# 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, Country 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 they 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.