**Prospecting for New Business Locations – Gym/Fitness Facilities**

**Margaret Platt**

**February 10th, 2020**

1. **Introduction**

This paper is designed to illustrate some of the ways in which the tools and techniques of data science can inform business decision-making, enabling business owners to achieve higher levels of growth and profitability.

For demonstration purposes, we examine the case of a (fictional) business, ‘Fitness4All’, which operates gyms in the state of New Jersey. Fitness4All is a family-owned business which began in south Jersey, offering a good range of fitness equipment in bright, comfortable surroundings and opportunities for customers to work with staff on personalized training plans to meet their goals. It does not compete with upscale facilities marketed to young professionals residing in major urban centers and their proximate commuter towns, and its fees are correspondingly lower.

The company would like to expand its operations into the state of Pennsylvania. The company’s owner has a presumptive belief that the townships with the best prospects for profitability have a population of between 20,000 and 50,000 (which are often overlooked by major national chains). The owner has commissioned this project to (1) identify and prioritize prospective PA townships for the location of new facilities, based on their conformance to population criteria and the degree of competition for commercial gym facilities which already exists; (2) use a statistical technique to identify locations in PA which share demographic characteristics with the New Jersey locations which have already proved successful.

**2 Data Acquisition, Cleansing & Feature Selection**

**2.1 Sources**

This project uses readily available population data from Wikipedia to identify a baseline list of possible PA townships and their populations. This data is available at: <https://en.wikipedia.org/wiki/List_of_populated_places_in_Pennsylvania>

It also uses US census data to identify key demographics for each county in NJ and PA (this data seems not to be available at a township level). We use the ‘City and County’ Data Book’ (2007 edition) for demographic data, which is available at: <https://www.census.gov/library/publications/2010/compendia/databooks/ccdb07.html>

Because this dataset was only available in .pdf format, it was scraped to an Excel spreadsheet and stored locally.

We also use Foursquare location data to explore existing locations in short-listed townships i.e. to discover competitor facilities and complimentary amenities.

See ‘Discussion’ section below for observations on the quality of data inputs used for this analysis

**2.2 Data Cleansing & Feature Selection**

With the exception of some semi-manual operations to represent the .pdf data described above in an Excel format, none of the data sources required extensive pre-processing outside of common Python operations to amend/simplify column headings and drop un-needed columns or rows e.g., for the demographic data:

* Dropped summary demographic values for the states of NJ and PA
* Dropped ‘Persons’ as this measure of population size is highly redundant with ‘Households’
* Addressed columns with NA values in numeric columns e.g. relating to the percentage population of Native American origin. Because most values for these features were either NA, or very small and unlikely to be material to the analysis, we eliminated these variables

**3 Methodology**

The key steps in the analysis are described below. Further details are documented with the results achieved in the ‘Results’ section:

1. Identified & reviewed the geographic locations of PA townships with a population between 20,000 – 50,000
2. Performed a Clustering analysis to compare NJ and PA counties based on demographic information, in order to identify PA counties which most resembled the NJ counties where Fitness4All locations have already been successful
3. Performed Foursquare venue searches and analysis to identify competitor gym facilities and complimentary amenities in short-listed PA townships (because locations with favorable demographics are not good business prospects if already subject to intense competition)
4. Used the information from II and III above, to further refine the shortlist, and demonstrate how Foursquare data can be used to investigate specific potential sites within townships

