Data Science Capstone Project Report

Assessing distribution / usage of Foursquare in German cities

Business problem

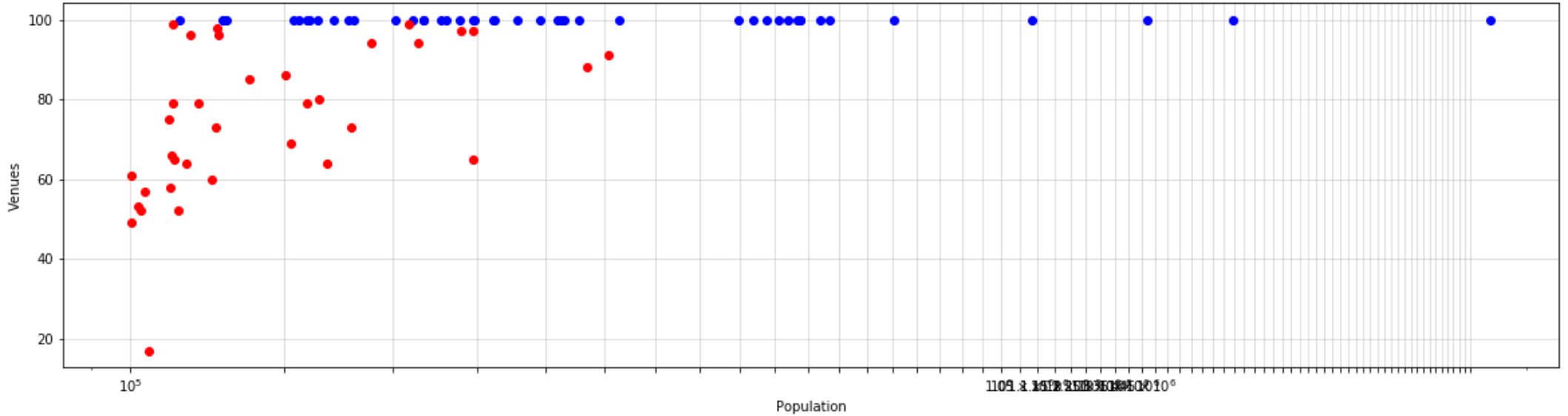
This project will assess the properties of German cities with a population of >100,000 people and their respective Foursquare venues data to give insights into the following questions:

- How much and what type of Foursquare Places data can a data scientist use to compare German cities?
- What are the limits of this comparison?
- Can one cluster the cities to enable data scientist to decide which cities are better suited for comparison than others? In which properties do the city clusters differ?
- Can one select cities that should be targeted with an ad campaign to promote the usage of Foursquare to expand the database?

Data acquisition and cleaning

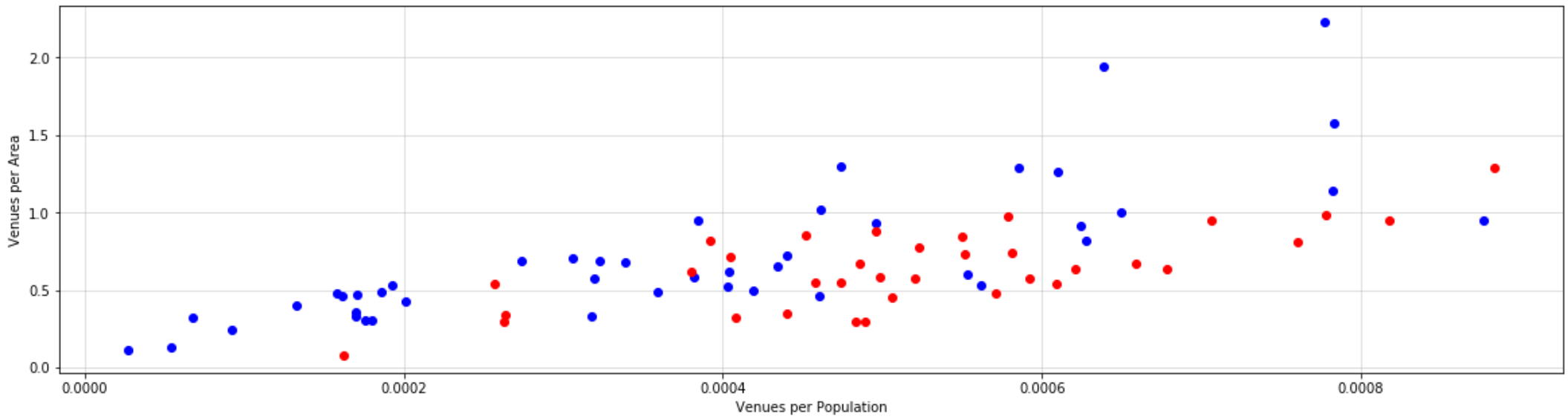
- List of German cities "Großstadt" from Wikipedia (Link)
  - Name of City, Population (latest data from 2018), Area of city (2016 data)
- Population by age groups by state (https://www.statistikportal.de/en/node/132)
  - State ("Bundesland"), Age groups
- OpenStreetMap with Geopy (OSM Nominatim), using the city name
  - Latitude, Longitude
- German federal states geojson (http://opendatalab.de/projects/geojson-utilities/)
  - Geojson file: bundeslaender_simplify200.geojson
- Foursquare GET Venue Search (https://developer.foursquare.com/docs/api-reference/venues/search/)
  - venue name, primary category of venue
- **Foursquare limits the venues that can be retrieved per call per city to 100**
- **Cities with 100 venues are discarded, Final dataframe consists of 35 cities**

