Final Project Report

Anyang Lu

Introduction

Nowadays, many people have begun to adopt pets, which can not only eliminate loneliness but also cultivate friendships between humans and animals. In the United States, many shelters and rescues have adopted stray dogs or some families have donated puppies that they cannot raise. On the website, like PetFinder, everyone can browse the information of these adoptable dogs. For families who are interested in adoption, through some simple websites For information, you can contact the foster home to adopt your own dog. Data we use in the project have already been scrapped from the website and collected in Github. By wrangling the data, I will show some visualized results in the report.

2. Data Description

This is data about adoptable dogs in each of the US states and the District of Columbia. The datasets can be viewed by the link: https://github.com/the-pudding/data/tree/master/dog-shelters. The main information on the three datasets in adoptable dogs are shown below:

allDogDescriptions.csv:

This file is collected from the PetFinder API for all adoptable dogs in each state posted by September 20, 2019. There are 58,180 rows, 36 columns in this dataset. Each row represents an individual adoptable dog in the US. Each row contains basic information about the adoptable dog, like age, sex, size, breed and etc.. Each dog has a unique ID number.

dogTravel.csv:

This file aims to show where those dogs are available and where they came from. There are 2,460 rows, 8 columns in this dataset. Each row represents a single dog that was available adoption somewhere in the US. Each of these dogs is described as having been moved from another location to their current location, mainly contain the state, city, zip code, and detailed description of the dogs.

movesByLocation.csv:

This file finds the total numbers of imports and exports for each location. There are 5 columns, 90 rows in this dataset. Each row represents how many adopted dogs are exported or imported in a specific US state or country.

3. Data Wrangling

3.1 Data Wrangling in allDogDescriptions

Dataset allDogDescriptions list basic information on every adopted dog. For example, considering age, sex, and size, we rank the number of adopted dogs of different ages, sex and size and draw a bar plot of the result. Dogs that are the most in the shelter or rescue are female, adult, and in medium size. And the male, adult in medium size dogs rank the second largest population in the adopted dogs.

age <chr>	sex <chr>	size <chr>	n <int>
Adult	Female	Medium	7246
Adult	Male	Medium	5797
Adult	Male	Large	5497
Young	Female	Medium	4912
Young	Male	Medium	4742
Adult	Female	Large	3707

