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 outake 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 outake 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 dataframe. I converted the list of dictionaries into a pandas dataframe 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 dataframes as CSV files as to have a reset point if I needed to start from scratch. A quick inspection of the dogs only intakes info and first few rows show that animals were all given a unique animal_id for distinguishment. After looking at the value counts of the unique animal_ids, it is clear that there were multiple intakes for some animals. The summary of issues found upon initial inspection of all the columns and dataframe 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 actually 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 dataframe 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

1) Multiple Intakes for the Same Pet

To take into account repeat intakes, I first sorted the data frame by animal_id and the intake date. Then, I looped through and labeled intakes with the corresponding number of the current intake for the dogs. 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 actually 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.

2) There are two date columns.

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

3) Color attribute contains multiple colors.

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

4) The intake age isn't machine digestible and is inconsistent.

The ages were expressed in years, months, weeks, and days. I chose to have all the intake ages be in months. I first split the age up into its number and time period moniquer. Then, I created and matched regular expressions for the different time periods. This, along with the number part of the intake age allowed for the conversion of the intake ages into months.

5) 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 asteriks from the values that had them.

6) 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.

7) Sex upon intake is actually two features in one column the sex and then the spay/neuter info as well.

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" for both.

8) 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, breed2 column value became "Mix". Lastly, if there was not

secondary breed info in the breed data, breed2 became “None”. Moreover, the total number of unique breeds were similar to the number of breeds recognised by the AKC.

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

B) Outtake Data Frame Wrangling

The outtake data had very 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 there would be no need for a shelter then. 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 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, assume that color and breed don't change from intake to outtake. Thus, these columns were dropped from the outtakes data frame along with the animal type, date of birth, and repeat outtake date column.

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

1. The age upon intake isn't machine digestable.
2. There are nulls, No Name, and values with asteriks in the name column.
3. Sex upon outtake is actually two features in one column the sex and then the spay/neuter info as well.
4. Need to parse the datetime column.
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 outtake 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 similar to 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 is the unique identifier created to take into account 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 outtake numbers. This means that the data frames when merged would not be misaligned. Also, I did a quick check to see if there were animal ids that showed up in the outtake data that weren't in the intake data. There were 381 entries that fit this description. 381 out of ~45000 are not a lot so, 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 and there are null values for the outtake data related columns. 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 just 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 will be needed in order to input the data into Machine Learning Models later. Also as will be discussed later in the initial findings section, other columns were created in the joined data frame to aid Exploratory Data Analysis. Remember, the intake age and outtake age are in months, and the shelter time is in days.

OTHER POTENTIAL DATASETS

- Use data from other shelters located all over the world.
- Wrangle data from adoption websites.
- Use the other animal information supplied from this shelter. Cat info could very well affect the adoption statistics of dogs in the same shelter.
- Use information from the shelters website to bring in factors like pictures, animal adoption events, etc.

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. Adoptability 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 other shelter outcome types.

1) **Now let's look at the adoptions as a whole.**

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 vast majority of outcome_subtypes were classified as normal with only a few other examples. The value counts of the outcome_type category are much more interesting as shown in the graph below:

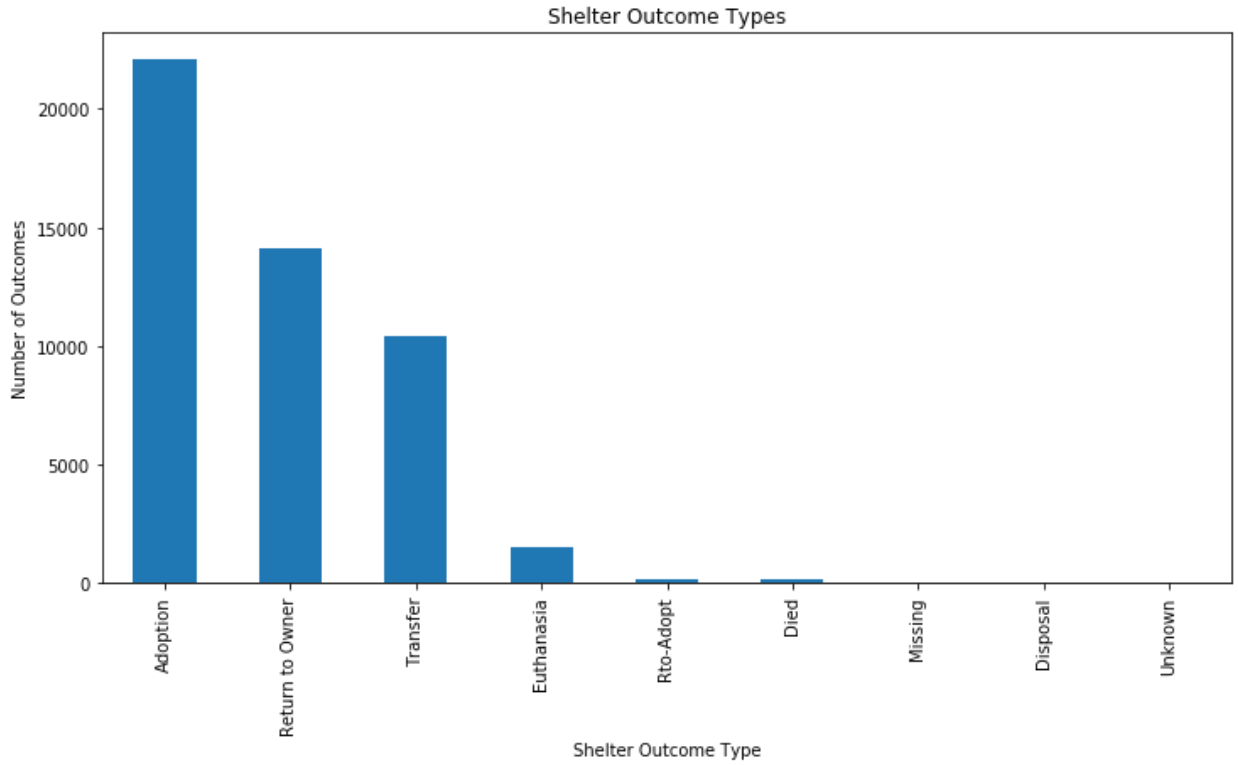

Figure 1: The outcome type value counts for dogs.

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 on, a deeper dive into the different adoption outcome type statistics is done.

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

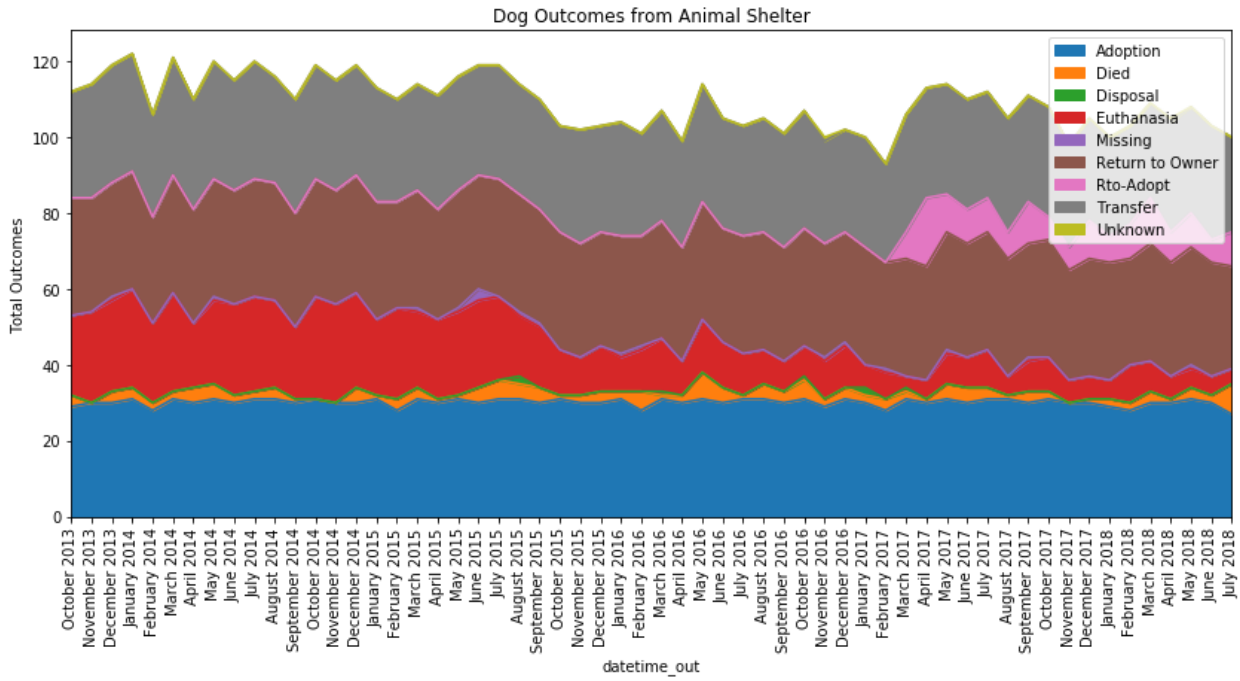

Figure 2: Time series plot of the shelter outcomes broken up by outcome subtype.

There are several insights that come from this plot:

- There seems to be about 30 adoptions a month or one a day, which is quite steady.
    - There seems to also be a dip in adoptions in the spring time around February March of each year.
- There are another 30-50 as return to owner/return to owner-adopt, this leaves about 40-60 dogs per month that could have had better outcomes.
- Transfers to other shelters (which is better but not counted as adopted) remain pretty 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 etc.. 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.
- Euthanasia dropped by about half from 30 to 10-15 per month around October-November 2015, and seems to gradually decreasing over time, which is great.

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 suggest that the time between adoptions aren't completely random and are affected by and correlated with outside influences. 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 look into the intake data.

I chose to break down intakes by dog features and get basic statistics.

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)

a) Breeds

The breeds were wrangled into two columns: a primary and a secondary breed. Let's look at the top 10 primary breeds value counts.

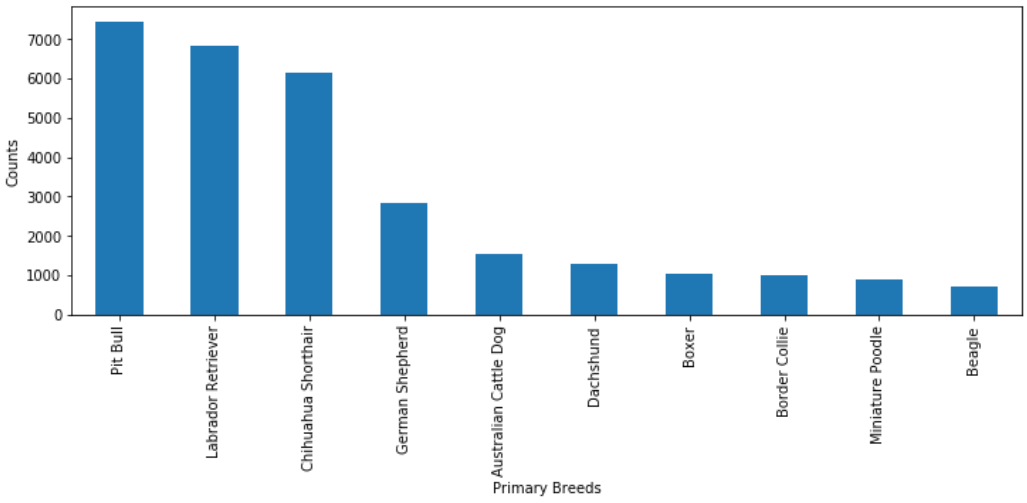

Figure 3: Top 10 primary breeds for intakes.

