Data sources for Data Science problem of analyzing cities as a collection of venues:

For the venues, the collected data are from the widely used location-based Social Network (LBSN) Foursquare. Users of Foursquare "check-in" to their current location on their mobile device by selecting it from a list of nearby named venues. Their check-in is then broadcast to their social connections. The venue information is easily accessible through a public API, and all venues are annotated with categories of different granularities which represent a natural semantic grouping for venues.

The best source for neighborhood data for USA cities is Zillow at <http://www.zillow.com/howto/api/neighborhood-boundaries.htm> Unfortunately, it does not match our purpose since does not embrace non-US cities. So, to get neighborhood allocation data (but Hong Kong), we parse files from "Airbnb Data Collection" of Tom Slee on the web site: <http://tomslee.net/airbnb-data-collection-get-the-data>. These data are collected in August 7, 2017. To obtain neighborhood-related information on Hong-Kong, we parse Wikipedia page "Hong Kong" on <https://en.wikipedia.org/wiki/Hong_Kong>

Discretization of venues into categories is performed naturally by using either the principal or secondary categories of the venues, depending on which analysis we perform. There are 9 principal venue categories and 231 secondary venue categories.

In total, once this processes was complete, the final data set consisted of near 400 000 venues from 6 cities, with the best represented city, New York, having near 70 000 venues.