



AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W
KRAKOWIE

WYDZIAŁ GEOLOGII, GEOFIZYKI I OCHRONY ŚRODOWISKA



Projekt EDA – raport

Feature Engineering i EDA na danych z eksploatacji dźwignicy

Tomasz Drozd

Kierunek: Inżynieria i Analiza Danych, III rok

Data oddania: 06.05.2025

Spis treści

1	Cel projektu	2
2	Opis danych	2
2.1	Zastosowanie danych w kontekście ML	3
3	Feature Engineering (Inżynieria cech)	3
3.1	Wstępna eksploracja danych	3
3.2	Usunięcie niepotrzebnych kolumn	4
3.3	Uzupełnienie braków danych	5
4	Exploratory Data Analysis (Eksploracyjna analiza danych)	6
4.1	Analiza map korelacji dla obu typów napędów	6
4.2	Rozkład cech dla silnika spalinowego	8
4.3	Wartości odstające	8
4.4	Analiza zależności poszczególnych zmiennych między sobą	9
4.5	Analiza trendów czasowych	11
4.6	Analiza porównawcza napędów	12
5	Wybór zmiennej docelowej (TARGET)	13
6	Wybór zmiennych wejściowych (FEATURES)	14
7	Wnioski	14

1. Cel projektu

Celem projektu jest przeprowadzenie wstępnej analizy danych pochodzących z pomiarów eksploatacyjnych dźwignicy. Analiza obejmuje inżynierię cech (Feature Engineering) oraz eksploracyjną analizę danych (Exploratory Data Analysis — EDA), której celem jest identyfikacja zależności pomiędzy zmiennymi, oczyszczenie i przygotowanie danych do dalszego modelowania oraz ocena ich jakości.

W ramach projektu zostanie również określona możliwość zastosowania danych w kontekście uczenia nadzorowanego. Przeanalizowane zostaną korelacje pomiędzy zmiennymi oraz wskazane zostaną potencjalne zmienne celu (target) i zmienne objaśniające (features), które mogłyby posłużyć do budowy modeli predykcyjnych.

2. Opis danych

Do wykonania projektu użyto danych ze strony <https://dane.gov.pl/>. Strona ta pozwala każdemu przeglądać dane publiczne z różnych dziedzin i kategorii. Analizie poddano dane z badań dźwignicy. Zbiór ten został wybrany, ze względu na trudność w znalezieniu innego odpowiedniego zestawu danych, a zestaw dotyczący dźwignicy był jednym z dwóch zaproponowanych przez prowadzącego.

Zbiór danych zawiera następujące kolumny:

- **Nr. Pomiaru** – numer pomiaru
- **Data** – data wykonania pomiaru,
- **Ciężar ładunku [T]** – masa podnoszonego ładunku w tonach,
- **Długość wysięgnika [m]** – długość wysięgnika dźwignicy w metrach,
- **Odległość od osi [m]** – odległość ładunku od osi obrotu,
- **Wysokość podnoszenia [m]** – wysokość, na jaką podnoszony jest ładunek,
- **Maksymalne, chwilowe zużycie ON [l/h]** – chwilowe maksymalne zużycie oleju napędowego (ON) w litrach na godzinę,
- **Dzienne zużycie ON [l/8h]** – zużycie ON podczas 8 godzin pracy,
- **Cena hurtowa ON 1000l [PLN]** – cena hurtowa oleju napędowego za 1000 litrów,
- **Maksymalne, chwilowe zużycie energii elektrycznej [kW]** – chwilowe maksymalne zużycie energii elektrycznej w kilowatach,
- **Dzienne zużycie energii elektrycznej [kW/8h]** – zużycie energii podczas 8 godzin pracy,
- **Cena energii elektrycznej [kWh]** – jednostkowa cena energii elektrycznej,

- **Koszt dzienny [PLN]** – szacowany dzienny koszt eksploatacji,
- **Prędkość wiatru [km/h]** – prędkość wiatru w kilometrach na godzinę,
- **Prędkość wiatru [m/s]** – prędkość wiatru w metrach na sekundę,
- **Temperatura [°C]** – temperatura otoczenia,
- **Ciśnienie [hPa]** – ciśnienie atmosferyczne.

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,	Dzienne	Cena energii elektrycznej [kWh]	Koszt dzienny [PLN]	Prędkość wiatru [km/h]	Prędkość wiatru [m/s]	Temperatura [C]	Ciśnienie [hPa]
										chwilowe zużycie energii elektrycznej [kW]	zużycie energii elektrycznej [kW/8h]						
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
10	11	11/7/2019	8	17.7	8	11	15.2	48.8	4245.0	NaN	NaN	NaN	207.1560	6.50	1.81	6.00	1008.50
11	12	11/7/2019	15	17.7	8	11	15.2	48.8	4245.0	NaN	NaN	NaN	207.1560	6.50	1.81	6.00	1008.50

Rysunek 1: Wygląd pierwszych 12 wierszy danych

2.1. Zastosowanie danych w kontekście ML

Dane z projektu nadają się do uczenia nadzorowanego, np. do budowy modelu regresyjnego przewidującego koszt dzienny lub zużycie paliwa/energii na podstawie warunków operacyjnych (ciężar ładunku, długość wysięgnika, temperatura itd.). Modele ML mogą pomóc w optymalizacji kosztów i efektywniejszym planowaniu pracy dźwignic.

3. Feature Engineering (Inżynieria cech)

3.1. Wstępna eksploracja danych

Na początku zostały sprawdzone podstawowe statystyki opisowe dotyczące danych. Ze względu na charakterystykę (brak informacji dotyczących kolumn związanych z energią elektryczną dla pierwszych 117 wierszy - 2019 rok oraz analogiczny brak informacji dotyczących zużycia i ceny ON dla wierszy od 118 wzwyż - od 2022 roku), zestaw został podzielony na dwa: jedna część uwzględnia statystyki związane z silnikami spalinowymi, a druga część związana jest z energią elektryczną.

