

# AudioProj Report

Maria Cordero    Yoon Hwang    Andrew Ramirez    Cherylyn Smith    Ai Yukino

2022-06-14

## The Relationship between Speech Acoustics and Gender

**Our Hypothesis:** Sounds where  $f_1 - f_0$  or  $f_1/f_0$  are larger are more likely to come from female speakers.

### R Packages/ Libraries

```
library(caret)
library(tidyverse)
library(factoextra)
library(cluster)
library(ggplot2)
library(e1071)
library(tidyr)
library(skimr)
library(readxl)
library(kernlab)
library(randomForest)
```

### Set working directory for R Studio

```
wd <- getwd() %>% str_replace("/scripts/R/praat_processing$", "")
setwd(wd)
getwd()

## [1] "C:/r/Ai-Yukino/smol-puhe"
rm(wd)

data <- read_csv("./data/filtered_data.csv")

## New names:
## Rows: 8564 Columns: 19
## -- Column specification
##   ----- Delimiter: ","
## (19): ...1, time, f1, b1, f2, b2, f3, b3, f4, b4, f1p, f2p, f3p, f4p, f0...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * ` `` -> `...1`
spec(data)

## cols(
##   ...1 = col_double(),
```

```

##   time = col_double(),
##   f1 = col_double(),
##   b1 = col_double(),
##   f2 = col_double(),
##   b2 = col_double(),
##   f3 = col_double(),
##   b3 = col_double(),
##   f4 = col_double(),
##   b4 = col_double(),
##   f1p = col_double(),
##   f2p = col_double(),
##   f3p = col_double(),
##   f4p = col_double(),
##   f0 = col_double(),
##   intensity = col_double(),
##   harmonicity = col_double(),
##   age = col_double(),
##   gender = col_double()
## )

```

## Supervised Model: Caret Prediction

```

data<- read_excel("./data/filtered_data.xls")

## New names:
## * ` ` -> `...`1` 

dim(data)

## [1] 8564    19

df <- data[c("f1", "f2", "f3", "f4", "f0", "intensity", "harmonicity", "age", "gender")]
head(df)

## # A tibble: 6 x 9
##       f1     f2     f3     f4     f0 intensity harmonicity    age gender
##   <dbl> <dbl> <dbl> <dbl> <dbl>      <dbl> <dbl> <dbl>
## 1  660. 2377. 2909. 4177. 232.      62.4   -215.    30     0
## 2  504. 2379. 2880. 4184. 232.      65.3   -213.    30     0
## 3  405. 2391. 2912. 4198. 231.      69      -200     30     0
## 4  388. 2440. 3053. 4205. 231.      72.5   -168.    30     0
## 5  438. 2497. 3150. 4159. 230.      75.5   -121     30     0
## 6  478. 2539. 3191. 4070. 230.      78.1   -68.8    30     0

# dimensions of data
dim(df)

## [1] 8564    9

#create output as target class
df$gender = as.factor(df$gender)

# Step 1: Get row numbers for the training data
trainRowNumbers <- createDataPartition(df$gender, p=0.8, list=FALSE)

# Step 2: Create the training data
trainData <- df[trainRowNumbers,]

```

```

# Step 3: Create the test data
testData <- df[-trainRowNumbers,]

# Store X and Y for later use.
x = trainData[ 1:5]
y = factor(trainData[[9]])

# Sys.setlocale( locale = 'Chinese')
Sys.setlocale('LC_ALL','C')

## [1] "C"

#skimmed <- skim_to_wide(trainData)
#skimmed[,]
skim(df)

```

Table 1: Data summary

Name	df
Number of rows	8564
Number of columns	9
<hr/>	
Column type frequency:	
factor	1
numeric	8
<hr/>	
Group variables	None

#### Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 5028, 1: 3536

#### Variable type: numeric

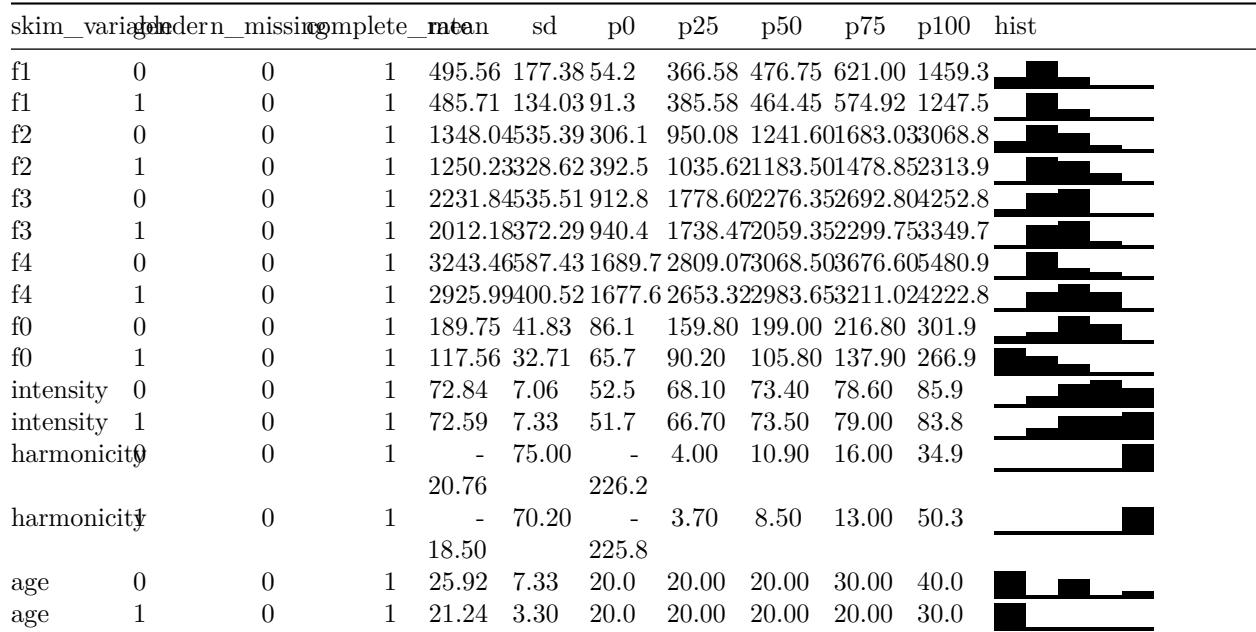
skim_variable	missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
f1	0	1	491.50	160.96	54.2	376.78	471.70	600.45	1459.3	
f2	0	1	1307.65	463.87	306.1	984.80	1202.65	1595.00	3068.8	
f3	0	1	2141.14	487.10	912.8	1757.47	2133.65	2537.83	4252.8	
f4	0	1	3112.38	541.52	1677.6	2750.18	3030.15	3381.98	5480.9	
f0	0	1	159.94	52.27	65.7	115.68	159.30	203.40	301.9	
intensity	0	1	72.74	7.17	51.7	67.50	73.45	78.80	85.9	
harmonicity	0	1	-	73.06	-	3.90	9.90	15.00	50.3	
			19.83		226.2					
age	0	1	23.99	6.43	20.0	20.00	20.00	30.00	40.0	

```
df %>% dplyr::group_by(gender) %>% skim()
```

Table 4: Data summary

Name	Piped data
Number of rows	8564
Number of columns	9
Column type frequency:	
numeric	8
Group variables	gender

### Variable type: numeric



```
#create a list of 80% of rows for training
validation_index <- createDataPartition(df$gender, p=0.80, list=FALSE)
validation <- df[-validation_index,]
df <- df[validation_index,]

# dimensions of data
dim(df)

## [1] 6852      9

#summarize attribute distributions
summary(df)

##          f1            f2            f3            f4
##  Min.   : 58.9   Min.   : 306.1   Min.   : 912.8   Min.   :1678
##  1st Qu.: 376.1   1st Qu.: 985.5   1st Qu.:1760.8   1st Qu.:2751
##  Median : 471.3   Median :1202.5   Median :2138.4   Median :3028
##  Mean   : 490.9   Mean   :1309.9   Mean   :2142.7   Mean   :3114
##  3rd Qu.: 599.4   3rd Qu.:1597.3   3rd Qu.:2536.1   3rd Qu.:3384
```

```

##   Max.    :1459.3    Max.    :3068.8    Max.    :4252.8    Max.    :5481
##   f0          intensity      harmonicity      age       gender
##   Min.    : 66.0    Min.    :51.70     Min.    :-226.20   Min.    :20.00   0:4023
##   1st Qu.:115.7   1st Qu.:67.50    1st Qu.: 3.90    1st Qu.:20.00   1:2829
##   Median  :159.2   Median  :73.40    Median : 9.80    Median :20.00
##   Mean    :159.9   Mean    :72.72    Mean   :-19.81   Mean   :24.01
##   3rd Qu.:203.5   3rd Qu.:78.80    3rd Qu.: 15.00   3rd Qu.:30.00
##   Max.    :301.9   Max.    :85.90    Max.   : 50.30   Max.   :40.00

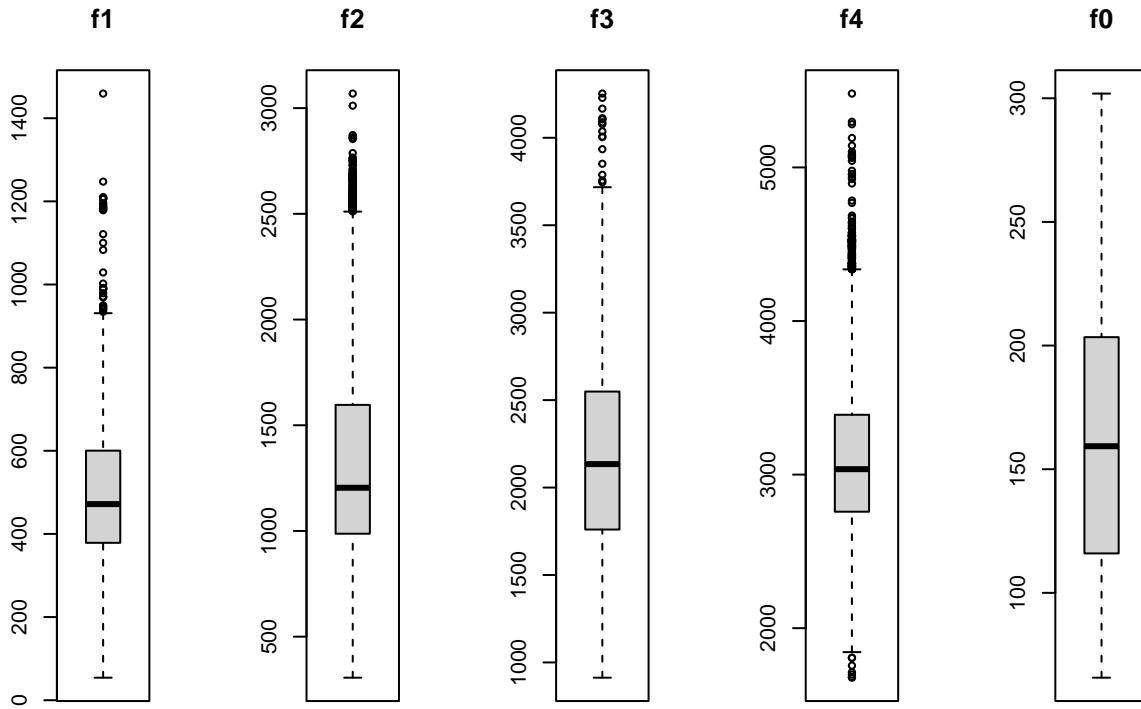
#summarize the class distribution
percentage <- prop.table(table(df$gender)) *100
y = factor(trainData[[9]])
cbind(freq=table(df$gender), percentage =percentage)

##   freq percentage
## 0 4023   58.71278
## 1 2829   41.28722

#split input and output
#x<-df[,1:5]
#y<-as.factor(df[[9]])

#boxplot for each attribute
par(mfrow=c(1,5))
for(i in 1:5){
  boxplot(x[,i], main=names(df)[i])
}

