**Project Code and GitHub Link**

Shawn Behrend

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Orenthio Goodwin

15 May 2022

**Project Code and GitHub Link**

GitHub Link: https://github.com/ShawnBehrend/MIS581\_Capstone

The following pages of this document use the R Markdown application to include R code, the results of running the code as well as other text added by the author. Each R code script is numbered by the order in which the scripts were run to complete the necessary steps of the project. The R code script names are as follows:

* 1. Split F M Names All Years
* 2. Compare M-F Names by Last Year
* 3. GN M-F Names By Year
* 4. Years in Decade Summary Table
* 5. All Decades Summary Table
* 6. Plot Quantity and Occurrences of GN versus All Names
* 7. All GN Names
* 8. Plot New and Common GN Names
* 9. Correlation Test and Plots

1. Split F M Names All Years

Shawn Behrend

2022-05-05

This code opens each YOB file per year and splits out the female names from the male names. Each file is saved as a separate CSV file named for the year and the following ending “\_m\_names\_top1000.txt” or “\_f\_names\_top1000.txt”.

#input what years will be analyzed, based on available data files  
allyears<-c(1880:2020)  
  
#establish first index for data frame based on years chosen for analysis  
firstyearminus1<-allyears[1]-1  
#get index for last year to be analyzed, used in plot later  
allyears\_last<-length(allyears)  
  
# set up a data frame full of "NA" to store the desired name and sex with each year and rank from the data files  
year\_rank<-data.frame(matrix(NA,length(allyears),4))  
  
# loop for each year to be searched  
for (year in allyears) {  
 #load names by year file  
 yob\_file <- read.csv(paste0("C:/Users/shawn/Shawn/CSU\_global/MIS581/names/yob",year,".txt"), header=FALSE)  
 #add headers  
 names(yob\_file) <- c("Name","Sex","Occurance")  
  
#which lines are female and male names  
is\_f<-yob\_file$Sex=="F"  
firstf<-min(which(is\_f==TRUE))  
lastf<-max(which(is\_f==TRUE))  
is\_m<-yob\_file$Sex=="M"  
firstm<-min(which(is\_m==TRUE))  
lastm<-max(which(is\_m==TRUE))  
#split file into female and male names  
yob\_file\_f<-yob\_file[firstf:lastf,]  
yob\_file\_m<-yob\_file[firstm:lastm,]  
# add columns/headers for year and rank to the file  
f\_rows<-nrow(yob\_file\_f)  
m\_rows<-nrow(yob\_file\_m)  
# year  
Year<-year  
#rank  
f\_Rank<-c(1:f\_rows)  
m\_Rank<-c(1:m\_rows)  
# add rank and year to dataset  
#cbind  
yob\_file\_f\_rank<-cbind(yob\_file\_f,f\_Rank,Year)  
yob\_file\_m\_rank<-cbind(yob\_file\_m,m\_Rank,Year)  
# subset only top 100 male names  
yob\_file\_m\_1000<-yob\_file\_m\_rank[c(1:1000),,]  
# subset only top 1000 female names  
yob\_file\_f\_1000<-yob\_file\_f\_rank[c(1:1000),,]  
# write file with top 1000 F or top 100 M names  
write.csv(x=yob\_file\_f\_1000,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",year,"\_f\_names\_top1000.txt"))  
write.csv(x=yob\_file\_m\_1000,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",year,"\_m\_names\_top1000.txt"))   
}

An excerpt of the 2020 year of birth (YOB) file showing the first 10 male names and first 10 female names is shown here.

head(yob\_file\_m\_1000,10)

## Name Sex Occurrence m\_Rank Year  
## 17361 Liam M 19659 1 2020  
## 17362 Noah M 18252 2 2020  
## 17363 Oliver M 14147 3 2020  
## 17364 Elijah M 13034 4 2020  
## 17365 William M 12541 5 2020  
## 17366 James M 12250 6 2020  
## 17367 Benjamin M 12136 7 2020  
## 17368 Lucas M 11281 8 2020  
## 17369 Henry M 10705 9 2020  
## 17370 Alexander M 10151 10 2020

head(yob\_file\_f\_1000,10)

## Name Sex Occurrence f\_Rank Year  
## 1 Olivia F 17535 1 2020  
## 2 Emma F 15581 2 2020  
## 3 Ava F 13084 3 2020  
## 4 Charlotte F 13003 4 2020  
## 5 Sophia F 12976 5 2020  
## 6 Amelia F 12704 6 2020  
## 7 Isabella F 12066 7 2020  
## 8 Mia F 11157 8 2020  
## 9 Evelyn F 9445 9 2020  
## 10 Harper F 8778 10 2020

2. Compare M-F Names by Last Year

Shawn Behrend

2022-05-06

This code compares every name in the Top 1000 CSV file for males with every name in the female, then compiles those names and their corresponding rank.

#input what years will be analyzed, based on available data files  
decade1<-c(1880:1889)  
decade2<-decade1+10  
decade3<-decade2+10  
decade4<-decade3+10  
decade5<-decade4+10  
decade6<-decade5+10  
decade7<-decade6+10  
decade8<-decade7+10  
decade9<-decade8+10  
decade10<-decade9+10  
decade11<-decade10+10  
decade12<-decade11+10  
decade13<-decade12+10  
decade14<-c(2010:2020)  
#grab first and last year for use in file names  
all\_years<-decade14  
first\_year<-min(all\_years)  
last\_year<-max(all\_years)  
  
# input names to be searched on  
comp\_names\_df<-read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",last\_year,"\_m\_names\_top1000.txt"), fileEncoding="UTF-8-BOM",colClasses=c("character"))  
comp\_names<-comp\_names\_df[[2]]  
  
# begin find\_name loop  
#choose number of searchnames  
inames<-c(1:length(comp\_names))  
#establish first index finding searchname  
iname<-1  
  
# set up a data frame full of "NA" to store the desired name and sex with each year and rank from the data files  
 name\_rank\_f<-data.frame(matrix(NA,length(comp\_names),5))  
 name\_rank\_m<-data.frame(matrix(NA,length(comp\_names),5))  
  
comp\_file\_f <- read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",last\_year,"\_f\_names\_top1000.txt"),colClasses=c(Name="character",Sex="character"))  
comp\_file\_m <- read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",last\_year,"\_m\_names\_top1000.txt"),colClasses=c(Name="character",Sex="character"))  
   
