PAD-ATA Call Log Analysis

#### Libraries used in the Project and uploading data

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(stringr)  
library(svMisc)

##   
## Attaching package: 'svMisc'

## The following object is masked from 'package:utils':  
##   
## ?

library(ggplot2)  
library(gtable)  
library("ggpubr")  
  
log\_data <- read\_csv("~/Desktop/PAE/github/data/logData.csv")

##   
## ── Column specification ───────────────────────────────────────────────────────────────────────────────────────────────────  
## cols(  
## callerId = col\_double(),  
## langId = col\_double(),  
## callTime = col\_datetime(format = ""),  
## lastCallTime = col\_character(),  
## noCallsMade = col\_double(),  
## noContentListened = col\_double(),  
## callId = col\_character(),  
## eventTime = col\_datetime(format = ""),  
## logInfo = col\_character(),  
## logInfoId = col\_double(),  
## inSurvey = col\_logical()  
## )

#MAKE IT ALL UPPER!!!  
  
log\_data$logInfo <- toupper(log\_data$logInfo)  
  
### GETTING THE DATE MONTH AND YEAR IF I WANT!!!  
extractdate <- function(date) {  
 day <- format(date, format="%d")  
 month <- format(date, format="%m")  
 year <- format(date, format="%Y")  
  
 cbind(day, month, year)  
}  
  
#making first calls  
first\_call<-extractdate(log\_data$callTime)  
  
  
#making last call  
last\_day <-substr(log\_data$lastCallTime, 9, 10)  
last\_month <- substr(log\_data$lastCallTime, 6, 7)  
last\_year <-substr(log\_data$lastCallTime, 1, 4)  
  
log\_data<-cbind(log\_data, first\_call,last\_day,last\_month,last\_year)  
  
#Converting columns to numeric  
log\_data$day <-as.numeric(log\_data$day)  
log\_data$last\_day <- as.numeric(log\_data$last\_day)

## Warning: NAs introduced by coercion

log\_data$month <-as.numeric(log\_data$month)  
log\_data$last\_month <- as.numeric(log\_data$last\_month)

## Warning: NAs introduced by coercion

log\_data$year <- as.numeric(log\_data$year)  
log\_data$last\_year <- as.numeric(log\_data$last\_year )

## Warning: NAs introduced by coercion

log\_data$year <- log\_data$year +log\_data$month/12  
log\_data$last\_year <- log\_data$last\_year + log\_data$last\_month/12  
  
log\_data$caller\_lifetime <- log\_data$last\_year - log\_data$year

*Cmd+Option+I*.

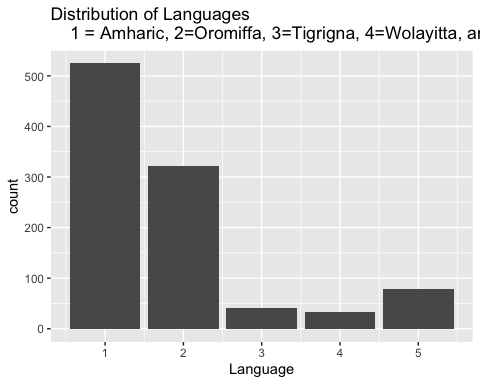
*Cmd+Shift+K*

### PRELIMINARY DISTRIBUTIONS

**In this section, we display some of the basic distributions of the data including languages, total lifeetime of a given caller, colls made, content accessed, and the ratio of content listened to calls made.**

**Based on the graphs below, we are nervous that some of the data is not random given the incredibly high ratio for content listened to calls made, but given that the data is distributed fairly evenly across thee different graphs, we are unsure**

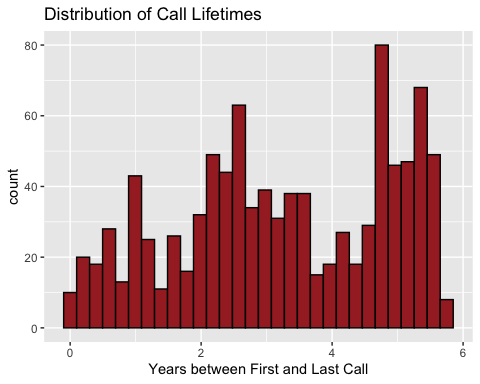
lang\_spoken <- log\_data %>%  
 group\_by(callerId, langId) %>%  
 count()  
  
#Ditribution of languages spoken  
ggplot(data=lang\_spoken, aes(x=langId)) +  
 geom\_bar()+  
 labs(  
 title = 'Distribution of Languages  
 1 = Amharic, 2=Oromiffa, 3=Tigrigna, 4=Wolayitta, and 5=Sidamigna',  
 x = 'Language',  
 y= 'count'  
 )



#Ditribution otiem difference beetween firs tand last call  
  
lifetime<- log\_data %>%  
 group\_by(callerId, caller\_lifetime) %>%  
 count()  
  
ggplot(data=lifetime, aes(x=caller\_lifetime)) +  
 geom\_histogram(fill='brown', color='black')+  
 labs(  
 title = 'Distribution of Call Lifetimes',  
 x = 'Years between First and Last Call',  
 y= 'count'  
 )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 17 rows containing non-finite values (stat\_bin).

