YingHu\_assignment3

Ying Hu

9/7/2019

install.packages("MASS", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/wingy/OneDrive/Documents/R/win-library/3.6'  
## (as 'lib' is unspecified)

## package 'MASS' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\wingy\AppData\Local\Temp\RtmpoLDYpg\downloaded\_packages

data(package="MASS")  
  
install.packages("dplyr", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/wingy/OneDrive/Documents/R/win-library/3.6'  
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'dplyr'

## Warning in file.copy(savedcopy, lib, recursive = TRUE):  
## problem copying C:\Users\wingy\OneDrive\Documents\R\win-  
## library\3.6\00LOCK\dplyr\libs\x64\dplyr.dll to C:  
## \Users\wingy\OneDrive\Documents\R\win-library\3.6\dplyr\libs\x64\dplyr.dll:  
## Permission denied

## Warning: restored 'dplyr'

##   
## The downloaded binary packages are in  
## C:\Users\wingy\AppData\Local\Temp\RtmpoLDYpg\downloaded\_packages

data(package="dplyr")  
  
install.packages("babynames", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/wingy/OneDrive/Documents/R/win-library/3.6'  
## (as 'lib' is unspecified)

## package 'babynames' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\wingy\AppData\Local\Temp\RtmpoLDYpg\downloaded\_packages

data(package="babynames")  
  
install.packages("ggplot2", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/wingy/OneDrive/Documents/R/win-library/3.6'  
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\wingy\AppData\Local\Temp\RtmpoLDYpg\downloaded\_packages

data(package="ggplot2")  
  
install.packages("ISwR", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/wingy/OneDrive/Documents/R/win-library/3.6'  
## (as 'lib' is unspecified)

## package 'ISwR' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\wingy\AppData\Local\Temp\RtmpoLDYpg\downloaded\_packages

data(package="ISwR")

#Question 1 (2 pts).  
  
#A. Install packages ISwR, dplyr, ggplot2, MASS, babynames  
  
#B. How many functions does each package offer?   
#Answer:  
#dplyr number of functions: 261  
#ISwR number of funcitons: 0  
#ggplot2 number of functions:351   
#MASS number of funcitons: 76  
#babynames number of funcitons: 0  
  
  
  
library(babynames)#pull out infomation from library  
grep("" ,lsf.str("package:babynames"), value = TRUE)#list all functions

## character(0)

ls("package:babynames") # list all objects

## [1] "applicants" "babynames" "births" "lifetables"

lsf.str("package:babynames")# details for function  
  
  
  
#C. What is the data structure of the output that contains only the data sets?  
#Answer: There is the index with datasets' names and with details/descreption.   
  
  
  
#D. How many datasets each package contains?  
#Answer:   
#dplyr number of dataset: 6  
#ISwR number of dataset: 38  
#ggplot2 number of dataset: 11  
#MASS number of dataset: 87  
#babynames number of dataset: 4  
  
  
library(babynames)#pull out infomation from library  
data(package = "babynames")$results #To list all datasets with brief description

## Package LibPath   
## [1,] "babynames" "C:/Users/wingy/OneDrive/Documents/R/win-library/3.6"  
## [2,] "babynames" "C:/Users/wingy/OneDrive/Documents/R/win-library/3.6"  
## [3,] "babynames" "C:/Users/wingy/OneDrive/Documents/R/win-library/3.6"  
## [4,] "babynames" "C:/Users/wingy/OneDrive/Documents/R/win-library/3.6"  
## Item Title   
## [1,] "applicants" "Applicants."  
## [2,] "babynames" "Baby names."  
## [3,] "births" "Births"   
## [4,] "lifetables" "Lifetables"

#E. What commands did you have to issue in order to get to your answer?What was the order of the commands and why?  
#Answer:Here is my order of the commands and reason:   
#1. Will pull out the info from library--make sure I can pull out the completed and correct information from package.   
#2. Then list all functions and data--run the results.   
#3. List all packages object--to double check my answer.   
#4. List all the details for functions--to double check my answer.

#Question 2 (4 pts).  
  
#A. Use package “babynames”. How many rows are there in the dataset? How many variables?  
#Answer: Using nrow(babynames)+nrow(Applicants)+nrow(Births)+nrow(lifetables)=total rows 1,927,930. and using ncol(babynames)+ncol(applicants)+ncol(births)+ncol(lifetables)=total 19 variables. Please see below:  
  
library(babynames)  
nrow(babynames)#1,924,665-row

## [1] 1924665

nrow(applicants)#276-row

## [1] 276

nrow(births)#109-row

## [1] 109

nrow(lifetables)#2880-row

## [1] 2880

ncol(babynames)#5-Var

## [1] 5

ncol(applicants)#3-Var

## [1] 3

ncol(births)#2-Var

## [1] 2

ncol(lifetables)#9-Var

## [1] 9

#B. How many unique female and male names are in the data set?  
#Answer: Female is 67046 and male is 40927.  
  
library(babynames)  
female<- babynames[babynames$sex=="F",]#filter female in package  
male<- babynames[babynames$sex=="M",]#filter male in package  
femaleuniq<- sort(unique(female$name))#output 67046 elements  
maleuniq<-sort(unique(male$name))#output 40927 elements  
length(femaleuniq)

## [1] 67046

length(maleuniq)

## [1] 40927

#C. What can be analyzed using this data set?   
#Answer: we can used this dataset analyze how many babies/female & male were born in a year; which were the popular names for both or each of them. We can also use this dataset to forecast the populations and the trend of the names.   
  
