

Predykcja Cen Samochodów



**Marzec
2025**

Przygotowane przez:
Łukasz Podoba
Bartosz Paszko

Przygotowane dla:



DATA
SCIENCE
CLUB
PJATK

SPIS TREŚCI

WPROWADZENIE	1
DATASET	2
EDA ILOŚCIOWE	3
EDA JAKOŚCIOWE	4
PREPROCESSING	5
MODELOWANIE	6
PODSUMOWANIE	7
PODZIĘKOWANIE	8

WPROWADZENIE

Konkurs Kaggle organizowany przez DSC PJATK stanowi drugi etap procesu rekrutacji do koła naukowego. Celem wydarzenia jest umożliwienie uczestnikom zaprezentowania umiejętności w zakresie analizy danych i modelowania predykcyjnego na przykładzie danych dotyczących rynku motoryzacyjnego.



DATA
SCIENCE
CLUB
PJATK

DSC PJATK

Data Science Club
PJATK

Cel konkursu

Stworzenie modelu regresyjnego do precyzyjnego przewidywania cen samochodów na podstawie dostępnych cech opisowych pojazdów.

Główne Wyzwania

- Jakość danych: Błędy, brakujące wartości oraz wartości odstające wynikające ze specyfiki web scrapingu.
- Analiza Eksploracyjna (EDA): Dokładna eksploracja danych, identyfikacja kluczowych czynników wpływających na cenę, zarządzanie danymi nietypowymi.
- Optymalne modelowanie: Budowa odpornego na overfitting algorytmu uwzględniającego złożoność oraz zmienność rynku samochodowego.

kaggle

Kontekst i Problem

- Domena: Handel samochodami używanymi i nowymi.
- Źródło danych: Web scraping ofert sprzedaży z popularnego serwisu ogłoszeniowego.
- Zakres danych: 25 zmiennych opisujących cechy pojazdów, m.in.: marka, model, rok produkcji, przebieg, moc silnika, emisja CO₂, rodzaj paliwa, wyposażenie.

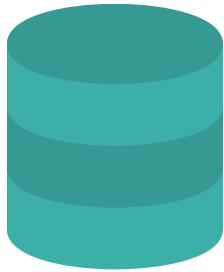
25

Zmiennych

1.385 MLN

Wierszy w 3 datasetach
(2 stworzone syntetycznie)

DATASET



Do finalnej predykcji użyty został wyłącznie oryginalny zbiór **sales_ads_train.csv**. Dane syntetyczne nie zostały wykorzystane ze względu na występowanie błędnych zależności między kolumnami (np. nieistniejące kombinacje typu „Audi Astra”) lub brak korelacji między kolumnami.

Zmienne

Komórki tabeli oznaczone na **zielono** wskazują zmienne, które zostały ostatecznie wykorzystane w modelu predykcyjnym ceny pojazdu, natomiast komórki oznaczone na **czerwono** przedstawiają zmienne, które **NIE** zostały wykorzystane.

Zmienne ilościowe	Zmienne jakościowe
Cena - Atrybut decyzyjny	ID - unikalny identyfikator ogłoszenia
Rok_produkcji	Waluta - głównie PLN, ale również EUR
Przebieg_km - w kilometrach	Stan - nowy lub używany
Moc_KM - moc silnika w koniach mechanicznych	Marka_pojazdu
Pojemnosc_cm3 - w centymetrach sześciennych	Model_pojazdu
Emisja_CO2 - emisja CO ₂ w g/km	Wersja_pojazdu
Liczba_drzwi	Generacja_pojazdu
	Rodzaj_paliwa
	Naped
	Skrzynia_biegow
	Typ_nadwozia
	Kolor
	Kraj_pochodzenia
	Pierwszy_własciciel - czy jest pierwszym właścicielem
	Data_pierwszej_rejestracji
	Data_publikacji_oferty
	Lokalizacja_oferty - podana przez sprzedającego
	Wyposażenie - lista wyposażenia pojazdu



ANALIZA EKSPLORACYJNA DANYCH (EDA)

W tej sekcji raportu zostanie przedstawiona Analiza Eksploracyjna Danych, podzielona na dwie części:

EDA ILOŚCIOWA

Analiza rozkładów zmiennych numerycznych oraz identyfikacja wartości odstających.

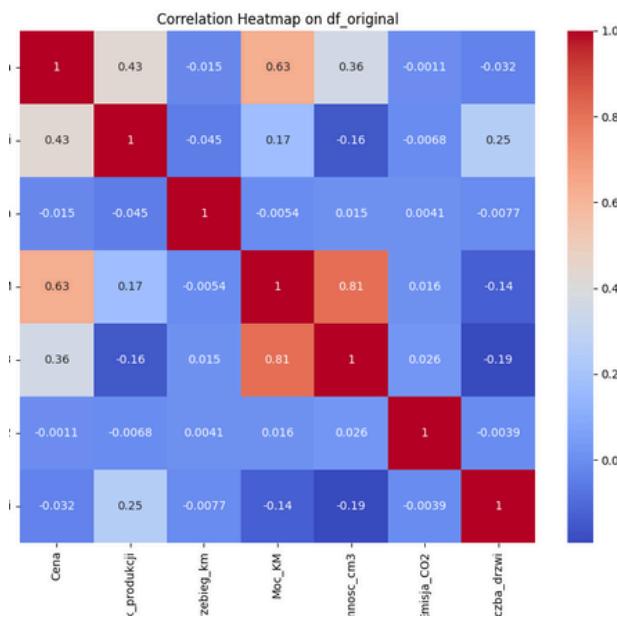
EDA JAKOŚCIOWA

Analiza rozkładów zmiennych kategorycznych oraz najważniejszych zależności jakościowych w danych.

