

R-Ladies for PAWS Datathon (2019)

R-Ladies Philly

May 01, 2019

Executive Summary

Problem definition and dataset

The 2019 R-Ladies for PAWS Datathon aimed to help the Philadelphia Animal Welfare Society (PAWS) improve its adoptions processes. For this data challenge, PAWS made 2018 data available containing adoption application form submissions, staff processing of applications, and animal outcome data. We developed analytic approaches to better understand the following topics:

1. An animal's trajectory at PAWS
2. An adoption application's trajectory at PAWS
3. Geographic characteristics that influence adoptions
4. Social media activity that could influence adoptions

The data included the following datasets: 1. PetPoint data for animals who had a recorded outcome at PAWS in 2018 (this means that animals who were active at PAWS at the time of download, i.e. January 2019, were not included).

Timeline and Workflow

Prior to kickoff, data were obtained and preprocessed. After downloading the data from multiple sources, individual entries were matched based on first name, last name, address, and other relevant variables. Once matched, any identifiable information was removed from the dataset. This occurred before the data were made available to the group for analysis.

February 12: Kickoff Meetup: At this event, the project was introduced and teams were formed to work on one of the 4 topics outlined above. All participants were encouraged to join the R-Ladies Philly Slack workspace and to set up a GitHub account (which was used as the main collaborative platform).

After the kickoff meetup, groups worked together online, getting together on an as-needed basis. Questions were asked and answered via Slack, with an occasional clarification email to PAWS.

March 26: Conclusion Meetup: At this meetup, teams presented their results and discussed their experiences. Individual team reports were finalized in the weeks following this meetup, and then integrated into the present report.

Results

1. Animal Trajectories

This analysis investigated factors relating to an animal's trajectory in the PAWS system using PetPoint data from 2018. The group operationalized animal trajectory as wait time and outcome (e.g. adoption), with wait time defined as time in days from intake to outcome. We restricted analyses to dogs and cats, as other animals' data points were sparse and compromised statistical power. Our primary factors of interest included animal characteristics (size, breed, health), intake type, and seasonal patterns.

Contributors

Alex Lesicko, PhD is a postdoctoral fellow studying auditory coding at the University of Pennsylvania. She recently moved to Philadelphia from Chicago, where she completed her PhD in neuroscience.

Jake Riley is a clinical data analyst at the Children's Hospital of Philadelphia (CHOP). He enjoys developing tools for analytic teams and specializes on data visualization and geospatial information systems (GIS).

Javier Jasso is a certified speech-language pathologist and a PhD candidate in communication sciences and disorders at the University of Texas at Austin. Javier has expertise in the assessment of culturally/linguistically diverse children, focusing on bilingual language acquisition.

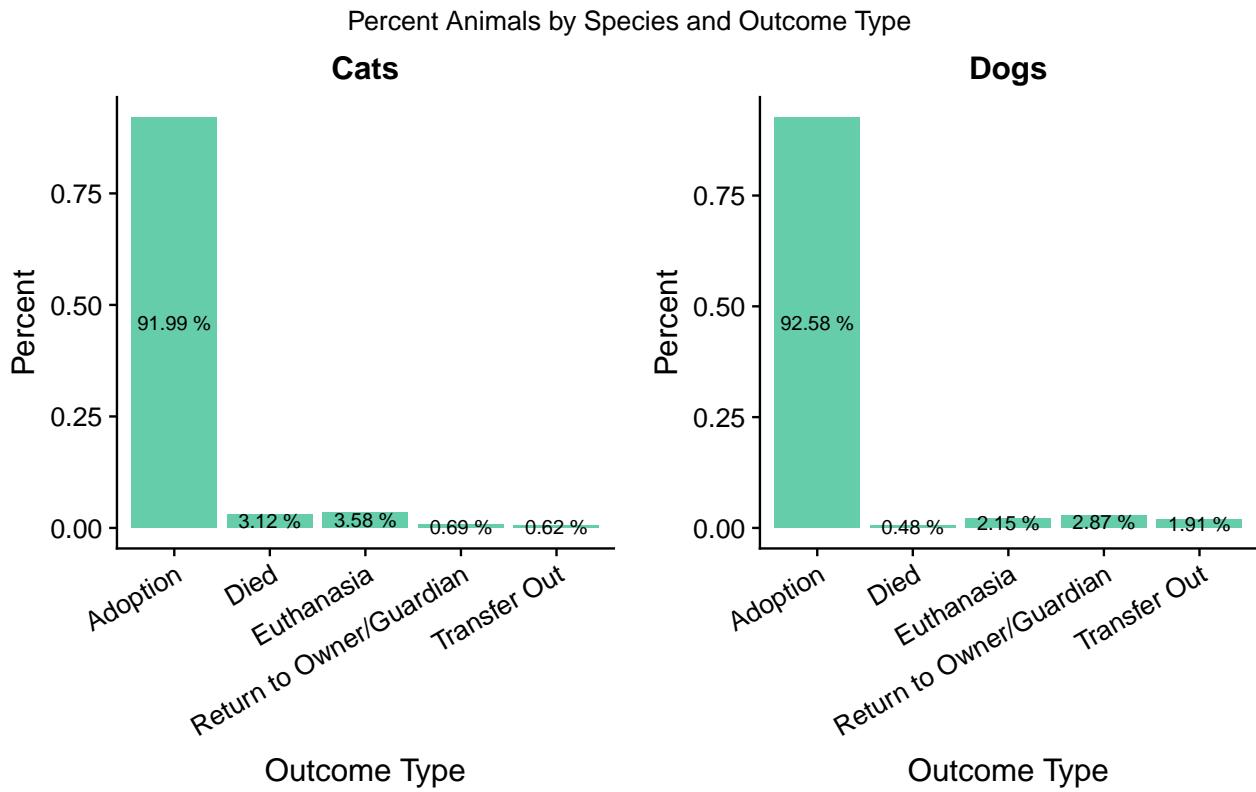
Katerina Placek is a PhD candidate in neuroscience at the University of Pennsylvania and a co-organizer of R-Ladies Philly. She enjoys integrating outreach with teaching and learning in the local data science community.

Summary

- Most animals' outcomes are adoption, and the 'wait time' from intake to outcome is longer for cats than for dogs
- Younger age and poorer health contributes to longer wait time for cats; and intake and wait time for younger, health-compromised cats is highest in spring/summer
- Dogs showed no differences in wait time based on age, health, size, or season
- Increased resources in spring/summer months for young, unhealthy cats may shorten wait times and alleviate staff burden

1. Initial Observations

First, we looked at the overall adoption percentages for cats vs. dogs, compared to other outcomes. Among the cats that were taken in at PAWS in 2018, 92% (2,390 cats) were adopted, 1.31% (34 cats) were transferred out or returned to their owner or guardian, and 6.7% (174 cats) died or were euthanized. For dogs, 93% (387 dogs) were adopted, 4.78% (20 dogs) were transferred out or returned to their owners or guardians, and 2.63% (11 dogs) died or were euthanized. In subsequent analyses, we focused only on those animals whose outcome was "Adoption", "Returned to Owner or Guardian" or "Transfer Out".

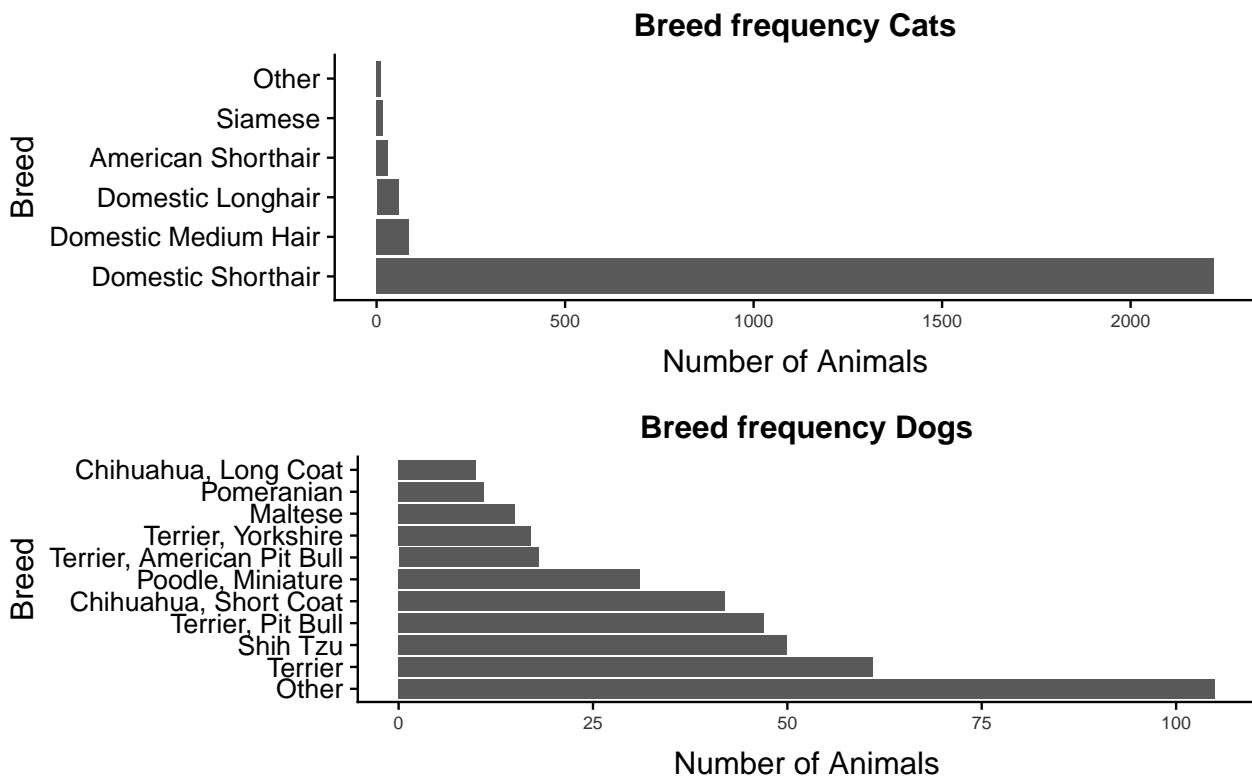


We then looked at the median ‘wait time’ (time in days from intake to release date) for an animal at PAWS in 2018. Across all animals, the median wait time was 45 days (51 days for cats, and 18 days for dogs). Cats waited approx. 18.5 days to be transferred out, 4.5 days to be reunited with their owner, and 52 days to find an adopter, while dogs waited 1.5 days to be transferred out, 2 days to be reunited with their owner, and 20 days to be adopted.

2. Animal Characteristics

Next, for each species, we visualized which PetPoint variables contributed to differences in wait time, and focused our subsequent analyses on the variables with the greatest contributions to wait times: primary breed, age group, and health at intake.

For animal characteristics contributing to wait time at PAWS, we first examined primary breed per species. The graph below shows the top 5 most frequent cat breeds and the top 10 most frequent dog breeds at PAWS in 2018. We found that most, if not all, cats from PAWS in 2018 were ‘domestic short hair’, while dogs’ breed variability was greater, with the most frequent dogs being labeled Terrier, Shih Tzu, Pit Bull Terrier, and Chihuahua. In both graphs, the “Other” category represents a total count of animals in breeds that were not in the top 5 or 10 most frequent.

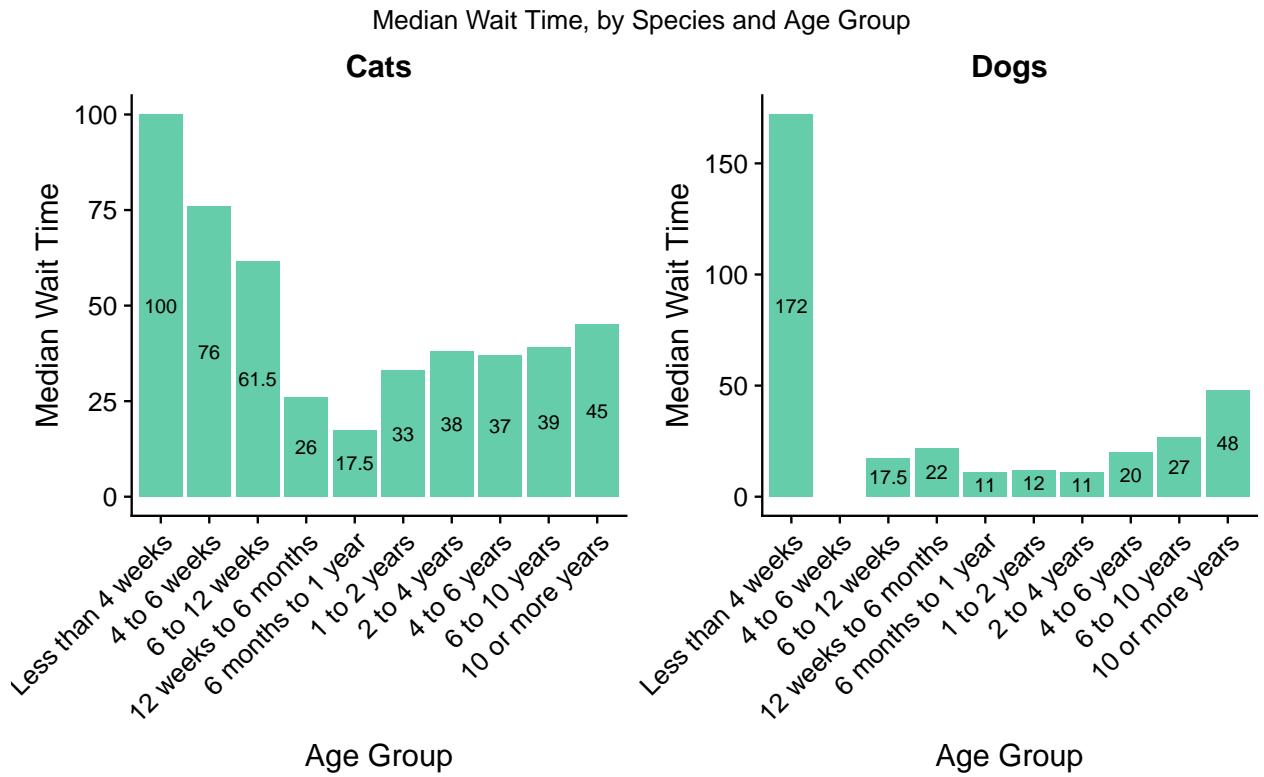


We then looked at wait times by breed, and found that among dogs, Shih Tzus tend to have shorter wait times whereas Terriers tend to have longer wait times. However, given the large number of unique dog breeds in the PetPoint dataset, it was difficult to draw definitive conclusions. Therefore, we classified dogs into 3 size categories based on average weight per breed:

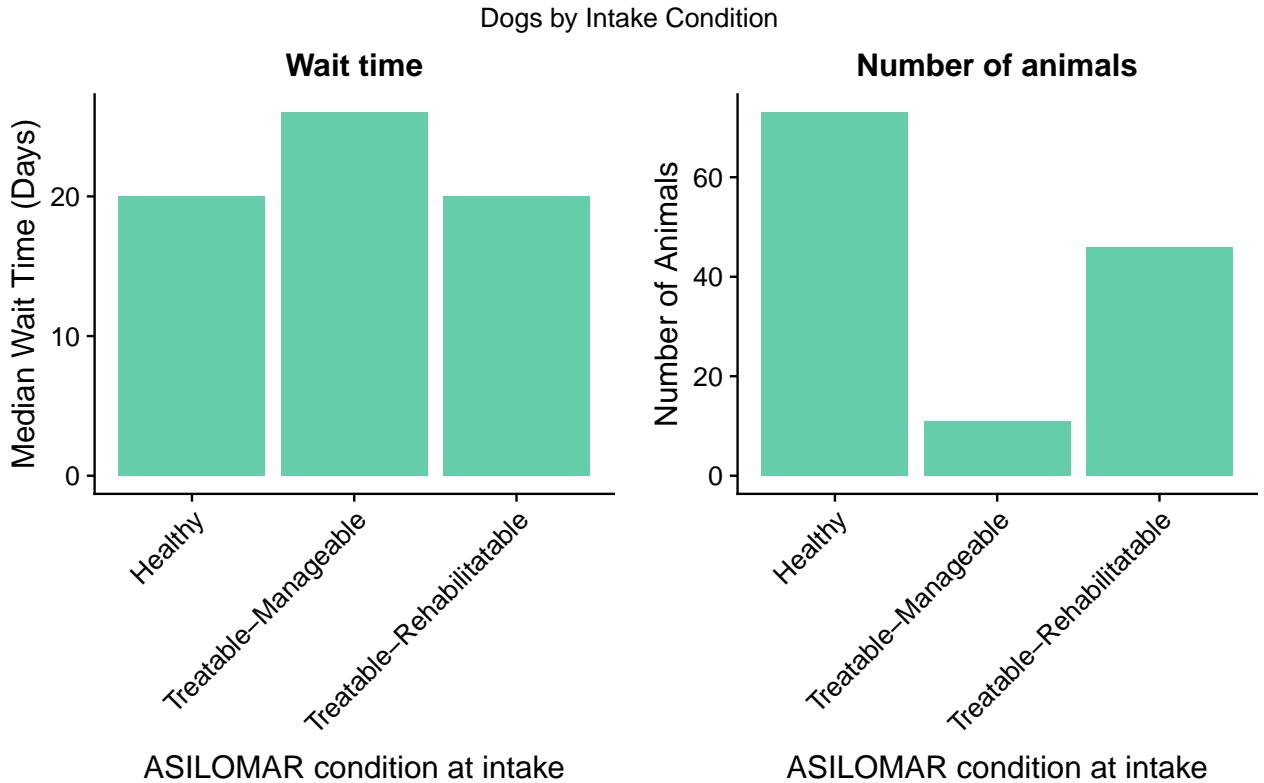
- small = under 24 lbs (Beagle, Bichon Frise, Chihuahua, Long Coat, Chihuahua, Short Coat, Dachshund, Miniature Long Haired, Dachshund, Miniature Smooth Haired, Dachshund, Standard Smooth Haired, Griffon, Brussels, Havanese, Kooikerhondje, Lhasa Apso, Maltese, Miniature Pinscher, Mixed Breed, Small (under 24 lbs fully grown), Papillon, Pekingese, Pomeranian, Poodle, Miniature, Poodle, Toy, Pug, Schnauzer, Miniature, Shiba Inu, Shih Tzu, Spaniel, Cavalier King Charles, Terrier, Terrier, Terrier, Cairn, Terrier, Jack Russell, Terrier, Russell, Terrier, Silky, Terrier, Yorkshire))
- medium = 24-44 lbs (American Eskimo, Mixed Breed, Medium (up to 44 lbs fully grown), Schnauzer, Standard, Spaniel, Spaniel, American Cocker, Spaniel, English Cocker, Terrier, Staffordshire Bull, Terrier, Tibetan, Welsh Corgi, Cardigan)
- large = 44+ lbs (Australian Shepherd, Border Collie, Boxer, Bulldog, Bulldog, American, Hound, Mixed Breed, Large (over 44 lbs fully grown), Norwegian Elkhound, Poodle, Standard, Retriever, Retriever, Golden, Retriever, Labrador, Shepherd, Terrier, American Pit Bull, Terrier, Bull, Terrier, Pit Bull)

While overall our analyses indicated no statistically significant differences in wait time based on dog size category, it is worth mentioning that medium sized dogs' wait time (11 days) trended towards being shorter than the wait time of small breeds (17 days) or large breeds (20 days).

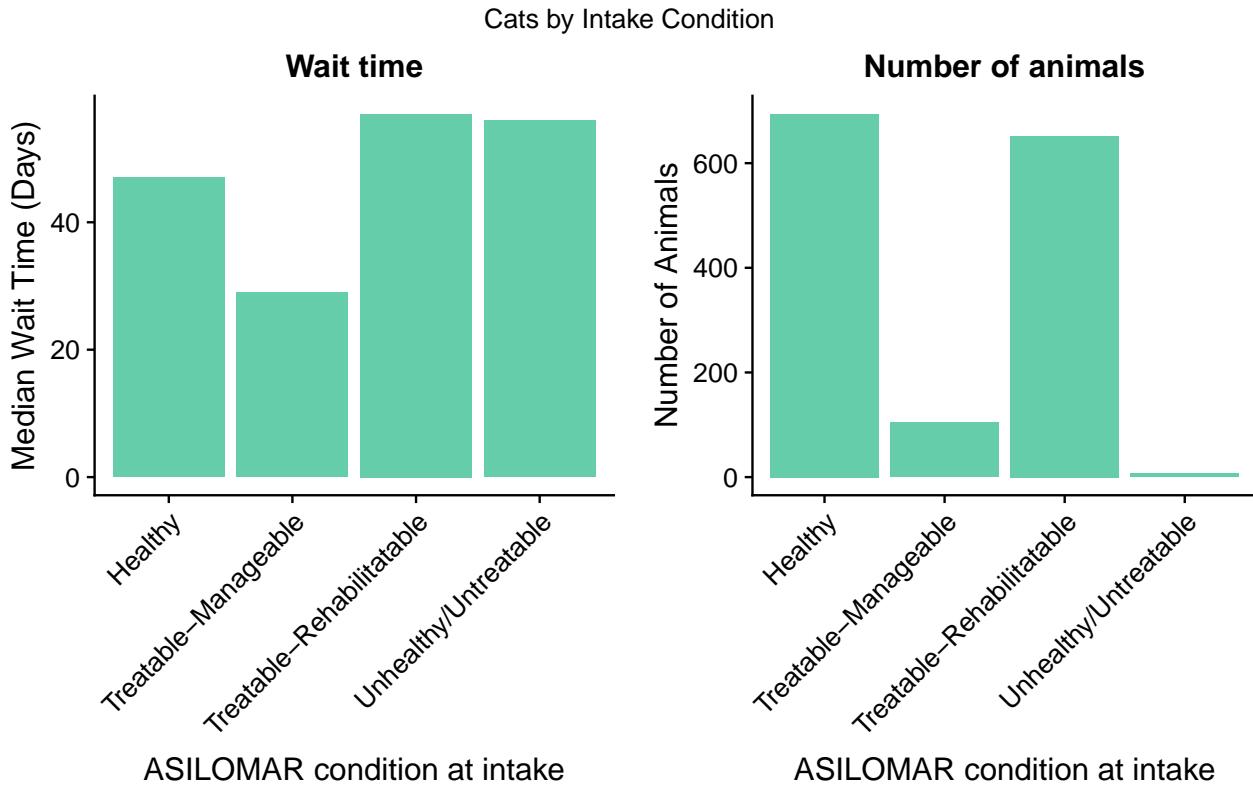
We then examined age group per species, to determine whether wait time different on age group for dogs and cats. For cats and dogs alike, the longest wait times were in the 0-4 weeks age group. Understandably, these kittens and puppies require more time at PAWS before they can be adopted out. Interestingly, the longer wait time holds for cats up to 12 weeks, after which it decreases from roughly 65 days to 35 days. Cats aged between 12 weeks and 1 year spend the least amount of time at PAWS (approx. 26 and 17.5 days respectively). Dogs aged 6 months to 4 years spent the least amount of time at PAWS, while the time at PAWS increased more sharply for dogs 10 or more years of age.



Last, we examined health status at intake per species relative to wait times. For dogs, we observed that the longest wait times were associated with dogs that came in as “treatable-manageable”. However, these dogs represent only a small fraction of the dogs coming in at PAWS, so this effect is unlikely to impact overall wait times for dogs.



