Influence on days an animal spends in shelter

Nokia

Table of contents

1 Introduction

A study is conducted using dataset from the Dallas animal shelter, aiming to uncover the factors affecting how long animals remain at the shelter before a final decision on their outcome is made. The insights gained may help improve animal welfare and shelter efficiency.

2 Description of the Dataset

Each of the 5 datasets contain a variety of information relating to each animal admitted to the shelter.

Table 1: Description of the Dataset

Variable	Description
type	The type of animal admitted to the shelter
month	Month the animal was admitted
year	Year the animal was admitted
intake	Reason for the animal being admitted
outcome	Final outcome for the admitted animal
chip	Did the animal have a microchip with owner information
duration	Days spent at the shelter between being admitted and the final outcome

Table 1 presents details about the variables in dataset. It is noteworthy that our dataset is complete, with no missing values across variables. We focus on the duration animals stay at the shelter before reaching their final outcome, which is quantified by the variable 'duration'.

3 Exploratory Data Analysis

3.1 numerical variables

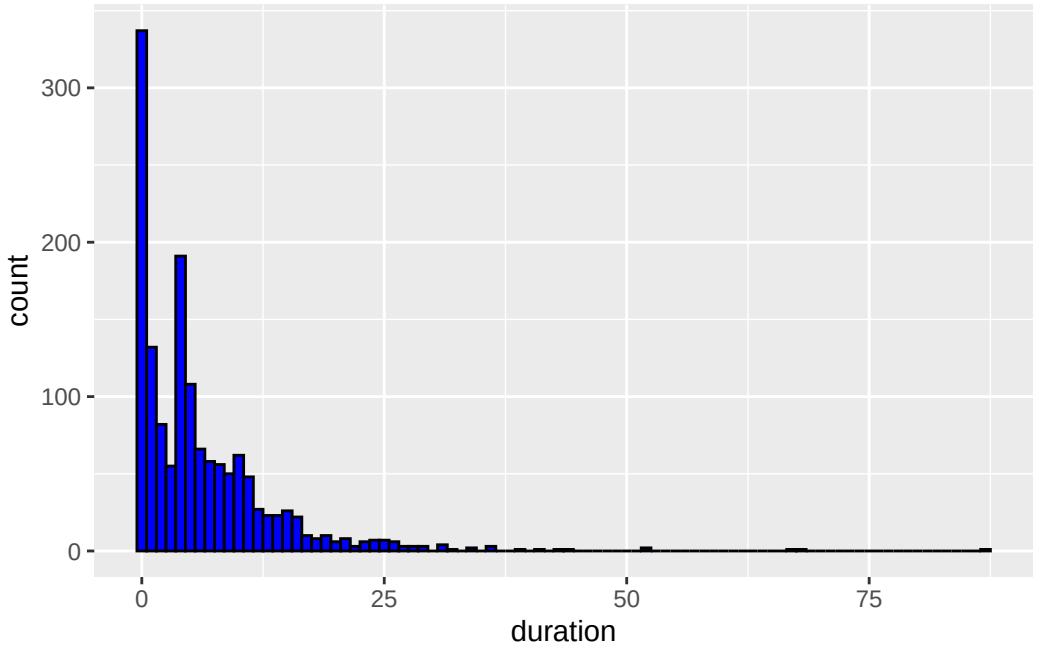
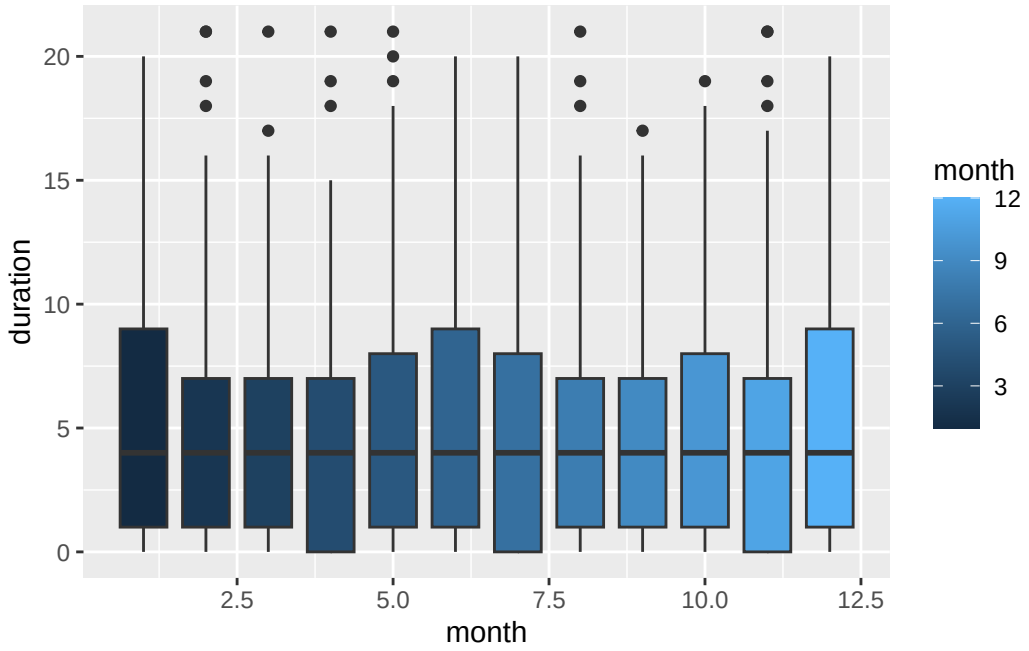
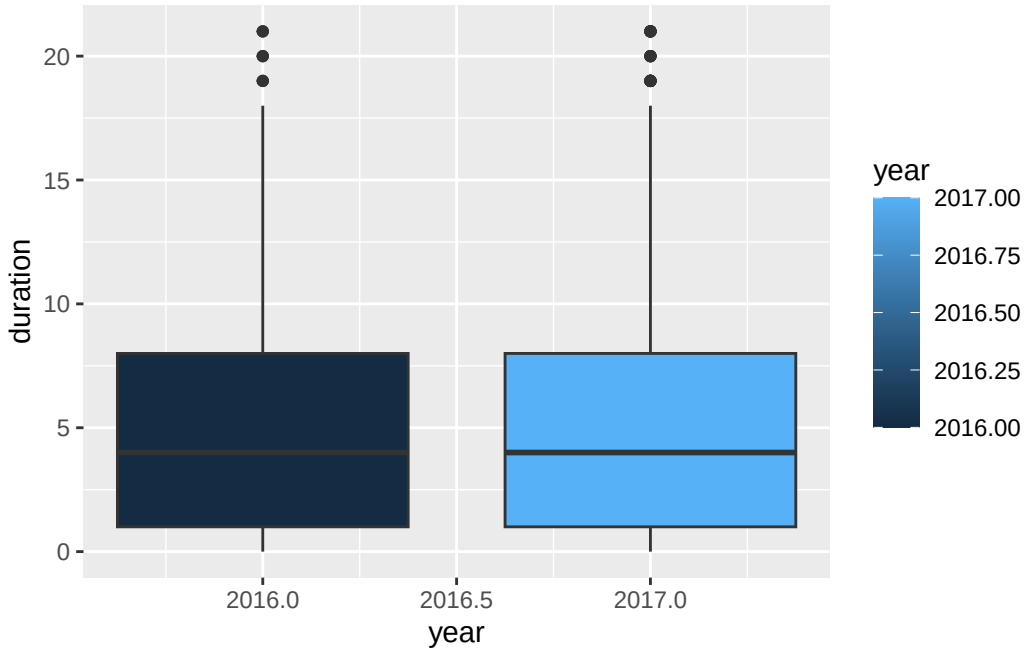


