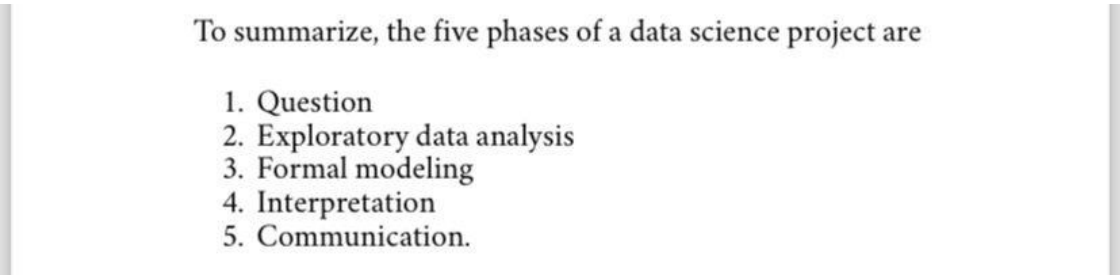
Data Science Project

Ceren Yıldız

2024-12-24



> To summarize, the five phases of a data science project are
>
> 1. Question
> 2. Exploratory data analysis
> 3. Formal modeling
> 4. Interpretation
> 5. Communication.

I would like to start by explaining two stages.

1.Question

In this step, we need to determine the questions we will analyze for later use.Our basic questions are:

- **Which animal species are most frequently found in the shelter?**
- **Are age and species effective in leaving the shelter?**
- **What are the types and reasons why animals leave the shelter? And in what age range are these reasons more common? What does it depend on?**

2.Exploratory Data Analysis(EDA)

In order to answer the questions we ask, we need to go through the EDA process. In this section, analysis and visualizations are made. First of all, it is necessary to do some coding in R to do exploratory data analysis. I am sorry about this image because it does not look very smooth in the file. I am not doing my analysis in detail at this stage. Despite this, I have listed my codes below in order to make them look neater.

```
options(repos = c(CRAN = "https://cran.rstudio.com/"))
install.packages("tinytex")

## Installing package into 'C:/Users/pc/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## also installing the dependency 'xfun'

## package 'xfun' successfully unpacked and MD5 sums checked
## Warning: cannot remove prior installation of package 'xfun'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\pc\AppData\Local\R\win-library\4.3\00LOCK\xfun\libs\x64\xfun.dll
to
## C:\Users\pc\AppData\Local\R\win-library\4.3\xfun\libs\x64\xfun.dll: Permis
sion
## denied
```

```
## Warning: restored 'xfun'
```

```
## package 'tinytex' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\pc\AppData\Local\Temp\RtmpUHWjj5\downloaded_packages
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(dplyr)
install.packages("tidyr")
```

```
## Installing package into 'C:/Users/pc/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'tidyr' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'tidyr'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\pc\AppData\Local\R\win-library\4.3\00LOCK\tidyr\libs\x64\tidyr.dl
l to
## C:\Users\pc\AppData\Local\R\win-library\4.3\tidyr\libs\x64\tidyr.dll:
## Permission denied
```

```
## Warning: restored 'tidyr'
```

```
##
## The downloaded binary packages are in
##  C:\Users\pc\AppData\Local\Temp\RtmpUHWjj5\downloaded_packages

library(tidyr)

## Warning: package 'tidyr' was built under R version 4.3.3

# Data Loading Phase
Processed_Austin_Animal_Center_Intakes <- read.csv("C:/Users/pc/Desktop/Proce
ssed_Austin_Animal_Center_Intakes.csv")

 head(Processed_Austin_Animal_Center_Intakes)

##   Animal.ID    Name            DateTime           MonthYear
## 1   A786884  *Brock 2019-01-03 16:19:00 2019-01-03 16:19:00
## 2   A706918   Belle 2015-07-05 12:59:00 2015-07-05 12:59:00
## 3   A724273 Runster 2016-04-14 18:43:00 2016-04-14 18:43:00
## 4   A665644 Unknown 2013-10-21 07:59:00 2013-10-21 07:59:00
## 5   A682524     Rio 2014-06-29 10:38:00 2014-06-29 10:38:00
## 6   A743852    Odin 2017-02-18 12:46:00 2017-02-18 12:46:00
##                         Found.Location     Intake.Type Intake.Condition
## 1 2501 Magin Meadow Dr in Austin (TX)           Stray           Normal
## 2    9409 Bluegrass Dr in Austin (TX)           Stray           Normal
## 3  2818 Palomino Trail in Austin (TX)           Stray           Normal
## 4                         Austin (TX)           Stray             Sick
## 5       800 Grove Blvd in Austin (TX)           Stray           Normal
## 6                         Austin (TX) Owner Surrender           Normal
##   Animal.Type Sex.upon.Intake Age.upon.Intake
## 1         Dog   Neutered Male         2 years
## 2         Dog   Spayed Female         8 years
## 3         Dog     Intact Male       11 months
## 4         Cat   Intact Female         4 weeks
## 5         Dog   Neutered Male         4 years
## 6         Dog   Neutered Male         2 years
##                                    Breed       Color Age.in.Days
## 1                             Beagle Mix    Tricolor         730
## 2               English Springer Spaniel White/Liver        2920
## 3                            Basenji Mix Sable/White         330
## 4                 Domestic Shorthair Mix      Calico          28
## 5 Doberman Pinsch/Australian Cattle Dog    Tan/Gray        1460
## 6                 Labrador Retriever Mix   Chocolate         730

 dim(Processed_Austin_Animal_Center_Intakes)

## [1] 124120     13

 colnames(Processed_Austin_Animal_Center_Intakes)

##  [1] "Animal.ID"        "Name"             "DateTime"          "MonthYear"
##  [5] "Found.Location"   "Intake.Type"      "Intake.Condition" "Animal.Type
"
```

```
##  [9] "Sex.upon.Intake"  "Age.upon.Intake"  "Breed"            "Color"
## [13] "Age.in.Days"
```