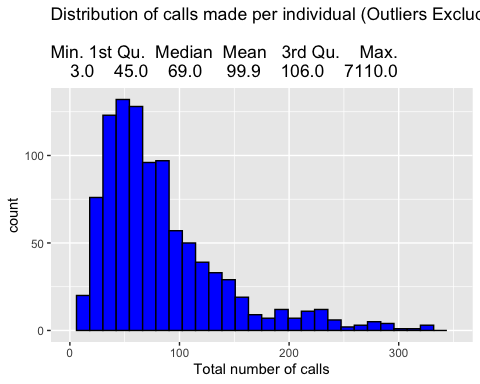


#distribution of calls made  
  
calls\_made <- log\_data %>%  
 group\_by(callerId, noCallsMade) %>%  
 count()  
  
ggplot(calls\_made, aes(x=noCallsMade)) +  
 geom\_histogram(color="black", fill="blue") +  
 xlim(0,350) +  
 labs(  
 title = 'Distribution of calls made per individual (Outliers Excluded) \n  
Min. 1st Qu. Median Mean 3rd Qu. Max.   
 3.0 45.0 69.0 99.9 106.0 7110.0 ',  
 x = 'Total number of calls'  
 )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 17 rows containing non-finite values (stat\_bin).

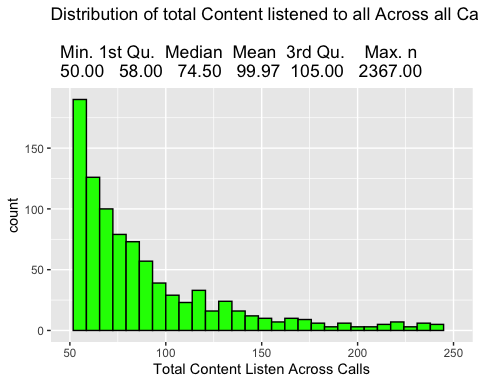
## Warning: Removed 2 rows containing missing values (geom\_bar).



#Distribution of contenet listened to  
  
content\_listened <- log\_data %>%  
 group\_by(callerId, noContentListened) %>%  
 count()  
  
#summary(content\_listened$noContentListened)  
  
ggplot(content\_listened, aes(x=noContentListened)) +  
 geom\_histogram(color="black", fill="green") +  
 xlim(50, 250) +  
 labs(  
 title = 'Distribution of total Content listened to all Across all Calls (Outliers Excluded) \n  
 Min. 1st Qu. Median Mean 3rd Qu. Max. n\  
 50.00 58.00 74.50 99.97 105.00 2367.00 ',  
 x = 'Total Content Listen Across Calls'  
 )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

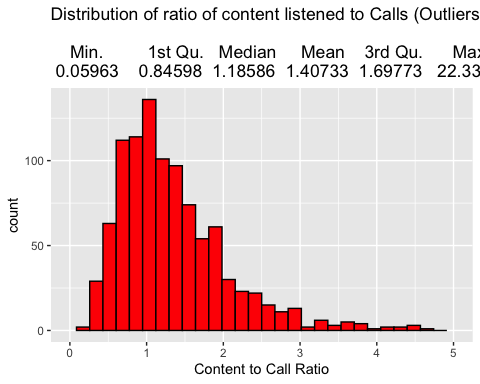
## Warning: Removed 37 rows containing non-finite values (stat\_bin).  
  
## Warning: Removed 2 rows containing missing values (geom\_bar).



# Ratio of content per call - creating the ratio and graphing the results  
content\_call\_ratio\_df <- inner\_join(calls\_made,content\_listened,by="callerId")  
content\_call\_ratio\_df$ratio <- content\_call\_ratio\_df$noContentListened/content\_call\_ratio\_df$noCallsMade  
  
#summary(content\_call\_ratio\_df$ratio)  
  
ggplot(content\_call\_ratio\_df, aes(x=ratio)) +  
 geom\_histogram(color="black", fill="red") +  
 xlim(0, 5) +  
 labs(  
 title = 'Distribution of ratio of content listened to Calls (Outliers Excluded) \n  
 Min. 1st Qu. Median Mean 3rd Qu. Max.  
 0.05963 0.84598 1.18586 1.40733 1.69773 22.33333 ',  
 x = 'Content to Call Ratio'  
 )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 11 rows containing non-finite values (stat\_bin).  
  