```



```

# Run algorithms using 10-fold cross validation
control <- trainControl(method="cv", number=10)
metric <- "Accuracy"

# a) linear algorithms
set.seed(7)
fit.lda <- train(gender~., data=df, method="lda", metric=metric, trControl=control)

# b) nonlinear algorithms
# CART
set.seed(7)
fit.cart <- train(gender~., data=df, method="rpart", metric=metric, trControl=control)

# kNN
set.seed(7)
fit.knn <- train(gender~., data=df, method="knn", metric=metric, trControl=control)

# c) advanced algorithms
# SVM
set.seed(7)
fit.svm <- train(gender~., data=df, method="svmRadial", metric=metric, trControl=control)

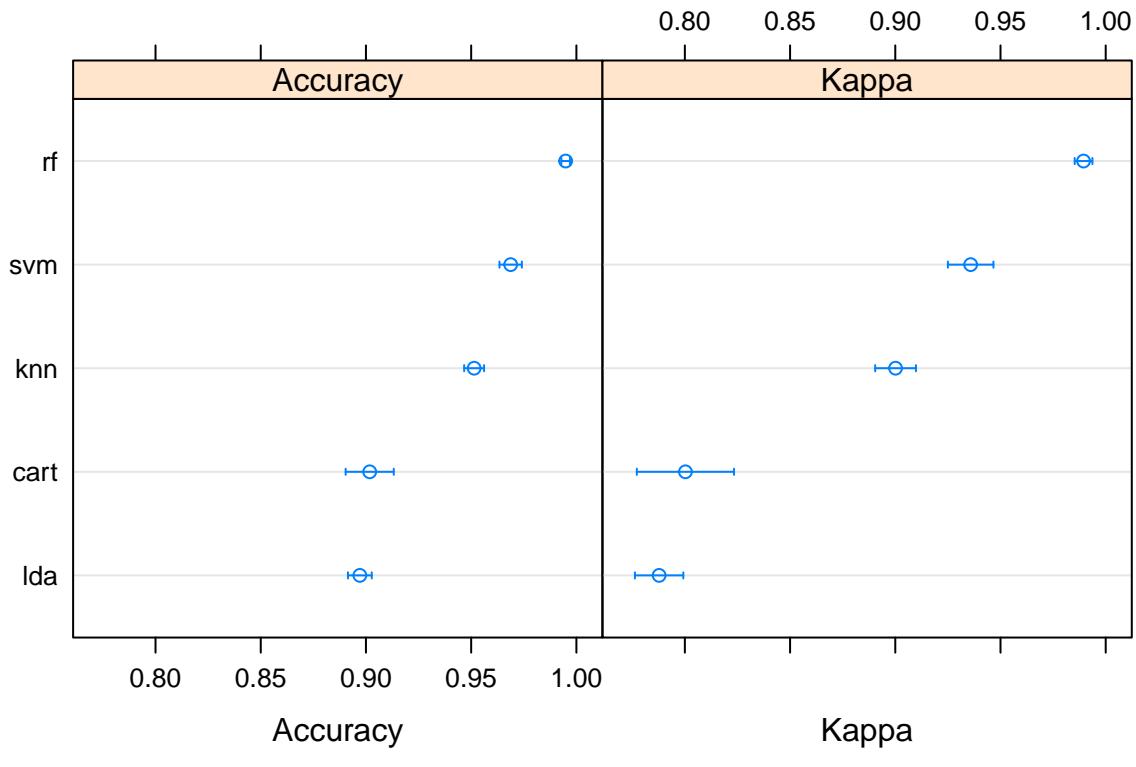
# Random Forest
set.seed(7)
fit.rf <- train(gender~., data=df, method="rf", metric=metric, trControl=control)

#summarize accuracy of models
results <- resamples(list(lda=fit.lda, cart=fit.cart, knn=fit.knn, svm=fit.svm, rf=fit.rf))
summary(results)

## 
## Call:
## summary.resamples(object = results)
##
## Models: lda, cart, knn, svm, rf
## Number of resamples: 10
##
## Accuracy
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## lda 0.8859649 0.8912409 0.8971558 0.8971068 0.9025915 0.9081633 0
## cart 0.8715328 0.8922562 0.9051808 0.9017826 0.9145985 0.9197080 0
## knn 0.9430657 0.9463695 0.9518248 0.9514037 0.9561755 0.9620438 0
## svm 0.9562044 0.9635451 0.9700730 0.9687670 0.9730321 0.9810219 0
## rf 0.9897959 0.9930715 0.9941606 0.9948924 0.9970803 0.9985423 0
##
## Kappa
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## lda 0.7654321 0.7758301 0.7871419 0.7877736 0.7988673 0.8110259 0
## cart 0.7402329 0.7794541 0.8079552 0.8002719 0.8252169 0.8360481 0
## knn 0.8827719 0.8900090 0.9008069 0.9001303 0.9097673 0.9221325 0
## svm 0.9104325 0.9250667 0.9384616 0.9358247 0.9445210 0.9608831 0
## rf 0.9789808 0.9856992 0.9879640 0.9894685 0.9939836 0.9969941 0

# compare accuracy of models
dotplot(results)

```



```
# best model is random forest
print(fit.rf)

## Random Forest
##
## 6852 samples
##     8 predictor
##     2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 6167, 6166, 6166, 6167, 6166, 6167, ...
## Resampling results across tuning parameters:
##
##     mtry  Accuracy   Kappa
##     2      0.9948924  0.9894685
##     5      0.9918274  0.9831368
##     8      0.9897836  0.9789223
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

*#analyze results in confusion matrix*

```
# estimate skill of rf on the validation data
predictions <- predict(fit.rf, validation)
confusionMatrix(predictions, validation$gender)
```

```

## Confusion Matrix and Statistics
##
##             Reference
## Prediction    0     1
##           0 1002     3
##           1     3 704
##
##                   Accuracy : 0.9965
##                   95% CI : (0.9924, 0.9987)
##       No Information Rate : 0.587
##       P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9928
##
## McNemar's Test P-Value : 1
##
##                   Sensitivity : 0.9970
##                   Specificity : 0.9958
##       Pos Pred Value : 0.9970
##       Neg Pred Value : 0.9958
##       Prevalence : 0.5870
##       Detection Rate : 0.5853
## Detection Prevalence : 0.5870
##       Balanced Accuracy : 0.9964
##
##       'Positive' Class : 0
##

```

## Supervised Learning: Classification

```

# Selecting specific columns
data <- data[, c('f0', 'f1', 'f2', 'gender')]
head(data)

## # A tibble: 6 x 4
##      f0     f1     f2 gender
##   <dbl> <dbl> <dbl> <dbl>
## 1 232.   660.  2377.     0
## 2 232.   504.  2379.     0
## 3 231.   405.  2391.     0
## 4 231.   388.  2440.     0
## 5 230.   438.  2497.     0
## 6 230.   478.  2539.     0

# Splitting data

library(caTools)

sample.split(data$gender, SplitRatio = 0.65) -> split_values
subset(data, split_values==T) -> train_set
subset(data, split_values==F) -> test_set

# Loading RPART to build classification model
library(rpart)

```

```

rpart(gender ~ ., data = train_set) -> mod_class
predict(mod_class, test_set, type = "matrix") -> result_class
table(test_set$gender, result_class)

##      result_class
##      0.0411985018726592 0.0431323283082077 0.114285714285714 0.369369369369369
##      0           134           1241            48            83
##      1              9             73            12            49
##      result_class
##      0.650602409638554 0.727355072463768 0.902349486049927
##      0           19           170            65
##      1           26           434            635

library(ggplot2)

