(Photo: Paul Finnerty)

Battle of the Neighborhoods
in
San Francisco, CA

Neeke Swart

April 14, 2021

This project is part of the ‘Applied Data Science Capstone’ Coursera Course and of IBM’s Professional Certificate ‘Data Scientist’.

What to expect?

- Client, goal and business question
- Data acquisition and preparation
- Methodology
- Analysis
- Results and discussion
- Conclusion

CBD, a fast-growing market

- Market: $4.67 million in 2017 to *$25 billion* in 2025
- California is market leader
- Products: oil/tinctures, lotion, gummies, bath bombs, bedsheets
- Use: mainly to reduce stress, anxiety, joint pain, insomnia (71%)
- CBD-only user: 33% is 20-34 years old

CBD, a fast-growing market

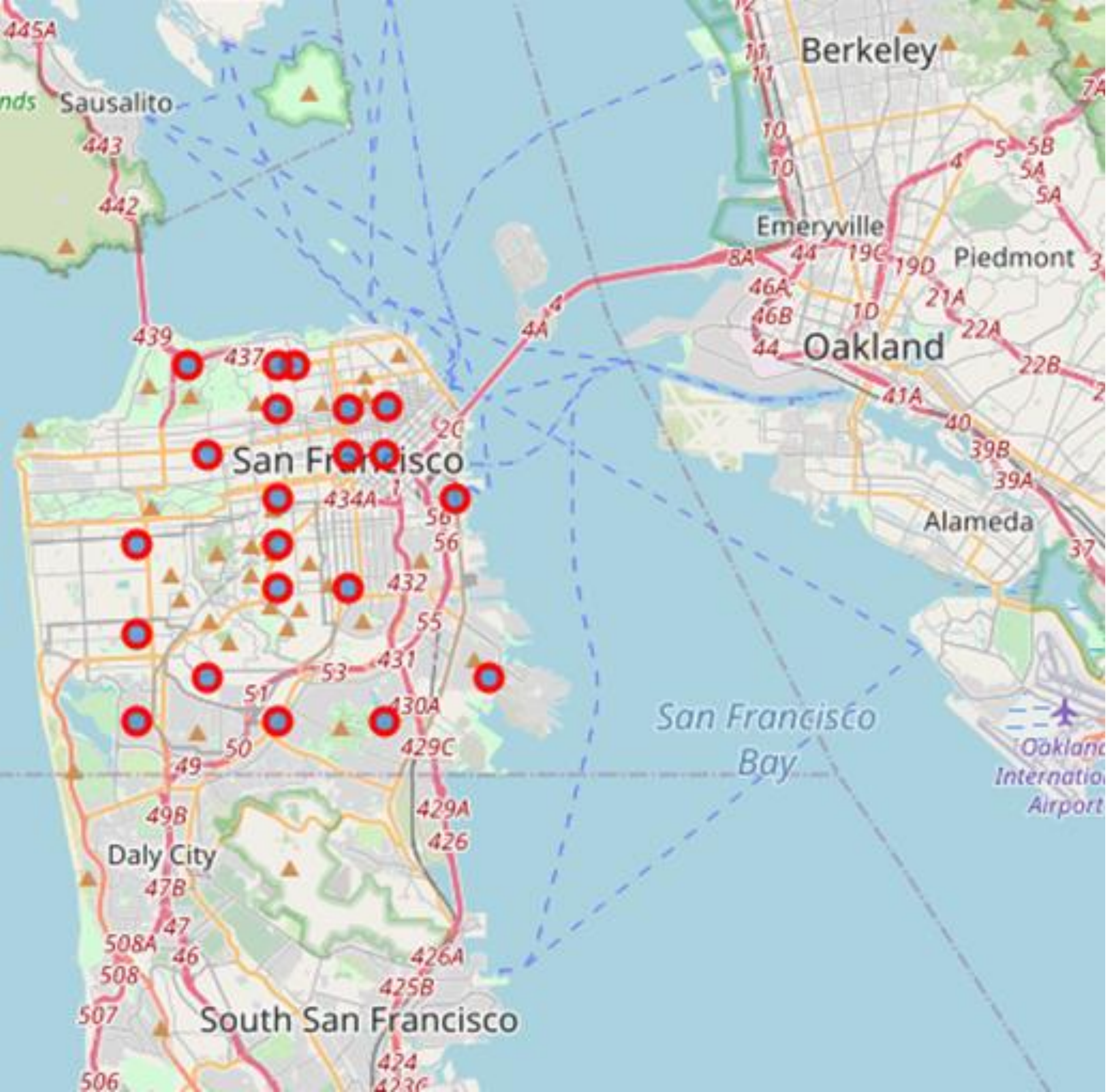
- Our client, CBD seller, wants to open a business in San Francisco, target age is 20–34 year
- Asks for data driven advice on location
- Location parameters (in order of importance):
    1. Cluster of neighborhoods
    2. Dominant age in neighborhood: 20-34 years
    3. Venues attract costumers, like pharmacies with CBD on shelves, coffeeshops, restaurants, nightlife

