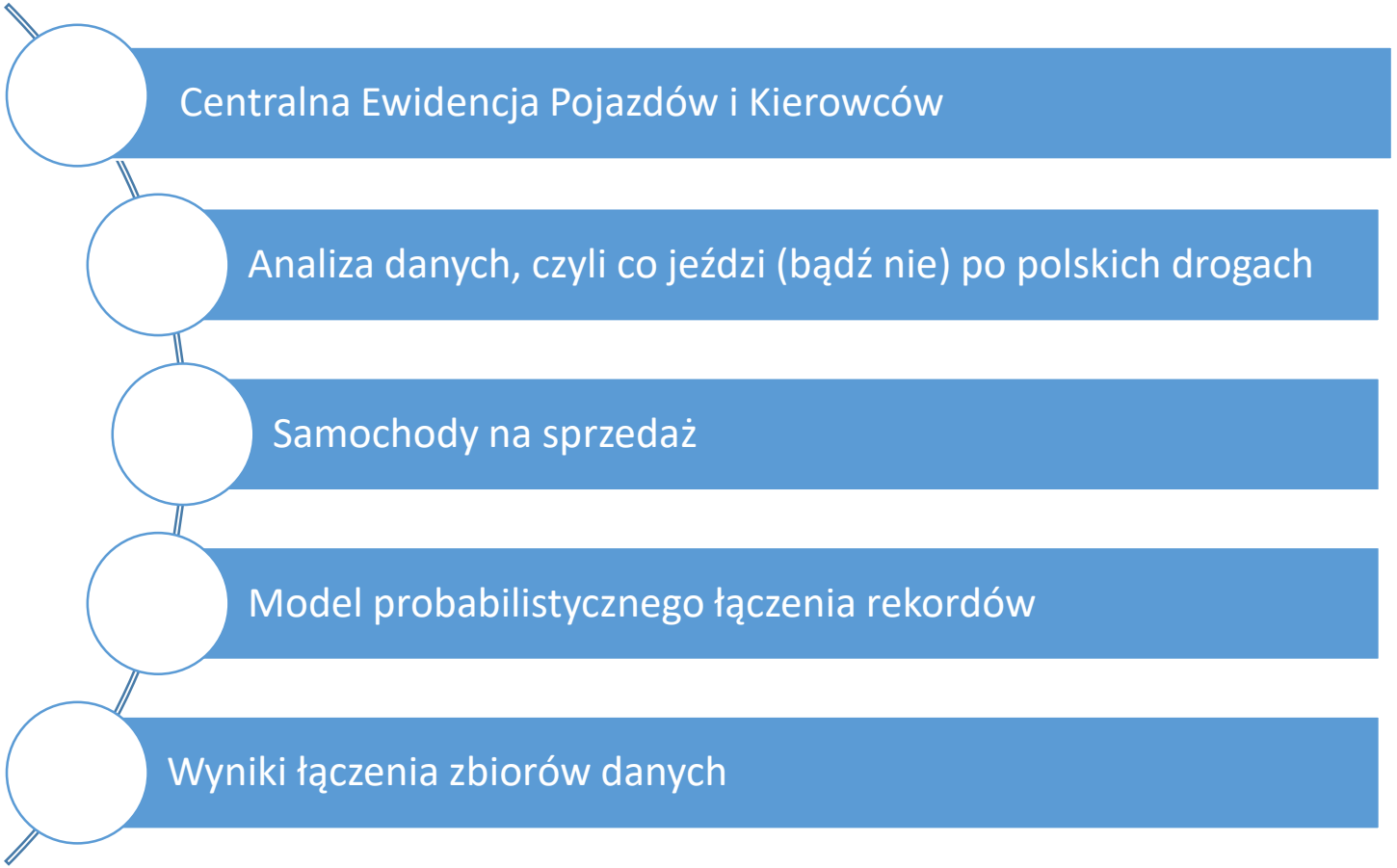
Rynek samochodów osobowych w Polsce

Agnieszka Giemza
specjalista
dr Sebastian Wójcik
kierownik działu
Dział Statystyki Matematycznej

Warszawa, 2-4.07.2024

Plan prezentacji



Wprowadzenie

- W ramach projektu Eurostatu pt. *Modernizacja Statystyki Energii*, jednym z naszych zadań jest analiza danych dotyczących samochodów osobowych poruszających się po polskich drogach oraz emisji zanieczyszczeń przez nie generowanych.
- Zaprezentujemy wstępne wyniki analizy rejestru CEPIK i ofert sprzedaży samochodów dostępnych na portalach motoryzacyjnych oraz wyniki połączenia obydwu zbiorów.