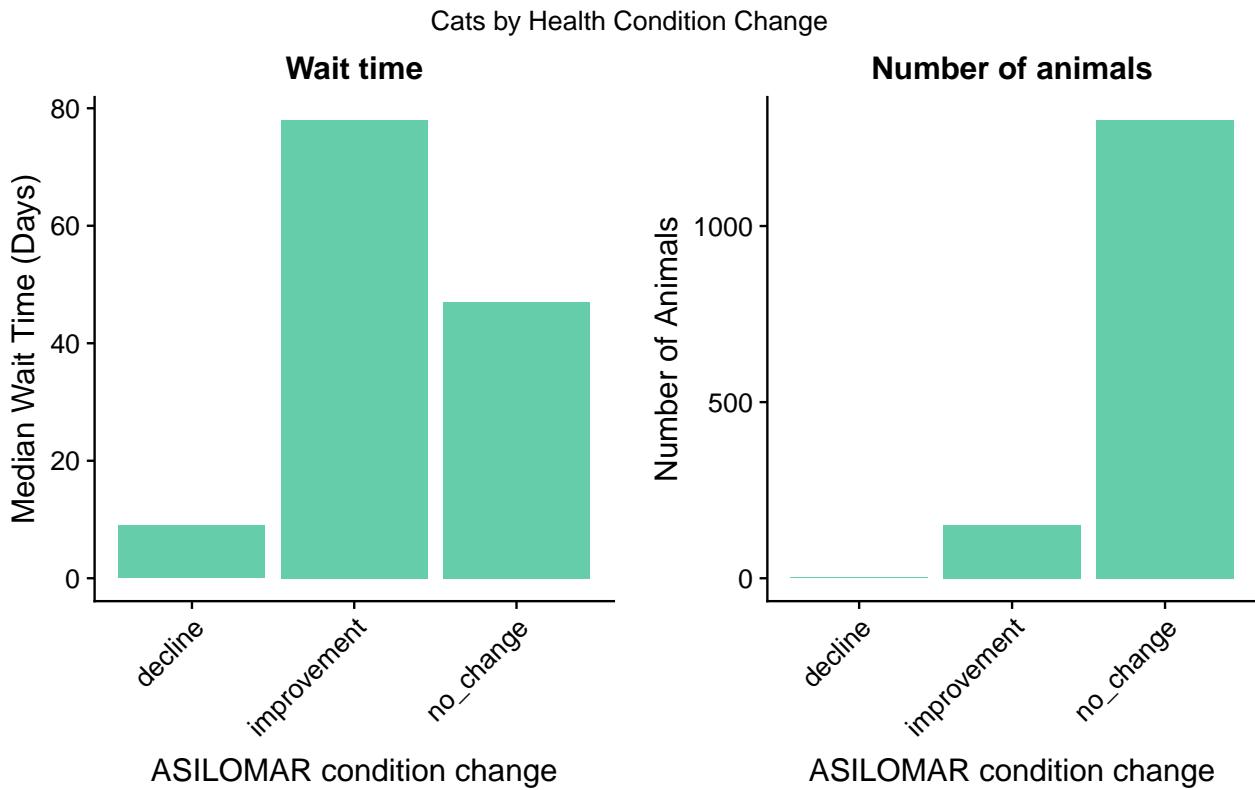
For cats, our analyses revealed that health condition at intake was associated with a longer median wait time. Specifically, cats classified as ‘Treatable-Rehabilitatable’ and ‘Unhealthy/Untreatable’ had longer wait times than cats classified as ‘Healthy’ or ‘Treatable-Manageable’. While cats whose intake status was “Unhealthy/Untreatable” were very few, PAWS took in a very high number of cats labeled as “Treatable-Rehabilitate”, and these were also the ones with the longest wait time. Because of this, efforts to reduce the wait time for these cats would likely impact overall wait times for cats.



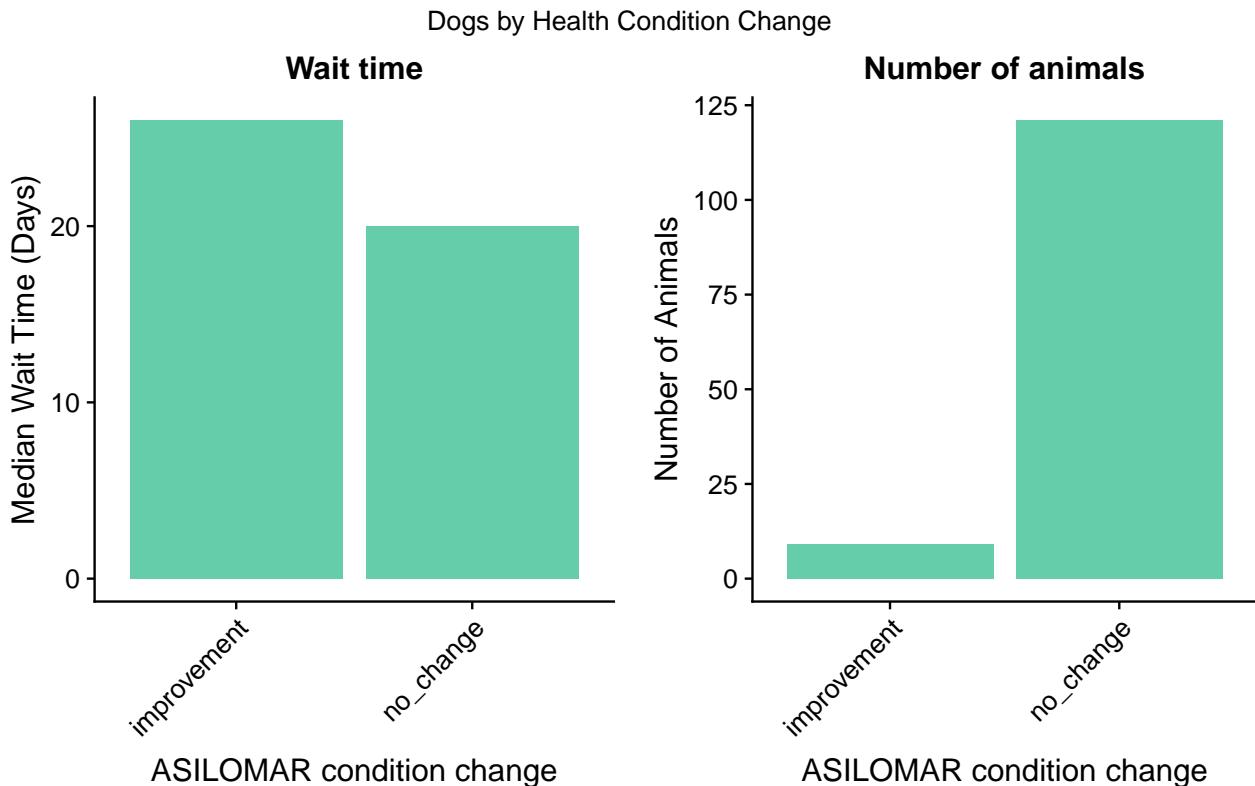
We also looked at health change from intake to outcome. Specifically, we created the following health change categories:

- no change = an animal’s intake and outcome ASILOMAR status were the same
- decline = an animal’s outcome ASILOMAR status changed from Healthy to Treatable or Unhealthy; or from Treatable to Unhealthy
- improvement = an animal’s outcome ASILOMAR status changed from Treatable or Unhealthy to Healthy; or from Unhealthy to Treatable.

For cats, we observed that the wait time is longest for those cats that experience an improvement in overall condition while at PAWS (78 days), although these cats represent only 10% of the total cats at PAWS. Cats that experienced no change in condition while at PAWS (89%) had a median wait time of 47 days.

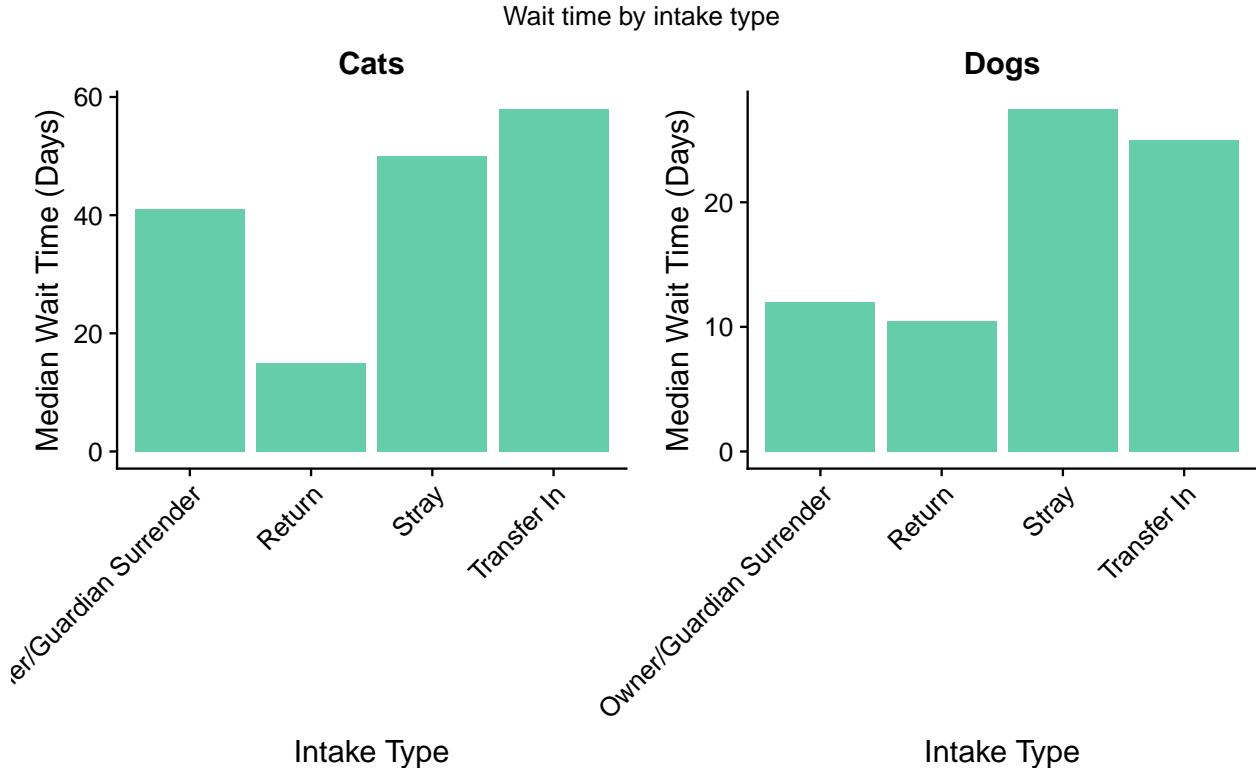


For dogs, we observed that the wait time is also longest for those dogs that experience an improvement in overall condition while at PAWS (26 days), although these dogs also represent only a small percentage (7%) of the total dogs at PAWS. Dogs that experienced no change in condition while at PAWS (93%) had a median wait time of 20 days.



3. Intake Characteristics

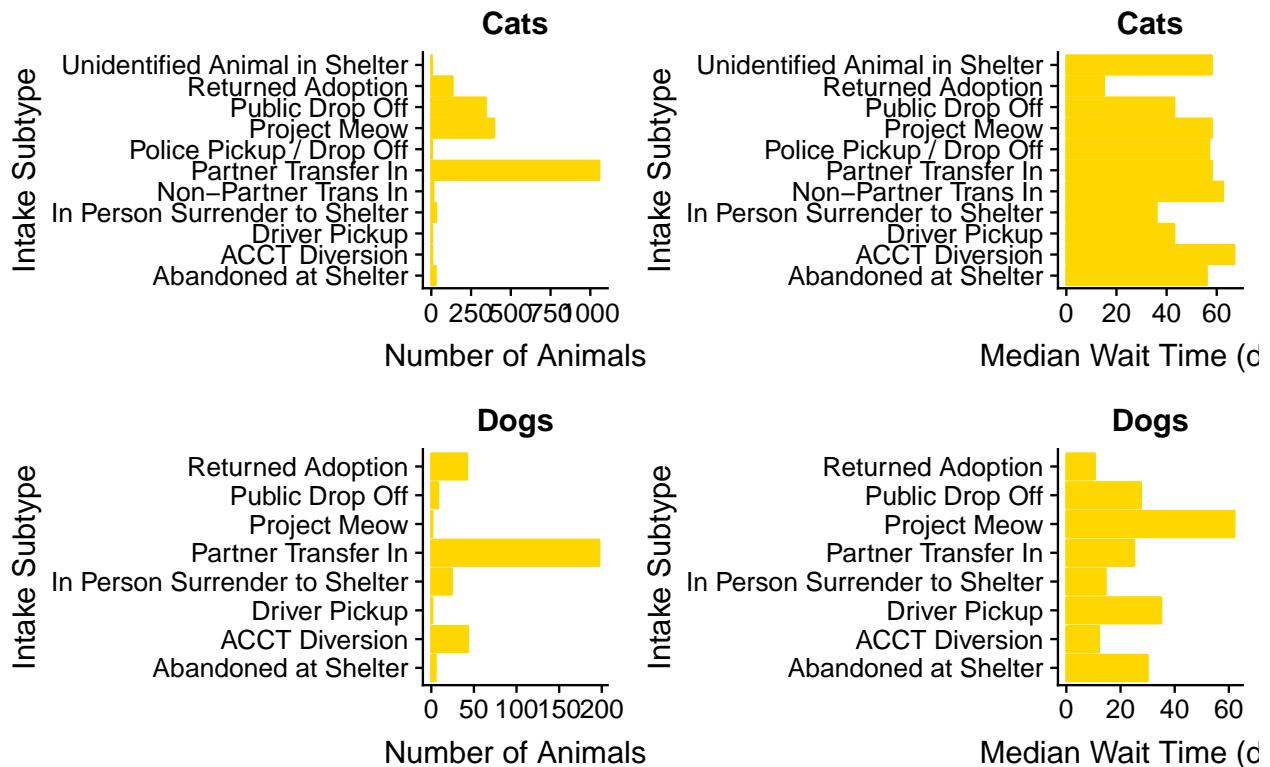
We examined intake characteristics across the petpoint dataset, by species:



We found that cats had longer wait times for each intake type relative to dogs. Furthermore, for both cats and dogs, strays and transfers in had the highest wait times. Interestingly, for cats (but not dogs) that were surrendered by their owner, these also had a high wait time (40 days).

And we examined whether the effect of intake type on wait time differed by animal species:

We next examined the intake subtype for all animals:



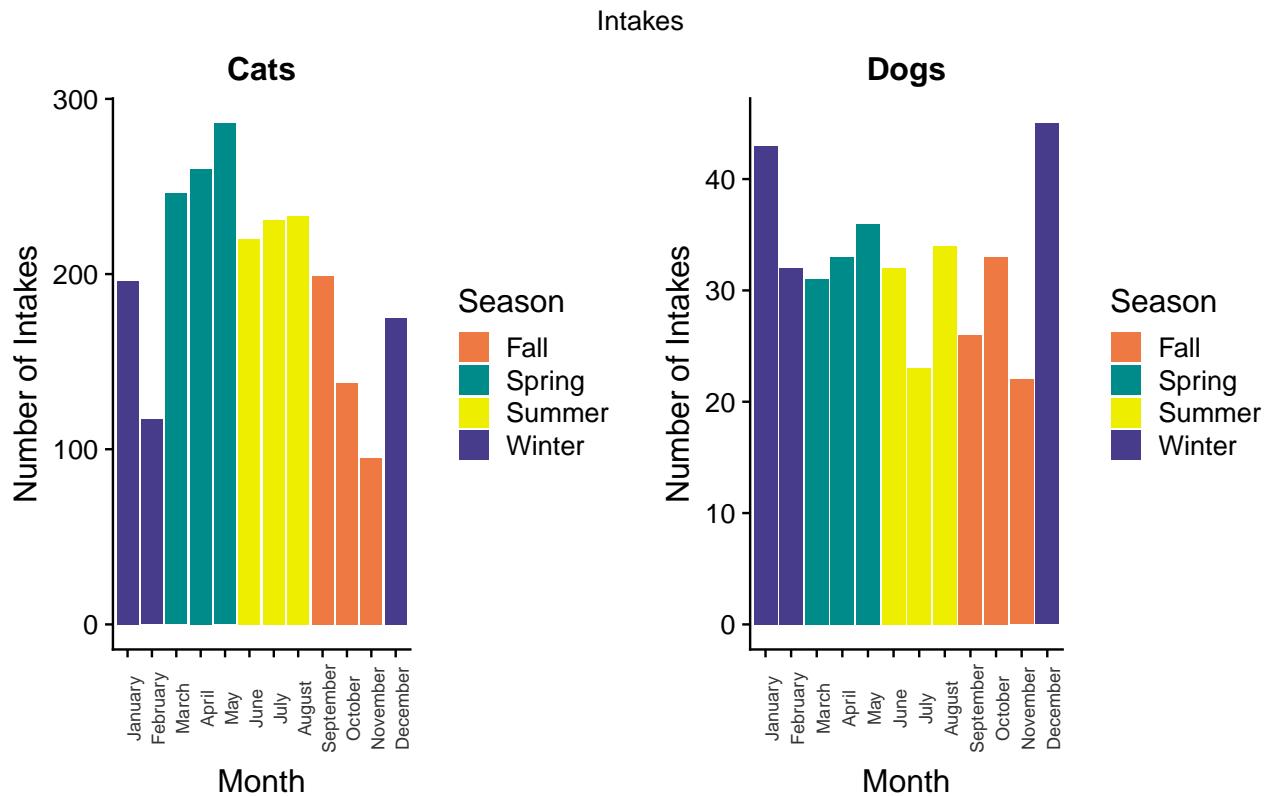
We found that for both cats and dogs, partner transfers in accounted for the vast majority of the intakes. However, differences were observed for the next highest intake subtypes: outside of partner transfers in, the majority of cats arrived at PAWS via Project Meow (395 cats, or 19.7%), Public Drop_off (342 cats, or 17%), and Returned Adoption (134 cats, or 7%). Meanwhile, for dogs that did not arrive at PAWS via Partner Transfer In, the majority were ACCT Diversion (43 dogs or 13%), Returned Adoptions (42 dogs or 13%), and In-Person Surrender to Shelter (24 dogs or 7%).

It is also worth noting that for both dogs and cats, returned adoptions experienced the speediest re-adoption (15 days median wait for cats, and 10 days median wait for dogs). The next intake subtypes that were associated with shortest wait times for cats were In-Person Surrender to Shelter (36 days), while for dogs these were ACCT Diversion (12 days) and In Person Surrender to Shelter (14.5 days).

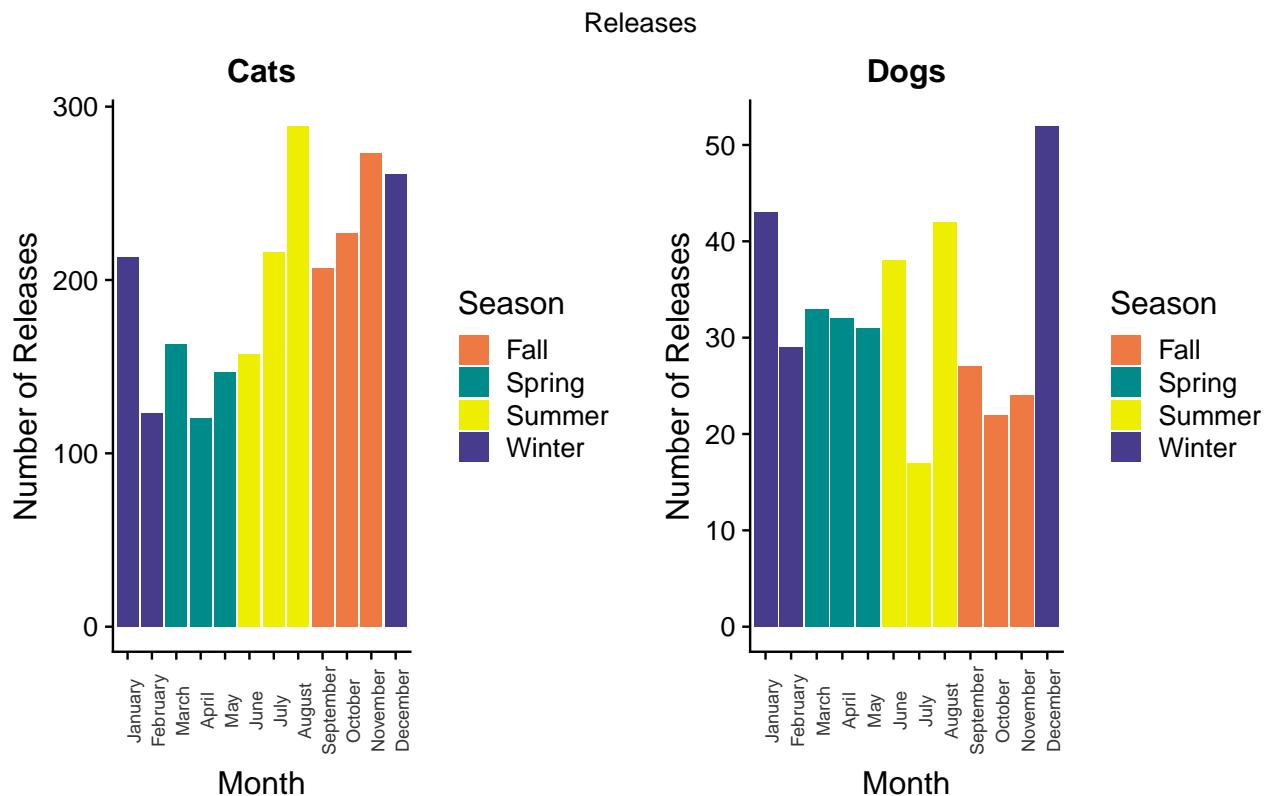
We found one record of a dog whose intake subtype was “Project Meow”, which may have been an input error - but we included it in the graph nonetheless.

4. Seasonal and Locational Patterns

Last, we examined animals’ wait time at PAWS by season and species. We examined the frequency of intakes and releases per month:

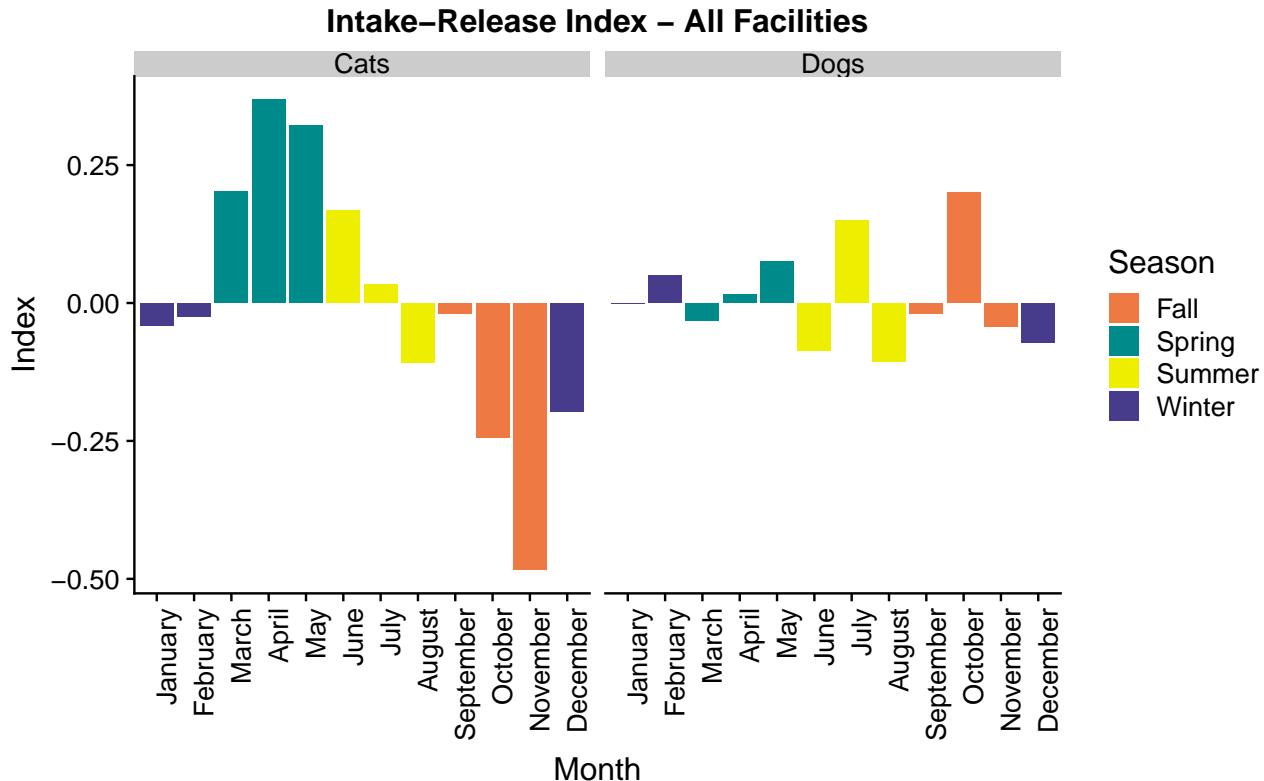


It appears that intakes are highest for cats in spring (peaking in May) and summer, and lowest in February and November. Meanwhile for dogs, the highest intakes are in winter (December and January), and the lowest in July and November.



On the other hand, for cats, releases from PAWS are highest in August, November and December, while for dogs the highest numbers of releases occur in December, January, August and June.

