**TBANLT 540**

**APPLIED REGRESSION MODELS**

**FINAL PROJECT: DATA ANALYSIS**

**ANALYSIS OF LIKELIHOOD OF ADOPTION OF A CAT**

**FOR**

A close up of a logo

Description automatically generated

**GROUP 16 –**

1. **AMER IMSIC**
2. **RATHANA SAN**
3. **SONIKA SHIVANI VIJAYKUMAR**

Table of Contents

[**Abstract:** 1](#_Toc26568027)

[**Introduction:** 1](#_Toc26568028)

[**Data Set:** 2](#_Toc26568029)

[Description of variables: 2](#_Toc26568030)

[Descriptive Statistics: 3](#_Toc26568031)

[**Analysis Plan:** 3](#_Toc26568032)

[Histograms: 4](#_Toc26568033)

[Residual Plots 10](#_Toc26568034)

[**Results:** 17](#_Toc26568035)

[**Limitations of study:** 18](#_Toc26568036)

[**Conclusion:** 19](#_Toc26568037)

[**References:** 19](#_Toc26568038)

# **Abstract:**

Before we started this analysis, we had the question of what kinds of cats were more, or less likely to be adopted. This question came to mind because one of us were told by a group of friends that if we were ever to adopt a cat, to adopt a black cat because they were the least likely to be adopted. After finding a dataset that worked for our question, we expanded our question from testing our theory with colors, to also considering age, breed, sex, etc. From our results, we found that kittens were more often adopted than cats. Cats were defined by being 6 months of age or older. When it came down to correlation significance, we didn’t find color to be a determining factor for adoption. We did find that some coat colors were less likely to be adopted but had a higher chance of being transferred, which meant the feline was moved to another shelter. According to our research, transfers can mean moving felines to other shelters due to capacity issues or someone looking for a specific sort of cat which can lead to adoption as well.

# **Introduction:**

The Austin Animal Center is the largest no-kill animal shelter in the United States that provides care and shelter to over 18,000 animals each year and is involved in a range of county, city, and state-wide initiatives for the protection and care of abandoned, at-risk, and surrendered animals. The agenda of this project is to spread awareness and help support and care for the animals who need it most. Also, this data analysis helps the animal welfare community to more effectively respond and identify animals that need more support to avoid unwanted outcomes.

The primary goal is to identify which variables have high correlation towards a cat being adopted. We’re working with adoption history of cats from 2013 to 2018 at Austin Animal Center Shelter. We expect cats that are younger, spayed/neutered, and have light colored coats to be more likely to be adopted.

The questions that we are interested in are:

1. What days or what time are the cats more likely to be adopted?
2. What color coated cats are more likely to be adopted?
3. What breeds of cats are most likely to be adopted?
4. Likelihood of adoption of cats versus kittens
5. Likelihood of adoption of Spayed versus neutered
6. How ‘Age upon outcome’, ‘outcome type’, ‘outcome age in years’, ‘sex age outcome’, ‘outcome month’ and ‘outcome weekday’ impact the adoption of a cat?

Our primary research question is **What attributes contribute to the likelihood of the adoption of a cat?**

# **Data Set:**

The dataset was obtained from Kaggle.

Link to the dataset: <https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and#aac_shelter_cat_outcome_eng.csv>

As part of the City of Austin Open Data Initiative, the Austin Animal Center makes available its collected dataset that contains statistics and outcomes of animals entering the Austin Animal Services system. The dataset contains shelter outcomes of several types of animals and breeds from 10/1/2013 to the present with an hourly time frequency.

The Austin Animal Center's original dataset includes columns for name, date of birth, outcome, animal type, sex and age at time of outcome, breed, and color. Outcomes range widely and include things like adoptions and transfers to other shelters.

The team worked on the dataset containing adoption of cats from the year 2013 to 2018. The dataset obtained was in CSV (comma-separated values) format. The table contained 27,944 rows (observations) and 37 columns (variables). Eleven variables that were considered not impacting the likelihood of adoption of a cat were removed. The cleaned dataset had 27,844 rows. 26 columns and 165658 data points.

