

# Analiza i prognozowanie cen samochodów używanych na aukcjach z wykorzystaniem regresji

Zuzanna Jużyniec  
[zuzanna.juzyniec@pk.edu.pl](mailto:zuzanna.juzyniec@pk.edu.pl)

Rafał Matusiak  
[rafal.matusiak@pk.edu.pl](mailto:rafal.matusiak@pk.edu.pl)

November 17, 2024

## 1 Abstract

## 2 Wprowadzenie

## 3 Zbiór danych

Zbiór danych został pobrany ze strony <https://www.kaggle.com/datasets/tunguz/used-car-auction-prices?resource=download> i zawiera historyczne ceny sprzedaży samochodów na aukcjach, zebrane z zewnętrznych źródeł internetowych. Został zebrany w 2015 roku i nie był już aktualizowany.

Zbiór danych zawiera następujące kolumny:

- year - rok produkcji pojazdu;
- make - marka pojazdu (np. Toyota, Ford);
- model - model pojazdu (np. Corolla, Mustang);
- trim - wersja wyposażenia pojazdu (np. SE, Base);
- body - rodzaj nadwozia pojazdu (np. sedan, SUV);
- transmission - rodzaj skrzyni biegów (np. automatyczna, manualna);
- vin - numer identyfikacyjny pojazdu (VIN);
- state - stan (lokalizacja geograficzna) w USA, gdzie odbyła się sprzedaż;
- condition - stan pojazdu oceniany w skali od 1 do 5, gdzie wyższe wartości oznaczają lepszy stan techniczny (zmienna ciągła);
- odometer - przebieg pojazdu w milach;
- color - kolor karoserii pojazdu;
- interior - kolor wnętrza pojazdu;
- seller - nazwa lub identyfikator sprzedającego pojazd;
- mmr - wartość pojazdu według Manheim Market Report (MMR), używana jako wskaźnik wartości rynkowej;
- sellingprice - rzeczywista cena sprzedaży pojazdu na aukcji;
- saledate - data sprzedaży pojazdu.

W celu poznania danych zostało wykonane wiele wykresów.

### 3.1 Wykres procentowej ilości brakujących wartości

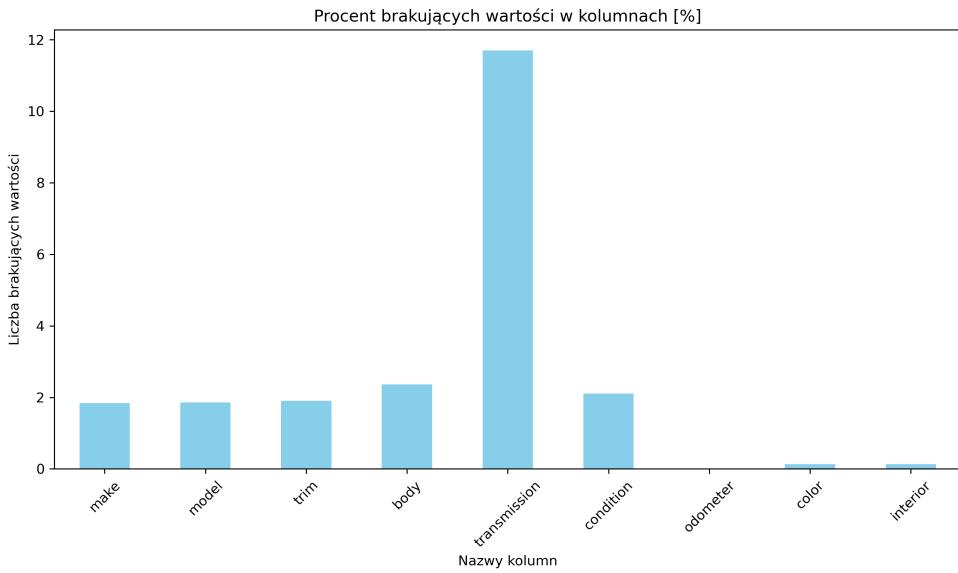


Figure 1: Histogramy zmiennych numerycznych.

Jak widać wartości NaN występują w 9 kolumnach, najczęściej w zmiennej 'transmission'.

### 3.2 Histogramy

Jak widać na Rys. 1, niektóre zmienne, takie jak sellingprice, mmr oraz odometer, charakteryzują się rozkładem długogonowym, co oznacza, że większość wartości koncentruje się w niższych przedziałach, jednak sporadycznie występują bardzo wysokie wartości odstające. Taki rozkład sugeruje, że typowe ceny sprzedaży, wartości rynkowe oraz przebiegi są zazwyczaj niskie, choć zdarzają się pojazdy o wyjątkowo wysokich cenach i dużym przebiegu.

Aby lepiej zobrazować te rozkłady, zastosowano przekształcenie logarytmiczne, co przedstawiono na Rys. 2 — po tej transformacji rozkłady stały się bardziej symetryczne i zbliżone do normalnego, co ułatwiało ich interpretację. Zmienna condition ma bardziej równomierny rozkład, wskazując na stabilny stan techniczny większości pojazdów, natomiast year wykazuje tendencję przesunięcia ku nowszym rocznikom, co sugeruje, że na aukcjach dominują młodsze samochody. Podczas dalszej analizy wyniknie, czy takie przekształcenie będzie konieczne w procesie modelowania, aby poprawić wyniki prognozowania, jednak na tym etapie pozostajemy przy badaniu oryginalnych danych.

