Furry Futures: Predictive Analytics for Adoption Outcomes in No-Kill Shelters

Tess Anderson

tanderson4@bellarmine.edu

February 2, 2024

1. Introduction

In this report, I explore data, and their patterns and relationships, on outcomes of cats in an animal shelter. The dataset comes from kaggle.com and includes data on over 28,000 cats at the Austin Animal Shelter in Austin, Texas. This dataset was chosen because of the potential for significant improvements in animal outcomes and animal shelter practices with the help of insights generated by data science. Animal shelters have limited resources and caring for animals, especially sick and vulnerable ones which most often end up in shelters, is costly but important. With insights into adoption trends and the ability to make informed decisions based on data, animal shelters can use their resources efficiently to find forever homes for the highest number of animals. This data, specifically, was collected by the Austin Animal Shelter and is a good resource for this analysis as it is fairly complete and quite thorough, containing 37 columns with relatively few null or missing values. Throughout this report, I explain the columns that exist in the dataset, the data that is missing, and explore the relationships between these columns using visualizations.

2. Dataset Description

To begin to understand and familiarize both the reader and researcher with this vast dataset, we start by summarizing the data it contains. First, the most important feature of this dataset is the Outcome Type column - the aim of future analysis is to use the other variables in the dataset to predict the Outcome Type. The types included in this data point are: adoption, transfer, euthanasia, return to owner, died, rto-adopt (also return to owner), missing, and disposal. There are 36 other columns which contain information on other features of each cat in the dataset. These columns are listed in the table below, and specific definitions and details are included to aid in understanding the structure and contents of this dataset. In addition to Outcome Type, a column called Outcome Subtype contains further information on the outcome where it is necessary or relevant – common subtypes in the dataset include partner, foster, SCRP, and suffering. There are several columns that contain information on the age of the cat in various formats – this allows the researcher to conduct various types of analysis and explore these data from several different approaches. Similarly, there are several variables which contain the same information on breed and color. Finally, there are many columns which deal with time or date – the time and date of the outcome, the date of birth, and the length of time a cat spent in the animal shelter.

Name	Definition	D Type	Range of Values	% Null
Age Upon Outcome	Cat's age approximated with numeric value and time interval such as weeks, months, and years.	Ordinal	1 day to 22 years	0
Animal ID	Unique ID number for cat.	Nominal	N/A	0
Animal Type	Type of animal – this dataset contains only cats.	Nominal	Cat	0
Breed	Breed of cat, most classified as domestic	Nominal	65 breeds and combinations of breeds	0
Color	Color and pattern of coat	Nominal	154 colors and patterns combined	12.33
Date of Birth	Cat's DOB, often only month and year.	Interval	06/25/1992 – 12/25/2017	0
Datetime	Date and time of outcome	Interval	20/01/2013 – 02/01/2018	0
Monthyear	Month and year of outcome	Interval	20/01/2013 – 02/01/2018	0
Name	Name of cat	Nominal	7409 distinct values	43.4
Outcome Subtype	Subtype for outcome	Nominal	Partner, Foster, SCRP, Suffering, SNR, In Kennel, In Foster, Offsite, Rabies Risk, Medical, At Vet, Enroute, In Surgery, Aggressive, Barn, Possible Theft, Underage	0
Outcome Type	Type of outcome	Nominal	Transfer, Adoption, Euthanasia, Return to Owner, Died, RTO-Adopt, Missing, Disposal	0
Sex Upon Outcome	Denotes cat's gender and if it was spayed or neutered upon shelter outcome	Nominal	Spayed/Intact Female, Neutured/Intact Male, Unknown	0
Count	Count of the specific cat, every entry = 1	Interval	1 for every entry	0
Sex	Gender derived from sex upon outcome	Nominal	Female, Male	0
Spay/Neuter	Status of spay/neuter derived from sex upon outcome	Nominal	Yes, No	0
Periods	Numeric value from age upon outcome: if 5 days, period=5; if 2 years, period=2.	Interval	0 – 22	0
Period Range	Value corresponding to time interval from age upon outcome: if days, period range=1; week, period range=7, etc.	Ordinal	1 – 365	0
Outcome Age (Days)	Age in days generated using Period and Period Range	Ratio	0 – 8030	0
Outcome Age (Years)	Age in years generated using Period and Period Range	Ratio	0 – 22	0
Cat/Kitten (Outcome)	Categorical variable designating cat as adult or kitten (6 months or less) upon outcome.	Nominal	Kitten/Cat	0

