Influential factors of days spent at the shelter

Hanfan Chen 2703334C, Zhaohao Li 2611602L, Zhenhao Qiao 2612824Q, Chao Wang 2518608W, Racha

1 Introduction

2 Exploratory Data Analysis

```
'data.frame':   1450 obs. of  7 variables:
 $ animal_type    : chr  "CAT" "DOG" "DOG" "DOG" ...
 $ month          : int  9 6 12 9 11 12 6 1 2 4 ...
 $ year           : int  2017 2017 2016 2017 2016 2016 2017 2017 2017 2017 ...
 $ intake_type    : chr  "STRAY" "STRAY" "STRAY" "STRAY" ...
 $ outcome_type   : chr  "ADOPTION" "EUTHANIZED" "ADOPTION" "ADOPTION" ...
 $ chip_status    : chr  "UNABLE TO SCAN" "SCAN NO CHIP" "SCAN NO CHIP" "SCAN NO CHIP" ...
 $ time_at_shelter: int  9 4 21 4 7 4 4 5 0 15 ...
```

```
Column 1 :
[1] "BIRD"     "CAT"      "DOG"      "WILDLIFE"

Column 2 :
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12"

Column 3 :
[1] "2016" "2017"

Column 4 :
[1] "CONFISCATED"     "OWNER SURRENDER" "STRAY"

Column 5 :
[1] "ADOPTION"          "DIED"              "EUTHANIZED"
[4] "FOSTER"            "RETURNED TO OWNER"

Column 6 :
[1] "SCAN CHIP"      "SCAN NO CHIP"   "UNABLE TO SCAN"
```

All the explanatory variable are categorical variable and each explanatory variable have multiple levels.

```
[1] 317
```

Over 300 zeros in raw data may cause overdispersion in Poisson regression. The negative binomial model is suggested to fit.

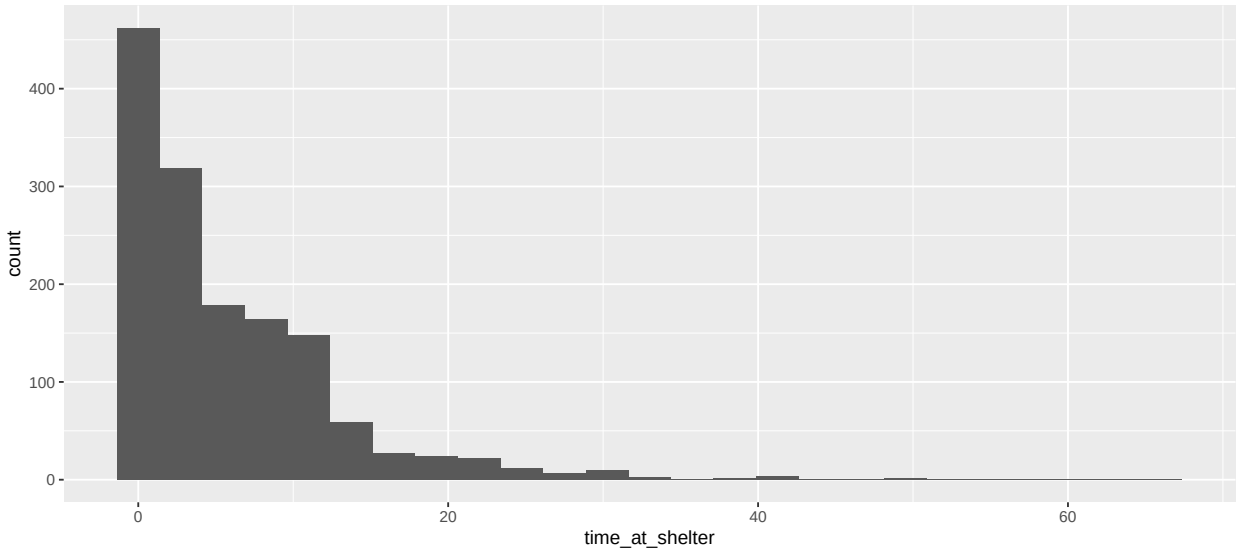


Figure 1: The histogram of day time at shelter

Figure 1 displays the histogram of the response variable, which is days time in the shelter. The histogram shows evidence of right-skewed and Poisson distribution of the response variable.

Figure 2: Pair plots of the variables

The explanatory variables are all categorical and their box plots are shown. The median time at shelter appears to be low for all the explanatory variables, which is due to the median time at shelter being 4.

Since in Figure1 the response variable is right-skewed, a median of the response variable is calculated. The figures below display the median of each category of the different explanatory variables.

