Capstone Project 1: Austin Animal Center – Milestone Report

1 Background

The Austin Animal Center is the largest no-kill animal shelter in the United States. Their major resources are monetary donation from public and volunteers. As the return, the public wants to learn more about their accomplishment. The Austin Animal Center posted monthly animal status report and yearly finance report. For the public, the reports fill the information gap between sponsors and animal center. For the administration of Austin Animal Center, they can allocate the resources better with past statistics data.

2 Clients

The Austin Animal Center wants to expand the facility. They want to know what is the best strategy. Besides, they also want to publish the data of the achievement to attract more financial support.

They want to know how many animals they saved every year and how they saved them. Also, they want to predict the trends in the future 5 years.

3 Date resources

The data is available as Kaggle site. There are 3 csv files, contains data from 2013 to 2018. The information includes animal categories, like animals' type, sex and age. There is also intake and outcome related information, like intake resources and time. The two files are intake and outcome data separately. The third file contains all the information. Because some animals entered and left animal center several times, the record are matched with special internal animal ID.

4 Data Wrangling

4.1 Import the data

The Austin Animal Center is the largest no-kill animal shelter in the United States that provides care and shelter to over 18,000 animals each year. As part of the AAC's efforts to help and care for animals in need, the organization makes available its accumulated data and statistics as part of the city of Austin's Open Data Initiative.

The dataset contains three csv. Files, which can be downloaded from Kaggle website. I imported Pandas package for further process.

Data source: https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes
The data is imported as data frame with code: pd.csv_read

4.2 Review the data

I review the data frame by reading it directly to get a general understanding. The first file is data of all animals that been taken in. The second file is the data of all animals that were sent out. The third data is the combination of another two data, merged on animal id and intake information. I mainly worked on the merged data frame, since it contains all the data.

I checked the information of the data frame, by df.info(). I got the shape, column, and object of the data frame.

4.3 Check the missing value

I counted the missing values of each column by code: df.isnull().sum().

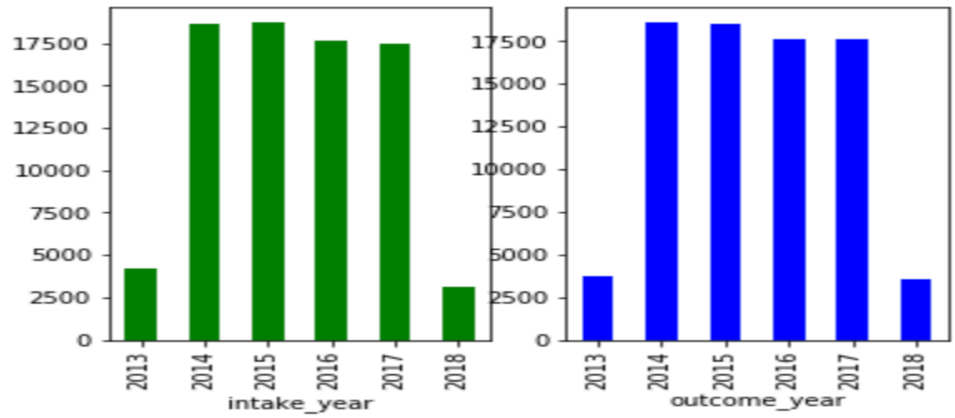
There are 4 columns with missing value: *outcome_subtype*, *outcome_type*, *sex_upon_outcome*, and *sex_upon_intake*. The *outcome_subtype* records the condition of animals and the missing value means a normal condition, so I will leave them with NaN. I found there is value in *outcome_date* column, while *outcome_type* value is missing. So those records are incomplete and invalid. Then I dropped them.

Code: combine = df.dropna(subset=['sex_upon_intake', 'sex_upon_outcome', 'outcome_type'])

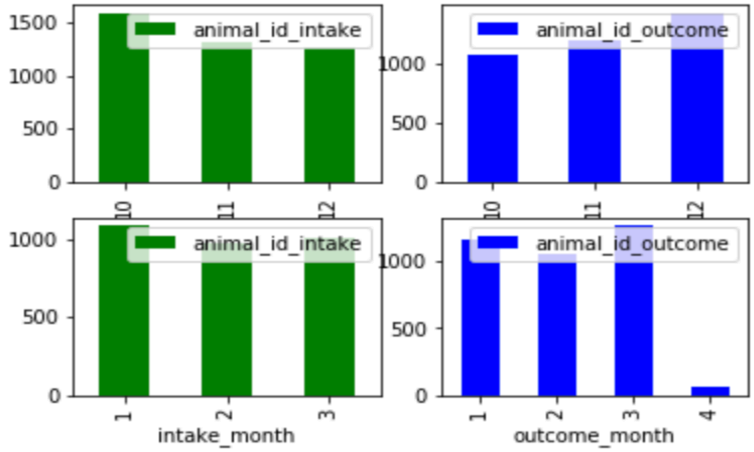
4.4 Find and remove outliers

Outliers may be caused by incorrect record or data abnormalities. It may lead to a wrong prediction or incorrect conclusion. We need to remove the outliers when necessary. Data visualization is a good method to determine outliers.

I plot the histogram of the count of the data each year, with df.plot(kind='bar'), and got the following result.

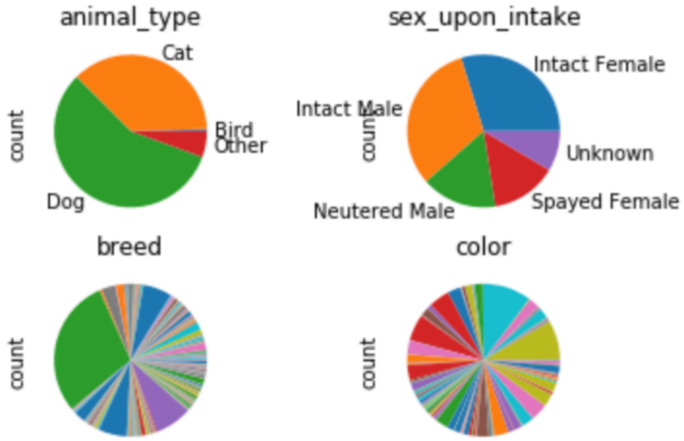


It contains data from 2013 to 2018, while the data of 2013 and 2018 are much less than data of other years. I guess it just records several months instead of all month for 2013 and 2018. So I plot the histogram of data count by month for 2013 and 2018 to confirm my thinking, the result shows as following:



Only October, November, December of 2013 and January, February, March, April of 2018 were included. I dropped the data of 2013-10 and 2018-04 because they are boundary value, which could be an incomplete record for the month.

Then I checked the *animal_type*, *animal_sex*, *animal_color,* and *animal_breed* by plotting pie chart to find out if there are outliers. Code: plot(kind='pie'). Result show as following:



There are hundreds of breed and color categories, so no outlier can be found. The 'other ' in *animal_type* and 'Unknown' in *sex* is unusual category, so I sliced the data frame with condition df['sex_upon_intake'] == 'Unknown' and df['animal_type'] == 'Other'.
The other types are bat or rabbit, which is specified in the breed column, so the data is valid. For the unknown sex, most animals are bat or rabbit, sex value of which is not required. There is also cat and dog with unknown sex, which is invalid input. I dropped the row that sex is unknown and animal type is cat or dog.

5. Storytelling

5.1 How long did animals stay in animal center?

This question is the top concern of client. Because the major cost of animal center operation is maintaining the life of animals, the time of animal stay determines financial requirement. Thus, the cost is significant positive related to average animal stay time.