## Warning: Removed 2 rows containing missing values (geom\_bar).



### TOP MENU

**Next, we want to Breakd down the initial Menu options and delve into the distributions of the selections that the farmer can make at a different state in the IVR System**

**For the first analysis, we looked at the distribution of the TOP MENU in the data. This comes in as follows**:

##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> ASSINGED TO EXPERIMENT (OPTIONAL) –> TOP MENU

**TOP MENU**

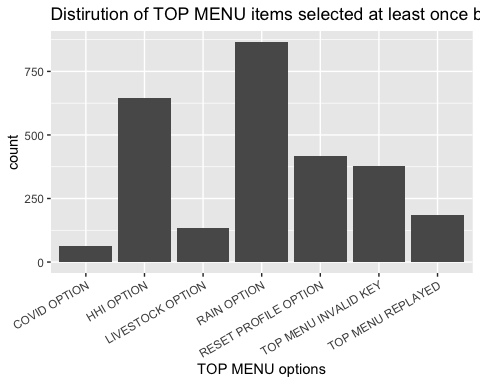
* RAIN OPTION
* HHI OPTION (household Irrigation)
* RESET PROFILE OPTION
* TOP MENU REPLAY
* LIVESTOCK OPTION
* COVID OPTION

*Please note, we are unsure of the order these are presented. For other menus, we do have the order*

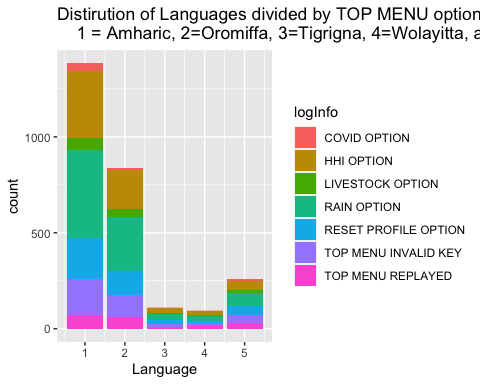
**Below, we have visualized the number of farmers, out of our sample, that in the life time of their calling history have accessed a given TOP MENU item at least once**

**Additionally, we also visualized this same selection divided by language to see if the patterns of seleciont are similar or different across langauges**v

# filtering out top Menu keys  
  
top\_menu <-filter(log\_data, grepl("TOP MENU",logInfo))  
  
#cleaning the axis  
top\_menu$logInfo<- str\_replace(top\_menu$logInfo, "TOP MENU - ", "")  
top\_menu$logInfo<- str\_replace(top\_menu$logInfo, " PRESSED", "")  
top\_menu$logInfo<- str\_replace(top\_menu$logInfo, " PRESSED", "")  
top\_menu$logInfo<- str\_replace(top\_menu$logInfo, " SELECTED", "")  
  
top\_menu <- top\_menu %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
  
ggplot(top\_menu, aes(logInfo)) +  
 theme(axis.text.x = element\_text(angle = 30, hjust =1))+  
 geom\_bar()+  
 labs(  
 title = 'Distirution of TOP MENU items selected at least once by a unique caller',  
 x = 'TOP MENU options'  
 )



ggplot(top\_menu, aes(langId, fill = logInfo)) +  
 geom\_bar()+  
 labs(  
 title = 'Distirution of Languages divided by TOP MENU option  
 1 = Amharic, 2=Oromiffa, 3=Tigrigna, 4=Wolayitta, and 5=Sidamigna',  
 x = 'Language'  
 )



**As we can easily see, the RAIN options and the HHI option dominate the choices from the top Menu. So, we decided to dive deeper into each of these menus to see if there are any outlier selections that are deep within the system. Additionally, ths distribution across languages seems to be fairly consistent at face value. We can run tests in the future, but for EDA this will be adequate. For 3/4 langauge options - having more data would be apprecaited**

### RAIN MENU

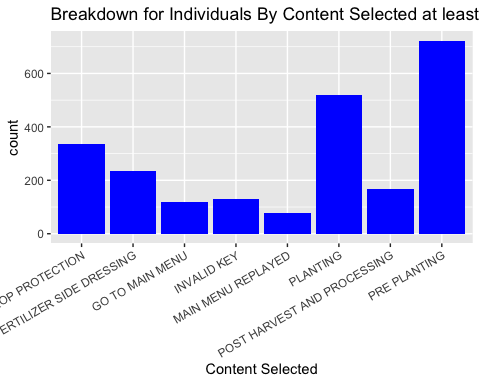
##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> RANDOM EXPERIMENT –> TOP MENU –> RAIN MENU

**Here, we look at the breakdown of individuals who have selected a given option in the rain menu at least once. In the data, MAIN MENU == RAIN MENU.**

**RAIN MENU**

* PRE PLANTING OPTION
* PLANTING OPTION
* CROP PROTECTION OPTION
* FERTILIZGER SIDE DREESSING
* POST HARVEST AND PROCESSING
* REPEAT RAIN MENU
* RETURN TO TOP MENU
* INVALID KEY PRESSED (ADDITIONAL OUTCOME)

# Pulling out the rain Menu  
  
#RAIN == HHI\_RAIN  
rain\_menu <- filter(log\_data, grepl("MAIN MENU - ",logInfo))  
rain\_menu <- filter(rain\_menu, !grepl("LIVESTOCK MAIN MENU",logInfo))  
rain\_menu <- filter(rain\_menu, !grepl("HHI MAIN MENU",logInfo))  
rain\_menu <- filter(rain\_menu, !grepl("COVID-19 MAIN MENU",logInfo))  
  
