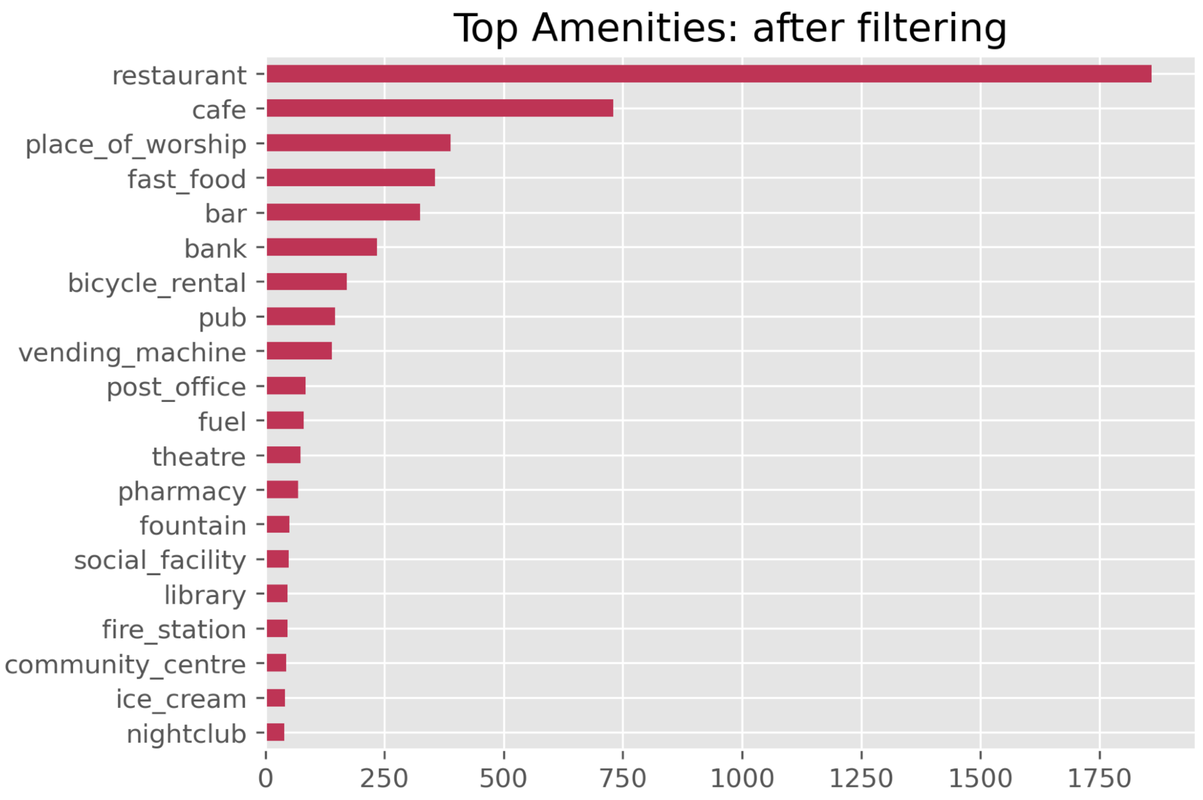
Observations: 13k

# Data Cleaning
## And Feature Engineering

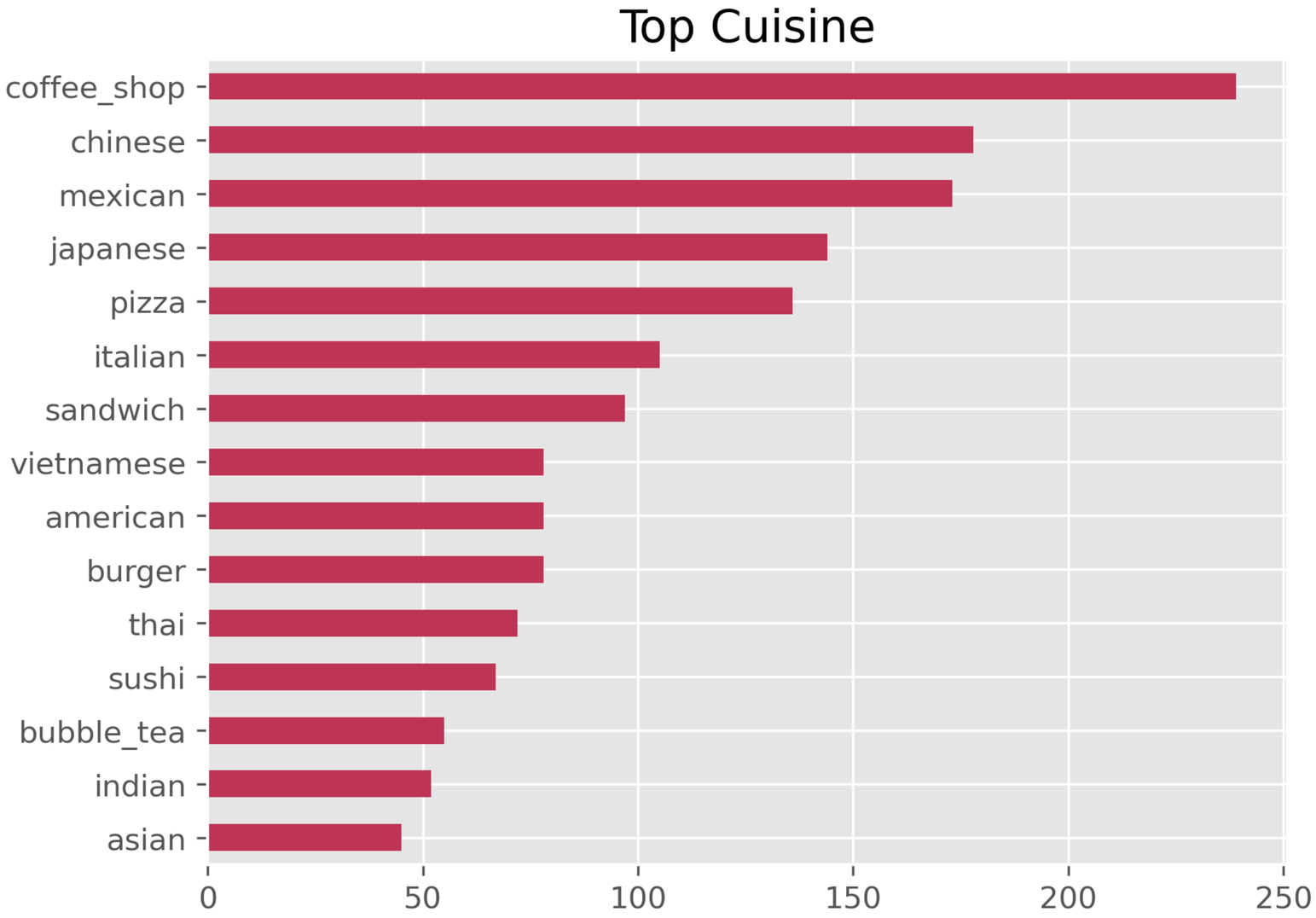
- **Excluded:**
  - Not interesting places (parking, clinic, ATM)
  - Closed places
  - Parks’ infrastructure (benches, BBQ, drinking fountains)
  - Transit (Bus stops, Ticket Validators)
  - Bicyclists’ infrastructure (Bicycle repair station,Bike Racks)
  - Places that were represented only once