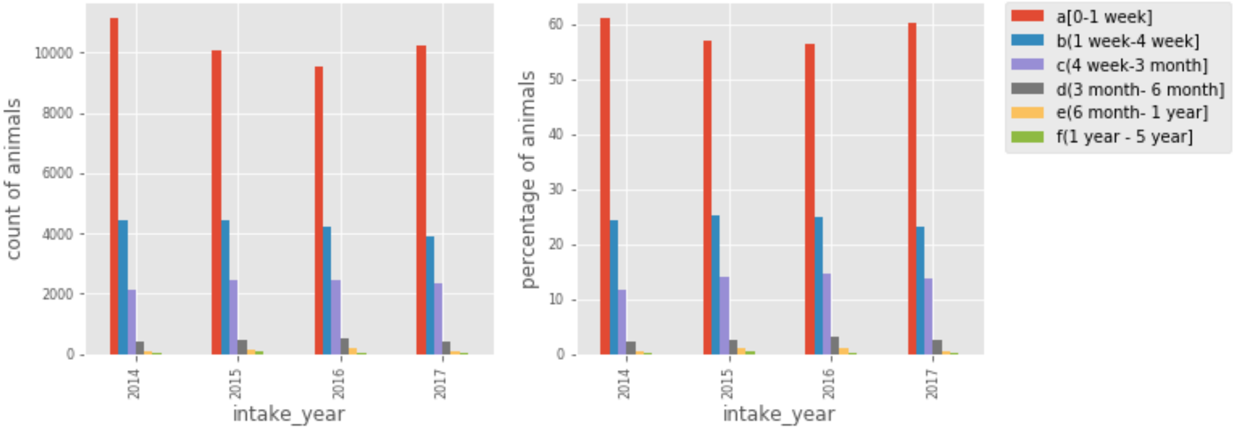


Fig5.1.1 Animals' time in shelter of year 2014,2015,2016,2017

60% animals stayed in shelter less then 1 week, and 85% animals stayed in shelter less than 1 month. This time distribution doesn’t change much from year 2014-2017. We can predict that 85% of animals will stay in shelter in the future five years.

5.2 Where and how many animals we will receive?

Now we know most animals stays in shelter less than a month, but how many animals we receive every month? Does the amount change monthly? Where do we receive them?

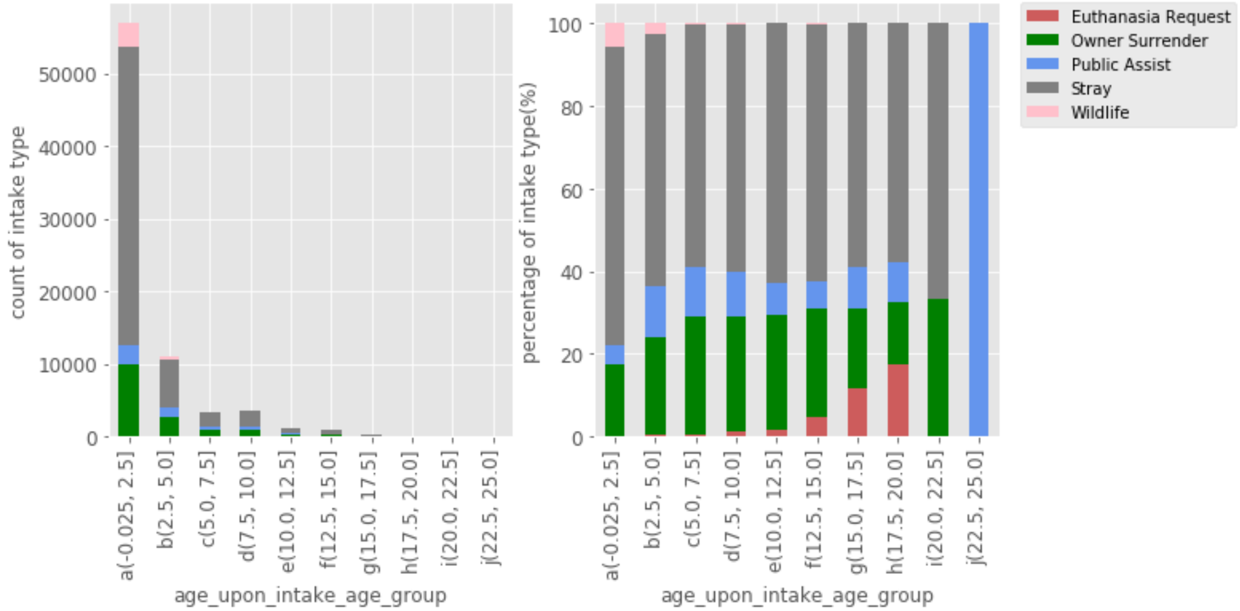


Fig5.2.1 Animal age and source when they were received

Most animals were younger than 2.5 years when they were received. Most of them are strayed animals, which means we should put much effort on searching for animals outside.

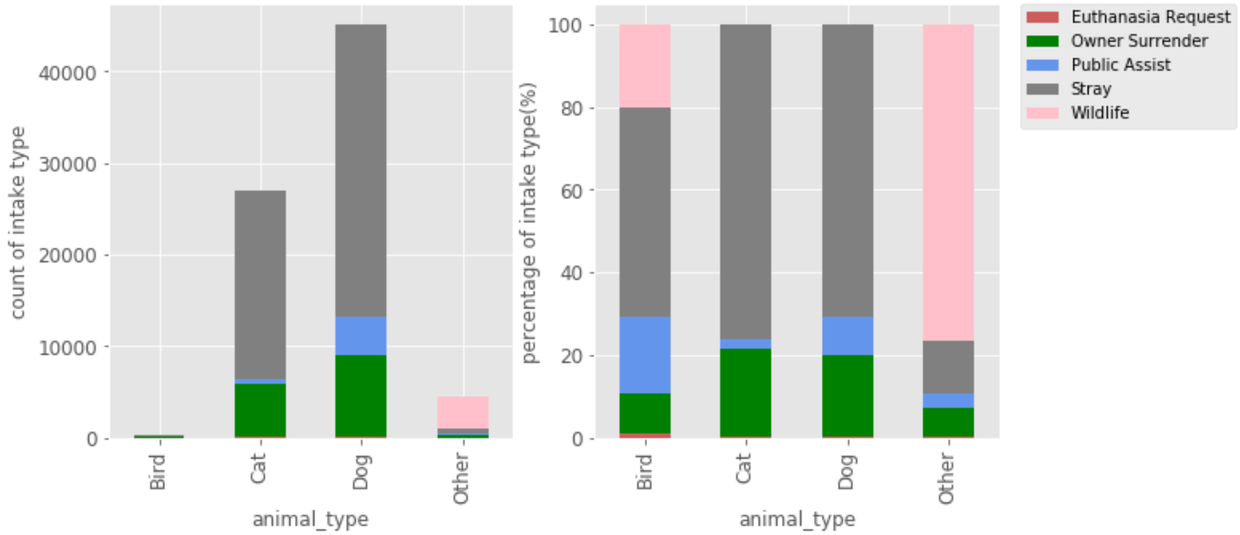


Fig5.2.2 Animal types and where they were received

Cats and dogs are domain animals received, in which dogs is more than cats.

