Predicting Shelter Stay Duration

Group 20

```
library(ggplot2)
library(tidyverse)
library(gt)
library(patchwork)
library(gridExtra)
library(viridis)
library(plotly)
library(dplyr)
library(GGally)
library(lubridate)
```

1 Introduction

Animal shelters play a critical role in managing stray and surrendered animals,yet the duration of an animal's stay before reaching its final outcome varies significantly. This study analyzes data from a Dallas animal shelter to investigate which factors impact the number of days an animal remains in the shelter before an outcome is determined.

To analyze this,we utilize descriptive statistics, data visualization, ANOVA, and a Generalized Linear Model (GLM) to assess the impact of animal type, intake type and other variables on shelter stay duration.

```
data <- read.csv("dataset20.csv")
data$animal_type <- as.factor(data$animal_type)
data$intake_type <- as.factor(data$intake_type)
data$outcome_type <- as.factor(data$outcome_type)
data$chip_status <- as.factor(data$chip_status)
data$season <- cut(data$month, breaks = c(2, 5, 8, 11, 12),
                   labels = c('Spring', 'Summer', 'Autumn', 'Winter'),
                   include.lowest = TRUE)
```

```r
data$season[data$month %in% c(12, 1, 2)] <- 'Winter'
data$season <- factor(data$season, levels = c("Spring", "Summer", "Autumn", "Winter"))
```

2 Exploratory data analysis

We have a final dataset consisting of 1405 animals with the following key attributes:

- **Animal_type** The type of animal admitted to the shelter
- **Month** Month the animal was admitted, recorded numerically with January=1
- **Year** Year the animal was admitted to the shelter.
- **Intake_type** Reason for the animal being admitted to the shelter
- **Outcome_take** Final outcome for the admitted animal
- **Chip_Status** Did the animal have a microchip with owner information?
- **Time_at_Shelter** Days spent at the shelter between being admitted and the final outcome.

```r
ggplot(data, aes(x = time_at_shelter)) +
  geom_histogram(binwidth = 5, fill = "pink", alpha = 0.6, color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Time Spent in Shelter", x = "Days in Shelter", y = "Count")
```

