Adoptable Animals

Milestone Report

8th June 2018

Christopher Seth Hill

OVERVIEW

Many animals wind up in shelters for various reasons. Some get adopted, some don't, and some even die in shelters. By looking at intake and outtake data and statistics, I can attempt to predict which dogs (based on breed, sex, health, etc.) tend to get adopted. With this information and prediction capability, the dogs that need a little more attention to get adopted can receive the extra time and attention needed to get them adoption ready.

All this will lead to better awareness and hopefully help find homes for the dogs that need it the most.

INTERESTED PARTIES

Shelters and dogs anywhere could benefit from this information. It could aid in optimizing the amount of time spent on animals to get them adoption ready. This in turn could help decrease the turnaround time on intakes and find homes for less fortunate animals. This would lead to better efficiency and allotment of resources on the shelter's part and could stand for potential savings.

DATA

The data for the adoption experiment will come from the Austin Animal Center Shelter API, which makes their intake and outtake information for their animals freely available and is consistently updated.

The adoption data comes freely available from the Austin Animal Center (AAC) shelter via the Socrata Open Data API (SODA). It is divided into two sets based upon the intake of the animals and when the animals left the shelter. This is a no kill shelter. Thus, the overarching outcome types are either adoption, transfer to a rescue, return to owner, or death of natural causes.

The Socrata API imports the data into python as a list of dictionaries with each dictionary representing a row in the data frame. I converted the list of dictionaries into a pandas data frame using the DataFrame method.

For the sake of time, I focused only on the dog adoptions.

Starting with the intake data, I filtered for dogs only and saved both the intake and the new dog intake only data frames as CSV files as to have a reset point if I needed to start from scratch. A quick inspection of the dogs only intakes data frame info and of the first few rows show that animals were all given a unique animal_id for unique identification. After looking at the value counts of the unique animal_ids, there were multiple intakes for some animals. The summary of issues found upon initial inspection of all the columns and data frame summary info are itemized below:

1) There are multiple intakes for the same pet.
2) There are two date columns.
3) The color attribute sometimes contains multiple colors.
4) The age upon intake isn't machine digestible.
5) There are nulls, No Name, and values with asterisks in the name column.
6) The found location is too specific in some cases.
7) Sex upon intake is two features in one column, the sex and then the spay/neuter info as well.
8) The breed sometimes contains mix and multiple breeds, which can be made to be its own column.

Moreover, the columns in the intake data frame are age_upon_intake, animal_id, breed, color, datetime, datetime2, found_location, intake_condition, intake_type, name, and sex_upon_intake.

A) Intake Data Frame Wrangling

There are multiple intakes for the same pet.

To account for repeat intakes, I first sorted the data frame by animal_id and the intake date. Then, I looped through and labeled the "intake" column with the corresponding number of the current intake for the dog. This allowed me to simultaneously label the repeat intakes and create unique labels for each pet, which will come in handy when joining the intake and outtake data frames. This step was done after attending to all the other issues. The result of this is shown below in Table 1:

	animal_id	datetime	Intake_condition	Intake_type	name	color1	color2	intake_age	found_loc	intake_sex	intake_fixed	breed1	breed2	intake
0	A006100	2014-03-07	Normal	Public Assist	Scamp	Yellow	White	72.0	Austin	Male	Neutered	Spinone Italiano	Mix	1
1	A006100	2014-12-19	Normal	Public Assist	Scamp	Yellow	White	84.0	Austin	Male	Neutered	Spinone Italiano	Mix	2

2	A006100	2017-12-07	Normal	Stray	Scamp	Yellow	White	120.0	Austin	Male	Neutered	Spinone Italiano	Mix	3
3	A047759	2014-04-02	Normal	Owner Surrender	Oreo	Tricolor	None	120.0	Austin	Male	Neutered	Dachshund	None	1
4	A134067	2013-11-16	Injured	Public Assist	Bandit	Brown	White	192.0	Austin	Male	Neutered	Shetland Sheepdog	None	1

Table 1: First five rows of the resulting fix for the repeat intake issue. Also, the finished wrangled intake data frame.

There are two date columns.

I deleted the second datetime column as it was a repeat of the first, and I also parsed the datetime entries into datetime objects.

Color attribute contains multiple colors.

The colors were delimited by "/". Thus, I separated out the secondary color into a color2 column and notated secondary color nulls as "None".

The intake age is not machine digestible and is inconsistent.

The ages were expressed in years, months, weeks, and days. I chose to format all the intake ages in months. I first split the age up into its number and period designator. Then, I created and matched regular expressions for the different time periods to each intake age in the data frame. This gave me the period information, and when combined with the number part of the intake age, allowed for the conversion of the intake ages into months.

There are nulls, No Name, and values with asterisks in the name column.

I replaced the nulls and "No Name" with "None". I also stripped the asterisks from the values that had them.

The found location is too specific in some cases.

All the locations were in Texas near Austin. I decided to retain only the city information from the intake location data by stripping off the extra information from the data frame location entries.

Sex upon intake is two features in one column, the sex and then the spay/neuter info.

I split the information in the column into two separate columns, the intake sex and the intake fixed information columns. Null values were labeled as "Unknown" in both columns.

The breed sometimes contains mix and multiple breeds, which can be made to be its own column.

The breed column information was split into the breed1 and breed2 if two breeds were listed. For the cases with mix in the breed info, the breed2 column value became "Mix". Lastly, if there was no secondary breed info in the breed data, breed2 became "None". Moreover, the total number of unique breeds were similar to the number of breeds recognized by the American Kennel Club (AKC).

This is all the wrangling done on the intake data frame at this point and the resulting first 5 rows is shown in Table 1.

B) Outtake Data Frame Wrangling

The outtake data had similar issues to the intake data. Also, a quick inspection of the data frame info shows that there are less outtakes than intakes. Moreover, the outtakes were also filtered for dogs only. There were also repeat outtakes, but they didn't exactly match all the repeat intakes. This is okay as not all animals have left the shelter yet. As we are trying to predict adoption outcomes based on intake data, the only relevant information from the outtake data is the outcome type and subtype, age upon outtake, and outtake date. It may however be interesting to do EDA on the outtake data as well and compare it to the intake data. So, the rest of the columns, minus the repeat information of the intake frame, will remain and be wrangled as well.

The data will be prepped for machine learning later in the machine learning step. For the EDA prep, I assumed that color and breed do not change from intake to outtake. Thus, these columns were dropped from the outtakes data frame along with the animal type, date of birth, and duplicate outtake date column.

The issues addressed in wrangling the outtake data are listed below:

1) The age upon intake is not machine digestible.
2) There are nulls, No Name, and values with asterisks in the name column.
3) Sex upon outtake is two features in one column: the sex and then the spay/neuter info as well.
4) The datetime column needs to be parsed.
5) Outcome type has 'nan' as a value and contains nulls.
6) Outcome subtype has 'nan' as a value and contains nulls.
7) There are multiple outtakes per animals.

Issues 1, 2, 3, 4, and 7 were fixed in the same manner as described above in wrangling the same issues in the intake data frame.

For the remaining issues 5 and 6, the "nan" and null values were replaced with "Unknown". The resulting data frame looks like the resulting wrangled intake data frame, but with the following columns instead: animal_id, breed, datetime, name, outcome_subtype, outcome_type, outtake_age, outtake_sex, outtake_fixed, and outtake. The "outtake" column is the unique identifier created to account for the repeat outtakes.

C) Join and Save the Data Frames

At this point, the wrangled intake and outtake data frames were saved as CSV files. The next thing to do was to join the two data frames on the unique combination of animal_id and intake and animal_id and outtake numbers. This means that the data frames would not misalign upon merger. Also, I did a quick check to see if there were animal ids that showed up in the outtake data that were not in the intake data. There were 381 entries that fit this description. 381 out of ~45000 is not a lot. Thus, I let the merge naturally exclude these entries.

After merging, a quick look at the joined data frame info shows that the data frame is the length of the intake data frame. There are null values for the outtake data related columns as well. This is okay as the rows with null values represent the animals still in the shelter. I left them as nulls for easier identification. Lastly, I added an additional column to the joined data frame which represented the time spent in the shelter between intake and outtake.

Upon inspection of this newly created column, I noticed that there were negative shelter time values. Closer inspection revealed that these appeared to be from data logging error where the intake and outtake dates were swapped. The fix was to loop through the joined data frame and swap the intake date and outtake date for the instances with negative shelter time values.

The Data Wrangling was complete enough at this point to do Exploratory Data Analysis. Further wrangling is needed to input the data into Machine Learning Models later. Also, as is discussed in the initial findings section, other columns were created in the joined data frame to aid the Exploratory Data Analysis. Remember, the intake age and outtake age are in months, and the shelter time is in days.

OTHER POTENTIAL DATASETS

There are some other data sources that could be utilized to help the statistical tests be more significant and more general to all dogs or animals everywhere. Similarly, more data might help the machine learning predictions generalize better to new data. Some of the other datasets that could be used would be data from other shelters located all over the world. I could also wrangle data from adoption websites. Moreover, enriching and expanding the feature set would potentially aid in the generalization process. Other features could be wrangled from the other animal information in the current shelter data. For example, the cat intake and outtake

information could very well affect the dog adoptions significantly and be useful features in predictions. More features could be pictures of the dogs and website descriptions wrangled from online adoption postings.

INITIAL EXPLORATORY DATA ANALYSIS FINDINGS

The goal is to find what features aid in the adoption of animals, both in proportion of intakes and time to adoption. Adoption efficiency is defined as the percent of all outtakes that are adoptions. The time to adoption is only considered for the animals that were adopted and not logged as other shelter outcome types.

1) Now, I will look at the shelter outcomes.

Considering the overall outcomes from the shelter can give insight into the overarching trends and stats of shelter operations. This will form the foundation for further analysis and can help identify potential troublesome trends that lie in the shelter's outcome data. Identifying the overarching problem gives a place to start to dive into the details where the interesting answers are located. Moreover, performance goals can be set based from the overarching statistics.

Initial data exploration led to interesting insights and raised some questions. The first action was to look at the value counts of the outcome_subtype. The clear majority of outcome_subtypes were classified as normal. There were only a few exceptions. Outcome_subtype will not be used in EDA. The value counts of the outcome_type category are much more interesting as shown in the graph below:

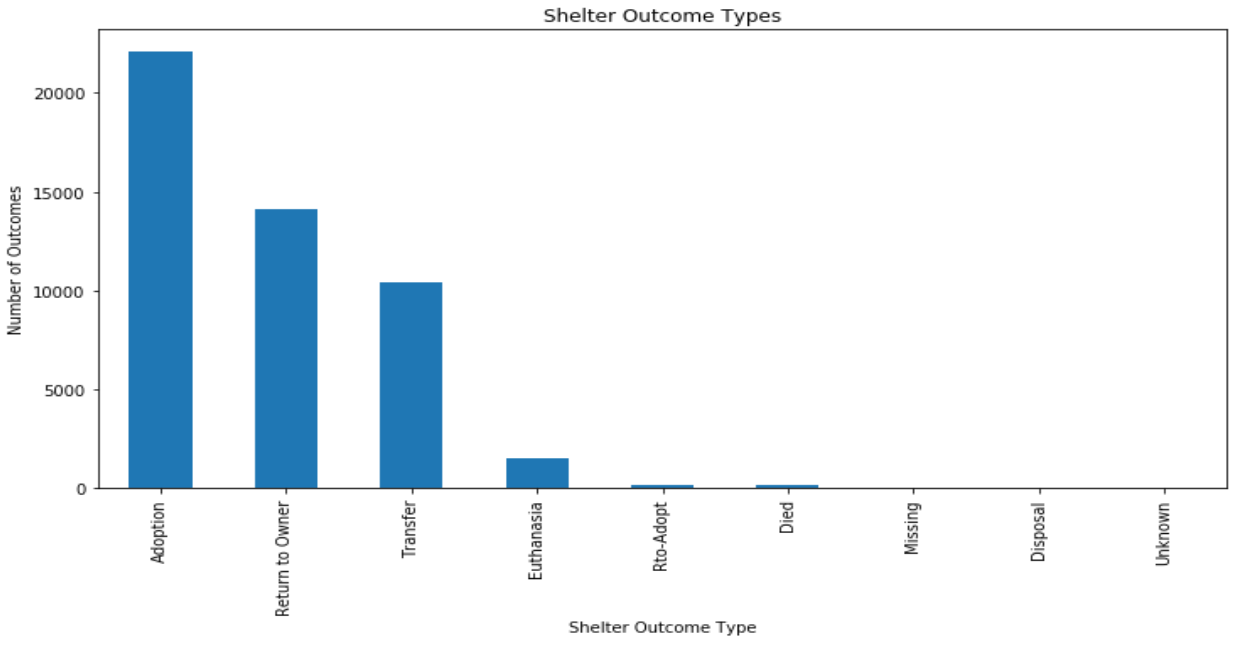

Figure 1: The outcome type value counts for dogs.

Figure 1 shows that the Adoptions, Return to Owner, and Transfers are the most likely outcome types. Thankfully, deaths don't seem to occur that often. Later, a deeper dive into the different adoption outcome type statistics will be done.

The next analysis done was to look at a time series broken up by the outcome subtypes. This is illustrated in Figure 2 below.

