Wczytanie PISA2012 do $\mathcal{R}$

Marcin Kosiński[1] [2]

marcin.kosinski@students.mimuw.edu.pl

kosinskim@student.mini.pw.edu.pl

26 lutego 2014

[1]Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski
[2]Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska

Poniżej mała instrukcja jak dokopać się do danych z PISA2012, aby działały w $\mathcal{R}$. Dane w formacie `.txt` pobieramy stąd. Następnie w systemie `SAS` tworzymy nowy program, którego 3 pierwsze linie można (ale nie trzeba) wpisać jak poniżej:

```
libname  MD "D:\PISA 2012";
filename STU "D:\PISA 2012\INT_STU12_DEC03.txt";
options nofmterr;
```

Kolejne linie w programie powinny byc przekopiowane z tego pliku. W tym momencie można już wywołać cały program w `SAS`, aby uzyskać pełną bazę danych `PISA2012`. Ponieważ baza zajmuje około 1,5 GB, ograniczymy się jedynie do danych dotyczących Polski, dzięki czemu program $\mathcal{R}$ będzia działał sprawniej na mniej pojemnym pliku. Posłuży do tego zapytanie `SQL`, które prezentuję poniżej:

```
proc sql;
create table POL as
select *
from Md.Stu
where CNT = 'POL'
;
```

Pomimo, że pierwsza kolumna bazy, z której wybieramy jedynie Polskę, ma widniejący podpis `Country code 3-character`, to jednak po wyświetleniu atrybutów kolumny widać, że jej nazwa to `CNT`, a `Country code 3-character` to jedynie etykieta. Dodatkowo można w ten sposób odczytać informację o długości znaków w tej kolumnie, która wynosi 3, dlatego ostatecznie w zapytaniu `SQL` widnieje linia `where CNT = 'POL'`.
Tak pomniejszoną bazę danych eksportuję do formatu `.csv` (możliwe, że bezmyślnie), dzięki procedurze `export`. Wszystkie dotychczasowe komendy i operacja odbywały się w systemie `SAS`.

```
proc export data=Pol
   outfile='D:\PISA 2012\polska.csv'
   dbms=csv
   replace;
run;
```

Ostatecznie z pliku `.csv` można już "tradycyjnie" wczytać dane do pakietu $\mathcal{R}$, używając prostego polecenia `read.csv`.

```
POL <- read.csv("D:/PISA 2012/polska.csv", sep = ",", h = TRUE)
```

Ostateczny wymiar bazdy danych, dotyczących jedynie Polski to:

```
dim(POL)

[1] 4607  634
```

A rozmiar, w bajtach:

```
file.info("D:/PISA 2012/polska.csv")$size

[1] 25376098
```

Dla porównania, cała baza danych `PISA2012` jeszcze w formacie `.txt`:

```
format(file.info("D:/PISA 2012/INT_STU12_DEC03.txt")$size, digits = 15)

[1] "1140901500"
```

Opisy poszczególnych kolumn można znaleźć w Codebook'u. Należy pamiętać, że powyższa baza danych dotyczyła jedynie kwestionariuszy wypełnianych przez uczniów.

Więcej na ten temat można znaleźć na stronie PISA2012.