Figure 1: distribution of duration

As seen in Figure 1, most animals stay in the shelter for a short period of time, while a few animals stay in the shelter for a long time. Although the histogram itself does not display outliers, the long tail suggests that there are relatively few cases that have stayed in the shelter for a very long time, which may be outliers.





We can see duration vary little across different months and years, so we may consider ignore

these two variables when modelling.

3.2 categorical variables

Table 2: type of animals

Var1	Freq
BIRD	0.0020478
CAT	0.2075085
DOG	0.7822526
WILDLIFE	0.0081911

Birds and wildlife only make up the 0.002 and 0.008 of the datasets, so they don't have a significant impact on the results. We then remove these observations.

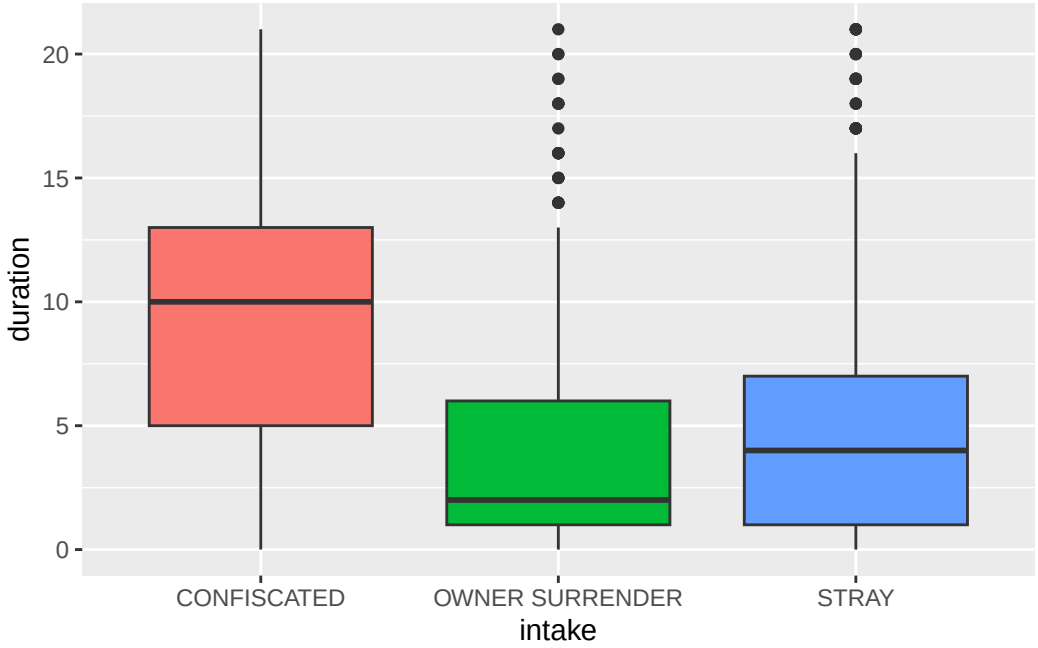


Figure 2: distribution of intake

Remove corresponding outliers according to the reason the animal was admitted to the shelter.

