Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

Podstawy analityki biznesowej

Ćwiczenie 1: Podstawy pracy z systemem SAS Studio

Niniejszy zestaw ćwiczeń oparto o środowisko SAS Studio. Dostęp do niego można uzyskać na trzy sposoby:

1. wykorzystanie wersji programu zainstalowanej na serwerze uczelnianym (http://sas.issi.uz.zgora.pl:7980/SASStudio/); dane logowania do przypisanego studentowi konta poda prowadzący zajęcia; dostęp do konta poza uczelnią wymaga zestawienia tunelu SSH, zob. opis pod adresem http://www.uz.zgora.pl/~agramack/files/SAS/SAS_tunel_SSH.pdf)

2. ściągnięcie i zainstalowanie *SAS University Edition* (http://www.sas.com/pl_pl/software/university-edition.html).
   Uwaga: ta wersja wymaga wcześniejszego zainstalowania środowiska do obsługi maszyn wirtualnych *Oracle VirtualBox* (https://www.virtualbox.org/);

3. dostęp poprzez przeglądarkę do zasobów *SAS OnDemand for Academics* (informacje na stronie http://support.sas.com/ondemand/, rejestracja na stronie https://odamid.oda.sas.com/SASODARegistration/, potem logowanie poprzez https://odamid.oda.sas.com/SASStudio/)

UWAGA: Poniższy tekst jest w pełni zgodny z SAS Studio w wersji 3.5. Nowsze wersje (np. w środowisku *SAS OnDemand for Academics* działa wersja 3.7) różnią się nieznacznie. W razie wątpliwości proszę pytać się prowadzącego zajęcia.

Zadanie 1. Jednym ze wskaźników stanu zdrowia jest procentowa zawartość tłuszczu w ciele człowieka. Poniższa tabela zawiera wyniki pewnej grupy ludzi po roku uczęszczania na zajęcia z aerobiku i/lub na siłowni (trzy razy tygodniowo):

Grupa	Zawartość procentowa tłuszczu
mężczyźni	13.3 8 20 12 12 19 18 31 16 24 20 22 21
kobiety	22 16 21.7 21 30 26 12 23.2 28 23

(a) Uruchomić SAS Studio. Okno programu podzielone jest na *panel nawigacyjny* (1/3 ekranu po lewej) oraz *obszar roboczy* (2/3 ekranu po prawej). Zaczniemy od utworzenia folderu na serwerze SAS, w którym będziemy zapisywać dane, a który nazwiemy **moje_dane**. W tym celu w panelu nawigacyjnym wybieramy **Foldery ►Moje foldery**, a następnie z widniejącego wyżej menu rozwijanego **Nowy** (nazwy przycisków pojawiają się, gdy najedziemy na nie myszą) wybieramy **Folder** i w oknie dialogowym wpisujemy **moje_dane**, po czym wybieramy **Zapisz**. Utworzony na serwerze SAS folder ma swoją ścieżkę dostępu, którą poznamy wybierając **Foldery ► Moje foldery ►moje_dane**, klikając prawym klawiszem myszy i wybierając **Właściwości**. Pełna ścieżka znajduje się w okienku **Położenie**. Proszę ją zapamiętać, ponieważ będzie potrzebna w kolejnym punkcie.

(b) SAS operuje nie tyle na fizycznych folderach, co na bibliotekach, których krótkie nazwy podajemy w rozmaitych poleceniach. Zwalnia to nas od podawania całych ścieżki dostępu do folderów, z którymi pracujemy. Zamiast tego wystarcza podanie krótkiej nazwy biblioteki. Bibliotekę możemy więc uważać za rodzaj synonimu (lub skrótu) do pełnej ścieżki dostępowej. Utwórzmy więc bibliotekę o nazwie **sasue**, poprzez którą będziemy odwoływać się do folderu. W tym celu w panelu nawigacyjnym wybieramy **Biblioteki** i wybieramy przycisk **Nowa biblioteka**. Ukaże się okno dialogowe, w którym należy wpisać nazwę biblioteki (tutaj **sasue**). W środkowym oknie należy wpisać pełną ścieżkę do folderu utworzonego w punkcie (a). Proszę też zaznaczyć **Przy uruchomieniu odtwórz tę bibliotekę** (jeśli tego nie zrobimy, po zakończeniu pracy z SAS Studio nazwa biblioteki zostanie zapomniana), a następnie wybrać **OK**.

