DALLAS ANIMAL SHELTER DATA ANALYSIS

Group20

1 Introduction

We are interested in studying the lifetime of a domestic animal in a shelter before the prospective adoption. For this reason, we obtained 1465 observations with respect to 7 different variables.

The variable of research is the time at the shelter which is considered as the response variable and the explanatory variables are described below:

First Explanatory Variable (discrete variable): The animal type: “Bird”, “Cat”, “Dog” & “Wildlife”.

Second Explanatory variable: (discrete variable): The months that the animal spent in a shelter within a year. The variable can take a value from 1 to 12.

Third Explanatory variable (discrete variable): We made this analysis based on the year 2017.

Fourth explanatory variable: (discrete variable): Intake type: “Confiscated”, “owner surrender” & “stray”.

Fifth explanatory variable: (discrete variable):outcome type: “adoption”, “died”, “euthanized”, “foster” & “returned to owner”.

Sixth variable: (discrete variable): chip status: “scan chip”, “scan no chip” & “unable to scan”.

2 Load Data

```r
dataset20 <- read.csv("dataset20.csv")
library(knitr)
dataset20$animal_type<-as.factor(dataset20$animal_type)
dataset20$intake_type<-as.factor(dataset20$intake_type)
dataset20$outcome_type<-as.factor(dataset20$outcome_type)
dataset20$chip_status<-as.factor(dataset20$chip_status)
dataset20$month<-as.factor(dataset20$month)
head20<-head(dataset20)#View the structure of data
library(kableExtra)
head20 %>%
  kable(caption = '\\label{tab:summary} Summary statistics on Dallas animal shelter.',
        align = 'c') %>%
  kable_styling(latex_option = "hold_position")
```

3 Exploratory Data Analysis

Table 1: Summary statistics on Dallas animal shelter.

animal_type	month	year	intake_type	outcome_type	chip_status	time_at_shelter
DOG	3	2017	STRAY	ADOPTION	SCAN NO CHIP	11
CAT	9	2017	STRAY	EUTHANIZED	SCAN NO CHIP	0
DOG	9	2017	CONFISCATED	RETURNED TO OWNER	SCAN NO CHIP	22
DOG	6	2017	STRAY	EUTHANIZED	SCAN NO CHIP	4
DOG	1	2017	STRAY	RETURNED TO OWNER	SCAN NO CHIP	5
DOG	4	2017	STRAY	RETURNED TO OWNER	SCAN CHIP	5

```
dataset20 %>%
  ggplot(aes(x = time_at_shelter)) +
  geom_histogram(binwidth = 5,fill="deep pink",color="white") +
  labs(x = "Time at shelter", y = "Number of days")
```

