Stat UN2104 Take-Home Quiz: Due Monday, April 30, or Thursday, May 3

The assignment is to select one of the two data sets described below, analyze it, and write a report from your work. No specific questions are provided for this exercise. You should treat it as an open-ended data analysis problem with your analytical approach driven by questions that seem to be relevant for your problem.

Your report should include quantitative information written at a technical level appropriate for this course. Selectively include some of your output but not all of it. Your report should be written to "business standards" in terms of its writing style; it is ok to include a limited amount of code in order to make it clear what you are doing, but do not simply paste together your code. The analysis process you choose is important, so your report should convey what you investigated, why, and what the results were. Communicating the results is also important, so provide interpretations of coefficients and use graphics where appropriate. The report should conclude with a brief, final paragraph summarizing in lay terms for the general public the most important results you found (i.e., brief executive summary). The length of the report should be around 5 to 8 pages with an absolute maximum of 10 pages.

The work is to be entirely your own. You are not to use any outside resource beyond the textbook, lecture notes and R. No group collaborations are permitted, unlike the case with homework assignments. If you have questions about your analysis and want to discuss it during office hours, I will be happy to do so.

This exercise is intended to take about as much time (or perhaps a bit more, but not a lot more) than a standard weekly homework assignment. Hence, it can be handed in at class on Monday, April 30. To avoid special requests and circumstances, however, **everyone** is given a three-day extension through the end of the reading period. Hence it can also be left by **5:00 pm on Thursday, May 3**, in the course dropbox in SSW 904. Please note that this is 5:00 pm and not later on that day. Your report is due as a printed copy, not an electronic file.

Option A: The problem concerns an employee satisfaction survey for a large company that has many employees spread across the country with various demographic characteristics. Suppose employees responded as Satisfied or Unsatisfied with their job, and they were categorized on the basis of geographic region, race, age and gender. The general goal is to understand employee satisfaction as it relates to the available variables. The data are in CourseWorks as jobsat-data.txt with the following coding for the variables:

> **region**: 1 = Northeast; 2 = Mid-Atlantic; 3 = Southern; 4 = Midwest; 5 = Northwest; 6 = Southwest; 7 = Pacific.
>
> **race**: w = white; o = other.
>
> **age**: 1 = less than 35; 2 = 35-44; 3 = greater than 44.
>
> **gender**: f = female; m = male.
>
> **sat**: count of the number of employees with these categories who report Satisfaction with their work.

unsat: count of the number of employees with these categories who report they are Unsatisfied with their work.

Thus, these data can be thought of as a $7 \times 2 \times 3 \times 2$ table along with the binary response.

Option B: This problem concerns credit scoring, that is, trying to predict if a consumer is credit worthy or not based on using various available explanatory variables. The goal is to build a model that could be used to predict whether or not a new customer is credit worthy. Suppose data are available on 1000 customers of a German bank, of whom 700 had been deemed credit worthy and 300 not, along with a set of 7 possible predictor variables. The response (credit worthy or not) and the 7 possible predictor variables are in CourseWorks as credit-data.txt. Each of the 7 explanatory variables was coded, though only limited information was supplied on what the codes mean. The available coding information is given below.

y: response, 1 = credit-worthy and 0 = not credit-worthy

running_acct: coded 1-4, information on balance of current account; (1-no running account; 2-no balance or debit; 3-small balance; 4-larger balance or checking account for at least 1 year)

duration: duration in months (continuous variable, not coded)

payment: coded 0-4, payment of previous credits (0-hesitant payment of previous credits; 1-problematic running account, credits running but at other banks; 2-no previous credits/paid back all previous credits; 3-no problem with current credits at this bank; 4-paid back previous credits at this bank)

purpose: coded 0-10, purpose of credit (0-other; 1-new car; 2-used car; 3-furniture; 4-radio/television; 5-household appliances; 6-repairs; 7-education; 8-vacation; 9-retraining; 10-business)

marital_and_gender: coded 1-4, combination of information on gender and marital status (1-male: divorced/living apart; 2-female: divorced/living apart/married or male: single; 3-male: married/widowed; 4-female: single)

current_employ: coded 1-5, length of time with current employer (1-unemployed; 2-less than one year; 3-one to four years; 4-four to seven years; 5-greater than seven years)

age: age in years (continuous variable, not coded)

With these data for the variables that have many levels, you might want to combine some coding levels based on their meanings. You also might want to treat these as discrete categories or create scores from them. Similarly, you might, or might not, want to create categories from the continuous variables of duration and age.