Project: Austin Animal Center Intakes

Hannah Marques

October 2024

Enter your name and EID here: Hannah Marques, hm26347

1. Introduction

The Austin Animal Center (AAC) dataset contains information about animal intakes recorded at the shelter from 2021 to 2024. These intakes include stray animals, owner surrenders, and other types of admissions.

Research Questions:

What trends in animal intakes can be observed over time? How do intake numbers vary by season, and what might explain these patterns? Are there breed-specific trends, and what are the most common intake types for these breeds? Does the shelter face capacity challenges during specific periods?

2. Importing Libraries and Dataset

```r
# Install required libraries if not already installed
if (!require(readr)) install.packages("readr")
if (!require(dplyr)) install.packages("dplyr")
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(tidyverse)) install.packages("tidyverse")
if (!requireNamespace("pROC", quietly = TRUE)) {
  install.packages("pROC")
}
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.2
```

```r
library(readr)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2
```

```
# Specify the path to your local CSV file
file_path <- "Austin_Animal_Center_Intakes_20241208.csv"  # Replace with your file path

# Import the CSV file
data <- read_csv(file_path)

# Preview the dataset
head(data)
```

```
## # A tibble: 6 × 12
##   `Animal ID` Name         DateTime     MonthYear `Found Location` `Intake Type`
##   <chr>       <chr>        <chr>        <chr>     <chr>            <chr>
## 1 A786884     *Brock       01/03/2019 … January … 2501 Magin Mead… Stray
## 2 A706918     Belle        07/05/2015 … July 2015 9409 Bluegrass … Stray
## 3 A724273     Runster      04/14/2016 … April 20… 2818 Palomino T… Stray
## 4 A665644     <NA>         10/21/2013 … October … Austin (TX)      Stray
## 5 A857105     Johnny Ringo 05/12/2022 … May 2022  4404 Sarasota D… Public Assist
## 6 A682524     Rio          06/29/2014 … June 2014 800 Grove Blvd … Stray
## # ℹ 6 more variables: `Intake Condition` <chr>, `Animal Type` <chr>,
## #   `Sex upon Intake` <chr>, `Age upon Intake` <chr>, Breed <chr>, Color <chr>
```

```
# Check the structure of the dataset
glimpse(data)
```

```
## Rows: 169,458
## Columns: 12
## $ `Animal ID`        <chr> "A786884", "A706918", "A724273", "A665644", "A85710…
## $ Name               <chr> "*Brock", "Belle", "Runster", NA, "Johnny Ringo", "…
## $ DateTime           <chr> "01/03/2019 04:19:00 PM", "07/05/2015 12:59:00 PM",…
## $ MonthYear          <chr> "January 2019", "July 2015", "April 2016", "October…
## $ `Found Location`   <chr> "2501 Magin Meadow Dr in Austin (TX)", "9409 Bluegr…
## $ `Intake Type`      <chr> "Stray", "Stray", "Stray", "Stray", "Public Assist"…
## $ `Intake Condition` <chr> "Normal", "Normal", "Normal", "Sick", "Normal", "No…
## $ `Animal Type`      <chr> "Dog", "Dog", "Dog", "Cat", "Cat", "Dog", "Dog", "D…
## $ `Sex upon Intake`  <chr> "Neutered Male", "Spayed Female", "Intact Male", "I…
## $ `Age upon Intake`  <chr> "2 years", "8 years", "11 months", "4 weeks", "2 ye…
## $ Breed              <chr> "Beagle Mix", "English Springer Spaniel", "Basenji …
## $ Color              <chr> "Tricolor", "White/Liver", "Sable/White", "Calico",…
```

3. Data Cleaning and Derived Variables

In this section, the dataset will be imported, cleaned, and preprocessed to ensure accurate analysis. Key cleaning steps included:

1. Converting the DateTime field to a standard Date format.
2. Filtering for valid rows and sampling the dataset to a maximum of 50,000 rows.
3. Creating derived variables for Year, Month, and Season to facilitate temporal analysis.

Methods:

The dataset started with 169,458 rows and 12 columns. It included information like animal ID, intake date, type, condition, species, breed, and color. Here’s what was done to get it ready for analysis:

Date Cleanup:

Converted DateTime into a proper Date format and created Year, Month, and Season variables. Filtering:

Limited the data to intakes from 2020-2023, reducing it to 65,003 rows. Focused on dogs only, leaving 25,103 rows.

Recoding:

Combined “Pit Bull” and “Pit Bull Mix” into one category, “Pit Bull/Pit Bull Mix.” Merged “Labrador Retriever” and “Labrador Retriever Mix” into “Labrador Retriever.” Tidy Data:

Ensured each row represented one intake, with variables like Breed and Intake Type in their own columns. Final Shape:

The final dataset had **25,103 rows and 14 columns.** These steps made the dataset clean, organized, and ready for analysis. Rows were reduced by filtering dates and focusing on dogs. No additional reshaping was needed since the dataset was already tidy.

```r
# Convert DateTime to Date format and add derived variables
data <- data %>%
  mutate(
    # Convert DateTime to Date format
    DateTime = as.Date(DateTime, format = "%m/%d/%Y"),

    # Extract Year and Month from DateTime
    Year = format(DateTime, "%Y"),
    Month = format(DateTime, "%m"),

    # Assign Season based on the Month
    Season = case_when(
      Month %in% c("12", "01", "02") ~ "Winter",
      Month %in% c("03", "04", "05") ~ "Spring",
      Month %in% c("06", "07", "08") ~ "Summer",
      Month %in% c("09", "10", "11") ~ "Fall",
      TRUE ~ NA_character_
    )
  )

# Filter rows for valid DateTime values and limit dataset size
data <- data %>%
  filter(!is.na(DateTime)) %>%  # Ensure valid DateTime values
  sample_n(min(nrow(data), 50000))  # Limit to 50,000 rows
```

