Analysis of the number of days an animal spends in the shelter

Group 17

```
# Import packages
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(ggplot2)
library(MASS)
library(car)
library(pscl)
```

1 Introduction

Data include animal type, month, year, intake type, outcome type, chip status, time at shelter. This study is dedicated to exploring which factors influence the number of days an animal spends in the shelter before its final outcome is decided.

2 Exploratory Data Analysis

2.1 Data Conversion

```
# read data
data <- read.csv("dataset17.csv", header = TRUE)
head(data)
```

```
  animal_type month year     intake_type outcome_type  chip_status
1         DOG     6 2017 OWNER SURRENDER     ADOPTION    SCAN CHIP
2         DOG     5 2017           STRAY   EUTHANIZED SCAN NO CHIP
3         CAT     6 2017 OWNER SURRENDER     ADOPTION SCAN NO CHIP
4         CAT     4 2017           STRAY   EUTHANIZED SCAN NO CHIP
5         DOG    11 2016           STRAY     ADOPTION SCAN NO CHIP
6         CAT     1 2017           STRAY     ADOPTION SCAN NO CHIP
  time_at_shelter
1               4
2               4
3              14
4               2
5              13
6               5
```

```
animal<- data %>%
  filter(animal_type == "DOG" |  animal_type=="CAT")%>%
  droplevels()

animal <- animal%>%
  mutate(animal_type=as.factor(animal_type),
         intake_type=as.factor(intake_type),
         outcome_type=as.factor(outcome_type),
         chip_status=factor(chip_status,levels=c("UNABLE TO SCAN","SCAN CHIP","SCAN NO CHIP")
         month =as.factor(month),
         year=as.factor(year))

# check the factors
levels(animal$animal_type)
```

```
[1] "CAT" "DOG"
```

```
levels(animal$intake_type)
```

```
[1] "CONFISCATED"     "OWNER SURRENDER" "STRAY"
```

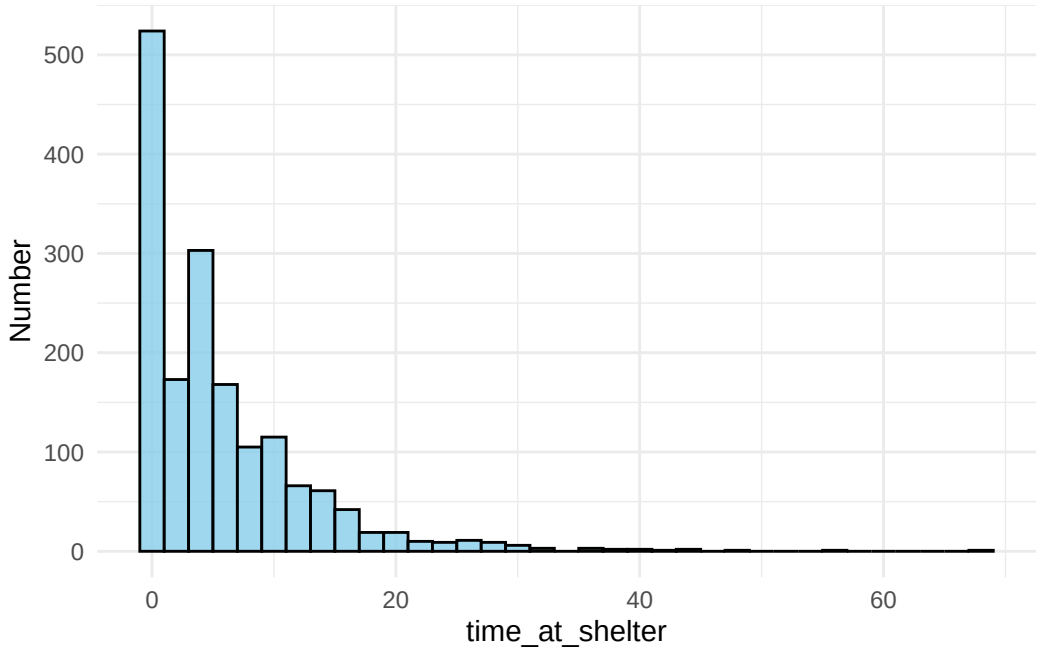
```
levels(animal$outcome_type)
```

```
[1] "ADOPTION"          "DIED"              "EUTHANIZED"
[4] "FOSTER"            "RETURNED TO OWNER"
```

```r
levels(animal$chip_status)
```

```
[1] "UNABLE TO SCAN" "SCAN CHIP"      "SCAN NO CHIP"
```

```r
#plot histogram
#| fig-cap: Histogram of the number of days an animal spends in the shelter
#| label: fig-1
#| fig-width:  8
#| fig-height: 6
#| fig-align: center
ggplot(animal, aes(x = time_at_shelter)) +
  geom_histogram(binwidth = 2, fill = "skyblue", color = "black", alpha = 0.8) +
  labs( x = "time_at_shelter", y = "Number") +
  theme_minimal()
```




```r
sum(animal$time_at_shelter == 0)
```

