Krzysztof Sołowiej

# Raport z projektu

## Wstęp

Celem projektu było zbadanie korelacji łączącej liczbę interwencji gdyńskiej straży pożarnej w latach 2015-2022 (93 obserwacje) ze średnią temperaturą oraz zbudowanie modelu predykcji.

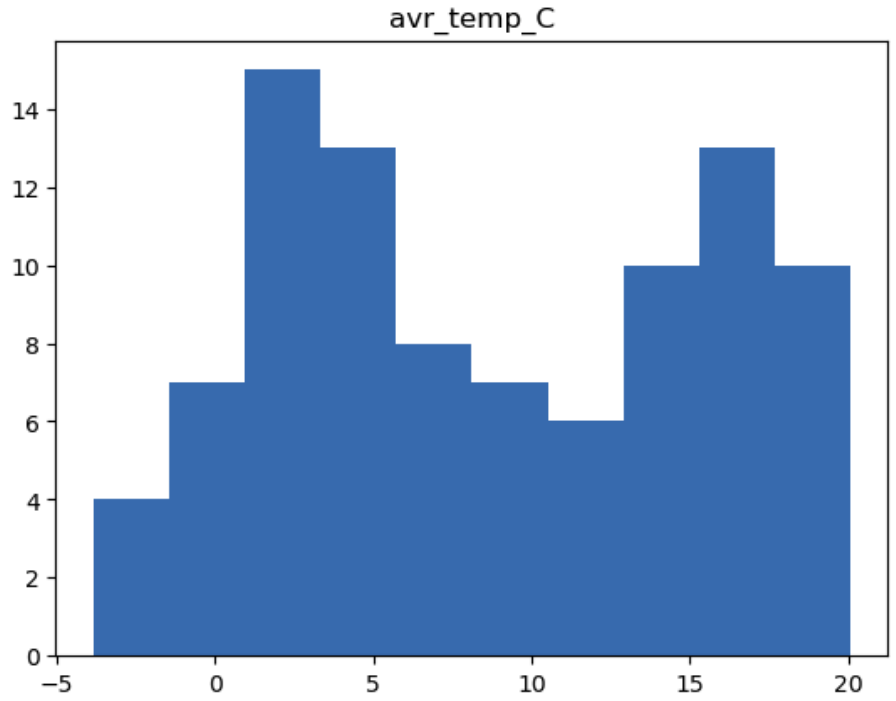
**Tabela 1.** Nieprzetworzone dane pobrano z: otwartedane.gdynia.pl i wunderground.com.

| | Data | Rok | Miesiąc | Łączna liczba interwencji | Średnia temperatura w C |
|---|---|---|---|---|---|
| **0** | 2015-01-01 | 2015 | 01 | 212 | 1.08 |
| **1** | 2015-02-01 | 2015 | 02 | 142 | 0.86 |
| **2** | 2015-03-01 | 2015 | 03 | 186 | 4.44 |
| **3** | 2015-04-01 | 2015 | 04 | 187 | 7.03 |
| **4** | 2015-05-01 | 2015 | 05 | 208 | 11.09 |
| **...** | ... | ... | ... | ... | ... |
| **88** | 2022-05-01 | 2022 | 05 | 205 | 11.33 |
| **89** | 2022-06-01 | 2022 | 06 | 230 | 16.80 |
| **90** | 2022-07-01 | 2022 | 07 | 224 | 17.68 |
| **91** | 2022-08-01 | 2022 | 08 | 269 | 20.06 |
| **92** | 2022-09-01 | 2022 | 09 | 210 | 11.61 |

