Assignment3\_08142015

Yan Li

August 14 2015

# Load packages, data and set the working library:

.libPaths("C:/work/Rpackages/R3.2.1")  
suppressMessages(library(ggplot2))  
suppressMessages(library(dplyr))  
suppressMessages(library(mosaicData))  
suppressMessages(library(babynames))  
suppressMessages(library(fastR))  
suppressMessages(library(tidyr))  
suppressMessages(library(ggvis))  
  
#turn off warning messages  
options(warn=-1)

# Define some functions

letter <- function(x, start=1, stop=start) {  
 # convert negative numbers to positions relative to end of string  
 n <- c(start,stop)  
 n <- ifelse( n < 0, nchar(x) + n + 1, n )  
 tolower(substr(x, min(n), max(n)))  
}  
  
vowels <- function(x) {  
 nchar(gsub("[^aeiouy]", "", tolower(x)))  
}

# Data: babynames

head(babynames)

## Source: local data frame [6 x 5]  
##   
## year sex name n prop  
## 1 1880 F Mary 7065 0.07238359  
## 2 1880 F Anna 2604 0.02667896  
## 3 1880 F Emma 2003 0.02052149  
## 4 1880 F Elizabeth 1939 0.01986579  
## 5 1880 F Minnie 1746 0.01788843  
## 6 1880 F Margaret 1578 0.01616720

# Problems

## 1 Which names were most popular in 1999?

First this need two fields information: Year=1999 and name.

Secondly it needs to reorder the counts and find the max.

#subset data by filter()  
Bnames1999<- babynames %>% filter(year == "1999")  
#reorder the dataframe  
arrange(Bnames1999,-Bnames1999$n)

## Source: local data frame [28,543 x 5]  
##   
## year sex name n prop  
## 1 1999 M Jacob 35344 0.01734586  
## 2 1999 M Michael 33900 0.01663719  
## 3 1999 M Matthew 30413 0.01492586  
## 4 1999 M Joshua 27249 0.01337306  
## 5 1999 F Emily 26535 0.01363875  
## 6 1999 M Nicholas 25633 0.01257997  
## 7 1999 M Christopher 25601 0.01256427  
## 8 1999 M Andrew 23847 0.01170345  
## 9 1999 M Joseph 23201 0.01138641  
## 10 1999 M Daniel 22667 0.01112434  
## .. ... ... ... ... ...

Thus the top 5 popular names in 1999 are Jacob, Michael,Mattew,Josha,Emily.

## 2 Which names have the highest usage (on average)?

Group data by name and average the use (number of person use / number of years).

BnamesByName <-  
babynames %>%  
group\_by(name) %>% # group by name   
summarise(sum.n = sum(n),  
 sum.year = length(unique(year)),  
 avg.n = sum.n/sum.year  
 ) %>%  
arrange(-avg.n)   
  
head(BnamesByName,10)

## Source: local data frame [10 x 4]  
##   
## name sum.n sum.year avg.n  
## 1 James 5114325 134 38166.60  
## 2 John 5095590 134 38026.79  
## 3 Robert 4809858 134 35894.46  
## 4 Michael 4315029 134 32201.71  
## 5 Mary 4127615 134 30803.10  
## 6 William 4054318 134 30256.10  
## 7 David 3578068 134 26702.00  
## 8 Joseph 2568379 134 19167.01  
## 9 Richard 2561839 134 19118.20  
## 10 Charles 2369238 134 17680.88

So the highest used names are: James,John,Robert, Micheal, Mary.

## 3 For each name and sex combination, find the year in which the name had its greatest popularity and its rank in that year. Print all names that have been the most popular name in at least one year.

First we need to group data by year and rank by proportion. The rank ==1 is the greatest popular one. Next we group the above data by name and sex and find the min rank. Last we filter the data by rank ==1.