(c) Dane niniejszego zadania znajdują się w pliku binarnym **bodyfat.sas7bdat**. W takim formacie SAS domyślnie zapisuje i odczytuje dane mające postać tabelaryczną. W celu ich przetwarzania przez SAS Studio, załadujmy ten plik na serwer SAS. W tym celu wybieramy **Foldery ►Moje foldery ►moje_dane**, a następnie wybieramy przycisk **Załaduj pliki**. W oknie dialogowym, które się ukaże, wybieramy przycisk **Wybierz pliki**, po czym odszukujemy na swoim komputerze pliku **bodyfat.sas7bdat** i wybieramy przycisk **Załaduj**.

(d) Teraz wypiszemy zawartość pliku. W tym celu w panelu nawigacyjnym wybieramy **Zadania ►Dane ►Listowanie danych**. Proszę zwrócić uwagę, jak zmienił się obszar roboczy. Przetestować, co powoduje wybranie w nim zakładek **Ustawienia**, **Kod/rezultaty** oraz **Podziel**. SAS Studio jest nie tyle programem bazującym na menu (takim, jak np. JMP lub Statistica), a wygodnym interfejsem do potężnego języka środowiska SAS. Efektem jest tworzony automatycznie kod w tym języku, który następnie jest wykonywany poprzez ***wybranie przycisku biegnącego ludzika*** lub przez naciśnięcie klawisza **F3**. W części **DANE** zakładki **Ustawienia** wybieramy plik **sasue.bodyfat**. Proszę sprawdzić, jak przekłada się to na kod języka SAS. Uruchomić ten kod i przeanalizować zawartość okien **REZULTATY** i **LOG**. Czy SAS rozróżnia duże i małe litery?

(e) Wybierając **Zadania ►Dane ►Atrybuty tabeli** proszę sprawdzić, jaka jest struktura danych w pliku **sasue.bodyfat**.

(f) Jak wypisać wartości jedynie zmiennej **fatpct**?

(g) Wybierając w panelu nawigacyjnym **Zadania ►Wykresy ►Wykres słupkowy** proszę w części **ROLE** pojawiającej się w zakładce **Ustawienia** w obszarze roboczym wybrać zmienną kategoryzującą **gender** i uruchomić wygenerowany kod. Przeanalizować to, co pokazuje wykres. Powtórzyć to samo dla wykresu kołowego.

(h) Jakie parametry wykresu można zmieniać w części **OPCJE** zakładki **Ustawienia**?

(i) Narysować histogram wartości zmiennej **fatpct**. Czy sensownym jest tworzenie jednego histogramu zarówno dla kobiet, jak i dla mężczyzn?

Zadanie 2. W zadaniu zajmiemy się bardziej szczegółowo tworzeniem zbiorów danych w systemie SAS. Tym razem dane dotyczą taryfikatora mandatów (w dolarach) w poszczególnych stanach USA za pierwsze przekroczenie dopuszczalnej prędkości na autostradzie o max. 39 km/h. Poszczególne stany reprezentuje się dwuliterowymi kodami, używanymi przez amerykańską pocztę.

(a) Tym razem dane zadania znajdują się w pliku **Tickets.XLS**, czyli w formacie MS Excel.

Proszę najpierw importować ten plik ze swojego komputera do biblioteki **sasue** na serwerze SAS. W panelu nawigacyjnym wybrać **Foldery ► Moje foldery ►moje_dane** i analogicznie jak w zadaniu 1(c) załadować plik **Tickets.XLS.** Następnie wybrać **Wstawki ►Dane ►Import pliku XLSX**. W obszarze roboczym zostanie wygenerowany szablon procedury importującej. Należy poprawić fragment **DATAFILE="<Your XLSX File>"** wpisując właściwą ścieżkę do załadowanego pliku. Ponieważ ładujemy plik w starszym formacie XLS, należy też linię **DBMS=XLSX** zmienić na **DBMS=XLS**. Możemy też zmodyfikować fragment **OUT=WORK.MYEXCEL** oraz **PROC PRINT DATA=WORK.MYEXCEL; RUN;** jeżeli chcemy zapisać plik w innej niż **work** bibliotece oraz pod inną nazwą. Teraz pozostaje tylko naciśnięcie przycisku z sylwetką biegnącego ludzika.
Biblioteka **work** jest domyślną biblioteką roboczą, w której SAS zapisuje wyniki naszej pracy o ile nie podamy innej nazwy biblioteki. Zapamiętajmy, że jej zawartość jest usuwana po zakończeniu sesji z SAS Studio. Jeśli więc nam zależy na trwałym zapamiętaniu pliku, zapisujmy go w innej bibliotece z naszymi danymi (np. **sasue**). Tu nie ma takiej potrzeby, więc biblioteka **work** może pozostać jako docelowa.