#grouping the observations  
rain\_menu <-rain\_menu %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
rain\_menu$logInfo<- str\_replace(rain\_menu$logInfo, "MAIN MENU - ", "")  
rain\_menu$logInfo<- str\_replace(rain\_menu$logInfo, " SELECTED", "")  
rain\_menu$logInfo<- str\_replace(rain\_menu$logInfo, " PRESSED", "")  
rain\_menu$logInfo<- str\_replace(rain\_menu$logInfo, " OPTION", "")  
  
#plotting rain menu breakdown  
ggplot(data=rain\_menu, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 30, hjust =1))+  
 geom\_bar(fill='blue')+  
 labs(  
 title = 'Breakdown for Individuals By Content Selected at least Once from the rain menu',  
 x = 'Content Selected',  
 y= 'count'  
 )



**We can see that the pre-planting menu, which still has over 75% of unique users accesss it at least once expereinces incredibly high usage. So, we now dive into this menu. We will breakdown conent selection by language and overall**

### RAIN MENU/PRE PLANTING MENU

##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> RANDOM EXPERIMENT –>

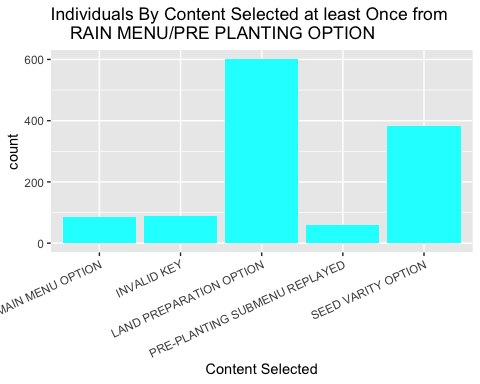
##### TOP MENU –> RAIN MENU –> PRE PLANTING MENU

**Please note, the pre planting menu under the rain menu is referred to as MENU 1 in the data**

**PRE PLANTING MENU**

* LAND PREPARATION
* SEED VARIETY
* REPEAT MENU
* GO TO MAIN MENU (RAIN MENU)
* INVALID KEY (ADDITIONAL OUTCOME)

#filtering out MENU 1 options as the pre planting  
MENU\_1 <- filter(log\_data, grepl("MENU 1",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("HHIMENU 1",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("APICULTURE MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("APICULTURE SUB4 MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("DAIRY MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("SMALL-SCALE SUB5 MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("DAIRY SUB2 MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("FATTENING MANU",logInfo))  
  
MENU\_1 <- filter(MENU\_1, !grepl("FATTENING SUB1 MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("FATTENING SUB2 MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("HOUSEHOLD MENU",logInfo))  
MENU\_1 <- filter(MENU\_1, !grepl("HOUSEHOLD SUB1 MENU ",logInfo))  
  
  
MENU\_1 <- MENU\_1 %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
MENU\_1$logInfo<- str\_replace(MENU\_1$logInfo, "MENU 1 - ", "")  
MENU\_1$logInfo<- str\_replace(MENU\_1$logInfo, " PRESSED", "")  
MENU\_1$logInfo<- str\_replace(MENU\_1$logInfo, " SELECTED", "")  
  
# plotting  
ggplot(data=MENU\_1, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 25, hjust= 1))+  
 geom\_bar(fill='cyan')+  
 labs(  
 title = 'Individuals By Content Selected at least Once from   
 RAIN MENU/PRE PLANTING OPTION',  
 x = 'Content Selected',  
 y= 'count'  
 )



### RAIN MENU/PRE PLANTING MENU/LAND PREPARATION

**We continue our pattern of following the most accessed submenu from the previous subment - land preparation! This is also the bottom of a tree. We will breakdown conent selection by language and overall**

##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> RANDOM EXPERIMENT –>

##### TOP MENU –> RAIN MENU –> PRE PLANTING MENU –> LAND PREPARATION

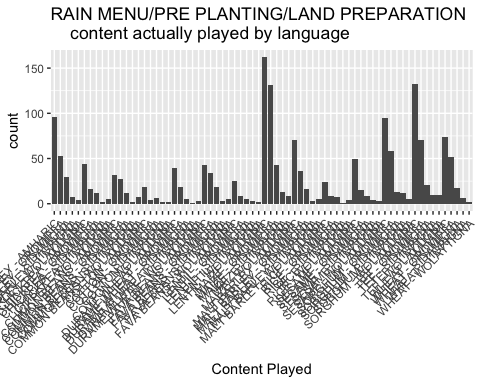
**LAND PREPARATION**

* BARLEY
* MAIZE
* SORGHUM
* TEF
* WHEAT
* SESAME
* FAVA BEAN
* CHICKPEA
* COMMON BEAN
* COTTON

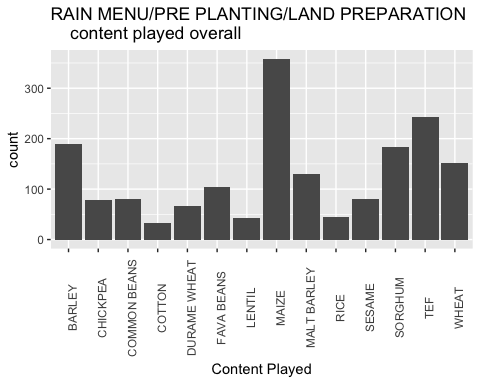
ORDER UNKOWN:

* RICE
* DURAME WHEAT
* LENTIL
* MALT BARLEY

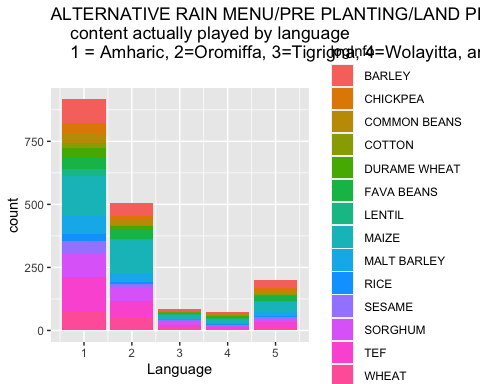
#GETTING THE CROP MENU FROM MENU 1 (rain/preplanting/land preparation sub menu)  
land\_prep <- filter(log\_data, grepl("CONTENT PLAYED - LAND PREPARATION - ",logInfo))  
  
land\_prep\_final <- land\_prep %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
land\_prep\_final$logInfo<- str\_replace(land\_prep\_final$logInfo, "CONTENT PLAYED - LAND PREPARATION -", "")  
  
ggplot(data=land\_prep\_final, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 45, hjust=1))+  
 geom\_bar()+  
 labs(  
 title = 'RAIN MENU/PRE PLANTING/LAND PREPARATION  
 content actually played by language',  
 x = 'Content Played',  
 y= 'count'  
 )



#get rid of the language breakdown  
land\_prep\_no\_lang <- land\_prep\_final  
  
land\_prep\_no\_lang$logInfo<- str\_replace(land\_prep\_no\_lang$logInfo, "- AMHARIC", "")  
land\_prep\_no\_lang$logInfo<- str\_replace(land\_prep\_no\_lang$logInfo, "- OROMIFFA", "")  
land\_prep\_no\_lang$logInfo<- str\_replace(land\_prep\_no\_lang$logInfo, "- TIGRIGNA", "")  
land\_prep\_no\_lang$logInfo<- str\_replace(land\_prep\_no\_lang$logInfo, "- WOLAYITTA", "")  
land\_prep\_no\_lang$logInfo<- str\_replace(land\_prep\_no\_lang$logInfo, "- WOLAYTIGNA", "")  
land\_prep\_no\_lang$logInfo<- str\_replace(land\_prep\_no\_lang$logInfo, "- SIDAMIGNA", "")  
  
ggplot(land\_prep\_no\_lang, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 90))+  
 geom\_bar()+  
 labs(  
 title = 'RAIN MENU/PRE PLANTING/LAND PREPARATION  
 content played overall',  
 x = 'Content Played',  
 y= 'count'  
 )



ggplot(land\_prep\_no\_lang, aes(langId, fill = logInfo)) +  
 geom\_bar()+  
 labs(  
 title = 'ALTERNATIVE RAIN MENU/PRE PLANTING/LAND PREPARATION  
 content actually played by language  
 1 = Amharic, 2=Oromiffa, 3=Tigrigna, 4=Wolayitta, and 5=Sidamigna  
 ',  
 x = 'Language'  
 )



**BIG TAKEAWAY: 35% of users end up accessing the maize menu at some point! Wheat, Tef, Sorghum, and Barley also appear to be qualitatively significant**

**There is no menu to go deeper in, so let’s go back up a level and check to see if other branches have a lot of access**

### RAIN MENU/PRE PLANTING MENU/SEED VARIETY

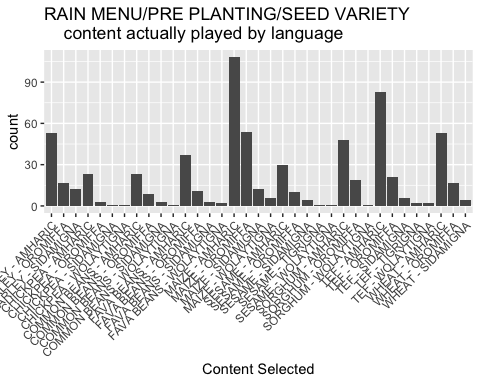
##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> RANDOM EXPERIMENT –>

