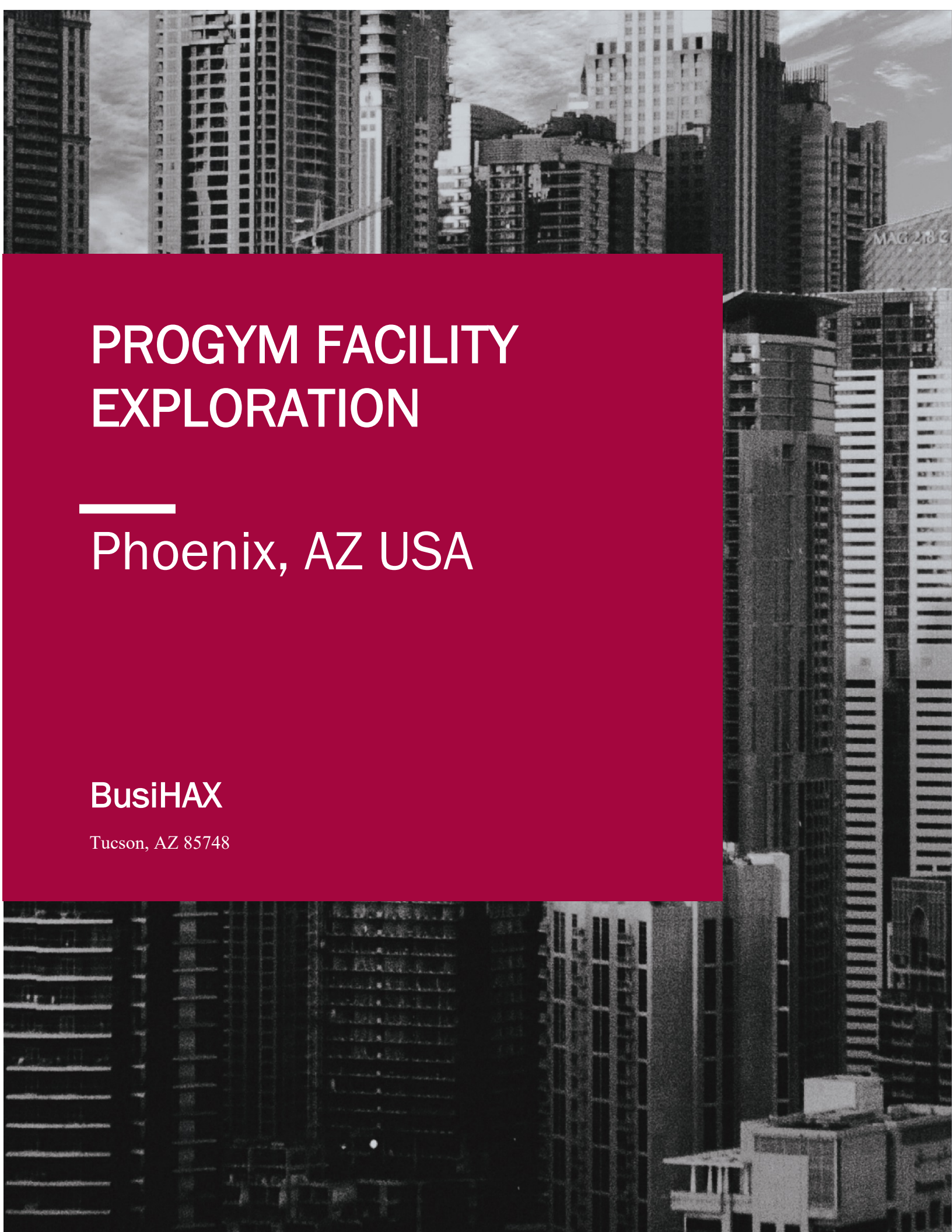


PROGYM FACILITY EXPLORATION

Phoenix, AZ USA

BusiHAX

Tucson, AZ 85748

Introduction

ProGym New Facility Exploration

ProGym is a fictional physical fitness and training company based in Los Angeles, CA. The ProGym Business Model caters primarily to physicians and members of the professional medical community. The ProGym management team has been growing the company steadily since 2014 and believes that Phoenix, AZ is the best location for the next facility. DataHAX is a fictional Business Intelligence firm based in Tucson, AZ. One of the core offerings by DataHAX is understanding business environments. DataHAX mission is simple - we turn data into intelligence, and intelligence into success. For almost 90 days we have been using our intelligence assets to help businesses win.

BUSINESS CHALLENGE

DataHAX has been hired by ProGym to help determine the best location for a new facility in the metro Phoenix area. ProGym has a deep understanding of their highest value facilities and what makes them profitable and successful. From their most successful facilities, ProGym as developed a Facility Profile. ProGym understands their clientele very well and have developed a very specific Client Profile based on their business model. The challenge for DataHAX will be to combine data about the Phoenix Metro Area with the ProGym Client and Facility Profiles to find the best location candidates.

STAKEHOLDERS

Interested parties for the ProGym-Phoenix project include the ProGym facilities team and the ProGym executive management team.

ProGym Facility Profile

ProGym Facility Profile
ProGym facilities located near hospitals, doctor's offices and other professional buildings have proven to be the most successful. Additionally, ProGym has noted that locations in close proximity to fine dining establishments tend to sell more personal training services. This correlation is strong enough to indicate that some weight should be given to proximity to fine dining.

- From the FourSquare.com Dataset API, will extract the following for each ZIP code (to allow an analysis of the ProGym Facility Profile):
  - Medical Facility Density
  - Fine Dining Restaurant Density

ProGym Client Profile

ProGym clientele are primarily medical professionals. This leads to the following demographics: age, education level, income level, and others.
Each dataset for this exploration consists of records indexed by a 5-digit ZIP code.

- From the US Census Dataset, will extract the following for each ZIP code:
  - Total age 35-55
  - Age 35-55 with bachelor’s degree
  - Age 35-55 with bachelor’s degree or higher
  - Age 35-55 Professional Degree
  - Annual Income over $90,000 US
- From the Internal Revenue Service Dataset, income level for each ZIP code will be added to the dataset
- From Zillow Dataset, the current median home value will be added to the dataset
- Each record in the new dataset will look like the following (to allow an analysis of the ProGym client profile):
  - ZIP Code (Index)
  - Total age 35-55
  - Age 35-55 with bachelor’s degree - count
  - Age 35-55 with bachelor’s degree or higher - count
  - Age 35-55 Professional Degree - count
  - Median Income

The Data

Each ProGym profile (Facility and Client) is composed of **profile indicators**. Profile indicators will be assigned to each of 115 ZIP codes in the Metro Phoenix area.

Data for this exploration will be drawn from multiple public sources as noted below. As will be seen later, records in the target dataset will map key profile

- FourSquare.com - Public venue information service: https://developer.foursquare.com/docs/places-api/
- US Census Data - https://data.census.gov
  - Age/Sex by ZIP Code - table **S0101**
  - Education by ZIP Code - table **S1501**
  - Income by ZIP Code - table **B19001**

- ZIP Code Latitude & Longitude: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/
- ZIP Code Geo Boundries https://catalog.data.gov/dataset/tiger-line-shapefile-2019-2010-nation-u-s-2010-census-5-digit-zip-code-tabulation-area-zcta5-na

Data Use

The datasets will be combined using the ZIP code as the index. The data will then be used to create Choropleth maps to better understand the population distribution based on each criteria. Next the data will be clustered using both K-Means and Support Vector methods to understand which ZIP Code groups best match all of the required criteria.

GEOJSON data was gleaned from this article:
https://medium.com/dataexplorations/generating-geojson-file-for-toronto-fsas-9b478a059f04

Data Wrangling

The data for this evaluation come from several sources. For analysis, each dataset needs to be cleaned up an normalized. The following code produces Pandas DataFrames with a ZIP code index for each dataset.

ZIP CODE ANOMALIES AND OUTLIERS

During the data analysis a couple of interesting issues were uncovered regarding ZIP codes **85309** and **85253**. 85309 is Luke Air Force Base. The US Census numbers are no available for this ZIP Code. 85253 is Paradise Valley. The median home values and income for this ZIP code are about 5x the median value for the rest of Phoenix.

