CAR RIDE-SHARE POTENTIAL IN MID-SIZE U.S. CITIES FROM GEOGRAPHIC SPREAD

Jens Koeplinger, 20 March 2019

This presentation is part of satisfying the IBM Data Science Specialization on Coursera, Course 9 ”Applied Data Science Capstone”

Objective

- Fictitious analysis requested by a car ride-share company startup, looking for the best suited places to roll out their service.
- Customers will be able to sign up for shared car rides across town.
- The cities need to be geographically spread out, with venues arranged in a non-compact manner as opposed to centered around a single core, and
- they need to have a significant number of venues of types that indicate a population open to trying out new transportation arrangements.

Data and success criterion

- Cities with 300,000 and 400,000 inhabitants according to the 2017 census estimate.
- Indicator venues include alternative eating, lifestyle, and work types. This may include coworking spaces, yoga studios, delicatessen, tea rooms, vegan and other nonstandard restaurants.
- A successful study clusters cities by geographic spread of their venues, and then ranks them within each cluster by relative occurrence of indicator venues as com- pared to total venues.

City: population 300,000 - 400,000 (2017)

- There are 19 cities (per 2017 U.S. Census estimate), across the US:
  - California: Bakersfield, Anaheim, Santa Ana, Riverside, Stockton.
  - Ohio: Cleveland, Cincinnati.
  - Texas: Arlington, Corpus Christi.
  - All other states: New Orleans, LA, Wichita, KA, Tampa, FL, Aurora, CO, Honolulu, HI, Lexington, KY, St. Louis, MO, Saint Paul, MN, Henderson, NV, and Pittsburgh, PA.

Indicator venues

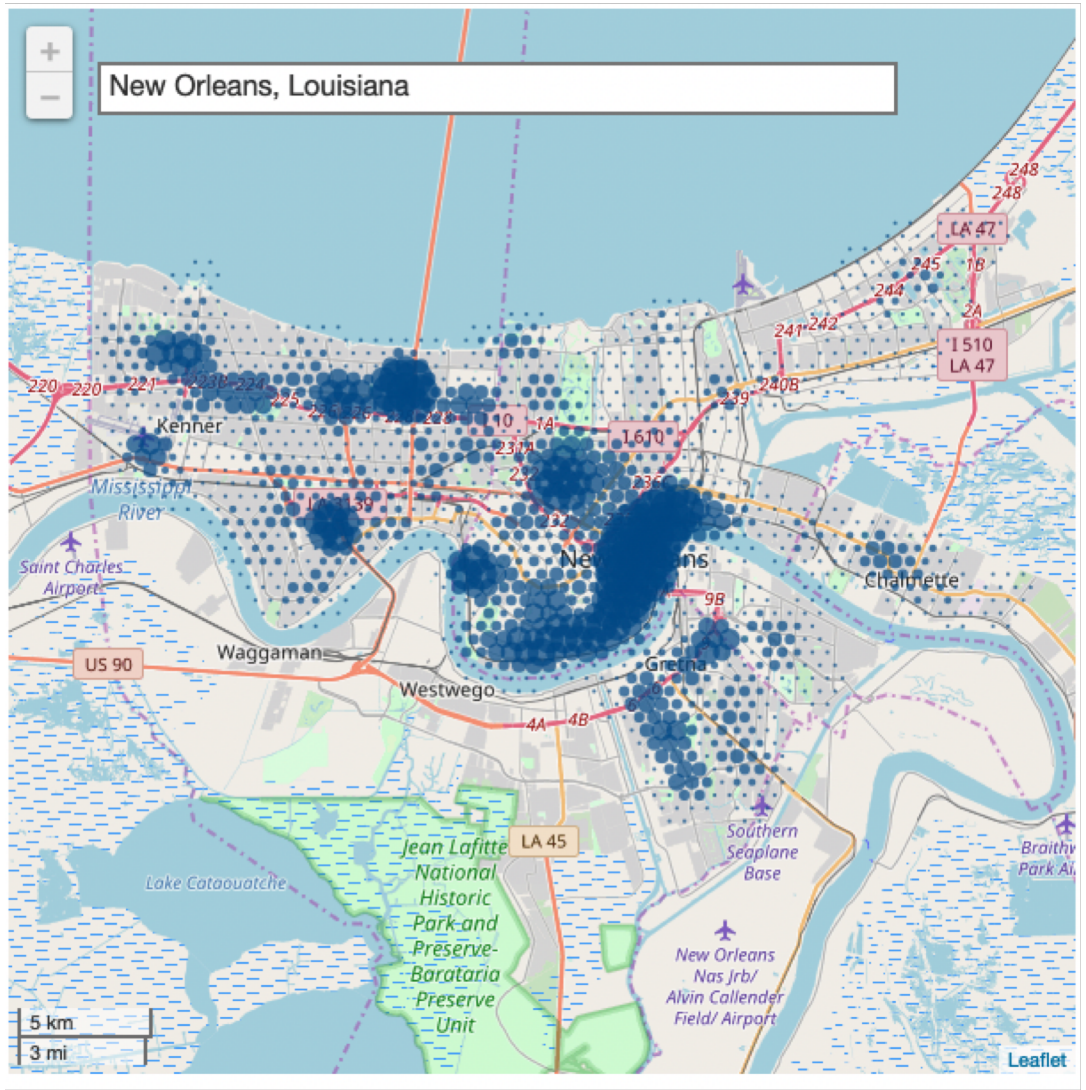
- The ficticious company provided a list of substrings to search for in venue type.
- If the substring is found, it is considered an “indicator venue” for a city population more open to trying out the company’s new service.
- The provided substring list is:
  - share, sharing, incubat, innovat, coworking, alternative, yoga, salad, bike, □tness, running, jogging, cycling, cycle, athletics, gluten, health, recreation, tennis, vegetarian, vegan, disc golf, pilates
- This is highly subjective.