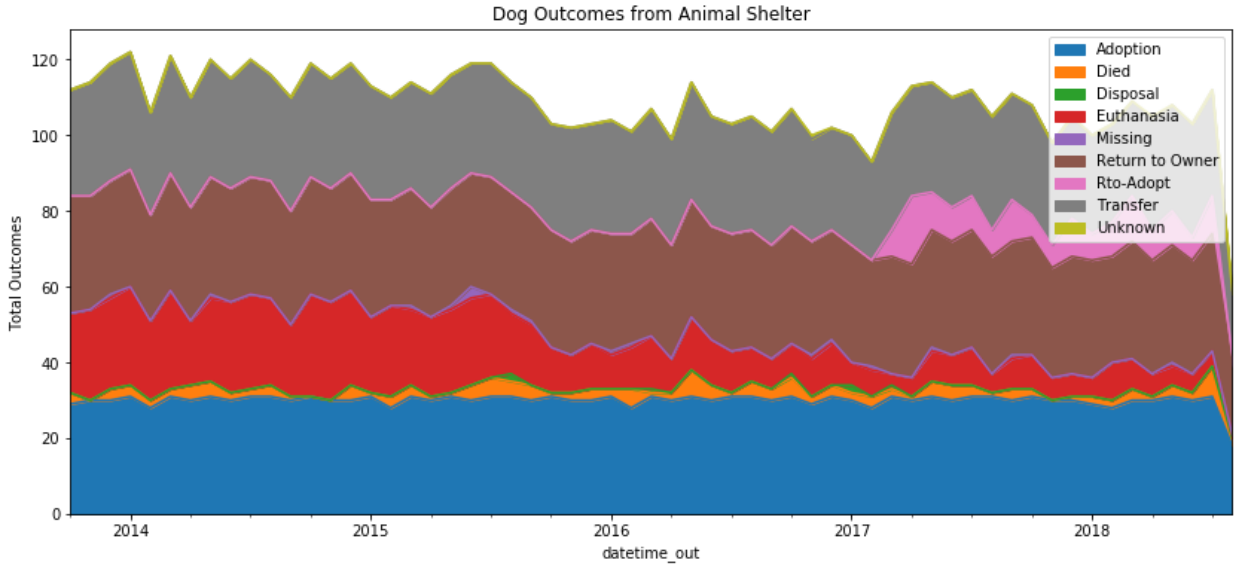

Figure 2: Time series plot of the shelter outcomes broken up by outcome subtype.

There are several insights that come from this plot. Mainly, we can gain the overall outcome trends from the shelter. There seems to be about 30 adoptions a month or one a day, which is quite steady over the years. There also seems to be a dip in adoptions in the spring time around February/March of each year. This systematic dip is a potential place of improvement to increase the shelter's adoptions and efficiency. There are another 30-50 outcomes per month that are return to owner/return to owner-adopt, which leaves about 40-60 dogs per month that were transfers or deaths. Transfers to other shelters (which is better but not counted as adopted) remain steady as well at about 30 per month. Cutting out transfers by increasing direct adoptions from the shelter would reduce strain and resources of the other animal shelters and adoption organizations. This reduces overall cost that goes into the adoption process and gets a dog to a happy home quicker. Died, Disposal, and Missing thankfully remain consistently low. Moreover, euthanasia rates dropped by about half from 30 to 10-15 per month around October-November 2015. Also, this rate seems to gradually decrease over time, which is great. The policy, procedural changes etc. affected at that time were effective in reducing the amount of euthanasia in the shelter.

Moving onto the time to adoption values, I considered only adoptions in the time to adoption statistics. Adoptions are comprised of any outtake with an outcome type of adoption, return to owner, or Rto-Adopt. A quick comparison of the distribution of the data revealed that it is very similar to an exponential distribution, which makes sense since the data we are plotting is the time between events. The slightly skewed part suggests that the time between adoptions aren't completely random and are affected by and correlated with outside influences. Deeper research into what factors influence the adoption rate will aid in finding which factors affect the adoption rate significantly. This will help the shelter become more efficient in adoptions by identifying which dogs need help to be "adoptable". Efforts can then be directed more appropriately. Also, there is a 95% probability that an adoption will occur in approximately 50 days or less.

Now it is time to delve deeper into the data. Let's take a closer look at the intake data on its own.

2) Let's consider the intake data.

Knowing the basic intake data statistics for dogs will allow the shelter to better understand its intake base. This is valuable information because the shelter can better understand the demographic of its population. Allowing the shelter to target their focus in marketing and adoption efforts on the right dogs. This is not the whole picture. However, this information when combined with the outtake data will give a basis to check for trouble areas when it comes to getting these dogs to a good home more efficiently and in less time. Moreover, the baseline data can be used as a type of gauge for normality in the shelter intakes.

I chose to break down intakes by dog features and get basic statistics. The features analyzed are listed below:

a) Breeds
    Primary (Top 10)
    Secondary (Top 10)
b) Age (Distribution)
c) Color
    Primary (Top 10)
    Secondary (Top 10)
d) Gender (Comparison)
e) Fixed (Comparison)

f) Intake Condition (Comparison)

g) Intake Type (Comparison)

We will start with the breeds feature.

The breeds were wrangled into two columns: a primary and a secondary breed. In exploring the secondary breed information, 84% of the data fit into either the "Mix" or "None" categories. This means that very few dogs are labeled with a distinct secondary breed. More breed data would aid in more accurate analysis and predictions. It would also add more real world meaning to the results. Being limited to the data and the labeling accuracy of the database users, it is probably okay to just lump the remaining breeds that are not "None" into the "Mix" label without causing too many issues.

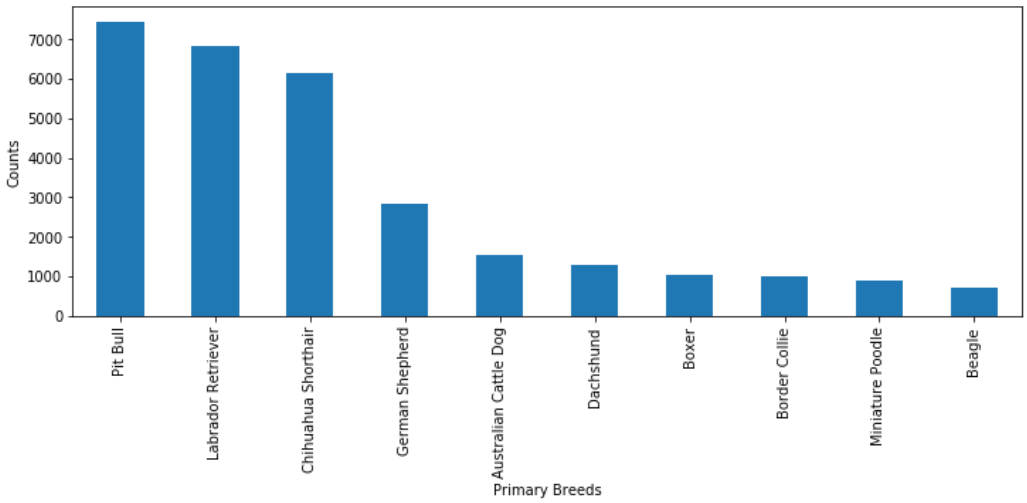

Figure 3: Top 10 primary breeds for intakes.

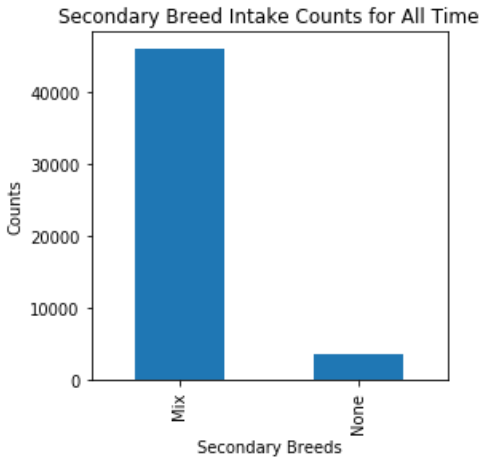

Figure 4: The secondary breed distribution for intakes.

A few observations can be obtained from the data. First, almost all the dogs in the shelter are mixed breed. The most popular primary breed taken in is pit bull, which from my limited experience with dog shelters seems to be the case. Moreover, most of the breed intakes are represented by the top 10 primary breed intakes. This information allows the shelter efforts and marketing to be more focused and less spread out. Similarly, knowing the top breeds in the shelter can help the shelter in determining the target audience for any adoption events that they could potentially have or are already conducting.

Now, I will look at the age upon intake data.

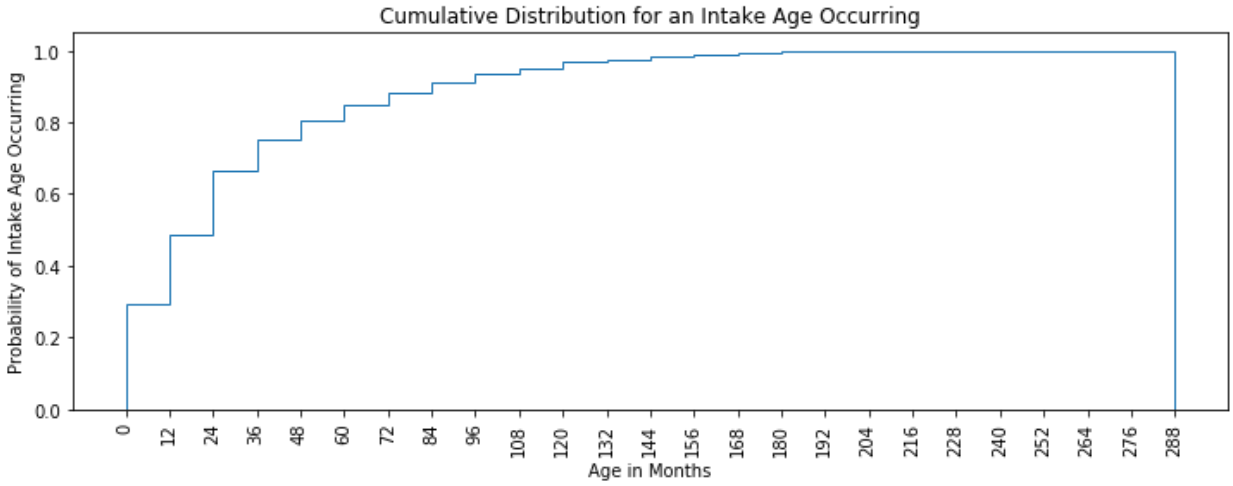

Figure 5: The cumulative distribution function of the age upon intake.

The plot above suggests that 80% of the dogs that come in are 5 years old or less. This means that most of the dogs that the shelter will encounter will be middle aged (for a dog) or younger. However, there are some outliers in the age of the dogs out there just as there are some old humans. One point stands out. The maximum intake age was 24 years old. I decided to keep the data point as the shelter had other intakes around 20 years old which is believable and makes the 24 years seem less odd. This information is useful regarding knowing the age demographic of the shelter's population. Future planning and ordering of supplies can be accurately speculated. This will lead to less waste in supplies such as dog food and medical supplies, which are all age specific items. Moreover, this is another demographic feature that can aid in better marketing for the adoption of the general dog population in the shelter.

Next, I will explore the primary and secondary color of the intakes.

The color was wrangled into primary and secondary coat color. The top 10 of each are below in Figures 6 and 7.

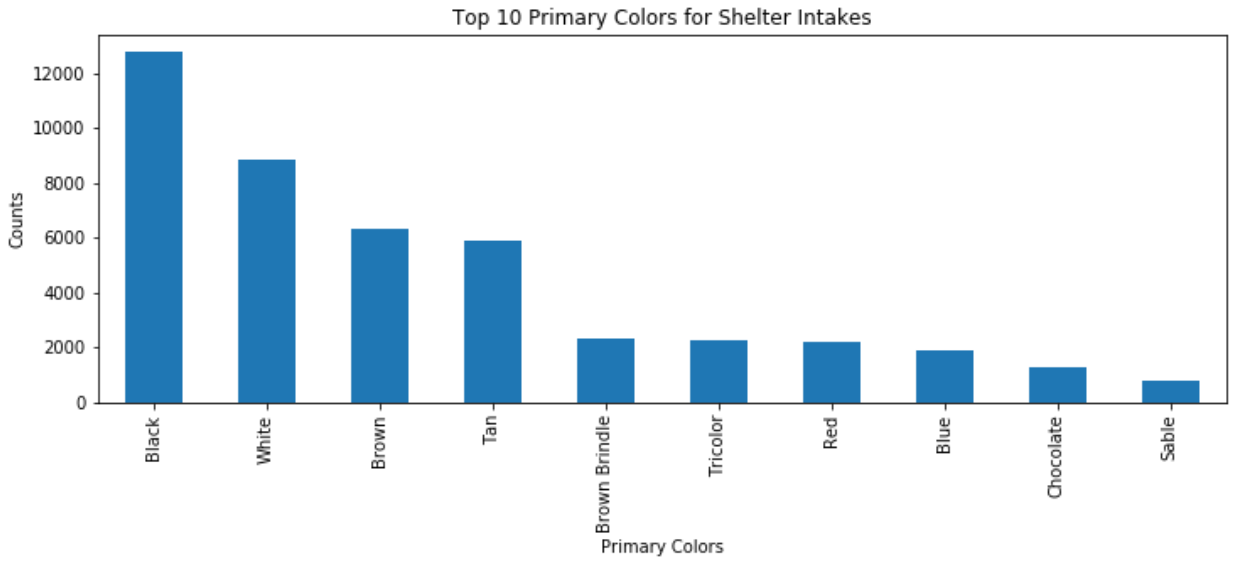

Figure 6: Top 10 Primary Breeds in the intake data.

