DAS_Group_19

1 Introduction

The question for Group 19 is: Which factors influence the number of days an animal spends in the shelter before their final outcome is decided?

We start by loading the required packages.

```
library(ggplot2)
library(tidyverse)
library(dplyr)
library(skimr)
library(moderndive)
library(sjPlot)
library(stats)
library(jtools)
library(readr)
```

The data we are required to use is read below.

```
animals<-read.csv("C:/Users/Yuchen/Documents/tmp/DAS-Group-19/dataset19.csv")
```

And we should always view the whole data at first.

```
glimpse(animals)
```

```
## Rows: 1,853
## Columns: 7
## $ animal_type     <chr> "CAT", "DOG", "DOG", "DOG", "DOG", "CAT", "DOG", "CAT"~
## $ month           <int> 11, 12, 5, 8, 10, 5, 12, 10, 4, 4, 6, 6, 6, 6, 5, 9, 3~
## $ year            <int> 2016, 2016, 2017, 2017, 2016, 2017, 2016, 2016, 2017, ~
## $ intake_type     <chr> "STRAY", "OWNER SURRENDER", "OWNER SURRENDER", "STRAY"~
## $ outcome_type    <chr> "ADOPTION", "ADOPTION", "EUTHANIZED", "RETURNED TO OWN~
## $ chip_status     <chr> "SCAN NO CHIP", "SCAN NO CHIP", "SCAN CHIP", "SCAN NO ~
## $ time_at_shelter <int> 21, 2, 2, 0, 0, 0, 1, 6, 5, 21, 7, 0, 22, 8, 4, 10, 2,~
```

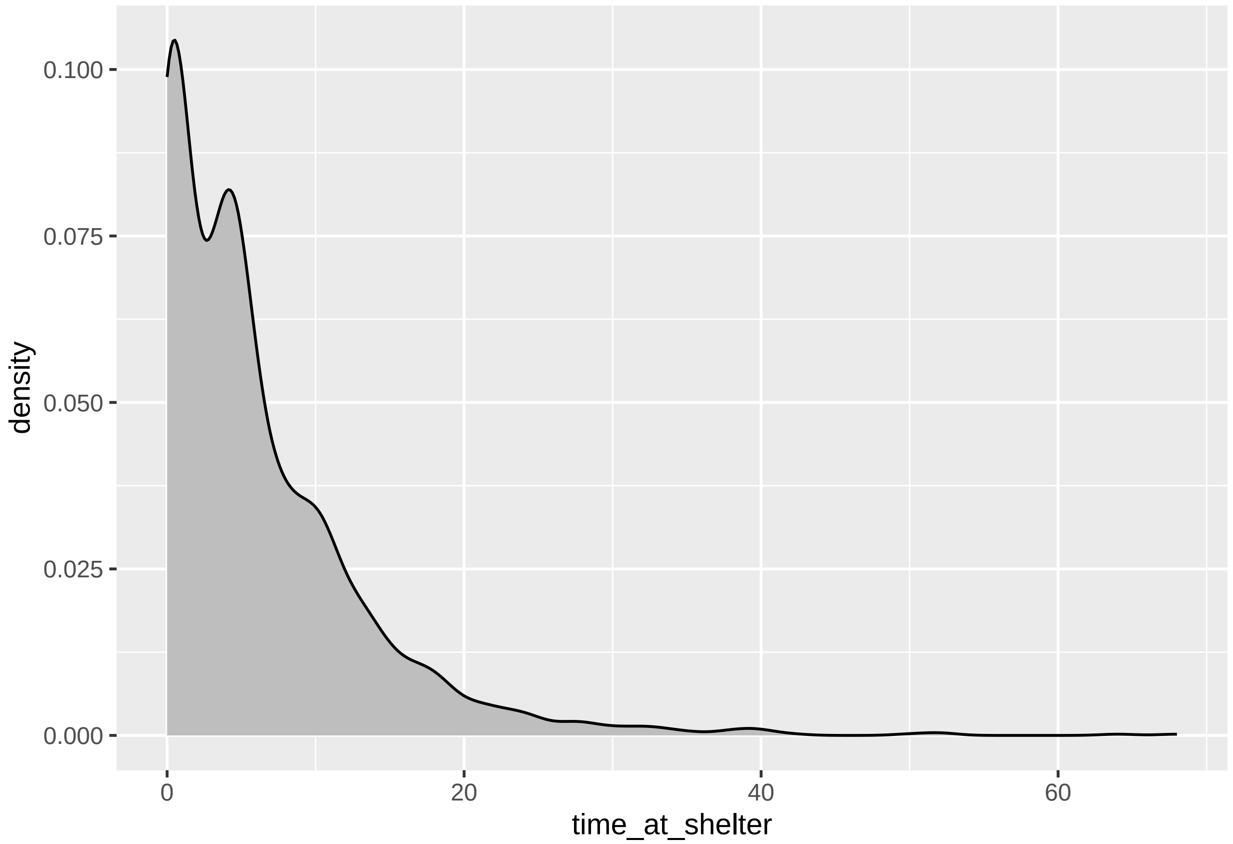
2 Exploratory Data Analysis

We need to choose which GLM should be used for the model by plotting the densities of manufacture (dependent variable y)