 #start loop through names  
 for (iname in inames) {  
#establish searchname  
 searchname<-comp\_names[iname]  
 find\_name\_f<-c("NA",searchname,"F","NA","NA","NA")  
 find\_name\_m<-c("NA",searchname,"M","NA","NA","NA")  
# Look for row that search name is in using grep.   
# If length of yob\_file\_sex$Name is greater than zero, replace dummy row with this row #  
#----female name files---  
 if(length(grep(paste("^",searchname,"$", sep=""),comp\_file\_f$Name))>0) {  
 #save row info in new variable  
 find\_name\_f<-c(comp\_file\_f[(which(comp\_file\_f$Name==searchname)),])  
 }  
 # Use index to save desired name data to the data frame  
 name\_rank\_f[iname,]<-c(searchname,find\_name\_f[3],find\_name\_f[4],find\_name\_f[5],  
 find\_name\_f[6])  
#----male name files---  
 if(length(grep(paste("^",searchname,"$", sep=""),comp\_file\_m$Name))>0) {  
 #save row info in new variable  
 find\_name\_m<-c(comp\_file\_m[(which(comp\_file\_m$Name==searchname)),])  
 }  
 # Use index to save desired name data to the data frame. Use female year to know which should be eliminated.  
 name\_rank\_m[iname,]<-c(searchname,find\_name\_m[3],find\_name\_m[4],find\_name\_m[5],find\_name\_f[6])  
 iname=iname+1  
 }  
 #add column names to data frame  
 names(name\_rank\_f)<-c("Name","Sex","Occurrance","Rank","Year")  
 names(name\_rank\_m)<-c("Name","Sex","Occurrance","Rank","Year")  
   
 #save names to data file  
 write.csv(x=name\_rank\_f,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_F\_",last\_year,".csv"))  
 write.csv(x=name\_rank\_m,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_M\_",last\_year,".csv"))

An excerpt of the matching names and corresponding ranks that were found for males and females is shown here. If the name did not appear on both the male and female list, then the occurrence, rank and year show "NA". As is seen here, Noah was found on both lists, with rank 876 on the female list and rank 2 on the male list.

head(name\_rank\_f,10)

## Name Sex Occurrance Rank Year  
## 1 Liam F NA NA NA  
## 2 Noah F 305 876 2020  
## 3 Oliver F NA NA NA  
## 4 Elijah F NA NA NA  
## 5 William F NA NA NA  
## 6 James F NA NA NA  
## 7 Benjamin F NA NA NA  
## 8 Lucas F NA NA NA  
## 9 Henry F NA NA NA  
## 10 Alexander F NA NA NA

head(name\_rank\_m,10)

## Name Sex Occurrance Rank Year  
## 1 Liam M 19659 1 NA  
## 2 Noah M 18252 2 2020  
## 3 Oliver M 14147 3 NA  
## 4 Elijah M 13034 4 NA  
## 5 William M 12541 5 NA  
## 6 James M 12250 6 NA  
## 7 Benjamin M 12136 7 NA  
## 8 Lucas M 11281 8 NA  
## 9 Henry M 10705 9 NA  
## 10 Alexander M 10151 10 NA

Before this file can be used for the next step in the process, the CSV file must be brought into Microsoft Excel, and all NA lines will be removed.

3. GN M-F Names By Year

Shawn Behrend

2022-05-07

After sorting the comp\_gn\_names files for males and females for the last year in the decade and removing all names that are NA for that decade, this code runs and makes a separate CSV for every year in the decade, using only the names from the last year in the decade.

#input what years will be analyzed, based on available data files  
decade1<-c(1880:1889)  
decade2<-decade1+10  
decade3<-decade2+10  
decade4<-decade3+10  
decade5<-decade4+10  
decade6<-decade5+10  
decade7<-decade6+10  
decade8<-decade7+10  
decade9<-decade8+10  
decade10<-decade9+10  
decade11<-decade10+10  
decade12<-decade11+10  
decade13<-decade12+10  
decade14<-c(2010:2020)  
#grab first and last year for use in file names  
all\_years<-decade14  
first\_year<-min(all\_years)  
last\_year<-max(all\_years)  
#dont write over comp\_gn file for last year in decade  
count\_allyearsminus1<-length(all\_years)-1  
allyearsminus1<-all\_years[1:count\_allyearsminus1]  
  
# input names to be searched on  
gn\_names\_df<-read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_M\_",last\_year,".csv"),fileEncoding="UTF-8-BOM",colClasses=c(Name="character",Sex="character"))  
gn\_names<-gn\_names\_df[[2]]  
#choose number of searchnames  
inames<-c(1:length(gn\_names))  
for (year in allyearsminus1) {  
# begin find\_name loop  
#establish first index finding searchname  
iname<-1  
  
  
# set up a data frame full of "NA" to store the desired name and sex with each year and rank from the data files  
 name\_rank\_f<-data.frame(matrix(NA,length(gn\_names),5))  
 name\_rank\_m<-data.frame(matrix(NA,length(gn\_names),5))  
  
 comp\_file\_f <- read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",year,"\_f\_names\_top1000.txt"),colClasses=c(Name="character",Sex="character"))  
   
 comp\_file\_m <- read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",year,"\_m\_names\_top1000.txt"),colClasses=c(Name="character",Sex="character"))  
   
 #start loop through names  
 for (iname in inames) {  
 #establish searchname  
 searchname<-gn\_names[iname]  
 find\_name\_f<-c("NA",searchname,"F","NA","NA","NA")  
 find\_name\_m<-c("NA",searchname,"M","NA","NA","NA")  
 # Look for row that search name is in using grep.   
 # If length of comp\_file\_sex$Name is greater than zero, replace dummy row with this row   
#----female name files---  
 if(length(grep(paste("^",searchname,"$", sep=""),comp\_file\_f$Name))>0) {  
 #save row info in new variable  
 find\_name\_f<-c(comp\_file\_f[(which(comp\_file\_f$Name==searchname)),])  
 }  
 # Use index to save desired name data to the data frame  
 name\_rank\_f[iname,]<-c(searchname,find\_name\_f[3],find\_name\_f[4],find\_name\_f[5],find\_name\_f[6])  
#----male name files---  
 if(length(grep(paste("^",searchname,"$", sep=""),comp\_file\_m$Name))>0) {  
 #save row info in new variable  
 find\_name\_m<-c(comp\_file\_m[(which(comp\_file\_m$Name==searchname)),])  
 }  
 # Use index to save desired name data to the data frame. Use female year to know which should be eliminated.  
 name\_rank\_m[iname,]<-c(searchname,find\_name\_m[3],find\_name\_m[4],find\_name\_m[5],find\_name\_f[6])  
 iname=iname+1  
 }  
 #add column names to data frame  
 names(name\_rank\_f)<-c("Name","Sex","Occurrance","Rank","Year")  
 names(name\_rank\_m)<-c("Name","Sex","Occurrance","Rank","Year")  
   
 #save names to data file  
 write.csv(x=name\_rank\_f,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_F\_",year,".csv"))  
 write.csv(x=name\_rank\_m,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_M\_",year,".csv"))  
 }

An excerpt of the output of this file is shown for year 2019. Even though the name Noah shows as "NA" in this year in the female list, it did appear in other years for both males and females.

head(name\_rank\_f,10)

## Name Sex Occurrence Rank Year  
## 1 Noah F NA NA NA  
## 2 Logan F 993 322 2019  
## 3 Carter F 610 513 2019  
## 4 Dylan F 728 436 2019  
## 5 Ezra F 326 830 2019  
## 6 Ryan F 773 399 2019  
## 7 Cameron F 562 544 2019  
## 8 Hunter F 308 866 2019  
## 9 Angel F 781 393 2019  
## 10 Jordan F 768 403 2019

head(name\_rank\_m,10)

## Name Sex Occurrence Rank Year  
## 1 Noah M 19097 2 NA  
## 2 Logan M 10520 16 2019  
## 3 Carter M 8652 30 2019  
## 4 Dylan M 7530 39 2019  
## 5 Ezra M 6521 49 2019  
## 6 Ryan M 6104 54 2019  
## 7 Cameron M 5236 67 2019  
## 8 Hunter M 5369 66 2019  
## 9 Angel M 4895 72 2019  
## 10 Jordan M 4317 89 2019

4. Years in Decade Summary Table

Shawn Behrend

2022-05-07

This code reads each GN file per decade to create a summary table of all applicable parameters for this study.