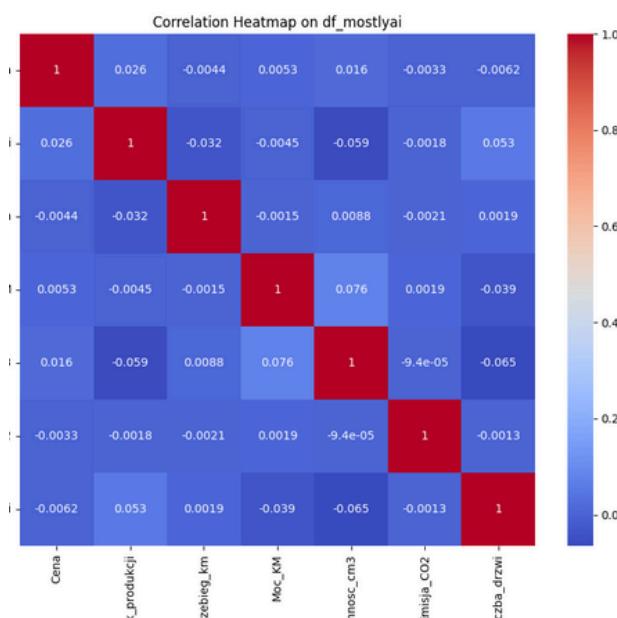
EDA ILOŚCIOWE

ANALIZA KORELACJI

* df_original - oryginalne dane z web scrapingu

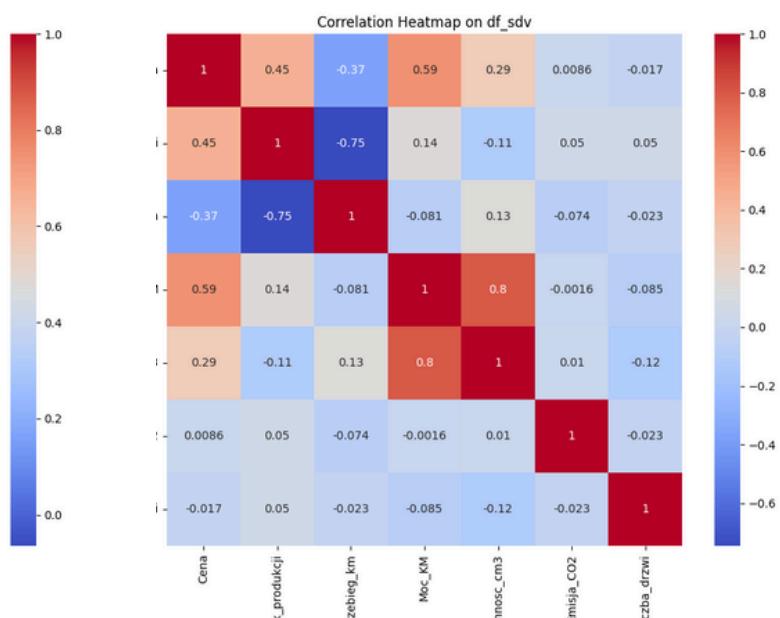


* df_mostlyai - syntetyczny (250k wierszy)

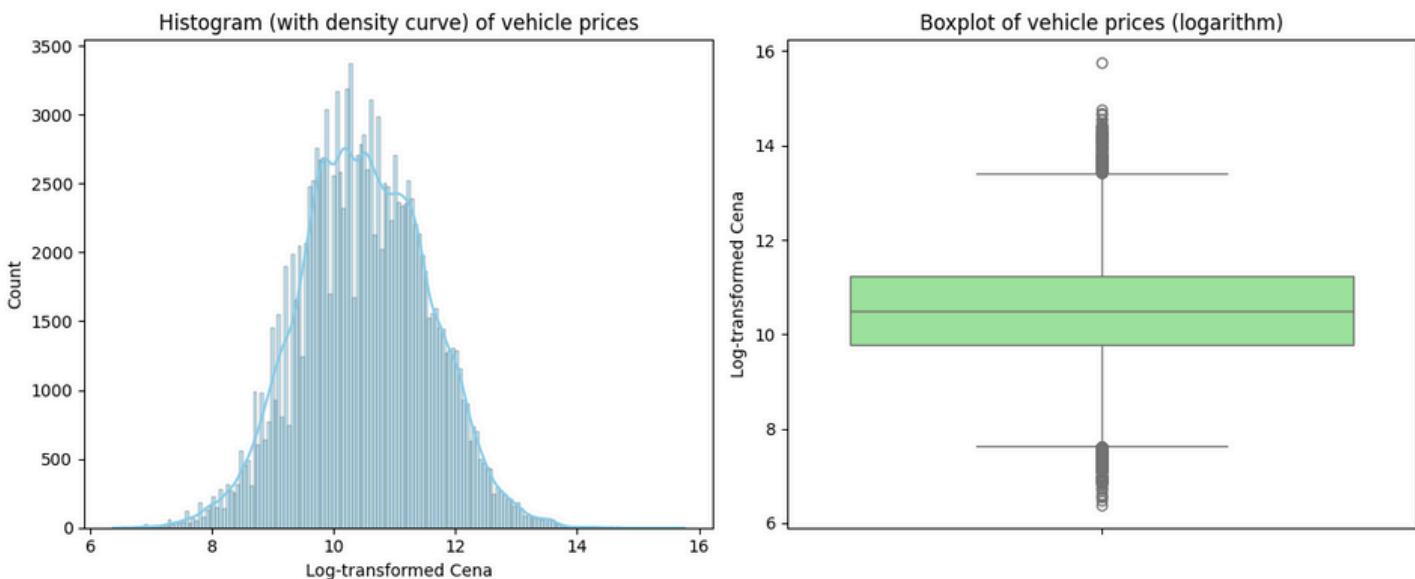


Rozpoczęliśmy badanie zależności między zmiennymi, analizując kilka macierzy korelacji. Wykorzystujemy oryginalny zbiór danych, ponieważ jedna z macierzy nie wykazuje żadnych istotnych korelacji, a inna, choć pokazuje korelacje, opiera się na kombinacjach Marka_pojazdu, Model_pojazdu i Generacja_pojazdu, które w rzeczywistości nie występują. Dlatego do dalszych analiz korzystamy wyłącznie z macierzy korelacji oryginalnego datasetu.

* df_original - syntetyczny (1mln wierszy)

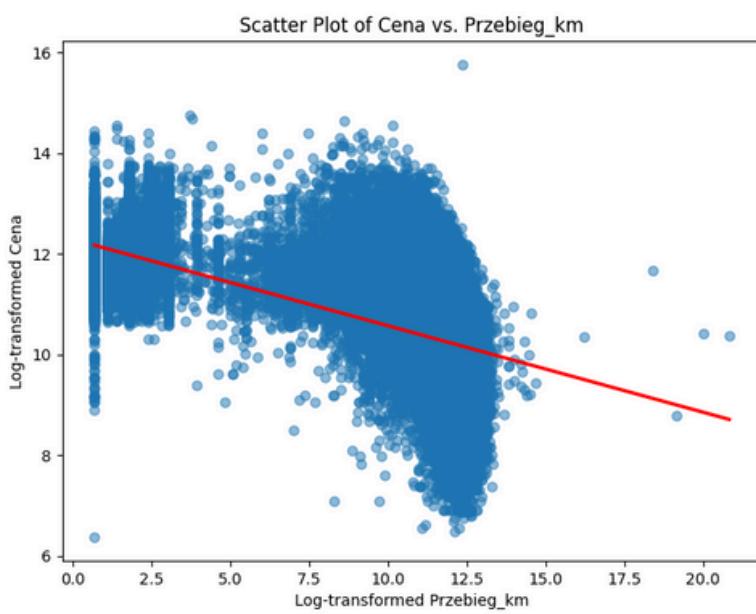


ROZKŁAD CEN



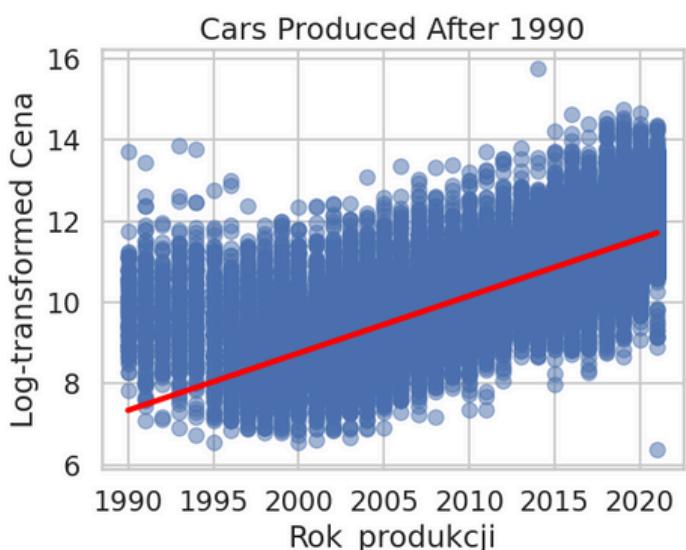
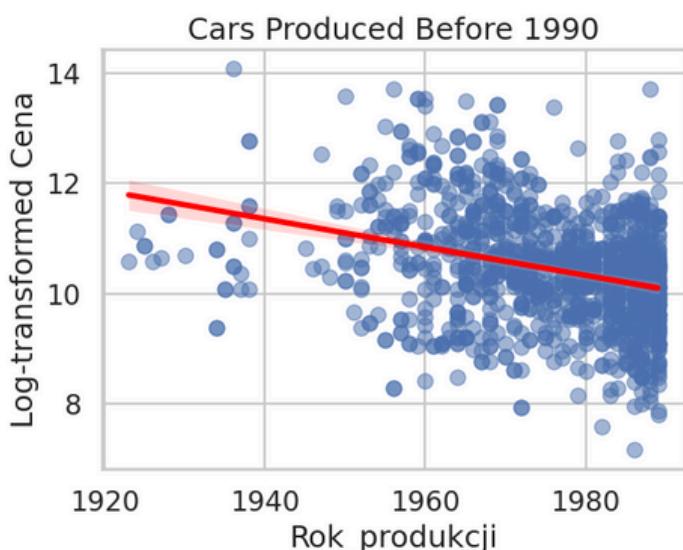
Rozkład cen pojazdów został przedstawiony za pomocą histogramu oraz boxplotu. Aby lepiej uchwycić charakterystyczny rozkład danych, ceny zostały przetransformowane logarytmicznie. Transformacja ta nie tylko ułatwia wizualizację, ale również jest stosowana w finalnym etapie trenowania modelu, co pozwala na uzyskanie bardziej symetrycznego rozkładu.