Most of the shelter intakes are accounted for in the top 10. A look into the secondary breed data shows that 84% of the data fell into either Mix or None categories. I decided to label the other values as Mix. The result is below in Figure 4.

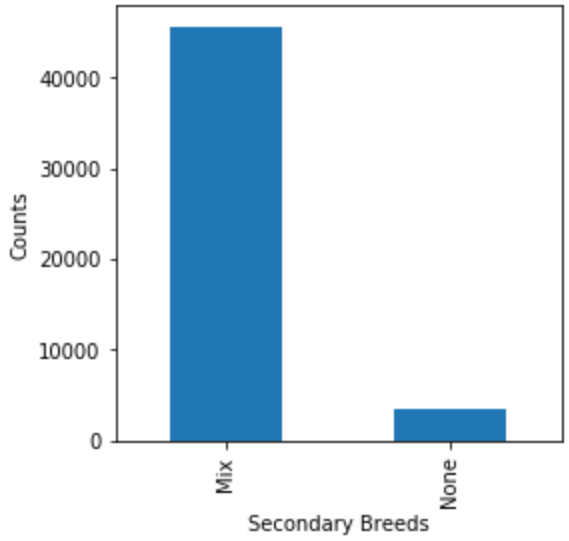

Figure 4: The secondary breed distribution for intakes.

Almost all the dogs in the shelter are mixed breed.

b) Age Upon Intake

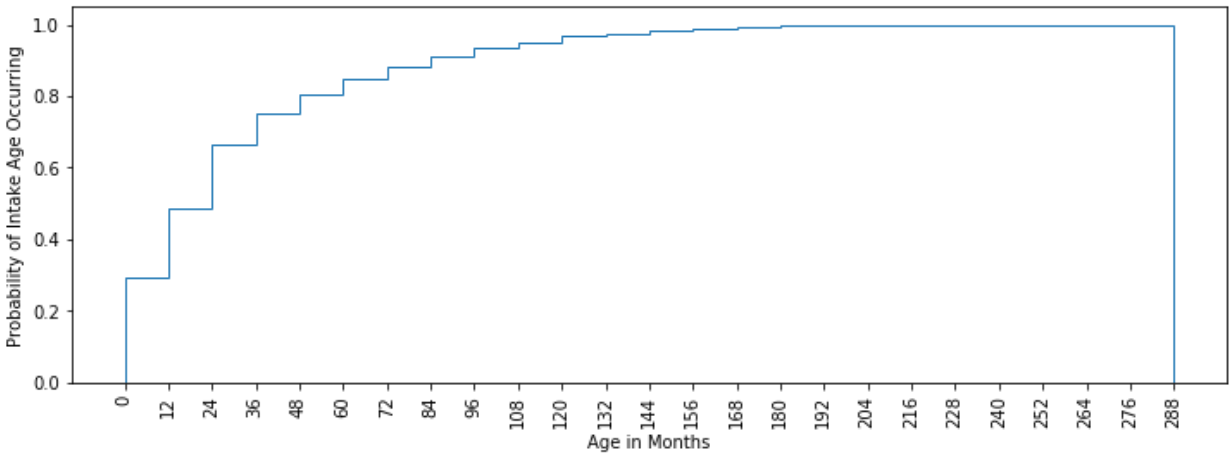

Figure 5: The cumulative distribution function of the age upon intake.

80% of the dogs that come in are 5 years old or less, which is middle age for a dog. Moreover, there are some outliers in age of dogs out there just as there are some really old humans. 24

years old dog may be a typo but there are some at 20 which is believable, and doesn't make the 24 years seem absurd.

c) Color

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

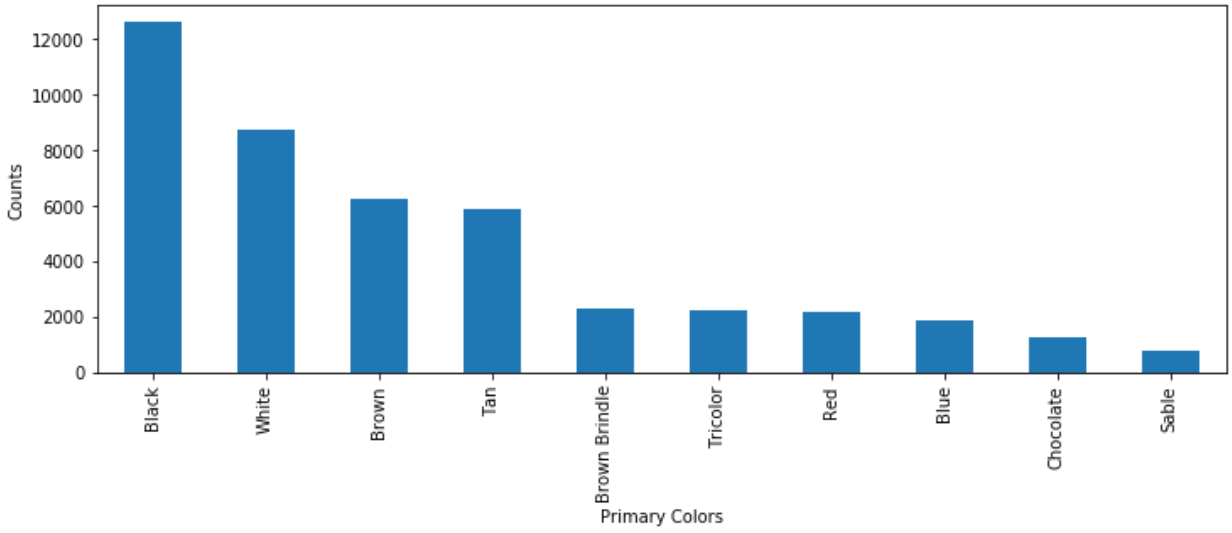

Figure 6: Top 10 Primary Breeds in the intake data.

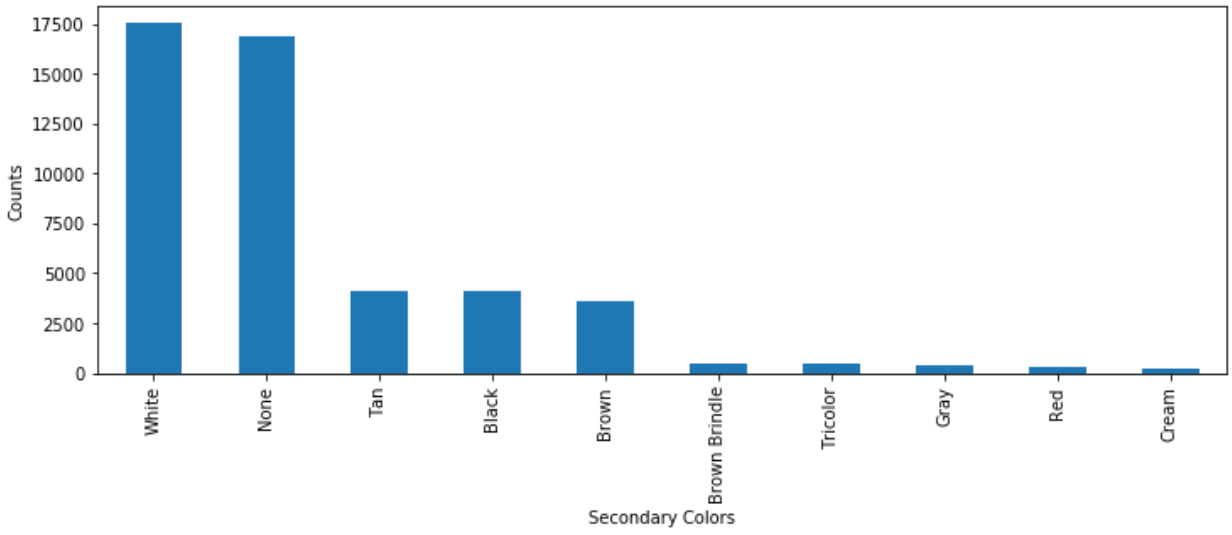

Figure 7: Top 10 secondary breeds in intakes.

It looks like the top colors in general are black and white. The top secondary color is white or none with trace amounts of other colors.

d) Gender

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

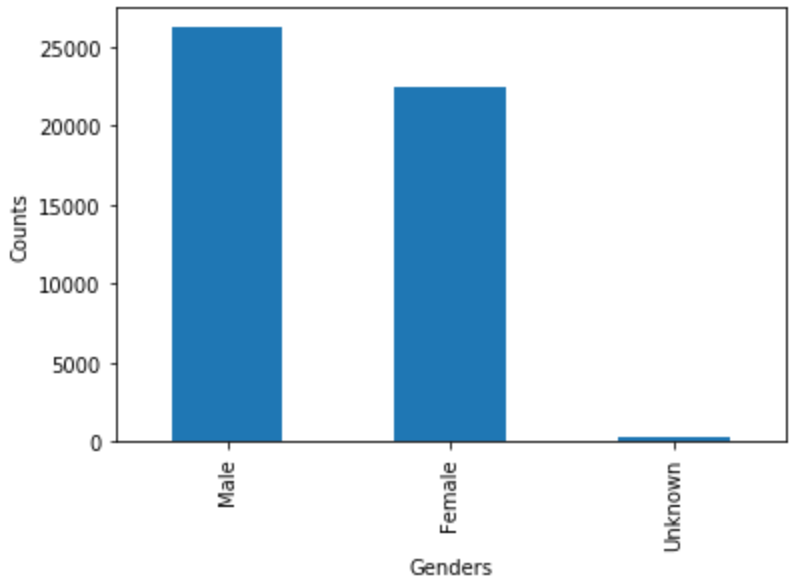

Figure 8: The distribution of Genders for intakes.

There are more intakes that are males than females but roughly 50/50. There are a small amount of unknowns probably missing data. I chose to leave the Unknowns as Unknowns.

e) Fixed

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

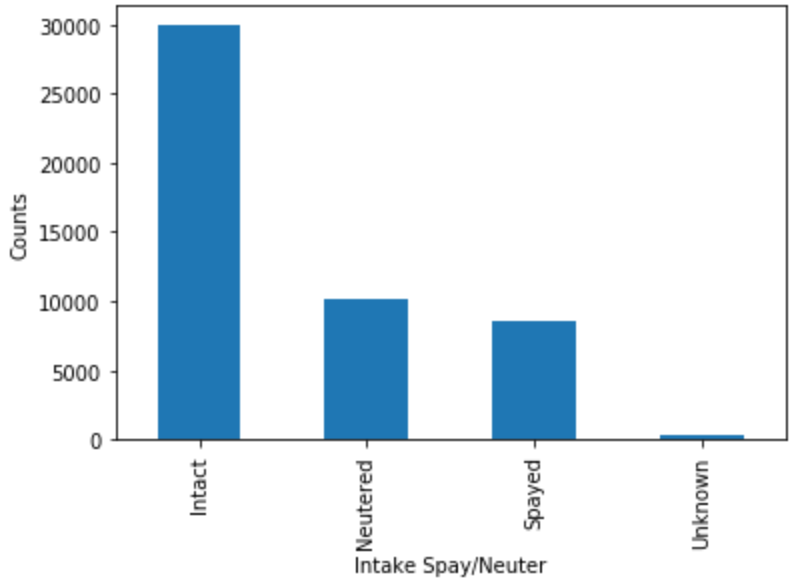

Figure 9: Fixed intake distributions.

