Projekt EDA - raport

Jakub Milasz

28 kwietnia 2025

Spis treści

Seaborn

Seaborn to biblioteka oparta na matplotlib — bibliotece Pythona służącej do wizualizacji danych. Dzięki szerokiej gamie typów wykresów, takich jak wykresy punktowe, histogramy, mapy ciepła czy wykresy skrzypcowe, pozwala przedstawić informacje w zrozumiały i atrakcyjny sposób.

Źródło danych

Dane do analizy pobrano ze strony `https://dane.gov.pl/`. Jest to państwowy portal otwartych danych, który umożliwia każdemu zainteresowanemu bezpłatny dostęp do informacji publicznej z różnych kategorii. Dostępne tam filtry ułatwiają skuteczne wyszukiwanie odpowiednich zbiorów danych (np. kategoria, dostawca, format, poziom otwartości danych).
Dane są dostępne w formatach: CSV, JSON, XLS lub przez różne API (np. REST API).

Czego dotyczą dane?

Analizie poddano dane z badań dźwignicy. Dane zostały zaproponowane przez prowadzącego, a ich poziom otwartości oceniono na 4 w skali 5-stopniowej. Te dwa aspekty wpłynęły na wybór tych danych.

Rys. 1 - Wygląd danych

	Nr. Pomiaru	Data	Ciężar ładunku [T]	Długość wysięgnika [m]	Odległość od osi [m]	Wysokość podnoszenia [m]	Maksymalne, chwilowe zużycie ON [l/h]	Dzienne zużycie ON [l/8h]	Cena hurtowa ON 1000l [PLN]	Maksymalne, chwilowe zużycie energii elektrycznej [kW]	Dzienne zużycie energii elektrycznej [kW/8h]	Cena energii elektrycznej [kWh]	Koszt dzienny [PLN]	Prędkość wiatru [km/h]	Prędkość wiatru [m/s]	Temperatura [C]	Ciśnienie [hPa]
0	1	11/4/2019	5	13.2	2	8	15.0	48.0	NaN	NaN	NaN	NaN	NaN	10.25	2.85	10.25	995.00
1	2	11/4/2019	8	13.2	2	8	15.0	48.0	NaN	NaN	NaN	NaN	NaN	10.25	2.85	10.25	995.00
2	3	11/4/2019	15	13.2	2	8	15.0	48.0	NaN	NaN	NaN	NaN	NaN	10.25	2.85	10.25	995.00
3	4	11/5/2019	5	13.2	1	10	14.9	47.2	4241.0	NaN	NaN	NaN	200.1752	7.00	1.94	8.00	999.50
4	5	11/5/2019	8	13.2	1	10	14.9	47.2	4241.0	NaN	NaN	NaN	200.1752	7.00	1.94	8.00	999.50
5	6	11/5/2019	15	13.2	1	10	14.9	47.2	4241.0	NaN	NaN	NaN	200.1752	7.00	1.94	8.00	999.50
6	7	11/6/2019	5	17.7	3	9	15.1	48.0	4244.0	NaN	NaN	NaN	203.7120	13.66	3.79	6.75	1004.25
7	8	11/6/2019	8	17.7	3	9	15.1	48.0	4244.0	NaN	NaN	NaN	203.7120	13.66	3.79	6.75	1004.25
8	9	11/6/2019	15	17.7	3	9	15.1	48.0	4244.0	NaN	NaN	NaN	203.7120	13.66	3.79	6.75	1004.25
9	10	11/7/2019	5	17.7	8	11	15.2	48.8	4245.0	NaN	NaN	NaN	207.1560	6.50	1.81	6.00	1008.50

Opis danych

Dane pochodzą z testów modernizowanego układu zasilania elektrycznego w dźwignicy żurawia. Celem modernizacji było zastąpienie silnika Diesla alternatywnym źródłem energii i analiza wpływu na:

- parametry pracy dźwignicy,
- koszty zużycia energii i paliw,
- efektywność w różnych warunkach pogodowych.

Opis kolumn — część 1

- **Nr pomiaru** – numer pomiaru,
- **Data** – data wykonania pomiaru,
- **Ciężar ładunku [T]** – ciężar w tonach,
- **Długość wysięgnika [m]** – długość w metrach,
- **Odległość od osi [m]** – odległość w metrach,
- **Wysokość podnoszenia [m]** – wysokość podnoszenia w metrach,
- **Maksymalne chwilowe zużycie ON [l/h]** – chwilowe zużycie oleju napędowego,
- **Dzienne zużycie ON [l/8h]** – zużycie paliwa w 8 godzin,
- **Cena hurtowa ON 1000l [PLN]** – cena hurtowa za 1000 litrów paliwa.