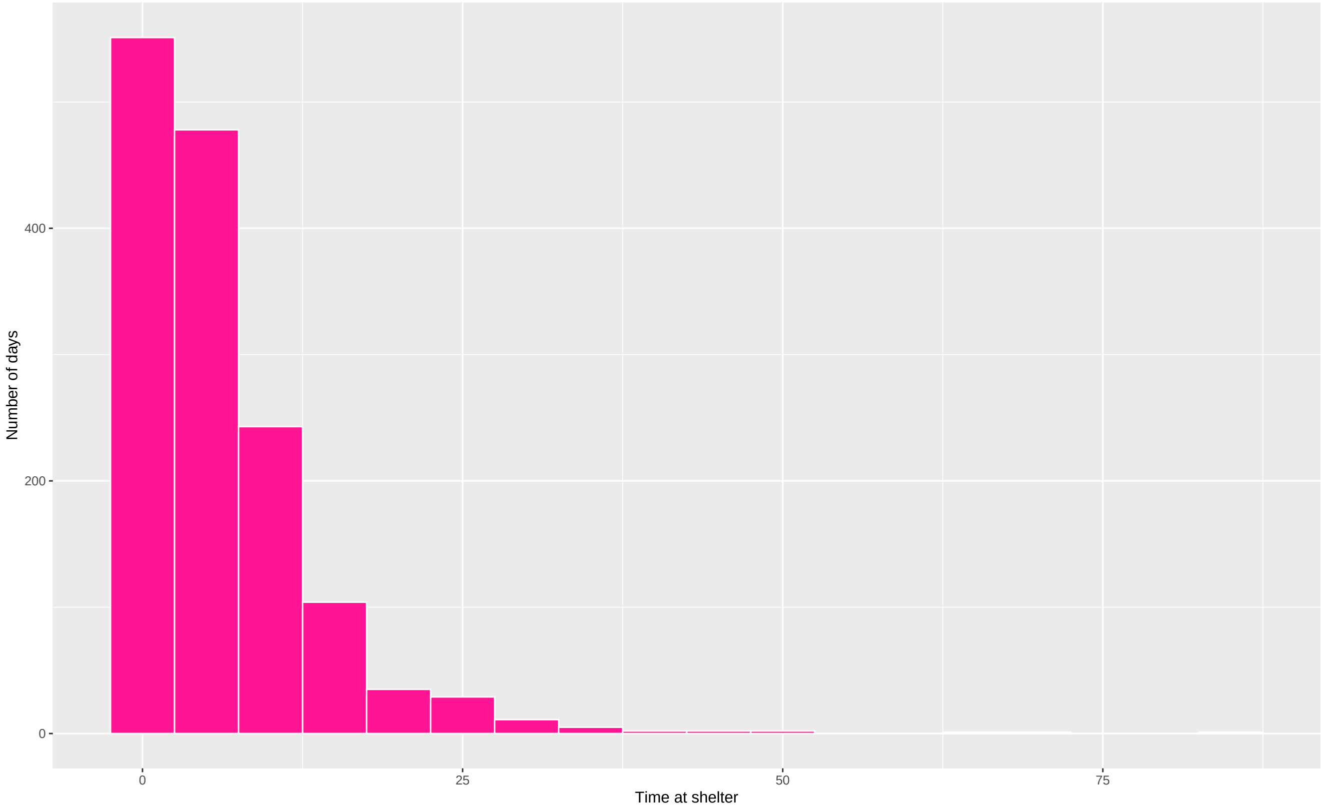

Figure 1: Days stay at shelter

In figure 1, we attempted to measure the overall time an animal remains at the shelter, in days. From the graph, we can easily see that an animal, according to its type, can spent from 0 (the animal might be adopted, returned to owner, or passed away within the same day) to more than 500 days in a shelter.

```
#Plot pie charts of 4 categorical variables
animal_count <- dataset20 %>%
  group_by(animal_type) %>%
  summarize(count = n())
```

```
total_count1 <- sum(animal_count$count)
animal_count$percent1 <- animal_count$count/ total_count1
pie1<-ggplot(animal_count, aes(x = "animal_type", y = percent1,fill =animal_type)) +
  geom_col(width = 1) + coord_polar(theta = "y") +
  geom_text(aes(label=paste0(round(percent1*100),"%")),position=position_stack(vjust=0.5)) +
  theme_void() + ggtitle("Pie chart of animal type")

outcome_count <- dataset20 %>%
  group_by(outcome_type) %>%
  summarize(count = n())

total_count2 <- sum(outcome_count$count)
outcome_count$percent2<- outcome_count$count/ total_count2

pie2<-ggplot(outcome_count, aes(x = "outcome_type",y = percent2,fill =outcome_type)) +
geom_col(width = 1) + coord_polar(theta = "y") +
  geom_text(aes(label=paste0(round(percent2*100),"%")),position=position_stack(vjust=0.5)) +
  theme_void() + ggtitle("Pie chart of outcome type")

chip_count <- dataset20 %>%
  group_by(chip_status) %>%
  summarize(count = n())

total_count3 <- sum(chip_count$count)
chip_count$percent3<- chip_count$count/total_count3

pie3<-ggplot(chip_count, aes(x = "chip_status",y = percent3,fill =chip_status))+
geom_col(width = 1) +
  coord_polar(theta = "y") +
  geom_text(aes(label=paste0(round(percent3*100),"%")),position=position_stack(vjust=0.5))+
  theme_void() + ggtitle("Pie chart of chip status")

intake_count <- dataset20 %>%
  group_by(intake_type) %>%
  summarize(count = n())

total_count4 <- sum(intake_count$count)

intake_count$percent4<- intake_count$count/total_count4

pie4<-ggplot(intake_count, aes(x = "intake_type",y = percent4,fill =intake_type)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  geom_text(aes(label=paste0(round(percent4*100),"%")),position=position_stack(vjust=0.5))+
  theme_void()+ggtitle("Pie chart of intake type")

grid.arrange(pie1,pie2,pie3,pie4,ncol=2,nrow=2)
```