Number of venues vs. population



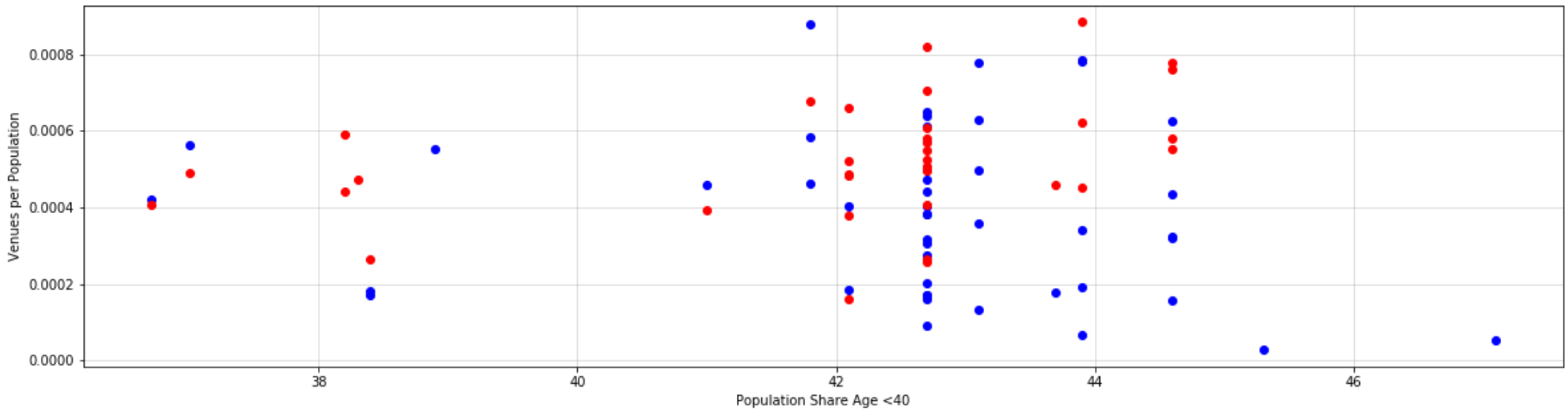
- except for two cities every city with a population bigger than 250k reaches the limit of 100 venue counts (blue dots)
- cities with <100 venues (red dots) are mainly small cities population-wise

‚Venue density‘



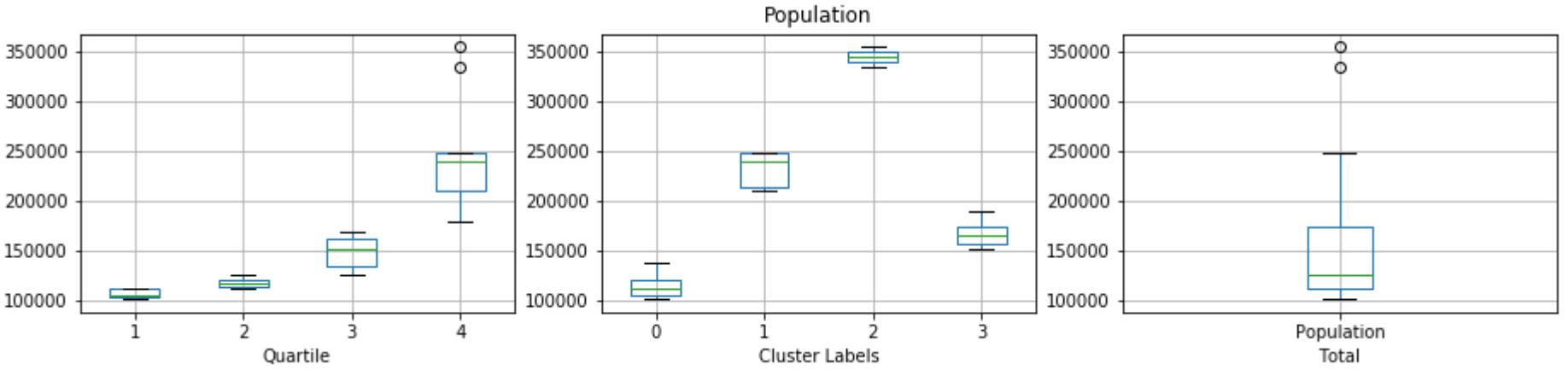
- No obvious difference between cities with >= 100 venues (blue dots) and <100 venues (red dots) regarding the ‚venue density‘ per population or per area (sqkm)