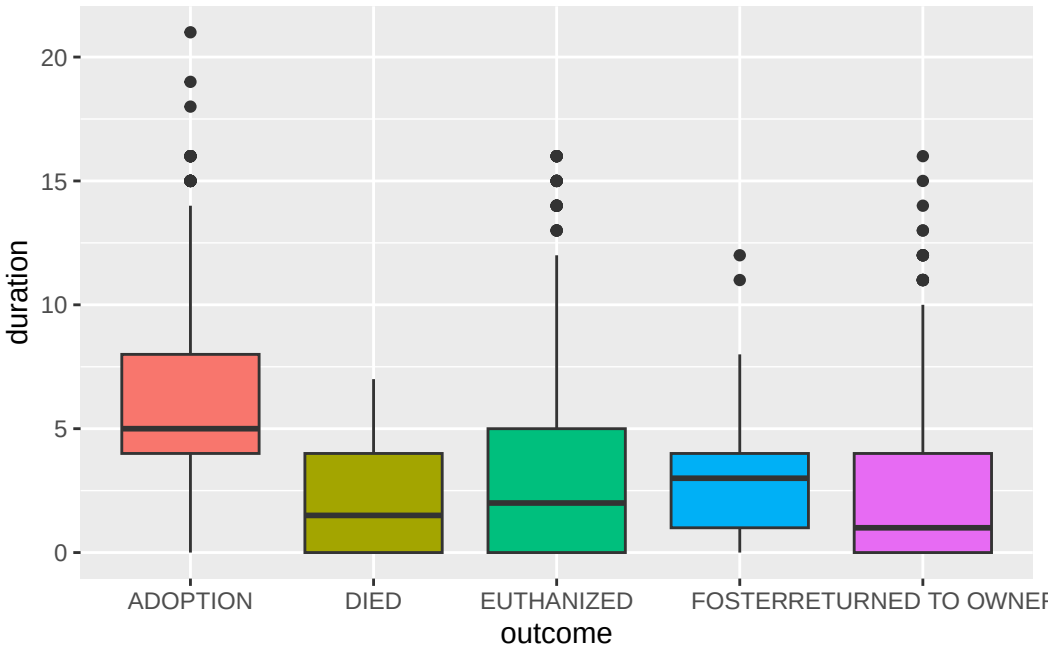


Figure 3: distribution of outcome

Remove corresponding outliers according final outcome.

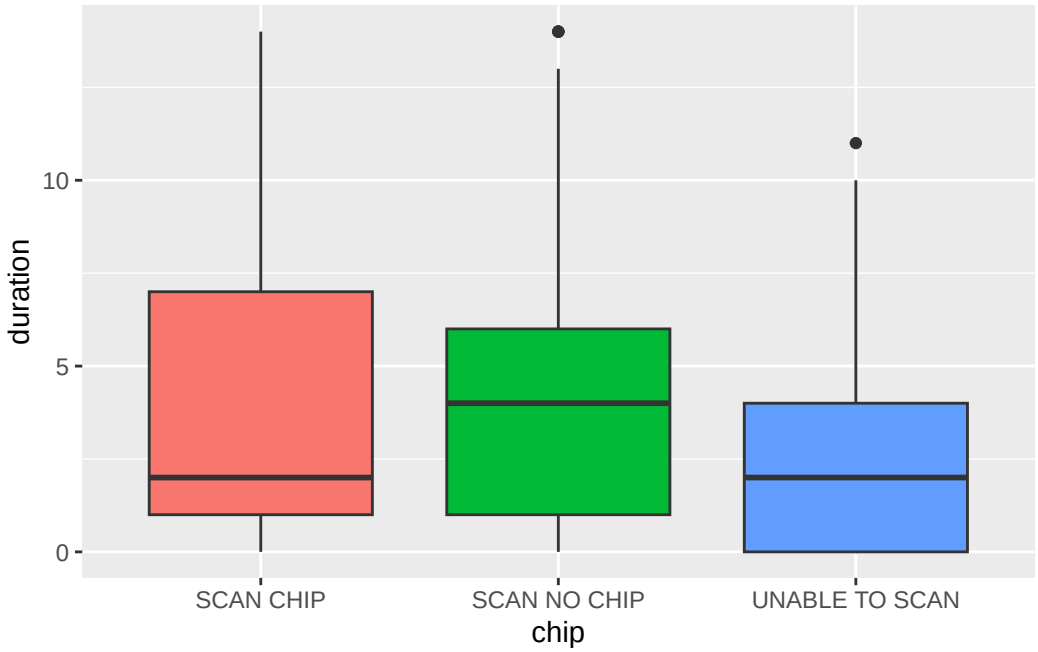


Figure 4: distribution of chip

Remove corresponding outliers according whether the animal have a microchip with owner information.

