Calvin's Paws Analysis

David Culhane

2024-05-09

Introduction - Rescue & Calvin's Paws (Part 1)

Calvin's Paws is a cat rescue charity based out of Raleigh, NC. They work in the Raleigh community to provide humane care of cats through medical care and adoption to work towards ending cat euthanasia in shelters and promoting no-kill philosophies and TNR (trap, neuter, and release) of feral cats. The organization was initially founded in 2006 under a different name and became Calvin's Paws in 2012.

Research Questions

- How much of a factor is a cat's breed in its adoptability?
- How much of a factor is a cat's color in its adoptability?
- How much of a factor is the sex of the cat in its adoptability?
- Are there other factors that positively affect a cat's adoptability?
- What factors make a cat less adoptable?
- Is there a correlation between the amount of money a rescue spends on medical expenses for a cat and the time a cat spends in the rescue?
- What reasons would cause an adoption applicant to have their application denied?

Approaching the Research Questions

In order to address the research questions, quite a bit of cutting through the data will be needed. Curating the data will require cutting older data from before 2014 and cutting animals that are either not cats or not belonging to Calvin's Paws. Histograms can be used to show frequencies of interest while bar graphs could be used to compare summary statistical values gathered through the analysis like means and standard deviations between breeds. Once the various frequencies of categorical data are found, other statistics can be found from those including means, standard deviations, and quartile ranges. These can be used to address how much, or little, time a cat spends in the rescue. This metric of time will be used to address ow adoptable a cat is. Coming up with some kind of formula to rate adoptability would become to subjective and time as a single value can work well in that role. Using time as our measure of choice will be the baseline for all analyses, whether they are pointed towards physical traits of the cats or others.

The cost data analysis will use money and time together to see if there is any correlation between length of adoptability.

Datasets and Packages

The data being used in this analysis came from my fiance, who ran Calvin's Paws' database from 2014-2020. She has continued to pitch in and help occasionally with managing/curating their database when asked. Since she still had access to the data, she asked if it was okay to be used for this analysis and was given the green light to send me the data.

When my fiance took over database maintenance in 2014, the rescue switched platforms. For that reason, we will only be using data from 2014 onward in this analysis. In addition to pre-2014 data, some entries in the data set are not actually cats. The occasional dog has gone through the rescue. They will be easily identifiable since their rescue ID begins with a D instead of an F. Other data with an M or a C as the first letter in the rescue ID will be removed as well since they are not affiliated with Calvin's Paws. M stands for "Medical care only" and C stands for "Courtesy post."

The files being used for analysis include an adoptability report ("Adoptability Report v2.csv"), medical expense report ("Adoptions Report with Cost.csv"), and a denied applications report ("Denied Applications with Comments.csv"). The adoptability report has fields including a cat's rescue ID, rescue/foster name, their birthday, gender, breed, coloration, status, adoption date, and a euthanasia date (if one occurred). The adoption report with cost has fields including cat's rescue ID, name, adopter, adoption date, time in the rescue, total spent on medical care, and fees paid for the cat's adoption. The denied applications report has rescue IDs, form IDs, and form submission comments for why the application was denied.

Packages to include for this analysis will include dplyr and magrittr, ggplot2 and pastecs, and string. String analysis will be needed for the denied applications report to try and process the text data from there. The stat.desc() function from the pastecs library will also be handy for quick analysis of the various histograms that will be drawn up. The readxl library will be needed to read the CSV data into R as well. ggplot2 will be handy in plotting the various data we collect and analyze as well.

Helpful Plots

Histograms will be heavily used in this analysis to track frequencies in different groups of cats - those that are adopted quickly, those in the program for a long time, the number of times various reasons appear in comments on the denied applications report, and more. These will be able to quickly identify notable factors that affect a cat's prospects. A good amount of the data is categorical as well, which histograms lend themselves nicely to.

Future Steps

At this point in the DSC 520 course, I believe I have the tools needed to clean the data and analyze it. Text parsing comes in the strings library while the other libraries previously mentioned will help bring the code into a workable state and process the data for analysis. A large amount of the data is categorical, so linear or logistical regressions are unlikely to come up without an unreasonable number of dummy variables being used with an incredibly long set of parameters in a multiple regression.

Data Cleaning (Part 2)

To clean the data, we'll need use of readxl (for loading), dplyr (for ease of use), and magrittr (for pipes). Then we can begin cleaning and putting the tables together. stringr will also be needed since the unique IDs for each cat's record are strings.

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(stringr)
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following object is masked from 'package:magrittr':
##
##     extract

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
adoptability_report_raw <- read.csv("Adoptability Report v2.csv")
cost_report_raw <- read.csv("Adoptions Report with Cost.csv")
denied_applications_raw <- read.csv("Denied Applications with Comments.csv")
```

The data being used for this analysis will begin in 2014. The unique ID for each cat that went through Calvin's Paws has the same structure, FYY-###. F indicates feline, YY, is the two digit year that the cat entered the rescue, and ### is the number of the animal for that year. We will want to filter through all Rescue.IDs for those that are F14-### or later. Since this will involve reading strings and repeating it 10 times for each of the three datasets, it would be wise to write a function to do this for us. I'll call it Rescue.ID.cleaner().

year.cleaner() will be able to take a dataset to select the Rescue.IDs in the dataset that match the string passed in via vector of strings. A for-loop will be used since it will allow the process to run each string one at a time since we're looking for individual parts of a string for each line.

```
Rescue.ID.cleaner <- function(dataset, strings)
{
  # To start, we will need an empty data frame with a number of columns equal
  # to that in the dataset being used in the function
  m <- matrix(ncol = ncol(dataset))
  df <- data.frame(m)
  colnames(df) <- colnames(dataset)
  # Initializes the dataframe to store the data into. We can use rbind() to
  # append the data we get later
  # We can use a for loop to be able to test each string supplied for the lines
  # that contain the string.
  for(prefix in strings)
  {
    indices <- str_detect(string=dataset$Rescue.ID, pattern=prefix)
    results <- dataset[indices, colnames(dataset)]
    df <- rbind(df, results)
  }
  return(df)
}
```

Now that Rescue.ID.cleaner() has been written, the years we're looking for can be specified as they're written in the Rescue.IDs and it can be run on our 3 datasets.

```
years_used <- c("F14", "F15", "F16", "F17", "F18", "F19", "F20", "F21", "F22", "F23")
adoptability_report_14on <- Rescue.ID.cleaner(adoptability_report_raw, years_used)
cost_report_14on <- Rescue.ID.cleaner(cost_report_raw, years_used)
denied_applications_14on <- Rescue.ID.cleaner(denied_applications_raw, years_used)
```

A side-effect of how Rescue.ID.cleaner() is written is that the first row of each of the 14on data frames is blank. They are NAs and can be removed.

```
adoptability_report_14on <- adoptability_report_14on %>% na.omit()
cost_report_14on <- cost_report_14on %>% na.omit()
denied_applications_14on <- denied_applications_14on %>% na.omit()
```

The adoptability report's Status column will need to be filtered to only have cats that were actually adopted. We will also want to make sure we can see, and compare, how long it was in the rescue. Functions from dplyr will be of use here for filtering and adding a variable column.

```
adoptability_report_14on <- adoptability_report_14on %>% filter(Status == 'Adopted')
```

Next should be converting the dates columns in each dataset into actual dates.

```
# For the adoptability report
adoptability_report_14on$Birthdate <- as.Date(
  adoptability_report_14on$Birthdate, format = "%m/%d/%Y")
adoptability_report_14on$Created <- as.Date(
  adoptability_report_14on$Created, format = "%m/%d/%Y")
adoptability_report_14on$Adopted.Date <- as.Date(
  adoptability_report_14on$Adopted.Date, format = "%m/%d/%Y")

# For the cost report
cost_report_14on$Created <- as.Date(
  cost_report_14on$Created, format = "%m/%d/%Y")
cost_report_14on$Adopted.Date <- as.Date(
  cost_report_14on$Adopted.Date, format = "%m/%d/%Y")
cost_report_14on$Date <- as.Date(
  cost_report_14on$Date, format = "%m/%d/%Y")
cost_report_14on$Submitted <- as.Date(
  cost_report_14on$Submitted, format = "%m/%d/%Y")

# For the denied applications
denied_applications_14on$Submitted <- as.Date(
  denied_applications_14on$Submitted, format = "%m/%d/%Y")
```

Now that the dates have been converted, we can create a field that would show the number of days each cat spent in Calvin's Paws. At the same time, the Euthanasia.Date field is mostly blank. We can remove it with dplyr's select.

```
adoptability_report_14on$Residency <- as.numeric(
  difftime(adoptability_report_14on$Adopted.Date, adoptability_report_14on$Created, units = "days"))
adoptability_report_14on <- adoptability_report_14on %>% select(-Euthanasia.Date, -Status)
```

The cost report document's costs also have an unfortunate translation from the database software's export process. Dollar amounts were translated to $ ##.##. The $ will need to be removed so the values can be treated as numbers instead of characters. This will have to be done for the Cost, and Fee.Paid columns.

```
cost_report_14on$Cost <- str_replace(
  string=cost_report_14on$Cost, pattern=" ", replacement="")
cost_report_14on$Cost <- str_replace(
  string=cost_report_14on$Cost, pattern="\\$", replacement="")
cost_report_14on$Cost <- as.numeric(cost_report_14on$Cost)
cost_report_14on$Fee.Paid <- str_replace(
  string=cost_report_14on$Fee.Paid, pattern=" ", replacement="")
cost_report_14on$Fee.Paid <- str_replace(
  string=cost_report_14on$Fee.Paid, pattern="\\$", replacement="")
cost_report_14on$Fee.Paid <- as.numeric(cost_report_14on$Fee.Paid)
```

With the dollar amounts now in numeric form, they will be easier to work with for graphing and other forms of analysis.

Now that the data has been cleaned, we will want to present some summary statistics about the data since simply displaying the copious amounts of data would be problematic.

Adoptability Report Statistics

```
length(unique(adoptability_report_14on$Rescue.ID))
```

```
## [1] 1430
```

```
length(unique(adoptability_report_14on$Primary.Breed))
```

```
## [1] 17
```

```
stat.desc(adoptability_report_14on$Residency)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##   1425.000000     11.000000      5.000000   -817.000000   2145.000000
##         range           sum        median          mean       SE.mean
##   2962.000000 279916.000000    107.000000    196.432281      7.317851
##  CI.mean.0.95           var       std.dev      coef.var
##     14.354926  76310.103733    276.242835      1.406301
```

```
summary(adoptability_report_14on$Residency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -817.0    65.0   107.0   196.4   195.0  2145.0       5
```

```
sum(adoptability_report_14on$Residency > 365, na.rm = TRUE)
```

```
## [1] 174
```

```r
sum(adoptability_report_14on$Residency > 561, na.rm = TRUE)
```

```
## [1] 108
```

The data in the Adoptability report from 2014-onward is the form of the original report after cleaning the data. There are 1430 different cats that have gone through Calvin's Paws from 2014-2023. Of those 1430 cats, only 5 did not have adoption dates listed. There are 11 cats who had 0 days of residency, they were adopted on the day that they entered the Calvin's Paws program. I ran pastecs' stat.desc to get a quick summary of the statistics surrounding the amount of time each cat spent with Calvin'sPaws before getting adopted. The range highlights the largest outliers within the dataset, a cat who did not enter the system until more than two years (817 days) after their adoption, Juane. Captain S on the other end of the spectrum spent almost 6 years (2145 days) as a foster cat before finally being adopted, Captain S. The median amount of time spent in the rescue by a Calvin's Paws cat was 107 days, around three and a half months. The mean time spent was around 196 days, around six and a half months. The standard deviation of time spent with Calvin's Paws for a cat is 276.24 days. The mean and standard deviation are likely driven up by various outliers. 174 cats had stays longer than a year and 108 of those were one standard deviation above the mean. The top of the upper quartile was 195 days, which means that some of these outliers are so far away that the mean was dragged outside of the interquartile range. When further cutting the data apart for analysis, this would be something to address. This would make the median a better statistic to use in certain cases when incorporating the entire dataset.

Additionally, 17 breeds of cats appear in the dataset. When doing the full analysis, it would be wise to compare how the breed of a cat can factor into the amount of time it stays in the rescue/how much it could contribute to it being adopted quickly.

The five NA's in the adoptability report can be removed since they will not contribute to our analysis.

