Dog Breeds Analysis

Project for Milestone 1, SIADS 591 & 592

May 21, 2021

Project team: Brian Minie (bminie@umich.edu), Shiran(Rio) Zhang (shiranz@umich.edu)
GitHub: https://github.com/bminie/SIADS_591_592

Motivation

Keeping pets has become a popular trend all over the world, and dogs are the number one choice for people to accompany them. Although many people call for adoption instead of buying, according to AKC's data, purebred dogs are still the primary consideration when people choose a companion dog. The AKC is a non-profit organization dedicated to the cause of purebred dogs. We will base their database on the popularity of these purebred dogs and other characteristics of the dogs, such as obedience, height and weight. To analyze what factors affect the popularity of dogs, and whether there is a relationship between these characteristics.
Our project will analyze the following issues:

1. Is there a relationship between the lifetime cost of ownership and popularity?
2. Is there a relationship between lifetime cost of ownership and lifespan?
3. What are the top 10 most popular dog breeds and what breeds have the largest changes in popularity?
4. Is the obedience of dogs related to their categories?
5. Which size of dog is more popular with the public?

Data Sources

1) Best in Show (data about dogs)

This data set is a csv file compiled by other users based on the
American Kennel Club (AKC) data, which contains a large amount of characteristic data about different breeds of dogs. The information in the file shows that it contains 68 columns (features). However, after checking, we found that half of the columns are null, and among the columns that are not null, a considerable part of the data is no data. Therefore, we will supplement the corresponding feature data (from other data sets) on the basis of this data set. The data is formatted as a CSV file that is 58kb. It can be found at https://www.kaggle.com/paultimothymooney/best-in-show-data-about-dogs and includes important variables such as dog breed, category, lifetime cost, weight (lbs), shoulder height (in), and 2 longitive.

2) Most Popular Dog Breeds – Full Ranking List

The AKC compiled the most popular dog breeds from 2014 to 2018 into a table. We carried out web scraping on the website, and extracted all the data in the table into a csv file that is 6.88kb. It can be found at https://www.akc.org/expert-advice/news/most-popular-dog-breeds-full-ranking-list/ and includes important variables Breed, 2018 Rank, 2017 Rank, 2016 Rank, 2015 Rank, and 2014 Rank.

3) AKC Breed Information

This data set is a data set compiled by other users based on the height and weight of different breeds of purebred dogs provided by the official website of AKC. Since a large amount of data about height and weight in our primary dataset is lost, we found this data set as a supplement of our primary data set. The data is formatted as a CSV file that is 4.51kb. It can be found at https://data.world/len/dog-size-intelligence-linked/workspace/file?filename=AKC+Breed+Info.csv and includes important variables breed, height_low_inches, height_high_inches, weight_low_lbs, and weight_high_lbs.

4) Dog Intelligence

This data set is a data set compiled by other users based on the intelligence of different breeds of purebred dogs provided by Wikipedia. Due to the large loss of data on intelligence in our primary dataset, we found this data set to be useful to our primary data set. The data is formatted as a CSV file that is 8kb. It can be found at https://data.world/len/dog-size-intelligence-linked/workspace/file?filename=dog_intelligence.csv and includes important variables breed and obey

5) The Most Popular Dog Breeds in Every State

This page summarizes the top three popular dog breeds in every state in the United States based on data provided by the AKC (except District of Columbia and Puerto Rico). The content we need is not presented in the form of tables on this website. Considering that if we do web scraping on this website, we need to perform a lot of complicated processing on the queries. Therefore, in order to improve efficiency and the amount of data to be extracted is not large, we finally decided to manually extract the required data into a csv file. The data is formatted as a HTML and then CSV file that is 2.92kb. It can be found at https://www.rd.com/list/most-popular-dog-breeds-in-every-state/ and includes important variables State, Top 1, Top 2, and Top 3.

6) States by Index in Altair

When we use geoshape in Altair to draw a map of the United States, we found that each state in Altair has a corresponding index. Therefore, we found a data set with the correct index based on the example given on the altair geoshape official website. The data is formatted as a CSV file that is 1.91kb. It can be found at https://github.com/vega/vega/blob/master/docs/data/population_engineers_hurricanes.csv and contains important variables state and id.

