R-Ladies for PAWS Datathon (2019)

R-Ladies Philly

April 15, 2019

Executive Summary

This section should have up to 5 bullet points summarizing the main conclusions from the analysis. These should be worded in such a way that PAWS management can easily understand what actios would be beneficial. It can be a shorter version of the conclusions section below.

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

Results

1. Animal Trajectories

2. Application Trajectories

Contributors

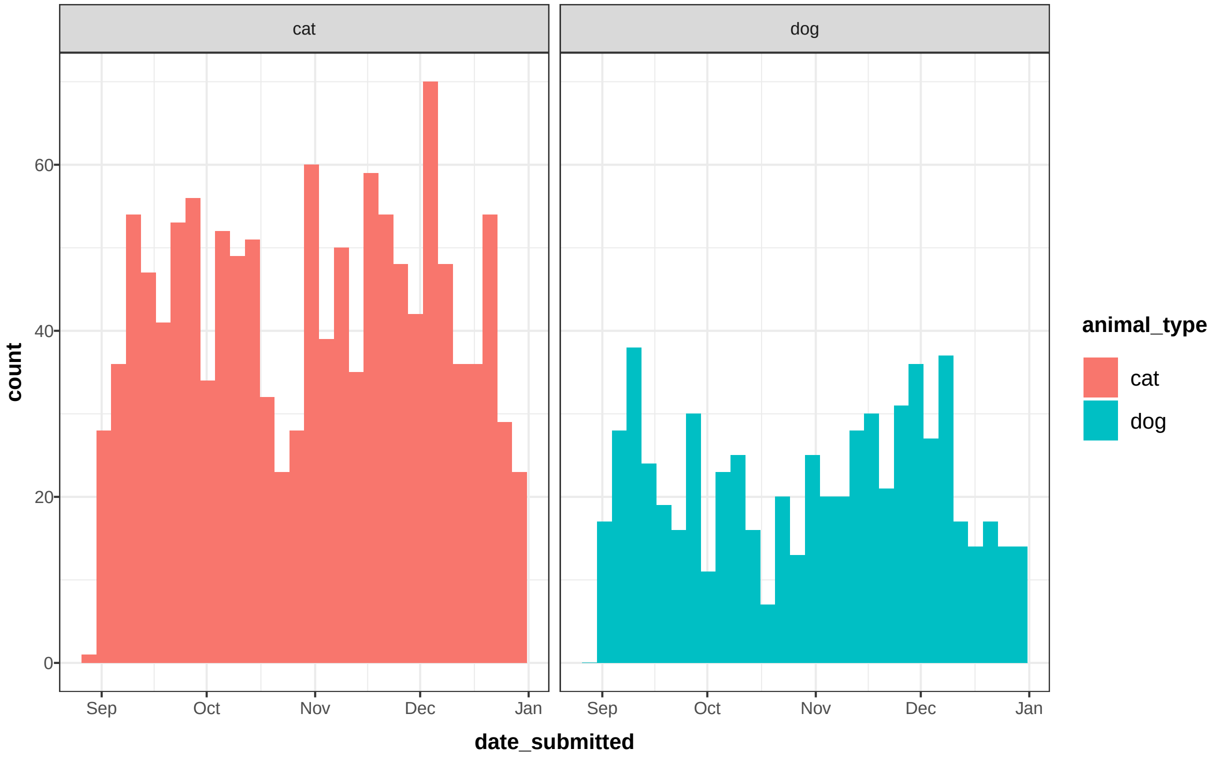
- **Ramaa Nathan** (group leader) is an aspiring data scientist with a PhD in Computer Science and an ongoing masters in Applied Statistics. Her background is in finance and healthcare.
- **Kate Connolly** is a digital analyst at the Philadelphia Inquirer where she helps to maintain the analytics framework 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.

Data Pre-processing

As our group focussed on questions about application trajectories, our starting point was an applications dataset comprised of `dog_apps.csv` and `cat_apps.csv`. For data pre-processing, the following steps were particularly important:

- Standardizing responses for `ideal_adoption_timeline`, `all_household_agree`, `home_pet_policy`, `experience` and `pet_kept`. For example, `ideal_adoption_timeline` had responses "next-few-weeks" and "few-weeks" which we standardised as one response ("few-weeks").
- For `children_in_home` and `adults_in_home`, ignoring "-" by taking the absolute value and replacing absurd values with NA (we replaced values greater than 15 with NAs).
- Capping `budget_monthly` and `budget_emergency` at $10000 and $20000 respectively.
- Addressing spelling variations in the `City` variable. For example, replacing the strings "PHILLY", "FILADELFIA", "PHILIDELPHIA", "PHIMADELPHIA", "PHIALADELPHIA", "PHIALDELPHIA", "PHILDELPHIA" with "PHILADELPHIA".
- Adding new indicator variables for variables containing lists of responses. For example, from `allergies` we created indicator variables for each response (`allergies_mildly.allergic_ind`, `allergies_no.allergies_ind`, `allergies_not.sure_ind`, `allergies_very.allergic_ind`).

Our cleaned applications dataset contained 1906 rows, 1594 unique trellos ids and the submitted dates ranged from 2018-08-30 to 2018-12-31:



To our applications dataset we added fields from the actions dataset (comprised of `dog_actions.csv` and `cat_actions.csv`), the cards dataset (comprised of `dog_cards.csv` and `cat_cards.csv`), and the petpoint dataset (`petpoint.csv`) to create our dataset for analysing successful applications. We also created another dataset comprised of the applications dataset and the cards dataset for analysing denied and red-flagged applications.

```
master_apps_report <- apps_with_indicators %>%
  filter(!is.na(trello_id)) %>%
  left_join(actions) %>%
  left_join(cards_with_indicators) %>%
  left_join(petpoint_with_indicators) %>%
  mutate(adoption = factor(ifelse((!is.na(outcome_type) & outcome_type=="Adoption"),TRUE,FALSE)),
         adoption_time = difftime(outcome_date, date_submitted, units = "days"),
         adoption_time = round(as.numeric(adoption_time), 2),
         budget_monthly_ranges =factor(budget_monthly_ranges,
                                       levels=c("<25","26-100","101-200","201-500","501-1000","1001-500
                                       ordered=TRUE))

masterapps_20190324 <- readRDS(here::here("/Analyses/2_Applicants/masterapps_20190324.rds"))
masterapps_20190324 <- masterapps_20190324 %>%
  mutate(adoption = factor(ifelse((!is.na(outcome_type) & outcome_type=="Adoption"),TRUE,FALSE)),
         adoption_time = difftime(outcome_date, date_submitted, units = "days"),
         adoption_time = round(as.numeric(adoption_time), 2),
         budget_monthly_ranges =factor(budget_monthly_ranges,
                                       levels=c("<25","26-100","101-200","201-500","501-1000","1001-500
                                       ordered=TRUE))

setdiff(colnames(master_apps_report), colnames(masterapps_20190324))
```

```
## character(0)
```

```
dim(master_apps_report)
```

```
## [1] 1684  251
```

```
dim(masterapps_20190324)
```

```
## [1] 1684  251
```

```
identical(master_apps_report, masterapps_20190324)
```

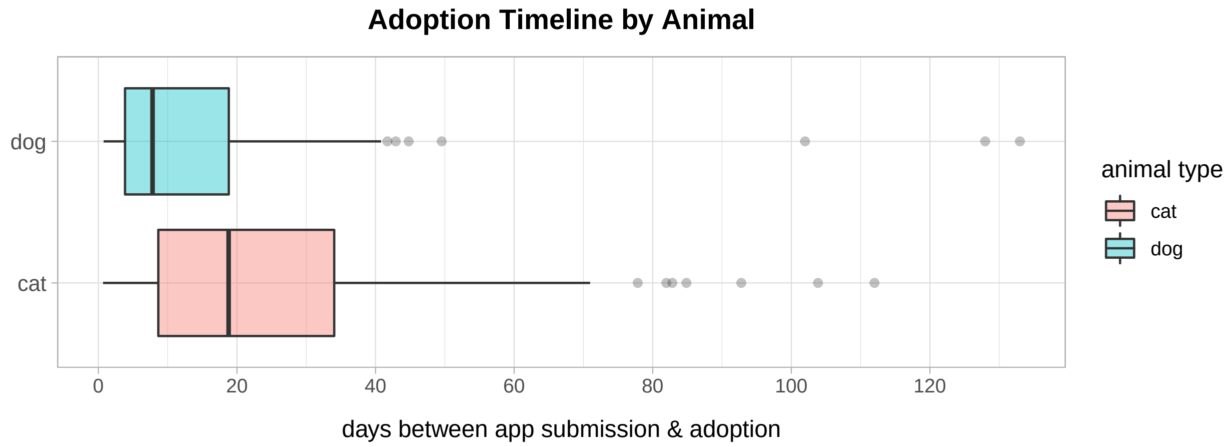
```
## [1] FALSE
```

Analysis of Time in Processing Applications

How Animal & Outcome Site Influence Application Timelines

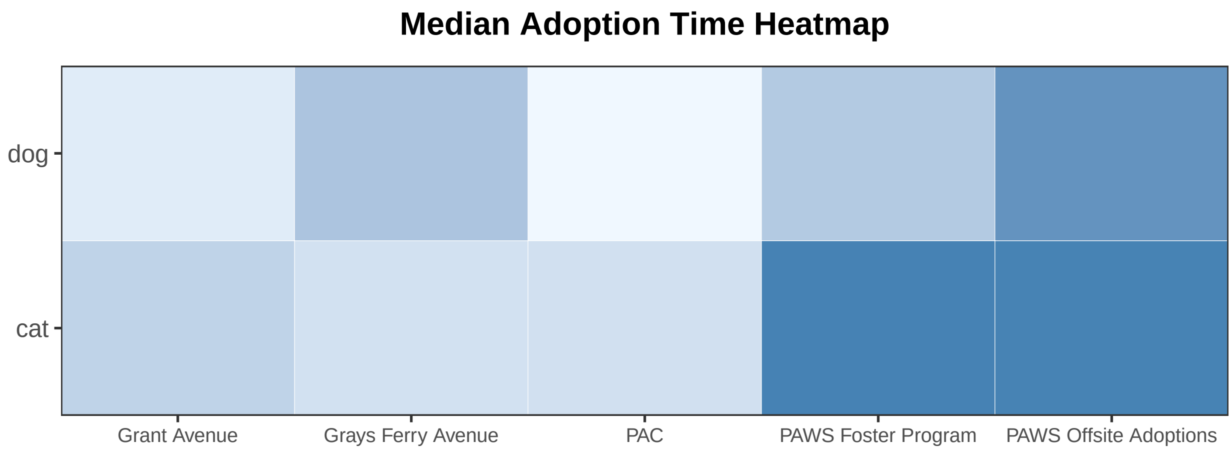
Application timelines were measured by taking the difference between the time an application was submitted and the time that application resulted in an adoption. Only applications that resulted in adoption were assessed; applications that were denied were not included in the analysis. This is a potential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