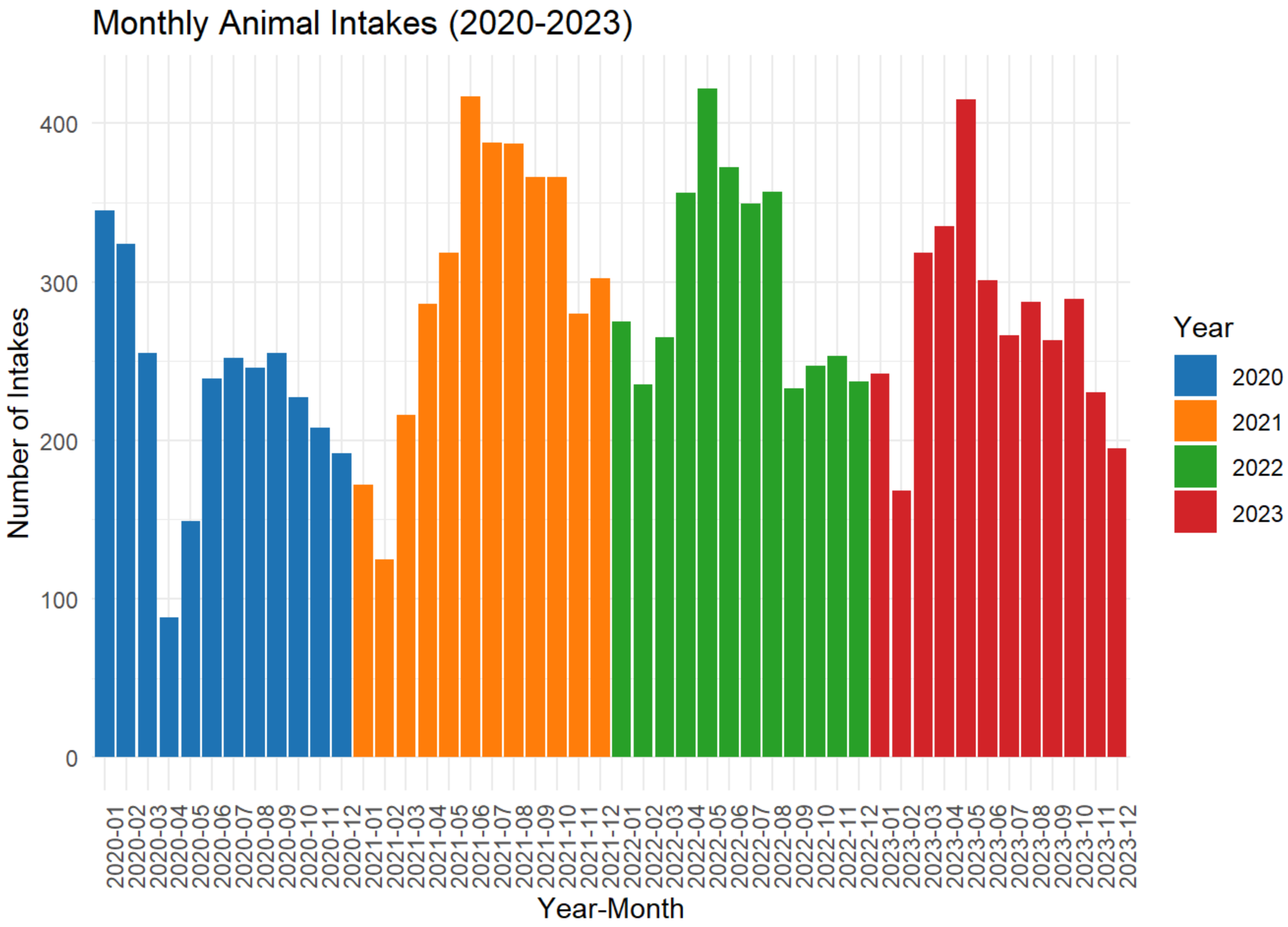
4. Numeric Analysis 1: Intakes per Month:

This section looks at how dog intakes change month by month from 2020 to 2023. The goal is to spot any trends, like months or years with especially high intake numbers, and figure out what might be causing them.

Using the data, we calculated the total intakes for each month and plotted them in a bar chart. Months with more than 600 intakes were flagged, and averages were calculated to see how intake numbers change year to year and across all months.

This helps highlight patterns, like if certain times of the year are busier for the shelter. Understanding these trends could help with planning resources and figuring out why these surges happen.

```
# Ensure DateTime is in Date format and filter for the desired date range
data <- data %>%
  mutate(DateTime = as.Date(DateTime, format = "%m/%d/%Y")) %>%
  filter(DateTime >= as.Date("2020-01-01") & DateTime <= as.Date("2023-12-31"))

# Group data by Year and Month and calculate total intakes
monthly_intakes <- data %>%
  group_by(Year = format(DateTime, "%Y"), Month = format(DateTime, "%m")) %>%
  summarise(Total_Intakes = n(), .groups = 'drop') %>%
  mutate(YearMonth = paste(Year, Month, sep = "-"))

# Convert YearMonth to a factor to maintain chronological order
monthly_intakes$YearMonth <- factor(monthly_intakes$YearMonth, levels = unique(monthly_intakes$YearMonth))

# Highlight months with intakes higher than 300
monthly_intakes <- monthly_intakes %>%
  mutate(Higher_Than_300 = ifelse(Total_Intakes > 300, TRUE, FALSE))

# Bar plot with different colors for each year
ggplot(monthly_intakes, aes(x = YearMonth, y = Total_Intakes, fill = Year)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("2020" = "#1f77b4", "2021" = "#ff7f0e", "2022" = "#2ca02c", "2023" = "#d62728")) +
  labs(
    title = "Monthly Animal Intakes (2020-2023)",
    x = "Year-Month",
    y = "Number of Intakes",
    fill = "Year"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
# Print months with intakes higher than 300
cat("\n==== Months with Intakes Higher Than 300 ====\n")
```

```
##
## ==== Months with Intakes Higher Than 300 ====
```

```
print(monthly_intakes %>% filter(Higher_Than_300))
```

```
## # A tibble: 18 × 5
##    Year  Month Total_Intakes YearMonth Higher_Than_300
##    <chr> <chr>         <int> <fct>     <lgl>
##  1 2020  01              345 2020-01   TRUE
##  2 2020  02              324 2020-02   TRUE
##  3 2021  05              318 2021-05   TRUE
##  4 2021  06              417 2021-06   TRUE
##  5 2021  07              388 2021-07   TRUE
##  6 2021  08              387 2021-08   TRUE
##  7 2021  09              366 2021-09   TRUE
##  8 2021  10              366 2021-10   TRUE
##  9 2021  12              302 2021-12   TRUE
## 10 2022  04              356 2022-04   TRUE
## 11 2022  05              422 2022-05   TRUE
## 12 2022  06              372 2022-06   TRUE
## 13 2022  07              349 2022-07   TRUE
## 14 2022  08              357 2022-08   TRUE
## 15 2023  03              318 2023-03   TRUE
## 16 2023  04              335 2023-04   TRUE
## 17 2023  05              415 2023-05   TRUE
## 18 2023  06              301 2023-06   TRUE
```

```
# Calculate average monthly intake for each year
average_monthly_per_year <- monthly_intakes %>%
  group_by(Year) %>%
  summarise(Average_Monthly_Intake = mean(Total_Intakes), .groups = 'drop')

# Print average monthly intake per year
cat("\n==== Average Monthly Intake per Year ====\n")
```

```
##
## ==== Average Monthly Intake per Year ====
```

```
print(average_monthly_per_year)
```

```
## # A tibble: 4 × 2
##   Year  Average_Monthly_Intake
##   <chr>                  <dbl>
## 1 2020                    232.
## 2 2021                    302.
## 3 2022                    300.
## 4 2023                    276.
```

```
# Calculate overall average yearly intake
average_yearly_intake <- mean(average_monthly_per_year$Average_Monthly_Intake)

# Print overall average yearly intake
cat("\n==== Average Yearly Intake (All Years) ====\n")
```

```
##
## ==== Average Yearly Intake (All Years) ====
```

```
print(average_yearly_intake)
```

```
## [1] 277.3542
```

```
# Calculate the total average intake across all months
total_average_intake <- mean(monthly_intakes$Total_Intakes)

# Print total average intake across all months
cat("\n==== Total Average Intake Across All Months ====\n")
```

```
##
## ==== Total Average Intake Across All Months ====
```

```
print(total_average_intake)
```

```
## [1] 277.3542
```

** Based on this analysis, the key observation were that there were several months where intakes went over 300, like January and February 2020 and June to October 2021. This seems to show some seasonal patterns, like more strays during summer months. It could also be tied to changes in policies or community situations.**