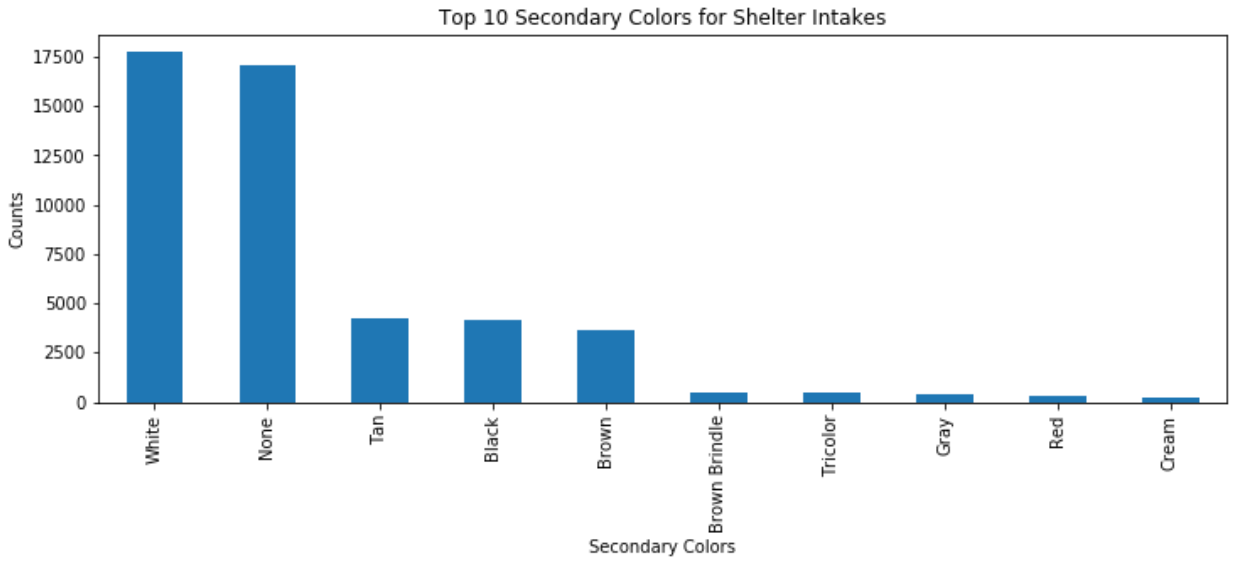

Figure 7: Top 10 secondary breeds in intakes.

It looks like most of the dogs can be identified as having either black, white, or brown primary color. The top secondary colors are white or none, which account for most of the dog intakes. Color is yet another feature that potential pet owners may or may not have a certain preference for. Pets with rarer colors or secondary colors may be more adoptable. Deeper analysis will be able to significantly speak to this topic. The results along with the other features will help in establishing general population trends and stats.

Now we will consider the intakes by gender.

The value counts for gender are shown below in Figure 8.

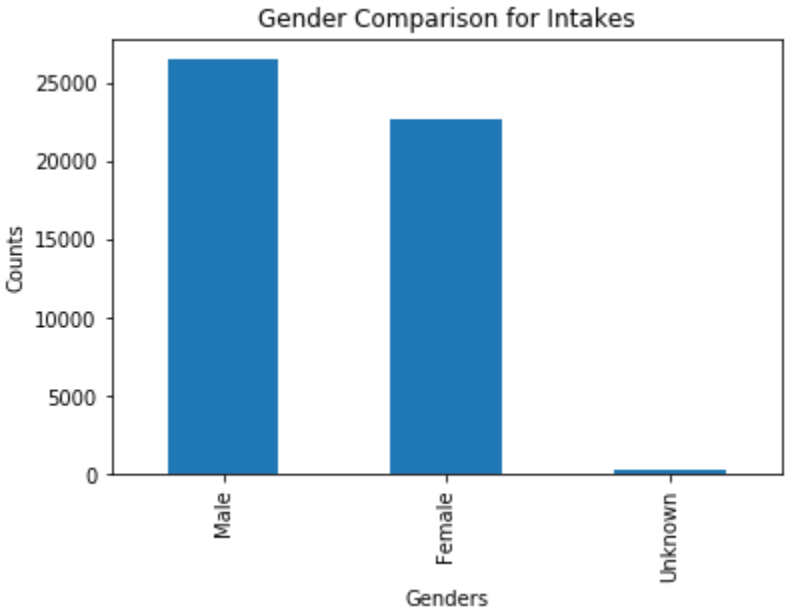

Figure 8: The distribution of Genders for intakes.

There are more intakes that are males than females, but the ratio is roughly 50/50. There are a small number of unknowns, which are probably missing data. I chose to leave the Unknowns as Unknowns for now. The gender is yet another baseline mark that can be used as a barometer for normality in the shelter.

Next, we will consider the fixed status data for intakes.

I considered the fixed intake and fixed outtake distributions. The results are summarized below in Figure 9 and 10.

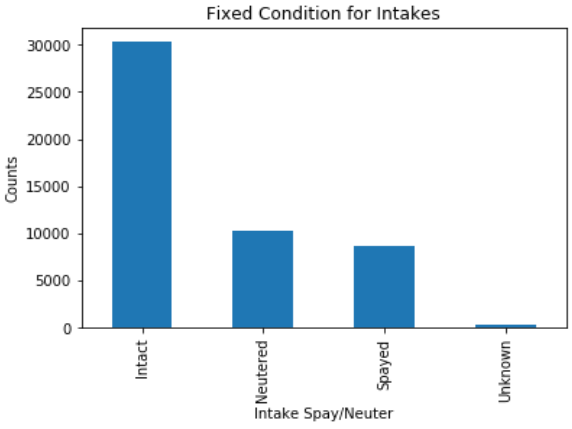

Figure 9: Fixed intake distributions.

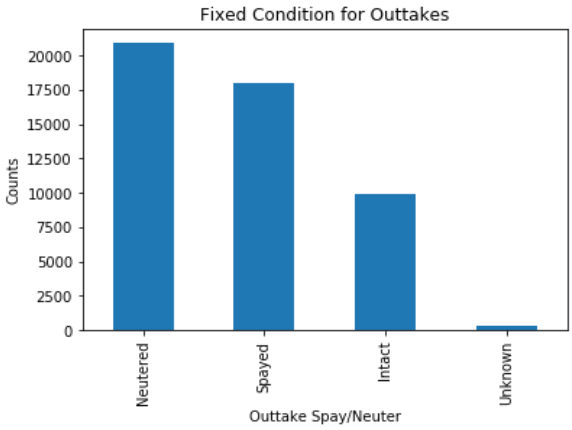

Figure 10: Fixed outtake distributions.

It seems that the majority or roughly 2/3 of the pets in the shelter that come in intact leave spayed or neutered. This is good for population control. Most of the intakes are intact, while most of the outcomes are fixed. There is not much to learn from these plots alone besides the

amount of the intake population that are getting fixed, which could be a performance metric for the shelter to increase their efforts in population control.

Now, we can consider the condition of the animals upon intake.

The value counts for the Intake Conditions are summarized below in Figure 11.

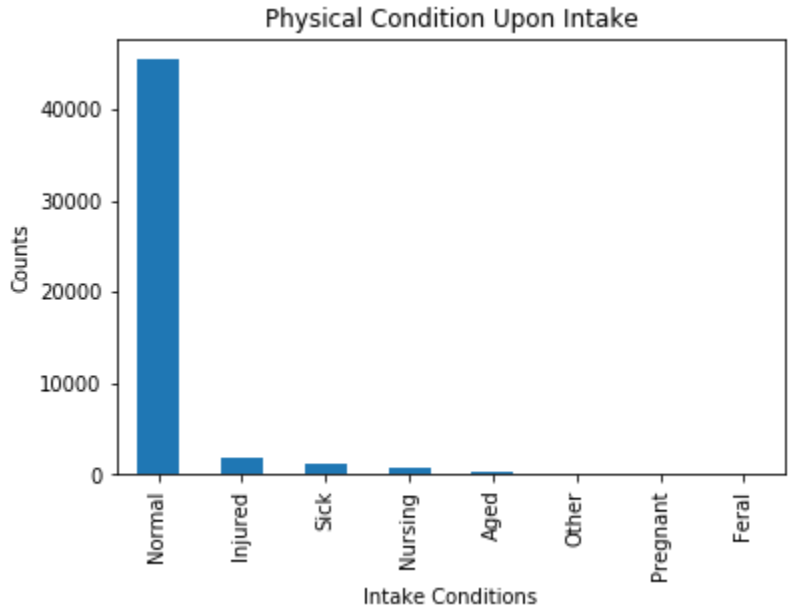

Figure 11: Distribution of the Intake Conditions.

Most of the intakes that come in are in normal condition. This would make any statistical analysis based on intake condition skewed and potentially subject to selection bias. Also, it is nice to see that nearly all the intakes come in virtually unharmed within this sample population.

Finally, we can consider the intake types.

The distribution of intake types is shown below in Figure 12.

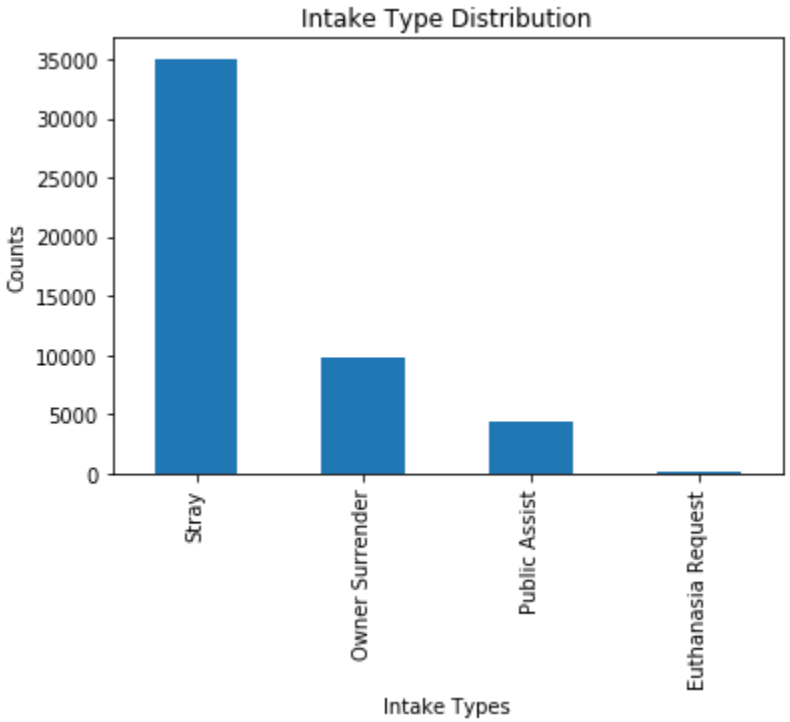

Figure 12: Distribution of Intake Types.

Upon inspection of the intake types I decided to consider the percent of Owner Surrenders that ended up as Return to Owner or Rto-Adopt. Only 4.4% of Owner Surrenders ended up getting returned to the owner. Most owner surrenders are not returned to the owners, suggesting that one should probably not use the animal shelter as a form of free animal boarding to be on the safe side. Moreover, only 9.6% of Euthanasia is requested. This is interesting for a no kill shelter. I'm hoping this is just some clerical error. Also, most of the intakes are strays, which makes sense for a shelter. This is yet another example of a normality barometer metric for the shelter.

3) Quick look at Intake Feature Correlations to Adoptions

Let's stick with the most interesting features that aren't heavily weighted to one feature grouping. Thus, we are dropping intake_type and intake_condition from this step. Also, we will look at the percent adoptions multiplied by the total number of adoptions for each category grouping (weighted adoption count). For example, the percent of Pit bulls adopted would be the number of Pit bulls adopted divided by the total number of Pit bulls in outtakes. This percent adopted will be multiplied by the number of Pit bulls adopted to produce the weighted adoption count for Pit bulls. This will hopefully give an idea of the efficiency and quantity of the adoptions in each category grouping. Additionally, we will also look at median time to adoption for the different category groupings. This information regarding how the intake features may or may not affect the Adoption percentages and time to adoptions will aid in identifying good and troublesome areas for dogs in the adoption process. This could lead to better direction of

efforts (whether that be marketing, increased animal care, etc.) to certain dogs that statistically need more help to be adopted more efficiently. This can save the shelter time and increase the turnover rate for adoptions which is a win-win situation for the shelter, surrounding animal homes, the potential adopters, and the potential adoptees.

Following the same flow as the previous section for intake features/category groupings, we have:

a) Breeds
    Primary (Top 10)
    Secondary (Comparison)
b) Age (Distribution)
c) Color
    Primary (Top 10)
    Secondary (Top 10)
d) Gender (Comparison)
e) Fixed (Comparison)
f) Seasons (Comparison)

Now we will consider the breeds adoption statistics.

Primary and secondary breed information is summarized in Figure 13, 14, 15, and 16.

