Adoptable Animals

Milestone Report

8th June 2018

Christopher Seth Hill

OVERVIEW

Many animals wind up in shelters for various reasons. Some are adopted, some are not, and for the very unfortunate ones, die in shelters. By looking at the 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. Moreover, insight into what type of dogs are left behind can be gained. This will hopefully serve as a kind of educational insight into the world of dog adoptions.

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

INTERESTED PARTIES

Shelters, people, and dogs anywhere could benefit from this information. It could aid in optimizing the amount of time spent and effort spent in advertising certain dogs for adoption by the shelter. This in turn could help decrease the turnaround time on outtakes, and aid in finding homes for less fortunate animals. This would lead to better efficiency and allotment of resources on the shelter's part. Moreover, potential adopters could be educated on which types of dogs may need their help more than others.

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, 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.

I started with the intake data. A quick inspection of the dogs only intake data frame showed that the animals were all given a unique animal_id for distinguishment. After looking at the value counts of the unique animal_ids, it was clear that there were multiple intakes for some animals. Further inspection of the various features in the data frame led to the discovery of other potential issues that needed to be fixed in order to make the data tidy and 100% useful. The summary of issues found upon initial inspection of the data frame 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

There are 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 the “intake” column 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’s visits to the shelter. This 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.

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 time period indicator. Then, I created and matched regular expressions for the different time periods to each intake age in the data frame. This gave me the time 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 dataframe location entries.

Sex upon intake is actually 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 breed1 and breed2 features 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 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 similar issues to the intake data. Also, a quick inspection of the data frame showed 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 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 data 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 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 duplicate outtake date column.

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

1) The age upon intake isn't machine digestible.
2) There are nulls, No Name, and values with asterisks 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 outtakes per animal.

Issues 1, 2, 3, 4, and 7 were fixed in the same manner as described above in wrangling 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” column 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 animal_id and outtake numbers. This means that the data frames would not misalign upon merging. 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 are not a lot. Thus, I let the merge naturally exclude these entries.

After merging, a quick look at the joined data frame info showed that the data frame length was the same as the intake data frame length. There were 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 in order to input the data into Machine Learning Models. These steps will be briefly explained in the Machine Learning section. Also, as is discussed in the initial findings section, other columns were created in the joined data frame to aid the Exploratory Data Analysis.

More detailed information and the code used to wrangle the data can be found in the ipython notebook found here.

OTHER POTENTIAL DATASETS

There are some other data sources that could be utilized to help the statistical tests to 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. Data could also be wrangled 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 taken from pictures of the dogs and website descriptions from online adoption postings.

INITIAL EXPLORATORY DATA ANALYSIS FINDINGS

The main goal is to find out what if anything helps increase animal adoptions based on the basic intake data recorded by most animal shelters. As such, we can also get a sense of the intakes and outtakes demographics. This can help in understanding if there is an imbalance in the features of the dogs that get adopted versus those that do not. In this way, we can better understand and educate people on which dogs may need more help and more love.

The main important findings from the exploratory data analysis give important insights into the world of the Austin Animal Shelter. We will start with the broader scope and narrow down slightly into some of the details.

To start, we will look at the time series of outcome types for the shelter shown in Figure 1.