Variable	P_Value
type	4.4e-06
intake	0.0e+00
outcome	0.0e+00
chip	0.0e+00

- The chi-square test for all categorical variables have p-value less than 0.05. So they are all significant associated with the response variable duration.

	GVIF	Df	GVIF^(1/(2*Df))
type	1.129852	1	1.062945
month	2.287867	1	1.512570
year	2.280442	1	1.510113
intake	1.211155	2	1.049059
outcome	1.357479	4	1.038943
chip	1.140194	2	1.033343

- VIF values for type, chip, intake and outcome are all less than 5, suggesting that the linearity problems between them are minor and are unlikely to affect the model's results
- However month and year have a relatively high value compared to other variables, which also contribute to the reason we drop them.

4 Formal Data Analysis

In this section, several formal statistical models will be conducted using a Generalized Model Analysis to infer the relationships between variables.

4.1 Poisson Model

Since the response variable represents the count data, the Poisson regression model is a starting point because it is the simplest model for the type of the data. It was conducted by this result:

Before presenting and interpreting the derived equation from our GLM analysis, it is important to note that validating the model is underlying assumptions is a critical step. Ensuring the assumptions hold true bolsters the reliability of our findings. Here's the equation based on the initial analysis, subject to further validation:

$$
\begin{aligned}
\log(\text{Expected Count of Time at Shelter}) = 2.48068 \\
&+ 0.18080 \times \text{TypeDog} \\
&- 1.11596 \times \text{IntakeOwnerSurrender} \\
&- 0.63153 \times \text{IntakeStray} \\
&- 0.99932 \times \text{OutcomeDied} \\
&- 0.70378 \times \text{OutcomeEuthanized} \\
&- 0.69878 \times \text{OutcomeFoster} \\
&- 1.31990 \times \text{OutcomeReturnedToOwner} \\
&- 0.14291 \times \text{ChipScanNoChip} \\
&- 0.38348 \times \text{ChipUnableToScan}
\end{aligned}
$$

where,

- TypeDog is the indicator variable for the animal types, taking the value 1 if the condition is true and 0 otherwise with the baseline category TypeCat.
- IntakeOwnerSurrender and IntakeStray are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category IntakeConfiscated.
- OutcomeDied, OutcomeEuthanized, OutcomeFoster, OutcomeReturnedToOwner are indicator variables for the outcome, taking the value 1 if the condition is true and 0 otherwise with the baseline category OutcomeAdoption.
- ChipScanNoChip and ChipUnableToScan are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category ChipScanChip.

The significance of a Generalized Linear Model (GLM) can be assessed by examining the p-values associated with the coefficients of the model, which indicate the strength of evidence against the null hypothesis that the corresponding coefficients are equal to zero. A small p-value (in this case, compared to the significance level 0.05) indicates strong evidence against the null hypothesis, suggesting that it is unlikely to observe such a significant effect if the predictor really had no impact on the response variable.

Conversely, a large p-value suggests insufficient evidence to reject the null hypothesis, indicating that the predictor may not have a significant effect on the response variable.

Derived from the model, all variables are significant predictors of the time an animal spends in the shelter.

4.1.1 Model Diagnostics and Assumptions Checking

The analysis utilized a Poisson regression model to investigate the impact of factors such as, including AnimalType, IntakeType, OutcomeType and ChipStatus on the duration of stay in a shelter. To ensure the robustness of our model, we conducted a thorough diagnostic evaluation using several methods. We checked the dispersion parameter to evaluate the presence of overdispersion, which would violate the Poisson assumption of equal mean and variance. We also examined the deviance to assess the model's goodness-of-fit, with a value near the degrees of freedom indicating a well-fitting model. Additionally, we conducted a residual analysis, including plotting residuals against fitted values and creating Q-Q plots of standardized residuals, to detect any systematic patterns that might indicate model misspecification. Together, these diagnostics help validate our model and confirm the reliability of our conclusions.

Dispersion

In a Poisson generalized linear model (GLM), the dispersion parameter is assumed to be fixed at 1. This assumption is crucial because the Poisson distribution is characterized by its mean being equal to its variance, a property known as *equidispersion*. When the observed variance is greater than the mean, the data exhibit over-dispersion. This is more common in practice and can arise from various sources, such as unobserved heterogeneity among observations, excess zeros, or violations of the Poisson model's assumptions (such as events not occurring independently).

Then, the hypothesis being tested relates to the dispersion of the data with a significance level 5% is stated below:

- Null Hypothesis (H0):

  The null hypothesis posits that the true dispersion parameter equals 1. This means the data follow a Poisson distribution accurately, where the mean and variance of the count data are equal (equidispersion). The model is adequately specified, and there's no extra variability in the data beyond what the Poisson model accounts for.

