Group Project 2_Group16

Group16 Zeyu Bai, Annika White, Hongze Shi, Wei Jin, Yijie Wu

```r
library(dplyr)
library(MASS)
```

1 Data Processing

1.1 Data Preprocessing

First, A new column 'season' is created based on the 'month' column, with the months categorized into seasons. For 'chip_status', we combined 'UNABLE TO SCAN' and 'SCAN NO CHIP' into a single category labelled 'No Chip'. Because the number of 'BIRD' and 'WILDLIFE' categories is much smaller compared to the other two, we filtered out 'BIRD' and 'WILDLIFE' and only studied 'CAT' and 'DOG'. The last line indicates that in the original dataset, the months are represented from 1 to 12, and the years are 2016 and 2017. Therefore, if the year is 2016, we use the month directly to represent time, but if the year is 2017, we represent time by adding 12 to the month.

```r
Animal <- read.csv("dataset16.csv")

Animal$year <- as.factor(Animal$year)
Animal$season <- ifelse(Animal$month %in% c(12, 1, 2), "Winter",
 ↪  Animal$month)
Animal$season <- ifelse(Animal$season %in% c(3, 4, 5), "Spring",
 ↪  Animal$season)
Animal$season <- ifelse(Animal$season %in% c(6, 7, 8), "Summer",
 ↪  Animal$season)
Animal$season <- ifelse(Animal$season %in% c(9, 10, 11), "Fall",
 ↪  Animal$season)
```

```
Animal$chip_status <- ifelse(Animal$chip_status == "SCAN CHIP", "Chip",
 ↪   "No Chip")
Animal <- subset(Animal, animal_type %in% c("CAT", "DOG"))
Animal$time <- Animal$month
Animal$time <- ifelse(Animal$year == 2017, Animal$month + 12,
 ↪   Animal$month)
```

2 EDA

2.0.0.1 Numerical summary

```
apply(Animal, 2, table)
```

```
$animal_type

 CAT  DOG 
 270 1163 

$month

  1   2   3   4   5   6   7   8   9  10  11  12 
 97  81 103 114 138 163 162 125 113 122 107 108 

$year

2016 2017 
 337 1096 

$intake_type

    CONFISCATED OWNER SURRENDER           STRAY 
             75             460             898 

$outcome_type

         ADOPTION              DIED        EUTHANIZED            FOSTER 
              627                24               482                29 
RETURNED TO OWNER 
              271 
```

```
$chip_status

   Chip No Chip
    285    1148

$time_at_shelter

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
311 143  86  57 174 105  73  62  63  37  62  52  32  19  25  15  14  12   9   8
 20  21  22  23  24  25  26  27  28  29  30  31  32  33  37  39  40  41  42  43
  7  13   5   3   2   9   1   2   5   4   2   4   1   2   1   2   2   1   1   1
 50  53  59  63  66
  2   1   1   1   1

$season

  Fall Spring Summer Winter
   342    355    450    286

$time

 10  11  12  13  14  15  16  17  18  19  20  21
122 107 108  97  81 103 114 138 163 162 125 113
```

```
library(skimr)
skim(Animal)
```

Table 1: Data summary

Name	Animal
Number of rows	1433
Number of columns	9
Column type frequency:	
character	5
factor	1
numeric	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
animal_type	0	1	3	3	0	2	0
intake_type	0	1	5	15	0	3	0
outcome_type	0	1	4	17	0	5	0
chip_status	0	1	4	7	0	2	0
season	0	1	4	6	0	4	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
year	0	1	FALSE	2	201: 1096, 201: 337

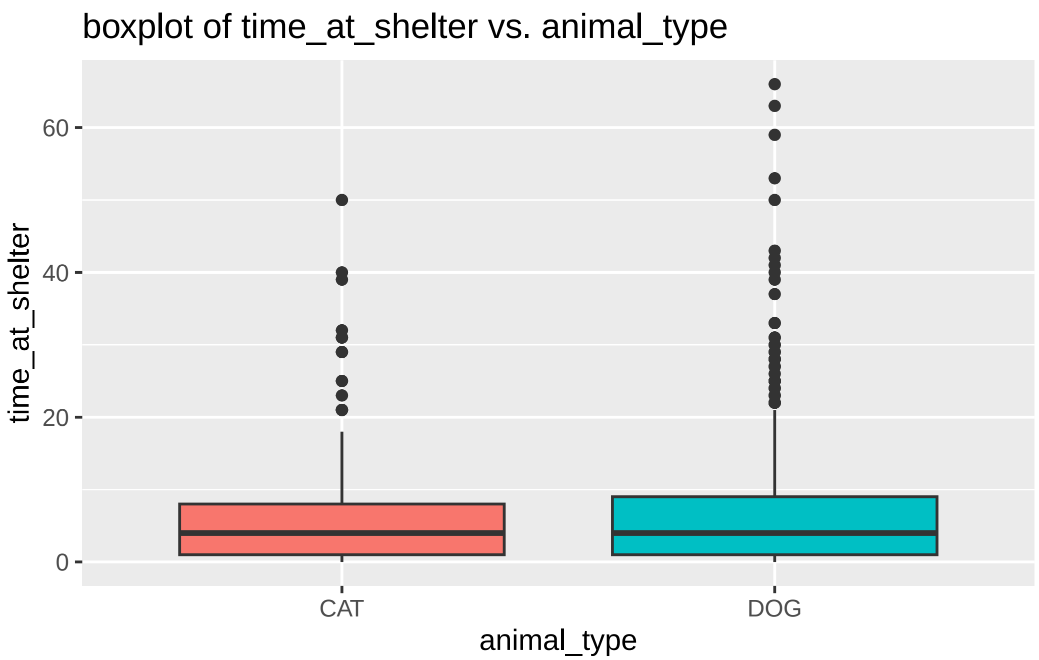
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
month	0	1	6.65	3.22	1	4	7	9	12	
time_at_shelter	0	1	6.07	7.37	0	1	4	9	66	
time	0	1	15.83	3.46	10	13	16	19	21	

2.0.0.2 Graphical summary

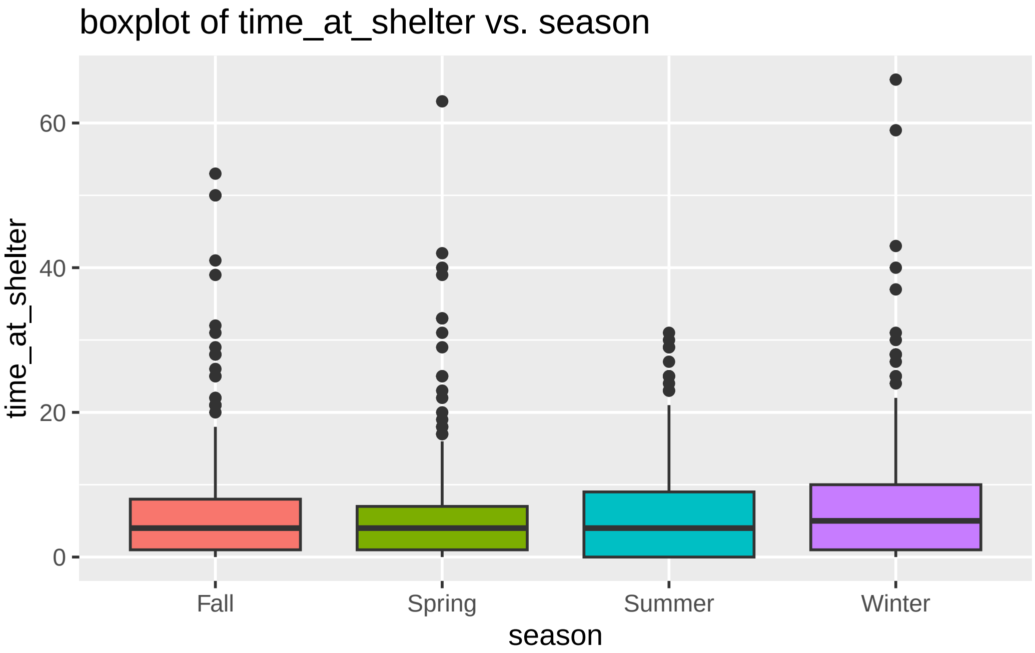
There are six boxplots for each explanatory variables, and two bar charts to show the relationship between the type of animal(cat or dog), the circumstances of their arrival at the shelter (intake_type), and their subsequent outcomes (outcome_type).

```
library(ggplot2)
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")+
  ggtitle("boxplot of time_at_shelter vs. animal_type")
```