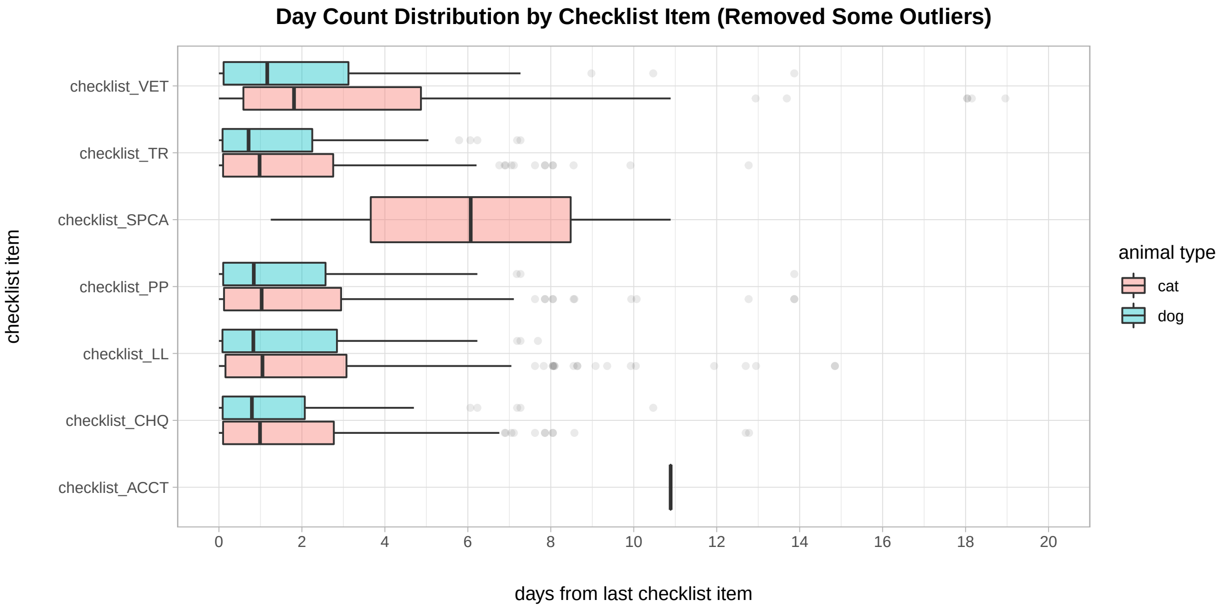
From the heatmap and table above, it's clear that overall average adoption times were higher at PAWS Foster Program & PAWS Offsite Adoptions locations. This is especially true for cat applications at those places

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

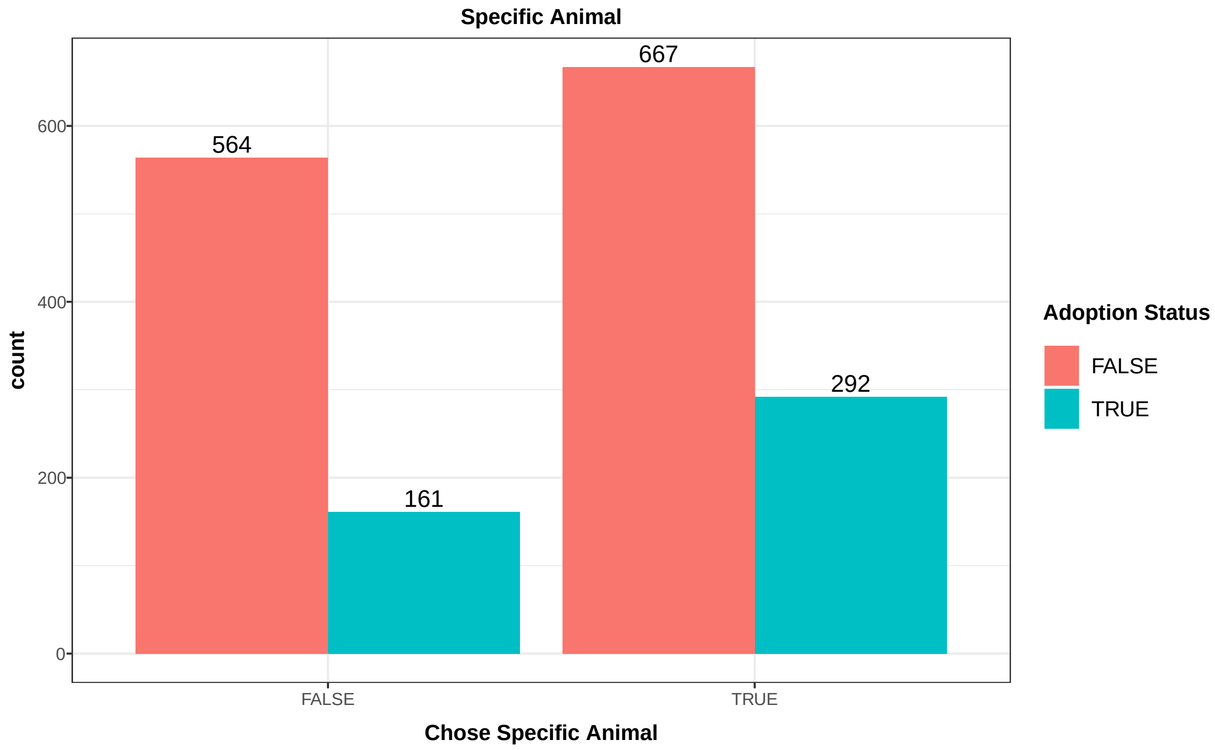


Most application items took between one and two days (median) to complete. While the animal type and outcome site didn't significantly impact the individual item times, cat applications generally exhibited slightly longer times between checklist items. Cat applications averaged about **1.2** days between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

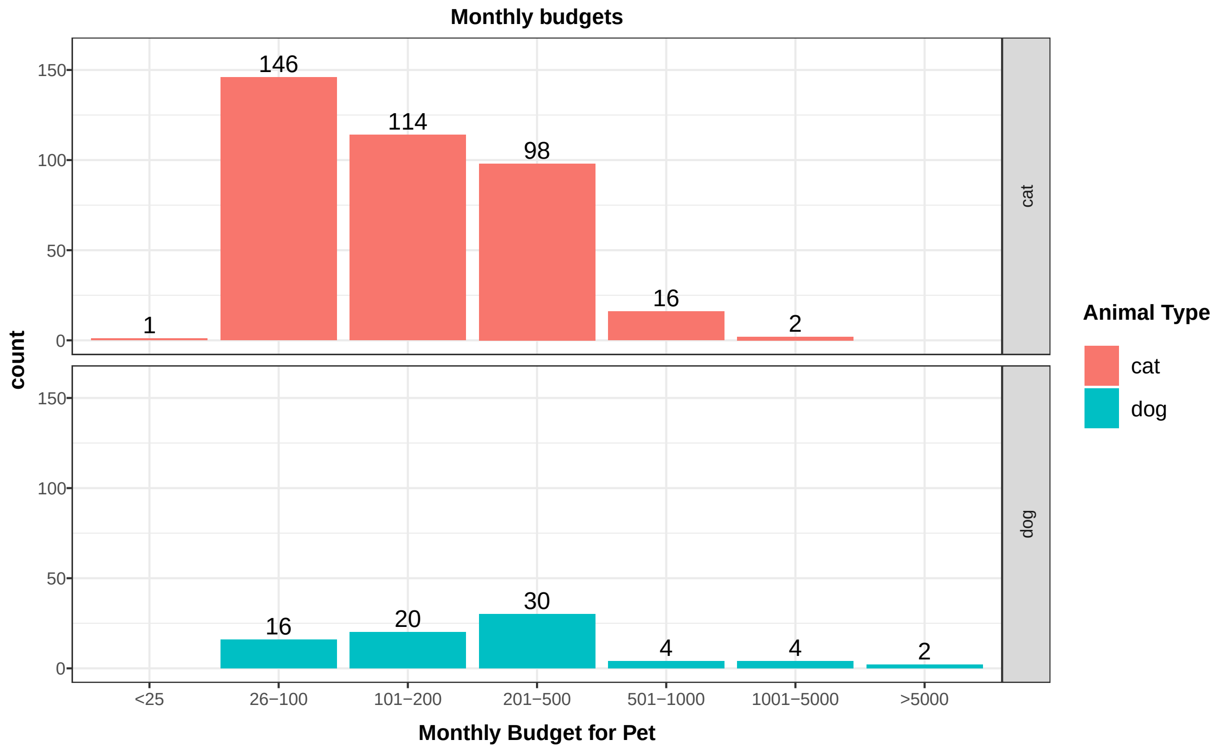
The chart above removed significant outliers, but further inspection of these outliers could be valuable. Understanding what causes certain application steps to take longer could help to streamline parts of the checklist process.