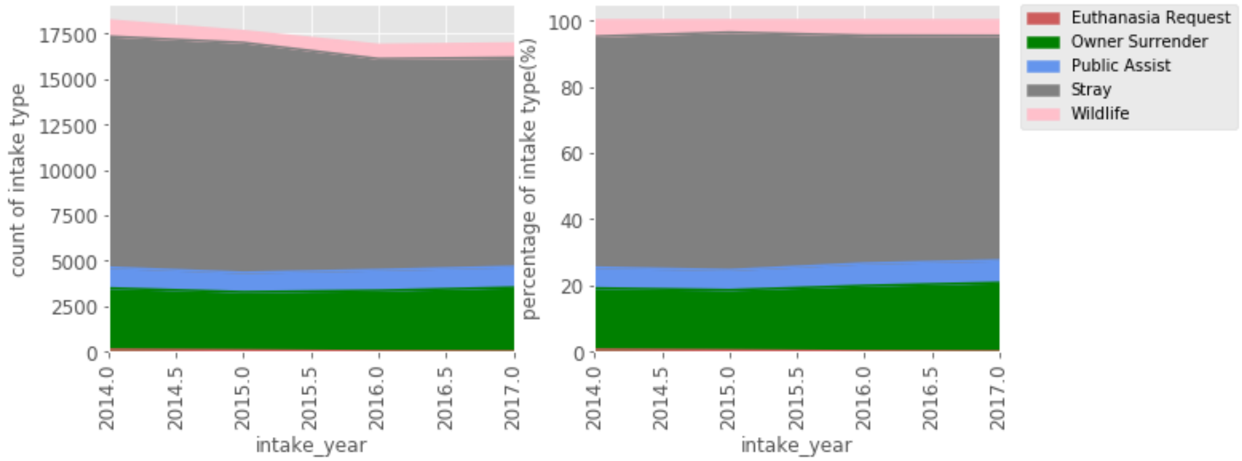


Fig5.2.3 Annually animal received of year 2014-2017

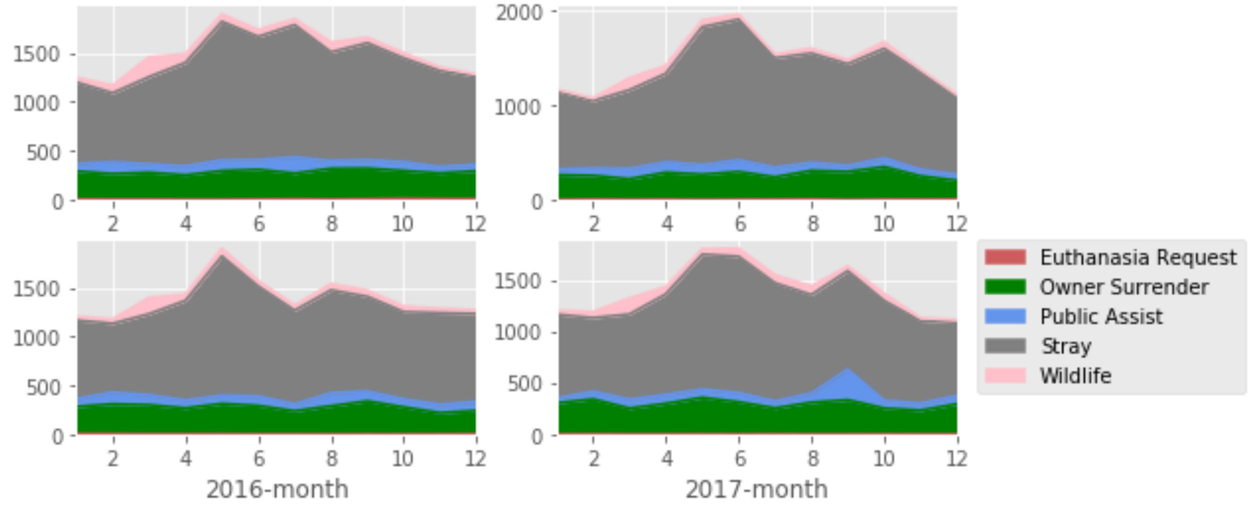


Fig5.2.4 Monthly animal received of year 2014-2017

There is no significant change of received animal amount each year, which are around 18000. However, there is large variety of monthly-received animal amount. Usually, May is peak and August is the second peak. We received least animals in January.

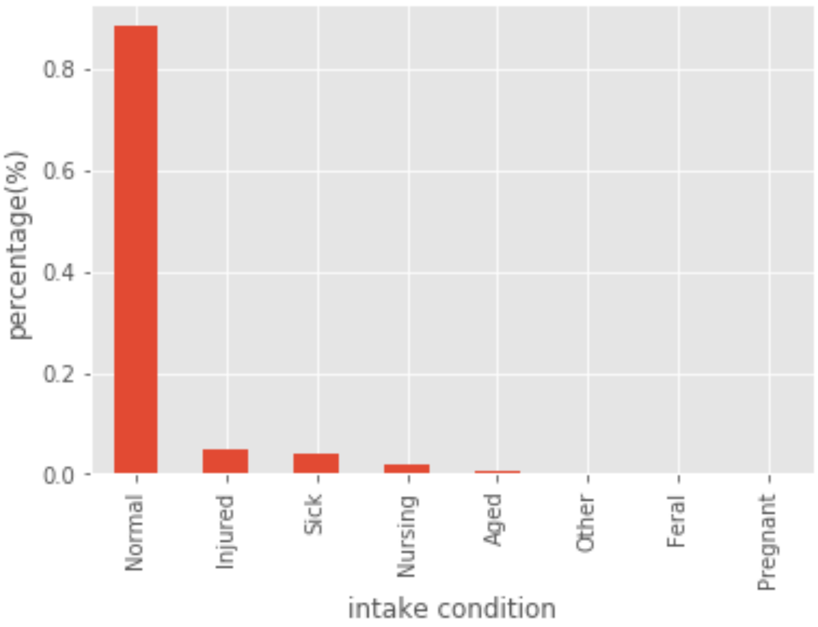


Fig5.2.5 Income condition %

Around 90% of animals were in normal condition when they entered shelter.