  
#D. How many unit name–sex combinations are there for 2014?  
#Answer:There are 493 name-sex combinations for 2014.Please see below:  
  
library(babynames)  
f\_namesex<-babynames[babynames$year==2014 & babynames$sex=="F",]#filter 2014 & female  
m\_namesex<-babynames[babynames$year==2014 & babynames$sex=="M",]#filter 2014 & male  
sum(f\_namesex$n,m\_namesex$n)# attached number of the names under female and male in 2014 and sum them together

## [1] 3696311

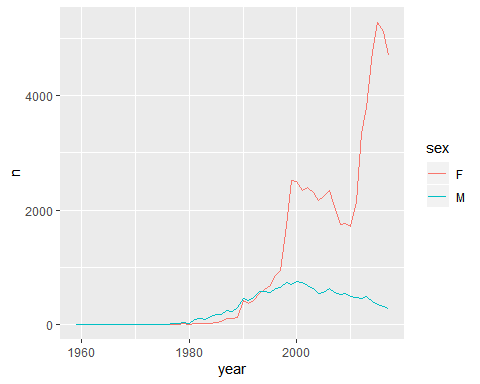
length(intersect(f\_namesex$n, m\_namesex$n))

## [1] 493

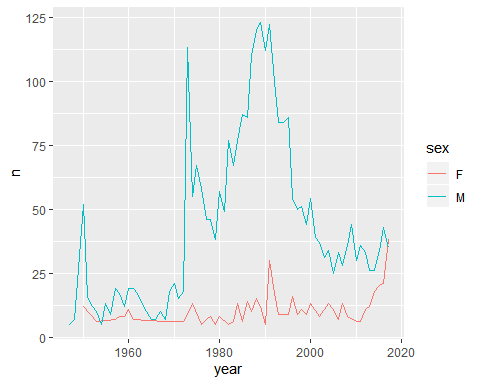
#E. Are there any names that are used for both male and female babies in 2017? How many?   
#Answer: Yes, there are 2559 names for both male and female babies in 2017. Please see below:  
library(babynames)  
fname<-babynames[babynames$year==2017 & babynames$sex=="F",]#filter female name from 2017  
mname<-babynames[babynames$year==2017 & babynames$sex=="M",]#filter male name from 2017  
length(intersect(fname$name, mname$name))

## [1] 2559

#F. Visualize the trend of “Skylar”, as both female and male name, across the years. Visualize similar trends for any other name of your choosing. Provide specific commands with options and arguments.  
#Answer: The graph shows female name Skylar is popular after the year 2010, and there are more than 5000 babies named Skylar in female group each year, and it was going down a little big around 2015. however, this name is not popular for male, it is going down after the year 2000. Another name is Jory, it is not popular for both female and male, but the population of male babies named Jory is larger than female; and female name Jory starts increasing around 2017 while male is decreasing.  
  
library(babynames)  
library(ggplot2)  
bothname<-babynames[babynames[['name']]=="Skylar",]#filter Skylar under name column   
ggplot( bothname, aes(x = year, y = n)) + geom\_line(aes(colour = sex))#turn it into graph



library(babynames)  
library(ggplot2)  
otherbname<-babynames[babynames[['name']]=="Jory",]#filter Jory under name column   
ggplot( otherbname, aes(x = year, y = n)) + geom\_line(aes(colour = sex))



#Question 3 (2 pts).  
  
#Consider the following relative frequency table on hurricanes that have made direct hits on the U.S. between 1851 and 2004.  
  
#Hurricanes are given a strength category rating based on the minimum wind speed generated by the storm. Some of the entries to the table are missing.  
  
#A. What is the relative frequency of direct hits of category 1?  
#Answer: total number of direct hits is 72/0.2637=273, there for the relative frequency is 109/273=0.3993. The acumulative relative frequency is also 0.3993.  
  
  
#B. What is the relative frequency of direct hits of category 4 or more?  
#Answer: total number of direct hits is 72/0.2637=273, there for the relative frequency is 18/273=0.0659. whichi is less than category 1.

#Question 4 (1 pt).  
  
#A. Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?  
  
#Answer:No, this will not be a good measure of household size since it only focuses on a specific group. and Yes, the average will be biased. It probably overestimate because some students are from a same large family, that will make double of the measure of household size.   
  
  
  
#Question 5 (1 pt).  
  
#The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.  
  
#Answer: The reasons that Gallup poll chooses to use random digit dialing because the phone book probably doesn't have most updated and completed info. Using the random digit dialing can make sure can call everyone and get the sample represent the population.