ZADA Solving Business Cases using Applied Data Analytics
Problem Set 1
Due Thursday April 27 5:00pm Germany time

Please work on this problem set as an individual, not with other students. Please let me know if I have not clearly worded a question, and I will be happy to clarify what the question is asking for. Because this is a 'take home' problem set designed to assess your learning during the first half of module 1, please know that I do not plan to answer questions about how to write code.

Because this problem set covers all material since the beginning of the course, you may need to look back at the videos, textbook or homework problems to review the material for some questions. If you are not able to complete all instructions for a question, complete the parts that you can and I can assign partial credit.

There are ten (10) data files for this problem set, including one comma-separated value file, one Excel files, six SAS files, and two text files.

Note: Some questions in this problem set will require more thought and creativity than the textbook homework questions. This is a natural part of the learning process, because we are preparing you to have the skills to perform the more comprehensive case studies in June.

Please name your code files as follows:

- *YourlastnamePS11*
- *YourlastnamePS12*
- *YourlastnamePS13*
- *YourlastnamePS14*
- *YourlastnamePS15*
- *YourlastnamePS16*

Question 1
*This question applies course concepts from the modules **Load data I** and **Load data II***

The text file *EventMedals* contains seven variables [*year, season, sport, event, gender, points, country*] related to the 2020 Summer Olympics in Tokyo Japan. Please write code to:

A. Load the data into SAS. *Hints: Text file does not include variable names (you can assign any variable names), and you will need to count the column numbers. Full credit for using the appropriate command to load the data, partial credit for copying the text file into the SAS command window*

B. Create a new variable *medal* based on values in the *points* variable. *Hint: Set the variable length at six to avoid truncating the values*
   1. Three points indicates a Gold medal
   2. Two points indicates a Silver medal
   3. One point indicates a Bronze medal

C. Display *gender* values to the screen as follows. *Hint: Do not create a new variable*
   - 'Male' when gender value is M
   - 'Female' when gender value is F
   - 'Mixed' when gender value is X

D. Display only five variables to the screen (in the order below) including the instructions in B and C above, showing the variable names as 'Olympic Sport', 'Olympic Event', 'Gender', 'Medal', and 'Country'. Add a title '2020 Tokyo Olympic Medals'

Question 2
This question applies course concepts from ***Format data I***

The Excel file *WinterMedalsTotal* contains the number of medals earned by each country in all Winter Olympics, plus some demographic statistics for each country. Please write code to:

A. Load the data into SAS

B. Create a variable named *populationmedals* to indicate the number of medals per 100,000 citizens of the country. The concept is 'what is the population density of medal winners in this country?'

C. Create a variable named *gdpmedals* to indicate the amount of Gross Domestic Product (GDP) per medal. The concept is 'what financial resources are required for each medal in this country?'

D. Display only the *country* variable and the new variables you created in parts B and C to the screen. Format the variable in part B to display with four digits after the decimal point, and format the variable in part C to display with commas separating thousands and zero digits after the decimal point

Question 3
This question applies course concepts from ***Format data II*** *and* ***Format data III***

The comma separated value file *OlympicHosts* contains host and date information for every Olympic Games from 1896 to 2022. Please write code to:

A. Load data into SAS. Full credit for using command to load data, partial credit for copy and paste

B. Create two new variables named *season* and *year* by separating data in the *seasonyear* variable

C. Create two new variables named *city* and *country* by separating data in the *citycountry* variable

D. Create a new variable named *enddatetime* by combining the month, day, year and time variables. *Hint: while there are different methods to do this, one method could be a two-step process first using the month-day-year command and then using the days-hours-minutes-seconds command*

E. Create a new variable named *hourslength* by computing the difference in hours between the *startdatetime* and *enddatetime* variables

F. Display only these variables to the screen in this order (*year*, *season*, *city*, *country*, *hourslength*), with results primary sorted in ascending order of *year* and secondary sorted with Winter before Summer

Question 4
This question applies course concepts from ***Format data III***

The SAS data file *SummerMedalsTotal* contains the total number of gold, silver and bronze medals for every country in the Summer Olympics from 1896 – 2020. Please write code to:

A. Load the data into SAS. Rotate the data such that rows become columns and vice versa. *Hint: You may save some time by copying in text from the file SummerOlympicCountries as part of your code*

B. Create a new variable named *totalpoints* awarding three points for each Gold medal, two points for each Silver medal, and one point for each Bronze medal. *Hint: Check whether you need to convert character variables into numeric values before you complete this step*

C. Display only the variables *country* and *totalpoints* to the screen, sorted in descending order by total points. Format the totalpoints variable with commas separating thousands, and zero digits after the decimal point

Question 5
This question applies course concepts from ***Report results II***

The SAS files *CountryMedals20022010* and *CountryMedals20142022* contain the number of Gold, Silver and Bronze medals per country for each Olympics from 2002 to 2022. Please write code to:

A. Load both files into SAS to create two datasets named *Years20022010* and *Years20142022*

B. Use a procedure to combine the two datasets to create a dataset named *Years20022022*. No credit for cut and paste

C. In the new dataset, create a variable named *totalmedals* as the total number of gold, silver and bronze medals

D. Display only three variables in this order (*country*, *year*, *totalmedals*) sorted in ascending order by country, then ascending order of year

Question 6
This question applies course concepts from ***Format data V***

The SAS file *ResultsOne* contains seven variables [*year*, *event*, *gender*, *medal*, *rank*, *gender*, *country code*] for the 1924-1972 Winter Olympics. The SAS file *Countries* contains two variables [*country code*, *country name*] for all countries. The SAS file *Events* contains two variables [*event*, *sport*] for all Winter Olympic sports. *Hint: Two of the variables above are each located in two different tables*

A. Use SQL to combine these three datasets into a single dataset. SQL is the only way to achieve full credit, partial credit may be possible with other commands

B. Use SQL to display only seven variables to the screen in this order [*year*, *sport*, *event*, *gender*, *medal*, *rank*, *country*]. *Hint: One of the variables above is located in two different tables*

C. From top to bottom, display the data sorted first by year, then by event, then by sport, then by rank

D. Give your output the title '1924-1972 Winter Olympics'