**INITIAL OBSERVATIONS YEAR**

## ###PACKAGES USED##

install.packages("ggplot2")

install.packages("esquisse")

install.packages("shiny")

library("esquisse")

library("ggplot2")

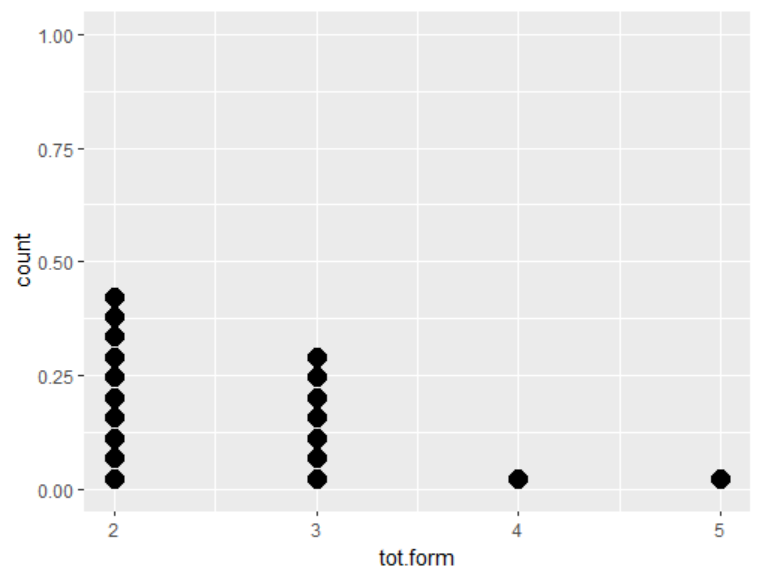
library("shiny")

# GENITIVE PL

We started by creating a plot for the occurrences.

ggplot(data=YrGenPlu.wide, aes(x=tot.form)) +geom\_dotplot()

We want to exclude occurrences lower than 2, so Chat GPT proposed these two codes for subsetting the data



### #CODE

ggplot(data = subset(YrGenPlu.wide, tot.form >= 2), aes(x = tot.form)) +

geom\_dotplot()

*In this code, we first use the subset() function to create a new data frame that only includes rows where the tot.form variable is greater than or equal to 2. We then pass this new data frame to ggplot() and specify tot.form as the x-axis variable, and geom\_dotplot() as the layer for the dot plot.*

*Note that you may need to adjust the x-axis limits depending on the range of values in your data after subsetting. You can do this by adding scale\_x\_continuous() to your ggplot code, like so:*

ggplot(data = subset(YrGenPlu.wide, tot.form >= 2), aes(x = tot.form)) +

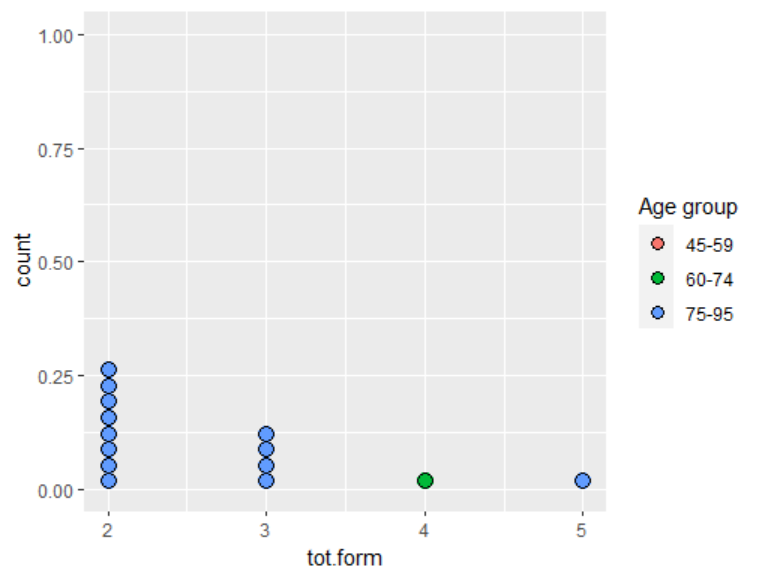
geom\_dotplot() +

scale\_x\_continuous(limits = c(2, max(YrGenPlu.wide$tot.form)))

Now that we have a subset of the plots, I wanted to start exploring my main dependant variables: age, gender and village of origin.