In many cases, the latitude and longitude did not match and had to be verified corrected via a 3rd party.

DUPLICATE VENUE ISSUES

The data pull from FourSquare has a significant number of duplicate entries for medical venues. For example, Phoenix Children's Hospital (PCH) has one entry for each department. The data from FourSquare contained almost 20 entries for PCH at the same location. and an individual doctor office at a medical center. The nature of the duplicate names is such that it was more efficient to manually select a list of facilities to keep than to try to filter them programmatically. Even after the list was trimmed, duplicates still existed - so I had to remove by the entry ID.

Primary Data Sets

VENUE_DF – THE VENUE DATASET

venue_df is generated using the FourSquare API, searching for medical key words: 'Hospital', "Medical Center", "Doctor's Office". Unique results from each data ZIP code were recorded. For the 'Restaurant' key word, an price tier was added to search only for the top price-tier restaurants. Again, this search was done for each ZIP Code, with uniqueness being defined by the FourSquare ID. Each venue_df record has the following fields:

ZIP: Venue ZIP Code
id: FourSquare ID
Name: Venue Name
Latitude: Venue Latitude
Longitude: Venue Longitude
Category: FourSquare Venue Category, may differ from search criteria
Search: key words used for this search
Price Tier: For Restaurants – should always be 4, the highest
Rating: For Restaurants, FourSquare user average rating
Likes: For Restaurants, FourSquare user Likes

PROFILE_DF – CLIENT DATA + MEDICAL FACILITIES

profile_df list of ProGym Profile Indicators for each ZIP code in the Phoenix Area. Each profile_df record contains the following records

ZIP: Target Zip Code
Age Pop: Number of people age 35-55
Edu Pop: Number of people with a bachelor's degree or higer
Income Pop: Mean Income
Med Count: Count of medical facilities in this

Maricopa County ZIP Codes (ZIP Code Set)

85003	85004	85006	85007	85008	85009	85012	85013	85014	85015	85016
85017	85018	85019	85020	85021	85022	85023	85024	85027	85028	85029
85031	85032	85033	85034	85035	85037	85040	85041	85042	85043	85044
85045	85048	85050	85051	85053	85054	85083	85085	85086	85140	85201
85202	85203	85204	85205	85206	85207	85208	85209	85210	85212	85213
85215	85224	85225	85226	85233	85234	85248	85249	85250	85251	85253
85254	85255	85256	85257	85258	85259	85260	85266	85268	85281	85282
85283	85284	85286	85295	85296	85297	85298	85301	85302	85303	85304
85305	85306	85307	85308	85310	85323	85331	85335	85338	85339	85340
85345	85351	85353	85355	85363	85373	85374	85375	85379	85381	85382
85383	85387	85388	85392	85395						

Methodology

Data exploration and review followed the same steps for each ProGym Profile Indicator.

- Create a dataset with attributes ZIP, Profile Indicator
- Generate a choropleth map to display the densities
- Note the top 5 ZIP Code densities

med_df:
ZIP: ZIP Code
Med Count: Count of Medical Facilities in this ZIP

age_df:
ZIP: ZIP Code
age pop: Number of People in the target age range, in this ZIP

edu_df:
ZIP: ZIP Code
edu pop: Number of People with bachelor's degree or higher

income_df:
ZIP: ZIP Code
Income Pop: Average Income for this ZIP code

profile_df – combines all data frames using ZIP as the join key. This data frame as then used to run a K-Means Clustering algorithm, to find the best-fit ZIP Codes.

cluster_df – groups the data above into K-Means cluters

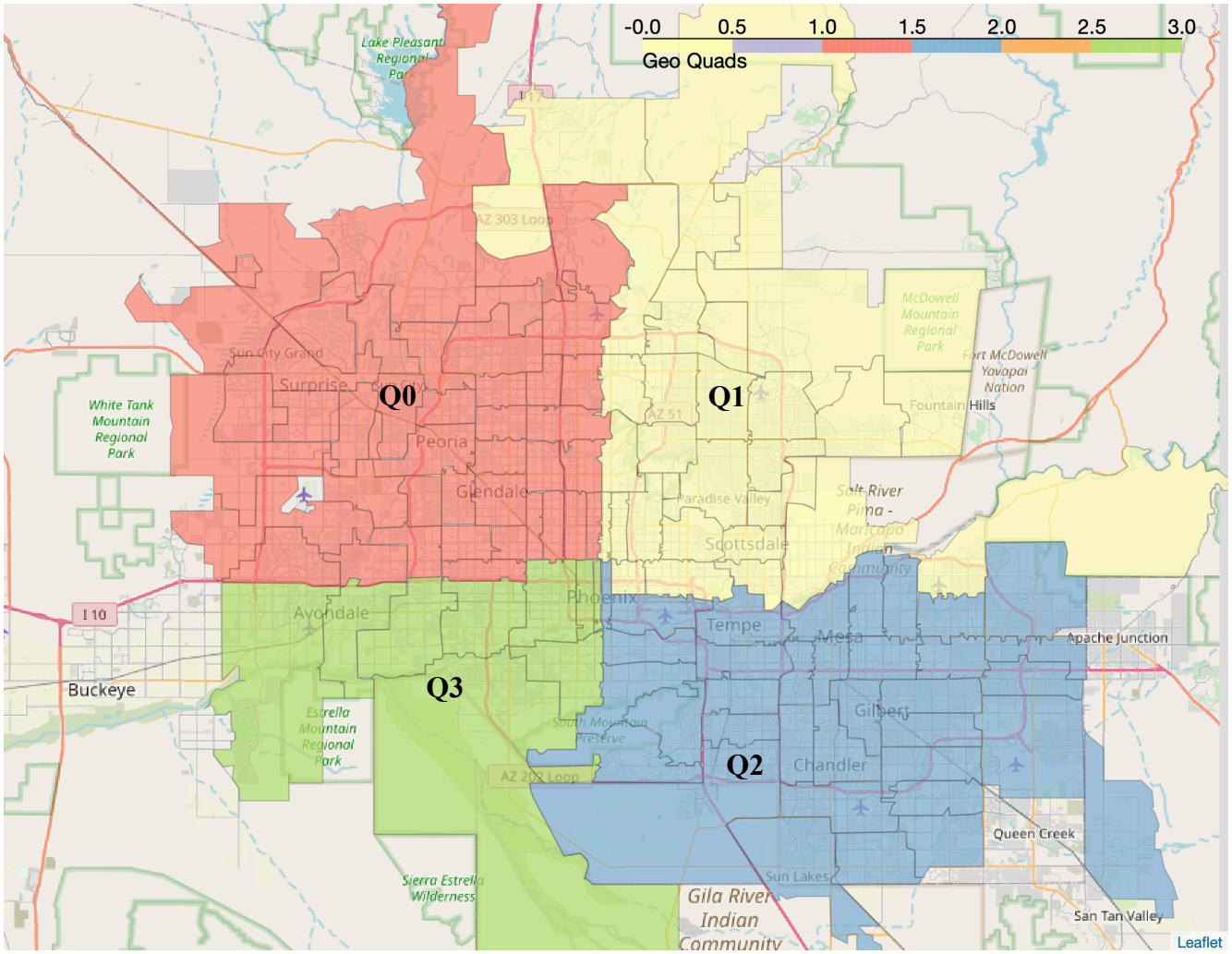
K-Means Clustering
K-Means Clustering is an unsupervised machine learning algorithm. K-Means reads a set of vectors and groups the vectors by minimizing the Euclidian distance between the members of each cluster and maximizing the distance between centroid of each cluster.

To summarize the top ZIP Codes, a K-Means algorithm was applied to the 'profile_df' data frame.
The K-Means
Using the venue_df , medical facilities were mapped to ZIP Codes a Choropleth map was generated to view the facility density for the ZIP Code Set. A list of ZIP Codes was created with the highest density of medical facilities.
A similar approach was taken for top tier restaurants for each zip code.