7) States Latitude and Longitude

This data set mainly contains the latitude and longitude of each state in the United States, so that we can correctly insert the abbreviations of each state into the drawn map of the United States. The data is formatted as a CSV file that is 1.84kb. It can be found at https://www.kaggle.com/washimahmed/usa-latlong-for-state-abbreviations and contains important variables State, Latitude, Longitude, and City.

Data Manipulation

Data Loading

Web Scraping and File Reading

One of our data sources is a website and we used web scraping to load the data into a Pandas dataframe. Our second data set comes from the official website of the American Kennel Club, but all we need is a table on the website, so we need to extract this table into a csv file by ourselves. We first save this website locally (in html format), then upload it to Jupyter Notebook, and read the html in the notebook. In order to extract website data, we need a package dedicated to web scraping, from bs4 import BeautifulSoup. Then, back to the website page and right-click on inspect, we can see that there is a class that contains all the contents of the table. After extracting what we need, we used StringIO to create an in-memory file.

All our other files are in CSV file format and we used Pandas read_csv() to load the data files to Pandas dataframes. When reading our Best in Show data set, we excluded the second row when loading the data as this row contained comments that were not used during our analysis. We also renamed the Dog Breed column to Breed as we will use the Breed column when merging dataframes.

Data Cleaning and Merging

Converting String to Numeric Data

Our breed info data set contains numeric data such as height and weight but it is in String format. In order to use it for downstream analysis, we converted all columns with the exception of Breed to numeric data types using pd.to_numeric(). We also removed rows that were missing data as this data set will be used in additional manipulations to fill in missing height and weight information.

Challenge: Cleaning Up and Standardizing Breed Names

During the initial exploration of the different data sets, we noticed that there were slight differences in the breed names across the data sets that were causing complications with our initial merge. A couple of examples of this included some breeds being in the plural vs. singular (ex. Labrador Retriever vs. Labrador Retrievers) and slight differences in name (ex. German Shepherd vs. German Shepherd

Dog). To handle these differences in breed names, we first found the unique breed names in each file. We then paired all the similar breed names together and created a dictionary where the key is the original breed name and the value is what to update the breed name to. We used this dictionary to update the Breed column of all the different data sets to standardize the breed names across all data sets. The four data sets that this strategy was applied to are Best in Show, Most Popular Dog Breeds - Full Ranking List, AKC Breed Information, and Dog Intelligence.

Calculate Year-to-Year Popularity Change

Our Most Popular Dog Breeds - Full Ranking List data set contains popularity ranking data for 2014-2018. One of the questions we are interested in is which dog breeds have the largest change in popularity. To enable this analysis, we calculated the year-to-year popularity change for each year from 2014-2018 by subtracting the popularity rank of the previous year from the popularity rank for the current year. We also calculated the max change in popularity for each breed which is defined as the max absolute value of the popularity changes.

Filling In Missing Weight and Height Data

Our Best in Show data set contains various columns we will be using for analysis such as lifetime cost and height/weight data. During initial exploration of that data set, we noticed that there were a lot of rows that were missing height and weight information (the data in this column is mean breed standards for height and weight). To fill in this information, we planned to use the data from our AKC Breed Information data set. First looking at the AKC Breed Information data set we see there are min and max values for height and weight but no mean data for those measurements, which is what we want to use to fill in the Best in Show missing values. We calculated the mean height and weight values for the AKC Breed Information data set using the existing min and max values already present in that data set. Once we had that information, we replaced the missing data in Best in Show for each breed with the information from AKC Breed Information. To identify which rows in Best in Show that were missing data, we found all rows that had either NaN, no data, or NA (3 classes) as values in the given column (weight (lbs) and shoulder height (in)) and then replaced it with the information for the corresponding breed in the AKC Breed Information data set (Mean weight and Mean height). Before we did our replacement we found 85 breeds missing weight and 13 breeds missing height. After our replacement, we only had 17 breeds missing weight and 1 missing height.

Data Merging

Earlier in our processing, I mentioned that all our data frames had the Breed column and we had standardized breed naming across all our data sets. When it comes to merge the data sets, we are able to merge our three cleaned data sets (Best in Show, Most Popular Dog Breeds - Full Ranking List, and Dog Intelligence) using Breed as the key merge column. During the merge, we also decided to keep all the rows resulting from the merge, meaning we did not drop rows that had missing information in any column. We decided to do this because we did not want to lose any breeds before we ran our analysis if as low as a single value was missing.

