Determining the Optimum Location for a Small to Medium Dog Kennel

Ian Fleury
Mar 20, 2021

1. Introduction/Business Problem

1.1 Background

I have been approached by a client who started a few years back to pursue her childhood dream of breeding American Cocker Spaniels. What started as a simple pursuit has turned, over time, into a burgeoning hobby. She has exceeded the dog limit per household enforced in her current location in Eastern Canada. The city where she resides has bylaws that only allow for 3 dogs and one litter per household. She has citizenship in both Canada and the United States and would like to move her kennel location to the U.S. after finding an optimum area that suits her needs. She wants to make an informed decision in determining where she should move next. A well-planned gathering and interpreting of the appropriate data should help to determine the best fit for her.

2. Data

2.1 Data requirements

The client would like to know which states are considered the best for raising dogs and which ones should be avoided. This would first require data related to which states contain the healthiest environments for animal livelihood. Secondly, there needs to be a comparison of each state's laws concerning dog breeding and whether they are favorable to the breeder or not. She does not currently make enough yearly from puppy sales to incur taxation, and she would prefer to stay somewhat under the radar of government regulations and fees if possible. Once she has selected which state to move to, she would like to have further data on a sampling of the cities within that state (starting with the most populated) to determine which area to begin looking for available real estate. She wants to be close to a populated city for the convenience of vet care, entertainment, dining establishments, etc., but prefers not to reside within the city limits due to the denser living spaces and higher probability of noise complaints.

2.2 Conditions

The client would like to find a state that is relatively lenient in its dog breeding laws and regulations. She has heard unsettling accounts from breeders that live in areas that have stiff legislations, intrusive inspections, high annual fees, and steep fines and penalties for non-compliant kennels. She currently owns 10 dogs, some that are breeding age, and others that have not reached maturity, but could see that number double in the next 5 – 10 years. She would like, if possible, to find a state that will allow her to kennel 20 dogs without much issue. She would also like to be able to produce at least 3 litters per year, leading to approximately 15 sales transactions annually without getting bogged down with too much red tape.

The client would also like her new residence to be able to meet her recreational needs as well. She expressed a desire to be near a city that has some good walking parks, some art exhibits, theatrical events, museums, and possibly even an animal preserve or a zoo. When asked about dining, she said that she is not keen on fine dining, but prefers good pizza places, taco vendors, domestic beer establishments, specialty coffee places (fresh roasted), and the occasional espresso vendor.

She has expressed a desire to be outside of the city limits, but would still like to be reasonably close to a well-rated vet hospital for when emergency care is needed and time is of the essence. She would also like to be within a short traveling distance to a good boarding facility where she can have her dogs looked after when she needs to travel or rotate breeding pairs.

2.3 Data Sources

To find the data that I require will involve seeking out expertise on where the best places to raise healthy pets are. I will need both experience-based health research and legal information on what each state allows for breeding animals. The American Kennel Club's (AKC) website is a good place to start for the first part, and a site that deals with animal law for the second. The client's choice of recreational establishments can be searched through the Foursquare API, and filtered according to keywords that the client has given.

Once the client is satisfied with a particular city area, then I can search for less populated areas on the outskirts of the city that are more suitable for raising dogs. I can use various mapping sites for this task. I will then look for a dog breeder database which details the location of local breeders. Knowing where the local breeders are clustered will help to determine which areas have the proper amount of acreage and the right terrain for raising and breeding dogs.

I can then use Foursquare and other databases, such as Yelp, to retrieve location data for local veterinary clinics, dog boarding facilities, and Pet supply stores. The client currently has membership with specific pet stores already, so she knows what she is looking for, but her choice of vets and boarding facilities will be dependent on how well they are rated and the customer reviews. Foursquare and Yelp both have business endpoints that may be helpful in that regard.

3. Methodology

3.1 Data acquisition

For the task of determining which state would be the best fit, I first visited the website for the American Kennel Club. The AKC registers over 250,000 pedigree and crossbreed dogs every year, so they have a good bit of information on how and where to start a successful kennel. I scraped an article from their expert-advice section called "*10 Best (and Worst) States for Your Pet's Health*." Next, I found an article from the Animal Legal & Historical Center in Michigan State University. This article gives an overview of commercial pet breeder laws per state and I was able to scrape a table from it. I used both BeautifulSoup and lxml for the website scraping process. Once a state was chosen, I used the Foursquare API to search through a selection of cities to look for city attractions that matched the keywords that the client had given.