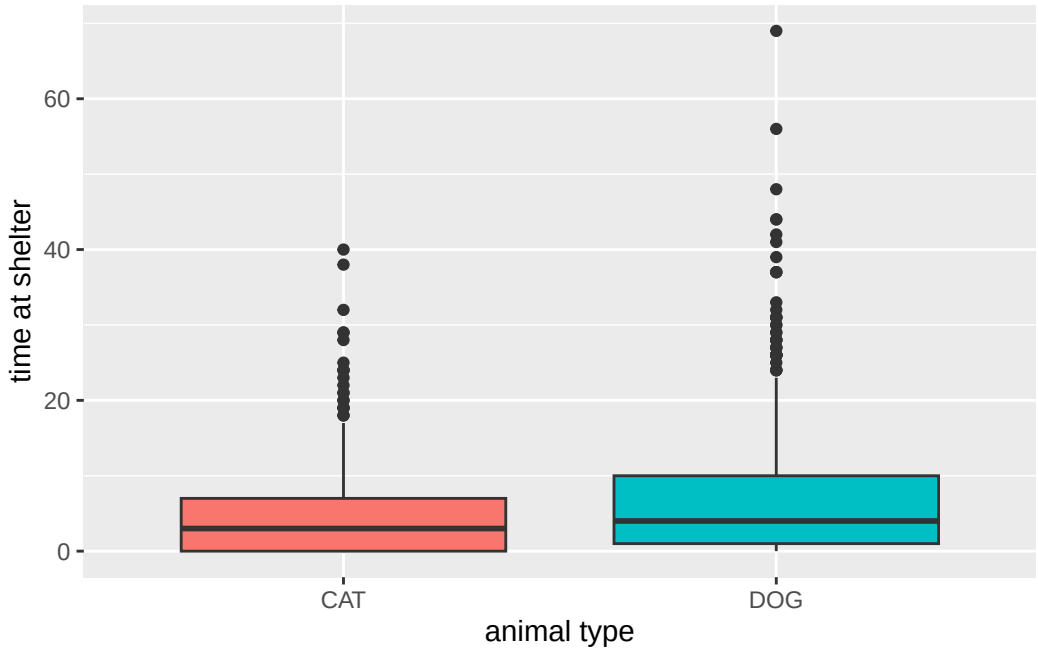
```
[1] 374
```

```r
mean(animal$time_at_shelter == 0)
```

```
[1] 0.2258454
```

The histogram presents a right-skewed distribution. Most animals stay in shelters for a relatively short period of time, concentrated in the range of 0-10 days, and very few animals stay for more than 30 days or even longer.And 374 observations (22.58%) have a `time_at_shelter` value of 0, suggesting that it should be carefully considered when fitting the model.

```r
# boxplot
#| fig-cap: Boxplot of time_at_shelter and animal_type
#| label: fig-2
#| fig-width:  8
#| fig-height: 6
#| fig-align: center
ggplot(data = animal, aes(x = animal_type, y = time_at_shelter, fill = animal_type)) +
  geom_boxplot() +
  labs(x = "animal type", y = "time at shelter")+
  theme(legend.position = "none")
```



The median of cats is smaller than that of dogs, but dogs have a wider distribution. The upper quartile range for dogs is larger, meaning that some dogs stay in shelters for longer periods of time.

```
#boxplot_2
ggplot(animal, aes(x = intake_type, y = time_at_shelter,fill = intake_type)) +
  geom_boxplot() +
  theme_minimal()
```