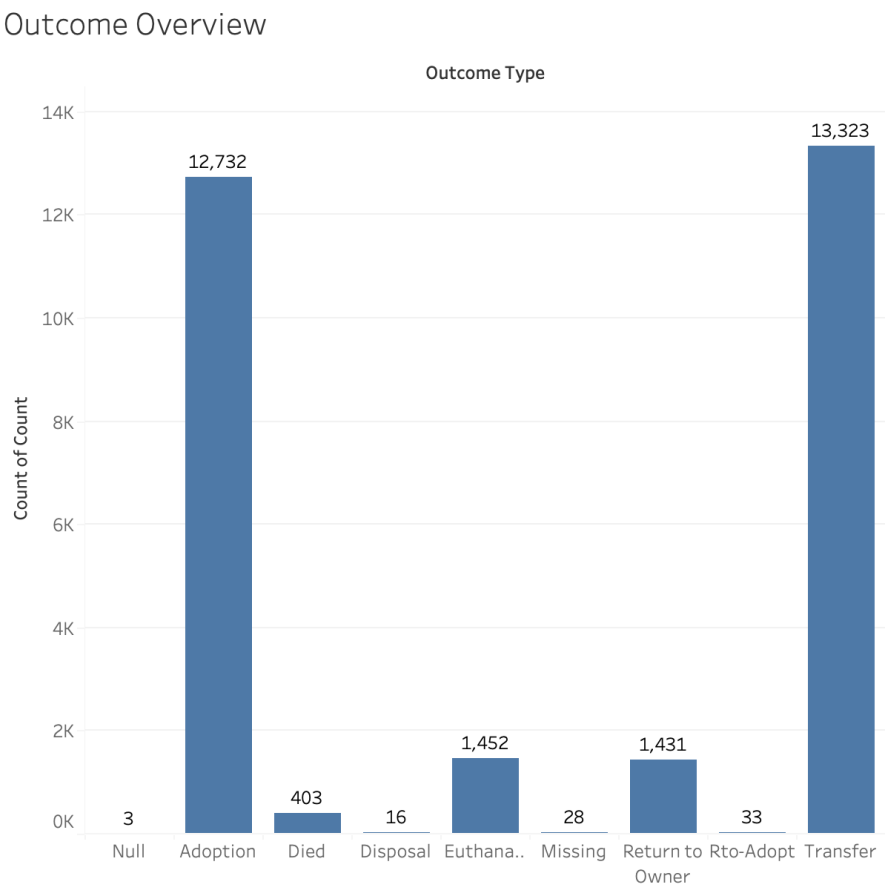
Sex Age Outcome	Describes neuter/spay status, sex, and kitten/adult status. ex: Neutered Male Kitten	Nominal	10 values Neutered/ spayed/intact, Male/ Female/Unknown, Kitten/ Cat	0
Age Group	Age of cat sorted into 10 bins by years	Interval	-0.022 – 2.2, 2.2 – 4.4, 4.4 – 6.6, 6.6 – 8.8, 8.8 – 11.0, 11.0 – 13.2, 12.2 – 15.4, 15.4 – 17.6, 17.6 – 19.8, 19.8 – 22.0	0
DOB Year	Year of birth extracted from DOB column	Interval	1992 – 2017	0
DOB Month	Month of birth extracted from DOB column	Interval	1 – 12	0
DOB Monthyear	Year and month of birth extracted from DOB column	Interval	10-2013 – 02/2018	0
Outcome Month	Year of outcome extracted from datetime column	Interval	1 – 12	0
Outcome Year	Month of outcome extracted from datetime column	Interval	2013 – 2018	0
Outcome Weekday	Day of week name of outcome	Nominal	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday	0
Outcome Hour	Hour of outcome extracted from datetime column	Interval	0 – 23	0
Breed1	First breed for records containing two breeds in Breed column	Nominal	41 breeds	0
Breed2	Second breed for records containing two breeds in Breed column	Nominal	14 breeds	99.8
CFA Breed	True if breed is one of 42 recognized by Cat Fancier's Association	Nominal	True, False	0
Domestic Breed	True if breed is not recognized by CFA	Nominal	True, False	0
Coat Pattern	Pattern split from "color" column	Nominal	Tabby, tortie, calico, point, torbie, smoke, agouti, brindle, tricolor	34.9
Color1	First color for records containing two colors in Color column	Nominal	40 colors	0
Color2	Second color for records containing two colors in Color column	Nominal	12 colors	64.8
Coat	Extracted colors and patterns combined	Nominal	44 coat categories	0