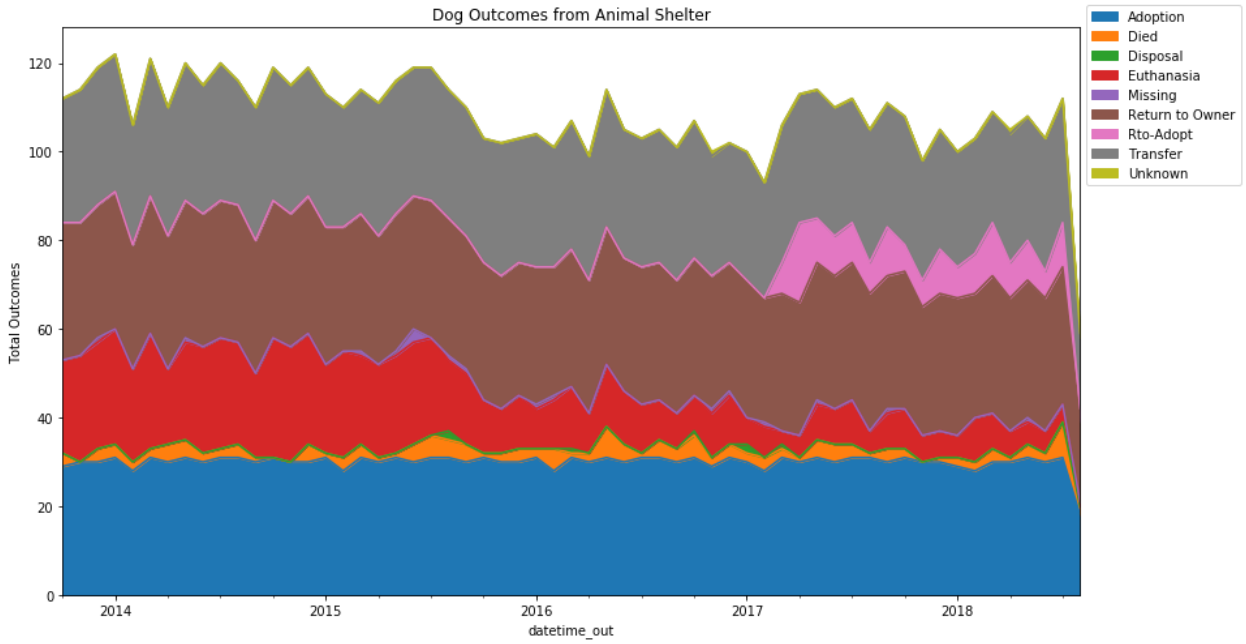

Figure 1: Time series of outcome types.

These outcome types can be efficiently grouped into 3 larger groups outlined below:

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

Under the construct of these outcome types, I noticed that most of the shelter outcome types are adoptions accounting for roughly 60% of the shelter's outcomes. Next, transfers account for about 25% of the shelter's outcomes. Last, deaths account for the remaining outcomes. Moreover, deaths seem to be trending down over time with adoptions remaining fairly steady. It is good that the amount of deaths are dropping. However, the adoptions remain surprisingly steady. There seems to be a need for work to be done to increase adoptions. A deeper look into the break down of the dog's attributes and the correlations to outcome types will yield valuable information.

We will now look at the distribution of time to adoption values for the dogs that have been adopted from the shelter. This information is shown in Figure 2.

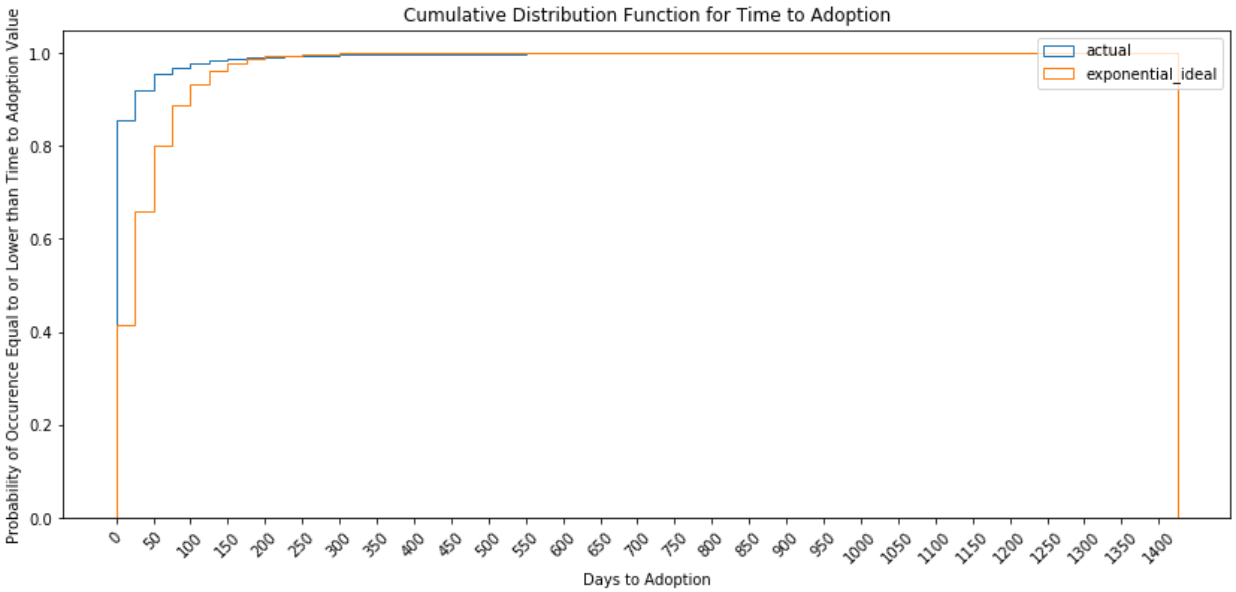

Figure 2: Cumulative distribution function for time to adoption values.