Liczba rekordów - napęd spalinowy: 117
Liczba rekordów - napęd elektryczny: 831

Nr. Pomiaru	Data	Ciepota 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]	Koszt dzienny [PLN]	Prędkość wiatru [km/h]	Prędkość wiatru [m/s]	Temperatura [C]	Ciepota [hPa]
0	1	11/4/2019	5	13.2	2	8	15.0	48.0	NaN	NaN	10.25	2.85	10.25
1	2	11/4/2019	8	13.2	2	8	15.0	48.0	NaN	NaN	10.25	2.85	10.25
2	3	11/4/2019	15	13.2	2	8	15.0	48.0	NaN	NaN	10.25	2.85	10.25
3	4	11/5/2019	5	13.2	1	10	14.9	47.2	4241.0	200.1752	7.00	1.94	8.00
4	5	11/5/2019	8	13.2	1	10	14.9	47.2	4241.0	200.1752	7.00	1.94	8.00
...
112	113	12/30/2019	1	72.0	9	37	15.0	48.0	NaN	NaN	21.50	5.97	3.75
113	114	12/30/2019	1	72.0	10	38	15.0	48.0	NaN	NaN	21.50	5.97	3.75
114	115	12/31/2019	1	72.0	5	40	14.9	47.2	4449.0	209.9928	24.75	6.88	3.00
115	116	12/31/2019	1	72.0	5	45	14.9	47.2	4449.0	209.9928	24.75	6.88	3.00
116	117	12/31/2019	1	72.0	5	50	14.9	47.2	4449.0	209.9928	24.75	6.88	3.00

117 rows x 14 columns

Nr. Pomiaru	Data	Ciepota ładunku [T]	Długość wysięgnika [m]	Odległość od osi [m]	Wysokość podnoszenia [m]	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]	Ciepota [hPa]
117	118	1/3/2022	5	13.2	2	3	39.6	79.2	0.4843	38.35656	34.0	9.44	8.5
118	119	1/3/2022	8	13.2	2	3	39.6	79.2	0.4843	38.35656	34.0	9.44	8.5
119	120	1/3/2022	15	13.2	1	3	39.6	79.2	0.4843	38.35656	34.0	9.44	8.5
120	121	1/4/2022	5	13.2	1	10	39.8	80.0	0.4843	38.74400	37.5	10.42	8.5
121	122	1/4/2022	8	13.2	1	10	39.8	80.0	0.4843	38.74400	37.5	10.42	8.5
...
943	944	2/27/2023	25	22.2	19	10	39.6	79.2	0.4843	38.35656	13.0	3.61	1.0
944	945	2/27/2023	40	22.2	19	10	39.6	79.2	0.4843	38.35656	13.0	3.61	1.0
945	946	2/28/2023	15	26.7	26	4	40.0	80.1	0.4843	38.79243	12.0	3.33	0.0
946	947	2/28/2023	25	26.7	26	4	40.0	80.1	0.4843	38.79243	12.0	3.33	0.0
947	948	2/28/2023	40	26.7	26	4	40.0	80.1	0.4843	38.79243	12.0	3.33	0.0

831 rows x 14 columns

Rysunek 2: Podział danych na dwie części ze względu na energię elektryczną i silniki spalinowe

3.2. Usunięcie niepotrzebnych kolumn

Usunięta została kolumna dotycząca prędkości wiatru przeliczana w jednostkach km/h, gdyż analogiczna kolumna w jednostkach układu SI czyli m/s jest wystarczająca. Ponadto przekształcono kolumnę Data w format daty i czasu. Następnie zostały przeanalizowane braki danych.

```
Braki danych dla energii elektrycznej:

Nr. Pomiaru          0
Data                 0
Ciepota ładunku [T]  0
Długość wysięgnika [m]  0
Odległość od osi [m]  0
Wysokość podnoszenia [m]  0
Maksymalne, chwilowe zużycie energii elektrycznej [kW]  0
Dzienne zużycie energii elektrycznej [kW/8h]  0
Cena energii elektrycznej [kWh]  0
Koszt dzienny [PLN]  0
Prędkość wiatru [m/s]  0
Temperatura [C]      0
Ciśnienie [hPa]      0
Efektywność EL (kWh/tonę)  0
Efektywność          0
Typ napędu           0
Efektywność_MJ       0
dtype: int64

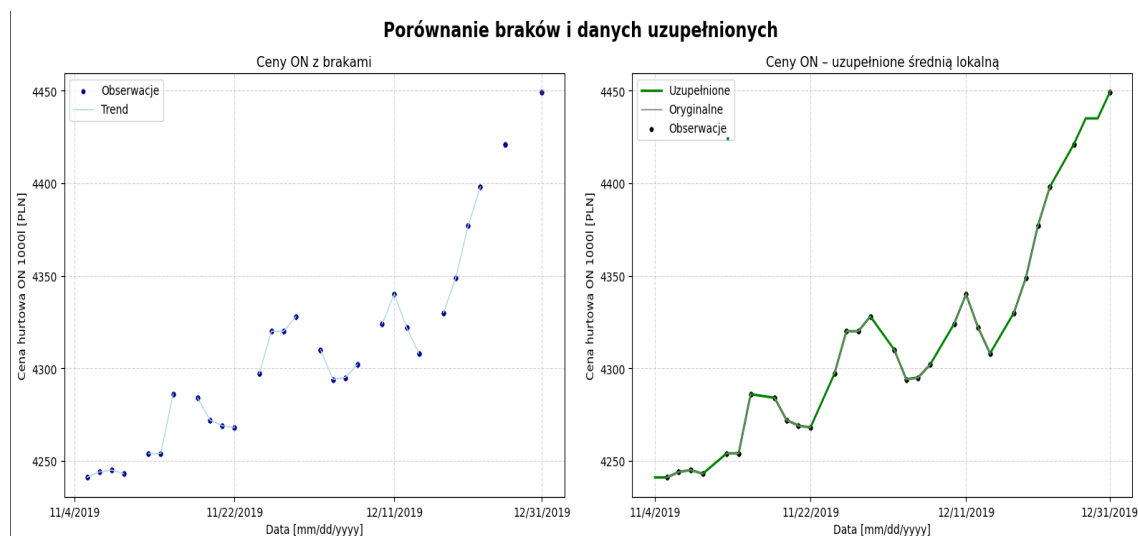
Braki danych dla ON:

Nr. Pomiaru          0
Data                 0
Ciepota ładunku [T]  0
Długość wysięgnika [m]  0
Odległość od osi [m]  0
Wysokość podnoszenia [m]  0
Maksymalne, chwilowe zużycie ON [l/h]  0
Dzienne zużycie ON [l/8h]  0
Cena hurtowa ON 1000l [PLN]  30
Koszt dzienny [PLN]  30
Prędkość wiatru [m/s]  0
Temperatura [C]      0
Ciśnienie [hPa]      0
dtype: int64
```