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%

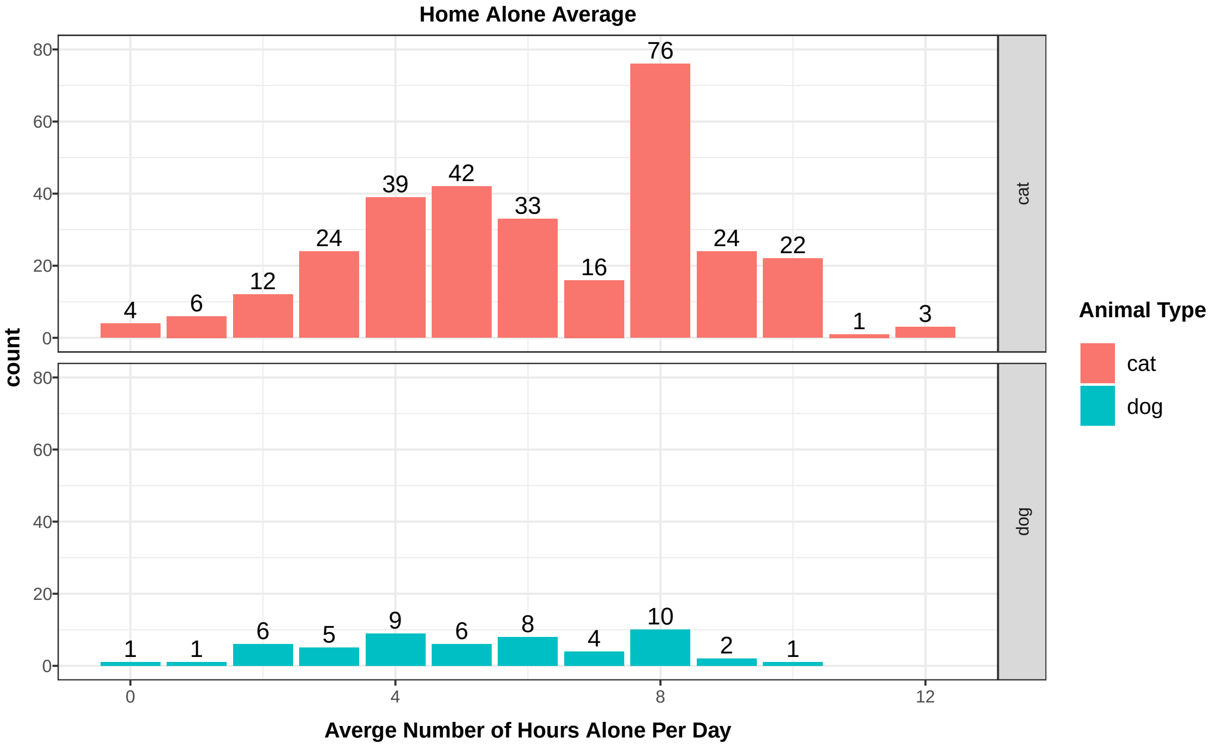
The table above shows the exceptions to the average checklist times. The ACCT and SPCA checklist items took considerably longer to complete than other items, but they also were present in less than 1% of applications. This low sample limits any sound conclusions, but does present an area for potential 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.



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



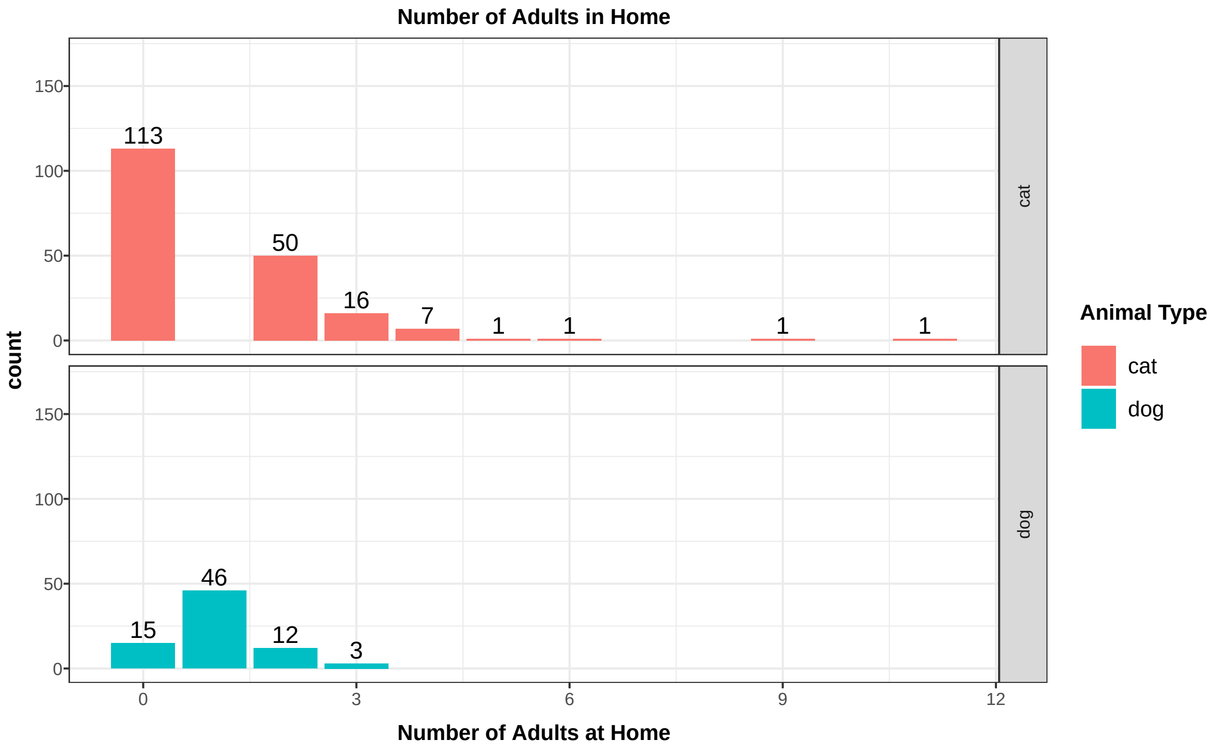
Most of the applicants who adopted a pet had allocated a monthly budget of less than $500.



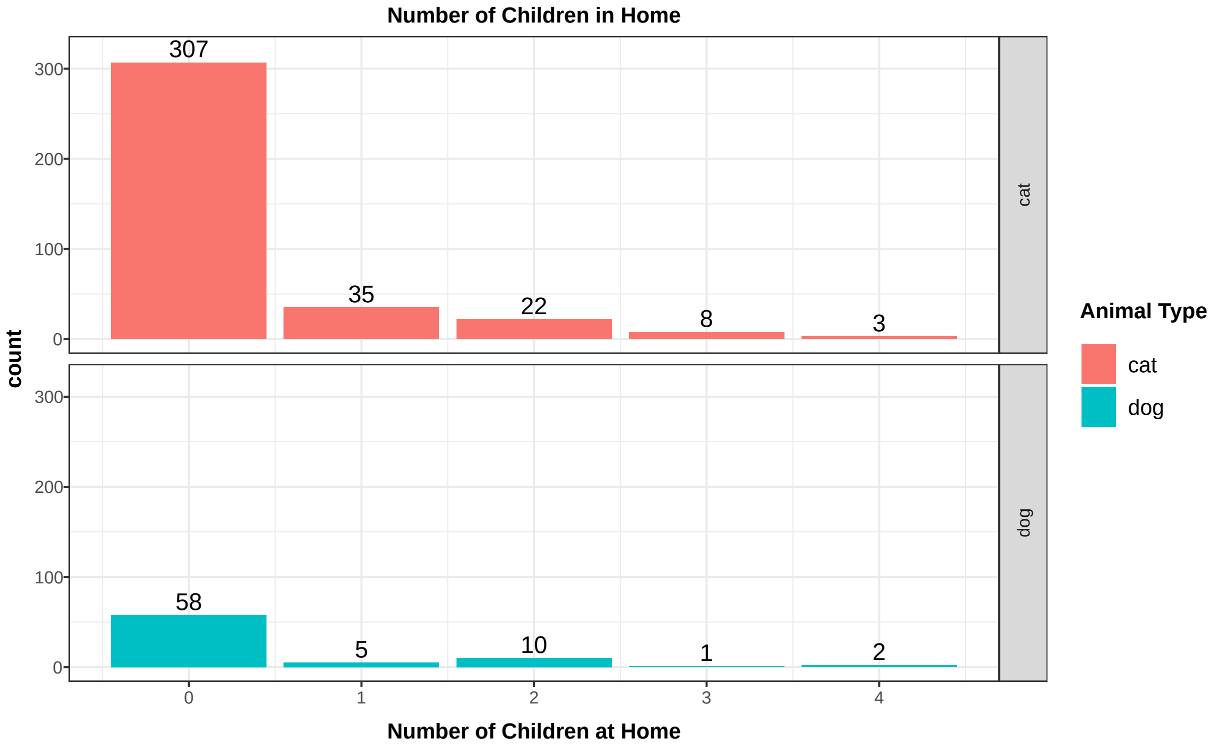
Applicants who expected to leave the animal alone at home for longer hours chose to adopt a cat. The largest number of applicants expected the animal to be alone for 8 hours, which would be typical of an applicant who works full time.

