DAS2021-Group-19

Xinyi Gao, Yiyang Li, Damien MacFarland, Neha Sinha, Jinda Zhang

1 Introduction

1.1 Dataset 19

Dataset 19 comes from the Dallas animal shelter. You will have access to the following variables, recorded by animal admission:

- `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_type` – 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.

1.2 Task

Imagine you have been asked by the shelter management to investigate the following questions of interest:

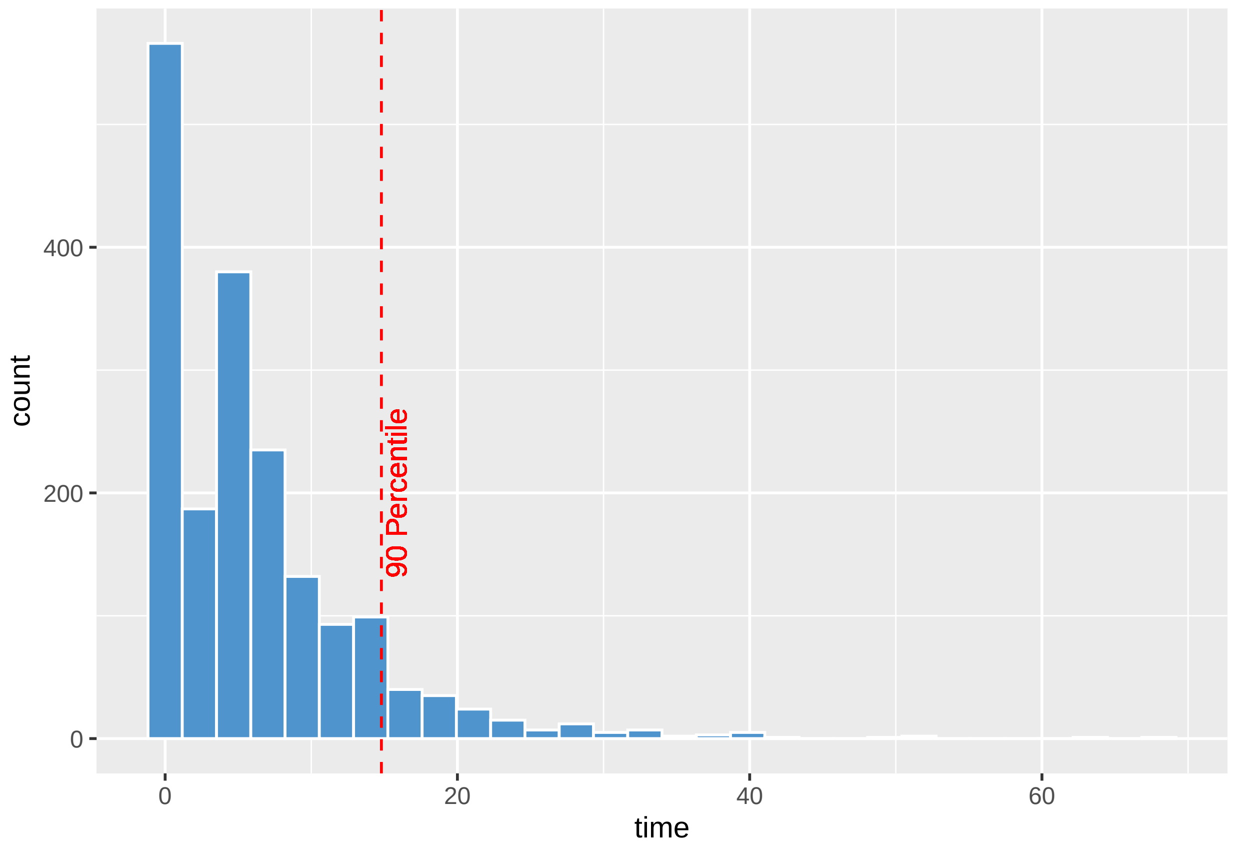
- Which factors influence the number of days an animal spends in the shelter before their final outcome is decided?

You should conduct an analysis to answer your question using a Generalised Linear Model (GLM). Following your analyses, you should then summarise your results in the form of a presentation.

2 Exploratory Data Analysis

2.1 Tidying Data

```
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,~
```



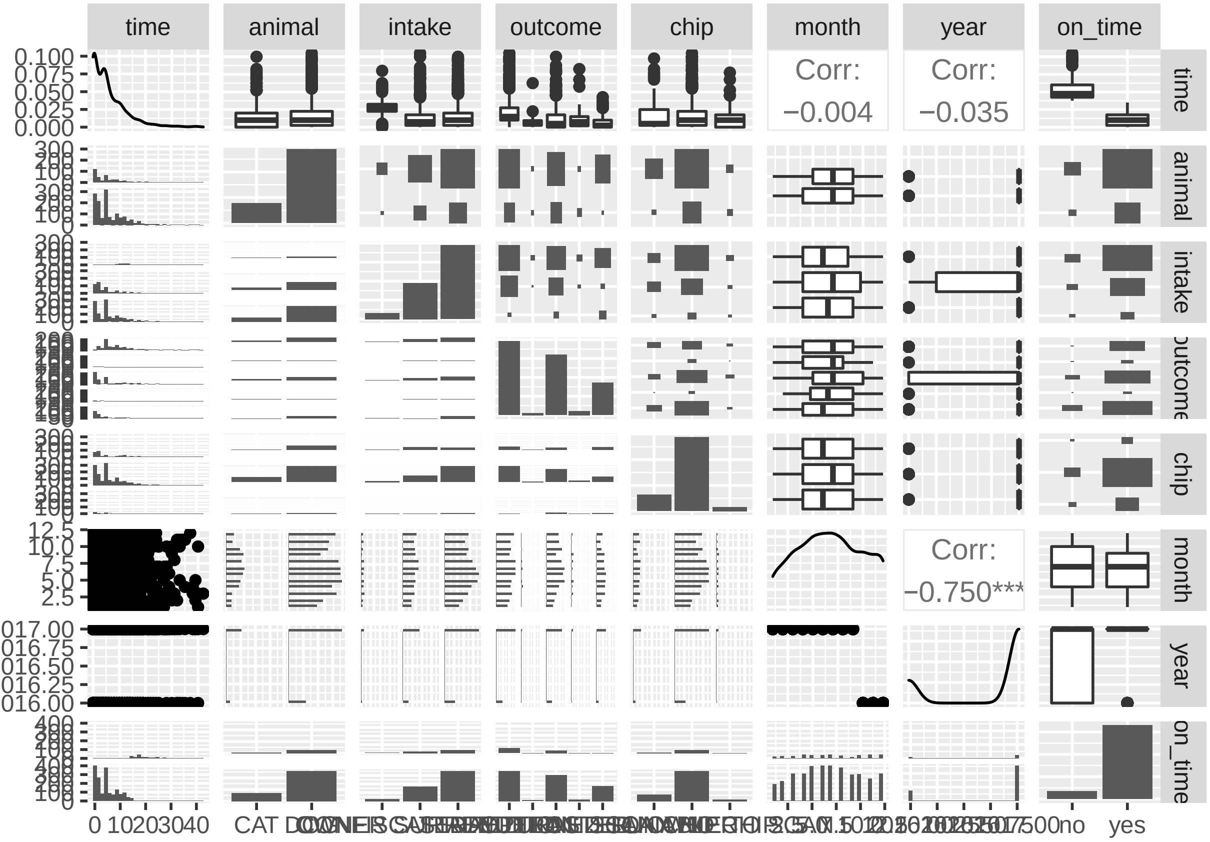
```
      mean      s.d min Q1 med Q3 max
1 6.132758 7.130583   0  1   4  9  68
```

The time spent in the shelter is skewed - most animals (90%) stay no longer than two weeks and very few after three. The majority seem leave the shelter within one week. We could speculate that there are factors (possibly categorical variables) that contribute to the time at the shelter being shorter, and therefore if these factors do not occur then the animal stays longer.

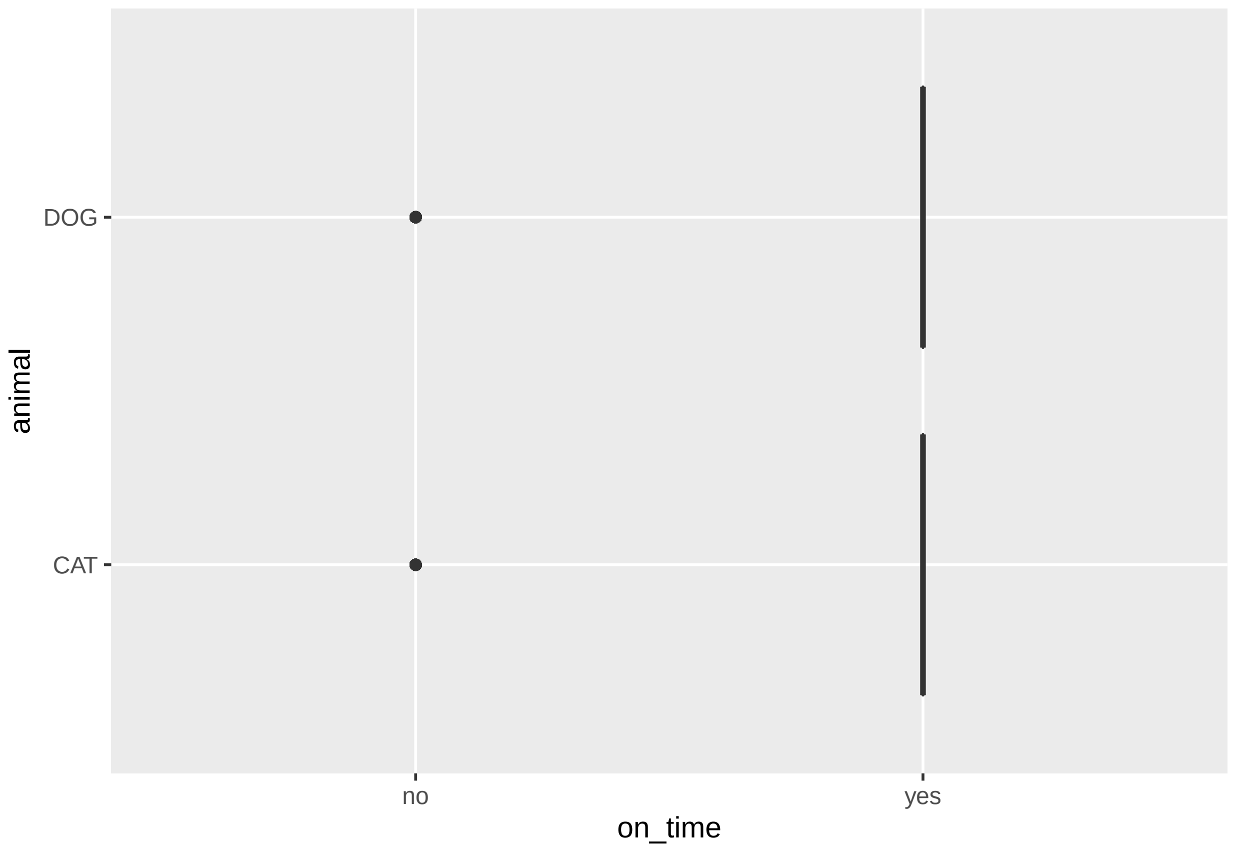
We see that 90% of animals' outcome is decided by 14.8 days at the shelter - this could be our research: what makes animals stay beyond 14 days?

