Consulting Assignment: Data dictionary, random sample

Task	Deliverables	Time Estimate and Deadline
1. Choose dataset (> 100,000 observations) and ***thirty*** variables. Variables should be chosen such that two different analyses can be performed from the list below. At least one categorical and one quantitative variable. ○ Dependent means or proportions ○ One-way ANOVA ○ Factorial ANOVA ○ Multivariable regression ○ Chi-square analysis	Provide an initial data dictionary for the 30 variables. Variable name, type, label, format, notes	
2. Produce a reproducible simple random sample (without replacement) of size 500 from your dataset.	Provide code to produce. Provide dataset in csv format	
3. Produce a reproducible stratified random sample (without replacement) using proportional sampling for a total sample size of 250 from your dataset. Choose a categorical variable to use as your stratification factor. For sampling, use only the code you developed for part 2. That is, do not add any additional options to the simple random sample.	Provide code to produce. Provide dataset in csv format	

File choices:

- The 2016 National Youth Tobacco Survey
- The 2013-14 National Adult Tobacco Survey
- The 2015-2016 National Health and Nutrition Examination Survey
- The 2016 National Health Interview Survey – sample adult
- The 2016 National Health Interview Survey – sample child
- The 2016 Medical Expenditures Panel Survey
- The National Longitudinal Survey of Youth - 1979
- The National Longitudinal Survey of Youth – 1997
- General Social Survey - 2016
- Or other large database file approved by instructor