PRZEBIEG VS. CENA



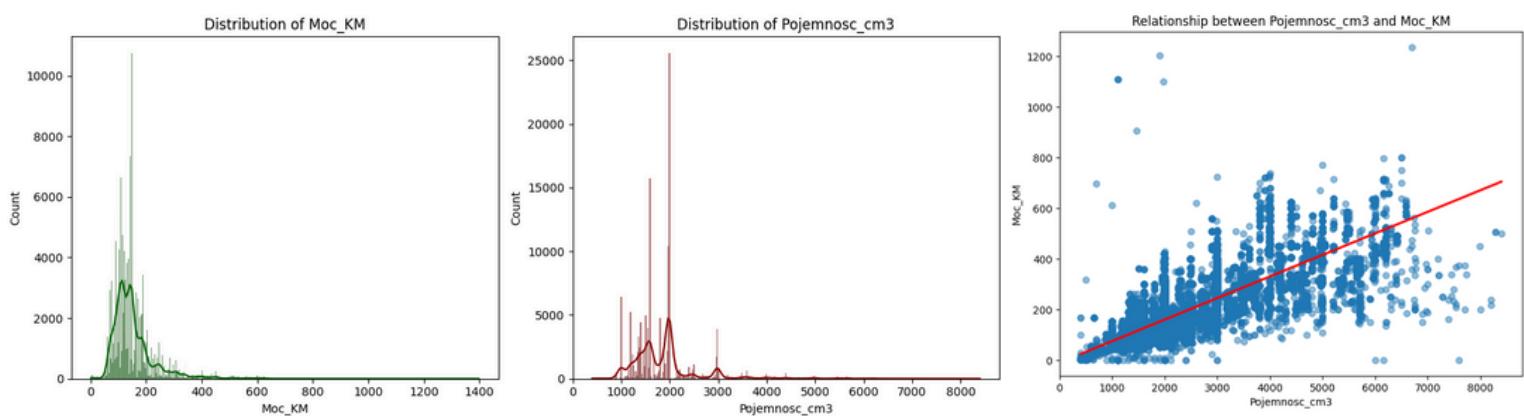
Wykres punktowy ilustrujący zależność między przebiegiem a ceną wykazuje naturalną tendencję im wyższy przebieg, tym niższa cena pojazdu. Jest to zgodne z oczekiwaniem spadkiem wartości pojazdu przy zwiększeniu zużyciu.

CENA VS. ROK PRODUKCJI



Analiza zależności między rokiem produkcji a ceną, przedstawiona na wykresie punktowym, pokazuje, że nowsze pojazdy generalnie są droższe. Jednak dla starszych samochodów trend ten nie jest już wyraźny, co może wynikać z obecności antyków lub pojazdów kolekcjonerskich.