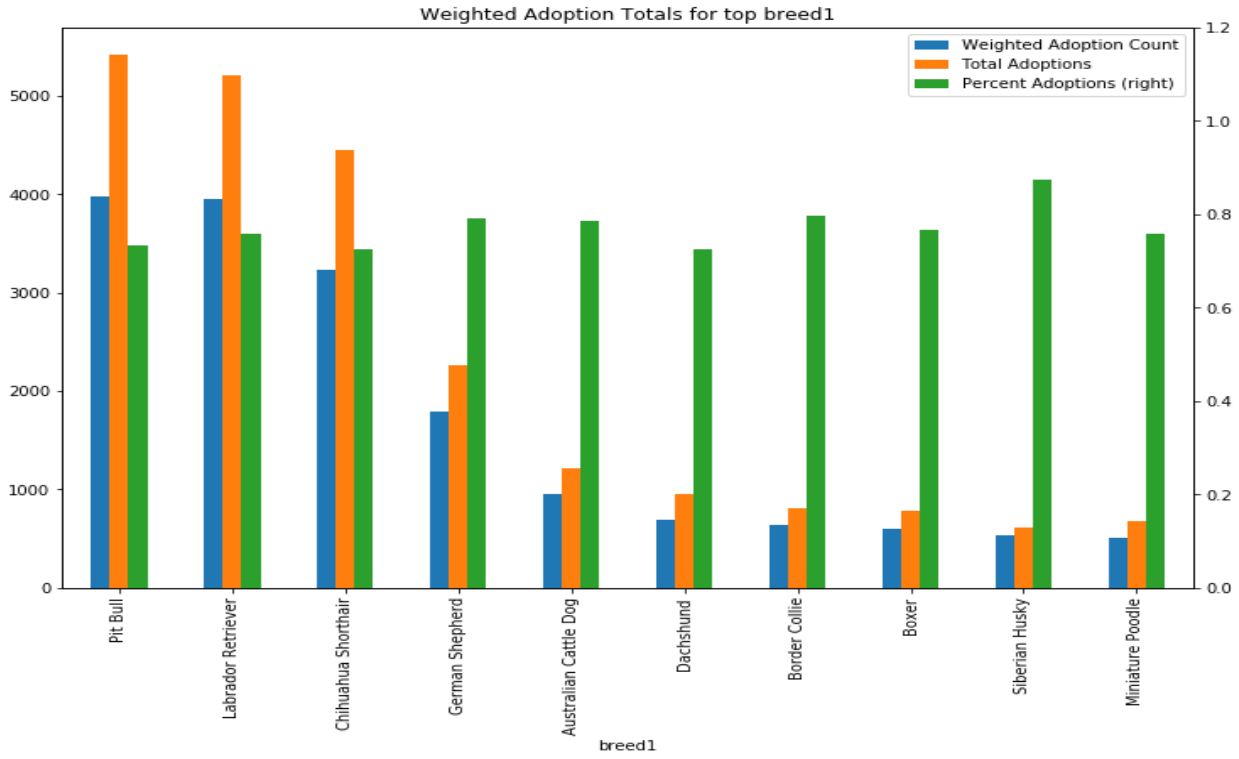

Figure 13: Primary breed adoptions distributions.

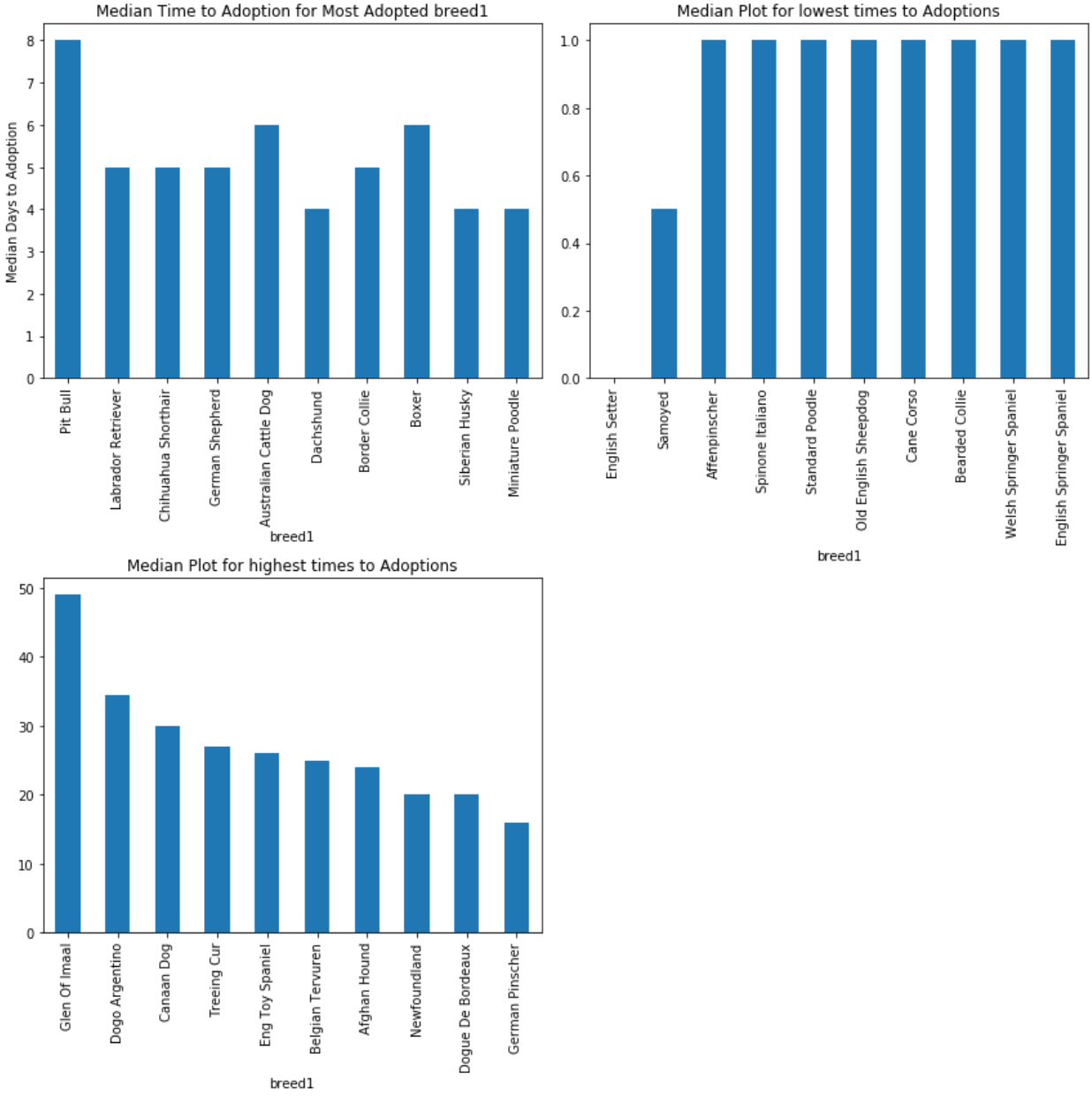

Figure 14: Primary breed time to adoption information.

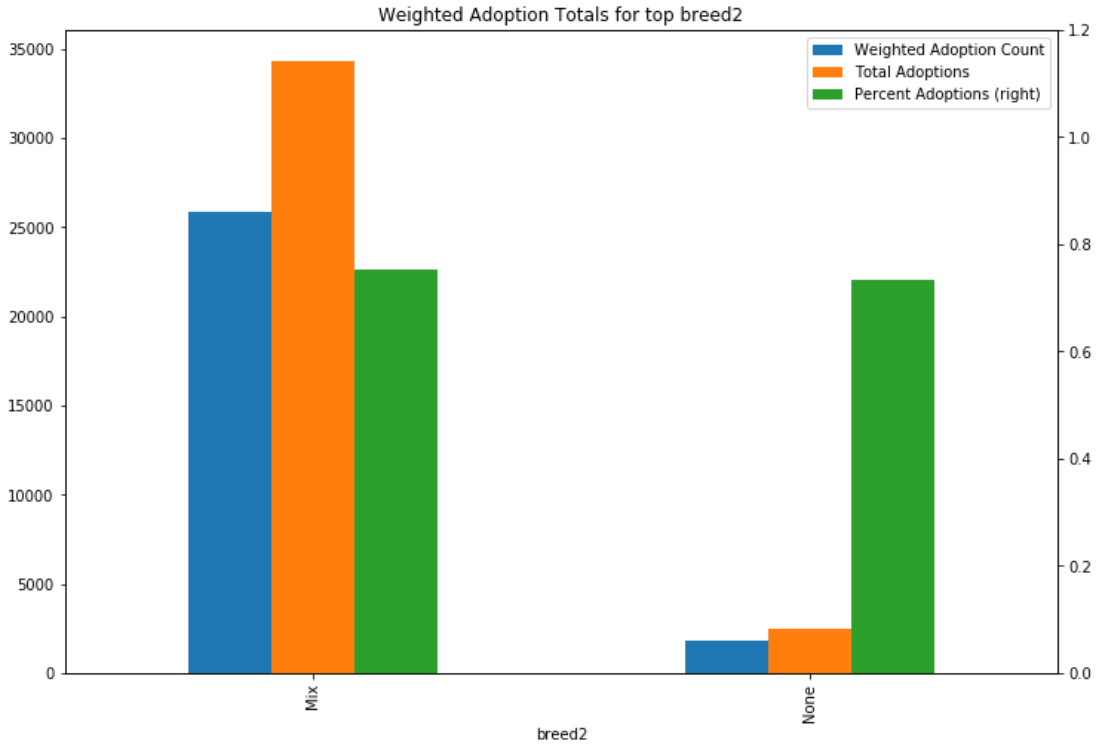

Figure 15: The percent adoptions for secondary breed.

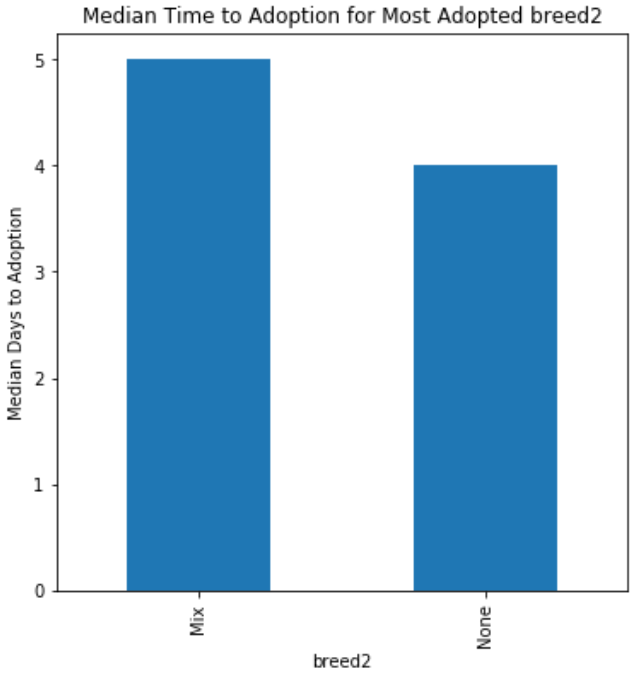

Figure 16: The time to adoption distributions for secondary breed.

It seems that the most popular adopted primary breeds are Labradors and Pit bulls, but these are also the most numerous primary breeds for intakes as well. Their efficiency isn't too bad

either, with about 72% of the pit bulls leaving the shelter under favorable means, and about 75% for the Labradors.

All the dogs in the top 10 weighted adoption counts are above 70% adoptions. Moreover, the more exotic are lesser seen breeds are usually adopted quicker as shown in the lowest median time to adoption plot. All the breeds in there are "exotic", i.e., not pit bull, labs, German shepherds, dachshund, chihuahua, etc.

However, the longest median time to adoption breeds are also exotics. These are both outlier groups with low numbers of total adoption. There may be other factors at play such as age, sex, etc.

The median time to adoption shows that it takes a bit longer for the pit bulls to be adopted at a median time to adoption of 8 days. This still isn't too bad, and the time to adoption for the top adopted breeds are all under 10 days median time to adoption.

The secondary breed information suggests that mixed vs. pure breed doesn't really play an impact on the adoptions or median time to adoption. Although, there are substantially more mixed breed dogs in the shelter than pure breeds. This makes sense if most of the intake types are strays, roaming free to breed with whomever, whenever.

In summary, it seems paying too much attention to the rare breeds would be wasted effort, and marketing efforts should be concentrated on the most populous breeds. Some of them have a negative stigmatism in this case. (Pit bulls, Boxers, etc.)

We will now move onto the age adoption stats.

Break age into groupings:
Puppy (0-12 Months Old)
Young Adult (13-36 Months Old)
Adult (37-72 Months Old)
Senior (73+ Months Old)

The analysis is summarized below in Figures 17 and 18.