We first investigate what kind of density function y (time in the shelter) obeys

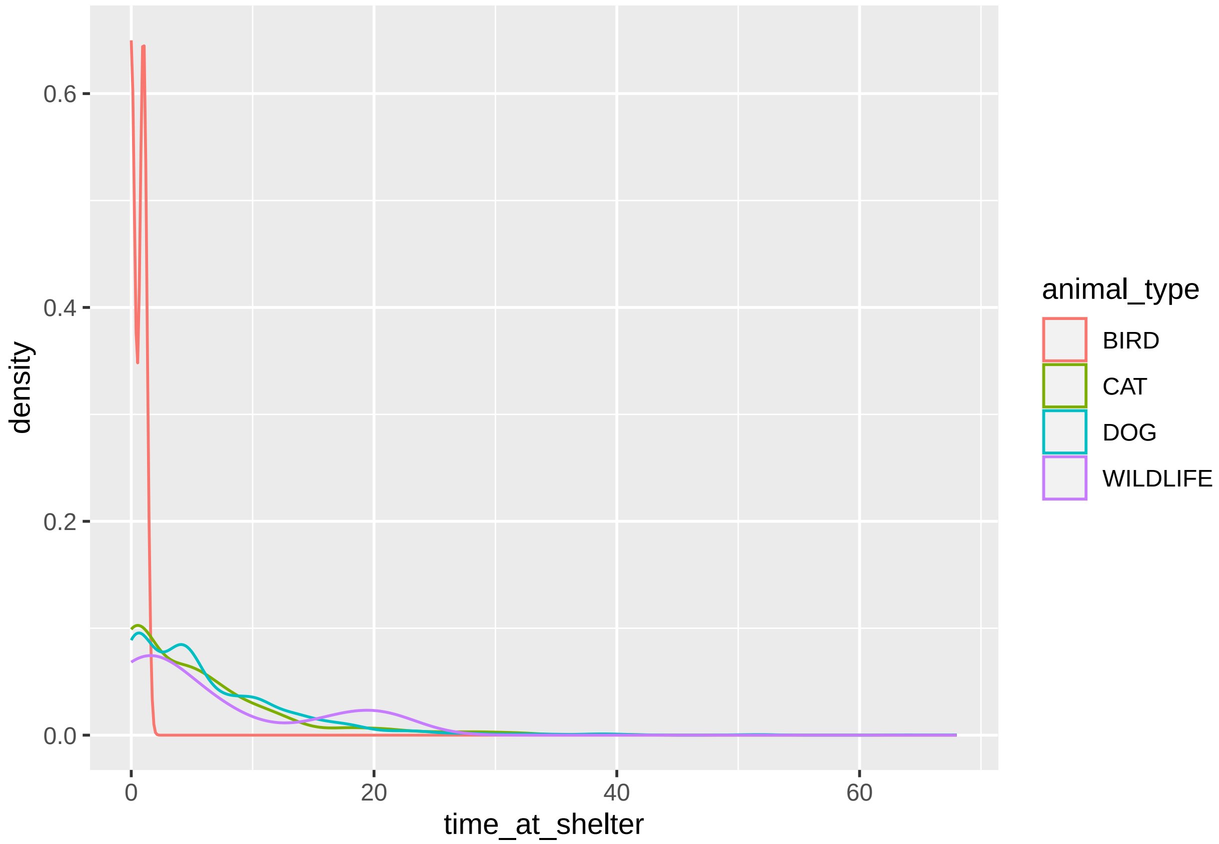
The overall density function of time at shelter is plotted below

```
p<-ggplot(animals, aes(x = time_at_shelter))
p + geom_density(color = "black", fill = "gray")
```



The density functions for different animals are plotted below

```
ggplot(animals, aes(x = time_at_shelter))+
  geom_density(aes(color = animal_type))
```

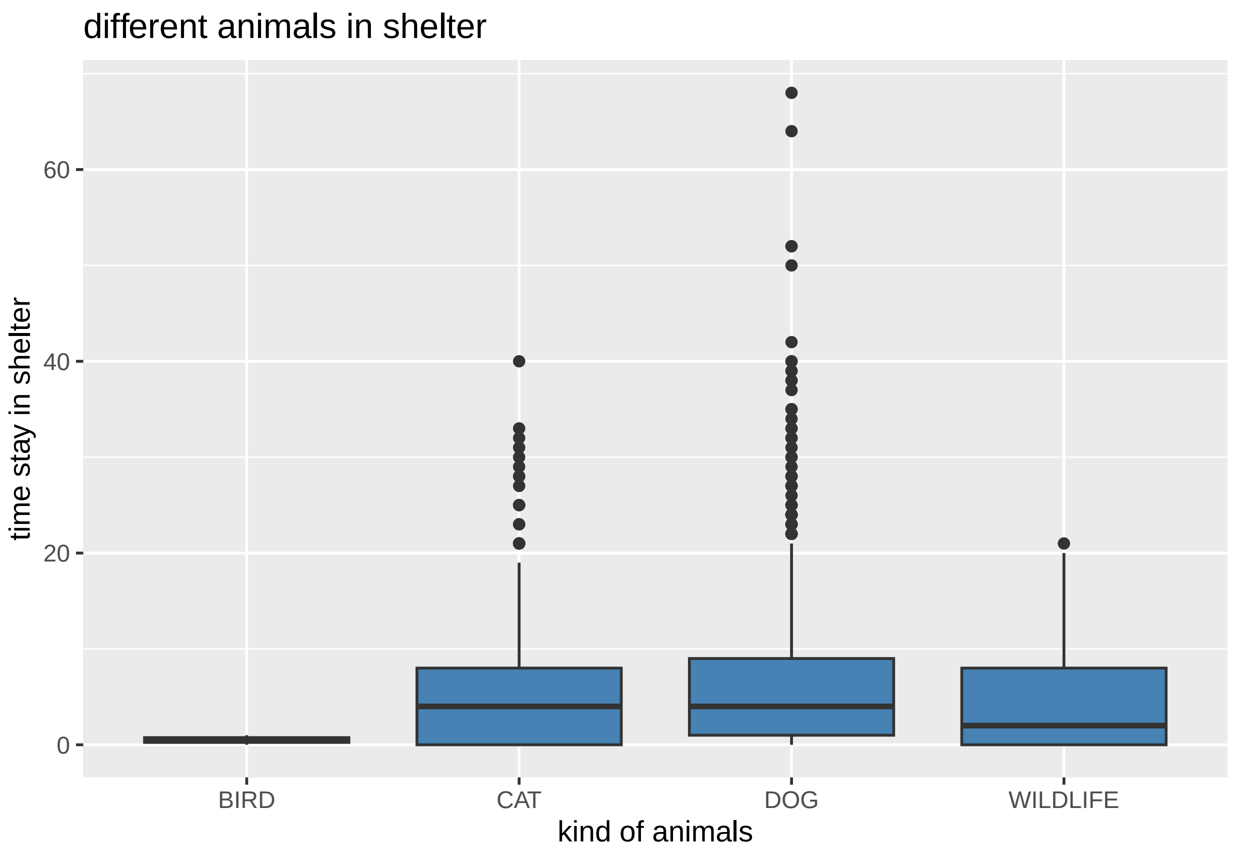


As we can see, these two graphs have shown a typical poisson distribution which give us the direction of using the **logit link function** combined with the assumption that the response y (time at shelter) obeys the **poisson distribution** for **GLM** method later on.