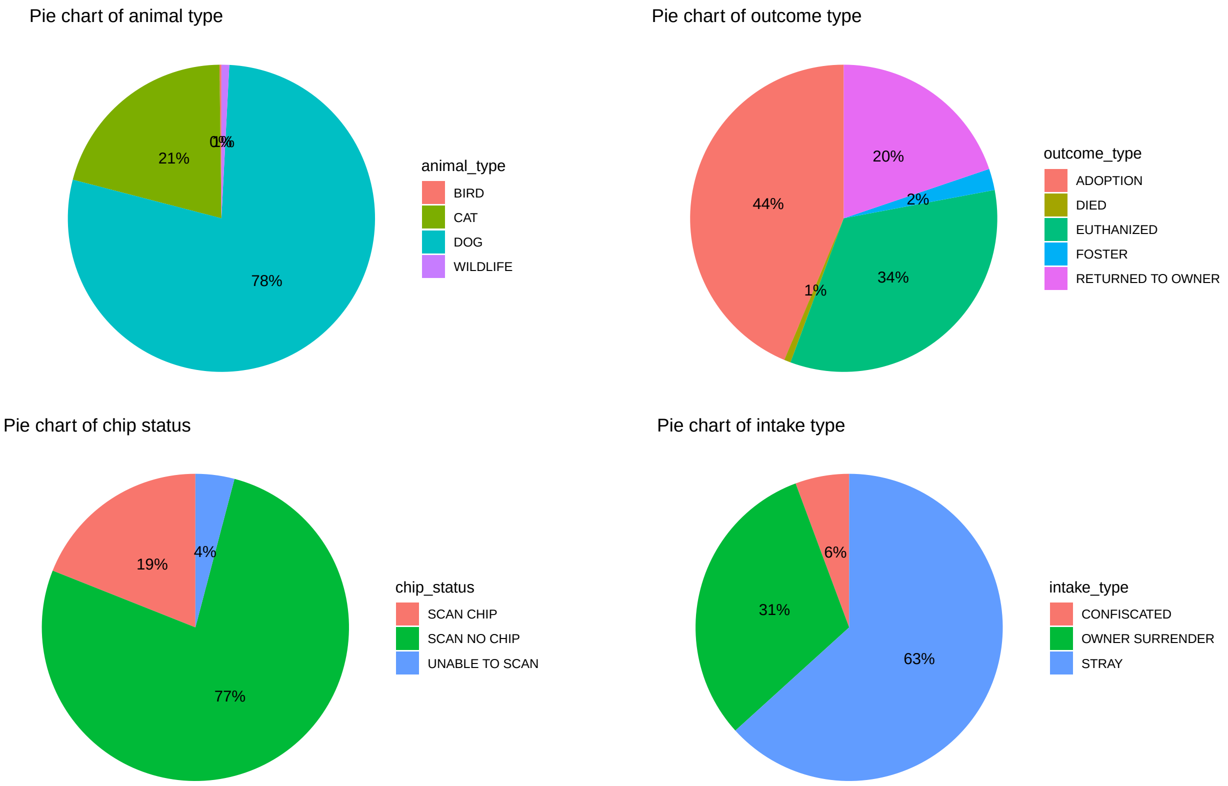

Figure 2: Different time at shelter of different categorical variables

```r
#Create 4 box plots of categorical variables and "time_at_shelter"
#with color mapped to categorical variables
plot1<-ggplot(data =dataset20, 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")
plot2<-ggplot(data =dataset20, 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")
plot3<-ggplot(data =dataset20, 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")
plot4<-ggplot(data =dataset20, 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")
library(gridExtra)
grid.arrange(plot1,plot2,plot3,plot4,ncol=2,nrow=2)
```

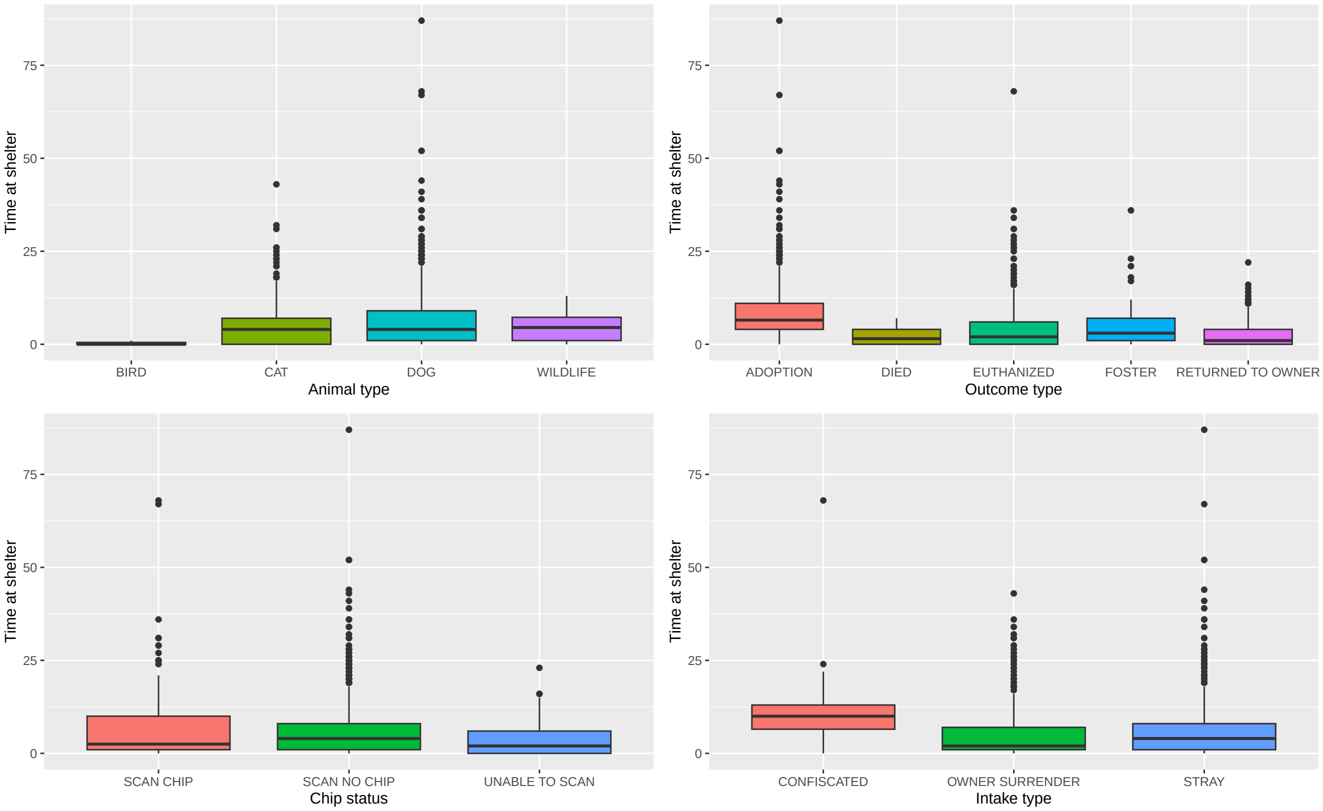

Figure 3: Boxplots of different time at shelter of different categorical variables

Time at shelter (vs) animal type:

We can see that the average time spent in a shelter is almost zero when the type is bird, compared to other domestic animals like cats and dogs, whose average life in a shelter is much higher.