5.3 When and where animal leave the animal center?

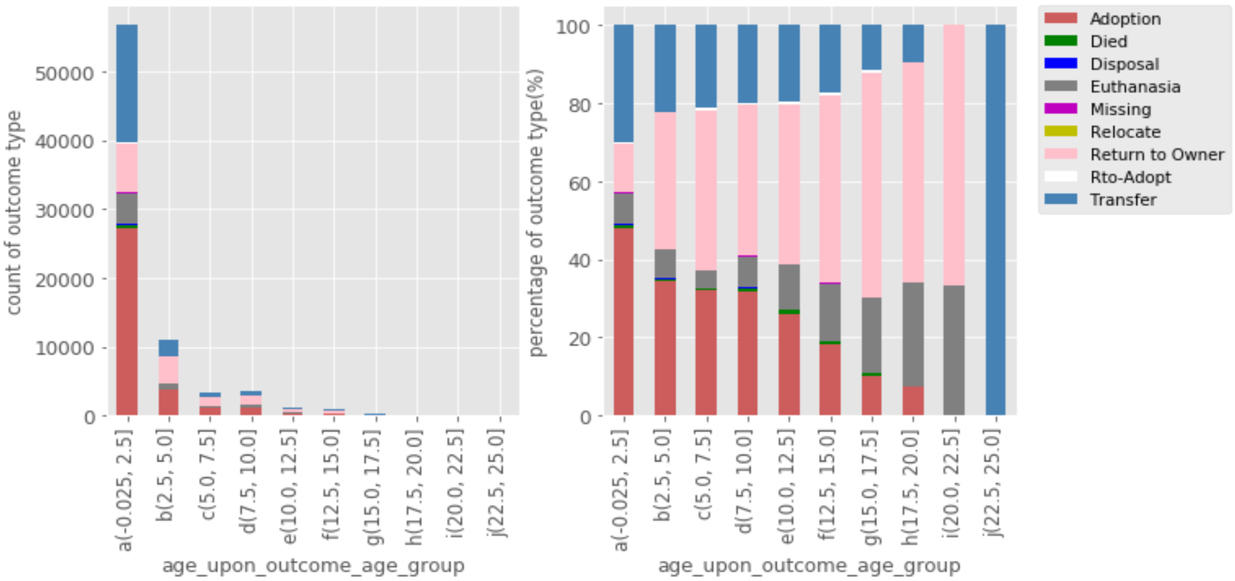


Fig5.3.1 The outcome type vs. outcome age group

Most animals are younger than 2.5 years old when they left the shelter, which is more than 80% of total animal amount. The animals of youngest group were more likely to be adopted than animals of other age groups. It is good to see the percentage of animal returned to owner increasing as percentage of dotation decreases, both animals outcome help them find a home. On the other hand, the euthanasia rate increases, as animals grow old.

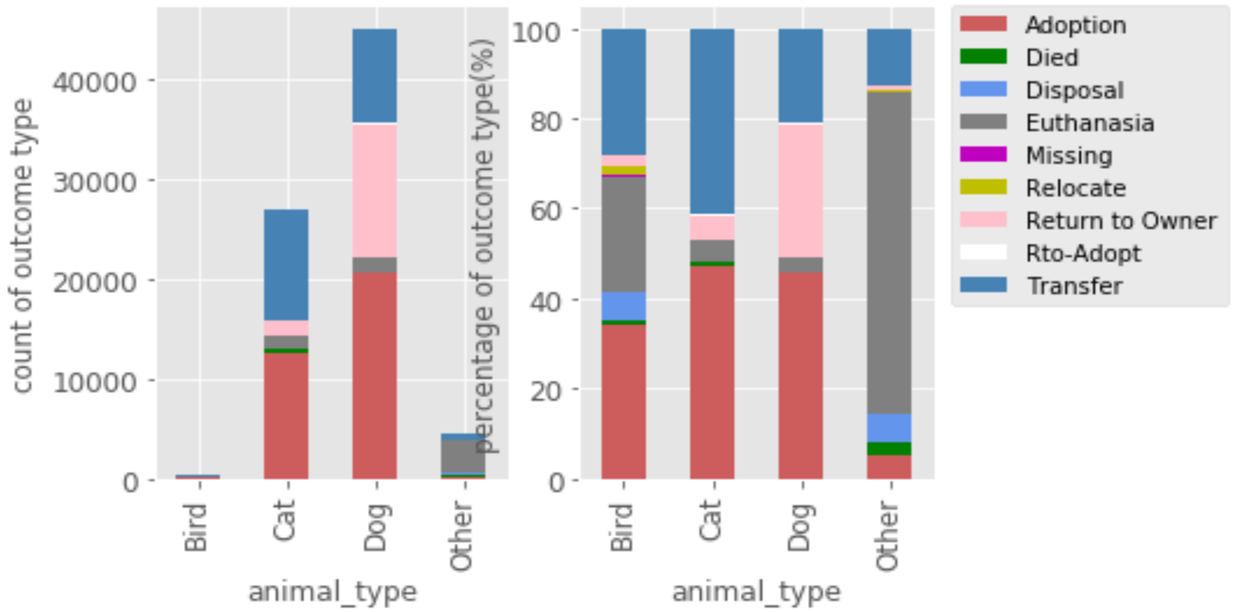


Fig5.3.2 Animal type and outcome type

The adoption rate for cat and dog are very similar, but much more dogs were returned to owner. Some of birds and other animals are disposal, while no cat and dog are disposal. These animals must be wild animals and keep ability to survive outside.