But we will still observe the distribution of each variable (via boxplots and histograms)

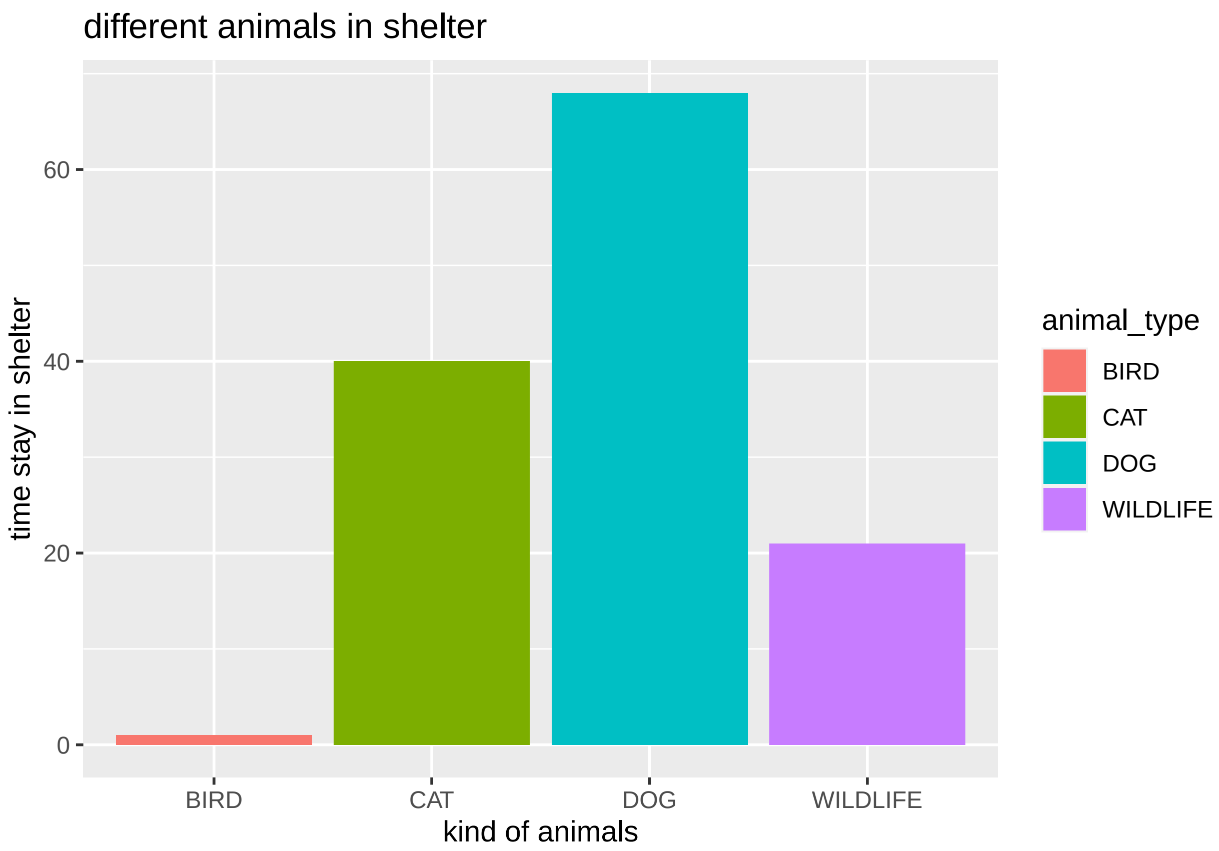
Using boxplot for observing the time at shelter for different kinds of animals.

```
ggplot(data = animals, mapping = aes(x = factor(animal_type), y = time_at_shelter)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "kind of animals", y = "time stay in shelter",
       title = "different animals in shelter")
```



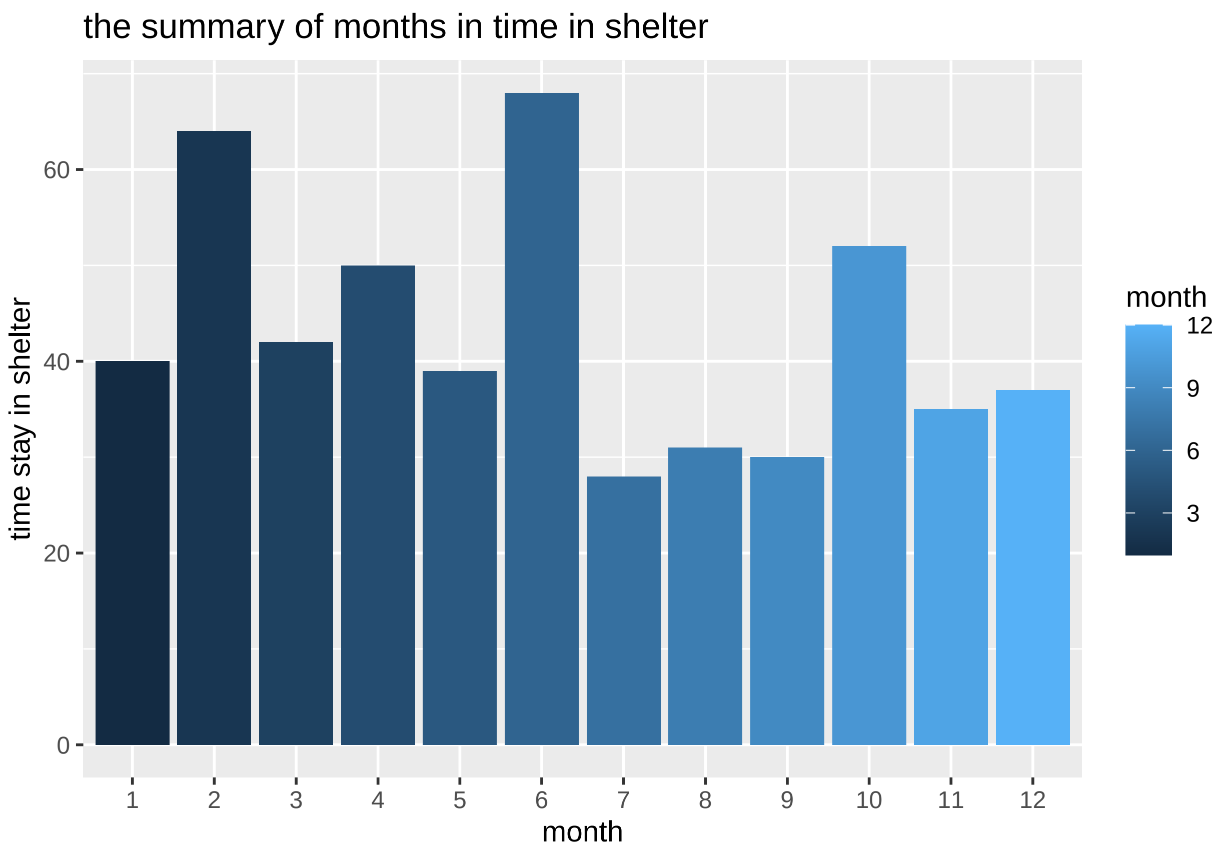
Observing the time at shelter for different kinds of animals by plotting the bar chart.

```
ggplot(data = animals, mapping = aes(x = factor(animal_type), y = time_at_shelter, fill = animal_type))
  geom_col(position = "dodge") +
  labs(x = "kind of animals", y = "time stay in shelter",
       title = "different animals in shelter")
```