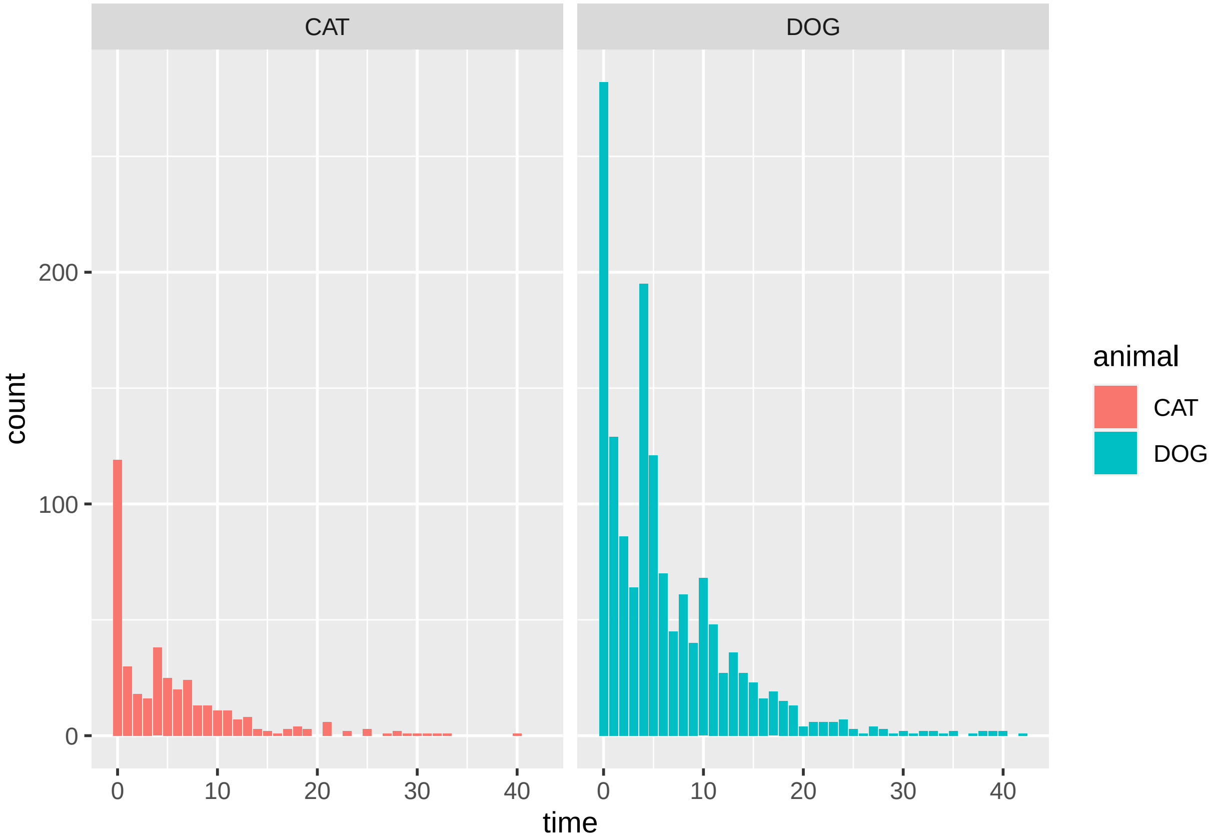
If we consider staying longer than 14 days as a failure, then we can reduce our problem to a binary GLM.

There are only two birds in the data and 13 wildlife. Therefore the impact these animal types will have on the other animal types will be minimal as they do not even contribute to 0.01% of the data, so they can be removed.

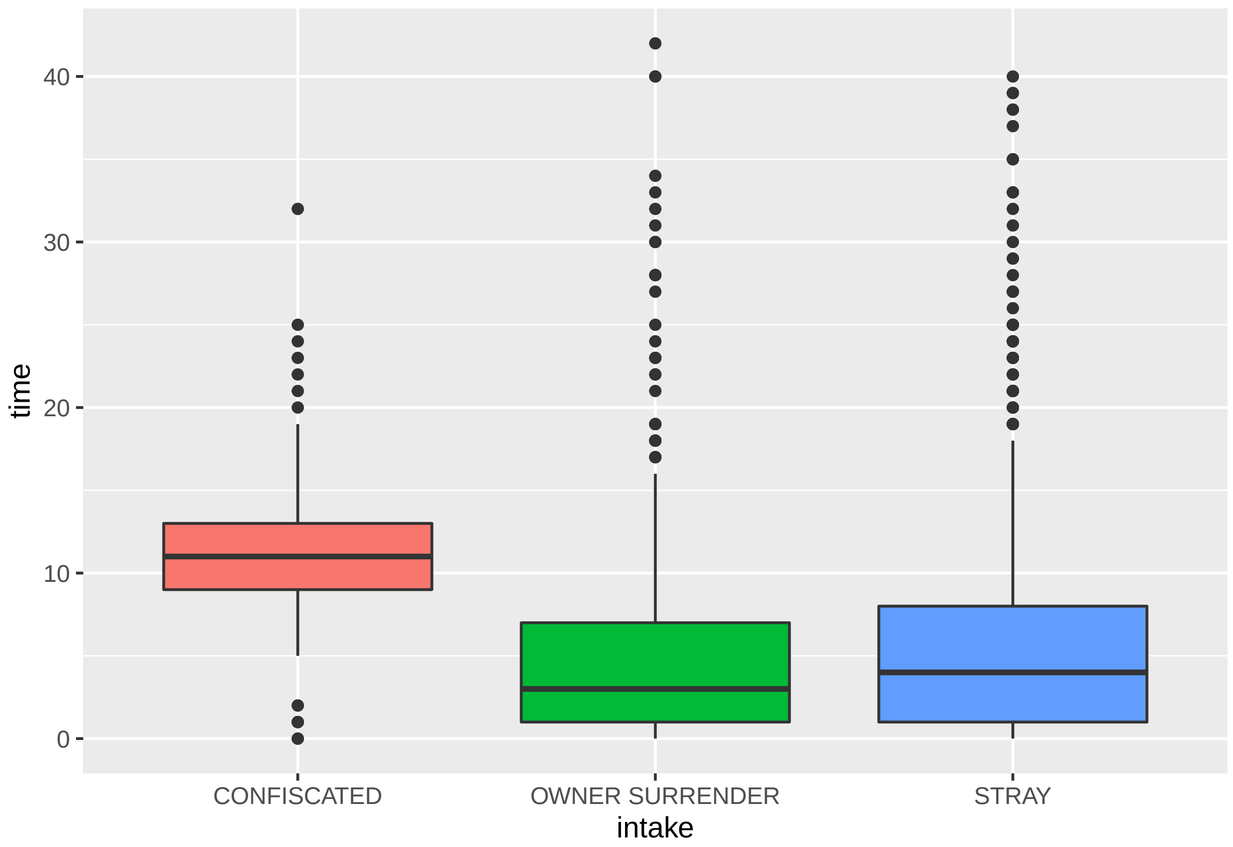


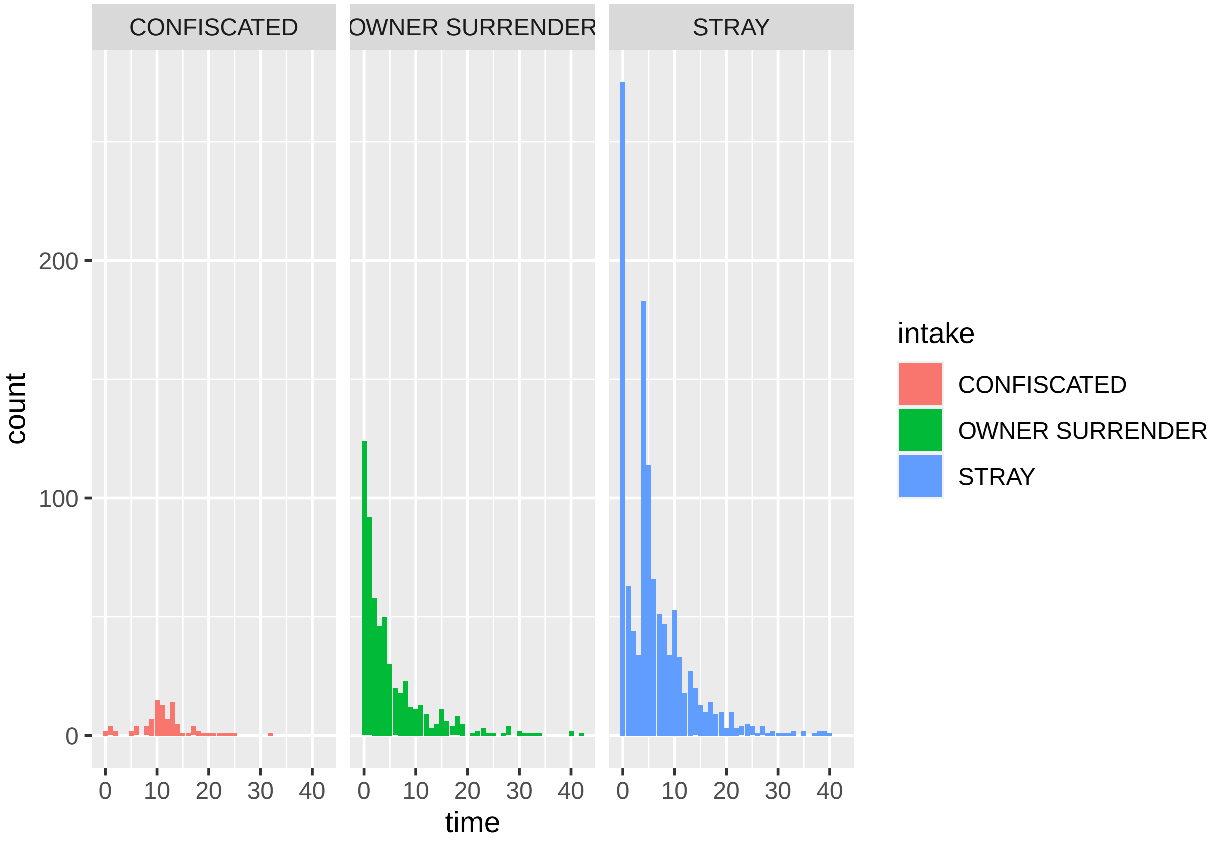
## Time response variable


DOG
CAT
animal
no
yes
on_time

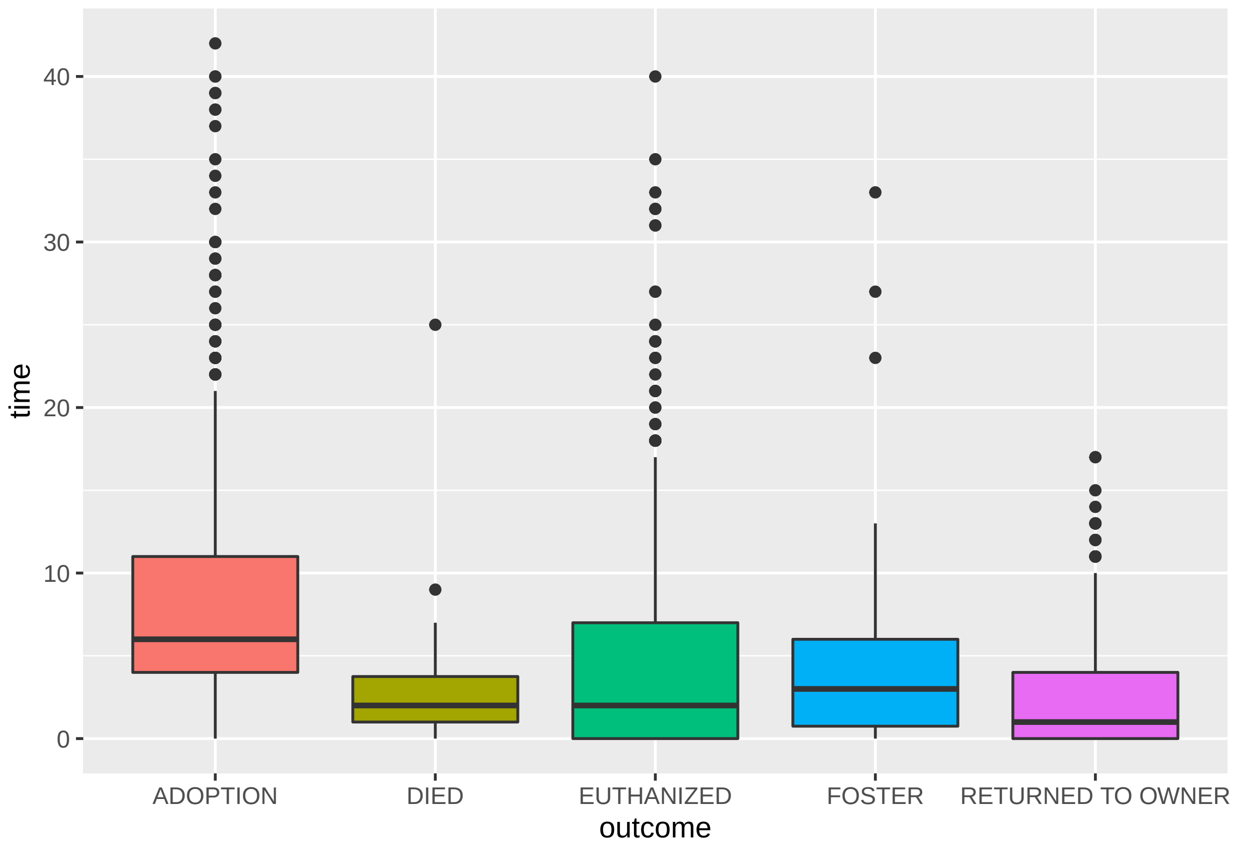


It would seem that there is a slight difference in how long dogs and cats spend at the shelter by looking at the histograms, but from the boxplots we can see that both animals have very similar distributions: the medians are in similar places, their ranges are the similar, both have long tails (skewed right) and lots of outliers. It could be possible that this variable will be omitted from the final model.