Foursquare foraging algorithm

- Start at a reference coordinate,
- get number of venues in certain distance,
- search a hex grid for nearby venues,
- continue at coordinate with most venues.
- Iterate until maximum number of coordinates checked.
- While at it, collect venue type and count indicator venues.

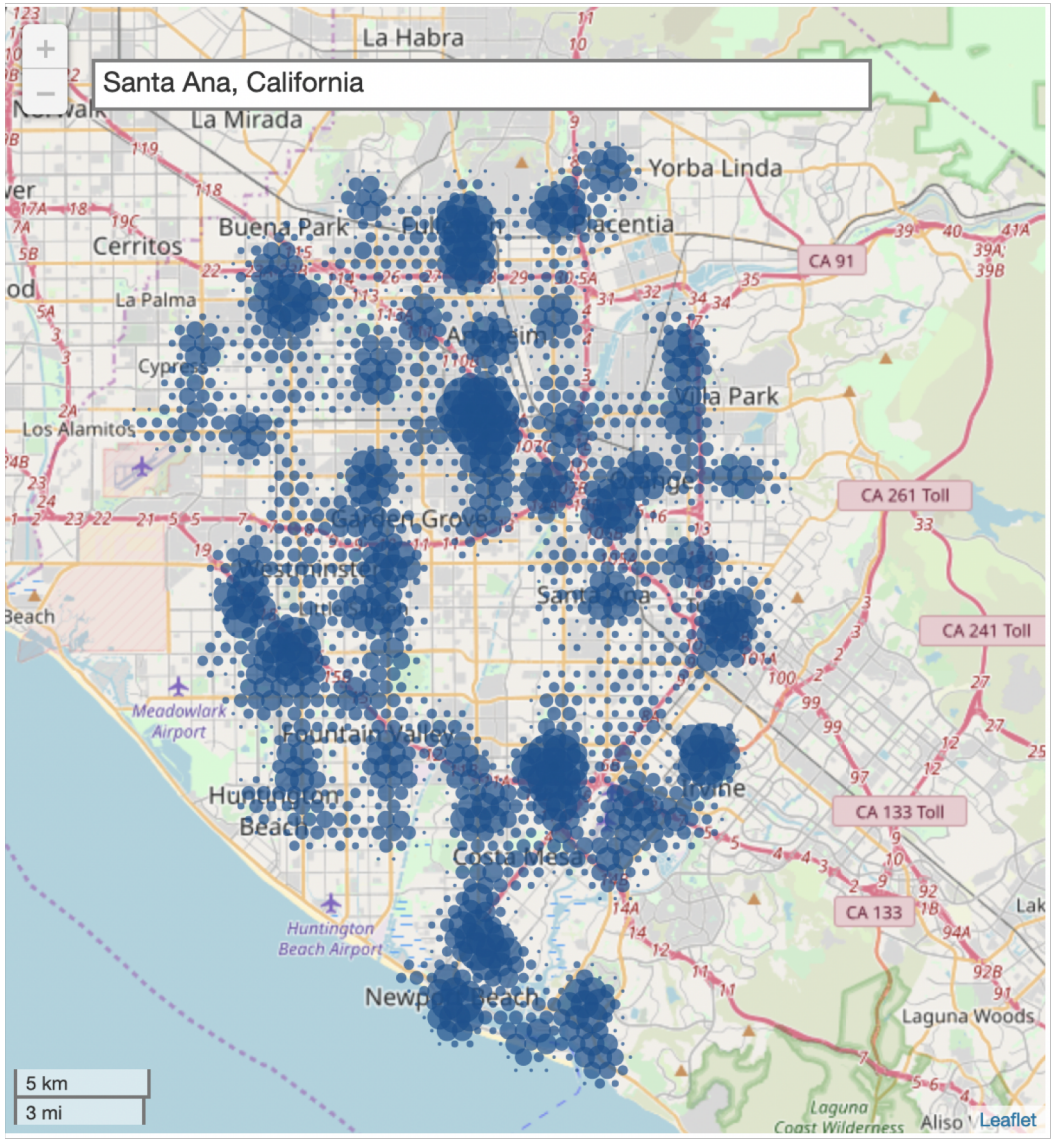
Example: New Orleans, LA

