Capstone Final Report
Dog Adoptions in the United States

Yenmin Young

1.) Introduction:

Background

In the United States, pet ownership has been steadily increasing for the past three decades. An estimated 65 million households in the U.S. own a dog as a pet and about 40% of them were adopted from shelters. Many people say that it is better to adopt a dog than buy one from a breeder, because there are so many stray dogs who need rescuing. In 2024, over 3 million dogs entered a shelter in the U.S. and only 2 million dogs were adopted from them.

Peter Zheutlin, author of *Rescue Road: One Man, Thirty Thousand Dogs, and a Million Miles on the Last Hope Highway*, documents his journey with Greg Mahle, who drives trucks full of stray dogs from the South to the North to help find the dogs' forever homes. He uncovers the issues that plague the South and how they affect dogs, shelters, and the families who live there.

The South has an overpopulation of dogs due to less stringent laws around leashing and neutering them. Poverty also influences a family's ability to care for their dogs, leaving many abandoned. Because of the limited space in shelters and the constant need to take in more strays, many "unadoptable" dogs end up being euthanized to make space for "more desirable" dogs. There is now a "Great Migration" of dogs, from the South to the North, as caravans transport dogs to states that have a larger network of shelters, more resources and support for animals, and a higher demand for pets.

We have data from all the dogs listed for adoption on PetFinder.com on September 20, 2019. Our goal is to analyze and visualize the data to better understand the movement patterns around dog adoption, and to see whether it is consistent with Peter Zheutlin's book.

Guiding Questions

Which states export the most dogs and where do they go?
Which type of dogs are overrepresented in the shelters?

2.) Data:

Datasets

Three data files were used in this project. dogTravel.csv, allDogDescription.csv, and movesByLocation.csv. They contain information regarding:

- Origin of the dog
- State they are currently listed for adoption in
- Amount of dogs exported and imported in each area
- Features and characteristics about the dog (breed, size, sex, fur color, environment suitability, medical records, etc.)

Disclaimer: This data represents a single day of data (9/20/2019). Information has changed since then. The data only includes information from PetFinder.com and does not include dogs listed on other adoption agencies, websites, or off-line transactions. This data does not tell which dogs are in highest demand, are most popular, or more “adoptable”, but rather what is available for adoption on that day in the market.

Data Wrangling

Standard data cleaning was necessary for all three data files. Many columns contained unique values (such as index, id number, dog name, PetFinder URL, etc.) and redundant information (such as the date the PetFinder posting was accessed, which was 2019-09-20 for all rows). These columns were deemed useless and dropped. Any columns that were missing a majority of the values (whether a dog was removed from the shelter) or didn’t seem relevant to the problem (city name of shelter) were also dropped. Many duplicate entries were discovered and removed. Some data types needed to be corrected (for example, turning “1.0” from a string into an integer).

While cleaning up allDogs.csv, I noticed that many rows had shifted data (a ‘zip code’ in the ‘state’ column). I identified the shifted rows and fixed the data. Fortunately, the missing cells that resulted from this correction were in an irrelevant column (the date posted) which we were going to drop anyway. I also needed to do some research in order to fill in some missing zip codes, by looking at the city and state and other instances that had similar values.

The dogTravel.csv had a ‘manual’ column that had a few values that needed to override the values in the ‘found’ column. I moved the values over and dropped the ‘manual’ column. The locations listed in the ‘found’ column were inconsistent in spelling, capitalization, and scope of region (state, country, city). Some also had phrases that didn’t seem like a location. By looking at the descriptions, I could deduce the correct location most of the time. I suspect that the outlier phrases came from an algorithm that grabbed the first proper noun with capitalized spelling and input that value as the ‘found’ location, even if it wasn’t even a location. I created a dictionary to replace the outlier values and input the deduced location that the dog comes from.

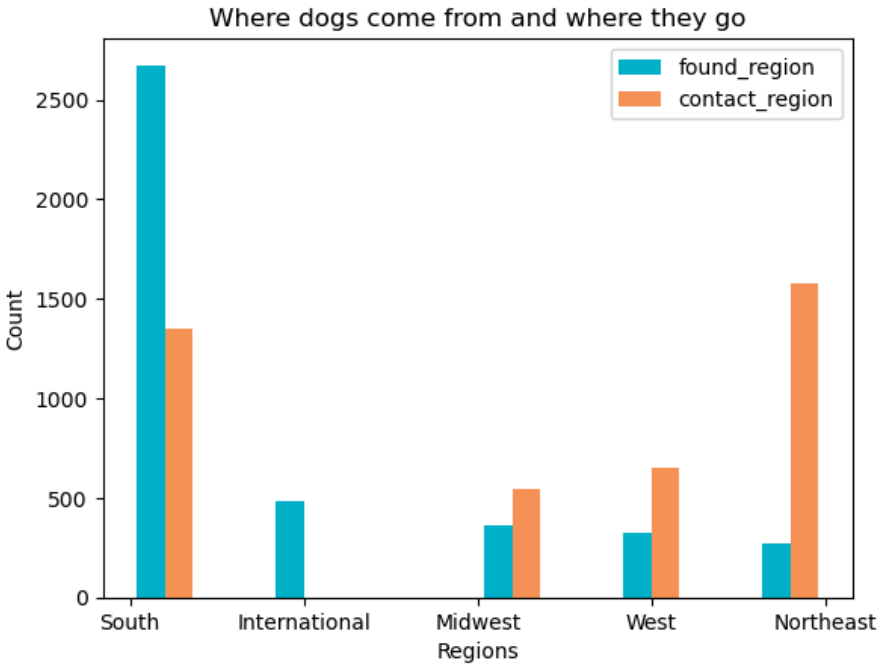
The movesByLocation.csv was relatively straight forward. There was only one incorrect row (‘Indianapolis’, which I consolidated with ‘Indiana’) that needed specific attention. I also consolidated the Caribbean islands into one region for simplicity.

