Applied Data Science Capstone Project Report

Problem & background

City segmentation (clustering)

Companies in touristic business may want to improve their offerings by segmenting destination cities and tailoring product for each segment. This may be relevant for city guide providers, tour makers, marketing companies etc. Building a tour to cities famous for their nightlife may be different from those famous for shopping.

Points of concentration of places of interest

Tourists when planning their visits may want to know points of concentration of places of interests depending on their category: Food, Shopping, Sightseeing etc. Visitors may choose accommodation close to those concentration points or plan transportation between them

Data

Foursquare venues data will be used for analysis. Currently analysis is performed by venue type only. With paid Foursquare account I could be further improved to take venue ratings into account.

Cities are taken from https://www.listchallenges.com/top-100-most-visited-cities-in-the-world

Methodology

Venue categories are rolled up to their top level according to https://developer.foursquare.com/docs/resources/categories . Only following categories are considered:

- **Arts & Entertainment**
- **College & University**
- **Food**
- **Nightlife Spot**
- **Outdoors & Recreation**
- **Shop & Service**

Others are deemed as not-so-important for tourists. Only primary category is considered. Venues without categories are excluded.

Analyzed are venues returned by Foursquere for 16 tiles closest to city center. Each tile is approximately 500m by 500m

City segmentation (clustering)

Each city is represented by ratio that corresponding venue category represent from all considered venues. Cities are grouped with KMeans

Points of concentration of places of interest

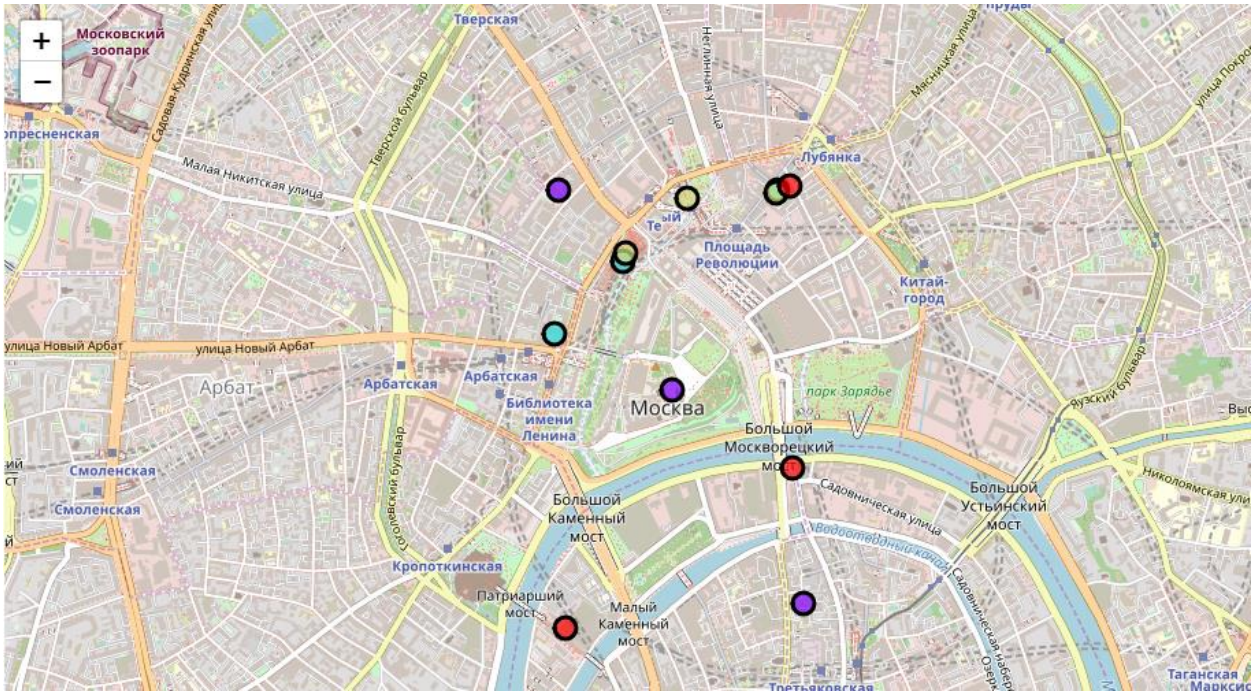
Analyses can be performed for any city. Venues for each category are clustered and then cluster centers from 3 top clusters in each category are shown. Clusters must have at least 5 venues in it to be shown on the map/

Results

Cities are clustered into 5 categories that were roughly labeled as “Eat”, “Shop”, “See” (Entertainment objects), “Dance” (Nightlife) and balanced.

Category	Cities
Eat	Guangzhou, China, Ho Chi Minh City, Vietnam, Macau, Seoul, South Korea, Shanghai, China, Shenzhen, China, Taipei, Taiwan
Shop	Antalya, Turkey, Istanbul, Turkey, Las Vegas, USA, Milan, Italy, Moscow, Russia, New York City, USA, Paris, France, Prague, Czech Republic, Sofia, Bulgaria
Dance	Bangkok, Thailand, Barcelona, Spain, Hong Kong, China, Pattaya, Thailand, Singapore
See	Beijing, China, Florence, Italy, Rome, Italy, Tokyo, Japan, Venice, Italy
Balanced	Amsterdam, Netherlands, Berlin, Germany, Budapest, Hungary, Dubai, United Arab Emirates, Johannesburg, South Africa, Kuala Lumpur, Malaysia, Lima, Peru, London, United Kingdom, Los Angeles, USA, Mecca, Saudi Arabia, Miami, USA, Orlando, USA, Phuket, Thailand, Vienna, Austria

Fore the second part of the analyses I took Moscow as an example. I live in Moscow so I can say that the picture does make sense:



Discussion

This was great exercise on getting, manipulating and visualizing geospatial data. The results are somewhat meaningful from practical point of view, yet there are many concerns that need to be solved before using this analysis in a real life – especially city clustering for touristic business.

- Foursquare data is heavily biased, and it is difficult to estimate magnitude the bias. For one API call returns some list of venues selected by some proprietary foursquare algorithm
- Secondly, Foursquare data are crowdsourced and principles of entering venues data are not consistent across cities.

- Quantity or especially ratio of venues is not a good indicator of anything. The analysis could be improved by counting only exceptionally high rated venues in each city. Unfortunately, rating is only returned by a separate query, individually for each venue
- In general domain knowledge is required to come up with meaningful set of features for city clusterization.

Conclusion

As per above, nice exercise on Data Science and indicates what can be used as a starting point for real-life analysis.