Final Report

Andrés Romero

2023-01-25

Loading the Dataset

```
library(googlesheets4)
dataset <- read_sheet('https://docs.google.com/spreadsheets/d/1G5XRN7adCeNlLVsgSiH3Cu2_6h2LXzZarpfgcMiB
```

```
## ! Using an auto-discovered, cached token.

##   To suppress this message, modify your code or options to clearly consent to
##   the use of a cached token.

##   See gargle's "Non-interactive auth" vignette for more details:

##   <https://gargle.r-lib.org/articles/non-interactive-auth.html>

## i The googlesheets4 package is using a cached token for
##   'soyandresromero@gmail.com'.

## v Reading from "Survey Results".

## v Range ''Final''.
```

```
dataset
```

```
## # A tibble: 6,440 x 27
##    Submission          Ich bestätige, dass ich die oben~1 Ich erkläre mich dam~2
##    <dttm>              <chr>                              <chr>
##  1 2023-01-09 11:09:45 Ja                                 Ja
##  2 2023-01-09 11:09:45 Ja                                 Ja
##  3 2023-01-09 11:09:45 Ja                                 Ja
##  4 2023-01-09 11:09:45 Ja                                 Ja
##  5 2023-01-09 11:09:45 Ja                                 Ja
##  6 2023-01-09 11:09:45 Ja                                 Ja
##  7 2023-01-09 11:09:45 Ja                                 Ja
##  8 2023-01-09 11:09:45 Ja                                 Ja
##  9 2023-01-09 11:09:45 Ja                                 Ja
## 10 2023-01-09 11:09:45 Ja                                 Ja
## # i 6,430 more rows
## # i abbreviated names:
```

```
## #   1: 'Ich bestätige, dass ich die oben genannten Informationen für Teilnehmer gelesen und versta..
## #   2: 'Ich erkläre mich damit einverstanden, an einem sprachwissenschaftlichen Forschungsprojekt ..
## # i 24 more variables:
## #   'Mir ist bekannt, dass ich meine Daten jederzeit und ohne Angabe von Gründen zurückziehen k...' <
## #   'Ich bin damit einverstanden, dass meine Antworten zu Forschungszwecken verwendet werden dü...' <
```

```
df <- dataset[which(!is.na(dataset$Alter)),]
df <- df[which(df$Muttersprache=="Deutsch"),]
cols <- c('Name','Alter','Geschlecht','Geburtsort','Höchster Bildungsabschluss','Fach','Beruf','Mutters
df <- df[cols]
df['nlang'] <- 4 - is.na(df$`Sprache 1`) - is.na(df$`Sprache 2`) - is.na(df$`Sprache 3`) - is.na(df$`Sp
df$Charaktereigenschaften <- as.factor(df$Charaktereigenschaften)
df$Geschlecht <- as.factor(df$Geschlecht)
df$Muttersprache <- as.factor(df$Muttersprache)
df$`Höchster Bildungsabschluss` <- factor(df$`Höchster Bildungsabschluss`,ordered = TRUE,
                                          levels = c("Unterer Schulabschluss", "Abitur oder gleichwertig
df['German_Pref'] <- (df$`Besten Aufnahme`=="Sprecher/in 1" | df$`Besten Aufnahme`=="Sprecher/in 2")
df$German_Pref_01 <- as.numeric(as.factor(df$German_Pref))-1
df$`Besten Aufnahme` <- as.factor(df$`Besten Aufnahme`)
df
```