- Most venues are at the city's most popular area



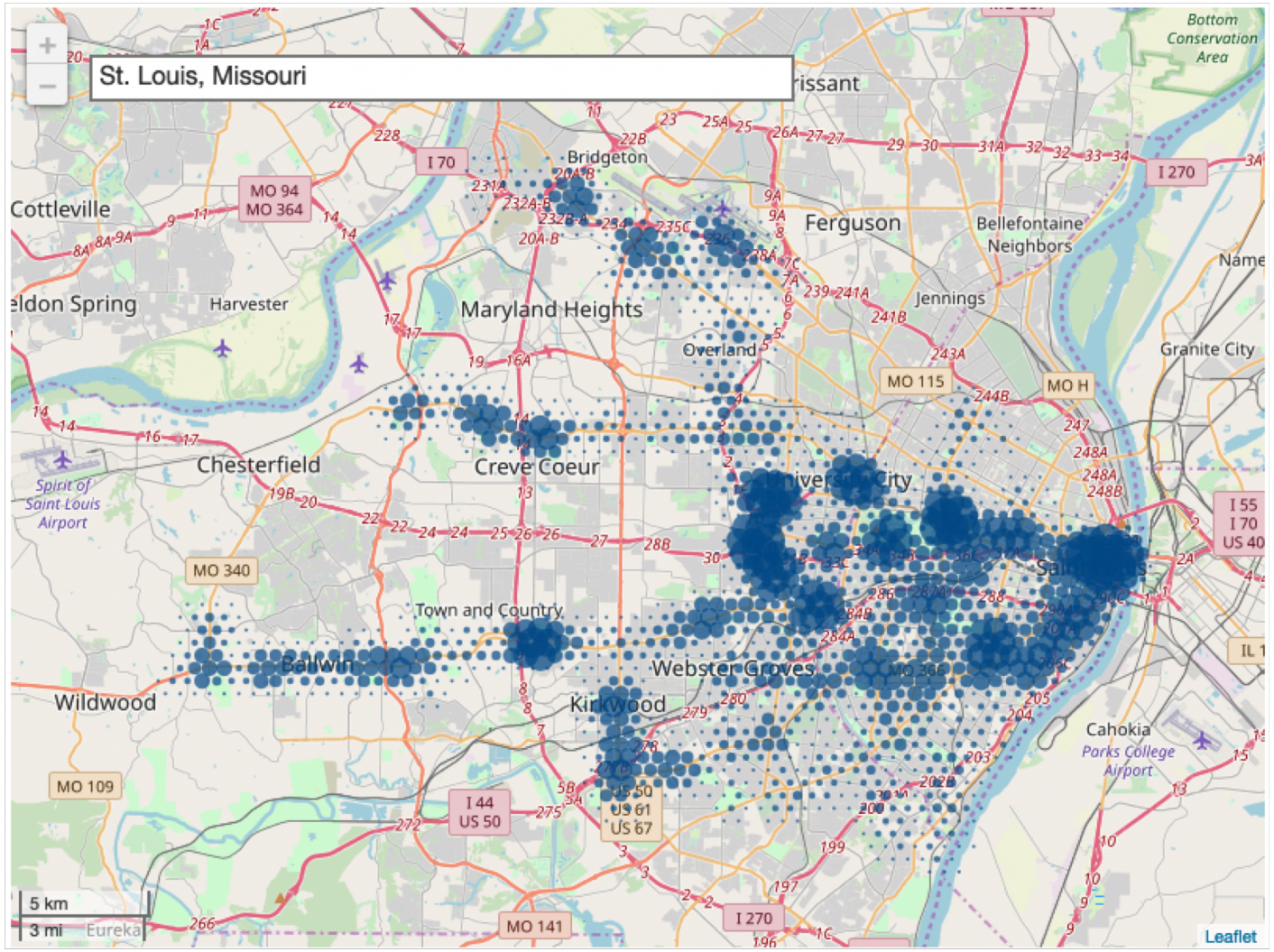
Example: Santa Ana and Anaheim, CA

- Cities run into one another



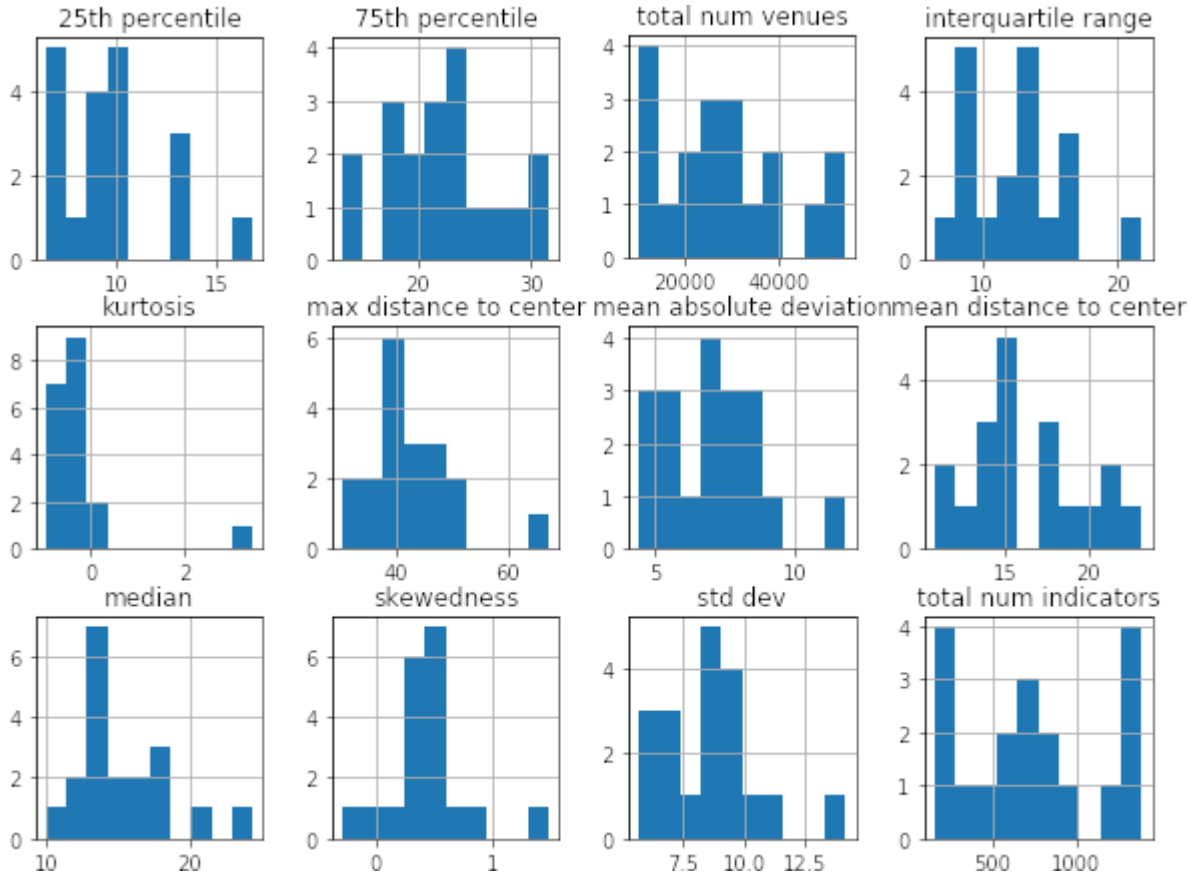
Example: St. Louis, MO

- Center with veins



Aggregate geographic spread

- Flatten dataset using descriptive analytics for each city, of venue distance to center:

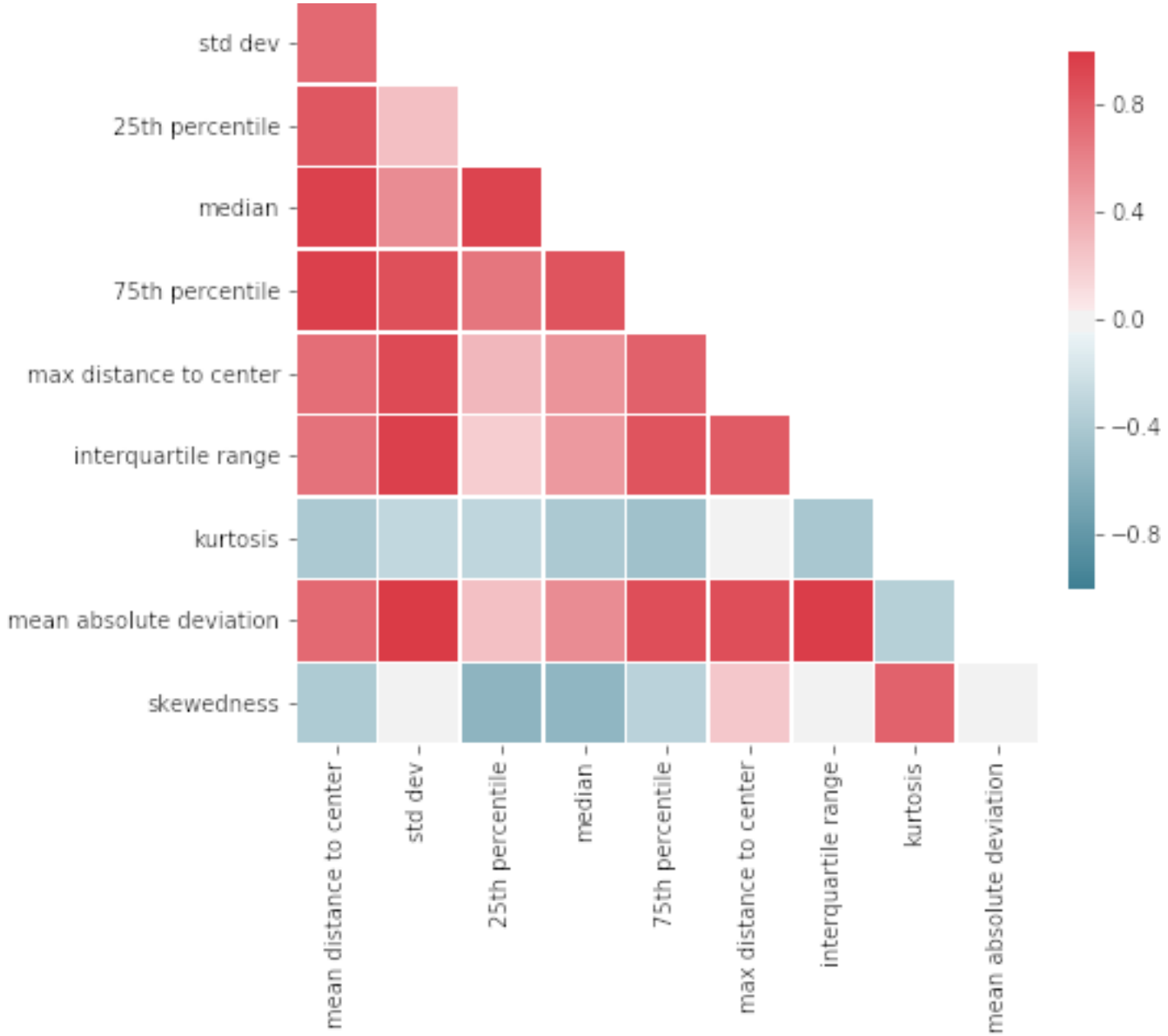


Feature engineering

- Remove highly correlated data points,
- build features by standardizing and bounding the data points:
  - Divide by standard deviation,
  - center, and
  - apply arctan to reduce the effect of outliers.