#input what years will be analyzed, based on available data files  
decade1<-c(1880:1889)  
decade2<-decade1+10  
decade3<-decade2+10  
decade4<-decade3+10  
decade5<-decade4+10  
decade6<-decade5+10  
decade7<-decade6+10  
decade8<-decade7+10  
decade9<-decade8+10  
decade10<-decade9+10  
decade11<-decade10+10  
decade12<-decade11+10  
decade13<-decade12+10  
decade14<-c(2010:2020)

Note that this code must be run separately per decade. Currently the code was set up for decade 14 (2010-2020), shown below.

#grab first and last year for use in file names  
all\_years<-decade14  
first\_year<-min(all\_years)  
last\_year<-max(all\_years)  
#establish first index for data frame based on years chosen for analysis  
firstyearminus1<-all\_years[1]-1  
#--- MAKE TABLE  
# set up a data frame full of "NA" to store the desired info by year from the data files  
years\_per\_decade\_table<-data.frame(matrix(NA,length(all\_years),13))  
names(years\_per\_decade\_table)<-c("Year","# of GN Names","Total Names","Ratio of GN Names to Total Names","F occurrances of GN names","Total F occurrances","Ratio of F GN Occur to Total F Occur","M occurrances of GN names","Total M occurrances","Ratio of M GN Occur to Total M Occur","Total GN Occurrances","Total Occurrances","Ratio of GN Occurrances to Total Occurrances")  
# start years loop  
for (year in all\_years) {  
# years per decade table  
# read in the gn names files for the year  
#female gn files  
year\_info\_gn\_f\_df<-read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_F\_",year,".csv"),fileEncoding="UTF-8-BOM",colClasses=c(Name="character",Sex="character"))  
year\_info\_gn\_f\_names<-year\_info\_gn\_f\_df[[2]]  
year\_info\_gn\_f\_occur<-year\_info\_gn\_f\_df[[4]]  
sum\_year\_info\_gn\_f\_occur<-sum(year\_info\_gn\_f\_occur,na.rm=TRUE)  
count\_year\_info\_gn\_f\_names<-length(year\_info\_gn\_f\_names)  
# male gn files  
year\_info\_gn\_m\_df<-read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/comp\_gn\_names\_M\_",year,".csv"),fileEncoding="UTF-8-BOM",colClasses=c(Name="character",Sex="character"))  
year\_info\_gn\_m\_names<-year\_info\_gn\_m\_df[[2]]  
year\_info\_gn\_m\_occur<-year\_info\_gn\_m\_df[[4]]  
sum\_year\_info\_gn\_m\_occur<-sum(year\_info\_gn\_m\_occur,na.rm=TRUE)  
count\_year\_info\_gn\_m\_names<-length(year\_info\_gn\_m\_names)  
# sum of gn occurrances  
sum\_gn\_occur<-sum\_year\_info\_gn\_f\_occur+sum\_year\_info\_gn\_m\_occur  
# read in the total names files for the year  
# female all names  
year\_info\_all\_f\_df<-read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",year,"\_f\_names\_top1000.txt"),fileEncoding="UTF-8-BOM",colClasses=c(Name="character",Sex="character"))  
year\_info\_all\_f\_names<-year\_info\_all\_f\_df[[2]]  
year\_info\_all\_f\_occur<-year\_info\_all\_f\_df[[4]]  
sum\_year\_info\_all\_f\_occur<-sum(year\_info\_all\_f\_occur,na.rm=TRUE)  
count\_year\_info\_all\_f\_names<-length(year\_info\_all\_f\_names)  
# male all names  
year\_info\_all\_m\_df<-read.csv(paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",year,"\_m\_names\_top1000.txt"),fileEncoding="UTF-8-BOM",colClasses=c(Name="character",Sex="character"))  
year\_info\_all\_m\_names<-year\_info\_all\_m\_df[[2]]  
year\_info\_all\_m\_occur<-year\_info\_all\_m\_df[[4]]  
sum\_year\_info\_all\_m\_occur<-sum(year\_info\_all\_m\_occur,na.rm=TRUE)  
count\_year\_info\_all\_m\_names<-length(year\_info\_all\_m\_names)  
# sum of all names and occur  
count\_year\_info\_all\_names<-sum(count\_year\_info\_all\_m\_names,count\_year\_info\_all\_f\_names)  
# sum of all occurrences  
sum\_all\_occur<-sum\_year\_info\_all\_f\_occur+sum\_year\_info\_all\_m\_occur  
#ratios  
ratio\_gn\_and\_all\_names<-count\_year\_info\_gn\_f\_names/count\_year\_info\_all\_names  
ratio\_f\_gn\_and\_all\_occur<-sum\_year\_info\_gn\_f\_occur/sum\_year\_info\_all\_f\_occur  
ratio\_m\_gn\_and\_all\_occur<-sum\_year\_info\_gn\_m\_occur/sum\_year\_info\_all\_m\_occur  
ratio\_gn\_and\_all\_occur<-sum\_gn\_occur/sum\_all\_occur  
# assemble row in table  
i<-year-firstyearminus1  
years\_per\_decade\_table[i,]<-c(year,count\_year\_info\_gn\_m\_names,count\_year\_info\_all\_names,ratio\_gn\_and\_all\_names,sum\_year\_info\_gn\_f\_occur,sum\_year\_info\_all\_f\_occur,ratio\_f\_gn\_and\_all\_occur,sum\_year\_info\_gn\_m\_occur,sum\_year\_info\_all\_m\_occur,ratio\_m\_gn\_and\_all\_occur,sum\_gn\_occur,sum\_all\_occur,ratio\_gn\_and\_all\_occur)  
}  
#save to data file  
write.csv(x=years\_per\_decade\_table,file=paste0("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/",first\_year,"\_",last\_year,"/",first\_year,"\_",last\_year,"\_summary\_table.csv"))

An excerpt of the first few rows from the Years per Decade summary table for Decade 14 are shown below.

head(years\_per\_decade\_table,20)