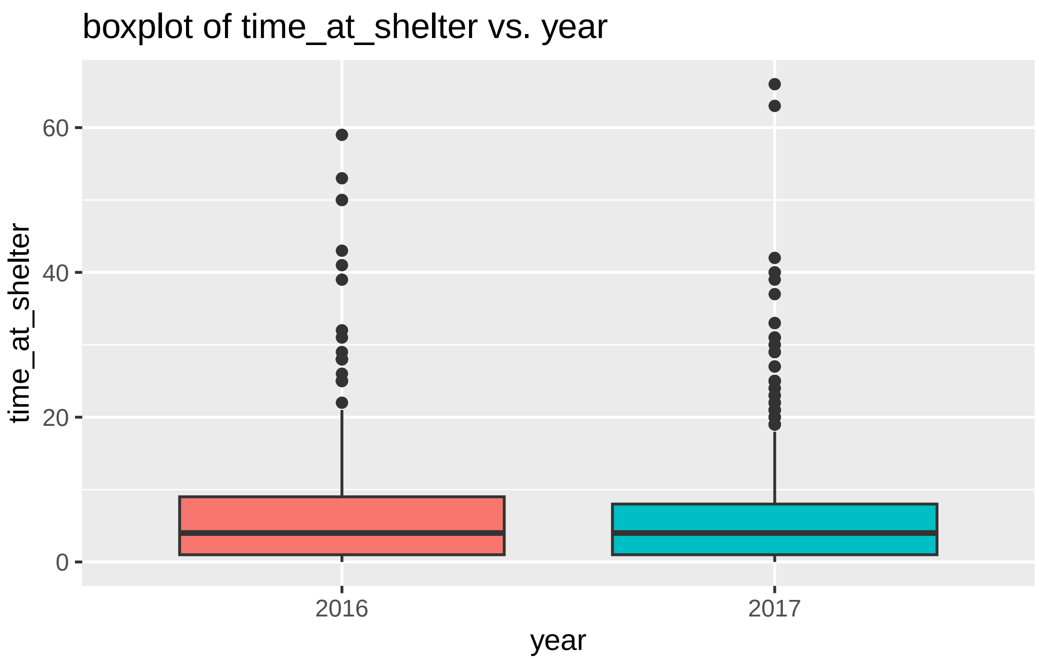
Dogs tend to have a wider range and potentially a higher median time at the shelter compared to cats. There are also more outliers for dogs, indicating that some stay significantly longer than the median.

```
ggplot(data = Animal, aes(x = season, y = time_at_shelter, fill =
↪  season))+
  geom_boxplot()+
  labs(x = "season", y = "time_at_shelter")+
  theme(legend.position = "none")+
  ggtitle("boxplot of time_at_shelter vs. season")
```



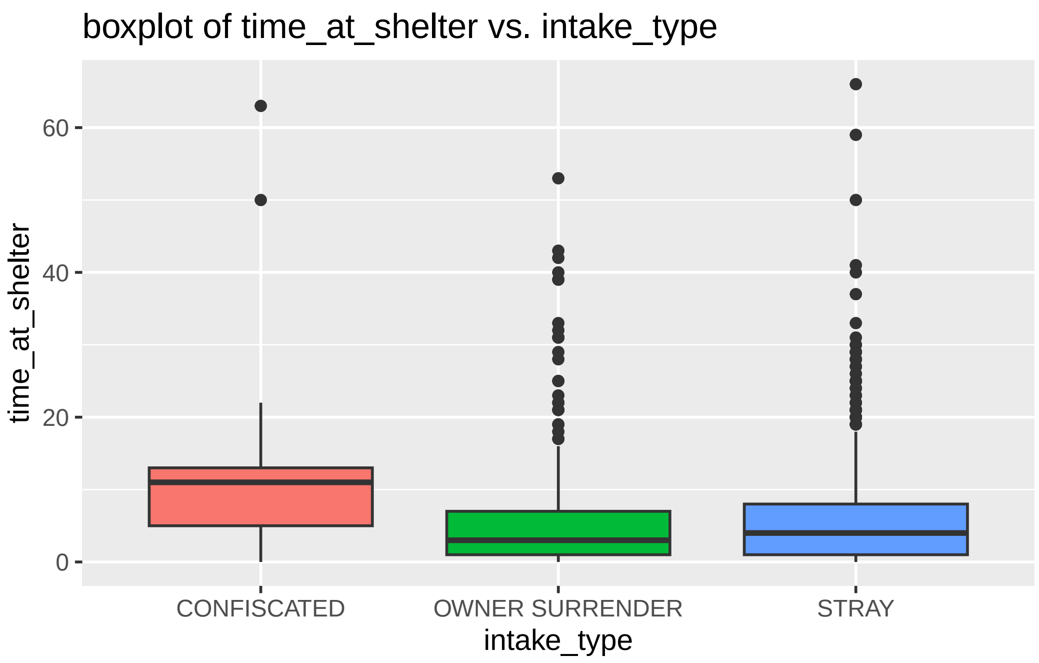
The boxplot shows varied medians across different outcomes. Adoption has a lower median time at the shelter compared to animals that were fostered or returned to the owner. Euthanized animals have a wide interquartile range, suggesting variable time frames before this outcome is reached.

```
ggplot(data = Animal, aes(x = year, y = time_at_shelter, fill = year))+
  geom_boxplot()+
  labs(x = "year", y = "time_at_shelter")+
  theme(legend.position = "none")+
  ggtitle("boxplot of time_at_shelter vs. year")
```



