House Price Index and Venue Based Location Selection

Capstone Project for Applied Data Science course by IBM/Coursera

Ravindranath Rao

2/2/2020

Capstone Project for Applied Data Science course by IBM/Coursera. Use of House Price Index and Venues by Neighborhood used to find matching retirement location.

Contents

Introduction: Business Problem

The Washington D.C. Metropolitan Area is the metropolitan area centered on Washington, D.C., the capital of the United States. The area includes all of the District of Columbia, and parts of the US States of Virginia, Maryland and West Virginia. Northern Virginia (locally referred to as NoVA or NOVA) comprises cities and counties in Virginia surrounding Washington D.C.

It is the most populous region of Virginia and Washington Metropolitan area. There are about 3.5million people working in various jobs in this area.

DC is home to all types of amenities and attractions. World famous Smithsonian Museums are based in D.C. D.C. Metro Area boasts of hundreds of Performing Arts/theaters, public transportation, metro rail systems, Convention Halls, and Resorts etc... List is endless.

When people want to retire, they move out of DC area to have a relaxed retirement life. Many though, want to live close to DC with similar amenities and if desired, make a trip to D. C. area.

According to Kiplinger Personal Finance magazine [1], in its list of ’12 Smart Places to Retire’ lists Richmond, Virginia as one of the top in the list.

When people move, they want to make sure place has needed amenities and can move out quickly, if it does not suit them. House Price Index (HPI) measures the price changes of residential housing. HPI is a weighted, repeat sales index, measuring average price changes in repeat sales or refinancing on the same properties [2].

HPI can be used to evaluate whether place is worth moving in without incurring a huge financial loss. The project tries to cluster Richmond area neighborhoods with Northern Virginia neighborhoods, based on Venues. Then, show areas with HPI and Venue along with cluster information, for making a informed decision about where to move within Richmond.

DATA

Following data will be used: For ease of identifying neighborhood, each Zip Code is considered as a Neighborhood. Thus, a file with Zip code and its longitude and latitude is needed for use with Foursquare API. US state of Virginia specific zipcode data is downloaded from website [3]

The file is edited to have just Zip, City, State, Latitude and Longitude, values for places in Northern Virginia (NOVA) and Richmond, VA , in a CSV format and saved as NOVA_city_richmond_selected.csv.

House Price Index is available by Zip Code at web site [4]. Downloaded Five-Digit ZIP Codes (Developmental Index; Not Seasonally Adjusted) XLSX file.

The file has following columns: Five-Digit ZIP Code, Year, Annual Change (%), HPI, HPI with 1990 base, HPI with 2000 base.

The downloaded file was filtered for data of Zip Code from 20001 thru 24000, year=2018 and created a CSV file HPI_AT_BDL_ZIP5_2018_VA.csv with following column : Five-Digit ZIP Code, Year, Annual Change (%),HPI.

NOVA_city_richmond_selected.csv fil merged with HPI_AT_BDL_ZIP5_2018_VA.csv file by Zip Code. This will make sure any combined data will have both geo-coordinates and HPI values. The combined data will be stored in pandas dataframe as nova_data.

Methodology

Using nova_data, call FourSquare API [6] to get Venue details for each Zip Code (renamed as Neighborhood for consistency with Battle of Neighborhood theme).

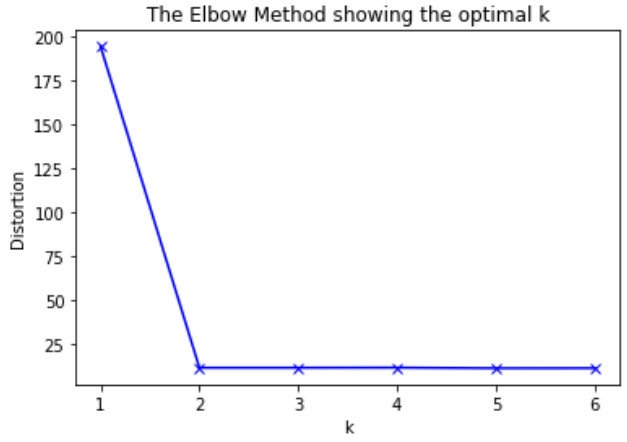
Once Venue data is obtained, do following steps:

1. Identify top 10 Venues for each Neighborhood.
2. Normalize the data,
3. Use K-Mean clustering to cluster the data
4. Show cluster on a folium map.
5. Create range for HPI