Rysunek 3: Braki danych w obu tabelach

3.3. Uzupełnienie braków danych

Widoczna jest dosyć spora liczba braków występująca w dwóch kolumnach tabeli z silnikiem spalinowym. Braki w cenie hurtowej są powodorem braków w koszcie dziennym, gdyż ciężko obliczyć wartość kosztów nie znając ceny hurtowej oleju napędowego. W celu uzupełnienia braków zdecydowano się na zastosowanie metody lokalnej średniej — polegającej na podstawieniu średniej ceny ON obliczonej dla najbliższych czasowo dni z dostępnymi danymi.



Rysunek 4: Dane z brakami oraz po uzupełnieniu metodą lokalnej średniej

Następnie mając już komplet danych w kolumnie Cena hurtowa ON mogliśmy wyliczyć dla tych wierszy Koszt dzienny. Zostało to wyliczone przemnażając cenę hurtową przez zużycie paliwa i dzieląc wszystko przez 1000. Po zastosowaniu tych działań liczba NaN została zniwelowana co prezentuje rysunek 5.

Braki danych dla ON:

Nr. Pomiaru	0
Data	0
Ciężar ładunku [T]	0
Długość wysięgnika [m]	0
Odległość od osi [m]	0
Wysokość podnoszenia [m]	0
Maksymalne, chwilowe zużycie ON [l/h]	0
Dzienne zużycie ON [l/8h]	0
Cena hurtowa ON 1000l [PLN]	0
Koszt dzienny [PLN]	0
Prędkość wiatru [m/s]	0
Temperatura [C]	0
Ciśnienie [hPa]	0
Efektywność ON (l/tonę)	0
Efektywność	0
Typ napędu	0
Efektywność_MJ	0
Cena hurtowa ON [zł/l]	0

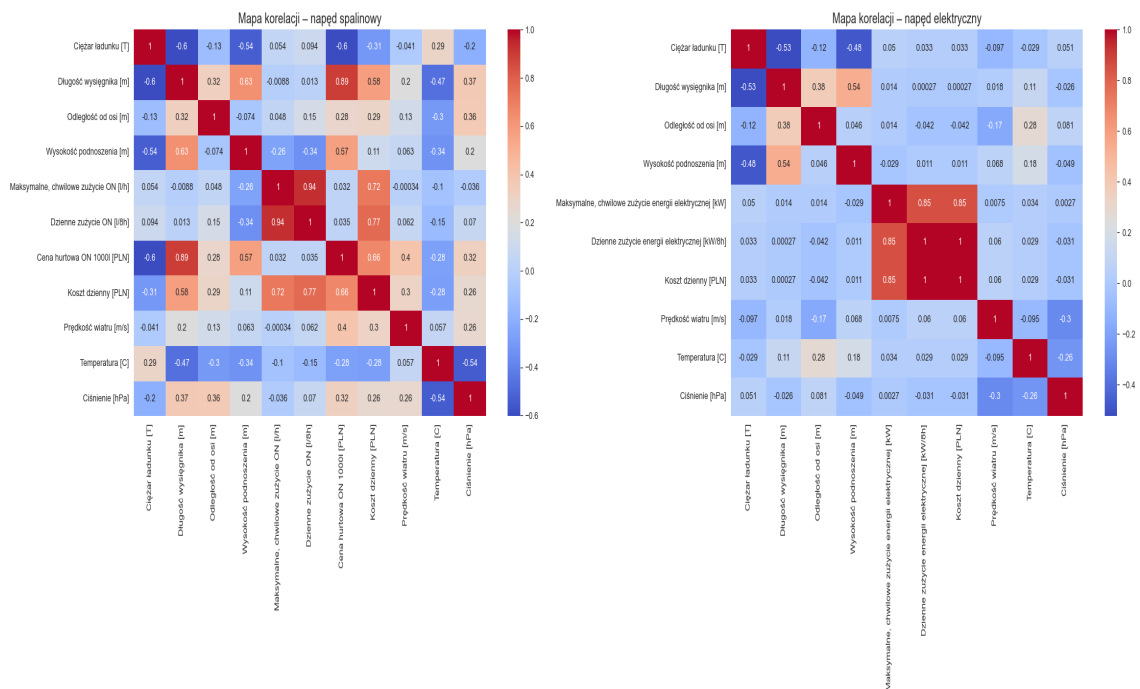
Rysunek 5: Liczba brakujących wartości (NaN) po uzupełnieniu danych

4. Exploratory Data Analysis (Eksploracyjna analiza danych)

4.1. Analiza map korelacji dla obu typów napędów

Ukazane zostały macierze korelacji dla silnika spalinowego oraz dla silnika elektrycznego.

