Eksploracyjna analiza danych z użyciem Pandas

Moduł 3 kursu Python LLM i analiza danych w notatnikach Jupyter

Materiał do samodzielnej nauki offline

Autor modułu mgr inż. Marcin Godziemba-Maliszewski

ENHANCER+ Politechnika Warszawska

Jak korzystać z podręcznika

W module 3 nauczysz się odpowiadać na pytania za pomocą danych przygotowanych w module 2. Poznasz rozkłady pomiarów, porównasz urządzenia i okresy oraz zbadasz zależności między zmiennymi. Każdy wynik połączysz z informacją o tym, jak powstał i czego nie pozwala stwierdzić. To podstawa dalszej pracy nad wykresami i narracją analityczną w module 4.

Czytaj rozdziały kolejno. Wyjaśnienia, tabele wyników i pytania „Sprawdź rozumienie” pozwalają uczyć się bez uruchamiania komputera. Odpowiedzi do tych pytań znajdziesz na końcu. Ćwiczenia w notebooku służą samodzielnemu zastosowaniu poznanej metody. Zanim porównasz wynik z omówieniem, spróbuj przewidzieć jego kształt, liczebność i znaczenie.

Do wykonania kodu offline przygotuj wcześniej środowisko Jupyter z Pandas, NumPy i Matplotlib oraz lokalny wynik M2. Wszystkie główne bloki kodu wykonuj kolejno w notebooks/M3_EDA.ipynb. Ćwiczenia oznaczono osobno i nie są warunkiem wykonania dalszych przykładów. Pomoc LLM jest opcjonalna; lektura i obliczenia nie wymagają korzystania z usługi internetowej.

Spis treści

Rozdział 1 Od problemu do pytań analitycznych

Rozdział 2 Rozkład i statystyki opisowe

Rozdział 3 Filtrowanie sortowanie i analiza liczebności

Rozdział 4 Grupowanie i agregacje

Rozdział 5 Tabele przestawne i porównania

Rozdział 6 Zależności korelacja i ostrożna interpretacja

Zapis wyników i przygotowanie do modułu 4

Odpowiedzi do sprawdzenia rozumienia

Słownik pojęć i samodzielna kontynuacja

Materiały modułu

Autor modułu: mgr inż. Marcin Godziemba-Maliszewski. Materiał towarzyszy prezentacji M3-v2.pptx „Eksploracyjna analiza danych z użyciem Pandas” i karcie M3_Instrukcja_EDA.docx. Przypadek przewodni „Eksploatacja urządzeń” wykorzystuje pomiary i kartotekę z pakietu startowego, przygotowane zgodnie z modułem 2.

Rozdział 1 Od problemu do pytań analitycznych

Eksploracyjna analiza danych, w skrócie EDA, pomaga poznać strukturę zbioru i wskazać zależności warte dalszego sprawdzenia. Punktem wyjścia jest problem, ale obliczenia wymagają precyzyjniejszego pytania. Zdanie „urządzenia w Hali A działają gorzej” nie określa jeszcze, co mierzyć. Może chodzić o drgania, energię albo produkcję, a każda z tych wielkości prowadzi do innej analizy.

Pytanie „Jak różnią się mediany drgań między lokalizacjami w okresie objętym zbiorem?” określa zmienną, miarę, grupy i zakres. Odpowiedzią będzie jedna mediana oraz liczebność dla każdej lokalizacji. Nadal nie ustalisz w ten sposób przyczyny różnic. Lokalizacje obejmują inne urządzenia, a dane nie zawierają wszystkich warunków ich pracy.

Pytanie opisowe porównawcze i dotyczące zależności

Pytanie opisowe dotyczy tego, co występuje w zbiorze: typowych wartości, rozproszenia albo częstości. Pytanie porównawcze zestawia tę samą miarę w kilku grupach. Pytanie o zależność bada, czy dwie zmienne współzmieniają się w obserwacjach. Dopiero pytanie przyczynowe dotyczyłoby mechanizmu zmiany; prosta korelacja nie daje na nie odpowiedzi.

Przykład pytania	Potrzebne pola	Oczekiwany wynik
Jakie są typowe drgania	drgania	Średnia, mediana, rozproszenie i n
Jak różnią się urządzenia	urzadzenie, drgania	Ta sama miara dla każdego urządzenia
Czy temperatura współzmienia się z drganiami	temperatura, drgania	Wykres, korelacja i liczba kompletnych par

Jednostka obserwacji i poziom odpowiedzi

Jeden wiersz wejścia oznacza pomiar jednego urządzenia w określonym czasie. Zbiór ma 720 takich rekordów, ale tylko osiem urządzeń. Po agregacji jeden wiersz może opisywać urządzenie, lokalizację albo urządzenie w miesiącu. Poziom agregacji określa, co reprezentuje wynik. Ranking najwyższych pojedynczych pomiarów nie odpowiada na pytanie o najwyższą medianę urządzenia.

Znaczenie miary także zależy od poziomu. Mediana energii opisuje typową energię na interwał pomiarowy, a suma dotyczy łącznej energii w wybranych interwałach. Sumy wymagają porównywalnego zakresu czasu i kontroli braków. W tym module pracujemy na danych dydaktycznych; wartości progowe z ćwiczeń nie są normami technicznymi.

Wspólny rejestr sześciu pytań

W całym module zachowaj poniższe identyfikatory. Rozdział 1 przygotowuje plan wszystkich odpowiedzi. Rozdział 4 obejmuje dwa pytania, dlatego numer pytania nie zawsze jest numerem rozdziału. Własne pytania dopisz jako rozszerzenia, bez zmiany znaczenia M3-P01–M3-P06.

ID	Rozdział	Pytanie
M3-P01	2	Jakie są typowe wartości i zróżnicowanie temperatury, drgań, energii i produkcji
M3-P02	3	Które pomiary przekraczają 1,5 mm/s i jak rozkładają się między urządzenia oraz miesiące
M3-P03	4	Które trzy urządzenia mają najwyższą medianę drgań i z ilu pomiarów obliczono każdą medianę
M3-P04	4	Jak różnią się mediany energii i produkcji między trybami oraz typami urządzeń
M3-P05	5	Jak różni się mediana energii w układzie rok_miesiac × lokalizacja
M3-P06	6	Czy wyższej temperaturze towarzyszą większe wartości drgań

Kontrakt odpowiedzi i hipoteza robocza

Przed napisaniem kodu zapisz zakres danych i czasu, potrzebne pola, jednostkę obserwacji, poziom grupowania, miarę z jednostką oraz formę wyniku. Dodaj kontrolę, po której poznasz, że odpowiedź jest kompletna. Dla M3-P03 będą to wszystkie osiem urządzeń, liczebność drgań w każdej grupie, pełny ranking median i wskazanie pierwszych trzech pozycji.

Hipoteza robocza jest przewidywaniem zapisanym przed obliczeniem, na przykład „wyższym temperaturom towarzyszą większe drgania”. Ustal także, jaki wynik mógłby jej przeczyć. Po analizie nadaj status: wsparta, niewsparta albo nierozstrzygnięta. Wsparcie hipotezy o współzmienności nie zamienia jej w dowód mechanizmu przyczynowego.

Pola mają różne role. Czas i urzadzenie identyfikują pomiar; rok_miesiac porządkuje okresy; typ_urzadzenia, lokalizacja i tryb określają grupy. Moc nominalna i wiek urządzenia dostarczają kontekstu. Status ma w całym zbiorze wartość OK. Nie ma tu rejestru zdarzeń ani liczby awarii, więc nie sprawdzisz pytania „czy starsze urządzenia częściej się psują”.

