Code Book – MSDS 6306 - Unit 5 Assignment

Author: R Chandna

Brief Description of Analyses.

This analysis took two raw text files as inputs: file named yob2015.txt and yob2016.txt, files provided as a part of homework assignment for this week. After some initial cleaning, data frames generated from both files were merged together and then analyzed to answer the questions of interest such as:

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

What are the top 10 most popular names?

What are the top 10 most popular girl's names?

Raw Data Files.

The raw data files for this analysis are placed and can be downloaded from following URL:

https://github.com/R-Chandna/SMU_MSDS_HomeWork/tree/master/MSDS%206306_Doing_Data_Science/Week5/Analysis/Data

- yob2015.txt: File containing information popular children's name for children born in USA in year 2015. The following three types of information were present in the file in a comma (",") separated format:
  - First Name
  - Gender
  - Number of children given this popular first name in year 2015.

- yob2016.txt: File containing information popular children's name for children born in USA in year 2016. The following three types of information were present in the file in a semi-colon (";") separated format:
  - First Name
  - Gender
  - Number of children given this popular first name in year 2016.

Variables Used in the Analysis.

- Name (character, string): It represents the First Name of children.

- Gender(character): It represents Gender of the children that were given corresponding First Name. It contains 2 levels:
  - M for Males
  - F for Females
- No_Of_Children_Given_This_Name_2015(numeric): It represents the total children given the corresponding popular name for the year 2015.
- No_Of_Children_Given_This_Name_2016(numeric): It represents the total children given the corresponding popular name for the year 2016.
- GivenIn2015(numeric): Short Name given to variable "No_Of_Children_Given_This_Name_2015"
- GivenIn2016(numeric): Short Name given to variable "No_Of_Children_Given_This_Name_2016"
- Total(numeric): GivenIn2015 + GivenIn2016. To contain the number of children given popular name in both 2015 and 2016 together.

Objects used in Analysis.

df (data frame): Data frame generated from import of "yob2016.txt". This contains unprocessed table with columns containing popular first name, gender, and number of children given the corresponding popular first name in year 2016. Variables of this frame are:

```
## 'data.frame':    32869 obs. of  3 variables:
##  $ Name                               : chr  "Emma" "Olivia" "Ava" "Sophia
" ...
##  $ Gender                             : chr  "F" "F" "F" "F" ...
##  $ No_Of_Children_Given_This_Name_2016: int  19414 19246 16237 16070 14722
14366 13030 11699 10926 10733 ...
```

y2016(data frame): Processed Data Frame containing all entries from "df" except the observation with misspelled name "Fionayyy". Structure of this data frame is as:

```
## 'data.frame':    32868 obs. of  3 variables:
##  $ Name                               : chr  "Emma" "Olivia" "Ava" "Sophia
" ...
##  $ Gender                             : chr  "F" "F" "F" "F" ...
##  $ No_Of_Children_Given_This_Name_2016: int  19414 19246 16237 16070 14722
14366 13030 11699 10926 10733 ...
```

y2015 (data frame): Data frame generated from import of “yob2015.txt”. This contains table with columns containing popular first name, gender, and number of children given the corresponding popular first name in year 2015. Variables of this frame are:

```
## 'data.frame':    33063 obs. of  3 variables:
##  $ Name                               : chr  "Emma" "Olivia" "Sophia" "Ava
" ...
##  $ Gender                             : chr  "F" "F" "F" "F" ...
##  $ No_Of_Children_Given_This_Name_2015: int  20415 19638 17381 16340 15574
14871 12371 11766 11381 10283 ...
```

final (data frame): Data frame generated from merge of y2015 and y2016 data frames. The merge was performed by union of Name and Gender columns and contains only those rows that are consistent in both data frames. This is done to avoid any NA values without losing any ability to answer questions of interest effectively. This contains table with columns containing popular first name, gender, number of children given the corresponding popular first name in year 2015 and number of children given the corresponding popular first name in year 2016. Structure of this frame is as:

```
## 'data.frame':    26550 obs. of  4 variables:
##  $ Name       : chr  "Aaban" "Aabha" "Aabriella" "Aadam" ...
##  $ Gender     : chr  "M" "F" "F" "M" ...
##  $ GivenIn2015: int  15 7 5 22 15 297 31 5 11 8 ...
##  $ GivenIn2016: int  9 7 11 18 11 194 28 6 5 14 ...
```

final$Total(numeric): Column(variable) added to final data frame to store total number of children given the corresponding popular name in both years 2015 and 2016. The structure of final data frame is changed to:

```
## 'data.frame':    26550 obs. of  5 variables:
##  $ Name       : chr  "Aaban" "Aabha" "Aabriella" "Aadam" ...
##  $ Gender     : chr  "M" "F" "F" "M" ...
##  $ GivenIn2015: int  15 7 5 22 15 297 31 5 11 8 ...
##  $ GivenIn2016: int  9 7 11 18 11 194 28 6 5 14 ...
##  $ Total      : int  24 14 16 40 26 491 59 11 16 22 ...
```

final_fem (data frame): Data frame generated from final data frame after retaining observations pertaining to Girl's Names only. Structure of this frame is as:

```
## 'data.frame':    15267 obs. of  5 variables:
##  $ Name       : chr  "Emma" "Olivia" "Sophia" "Ava" ...
##  $ Gender     : chr  "F" "F" "F" "F" ...
##  $ GivenIn2015: int  20415 19638 17381 16340 15574 14871 11381 12371 11766
10283 ...
##  $ GivenIn2016: int  19414 19246 16070 16237 14722 14366 13030 11699 10926
10733 ...
##  $ Total      : int  39829 38884 33451 32577 30296 29237 24411 24070 22692
21016 ...
```

Top10GirlsNamesFor2015-16.csv: CSV file generated after top 10 popular girl's name along with the corresponding frequency with which these names were given during year 2015 and 2016 were written to a .csv file. This file can be downloaded from following URL:

https://github.com/R-Chandna/SMU_MSDS_HomeWork/tree/master/MSDS%206306_Doing_Data_Science/Week5/Analysis/Data