**R assignment 1 -- Epi 510, Fall 2022**

For this assignment, please submit two files: (1) **a file containing text and tables** (.doc, .docx, .pdf or .txt) answering questions posed in the assignment, and (2) **an R script** that performs the requested operation (.R).

**50 points possible**

1. Let’s setup our session:
   1. Change the working directory to be your Epi 510 R directory (e.g. C:/epi510/r). **(1 point)**
   2. Create a data frame by reading data from “vipcls.csv” into R. **(2 points)**   
      All datasets for the R portion of the class are on the canvas site (<https://canvas.uw.edu/courses/1580161/pages/r-data-and-program-files-for-download>). Note that a description of the data and data dictionary are given on the course website (<https://canvas.uw.edu/courses/1580161/files/95696503/download?wrap=1> ). Read through this!
2. Now, let’s examine the data:
   1. View the dataset to visually inspect the data. **(1 point)**

Done in R - view(vipcls)

* 1. What is the class of vipcls? **(1 point)**

Data frame.

* 1. List the names of variables in vipcls. **(1 point)**

[1] "patid" "delmo" "deldy" "delyr" "enrmo" "enrdy"

[7] "enryr" "momage" "raceth" "grade" "marstat" "cigs1"

[13] "cigs2" "etoh1" "etoh2" "partyr" "pregnum" "delges"

[19] "bw" "deltype" "induclab" "auglab" "intrapih"

* 1. What are the dimensions of vipcls? **(1 point)**

13285, 23

* 1. View the structure of vipcls. **(2 points)**

'data.frame': 13285 obs. of 23 variables:

$ patid : int 1200005 1200104 1200203 1200302 1200401 1200500 1200609 1200708 1200807 1201003 ...

$ delmo : int -1 11 10 11 11 11 10 -1 10 10 ...

$ deldy : int -1 9 31 3 3 1 24 -1 4 2 ...

$ delyr : int -1 1987 1987 1987 1987 1987 1987 -1 1987 1987 ...

$ enrmo : int 4 5 3 7 6 4 5 4 3 3 ...

$ enrdy : int 27 13 31 1 2 3 8 10 18 23 ...

$ enryr : int 1987 1987 1987 1987 1987 1987 1987 1987 1987 1987 ...

$ momage : int 24 26 22 26 21 25 20 35 20 29 ...

$ raceth : int 1 2 2 1 2 1 2 1 1 1 ...

$ grade : int 13 11 12 16 12 12 11 8 12 12 ...

$ marstat : int 1 5 5 3 5 5 5 3 1 2 ...

$ cigs1 : int 0 0 0 0 0 0 3 0 0 0 ...

$ cigs2 : int 0 0 0 0 0 0 3 0 0 0 ...

$ etoh1 : int 6 4 6 6 6 6 6 6 6 6 ...

$ etoh2 : int 6 4 6 6 6 6 6 6 6 6 ...

$ partyr : int 1 1 1 1 1 1 1 1 1 1 ...

$ pregnum : int 2 4 2 2 1 3 2 8 1 4 ...

$ delges : int 40 40 39 38 38 38 37 37 38 40 ...

$ bw : int 2650 2725 3232 3195 2720 3629 2090 3015 3430 3686 ...

$ deltype : int 1 1 1 1 1 1 1 1 1 1 ...

$ induclab: int 2 2 2 2 1 2 2 2 2 2 ...

$ auglab : int 2 2 2 2 2 1 2 2 1 2 ...

$ intrapih: int 2 2 2 2 1 2 2 2 2 2 ...

* 1. Use the summary function to summarize all variables in vipcls to see the distribution and range of values for each variable. Look at the range of values for each variable. You should see a value that is impossible for most of the variables. What is that value? **(3 points)**

-1 (indicating missing data)

1. Write an R script that cleans the new data frame in the following way:
   1. Set all -1 values to missing **(2 points)**

vipcls[vipcls==-1] <- NA

* 1. Set mother’s age (momage) of ≤15 to missing **(2 points)**

vipcls$momage[vipcls$momage<=15] <- NA

* 1. Set gestational age (delges) above 52 to missing **(2 points)**

vipcls$delges[vipcls$delges>52] <- NA

* 1. Set birth weight of babies (bw) below 300 to missing **(2 points)**

vipcls$bw[vipcls$bw<300] <- NA

* 1. Set birth weight of babies (bw) 6000 and above to missing **(2 points)**  
     vipcls$bw[vipcls$bw>=6000] <- NA