## 1.1 Age group

There are no speakers younger than 60 producing more than 2 forms, so the results may not be very conclusive. Or else, the older the people they get, the more variation we can expect?



### #CODE

I had some issues with the code, because of the name of the dependant, but this one should work.

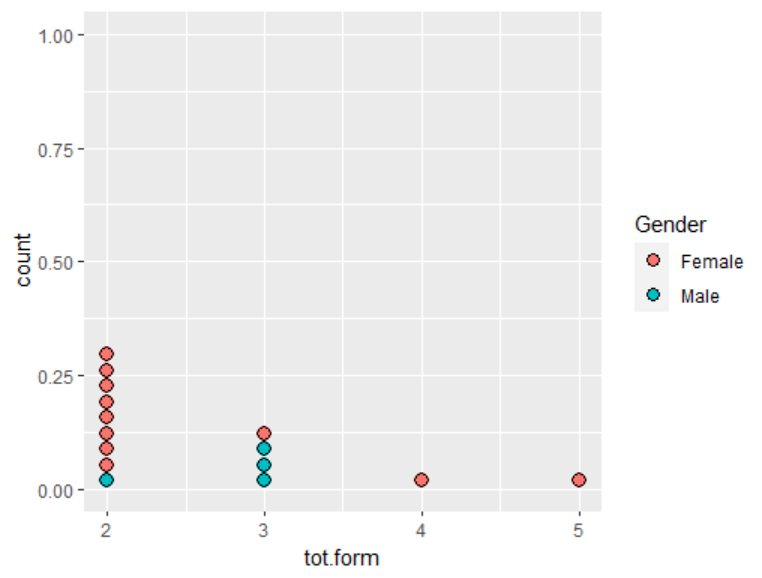
ggplot(data = subset(YrGenPlu.wide, tot.form >= 2), aes(x = tot.form, fill = `Age group`)) +

geom\_dotplot() +

scale\_x\_continuous(limits = c(2, max(YrGenPlu.wide$tot.form)))

## 1.2 Gender

Concerning gender, if we didn’t know anything else, it would seem that women produce more forms than men. The reality is that there are a lot more women on the subset, so I would be very careful with the conclusions.



### #CODE

The code for this one looks almost the same as the previous one.

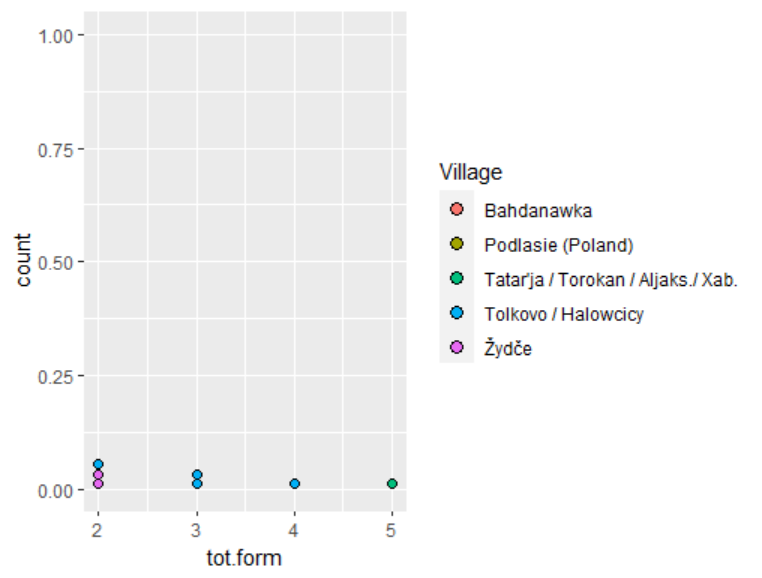
ggplot(data = subset(YrGenPlu.wide, tot.form >= 2), aes(x = tot.form, fill = `Gender`)) +

