

# STAT 331 Lab 4

## Instructions

Please submit an HTML document created using R Notebook, but you are more than welcome to test your code out in an R script first.

For this exercise, you will work with the American Time Use Data Set. This is a survey of how a sample of Americans spent their time during a 24-hour period. Respondents record their activities during different time steps, and also report on their experience during the activity and who they interacted with. We will work with two data sets located at <http://www.bls.gov/tus/wbdatafiles.htm>, which are also available on PolyLearn. Variable descriptions are in the following codebook: <http://www.bls.gov/tus/wbmintcodebk.pdf>. You will need this to solve some of the problems.

*BE SURE TO SAVE YOUR WORK REGULARLY!!!*

## Exercises

1. Read the ATUS 2010 WB Respondent file into R. This data set has information about each of the respondents who were surveyed. Name your data frame “respondents.”
2. Display summary statistics of all variables in the data frame.
3. Find the dimensions of the data frame.
4. Display the first 6 observations of the data frame.
5. Display the variable names and types.
6. Perform a cross tabulation of werest (how well-rested respondents felt the day before the survey) and wgenhth (general health). Can you interpret the results? Why or why not?
7. Now let's working on making these variables easier to interpret in R. The variable werest is the response to how well-rested respondents felt the day before the survey (1=very, 2=somewhat, 3=a little, 4=not at all). Create a factor variable, named “rested”, which has the categories of how well-rested respondents were. Cross-tabulate rested with werest. Is your coding correct?
8. Remove werest from your data frame, and add rested to it. Re-examine the variable names and types.
9. Remove the variable rested from your workspace.
10. Create a table of rested values.
11. Following the same steps as you did with werest, create a factor variable named “genhealth”, for wgenhth (general health). You will need to locate the codes in the codebook.
12. Now cross-tabulate rested and general health again, this time with the factor variables. Do you see evidence for a relationship? (*You do not need to do a formal statistical test.*)
13. Let's also examine reports on each individual's activity. Download the ATUS 2010 WB Activity file, and read the activity data file into R. Name your data frame “activity”. While in the previous data set, each row was a respondent, here each row is an activity, with respondents reporting on multiple activities per day.
14. Compute summary statistics of the activity data frame. Find its dimensions. Examine the first 6 rows to make sure you've read it in correctly. Determine the type of each variable.

15. In this data set, missing values were coded as numbers. Recode these as NA in your data frame: -1 = Blank ; -2 = Don't know ; -3 = Refused. How many missing values are in this data set?
16. The variable "wuhappy" is a happiness score ranging from 0 to 6 that respondents recorded for each activity. Create a variable, "happy" which is the mean happiness score for each respondent over the course of the day.
17. Compare the length of "happy" to the dimension of respondents. Is it correct?
18. Compute the median happiness score for each activity performed (TUACTIVITY\_N), ignoring missing values.
19. Compute the variance of each variable in the activity data frame, except for TUCASEID. Exclude missing values from your calculation.
20. Merge the two data frames together by TUCASEID, making sure you keep all observations in the activity data frame. Call your new data frame "ATUS".
21. Compute the dimensions of ATUS.