1. Add code to your script that that does the following:
2. Converts binary variables to logicals. **(8 points)**

* Step 1: Convert deltype cesarean (2) to 0 (false) and convert to logical

Vipcls$deltype[vipcls$deltype==2] <- 0

vipcls$deltype <- as.logical(vipcls$deltype)

* Step 2: Convert 2 (No) to 0 (false) for induclab and convert to logical

vipcls$induclab[vipcls$induclab==2] <- 0

vipcls$induclab <- as.logical(vipcls$induclab)

* Step 3: Convert 2 (No) to 0 (false) for auglab and convert to logical

vipcls$auglab[vipcls$auglab==2] <- 0

vipcls$auglab <- as.logical(vipcls$auglab)

* Step 4: Convert 2 (No) to 0 (false) for intrapih and convert to logical

vipcls$intrapih[vipcls$intrapih==2] <- 0

vipcls$intrapih <- as.logical(vipcls$intrapih)

1. Converts categorical variables with more than two levels to factors (excluding date variables). **(8 points)**

* Convert etoh 1

vipcls$etoh1 <- factor(vipcls$etoh1, levels = 1:6, labels=c("every day", "3-5/wk", "1/wk","<1/wk", "<1/month", "never"))

* Convert etoh 2

vipcls$etoh2 <- factor(vipcls$etoh2, levels = 1:6, labels=c("every day", "3-5/wk", "1/wk","<1/wk", "<1/month", "never"))

* Convert marstat

vipcls$marstat <- factor(vipcls$marstat, levels = 1:5, labels=c("married", "separated", "divorced","widowed", "never married"))

* Convert raceth

vipcls$raceth <- factor(vipcls$raceth, levels = 0:2, labels=c("White", "Hispanic", "Black"))

1. Add code to your script that does the following:

* For each numeric (i.e. discrete or continuous) variable in your dataset, determine the minimum, maximum, mean and the number of missing values **(2 points)**.
  + cigs1:
    - Min: 0
    - Max: 92
    - Mean: 4
    - Number of missing values: 1
  + Cigs2:
    - Min: 0
    - Max: 70
    - Mean: 3.07
    - Number of missing values: 1
  + partyr:
    - Min: 0
    - Max: 200
    - Mean: 1.26
    - Number of missing values: 223
  + pregnum:
    - Min: 1
    - Max: 16
    - Mean: 2.642
    - Number of missing values: 0
  + Delges
    - Min: 23
    - Max: 52
    - Mean: 39.06
    - Number of missing values: 5
  + Grade:
    - Min: 0
    - Max: 27
    - Mean: 11.52
    - Number of missing values: 30
  + Momage:
    - Min: 16
    - Max: 46
    - Mean: 24.32
    - Number of missing values: 6
  + bw:
    - Min: 320
    - Max: 5650
    - Mean: 3268
    - Number of missing values: 477
* For each binary and categorical variable, determine the levels, the number of observations in each level and the number of missing values **(2 points)**.
  + Raceth:
    - White: 4219
    - Hispanic: 4023
    - Black: 5043
    - Missing: 0
  + Marstat:
    - Married: 5699
    - Separated: 1206
    - Divorced: 811
    - Widowed: 58
    - Never married: 5499
    - Missing: 12
  + Etoh1:
    - Every day: 74
    - 3-5/wk: 123
    - 1/wk: 558
    - <1/wk: 624
    - <1/month: 1932
    - Never: 9972
    - Missing: 2
  + Etoh2:
    - Every day: 32
    - 3-5/wk: 40
    - 1/wk: 348
    - <1/wk: 548
    - <1/month: 1872
    - Never: 10442
    - Missing: 3
  + Deltype:
    - True (vaginal delivery): 12636
    - Missing: 649
  + Induclab:
    - False: 11660
    - True: 1160
    - Missing: 465
  + Auglab:
    - False: 8325
    - True: 4498
    - Missing: 462
  + Intrapih:
    - False: 12022
    - True: 925
    - Missing: 338

Check that your data frame matches the information on the following page and correct your code, if necessary. **You do NOT need to submit output or a table.**