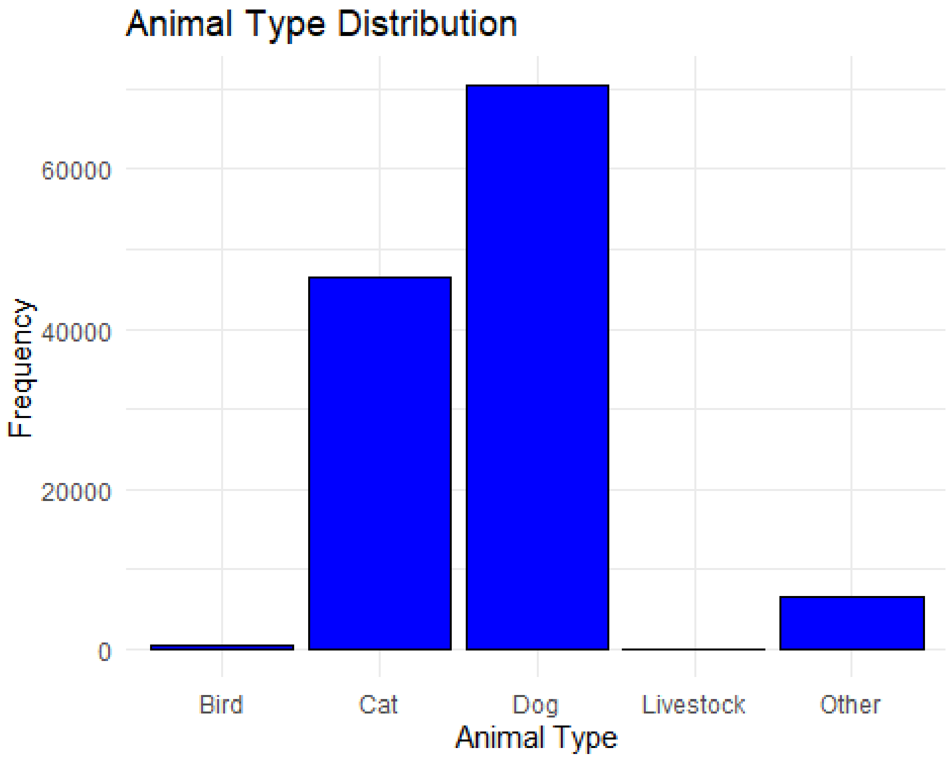
```
summary(Processed_Austin_Animal_Center_Intakes)
```

```
##   Animal.ID             Name             DateTime          MonthYear
##  Length:124120      Length:124120      Length:124120      Length:124120
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Found.Location     Intake.Type        Intake.Condition   Animal.Type
##  Length:124120      Length:124120      Length:124120      Length:124120
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Sex.upon.Intake    Age.upon.Intake       Breed              Color
##  Length:124120      Length:124120      Length:124120      Length:124120
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Age.in.Days
##  Min.   :-1095.0
##  1st Qu.:   60.0
##  Median :  365.0
##  Mean   :  751.9
##  3rd Qu.:  730.0
##  Max.   : 9125.0
```

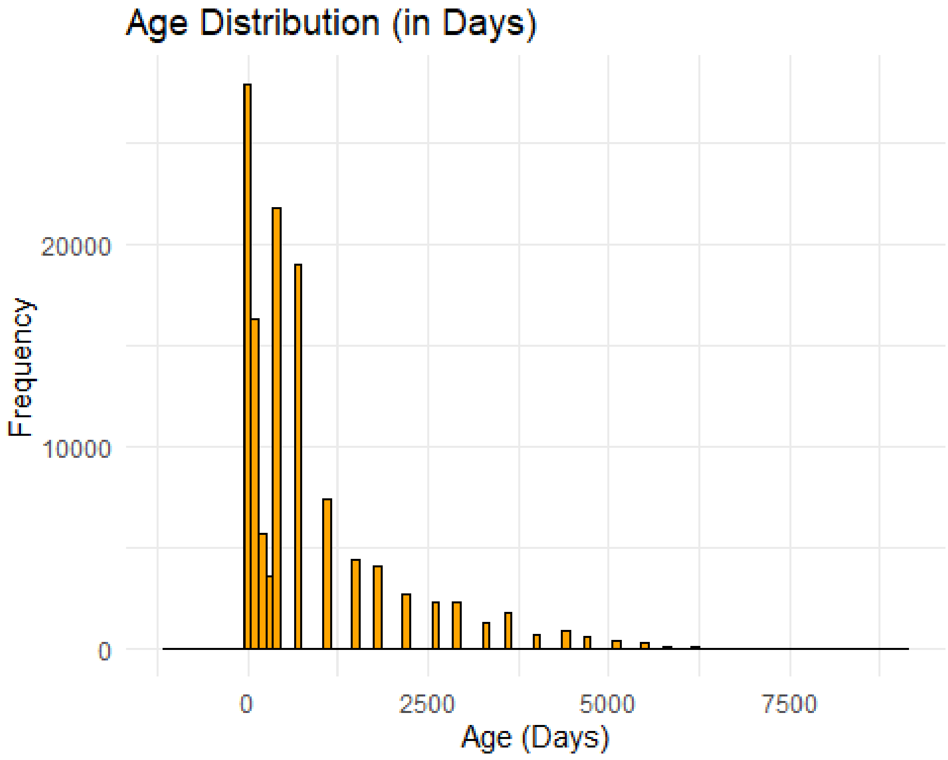
```
str(Processed_Austin_Animal_Center_Intakes)
```

```
## 'data.frame':    124120 obs. of  13 variables:
##  $ Animal.ID       : chr  "A786884" "A706918" "A724273" "A665644" ...
##  $ Name            : chr  "*Brock" "Belle" "Runster" "Unknown" ...
##  $ DateTime        : chr  "2019-01-03 16:19:00" "2015-07-05 12:59:00" "201
6-04-14 18:43:00" "2013-10-21 07:59:00" ...
##  $ MonthYear       : chr  "2019-01-03 16:19:00" "2015-07-05 12:59:00" "201
6-04-14 18:43:00" "2013-10-21 07:59:00" ...
##  $ Found.Location  : chr  "2501 Magin Meadow Dr in Austin (TX)" "9409 Blue
grass Dr in Austin (TX)" "2818 Palomino Trail in Austin (TX)" "Austin (TX)" .
..
##  $ Intake.Type     : chr  "Stray" "Stray" "Stray" "Stray" ...
##  $ Intake.Condition: chr  "Normal" "Normal" "Normal" "Sick" ...
##  $ Animal.Type     : chr  "Dog" "Dog" "Dog" "Cat" ...
##  $ Sex.upon.Intake : chr  "Neutered Male" "Spayed Female" "Intact Male" "I
ntact Female" ...
```