- Alternative Hypothesis (Ha):

  The alternative hypothesis suggests that the true dispersion parameter is greater than 1, indicating over-dispersion in the data. Overdispersion occurs when the observed variance in the count data is greater than what the Poisson model would predict based on the mean.

```
    Overdispersion test

data:  glm_poisson
z = 14.143, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
```

```
sample estimates:
dispersion
  2.584389
```

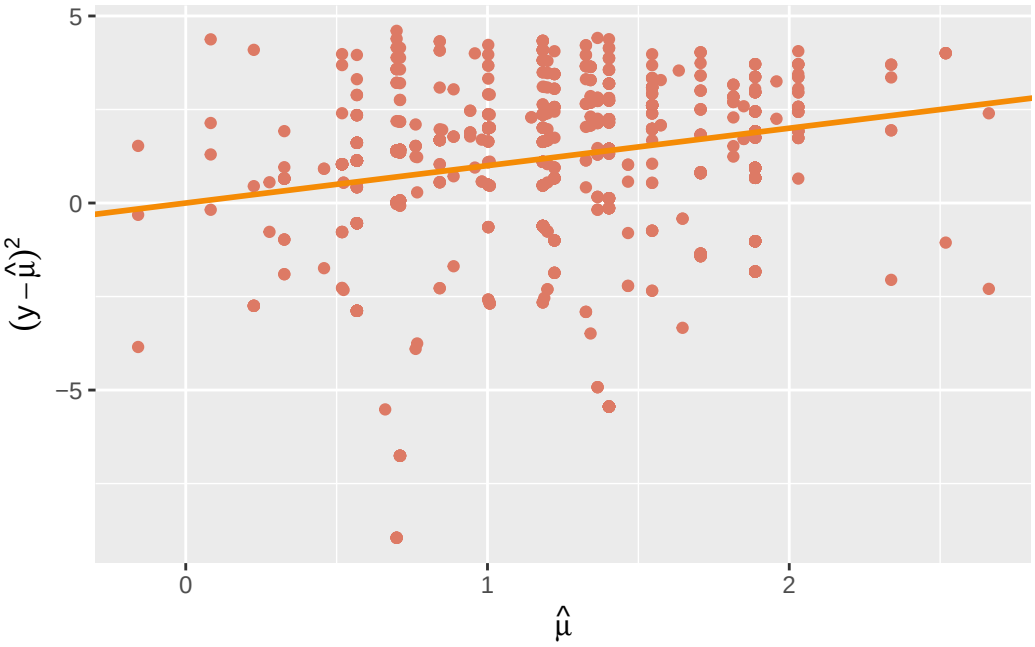


Figure 5: residual

Interpretation:

- p-value is extremely small, below the significance level, which indicates that the result is highly statistically significant. Hence, this supports the alternative hypothesis that the true dispersion is greater than 1.
- The sample estimated - dispersion is 2.584389 which is substantially greater than 1, indicating over-dispersion in the data.
- Meanwhile, a common way to assess dispersion in a Poisson model is through a plot of the residuals which is displayed in Figure 5. It suggests that as the predicted values increase, the variance of the residuals also increases, which is a classic sign of over-dispersion. Over-dispersion is when the variance is greater than the mean, which often occurs with count data.
- Given this evidence of over-dispersion, it would be prudent to consider alternative models that can accommodate the extra variability, such as a Negative Binomial regression model which will be conducted later.

Deviance:

Deviance is used to quantify the difference between a fitted model and a perfect model (a saturated model that fits the data exactly). The deviance essentially quantifies the discrepancy between the observed data and the values predicted by the model under the assumption that the model is correct. A lower deviance indicates a better fit of the model to the data.

Hence, based on the generated model result, we got:

- Null Deviance: this represents the goodness of fit of a model that includes only the intercept (no predictors). It is 5122.4 on 1449 degrees of freedom.
- Residual Deviance: This is the goodness of fit of the model that includes predictors (AnimalType, IntakeType, OutcomeType and ChipStatus). It is 3759.4 on 1440 degrees of freedom.

The residual deviance is used to assess the fit of the model to the data. For a well-fitting model, the residual deviance should be close to the degrees of freedom (relatively low). Here, the residual deviance (3759.4) is quite high compared to the Chi-square distribution with the 1440 - degrees of freedom, which might indicate that the model does not fit the data perfectly. This could be a sign of over-dispersion or that the model is missing some key explanatory variables.