CEPIK

CEPIK integruje dane pochodzące z różnych źródeł, zapewnia wsparcie dla:

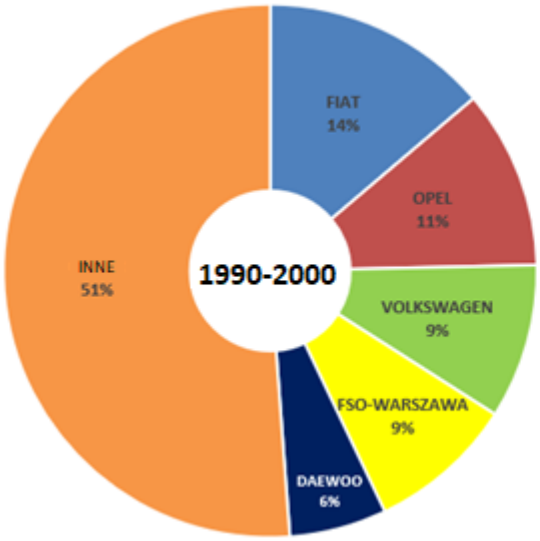
- procesów związanych z rejestracją pojazdów i wydawaniem dokumentów potwierdzających uprawnienia do kierowania pojazdami,
- procesów związanych z przeprowadzaniem badań technicznych pojazdów,
- działań organów odpowiedzialnych za bezpieczeństwo państwa i obywateli.

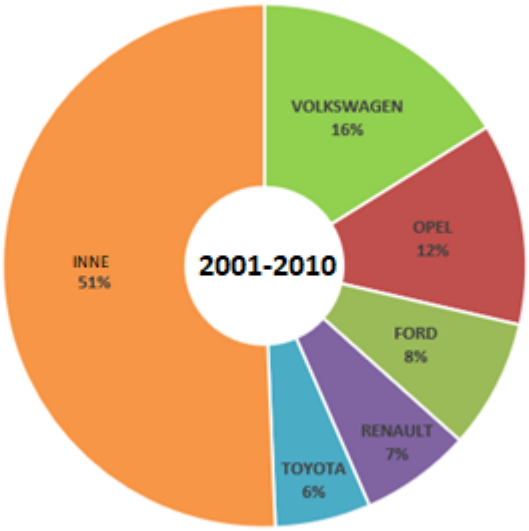
Baza danych zawiera

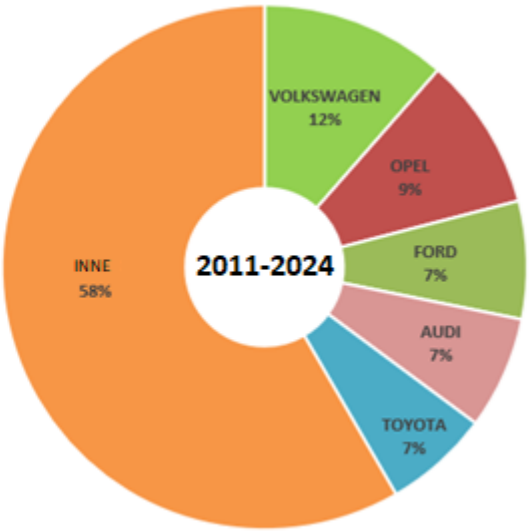
Baza CEPIK, którą pozyskaliśmy zawiera około 18 mln samochodów osobowych z określeniem m.in.:

- Marki
- Modelu
- Pojemności skokowej silnika
- Rodzaju paliwa
- Rodzaju paliwa alternatywnego
- Województwa, powiatu oraz gminy rejestracji
- Daty rejestracji
- Masy własnej
- Pochodzenia pojazdu

Najpopularniejsze marki na przestrzeni 30 lat







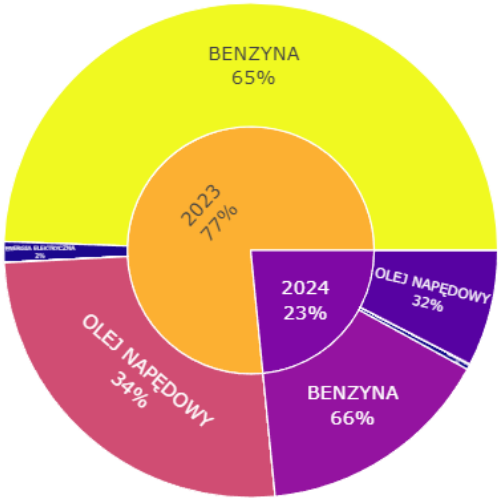
W dalszej części prezentacji, aby określić zachowania na rynku motoryzacyjnym, zbiór danych został ograniczony do okresu 2023-2024.