3. Dataset Summary Statistics

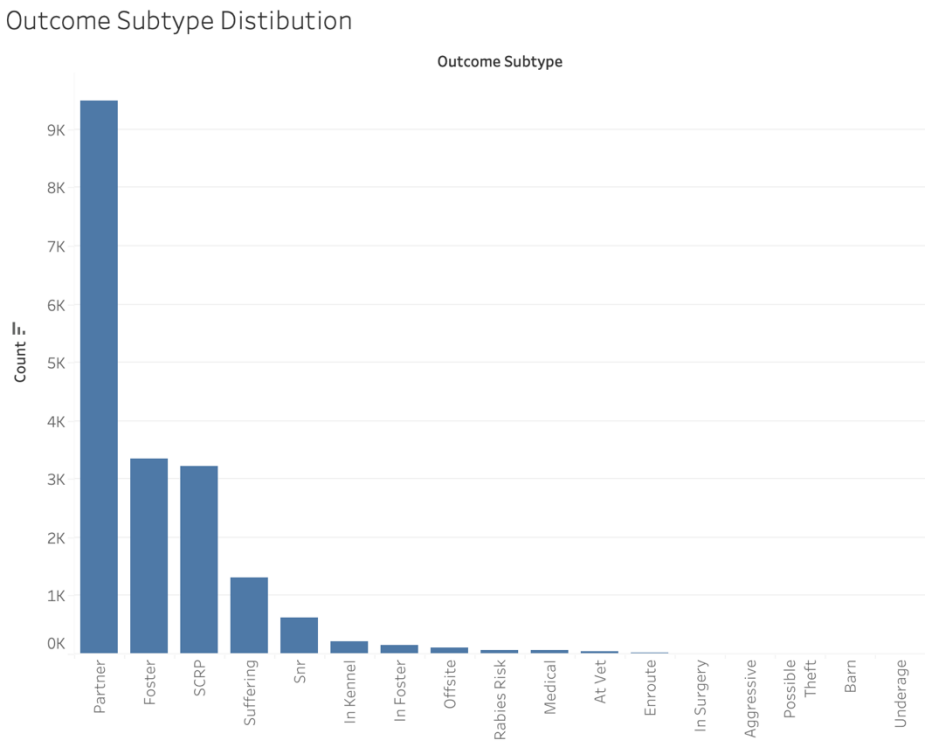
The numerical variables in this dataset are fewer and, in many cases, less important for analysis than the categorical varaibles. However, we will start with an overview of the statistics for these numerical variables as it is still important to gain an understanding of all the variables and how they are distributed throughout the dataset. The table below shows relevant summary statistics for the numerical variables in the dataset.

	Mean	St. Dev.	Min	25%	50%	75%	Max
Periods	3.2	2.6	0	2	2	4	22
Period Range	143.2	162.9	1	30	30	365	365
Outcome Age (Days)	509.45	966.8	0	60	90	365	8030
Outcome Age (Years)	1.4	2.65	0	0.16	0.25	1	22
DOB Year	2014	2.97	1992	2013	2015	2016	2017
DOB Month	6.19	2.82	1	4	6	8	12
Outcome Month	7.15	3.14	1	5	7	10	12
Outcome Year	2015.4	1.24	2013	2014	2015	2016	2018
Outcome Hour	14.01	3.86	0	12	14	17	23