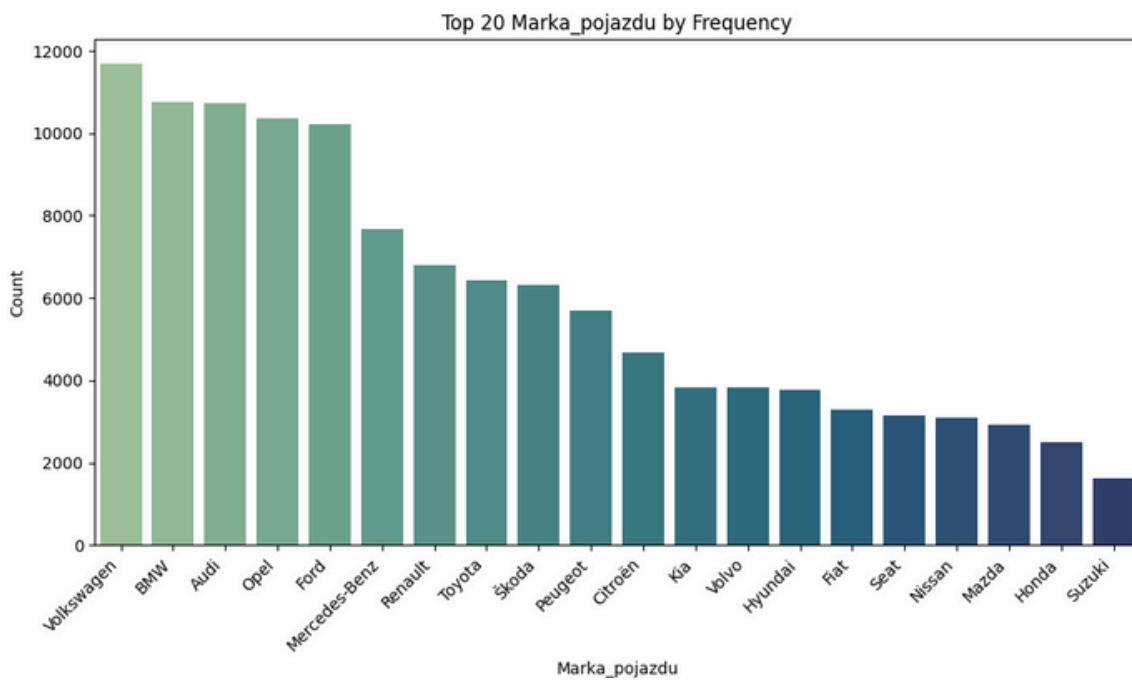
SPECYFIKACJE SILNIKA



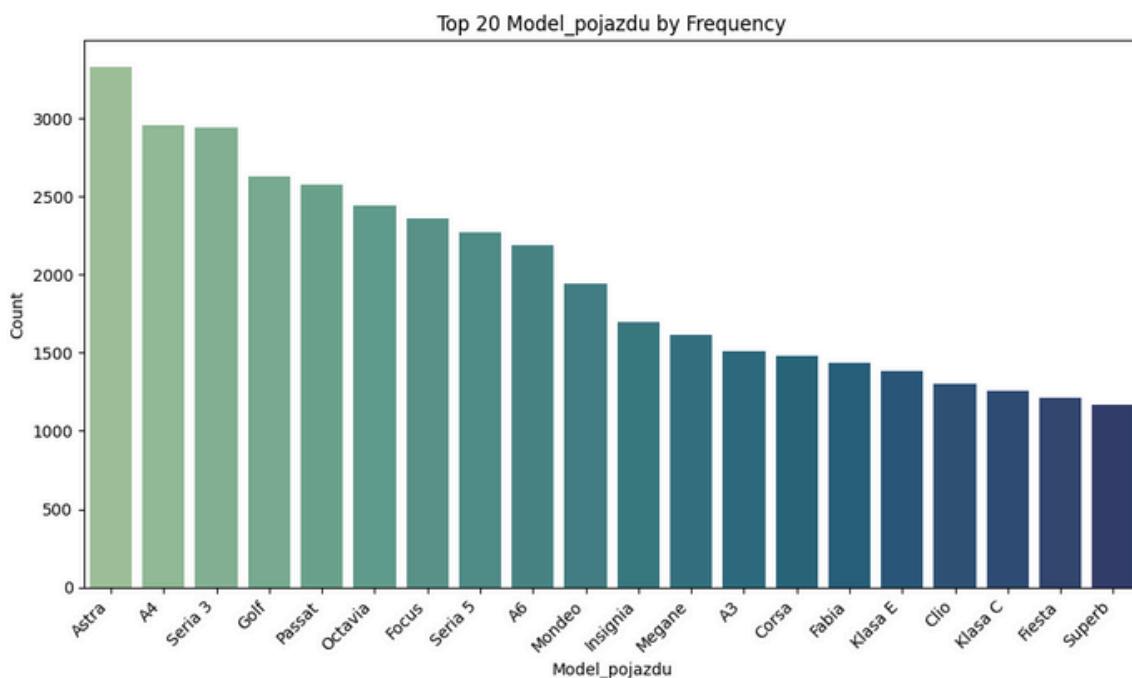
Rozkłady mocy silnika oraz pojemności silnika wskazują, że większość pojazdów koncentruje się wokół określonych, popularnych wartości. Dodatkowo, wykres punktowy pokazujący zależność między pojemnością a mocą silnika ujawnia wyraźną dodatnią korelację – im większa pojemność, tym zwykle wyższa moc.

EDA JAKOŚCIOWE

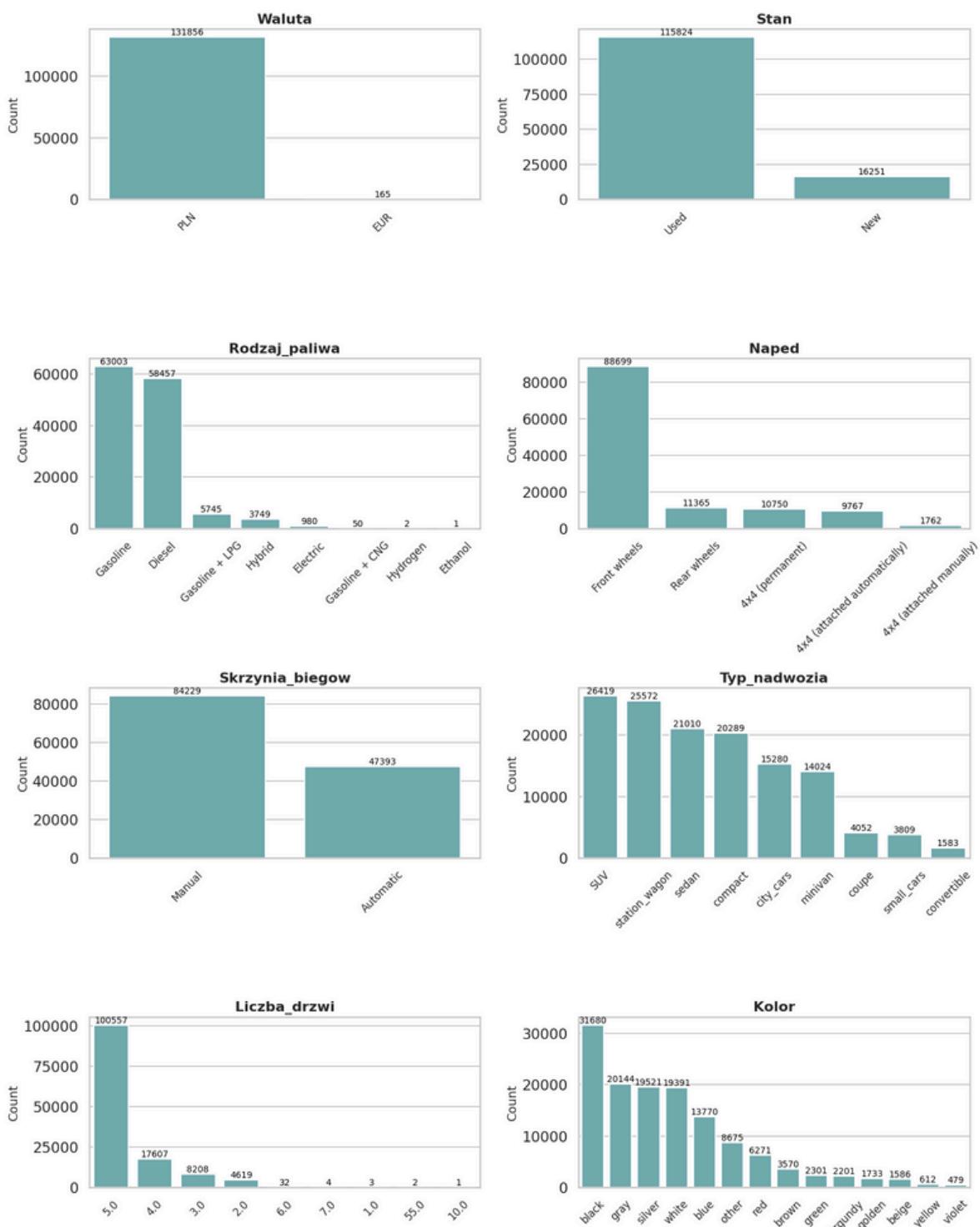
NAJPOPULARNIEJSZE MARKI I MODELE



Na pierwszych dwóch wykresach słupkowych zaprezentowano Top 20 marek oraz Top 20 modeli samochodów. Widzimy, które marki i modele dominują w naszym zbiorze danych, co może wskazywać na ich większą dostępność lub popularność na rynku wtórnym.