Porównanie korelacji dla dwóch typów napędu



Rysunek 6: Macierze korelacji dla silnika ON i silnika elektrycznego

Na podstawie macierzy korelacji jesteśmy w stanie zauważyć zależności pomiędzy zmiennymi.

Niepotrzebne kolumny takie jak numer pomiaru, czy data zostały pominięte, a korelacja związana z tymi zmiennymi nie ma znaczenia przyczynowo-skutkowego.

Z lewego rysunku widać silną korelację kosztu dziennego z takimi parametrami jak: dzienne zużycie ON, maksymalne, chwilowe zużycie ON oraz z ceną hurtową ON za 1000l.

Logicznym powiązaniem jest silna korelacja dziennego zużycia ON i maksymalnego chwilowego zużycia ON.

Ciężar ładunku ujemnie koreluje z długością wysięgnika (-0.60) i wysokością podnoszenia (-0.54) – może sugerować, że większe ciężary są podnoszone przy krótszym wysięgniku i mniejszej wysokości.

Zauważalny jest również niewielki związek między temperaturą, prędkością wiatru, a dziennym zużyciem ON oraz liniowe zależności między wysokością podnoszenia, odległością od osi, a dziennym zużyciem ON.

Rysunek przedstawiający macierz korelacji dla silnika elektrycznego również wskazuje zależności między zmiennymi.

Koszt dzienny, Dzienne zużycie energii, Maksymalne zużycie energii – wszystkie między sobą bardzo silnie skorelowane (0.85–1). Oznacza to, że koszt zależy głównie od zużycia energii.

Parametry maszyny (Ciężar ładunku, Długość wysięgnika, Wysokość podnoszenia) mają bardzo słabą korelację z zużyciem i kosztem.

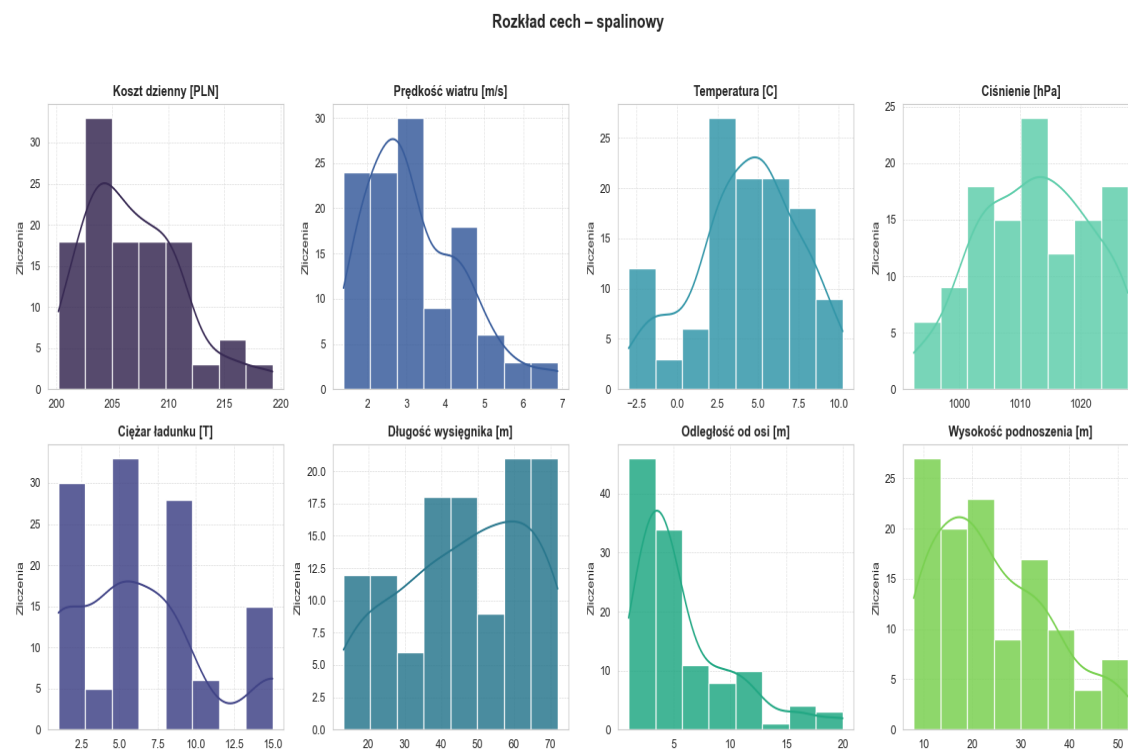
Warunki atmosferyczne także nie wpływają znacząco na zmienne docelowe.

W przypadku napędu spalinowego, koszty i zużycie paliwa są mocno uzależnione od

warunków pracy maszyny oraz ceny ON, natomiast w napędzie elektrycznym zużycie i koszty są dużo bardziej stabilne i zdeterminowane niemal wyłącznie przez ilość zużytej energii.

4.2. Rozkład cech dla silnika spalinowego

W celu analizy struktury danych wejściowych dla napędu spalinowego, wygenerowano histogramy przedstawiające rozkłady wartości poszczególnych cech. Wykresy pozwalają na m.in: ocenę symetrii rozkładów, a także kształtu funkcji gęstości.