Data acquisition

- Zip codes and neighborhoods scraped from San Francisco Department of Public Health (2004)
- Latitudes and Longitudes from uszipcode and Geopy (both Python libraries)
- Age distribution per neighborhood from US Census Bureau (2019)
- Venues per neighborhood from Foursquare API

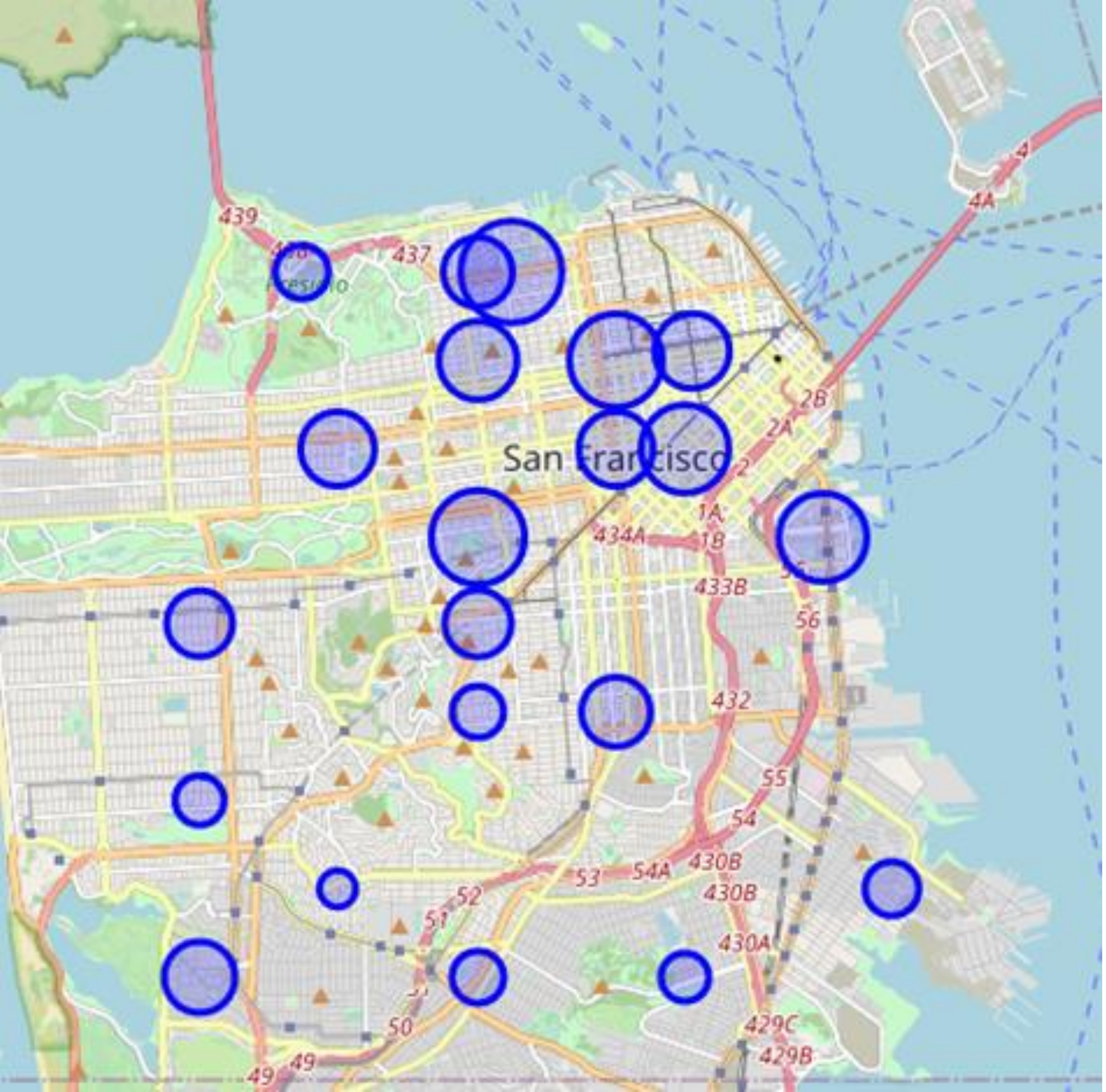
Data preparation

- Two longitudes needed to be corrected
- Absolute age distribution columns were dropped
- ‘Age-Zip Code’ table was transposed
- Column name ‘Zip Code’ was added
- Venues per neighborhood (223 unique categories) were normalized and grouped per neighborhood
- Result: dataframe with 21 rows and 15 features