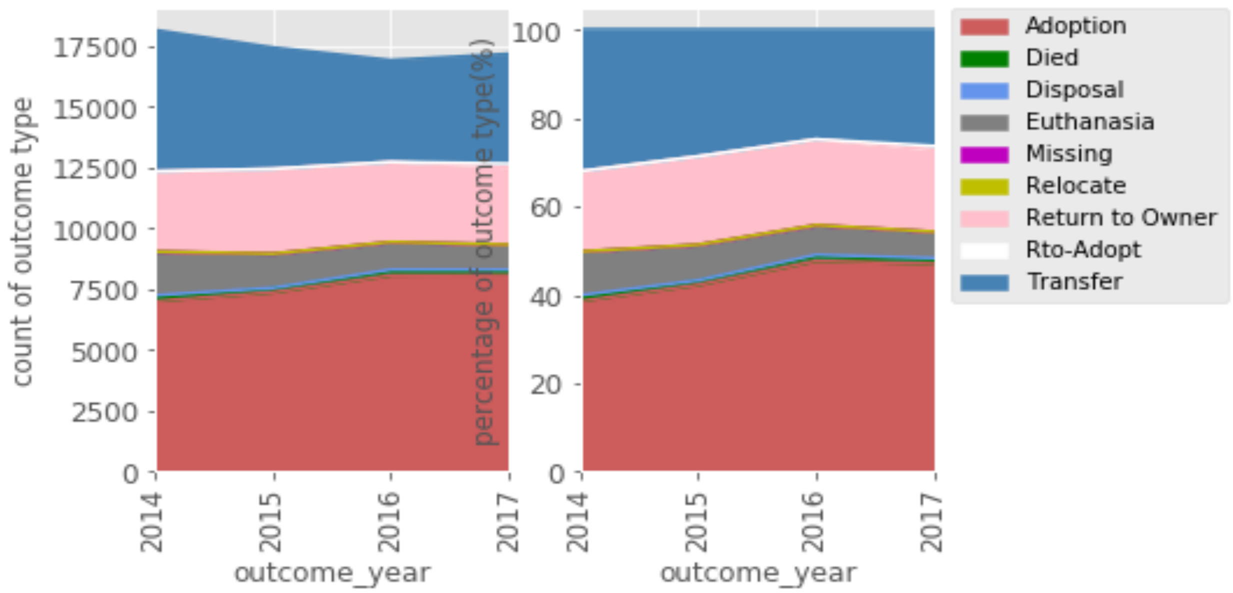


Fig5.3.3 Annually animal outcome

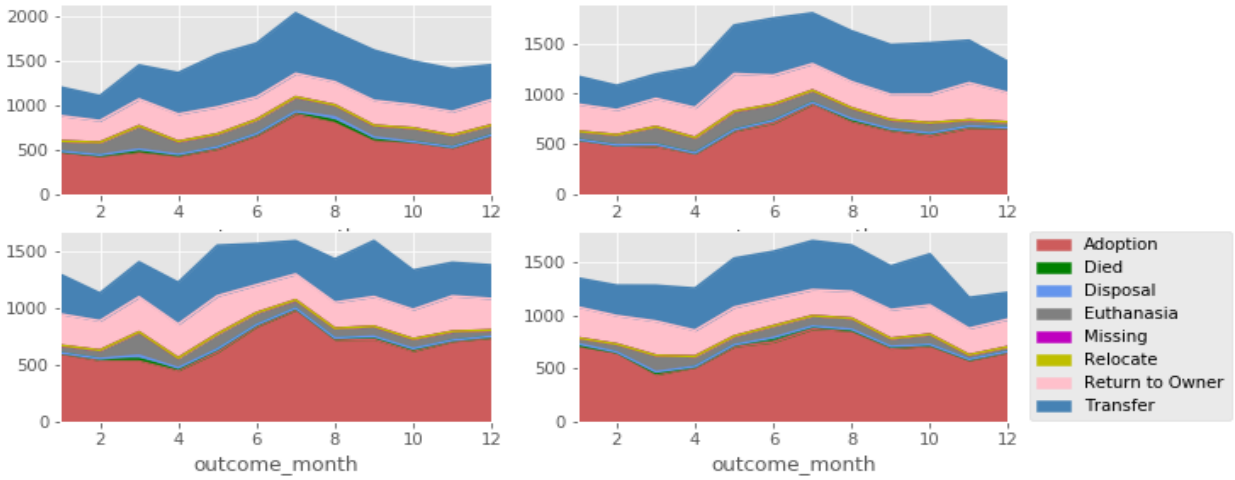


Fig5.3.4 Monthly animal outcome from 2014 to 2017

There is not much change for total amount annually, but there is pattern for monthly change. Usually June to August is the peak time, more people adopted animals then. There is no time-related change for other outcome type. There is a little increase of adoption rate from 2016 to 2017.

5.4 Why animals are conducted euthanasia?

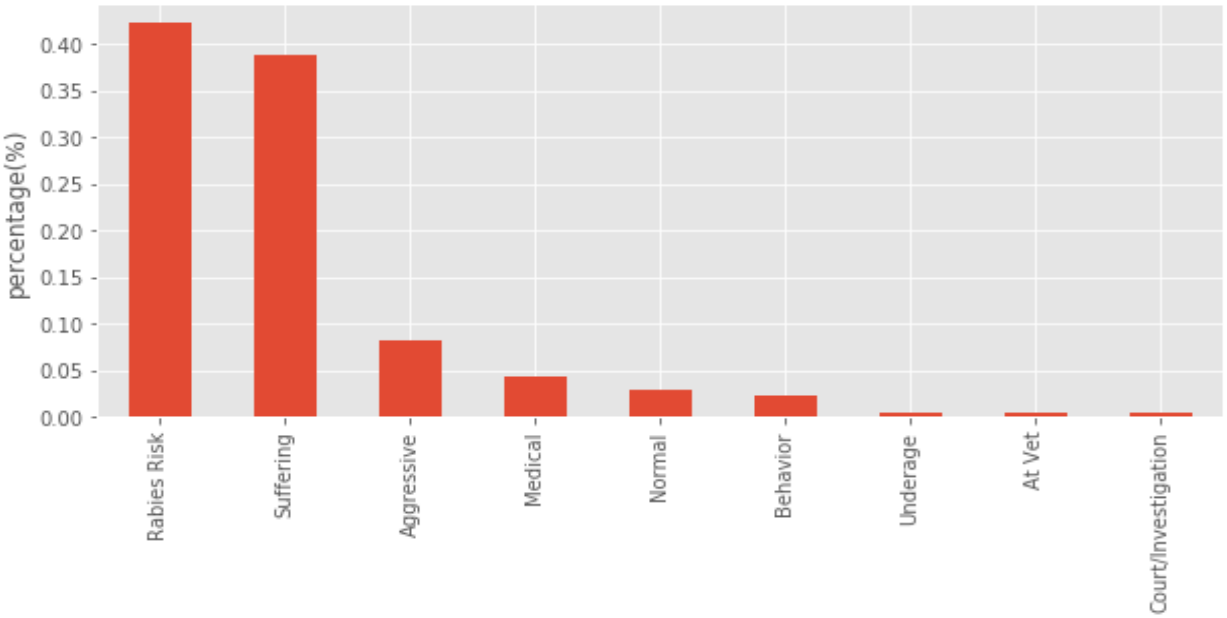


Fig5.4.1 Reasons of euthanasia

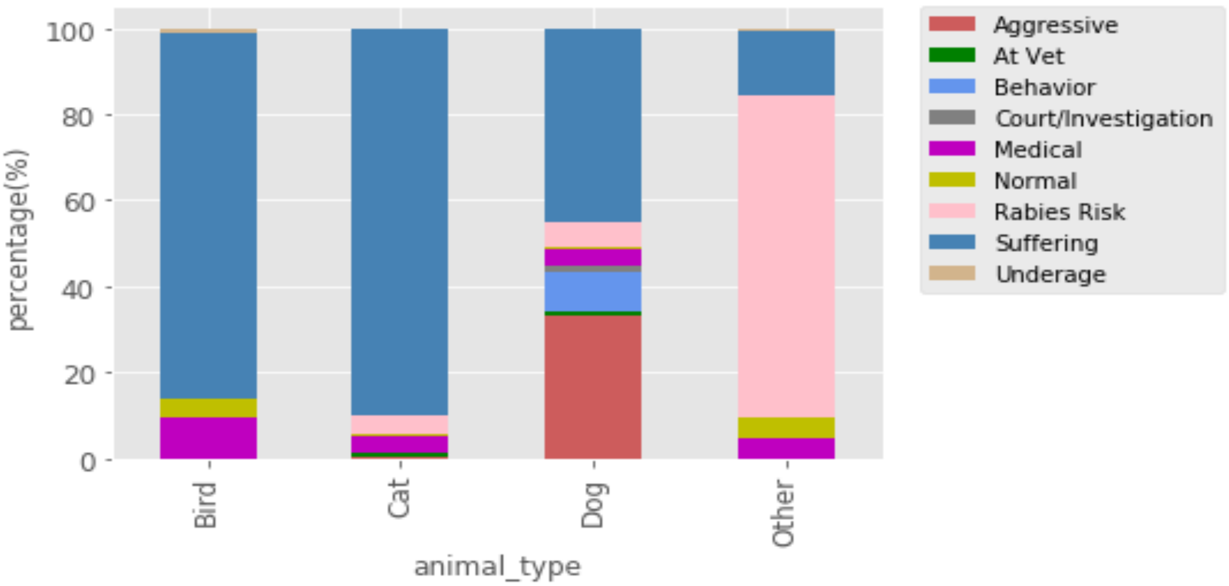


Fig5.4.2 Reasons of euthanasia of each animal type

Top reason for euthanasia is rabies and suffering, which takes 80%. In terms of each animal type, more than 85% cats and birds were conducted euthanasia because of suffering, while only 45% of dogs were conducted euthanasia for same reason. Aggressive and behavior issue is an important cause of euthanasia. Around 80% of other animals were conducted euthanasia due to rabies risk. They were mostly wild animals, like rabbits and bats, having more accessible to virus.

