Raport z projektu zaliczeniowego (NPD)

Paweł Brachaczek

17 czerwca 2021

Celem projektu zaliczeniowego było stworzenie programu służącego analizie danych. Naszym zadaniem było przetworzenie oficjalnych danych na temat wszystkich szkół w Polsce oraz rozkładu populacji w gminach, oraz policzenie na ich podstawie podstawowych statystyk dotyczących

- liczby uczniów na 1 nauczyciela w zależności od typu szkoły (w każdej gminie z osobna oraz łącznie dla każdego typu gminy);
- liczby uczniów na 1 szkołę w zależności od rocznika (łącznie dla każdego typu gminy).

Podczas obliczeń i szeroko pojętej pracy na strukturach danych korzystaliśmy przede wszystkim z biblioteki `pandas`. W raporcie omówimy zaś kluczowe decyzje podjęte podczas analizy danych.

1 Liczba uczniów w szkole na etat nauczycielski

Pierwszym krokiem podczas analizy było przetworzenie pliku z danymi o szkołach. Nie wszystkie kolumny arkusza są dla nas istotne. Na przykład liczbę nauczycieli, rozłożoną na trzy kolumny, można zapisać w formie jednej kolumny sumującą liczbę pełnych etatów (być może niecałkowitą).

Zakładamy, że gminę identyfikuje nie nazwa, lecz ciąg trzech liczb, oznaczających odpowiednio numer województwa, powiatu oraz gminy. Szkoły indeksujemy zaś po numerze REGON, o stosunkowo niewielkiej liczbie duplikatów (11 szkół i 197 uczniów), które usunęliśmy z danych.

Pierwszym problemem, który napotkaliśmy podczas analizy, jest zagadnienie zespołów szkół (złożoność: 2) oraz szkół filialnych (złożoność: 3). W przypadku placówek, które prowadzą np. zarówno szkołę podstawową, jak i liceum, uczniowie są przypisani do jednej z tych dwóch szkół w sposób standardowy – jednak nauczyciele zatrudnieni są nie przez liceum, lecz przez zespół szkół, który widnieje w osobnym wierszu. Aby policzyć liczbę uczniów na 1 nauczyciela w takich placówkach, konieczne jest więc „przesunięcie" nauczycieli.

Połączenia pomiędzy placówkami identyfikujemy na podstawie numerów REGON. Podczas realokacji nauczycieli zakładamy, że w danym zespole liczba nauczycieli na jeden oddział (klasę) jest stała, i każdej szkole przypisujemy nauczycieli w sposób proporcjonalny. Dopiero jeśli w danym zespole liczba oddziałów wynosi 0 (być może wskutek źle wypełnionego raportu), próbujemy dokonać takiego proporcjonalnego podziału w zależności od liczby uczniów. W przypadku zespołów, które nie mają ani oddziałów, ani uczniów, nie robimy nic (natomiast odfiltrowujemy takie placówki, w tym także delegatury o złożoności równej 4, z dalszej analizy).

W praktyce taka decyzja może wpłynąć na ostateczne wyniki. Obserwujemy na przykład, że w szkołach podstawowych uczniów na 1 etat nauczycielski jest istotnie mniej niż w liceach, pomimo mniejszej liczby godzin lekcyjnych. Oznacza to, że za taką dysproporcję musi odpowiadać m.in. większa liczebność licealnych klas. W dalszej analizie należałoby sprawdzić, który z tych dwóch czynników – liczba godzin

lekcyjnych czy liczebność klas – bardziej różni licea i szkoły podstawowe. Jeśli byłaby to liczebność klas, to założenie o stałości stosunku liczby nauczycieli do liczby klas w danym zespole szkół byłoby niestosowne.

Dalsze obliczenia w pierwszej części projektu przebiegły standardowo. Ostatecznym wynikiem są dwa arkusze z wynikami. Zawierają one statystyki: maksimum, minimum, medianę oraz średnią (liczoną z całości, tzn. ważoną po liczbie uczniów) liczby uczniów na 1 etat nauczycielski w zależności od typu szkoły. W pierwszym rozważamy każdą gminę z osobna, w drugim zaś rozważamy łącznie trzy typy: gminy miejskie, miejsko-wiejskie, oraz wiejskie. Statystyki obliczane są jedynie dla tych typów szkoły, które występują odpowiednio w danej gminie lub w danym typie gminy.

Na podstawie wstępnej analizy wyników można kontynuować rozważania na temat liczebności grup i średniej liczby godzin spędzanej w szkole przez ucznia, na przykład na podstawie szkół policealnych (które, oferując np. kursy zaoczne, mogą mieć w programie mniej godzin lekcyjnych, a więc i mniej etatów). Możemy również wysunąć hipotezę, że liczba uczniów na nauczyciela mocniej zależy od typu szkoły, niż od tego, w jakiego rodzaju gminie znajduje się ta szkoła.

2 Liczba uczniów w roczniku na szkołę

Prace w drugiej części projektu zaczynamy od przetworzenia danych o szkołach, startując z punktu, na którym poprzednio skończyliśmy. Operacje, które wykonujemy, mają posłużyć temu, by po pierwsze móc zidentyfikować, w jakim wieku są uczniowie danej szkoły, a po drugie, by łatwo i w sposób zautomatyzowany odnaleźć dane o gminie w tabeli o rozkładzie wiekowym populacji (tabeli nr 12).