Opis kolumn — część 2

- **Maksymalne chwilowe zużycie energii elektrycznej [kW]** - maksymalne, chwilowe zużycie energii elektrycznej wyrażone w kilowatach,
- **Dzienne zużycie energii elektrycznej [kW/8h]** - dzienne (czas pracy 8 godzin) zużycie energii elektrycznej wyrażone w kilowatach,
- **Cena energii elektrycznej [PLN/kWh]** - cena energii elektrycznej wyrażona w cenie w polskim złotym za kilowatogodzinę,
- **Koszt dzienny [PLN]** - dzienny koszt użycia dźwignicy wyrażony w polskim złotym,
- **Prędkość wiatru [km/h] oraz [m/s]** - prędkość wiatru,
- **Temperatura [°C]** - temperatura otoczenia wyrażona w stopniach celsjusza,
- **Ciśnienie [hPa]** - ciśnienie atmosferyczne wyrażone w hektopaskalach.

Potencjalne zastosowania w ML

Możliwość wykorzystania danych w zadaniach klasyfikacyjnych (uczenie nadzorowane), np.:

- Klasyfikacja dziennego zużycia paliwa lub energii,
- Optymalizacja kosztów operacyjnych dźwignicy.

Problemy z danymi

Ze względu na 2 rodzaje silników w dźwignicach, napotkano następujące problemy:

- Brak danych dotyczących zużycia i cen energii dla roku 2019,
- Brak danych o paliwie od roku 2022 wzwyż.

Podzielono dane na dwa podzbiory odpowiednio do rodzaju braków. Pominięto również kolumnę z prędkością wiatru w km/h, ponieważ jest przeskalowaną wartością prędkości w m/s.

Braki danych

Przeanalizowano dane dotyczące dźwignicy z silnikiem spalinowym.

Rys. 2 - Braki danych dla dźwignicy z silnikiem spalinowym

Nr. Pomiaru	0.00000
Data	0.00000
Ciężar ładunku [T]	0.00000
Długość wysięgnika [m]	0.00000
Odległość od osi [m]	0.00000
Wysokość podnoszenia [m]	0.00000
Maksymalne, chwilowe zużycie ON [l/h]	0.00000
Dzienne zużycie ON [l/8h]	0.00000
Cena hurtowa ON 1000l [PLN]	0.25641
Maksymalne, chwilowe zużycie energii elektrycznej [kW]	1.00000
Dzienne zużycie energii elektrycznej [kW/8h]	1.00000
Cena energii elektrycznej [kWh]	1.00000
Koszt dzienny [PLN]	0.25641
Prędkość wiatru [km/h]	0.00000
Prędkość wiatru [m/s]	0.00000
Temperatura [C]	0.00000
Ciśnienie [hPa]	0.00000

dtype: float64

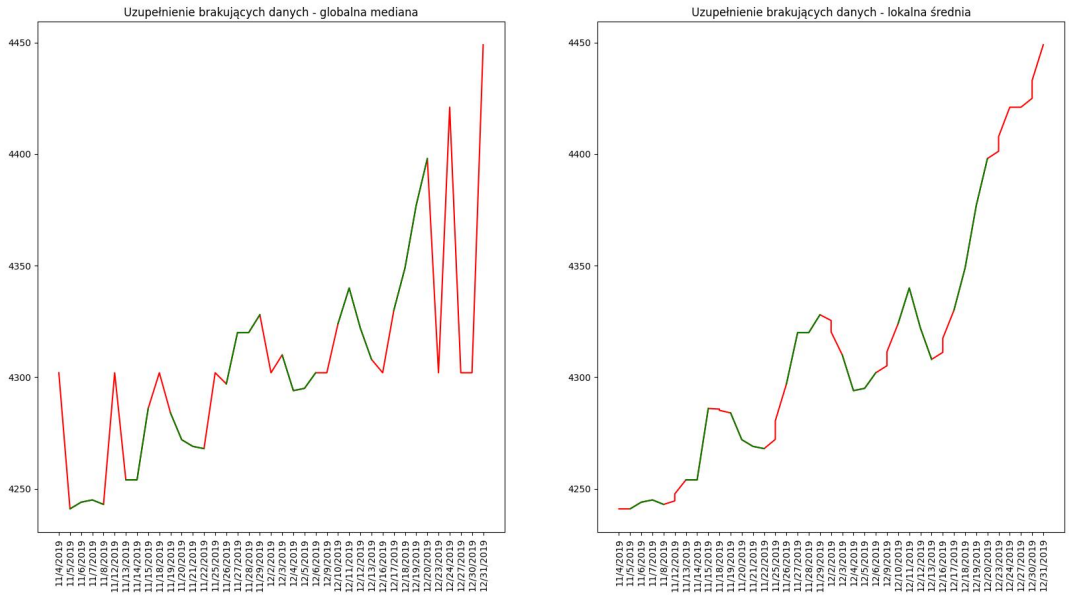
Widzimy brak 100% danych dla zużycia i cen energii elektrycznej oraz 26% braków dla ceny hurtowej ON. Z tych braków wynika również brak kosztu dziennego.