6 rows

Figure 3.1: table of grouping by age, sex, size

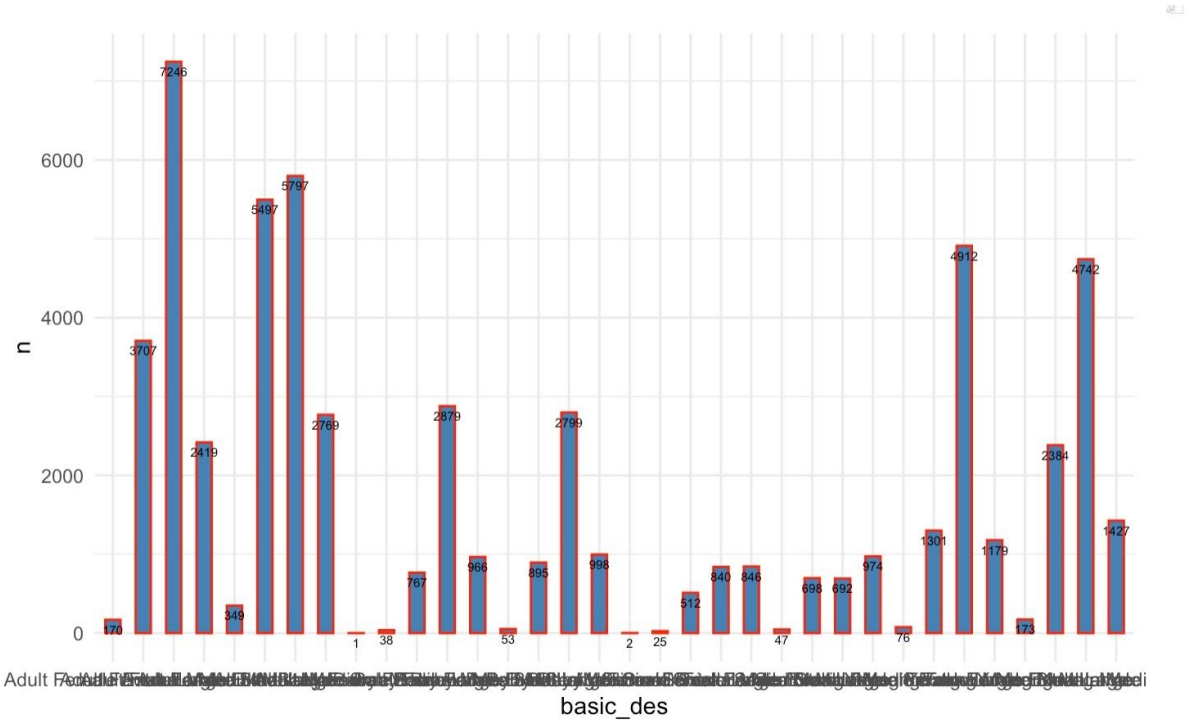

Figure 3.2: bar plot of grouping by age, sex, and size

Only considering age and sex aspects, ggplot perform a dodge-position bar plot and a stack-position bar plot in this case. In general, among the adoptable dogs, adult dogs and young dogs are common than dogs in other ages. And male dogs are more than female dogs.

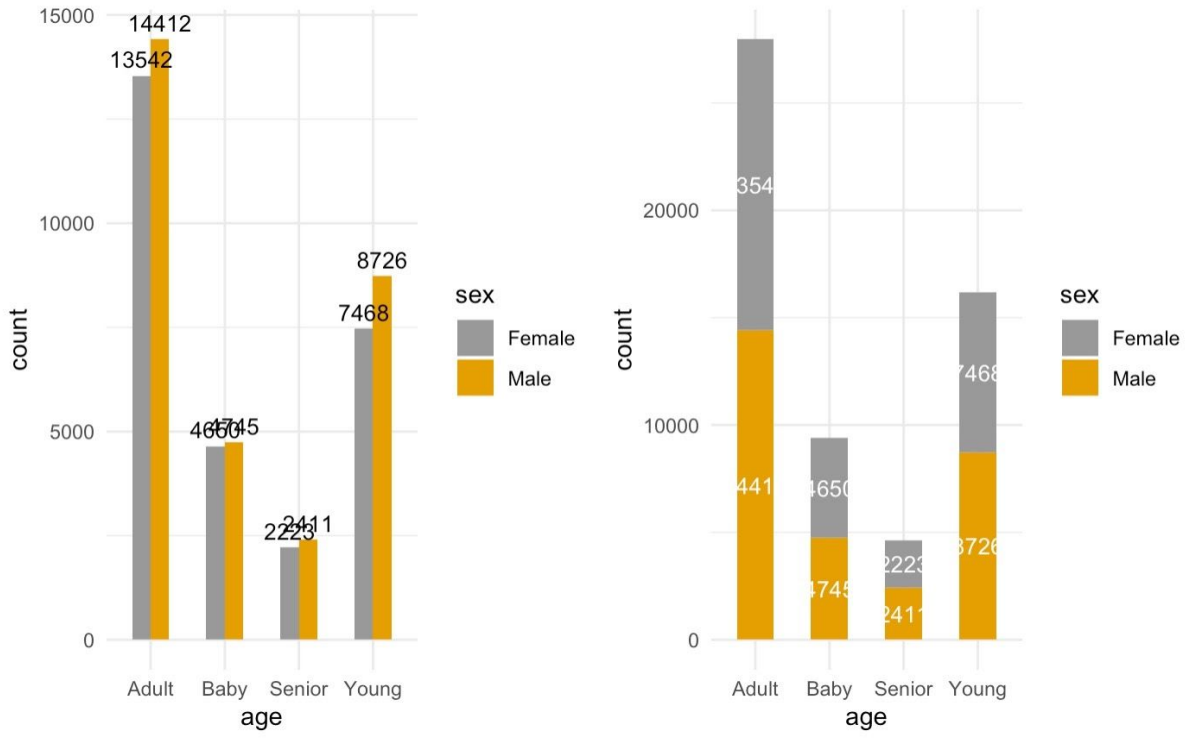

Figure 3.3: dodged-position and stacked position bar plot of grouping by age and sex

Spread() function allows the table spread sex(the key-value pair) across multiple columns. In this case, numbers in the table represent the number of adopted dogs according to the row value and column value.

age <chr>	Female <int>	Male <int>
Adult	13542	14412
Baby	4650	4745
Senior	2223	2411
Young	7468	8726

4 rows

Figure 3.4: table of spreading sex

As far as all the states in the US are concerned, a choropleth map performs better-visualized sights when it comes to the number of adopted dogs in the states.