After finding a satisfactory area, I scraped data from www.gomapper.com and www.withinhours.com , which both gave a list of surrounding towns and cities within 25 miles of Wichita. Next, I used BeautifulSoup to scrape the AKC website again to find a list of some of the other breeders in the area and map out where they are situated. Lastly, I used both the Foursquare and Yelp APIs to discover the locations of nearby vet clinics, pet stores, and boarding kennels. Although Foursquare sometimes provides ratings, and customer reviews for venues in its database, it did not give any helpful information in these particular cases, so the Yelp API was used for the purpose of retrieving ratings and reviews.

3.2 Data cleaning

The data scraped from the AKC website came in the form of two lists, one being the top ten best states for pets' health, and the second being the top ten worst. When converting the lists to data frames, I added a 'score' column to give a positive weight of +1 to the best states and a negative weight of -1 to the worst states. The data from the article detailing breeder laws per state contained some columns with unnecessary information, which were removed, and a good deal of empty and incomplete rows, which were also removed. Long passages of text were converted to be able to be represented solely by integers.

When gathering data to compare the attractions of ten sample cities, there was a redundancy in some of the results due to the same keywords being queried in several different ways. The duplicates were dropped for easier comparison. When retrieving the locations of surrounding towns within 25 miles, there were some redundancies, some outliers that did not belong, and a few wrong latitude/longitude coordinates. All of these were either fixed or removed. The data regarding other dog breeders in the area contained several locations having different names for the same site, but that was due to the same breeder being listed under more than one breed and did not harm the results in any way.

The vet clinic location data required the removal of unnecessary columns of information, as did the pet store locations. The pet store data also had an entry with the wrong latitude/longitude coordinates, which was fixed. For the data regarding boarding kennels, I looked at results from both Foursquare and Yelp for comparison. Foursquare did not have any available rating information on the boarders, but Yelp did, so I went with the results from the Yelp API, removed unnecessary columns and then made an additional column to indicate which locations also appeared on Foursquare.

3.3 Feature selection

Only two features were used for the top ten best and worst states: the name of the states, and the value indicators (+1 or -1) that I added under the heading of 'Pet Health Score.' I did not need any other features for this section, since I intended to merge these columns with the next dataset. The next set, which details the pet breeder laws per state, has a common feature with the first called "State." It is this feature that was used to merge the two data sets. The other key feature is called "Definition of commercial breeder and licensing requirements." The information contained in this column was redistributed as three new features: "Maximum number of dogs", "Maximum puppy sales per year", and "Maximum litters per year." After the "Pet Health Score" feature was merged to the data frame, another feature was added called "Total Score" to sum the total scores for each row.

The data frame used when exploring the attractions in the 10 Kansas cities contained solely the query results for Wichita, Overland Park, Kansas City, Olathe, Topeka, Lawrence, Shawnee, Manhattan, Lenexa, and Salina. The features used for the surrounding cities, breeders, and pet shops were all set up essentially the same way with one feature being the name, and the others being the latitude and longitude coordinates, and the distance in miles from the city center (Wichita.) The vet clinic and boarding facility data follows the same suit, but two more features are added: "rating", and "review_count." These features represent customer feedback, which is an essential indicator of both the quality and service of these two venue types.

3.4 Calculation of target variables

To determine which state would be the best fit for the client's needs, a target variable was created that would give a total score for each state listed. This variable contained the sum, for each state, of the four numeric features: "Maximum number of dogs", "Maximum puppy sales per year", "Maximum litters per year", and "Pet Health Score." Having a total score per state made it easier to use a bar chart to visualize which states have the highest scores. Similarly, when determining which city to move near to, I could calculate the best choice by doing a simple count of the number of keyword matches generated per sample city.

When determining the best vet clinics and pet boarders, the target variable was a combination of two features: "rating" and "review_count." When sorting by rating, the Yelp API, does not just return the highest ratings, but also takes into account the number of people rating/reviewing it, similar to a Bayesian average, so that the results are not improperly weighted.

3.5 Data manipulation

After parsing the "*pet breeder laws per state*" table from its website, I had to extract the numbers contained within the text. For instance, there would be a passage such as: "*The sale or exchange of up to, and including, three litters of puppies within a twelve month period shall not be considered a kennel*." I would then add the integer 3 to the "Maximum litters per year" column in that row. If the text mentioned other measurements as well, then I would likewise fill them in. If the other measurements were not mentioned, then I was able to infer some of their values. I made an assumption that, since the average cocker spaniel litter is 5 puppies, that a "Maximum litters per year" value of 3 would be equivalent to a "Maximum puppy sales per year" of 15 (3 litters x 5 puppies) and vice versa. If the text did not state a value for "Maximum number of dogs", but allowed the breeder to have a certain amount of litters, then I would put a value of 2 for "Maximum number of dogs" (the minimum amount of breeding stock it takes to produce litters.)