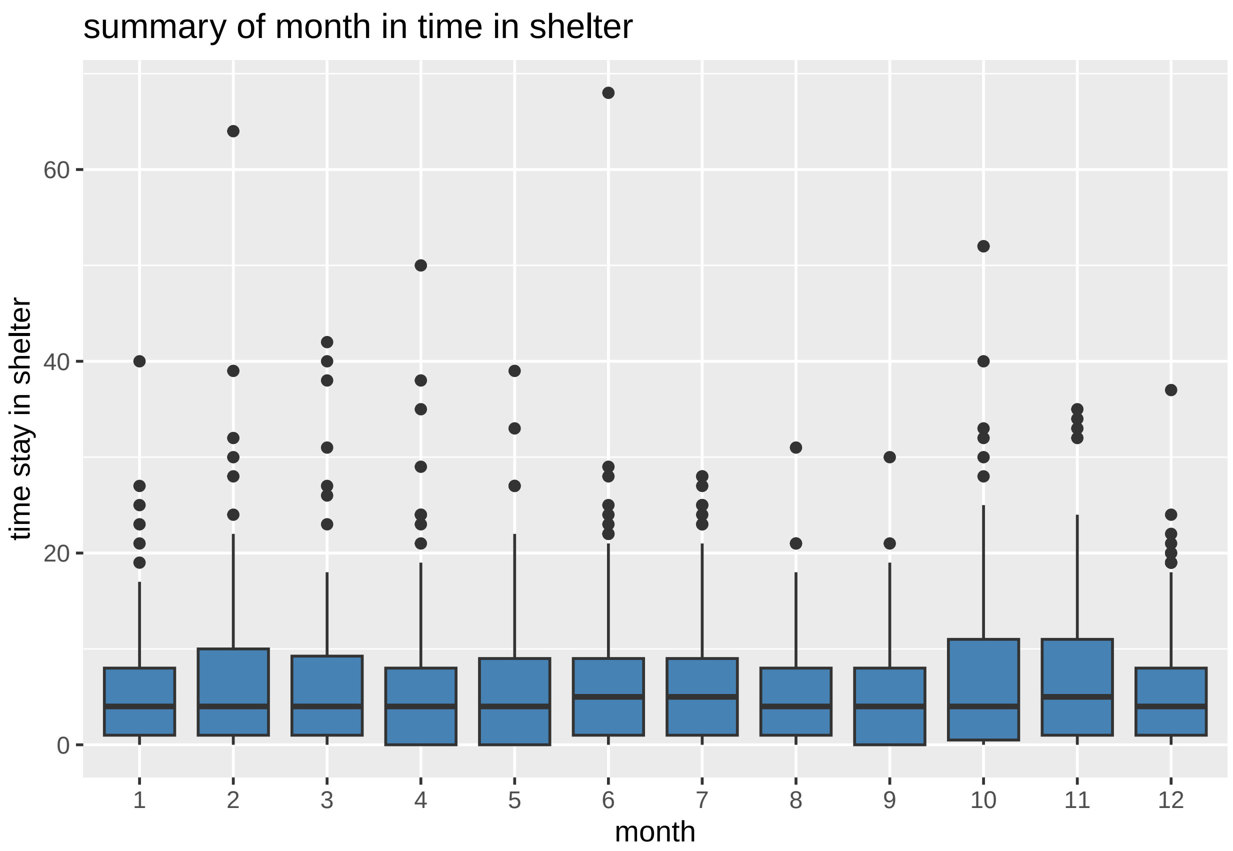


Observing the time that animals stay in shelter in different month.

```
ggplot(data = animals, mapping = aes(x = factor(month), y = time_at_shelter, fill = month)) +
  geom_col(position = "dodge") +
  labs(x = "month", y = "time stay in shelter",
       title = "the summary of months in time in shelter")
```



```r
ggplot(data = animals, mapping = aes(x = factor(month), y = time_at_shelter)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "month", y = "time stay in shelter",
       title = "summary of month in time in shelter")
```



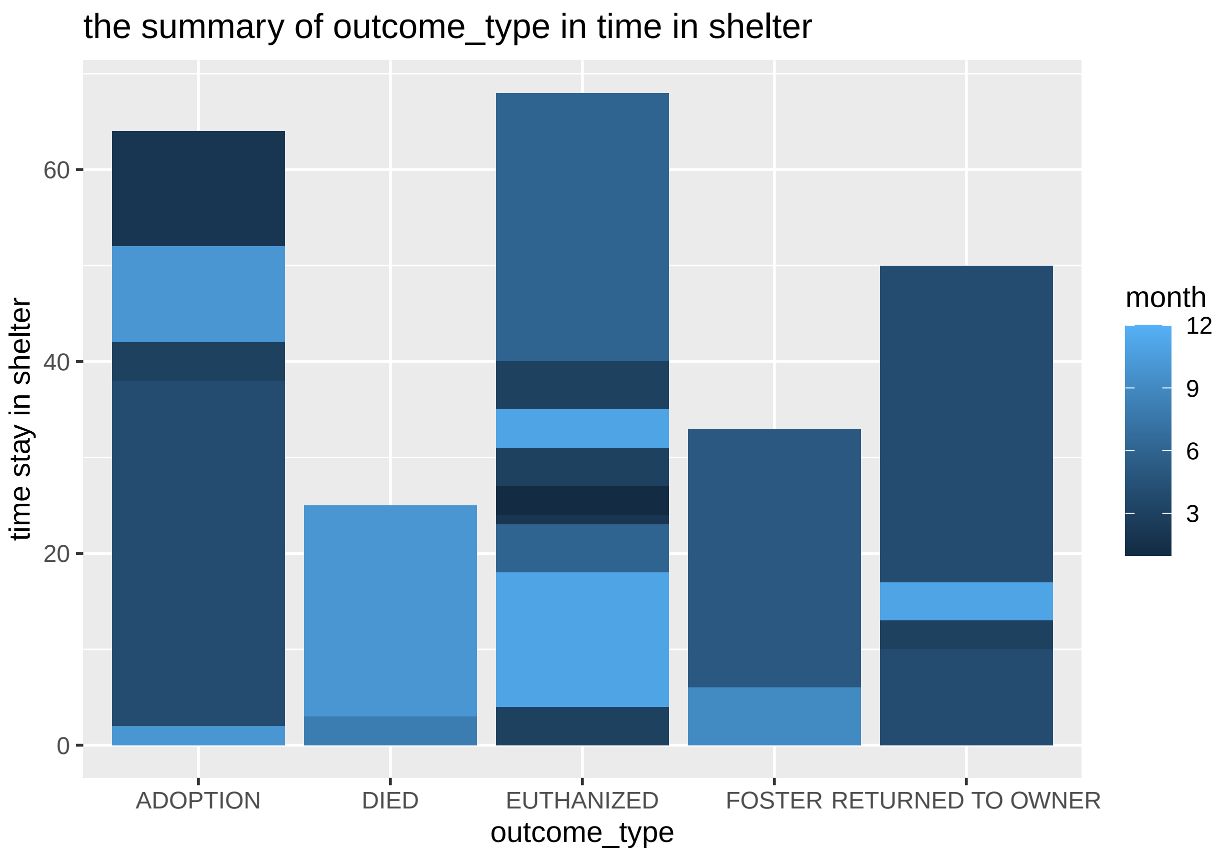
After visualizing the data, we fond that there was no obvious changes but slightly different between the **first half year** and the **second half year** so we may classify time into two parts. (The month from 1 to 6 will be labelled as first half year and the month from 7 to 12 will be labelled as second half year)

```
animals_used<-animals%>%
  mutate(time = ifelse(month > 6, "second half year", "first half year"))
```