## Year # of GN Names Total Names Ratio of GN Names to Total Names  
## 1 2010 85 2000 0.0425  
## 2 2011 85 2000 0.0425  
## 3 2012 85 2000 0.0425  
## 4 2013 85 2000 0.0425  
## 5 2014 85 2000 0.0425  
## 6 2015 85 2000 0.0425  
## 7 2016 85 2000 0.0425  
## 8 2017 85 2000 0.0425  
## 9 2018 85 2000 0.0425  
## 10 2019 85 2000 0.0425  
## 11 2020 85 2000 0.0425  
## F occurrances of GN names Total F occurrances  
## 1 73973 1307926  
## 2 73607 1295259  
## 3 75797 1294652  
## 4 80402 1296592  
## 5 85416 1325930  
## 6 89910 1322802  
## 7 93287 1312169  
## 8 91657 1275695  
## 9 92724 1257343  
## 10 93043 1235481  
## 11 93268 1180696  
## Ratio of F GN Occur to Total F Occur M occurrances of GN names  
## 1 0.05655748 149710  
## 2 0.05682802 149875  
## 3 0.05854623 151182  
## 4 0.06201025 152187  
## 5 0.06441969 160444  
## 6 0.06796936 162211  
## 7 0.07109374 155867  
## 8 0.07184868 155443  
## 9 0.07374599 150056  
## 10 0.07530913 146688  
## 11 0.07899408 142672  
## Total M occurrances Ratio of M GN Occur to Total M Occur  
## 1 1615800 0.09265379  
## 2 1599083 0.09372559  
## 3 1594004 0.09484418  
## 4 1592997 0.09553502  
## 5 1616173 0.09927403  
## 6 1609215 0.10080132  
## 7 1585894 0.09828337  
## 8 1538982 0.10100378  
## 9 1501581 0.09993201  
## 10 1473990 0.09951764  
## 11 1404385 0.10159038  
## Total GN Occurrances Total Occurrances  
## 1 223683 2923726  
## 2 223482 2894342  
## 3 226979 2888656  
## 4 232589 2889589  
## 5 245860 2942103  
## 6 252121 2932017  
## 7 249154 2898063  
## 8 247100 2814677  
## 9 242780 2758924  
## 10 239731 2709471  
## 11 235940 2585081  
## Ratio of GN Occurrances to Total Occurrances  
## 1 0.07650614  
## 2 0.07721340  
## 3 0.07857599  
## 4 0.08049207  
## 5 0.08356608  
## 6 0.08598893  
## 7 0.08597260  
## 8 0.08778982  
## 9 0.08799807  
## 10 0.08847889  
## 11 0.09126987

5. All Decades Summary Table

Shawn Behrend

2022-05-07

This code reads each of the Years per Decade summary tables created in the previous piece of code to create an overall summary table.

# read summary table files  
sum\_1880\_1889\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1880\_1889/1880\_1889\_summary\_table.csv")  
sum\_1890\_1899\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1890\_1899/1890\_1899\_summary\_table.csv")  
sum\_1900\_1909\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1900\_1909/1900\_1909\_summary\_table.csv")  
sum\_1910\_1919\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1910\_1919/1910\_1919\_summary\_table.csv")  
sum\_1920\_1929\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1920\_1929/1920\_1929\_summary\_table.csv")  
sum\_1930\_1939\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1930\_1939/1930\_1939\_summary\_table.csv")  
sum\_1940\_1949\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1940\_1949/1940\_1949\_summary\_table.csv")  
sum\_1950\_1959\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1950\_1959/1950\_1959\_summary\_table.csv")  
sum\_1960\_1969\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1960\_1969/1960\_1969\_summary\_table.csv")  
sum\_1970\_1979\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1970\_1979/1970\_1979\_summary\_table.csv")  
sum\_1980\_1989\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1980\_1989/1980\_1989\_summary\_table.csv")  
sum\_1990\_1999\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1990\_1999/1990\_1999\_summary\_table.csv")  
sum\_2000\_2009\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/2000\_2009/2000\_2009\_summary\_table.csv")  
sum\_2010\_2020\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/2010\_2020/2010\_2020\_summary\_table.csv")  
  