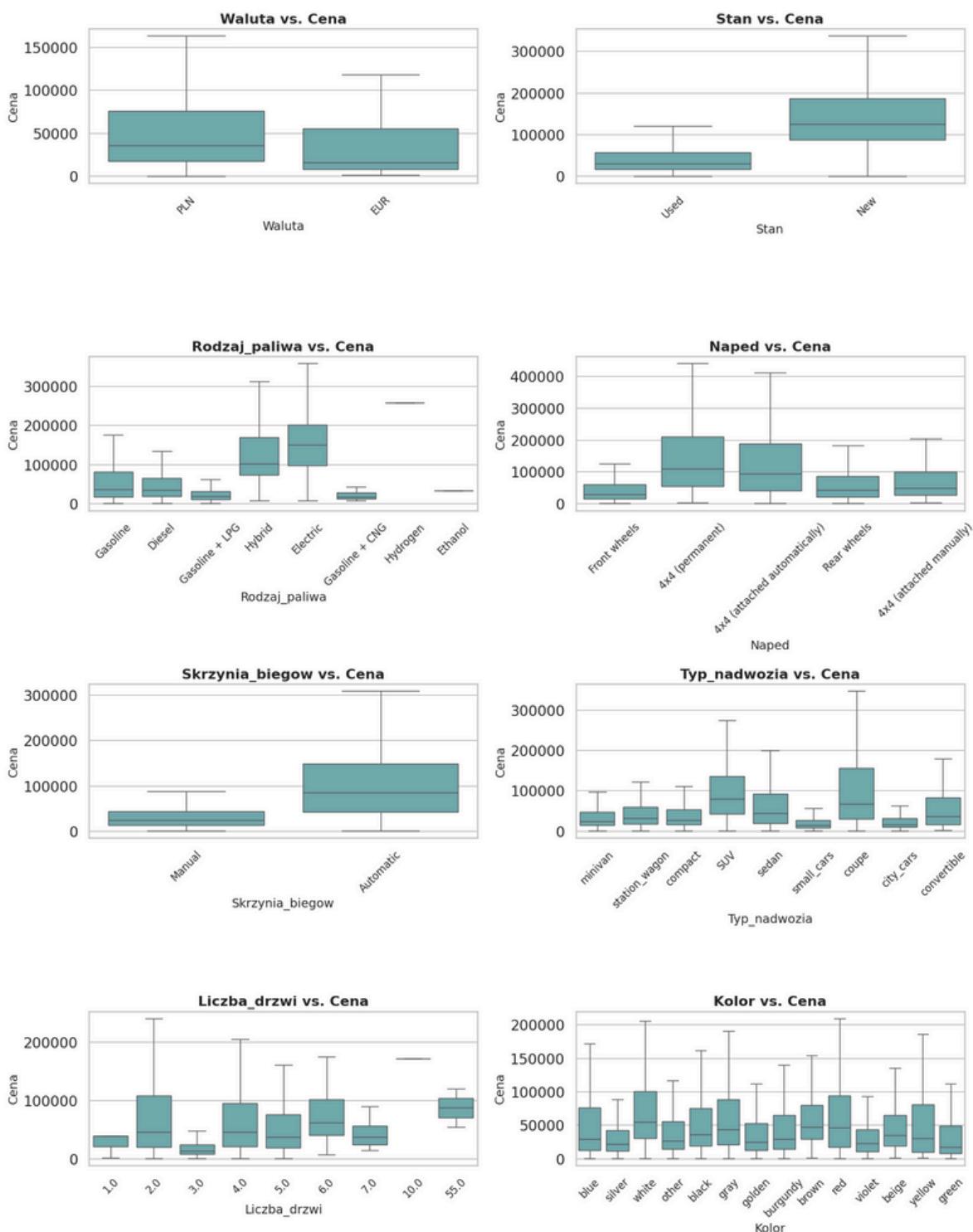


ROZKŁAD WARTOŚCI W KATEGORIACH



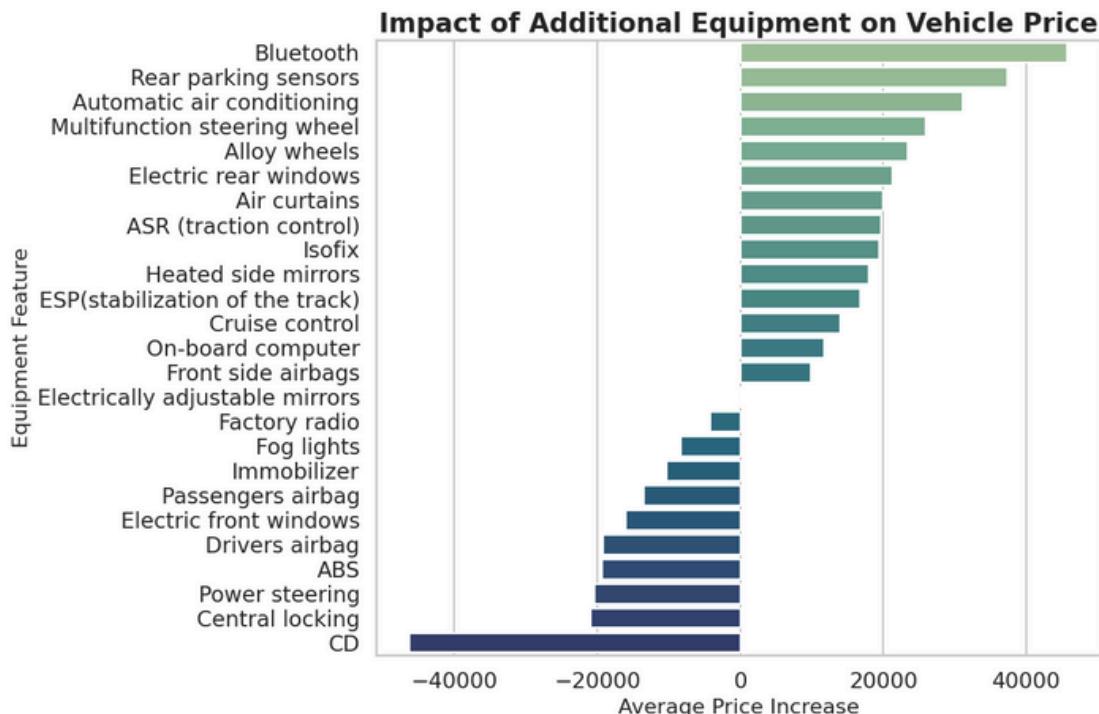
Kolejny, większy wykres przedstawia osiem kolumn kategorycznych w formie countplotów. Dzięki temu możemy szybko ocenić, które kategorie (np. rodzaj paliwa, kolor czy typ nadwozia) występują najczęściej. Jest to przydatne, by zrozumieć dominujące cechy w zbiorze danych i wychwycić ewentualne braki lub nierównowagi w kategoriach.

ZALEŻNOŚĆ KATEGORII OD CENY



Następnie widzimy osiem boxplotów pokazujących, jak poszczególne kategorie wpływają na rozkład ceny. Boxploty pozwalają zaobserwować, czy w obrębie danej kategorii cena jest wyższa, niższa czy może mocno rozproszona. Dzięki temu można wychwycić np. czy samochody z określonym rodzajem paliwa lub określonym stanem technicznym mają zauważalnie wyższe ceny.

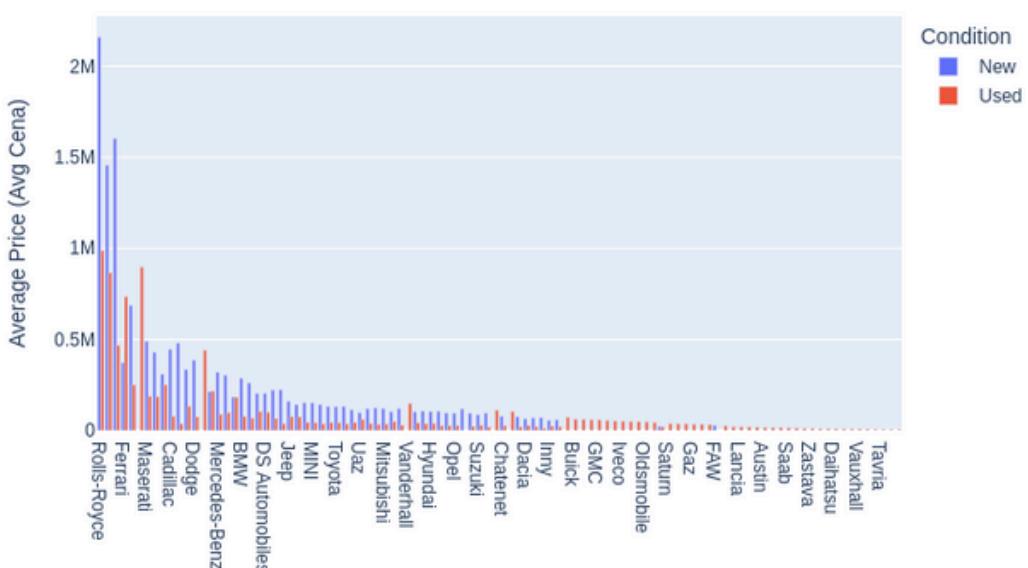
WPŁYW WYPOSAŻENIA DODATKOWEGO NA CENĘ



Kolejny wykres słupkowy pokazuje, w jaki sposób obecność określonych elementów wyposażenia (np. Bluetooth, czujniki parkowania) wiąże się z różnicą w średniej cenie pojazdu. Słupki powyżej zera oznaczają, że posiadanie danego wyposażenia wiąże się z wyższą średnią ceną, a wartości ujemne sugerują, że pojazdy z tym elementem są w naszym zbiorze danych średnio tańsze.