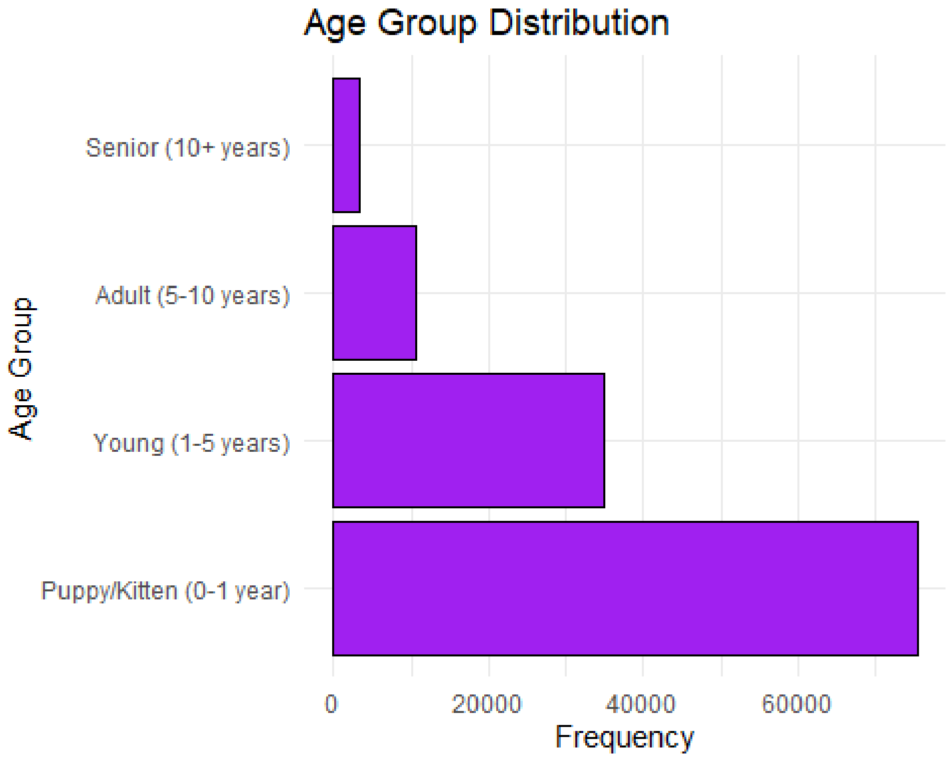
```
##  $ Age.upon.Intake : chr  "2 years" "8 years" "11 months" "4 weeks" ...
##  $ Breed           : chr  "Beagle Mix" "English Springer Spaniel" "Basenji
Mix" "Domestic Shorthair Mix" ...
##  $ Color           : chr  "Tricolor" "White/Liver" "Sable/White" "Calico"
...
##  $ Age.in.Days     : int  730 2920 330 28 1460 730 2190 730 28 28 ...

# Animal Type Distribution
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Animal.Type)) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Animal Type Distribution", x = "Animal Type", y = "Frequency"
) +
  theme_minimal()
```



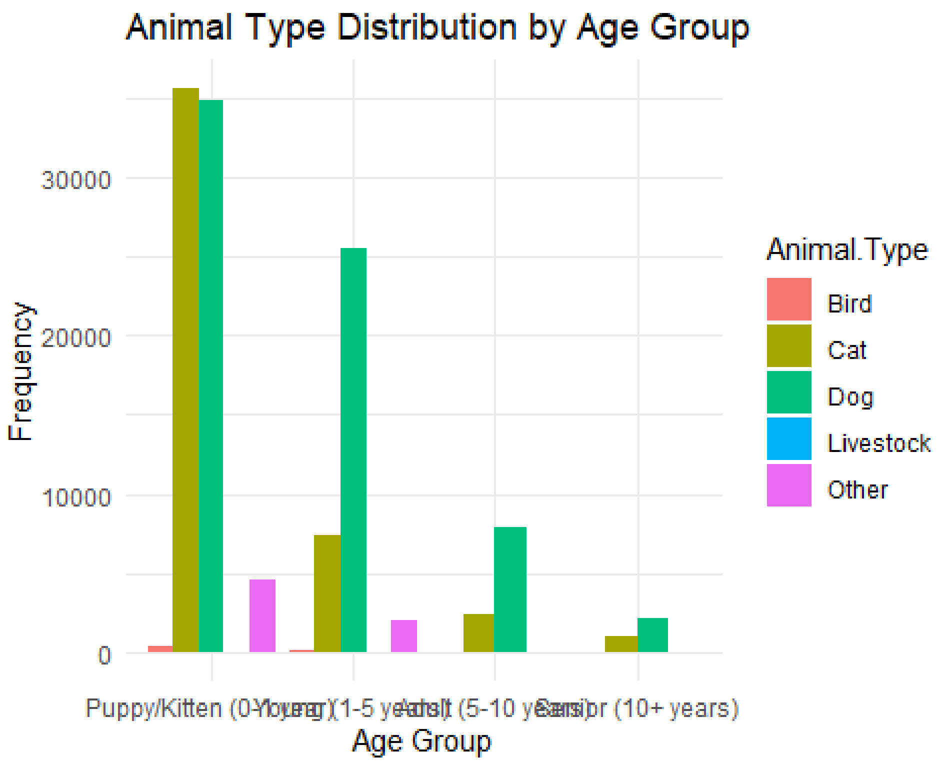
```
# Age Distribution in Days
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Age.in.Days)) +
  geom_histogram(binwidth = 100, fill = "orange", color = "black") +
  labs(title = "Age Distribution (in Days)", x = "Age (Days)", y = "Frequency
") +
  theme_minimal()
```