Once I felt that I had the best approximation of values for all of the categories, I compared those values to the client's desired goal of 20 dogs in residency, 3 litters per year, and 15 puppy sales per year. If a state's laws allowed the client to operate within her goal, then I would give that state a score of 1 for each feature that met or exceeded the client's upper limits, and 0 for each feature that did not meet it. The scores that each state received would be added, along with their "Pet Health Score", to give a "Total Score" result. This would be the target variable used in the bar graph for determining the most lenient, and likely healthy, state.

Another instance of data manipulation was the use of “vincenty”, a python module that performs distance calculation between latitude and longitude coordinates through the use of a complex geodesy formula. I used this module when the data that was scraped did not include distance measurements, e.g. the distance from Wichita’s center point. When preparing the data for the predictive models, I used vincenty heavily to retrieve distance measurements from each surrounding city/town to all other points of interest, namely the other breeders in the area, the top ten vet clinics in the area, the top ten boarding facilities in the area, and the client’s preferred pet stores. Having the data in this form helped to prepare a basis for running clustering algorithms in the models.

3.6 Predictive Modeling

I used two types of classification models for predictive modeling in this case study: DBSCAN and K-Means. Both are popularly used for clustering data, and both have different approaches and produce somewhat different results.

To use DBSCAN, it was not necessary for the data to be normalized or fit, but rather the process relied on the latitude/longitude coordinates from each city to position the cluster points. The coordinates were made into arrays and converted to pixel locations, under the headings of “xm” and “ym”, to be used with Basemap. The points were then projected onto a map-like background for visualization. Each row, containing the names of cities, was classified into one of the following labels: “-1”, “0”, “1”, “2”, or “3”, which was then placed under the heading of “Clus_Db”. Since DBSCAN accounts for both noise and outliers, the label “-1” was given to the entries that were considered to be outliers. Using Basemap again, I was able to visualize the resulting five clusters, which were labeled by number and color coded in light blue, green, dark blue, orange, and gray (outliers) to be able to distinguish them.

Using K-Means required the data to be put into an array and then run through the preprocessing.StandardScaler().fit().transform() function. I had to manually choose how many clusters to split the data into, since it doesn’t have an algorithm for determining that. I chose four clusters, which is the same amount that the DBSCAN model produced (if the outliers are ignored). K-Means has a fit() function, which further fits the array data to the model. The model then creates labels “0” through “3”, in a similar fashion to DBSCAN, but without using the “-1” for outliers, and places them under the heading of “labels” on the data frame.

4. Results

4.1 Interpretive functions

To help interpret the DBSCAN and K-Means models, I created two functions called “avg_prox()” and “show_clus()”. The avg_prox() function takes any feature from the dataset as an input (e.g. “vet1”, “breeder1”, “boarder1”, etc.) and, using the clusters of cities created by both DBSCAN and K-Means, prints the average distance of each cluster to that input. The show_clus() function takes a cluster name from each model as an input and displays all of the cities and towns in those two clusters. Using “vet1” (the top-rated veterinarian) and “boarder1” (the top-rated boarding facility) as examples, I called the avg_prox() function to determine which optimized clusters had the lowest average distance to the inputs. For “vet1”, the closest clusters

were Cluster 0 in the DBSCAN model, and Cluster 1 in the K-Means model. The lowest average distance to “boarder1” was Cluster1 for DBSCAN, and Cluster3 for K-Means. The following table gives the results of both models side by side.

DBSCAN Cluster 0 (avg dist to ‘vet1’)	**K-Means Cluster 1 (avg dist to ‘vet1’)**	**DBSCAN Cluster 1 (avg dist to ‘boarder1’)**	**K-Means Cluster 3 (avg dist to ‘boarder1’)**
Eastborough	Bentley	Derby	Peck
Valley Center	Colwich	Haysville	Mulvane
Andover	Goddard	Rose Hill	Haysville
Furley	Valley Center	Peck	Derby
Maize	Furley	Mulvane	Benton
Brainerd	Maize	-	Park City
Bel Aire	-	-	Bel Aire
Park City	-	-	Andover
Kechi	-	-	Eastborough
-	-	-	Rose Hill
-	-	-	Kechi

As can be seen, there are <u>not</u> a lot of similarities between the clusters, i.e. they do not have very many cities in common. This is because the two venues are too far away from each other to estimate optimum locations that would satisfy both requests. Map_wichita7 verifies that these two locations (vet1 and boarder1) are a good distance apart. So, I tried running two more locations: the number two rated vet clinic and the number two rated boarding facility. The following table shows their results:

DBSCAN Cluster 0 (avg dist to ‘vet2’)	**K-Means Cluster 3 (avg dist to ‘vet2’)**	**DBSCAN Cluster 0 (avg dist to ‘boarder2’)**	**K-Means Cluster 3 (avg dist to ‘boarder2’)**
Eastborough	Peck	Eastborough	Peck
Valley Center	Mulvane	Valley Center	Mulvane
Andover	Haysville	Andover	Haysville
Furley	Derby	Furley	Derby

Maize	Benton	Maize	Benton
Brainerd	Park City	Brainerd	Park City
Bel Aire	Bel Aire	Bel Aire	Bel Aire
Park City	Andover	Park City	Andover
Kechi	Eastborough	Kechi	Eastborough
-	Rose Hill	-	Rose Hill
-	Kechi	-	Kechi

There are definitely more commonalities among these results. If these two unsupervised machine-learning models are running their algorithms accurately, then these cities represent optimum positions, relative to all points of interest on the map, with special emphasis on being in close proximity to the number two rated vet and number two rated boarding facility. Since locations like this could be ideal for the client, these might be good choices for her to start looking at real estate. The cities that are in common in both models, for both inputs, are: Eastborough, Andover, Bel Aire, Park City, and Kechi.

5. Discussion

I can appreciate how the combination of Basemap and DBSCAN allowed the results to be mapped out through latitude/longitude coordinates into a color-coded display of clusters. It is a decent model for presentation, but I did not find it to be very easy to manipulate. If I had wanted to use either more or less clusters in the model, then I would have to change the epsilon value and/or the number of minimum samples being used. Unfortunately, this produced large and sometimes unpredictable changes in how the labels were determined, greatly affecting the reliability of the clusters. DBSCAN designated a good deal of cities as outliers, which I did not require as a category, but could not find a way to avoid it. Not only was it unnecessary, but I found the model's choice of outliers to be illogical much of the time. I would have liked to eliminate the -1 label altogether and incorporate those entries into the other four clusters, but there was no apparent way of doing that.

The K-Means model was easier to manipulate, seeing it allows a choice in the amount of clusters that are used, and incorporates all of the data points within the clusters (which is useful in this particular study.) However, I did not appreciate the fact that, unlike DBSCAN, the K-Means algorithm frequently produced different results when run multiple times. The goal of the predictive model was to find the optimum cluster of locations, having the lowest mean distance from important locations, which should produce static results when the same points are being used over and over, not dynamic results. Since the outcome of the K-Means model was not consistently the same, it was difficult to know if it was actually producing the optimum clusters. Had the model been run a large amount of times, then the clusters could be fine-tuned statistically, but, in this study, it was not necessary to be that precise.

Since neither of these clustering models were without faults, I found that comparing the results of both of them together helped to give weight to any entries that were common to the two models. Sifting through the cities that show up in both algorithms, should produce a few good cities to start looking for available real estate. This is with the knowledge, of course, that there are many factors involved in choosing the right home, and the availability and desirability of property can override the necessity of it being chosen from an "optimum" cluster. Should the client want to adjust her priorities, at any time, there is definitely flexibility within the models to do so.

6. Conclusions

Using various procedures from the Data Science Methodology, I was able to take the client from having no particular destination in which to move her kennel to having a finely-tuned set of results to choose from.

The available data revealed that Kansas was the best choice of states from the list of states provided. Next, careful analysis made it clear that Wichita was the most desirable city to be in proximity to in order to fit the client's specifications. Once a suitable state and city was chosen, it was just a matter of using the city's central point as a central hub to search out the surrounding cities and towns in which to live. From there, I could also scope out how the breeders in the area were clustered, and find specific locations of veterinarians, pet boarders, and pet stores. Once all of this data was retrieved and stored, I was able to fit the dataset into two unsupervised, machine-learning models: DBSCAN and K-Means. These two models produced optimum clusters, which when measured in relation to their proximity to the number two rated vet and number two rated boarding facility, gave a common set of cities. These reoccurring cities are: Eastborough, Andover, Bel Aire, Park City, and Kechi.

Upon close examination of the results, I would eliminate Eastborough as an option since it is an enclave located within the city of Wichita and, therefore, not to the client's original specifications. Bel Aire is also a bit too close to the city, and seems to lack the amount of acreage that the client would need. My recommendation is that the client begins her search for real estate in one of the following areas first: Andover, Park City, or Kechi. If suitable land is not available in any of those locations, then I would suggest moving the search out to the second closest clusters produced by both the models and working her way out from there.