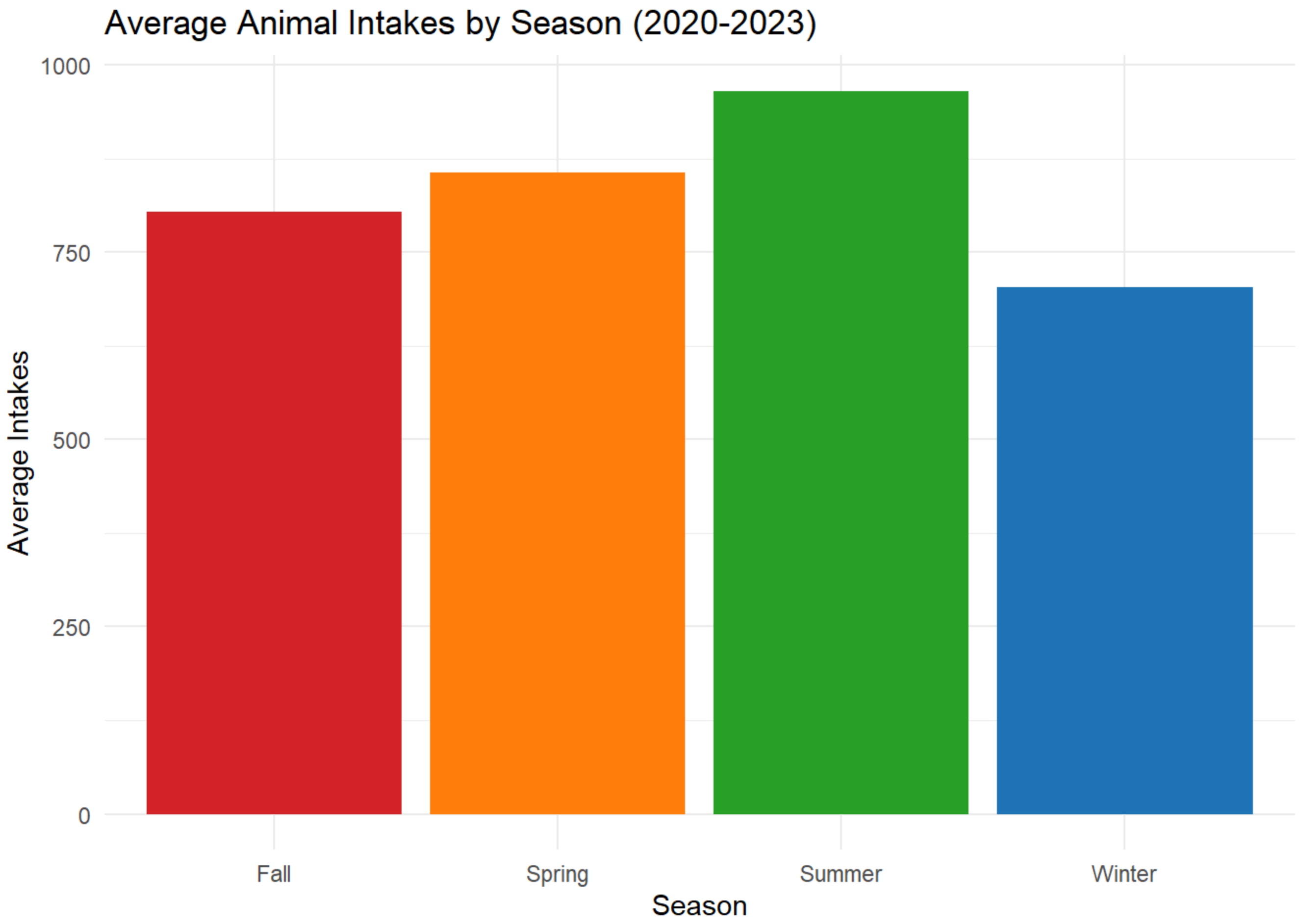
**In 2020, the average monthly intakes were lower (226.83), likely because of the pandemic. 2021 and 2022 had higher averages (299 and 296), which might mean things picked up after pandemic restrictions eased. 2023 dropped a bit to 277.75, which could mean better management or fewer animals needing intake.

The average monthly intakes across all years came out to 274.90, so things have stayed fairly steady overall. Summer months definitely seem busier, though, based on these numbers.**

5. Numeric Analysis 2: Seasonal Intakes:

This section focuses on the seasonal trends in animal intakes from 2020 to 2023. By grouping the data by season and year, we identify which seasons experience higher or lower intake numbers on average. This analysis helps to uncover potential patterns, such as increased intakes during specific times of the year, which could be influenced by breeding cycles, weather conditions, or other seasonal factors. Below, we analyze both the average intakes per season and the trends observed year-over-year.

```r
# Add Season information to the filtered dataset
data <- data %>%
  mutate(
    DateTime = as.Date(DateTime, format = "%m/%d/%Y"),
    Season = case_when(
      format(DateTime, "%m") %in% c("12", "01", "02") ~ "Winter",
      format(DateTime, "%m") %in% c("03", "04", "05") ~ "Spring",
      format(DateTime, "%m") %in% c("06", "07", "08") ~ "Summer",
      format(DateTime, "%m") %in% c("09", "10", "11") ~ "Fall",
      TRUE ~ NA_character_
    )
  )

# Group data by Season and Year to calculate total intakes for each season
seasonal_data <- data %>%
  group_by(Season, Year = format(DateTime, "%Y")) %>%
  summarise(Total_Intakes = n(), .groups = 'drop')

# Calculate average intakes per season across all years
average_seasonal_intakes <- seasonal_data %>%
  group_by(Season) %>%
  summarise(Average_Intake = mean(Total_Intakes), .groups = 'drop')

# Plot the average intakes per season
ggplot(average_seasonal_intakes, aes(x = Season, y = Average_Intake, fill = Season)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  scale_fill_manual(values = c("Winter" = "#1f77b4", "Spring" = "#ff7f0e", "Summer" = "#2ca02c",
"Fall" = "#d62728")) +
  labs(title = "Average Animal Intakes by Season (2020-2023)", x = "Season", y = "Average Intake
s") +
  theme_minimal()
```