##### TOP MENU –> RAIN MENU –> PRE PLANTING MENU –> SEED VARIETY

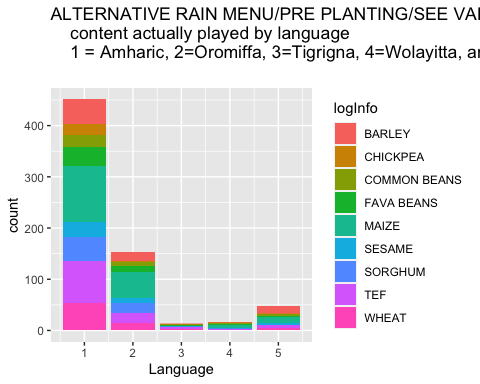
**SEED VARIETY**

* BARLEY
* MAIZE
* SORGHUM
* TEF
* WHEAT
* SESAME
* FAVA BEAN
* CHICKPEA
* COMMON BEAN

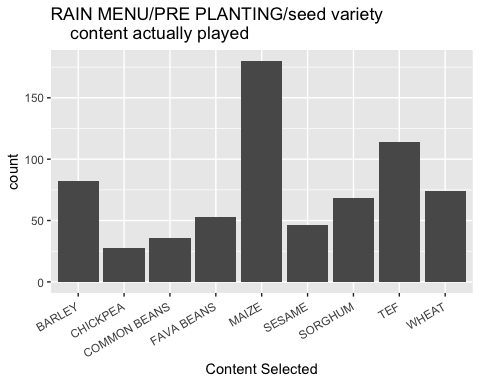
#GETTING THE CROP MENU FROM MENU 1 (rain/preplanting/seed variety sub menu)  
seed\_variety <- filter(log\_data, grepl("CONTENT PLAYED - SEED VARITY - ",logInfo))  
  
seed\_variety\_final <- seed\_variety %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
seed\_variety\_final$logInfo<- str\_replace(seed\_variety\_final$logInfo, "CONTENT PLAYED - SEED VARITY - ", "")  
  
ggplot(data=seed\_variety\_final, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 45, hjust=1))+  
 geom\_bar()+  
 labs(  
 title = 'RAIN MENU/PRE PLANTING/SEED VARIETY  
 content actually played by language',  
 x = 'Content Selected',  
 y= 'count'  
 )



#get rid of the language breakdown  
seed\_variety\_no\_lang <- seed\_variety\_final  
  
seed\_variety\_no\_lang$logInfo<- str\_replace(seed\_variety\_no\_lang$logInfo, "- AMHARIC", "")  
seed\_variety\_no\_lang$logInfo<- str\_replace(seed\_variety\_no\_lang$logInfo, "- OROMIFFA", "")  
seed\_variety\_no\_lang$logInfo<- str\_replace(seed\_variety\_no\_lang$logInfo, "- TIGRIGNA", "")  
seed\_variety\_no\_lang$logInfo<- str\_replace(seed\_variety\_no\_lang$logInfo, "- WOLAYITTA", "")  
seed\_variety\_no\_lang$logInfo<- str\_replace(seed\_variety\_no\_lang$logInfo, "- WOLAYTIGNA", "")  
seed\_variety\_no\_lang$logInfo<- str\_replace(seed\_variety\_no\_lang$logInfo, "- SIDAMIGNA", "")  
  
#alternative with langauge view  
ggplot(seed\_variety\_no\_lang, aes(langId, fill = logInfo)) +  
 geom\_bar()+  
 labs(  
 title = 'ALTERNATIVE RAIN MENU/PRE PLANTING/SEE VARIETY  
 content actually played by language  
 1 = Amharic, 2=Oromiffa, 3=Tigrigna, 4=Wolayitta, and 5=Sidamigna  
 ',  
 x = 'Language'  
 )



#no language - overall  
ggplot(seed\_variety\_no\_lang, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 30, hjust=1))+  
 geom\_bar()+  
 labs(  
 title = 'RAIN MENU/PRE PLANTING/seed variety  
 content actually played',  
 x = 'Content Selected',  
 y= 'count'  
 )



**We have mostly exhausted the pre-planting option menu, and if we go back up to the original rain menu (aka MAIN MENU), we can see that planting also has a large porportion of users - over half**

### RAIN MENU/PLANTING MENU

##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> RANDOM EXPERIMENT –>

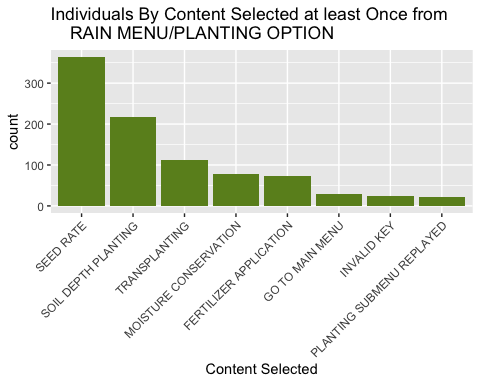
##### TOP MENU –> RAIN MENU –> PLANTING MENU

**The planting option as a subsection of the rain option is known as MENU 2 in the data**

PLANTING MENU

* SEED RATE
* SOIL DEPTH PLANTING
* TRANSPLANTING
* MOISTURE CONSERVATION
* FERTILIZER APPLICATION
* REPLAY MENY
* RETURN TO RAIN MENY
* INVALID KEY (ADDITIONAL OUTCOME)