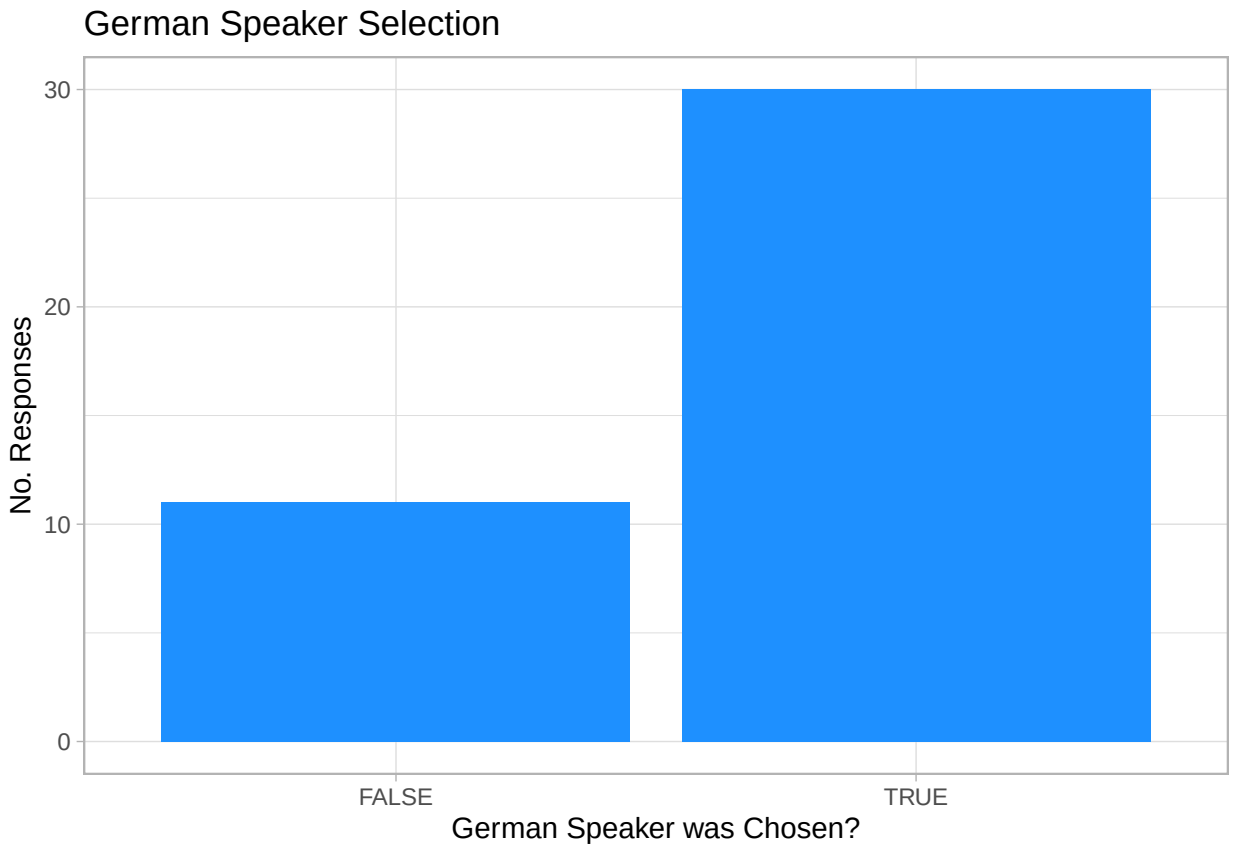
```
## # A tibble: 2,460 x 23
##    Name  Alter Geschlecht Geburtsort  'Höchster Bildungsabschluss'   Fach  Beruf
##    <chr> <dbl> <fct>      <chr>       <ord>                          <chr> <chr>
##  1 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  2 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  3 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  4 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  5 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  6 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  7 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  8 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
##  9 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
## 10 Jan T    21 männlich   Deutschland Abitur oder gleichwertiger Ab~ <NA>  <NA>
## # i 2,450 more rows
## # i 16 more variables: Muttersprache <fct>, 'Besten Aufnahme' <fct>,
## #   Charaktereigenschaften <fct>, Aufnahme <dbl>, Score <dbl>,
## #   'Sprache 1' <chr>, 'Level 1' <chr>, 'Sprache 2' <chr>, 'Level 2' <chr>,
## #   'Sprache 3' <chr>, 'Level 3' <chr>, 'Sprache 4' <chr>, 'Level 4' <chr>,
## #   nlang <dbl>, German_Pref <lgl>, German_Pref_01 <dbl>
```

```
cols_r <- c('Name','Alter','Geschlecht','Geburtsort','Höchster Bildungsabschluss','Fach','Beruf','Mutte
df_reduced <- df[cols_r]
df_reduced <- df_reduced[!duplicated(df_reduced$Name),]
df_reduced
```

