ZADA Solving Business Cases using Applied Data Analytics
Problem Set 2
Due Monday June 5 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. If you are not able to complete all instructions for a question, complete the parts that you can and I can assign partial credit. Because this problem set covers all material since the first problem set, you may need to look back at the videos, textbook or homework problems to review the material.

Hint: For some questions, I ask you to prepare output based on a subset of the full dataset. In these cases, you may (or may not) find it helpful to create one or more additional dataset(s) for specific questions (beyond the original dataset that you create by loading the datafile), and then run your commands on these additional dataset(s).

Note: Some questions in this problem set will require more thought and creativity than questions in the first problem set. This is a natural part of the learning process, because we are now preparing for you to have the skills to perform a comprehensive final project.

Please name your code files as follows:

- *YourlastnamePS21*
- *YourlastnamePS22*
- *YourlastnamePS23*
- *YourlastnamePS24*
- *YourlastnamePS25*
- *YourlastnamePS26*

Question 1
This question applies course concepts from ***Analyze data II***

The SAS datafile *olympicmedals* contains nine fields for all Summer and Winter Olympic games from 1896 – 2022 [*year, hostcity, gametype, event, sport, gender, medal, swedishscore, country*]. The *swedishscore* variable uses the 'Swedish' scoring system to assign points for each gold:silver:bronze medal on a 3:2:1 basis (1912 Olympics).

A. Produce a table that lists each country on the vertical axis, plus the number of medals won by that country for all Summer Olympics from 1936-2020. Sort this table in descending order based on the number of medals won.

B. Produce a table that lists each country on the vertical axis, each sport on the horizontal axis, with cells containing the number of medals won by each country in each sport for all Summer Olympics from 1936-2020. Sort this table in descending order based on the number of medals won.

C. Produce a separate table for each gender, that lists each country on the vertical axis, each sport on the horizontal axis, with cells containing the number of medals won by each gender for each country in each sport for all Summer Olympics from 1936-2020. Sort this table in descending order based on the number of medals won.

Hint: The maximum number of columns for parts B and C would be the number of Summer Olympic sports plus a total column.

Question 2
This question applies course concepts from ***Analyze data III***

The comma-separated value file *wintermedals* contains five variables for 32 countries that won medals in the 2002–2022 Winter Olympics [c*ountry, medals, per capita GDP, population, average temperature in Celsius*].

A. Load the data into SAS. *Hint: After loading the data, use a data statement to initialize the data set*

B. Generate statistical output that shows how the four numerical variables are correlated with each other. Add a title that states the correlation coefficient and *p* value for the variable that has the highest absolute (positive or negative) correlation with *medals*

C. Prepare a linear regression with *medals* as the dependent variable, and [*percapitagdp, population, temperature*] as independent variables. Add a title that states the regression coefficient and *p* value for the variable has the highest explanatory power (positive or negative) for rank. *Hint: The command for linear regression is provided in the assigned reading for this class meeting*

Hint: You will need to run your code for parts B and C first before you have information to add the titles. Please format the titles as follows: "Variable X, coefficient #, p #", with actual data instead of X and # symbols

Question 3
This question applies course concepts from ***Analyze data V***

The comma-separated value file *olympicceremony* contains comments from International Olympic Committee President Thomas Bach at the opening and closing ceremonies for the 2022 Beijing Winter Olympics

A. Please write code to:

- Import the CSV file into SAS
- Convert all text into lower case
- Produce output to show the total number of words in the text
- Create a separate record for each word
- Produce output to show how many times each word is used, and display this output in descending order

B. Hints:

- *All punctuation has already been removed from the file*
- *Full credit for importing file, partial credit for copying/pasting text*
- *Variable name 'speech' on first line of CSV file*
- *One line of your code will be as follows:*
  array word {1:1000} $20. _TEMPORARY_;

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

Based on the SAS data file *olympicmedals* (same datafile as question 1).

A. Load the file into SAS. Use a procedure to create a data set that contains one variable named *country* with the country name, and another variable with the total number of medals won by that country across all Summer and Winter Olympics from 1896-2022. *Hint: This data set will have 142 rows*

B. Use *mapsgfk.world* to create another data set. *Hint: This data set will have 20,000+ rows, so if you print this dataset to the screen please limit the number of observations that you print*

C. Bring the data sets from parts A and B together to produce a world map showing the total number of medals won by each country across all Summer and Winter Olympics from 1896-2022. Use seven levels to show seven gradations of color on the map

D. Add a title 'Total Olympic medals by country'

Question 5
This question applies course concepts from ***Report Results I***

Based on the SAS data file *olympicmedals* (same data file as question 1):

A. Generate output that lists:
- Country on vertical axis
- Sport on horizontal axis
- Cells containing total number of medals for each country for each sport for all Summer and Winter Olympics from 1994-2022
- *Hint: This table will have one row per country and one column per sport*

B. Generate output that lists:
- Country on vertical axis
- Year on horizontal axis
- Cells containing total Swedish score for each country in each year for each Summer and Winter Olympics from 1994-2022
- Include total row and total column, and include percentage for each cell in the table
- *Hint: This table will have one row per country and two columns per year*

Question 6
This question applies course concepts from ***Report results II***

Based on the SAS data file *olympicmedals* (same data file as question 1):

A. Produce output with the following characteristics:
- Country on vertical axis
- Medal on horizontal axis
- Cells containing total number of each type of medal (bronze, gold silver) for each country
- *Hint: This output will have one row per country, and a separate columns for each type of medal (bronze, gold and silver)*

B. Produce output with the following characteristics:
- Country on vertical axis
- Game type (Summer/Winter) on vertical axis below each country
- Sport on horizontal axis
- Cells containing number of medals for each country for each sport
- *Hint: This output will have up to two rows per country, and a separate column for each sport*