141XP EDA

TEAM 4

2025-05-14

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.4.3
```

```
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.4.3
```

```
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.4.3
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```
## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
## v readr     2.1.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
belly_pain <- read.csv("belly_pain_features_osm_affect (1).csv", header = TRUE)
burping <- read.csv("burping_features_osm_affect (1).csv", header = TRUE)
discomfort <- read.csv("discomfort_features_osm_affect (1).csv", header = TRUE)
full_data_odd <- read.csv("filtered_full_data_odd.csv", header = TRUE)
tired <- read.csv("tired_features_osm_affect.csv", header = TRUE)
demographics <- read.csv("demographics_students.csv", header = TRUE)
```

```
head(demographics)
```

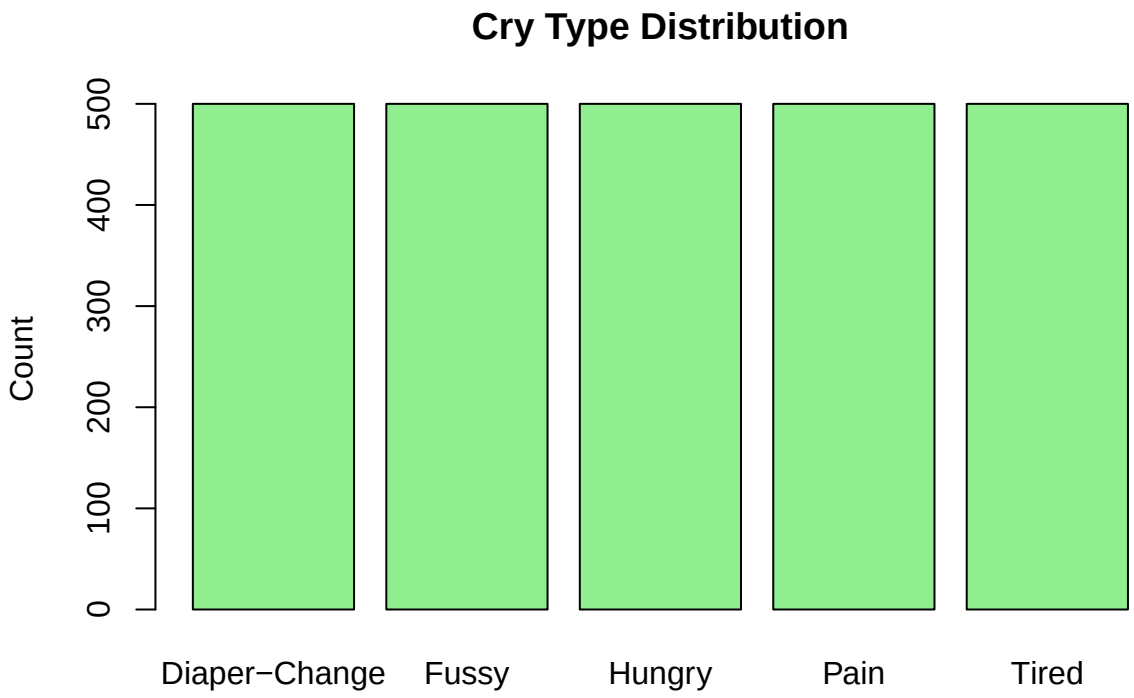
```
##                                 ID        Reason    Age    Gender   Date
## 1 bfb4662ea7ea4b8468d74c7ad1909ef1 Diaper-Change     49    female 181002
## 2 79eb1cf511da7ca57dd1996f0e0dca9e Diaper-Change    122    female 210811
## 3 1bb7c3a247deb74ec63b50048d97295b Diaper-Change NO-AGE      male 210609
## 4 aefc074bf9d634beeb762f45600060b7 Diaper-Change NO-AGE    female 220223
## 5                         NO-EMAIL Diaper-Change NO-AGE NO-GENDER 181223
## 6 5c78e65a7f0c779bc56ef188171ec829 Diaper-Change    241    female 180810
##    sample
## 1  340074
## 2 1099184
## 3 1048016
## 4 1306174
## 5  402716
## 6  283764
```

Reason Distribution

```
table(demographics$Reason)
```

```
##
## Diaper-Change         Fussy        Hungry          Pain         Tired
##           500           500           500           500           500
```