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

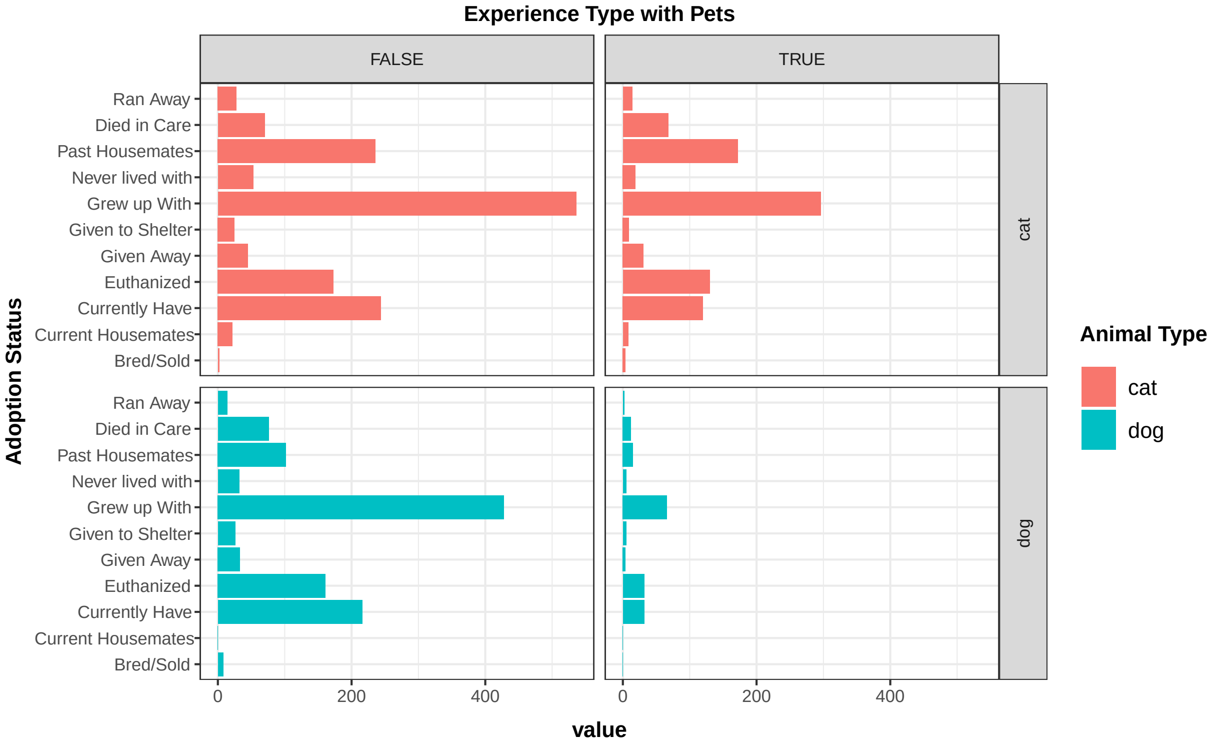
```
## Warning: Removed 1 rows containing missing values (geom_text).
```



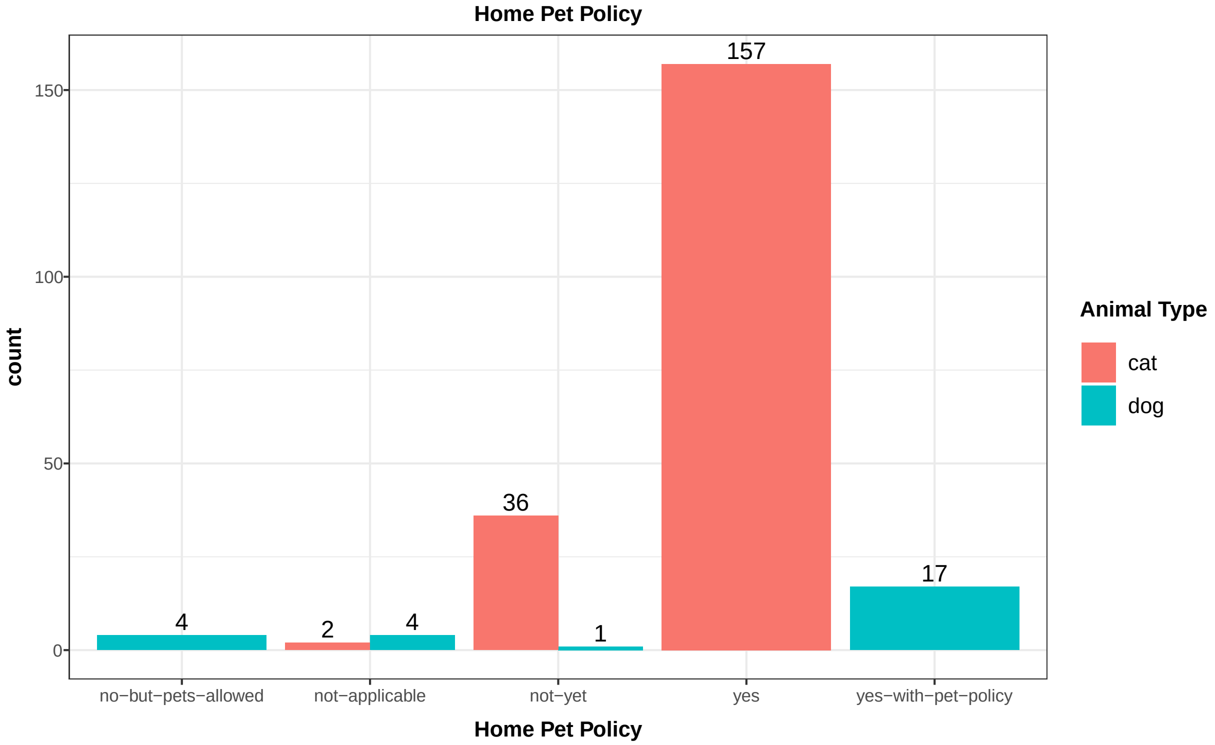
Singles overwhelmingly seem to prefer to adopt a pet.



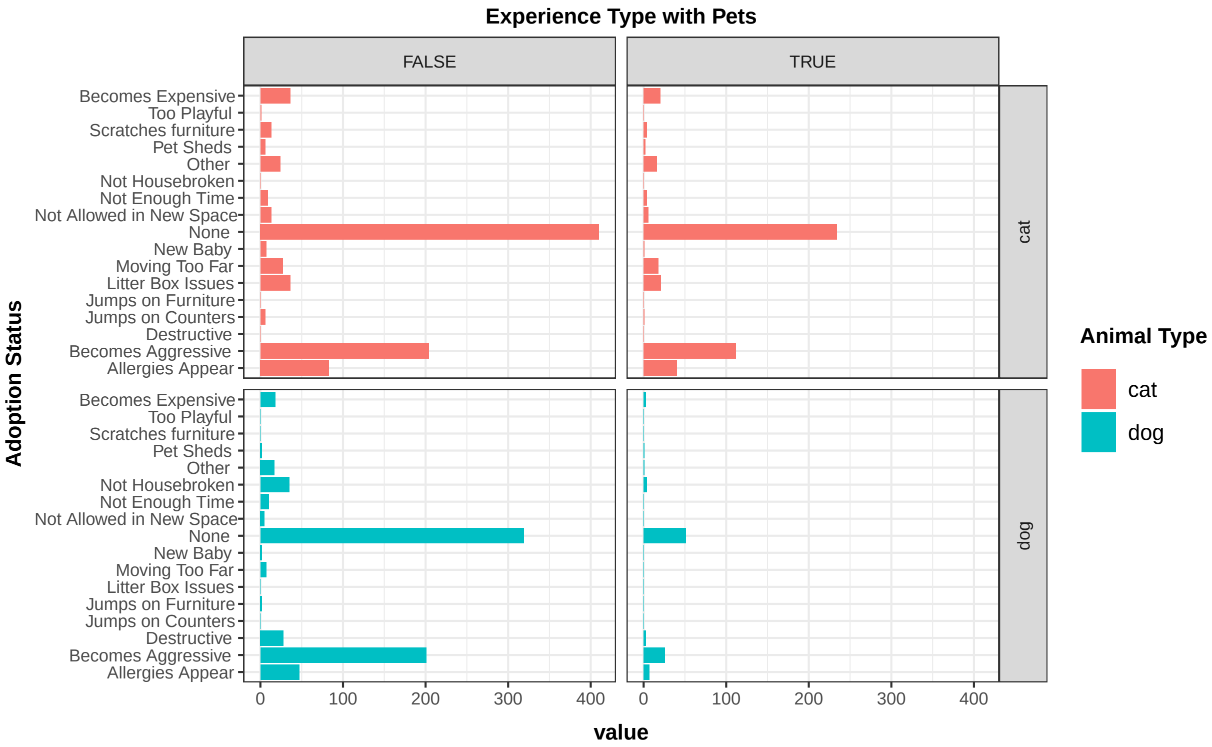
Again, families with no children at home seem to be the largest number of applicants. This correlates with mostly singles wanting to adopt.



