# Data Science Capstone Project Report

Title

# Business Location Recommendation in Richmond City, Virginia using FourSquared API & K-means clustering

By

Harry Jose Manjiyil

IBM Coursera Data Science Professional Student

# Description of the Problem & Background

A nationwide ‘Pool Equipment Supplies’ company is looking to open their first store in the city of Richmond, Virginia, USA. The company has decided to take a data-based approach in selecting the location of their store. The company has requested a data science project to give recommendations on where to start their new store. From the preliminary analysis of the counties around the Richmond City, the company executives have narrowed down to 3 potential candidate counties for them to start the store. The data science project will focus on analyzing the locations within these 3 counties and come up with the recommendations and reasoning behind the recommendations. The counties selected are Henrico County, Chesterfield County and Richmond City County which are adjacent counties around Richmond downtown metro area.

The data science project aims to build a model that recommends a business entrepreneur the exact geographical location to start his business in the above said 3 counties of Richmond, VA given a few parameters. The code will be built to make recommendation specifically for a Pool Equipment Supplies company but minor changes to some parameters can be made to make it more generic to any competing business categories currently in the 3 counties of Richmond.

The company has requested a ranking of the locations within these 3 counties in Richmond city where they can start their Pool Equipment Supplies Store. The company also has requested a visualization of the data of the locations considered and wanted to have a visual comparison of the locations in contention.

# Description of the data and how it will be used to solve the problem

The solution to the problem described here is based on a data science project which does the analysis of a geographical location based on its location data. The project would explore the venues of a location based on the area name, latitude and the longitude of the location using the Four Squared API. Based on the analysis of this data and using some statistical and clustering algorithms, the selection of the best location for starting a pool supplies and equipment store will be made. This will be explained in the subsequent sections of this report. The data required for this kind of analysis was obtained as shown below.

The exploration of the location has to cover all the 3 counties and all possible locations in the 3 counties of Henrico, Chesterfield and Richmond city. A data source to start such an analysis was found at the website of ‘United States Board on Geographic Names’’ . The link for this file is provided here : <https://geonames.usgs.gov/domestic/download_data.htm>

On this site, under the section ‘**State Files with Federal Codes’** the file for Virginia state was available to be downloaded as a pipe delimited text file. On further analysis it provided features on6528 locations or areas called as Feature Names in the state of VA. Below link has the file format and the explanation of all the columns in the pipe delimited file as provided by the United States Board on Geographic Names website: <https://geonames.usgs.gov/domestic/fips55_fileformat.htm>

The file format with field characteristics and descriptions is explained below:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Length | Description |
| Feature ID | Number | 10 | Permanent, unique feature record identifier and official feature name as defined in INCITS 446-2008, Identifying Attributes for Named Physical and Cultural Geographic Features (Except Roads and Highways) of the United States, Its Territories, Outlying Areas, and Freely Associated Areas, and the Waters of the Same to the Limit of the Twelve-Mile Statutory Zone. |
| Feature Name | Character | 120 |
| Feature Class | Character | 50 | Definitions |
| Census Code | Character | 5 | Formerly the FIPS55 Place Code and FIPS55 Class Code. See Census ANSI Codes, for INCITS 446-2008, Identifying Attributes for Named Physical and Cultural Geographic Features (Except Roads and Highways) of the United States, Its Territories, Outlying Areas, and Freely Associated Areas, and the Waters of the Same to the Limit of the Twelve-Mile Statutory Zone. |
| Census Class Code | Character | 2 | Class Code Definitions |
| GSA Code | Character | 4 | General Services Administration (GSA) location code. Maintained by GSA as a place reference for administrative and personnel purposes. Not otherwise used or referenced by GNIS. |
| OPM Code | Character | 9 | Office of Personnel Management (OPM) location code. Maintained by OPM as a place reference for administrative and personnel purposes. Not otherwise used or referenced by GNIS. |
| State Numeric | Character | 2 | The unique two number code and the unique two letter alphabetic code for a US State as specified in INCITS 38:200x, (Formerly FIPS 5-2) Codes for the Identification of the States, the District of Columbia, Puerto Rico, and the Insular Areas of the United States. |
| State Alpha | Character | 2 |
| County Sequence | Number | 3 | Sequence of counties containing the feature. #1 contains feature location. |
| County Numeric | Character | 3 | The unique three number code and name for a county or county equivalent as specified in INCITS 31:200x, (Formerly FIPS 6-4) Codes for the Identification of Counties and Equivalent Entities of the United States, its Possessions, and Insular Areas |
| County Name | Character | 100 |  |
| Primary Latitude | Number | 11 by 7 | The official feature location as defined in INCITS 446-2008, Identifying Attributes for Named Physical and Cultural Geographic Features (Except Roads and Highways) of the United States, Its Territories, Outlying Areas, and Freely Associated Areas, and the Waters of the Same to the Limit of the Twelve-Mile Statutory Zone. |
| Primary Longitude | Number | 12 by 7 |
| Date Created | Date |  | The date the feature was initially committed to the database. |
| Date Edited | Date |  | The date any attribute of an existing feature was last edited. |