#PLANTING OPTION - second most accessed MENU 2  
MENU\_2 <- filter(log\_data, grepl("MENU 2",logInfo))  
MENU\_2 <- filter(MENU\_2, !grepl("HHIMENU 2",logInfo))  
MENU\_2 <- filter(MENU\_2, !grepl("APICULTURE MENU",logInfo))  
MENU\_2 <- filter(MENU\_2, !grepl("APICULTURE SUB2 MENU",logInfo))  
MENU\_2 <- filter(MENU\_2, !grepl("DAIRY MENU",logInfo))  
MENU\_2 <- filter(MENU\_2, !grepl("FATTENING MANU",logInfo))  
  
MENU\_2 <- MENU\_2 %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
#cleaning  
MENU\_2$logInfo<- str\_replace(MENU\_2$logInfo, "MENU 2 - ", "")  
MENU\_2$logInfo<- str\_replace(MENU\_2$logInfo, " SELECTED", "")  
MENU\_2$logInfo<- str\_replace(MENU\_2$logInfo, " PRESSED", "")  
MENU\_2$logInfo<- str\_replace(MENU\_2$logInfo, " OPTION", "")  
  
  
#plotting  
ggplot(data=MENU\_2, aes(x=fct\_infreq(logInfo))) +  
 theme(axis.text.x = element\_text(angle = 45, hjust=1))+  
 geom\_bar(fill='olivedrab')+  
 labs(  
 title = 'Individuals By Content Selected at least Once from   
 RAIN MENU/PLANTING OPTION',  
 x = 'Content Selected',  
 y= 'count'  
 )



**As per usual, we will go into the most common selection: seed rate!**

### RAIN MENU/PLANTING MENU/SEED RATE

##### INCOMING CALL STARTED –> WELCOME MESSAGE PLAYED –> RANDOM EXPERIMENT –>

##### TOP MENU –> RAIN MENU –> PLANTING MENU –> SEED RATE

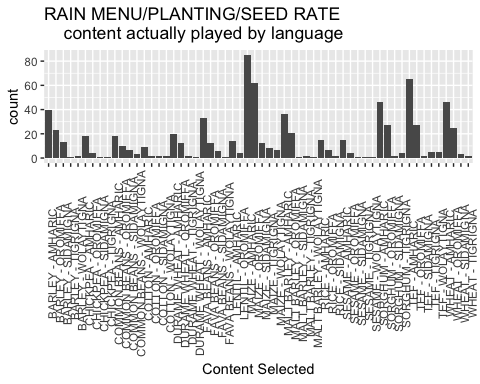
**SEED RATE**

* BARLEY
* MAIZE
* SORGHUM
* TEF
* WHEAT
* SESAME
* FAVA BEAN
* CHICKPEA
* COMMON BEAN
* COTTON

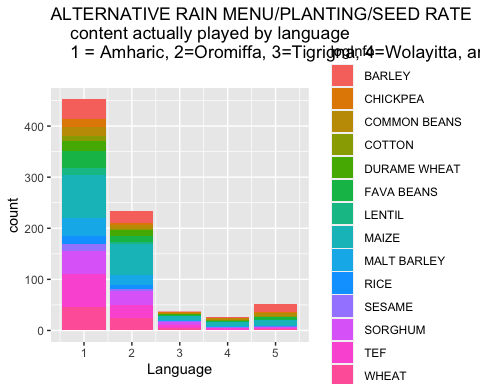
ORDER UNKOWN:

* RICE
* DURAME WHEAT
* LENTIL
* MALT BARLEY

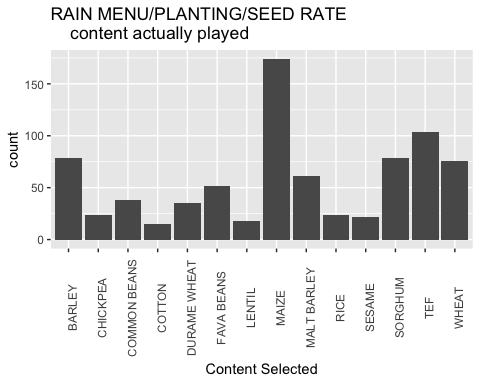
seed\_rate <- filter(log\_data, grepl("CONTENT PLAYED - SEED RATE",logInfo))  
  
seed\_rate\_final <- seed\_rate %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
seed\_rate\_final$logInfo<- str\_replace(seed\_rate\_final$logInfo, "CONTENT PLAYED - SEED RATE -", "")  
  
ggplot(data=seed\_rate\_final, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 90))+  
 geom\_bar()+  
 labs(  
 title = 'RAIN MENU/PLANTING/SEED RATE  
 content actually played by language',  
 x = 'Content Selected',  
 y= 'count'  
 )