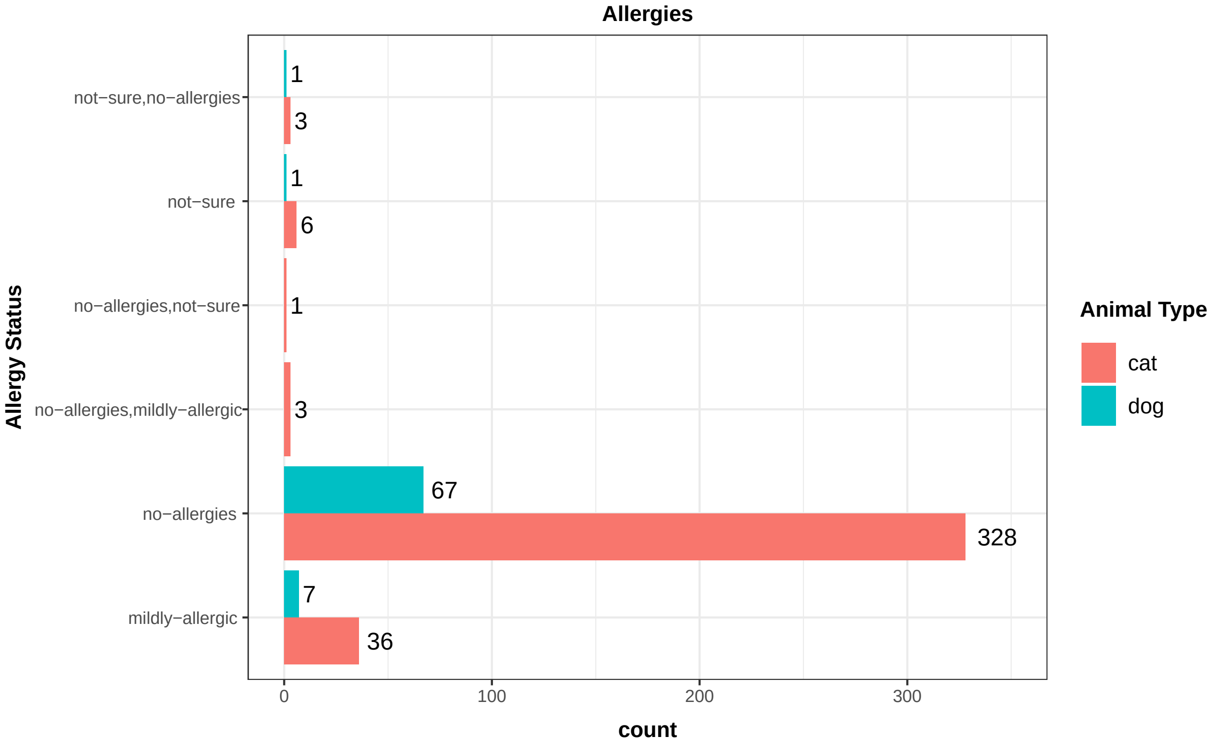
Interestingly, more number of applicants who were able to successfully adopt had less expereince in each of the types of experiences.



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



The main reason that people would return a pet in the future seem to be if the pet sheds or if they moved too far away. Of these, more number of people who would return if the pet sheds did not adopt and of hte ones who adopted, they mainly adopted a cat.

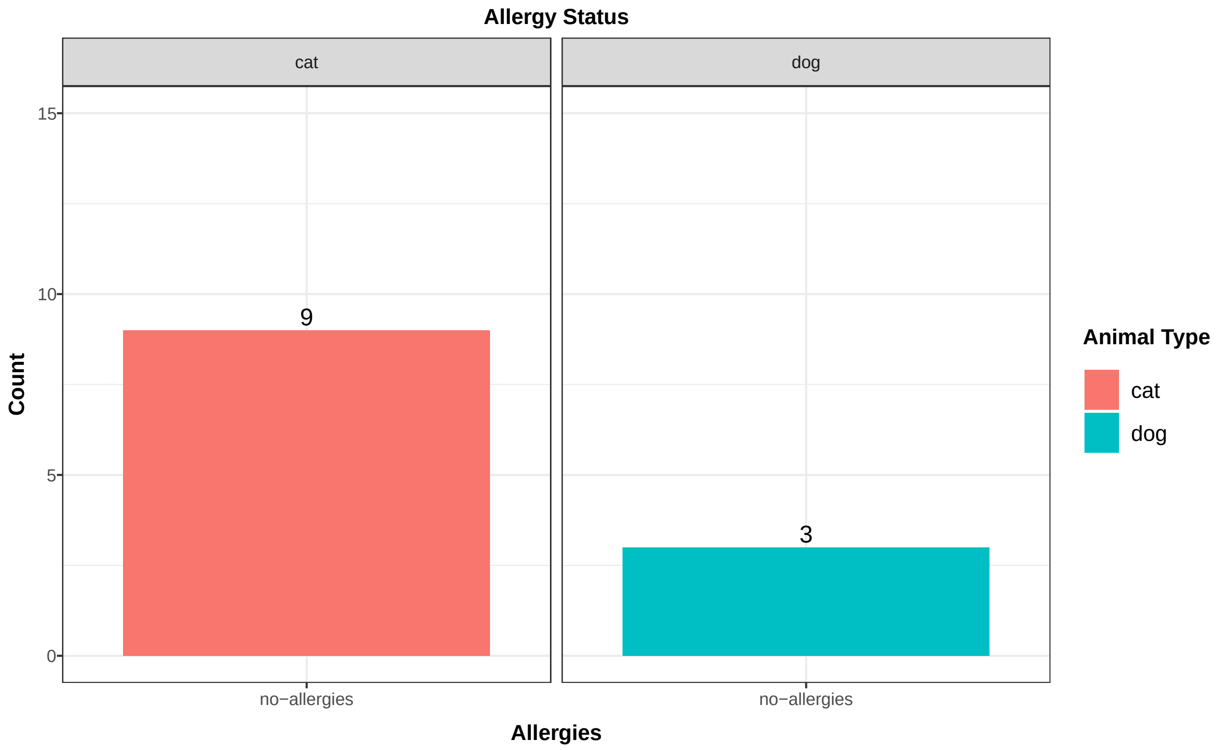


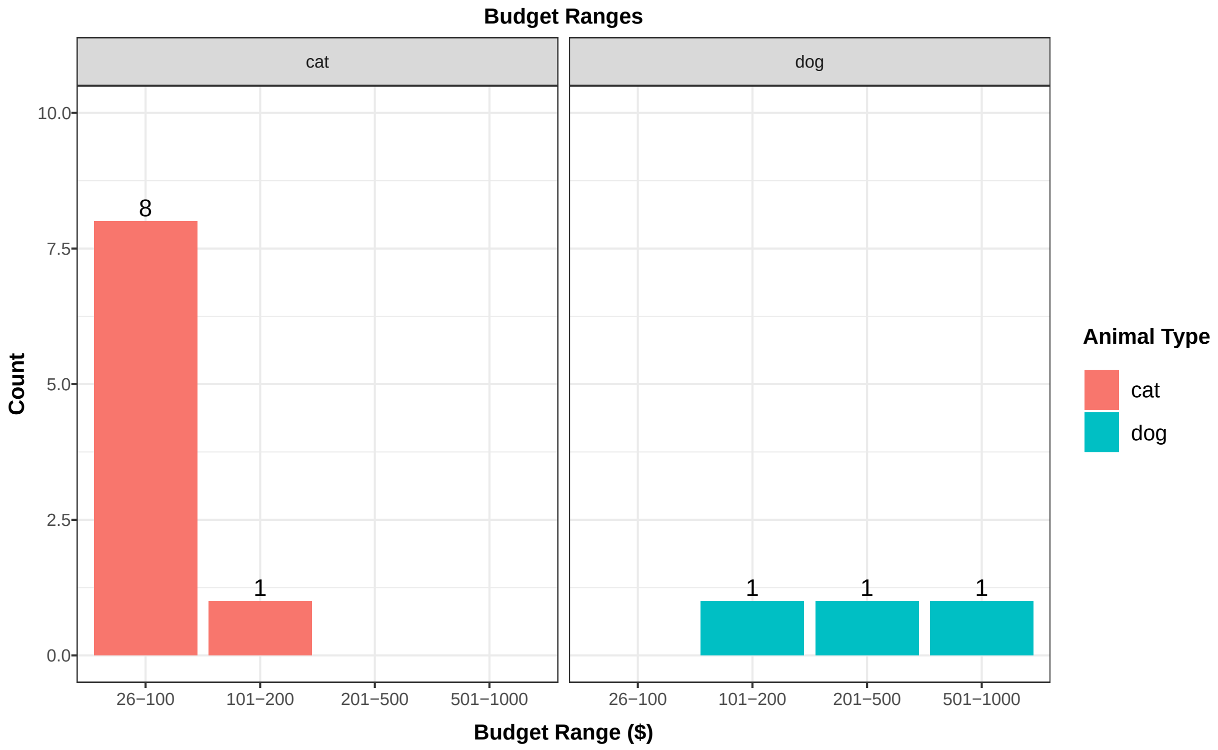
Most of the people who adopted a pet had no allergies.