# Shortlisted Amenities



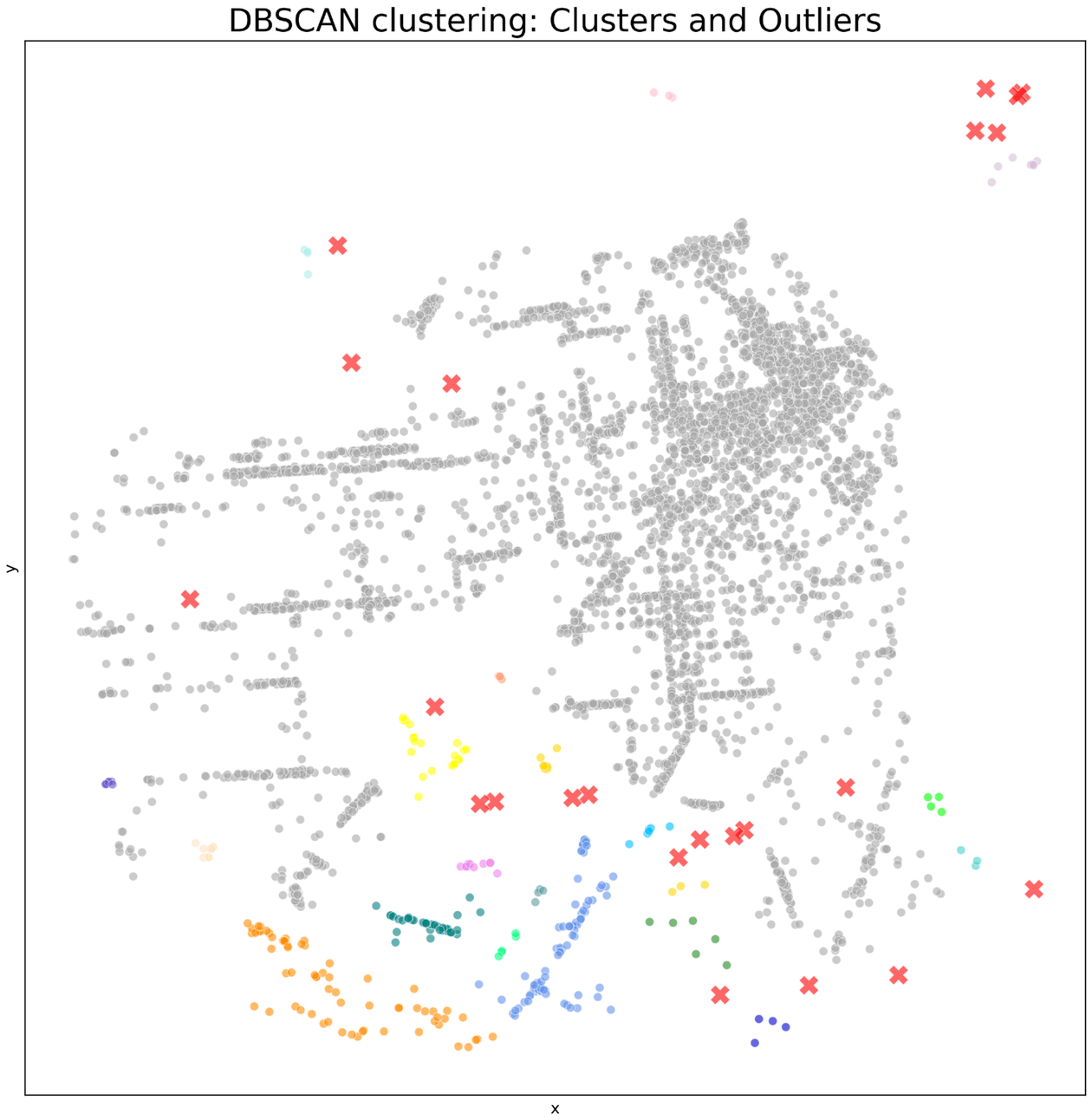
# Feature Engineering