```r
adoptability_report_14on <- adoptability_report_14on %>%
  filter(is.na(Adopted.Date) == FALSE)
```

Cost Report Statistics

```r
length(cost_report_14on$Rescue.ID)
```

```
## [1] 1405
```

```r
length(unique(cost_report_14on$Rescue.ID))
```

```
## [1] 1327
```

```r
stat.desc(cost_report_14on$Cost)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.405000e+03 5.900000e+01 0.000000e+00 0.000000e+00 4.627900e+03 4.627900e+03
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 2.060826e+05 1.160400e+02 1.466780e+02 5.313594e+00 1.042344e+01 3.966916e+04
##      std.dev     coef.var
## 1.991712e+02 1.357880e+00
```

```r
summary(cost_report_14on$Cost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   90.22  116.04  146.68  142.93 4627.90
```

The Cost Report is a cousin of the Adoptability Report. It only contains cats that were adopted but it also includes the total cost spent on each individual cat during their stay with Calvin's Paws. There are 1405 total lines in the report but only 1327 unique Rescue.IDs. This means that there are 78 cats who appear more than once in the report for one reason or another. In the next step, investigating these instances will provide additional information. If this table is to be merged with the Adoptability Report's, this will have to be addressed since duplicate Rescue.IDs would prove to be problematic.

The summary/stat.desc() statistics for this report focus on the cost. Costs accrued for cats with Calvin's Paws include spay/neuter, vaccinations, and disease testing (FIV and FeLV, mostly). 59 cats had null/0 values, indicating no money was spent according to the record. The highest amount at the top of the range was $4,627.90. The median cost for medical expenses for an individual cat came to $116.04 while the mean was $146.68. Unfortunatley this also appears outside of the interquartile range. This would be worth investigating again after excluding potential outliers. The standard deviation of money spent on each cat came to $199.17. This statistic would also need to be re-evaluated if a look without outliers is taken later.

Denied Applications Statistics

Since the denied applications report reasons are purely text, sample statistics for these will involve screening the comments of denial reasons for important factors. Calvin's Paws has a few reasons why they would deny an applicant and the include:

- Intention to declaw the cat or has declawed a cat before
- Intention to adopt only a single kitten into a home with no other cats
- Intention to allow the cat to go outdoors
- Living outside a 60 mile radius from Calvin's Paws
- The cat being applied for already having a pending adoption
- The applicant being under 21
- The applicant living in under a lease where either no or no more pets are allowed or other landlord issues.
- The applicant not vaccinating their own cats or wanting the cat being adopted to be vaccinated

So let's see how things break down for the denied applications after the start of 2014. It is also worth noting that any one application denial could have more than one reason listed as well.

```r
length(denied_applications_14on$Rescue.ID)
```

```
## [1] 382
```

```r
length(unique(denied_applications_14on$Rescue.ID))
```

```
## [1] 265
```

```r
sum(str_detect(denied_applications_14on$Comments..inline., pattern="declaw"))
```

```
## [1] 74
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="one"))
```

```
## [1] 48
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="outdoor"))
```

```
## [1] 17
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="Lives in"))
```

```
## [1] 13
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="pending"))
```

```
## [1] 2
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="under 21"))
```

```
## [1] 9
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="apartment")
    + str_detect(denied_applications_14on$Comments..inline., pattern="condo")
    + str_detect(denied_applications_14on$Comments..inline., pattern="landlord"))
```

```
## [1] 16
```

```
sum(str_detect(denied_applications_14on$Comments..inline., pattern="vaccine"))
```

```
## [1] 9
```

There are 382 applications to adopt sent to Calvin's Paws that were denied between 2014 and 2023. Of those, 265 different cats were applied for. That means 117 different cats received multiple applications!

Declawing is now largely seen as inhumane and is grounds for summary disqualification. If an adopter states in their application that they intend to declaw the cat they are going to adopt, they are essentially blacklisted from the adoption process with Calvin's Paws. 74 of the denied applications mentioned declaw in their comments. Adopting only a single kitten can lead to behavioral issues because kittens are incredibly social creatures - it's how they learn. Adopting a kitten under the age of 6 months can lead to aggressive or other behavioral issues and can commonly lead to those cats being relinquished to shelters, returned to rescues, or even worse - just kicked out the door. For that reason, 48 applications were denied because the applicant wanted to adopt a single kitten into a home with no other cats. Seventeen applicants mentioned wishing to allow their adopted cat to live outdoors and were denied for that reason. Thirteen applications came from people who lived outside the area that Calvin's Paws serves, Raleigh, NC, and were denied for that reason. Sixteen applications were denied for reasons related to renting/leasing and landlord agreements. Nine applications were denied since Calvin's Paws prefers to adopt to those older than 21 and nine others were denied for wishing to not vaccinate their cats.

Questions for Part 3

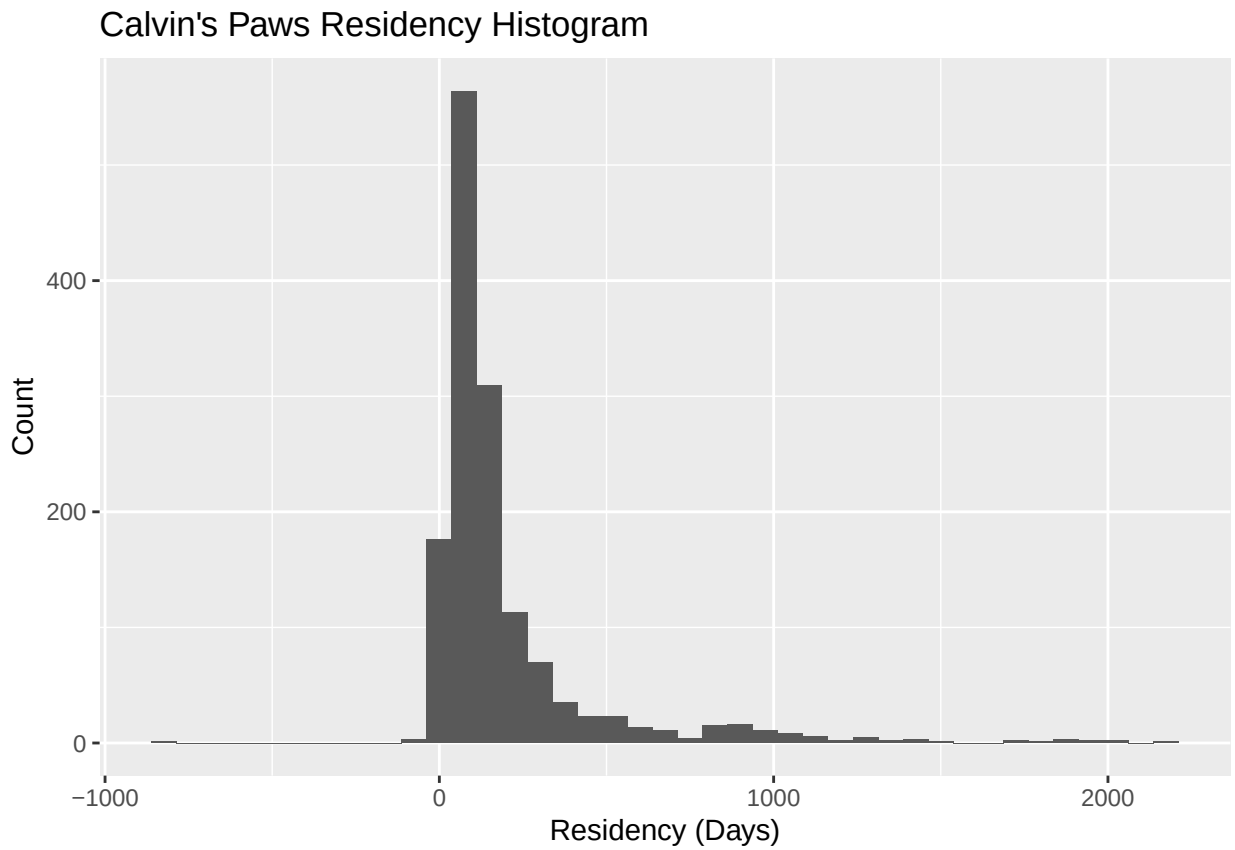
When going through final analysis, I plan on removing superfluous and sensitive data. The Adoptability Report has a euthanasia field that is mostly empty, so it would make sense to get rid of it. The Cost and Denied Application reports have personal info that should be scrubbed. I also plan on plotting and graphing for the various subsets of the data - breed and coloring within breeds if there is sufficient data to do so. Sex of the cat will also be analyzed. And if possible, an analysis comparing the names of the cats with their residency times will also be done to see if there are any trends that can be identified.

Plotting, Investigation, and the Narrative (Part 3)

General Adoptability Analysis

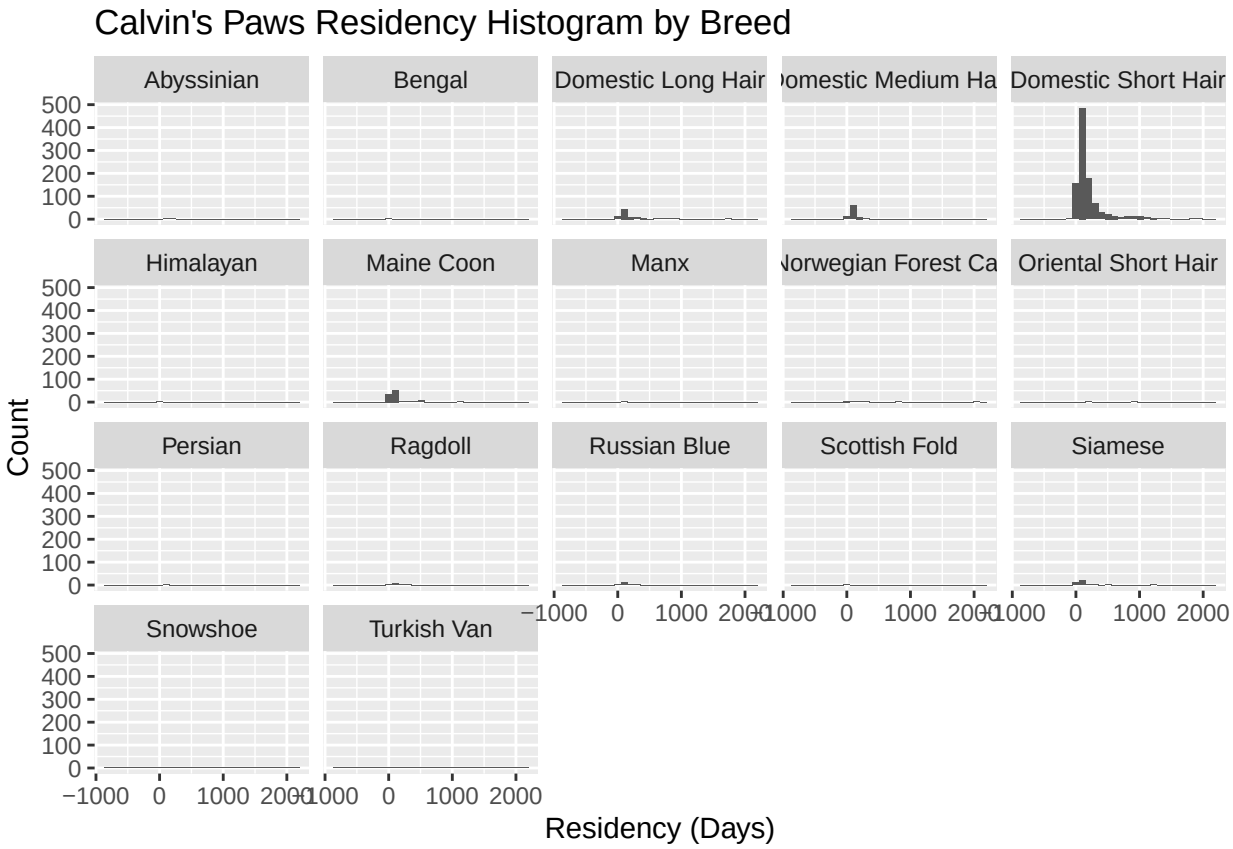
We have now cleaned our data and gotten some summary statistics regarding Calvin's Paws adoption data from 2014-2023. To visualize it and look at how much time cats spent on average in the rescue, we can use histograms for the entire population as well as by breed.