#write\_summary table vectors for each decade  
tot\_1880\_1889\_summary<-c("1880-1889",sum\_1880\_1889\_df[10,3],sum(sum\_1880\_1889\_df[6]), sum(sum\_1880\_1889\_df[7]),sum(sum\_1880\_1889\_df[6])/sum(sum\_1880\_1889\_df[7]),sum(sum\_1880\_1889\_df[9]),sum(sum\_1880\_1889\_df[10]), sum(sum\_1880\_1889\_df[9])/sum(sum\_1880\_1889\_df[10]),sum(sum\_1880\_1889\_df[12]),sum(sum\_1880\_1889\_df[13]),sum(sum\_1880\_1889\_df[12])/sum(sum\_1880\_1889\_df[13]))  
tot\_1890\_1899\_summary<-c("1890-1899",sum\_1890\_1899\_df[10,3],sum(sum\_1890\_1899\_df[6]),sum(sum\_1890\_1899\_df[7]), sum(sum\_1890\_1899\_df[6])/sum(sum\_1890\_1899\_df[7]),sum(sum\_1890\_1899\_df[9]),sum(sum\_1890\_1899\_df[10]), sum(sum\_1890\_1899\_df[9])/sum(sum\_1890\_1899\_df[10]),sum(sum\_1890\_1899\_df[12]),sum(sum\_1890\_1899\_df[13]),sum(sum\_1890\_1899\_df[12])/sum(sum\_1890\_1899\_df[13]))  
tot\_1900\_1909\_summary<-c("1900-1909",sum\_1900\_1909\_df[10,3],sum(sum\_1900\_1909\_df[6]),sum(sum\_1900\_1909\_df[7]), sum(sum\_1900\_1909\_df[6])/sum(sum\_1900\_1909\_df[7]),sum(sum\_1900\_1909\_df[9]),sum(sum\_1900\_1909\_df[10]), sum(sum\_1900\_1909\_df[9])/sum(sum\_1900\_1909\_df[10]),sum(sum\_1900\_1909\_df[12]),sum(sum\_1900\_1909\_df[13]),sum(sum\_1900\_1909\_df[12])/sum(sum\_1900\_1909\_df[13]))  
tot\_1910\_1919\_summary<-c("1910-1919",sum\_1910\_1919\_df[10,3],sum(sum\_1910\_1919\_df[6]),sum(sum\_1910\_1919\_df[7]), sum(sum\_1910\_1919\_df[6])/sum(sum\_1910\_1919\_df[7]),sum(sum\_1910\_1919\_df[9]),sum(sum\_1910\_1919\_df[10]), sum(sum\_1910\_1919\_df[9])/sum(sum\_1910\_1919\_df[10]),sum(sum\_1910\_1919\_df[12]),sum(sum\_1910\_1919\_df[13]),sum(sum\_1910\_1919\_df[12])/sum(sum\_1910\_1919\_df[13]))  
tot\_1920\_1929\_summary<-c("1920-1929",sum\_1920\_1929\_df[10,3],sum(sum\_1920\_1929\_df[6]),sum(sum\_1920\_1929\_df[7]),sum(sum\_1920\_1929\_df[6])/sum(sum\_1920\_1929\_df[7]),sum(sum\_1920\_1929\_df[9]),sum(sum\_1920\_1929\_df[10]), sum(sum\_1920\_1929\_df[9])/sum(sum\_1920\_1929\_df[10]),sum(sum\_1920\_1929\_df[12]),sum(sum\_1920\_1929\_df[13]),sum(sum\_1920\_1929\_df[12])/sum(sum\_1920\_1929\_df[13]))  
tot\_1930\_1939\_summary<-c("1930-1939",sum\_1930\_1939\_df[10,3],sum(sum\_1930\_1939\_df[6]),sum(sum\_1930\_1939\_df[7]), sum(sum\_1930\_1939\_df[6])/sum(sum\_1930\_1939\_df[7]),sum(sum\_1930\_1939\_df[9]),sum(sum\_1930\_1939\_df[10]), sum(sum\_1930\_1939\_df[9])/sum(sum\_1930\_1939\_df[10]),sum(sum\_1930\_1939\_df[12]),sum(sum\_1930\_1939\_df[13]),sum(sum\_1930\_1939\_df[12])/sum(sum\_1930\_1939\_df[13]))  
tot\_1940\_1949\_summary<-c("1940-1949",sum\_1940\_1949\_df[10,3],sum(sum\_1940\_1949\_df[6]),sum(sum\_1940\_1949\_df[7]), sum(sum\_1940\_1949\_df[6])/sum(sum\_1940\_1949\_df[7]),sum(sum\_1940\_1949\_df[9]),sum(sum\_1940\_1949\_df[10]),sum(sum\_1940\_1949\_df[9])/sum(sum\_1940\_1949\_df[10]),sum(sum\_1940\_1949\_df[12]),sum(sum\_1940\_1949\_df[13]),sum(sum\_1940\_1949\_df[12])/sum(sum\_1940\_1949\_df[13]))  
tot\_1950\_1959\_summary<-c("1950-1959",sum\_1950\_1959\_df[10,3],sum(sum\_1950\_1959\_df[6]),sum(sum\_1950\_1959\_df[7]),sum(sum\_1950\_1959\_df[6])/sum(sum\_1950\_1959\_df[7]),sum(sum\_1950\_1959\_df[9]),sum(sum\_1950\_1959\_df[10]),sum(sum\_1950\_1959\_df[9])/sum(sum\_1950\_1959\_df[10]),sum(sum\_1950\_1959\_df[12]),sum(sum\_1950\_1959\_df[13]),sum(sum\_1950\_1959\_df[12])/sum(sum\_1950\_1959\_df[13]))  
tot\_1960\_1969\_summary<-c("1960-1969",sum\_1960\_1969\_df[10,3],sum(sum\_1960\_1969\_df[6]),sum(sum\_1960\_1969\_df[7]),sum(sum\_1960\_1969\_df[6])/sum(sum\_1960\_1969\_df[7]),sum(sum\_1960\_1969\_df[9]),sum(sum\_1960\_1969\_df[10]),sum(sum\_1960\_1969\_df[9])/sum(sum\_1960\_1969\_df[10]),sum(sum\_1960\_1969\_df[12]),sum(sum\_1960\_1969\_df[13]),sum(sum\_1960\_1969\_df[12])/sum(sum\_1960\_1969\_df[13]))  
tot\_1970\_1979\_summary<-c("1970-1979",sum\_1970\_1979\_df[10,3],sum(sum\_1970\_1979\_df[6]),sum(sum\_1970\_1979\_df[7]),sum(sum\_1970\_1979\_df[6])/sum(sum\_1970\_1979\_df[7]),sum(sum\_1970\_1979\_df[9]),sum(sum\_1970\_1979\_df[10]),sum(sum\_1970\_1979\_df[9])/sum(sum\_1970\_1979\_df[10]),sum(sum\_1970\_1979\_df[12]),sum(sum\_1970\_1979\_df[13]),sum(sum\_1970\_1979\_df[12])/sum(sum\_1970\_1979\_df[13]))  
tot\_1980\_1989\_summary<-c("1980-1989",sum\_1980\_1989\_df[10,3],sum(sum\_1980\_1989\_df[6]),sum(sum\_1980\_1989\_df[7]),sum(sum\_1980\_1989\_df[6])/sum(sum\_1980\_1989\_df[7]),sum(sum\_1980\_1989\_df[9]),sum(sum\_1980\_1989\_df[10]),sum(sum\_1980\_1989\_df[9])/sum(sum\_1980\_1989\_df[10]),sum(sum\_1980\_1989\_df[12]),sum(sum\_1980\_1989\_df[13]),sum(sum\_1980\_1989\_df[12])/sum(sum\_1980\_1989\_df[13]))  
tot\_1990\_1999\_summary<-c("1990-1999",sum\_1990\_1999\_df[10,3],sum(sum\_1990\_1999\_df[6]),sum(sum\_1990\_1999\_df[7]),sum(sum\_1990\_1999\_df[6])/sum(sum\_1990\_1999\_df[7]),sum(sum\_1990\_1999\_df[9]),sum(sum\_1990\_1999\_df[10]),sum(sum\_1990\_1999\_df[9])/sum(sum\_1990\_1999\_df[10]),sum(sum\_1990\_1999\_df[12]),sum(sum\_1990\_1999\_df[13]),sum(sum\_1990\_1999\_df[12])/sum(sum\_1990\_1999\_df[13]))  
tot\_2000\_2009\_summary<-c("2000-2009",sum\_2000\_2009\_df[10,3],sum(sum\_2000\_2009\_df[6]),sum(sum\_2000\_2009\_df[7]),sum(sum\_2000\_2009\_df[6])/sum(sum\_2000\_2009\_df[7]),sum(sum\_2000\_2009\_df[9]),sum(sum\_2000\_2009\_df[10]),sum(sum\_2000\_2009\_df[9])/sum(sum\_2000\_2009\_df[10]),sum(sum\_2000\_2009\_df[12]),sum(sum\_2000\_2009\_df[13]),sum(sum\_2000\_2009\_df[12])/sum(sum\_2000\_2009\_df[13]))  
tot\_2010\_2020\_summary<-c("2010-2020",sum\_2010\_2020\_df[10,3],sum(sum\_2010\_2020\_df[6]),sum(sum\_2010\_2020\_df[7]),sum(sum\_2010\_2020\_df[6])/sum(sum\_2010\_2020\_df[7]),sum(sum\_2010\_2020\_df[9]),sum(sum\_2010\_2020\_df[10]),sum(sum\_2010\_2020\_df[9])/sum(sum\_2010\_2020\_df[10]),sum(sum\_2010\_2020\_df[12]),sum(sum\_2010\_2020\_df[13]),sum(sum\_2010\_2020\_df[12])/sum(sum\_2010\_2020\_df[13]))  
#--- Assemble table for all decades  
decades\_table<-t(data.frame(tot\_1880\_1889\_summary,tot\_1890\_1899\_summary,tot\_1900\_1909\_summary,tot\_1910\_1919\_summary,tot\_1920\_1929\_summary,tot\_1930\_1939\_summary,tot\_1940\_1949\_summary,tot\_1950\_1959\_summary,tot\_1960\_1969\_summary,tot\_1970\_1979\_summary,tot\_1980\_1989\_summary,tot\_1990\_1999\_summary,tot\_2000\_2009\_summary,tot\_2010\_2020\_summary))  
colnames(decades\_table)<-c("Years","# of GN Names","F occurrances of GN names","Total F occurrances","Ratio of F GN Occur to Total F Occur","M occurrances of GN names","Total M occurrances","Ratio of M GN Occur to Total M Occur","Total GN Occurrances","Total Occurrances","Ratio of GN Occurrances to Total Occurrances")  
rownames(decades\_table)<-NULL  
#save to data file  
write.csv(x=decades\_table,file=("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/all\_decades\_summary\_table.csv"))