Uzupełnianie braków

Przetestowano dwa sposoby uzupełnienia brakujących danych:

- globalna mediana,
- lokalna średnia.

Lokalna średnia lepiej oddaje trend danych, szczególnie na końcach przedziału, co zostało zwizualizowane na wykresach w następnym slajdzie.

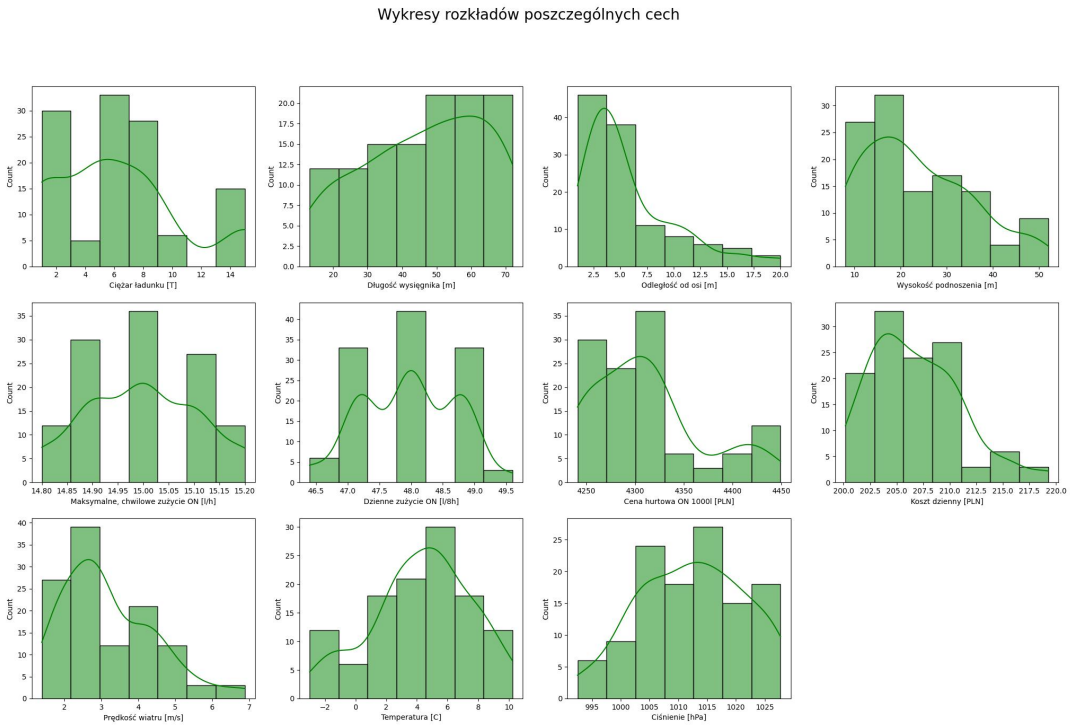


Rys. 3 - Porównanie metod uzupełniania braków

Wybrano uzupełnienie lokalną średnią. Brakujący koszt dzienny wyliczono jako:

`zużycie paliwa × cena hurtowa / 1000`.