To add simplicity and see larger trends and patterns, I added a ‘region’ column to all the data files (this applied twice for dogTravel.csv since they have a column for where the dog was found and another for where the dog is currently listed).

Exploratory Data Analysis

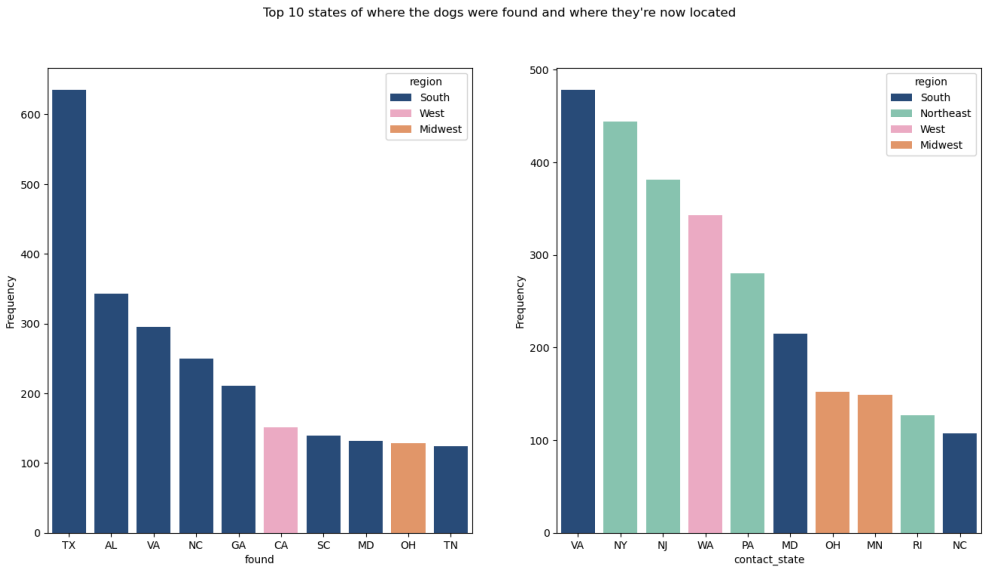
Location and Movement of Dogs

The issue of dog adoption is heavily influenced by the geo cultural landscape.