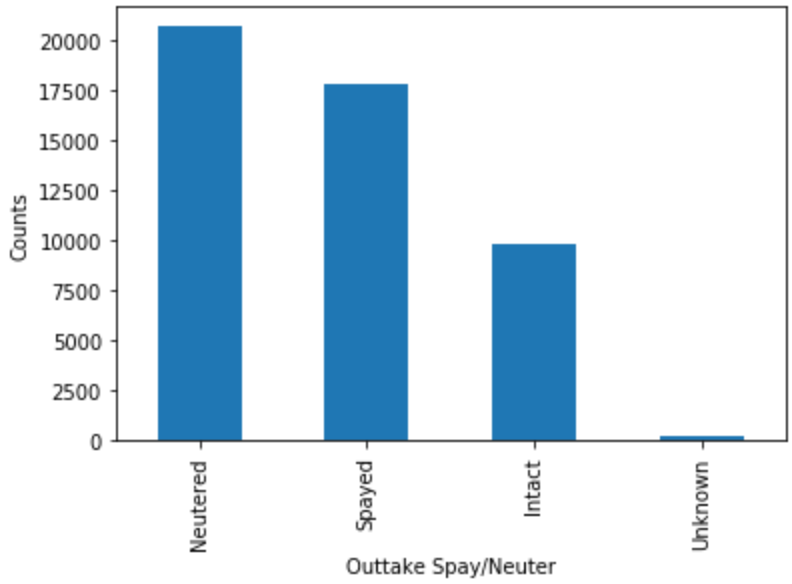

Figure 10: Fixed outtake distributions.

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.

f) Intake Condition

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

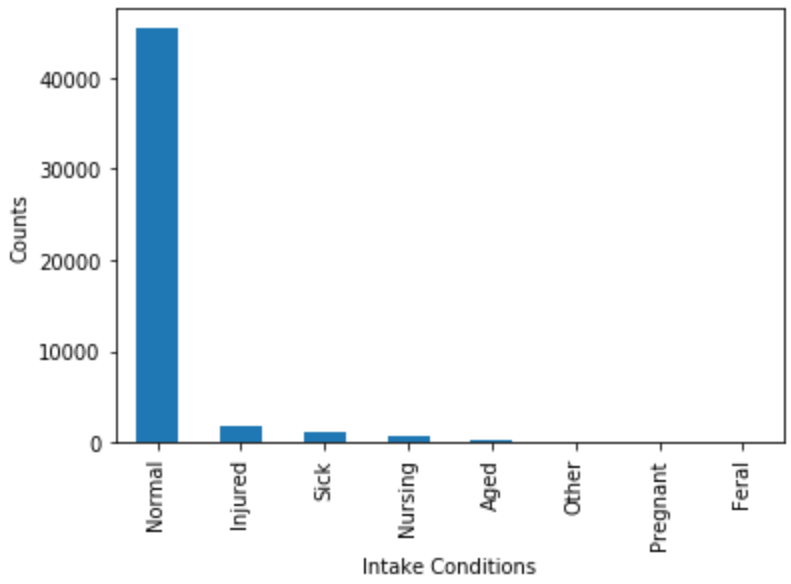

Figure 11: Distribution of the Intake Conditions.

The vast majority of the intakes that come in are in normal condition. This would make any statistical analysis base on intake condition skewed.

g) Intake Type

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

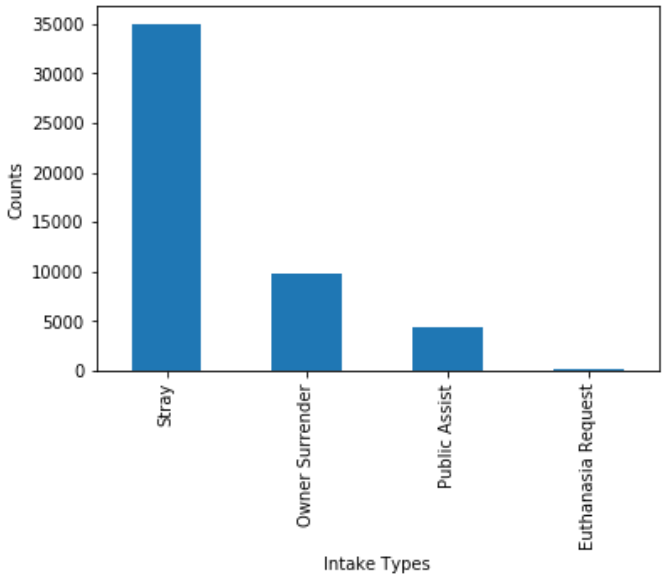

Figure 12: Distribution of Intake Types.

Upon inspection of the intake types I decided to look into the percent of Owner Surrenders that ended up as Return 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 Euthanasias were requested. 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.

3) Quick look at Intake Feature Correlations to Adoptions

Let's keep with the most interesting features that aren't heavily weighted to one feature grouping. Thus, I chose to drop intake_type and intake_condition. We will look at the percent adoptions multiplied by the total number of adoptions for each category grouping. For example, percent of Pitbulls from intake that are adopted. This will hopefully give an idea of the efficiency and quantity of the adoptions in each category grouping. We will also look at median time to adoption for the different category groupings.

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)

a) Breeds

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

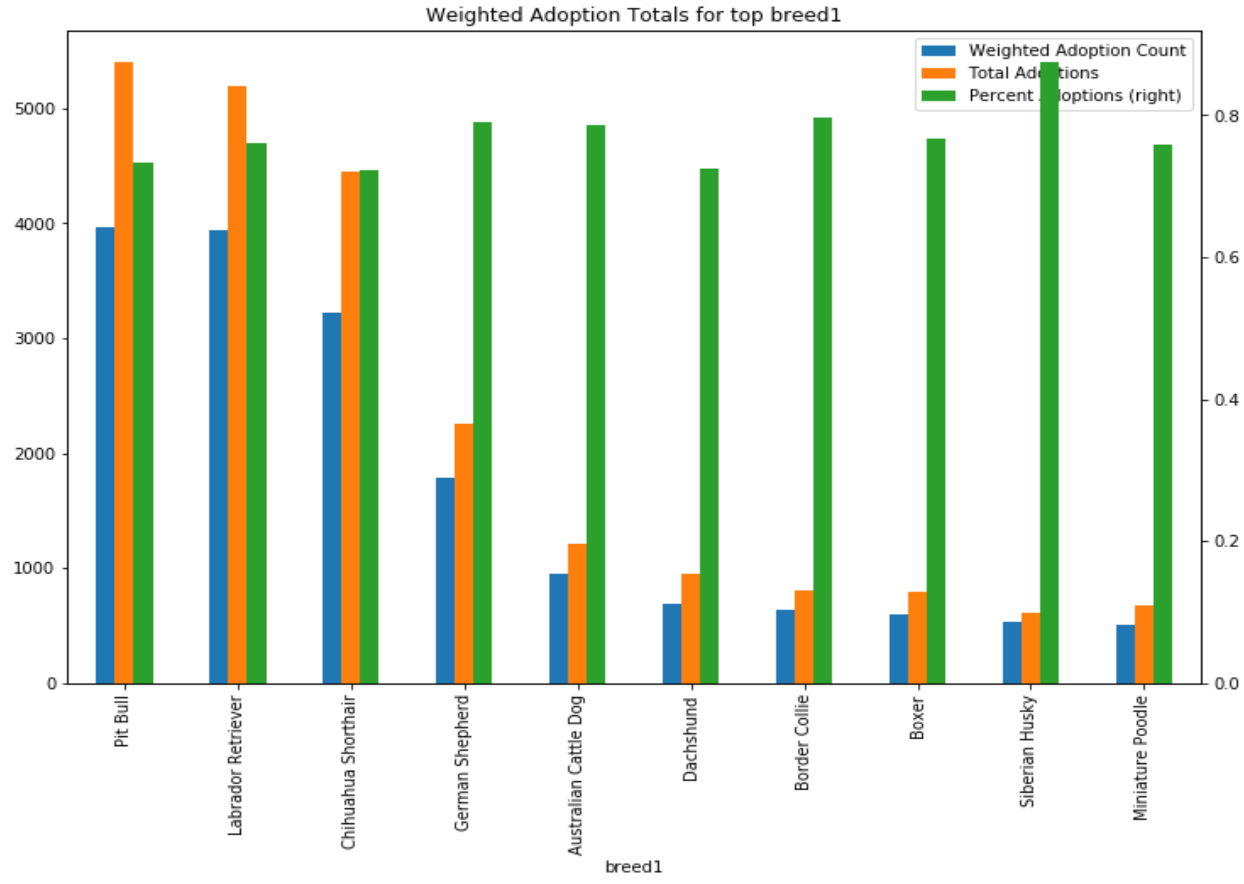

Figure 13: Primary breed adoptions distributions.

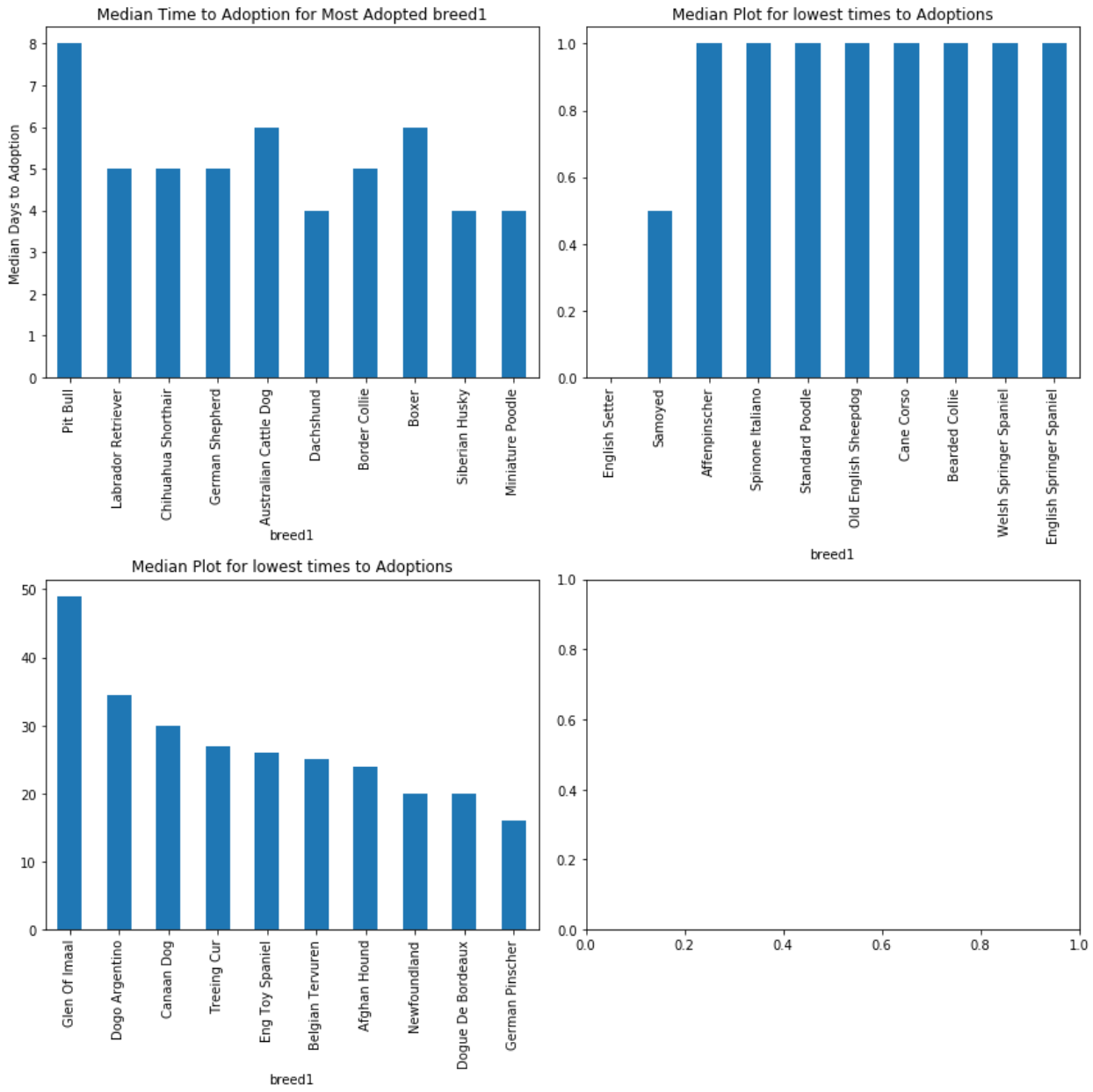

Figure 14: Primary breed time to adoption information.