```
library(ggplot2)
adoptability_hgram <- ggplot(adoptability_report_14on) +
  geom_histogram(aes(x=Residency), binwidth = 75) + labs(x="Residency (Days)",
                                                         y="Count",
                                                         title="Calvin's Paws Residency Histogram")
adoptability_hgram
```



```
adoptability_hgram_multiplot <- ggplot(adoptability_report_14on) +
  geom_histogram(aes(x=Residency)) + facet_wrap(~Primary.Breed) +
  labs(x="Residency (Days)",
       y="Count",
       title="Calvin's Paws Residency Histogram by Breed")
adoptability_hgram_multiplot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



With the histograms plotted, we see a couple of things. From the complete histogram, we see most cats are adopted within 250 days and that there are some outliers. We could identify those by adding Residency z-scores and get a better picture by excluding them.

Breed and Post-Outlier Analysis

With the histograms broken up by breed, most of the non-Domestic labeled breeds barely even register enough to warrant a histogram of their own. Another could be made with each of those breeds and another feature, like color, could be used to differentiate between breeds in that histogram. For outliers, we will say that any datapoint whose z-score's absolute value is greater than 3 is an outlier.

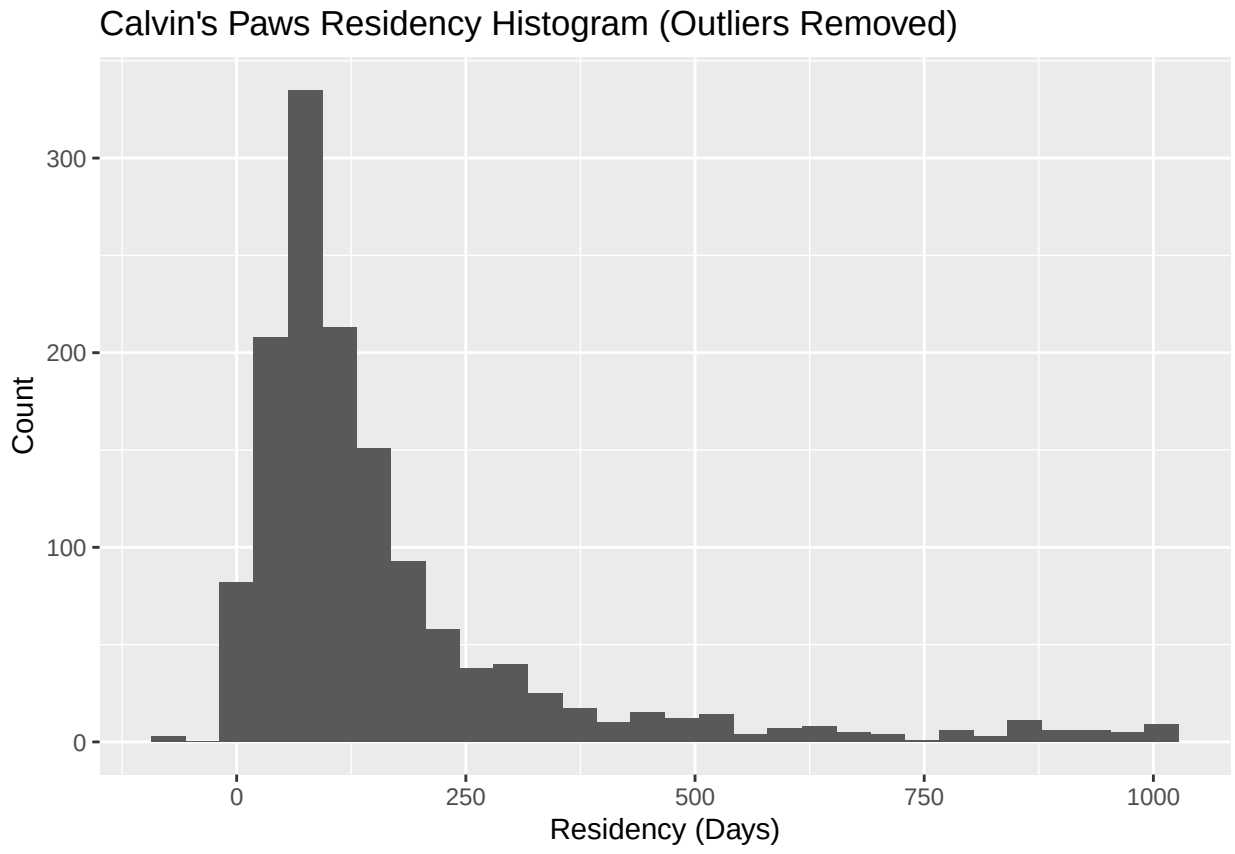
```
adoptability_report_14on$Residency_z <-
  (adoptability_report_14on$Residency - mean(adoptability_report_14on$Residency)) /
  sd(adoptability_report_14on$Residency)
adoptability_report_14on_reg <-
```

```
  adoptability_report_14on %>% filter(abs(Residency_z) < 3)
adoption_outliers <- adoptability_report_14on %>% filter(abs(Residency_z) >= 3)
other_breeds <- adoptability_report_14on_reg %>% filter(
  Primary.Breed != "Domestic Short Hair" &
    Primary.Breed != "Domestic Medium Hair" &
    Primary.Breed != "Domestic Long Hair")
```

With the outliers now filtered out of the main set and a set for non-domestic breeds created, we can check their histograms as well.

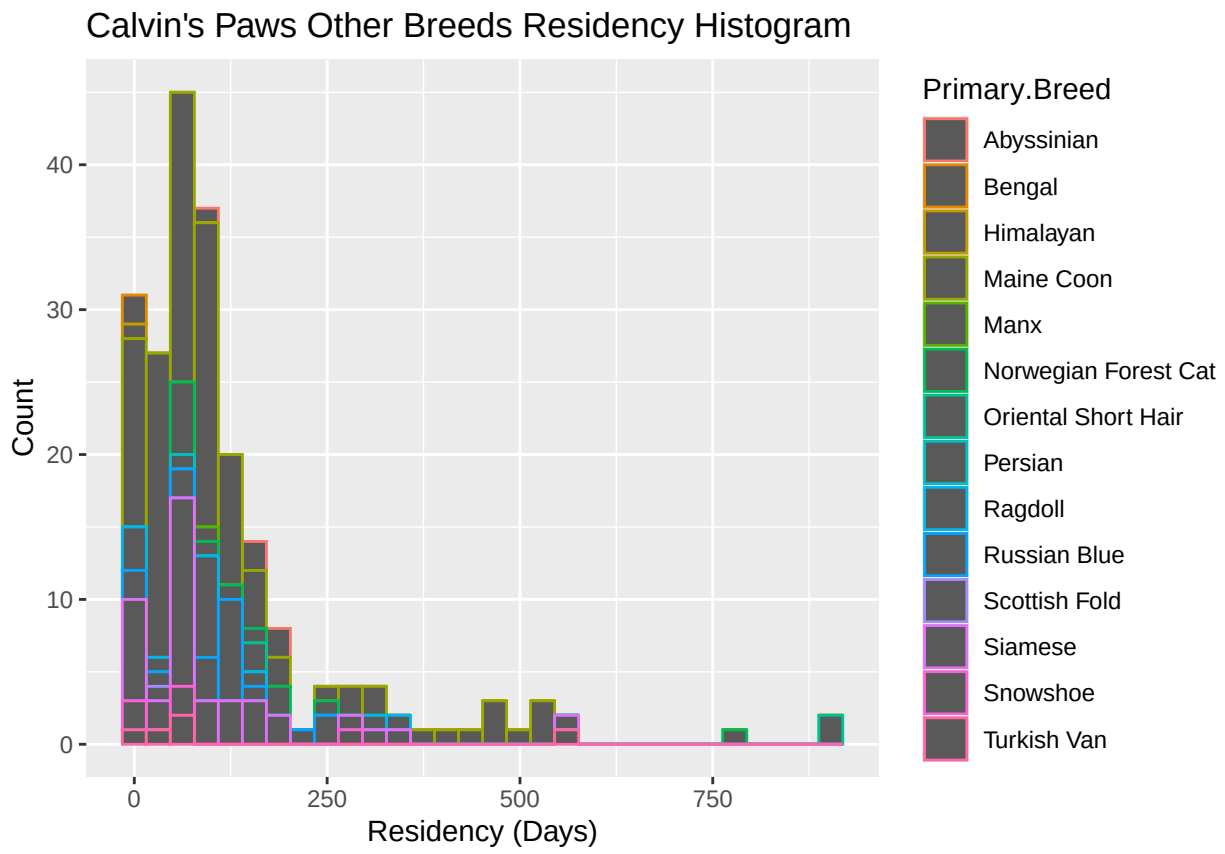
```
filtered_hgram <- ggplot(adoptability_report_14on_reg) +
  geom_histogram(aes(x=Residency)) +
  labs(x="Residency (Days)", y="Count",
       title="Calvin's Paws Residency Histogram (Outliers Removed)")
filtered_hgram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
others_hgram <- ggplot(other_breeds) +
  geom_histogram(aes(x=Residency, color=Primary.Breed)) +
  labs(x="Residency (Days)", y="Count", title="Calvin's Paws Other Breeds Residency Histogram")
others_hgram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



With the adjusted histograms plotted, we can see most adoption activity is taking place within 125 days for all cats and possibly even faster for non-domestic breeds. We can re-check summary statistics to confirm this with the adjusted data.