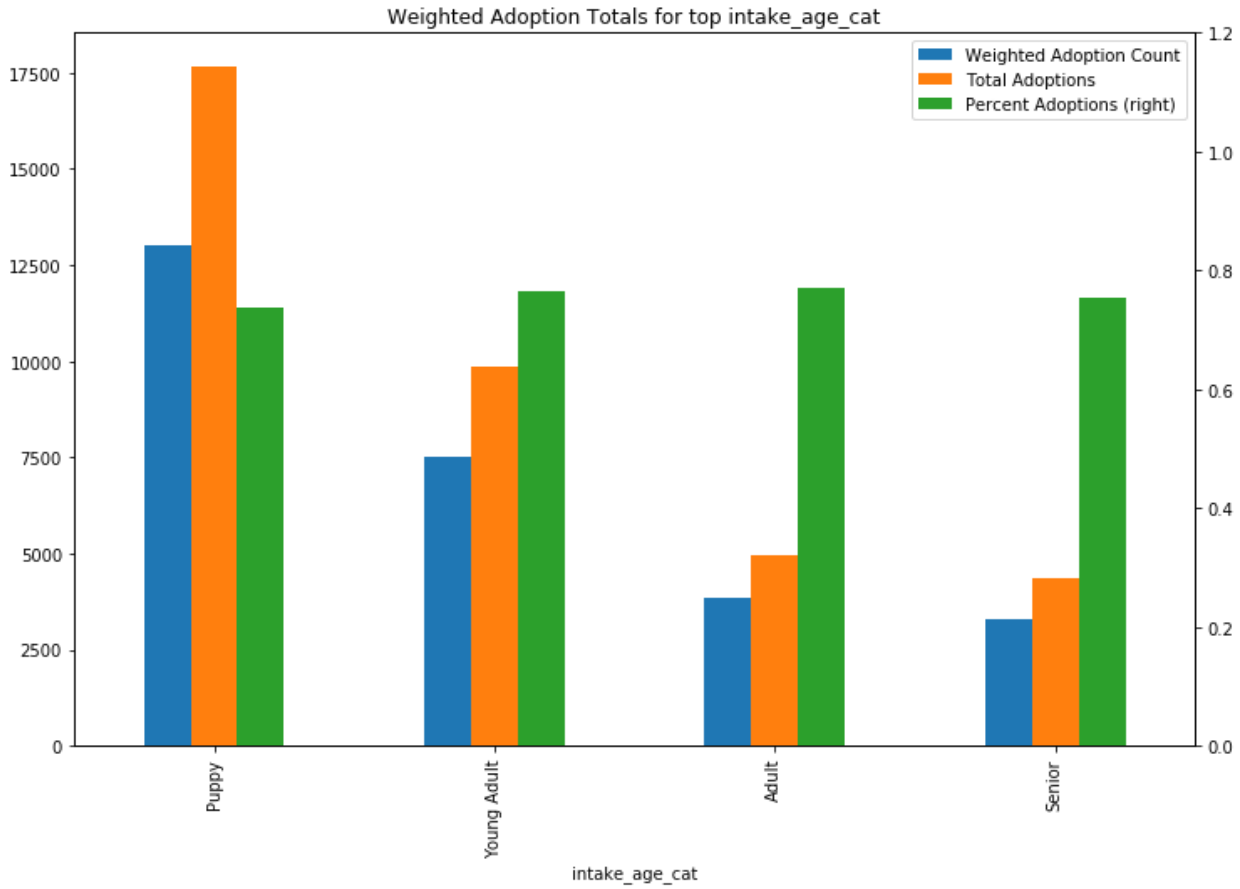

Figure 17: The adoption percentages for intake ages.

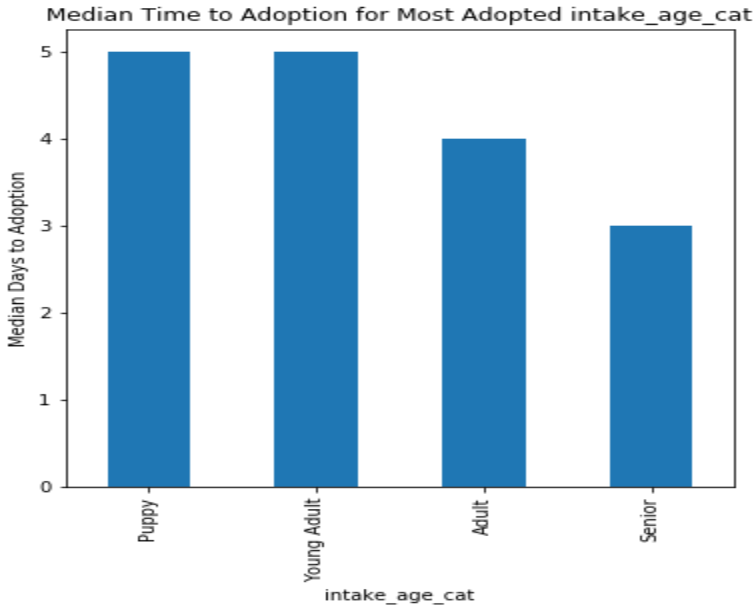

Figure 18: The time to adoption for intake ages.

There seems to be an equal efficiency in adoptions between ages. Also, there appears to be a negative correlation between intake age and time to adoption. This information is useful in determining if older vs younger dogs should need more aid in getting adopted.

Next, we will look at the color adoption stats.

The color analysis is summarized below in Figures 19 - 22.

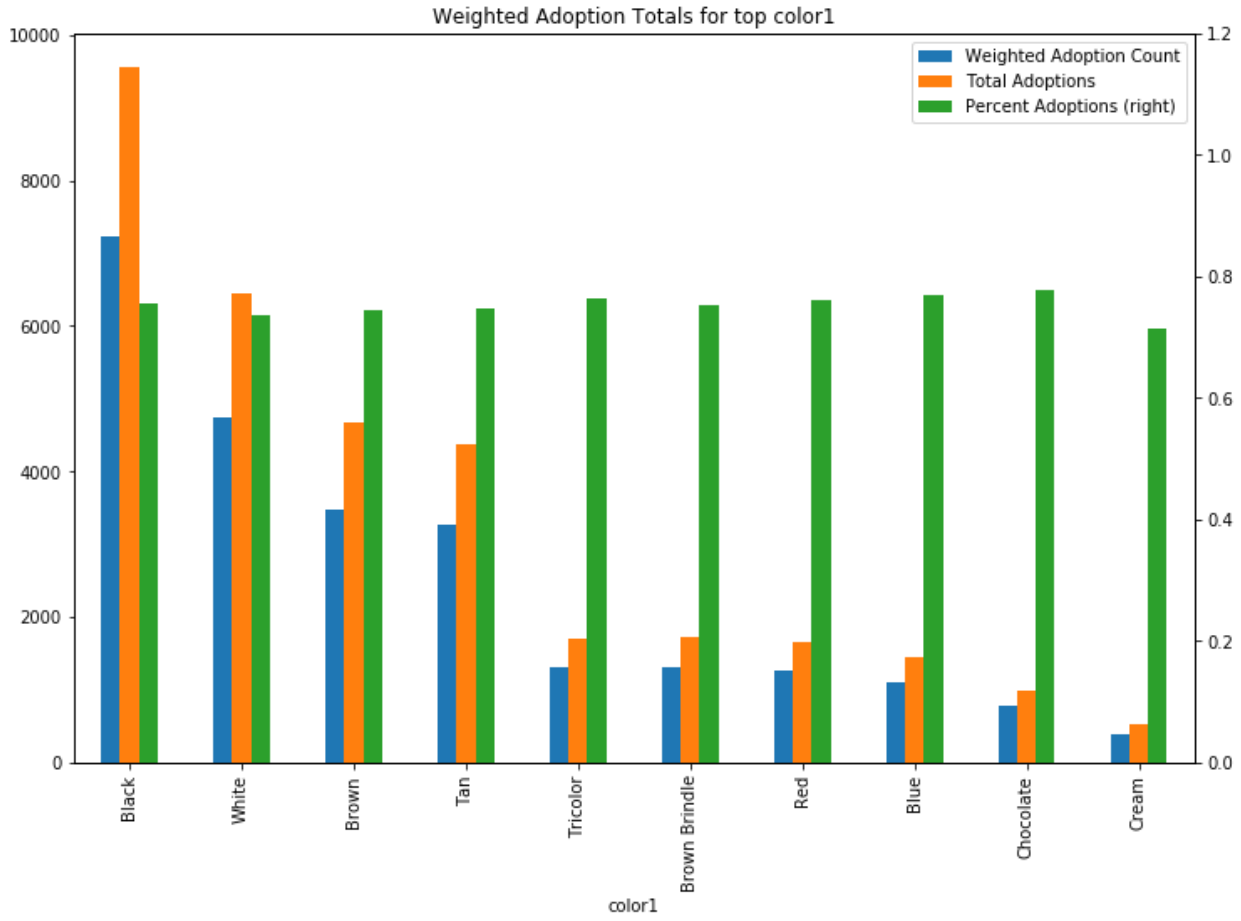

Figure 19: Primary color adoption statistics.

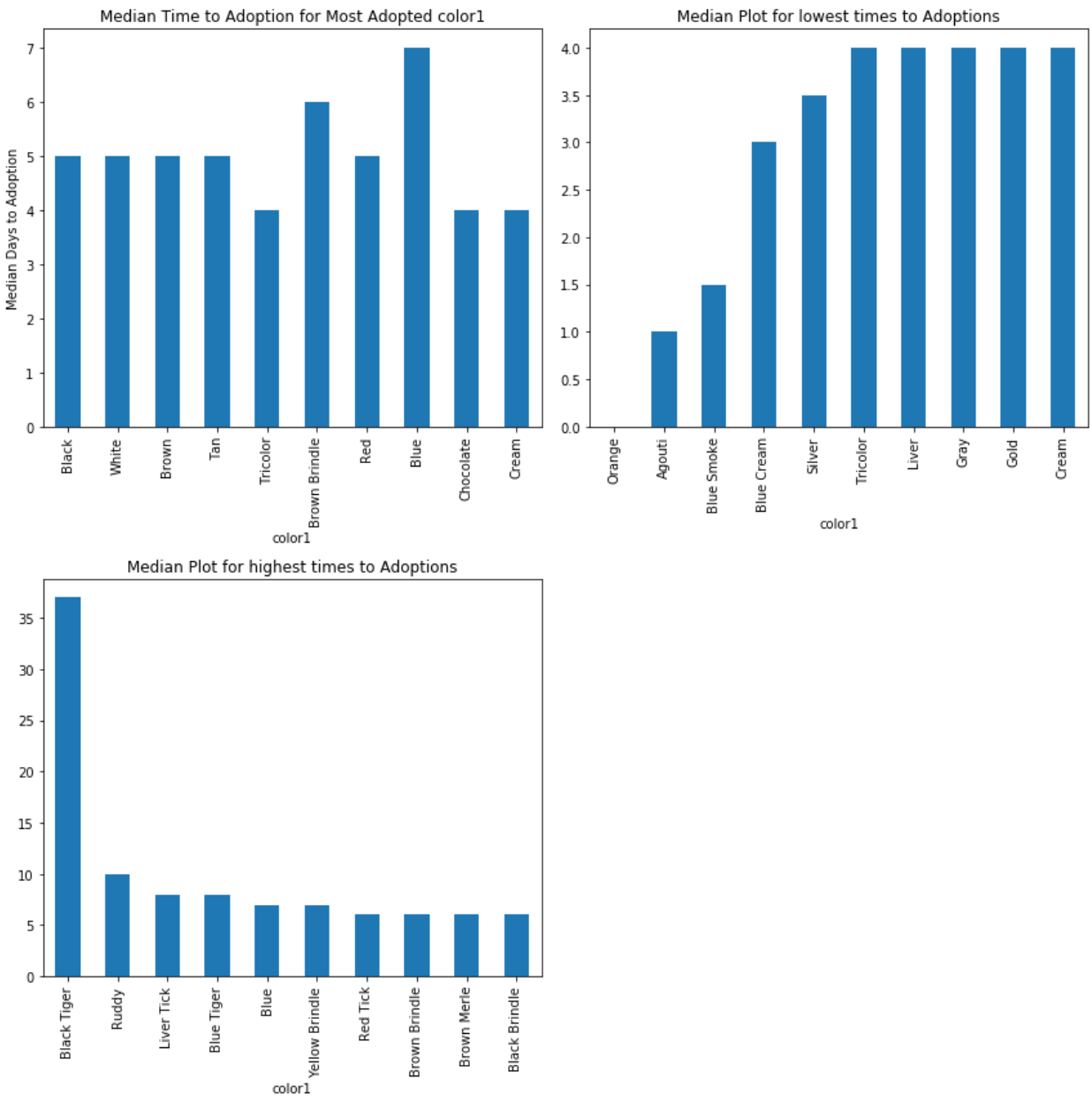

Figure 20: Time to adoption statistics for primary color.