Venues number vs.age group



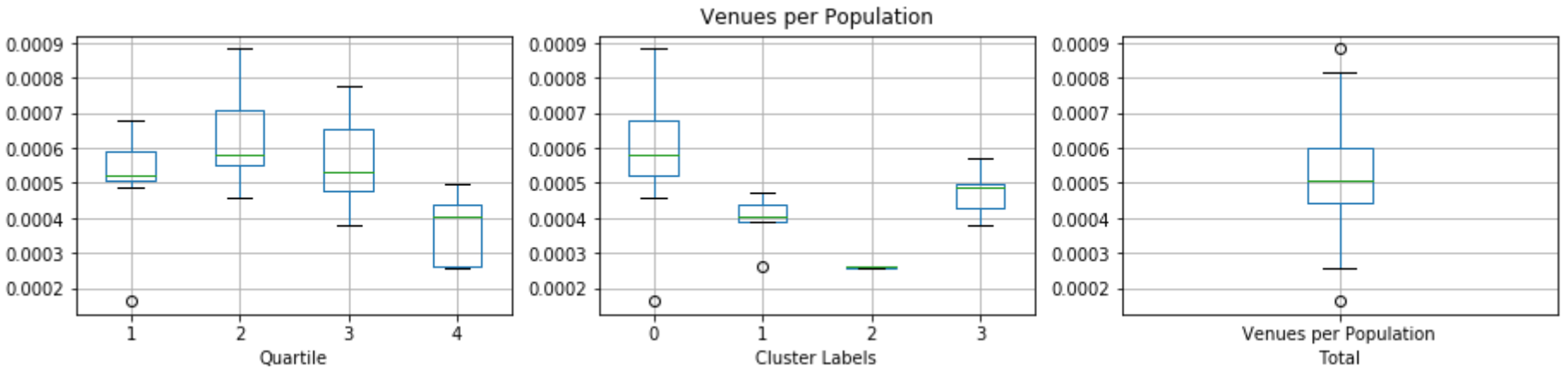
- A slight tendency of the cities with <100 venues towards a lower share of young people, but still not an obvious factor

Manual clustering vs. K-means



- Manual clustering by quartiles of the population size reflects in the boxplot on the left
- K-Means method clustering yields quite similar results (middle): significant difference between population size between clusters
- The two largest cities (by population), that are outliers in the ‘manual’ clustering method quartile 4 seem to fall into K-means cluster 2.

Manual clustering vs. K-means



- K-Means method clustering (middle) results in a more significant split between the ‘Venues per Population’ data than the manual clustering method

Comparing top 5 venues of K-Means cluster #0 and #2

	Cluster Labels	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	0	Ingolstadt	Supermarket	German Restaurant	Café	Italian Restaurant	Drugstore
9	0	Erlangen	German Restaurant	Italian Restaurant	Supermarket	Café	Beer Garden
32	0	Bottrop	Supermarket	Clothing Store	Gym / Fitness Center	Drugstore	Coffee Shop
33	0	Recklinghausen	Supermarket	Café	Drugstore	Italian Restaurant	Clothing Store
34	0	Bergisch Gladbach	Supermarket	Drugstore	Hotel	Italian Restaurant	Hardware Store
16	2	Wuppertal	Supermarket	Café	Bar	Gas Station	Park
17	2	Bielefeld	Supermarket	Hotel	Café	Bar	Bakery

- The supermarket category dominates the 1st most common venues.
- Group 'cluster label 0' has more 'store-like' venues (Drugstore, Clothing Store, Hardware Store,...) are than group 'cluster label 2'.
- On the other hand, in group 'cluster label 2' more 'travel-related' and 'touristic' venues like hotels and cafés dominated the 2nd to 5th most common venues.

Conclusion

	Cluster Labels	Quartile	City	Population	Venues	Venues per Population	Population Share Age <40	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
16	2	4	Wuppertal	354382	91	0.000257	42.7	Supermarket	Café	Bar	Gas Station	Park
17	2	4	Bielefeld	333786	88	0.000264	42.7	Supermarket	Hotel	Café	Bar	Bakery

- The Foursquare's API limits each search to 100 venues per call. This limit may be bypassed by using a smaller search radius and multiple searches per city.
- When using the K-means method to cluster the cities with less than 100 venues, cluster #2 with only two cities sticks out.
- This cluster includes the cities **Bielefeld** and **Wuppertal**, which have large population but also a low count of venues. Thus resulting in very low 'Venues per Population' values.
- **Wuppertal and Bielefeld should be targeted by the advertisement, because they have quite a large population (>300,000 inhabitants) and a high share of young people, which, to me, are the factors for a large potential Foursquare user base**.