```
# Summary Statistics for Adjusted Adoptability Report
stat.desc(adoptability_report_14on_reg$Residency)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##   1389.000000     11.000000      0.000000    -64.000000   1020.000000
##         range           sum        median          mean       SE.mean
##   1084.000000 230420.000000    105.000000    165.889129      5.003540
##  CI.mean.0.95           var       std.dev      coef.var
##      9.815317  34774.187987    186.478385      1.124115
```

```
summary(adoptability_report_14on_reg$Residency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -64.0    64.0   105.0   165.9   184.0  1020.0
```

```
# Summary Statistics for Other Breeds
stat.desc(other_breeds$Residency)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
##   212.000000     3.000000     0.000000    -6.000000   897.000000   903.000000
```

```
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 25913.000000    78.000000   122.231132    10.090928    19.891950 21587.287557
##      std.dev     coef.var
##   146.926130     1.202035
```

```
summary(other_breeds$Residency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -6.0    42.0    78.0   122.2   137.2   897.0
```

The summary statistics we have found confirmed our suspicions. For all breeds, the median residency was 105 days. This value goes as low as 78 days for non-domestic breeds. This would provide evidence that a cat's listed breed can affect how long it takes to adopt the cat in that non-domestic breeds are adopted faster.

To check how a cat's color affects adoptability, we can analyze by color within the domestic breed cats. So we can create a domestic dataset for these purposes.

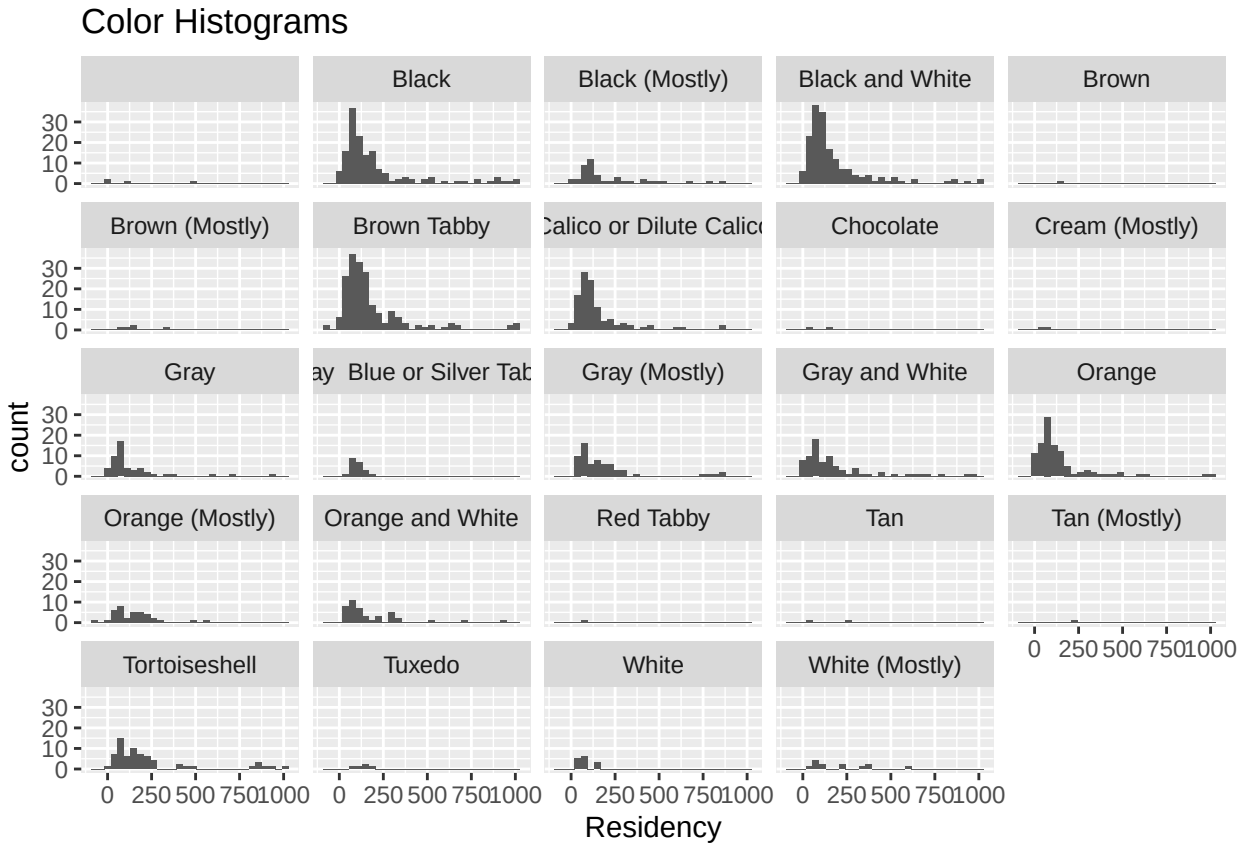
```
domestics <- adoptability_report_14on_reg %>% filter(
  Primary.Breed == "Domestic Short Hair" |
    Primary.Breed == "Domestic Medium Hair" |
    Primary.Breed == "Domestic Long Hair")
unique(domestics$Color..General.)
```

```
##  [1] "Black"                     "Black and White"
##  [3] "Orange"                    "Tortoiseshell"
##  [5] "Brown (Mostly)"            "Gray and White"
##  [7] "Calico or Dilute Calico"   "Orange (Mostly)"
##  [9] "Brown Tabby"               "Gray (Mostly)"
## [11] "Red Tabby"                 "Tuxedo"
## [13] "Tan"                       "Orange and White"
## [15] "Black (Mostly)"            "Gray  Blue or Silver Tabby"
## [17] "White (Mostly)"            "White"
## [19] "Gray"                      "Tan (Mostly)"
## [21] ""                          "Cream (Mostly)"
## [23] "Chocolate"                 "Brown"
```

Of the 1,177 cats listed with a domestic breed, only four do not have a color listed. We can use a multiplot to see which colors should have their summary statistics checked.

```
ggplot(domestics) + geom_histogram(aes(x=Residency)) +
  facet_wrap(~Color..General.) + labs(title="Color Histograms")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see from the multiple plots of histograms for the various colors that not all colors have a lot of representation in the data. If we would like to see the median and mode of colors with at least 10 cats represented, we can filter for those colors and list out a summary of mean and median residencies for each color represented with at least 10 cats.

```
domestics_threshold <- domestics %>% count(Color..General.) %>% filter(n > 10)
domestics_threshold
```

```
##               Color..General.   n
## 1                       Black 150
## 2              Black (Mostly)  44
## 3             Black and White 173
## 4                 Brown Tabby 189
## 5     Calico or Dilute Calico 106
## 6                        Gray  50
## 7  Gray  Blue or Silver Tabby  21
## 8                Gray (Mostly)  64
## 9              Gray and White  78
## 10                      Orange 105
## 11             Orange (Mostly)  37
## 12            Orange and White  43
## 13               Tortoiseshell  67
## 14                      White  14
## 15             White (Mostly)  13
```

```
domestics %>% group_by(Color..General.) %>%
  summarize(MedianResidency=median(Residency),
            MeanResidency=mean(Residency),
            SDResidency=sd(Residency)) %>%
  filter(Color..General. %in% domestics_threshold[,1])
```

```
## # A tibble: 15 x 4
##    Color..General.            MedianResidency MeanResidency SDResidency
##    <chr>                                <dbl>         <dbl>       <dbl>
##  1 Black                                  116          203.        228.
##  2 Black (Mostly)                         118.         208.        204.
##  3 Black and White                        116          178.        189.
##  4 Brown Tabby                            115          176.        192.
##  5 Calico or Dilute Calico                106.         144.        148.
##  6 Gray                                    75          138.        177.
##  7 Gray  Blue or Silver Tabby             102          104.         35.2
##  8 Gray (Mostly)                          125          187.        200.
##  9 Gray and White                         118.         187.        216.
## 10 Orange                                  91          142.        172.
## 11 Orange (Mostly)                       132          148.        126.
## 12 Orange and White                      102          176.        184.
## 13 Tortoiseshell                         140          225.        248.
## 14 White                                  82           79.7        43.4
## 15 White (Mostly)                        111          205.        172.
```

```
stat.desc(domestics$Residency)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##   1177.000000      8.000000      0.000000    -64.000000   1020.000000
##         range           sum        median          mean       SE.mean
##   1084.000000 204507.000000    110.000000    173.752761      5.588880
##  CI.mean.0.95           var       std.dev      coef.var
##     10.965290  36764.283209    191.740145      1.103523
```

```
summary(domestics$Residency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -64.0    68.0   110.0   173.8   192.0  1020.0
```

Here we get a decent picture from the summary statistics of the domestic labeled cats. The overall median residency is 110 days with a mean residency of 173.8 days. Grey, orange, and white cats have median residencies far below that of the overall group, suggesting that those colors may adopt faster than others. On the other hand, tortoiseshell (torties) cats have a median residency a whole month later (30 days) than the overall rate and their mean residency is almost two months later. Torties are known to have attitude (usually called "tortitude"), so this reputation may affect their residency times as a lengthening effect.

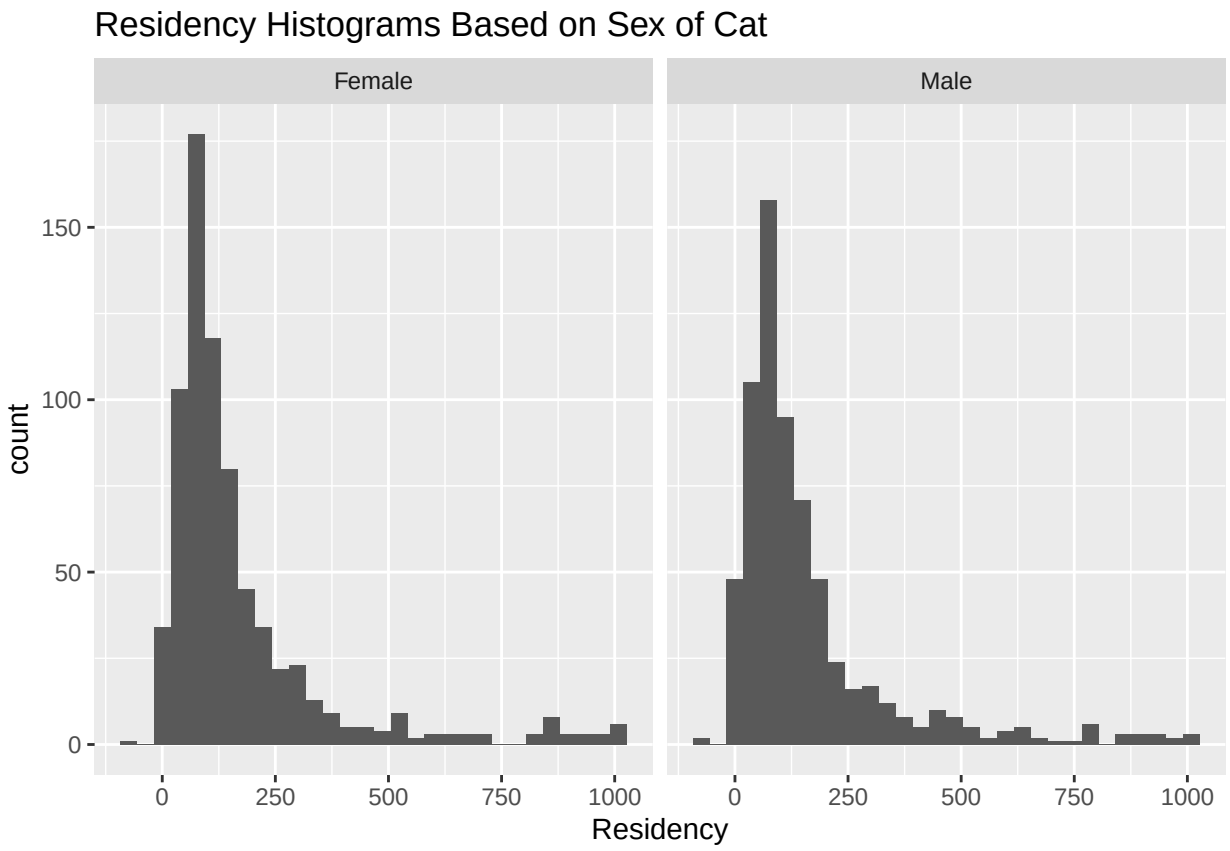
The standard deviations of the included colorations tend to be fairly consistent outside of gray/blue or silver Tabby and white cats, whose standard deviations are significantly lower. This would indicate a much more consistent amount of time spent in the rescue for those colorations. This should be especially true since there were 21 and 14 cats of those colors, respectively.