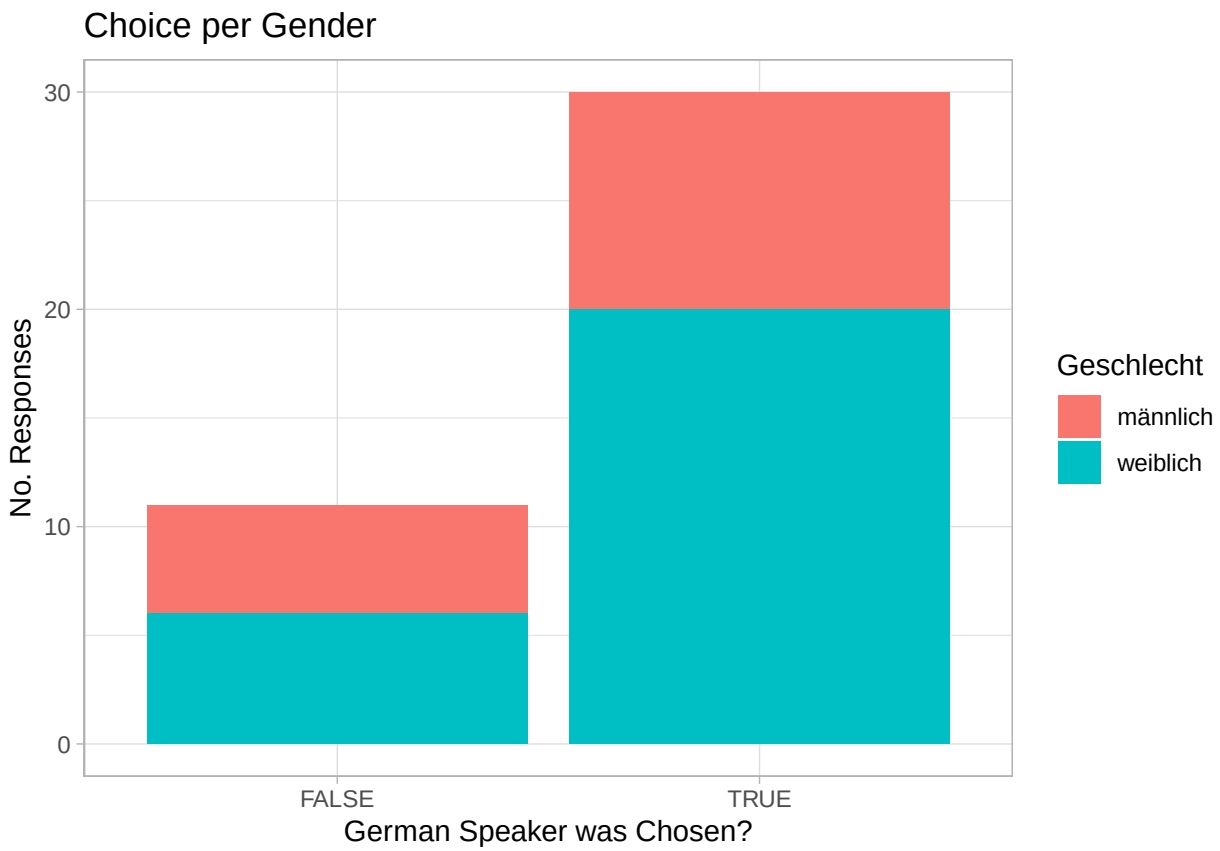
```
## # A tibble: 41 x 12
##    Name      Alter Geschlecht Geburtsort  Höchster Bildungsabschlu~1 Fach  Beruf
##    <chr>     <dbl> <fct>      <chr>       <ord>                      <chr> <chr>
##  1 Jan T        21 männlich   Deutschland Abitur oder gleichwertige~ <NA>  <NA>
##  2 Maya R.      22 weiblich   Deutschland Bachelor                   Biol~ <NA>
##  3 Denise N.    25 weiblich   <NA>        Abitur oder gleichwertige~ <NA>  Cust~
##  4 Omisha B     21 weiblich   Deutschland Abitur oder gleichwertige~ <NA>  <NA>
```

```
##  5 Gesa T.      25 weiblich   Deutschland Master-Abschluss           Nach~ Stel~
##  6 Pia R.       24 weiblich   Deutschland Abitur oder gleichwertige~ <NA>  <NA>
##  7 Alicia H.    22 weiblich   <NA>        Abitur oder gleichwertige~ <NA>  <NA>
##  8 Julia K.     23 weiblich   Deutschland Abitur oder gleichwertige~ <NA>  <NA>
##  9 Anna E.      23 weiblich   <NA>        Bachelor                   Land~ <NA>
## 10 Lukas V.     23 männlich   Deutschland Abitur oder gleichwertige~ <NA>  Stud~
## # i 31 more rows
## # i abbreviated name: 1: 'Höchster Bildungsabschluss'
## # i 5 more variables: Muttersprache <fct>, 'Besten Aufnahme' <fct>,
## #   nlang <dbl>, German_Pref <lgl>, German_Pref_01 <dbl>
```

