70% of your grade is programming accuracy and elegance (appropriate utilization of SQL and other functions) and 30% of your grade is on presentation and clarity of your answers.

Consider the HW DATA1 and HW DATA2 datasets

a. Create a single dataframe. Only keep records (identified by MATCHKEY) in BOTH datasets. Select AGE, RBAL, TRADES, MATCHKEY, CRELIM, DELQID <create a screenshot>.

b. Prove that there are no duplicate MATCHKEYS in the file.

c. Report the structure and overall summary statistics from your new dataset. <create a screenshot>.

d. Create the following new variables:

   • If the DELQID = 0 then RISK = NEW, if the DELQID = 1 or 2 then RISK = GOOD, if the DELQID = 3 + then RISK = BAD.

   • If the CRELIM <=2000 then LIMIT = 1, if the CRELIM 2000 <=5000 then LIMIT = 2, if the CRELIM >5000 then LIMIT = 3.

   e. Any value that is more than 4 standard deviations above the mean should be imputed with the median.
f. Answer the following questions (BASED UPON THE DATA PROVIDED):
   - What is the average number of trades for each risk category? Does trades appear to be related to risk?
   - Do riskier people have lower credit limits?
   - Does age appear to be related to risk?
   - Does revolving balance appear to be related to risk?

g. Export your final file, sorted by CRELIM within RISK to a CSV file <create a screenshot>. Please format your answers to all questions neatly and professionally (you are encouraged to provide tables as appropriate).

h. Provide all of your annotated code as an appendix. Your code MUST match your output.