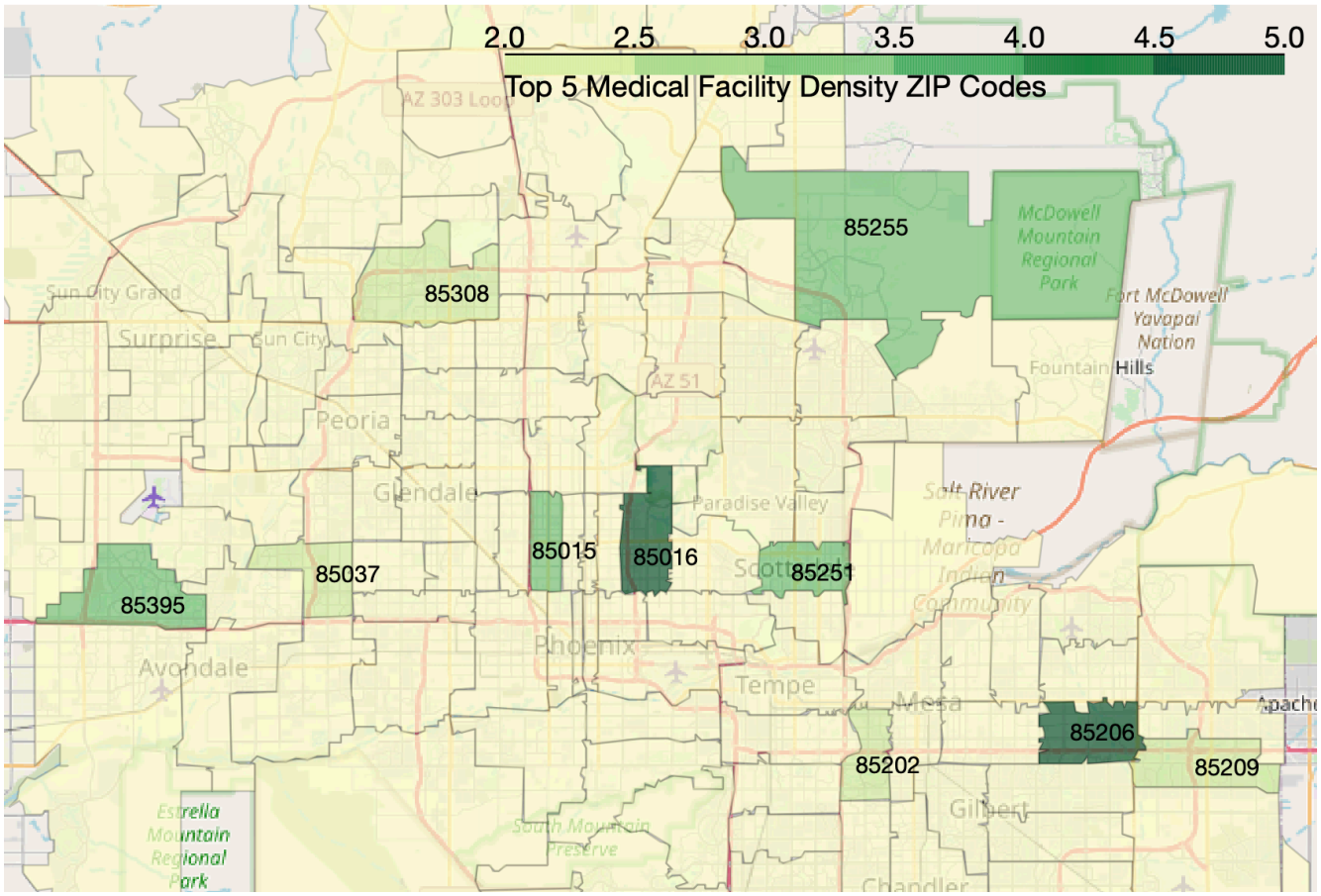
Results: Exploratory Investigation

To get an initial understanding of the demographic distribution in the Phoenix area, the city was divided int quadrants, Northeast, Southeast, Northwest, and Southwest. Median values from breakdown shown below – the boarders on this investigation are arbitrary. Using just the data in this table, “Quad 1” the highest income, target Education population, target age, and the highest Medical Facility count per ZIP code.

Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	color
8,761.93	4,841.81	74,383.64	0.43	33.59	-112.23	0
8,335.29	9,696.71	110,828.75	0.86	33.59	-111.95	1
10,426.71	8,753.40	86,528.74	0.60	33.36	-111.83	2
9,484.30	3,756.80	76,619.40	0.20	33.41	-112.19	3



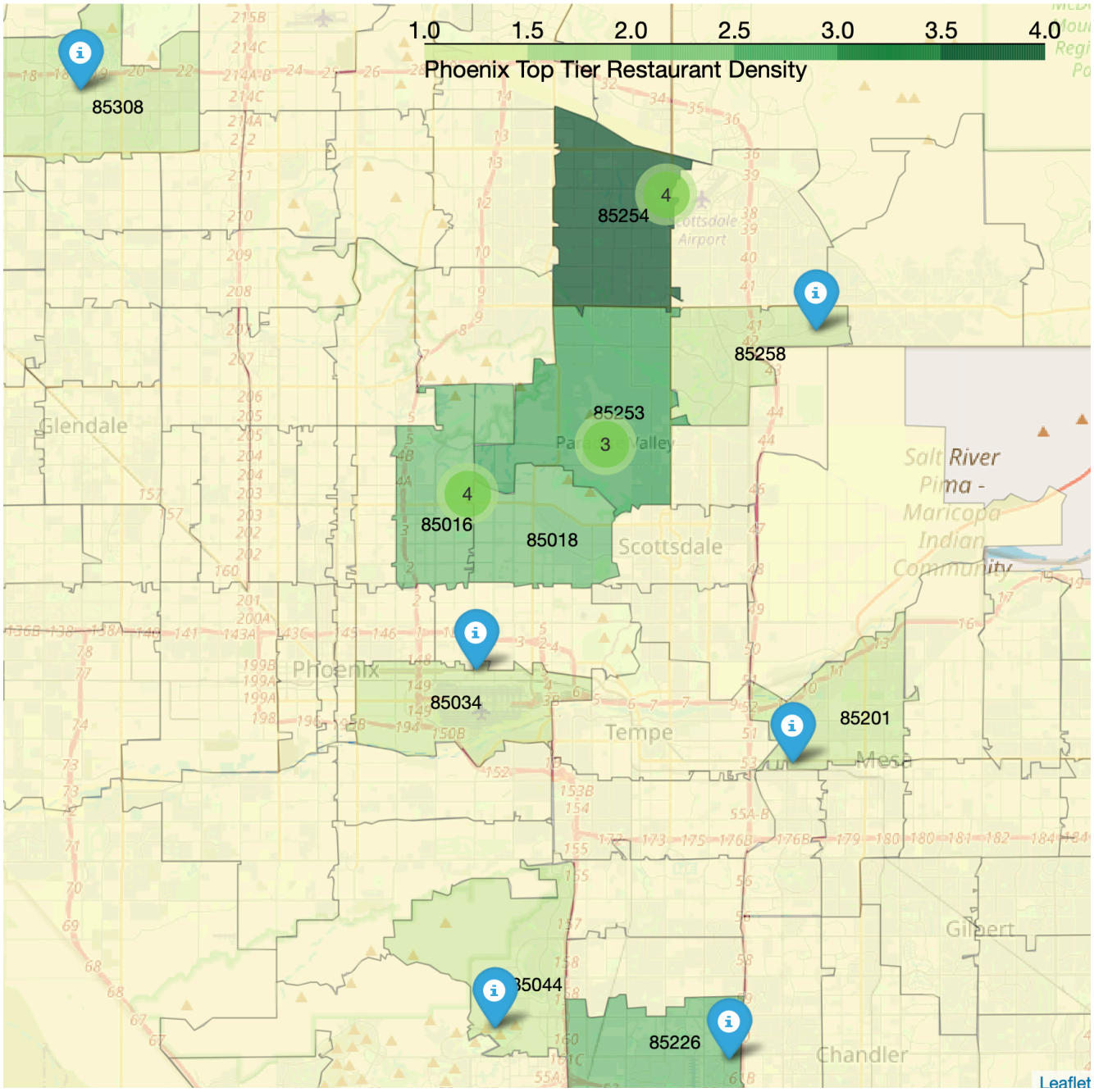
Facility Profile – Medical Facility Density