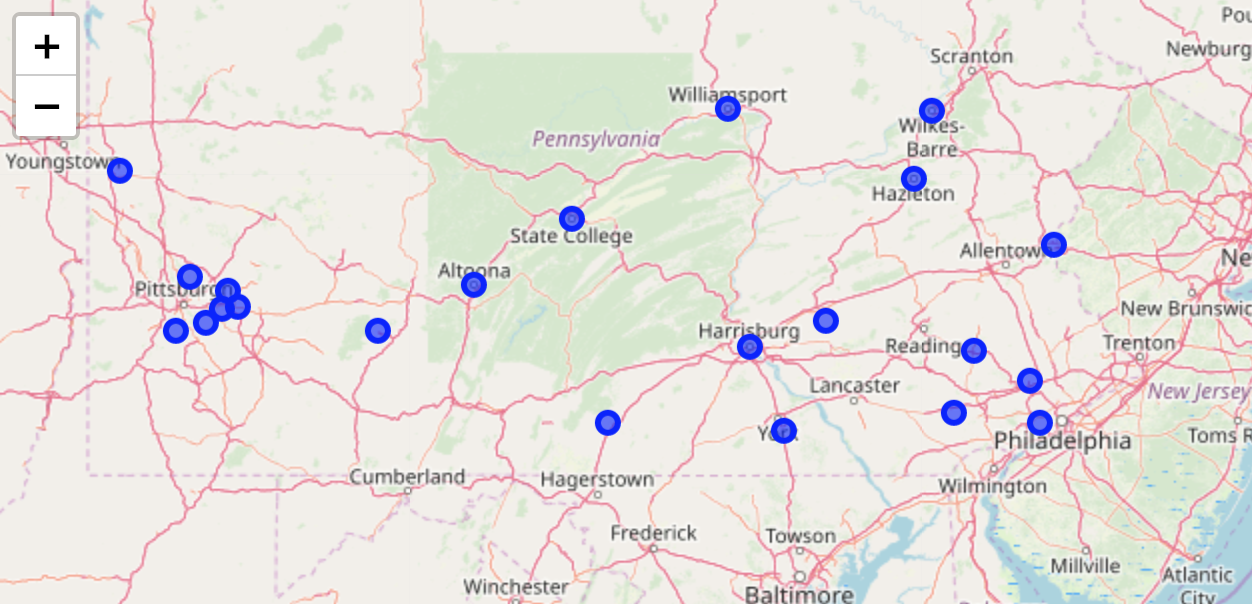
From a technical perspective, the key statistical technique used was K-means clustering as implemented by Scikit-learn, to characterize and compare like counties across NJ and PA. Metrics were used to find the optimal value of k.

**4 Analysis Detail and Results**

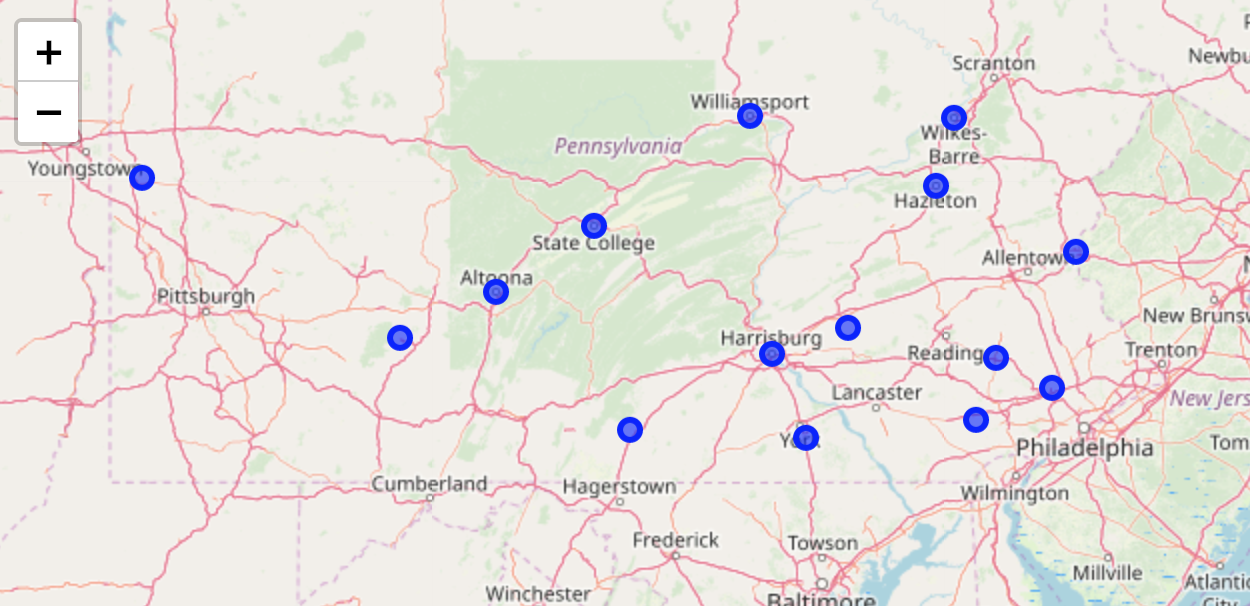
1. *Identified & reviewed the geographic locations of PA Townships with a population between 20,000 – 50,000*

From the Wikipedia PA township information, a data frame of townships with the specified population criteria was created, with spatial coordinates appended, and plotted on a Folium map of Pennsylvania. It was immediately apparent that some ‘townships’ were effectively suburbs close to the center of major urban areas (Philadelphia or Pittsburgh). As this is not our target market, these townships were filtered out from the focus-list data frame, and the map was re-generated. As a result of these operations, a short-list of 15 townships was identified.