In addition, in order to draw a map of the United States, we used the States by Index data set to insert the specific index of each state in Altair into The Most Popular Dog Breeds in Every State, and combine it with the data set States Latitude and Longitude Merge according to state. In this way, we get a data frame that can draw the map of the United States correctly.

Analysis and Visualization

Lifetime Cost vs. Popularity and Lifespan

One of the things we wanted to explore was the relationship between lifetime cost and popularity. Are the dogs that are the most popular the most expensive? Is there a correlation between popularity and lifetime cost of ownership? There does not look to be any trend in the data (Figure 1). The most popular breed, Labrador Retriever, is in the middle of the pack for lifetime cost of ownership at $21,299. Interestingly, a chihuahua has a very high cost of ownership of $26,250. Maybe there is some relationship between lifespan and cost of ownership. Let's take a look at that.

When visualizing lifetime cost of ownership in U.S. dollars against average lifespan in years we see that the dogs with longer life spans in general have a higher cost of ownership (Figure 1). We also see some interesting groupings when we color the data by size category as breeds of the same size category tend to be close together almost in bands. In order to understand this relationship better, we ran a linear regression which reported a R-squared value of 0.59 indicating a positive relationship between lifetime cost of ownership and average lifespan.

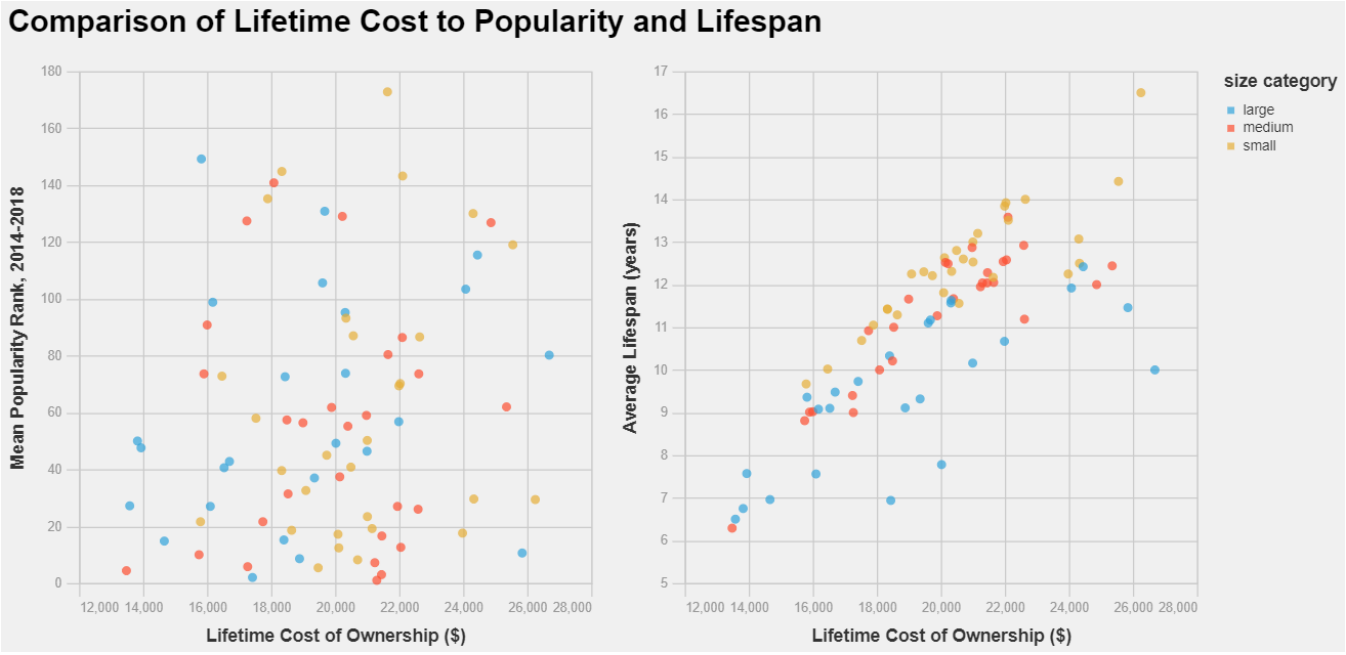


Figure 1: Comparison of lifetime cost of ownership in U.S dollars against the mean popularity rank for 2014-2018 and Average Lifespan (years)