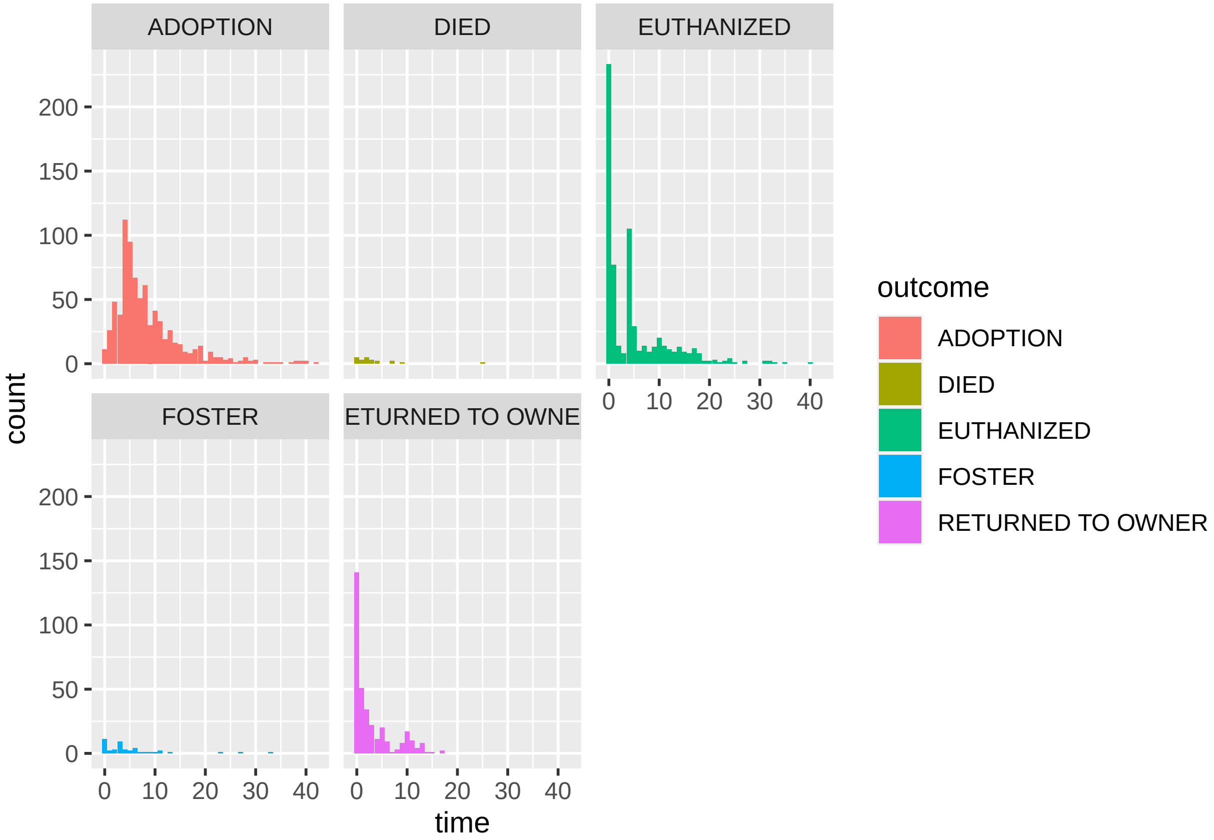

time
40
30
20
10
0
CONFISCATED
OWNER SURRENDER
STRAY
intake



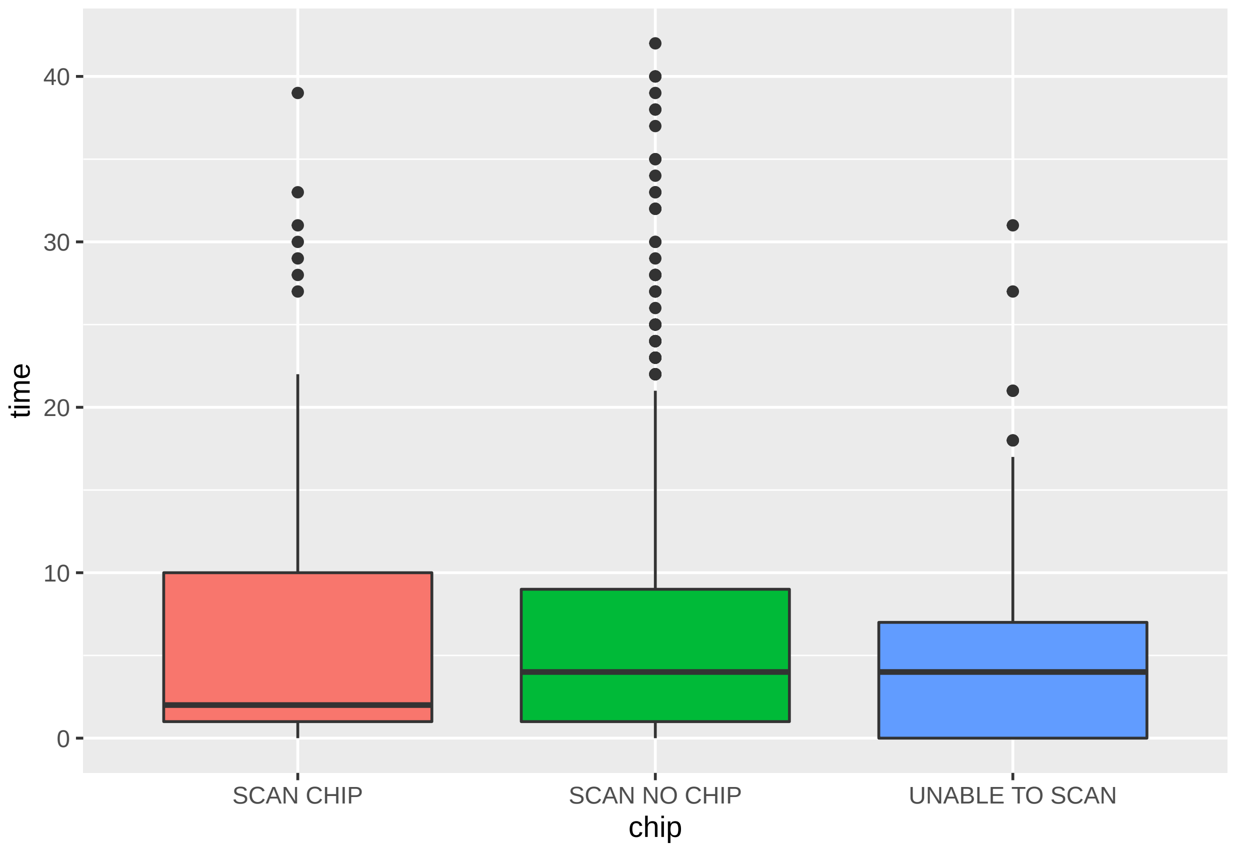
Confiscated animals seems to stay longer at the shelter than those animals that are surrendered by their owner or found as strays. There is a clear separation of one group from the other two. Strays seem to stay slighlty longer than animals that are surrendered.