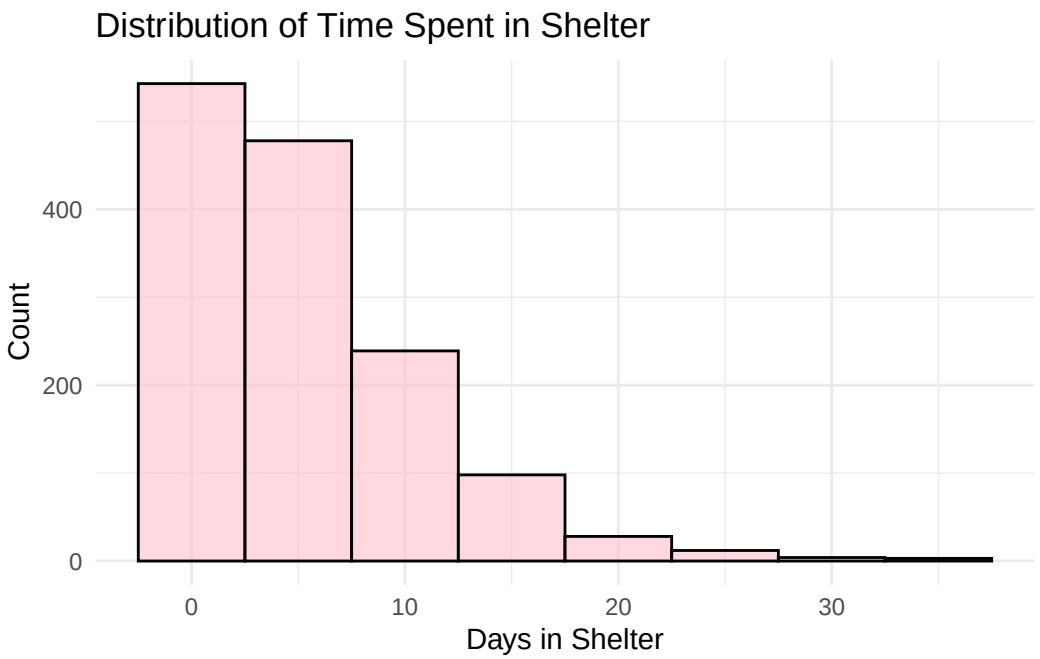

Figure 1: Distribution of Time Spent in Shelter

Firstly, figure 1 displays the distribution of time spent in shelter by animals and it shows right-skewed, indicating most animals stay for fewer than 10 days and small number of animals remain for extend periods.

```
ggplot(data, aes(x = animal_type, y = time_at_shelter, fill = animal_type)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Time Spent in Shelter by Animal Type", x = "Animal Type", y = "Days in Shelte
```

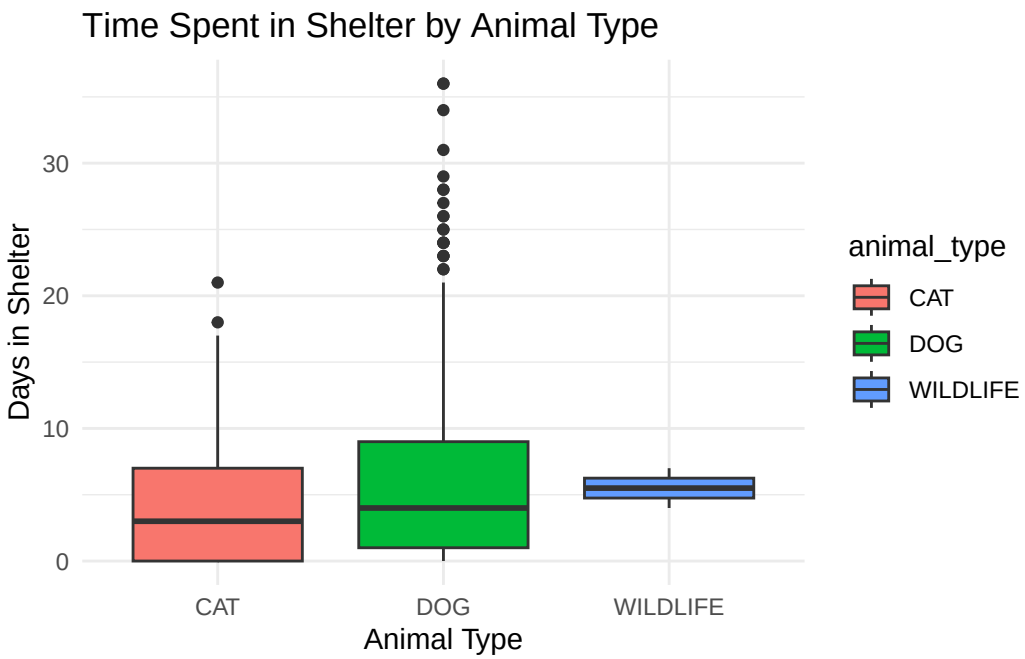

Figure 2: Time Spent in Shelter by Animal Type

The boxplot of Figure 2 visualizes the distribution of time spent in the shelter for different animal types. Dogs and cats occupy a large proportion of all animals in the shelter and they exhibit the widest range of shelter stay, with a considerable number of outliers indicating that some of them stay significantly longer than others. In contrast,birds and wildlife tend to have shorter and more consistent stay duration. However, The median stay duration across all animal types appears relatively low, indicating that most animals are processed efficiently, though certain cases, particularly among dogs and cats, experience extended stays.

```
ggplot(data, aes(x = intake_type, y = time_at_shelter, fill = intake_type)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Time in Shelter by Intake Type", x = "Intake Type", y = "Days in Shelter")
```