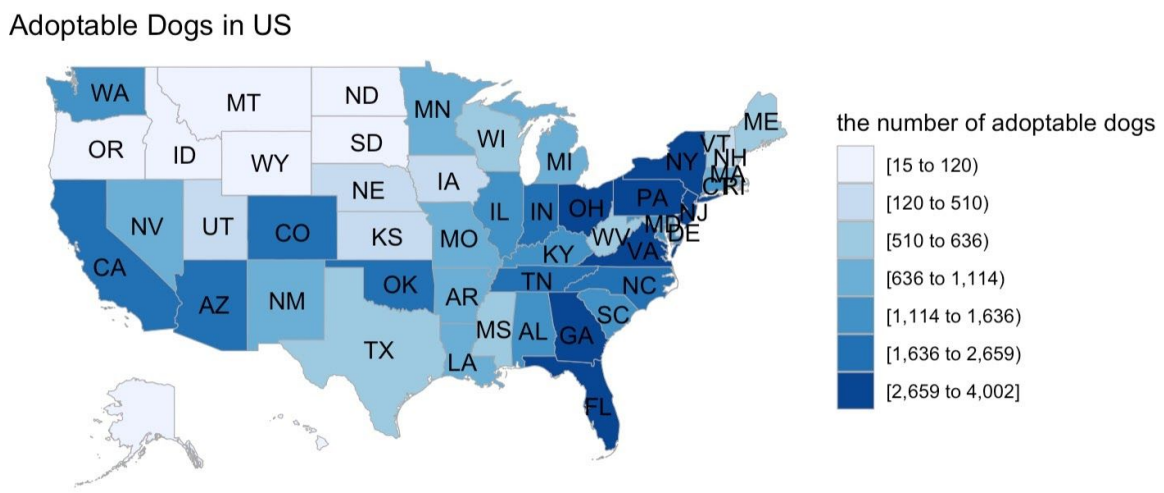

Figure 3.5: choropleth map of adoptable dogs in the US

3.2 Data Wrangling in dogTravel

Dataset dogTravel lists the information about the location and a detailed description about the adopted dogs. To extract the text from the column 'description', we use unnest_tokens() function. There are 2106372 words in the text from 'description'. After removing the stop_words, word cloud about the text shows us the top words in the description.

Figure 3.6: word cloud of dogs' description

3.3 Data Wrangling in movesByLocation

Dataset movesByLocation lists the information about the number of exported, imported adopted dogs in a specific US state or country. After reading the data, we assign NA equal to 0 in this dataset.

State choropleth map of the exported adopted dogs and imported adopted dogs are shown as below. Among the states, Tennessee is the largest exporter and New Mexico is the largest importer. So maybe people in New Mexico, Virginia, New Hampshire are more likely to adopt a dog in their family.