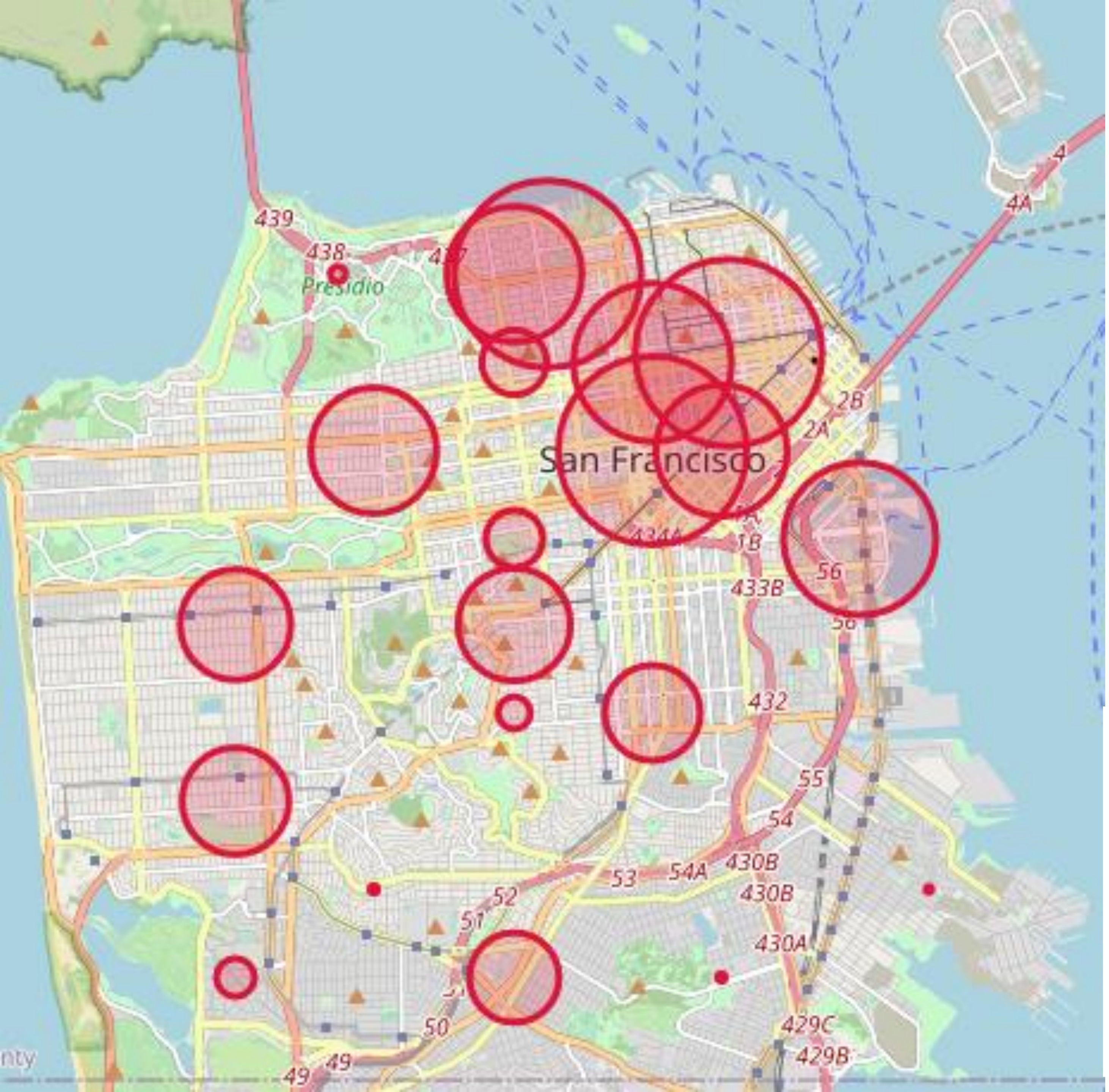
San Francisco neighborhoods

- Map shows 21 neighborhoods on Peninsula



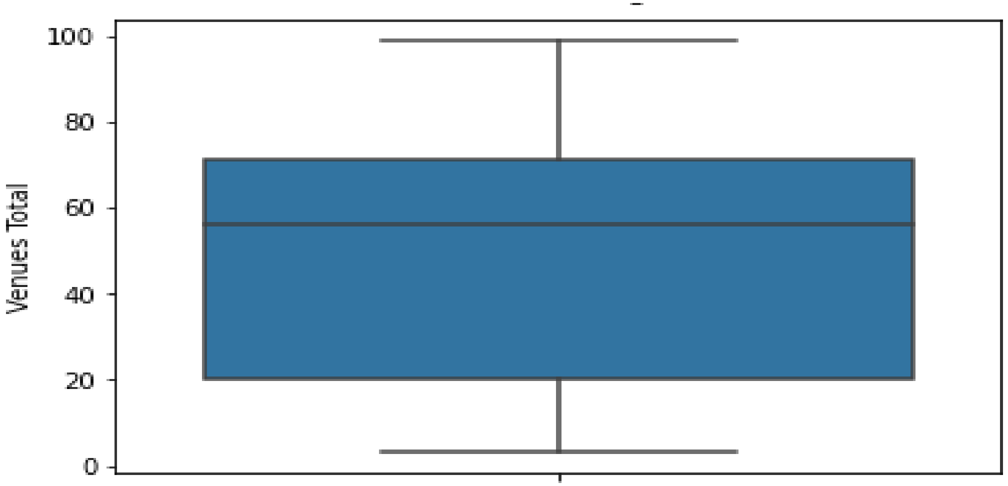
Age distribution 20-34 years

- Bubble map shows higher concentration of 20-34 age group in northeast
- Six neighborhoods where 20-34 age group >=33% :
  - Marina (0.43)
  - Haight-Ashbury (0.39)
  - Polk/Russian Hill (0.39)
  - Portrero Hill (0.37)
  - South of Market (0.37)
  - Western Addition/Japantown (0.33)