This tells us that roughly 80-85% of the dogs that are adopted are adopted in 25 days or less. This is great. It means that the dogs that do enter the shelter and get adopted generally don't stay long. It is nice to see quick turnarounds here as it hopefully means dogs and owners find each other that much sooner.

However, this plot does not give insight into the dogs that are less fortunate. We will now examine the dog features and how they match up or change between dogs that are currently in

the shelter, intakes for all time, and adoptions for all time. These will be displayed in the following figures.

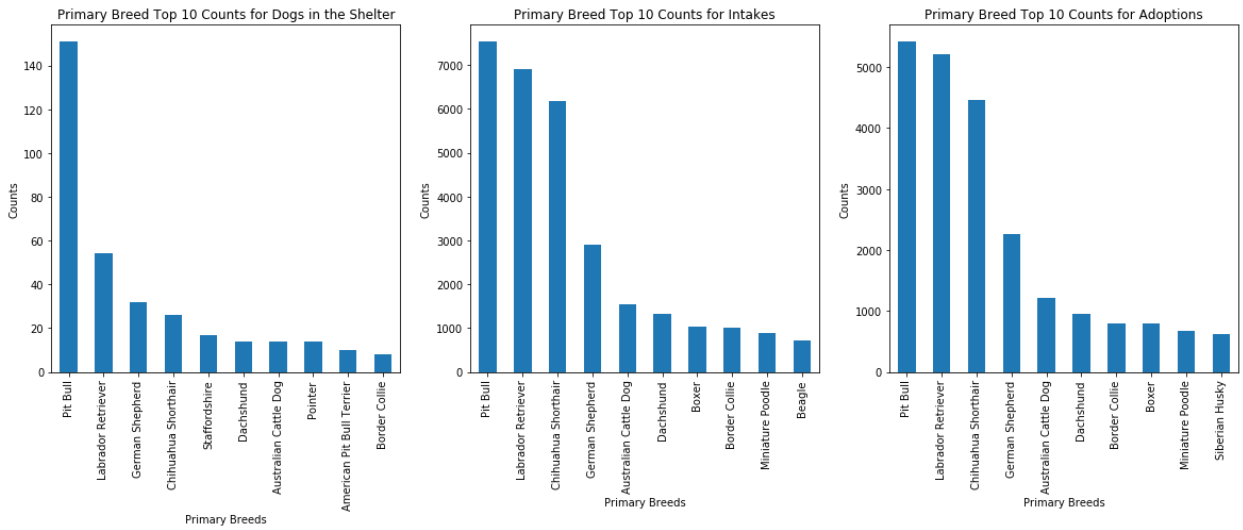

Figure 3: Top 10 primary breed comparisons for dogs currently in the shelter, intakes for all time, and adoptions for all time.

It seems that for the most part that the dogs that come in the shelter are moved through the shelter. Most of top 10 breed intakes are moved through the shelter as adoptions and not as other outcome types. However, approximately 2000 of each top primary breed are not adopted and not in the shelter. This suggests that there is still some room for improvement in getting dogs adopted. The fact that most of the same breeds are in the shelter currently suggests that there is not a overwhelming amount of one breed that gets "stuck" in the shelter or ends up overwhelmingly as the other outcome types. However, the Staffordshire terriers show up in the shelter top 10 but not in the intakes or the adoptions. This suggests that the shelter might keep some of this breed longer than others. Maybe, they are not adopted as often and end up as transfers or deaths. From my time in shelters (not the one for this data), I've noticed that quite a bit of the dogs are labeled as staffordshire. There is a lot negative stigmatism surrounding certain “aggressive” breeds. With proper handling and education, these dogs can be wonderful mild tempered dogs. More can be done to remove this stigmatism and help these dogs get wonderful loving owners and homes too.