```r
barplot(table(demographics$Reason),
        col = "lightgreen",
        main = "Cry Type Distribution",
        ylab = "Count")
```

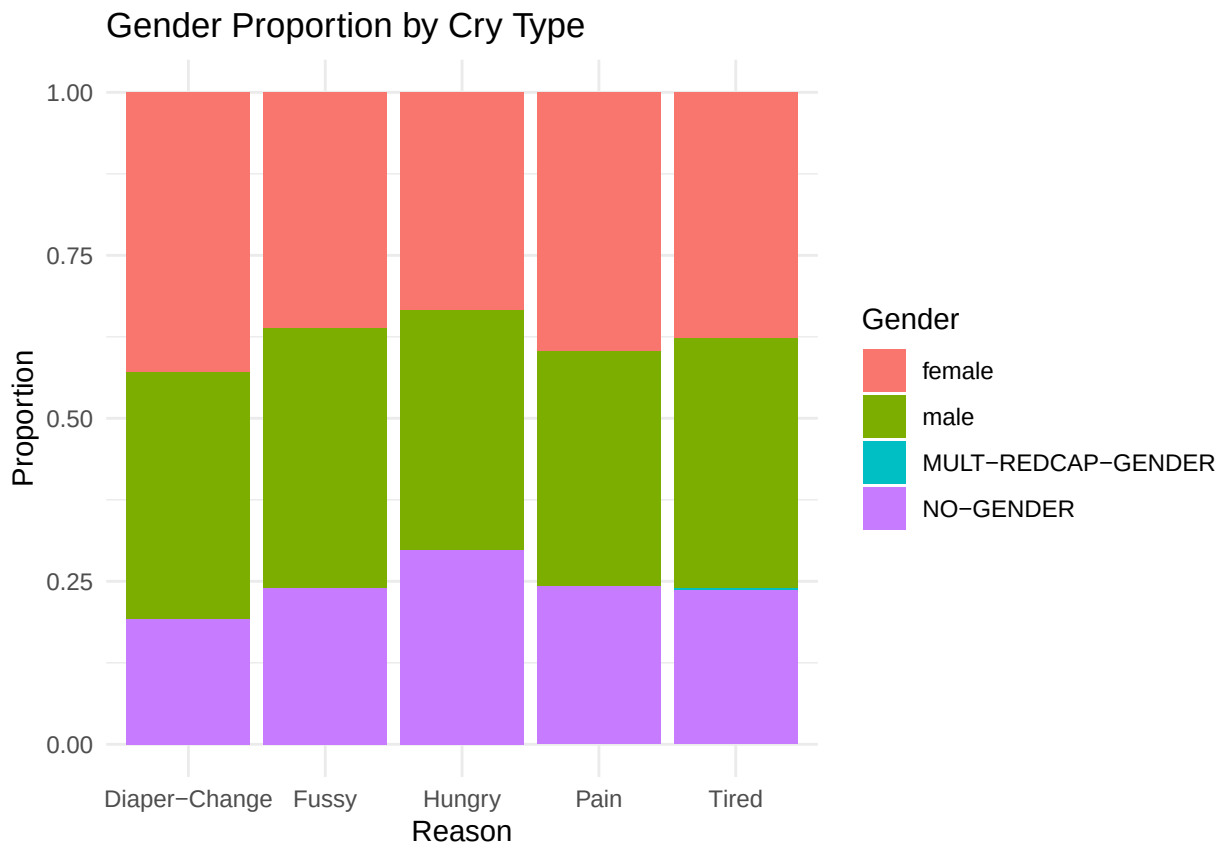


Reason with Gender Proportion

```r
table(demographics$Gender, demographics$Reason)
```

```
##
##                       Diaper-Change Fussy Hungry Pain Tired
##   female                        215   181    167  198   189
##   male                          189   199    184  181   191
##   MULT-REDCAP-GENDER              0     0      0    0     2
##   NO-GENDER                      96   120    149  121   118
```

```r
library(ggplot2)
ggplot(demographics, aes(x = Reason, fill = Gender)) +
  geom_bar(position = "fill") +
  labs(title = "Gender Proportion by Cry Type", y = "Proportion") +
  theme_minimal()
```



Chi-Square