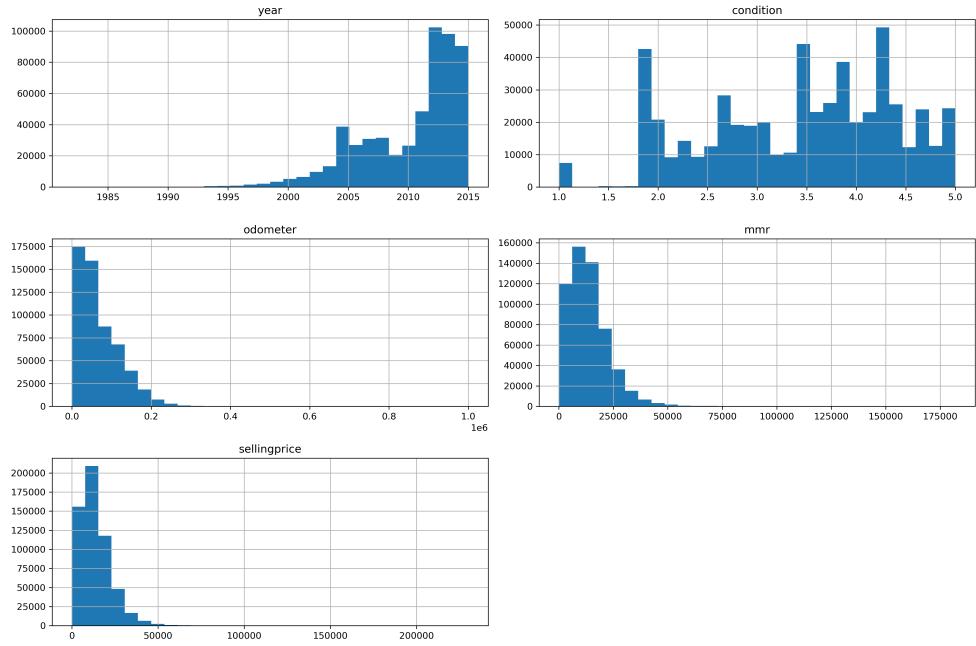


Figure 2: Histogramy zmiennych numerycznych.

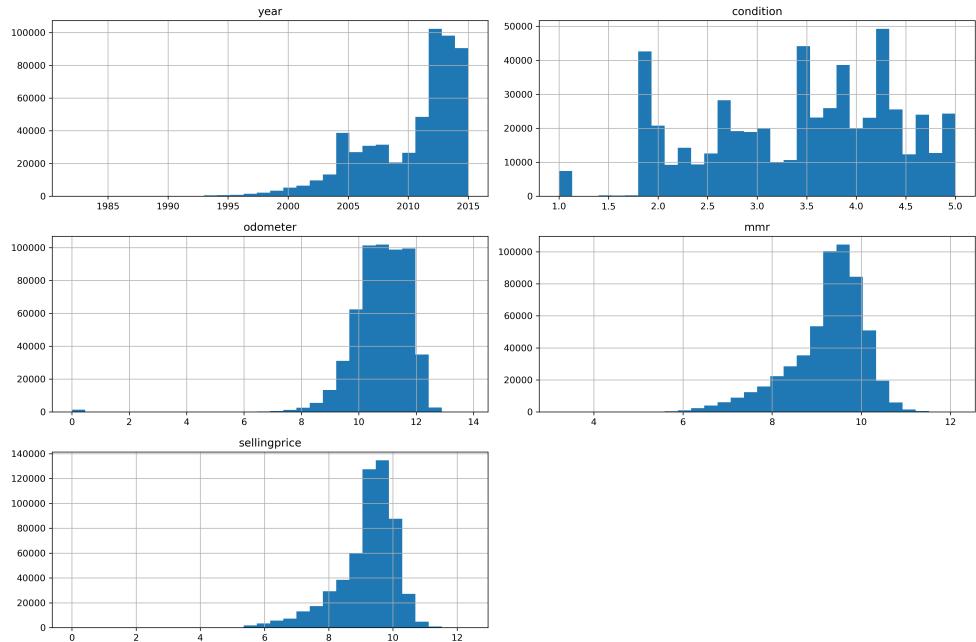


Figure 3: Histogramy zmiennych numerycznych po przekształceniu logarytmicznym.

### 3.3 Boxploty

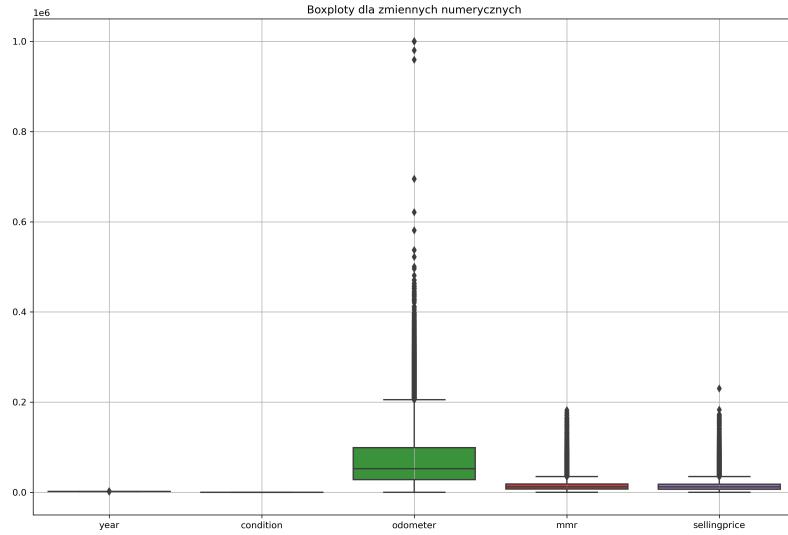


Figure 4: Boxploty dla wszystkich zmiennych numerycznych.

Na powyższym wykresie (Rys. 4) przedstawiono boxploty dla zmiennych numerycznych, jednak ze względu na dużą różnicę w wartościach pomiędzy zmiennymi, szczegóły niektórych cech są trudne do dostrzeżenia. Aby lepiej zobrazować rozkład wartości, w tym wartości odstające, zostaną przedstawione oddzielne boxploty dla każdej z cech numerycznych. Dodatkowo, wykresy te zostaną wygenerowane zarówno z uwzględnieniem danych odstających, jak i bez nich, aby zapewnić pełniejszą analizę danych.