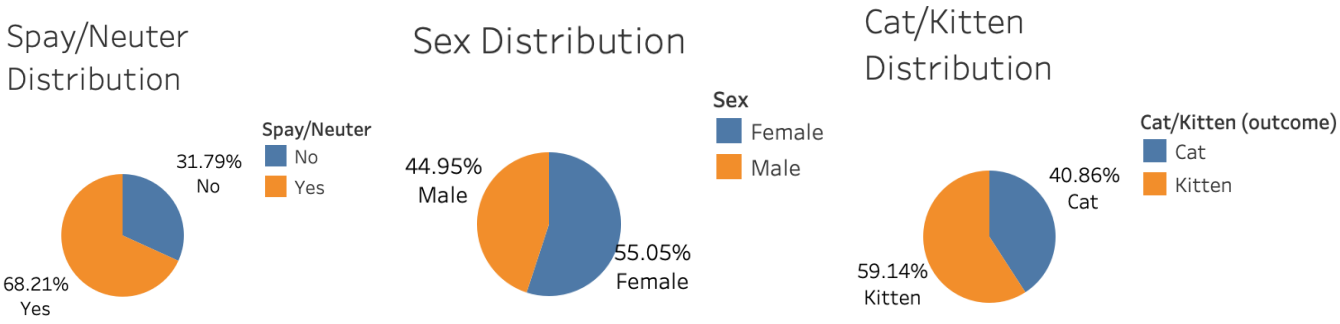
As is clear after understanding the definitions and distributions of these variables, many are not important for the future analysis – a large part of this is because many of the columns in this dataset, Periods and Period Range for example, were used simply to generate another column, so their insights independently are not helpful. These columns should be dropped further in this analysis process. To explore the other, more relevant variables, visualizations are generated as these are one of the best ways to understand distributions of categorical variables. First, an overview of the distribution of outcomes which serves as our dependent variable in the upcoming analysis.



Clearly, the most common outcomes are Adoption and Transfer, followed by Euthanasia and Return to Owner. Adoption indicates that the cat has been adopted to a, hopefully, forever home. Transfer indicates that the cat has been transferred to another facility, often to a vet to be neutered/spayed or for other medical treatment.

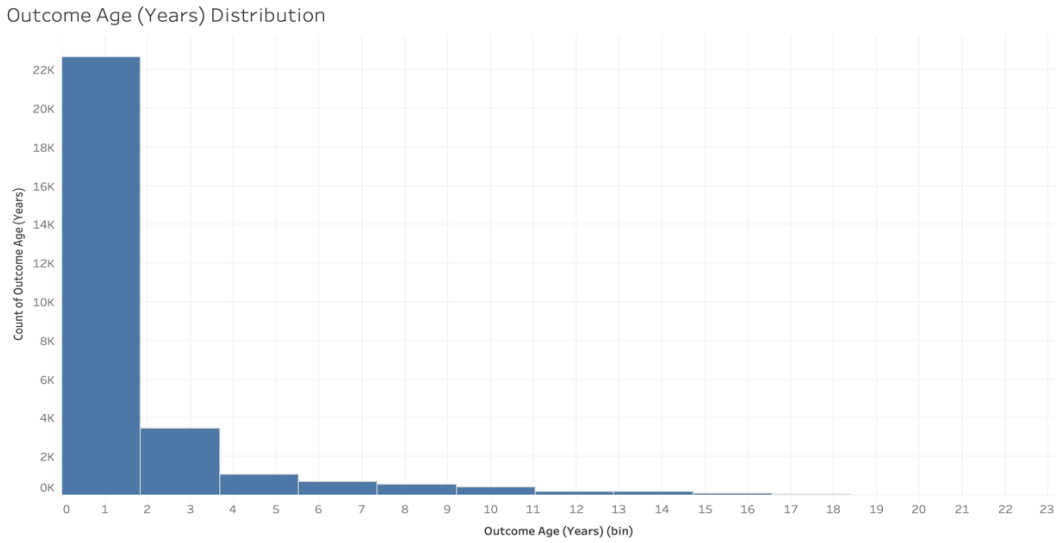


The outcome subtypes give more specific information into the outcome where it is necessary or relevant. All outcomes in Transfer category contain further information in the subtype column, as do many adoption and euthanasia entries. All return to owner, RTO-adopt, and disposal entries do not contain any subtype information.

