Predicting Shelter Stay Duration

Group 20

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

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.

```r
# The process of cleaning the data is reflected in the lines 264-267.
df <- read.csv("dataset20_cleaned.csv")

# transform the type of variables
df$animal_type <- as.factor(df$animal_type)
df$intake_type <- as.factor(df$intake_type)
df$outcome_type <- as.factor(df$outcome_type)
```

```r
df$chip_status <- as.factor(df$chip_status)
df$season <- cut(df$month, breaks = c(2, 5, 8, 11, 12),
                 labels = c('Spring', 'Summer', 'Autumn', 'Winter'),
                 include.lowest = TRUE)
df$season[df$month %in% c(12, 1, 2)] <- 'Winter'
df$season <- factor(df$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.
- **Season** Season the animal was admitted

```r
ggplot(df, 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")
```

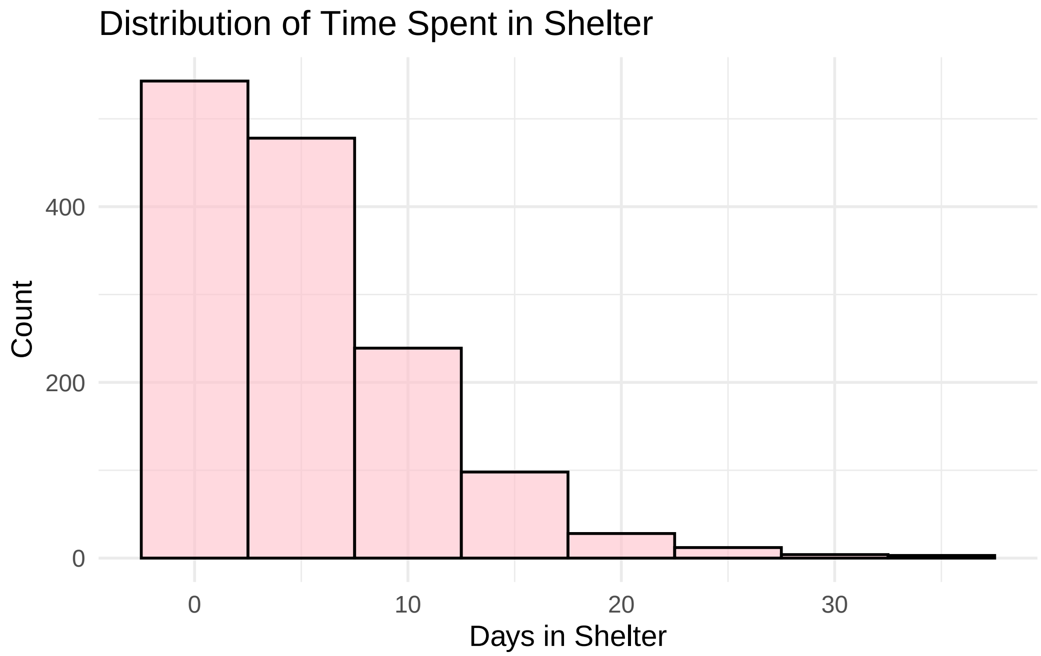


Figure 1: Distribution of Time Spent in Shelter

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(df, 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
```