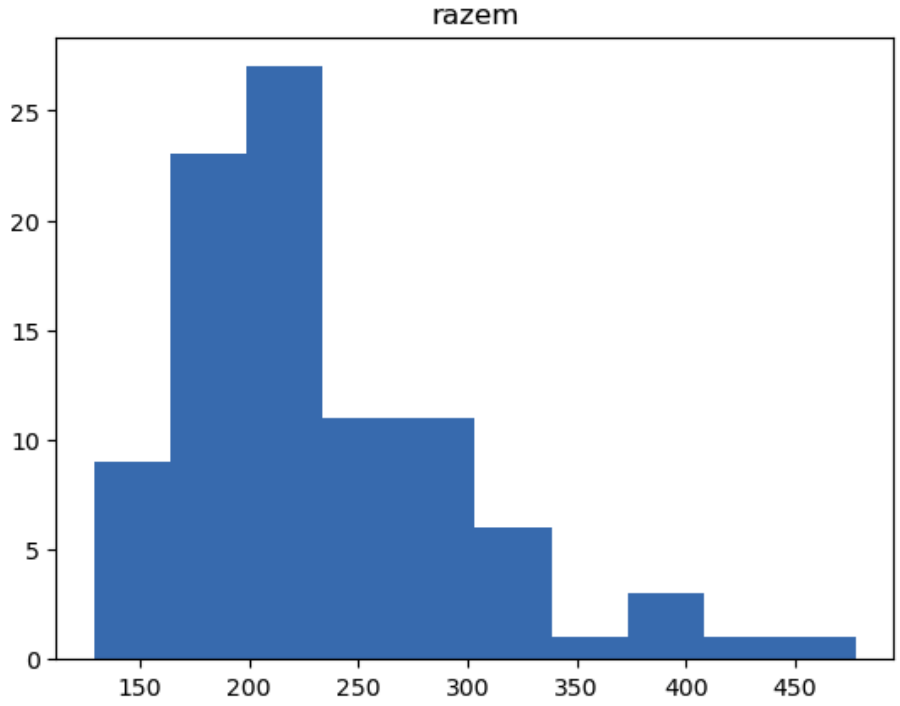
**Tabela 2.** Typy danych

| | |
|---|---|
| Data | datetime64[ns] |
| Rok | object |
| Miesiac | object |
| Łączna liczba interwencji | int64 |
| Średnia temperatura w C | float64 |

**Rys. 1.** Bimodalny rozkład danych z kolumny “Średnia temperatura w C” odzwierciedlający istnienie dwóch, wyraźnie różnych pór roku.




**Rys. 2.** Prawostronnie skośny rozkład danych z kolumny “Łączna liczba interwencji”, który sugeruje zastosowanie transformacji logarytmicznej.



# Rozdział 1 - Usunięcie najbardziej wpływowej obserwacji

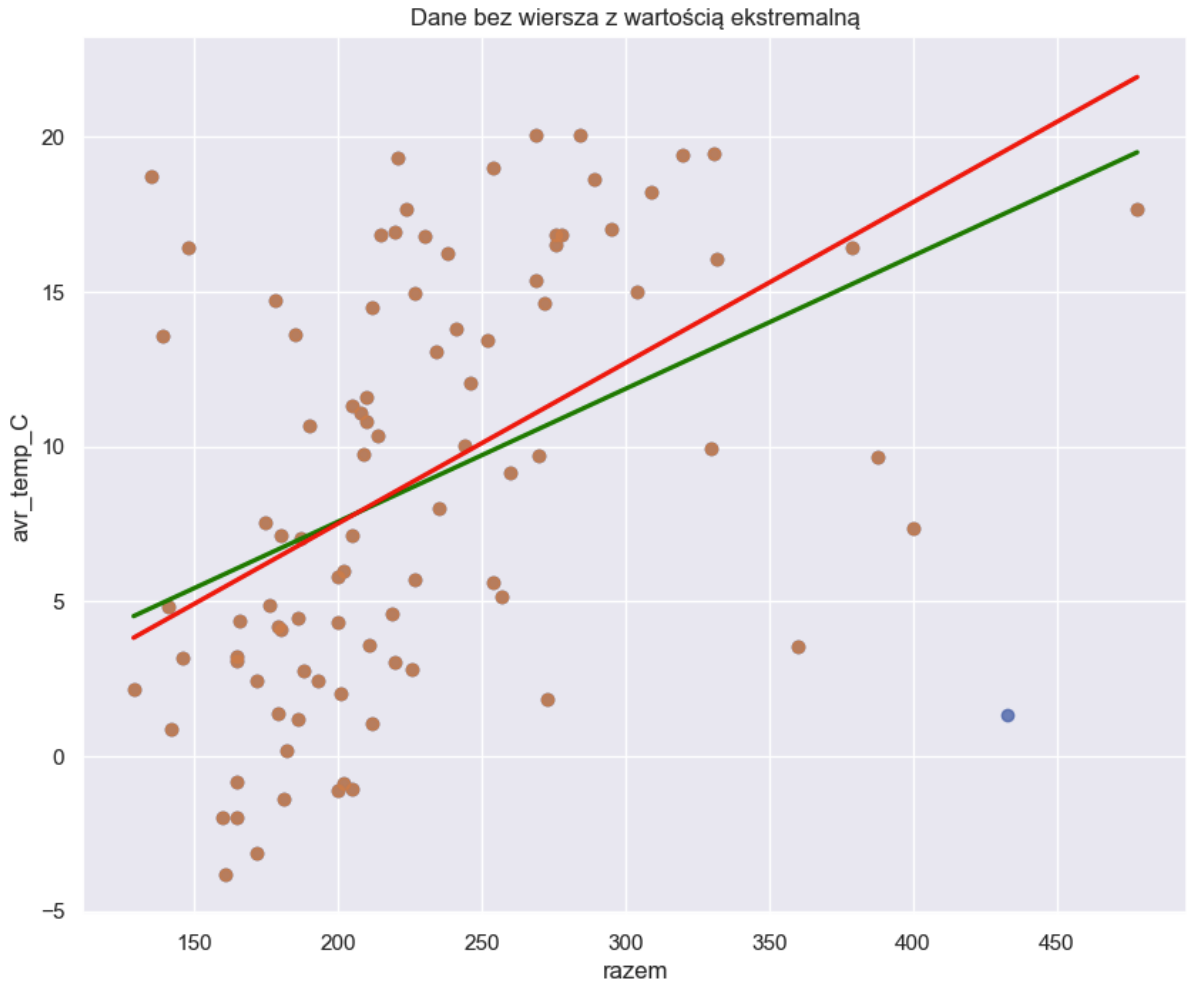
Aby zbudować dobry model, postanowiłem zidentyfikować i usunąć najbardziej wpływową obserwację. W tym celu użyłem funkcji służącej do obliczania odległości Cooka.