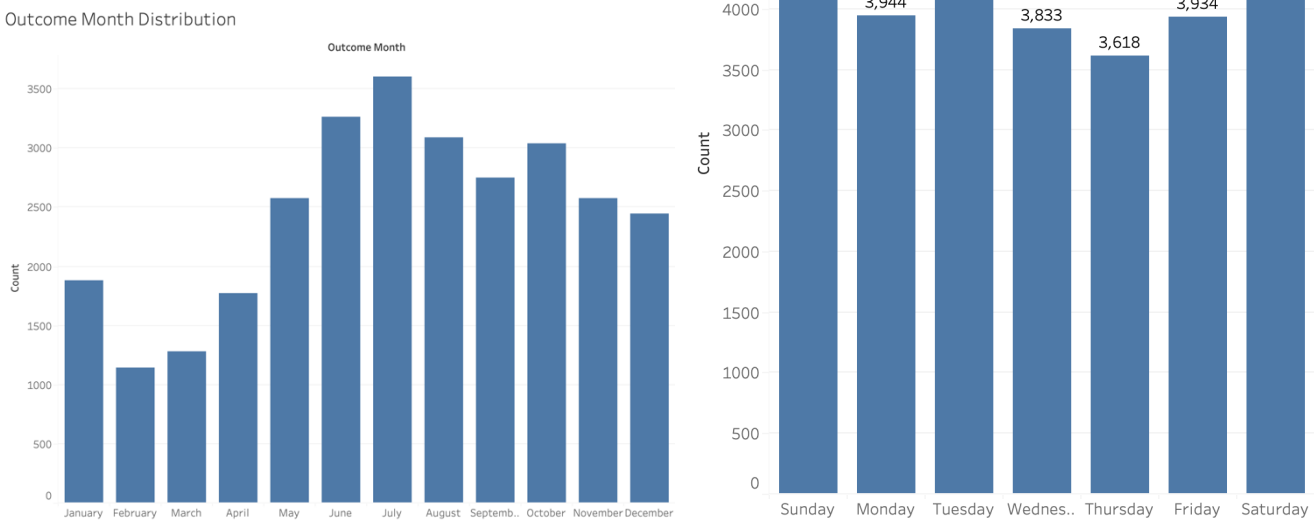


Shown above are proportion distributions as pie charts for three variables which each contain two categories: spay/neuter which is indicated by yes/no, sex, and cat/kitten. The majority of the cats at 68.2% are spayed or neutered by the time they reach their outcome. Most of the cats in the shelter are female, but only by a fairly slim margin. Addtionally, most of the cats are kittens at 59.1%.

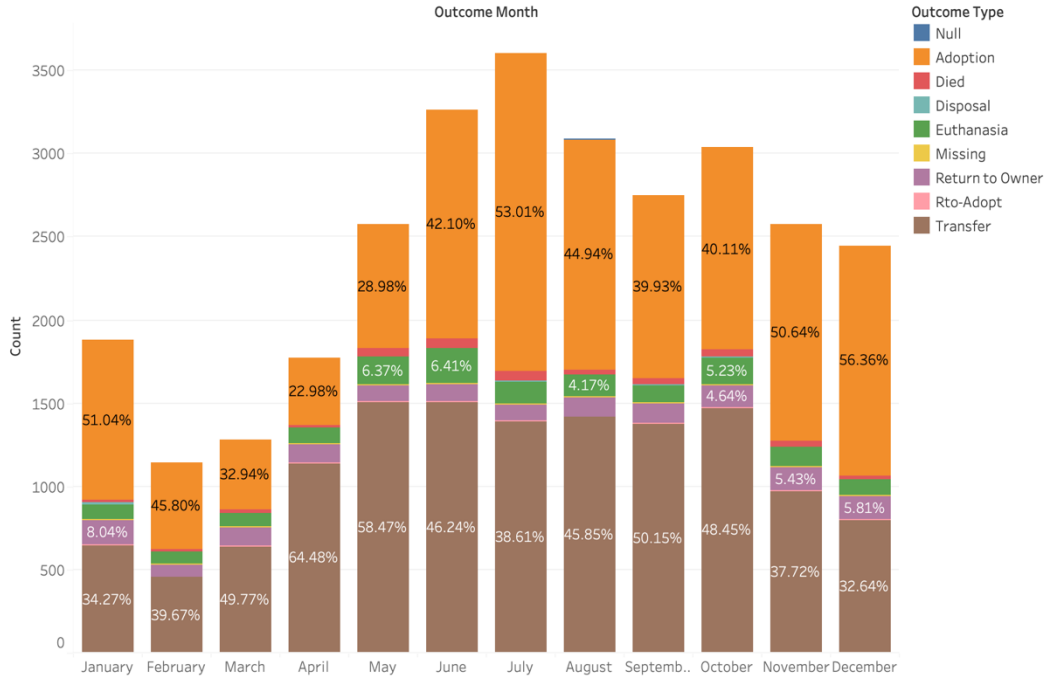
As shown on the following page and reflected in the cat/kitten distribution above, there are more young cats than old. The histogram on the following page shows a vast majority of cats belonging to the 0-2 year age group. There are four entries with extreme values of age at 22 years, but these are not considered outliers.