BnamesByYearNameSex <-  
babynames %>%  
group\_by(year)%>% # group by name year  
mutate(rank = rank(prop, ties.method = "first"))%>% # missing values will be put last  
arrange(rank) %>%  
select(year, name, sex, rank) %>%  
group\_by(name, sex)%>%  
summarise(top.rank = min(rank) ,  
 top.year = year[which(rank==top.rank)][1]) %>%  
filter(top.rank == 1)  
   
#head(BnamesByYearNameSex, 10)  
   
BnamesByYearNameSex$name

## [1] "Aaban" "Aabid" "Aadam" "Aadarsh" "Aadesh" "Aadhav"   
## [7] "Aadhavan" "Aadil" "Aadin" "Aadith" "Aadrit" "Aagot"   
## [13] "Aahil" "Aakarsh" "Aakash" "Aakeem" "Aalok" "Aalon"   
## [19] "Aamer" "Aamil" "Aanand" "Aaqil" "Aaran" "Aaric"   
## [25] "Aarik" "Aarin" "Aaron" "Aarron" "Aashish" "Ab"   
## [31] "Abayomi" "Abb" "Abbey" "Abbey" "Abbie" "Abbott"   
## [37] "Abby" "Abby" "Abdalla" "Abdallah" "Abdel" "Abdon"   
## [43] "Abdul" "Abdullah" "Abelina" "Abelino" "Abie" "Able"   
## [49] "Abrahan" "Abundio" "Achille" "Achsa" "Achsah" "Acie"   
## [55] "Adabelle" "Adair" "Adala" "Adalaide" "Adalee" "Adalene"   
## [61] "Adalia" "Adalyn" "Adamae" "Adante" "Add" "Addis"   
## [67] "Addis" "Adel" "Adelaida" "Adelma" "Adelyn" "Ader"   
## [73] "Adina" "Adlena" "Adra" "Adrien" "Adrienne" "Agness"   
## [79] "Agusta" "Aida"

## 4 What name has been in the top 10 most often? (Hint: you may want to think about doing this in multiple steps. Think about what they should be before you start coding.)

First we group data by year and rank the names based on their proportions, so we will have a rank field to record this and the smaller the number the more popular. Secondly, we group the data by name and summerise the rank and sort the sum. Again, the smaller of the sum means the more popular of the names.

#Group data by year and get rank field  
#Filter data by rank <=10: find the top 10 popular names  
BnamesByYearRank <-  
babynames %>%  
group\_by(year)%>% # group by name year  
mutate(rank = rank(prop, ties.method = "first"))%>% # missing values will be put last  
#arrange(rank) %>%  
 filter(rank <=10)  
  
head(BnamesByYearRank,10)

## Source: local data frame [10 x 6]  
## Groups: year  
##   
## year sex name n prop rank  
## 1 1880 M Ab 5 4.222973e-05 1  
## 2 1880 M Abbott 5 4.222973e-05 2  
## 3 1880 M Agustus 5 4.222973e-05 3  
## 4 1880 M Albertus 5 4.222973e-05 4  
## 5 1880 M Almer 5 4.222973e-05 5  
## 6 1880 M Alphonso 5 4.222973e-05 6  
## 7 1880 M Alvia 5 4.222973e-05 7  
## 8 1880 M Artie 5 4.222973e-05 8  
## 9 1880 M Arvid 5 4.222973e-05 9  
## 10 1880 M Ashby 5 4.222973e-05 10

#The number of names appear in the above data == the number of years that name appear in top 10  
#Get the top 10 most popular names  
  
tail(sort(table(BnamesByYearRank$name)),10)

##   
## Abundio Adelaido Adrienne Afton Agusta Adlene Adair Addis   
## 9 9 9 9 10 11 12 12   
## Abbott Abby   
## 13 16

## 5 What name has had the biggest increase in popularity from one year to the next? What about the biggest decrease? In what years did these changes occur? Repeat for both boys and girls Is it better to use rank or proportion for this task? Why?

It is better to use rank other than proportion since proportion may have missing values. use diff(lag = 1) to claculate the change of rank over year.