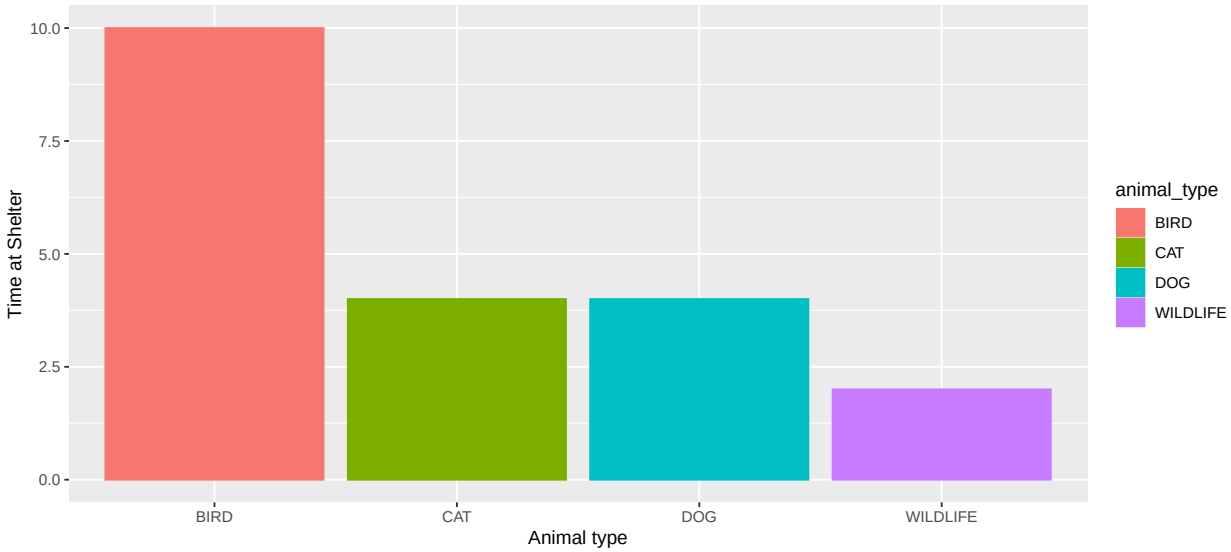


Figure 3: Bar plot of animal type vs time at shelter

Table 1: Summary statistics on the time at shelter by animal type

animal_type	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
BIRD	3	9.333333	8.020806	1	5.5	10	13.5	17
CAT	270	5.903704	7.366027	0	1.0	4	8.0	50
DOG	1163	6.110920	7.375513	0	1.0	4	9.0	66
WILDLIFE	14	4.500000	6.525099	0	0.0	2	6.5	23

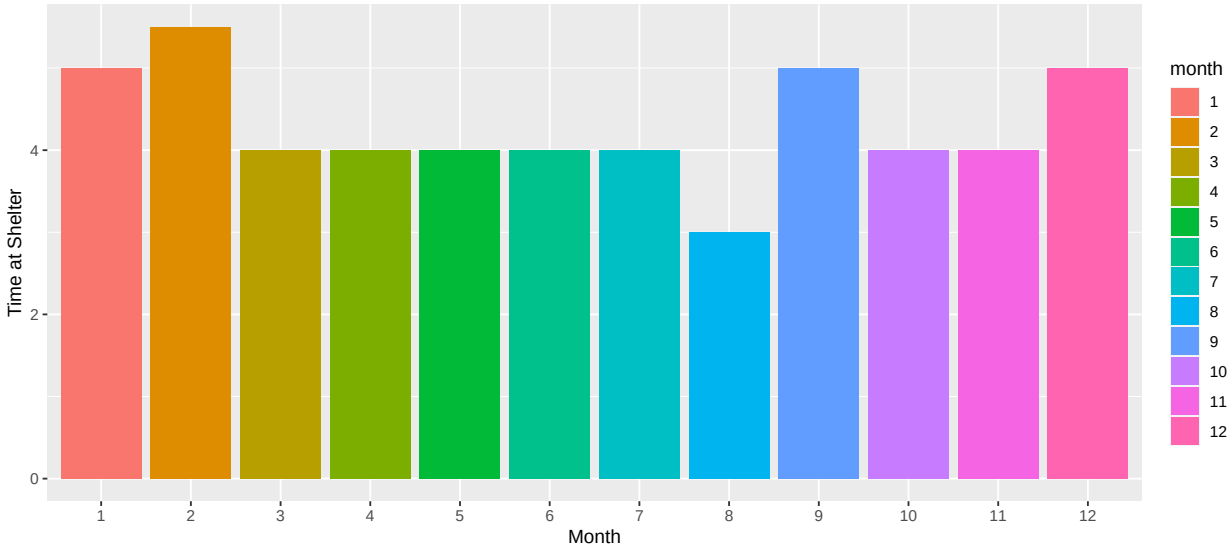


Figure 4: Bar plot of month vs time at shelter

Table 2: Summary statistics on the time at shelter by month

month	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
1	99	6.888889	7.618303	0	1	5.0	10	40
2	82	7.707317	9.646195	0	2	5.5	10	66
3	108	5.287037	7.163055	0	1	4.0	7	42
4	115	5.069565	5.549967	0	1	4.0	6	31
5	139	6.000000	8.062258	0	0	4.0	8	63
6	163	6.184049	6.325765	0	1	4.0	9	29
7	162	5.845679	6.315289	0	0	4.0	10	30
8	127	4.078740	4.922585	0	0	3.0	6	31
9	114	5.456140	4.954912	0	1	5.0	8	22
10	123	6.967480	9.716418	0	1	4.0	8	50
11	110	6.236364	7.911120	0	1	4.0	7	53
12	108	7.888889	9.075317	0	2	5.0	11	59