Moreover, cats are spending a higher time in a shelter than birds but shorter than dogs and wild animals; However, the outlier number of cats remained in a shelter is lower that this of dogs and higher than this of wild animals. Also, the average life of dogs within a shelter is almost the same as those animals who belong to “wildlife” category, however the wild animals does not exhibit a higher outlier than the dogs.

To conclude, the overall picture is that dogs remain in a shelter for a longer period compared to all other animals.

Time at shelter (vs) outcome type:

We can see from the graph that lost animals are being returned to the owners almost within the same day.

Except of those returned to the owner, the shortest life in the shelter is exhibited by those animals that need to be euthanized, due to health anomalies. However, we can see a reasonably high outlier number of days in this category which enhances the hypothesis that animals are getting a good treatment in the shelters and the euthanasia is considered the last resort.
Finally, we can notice that animals whose odds are in favour to get adopted or fostered, present higher divergence in days along with the time spent in a shelter.

Time at shelter (vs) chip status:

As excepted, animals that have been chipped are spending on average, a shorter time in a shelter as they can be easily found, cured and returned to owners; However, animals that have not been chipped, are spending a longer time in a shelter and present higher outliers.

Time at shelter (vs) Intake type:

Life in a shelter is extended for animals that are confiscated compared to those that have been surrounded by an owner or strayed. Among the later two categories, higher average time present the surrounded animals.

```
ggplot(data = dataset20, mapping = aes(x = factor(month), y = time_at_shelter)) +
  geom_boxplot(fill = "purple") +
  labs(x = "Month", y = "time_at_shelter",
       title = "Boxplots of time at shelter by month")  +
  scale_x_discrete(labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                              "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
```