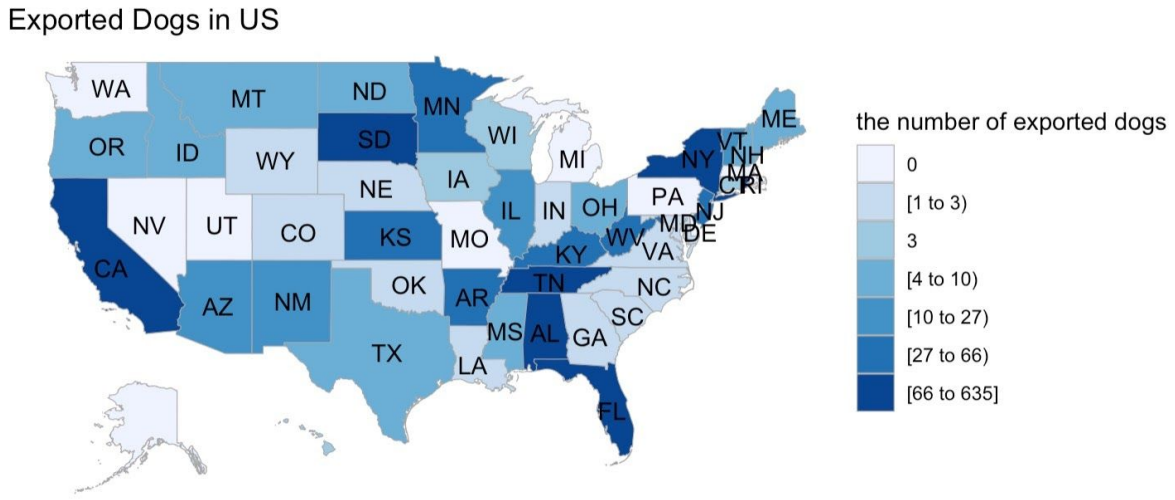

Figure 3.7: choropleth map of exported dogs in the US

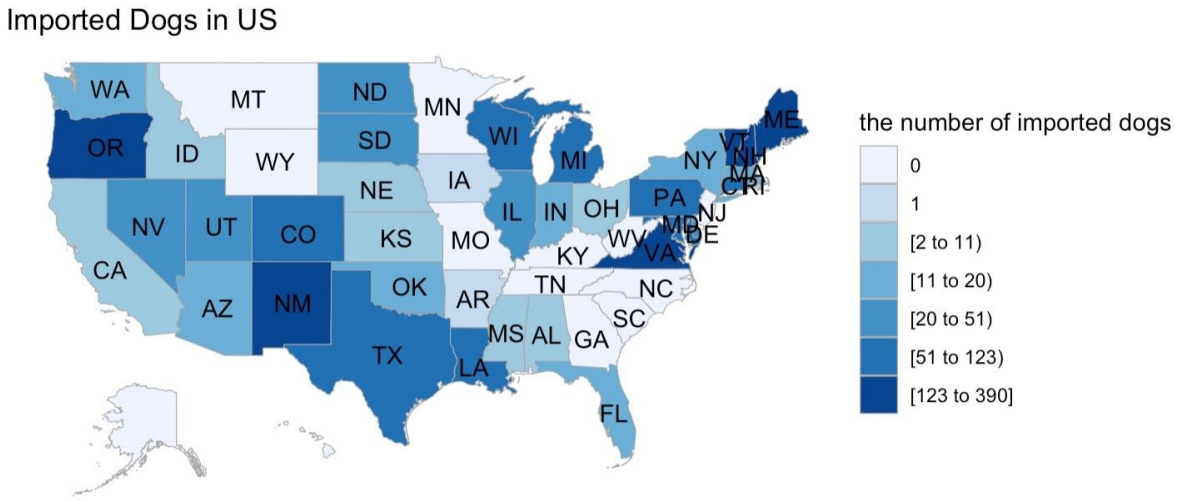

Figure 3.8: choropleth map of imported dogs in the US

3.4 Join Dataset

Because each dog has a unique ID number, we inner join the dataset allDogDescriptions and dogTravel by the key 'id'. After inner join, there are 6194 rows of the

new dataset and we add 'dogTravel$found' in the new dataset, which represents where the dog was found.

There are some tags simplified to describe the adopted dogs. All of the words used to describe dogs are positive. Here is the word frequency barplot and word cloud of the tags.

