Names

# Libraries

library(readxl)

## Warning: package 'readxl' was built under R version 4.1.2

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(stringr)  
library(writexl)  
library(rlang)

# Loading in the Data

#raw\_data <- read\_excel("Preferred Name QA\_QC 20211007-090939.xlsx")  
raw\_data <- read\_excel("Quick Query 20220322-142208.xlsx")  
clean\_data <- raw\_data

# Creation of the function that will find common errors in a dataset

Uncleaned <- function(Raw\_data,Column\_Name) { # Creation of function   
df1 <- Raw\_data%>% # I created this first dataset to filter for the peple who have an unknown character   
 filter(str\_detect(Column\_Name,"1|2|3|4|5|6|7|8|9|0|\\(|\\)")) %>%   
 mutate(Reason\_4\_Review = "Unknown Character")  
df2 <- Raw\_data%>% # I created this second dataset to find the people who have an "or" or a ","  
 filter(str\_detect(Column\_Name," or |,")) %>%  
 mutate(Reason\_4\_Review = "Or/,")  
Return <- union(df1,df2) # I combined the data  
 return(Return)  
}  
  
Length\_Check <- function(Raw\_data,Column\_Name){  
df1 <- Raw\_data%>%   
 filter(str\_length(Column\_Name) > 12) %>%   
 mutate(Reason\_4\_Review = "Greater than 12 characters")  
}

# Finding the data that needs to be clean

Needs\_Review\_p <- Uncleaned(raw\_data,raw\_data$Preferred) %>% # Execution of the data through the function  
 mutate(Error\_Location = "Preferred")  
Needs\_Review\_f <- Uncleaned(raw\_data,raw\_data$First) %>%   
 mutate(Error\_Location = "First")  
Needs\_Review\_m <- Uncleaned(raw\_data,raw\_data$Middle)%>%   
 mutate(Error\_Location = "Middle")  
Needs\_Review\_l <- Uncleaned(raw\_data,raw\_data$Last)%>%   
 mutate(Error\_Location = "Last")  
Needs\_Review\_length\_p <- Length\_Check(raw\_data,raw\_data$Preferred) %>%  
 mutate(Error\_Location = "Preferred")  
  
#Needs\_review\_pfl <- clean\_data%>% #This was to find when when the preferred name consisted of the first name plus last name.   
 #filter(str\_detect(Preferred, First)) #%>%  
 #filter(str\_detect(Preferred, Last)) %>%  
 #mutate(Reason\_4\_Review = "Preferred = First+Last Name")  
  
Needs\_review <- bind\_rows(Needs\_Review\_length\_p,Needs\_Review\_f,Needs\_Review\_l,Needs\_Review\_m,Needs\_Review\_p) # Combining the data so all the uncleaned data is in one place.

#Exporting data

#write\_xlsx(Needs\_review,"C:\\Users\\foilbn\\Documents\\R workspace\\Needs\_review.xlsx") #This exports the data as a excel file.

# Preferred Name if unexpected number

Fixed\_Pre <- Needs\_Review\_p %>%  
 filter(str\_detect(Preferred,"1|2|3|4|5|6|7|8|9|0")) %>%  
 mutate(Preferred = First)  
  
clean\_data <- full\_join(clean\_data,Fixed\_Pre)%>%  
 filter(str\_detect(Preferred,"1|2|3|4|5|6|7|8|9|0") == FALSE)

## Joining, by = c("Ref", "First", "Preferred", "Middle", "Last", "Entry Term")

# Preferred Name if format: First(Preferred)

# Preferred Name if

#Fixed\_pfl <- clean\_data%>%   
 # filter(str\_detect(Preferred," or ")) %>%  
 #filter(str\_detect(Preferred, First)) %>%  
 #filter(str\_detect(Preferred, Last)) %>%  
 #mutate(Preferred = First)  
  
#Tst <- full\_join(clean\_data,Fixed\_pfl)#%>%  
 # filter(!str\_detect(Preferred, First) == FALSE & !str\_detect(Preferred, Last) == FALSE)  
   
  
#clean\_data[str\_detect(clean\_data$Preferred,clean\_data$Last)] #<- Fixed\_pfl$Preferred

# Middle Name

raw\_data %>%  
 filter(str\_length(raw\_data$Preferred)>12)

## # A tibble: 1,327 x 6  
## Ref First Preferred Middle Last `Entry Term`  
## <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 624573678 Ava Koistinen Ava Koistinen <NA> <NA> Fall 2022   
## 2 258447623 Megan Gabrielle Megan Gabrielle O Abano Fall 2022   
## 3 370392304 Desirae-Marie Desirae-Marie B Abayab~ Fall 2022   
## 4 886156255 Mary Ryan Elizabeth Mary Ryan Elizabeth S Abbago Fall 2022   
## 5 254491583 Muhammed Rayyan Muhammed Rayyan <NA> Abdul ~ Fall 2022   
## 6 114356641 Danika Gloria Danika Gloria P Abelar~ Fall 2022   
## 7 973734494 Mosinmiloluwa Mosinmiloluwa <NA> Abidoye Fall 2022   
## 8 988574768 Oluwatobiloba Oluwatobiloba M Abidoye Fall 2022   
## 9 120059970 Bonaventura/Bagas Bonaventura/Bagas <NA> Abiman~ Fall 2022   
## 10 480949466 Sanmathi Priya Sanmathi Priya <NA> Abiram~ Fall 2022   
## # ... with 1,317 more rows

# Creation of the cleaning function

Cleaned <- function(Raw\_data, Needs\_review, Column\_Name) {  
   
Raw\_data %>%  
 filter(str\_detect(Column\_Name,"1|2|3|4|5|6|7|8|9|0")) %>%   
 mutate(Column\_Name, )  
   
Raw\_data %>%   
 filter()  
   
   
   
   
   
}

# Cleaning the data

Cleaning <- function(Unclean\_Dataset,Column\_Name){  
 Unclean\_Dataset %>%  
 filter(str\_detect(Column\_Name,"1|2|3|4|5|6|7|8|9|0")) %>%  
 mutate(Column\_Name = NA)  
}

# Cleaning the data

#Fixed\_Pre <- uncleaned\_Pre %>%  
# filter(str\_detect(Preferred,"1|2|3|4|5|6|7|8|9|0")) %>%  
# mutate(Preferred = NA)  
  
#clean\_data <- full\_join(clean\_data,Fixed\_Pre)%>%  
# filter(str\_detect(Preferred,"1|2|3|4|5|6|7|8|9|0") == FALSE | is.na(Preferred) == TRUE)  
  
  
  
  
#Fixed\_Mid <- uncleaned\_Mid %>%  
# filter(str\_detect(Middle,"1|2|3|4|5|6|7|8|9|0")) %>%  
# mutate(Middle = NA)  
  
#clean\_data <- full\_join(clean\_data,Fixed\_Mid)%>%  
# filter(str\_detect(Middle,"1|2|3|4|5|6|7|8|9|0") == FALSE | is.na(Middle) == TRUE)

#Luiza Acioly - look her frist name and last name # I need to fix the () and “” issue for reusability # Ned to clean the data for a theoretical missinput for last name and first name # For