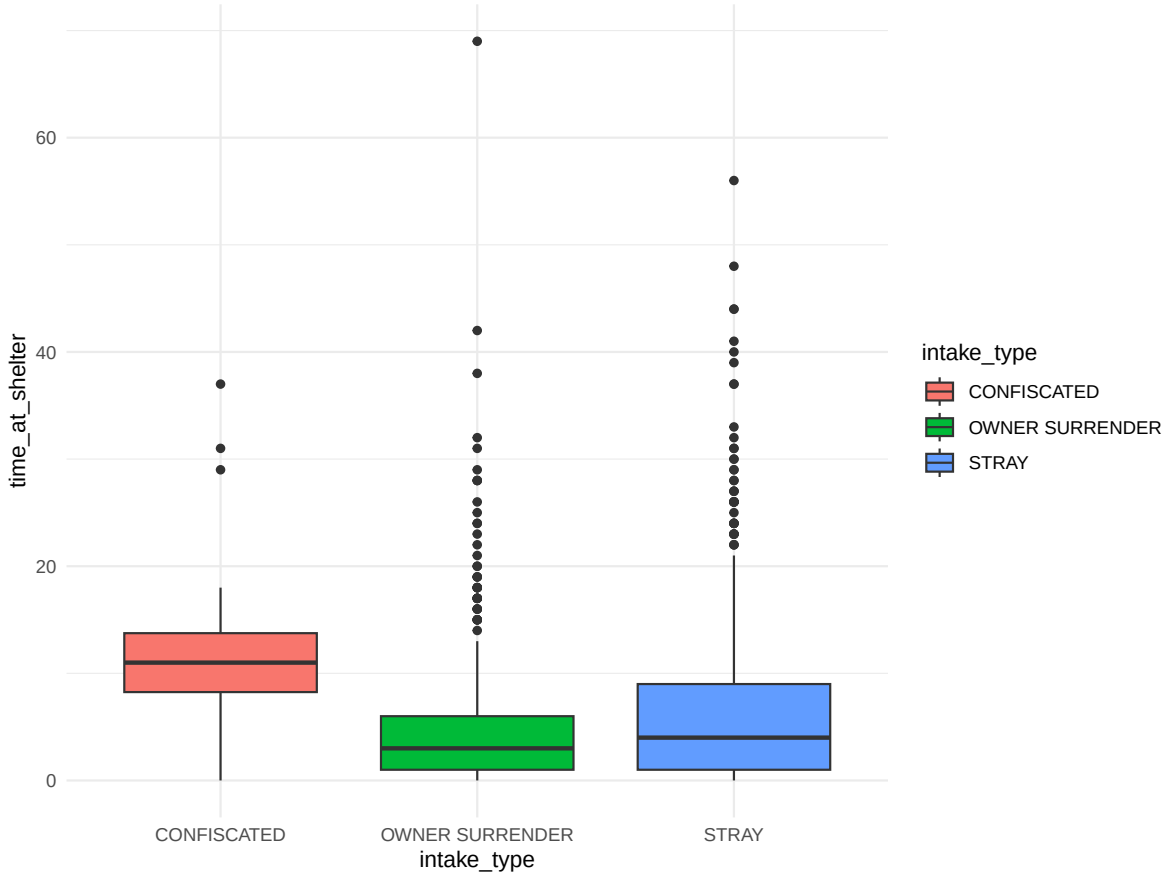


Figure 1: Boxplot of time_at_shelter and intake_type

The median of confiscated animals was significantly higher than that of the other two groups, indicating that confiscated animals stayed concentrated and longer.

```
#boxplot_3
ggplot(animal, aes(x = chip_status, y = time_at_shelter,fill = chip_status)) +
  geom_boxplot() +
  theme_minimal()
```

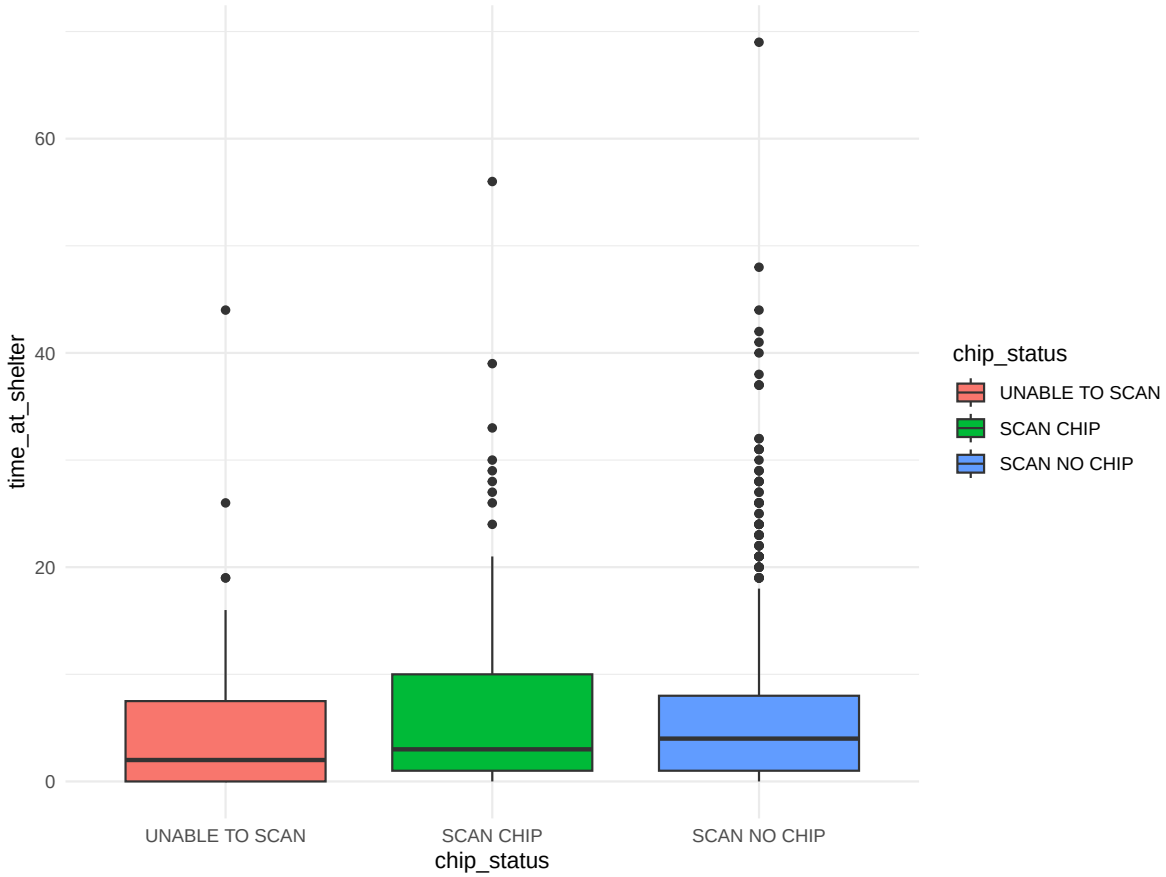
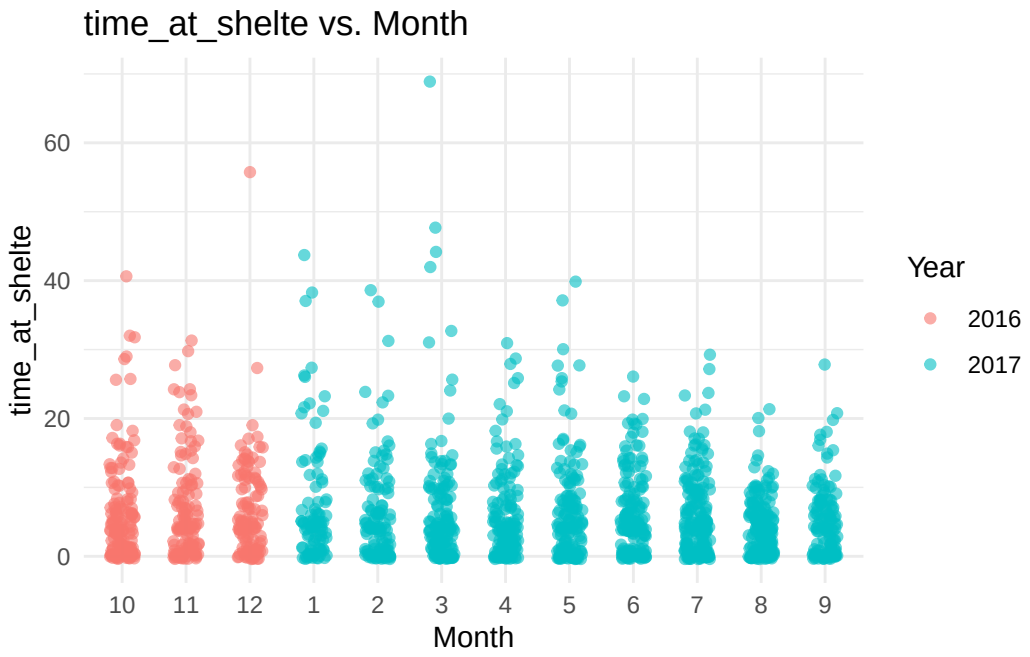