Trends in Popularity Rankings

Now let's take a look at the most popular dog breeds. We are interested in knowing which dogs had the largest change in popularity from year to year as well as which dogs maintain a high popularity. Using popularity data from 2014-2018, we see that the top 3 most popular dogs have stayed consistent within the past 3 years: Labrador Retriever, German Shepherd Dog, and Golden Retriever. There have been some changes however as we can see the French Bulldog is increasing in popularity while both the Yorkshire Terrier and Boxer are decreasing in popularity (Figure 2). Let's take a look at the breeds with the biggest year to year changes.

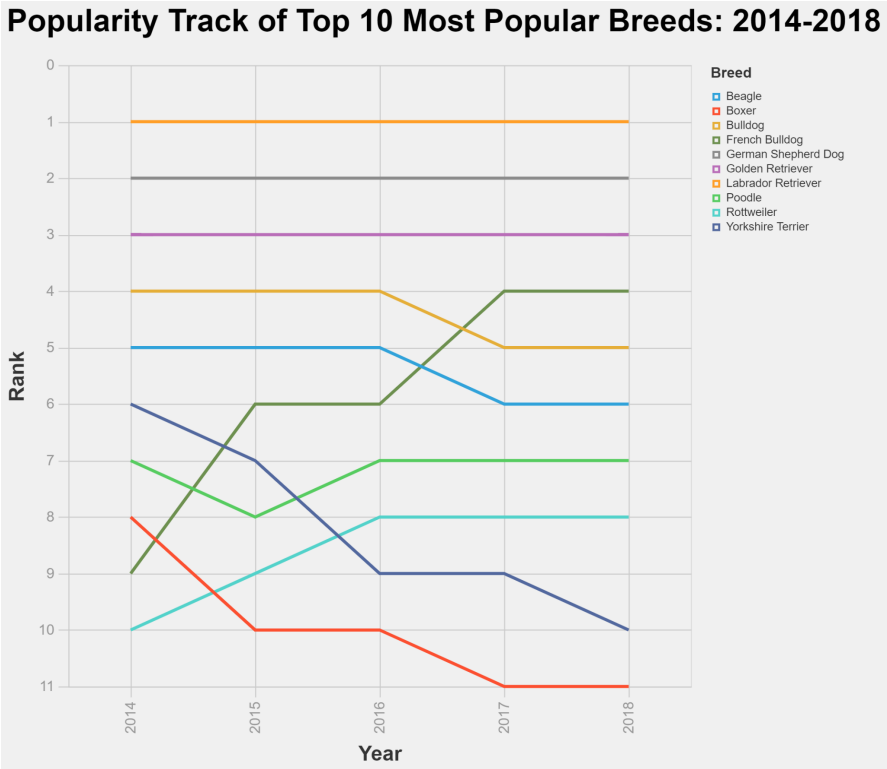


Figure 2: Top 10 most popular dogs from 2014-2018. Each line is colored by dog breed with the most popular breeds at the top.

When determining which breed had the largest popularity change from year to year, we first calculated the year-to-year change in popularity from 2014-2018 and then took the breeds with the 10 highest changes in popularity for any given year. There are several breeds that have similar max popularity change which results in around 29 breeds with similar elevated changes in popularity. We can see that there is a maximum increase in popularity of 20 spots for the Afghan Hound in 2017 and a maximum decrease in popularity of 22 spots for the Otterhound (Figure 3). Even with these big swings in popularity, we see that none of these breeds move into the top 40 most popular breeds.