BnamesByYearRankDiff <-  
babynames %>%  
group\_by(year)%>% # group by name year  
mutate(rank = rank(prop, ties.method = "first"))%>% # missing values will be put last  
group\_by(name) %>%  
 arrange(year) %>%  
mutate(rank.diff= c(0,diff(rank)))%>%  
summarise(max.diff = max(rank.diff),  
 max.year = year[which(rank.diff == max.diff)][1],  
 min.diff = min(rank.diff),  
 min.year = year[which(rank.diff == min.diff)][1])  
   
   
head(BnamesByYearRankDiff %>% arrange(-max.diff))

## Source: local data frame [6 x 5]  
##   
## name max.diff max.year min.diff min.year  
## 1 Allison 34860 2008 -34719 2007  
## 2 Anna 34744 2007 -33892 2006  
## 3 Amelia 34664 2007 -34425 2009  
## 4 Adrianna 34537 2007 -33662 2006  
## 5 Alexandria 34516 2007 -33684 2006  
## 6 Alejandra 34462 2008 -34359 2007

head(BnamesByYearRankDiff %>% arrange(min.diff))

## Source: local data frame [6 x 5]  
##   
## name max.diff max.year min.diff min.year  
## 1 Arianna 34398 2009 -34741 2008  
## 2 Allison 34860 2008 -34719 2007  
## 3 Bella 34309 2009 -34570 2008  
## 4 Abby 34147 2009 -34518 2008  
## 5 Amelia 34664 2007 -34425 2009  
## 6 Alejandra 34462 2008 -34359 2007

The biggest increase of popularity is for name : Allison for year 2007-2008. The biggest decrease of popularity is for name : Arianna for year 2007-2008.

BnamesBoysDiff <-  
babynames %>%  
group\_by(year)%>% # group by name year  
mutate(rank = rank(prop, ties.method = "first"))%>% # missing values will be put last  
filter(sex == "M") %>%  
 group\_by(name) %>%  
 arrange(year) %>%  
mutate(rank.diff= c(0,diff(rank)))%>%  
summarise(max.diff = max(rank.diff),  
 max.year = year[which(rank.diff == max.diff)][1],  
 min.diff = min(rank.diff),  
 min.year = year[which(rank.diff == min.diff)][1])  
  
   
head(BnamesBoysDiff %>% arrange(-max.diff))

## Source: local data frame [6 x 5]  
##   
## name max.diff max.year min.diff min.year  
## 1 Barack 29006 2008 -5230 2011  
## 2 Kipton 28441 2009 -2651 1986  
## 3 Omarion 28111 2002 -1225 2009  
## 4 Jakwon 26893 2004 -12723 2003  
## 5 Kanye 26651 2003 -4186 2010  
## 6 Renner 26425 2007 -7264 2005

head(BnamesBoysDiff %>% arrange(min.diff))

## Source: local data frame [6 x 5]  
##   
## name max.diff max.year min.diff min.year  
## 1 Brasen 10270 2013 -25374 2007  
## 2 Kaitlyn 15441 2006 -24065 2005  
## 3 Autumn 13429 2004 -23490 2005  
## 4 Jurem 0 2008 -23313 2009  
## 5 Khale 23169 2010 -22814 2009  
## 6 Izaiha 10166 1997 -22703 2012

BnamesGirlsDiff <-  
babynames %>%  
group\_by(year)%>% # group by name year  
mutate(rank = rank(prop, ties.method = "first"))%>% # missing values will be put last  
filter(sex == "F") %>%  
 group\_by(name) %>%  
 arrange(year) %>%  
mutate(rank.diff= c(0,diff(rank)))%>%  
summarise(max.diff = max(rank.diff),  
 max.year = year[which(rank.diff == max.diff)][1],  
 min.diff = min(rank.diff),  
 min.year = year[which(rank.diff == min.diff)][1])  
  
   
head(BnamesGirlsDiff %>% arrange(-max.diff))