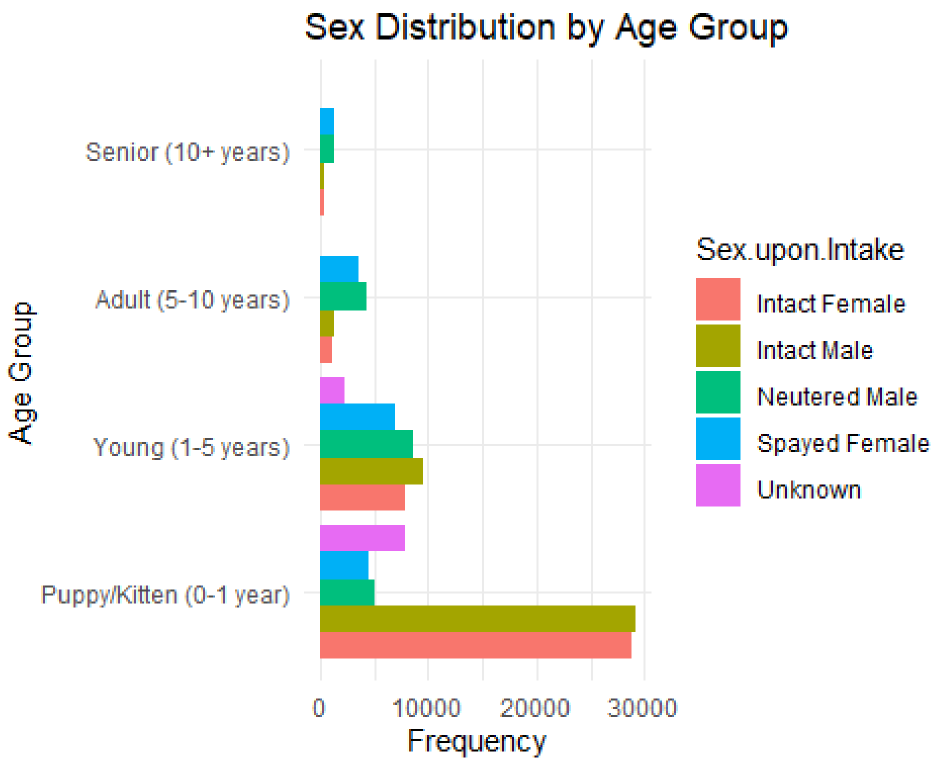
```r
# For create age groups and dataframe
Processed_Austin_Animal_Center_Intakes$Age.Group <- cut(
  Processed_Austin_Animal_Center_Intakes$Age.in.Days,
  breaks = c(-Inf, 365, 1825, 3650, Inf),
  labels = c("Puppy/Kitten (0-1 year)", "Young (1-5 years)", "Adult (5-10 yea
rs)", "Senior (10+ years)")
)


ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Age.Group)) +
  geom_bar(fill = "purple", color = "black") +
  labs(title = "Age Group Distribution", x = "Age Group", y = "Frequency") +
  theme_minimal() +
  coord_flip() # Flip for better readability
```



```r
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Age.Group, fill = Anim
al.Type)) +
  geom_bar(position = "dodge") +
  labs(title = "Animal Type Distribution by Age Group", x = "Age Group", y =
"Frequency") +
  theme_minimal()
```

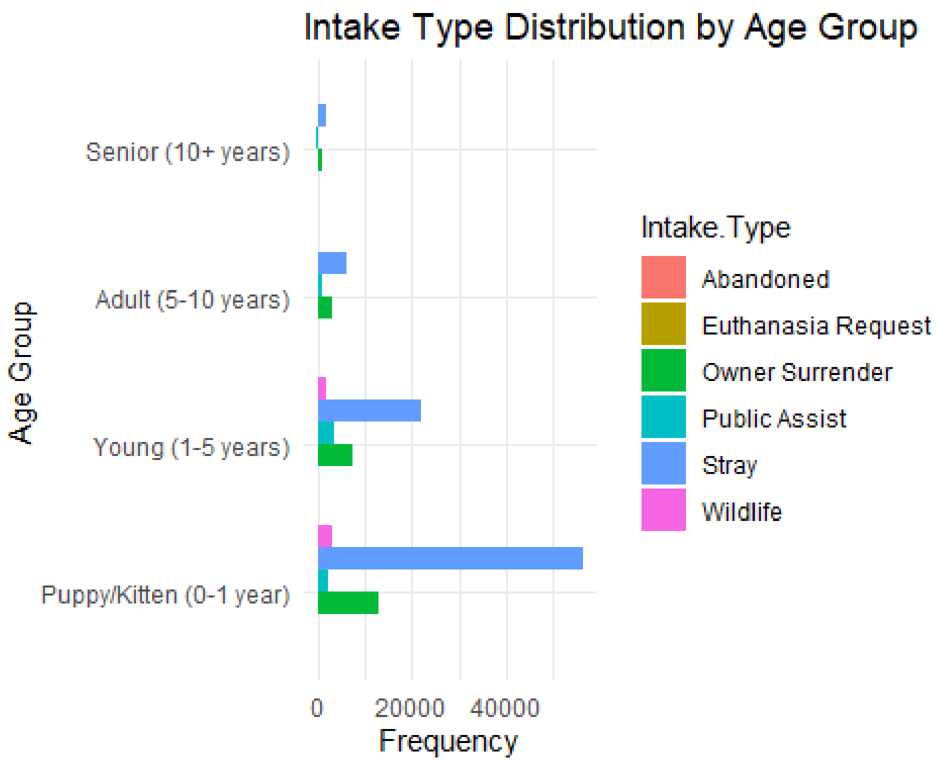


```
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Age.Group, fill = Sex.
upon.Intake)) +
  geom_bar(position = "dodge") +
  labs(title = "Sex Distribution by Age Group", x = "Age Group", y = "Frequen
cy") +
  theme_minimal() +
  coord_flip() # Better readability
```

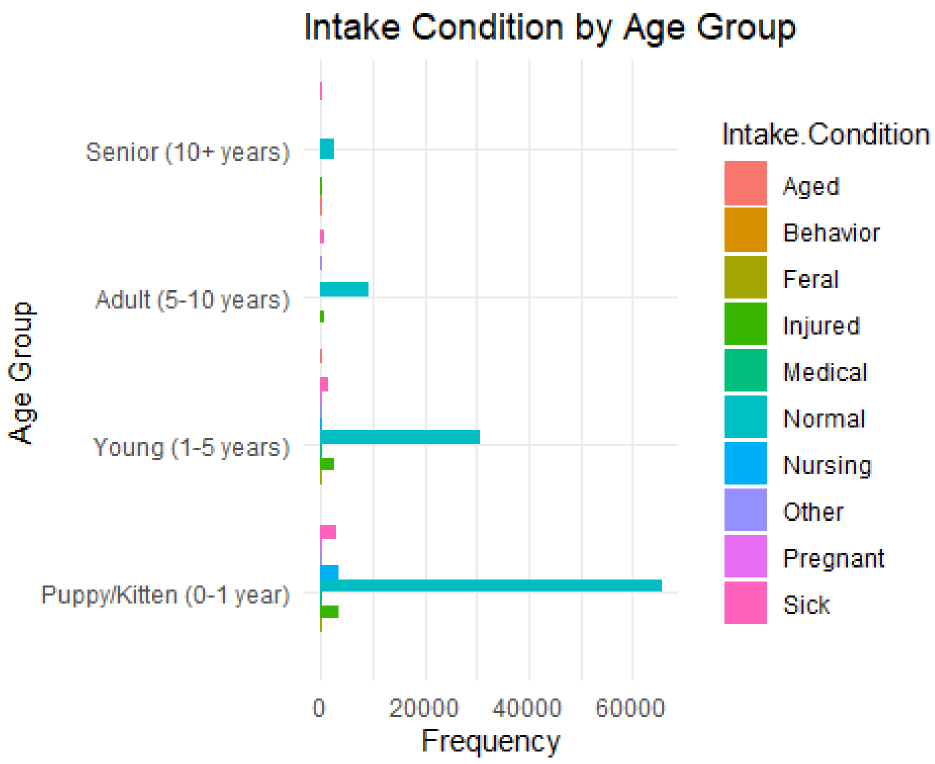


```r
Processed_Austin_Animal_Center_Intakes$DateTime <- as.Date(Processed_Austin_Animal_Center_Intakes$DateTime)
```