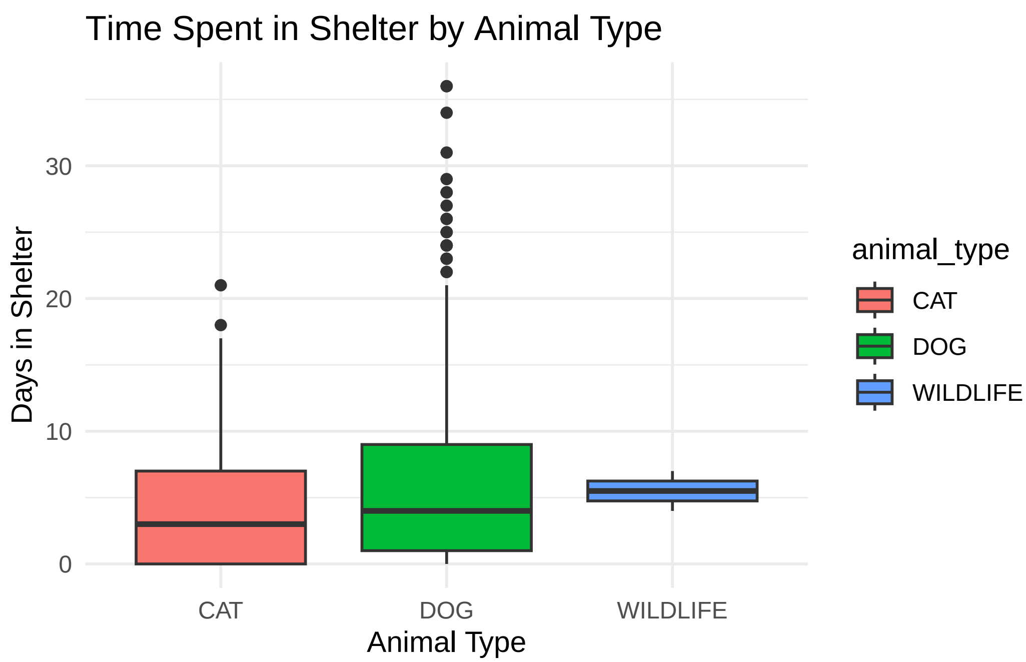


Figure 2: Time Spent in Shelter by Animal Type

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(df, 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")
```

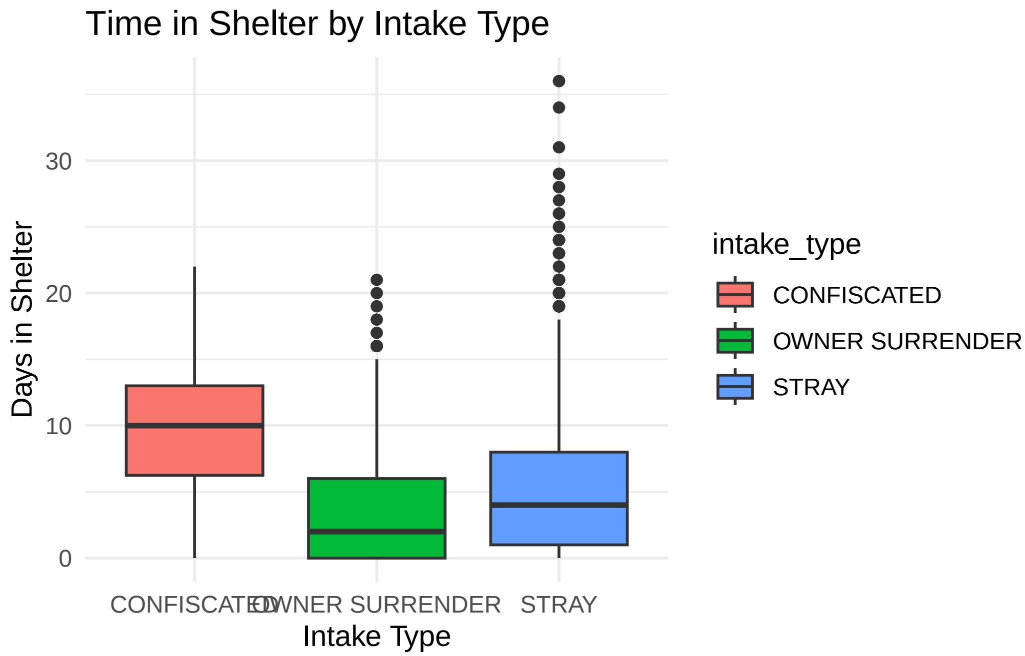


Figure 3: Time in Shelter by Intake Type

We also explore the distribution of time spent in the shelter based on different intake types shown as figure 3, highlighting notable variations in shelter stay duration.The boxplot shows 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(df, 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")
```

