**Neighborhood Clustering – Virginia Beach, VA**

# Introduction

## Background

Virginia Beach is an independent city located on the southeastern coast of the Commonwealth of Virginia in the United States. It is the most populous city in Virginia and the 44th most populous city in the nation. Located on the Atlantic Ocean at the mouth of the Chesapeake Bay, Virginia Beach is included in the Hampton Roads metropolitan area. This area, known as "America's First Region", also includes the independent cities of Chesapeake, Hampton, Newport News, Norfolk, Portsmouth, and Suffolk, as well as other smaller cities, counties, and towns of Hampton Roads.

Virginia Beach is a resort city with miles of beaches and hundreds of hotels, motels, and restaurants along its oceanfront. A 3-mile boardwalk stretches along its beach-lined oceanfront. The bayside First Landing State Park marks the 1607 arrival of the Jamestown colonists from England. The Virginia Aquarium & Marine Science Center exhibits ocean life including sharks, rays and sea turtles in globally themed habitats

## Neighborhood clustering

Neighborhood clustering is an important process by which the different neighborhoods in a city are grouped together based on various parameters like property price, safety, venues of interest etc. Data from various sources will be combined and aggregated to cluster the neighborhoods based on the parameters they are good at as well as the parameters they need to improve on.

Once the neighborhoods are classified based on features people will be able to make informed decisions about which neighborhood to choose based on their priorities

For instance, budget will be a crucial factor for some while availability of good schools will be a detrimental factor for some others and so on.

# Interest

The process of clustering the neighborhoods helps people who are moving into the city or looking to buy a home in the city to understand the city in detail and the livability of each neighborhoods in the city. They will be able to gauge based on the data

* How costly a home would be in an area or what are the different areas within the city where they can find homes in comparable budget.
* What are the neighborhoods that have more choices of schools or easy access to healthcare?
* What are the neighborhoods which have the lowest crime rate?
* Neighborhoods having an acceptable threshold for more than one of the desired parameters

# Data acquisition, cleaning and exploratory data analysis

The neighborhood clustering for the city of VA Beach in this project is done by using data from multiple sources

|  |  |
| --- | --- |
| Parameter | Source |
| Home value index(property price ranges) | zillow.com(<https://www.zillow.com/research/data/>) |
| Safety (number of crime related incidents reported to local law enforcement) | Virginia Beach city electronic police reports online - <https://eprodmz.vbgov.com/MainUI/Crimes/CrimeSearch.aspx> |
| Availability of schools | Foursquare |
| Access to health care (hospitals, urgent care, Doctor’s office etc.) | Foursquare |
| Category and frequency of different venues in and around neighborhood | Foursquare |
| Zip code data | Virginia Beach city property sales data, VB city website <https://s3.amazonaws.com/vbgov-ckan-open-data/Property_Sales.csv> |

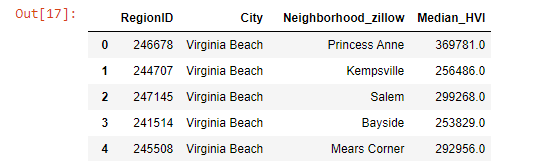
## Step 1 - Neighborhood Zip code data analysis and mapping

Zillow research data was downloaded and used in the project as the base data for list of neighborhoods and the respective median Home Value Index. Since neighborhood classification and zip code data was not readily available for VA Beach an alternative source also was used in conjunction with Zillow data to map the neighborhood data in the zip code. Below steps were performed to do the mapping

* + 1. Read the property sales csv into a pandas data frame.
    2. The neighborhood name and 5-digit zip code were extracted into another DF
    3. Since property sales DB was large with several years of transactions there were multiple zip codes entered for one neighborhood. This could be due to incorrect zip code or in some cases due to the borderline location of properties.
    4. In order to improve accuracy, the occurrence count was taken for each neighborhood and zip code. Resulting DF looked like below –