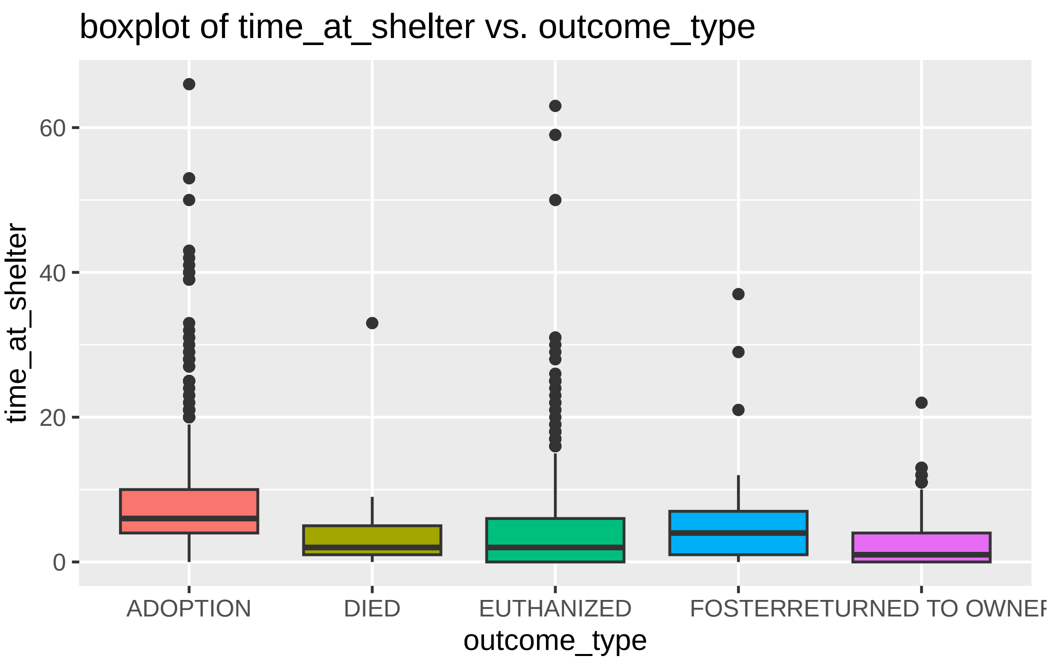
The boxplot shows a lower median for the year 2016 compared to 2017, with fewer extreme outliers in 2016. This could imply a change in the shelter's operations or external factors affecting the length of stay.

```
ggplot(data = Animal, aes(x = intake_type, y = time_at_shelter, fill =
↪  intake_type))+
  geom_boxplot()+
  labs(x = "intake_type", y = "time_at_shelter")+
  theme(legend.position = "none")+
  ggtitle("boxplot of time_at_shelter vs. intake_type")
```



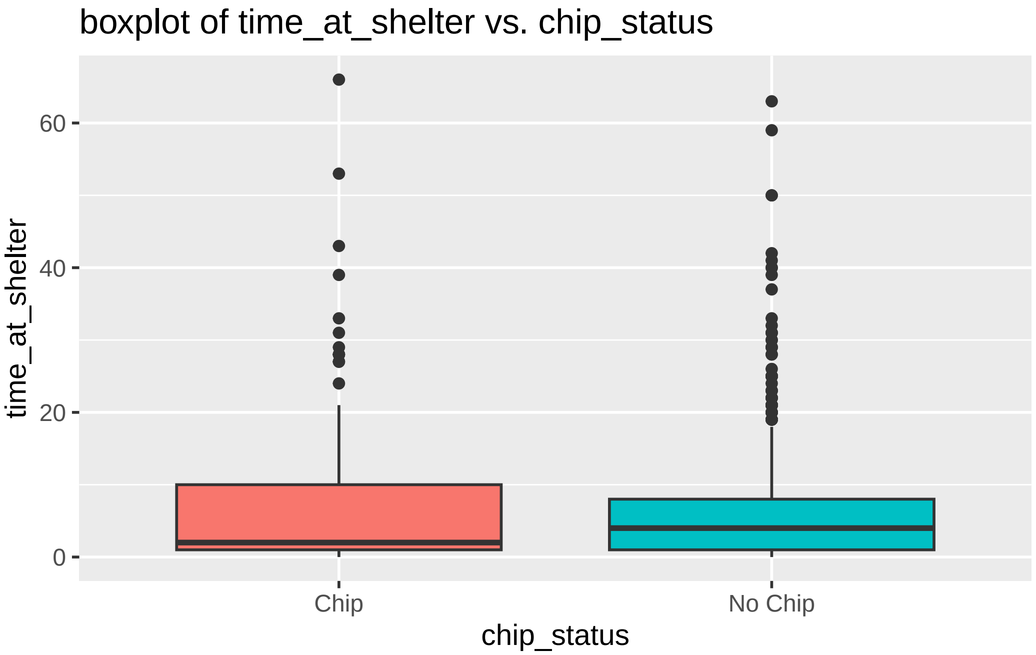
Stray animals show a higher median and wider interquartile range of time spent at the shelter. Confiscated animals have the shortest stay, while owner surrenders are in between but closer to confiscated in terms of the median time at the shelter.

```
ggplot(data = Animal, aes(x = outcome_type, y = time_at_shelter, fill =
 ↪  outcome_type))+
  geom_boxplot()+
  labs(x = "outcome_type", y = "time_at_shelter")+
  theme(legend.position = "none")+
  ggtitle("boxplot of time_at_shelter vs. outcome_type")
```



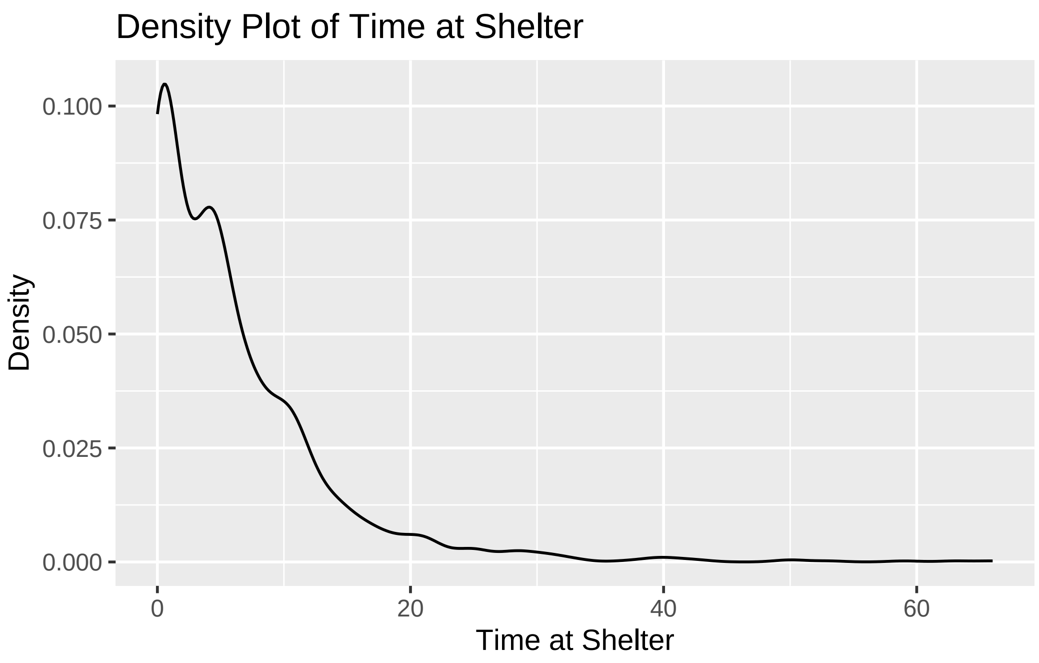
The boxplot shows varied medians across different outcomes. Adoption has a lower median time at the shelter compared to animals that were fostered or returned to the owner. Euthanized animals have a wide interquartile range, suggesting variable time frames before this outcome is reached.

```
ggplot(data = Animal, aes(x = chip_status, y = time_at_shelter, fill =
↪  chip_status))+
  geom_boxplot()+
  labs(x = "chip_status", y = "time_at_shelter")+
  theme(legend.position = "none")+
  ggtitle("boxplot of time_at_shelter vs. chip_status")
```