Initial Plot:

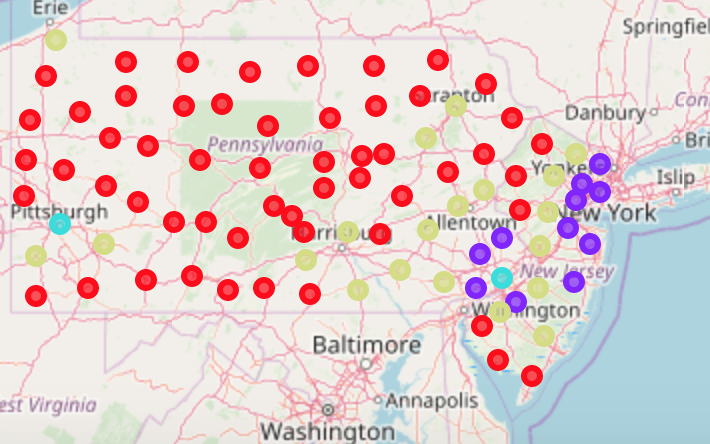


Revised Plot:



1. ***Performed a Clustering analysis to compare NJ and PA counties based on demographic information, in order to identify PA counties which most resembled those NJ counties where Fitness4All locations have already been successful***

US census data for NJ and PA counties was imported and cleansed (see above). This process resulted in data for 88 counties with 23 features. The variables were normalized using StandardScaler and a Scikit k-means model was fitted to the data, initially using a k value of 4. Assigning the generated cluster IDs to the counties and their spatial coordinates, we created the below cluster map



This result was intuitive, and provided clusters that could be characterized as follows, based on the mean value of the features for each cluster.

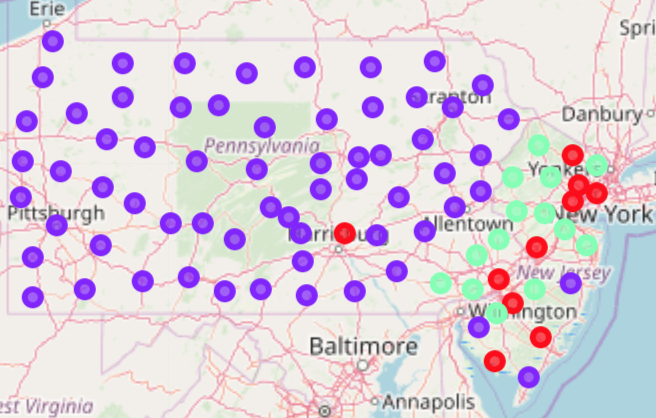
Cluster 0 (Red) – “Older, small town communities”: highest proportion of over 55s, least ethnically diverse, lowest level of college education and high income households; geographically located in western Pennsylvania and south Jersey