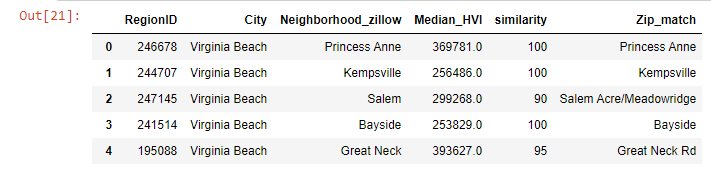
* + 1. Next the latest Zillow research data was downloaded and read into a Pandas DF. Most recent HVI and neighborhood name was extracted



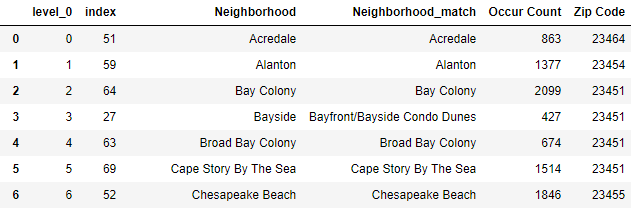
* + 1. Since the sources were different the neighborhood names were different too. In order to improve the accuracy fuzzy matching techniques was used between the property sales csv and Zillow data to identify the correct neighborhood and zip code
    2. This was done with a package called Fuzzywuzzy.



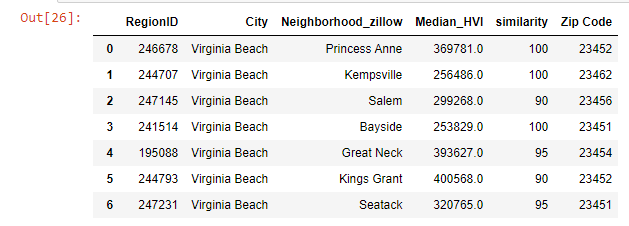
* + 1. For each neighborhood in Zillow data a fuzzy matching was done in the property sales csv to find the neighborhood name same or most matching with Zillow name. a new column was added to indicate the most matching neighborhood name from property sales csv. A similarity score of 89 was selected based on data match results



* + 1. Next step the matching neighborhood name is iterated through the zip code occurrence DF to identify the zip code that occurs most times for the neighborhood and that zip code is assigned to the neighborhood

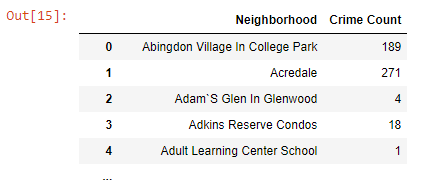


This DF is merged with the Zillow DF to get the neighborhood, Zip code and the HVI

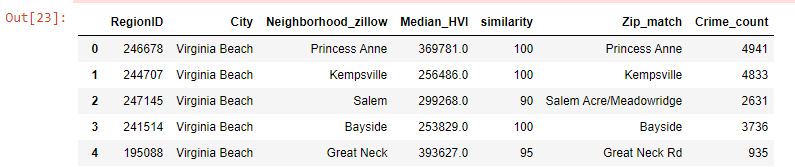


## Step 2 Crime data acquisition and analysis

* + 1. Crime data from VB police reports website was downloaded. The rows in the text file was split and loaded into a pandas DF.
    2. A neighborhood-wise summary count was taken to obtain a result like below –



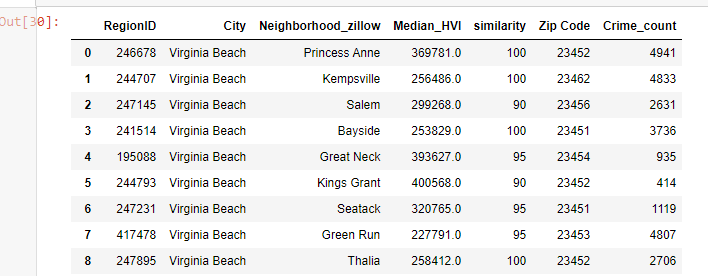
* + 1. Next step each neighborhood in Zillow DF is searched in the crime count summary data frame and when the name matches the total crime count is updated by adding the crime count from each matching neighborhood in the crime data frame. Final DF will be like below