Niezależny start notebooka M3

Zachowaj projekt z poprzednich modułów. Notebook M2 dokumentuje przygotowanie danych, a nowy M3 zaczyna od data/processed/pomiary_clean.csv. Nie musi korzystać ze zmiennych zapamiętanych przez M2. Nie powtarzaj czyszczenia wewnątrz EDA i nie nadpisuj zatwierdzonego CSV. Słownik i dziennik decyzji pozwolą wyjaśnić ewentualną niezgodność.

```
from pathlib import Path
from hashlib import sha256
import pandas as pd
import numpy as np

print("Katalog roboczy:", Path.cwd())
assert Path.cwd().name == "notebooks"
PROJECT_DIR = Path.cwd().parent
TABLES_DIR = PROJECT_DIR / "outputs" / "tables"
FIGURES_DIR = PROJECT_DIR / "outputs" / "figures"
plik = PROJECT_DIR / "data" / "processed" / "pomiary_clean.csv"
assert plik.is_file(), "Najpierw przygotuj i sprawdź wynik M2"
hash_wejscia = sha256(plik.read_bytes()).hexdigest()
pomiary_clean = pd.read_csv(
    plik, encoding="utf-8", parse_dates=["czas", "data_instalacji"]
)
KOLUMNY = [
    "czas", "rok_miesiac", "urzadzenie", "typ_urzadzenia",
    "lokalizacja", "moc_nominalna_kw", "data_instalacji",
    "wiek_urzadzenia_dni", "temperatura", "drgania", "energia",
    "produkcja", "tryb", "status"
]
assert pomiary_clean.columns.tolist() == KOLUMNY
SCHEMAT_WYNIK = dict.fromkeys([
    "rok_miesiac", "urzadzenie", "typ_urzadzenia", "lokalizacja"
], "string")
SCHEMAT_WYNIK.update(dict.fromkeys([
    "temperatura", "drgania", "energia", "moc_nominalna_kw"
], "float64"))
SCHEMAT_WYNIK.update({
    "produkcja": "Int64", "wiek_urzadzenia_dni": "Int64",
    "tryb": "category", "status": "category"
})
pomiary_clean = pomiary_clean.astype(SCHEMAT_WYNIK)
```

Samo umieszczenie notebooka w folderze notebooks nie gwarantuje, że jest on katalogiem roboczym. Sprawdź Path.cwd(), zamiast wpisywać ścieżkę z dyskiem i nazwą konta. Gdy kontrola zatrzyma wykonanie, popraw punkt startu. Nie przenoś plików przypadkowo, aż program przestanie zgłaszać błąd.

CSV zachowuje wartości tekstowe, ale nie przechowuje schematu Pandas. Dlatego jawnie odtwarzasz daty i typy. Int64 z wielką literą obsługuje brak produkcji. Typ category sam nie potwierdza dozwolonych etykiet; to osobna kontrola na następnej stronie.

Kontrola danych przed pierwszym pytaniem

```
KLUCZ = ["czas", "urzadzenie"]
assert pomiary_clean.shape == (720, 14)
assert pomiary_clean[KLUCZ].notna().all().all()
assert not pomiary_clean.duplicated(KLUCZ).any()
assert pomiary_clean["urzadzenie"].nunique() == 8
assert pomiary_clean["urzadzenie"].value_counts().eq(90).all()
for pole in ["czas", "data_instalacji"]:
    assert pd.api.types.is_datetime64_any_dtype(pomiary_clean[pole])
    assert pomiary_clean[pole].notna().all()
assert pomiary_clean["czas"].min().date().isoformat() == "2026-01-01"
assert pomiary_clean["czas"].max().date().isoformat() == "2026-02-14"
assert pomiary_clean["rok_miesiac"].eq(
    pomiary_clean["czas"].dt.strftime("%Y-%m")
).all()
assert pomiary_clean["energia"].isna().sum() == 2
assert pomiary_clean["produkcja"].isna().sum() == 1
assert pomiary_clean[["temperatura", "drgania"]].notna().all().all()
assert set(pomiary_clean["tryb"]) == {"auto", "oszczedny"}
assert pomiary_clean["status"].eq("OK").all()
print("SHA-256:", hash_wejscia)
```

Te liczby odnoszą się do kursowego wyniku M2. Nie są uniwersalną regułą dla każdego zbioru. Dwa braki energii pozostały bez uzupełnienia, a jeden brak produkcji zastępuje błędną wartość −12. Nietypowe drgania 4,850 mm/s zachowano do dalszego sprawdzenia. Zmienna clean oznacza udokumentowany stan danych, a nie brak wszystkich ograniczeń.

Porównaj SHA-256 z wpisem M2 w README.md. Zgodność potwierdza tożsamość pliku, lecz sama nie dowodzi jego jakości. Przy różnicy wróć do procesu przygotowania danych. Pamiętaj również, że luty obejmuje tylko dni 1–14, podczas gdy styczeń jest kompletny. To ograniczenie będzie wracać w porównaniach okresów.

Ćwiczenie w notebooku

Uzupełnij rejestr sześciu pytań o kontrakty odpowiedzi. Do jednego dopisz hipotezę oraz możliwy wynik przeciwny. Jeśli używasz LLM, podaj rzeczywiste nazwy pól i ograniczenia. Odrzuć pytania wymagające liczby awarii, a pytania o przyczyny przepisz na porównania możliwe do sprawdzenia.

Sprawdź rozumienie 1

Czy wybranie trzech najwyższych drgań z 720 rekordów daje odpowiedź na M3-P03? Co powinien reprezentować jeden wiersz właściwego rankingu?

Rozdział 2 Rozkład i statystyki opisowe

Pytanie M3-P01 dotyczy typowych wartości i zróżnicowania czterech miar. Jedna średnia nie opisuje wszystkich własności rozkładu. Potrzebujesz informacji o środku, rozproszeniu, granicach oraz liczbie dostępnych obserwacji. Każda z tych perspektyw pomaga wykryć inną pułapkę interpretacji.

Średnia arytmetyczna wykorzystuje każdą dostępną wartość. Pojedynczy skrajny pomiar może więc ją przesunąć. Mediana wskazuje środek uporządkowanych danych i słabiej reaguje na pojedyncze maksimum. Dla parzystej liczby obserwacji jest średnią dwóch środkowych wartości, dlatego nie musi odpowiadać jednemu rzeczywistemu pomiarowi.

Podobieństwo średniej i mediany nie dowodzi symetrii ani braku wartości nietypowych. Dla drgań obie miary są bliskie 1,24 mm/s, mimo obecności maksimum 4,850 mm/s. W energii średnia około 13,929 kWh jest niższa od mediany 14,235 kWh. Tę różnicę warto zbadać, lecz sama nie wyjaśnia kształtu ani przyczyny rozkładu.

Kwartyle percentyle i rozproszenie

Pierwszy kwartyl Q1 odpowiada pozycji około 25%, a trzeci Q3 około 75% uporządkowanych danych. Pas pomiędzy nimi opisuje środkową połowę obserwacji. Rozstęp międzykwartylowy IQR jest różnicą Q3 i Q1. W odróżnieniu od rozstępu, czyli różnicy maksimum i minimum, nie opiera się wyłącznie na dwóch skrajnych wartościach.

Percentyl P90 wskazuje próg, którego nie przekracza około 90% wartości w analizowanym zbiorze. Dla drgań wynosi około 1,438 mm/s. Nie oznacza to 90% prawdopodobieństwa dla następnego pomiaru. Kwantyle mogą być interpolowane między obserwacjami; zaokrąglenie i powtarzające się wartości także wpływają na dosłowne liczenie obserwacji przy progu.

Odchylenie standardowe opisuje rozproszenie wokół średniej i reaguje na wartości skrajne. Pandas w std() i describe() domyślnie używa ddof=1, czyli dzielnika n−1 przy obliczaniu wariancji. Odchylenie, IQR i rozstęp zachowują jednostkę zmiennej. Nie porównuj ich wielkości bezpośrednio między temperaturą, energią i produkcją.

Liczebność jako część wyniku

Wierszy jest 720, ale count energii wynosi 718, a produkcji 719. Podstawy statystyk nie są zatem identyczne. Usunięcie całych wierszy z brakami energii pozbawiłoby inne analizy poprawnych temperatur i drgań. Każdą miarę licz na właściwym zbiorze dostępnych wartości i podawaj jej liczebność.

Tabela odpowiadająca na M3 P01

```
MIARY = ["temperatura", "drgania", "energia", "produkcja"]
jednostki = {"temperatura": "°C", "drgania": "mm/s",
             "energia": "kWh", "produkcja": "szt. na interwał"}
statystyki = pomiary_clean[MIARY].describe(
    percentiles=[.10, .25, .50, .75, .90]
).T
statystyki["IQR"] = statystyki["75%"] - statystyki["25%"]
statystyki["braki"] = pomiary_clean[MIARY].isna().sum()
statystyki["braki_proc"] = statystyki["braki"] / len(pomiary_clean) * 100
statystyki["jednostka"] = pd.Series(jednostki)
statystyki.index.name = "zmienna"
display(statystyki.round(3))
```

Describe() zwraca liczbę niebrakujących wartości, średnią, odchylenie, minimum, maksimum i wskazane percentyle. Transpozycja .T sprawia, że jedna zmienna zajmuje jeden wiersz. Nazwy 25%, 50% i 75% oznaczają Q1, medianę i Q3. Dołączone kolumny pokazują brakującą część danych oraz jednostkę.