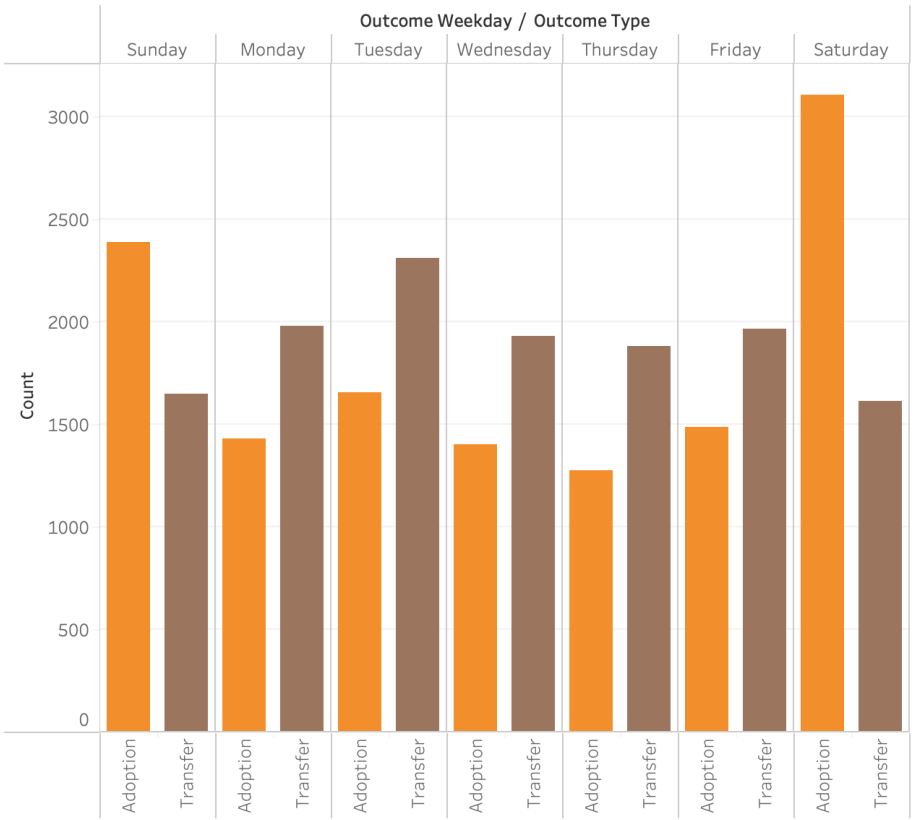
Next, distributions related to time of outcome were generated. On the right, weekday distribution is shown and Saturday is clearly the day which sees the highest frequency of outcomes while Wednesday, Thursday, and Friday are the lowest. On the left, the month distribution is shown and there are notable differences between June and July, and February and March. February is the lowest frequeny, by far, while July contains the highest by a strong margin. This distribution is further visualized by Outcome Type and it becomes clear that adoptions are highest on weekends and in December, while transfers make up the difference during the weekdays. These distributions are shown on the following page.