**Tabela 3.** Obserwacje wraz z obliczoną odległością Cooka (najwyższe wartości na początku).

| | Data | Rok | Miesiąc | Łączna liczba interwencji | Średnia temperatura w C | Odległość Cooka |
|---|---|---|---|---|---|---|
| **84** | 2022-01-01 | 2022 | 01 | 433 | 1.34 | 4.822045e-01 |
| **63** | 2020-04-01 | 2020 | 04 | 400 | 7.38 | 9.653361e-02 |
| **62** | 2020-03-01 | 2020 | 03 | 360 | 3.53 | 9.003168e-02 |
| **55** | 2019-08-01 | 2019 | 08 | 135 | 18.71 | 8.574067e-02 |
| **54** | 2019-07-01 | 2019 | 07 | 148 | 16.40 | 4.432166e-02 |
| **...** | ... | ... | ... | ... | ... | ... |
| **21** | 2016-10-01 | 2016 | 10 | 205 | 7.13 | 6.896515e-05 |
| **45** | 2018-10-01 | 2018 | 10 | 244 | 10.02 | 4.777007e-05 |
| **9** | 2015-10-01 | 2015 | 10 | 180 | 7.15 | 4.263168e-05 |
| **58** | 2019-11-01 | 2019 | 11 | 141 | 4.83 | 1.736161e-05 |
| **3** | 2015-04-01 | 2015 | 04 | 187 | 7.03 | 6.960930e-08 |

W styczniu 2022 roku zaobserwowano wyjątkowo wysoką liczbę interwencji straży pożarnej w stosunku do średniej temperatury. W związku z tym, przy budowaniu modelu predykcji, postanowiłem nie brać tej obserwacji pod uwagę. Liczba obserwacji skurczyła się do 92.

**Rys. 3.** Wizualizacja regresji liniowej przed (zielona linia) i po (czerwona linia) usunięciu najbardziej wpływowej obserwacji (zaznaczonej na niebiesko).



# Rozdział 2 - Przekształcenie zmiennej

Współczynnik korelacji obu (tzn. “Łączna liczba interwencji” i “Średnia temperatura w C”) nieprzetworzonych zmiennych wynosił zaledwie 0,42, co nie wróżyło sukcesu w budowie modelu predykcji. Po przekształceniu logarytmicznym zmiennej “Łączna liczba interwencji”, udało się poprawić współczynnik korelacji do poziomu 0,51 oraz zbliżyć dystrybucję zmiennej do rozkładu normalnego.