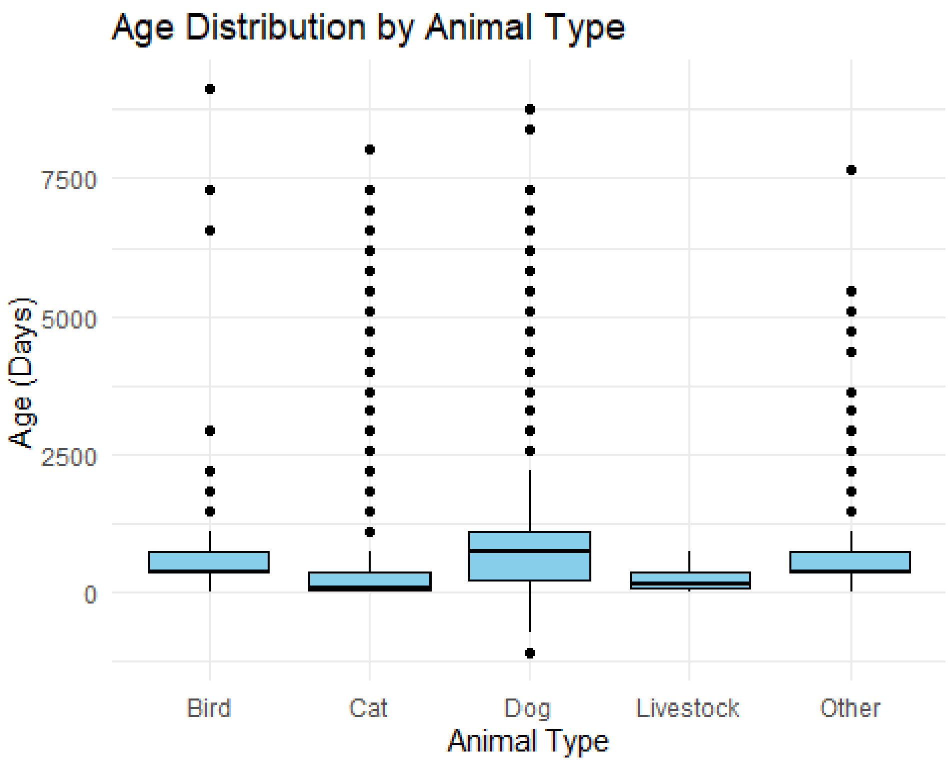
```r
# Age Group vs Intake Type Distribution
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Age.Group, fill = Intake.Type)) +
  geom_bar(position = "dodge") +
  labs(title = "Intake Type Distribution by Age Group", x = "Age Group", y = "Frequency") +
  theme_minimal() +
  coord_flip() # Better readability
```



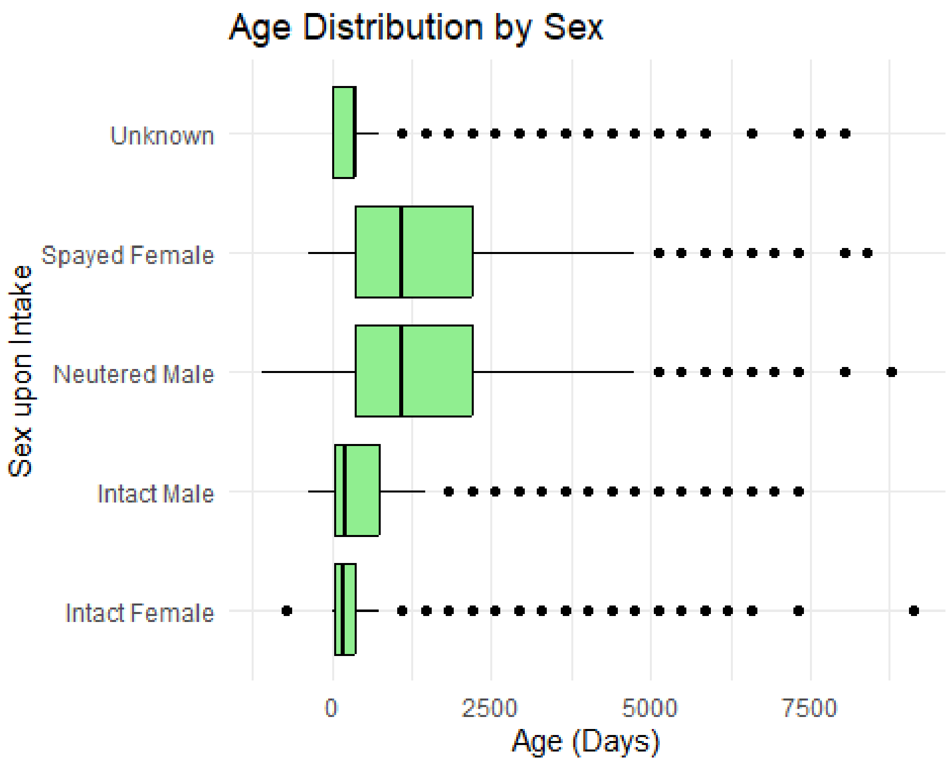
```
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Age.Group, fill = Inta
ke.Condition)) +
  geom_bar(position = "dodge") +
  labs(title = "Intake Condition by Age Group", x = "Age Group", y = "Frequen
cy") +
  theme_minimal() +
  coord_flip() # Better readability
```