## Description of variables:

1. **age\_upon\_outcome** - age of cat at outcome in week, month or year.
2. **animal\_id** – unique identifier
3. **breed**
4. **color**
5. **datetime** – time and date the outcome happened
6. **outcome\_subtype** – detailed outcome
7. **outcome\_type** – categorized outcome types (adoption, died, euthanasia, rto-adopt and transfer)
8. **sex\_upon\_outcome** – spayed/neutered, male/female
9. **sex**
10. **Spay/Neuter** – yes/no
11. **outcome\_age\_(days)** - age of cat in days
12. **outcome\_age\_(years)** – age of cat in years
13. **Cat/Kitten (outcome)** – cat/kitten. If 6 months or older, it’s a cat, otherwise, a kitten.
14. **sex\_age\_outcome** - spayed/neutered, male/female, cat/kitten
15. **outcome\_month** – month of outcome
16. **outcome\_year** – year of outcome
17. **outcome\_weekday** – weekday of outcome
18. **outcome\_hour** – hour of outcome
19. **breed1** – first breed
20. **breed2** – potential second breed
21. **cfa\_breed** – cfa standard true/false
22. **domestic\_breed** – domestic true/false
23. **coat\_pattern** – pattern of coat (tabby, torbie, point….)
24. **color1** – primary color
25. **color2** - secondary color
26. **coat** – overall color

Variable types – **Categorical, Numerical, Boolean and Datetime**

## Descriptive Statistics:

A screenshot of a cell phone

Description automatically generated

**Analysis Plan:**

We used logistic regression modeling to find out the attributes that contribute to the adoption of a cat. We plotted residual plots and scatter plots with Gretl to check for linearity/non-linearity. We gave numeric codes to the non-numeric data. We created dummy code for our categorical variables in order to run regression models using our data. We used ordinary least squares (OLS) regression to find the significance between outcome age in years and outcome type. We used OLS because our data was continuous and dummy coded. We conducted Multinomial Regression in an attempt to find significant correlation between ‘outcome type’ with the predictive variables. We used Multinomial Logistic Regression in our analysis as the dependent variable was nominal with more than 2 levels (more than 2 numeric codes). We

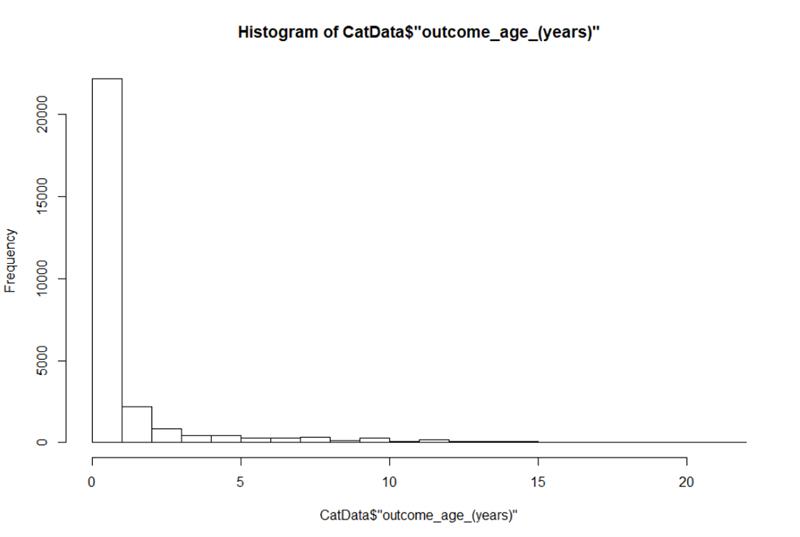
A two-class regression (Azure ML Studio) was performed using the outcomes died and adopted. The independent variables used were age\_upon\_outcome, outcome\_type, cat/kitten, Spay/Neutered, Breed, outcome\_weekday, outcome\_month, and coat.

We looked for lower p-values in the predictive variables. We selected the most parsimonious model by comparing different regression techniques. Predictive variables: Breed, Spay/Neutered, age\_upon\_outcome, color, datetime, coat\_pattern.