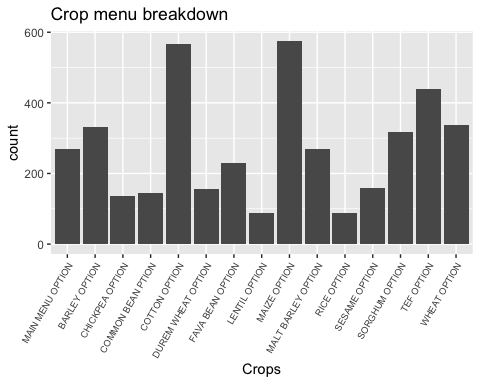
#get rid of the language breakdown  
seed\_rate\_no\_lang <- seed\_rate\_final  
  
seed\_rate\_no\_lang$logInfo<- str\_replace(seed\_rate\_no\_lang$logInfo, "- AMHARIC", "")  
seed\_rate\_no\_lang$logInfo<- str\_replace(seed\_rate\_no\_lang$logInfo, "- OROMIFFA", "")  
seed\_rate\_no\_lang$logInfo<- str\_replace(seed\_rate\_no\_lang$logInfo, "- TIGRIGNA", "")  
seed\_rate\_no\_lang$logInfo<- str\_replace(seed\_rate\_no\_lang$logInfo, "- WOLAYITTA", "")  
seed\_rate\_no\_lang$logInfo<- str\_replace(seed\_rate\_no\_lang$logInfo, "- WOLAYTIGNA", "")  
seed\_rate\_no\_lang$logInfo<- str\_replace(seed\_rate\_no\_lang$logInfo, "- SIDAMIGNA", "")  
  
#alternative with langauge view  
ggplot(seed\_rate\_no\_lang, aes(langId, fill = logInfo)) +  
 geom\_bar()+  
 labs(  
 title = 'ALTERNATIVE RAIN MENU/PLANTING/SEED RATE  
 content actually played by language  
 1 = Amharic, 2=Oromiffa, 3=Tigrigna, 4=Wolayitta, and 5=Sidamigna  
 ',  
 x = 'Language'  
 )



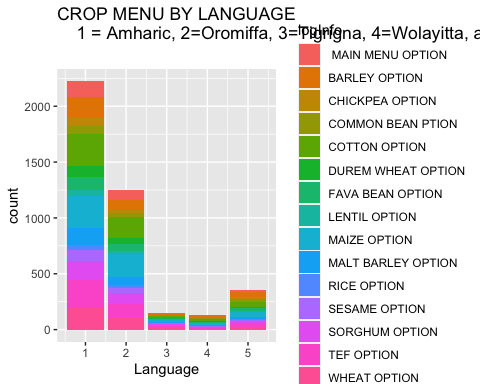
#no language  
ggplot(seed\_rate\_no\_lang, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 90))+  
 geom\_bar()+  
 labs(  
 title = 'RAIN MENU/PLANTING/SEED RATE  
 content actually played',  
 x = 'Content Selected',  
 y= 'count'  
 )



crop\_menu <- log\_data %>%  
 filter(grepl("CROP MENU - ",logInfo)) %>%  
 filter(!grepl("HHI",logInfo)) %>%  
 filter(!grepl("REPLAYED",logInfo))  
  
crop\_menu <- crop\_menu %>%  
 group\_by(callerId, logInfo, langId) %>%  
 count()  
  
crop\_menu$logInfo<- str\_replace(crop\_menu$logInfo, "CROP MENU - ", "")  
crop\_menu$logInfo<- str\_replace(crop\_menu$logInfo, " SELECTED", "")  
crop\_menu$logInfo<- str\_replace(crop\_menu$logInfo, "GO TO", "")  
  
  
ggplot(crop\_menu, aes(x=logInfo)) +  
 theme(axis.text.x = element\_text(angle = 60, hjust =1, size = 7))+  
 geom\_bar()+  
 labs(  
 title = 'Crop menu breakdown',  
 x = 'Crops',  
 y= 'count'  
 )



ggplot(crop\_menu, aes(langId, fill = logInfo)) +  
 geom\_bar()+  
 labs(  
 title = 'CROP MENU BY LANGUAGE  
 1 = Amharic, 2=Oromiffa, 3=Tigrigna, 4=Wolayitta, and 5=Sidamigna  
 ',  
 x = 'Language'  
 )

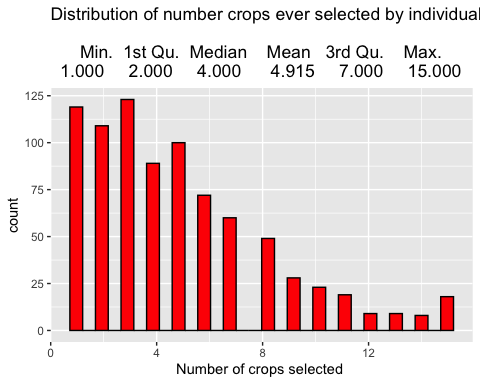


num\_crop\_types <- data.frame(as.numeric(table(crop\_menu$callerId)))  
num\_crop\_types$frequencies <- num\_crop\_types$as.numeric.table.crop\_menu.callerId.  
  
summary(num\_crop\_types$frequencies)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 4.000 4.915 7.000 15.000

ggplot(num\_crop\_types, aes(frequencies)) +  
 geom\_histogram(color="black", fill="red") +  
 labs(  
 title = 'Distribution of number crops ever selected by individual \n  
 Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1.000 2.000 4.000 4.915 7.000 15.000 ',  
 x = 'Number of crops selected '  
 )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



HOW MANY CROPS PER USER ARE LOOKED AT??????????