Figure 2: Boxplot of time_at_shelter and chip_status

The medians of the three sets of data were relatively close, indicating that animals with different chip statuses stayed in the shelter for roughly similar lengths of time. The IQR of the animals scanned to the chip was wider, indicating that the data distribution in this group was more dispersed.

```
# Scatter Plot
#| fig-cap: Scatter plot of time_at_shelter and intake_type
#| label: fig-5
#| fig-width:  8
#| fig-height: 6
#| fig-align: center
ggplot(animal, aes(x = factor(month, levels = c(10:12, 1:9)), y = time_at_shelter, color = fa
  geom_jitter(alpha = 0.6, width = 0.2) +
  labs(title = "time_at_shelte vs. Month", x = "Month", y = "time_at_shelte", color = "Year")
  theme_minimal()
```



The time range of the dataset is from October to December 2016 and January to September 2017. The monthly data is complete so we consider it as an explanatory variable. The length of time animals spent in shelters did not show significant seasonal variations across years, with most animals staying short (less than 20 days).

3 Formal Analysis

3.1 Method1 : Poisson regression

Poisson regression is one of the most commonly used generalized linear models when analyzing numerical data, so the study starts with poisson regression.

```
# set up "month" in order
animal <- animal %>%
  mutate(month = as.numeric(month),
         month_ordered = ifelse(month >= 10, month - 9, month + 3))


# Poisson regression
glm_model <- glm(time_at_shelter ~ animal_type +  month_ordered + intake_type  + chip_status
                 data = animal,
                 family = poisson(link = "log"))
```

```r
summary(glm_model)
```

```
Call:
glm(formula = time_at_shelter ~ animal_type + month_ordered +
    intake_type + chip_status, family = poisson(link = "log"),
    data = animal)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.148086   0.066150  32.473  < 2e-16 ***
animal_typeDOG              0.156898   0.026786   5.857 4.70e-09 ***
month_ordered              -0.033369   0.002898 -11.515  < 2e-16 ***
intake_typeOWNER SURRENDER -0.837684   0.039635 -21.135  < 2e-16 ***
intake_typeSTRAY           -0.555981   0.036360 -15.291  < 2e-16 ***
chip_statusSCAN CHIP        0.295779   0.059474   4.973 6.58e-07 ***
chip_statusSCAN NO CHIP     0.381497   0.055719   6.847 7.55e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11749  on 1655  degrees of freedom
Residual deviance: 11093  on 1649  degrees of freedom
AIC: 15692

Number of Fisher Scoring iterations: 6
```

```r
# overdivergence of Poisson regression
dispersion <- sum(residuals(glm_model, type = "pearson")^2) / glm_model$df.residual
dispersion
```

```
[1] 8.007112
```

The dispersion parameter (8.007) is much larger than 1, which means our GLM may have overdispersion. Therefore, we consider negative binomial regression to improve the model's fit.