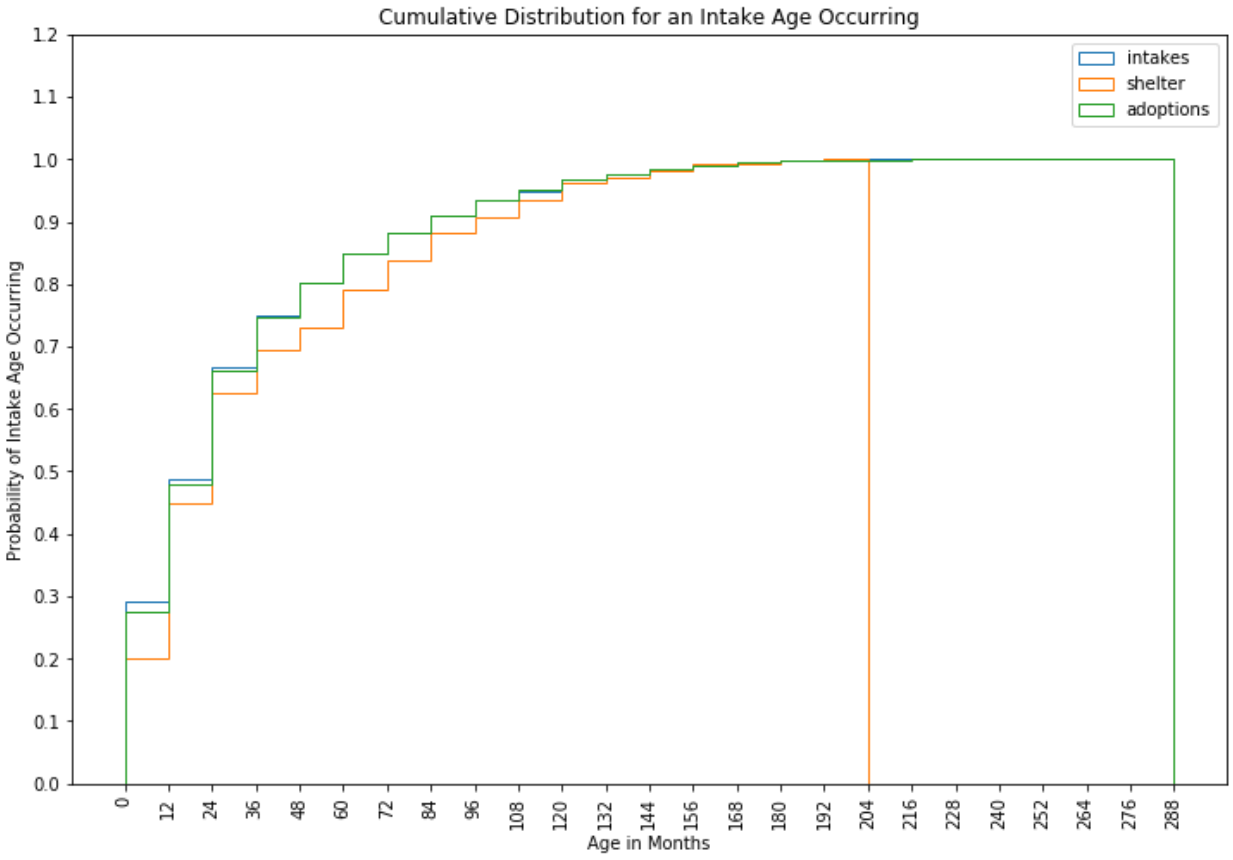

Figure 4: Intake age distributions for intakes, dogs in the shelter, and adoptions.

The age distribution for intakes and adoptions seem mostly the same. However, the intakes seem to be slightly older in general than the adoptions. This suggests that younger dogs are adopted more than older dogs. The age distribution for the dogs in the shelter currently is quite different from the intakes and the adoptions. The max age is much less than the max age for adoptions and intakes. However, the age at 80% is higher than the age at 80% for the adoptions and the intakes. This suggests once again that younger dogs are adopted more than older dogs.

Puppies are generally more desirable and adopted more often than older dogs. This is hard to counteract or improve. More effort can be made in fully explaining the pros and cons of both puppies and older dogs. Older dogs need loving homes just as much as puppies if not more than puppies.