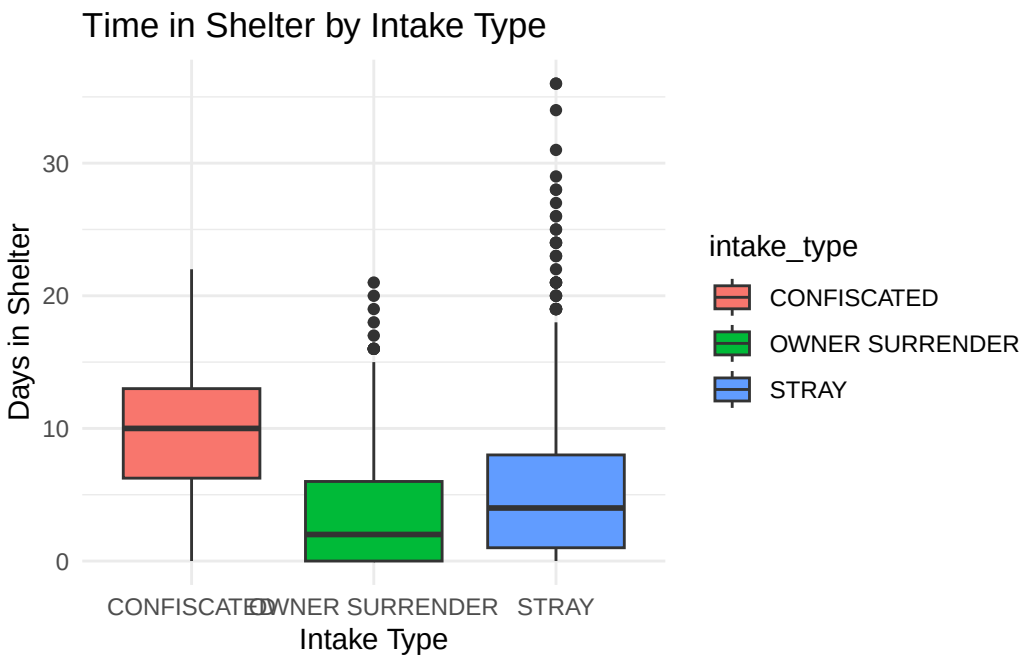

Figure 3: Time in Shelter by Intake Type

We also explore the distribution of time spent in the shelter based on different intake types, as shown in Figure 3, highlighting notable variations in shelter stay duration. The boxplot indicates that confiscated animals tend to stay in the shelter longer than those that are owner-surrendered or stray. Additionally, stray animals exhibit a wider spread and more outliers, suggesting that some cases remain in the shelter significantly longer than the majority.

```
ggplot(data, aes(x = chip_status, y = time_at_shelter, fill = chip_status)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Time in Shelter by Chip Status", x = "Chip Status", y = "Days in Shelter")
```