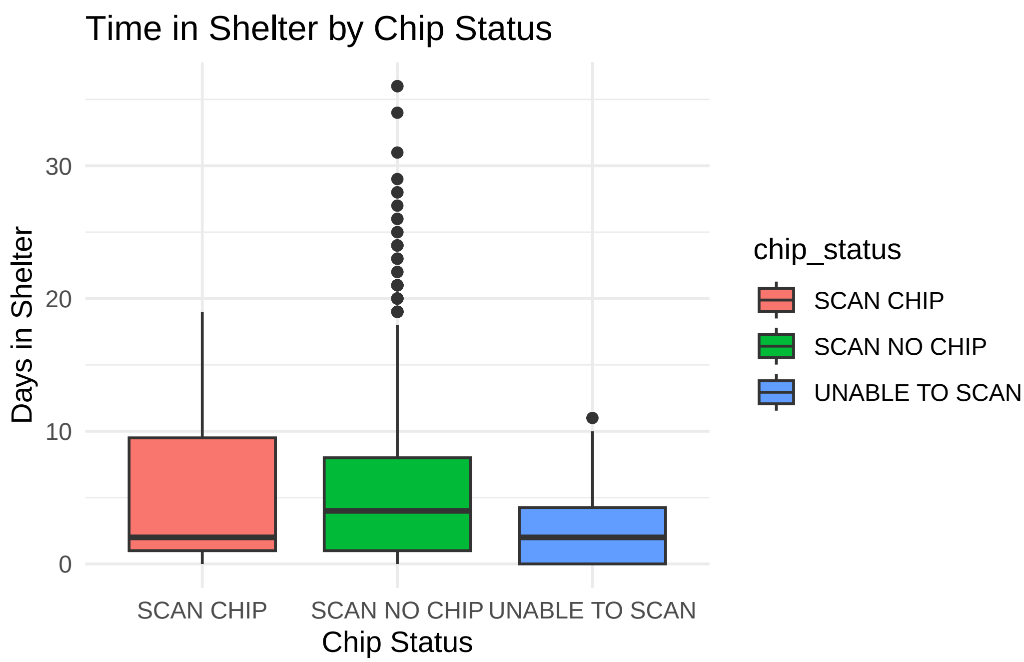


Figure 4: Time in Shelter by Chip Status

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.

```
df$admission_date <- make_date(year = df$year, month = df$month)

df$year_month <- format(df$admission_date, "%Y-%m")

ggplot(df, 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))
```