In exploring the data, I noticed that there were a lot of dogs that would make return trips to the shelter. In digging deeper, I found some interesting insights.

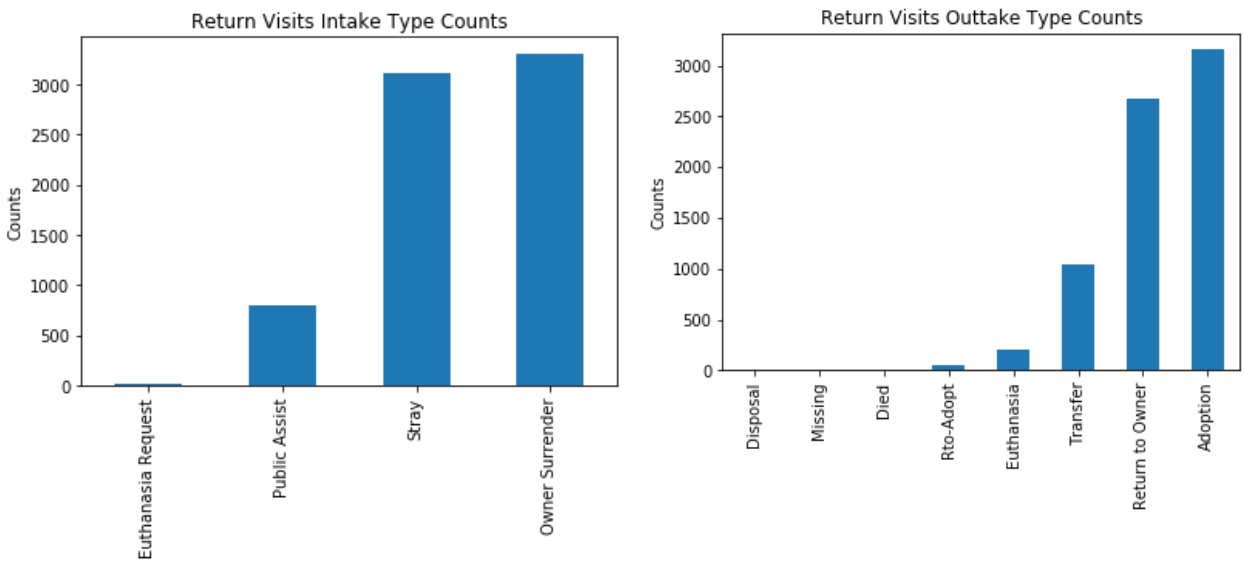

Figure 5: Comparison between return visits intake type and outcome type.

It seems that a fair amount of return visits are due to owners giving their pets up and trying to reclaim them. Maybe they were trying to utilize the shelter as "free pet boarding". As shown in the figures above, there is no guarantee that the dogs that are owner surrenders will end up as return to owner. Actually, it appears that more of the return visit outcomes are adoptions. Also, this eats up a lot of valuable shelter resources and is potentially a cause for lots of unnecessary heartbreak.

Also, a lot of return visit intakes are strays suggesting that the something better could be done to screen potential adopters or more emphasis placed on educating new potential pet owners before adopting. This might aid in reducing the number of “dog returns” to the shelter.

In summary, the main findings found while doing exploratory data analysis, are that older dogs and certain breeds deemed “aggressive” have a harder time becoming adopted. In addition, pet owners should not try to use the shelter as a means of free boarding. There is no guarantee that they will get their pets back.

These are only the main highlights found in the exploratory data analysis. The full analysis and ideas for further examination can be found in the ipython notebook located here.

INITIAL STATISTICAL ANALYSIS FINDINGS

Enough insights were gained to suggest there might be correlations between dog intake features, time to adoption, and chance of adoption. The main findings will be explored in this section. For full details regarding methodology and testing, please look here.

To understand the results, I will first define the result categories. Chance of adoption is calculated as the number of adoption outcomes divided by the total number of outcomes. Time to adoption is the time spent in the shelter. Several different features were tested to see if there were significant differences in chance of adoption or time to adoption for different groups of dogs. The significant results will be discussed below.