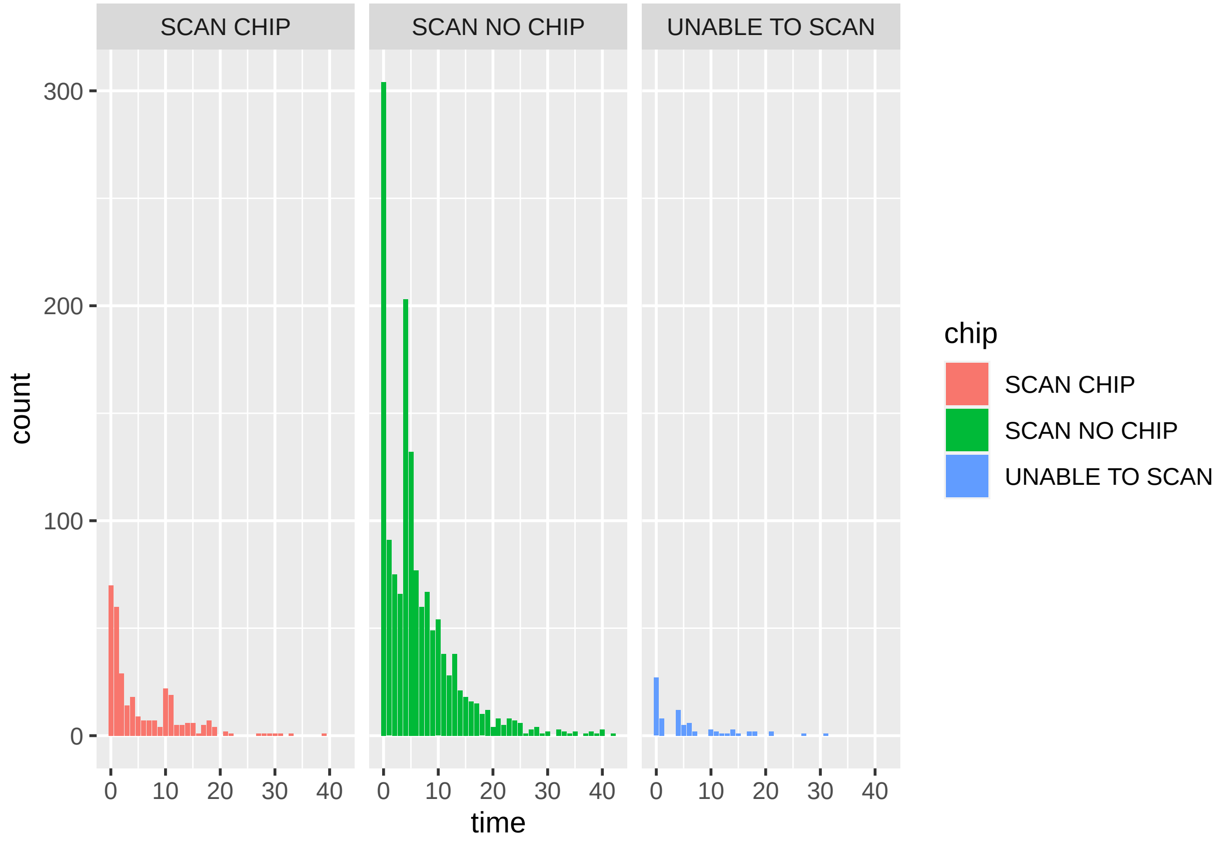


time
40
30
20
10
0
ADOPTION
DIED
EUTHANIZED
FOSTER
RETURNED TO OWNER
outcome

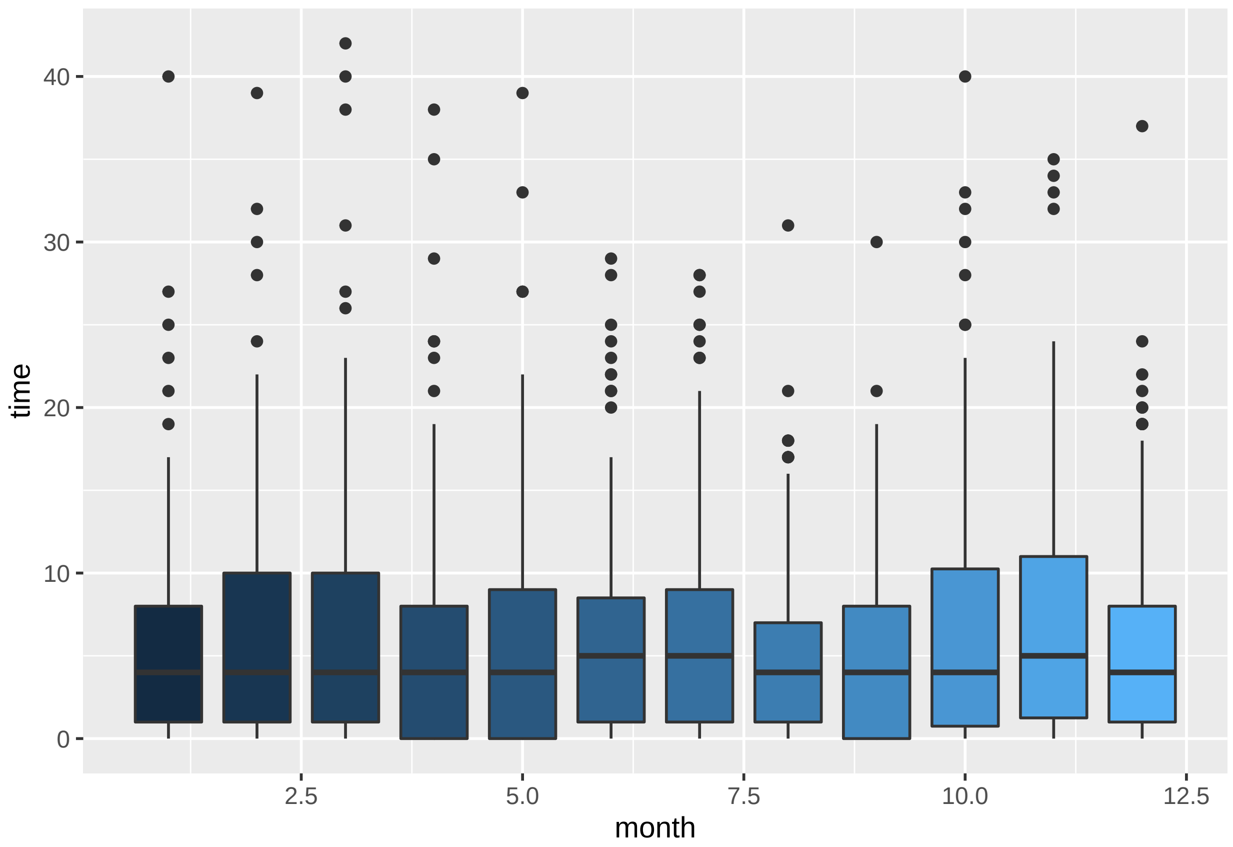


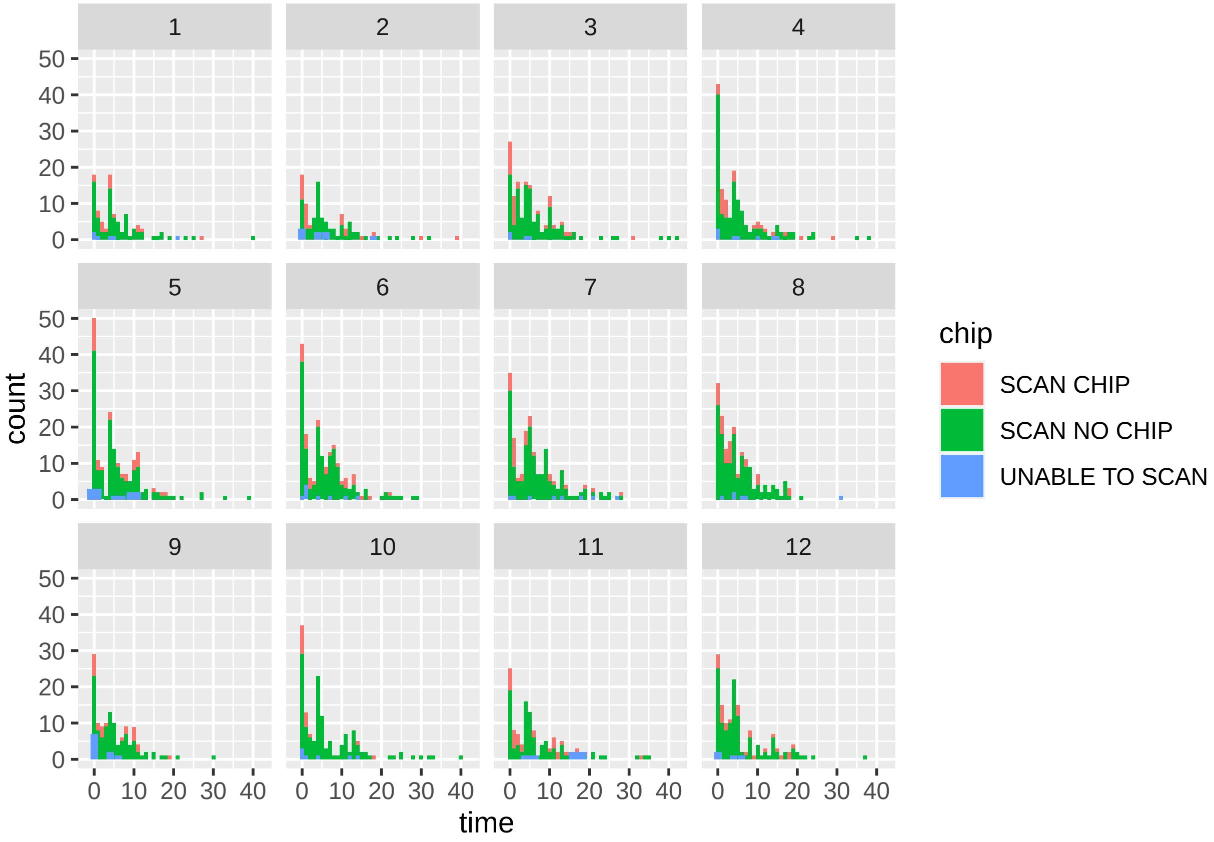
There is some variation across outome - animals returned to owner or euthanized tend to spend short times at the shelter. Those that are adopted tend to be at the shelter for a slightly longer time and the distribution of this subgroup is more normally distributed. There is less data for died and fostered to make an assumption.


time
40
30
20
10
0
SCAN CHIP
SCAN NO CHIP
UNABLE TO SCAN
chip

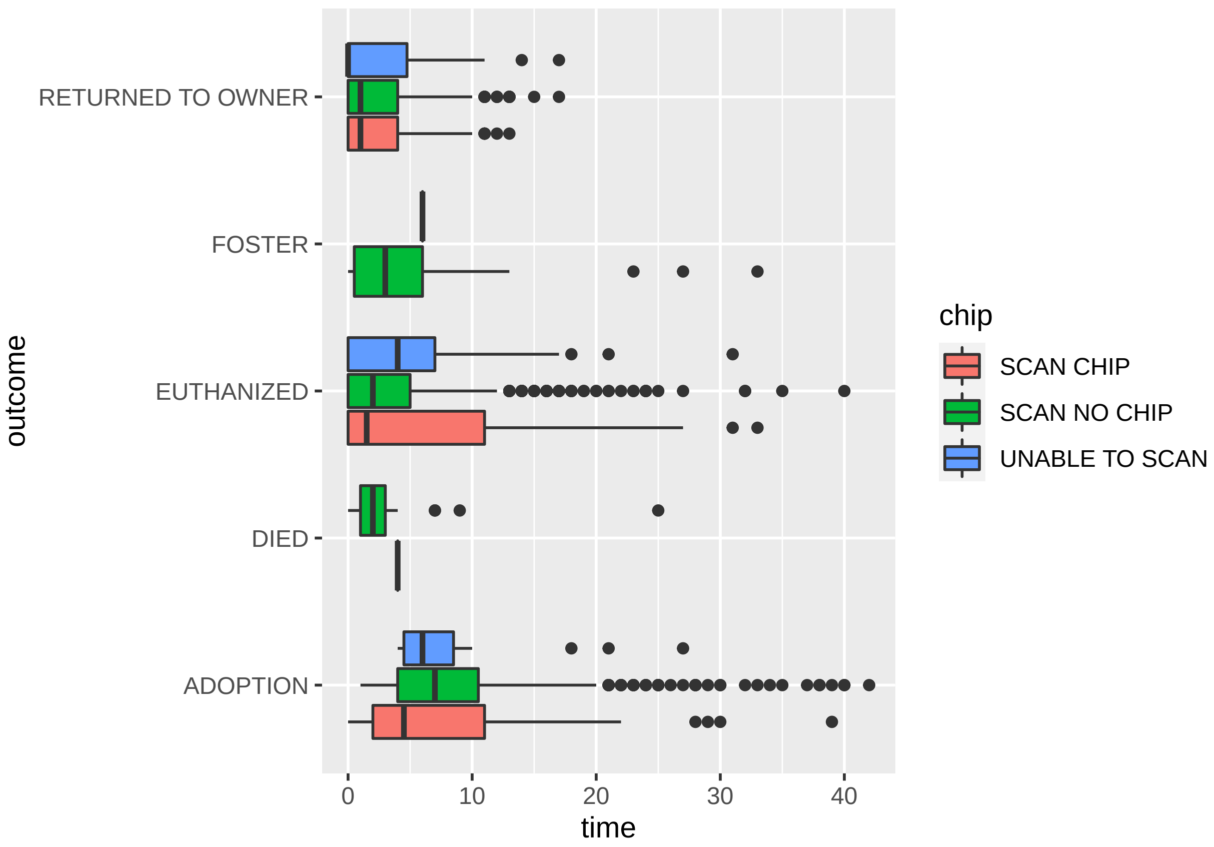


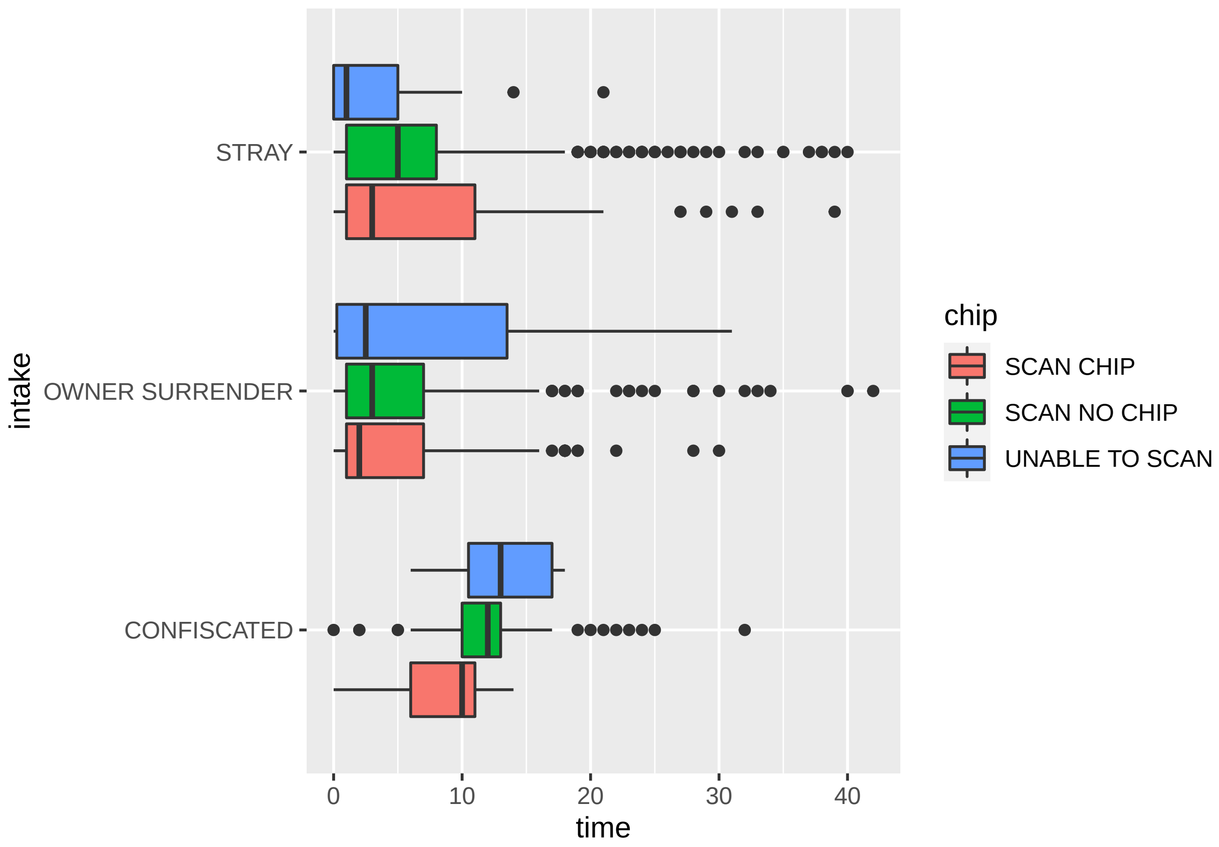
Again all the data is skewed to the right, meaning most types stay for a short period of time. However, it would seem that animals with a chip is a lot more skewed that those that are not. We could speculate that these animals are rescued early and if not then held onto for longer, thus this variable could be a good predictor if combined with outcome perhaps.