* + 1. Any non-matching neighborhood is done a fuzzy matching against the crime DF and crime counts updated

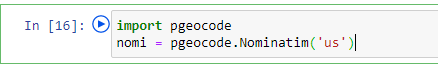
## Merge Home value and crime data frames

* + 1. Dataframes obtained from a and b above are merged to obtain the below dataframe with the data elements –
* City
* Neighborhood
* Median Home Value Index (HVI)
* Zip Code
* Crime count

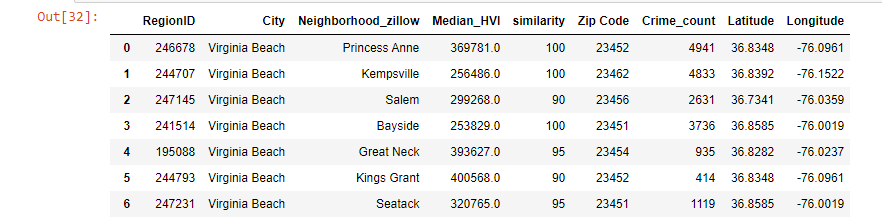


## Using Foursquare for Venue Data

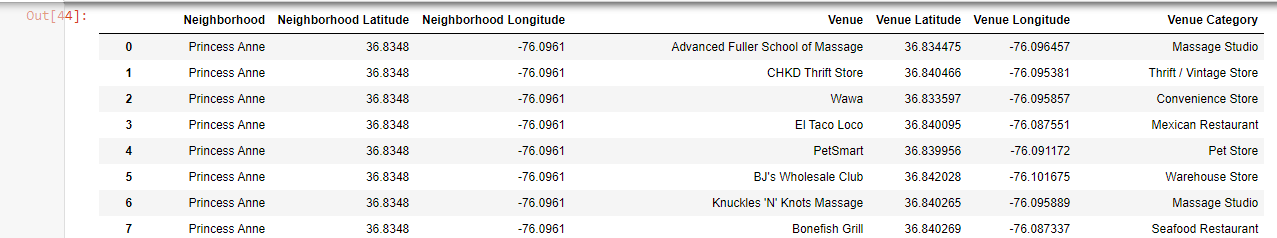
* + 1. Next step is to use Four Square API to obtain venue data in and around each neighborhood. For using the Foursquare the latitude and longitude data for each neighborhood need to be fetched. For this a package called pgeocode is imported



* + 1. Each of the zip code passed to nominatim to obtain the latitude and longitude and merged with the data frame from c)



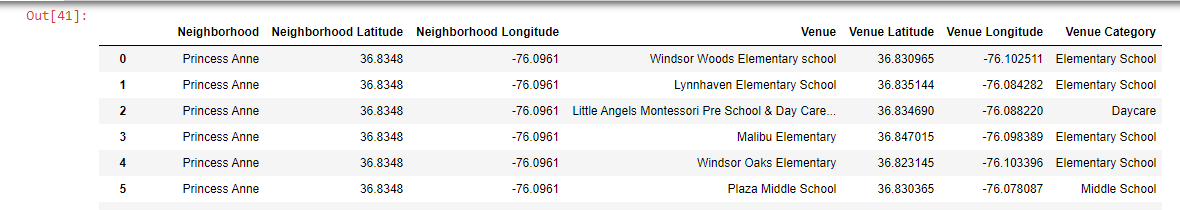
* + 1. After obtaining the latitude and longitude a series of Foursquare calls are made to obtain all the venues in a 2.5 mile radius with maximum limit 100 venues and all the venues are stored in a dataframe like below



