Projekt zaliczeniowy (Raport) - Podstawowy Warsztat AI

Zuzanna Kurek i Szymon Gajdziszewski

Styczeń 2026

Spis treści

1 Wstęp

Projekt obejmuje pobranie i skonsolidowanie historycznych danych pogodowych udostępnianych przez Instytut Meteorologii i Gospodarki Wodnej (IMGW) oraz wykonanie prostej analizy wybranych danych. W poniższym raporcie zostały opisane

- wykonane kroki;
- fragmenty użytego kodu;
- uzyskane w konsekwencji wyniki;
- oraz wyciągnięte wnioski.

2 Konsolidacja danych

2.1 Pobieranie

Pliki do późniejszej analizy należało pobrać ze strony o adresie `https://danepubliczne.imgw.pl/data/dane_pomiarowo_obserwacyjne/dane_meteorologiczne/dobowe/klimat/`. Do tego celu napisano krótki skrypt w Pythonie. Zaczęto od stworzenia docelowego folderu:

```
OUTPUT_DIR = "dane_imgw"
os.makedirs(OUTPUT_DIR, exist_ok=True)
```

Następnie, za pomocą pętli `for`, biblioteki `urllib.request` i `zipfile`, pobrano osobno pliki zip z każdego miesiąca każdego roku i rozpakowano je do `dane_imgw`.

2.2 Scalanie

Napisano skrypt w Pythonie, w którym zaczęto od stworzenia pustej listy o odpowiednim rozmiarze:

```
full_data = np.empty((0, 18), dtype=object)
```

Następnie pętlą `for` wpisano do niej odpowiednie dane z wcześniej uzyskanych plików:

```
for year in range(2001, 2024):
    for month in range(1, 13):
        filename = f"k_d_{month:02d}_{year}.csv"
        data = np.loadtxt(filename, delimiter=',', dtype=str,
        encoding="cp1250", quotechar='"')
        full_data = np.vstack((full_data, data))
```

Usunięto zbędnę kolumny przy użyciu `np.delete`, usunięto cudzysłowia i zapisano do nowego pliku `.csv`.

3 Proste charakterystyki i statystyki danych

3.1 Stacje z pełną historią danych

Kluczowym krokiem było odpowiednie wczytanie danych w następujący sposób:

```
station_ids = np.genfromtxt(filename, dtype=str, delimiter=None,
usecols = 0, encoding="cp1250", filling_values="")
```

`usecols = 0` pozwala na wczytanie jedynie pierwszej kolumny (a więc tej zawierającej numer stacji. Następnie zliczono ilość występowania każdego numeru i sprawdzeniu czy jest równa `8400` (ponieważ tyle wystąpień powinna mieć stacja, która posiada pełną historię). Do tego użyte zostało `np.unique` i maski:

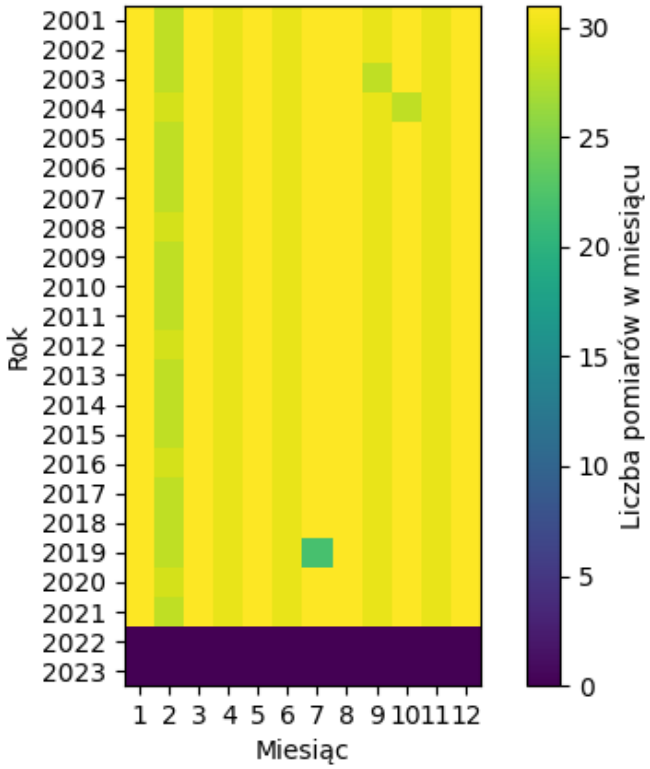
```
ID, counts = np.unique(station_ids, return_counts=True)

full_days = 8400
mask_full_history = counts == full_days
stacje_full = ID[mask_full_history]
stacje_nfull = ID[~mask_full_history]
```