For use in this data science project this txt file will be converted to a csv file and used in the analysis of the location data. From the file, the below 5 fields are selected for the location data analysis: "STATE\_ALPHA", "COUNTY\_NAME", "FEATURE\_NAME", "PRIMARY\_LATITUDE" ,"PRIMARY\_LONGITUDE”. These corresponds to the State Name, County Name, Area Name, Latitude of the Area and the Longitude of the Area respectively.

The file format is explained above, and the five fields selected for the data analysis from the State Files with Federal Codes file for Virginia are shaded in grey. From the file the areas within the 3 counties of Henrico, Chesterfield and Richmond city can be easily filtered using a sorting logic for the data frame in python pandas. Thus, we are equipped with the Area Name, and their Latitude and Longitude Information which can be used to analyze them.

The geographical information can be read into a python pandas data frame and used as the input for the exploration of these locations using the Foursquared API. Using Four Squared API’s **explore** feature, the different ‘category of venues in the JSON file’ otherwise the type of venues are extracted from the Foursquared API’s JSON file output. The locations with the geographic information of each venue and the category type of each venue are tabulated into a pandas data frame. This data frame will be used as input to the one hot encoding algorithm. The resultant data frame from the one hot encoding is used as an input to the k-means clustering algorithm.

A pandas data frame with the one hot encoding of the type of venues is sorted for each location for its mean value and used as the input data frame for k-means clustering. The dataset values would be the n observations which can be partitioned to k clusters in the k-mean clustering algorithm. The clustering algorithm will move each location into its own ‘location type’ cluster.

With the venues data frame, logic can be written to list the number of pools as venues for each location. Also, from the output of the one-hot encoding logic can be written to understand the relative importance of pools and rank them amongst the other venues in a location. These 2 derived data points along with the cluster information of each location is tabulated and is used to make a data driven decision on the final location where it makes the most business sense to open a Pool Equipment Supplies Store.

# Methodology section

The geographical information of the 3 counties of Richmond is read into a csv file and the required columns for the data science project is selected for the 3 counties of Henrico', Chesterfield and Richmond City.

## Some general considerations in the data science methodology

* The data is verified by visual and statistical methods after each data cleaning is performed on the data. Always the project looks at the raw data that is being analyzed to understand what the next steps should be to make the analysis as meaningful and accurate as possible. Validation is performed in after a few steps to ensure the code is preforming the intended function.
* The geographical co-ordinates of Richmond can be selected using the Nominatim API from the to convert an address into latitude and longitude values.
* Using Folium and matplotlib we map the geographical map of Richmond and added the required markers to it and thus we are able to visualize the locations of the different Area of Richmond on a map.
* By Applying one-hot encoding on the venue categories and then applying K-means algorithm to the one hot data frame a clustering of the locations based on the venue type is obtained
* Other data points like the number of pools in a given area and the ranking of the pools amongst the different venue types in an Area was obtained through Python coding using Pandas and Numpy arrays.