Sex of the Cat and Adoptability

When adopting a new pet, a lot of pet owners prefer one gender over another for their new pet for one reason or another. We can check if there is a bias towards male or female cats.

```
# Histogram to compare the Residencies based on sex of the cat
ggplot(adoptability_report_14on_reg) +
  geom_histogram(aes(x=Residency)) +
  facet_wrap(~Sex) +
  labs(title="Residency Histograms Based on Sex of Cat")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Creating separate datasets to check summary statistics
males <- adoptability_report_14on_reg %>% filter(Sex == "Male")
stat.desc(males$Residency)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##    667.000000      3.000000      0.000000    -64.000000   1020.000000
##         range           sum        median          mean       SE.mean
##   1084.000000 108195.000000    100.000000    162.211394      7.082344
##  CI.mean.0.95           var       std.dev      coef.var
##     13.906411  33456.449239    182.911042      1.127609
```

```
summary(males$Residency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -64.0    59.0   100.0   162.2   183.0  1020.0
```

```
females <- adoptability_report_14on_reg %>% filter(Sex == "Female")
stat.desc(females$Residency)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##    722.000000      8.000000      0.000000    -62.000000   1020.000000
##         range           sum        median          mean       SE.mean
##   1082.000000 122225.000000    107.000000    169.286704      7.062794
##  CI.mean.0.95           var       std.dev      coef.var
##     13.866098  36015.564012    189.777670      1.121043
```

```
summary(females$Residency)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -62.0    68.0   107.0   169.3   186.0  1020.0
```

Looking at the histograms and the summary statistics, I think it would be fair to say that the sex of the cat being adopted is not a factor that affects adoptability overall. The mean, median, and standard deviation each only differ by 7 days. The standard errors of each sex is also 7 days and confidence intervals are just under 14 days. This factor should not impact adoptability.

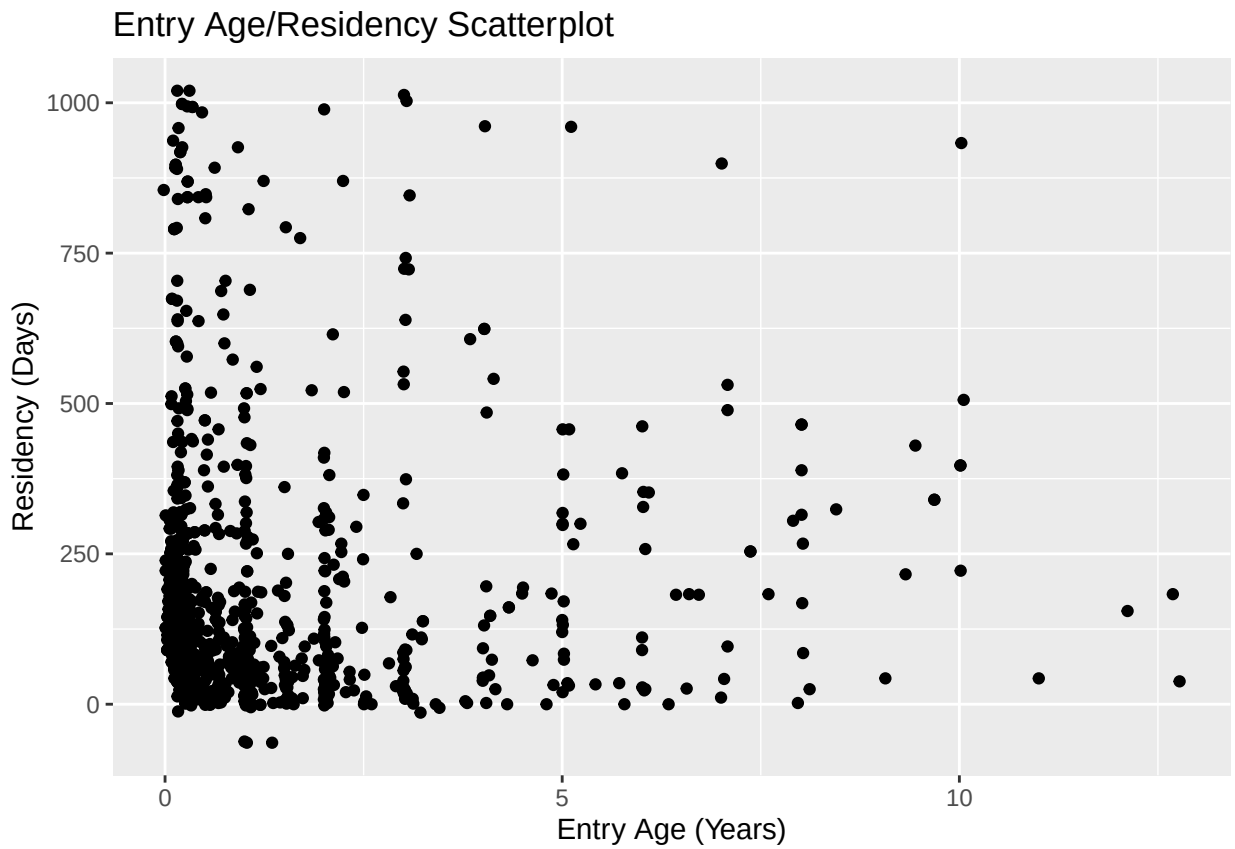
Other Adoptability Factors

We would also like to explore if other factors that can be derived from the data can impact adoptability. The first factor to look at should be the age of the cat. To check that, we can calculate the age of the cat when it entered the rescue and the age of the cat when it was adopted. Once the ages have been computed, we can make scatterplots to check if there is a correlation between the age of a cat when it enters the rescue and its residency.

It is worth noting here that in cat rescue, sometimes the birthdate of the cat is an estimate. Rescues and foster families are not always present for the birth of a cat. So there may be some estimated ages based on guessed birthdates depending on the cat.

```
# Creating fields for age at entrance to the rescue
adoptability_report_14on_reg$`Entry Age D` <- as.numeric(
  difftime(adoptability_report_14on_reg$Created,
           adoptability_report_14on_reg$Birthdate,
           units = "days"))
adoptability_report_14on_reg$`Entry Age Y` <-
  adoptability_report_14on_reg$`Entry Age D` / 365
# Now we should make sure we have fields for the age at adoption
adoptability_report_14on_reg$`Adoption Age D` <-
  adoptability_report_14on_reg$Residency + adoptability_report_14on_reg$`Entry Age D`
adoptability_report_14on_reg$`Adoption Age Y` <-
  adoptability_report_14on_reg$`Adoption Age D` / 365
```

```
entry_plot <- ggplot(adoptability_report_14on_reg, aes(x=`Entry Age Y`, y=Residency))
entry_plot + geom_point() + labs(x="Entry Age (Years)",
                                 y="Residency (Days)",
                                 title="Entry Age/Residency Scatterplot")
```



The first thing that can be noticed from the Entry Age/Residency scatterplot is the cloud of points in the lower left corner. This is likely heavily populated with kittens and they could probably use their own analysis. The second thing to notice is an unexpectedly high number of points in the upper left corner. These are kittens who stayed in Calvin's Paws for a long time for one reason or another. We can look into them in conjunction with the other datasets.

We also want to see what the general correlation is between these variables.

```
# Creating the long-haul kitten dataset for later
kittenhaulers <- adoptability_report_14on_reg %>% filter(`Entry Age Y` <= 1 & Residency >= 375) %>%
  select(-Residency_z)
# Finding the correlation value for this scatterplot.
cor(adoptability_report_14on_reg$`Entry Age Y`, adoptability_report_14on_reg$Residency)
```

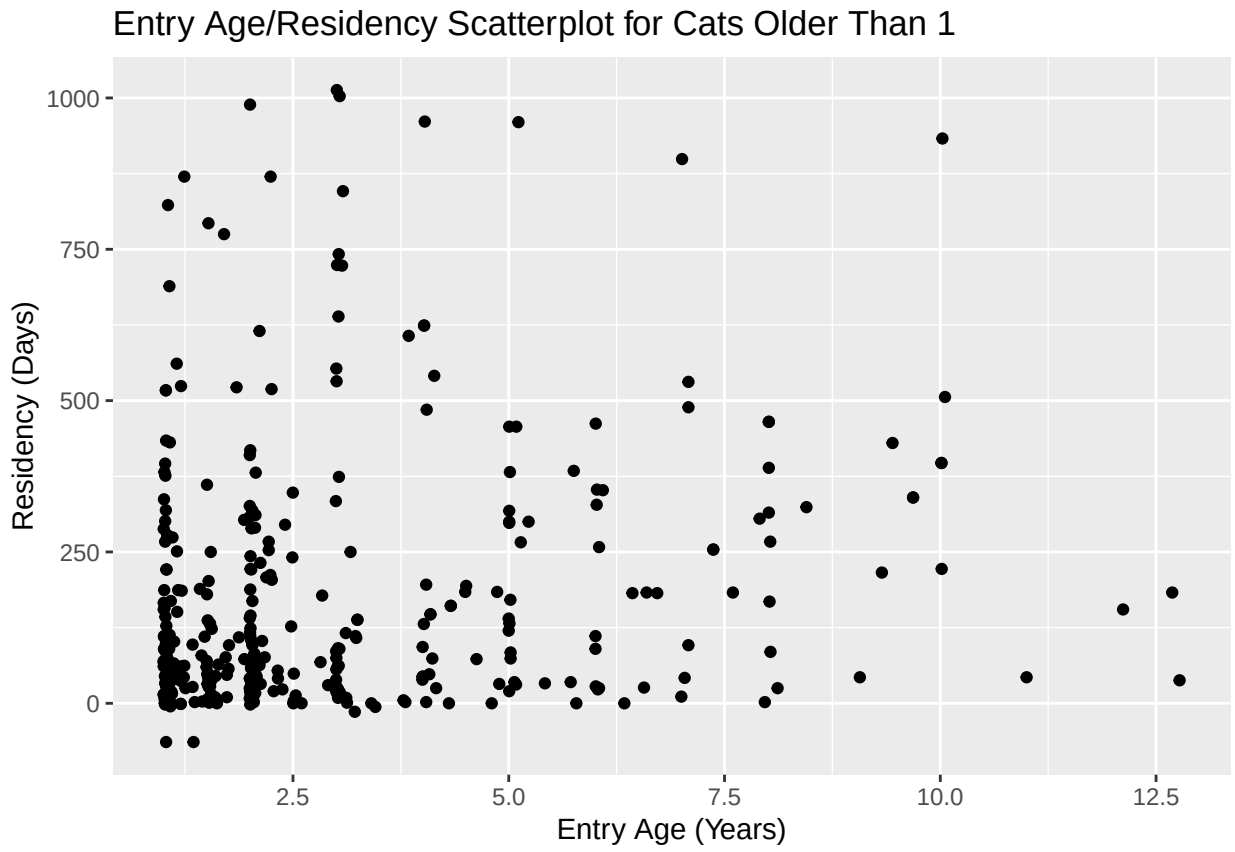
```
## [1] 0.1057277
```

```
length((adoptability_report_14on_reg %>% filter(`Entry Age Y` > 1))[,1])
```

```
## [1] 352
```

The base Pearson correlation value of around 0.106 should not be terribly surprising. Most of the data appears situated in the lower left corner with a decent amount of data above and to the right of that cluster. A negatively sloped line could be imagined from the upper left to the lower right, but it would be too far away from the kitten-cloud in the lower left corner. There are 352 cats who entered the rescue and were older than exactly one year. So it may be smart to check the correlation for those cats separately.

```
older_than_ones <- adoptability_report_14on_reg %>% filter(`Entry Age Y` > 1)
ggplot(older_than_ones, aes(x=`Entry Age Y`, y=Residency)) +
  geom_point() + labs(x="Entry Age (Years)",
                      y="Residency (Days)",
                      title="Entry Age/Residency Scatterplot for Cats Older Than 1")
```




```
cor(older_than_ones$`Entry Age Y`, older_than_ones$Residency)
```

```
## [1] 0.1812658
```

The correlation value only increases to 0.181 when omitting the kittens, so this weak positive correlation is unlikely to be meaningful.