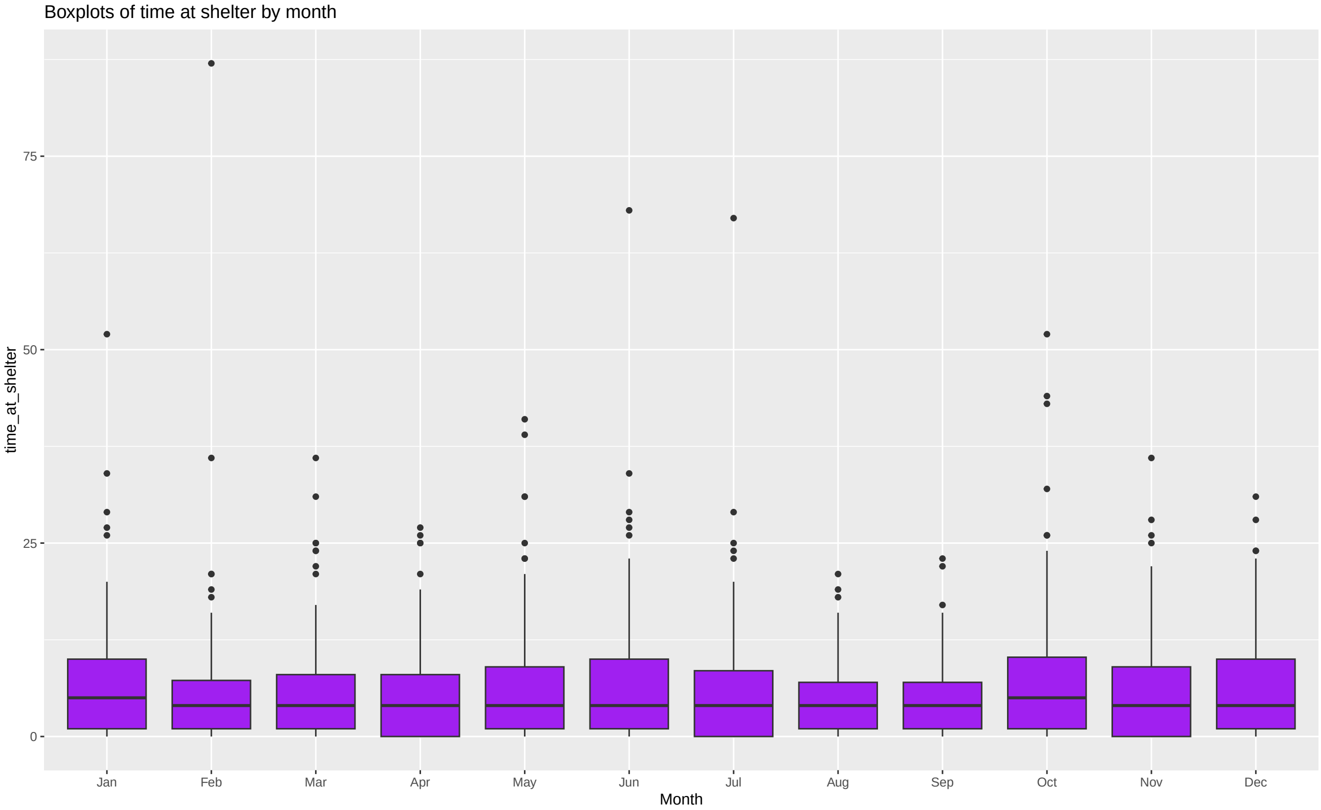

Figure 4: Boxplots of time at shelter by month

From the figure 4, the average time is similar across all months and there is not a specific pattern with respect to which month an animal is spending in a shelter.

4 Application Condition Judgement

4.1 Check of outliers

```
# detection of outliers
boxplot(dataset20$time_at_shelter,
        col=c('orange'),
        ylab="Days",xlab="Time at shelter")
```

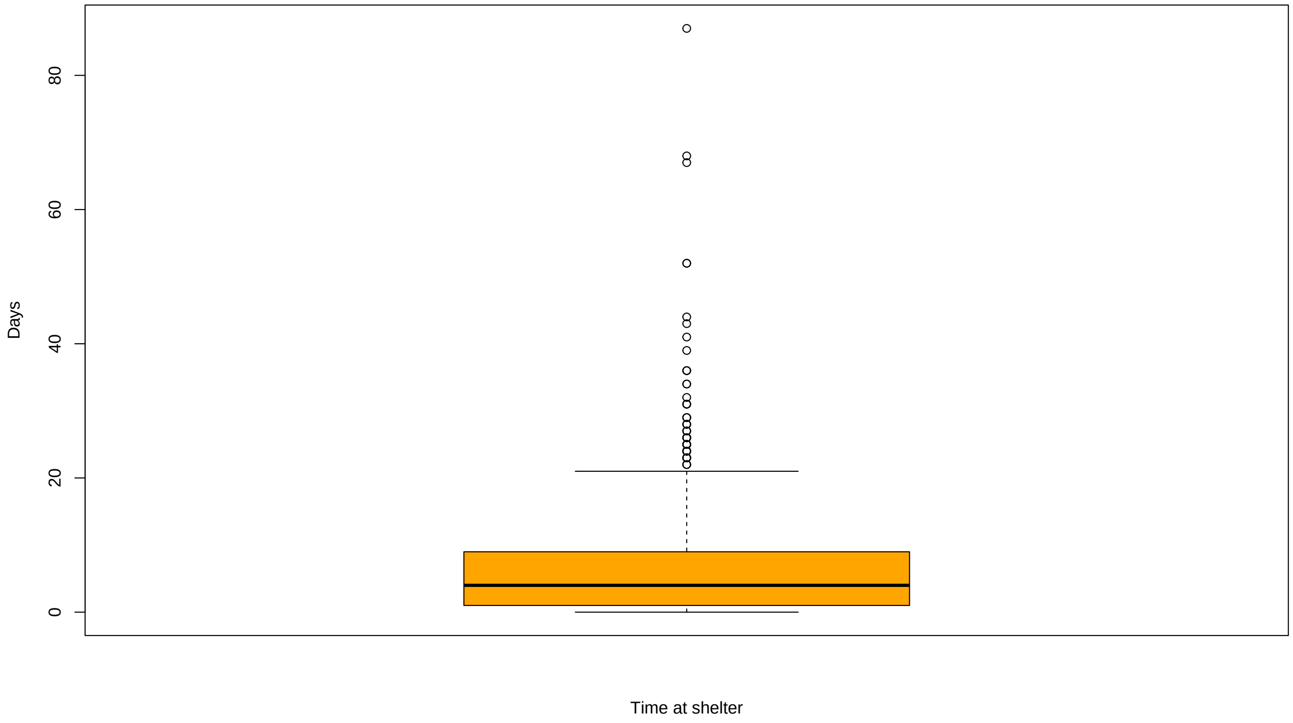

Figure 5: Boxplot of number of days at shelter to check for outliers

The figure 5 indicates multiple outliers, by looking at the maximum values which is 87 which means that the days the animal spent at the shelter was 87. In this case, the 'outliers' we find are consistent with the actual conditions, thus these values can be retained.

4.2 Check of Overdispersion

Initially we run a Poison generalized model and we attempted to discover if the Poisson function is the most appropriate to be used as the linear function to fit the data.

For this purpose, we examined the significance of slope coefficients individually and the validity of the overall model.