Rysunek 7: Rozkłady wartości cech

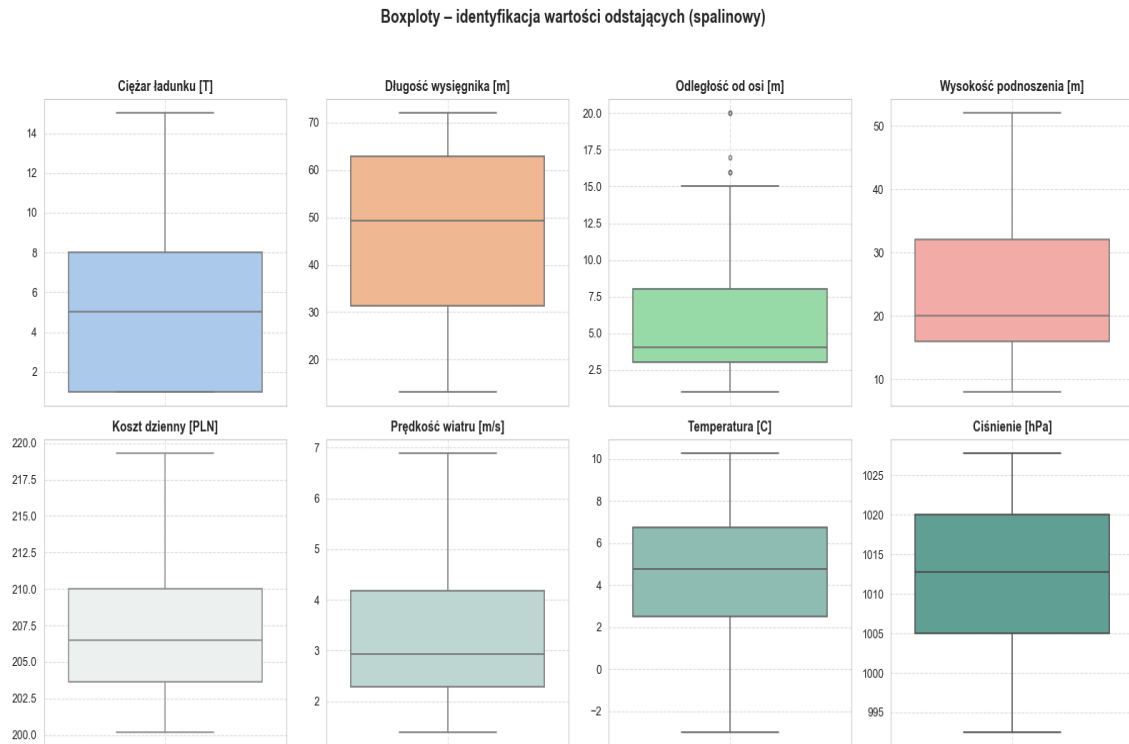
Większość analizowanych cech nie wykazuje rozkładu zbliżonego do normalnego – są zauważalne asymetrie, skupienia wartości, a także wartości odstające.

Cechy takie jak Ciśnienie oraz Temperatura wykazują rozkłady najbardziej zbliżone do rozkładu normalnego – są względnie symetryczne, bez wyraźnych wartości odstających. Może to świadczyć o ich naturalnym pochodzeniu i mniejszym wpływie czynników zakłócających.

Wstępna analiza rozkładu danych pozwala na ocenę jakości pomiarów.

4.3. Wartości odstające

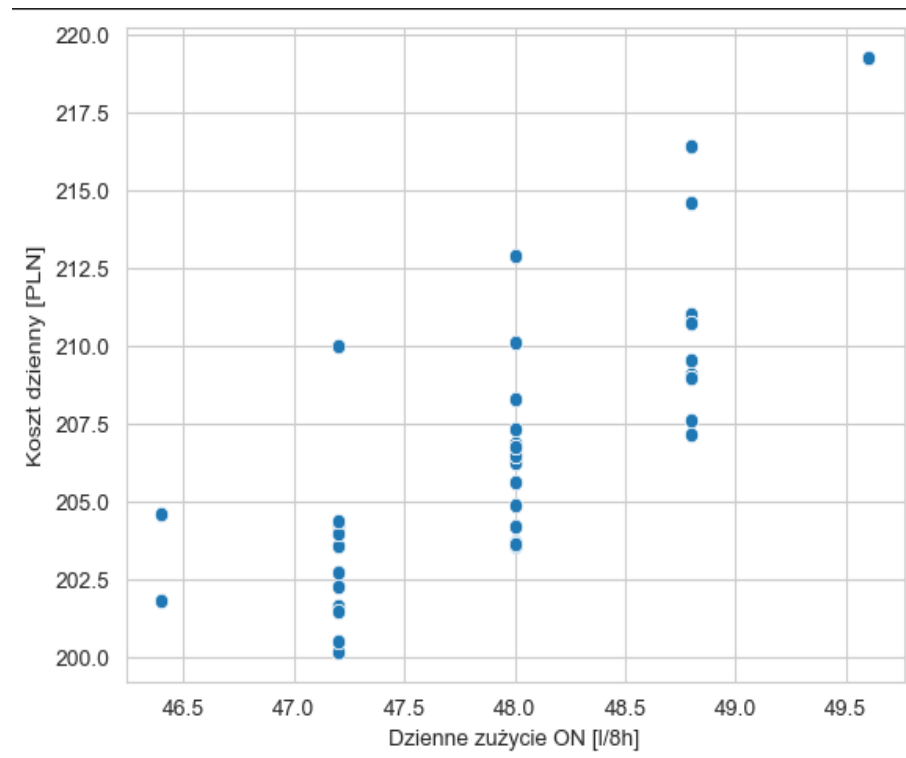
Wykonano boxploty w celu identyfikacji cech z wartościami odstającymi. Anomalie zaobserwowano dla odległości od osi. Wartości odstające pozostawiono w zbiorze.