## Source: local data frame [6 x 5]  
##   
## name max.diff max.year min.diff min.year  
## 1 Jaslene 30650 2007 -5111 1998  
## 2 Allysson 27846 2008 -3737 2004  
## 3 Colie 27680 2007 -11536 2010  
## 4 Akeelah 26980 2006 -5202 1995  
## 5 Baya 26838 2009 -3403 2008  
## 6 Ambrielle 26695 2007 -14712 2003

head(BnamesGirlsDiff %>% arrange(min.diff))

## Source: local data frame [6 x 5]  
##   
## name max.diff max.year min.diff min.year  
## 1 Jenascia 22041 2004 -24831 2005  
## 2 Betzi 13455 2008 -22344 2007  
## 3 Carin 11537 2009 -22146 2008  
## 4 Jnaya 16023 2001 -21894 2003  
## 5 Anallely 9011 2003 -21872 2002  
## 6 Adasyn 16791 2009 -21797 2008

For boys: the biggest increase is for name: Barack from 2007-2008. The biggest decrease is for name: Brasen from 2006-2007.

For girls: the biggest increase is for name: Jaslene from 2006-2007. The biggest decrease is for name: Jenascia from 2004-2005.

## 6 Make a plot showing the popularity of the 5 most popular names for each sex over time.

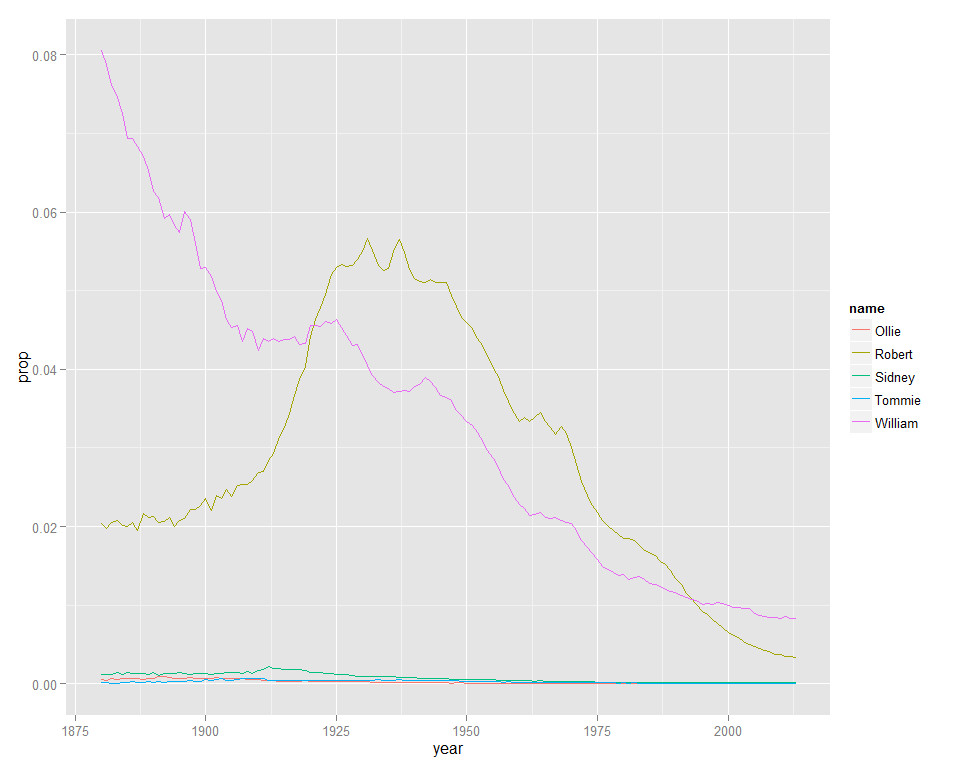
#First : rank the name   
BnamesByYearRank <-  
babynames %>%  
group\_by(year)%>% # group by name year  
mutate(rank = rank(prop, ties.method = "first"))%>% # missing values will be put last  
arrange(rank)  
  
#get the top 5 'boys' names based on rank  
  
BnamesBoysRank<- BnamesByYearRank %>%  
 filter(sex == "M" && rank <=10)  
  
#head(BnamesBoysRank)  
  
