Moving to Silicon Valley? Explore the Area by Data Science Methods

1. Introduction

1.1. Description and Discussion of the Background

Silicon Valley is a region in southern part of San Francisco Bay Area in Northern California that is considered as a global center for high technology and innovation. Most of Silicon Valley is in Santa Clara County, although it slightly extends to the neighboring counties such as San Mateo and Alameda Counties. The focus of this project is to explore Santa Clara County, the center point of Silicon Valley.

Many of the most well-known technology companies such as Google, Apple, Microsoft, Intel, IBM, Facebook, eBay, Cisco, Qualcomm, Texas Instruments, Netflix, and Oracle reside in Santa Clara County. Economic prosperity, world class universities, cultural diversity, a support system for entrepreneurs, and a rich engineering tradition attracts many tech talents to the region. The region's economy is heavily service based. Although technology, both hardware and software, dominates the service sector by value, Santa Clara County has its share of retail and office support workers.

In this project, I used GitHub repository as a database. I utilized geopy to convert addresses into latitude and longitude values. I used Foursquare API to explore cities in Santa Clara County specifically to retrieve the most common venue categories in each city and then used this feature to group cities into clusters. This task was completed using k-means clustering algorithm. I also used python Folium library to visualize cities and their respective clusters. Finally, I used open datasets to analyze demographics, economics, and other social aspects in Santa Clara County.

The main audience of this project are people who come to live and work in Santa Clara County. People consider many factors when thinking about a new destination to call home. Such factors include amenities such as popular venues in each city, demographics, education, housing, and school rating. This project uses a data science approach to explore these factors. In addition to new residents, home buyers, investors, policy makers or anybody interested in learning about Santa Clara County can benefit from this project.

1.2. Datasets and Toolkits

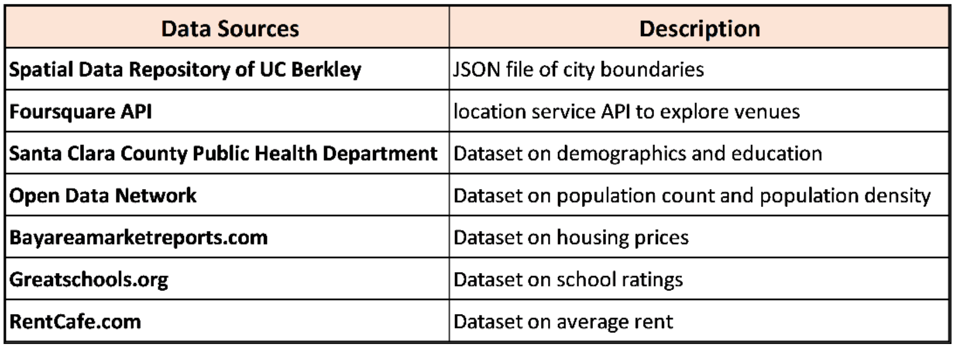
1.2.1. Prerequisite

This project requires a Foursquare developer account. To create an account, go to https://developer.foursquare.com/

1.2.2. Datasets

Table 1 lists data sources that were used in this project along with their description.

Table 1 List of data sources that are used in this project.

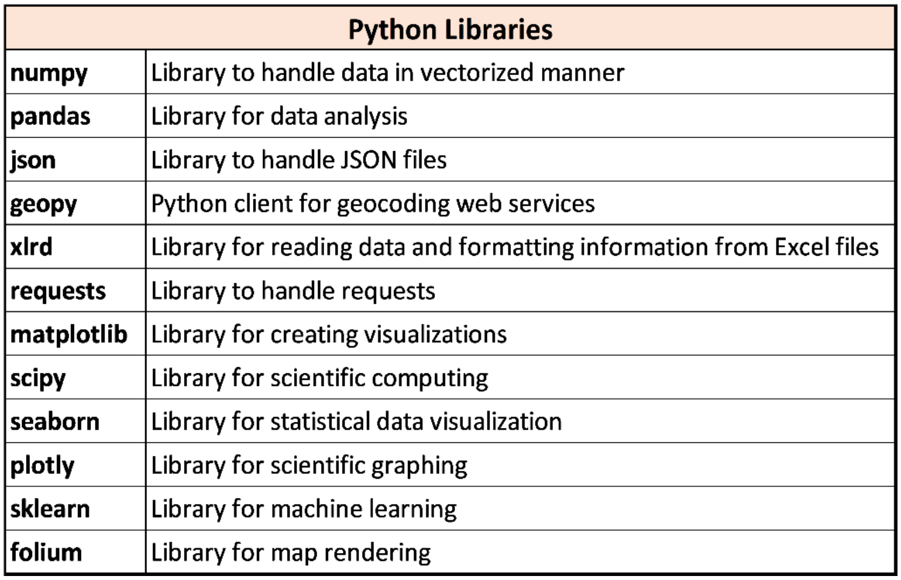


Data Sources	Description
Spatial Data Repository of UC Berkley	JSON file of city boundaries
Foursquare API	location service API to explore venues
Santa Clara County Public Health Department	Dataset on demographics and education
Open Data Network	Dataset on population count and population density
Bayareamarketreports.com	Dataset on housing prices
Greatschools.org	Dataset on school ratings
RentCafe.com	Dataset on average rent

1.2.3. Python Libraries

Table 2 shows a list of Python libraries used in this project.

Table 2 List of python libraries used in this project.



Python Libraries	
numpy	Library to handle data in vectorized manner
pandas	Library for data analysis
json	Library to handle JSON files
geopy	Python client for geocoding web services
xlrd	Library for reading data and formatting information from Excel files
requests	Library to handle requests
matplotlib	Library for creating visualizations
scipy	Library for scientific computing
seaborn	Library for statistical data visualization
plotly	Library for scientific graphing
sklearn	Library for machine learning
folium	Library for map rendering