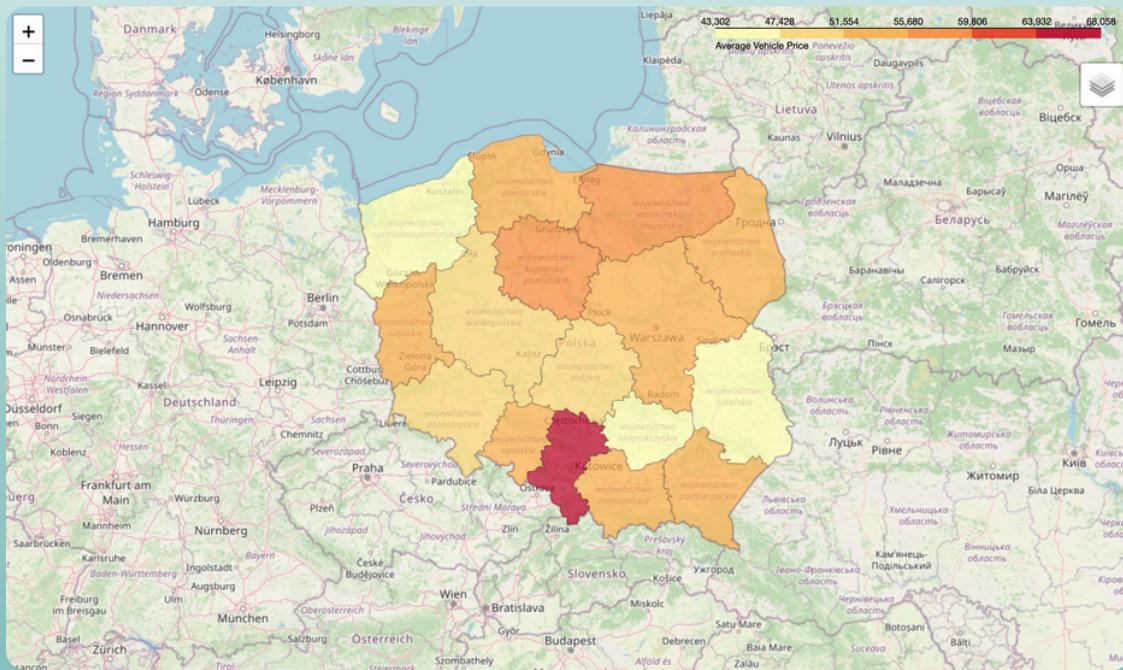
PORÓWNANIE CEN NOWYCH I UŻYWANYCH AUT WEDŁUG MARKI

Comparison of New vs. Used Vehicle Prices (Cena) by Marka_pojazdu

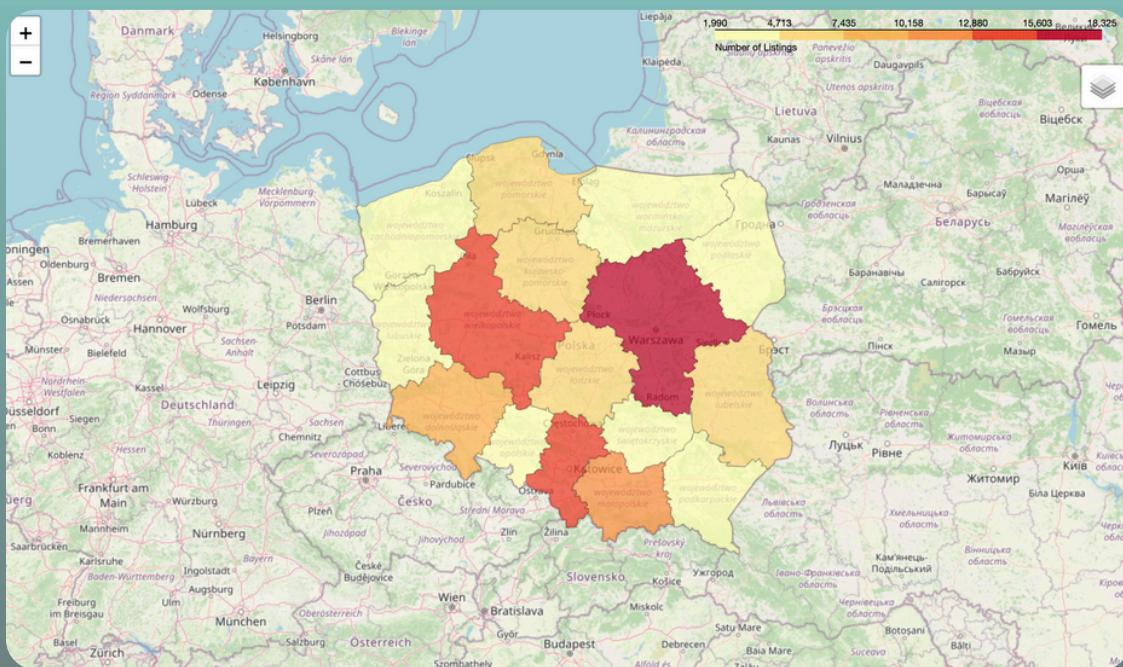


MAPA POLSKI

ŚREDNIA CENA AUTA WEDŁUG WOJEWÓDZTWA



LICZBA OFERT WEDŁUG WOJEWÓDZTWA





PREPROCESSING



IMPUTOWANIE

W ramach przygotowania danych do modelowania dokonano imputacji brakujących wartości w następujący sposób:

- **Waluta** – uzupełniona na podstawie porównania ceny ogłoszenia z medianą cen w PLN i EUR dla danego modelu, marki i roku produkcji. Użyto kursu EUR/PLN.
- **Stan pojazdu** – określony na podstawie przebiegu: pojazdy z przebiegiem ≤ 5 km uznane za New, a $> 10\,000$ km za Used.
- **Model pojazdu** – uzupełniony na podstawie marki i generacji, korzystając z danych referencyjnych.
- **Marka pojazdu** – uzupełniona na podstawie najczęściej występującej marki dla danego modelu.
- **Generacja pojazdu** – przypisana na podstawie pasującego przedziału lat dla danej marki i modelu z danych referencyjnych.
- **Rok produkcji** – brakujące wartości uzupełnione średnią z początku i końca zakresu danej generacji

6

Kolumn zostało
uzupełnionych



ENCODOWANIE

W ramach przygotowania danych do modelowania zastosowano różne techniki kodowania zmiennych:

- **Waluta** – zakodowano binarnie.
- **Stan** – zakodowano binarnie.
- **Model_pojazdu_encoded** – Target Encoding.
- **Marka_pojazdu_encoded** – Target Encoding.
- **Generacja_pojazdu_encoded** – Target Encoding.
- **Naped_encoded** – Target Encoding.
- **Wyposazenie** – One-Hot Encoding.
- **Lokalizacja** – zastosowano One-Hot Encoding na podstawie wyodrębnionego województwa z lokalizacji oferty.
- **Skrzynia_biegow** – zakodowano z użyciem One-Hot Encoding.
- **Pierwszy_wlasciciel** – zakodowano binarnie.

