STAT 440 – Homework 5

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-ace, 2-abc, 3-bd

Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\hw5. 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\hw5 folder elsewhere.
b. Assign the library reference **hw5** to the folder 'C:\440\hw5'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **hw5** to your …\440\hw5 folder.
   Note: If you are using a folder other than 'C:\440\hw5', you must change any pathname references in your program file to 'C:\440\hw5' 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 HW4 assignment would be HW4_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. **Employee data**
   You will be working with the SAS data file **employee_roster5** which contains many variables regarding the employees of the Orion corporation, but it has been tampered with.

   a. If you check the values of Employee_ID in the data set, you will find duplicates. Print a table that succinctly summarizes the duplicate Employee_ID values. No need to include values of Employee_ID in the table that are unique. (Include results in the HW Report.)

   b. Clean the values of Employee_ID in a DATA step that creates a new permanent data set called **employee_roster_*NetID***.
      - Note that if there are duplicate occurrences of an ID, only one is correct and the others are incorrect.
      - Hint: To figure out the correct values for Employee_ID, you will have to do some addition investigation of Employee_ID and some of the other variables to figure out how they need to be fixed. Employees with ID's close together in enumeration share similar values in other variables.

   c. Rerun the same step from part (a) that generated the table in your report, but this time the data set being used should be **employee_roster_*NetID*** (the one created in part b). If part (b) was done correctly, running the step again should reveal that all issues have been fixed. (Include results in the HW Report. Either include the table or the portion of the SAS Log generated by the table.)

   d. Create a temporary user-defined format for Gender. Possible values are Male, Female, and Not Given.

   e. Create a two-way table of Department (rows) and Gender (columns). (Include results in the HW Report.)
      - Suppress the row and column percentages.
      - Identify which department employs the most males and which employs the most females in one sentence after the table.

2. Baseball data

The SAS data set **batting** contains a complete history of Major League Baseball's (MLB) batting data from 1871 through the 2010 season. Each observation holds a single season of batting statistics for a single player. So each observation contains a unique combination of PlayerID and YearID.

The SAS data set **master** contains a complete list of every player, every manager (i.e., head coach), and other notable people in MLB history from 1871 through 2010. Each observation holds a single person.

a. Write a single DATA step to create the following temporary data sets from **batting** and **master**.
   i. **mostruns_*NetID***
      - This should contain all observations with at least 100 runs.
      - Include only variables for First Name, Last Name, Year, Team, and Runs.

   ii. **power_*NetID***
      - This should include all observations with at least 1 homerun.
      - Include only variables for First Name, Last Name, Bats, Year, HR, and RBI.

   iii. **bestavg_*NetID***
      - This should include a variable for Plate Appearances (PA) which is defined as the sum of these other variables: PA = AB + BB + HBP + SH + SF.
      - Only players with at least 3.1 plate appearances per game played can be considered for the batting title. So, only include observations where PA/G is at least 3.1.
      - This data set should also include a variable called Batting Average (AVG) where AVG = Hits/AB.
      - Include only variables for First Name, Last Name, BirthDate, Year, PA, and AVG.

   Include the SAS Log notes (just the blue ones at the end) that display how many observations and variables are in each of the newly created data sets.

b. Print a table that includes the mean, median, and maximum number of HRs partitioned by Bats. Print a similar table for RBIs. (Include results in the HW Report.)

c. Print a table that includes the five-number summary (min, Q1, median, Q3, max) for AVG partitioned by decade. That is, every observation from 1871-1879 should be included in the "1870s"; every observation from 1880-1889 should be included in the "1880s"; … every observation from 2001-2009 so be included in the "2000s". Exclude observations from 2010. (Include results in the HW Report.)

3. Survey data

You will be working with the SAS data files **demographic**, **survey1**, and **survey2**.

a. Merge the **demographic** and **survey1** data sets based on a common identifier to create a new, temporary SAS data set called **demo1_*NetID***. In **demographic**, the identifier is called ID. In **survey1**, the identifier is called Subj. Both are character variables.
b. Print the data portion of **demo1_*NetID***. (Include results in the HW Report.)

c. Merge the **demographic** and **survey2** data sets based on a common identifier to create a new, temporary SAS data set called **demo2_*NetID***. In **demographic**, the identifier is called ID and it is a character variable. In **survey2**, the identifier is also called ID, but it is a numeric variable.
d. Print the data portion of **demo2_*NetID***. (Include results in the HW Report.)