Zmienna	n i braki	Średnia	Mediana
Temperatura °C	720 i 0	59,275	59,330
Drgania mm/s	720 i 0	1,246	1,243
Energia kWh	718 i 2	13,929	14,235
Produkcja szt.	719 i 1	152,676	153,000

Zmienna	Q1 do Q3	Minimum do maksimum	Odchylenie standardowe
Temperatura °C	57,17–61,60	52,70–66,25	2,880
Drgania mm/s	1,132–1,3515	0,781–4,850	0,205
Energia kWh	12,84–14,9775	11,48–15,89	1,233
Produkcja szt.	147–158	133–173	7,144

Tabele pokazują zaokrąglone wyniki. W obiekcie statystyki zachowaj pełną precyzję do dalszych obliczeń. Z podanych kwartyli otrzymasz IQR temperatury 4,43 °C, drgań 0,2195 mm/s, energii 2,1375 kWh i produkcji 11 sztuk. Określenie „duże zróżnicowanie” wymagałoby porównywalnego punktu odniesienia.

Wpływ maksimum i odpowiedzialna interpretacja

Sprawdź, jak zmienią się podsumowania drgań po tymczasowym wyłączeniu jednego wskazanego maksimum. Eksperyment wykonaj na osobnej serii. Nie zmieniaj pomiary_clean ani pliku z M2. Wariant odpowiada na pytanie o wrażliwość wyniku, nie rozstrzyga poprawności samego pomiaru.

```
pelne = pomiary_clean["drgania"]
bez_max = pelne.drop(index=pelne.idxmax())
wrazliwosc_statystyk = pd.DataFrame({
    "pelny": pelne.agg(["count", "mean", "median", "std"]),
    "bez_maksimum": bez_max.agg(["count", "mean", "median", "std"])
}).T
display(wrazliwosc_statystyk.round(3))
```

Dla pełnych 720 obserwacji średnia wynosi około 1,246 mm/s, mediana 1,243, a odchylenie 0,205. Dla 719 wartości otrzymasz odpowiednio 1,241, 1,242 i 0,154 mm/s. Mediana zmienia się minimalnie, podczas gdy odchylenie maleje o około jedną czwartą. Bliskość średniej i mediany nie wystarczyła więc do oceny wpływu maksimum.

Cztery warstwy odpowiedzi

Wynik obliczenia zawiera konkretną miarę, jednostkę i liczebność. Obserwacja mówi, co widać w wyniku. Interpretacja wskazuje możliwe znaczenie lub kierunek dalszego sprawdzenia. Ograniczenie wyznacza granicę tego, co można stwierdzić. Zachowaj ten podział w każdej odpowiedzi modułu.

Dla temperatury możesz zapisać: „Mediana wynosi 59,33 °C dla 720 pomiarów, a Q1 i Q3 to 57,17 i 61,60 °C. Środkowa połowa danych mieści się w tym przedziale. Warto sprawdzić, czy podobny poziom występuje we wszystkich urządzeniach. Podsumowanie całego zbioru nie rozstrzyga różnic między grupami ani warunków ich pracy”.

Ćwiczenie w notebooku

Przygotuj po jednej obserwacji dla każdej z czterech zmiennych, z jednostką i liczebnością. Dla drgań uwzględnij eksperyment wrażliwości. Odczytaj P10 i P90 i wyjaśnij je własnymi słowami. Tabelę statystyk zachowaj do eksportu na końcu książki.

Sprawdź rozumienie 2

Średnia i mediana drgań prawie się pokrywają. Czy można pominąć analizę wartości skrajnych? Dlaczego P90 nie jest prognozą następnego odczytu?

Rozdział 3 Filtrowanie sortowanie i analiza liczebności

Pytanie M3-P02 dotyczy pomiarów drgań ściśle powyżej 1,5 mm/s. Próg jest roboczym kryterium ćwiczenia. Nie zastępuje normy technicznej ani progu diagnostycznego z M2. Najpierw wybierzesz rekordy, później uporządkujesz je i sprawdzisz, jak rozkładają się między urządzenia i okresy.

Maska logiczna i wybór przez loc

Maska zawiera jedną odpowiedź True lub False dla każdego wiersza. Jej suma jest liczbą spełnionych warunków, ponieważ True jest przy sumowaniu traktowane jak jeden. Wyrażenie .loc[maska, kolumny] wybiera równocześnie wiersze i potrzebne pola. Osobna kopia pozwala pracować na podzbiorze bez nadpisania pełnych danych.

```
prog = 1.5
maska = pomiary_clean["drgania"] > prog
kolumny = ["czas", "urzadzenie", "drgania", "rok_miesiac"]
powyzej_progu = pomiary_clean.loc[maska, kolumny].copy()
assert len(powyzej_progu) == maska.sum() == 28
assert powyzej_progu["drgania"].gt(prog).all()
assert pomiary_clean.shape == (720, 14)
display(powyzej_progu.head())
```

W podzbiorze znajduje się 28 wierszy, od 1,502 do 4,850 mm/s. Pełna tabela nadal ma 720 rekordów. Filtrowanie nie usuwa wierszy ze źródła. Dopiero przypisanie wyniku do tej samej nazwy mogłoby sprawić, że dalszy kod nieświadomie korzystałby tylko z wybranych pomiarów.

Granice i kilka warunków

Operator > wyłącza wartość równą progowi, a >= ją włącza. W danych nie ma dokładnie 1,500 mm/s, więc oba zapisy dają 28 wierszy. To zgodność na konkretnym zbiorze, nie równoważność warunków. Metoda between(1.5, 2.0, inclusive="right") opisuje przedział większy od 1,5 i nie większy od 2,0; obejmuje 27 pomiarów.

Dwa warunki na seriach łącz operatorem & i umieszczaj każdy w nawiasach. Operator | oznacza co najmniej jeden spełniony warunek, a ~ negację. Metoda isin() wskazuje członkostwo na liście, a notna() dostępność wartości. Nie zastępuj & i | słowami and i or w zwykłym wyrażeniu na seriach Pandas.

Czytelny filtr i powtarzalne sortowanie

Query() przyjmuje tekst opisujący warunek. Znak @ odwołuje się do zmiennej Pythona spoza tego tekstu. Słowo and jest tutaj częścią wyrażenia interpretowanego przez query(), dlatego jego użycie różni się od and pomiędzy dwiema seriami.

```
miesiac = "2026-02"
maska_luty = ((pomiary_clean["drgania"] > prog)
              & (pomiary_clean["rok_miesiac"] == miesiac))
powyzej_luty = pomiary_clean.query(
    "drgania > @prog and rok_miesiac == @miesiac"
)
pd.testing.assert_frame_equal(
    pomiary_clean.loc[maska_luty], powyzej_luty
)
assert len(powyzej_luty) == 13
ranking_pomiarow = powyzej_progu.sort_values(
    ["drgania", "czas", "urzadzenie"],
    ascending=[False, True, True]
)
display(ranking_pomiarow.head())
```

Sortowanie nie zmienia liczby rekordów. Najpierw ustawia malejąco drgania, a przy równych wartościach rozstrzyga kolejność czasem i identyfikatorem. To ważne dla powtarzalnego wyboru pierwszych pozycji. Trzy odczyty U-07 mają drgania 1,595 mm/s, lecz pochodzą z różnych terminów.

Najwyższy odczyt to 4,850 mm/s dla U-05 z 8 lutego 2026 roku o 20:00. Następny to 1,620 mm/s dla U-07 w tym samym czasie. Pozostałych 27 przekroczeń mieści się w przedziale 1,502–1,620 mm/s. Ranking wskazuje rekord do wyjaśnienia, ale nie mówi, czy jego przyczyną jest błąd czujnika, czy rzeczywiste zdarzenie.