10

Kolumn zostało
zakodowanych



FEATURE ENGINEERING

W ramach Feature Engineering utworzono nowe zmienne:

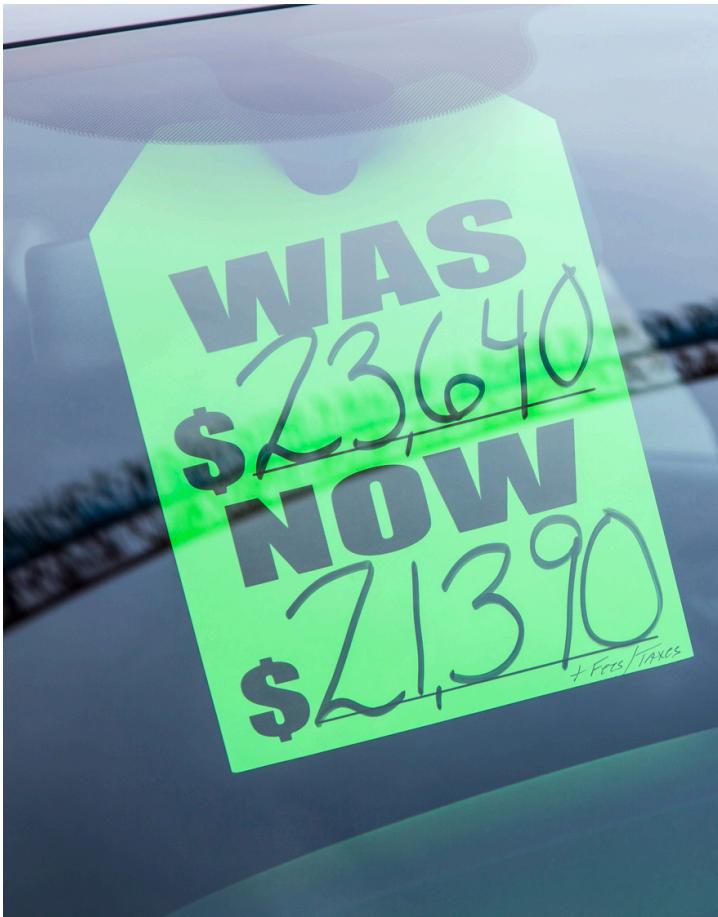
- **Moc_do_Pojemnosc** – stosunek mocy silnika do jego pojemności, przydatny do określenia wydajności silnika niezależnie od jego wielkości.
- **Pojemnosc_bin** – podział pojemności silnika na 10 równolicznych przedziałów (kwantyle).
- **Moc_bin** – podział mocy silnika na 10 przedziałów (kwantyle).
- **Duza_pojemnosc** – zmienna binarna, która przyjmuje wartość 1 dla silników o pojemności powyżej 4000.
- **Duza_moc** – zmienna binarna, która przyjmuje wartość 1 dla pojazdów o mocy powyżej 400 KM.
- **Nowy_pojazd** – zmienna binarna wskazująca, czy pojazd został wyprodukowany po 1980 roku.
- **Województwo** – lokalizacja zakodowana metodą one-hot encoding wyłuskana z kolumny 'Lokalizacja_oferty'.

7

Nowych kolumn zostało
utworzonych



MODELOWANIE



XGBOOST

Modelowanie predykcyjne

OPTUNA

Optymalizacja hiperparametrów

RMSE

Ocena dokładności

OPTYMALIZACJA HIPERPARAMETRÓW

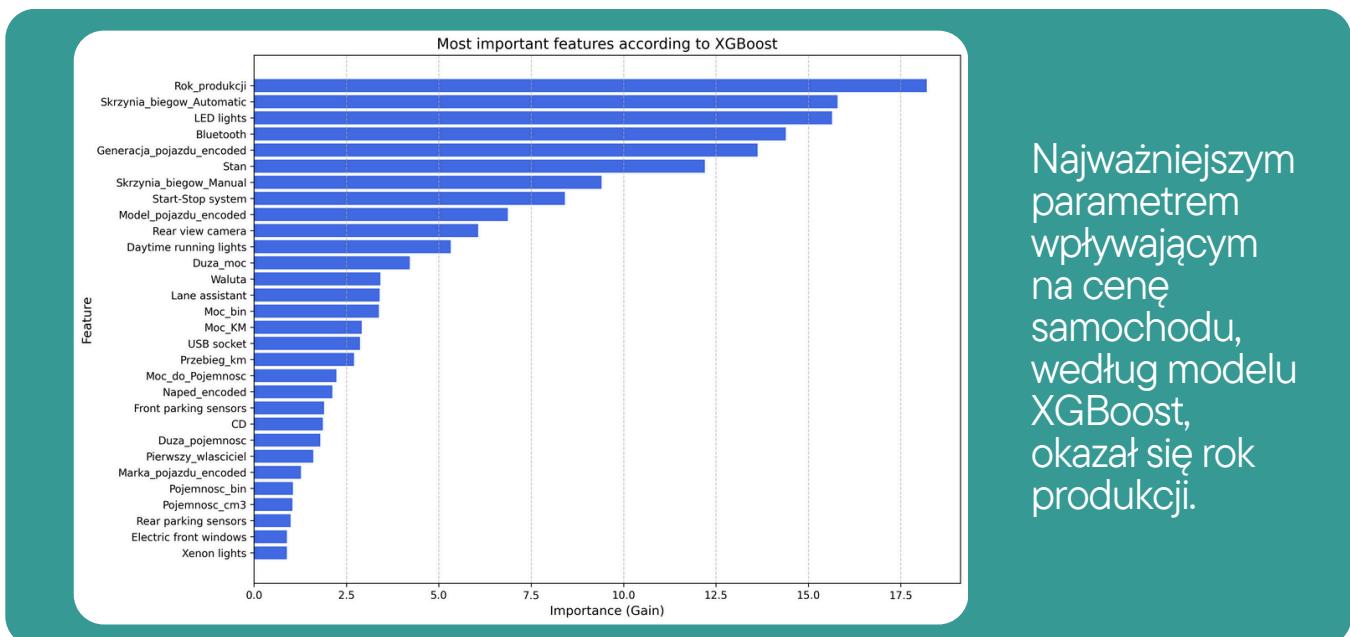
Do znalezienia najlepszych hiperparametrów wykorzystana została biblioteka **Optuna**, która automatycznie optymalizowała parametry modelu **XGBoost**. Finalne, optymalne parametry przedstawiają się następująco:

```
params = {  
    'booster': 'gbtree',  
    'eta': 0.015585502897414984,  
    'max_depth': 8,  
    'min_child_weight': 3,  
    'gamma': 0.014398001187365343,  
    'subsample': 0.5725110488549374,  
    'colsample_bytree': 0.4086665476849841,  
    'lambda': 4.745204180610975e-06,  
    'alpha': 2.5350801304973753e-05,  
}  
num_boost_round=4544
```

FEATURE IMPORTANCE

Najbardziej istotnymi cechami wpływającymi na cenę pojazdu okazały się:

- Rok produkcji – najważniejsza zmienna wpływająca na cenę.
- Wyposażenie dodatkowe (np. Bluetooth, światła dzienne LED) – silnie wpływają na wartość pojazdu.
- Generacja pojazdu (zakodowana) – potwierdza znaczenie specyfikacji modelowej auta.
- Skrzynia biegów - występowanie skrzyni automatycznej znacząco zwiększa cenę pojazdu



Najważniejszym parametrem wpływającym na cenę samochodu, według modelu XGBoost, okazał się rok produkcji.