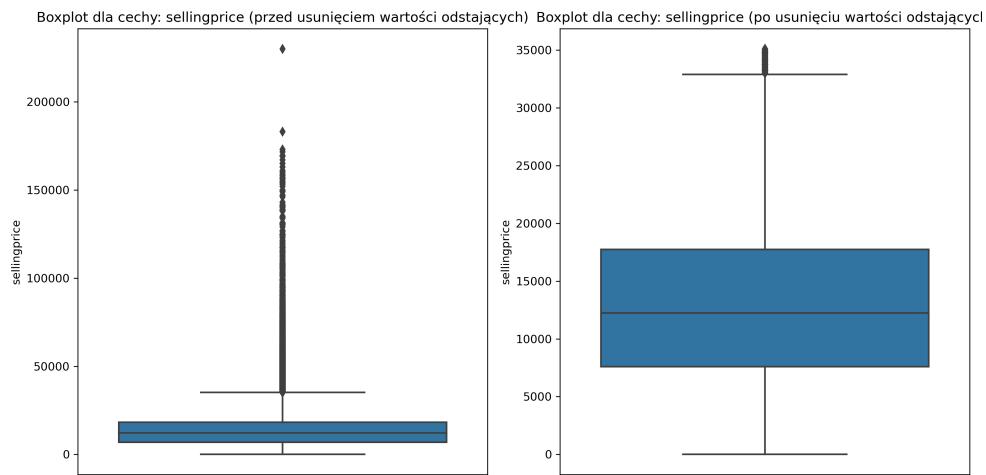


Figure 5: Boxploty dla zmiennej 'sellingprice'.

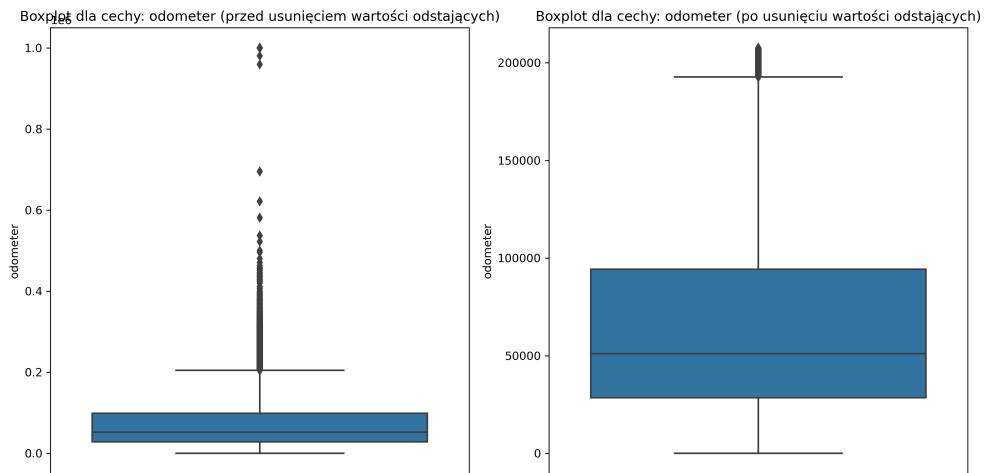


Figure 6: Boxploty dla zmiennej 'odometer'.

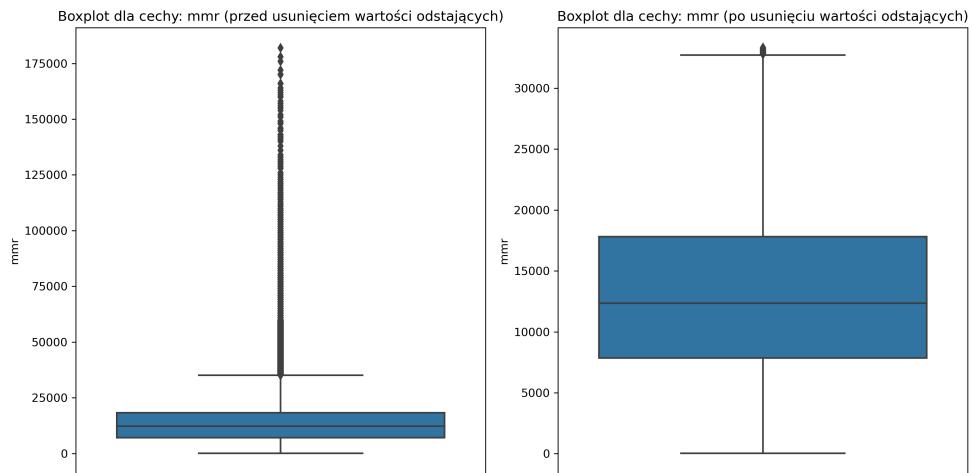


Figure 7: Boxploty dla zmiennej 'mmr'.

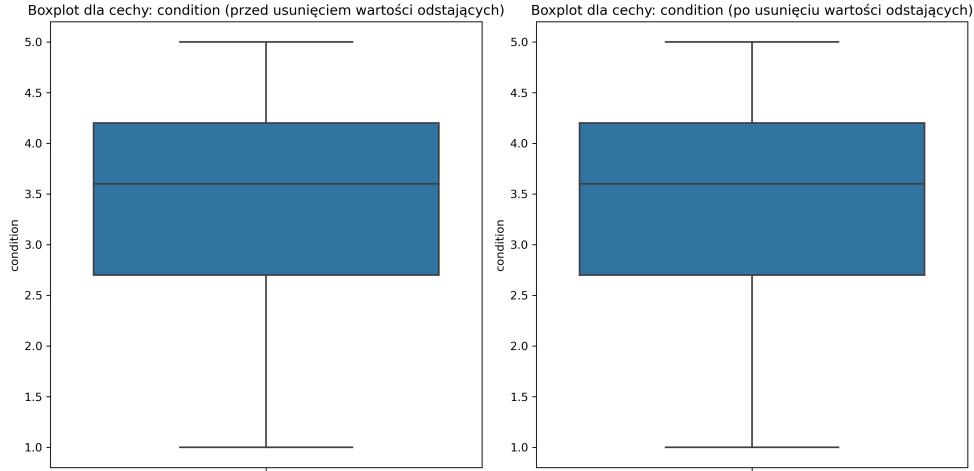


Figure 8: Boxploty dla zmiennej 'condition'.