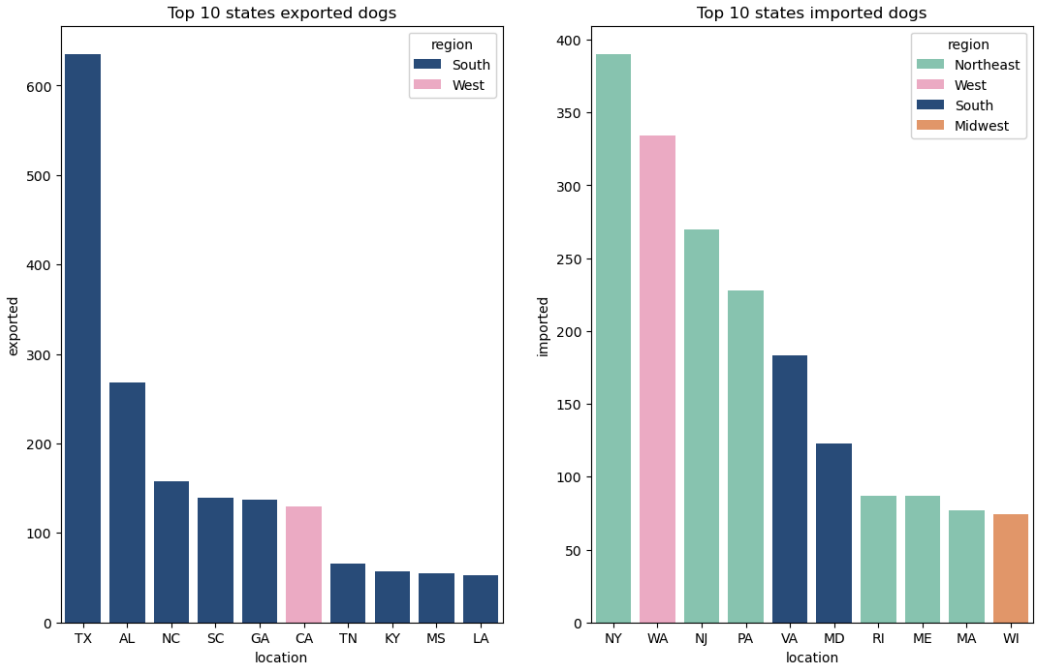


A quick count shows where dogs are found (blue) and where they are listed for adoption (orange) by region. A vast majority of dogs come from the South. By raw numbers, it seems that the Northeast and South have about the same number of dogs listed in shelters. However, it is clear that the Northeast does not supply most of their own dogs and mainly gets them from the South. The Midwest and West also rely, albeit with less intensity, on the South and other countries for dogs.

We can give a breakdown by state with the dogTravel.csv dataset.

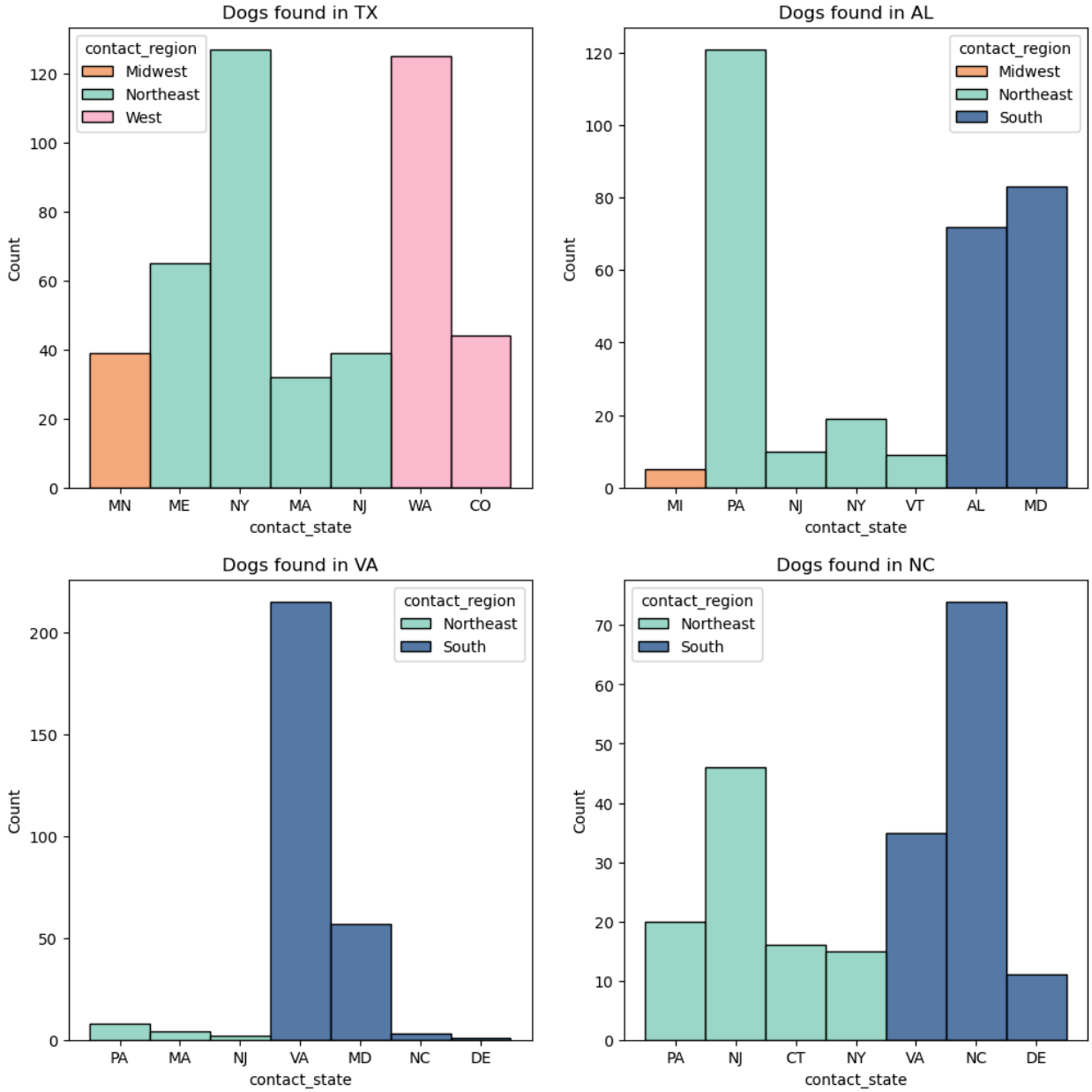


Here are the top 10 states where dogs are found. The columns are color coded by region and one can easily see that most of these states are from the South. It is quite common for a rescue owner to say that their dog is from Texas, as supported by this chart. The chart on the right shows the states that have the most listings for adoption. While Virginia has the most, the Northeast overall is still the most represented..