(b) Wypisać zawartość pliku **work.tickets**. Kolumny odpowiadają *zmiennym*, a wiersze – *obserwacjom*.

(c) Zapamiętać, że nazwy zbiorów danych muszą zaczynać się od litery lub podkreślenia (tego ostatniego lepiej unikać nie tylko na początku, ale i na końcu nazwy, bo często takie zbiory SAS rezerwuje do wewnętrznego użytku), mogą mieć długość co najwyżej 32 znaków i nie mogą zawierać odstępów.

(d) Brakujące dane są reprezentowane przez kropkę (sprawdzić, że dotyczy to jednego ze stanów).

(e) Jak nie wypisywać numeru obserwacji? (Wskazówka: w zakładce **Ustawienia** sprawdzić **Opcje**).

(f) Postortować dane ze względu na zmienna **amount** (**Zadania ►Dane ►Sortowanie danych**), zapisując wynik w pliku **tickets_sorted** w bibliotece **work** (w zakładce **Ustawienia** wykorzystać część **Rezultaty**).

Zadanie 3. W zadaniu zajmiemy się podsumowaniem zawartości zbioru danych **tickets** w oparciu o zestaw parametrów statystycznych.

(a) W celu uzyskania poszerzonego podsumowania danych zawartych w zbiorze danych **`tickets`**, wywołujemy **Zadania ►Statystyka ►Statystyki agregujące**. Jako zmienną analizowaną wybieramy **amount** i uruchamiany kod.

(b) Jak spowodować wypisanie mediany? Jak spowodować wypisanie wariancji i mody?

(c) Jak jednocześnie narysować histogram?

Zadanie 4. Poniższa tabela zawiera liczbę razy, jaką ludzie z sąsiedztwa wyprowadzają swoje psy każdego dnia:

3	1	2	0	1
2	3	1	1	2
1	2	2	2	1
3	2	4	2	1

Dane te zawiera plik **dogwalk.sas7bdat**.

(a) Importować ten plik do biblioteki **sasue**.

(b) Narysować wykres słupkowy ze zmienną **walks** jako zmienną kategoryzującą.

(c) Czy sensownym jest w tym przypadku analiza tych danych z wykorzystaniem **Zadania ►Statystyka ►Statystyki agregujące**? (Jako zmienną analizowaną proszę wybrać walks.)

(d) Powtórzyć poprzedni podpunkt, ale tym razem wykorzystać **Zadania ►Statystyka ►Jednoczynnikowe liczebności.** Czy tym razem analiza jest taka, jakiej oczekiwaliśmy?

Zadanie 5. Plik **Market_campaign.csv** jest plikiem tekstowym w formacie Comma-Separated Values (CSV). W dowolnym edytorze tekstowym proszę podglądnąć jego zawartość i wykryć regułę, według której zapisywane są dane w tym formacie.

(a) Importować ten plik na serwer SAS, a następnie przekonwertować go na plik **market_campaign.sas7bdat** zapisany w folderze **work** (robimy to analogicznie jak w czasie importowania pliky XLS).

(b) Wyznaczyć statystyki agregujące dla zmiennej **budget**. Jako zmienną klasyfikującą wybrać **Category**. Która kategoria jest najliczniejsza? W której kategorii średni budżet był największy? W której kategorii rozproszenie wartości budżetu było największe?

(c) Wykorzystując **Zadania ►Statystyka ►analiza rozkładu** przeanalizować dane histogramy zmiennej **budget** grupując dane wg zmiennej **Category**.

(d) Jeden z plików danych wygenerowany w poprzednim podpunkcie proszę przekształcić do formatu CSV i skopiować z serwera SAS na swój komputer. W tym celu wykorzystać **Wstawki ►Dane ►Generacja pliku CSV** w panelu nawigacyjnym, modyfikując i uruchoamiając kod SAS, który się wyświetli. Wyświetlić go za pomocą MS Excel lub **calc** z pakietu Libre Office)