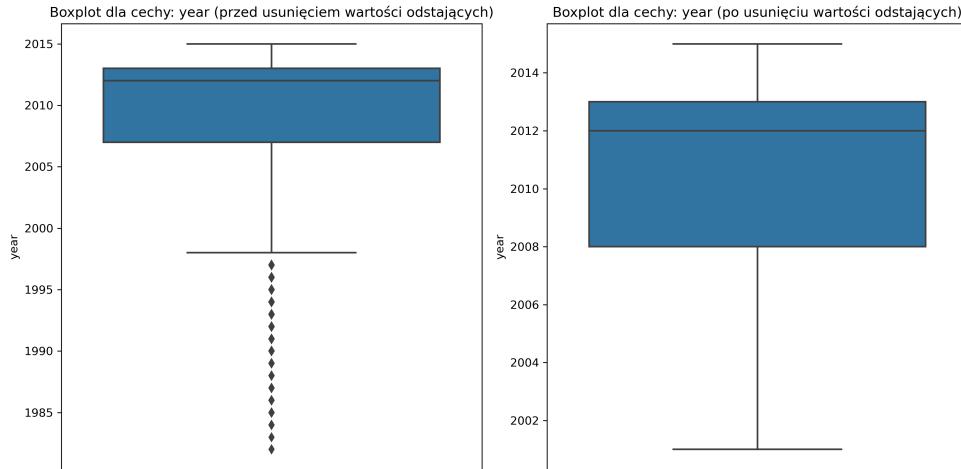


Figure 9: Boxploty dla zmiennej 'year'.

Wartości odstające mają największy wpływ na zmienne takie jak cena sprzedaży (sellingprice), przebieg (odometer) i wartość rynkowa (mmr), które charakteryzują się nielicznymi, ale bardzo skrajnymi wartościami odstającymi. Po ich usunięciu uzyskujemy bardziej klarowny obraz typowego rozkładu danych, co pozwala lepiej zrozumieć, jakie są rzeczywiste wartości dominujące w zbiorze danych.

W przypadku zmiennych rok produkcji (year) i stan techniczny (condition), wartości odstające mają mniejszy wpływ. Boxplot dla stanu technicznego jest bardzo jednorodny, co wskazuje, że ta zmienna nie zawiera wielu ekstremalnych wartości, a typowy stan techniczny pojazdów waha się od 2.5 do 4.5.

### 3.4 Pairplot

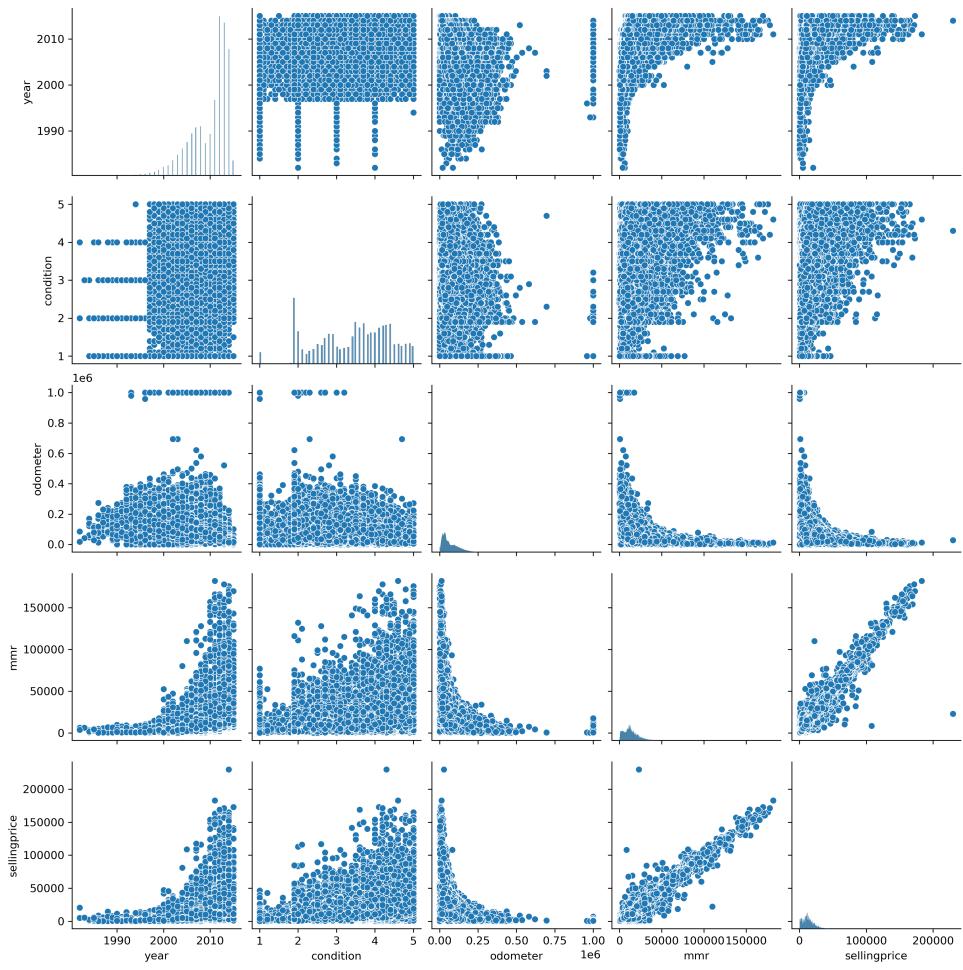


Figure 10: Pairplot.

Patrząc na pairplot, widzimy wyraźne zależności między niektórymi zmiennymi (np. cena sprzedaży a rok produkcji, przebieg czy wartość rynkowa). Aby lepiej zrozumieć, jak silne są te zależności, warto obliczyć współczynniki korelacji dla tych zmiennych. Pozwoli to na dokładniejszą ocenę siły tych relacji.