Kategorie bez trafień

Value_counts() dla tekstowych identyfikatorów pokaże tylko urządzenia obecne w podzbiorze. Brak wiersza z U-01 nie wyjaśnia, czy urządzenie nie było badane, czy nie miało przekroczeń. Reindex z pełną listą grup pozwala wpisać zero tam, gdzie nie było trafień. W naszym kontrakcie wszystkie osiem urządzeń występuje w pomiarach. Dla nowego zbioru pełną listę zakresu badania należy uzgodnić z kartoteką.

Zero przekroczeń ma sens tylko przy istniejącej podstawie obserwacji. Urządzenie bez pomiarów ma nieznany odsetek, a nie 0%. Nie wypełniaj zerem brakującego mianownika.

Mianownik nadaje znaczenie odsetkowi

Liczba przekroczeń mówi, ile rekordów wybrano. Odsetek wymaga wskazania, do jakiej liczby pomiarów ją odnosisz. Funkcja poniżej oblicza licznik, liczbę dostępnych drgań i odsetek w tej samej grupie. Zachowuje pełną listę grup wynikającą ze zbioru wejściowego.

```
def zestaw_przekroczenia(pole):
    baza = pomiary_clean.groupby(pole)["drgania"].count()
    trafienia = powyzej_progu[pole].value_counts().reindex(
        baza.index, fill_value=0
    )
    wynik = pd.DataFrame({"przekroczenia": trafienia,
                          "pomiary_n": baza})
    wynik["udzial_proc"] = wynik["przekroczenia"].div(
        wynik["pomiary_n"].replace(0, np.nan)
    ).mul(100)
    assert wynik["przekroczenia"].sum() == len(powyzej_progu)
    assert wynik["przekroczenia"].le(wynik["pomiary_n"]).all()
    return wynik.rename_axis(pole).reset_index()

przekroczenia_urzadzenia = zestaw_przekroczenia("urzadzenie")
przekroczenia_miesiace = zestaw_przekroczenia("rok_miesiac")
display(przekroczenia_urzadzenia.round(1))
display(przekroczenia_miesiace.round(1))
```

U-07 ma 13 przekroczeń na 90 pomiarów, czyli 14,4%; U-05 ma 8 na 90, czyli 8,9%; U-08 ma 7 na 90, czyli 7,8%. Pozostałe pięć urządzeń ma po zero na 90 pomiarów. Cały zbiór zawiera 28 przekroczeń na 720 rekordów, około 3,9%.

Okres	Przekroczenia i podstawa	Odsetek
Styczeń 2026	15 z 496 pomiarów	3,0%
Od 1 do 14 lutego 2026	13 z 224 pomiarów	5,8%

W dostępnej części lutego licznik jest mniejszy, a odsetek większy. Różna liczba pomiarów wyjaśnia, dlaczego sam licznik mógłby prowadzić do innego odczytania wyniku. Nadal porównujesz tylko dwa okresy o różnej długości. Nie nazywaj tego potwierdzonym trendem ani dowodem pogorszenia stanu urządzeń.

Ćwiczenie w notebooku i sprawdzenie rozumienia 3

Powtórz wybór dla temperatury > 64 °C. Wyświetl pięć najwyższych wyników, policz trafienia i udziały dla wszystkich urządzeń. Potwierdź zgodność sumy maski, liczby wybranych wierszy i sumy trafień. Wyjaśnij bez komputera: czy mniejsza liczba zdarzeń zawsze oznacza mniejszą częstość ich występowania?

Rozdział 4 Grupowanie i agregacje

Grupowanie dzieli rekordy według klucza i oblicza miary dla każdej grupy. W M3-P03 jeden wiersz wyniku opisuje urządzenie. W M3-P04 porównasz tryby, typy oraz ich kombinacje.

Wróć do pełnych 720 pomiarów. Podzbiór powyzej_progu dałby mediany tylko wysokich drgań trzech urządzeń. Poprawny technicznie kod odpowiedziałby na inne pytanie.

Ranking median dla M3 P03

```
ranking_urzadzen = (
    pomiary_clean.groupby("urzadzenie", as_index=False)
    .agg(liczba_pomiarow=("drgania", "count"),
         mediana_drgan=("drgania", "median"))
    .sort_values(["mediana_drgan", "urzadzenie"],
                 ascending=[False, True])
)
assert ranking_urzadzen.shape == (8, 3)
assert ranking_urzadzen["liczba_pomiarow"].eq(90).all()
assert ranking_urzadzen["liczba_pomiarow"].sum() == 720
display(ranking_urzadzen)
```

Agregacja nazwana łączy nazwę wyniku, pole i funkcję, np. mediana_drgan=("drgania", "median"). As_index=False zachowuje klucz jako kolumnę do sortowania i eksportu.

Urządzenie	Mediana drgań mm/s	Liczba pomiarów
U-05	1,3735	90
U-08	1,3605	90
U-07	1,3460	90
U-03	1,2520	90
U-06	1,2500	90
U-02	1,1575	90
U-04	1,1250	90
U-01	1,0380	90

Pierwsze trzy urządzenia odpowiadają na M3-P03, ale zachowaj cały ranking. Mediana U-05 przewyższa medianę U-07 o 0,0275 mm/s. Wynik opisuje typowy poziom drgań w badanym okresie. Bez norm i historii zdarzeń nie uzasadnia określenia urządzenia jako uszkodzonego ani zalecenia naprawy.

Liczba rekordów i liczba dostępnych wartości

Size liczy wszystkie wiersze grupy. Count liczy niebrakujące wartości wskazanego pola. W przypadku energii i produkcji obie wielkości mogą się różnić. Jedna wspólna kolumna n dla całej tabeli nie pokaże wtedy, z ilu obserwacji powstała każda mediana.

Pytanie M3-P04 wymaga trzech poziomów porównania. Zaczynamy od dwóch trybów pracy, potem czterech typów urządzeń, a na końcu ośmiu kombinacji typu i trybu. Funkcja pozwala zachować ten sam zestaw miar przy zmianie klucza.

```
def agreguj_prace(klucz):
    return pomiary_clean.groupby(
        klucz, as_index=False, observed=True, dropna=False
    ).agg(
        rekordy=("urzadzenie", "size"),
        energia_n=("energia", "count"),
        energia_mediana=("energia", "median"),
        produkcja_n=("produkcja", "count"),
        produkcja_mediana=("produkcja", "median")
    )

wedlug_trybu = agreguj_prace("tryb")
wedlug_typu = agreguj_prace("typ_urzadzenia")
wedlug_typu_trybu = agreguj_prace(["typ_urzadzenia", "tryb"])
for tabela in [wedlug_trybu, wedlug_typu, wedlug_typu_trybu]:
    assert tabela["rekordy"].sum() == 720
    assert tabela["energia_n"].sum() == 718
    assert tabela["produkcja_n"].sum() == 719
assert len(wedlug_typu_trybu) == 8
display(wedlug_trybu)
```

Observed=True ogranicza grupy kategoryczne do obserwowanych kombinacji. Dropna=False pozwala nie pomijać grup z brakującym kluczem. W danych kursowych klucze są kompletne, a wszystkie oczekiwane kombinacje istnieją. Przy innym zbiorze trzeba porównać wynik z zakresem pytania, zwłaszcza gdy niektóre kombinacje nie występują.

Tryb	Rekordy	n energii i produkcji	Mediany energii i produkcji
auto	360	359 i 359	14,25 kWh i 152 szt.
oszczedny	360	359 i 360	14,23 kWh i 153 szt.

Tryby obejmują po tyle samo rekordów, ale mediany produkcji mają różne podstawy. Dwa braki energii rozkładają się po jednym na tryb, a brak produkcji należy do auto. Sumy liczebności powinny zgadzać się z wejściem. Median grup nie sumuje się w celu uzyskania mediany całego zbioru.

Porównanie ogólne i porównanie wewnątrz typów

Ogólne mediany energii trybów różnią się o 0,02 kWh, a rozpiętość median typów wynosi około 1,70 kWh. Przenośniki mają medianę 14,58 kWh, sprężarki 14,25, wentylatory 13,91, a pompy 12,88. Ogólne porównanie trybów nie pokazuje tej struktury.

Każda z ośmiu kombinacji obejmuje 90 rekordów. W nawiasach przy medianach podano liczby dostępnych wartości. Tabela nie porównuje sprawności urządzeń.