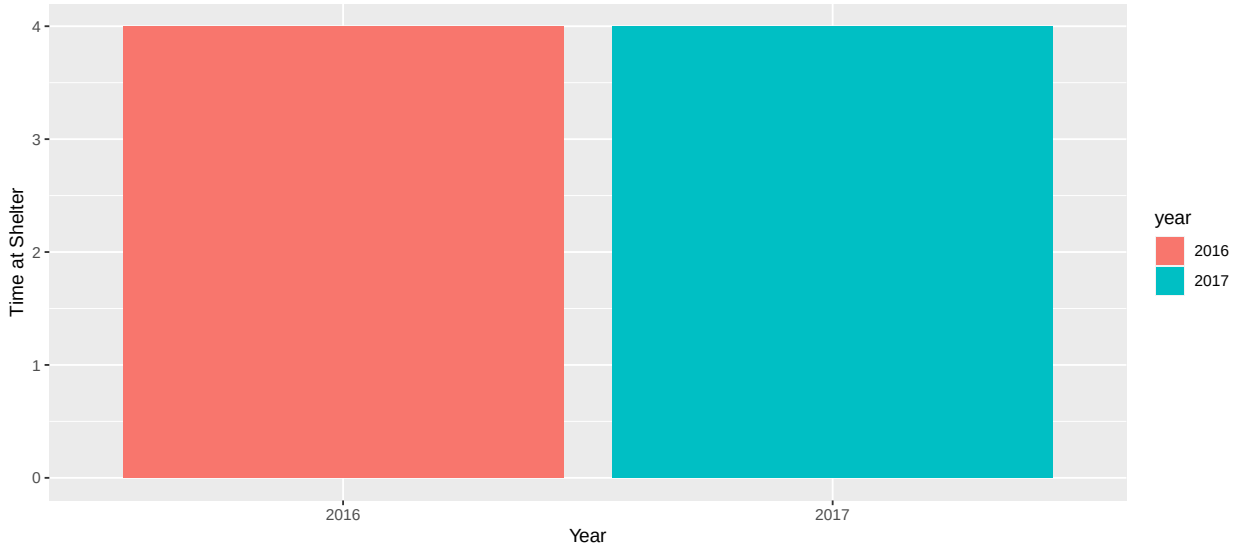


Figure 5: Bar plot of year vs time at shelter

```
[1] FALSE
```

No overlap between the months and years, according to the bar plot5, no obvious difference between two years and the relationship between the response variable and month variable is similar to that relationship between the response variable and the year variable. Therefore, the year variable is removed.

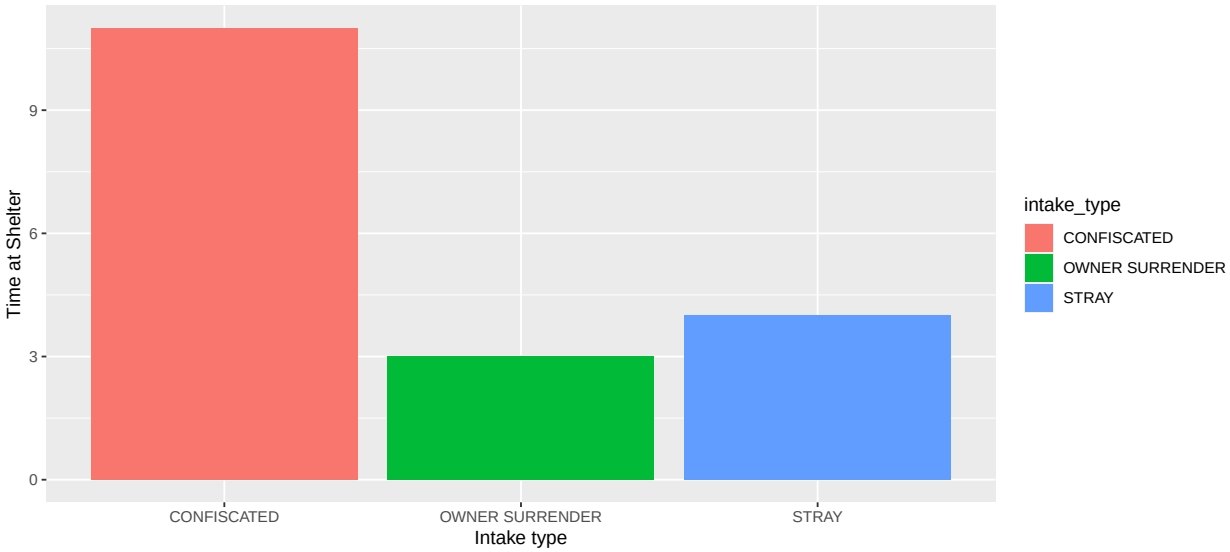


Figure 6: Bar plot of intake type vs time at shelter

Table 3: Summary statistics on the time at shelter by intake type

intake_type	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
CONFISCATED	77	10.896104	9.564992	0	5	11	13	63
OWNER SURRENDER	467	5.141328	7.215962	0	1	3	7	53
STRAY	906	6.128035	7.063027	0	1	4	8	66