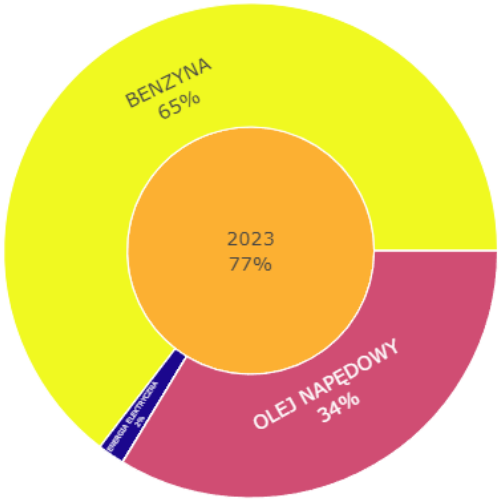
Najchętniej wybierani - najczęściej rejestrowani w ciągu ostatniego 1,5 roku wszystkich samochodów

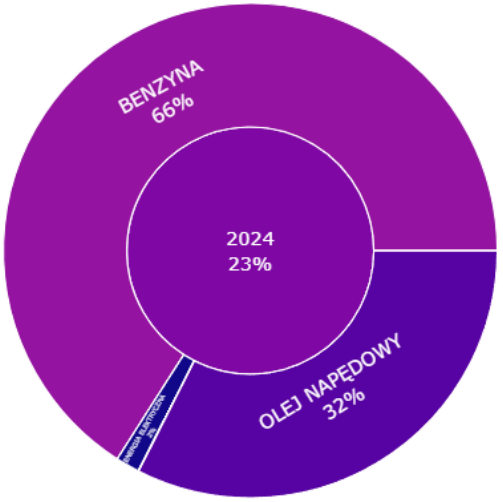
1. Volkswagen Golf 24745
2. Skoda Octavia 21151
3. Kia Sportage 20591
4. Ford Focus 17317
5. Audi A4 16732

