Analiza danych

Dataset zawiera 10 kolumn i 91523 wierszy, nie posiada brakujących wartości.

Kolumny:

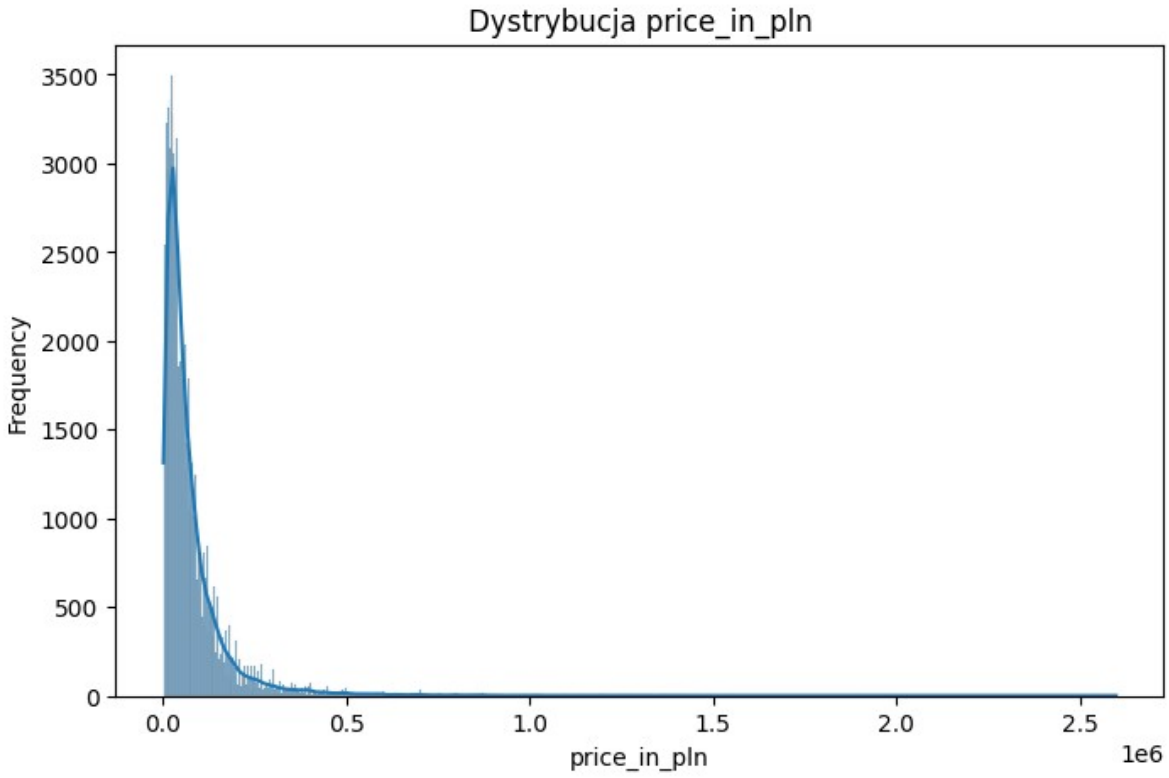
1. brand – marka samochodu
2. model – nazwa konkretnego modelu
3. price_in_pln – cena w PLN
4. mileage – przebieg w kilometrach (np. „133 760 km”)
5. gearbox – rodzaj skrzyni biegów, manualna lub automatyczna
6. engine_capacity – pojemność silnika w centymetrach sześciennych (np. „1970 cm3”)
7. fuel_type – rodzaj paliwa (np. „Benzyna”, „Benzyna+LPG”, „Diesel”)
8. city – miejscowość wystawienia oferty
9. voivodship – województwo wystawienia oferty
10. year – rok produkcji samochodu

4 kolumny to kolumny numeryczne – price_in_pln, mileage, engine_capacity, year.
2 kolumny to kolumny tekstowe – model, city
4 kolumny to kolumny kategoryczne – brand, gearbox, fuel_type, voivodship