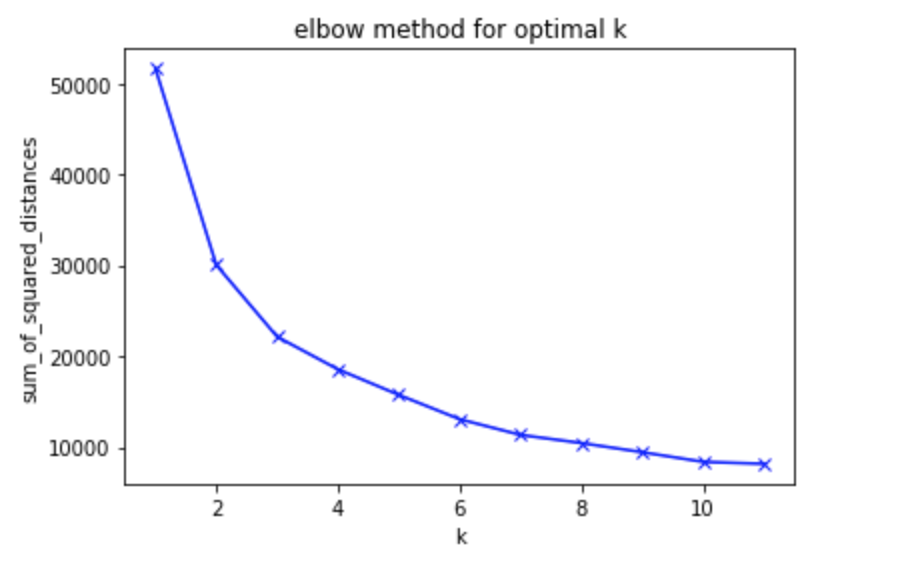
Cluster 1 (Purple) - “Commuter belt communities”: relatively young population, the most ethnically diverse outside of major cities; highest percent of college educated and high income; geographically located in the commuter belt close to New York and Philadelphia

Cluster 2 (Blue) – “Major urban centers”: greater Philadelphia and Pittsburgh

Cluster 3 (Gold) – “Middle America” – average scores on most attributes, generally located too far west for commuting to New York and Philadelphia (a couple of these locations are close to Pittsburgh)

However, this first pass was not fully satisfactory, because it appeared that the ‘population size’ attribute (‘Households’) was dominating the results. While township population size was an important consideration in location selection, it’s not a pure demographic measure. So the size attribute was dropped and the analysis was repeated. In this version of the analysis we found that k=3 provided an insightful differentiation of cluster characteristics (a plot of sum-of-squared-differences also suggests diminishing information value for higher than 3).





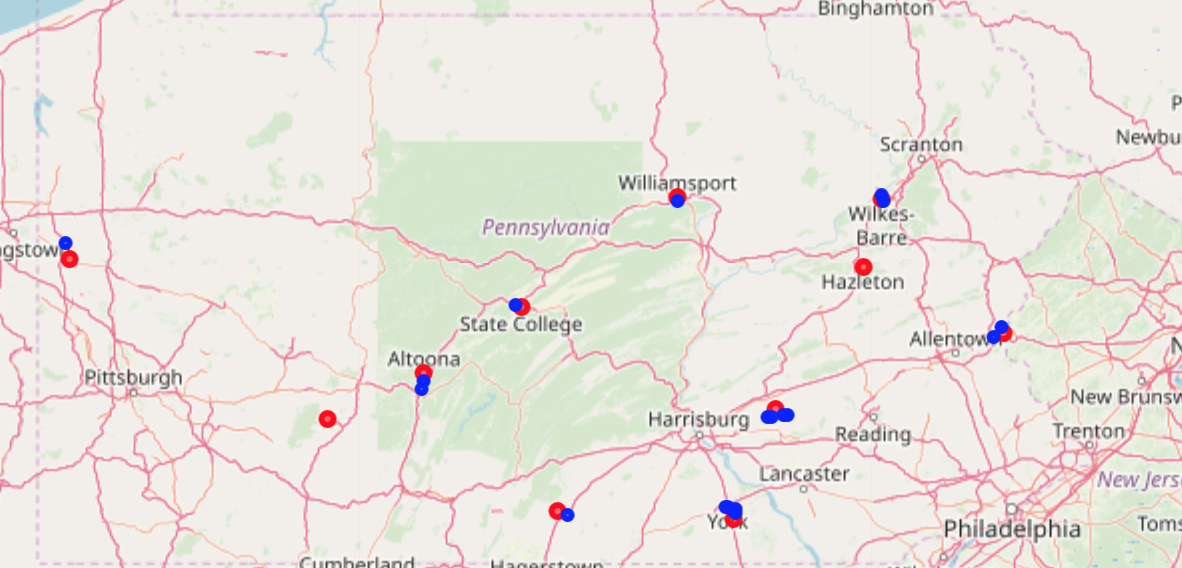
Cluster 0 (Red) – Relatively young population, the most ethnically diverse group, median score among these clusters for high-income households and college education