Foster Cat Names and Adoptability

The last thing to check as a possible factor would be the names of the cats within the rescue. Cats in Calvin's Paws have foster names to be used for the adoption process. A lot of times, the names of these cats could be based on themes (with litters of kittens) and/or interests. A potential adopter could be more interested

in a cat named after a Star Wars or Game of Thrones character if the adopter is also a fan. Fandom and theme naming would be harder to quantify, but Calvin's Paws does list their cats available for adoption in alphabetical order. We can then check how the first letter of a cat's name impacts their residency.

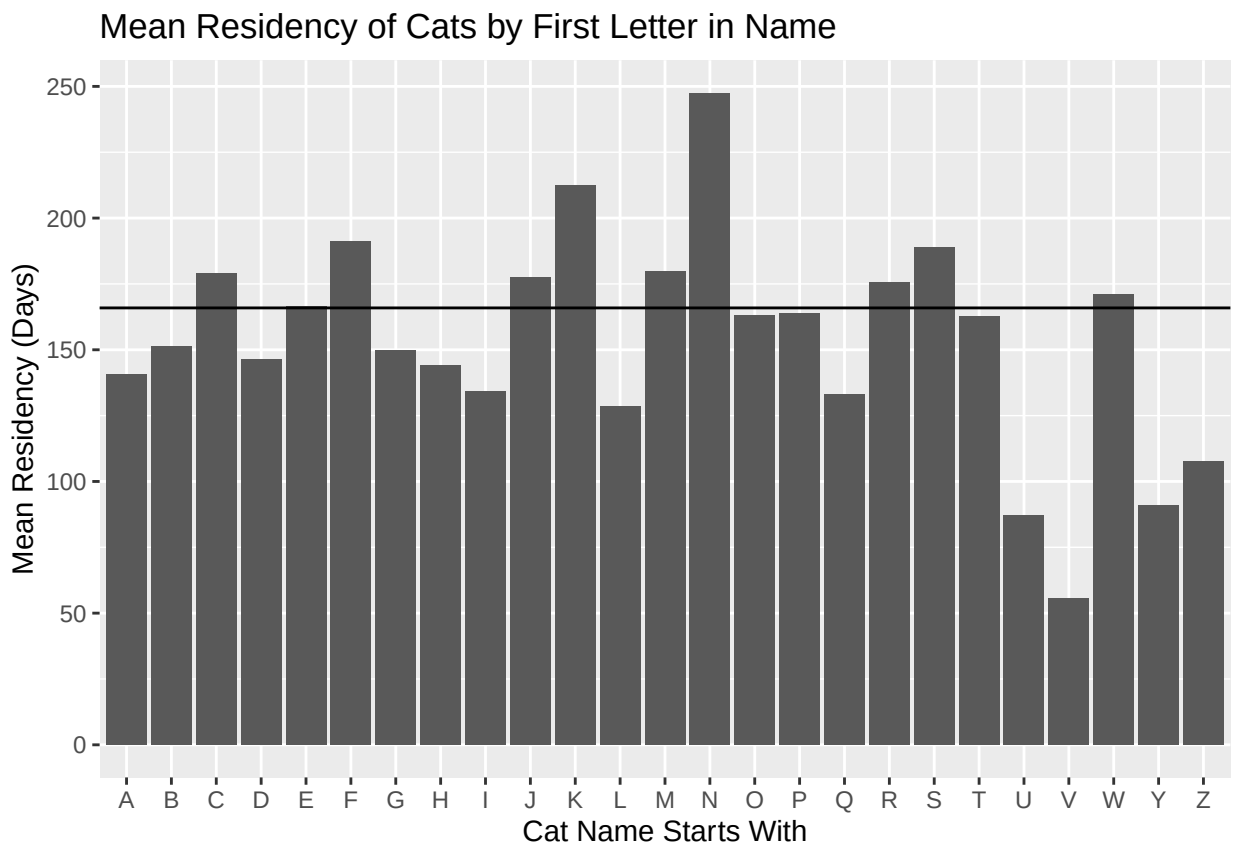
We will want to do this with some of the different datasets that have been created for various interests as well.

```
adoptability_report_14on_reg$`Name Start` <- substr(adoptability_report_14on_reg$Name, 1, 1)
kittenhaulers$`Name Start` <- substr(kittenhaulers$Name, 1, 1)
older_than_ones$`Name Start` <- substr(older_than_ones$Name, 1, 1)
cost_report_14on$`Name Start` <- substr(cost_report_14on$Animal, 1, 1)
domestics$`Name Start` <- substr(domestics$Name, 1, 1)
# Dataset showing values of median, mean, and standard deviation of Residency
# for the cats sorted by first letter of their name
letters <- adoptability_report_14on_reg %>% group_by(`Name Start`) %>%
  summarize(MedianResidency=median(Residency),
            MeanResidency=mean(Residency),
            SDResidency=sd(Residency))
letters <- cbind(letters, (adoptability_report_14on_reg %>% count(`Name Start`))[,2])
colnames(letters) <- c("Starts With", "Median Residency",
                       "Mean Residency", "SD of Residency", "Count for Letter")
letters
```

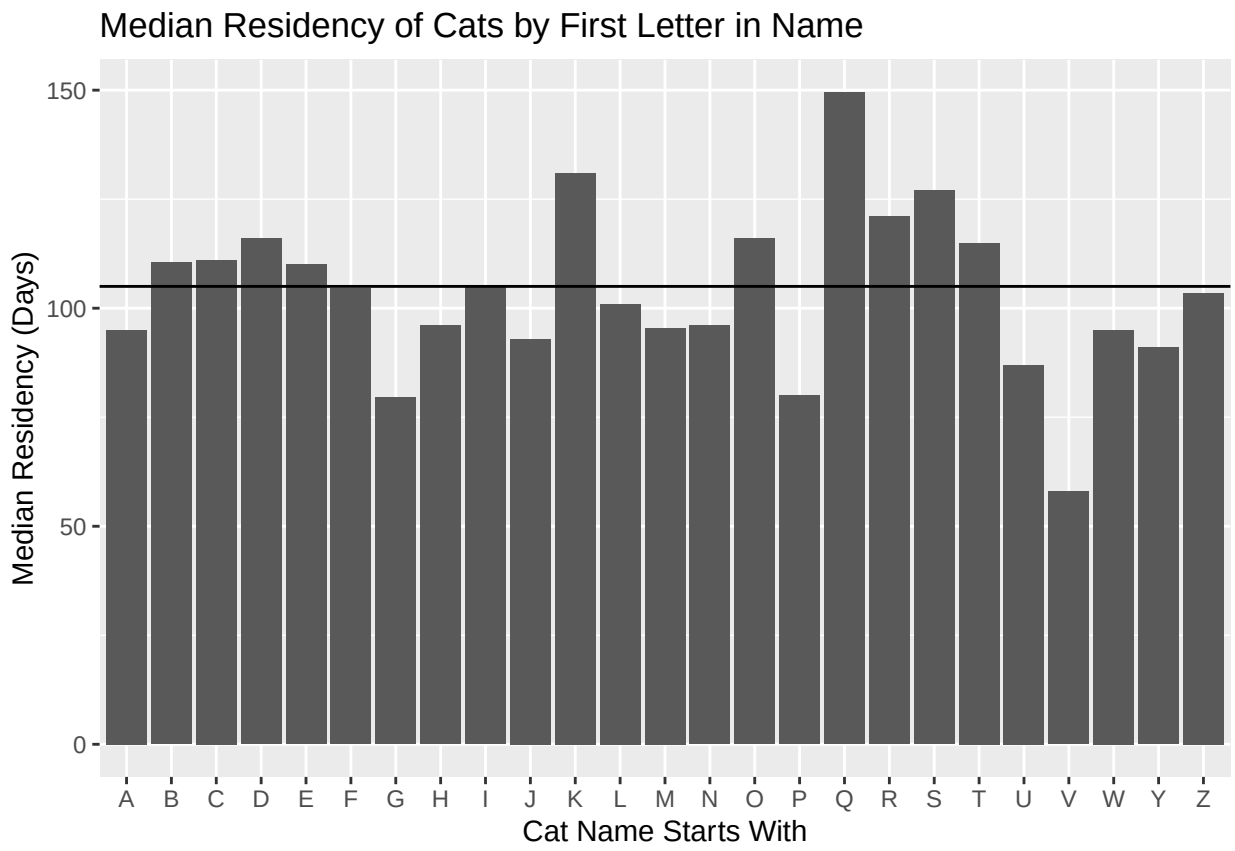
```
##    Starts With Median Residency Mean Residency SD of Residency Count for Letter
## 1            A             95.0      140.68966       160.93255               87
## 2            B            110.5      151.17544       148.05741              114
## 3            C            111.0      179.14483       198.95215              145
## 4            D            116.0      146.18310       141.73449               71
## 5            E            110.0      166.44444       180.71674               36
## 6            F            104.5      191.28947       225.86649               38
## 7            G             79.5      149.84091       188.84902               44
## 8            H             96.0      144.19048       164.97779               42
## 9            I            105.0      134.33333        96.58283               12
## 10           J             93.0      177.47170       213.82601               53
## 11           K            131.0      212.56757       249.89671               37
## 12           L            101.0      128.62366       113.69007               93
## 13           M             95.5      179.90278       215.05928              144
## 14           N             96.0      247.54286       301.59901               35
## 15           O            116.0      163.08696       146.59062               23
## 16           P             80.0      163.65116       225.56980               86
## 17           Q            149.5      133.25000        47.06290                4
## 18           R            121.0      175.60345       173.17225               58
## 19           S            127.0      188.79845       188.70572              129
## 20           T            115.0      162.58427       166.22950               89
## 21           U             87.0       87.00000              NA                1
## 22           V             58.0       55.53333        33.10776               15
## 23           W             95.0      171.13043       166.95326               23
## 24           Y             91.0       91.00000        35.35534                2
## 25           Z            103.5      107.50000        68.43558                8
```

```
# Graphs of the letters data by mean, median, and standard deviation
ggplot(adoptability_report_14on_reg, (aes(x=`Name Start`, y=Residency))) +
  stat_summary(fun=mean, geom="bar", position="dodge") +
  geom_hline(yintercept=mean(adoptability_report_14on_reg$Residency)) +
```

```
  labs(x="Cat Name Starts With",
       y="Mean Residency (Days)",
       title="Mean Residency of Cats by First Letter in Name")
```