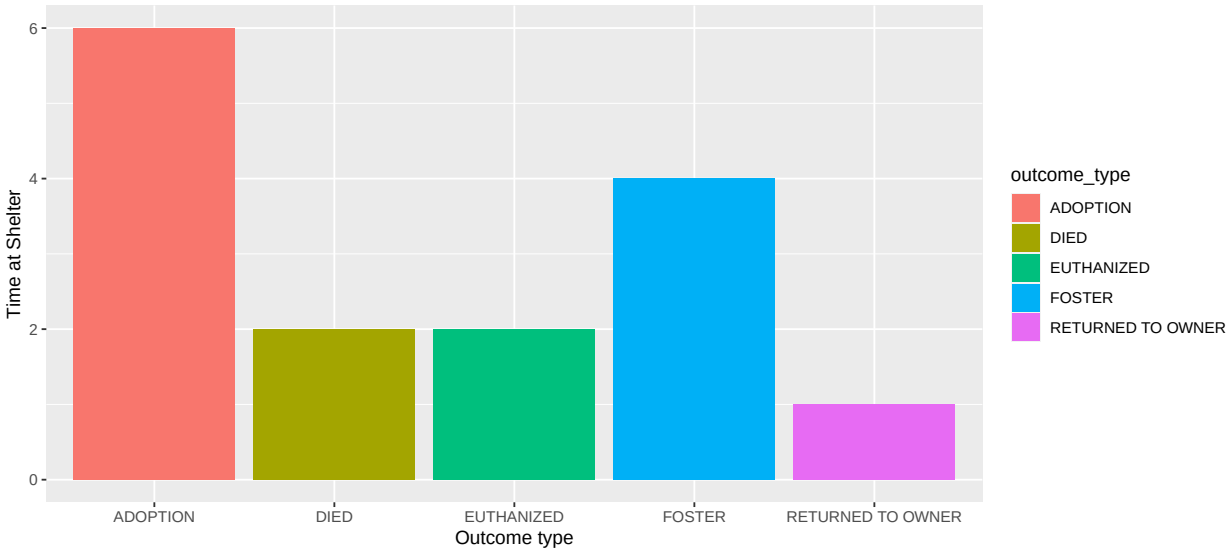


Figure 7: Bar plot of outcome type vs time at shelter

Table 4: Summary statistics on the time at shelter by outcome type

outcome_type	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
ADOPTION	636	8.523585	7.618321	0	4	6	10.25	66
DIED	25	4.360000	6.531207	0	1	2	5.00	33
EUTHANIZED	489	4.777096	7.380844	0	0	2	6.00	63
FOSTER	29	6.482759	8.708045	0	1	4	7.00	37
RETURNED TO OWNER	271	2.723247	3.952610	0	0	1	4.00	22

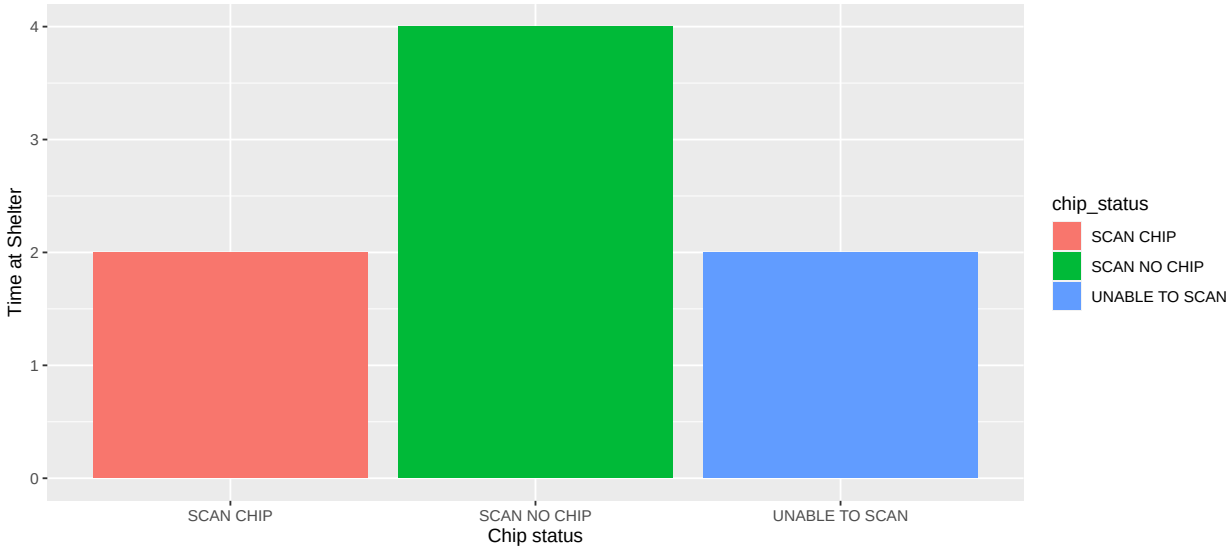


Figure 8: Bar plot of chip status vs time at shelter

Table 5: Summary statistics on the time at shelter by chips status

chip_status	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
SCAN CHIP	285	6.000000	8.582655	0	1	2	10	66
SCAN NO CHIP	1110	6.141441	7.038910	0	1	4	8	63
UNABLE TO SCAN	55	4.818182	6.944465	0	0	2	6	31

3 Formal Data Analysis

3.1 Fitting a Poisson model

```
Call:
glm(formula = time_at_shelter ~ ., family = "poisson", data = data10)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9146  -1.9976  -0.8903   0.6306  12.7550

Coefficients:
```

```
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     2.997158   0.197263  15.194  < 2e-16 ***
animal_typeCAT                  0.441668   0.195885   2.255 0.024150 *
animal_typeDOG                  0.485824   0.194425   2.499 0.012462 *
animal_typeWILDLIFE             0.225305   0.231453   0.973 0.330336
month2                          0.075718   0.055370   1.367 0.171470
month3                         -0.132108   0.057115  -2.313 0.020721 *
month4                         -0.193819   0.056691  -3.419 0.000629 ***
month5                         -0.005919   0.052007  -0.114 0.909386
month6                         -0.035721   0.050097  -0.713 0.475818
month7                         -0.057427   0.050613  -1.135 0.256526
month8                         -0.413755   0.058842  -7.032 2.04e-12 ***
month9                         -0.082308   0.056140  -1.466 0.142617
month10                         0.101852   0.051801   1.966 0.049273 *
month11                        -0.055580   0.054389  -1.022 0.306833
month12                         0.114138   0.051633   2.211 0.027065 *
intake_typeOWNER SURRENDER     -1.451530   0.043649 -33.254  < 2e-16 ***
intake_typeSTRAY               -1.031365   0.039395 -26.180  < 2e-16 ***
outcome_typeDIED               -0.649881   0.097578  -6.660 2.74e-11 ***
outcome_typeEUTHANIZED         -0.592552   0.025262 -23.456  < 2e-16 ***
outcome_typeFOSTER             -0.279520   0.076201  -3.668 0.000244 ***
outcome_typeRETURNED TO OWNER  -1.531722   0.042358 -36.161  < 2e-16 ***
chip_statusSCAN NO CHIP        -0.171716   0.028935  -5.934 2.95e-09 ***
chip_statusUNABLE TO SCAN      -0.247414   0.068726  -3.600 0.000318 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10551.2  on 1449  degrees of freedom
Residual deviance:  8079.3  on 1427  degrees of freedom
AIC: 12147

Number of Fisher Scoring iterations: 6
```

The rootogram could be used to check the overdispersion. The line at 0 allows us to easily visualize where the model is over-fitting or under-fitting, if the bar is below the zero line then that value has been under-fitting. And if there is a space between the zero line and the bar then it has been over-fitting. For the model to be fitted correctly, the bar should sit as close to the zero line as possible.