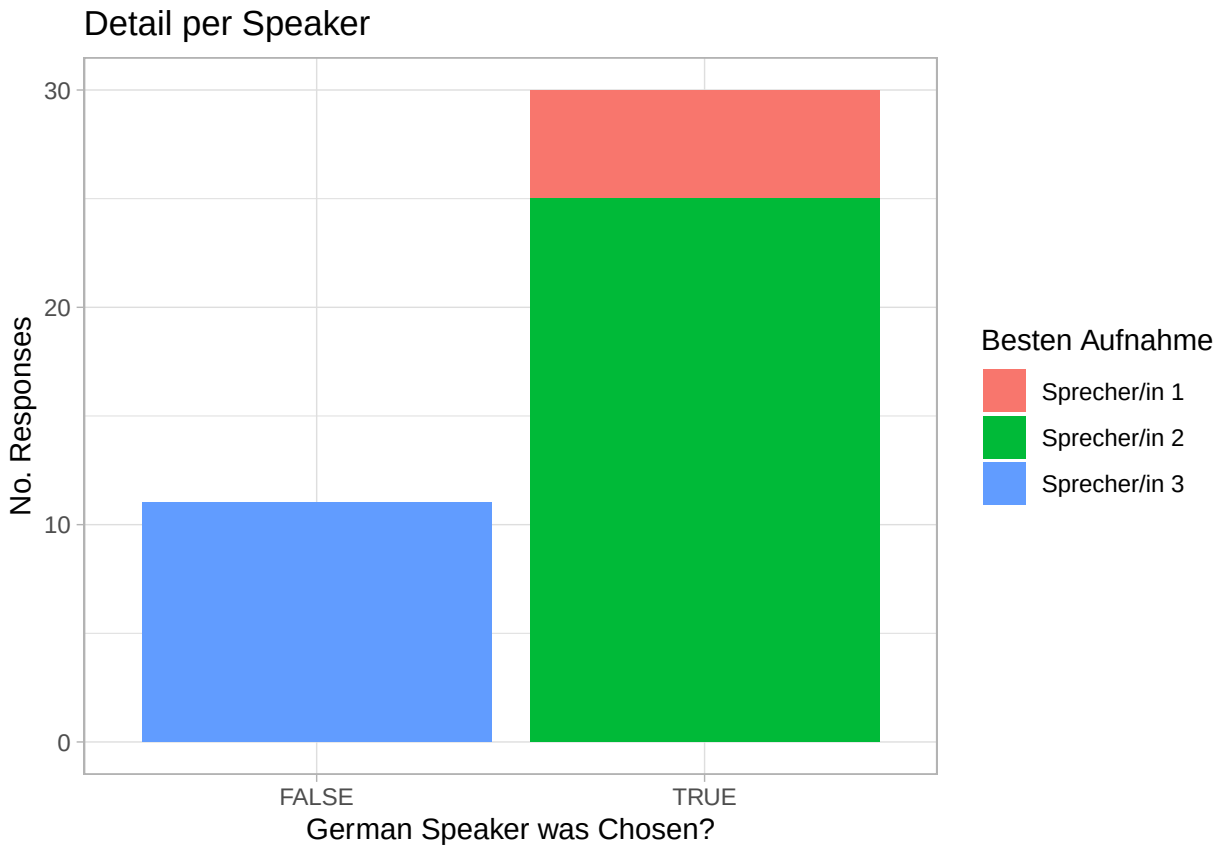
```r
library(ggplot2)
ggplot(df_reduced, aes(x=German_Pref))+
    geom_bar(fill="dodgerblue")+
    labs(title="German Speaker Selection", x = "German Speaker was Chosen?", y = "No. Responses")+
    theme_light()
```