For this purpose, we employee the Wald test and the Null Hypothesis is that each slope coefficient individually equals to zero and the alternative Hypothesis is that each slope coefficient individually is not equal to zero. At 5% level of significance, we concluded that the slope coefficients related to the below covariates are significantly different from zero. In other words, they are able to predict the value of time that an animal remains in a shelter:

animal_typeCAT, animal_typeDOG, animal_typeWILDLIFE
intake_typeOWNER SURRENDER , intake_typeSTRAY
outcome_typeDIED , outcome_typeEUTHANIZED, outcome_typeFOSTER , outcome_typeRETURNED
chip_statusSCAN NO CHIP, chip_statusUNABLE TO SCAN
month2, month4, month5, month6, month7, month8, month9, month10, month11, month12

We can now evaluate the significance of the estimated regression model as a whole and assess whether all regression coefficients as a group explain the variation in the dependent variable. To achieve this, we employee the concept of deviance. The residual deviance for the fitted model is 8018.8 on 1442 degrees of freedom. The value of the deviance is fairly low compared to the degrees of freedom and for this reason we can conclude that the model fits the data well.

```
fit1<-glm(time_at_shelter~animal_type+intake_type+outcome_type+chip_status+month,
          data=dataset20,family=poisson(link = "log")) #build poisson glm model
summary(fit1) #view the coefficients
```

```
Call:
glm(formula = time_at_shelter ~ animal_type + intake_type + outcome_type +
    chip_status + month, family = poisson(link = "log"), data = dataset20)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2744  -2.1217  -0.8758   0.7295  14.9771

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.94079    1.00176   0.939 0.347663
animal_typeCAT                 2.31600    1.00071   2.314 0.020648 *
animal_typeDOG                 2.58007    1.00048   2.579 0.009913 **
animal_typeWILDLIFE            2.05787    1.00931   2.039 0.041462 *
intake_typeOWNER SURRENDER    -1.38232    0.04242 -32.583  < 2e-16 ***
intake_typeSTRAY              -0.92723    0.03822 -24.262  < 2e-16 ***
outcome_typeDIED              -1.32437    0.20950  -6.322 2.59e-10 ***
outcome_typeEUTHANIZED        -0.64747    0.02531 -25.585  < 2e-16 ***
outcome_typeFOSTER            -0.23164    0.07356  -3.149 0.001638 **
outcome_typeRETURNED TO OWNER -1.56777    0.04075 -38.476  < 2e-16 ***
chip_statusSCAN NO CHIP       -0.17566    0.02873  -6.115 9.67e-10 ***
chip_statusUNABLE TO SCAN     -0.35349    0.07069  -5.001 5.72e-07 ***
month2                        -0.12768    0.05605  -2.278 0.022728 *
month3                        -0.09897    0.05434  -1.821 0.068558 .
month4                        -0.21679    0.05433  -3.990 6.60e-05 ***
month5                        -0.10257    0.05175  -1.982 0.047468 *
month6                         0.09563    0.04963   1.927 0.053988 .
month7                        -0.12955    0.05108  -2.536 0.011205 *
month8                        -0.38149    0.05598  -6.815 9.44e-12 ***
month9                        -0.22081    0.05986  -3.689 0.000225 ***
month10                        0.07004    0.05133   1.364 0.172423
month11                       -0.14477    0.05559  -2.604 0.009209 **
month12                       -0.12170    0.05605  -2.171 0.029921 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 10754.2  on 1464  degrees of freedom
Residual deviance:  8018.8  on 1442  degrees of freedom
AIC: 12096

Number of Fisher Scoring iterations: 6
```

```
c<-deviance(fit1)/df.residual(fit1) #calculate the deviance over the degree of residual
c
```

```
[1] 5.560922
```

```
ggplot(fit1,aes(x=log(fitted(fit1)),y=log((dataset20$time_at_shelter-fitted(fit1))^2)))+
  geom_point(col="#f46d43") +
  geom_abline(slope=1, intercept=0, col="#a6d96a", size=1) +
  ylab(expression((y-hat(mu))^2)) + xlab(expression(hat(mu)))
```

