Analiza modelów regresji do przewidywania liczby ludności w Polsce

Krzysztof Kulka
272667@student.pwr.edu.pl
MSiD Lab Wtorek 9.15 NP

May 15, 2024

Spis treści

1 Wstęp **3**

2 Zbiór danych i jego analiza **3**

1 Wstęp

Problemem projektu jest analiza możliwości modelów regresji liniowej do przewidywania liczby ludności w Polsce. W tym celu wykorzystane zostaną dane historyczne dotyczące demografi, oraz innych czynników wpływających na liczebność populacji. Przedstawiona analiza ma na celu rozstrzygnięcie czy model regresji liniowej jest odpowiedni do przewidywania liczby ludności w Polsce, oraz jakie modele sprawdzają się do tego najlepiej. Analizie zostaną poddane następujące czynniki:

- Historyczna liczba ludności
- Imigracja do kraju
- Wskaźnik dzietności
- Oczekiwana długość życia
- Urbanizacja
- Wskaźnik zmiany populacji na przestrzeni ostatnich 5 lat

Zbadane zaś zostaną następujące modele regresji:

- Regresja liniowa
- Regresja typu Ridge
- Regresja Decisions Trees
- Regresja Random Forest
- Regresja Lasso

2 Zbiór danych i jego analiza

Opis zbioru danych

Zbiór danych zawiera informacje na temat historycznej liczby ludności, imigracji do kraju, wskazniku dzietnosci, oczekiwanej długości życia w momencie urodzenia na przestrzeni lat 1960-2023. Dane zostały pobrane z serwisu internetowego World Bank[2] Dane dotyczą około 260 krajów. Dodatkowo informacje na temat urbanizacji zostały pobrane z serwisu internetowego Zintegrowana Platforma Edukacyjna Ministerstwa Edukacji Narodowej[3], a wskaźnik zmiany populacji na przestrzeni ostatnich 5 lat został obliczony na podstawie danych historycznych.

Analiza danych

Do analizy eksploracyjnej danych wykorzystano bibliotekę pandas-profiling[4]