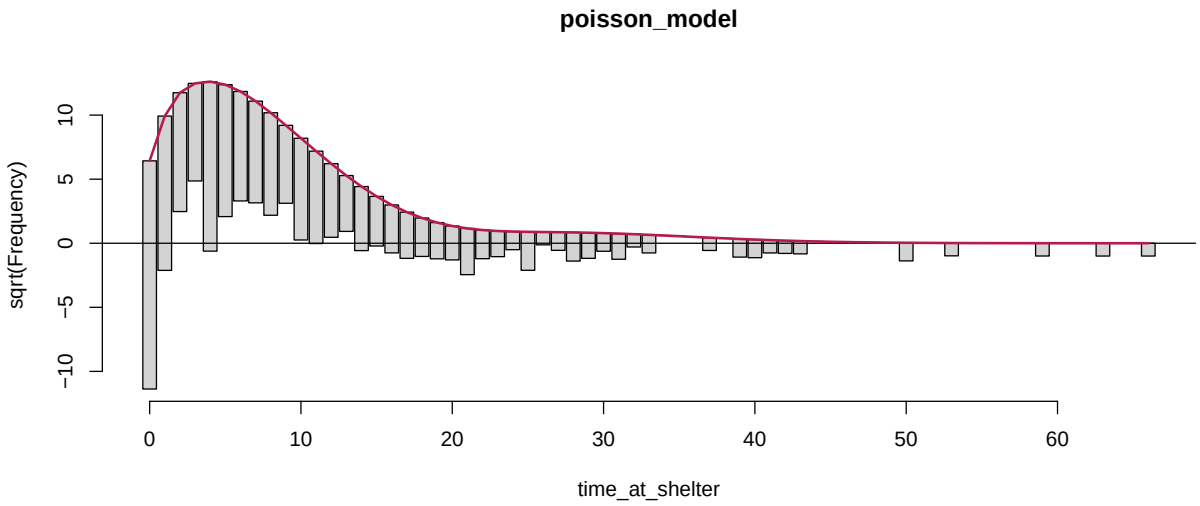


Figure 9: Rootogram of Poisson Model

From Figure9, the Poisson model is severely under-fitting zero counts. There were 317 zero counts observed in the data set but the model only fitted 41. It is also over-fitting the lower positive counts and under-fitting the higher counts, suggesting there is over-dispersion in the model. Hence a hurdle model will be fitted to provide a better fit.

3.2 Fitting a Hurdle model

```
Call:
hurdle(formula = time_at_shelter ~ ., data = data10, dist = "poisson",
    zero.dist = "binomial")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.3608 -1.0287 -0.5823  0.4795 14.9926

Count model coefficients (truncated poisson with log link):
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 2.9579923  0.1983275  14.915  < 2e-16 ***
animal_typeCAT              0.3743137  0.1965591   1.904 0.056867 .
animal_typeDOG              0.3213099  0.1951832   1.646 0.099723 .
animal_typeWILDLIFE         0.4412799  0.2325810   1.897 0.057786 .
month2                     -0.0007866  0.0555725  -0.014 0.988706
month3                     -0.1913094  0.0574189  -3.332 0.000863 ***
month4                     -0.2968745  0.0570389  -5.205 1.94e-07 ***
month5                     -0.0358694  0.0522504  -0.686 0.492405
month6                     -0.1290100  0.0505296  -2.553 0.010675 *
month7                     -0.0908291  0.0508464  -1.786 0.074043 .
month8                     -0.3531232  0.0594007  -5.945 2.77e-09 ***
month9                     -0.1700644  0.0563869  -3.016 0.002561 **
month10                     0.0425144  0.0518410   0.820 0.412164
month11                    -0.0777278  0.0545280  -1.425 0.154023
month12                     0.0460268  0.0517740   0.889 0.374006
```

```
intake_typeOWNER SURRENDER    -1.1067328  0.0453104 -24.426  < 2e-16 ***
intake_typeSTRAY              -0.7609702  0.0407405 -18.678  < 2e-16 ***
outcome_typeDIED              -0.6233442  0.0998502  -6.243 4.30e-10 ***
outcome_typeEUTHANIZED        -0.2197569  0.0254704  -8.628  < 2e-16 ***
outcome_typeFOSTER            -0.1110361  0.0769153  -1.444 0.148847
outcome_typeRETURNED TO OWNER -0.9857031  0.0450846 -21.863  < 2e-16 ***
chip_statusSCAN NO CHIP       -0.2019465  0.0290236  -6.958 3.45e-12 ***
chip_statusUNABLE TO SCAN     -0.2152199  0.0686741  -3.134 0.001725 **
Zero hurdle model coefficients (binomial with logit link):
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.905e+01  6.099e+02   0.031    0.975
animal_typeCAT                -1.328e+01  6.099e+02  -0.022    0.983
animal_typeDOG                -1.266e+01  6.099e+02  -0.021    0.983
animal_typeWILDLIFE           -1.454e+01  6.099e+02  -0.024    0.981
month2                         7.990e-01  4.898e-01   1.631    0.103
month3                         3.817e-01  4.040e-01   0.945    0.345
month4                         3.724e-01  4.020e-01   0.926    0.354
month5                        -9.406e-04  3.735e-01  -0.003    0.998
month6                         4.541e-01  3.702e-01   1.227    0.220
month7                         1.809e-01  3.643e-01   0.497    0.620
month8                        -2.548e-01  3.782e-01  -0.674    0.500
month9                         3.331e-01  3.984e-01   0.836    0.403
month10                        3.409e-01  3.981e-01   0.856    0.392
month11                        5.129e-02  4.062e-01   0.126    0.900
month12                        4.482e-01  4.345e-01   1.032    0.302
intake_typeOWNER SURRENDER    -3.171e+00  5.161e-01  -6.143 8.07e-10 ***
intake_typeSTRAY              -2.406e+00  4.857e-01  -4.955 7.25e-07 ***
outcome_typeDIED              -8.929e-01  8.223e-01  -1.086    0.278
outcome_typeEUTHANIZED        -2.999e+00  2.661e-01 -11.273  < 2e-16 ***
outcome_typeFOSTER            -2.137e+00  5.383e-01  -3.969 7.21e-05 ***
outcome_typeRETURNED TO OWNER -4.203e+00  3.115e-01 -13.491  < 2e-16 ***
chip_statusSCAN NO CHIP       -1.024e-01  1.978e-01  -0.518    0.605
chip_statusUNABLE TO SCAN     -6.084e-01  3.793e-01  -1.604    0.109
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 30
Log-likelihood: -5193 on 46 Df
```

