Capstone Project Abstract

CS1951A - Lorenzo De Stefani
David Han, Kota Soda, Robert Bellaire

Canine Companions

Goal

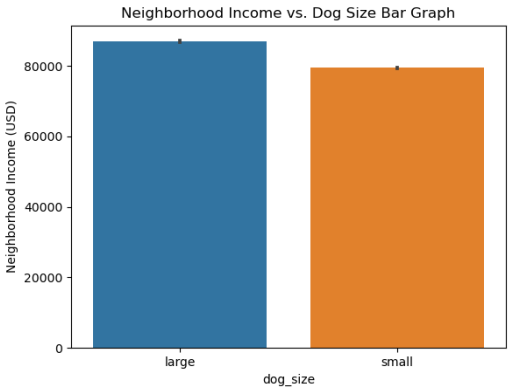
As a group of dog lovers, we sought to gather data for, evaluate, and visualize the correlation between the dog population of NYC and their neighborhood factors. By investigating this correlation, we hoped to learn more about the distribution and preferences of New York dog owners to build a suggestion algorithm. Specifically, we wanted to test the relationships between dog size vs. neighborhood income, dog energy vs. neighborhood income, and dog gender vs. neighborhood park size.

Data

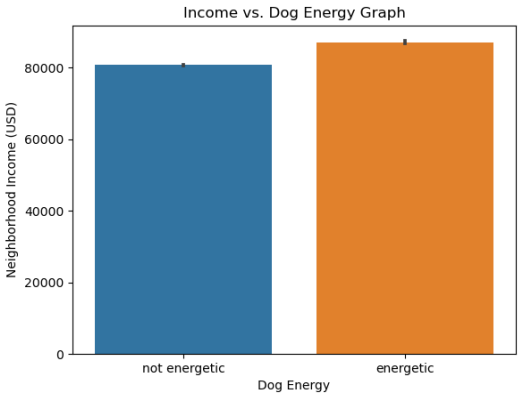
We gathered our data from a variety of datasets, filtering and combining them to create our own database with all the information that we needed. NYC Open Data provided a lot of information about the city's zip codes/neighborhoods such as public facilities, park information, and individual dogs. Additionally, to get traits like height and weight of a particular breed, we used the kaggle dataset: Dog Breed Details. Once we had this data, we wanted to clean it and filter out information that we didn't need like non-school facilities, dogs with an unknown breed, etc. Finally, with our cleaned data we created a database with individual tables and a consolidated table which joined all the information on the dog breed. Thus, for each dog in New York, we had all the information about its breed and the zipcode that it was in.

Findings

Claim #1: There is a correlation between dog size and neighborhood income (wealthier neighborhoods tend to have larger dogs)
Support for Claim #1: Conducting a two sample t-test shows that we end up with a very large T-statistic of -47 and a p-value close to 0



Claim #2: There is a correlation between dog energy and neighborhood income (wealthier neighborhoods tend to have higher energy dogs)
Support for Claim #2: Conducting a two sample t-test shows that we end up with a very large T-statistic of -31.6 and a p-value close to 0



Claim #3: There is no correlation between dog gender and neighborhood park size
Support for Claim #3: Conducting a two sample t-test shows that we end up with a p-value of 0.16, exceeding our p-threshold of 0.05.

Dog Size v. Neighborhood Income	Dog Gender v. Neighborhood Park Size	Dog Energy Level v. Neighborhood Income
T = -47.20	T = 1.38	T = -31.6
P = 0.000	P = 0.168	P = 0.000
Reject Null	Fail to Reject Null	Reject Null

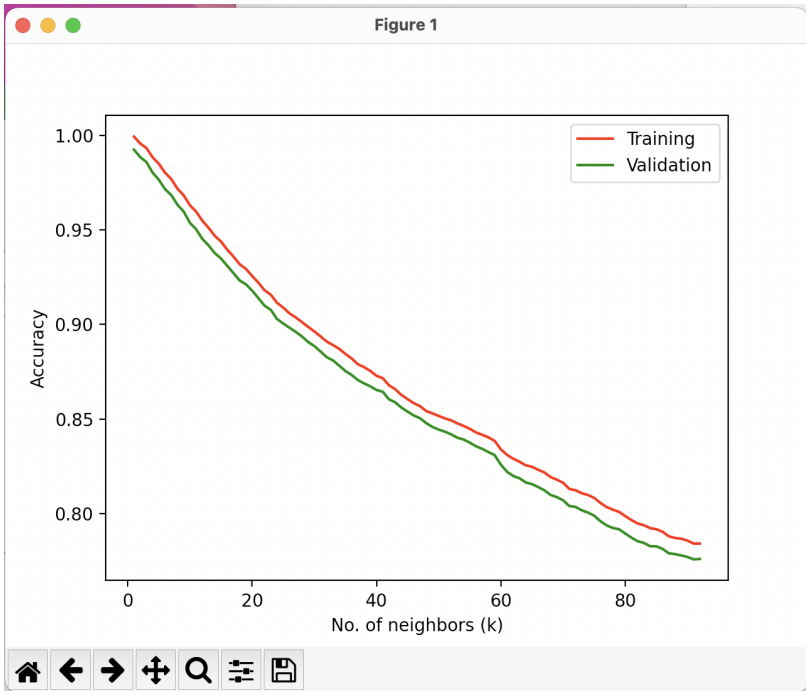
Model+Evaluation Setup

Taking the project beyond testing our hypotheses, we were hoping to use ML to create a dog breed predictor based on information and preferences of the user. With this predictor, we hoped to better match New York owners to dogs that would fit their circumstances. To start, we used a Principal Component Analysis (PCA) which took in our normalized features and reduced them to k = 6 values to decrease overfitting and computation time. We chose k = 6 because the explained ratio sum at k = 6 is 0.86. The explained ratio represents the variance that the dataset retains as the number of features decreases. The lower the explained ratio sum is, the less information we retain. Thus, we set 0.86 as a threshold for the explained ratio sum (also since we can achieve high accuracy without overfitting). Then, we used the supervised learning K-Nearest Neighbors (KNN) algorithm to train, validate, and test our model for predicting dog breeds. This algorithm worked best for our task because we wanted to make classifications based on the training of labeled data and the similarity of the new data's features. Although we considered other supervised classification algorithms such as decision trees, we went with KNN since we had a high number of dimensions in our original data, and we also wanted to prevent overfitting.

Results and Analysis

Claim #1: Decreasing the number of k values increases the accuracy of the model (reducing k in this case would be generalizing dogs into k categories).

Support for Claim #1: By plotting the number of k's vs accuracy, we can see that we achieve much higher accuracy if the k value is less.



Claim #2: 4 features is the sweet spot for PCA dimensionality reduction without overfitting or losing too much information.
Support for Claim #2: We can tweak the number of components until we find a sweet spot that gives us an explained ratio sum of 0.86 (amount of variance retained from the original data set). The value 0.8 is chosen so most information is retained while preventing overfitting.

Num Components	Explained Ratio Sum
14	1
13	0.9995448191310229
12	0.9986467138958508
11	0.9940817597418046
10	0.9816975990015423
9	0.9623978639985292
8	0.9333083789402821
7	0.9011484761739379
6	0.8603383963875986
5	0.7961332766810949
4	0.7284501807422128
3	0.6345026148760988
2	0.5141005131773999
1	0.37539511516426716