The All Decades Summary Table that weas saved to CSV above is shown here. Each row is wrapper around due to the size of the table.

head(decades\_table,20)

## Years # of GN Names F occurrances of GN names Total F occurrances  
## [1,] "1880-1889" "140" "683748" "1301480"   
## [2,] "1890-1899" "143" "1180592" "2171767"   
## [3,] "1900-1909" "134" "1442681" "2831578"   
## [4,] "1910-1919" "116" "2074047" "7748618"   
## [5,] "1920-1929" "138" "4552460" "11300635"   
## [6,] "1930-1939" "127" "4305909" "10133750"   
## [7,] "1940-1949" "107" "4115658" "13800142"   
## [8,] "1950-1959" "92" "5287603" "18224976"   
## [9,] "1960-1969" "96" "4170675" "16973850"   
## [10,] "1970-1979" "93" "2697790" "13647127"   
## [11,] "1980-1989" "94" "4789294" "14816745"   
## [12,] "1990-1999" "75" "1364492" "14785246"   
## [13,] "2000-2009" "69" "1038179" "14098356"   
## [14,] "2010-2020" "85" "943084" "14104545"   
## Ratio of F GN Occur to Total F Occur M occurrances of GN names  
## [1,] "0.525361895687986" "585762"   
## [2,] "0.543608959893027" "487537"   
## [3,] "0.509497177898684" "558825"   
## [4,] "0.267666698758411" "3096960"   
## [5,] "0.402849928344735" "6217959"   
## [6,] "0.424907758727026" "5064778"   
## [7,] "0.298233018181987" "6545466"   
## [8,] "0.290129490431153" "7525882"   
## [9,] "0.245711786070927" "7996958"   
## [10,] "0.19768190037361" "5937839"   
## [11,] "0.323235231489777" "4413902"   
## [12,] "0.0922874059721428" "1793626"   
## [13,] "0.0736383022247417" "1419144"   
## [14,] "0.0668638371532013" "1676335"   
## Total M occurrances Ratio of M GN Occur to Total M Occur  
## [1,] "1091006" "0.536900805311795"   
## [2,] "1129420" "0.431670237821183"   
## [3,] "1331528" "0.419687006206403"   
## [4,] "6358747" "0.48703934910447"   
## [5,] "10501234" "0.592116983584977"   
## [6,] "10159661" "0.49851840528931"   
## [7,] "14477966" "0.452098450845927"   
## [8,] "19722553" "0.38158761697839"   
## [9,] "18732541" "0.426901934980417"   
## [10,] "15728444" "0.377522341052936"   
## [11,] "17358610" "0.254277387417541"   
## [12,] "17728340" "0.101172811442019"   
## [13,] "17299564" "0.0820335125208936"   
## [14,] "17132104" "0.0978475848617309"   
## Total GN Occurrances Total Occurrances  
## [1,] "1269510" "2392486"   
## [2,] "1668129" "3301187"   
## [3,] "2001506" "4163106"   
## [4,] "5171007" "14107365"   
## [5,] "10770419" "21801869"   
## [6,] "9370687" "20293411"   
## [7,] "10661124" "28278108"   
## [8,] "12813485" "37947529"   
## [9,] "12167633" "35706391"   
## [10,] "8635629" "29375571"   
## [11,] "9203196" "32175355"   
## [12,] "3158118" "32513586"   
## [13,] "2457323" "31397920"   
## [14,] "2619419" "31236649"   
## Ratio of GN Occurrances to Total Occurrances  
## [1,] "0.530623794663793"   
## [2,] "0.505311877212651"   
## [3,] "0.480772288767089"   
## [4,] "0.366546622987354"   
## [5,] "0.494013563699516"   
## [6,] "0.461760075721129"   
## [7,] "0.377009805606514"   
## [8,] "0.337663224395981"   
## [9,] "0.340769051680412"   
## [10,] "0.293973145236904"   
## [11,] "0.286032461801898"   
## [12,] "0.0971322572662394"   
## [13,] "0.0782638786263549"   
## [14,] "0.0838572344940073"

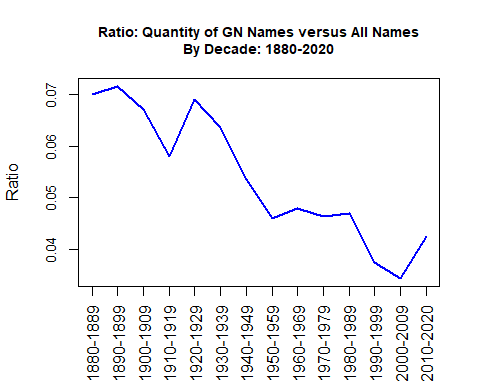
6. Plot Quantity and Occurrences of GN versus All Names

Shawn Behrend

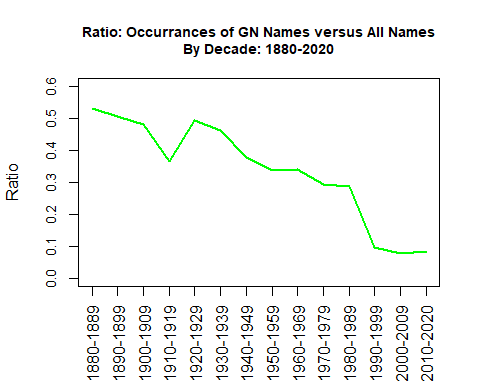
2022-05-07

This code produces two plots. The first is the Quantity of GN Names versus All Names. The second is the Occurrences of GN Names versus All Names.

sum\_all\_decades\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/all\_decades\_summary\_table.csv")  
Ratio.of.GN.Names.to.All.Names<-sum\_all\_decades\_df$X..of.GN.Names/2000  
q\_test\_set<-data.frame(sum\_all\_decades\_df$Years,Ratio.of.GN.Names.to.All.Names,sum\_all\_decades\_df$Ratio.of.GN.Occurrances.to.Total.Occurrances)  
xaxis\_labels<-unlist(q\_test\_set[1])  
xaxis<-c(1:length(xaxis\_labels))  
  