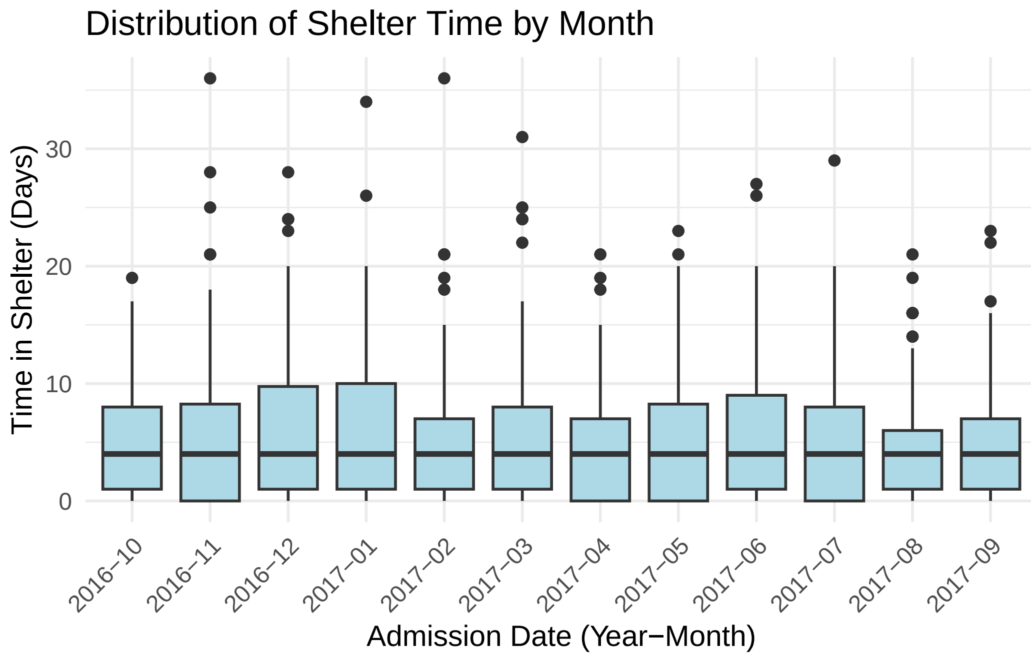


Figure 5: Distribution of Shelter Time by Month

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(df, 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")
```

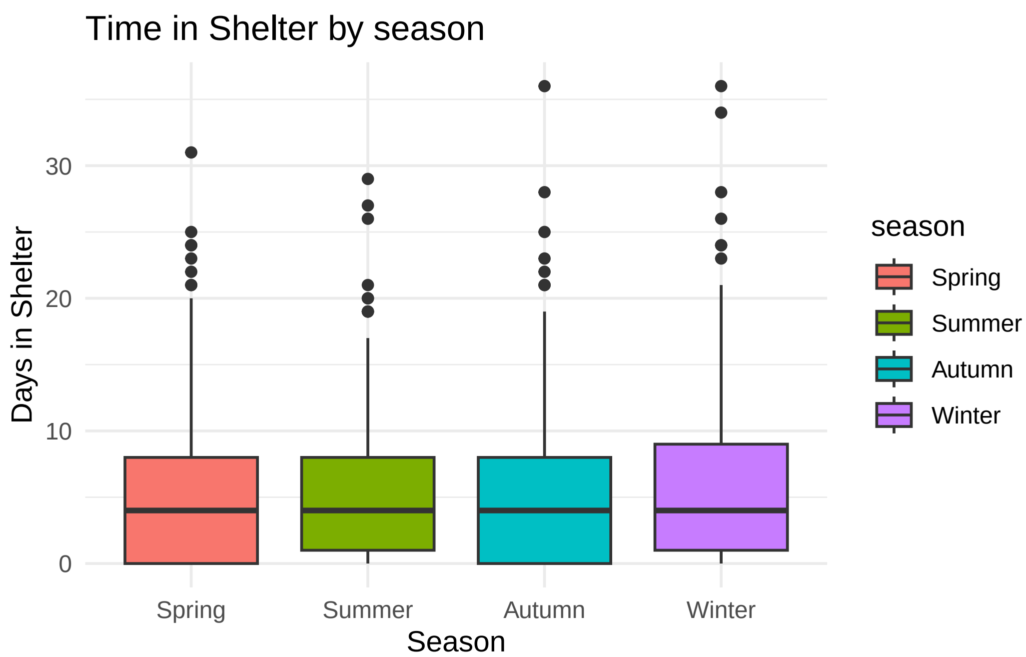


Figure 6: Time in Shelter by Season

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).

3 Modeling Framework

```r
# Read the CSV file into a dataframe
df <- read.csv("dataset20.csv")

# Ensure the 'month' column is numeric
df$month <- as.numeric(df$month)

 # Categorize months into seasons
df$season <- cut(df$month, breaks = c(2, 5, 8, 11, 12),
                 labels = c('Spring', 'Summer', 'Autumn', 'Winter'),
                 include.lowest = TRUE)

# Correct the classification of December, January, and February as Winter
df$season[df$month %in% c(12, 1, 2)] <- 'Winter'

# Convert 'season' to a factor and set the order
df$season <- factor(df$season, levels = c("Spring", "Summer", "Autumn", "Winter"))
```

Spring is the breeding season for many animals, which may lead to a surge in the number of stray animals, putting shelters under immense pressure. As resources become strained and workloads increase, staff efficiency may decline, resulting in longer stays for animals in shelters.

Summer is a peak travel season, and as people are away from home, the demand for pet adoption decreases. Additionally, the hot and humid environment increases the risk of diseases, further prolonging the stay of animals in shelters.

In winter, the holiday season, including Christmas and New Year, may lead to an adoption surge, reducing the time animals spend in shelters.

Given the impact of these seasonal factors on animal sheltering and adoption trends, we have decided to incorporate seasonal effects as an explanatory variable

```r
# Fit a negative binomial regression model
full_model <- glm.nb(time_at_shelter ~ animal_type + intake_type  + chip_status + season + ye
```

Outcome_type is an outcome variable rather than a factor influencing the length of stay in the shelter, and therefore should not be used as an explanatory variable. In this study, we use animal_type, chip_status, intake_type, season, and year as explanatory variables, while time_at_shelter serves as the outcome variable to construct a negative binomial regression model.

```r
# Perform backward stepwise selection to simplify the model
selected_model1 <- step(full_model, direction = "backward", trace = TRUE)
```

```
Start:  AIC=8356.35
time_at_shelter ~ animal_type + intake_type + chip_status + season +
    year

              Df Deviance    AIC
- season       3   1689.8 8353.1
<none>             1687.0 8356.4
- year         1   1689.1 8356.4
- animal_type  3   1695.1 8358.4
- chip_status  2   1694.5 8359.8
- intake_type  2   1718.7 8384.1

Step:  AIC=8353.09
time_at_shelter ~ animal_type + intake_type + chip_status + year

              Df Deviance    AIC
<none>             1686.8 8353.1
- year         1   1689.3 8353.6
- animal_type  3   1695.4 8355.7
- chip_status  2   1693.8 8356.1
- intake_type  2   1717.7 8380.0
```

To select the most appropriate explanatory variables, we employ a backward stepwise regression approach using the Akaike Information Criterion (AIC) as the evaluation standard. Specifically, we start with a full model that includes all candidate explanatory variables and iteratively remove variables with lower contributions to the model until we identify the optimal model with the lowest AIC. The results indicate that removing the season variable leads to a lower AIC value for the model. Therefore, we exclude the season variable to improve model fit.