Typ urządzenia	Tryb	Energia kWh i n	Produkcja szt. i n
Pompa obiegowa	auto	12,880 (89)	156,0 (89)
Pompa obiegowa	oszczedny	13,600 (90)	157,0 (90)
Przenośnik	auto	14,580 (90)	150,5 (90)
Przenośnik	oszczedny	14,545 (90)	150,0 (90)
Sprężarka	auto	14,240 (90)	151,0 (90)
Sprężarka	oszczedny	14,270 (90)	152,0 (90)
Wentylator przemysłowy	auto	13,195 (90)	153,5 (90)
Wentylator przemysłowy	oszczedny	13,910 (89)	153,0 (90)

Różnica energii oszczedny względem auto jest dodatnia dla pompy, sprężarki i wentylatora, a ujemna dla przenośnika. Różnice produkcji też nie mają jednego kierunku. Twierdzenie „tryb oszczędny jest zawsze lepszy” wykraczałoby poza dane.

Podział na typy nadal nie kontroluje obciążenia, mocy ani konkretnego urządzenia. Mediana energii nie mierzy efektywności energetycznej. Taka ocena wymaga definicji efektu pracy, porównywalnych warunków i odpowiedniej miary.

Ćwiczenie w notebooku

Porównaj temperaturę między lokalizacjami przez size, count i medianę. Posortuj mediany malejąco. Sprawdź sumę 720 obu liczebności i wyjaśnij ich równość. Zapisz obserwację oraz ograniczenie dotyczące składu urządzeń.

Sprawdź rozumienie 4

Dlaczego mediany obliczone na 28 przekroczeniach nie odpowiadają na M3-P03? Czy prawie identyczna mediana energii dwóch trybów wyklucza różnice wewnątrz typów urządzeń?

Rozdział 5 Tabele przestawne i porównania

Tabela przestawna zestawia jedną miarę według dwóch wymiarów. W pytaniu M3-P05 wiersze reprezentują miesiące, kolumny lokalizacje, a komórki medianę energii. Dzięki temu możesz porównywać lokalizacje w tym samym okresie i okresy w tej samej lokalizacji. Zawsze zaczynaj od pełnej ramki pomiary_clean.

Znaczenie komórki i wybór agregacji

```
energia_ml = pomiary_clean.pivot_table(
    index="rok_miesiac", columns="lokalizacja",
    values="energia", aggfunc="median"
).sort_index()
kontrola_ml = pomiary_clean.groupby(
    ["rok_miesiac", "lokalizacja"]
)["energia"].median().unstack().sort_index()
pd.testing.assert_frame_equal(energia_ml, kontrola_ml)
assert energia_ml.shape == (2, 3)
assert energia_ml.notna().all().all()
display(energia_ml.round(3))
```

Index i columns określają układ, values wskazuje pole liczbowe, a aggfunc definiuje sposób połączenia wielu pomiarów w jedną wartość. Pivot() bez agregacji nie wystarczy: dla stycznia i Hali A istnieje 186 rekordów, a nie jedna wartość. Pivot_table() jawnie oblicza z nich medianę. To agregowanie, nie usuwanie duplikatów.

Okres	Hala A kWh	Hala B kWh	Hala C kWh
Styczeń 2026	13,160	14,790	14,645
Od 1 do 14 lutego 2026	13,080	14,695	14,760

W obu okresach Hala A ma najniższą medianę energii. Kolejność Hal B i C się zmienia. Pojedyncza komórka opisuje medianę wielu pomiarów, a nie jeden odczyt. Kontrola drugą metodą, przez groupby() i unstack(), potwierdza zgodność wartości i układu wyników.

```
roznica_median = energia_ml.loc["2026-02"] - energia_ml.loc["2026-01"]
display(roznica_median.round(3))
```

Różnice wynoszą −0,080 kWh dla Hali A, −0,095 dla B i +0,115 dla C. Opisują dwa dostępne okresy o różnej długości. Nie stanowią dowodu trendu miesięcznego. Odmienny skład urządzeń także ogranicza przypisywanie różnic samej lokalizacji.

Liczebności braki i marginesy

Mediana może wyglądać wiarygodnie nawet wtedy, gdy w jej grupie brakuje pomiarów. Dlatego obok wartości pokaż liczby wszystkich rekordów oraz dostępnych energii. Ich różnica wskazuje, gdzie występują braki.

```
energia_wiersze = pomiary_clean.pivot_table(
    index="rok_miesiac", columns="lokalizacja",
    aggfunc="size", fill_value=0
)
energia_n = pomiary_clean.pivot_table(
    index="rok_miesiac", columns="lokalizacja",
    values="energia", aggfunc="count", fill_value=0
)
display(energia_wiersze)
display(energia_n)
energia_ml_razem = pomiary_clean.pivot_table(
    index="rok_miesiac", columns="lokalizacja", values="energia",
    aggfunc="median", margins=True, margins_name="Razem"
)
assert np.isclose(energia_ml_razem.loc["Razem", "Razem"],
                  pomiary_clean["energia"].median())
display(energia_ml_razem.round(3))
```

Okres i miara	Hala A	Hala B	Hala C
Styczeń wszystkie rekordy	186	186	124
Styczeń dostępna energia	186	184	124
Luty wszystkie rekordy	84	84	56
Luty dostępna energia	84	84	56

Dwa braki energii występują w styczniu w Hali B. Zero jest właściwym uzupełnieniem liczebności nieobserwowanej kombinacji, ale nie mediany. Gdy brak dostępnych wartości energii, wynik mediany powinien pozostać nieznany. W naszych sześciu kombinacjach każda mediana istnieje.

Margins=True dodaje podsumowania obliczone ponownie z rekordów źródłowych. Mediana wszystkich 718 dostępnych energii wynosi 14,235 kWh i trafia do komórki Razem × Razem. Średnia sześciu median wynosi około 14,188 kWh. Nie jest tym samym wynikiem, ponieważ mediana nie daje się w ten sposób agregować. Analogicznie margines miesiąca lub lokalizacji nie jest sumą widocznych median.

Tabela krzyżowa i jawny mianownik

Crosstab() domyślnie liczy współwystępowanie kategorii. Dla lokalizacji i typu urządzenia pokaże liczbę rekordów pomiarowych w każdej parze. To inne pytanie niż mediana energii. Każde urządzenie wnosi tutaj 90 pomiarów, więc liczba rekordów nie jest bezpośrednio liczbą urządzeń.

```
typy_lokalizacje = pd.crosstab(
    pomiary_clean["lokalizacja"], pomiary_clean["typ_urzadzenia"],
    margins=True, margins_name="Razem"
)
udzialy_rekordow_typow = pd.crosstab(
    pomiary_clean["lokalizacja"], pomiary_clean["typ_urzadzenia"],
    normalize="index"
).mul(100)
assert typy_lokalizacje.loc["Razem", "Razem"] == 720
assert np.allclose(udzialy_rekordow_typow.sum(axis=1), 100)
display(typy_lokalizacje)
display(udzialy_rekordow_typow.round(1))
```

Normalize="index" odnosi każdą komórkę do sumy jej wiersza, tutaj liczby rekordów lokalizacji. Dla Hali A pompa, sprężarka i wentylator mają po około 33,3%, a przenośnik 0%. Dla Hali B po około 33,3% mają pompa, przenośnik i wentylator. Hala C dzieli się po 50% na przenośnik i sprężarkę.

Po zaokrągleniu trzy wartości 33,3% sumują się do 99,9%. To efekt prezentacji liczb, dlatego kontrolę wykonujesz przed zaokrągleniem. W tym zbiorze równe liczby pomiarów na urządzenie sprawiają, że udziały rekordów odpowiadają też strukturze urządzeń. Przy nierównej częstości pomiarów ta zgodność mogłaby zniknąć.

Normalize="columns" używa sumy kolumny i pyta o rozkład danego typu między lokalizacje. Normalize="all" odnosi komórkę do całego zbioru. Brak normalizacji daje liczebności. Najpierw wybierz pytanie i podstawę porównania, dopiero potem parametr funkcji.

Ćwiczenie w notebooku

Zbuduj analogiczną tabelę median produkcji z marginesami Razem oraz tabelę count. Obie z marginesami powinny mieć 3 wiersze i 4 kolumny, a globalna liczebność produkcji wynosić 719. Zapisz obserwację z jednostką oraz ograniczenie dotyczące niepełnego lutego.