1. Answer the following questions:
   1. What percentage of mothers drank alcohol exactly one time per week during their second trimester? **(2 points)**

2.6200873%

* 1. What are the median and 95th percentile of years of education? **(2 points)**
  + Median: 12
  + 95th percentile: 15
  1. What are the mean, standard deviation, and sample sizes for mother’s age in each of the five marital groups? For clarity, we are looking for you to produce the output needed to fill in the table below. **(4 points)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Marital status** | **Age (mean)** | **Age (SD)** | **Sample size** |
| Married | 25.29853 | 5.49711 | 46 |
| Separated | 27.15685 | 5.395131 | 44 |
| Divorced | 27.81504 | 5.32456 | 45 |
| Widowed | 30.17241 | 5.089163 | 40 |
| Never married | 22.11265 | 4.648772 | 43 |

BONUS QUESTION: What are the patient IDs of the 5 subjects whose children had the smallest birthweights? **(5 points extra credit)**

* Assuming we are still excluding babies below 300 grams:
  + 9170101
  + 9293903
  + 1276203
  + 3214103
  + 9145905

patid delmo deldy delyr enrmo

Min. :1200005 Min. : 1.000 Min. : 1.00 Min. :1984 Min. : 1.000

1st Qu.:5269709 1st Qu.: 3.000 1st Qu.: 8.00 1st Qu.:1986 1st Qu.: 3.000

Median :7076300 Median : 6.000 Median :16.00 Median :1987 Median : 7.000

Mean :6380925 Mean : 6.458 Mean :15.67 Mean :1987 Mean : 6.452

3rd Qu.:7578503 3rd Qu.: 9.000 3rd Qu.:23.00 3rd Qu.:1988 3rd Qu.: 9.000

Max. :9500406 Max. :12.000 Max. :31.00 Max. :1989 Max. :12.000

NA's :610 NA's :610 NA's :610 NA's :20

enrdy enryr momage raceth grade

Min. : 1.00 Min. :1984 Min. :16.00 white :4219 Min. : 0.00

1st Qu.: 8.00 1st Qu.:1985 1st Qu.:20.00 hispanic:4023 1st Qu.:10.00

Median :15.00 Median :1986 Median :23.00 black :5043 Median :12.00

Mean :15.51 Mean :1986 Mean :24.32 Mean :11.52

3rd Qu.:23.00 3rd Qu.:1987 3rd Qu.:28.00 3rd Qu.:12.00

Max. :31.00 Max. :1989 Max. :46.00 Max. :27.00

NA's :20 NA's :20 NA's :6 NA's :30

marstat cigs1 cigs2 etoh1 etoh2

married :5699 Min. : 0 Min. : 0.00 every day: 74 every day: 32

separated :1206 1st Qu.: 0 1st Qu.: 0.00 3-5/wk : 123 3-5/wk : 40

divorced : 811 Median : 0 Median : 0.00 1/wk : 558 1/wk : 348

widowed : 58 Mean : 4 Mean : 3.07 <1/wk : 624 <1/wk : 548

never married:5499 3rd Qu.: 3 3rd Qu.: 2.00 <1/month :1932 <1/month : 1872

NA's : 12 Max. :92 Max. :70.00 never :9972 never :10442

NA's :1 NA's :1 NA's : 2 NA's : 3

partyr pregnum delges bw induclab

Min. : 0.00 Min. : 1.000 Min. :23.00 Min. : 320 Mode :logical

1st Qu.: 1.00 1st Qu.: 1.000 1st Qu.:38.00 1st Qu.:2960 FALSE:11660

Median : 1.00 Median : 2.000 Median :39.00 Median :3295 TRUE :1160

Mean : 1.26 Mean : 2.642 Mean :39.06 Mean :3268 NA's :465

3rd Qu.: 1.00 3rd Qu.: 3.000 3rd Qu.:41.00 3rd Qu.:3630

Max. :200.00 Max. :16.000 Max. :52.00 Max. :5650

NA's :223 NA's :5 NA's :477

auglab intrapih csection

Mode :logical Mode :logical Mode :logical

FALSE:8325 FALSE:12022 FALSE:10008

TRUE :4498 TRUE :925 TRUE :2628

NA's :462 NA's :338 NA's :649