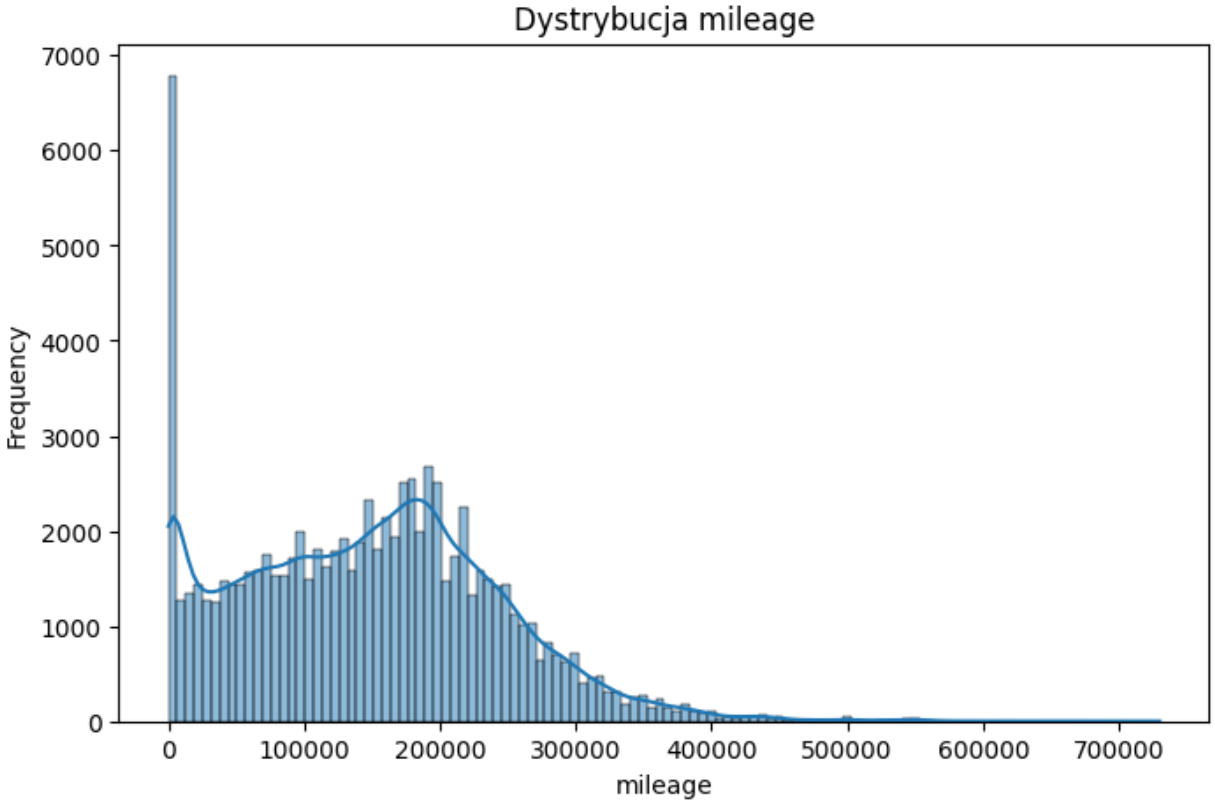
Po wstępnym formatowaniu kolumn okazało się że dataset zawiera wiersze w których dane w poszczególnych kolumnach są pomieszane. Po ich usunięciu zostało 85677 wierszy.