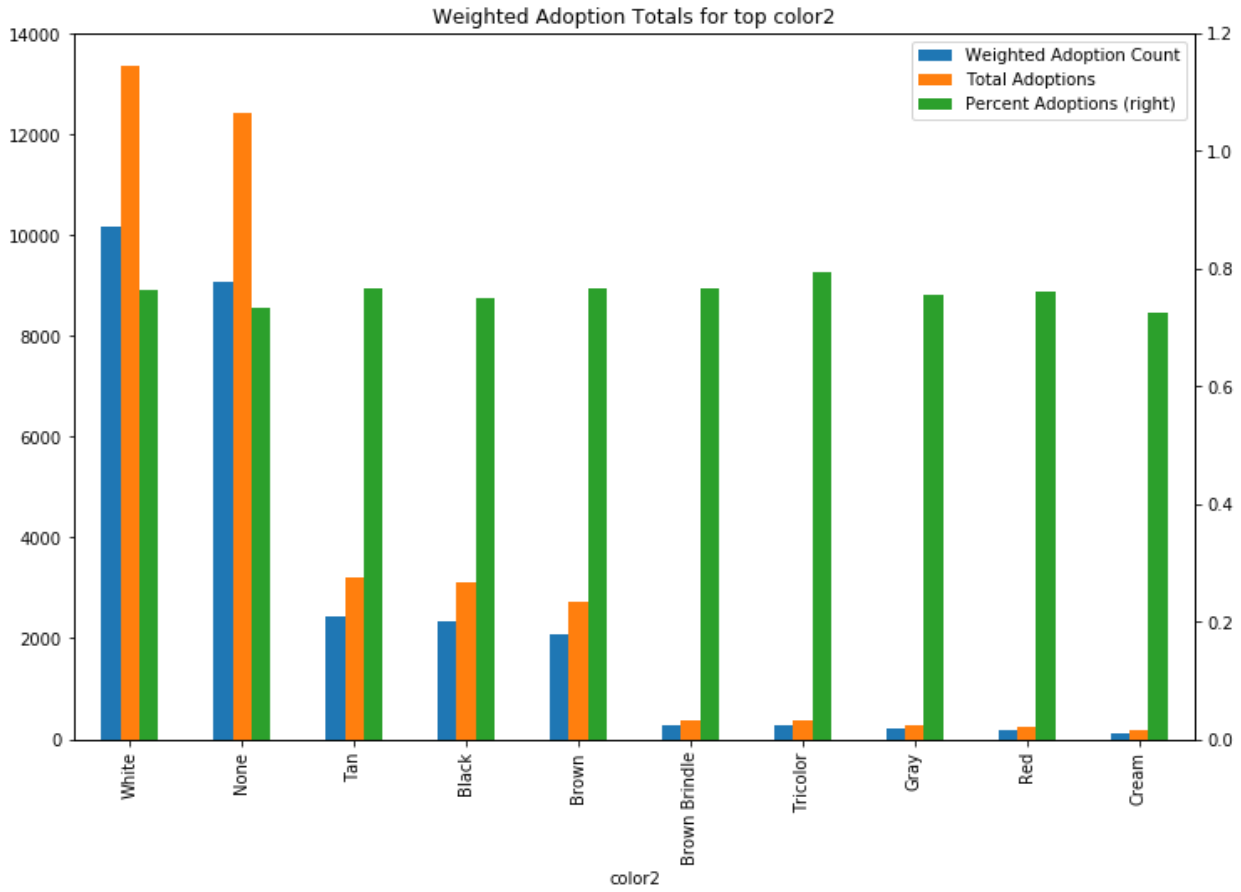

Figure 21: Secondary color adoption statistics.

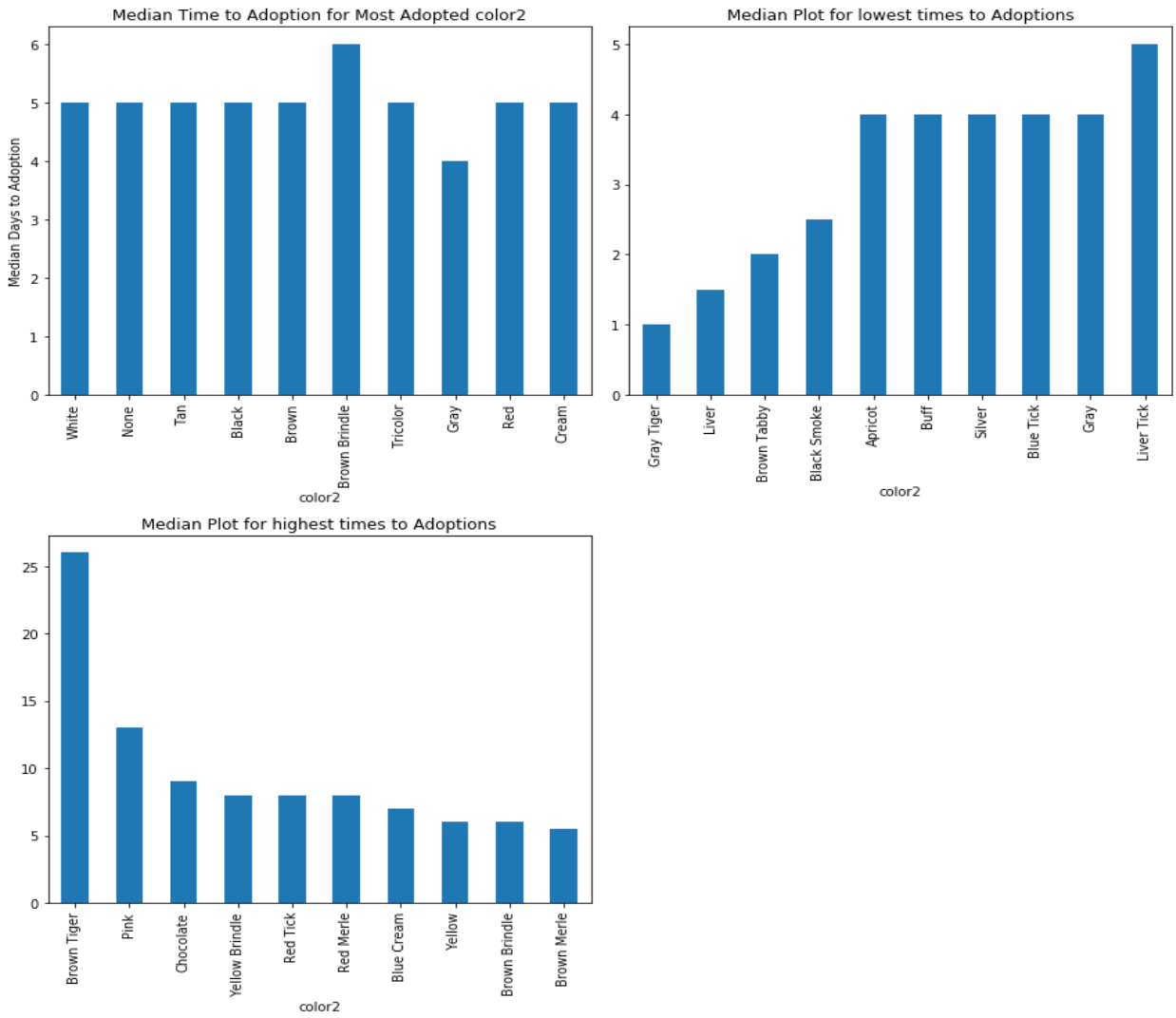

Figure 22: Time to adoption statistics for secondary color.

It seems that neither primary color or secondary color have a big impact on adoptions or time to adoptions shown by the fact that most colors have a median time to adoption of around 5 days. Also, the percent adoptions are all similar. This shows that color might not be a significant factor in determining the adoptability of the dogs. This could mean that color is not important to think about when trying to optimize adoption rates.

Now we can consider how gender affects adoptions.

The gender comparison is summarized in Figures 23-24.

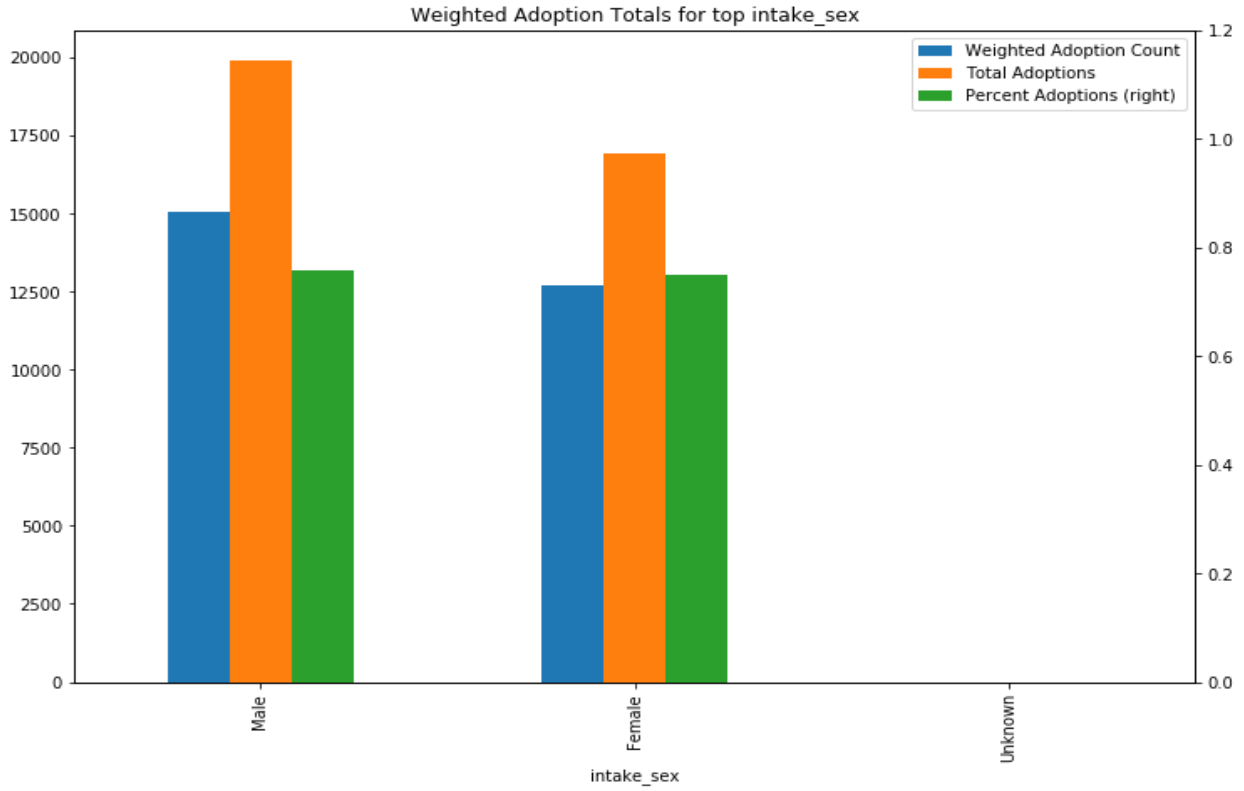

Figure 23: Adoption statistics for gender.

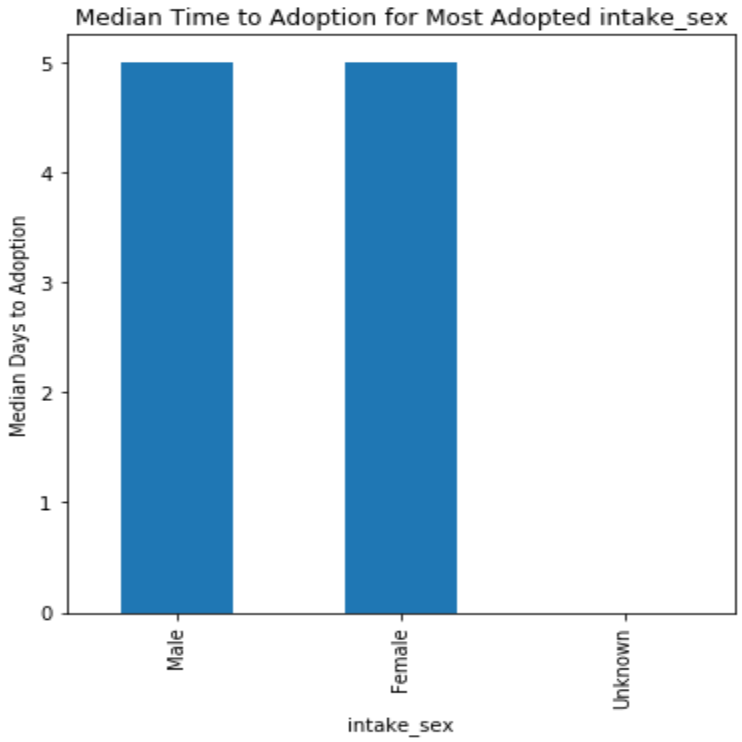

Figure 24: Time to adoption statistics for gender.

On the surface, the gender doesn't seem to have an appreciable impact either. This based on the fact that there are the exact same results for both males and females. This is not the whole story since it is only based on a couple of numbers and plots. However, at this point it seems that gender is not worthwhile to consider in adoption performance or direction of aid.

Now we can look at the fixed adoption stats.

The fixed comparison is summarized in Figures 25 and 26.