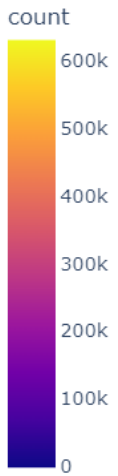
Zwycięzcy w województwach

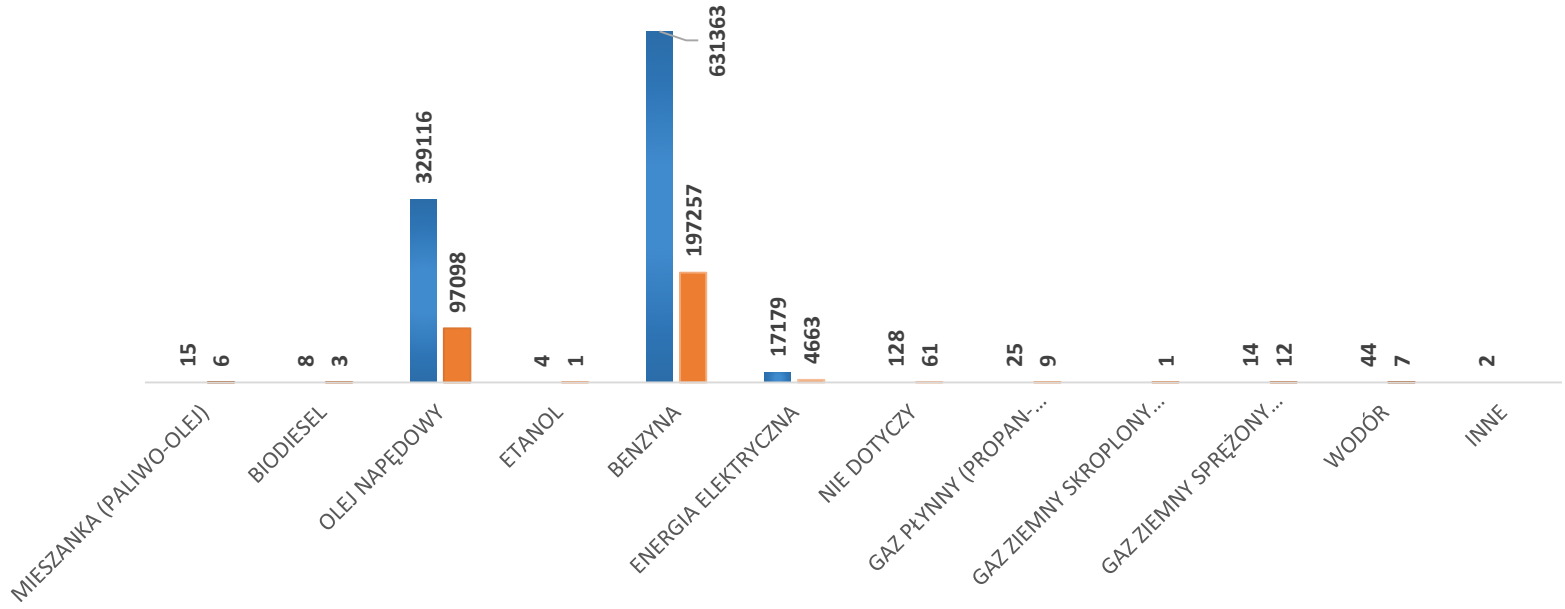
Najczęściej wybierane rodzaje paliwa



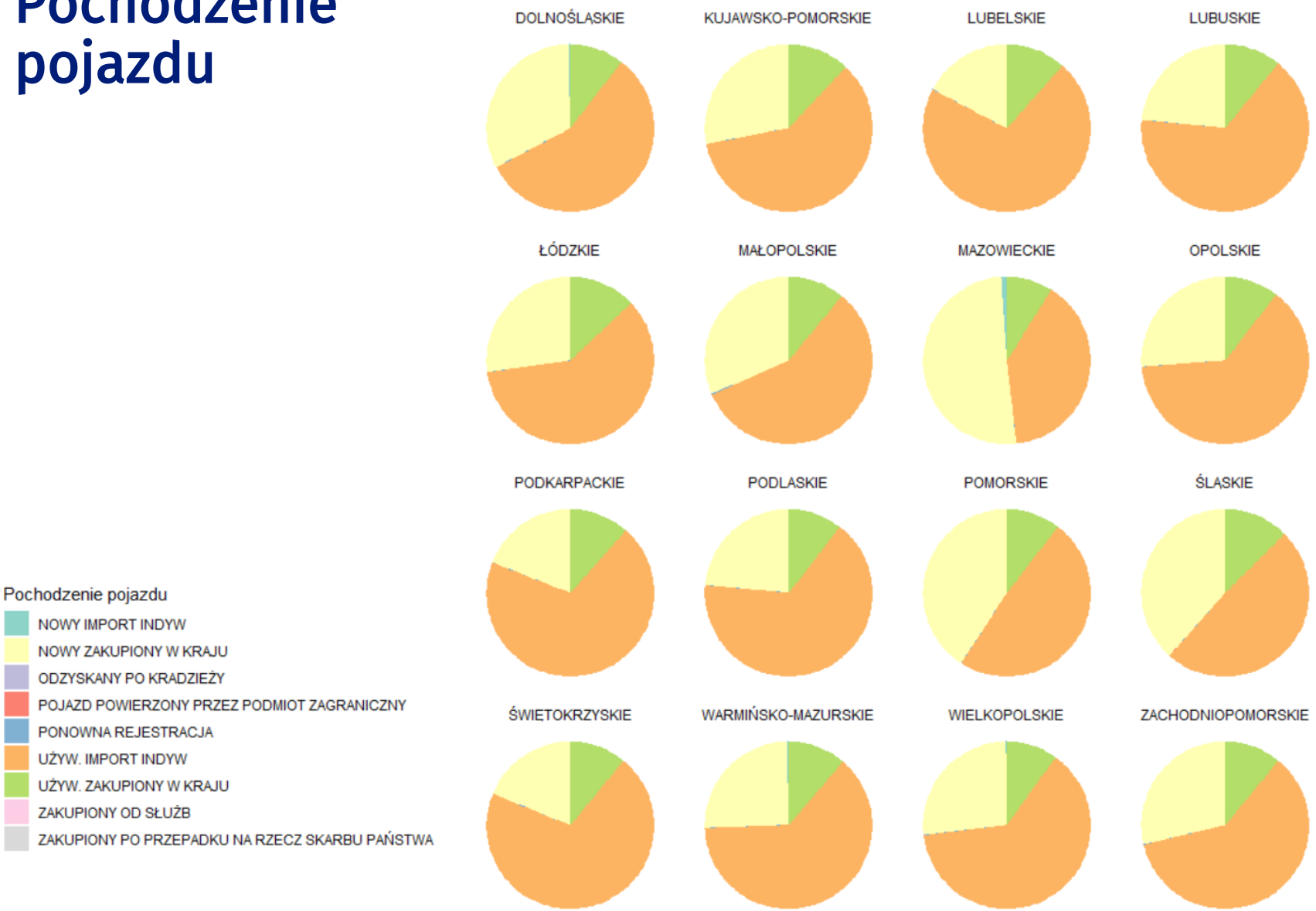




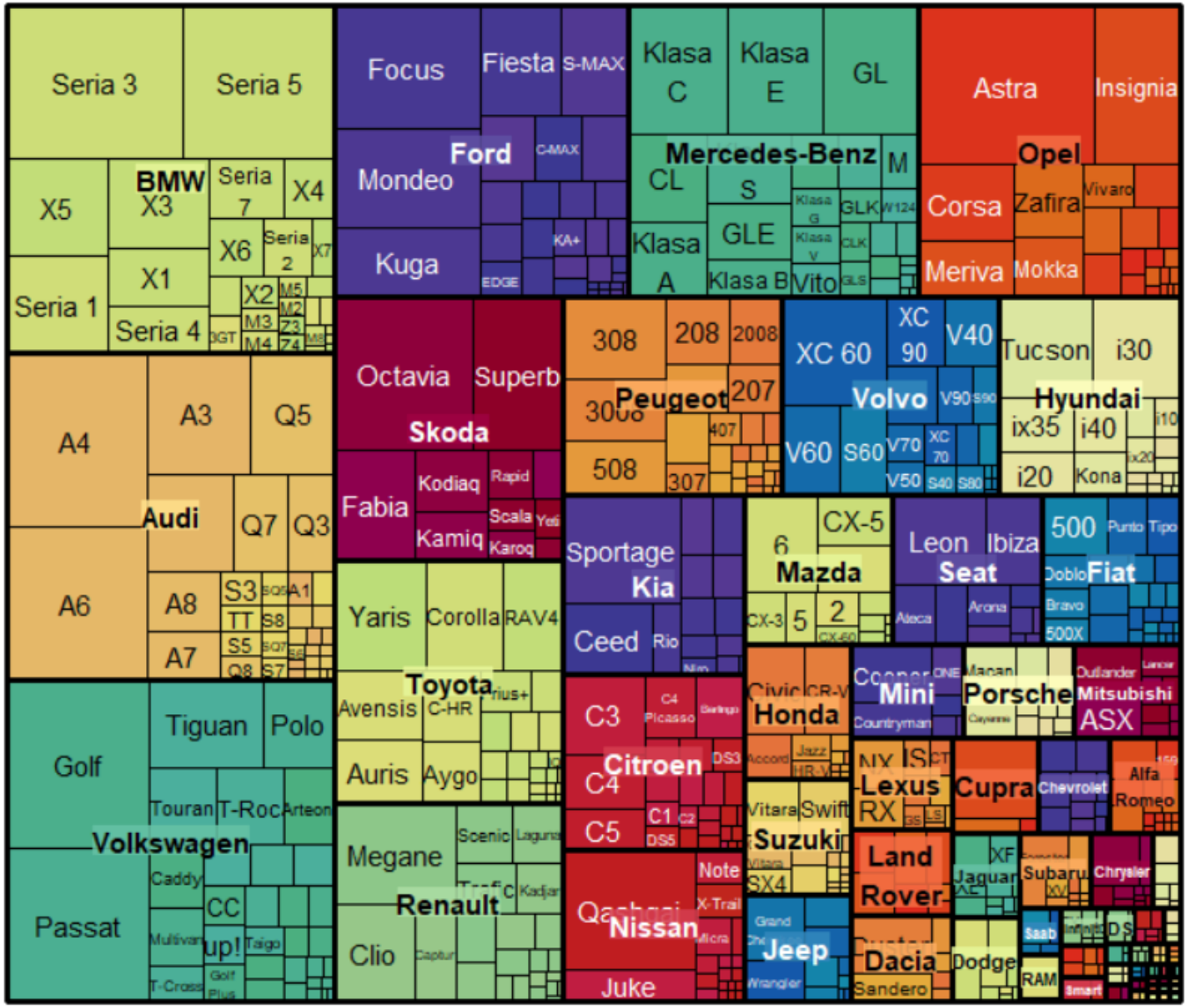




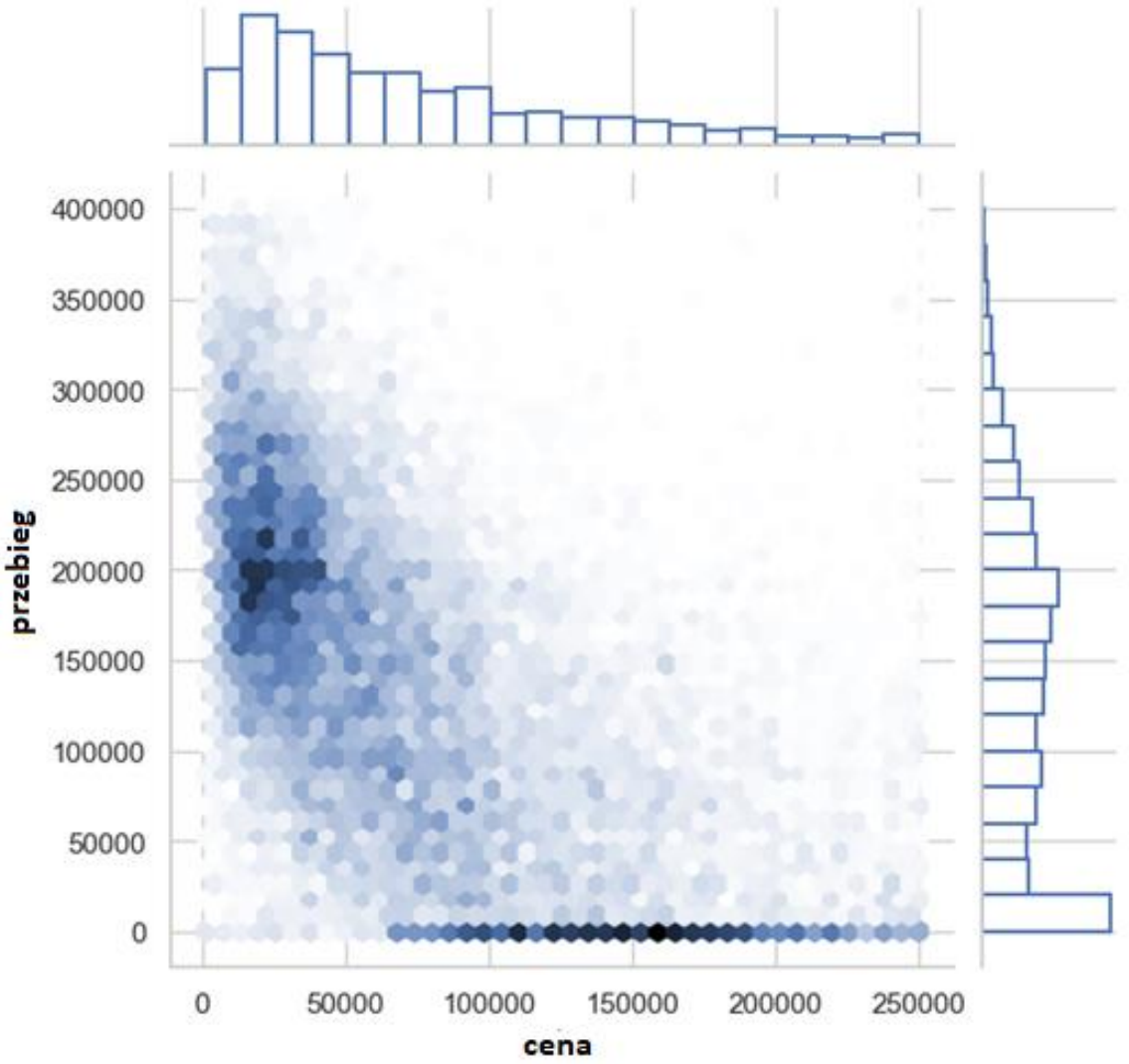
Pochodzenie pojazdu



