HW5

Emil Ramos

February 12, 2018

## 1. Data Munging (30 points):

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’re dealing with. Assign the resulting data frame to an object, df, that consists of three columns with human-readable column names for each.

View(Book3) children <- Book3 str(children)

1. Display the summary and structure of df

str(children) Classes ‘tbl\_df’, ‘tbl’ and ‘data.frame’: 32868 obs. of 3 variables: $ Name : chr “Emma” “Olivia” “Ava” “Sophia” … $ Gender: chr “F” “F” “F” “F” … $ Total : num 19414 19246 16237 16070 14722 …

summary(children) Name Gender Total  
Length:32868 Length:32868 Min. : 5.0  
Class :character Class :character 1st Qu.: 7.0  
Mode :character Mode :character Median : 12.0  
Mean : 110.7  
3rd Qu.: 30.0  
Max. :19414.0

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’s at the end of the name. Write an R command to figure out which name it is and display it.  
   > which(children$Name == “Fionayyy”) [1] 212
2. Upon finding the misspelled name, please remove this particular observation, as the client says it’s redundant. Save the remaining dataset as an object: y2016

remove(children$Name = “Fionayyy”) y2016 <- Book3 y2016

## 2. Data Merging (30 points):

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.

y2015 <- Book5 y2015

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

tail(y2015,10) # A tibble: 10 x 3 Name Gender TotalChildren 1 Ziyu M 5.00 2 Zoel M 5.00 3 Zohar M 5.00 4 Zolton M 5.00 5 Zyah M 5.00 6 Zykell M 5.00 7 Zyking M 5.00 8 Zykir M 5.00 9 Zyrus M 5.00 10 Zyus M 5.00

The last 10 rows share the same amount of 5. Also, the children names begin with a Z.

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 amount of children rows after merging.

MergeData01 <- merge(x = y2016, y = y2015, by = “Name”, all = TRUE) final <- MergeData01 final

## 3. Data Summary (30 points

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

TotalX <- sum(finalTotal.y, na.rm = TRUE) [1] 5773122 TotalY <- sum(final$Total.y, na.rm = TRUE) TotalY [1] 5773122 TotalX + TotalY [1] 11478340 Total <- TotalX + TotalY Total [1] 11478340

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

Emma Olivia Noah Liam Sophia Mason Ava Jacob William Isabella

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

Emma Olivia Sophia Isabella Mia Harper Madison Evelyn Elizabeth Ella

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