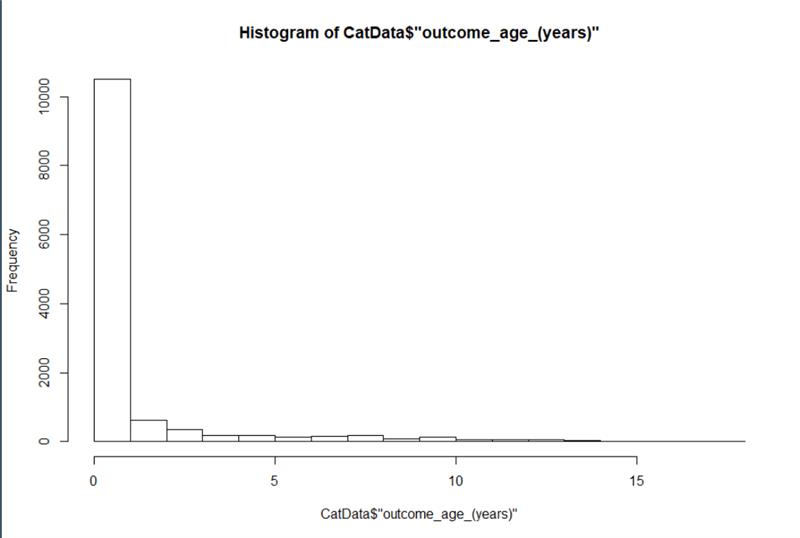
Dependent variable: outcome\_type

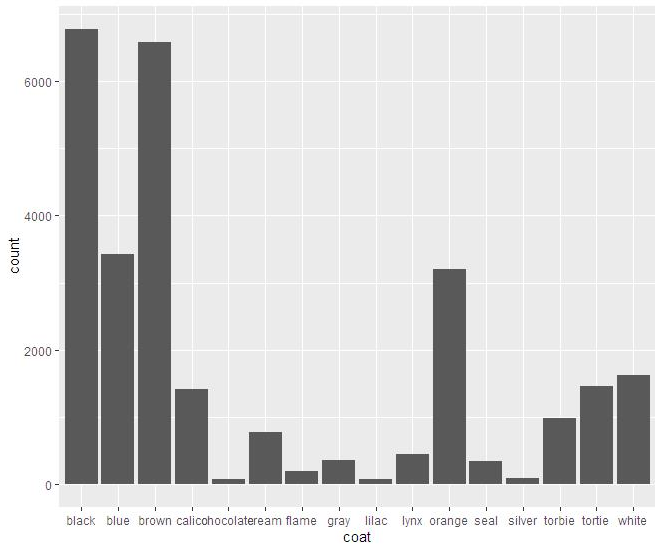
## Histograms:

