DATA-LA 320

Midterm Exam

In a R script file, answer the following questions. Provide the code to re-create your answers, datasets, figures, etc. Be sure to include comments – particularly when asked to provide interpretation of the graphs. If you get stuck figuring out one of the questions, but you need that variable or dataset in the next question, select a different variable to use and indicate it in your code. For example, if you cannot create the dataframe in #7, continue to use the NHANES dataset. If you cannot create AvgNumSexPar in #8, use SexNumPartnLife for the remainder of the assignment. Your midterm is due: Tuesday, October 15th by the start of class (noon).

1. Load the NHANES dataset (it’s in the NHANES library, which you may have to install). Provide an overview of this dataset. How many variables (columns) are there and how many observations (rows) do we have?
2. What is the data type for the **Education** variable? What are the possible values of **Education**?
3. What is the largest head circumference (**HeadCirc**) in this dataset? What is the ID number **(ID)** and age of the child (in months – **AgeMonths**) with the largest head circumference (**HeadCirc**)?
4. If you look at the relationship between Age and AgeDecade

**table(NHANES$Age, NHANES$AgeDecade)**

you will see that those people who are listed as age 80 (which actually represents 80 or older), do not have an **AgeDecade** value. Create a new variable: **AgeDecade2** which categorizes age into 10-year increments and ends with 80+.

1. How many people in this survey have had more than 4 children (**nBabies**) and sleep more than 8 hours per night (**SleepHrsNight**). Note: You will have to exclude those who have missing values for either nBabies or SleepHrsNight. You will also have to account for duplicate IDs (the same people surveyed more than one year).
2. Of those people who have had more than 4 children (**nBabies**) and sleep more than 8 hours (**SleepHrsNight**), what are the ages ranges of those people? (Use the AgeDecade2 variable you just created in #5). Present this information as a table. Interpret your findings.
3. Create a new dataframe that only includes adults ages 18 to 45, only including the following variables: **ID, Age, AgeDecade2, Gender, Education, MaritalStatus, SexNumPartnLife, SexOrientation**. How many adults are in this dataset? (Note: the select() function occurs in many packages, so you will need to utilize the dplyr package if it’s not your default).
4. Create a new variable (in your new dataframe), **AvgNumSexPar**, defined as the number of sexual partners per year beginning at age 18. As an example, a person who is 25 and has had 10 sexual partners, would have an average number of sexual partners of 10 / (25-17) = 10/8 = 1.25 sexual partners / year (since 18). (Note: We will divide by Current Age – 17 so that those individuals who are currently 18 don’t have a 0 in the denominator).
5. Using your new dataframe, create a histogram of **AvgNumSexPar**. Describe the distribution. Is there an outlier? Provide details on this outlier? (What is their age, gender, total number of sexual partners)?
6. Using your new dataframe, create a density plot of AvgNumSexPar with fill by Gender. Convert the x axis to a log2 scale. Interpret your findings.
7. Using your new dataframe, create a scatter plot of Age by AvgNumSexPar, with the y axis on a log2 scale. Indicate gender by color.
8. Using your new dataframe, create a linear model ‘predicting’ AvgNumSexPar (y) by Age (x). Use your linear model’s predictions to plot a green line over the scatter plot you made in question 11. (Note that the log2 transformation will make your straight line look curved). Does the linear model make sense on this dataset? Why, or why not?