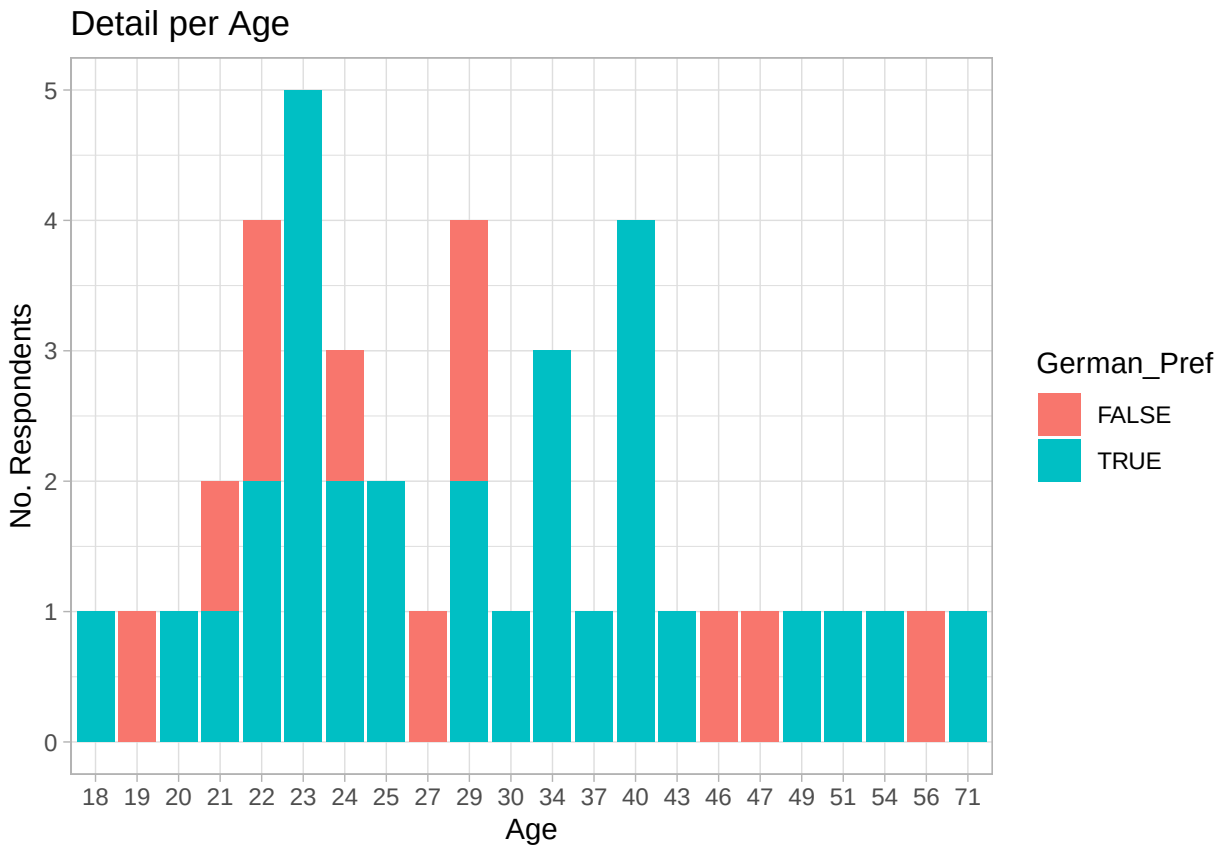

```r
library(ggplot2)
ggplot(df_reduced, aes(x=German_Pref, fill=Geschlecht))+
    geom_bar()+
    labs(title="Choice per Gender", x = "German Speaker was Chosen?", y = "No. Responses")+
    theme_light()
```