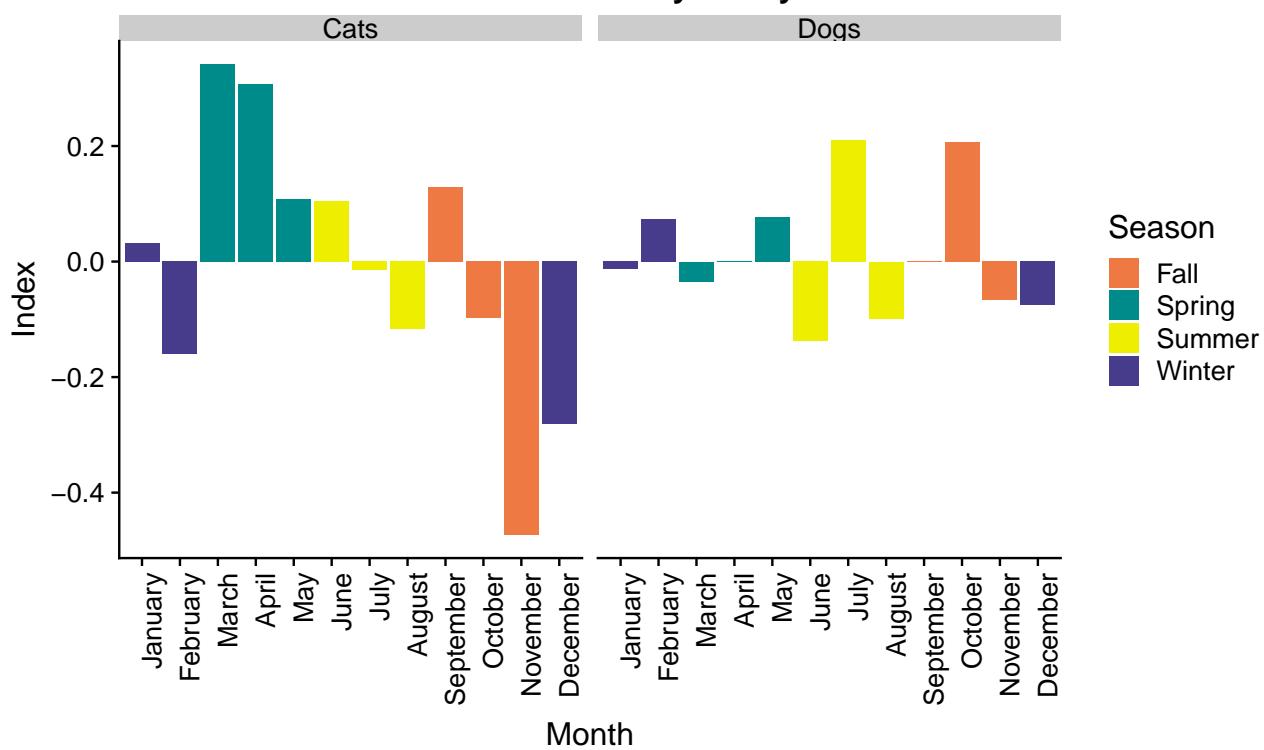
We then looked at monthly trends in intakes and releases by species. Specifically, the graphs below show the difference between total number of intakes and releases by month, divided by the total number of intakes and releases per month. The resulting bars can be interpreted as follows: - if >0 , the number of intakes is greater than the number of releases - if <0 , the number of releases is greater than the number of intakes - the closer to 0 a bar, the more evenly distributed intakes and releases were in that month



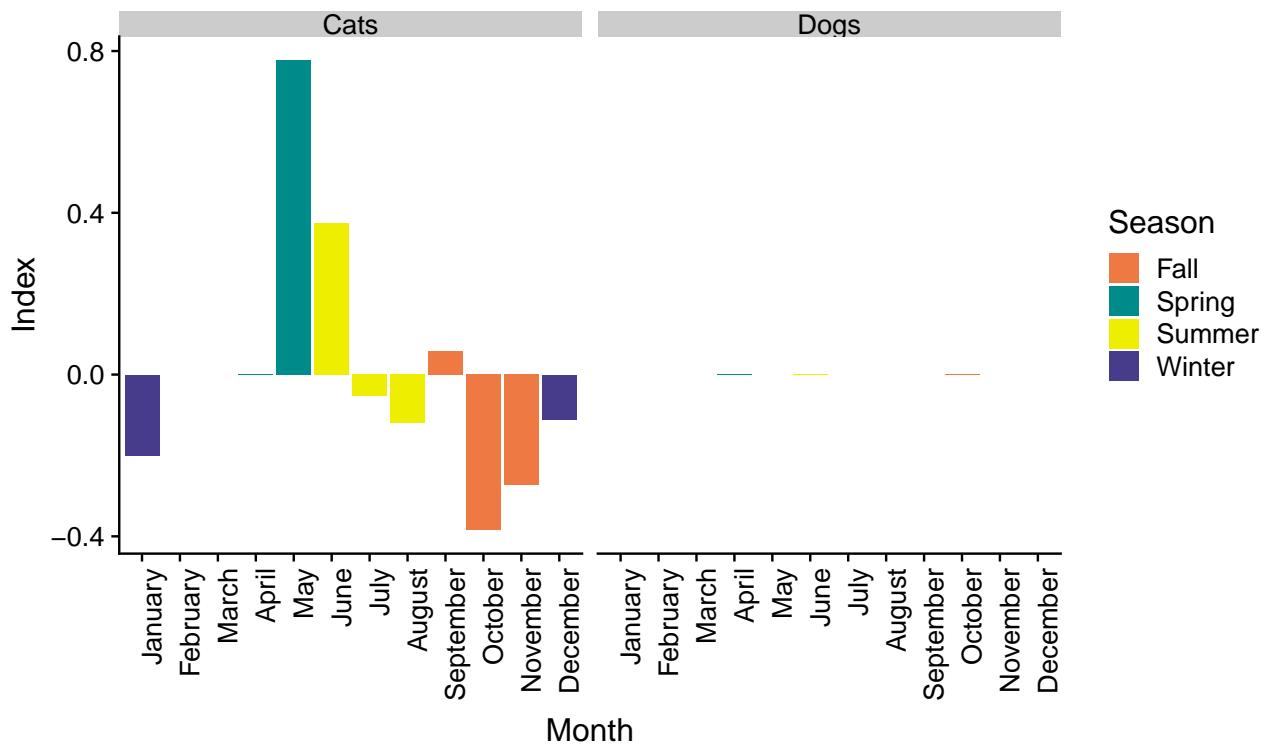
We found that cats have higher intakes than releases in March - June, while more releases than intakes were observed in October-December. For dogs, June and September represent spikes in intakes compared to releases. These findings may be useful for volunteer resourcing considerations during these times.

We also examined seasonal patterns in intake-release index by location for each species. It is important to note here that these patterns only take into account the location of intake and the location at time of release, and do not track an animal's transfers within PAWS locations. Therefore, the location-specific results should be interpreted with caution.

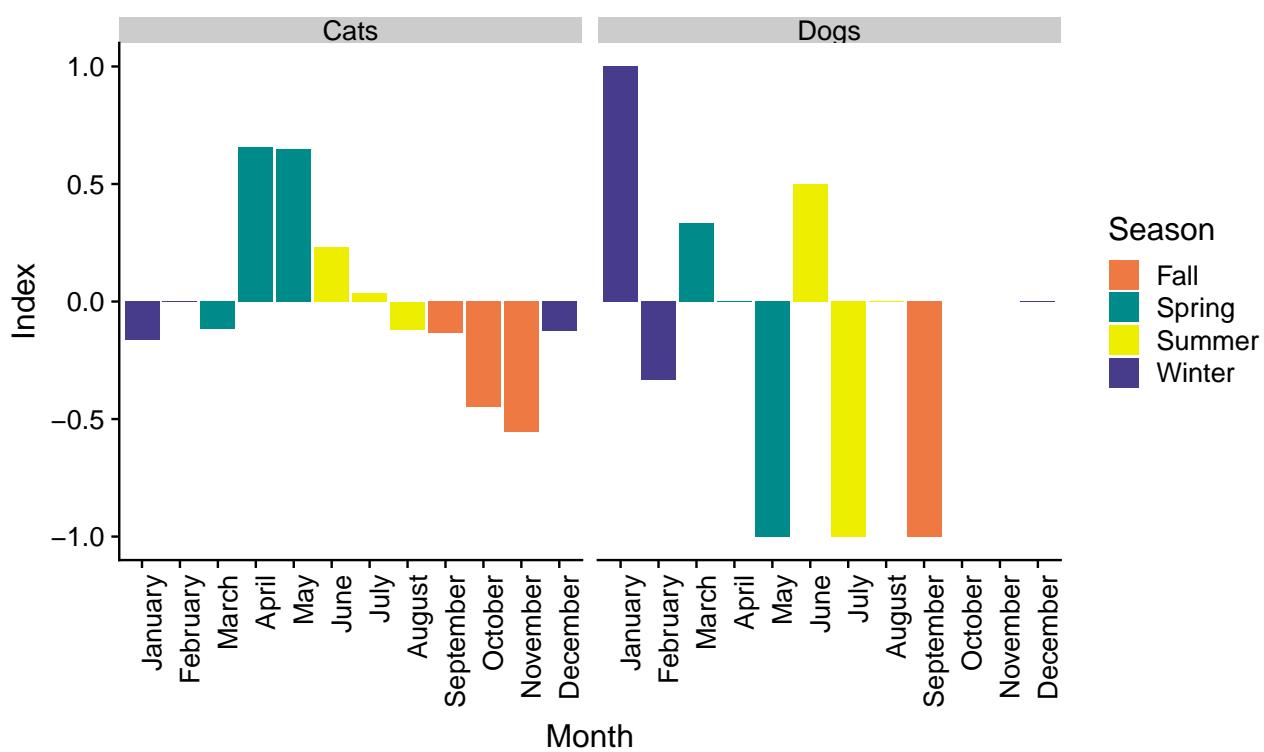
Intake–Release Index – Grays Ferry Avenue



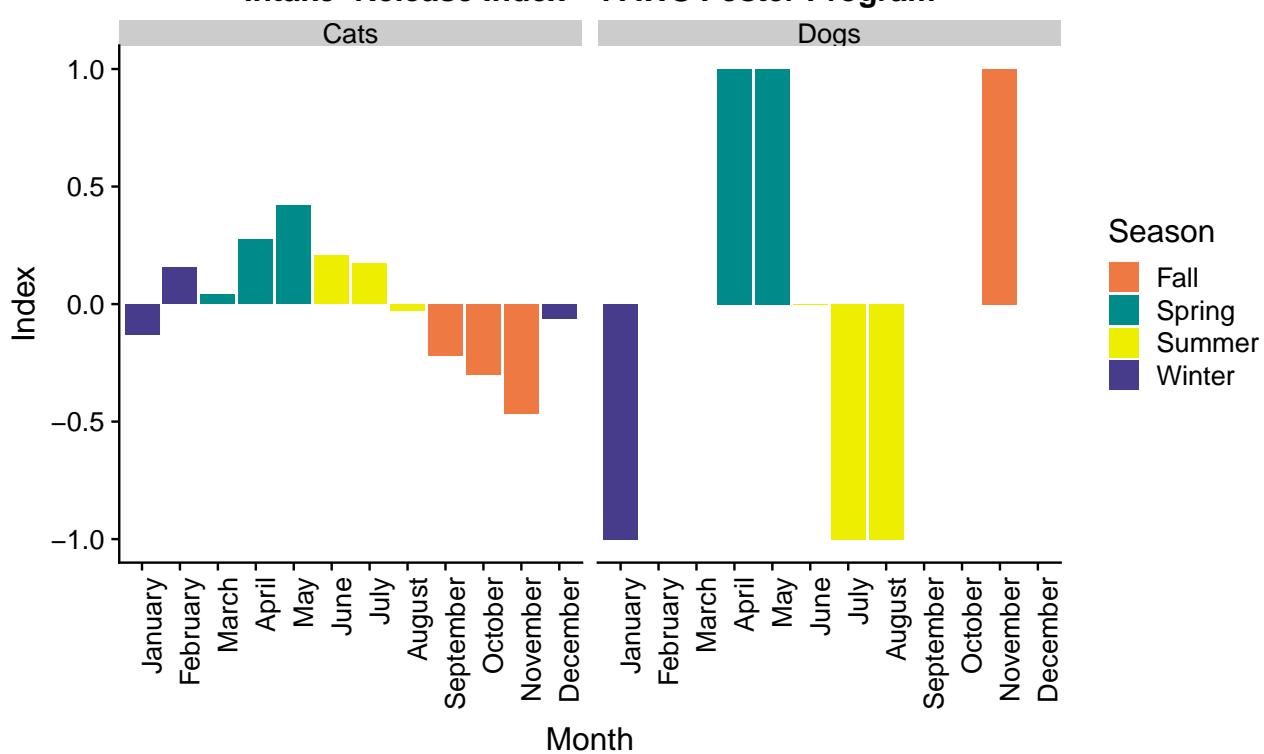
Intake–Release Index – PAC

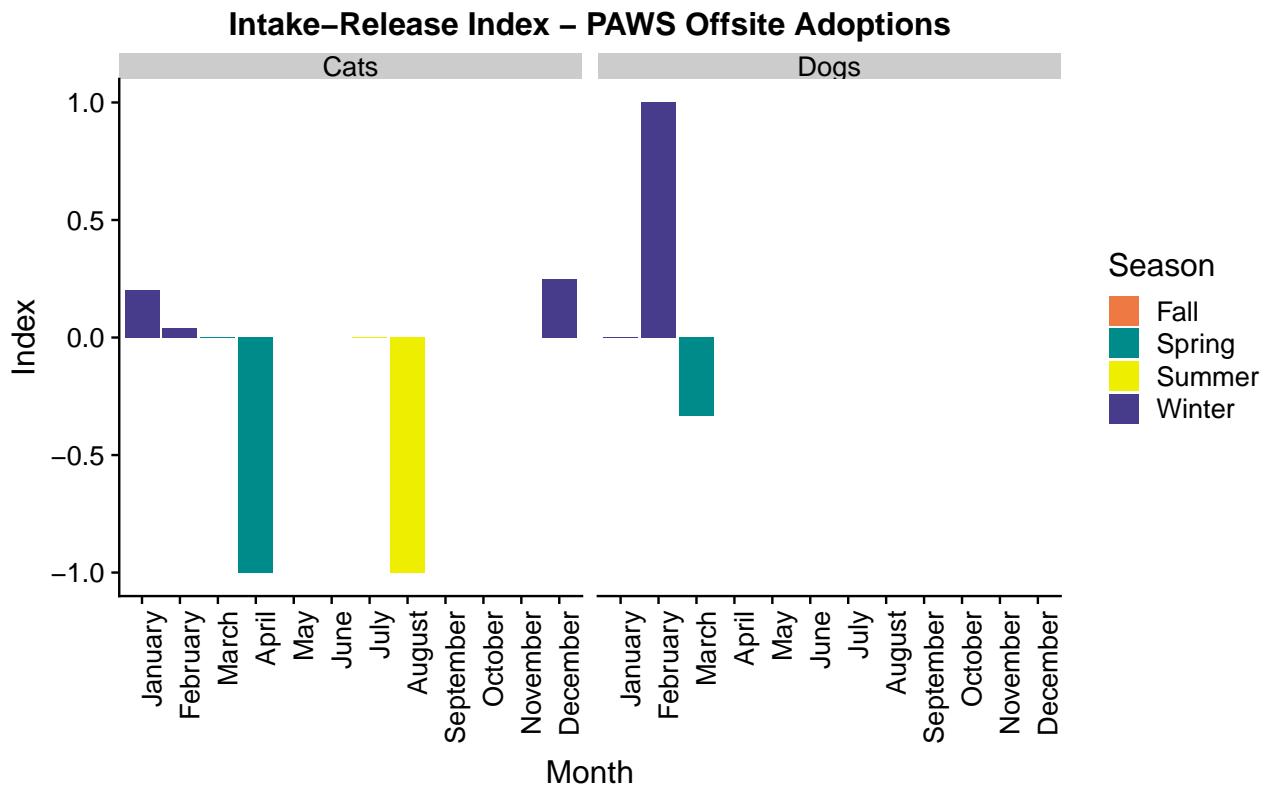


Intake–Release Index – Grant Avenue

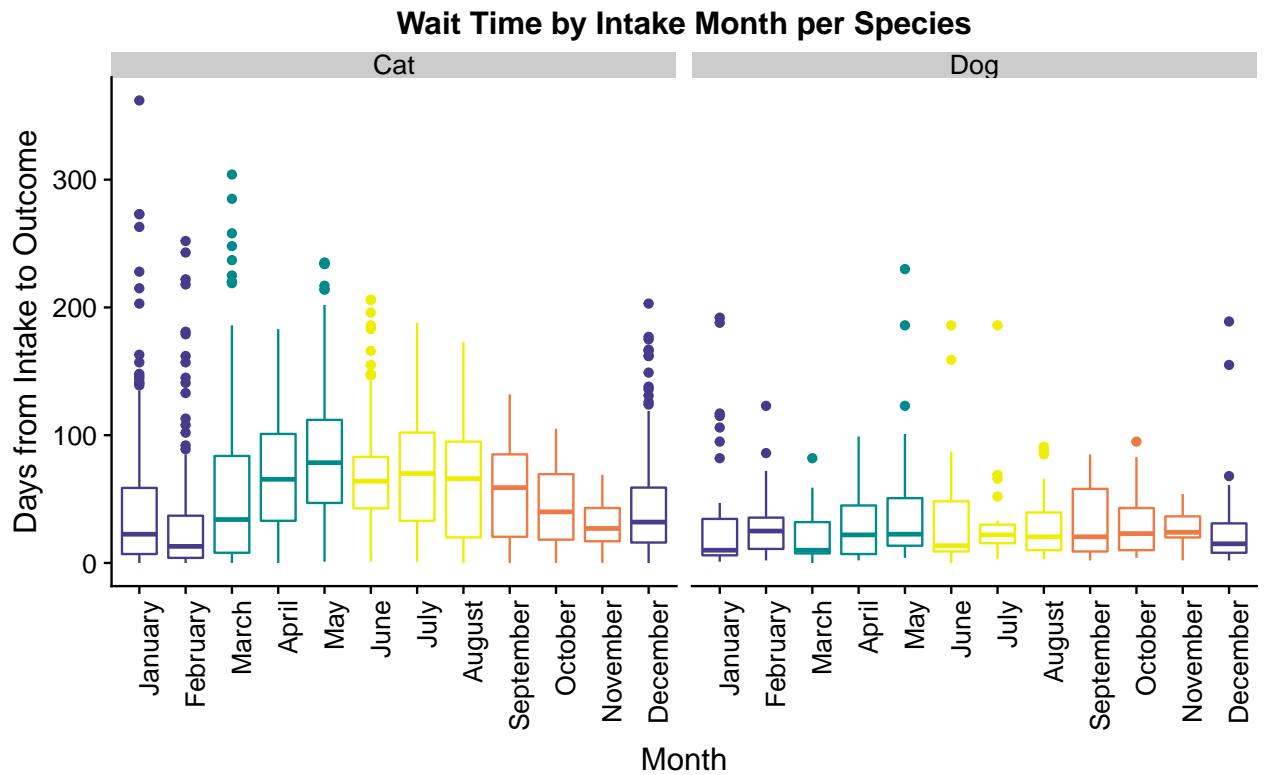


Intake–Release Index – PAWS Foster Program

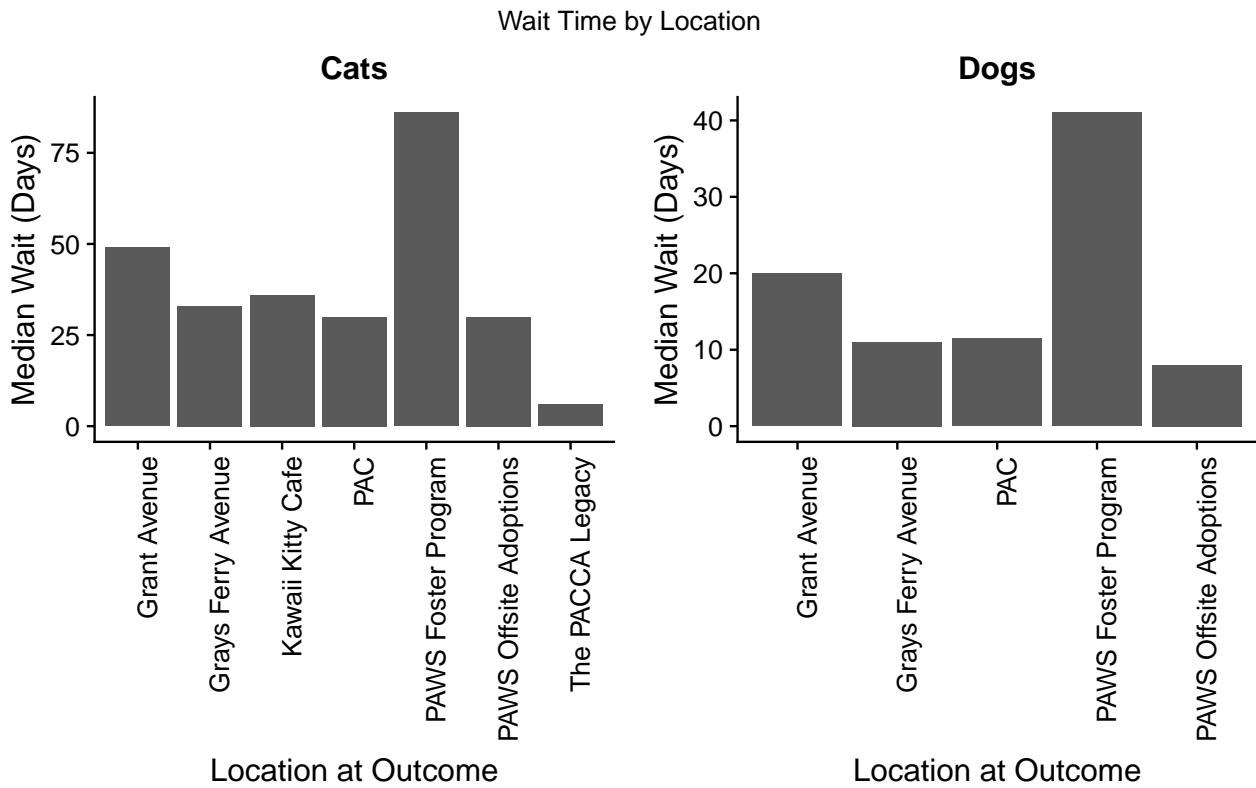




We then examined seasonal patterns in wait time, by species:

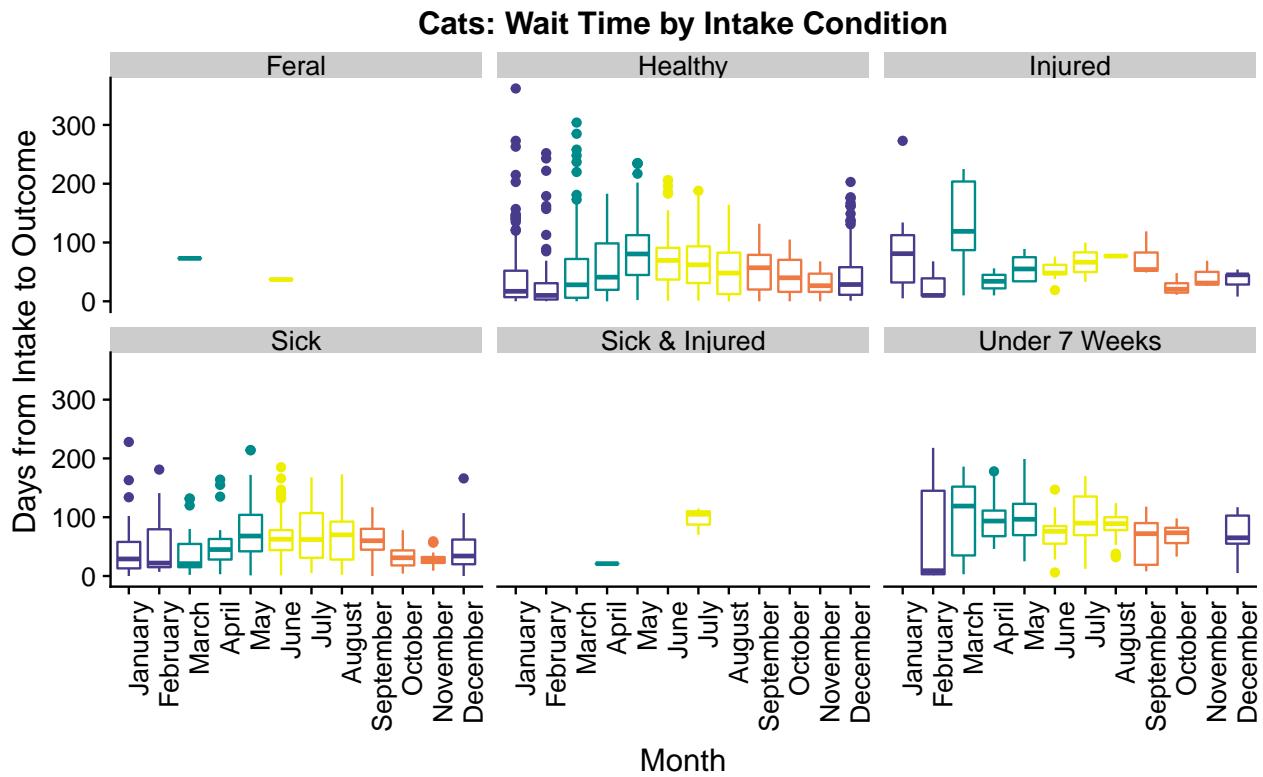


We also estimated wait time by location for each species.