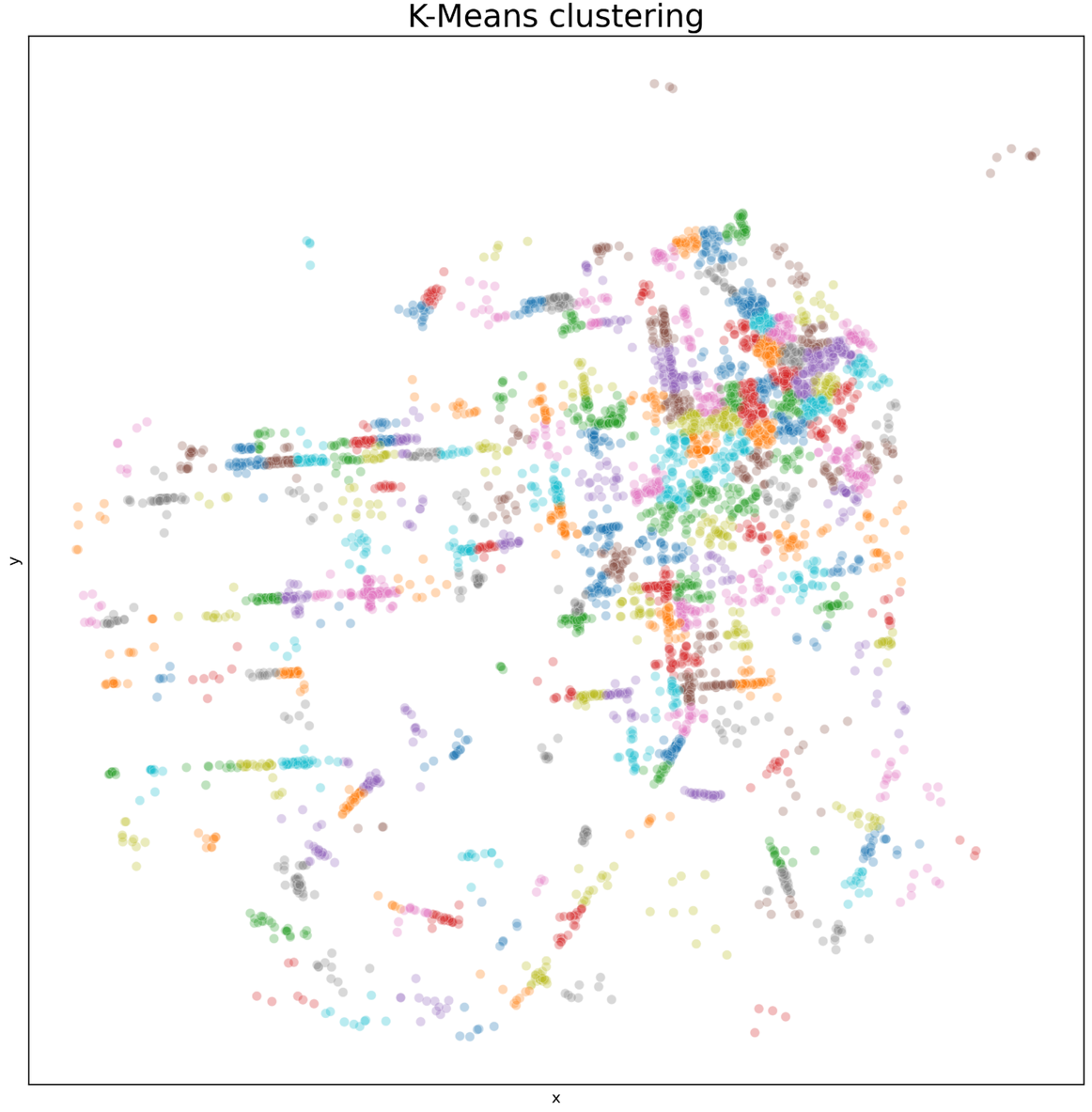


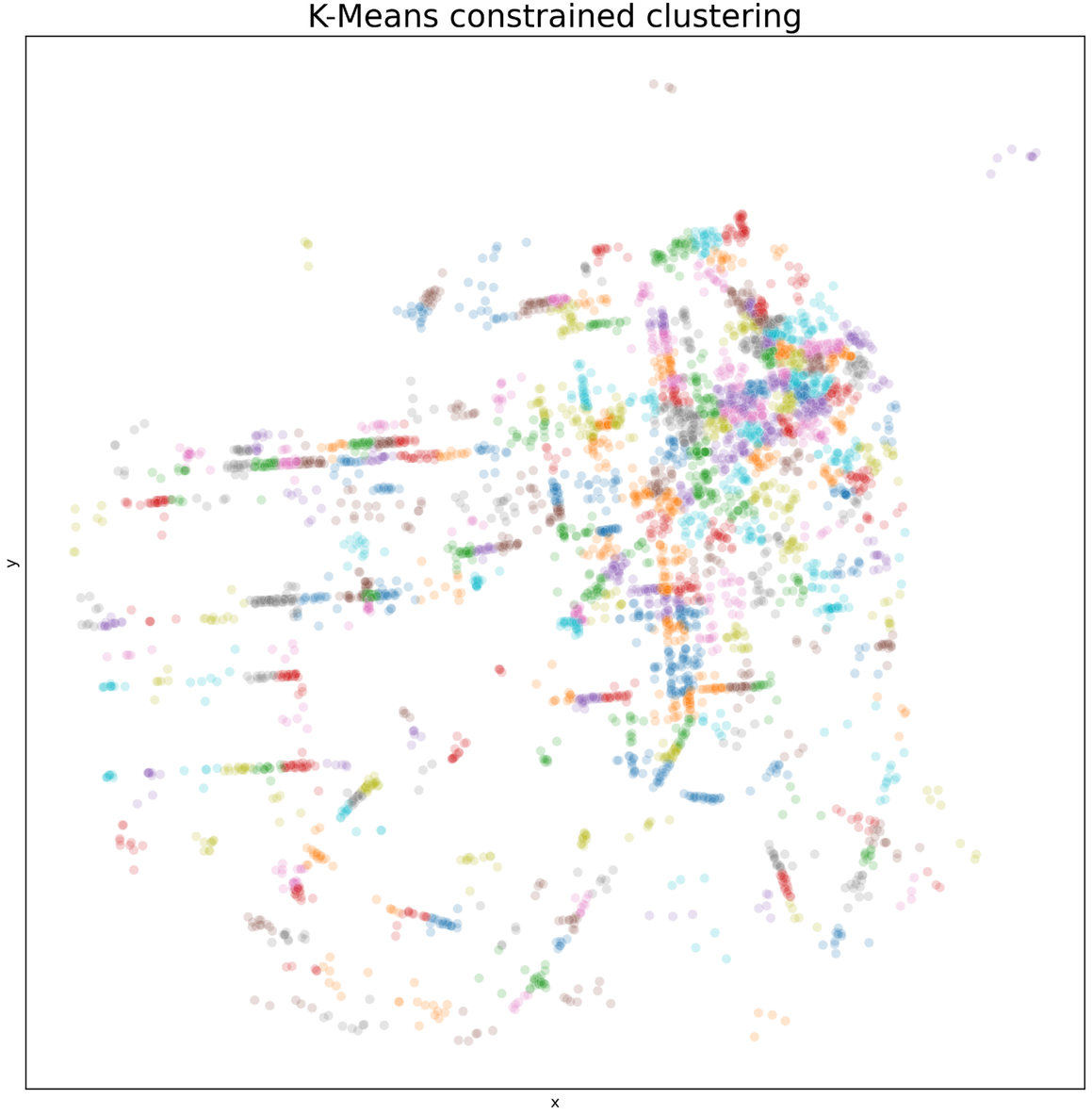
# Clustering

## Using only XY

- **Small clusters: max 1 km**
- **Minimum 3 points**
- **Density and number of clusters aren’t important**