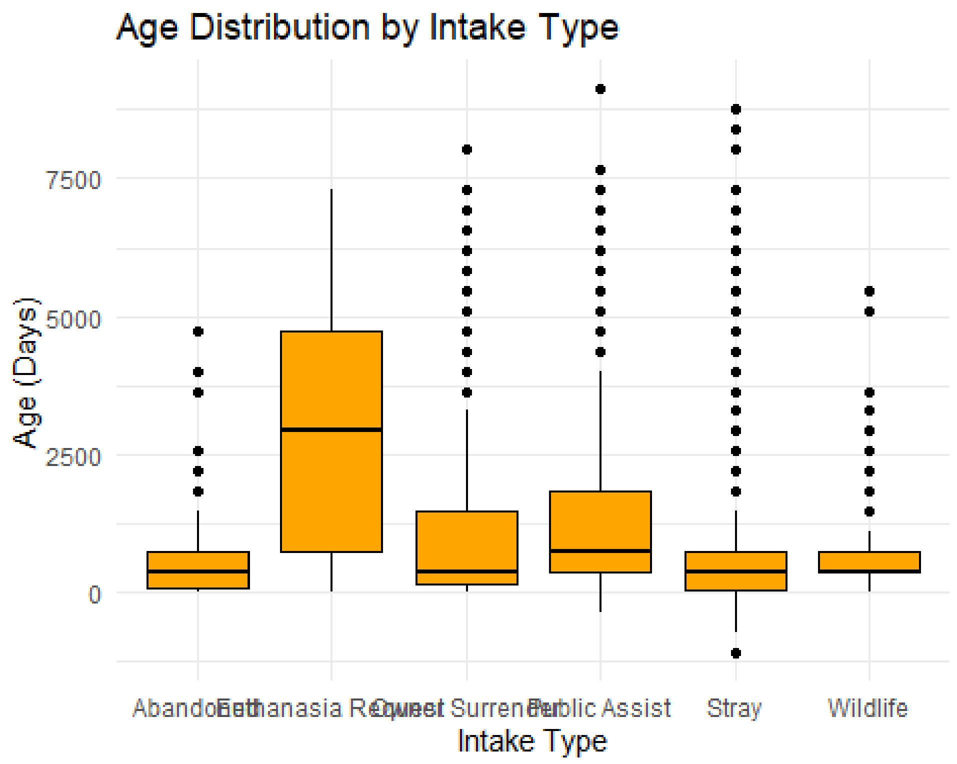
```r
# Boxplot of  Age by Animal Type
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Animal.Type, y = Age.in.Days)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Age Distribution by Animal Type", x = "Animal Type", y = "Age (Days)") +
  theme_minimal()
```



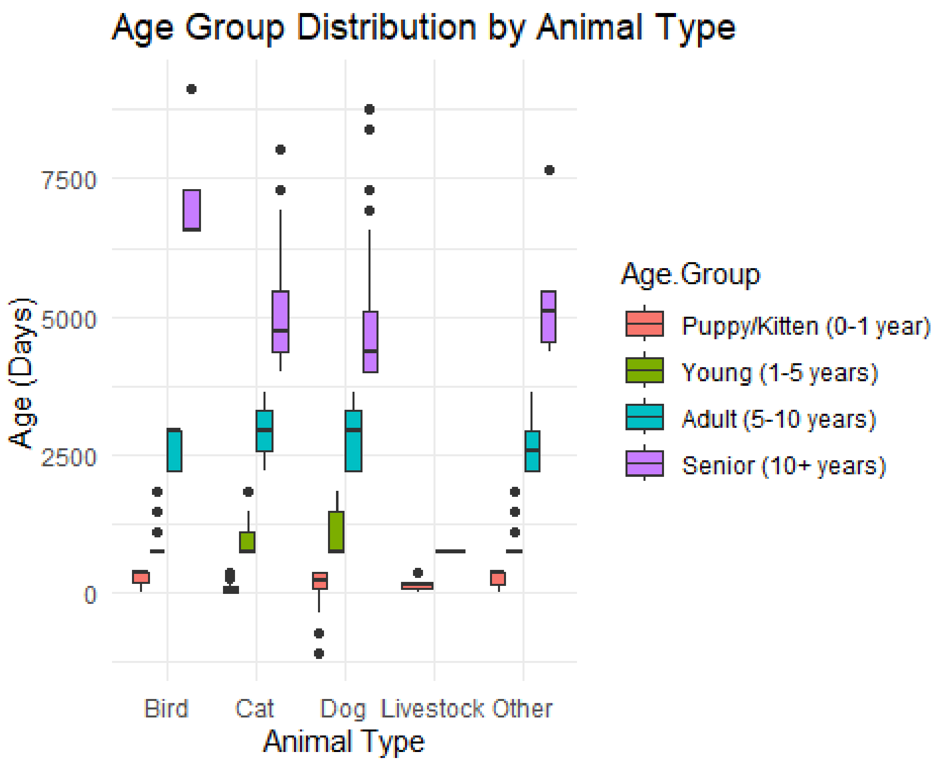
```r
# Boxplot of Age by Sex
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Sex.upon.Intake, y = A
ge.in.Days)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
  labs(title = "Age Distribution by Sex", x = "Sex upon Intake", y = "Age (Da
ys)") +
  theme_minimal() +
  coord_flip() # Better readability
```



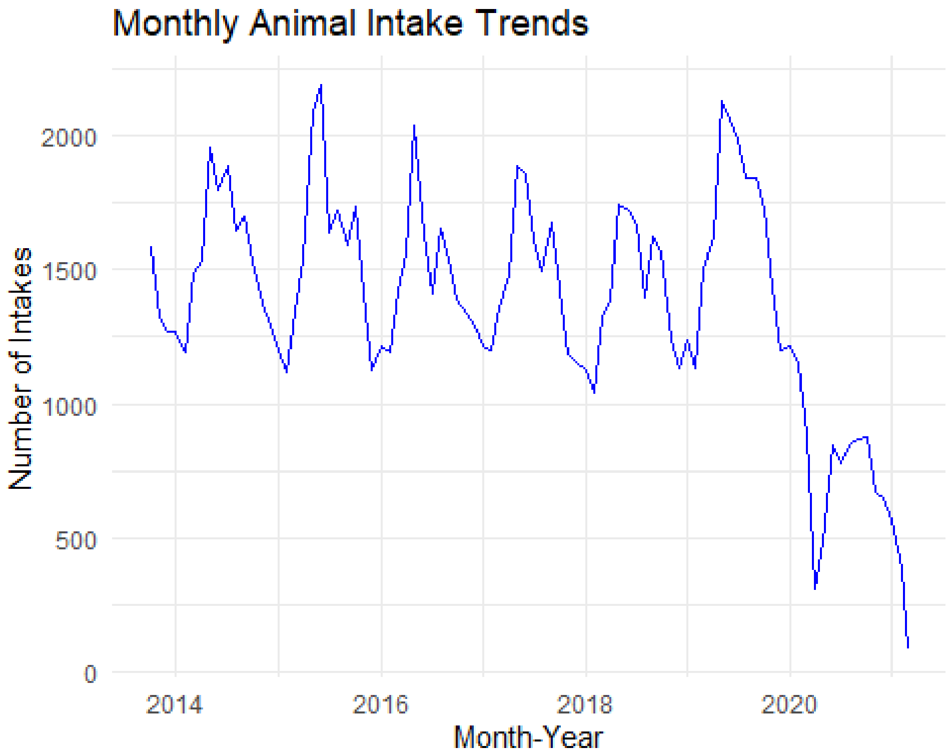
```
# Boxplot of  Age by Intake Type
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Intake.Type, y = Age.in.Days)) +
  geom_boxplot(fill = "orange", color = "black") +
  labs(title = "Age Distribution by Intake Type", x = "Intake Type", y = "Age (Days)") +
  theme_minimal()
```