Populacja w Polsce

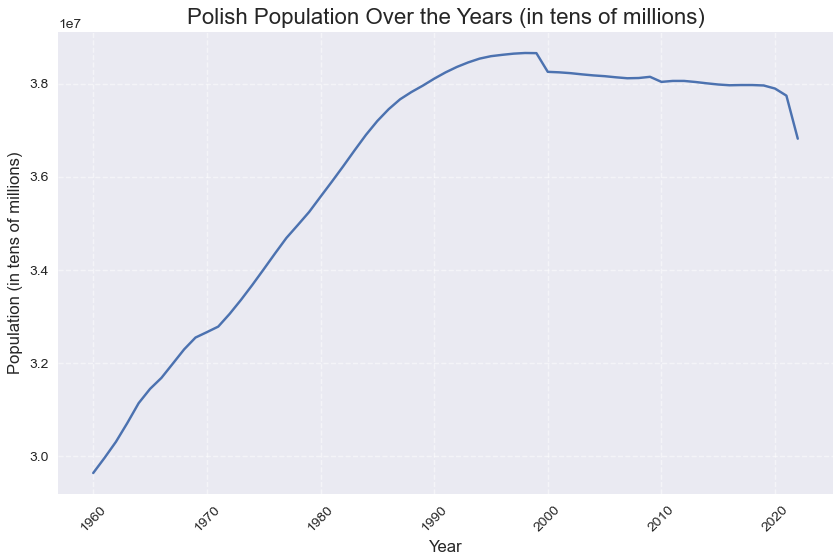

Figure 1: Wizualizacja liczby ludności w Polsce na przestrzeni lat

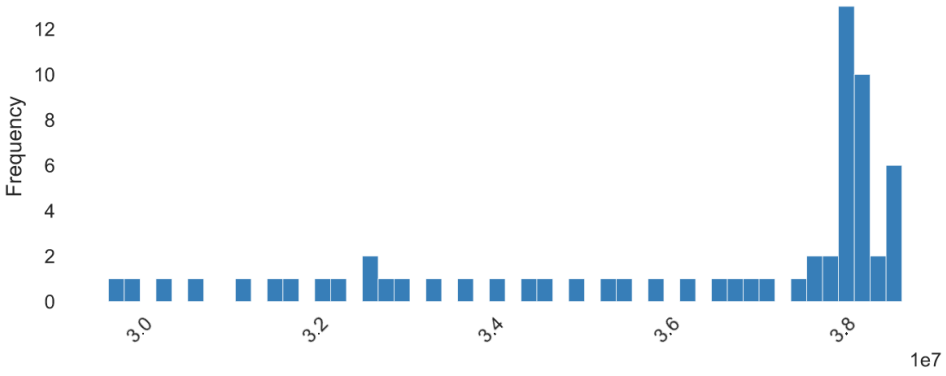

Figure 2: Histogram liczby ludności w Polsce

Minimum	Maximum	Mediana
29637450	38663481	37899070

Table 1: Statystyki liczby ludności w Polsce

Imigracja do Polski

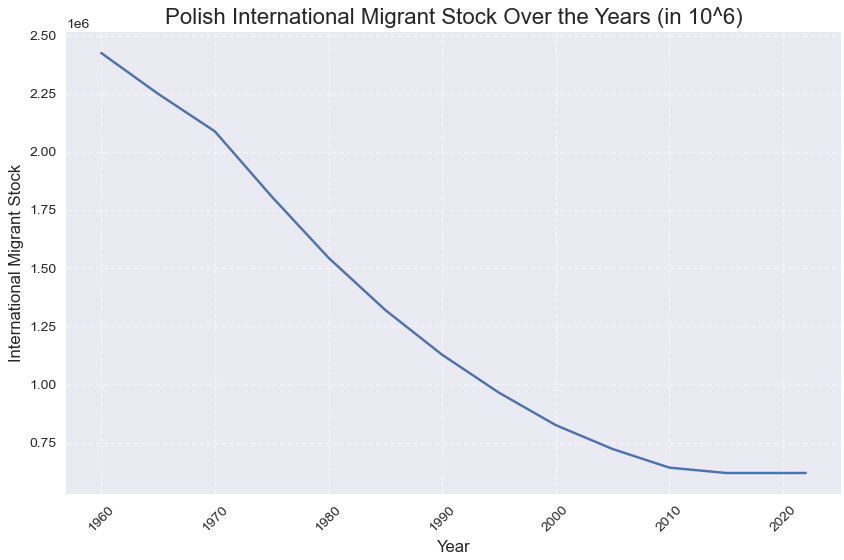

Figure 3: Wizualizacja imigracji do Polski na przestrzeni lat

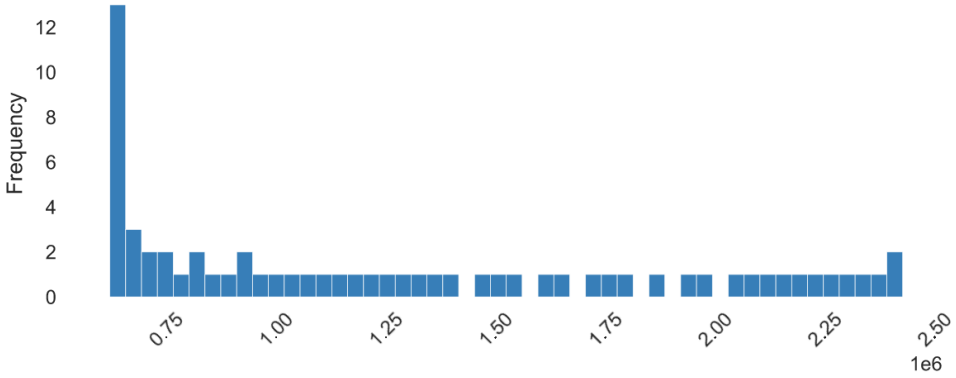

Figure 4: Histogram imigracji do Polski na przestrzeni lat

Minimum	Maximum	Mediana
0	0	0

Table 2: Statystyki imigracji do Polski

Współczynnik dzietności

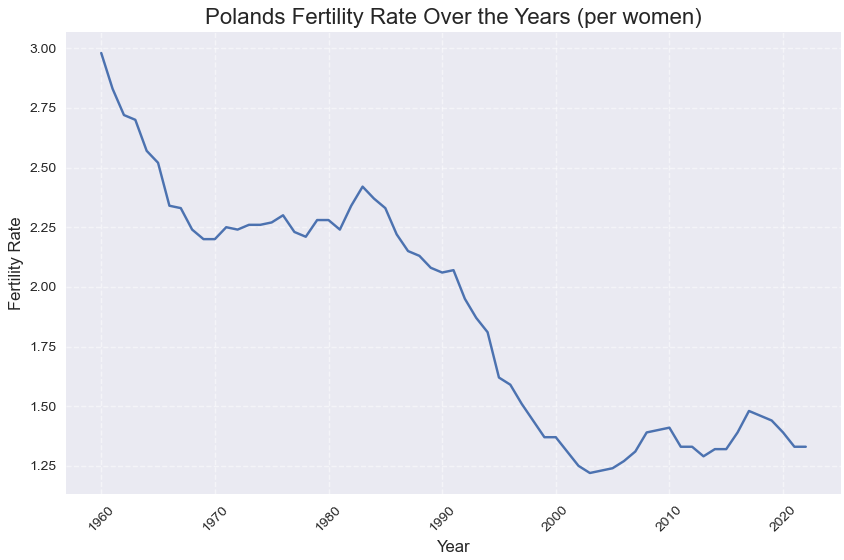

Figure 5: Wizualizacja współczynnika dzietności Polski na przestrzeni lat

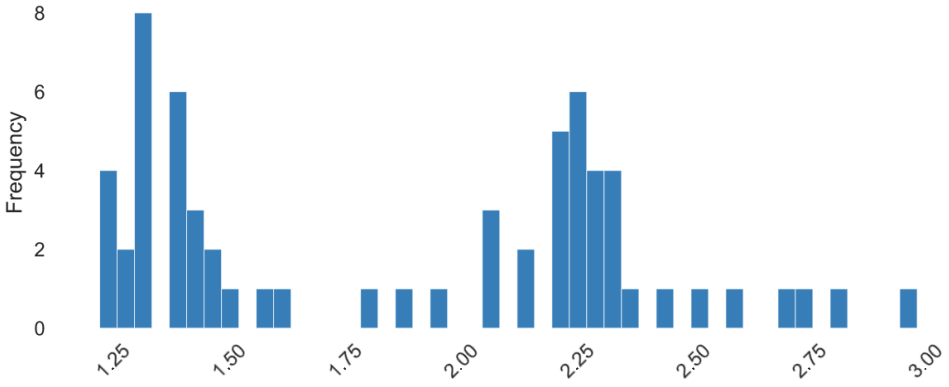

Figure 6: Histogram współczynnika dzietności na przestrzeni lat

Minimum	Maximum	Mediana
1.4	1.6	1.5

Table 3: Statystyki współczynnika dzietności