Cluster 1 (Purple) – Highest percentage of over-55s, least ethnically diverse, lower incidence of high-income households, less likely to have a college education

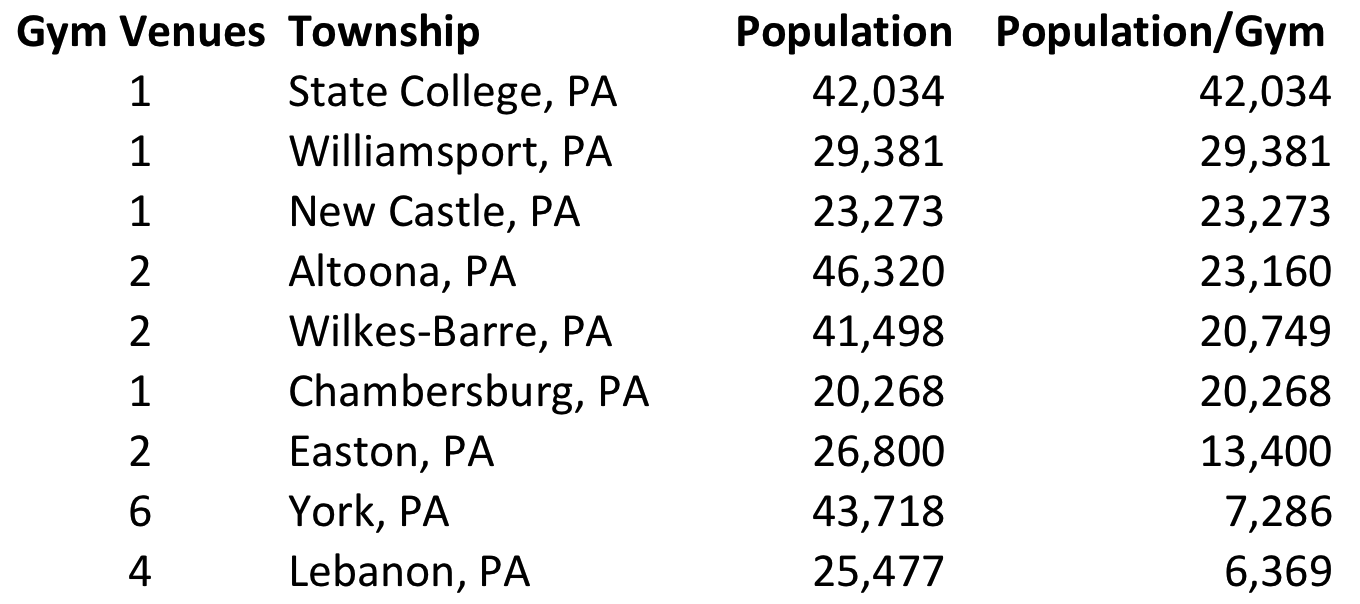
Cluster 2 (Green) – Significantly higher incidence of college education and high-income households, average profile for age distribution and ethnicity

Cluster 1 reflects the demographic profile of existing south New Jersey locations. We therefore drop Harrisburg, Norristown, Chester & Pottstown, leaving a short-list of 11 towns in counties with optimal demographics.