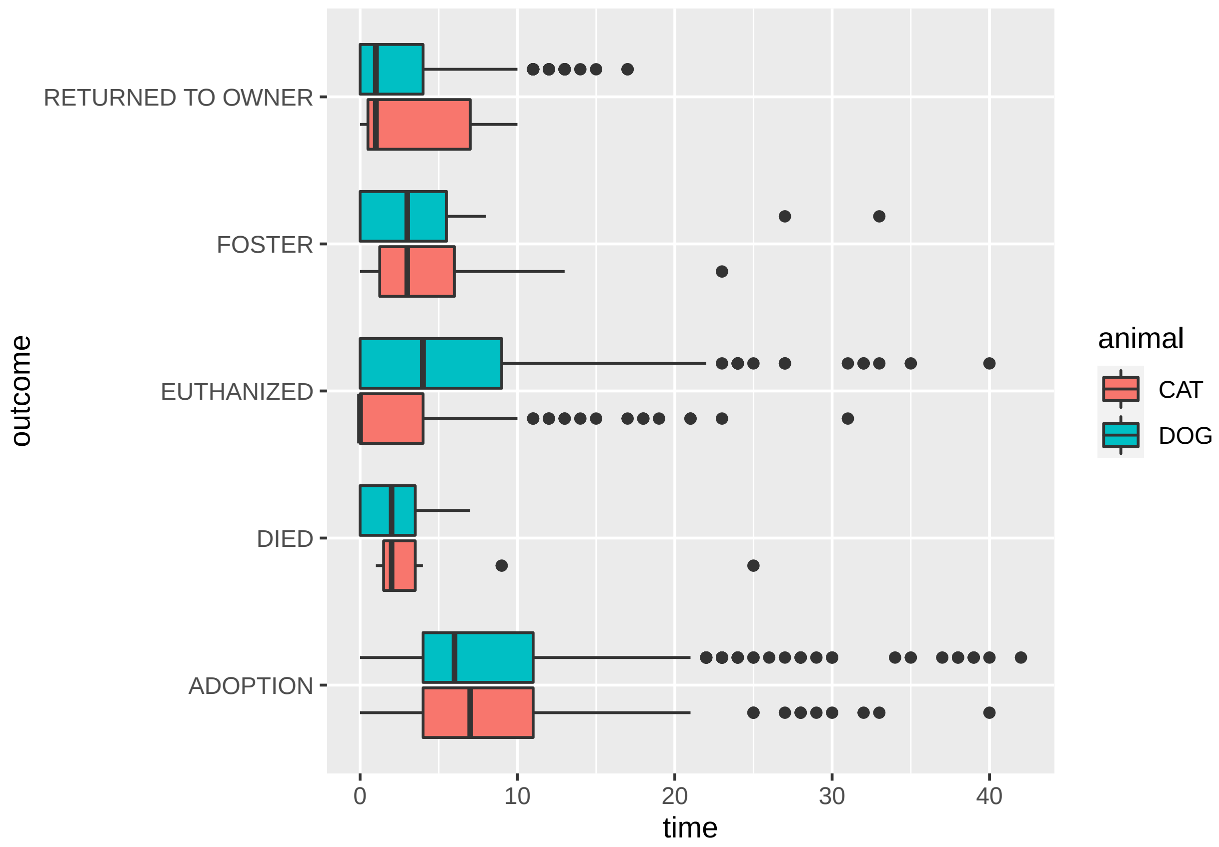

time
40
30
20
10
0
2.5
5.0
7.5
10.0
12.5
month

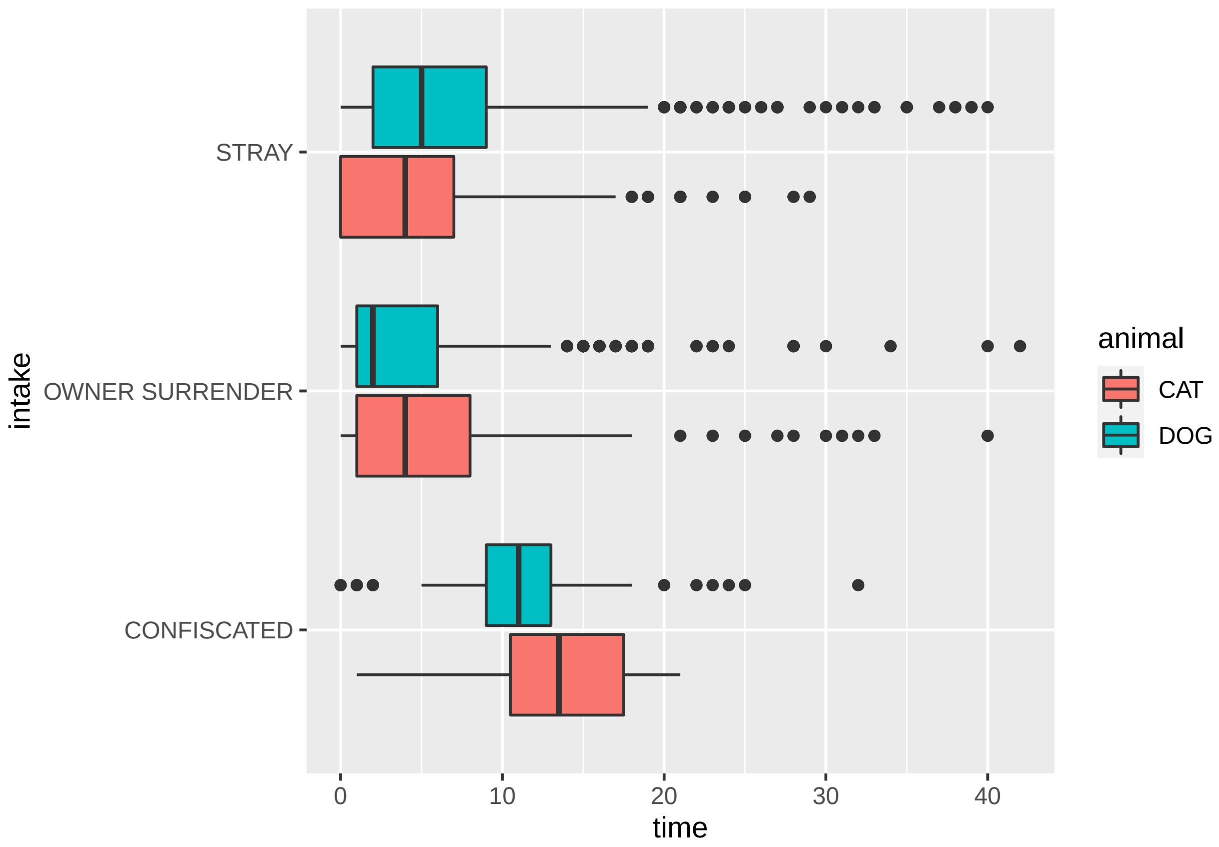


The distributions for month seem to all be similar. We do not envisage that this variable will have an impact on the time variable. Similarly the variable year, which only has 2 levels, may not be a good predictor of time. that means we could potentially have 4 categorical variables explaining one continuous variable - **possible loglinear model**.