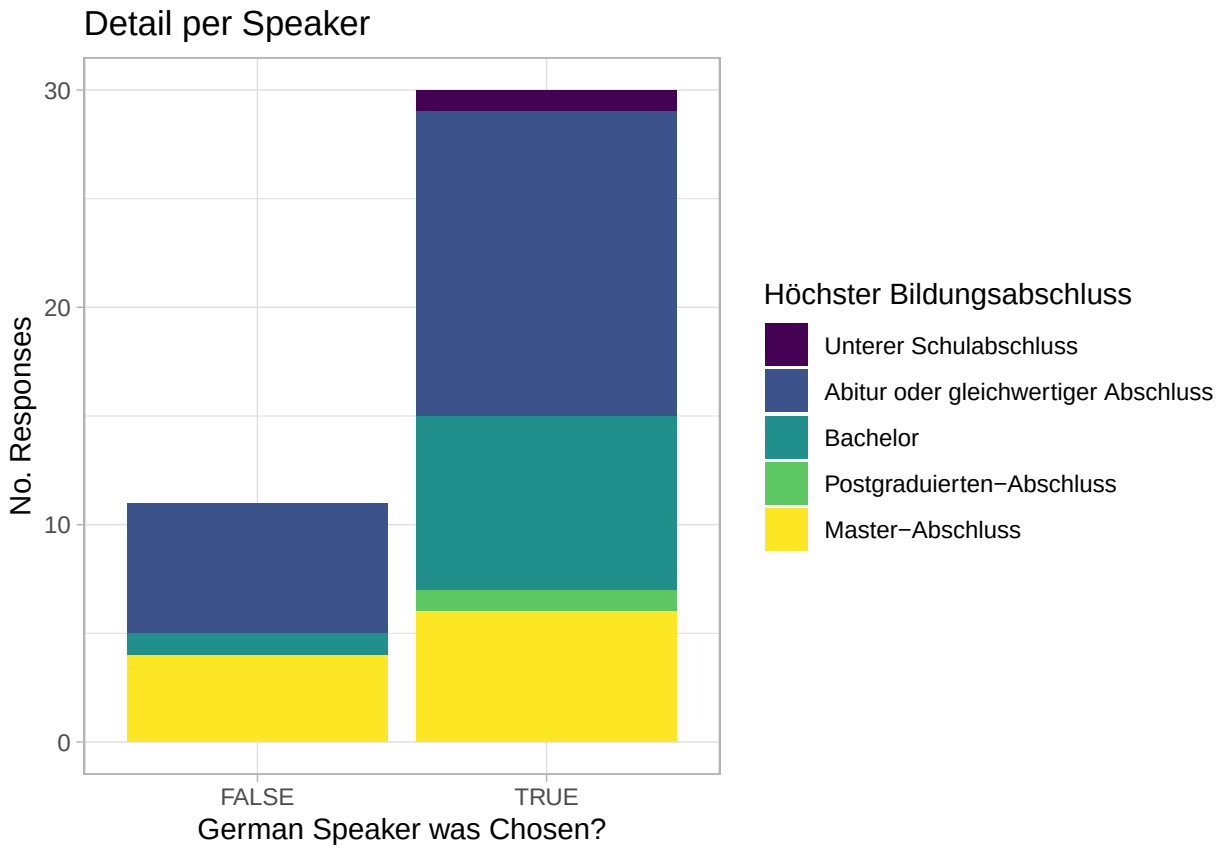

```
ggplot(df_reduced, aes(x=German_Pref, fill=`Besten Aufnahme`))+
    geom_bar()+
    labs(title="Detail per Speaker", x = "German Speaker was Chosen?", y = "No. Responses")+
    theme_light()
```




```
ggplot(df_reduced, aes(x=as.factor(Alter), fill = German_Pref))+
    geom_bar()+
    labs(title="Detail per Age", x = "Age", y = "No. Respondents")+
    theme_light()
```




```
ggplot(df_reduced, aes(x=German_Pref, fill=`Höchster Bildungsabschluss`))+
    geom_bar()+
    labs(title="Detail per Speaker", x = "German Speaker was Chosen?", y = "No. Responses")+
    theme_light()
```