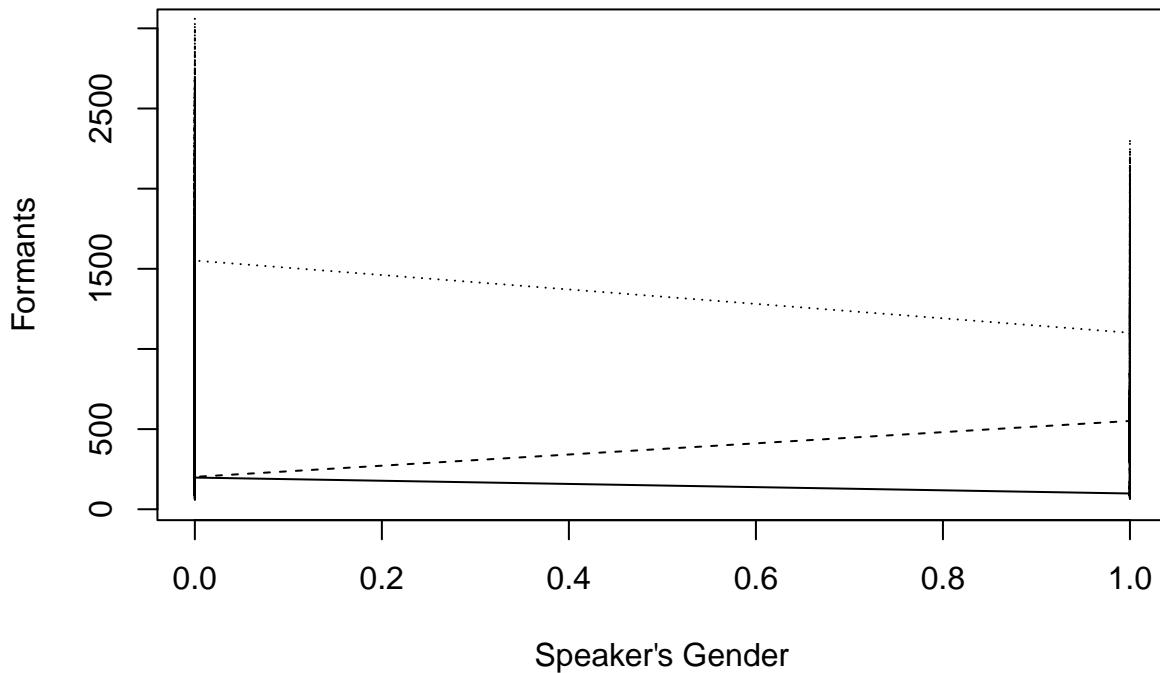
data <- read_csv("./data/filtered_data.csv")

## New names:
## Rows: 8564 Columns: 19
## -- Column specification
## ----- Delimiter: ","
## (19): ...1, time, f1, b1, f2, b2, f3, b3, f4, b4, f1p, f2p, f3p, f4p, f0...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`

plot(data$gender, data$f0,
      main = "Formants by Speaker's Gender",
      xlab = "Speaker's Gender",
      ylab = "Formants",
      type = "l",
      ylim = c(50, 3000))
lines(data$gender, data$f1,
      lty = "dashed")
lines(data$gender, data$f2,
      lty = "dotted")

```

## Formants by Speaker's Gender



## Unsupervised Learning: Principal Components Analysis

```
data <- data[, c('f0', 'f1', 'f2', 'gender')]

#Calculate the Principal Components
results <- prcomp(data, scale = TRUE)

#reverse the signs
results$rotation <- -1*results$rotation

#display principal components
results$rotation

##          PC1         PC2         PC3         PC4
## f0      0.69431289 -0.14718505  0.07206468  0.70076591
## f1      0.04984055  0.70981372  0.70208544  0.02750343
## f2      0.13751357  0.68771333 -0.70821956  0.08102747
## gender -0.70465279  0.03938812 -0.01754345  0.70824096

#reverse the signs of the scores
results$x <- -1*results$x

#display the first six scores
head(results$x)

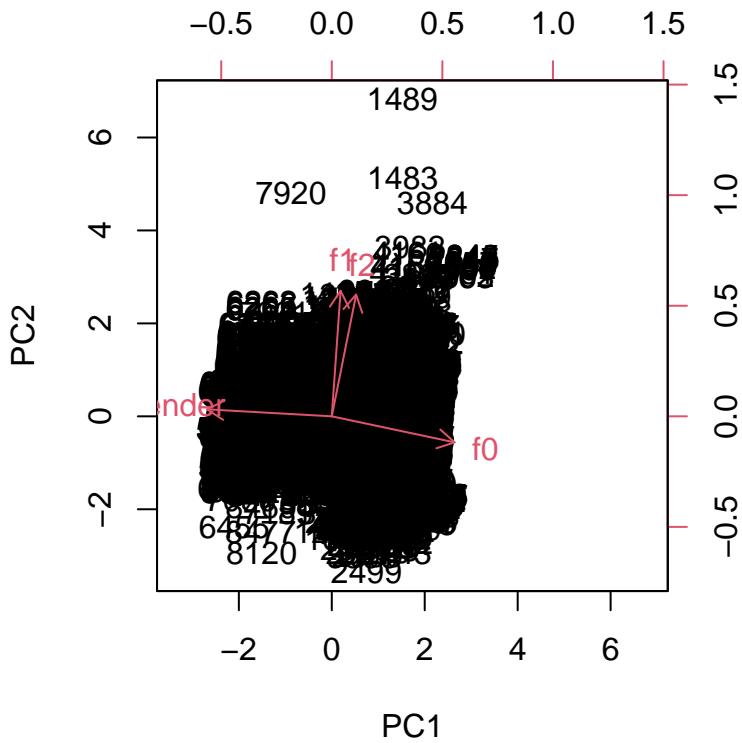
##          PC1         PC2         PC3         PC4
## [1,] 1.920106 2.0947255 -0.7817518  0.5905168
```

```

## [2,] 1.865724 1.4107358 -1.4663625 0.5574911
## [3,] 1.831773 0.9911598 -1.9184719 0.5358478
## [4,] 1.834400 0.9917099 -2.0651772 0.5347974
## [5,] 1.860228 1.2985551 -1.9352620 0.5466468
## [6,] 1.876848 1.5356663 -1.8286434 0.5526373

biplot(results, scale = 0)

```



```

#calculate total variance explained by each principal component
results$sdev^2 / sum(results$sdev^2)

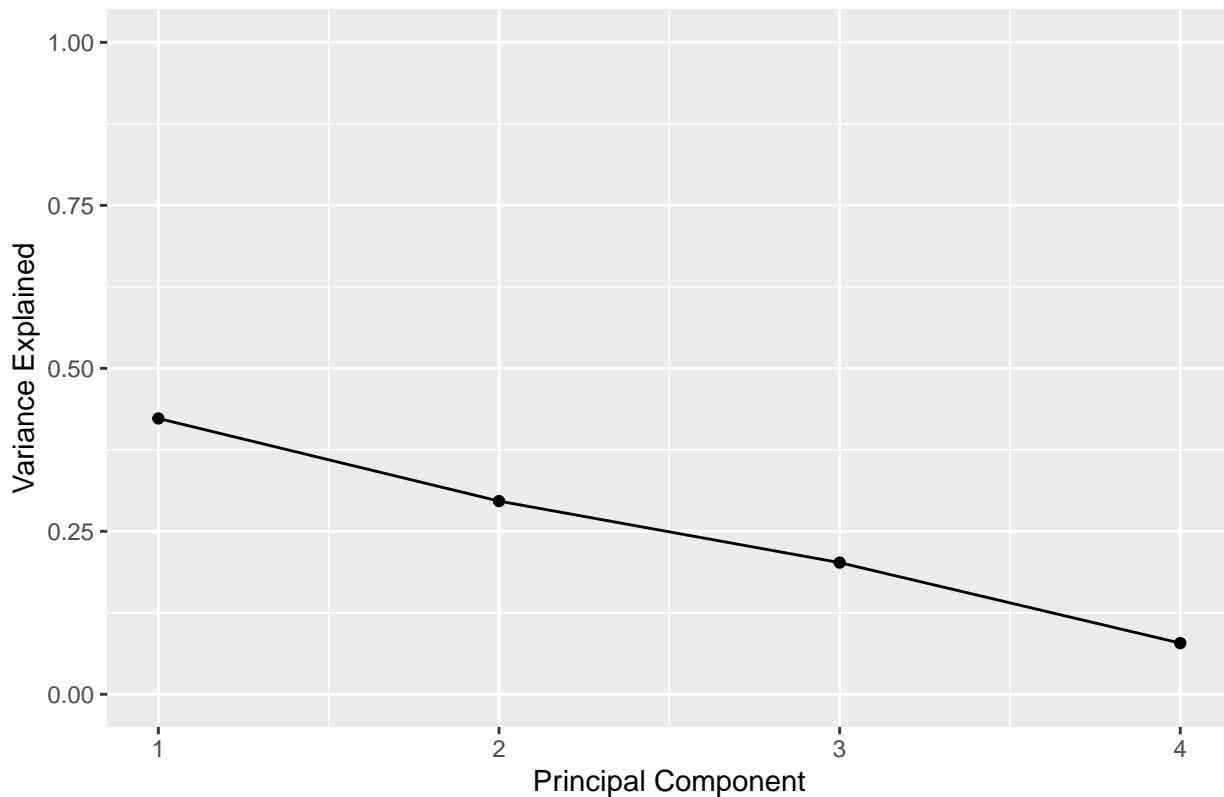
## [1] 0.42310653 0.29636698 0.20199740 0.07852909

#calculate total variance explained by each principal component
var_explained = results$sdev^2 / sum(results$sdev^2)

#create scree plot
qplot(c(1:4), var_explained) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1)

```

### Scree Plot



### Kmeans()

```
data <- read_csv("./data/filtered_data.csv")

## New names:
## Rows: 8564 Columns: 19
## -- Column specification
##   Delimiter: ","
## (19): ...1, time, f1, b1, f2, b2, f3, b3, f4, b4, f1p, f2p, f3p, f4p, f0...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`

data$difference = data$f1 - data$f0
data$quotient = data$f1/data$f0
dataMod = subset(data, select = -c(b1,b2,b3,b4,f1p,f2p,f3p,f4p, intensity, harmonicity, ...1, time))

model1 = lm(difference~gender, data = data)
summary(model1)

##
## Call:
## lm(formula = difference ~ gender, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.0000  -0.0000  -0.0000  -0.0000  -0.0000
```

```

## -450.31 -113.88 -16.36 113.09 985.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 305.813     2.360 129.56 <2e-16 ***
## gender       62.344     3.673 16.97 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 167.4 on 8562 degrees of freedom
## Multiple R-squared:  0.03255, Adjusted R-squared:  0.03244
## F-statistic: 288.1 on 1 and 8562 DF, p-value: < 2.2e-16
model2 = lm(quotient~gender, data = data)
summary(model2)

