**Me@morandum**

To: NTRHD Intern

From: Dr. Brad Cannell, Director, NTRHD

Re: Analyze Infant Birth Data

**Overview:**

The Health Department needs your help analyzing a new data set containing birth and infant death data. The North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill make publicly available birth and infant death data for all children born in the state of North Carolina. Records on birth data go back to 1968. The comprehensive data set for the births in 2001 contains 120,300 records. The data below represents a random sample of 29 of those births and selected variables. [Please click here to view the codebook](https://www.dropbox.com/s/0x3zylubecnszwg/NCBIRTH%20Codebook.pdf?dl=0) and use this data set to complete the following tasks.

1 1 1 32 40 1 1 N 38 0 0 111 3146.85 0 0

2 1 2 32 37 1 1 N 34 0 0 116 3288.6 0 0

3 1 1 27 39 1 1 N 12 0 0 138 3912.3 0 0

4 1 1 27 39 1 1 N 15 0 0 136 3855.6 0 0

5 1 1 25 39 1 1 N 32 0 0 121 3430.35 0 0

6 1 1 28 43 1 1 N 32 0 0 117 3316.95 0 0

7 1 2 25 39 1 1 N 75 0 0 143 4054.05 0 0

8 1 2 15 42 2 1 N 25 0 0 113 3203.55 0 0

9 1 2 37 41 1 8 N 31 0 0 139 3940.65 0 0

10 1 2 21 39 1 1 N 28 0 0 120 3402 0 0

11 1 2 28 37 1 2 N 18 0 0 110 3118.5 0 0

12 1 2 27 40 2 1 N 37 0 0 124 3515.4 0 0

13 1 1 26 41 1 1 N 45 0 0 121 3430.35 0 0

14 1 2 20 41 2 1 N 52 1 0 112 3175.2 0 0

15 1 2 21 41 2 2 N 15 0 0 105 2976.75 0 0

16 1 2 19 40 1 1 N 26 1 0 131 3713.85 0 0

17 1 2 18 36 2 1 P 31 1 0 88 2494.8 1 1

18 1 2 26 31 2 1 N 40 1 1 93 2636.55 1 1

19 1 1 24 40 1 1 M 51 0 0 140 3969 0 0

20 1 1 24 38 1 1 N 45 1 0 105 2976.75 0 0

21 1 2 17 40 2 2 N . 0 0 121 3430.35 0 0

22 1 1 36 37 1 1 N 25 0 0 91 2579.85 0 0

23 1 1 21 30 1 1 N 0 1 1 51 1445.85 1 1

24 1 1 21 38 1 1 P 20 0 0 72 2041.2 1 0

25 1 2 26 39 1 1 N 20 0 0 116 3288.6 0 0

26 1 2 30 38 1 1 N 38 0 0 114 3231.9 0 0

27 1 2 32 43 2 1 M 15 0 0 114 3231.9 0 0

28 1 1 18 40 2 2 N 20 0 0 116 3288.6 0 0

29 1 2 21 39 2 2 N 27 0 0 120 3402 0 0

**Task 1**. Please create a data frame in R from this data.

* Name the data frame **ncbirth30**.
* In addition to the variables listed in the codebook, I’ve added a participant id column to the data above. It is the first column. Please name it **p\_id**.

**Task 2.** Use R to calculate the number of missing values, mean, median, standard deviation, minimum value, and maximum value for the variables **mage**, **weeks**, **gained**, **tounces**, and **tgrams**.

**Questions**

**1.** Which variable is missing 1 observation?

**2.** Which variable appears to have the largest dispersion (in absolute terms)?

**Task 3.** Use R to calculate the frequencies and percentages of each non-missing category of the variables **smoke**, **drink**, **low**, and **premie**.

**Questions**

**3.** What percentage of participants smoked during pregnancy (rounded to the nearest integer)?

**4.** What percentage of infants were not born premature (rounded to the nearest integer)?

**5.** How many non-missing values are there for the variable **drink**?

**Task 4.** Use the ggplot2 package to create histograms of the variables **mage**, **weeks**, **gained**, **tounces**, and **tgrams**. Please set the bin argument to the value 10.

**Questions**

**6.** Which of the variables you created a histogram (i.e., **mage**, **weeks**, **gained**, **tounces**, and **tgrams)** for appear to have long tails out to the left?

**Task 5.** Use the ggplot2 package to create boxplots for the variables **mage**, **weeks**, **gained**, and **tounces**.

**Questions**

**7.** Which of the following variables listed above (i.e., **mage**, **weeks**, **gained**, and **tounces)** has the greatest number of outliers?

**8.** All of the outlying values for the variable **weeks** occur [above\_below] 32 weeks?

**NOTE**: You can just resubmit the same code you submitted in the previous task above. Submitting it again let's me know that I should check it for the bonus.