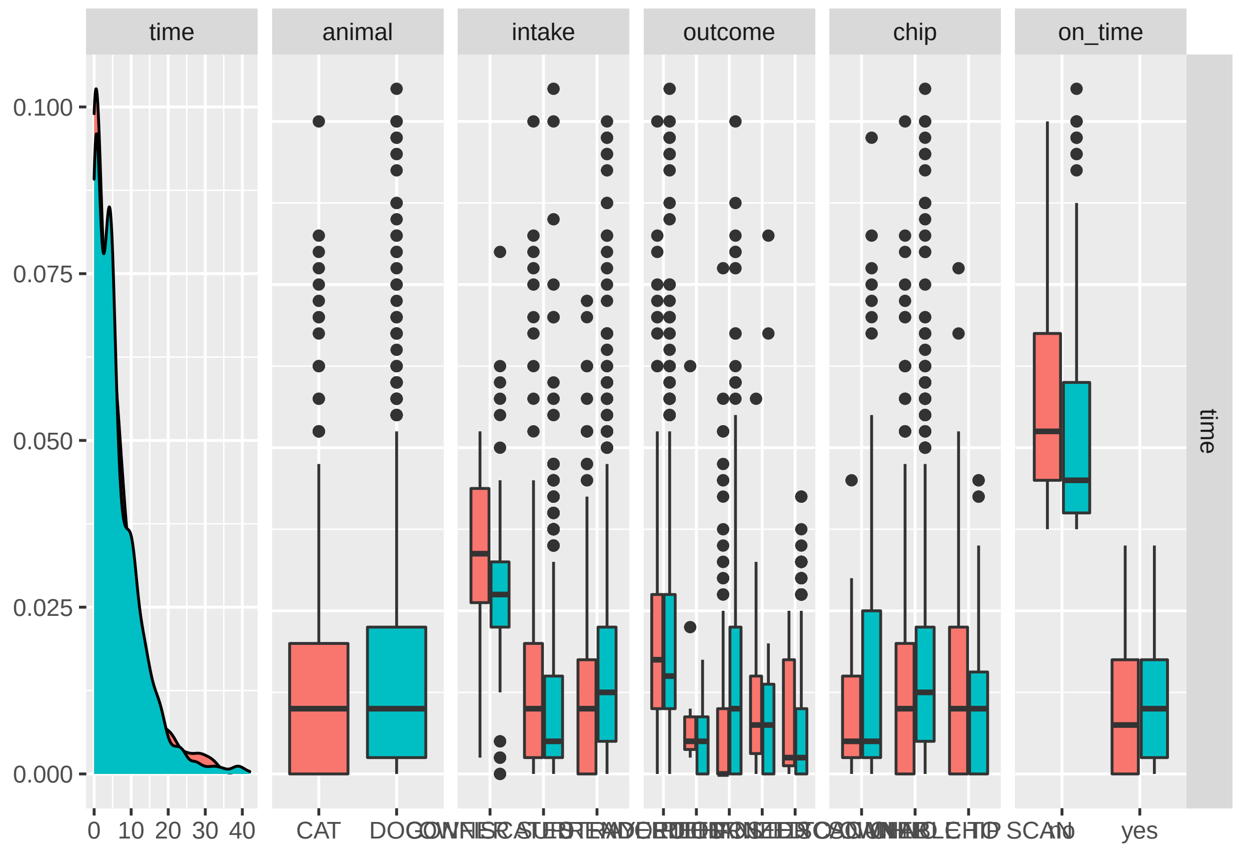

RETURNED TO OWNER
FOSTER
EUTHANIZED
DIED
ADOPTION
outcome
0
10
20
30
40
time
chip
SCAN CHIP
SCAN NO CHIP
UNABLE TO SCAN


STRAY
OWNER SURRENDER
CONFISCATED
intake
0
10
20
30
40
time
chip
SCAN CHIP
SCAN NO CHIP
UNABLE TO SCAN

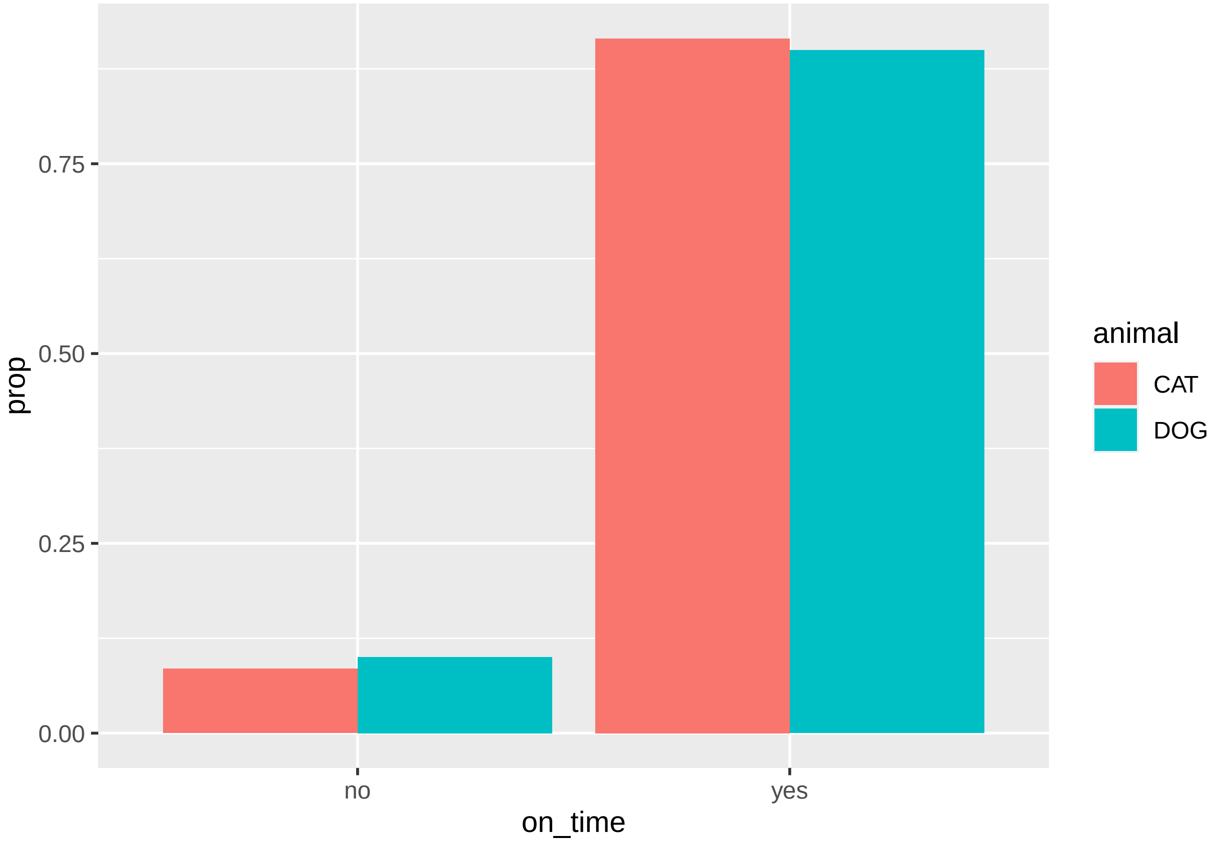

RETURNED TO OWNER
FOSTER
EUTHANIZED
DIED
ADOPTION
outcome
0
10
20
30
40
time
animal
CAT
DOG

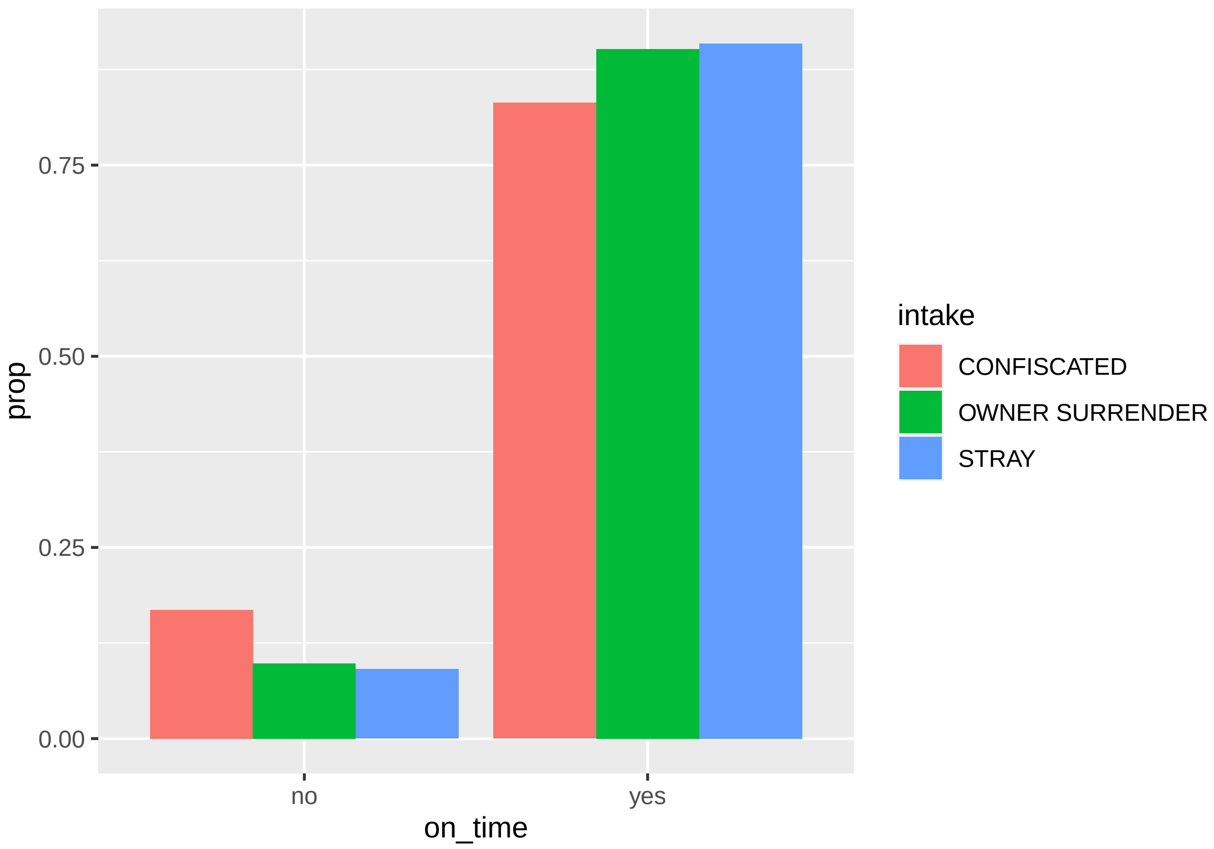


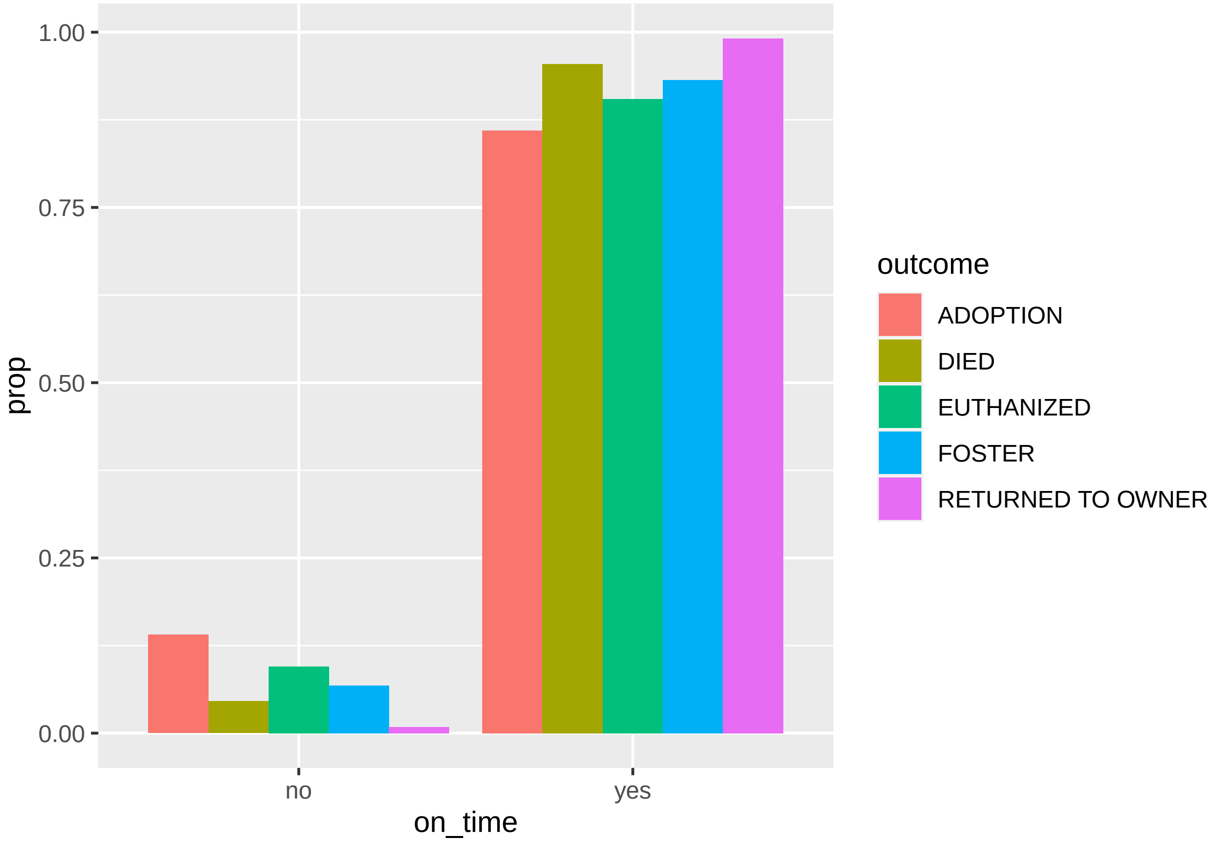
Looking at the above plots, we see some change in distributions across different combinations of categories.

