STAT 440 – Midterm Exam
Due Friday, March 4 at 10:00am

Sharing or copying any part of the exam is an infraction of the University's rules on Academic Integrity and will be disciplined accordingly. Neither the TA nor I will be available to offer help with solving the exercises as we do with HW. Further, you are not to consult with fellow students or any other person regarding the midterm. You are allowed to consult e-Learning modules, online notes, lecture notes, SAS documentation, and the recommended textbooks – just not people.

You must complete the exercises and turn in the SAS program file and Report just like with HW. Submissions must be uploaded to our Compass 2g site on the Exams page. No email, hardcopy, or late submissions will be accepted.

Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\midterm. Save all data files accompanying this assignment in that folder. If you cannot create the folder because you are working on a university computer and don't have permission, create the …\440\midterm folder elsewhere.
b. Assign the library reference **midterm** to the folder 'C:\440\midterm'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **midterm** to your …\440\midterm folder.
   Note: If you are using a folder other than 'C:\440\midterm', you must change any pathname references in your program file to 'C:\440\midterm' before submitting your homework.

Submitting your work to Compass 2g

You are to submit two (and only two) files for your homework submission.

1. Your SAS program file which should be saved as **midterm_*YourNetID*.sas**. For example, my file for the midterm would be midterm_dunger.sas. All program statements and code should be included in one program file.
2. Your Report including all relevant output to address the exercises. For this, use ODS to send your results to a Portable Document Format (PDF) file called **midterm_*YourNetID*.pdf**. Only include your final set of output. Do not include output for every execution of your SAS program. Use the template file **hw1 template.sas** as your guide.

You be allowed a maximum of ***THREE*** submissions for this midterm, but only the last one will be viewed and graded. Midterm submissions must come as a pair of files, as described above.

*** SAVE OFTEN ***

1. In this exercise, you will work with a data set containing information from the 124 seasons of University of Illinois football. The raw data set **illinifb.dat** contains data from 1892 to 2015.

Field	Description	Notes
1	Obs	Observation number
2	Season	
3	Conf	Conference
4	W	Wins
5	L	Losses
6	T	Ties
7	Pct	Win percentage
8	SRS	Simple Rating System: A rating that takes into account average point differential and strength of schedule. Average of all teams in a season is 0.
9	SOS	Strength of Schedule: Average of all teams in a season is 0.
10	AP_pre	Rank in pre-season AP poll. Possible values are 1-25 and missing if unranked.
11	AP_high	Highest rank of the team in the AP poll during that season.
12	AP_post	Rank in final AP poll at the end of the season.
13	Coach	Head coach (or coaches)
14	Record	Coach's record
15	Bowl	Post-season bowl game played in, or missing
16	BowlResult	Result of Bowl game: W or L

   a. Write a DATA step to read the values of **illinifb.dat** into SAS. The output data set is to be a SAS data file called **illinifb*_NetID***. Perform the following tasks in the DATA step as well.

      - Use the Description column above to name the variables.
      - Choose appropriate labels for variables whose names are abbreviated.
      - Apply a format to Pct so that the values display in the unit of percentages. For example, 0.123 would be 12.3%.
      - Prevent the Obs variable from being displayed in the output data set.

   b. Include output showing the descriptor portion of **illinifb*_NetID*** in the Midterm Report.

*** SAVE OFTEN ***

Perform data validation to check the following. These results from parts c-f should each appear in the Midterm Report.

c. Print a table which succinctly confirms that each value of Season is unique.

d. Print a table to check the values of W, L, T, and Pct.

e. Winning percentage is equal to the number of wins (W) divided by the total number of games (W+L+T). Using a procedure, find all observations that violate this requirement and print only the variables Season, W, L, T, Pct, and Record.

   Note: There might be no violations of this rule. If so, print only the observation from 2016 and include only the variables mentioned above.

f. If there are any typos in a coach's name, each unique spelling would appear in a frequency report. Print a frequency report to check for this and to determine who presided as head coach for the most seasons.

Perform data cleaning to correct any anomalies you've discovered thus far.

g. Write a DATA step to clean the values of **illinifb_*NetID*** and create an output data set called **illinifb_clean_*NetID***.

   - If you found inconsistencies from your results in parts (d) and (e), fix them. Note that the values in Record are correct and reliable.
   - If you found any issues within your results in part (f), fix them. These may or may not include the following.
     - If more than one coach is listed for a season, clean the Coach and Record variables to note which coach had more wins that season.
     - If no coach is listed, it means that more than one person shared the duties of coach. Learn who they were by searching the internet and clean the Coach variable to contain the name of the coach whose last name comes first in alphabetical order. Clean the Record variable to match the W, L, and T entries.

Perform re-validation to check parts (e) and (f) by running that same code on **illinifb_clean_*NetID***. These results from parts h-i should each appear in the Midterm Report.

h. Redo of part (e).
   Note: There might be no violations of this rule. If so, print only the observation from 2016 and include only the variables mentioned above.

i. Redo of part (f).

*** SAVE OFTEN ***