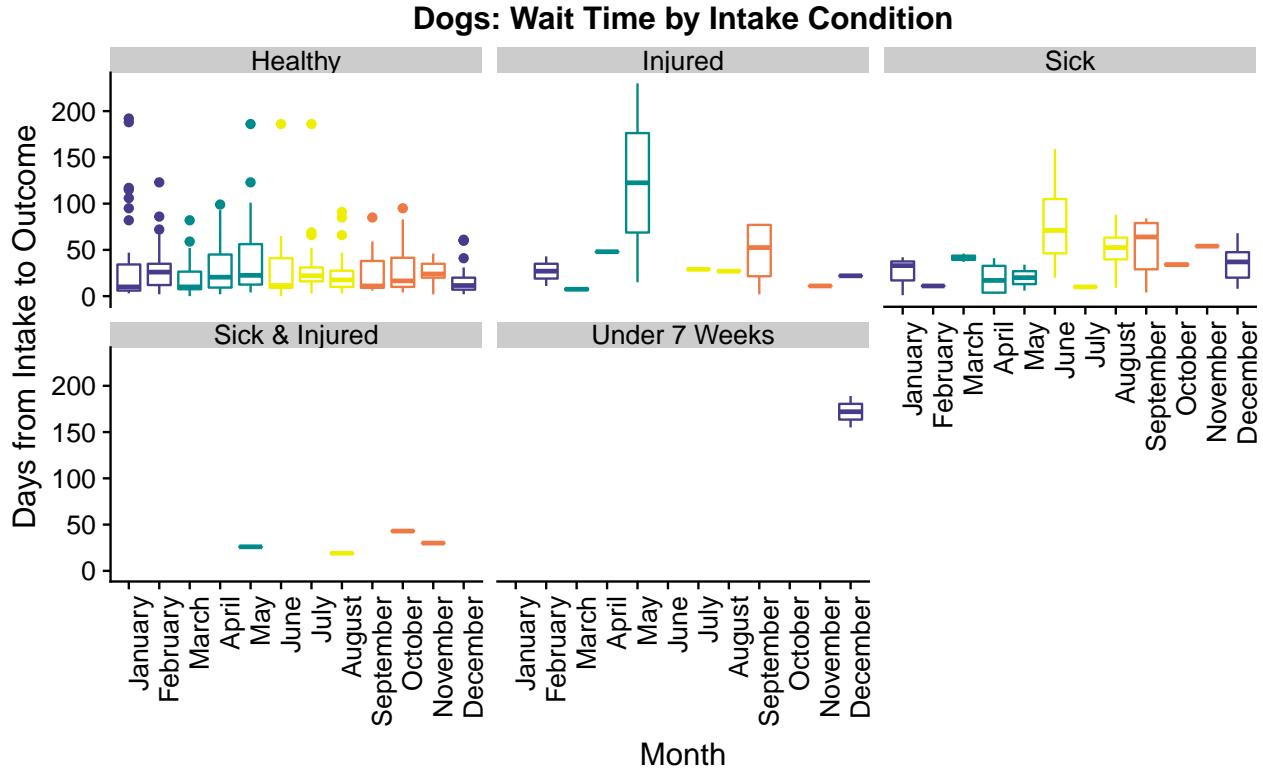


Our findings demonstrate that cats adopted from PAC or from Offsite Adoptions spend the shortest time at PAWS, while cats in the Foster Program spend the longest time at PAWS. Similarly, dogs adopted at Offsite Adoptions spend the shortest time at PAWS, while dogs in the Foster Program spend the longest time at PAWS. Interestingly, for both cats and dogs, animals adopted from Grant Avenue spend longer in the shelter than animals adopted from any other PAWS location.

Last, we examined wait time by intake health condition for each species.



For cats, we found that animals classified as 'Sick', 'Injured', and 'Under 7 Weeks' tended to have longer wait times in the spring and summer months.



For dogs, we found that animals classified as 'Injured' and 'Sick' also tended to have longer wait times in the spring and summer months.

Conclusions and Next Steps

In conclusion, our analyses revealed that animal characteristics, intake characteristics, and seasonal and locational patterns contribute to animals' wait times at PAWS in the 2018 year. Almost all animals' outcomes from 2018 were adoptions, meaning that PAWS is fulfilling its goal of finding homes for needy animals in the Philadelphia area. The median wait time for cats (51 days) was longer than the median wait time for dogs (18 days), and this was likely due to 1) greater number of cats vs dogs, and 2) longer wait times for sick, young cats in the spring and summer months. We also observed differences in wait time by PAWS location for each species, but this is likely due to the number and species of animals at each location. Overall, our findings indicate that PAWS may want to focus resources on young and sick cats in the spring and summer months in order to reduce wait times.

2. Application Trajectories

This analysis investigated factors relating to an application's trajectory at PAWS - from the time when an applicant submits an application, through the validation checks done by PAWS, to the time the application is marked as complete. The dataset analyzed here included trello IDs (1594 unique trello IDs) and their matched application data from the online application form (applications submitted dates between 2018-08-30 to 2018-12-31).

Contributors

- **Ramaa Nathan** (group leader) is an aspiring data scientist with a PhD in Computer Science and an ongoing masters in Applied Statistics. Her background is in finance and healthcare.
- **Kate Connolly** is a digital analyst at the Philadelphia Inquirer where she helps to maintain the analytics framework and to provide data-driven support and decisions across the organization.
- **Veena Dali** is a senior business intelligence analyst at Comcast working to provide data solutions to support business decisions. Her background is in Neuroscience and Computer Science.
- **Amy Goodwin Davies** is a data scientist with a background in psycholinguistics.
- **Brendan Graham** is a clinical data analyst at The Children's Hospital of Philadelphia with a background in applied statistics.
- **Ambika Sowmyan** heads the Marketing data analytics group at Hartford Funds. Her background is in Finance and Retail and has a graduate degree in Management and Predictive Analytics.

Summary

- For applications that resulted in adoption, cat applications took longer to process than dog applications. Cat apps took 19 days and dog apps took 8 days on average.
- Based on the animal's outcome site, adoption times were faster at PAWS Foster Program & PAWS Offsite Adoptions locations.
- Singles seem to prefer to adopt a pet.
- There were 12 denied applications and 133 red flagged
- For the denied applications, the applicants had no known allergies and many of them had unfortunate incidents with prior pets
- There was a lot of missing data in the applications especially for the home pet policy question.
- We recommend redesigning the application to enforce standardized, limited, and logical responses. For example, for certain questions, allow only a single response or provide a drop down menu. Doing so will save PAWS staff time when reviewing applications and help with future analyses.

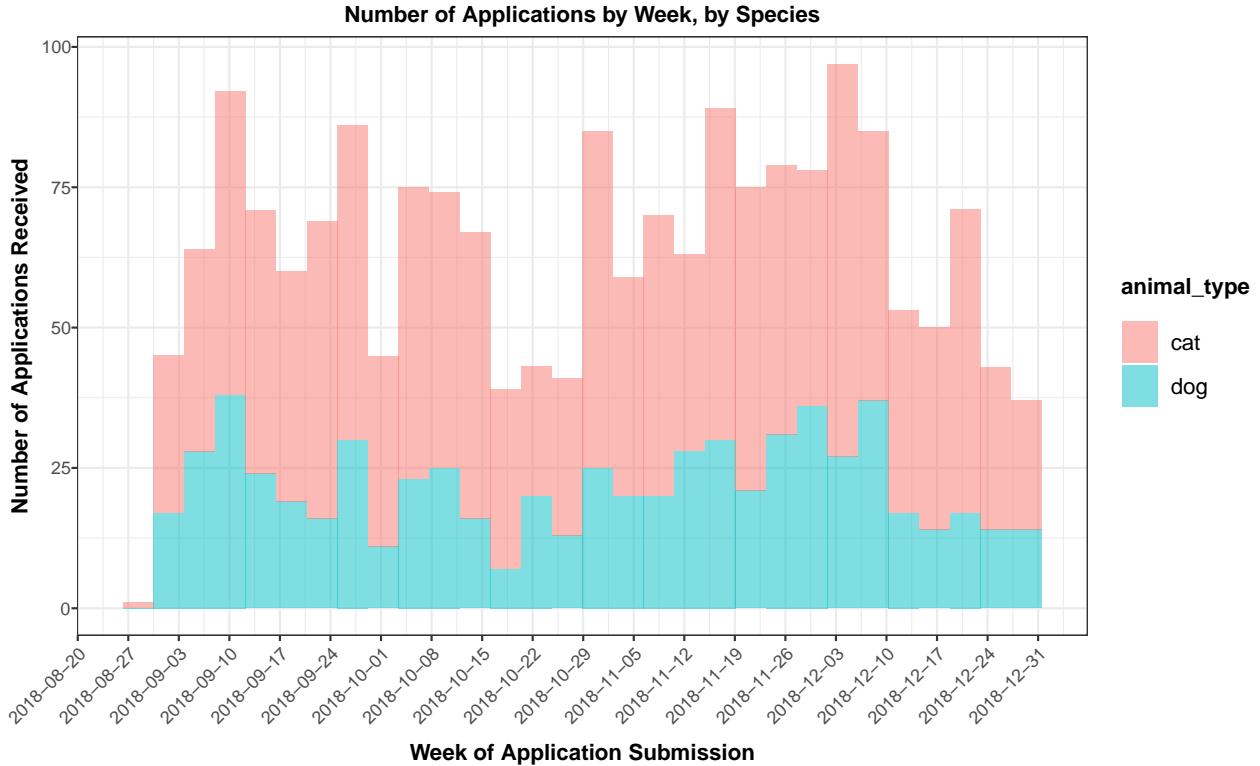
Data Pre-processing

As our group focused on questions about application trajectories, our starting point was an applications dataset comprised of online dog and cat applications (`dog_apps.csv` and `cat_apps.csv`). Several data pre-processing steps are important to highlight:

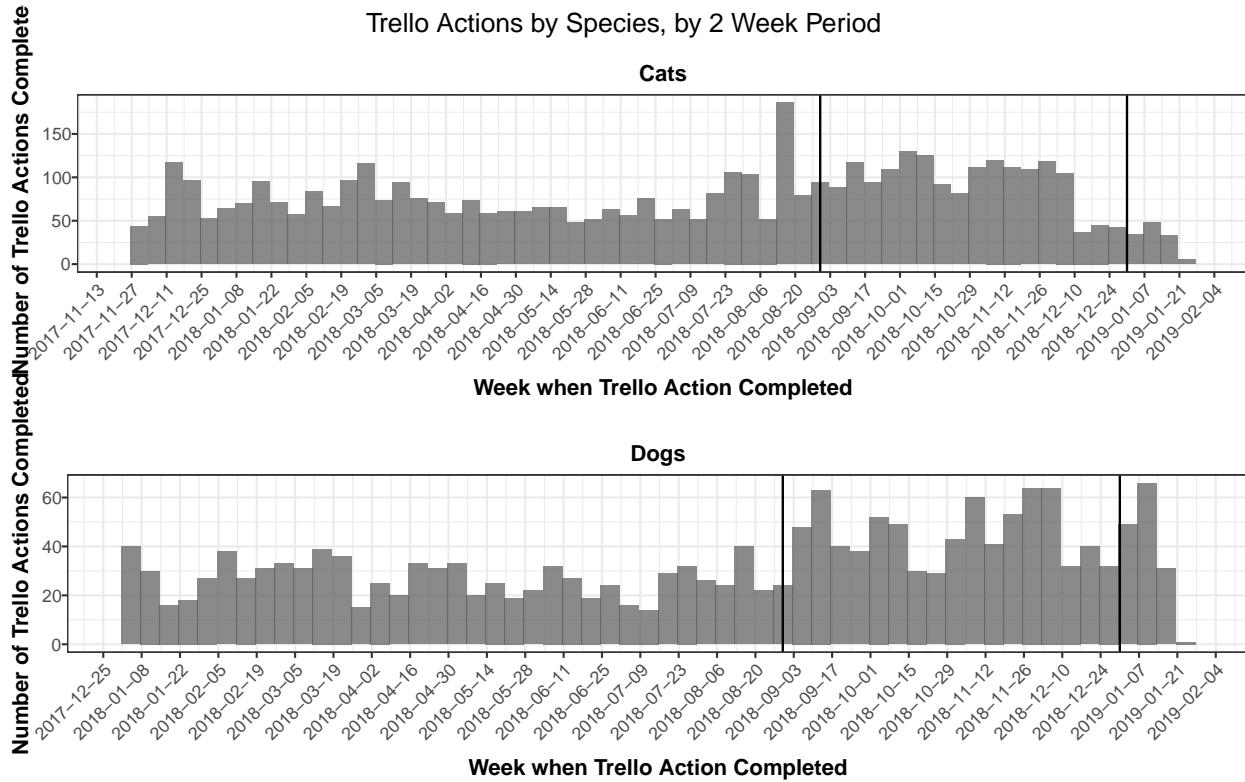
- We standardized some responses that differed between cat and dog applications but that we believe had the same meaning (e.g. `ideal_adoption_timeline` had responses “next-few-weeks” and “few-weeks” which we standardised as one response (“few-weeks”). (See further discussion of this issue in the *Data Issues affecting Analyses* section)
- We replaced text values for `children_in_home` and `adults_in_home` to our best interpretation of what the responder meant (e.g. “-2” we interpreted as “2”) and considered values that were excessive (e.g. 15 children in the home) as missing data.

- Due to a high variability of responses on monthly and emergency budgets, we capped these responses at \$10,000 (monthly) and \$20,000 (emergency). Values higher than these were rounded down to the maximum.

Our cleaned applications dataset contained 1906 rows, 1594 unique Trello IDs, and the submitted dates ranged from 2018-08-30 to 2018-12-31.



To our applications dataset, we added fields from Trello actions and cards. A card contains information about when the application started being processed by PAWS and what label PAWS determined it should have (e.g. “ready to adopt”, “red flag”, etc.), and Trello actions contained information about the steps in the background check process (e.g. vet check, Pet Point check, landlord check, etc.). We also matched applications with PetPoint outcome (`petpoint.csv`) to be able to analyze successful applications. One issue we encountered was that the date range for the applications dataset (123 days) was considerably smaller than the actions and petpoint datasets (417 and 413 days respectively), which was a result of when the online applications were re-initialized in the PAWS website. Similar data-preprocessing steps as we took for the applications dataset were taken for the actions, cards, and petpoints datasets.



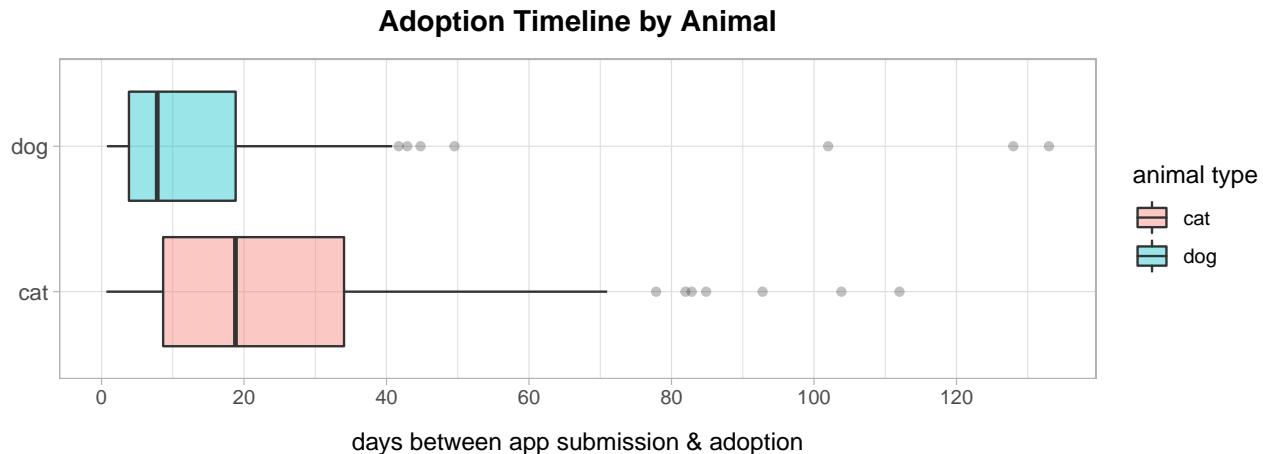
We noticed that for both cat and dog applications, there was an increase in Trello actions that correlated with the re-introduction of online applications at the end of August 2018. At the minimum, this indicates an increase in tracking/recording overhead that coincided with the re-introduction of online applications.

Analysis of Time in Processing Applications

How Animal & Outcome Site Influence Application Timelines

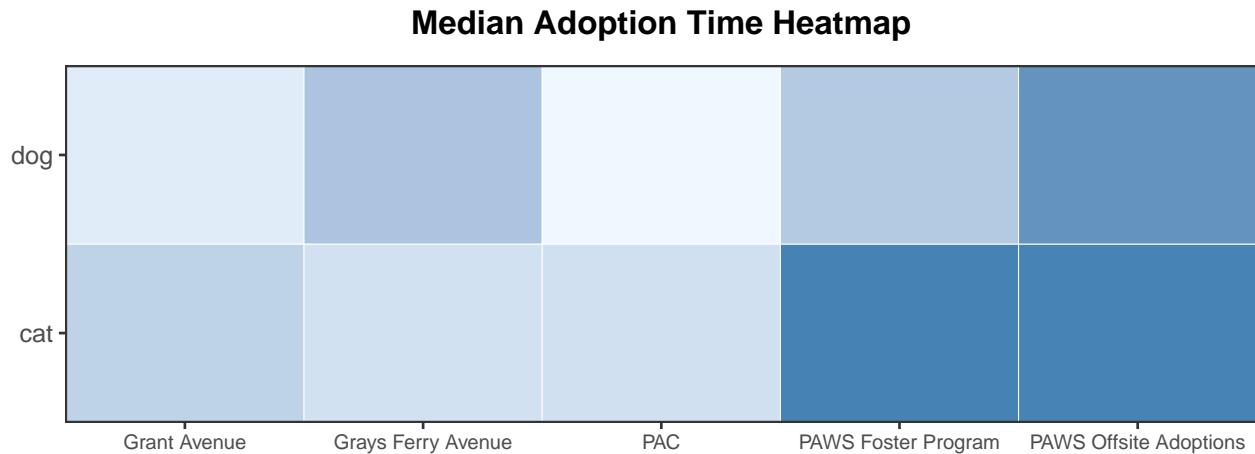
Application timelines were measured by taking the difference between the time an application was submitted and the time that application resulted in an adoption. Only applications that resulted in adoption were assessed; applications that were denied were not included in the analysis. This is a potential area of further investigation.

In general, cat applications typically take longer than dog applications. The chart below shows that the median adoption timeline for **cats** is approximately **19** days (vertical black line inside red box), while **dog** applications average about **8** to result in an adoption (vertical black line inside blue box).



The chart also illustrates that for longer-than-average application timelines, animal type may influence just *how much longer* those above-average timelines are. Of the longer-than-usual applications, cat ones took between 35 days and 70 days compared to about 18 days to 40 days for dogs.

The outcome site for an adoption also influences the timeline of an application. It's important to note that this analysis does not consider all the potential locations that an animal spent its time during the application process; it is strictly based on the animal's outcome site.



outcome_sitename	animal_type	n	median adoption time (days)
Grant Avenue	cat	74	10
	dog	19	6
Grays Ferry Avenue	cat	2	8
	dog	18	13
PAC	cat	70	8
	dog	17	4
PAWS Foster Program	cat	187	25
	dog	20	12
PAWS Offsite Adoptions	cat	44	25
	dog	1	22

From the heatmap and table above, it's clear that overall median processing times for applications that resulted in adoptions were higher at PAWS Foster Program & PAWS Offsite Adoptions locations. This is

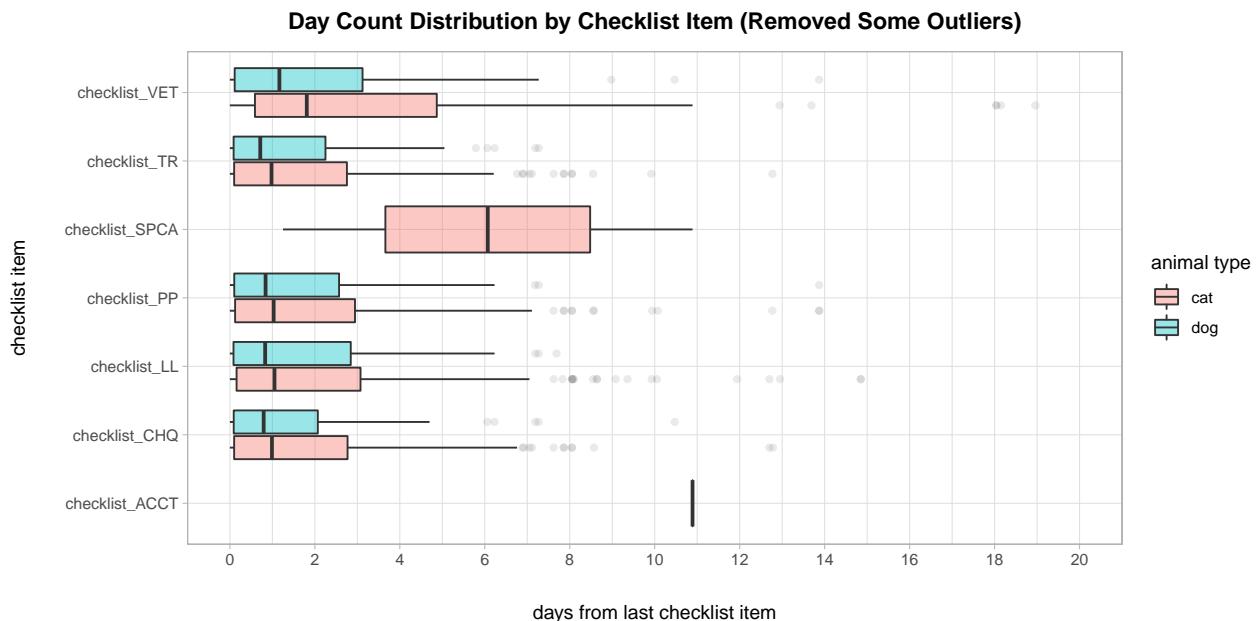
especially true for cat applications at those places.

Based on median values, here are the fastest & slowest time-to-adoption sites:

- **Cats**
 - Slowest: PAWS Foster Program
 - Fastest: Grays Ferry Avenue
- **Dogs**
 - Slowest: PAWS Foster Program
 - Fastest: PAC

Only one site had a higher median adoption time for dogs than for cats—Grays Ferry Avenue. This site also had the fewest cat adoptions, though ($n=2$). It's also important to note the small n size for dog apps at PAWS Offsite Adoptions ($n=1$).

How Animal & Outcome Site Influence Application Checklist Items



Most application items took between one and two days (median) to complete. While the animal type and outcome site didn't significantly impact the individual item times, cat applications generally exhibited slightly longer times between checklist items. Cat applications averaged about **1.2** days between checklist item, compared to **0.9** for dogs (excluding SPCA & ACCT items). The VET checklist item had the greatest difference between animals, while modest, could contribute to longer submission-to-adoption times for cat applications.