Sprawdź rozumienie 5

Dlaczego komórka Razem × Razem z medianą energii nie jest średnią sześciu median? Co stanowi mianownik normalize="index" w tabeli lokalizacja × typ urządzenia?

Rozdział 6 Zależności korelacja i ostrożna interpretacja

Pytanie M3-P06 brzmi: czy wyższej temperaturze towarzyszą większe wartości drgań? Badamy współzmienność dwóch liczb z tego samego pomiaru. Współczynnik korelacji opisuje wzorzec w danych. Nie ustala kierunku oddziaływania ani nie wyklucza wspólnego wpływu obciążenia, urządzenia lub czasu.

Kompletne pary z tych samych wierszy

Ponownie wróć do pełnej tabeli, a nie do 28 przekroczeń ani do median lokalizacji. Jeden punkt powinien reprezentować jedną temperaturę i drgania zmierzone w tym samym czasie dla tego samego urządzenia. Oddzielne usuwanie braków z dwóch list mogłoby zerwać to powiązanie.

```
pary = pomiary_clean[
    ["czas", "urzadzenie", "temperatura", "drgania"]
].dropna(subset=["temperatura", "drgania"]).copy()
assert len(pary) == 720
assert pary[["temperatura", "drgania"]].notna().all().all()
print("Pominięte rekordy:", len(pomiary_clean) - len(pary))
```

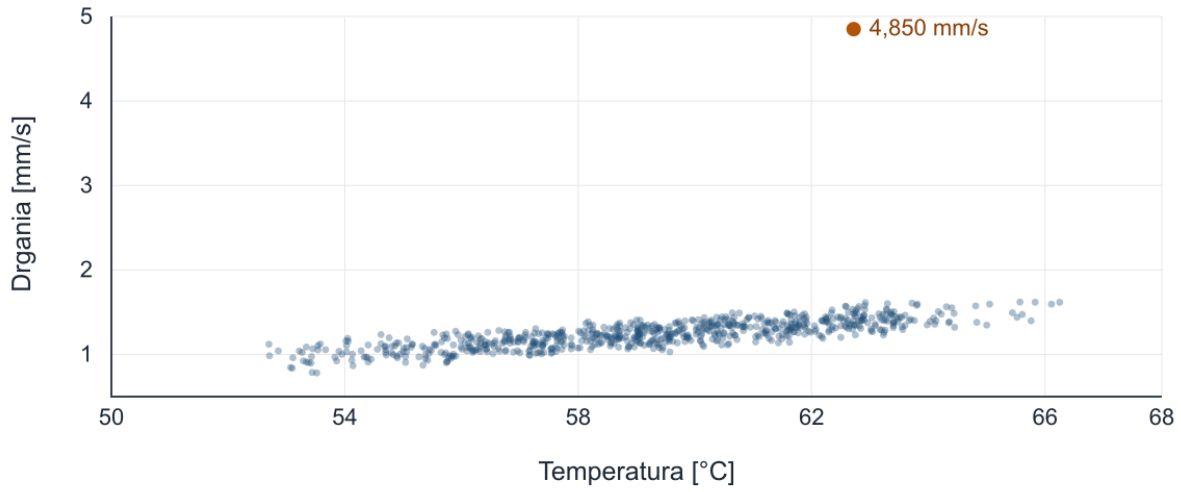
Temperatura i drgania są kompletne, więc pozostaje 720 par. To 720 powtarzanych pomiarów z ośmiu urządzeń, nie 720 niezależnych urządzeń. Liczba rekordów i liczba obiektów badania to różne informacje, istotne przy uogólnianiu wyniku.

Wykres jako kontrola zależności

Przed obliczeniem korelacji obejrzyj wykres punktowy. Opisz kierunek, kształt, rozproszenie i przypadki nietypowe. Czy widzisz pas zbliżony do prostej, krzywą albo kilka skupień? Czy jeden punkt leży daleko od pozostałych? Korelacja liniowa nie zastępuje tej kontroli.

```
import matplotlib.pyplot as plt
fig, ax = plt.subplots(figsize=(7.5, 4.5))
ax.scatter(pary["temperatura"], pary["drgania"], alpha=.35, s=28)
punkt = pary.loc[pary["drgania"].idxmax()]
ax.scatter(punkt["temperatura"], punkt["drgania"],
           color="tab:orange", s=80)
ax.set_xlabel("Temperatura [°C]")
ax.set_ylabel("Drgania [mm/s]")
ax.set_title(f"Temperatura a drgania — {len(pary)} pomiarów")
ax.grid(alpha=.2)
FIGURES_DIR.mkdir(parents=True, exist_ok=True)
fig.savefig(FIGURES_DIR / "M3_P06_temperatura_drgania.png",
            dpi=150, bbox_inches="tight")
plt.show()
```

Kierunek kształt i siła zależności



Rysunek 1. Temperatura i drgania dla 720 pomiarów kursowych. Wyróżniono maksimum 4,850 mm/s. Punkty przedstawiają rzeczywiste pary z danych, a nie schemat poglądowy.

Wyższym temperaturom w tym zbiorze zwykle towarzyszą wyższe drgania. Chmura nie tworzy jednak jednej wąskiej prostej; widać zróżnicowanie punktów i odseparowane maksimum. Taki obraz uzasadnia obliczenie współczynnika oraz sprawdzenie wpływu punktu i struktury urządzeń.

Współczynnik Pearsona r mieści się od −1 do 1, gdy jest określony. Znak wskazuje kierunek, a wartość bezwzględna opisuje siłę zależności liniowej. Wynik nie ma jednostki. R bliskie zeru nie wyklucza zależności nieliniowej; nie ma też jednego uniwersalnego progu „silnej” korelacji dla wszystkich zastosowań.

```
r_pelny = pary["temperatura"].corr(pary["drgania"], method="pearson")
n_pelny = len(pary)
assert -1 <= r_pelny <= 1
macierz_r = pomiary_clean[MIARY].corr(method="pearson")
display(macierz_r.round(3))
```

Dla temperatury i drgań r wynosi około 0,644 przy n = 720. W macierzy temperatura i energia mają r około 0,846, temperatura i produkcja −0,762, a energia i produkcja −0,658. To tropy do dalszej analizy. Każda komórka może korzystać z innej liczby par: temperatura lub drgania z energią mają 718, z produkcją 719, a energia z produkcją 717. Sama macierz współczynników tej różnicy nie pokazuje.

Analiza wrażliwości i poziom urządzenia

Wyłącz w osobnym wariancie jeden wskazany pomiar maksimum drgań. Nie usuwaj wszystkich rekordów powyżej arbitralnego progu i nie zmieniaj źródła. Zanotuj klucz wyłączonego rekordu, aby eksperyment można było odtworzyć.

```
idx_punktu = pary["drgania"].idxmax()
display(pary.loc[[idx_punktu]])
pary_wariant = pary.drop(index=idx_punktu).copy()
r_wariant = pary_wariant["temperatura"].corr(
    pary_wariant["drgania"], method="pearson"
)
assert len(pary_wariant) == 719
wiersze_r = []
for urzadzenie, grupa in pary.groupby("urzadzenie"):
    r = grupa["temperatura"].corr(grupa["drgania"], method="pearson")
    wiersze_r.append({"urzadzenie": urzadzenie, "n": len(grupa), "r": r})
r_urzadzenia = pd.DataFrame(wiersze_r)
display(r_urzadzenia.round(3))
```

R zmienia się z około 0,644 dla 720 par do 0,817 dla 719 par. Kierunek pozostaje dodatni, ale wartość współczynnika jest wrażliwa na pojedynczy pomiar. Większe r nie jest powodem, aby wybrać wariant jako „lepszy” i ukryć wynik pełny. Podaj oba. Status wartości 4,850 mm/s nadal wymaga danych źródłowych lub technicznego kontekstu.