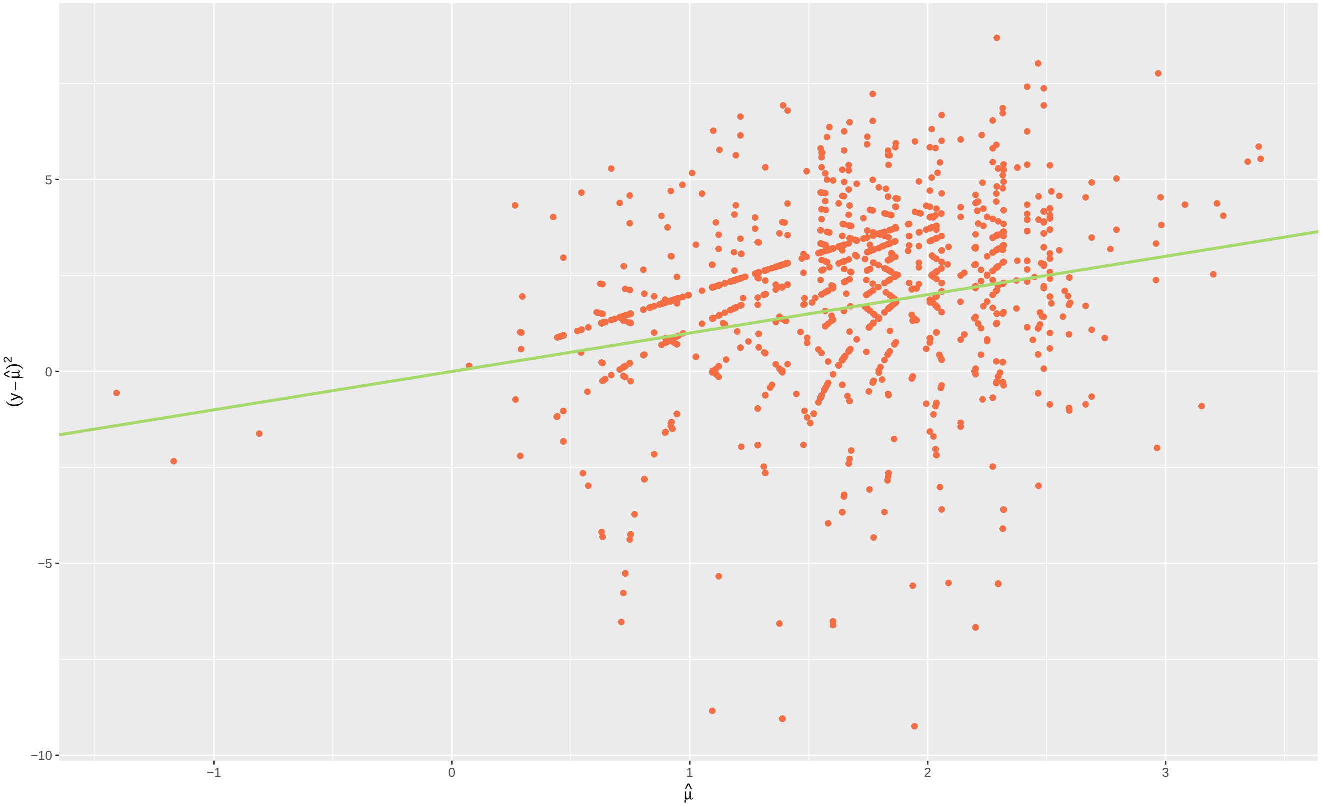

Figure 6: Scatterplot of overdispersion detection

We employed the scatterplot of overdispersion, and we discovered that our model suffers from overdispersion and for this reason it cannot be used for out of sample data.
The proportion is much greater than 1 From the figure 6, this appears that most of the points lie above the line of quality for mean and variance. These both imply that may have overdispersion of the data. Then use the qcc library to test whether the data is overdispersion.

```
qcc.overdispersion.test(dataset20$time_at_shelter,type="poisson")
```

```
Overdispersion test Obs.Var/Theor.Var Statistic p-value
       poisson data          9.127444  13362.58       0
```

The p-value of significance testing is less than 0.05, further indicating overdispersion in data.

4.3 Check of Multicollinearity

```
fit2<- lm(time_at_shelter~animal_type+intake_type+outcome_type+chip_status+month,
          data=dataset20) #build the linear model
```

```
vif(fit2) %>% #calculate VIF
  kable(caption = '\\label{tab:vif} Variance Inflation Factor table', align = 'c') %>%
  kable_styling(latex_option = "hold_position")
```

Table 2: Variance Inflation Factor table

	GVIF	Df	GVIF^(1/(2*Df))
animal_type	1.176723	3	1.027493
intake_type	1.232188	2	1.053584
outcome_type	1.421322	4	1.044928
chip_status	1.159105	2	1.037602
month	1.130833	11	1.005605

From the above table, it lists the VIF of the independent variables. It indicates that the VIF of the variables are all less than 10 which means that the independent variables doesn't exist serious multicollinearity.

5 Model fitting

We rejected our previous model, given that it suffers from overdispersion, and we now attempt to fit the data using the negative binomial as the linear function.

Following an identical methodology as in the previous model, we examined the significance of slope coefficients individually and the validity of the overall model.

Employing the Wald test, the first Null Hypothesis is that each slope coefficient individually equals to zero and the alternative Hypothesis is that each slope coefficient individually is not equal to zero. At 5% level of significance (critical value: 1.94), we concluded that the slope coefficients related to the below covariates are significantly different from zero. In other words, they are able to predict the value of time that an animal remains in a shelter:

animal_typeCAT, animal_typeDOG, animal_typeWILDLIFE
intake_typeOWNER SURRENDER , intake_typeSTRAY
outcome_typeDIED , outcome_typeEUTHANIZED, outcome_typeFOSTER , outcome_typeRETURNED

To further assess the above aspect, we could investigate the p-value. Based on our knowledge, the lower the p-value for a regression coefficient, the stronger the case of rejecting the Null Hypothesis. In our model, the p-values for the regression coefficients corresponding to above covariates are close to zero and this constitutes a strong evidence that these are the covariates that we need to keep in our ultimate model which we did.