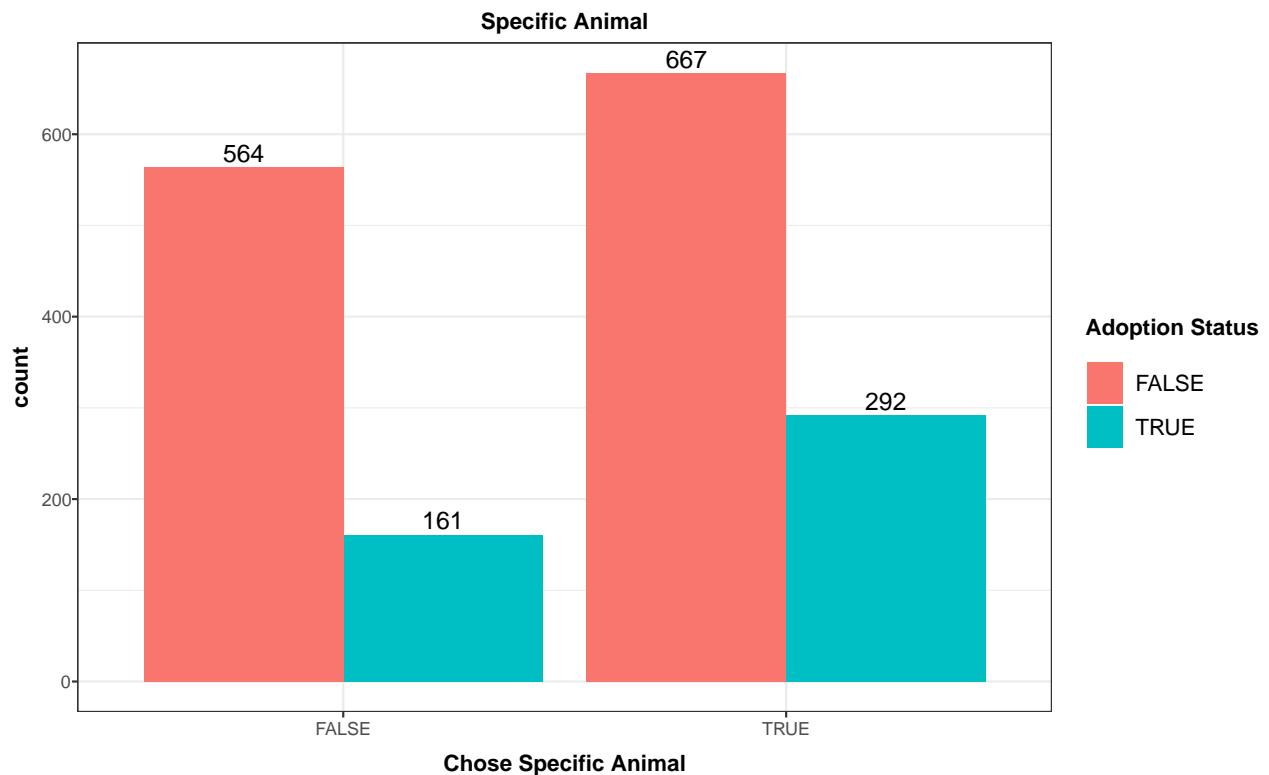
The chart above removed significant outliers, but further inspection of these outliers could be valuable. Understanding what causes certain application steps to take longer could help to streamline parts of the checklist process.

checklist item	n	median days from last item	percent of cards with item checked
checklist_ACCT	1	10.89	0.2%
checklist_SPCA	2	6.07	0.4%
checklist_VET	425	1.80	93.8%
checklist_CHQ	432	0.97	95.4%
checklist_LL	433	1.03	95.6%
checklist_PP	433	1.03	95.6%
checklist_TR	435	0.95	96.0%

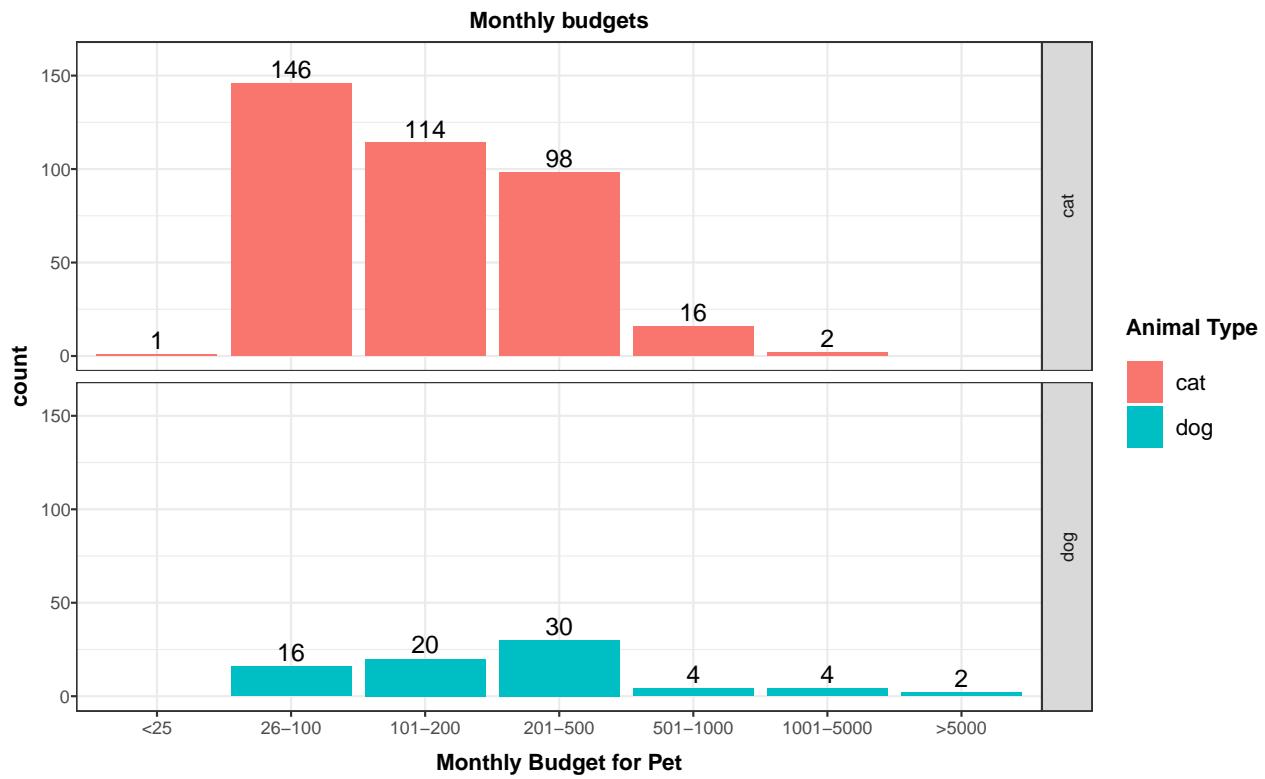
The table above shows the exceptions to the average checklist times. The ACCT and SPCA checklist items took considerably longer to complete than other items, but they also were present in less than 1% of applications. This low sample limits any sound conclusions, but does present an area for potential further exploration. It may be valuable to assess if other components of an application—like red flags or particular animal information—lead to this item being more mandatory. But more data would be needed for this analysis.

Analysis of Application Characteristics that Result in Adoption

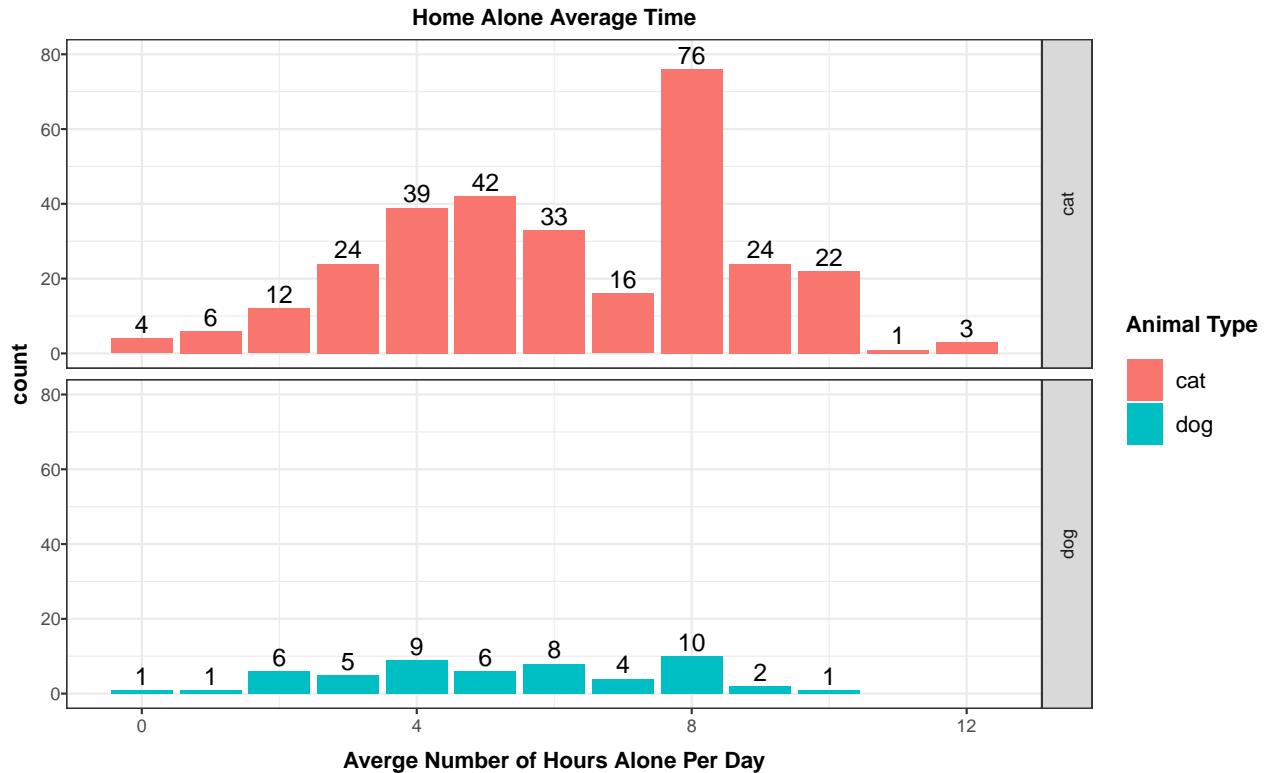
We analysed the different factors of the applications that ended with a successful adoption.



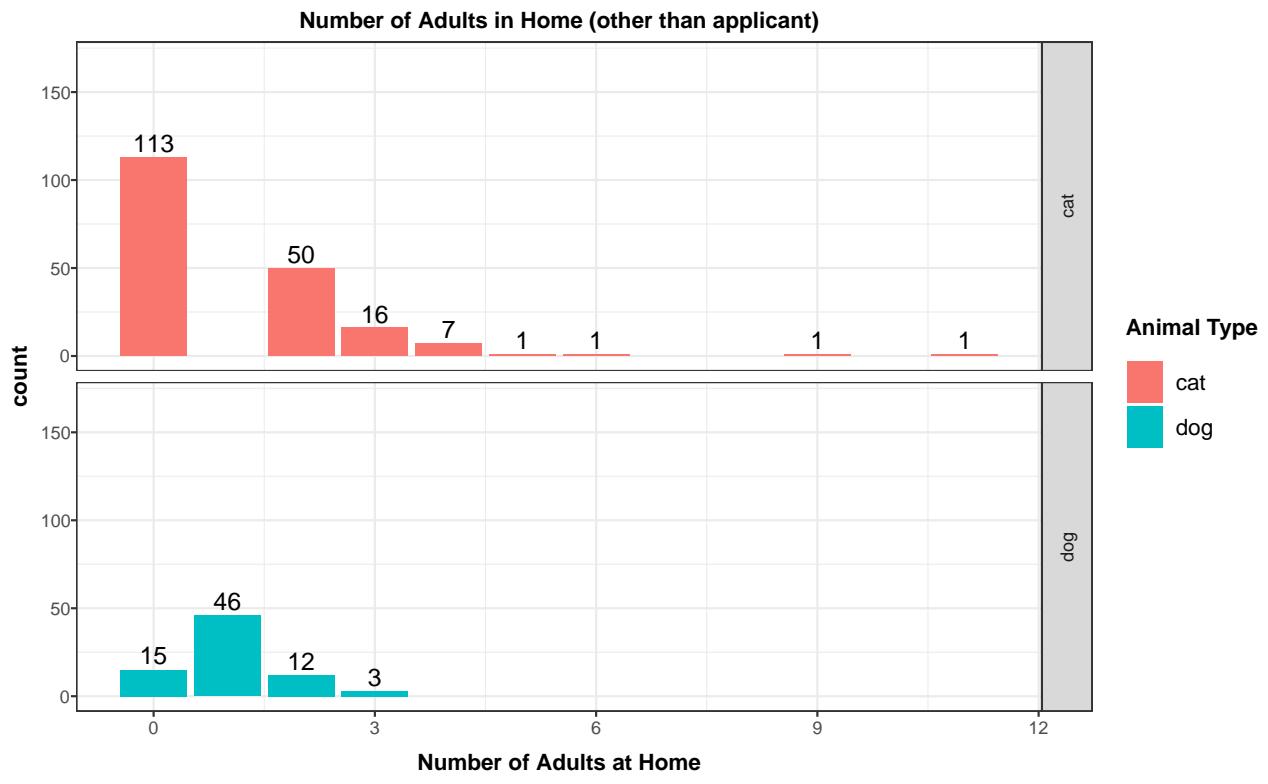
When applicants requested a specific animal, 30% of applications resulted in an adoption vs. only 22% of the applications resulted in an adoption when the applicants did not request a specific animal. This seems surprising as we would expect an applicant who is not specific about the type of animal to be able to adopt easily.



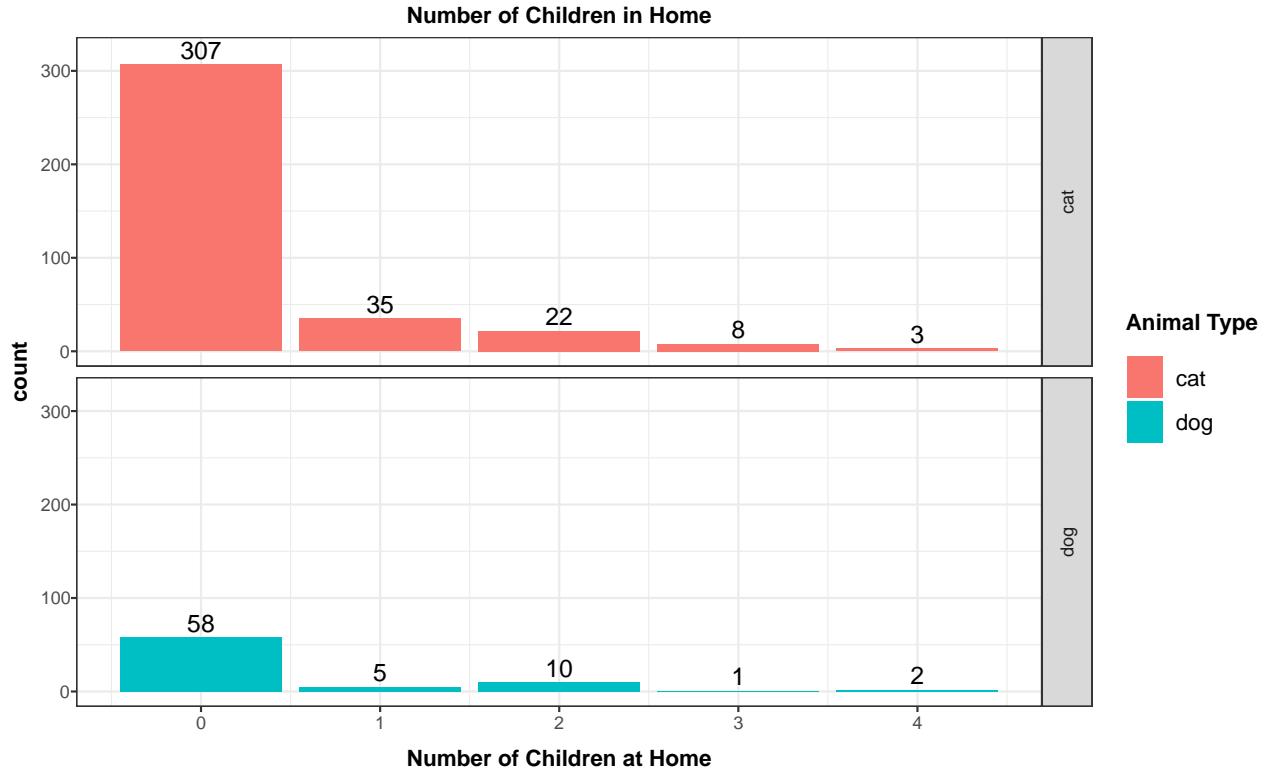
Most of the applicants who adopted a pet had allocated a monthly budget of less than \$500.



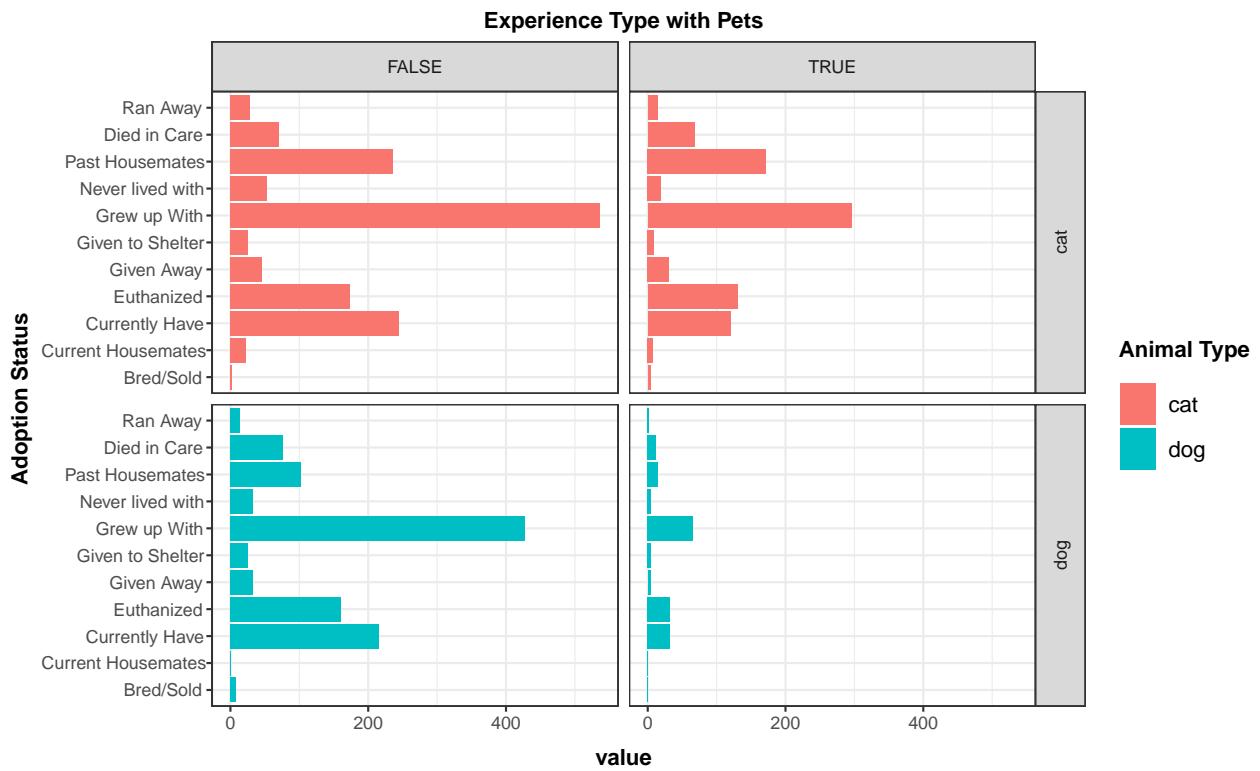
Applicants who expected to leave the animal alone at home for longer hours chose to adopt a cat. The largest number of applicants expected the animal to be alone for 8 hours, which would be typical of an applicant who works full time.



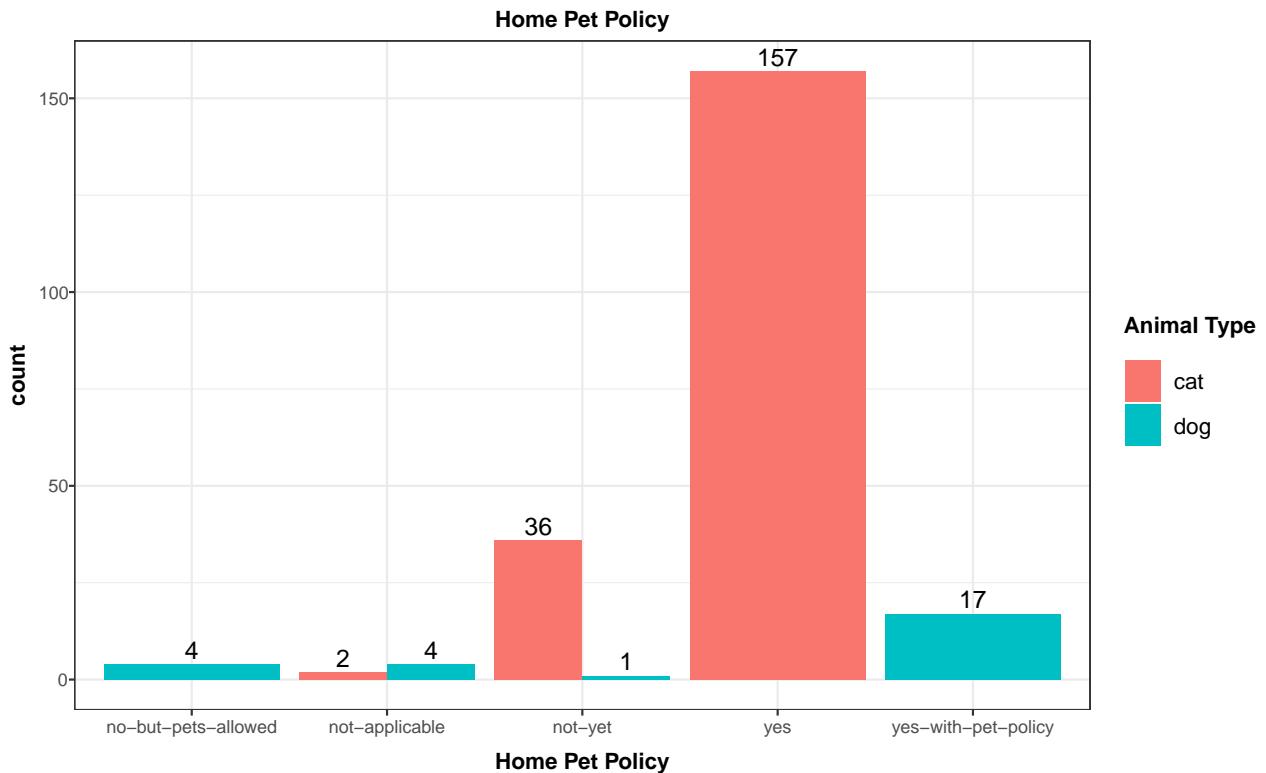
We found that the majority of applications where the applicant lived alone (0 other adults in the home) preferred cats, while the the majority of applicants who lived with 1 other person preferred a dog.



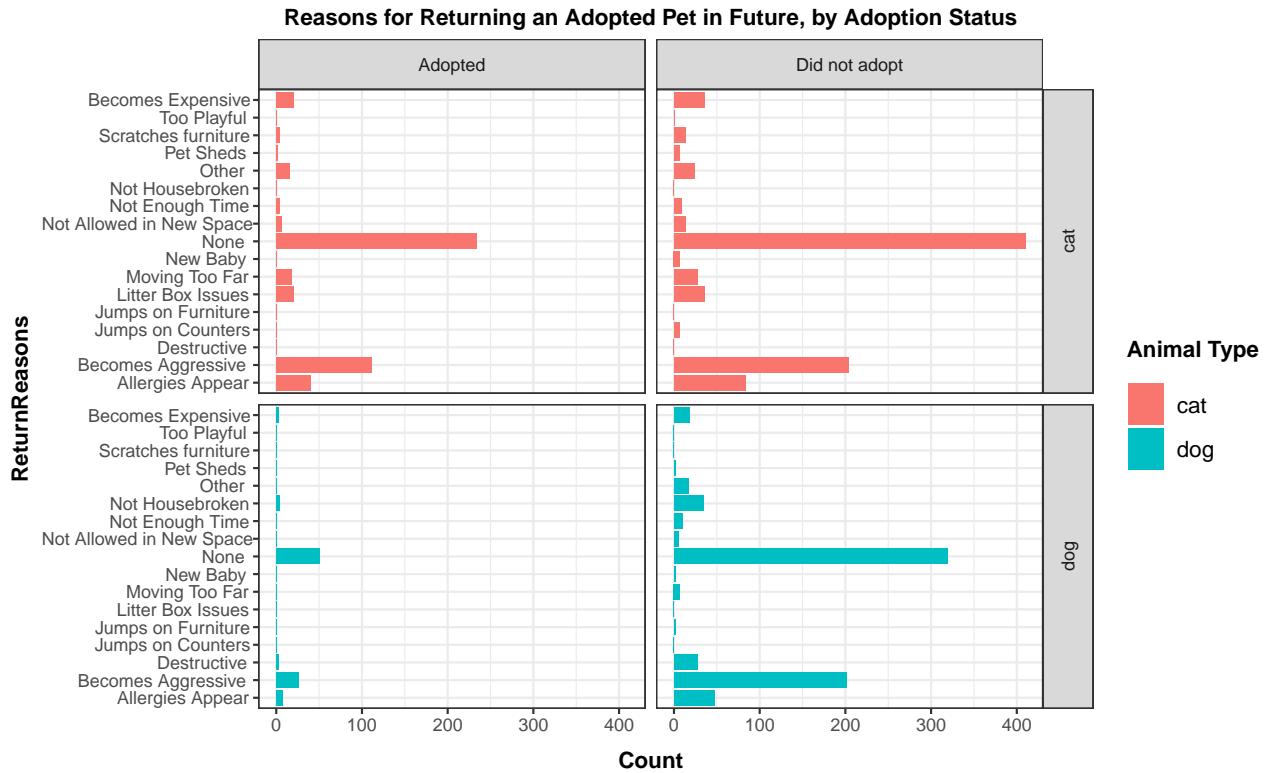
We also found that individuals with no children at home seem to make up the largest number of applicants.