geom\_dotplot() +

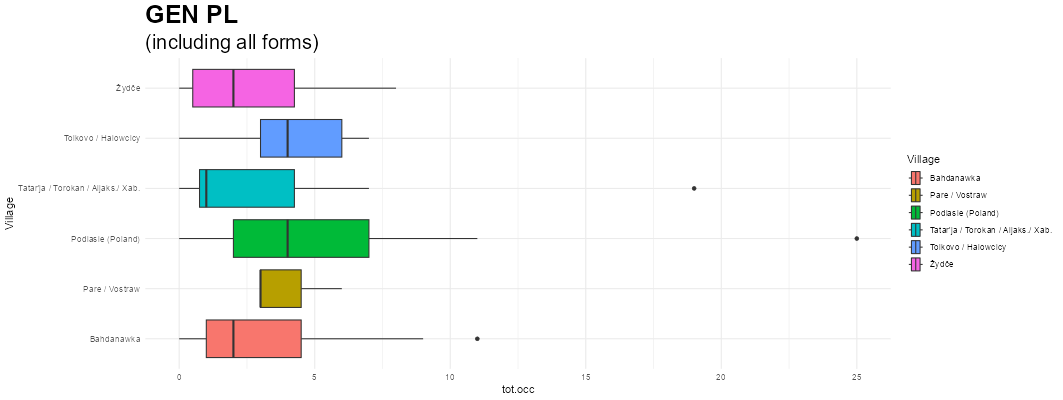
scale\_x\_continuous(limits = c(2, max(YrGenPlu.wide$tot.form)))

## 1.3. Village of origin

It seems that one village group (Tatar’ja/Torokan) is particularly prone to overabundance, but I would be very careful. For some reason, I can’t see data from Podlasie (and I could easily imagine people using more than one GEN PL). So I have tried running a boxplot, but including the subset we had initially excluded, so there, data from Podlasie are not hidden.



Boxplot with ALL the forms



### #CODE

ggplot(data = subset(YrGenPlu.wide, tot.form >= 2), aes(x = tot.form, fill = `Village`)) +

geom\_dotplot() +

scale\_x\_continuous(limits = c(2, max(YrGenPlu.wide$tot.form)))

### #BOXPLOT

ggplot(YrGenPlu.wide) +

aes(x = tot.occ, y = Village, fill = Village) +

geom\_boxplot() +

scale\_fill\_hue(direction = 1) +

labs(title = "GEN PL", subtitle = "(including all forms)") +

theme\_minimal() +

theme(

plot.title = element\_text(size = 25L,

face = "bold"),

plot.subtitle = element\_text(size = 20L)

)

# ⚠ TASKS TO DO FROM NOW ON

## STEP #1 GRPH

Create a plot for each number/case combinations.

First create a table of frequencies based on a +2 occurrence subset.

### #these create the data frames to draw the dot plot

genplu.gph1<- subset(YrGenPlu.wide, tot.occ >= 2 )

genplu.gph<-ddply(genplu.gph1, .variable=c("tot.occ","tot.form"),nrow)

ggplot(data = genplu.gph, aes(x = tot.occ, y = tot.form, size = V1)) +

geom\_point()

Afterwards, create a graph of scattered plots where size is adjusted for frequency. We want to answer two questions:  
 a) How many people have used more than 2 occurrences for each parameter?

b) Which percentage of the +2 occurrences (i.e. use) use +1 form?

## STEP #2

We will be using multilevel logistic regressions (instead of Poisson) to see which form is used in each dependent variable.

For that, we will use the “long” datasets, but given that the presence seems a bit problematic, we will first create a subset which excludes all the false values for presence. [warning sign!] ⚠ IMPORTANT, save the subset as a new data frame!

Then, we will go to the “Esquisse” package and using “esquisser” we open the modified data frames and use the following values for the arguments:

x= form; facet= “Village”

We will replace the facet argument depending on the dependent we want to study.