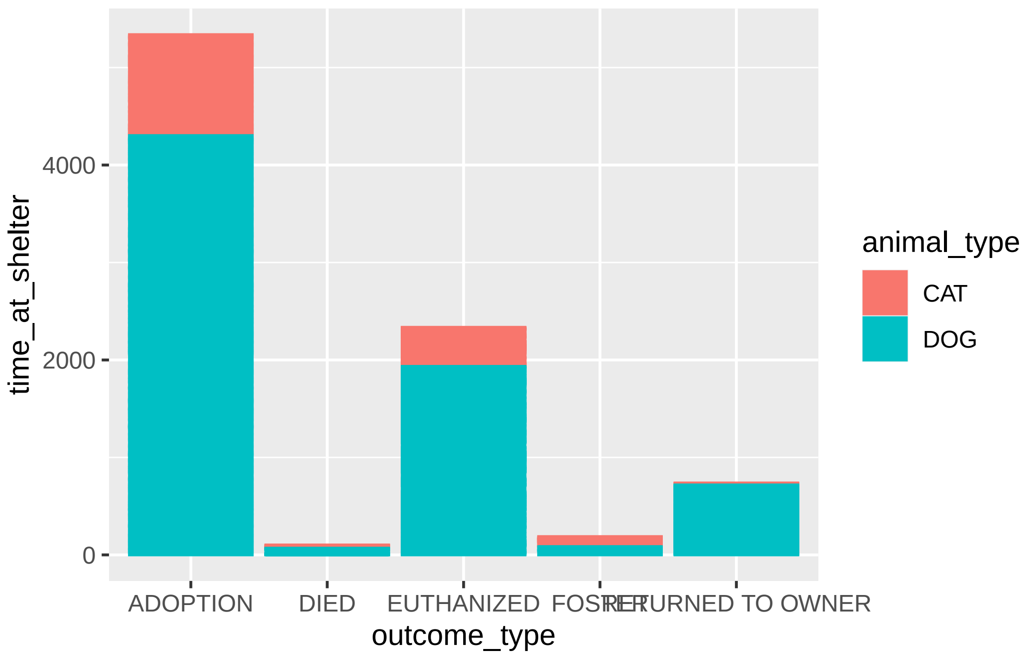
Animals with a chip have a more compact interquartile range of days spent at the shelter and fewer outliers, suggesting they tend to stay for a shorter and more consistent period. In contrast, animals without a chip show a wider range and more outliers, indicating longer and more variable stays.

```
ggplot(Animal, aes(x=time_at_shelter)) +
  geom_density() +
  labs(title="Density Plot of Time at Shelter",
       x="Time at Shelter",
       y="Density")
```



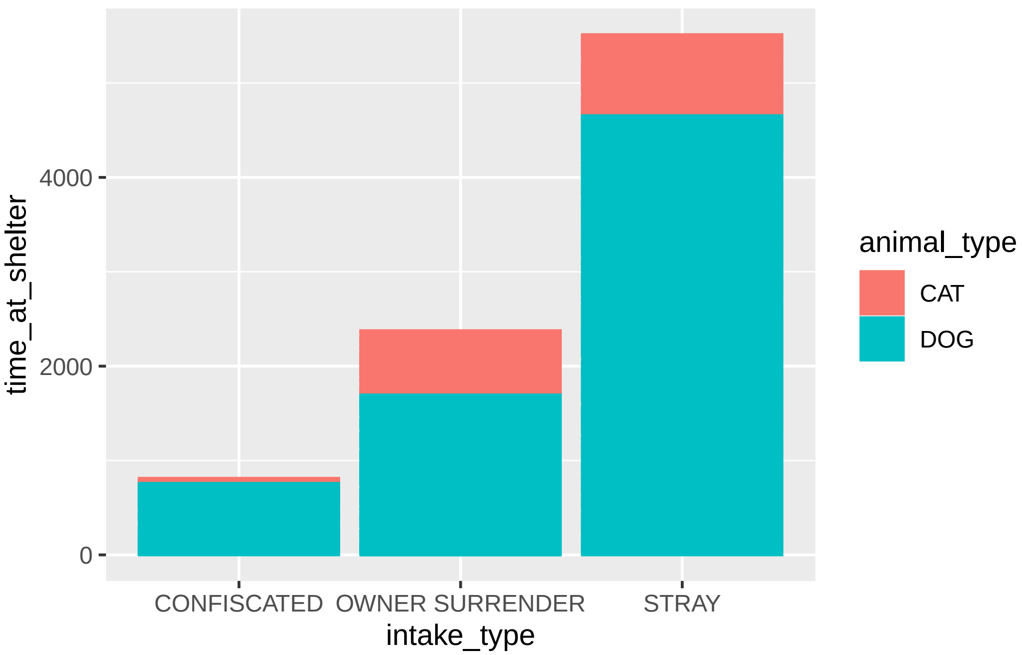
This plot shows that a large proportion of animals have a very short stay at the shelter, with a sharp decrease in density as time increases. This suggests that most animals are not at the shelter for an extended period.

```
ggplot(Animal, aes(x = outcome_type, y = time_at_shelter, fill =
 ↪  animal_type, colour = animal_type)) +
  geom_bar(stat = "identity")
```



The bar plot indicates that for both cats and dogs, adoption and return to owner are the most common outcomes. However, dogs have a much higher rate of being returned to their owners, which might be related to longer stays in the shelter.

```
ggplot(Animal, aes(x = intake_type, y = time_at_shelter, fill =
↪  animal_type, colour = animal_type)) +
  geom_bar(stat = "identity")
```



The barchart compares the time_at_shelter by cats and dogs, broken down by intake_type. The intake types shown are “Confiscated”, “Owner Surrender”, and “Stray”. It appears that the majority of animals in the shelter are strays, and among these, dogs tend to stay longer in the shelter than cats. The “Owner Surrender” category has a more even distribution between cats and dogs, but again, dogs show a longer shelter time overall. The “Confiscated” category has the least number of animals, but similar to the other categories, dogs have a longer shelter time than cats.