5.5 Why do we need to neuter the animals?

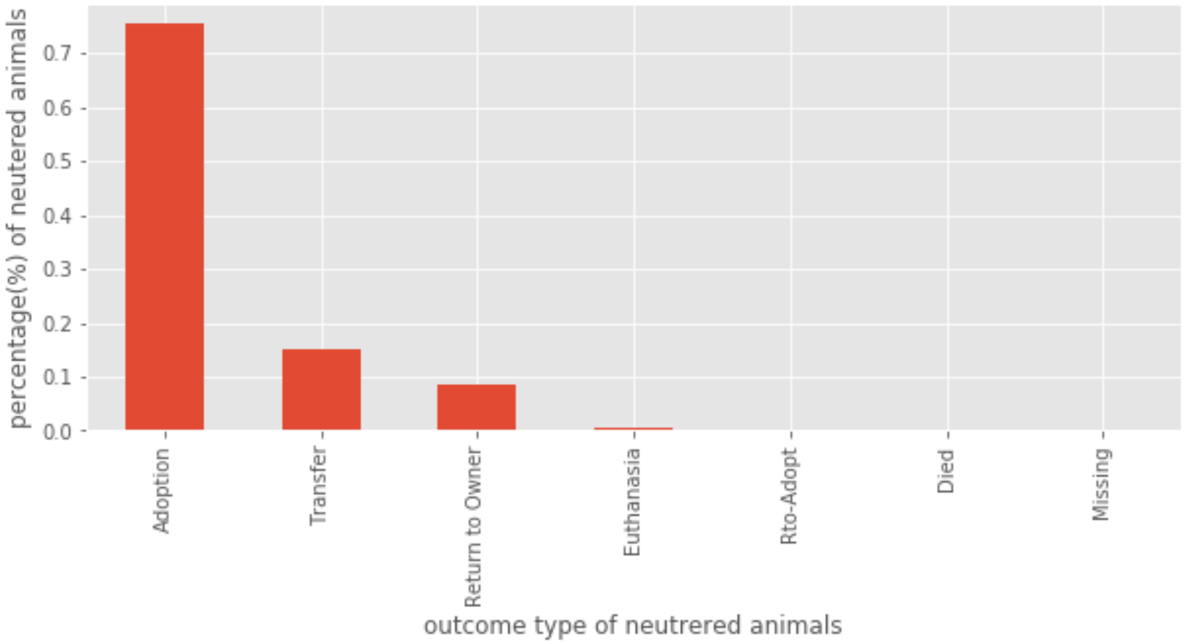


Fig 5.5.1 Outcome type of neutered animals

More than 75% of animals neutered for adoption. Other animals were returned to owner or transfer. They rarely neutered animals for other purpose.