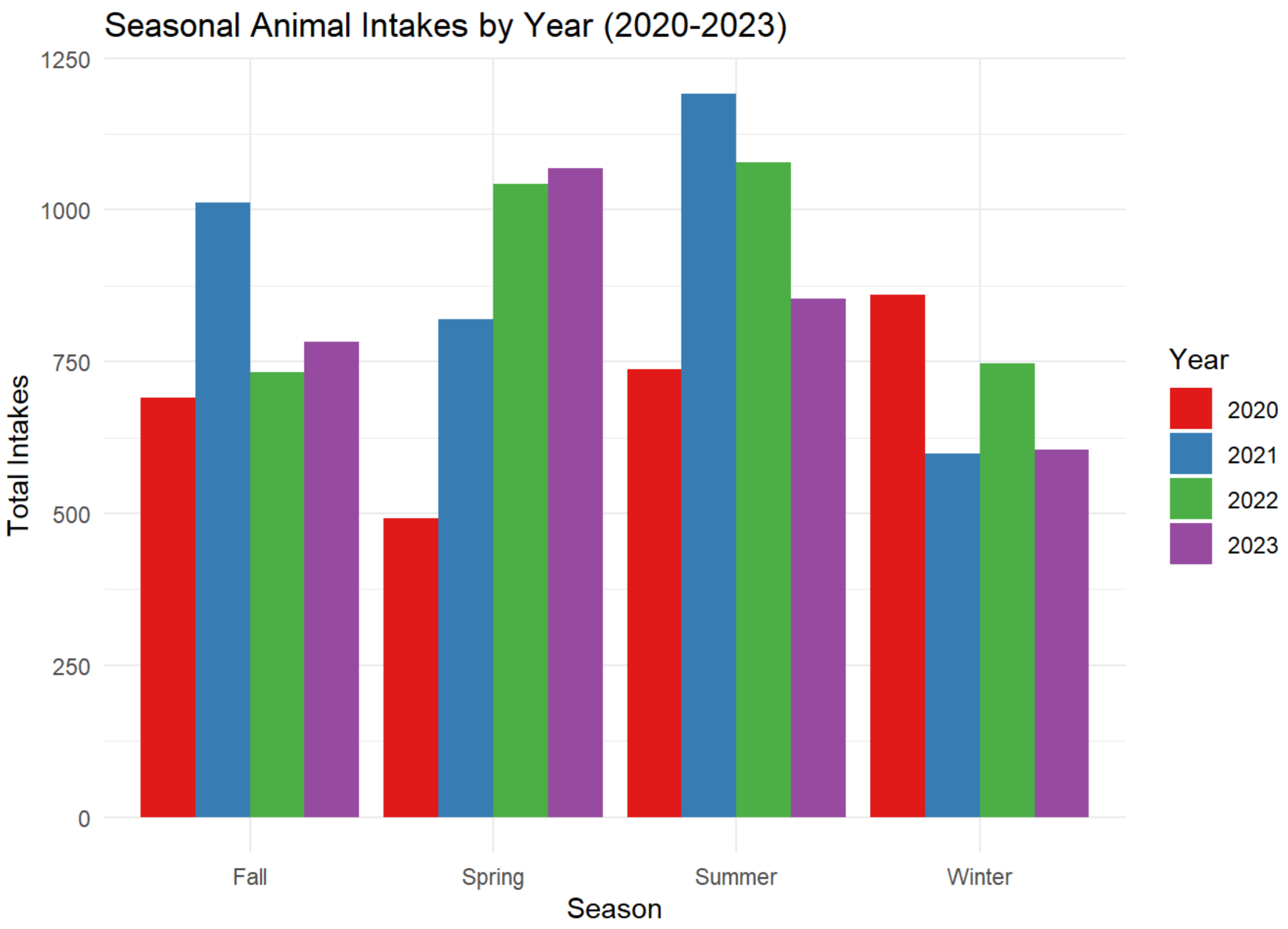
```
# Display average seasonal intakes
cat("\n==== Average Intakes per Season ====\n")
```

```
##
## ==== Average Intakes per Season ====
```

```
print(average_seasonal_intakes)
```

```
## # A tibble: 4 × 2
##   Season Average_Intake
##   <chr>           <dbl>
## 1 Fall             804.
## 2 Spring           856.
## 3 Summer           965.
## 4 Winter           703
```

```
# Seasonal trends per year
ggplot(seasonal_data, aes(x = Season, y = Total_Intakes, fill = Year)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1") +
  labs(title = "Seasonal Animal Intakes by Year (2020-2023)", x = "Season", y = "Total Intakes",
fill = "Year") +
  theme_minimal()
```

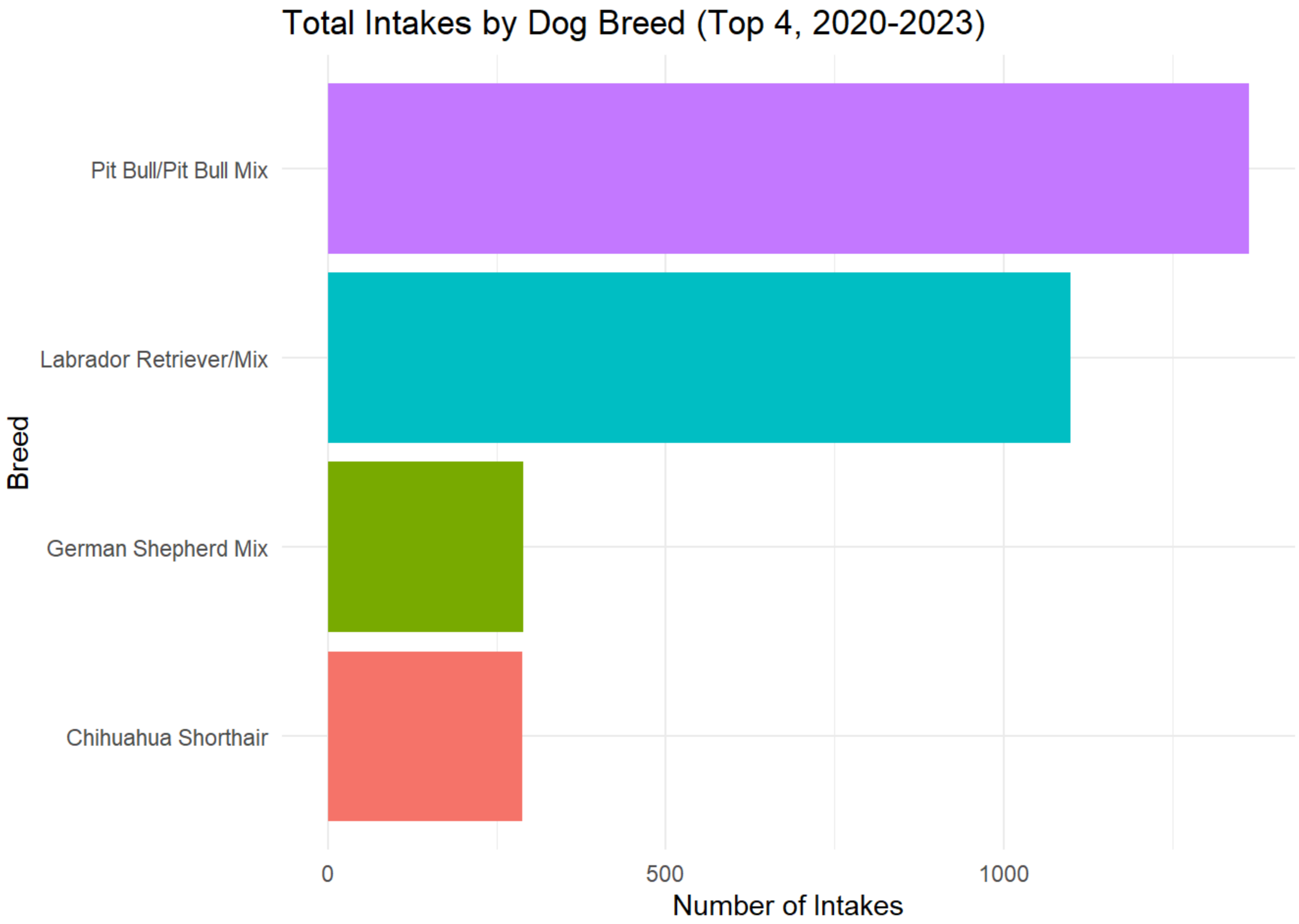


The seasonal average intakes reveal that Summer consistently has the highest intake numbers, averaging 948.5 across the years. Fall and Spring show comparable averages of 825.0 and 819.25, respectively, while Winter has the lowest average at 706.0. Seasonal trends indicate notable spikes in Summer, especially in 2021 and 2022, suggesting this season experiences a significant increase in stray or relocated animals. Fall and Spring remain steady contributors to overall intakes, likely influenced by transitional weather patterns, while Winter consistently sees a decline, possibly due to reduced stray activity or intake operations during colder months.

