Assignment 5 - V2

Jonathan Flores

February 10, 2018

## Question 1

Data Munging (30 points): Utilize yob2016.txt for this question. This file is a series of popular childrens names born in the year 2016 in the United States. It consists of three columns with a first name, a gender, and the amount of children given that name. However, the data is raw and will need cleaning to make it tidy and usable.

1. First, import the .txt file into R so you can process it. Keep in mind this is not a CSV file. You might have to open the file to see what you are dealing with. Assign the resulting data frame to an object, df, that consists of three columns with human-readable column names for each.

## Setting working directory where yop2016 resides  
  
require(RCurl)

## Loading required package: RCurl

## Warning: package 'RCurl' was built under R version 3.4.3

## Loading required package: bitops

## Warning: package 'bitops' was built under R version 3.4.1

## Importing the text files into a data frame and assigning it to variable df  
  
df <- data.frame(read.delim(text = getURL("https://raw.githubusercontent.com/jfzjonathan/Homework5/master/RawData/yob2016.txt"),header = FALSE,sep=";"))  
  
  
## Naming the columns of the table  
  
colnames(df) <- c("Names","Gender","AmountOfNames")

1. Display the summary and structure of df

library(knitr)  
knitr::kable(head(df))

|  |  |  |
| --- | --- | --- |
| Names | Gender | AmountOfNames |
| Emma | F | 19414 |
| Olivia | F | 19246 |
| Ava | F | 16237 |
| Sophia | F | 16070 |
| Isabella | F | 14722 |
| Mia | F | 14366 |

1. Your client tells you that there is a problem with the raw file. One name was entered twice and misspelled. The client cannot remember which name it is; there are thousands he saw! But he did mention he accidentally put three y at the end of the name. Write an R command to figure out which name it is and display it.

## Find yyy in the Names column  
  
grep("yyy",df$Names)

[1] 212

## Display row 212  
  
df$Names[212]

[1] Fionayyy 30295 Levels: Aaban Aabha Aabid Aabir Aabriella Aadam Aadarsh … Zyva

1. Upon finding the misspelled name, please remove this particular observation, as the client says it is redundant. Save the remaining dataset as an object: y2016

## Validate the row to be removed  
  
df[212,]

Names Gender AmountOfNames

212 Fionayyy F 1547

## Remove row 212 and display the new row 212  
y2016 = df[-212,]  
y2016[212,]

Names Gender AmountOfNames

213 Callie F 1531

## Question 2

Data Merging (30 points): Utilize yob2015.txt for this question. This file is similar to yob2016, but contains names, gender, and total children given that name for the year 2015.

1. Like 1a, please import the .txt file into R. Look at the file before you do. You might have to change some options to import it properly. Again, please give the dataframe human-readable column names. Assign the dataframe to y2015.

## Setting working directory where y2015 resides  
  
  
  
y2015 <- data.frame(read.delim(text = getURL("https://raw.githubusercontent.com/jfzjonathan/Homework5/master/RawData/yob2015.txt"),header = FALSE,sep=","))  
  
  
## Naming the columns of the table  
  
colnames(y2015) <- c("Names","Gender","AmountOfNames")  
  
knitr::kable(head(y2015))

|  |  |  |
| --- | --- | --- |
| Names | Gender | AmountOfNames |
| Emma | F | 20415 |
| Olivia | F | 19638 |
| Sophia | F | 17381 |
| Ava | F | 16340 |
| Isabella | F | 15574 |
| Mia | F | 14871 |

1. Display the last ten rows in the dataframe. Describe something you find interesting about these 10 rows.

knitr::kable(tail(y2015,10))

|  |  |  |  |
| --- | --- | --- | --- |
|  | Names | Gender | AmountOfNames |
| 33054 | Ziyu | M | 5 |
| 33055 | Zoel | M | 5 |
| 33056 | Zohar | M | 5 |
| 33057 | Zolton | M | 5 |
| 33058 | Zyah | M | 5 |
| 33059 | Zykell | M | 5 |
| 33060 | Zyking | M | 5 |
| 33061 | Zykir | M | 5 |
| 33062 | Zyrus | M | 5 |
| 33063 | Zyus | M | 5 |

## Somthing interesting is that all these names were given to 5 children each

1. Merge y2016 and y2015 by your Name column; assign it to final. The client only cares about names that have data for both 2016 and 2015; there should be no NA values in either of your number of children rows after merging.