Najpierw grupujemy szkoły pod względem tego, w jakim wieku są jej uczniowie. Oznacza to np. unifikację typów takich jak „przedszkole" i „punkt przedszkolny" (lata 3-6), albo „liceum ogólnokształcące" i „czteroletnie liceum plastyczne" (lata 15-18). Odrzucamy jednak szkoły, dla których nie możemy poczynić takich założeń. Są wśród nich szkoły unikalne w skali kraju, jak np. „bednarska szkoła realna", ale także szkoły przy zakładach poprawczych czy szkoły policealne. Łącznie, odrzucamy dane o 3316 szkołach, uczęszczanych przez 313783 uczniów (około 5%).

Dalej każdej gminie przyporządkowujemy siedmiocyfrowy kod, pod którym mamy nadzieję znaleźć ją w tabeli o rozkładzie wiekowym populacji. Pierwsze dwie cyfry to numer województwa, kolejne dwie – powiatu, kolejne dwie – gminy (wszystkie potencjalnie z wiodącym zerem), zaś ostatnia to typ gminy (gdzie 1 – gmina miejska, 2 – gmina wiejska, 3 – gmina miejsko-wiejska). Później nie każde takie przyporządkowanie okazało się skuteczne; prawdopodobnie z powodu ostatniej cyfry, która różniła się pomiędzy arkuszami na więcej niż jeden sposób. Odrzuciliśmy w ten sposób dane z ponad 20 gmin, gdzie do szkół chodziło 15145 osób.

Już po zgrupowaniu placówek każdej „kategorii wiekowej" wewnątrz danej gminy, ostatnia modyfikacja danych o szkołach dotyczyła uczniów, którzy, przypisani do szkoły podstawowej, w rzeczywistości należą do oddziałów przedszkolnych. Aby uprościć dalszą analizę, uczniów z takiej rubryki odjęto ze szkół podstawowych i dodano do przedszkola (w tej samej gminie), dokąd „przynależą wiekowo".

Następnie każdej „kategorii wiekowej" przyporządkowujemy listę wieków, których oczekujemy od jej uczniów. Nie bierzemy pod uwagę powtarzania klasy, ani uczniów, którzy poszli do szkoły (lub przedszkola) o rok wcześniej. Zakładając, że w każdej gminie w każdej jej szkole rozkład wiekowy uczniów odpowiada rozkładowi wiekowemu populacji, sumujemy liczbę uczniów z każdego rocznika. Rozważane roczniki to 1999-2015, jako że dane o szkołach pochodzą z 2018, jednak odczytujemy dane o populacji z 2020 – tym samym patrząc na uczniów w wieku 5-21 lat oraz ignorując ewentualne migracje pomiędzy gminami.

Na tej podstawie możemy już dojść do ostatecznego wyniku. Arkusz zawiera statystyki: maksimum, minimum, medianę oraz średnią (liczoną traktując gminy jako równoliczne) liczby uczniów z danego rocznika przypadającej na 1 szkołę w gminach każdego z trzech typów.

Wstępna analiza wyników pozwala przypuszczać, że na terenie każdej gminy znajduje się szkoła podstawowa, natomiast dla każdego typu znajdą się też takie gminy, gdzie nie ma szkoły średniej lub przed-

szkola. Jednak na wsiach takie zjawisko zdarza się częściej – co można wnioskować po tym, że średnia liczba i przedszkolaków, i uczniów szkół średnich jest tam dużo niższa niż w miastach (podczas gdy średnie dla roczników ze szkół podstawowych są bardzo podobne).

3 Podsumowanie

Podczas analizy zdecydowaliśmy się łącznie zignorować dane na temat około 5.15% wszystkich uczniów, z czego prawie wszyscy to uczniowie z nieoczywistych typów szkół pod względem wiekowym, np. szkół policealnych. Co więcej, uczniowie ci zostali uwzględnieni podczas pierwszej części obliczeń. Dzięki użyciu biblioteki `pandas`, wyniki uzyskaliśmy względnie szybko. Dodatkowo zapisanie ich w excelowym formacie pozwala innym użytkownikom łatwo sporządzić na ich podstawie np. stosowny wykres.

Log tekstowy (z konsoli)

```
197 students from 11 schools with duplicated REGON found. Dropping...
0 students from 36 delegatures found. Dropping...
203 facilities are not connected to any classes or students; teachers not reallocated.
Printing out to results\gmina_per_teacher.xlsx...
Printing out to results\gminatype_per_teacher.xlsx...

313783 students are enrolled in 3316 schools with unclear age range. Dropping...
No gmina with code 1420042 in population data. Dropping 790 students...
No gmina with code 1437032 in population data. Dropping 835 students...
No gmina with code 1214022 in population data. Dropping 484 students...
No gmina with code 1609103 in population data. Dropping 1129 students...
No gmina with code 2005022 in population data. Dropping 317 students...
No gmina with code 2817082 in population data. Dropping 801 students...
No gmina with code 3203042 in population data. Dropping 271 students...
No gmina with code 3206083 in population data. Dropping 611 students...
No gmina with code 3208033 in population data. Dropping 884 students...
No gmina with code 3215033 in population data. Dropping 791 students...
No gmina with code 1004062 in population data. Dropping 810 students...
No gmina with code 1010113 in population data. Dropping 1173 students...
No gmina with code 1018042 in population data. Dropping 844 students...
No gmina with code 2601032 in population data. Dropping 486 students...
No gmina with code 2601042 in population data. Dropping 797 students...
No gmina with code 2603042 in population data. Dropping 322 students...
No gmina with code 2604132 in population data. Dropping 1190 students...
No gmina with code 2604152 in population data. Dropping 463 students...
No gmina with code 2609032 in population data. Dropping 946 students...
No gmina with code 2612032 in population data. Dropping 415 students...
No gmina with code 2612082 in population data. Dropping 458 students...
Printing out to results\gminatype_per_age.xlsx...

5.15% of information about students dropped during all calculations.
```