```r
# Display the summary of the selected model
summary(selected_model1)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    chip_status + year, data = df, init.theta = 0.7592685117,
    link = log)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                243.35450  154.43465   1.576   0.1151
animal_typeCAT               2.72164    1.20350   2.261   0.0237 *
animal_typeDOG               2.84387    1.20218   2.366   0.0180 *
animal_typeWILDLIFE          2.54467    1.25404   2.029   0.0424 *
intake_typeOWNER SURRENDER  -0.74503    0.14350  -5.192 2.08e-07 ***
intake_typeSTRAY            -0.54367    0.13757  -3.952 7.75e-05 ***
chip_statusSCAN NO CHIP      0.01149    0.08369   0.137   0.8908
chip_statusUNABLE TO SCAN   -0.45268    0.18023  -2.512   0.0120 *
year                        -0.12089    0.07658  -1.579   0.1144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7593) family taken to be 1)

    Null deviance: 1737.8  on 1464  degrees of freedom
Residual deviance: 1686.8  on 1456  degrees of freedom
AIC: 8355.1

Number of Fisher Scoring iterations: 1

              Theta:  0.7593
          Std. Err.:  0.0346

 2 x log-likelihood:  -8335.0890
```

After removing the season variable, we reconstructed the negative binomial regression model. The results indicate that at a 95% confidence level, the p-value of the year variable is greater than 0.05, suggesting that it is not statistically significant. Additionally, based on the AIC evaluation, the model with the year variable has an AIC of 8353.1, while the model without it has an AIC of 8353.6, indicating that removing the year variable has a minimal impact on the model. Therefore, we decide to exclude the year variable from the explanatory variables.

```
# Fit a negative binomial regression model with a reduced set of predictors
selected_model2 <- glm.nb(time_at_shelter ~ animal_type + intake_type  + chip_status,  data =
```

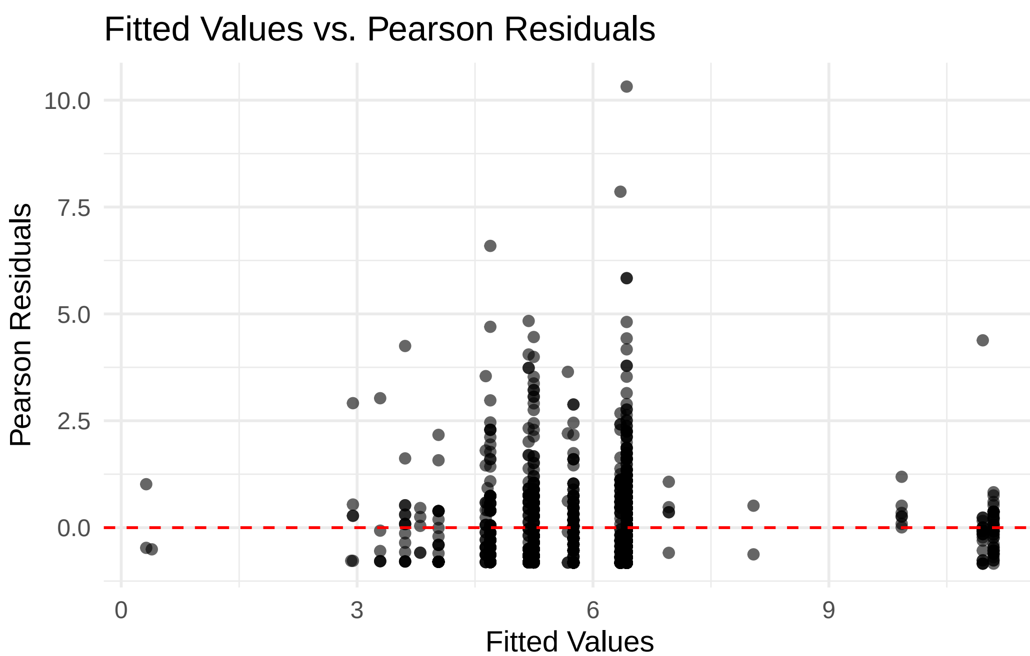
```r
# Display the summary of the model
summary(selected_model2)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    chip_status, data = df, init.theta = 0.7575663959, link = log)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.41057    1.20172  -0.342   0.7326
animal_typeCAT              2.69334    1.19365   2.256   0.0240 *
animal_typeDOG              2.80462    1.19231   2.352   0.0187 *
animal_typeWILDLIFE         2.48278    1.24462   1.995   0.0461 *
intake_typeOWNER SURRENDER -0.74886    0.14364  -5.213 1.85e-07 ***
intake_typeSTRAY           -0.54570    0.13770  -3.963 7.40e-05 ***
chip_statusSCAN NO CHIP     0.01240    0.08376   0.148   0.8823
chip_statusUNABLE TO SCAN  -0.45322    0.18040  -2.512   0.0120 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7576) family taken to be 1)