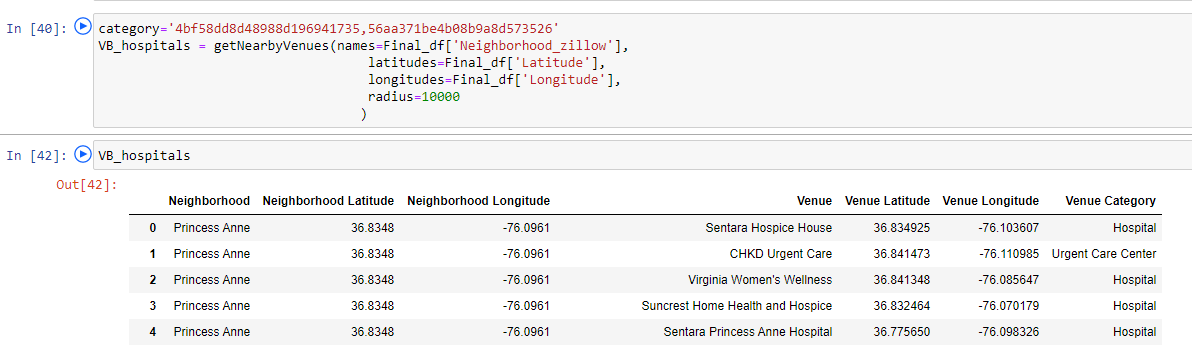
## Using Foursquare for school and hospital data

* + 1. Foursquare calls are made passing specific category codes to obtain the list of primary schools, middle schools, private schools etc.