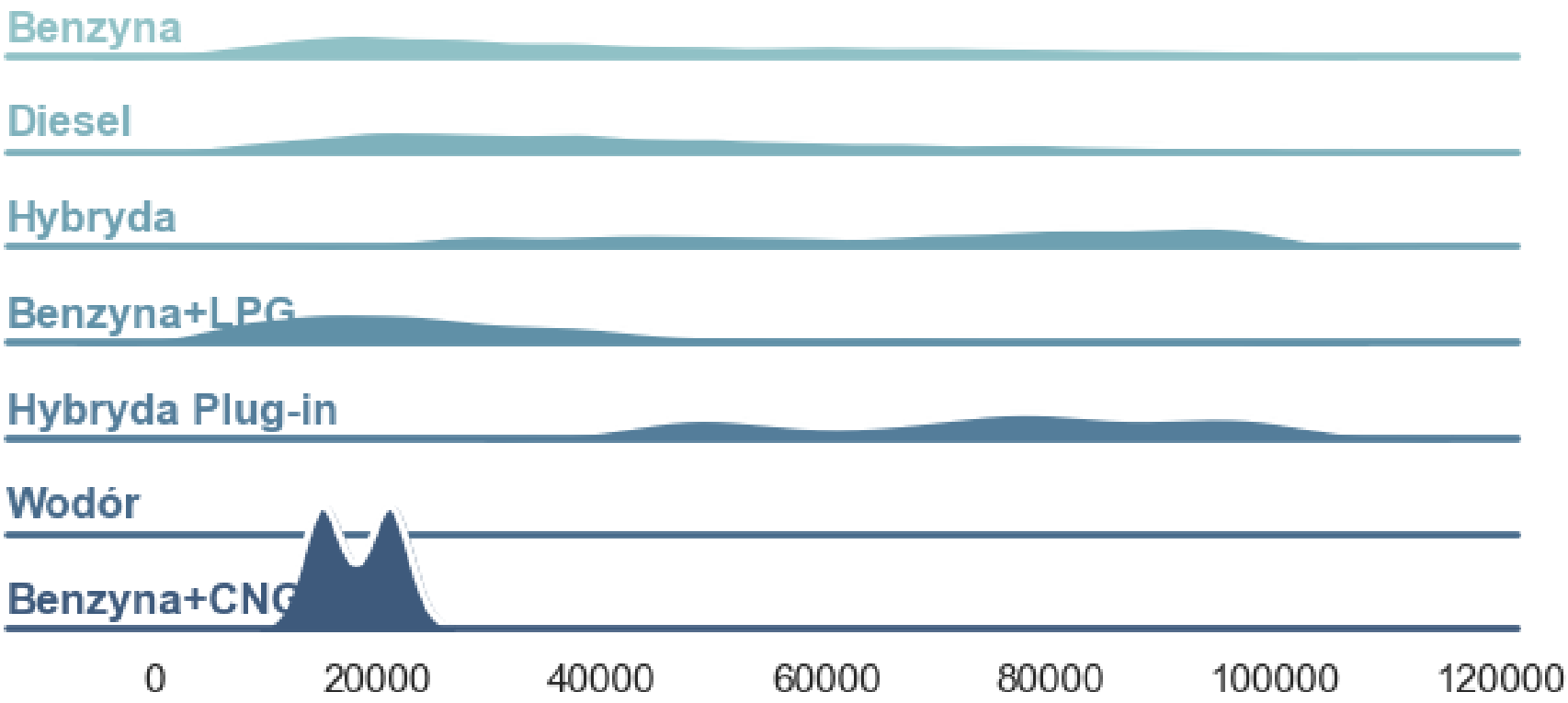
Samochody osobowe na sprzedaż



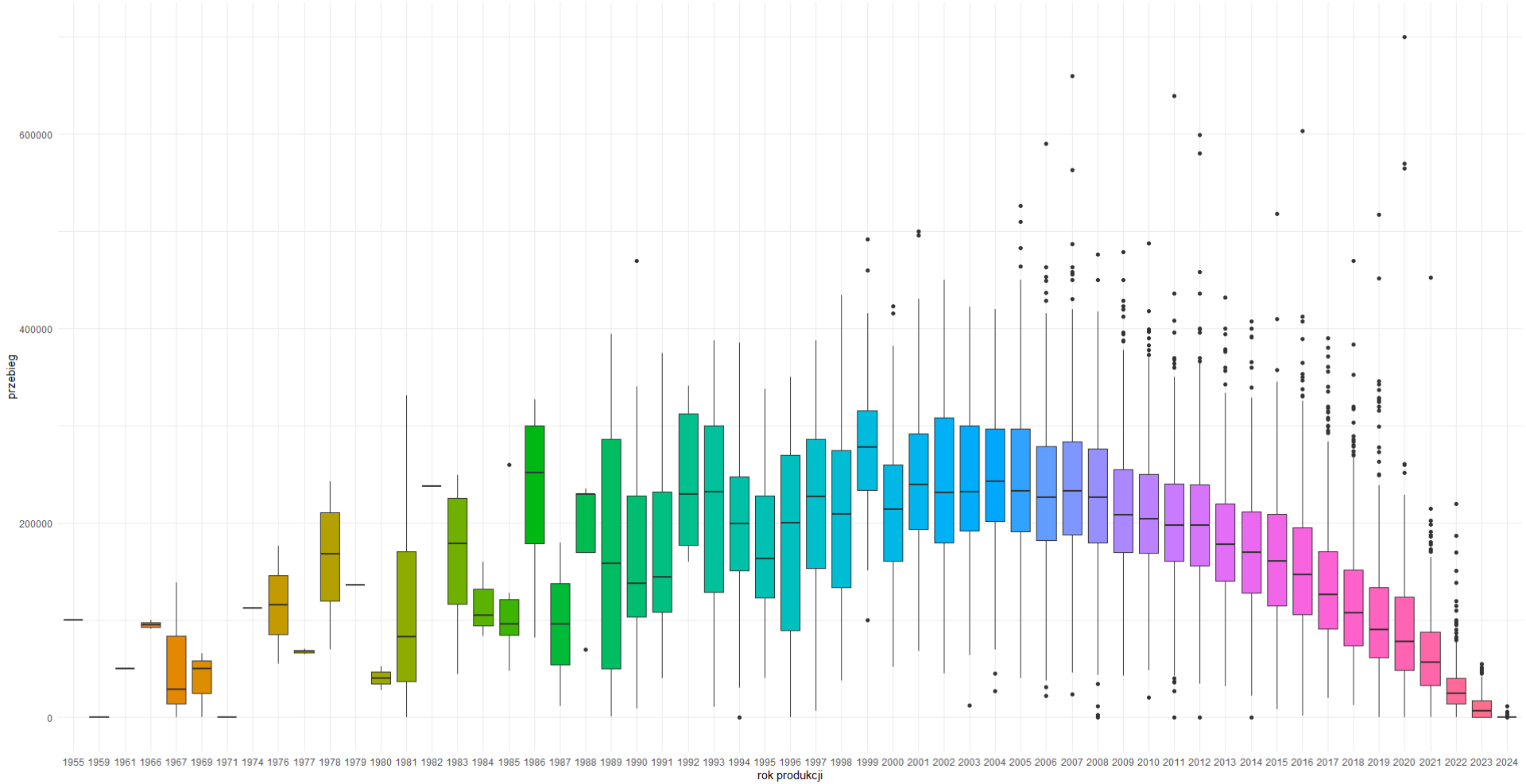
Czy cena zależy od przebiegu?



Czy cena zależy od rodzaju paliwa?



Przebieg a rok produkcji



Probabilistyczne łączenie zbiorów danych

- Tradycyjne podejścia do łączenia rekordów często opierają się na ścisłych