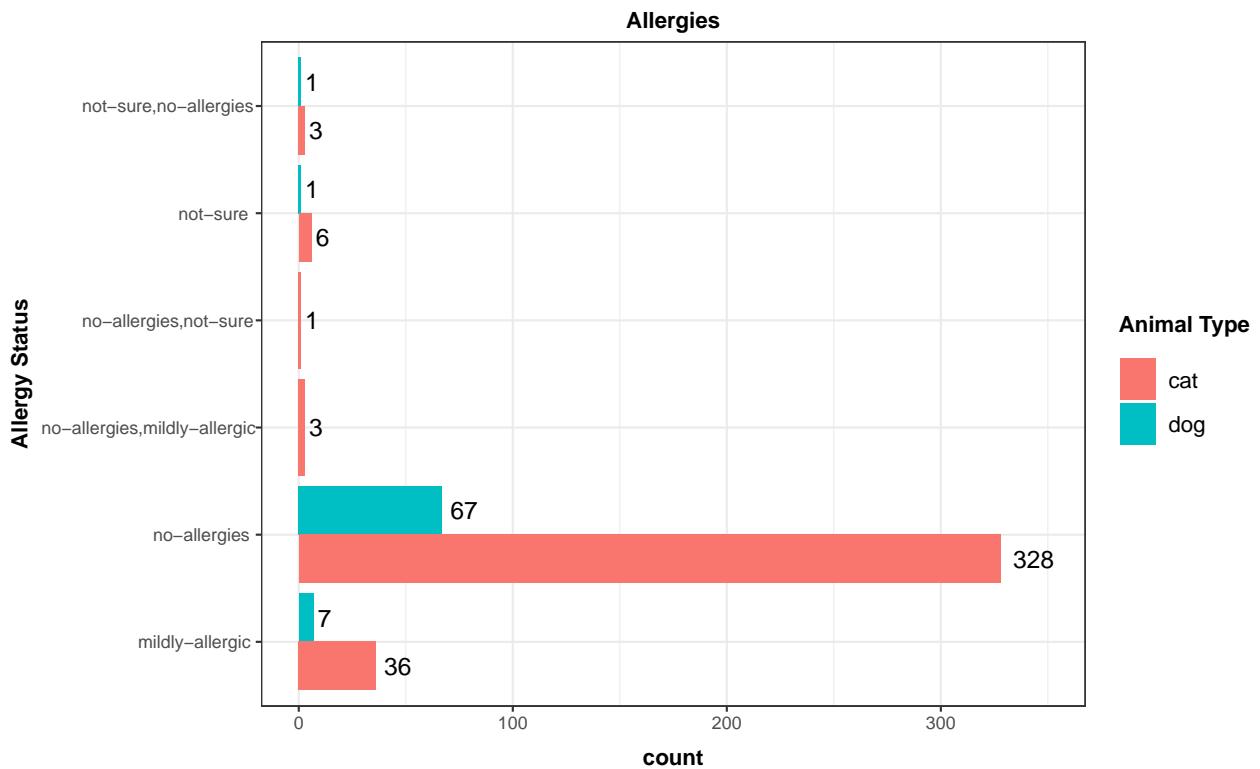
When we looked at the type of experiences applicants had with animals, we noticed that a higher number of applicants who were able to successfully adopt had less experience with animals in each area. Also important to note is the very low number of successful adoptions for applicants who reported that their pet ran away, was given away, was given to a shelter, or who reported having experience breeding or selling animals.



Not surprisingly, the highest number of successful adoptions were associated with a home policy that allowed pets.



The main reason that people would return a pet in the future is if the pet becomes aggressive or if allergies appeared. Regarding allergies, the vast majority of people who adopted a pet from PAWS reported no allergies.



Analysis of Denied and Red Flagged Applications

We further investigated the characteristics of applications that were denied or red flagged. There were 12 applications that were denied, 19 that were withdrawn, and 133 that were red flagged.

Denied Applications We only have data for 12 denied applications so the analysis is limited. In the future when we have more data, we could compare the denied applications to the adopted ones.

Key takeaways:

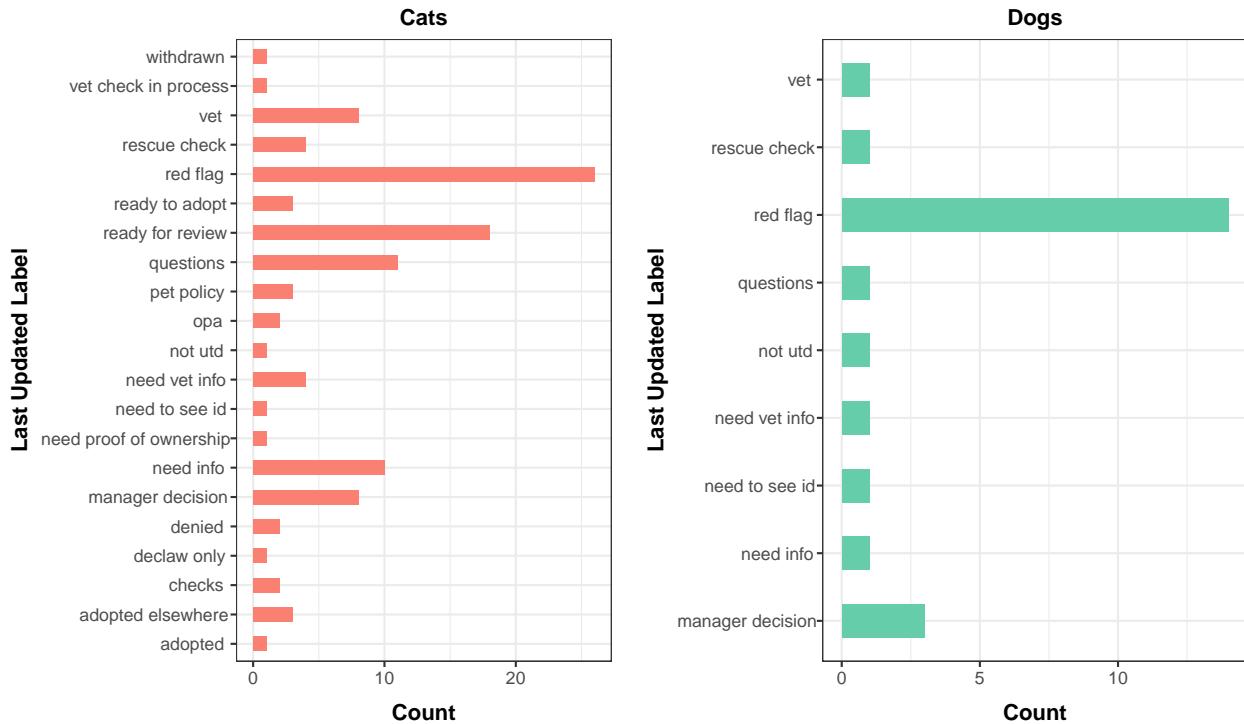
- No known allergies for the applicants
- Budget had no impact (same budget range for approved applications)
- All household members agreed to get a pet
- Majority of the applicants did not enter a home pet policy and not everyone is the home owner
- Many applicants had unfortunate incidents with prior pets (e.g. ran away, died in care)

Red Flagged Applications

There were 133 applications that were red flagged. 129 of the 133 have not yet resulted in an adoption or are still being processed. Two of the applications that were flagged were denied but that does not mean that the rest are going to result in adoption. Since the data set for the applications is from the end of 2018, many of the applications are still in progress. We do not have the final status of all the applications so we cannot conclude what happened to the red flagged applications. As a further project, I think it would be interesting to track the final status of the applications that were red flagged.

Below is a visualization that shows the last updated status for applications that were red flagged. After being flagged, the applications were sent to the manager to make a decision or the applicant was requested to provide more information (e.g. in many cases the applicant was required to provide more information about the vet).

Last Updated Label for Red Flagged Applications



Important Features for Prediction

After separately analyzing the different characteristics that affect whether an application results in an adoption, we wanted to further understand how the different features in the dataset could have had a combined effect on the adoption status. We therefore ran a basic Random Forests model on the dataset. A Random Forest is basically a tree-based algorithm where a random subset of predictors (or features) are evaluated at each node and the observed data is split into two regions using one of the predictors and a threshold value for that predictor such that the error in predicting the adoption status is minimized. Starting from the top of the tree with one node, two new nodes are created with each split and the tree is grown recursively till there are only a few observations in each leaf node. Multiple trees are built similarly and the results are combined together to predict the adoption status for any given set of characteristics.

To successfully build a Random Forest, we further cleaned the data to take care of all the missing values. We used 1665 observations and 90 variables out of a total of 1684 observations and 251 variables.

The combined effect of different characteristics on the adoption status can be studied by considering one of the important outputs generated by the Random Forests, the subset of predictor values that are found to be most commonly used as a criteria for splitting the dataset into two smaller regions at each node. This subset of predictor values, referred to as Important Variables, are shown in the plot below. As seen in the list, we find that the top three characteristics are number of children in a home, the type of dog, and the date the application was submitted. Improved results or a different set of important characteristics can be obtained from better and more complete data.

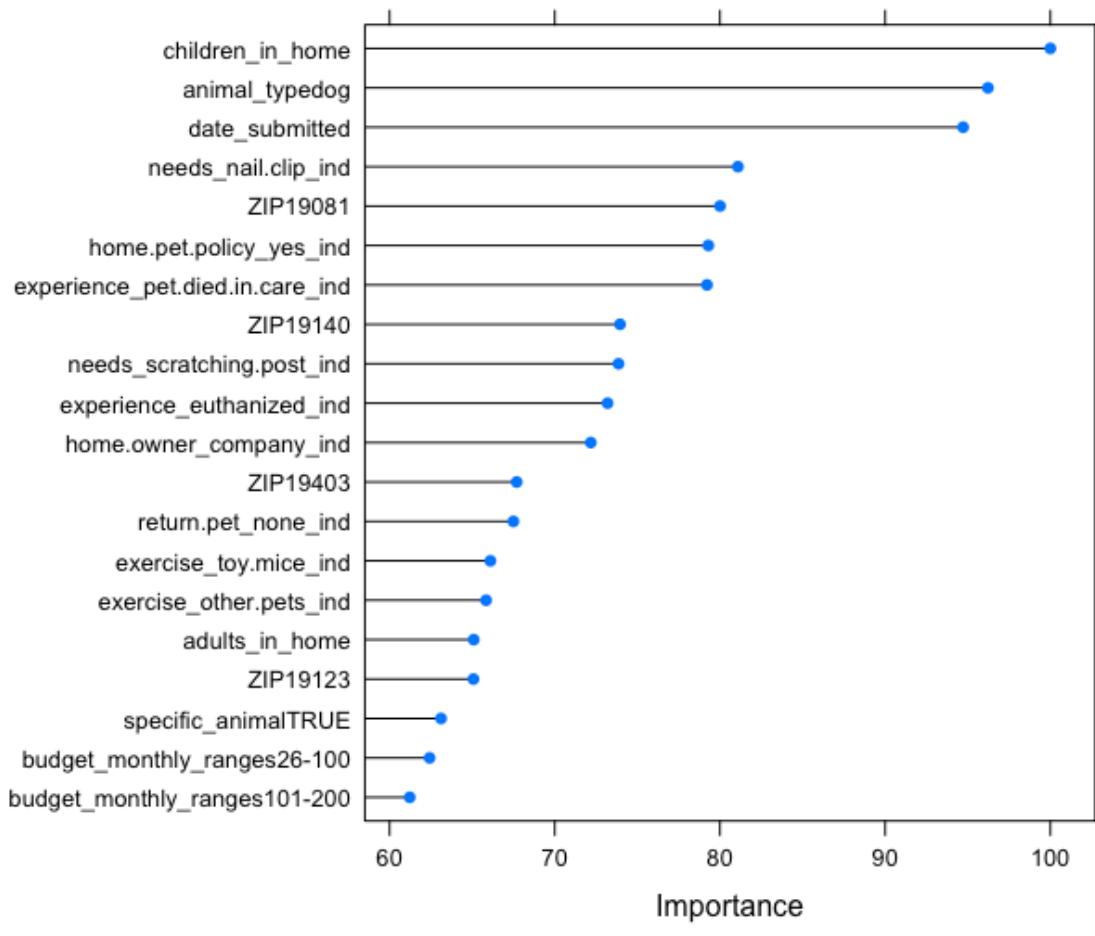


Figure 1: Important variables

3. Geographic factors

These analyses examined the data in relation to geographic and population parameters. We sought to understand how the socioeconomic environment of applicants might affect the likelihood of adoption, and whether there were factors contributing to regional differences in application processing speed.

Contributors

Joy Payton, MS is the Supervisor of Data Education in the Department of Biomedical and Health Informatics at the Children's Hospital of Philadelphia. She leads the development and implementation of education and outreach programs to help CHOP scientists become data-savvy and make the best, most informed use of the tools they have available.

Karla Fettich, PhD is Head of Algorithm Development at Orchestral, Inc. She leads efforts to develop data analytics solutions, predictive models and optimization approaches to create sustainable changes that improve operations and outcomes in long term care facilities.

Summary

Datasets

The following datasets were used:

1. Online applications for both cats and dogs In addition to the data collected via the online forms, applicants' addresses were extracted and associated with their respective census tracts. Census tracts are areas roughly equivalent to a neighborhood established by the Bureau of Census for analyzing populations, and generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people. Prior to making the PAWS data available, individual applicants' names, addresses and other identifiable data were removed from the dataset, keeping only census tract data and ZIP codes.
2. Trello cards and actions
3. Census data from the 2017 five-year American Community Survey via the American Fact Finder for the following areas:
 - Economic characteristics
 - Education characteristics
 - Median rent
 - Computer and networking characteristics

Results

Economic Considerations in Processing Applications

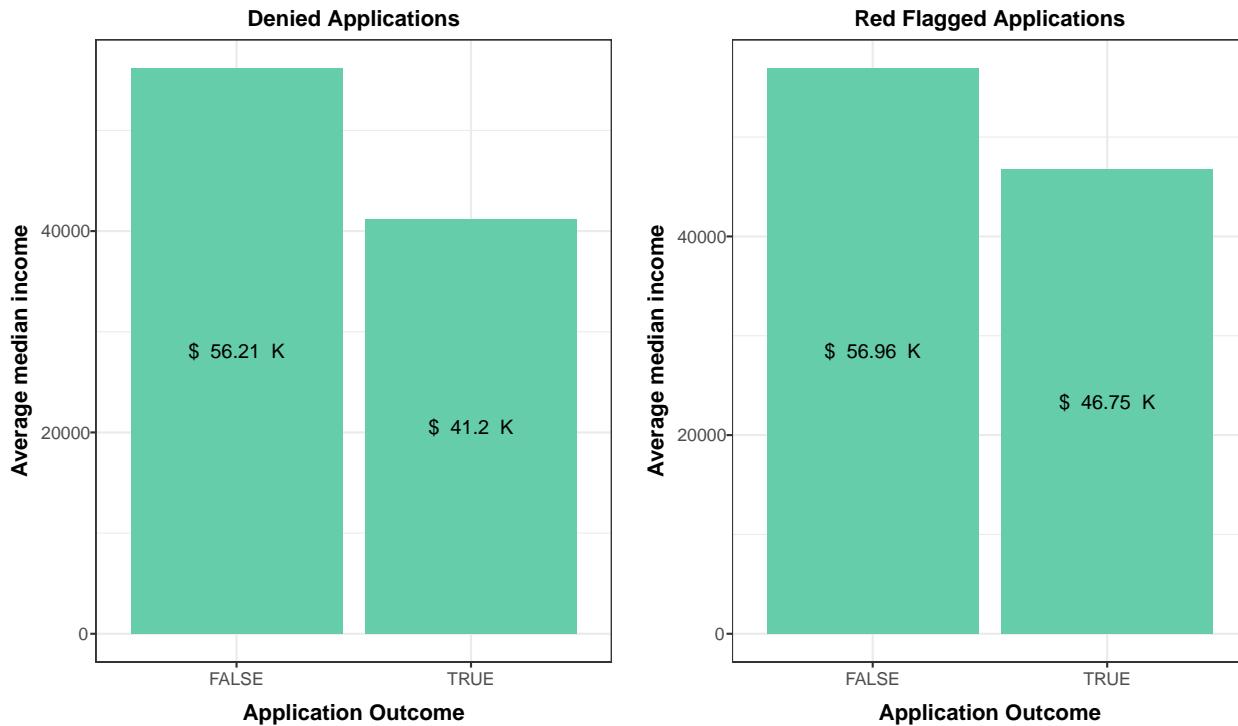
On average, we observed that dog applicants live in areas where the median income is higher compared to cat applicants (around \$60,000/year for dog applicants vs. \$54,000/year for cat applicants) and where the percent of households living under the poverty level is lower (18% for dog applicants vs. 22% for cat applicants). This suggests that dog applicants are from slightly wealthier neighborhoods. We further observed that dog applicants have more range between lower middle class and upper middle class, while cat applicants tend to skew more toward lower incomes. We also examined whether unemployment and reliance on public transit might play a role in applicants' preferences for dogs vs. cats. We observed no differences in mean neighborhood estimates of number of unemployed people of dog vs. cat applicants, but we did notice that in

neighborhoods where cat applicants live, more people tend to rely on public transit for their commute to work.

This finding aligns with the previous group's findings that a greater proportion of cat applicants had a lower monthly budget for pet care, compared to dog applicants, as well as with the pet care cost estimates provided by the ASPCA which suggest that the first year total costs of owning a dog (\$1,471 - \$1,779) exceed those of owning a cat (\$1,174).

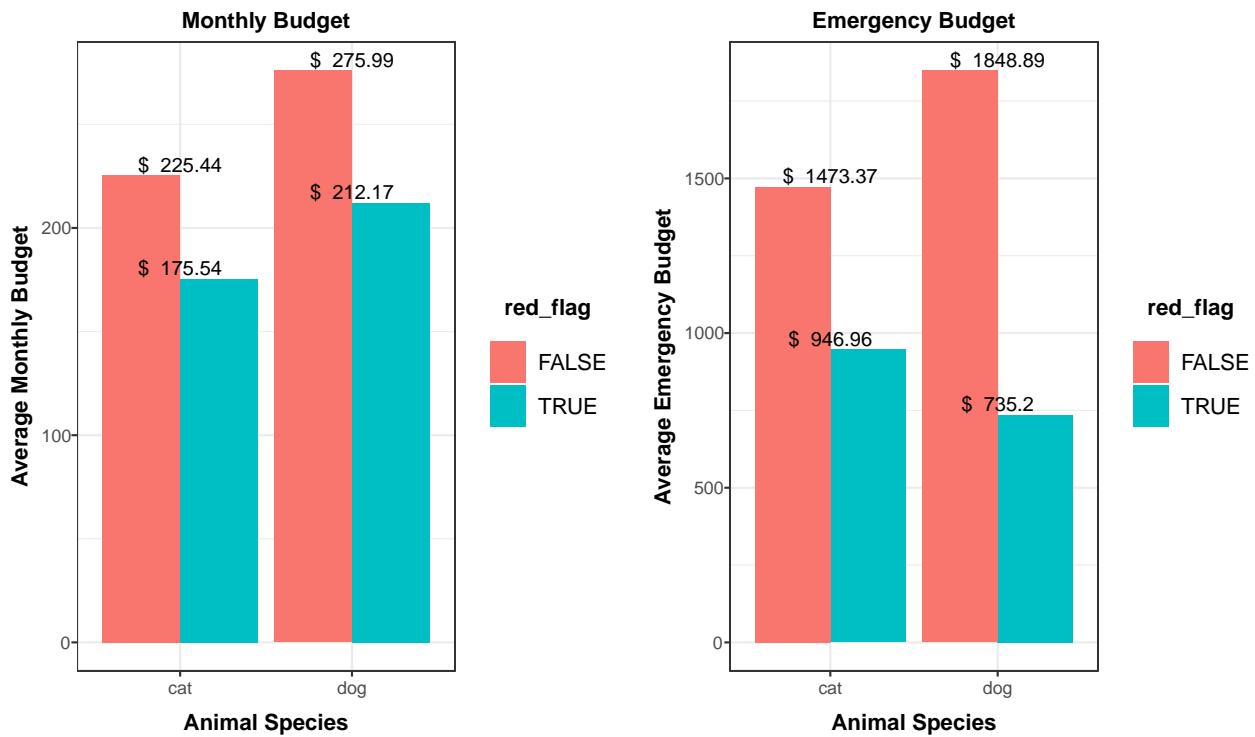
While we observed neighborhood wealth differences in animal species preference by applicants, we did not observe this factor to be biasing PAWS when processing applications (for this comparison, we used the “complete” status of a trello card at the time when the data were pulled - although this is fairly vague indicator because it does not describe the outcome of an application beyond having been processed).

Neighborhood wealth differences in application outcomes



We further looked into some of the outcomes of application processing, specifically *red flags* and *denied* applications. Applications from neighborhoods with a lower household median income (under \$50,000/year) are more likely to be red flagged and denied, compared to those with a higher household median income (over \$50,000/year).

Budgets by Red Flag Status



Additionally, red flagged **cat** applicants have a lower estimated monthly budget than their non-red-flagged counterparts (\$176 vs. \$225). For **dogs**, a similar trend was observed, but it did not reach the statistical significance threshold (\$212 vs. \$276). This pattern also holds when it comes to emergency budgets: red flagged applicants have a lower estimated emergency budget than their non-red-flagged counterparts (\$947 vs. \$1,473 for **cats** and \$735 vs. \$1,849 for **dogs**).

While we found that living in a lower income neighborhood does impact the estimated emergency budget at a statistically significant level, it only accounts for about 7% of the observed pattern. This indicates that there are additional factors that may play a role in how much money an applicant is able to set aside on a regular basis for pet care.

Efficiency Analysis in Philadelphia County

We also looked at applications that were processed within an efficient timeframe (defined here as 10 days), vs those that did not. An application was considered efficient if it was given a decision label (“denied”, “do not follow up”, “adopted”, “adoption follow up”, “approved”, “ready to adopt”, “ready for review”, “reviewed with handouts only”, “approved with limitation”, “dog meet”, “returned”, “adopted elsewhere”) and the last trello checklist item was checked off 10 days or less from the date of application submission.

Dogs

We found that in neighborhoods with a higher percentage of people who have a cell data plan and no other type of internet subscription, there was also a trend for a lower proportion of efficient applications, this effect being more pronounced in north and northeast Philadelphia. There could be many reasons for this: applicants who live in areas where many people do not have easy access to the internet may not be as familiar with filling out an online application (which represents the current application dataset); they may also not be able to easily find the information they need (since not all websites are mobile friendly); or they may be filling out the application form on a mobile device, which might be too long/detailed to adequately complete on a small screen.

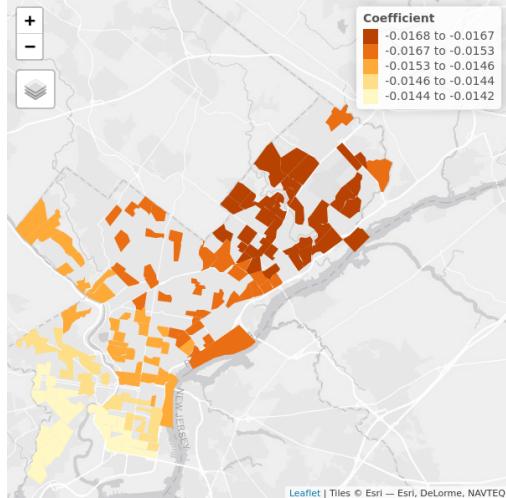


Figure 2: Cell data plan only coefficients

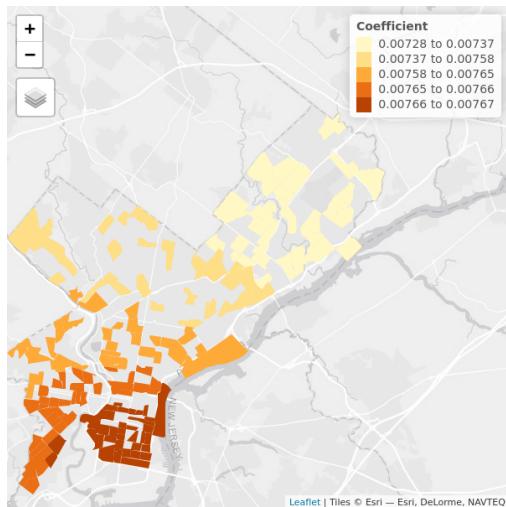


Figure 3: Population 25-34 enrolled in school coefficients

Additionally, in neighborhoods with a higher percentage of the population 25 to 34 year old enrolled in school, we also observed a significantly higher proportion of efficient applications. It is unclear what the reasons behind this might be, but possible options include the applicants' level of comfort with online applications, access to information, or other factors that are more specific to the life circumstances of individuals enrolled in school. This effect was less pronounced in northeast Philly.