**Rys. 4.** Dystrybucje zmiennej “Łączna liczba interwencji” przed i po transformacji logarytmicznej.

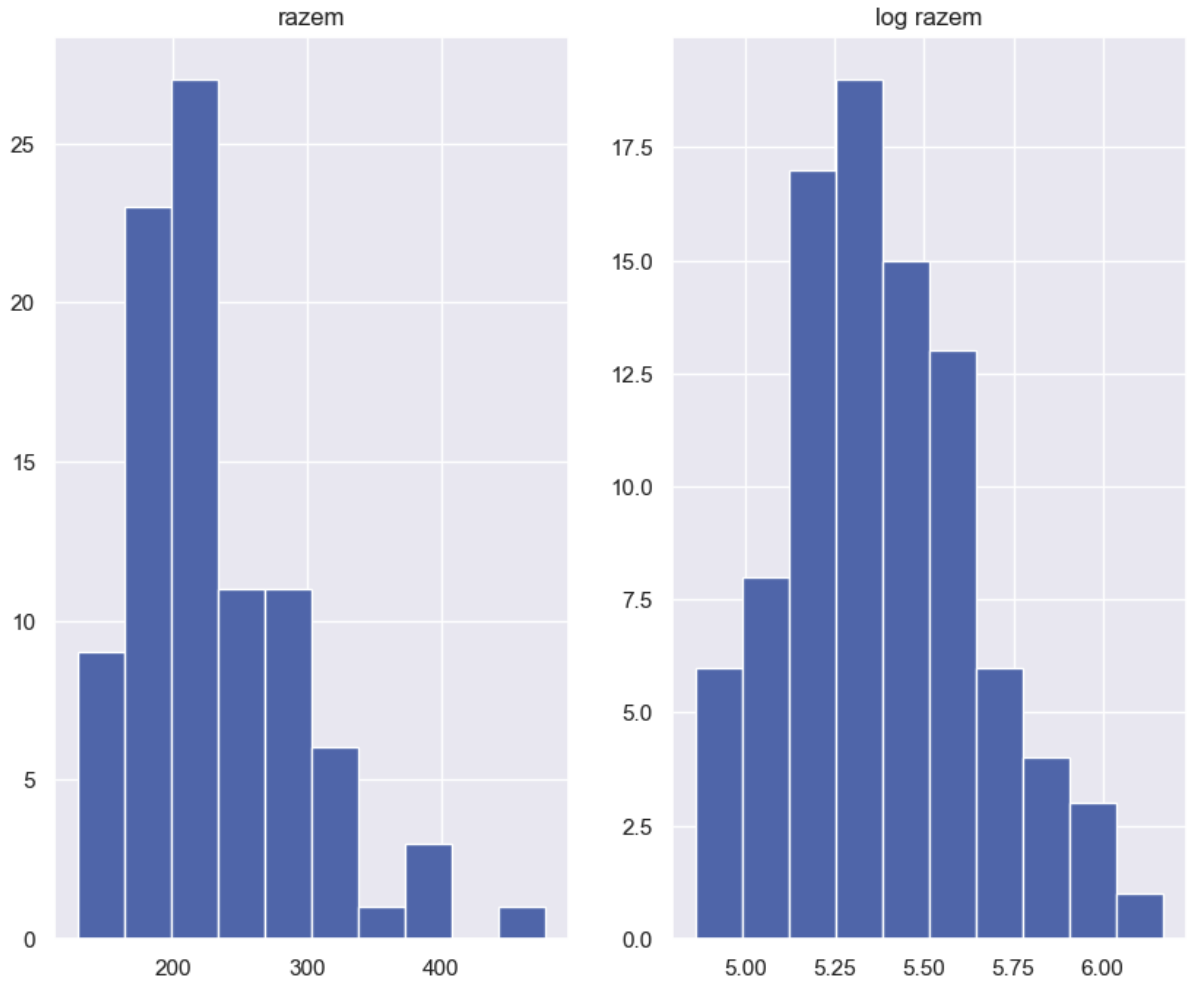


**Tabela 4.** Przekształcenie logarytmiczne zmiennej “Łączna liczba interwencji”.

| | Łączna liczba interwencji | Łączna liczba interwencji po przekształceniu logarytmicznym |
|---|---|---|
| **0** | 212 | 5.356586 |
| **1** | 142 | 4.955827 |
| **2** | 186 | 5.225747 |
| **3** | 187 | 5.231109 |
| **4** | 208 | 5.337538 |
| **...** | ... | ... |
| **88** | 205 | 5.323010 |
| **89** | 230 | 5.438079 |
| **90** | 224 | 5.411646 |
| **91** | 269 | 5.594711 |
| **92** | 210 | 5.347108 |