**A) Outcome-age of all the cats**

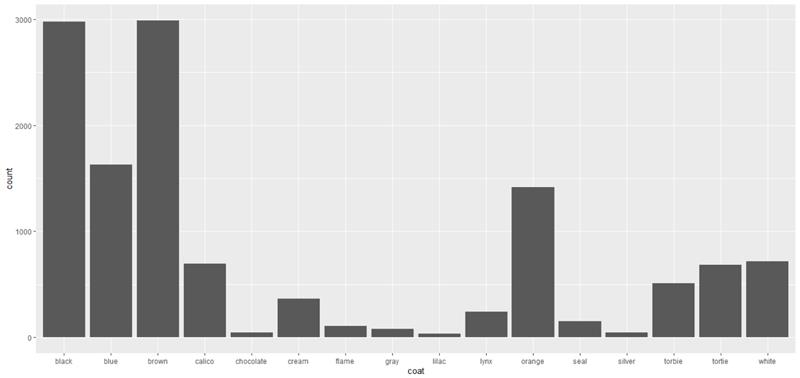


**B) Outcome age of adopted cats**

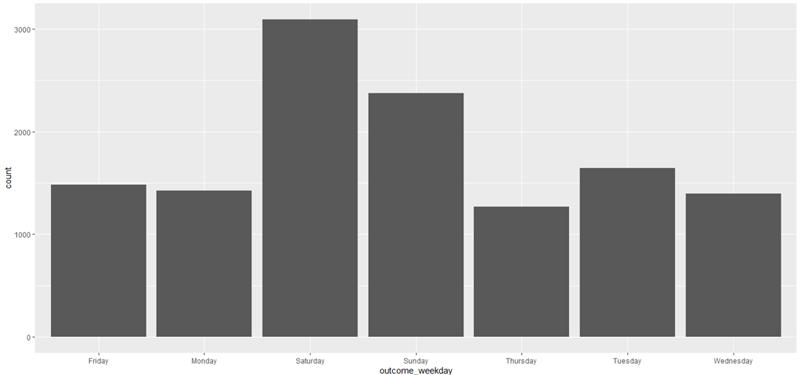


**C) Significant coat colors** 

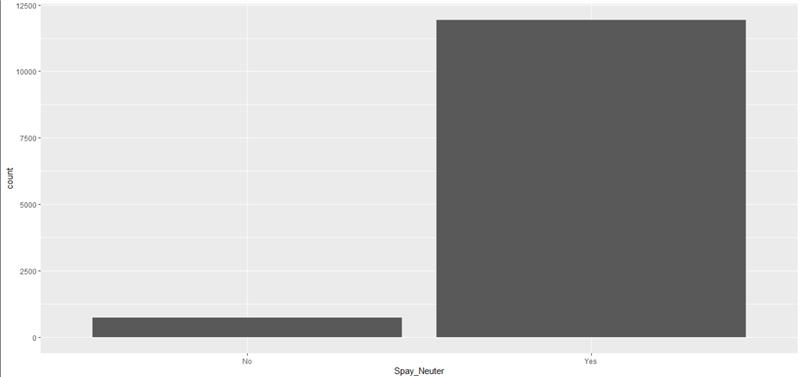
**D) Coat color of adopted cats**



**E) Day distributions for adoptions (does not include return to owner adopt)**

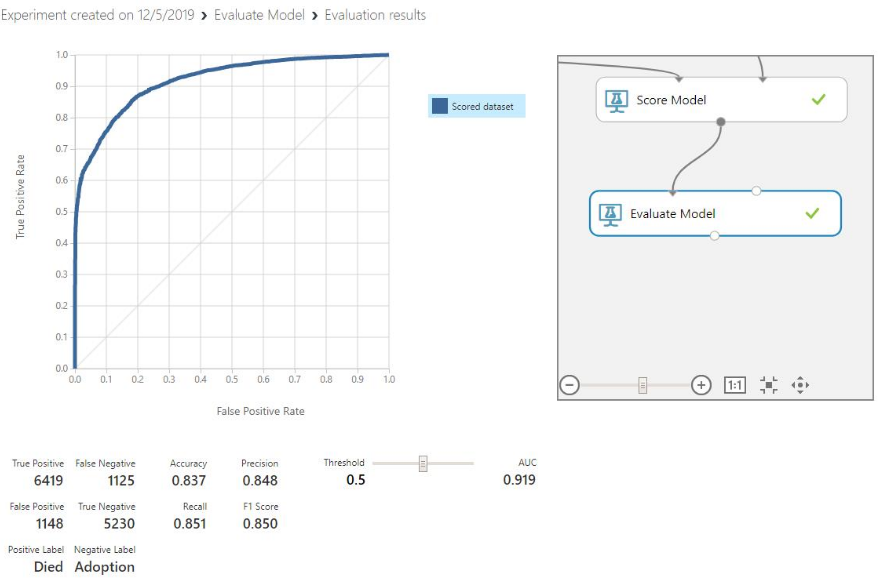


**F) Spay/Neuter adoptions**



**G) Two class logistic regression (in Azure Machine Learning Studio)**

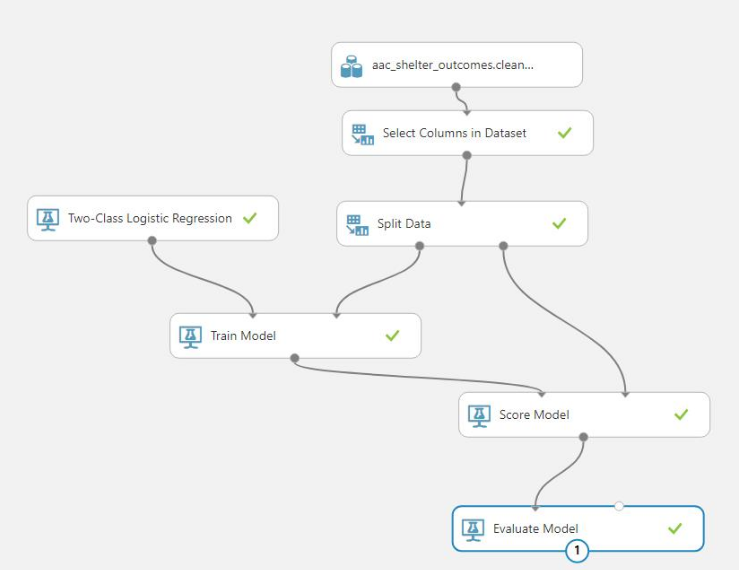
**Part-A**



**Part-B**



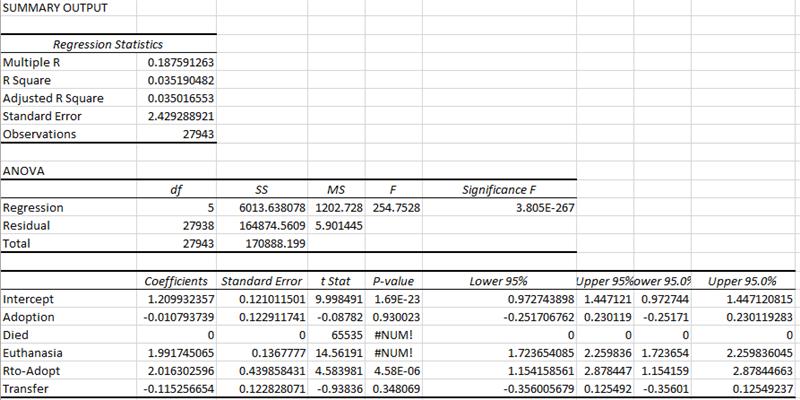