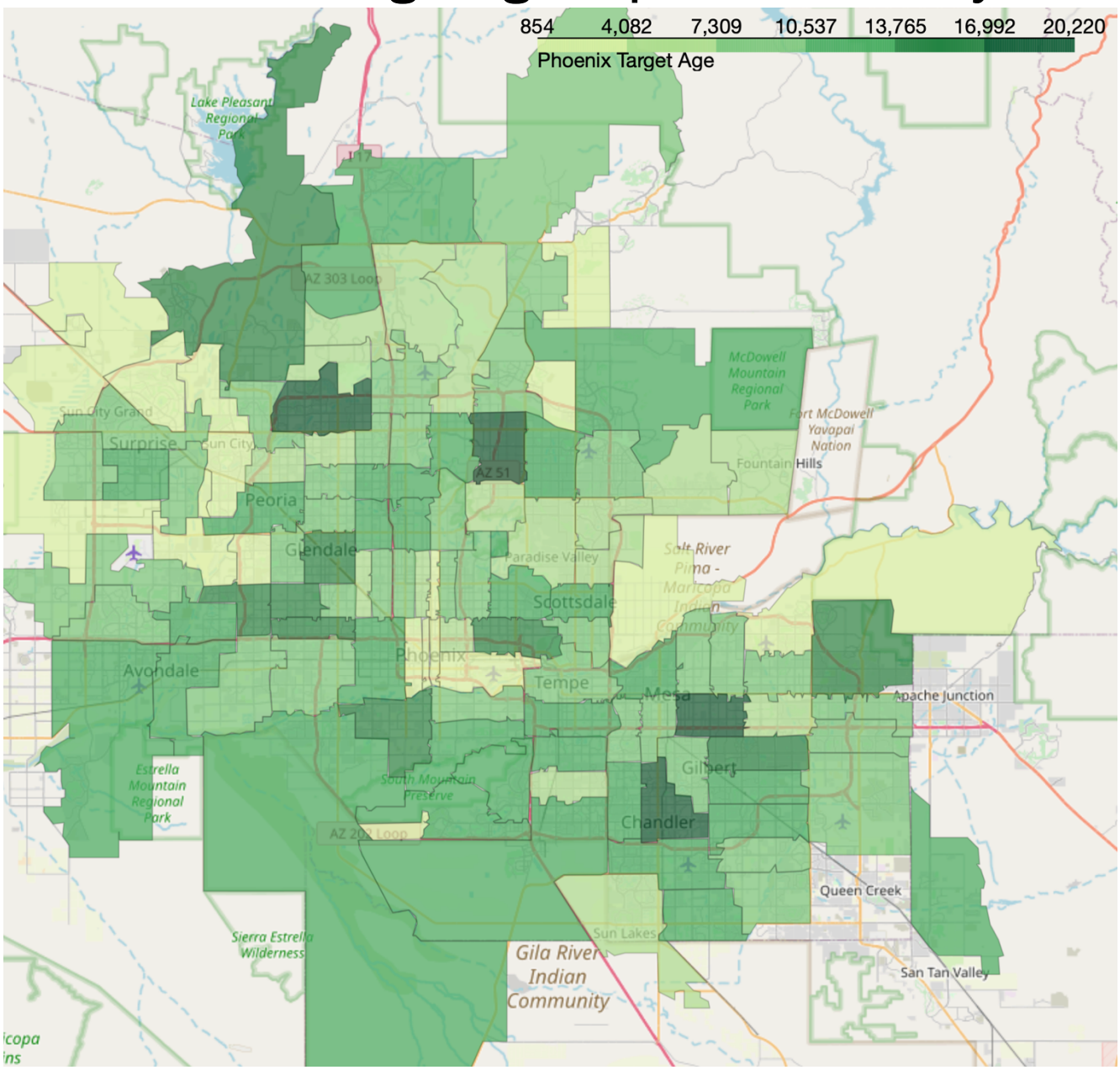
ZIP 85206 is the home of Banner Medical Center, the largest private employer in Arizona. 85016 is a Phoenix ZIP Code is the home of Phoenix Children's Hospital. 85255 is a North Scottsdale ZIP Code, home to the Mayo Clinic, Phoenix and Honor Health. 85251 is a Scottsdale ZIP Code and is home to Honor Health Scottsdale and the OneMedial Group. Note that 3 of these ZIP Codes – 85016, 86251 and 85255 surround Paradise Valley, home of the highest median income in the Phoenix Metro Area. 85395 has sufficient medical facilities, but the median income and education levels are not as high as the east side.

Facility Profile: Top Tier Restaurants

This map indicates the density of top-tier restaurants. A blue arrow is a single restaurant, a green circle indicates the number of restaurants in close proximity. A pattern is beginning to emerge around east Phoenix, Scottsdale, Paradise Valley and North Scottsdale.