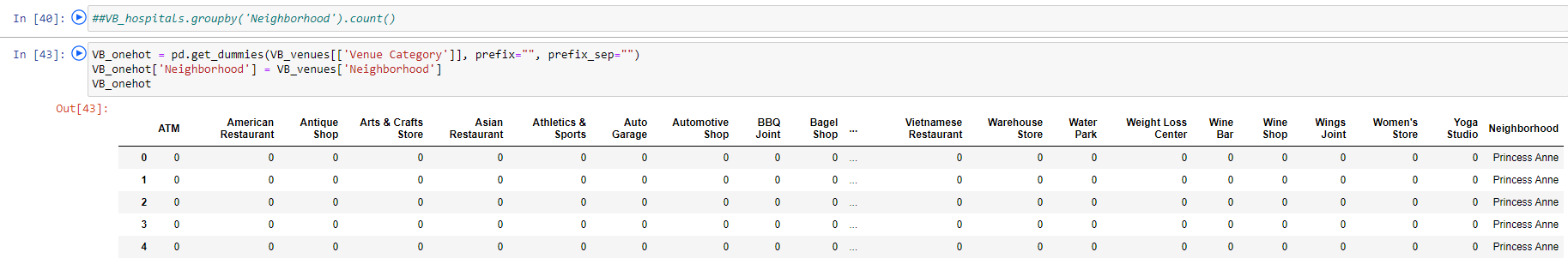


* + 1. Next Foursquare calls are made using hospital category codes

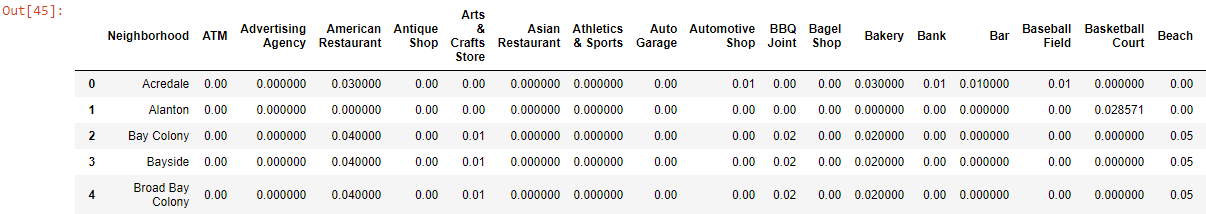


# Preparing data for Clustering

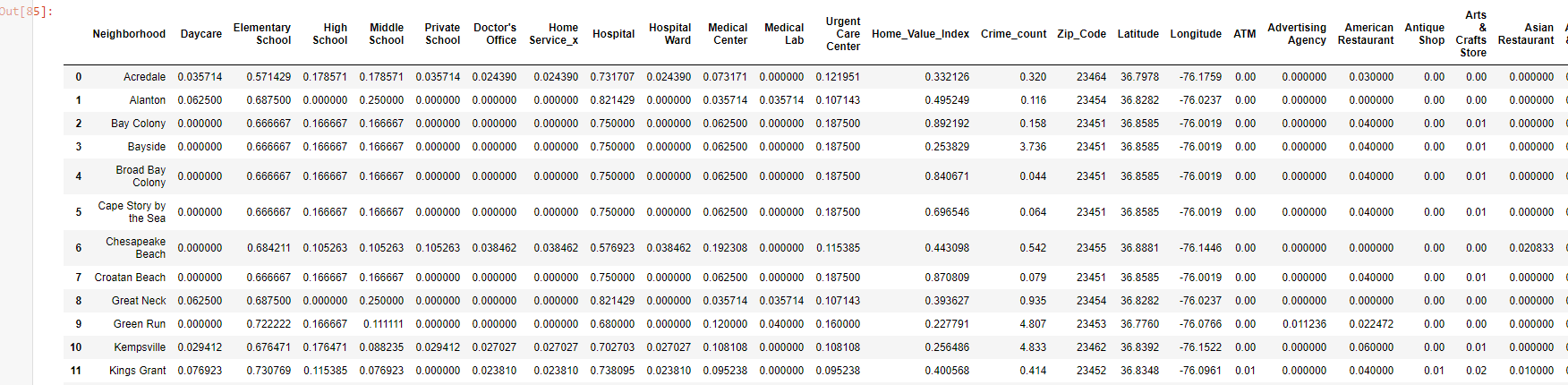
* + 1. In order to do the clustering as a first step all the venues are given a weighted average for each Neighborhood based on their frequency of occurrence in each neighborhood. This is done by applying one hot encoding through get dummies function in pandas. One hot encoding returns a 0 or 1 for each row in the venues dataframe based on what venue category that row represents



* + 1. After one hot encoding a neighborhood level grouping and averaging is done to obtain a value representing the frequency for each neighborhood per venue category. As the value gets closer to 1 it indicates a better availability or higher frequency of that particular venue



* + 1. Similar one hot encoding and grouping is done for school and hospital data as well
    2. Finally, all the below data elements are merged into a single dataframe
* Neighborhood name
* Zip code
* Home Value Index
* Crime count
* School data encoded and averaged
* Hospital data encoded and averaged
* Venue data encoded and averaged



* + 1. Additionally, a separate DF is created with each neighborhood and top 10 venues for that neighborhood. After clustering the results will be displayed using this DF instead of the grouped DF with all venues



# Running K means Clustering algorithm

* + 1. Important step in K means algorithm is identifying the right value of K. Here a random trial and error method has been used to run the model multiple times to get the best value of K.
    2. Starting value of 5 was assigned to K and then the values of the different parameters (HVI, crime count, school and hospital data) observed.
    3. Then model was re-ran with the value of K was changed to 9 with an increment of 1 in each step. The value of 8 yielded better results. Other values gave more clusters with only one neighborhood or clusters which had nonmatching values with significant difference. Detailed results attached



# Discussion of Results

Based on the value of K = 8 clusters were identified to cluster the neighbor hoods in Virginia Beach. Below is an overview of the different clusters with some of the key parameters highlighted. Please note the clustering has been done based on many different categories of venues and their scores while only 10 venues have been displayed in the report. Also schools have been searched with a 5000 m radius and hospitals with a 10000m radius

Cluster 0 –

Home Value mostly ~230K-300K with low crime rates around 2700 with similar school and hospital availability

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Oceana | 0.6875 | 0 | 0 | 0.821429 | 231062 | 2686 | Discount Store |
| Salem | 0.545455 | 0.090909 | 0 | 0.785714 | 299268 | 2631 | Golf Course |
| Thalia | 0.730769 | 0.115385 | 0.02381 | 0.738095 | 258412 | 2706 | Convenience Store |