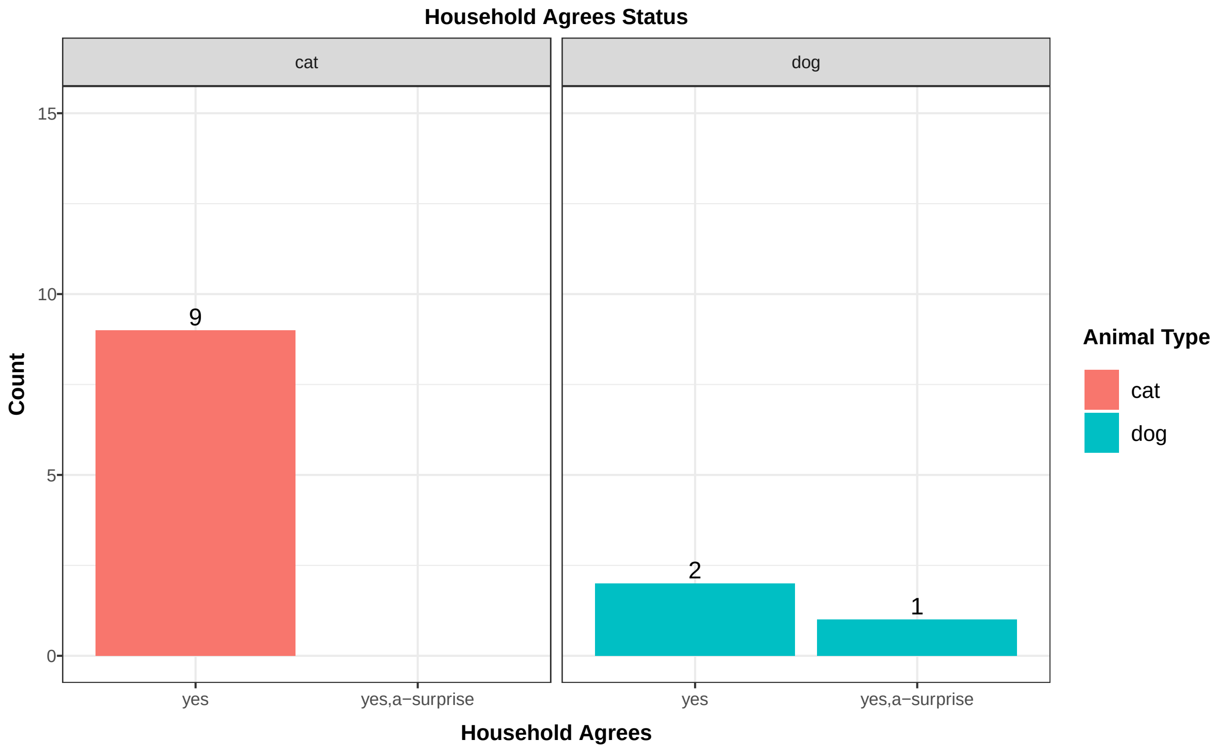
Denied and Red Flagged Applications

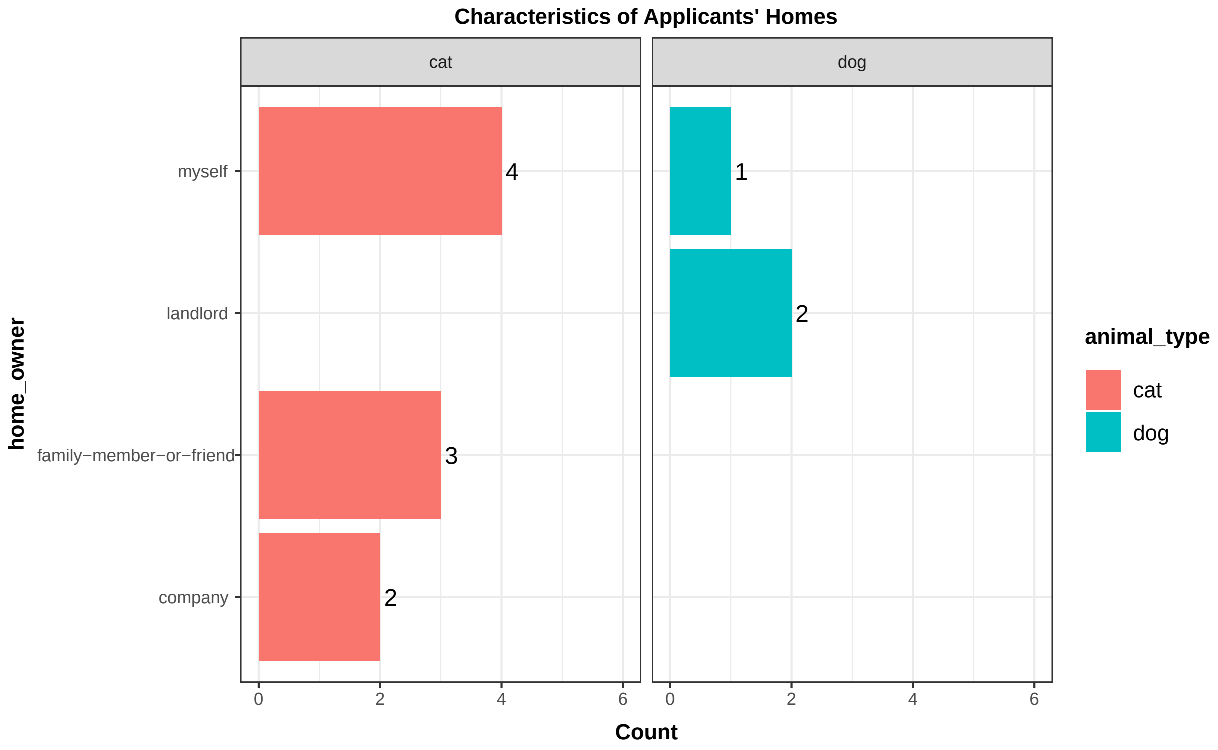
We further investigated the characteristics of applications that were denied or red flagged. There were 12 applications that were denied, 19 that were withdrawn, and 133 that were red flagged. **Denied Applications** Below are visualizations that illustrate the applicants' characteristics (e.g. allergies, budget, home pet policy, etc.). 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.

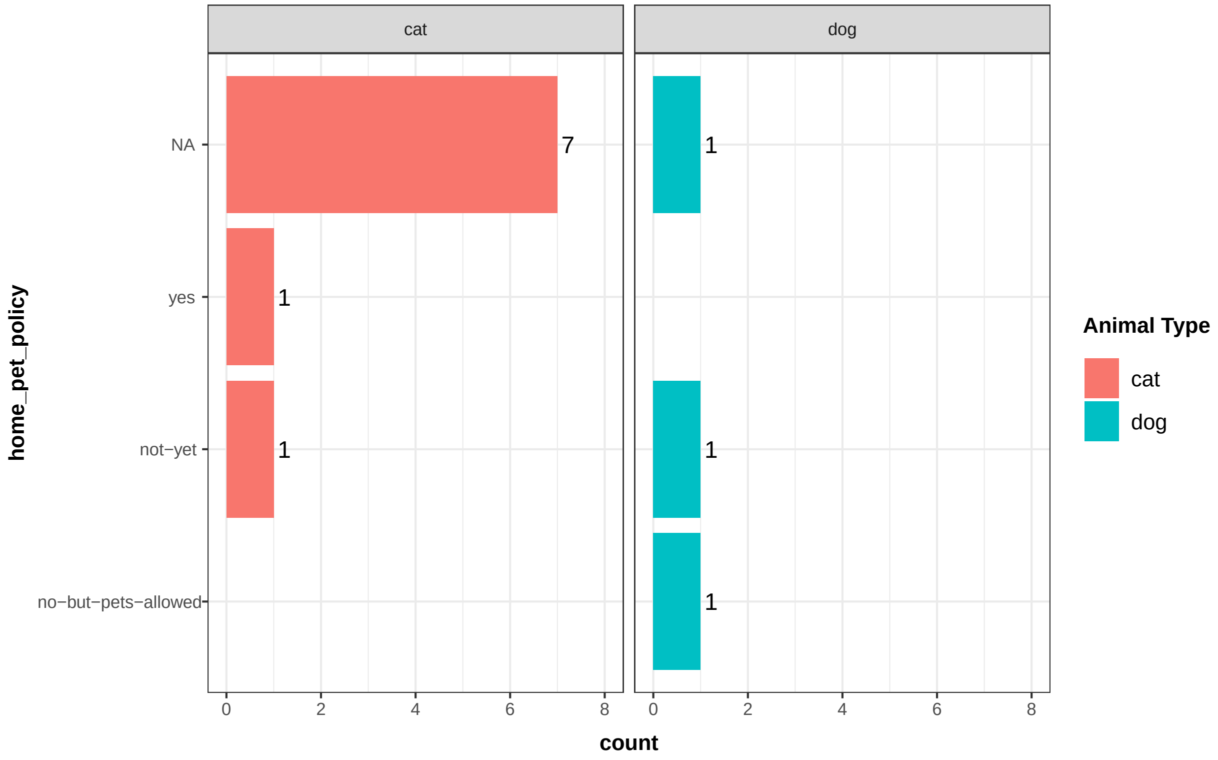
Key takeaways: * No known allergies for the applicants * Budget had no impact (same budget range for approved applications) * All household members agreed to get a pet * Majority of the applicants did not enter a home pet policy and not everyone is the home owner * Many applicants had unfortunate incidents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

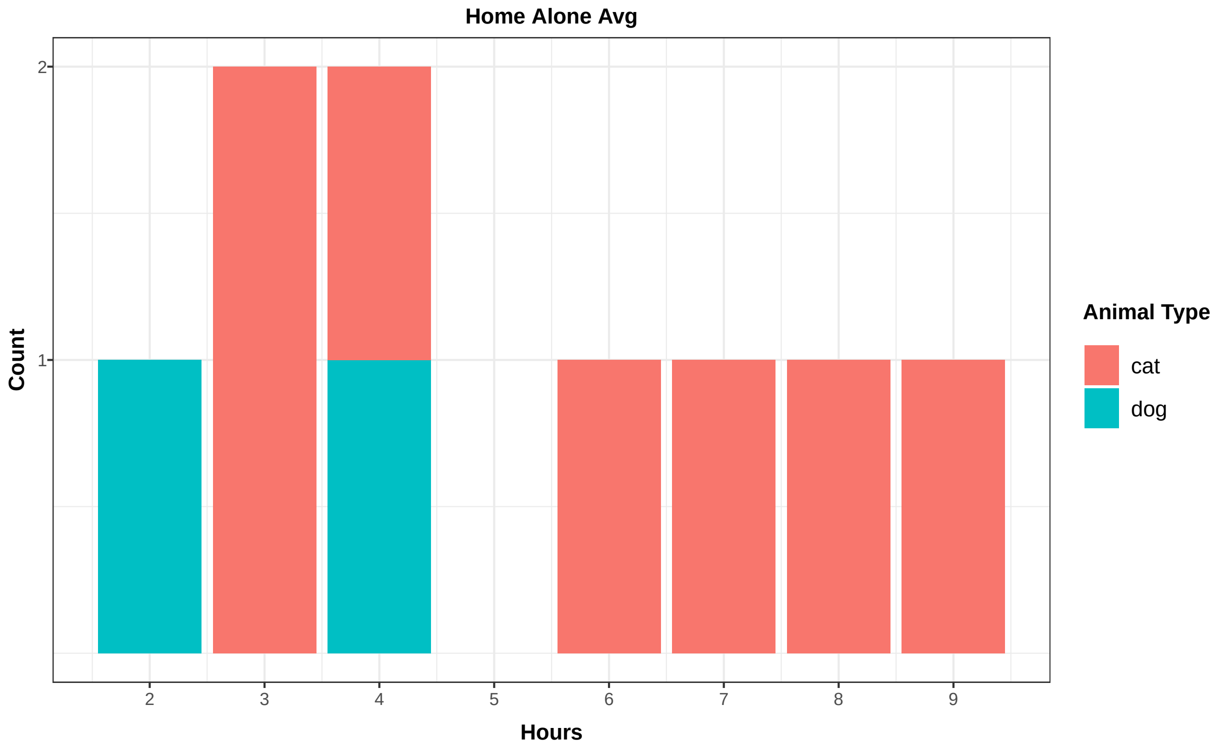

Allergy Status
cat
dog
15
10
5
0
Count
9
3
no−allergies
no−allergies
Allergies
Animal Type
🟥 cat
🟦 dog