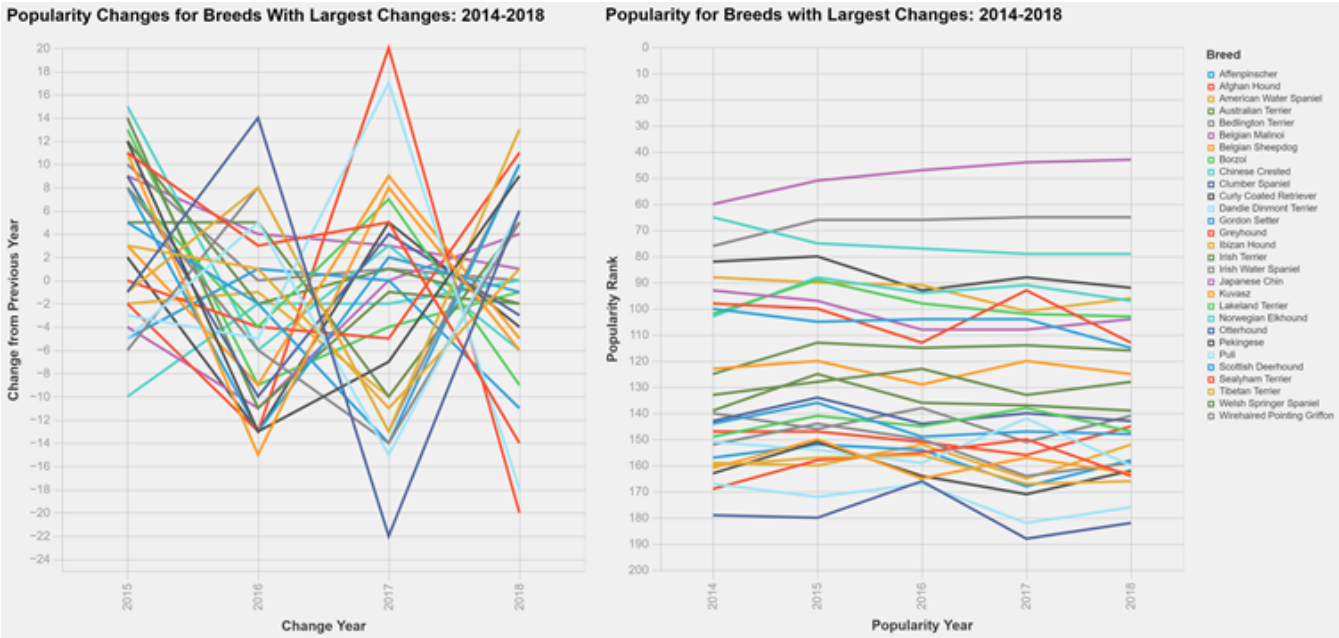
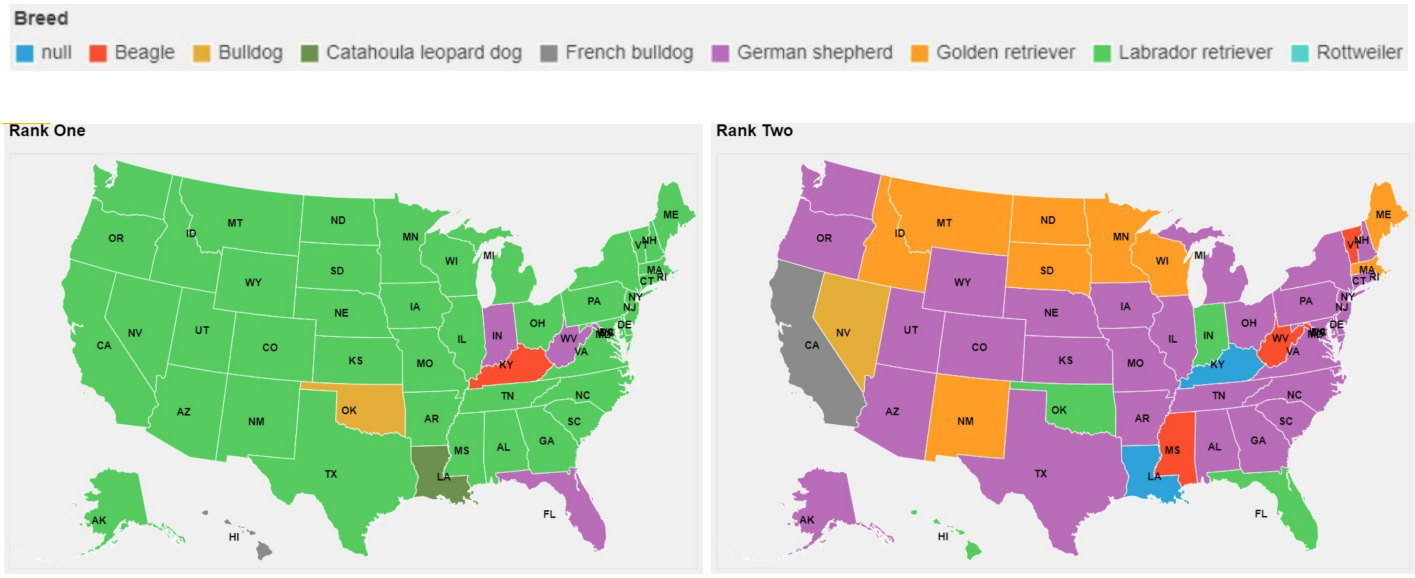