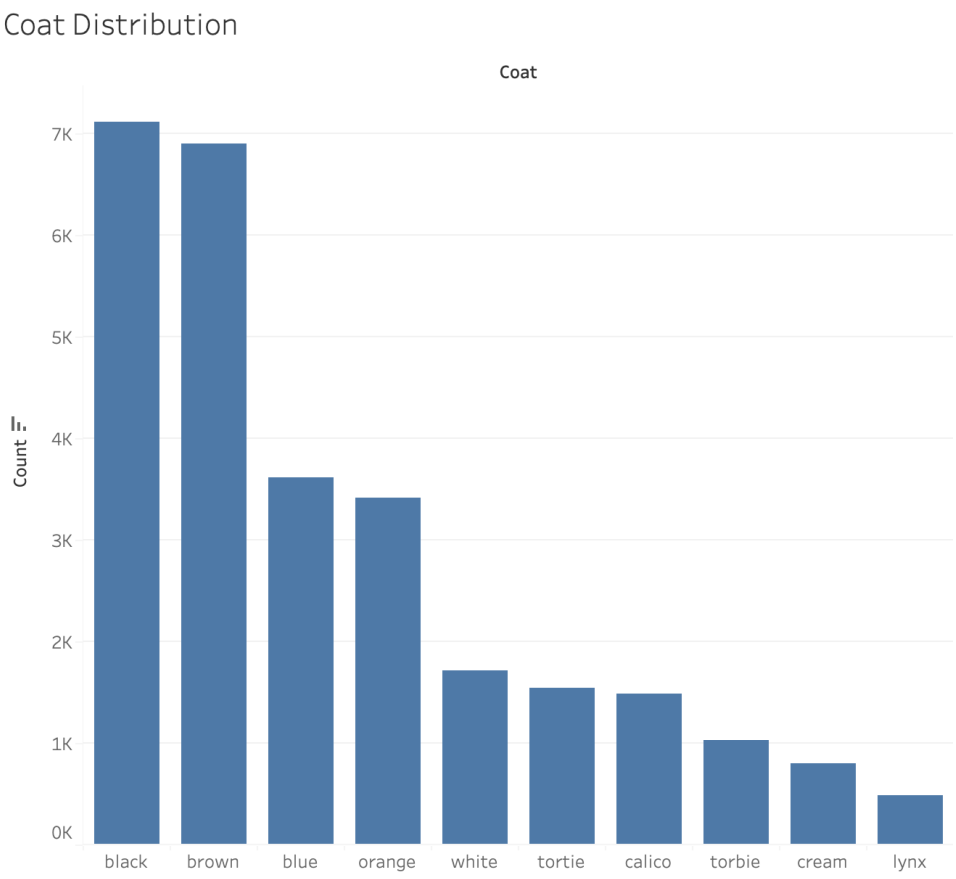
Outcome by Month



Outcome by Weekday



The breed and color columns are not very appropriate for visualizations and have many possible categories with relatively few entries in each, so they are not very useful for exploring the data and may not be useful for analysis. The coat column, however, provides a combination of the other columns to provide insight into the appearance of the cats that could prove to be useful in analysis. The distribution of the coat variable is below. The most frequent, by far, are black and brown.



4. Summary of Findings

The findings that resulted from this Exploratory Data Analysis stage were interesting and will provide a solid foundation on which to construct the future predictive analysis for cat adoptions and outcomes. In exploring the columns of this dataset, it becomes clear that many of these variables are redundant or were used to generate a more useful format of the data. Many of the datetime objects as well as Period and Period Range, for example, can be removed from the dataset as they were used to generate the more useful measures of time like Outcome Age (Days) and Outcome Age (Years). Period and Period Range, in fact, are quite meaningless without the transformation to the

Outcome Age variables. The same is true for many of the color and breed variables – through this exploration I determined that these columns were used to generate the coat column which proves to be the only useful representation of appearance for the cats. Many columns will be dropped in the next stage of this research before the analysis and modeling begins.

The distribution of outcomes in the dataset is quite promising in terms of finding homes for the cats in the dataset – the most frequent outcomes, by far, are Transfer and Adoption. According to the dataset source, transfer usually indicates movement for medical treatment or spaying/neutering, so this is generally a positive outcome. The other most common outcomes are return to owner, another positive result, and euthanasia. It is worth noting, in regard to euthanasia, that this data comes from a no-kill animal shelter so their frequency of euthanasia is likely much lower than other shelters – this won't affect this research but should be considered when thinking of implications of these results on animal shelters in general.

Other distributions show interesting patterns, like outcome month and weekday. These distributions give insight into temporal trends – Saturdays witness the highest frequency of outcomes, while Wednesdays, Thursdays, and Fridays sees the lowest. Furthermore, there are notable differences in outcome frequency across months, with February having the lowest and July having the highest frequency. Additionally, there's a clear association between outcome types and time, with adoptions peaking on weekends and during December. This makes sense as the holidays are likely to see more adoptions as pets are frequently gifted within families, and the public is generally more available to spend time on pet adoptions and welcoming them home during the weekend when they are less likely to be working and free time is more abundant.

Finally, there were significant insights into the appearances of cats in this dataset found using the coat column. The distribution shows that black and brown are the most common coat colors among sheltered cats. While the breed and color columns may not be suitable for detailed visualizations due to the numerous categories and sparse entries, the coat column offers valuable insights into the cats' appearances, which could be useful for further analysis. I am interested to learn about the impact of appearance on adoption during later stages of analysis as "cuteness" may be one of the main decision factors when adopting pets.

Overall, these findings provide a comprehensive overview of the dataset, shedding light on various aspects such as outcome distribution, spay/neuter status, age distribution, time related trends, and appearance characteristics. These

insights will serve as the foundation of subsequent predictive modeling and analysis to optimize shelter operations and improve animal outcomes.