6. Show HPI data, cluster data on a Folium map

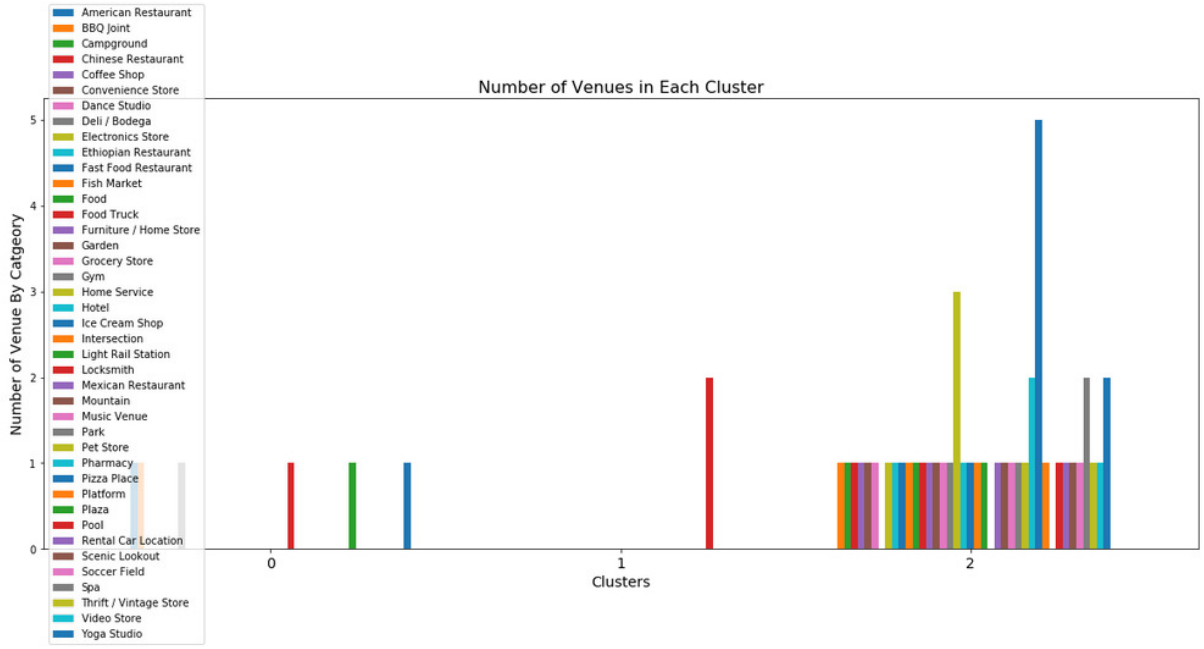
Result

The Elbow method is used to identify the optimal cluster value for K-Mean.



Above chart shows that cluster size of **2** is good enough. Running K-Mean with clusters=3 will clearly shows the outliers.

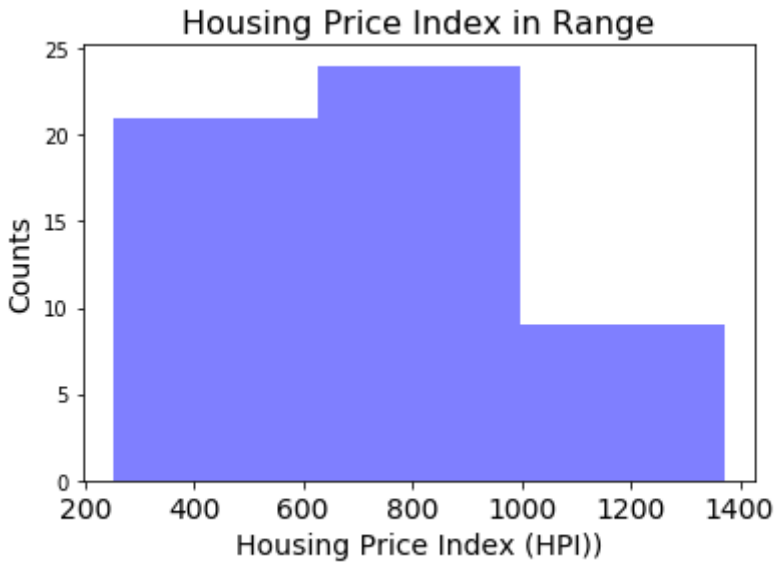
By checking on Venue type and their concentration/count , a label is assigned to the clusters.



When we examine above graph we can label each cluster as follows:

- Cluster 0 : "Good Locality " Based on high percentagee of increase and good HPI
- Cluster 1 : "Limited choice"
- Cluster 2 : "Moderate " Based on lower perentagee of increase and good HPI

Similarly, House Price Index (HPI) range is identified by plotting a histogram as shown below.



As it seems in above histogram, we can define the ranges as below:

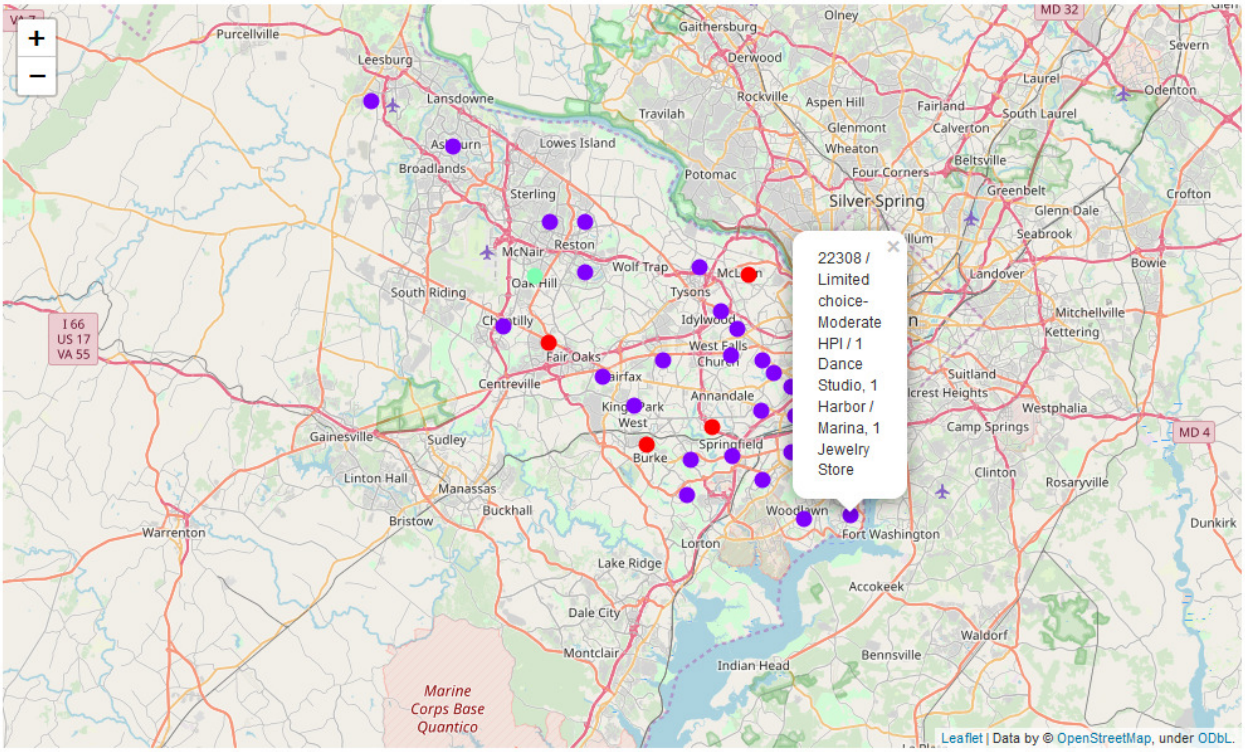
- 200-600 HPI : "Low HPI"
- 600-1000 HPI : "Moderate HPI"
- 1000-1400 HPI : "High HPI"

In this case, we can create **"Level_labels"** with those levels.

We can also identify top 3 venues for each neighborhood. Sample is given below:

	Neighborhood	Top_3
0	20147	1 BBQ Joint, 1 Deli / Bodega, 1 Furniture / Ho...
1	20151	2 Furniture / Home Store, 2 Gym / Fitness Cent...
2	20170	1 Food Truck, 1 Playground
3	20171	1 Pool
4	20175	1 Athletics & Sports, 1 BBQ Joint, 1 Construct...

Folium map is created with markers on Neighborhood, cluster labels and HPI related labels as shown below:



By clicking on each circle mark, we can see the Neighborhood details.

Discussion

Most of the neighborhood selected belongs to cluster 3. Richmond, VA also has much area that are part of cluster 3. This shows that it is easier for people in Northern Virginia to retire in Richmond as similar amenities are available in Richmond. Also, there are many areas in Richmond with positive price change and higher HPI.

To build more clusters, additional attributes like house price, specific amenities like transit services, night life, golf courses etc.. can be considered, based on individual preferences.

The result clearly shows that clustering techniques can be easily and accurately used to identify the place to retire.

Conclusion

When people retire, they try to move to a place where cost of living is less, any tax break and most importantly, amenities. People get accustomed to certain life style and it is difficult to

change immediately. Moving to closer place which meets the needs is one option. For people in NoVA area, Richmond, VA, seems to match the criteria mentioned.

The NoVA and Richmond area resemble each other. This is shown by the optimum cluster value being just 2. It also within 100 miles from NoVA. Retirees can get best of both places.

Main purpose of the project is to help retirees. The methodology can also be used by service provider to develop web page that allows filtering based on HPI. The criterion of retirement place selection varies from individual to individual.

Reference

Note: The web address is shown in parenthesis, instead of hyperlink. This will help, in case hyperlink does not work on pdf version.

[1] [Kiplinger Finance Magazine](https://www.kiplinger.com/slideshow/retirement/T006-S002-12-smart-places-to-retire/index.html)

[2] [HPI Definition](https://www.investopedia.com/terms/h/house-price-index-hpi.asp)

[3] [Virginia zipcode level geo spatial data](https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/)

[4] [FHFA House Price Index](https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx#qpo)

[5] [NOVA Wiki](https://en.wikipedia.org/wiki/Northern_Virginia)

[6] [Washington Metroploitan Area Wiki](https://en.wikipedia.org/wiki/Washington_metropolitan_area)

[7] [Forsquare API](https://developer.foursquare.com/docs)

Thank You

The project work is immensely helped by excellent training Notebook, titled "Segmenting and Clustering Neighborhoods in New York City" created by Alex Aklson (https://www.linkedin.com/in/aklson/) and Polong Lin (https://www.linkedin.com/in/polonglin/), which was part of **Coursera** course **Applied Data Science Capstone**.

Also reviewed, blogpost example provided [Blogpost:](https://cocl.us/coursera_capstone_blogpost)