Oczekiwana długość życia

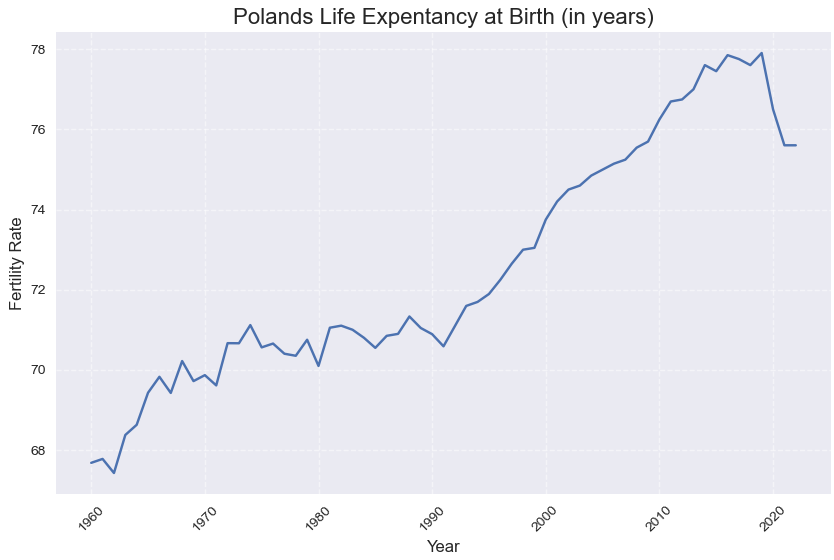

Figure 7: Wizualizacja oczekiwanej długości życia na przestrzeni lat

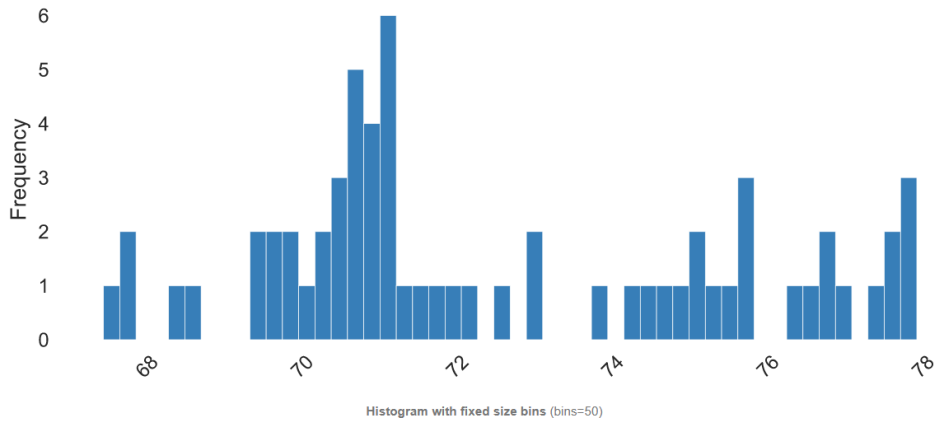

Figure 8: Histogram oczekiwanej długości życia na przestrzeni lat

Minimum	Maximum	Mediana
70.0	80.0	75.0

Table 4: Statystyki oczekiwanej długości życia

Urbanizacja

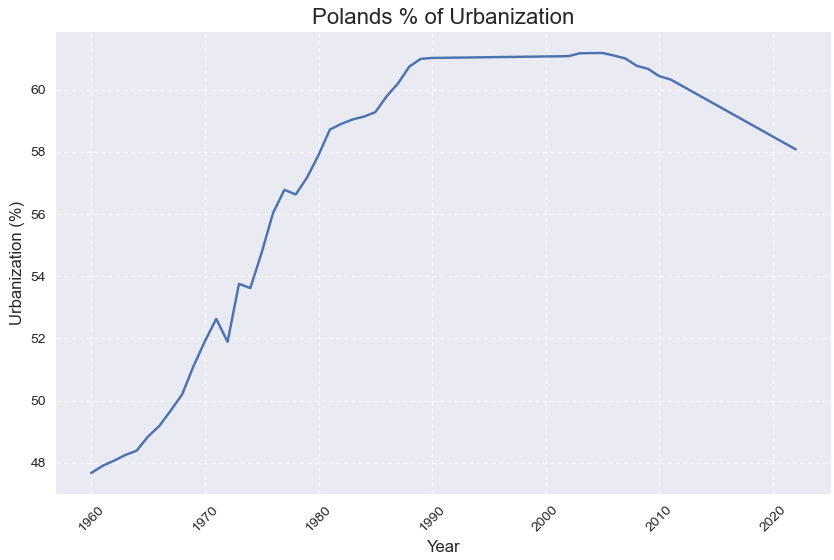

Figure 9: Wizualizacja urbanizacji na przestrzeni lat

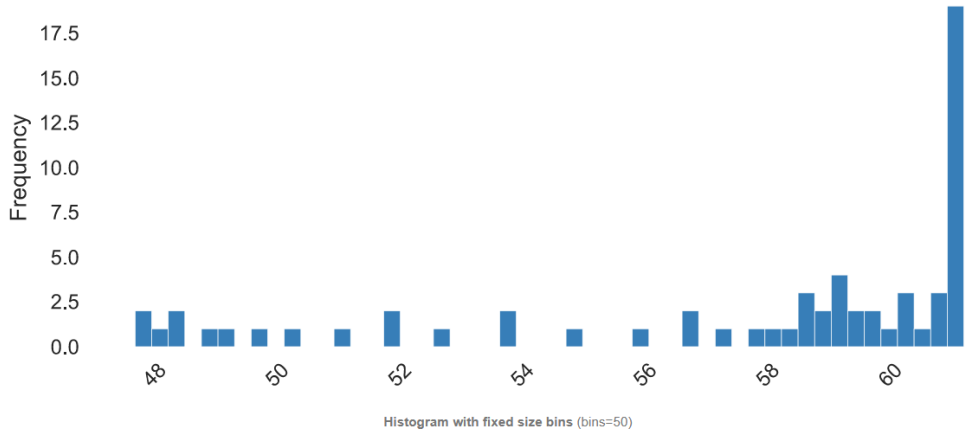

Figure 10: Wizualizacja urbanizacji na przestrzeni lat

Minimum	Maximum	Mediana
0.0	0.0	0.0

Table 5: Statystyki urbanizacji

Wskaźnik zmiany populacji

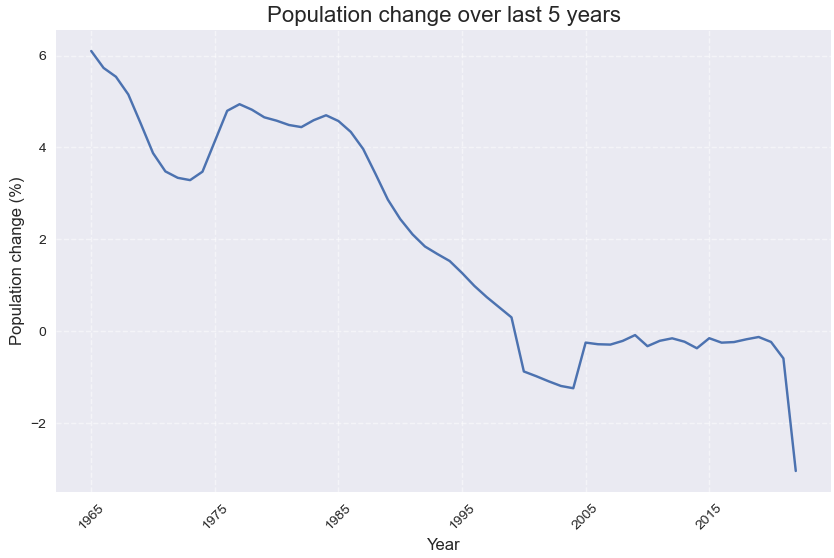

Figure 11: Wizualizacja wskaźnika zmiany populacji na przestrzeni lat

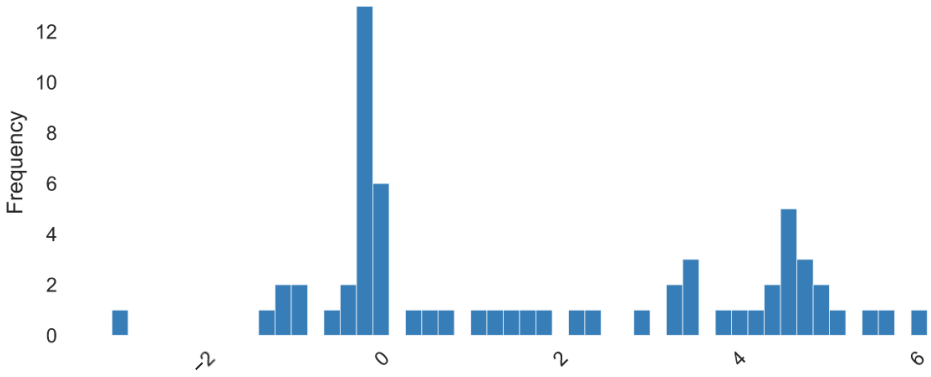

Figure 12: Histogram zmiany polskiej populacji na przestrzeni ostatnich 5 lat

Minimum	Maximum	Mediana
-0.0001	0.0001	0.0

Table 6: Statystyki wskaźnika zmiany populacji