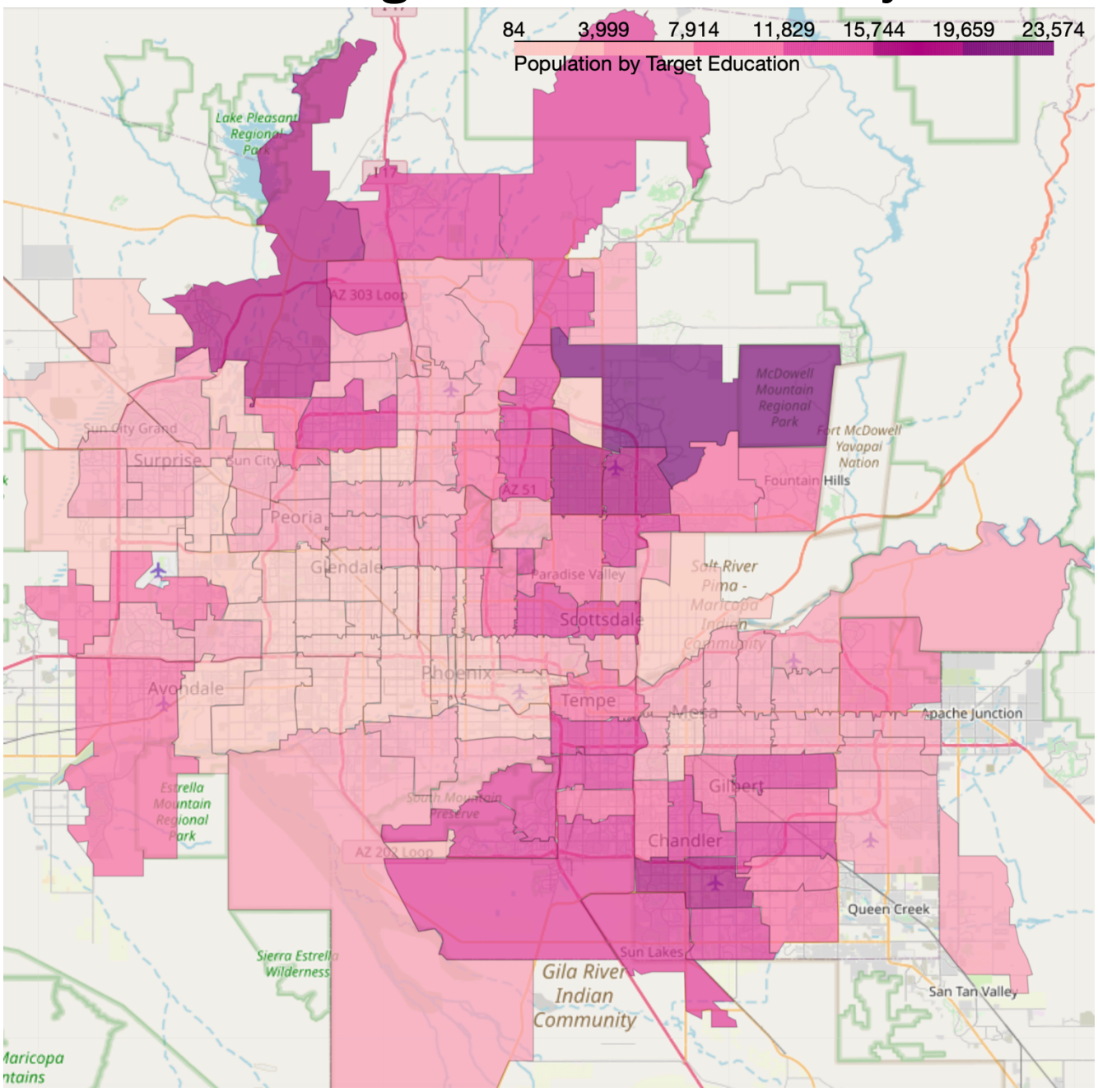


Client Profile: Target Age Population Density



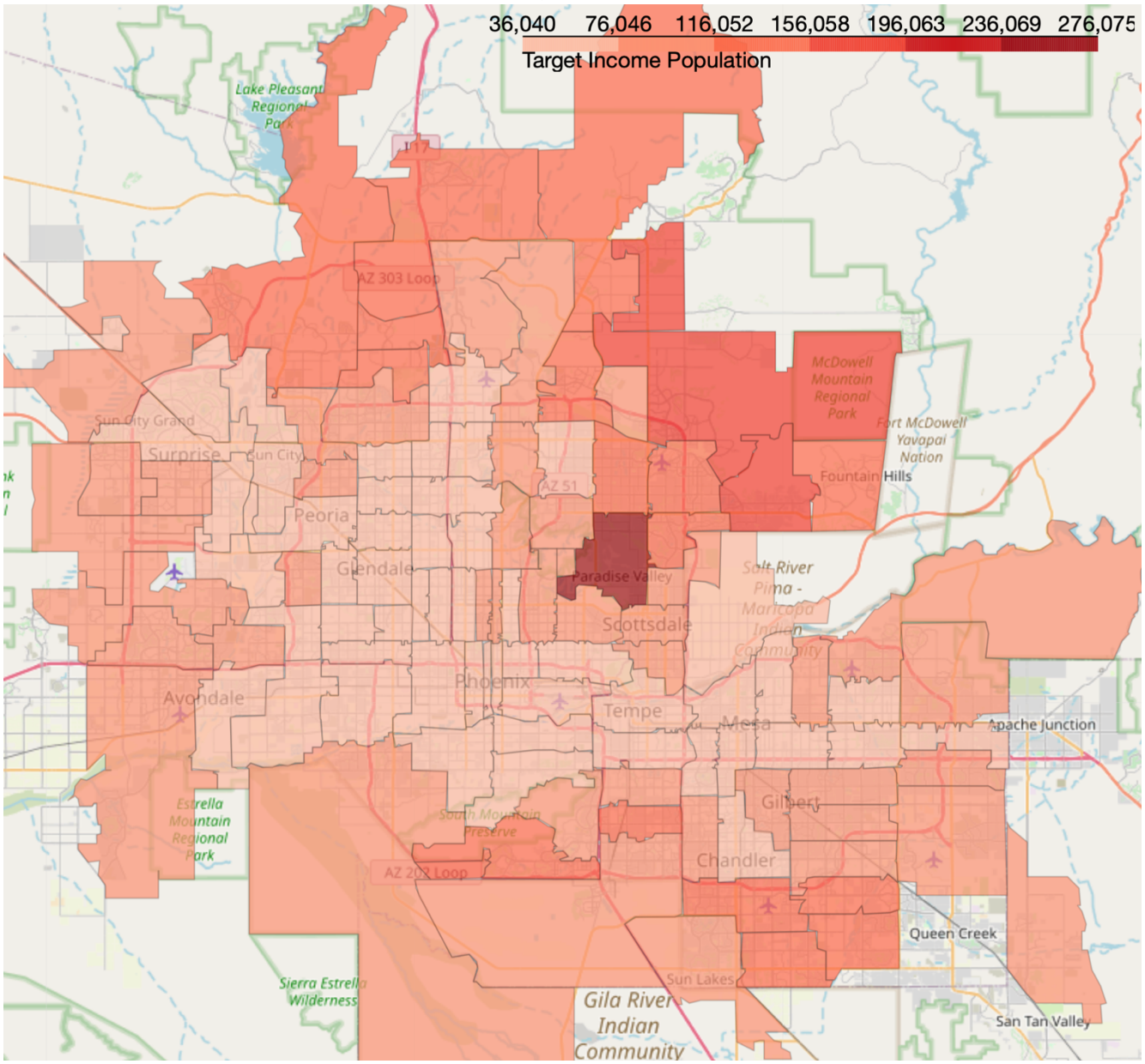
This chart is not exceedingly revealing but, it does as all of our data sets do, have a high data point in that central east to north east corridor.

Client Profile: Target Educational Density



Once again, our target population by education tends to be in the area between east Pheonix and Fountain Hills – focusing our search on that central east to north east corridor.

Client Profile: Income by ZIP Code



The pattern continues – with the higher income ZIP Codes being in the same general pattern as the target education range. Paradise Valley is an outlier in the income dataset with almost 4x the overall median income for the Metro Phoenix area.

K-Means Clustering

K-Means is a method of grouping data by minimizing the “distance” between data points. Details follow this link: https://scikit-learn.org/stable/modules/clustering.html#k-means

In this case, K-Means allows us to visualize clusters with similar attributes, both demographic and geographic. For this investigation, the algorithm was run on the ‘profile_df’ data set and split into 6 groups. The mean values of those groupings are below:

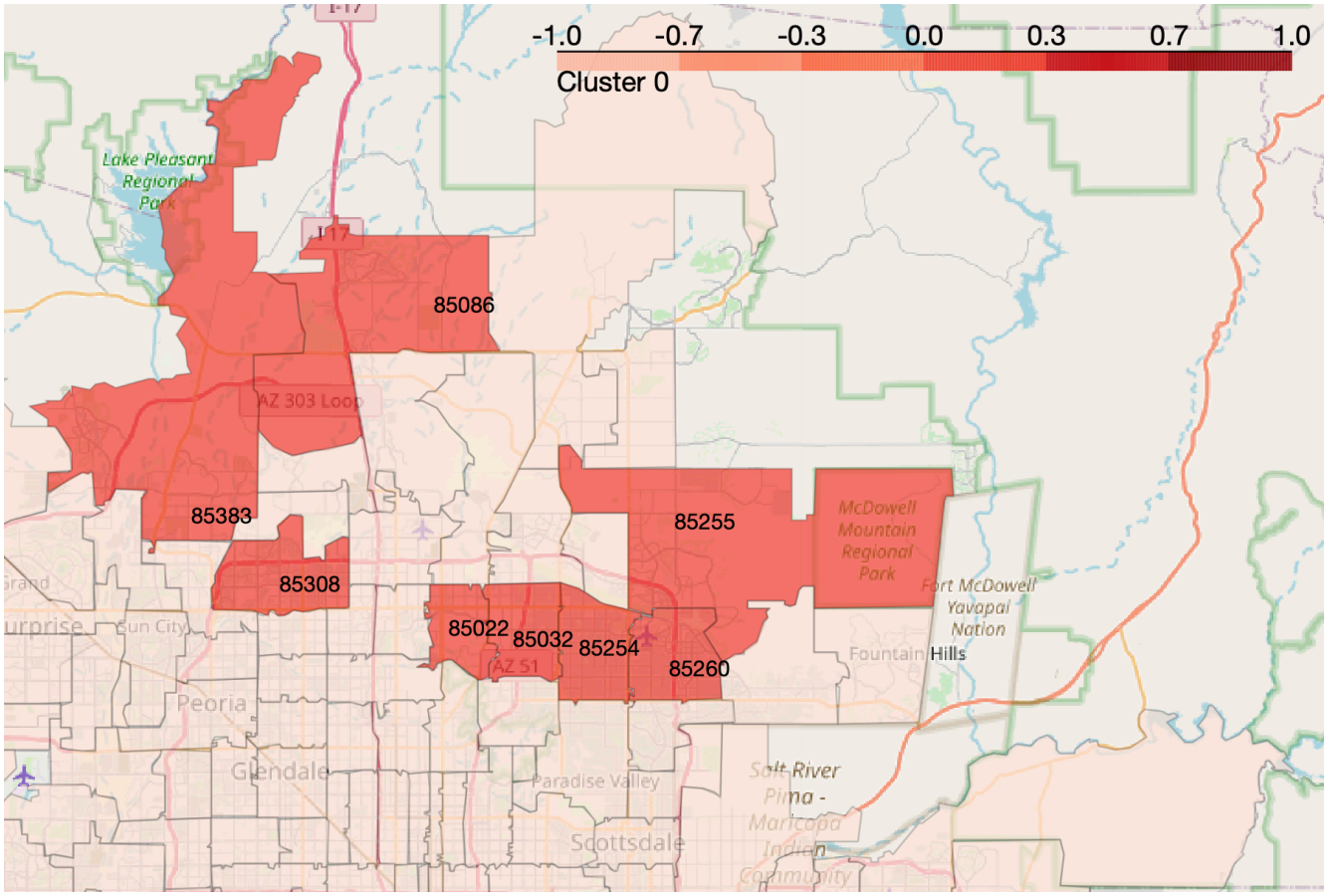
Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
14,207.50	16,126.50	114,757.75	1.00	33.67	-112.03	0
3,300.80	2,649.00	70,631.45	0.55	33.56	-112.17	1
10,702.11	6,675.78	73,116.33	3.00	33.49	-112.05	2
10,959.17	4,711.46	66,199.29	0.09	33.51	-112.17	3
5,794.87	8,127.33	135,393.00	0.13	33.62	-111.96	4
11,238.50	10,354.14	95,907.50	0.50	33.35	-111.80	5

