KARTA PRACY UCZESTNIKA

Moduł 3 Eksploracyjna analiza danych

Od pytania do sprawdzonej obserwacji w Jupyter Notebook

Czas	Pracujesz z	Rezultat
70–90 min	danymi przygotowanymi w M2 i biblioteką Pandas	sześcioma odpowiedziami, tabelami i kontrolami w notebooku M3

Cel i rezultat modułu

W M2 sprawdziłeś jakość danych. Teraz zbadasz ich rozkłady, porównasz grupy i sprawdzisz zależności. Każdą odpowiedź powiążesz z kodem, kontrolą i ostrożną interpretacją. Tabele wykorzystasz ponownie w module 4.

Zanim zaczniesz

Kontynuuj projekt eksploatacja_urzadzen z M1. Utwórz notebooks/M3_EDA.ipynb. Notebook M2 pozostaw jako dokumentację przygotowania danych; M3 ma działać bez otwierania ani wykonywania jego komórek.

Potrzebny plik w projekcie	Do czego wracasz
data/processed/pomiary_clean.csv	Zweryfikowany wynik M2, tylko do odczytu w M3.
docs/slownik_danych.md	Nazwy, typy, jednostki i dopuszczone braki wyniku M2.
notebooks/M2_przygotowanie_danych.ipynb	Dziennik decyzji i kontrole zbioru wejściowego.

Pytania prowadzące przez analizę

Zachowaj identyfikatory w notebooku. Rozdział 1 planuje wszystkie pytania; rozdział 4 obejmuje dwa pytania.

ID i rozdział	Na co odpowiesz
M3-P01 · R2	Jakie są typowe wartości i zróżnicowanie temperatury, drgań, energii i produkcji?
M3-P02 · R3	Które pomiary przekraczają 1,5 mm/s i jak rozkładają się między urządzenia oraz miesiące?
M3-P03 · R4	Które trzy urządzenia mają najwyższą medianę drgań i ile pomiarów tworzy każdą medianę?
M3-P04 · R4	Jak różnią się mediany energii i produkcji między trybami pracy oraz typami urządzeń?
M3-P05 · R5	Jak różni się mediana energii w układzie rok_miesiac × lokalizacja?
M3-P06 · R6	Czy wyższej temperaturze towarzyszą większe wartości drgań?

Czas jest orientacyjny. Wyniki oblicz samodzielnie — karta nie jest kluczem odpowiedzi. Rozszerzenia wykonaj we własnym tempie.

Workflow pracy krok po kroku

Krok 1 Uruchom niezależny notebook M3

Sprawdź Path.cwd(). Poniższy kod zakłada katalog roboczy notebooks. Jeśli uruchomiłeś notebook z innego miejsca, popraw punkt startu. Nie wpisuj ścieżek zawierających dysk lub nazwę swojego konta.

```
from pathlib import Path
from hashlib import sha256
import pandas as pd
import matplotlib.pyplot as plt

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(), "Sprawdź katalog roboczy i wynik M2"
hash_wejscia = sha256(plik.read_bytes()).hexdigest()
pomiary_clean = pd.read_csv(
    plik, sep=",", encoding="utf-8",
    parse_dates=["czas", "data_instalacji"]
)
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)
```

CSV nie zachowuje typów. Po odczycie porównaj wszystkie 14 nazw i ich kolejność ze słownikiem M2, sprawdź dtypes oraz wartości kategorii. Upewnij się, że obie daty są typu datetime i nie powstały nowe braki.

Krok 2 Potwierdź kontrakt wejściowy

- Dla danych kursowych po M2 sprawdź 720 wierszy, 14 kolumn, 8 urządzeń oraz unikalny i kompletny klucz czas + urzadzenie.
- Potwierdź zakres od 1 stycznia do 14 lutego 2026 r., dwa okresy rok_miesiac i zgodność tej kolumny z czasem. Luty nie jest pełnym miesiącem.
- Sprawdź dwa braki energii, jeden brak produkcji i brak braków temperatury oraz drgań. Status ma tylko wartość OK — nie porównasz tu grup awarii.
- Porównaj SHA-256 pliku z wpisem M2 w README.md. Niezgodność wyjaśnij w M2; nie zmieniaj danych, aby przejść kontrolę w M3.