Comparing this data to movesbyLocation.csv, we can see that despite some minor differences, the main trend is the same: the South certainly exports the most dogs while the Northeast imports the most.

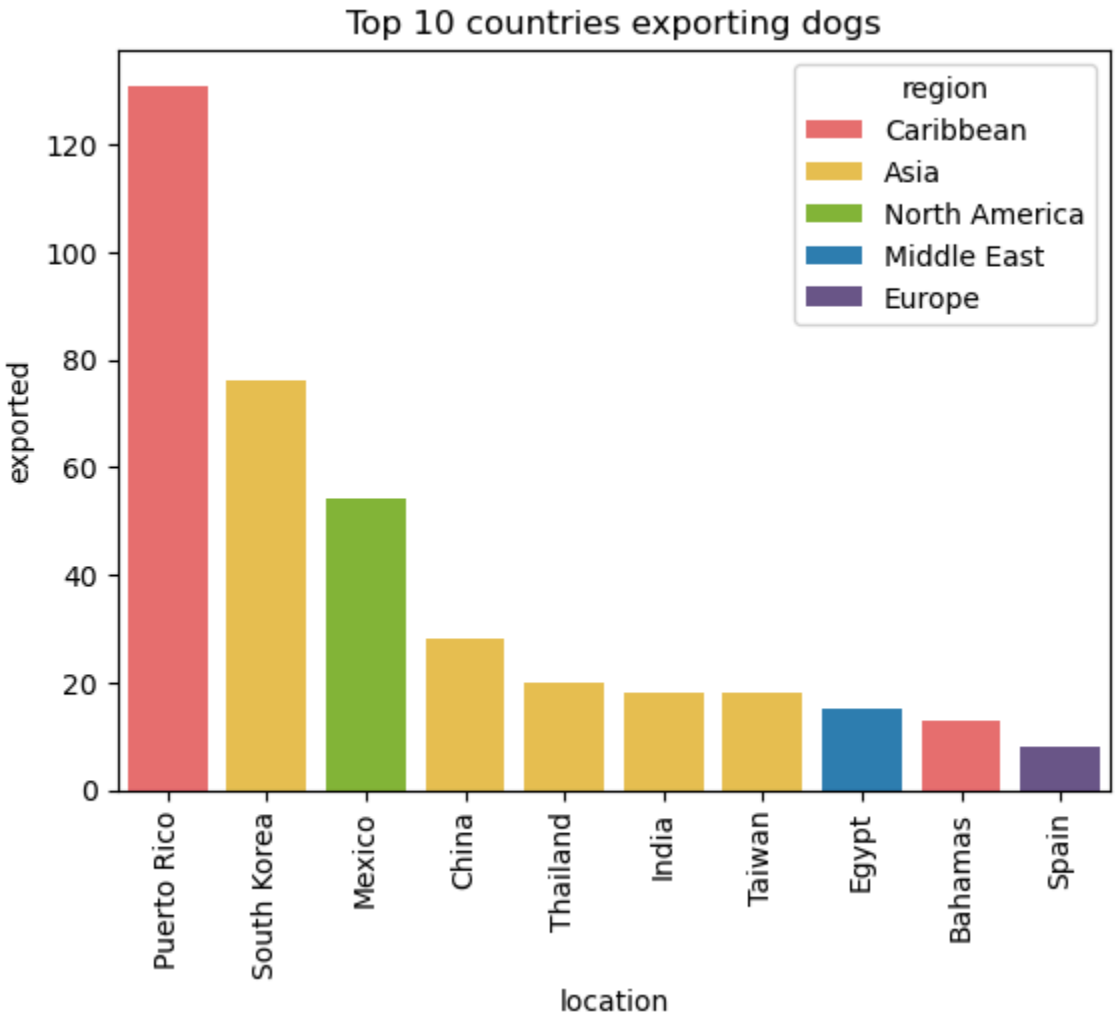
Let’s look at the top 4 states where the dogs are found. Where do they tend to go?



Each state paints its own story. Notice how Texas supplies mainly to New York and Washington, and Alabama supplies to Pennsylvania and Maryland. This may reflect the partnerships between rescue organizations between these states. Texas also supplies many Northeastern states because they have an “adoption caravan”, where rescue organizations make stops along their route to Maine, hosting adoption events as they go.

On the other hand, it looks like Virginia and North Carolina mainly supply themselves - their dogs stay local. (Maryland, Virginia, and North Carolina are all adjacent and can be considered local to each other.)

Let’s look at the dogs crossing international borders to arrive in the United States.

