

Przetwarzanie i analiza danych w systemie SAS®

Kolokwium nr 1 - 13.11.2014

- Zapisać plik **szablon.sas** pod nazwą **nazwisko.sas** (gdzie **nazwisko** to nazwisko piszącego kolokwium). Na początku pliku należy wpisać w komentarzu własne imię i nazwisko.
- Rozwiązania zadań należy wpisywać do pliku **nazwisko.sas**. Plik należy we własnym interesie często zapisywać.
- Rozpakowane pliki z danymi wejściowym mają być umieszczone w bibliotece o nazwie **KOLO**.
- Wszelkie tworzone przez siebie zbiory sasowe należy umieszczać w bibliotece **WORK**.
- Tworzone programy powinny działać poprawnie bez żadnych zmian dla dowolnych zbiorów o takiej samej strukturze (tj. o takich samych zmiennych i ich atrybutach), jak wymienione w treści zadań. W szczególności, rozwiązania będą testowane na zbiorach różnych od podanych.
- Rozwiązania (tylko zawartość pliku **nazwisko.sas**, bez tworzonych zbiorów sasowych) należy przesłać na adres prowadzącego laboratorium, tzn.:

B.Kolodziejek@mini.pw.edu.pl lub K.Szpojankowski@mini.pw.edu.pl lub matysiak@mini.pw.edu.pl

List ma nie zawierać żadnych załączników – innymi słowy kod program musi być bezpośrednio wklejony do listu. We własnym interesie należy poczekać na potwierdzenie odbioru.

- Powodzenia!

1. (pkt) Plik tekstowy **a.txt** składa się z nieznanej liczby dwuwierszowych bloków. Liczby (oddzielone różnymi liczbami spacji) w pierwszym wierszu bloku oznaczają numery tych kolumn w drugim wierszu, w których zawarte są wartości dwóch zmiennych numerycznych x i y (x i y mogą występować w różnej kolejności - podana jest ona w pierwszych dwóch kolumnach pierwszego wiersza bloku). Wczytać plik **a.txt** do zbioru **a**.
2. (pkt) Załóżmy, że zbiór sasowy **b** ma znaną liczbę (10) zmiennych numerycznych o znanych nazwach x_1, \dots, x_{10} i nieznana liczbę obserwacji. Za pomocą jednego DATA STEPu, na podstawie zbioru **b**, utworzyć zbiór **bb**. Zbiór **bb** ma mieć dwie obserwacje i zmienne o nazwach x_1, \dots, x_{10} (takie same nazwy, jak nazwy zmiennych w **b**). Dla każdego $i = 1, \dots, 10$, i -ty element pierwszego wiersza zbioru **bb** to liczba wierszy, w których i -ta zmienna miała największą wartość wśród zmiennych x_1, \dots, x_{10} w zbiorze **b**. Analogicznie, drugi wiersz zbioru **bb** ma informować, w ilu wierszach dana zmienna miała najmniejszą wartość wśród zmiennych x_1, \dots, x_{10} w zbiorze **b**. Uwaga: może się zdarzyć, że wartości najmniejsze i największe w niektórych wierszach występują wielokrotnie.
3. (pkt) Tabela **c** zawiera rezultaty (zmienna w) pewnego testu, któremu poddawane były różne osoby (zmienna o) w różnych latach (zmienna r) (każda osoba poddawana była testowi co najwyżej raz w roku). Używając tylko PROC SQL znaleźć rok (ewentualnie lata), w których największa liczba osób osiągnęła swoje najlepsze wyniki w historii.
4. (pkt) Zbiór **d** ma zmienną tekstową x i zmienną numeryczną y i jest posortowany rosnąco po x . Zmienną x traktujemy jako zmienną grupującą. Wiadomo, że w każdej grupie jest co najwyżej pięć dodatnich wartości y . Za pomocą jednego DATA STEPu stworzyć zbiór **dd** ze zmienną x i zmiennymi numerycznymi y_1, \dots, y_5 , o liczbie obserwacji równej (nieznanej) liczbie różnych wartości zmiennej x . Jeśli w danej grupie jest i dodatnich wartości zmiennej y , to mają one się znaleźć w wartościach pierwszych i zmiennych y_1, \dots, y_5 ; jeśli $i < 5$, to zmienne o numerach $i + 1, \dots, 5$ mają zawierać braki danych.