Wprowadzenie do R i ćwiczenia z danymi.

Sebastian Zalas

s.zalas@uw.edu.pl

Zajęcia nr 3.

Dane o firmach - zadanie.

Jako wprowadzenie do pracy z danymi, przygotujemy bazę danych o firmach, oraz wykonamy na niej kilka operacji. W repozytorium `GitHub` znajdują się dwa zbiory danych: *firms_long_2002.csv* oraz *industry2002.csv*.

Plik *firms_long_2002.csv* zawiera zmienne:

- `year` - rok
- `id` - numer identyfikacyjny
- `EMPL` - liczba pracowników
- `STAF` - suma płac w firmie, koszt pracy, tys. $ USD
- `AV` - wartość dodana, tys. $ USD
- `TURN` - wartość sprzedaży, tys. $ USD
- `TOAS` - wartość aktywów, tys. $ USD

Plik *industry2002.csv* zawiera zmienne:

- `id` - numer identyfikacyjny
- `nacerev1primarycode` - kod sektora, w którym operowała firma
- `dateofincorporation` - data powstania firmy

Zadania:

(i) Wczytaj pliki do `R` jako dwie ramki danych. Możesz wczytać je bezpośrednio z sieci. (Pamiętaj o odpowiednim separatorze!)

(ii) Połącz obie ramki danych po zmiennej `id`. (Możesz korzystać z funkcji `merge` albo `full_merge`)

(iii) Przygotuj zmnienną `year_inc`, która ma opisywać rok powstania firmy. (Skorzystaj z np. `substring` oraz `nchar`)

(iv) Usuń obserwacje z brakującymi danymi w `nacerec1primarycode`. Utwórz zmienną, która opisuje sektor na poziomie dwóch cyfr. Korzytając z tej zmiennej, utwórz zmienną która zalicza firmy do przemysłu (kody pomiędzy 15 a 45) lub do usług (kody pomiędzy 50 a 99). Usuń obserwacje z kodami poniżej 15.

(v) Usuń obserwacje z brakującymi danymi w `STAF`, `AV`, `EMPL`, `STAF`. Czy występują ujemne wartości w tych zmiennych? Czy one mają sens?

(vi) Sprawdź rozkłady zmiennych (szczególnie `AV` oraz `EMPL`) np. używając histogramu. Czy te wykresy są zrozumiałe? Porównaj je z logarytmami zmiannych.

(vii) Przygotuj tabelę ze statystykami podsumowującymi utworzoną bazę danych. Niech zamiera ona liczę obserwacji, średnią, medianę, 25-ty i 75-ty percentyl, minimum oraz maximum.

(viii) Teraz oblicz tzw. *labor share* czyli stosunek płac do wartości dodanej ($\frac{STAF}{AV}$) na poziomie sektora (przemysł vs usługi), oraz dla całego zbioru danych. Narysuj wykres przebiegu obliczonych wskaźników w czasie.

(ix) Spróbuj utworzyć zmienną opisującą wiek firmy. Czy istnieje zależność między wiekiem firmy a np. *labor share*. Narysuj wykres rozrzutu (*scatterplot*). Zrób takie wykresy również dla innych zmiennych.

(x) Korzystając z funkcji `lm()`, oszacuj funkcję produkcji. Co oznaczają współczynniki?

Pamiętaj aby opisywać swój kod w komentarzach, oraz aby wczytać odpowiednie pakiety na początku kodu. Korzytasaj z tzw. `helpa` oraz ze źródeł internetowych. Staraj się dbać o stronę estetyczną wykonywanych wykresów.