Residuals Analysis

- Plotting residuals vs. fitted values

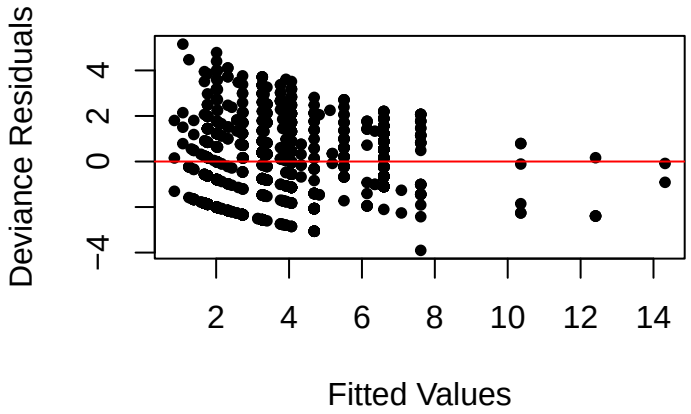


Figure 6: Residuals vs Fitted Values

Based on the plot Figure 6 The residuals plot here shows that as the fitted values increase, the spread of the residuals also increases. The increase in the spread of residuals as the fitted values increase suggests that the variance of the residuals is not constant. This pattern indicates potential over-dispersion in the data where the variance exceeds the mean.

- Scale location plot (spread vs. level plot)

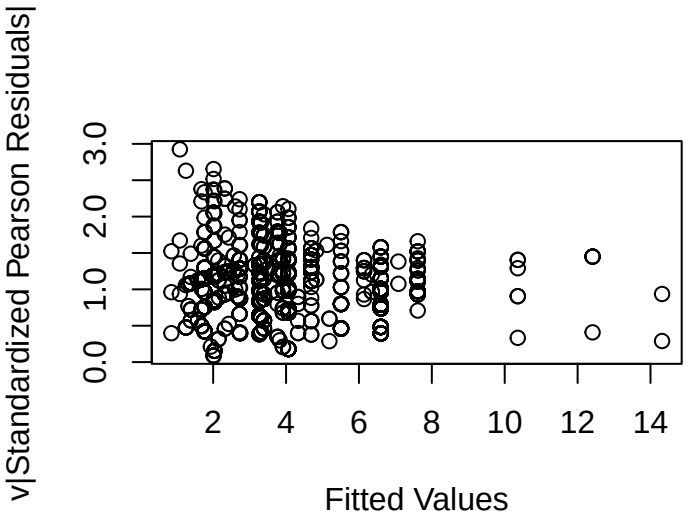


Figure 7: Scale-Location

The plot Figure 7 represents a scatter plot of the square root of the absolute standardized Pearson residuals versus the fitted values from a Generalized Linear Model (GLM). It will give the information about the move of the variance across different levels of the mean, making it easier to see whether there is a consistent spread of residuals across all counts or whether the spread increases with the count (which would indicate over-dispersion).

This pattern suggests possible over-dispersion in the data and will be advisable to consider model alternatives like the negative binomial regression.

- Residual vs. leverage

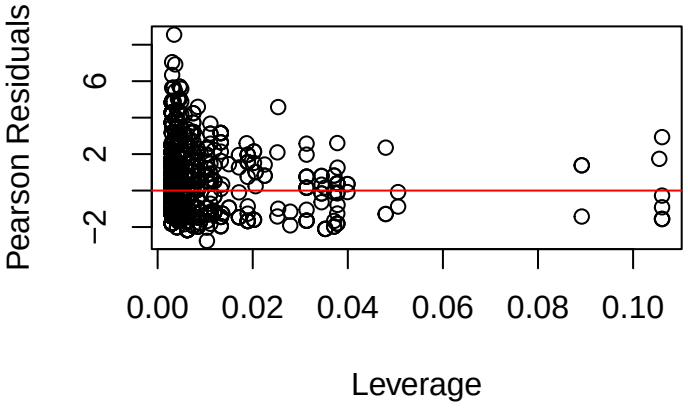


Figure 8: Residuals vs Leverage

Most of the data points cluster at the left side of the plot, suggesting that these observations have lower leverage and smaller residuals, shown by the plot Figure 8. Then, they don't have an undue influence on the model which seems the model fits well for the majority of the data.

However, the presence of points with high residuals and high leverage in the right side suggests there are some exceptions where the model's fit is not as good. Hence, it would be that while the model appears to provide a good fit for most of the data, there are specific observations that require further investigation to ensure the model's robustness and to potentially improve its accuracy.

When diagnostic plots provide different insights, it is crucial to consider the overall evidence and the specific research context. If over-dispersion is a consistent concern across multiple diagnostics (like the Residuals vs. Fitted Values Plot and the Scale-Location Plot), it often outweighs indications from other plots that the model might be adequate for most data points. Therefore, even if the Residuals vs. Leverage Plot suggests the model is mostly fitting well, the evidence of over-dispersion from the other plots is a strong indicator that the Poisson model might not be the best choice. Exploring alternative models like the Negative Binomial, which can accommodate over-dispersion, is likely a prudent step to improve model fit and ensure more reliable inference from the model.