**Part-C**



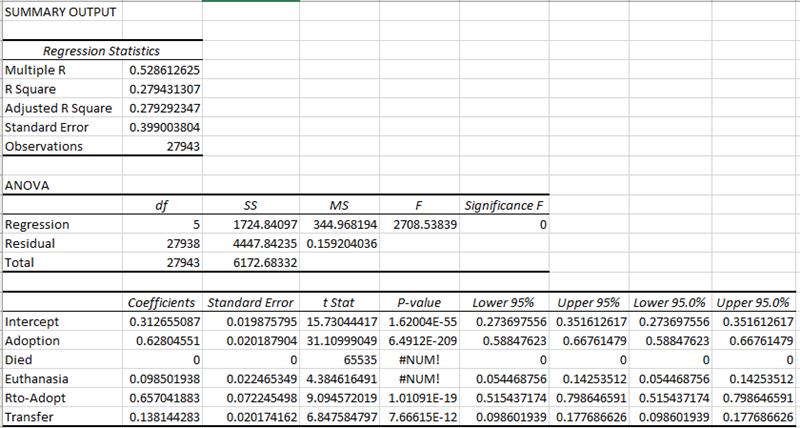
**H) Outcome regression for outcome-age in years, in excel**

Variable Y – Outcome age

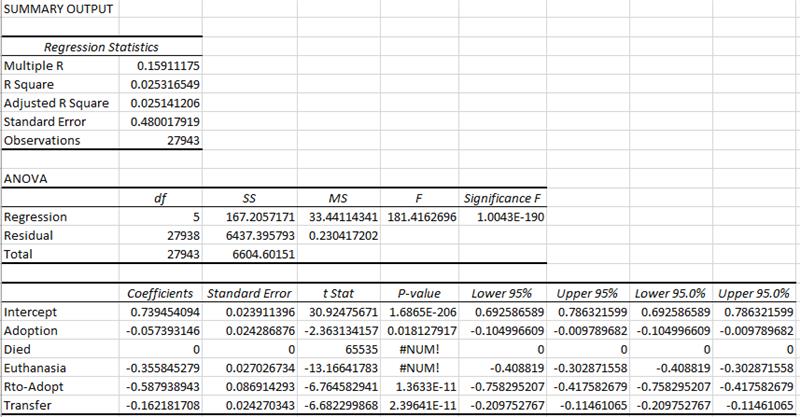
Variable X – Outcome type



**I) Outcome regression for Kitten/Cat (Kitten=1, Cat=0)**

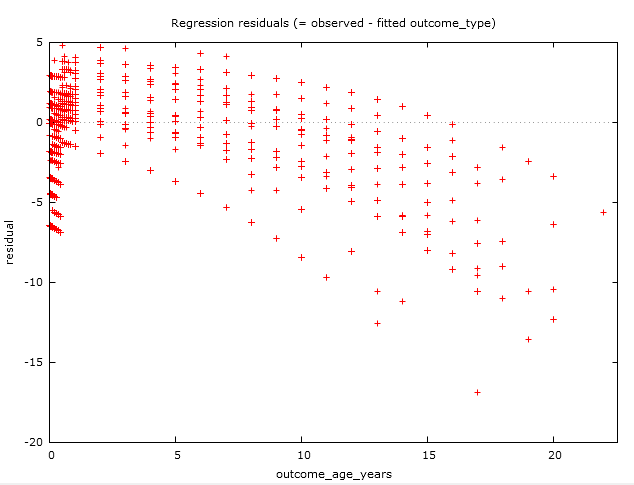


**J) Outcome regression for Female/Male (Female=1, Male=0)**

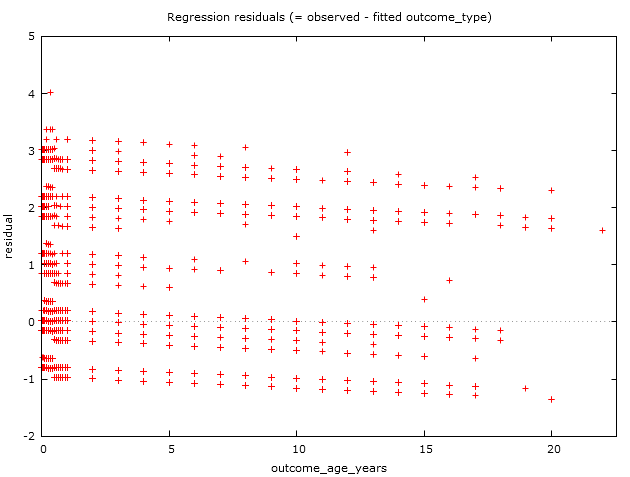


## Residual Plots

The first residual plot is using logistic regression and shows a cluster of points where the age is 0-1. As the age grows, the residual thins out and gets smaller.