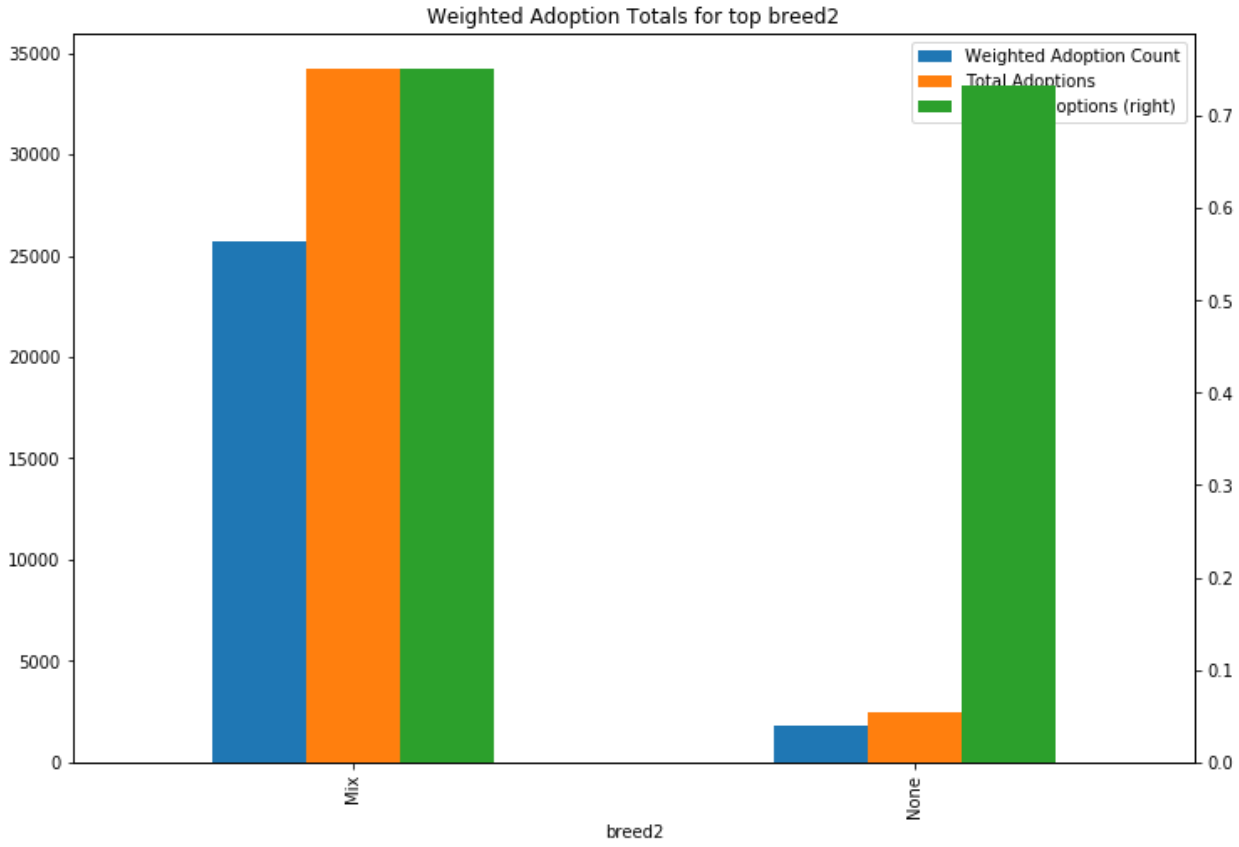

Figure 15: The percent adoptions for secondary breed.

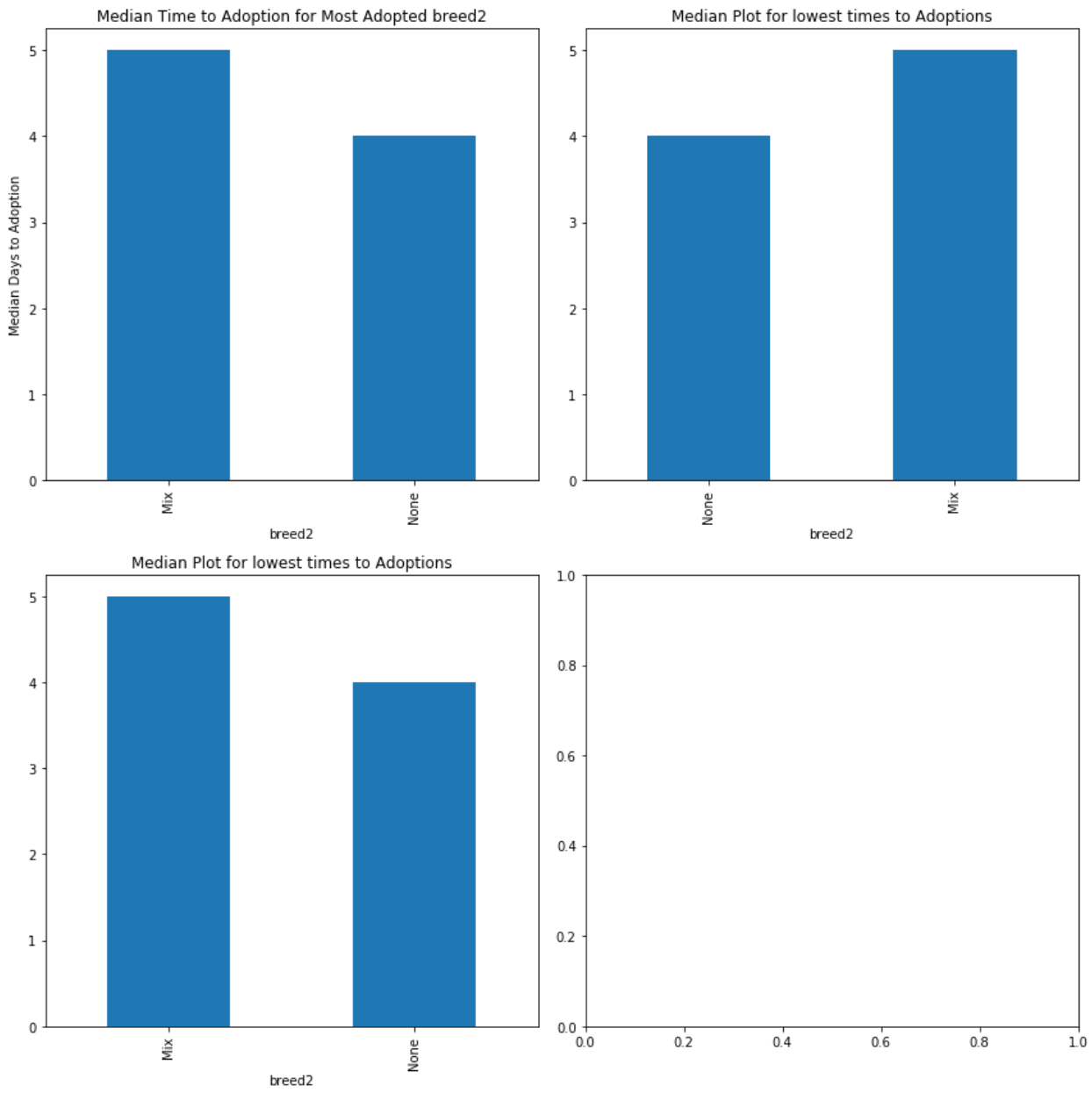

Figure 16: The time to adoption distributions for secondary breed.

It seems that the most popular adopted primary breeds are Labradors and Pitbulls, but these are also the most numerous primary breeds for intakes as well. There efficiency isn't too bad either, with about 72% of the pitbulls leaving the shelter under favorable means, and about 75% for the labradors.

All the dogs in the top 10 weighted adoption count 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 pitbull, labs, german shephards, daschund, 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 pitbulls 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, which makes sense if the majority of the intake type are strays, roaming free to breed with whomever, whenever.

b) Age

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.

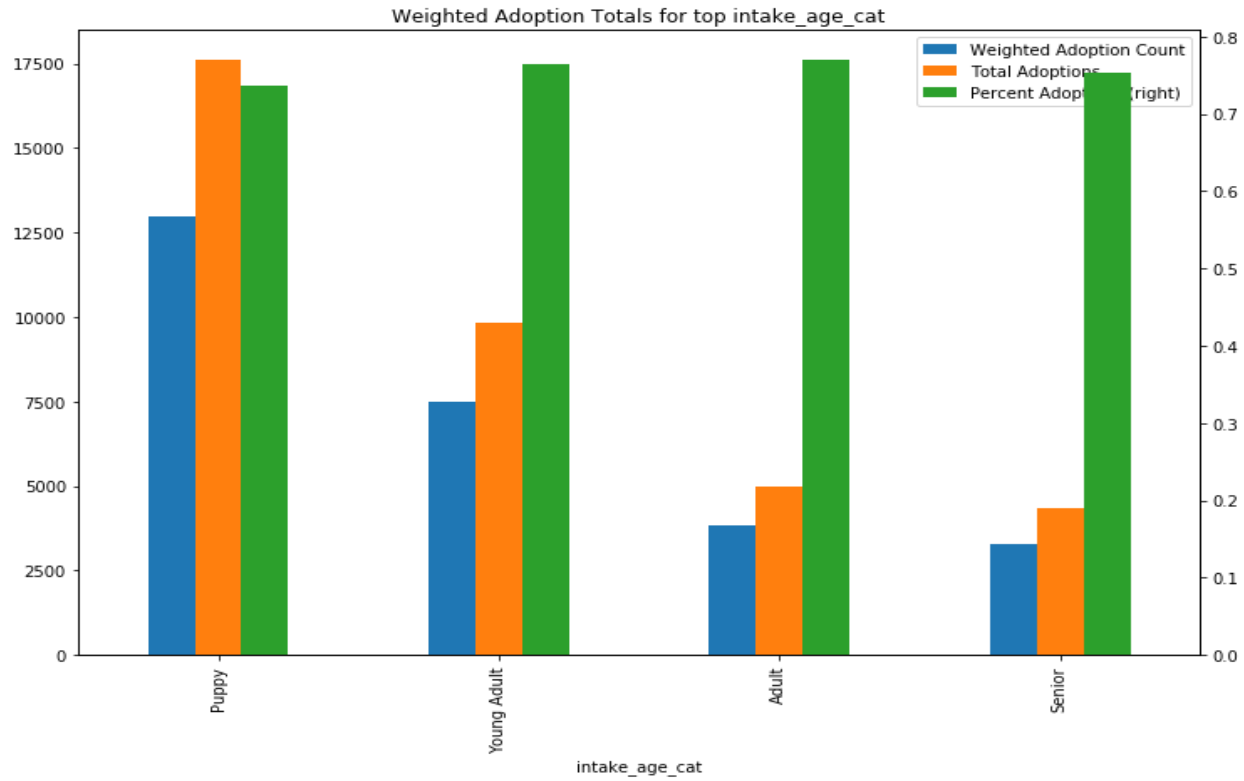

Figure 17: The adoption percentages for intake ages.