```
chisq.test(table(demographics$Gender, demographics$Reason))
```

```
## Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(demographics$Gender, demographics$Reason)
## X-squared = 27.608, df = 12, p-value = 0.00631
```

ANOVA w/ removed missing ages

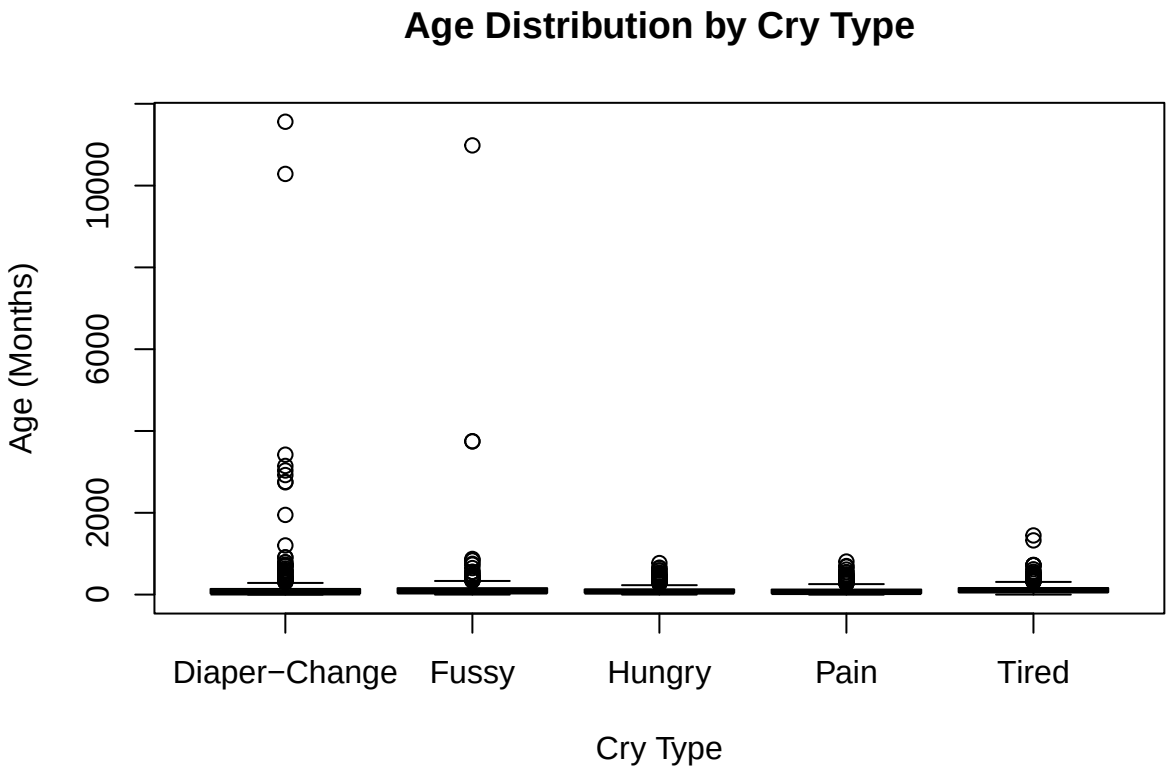
```
demographics$Age[demographics$Age == "NO-AGE"] <- NA
demographics$Age <- as.numeric(demographics$Age)
```

```
## Warning: NAs introduced by coercion
```

```r
anova_result <- aov(Age ~ Reason, data = demographics)
summary(anova_result)
```

```
##               Df    Sum Sq Mean Sq F value  Pr(>F)
## Reason         4   3832052  958013   3.648 0.00575 **
## Residuals   1761 462429901  262595
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 734 observations deleted due to missingness
```

```r
boxplot(Age ~ Reason, data = demographics,
        main = "Age Distribution by Cry Type",
        xlab = "Cry Type",
        ylab = "Age (Months)",
        col = "lightgreen")
```



Cry Acoustics Dimensions

```r
library(factoextra)
```

```
## Warning: package ’factoextra’ was built under R version 4.4.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
acoustic_features <- full_data_odd %>%
  select(where(is.numeric)) %>%
  na.omit()

scaled_features <- scale(acoustic_features)
pca <- prcomp(scaled_features, center = TRUE, scale. = TRUE)

fviz_eig(pca, addlabels = TRUE, barfill = "steelblue") +
  labs(title = "Scree Plot")
```