Clusters 1,2 and 3 are below our target income level, but cluster 3 has the highest density of Medical Facilities. Clusters 0 and 4 have the highest income levels, low density of Medical Facilities. A visual representation will better show how these clusters are related.

While reviewing the cluster maps – keep in mind that Cluster 0 and 4 have the highest income levels and our initial investigation indicates that on average, or target population in is the north and east quadrant of the Metro Phoenix area.

Cluster 0

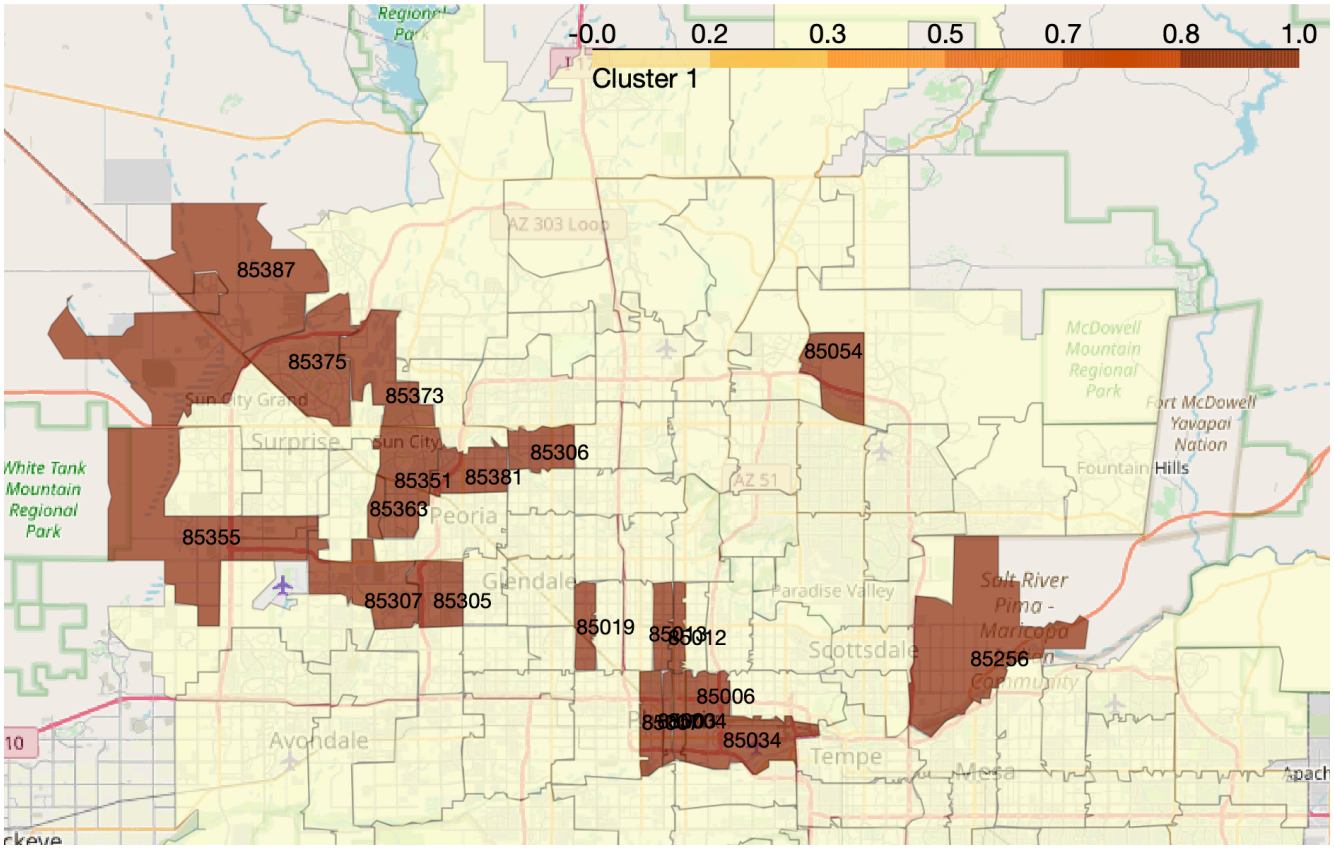
Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
14,207.50	16,126.50	114,757.75	1.00	33.67	-112.03	0



This cluster has a high mean income and education level. As expected, it is located primarily in the north, but split almost evenly east and west. Half of the cluster is in the Northeast Quadrant.

Cluster 1

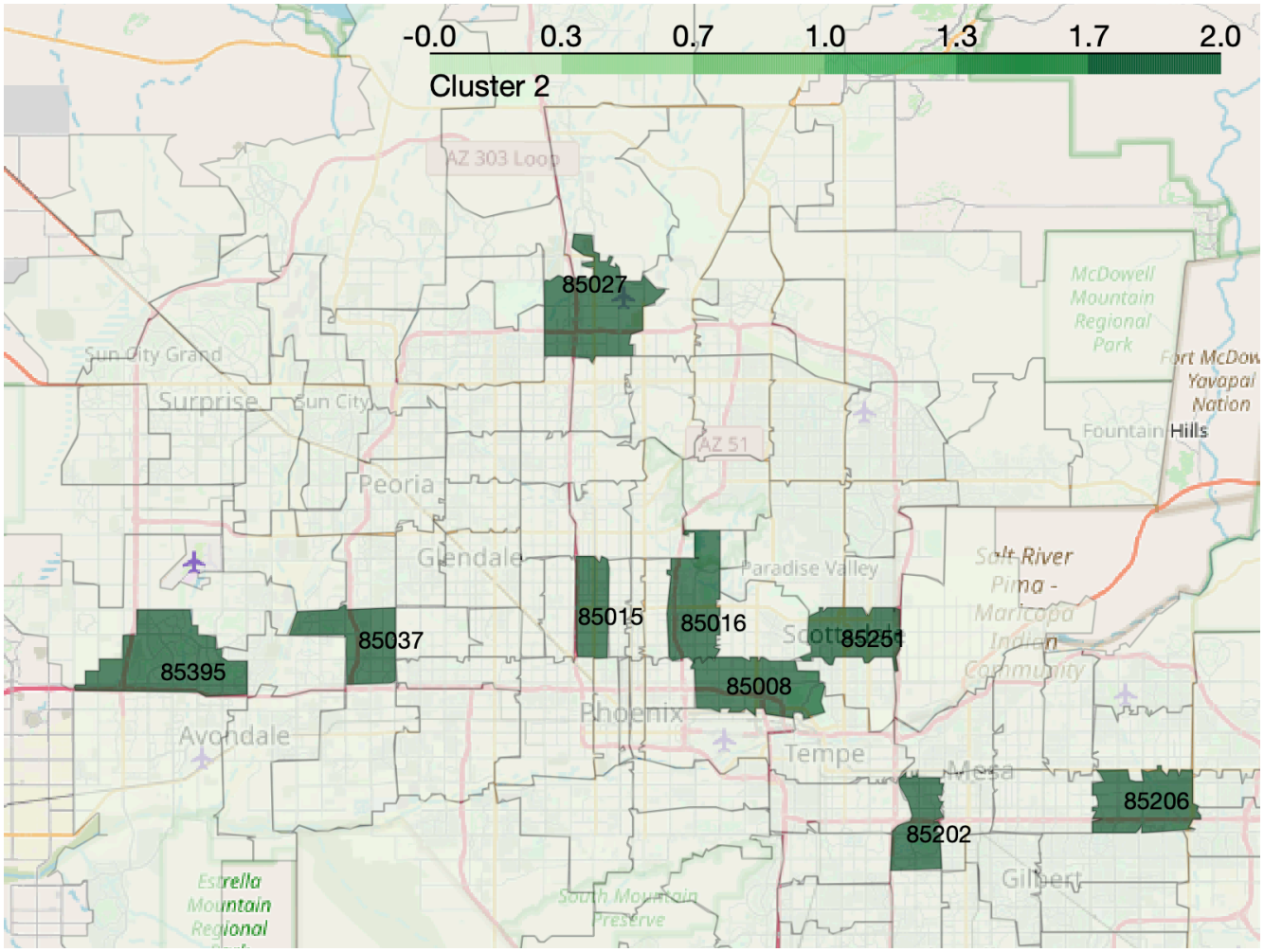
Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
3,300.80	2,649.00	70,631.45	0.55	33.56	-112.17	1