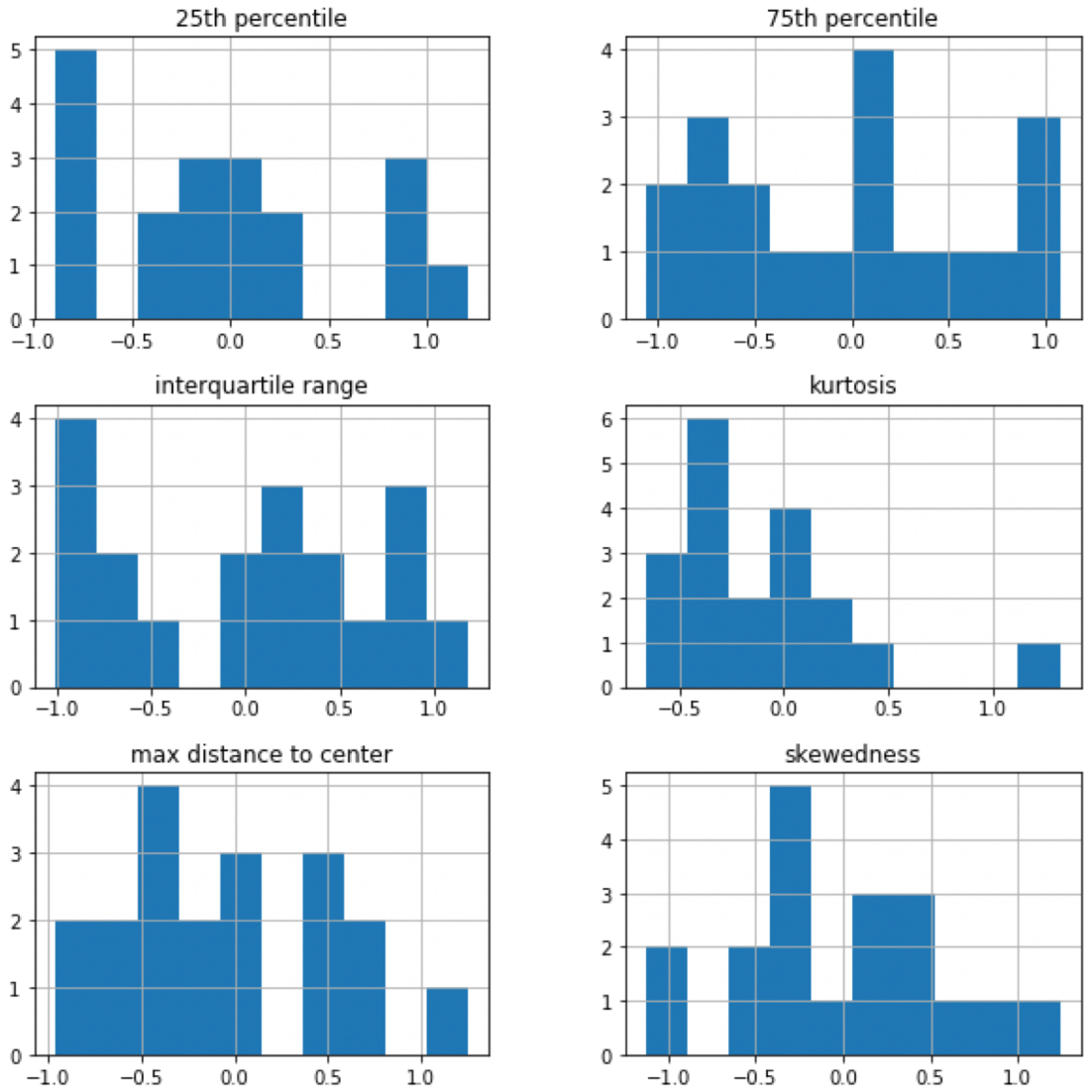
Correlations

- Data points to exclude:
  - Mean absolute deviation,
  - mean distance to center,
  - median,