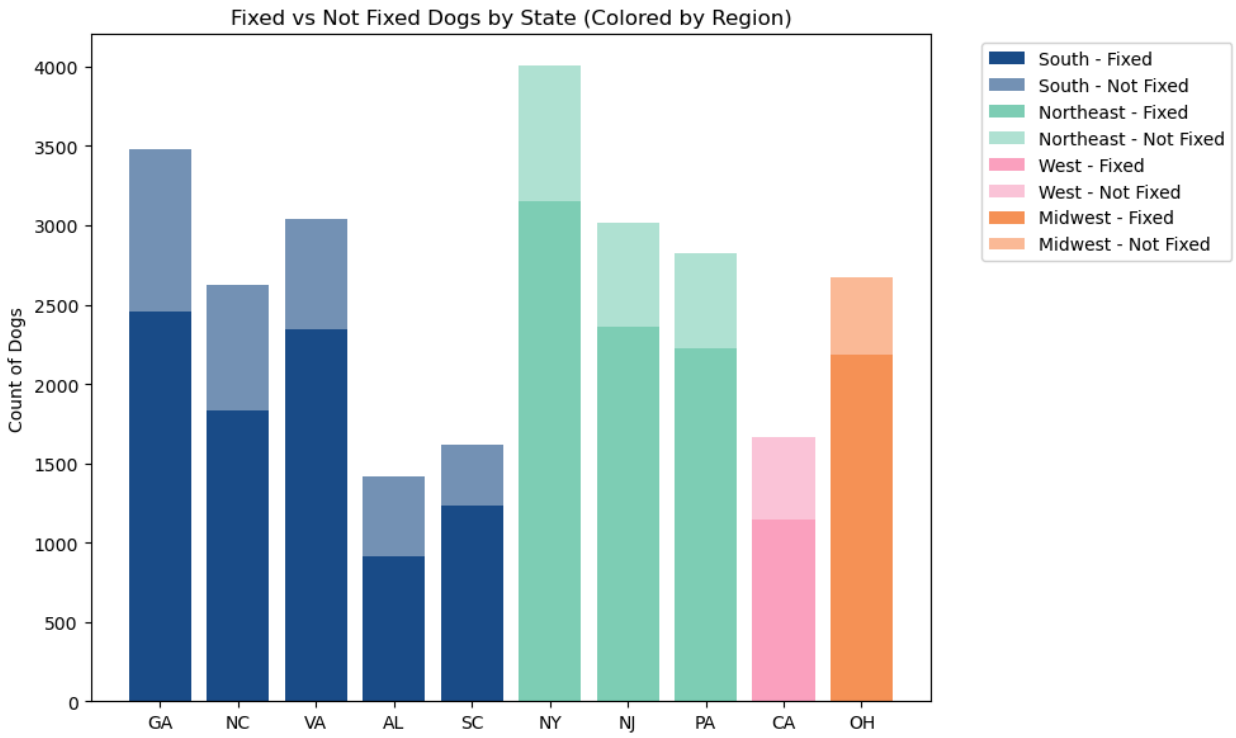


A majority come from Puerto Rico, which also has an overabundance of stray dogs (called ‘satos’) due to reasons similar to the Southern continental US. A combination of poverty and resources, and a lack of neutering and government infrastructure to manage animals has led to an overpopulation of stray dogs.

South Korea also provides a lot of dogs, though more for cultural reasons. While recent years have seen a decline in this practice due to controversial discussion, Korea has historically consumed dogs in their cuisine. Nowadays, there are organizations aimed at rescuing dogs in South Korea.