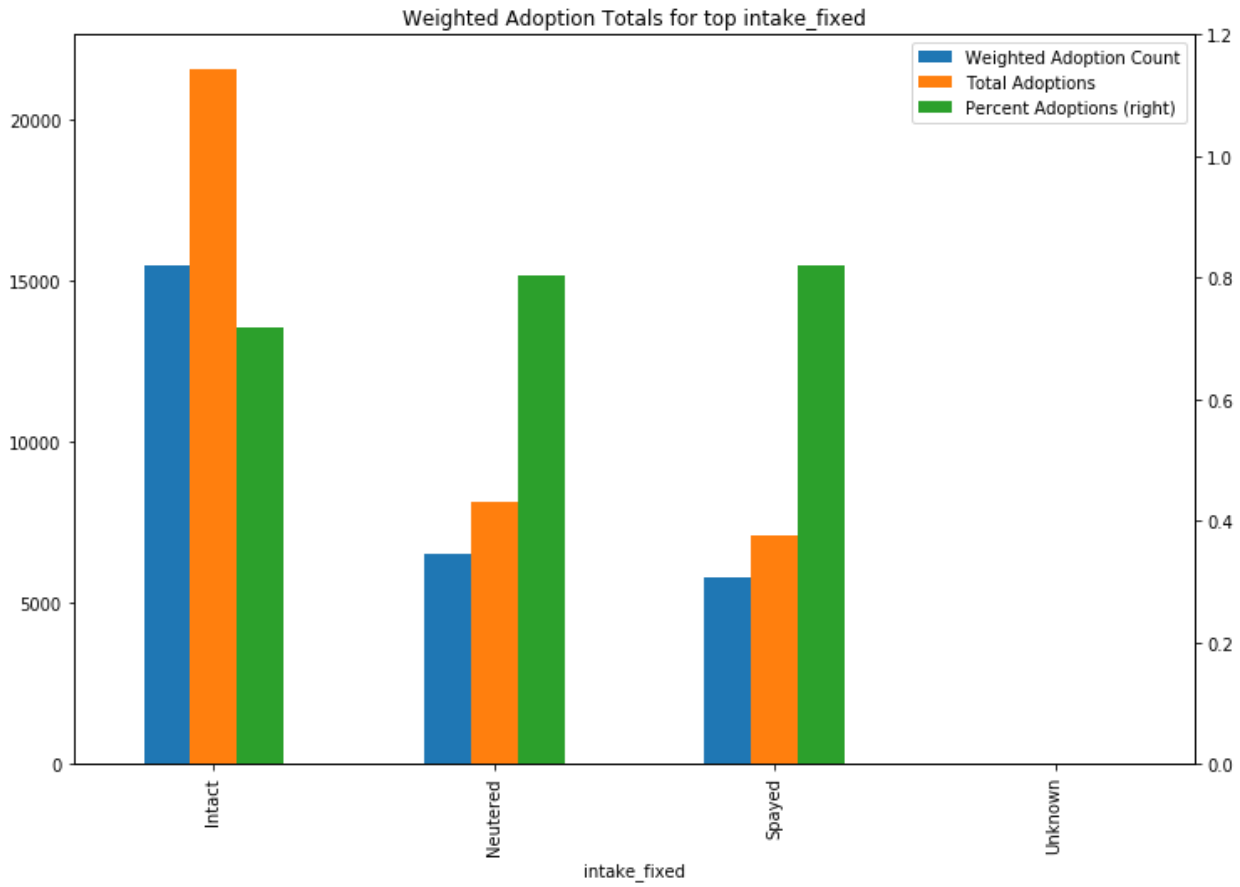

Figure 25: Fixed adoption statistics.

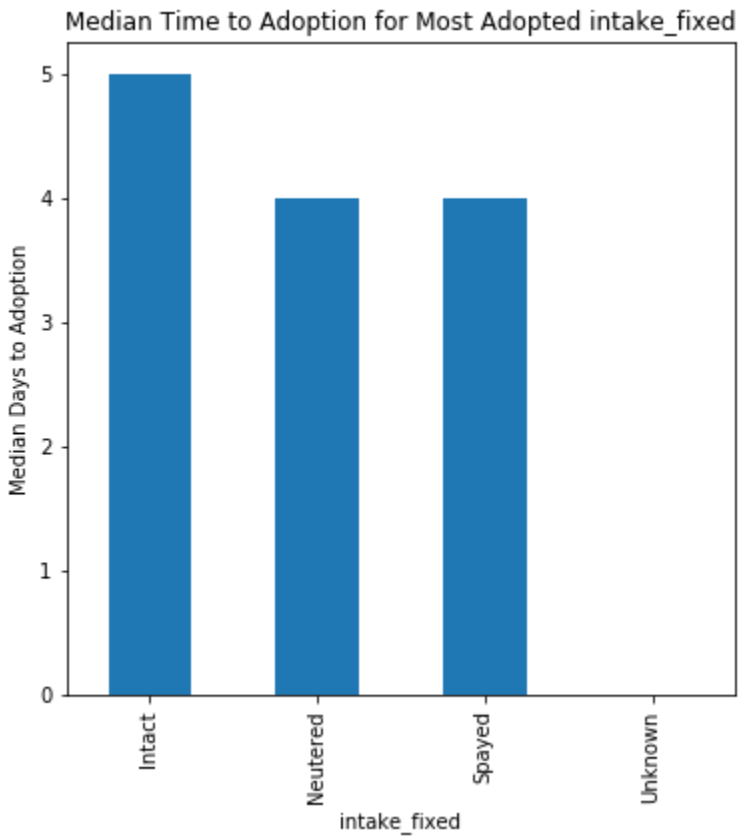

Figure 26: Time to adoption statistics for fixed.

It seems there may be a slight favor for the fixed dogs to get adopted faster and more often than the intact intakes. The median time to adoption is shorter for fixed dogs. The percent adopted is slightly higher for the fixed dogs as well. This apparent time reduction to adoption could be from the fact that the recorded adopted times factor in surgery time as most dogs leave the shelter fixed even when they came in intact. This information may indicate that the shelter should push for all pets to be spayed or neutered to increase adoption efficiency and decrease time to adoption.

f) Seasons

The seasons analysis is outlined in Figures 27 and 28.

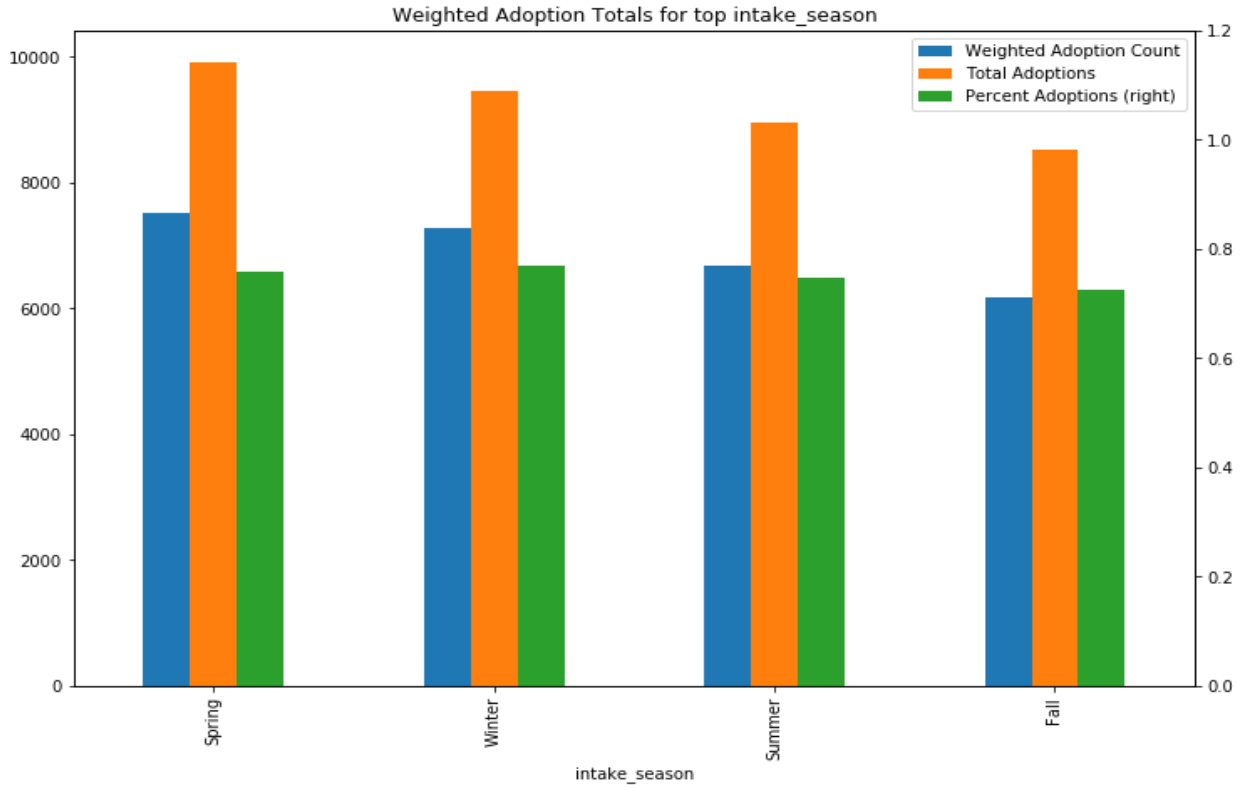

Figure 27: Seasons adoption statistics.

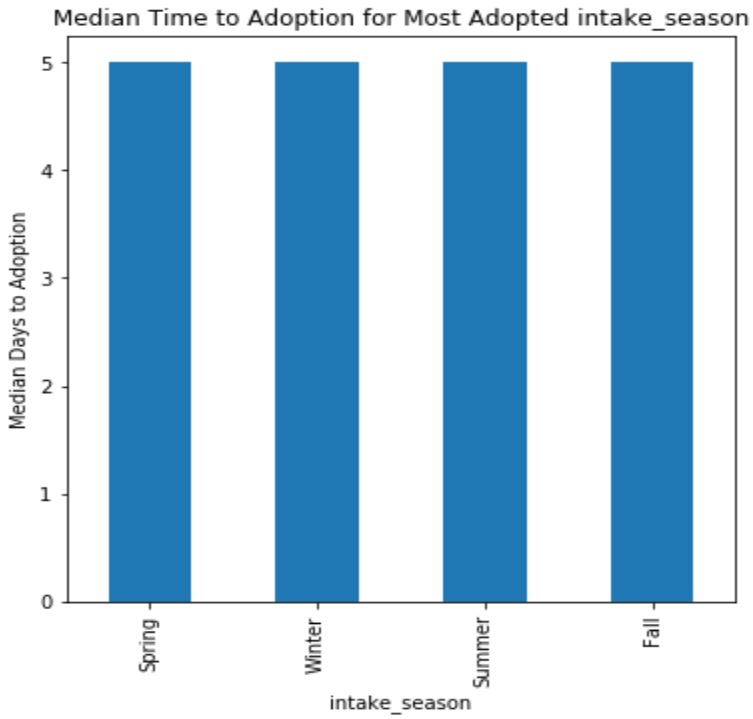

Figure 28: Time to adoption statistics for seasons.

As shown in the time series plot at the beginning, the time of year seems to have a slight effect on the adoption percent. However, the median time to adoption seems to be the same for

each season. Moreover, the total number of adoptions seems to be affected by the time of year. This could be useful information. It could mean that adoption events need to occur more often in the Fall to boost adoptions at that time.

4) Now we will consider the outcome types' correlations with intake features.

Considering what intake features such as breed and gender that end up as different shelter outcomes can draw attention to what dogs need attention to end up as Adoptions. This can also reduce the number of dogs that get transferred from the shelter or die in the shelter. Transfers create additional strain on surrounding resources and on the pet. Transfers are also a longer process to adoption and thus are more expensive and less efficient of a process.

To start, we will break the outcome types into three main groups: Adoptions (Adoptions, Return to Owner, Rto-Adopt), Transfers (Transfer), and Deaths (Died, Euthanasia, Disposal, Missing, Unknown).

We will also look at the following intake features:

a) Breeds
    Primary (Top 10)

b) Gender (Comparison)

First, we will consider the primary breeds top 10 stats.

A look into the top 10 primary breeds by outcome type shows that there are a lot of top breeds in all outcome types. The only non-popular breeds that stand out are that Shih Tzus are transferred a lot to other shelters or organizations. Also, Chow Chow is on the top 10 deaths list.
There also seems to be few deaths in the shelter which is good. This type of information shows the typical breed distribution for the outcome types, which is mostly the populous breeds in the shelter. There does not seem to be any abnormalities here. For example, there is not a breed that has overwhelmingly larger transfer or death outcomes when compared to adoptions. Information like this can be used to set performance and abnormality check indicators.

Next, we will consider the gender stats.

Considering the gender by outcome types revealed that it seems to be evenly split in all outcome types. Also, there are not a disproportionate number of negative outcomes compared to adoptions for either gender. This too can be used as a normality barometer and gauge for performance.

5) We will now move on to a special case study on return visits.

I decided to see what the intake and outtake data looked like for the animals that make two or more visits to the shelter. It seems that a fair amount of return visits are due to owners giving their pets up and trying to reclaim them. They may have been trying to utilize the shelter as "free boarding". As shown earlier, a low percentage of owner surrenders end up back as return to owners. Also, this would eat up a lot of valuable shelter resources and potentially cause lots of unnecessary heartbreak. Return visits can usurp resources and take away efficiency from the shelter's operations. Also, return visits can hamper the shelter's ability to serve new pets. Identifying potential return cases and the circumstances around them can help in changing policies or procedures to prevent them.

6) A little Time Series Analysis

We will look at the trends of Adoptions over time for the Top 5 breeds Identified earlier: Pit Bull, Labrador Retriever, Chihuahua Shorthair, German Shepherd, and Australian Cattle Dog. Breeds seem to be the most interesting statistically to look at by first inspection.

