Anna Perelman
IBM Data Science
Professional Certificate
Coursera.com

Applied Data Science Capstone Project

28th May 2020

1. Introduction

This capstone project concludes a series of data science and machine learning courses provided by IBM company via Coursera.com platform and gives its participants an opportunity to come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of their choice or to come up with a problem that can use the Foursquare location data to be solved using data analysis techniques presented within the specialization.

I decided to implement the following set up: a well-known big chain of private schools would like to open a new branch in San Francisco, CA. They would like to receive recommendations on the best neighborhood to start with. Their branch usually includes a preschool/daycare unit for ages 6 months to 5 years, and an elementary+middle school unit for grades K-8.

In order to provide this theoretical audience with an appropriate set of recommendations it seems wise to collect some data on the usual set of parameters used by this particular chain and its competitors to decide on the best location for their branches. Let's suppose that after some investigation and discussion with the client it was decided to take into account each neighborhood's population, a number of schools already existing within its borders, its safety level and a number of recreational areas (parks and playgrounds) in the proximity. This layer of information should be then combined with the data on different numbers and categories of venues derived from Foursquare, which will give the audience a wide variety of factors to be taken into account for their decision on where to open a new branch.

To help my theoretical client with this decision an unsupervised machine learning clustering technique will be used along with colorful maps of neighborhoods and clusters, so they can analyze and compare groups of neighborhoods, and then zoom in into particular ones.

.

2. Data

This analysis is based on several freely accessible data sources:

a. Information on San Francisco, CA neighborhoods with their zip codes and population can be found at http://www.healthysf.org/bdi/outcomes/zipmap.htm. This simple data set is a table with three columns showing a list of neighborhoods, their zip codes and their population data as can be seen at the following screenshot:

Zip Code	Neighborhood	Population (Census 2000)
94102	Hayes Valley/Tenderloin/North of Market	28,991
94103	South of Market	23,016
94107	Potrero Hill	17,368
94108	Chinatown	13,716
94109	Polk/Russian Hill (Nob Hill)	56,322
94110	Inner Mission/Bernal Heights	74,633
94112	Ingelside-Excelsior/Crocker-Amazon	73,104
94114	Castro/Noe Valley	30,574
94115	Western Addition/Japantown	33,115
94116	Parkside/Forest Hill	42,958
94117	Haight-Ashbury	38,738
94118	Inner Richmond	38,939
94121	Outer Richmond	42,473
94122	Sunset	55,492
94123	Marina	22,903
94124	Bayview-Hunters Point	33,170
94127	St. Francis Wood/Miraloma/West Portal	20,624
94131	Twin Peaks-Glen Park	27,897
94132	Lake Merced	26,291
94133	North Beach/Chinatown	26,827
94134	Visitacion Valley/Sunnydale	40,134
All Zips	(all of San Francisco, including very small population zips, such as Treasure Island or the Presidio, which are not listed above)	776,733

To obtain this table in a format enabling its data analysis using Python language web scraping techniques can be used (BeautifulSoup library is used in this project).

b. In order to provide geographical maps of the neighborhoods latitude and longitude figures should be added to the above table from Python package *uszipcode*.

c. The information on San Francisco schools with their zip codes, public/private type and grade levels can be found at: https://data.sfgov.org/Economy-and-Community/Schools/tpp3-epx2. This dataset consolidates Infant, Pre-K, and K-14 education points for facilities both public and private, and can be freely downloaded in a variety of formats (CSV format is used in this project).

d. The information on the number of crimes committed in San Francisco along with their latitude and longitude to be converted into zip codes can be found at: https://www.kaggle.com/psmavi104/san-francisco-crime-data#Police_Department_Incident_Reports__2018_to_Present.csv. The data set contains all the crimes listed in the public database from 2018 to present. The data can be downloaded in CSV format for further analysis.

e. The information on parks and playgrounds located in SF neighborhoods can be found at https://www.kaggle.com/san-francisco/sf-recreation-park-department-park-info-dataset. This data set lists all parks, playgrounds, and stadiums within SF borders.
f. The data on different venues located in each neighborhood can be derived from Foursquare location data. Using a free developer account their Places API offers real-time access to Foursquare's global database of rich venue data and user content to power location-based experiences in an app or a website.