Fixed Dogs

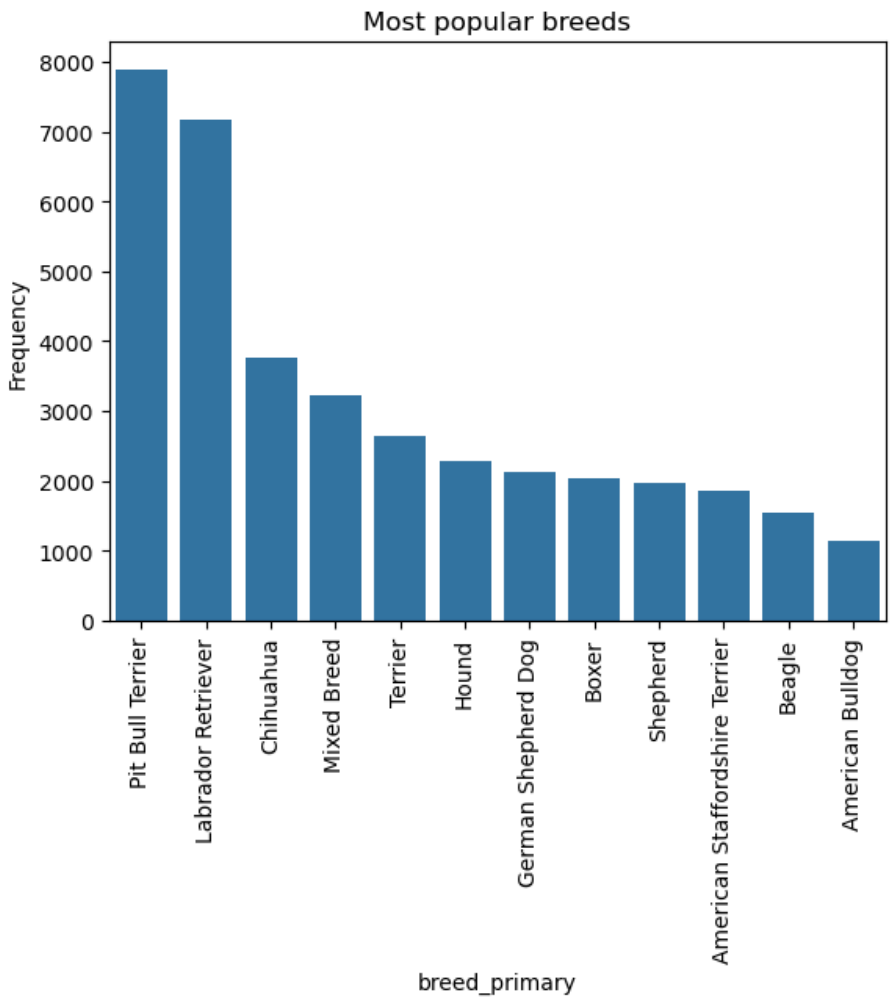
A common reason for overpopulation is due to lack of spaying and neutering of stray dogs. Here are the top 10 states with the most unfixed dogs listed.



Please note though that the states listed are of the shelter, and not where the dog was found. Therefore, we cannot conclude anything aligned with our original prediction.

Dog Breeds

Here are the breeds most commonly listed on PetFinder.com.



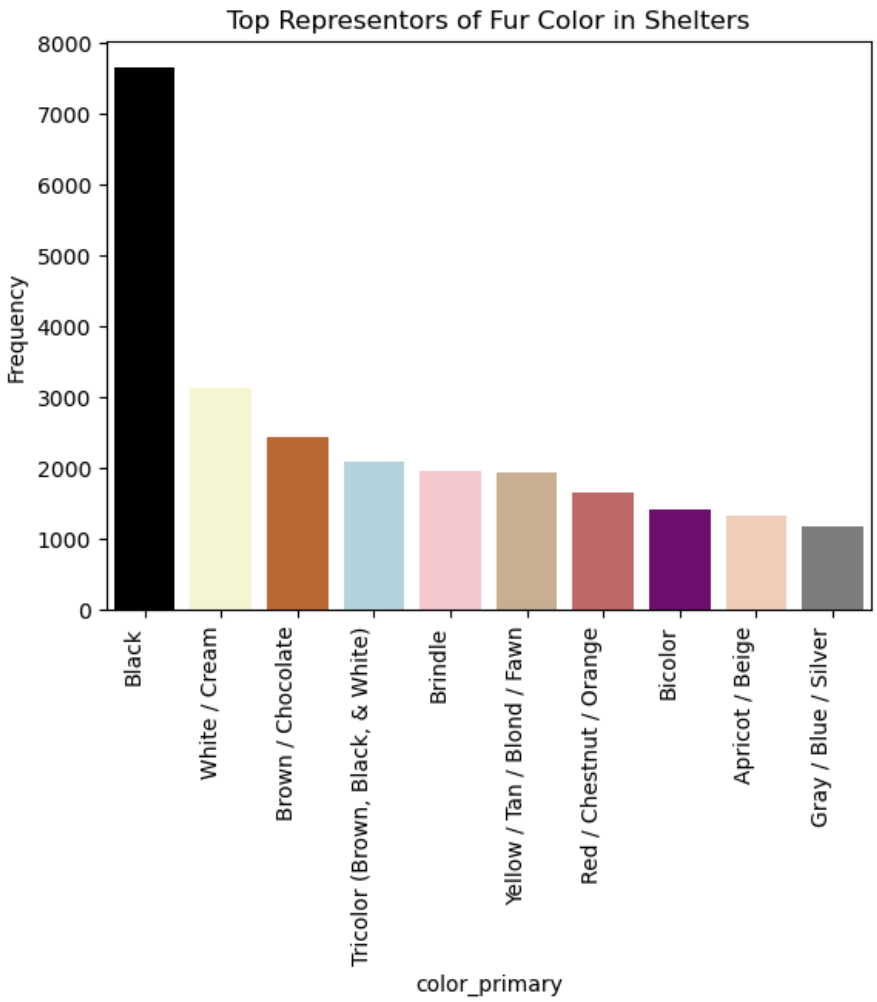
There are a significant number of Pit Bull Terriers and Labrador Retrievers. Why is this the case?