    Null deviance: 1735.1  on 1464  degrees of freedom
Residual deviance: 1686.7  on 1457  degrees of freedom
AIC: 8355.6

Number of Fisher Scoring iterations: 1

              Theta:  0.7576
          Std. Err.:  0.0345

 2 x log-likelihood:  -8337.5990
```

Finally, we select animal_type, chip_status, and intake_type as explanatory variables, with time_at_shelter as the outcome variable to construct a negative binomial regression model.

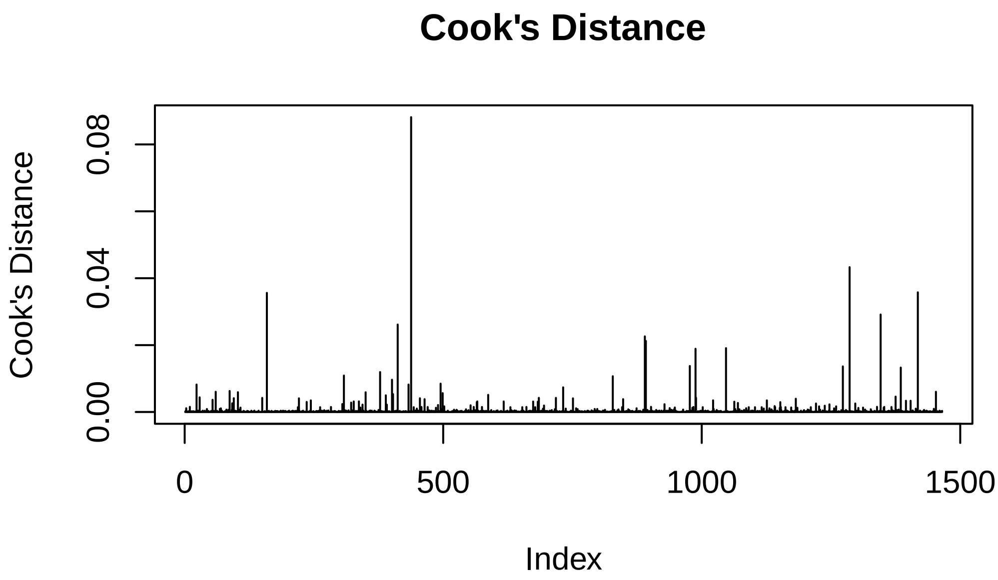
```r
# Compute Pearson residuals from the model
df$residuals <- residuals(selected_model2, type = "pearson")
# Extract fitted values from the model
df$fitted_values <- fitted(selected_model2)
```

```r
# Create scatter plot of fitted values vs. residuals
ggplot(df, aes(x = fitted_values, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Fitted Values vs. Pearson Residuals",
       x = "Fitted Values",
       y = "Pearson Residuals")
```



The plot shows some points with Pearson Residuals > 6, indicating the possible presence of outliers in the data. These outliers may have a significant impact on the model, requiring further investigation to determine whether any adjustments or modifications to the model are necessary.

```r
# Create a stem plot of Cook's Distance values
plot(cooks.distance(selected_model2), type="h",
     main="Cook's Distance", ylab="Cook's Distance")

# Add a red dashed horizontal line at Cook's Distance = 1
abline(h = 1, col = "red", lty = 2)
```



The highest Cook's Distance in the plot appears to be less than 0.1, which is far below 1, indicating that there are no particularly severe high-influence points. However, some points still exhibit a relatively large influence. Therefore, in the subsequent steps, we will identify these high-influence points and attempt to remove the outliers before refitting the model to assess their impact.