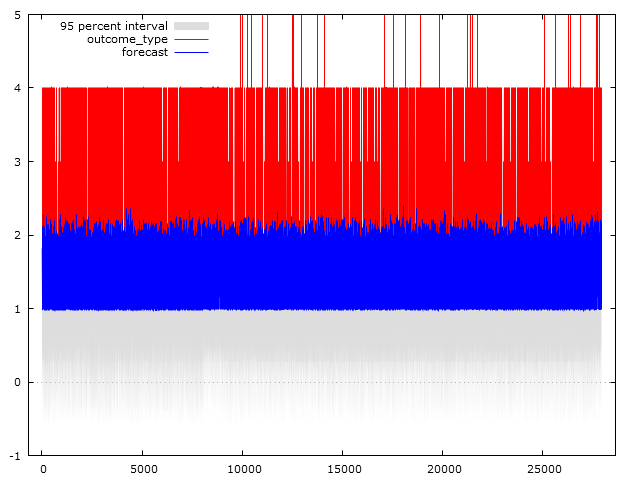


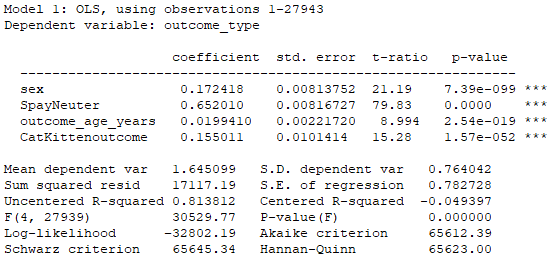
This residual plot was done with OLS regression. We see a similar clustering fora smaller value of age to the logistic regression, but the points are not the residuals are not as low as the logistic regression.

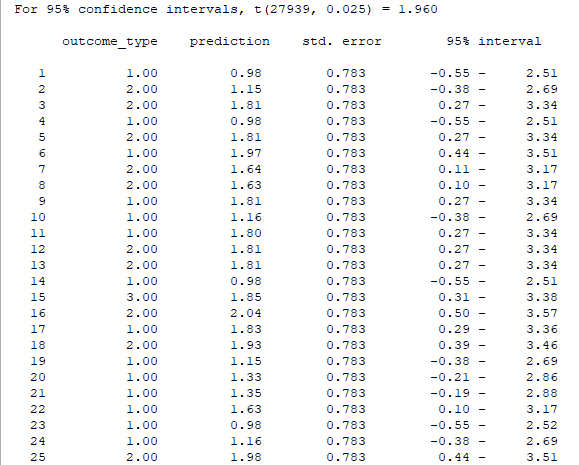


**Forecasts with OLS regression:**

Below is the forecast graph provided by Gretl. It’s predicting outcome type based on sex, outcome age in years, cat or kitten outcome, and spayed or neutered value. Below it you can see prediction values. We found that some predicted values were very close to the actual value, but it wasn’t consistent enough to be certain of a direct correlation.







**Table 1**

This table represents the percentage of outcomes for each group (Cat and Kitten)

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of Cat/Kitten (outcome)** | **Column Labels** |  |  |
| **Row Labels** | **Cat** | **Kitten** | **Grand Total** |
| Adoption | 37.82% | 50.40% | 45.59% |
| Died | 0.98% | 1.73% | 1.44% |
| Euthanasia | 8.32% | 3.21% | 5.17% |
| Rto-Adopt | 0.25% | 0.03% | 0.11% |
| Transfer | 52.63% | 44.63% | 47.69% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |

**Table 2**

This table represents the percentage of outcomes for the combined groups of Cat and Kitten

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of Cat/Kitten (outcome)** | **Column Labels** |  |  |
| **Row Labels** | **Cat** | **Kitten** | **Grand Total** |
| Adoption | 14.47% | 31.11% | 45.59% |
| Died | 0.37% | 1.07% | 1.44% |
| Euthanasia | 3.19% | 1.98% | 5.17% |
| Rto-Adopt | 0.10% | 0.02% | 0.11% |
| Transfer | 20.14% | 27.55% | 47.69% |
| **Grand Total** | **38.27%** | **61.73%** | **100.00%** |