BnamesBoysRankName<- names(tail(sort(table(BnamesBoysRank$name)),5))  
  
#get the top 5 girls' names based on rank  
  
BnamesGirlsRank<- BnamesByYearRank %>%  
 filter(sex == "F" && rank <=10)  
  
#head(BnamesGirlsRank)  
  
BnamesGirlsRankName<- names(tail(sort(table(BnamesGirlsRank$name)),5))  
  
#Second: Filter the data with the top names and sex  
#boys  
BnamesBoysTop <- babynames[which(babynames$name %in% c(BnamesBoysRankName)),] %>% filter(sex == "M")  
  
head(BnamesBoysTop)

## Source: local data frame [6 x 5]  
##   
## year sex name n prop  
## 1 1880 M William 9532 0.0805067568  
## 2 1880 M Robert 2415 0.0203969595  
## 3 1880 M Sidney 142 0.0011993243  
## 4 1880 M Ollie 63 0.0005320946  
## 5 1880 M Tommie 15 0.0001266892  
## 6 1881 M William 8524 0.0787181973

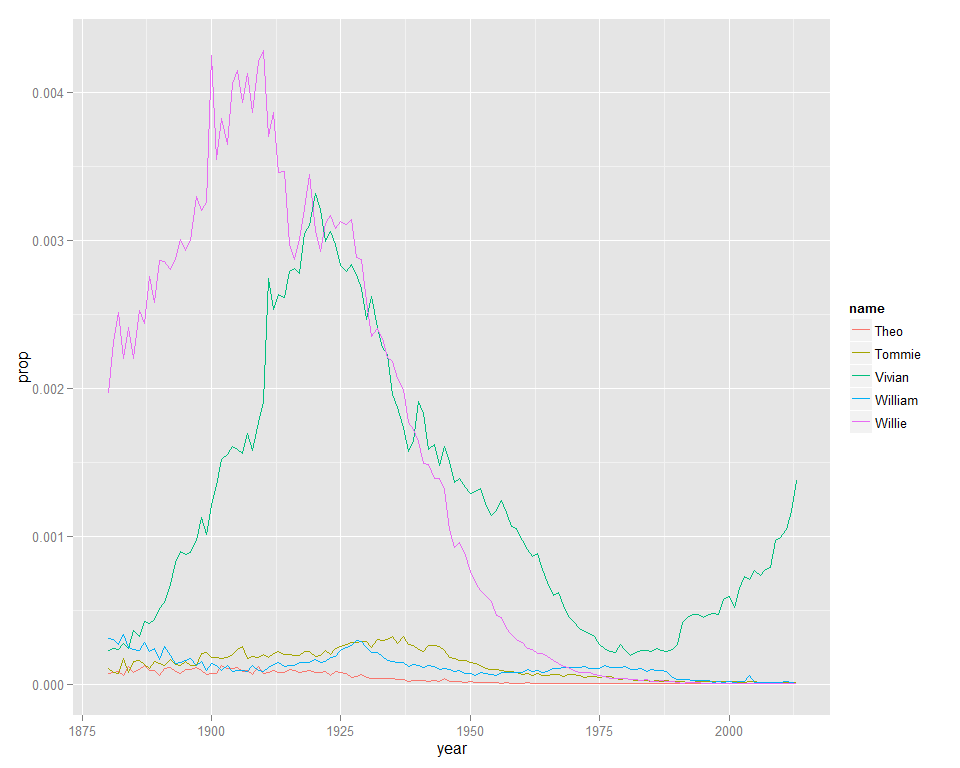
#girls  
BnamesGirlsTop <- babynames[which(babynames$name %in% c(BnamesGirlsRankName)),] %>% filter(sex == "F")  
  
head(BnamesGirlsTop)

## Source: local data frame [6 x 5]  
##   
## year sex name n prop  
## 1 1880 F Willie 192 1.967112e-03  
## 2 1880 F William 30 3.073613e-04  
## 3 1880 F Vivian 22 2.253983e-04  
## 4 1880 F Tommie 10 1.024538e-04  
## 5 1880 F Theo 7 7.171764e-05  
## 6 1881 F Willie 229 2.316477e-03