Figure 3: Analysis of the dog breeds with the largest change in popularity from year to year and their popularity ranks. The chart on the left shows the breeds with the largest changes and their change from year to year. The chart on the right shows the ranking of those sample breeds from 2014-2018. Lines are colored by breed.

The Most Popular Dog Breeds in Every State

When we knew that the Labrador Retriever, German Shepherd and Golden Retriever continued to be the top, second and third place in America's most popular dog breeds from 2014 to 2018, we became interested in the most popular dog breeds in every state. It is believed that some states have different results from the national data.



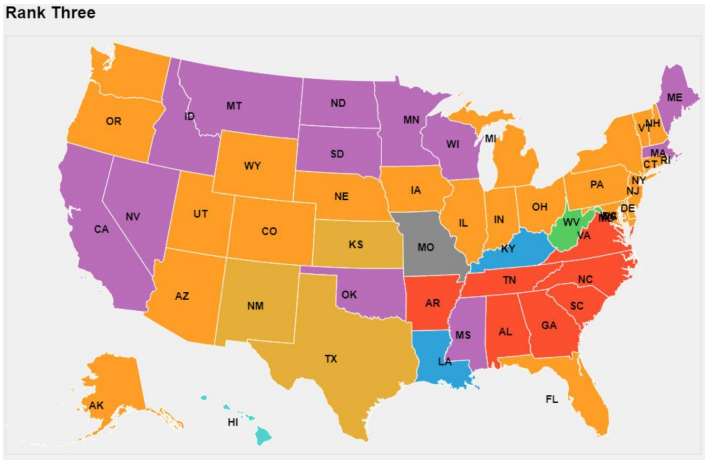


Figure 4: Used to analyze the most popular dog breeds in every state, the first graph represents the distribution of rank one dog breeds, the second graph represents the distribution of rank two dog breeds, and the third graph represents the distribution of rank three dog breeds. Dog breeds are distinguished by different colors.

Through the image, it is not difficult to see that the Labrador Retriever can be said to occupy the first place among the most popular dog breeds in almost all states without any suspense. Only a few states have other results. For the second most popular dog breed, although most states still show that the German Shepherd is the second most popular dog breed, it is clear that the second most popular dog breed is more diverse. Among them, there are not a few states that rank the Golden Retriever as the second place. Let's take a look at the third place. Although there are more breeds, the popularity of Golden Retriever and German Shepherd is still far ahead of other breeds. In addition, Beagle and Bulldog also occupy a place in many states.

The Relationship between Dog Categories and Obedience

People divide dogs into different categories according to their characteristics, such as herding, hound and sporting. Does the category of dog also reflect their obedience? For example, are dogs in the category of shepherd more obedient to instructions?