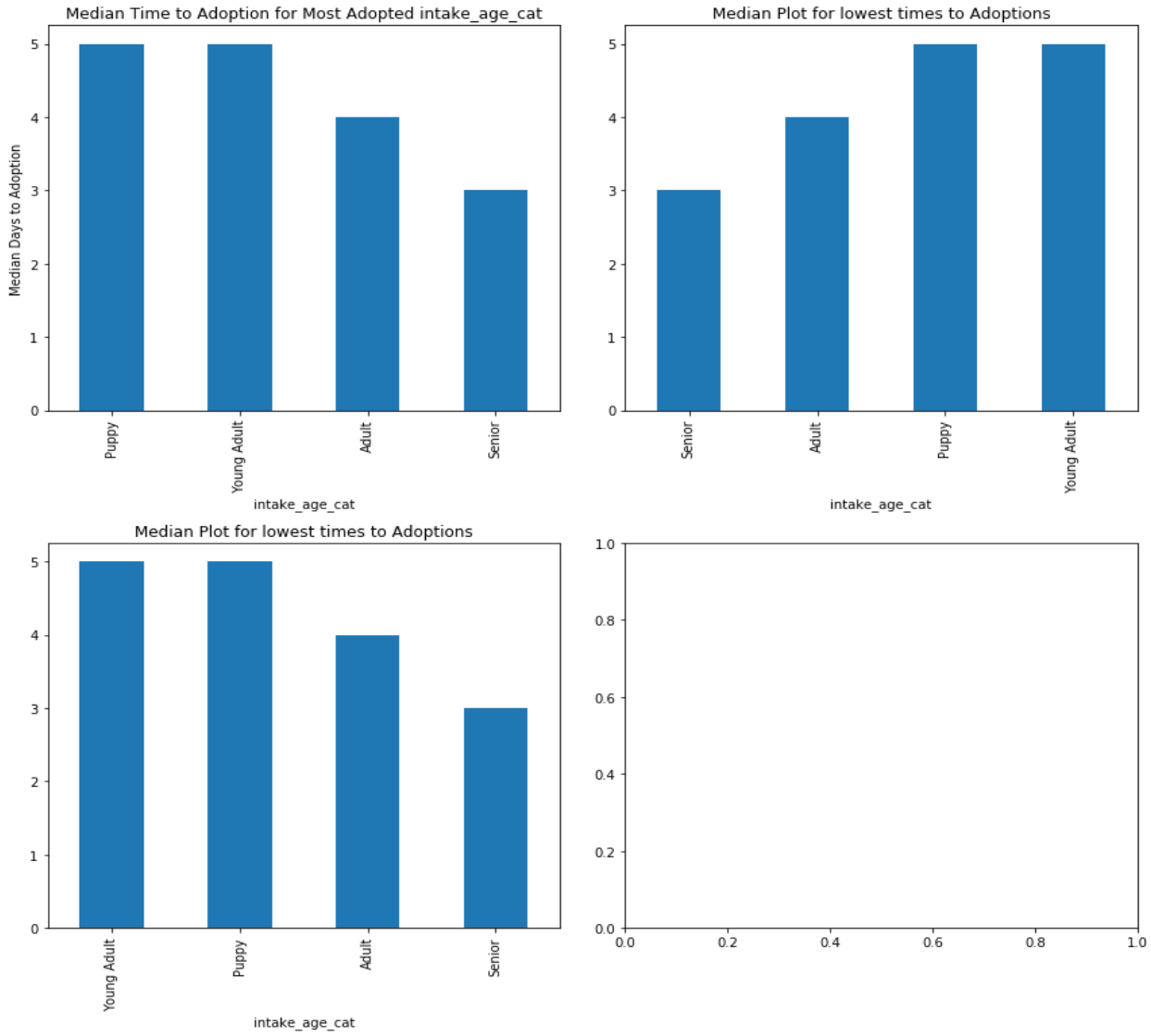

Figure 18: The time to adoption for intake ages.

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.

c) Color

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

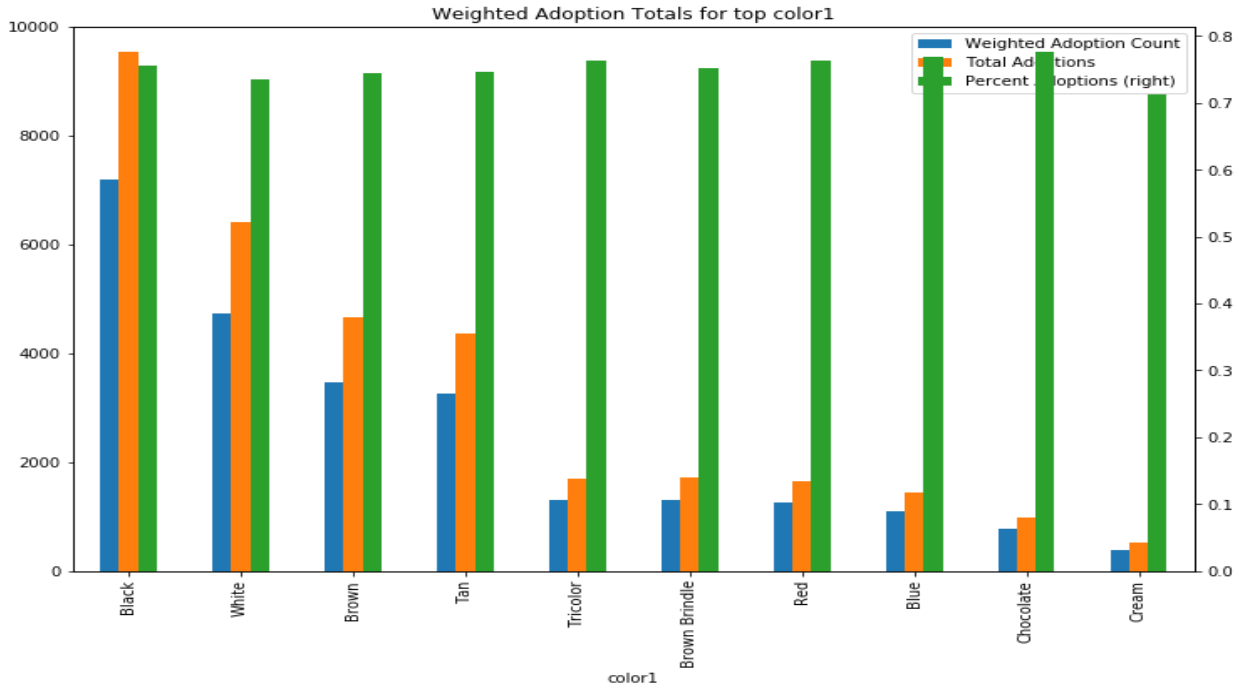

Figure 19: Primary color adoption statistics.

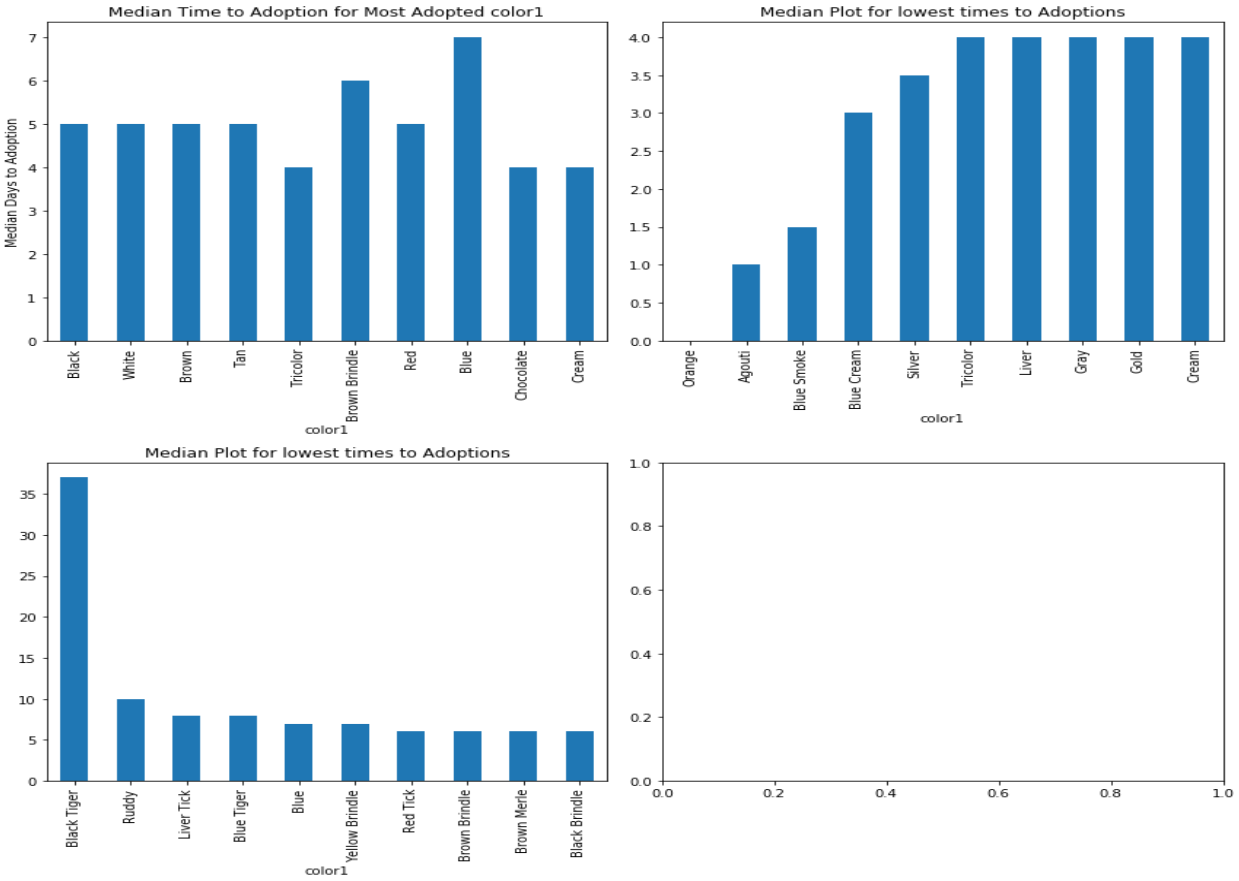

Figure 20: Time to adoption statistics for primary color.