The Negative Binomial distribution extends the Poisson model by allowing for overdispersion,which means the variance is significantly greater than the mean.

3.2 Method2 : Negative binomial regression model

Negative binomial distribution is used when the data is too discrete

```
# Negative binomial regression model 1
nb_model_1 <- glm.nb(time_at_shelter ~ animal_type + intake_type  +chip_status + month_ordere
summary(nb_model_1)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    chip_status + month_ordered, data = animal, init.theta = 0.7903189393,
    link = log)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.154612   0.200866  10.727  < 2e-16 ***
animal_typeDOG               0.122268   0.074224   1.647  0.09950 .
intake_typeOWNER SURRENDER -0.844662   0.141015  -5.990 2.10e-09 ***
intake_typeSTRAY           -0.580413   0.135640  -4.279 1.88e-05 ***
chip_statusSCAN CHIP         0.343540   0.163576   2.100  0.03571 *
chip_statusSCAN NO CHIP      0.426423   0.151582   2.813  0.00491 **
month_ordered               -0.034058   0.008612  -3.955 7.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7903) family taken to be 1)

    Null deviance: 1985.2  on 1655  degrees of freedom
Residual deviance: 1913.6  on 1649  degrees of freedom
AIC: 9440

Number of Fisher Scoring iterations: 1

              Theta:  0.7903
          Std. Err.:  0.0342

 2 x log-likelihood:  -9423.9930
```

```
# Fit the full model (including all explanatory variables)
full_model <- glm.nb(time_at_shelter ~ animal_type + intake_type  +chip_status + month_ordere
```

```r
# Fit an empty model (only intercept terms)
null_model <- glm.nb(time_at_shelter ~ 1, data = animal)

# Model selection using stepwise regression (two-way selection), based on AIC index
selected_model <- stepAIC(null_model,
                          scope = list(lower = null_model, upper = full_model),
                          direction = "both")
```

```
Start:  AIC=9495.81
time_at_shelter ~ 1

                Df    AIC
+ intake_type    2 9454.9
+ month_ordered  1 9483.2
+ animal_type    1 9488.1
<none>             9495.8
+ chip_status    2 9497.0

Step:  AIC=9454.85
time_at_shelter ~ intake_type

                Df    AIC
+ month_ordered  1 9442.8
+ chip_status    2 9452.3
+ animal_type    1 9453.3
<none>             9454.9
- intake_type    2 9495.8

Step:  AIC=9442.76
time_at_shelter ~ intake_type + month_ordered

                Df    AIC
+ chip_status    2 9438.6
+ animal_type    1 9441.9
<none>             9442.8
- month_ordered  1 9454.9
- intake_type    2 9483.2

Step:  AIC=9438.55
time_at_shelter ~ intake_type + month_ordered + chip_status

                Df    AIC
```

```
+ animal_type    1 9438.0
<none>              9438.6
- chip_status    2 9442.8
- month_ordered  1 9452.3
- intake_type    2 9483.2

Step:  AIC=9437.99
time_at_shelter ~ intake_type + month_ordered + chip_status +
    animal_type

                Df    AIC
<none>              9438.0
- animal_type    1 9438.6
- chip_status    2 9441.9
- month_ordered  1 9451.0
- intake_type    2 9476.4
```

```
summary(selected_model)
```

```
Call:
glm.nb(formula = time_at_shelter ~ intake_type + month_ordered +
    chip_status + animal_type, data = animal, init.theta = 0.7903189298,
    link = log)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.154611   0.200866  10.727  < 2e-16 ***
intake_typeOWNER SURRENDER -0.844661   0.141015  -5.990 2.10e-09 ***
intake_typeSTRAY           -0.580414   0.135640  -4.279 1.88e-05 ***
month_ordered              -0.034058   0.008612  -3.955 7.66e-05 ***
chip_statusSCAN CHIP        0.343540   0.163576   2.100  0.03571 *
chip_statusSCAN NO CHIP     0.426424   0.151582   2.813  0.00491 **
animal_typeDOG              0.122270   0.074224   1.647  0.09949 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7903) family taken to be 1)

    Null deviance: 1985.2  on 1655  degrees of freedom
Residual deviance: 1913.6  on 1649  degrees of freedom
AIC: 9440
```

```
Number of Fisher Scoring iterations: 1

              Theta:  0.7903
          Std. Err.:  0.0342

 2 x log-likelihood:  -9423.9930
```

```r
# Negative binomial regression model 2
nb_model_2 <- glm.nb(time_at_shelter ~  intake_type  +chip_status + month_ordered ,data = ani

summary(nb_model_2)
```

```
Call:
glm.nb(formula = time_at_shelter ~ intake_type + chip_status +
    month_ordered, data = animal, init.theta = 0.7885088453,
    link = log)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.254340   0.191485  11.773  < 2e-16 ***
intake_typeOWNER SURRENDER -0.879156   0.140286  -6.267 3.68e-10 ***
intake_typeSTRAY           -0.591778   0.135576  -4.365 1.27e-05 ***
chip_statusSCAN CHIP        0.379564   0.162744   2.332  0.01969 *
chip_statusSCAN NO CHIP     0.446105   0.151474   2.945  0.00323 **
month_ordered              -0.034902   0.008606  -4.055 5.00e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7885) family taken to be 1)

    Null deviance: 1982.0  on 1655  degrees of freedom
Residual deviance: 1913.1  on 1650  degrees of freedom
AIC: 9440.6

Number of Fisher Scoring iterations: 1

              Theta:  0.7885
          Std. Err.:  0.0340

 2 x log-likelihood:  -9426.5530
```