Cluster 1 is not of much interest; the income level is well below our target ($90K) and mostly central and west.

Cluster 2

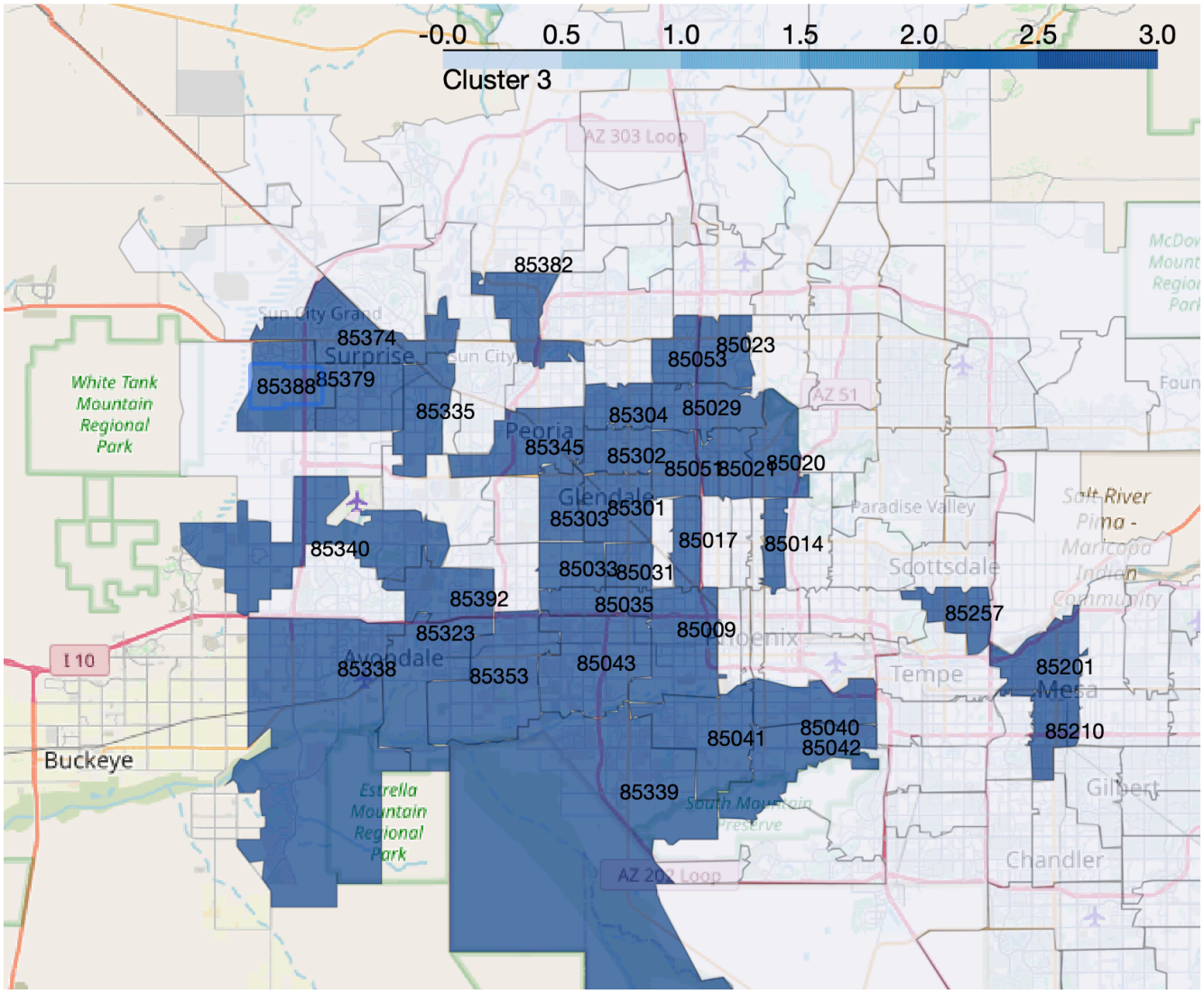
Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
10,702.11	6,675.78	73,116.33	3.00	33.49	-112.05	2



While Cluster 2 has a lower than desired target income level, the medical facility density is impressive. Note the cluster of ZIPs around the town of Scottsdale. 85008, 85015, 85016 and 85251. This area should to be considered in the final review.