3.2 Prezentacja graficzna ilości pomairów wybranej stacji

Za pomocą funkcji `np.genfromtxt` wczytano kolumny odpowiadające za numer ID stacji, rok oraz miesiąc. Stosując maskę logiczną oddzielono dane pasujące tylko do analizowanej stacji. Zagnieżdżoną pętlą dla każdego miesiąca w każdym roku zaliczono liczbę wystąpień pomariarów. Wynik zapisywany jest w odpowiedniej komórce wcześniej zrobionej macierzy 23×12.

```
for i, rok in enumerate(years):
for j, miesiac in enumerate(months):
    warunek = (dane_stacji[:, 1] == rok) & (dane_stacji[:, 2] == miesiac)
    macierz_pomiarów[i, j] = np.sum(warunek)
```



Rysunek 1: Wykres ilości pomiarów stacji 249190480

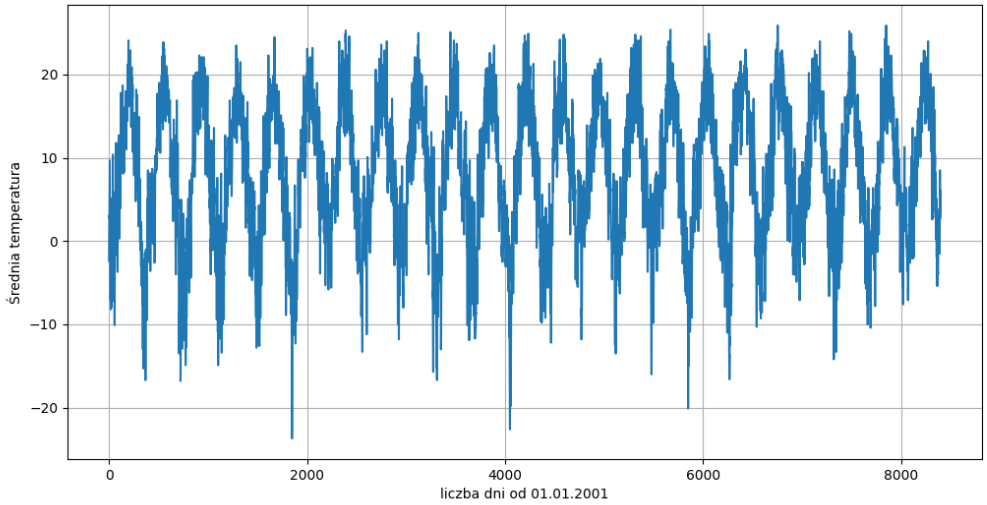
Do sporządzenia wykresu stacji o ID 249190480 użyto

```
plt.colorbar(label='Liczba pomiarów w miesiącu')
plt.xticks(np.arange(len(months)), months)
plt.yticks(np.arange(len(years)), years)
```

3.3 Wykres średniej dziennej temperatury w zależności od liczby dni, jakie upłynęły od 01.01.2001 (wersja dokładna)

Po wczytaniu danych za pomocą `np.genfromtxt` stworzono maski:

```
maska_stacji = data_subset[:, 0] ==int(wybrana_stacja)
dane_stacji = data_subset[maska_stacji]
```



Rysunek 2: Wykres średniej dziennej temperatury w zależności od liczby dni, jakie upłynęły od 01.01.2001