Korelacja

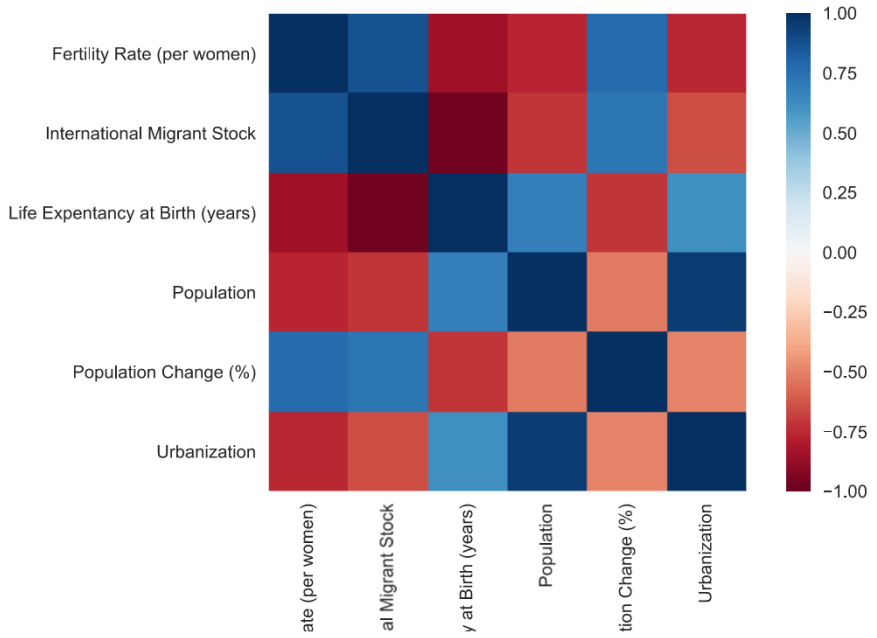

Figure 13: Macierz korelacji

Z racji ręcznego dobioru danych i ich selekcji, analiza wykazała bardzo dużą korelację (oraz antykorelacje) pomiędzy wybranymi współczynnikami. Pojawia się bardzo zaskakująca, przecząca logice korelacja, pomiędzy współczynnikiem dzietności a populacją -0.763. Prawdopodobnie wynika ona z tego że przez większość badanego okresu, współczynnik był nadal na bardzo wysokim poziomie, więc mimo że nie malał, to populacja stale się zwiększała. Analogicznie zaskakuje negatywna korelacja migracji z populacją, co jest zaskakujące, ponieważ zgodnie z intuicją, imigracja powinna zwiększać populację. Prawdopodobnie wynika to z faktu, że dane dotyczące imigracji są niewielkie, co powoduje że słabo, choć wciąż, przekładają się na polską populację. Jeśli zaś chodzi o spodziewane korelacje, należy szczególnie zwrócić uwagę:

- Oczekiwana długość życia z populacją: 0.683