## The decision on the location for the new store was decided based on

1. Cluster name that area belonged
2. Count of Pools in the Area
3. Ranking of Pools amongst the venues in an Area

## High-level flow of the coding sequences is shown in the below steps:

* Identify the fields required for the data analysis from the dataset from United States Board on Geographic Names’
* Clean up the data with proper columns names, data formats, and verify that there is no missing data
* Select only the Areas or neighborhoods of the 3 counties in question around Richmond
* Visualize the areas on the geo map using Folium and Matplotlib libraries and ensure they are relevant and adjacent
* Using FourSquared API find the different type of venues for each area with say a limit of 100 and radius of 500
* From the JSON format select the required fields for further data processing, like the type of venues -category field in JSON for each of the venues
* Apply one hot encoding on the data frame with the type or the category of the venues for the many venues in the all the areas
* The output of one hot encoding is grouped for each area for its mean value and the resultant data frame can be used as input to k-means clustering algorithm
* K-means clustering will segregate Areas into different clusters based on the category or the venue
* By analyzing the type of venues in the cluster output from k-means, a name can be given to clusters
* Verify if the data we have would help make a confident and data driven decision to the problem in hand
* Visually analyze the clusters by mapping the clusters on to a Folium geo map using matplotlib and Folium
* Focus on the Areas which has Pools and also collect further data points like the count of Pools in an area and the ranking of Pools amongst the different venues in the Area
* Come up with a result table which clearly explains the different criteria and the Areas ranked based on the data leading us into a conclusion on where the Pool Equipment and Supplies shop should be.
* Provide the stakeholders with the final recommendation of the best location to start the store and also provide them with a ranking of locations that they can consider.

## Diagrammatic Representation of the Data Processing Flow:

Download the txt file and create CSV File with Geo information of Areas in all the counties in the State of Virginia

Clean the Data – Select relevant features, verify formats and ensure no missing data

Evaluation and Validation of Data for fitment

Find Co-ordinates of Richmond using Nominatim API from geopy.geocoders library

Plot Map of Areas of the 3 Counties in Richmond that are considered

Visual Evaluation and Validation of Data for fitment for the Data Science Analysis

Using FourSquared API

Find venues for each area and their category field from JSON

Limit=100, Radius =500

Clean the dataset and prepare for One Hot Encoding Logic

Apply One Hot Encoding

Clean Data Again to prepare for k-means clustering- Group for each Area the category type and then assign the mean value for category

Using the data frame apply

k-means clustering

Filter areas with Pools and further analyze

Visualize the cluster on a Geo map using Folium and matplotlib and apply markers

Find the Count of Pools for areas with Pools

Get the ranking of Pools amongst the different venues in the Area

Create a result table with the 3 criteria – Cluster Label, Pool Count, Ranking of Pools