3 Poisson Model Fitting

```
# Full Model
poisson_model1 <- glm(time_at_shelter ~ animal_type + year + intake_type
 ↪  + outcome_type + chip_status + month, family = poisson, data =
 ↪  Animal)
summary(poisson_model1)
```

```
Call:
glm(formula = time_at_shelter ~ animal_type + year + intake_type +
    outcome_type + chip_status + month, family = poisson, data = Animal)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6262  -1.9624  -0.8582   0.6309  13.0556

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.762396   0.081583  46.117  < 2e-16 ***
animal_typeDOG                 0.046319   0.029011   1.597 0.110350
year2017                      -0.282622   0.038381  -7.364 1.79e-13 ***
intake_typeOWNER SURRENDER    -1.457518   0.043572 -33.451  < 2e-16 ***
intake_typeSTRAY              -1.034717   0.039272 -26.348  < 2e-16 ***
outcome_typeDIED              -0.713392   0.100568  -7.094 1.31e-12 ***
outcome_typeEUTHANIZED        -0.578425   0.025049 -23.092  < 2e-16 ***
outcome_typeFOSTER            -0.291184   0.075843  -3.839 0.000123 ***
outcome_typeRETURNED TO OWNER -1.541441   0.042153 -36.568  < 2e-16 ***
chip_statusNo Chip            -0.171638   0.028761  -5.968 2.41e-09 ***
month                         -0.024618   0.005053  -4.872 1.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10414.5  on 1432  degrees of freedom
Residual deviance:  8083.4  on 1422  degrees of freedom
AIC: 12087

Number of Fisher Scoring iterations: 6
```

```r
# Remove `animal_type`
poisson_model2 <- glm(time_at_shelter ~ year + intake_type +
 ↪  outcome_type + chip_status + month, family = poisson, data = Animal)
summary(poisson_model2)
```

```

Call:
glm(formula = time_at_shelter ~ year + intake_type + outcome_type +
    chip_status + month, family = poisson, data = Animal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-6.6136  -1.9595  -0.8591   0.6278  13.1035

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.818469   0.073627  51.862  < 2e-16 ***
year2017                      -0.288872   0.038182  -7.566 3.86e-14 ***
intake_typeOWNER SURRENDER    -1.461716   0.043507 -33.597  < 2e-16 ***
intake_typeSTRAY              -1.035218   0.039276 -26.358  < 2e-16 ***
outcome_typeDIED              -0.726402   0.100245  -7.246 4.28e-13 ***
outcome_typeEUTHANIZED        -0.581238   0.024990 -23.258  < 2e-16 ***
outcome_typeFOSTER            -0.312638   0.074653  -4.188 2.82e-05 ***
outcome_typeRETURNED TO OWNER -1.536883   0.042076 -36.526  < 2e-16 ***
chip_statusNo Chip            -0.176727   0.028595  -6.180 6.40e-10 ***
month                         -0.025708   0.005008  -5.133 2.85e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10414  on 1432  degrees of freedom
Residual deviance:  8086  on 1423  degrees of freedom
AIC: 12088

Number of Fisher Scoring iterations: 6
```

```r
# Include `time`
poisson_model3 <- glm(time_at_shelter ~ animal_type + intake_type +
 ↪  outcome_type + chip_status + time, family = poisson, data = Animal)
summary(poisson_model3)
```

```

Call:
glm(formula = time_at_shelter ~ animal_type + intake_type + outcome_type +
    chip_status + time, family = poisson, data = Animal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6275  -1.9612  -0.8635   0.6307  13.0355

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)                      3.751120   0.073898  50.761  < 2e-16 ***
animal_typeDOG                   0.047338   0.028842   1.641 0.100737    
intake_typeOWNER SURRENDER      -1.458011   0.043545 -33.483  < 2e-16 ***
intake_typeSTRAY                -1.035248   0.039237 -26.384  < 2e-16 ***
outcome_typeDIED                -0.712647   0.100542  -7.088 1.36e-12 ***
outcome_typeEUTHANIZED          -0.578558   0.025045 -23.101  < 2e-16 ***
outcome_typeFOSTER              -0.291232   0.075843  -3.840 0.000123 ***
outcome_typeRETURNED TO OWNER -1.541390   0.042151 -36.568  < 2e-16 ***
chip_statusNo Chip              -0.171373   0.028750  -5.961 2.51e-09 ***
time                            -0.023320   0.003116  -7.485 7.16e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10414.5  on 1432  degrees of freedom
Residual deviance:  8083.5  on 1423  degrees of freedom
AIC: 12085

Number of Fisher Scoring iterations: 6
```

```r
# Remove `animal_type`
poisson_model4 <- glm(time_at_shelter ~ intake_type + outcome_type +
 ↪   chip_status + time, family = poisson, data = Animal)
summary(poisson_model4)
```

```
Call:
glm(formula = time_at_shelter ~ intake_type + outcome_type + 
    chip_status + time, family = poisson, data = Animal)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.6153  -1.9576  -0.8505   0.6462  13.0740  

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      3.802874   0.066787  56.940  < 2e-16 ***
intake_typeOWNER SURRENDER      -1.462632   0.043468 -33.648  < 2e-16 ***
intake_typeSTRAY                -1.036064   0.039239 -26.404  < 2e-16 ***
outcome_typeDIED                -0.725711   0.100236  -7.240 4.49e-13 ***
```

```
outcome_typeEUTHANIZED        -0.581539   0.024983 -23.277  < 2e-16 ***
outcome_typeFOSTER            -0.313446   0.074636  -4.200 2.67e-05 ***
outcome_typeRETURNED TO OWNER -1.536645   0.042072 -36.524  < 2e-16 ***
chip_statusNo Chip            -0.176501   0.028592  -6.173 6.70e-10 ***
time                          -0.023729   0.003103  -7.646 2.07e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10414.5  on 1432  degrees of freedom
Residual deviance:  8086.2  on 1424  degrees of freedom
AIC: 12086

Number of Fisher Scoring iterations: 6
```

```r
# Step Selection for Final Model
null.model <- glm(time_at_shelter ~ 1, family = poisson, data = Animal)
step.model <- step(null.model,scope = list(lower = null.model, upper =
 ↪  poisson_model3), direction = "both")
```

```
Start:  AIC=14398.4
time_at_shelter ~ 1

               Df Deviance   AIC
+ outcome_type  4   9092.4 13084
+ intake_type   2  10122.1 14110
+ time          1  10295.6 14282
<none>             10414.5 14398
+ animal_type   1  10412.9 14399
+ chip_status   1  10414.2 14400

Step:  AIC=13084.37
time_at_shelter ~ outcome_type

               Df Deviance   AIC
+ intake_type   2   8179.9 12176
+ time          1   9040.3 13034
+ animal_type   1   9074.0 13068
<none>              9092.4 13084
+ chip_status   1   9091.3 13085
```

```
- outcome_type  4  10414.5 14398

Step:  AIC=12175.81
time_at_shelter ~ outcome_type + intake_type

               Df Deviance   AIC
+ time          1   8123.3 12121
+ chip_status   1   8144.6 12142
+ animal_type   1   8171.6 12170
<none>              8179.9 12176
- intake_type   2   9092.4 13084
- outcome_type  4  10122.1 14110

Step:  AIC=12121.24
time_at_shelter ~ outcome_type + intake_type + time

               Df Deviance   AIC
+ chip_status   1   8086.2 12086
+ animal_type   1   8118.1 12118
<none>              8123.3 12121
- time          1   8179.9 12176
- intake_type   2   9040.3 13034
- outcome_type  4   9993.1 13983

Step:  AIC=12086.19
time_at_shelter ~ outcome_type + intake_type + time + chip_status

               Df Deviance   AIC
+ animal_type   1   8083.5 12086
<none>              8086.2 12086
- chip_status   1   8123.3 12121
- time          1   8144.6 12142
- intake_type   2   9039.1 13035
- outcome_type  4   9992.7 13985

Step:  AIC=12085.48
time_at_shelter ~ outcome_type + intake_type + time + chip_status +
    animal_type

               Df Deviance   AIC
<none>              8083.5 12086
- animal_type   1   8086.2 12086
- chip_status   1   8118.1 12118
- time          1   8139.4 12139
- intake_type   2   9026.1 13024
- outcome_type  4   9991.2 13985
```

```r
summary(step.model)
```

```
Call:
glm(formula = time_at_shelter ~ outcome_type + intake_type +
    time + chip_status + animal_type, family = poisson, data = Animal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6275  -1.9612  -0.8635   0.6307  13.0355

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.751120   0.073898  50.761  < 2e-16 ***
outcome_typeDIED              -0.712647   0.100542  -7.088 1.36e-12 ***
outcome_typeEUTHANIZED        -0.578558   0.025045 -23.101  < 2e-16 ***
outcome_typeFOSTER            -0.291232   0.075843  -3.840 0.000123 ***
outcome_typeRETURNED TO OWNER -1.541390   0.042151 -36.568  < 2e-16 ***
intake_typeOWNER SURRENDER    -1.458011   0.043545 -33.483  < 2e-16 ***
intake_typeSTRAY              -1.035248   0.039237 -26.384  < 2e-16 ***
time                          -0.023320   0.003116  -7.485 7.16e-14 ***
chip_statusNo Chip            -0.171373   0.028750  -5.961 2.51e-09 ***
animal_typeDOG                 0.047338   0.028842   1.641 0.100737
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10414.5  on 1432  degrees of freedom
Residual deviance:  8083.5  on 1423  degrees of freedom
AIC: 12085

Number of Fisher Scoring iterations: 6
```