We applied stepwise model selection using the `stepAIC()` function, starting from a null model and considering both forward and backward selection. The algorithm selected a model including `intake_type`, `month_ordered`, `chip_status`, and `animal_type`, with the lowest AIC (9438.0). Although `animal_type` was only marginally significant (p = 0.099), it contributed to lowering the AIC.

To evaluate model simplicity , we constructed a second Negative Binomial model (`nb_model_2`) excluding `animal_type`. This simplified model had an AIC of 9440.6,only slightly higher than the previous one,suggesting that `animal_type` may be safely excluded for interpretation purposes. Given the marginal significance and the small AIC increase, we decided to focus on the simpler model in our conclusions.

3.3 Method3 : Zero-inflated model

Given the large number of zero counts in the outcome variable, we fitted a Zero-Inflated Negative Binomial (ZINB) model to account for excess zeros.This model combines a count process (Negative Binomial regression) and a zero-inflation process (logistic regression).

```
# Zero-inflated model

zeroflated_model <- zeroinfl(time_at_shelter ~ animal_type + intake_type + chip_status + mont

summary(zeroflated_model)
```

```
Call:
zeroinfl(formula = time_at_shelter ~ animal_type + intake_type + chip_status +
    month_ordered, data = animal, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.1220 -0.7552 -0.2852  0.3403 11.7376

Count model coefficients (negbin with log link):
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.513082   0.178590  14.072  < 2e-16 ***
animal_typeDOG              0.030269   0.069855   0.433 0.664792
intake_typeOWNER SURRENDER -0.657436   0.113270  -5.804 6.47e-09 ***
intake_typeSTRAY           -0.345892   0.105535  -3.278 0.001047 **
chip_statusSCAN CHIP       -0.012539   0.161918  -0.077 0.938271
chip_statusSCAN NO CHIP     0.033998   0.152607   0.223 0.823704
month_ordered              -0.026262   0.007719  -3.403 0.000668 ***
```

```
Log(theta)                  0.359461   0.065660   5.475 4.39e-08 ***

Zero-inflation model coefficients (binomial with logit link):
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -12.65571  148.57876  -0.085 0.932120
animal_typeDOG             -0.49317    0.20479  -2.408 0.016033 *
intake_typeOWNER SURRENDER 12.44402  148.57884   0.084 0.933252
intake_typeSTRAY           12.66444  148.57872   0.085 0.932073
chip_statusSCAN CHIP       -1.36655    0.36766  -3.717 0.000202 ***
chip_statusSCAN NO CHIP    -1.59160    0.31069  -5.123 3.01e-07 ***
month_ordered               0.04147    0.02570   1.613 0.106642
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.4326
Number of iterations in BFGS optimization: 32
Log-likelihood: -4653 on 15 Df
```

```r
# AIC, BIC and mcfadden R²

aic_value <- AIC(zeroflated_model)
bic_value <- BIC(zeroflated_model)
print(aic_value)
```

```
[1] 9335.147
```

```r
print(bic_value)
```

```
[1] 9416.329
```

```r
null_model <- update(zeroflated_model, . ~ 1)
mcfadden_r2 <- 1 - as.numeric(logLik(zeroflated_model)) / as.numeric(logLik(null_model))
print(mcfadden_r2)
```

```
[1] 0.01301562
```

```r
# Comment:
# AIC is 9335 and BIC is 9416 , both of them are lower than the model above. Mcfadden R² is (
```

The count part predicts how long an animal stays, given it was truly admitted, while the zero-inflation part models the likelihood that an observation is a structural zero. Results suggest that intake_type and month_ordered are associated with actual time spent, whereas animal_type and chip_status influence the likelihood of being a structural zero.