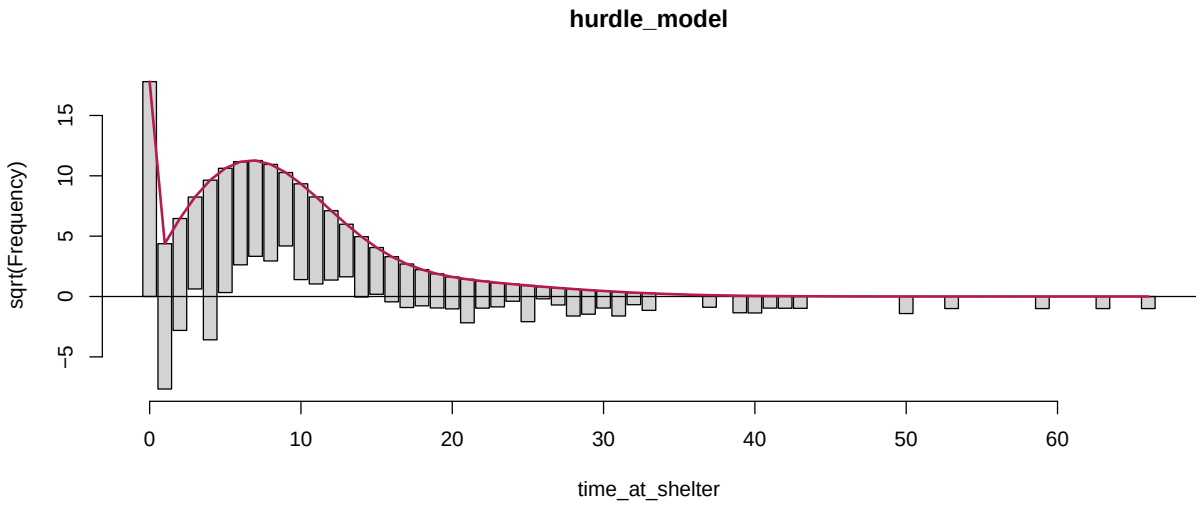


Figure 10: Rootogram of Binomial Hurdle Model

In Figure10 Counts 1,2 and 4 are being severely under-fitted, while 6-9 are being over-fitted. There is also under-fitting at the higher counts which suggests over-dispersion. Therefore, a negative binomial hurdle model shall be fitted to address this.

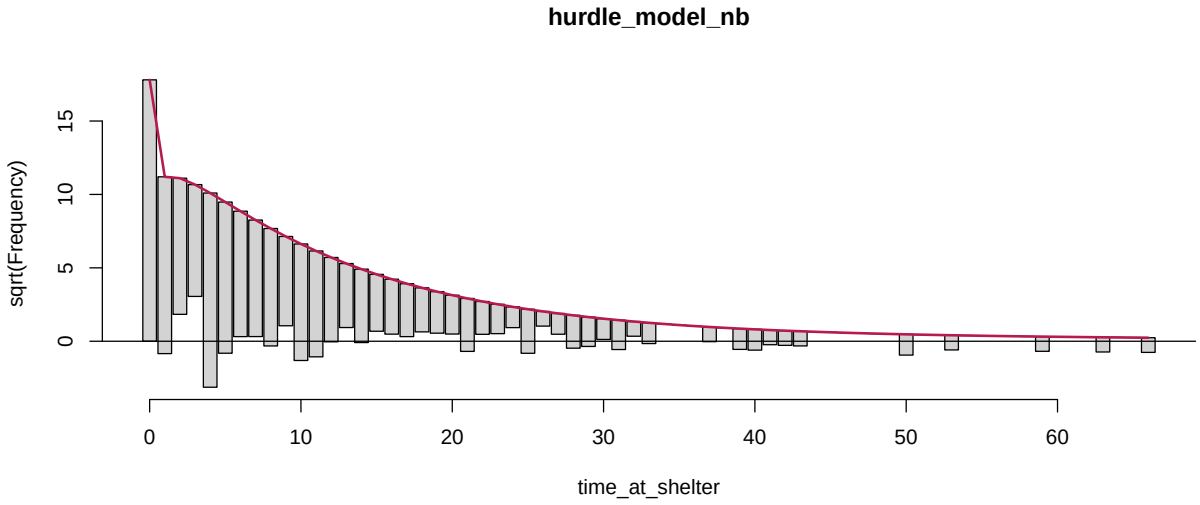


Figure 11: Rootogram of Negative Binomial Hurdle Model

The AIC of the hurdle model is 10478 and the AIC of the negative binomial hurdle model is 7781. From this, the negative binomial model shows a much better fit to the data. However, in Figure11 some values are still being under-fitted.