Cluster 1 –

Home Value mostly 700K-890K with very low crime rates around 50-150, similar school and hospital availability due to same zip code. Highest in terms of HVI due to safety as well as good school and hospital availability

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Bay Colony | 0.666667 | 0.166667 | 0 | 0.75 | 892192 | 158 | Seafood Restaurant |
| Broad Bay Colony | 0.666667 | 0.166667 | 0 | 0.75 | 840671 | 44 | Seafood Restaurant |
| Cape Story by the Sea | 0.666667 | 0.166667 | 0 | 0.75 | 696546 | 64 | Seafood Restaurant |
| Croatan Beach | 0.666667 | 0.166667 | 0 | 0.75 | 870809 | 79 | Seafood Restaurant |

Cluster 2 –

Home Value 340K with crime rate 5400

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Pembroke | 0.676471 | 0.176471 | 0.027027 | 0.702703 | 341492 | 5400 | American Restaurant |

Cluster 3 –

Home Value mostly 330K-500K with low crime rates 100-550, similar school and hospital availability

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Acredale | 0.571429 | 0.178571 | 0.02439 | 0.731707 | 332126 | 320 | Sandwich Place |
| Alanton | 0.6875 | 0 | 0 | 0.821429 | 495249 | 116 | Discount Store |
| Chesapeake Beach | 0.684211 | 0.105263 | 0.038462 | 0.576923 | 443098 | 542 | Pizza Place |
| Kings Grant | 0.730769 | 0.115385 | 0.02381 | 0.738095 | 400568 | 414 | Convenience Store |
| Rudee Heights | 0.666667 | 0.166667 | 0 | 0.75 | 380966 | 242 | Seafood Restaurant |
| Thoroughgood | 0.684211 | 0.105263 | 0.038462 | 0.576923 | 414709 | 362 | Pizza Place |

Cluster 4 –

Home Value mostly 230K-370K with crime rates around 4800-5000, similar school and hospital availability

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Green Run | 0.722222 | 0.166667 | 0 | 0.68 | 227791 | 4807 | Gym / Fitness Center |
| Kempsville | 0.676471 | 0.176471 | 0.027027 | 0.702703 | 256486 | 4833 | American Restaurant |
| Princess Anne | 0.730769 | 0.115385 | 0.02381 | 0.738095 | 369781 | 4941 | Convenience Store |

Cluster 5 –

Home Value mostly ~250K with crime rates around 3200-3500, similar school and hospital availability

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Bayside | 0.666667 | 0.166667 | 0 | 0.75 | 253829 | 3736 | Seafood Restaurant |
| Princess Anne Plaza | 0.730769 | 0.115385 | 0.02381 | 0.738095 | 245260 | 3265 | Convenience Store |

Cluster 6 –

Home Value mostly 320 - 390K with crime rates around 1000, similar school and hospital availability

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Great Neck | 0.6875 | 0 | 0 | 0.821429 | 393627 | 935 | Discount Store |
| Seatack | 0.666667 | 0.166667 | 0 | 0.75 | 320765 | 1119 | Seafood Restaurant |

Cluster 7

Only one neighborhood with Home Value ~700K and crime count ~700

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Elementary School** | **High School** | **Doctor's Office** | **Hospital** | **HVI** | **Crime Count** | **1st Most Common Venue** |
| Sandbridge | 0.545455 | 0.090909 | 0 | 0.785714 | 705299 | 713 | Golf Course |

# Conclusion

In this project I gathered data on city of Virginia Beach, VA, USA and clustered the neighborhoods based on some of key parameters indicating the livability of the neighborhood. Clustering was done based on the cost of home ownership, availability of schools, hospitals, safety, and venues in and around neighborhood. Clusters indicate the neighborhood with higher safety and higher cost of home ownership as well neighborhoods with good school and hospital availability. This type of clustering helps people to understand the neighborhoods in detail and what to expect when buying a home in the nrighborhood