### 3.5 Macierz korelacji

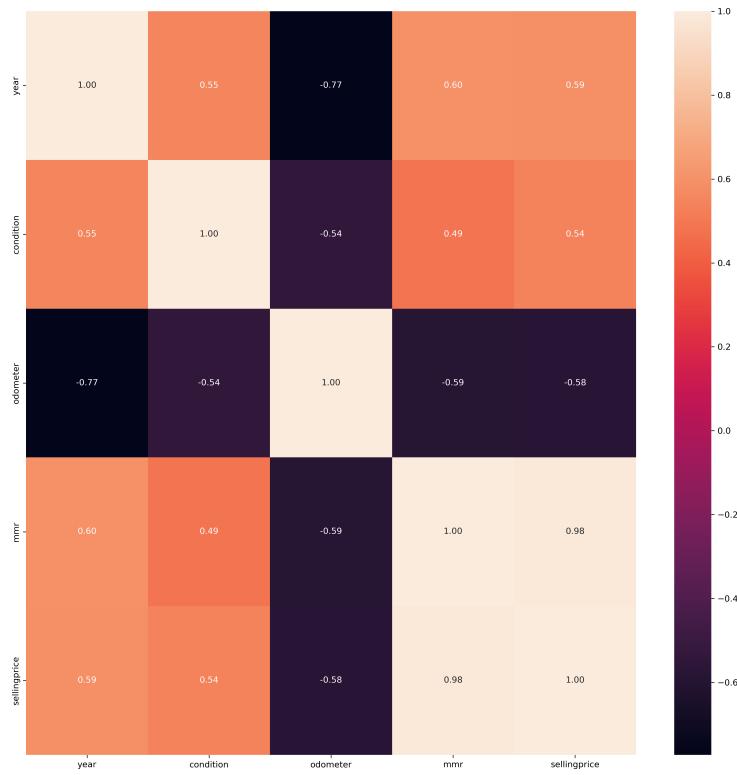


Figure 11: Macierz korelacji.

Rok produkcji (year) ma silną dodatnią korelację z ceną sprzedaży (0.59), co oznacza, że nowsze samochody zazwyczaj sprzedają się za wyższą cenę. Stan techniczny (condition) ma dodatnią korelację z ceną sprzedaży (0.54), co sugeruje, że samochody w lepszym stanie są droższe. Przebieg (odometer) wykazuje silną ujemną korelację z ceną sprzedaży (-0.58), co oznacza, że większy przebieg zmniejsza wartość samochodu. MMR (wartość rynkowa) ma bardzo silną dodatnią korelację z ceną sprzedaży (0.98), co wskazuje, że wartość rynkowa samochodu dobrze prognozuje jego cenę na aukcji.

### 3.6 Wykres sprzedaży samochodów - Cena, Przebieg i Stan pojazdu dla 1000 losowych próbek

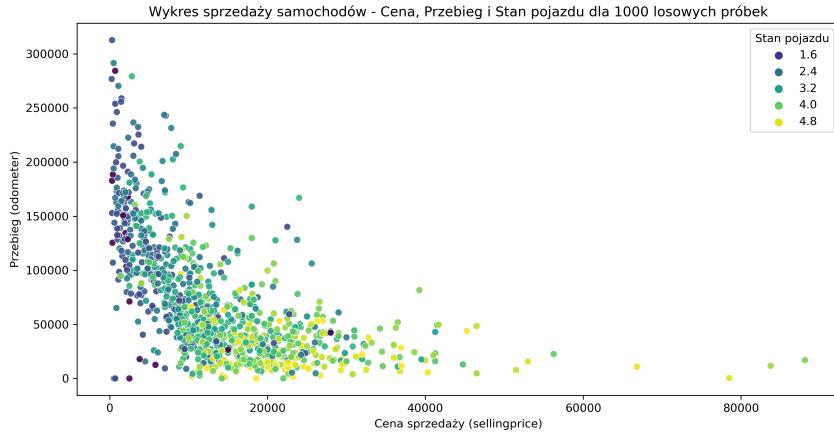


Figure 12: Wykres sprzedaży samochodów.

Na powyższym wykresie zbadano zależność między przebiegiem pojazdów a ceną sprzedaży, uwzględniając jednocześnie ich stan techniczny. Zauważalna jest silna ujemna zależność – samochody z większym przebiegiem są zazwyczaj tańsze. Lepszy stan techniczny pojazdu wpływa pozytywnie na cenę sprzedaży, co pozwala niektórym pojazdom osiągnąć wyższe ceny, nawet przy relatywnie dużym przebiegu. Najwyższe ceny uzyskują samochody w najlepszym stanie, podczas gdy pojazdy w najgorszym stanie sprzedają się za niższe kwoty, niezależnie od przebiegu.

### 3.7 Wykres sprzedaży samochodów - Cena, Przebieg i Rok produkcji pojazdu dla 1000 losowych próbek

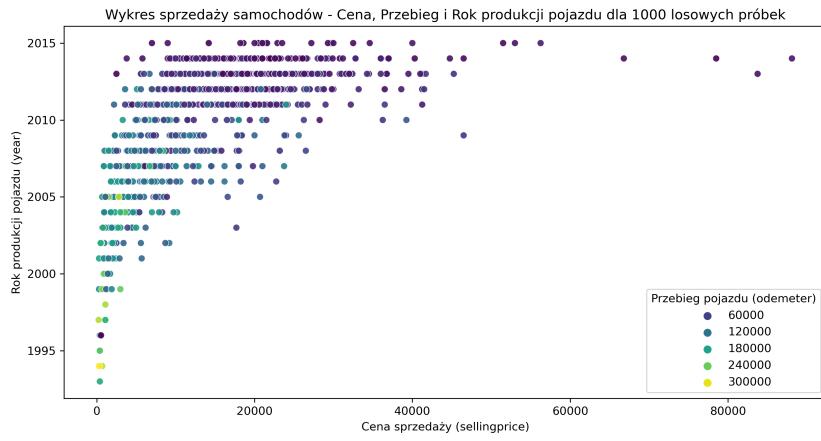


Figure 13: Wykres sprzedaży samochodów.

Wykres przedstawia zależność między rokiem produkcji a ceną sprzedaży samochodów, uwzględniając jednocześnie ich przebieg. Zauważalna jest wyraźna tendencja – nowsze samochody (z lat 2010–2015) osiągają wyższe ceny, podczas gdy starsze pojazdy (sprzed 2005 roku) sprzedają się za znacznie niższe kwoty. Przebieg również ma istotny wpływ – samochody

z mniejszym przebiegiem mają tendencję do wyższych cen, zwłaszcza w przypadku nowszych roczników. Samochody z większym przebiegiem, niezależnie od roku produkcji, zazwyczaj mają niższą wartość.

