2 Data Acquisition and Cleaning

2.1 Data Sources

All data related to Rio's neighborhoods was acquired in the following wikipedia article: `https://pt.wikipedia.org/wiki/Lista_de_bairros_do_Rio_de_Janeiro_por_IDH`. The venues information were acquired using the Foursquare api and the coordinates of each neighborhoods were acquired using GeoPy library.

2.2 Data Cleaning

The data cleaning consisted basically in removing the rows of the data frame found in the article that weren't neighborhoods, removing the neighborhood Mangueira, whose coordinates weren't correctly found by GeoPy, removing the duplicates in the venues data frame and removing the neighborhoods with no venues located by the Foursquare api for the cluster analysis. Also, the HDI values needed to be converted into float type and the columns name were translated to English.

2.3 Feature Selection

From the data frame of the article, only the neighborhoods' names and the HDI values were selected, it was also necessary to separate the rows containing multiple neighborhoods (the ones with the same HDI). After that, GeoPy was used to gather the coordinates of each neighborhood. Having the names of the neighborhoods and the coordinates, it was possible to find the venues in a 1km radius using the Foursquare api.