Krok 3 Zaplanuj odpowiedzi w rozdziale 1

Przed obliczeniami utwórz rejestr M3-P01–M3-P06. Przy każdym pytaniu zapisz: zakres danych i czasu, zmienne, jednostkę obserwacji, grupowanie, miarę z jednostką, formę wyniku i kontrolę. Jeden rekord wejściowy to pomiar jednego urządzenia w danym czasie, nie osobne urządzenie.

Wybierz jedną hipotezę roboczą i zapisz, jaki wynik mógłby jej przeczyć. Na końcu nadaj jej status: wsparta, niewsparta albo nierozstrzygnięta. W notebooku zachowaj kolejność: konfiguracja, kontrola wejścia, rejestr pytań, sześć odpowiedzi, podsumowanie i rejestr LLM.

Od opisu zbioru do wybranych przypadków

Rozdział 2 Statystyki opisowe dla M3 P01

Pracuj na pełnej ramce pomiary_clean. Dla temperatury [°C], drgań [mm/s], energii [kWh] i produkcji [sztuki na pomiar] przygotuj tabelę, w której jeden wiersz opisuje jedną zmienną.

```
miary = ["temperatura", "drgania", "energia", "produkcja"]
statystyki = pomiary_clean[miary].describe().T
statystyki["braki"] = pomiary_clean[miary].isna().sum()
statystyki["braki_proc"] = statystyki["braki"] / len(pomiary_clean) * 100
statystyki["IQR"] = statystyki["75%"] - statystyki["25%"]
```

- Dodaj jednostki oraz percentyle P10 i P90. Porównaj średnią z medianą, minimum i maksimum z kwartylami oraz odchylenie standardowe z IQR.
- Sprawdź count osobno dla każdej zmiennej. Brak energii nie oznacza braku całego rekordu; nie usuwaj tych wierszy z wszystkich analiz.
- Zapisz po jednej obserwacji dla każdej zmiennej. Wskaż konkretną wartość, jednostkę, liczebność i ograniczenie. Nie porównuj bezpośrednio zmienności wielkości w różnych jednostkach.

Sprawdź wrażliwość: porównaj n, średnią, medianę i odchylenie standardowe drgań w pełnym zbiorze oraz w kopii bez jednego wskazanego maksimum. Zachowaj oba wyniki. Wyłączenie diagnostyczne nie jest decyzją o usunięciu pomiaru ze źródła.

Punkt kontrolny: potrafisz wyjaśnić, co wnoszą mediana i kwartyle, czego nie pokazuje sama średnia.

Rozdział 3 Filtrowanie i liczebności dla M3 P02

Zapisz próg roboczy drgań 1,5 mm/s i wybierz pomiary ściśle powyżej niego. To kryterium ćwiczenia, a nie norma techniczna ani rozpoznanie awarii. Nie zastępuj go progiem diagnostycznym używanym w M2.

- Utwórz osobną kopię powyzej_progu przez .loc[warunek, kolumny].copy(). Zachowaj czas, urzadzenie, drgania i rok_miesiac. Nie nadpisuj pomiary_clean.
- Uporządkuj pomiary malejąco po drganiach; przy remisie rosnąco po czasie i identyfikatorze. Sprawdź warunek w każdym wybranym rekordzie.
- Policz przekroczenia dla urządzeń i miesięcy. Uwzględnij także grupy z zerem przekroczeń, np. przez reindex względem grup w pełnym zbiorze.
- Oblicz odsetek przekroczeń w każdej grupie: liczba przekroczeń / liczba dostępnych pomiarów drgań w tej samej grupie × 100%. Podaj licznik i mianownik.

Porównaj styczeń z okresem 1–14 lutego. Sama liczba przekroczeń nie uwzględnia różnej liczby pomiarów. Nawet porównanie odsetków w dwóch okresach nie wystarcza do stwierdzenia trendu.

Punkt kontrolny: sumy liczebności przekroczeń według urządzeń i miesięcy zgadzają się z długością podzbioru, a żaden udział nie przekracza 100%.

Zapisz odpowiedź w czterech warstwach

Wynik: „Obliczyłem … dla n = …, w jednostce …”. Obserwacja: „W tym zbiorze …”. Interpretacja: „Warto sprawdzić, czy …”. Ograniczenie: „Na tej podstawie nie można stwierdzić …”. Ten sam układ zastosuj do wszystkich sześciu pytań.

