Applied Data Science Capstone Project

Week 1: Introduction/Business Problem and Data

Identifying the Best Neighborhoods to Move to in Chicago

Introduction/Business Problem

Moving to a new city is exciting but at the same time can be a little bit of scary. Depending on the family size, age, style of living, etc. considerations and priorities differ when choosing the best neighborhood to live.

It takes a lot of planning especially for a family with kids to find the best neighborhood to live. Moving is expensive, thus you want to minimize the likelihood of having to move to another neighborhood after your first move to the city.

There are websites that list best neighborhoods in cities, however these are not sufficient in making decisions and don't allow one to weigh each neighborhood based on specific preferences.

In this project, we'll use our data science powers to help a family moving to Chicago to locate the best neighborhoods that match their priorities/preferences. Here are the considerations of this family:

- Neighborhood safety (crime rates)
- School (K-12) rankings
- Proximity to venues such as parks, martial art schools, and music schools
- Housing prices
- Proximity to new job location (which is Downtown Chicago)

After gathering all necessary and up-to-date data, we'll cluster the neighborhoods and provide a list of options to the family based on their interests.

Data

Chicago is one of the largest metropolises in the United States with a population of 2.706 million as of 2018. It has more than 200 neighborhoods and 77 community areas. Community areas are used for city-wide statistical and planning purposes. As most data we need is available at the community area level, we'll start our analysis at the community area level and then focus on the neighborhoods.

There are four main sources of data we'll use in our project:

- Wikipedia
- Chicago Data Portal
- Zillow Housing Data
- Foursquare API

Wikipedia

We'll be scraping data from the following pages:

1. List of Chicago community areas with associated community numbers, population, area, and population density.
   https://en.wikipedia.org/wiki/Community_areas_in_Chicago#cite_note-City_basics-9

2. Relation between Chicago neighborhoods and community areas. Our crime data is available at the community area level and our housing prices are available at the neighborhood level.
   https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago#cite_note-1

Chicago Data Portal

Most of the data we need to conduct our study is available in the "Chicago Data Portal" at https://data.cityofchicago.org. The nice feature of this portal is that it allows access to some datasets via SODA API.
Here's a list of data files available in the Chicago Data Portal:

1. **Current community area boundaries in Chicago:** We can export the data as a geoJson file.
   We'll use this dataset for visualizations.
   https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

2. **Chicago crime data:** This data set displays the incidents of crime reported in Chicago from 2001 to present (minus the most recent 7 days). In addition to details on the nature of the crime, this data set also reports the community area that the crime took place.

   Since this is a huge dataset, we'll use SODA API to filter results to focus on certain columns and years (2019 and 2020), and only download the relevant data.
   https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

   We'll derive crime statistics for the community areas such as total number of crimes, distribution of crime types (theft, assault, etc.). These statistics will be useful in determining safer community areas.

3. **School Progress Report Ratings for Chicago Public schools:** This data set does not report the community area or neighborhood a school belongs to. We'll use this data set together with the Foursquare API. For the schools returned by the Foursquare API in a location of interest, we'll use this data set to extract the progress reports.

   https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/dw27-rash

Zillow Housing Data

Zillow has time series data available for download:

https://www.zillow.com/research/data/

Among several datasets listed here, we will use **ZHVI Single-Family Homes Time Series** data at the **neighborhood level**. If need arises, we can also download this data set at **the zip code level**. ZVHI stands for Zillow Home Value Index. This data set displays typical value in dollars for all single-family homes in a given region. After downloading this data set, we'll extract the entries for Chicago Neighborhoods. In Zillow dataset, there are 200 Chicago Neighborhoods.

Foursquare API

We'll use Foursquare API to search for the venues that the family is interested in:

1. We'll search for the schools in the neighborhood by using the following venue types: Elementary School, Middle School, High School
2. We'll search for the following venues: Park, Martial Arts School, Music School

After gathering all the necessary data, we'll cluster the neighborhoods and determine the best neighborhoods that match family preferences.

Some data examples:

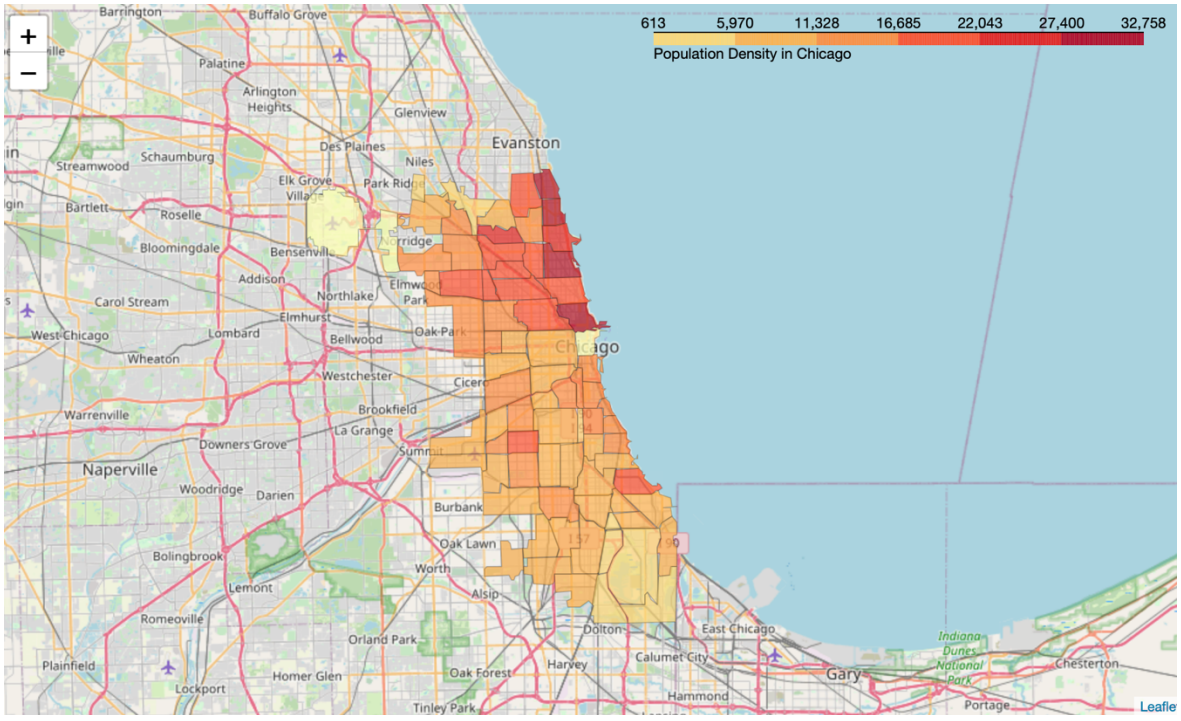
Here's a screenshot of crime data covering years 2019 and 2020:

	date	primary_type	description	community_area	year
0	2019-03-21T16:39:00.000	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	43	2019
1	2019-11-18T08:00:00.000	THEFT	$500 AND UNDER	31	2019
2	2019-12-17T12:00:00.000	MOTOR VEHICLE THEFT	TRUCK, BUS, MOTOR HOME	61	2019
3	2019-12-27T10:27:00.000	NARCOTICS	POSSESS - HEROIN (WHITE)	61	2019
4	2019-12-27T08:01:00.000	NARCOTICS	MANUFACTURE / DELIVER - HEROIN (WHITE)	25	2019

Wikipedia data:

	Community_Number	Community_Name	2017_Population	Area_sq_mi	2017_Population_Density
0	01	ROGERS PARK	55062	1.84	29925.00
1	02	WEST RIDGE	76215	3.53	21590.65
2	03	UPTOWN	57973	2.32	24988.36
3	04	LINCOLN SQUARE	41715	2.56	16294.92
4	05	NORTH CENTER	35789	2.05	17458.05
...	...	...	...	...	...
73	74	MOUNT GREENWOOD	19277	2.71	7113.28
74	75	MORGAN PARK	22394	3.30	6786.06
75	76	O'HARE	12377	13.34	927.81
76	77	EDGEWATER	55965	1.74	32163.79
77	Total	CHICAGO	2722586	227.34[12]	11975.83

Here's a choropleth map of the Chicago community areas using population density in 2017. (Using community area geojson data and Wikipedia data)



Here's a screenshot of Zillow Housing Data:

RegionName	RegionType	StateName	State	City	Metro	CountyName	1996-01-31	...	2020-01-31	2020-02-29	2020-03-31	2020-04-30	2020-05-31	2020-06-30	202
Logan Square	Neighborhood	IL	IL	Chicago	Chicago-Naperville-Elgin	Cook County	154086.0	...	499367.0	503470.0	508541.0	513503.0	516831.0	520610.0	525
Little Village	Neighborhood	IL	IL	Chicago	Chicago-Naperville-Elgin	Cook County	NaN	...	173503.0	173270.0	173594.0	173476.0	175075.0	177939.0	181
West Rogers Park	Neighborhood	IL	IL	Chicago	Chicago-Naperville-Elgin	Cook County	157304.0	...	364883.0	367482.0	370122.0	372114.0	373802.0	375213.0	378
South Austin	Neighborhood	IL	IL	Chicago	Chicago-Naperville-Elgin	Cook County	NaN	...	178077.0	179646.0	181244.0	182633.0	184235.0	186681.0	190
Albany Park	Neighborhood	IL	IL	Chicago	Chicago-Naperville-Elgin	Cook County	146045.0	...	372049.0	373105.0	373382.0	374396.0	375792.0	377530.0	380
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Mount Greenwood Heights	Neighborhood	IL	IL	Chicago	Chicago-Naperville-Elgin	Cook County	130780.0	...	272249.0	274142.0	276192.0	277347.0	278117.0	278338.0	280