#plot quantity of names  
ratio\_quan\_gn\_to\_all<-as.numeric(unlist(q\_test\_set[2]))  
ratio\_occur\_gn\_to\_all<-as.numeric(unlist(q\_test\_set[3]))  
  
#plot ratio of quantity  
plot(xaxis,ratio\_quan\_gn\_to\_all,main="Ratio: Quantity of GN Names versus All Names\nBy Decade: 1880-2020", xlab="",xaxt="n",ylab="Ratio",type="l",lwd=2,col="blue",cex.main=0.9,cex.axis=0.8)  
axis(1, at=seq\_along(xaxis),labels=as.character(xaxis\_labels), las=2,cex=0.6)



#plot ratio of occurrences  
plot(xaxis,ratio\_occur\_gn\_to\_all,main="Ratio: Occurrances of GN Names versus All Names\nBy Decade: 1880-2020",xlab="",xaxt="n",ylab="Ratio",type="l",lwd=2,col="green",ylim=c(0,0.6),cex.main=0.9,cex.axis=0.8)  
axis(1, at=seq\_along(xaxis),labels=as.character(xaxis\_labels), las=2,cex=0.6)



7. All GN Names

Shawn Behrend

2022-05-07

This code compiles all gender neutral names in each comp\_gn\_names file to create a long list of all gender neutral names in every decade along with the accompanying year.

#load data files  
gn\_1880\_1889\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1880\_1889/comp\_gn\_names\_F\_1889.csv")  
gn\_1890\_1899\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1890\_1899/comp\_gn\_names\_F\_1899.csv")  
gn\_1900\_1909\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1900\_1909/comp\_gn\_names\_F\_1909.csv")  
gn\_1910\_1919\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1910\_1919/comp\_gn\_names\_F\_1919.csv")  
gn\_1920\_1929\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1920\_1929/comp\_gn\_names\_F\_1929.csv")  
gn\_1930\_1939\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1930\_1939/comp\_gn\_names\_F\_1939.csv")  
gn\_1940\_1949\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1940\_1949/comp\_gn\_names\_F\_1949.csv")  
gn\_1950\_1959\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1950\_1959/comp\_gn\_names\_F\_1959.csv")  
gn\_1960\_1969\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1960\_1969/comp\_gn\_names\_F\_1969.csv")  
gn\_1970\_1979\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1970\_1979/comp\_gn\_names\_F\_1979.csv")  
gn\_1980\_1989\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1980\_1989/comp\_gn\_names\_F\_1989.csv")  
gn\_1990\_1999\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/1990\_1999/comp\_gn\_names\_F\_1999.csv")  
gn\_2000\_2009\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/2000\_2009/comp\_gn\_names\_F\_2009.csv")  
gn\_2010\_2020\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/2010\_2020/comp\_gn\_names\_F\_2020.csv")  
  
# save name and year columns only  
gn\_1889<-data.frame(c(gn\_1880\_1889\_df[2],gn\_1880\_1889\_df[6]))  
gn\_1899<-data.frame(c(gn\_1890\_1899\_df[2],gn\_1890\_1899\_df[6]))  
gn\_1909<-data.frame(c(gn\_1900\_1909\_df[2],gn\_1900\_1909\_df[6]))  
gn\_1919<-data.frame(c(gn\_1910\_1919\_df[2],gn\_1910\_1919\_df[6]))  
gn\_1929<-data.frame(c(gn\_1920\_1929\_df[2],gn\_1920\_1929\_df[6]))  
gn\_1939<-data.frame(c(gn\_1930\_1939\_df[2],gn\_1930\_1939\_df[6]))  
gn\_1949<-data.frame(c(gn\_1940\_1949\_df[2],gn\_1940\_1949\_df[6]))  
gn\_1959<-data.frame(c(gn\_1950\_1959\_df[2],gn\_1950\_1959\_df[6]))  
gn\_1969<-data.frame(c(gn\_1960\_1969\_df[2],gn\_1960\_1969\_df[6]))  
gn\_1979<-data.frame(c(gn\_1970\_1979\_df[2],gn\_1970\_1979\_df[6]))  
gn\_1989<-data.frame(c(gn\_1980\_1989\_df[2],gn\_1980\_1989\_df[6]))  
gn\_1999<-data.frame(c(gn\_1990\_1999\_df[2],gn\_1990\_1999\_df[6]))  
gn\_2009<-data.frame(c(gn\_2000\_2009\_df[2],gn\_2000\_2009\_df[6]))  
gn\_2020<-data.frame(c(gn\_2010\_2020\_df[2],gn\_2010\_2020\_df[6]))  
  
# row bind all names/years together  
gn\_all<-rbind(gn\_1889,gn\_1899,gn\_1909,gn\_1919,gn\_1929,gn\_1939,gn\_1949,gn\_1959,gn\_1969,gn\_1979,gn\_1989,gn\_1999,gn\_2009,gn\_2020)  
  
#write the appended file   
write.csv(gn\_all,"C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/gn\_all.csv")

An excerpt of the head and tail of the CSV file that was written, which shows all names along with the year that matched is shown here.

head(gn\_all,10)

## Name Year  
## 1 Ada 1889  
## 2 Addie 1889  
## 3 Alice 1889  
## 4 Allie 1889  
## 5 Alma 1889  
## 6 Alpha 1889  
## 7 Alta 1889  
## 8 Alva 1889  
## 9 Anna 1889  
## 10 Annie 1889

tail(gn\_all,10)

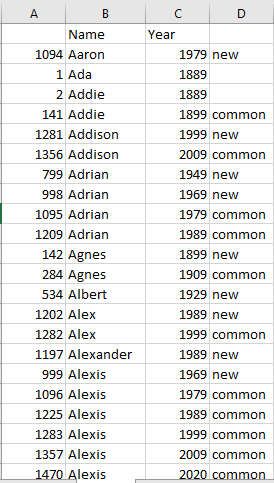
## Name Year  
## 1500 Genesis 2020  
## 1501 Kamryn 2020  
## 1502 Robin 2020  
## 1503 Dior 2020  
## 1504 London 2020  
## 1505 Marley 2020  
## 1506 Landry 2020  
## 1507 Lyric 2020  
## 1508 Murphy 2020  
## 1509 Frankie 2020

8. Plot New and Common GN Names

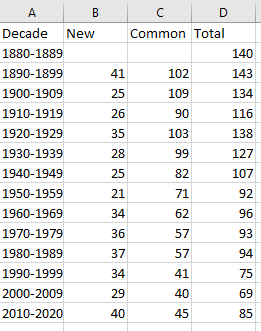
Shawn Behrend

2022-05-07

After creation of the gn\_all.csv which contains all the gender-neutral names along with the last year of the decade they are a part of, this CSV is then processed in Microsoft Excel. It is sorted by Name and then by Year to determine if each name is new within a decade or was common with the previous decade. An excerpt is shown here. Names from 1889 are not considered to be new or common since there is no previous data.