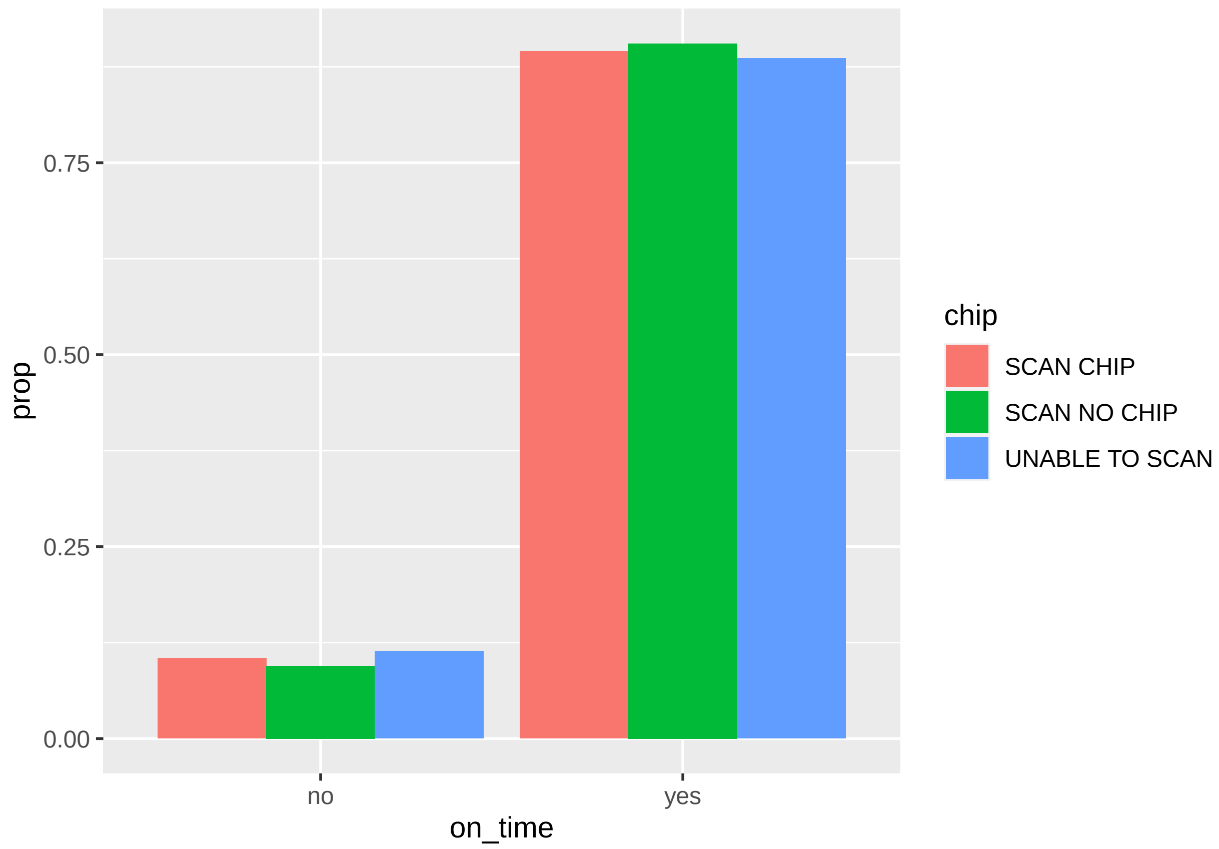


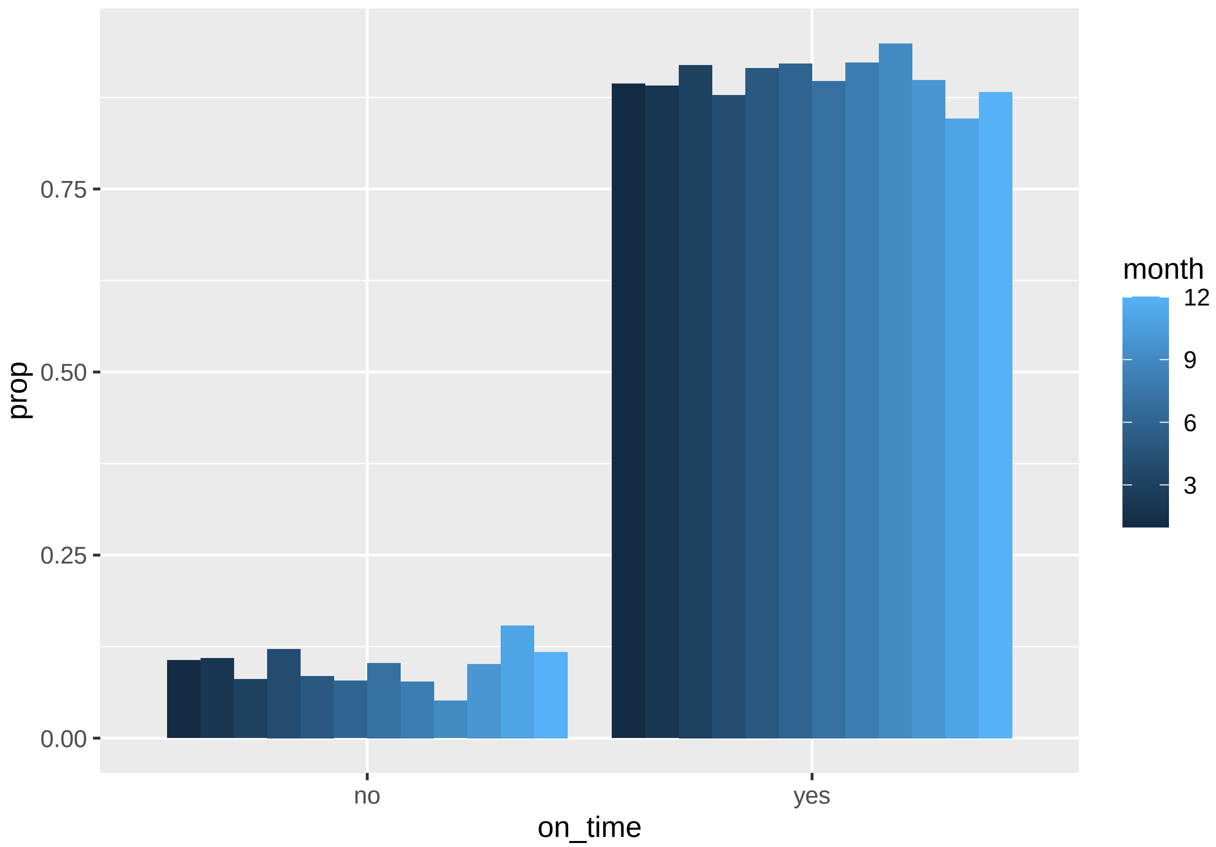
## on_time response variable


0.75
0.50
0.25
0.00
prop
no
yes
on_time
animal
CAT
DOG

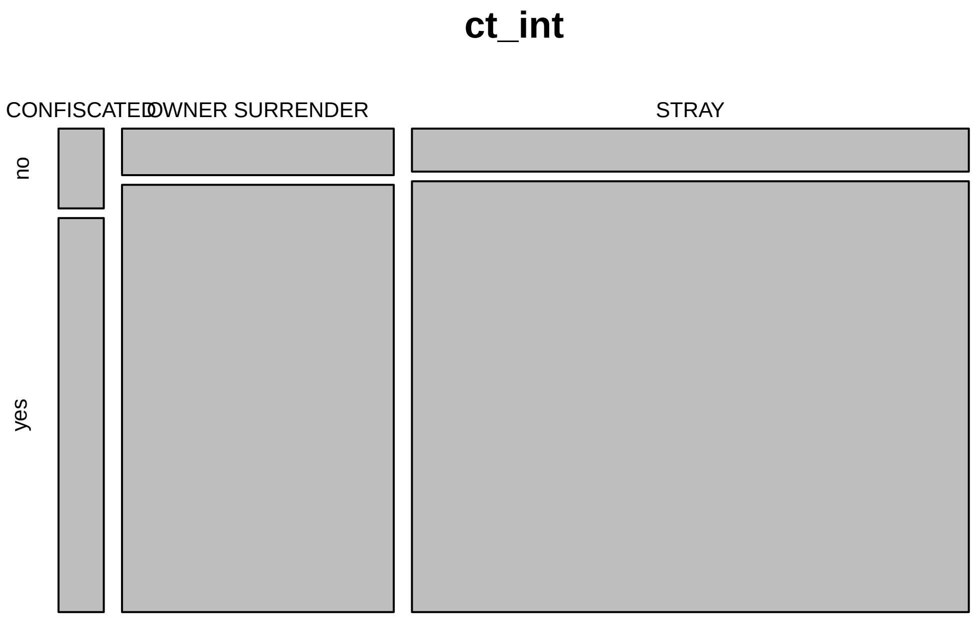

prop
0.00
0.25
0.50
0.75
no
yes
on_time
intake
CONFISCATED
OWNER SURRENDER
STRAY

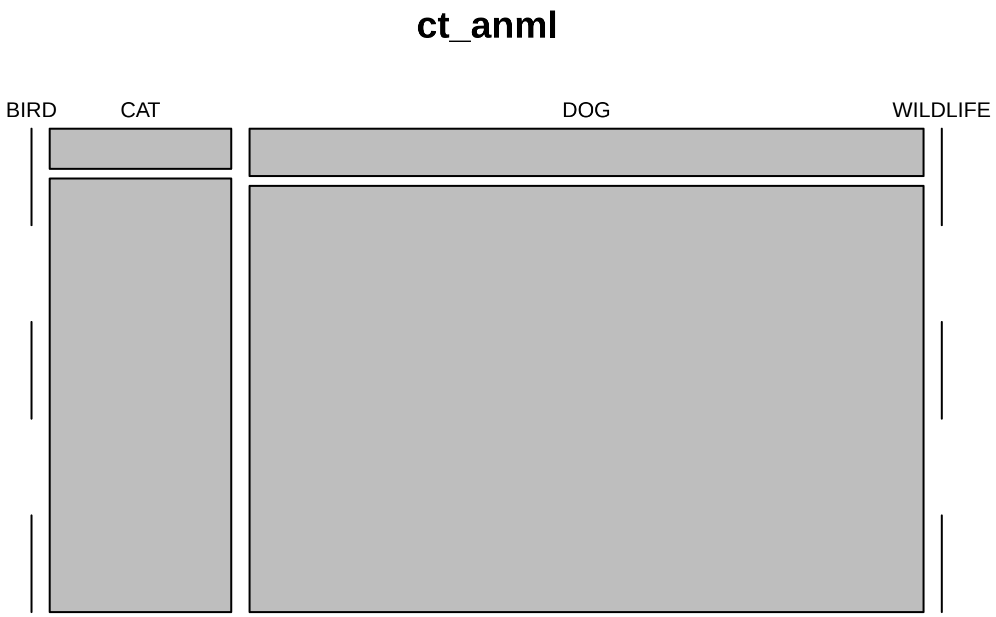

1.00
0.75
0.50
0.25
0.00
prop
no
yes
on_time
outcome
ADOPTION
DIED
EUTHANIZED
FOSTER
RETURNED TO OWNER

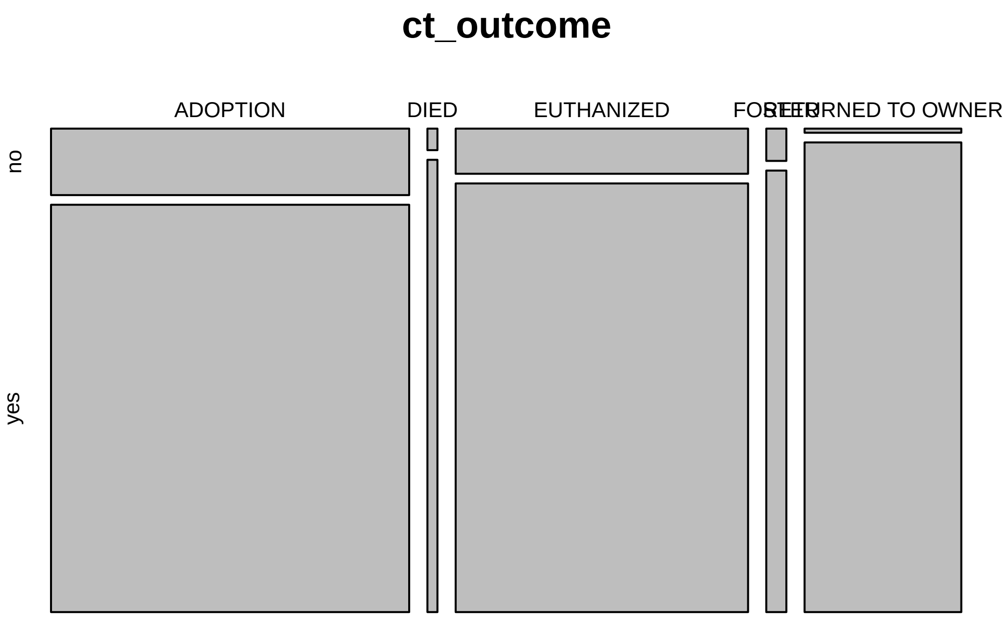

prop
0.00
0.25
0.50
0.75
no
yes
on_time
chip
SCAN CHIP
SCAN NO CHIP
UNABLE TO SCAN

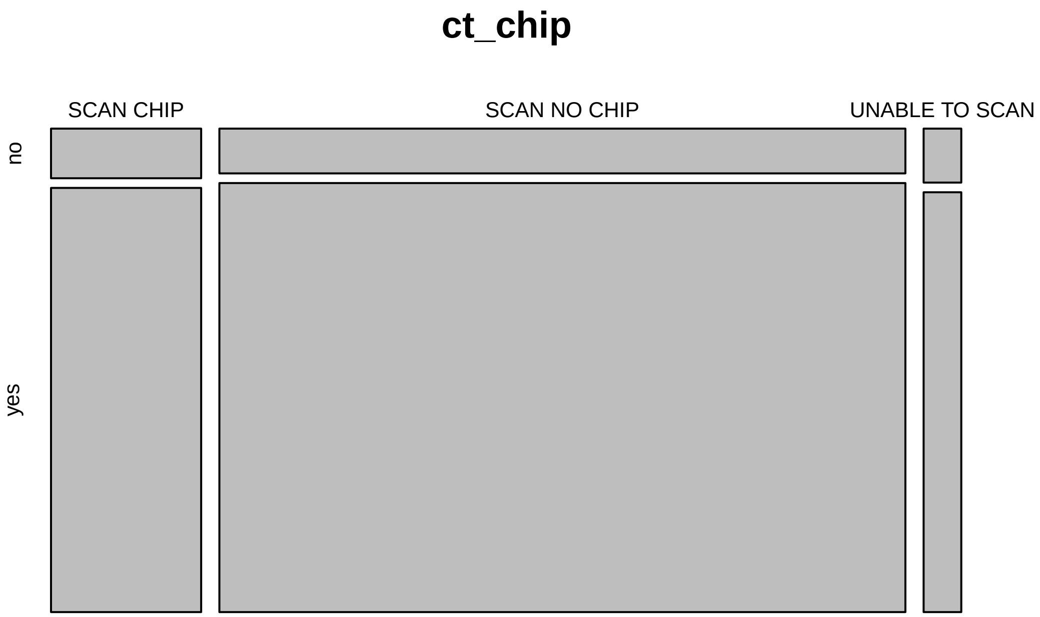


There does not seem to be much difference in proportions between the two types of animals. There is some difference in proportion between confiscated and the other levels of intake. There is some differences from one group to another in the variable outcome. There does not seem to be much differences in proportions for chip. For month, the proportions seem fairly even - cannot see any trend that could relate to seasonal activity.


ct_int
CONFISCATED
OWNER SURRENDER
STRAY
no
yes


ct_anml
BIRD
CAT
DOG
WILDLIFE


ct_outcome
ADOPTION
DIED
EUTHANIZED
FOSTER
RETURNED TO OWNER
no
yes



Very little difference in proportions for yes/no across different levels of each categorical explanatory - maybe do not go for log-odds model.