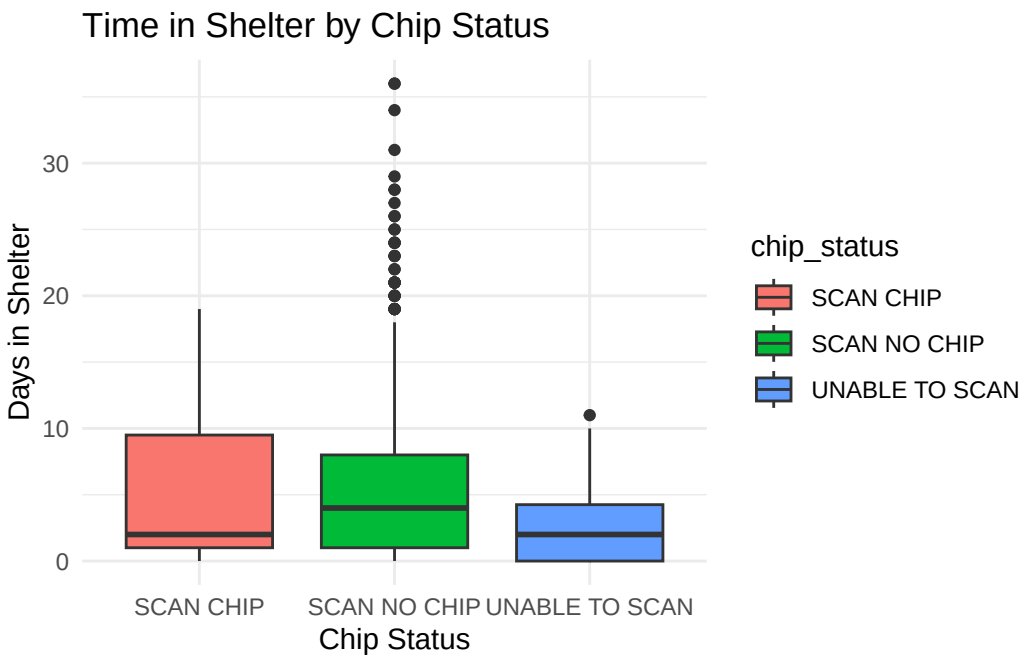

Figure 4: Time in Shelter by Chip Status

The relationship between chip status and shelter stay duration shows that animals with a scannable chip, no chip, or an unreadable chip all exhibit similar median shelter stays, so we assume that they might slightly affect the days in shelters.

```
data$admission_date <- make_date(year = data$year, month = data$month)
data$year_month <- format(data$admission_date, "%Y-%m")
ggplot(data, aes(x = year_month, y = time_at_shelter)) +
  geom_boxplot(fill = "lightblue") +
  theme_minimal() +
  labs(title = "Distribution of Shelter Time by Month",
       x = "Admission Date (Year-Month)",
       y = "Time in Shelter (Days)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

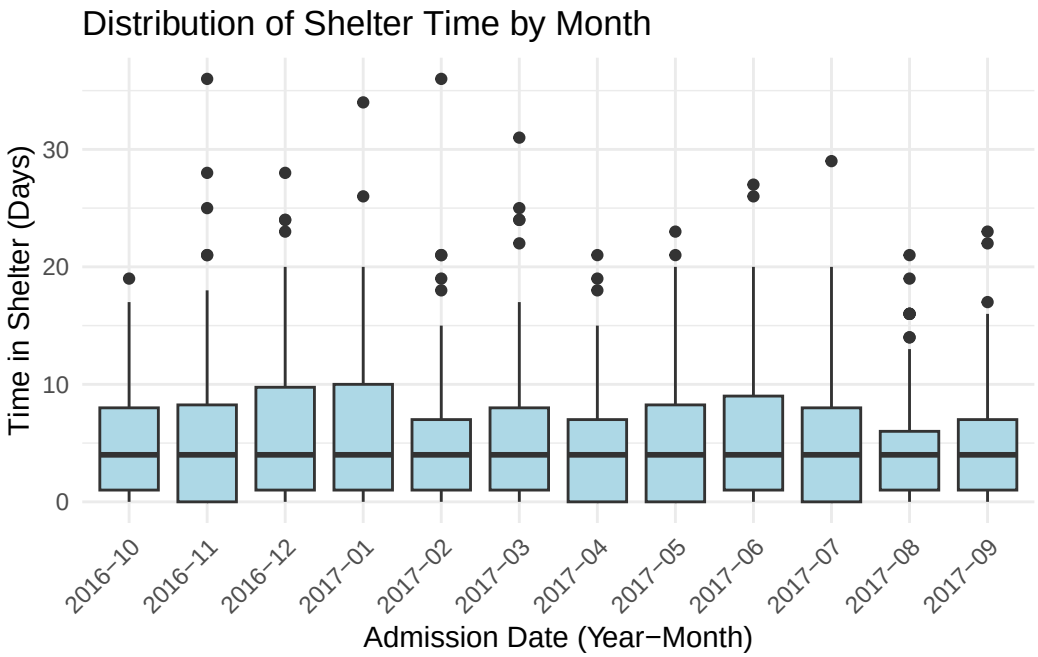

Figure 5: Distribution of Shelter Time by Month

Additionally, we found that there is no significant difference in shelter stay duration across different admission months. The median shelter stay remains relatively stable throughout the observed period, with only slight variations.

Also, animals spend slightly more time in shelter in winter than other season and there is no apparent different median among all seasons from the figure 6.