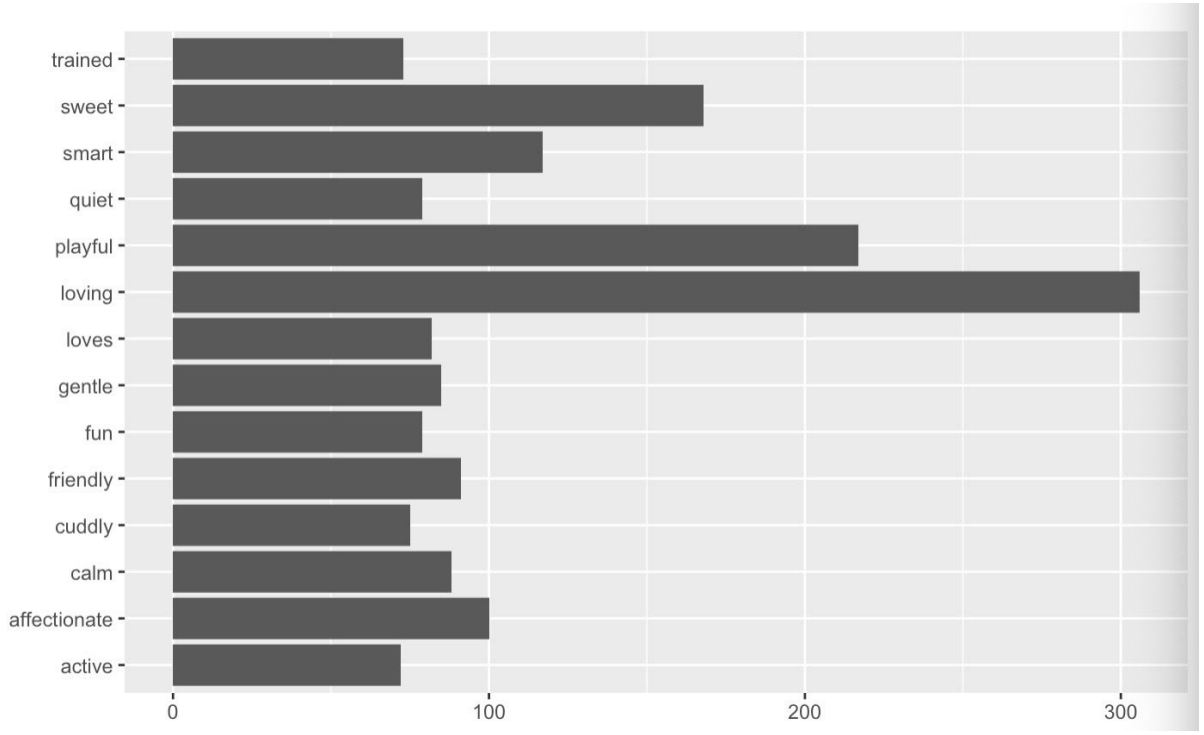

Figure 3.9: top words in tags

Figure 3.10: word cloud of tags

In the new dataset, the column 'posted' are in the same format as that in the dataset allDogDescriptions. If we want to count how many adopted dogs are posted on the PetFinder API every year, separate the 'posted' into 'year' and 'others', then group by 'year' and we may get the result. Here is a line plot to directly show the result.

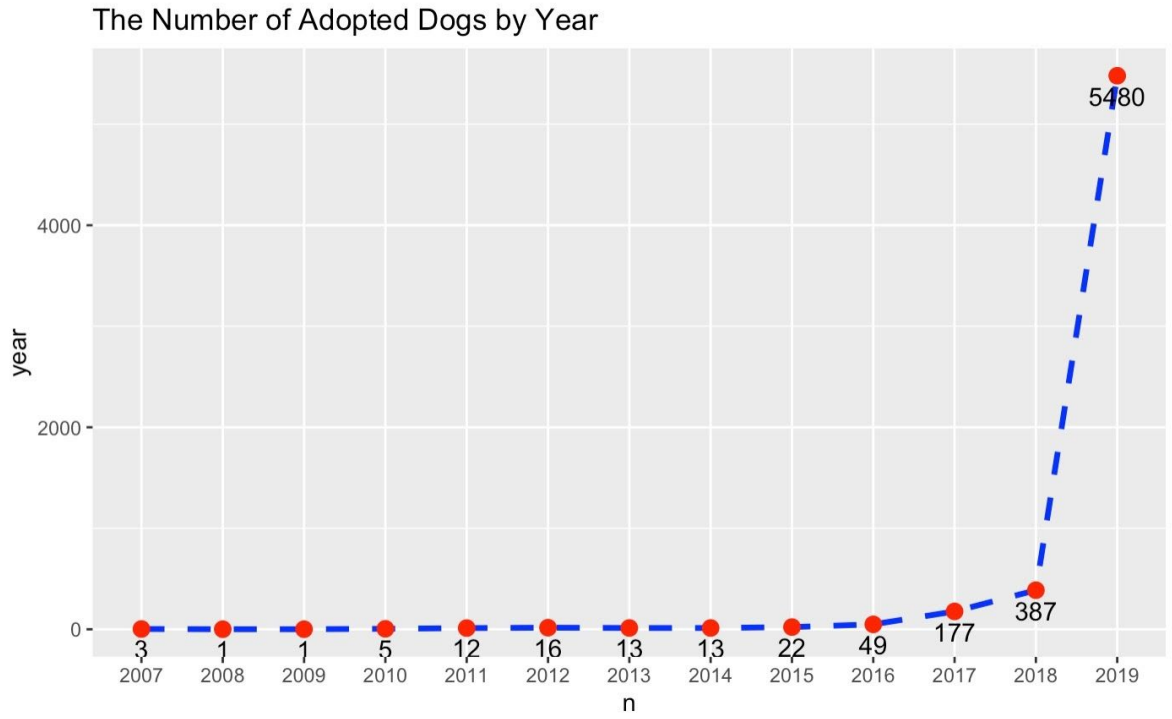

Figure 3.11: the number of dogs posted by year

4. Comment

When I was doing the data wrangling, I realized that my datasets are not very well selected. First, dataset allDogDescriptions and dogTravel have a lot of the same information. It caused the dataset after joining is of no practical use. Second, there is too little information on the aspect of statistics. It would be nice if a small amount of data analysis could be added.