```
# Remove highly influential observations based on Cook's Distance
df_cleaned <- df[-which(cooks.distance(selected_model2) > 4 / nrow(df)), ]
# Write a cleaned dataset
write.csv(df_cleaned, 'D:\\Glasgow\\DA\\Group Assignment 2\\dataset20_cleaned.csv')

# Refit the model using cleaned data
final_model <- glm.nb(time_at_shelter ~ animal_type + intake_type  + chip_status, data = df_c

# Display summary of the final refined model
summary(final_model)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    chip_status, data = df_cleaned, init.theta = 0.9022135578,
    link = log)
```

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 2.01893    0.15799  12.778  < 2e-16 ***
animal_typeDOG              0.20473    0.07799   2.625 0.008659 ** 
animal_typeWILDLIFE         0.37388    0.57373   0.652 0.514620    
intake_typeOWNER SURRENDER -0.93785    0.13783  -6.804 1.01e-11 ***
intake_typeSTRAY           -0.57653    0.13135  -4.389 1.14e-05 ***
chip_statusSCAN NO CHIP     0.14810    0.08121   1.824 0.068190 .  
chip_statusUNABLE TO SCAN  -0.66838    0.18756  -3.564 0.000366 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9022) family taken to be 1)

    Null deviance: 1728.7  on 1404  degrees of freedom
Residual deviance: 1636.9  on 1398  degrees of freedom
AIC: 7642.4

Number of Fisher Scoring iterations: 1

              Theta:  0.9022 
          Std. Err.:  0.0454 

 2 x log-likelihood:  -7626.4060 
```

The results show that after removing the outliers, the refitted model has an AIC of 7642.4, which is 695.199 lower than the previous model's AIC, indicating an improvement in model fit.

When all categorical variables are set to their baseline categories (animal_type = cat, intake_type = confiscated, chip_status = scan chip), the model's log-predicted value is 2.01893. Thus, the estimated shelter stay for the baseline group (cats) is approximately 7.52 days.

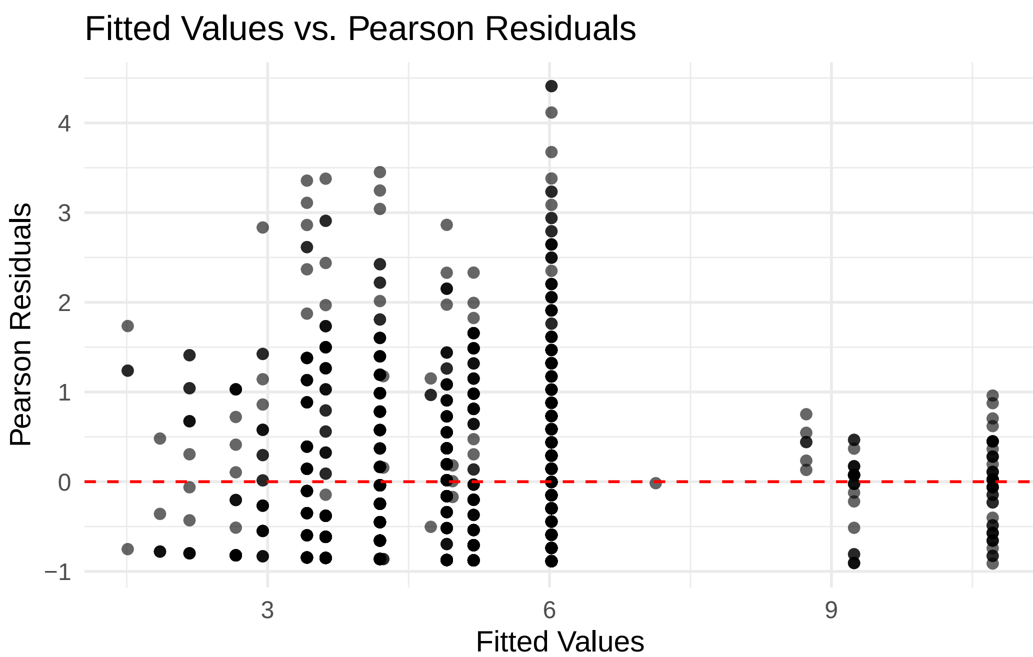
The coefficient for animal_type = DOG is 0.20473, indicating that with other variables held constant, dogs stay longer in the shelter compared to the baseline category (cats), with an estimated increase of approximately exp(0.20473)   1.23 times. Additionally, animal_type = WILDLIFE is not statistically significant at the 95% confidence level, suggesting that the shelter stay duration for wildlife does not significantly differ from that of the baseline category (cats)

intake_type = OWNER SURRENDER has a significant impact on shelter stay duration ($p < 0.001$). With other variables held constant, animals surrendered by their owners stay in the shelter for a shorter duration compared to the baseline category (confiscated), with a stay duration of exp(-0.93785) = 39% of the baseline category. intake_type = STRAY also has a significant effect on shelter stay duration ($p < 0.001$). With other variables held constant, stray animals stay in the shelter for exp(-0.57653) = 56% of the baseline category (confiscated).

chip_status = SCAN NO CHIP is not statistically significant at the 95% confidence level ($p = 0.068$), indicating that the shelter stay duration of animals without a chip does not significantly differ from that of the baseline category (SCAN CHIP). chip_status = UNABLE TO SCAN has a significant impact on shelter stay duration ($p = 0.0004$). With other variables held constant, animals whose chips cannot be scanned have a stay duration of only exp(-0.66838) = 0.51 of the baseline category (SCAN CHIP), meaning their stay is approximately 51% of the baseline category.