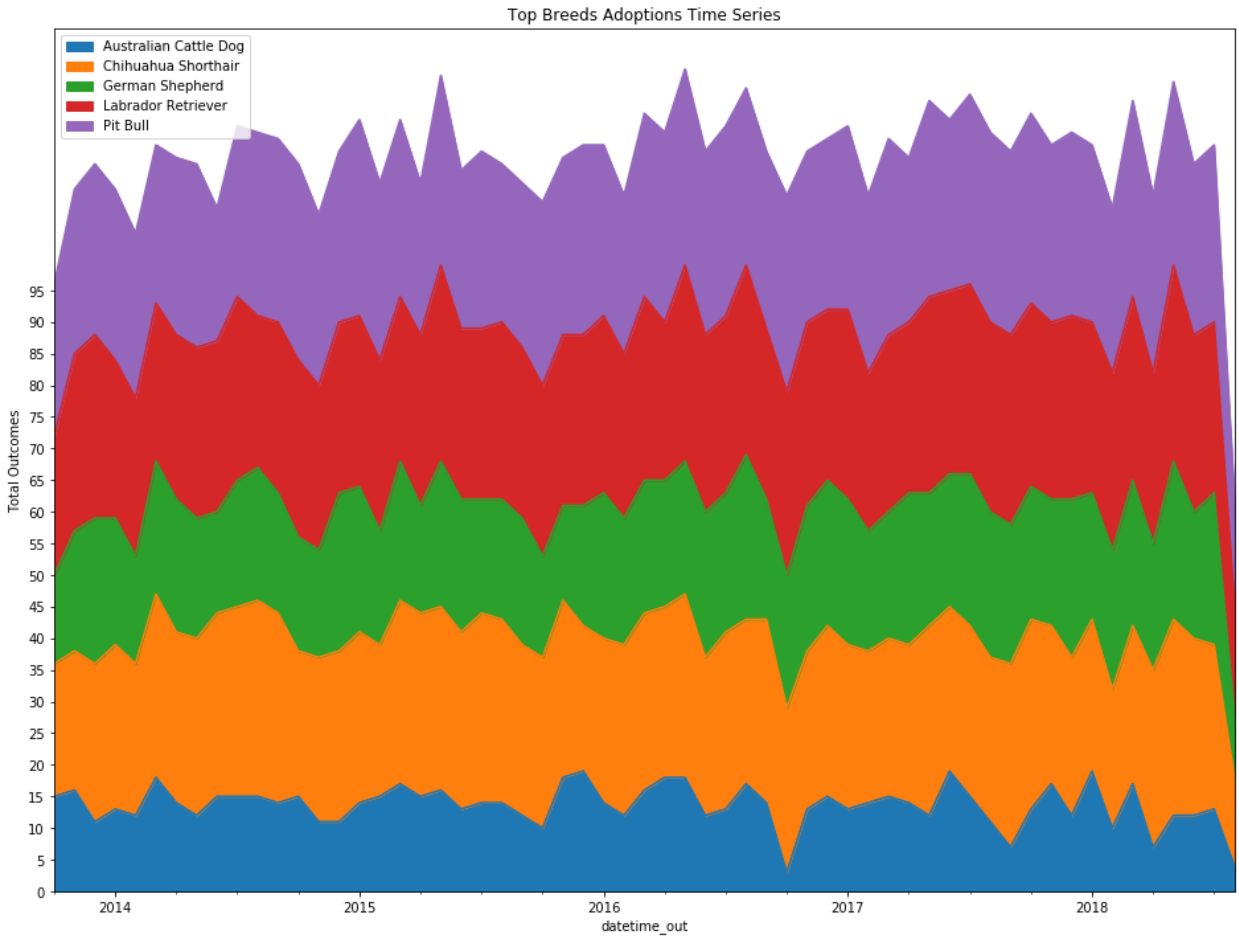

Figure 29: Time series plot of the top 5 primary dog breeds.

Interestingly, it looks like they may run adoption events in the fall and spring. A lot of the peaks seem cyclical and line up well with the fall and spring months. Contrastingly, this could be

because people feel like adopting more before and after winter. By looking at the breed adoptions over time, a general overview can be obtained of when adoption/marketing efforts can be intensified for certain breeds to get them adopted sooner and more reliably. Also, more analysis can be done into the time behavior for the other features as well.

INITIAL STATISTICAL ANALYSIS FINDINGS

To start, I will go over a few reminders and definitions. Adoptability will be calculated as the percent of adoptions in all other shelter outcomes. Also, I will employ hacker statistics for all testing as it is widely applicable, and I have enough computing power to produce viable results.

For both tests, the null hypothesis is that there is not a difference, and the test will be set up under this assumption. The assumption being that the distribution of adoptions and time to adoptions between the two groups are not different. I will use permutation replicates to generate samples to create a distribution of differences for both adoptability and standard deviations of time to adoptions. Time to adoption will be assessed via the difference of expected values of exponential distributions (standard deviation) of the time to adoption values for each group. This is done because time to adoption is roughly exponentially distributed as shown earlier and this is a more accurate summary statistic for time to adoption. The p-value of the test will be calculated by taking the number of permutation replicates that have a difference as extreme or more extreme than the empirical difference divided by the total number of permutation replicates generated for both adoptability and time to adoption.

The questions/hypotheses drawn from the EDA are:
1. Does age play an important role in the adoptability/time to adoption of the dogs?
Puppies are defined as being less than or equal to 1 year old, and all other ages or defined as non-puppies. The test significance level is $\alpha = 0.05$. The null hypothesis is that there is not a difference in distributions of adoptability or time to adoption.
The main result is that we can reject the null hypothesis. The difference in adoption percentages is significant. Non-Puppies seem to have a higher adoptability than older dogs by 2% empirically. However, the expected value for time to adoption for puppies is significantly shorter with an empirical difference of 28 days shorter than non-puppies. The results suggest that more time and attention should be spent on trying to get older dogs adopted sooner. Puppies don't seem to have a very hard time getting adopted as compared to the older dogs.

2. Does having a secondary color make dogs more adoptable and decreases the time to adoption?
Solids are defined as having a secondary color value of 'None', and all other dogs have mixed coat color. The test significance level is $\alpha = 0.05$. The null hypothesis is that there is not a difference in distributions of adoptability or time to adoption.

The main result is that we can reject the null hypothesis for adoption percentage. The empirical difference of 2.4% greater for Mixed color dogs is significant. The time to adoption result is not significant, and the difference of 2 days expected value is within normal variation. Although the result is significant that mixed color dogs are adopted more than solid colored dogs, it is only by 2.4% empirically. I would not devote too much time focusing on color as the key factor in adoptions. Moreover, there are a lot more mixed color dogs than solid color dogs. Plus, the difference in time to adoption is not significant. Maybe try bringing roughly 2.4% more solid color dogs to adoption events as a test.

3. Is there a difference in adoptability/time to adoption for males as compared to females?

Since there are only 282 unknown values for gender, I will lump these in together with females as the count of females is lower than males. The test significance level is $\alpha = 0.05$. The null hypothesis is that there is not a difference in distributions of adoptability or time to adoption. The main result is that the test for adoptability was significant, and the null hypothesis can be rejected. The empirical difference suggests that Male dog adoptability is 1.3% higher. The time to adoption result is not significant, and the difference of 4 days expected value is within normal variation. The result shows that there is not a significant difference in the time to adoptions for males and females. The significant result in percent adoptions suggests that there are 1.3% higher adoptions for males than females. I would not focus too much on the adoptability of males to females. Simply presenting an equal ratio of males to females or slightly shifted to more females would probably be the best course of action. Also, there are more males in the shelter historically.

4. Does the population size of a breed in the shelter play an important role in adoptability?

To test this, the groups will be the top 5 primary breeds in intakes vs. the rest of the breeds. The test significance level is $\alpha = 0.05$. The null hypothesis is that there is not a difference in distributions of adoptability or time to adoption.

The main result is that the test for adoptability is significant and we can reject the null hypothesis. The empirical difference suggests that the less populous breeds have 0.7% higher adoptability. The time to adoption test has a significant result as well, and the null hypothesis can be rejected. The empirical difference of 11 days in expected value suggests that the less populous breeds are adopted sooner. The result of this test suggests that populous breeds in the shelter need more attention than the non-populous breeds. It seems that, even with dogs, being rare is desirable. Both the adoptability and time to adoption are favored to the non-populous breeds significantly. The rarer breeds seem to handle themselves. While more attention and effort in marketing/adoption events needs to be geared toward the more populous breeds in the shelter. There are more of those to adopt as well. Thus, moving them would more greatly reduce the shelter population than moving the less populous breeds.

5. Does the primary breed American Kennel Club group play an influential role in adoptability/time to adoption?

To test this, the breeds are separated into the AKC breed groups and tested one vs. all for each AKC group. The test significance level is α = 0.05. The null hypothesis is that there is not a difference in distributions of adoptions or time to adoption.

The results are a mixed bag of significant and non-significant results. No one AKC group stands out among the rest. However, some groups pop up as being potentially troublesome. These include the groups such as Toy with a significant adoption ratio value suggesting their adoption ratio is lower as compared to the others. More time should be spent in trying to get the groups with lower adoptability represented. Also, more educational events could be conducted to educate potential pet owners on the difference between the groups and breeds. This could potentially lead to a reduction in return visits as the pet owner would adopt on an educated basis. They would more likely adopt the type of dog that suits their lifestyle better. Also, Misc and Terrier groups seem to take longer to adopt than the other groups. More attention should be paid in getting these dogs adopted. The results are summarized in Table 2 below.

Group	Adoption Ratio	Time to Adoption
Sporting	Not Significant	Not Significant
Hound	Not Significant	Significant(less)
Herding	Significant(more)	Significant(less)
Terrier	Not Significant	Significant(more)
Non-Sporting	Not Significant	Not Significant
Toy	Significant(less)	Significant(less)
Working	Significant(more)	Not Significant
Misc	Not Significant	Significant(more)

Table 2: The results for the AKC statistical analysis tests.

6. Is there a correlation between time to adoption and intake age?

To test this, all the dogs who have outcome_type of adoption will be organized by intake age and correlated with time to adoption. The test significance level is α = 0.05. The null hypothesis is that there is not a correlation. The test setup is a little different than the previous tests. To simulate and test the null hypothesis, we need to first calculate the empirical

correlation coefficient. Next, we permute one of the variables and keep the other steady. Then, we calculate the correlation coefficient of the permuted data. We repeat this many times until a distribution of correlation coefficients are generated. Finally, we calculate the p-value, which is the count of instances where the permuted correlation coefficient is at least as extreme as the empirical correlation coefficient.

The empirical correlation coefficient was 0.03. This is not very good, but it is significant. The null hypothesis can be rejected. This suggests that the older the dog is upon intake, the longer it will take to get adopted. This could be a trouble area to look out for in the shelter as well. Focusing more time and attention promoting older dogs vs puppies is probably a good idea. However, looking at the scatterplot of Average time to adoption vs. intake age in months the conclusion becomes convoluted. The maximum time to adoption seems to occur for young puppies. This number may factor in time needed to ween and/or fix the puppy. Then the graph grows almost logarithmically, plateauing at around 50 months intake age. After this the time to adoption drops significantly. There are a lot fewer cases of old dogs in the shelter, and these can be considered outliers. The general trend is that puppies are adopted sooner than older dogs as shown previously.

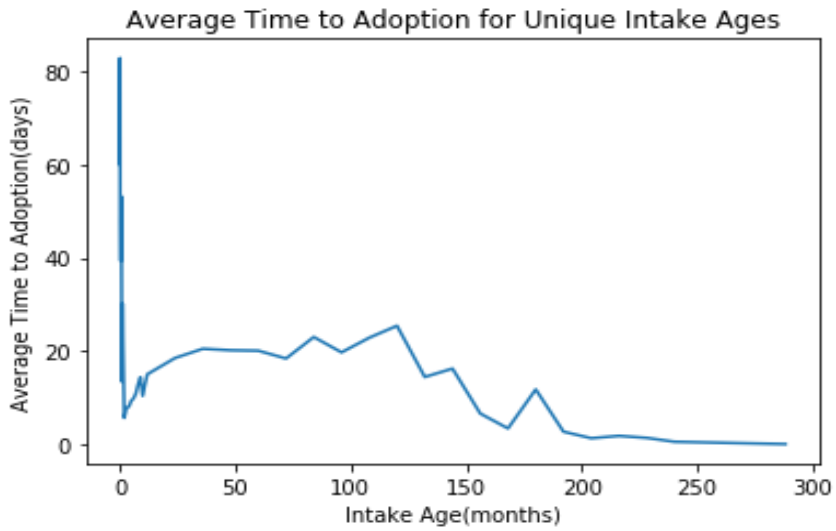

Figure 30: Time to adoption (averaged for each unique age) in days vs intake age in months.

7. Are adoptions affected by seasons?

To test this, the groups will be the seasons, and they will be tested one vs. all for each season. The test significance level is $\alpha = 0.05$. The null hypothesis is that there is not a difference in distributions of adoptability or time to adoption.

The results are all significant for the adoptability tests. This means that we can reject the null hypothesis that the distribution of adoption ratios is not different. The test results suggest that Summer and Fall may be times with less adoptability and that Spring and Winter have higher adoptability. More time and effort could be spent in the summer and fall to even out the adoptability of dogs throughout the year.

For time to adoption, the results are mixed significance. However, spring seems to have lower times to adoption (expected value) as compared to the other seasons. Moreover, Fall seems to have higher times to adoption (expected value) as compared to the other seasons. Once again, more can be done in the months with higher time to adoption to aid in increasing the efficiency at which the dogs are adopted throughout the year. The results are summarized below in Table 3.

Season	Adoption Ratio	Time to Adoption
Spring	**Significant(More)**	**Significant(less)**
Winter	**Significant(More)**	**Not Significant**
Fall	**Significant(less)**	**Significant(more)**
Summer	**Significant(less)**	**Not Significant**

Table 3: Results for the seasons statistical analysis tests.

MACHINE LEARNING

Under construction.