The first feature tested was age. I grouped all the dogs into two groups: puppies (less than or equal to 1 year old) and not puppies (greater than 1 year old). The results showed that there was an average difference in time to adoption of 28 days favoring quicker adoptions for puppies. This was the exact result guessed from EDA done earlier. It suggests that more attention needs to be given to older dogs so that all dogs have a fair chance at finding a forever home, even if forever is shorter for them.

Breed and breed popularity (number of a breed in outtakes) is significant in determining the chance of adoption and time to adoption for dogs. Breeds were divided into top 5 breeds by count and non top 5 breeds. The top 5 breeds showed a lower chance of adoption, but a lesser time to adoption. This suggests that the more unique or less popular breeds get adopted at higher percentage than the popular breeds. A look into the top 5 breeds done in the EDA shows that most are “aggressive” breeds. This can be balanced with education and time to create equal playing fields for adoptions.

Further intrigued by breed being a main factor in determining outcome type and time to adoption. I put all the dogs into their respective American Kennel Club breed groups using the dogs’ primary breeds. The results are summarized in the table below.

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

Table 2: AKC breed group statistical testing significant results.

The results show that indeed breed is a significant factor in determining the shelter outcomes. The Herding group is significantly more likely to have a higher adoption chance and be adopted sooner. The Hound and the Toy group are also more likely to be adopted sooner as well. Contrastingly, the Terrier group is less likely to be adopted sooner. With this information we can see in finer detail which breeds may not have the fairest chance at being adopted.

Next, I looked into age a little more, and tested whether time to adoption and intake age are correlated. The result is that there is a slight positive correlation between intake age and time to adoption. This suggests once again that older dogs will sit in the shelter longer before being adopted. The correlation coefficient was 0.03, which is almost no correlation. Thus, I decided to take a closer look at the data. The plot below shows the story in more detail.

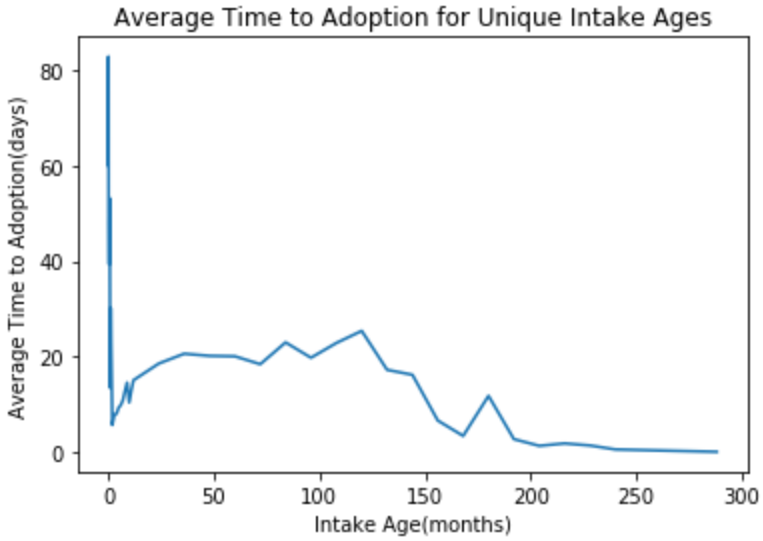

Figure 7: Intake Age vs. Average Time to Adoption for each unique Intake Age.

The plot shows that really young dogs on average take the longest to get adopted. This may because of things like weaning and spaying or neutering. Then the plot trends up logarithmically with intake age. This suggests that for most of the dogs in the shelter, older dogs are adopted less quickly. However, the really old dogs (outliers in age with only a few instances) are adopted quicker than the younger dogs on average.

I decided to lastly take a look into whether the time of the year had any significance in time to adoption or chance of adoption. The results are outlined in the table below.

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 of the statistical testing of seasons.

The time of year is a significant factor in time to adoption and chance of adoption. Spring and Winter are better times of the year for adoption compared to Fall and Summer. To even out adoptions throughout the year more adoption events and community outreach events could be organized for the Fall and Summer months.

In summary, the main factors in determining chance of adoption and time to adoption for the dogs are age and breed. In regards to shelter operations, it seems dogs are adopted more in the Spring and Winter than in the Fall and Summer. These are all topics and insights that will hopefully aid in making adoptions more even throughout the year and fair for all dogs no matter the breed or the age.