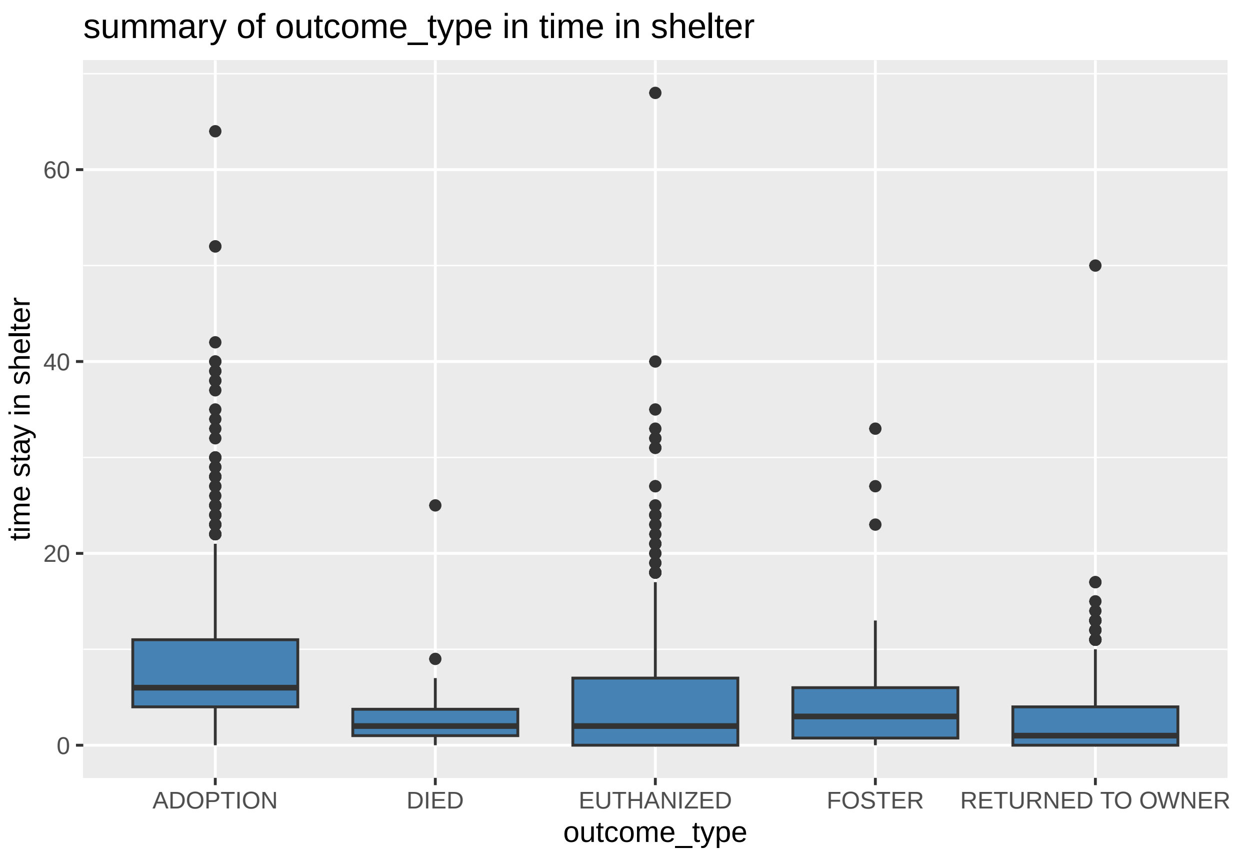
Observing the time that animals stay in shelter related to intake type.

Observing the time that animals stay in shelter related to outcome type.

```
ggplot(data = animals, mapping = aes(x = factor(outcome_type), y = time_at_shelter, fill = month)) +
  geom_col(position = "dodge") +
  labs(x = "outcome_type", y = "time stay in shelter",
       title = "the summary of outcome_type in time in shelter")
```

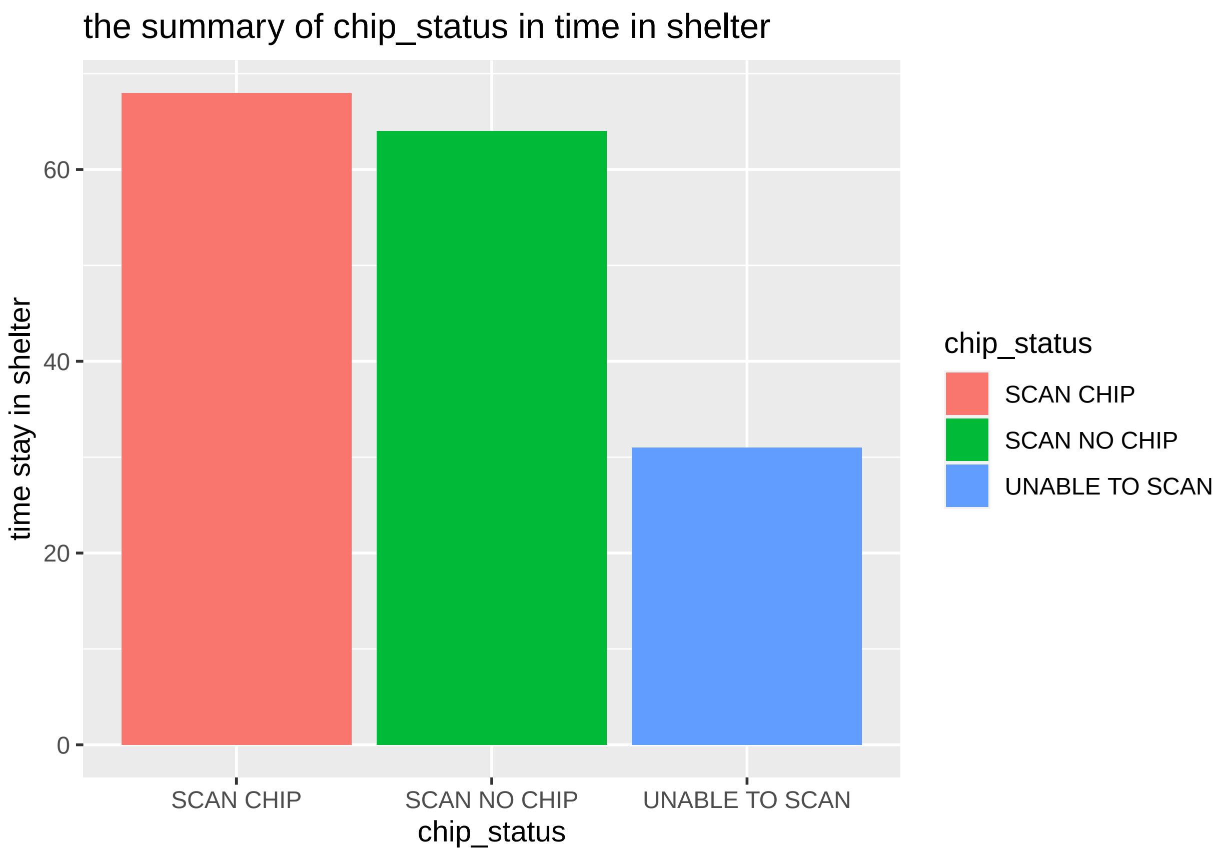


```
ggplot(data = animals, mapping = aes(x = factor(outcome_type), y = time_at_shelter)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "outcome_type", y = "time stay in shelter",
       title = "summary of outcome_type in time in shelter")
```

