STAT 440 – Homework 3

Students are encouraged to work together on homework. However, sharing or copying any part of the homework is an infraction of the University's rules on Academic Integrity.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

The HW Report should include the output generated from the following exercises:
1-b, 2-bcdf

Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\hw3. 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\hw3 folder elsewhere.

b. Assign the library reference **hw3** to the folder 'C:\440\hw3'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **hw3** to your …\440\hw3 folder.
Note: If you are using a folder other than 'C:\440\hw3', you must change any pathname references in your program file to 'C:\440\hw3' 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 **HW*n*_*YourNetID*.sas**. For example, my file for the HW3 assignment would be HW3_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 homework, use ODS to send your results to a Rich Text Format (RTF) file called ***YourNetID*_HW*n*.rtf**. Only include your final set of output. Do not include output for every execution of your SAS program. Use the template file **hw3 template.sas** as your guide.

Once the results have been sent to the .rtf file, you may open it in Word and include your own responses in the relevant areas (as directed in the exercises).

You have an unlimited number of submissions, but only the last one will be viewed and graded. Homework submissions must always come as a pair of files, as described above.

1. **Cleaning Data from shoes_tracker**
   In this exercise, you will continue to work with a data set from the Chapter 8 Lab Exercises. In the lab exercises, you uncover some data issues during the data validation procedures. Time to clean them up.

   a. Write a DATA step to read **shoes_tracker** to create a temporary SAS data set called **shoes_tracker_*NetID***. In the DATA step, include statements to correct the invalid data found in the lab exercises. Here are some notes to assist in the cleaning.
      - Most corrections can be inferred from the data set as a whole.
      - When present, the Supplier_ID can be assumed to be correct.
      - If the Product_ID is too short, add a 9 to the end. If the Product_ID is too long, drop the last digit from the end.
      - Make sure all character values appear in their proper case.

   b. Write additional steps to verify that all the data requirements are met, i.e. no more errors exist in **shoes_tracker_*NetID***. (Include results in the HW Report.)

      To be specific, check that…

      - All variables that originally had missing values no longer have any missing values.
      - Supplier_Country must have a value of GB or US.
      - A Supplier_ID of 2963 corresponds to 3Top Sports and 14682 corresponds to Greenline Sports Ltd.
      - Product_ID must have only 12 digits.
      - Any adjustments made to variables in part (a) have been corrected.

2. Rushing data

In this exercise, you will work with a data set from the National Football League containing 1578 unique observations. The SAS data set **badrush.txt** contains data from the 2010-2014 seasons of professional American football for every player who record rushing yards (ran with the ball instead of being thrown the ball). If the player moves the ball forward, he receives positive rushing yards. If he's knocked backward, he records negative rushing yards for that attempt. But this data set has been tampered with and contains many data issues.

Variable	Name	Description
1	Season	
2	Player	Full name
3	Team	Player's team; max of 3 letters
4	Games	Number of games in which that player appeared that season
5	Att	Number of rushing attempts
6	Yds	Number of rushing yards
7	Avg	Average rushing yards per attempt, rounded to nearest 0.01
8	YPG	Average rushing yards per game, rounded to nearest 0.1
9	Lg	Longest rushing attempt
10	TD	Number of rushing touchdowns (i.e., scores, "goals")
11	FD	Rushing first-downs

Data Entry

a. Write a DATA step to read the values into SAS. The output data set is to be a temporary SAS data file called **rushing_*YourNetID***. Include code to ensure that variables have appropriate attributes.

   - This can just be a "first attempt" DATA step to just get the raw data into SAS. It does not have to fix all value errors; it just needs to get all 1578 observations into SAS.
   - Note that there are issues that invalid data issues that will need to be discovered through verification and then cleaned. You don't need to address those now.
   - You should not alter the raw data file to read it in.

b. Print the descriptor portion of your new SAS data file once completed. There should be 1578 observations. (Include results in the HW Report.)

Data Validation

c. After your "first attempt," examine the log for invalid data errors or other notes identifying issues. (Include just the log notes from the DATA step execution that results in 1578 observations in the HW Report by using copy-paste. Do not include the entire session log!)

d. Write any and all necessary steps to check the values of the newly created SAS data set. For each table you create, write one and only one sentence for each variable that has an issue. For example,…

   - Var1 has missing values.
   - Var2 has at least one observation out of range.

Some notes and things worth checking:

- The data set contains observations from 2010 to 2014.
- A player can appear in at most 16 games per season.
- The calculations of Avg and YPG should be inspected.
- The field is only 100 yards long, so the longest rush can't exceed that.
- If the longest rush is a negative number then the value of total yards is negative. (The converse is not true.)

(Please include only one relevant table for each variable (i.e., those that identify possible data issues) in the HW Report.)

Data Cleaning

e. Write a DATA step that programmatically corrects all invalid data issues according to your table in part (e). The output data set is to be a permanent SAS data file called **rush_*YourNetID***.

Some notes on cleaning:

- Some invalid values many have too many 0's.
- The missing player in the 2013 NFL leader in total rushing yards.
- If the longest rush is only a negative number then the value of total yards is negative. (The converse is not true.)
- There should be no missing values when cleaning is completed.

f. Run the same validation steps you wrote in part (d) again and verify that all invalid data issues that you can control have been corrected. (Include only relevant tables and results, i.e., those that identify possible data issues, in the HW Report.)