Rysunek 8: Wartości odstające

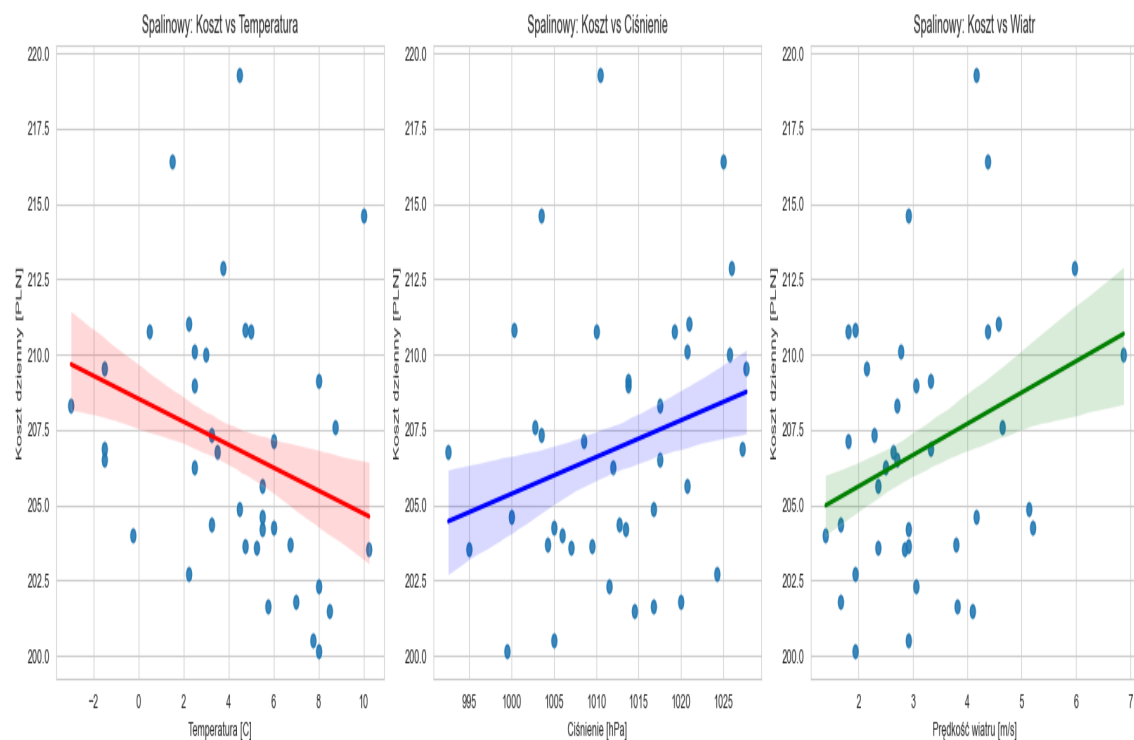
4.4. Analiza zależności poszczególnych zmiennych między sobą

Dla danych z silnikiem spalinowym dokonano analizy korelacji dla zmiennej Koszt dzienny [PLN]. Najwyższą wartość ma ona ze zmienną Dzielne zużycie ON. Wynosi aż 0.770172.0 oznacza silną dodatnią korelację liniową między tymi dwiema zmiennymi. Można więc pokusić się o stwierdzenie, że wraz ze wzrostem dziennego zużycia ON rośnie również koszt dzienny. Poniższy rysunek przedstawia tę zależność



Rysunek 9: Zależność Kosztu dziennego od dziennego zużycia ON

Następnie przeanalizowano dla silnika spalinowego wpływ różnych cech meteorologicznych na koszt dzienny co przedstawia rysunek 10.

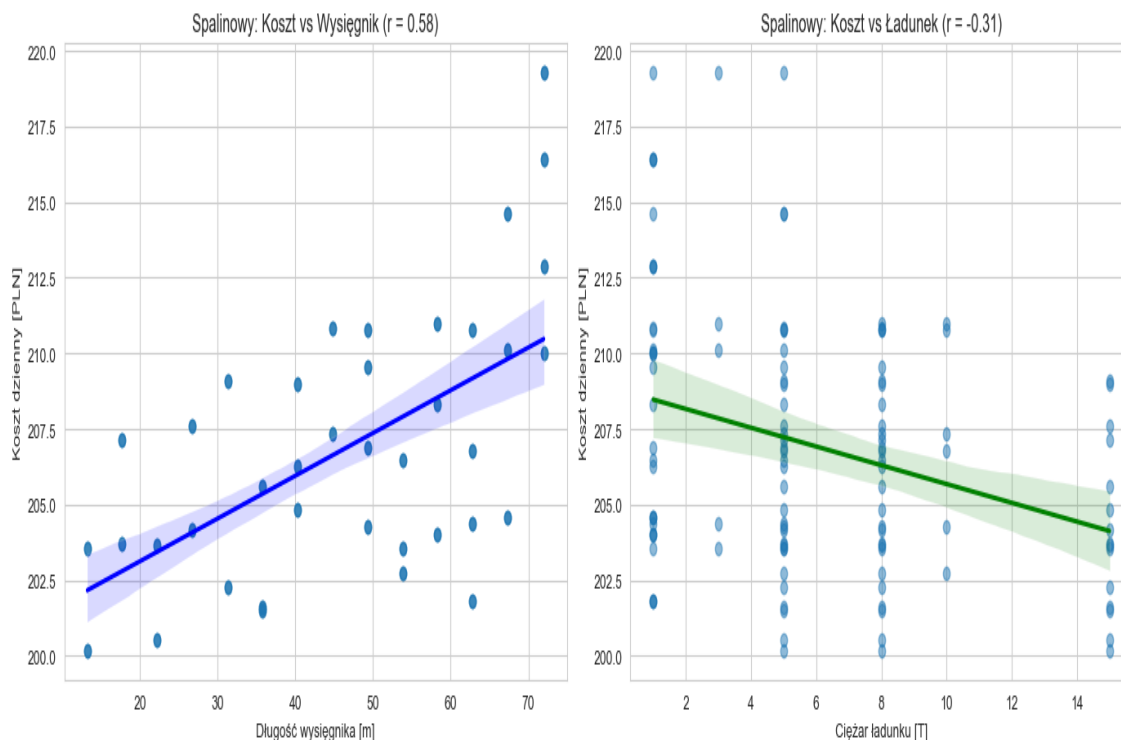


Rysunek 10: Cechy meteorologiczne a dzienny koszt

Zauważono niewielką zależność liniową i w przypadku, gdy temperatura maleje to

koszt również maleje. Gdy rośnie ciśnienie oraz zwiększa się prędkość wiatru, rośnie koszt dzienny. Jest to marginalna zależność.