5.6 What is the animal breed and color should we expect?

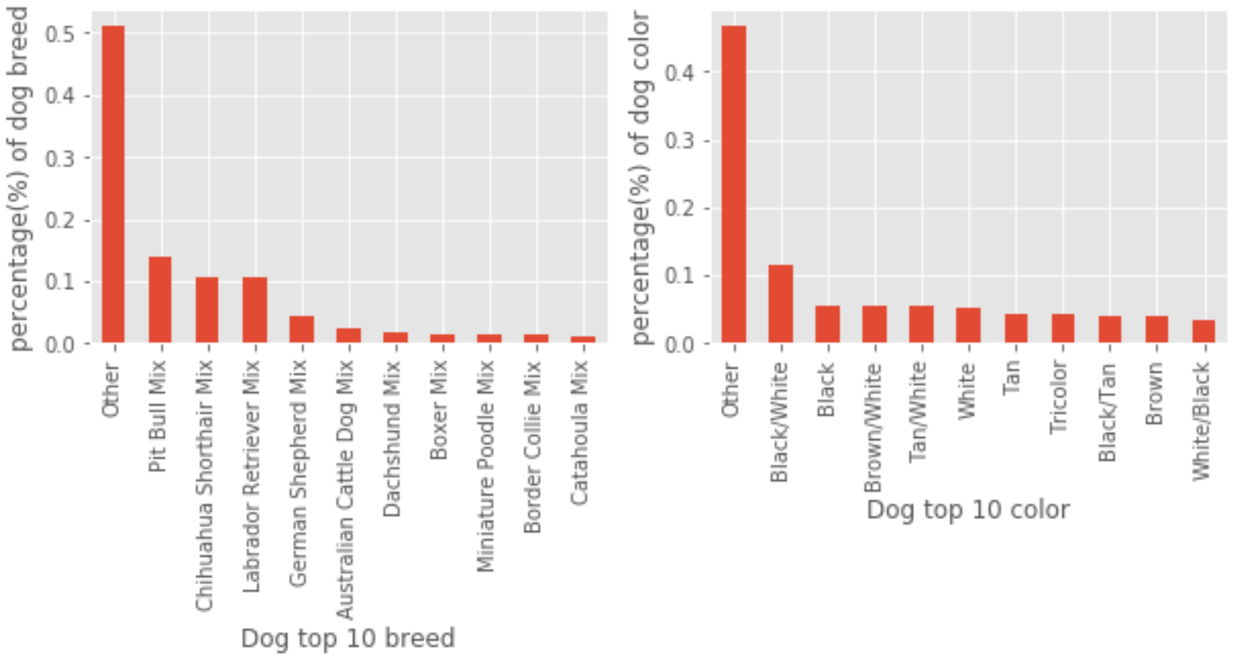


Fig 5.6.1 Top 10 dog breed and color

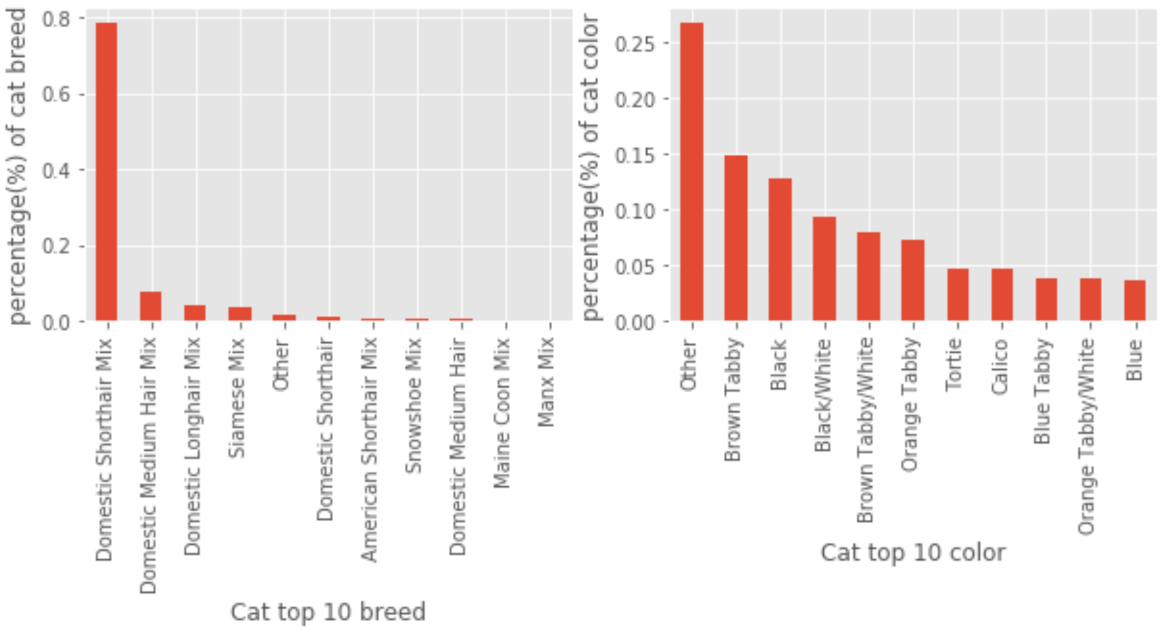


Fig5.6.2 Top 10 cat breed and color

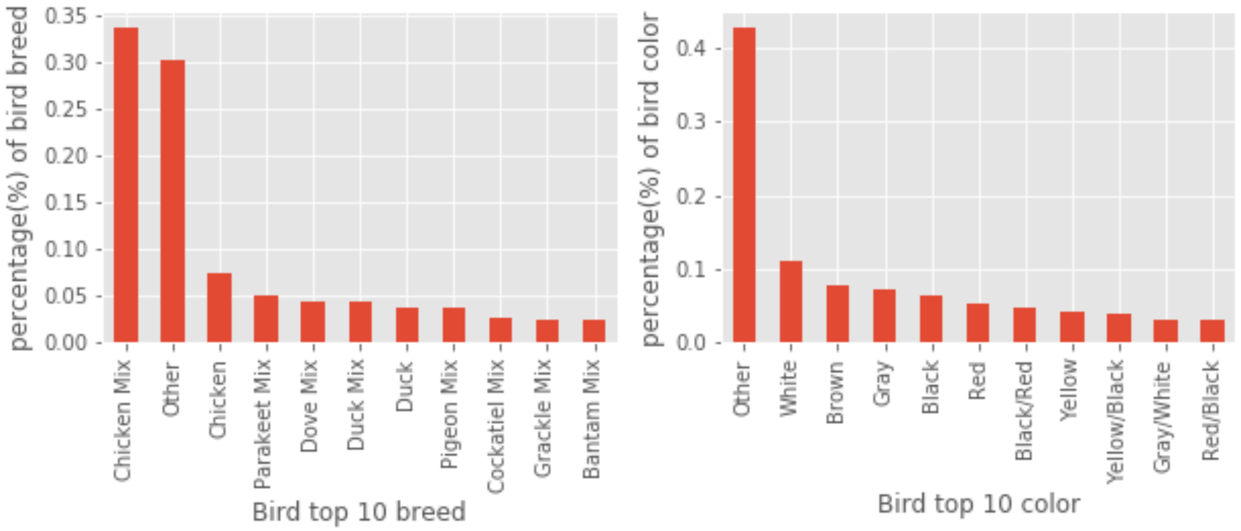


Fig5.6.3 Top 10 bird breed and color

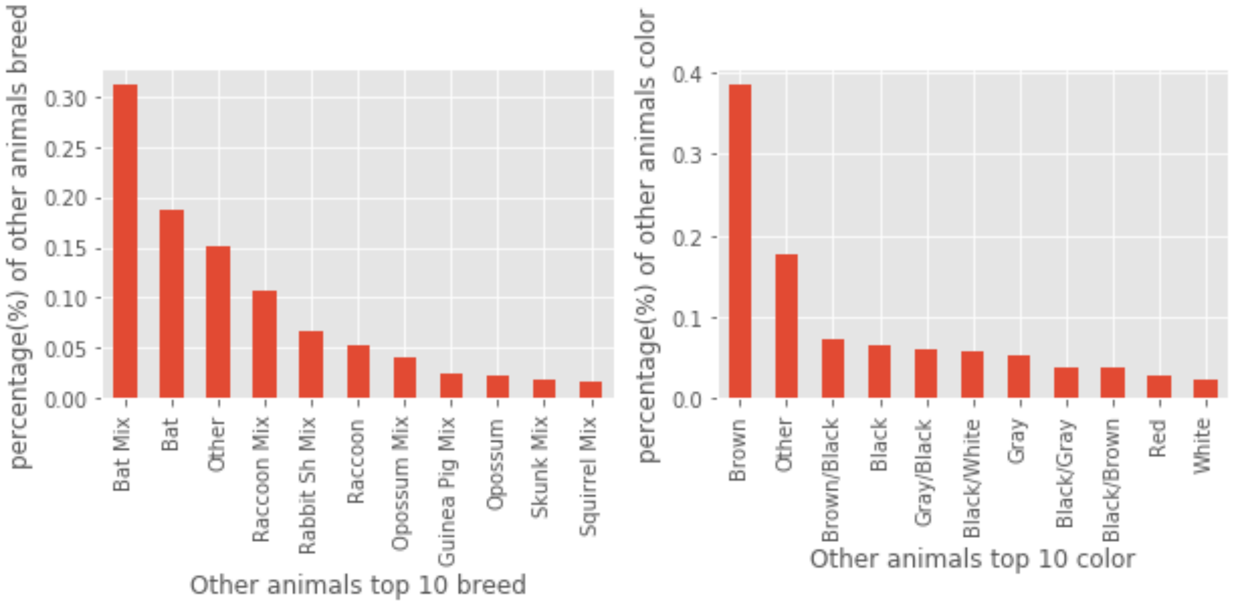


Fig5.6.4 Top 10 other animals breed and color

6 Conclusions and Discussion

6.1 The animal center should be preparing for contain and feed at least 2000 animals, 57% of which are dogs, 40% of which are cats and 3% are birds and other animals. Because over 85% animals stay in shelter for less than a month, we can estimate the animal's amount in shelter by counting how many animals received within a month. Since there is large variety of animal receiving amount for different month, the month of highest volume should be considered. The peak volume of both animal receiving and leaving happens on June to July, and the volume is 2000. Besides, cats, dogs and other animals need different container and food, the animals type distribution also need to be considered. There is no much change for the overall amount and animal type distribution over four years, so it can be assumed that no much change would happen in next four to five years.

6.2 More than 90% animal are younger than 2.5 years old, and more than 85% animals are strayed animals. Around 95% of animals are healthy. The animal center can make poster on the website to encourage people report and deliver strayed animals, which may help them find and help more animals. The good news is the health condition of animals is good, so they don't have to prepare too much first aid method.

6.3 People tend to adopt animals at younger age, and dogs are more likely to be returned to owner. The animal center can post animals pictures of younger animals, which can attract more people for adoption. Also, they should try to put the dog pictures prior cat pictures on the lost and found page, it may helps owner to find their pets more quickly.

6.4 More animals were adopted from June to August than other months. The animal center can host animal event to encourage people visit the animal center and increase adoption rate.

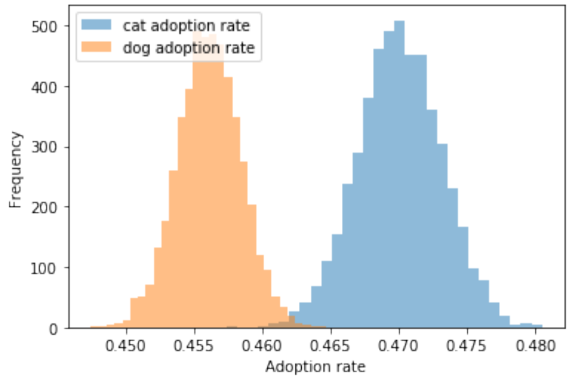
6.5 Aggressive and behavior issue is a common reason for dog euthanasia. It is regret to see euthanasia of any animal. The Austin animal center can train the dog before adoption. An effective training can decrease aggressive behavior.

6.6 Neutered animals and find them a home. More than 75% of neutered animals were adopted. Compared with the adoption rate of 50% of animals, neutered animals were more likely to be adopted.

7. Statistical_Data_analysis

7.1 Is adoption rate of cats the same as dogs?

7.1.1 Bootstrap



According to the histogram plot, cats adoption rate is higher than dog adoption rate.