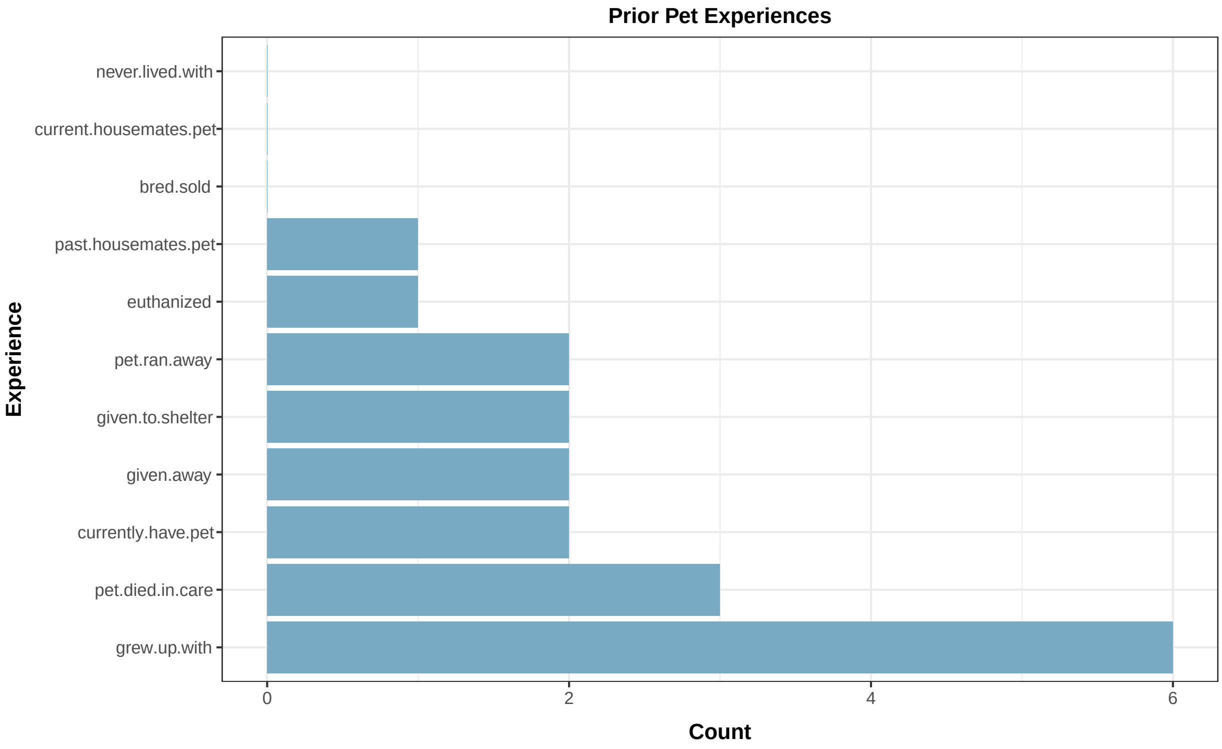

Budget Ranges
cat
dog
10.0
7.5
5.0
2.5
0.0
Count
8
1
1
1
1
26−100
101−200
201−500
501−1000
26−100
101−200
201−500
501−1000
Budget Range ($)
Animal Type
🟥 cat
🟦 dog


Household Agrees Status
cat
dog
Count
15
10
5
0
9
2
1
yes
yes,a−surprise
yes
yes,a−surprise
Household Agrees
Animal Type
cat
dog


Characteristics of Applicants' Homes
cat
dog
home_owner
myself
landlord
family−member−or−friend
company
4
3
2
1
2
0
2
4
6
0
2
4
6
Count
animal_type
cat
dog


cat
dog
home_pet_policy
NA
yes
not−yet
no−but−pets−allowed
7
1
1
1
1
1
0
2
4
6
8
count
Animal Type
cat
dog

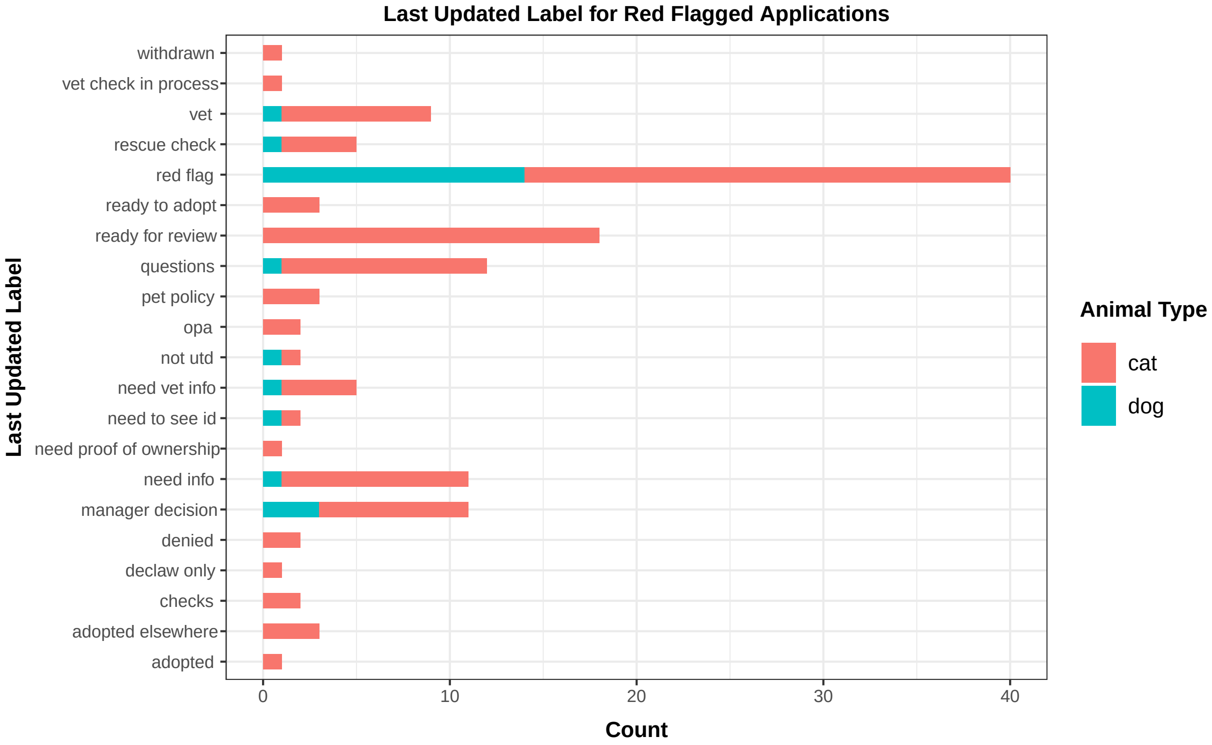

Home Alone Avg
Count
2
1
Hours
2
3
4
5
6
7
8
9
Animal Type
cat
dog



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 procesed. 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.