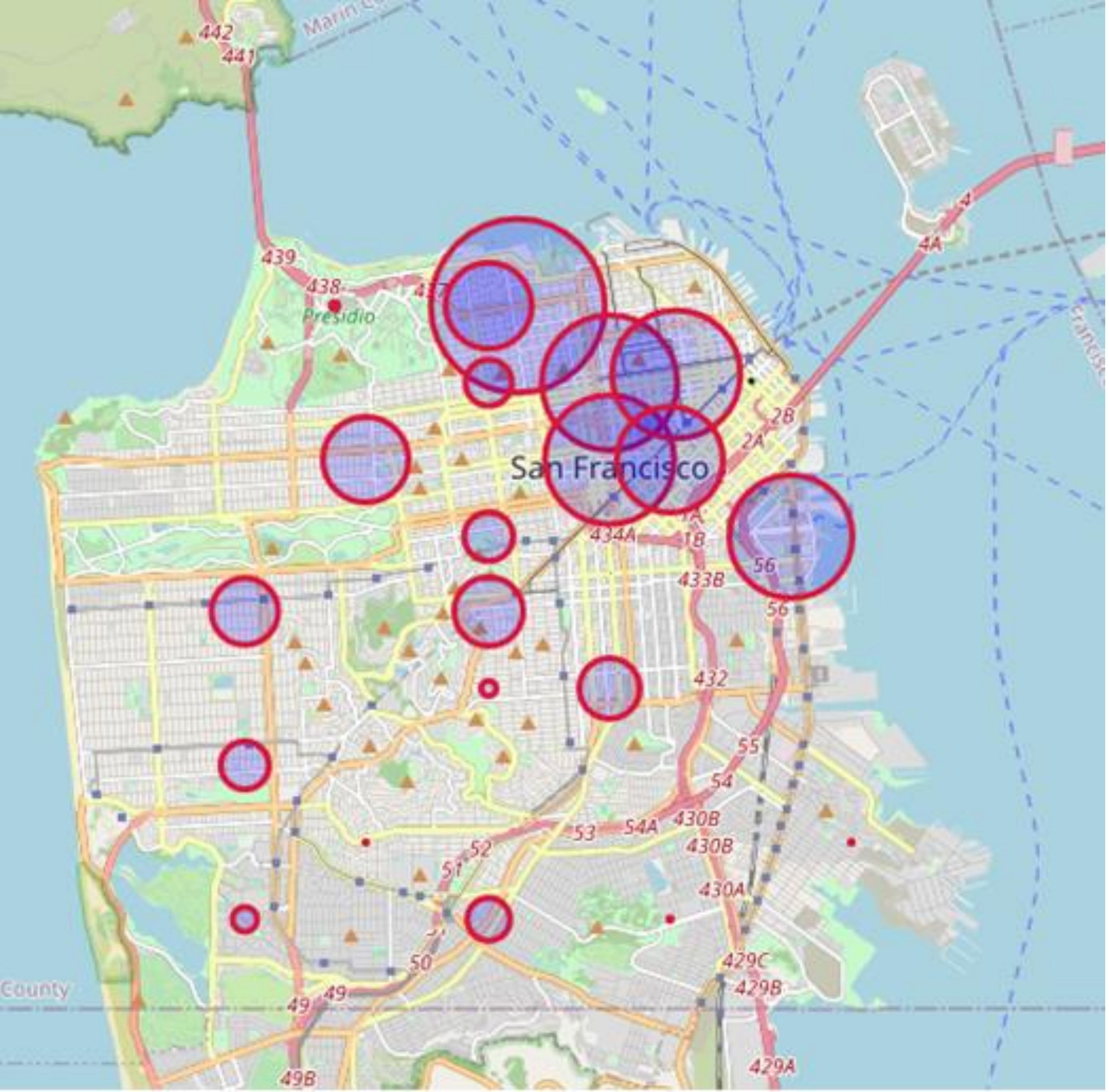


Spread of venues

- Bubble map shows higher concentration of venues in northeast
- Boxplot shows wide range: 3-99, median=56
- Neighborhoods in upper quartile:
  - Hayes Valley/Tenderloin/North of Market (99)
  - Chinatown (98)
  - Marina (98)
  - Polk/Russian Hill (Nob Hill) (83)
  - Portrero Hill (80)