3.4 Minimalna i maksymalna temperatura

Wybrano kolumny odpowiadające za minimalną i maksymalną temperaturę, następnie wyszukano odpowiednio najmniejszych i największych wartości. Maski umożliwiły na znalezienie wiersza, w którym znajduje się wartość i z tych wierszy odczytano potrzebne dane (ID stacji, nazwę stacji, dzień, miesiąc, rok) z odpowiednich kolumn. Użycie masek:

```
maska_min = (t_min_values == min_temp)
maska_max = (t_max_values == max_temp)
rekordy_min = data_subset[maska_min]
rekordy_max = data_subset[maska_max]
```

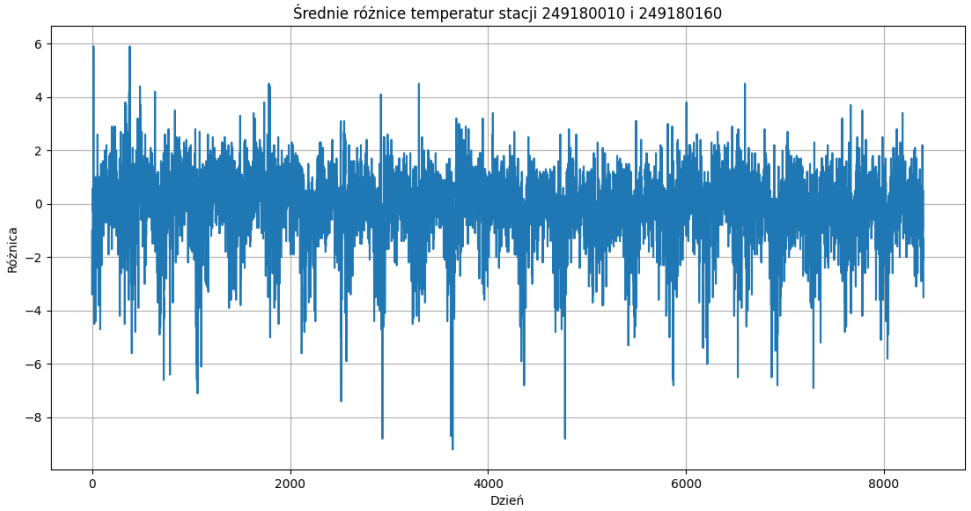
Wynik:

```
Stacja: 249190560, JABŁONKA
Data: 8/1/2017
Stacja: 251150060,CEBER
Data: 08/08/2015
```

3.5 Różnica dziennych średnich temperatur dwóch stacji

W tym podpunkcie zastosowane zostały maski, pętla `for` oraz funkcje numpy `np.argpartition()` i `np.argsort` do znalezienia indeksów odpowiadającym dniom z największą różnicą średnich temperatur.

```
top_idx = np.argpartition(s_roznica_temp, -5)[-5:]
top_idx = top_idx[np.argsort(s_roznica_temp[top_idx])[::-1]]
```



Rysunek 3: Wykres różnicy średnich temperatur dla dwóch stacji z pełnymi historiami

Skrypt zwraca też odpowiedź na pytanie "W jakich dniach różnice są największe?", podając pięć największych wartości:

- 5.90 stopni – w dniu 2001 - 1 - 10
- 5.90 stopni – w dniu 2001 - 1 - 15
- 5.10 stopni – w dniu 2001 - 1 - 11
- 5.10 stopni – w dniu 2001 - 1 - 16
- 4.50 stopni – w dniu 2001 - 2 - 16