K-Means clustering
y
x


K-Means constrained clustering
y
x

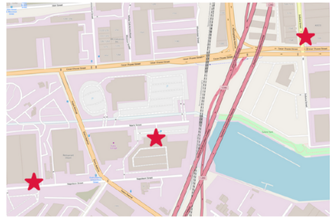
# Biggest Clusters

Cluster 55

Cluster 274

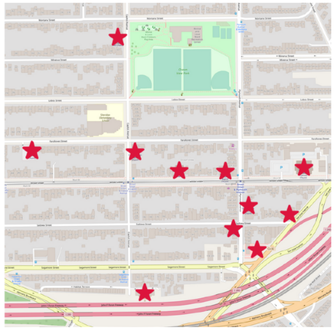
Cluster 154

Cluster 280



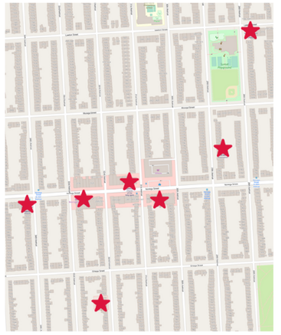
Cluster 6

Cluster 106



Cluster 299

Cluster 161



Cluster 181

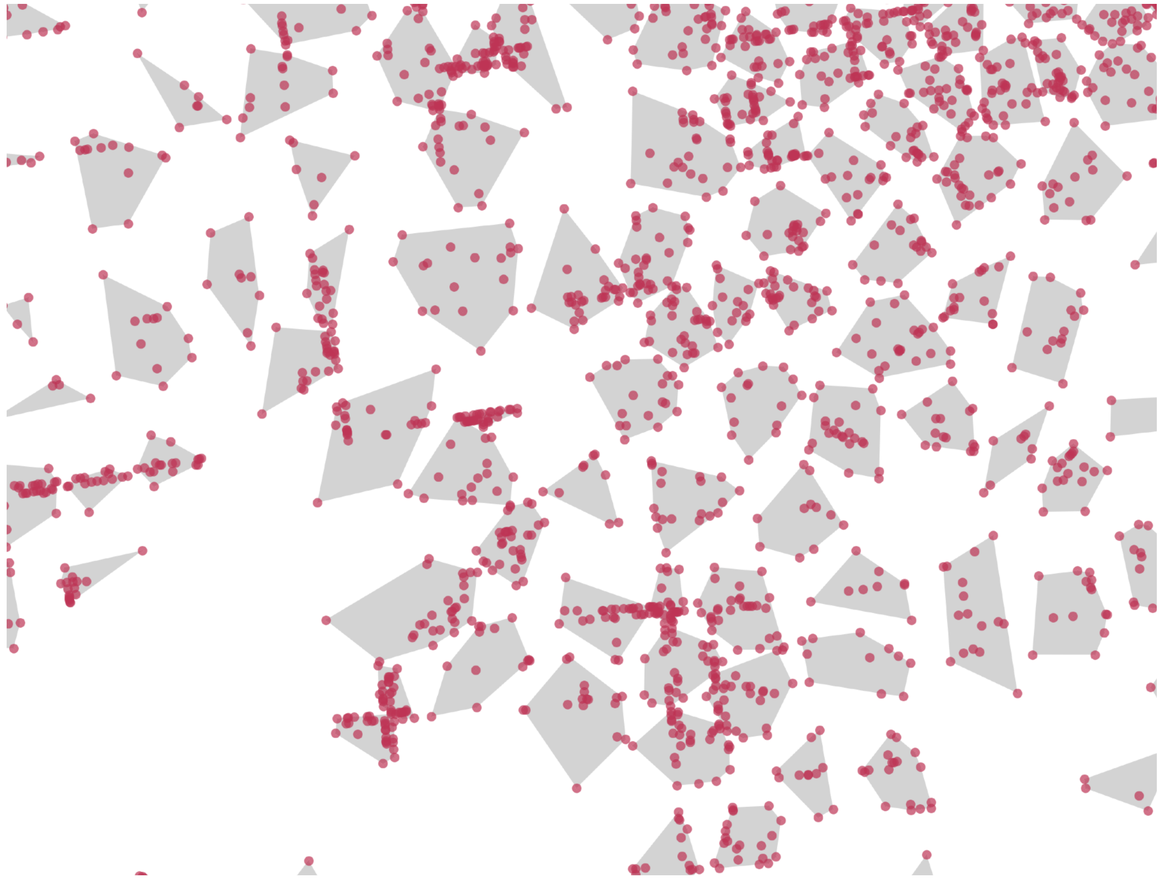
Cluster 109

Cluster 131

Cluster 102

# Putting it all together

- Points to polygons
- Summing all tags' presence per cluster
- Extracting top tags and major streets to display on a map, so recommendation is more informative
- Cosine similarity used to create a Recommender System of places to check out



# Potential Improvements

- Include more places - like grocery shops.
- Collect user feedback on if recommendations make sense and what parts to improve
- Adding more data - proximity to parks and water?