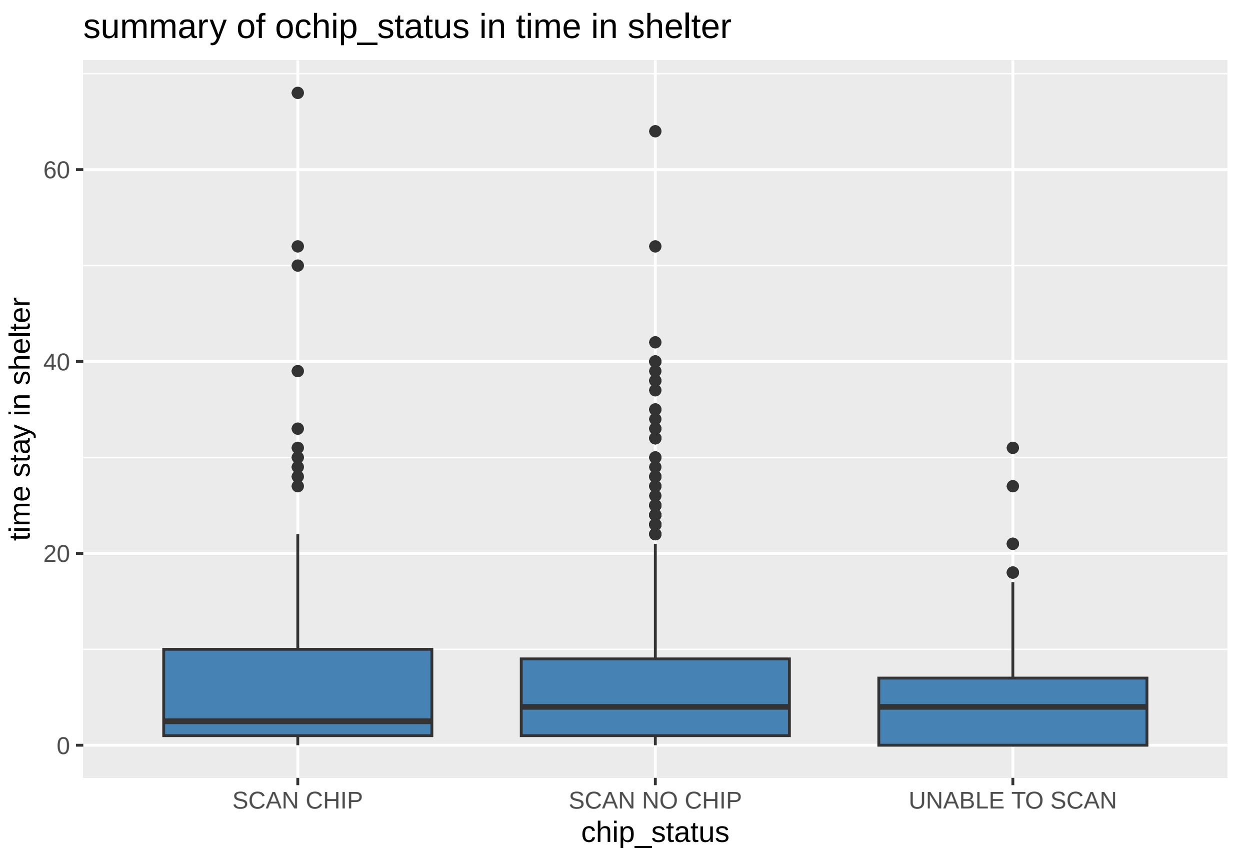


Observing the time that animals stay in shelter related to chip status.

```
ggplot(data = animals, mapping = aes(x = factor(chip_status), y = time_at_shelter, fill = chip_status))
  geom_col(position = "dodge") +
  labs(x = "chip_status", y = "time stay in shelter",
       title = "the summary of chip_status in time in shelter")
```



```
ggplot(data = animals, mapping = aes(x = factor(chip_status), y = time_at_shelter)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "chip_status", y = "time stay in shelter",
       title = "summary of ochip_status in time in shelter")
```



3 Model Analysis

After observing all the variables, we fond that there was no obvious linear relationship. But as we mentioned before, we can use **GLM** method with family=poisson(link="log") for fitting the data. And the formula we will use is given by:

$$y_i$$

$$\sim$$

$$Poisson\left(\lambda_i\right)$$

$$log\left(\lambda_i\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji}$$

Where

$$\lambda_i = exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji}\right)$$

and

$$\frac{\lambda_{ji}}{\lambda_0} = exp\left(\beta_j\right)$$

where λ_{ji} is the value of λ_i when only x_{ji} is 1 and λ_0 is the exp of β_0.

Since all the explanatory variables are characteristic, the exp of β_j means that the mean of y_i will be e^{β_j} times of the mean where the j_{th} factor changes from 0 to 1.

As a result, if the exp of coefficient is closer to 1, the more unlikely the factor will affect the response term.