3.6 Wykywanie anomalii

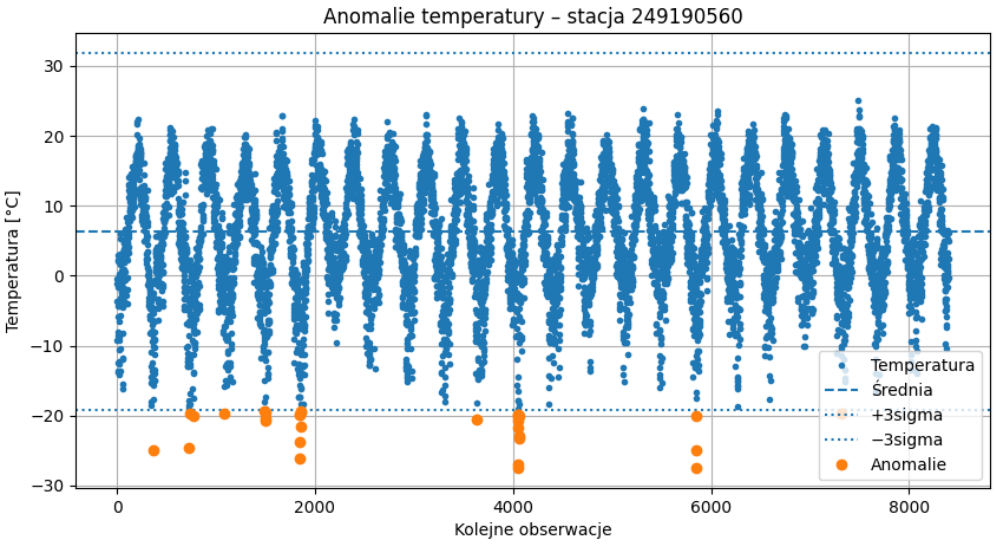
Anomalię zdefiniowano jako obserwację, dla której wartość temperatury odbiega od średniej temperatury dla danej stacji o więcej niż trzykrotność odchylenia standardowego. Dla stacji o ID 249190560 obliczono średnią temperaturę oraz odchylenie standardowe, a następnie wyznaczono anomalie.

```
srednia = np.mean(temperatury)
sigma = np.std(temperatury)
prog = 3
maska_anomalii = np.abs(temperatury - srednia) > prog * sigma
anomalie = dane_stacji[maska_anomalii]
```

Wykryto 31 anomalii, z których wypisano 5:

- 04.01.2002 – -24.9°C
- 25.12.2002 – -24.6°C
- 09.01.2003 – -19.7°C
- 13.02.2003 – -20.0°C
- 25.12.2003 – -19.7°C

oraz sporządzono wykres:



Rysunek 4: Wykres anomalii

4 Wnioski

4.1 Stacje z pełną historią danych

Znacznie więcej wykryto stacji z niepełną historią co wskazuje na występowanie wydarzeń typu awarie, sugeruje, że niektóre stacje mogły zostać zamknięte lub otwarte w okresie, który był analizowany

4.2 Prezentacja graficzna ilości pomairów wybranej stacji

Dla wybranej stacji można dostrzec, że mniej pomiarów wykonywanych było w miesiące o mniejszej ilości dni (lipiec 2019). Widać też pojedyncze miesiące o mniejszej ilości wykonanych pomiarów (luty), w których najpewniej doszło do awarii lub wykonywano prace konserwacyjne.

4.3 Wykres średniej dziennej temperatury w zależności od liczby dni, jakie upłynęły od 01.01.2001 (wersja dokładna)

Zgodnie z podejrzeniami średnie temperatury były co roku najniższe w miesiącach zimowych, a najwyższe w letnich. Zaskakujące jest występowanie aż trzech dni w przeciągu analizowanych lat, w których średnia temperatura wynosiła -20 stopni.

4.4 Minimalna i maksymalna temperatura

Z uzyskanych danych można zgodnie z przewidywaniami wywnioskować, że najniższa zarejestrowana temperatura panowała w Jabłonce (miejscowość na południu Małopolski) w styczniu, a najwyższa w Ceber (Dolny Śląsk - potencjalnie najcieplejszy obszar w Polsce) w sierpniu.

4.5 Różnica dziennych średnich temperatur dwóch stacji

Z wykresu można wywnioskować, że różnica średnich zmierzonych temperatur pomiędzy wybranymi stacjami zawiera się w przedziale od -9 do 6 stopni. Są to istotne różnice, co wstazuje na to prawdopodobny duży dystans pomiędzy dwiema stacjami. Wielkość różnic zmieniała się rocznie z pewną regularnością.

4.6 Wykywanie anomalii

Większość pomiarów temperatury mieści się w typowym zakresie wahań wokół średniej, natomiast wykryte anomalie odpowiadają pojedynczym, silnym odchyleniom. Anomalie występują głównie w okresach zimowych i mają charakter sporadyczny.