# Rozdział 3 - Budowa modelu predykcji

Po przekształceniu danych zająłem się budowaniem modelu regresji liniowej, przekazując algorytmowi dane widoczne w Tabeli 4 (patrz wyżej).

Z uwagi na to, że model miał posłużyć do zbudowania aplikacji, która przyjmuje dane pochodzące od użytkownika, potraktowałem cały zbiór jako zbiór treningowy.

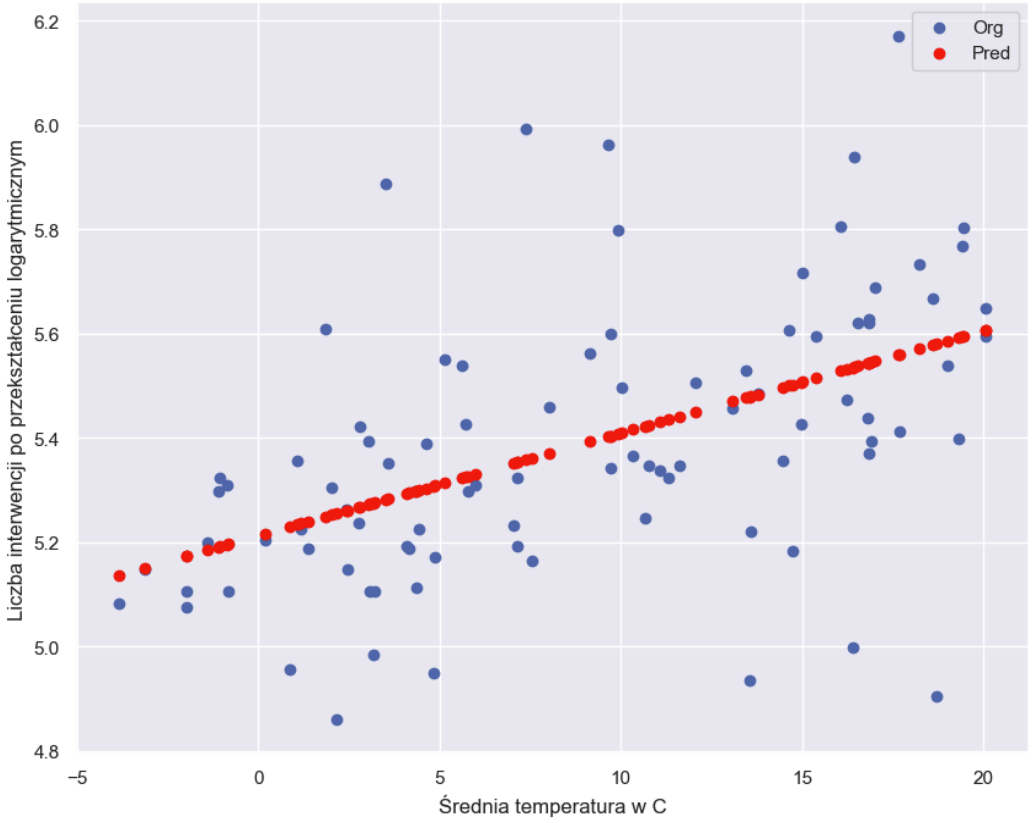
**Tabela 5.** Opis otrzymanego modelu

| Współczynnik | Wartość |
|---|---|
| Współczynnik kierunkowy (slope) | 0.01965732 |
| Wyraz wolny (intercept) | 5.212316949807735 |
| RMSE | 0.22 |