We first try the full model for fitting the whole data.

```
animals_used<-animals_used%>%
  select(-c("month"))
animals_used$year<-as.character(animals_used$year)
mod.loglinear <- glm(time_at_shelter ~ year + animal_type + intake_type + outcome_type + chip_status + t
summary(mod.loglinear)
```

```
##
## Call:
## glm(formula = time_at_shelter ~ year + animal_type + intake_type +
##     outcome_type + chip_status + time, family = poisson(link = "log"),
##     data = animals_used)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5705  -2.1015  -0.9003   0.5869  13.2700
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   1.47966    1.00153   1.477 0.139568
## year2017                     -0.07657    0.02645  -2.895 0.003788 **
## animal_typeCAT                1.99820    1.00051   1.997 0.045805 *
## animal_typeDOG                2.15471    1.00033   2.154 0.031241 *
## animal_typeWILDLIFE           1.72568    1.00668   1.714 0.086488 .
## intake_typeOWNER SURRENDER   -1.47194    0.03623 -40.632  < 2e-16 ***
## intake_typeSTRAY             -1.09020    0.03221 -33.842  < 2e-16 ***
## outcome_typeDIED             -0.87456    0.11427  -7.654 1.95e-14 ***
## outcome_typeEUTHANIZED       -0.65537    0.02199 -29.803  < 2e-16 ***
## outcome_typeFOSTER           -0.39429    0.06788  -5.809 6.30e-09 ***
## outcome_typeRETURNED TO OWNER -1.47359    0.03594 -41.004  < 2e-16 ***
## chip_statusSCAN NO CHIP      -0.19965    0.02679  -7.451 9.25e-14 ***
## chip_statusUNABLE TO SCAN    -0.18520    0.05460  -3.392 0.000694 ***
## timesecond half year         -0.05483    0.02295  -2.389 0.016913 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13111.4  on 1852  degrees of freedom
## Residual deviance:  9973.7  on 1839  degrees of freedom
## AIC: 15185
##
## Number of Fisher Scoring iterations: 6
```

```
AIC(mod.loglinear)
```

```
## [1] 15184.51
```

We fond that collinearity exists between different animal types. We then check the data grouped by animal type and fond that there were only 2 rows of data for **BIRD**. Since we know that small size of data would cause collinearity, we decided to include the data for BIRD into WILDLIFE.

```
animals_used<-animals_used%>%
  filter(animal_type != "BIRD")
```

Then we use the new version of data for building the model.

```
mod.loglinear_alt <- glm(time_at_shelter ~ year + animal_type + intake_type + outcome_type + chip_statu
summary(mod.loglinear_alt)
```

```
##
## Call:
## glm(formula = time_at_shelter ~ year + animal_type + intake_type +
##     outcome_type + chip_status + time, family = poisson(link = "log"),
##     data = animals_used)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5706  -2.1010  -0.9005   0.5868  13.2710
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   3.47797    0.05575  62.390  < 2e-16 ***
## year2017                     -0.07657    0.02645  -2.895 0.003791 **
## animal_typeDOG                0.15649    0.02528   6.190 6.03e-10 ***
## animal_typeWILDLIFE          -0.27250    0.11432  -2.384 0.017144 *
## intake_typeOWNER SURRENDER   -1.47218    0.03623 -40.637  < 2e-16 ***
## intake_typeSTRAY             -1.09013    0.03221 -33.840  < 2e-16 ***
## outcome_typeDIED             -0.87463    0.11427  -7.654 1.94e-14 ***
## outcome_typeEUTHANIZED       -0.65538    0.02199 -29.803  < 2e-16 ***
## outcome_typeFOSTER           -0.39432    0.06788  -5.809 6.28e-09 ***
## outcome_typeRETURNED TO OWNER -1.47367    0.03594 -41.006  < 2e-16 ***
## chip_statusSCAN NO CHIP      -0.19975    0.02679  -7.455 8.98e-14 ***
## chip_statusUNABLE TO SCAN    -0.18531    0.05460  -3.394 0.000689 ***
## timesecond half year         -0.05481    0.02295  -2.388 0.016937 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13092.4  on 1850  degrees of freedom
## Residual deviance:  9971.9  on 1838  degrees of freedom
## AIC: 15179
##
## Number of Fisher Scoring iterations: 6
```

```
AIC(mod.loglinear_alt)
```

```
## [1] 15178.7
```

We then have a new model where no p values for explanatory variables are greater than 0.05.

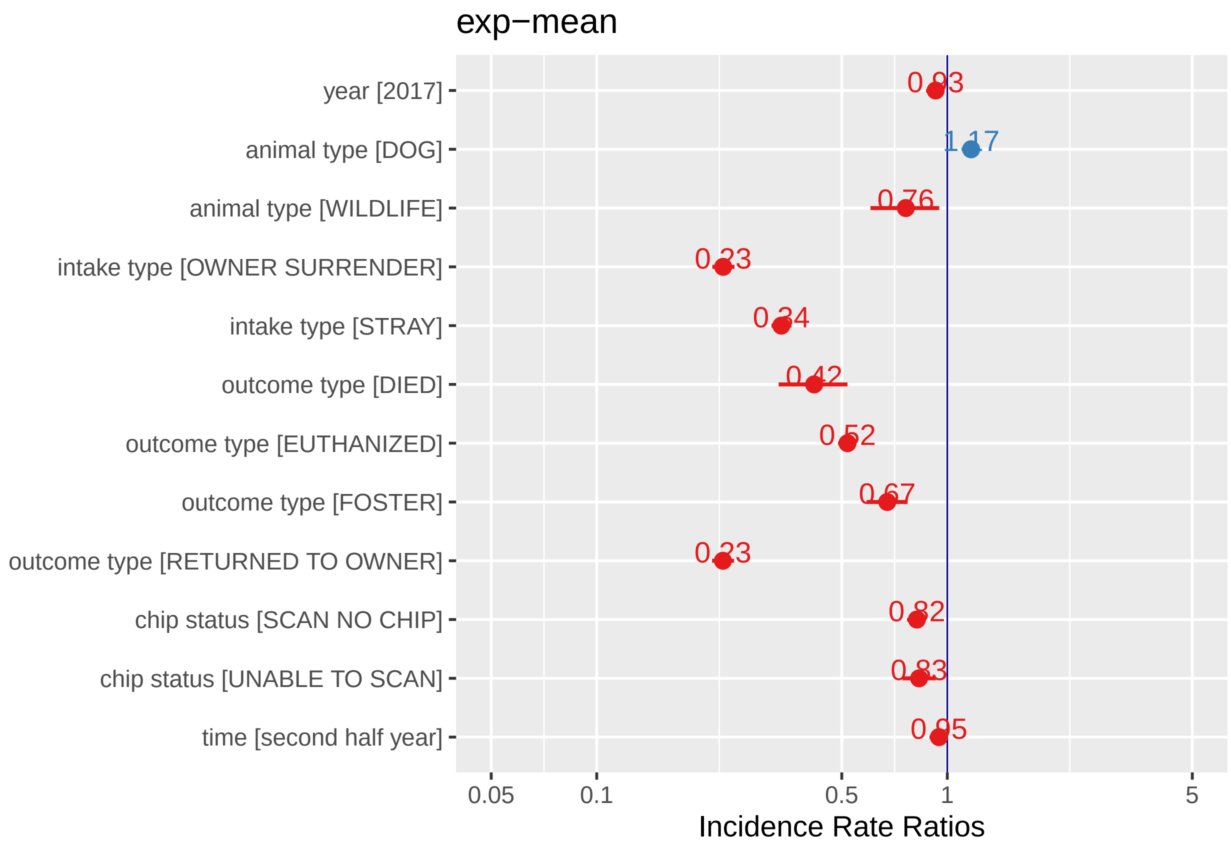
We can now have a look at the exp of coefficient plotted and listed below:

```
confint(mod.loglinear_alt)
```

```
## Waiting for profiling to be done...
```

```
##                                     2.5 %       97.5 %
## (Intercept)                     3.3684538  3.586977507
## year2017                       -0.1283984 -0.024721725
## animal_typeDOG                  0.1071588  0.206269752
## animal_typeWILDLIFE            -0.5043365 -0.055614885
## intake_typeOWNER SURRENDER     -1.5429209 -1.400900756
## intake_typeSTRAY               -1.1528764 -1.026589780
## outcome_typeDIED               -1.1069610 -0.658437904
## outcome_typeEUTHANIZED         -0.6985896 -0.612386564
## outcome_typeFOSTER             -0.5299808 -0.263774974
## outcome_typeRETURNED TO OWNER -1.5445880 -1.403702244
## chip_statusSCAN NO CHIP        -0.2520519 -0.147015245
## chip_statusUNABLE TO SCAN      -0.2934223 -0.079348945
## timesecond half year           -0.0999038 -0.009924671
```

```
plot_model(mod.loglinear_alt, show.values = TRUE, title = "exp-mean", show.p = FALSE,vline.color = "dar
```



```
mod.loglinear_alt %>%
   coef() %>%
   exp()
```