3.2.1 Variable selection using AIC

```
Start:  AIC=7780.7
```

```
time_at_shelter ~ animal_type + month + intake_type + outcome_type +
    chip_status

               Df    AIC
- month        22 7767.3
<none>            7780.7
- chip_status   4 7782.2
- animal_type   6 7787.7
- intake_type   4 7942.5
- outcome_type  8 8245.8

Step:  AIC=7767.26
time_at_shelter ~ animal_type + intake_type + outcome_type +
    chip_status

               Df    AIC
<none>            7767.3
- chip_status   4 7767.7
- animal_type   6 7776.1
+ month        22 7780.7
- intake_type   4 7931.5
- outcome_type  8 8248.1

Call:
hurdle(formula = time_at_shelter ~ animal_type + intake_type + outcome_type +
    chip_status, data = data10, dist = "negbin")

Count model coefficients (truncated negbin with log link):
                  (Intercept)                 animal_typeCAT
                       2.4956                         0.9004
               animal_typeDOG            animal_typeWILDLIFE
                       0.8454                         0.9344
   intake_typeOWNER SURRENDER               intake_typeSTRAY
                      -1.3568                        -0.9797
            outcome_typeDIED         outcome_typeEUTHANIZED
                      -0.7449                        -0.2824
          outcome_typeFOSTER  outcome_typeRETURNED TO OWNER
                      -0.1796                        -1.2008
      chip_statusSCAN NO CHIP      chip_statusUNABLE TO SCAN
                      -0.1833                        -0.1427
Theta = 1.5067

Zero hurdle model coefficients (binomial with logit link):
                  (Intercept)                 animal_typeCAT
                      19.1526                       -13.1510
               animal_typeDOG            animal_typeWILDLIFE
                     -12.4842                       -14.4181
   intake_typeOWNER SURRENDER               intake_typeSTRAY
                      -3.2086                        -2.4313
            outcome_typeDIED         outcome_typeEUTHANIZED
                      -0.9783                        -2.9986
          outcome_typeFOSTER  outcome_typeRETURNED TO OWNER
                      -2.0942                        -4.2473
      chip_statusSCAN NO CHIP      chip_statusUNABLE TO SCAN
```

```
                -0.1077                         -0.5265
```

Using AIC as a selection criteria, the model with the minimum AIC and hence the best fit for the data is the model with animal type, chip status, intake type and outcome type as the explanatory variables.