Analiza zmiennych

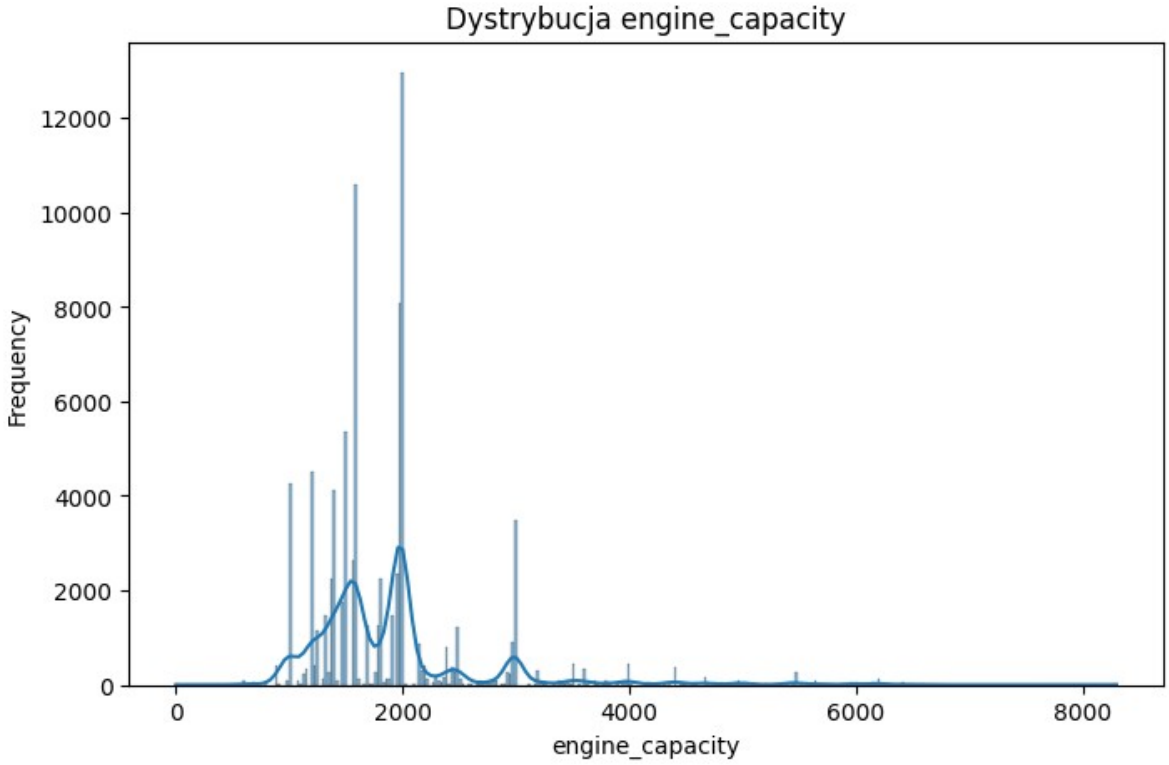
- price_in_pln – ceny wahają się od 1111 do 2 599 000. Mediana to 46 500, ponad 75% cen jest poniżej 100 000 złotych.