Następnie sprawdzono jaki wpływ ma długość wysięgnika oraz ciężar ładunku na koszt dzienny. Widoczna jest większa liniowa zależność i w momencie gdy rośnie długość wysięgnika, zwiększa się koszt dzienny



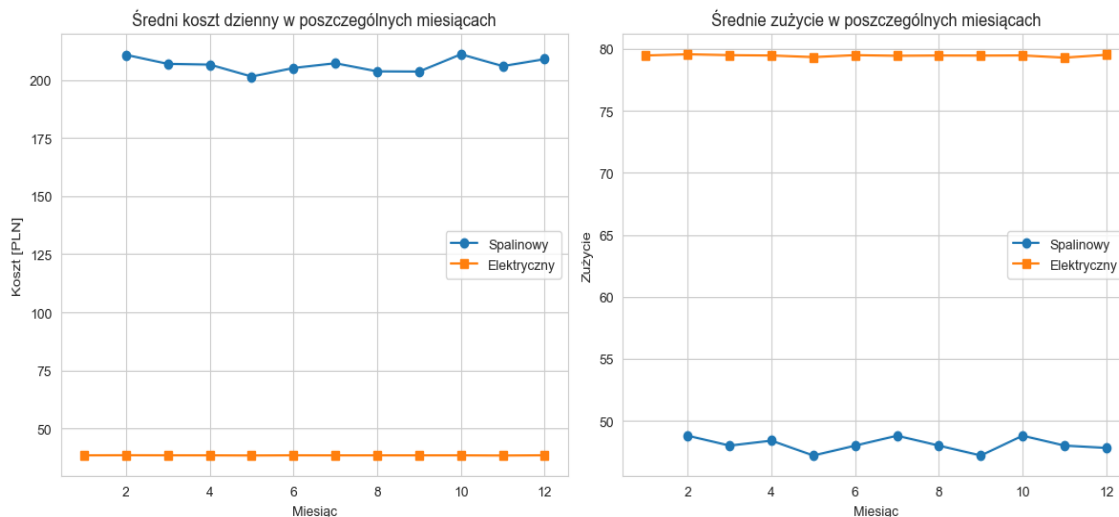
Rysunek 11: Wysięgnik oraz ciężar a koszt dzienny

4.5. Analiza trendów czasowych

W ramach analizy sezonowości danych, edytowano ramki danych w taki sposób, aby umożliwić grupowanie po miesiącach. W tym celu kolumna zawierająca daty została przekonwertowana do formatu typu datetime, a następnie utworzono nową kolumnę Miesiąc, zawierającą numery miesięcy pozyskane z daty pomiaru.

Na tej podstawie przeanalizowano, czy miesiąc (a tym samym potencjalnie pora roku) ma wpływ na koszt dzienny oraz zużycie energii/paliwa. W tym celu przygotowano wykresy liniowe i/lub słupkowe przedstawiające średnie wartości kosztów oraz zużycia dla każdego miesiąca w roku.

Z uzyskanych wykresów wynika, że miesiąc nie wywiera istotnego wpływu na poziom kosztów ani zużycie paliwa/energii. Obserwowane różnice pomiędzy poszczególnymi miesiącami są niewielkie, a wartości analizowanych metryk utrzymują się na zbliżonym poziomie przez cały rok. Można więc wnioskować, że sezonowość nie jest dominującym czynnikiem w kontekście kosztów eksploatacyjnych dźwignic.



Rysunek 12: Sezonowe zmiany kosztów i zużycia

4.6. Analiza porównawcza napędów

W celu określenia, który rodzaj napędu – spalinowy czy elektryczny – charakteryzuje się wyższą efektywnością operacyjną, przeprowadzono analizę porównawczą wybranych metryk opisujących zużycie energii, koszty oraz efektywność pracy.

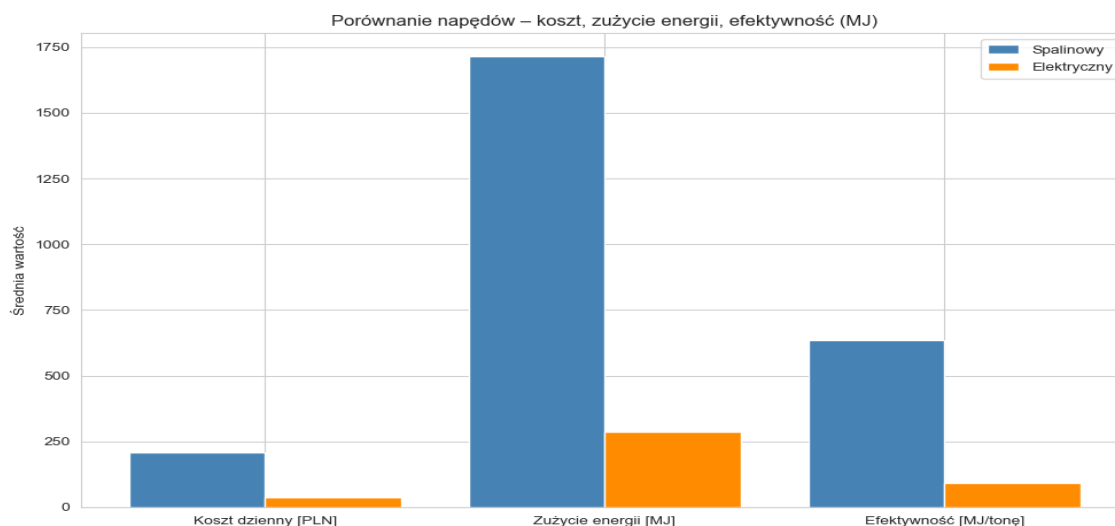
Analizie poddano następujące wskaźniki:

- średnią efektywność energetyczną,
- średnie dzienne zużycie energii,
- średni koszt dzienny eksploatacji.