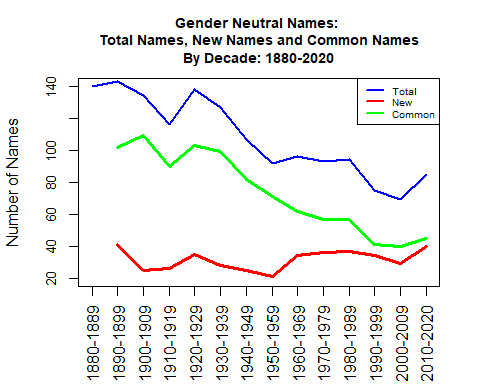


Once that is complete, the count of how many names were new per decade versus common with the previous decade was done. This is shown here.



Next those counts are plotted per decade.

#load data files  
new\_common\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/gn\_all\_with\_status.csv")  
xaxis\_labels<-unlist(new\_common\_df[1])  
xaxis<-c(1:length(xaxis\_labels))  
new<-as.numeric(unlist(new\_common\_df[2]))  
common<-as.numeric(unlist(new\_common\_df[3]))  
total<-as.numeric(unlist(new\_common\_df[4]))  
  
# x axis for all plots  
  
#plot first data line  
plot(xaxis,total,main="Gender Neutral Names: \nTotal Names, New Names and Common Names\nBy Decade: 1880-2020",xlab="",xaxt="n",ylab="Number of Names",type="l",lwd=2,col="blue",ylim=c(20,140),cex.main=0.9,cex.axis=0.8)  
# plot additional lines  
lines(xaxis,new,lwd=3,col="red")  
lines(xaxis,common,lwd=3,col="green")  
axis(1, at=seq\_along(total),labels=as.character(xaxis\_labels), las=2,cex=0.6)  
# legend  
legend("topright",legend=c("Total","New", "Common"),cex=0.6,col=c("blue", "red", "green"),lwd=2,lty=1)



9. Correlation Test and Plots

Shawn Behrend

2022-05-07

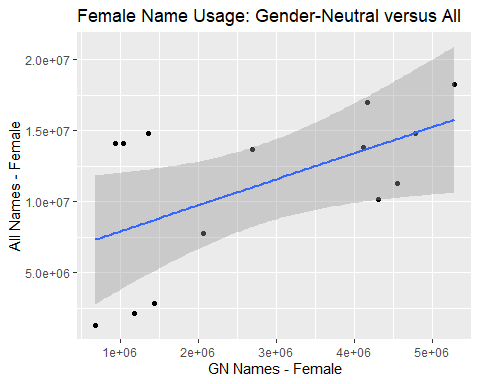
Using the All Decades Summary Table created in a previous step, correlation tests were run for the occurrences of gender neutral names versus the total occurrences of any name, for both males and females.

# Load ggplot2  
library("ggplot2")

## Registered S3 methods overwritten by 'tibble':  
## method from   
## format.tbl pillar  
## print.tbl pillar

sum\_all\_decades\_df<-read.csv("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/all\_decades\_summary\_table.csv")  
  
#correlation for female names  
f\_test\_set<-data.frame(sum\_all\_decades\_df$Years,sum\_all\_decades\_df$F.occurrances.of.GN.names,sum\_all\_decades\_df$Total.F.occurrances)  
f\_gn\_names<-as.numeric(unlist(f\_test\_set[2]))  
f\_all\_names<-as.numeric(unlist(f\_test\_set[3]))  
f\_gn\_all<-data.frame(f\_gn\_names,f\_all\_names)  
# correlation tests  
f\_gn\_scatter\_plot <- ggplot(f\_gn\_all, aes(f\_gn\_names, f\_all\_names))  
f\_gn\_scatter\_plot + geom\_point() + labs(title="Female Name Usage: Gender-Neutral versus All",x = "GN Names - Female", y = "All Names - Female") + geom\_smooth(method="lm")

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

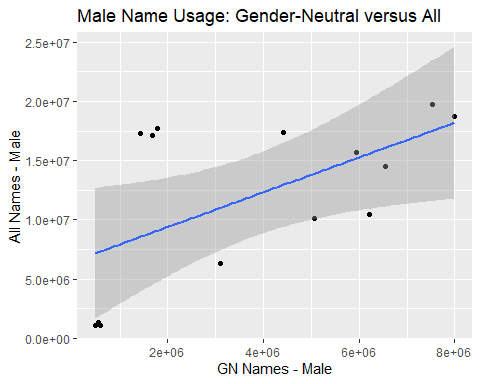


cortest\_f<-cor.test(f\_gn\_names, f\_all\_names, method = "pearson")  
  
sink("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/f\_gn\_names\_corr.txt")  
print(cortest\_f)

##   
## Pearson's product-moment correlation  
##   
## data: f\_gn\_names and f\_all\_names  
## t = 2.3283, df = 12, p-value = 0.03819  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.03870847 0.83984009  
## sample estimates:  
## cor   
## 0.5578313

sink() # returns output to the console  
#----  
#correlation for male names  
m\_test\_set<-data.frame(sum\_all\_decades\_df$Years,sum\_all\_decades\_df$M.occurrances.of.GN.names,sum\_all\_decades\_df$Total.M.occurrances)  
m\_gn\_names<-as.numeric(unlist(m\_test\_set[2]))  
m\_all\_names<-as.numeric(unlist(m\_test\_set[3]))  
m\_gn\_all<-data.frame(m\_gn\_names,m\_all\_names)  
# correlation tests  
m\_gn\_scatter\_plot <- ggplot(m\_gn\_all, aes(m\_gn\_names, m\_all\_names))  
m\_gn\_scatter\_plot + geom\_point() + labs(title="Male Name Usage: Gender-Neutral versus All",x = "GN Names - Male", y = "All Names - Male") + geom\_smooth(method="lm")

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



cortest\_m<-cor.test(m\_gn\_names, m\_all\_names, method = "pearson")  
#---  
sink("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/f\_gn\_names\_corr.txt")  
print(cortest\_f)

##   
## Pearson's product-moment correlation  
##   
## data: f\_gn\_names and f\_all\_names  
## t = 2.3283, df = 12, p-value = 0.03819  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.03870847 0.83984009  
## sample estimates:  
## cor   
## 0.5578313

sink() # returns output to the console  
  
sink("C:/Users/shawn/OneDrive/Shawn/CSU\_global/MIS581/project\_r\_code/name\_files/m\_gn\_names\_corr.txt")  
print(cortest\_m)

##   
## Pearson's product-moment correlation  
##   
## data: m\_gn\_names and m\_all\_names  
## t = 2.4599, df = 12, p-value = 0.03004  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.06986763 0.84881121  
## sample estimates:  
## cor   
## 0.578984

sink() # returns output to the console