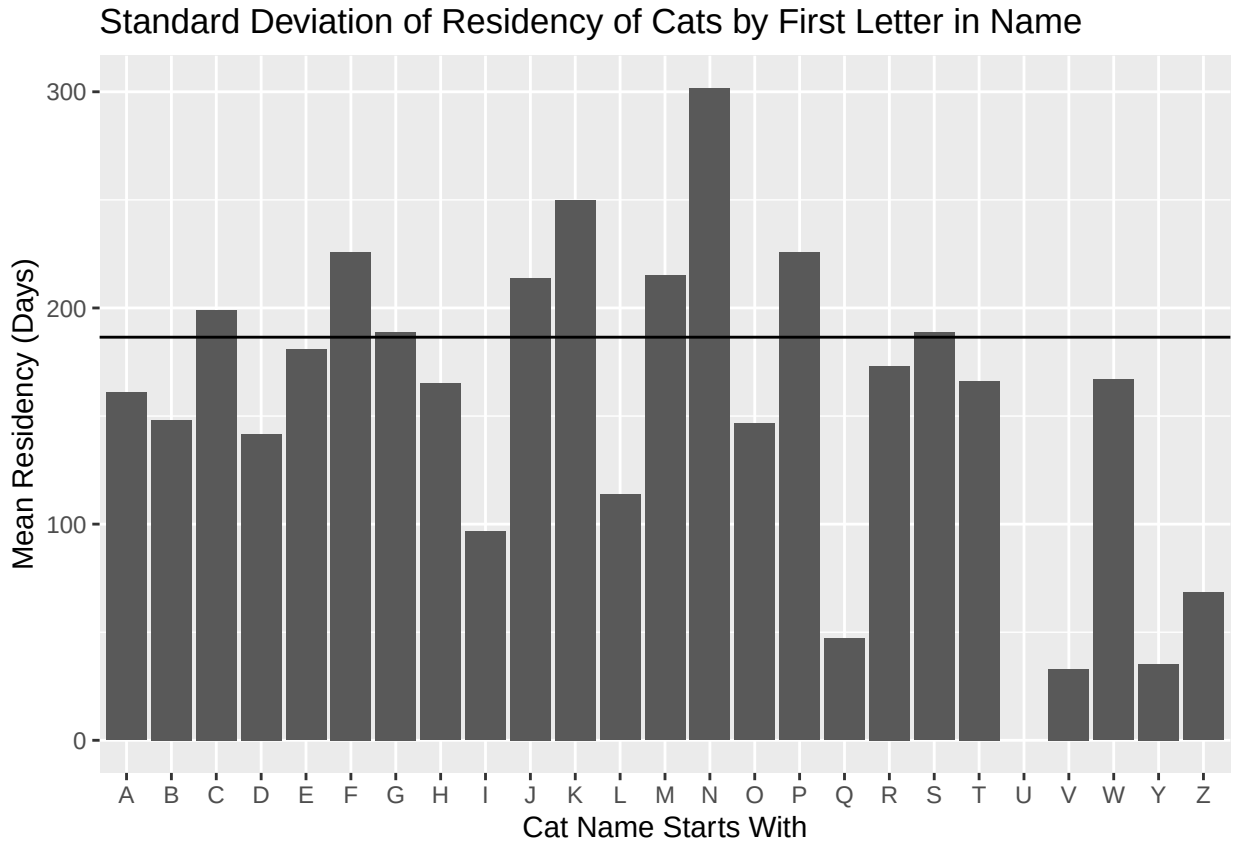



```
ggplot(adoptability_report_14on_reg, (aes(x=`Name Start`, y=Residency))) +
  stat_summary(fun=median, geom="bar", position="dodge") +
  geom_hline(yintercept=median(adoptability_report_14on_reg$Residency)) +
  labs(x="Cat Name Starts With",
       y="Median Residency (Days)",
       title="Median Residency of Cats by First Letter in Name")
```




```
ggplot(adoptability_report_14on_reg, (aes(x=`Name Start`, y=Residency))) +
  stat_summary(fun=sd, geom="bar", position="dodge") +
  geom_hline(yintercept=sd(adoptability_report_14on_reg$Residency)) +
  labs(x="Cat Name Starts With",
       y="Mean Residency (Days)",
       title="Standard Deviation of Residency of Cats by First Letter in Name")
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



From the printed table and graphs, we can see what the median, mean, and standard deviation of residency time grouped by the first letter of the cat's name. The table also contains the number of cats with each letter for this analysis. Each graph also has a horizontal bar that represents the graph's pictured statistic (mean, median, or standard deviation) for the entire dataset for easy comparison. For anecdotal comparison, cats whose names start with A tend to have residencies shorter than the population as a whole according while those whose names start with S have residencies have a higher mean and median than the population. Further statistical testing would be needed in order to provide additional information.

Cost Report Analysis

The Cost Report contains cost information for just about cat that has gone through Calvin's Paws. It does contain some sensitive information that should be scrubbed as well. Some of the variables in there may seem redundant as well, but they do have their purposes. For example, there are fields for Adopted.Date and Date. For these fields, the Adopted.Date field is a cat's first adoption and the Date field is the cat's most recent adoption. (In some instances a cat may have been returned and then was later adopted again) The ContactID field is a unique and anonymous identifier for the adopter.

One thing to also note from earlier is that we do know that some cats appear more than once in the cost report. The cost listed in the Cost field is the total money spent on the cat throughout their tenure with Calvin's Paws. This happened due to return and re-adoption of the repeating cats. When this happens, the adoption later than the first has a higher residency (Length.of.Stay in the cost report). The highest residency appearing in the cost report is the value that appears in the adoption report. So for consistency, duplicates would need to be removed and the entry to keep will be the entry with the latest Date (the Adoption.Date for each duplicate will also reflect the most recent adoption).

```
cost_report_14on <- cost_report_14on %>% select(-Adopter)
duplicates <- cost_report_14on %>% count(Rescue.ID) %>% filter(n > 1)
cost_report_14on <- cost_report_14on %>% filter(Adopted.Date == Date)
stat.desc(cost_report_14on$Cost)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.316000e+03 5.800000e+01 0.000000e+00 0.000000e+00 4.627900e+03 4.627900e+03
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 1.883085e+05 1.157500e+02 1.430915e+02 5.364367e+00 1.052365e+01 3.786979e+04
##      std.dev     coef.var
## 1.946016e+02 1.359980e+00
```

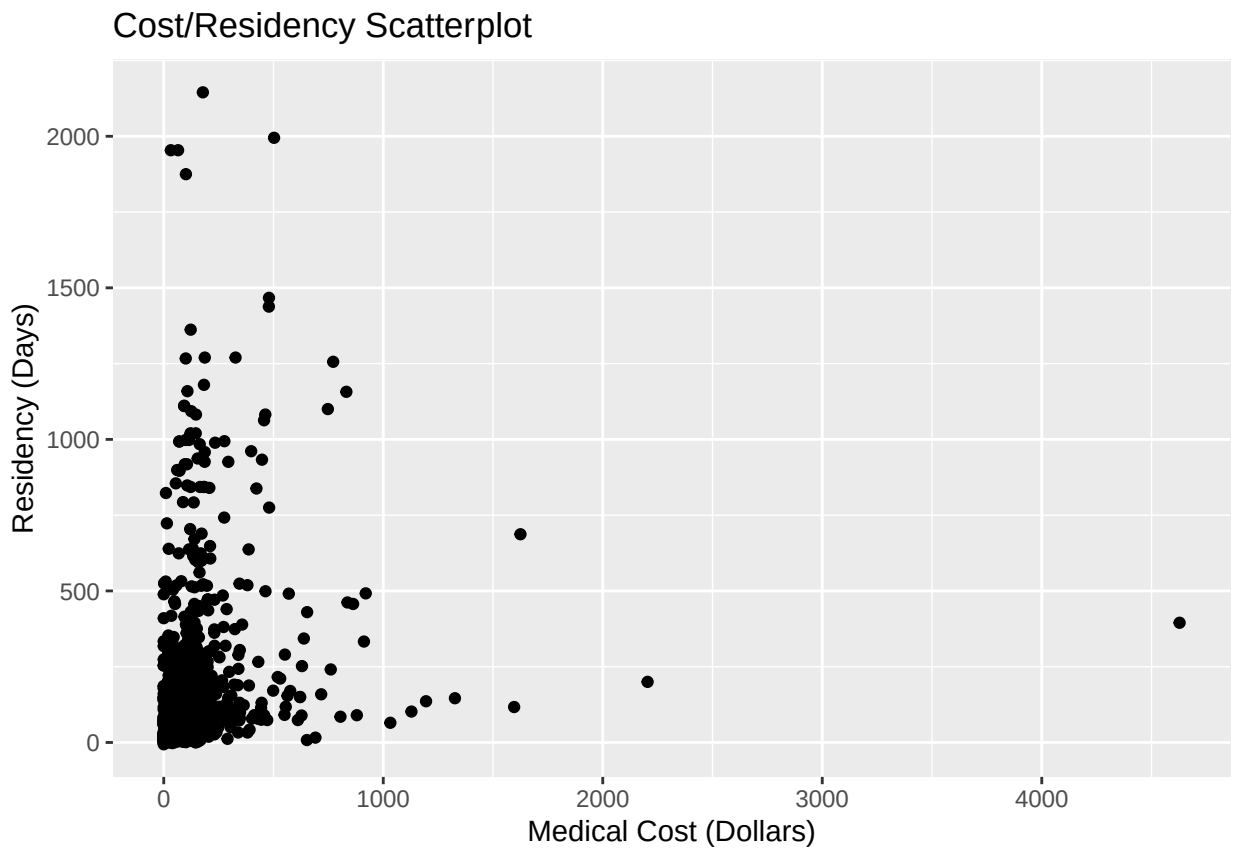
```
summary(cost_report_14on$Cost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   89.77  115.75  143.09  142.93 4627.90
```

After the duplicate entries are filtered out, we see the median cost for a cat's medical care comes to $115.75, the mean cost is $143.09, and the standard deviation of the cost is $194.60.

With the duplicate entries removed and a value for Length.of.Stay decided on, we can make a scatterplot to see if there is any kind of correlation between the time a cat spends in the rescue and the amount of money Calvin's Paws spent on the cat.

```
ggplot(cost_report_14on, aes(x=Cost, y=Length.of.Stay)) +
  geom_point() + labs(x="Medical Cost (Dollars)",
                      y="Residency (Days)",
                      title="Cost/Residency Scatterplot")
```



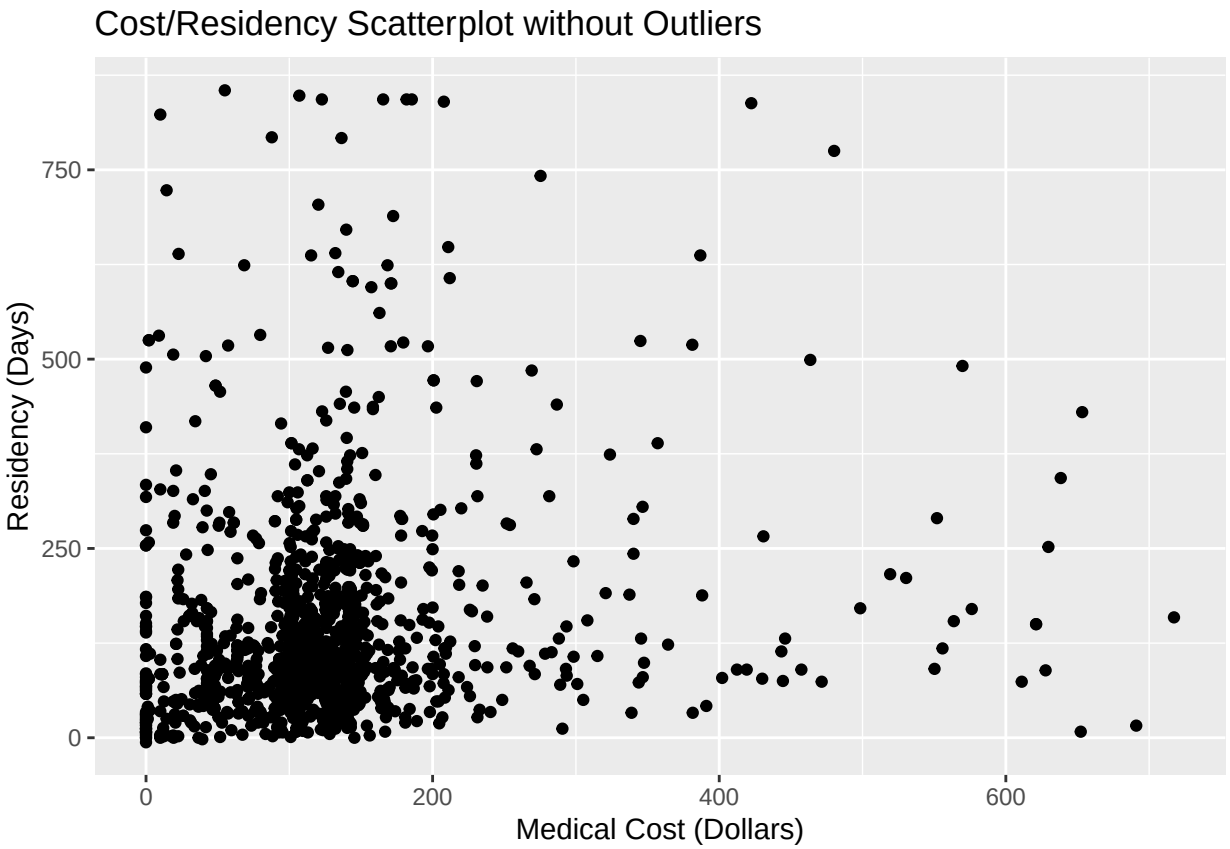

```
cor(cost_report_14on$Cost, cost_report_14on$Length.of.Stay)
```

```
## [1] 0.1597802
```

The scatterplot has a cloud of points in the lower left, indicating that most of the cats included here end up with low vet costs and low residencies. The points outside the cloud tend to be spread out randomly with some points further to the right with higher expenses, and some points further up with higher residencies. If we want to try and exclude outliers and get a closer picture, that could be done for residencies and costs.