Na potrzeby tej analizy:

- połączono dane z obu typów napędu (spalinowy i elektryczny) w jedną ramkę danych,
- ujednolicono jednostki zużycia energii do megadżuli (MJ), stosując przeliczniki: 1 litr ON = 35,8 MJ oraz 1 kWh = 3,6 MJ,
- dodano następujące kolumny:
 - Efektywność_MJ – skrótowa nazwa efektywności energetycznej,
 - Typ napędu – etykieta klasyfikująca rodzaj napędu (spalinowy lub elektryczny),
 - Zużycie energii [MJ] – dzienne zużycie energii przeliczone na MJ,
 - Efektywność [MJ/tonę] – zużycie energii (MJ) w przeliczeniu na tonę przewożonego ładunku.

W celu przejrzystej prezentacji wyników, przygotowano wykres słupkowy przedstawiający średnie wartości analizowanych wskaźników dla obu typów napędów.



Rysunek 13: Porównanie napędów: koszt, zużycie, efektywność

Wnioski z analizy:

- Napęd **elektryczny** wykazuje niższe wartości w kolumnie **Efektywność**, co oznacza wyższą efektywność energetyczną – zużywa mniej energii na przetransportowanie jednej tony ładunku.
- Koszt dzienny użytkowania napędu elektrycznego jest niższy, co wynika z niższej ceny jednostkowej energii elektrycznej względem oleju napędowego.
- Zużycie energii (w MJ) w ujęciu dziennym również wypada korzystniej dla napędu elektrycznego, co potwierdza jego wyższą sprawność.

Podsumowując, przy obecnych warunkach eksploatacyjnych i cenach energii, napęd elektryczny okazuje się bardziej efektywny zarówno energetycznie, jak i kosztowo.

5. Wybór zmiennej docelowej (TARGET)

Dobrym kandydatem na zmienną docelową (TARGET) do zastosowania w modelu uczenia nadzorowanego jest:

Efektywność [MJ/tonę]

Uzasadnienie wyboru:

- Jest to zmienna ciągła i ilościowa, co umożliwia zastosowanie algorytmów regresyjnych. Dzięki temu można przewidywać konkretne wartości efektywności na podstawie innych cech.
- Efektywność energetyczna jest jednym z kluczowych wskaźników wydajności operacyjnej, szczególnie istotnym przy analizie porównawczej różnych typów napędów.
- Predykcja tej zmiennej może być wykorzystana do optymalizacji operacji – np. planowania pracy maszyn w sposób pozwalający na minimalizację zużycia energii na jednostkę ładunku, co przekłada się na niższe koszty eksploatacyjne.

6. Wybór zmiennych wejściowych (FEATURES)

Na podstawie analizy dostępnych danych, bez wykorzystania technik obliczających wagę cech, do wyznaczenia zmiennej docelowej **Efektywność [MJ/tonę]** wybrano następujący podzbiór zmiennych wejściowych:

- **Ciężar ładunku [T]** – bezpośrednio wpływa na mianownik wzoru na efektywność; im większy ładunek przy tym samym zużyciu, tym niższe zużycie energii na tonę.
- **Zużycie energii [MJ]** – główny składnik licznika w wyliczaniu efektywności. Zmienna ta zawiera już przeliczone dane energetyczne zarówno dla napędu elektrycznego, jak i spalinowego, co ujednolica analizę.
- **Temperatura [°C], Ciśnienie [hPa], Prędkość wiatru [m/s]** – warunki atmosferyczne mogą w jakiś sposób wpływać na sprawność maszyn i opory pracy, co może pośrednio modyfikować efektywność.
- **Typ napędu** – istotna zmienna kategoryczna, ponieważ typ napędu (elektryczny lub spalinowy) wprowadza inne źródła energii i charakterystyki pracy.

7. Wnioski

Celem projektu była eksploracyjna analiza danych (EDA) oraz przygotowanie zbioru danych pod kątem dalszych zastosowań w modelowaniu predykcyjnym. W toku analizy dokonano szeregu transformacji, oczyszczenia i wzbogacenia danych, co umożliwiło lepsze zrozumienie zależności między cechami a zmienną docelową.

W projekcie:

- Uzupełniono brakujące dane oraz przekształcono jednostki energetyczne do wspólnej skali (MJ), co pozwoliło na bezpośrednie porównanie różnych typów napędu.
- Stworzono nowe cechy (np. efektywność energetyczna), które lepiej opisują zjawiska zachodzące w systemie.
- Wykorzystano wizualizacje (histogramy, wykresy słupkowe, boxploty), które pozwoliły zidentyfikować rozkłady zmiennych, wartości odstające oraz potencjalne zależności.
- Przeprowadzona analiza porównawcza napędów spalinowego i elektrycznego wykazała, że napęd **elektryczny** charakteryzuje się wyższą efektywnością energetyczną oraz niższymi kosztami eksploatacyjnymi w przeliczeniu na tonę ładunku.
- Dokonano wstępnego wyboru zmiennej docelowej (target) oraz cech wejściowych (features), co przygotowuje dane do użycia w modelach uczenia maszynowego.

Projekt umożliwił uzyskanie pogłębionego wglądu w dane, co stanowi solidną bazę do dalszych analiz predykcyjnych, takich jak budowa modelu regresyjnego.

Ponadto, wyniki analizy mogą stanowić podstawę do podejmowania decyzji operacyjnych i inwestycyjnych w zakresie wyboru rodzaju napędu w sprzęcie technicznym.