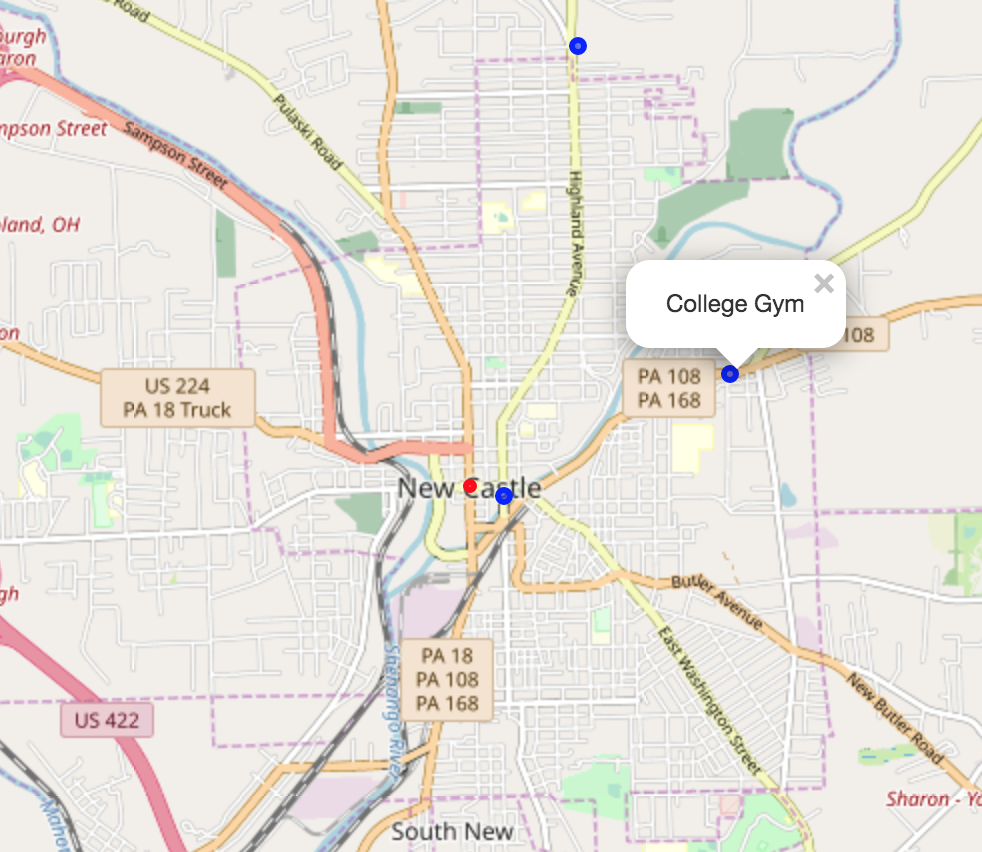
We then retrieve FourSquare venue information, filter it down to gym facilities, & plot these facilities against our 11 towns:



We note right away that two of our short-listed towns (Red dots) have no gyms (Blue dots) nearby: Hazleton and Johnstown – this is worthy of further investigation! For the remaining 9 Townships, we compute the size of the population divided by existing gym venues to prioritize locations where competition is relatively less intense:



Finally, Foursquare data can be used to look for potential sites within our top-prospect townships, by looking at proximity to any existing gyms and the locations of complimentary facilities, such as coffee-shops. For example, the below plot of venues in New Castle, PA shows that the existing gym is some ways out of the center of town (and is a College Gym, so may not be attractive/available to commercial customers). There are a couple of coffee shops for potential co-location: one right in the center of town, and another at the northern end.



**5 Discussion – Observations**

Several challenges were encountered in producing this analysis:

* More time and experience may be needed to locate the most accurate and relevant data at a reasonable cost. The ‘City and County Data Book’ used has not been issued by the US Census Bureau since 2012, and an equivalent publication produced privately costs hundreds of dollars (and is in book or .pdf form). It seems there is still a lot of information interesting to historians and social scientists available only in in paper and microfiche formats.
* In the time available, I could only find demographic data at a county level, so had to make the assumption that county demographics were good representations of the demographics of the townships within them. This seemed not unreasonable – in that the townships analyzed were generally the dominant communities within their counties – but gathering information at a township level would have been a more consistent approach.
* I had some difficulty running the function to extract gym information from Foursquare for each of the townships using the Foursquare ‘search?’ function. The code ran to completion successfully only when I severely constrained the LIMIT parameter to the point where the data retrieved was not sufficiently comprehensive. I could not determine if I had introduced a problem in modifying Lab code, or if the problem was related to Foursquare api developer constraints. I was able to get the needed data by reverting to the ‘explore?’ format of API calls, but this seemed very inefficient from a data transfer perspective because it involved requests for large volumes of venue information, most of which was filtered out by venue category to produce the data frame of interest.

**6 Conclusions & Opportunities for Further Analysis**

For this analysis, the goal was simply to create a focus list of locations for new business opportunities, and to limit the risk of business failure by identifying which locations have similar population characteristics to existing locations. However, the clustering of communities by demographic characteristics offers some interesting information for future consideration. Fitness4All is a currently small business, without access to sophisticated market research or the scale to contemplate targeting branding strategies to different market segments. As the business grows, the management team may wish to further explore the requirements and preferences of market segments different from their current target market, to expand the range of potential locations where they can operate successfully.