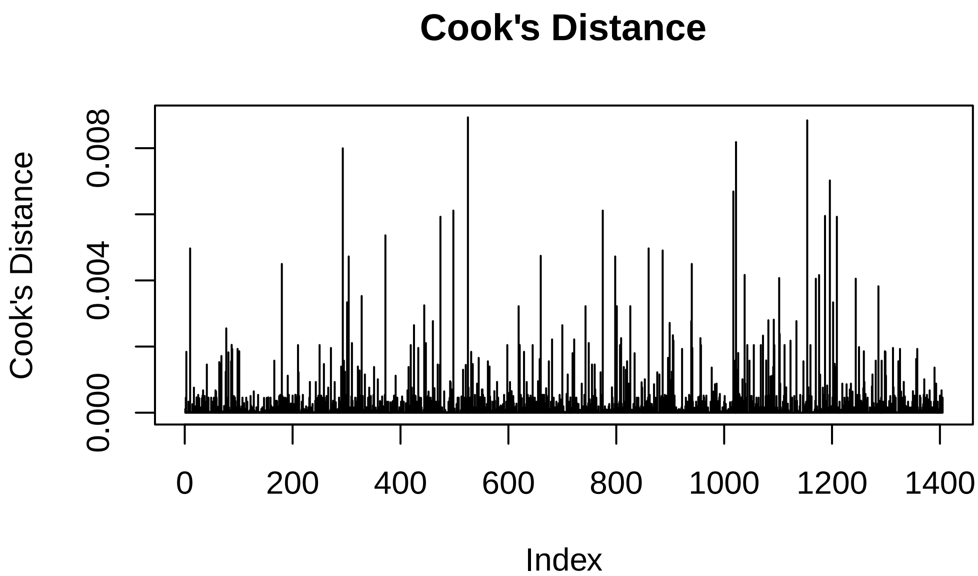
4 Assessing model fit

```
# Compute Pearson residuals from the final model
df_cleaned$residuals <- residuals(final_model, type = "pearson")

# Extract fitted values from the final model
df_cleaned$fitted_values <- fitted(final_model)

# Create scatter plot of fitted values vs. residuals
ggplot(df_cleaned, aes(x = fitted_values, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Fitted Values vs. Pearson Residuals",
       x = "Fitted Values",
       y = "Pearson Residuals")
```



The Pearson residuals are distributed around 0 without any distinct U-shaped or V-shaped patterns, indicating that the model does not suffer from severe systematic bias and has a good overall fit.

```
# Create a stem plot of Cook's Distance values
plot(cooks.distance(final_model), type="h",
     main="Cook's Distance", ylab="Cook's Distance")

# Add a red dashed horizontal line at Cook's Distance = 1
abline(h = 1, col = "red", lty = 2)
```



Cook's Distance values are generally low, indicating that no single data point has an excessively large influence on the model, ensuring stable model fitting.

```
# Compute dispersion parameter
deviance(final_model) / df.residual(final_model)
```

```
[1] 1.17089
```

The calculated value of 1.17089 is slightly greater than 1, indicating a mild degree of over dispersion in the data. However, overall, the model fits well and can still accurately describe the data distribution.

5 Conclusion

This study analyzed factors affecting the shelter stay duration of animals using data from a Dallas animal shelter. Through exploratory data analysis and statistical modeling, we found that intake type and animal type significantly influence shelter stay duration, while chip status has a limited impact.

The final negative binomial regression model included animal type, intake type, and chip status as explanatory variables. After removing outliers, the model's AIC decreased by 695.199, indicating an improved fit.

Key findings:

- Dogs stay **1.23 times longer** in the shelter than cats.
- Stray animals and owner-surrendered animals stay shorter than confiscated animals (56% and 39% of the baseline category, respectively).
- Chip status generally does not significantly impact stay duration, except for "unable to scan" cases, where animals had 51% of the baseline stay duration.

The Pearson residuals analysis showed no severe systematic bias, and Cook's Distance analysis confirmed model stability. The slightly over dispersed data ($1.17089 > 1$) suggests a good model fit.