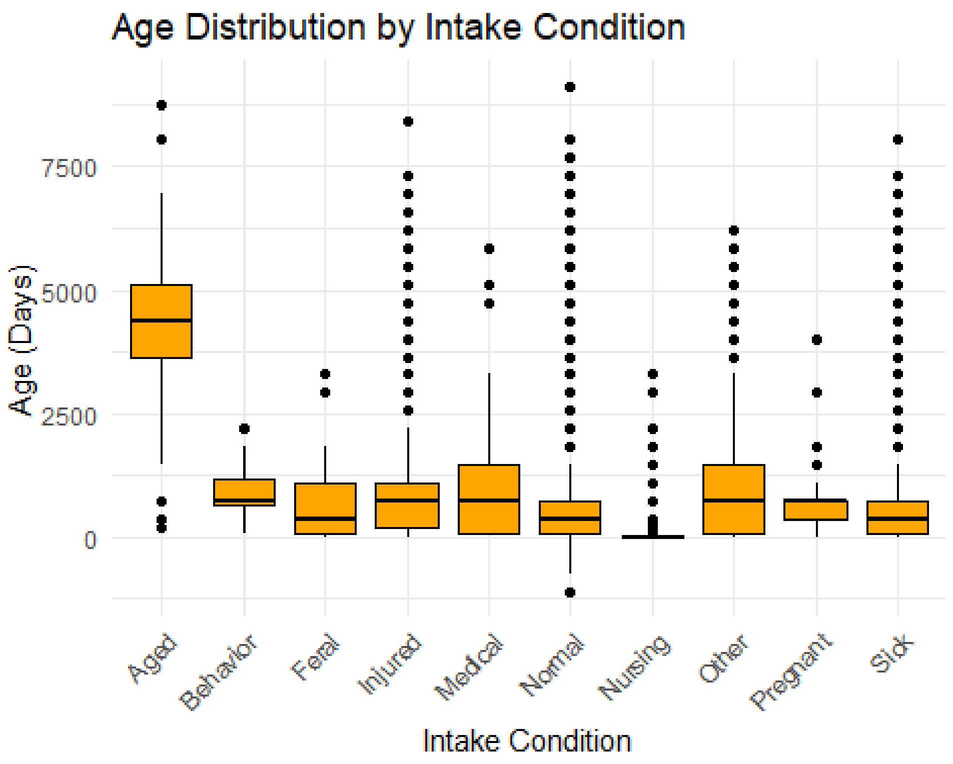
```r
# Boxplot of  Age Group by Animal Type
ggplot(Processed_Austin_Animal_Center_Intakes, aes(x = Animal.Type, y = Age.in.Days, fill = Age.Group)) +
  geom_boxplot() +
  labs(title = "Age Group Distribution by Animal Type", x = "Animal Type", y = "Age (Days)") +
  theme_minimal()
```



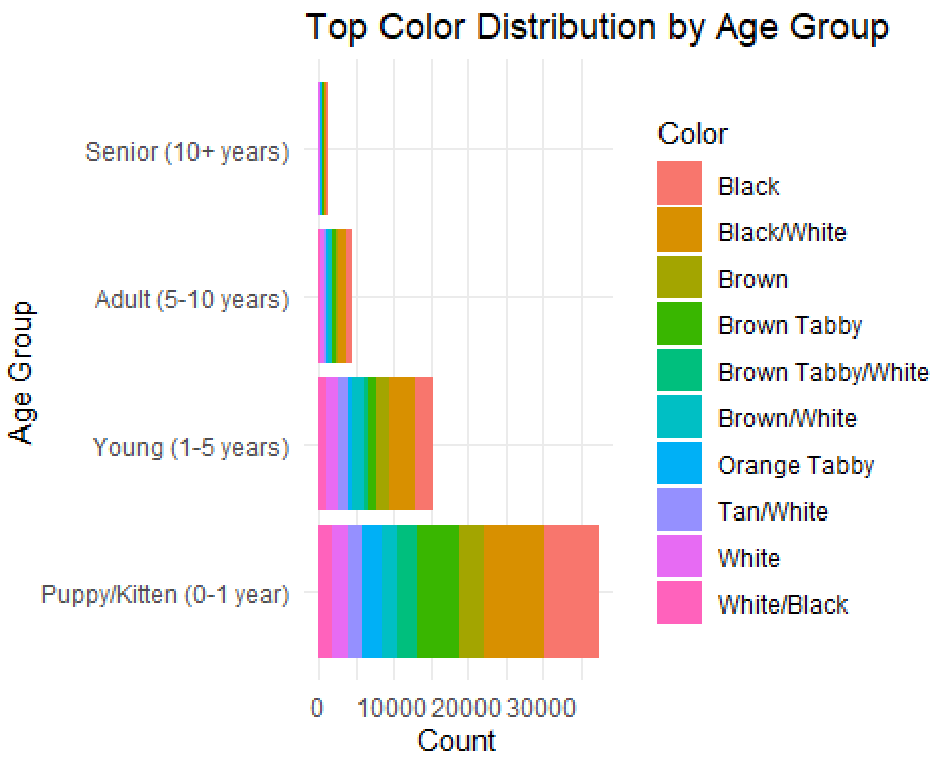
```
library(ggplot2)
library(dplyr)
install.packages("tidyr")

## Warning: package 'tidyr' is in use and will not be installed

library(tidyr)

data <- read.csv("Processed_Austin_Animal_Center_Intakes.csv")

data$DateTime <- as.Date(data$DateTime, format = "%Y-%m-%d")
data$Age.in.Days <- as.numeric(data$Age.in.Days)


data <- data[!is.na(data$DateTime) & !is.na(data$Age.in.Days), ]

data$Age.Group <- cut(
  data$Age.in.Days,
  breaks = c(-Inf, 365, 1825, 3650, Inf),
  labels = c("Puppy/Kitten (0-1 year)", "Young (1-5 years)", "Adult (5-10 yea
rs)", "Senior (10+ years)")
)

# 1. Time Series Graph
monthly_data <- data %>%
  mutate(MonthYear = format(DateTime, "%Y-%m")) %>%
  group_by(MonthYear) %>%
  summarise(Count = n())

ggplot(monthly_data, aes(x = as.Date(paste0(MonthYear, "-01")), y = Count)) +
  geom_line(color = "blue") +
  labs(title = "Monthly Animal Intake Trends", x = "Month-Year", y = "Number
of Intakes") +
  theme_minimal()
```