Cluster 3

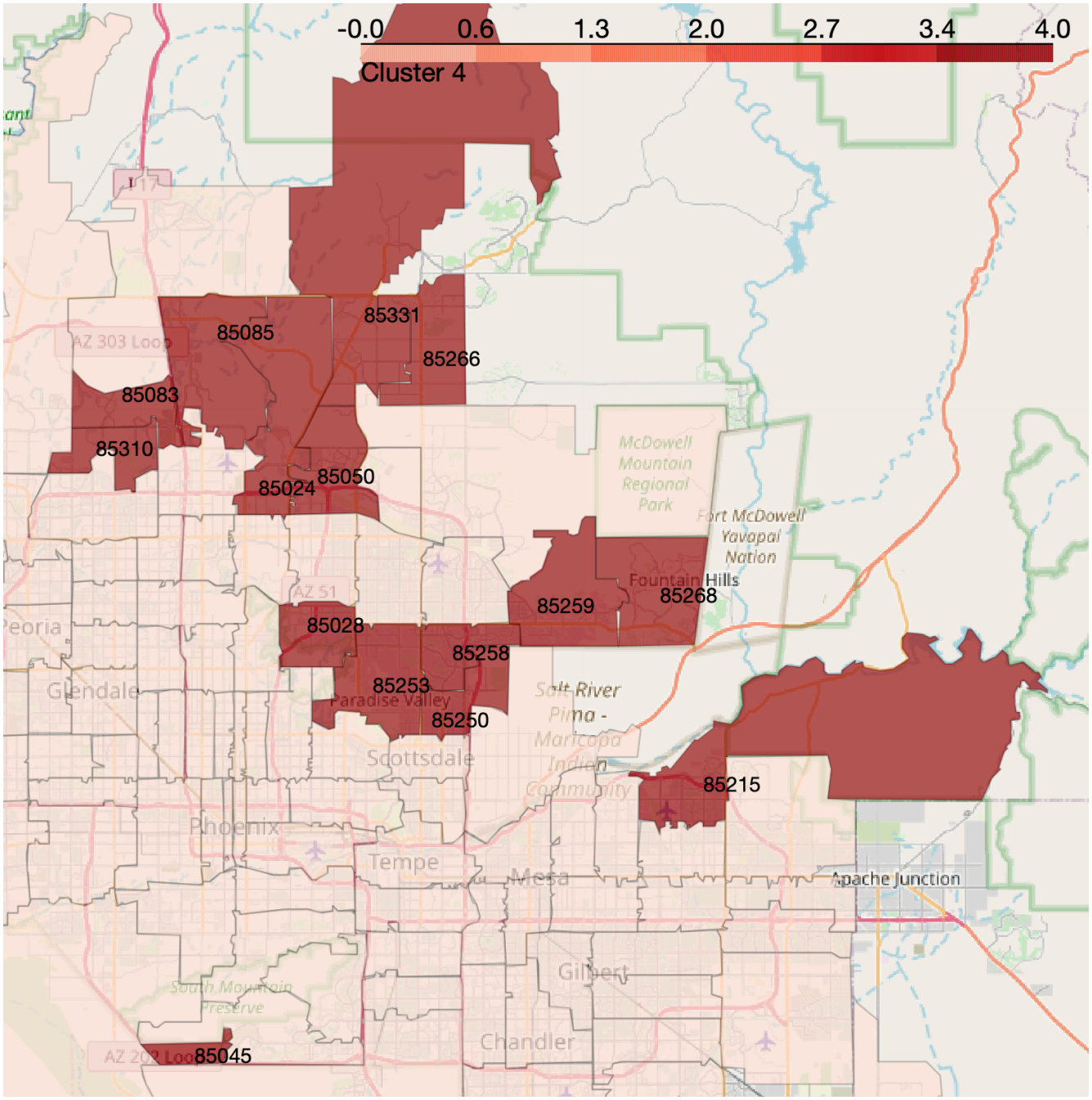
Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
10,959.17	4,711.46	66,199.29	0.09	33.51	-112.17	3



Cluster 3 does not have sufficient education or income levels to be of interest.

Cluster 4

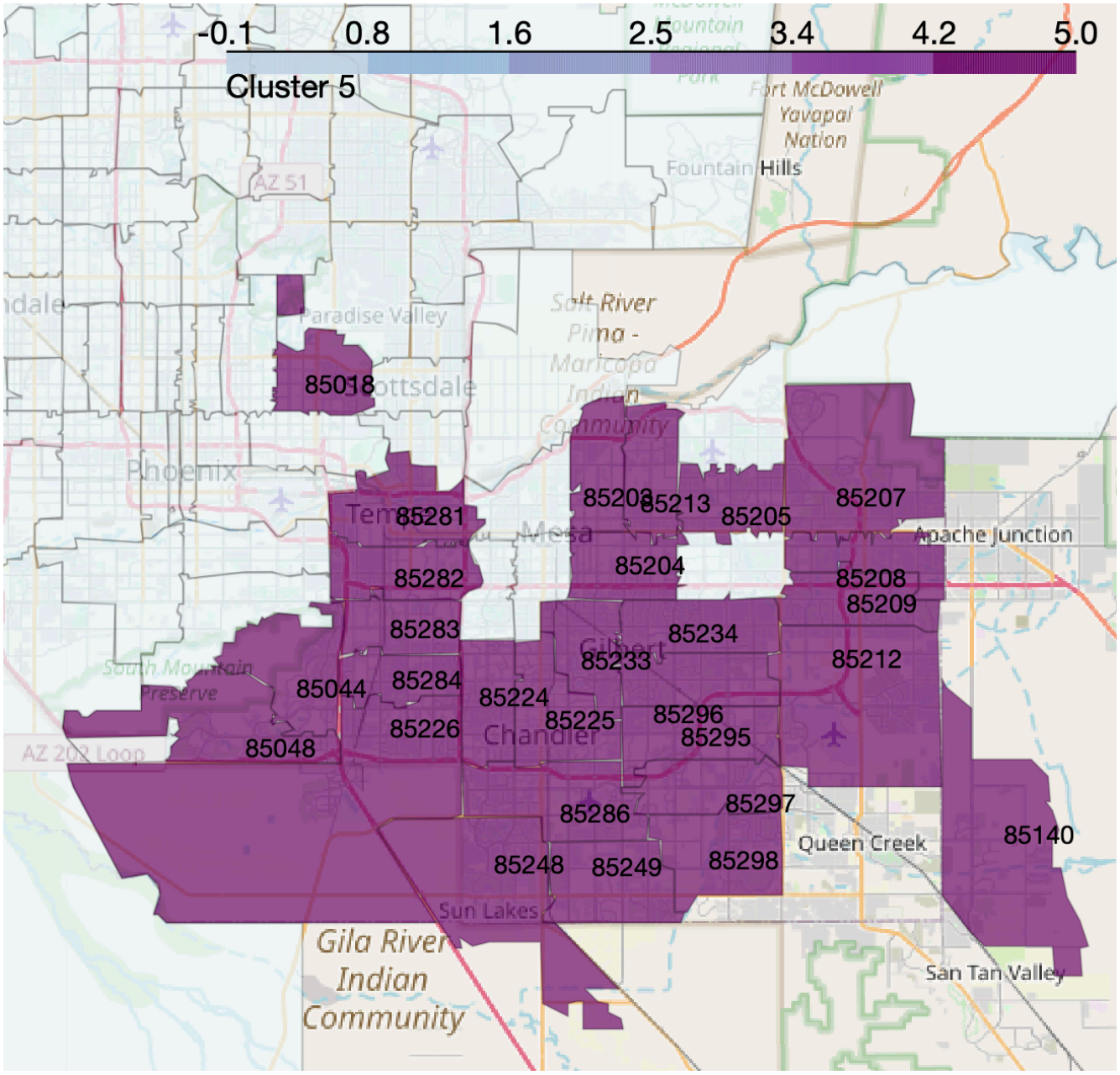
Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
5,794.87	8,127.33	135,393.00	0.13	33.62	-111.96	4



Cluster 4 has the highest income level the K-Means clusters. 84045 is an outlier, and not of interest to ProGym. The cluster of ZIP codes north of Scottsdale, again adds weight to this area. The areas to the far north, do not have a sufficient density of medical facilities, as does the east outlier, 85215.

Cluster 5

Age Pop	Edu Pop	Income Pop	Med Count	Latitude	Longitude	Cluster
11,238.50	10,354.14	95,907.50	0.50	33.35	-111.80	5



This cluster has sufficient income and education to be of interest, however, the dominant high-paying professional positions in the part of Phoenix are high-tech, and not medical. Both Chandler and Gilbert are home to electronics research, development and manufacturing. This cluster shouldn't be ignored, but it is of lower priority than Cluster 4 and Cluster 2.

Facility Placement Discussion

Metro Phoenix is a diverse and vibrant metropolitan area. A community of people who fit the ProGym client profile as well as geographies that fit the ProGym facility profile are numerous. The data indicate that a ProGym facility could be successful in several areas, but the focus of this exploration is to find the best area.

As we look around the Phoenix ZIP code data, the area to the North East of central Phoenix is the prime area for further investigation. The surrounding ZIP areas – South and West of central Phoenix, are missing the key elements of the ProGym Profile, particularly – higher education and dense medical facilities.

Focusing on the North-east quadrant, and combining Clusters 2 (high medical density) and Cluster 4 (highest education and income) yields a target area somewhere north of the city of Scottsdale.

Conclusion

It is the recommendation of BusiHAX that the new facility be constructed in or near ZIP Codes 85250 or 85258. Both of these ZIPs were grouped in Cluster 2. They are within 15-minute driving distance from two Cluster 1 ZIPs – 85251 and 85215. Interstate 101 passes directly through both of these areas, making travel from other areas much easier.

Furthermore, ZIPs 85250 and 85258 are within 15 driving minutes from the Mayo Clinic, Phoenix Children's hospital, HonorHealth Medical Center and Lincoln Medical. The area surrounding these ZIP Codes (mapped below) is home to over 120,000 individuals who match the ProGym client profile.