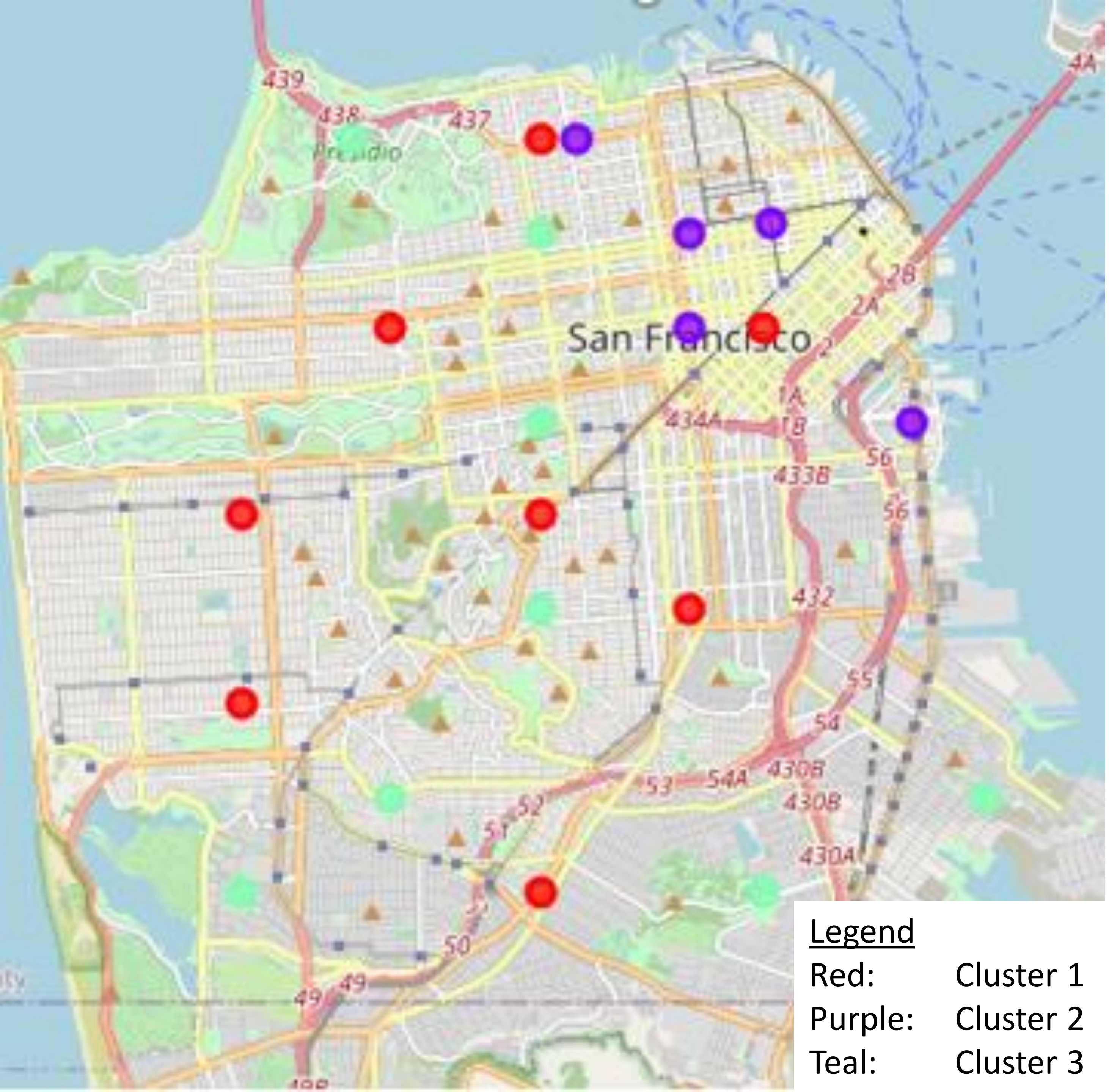


AVE scores



- Age is more important than venues
- AVE: expresses weighted venues. Total venues are multiplied with proportion of 20-34 years age group
- AVE = ('20-34 years' * 'Venues Total')/100
- Bubble map shows concentration of higher AVE scores in northeast

K-Means Clustering



- Optimal K, k=3, has:
  - northeast cluster,
  - homogeneity in age distribution, in the number and in types of venues
- Map shows concentration of Cluster 2 neighborhoods in northeast
- Cluster 1: Mixed-Use Cluster (8 neighborhoods)
- Cluster 2: Hospitality, Health and Luxury Cluster (5 neighborhoods)
- Cluster 3: Residential Cluster (8 neighborhoods)

Results and discussion

- Cluster 3 best starting point to look for location:
  - Cluster of 5 adjacent neighborhoods
  - Age group 20-34 years is dominant in all 5 neighborhoods
  - Highest numbers of venues and top 25% AVE scores of San Francisco, types of venues fit with profile
- Follow-up research:
  - Keep South of Market, North Beach, and Western Addition in mind: location within/next to cluster 3, high 20-34 population, many businesses
  - Research on street level: Real estate availability & prices, existing selling points vs market size, movements of (potential) customers
- Discussion points:
  - Future research should use multiple data sources
  - Difficulty in data replication

Conclusion

- Location search best starting point in:
  - Cluster 3: Marina, Polk/Russian Hill, Hayes Valley/Tenderloin/north of Market, Chinatown, and Portrero Hill
  - Possible additions: South of Market, North Beach, and Western Addition
- Current research might be interesting for businesses who target 20-34 years old, health oriented, customers.

(Photo: Paul Finnerty)

Thank you for your attention!