### 3.8 Liczba sprzedanych pojazdów w poszczególnych stanach USA

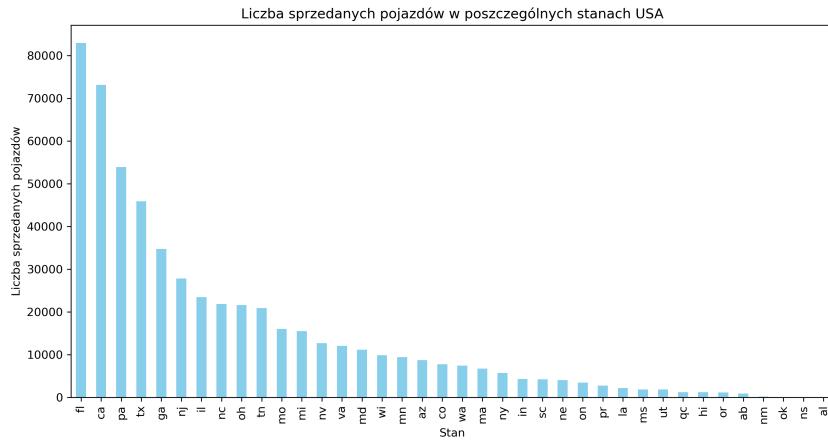


Figure 14: Liczba sprzedanych pojazdów w poszczególnych stanach USA.

Największą liczbę sprzedanych pojazdów odnotowano w stanach Floryda (FL) oraz Kalifornia (CA). Te dwa stany zdecydując wyróżniają się jako liderzy. W czołówce są także Teksas (TX), Pensylwania (PA) oraz Georgia (GA).

### 3.9 Top 10 najczęściej sprzedawanych marek samochodów

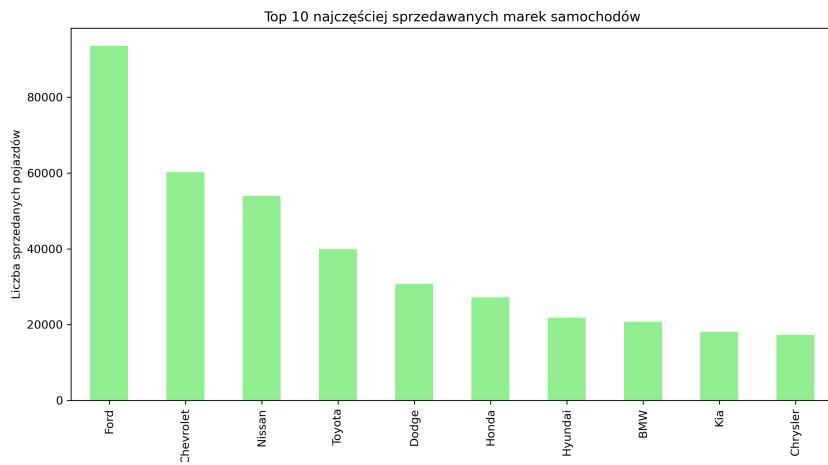


Figure 15: Top 10 najczęściej sprzedawanych marek samochodów

Ford zdecydując dominuje jako najczęściej sprzedawana marka, z wyraźną przewagą nad Chevroletem oraz Nissanem. Toyota oraz Dodge zamykają czołową piątkę. Wynik ten może odzwierciedlać popularność tych marek na rynku amerykańskim, gdzie dominują pojazdy amerykańskie i azjatyckie.

### 3.10 Liczba sprzedanych pojazdów według roku produkcji

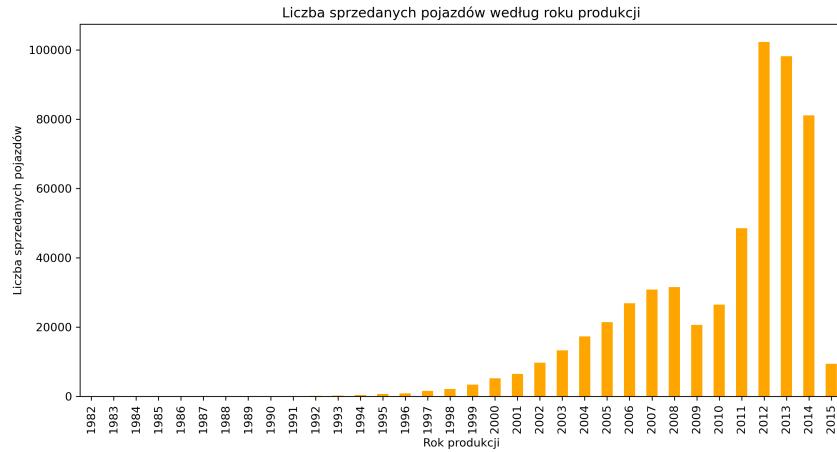


Figure 16: Liczba sprzedanych pojazdów według roku produkcji

Największa liczba sprzedanych pojazdów pochodzi z lat 2011-2014, co sugeruje, że rynek używanych samochodów obejmuje głównie nowsze modele, które mają jeszcze sporo lat eksploatacji przed sobą. Starsze roczniki, szczególnie te sprzed 2000 roku, są rzadziej spotykane, co jest zgodne z trendem zmniejszania się podaży starszych aut.