##
## Call:
## lm(formula = quotient ~ gender, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6579 -1.0880 -0.2212  0.7798  7.5781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.81051    0.02086 134.73 <2e-16 ***
## gender      1.58823    0.03246  48.92 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 8562 degrees of freedom
## Multiple R-squared:  0.2185, Adjusted R-squared:  0.2184
## F-statistic: 2393 on 1 and 8562 DF, p-value: < 2.2e-16
model3 = lm(difference~., data = subset(dataMod, select = -c(quotient, f0, f1)))
summary(model3)

##
## Call:
## lm(formula = difference ~ ., data = subset(dataMod, select = -c(quotient,
## f0, f1)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -425.60 -114.10 -27.02  116.59  750.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.558748  12.059524  3.032  0.00244 **
## f2          0.047345  0.004416 10.722 < 2e-16 ***
## f3         -0.051975  0.005376 -9.669 < 2e-16 ***
## f4          0.104182  0.004649 22.411 < 2e-16 ***
## age        -0.635777  0.302496 -2.102  0.03560 *
## gender      85.659570  3.827000 22.383 < 2e-16 ***

```





















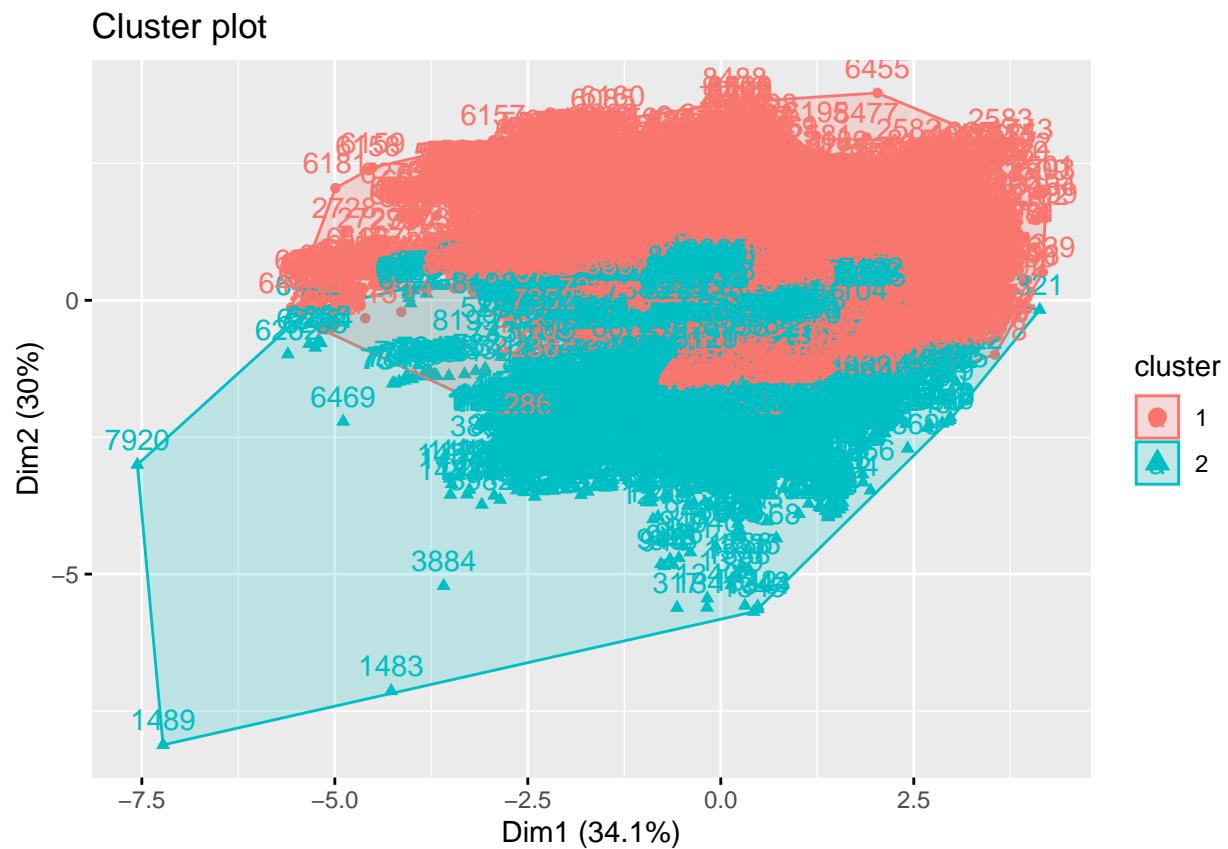






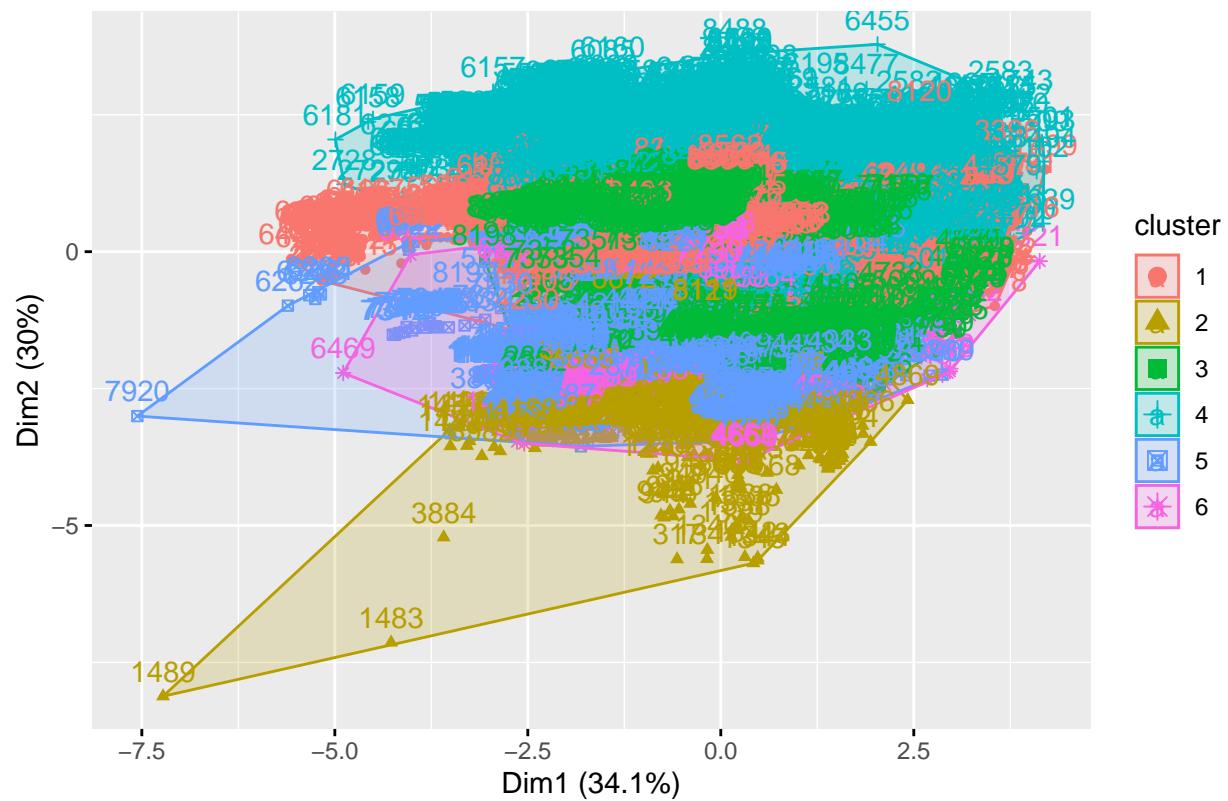




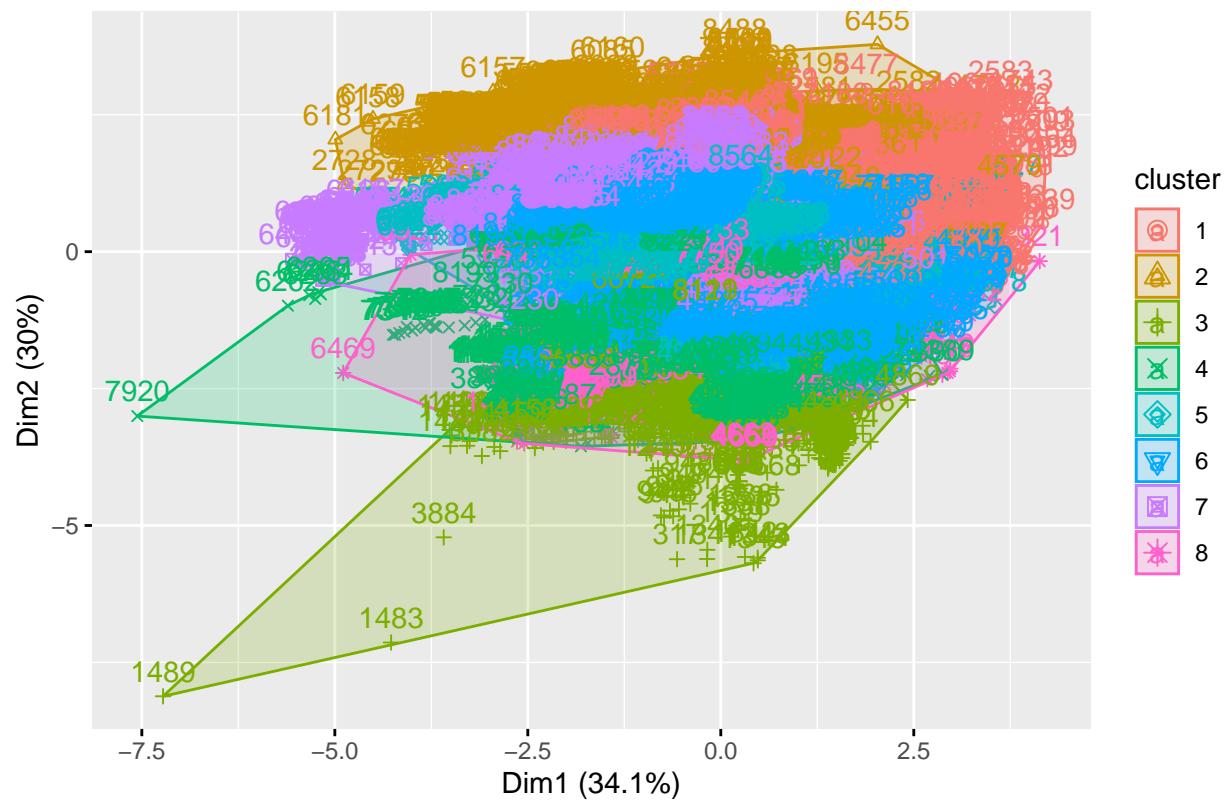