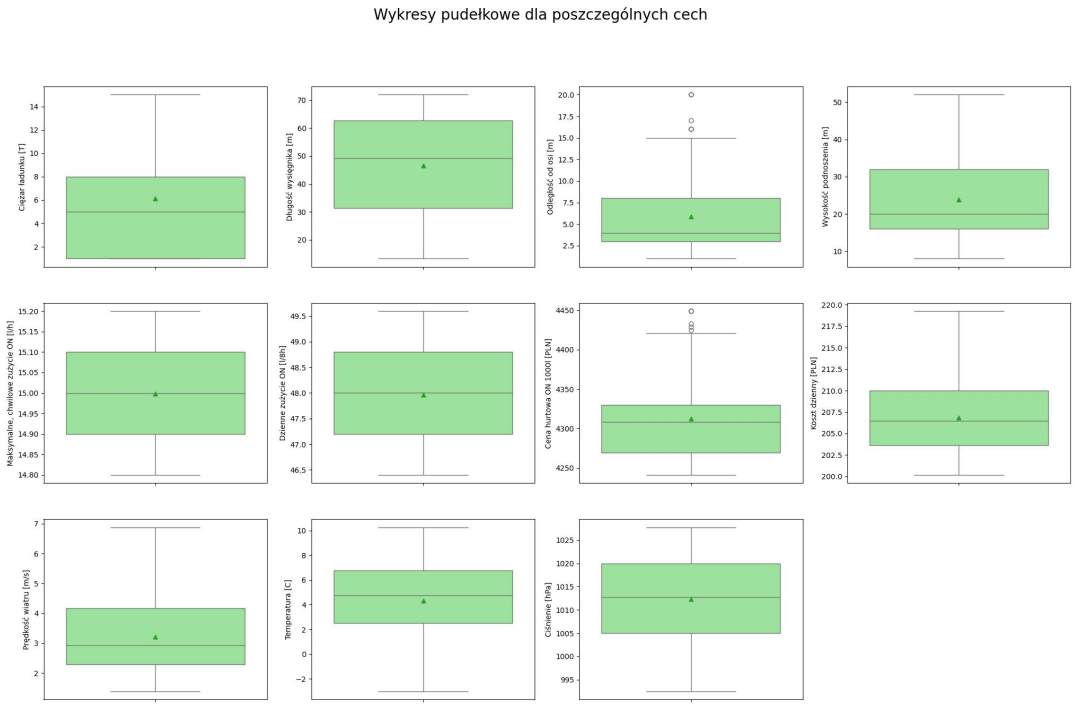
Rozkład cech



Rys. 4 - Histogramy wybranych cech

Dzięki histogramom można lepiej zrozumieć dane poprzez zobaczenie typowych wartości zmiennych. Ilość histogramów dobrano na podstawie reguły Sturge'a. Rozkład cech odbiega od rozkładu normalnego. Najbliższe są: temperatura i ciśnienie.