The final model chosen in the step selection process includes 4 independent variables: intake type, outcome type, chip status, and time. The model equation is as follows:

$$
\begin{aligned}
\text{time at shelter } =\ & \beta_0 + \beta_1 \times \mathbb{1}_{\text{intake type: owner surrender}} + \beta_2 \times \mathbb{1}_{\text{intake type: stray}} + \beta_3 \times \mathbb{1}_{\text{outcome type: died}} \\
& + \beta_4 \times \mathbb{1}_{\text{outcome type: euthanized}} + \beta_5 \times \mathbb{1}_{\text{outcome type: foster}} + \beta_6 \times \mathbb{1}_{\text{outcome type: returned to owner}} \\
& + \beta_7 \times \mathbb{1}_{\text{chip status: no chip}} + \beta_8 \times \text{time}
\end{aligned}
$$

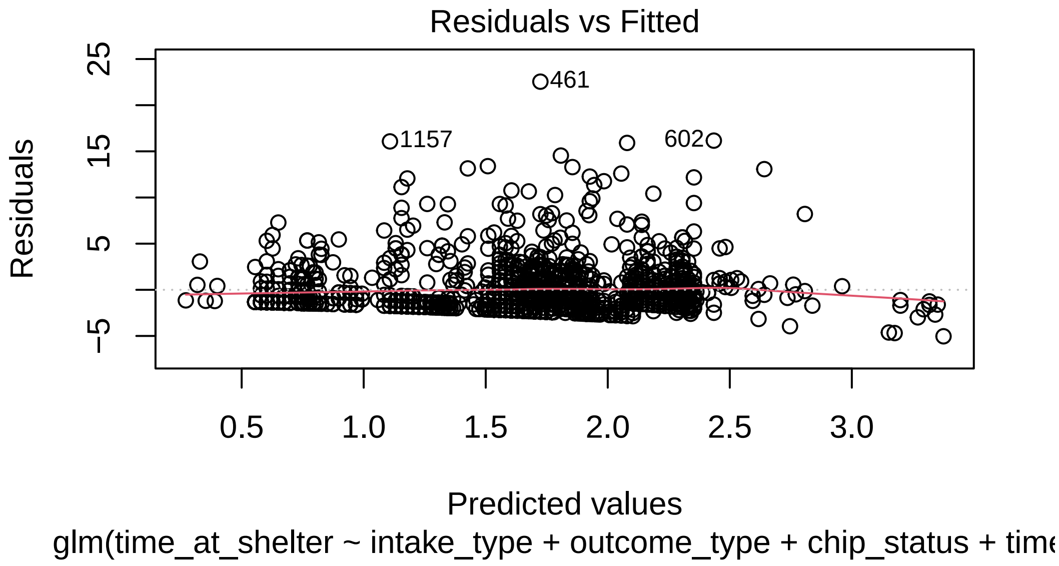
where,

$$
\begin{aligned}
\beta_0 &= \text{mean time at shelter when intake type is confiscated and outcome type is adopted} \\
&\quad \text{and the animal has a chip} \\
\beta_1 \text{ and } \beta_2 &= \text{effects of different intake types on time at shelter} \\
\beta_3 \text{ and } \beta_6 &= \text{effects of different outcome types on time at shelter} \\
\beta_7 &= \text{effect of an animal having no chip on time at shelter} \\
\beta_8 &= \text{effect of how many months since January 2016 since the animal arrived} \\
&\quad \text{at the shelter on time at shelter}
\end{aligned}
$$

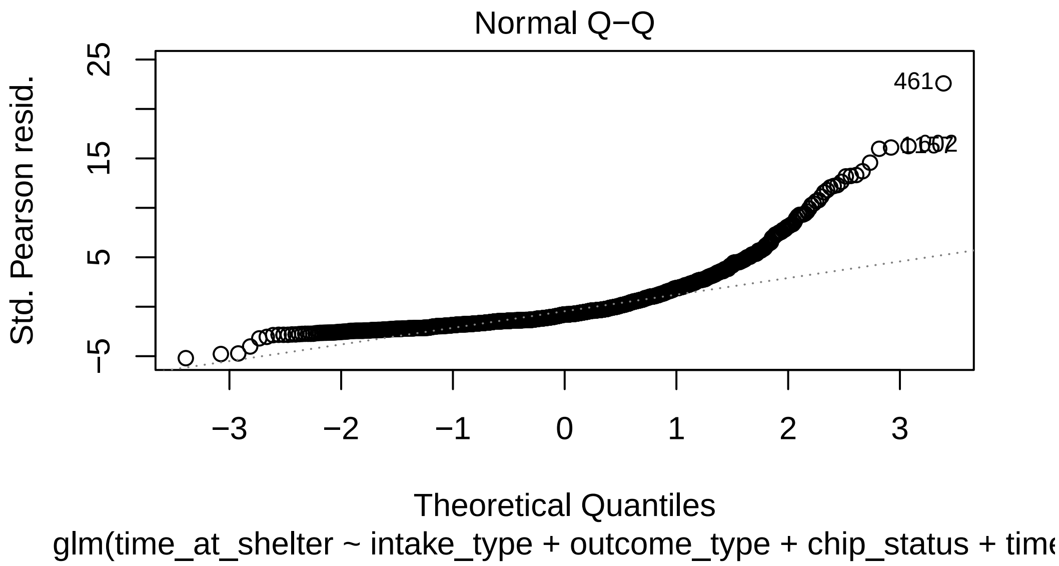
3.0.0.1 Model Diagnostics for Poisson Regression

```r
# Check Assumptions
residuals <- residuals(poisson_model4)

plot(poisson_model4, which = 1)
```



```r
plot(poisson_model4, which = 2)
```



The residuals against fitted values indicate the the residuals are scattered around the zero line indicating this assumptions appears valid.

From the QQ plot, it seems fairly reasonable that the normality assumption holds but there is some deviations in the tails of the distribution.

```
# Distribution Assumption
mean(Animal$time_at_shelter)
```

```
[1] 6.071877
```

```
var(Animal$time_at_shelter)
```

```
[1] 54.3405
```

Given that the mean is approximately 6.07 and the variance is substantially higher at 54.34, the data demonstrate overdispersion (where the variance is greater than the mean), which is a common characteristic in count data and can invalidate models that assume equal mean and variance such as the Poisson distribution. The model data distribution assumption here is most likely a negative binomial distribution.

4 Negative-Binomial Model Fitting

```
# Full Model
negbin_model1 <- glm.nb(time_at_shelter ~ animal_type + intake_type +
 ↪  outcome_type + chip_status + year + month, data = Animal)
summary(negbin_model1)
```

```

Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    outcome_type + chip_status + year + month, data = Animal,
    init.theta = 1.044921073, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5536  -1.0748  -0.3451   0.2255   3.7839
```

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    4.09360    0.23186  17.655  < 2e-16 ***
animal_typeDOG                 0.03798    0.07664   0.496  0.62022    
intake_typeOWNER SURRENDER    -1.81079    0.14099 -12.844  < 2e-16 ***
intake_typeSTRAY              -1.39268    0.13016 -10.700  < 2e-16 ***
outcome_typeDIED              -0.73274    0.22915  -3.198  0.00139 ** 
outcome_typeEUTHANIZED        -0.63002    0.06527  -9.653  < 2e-16 ***
outcome_typeFOSTER            -0.34085    0.20586  -1.656  0.09777 .  
outcome_typeRETURNED TO OWNER -1.76909    0.09381 -18.859  < 2e-16 ***
chip_statusNo Chip            -0.20330    0.07644  -2.659  0.00783 ** 
year2017                      -0.22941    0.10201  -2.249  0.02452 *  
month                         -0.01912    0.01349  -1.418  0.15632    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0449) family taken to be 1)