```
# Length of Stay Z-Score
cost_report_14on$Length_z <- (cost_report_14on$Length.of.Stay -
                                mean(cost_report_14on$Length.of.Stay)) /
  sd(cost_report_14on$Length.of.Stay)
# Cost Z-Score
cost_report_14on$Cost_z <- (cost_report_14on$Cost - mean(cost_report_14on$Cost)) /
  sd(cost_report_14on$Cost)
cost_report_14on_reg <- cost_report_14on %>% filter(
  abs(Length_z) < 3 & abs(Cost_z) < 3)
cost_outliers_cost <- cost_report_14on %>% filter(abs(Cost_z) >= 3)
cost_outliers_length <- cost_report_14on %>% filter(abs(Length_z) >= 3)

ggplot(cost_report_14on_reg, aes(x=Cost, y=Length.of.Stay)) +
  geom_point() + labs(x="Medical Cost (Dollars)",
                      y="Residency (Days)",
                      title="Cost/Residency Scatterplot without Outliers")
```




```
cor(cost_report_14on_reg$Cost, cost_report_14on_reg$Length.of.Stay)
```

```
## [1] 0.1622052
```

Removing the cost and residency outliers from the scatterplot essentially allowed us to zoom in on the cloud of points in the lower part of the graph. The Pearson correlation value for this second plot only increased by a measly .003. The small positive correlation between Cost and Length of Stay/Residency does not seem to indicate a relationship between the variables.

Denied Applications Analysis

In Part 2 of this analysis, a small analysis almost started for the Denied Applications Report. A set of string searched were conducted in the Comments field of the report. There were 382 total denied applications for 265 different cats. From the searches, it was found that:

- 74 denial comments mentioned declawing, 19.3% of denied applications
- 48 denial comments mentioned wanting to adopt a single kitten as an only pet, 12.6 % of denied applications
- 17 denial comments mentioned allowing the cat to go outdoors, 4.5% of denied applications.
- 13 denial comments mentioned that the applicant lived outside the Calvin's Paws area, 3.4% of denied applications
- 2 denial comments mentioned that the cat being applied for already had an approved application, 0.5% of denied applications
- 9 denial comments mentioned the applicant being under 21, 2.4% of denied applications

- 16 denial comments mentioned that approving the adoption would break lease agreements for condos or apartments or otherwise get the applicant in trouble with their landlord, 4.2% of denied applications
- 9 denial comments mentioned the applicant did not want to vaccinate their cat, 2.4$ of denied applications

These string searches were also not mutually exclusive of each other and are not an exhaustive search of the comments across all 382 denied applications. A function used for total textual analysis would likely be needed for other conclusions to be drawn.

Outliers and Duplicates Check

Part of this analysis should include checking in on the outliers and duplicates from the cost report as well as the cats who received denied applications. The duplicates dataframe would be a good place to place our checks for easy analysis since it only has two fields at this time.

```
duplicates$`Adoption Outlier` <- duplicates$Rescue.ID %in% adoption_outliers$Rescue.ID
duplicates$`Cost Outlier Cost` <- duplicates$Rescue.ID %in% cost_outliers_cost$Rescue.ID
duplicates$`Cost Outlier Length` <- duplicates$Rescue.ID %in% cost_outliers_length$Rescue.ID
duplicates$`Kittenhauler` <- duplicates$Rescue.ID %in% kittenhaulers$Rescue.ID
duplicates$`Denied Application` <- duplicates$Rescue.ID %in% denied_applications_14on$Rescue.ID
colnames(duplicates) <- c("Rescue.ID", "Adoptions", "Adoption Outlier", "Cost Outlier Cost",
                          "Cost Outlier Length", "Kittenhauler", "Denied Application")
sum(duplicates$`Adoption Outlier`)
```

```
## [1] 14
```

```
sum(duplicates$`Cost Outlier Cost`)
```

```
## [1] 5
```

```
sum(duplicates$`Cost Outlier Length`)
```

```
## [1] 21
```

```
sum(duplicates$Kittenhauler)
```

```
## [1] 22
```

```
sum(duplicates$`Denied Application`)
```

```
## [1] 21
```

With the Duplicates dataframe now holding a number of logical checks for if they appear as outliers or in other specific datasets, we can check those values. Of the cats adopted multiple times,

- 14 had residencies in the adoption report considered outliers, 19.2% of the cats with multiple adoptions
- 5 had costs in the cost report considered outliers, 6.8% of the cats with multiple adoptions
- 21 of them had residencies/lengths of stay in the cost report considered outliers, 28.8% of the cats with multiple adoptions

- 22 of them were Kittenhaulers (kittens with long stays in the rescue), 30.1% of the cats with multiple adoptions
- 21 of them had denied applications, 28.8% of the cats with multiple adoptions

It is worth noting that the 22 kittenhaulers that appear in the Duplicates dataframe are also 27% of the Kittenhauler dataframe.

The Narrative Following the Analsyis

The analysis covered here took almost a decade's worth of data from Calvin's Paws in order to look through data to help identify factors that make adoptions happen faster. A second goal was to identify factors that may negatively affect a cat's adoption prospects. Outside the qualities of the cats themselves, this analysis also looked into the various reasons why an adoption application would be declined and if there were any correlations present between medical expenses and a cat's adoptablity.

Methods

To accomplish this, data from 2014-2023 that contained each cat's breed, coloration, birthday, adoption day, and the date the cat entered Calvin's Paws was worked with. New fields were derived from the existing data, outliers were identified to see how they impacted the summary statistics used on the data, and plots were made to visualize and interpret the data.

Analysis and Insights

Insights from this analysis included things that were not shocking, like the fact that if a cat is listed as something other than a domestic breed, it will be adopted faster and that there is no real preference towards male or female cats at large. For cats that are listed as domestics, three different colorations performed significantly different from the others. Solid gray cats and solid white cats tended to be adopted at rates faster than the median and mean times cats spent with Calvin's Paws. Orange cats also performed better than the median but their mean was closer to the other colorations.

On the other hand, tortoiseshell/tortie cats tended to stay in the rescue for longer periods of time. The median, mean, and standard deviation of residency were all markedly higher than their peers. Torties have been stereotyped as having attitude to the point that a portmanteau was coined for it, tortitude. Their tortitude isn't always present but can manifest itself as tending to be independent.

Through the analysis, the median usually felt to be more representative of the data than the mean due to the presence of cats who happened to stay in the rescue longer.

No significant correlation was found between the age at which a cat entered the Calvin's Paws program and how long they stayed in the rescue. A couple of factors likely had to do with this. First, adopters really do like kittens. This was seen in scatterplots where an absolute cloud of points were clustered in the lower left corner of the graph. These were the cats who were either born into the rescue and adopted while they were kittens. This put significant weight of any correlation to take them into account.

A scatterplot that was initially drawn up but eventually not included in this analysis was that of an age at adoption versus residency. It was not included beacuse it was essentially a shear of the entry age vs residency plot. The residencies didn't change, so the only thing that did change was the addition of time to each point based on the residency of the cat. So each point was shifted to the right by an amount equal to its y-value/residency. For that reason, the plot was excluded.

Outside the kitten cloud, points were spread out but there weren't enough to bring weight to a correlation analysis. After the removal of the kitten cloud from the scatterplot and viewing it, I would have predicted a negative correlation. This would go through the points in the upper left corner (kittens who stayed in

the rescue for a long time - kittenhaulers) through the middle to the lower right corner (cats that entered the program at older ages and adopted faster). Older cats would have been adopting better than younger ones, something that would have defied conventional wisdom. This did not come to fruition though, since the Pearson correlation constant for that scatterplot was still low.

A second scatterplot was drawn up to try and test this theory. A second data subset was derived that included cats that entered the rescue at later than one year old. The correlation between age and residency did increase from 0.106 to 0.181, a difference of more than 50%. Even still though, a Pearson correlation score of that magnitude indicates that the positive correlation is weak at best. So it appears that the age of a cat (outside of kittens) entering the rescue should not necessarily impact how long it will stay in the rescue.

Additional analysis was given for the names given to the rescue's cats. Since Calvin's Paws lists their cats in alphabetical order, cats with names higher up the alphabet would theoretically be noticed sooner on the Calvin's Paws website. To check this, the adoptability data was grouped by the first letter of the cats' names and median, mean, and the standard deviation were found for the groups. Plots were made as well, each with a horizontal line showing the value for the entire population as a whole for comparison. Residencies for the statistics tended to vary across the whole alphabet without following a pattern. Early letters like a, b, and c did not appear to perform significantly better than w, y, or z. X is the only letter that did not appear and u had only one cat (and thus did not have a standard deviation for that reason). It likely cannot be concluded that early letters perform better than later ones. Other testing would likely be needed to investigate which letters are "best" to work with.

No correlation was found between total medical cost for cats and their adoption times as well.

It was found that in the Cost Report, multiple lines of data could appear for a single cat if they were returned and re-adopted. The way this was handled by the rescue, adding a date column to differentiate between adoptions, worked okay for the Cost Report but it caused any summary statistics to be inflated by those cats. The dataset needed to be filtered so the Adopted.Date and Date fields matched to remedy the issue. More on the ramifications of this in the limitations section.

Limitations

A cost of the way returns and adoptions were handled by Calvin's Paws is that the length of time a cat spent with the rescue dramatically increased if they were returned. A better way to handle it could have involved additional fields to specify return dates and other adoption dates to signify that it was not the first adoption. The approach taken in this analysis is akin to seeing the returns and re-adoptions were part of a much longer journey, for some cats, with the rescue before being adopted for good by the cat's current owner. That is not necessarily in the spirit or intention of the analysis, but it is what could be done with the data available to me.

Analysis of factors outside of breed/coloration were not present to help identify other factors that could negatively impact a cat's adoptability, so not much could be done to explore that space. The area that could be explored and identified was looked into though.

A fully functional text analysis function would have been something that would have made analysis of the comments in the Denied Applications Report better. The way the comments in the report were structured was that each value for the denied application's comments were all comments made by rescue personnel during the process of the application until it was denied. The current analysis searched for selected buzzwords instead of doing a brute-force compilation and tally of each word used in the comments for an application. If a bulk text analyzer were used for this analysis, a graphical approach likely could have been taken

Concluding Remarks

The major takeaways from this analysis come from noting that non-domestic cats tend to adopt more quickly and certain colorations perform better or worse than others. However, these conclusions are unlikely to be

revelatory to those who have worked in rescue for long periods of time. Rescue veterans likely had an intuitive sense of these things already but confirmation through data is helpful nevertheless. I should note though that the mishandling of this information to intentionally mislabel the breeds of cats or their colorations should not be done in order to speed up adoption rates for rescues. It should however be heartening that there is no significant or apparent correlation between the age of a cat when it enters the rescue and how long they will stay within the rescue. Older cats have good chances of reasonably timed adoptions just like kittens do.

The best thing for a rescue like Calvin's Paws to do is make sure they can love and care for the cats taken into their care as best as they can to provide the best life possible for their cats. The clients of the rescue, the cats, are making their way through life and deserve to have the best experience possible and go to a loving family that will care for them.