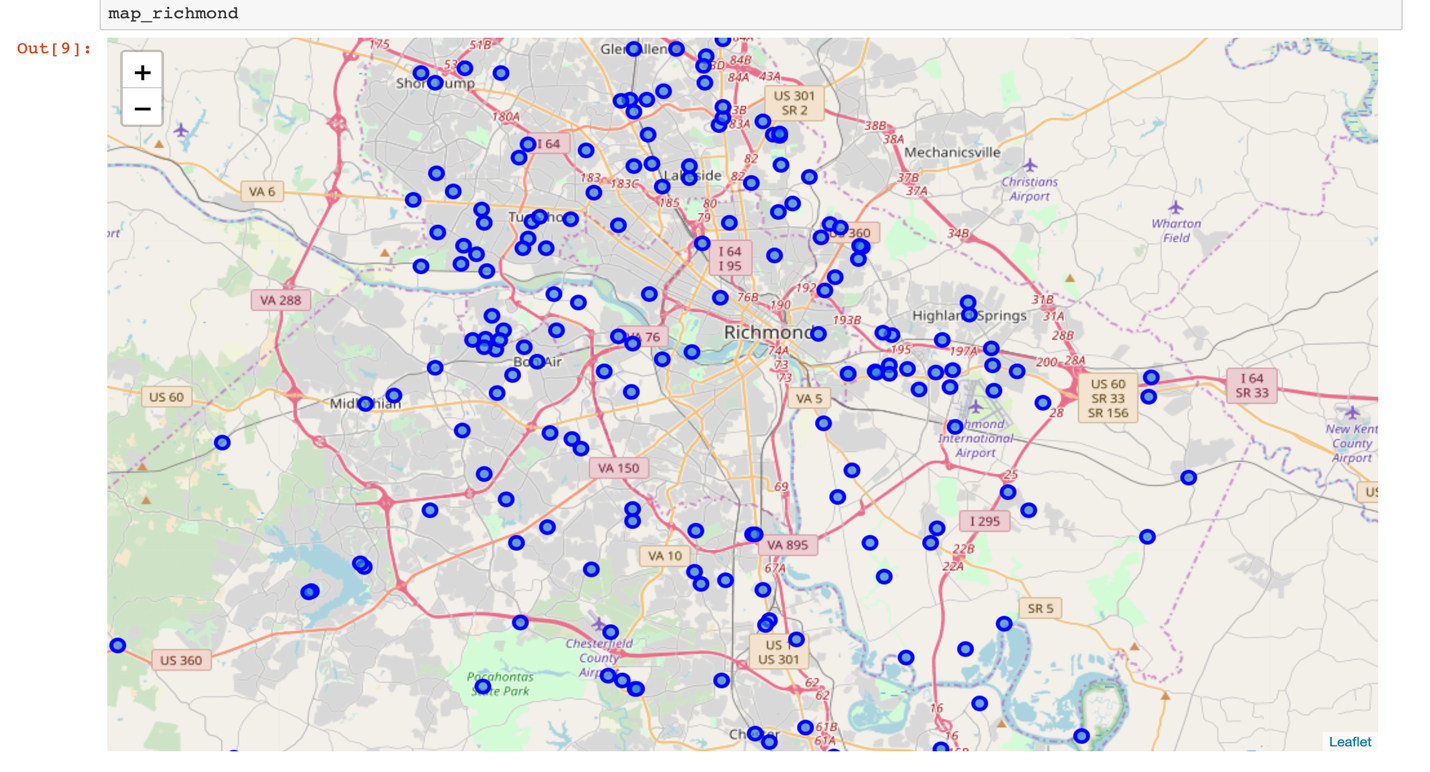
Final recommendation of the new Store Location

* Results section

There were 6528 locations or Area Names for the state of Virginia that was provided in the file from United States Board on Geographic Names ‘Out of these 6528 Areas ,190 Areas were from the 3 counties of Henrico, Chesterfield and Richmond City which was considered for the data science project.947 venues were identified by the Foursquared API falling into 194 unique categories when the input was for 190 Areas. 40 Areas did not have any venues identified by the FourSquared API. Thus, we considered 150 Areas with 947 venues in total falling into 194 unique venue categories.