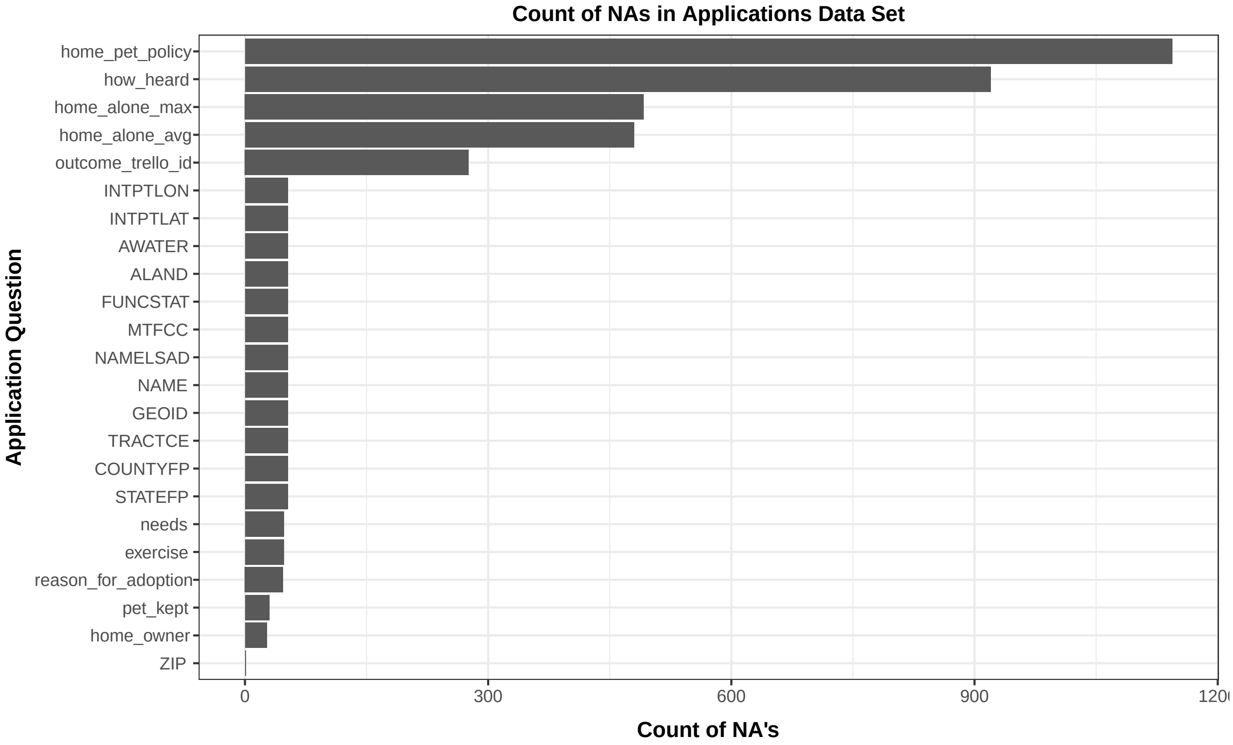
Below is a visualization that shows the last updated status for applications that were red flagged. After being flagged, the applications were sent to the manager to make a decision or the applicant was requested to provide more information (e.g. in many cases the applicant was required to provide more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



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
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 a 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.

Important Features for Prediction

```
## Warning in min.default(structure(c(NA_real_, NA_real_, NA_real_,
## NA_real_, : no non-missing arguments to min; returning Inf

## Warning in max.default(structure(c(NA_real_, NA_real_, NA_real_,
## NA_real_, : no non-missing arguments to max; returning -Inf

## Warning in min.default(structure(c(NA_real_, NA_real_, NA_real_,
## NA_real_, : no non-missing arguments to min; returning Inf

## Warning in max.default(structure(c(NA_real_, NA_real_, NA_real_,
## NA_real_, : no non-missing arguments to max; returning -Inf
```

Until now, we have separately analysed the different characteristics that affect adoption or decline. In an attempt to understand how the different features in the dataset could have had a combined effect on the adoption status, we 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.

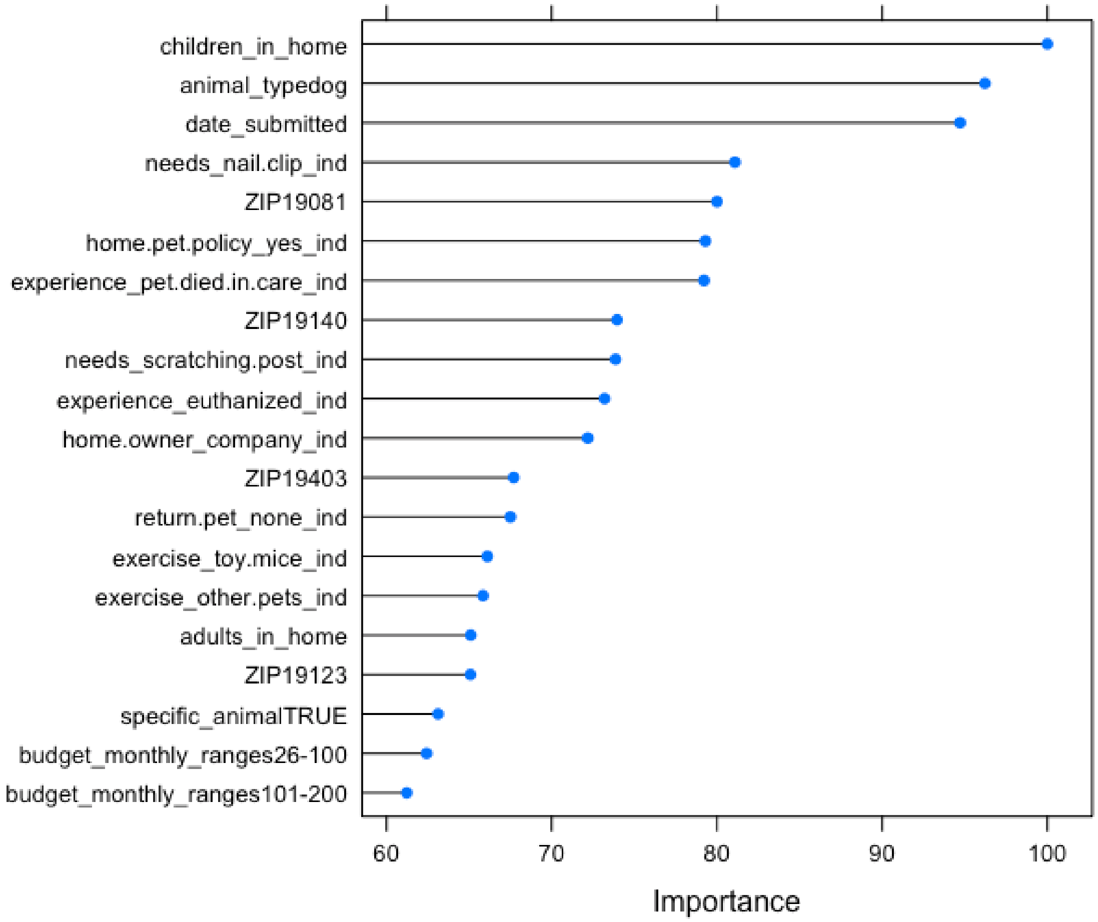


Figure 1: Important variables

3. Geographic factors

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 Orchestrall, 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.

Goals

These analyses examined the data in relation to geographic and population parameters, with two main objectives:

1) identify an initial set of variables that may be informative for application processing
2) provide a basis for discussion around the usefulness of geographical data analysis for PAWS at a broader level

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, 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. This finding aligns 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) - although it is unclear how recent these estimates are.

Using the "complete" status of a trello card at the time when the data were pulled, we did not observe a neighborhood wealth difference between completed and non-completed applications. While a "complete" status is fairly vague (it does not indicate the outcome of an application), and several trello cards may have been incomplete due to them being fairly recent, the data do not indicate an economic bias when processing applications.

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). Additionally, red flagged **cat** applicants have a lower estimated monthly budget than their non-red-flagged counterparts ($176 vs. $224). For **dogs**, a similar trend was observed, but it did not reach the statistical significance threshold ($212 vs. $277). 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. $1,446 for **cats** and $735 vs. $1,848 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 Philadephia. 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.

Additionally, in neighborhoods with a higher percentage of the population 25 to 34 year old enrolled in school, we also observed a significantly higher proprotion 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).

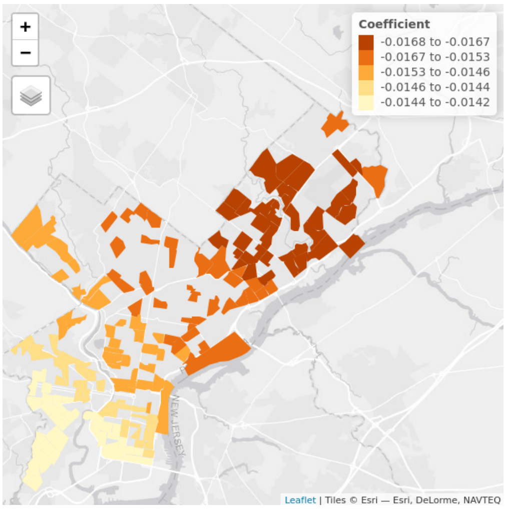


Figure 2: Cell data plan only coefficients

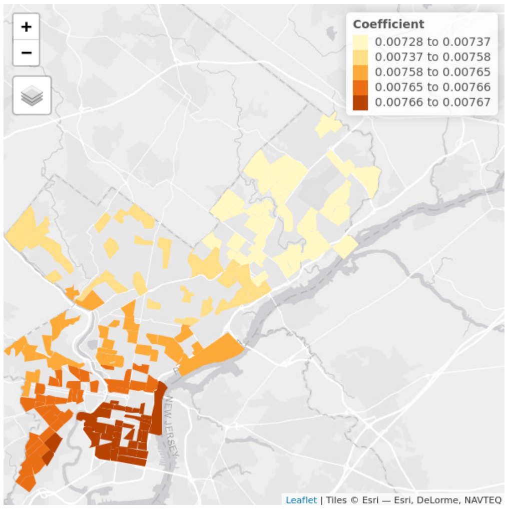


Figure 3: Population 25-34 enrolled in school coefficients

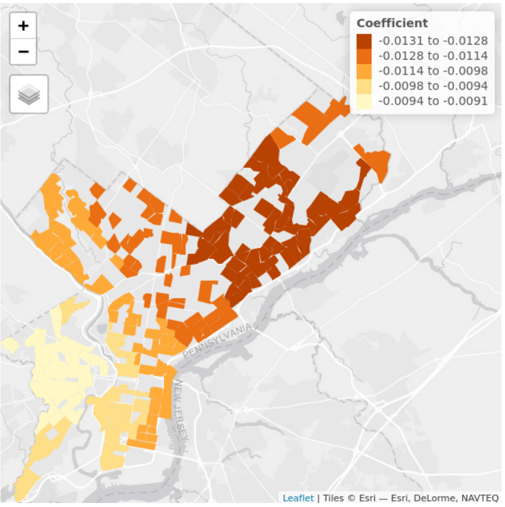


Figure 4: Children grades 5-8 coefficients

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 mbile-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

5. Data considerations

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.