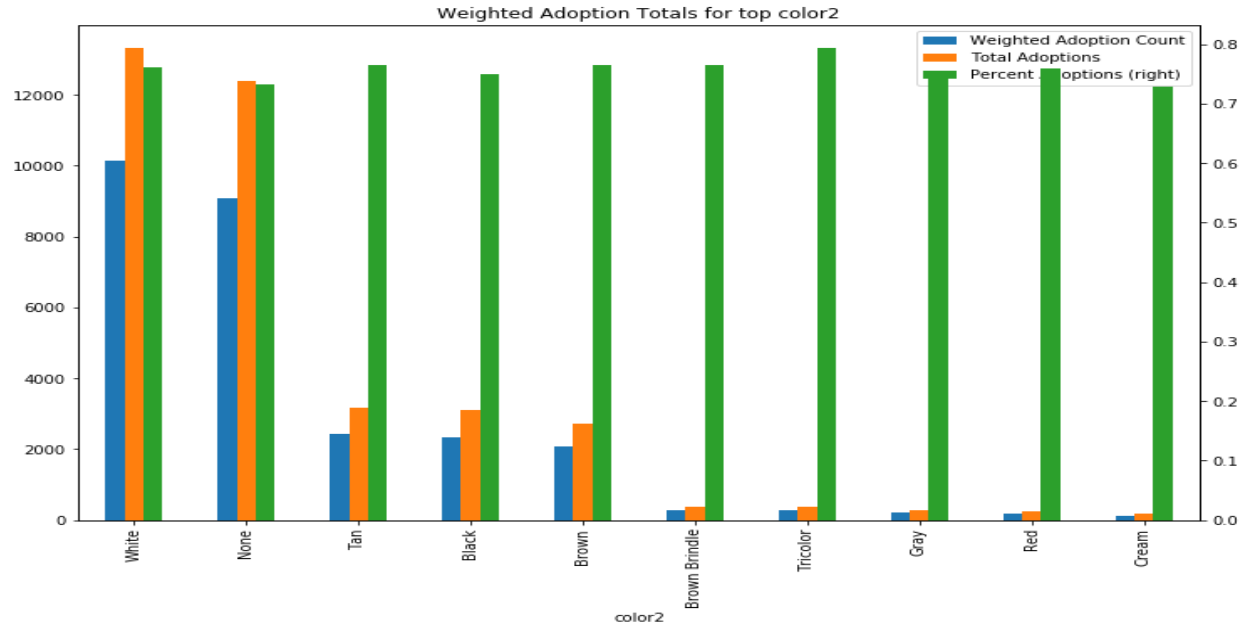

Figure 21: Secondary color adoption statistics.

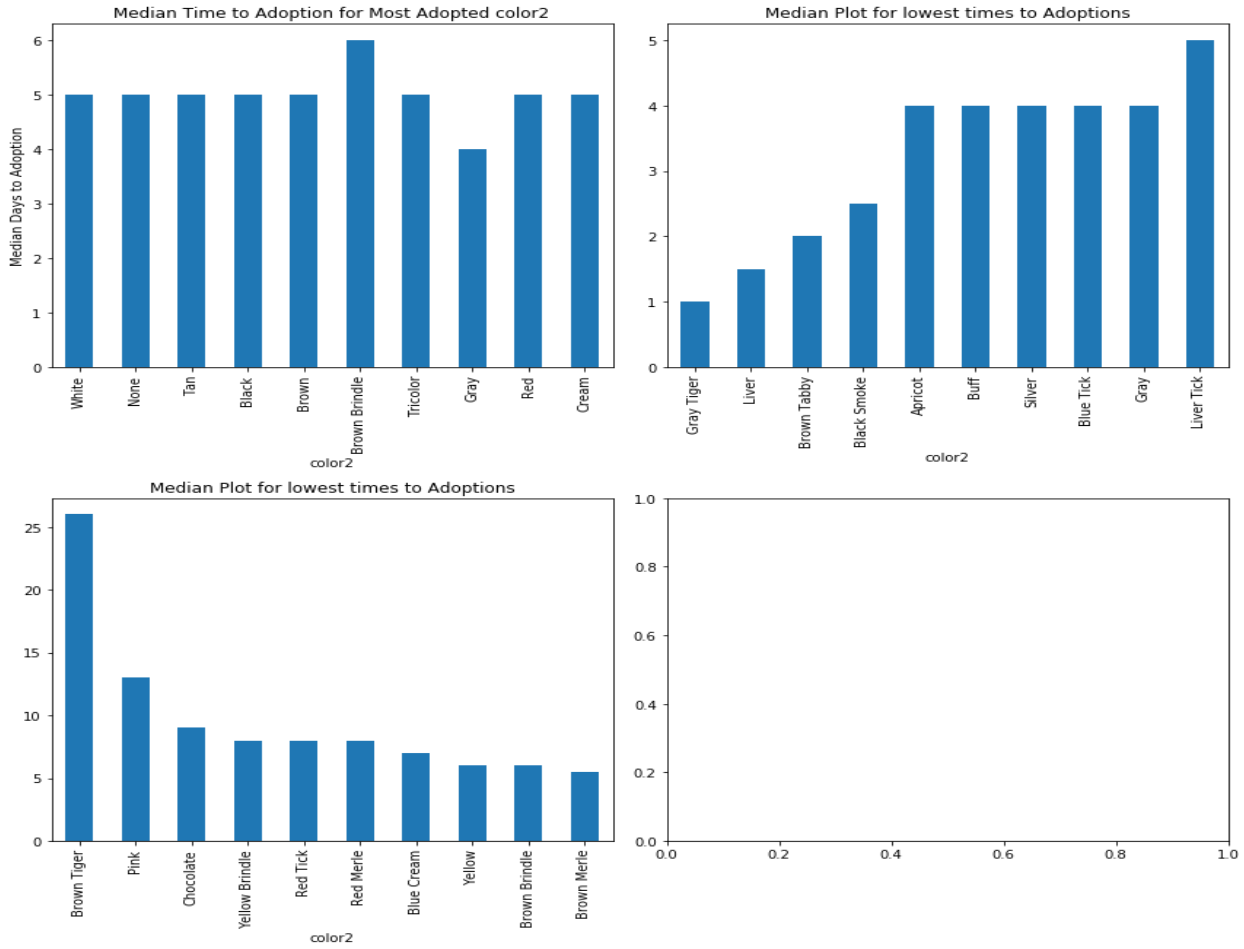

Figure 22: Time to adoption statistics for secondary color.

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.

d) Gender

The gender comparison is summarized in Figures 23-24.

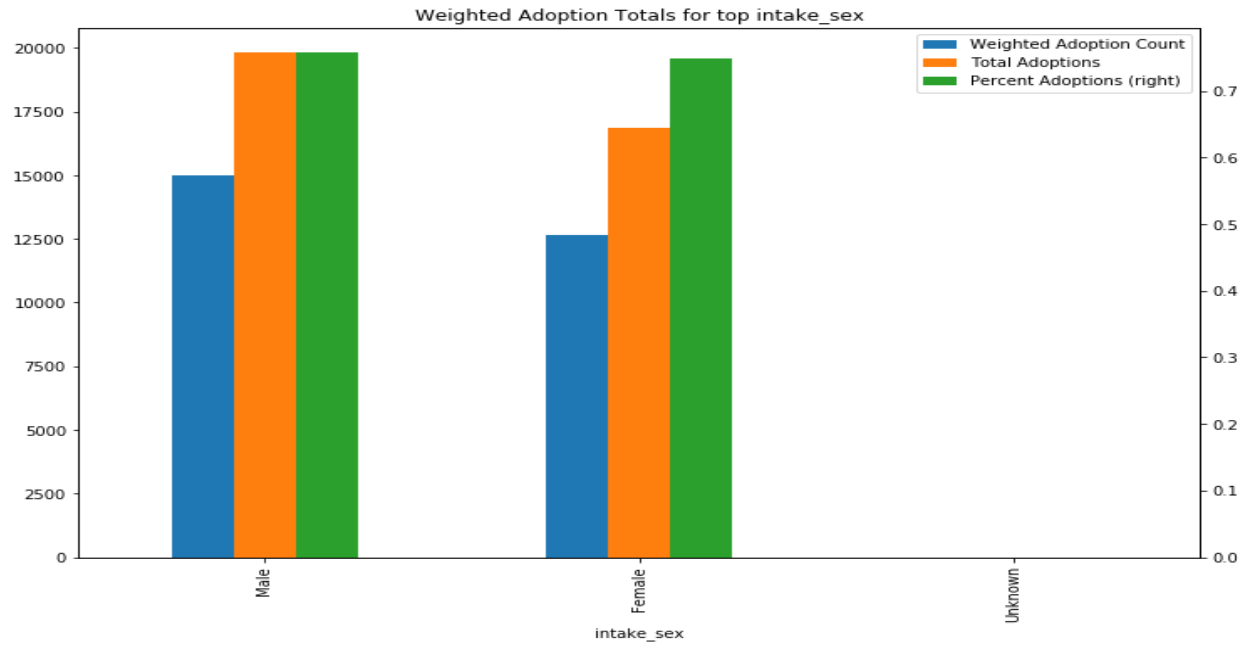

Figure 23: Adoption statistics for gender.