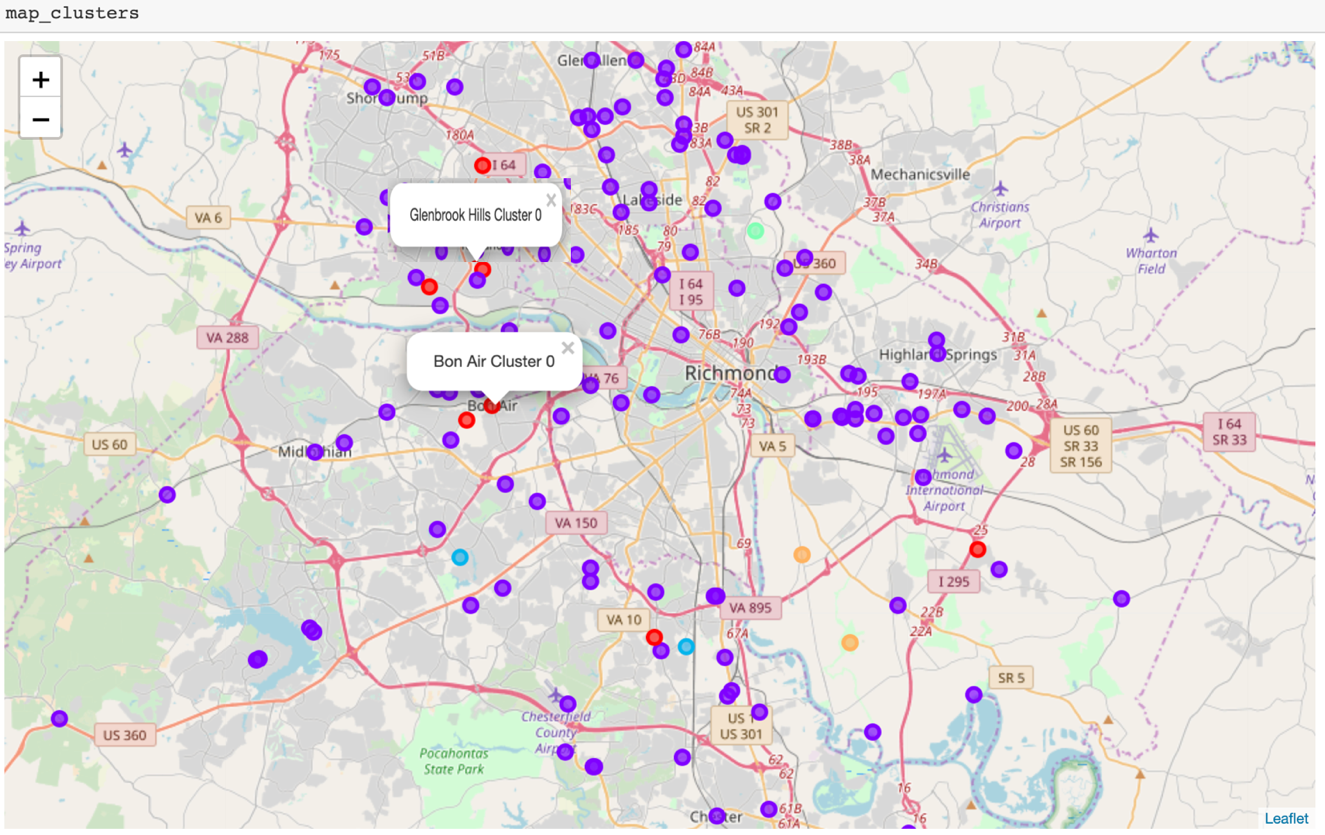
|  |  |
| --- | --- |
| Description | Count |
| Areas in the state of VA | 6528 |
| Areas in the 3 Counties of Henrico, Chesterfield, Richmond City | 190 |
| Total number of Venues selected for the 190 Areas | 947 |
| Areas with no venues identified by FourSquared API | 40 |
| Total Areas considered | 150 |
| Unique Venue Categories | 194 |
| Total number of Venues | 947 |