3.4 Model comparison

```
# AIC
aic_values <- AIC(glm_model, nb_model_2,zeroflated_model)

#df
df_values <- c(glm_model$df.residual, nb_model_2$df.residual, zeroflated_model$df.residual)

# merge
model_comparison <- data.frame(
  Model = c("GLM (Poisson)", "NB2 (Negative Binomial)", "ZERO (Zero-inflated model)"),
  DF = df_values,
  AIC = aic_values$AIC
)

print(model_comparison)
```

```
                       Model   DF       AIC
1               GLM (Poisson) 1649 15692.180
2    NB2 (Negative Binomial) 1650  9440.553
3 ZERO (Zero-inflated model) 1641  9335.147
```

By comparing AIC, the negative binomial regression model performs better.

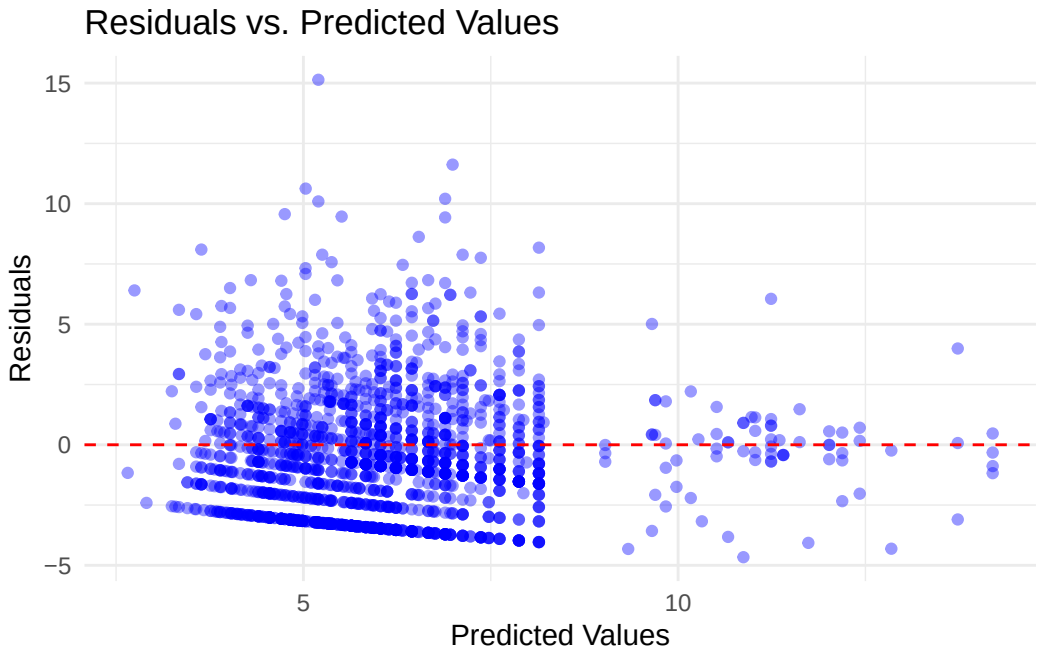
3.5 Residual Plot

3.5.1 Model1 : Possion

```
animal$residuals <- residuals(glm_model, type = "deviance")
predicted_values <- predict(glm_model, type = "response")

ggplot(animal, aes(x = predicted_values, y = residuals)) +
  geom_point(alpha = 0.4, color = "blue") +
```

```
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Predicted Values", x = "Predicted Values", y = "Residuals") +
  theme_minimal()
```