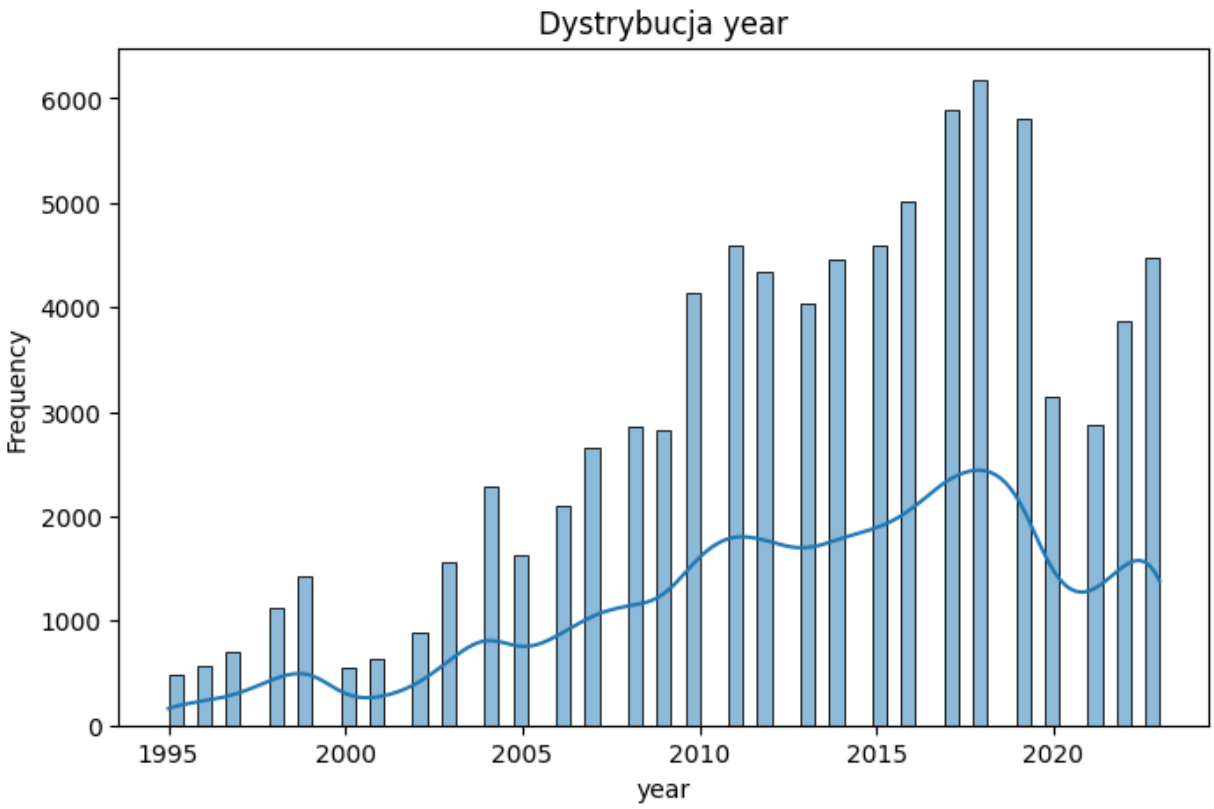
- mileage



- engine_capacity

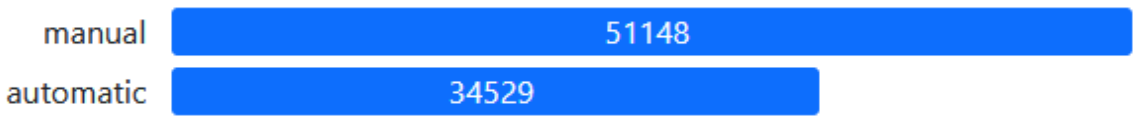


- year

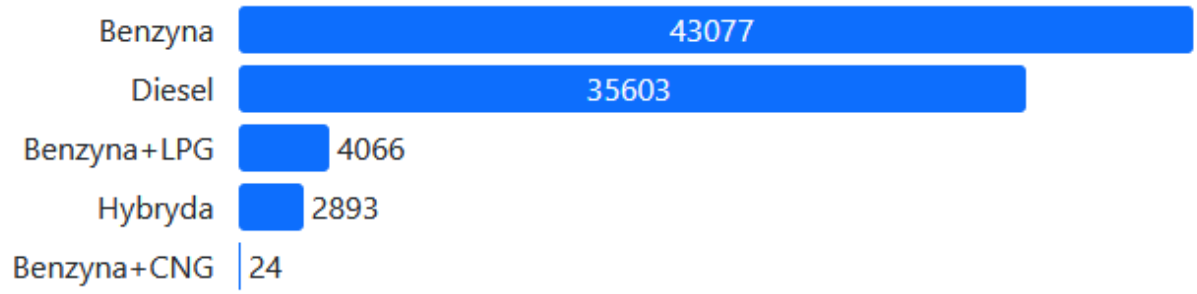


- brand – 43 unikalne wartości

- model – 18021 unikalnych wartości
- gearbox

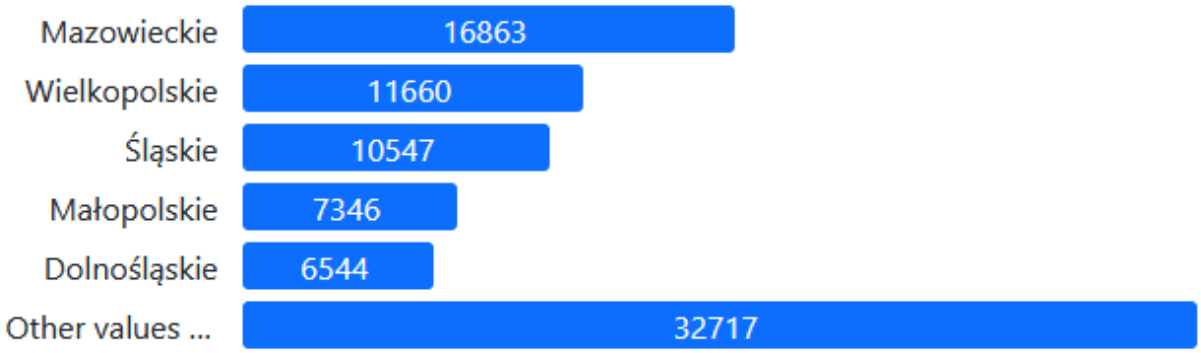


- fuel_type – 6 unikalnych wartości, wyraźna większośc to benzyna lub diesel

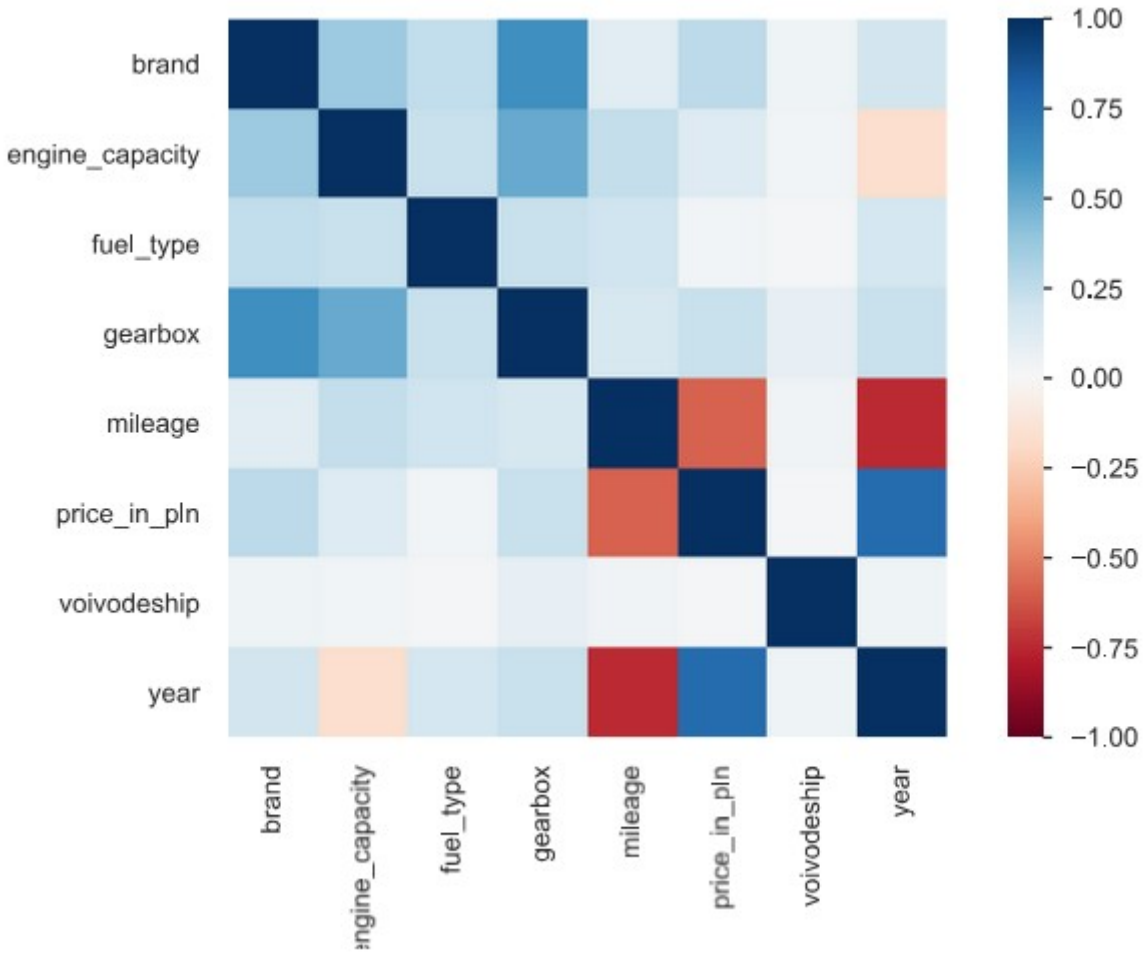


- city – 4353 unikalnych wartości, najpopularniejsze to duze miasta wojewódzkie jak Warszawa, Kraków czy Wrocław

- voivodship - 25 unikalnych wartości???- trzeba sprawdzić



Macierz korelacji



Największy wpływ na cenę samochodu mają przebieg i rok produkcji.

Marka samochodu ma całkiem wysoki wpływ na rodzaj skrzyni biegów.

Wiek samochodu ma bardzo wysoki wpływ na jego przebieg.