## 190 Areas of Richmond Plotted on a Folio Map

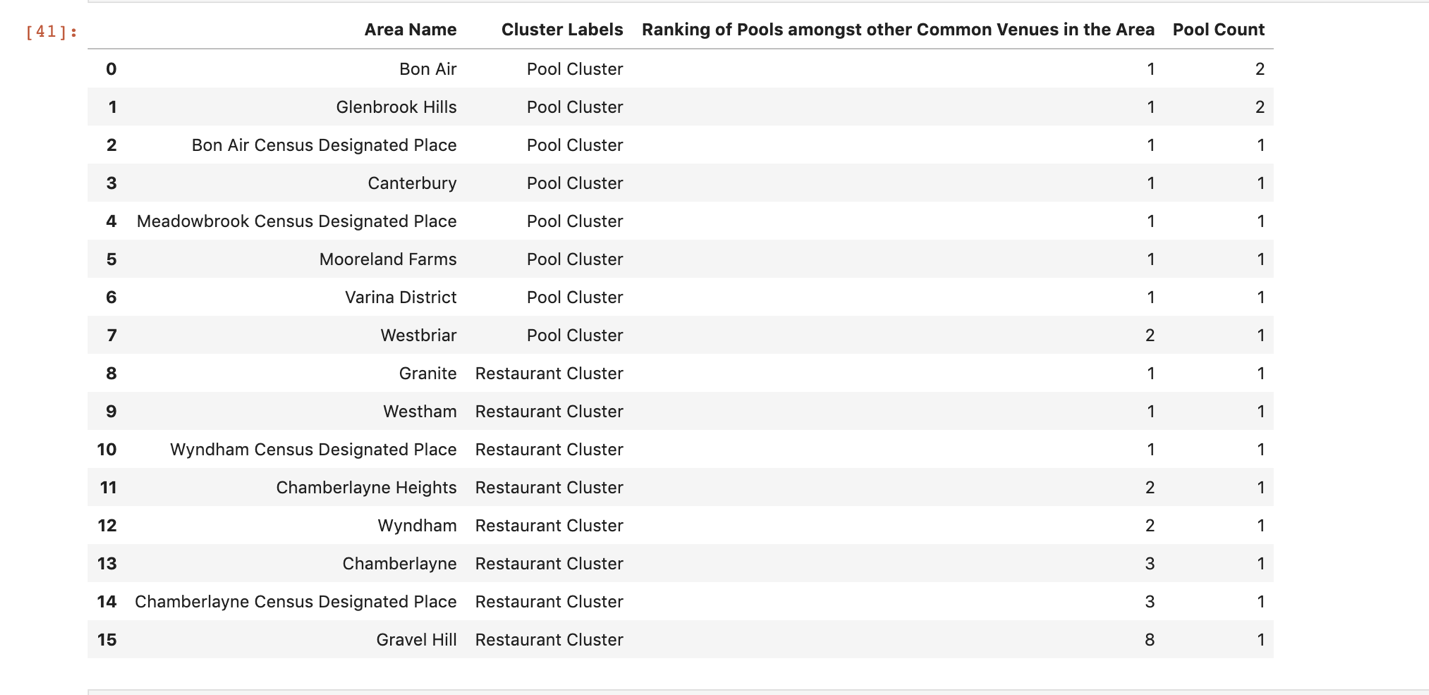


## 150 Areas of Richmond Clustered using k-means based on Category type of Venues in an Area and Plotted on a Folio Map – Pool Cluster represented by the Reds

**Recommendation of the shop is between Glenbrook Hills and of Bon Air**



## Final Result Table



## Recommendation of the Destination for Pools Equipment and Supplies Store

**A location south of Glenbrook Hills and north of Bon Air ideally between them and equidistant from both of them**

# Discussion Section

# It was observed that the number of venues reported by the FourSquared location API were limited. It would be advisable to ensure that the FourSquared API has the best venue information available for the 3 counties considered for Richmond.

The data available from FourSquared API was very consistent throughout and data validations always passed without issues.

It was observed that the number of venues for each location in Richmond counties were less in number. From this data point it can be concluded that the density of venues in many areas of the counties are less meaning the density of the population is also less. This leads us to recommend that the density of population of the areas be also considered in the final decision making.

The density of population and the per capita income were not considered here for the analysis because the number of Areas which had pools were very less [15 Areas]. So, this was a data point that can be manually considered before choosing the venue. Also, a relative grading of all the 15 Areas are provided which can be taken into consideration when deciding the store location.

# Since this is the first store that the company is planning to open a conservative approach of opening the store where Pools are currently present was taken along with the ranking of the ranking of the pools in an area belonging to a pool cluster.

# Conclusion Section

It can be concluded from the data that the level of confidence that I have in suggested the location for the Pool Equipment and Supplies Store is very high. The data points are very consistent and interestingly the 3 data criteria selected here pointed towards the same 2 Areas where the store should be opened which are BonAir and Glenbrook Hills. So, the final recommendation is to open the store in between and possibly equidistant from both the locations. It was also found that the other Pools and Pools clusters fall close by to these 2 locations and thus the decision can be final.