```
fviz_cluster(km2, dataMod)
```

Cluster plot



## Cluster plot























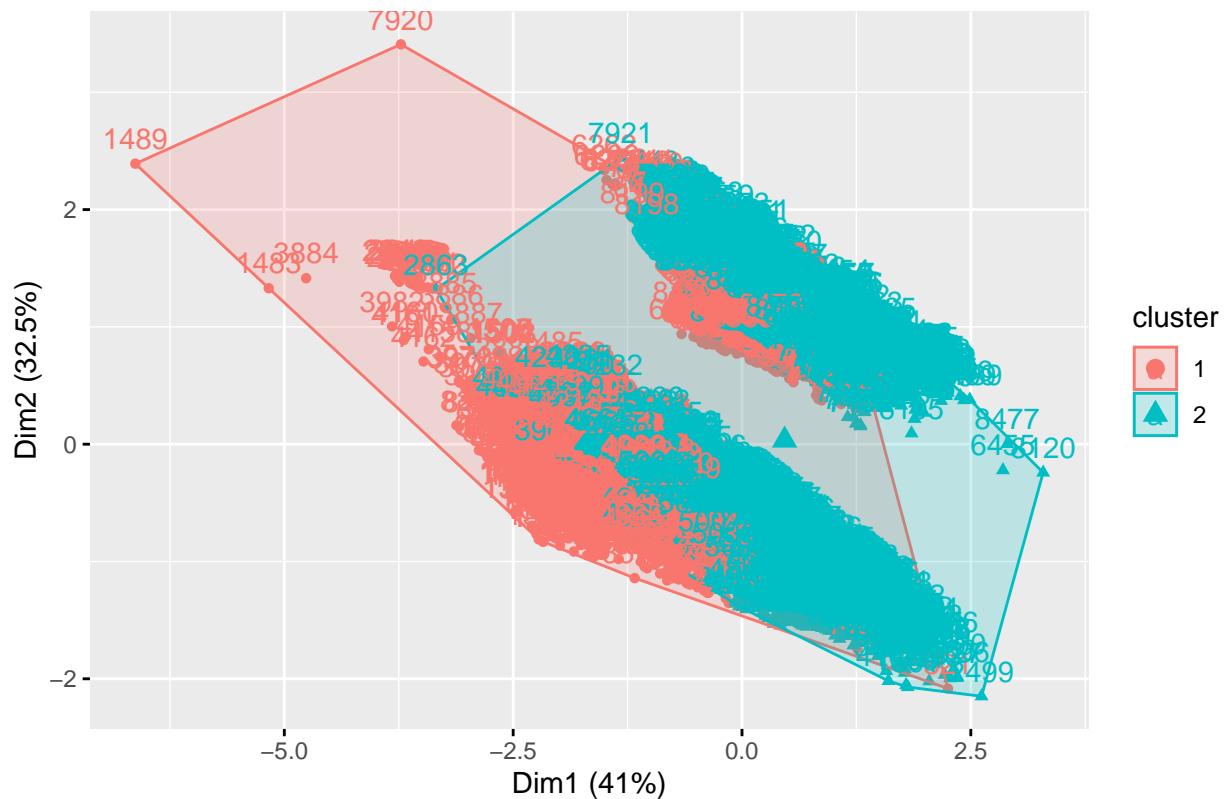






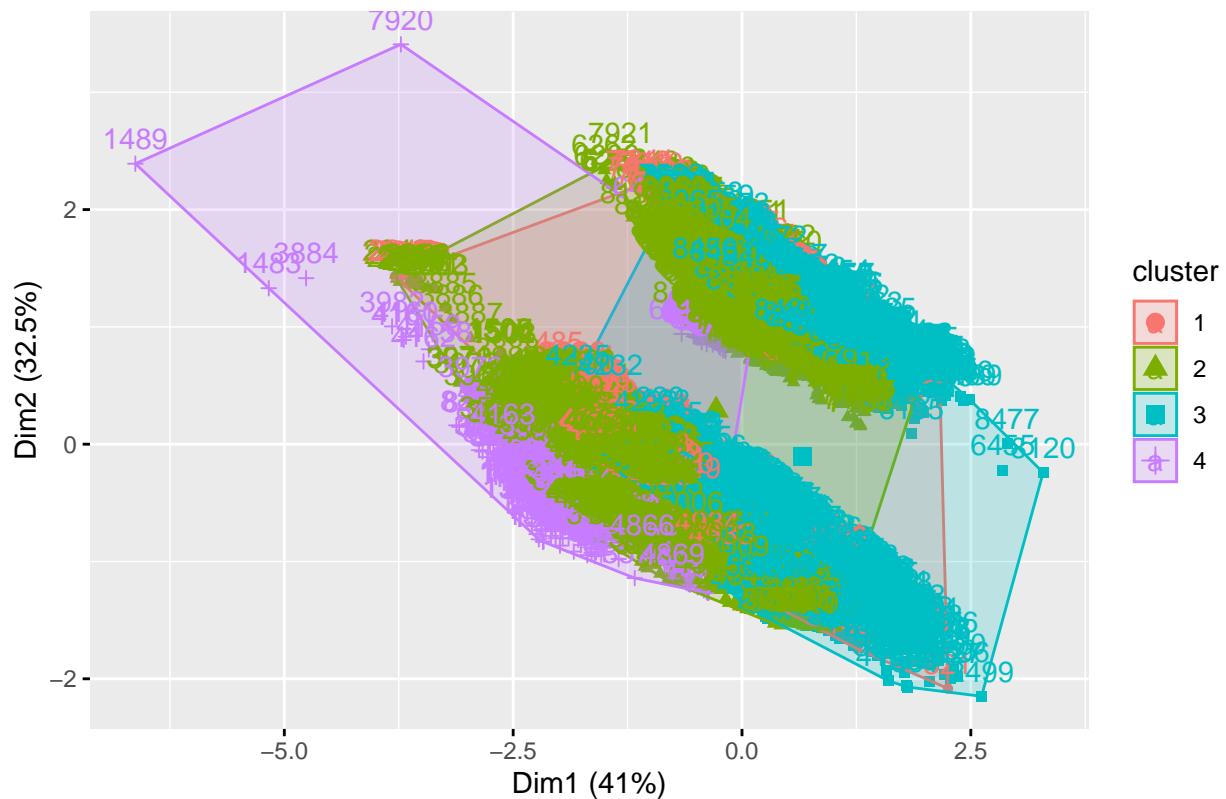


## Cluster plot

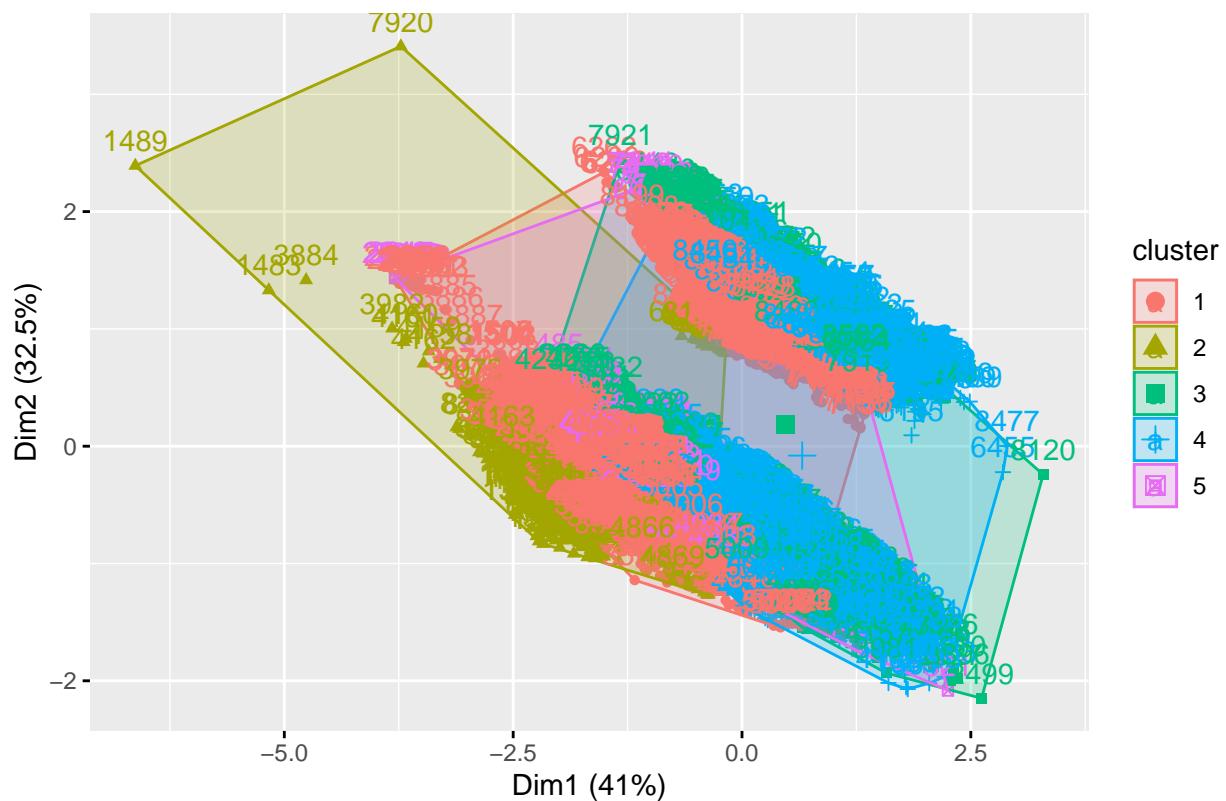


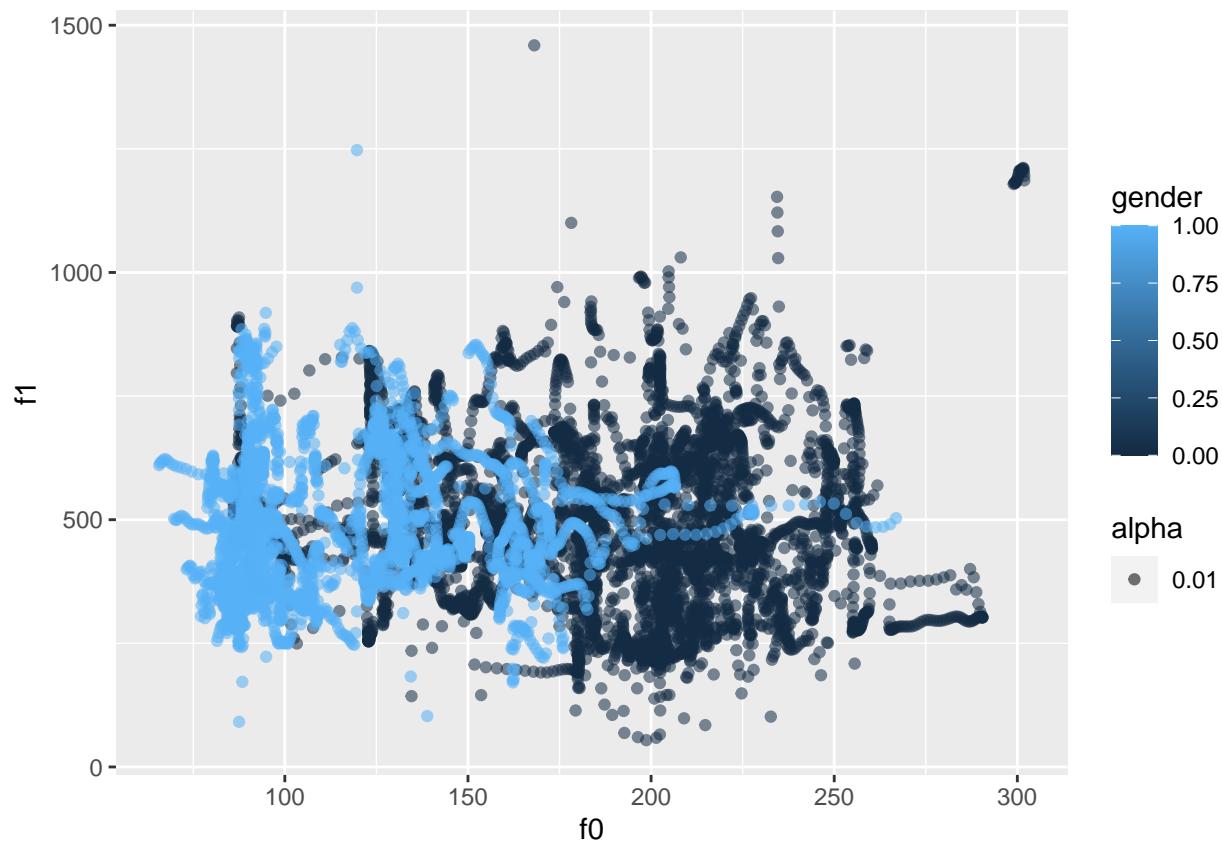
```
fviz_cluster(km5, subset(dataMod, select = c(f1, f2, gender)))
```

Cluster plot



Cluster plot