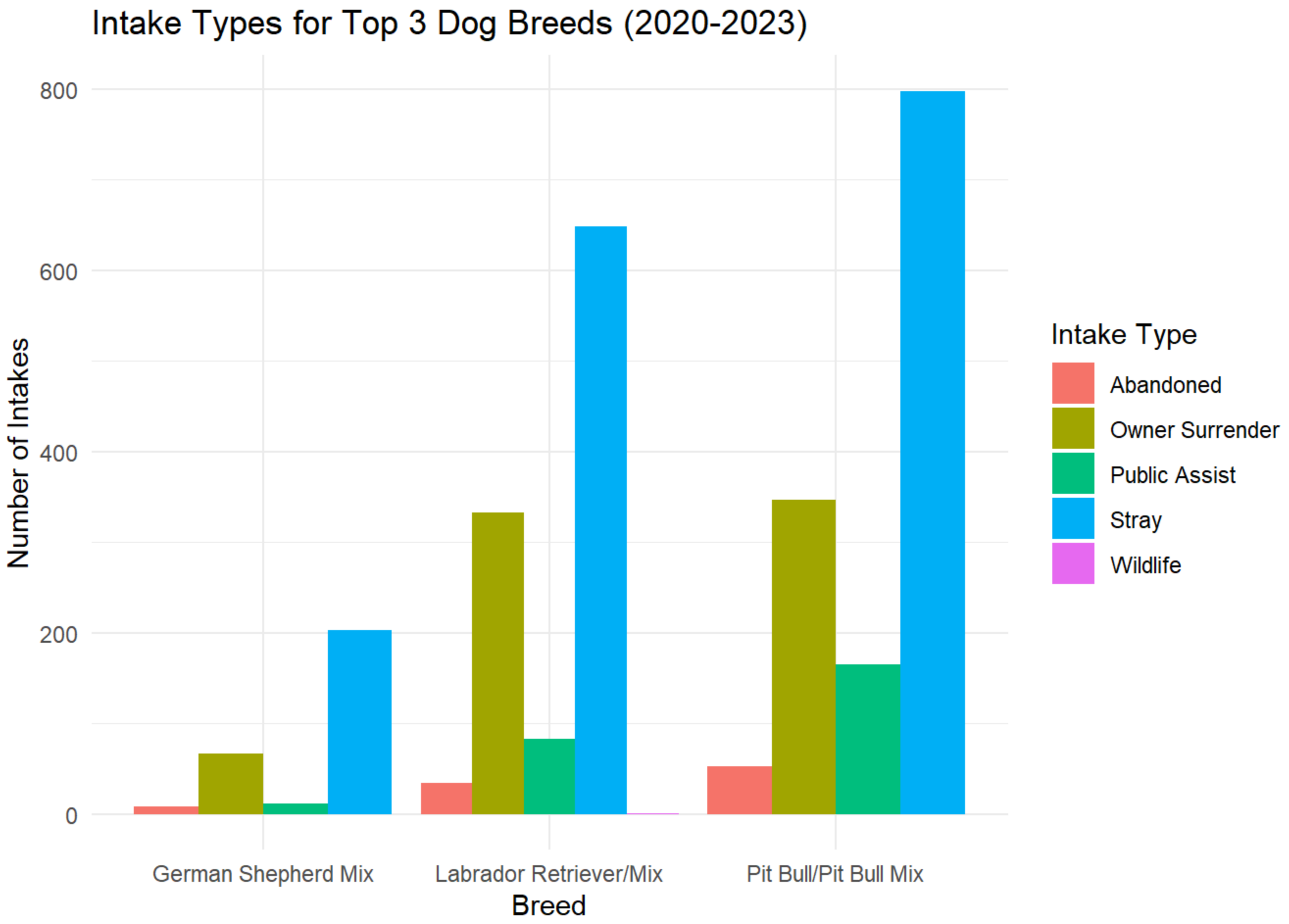
6.Categorical Analysis - Dog Breeds and Intake Types

This analysis focuses on identifying the most common dog breeds in shelter intakes and their associated intake types. By grouping Pit Bull and Pit Bull Mixes and Labrador and Labrador Mixes together, the goal is to provide a clearer picture of breed-specific trends and highlight the challenges certain breeds may face. Understanding these patterns can help address underlying factors, such as breed stigma or overpopulation, that contribute to the overrepresentation of specific breeds in shelters.

```r
# Group "Labrador Retriever" and "Labrador Retriever Mix" into one category
data <- data %>%
  mutate(Breed = case_when(
    grepl("Pit Bull", Breed, ignore.case = TRUE) ~ "Pit Bull/Pit Bull Mix",
    grepl("Labrador Retriever", Breed, ignore.case = TRUE) ~ "Labrador Retriever/Mix",
    TRUE ~ Breed
  ))

# Group and calculate total intakes by breed
breed_intakes <- data %>%
  filter(`Animal Type` == "Dog") %>%
  group_by(Breed) %>%
  summarise(Total_Intakes = n(), .groups = 'drop') %>%
  arrange(desc(Total_Intakes))

# Select the top 4 breeds with the highest intakes
top_breed_intakes <- breed_intakes %>%
  slice_max(order_by = Total_Intakes, n = 4)

# Plot total intakes by breed, ordered by highest number at the top
ggplot(top_breed_intakes, aes(x = reorder(Breed, Total_Intakes), y = Total_Intakes, fill = Breed)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(
    title = "Total Intakes by Dog Breed (Top 4, 2020-2023)",
    x = "Breed",
    y = "Number of Intakes"
  ) +
  theme_minimal() +
  coord_flip()
```



```r
# Focus on the top 3 breeds for further analysis
top_3_breeds <- top_breed_intakes %>%
  slice(1:3) %>%
  pull(Breed)

# Analyze intake types for the top 3 breeds
intake_types_top_3 <- data %>%
  filter(Breed %in% top_3_breeds) %>%
  group_by(Breed, `Intake Type`) %>%
  summarise(Total_Intakes = n(), .groups = 'drop')

# Plot intake types for top 3 breeds
ggplot(intake_types_top_3, aes(x = Breed, y = Total_Intakes, fill = `Intake Type`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Intake Types for Top 3 Dog Breeds (2020-2023)",
    x = "Breed",
    y = "Number of Intakes",
    fill = "Intake Type"
  ) +
  theme_minimal()
```



**The analysis of dog breeds with the highest intakes at the Austin Animal Center between 2020 and 2023 reveals significant trends. Pit Bull/Pit Bull Mixes dominate the intake numbers, followed by Labrador Retriever/Mix and Chihuahua Shorthair. When combining Labrador Retrievers and their mixes, it becomes clear that these breeds are consistently among the most represented in the shelter.