Porównania grup i okresów

Rozdział 4 Ranking urządzeń dla M3 P03

Wróć do pełnej ramki pomiary_clean, nie do podzbioru przekroczeń. Pogrupuj dane po urzadzenie i oblicz medianę drgań oraz liczbę ich dostępnych pomiarów. Posortuj mediany malejąco, przy remisie identyfikatory rosnąco, i wskaż pierwsze trzy urządzenia.

Punkt kontrolny: pełna tabela ma jeden wiersz na urządzenie. Ranking median nie jest rankingiem trzech najwyższych pojedynczych pomiarów. Zachowaj pełną tabelę, a nie tylko pierwsze trzy wiersze.

Rozdział 4 Tryby i typy urządzeń dla M3 P04

Porównaj energię i produkcję na trzech poziomach: tryb, typ_urzadzenia oraz para typ_urzadzenia + tryb. Dla każdej grupy podaj liczbę rekordów, mediany i oddzielne liczby dostępnych wartości obu miar.

```
wedlug_trybu = pomiary_clean.groupby(
    "tryb", observed=True, dropna=False, as_index=False
).agg(
    rekordy=("urzadzenie", "size"),
    energia_n=("energia", "count"), energia_mediana=("energia", "median"),
    produkcja_n=("produkcja", "count"),
    produkcja_mediana=("produkcja", "median")
)
```

Dostosuj przykład do dwóch pozostałych poziomów. size liczy rekordy, count pomija braki wskazanego pola. W każdej tabeli suma rekordów ma odtwarzać wejście; sumy energia_n i produkcja_n porównaj z count pełnych kolumn. Nie sumuj median.

Zastanów się: czy różnicę między trybami możesz przypisać samemu trybowi pracy? Uwzględnij typ i skład urządzeń, moc nominalną oraz brak danych o obciążeniu. Mediana energii nie jest miarą efektywności energetycznej.

Rozdział 5 Tabela przestawna dla M3 P05

Z pełnych pomiarów zbuduj pivot_table: wiersze rok_miesiac, kolumny lokalizacja, wartości energia, agregacja median. Uporządkuj okresy. Jedna komórka opisuje medianę energii pomiarów z danego okresu i lokalizacji.

- Odtwórz te same wartości przez groupby().median().unstack() i porównaj wyniki. Sprawdź dwa okresy, trzy lokalizacje i sześć obserwowanych kombinacji.
- Obok median pokaż dwie tabele: liczbę rekordów oraz count energii. Ich różnica ujawnia braki. Pustej mediany nie wypełniaj zerem.
- Dodaj osobny wariant z margins=True i margins_name="Razem". Margines mediany liczy się ponownie z rekordów, nie jako suma ani średnia widocznych median.
- Porównaj lokalizacje w okresie i różnice między okresami w lokalizacji. W ograniczeniu uwzględnij niepełny luty oraz odmienny skład urządzeń.

Uzupełnij kontekst tabelą pd.crosstab lokalizacja × typ_urzadzenia: najpierw liczebności, potem udziały z normalize="index". To udziały rekordów w lokalizacji, nie bezpośrednio liczby urządzeń. Sprawdź sumę liczebności i sumy wierszy udziałów przed zaokrągleniem.

Punkt kontrolny: umiesz powiedzieć jednym zdaniem, co oznacza komórka każdej tabeli i z ilu obserwacji powstała.

Zależności i granice interpretacji

Rozdział 6 Korelacja dla M3 P06

Ponownie zacznij od pełnych pomiarów. Wybierz czas, urzadzenie, temperatura i drgania. W osobnej ramce pary zachowaj kompletne pary temperatura–drgania z tych samych wierszy. Podaj liczbę par oraz liczbę pominiętych rekordów.