Many backyard (and unregulated) breeders often reproduced pit bulls because of their high demand and promise of high profit. Owners wanted a strong and loyal guard dog, or good contenders for dog fights. However, some owners were not ready to commit to the high effort it takes to care for a dog, or did not see their dogs fit for the fighting ring, leading to many abandoned dogs at the shelter. Unfortunately, due to misrepresentation in the media of pitbulls as ‘aggressive’ and ‘dangerous’, many potential adopters were deterred. On top of that, some housing and even city regulations banned certain ‘dangerous’ breeds, making it more difficult for pit bulls to be adopted.

Labrador retrievers are another popular breed for breeders to profit from, since they are seen as an “easy starter pet” for families. However, they often have high energy levels and require a lot of exercise and attention, making it difficult for families who don’t know how to care for a dog. Labrador retrievers are often returned to the shelter by surrendering families.

Fur Color

A vast majority of adoptable dogs have black fur.



Unfortunately, adoption shelters often see black dogs and cats left behind at higher rates than their lighter shelter mates. This could be for a variety of reasons, ranging from the negative perception of the color black, to how (in)visible they are in dimly lit shelters, or how difficult they are to distinguish between other black dogs.

3.) Modeling:

Predicting where a dog came from

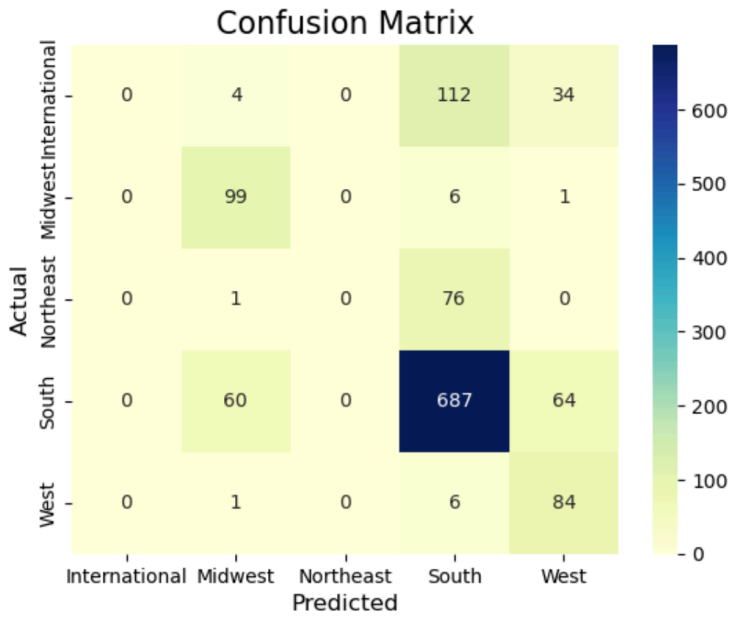
I was curious to see if I could build a model that could predict which region a dog came from based on where the shelter was. I used the second data set (dogTravel.csv, which is the only set to have both the current location and where the dog is from) to extract the features, which was simply the region that the dog came from, one-hot encoded. I experimented with models that are suited for multiclass categorical datasets, since the answer could be any of the four regions (Northeast, Midwest, West, South): Logistic regression, Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbors.

Unfortunately, the data and algorithms must've been too simple because all of the evaluation metrics churned out almost equivalent scores. This may also be because of the multiclass classification problem, in which the meaning of true and false positives gets muddled.

Model	Accuracy	Precision	Recall	F1 score	ROC_AUC	mean CV score
Logistic Regression	0.70445	0.70445	0.70445	0.70445	0.92038	0.70693
Support Vector Machine	0.70445	0.70445	0.70445	0.70445	0.92038	0.70693
Decision Tree	0.70445	0.70445	0.70445	0.70445	0.92042	0.70693
Random Forest	0.70445	0.70445	0.70445	0.70445	0.92050	0.70693
K-Nearest Neighbors	0.70445	0.70445	0.70445	0.70445	0.83638	0.68846

Ultimately, I would choose the Random Forest model due to its higher ROC_AUC score, and which tends to be more accurate than a single Decision Tree.

Looking at the Confusion Matrix, it appears that it did a fantastic job of predicting Midwestern, Southern, and Western dogs. However, it did not identify any International or Northeastern dogs.