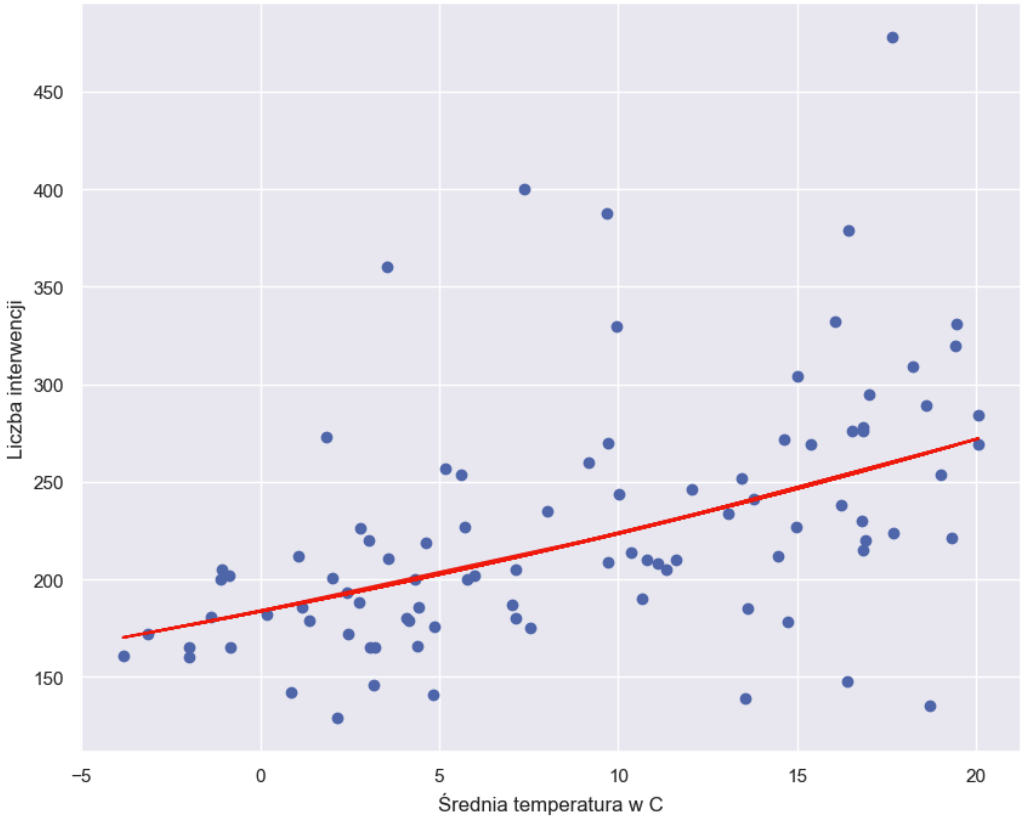
**Tabela 6.** Próba przewidzenia liczby interwencji straży pożarnej

| | Średnia temperatura w C | Łączna liczba interwencji | Łączna liczba interwencji po przekształceniu logarytmicznym | Wyniki predykcji w postaci logarytmicznej | Wyniki predykcji w postaci wykładniczej |
|---|---|---|---|---|---|
| **0** | 1.08 | 212 | 5.356586 | 5.233547 | 187.456508 |
| **1** | 0.86 | 142 | 4.955827 | 5.229222 | 186.647582 |
| **2** | 4.44 | 186 | 5.225747 | 5.299595 | 200.255784 |
| **3** | 7.03 | 187 | 5.231109 | 5.350508 | 210.715301 |
| **4** | 11.09 | 208 | 5.337538 | 5.430317 | 228.221506 |
| **...** | ... | ... | ... | ... | ... |
| **88** | 11.33 | 205 | 5.323010 | 5.435034 | 229.300743 |
| **89** | 16.80 | 230 | 5.438079 | 5.542560 | 255.330809 |
| **90** | 17.68 | 224 | 5.411646 | 5.559858 | 259.786059 |
| **91** | 20.06 | 269 | 5.594711 | 5.606643 | 272.228795 |
| **92** | 11.61 | 210 | 5.347108 | 5.440538 | 230.566306 |

**Rys. 5.** Wizualizacja wytrenowanego modelu




**Rys. 6.** Wizualizacja wytrenowanego modelu ze zmienną w postaci wykładniczej



# Wnioski końcowe

Wyżej opisany projekt jest jedynie pierwszą próbą eksploracji technik machine learningu i nie posiada żadnego praktycznego zastosowania (ze względu na niski współczynnik korelacji). Działający model można przetestować pod adresem:

http://krsolowiej.pythonanywhere.com/

Dziękuję za uwagę.