Cats

Interestingly, for cats we found that in neighborhoods where a higher percentage of the population is children in grades 5-8, the proportion of efficient applications was lower, this effect being more pronounced in the north and northeast. While we do not know the reasons for this effect, it may be worth noting that ownership of and interest in pets tend to peak in middle childhood (i.e., 8–12 years). It may be that this effect influences the decision to submit an application, but that other barriers interfere with the application's timely processing (e.g. incomplete information, lack of responsiveness to provide additional information, change of mind).

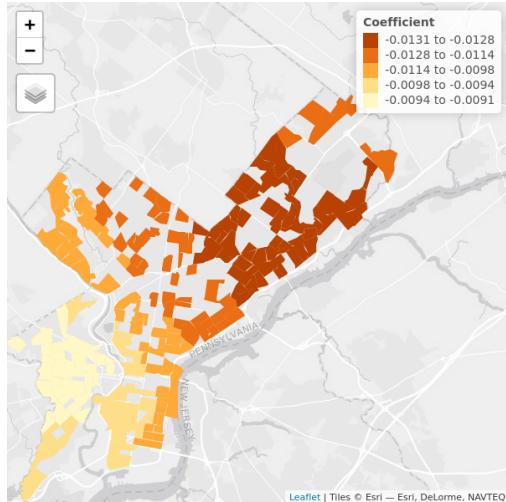


Figure 4: Children grades 5-8 coefficients

Conclusions and Next Steps

1. PAWS could develop a “smart” online application, that automatically educates the applicant on the cost of pet ownership when the budget is too low.

Since red flagged and denied applications still require processing by PAWS staff, and possibly even more intense processing than approved applications, it may be worth automatically screening and educating applicants who may have unrealistic budgeting expectations. Thus, perhaps a pop-up chart could appear when the budget is too low, *while* the applicant fills out the form. If the applicant proceeds to submit the application with a too-low budget, this application could be automatically labeled a red flag and sent for processing to a more experienced staff for further processing.

2. PAWS could provide applicants with a detailed breakdown of costs for a new pet, and have an adoption counselor go through the itemized list with the applicant to identify how each item could be covered.

Taking for instance the pet care cost estimates provided by the ASPCA, PAWS could identify which categories might be most difficult for an applicant to cover. Then, PAWS could provide a set of options (e.g. list of lower cost vets, cheaper options for enrichment using household items, list of affordable dog trainers) that might make the costs more manageable for those who are on a tighter budget.

3. PAWS could promote sharing or pooling of resources among its adopters.

Many pets have preferences when it comes to food, treats and toys, and it takes a while for a new adopter to learn them. This can result in a lot of wasted money. PAWS could facilitate and promote sharing of these resources (including any other accessories, or even transport help), at the adopters’ own risk, via an online community.

4. PAWS could assess the user-friendliness of its online application form on different platforms.

While the PAWS website might be mobile-friendly, PAWS could further assess whether the application form itself is represented in the most efficient way on a mobile device. To do this, information would first need to

be collected on the number of applicants who submit the application from a mobile device, as a revamping of the mobile interface for the application form may only be necessary if application quality is dependent on the device from which the application was submitted. An additional indicator of user friendliness could be the amount of time applicants spend on an application. PAWS could consider a ‘smart’ approach in sequencing and presenting questions so that the process is relatively speedy for the applicant, while also ensuring quality data.

5. PAWS could consider creating programs that are aimed at families with middle-schoolers.

Given that there is a spike in children’s interest in animals in middle school, PAWS could consider some ways to increase involvement of children in the animal care process, either by creating kid-friendly volunteer opportunities, kid-friendly community groups among adopters, or even informational events where people who are interested in adopting can ask questions and discuss experiences with adopters and PAWS representatives.

4. Social media factors

- @phillypaws tweets were analyzed to understand any patterns in twitter activity and whether that could be linked to application or pet information
- No strong trends or observations were gained, but it was fun to look at the data
- We did not see a strong association between twitter activity and applications
- We identified the most commonly tweeted words such as “home”, “adoption”, and “meet”
- We found that tweets with photos were more often favorited and retweeted

Contributors

Alice M Walsh, PhD is a computational biologist in the pharmaceutical industry. She enjoys analyzing patient data and trying to make informed decisions using data. She also enjoys walking and training her dog, Pebbles.

Problem definition and dataset

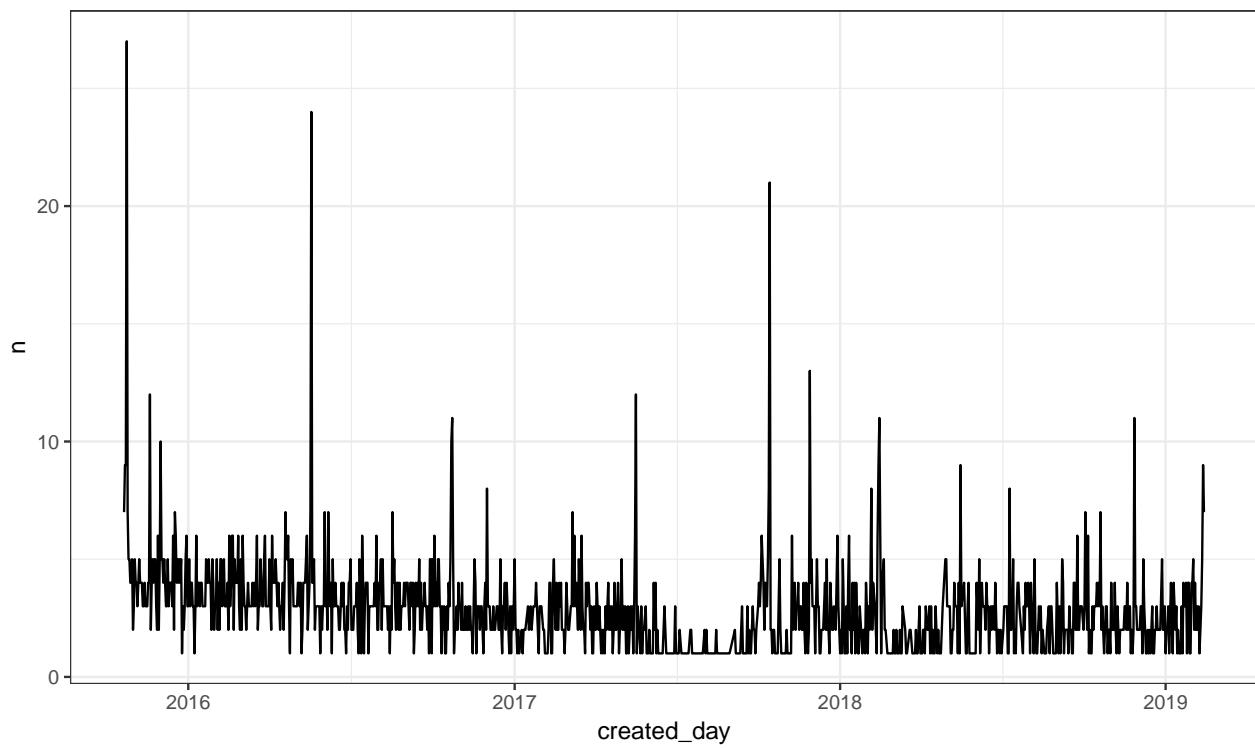
We examined the data from the PAWS twitter account, @phillypaws. We used the twitter API to download the most recent 3200 tweets, which included all tweets from 2018. Quotes and retweets were not excluded from the dataset.

Results

```
# Load in data - previously pulled with rtweet package
tweets <- readRDS(here::here('/Analyses/4_Other/tweets_13FEB2019.Rds'))
# Let's focus on 2018 to match other datasets
tweets_18 <- filter(tweets, created_at < as.Date("2019-01-01"), created_at > as.Date("2018-01-01"))

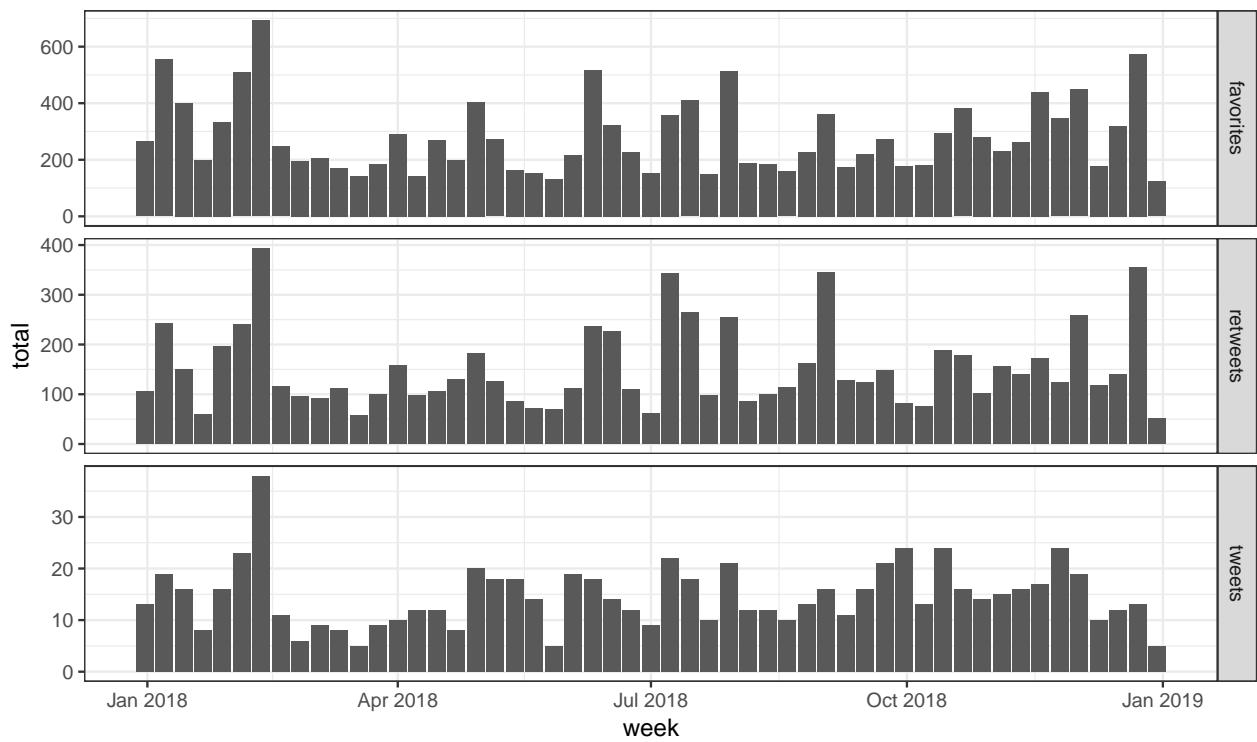
tweets %>%
  mutate(created_day = lubridate::floor_date(created_at, unit = "day")) %>%
  count(created_day) %>%
  ggplot(aes(x=created_day, y=n)) +
  geom_line() +
  theme_bw() +
  ggtitle("Volume of recent @phillypaws tweets")
```

Volume of recent @phillypaws tweets



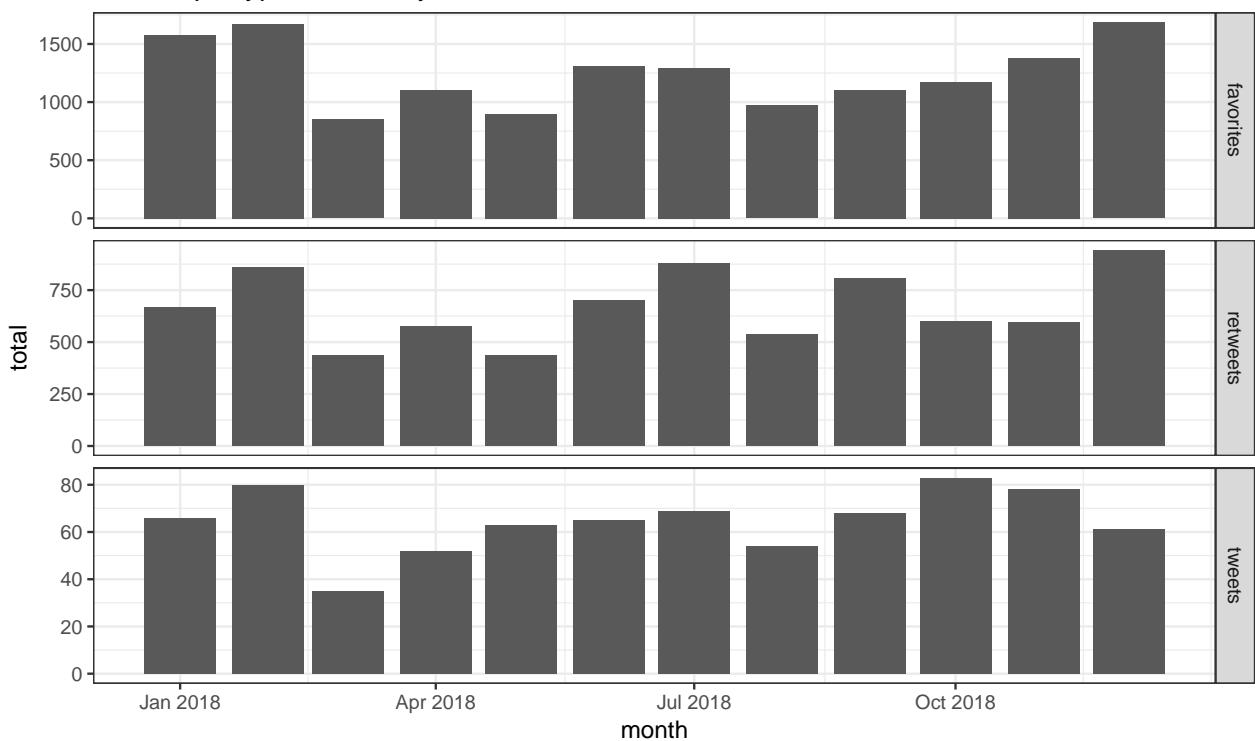
```
# Let's count tweets by day-of-the-week, week, month, and look for trends
# Plot by week:
tweets_18 %>%
  mutate(week = lubridate::floor_date(created_at, unit = "week")) %>%
  group_by(week) %>%
  summarise(favorites = sum(favorite_count),
            tweets = n(),
            retweets = sum(retweet_count)) %>%
  gather(metric, total, favorites:retweets) %>%
  ggplot(aes(x=week, y=total)) +
  geom_col() +
  facet_grid(metric~., scales = "free") +
  theme_bw() +
  ggtitle("2018 @phillypaws stats by week")
```

2018 @phillypaws stats by week



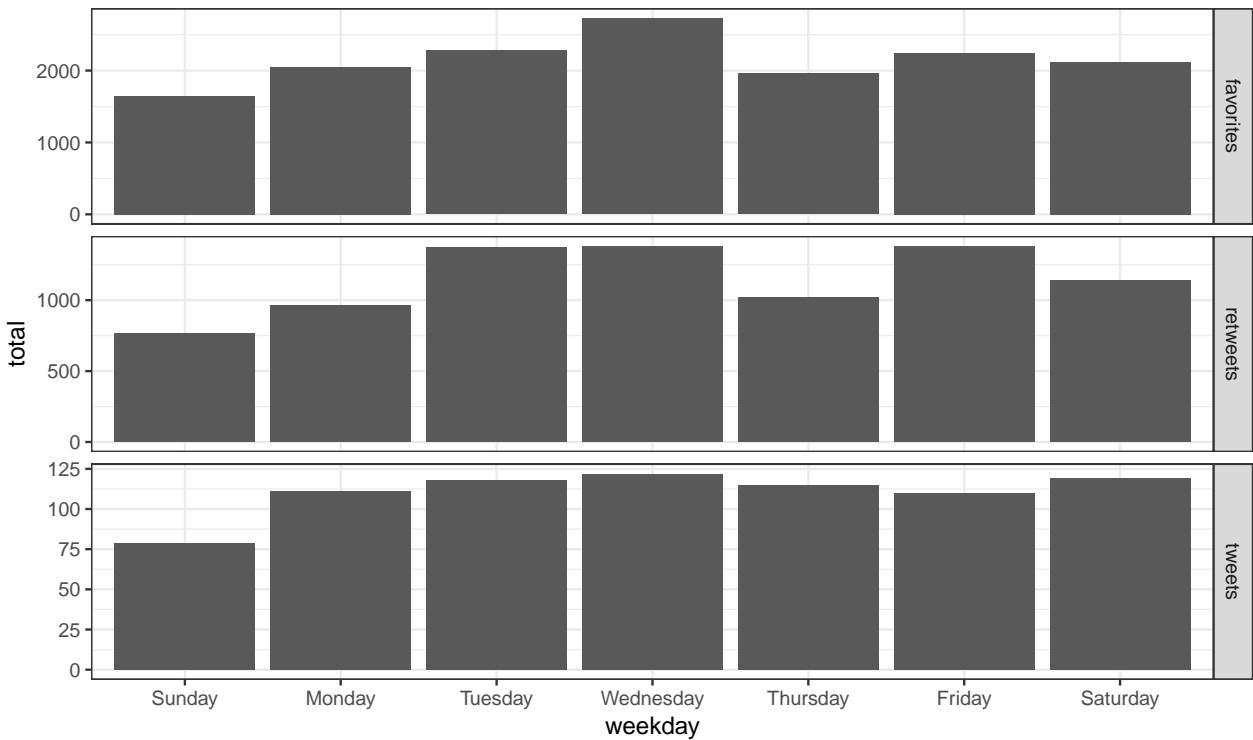
```
# Plot by month:
tweets_18 %>%
  mutate(month = lubridate::floor_date(created_at, unit = "month")) %>%
  group_by(month) %>%
  summarise(favorites = sum(favorite_count),
            tweets = n(),
            retweets = sum(retweet_count)) %>%
  gather(metric, total, favorites:retweets) %>%
  ggplot(aes(x=month, y=total)) +
  geom_col() +
  facet_grid(metric~., scales = "free") +
  theme_bw() +
  ggtitle("2018 @phillypaws stats by month")
```

2018 @phillypaws stats by month



```
# Plot by day-of-the-week:
tweets_18 %>%
  mutate(weekday = weekdays(created_at)) %>%
  mutate(weekday = factor(weekday,
    levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")))
group_by(weekday) %>%
  summarise(favorites = sum(favorite_count),
            tweets = n(),
            retweets = sum(retweet_count)) %>%
gather(metric, total, favorites:retweets) %>%
ggplot(aes(x=weekday, y=total)) +
  geom_col() +
  facet_grid(metric ~ ., scales = "free") +
  theme_bw() +
  ggtitle("2018 @phillypaws stats by weekday")
```

2018 @phillypaws stats by weekday



```
# Compare twitter activity to application activity?
cat_apps <- read.csv(here::here('/Data/cat_apps.csv'),
                      na.strings = c(" ","","", "na", "NA"), stringsAsFactors = F) %>%
  janitor::clean_names() %>%
  mutate(date_submitted = as.Date(date_submitted, "%m/%d/%Y"))

dog_apps <- read.csv((here::here('/Data/dog_apps.csv')),
                      na.strings = c(" ","","", "na", "NA"), stringsAsFactors = F) %>%
  janitor::clean_names() %>%
  mutate(date_submitted = as.Date(date_submitted, "%m/%d/%Y"))

compare_cat_apps <- cat_apps %>%
  count(date_submitted) %>%
  mutate(type = "cat_apps") %>%
  rename(created_day = date_submitted)

compare_dog_apps <- dog_apps %>%
  count(date_submitted) %>%
  mutate(type = "dog_apps") %>%
  rename(created_day = date_submitted)

compare_tweets <- tweets_18 %>%
  mutate(created_day = lubridate::floor_date(created_at, unit = "day")) %>%
  filter(created_day >= min(compare_cat_apps$created_day)) %>%
  count(created_day) %>%
  mutate(type = "tweets")

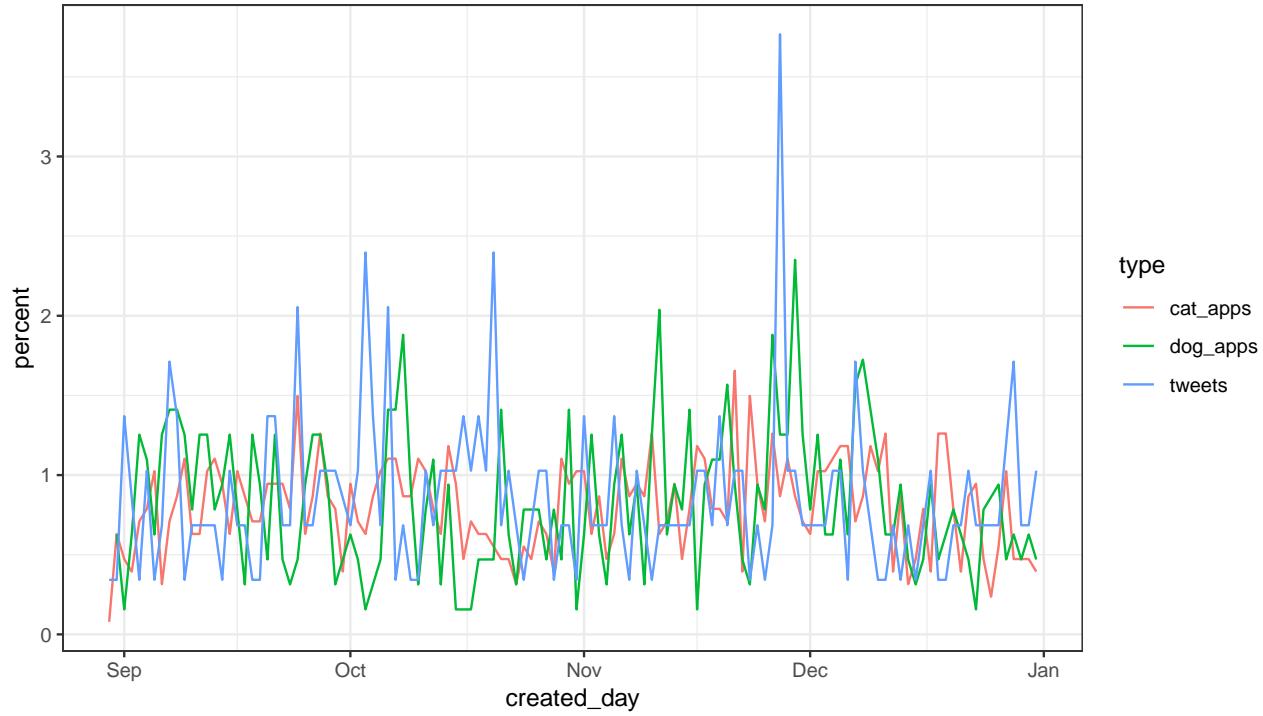
compare_days <- rbind(compare_tweets, compare_cat_apps, compare_dog_apps)
```

```

# Line by day
compare_days %>%
  group_by(type) %>%
  mutate(percent = n/sum(n) * 100) %>%
  ggplot(aes(x=created_day, y=percent, color=type)) +
  geom_line() +
  theme_bw() +
  ggtitle("Frequency of applications and tweets \nby day")

```

Frequency of applications and tweets
by day

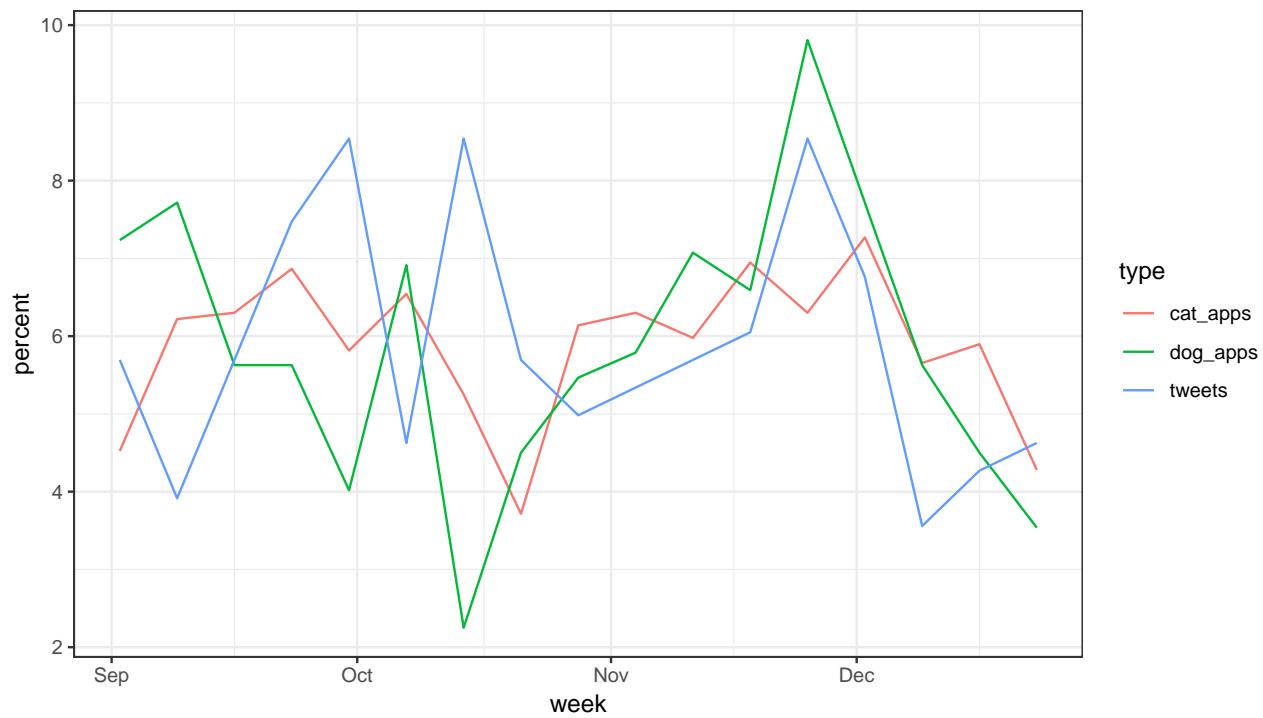


```

# Line by week
compare_days %>%
  # remove the short week - 2018-08-26 and 2018-12-30
  filter(created_day > as.Date("2018-09-02"), created_day < as.Date("2018-12-30")) %>%
  mutate(week = lubridate::floor_date(created_day, unit = "week")) %>%
  group_by(type, week) %>%
  summarise(total = sum(n)) %>%
  mutate(percent = total/sum(total) * 100) %>%
  ggplot(aes(x=week, y = percent)) +
  # geom_bar(aes(fill = type), position="dodge", stat = "identity") +
  geom_line(aes(color = type))+
  theme_bw() +
  ggtitle("Frequency of applications and tweets \nby week")