We fit a negative binomial model here:

```
fit3<- glm.nb(time_at_shelter~animal_type+intake_type+outcome_type+chip_status+month,
              data = dataset20)
summary(fit3)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type +
    outcome_type + chip_status + month, data = dataset20, init.theta = 1.050922744,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2124  -1.2374  -0.3709   0.2697   3.2889

Coefficients:
```

```
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.29295    1.13165   1.143   0.2532
animal_typeCAT                2.15545    1.12196   1.921   0.0547 .
animal_typeDOG                2.46527    1.12079   2.200   0.0278 *
animal_typeWILDLIFE           1.81693    1.16601   1.558   0.1192
intake_typeOWNER SURRENDER   -1.58362    0.13299 -11.907   <2e-16 ***
intake_typeSTRAY             -1.17072    0.12353  -9.477   <2e-16 ***
outcome_typeDIED             -1.27107    0.37500  -3.390   0.0007 ***
outcome_typeEUTHANIZED       -0.71708    0.06541 -10.964   <2e-16 ***
outcome_typeFOSTER           -0.27218    0.19453  -1.399   0.1618
outcome_typeRETURNED TO OWNER -1.72346    0.09030 -19.086   <2e-16 ***
chip_statusSCAN NO CHIP      -0.14139    0.07698  -1.837   0.0663 .
chip_statusUNABLE TO SCAN    -0.27869    0.16440  -1.695   0.0900 .
month2                       -0.11840    0.15584  -0.760   0.4474
month3                       -0.09213    0.15218  -0.605   0.5449
month4                       -0.24728    0.14782  -1.673   0.0944 .
month5                       -0.09839    0.14545  -0.676   0.4987
month6                        0.09050    0.14226   0.636   0.5247
month7                       -0.14917    0.14305  -1.043   0.2971
month8                       -0.38010    0.15001  -2.534   0.0113 *
month9                       -0.21585    0.15874  -1.360   0.1739
month10                       0.10131    0.14852   0.682   0.4951
month11                      -0.18748    0.15444  -1.214   0.2248
month12                      -0.10305    0.15879  -0.649   0.5164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0509) family taken to be 1)

    Null deviance: 2167.9  on 1464  degrees of freedom
Residual deviance: 1709.0  on 1442  degrees of freedom
AIC: 8043.1

Number of Fisher Scoring iterations: 1

              Theta:  1.0509
          Std. Err.:  0.0530

 2 x log-likelihood:  -7995.1460
```

In the original quasi-Poisson model, we detected that only the *p*-value for the month8 is significant (<0.05), hence, we removed the covariate of the month from the model.

```
fit4<- glm.nb(time_at_shelter~animal_type+intake_type+outcome_type,
              data = dataset20)
summary(fit4)
```

```
Call:
glm.nb(formula = time_at_shelter ~ animal_type + intake_type + 
    outcome_type, data = dataset20, init.theta = 1.023256193, 
    link = log)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2273  -1.2870  -0.3668   0.2898   3.4882  
```

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.09708    1.14668   0.957 0.338699
animal_typeCAT                 2.12358    1.13987   1.863 0.062461 .
animal_typeDOG                 2.45345    1.13884   2.154 0.031213 *
animal_typeWILDLIFE            1.69066    1.18341   1.429 0.153108
intake_typeOWNER SURRENDER    -1.58046    0.13373 -11.818  < 2e-16 ***
intake_typeSTRAY              -1.19621    0.12419  -9.632  < 2e-16 ***
outcome_typeDIED              -1.34650    0.37706  -3.571 0.000356 ***
outcome_typeEUTHANIZED        -0.71878    0.06542 -10.987  < 2e-16 ***
outcome_typeFOSTER            -0.32752    0.19418  -1.687 0.091671 .
outcome_typeRETURNED TO OWNER -1.69339    0.08857 -19.120  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for Negative Binomial(1.0233) family taken to be 1)

    Null deviance: 2129.8  on 1464  degrees of freedom
Residual deviance: 1706.7  on 1455  degrees of freedom
AIC: 8043.4

Number of Fisher Scoring iterations: 1

              Theta:  1.0233
          Std. Err.:  0.0511

 2 x log-likelihood:  -8021.3640
```

In this model, we kept only the significant slope coefficients, and we are now able to evaluate the significance of the estimated regression model as a whole and assess whether all regression coefficients as a group explain the variation in the dependent variable. To achieve this, we employee the concept of deviance. The residual deviance for the fitted model is 1706 on 1445 degrees of freedom. The value of the deviance is higher than the degrees of freedom however not so high to support that the model does not fit the data well. For this reason, given that the residual deviance is close to the value of the degrees of freedom we conclude that the model does a good join explaining the data.

We also calculated the AIC value for the two negative binomial models, the AIC value of the second model is lower than the AIC value of the first model and this constitutes a strong evidence that the second binomial model is better than the first.

6 Conclusion

The negative binomial model fits the data compared to Poisson model which exhibits overdispersion.

The significant explanatory variables with respect to the negative binomial model are the following:

animal_typeCAT, animal_typeDOG, animal_typeWILDLIFE
intake_typeOWNER SURRENDER , intake_typeSTRAY
outcome_typeDIED , outcome_typeEUTHANIZED, outcome_typeFOSTER , outcome_typeRETURNED

The AIC of the second binomial model is lower than this of the first binomial model and for this reason we can support that our ultimate negative binomial model fits the data with the most efficient way.

Hence, we can conclude that animal types, intake types and outcome types will affect the number of days staying at the shelter.