**Stat 301 – Working with a large data file**

In Investigation 2.1, we want to explore birth weights of newborns in the United States. The CDC’s Vital Statistics Online Data Portal ([www.cdc.gov/nchs/data\_access/vitalstatsonline.htm](http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm)) gives you access to ALL births in a year.

(a) Open this website and follow the Downloadable Data Files link for Births. How large is the zip file for the U.S. data?

223MB

Open the User’s Guide for 2018. How many births were recorded in the U.S. in 2018?

*3,801,534*

According to this documentation file, how many total “positions” are there in the data file? What information is in the first 8 positions? What information is in the next 4 positions? How would you determine the birth month?

*There are 1330 positions. First 8 positions are filler, next four positions convey the year. The birth month is represented by the two digits in positions 13 and 14.*

If you were to download this file, it could take 10-20 minutes. So we have put the .zip file on a USB drive. Copy the BirthDataActivity folder to your desktop.

(b) Unzip the contents of the .zip file.

* In the BirthDataActivity folder on the desktop, right click on the .zip file and select Extract All.
* Next, move the awk.exe and getMonth.bat files into that new folder.

The unzipped file 5GB! This can be problematic to work with, especially if you were most interested in just a few of the variables. It can also be cumbersome to deal with all of the rows, especially if you just want to see general patterns. Most packages can now work with datafiles this large, but we are still pushing the limit a bit. JMP and R still take several minutes to read in the data, and each row is still a single string (not separated into columns or tabs).One approach is to preprocess the data a bit first.

To read a fixed width data file, one approach is to use the “awk” language

* Right click on GetMonth.bat and select Edit (should open NotePad)
	+ The goal of this program is to extract all the records for one month of the year. This happens by extracting out the rows where strings in positions 13 and 14 are equal to 01. The results are stored in Jan.txt.

(c) What does substr($0, 13,2) ==\"01\" do? (*Hint*: I just told you, so what do the 2 and 01 do?)

*This is telling the computer to read in the information in position 13 and position 14 (2 positions) and seeing where these birth month codes match 01, or January. “substr” stands for substring.*

* To run this batch file, double click on it (GetMonth.bat). (A “Command Prompt window” should open). After a few minutes, you will see the pause command. Press any key to continue. Now, in the folder, you should see a new file: Jan.txt

This is still a pretty large file and each row is one long string (“fixed width”). If you open this file in JMP or Excel, they at least try to separate the columns. You can use read.fwf in R to specify about the positions/variable, but you need to know all of the widths, e.g.,

w <- c(7, 11, 10, 5, 22, 10, 9, 9, 9, 9, 12)

read.fwf("myfile.dat", w)

(d) Where is information on how much weight the mother gained during pregnancy?

304-305

(e) How many positions are between birth month and mom’s weight gain? *Hint*: What is the “length” between these two variables?

15 to 303 (inclusive) = length of 289

(f) Where is information on the “5 minute Apgar score”? What is “apgar score”?

444-445

(g) How many positions between mom weight gain and 5 min Apgar score?

306 to 443 = 138

(h) Find OEGest\_R3. Where is it and what is it? How many positions between?

503, 446 to 502 => 57

(i) Which positions contain the baby birth weight? What are the measurement units?

504-507, grams

So now we want to extract data on birth weight, whether the baby was full term (gestation over 36 weeks), the 5 minute apgar score (an immediate measure of the infant’s health), and the amount of weight gained by the mother during pregnancy (in lbs).

(j) Confirm the following widths

 14 to include positions 1-14

 488 to include positions 15-502

 1 to include OEGest\_R3 (503)

 4 to include birthweight (504-507)

 823 to include 508-1330

> w = c(14, -488, 1, 4, -823)

> JanSub = read.fwf("C:/Users/bchance/Downloads/Nat2018us/Jan.txt", w)

> names(JanSub) = c("date", "OEG", "birthweight")

This should take 2-3 minutes. Let’s extract less data instead.

* In the text editor of the getMonth.bat file, type rem in front of the existing line.
* Copy/type the following as the next line (before pause)

rem awk "{ if (substr($0, 13,2) ==\"01\") print $0 }" < Nat2018PublicUS.c20190509.r20190717.txt > Jan.txt

awk "{ if (substr($0, 13,2) ==\"01\") print (substr($0, 504, 4) \",\" substr($0, 503, 1) ) }" < Nat2018PublicUS.c20190509.r20190717.txt > Jan.txt

pause

Note: I had the closing parenthesis for the print statement in the wrong location.

This takes out a minute and then prints out positions 503, comma separated, positions 504-507. Run the batch file.



(k) Read these data into R or JMP (use a file open dialog and navigate to this folder)

RStudio: Import Dataset > From Text (base)



> names(Jan) = c("birthweight”, "OEGest")

Alternatively, if the datafile is on the web, you can use the Import Dataset > From Text (readr). This may require installing the readr package. Then you can paste in the URL. You may need to clarify tab vs. csv delimiting.

JMP: File > Open (probably then “Data, using best guess), and name the columns

For a datafile on the web, download it to your computer (e.g., save target as) and then File > Open in JMP.

(l) Now extract the birth weights, mother’s weight gain, Apgar score, and OEGest\_R3 for your assigned month.

We didn’t do this, but try

awk "{ if (substr($0, 13,2) ==\"02\") print (substr($0, 504, 4) \",\", substr($0, 304, 2) \",\" substr($0, 444, 2) \",\" substr($0, 503, 1) ) }" < Nat2018PublicUS.txt > Feb2.csv

This will give the four variables for February