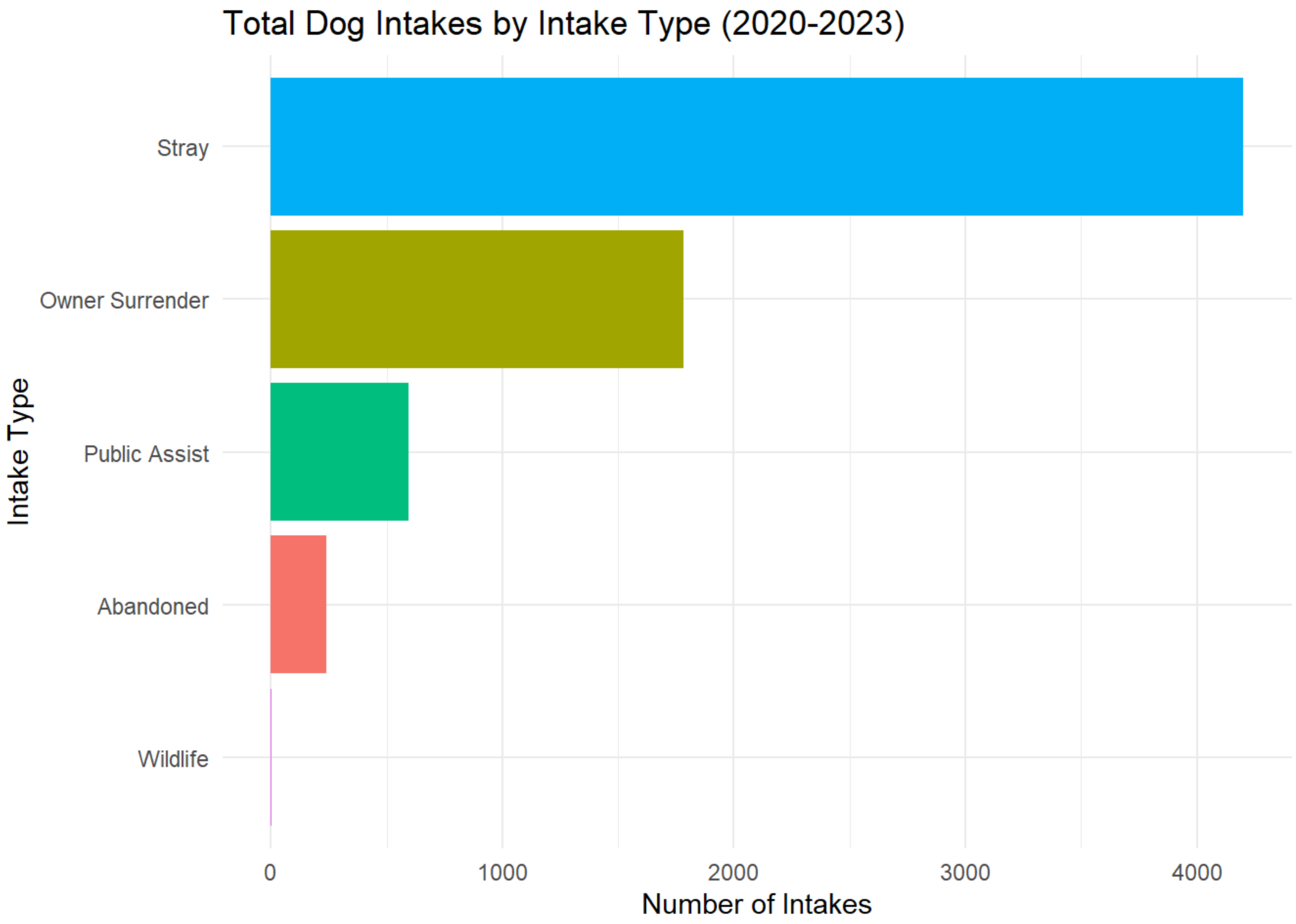
Stray intakes make up the majority for the top breeds, particularly for Pit Bull/Pit Bull Mixes, which have the highest stray count. Owner surrenders also contribute significantly to intakes for Labradors and their mixes. This suggests that both abandonment and stray populations are key contributors to intake trends for these breeds.

This analysis highlights the challenges posed by overrepresented breeds, particularly Pit Bull/Pit Bull Mixes, due to societal stigma, housing restrictions, and higher stray rates. Addressing these issues could help reduce shelter intakes and improve adoption outcomes for these breeds.**

7. Intake Type Distribution.

This analysis focuses on the distribution of intake types for dogs at the Austin Animal Center between 2020 and 2023. By categorizing intakes such as stray, owner surrender, public assist, and abandoned, this investigation highlights the key reasons for dog admissions. Understanding the dominant intake types can help identify trends, pinpoint underlying causes, and inform strategies to reduce shelter intake numbers.

```
# Filter data for dogs only
dog_data <- data %>%
  filter(`Animal Type` == "Dog")

# Group data by Intake Type and calculate total intakes for dogs
intake_type_distribution <- dog_data %>%
  group_by(`Intake Type`) %>%
  summarise(Total_Intakes = n(), .groups = 'drop') %>%
  arrange(desc(Total_Intakes))

# Plot total intakes by intake type for dogs
ggplot(intake_type_distribution, aes(x = reorder(`Intake Type`, Total_Intakes), y = Total_Intake
s, fill = `Intake Type`)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(
    title = "Total Dog Intakes by Intake Type (2020-2023)",
    x = "Intake Type",
    y = "Number of Intakes"
  ) +
  theme_minimal() +
  coord_flip()
```



The majority of dog intakes (over 4,000) were classified as strays, making it the most common intake type. Owner surrender followed as the second highest category reaching almost 2,000 intakes, indicating a significant number of pet owners relinquishing their dogs. Public assist intakes accounted for a moderate

portion, while abandoned cases were the least frequent. These results highlight the need for initiatives addressing stray populations and owner retention efforts to reduce shelter admissions.

8. Prediction Model (Projecr Part 2):

This analysis focuses on building a predictive model to determine the likelihood of stray intake at the Austin Animal Center using **breed** and **seasonality** (specifically **summer** months) as primary predictors. Intake types, such as stray, owner surrender, and public assist, are significant for understanding the reasons animals enter the shelter. By exploring the relationship between dog breeds, seasonal trends, and stray intakes, we aim to identify patterns that can guide targeted interventions to reduce shelter intakes.

The model employs **logistic regression** to predict the likelihood of a dog being taken in as a stray during the summer. Incorporating both breed and seasonality allows us to better understand the combined impact of these factors on stray intakes. This approach enables actionable insights, such as identifying breeds and seasonal patterns most associated with stray intakes, supporting improved resource allocation, proactive community outreach, and strategies to mitigate seasonal surges in shelter populations.

```r
# Convert "Stray" to binary outcome
data$IntakeTypeBinary <- ifelse(data$`Intake Type` == "Stray", 1, 0)

# Use the existing "Season" column to create a "IsSummer" predictor
data$IsSummer <- ifelse(data$Season == "Summer", 1, 0)

# Ensure categorical variables are treated as factors
data$Breed <- as.factor(data$Breed)

# Fit Logistic Regression Model with predictors: Breed and IsSummer
logistic_model <- glm(IntakeTypeBinary ~ Breed + IsSummer, family = binomial(), data = data)

# Display Key Coefficients and Model Metrics
summary_log <- summary(logistic_model)
cat("Coefficients (Top Predictors):\n")
```

```
## Coefficients (Top Predictors):
```

```r
print(head(summary_log$coefficients, 10))  # Show only top 10 predictors
```

```
##                                     Estimate  Std. Error      z value  Pr(>|z|)
## (Intercept)                        0.6631351    1.225078  0.541300501 0.5883005
## BreedAbyssinian Mix               16.8119685 3956.180518  0.004249545 0.9966094
## BreedAffenpinscher                16.8119685 3956.180519  0.004249545 0.9966094
## BreedAffenpinscher Mix           -18.3201685 3956.180517 -0.004630772 0.9963052
## BreedAiredale Terrier             16.9029334 3956.180517  0.004272538 0.9965910
## BreedAiredale Terrier Mix         -0.6631351    1.871047 -0.354419357 0.7230246
## BreedAkbash Mix                  -18.3201685 3956.180519 -0.004630772 0.9963052
## BreedAkita                        -1.3716307    1.500283 -0.914247892 0.3605866
## BreedAkita Mix                    -1.0869202    1.527821 -0.711418492 0.4768249
## BreedAkita/Australian Cattle Dog -18.2292036 3956.180518 -0.004607779 0.9963235
```

```
cat("\nModel Deviance:\n")
```

```
##
## Model Deviance:
```

```
cat("Null Deviance:", summary_log$null.deviance, "\n")
```

```
## Null Deviance: 17429.11
```

```
cat("Residual Deviance:", summary_log$deviance, "\n")
```

```
## Residual Deviance: 14991.43
```