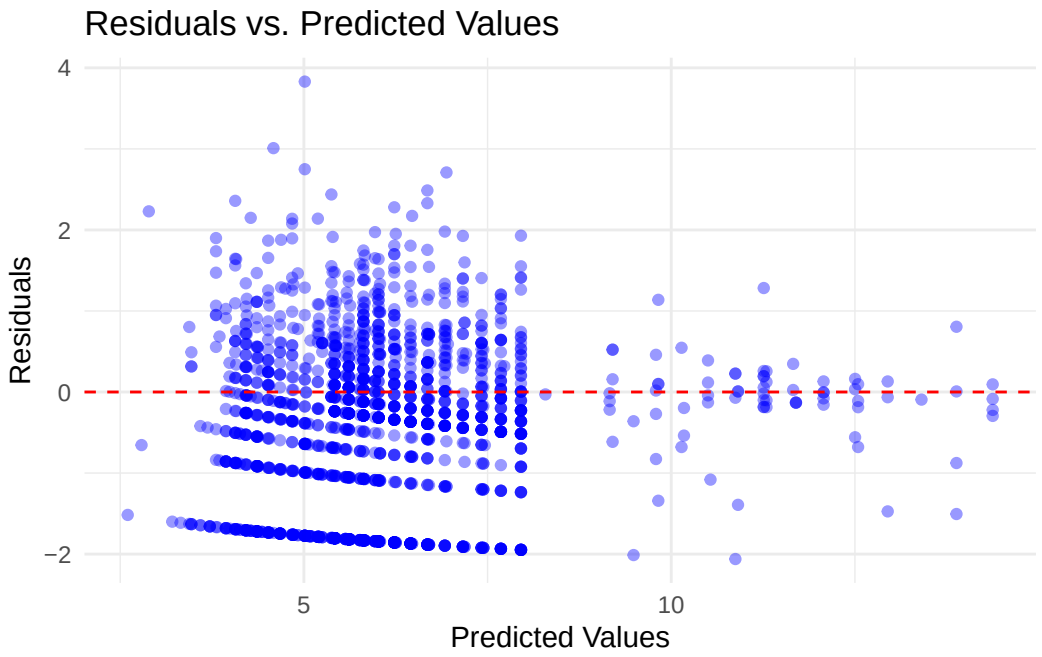


The variability of the residuals is larger at smaller predicted values, which indicates excessive dispersion, i.e., the variance is larger than the mean of Poisson's assumptions. This is consistent with the dispersion parameter we calculated earlier.

3.5.2 Model2 : nb_model_2

```
animal$residuals <- residuals(nb_model_2, type = "deviance")
predicted_values <- predict(nb_model_2, type = "response")

ggplot(animal, aes(x = predicted_values, y = residuals)) +
  geom_point(alpha = 0.4, color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Predicted Values", x = "Predicted Values", y = "Residuals") +
  theme_minimal()
```

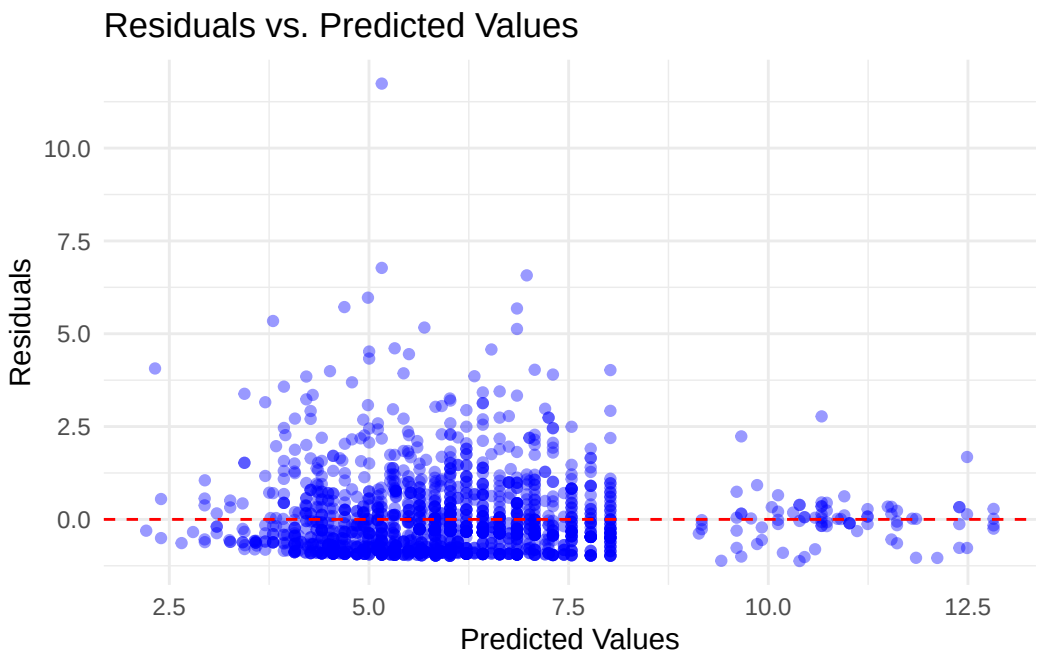


The variance of the residuals is generally in the range of -2 to +2. The negative binomial regression model basically solves the overdispersion problem.

3.5.3 Model3 : Zero-flated_model

```
res <- residuals(zeroflated_model, type = "pearson")
fitted_vals <- fitted(zeroflated_model)

ggplot(animal, aes(x = fitted_vals, y = res)) +
  geom_point(alpha = 0.4, color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Predicted Values", x = "Predicted Values", y = "Residuals") +
  theme_minimal()
```



4 Summary

This report analyzes the number of days animals stay in shelters. The data show a right-skewed distribution, with most animals staying fewer than 10 days and many recorded as 0 days.

To explore influencing factors, we applied Poisson and Negative Binomial regression models using predictors such as `intake_type`, `animal_type`, `chip_status`, and `month`. Results show that animals surrendered by owners or found as strays tend to stay for fewer days. Dogs tend to stay slightly longer, while chip status and intake month showed weaker effects.

Due to the large number of zeros, we also fitted a Zero-Inflated Negative Binomial model. This model achieved the best fit and helped distinguish structural zeros from regular counts, offering deeper insight into the shelter process.

5 Future work

Considering the large number of 0 values, the presence of multiple explanatory variables in the data, and the fact that glm's fitting is not very good, we can try to use ANN to predict the time of animals in shelters.

How to transform non-numerical explanatory variables into numerical explanatory variables is still in the process of research, or will the frequency of observer occurrence be considered as the conversion criterion.