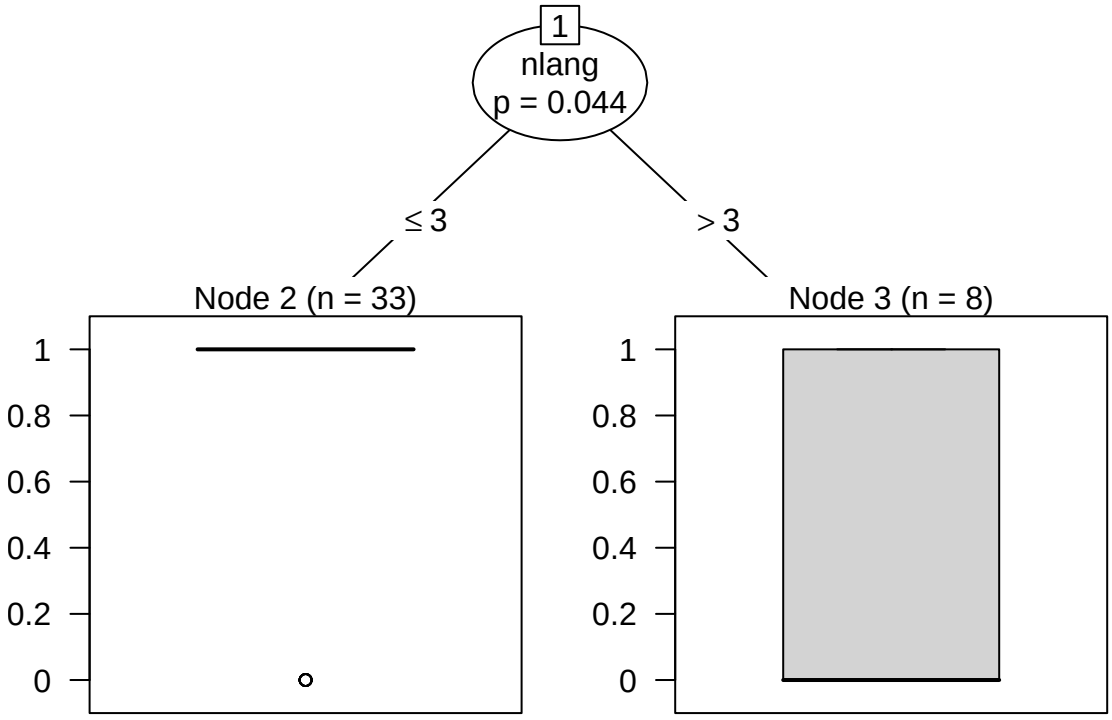

```
library(party)
```

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
tree <- ctree(`German_Pref` ~  Alter + nlang + Geschlecht + `Höchster Bildungsabschluss`, data = df_redu
tree
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  German_Pref
## Inputs:  Alter, nlang, Geschlecht, Höchster Bildungsabschluss
## Number of observations:  41
##
## 1) nlang <= 3; criterion = 0.956, statistic = 6.433
##   2)*  weights = 33
## 1) nlang > 3
##   3)*  weights = 8
```