## Merging both tables  
  
final <- merge(y2015,y2016, by ="Names")  
knitr::kable(head(final))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Names | Gender.x | AmountOfNames.x | Gender.y | AmountOfNames.y |
| Aaban | M | 15 | M | 9 |
| Aabha | F | 7 | F | 7 |
| Aabriella | F | 5 | F | 11 |
| Aadam | M | 22 | M | 18 |
| Aadarsh | M | 15 | M | 11 |
| Aaden | M | 297 | M | 194 |

## Checking if there are blanks on the Names column  
  
sum(is.na(final$Names))

[1] 0

## Checking if there are blanks on the Gender column  
sum(is.na(final$Gender))

## Warning in is.na(final$Gender): is.na() applied to non-(list or vector) of  
## type 'NULL'

[1] 0

## Checking if there are blanks on the AmountOfNames column  
sum(is.na(final$AmountOfNames))

## Warning in is.na(final$AmountOfNames): is.na() applied to non-(list or  
## vector) of type 'NULL'

[1] 0

## Since they are all 0, means there are no NAs

## Question 3

Data Summary (30 points): Utilize your data frame object final for this part.

1. Create a new column called “Total” in final that adds the number of children in 2015 and 2016 together. In those two years combined, how many people were given popular names?

final$Total <- (final$AmountOfNames.x + final$AmountOfNames.y)  
knitr::kable(head(final))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Names | Gender.x | AmountOfNames.x | Gender.y | AmountOfNames.y | Total |
| Aaban | M | 15 | M | 9 | 24 |
| Aabha | F | 7 | F | 7 | 14 |
| Aabriella | F | 5 | F | 11 | 16 |
| Aadam | M | 22 | M | 18 | 40 |
| Aadarsh | M | 15 | M | 11 | 26 |
| Aaden | M | 297 | M | 194 | 491 |

1. Sort the data by Total. What are the top 10 most popular names?

library(plyr)

## Warning: package 'plyr' was built under R version 3.4.3

knitr::kable(head(arrange(final,desc(Total)),10))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Names | Gender.x | AmountOfNames.x | Gender.y | AmountOfNames.y | Total |
| Emma | F | 20415 | F | 19414 | 39829 |
| Olivia | F | 19638 | F | 19246 | 38884 |
| Noah | M | 19594 | M | 19015 | 38609 |
| Liam | M | 18330 | M | 18138 | 36468 |
| Sophia | F | 17381 | F | 16070 | 33451 |
| Ava | F | 16340 | F | 16237 | 32577 |
| Mason | M | 16591 | M | 15192 | 31783 |
| William | M | 15863 | M | 15668 | 31531 |
| Jacob | M | 15914 | M | 14416 | 30330 |
| Isabella | F | 15574 | F | 14722 | 30296 |

1. The client is expecting a girl! Omit boys and give the top 10 most popular girl’s names.

finalgirls <- subset.data.frame(final,final$Gender.x=="F"|final$Gender.y=="F")  
knitr::kable(head(arrange(finalgirls,desc(Total)),10))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Names | Gender.x | AmountOfNames.x | Gender.y | AmountOfNames.y | Total |
| Emma | F | 20415 | F | 19414 | 39829 |
| Olivia | F | 19638 | F | 19246 | 38884 |
| Sophia | F | 17381 | F | 16070 | 33451 |
| Ava | F | 16340 | F | 16237 | 32577 |
| Isabella | F | 15574 | F | 14722 | 30296 |
| Mia | F | 14871 | F | 14366 | 29237 |
| Charlotte | F | 11381 | F | 13030 | 24411 |
| Abigail | F | 12371 | F | 11699 | 24070 |
| Emily | F | 11766 | F | 10926 | 22692 |
| Harper | F | 10283 | F | 10733 | 21016 |

1. Write these top 10 girl names and their Totals to a CSV file. Leave out the other columns entirely.

TopCommonGirlNames <- head(arrange(finalgirls,desc(Total)),10)  
write.csv(TopCommonGirlNames,file="TopCommonGirlNames.csv",sep = ",")

## Warning in write.csv(TopCommonGirlNames, file = "TopCommonGirlNames.csv", :  
## attempt to set 'sep' ignored

1. Upload to GitHub (10 points): Push at minimum your RMarkdown for this homework assignment and a Codebook to one of your GitHub repositories (you might place this in a Homework repo like last week). The Codebook should contain a short definition of each object you create, and if creating multiple files, which file it is contained in. You are welcome and encouraged to add other files-just make sure you have a description and directions that are helpful for the grader.

## GitHub repo for assignment 5: https://github.com/jfzjonathan/Homework5