#Third : plot  
#Boys  
ggplot(aes(x= year), data = BnamesBoysTop)+  
 geom\_line(aes(y = prop, color = name))



#Boys  
ggplot(aes(x= year), data = BnamesGirlsTop)+  
 geom\_line(aes(y = prop, color = name))



## 7 Some names have been relatively popular for both boys and girls (sometimes in different years). Find the top 10 such names (you will have to define ?top?) and create a plot showing the popularity of these ten names for boys and girls over time. Are there any patterns?

#Get the popular names for both boys and girls  
#Boys name and count  
BnamesBoysRankName<-sort(table(BnamesBoysRank$name))  
  
BnamesBoys <- matrix(0, nrow= length(BnamesBoysRankName), ncol = 2 )  
  
colnames(BnamesBoys) <- c("name","count")  
  
BnamesBoys[,1]<- names(BnamesBoysRankName)  
  
BnamesBoys[,2] <- as.numeric(BnamesBoysRankName)  
  
head(BnamesBoys)

## name count  
## [1,] "Aabid" "1"   
## [2,] "Aabriella" "1"   
## [3,] "Aadhi" "1"   
## [4,] "Aadhyan" "1"   
## [5,] "Aadiv" "1"   
## [6,] "Aadrian" "1"

#Girls: name and count  
BnamesGirlsRankName<-sort(table(BnamesGirlsRank$name))  
  
BnamesGirls <- matrix(0, nrow= length(BnamesGirlsRankName), ncol = 2 )  
  
colnames(BnamesGirls) <- c("name","count")  
  
BnamesGirls[,1]<- names(BnamesGirlsRankName)  
  
BnamesGirls[,2] <- as.numeric(BnamesGirlsRankName)  
  
head(BnamesGirls)

## name count  
## [1,] "Aage" "1"   
## [2,] "Aagot" "1"   
## [3,] "Abba" "1"   
## [4,] "Abbot" "1"   
## [5,] "Aberham" "1"   
## [6,] "Abney" "1"

#Find the popular name for both boys and girls by merge()  
BnamesBoth<-merge(BnamesBoys,BnamesGirls ,by = "name",all.x=FALSE, all.y=FALSE)   
  
BnamesBoth$sum.count <- as.numeric(BnamesBoth$count.x)+as.numeric(BnamesBoth$count.y)  
  
  
BnamesBothSort<-arrange(BnamesBoth, -sum.count)  
  
head(BnamesBothSort,10)

## name count.x count.y sum.count  
## 1 Sylvester 98 98 266  
## 2 Jennie 97 97 264  
## 3 Lessie 99 95 264  
## 4 Ernie 95 97 262  
## 5 Willard 95 97 262  
## 6 Thelma 95 96 261  
## 7 Mae 91 99 260  
## 8 Pauline 95 94 259  
## 9 Verna 99 90 259  
## 10 Roma 98 90 258

#Filter data by the above names  
BnamesBothData <- babynames[which(babynames$name %in% c(as.character(BnamesBothSort[1:10,]$name))),]   
  
head(BnamesBothData,10)

## Source: local data frame [10 x 5]  
##   
## year sex name n prop  
## 1 1880 F Jennie 793 8.124584e-03  
## 2 1880 F Mae 344 3.524410e-03  
## 3 1880 F Pauline 144 1.475334e-03  
## 4 1880 F Verna 46 4.712873e-04  
## 5 1880 F Lessie 18 1.844168e-04  
## 6 1880 F Thelma 5 5.122688e-05  
## 7 1880 M Sylvester 89 7.516892e-04  
## 8 1880 M Willard 88 7.432432e-04  
## 9 1880 M Jennie 7 5.912162e-05  
## 10 1880 M Ernie 6 5.067568e-05

# Plot  
ggplot(aes(x= year), data = BnamesBothData)+  
 geom\_line(aes(y = prop, color = name))