```
ggplot(data, aes(x = season, y = time_at_shelter, fill = season)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Time in Shelter by season", x = "Season", y = "Days in Shelter")
```

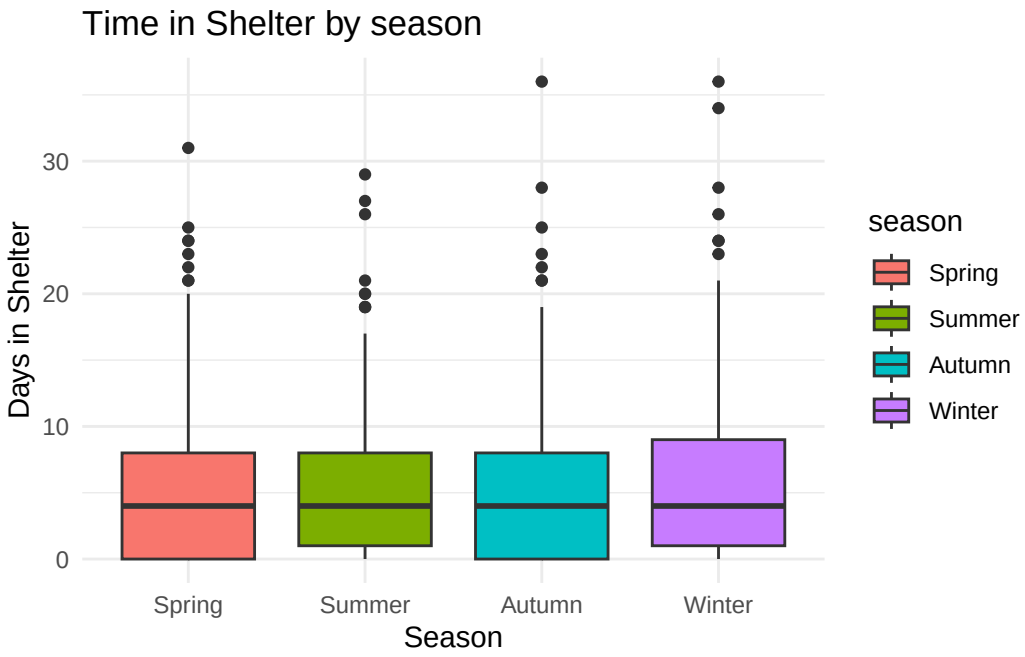

Figure 6: Time in Shelter by Season

To further explore the impact of those variable on the time of animals staying at shelter, we draw a ANOVA table to validate it. The ANOVA results indicate that intake type ($p < 0.001$) and outcome type ($p < 0.001$) have a highly significant impact on shelter stay duration. This aligns with the boxplots we analyze before, where different intake methods (e.g.strays vs. owner surrenders) and outcomes (e.g. adoption vs. euthanasia) showed clear differences in stay duration. And animal type also has a moderate effect ($p = 0.0322$), which means there have some differences across species. However, chip status is not significant ($p = 0.0740$), supporting the earlier boxplot observation that having a chip does not strongly influence shelter stay duration.

```
anova_model <- aov(time_at_shelter ~ animal_type + intake_type + outcome_type + chip_status,
summary(anova_model)
```

```
              Df Sum Sq Mean Sq F value   Pr(>F)
animal_type    2    469   234.4  10.651 2.57e-05 ***
intake_type    2   2344  1172.0  53.269  < 2e-16 ***
outcome_type   4   9126  2281.5 103.695  < 2e-16 ***
chip_status    2    120    59.9   2.721   0.0662 .
Residuals   1394  30671    22.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To further quantify these relationships and predict shelter stay duration, we will now construct a Generalized Linear Model (GLM).