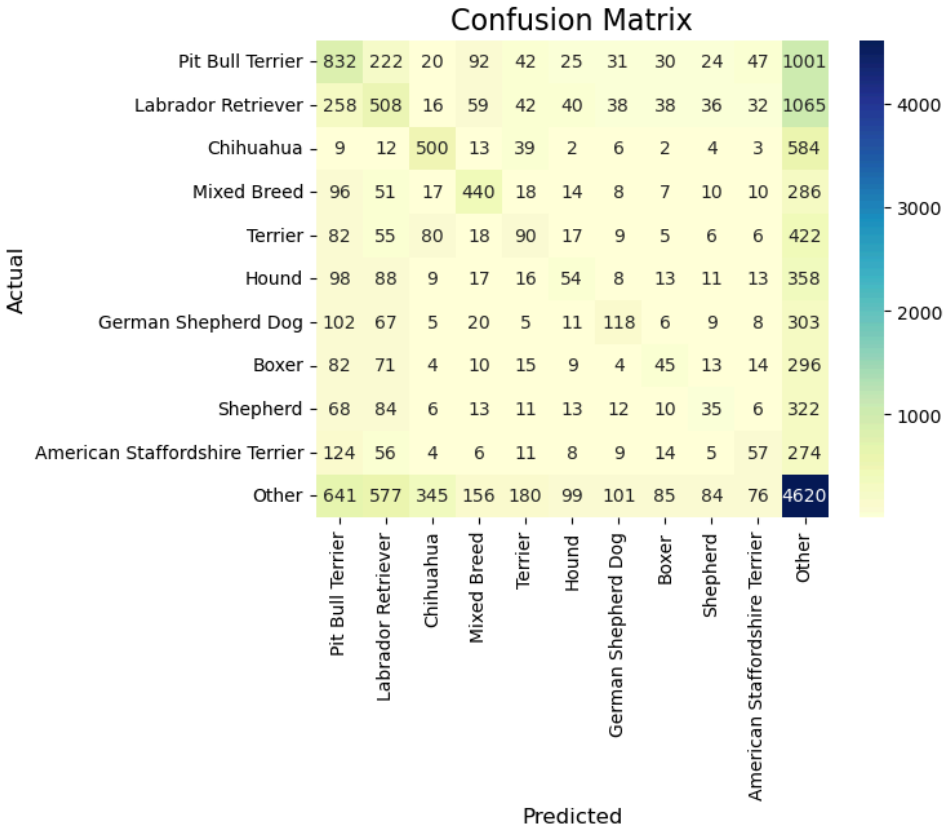
Predicting the breed of a dog

I was also curious about predicting the breed of the dog based on the traits provided in the first data set (allDogs.csv). I retained all the features that were either categorical or Boolean and gave unique categorical codes to the top 10 breeds. Any other breed listed was coded with an 'Other' code.

Not surprisingly, it is difficult to predict the breed of a dog with features that don't directly pertain to the dog's breed, and especially when even within breeds, there is much variability. But this was a fun exercise to try regardless.

	Model	Accuracy	Precision	Recall	F1 score	ROC_AUC	mean CV score
0	Logistic Regression	0.418803	0.418803	0.418803	0.418803	0.849813	0.295243
1	Decision Tree	0.389204	0.389204	0.389204	0.389204	0.748191	0.268122
2	Random Forest	0.418689	0.418689	0.418689	0.418689	0.823618	0.305397
3	K-Nearest Neighbors	0.351230	0.351230	0.351230	0.351230	0.730177	0.240484

Again, I would choose the Random Forest model because of its higher scores and its ability to incorporate randomness. While it has nearly identical values as the Logistic Regression metrics, the CV score has ultimately tipped the scales in favor of the Random Forest model.



4.) Conclusion:

Takeaways

As expected, dogs mainly come from the South and are exported to other states throughout the U.S., and especially to the Northeast where supply is low. Some states, like Virginia and North Carolina, mainly keep their dogs local. Other states, like Texas and Alabama, heavily rely on other states to take in their stray dogs.

There is an overabundance of pit bulls, labrador retrievers, and black dogs due to selective overbreeding, bias, and inability to give proper care.

Any stakeholder who is interested in addressing this issue can approach it through a number of avenues: educating potential adopters about how to properly care for a dog to prevent surrender, providing more resources to neuter dogs in Texas, and campaigning against the discrimination of pit bulls and black dogs. These are just to name some examples to start. Truly, this is a multifaceted and complex issue that has many aspects to solve.

Future Research

With stronger technical skills, I'd like to make an interactive chart overlaid on top of a map of the United States, with weighted vector arrows indicating the volume of dogs traveling from state to state. For example, if you hover over Texas, you would see arrows pop up, pointing towards the different states that the dogs are going to, with its thickness proportional to the volume. This visual representation would help paint a clearer picture of the flow of dogs than color coded charts.

The following question would require data from shelters with more extensive knowledge on the dog and its background. What traits make a dog more adoptable? Is it their looks? Their age and sex? Their behavior? Their medical background? We would need data such as how long they've been in the shelter for, and whether or not they've been adopted, as well as the reason why they entered the shelter to begin with. This would help shelters determine which dogs are more "adoptable". Unfortunately, many overcrowded shelters need to make difficult decisions to euthanize "less desirable" dogs in order to allocate resources to other dogs that do actually have a chance of survival and adoption. This knowledge would also help organizations advocate for the "less desirable" dogs and find them a forever home.