Live Session Homework Codebook

All code made in R

# CBass\_Livesession5assignment.Rmd

R-Markdown file for Unit 5 Homework. Data frames and files created are listed below.

## Data frame – df

Made from yob2016.txt file using the following code:

yob2016 <- read.table('C:\\Users\\Clovis\\Desktop\\School\\SMU Data Science\\Spring 2018\\Doing Data Science MSDS 6306\\Unit 5\\yob2016.txt')

library(tidyr)

df <- separate(yob2016, V1, c("Name", "Sex", "Count"))

yob2016.txt contained untidy data listing the name, sex, and number of newborns that have such name. Columns were separated to make the data frame.

### Variables:

Name – character

Sex – character

Count – character

## Data frame – df2

Made from yob2015.txt file using the following code:

yob2015 <- read.table('C:\\Users\\Clovis\\Desktop\\School\\SMU Data Science\\Spring 2018\\Doing Data Science MSDS 6306\\Unit 5\\yob2015.txt')

df2 <- separate(yob2015, V1, c("Name", "Sex", "Count"))

yob2015.txt contained untidy data listing the name, sex, and number of newborns that have such name. Columns were separated to make the data frame.

### Variables:

Name – character

Sex – character

Count – character

## Data frame – mergedDF

Made from merging df and df2 data frames under the Name and Sex variables using the following code:

mergedDF <- merge(df, df2, union("Name", "Sex"), all=TRUE)

mergedDF$Count.x <- as.numeric(mergedDF$Count.x)

mergedDF$Count.y <- as.numeric(mergedDF$Count.y)

mergedDF$Count.x[is.na(mergedDF$Count.x)] <- 0

mergedDF$Count.y[is.na(mergedDF$Count.y)] <- 0

library(dplyr)

mergedDF <- mutate(mergedDF, Total = Count.x+Count.y)

names(mergedDF) = c("Name", "Sex", "Count2016", "Count2015", "Total")

Newly merged data frame contains name counts from both yob2016.txt (df) and yob2015.txt (df2). They were then summed in the column “Total”.

### Variables:

Name – character

Sex – character

Count2016 – numeric

Count2015 – numeric

Total – numeric

## Data frame – sortedmergedDF

Made taking mergedDF data frame and sorting it in order from most common names from to least common using the following code:

sortedmergedDF <- mergedDF[order(mergedDF$Total, decreasing=TRUE),]

Name counts come from both yob2016.txt (df) and yob2015.txt (df2). They are summed in the column “Total”.

### Variables:

Name – character

Sex – character

Count2016 – numeric

Count2015 – numeric

Total – numeric

## Data frame – sorted.gender.mergedDF

Made using sortedmergedDF data frame to sort also by gender (females) using the following code:

sorted.gender.mergedDF <- arrange(sortedmergedDF, Sex)

All genders remain in this data frame, only that females are listed first.

### Variables:

Name – character

Sex – character

Count2016 – numeric

Count2015 – numeric

Total – numeric

## Data frame – top10girlsnames

Made using the first ten values from the data frame sorted.gender.mergedDF representing the top 10 female names from newborns in 2015 and 2016 using the following code:

top10girlsnames <- sorted.gender.mergedDF

top10girlsnames$Count2016 <- NULL

top10girlsnames$Count2015 <- NULL

top10girlsnames$Sex <- NULL

top10girlsnames <- head(top10girlsnames,10)

Only “Name” and “Total” variables are within this data frame.

### Variables:

Name – character

Total – numeric

### File – “Top10 girls names.csv”

Made from top10girlsnames data frame using the following code:

write.csv(top10girlsnames, file = "Top10 girls names.csv")

Data columns include “Name” and “Total” count of the top 10 girls names from 2015 (yob2015.txt and “df” dataframe) and 2016 (yob2016.txt and “df2” dataframe) combined.