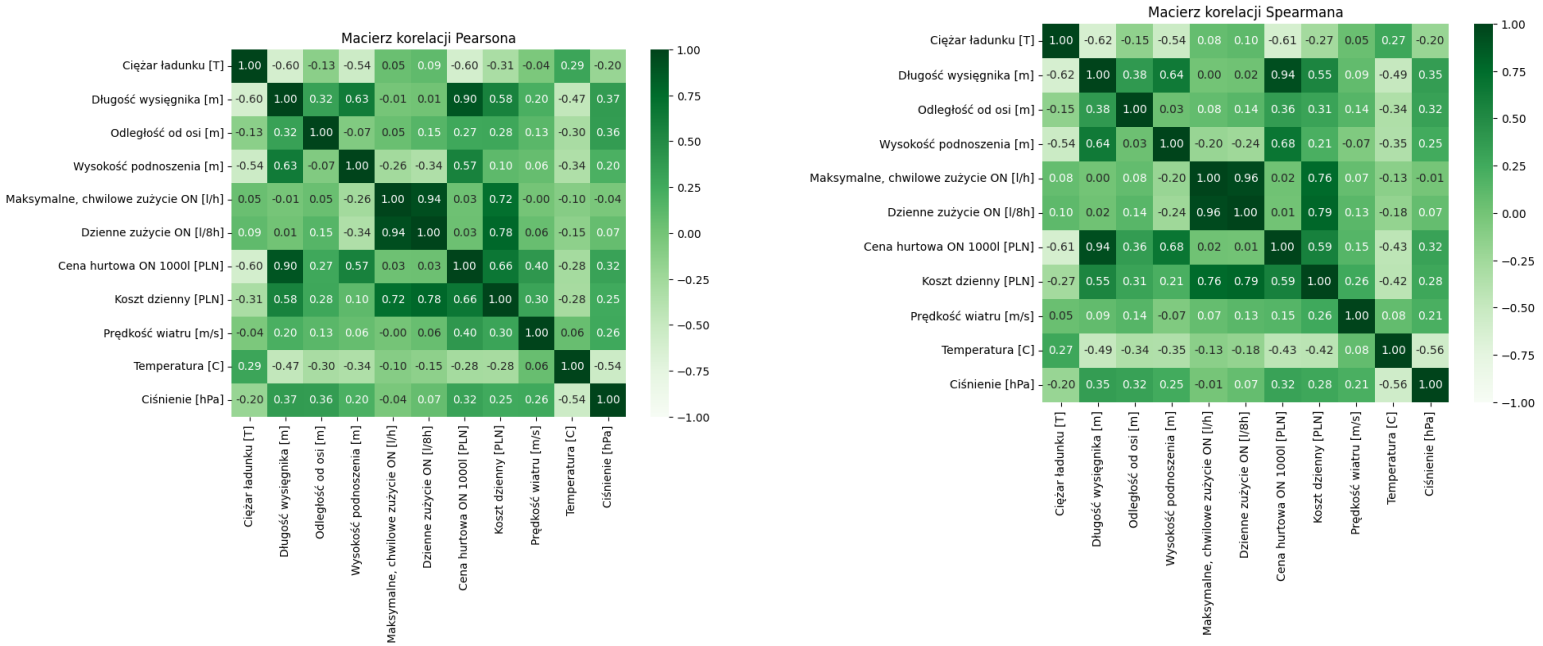
Anomalie



Rys. 5 - Wykresy pudełkowe wybranych cech

Wykresy pudełkowe pomagają w detekcji anomalii. Zaobserwowano je dla ceny hurtowej oraz odległości od osi.

Macierze korelacji



Rys. 6 - Macierze korelacji

Macierze korelacji pozwala zauważyć zależności między zmiennymi. Należy wziąć pod uwagę, że niektóre relacje w takim badaniu mogą być przypadkowe, np. silna korelacja cenu hurtowej paliwa z długością wysięgnika.

Zmienna target

Utworzono nową zmienną docelową — klasy zużycia paliwa:

- Małe: $[46.4, 47.5)$,
- Średnie: $[47.5, 48.5)$,
- Duże: $[48.5, 49.6]$.

Wybrano ją, ponieważ może ona pomóc w optymalizacji zużycia paliwa oraz kosztów.

Wybór cech

Wybrane cechy techniczne:

- Ciężar ładunku - im cięższy ładunek, tym więcej energii będzie potrzebowała maszyna,
- Długość wysięgnika - dłuższy wysięgnik powoduje większy moment siły, tym maszyna zużywa więcej energii, by utrzymać stabilność i bezpiecznie podnieść ładunek,
- Odległość od osi - im dalej od osi podnosisz ciężar, tym większe obciążenie konstrukcji i silników oraz rośnie zapotrzebowanie na energię,
- Wysokość podnoszenia - wyższe podnoszenie = więcej pracy do wykonania = większe zużycie energii (bo trzeba pokonać grawitację na większej odległości).

Selekcja cech — dodatkowe

Wybrane cechy środowiskowe:

- Prędkość wiatru - przy wyższej prędkości wiatru maszyna będzie wymagać więcej energii do stabilizacji ładunku,
- Temperatura - bardzo niskie temperatury wpływają na sprawność maszyny, co może powodować większe zużycie paliwa,
- Ciśnienie - powiązane z temperaturą (korelacja Pearsona $r = -0{,}54$), ale warto sprawdzić.

Uzasadnienie wyboru cech

Wybrano cechy, na które człowiek może mieć wpływ lub które mogą być użyteczne przy planowaniu pracy (np. odpowiednia pora dnia). Pomięto cechy związane bezpośrednio z kosztami paliwa. Może być też potrzeba utworzenia nowej zmiennej, jednak to zadanie najlepiej oddać osobom mającym wiedzę techniczną z zakresu pracy dźwignicy.

Techniki określające ważność cech

- Korelacja zmiennych niezależnych,
- Silne powiązanie ze zmienną zależną — zachować,
- Redundancja zmiennych — eliminacja jednej z pary,
- Uwzględnienie wniosków wynikających z wykresów, np. usunięcie anomalii.

Analiza zużycia energii — część 1

Analogiczna procedura jak w przypadku paliwa. Analizując dane dotyczące zużycia energii można zauważyć, że różnice w dziennym zużyciu są niewielkie, zatem cechy mają mniejszy wpływ na koszt pracy maszyny. Ponadto koszt dzienny użytkowania jest dużo mniejszy niż w przypadku zasilania olejem napędowym.

Analiza zużycia energii — część 2

Przy analizie zużycia energii elektrycznej warto zwrócić uwagę na:

- Cenę energii - stała dla całości danych,
- Wybór przedziałów do klasyfikacji,
- Obliczanie brakującego kosztu dziennego.