Urządzenie	Korelacja Pearsona	Liczba par
U-01	0,393	90
U-02	0,494	90
U-03	0,325	90
U-04	0,352	90
U-05	0,102	90
U-06	0,356	90
U-07	0,778	90
U-08	0,386	90

Wszystkie kierunki są dodatnie, lecz siedem współczynników jest niższych od ogólnego 0,644. Wynik zbiorczy łączy różnice poziomów między urządzeniami ze zmianami wewnątrz nich. U-05 ma r około 0,102, a U-07 około 0,778; sama ta różnica nie wyjaśnia mechanizmu. Dla stałej zmiennej lub zbyt małej liczby par współczynnik może być nieokreślony. NaN nie zastępuj zerem.

Hipotezy modelu językowego i werdykt analizy

LLM może rozszerzyć listę możliwych wyjaśnień. Poproś o trzy alternatywne hipotezy dodatniej zależności temperatury i drgań oraz dane lub obliczenia potrzebne do ich sprawdzenia. Zaznacz, że model nie może wymyślać wyników ani przedstawiać hipotez jako potwierdzonych przyczyn. Przekaż tylko potrzebny kontekst, bez danych poufnych.

Hipotezę o wpływie różnic między urządzeniami możesz sprawdzić przez porównanie korelacji ogólnej i wewnątrz grup. Hipotezę o wpływie maksimum sprawdza wariant diagnostyczny. Hipotezy o obciążeniu nie rozstrzygniesz bez danych o obciążeniu lub historii pracy. Zapis „nierozstrzygnięta z powodu braku danych” jest poprawnym wynikiem tej oceny.

Odpowiedź modelu „korelacja dowodzi, że temperatura powoduje wzrost drgań; usuń maksimum, bo pogarsza wynik; U-07 wymaga naprawy” zawiera trzy nieuzasadnione kroki. Korelacja nie ustala przyczyny, zwiększenie r nie uzasadnia usuwania obserwacji, a współczynnik nie jest diagnozą stanu urządzenia. Oceniaj każdy fragment osobno. Część o dodatniej współzmienności może być zgodna z obliczeniami, mimo błędów w reszcie wypowiedzi.

Odpowiedź na M3 P06

Wynik: dla 720 kompletnych par korelacja Pearsona wynosi około 0,644. Obserwacja: wyższym temperaturom w tym zbiorze zwykle towarzyszą większe drgania. Interpretacja: warto zbadać udział różnic między urządzeniami, obciążenia i zmian w czasie. Ograniczenie: wariant bez jednego maksimum daje około 0,817, a wyniki dla urządzeń mieszczą się od 0,102 do 0,778. Dane nie dowodzą związku przyczynowego ani awarii.

Werdykt może brzmieć: „Hipoteza dodatniej współzmienności jest wsparta w badanym zbiorze. Jej kierunek pozostaje zgodny, lecz siła zależy od zakresu i struktury grup. Przyczyna nie została ustalona”. W docs/rejestr_uzycia_LLM.csv zapisz rzeczywisty prompt, odpowiedź, kontrolę, werdykt i uzasadnienie z oznaczeniem modul = M3. Przy pracy bez LLM odnotuj brak jego użycia.

Ćwiczenie w notebooku

Zbadaj energię i produkcję. Zachowaj kompletne pary z tych samych wierszy, sprawdź ich liczbę 717, przygotuj wykres, odtwórz r około −0,658 i porównaj osiem wyników dla urządzeń. Nie przenoś do tego zadania gotowej ramki pary temperatury i drgań. Dla chętnych: porównaj Pearsona i Spearmana, który bada monotoniczną zależność rang.

Sprawdź rozumienie 6

Po pominięciu maksimum korelacja rośnie. Czy należy usunąć pomiar i opublikować wyłącznie większy współczynnik? Jakich informacji nie dostarcza nawet dodatnie r we wszystkich urządzeniach?

Zapis wyników i przygotowanie do modułu 4

Wynik EDA powinien dać się wykorzystać bez odtwarzania ręcznych poprawek. Zachowaj notebook z kodem, tabelami i komentarzami oraz eksporty powiązane z pytaniami. Dane wejściowe w processed pozostają niezmienione. Wykres z rozdziału 6 jest diagnostyczny; finalne wizualizacje i narrację rozwiniesz w M4.

Stałe nazwy i zawartość eksportów

Używaj katalogu outputs/tables, zgodnie z kartami M1–M3. Output/tables ze slajdów i notatek oznacza miejsce o tej samej roli. Nie twórz obu katalogów. Prezentacyjne statystyki_opisowe.csv i energia_miesiac_lokalizacja.csv zapisujemy z prefiksami M3_P01 i M3_P05. W tekście pytanie ma zapis M3-P01, a w nazwie pliku M3_P01.

Plik w outputs/tables	Zawartość
M3_P01_statystyki_opisowe.csv	Cztery zmienne, statystyki, jednostki i braki
M3_P02_przekroczenia_urzadzenia.csv	Licznik, mianownik i udział dla urządzeń
M3_P02_przekroczenia_miesiace.csv	Licznik, mianownik i udział dla okresów
M3_P03_ranking_drgan.csv	Osiem urządzeń, mediana drgań i liczebność
M3_P04_tryb.csv	Agregacje energii i produkcji według trybu
M3_P04_typ.csv	Te same miary według typu urządzenia
M3_P04_typ_tryb.csv	Te same miary dla kombinacji typu i trybu
M3_P05_energia_miesiac_lokalizacja.csv	Sześć kombinacji okresu i lokalizacji z medianą i liczebnościami
M3_P06_korelacje.csv	Zakres, urządzenie, wariant, metoda, n i r

Notebook pozostaje w notebooks/M3_EDA.ipynb, słownik w docs/slownik_danych.md, a rejestr LLM w docs/rejestr_uzycia_LLM.csv. Wykres pomocniczy zapisujesz jako outputs/figures/M3_P06_temperatura_drgania.png. Pozostałe macierze, kontrole i ćwiczenia mogą pozostać w notebooku.

Nazwy plików nie zawierają spacji ani polskich znaków. Unikaj nazw final, final2 i ostateczny. Prefiks wskazuje moduł i pytanie, a reszta opisuje zawartość. W README.md dopisz sposób uruchomienia M3, wersje bibliotek, listę wyników oraz związek z zatwierdzonym wejściem M2.

Tabele w układzie gotowym do zapisu

W notebooku macierz jest wygodna do porównywania, a w pliku często przydatny jest układ długi: jeden wiersz na kombinację kluczy. Eksport P05 będzie więc zawierał okres, lokalizację, medianę energii oraz obie liczebności. Jest to odpowiednik macierzy z rozdziału 5 wzbogacony o kontrole, a nie inna analiza.

```
energia_do_zapisu = pomiary_clean.groupby(
    ["rok_miesiac", "lokalizacja"], as_index=False
).agg(
    energia_mediana=("energia", "median"),
    rekordy=("urzadzenie", "size"), energia_n=("energia", "count")
)
korelacje = pd.DataFrame([
    {"zakres": "caly_zbior", "urzadzenie": "nie_dotyczy",
     "wariant": "pelny", "metoda": "pearson", "n": len(pary),
     "r": r_pelny},
    {"zakres": "caly_zbior", "urzadzenie": "nie_dotyczy",
     "wariant": "bez_maksimum", "metoda": "pearson",
     "n": len(pary_wariant), "r": r_wariant}
])
lokalne = r_urzadzenia.assign(
    zakres="urzadzenie", wariant="pelny", metoda="pearson"
)
korelacje = pd.concat([korelacje, lokalne], ignore_index=True)
eksporty = {
    "M3_P01_statystyki_opisowe.csv": statystyki.reset_index(),
    "M3_P02_przekroczenia_urzadzenia.csv": przekroczenia_urzadzenia,
    "M3_P02_przekroczenia_miesiace.csv": przekroczenia_miesiace,
    "M3_P03_ranking_drgan.csv": ranking_urzadzen,
    "M3_P04_tryb.csv": wedlug_trybu,
    "M3_P04_typ.csv": wedlug_typu,
    "M3_P04_typ_tryb.csv": wedlug_typu_trybu,
    "M3_P05_energia_miesiac_lokalizacja.csv": energia_do_zapisu,
    "M3_P06_korelacje.csv": korelacje
}
```

Nazwy grup muszą być kolumnami, zanim użyjesz index=False. Dla statystyk reset_index() przenosi nazwę zmiennej z indeksu do pola zmienna. Brak tego kroku pozbawiłby eksport etykiet. Pozostałe tabele przygotowano wcześniej w układzie kolumnowym. Zapis r powinien zawsze zawierać n, zakres i wariant, aby większy współczynnik nie został pomylony z wynikiem pełnym.