```

fullMod = lm(gender~f0 + f1 + f2, dataMod)
redMod = lm(gender~f0, dataMod)
anova(redMod, fullMod)

## Analysis of Variance Table
##
## Model 1: gender ~ f0
## Model 2: gender ~ f0 + f1 + f2
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   8562 1116.0
## 2   8560 1097.3  2     18.742 73.103 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

test = lm(gender~f0, dataMod)
summary(test)

##
## Call:
## lm(formula = gender ~ f0, data = dataMod)
##
## Residuals:
##       Min     1Q Median     3Q    Max 
## -0.88586 -0.24021  0.02176  0.19727  1.27221 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.02176   0.02176   1.000   0.3173    
## f0          -0.24021   0.02176  -11.000  < 2.2e-16 ***
## f1           0.19727   0.02176   9.000  < 2.2e-16 ***
## f2           0.12858   0.02176   5.900  < 2.2e-16 ***
## 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## (Intercept) 1.437e+00 1.256e-02 114.45 <2e-16 ***
## f0 -6.405e-03 7.464e-05 -85.82 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.361 on 8562 degrees of freedom
## Multiple R-squared: 0.4624, Adjusted R-squared: 0.4624
## F-statistic: 7365 on 1 and 8562 DF, p-value: < 2.2e-16
test2 = lm(gender~f0 + f1, dataMod)
summary(test2)

##
## Call:
## lm(formula = gender ~ f0 + f1, data = dataMod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89008 -0.23932  0.02505  0.19688  1.27451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.503e+00 1.743e-02 86.185 < 2e-16 ***
## f0 -6.413e-03 7.453e-05 -86.044 < 2e-16 ***
## f1 -1.304e-04 2.420e-05 -5.387 7.37e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3604 on 8561 degrees of freedom
## Multiple R-squared: 0.4642, Adjusted R-squared: 0.4641
## F-statistic: 3709 on 2 and 8561 DF, p-value: < 2.2e-16
test3 = lm(gender~f0 + f1 + f0*f1, dataMod)
summary(test3)

##
## Call:
## lm(formula = gender ~ f0 + f1 + f0 * f1, data = dataMod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90859 -0.23638  0.01256  0.21176  1.27061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.846e+00 3.979e-02 46.392 <2e-16 ***
## f0 -8.392e-03 2.195e-04 -38.240 <2e-16 ***
## f1 -8.241e-04 7.628e-05 -10.804 <2e-16 ***
## f0:f1 4.007e-06 4.181e-07  9.584 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3585 on 8560 degrees of freedom
## Multiple R-squared: 0.4699, Adjusted R-squared: 0.4697
## F-statistic: 2530 on 3 and 8560 DF, p-value: < 2.2e-16

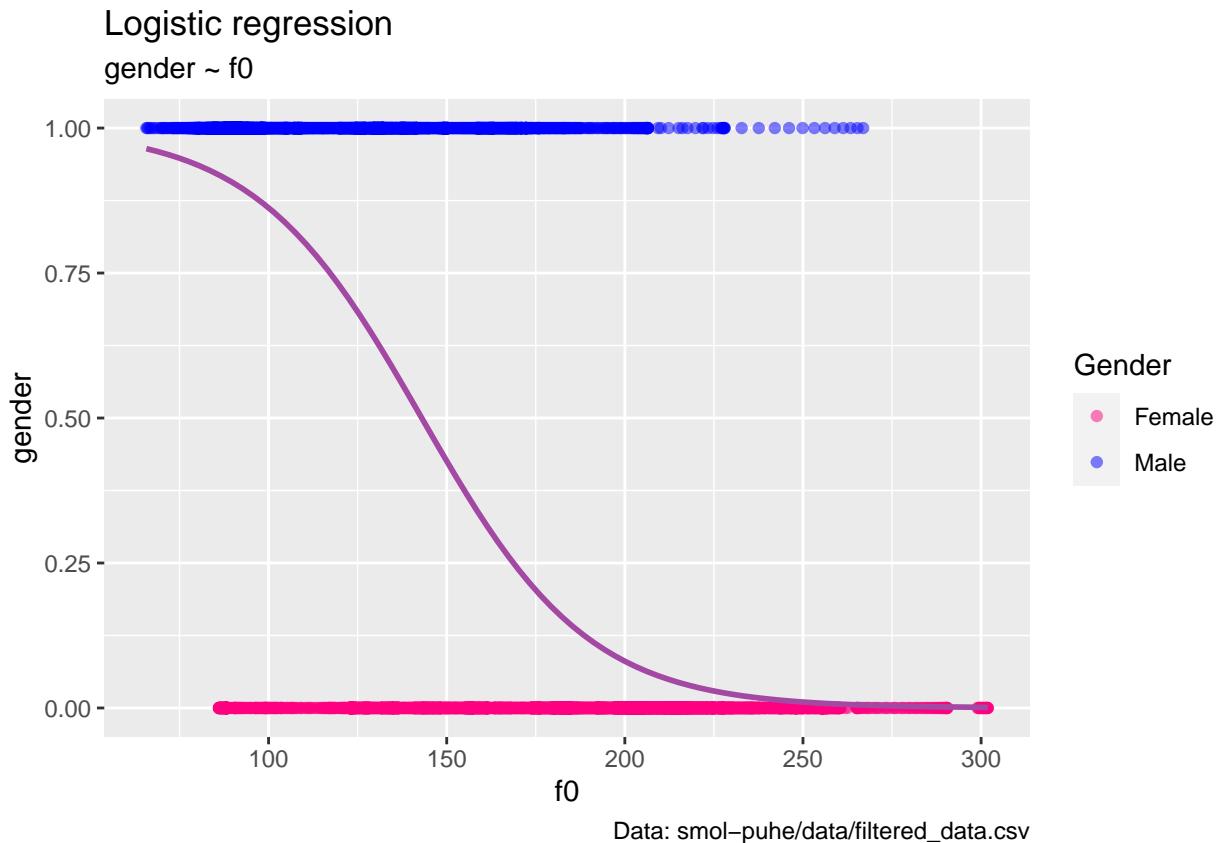
```

## Logistic Regression

```
## Logistic regression
## gender ~ f0

ggplot(data, aes(x=f0, y=gender)) +
  geom_point(alpha=.5, aes(color = cut(data$gender, c(-1, 0.5, 1)))) +
  scale_color_manual(name = "Gender",
                     values = c("(-1,0.5]" = "#ff0080",
                               "(0.5,1]" = "#0000ff"),
                     labels = c("Female", "Male")) +
  stat_smooth(method="glm", se=FALSE,
              method.args = list(family=binomial), col="#a349a4") +
  labs(title = "Logistic regression",
       subtitle = "gender ~ f0",
       caption = "Data: smol-puhe/data/filtered_data.csv")