    Null deviance: 2063.8  on 1432  degrees of freedom
Residual deviance: 1660.1  on 1422  degrees of freedom
AIC: 7893.1

Number of Fisher Scoring iterations: 1


              Theta:  1.0449 
          Std. Err.:  0.0521 

 2 x log-likelihood:  -7869.0830 
```

```r
# Use `season` to replace `month`
negbin_model2 <- glm.nb(time_at_shelter ~ animal_type + intake_type +
 ↪  outcome_type + chip_status + year + season, data = Animal)
summary(negbin_model2)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type + 
    outcome_type + chip_status + year + season, data = Animal, 
    init.theta = 1.047354542, link = log)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5649  -1.1030  -0.3423   0.2325   3.5137

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    3.83676    0.17538  21.877  < 2e-16 ***
animal_typeDOG                 0.02990    0.07648   0.391  0.69585
intake_typeOWNER SURRENDER    -1.81034    0.14090 -12.849  < 2e-16 ***
intake_typeSTRAY              -1.39853    0.13003 -10.756  < 2e-16 ***
outcome_typeDIED              -0.70877    0.22917  -3.093  0.00198 **
outcome_typeEUTHANIZED        -0.62229    0.06557  -9.490  < 2e-16 ***
outcome_typeFOSTER            -0.33689    0.20583  -1.637  0.10169
outcome_typeRETURNED TO OWNER -1.76726    0.09387 -18.827  < 2e-16 ***
chip_statusNo Chip            -0.19878    0.07650  -2.598  0.00936 **
year2017                      -0.08597    0.09020  -0.953  0.34055
seasonSpring                  -0.01248    0.10271  -0.122  0.90327
seasonSummer                  -0.01579    0.09928  -0.159  0.87359
seasonWinter                   0.14947    0.09049   1.652  0.09861 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0474) family taken to be 1)

    Null deviance: 2067.0  on 1432  degrees of freedom
Residual deviance: 1659.8  on 1420  degrees of freedom
AIC: 7894.3

Number of Fisher Scoring iterations: 1


              Theta:  1.0474
          Std. Err.:  0.0523

 2 x log-likelihood:  -7866.3020
```

```r
# Use `time` to replace `year` & `season`
negbin_model3 <- glm.nb(time_at_shelter ~ animal_type + intake_type +
 ↪  outcome_type + chip_status + time, data = Animal)
summary(negbin_model3)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    outcome_type + chip_status + time, data = Animal, init.theta = 1.044920979,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5536  -1.0748  -0.3451   0.2255   3.7838

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    4.093527   0.213986  19.130  < 2e-16 ***
animal_typeDOG                 0.037985   0.076196   0.499  0.61812
intake_typeOWNER SURRENDER    -1.810787   0.140930 -12.849  < 2e-16 ***
intake_typeSTRAY              -1.392681   0.130084 -10.706  < 2e-16 ***
outcome_typeDIED              -0.732733   0.229056  -3.199  0.00138 **
outcome_typeEUTHANIZED        -0.630021   0.065267  -9.653  < 2e-16 ***
outcome_typeFOSTER            -0.340850   0.205858  -1.656  0.09777 .
outcome_typeRETURNED TO OWNER -1.769096   0.093809 -18.859  < 2e-16 ***
chip_statusNo Chip            -0.203295   0.076412  -2.661  0.00780 **
time                          -0.019116   0.008356  -2.288  0.02216 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0449) family taken to be 1)

    Null deviance: 2063.8  on 1432  degrees of freedom
Residual deviance: 1660.1  on 1423  degrees of freedom
AIC: 7891.1

Number of Fisher Scoring iterations: 1

              Theta:  1.0449
          Std. Err.:  0.0521

 2 x log-likelihood:  -7869.0830
```

```r
# Remove `animal_type`
negbin_model4 <- glm.nb(time_at_shelter ~ intake_type + outcome_type +
 ↪  chip_status + time, data = Animal)
summary(negbin_model4)
```

```
Call:
glm.nb(formula = time_at_shelter ~ intake_type + outcome_type +
    chip_status + time, data = Animal, init.theta = 1.044545676,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5523  -1.0788  -0.3440   0.2282   3.7946

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    4.13871    0.19783  20.920  < 2e-16 ***
intake_typeOWNER SURRENDER    -1.81689    0.14090 -12.895  < 2e-16 ***
intake_typeSTRAY              -1.39365    0.13010 -10.712  < 2e-16 ***
outcome_typeDIED              -0.73862    0.22810  -3.238  0.00120 **
outcome_typeEUTHANIZED        -0.62970    0.06513  -9.668  < 2e-16 ***
outcome_typeFOSTER            -0.35616    0.20252  -1.759  0.07863 .
outcome_typeRETURNED TO OWNER -1.76429    0.09323 -18.925  < 2e-16 ***
chip_statusNo Chip            -0.20865    0.07611  -2.741  0.00612 **
time                          -0.01962    0.00832  -2.359  0.01834 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0445) family taken to be 1)

    Null deviance: 2063.3  on 1432  degrees of freedom
Residual deviance: 1660.0  on 1424  degrees of freedom
AIC: 7889.3

Number of Fisher Scoring iterations: 1

              Theta:  1.0445
          Std. Err.:  0.0521

 2 x log-likelihood:  -7869.3260
```

```r
# Step Selection for Final Model
null.model <- glm.nb(time_at_shelter ~ 1, data = Animal)
step.model <- step(null.model, scope = list(lower = null.model, upper =
 ↪  negbin_model3), direction = "both")
```

```
Start:  AIC=8221.3
time_at_shelter ~ 1

               Df Deviance    AIC
+ outcome_type  4   1488.5 8070.3
+ intake_type   2   1618.9 8196.7
+ time          1   1634.3 8210.1
<none>              1647.5 8221.3
+ animal_type   1   1647.3 8223.1
+ chip_status   1   1647.5 8223.3

Step:  AIC=8059.3
time_at_shelter ~ outcome_type

               Df Deviance    AIC
+ intake_type   2   1499.4 7906.5
+ time          1   1650.5 8055.5
+ animal_type   1   1651.9 8056.9
<none>              1656.3 8059.3
+ chip_status   1   1656.1 8061.1
- outcome_type  4   1838.6 8233.6

Step:  AIC=7896.4
time_at_shelter ~ outcome_type + intake_type

               Df Deviance    AIC
+ chip_status   1   1652.5 7890.8
+ time          1   1655.0 7893.3
<none>              1660.1 7896.4
+ animal_type   1   1658.8 7897.1
- intake_type   2   1838.1 8070.4
- outcome_type  4   2008.6 8236.9

Step:  AIC=7890.79
time_at_shelter ~ outcome_type + intake_type + chip_status

               Df Deviance    AIC
+ time          1   1654.4 7887.3
<none>              1659.9 7890.8
+ animal_type   1   1659.3 7892.2
- chip_status   1   1667.5 7896.4
- intake_type   2   1846.2 8073.1
- outcome_type  4   2017.7 8240.6
```

```
Step:  AIC=7887.33
time_at_shelter ~ outcome_type + intake_type + chip_status +
    time

               Df Deviance    AIC
<none>              1660.0 7887.3
+ animal_type   1   1659.7 7889.1
- time          1   1665.5 7890.8
- chip_status   1   1667.9 7893.3
- intake_type   2   1845.7 8069.1
- outcome_type  4   2006.4 8225.8
```