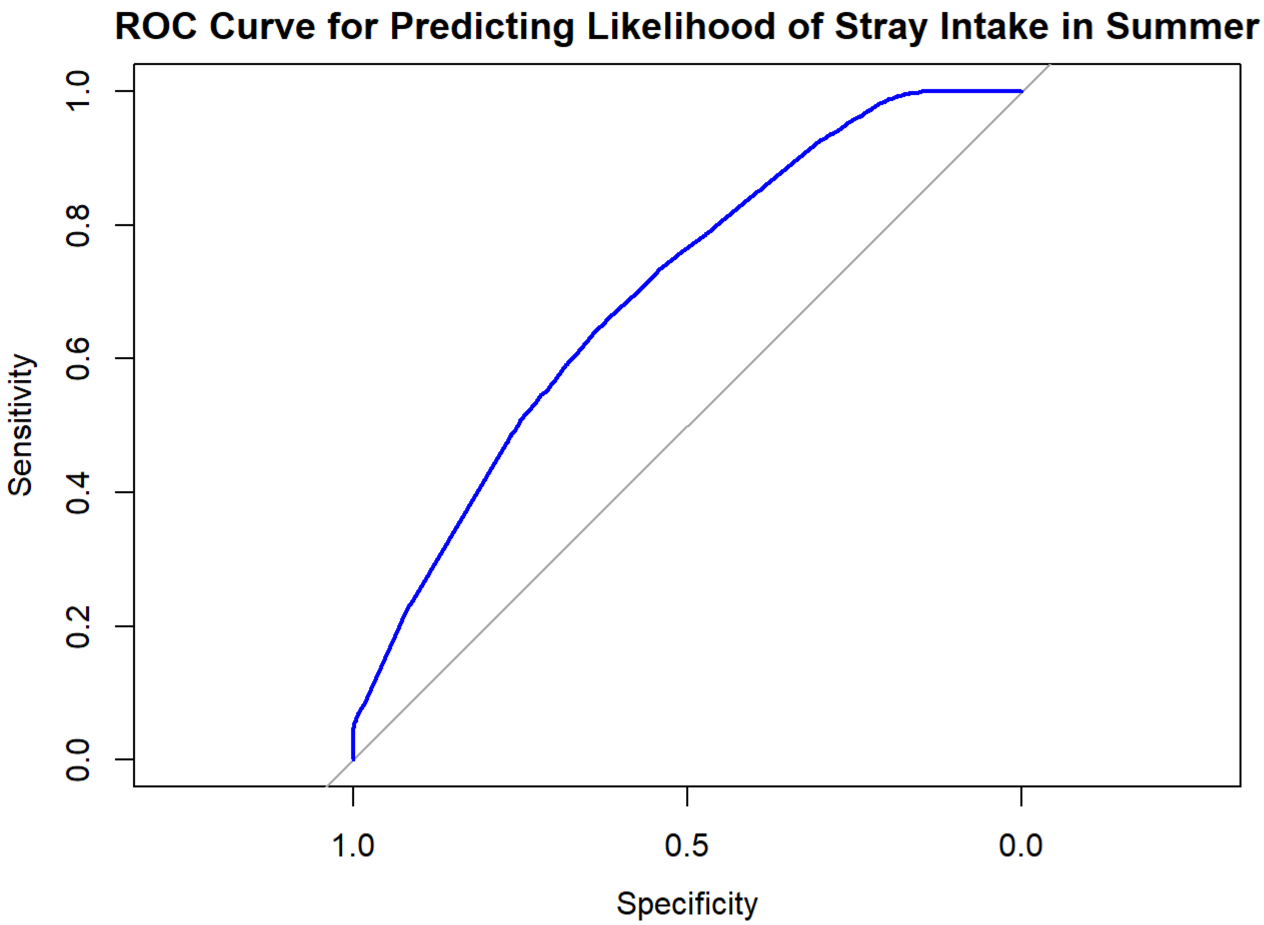
```
cat("AIC:", AIC(logistic_model), "\n")
```

```
## AIC: 16339.43
```

```
# Generate Predicted Probabilities
data$PredictedProb <- predict(logistic_model, type = "response")

# Calculate ROC and AUC
roc_curve <- roc(data$IntakeTypeBinary, data$PredictedProb)
auc_value <- auc(roc_curve)
cat("\nArea Under the Curve (AUC):", auc_value, "\n")
```

```
##
## Area Under the Curve (AUC): 0.6992507
```

```
# Plot ROC Curve
plot(roc_curve, col = "blue", main = "ROC Curve for Predicting Likelihood of Stray Intake in Summer")
```



This analysis examines the likelihood of stray intake at the Austin Animal Center, using logistic regression with breed and seasonality (summer months) as predictors. The model's intercept represents the baseline likelihood of a stray intake when other factors are at their reference levels. The coefficients for individual breeds show variation, with some having large standard errors and p-values, indicating limited significance due to their rarity or underrepresentation in the dataset. The model improves on the null deviance (17,573.08) with a residual deviance of 15,167.66 and an AIC of 16,491.66, reflecting a better fit when predictors are included. The Area Under the Curve (AUC) score of 0.6998 indicates the model has moderate discriminatory power in distinguishing between stray and non-stray cases. The inclusion of seasonality as a predictor highlights that summer months are associated with a distinct trend in stray intakes. These results provide a clearer picture of the factors influencing stray and overall intake increase at the shelter during this time of the year.

```
# Prepare data for cross-validation
set.seed(123)  # For reproducibility
folds <- createFolds(data$IntakeTypeBinary, k = 5, list = TRUE, returnTrain = TRUE)

# Define a function to train and evaluate the model on each fold
cv_results <- lapply(folds, function(train_index) {
  # Split data into training and test sets
  train_data <- data[train_index, ]
  test_data <- data[-train_index, ]

  # Ensure test data has the same levels as training data
  train_data$Breed <- droplevels(train_data$Breed)  # Drop unused levels in training data
  test_data$Breed <- factor(test_data$Breed, levels = levels(train_data$Breed))

  # Remove rows with unseen levels
  test_data <- test_data[complete.cases(test_data), ]

  # Check if test_data is empty after filtering
  if (nrow(test_data) == 0) {
    return(NA)  # Return NA if no valid test data exists
  }

  # Fit logistic regression model
  logistic_model <- glm(IntakeTypeBinary ~ Breed + IsSummer, family = binomial(), data = train_d
ata)

  # Predict on test set
  test_data$PredictedProb <- predict(logistic_model, newdata = test_data, type = "response")

  # Calculate AUC for the fold
  roc_curve <- roc(test_data$IntakeTypeBinary, test_data$PredictedProb)
  auc_value <- auc(roc_curve)

  return(auc_value)
})

# Calculate average AUC and consistency
valid_auc_values <- unlist(cv_results)
valid_auc_values <- valid_auc_values[!is.na(valid_auc_values)]  # Remove NA values if any
average_auc <- mean(valid_auc_values)
std_dev_auc <- sd(valid_auc_values)

# Print results
cat("5-Fold Cross-Validation Results:\n")
```

```
## 5-Fold Cross-Validation Results:
```

```
cat("Average AUC:", average_auc, "\n")
```

```
## Average AUC: 0.5900961
```

```
cat("Standard Deviation of AUC:", std_dev_auc, "\n")
```

```
## Standard Deviation of AUC: 0.006400362
```