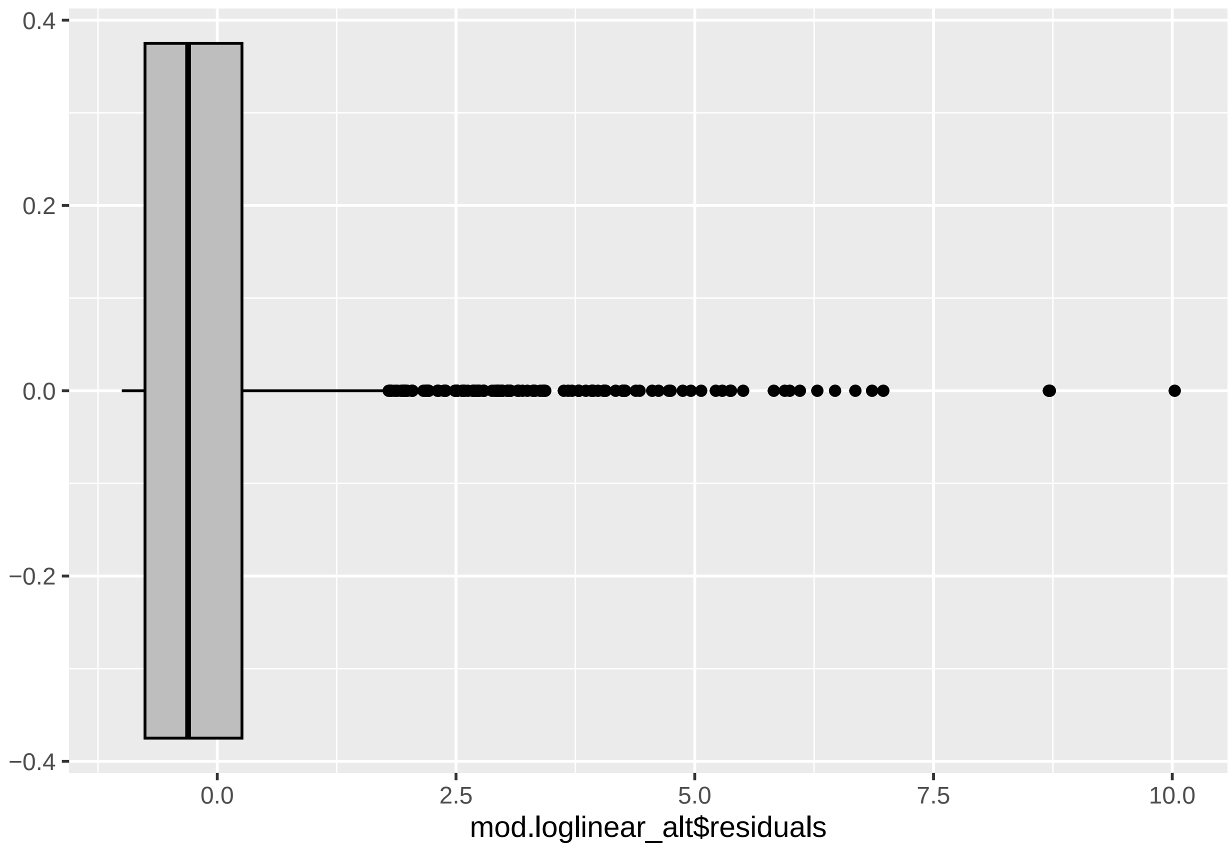
```
##                   (Intercept)                      year2017 
##                    32.3940141                     0.9262924 
##               animal_typeDOG            animal_typeWILDLIFE 
##                     1.1694028                     0.7614738 
##    intake_typeOWNER SURRENDER              intake_typeSTRAY 
##                     0.2294241                     0.3361741 
##              outcome_typeDIED        outcome_typeEUTHANIZED 
##                     0.4170183                     0.5192463 
##            outcome_typeFOSTER outcome_typeRETURNED TO OWNER 
##                     0.6741402                     0.2290823 
##        chip_statusSCAN NO CHIP     chip_statusUNABLE TO SCAN 
##                     0.8189335                     0.8308505 
##          timesecond half year 
##                     0.9466609
```

4 Conclusion

We then conclude that the factors: intake type and outcome type affect the time at shelter the most.

Residuals can be seen as follow the assumption. Actually, the residuals should follow the assumptions theoretically by the theory of GLM.

```
animals_used <- animals_used %>%
  mutate(counts_pred = fitted(mod.loglinear_alt))
p1<-ggplot(mod.loglinear_alt, aes(x = mod.loglinear_alt$residuals))
p1 + geom_boxplot(color = "black", fill = "gray")
```


0.4
0.2
0.0
−0.2
−0.4
0.0
2.5
5.0
7.5
10.0
mod.loglinear_alt$residuals