### 3.11 Top 10 najlepiej sprzedających się typów nadwozia

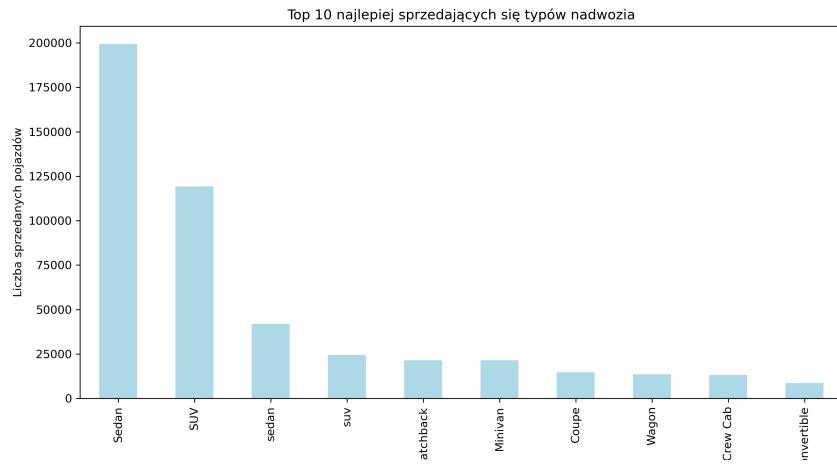


Figure 17: Top 10 najlepiej sprzedających się typów nadwozia

Sedan jest zdecydowanie najpopularniejszym typem nadwozia, z liczbą sprzedanych egzemplarzy wyraźnie przewyższającą inne typy. SUV zajmuje drugie miejsce, co świadczy o rosnącej popularności pojazdów tego segmentu. Inne typy nadwozia, takie jak hatchback czy minivan, również znajdują się w czołówce, ale z mniejszą liczbą sprzedanych egzemplarzy.

### 3.12 Top 10 kolorów z największą liczbą sprzedanych pojazdów

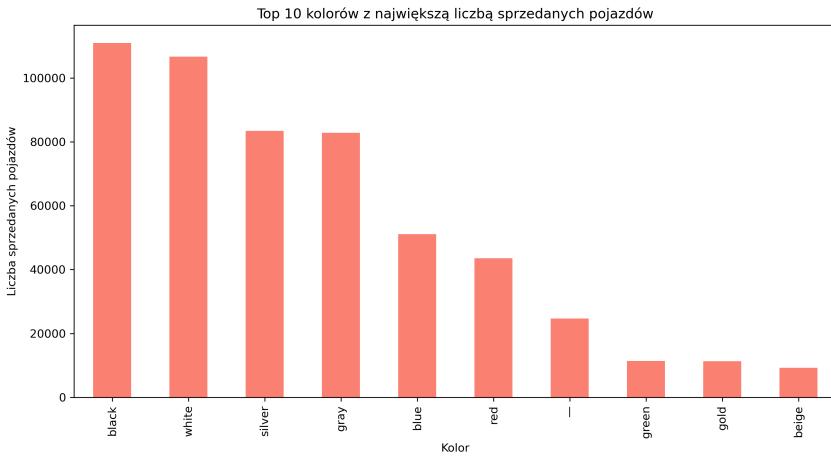


Figure 18: Top 10 kolorów z największą liczbą sprzedanych pojazdów

Kolory takie jak czarny, biały i srebrny są najczęściej wybierane przez nabywców, co nie jest zaskoczeniem, biorąc pod uwagę ich popularność na rynku. Są to kolory neutralne, które zazwyczaj lepiej utrzymują swoją wartość na rynku wtórnym. Inne kolory, takie jak niebieski i czerwony, również znajdują się w top 10, ale w mniejszych ilościach.

## 4 Przegląd literatury

Zbiór nie był wykorzystywany w publikacjach. Dostępne jest kilka notatek z przykładowymi podejściami do problemu.

- Notatnik Kaggle autorstwa Zabihullah18 dotyczy predykcji cen samochodów za pomocą uczenia maszynowego. Zawiera kroki takie jak wstępne przetwarzanie danych, wybór cech i trenowanie modeli. Używa algorytmu regresji liniowej do przewidywania cen na podstawie danych o samochodach (np. marka, model, przebieg). Wyniki są oceniane przy użyciu wskaźników takich jak błąd średniokwadratowy, który pozwala mierzyć dokładność prognoz. Pełen notatnik dostępny jest pod adresem: <https://www.kaggle.com/code/zabihullah18/car-price-prediction---5---Train-the-Model->
- Notatnik na Kaggle dotyczący predykcji cen używanych samochodów autorstwa Midael3ila wykorzystuje techniki uczenia maszynowego do przewidywania cen samochodów na podstawie ich cech. Proces obejmuje wstępne przetwarzanie danych, eksplorację cech takich jak marka, model, przebieg i wiek pojazdu, a następnie budowanie modeli predykcyjnych. Algorytmy takie jak regresja liniowa i lasy losowe są stosowane do trenowania modelu. Notatnik ocenia dokładność przewidywań za pomocą miar takich jak błąd średniokwadratowy. Pełen notatnik dostępny jest pod adresem: <https://www.kaggle.com/code/midael3ila/car-used-priced-prediction-using-ml>

## 5 Motywacja

Projekt dotyczący prognozowania cen samochodów używanych jest interesujący, ponieważ może przynieść korzyści zarówno kupującym, jak i sprzedającym, pomagając przewidzieć realistyczne ceny rynkowe. W społeczeństwie, gdzie kupno używanych pojazdów jest powszechnie, dokładne modele predykcyjne mogą wspierać transparentność i zaufanie w transakcjach. W przemyśle, takie prognozy mogą być użyteczne dla wypożyczalni samochodów, dealerów oraz platform aukcyjnych, pomagając w lepszym oszacowaniu wartości pojazdów i optymalizacji strategii sprzedaży.

## 6 Ewaluacja

Oczekiwany model będzie w stanie dokładnie przewidzieć ceny samochodów używanych na aukcjach na podstawie dostępnych danych. Satysfakcjonujący wynik to taki, w którym błąd predykcji (np. błąd średniokwadratowy) będzie niski, co

oznacza, że przewidywane ceny są bliskie rzeczywistym cenom sprzedaży. Osiągniety wynik, powinien być wystarczająco precyzyjny, aby umożliwić zastosowanie modelu w rzeczywistych warunkach — np. na platformach aukcyjnych lub w firmach dealerskich — gdzie może pomóc w wycenie pojazdów. Oczywiście projekt ma charakter wyłącznie edukacyjny i nie jest przeznaczony do komercyjnego wykorzystania ani sprzedaży. Celem jest nauka i zrozumienie metod prognozowania cen samochodów używanych z wykorzystaniem modeli uczenia maszynowego. Chociaż dążymy do uzyskania dokładnych wyników, system ten nie będzie wdrażany w żadnym rzeczywistym środowisku komercyjnym. Skupiamy się na poszerzaniu wiedzy i umiejętności w zakresie analizy danych oraz budowania modeli predykcyjnych.