Kontrola zapisu i audyt odpowiedzi

```
TABLES_DIR.mkdir(parents=True, exist_ok=True)
for nazwa, tabela in eksporty.items():
    tabela = tabela.reset_index(drop=True).copy()
    tabela.to_csv(TABLES_DIR / nazwa, index=False, encoding="utf-8")
    odczyt = pd.read_csv(TABLES_DIR / nazwa, encoding="utf-8")
    wzorzec = tabela.copy()
    for pole in wzorzec.select_dtypes(include="category").columns:
        wzorzec[pole] = wzorzec[pole].astype("string")
    pd.testing.assert_frame_equal(
        wzorzec, odczyt, check_dtype=False,
        check_exact=False, rtol=1e-10, atol=1e-10
    )
assert sha256(plik.read_bytes()).hexdigest() == hash_wejscia
```

Ponowny odczyt sprawdza wartości, etykiety i braki, nie tylko liczbę kolumn. CSV nie przechowuje typów Pandas, dlatego kontrola wartości nie wymaga identycznego dtype. Kategorie w kopii wzorcowej porównujemy jako tekst. Tolerancja uwzględnia drobne różnice reprezentacji liczb, a nie merytoryczne rozbieżności. Przy kolejnym wykorzystaniu pliku typy odtworzysz według jego schematu.

Ponowne wykonanie zapisu zastępuje eksporty pod tymi samymi nazwami. Jeśli chcesz zachować wcześniejszy etap, zarchiwizuj go przed uruchomieniem komórek. Nie nadpisuj wejścia pomiary_clean.csv ani wyników poprzednich modułów.

Audyt jednej odpowiedzi

Wybierz jedną z odpowiedzi M3-P01–M3-P06. Wskaż pytanie i kontrakt, konkretny kod oraz wynik, kontrolę liczebności i zakresu, obserwację, interpretację i ograniczenie. Nie wykonujesz nowej analizy. Sprawdzasz, czy osoba czytająca notebook potrafi prześledzić uzasadnienie od pytania do werdyktu.

Gotowość do dalszej pracy

- Sześć pytań ma odpowiedzi, a co najmniej pięć obserwacji wskazuje konkretne wyniki kodu.
- Każda miara ma jednostkę, liczebność i właściwy zakres; procenty mają jawny mianownik.
- Porównania grup zaczynają się od pełnych danych, a nie od filtra przekroczeń.
- Korelacja ma oba warianty, wyniki dla urządzeń i ograniczenie przyczynowości.
- Eksporty odczytują się poprawnie; źródło i zatwierdzony wynik M2 pozostały bez zmian.
- Sugestie LLM mają weryfikację albo odnotowano brak użycia modelu.

Zapisz notebook, wykonaj Restart Kernel and Run All Cells, popraw pierwszy błąd i powtórz przebieg. Po udanym wykonaniu zapisz aktualne wyniki. Do M4 wybierz dwie obserwacje, które warto pokazać na wykresach. Przy każdej wskaż tabelę, jednostkę i ograniczenie, które powinno pozostać widoczne w opisie.

Odpowiedzi do sprawdzenia rozumienia

Pytania służą samokontroli, nie ocenie. Porównaj przede wszystkim uzasadnienie. Jeśli wynik Cię zaskoczył, wróć do przykładu i wskaż, które założenie albo ograniczenie zmienia jego sens.

1. Trzy najwyższe pojedyncze pomiary nie są rankingiem urządzeń według mediany. Właściwy wynik ma po jednym wierszu na urządzenie. Najpierw agregujesz wszystkie jego drgania, podajesz liczebność, następnie sortujesz osiem median i wybierasz pierwsze trzy. Kilka skrajnych odczytów tego samego urządzenia nie zastępuje takiego porównania.

2. Bliskie średnia i mediana nie wykluczają wpływu wartości skrajnej. W danych kursowych wyłączenie jednego maksimum zmienia odchylenie standardowe z około 0,205 do 0,154 mm/s, choć mediana prawie się nie zmienia. P90 opisuje pozycję w obserwowanym zbiorze. Do prognozy przyszłego pomiaru potrzebne byłyby dodatkowe założenia o procesie i jego stabilności.

3. Mniejszy licznik może odpowiadać większemu odsetkowi, jeśli mianownik jest odpowiednio mniejszy. Styczeń ma 15 przekroczeń z 496, a dostępna część lutego 13 z 224. To około 3,0% i 5,8%. Wniosek dotyczy dostępnych pomiarów w tych okresach, nie potwierdzonego trendu ani liczby awarii.

4. Podzbiór 28 przekroczeń pomija pomiary nieprzekraczające progu i pięć urządzeń. Jego mediany opisują wysokie drgania wybranej części danych. M3-P03 wymaga pełnych grup. Z kolei podobne mediany ogólne dwóch trybów nie wykluczają różnic wewnątrz typów; w rozbiciu kierunek różnicy energii zależy od typu.

5. Margines mediany powstaje z właściwych rekordów źródłowych. Mediana wszystkich 718 dostępnych energii to 14,235 kWh, podczas gdy średnia sześciu median to około 14,188 kWh. Dla normalize="index" mianownikiem jest liczba rekordów danego wiersza, czyli tutaj lokalizacji. To nie liczba wszystkich rekordów ani automatycznie liczba urządzeń.

6. Należy pokazać oba warianty i wyjaśnić różnicę. Wzrost r nie dowodzi, że pomiar jest błędem. Dodatni kierunek w każdym urządzeniu także nie ustala przyczyny zależności, nie wyklucza wpływu obciążenia lub czasu i nie potwierdza awarii. Rozstrzygnięcie wymaga innego rodzaju dowodów niż sam współczynnik.

Słownik pojęć i samodzielna kontynuacja

EDA to eksploracyjna analiza danych: badanie rozkładów, grup i zależności połączone z oceną ograniczeń. Kontrakt odpowiedzi określa zakres, pola, poziom analizy, miary, formę wyniku i kontrole. Hipoteza robocza jest przewidywaniem do sprawdzenia; obserwacja opisuje wynik, a interpretacja proponuje jego możliwe znaczenie.

Jednostka obserwacji mówi, co reprezentuje wiersz wejścia. Poziom agregacji mówi, co reprezentuje wiersz wyniku. Maska logiczna wybiera rekordy spełniające warunek, a agregacja oblicza miarę w grupie. Size to liczba rekordów, count to liczba dostępnych wartości konkretnego pola.

Mediana jest środkiem uporządkowanych danych. Kwartyle dzielą je według pozycji, a IQR mierzy szerokość środkowej połowy. Percentyl opisuje próg pozycyjny w zbiorze. Odchylenie standardowe mierzy rozproszenie wokół średniej. Wszystkie te miary czytaj razem z jednostką, liczebnością i zakresem danych.

Tabela przestawna agreguje wielkość liczbową w macierzy kategorii. Tabela krzyżowa domyślnie zlicza pary kategorii. Margines podsumowuje rekordy właściwe dla całego wiersza lub kolumny. Normalizacja zamienia liczebność w udział względem jawnej podstawy.

Korelacja Pearsona opisuje zależność liniową kompletnych par. Analiza wrażliwości sprawdza, jak wynik zmienia się po jawnie określonej zmianie zakresu. Zmienna zakłócająca może wpływać na porównanie albo współzmienność, choć nie jest głównym przedmiotem pytania. Współzmienność nie ustala mechanizmu przyczynowego.

Pytania do powrotu po module

Która kontrola najbardziej zmieniła Twoją pierwszą interpretację danych? Czy była to liczebność, mianownik, podział na grupy czy wariant bez maksimum? Wybierz konkretny przykład i opisz, co stwierdziłbyś bez tej kontroli oraz jak sformułujesz obserwację po jej wykonaniu.

Przed modułem 4 przejrzyj sześć odpowiedzi i wybierz te, dla których wykres rzeczywiście pomoże czytelnikowi. Zastanów się, czy chcesz pokazać rozkład, porównanie grup, różnicę okresów czy zależność. Zachowaj ustalone jednostki i ograniczenia. Zmiana sposobu prezentacji wyniku nie powinna zmieniać pytania, na które odpowiadasz.