Assignment_05

Stephen Johnson

February 5, 2019

1. Data Munging

a. Import the file into R

```
library(tidyr)
child_names <- read.delim("yob2016.txt", header = FALSE)
df <- separate(child_names, V1, c("Name", "Sex", "Count"), sep = ";")
```

b. Display the summary and structure of df

```
summary(df)
```

```
##      Name               Sex                Count          
##  Length:32869       Length:32869       Length:32869      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character
```

```
str(df)
```

```
## 'data.frame':    32869 obs. of  3 variables:
##  $ Name : chr  "Emma" "Olivia" "Ava" "Sophia" ...
##  $ Sex  : chr  "F" "F" "F" "F" ...
##  $ Count: chr  "19414" "19246" "16237" "16070" ...
```

c. ID name misspelled with “yyy”

```
yyySubset <- df[grep("yyy", df$Name),]
yyySubset
```

```
##         Name Sex Count
## 212 Fionayyy   F  1547
```

d. Remove misspelled observation

```
df$Name <- gsub("yy","",df$Name)
y2016 <- df
head(y2016)
```

```
##       Name Sex Count
## 1     Emma   F 19414
## 2   Olivia   F 19246
## 3      Ava   F 16237
## 4   Sophia   F 16070
## 5 Isabella   F 14722
## 6      Mia   F 14366
```

2. Data Merging

a. Import the file into R

```
child_names2 <- read.delim("yob2015.txt", header = FALSE)
y2015 <- separate(child_names2, V1, c("Name", "Sex", "Count"), sep = ",")
```

b. Display the last 10 rows

```
tail(y2016, 10)
```

```
##         Name Sex Count
## 32860   Zinn   M     5
## 32861  Zirui   M     5
## 32862   Ziya   M     5
## 32863 Ziyang   M     5
## 32864   Zoel   M     5
## 32865 Zolton   M     5
## 32866 Zurich   M     5
## 32867 Zyahir   M     5
## 32868   Zyel   M     5
## 32869  Zylyn   M     5
```

Names that start with letter Z all Male with 5 count

c. Merge y2015 and y2016 by Name

```
names2015 <- c("Name", "Sex", "Count2015")
names2016 <- c("Name", "Sex", "Count2016")
colnames(y2015) <- names2015
colnames(y2016) <- names2016
final <- merge(y2015, y2016, by = c("Name","Sex" ))
```

3. Data Summary

a. Create a new column for total

```
final$Count2015 <- as.numeric(as.character(final$Count2015))
final$Count2016 <- as.numeric(as.character(final$Count2016))
final$total <- final$Count2015 + final$Count2016
head(final)
```

```
##        Name Sex Count2015 Count2016 total
## 1     Aaban   M        15         9    24
## 2     Aabha   F         7         7    14
## 3 Aabriella   F         5        11    16
## 4     Aadam   M        22        18    40
## 5   Aadarsh   M        15        11    26
## 6     Aaden   M       297       194   491
```

```
write.csv(final, "final.csv")
```

b. In those two years combined, how many people were given popular names?

```
print("Number of people who were given popular names are ");sum(final$total)
```

```
## [1] "Number of people who were given popular names are "
```

```
## [1] 7238859
```

c. Sort the data by total - What are the top 10 most popular names?

```
library(plyr)
head(arrange(final,desc(total)), n=10)
```

```
##        Name Sex Count2015 Count2016 total
## 1      Emma   F     20415     19414 39829
## 2    Olivia   F     19638     19246 38884
## 3      Noah   M     19594     19015 38609
## 4      Liam   M     18330     18138 36468
## 5    Sophia   F     17381     16070 33451
## 6       Ava   F     16340     16237 32577
## 7     Mason   M     16591     15192 31783
## 8   William   M     15863     15668 31531
## 9     Jacob   M     15914     14416 30330
## 10 Isabella   F     15574     14722 30296
```

d. Omit boys and provide the top 10 most popular girl's names

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
top10_girlnm <- final %>%  filter(Sex=="F") %>% arrange(desc(total)) %>% head(10)
head(top10_girlnm)
```

```
##       Name Sex Count2015 Count2016 total
## 1     Emma   F     20415     19414 39829
## 2   Olivia   F     19638     19246 38884
## 3   Sophia   F     17381     16070 33451
## 4      Ava   F     16340     16237 32577
## 5 Isabella   F     15574     14722 30296
## 6      Mia   F     14871     14366 29237
```

e. Write these top 10 girl names and their Totals to a CSV file

```
write.csv(top10_girlnm, "top10_girlnm.csv", row.names = FALSE)
```

6 Codebook

Local directory for Homework: “C:103_Working_05" ## Link to GitHub: https://github.com/sjohnson1039/MSDS_Assignments (https://github.com/sjohnson1039/MSDS_Assignments)