3 Formal Data Analysis

3.1 Binomial models for on_time

Observations	1833
Dependent variable	on_time
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(9)$	92.20
Pseudo-R^2 (Cragg-Uhler)	0.10
Pseudo-R^2 (McFadden)	0.08
AIC	1096.09
BIC	1151.23

BINOMIAL MODEL:From the exploratory analysis, it seemed fitting to try models based on intake and/or outcome, and then build up from there. Having fitted some models, model1 seems to be the best, which is simply intake as the only explanatory variable for the log-odds.

	Est.	S.E.	z val.	p
(Intercept)	0.24	0.45	0.54	0.59
animalDOG	-0.31	0.21	-1.44	0.15
intakeOWNER SURRENDER	1.69	0.35	4.80	0.00
intakeSTRAY	1.39	0.33	4.18	0.00
outcomeDIED	1.13	1.03	1.09	0.27
outcomeEUTHANIZED	0.53	0.18	2.99	0.00
outcomeFOSTER	0.59	0.62	0.96	0.34
outcomeRETURNED TO OWNER	3.51	0.61	5.77	0.00
chipSCAN NO CHIP	0.44	0.23	1.93	0.05
chipUNABLE TO SCAN	-0.02	0.44	-0.05	0.96

Standard errors: MLE

Observations	1833
Dependent variable	on_time
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(4)$	65.23
Pseudo-R² (Cragg-Uhler)	0.07
Pseudo-R² (McFadden)	0.06
AIC	1113.07
BIC	1140.63

	Est.	S.E.	z val.	p
(Intercept)	1.81	0.10	17.63	0.00
outcomeDIED	1.23	1.03	1.20	0.23
outcomeEUTHANIZED	0.44	0.17	2.58	0.01
outcomeFOSTER	0.80	0.61	1.32	0.19
outcomeRETURNED TO OWNER	2.92	0.59	4.95	0.00

Standard errors: MLE

Observations	1833
Dependent variable	on_time
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(2)$	5.18
Pseudo-R² (Cragg-Uhler)	0.01
Pseudo-R² (McFadden)	0.00
AIC	1169.11
BIC	1185.65

3.2 Poisson models for time

```
Call:
glm(formula = time ~ intake + animal + chip + outcome, family = poisson(),
```

	Est.	S.E.	z val.	p
(Intercept)	1.60	0.27	5.82	0.00
intakeOWNER SURRENDER	0.62	0.31	2.01	0.04
intakeSTRAY	0.71	0.29	2.42	0.02

Standard errors: MLE

Observations	1833
Dependent variable	on_time
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(6)$	84.96
Pseudo-R^2 (Cragg-Uhler)	0.10
Pseudo-R^2 (McFadden)	0.07
AIC	1097.33
BIC	1135.93

	Est.	S.E.	z val.	p
(Intercept)	0.32	0.33	0.97	0.33
outcomeDIED	1.25	1.03	1.21	0.23
outcomeEUTHANIZED	0.53	0.17	3.07	0.00
outcomeFOSTER	0.79	0.61	1.31	0.19
outcomeRETURNED TO OWNER	3.34	0.60	5.52	0.00
intakeOWNER SURRENDER	1.65	0.35	4.78	0.00
intakeSTRAY	1.45	0.33	4.41	0.00

Standard errors: MLE

Observations	1833
Dependent variable	on_time
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(4)$	5.42
Pseudo-R^2 (Cragg-Uhler)	0.01
Pseudo-R^2 (McFadden)	0.00
AIC	1172.88
BIC	1200.45

	Est.	S.E.	z val.	p
(Intercept)	1.55	0.32	4.86	0.00
chipSCAN NO CHIP	0.08	0.21	0.37	0.71
chipUNABLE TO SCAN	-0.05	0.40	-0.13	0.89
intakeOWNER SURRENDER	0.61	0.31	1.98	0.05
intakeSTRAY	0.69	0.29	2.34	0.02

Standard errors: MLE

Observations	1833
Dependent variable	on_time
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(3)$	5.77
Pseudo-R² (Cragg-Uhler)	0.01
Pseudo-R² (McFadden)	0.00
AIC	1170.52
BIC	1192.58

	Est.	S.E.	z val.	p
(Intercept)	1.74	0.33	5.24	0.00
animalDOG	-0.15	0.20	-0.76	0.45
intakeOWNER SURRENDER	0.60	0.31	1.94	0.05
intakeSTRAY	0.69	0.29	2.35	0.02

Standard errors: MLE

```
    data = shelter)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.0878  -2.0942  -0.8433   0.5840   9.8204  

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)               3.25588    0.04922  66.151  < 2e-16 ***
intakeOWNER SURRENDER    -1.40129    0.03806 -36.820  < 2e-16 ***
intakeSTRAY              -1.01833    0.03393 -30.013  < 2e-16 ***
animalDOG                 0.15174    0.02523   6.015 1.80e-09 ***
chipSCAN NO CHIP         -0.14071    0.02783  -5.056 4.28e-07 ***
chipUNABLE TO SCAN       -0.10393    0.05517  -1.884   0.0596 .  
outcomeDIED              -0.87212    0.11426  -7.633 2.29e-14 ***
outcomeEUTHANIZED        -0.64037    0.02225 -28.775  < 2e-16 ***
outcomeFOSTER            -0.39698    0.06786  -5.850 4.92e-09 ***
outcomeRETURNED TO OWNER -1.46341    0.03663 -39.953  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12201.6  on 1832  degrees of freedom
Residual deviance:  9378.3  on 1823  degrees of freedom
AIC: 14517

Number of Fisher Scoring iterations: 6


                              2.5 %      97.5 %
(Intercept)               3.1590085  3.35194721
intakeOWNER SURRENDER    -1.4755537 -1.32635806
```

```
intakeSTRAY              -1.0843667 -0.95135276
animalDOG                 0.1025139  0.20141301
chipSCAN NO CHIP         -0.1950108 -0.08591295
chipUNABLE TO SCAN       -0.2131352  0.00317712
outcomeDIED              -1.1044422 -0.65595751
outcomeEUTHANIZED        -0.6840986 -0.59686078
outcomeFOSTER            -0.5326087 -0.26647242
outcomeRETURNED TO OWNER -1.5357062 -1.39211393
```

```
Call:
glm(formula = time ~ intake + outcome, family = poisson(), data = shelter)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0660  -2.0874  -0.9047   0.5466   9.9237

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.26169    0.03392  96.147  < 2e-16 ***
intakeOWNER SURRENDER    -1.38194    0.03731 -37.042  < 2e-16 ***
intakeSTRAY              -1.03047    0.03372 -30.561  < 2e-16 ***
outcomeDIED              -0.92431    0.11401  -8.107 5.18e-16 ***
outcomeEUTHANIZED        -0.65207    0.02207 -29.551  < 2e-16 ***
outcomeFOSTER            -0.48740    0.06674  -7.303 2.81e-13 ***
outcomeRETURNED TO OWNER -1.40548    0.03578 -39.284  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12201.6  on 1832  degrees of freedom
Residual deviance:  9448.5  on 1826  degrees of freedom
AIC: 14581

Number of Fisher Scoring iterations: 6
```

```
                              2.5 %     97.5 %
(Intercept)               3.1946881  3.3276782
intakeOWNER SURRENDER    -1.4547129 -1.3084578
intakeSTRAY              -1.0960888 -0.9639035
outcomeDIED              -1.1561811 -0.7086556
outcomeEUTHANIZED        -0.6954363 -0.6089368
outcomeFOSTER            -0.6209137 -0.3591875
outcomeRETURNED TO OWNER -1.4761247 -1.3358652
```

```
[1] FALSE
```

```
[1] FALSE
```

POISSON MODEL: looking at model6, using it has a very high pearson residual which is higher than the chi-squared statistic meaning it is probably suffering from overdispersion. To deal with this could be to introduce a dispersion parameter and have a quasi-poisson model OR we go for a negative binomial model.

```
Call:
glm(formula = time ~ intake + outcome, family = poisson(), data = shelter)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0660  -2.0874  -0.9047   0.5466   9.9237

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.26169    0.08524  38.265  < 2e-16 ***
intakeOWNER SURRENDER    -1.38194    0.09374 -14.742  < 2e-16 ***
intakeSTRAY              -1.03047    0.08472 -12.163  < 2e-16 ***
outcomeDIED              -0.92431    0.28647  -3.227  0.00125 **
outcomeEUTHANIZED        -0.65207    0.05544 -11.761  < 2e-16 ***
outcomeFOSTER            -0.48740    0.16769  -2.907  0.00365 **
outcomeRETURNED TO OWNER -1.40548    0.08990 -15.634  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 6.313406)

    Null deviance: 12201.6  on 1832  degrees of freedom
Residual deviance:  9448.5  on 1826  degrees of freedom
AIC: 14581

Number of Fisher Scoring iterations: 6


Single term deletions

Model:
time ~ intake + outcome
        Df Deviance   AIC F value    Pr(>F)
<none>       9448.5 14581
intake   2  10587.1 15715  110.03 < 2.2e-16 ***
outcome  4  11750.8 16875  111.24 < 2.2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Coefficients still significant (by F-tests) but deviance still above chi-squared.

```
Call:
glm.nb(formula = time ~ intake + outcome + chip + animal, data = shelter,
    init.theta = 1.114925073, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1487  -1.0149  -0.3325   0.2207   3.6358

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.48306    0.14330  24.306  < 2e-16 ***
intakeOWNER SURRENDER    -1.62161    0.11844 -13.692  < 2e-16 ***
intakeSTRAY              -1.25740    0.11014 -11.416  < 2e-16 ***
```

```
outcomeDIED              -0.86714    0.23541  -3.684  0.00023 ***
outcomeEUTHANIZED        -0.71229    0.05612 -12.691  < 2e-16 ***
outcomeFOSTER            -0.43965    0.16496  -2.665  0.00769 **
outcomeRETURNED TO OWNER -1.61127    0.07887 -20.430  < 2e-16 ***
chipSCAN NO CHIP         -0.12566    0.07004  -1.794  0.07279 .
chipUNABLE TO SCAN       -0.07656    0.13781  -0.556  0.57852
animalDOG                 0.17273    0.06356   2.718  0.00658 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for Negative Binomial(1.1149) family taken to be 1)

    Null deviance: 2660.0  on 1832  degrees of freedom
Residual deviance: 2156.3  on 1823  degrees of freedom
AIC: 10079

Number of Fisher Scoring iterations: 1

              Theta:  1.1149
          Std. Err.:  0.0507

 2 x log-likelihood:  -10057.4320

Call:
glm.nb(formula = time ~ intake + outcome, data = shelter, init.theta = 1.103703043,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1038  -0.9345  -0.4035   0.2407   3.3850

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               3.51491    0.11394  30.849  < 2e-16 ***
intakeOWNER SURRENDER    -1.62341    0.11796 -13.762  < 2e-16 ***
intakeSTRAY              -1.26185    0.11024 -11.446  < 2e-16 ***
outcomeDIED              -0.95171    0.23616  -4.030 5.58e-05 ***
outcomeEUTHANIZED        -0.70648    0.05586 -12.647  < 2e-16 ***
outcomeFOSTER            -0.52936    0.16266  -3.254  0.00114 **
outcomeRETURNED TO OWNER -1.55583    0.07675 -20.273  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for Negative Binomial(1.1037) family taken to be 1)

    Null deviance: 2642.4  on 1832  degrees of freedom
Residual deviance: 2154.8  on 1826  degrees of freedom
AIC: 10085

Number of Fisher Scoring iterations: 1

              Theta:  1.1037
          Std. Err.:  0.0500
 2 x log-likelihood:  -10069.4000
```

4 Conclusions