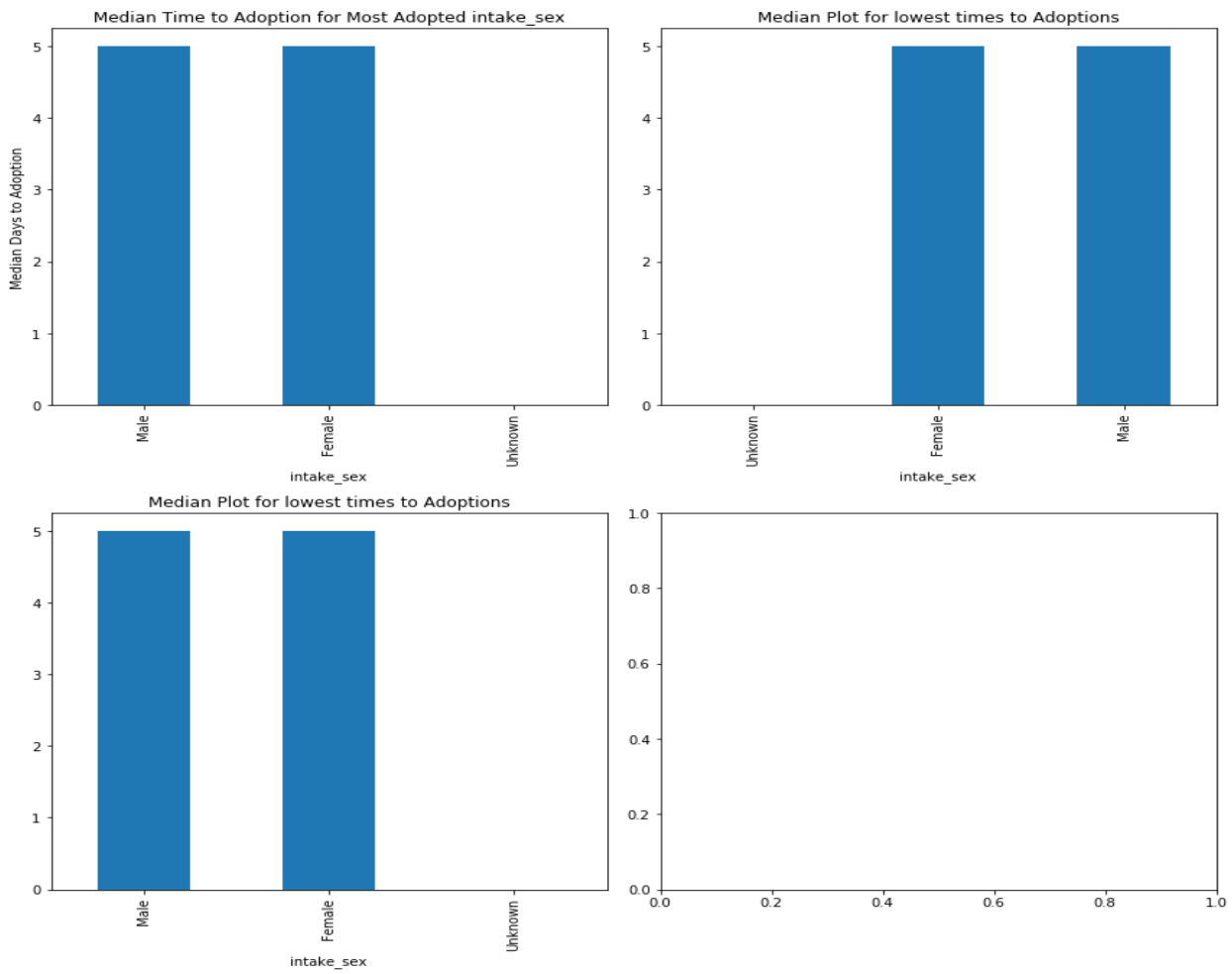

Figure 24: Time to adoption statistics for gender.

The gender doesn't seem to have an appreciable impact either based on the exact same results for both males and females.

e) Fixed

The fixed comparison is summarized in Figures 25 and 26.

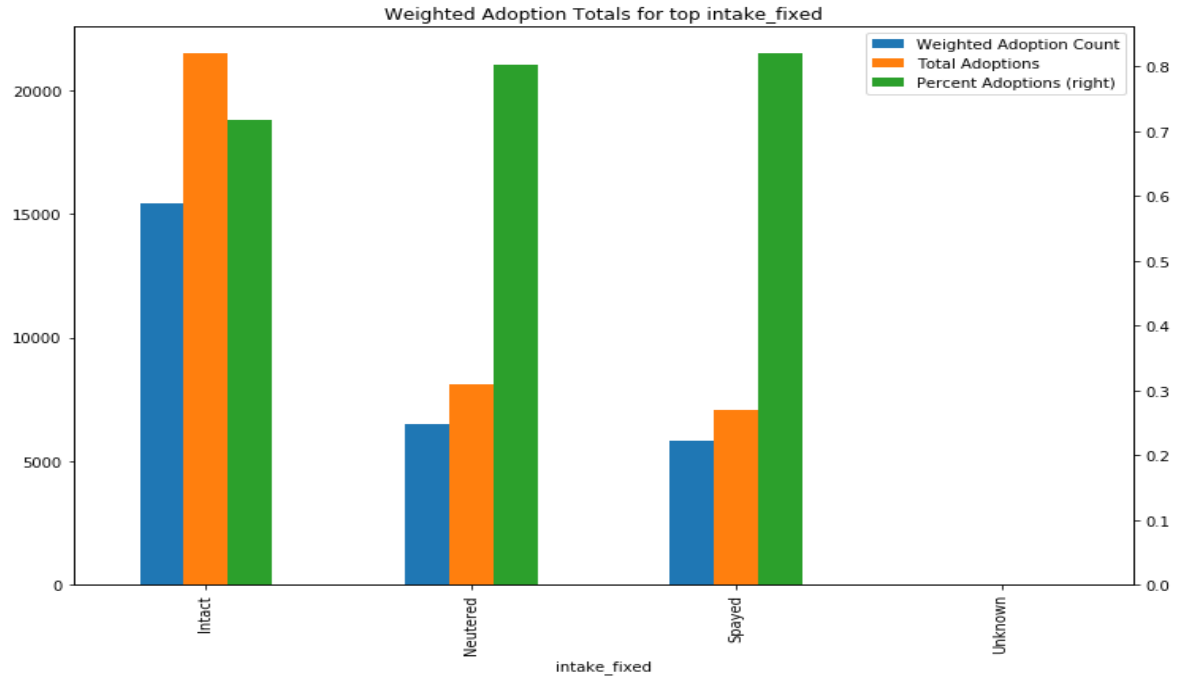

Figure 25: Fixed adoption statistics.

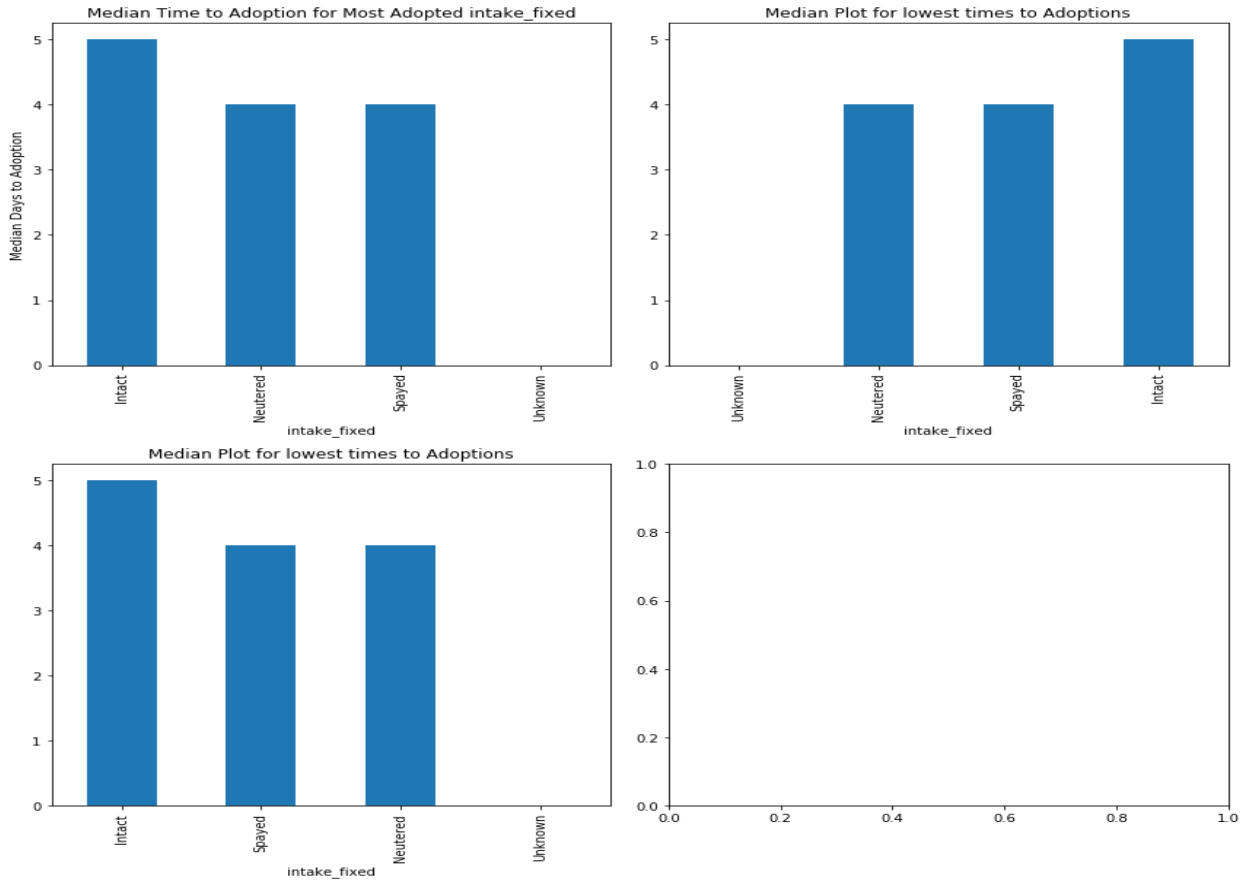

Figure 26: Time to adoption statistics for fixed.

It seems there may be a slight favor for the fixed dogs to get adopted faster and more often then the intact intakes. The median time to adoption is shorter for fixed dogs. The percent adopted is slightly higher for the fixed dogs. This apparent time reduction to adoption could be from the fact that the recorded adopted factors in surgery time as most dogs leave the shelter fixed even when they came in intact.

f) Seasons

The seasons analysis is outlined in Figures 27 and 28.

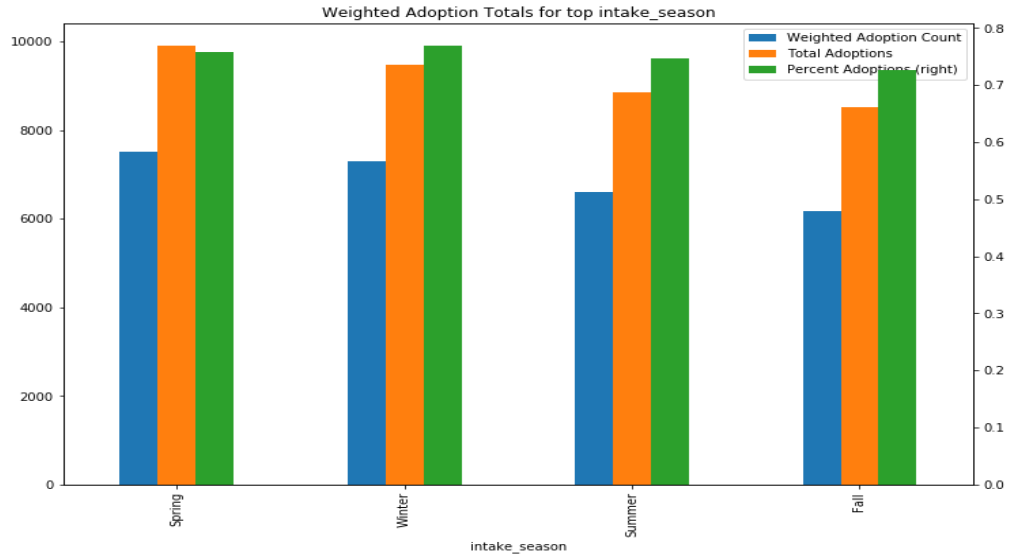

Figure 27: Seasons adoption statistics.