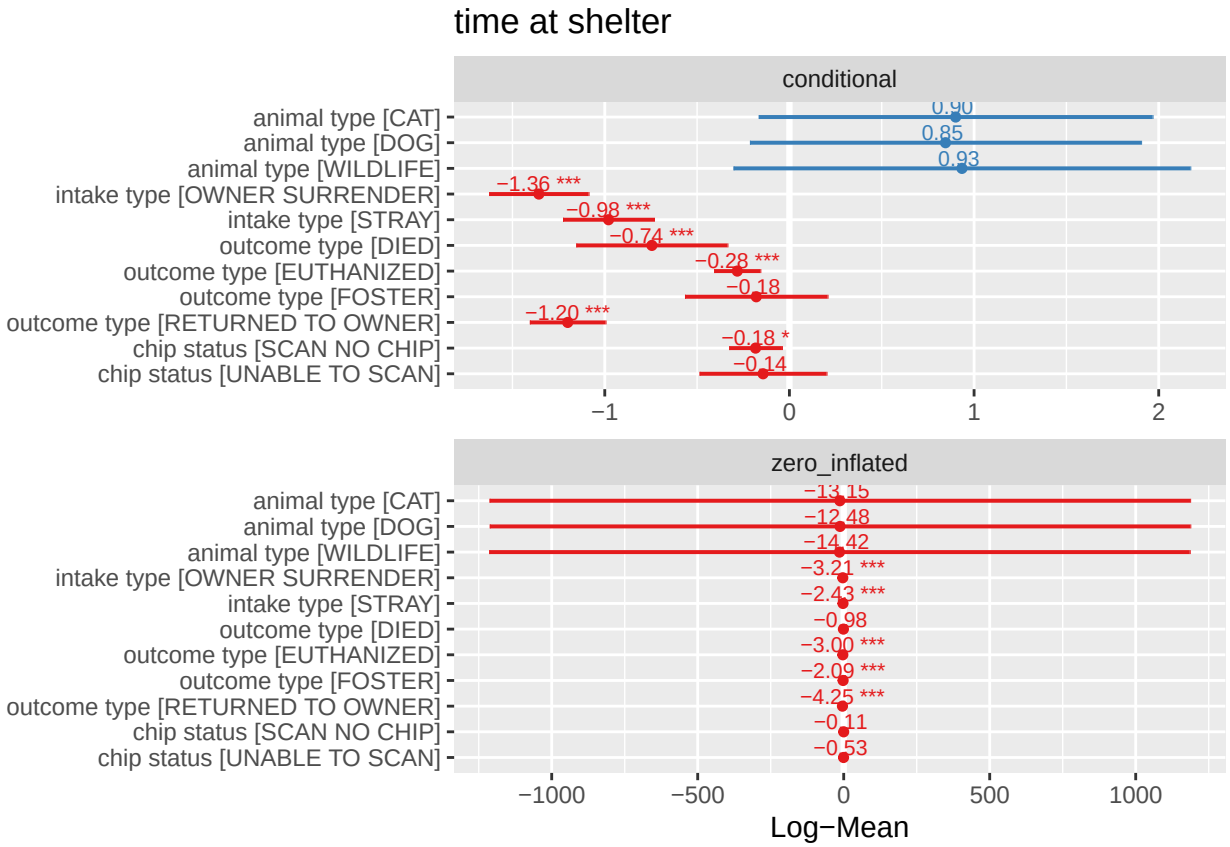
3.2.2 P-value and confidence intervals

```
Call:
hurdle(formula = time_at_shelter ~ animal_type + intake_type + outcome_type +
    chip_status, data = data10, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.1815 -0.6457 -0.3219  0.2380  8.9096

Count model coefficients (truncated negbin with log link):
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   2.49559    0.53328   4.680 2.87e-06 ***
animal_typeCAT                0.90035    0.54405   1.655 0.097943 .
animal_typeDOG                0.84537    0.54038   1.564 0.117726
animal_typeWILDLIFE           0.93442    0.63104   1.481 0.138667
intake_typeOWNER SURRENDER   -1.35684    0.13723  -9.887  < 2e-16 ***
intake_typeSTRAY             -0.97973    0.12565  -7.797 6.33e-15 ***
outcome_typeDIED             -0.74487    0.20889  -3.566 0.000363 ***
outcome_typeEUTHANIZED       -0.28239    0.06371  -4.432 9.32e-06 ***
outcome_typeFOSTER           -0.17956    0.19697  -0.912 0.361973
outcome_typeRETURNED TO OWNER -1.20077    0.10457 -11.483  < 2e-16 ***
chip_statusSCAN NO CHIP      -0.18330    0.07284  -2.517 0.011851 *
chip_statusUNABLE TO SCAN    -0.14273    0.17540  -0.814 0.415789
Log(theta)                    0.40994    0.07215   5.682 1.33e-08 ***
Zero hurdle model coefficients (binomial with logit link):
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   19.1526   612.0501   0.031    0.975
animal_typeCAT               -13.1510   612.0498  -0.021    0.983
animal_typeDOG               -12.4842   612.0498  -0.020    0.984
animal_typeWILDLIFE          -14.4181   612.0502  -0.024    0.981
intake_typeOWNER SURRENDER    -3.2086     0.5150  -6.231 4.64e-10 ***
intake_typeSTRAY              -2.4313     0.4848  -5.016 5.29e-07 ***
outcome_typeDIED              -0.9783     0.8054  -1.215    0.225
outcome_typeEUTHANIZED        -2.9986     0.2648 -11.322  < 2e-16 ***
outcome_typeFOSTER            -2.0942     0.5372  -3.898 9.69e-05 ***
outcome_typeRETURNED TO OWNER -4.2473     0.3101 -13.697  < 2e-16 ***
chip_statusSCAN NO CHIP       -0.1077     0.1944  -0.554    0.579
chip_statusUNABLE TO SCAN     -0.5265     0.3724  -1.414    0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 1.5067
Number of iterations in BFGS optimization: 20
Log-likelihood: -3859 on 25 Df
```



3.3 Goodness of fit

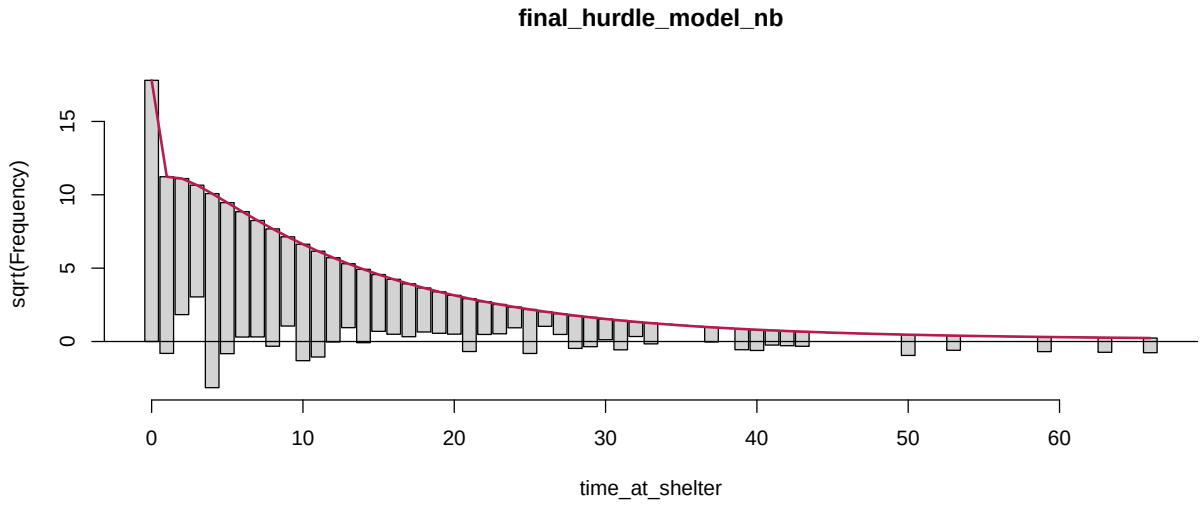


Figure 12: Rootogram of Negative Binomial Hurdle Model with reduced variables

The final model provides an adequate fit to the data. It has the lowest AIC of 7767.26 and as seen from Figure 12 fits most values well.

4 Conclusions

The model which provides the best fit to the data is one which includes animal type, intake type, outcome type and chip status as explanatory variables. Hence these factors are the most influential in determining the number of days an animal spends at the shelter before their final outcome is decided.