- Przygotuj diagnostyczny wykres punktowy: temperatura na osi X, drgania na osi Y, jednostki przy osiach. Wskaż maksimum drgań i opisz kierunek, kształt, rozproszenie oraz nietypowy punkt.
- Oblicz współczynnik Pearsona dla pełnych par. Podaj r i n razem. Współczynnik nie ma jednostki; wynik bliski zeru nie wyklucza zależności nieliniowej.
- W wariancie diagnostycznym wyłącz jeden wskazany pomiar maksimum, a nie wszystkie przekroczenia progu. Porównaj oba wyniki i obie liczebności.

```
r_pelny = pary["temperatura"].corr(pary["drgania"], method="pearson")
idx_punktu = pary["drgania"].idxmax()
pary_wariant = pary.drop(index=idx_punktu).copy()
r_wariant = pary_wariant["temperatura"].corr(
    pary_wariant["drgania"], method="pearson"
)
print({"pelny": (len(pary), r_pelny),
       "wariant": (len(pary_wariant), r_wariant)})
```

Przed tym fragmentem utwórz pary zgodnie z poleceniem. Wskaż w notebooku klucz czas + urzadzenie pominiętego rekordu. Pomiar 4,850 mm/s pozostaje w danych wejściowych; wariant służy wyłącznie sprawdzeniu wrażliwości.

- Oblicz korelację i n osobno dla każdego urządzenia. Porównaj je z wynikiem zbiorczym: jedna liczba miesza różnice między urządzeniami ze zmianami wewnątrz nich.
- Sprawdź kompletność par, zakres r od −1 do 1 i ewentualne NaN. Brak wartości współczynnika może oznaczać zbyt małą próbę lub stałą zmienną; nie zamieniaj go na zero.
- Zapisz cztery warstwy odpowiedzi i werdykt hipotezy. Nie wyprowadzaj z korelacji przyczyny zmian, diagnozy awarii ani zaleceń naprawczych.

Punkt kontrolny: pokazujesz oba warianty nawet wtedy, gdy jeden daje większą korelację. Pomiary powtarzają się dla tych samych urządzeń — nie traktuj ich jak tylu niezależnych urządzeń.

Sprawdź własną odpowiedź

Wybierz jedną odpowiedź M3-P01–M3-P06 i zapisz krótki audyt: pytanie, wynik kodu, kontrola, obserwacja, interpretacja i ograniczenie. Wskaż konkretną tabelę lub komórkę notebooka, aby inna osoba mogła odtworzyć Twoje rozumowanie.

Utrwal umiejętności z prezentacji

W osobnych sekcjach wykonaj ćwiczenia: R3 — temperatura > 64 °C, ranking i udziały; R4 — mediana temperatury oraz size i count według lokalizacji; R5 — mediana produkcji i count z marginesami; R6 — korelacja energii i produkcji dla kompletnych par, ogółem i według urządzeń. Przy każdym zapisz kontrolę, obserwację i ograniczenie. Rozłóż tę pracę na krótsze sesje.

Dla chętnych: porównaj współczynniki Pearsona i Spearmana. Sprawdź, czy zmienia się interpretacja, a nie tylko wartość liczby.

Zapis rezultatów i samokontrola

Nazewnictwo plików

Zachowaj strukturę M1 i M2: outputs/tables zamiast output/tables ze slajdów i notatek. Poniżej nazwy kursowe z prefiksami pytań. Identyfikator M3-P01 w tekście odpowiada M3_P01 w pliku. Nie twórz dwóch folderów wynikowych ani duplikatów pod obiema nazwami.

Zapisz w outputs/tables	Zawartość
M3_P01_statystyki_opisowe.csv	Odpowiednik statystyki_opisowe.csv ze slajdów 31 i 50; dodaj jednostki i braki.
M3_P02_przekroczenia_urzadzenia.csv M3_P02_przekroczenia_miesiace.csv	Licznik, mianownik i odsetek przekroczeń w grupie.
M3_P03_ranking_drgan.csv	Pełny ranking urządzeń: mediana i liczebność.
M3_P04_tryb.csv M3_P04_typ.csv M3_P04_typ_tryb.csv	Trzy poziomy agregacji energii i produkcji z liczebnościami.
M3_P05_energia_miesiac_lokalizacja.csv	Odpowiednik energia_miesiac_lokalizacja.csv ze slajdu 124, w układzie długim z liczebnościami.
M3_P06_korelacje.csv	Zakres analizy, urządzenie (jeśli dotyczy), wariant, metoda, n i r.

Eksport P05 ma sześć wierszy i kolumny: rok_miesiac, lokalizacja, energia_mediana, rekordy, energia_n. Macierze i kontrole pozostaw w notebooks/M3_EDA.ipynb. Eksport P06 obejmuje pełny zbiór, wariant diagnostyczny i urządzenia. Utwórz FIGURES_DIR i zapisz wykres jako outputs/figures/M3_P06_temperatura_drgania.png.