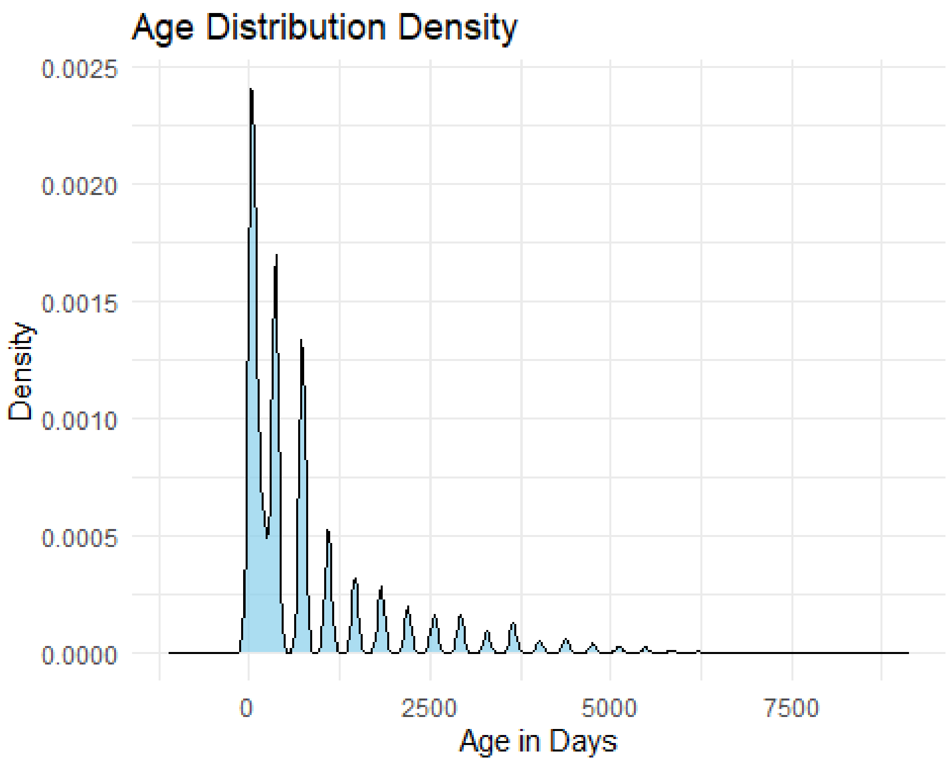
```
ggplot(data, aes(x = Intake.Condition, y = Age.in.Days)) +
  geom_boxplot(fill = "orange", color = "black") +
  labs(title = "Age Distribution by Intake Condition", x = "Intake Condition"
, y = "Age (Days)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



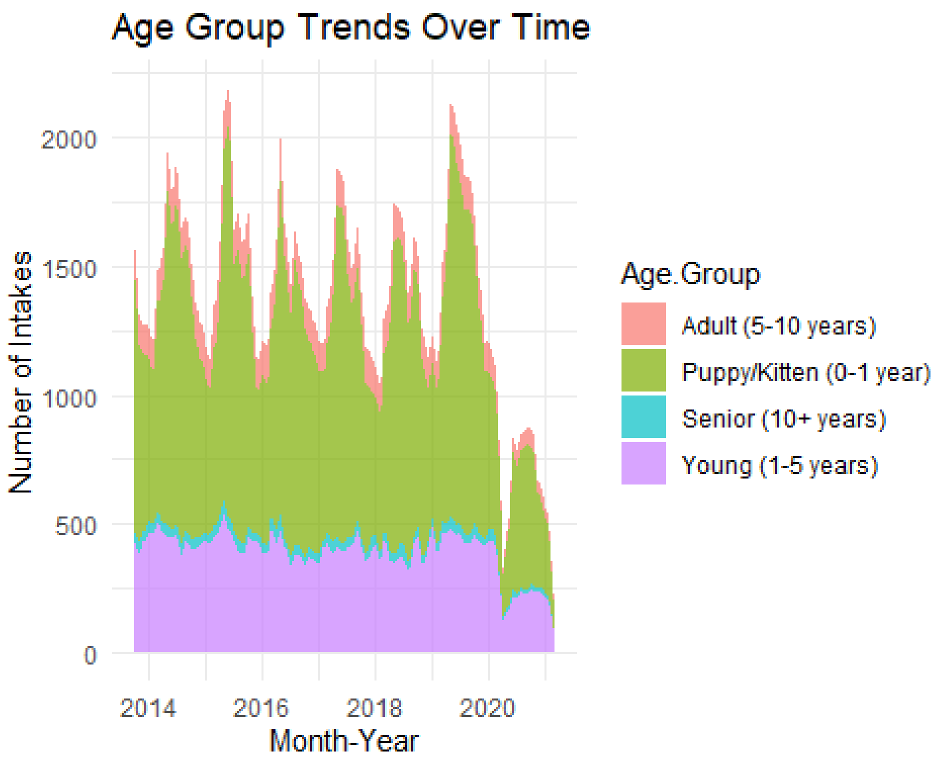
```
top_colors <- data %>%
  count(Color) %>%
  arrange(desc(n)) %>%
  slice_head(n = 10)

filtered_data <- data[data$Color %in% top_colors$Color, ]

ggplot(filtered_data, aes(x = Age.Group, fill = Color)) +
  geom_bar(position = "stack") +
  labs(title = "Top Color Distribution by Age Group", x = "Age Group", y = "C
ount") +
  theme_minimal() +
  coord_flip()
```



```
#  Density Graph

ggplot(data, aes(x = Age.in.Days)) +
  geom_density(fill = "skyblue", alpha = 0.7) +
  labs(title = "Age Distribution Density", x = "Age in Days", y = "Density")
+
  theme_minimal()
```



```
#  Time Series Density (Age Distribution)
data$YearMonth <- format(data$DateTime, "%Y-%m")

age_group_distribution <- data %>%
  group_by(YearMonth, Age.Group) %>%
  summarise(Count = n()) %>%
  pivot_wider(names_from = Age.Group, values_from = Count, values_fill = 0)

## `summarise()` has grouped output by 'YearMonth'. You can override using th
e
## `.groups` argument.

age_group_long <- pivot_longer(age_group_distribution, cols = -YearMonth, nam
es_to = "Age.Group", values_to = "Count")

ggplot(age_group_long, aes(x = as.Date(paste0(YearMonth, "-01")), y = Count,
fill = Age.Group)) +
  geom_area(alpha = 0.7) +
  labs(title = "Age Group Trends Over Time", x = "Month-Year", y = "Number of
Intakes") +
  theme_minimal()
```



NOTE: Some graphics may be similar to each other. I added both because there are minor differences between the two when viewed from a superficial and deep perspective.