7.1.2 Significance test

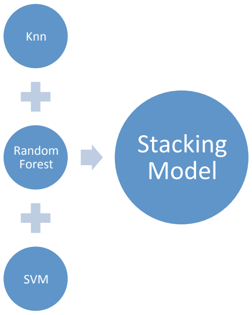
Null hypothesis: Dog adoption rate is equal to cat adoption rate.
Alternative hypothesis: Dog adoption rate is different from cat adoption rate.
p-value=0. P-value is small enough to reject Null hypothesis, so it is confirmed cat adoption rate is higher.

8. Machine Learning

8.1 Goals and model design

The goal of Austin Animal Center is helping each animal a forever home. The home can be various. Some animals will be adopted, and some animals could be transferred to other facilities. Some lucky animals and owner can reunion together. There are several pages of Austin Animal Center website to claim the animals: Adoption, Lost & Found and Animal protection. People can see animal pictures on these pages, depending on what their requests is. If we can predict the most possible outcome of animals, we can put their pictures on correct page, so people can find their pets more efficiently and animals can find their home easier.

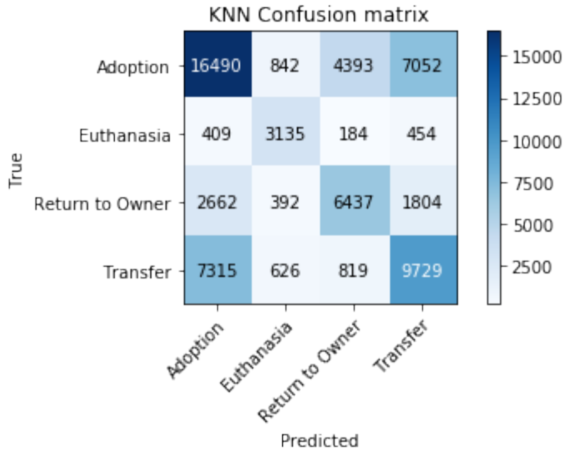
Based on the request, we developed a model to predict the outcome type of animals. It consists of three basic models, K-Nearest Neighbors, Random Forest, Support Vector Machine(SVM), and a stacking model.



This model classifies the animals' outsources types into four classes: adoption, transfer, return to the owner and Euthanasia. With the prediction result, we can help animals find their home more efficiently.

8.2 Models and metrics

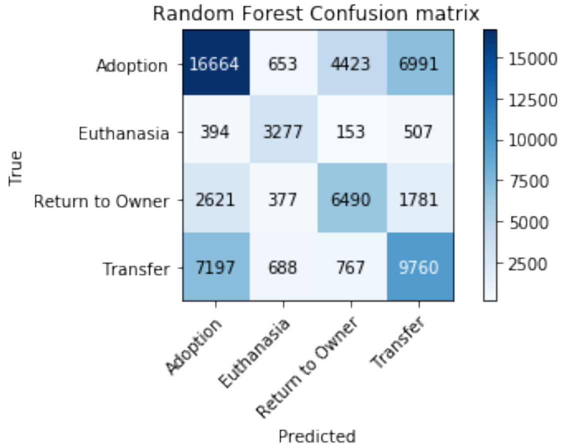
8.2.1 K-Nearest Neighbors result



	Adoption	Euthanasia	Return to Owner	Transfer
Adoption	16490	842	4393	7052
Euthanasia	409	3135	184	454
Return to Owner	2662	392	6437	1804
Transfer	7315	626	819	9729

	precision	recall	f1-score	support
Adoption	0.61	0.57	0.59	28777
Euthanasia	0.63	0.75	0.68	4182
Return to Owner	0.54	0.57	0.56	11295
Transfer	0.51	0.53	0.52	18489
accuracy			0.57	62743
macro avg	0.57	0.6	0.59	62743
weighted avg	0.57	0.57	0.57	62743

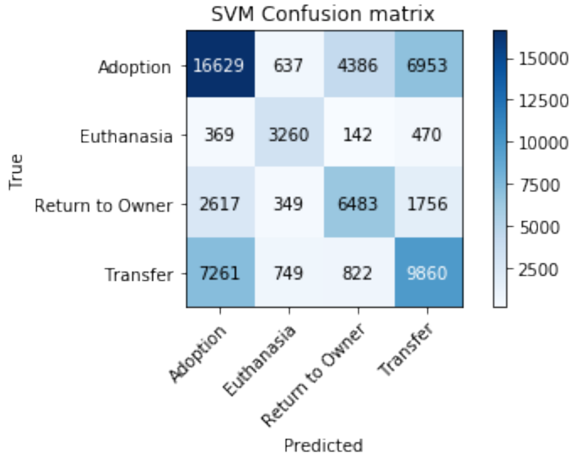
8.2.2 Random Forest result



	Adoption	Euthanasia	Return to Owner	Transfer
Adoption	16664	653	4423	6991
Euthanasia	394	3277	153	507
Return to Owner	2621	377	6490	1781
Transfer	7197	688	767	9760

	precision	recall	f1-score	support
Adoption	0.62	0.58	0.6	28731
Euthanasia	0.66	0.76	0.7	4331
Return to Owner	0.55	0.58	0.56	11269
Transfer	0.51	0.53	0.52	18412
accuracy			0.58	62743
macro avg	0.58	0.61	0.6	62743
weighted avg	0.58	0.58	0.58	62743

8.2.3 Support Vector Machine result



	Adoption	Euthanasia	Return to Owner	Transfer
Adoption	16629	637	4386	6953
Euthanasia	369	3260	142	470
Return to Owner	2617	349	6483	1756
Transfer	7261	749	822	9860

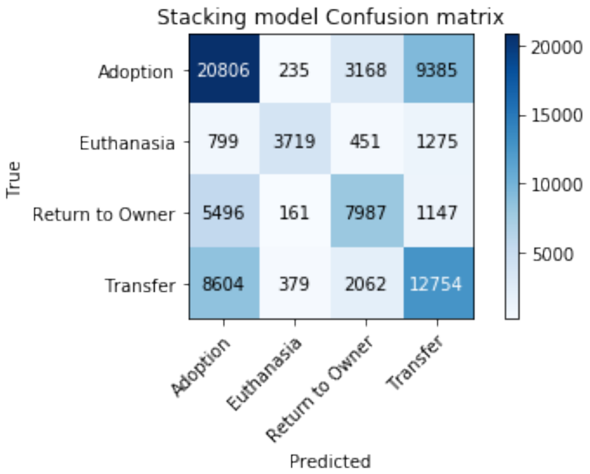
	precision	recall	f1-score	support
Adoption	0.62	0.58	0.6	28605
Euthanasia	0.65	0.77	0.71	4241
Return to Owner	0.55	0.58	0.56	11205
Transfer	0.52	0.53	0.52	18692
accuracy			0.58	62743
macro avg	0.58	0.61	0.6	62743
weighted avg	0.58	0.58	0.58	62743

8.2.4 Three models comparison

Compared with confusion matrix and accuracy score, random forest model and support vector machine model have batter performance than knn model. The overall accuracy score is 58% now.

8.3 Stacking Model

I combined the prediction result of the three basic models, and use as new X data set to train stacking model.



	precision	recall	f1-score	support
Adoption	0.58	0.62	0.6	33594
Euthanasia	0.83	0.6	0.69	6244
Return to Owner	0.58	0.54	0.56	14791
Transfer	0.52	0.54	0.53	23799
accuracy			0.58	78428
macro avg	0.63	0.57	0.6	78428
weighted avg	0.58	0.58	0.58	78428

The stacking model only shows little improvement, only the Euthanasia got much higher precision rate. To predict animals of all outcomes, random forest and SVM are better choices.