Eksport który zachowuje znaczenie tabeli

Zapisuj tabele w postaci wierszowej, z nazwami grup jako kolumnami. Jeśli nazwa zmiennej lub grupy jest w indeksie, najpierw przenieś ją przez reset_index(). Poniżej przykład dla tabeli statystyki po uzupełnieniu jej o jednostki i percentyle.

```
TABLES_DIR.mkdir(parents=True, exist_ok=True)
do_zapisu = statystyki.rename_axis("zmienna").reset_index()
plik_tabeli = TABLES_DIR / "M3_P01_statystyki_opisowe.csv"
do_zapisu.to_csv(plik_tabeli, index=False, encoding="utf-8")
odczyt = pd.read_csv(plik_tabeli, encoding="utf-8")
pd.testing.assert_frame_equal(
    do_zapisu, odczyt, check_dtype=False, rtol=1e-10, atol=1e-10
)
assert sha256(plik.read_bytes()).hexdigest() == hash_wejscia
```

Analogicznie sprawdź inne eksporty: nazwy kolumn, klucze grup, braki i wartości. W README.md dopisz notebook M3, sposób uruchomienia i listę wyników. Nie nadpisuj pomiary_clean.csv ani rezultatów wcześniejszych modułów.

Domknięcie pracy i przygotowanie do M4

LLM jako partner do sprawdzania hipotez

Jeśli korzystasz z LLM, przekaż minimalny kontekst bez danych poufnych. Poproś: „Zaproponuj trzy alternatywne hipotezy dotyczące związku temperatury i drgań. Dla każdej wskaż kontrolę lub potrzebne dane. Nie wymyślaj wyników ani nie przedstawiaj hipotez jako przyczyn”.

Sprawdź co najmniej jedną sugestię kodem lub wskaż brakujące dane. Zapisz cel, prompt, odpowiedź, test, werdykt i uzasadnienie w docs/rejestr_uzycia_LLM.csv, z modul = M3. Oceniaj fragmenty osobno: przyjmij, zmodyfikuj lub odrzuć. Jeśli nie używasz LLM, odnotuj to w notebooku.

Lista kontroli przed zakończeniem modułu

- □ Sześć pytań ma odpowiedzi powiązane z kodem, kontrolą i ograniczeniem.
- □ Wyniki zawierają jednostki, właściwe liczebności i jawne mianowniki.
- □ Porównania P03–P06 zaczynają się od pełnego zbioru, nie od filtra P02.
- □ Zapisano oba warianty korelacji, wyniki dla urządzeń i werdykt hipotezy.
- □ CSV odczytują się poprawnie, a dane wejściowe pozostały bez zmian.
- □ Rejestr LLM zawiera weryfikację sugestii albo odnotowano brak użycia LLM.
- □ Restart Kernel and Run All Cells działa; notebook zapisano po przebiegu.

Gdy pojawi się problem

Objaw	Najpierw sprawdź
FileNotFoundError	Path.cwd(), plik.is_file() i obecność wyniku M2 w data/processed.
NameError po restarcie	Czy import, zmienna lub ramka pary powstają przed użyciem.
Niespodziewanie mała liczba pomiarów	Czy analizujesz pełną ramkę, a nie filtr z P02; czy count pomija braki.
Błąd pivot lub puste komórki	Czy potrzebna jest agregacja pivot_table; jakie kombinacje i wartości rzeczywiście istnieją.
NaN zamiast współczynnika	Liczbę kompletnych par i zmienność obu kolumn w danej grupie.
Brak nazw grup w CSV	Czy etykiety pozostały w indeksie pominiętym przy index=False.

Twój następny krok

Dokończ: „Najbardziej zmieniła moje rozumienie danych kontrola …, ponieważ …”. Wskaż dwie obserwacje, które warto zilustrować w M4, i dobierz do nich sprawdzone tabele. Nie musisz jeszcze tworzyć finalnych wykresów — ich przygotowanie jest zadaniem następnego modułu.

Zasada końcowa: zapisz notebook, wykonaj Restart Kernel and Run All Cells, popraw pierwszy ewentualny błąd i powtórz przebieg. Po udanym wykonaniu zapisz notebook ponownie, aby zachować aktualne wyniki.