  - standard deviation.



Final features

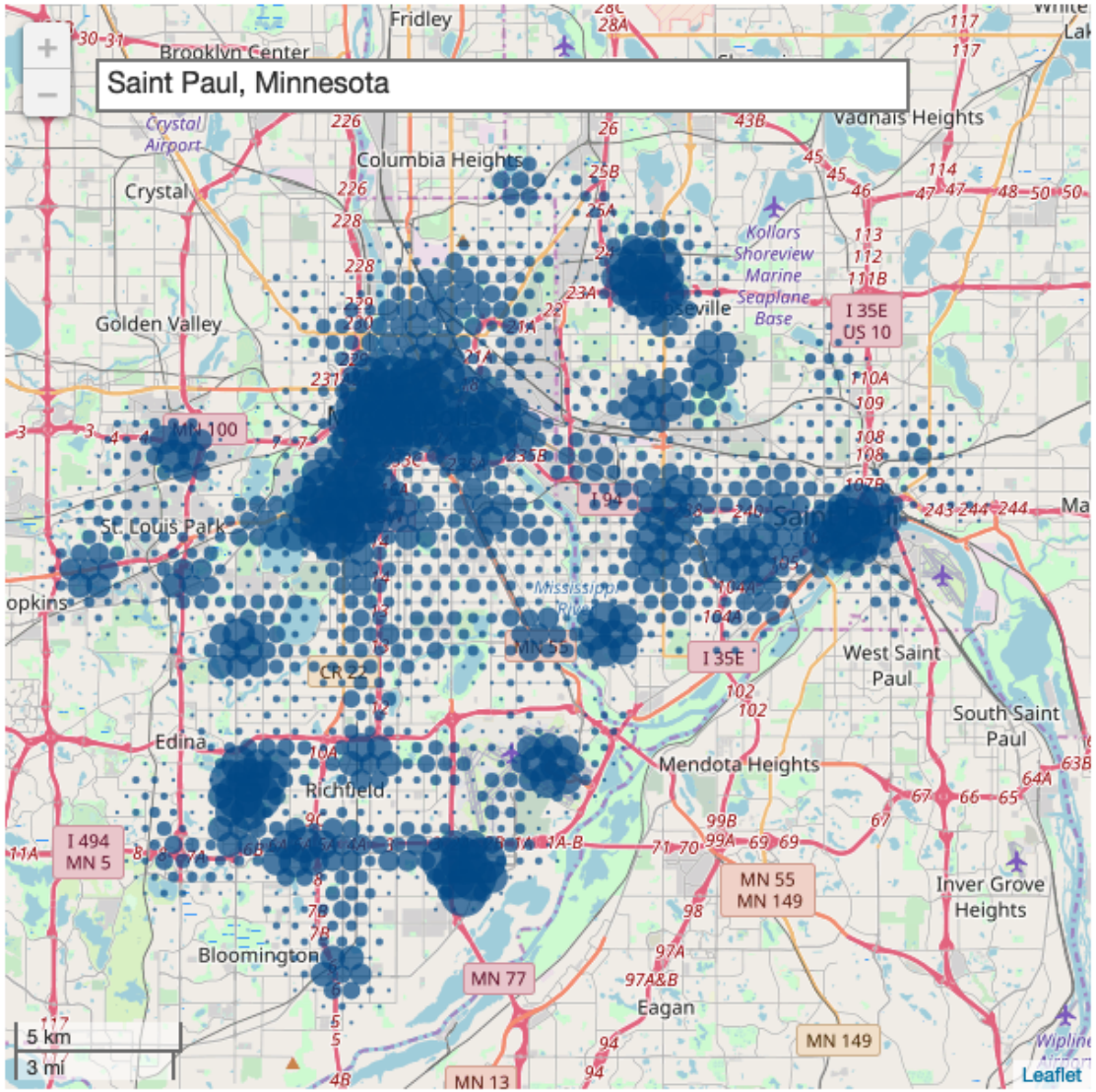
- Feature value distribution after selection, standardization and bounding:



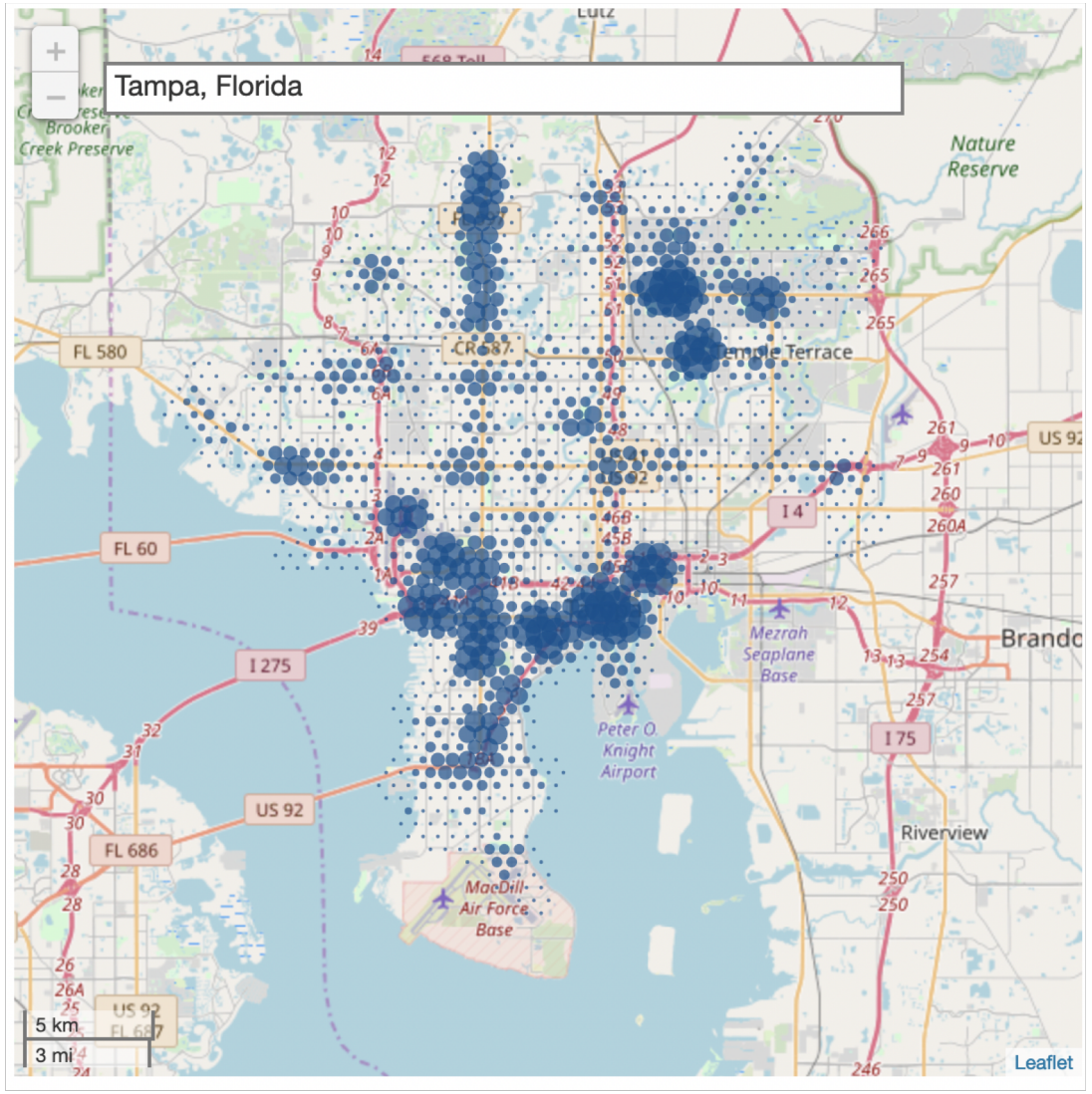
K-means clustering

- Saint Paul, MN is the “most indicative city” (highest relative percentage of indicator values).
- Apply k-means on the features:
  - Appears stable for Saint Paul, consistently putting it together with the same four or five other cities,
  - The cluster contains other cities with high relative percentage of indicators.
- Methodology overall appears stable in the result from clustering, and there is a single cluster that contains the cities with highest number of indicator venues:

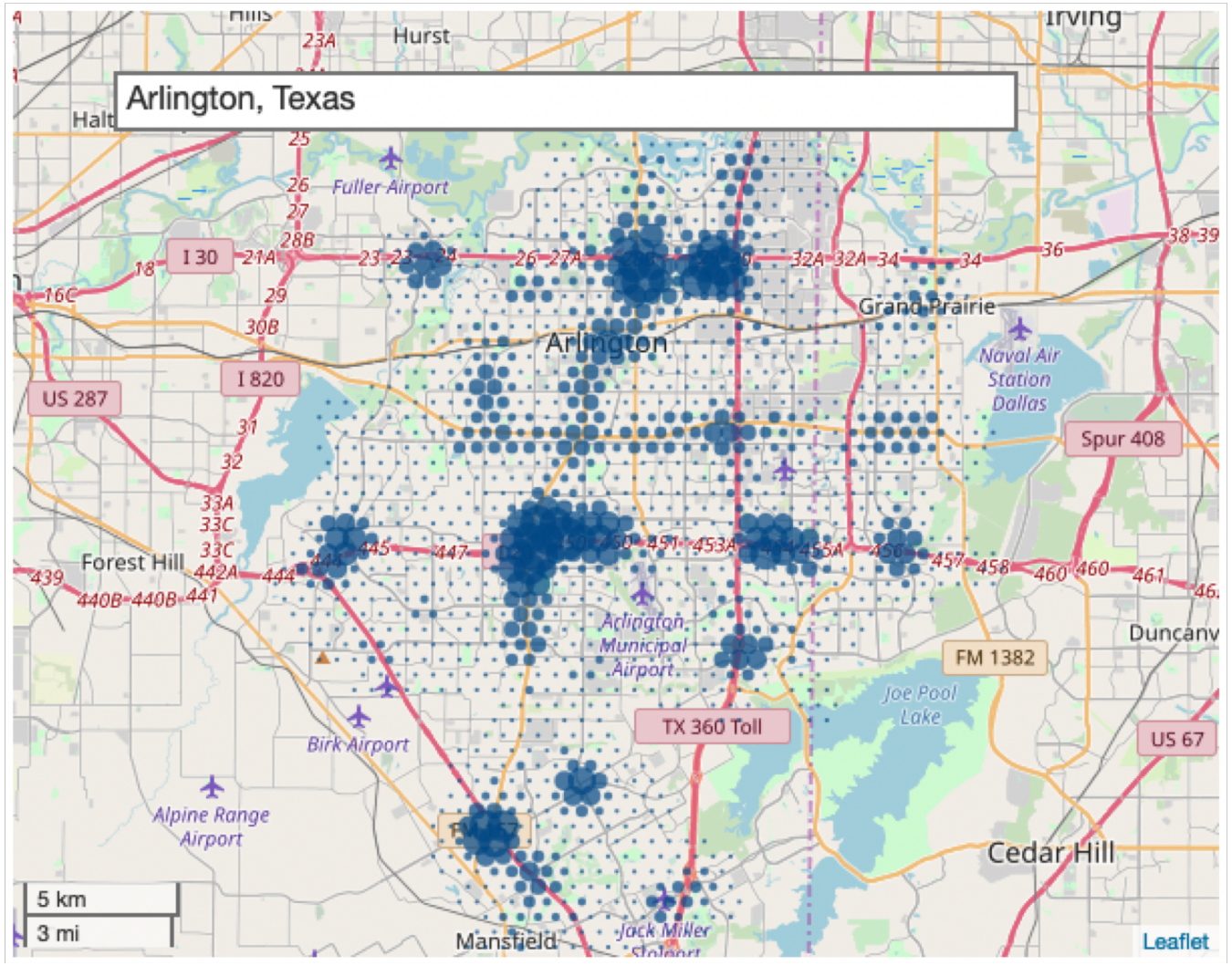
Result #1 (of 3): Saint Paul, MN



Result #2 (of 3): Tampa, FL



Result #3 (of 3): Arlington, TX



Cross check: Geographic spread

- By visually inspecting the venue spread for the top 3 cities, we confirm the desired geographic spread of venues (as opposed to centralization).
- The methodology is consistent with the fictitious company's requirements.

Bias

- Various factors bias the result:
  - Indicator venue selection is highly subjective.
  - Foursquare venue search by radius double-counts venues due to overlap.
  - Foursquare venue collection depends on contributing users.
  - Algorithm avoids natural barriers (e.g. broad rivers).
  - Algorithm runs into neighboring cities (thereby violating the study premise of population size) by following areas with a higher density of venues.
- Recommendation: Perform another study with modified methodology, and compare the results.

References

- Report (including source references): https://github.com/koeplinger/sandbox-ibm-ds/blob/master/carRideShareAnalysisReport.pdf
- Investigation notebooks: https://github.com/koeplinger/sandbox-ibm-ds/blob/master/ibmDsSpecCourse9capstone.ipynb