4.2 Negative Binomial Model

It will be conducted the Negative Binomial Model which is a good alternative model when the assumption of equidispersion (mean equals variance) in Poisson regression doesn't hold. The result was shown by:

$$\begin{aligned}\log(\text{Expected Count of Time at Shelter}) = {} & 2.58882 \\ & + 0.22863 \times \text{TypeDog} \\ & - 1.23810 \times \text{IntakeOwnerSurrender} \\ & - 0.75541 \times \text{IntakeStray} \\ & - 0.99159 \times \text{OutcomeDied} \\ & - 0.75098 \times \text{OutcomeEuthanized} \\ & - 0.67674 \times \text{OutcomeFoster} \\ & - 1.39817 \times \text{OutcomeReturnedToOwner} \\ & - 0.15081 \times \text{ChipScanNoChip} \\ & - 0.40916 \times \text{ChipUnableToScan}\end{aligned}$$

where,

- TypeDog is the indicator variable for the animal types, taking the value 1 if the condition is true and 0 otherwise with the baseline category TypeCat.

- IntakeOwnerSurrender and IntakeStray are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category IntakeConfiscated.
- OutcomeDied, OutcomeEuthanized, OutcomeFoster, OutcomeReturnedToOwner are indicator variables for the outcome, taking the value 1 if the condition is true and 0 otherwise with the baseline category OutcomeAdoption.
- ChipScanNoChip and ChipUnableToScan are indicator variables for the intake types, taking the value 1 if the condition is true and 0 otherwise with the baseline category ChipScanChip.

Derived from the model, all variables are significant predictors of the time an animal spends in the shelter.

4.2.1 Model Diagnostics and Assumptions Checking

To ensure the reliability and validity of our Negative Binomial regression model findings, it is imperative to rigorously examine the underlying assumptions through a series of diagnostic checks.

Over-dispersion Check

Attained from the model summary, we can draw the result that the parameter Theta: 2.189 along with a relatively small standard error Std.Err: 0.16 which indicates that the model has identified and is accounting for over-dispersion in the data. It suggest that the Negative Binomial model is a suitable choice for the data, given the presence of over-dispersion as a positive outcome in terms of assumptions checking for the model.

Residual Deviance

Deviance in GLMs is a measure of the goodness-of-fit of a model. It is based on the likelihood function and compares two models between the fitted model and a reference model. The lower the deviance, the closer the fitted model is to the reference model in terms of likelihood. Comparing the residual deviance to the degrees of freedom helps assess if the fitted model is adequately fitting the data without overfitting. If the residual deviance is close to the degrees of freedom for the fitted model, it suggests that the model is adequately fitting the data.

Then, in the model summary, attained the Residual Deviance: 1795.2 with the 1440 degrees of freedom and the Null Deviance: 2316.9 with the 1449 degrees of freedom, this shows how much better the model fits the data compared to the model with only the intercept.

Residual Analysis

In evaluating the adequacy of our Negative Binomial regression model, a series of residual diagnostic plots were examined, including Residuals vs Fitted, Normal Q-Q, Scale-Location, and

Residuals vs Leverage, each offering insights into different aspects of model fit and underlying assumptions.

- Residuals vs. Fitted Values Plot

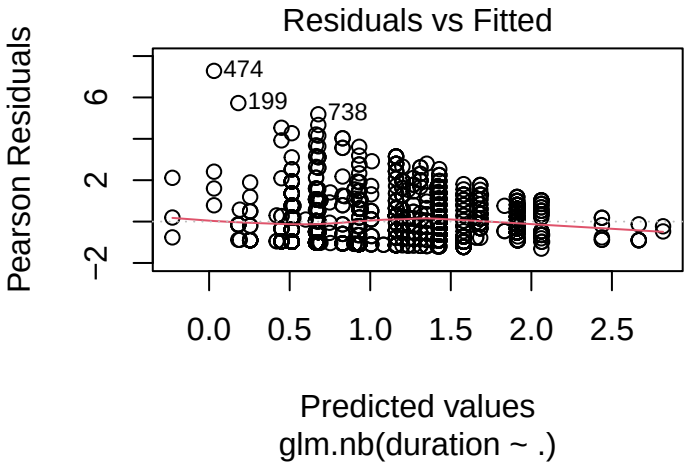


Figure 9: Residuals vs Fitted Values

Shown in the plot Figure 9, the residuals seem to increase slightly in variance with smaller fitted values, suggesting potential mild heteroscedasticity. However, since this is count data being modeled with a Negative Binomial regression, some of this is to be expected and the model accounts for it. There may be some potential outliers, but their influence would need further investigation, possibly with additional diagnostics like Cook's distance.

- Normal Q-Q Plot:

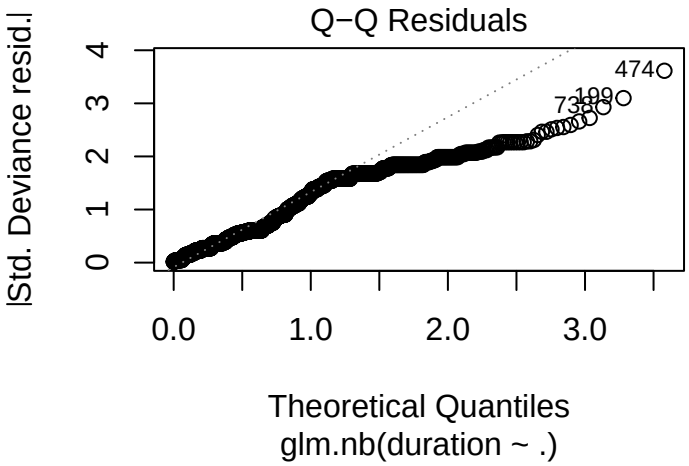


Figure 10: Normal Q-Q Plot

In the Q-Q plot Figure 10, the residuals deviate from the line at both ends, indicating that the residuals might not be following a normal distribution, which is a common issue for count data and is often the reason for using a Negative Binomial model instead of Poisson.