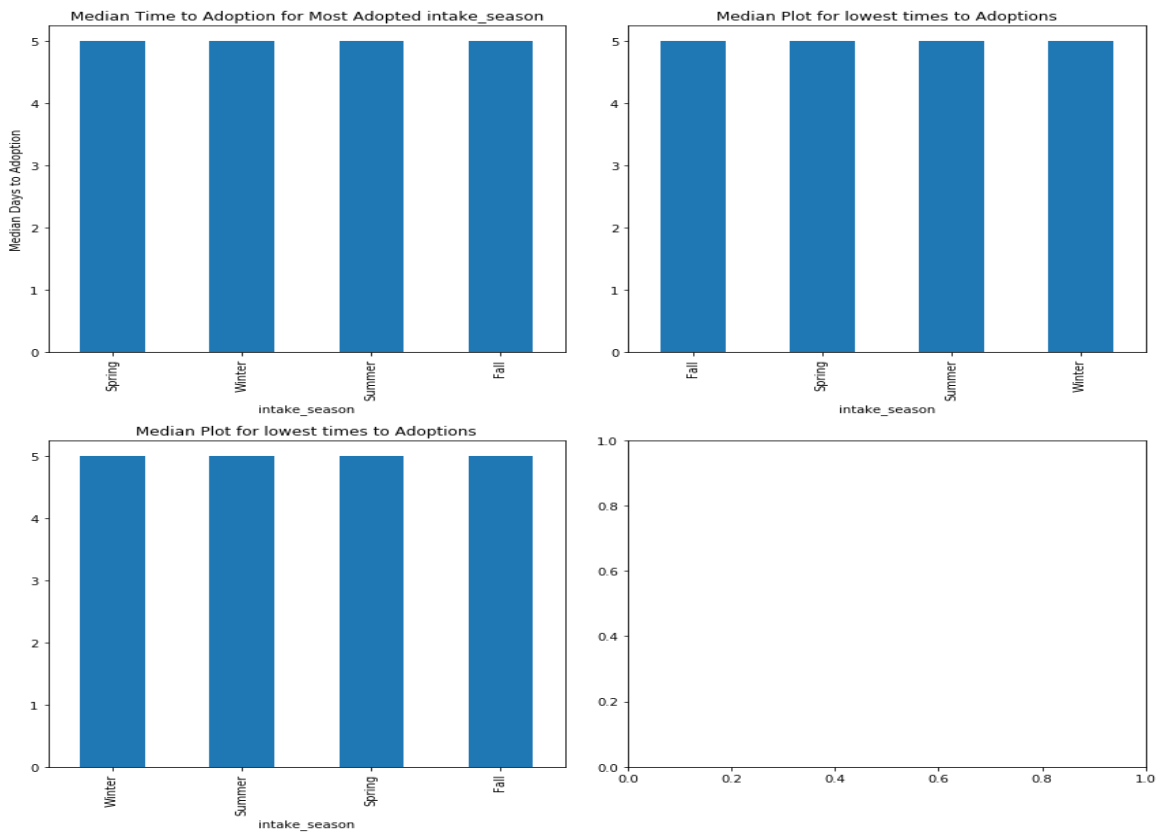

Figure 28: Time to adoption statistics for seasons.

As shown in the time series plot at the beginning the time of year doesn't really seem to have an effect on the adoption percent or the median time to adoption. However, the total number of adoptions seems to be affected by the time of year.

4) Look into Outcome Types Correlations to Intake Features

We will break the outcome types into Main Groups.

- Adoptions (Adoptions, Return to Owner, Rto-Adopt)
- Transfers (Transfer)
- Deaths (Died, Euthanasia, Disposal, Missing, Unknown)

We will also look at the following intake features:

- a) Breeds
  Primary (Top 10)
- b) Gender (Comparison)

a) Breeds

A look into the top 10 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 get transfered a lot to other shelters or organizations. Also, Chow Chow are on the top 10 deaths list.
There also seems to be few deaths in the shelter which is good.

b) Gender

Looking into the gender by outcome types revealed that it seems to be evenly split in all outcome types.

5) Special Case Study: 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", but as shown earlier a low percentage of owner surrenders end up back as return to owners. A risk worth taking? Also, this would eat up a lot of valuable shelter resources and potentially cause lots of unnecessary heartbreak.

6) A little Time Series Analysis

Let’s take a look at the trends of Adoptions over time for the Top 5 breeds Identified earlier.

- Pit Bull
- Labrador Retriever
- Chihuahua Shorthair
- German Shepherd
- Australian Cattle Dog

Breeds seem to be the most interesting statistically to look at by first inspection.

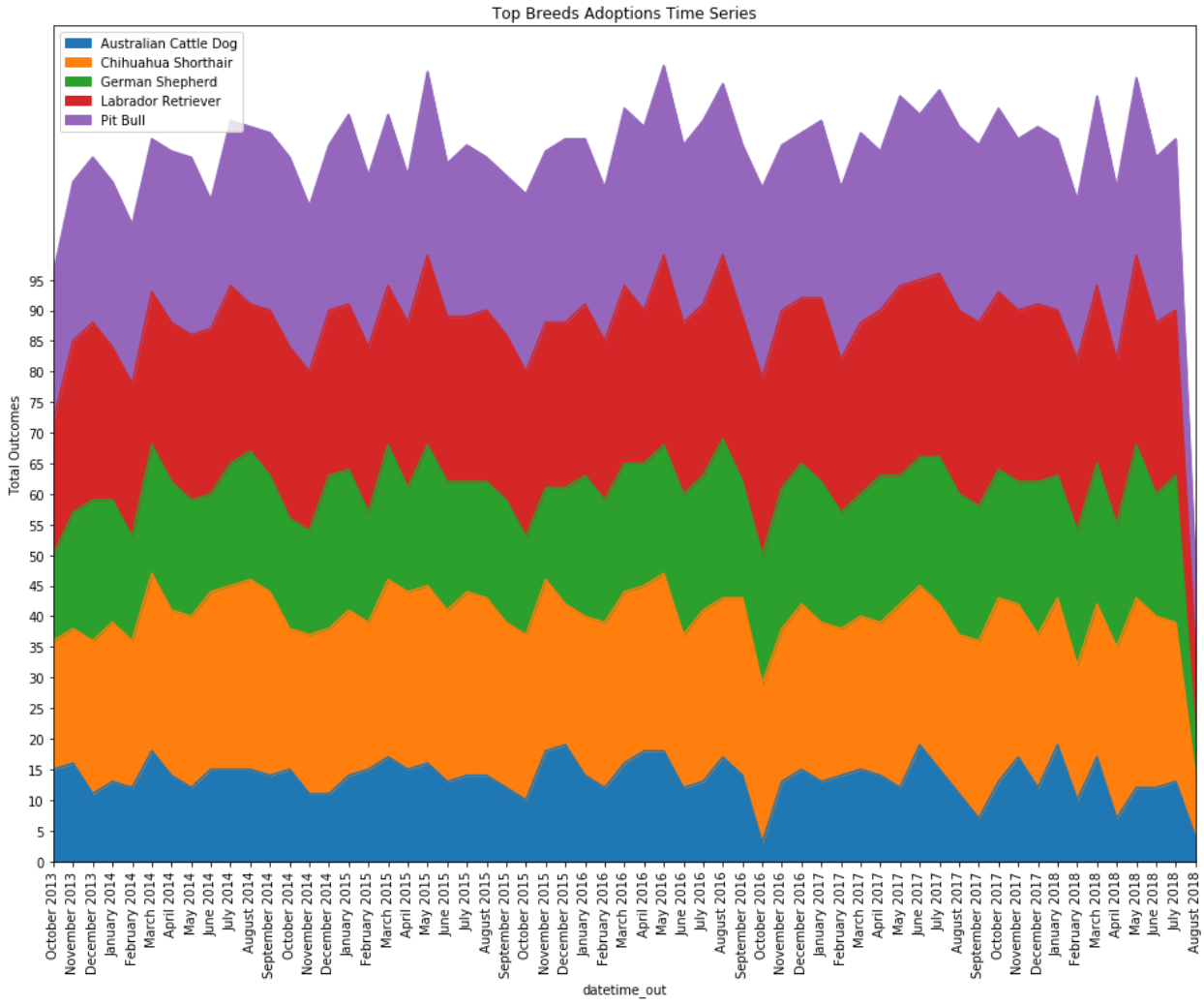

Figure 29: Time series plot of the top 5 primary dog breeds.

Interestingly it looks like they may run adoption events or something in the fall and spring. A lot of the peaks seem cyclical and line up well with fall and spring months. Contrastingly, this could be because people feel like adopting more before and after winter.

INITIAL STATISTICAL ANALYSIS FINDINGS

Notes: Adoptability will be calculated as the percent of Adoptions in All other shelter outcomes.
Notes: I will employ hacker statistics for all testing as they are 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 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 (as the time to adoptions are loosely exponentially distributed). 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.

The quetions/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.
- Test significance level is $\alpha = 0.05$.
- Null hypothesis is that there isn't a difference in distributions of adoptions or time to adoption.
- **Result: 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.**

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.
- Test significance level is $\alpha = 0.05$.
- Null hypothesis is that there isn't a difference in distributions of adoptions or time to adoption
- **Result: 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.**

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.
- Test significance level is $\alpha\alpha = 0.05$.
- Null hypothesis is that there isn't a difference in distributions of adoptions or time to adoption.
- **Result: The test for adoptability was significant so 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.

4. Does the populous 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.
- Test significance level is $\alpha\alpha$ = 0.05.
- Null hypothesis is that there isn't a difference in distributions of adoptions or time to adoption.
- **Result: 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. The null hypothesis can be rejected. The empirical difference of 11 days in expected value suggests that the less populous breeds are adopted sooner.**

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

- To test this, the groups will be the breeds are separated into the AKC groups and tested one vs. all for each AKC group.
- Test significance level is $\alpha\alpha$ = 0.05.
- Null hypothesis is that there isn't a difference in distributions of adoptions or time to adoption.
- **Results: 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, the ones such as Toy with a significant Adoption Ratio value suggesting their adoption ratio is lower compared to the others.**
- **Also, the Terrier and Misc groups seem to take longer to adopt than the other groups.**

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)**

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.
- Test significance level is $\alpha\alpha = 0.05$.
- Null hypothesis is that there isn't a correlation.
- Test setup:
    - To simulate the null hypothesis, we need to:
    - Calculate the empirical correlation coefficient.
    - Permute one of the variables and keep one steady.
    - Then, then calculate the correlation coefficient(test statistic).
    - Repeat this many times until a distribution of correlation of coefficients are generated.
    - The p-value will be the count of instances where the permuted correlation coefficient is at least as extreme as the empirical correlation coefficient.
- **Result: The empirical correlation coefficient was 0.03. This isn't 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 as the puppy vs non puppy suggested as well. However, looking at the scatterplot it seems more random at the left side of the graph, then it seems like the really old dogs have shorter times to adoptions but this is based on fewer samples.**

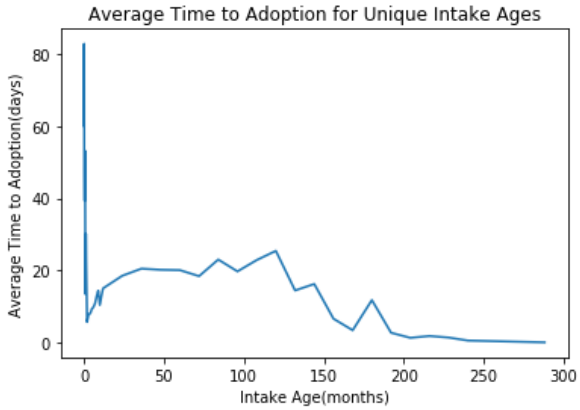

Figure 30: Time to adoption (averaged for each unique age) in dasy vs intake age in months.

7. Are adoptions affected by seasons?
    - To test this, the groups will be the seasons and will be tested one vs. all for each season.
    - Test significance level is $\alpha = 0.05$.
    - Null hypothesis is that there isn't a difference in distributions of adoptions or time to adoption.
    - **Result: The results are all significant results 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.**
      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.

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**

MACHINE LEARNING

Under construction.