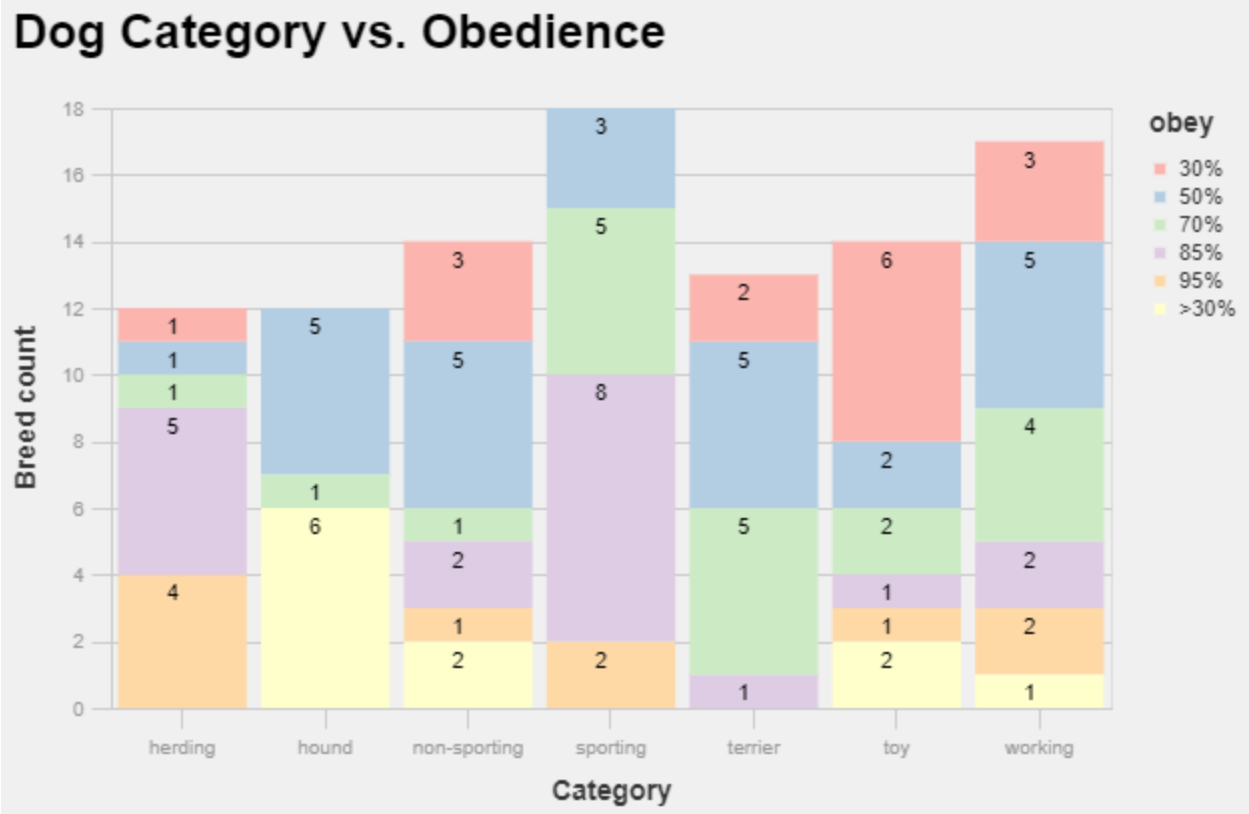


Figure 5: This chart is used to compare the analysis of dog breed obedience in different categories. Obedienceis divided into five different groups and distinguished by different colors.

We divide compliance into six levels, the highest level is 95%, and the lowest level is <30%. Obviously, the obedience of herding and sporting dogs is basically greater than or equal to 85%, while most of the obedience of hound, non-sporting and toy is less than 70%. Even half of the dogs in the hound category have an obedience of less than 30%. To our surprise, we originally thought that most of the dogs in the working category should also be highly obedient, just like herding and sporting, but the data shows that most of them are less than 85%. We need to continue to investigate the reasons for this conclusion.

The Most Popular Dog Height and Weight Range

We divided the top 100 most popular dog breeds into five groups, and wanted to see if the size of these dogs would stabilize in a range in the same group.

Figure 6: This chart is used to analyze whether the top 100 popular dog breeds are related to their body size, and at the same time, by observing the distribution of scatter to analyze which range of dog body sizes are more popular.

Unfortunately, in the same group, we found that the size of dogs can vary greatly. For example, in the top 1-20 group, both smaller dog breeds and larger dog breeds are included. This situation also appeared in other groups. Therefore, we cannot say that the dog size of each group is stable in a range. However, by analyzing the top 100 popular dog breeds, we can see that most dogs are 7 to 26 inches tall and weigh 5 to 90 lbs.

Statement of Work

Brian Minie came up with the original project idea and Shiran (Rio) Zhang expanded upon it to include location popularity analysis. Rio was responsible for web scraping used to gather the data sets and Brian was responsible for cleaning and merging the data. Brian was responsible for analyzing lifetime cost vs. popularity and lifespan as well as trends in popularity rankings. Rio was responsible for analyzing the most popular dog breeds by state as well as the relationship between dog categories and obedience. Brian created the GitHub repo used to manage project files. Both Brian and Rio reviewed each other's analysis and conclusions and each contributed equally to writing the final report.