```

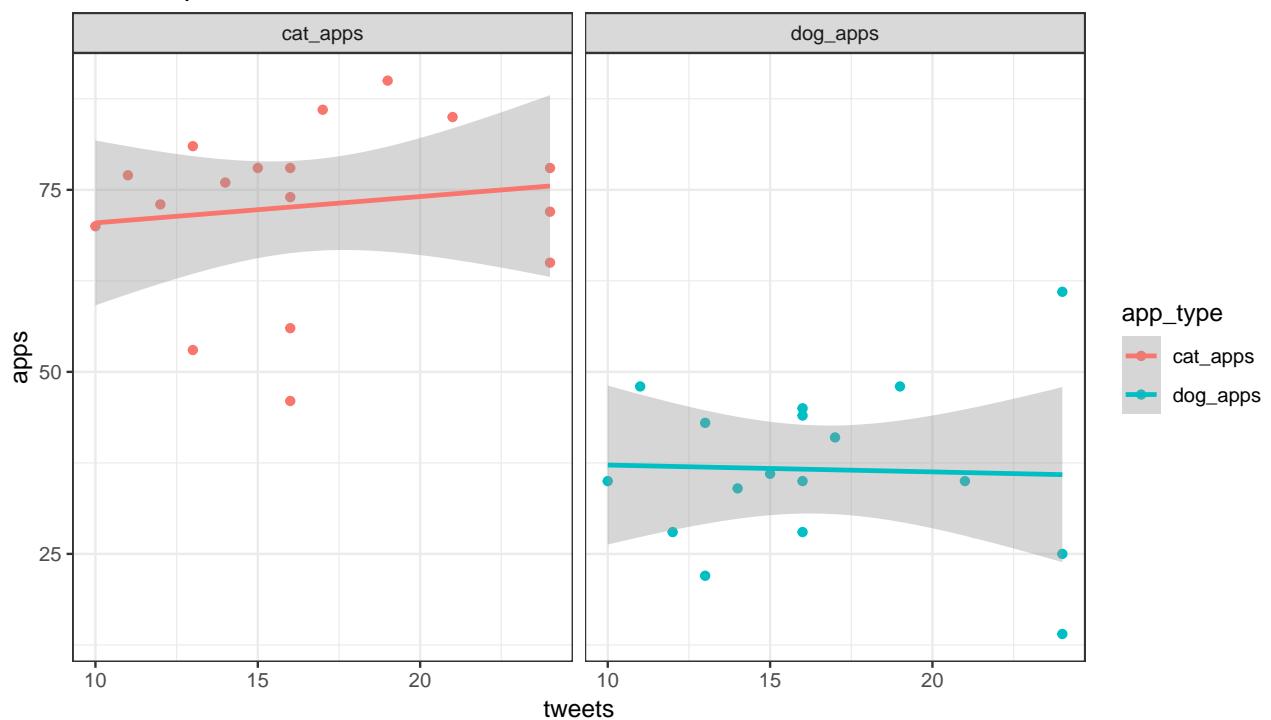
Frequency of applications and tweets
by week



```
# Scatter plot grouped by week
compare_days %>%
  # remove the short week - 2018-08-26 and 2018-12-30
  filter(created_day > as.Date("2018-09-02"), created_day < as.Date("2018-12-30")) %>%
  mutate(week = lubridate::floor_date(created_day, unit = "week")) %>%
  group_by(type, week) %>%
  summarise(total = sum(n)) %>%
  ungroup() %>%
  tidyr::spread(key = type, value = total) %>%
  tidyr::gather(app_type, apps, cat_apps:dog_apps) %>%
  ggplot(aes(x=tweets, y = apps, color=app_type)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~app_type) +
  theme_bw() +
  ggtitle("Applications by tweets\nNumber per week")
```

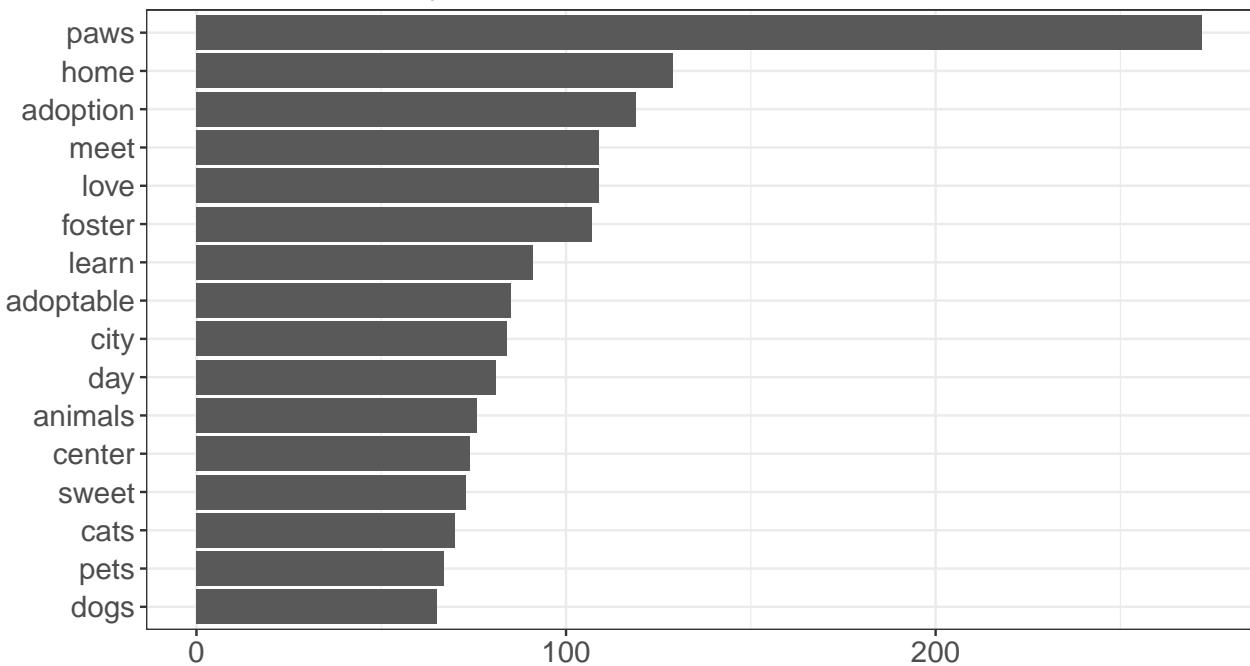
Applications by tweets

Number per week



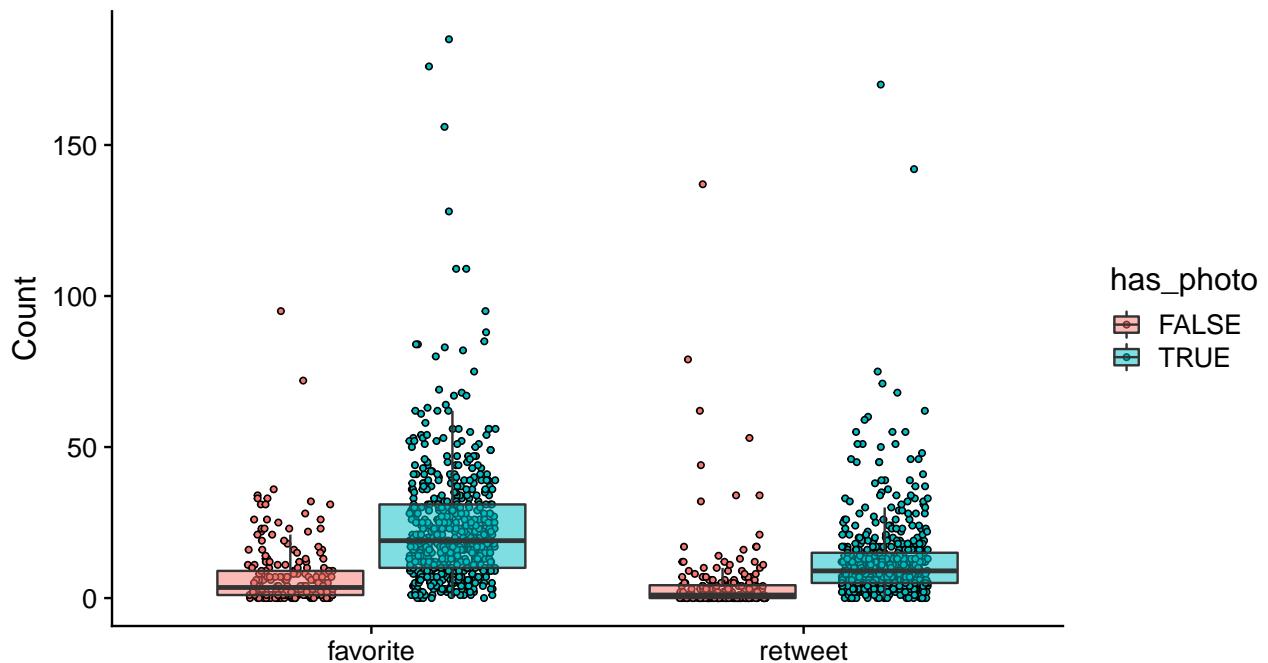
```
# What are the top words in all the tweets?  
# Wrote a function to plot for a given dataset  
# thanks to https://www.williamrchase.com/post/is-hadley-wickham-a-cat-or-dog-person-a-twitter-tidytext  
source(here::here('/Analyses/4_Other/alice_plot_top_words.R'))  
  
# plot_top_words(tweets)  
plot_top_words(tweets_18) + ggtitle("@phillypaws 2018 tweets\nWord frequency") + theme_bw() + xlab("")
```

@phillypaws 2018 tweets Word frequency



```
tweets_18 %>%
  mutate(has_photo = !is.na(media_type)) %>%
  gather(metric, count, favorite_count:retweet_count) %>%
  mutate(metric = gsub("_count","",metric)) %>%
  ggplot(aes(x=metric, y=count, fill = has_photo)) +
  geom_point(pch=21, size=1,
             position = position_jitterdodge(jitter.width = 0.2, jitter.height = 0)) +
  geom_boxplot(alpha = 0.5, outlier.shape = NA) +
  ylab("Count") + xlab("") +
  ggtitle("Tweets with photos \nare liked and retweeted more")
```

Tweets with photos are liked and retweeted more



```
# Merge in cat/dog predictions
pred_img <- readRDS(here::here('Analyses/4_Other/alice_predict.Rds'))
pred_img <- pred_img %>%
  mutate(status_id = gsub(".jpg","",gsub("predict/","",filenames)))

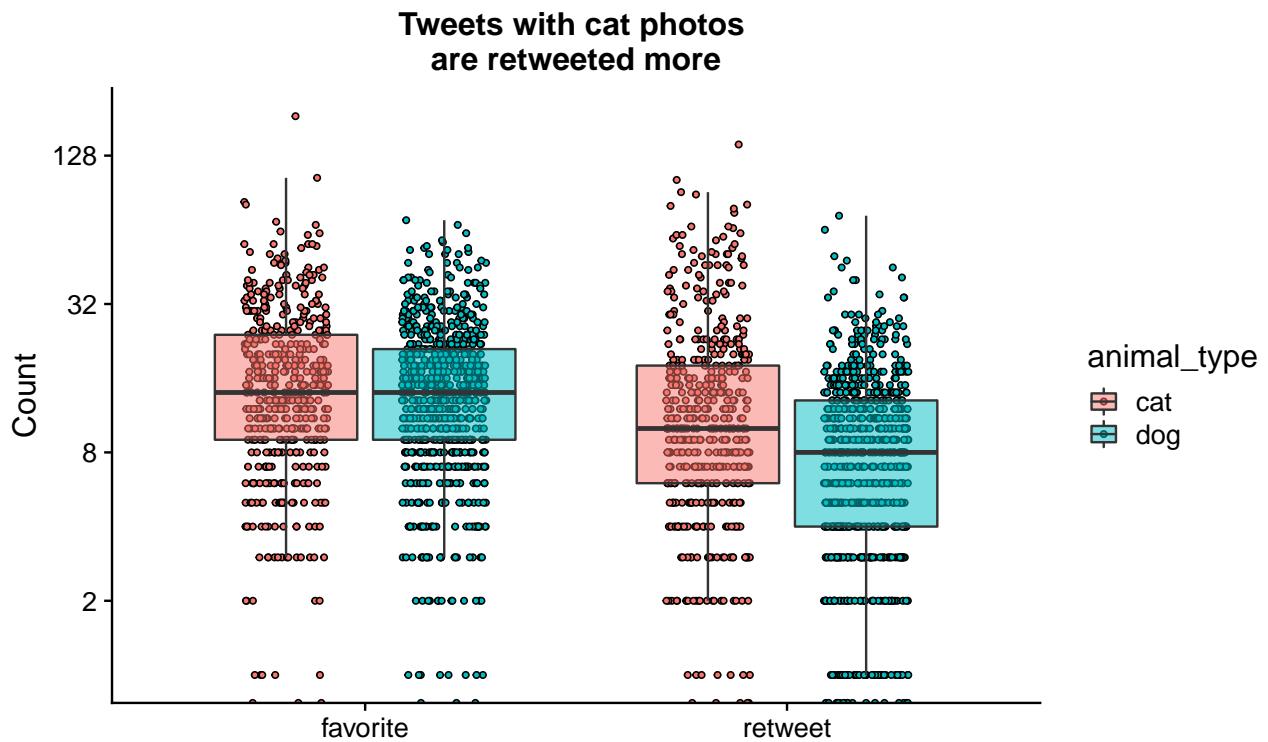
tweets_pred <- tweets %>%
  left_join(pred_img, by = "status_id") %>%
  mutate(has_photo = !is.na(media_type)) %>%
  # filter(has_photo == T) %>%
  mutate(animal_type = case_when(
    predict < 0.2 ~ "cat",
    predict > 0.9 ~ "dog",
    is.na(predict) ~ "no photo",
    TRUE ~ "other"
  ),
  animal_text = case_when(
    grepl("cat|kitten|Cat|Kitten", text) ~ "cat",
    grepl("dog|pup|Dog|Pup", text) ~ "dog",
    TRUE ~ "other"
  )) %>%
  filter(is_quote == F, is_retweet == F)

tweets_pred %>%
  filter(animal_type %in% c("cat","dog")) %>%
  # filter(retweet_count < 200) %>%
  gather(metric, count, favorite_count:retweet_count) %>%
  mutate(metric = gsub("_count","",metric)) %>%
  ggplot(aes(x=metric, y=count, fill = animal_type)) +
  geom_point(pch=21, size=1,
```

```

    position = position_jitterdodge(jitter.width = 0.2, jitter.height = 0)) +
geom_boxplot(alpha = 0.5, outlier.shape = NA) +
scale_y_continuous(trans = "log2") +
# ylim(c(0,100)) +
ylab("Count") + xlab("") +
ggtitle("Tweets with cat photos \nare retweeted more")

```

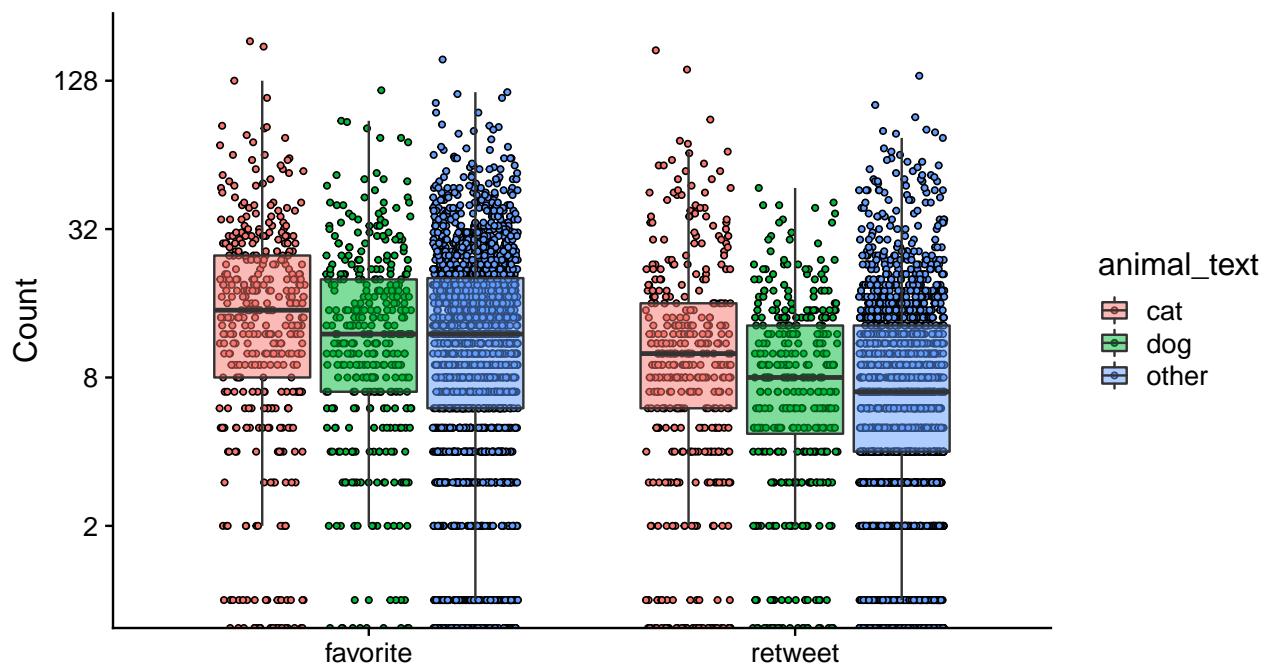


```

tweets_pred %>%
  # filter(animal_type %in% c("cat", "dog")) %>%
  # filter(retweet_count < 200) %>%
  gather(metric, count, favorite_count:retweet_count) %>%
  mutate(metric = gsub("_count", "", metric)) %>%
  ggplot(aes(x=metric, y=count, fill = animal_text)) +
  geom_point(pch=21, size=1,
             position = position_jitterdodge(jitter.width = 0.2, jitter.height = 0)) +
  geom_boxplot(alpha = 0.5, outlier.shape = NA) +
  # ylim(c(0,100)) +
  scale_y_continuous(trans = "log2") +
  ylab("Count") + xlab("") +
  ggtitle("Tweets with cat text \nare retweeted more")

```

Tweets with cat text are retweeted more



Conclusions and Next Steps

Keep on tweeting! If someone wanted to dig deeper into this data, they could:

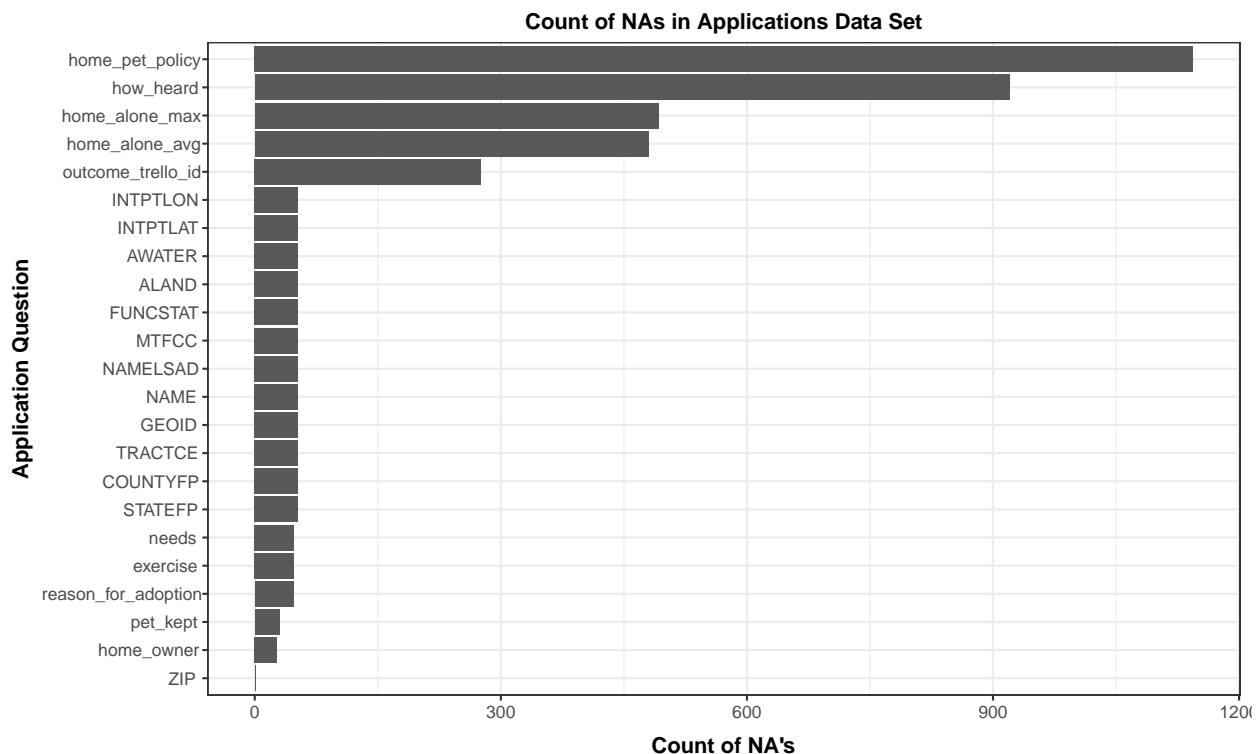
- Perform better image analysis on photos tweeted
- Try to link tweets about certain topics (a specific animal, request for donations) to specific outcomes (did the animal get adopted? did they receive more donations?)

5. Data considerations

Data Issues affecting Analyses

Missing Data

Overall we were able to achieve some insights given the application data. However, we were at times limited due to missing data in the applications data set. Below is a plot that shows counts of NA's in each column of the data set.



The question with the most missing data is one regarding the home pet policy. This seems like an important question, especially for renters, and a non-response here may require manual follow up by PAWS staff. Making this a required question could save some time in the future.

Unlimited Responses and Response Validation

Like many of the other teams, we ran into several challenges as a result of questions having a wide range of possible responses and illogical answers. For example, the 12 different responses below are for the Allergy question:

Response	Count
no-allergies	1,694
mildly-allergic	130
not-sure	38
not-sure,no-allergies	16
very-allergic	10
no-allergies,mildly-allergic	5
no-allergies,not-sure	5
mildly-allergic,no-allergies	3

Response	Count
mildly-allergic,very-allergic	3
mildly-allergic,not-sure	1
very-allergic,mildly-allergic	1
very-allergic,no-allergies	1

In one case the responses conflict with each other: “very-allergic,no-allergies”. This make grouping the data after the fact almost impossible because its not clear if this applicant has allergies or not. This is one example, but there were some other cases where this problem occurred as well, such as for the questions relating to Experience and Where the Pet Will be Kept.

For the monthly budget question, there were several negative numbers and some extremely large, strange values (i.e \$150,159.00). Utilizing some kind of response validation logic (i.e.only allow positive values) and limiting the range of responses to a reasonable size given the question (in this case maybe between 200 and 1,000) would also make future analysis much more efficient.

Recommendations for Collecting Clean Data

One of the most important recommendations moving forward would be to redesign the application to enforce standardized, limited and logical responses. Allowing only a single response combined with a limited response set would make analysis much easier in the future. Doing so will save PAWS staff time when reviewing applications *and* make future analyses easier and can lead to better insights.

Conclusions and Next Steps

This section should have a bulletpoint list of what conclusions can be drawn from the analyses that were performed, and what next steps should be taken, both by PAWS and by R-Ladies.