OCENA MODELU

Do oceny modeli wykorzystano następującą metrykę jakości regresji:

- RMSE (Root Mean Squared Error) – średni błąd kwadratowy pierwiastkowy.



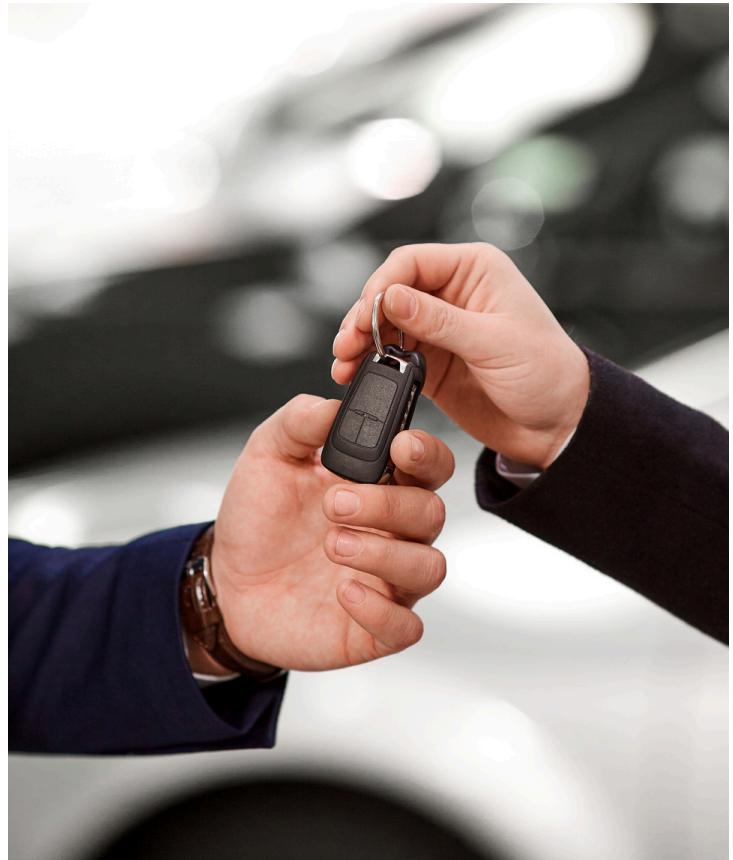
19085

Finalne RMSE

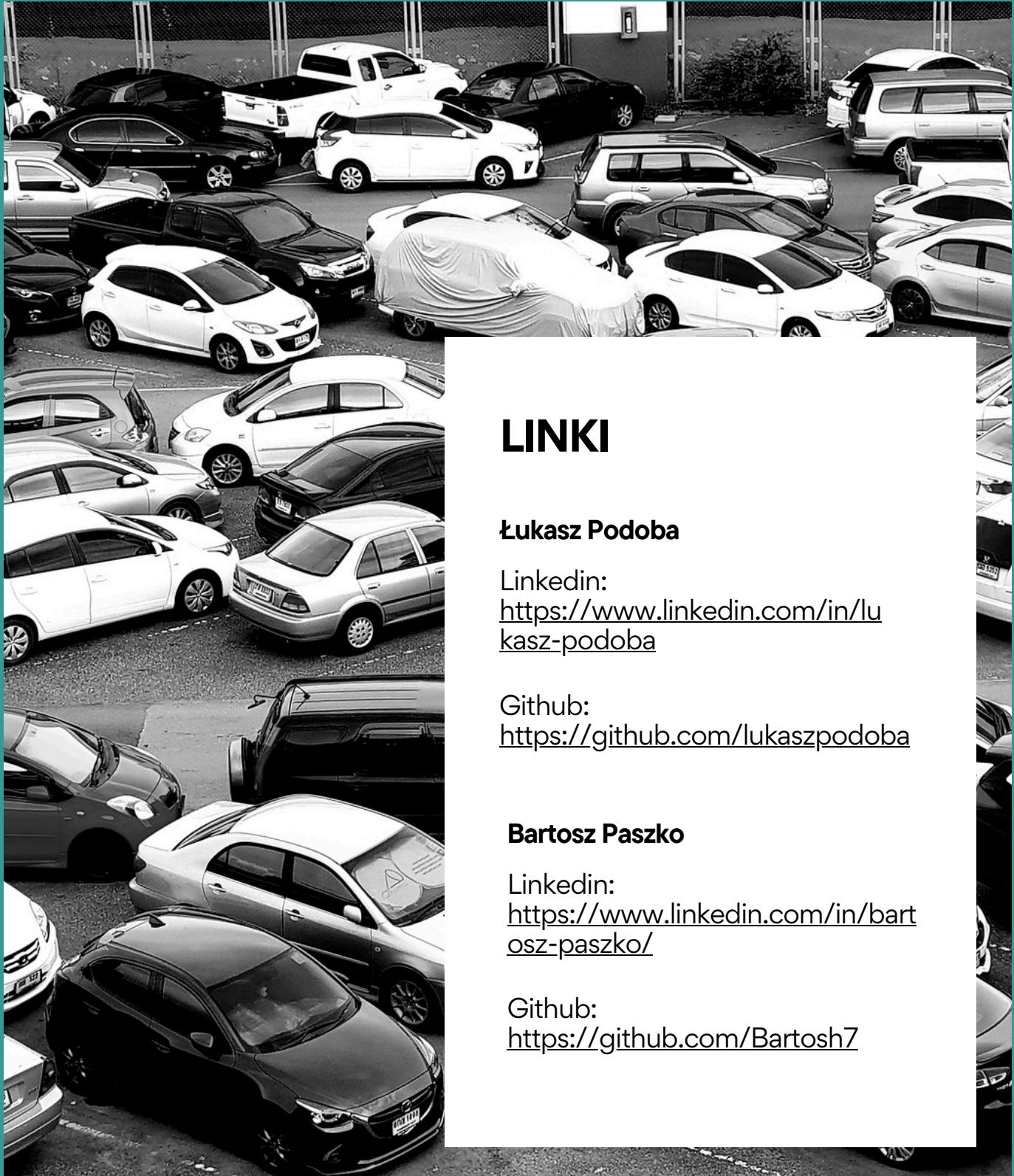
PODSUMOWANIE

NAJWAŻNIEJSZE INFORMACJE

- Cel projektu: Stworzenie modelu predykcji cen samochodów na podstawie rzeczywistych danych ofert sprzedaży (web scraping).
- Wybrany model: XGBoost, zoptymalizowany biblioteką Optuna, zapewnił wysoką skuteczność predykcji (niski RMSE).
- Kluczowe wnioski: Najważniejszą cechą wpływającą na cenę pojazdu jest rok produkcji, a istotną rolę odgrywają również generacja, wyposażenie, typ skrzyni biegów oraz stan pojazdu.
- Uwagi: Wykorzystanie syntetycznych danych nie poprawiło wyników z powodu błędnych kombinacji cech; znaczną poprawę dało skalowanie cen, kodowanie zmiennych kategorycznych (target encoding) oraz feature engineering



DZIĘKUJEMY ZA UWAGĘ



LINKI

Łukasz Podoba

Linkedin:

<https://www.linkedin.com/in/lukasz-podoba>

Github:

<https://github.com/lukaszpodoba>

Bartosz Paszko

Linkedin:

<https://www.linkedin.com/in/bartosz-paszko/>

Github:

<https://github.com/Bartosh7>