```r
summary(step.model)
```

```

Call:
glm.nb(formula = time_at_shelter ~ outcome_type + intake_type +
    chip_status + time, data = Animal, init.theta = 1.044545676,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5523  -1.0788  -0.3440   0.2282   3.7946

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    4.13871    0.19783  20.920  < 2e-16 ***
outcome_typeDIED              -0.73862    0.22810  -3.238  0.00120 **
outcome_typeEUTHANIZED        -0.62970    0.06513  -9.668  < 2e-16 ***
outcome_typeFOSTER            -0.35616    0.20252  -1.759  0.07863 .
outcome_typeRETURNED TO OWNER -1.76429    0.09323 -18.925  < 2e-16 ***
intake_typeOWNER SURRENDER    -1.81689    0.14090 -12.895  < 2e-16 ***
intake_typeSTRAY              -1.39365    0.13010 -10.712  < 2e-16 ***
chip_statusNo Chip            -0.20865    0.07611  -2.741  0.00612 **
time                          -0.01962    0.00832  -2.359  0.01834 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0445) family taken to be 1)
```

```
    Null deviance: 2063.3  on 1432  degrees of freedom
Residual deviance: 1660.0  on 1424  degrees of freedom
AIC: 7889.3

Number of Fisher Scoring iterations: 1


              Theta:  1.0445
          Std. Err.:  0.0521

 2 x log-likelihood:  -7869.3260
```

The final negative-binomial model equation is represented as:

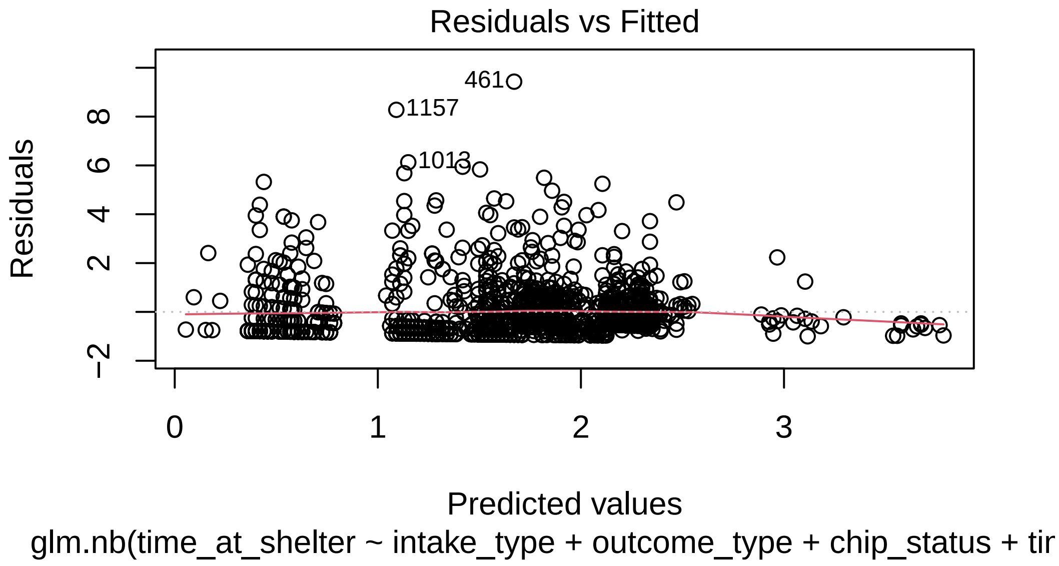
$$\begin{aligned}\text{time at shelter } &= \beta_0 + \beta_1 \times \mathbb{1}_{\text{intake type: owner surrender}} + \beta_2 \times \mathbb{1}_{\text{intake type: stray}} + \beta_3 \times \mathbb{1}_{\text{outcome type: died}} \\ &\quad + \beta_4 \times \mathbb{1}_{\text{outcome type: euthanized}} + \beta_5 \times \mathbb{1}_{\text{outcome type: foster}} + \beta_6 \times \mathbb{1}_{\text{outcome type: returned to owner}} \\ &\quad + \beta_7 \times \mathbb{1}_{\text{chip status: no chip}} + \beta_8 \times \text{time}\end{aligned}$$

where,

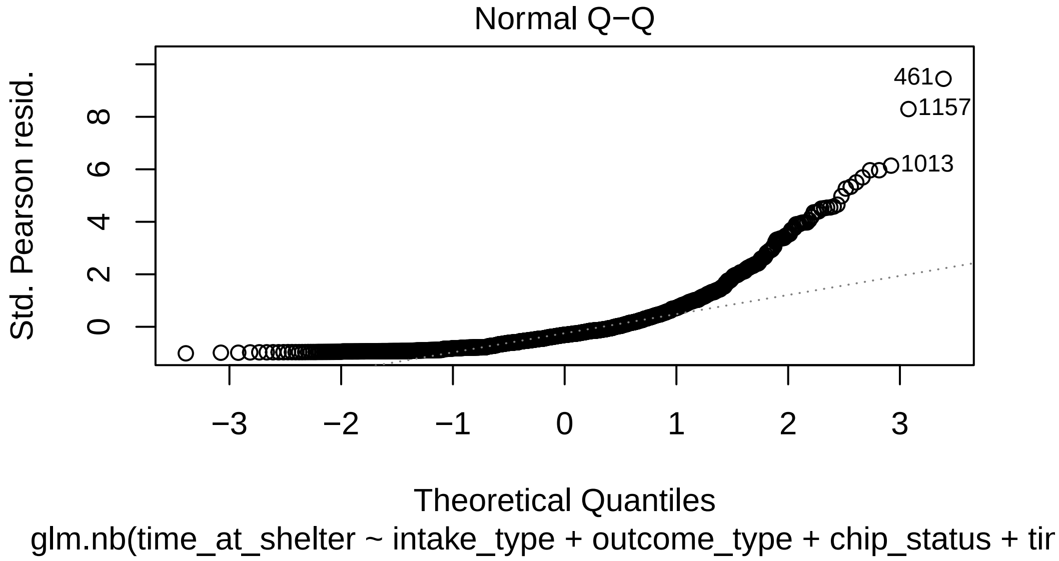
$$\begin{aligned}\beta_0 &= \text{mean time at shelter when intake type is confiscated and outcome type is adopted} \\ &\quad\ \text{and the animal has a chip} \\ \beta_1 \text{ and } \beta_2 &= \text{effects of different intake types on time at shelter} \\ \beta_3 \text{ to } \beta_6 &= \text{effects of different outcome types on time at shelter} \\ \beta_7 &= \text{effect of an animal having no chip on time at shelter} \\ \beta_8 &= \text{effect of how many months since January 2016 since the animal arrived} \\ &\quad\ \text{at the shelter on time at shelter}\end{aligned}$$

```r
# Check Assumptions
residuals <- residuals(negbin_model4)

plot(negbin_model4, which = 1)
```



```
plot(negbin_model4, which = 2)
```



The residuals appear randomly dispersed around the horizontal line without evident patterns, suggesting no major violations of model assumptions. However, slight curvilinear trends at lower fitted values may indicate potential nonlinear relationships or heteroscedasticity.

While the bulk of points follow the reference line, deviations in the tails suggest the residuals may have a non-normal distribution, which is not unexpected for count data modelled with a negative binomial distribution.

5 Conclusion

In conclusion, for the distribution and model assumption, negative binomial model performs better. For the variable, outcome, intake types and general time trend without chips are all significantly negative impact with time at shelter.

In terms of further explorations, outcome type may be a confounding variable since once we know the outcome type, we also know how long the time spent at shelter is. We may also introduce new variables to replace the outcome type.