- Scale-Location Plot

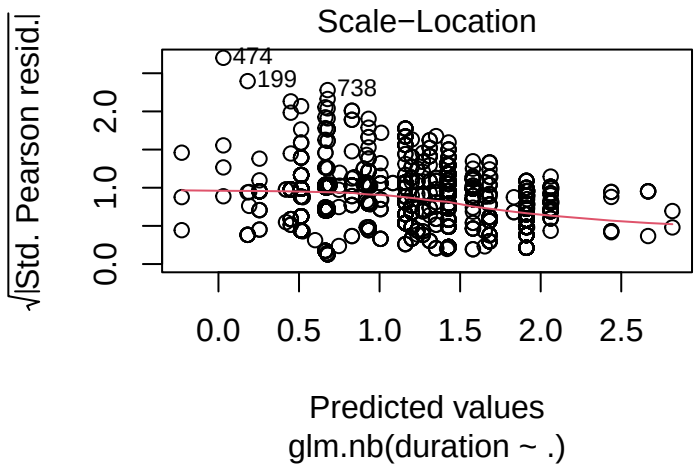


Figure 11: Scale-Location

The red line, captured on the plot Figure 11, shows some fluctuation but does not exhibit a clear or strong trend which indicates potential mild heteroscedasticity.

- Residuals vs. Leverage

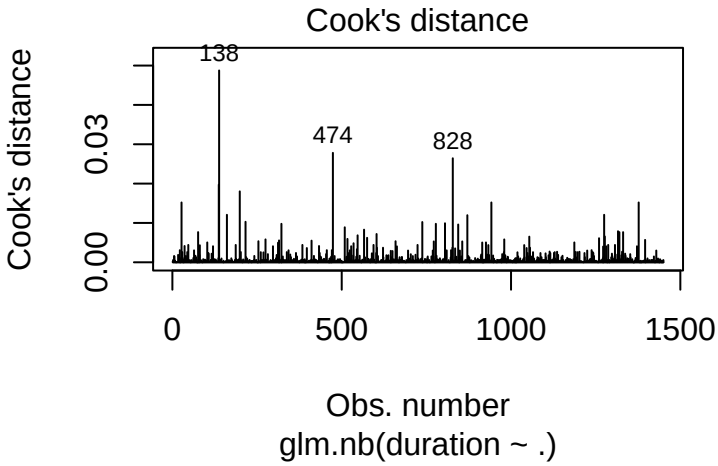


Figure 12: Residuals vs Leverage

Observations with higher Cook's distance values are worth examining because they might be outliers or influential data points that could unduly affect the model's coefficients, predictions, and overall fit. Based on the plot Figure 12, most observations have a Cook's distance close to zero, indicating they have little influence on the regression model. However, there are a few observations, labeled with their observation numbers, that stand out with higher Cook's distance values yet not necessarily a problem for the fitting model.

Summary of Model Diagnostic

In conclusion, the model appears to perform well in terms of fitting the central tendency and dispersion of the data, as suggested by the lack of systematic patterns in residuals and the appropriate handling of over-dispersion.

4.3 Model selection

We employ a backward elimination process guided by the Akaike Information Criterion (AIC) to identify the most parsimonious model that adequately explains the variation in the time animals spend at the shelter. This approach systematically removes the least significant predictors from the full model, optimizing the balance between model complexity and fit to the data.

```
Start:  AIC=6756.64
duration ~ type + intake + outcome + chip

          Df    AIC
<none>       6756.6
- chip     2 6763.7
- type     1 6769.3
- intake   2 6902.7
- outcome  4 7128.5
```

The output of the stepAIC function with the direction set to "backward" suggests that the most parsimonious model according to the Akaike Information Criterion (AIC) is actually the initial model. The result of <none> : 6756.6 as the lowest AIC indicates to remove no predictors from the current model.

5 Conclusion

This study analyzed the factors affecting the length of stay in the Dallas animal shelter. From exploratory data analysis, we found that most animals stay in shelters for a short time, and only a few stay for a long time.

In formal data analysis, we first tried the Poisson Regression Model, because the response variable represents the count data. However, diagnostic tests showed that there is over-dispersion in the data, suggesting that the Poisson Regression Model may not be the best option. Then we used the Negative Binomial Model, which adapts to the over-dispersion phenomenon.

Diagnostic analyses of Negative Binomial Model reveal that the model performs well, despite mild heteroscedasticity and potential outliers.

In summary, our analysis shows that lots of important predictive variables influence the length of time an animal remain at the shelter before a final decision on their outcome is made, including the type of animal, the reason it was sent to the shelter, the final outcome, and if the animal has a microchip with owner information.