## `geom_smooth()` using formula 'y ~ x'
```



```
# ggsave("logistic_gender_f0.png")

model_f0 <- glm(gender~f0, data = data)
summary(model_f0)

## Call:
## glm(formula = gender ~ f0, data = data)
```

```

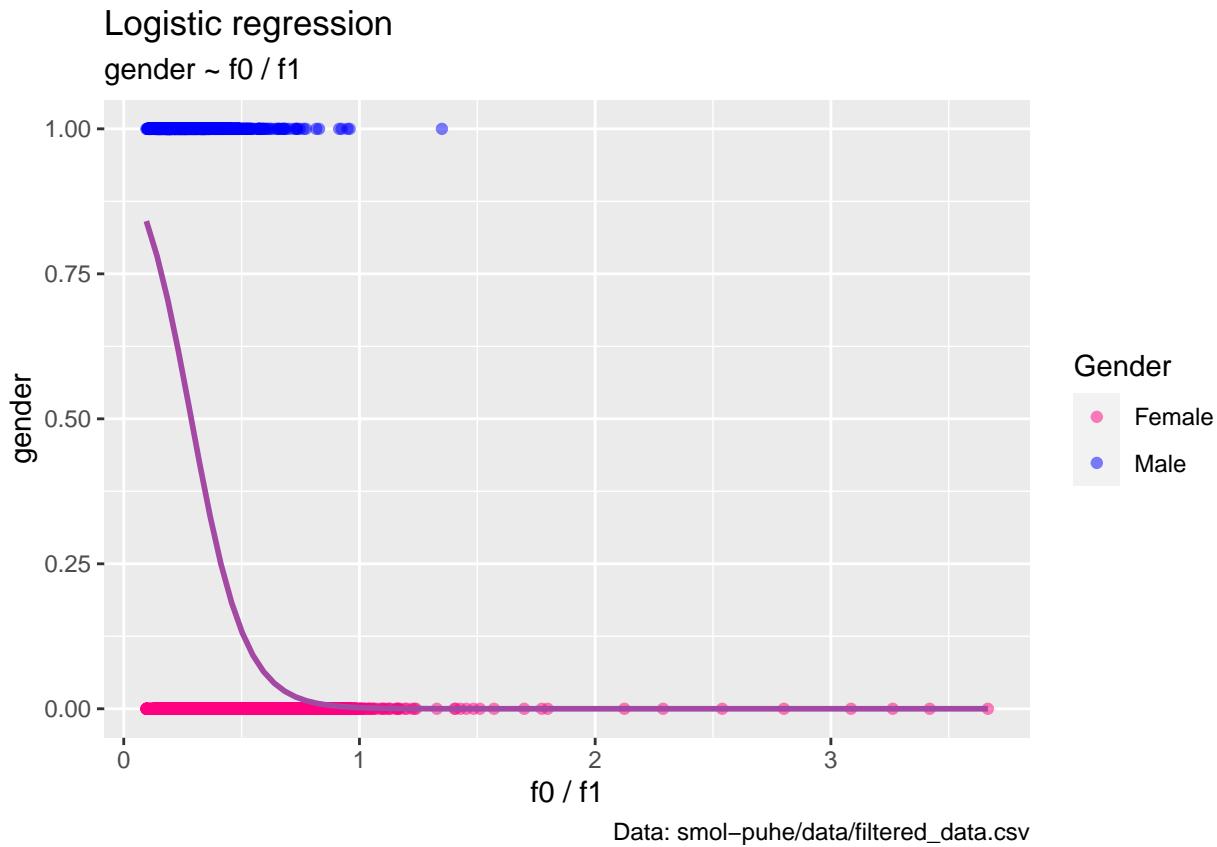
## 
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88586 -0.24021  0.02176  0.19727  1.27221
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.437e+00 1.256e-02 114.45 <2e-16 ***
## f0          -6.405e-03 7.464e-05 -85.82 <2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.1303468)
## 
## Null deviance: 2076 on 8563 degrees of freedom
## Residual deviance: 1116 on 8562 degrees of freedom
## AIC: 6857.9
## 
## Number of Fisher Scoring iterations: 2
## Logistic regression -----
## gender ~ f0 / f1

data$f0_over_f1 <- data$f0 / data$f1

ggplot(data, aes(x=f0_over_f1, y=gender)) +
  geom_point(alpha=.5, aes(color = cut(data$gender, c(-1, 0.5, 1)))) +
  scale_color_manual(name = "Gender",
                     values = c("(-1,0.5]" = "#ff0080",
                               "(0.5,1]" = "#0000ff"),
                     labels = c("Female", "Male")) +
  stat_smooth(method="glm", se=FALSE,
              method.args = list(family=binomial), col="#a349a4") +
  labs(title = "Logistic regression",
       subtitle = "gender ~ f0 / f1",
       caption = "Data: smol-puhe/data/filtered_data.csv") +
  xlab("f0 / f1")

## `geom_smooth()` using formula 'y ~ x'

```



```
# ggsave("logistic_gender_f0_over_f1.png")

model_f0_over_f1 <- glm(gender~f0_over_f1, data = data, family="binomial")
summary(model_f0_over_f1)
```

```
##
## Call:
## glm(formula = gender ~ f0_over_f1, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.9164   -0.9478   -0.2153    0.9317    4.3118
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.50409   0.07563  33.11   <2e-16 ***
## f0_over_f1 -8.74146   0.23893 -36.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 11611.0  on 8563  degrees of freedom
## Residual deviance: 9137.9  on 8562  degrees of freedom
## AIC: 9141.9
##
```

```

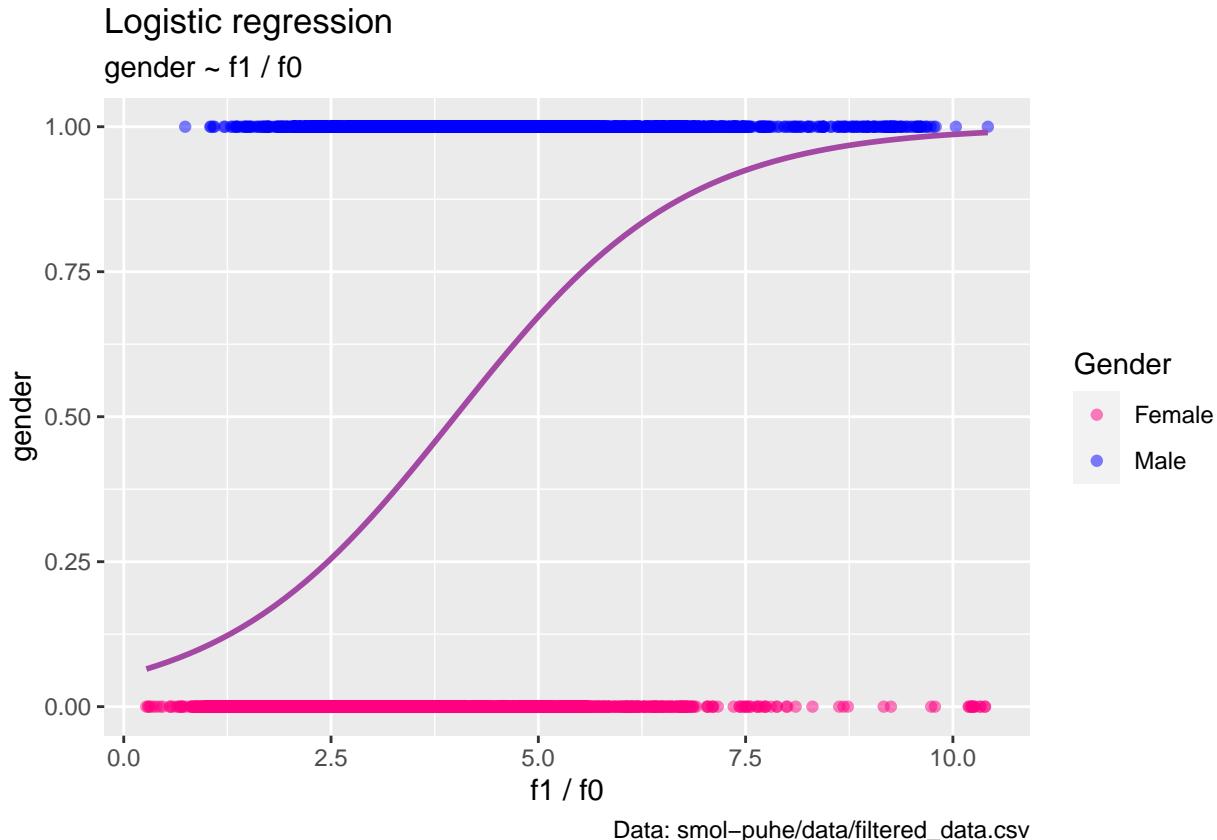
## Number of Fisher Scoring iterations: 5
## Logistic regression #####
## gender ~ f1 / f0

data$f1_over_f0 = data$f1 / data$f0

ggplot(data, aes(x=f1_over_f0, y=gender)) +
  geom_point(alpha=.5, aes(color = cut(data$gender, c(-1, 0.5, 1)))) +
  scale_color_manual(name = "Gender",
                     values = c("(-1,0.5]" = "#ff0080",
                               "(0.5,1]" = "#0000ff"),
                     labels = c("Female", "Male")) +
  stat_smooth(method="glm", se=FALSE,
              method.args = list(family=binomial),
              col="#a349a4") +
  labs(title = "Logistic regression",
       subtitle = "gender ~ f1 / f0",
       caption = "Data: smol-puhe/data/filtered_data.csv") +
  xlab("f1 / f0")

```

## `geom\_smooth()` using formula 'y ~ x'



```

# ggsave("logistic_gender_f1_over_f0.png")

model_f1_over_f0 <- glm(gender~f1_over_f0, data = data, family="binomial")
summary(model_f1_over_f0)

```