**Table 3**

The below table is split in parts. These are the percentages of outcomes per color. The percentage grouping is done by color.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Count of coat** | **Column Labels** |  |  |  |  |  |  |  |
| **Row Labels** | **black** | **blue** | **brown** | **calico** | **chocolate** | **cream** | **flame** | **gray** |
| Adoption | 43.88% | 47.48% | 45.42% | 49.22% | 56.96% | 47.29% | 55.44% | 22.03% |
| Died | 1.61% | 1.49% | 1.47% | 0.92% | 1.27% | 2.58% | 1.55% | 2.26% |
| Euthanasia | 5.56% | 4.40% | 4.92% | 4.38% | 5.06% | 3.61% | 4.15% | 13.28% |
| Rto-Adopt | 0.09% | 0.09% | 0.11% | 0.14% | 0.00% | 0.13% | 0.52% | 0.00% |
| Transfer | 48.86% | 46.55% | 48.08% | 45.34% | 36.71% | 46.39% | 38.34% | 62.43% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
| **lilac** | **lynx** | **orange** | **seal** | **silver** | **torbie** | **tortie** | **white** | **Grand Total** |
| 44.30% | 52.53% | 44.31% | 43.90% | 49.47% | 52.19% | 47.02% | 44.29% | 45.59% |
| 0.00% | 1.54% | 1.10% | 2.03% | 0.00% | 0.71% | 1.03% | 1.73% | 1.44% |
| 8.86% | 5.05% | 5.35% | 8.14% | 4.21% | 3.26% | 4.18% | 6.92% | 5.17% |
| 0.00% | 0.22% | 0.19% | 0.00% | 0.00% | 0.00% | 0.00% | 0.31% | 0.11% |
| 46.84% | 40.66% | 49.06% | 45.93% | 46.32% | 43.83% | 47.77% | 46.76% | 47.69% |
| **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** |

**The tables**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Count of outcome\_type** | **Column Labels** |  |  |  |  |  |  |  |  |
|  | **black** |  | **black Total** | **blue** |  | **blue Total** | **brown** |  | **brown Total** |
| **Row Labels** | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  |
| Adoption | 36.06% | 48.30% | 43.88% | 38.76% | 52.76% | 47.48% | 36.86% | 50.32% | 45.42% |
| Died | 1.27% | 1.80% | 1.61% | 0.62% | 2.01% | 1.49% | 1.13% | 1.67% | 1.47% |
| Euthanasia | 8.94% | 3.65% | 5.56% | 7.26% | 2.67% | 4.40% | 7.80% | 3.27% | 4.92% |
| Rto-Adopt | 0.24% | 0.00% | 0.09% | 0.23% | 0.00% | 0.09% | 0.25% | 0.02% | 0.11% |
| Transfer | 53.49% | 46.25% | 48.86% | 53.13% | 42.56% | 46.55% | 53.96% | 44.71% | 48.08% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **calico** |  | **calico Total** | **chocolate** |  | **chocolate Total** | **cream** |  | **cream Total** | **flame** |  | **flame Total** |
| **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  |
| 41.76% | 56.11% | 49.22% | 56.25% | 57.45% | 56.96% | 34.07% | 54.35% | 47.29% | 53.19% | 57.58% | 55.44% |
| 0.74% | 1.09% | 0.92% | 3.13% | 0.00% | 1.27% | 1.48% | 3.16% | 2.58% | 1.06% | 2.02% | 1.55% |
| 7.21% | 1.77% | 4.38% | 12.50% | 0.00% | 5.06% | 7.41% | 1.58% | 3.61% | 4.26% | 4.04% | 4.15% |
| 0.29% | 0.00% | 0.14% | 0.00% | 0.00% | 0.00% | 0.37% | 0.00% | 0.13% | 1.06% | 0.00% | 0.52% |
| 50.00% | 41.03% | 45.34% | 28.13% | 42.55% | 36.71% | 56.67% | 40.91% | 46.39% | 40.43% | 36.36% | 38.34% |
| **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **gray** |  | **gray Total** | **lilac** |  | **lilac Total** | **lynx** |  | **lynx Total** | **orange** |  | **orange Total** |
| **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  |
| 23.48% | 21.17% | 22.03% | 34.78% | 48.21% | 44.30% | 50.53% | 53.96% | 52.53% | 36.23% | 48.85% | 44.31% |
| 0.76% | 3.15% | 2.26% | 0.00% | 0.00% | 0.00% | 1.05% | 1.89% | 1.54% | 0.61% | 1.37% | 1.10% |
| 22.73% | 7.66% | 13.28% | 30.43% | 0.00% | 8.86% | 6.32% | 4.15% | 5.05% | 9.64% | 2.93% | 5.35% |
| 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.53% | 0.00% | 0.22% | 0.26% | 0.15% | 0.19% |
| 53.03% | 68.02% | 62.43% | 34.78% | 51.79% | 46.84% | 41.58% | 40.00% | 40.66% | 53.26% | 46.70% | 49.06% |
| **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **seal** |  | **seal Total** | **silver** |  | **silver Total** | **torbie** |  | **torbie Total** | **tortie** |  | **tortie Total** |
| **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  | **Cat** | **Kitten** |  |
| 37.72% | 49.72% | 43.90% | 44.19% | 53.85% | 49.47% | 42.23% | 59.40% | 52.19% | 38.37% | 54.20% | 47.02% |
| 0.00% | 3.95% | 2.03% | 0.00% | 0.00% | 0.00% | 0.73% | 0.70% | 0.71% | 0.91% | 1.13% | 1.03% |
| 8.98% | 7.34% | 8.14% | 6.98% | 1.92% | 4.21% | 4.61% | 2.28% | 3.26% | 5.59% | 3.01% | 4.18% |
| 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 53.29% | 38.98% | 45.93% | 48.84% | 44.23% | 46.32% | 52.43% | 37.61% | 43.83% | 55.14% | 41.66% | 47.77% |
| **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** | **100.00%** |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **white** |  | **white Total** | **Grand Total** |
| **Cat** | **Kitten** |  |  |
| 38.69% | 48.13% | 44.29% | 45.59% |
| 1.21% | 2.08% | 1.73% | 1.44% |
| 11.53% | 3.75% | 6.92% | 5.17% |
| 0.61% | 0.10% | 0.31% | 0.11% |
| 47.95% | 45.94% | 46.76% | 47.69% |
| **100.00%** | **100.00%** | **100.00%** | **100.00%** |