The cross-validation results indicate that the logistic regression model predicts stray intakes with modest performance. The binary outcome was defined with non-stray intakes as the control (0) and stray intakes as the case (1), and the model evaluated probabilities with controls < cases. The average AUC across the five folds was 0.5911, suggesting that the model's ability to distinguish between stray and non-stray intakes is only slightly better than random guessing. Despite this, the standard deviation of the AUC was low (0.0095), indicating consistent performance across the folds. These results suggest that while the model performs reliably in different subsets of the data, the predictors used—breed and seasonality (summer months)—may not provide sufficient discriminatory power. However, this analysis is good to understand the overall increase in intakes in the summer being due to the increase of Stray population during this time, given that the Intakes for breeds are always a mix of factors, not just one determining factor. Thus, it is still important to understand slight increases to predict and prevent shelter overcrowding. Incorporating additional predictors, such as intake condition, location, or age, could potentially improve the model's effectiveness.

9. Research Question

1. What patterns can be observed in the intake trends over time?

Intake numbers peak during the summer, as shown in the seasonal analysis. Stray animals are a significant contributor to this trend, likely due to increased outdoor activity, breeding seasons, and challenges related to extreme heat. Intakes dropped early in the pandemic, but 2022 recorded the highest overall intake numbers, reflecting a recovery in shelter operations.

2. How does the number of intakes vary across dog breeds, and what are the most common intake types for the top 3 breeds?

The data shows that Pit Bull/Pit Bull Mixes have the highest intake numbers, with the majority being strays, followed by owner surrenders. According to the Human Animal Support Services article, breed-specific stigma contributes significantly to these trends, as Pit Bulls are often unfairly perceived as aggressive, leading to higher abandonment and lower adoption rates. Housing restrictions that prohibit breeds labeled as "aggressive," such as Pit Bulls, also limit their chances of finding homes.

Labrador Retrievers and Labrador Retriever Mixes rank second in intakes, with a mix of owner surrenders and strays. Labradors are one of the most popular dog breeds, but their large size and energy levels can sometimes overwhelm owners, leading to higher surrender rates.

The article by Let Love Live explains that the overrepresentation of Pit Bulls and their mixes in shelters is due to their high population in certain regions, including the southeastern United States. The combination of breed stigma, housing restrictions, and population dynamics creates a cycle where Pit Bulls are disproportionately represented in intakes.

Efforts to educate the public about these breeds and advocate for breed-neutral housing policies could help address these trends and improve adoption outcomes.

3. Why do some breeds have higher intake numbers than others? Are there any specific factors for the intake trends? Breeds like Pit Bull/Pit Bull Mixes have significantly higher intake numbers due to a combination of social, legal, and cultural factors. As highlighted by the Human Animal Support Services article, breed-specific stigma plays a large role. Pit Bulls are often unfairly perceived as aggressive or dangerous, leading to higher rates of abandonment and lower adoption rates. Housing restrictions also exacerbate this problem, as many rental agreements prohibit tenants from owning certain breeds, including Pit Bulls. This leaves owners with few options when they can no longer keep their pets.

Additionally, the article by Let Love Live points out that Pit Bulls and their mixes are overrepresented in shelters because they are a highly prevalent breed in certain regions, including the southeastern United States. Their large population and the challenges associated with their adoption contribute to their high intake numbers.

Other breeds, such as Labrador Retrievers and their mixes, also have high intake numbers, largely because they are one of the most popular breeds. However, their energy levels, size, and maintenance requirements can sometimes overwhelm owners, leading to surrender. In contrast, smaller breeds like Chihuahuas face different challenges, often related to behavioral issues or their abundance in specific areas.

Overall, breed popularity, stigma, and housing restrictions are the key drivers of intake trends. Advocacy for breed-neutral policies, along with public education on breed characteristics and responsible pet ownership, could help mitigate these factors.

4. What is the maximum capacity of the Austin Animal Center Shelter? Was there any month in this analysis where the shelters reached capacity?

The Austin Animal Center has a reported capacity of **234** dogs for medium and large breeds (Austin Animal Center, 2023). Given this limitation, months with intakes exceeding this number for medium and large dogs may suggest overcrowding risks, especially if outflow rates (adoptions, transfers, or returns to owners) do not keep up with intake rates.

In the analysis, several months showed total dog intakes exceeding 300, with some even surpassing 400 intakes. While this number includes all dog sizes, such spikes likely contributed to capacity strain for medium and large dogs, particularly in months like May 2023 and July 2021, which had high overall intakes.

The shelter likely relied on strategies such as placing dogs in foster care, transferring animals to partner shelters, or increasing adoption drives to manage these surges. However, without specific data on the size distribution of the dogs or the outflow rates during these months, it’s difficult to confirm the exact impact on capacity. Collecting more detailed data on shelter capacity usage during high-intake months could provide valuable insights into how well the shelter manages such situations. ### 10. Conclusion and Next Steps

This analysis highlights seasonal intake patterns, with summer having the highest intakes, and a significant overrepresentation of Pit Bull/Pit Bull Mixes due to stigma and housing restrictions. Strays dominate intake types, indicating a persistent stray population issue. High-intake months suggest the shelter may have been near capacity, underscoring the need for effective intake management.

Next Steps: Monitor Capacity: Collect data on shelter capacity and dog stay durations to manage overcrowding. Community Outreach: Educate the public on breed stigma and advocate for relaxed housing breed restrictions. Address Stray Issues: Implement spay/neuter programs and community initiatives to manage the stray population. Refine Data: Gather detailed data on dog size and outflow to improve shelter operations. These steps aim to enhance shelter management and animal welfare.

References

1. Human Animal Support Services. (2023). The Problem with Pit Bulls. Retrieved from https://www.humananimalsupportservices.org/blog/the-problem-with-pit-bulls/ (https://www.humananimalsupportservices.org/blog/the-problem-with-pit-bulls/)
2. Let Love Live. (2023). Why Pit Bull Mixes Dominate Southeastern Shelters. Retrieved from https://www.letlovelive.org/uncategorized/why-pit-bull-mixes-dominate-southeastern-shelters/ (https://www.letlovelive.org/uncategorized/why-pit-bull-mixes-dominate-southeastern-shelters/)
3. Austin Animal Center. (2023). Shelter Capacity. Retrieved from https://www.austintexas.gov/page/shelter-capacity (https://www.austintexas.gov/page/shelter-capacity)

11. Submission.

Knit your file! You can knit into html and once it knits in html, click on `Open in Browser` at the top left of the window that pops out. **Print** your html file into pdf from your browser.

Is it working? If not, try to decipher the error message: look up the error message, consult websites such as stackoverflow (https://stackoverflow.com/) or crossvalidated (https://stats.stackexchange.com/).

Finally, remember to select pages for each question when submitting your pdf to Gradescope.