```

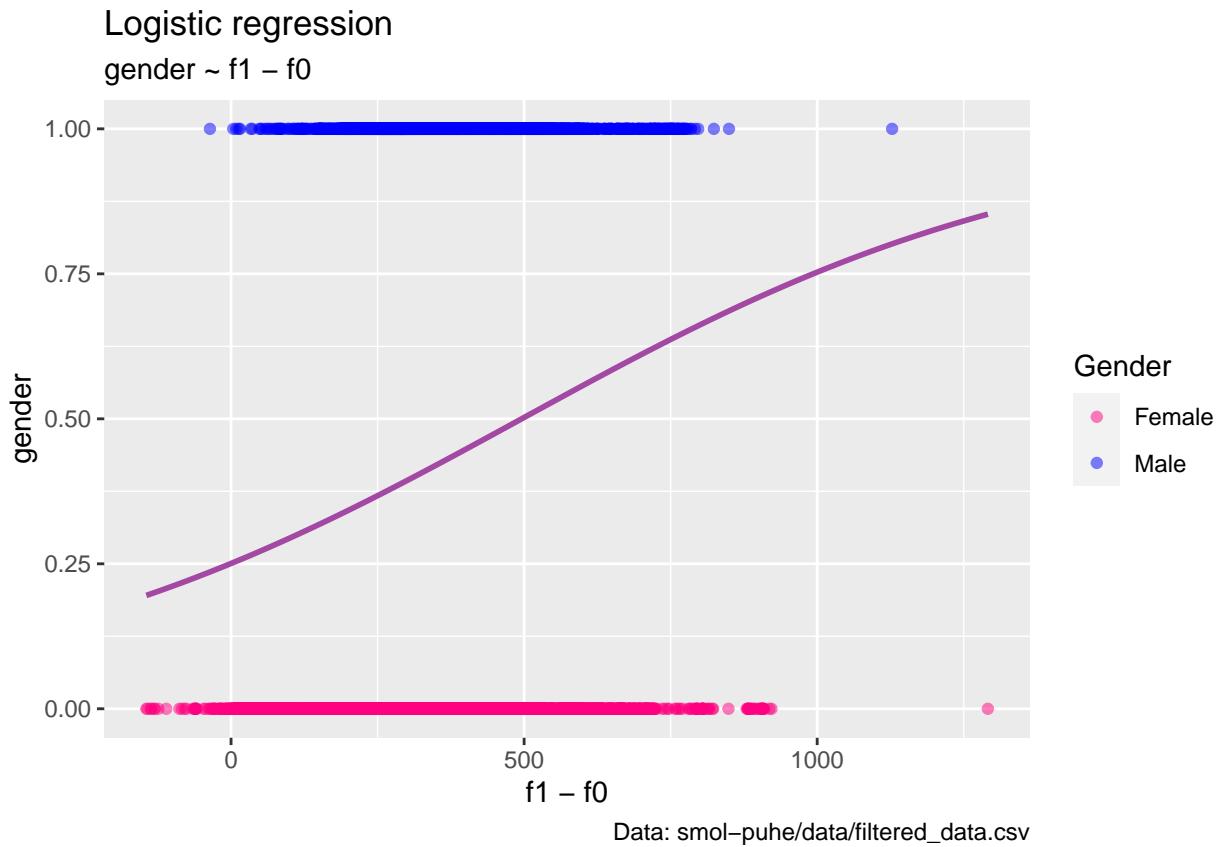
## 
## Call:
## glm(formula = gender ~ f1_over_f0, family = "binomial", data = data)
## 
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max 
## -3.0306  -0.8531  -0.5374   0.9956   2.2022 
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.86315   0.06974 -41.05 <2e-16 ***
## f1_over_f0    0.71667   0.01873   38.27 <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 11611  on 8563  degrees of freedom
## Residual deviance: 9539  on 8562  degrees of freedom
## AIC: 9543
## 
## Number of Fisher Scoring iterations: 4
## Logistic regression -----
## gender ~ f1 - f0

data$f1_minus_f0 = data$f1 - data$f0

ggplot(data, aes(x=f1_minus_f0, y=gender)) +
  geom_point(alpha=.5, aes(color = cut(data$gender, c(-1, 0.5, 1)))) +
  scale_color_manual(name = "Gender",
                     values = c("(-1,0.5]" = "#ff0080",
                               "(0.5,1]" = "#0000ff"),
                     labels = c("Female", "Male")) +
  stat_smooth(method="glm", se=FALSE,
              method.args = list(family=binomial),
              col="#a349a4") +
  labs(title = "Logistic regression",
       subtitle = "gender ~ f1 - f0",
       caption = "Data: smol-puhe/data/filtered_data.csv") +
  xlab("f1 - f0")

## `geom_smooth()` using formula 'y ~ x'

```



```
# ggsave("logistic_gender_f1_minus_f0.png")

model_f1_minus_f0 <- glm(gender~f1_minus_f0, data = data, family="binomial")
summary(model_f1_minus_f0)

##
## Call:
## glm(formula = gender ~ f1_minus_f0, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.9579   -1.0346   -0.8241    1.2815    1.6996
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.0956149  0.0510089 -21.48   <2e-16 ***
## f1_minus_f0  0.0022098  0.0001346   16.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 11611  on 8563  degrees of freedom
## Residual deviance: 11329  on 8562  degrees of freedom
## AIC: 11333
##
```

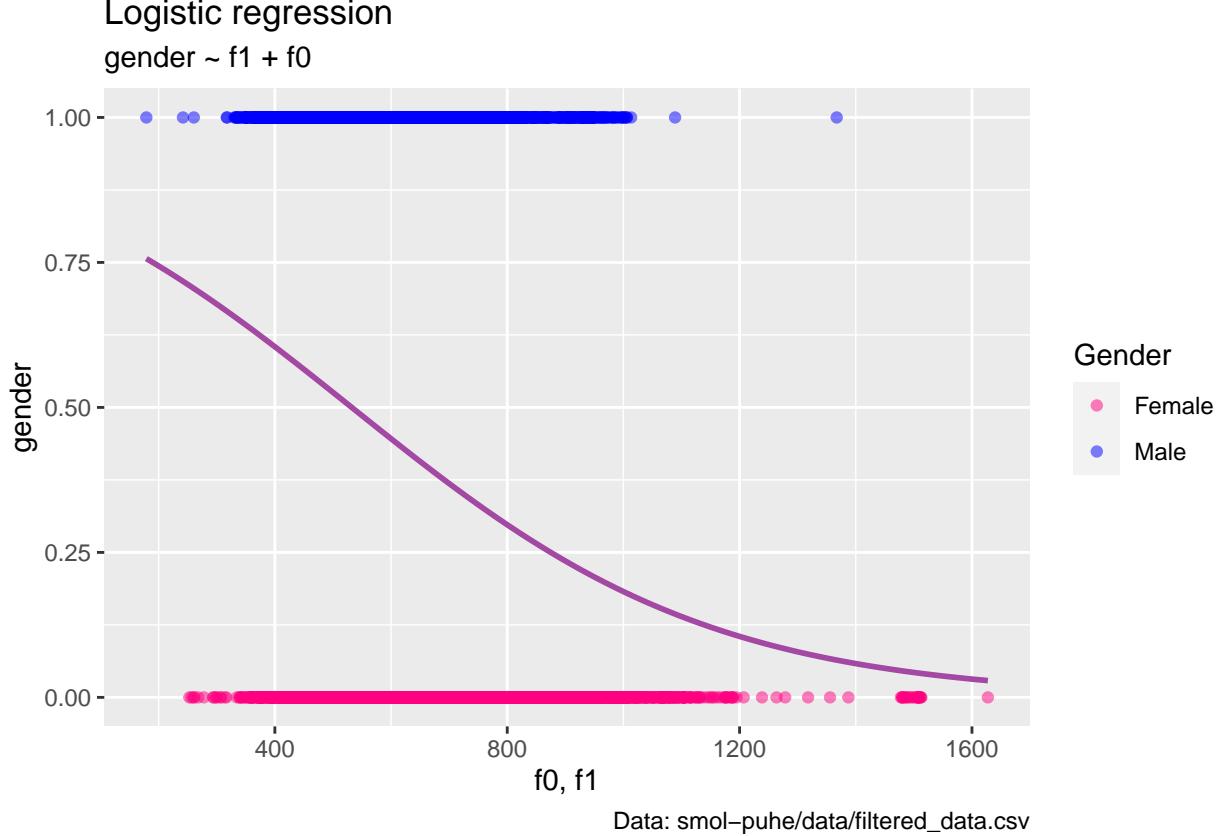
```

## Number of Fisher Scoring iterations: 4
## Logistic regression #####
## gender ~ f0 + f1

ggplot(data, aes(x=f0 + f1, y=gender)) +
  geom_point(alpha=.5, aes(color = cut(data$gender, c(-1, 0.5, 1)))) +
  scale_color_manual(name = "Gender",
                     values = c("(-1,0.5]" = "#ff0080",
                               "(0.5,1]" = "#0000ff"),
                     labels = c("Female", "Male")) +
  stat_smooth(method="glm", se=FALSE,
              method.args = list(family=binomial),
              col="#a349a4") +
  labs(title = "Logistic regression",
       subtitle = "gender ~ f1 + f0",
       caption = "Data: smol-puhe/data/filtered_data.csv") +
  xlab("f0, f1")

```

## `geom\_smooth()` using formula 'y ~ x'



```

# ggsave("logistic_gender_f1_minus_f0.png")

model <- glm(gender~f0 + f1 + f0*f1, data = data, family="binomial")
summary(model)

```

```

## 
## Call:

```

```

## glm(formula = gender ~ f0 + f1 + f0 * f1, family = "binomial",
##      data = data)
##
## Deviance Residuals:
##       Min      1Q  Median      3Q     Max
## -2.3819 -0.5314 -0.2197  0.4901  3.2730
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 9.773e+00 4.378e-01 22.321 < 2e-16 ***
## f0          -6.141e-02 2.757e-03 -22.274 < 2e-16 ***
## f1          -6.988e-03 7.830e-04 -8.924 < 2e-16 ***
## f0:f1       3.561e-05 5.029e-06  7.080 1.44e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 11611.0 on 8563 degrees of freedom
## Residual deviance: 6581.4 on 8560 degrees of freedom
## AIC: 6589.4
##
## Number of Fisher Scoring iterations: 5

```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.