**Results:**

As seen in table 1-3, one of our approaches in analyzing the data was finding the percentages of the outcomes for cats. We looked at percentages within each group we observed and the percentage of the grand total. For example, table 1 has percentages of the individual groups, so out of the total group of just kittens, 50.4% were adopted. However, when we account for all the cats by including cats in our calculations, only 31.11% of kittens were adopted. Kittens make up 61.73% of our total data, so relying on the percentage of the total alone can be misleading. In both instances, kittens had a higher rate of natural death while cats had a higher rate of euthanasia in both instances. With the percentage of each group, we can conclude that there is a correlation between age of the feline and adoption rate. According to this data set, about 13% more kittens will get adopted when compared to grown cats.

The two-class regression performed in Azure inaccurately categorized variables into “died” for a true positive. Based on the percentage data from table 1 the outcomes are split evenly between transfer (47%) and adoption (45%). However, when the other categories euthanasia, rto-transfer, and died are clumped into those two it is roughly a 50/50 split. This explains the high accuracy (0.83) and precision (0.83). Therefor this graph does not accurately display the distribution of all the outcomes. This explains the large area under the curve which is misleading in making “died” (true positive) a highly probable outcome.

Kitten versus cat is the only variable correlation of significance. The logistic regression performed in Excel showed the R-squared value to be 0.279. Which was the highest value obtained through different methods of regression analysis performed. The other two variables used for the Excel regression were male versus female and age of outcome(months). Both had R-squared values of under 0.03.

The residual plots in section J show the negative correlation between age and outcome type. As age increases the cat is less likely to be adopted. This is seen in the heavy clustering of data points for cats under the age of one (kitten).

**Limitations of study:**

One of our limitations was that we had categorical variables and a lack of useful numerical. This made doing a regression model and plotting it tougher. We had to create “dummy code” as a work-around to this issue. This doesn’t represent the most accurate results but the results we got were so far away from being useful that we can conclude that there wasn’t any correlation in those sets. This caused issues when running logistic regressions as strange clustering occurred (in Gretl). Data points were binary and clustering incorrectly. For this reason, many graphs could not be used. There was difficulty creating dummy code for variables such as coat color which had 16 types. The binary translation of complex data caused the regressions to run with many errors.

Another limitation was a limitation in the data. It seems like the outcome types were only from this specific shelter. Once a cat or kitten got transferred, we lose any sort of outcome after the fact. This means that the cat may have gotten euthanized, died, or adopted in the shelter it was transferred to. For the purpose of data analysis, having 1 database between all the shelters would prove most useful, but that doesn’t make sense from an operational or business standpoint.

**Conclusion:**

From our analysis, we found that our initial thought of cat color being a predictor of adoption was wrong. However, we were right to assume that younger cats were more likely to get adopted, as well as cats that were spayed or neutered. Using this data, the shelter can take some actions to improve the adoption rates of cats that are 6 months and older (not kittens). The shelter can spay and neuter the cats since they are more likely to be adopted than those cats who are not. Another action they can take is to promote older cats more. From a buyer’s standpoint, buying a kitten is the choice that makes sense on initial thought since they want the kitten to grow up around them and be friendly with them. On the other hand, adult cats may be more likely to be potty trained, can be just as friendly, may already be disciplined to avoid bad habits, and you can judge their behavior better since they are matured. If the shelter promotes these positive traits as selling/adoption point, they may have a higher rate of adoption for cats and lower rate of euthanization.

# **References:**

1. <https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-outcomes-and#aac_shelter_cat_outcome_eng.csv>