## 7 Zasoby

W projekcie zostanie wykorzystany język programowania Python oraz środowisko Jupyter Notebook do analizy danych i budowy modeli predykcyjnych. Python oferuje bogaty zestaw bibliotek do uczenia maszynowego, takich jak scikit-learn do budowy i oceny modeli, pandas do manipulacji danymi, oraz numpy do obliczeń numerycznych. Dodatkowo, zasadniczo wykorzystana biblioteka matplotlib i seaborn do wizualizacji danych, co pozwoli na lepsze zrozumienie zależności i wzorców w zebranych informacjach.

## 8 Zastosowane metody

## 9 Eksperyment

### 9.1 Preprocessing

W celu przygotowania danych do modelowania i analizy predykcyjnej, przeprowadzono zaawansowany preprocessing, który obejmował różne etapy mające na celu poprawienie jakości i spójności zbioru danych. Szczegółowe kroki opisano poniżej.

**1. Obliczenie wieku samochodu (car\_age)** Wiek samochodu w momencie sprzedaży został obliczony na podstawie kolumny `saledate` poprzez odjęcie od roku sprzedaży (`sale_year`) roku produkcji pojazdu (`year`). Dodatkowo, z tej samej kolumny (`saledate`) wyodrębniono dodatkowe cechy, takie jak:

- Miesiąc sprzedaży (`sale_month`),
- Rok sprzedaży (`sale_year`),
- Godzina sprzedaży (`sale_hour`),
- Minuta sprzedaży (`sale_minute`).

Po tej transformacji przeanalizowano sprzedaż samochodów w różnych miesiącach i dniach tygodnia. Wyniki wskazują, że najczęściej sprzedaży miało miejsce w styczniu (140,815 samochodów) oraz lutym (163,054 samochody). Z kolei analiza sprzedaży według dnia tygodnia wykazała, że transakcje najczęściej odbywają się na początku tygodnia, zwłaszcza we wtorki (180,158 sprzedaży), natomiast najmniej w niedziele (11,868 sprzedaży).

Podczas analizy wieku pojazdów wykryto, że 201 rekordów miało ujemne wartości w kolumnie `car_age`. Można zatem założyć, że taka liczba samochodów została zakupiona w przedsprzedaży.

**2. Ekstrakcja informacji z numeru VIN** Z numeru VIN pojazdu przy użyciu biblioteki `vininfo` wyodrębniono informacje o kraju produkcji (`vin_country`). Na tym etapie zrezygnowano z pozyskiwania dodatkowych danych, takich jak typ pojazdu, ponieważ były one już zawarte w innych kolumnach (`model`, `body`). Po wyodrębnieniu kraju, kolumna `vin` została usunięta, gdyż nie była już potrzebna do dalszej analizy.

**3. Usunięcie wartości odstających** W celu usunięcia wartości odstających przeprowadzono analizę cech numerycznych (`year`, `condition`, `odometer`, `mmr`, `sellingprice`) za pomocą metody IQR (Interquartile Range). Wartości spoza zakresu  $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$  zostały usunięte. Łącznie usunięto około 8.3% danych, co było akceptowalne, ponieważ nie przekraczało 10% zbioru.

**4. Przekształcenie zmiennych kategorycznych na wartości liczbowe** Wszystkie zmienne kategoryczne (make, model, trim, body, transmission, state, color, interior, seller, vin\_country) zostały przekonwertowane na wartości liczbowe. Dla kolumny transmission zastosowano kodowanie binarne: 0 dla manualnej skrzyni biegów oraz 1 dla automatycznej. Dla pozostałych kolumn zastosowano "label encoding" co w kolejnych etapach może ograniczyć możliwość wykorzystania niektórych modeli. Przy tak dużych danych i wielu kategoriach niemożliwe było zastosowanie one-hot encodingu.

**5. Sprawdzenie brakujących wartości** Po zakończeniu wcześniejszych kroków przeanalizowano brakujące wartości w każdej z kolumn. Wyniki prezentowały się następująco:

year	0.000000%
make	1.843378%
model	1.860915%
trim	1.906011%
body	2.361263%
transmission	11.695010%
state	0.000000%
condition	2.110553%
odometer	0.016821%
color	0.134035%
interior	0.134035%
seller	0.000000%
mmr	0.000000%
sellingprice	0.000000%
car_age	0.000000%
sale_year	0.000000%
sale_month	0.000000%
sale_hour	0.000000%
sale_minute	0.000000%
sale_weekday	0.000000%
vin_country	0.000000%

**6. Uzupełnianie brakujących wartości** Do imputacji brakujących danych zastosowano trzy metody:

- Średnia dla danych numerycznych oraz moda dla danych kategorycznych.
- Mediana dla danych numerycznych oraz moda dla danych kategorycznych.
- Imputacja metodą k-Nearest Neighbors (KNNImputer) z 3 sąsiadami.

**7. Skalowanie i standaryzacja danych** W celu ujednolicenia wartości numerycznych zastosowano dwa podejścia:

- Skalowanie Min-Max w zakresie [0, 1]
- Standaryzacja do średniej 0 i odchylenia standardowego 1

Dla każdego z trzech zbiorów (imputacja średnią, medianą, KNN) zastosowano skalowanie i standaryzację, co doprowadziło do utworzenia sześciu różnych zestawów danych:

```
mean_min_max.csv  
median_min_max.csv  
knn_min_max.csv  
mean_standard.csv  
median_standard.csv  
knn_standard.csv
```

**8. Zapis przetworzonych danych** Wszystkie powyższe zbiory zostały zapisane jako pliki CSV. Zbiory te są gotowe do użycia w modelach predykcyjnych i zapewniają lepszą jakość danych, co przekłada się na większą dokładność modeli.

9.2 Trening modelu

9.3 Optymalizacja

10 Podsumowanie

References