```
plot(tree)
```

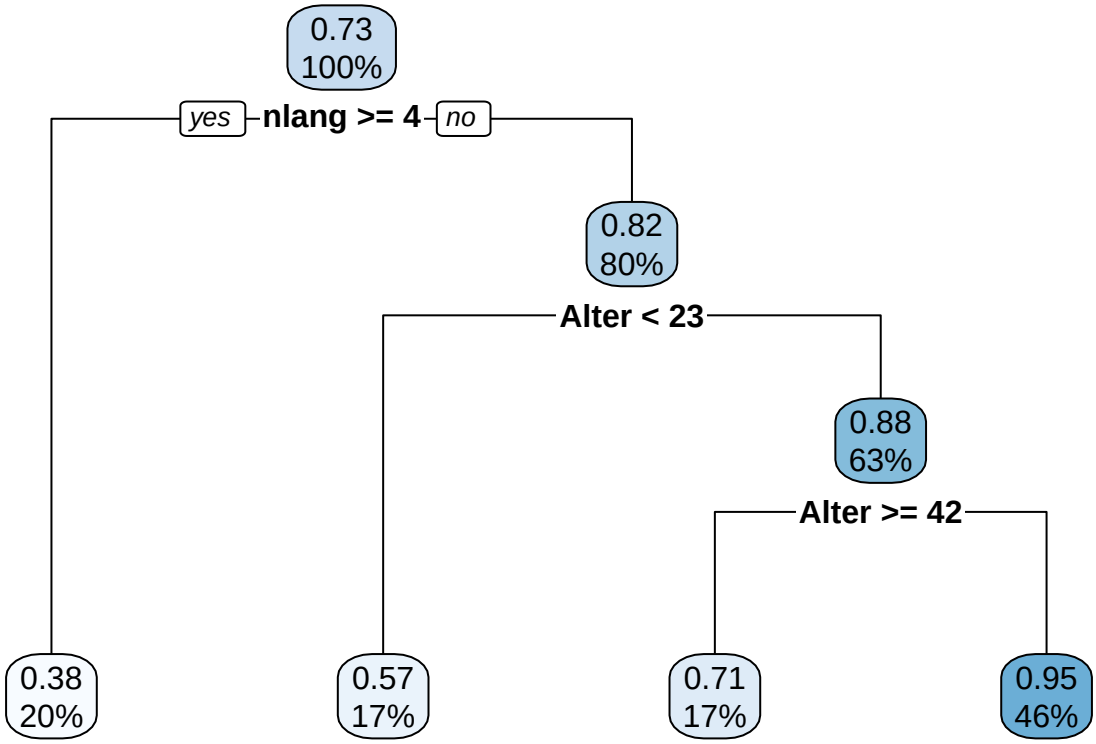



```
library(rpart)
library(rpart.plot)
tree <- rpart(`German_Pref` ~  Alter + nlang + Geschlecht + `Höchster Bildungsabschluss`, data = df_redu
tree
```

```
## n= 41
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 41 8.0487800 0.7317073
##    2) nlang>=3.5 8 1.8750000 0.3750000 *
```

```
##    3) nlang< 3.5 33 4.9090910 0.8181818
##      6) Alter< 22.5 7 1.7142860 0.5714286 *
##      7) Alter>=22.5 26 2.6538460 0.8846154
##       14) Alter>=41.5 7 1.4285710 0.7142857 *
##       15) Alter< 41.5 19 0.9473684 0.9473684 *
```

```
rpart.plot(tree)
```




```
printcp(tree)
```

```
##
## Regression tree:
## rpart(formula = German_Pref ~ Alter + nlang + Geschlecht + `Höchster Bildungsabschluss`,
##     data = df_reduced)
##
## Variables actually used in tree construction:
## [1] Alter nlang
##
## Root node error: 8.0488/41 = 0.19631
##
## n= 41
##
##         CP nsplit rel error xerror    xstd
## 1 0.157128      0   1.00000 1.0854 0.17842
## 2 0.067210      1   0.84287 1.1522 0.22270
## 3 0.034528      2   0.77566 1.1420 0.23760
## 4 0.010000      3   0.74113 1.1712 0.24404
```

```r
an <- anova(glm(`German_Pref` ~  Alter + nlang + Geschlecht + `Höchster Bildungsabschluss`-1, data=df_r
an
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: German_Pref
##
## Terms added sequentially (first to last)
##
##
##                               Df Deviance Resid. Df Resid. Dev
## NULL                                             41    30.0000
## Alter                          1  19.6323        40    10.3677
## nlang                          1   0.0149        39    10.3528
## Geschlecht                     2   3.8390        37     6.5138
## `Höchster Bildungsabschluss`  4   0.4732        33     6.0405
```