- Urbanizacja z populacją: 0.949
- Urbanizacja z oczekiwaną długością życia: 0.611

Korelacje te dobrze wróżą dla modeli regresji, ponieważ są one na tyle silne, że powinny pozwolić na skuteczne przewidywanie liczby ludności w Polsce.

Obróbka danych

Pozyskanie danych

Najważniejszym krokiem w obróbce danych było wyizolowanie danych dotyczących Polski, oraz usunięcie kolumn, które nie były istotne dla analizy. Dodatkowo, z racji tego że dane dotyczące imigracji rejestrowane co pięć lat, skorzystano z interpolacji liniowej, aby uzupełnić brakujące dane.

Największe wyzwanie pojawiło się z danymi dotyczącymi urbanizacji. Jedyne z zaufanego oficjalnego źródła były w formie obrazka .jpg. Aby więc uzyskać dane, i móc zamienić je w data frame, użyto następujących kroków:

- Scraping obrazka za pomocą skryptu pythonowego
- Następnie przekazanie zescrapowanego obrazka do zewnętrznego programu WebPlotDigitizer[1]
- Finalnie, z racji niedoskonałości otrzymanego wyniku (m.in. potraktowanie lat jako liczb rzeczywistych, a nie całkowitych), dane zostały poprawione za pomocą kolejnego skryptu napisanego w pythonie.

Eliminacja danych odstających

Podczas przeglądania grafów reprezentujących zebrane dane, nie trudno było zauważyć że dane po 2019 roku są znacznie odstające od reszty. Oczywistym jest że w 2020 roku, z racji pandemii, wiele wskaźników uległo zmianie, co sprawia że dane z tego roku są nieprzydatne do analizy. Z tego powodu, dane z 2020 roku wzwyż zostały usunięte.

References

[1] WebPlotDigitizer. `https://automeris.io/`.

[2] World Bank Data poland. `https://www.worldbank.org/pl/country/poland`.

[3] Zintegrowana Platforma Edukacyjna Ministerstwa Edukacji Narodowej urbanizacja w polsce. `https://zpe.gov.pl/a/zroznicowanie-poziomu-urbanizacji-w-polsce/D19MUchJD`.

[4] Fabian Clemente. YData-Profiling. `https://github.com/ydataai/ydata-profiling`.