regułach dopasowania. Najczęściej stosowany jest unikalny identyfikator np. PESEL
- W przypadku braku takiego identyfikatora często używa się kilku zmiennych jednocześnie np. imię, nazwisko, data urodzenia.
- Braki danych lub błędy literowe uniemożliwiają sparowanie np. „Toyota C-HR” a „Toyota CHR”
- Zastosowanie komparatorów tekstowych i technik probabilistycznych pozwala oszacować prawdopodobieństwo, że dwa rekordy dotyczą tej samej jednostki, mimo różnic w zapisach.

Probabilistyczne łączenie zbiorów danych

- Zbiór wszystkich możliwych par łączonych zbiorów A, B

$$A \times \mathrm{B} = \{(a, b):\ a \in \mathrm{A}, b \in B\}$$

jest sumą dwóch rozłącznych zbiorów

$M = \{(a, b): a = b, a \in \mathrm{A}, b \in B\}$, $\quad U = \{(a, b): a \neq b, a \in \mathrm{A}, b \in B\}$

- Załóżmy, że chcemy je połączyć na podstawie k wspólnych zmiennych.

$X(a) = \big(X_1(a), \dots, X_k(a)\big), Y(b) = \big(Y_1(b), \dots, Y_k(b)\big)\ dla\ a \in A, b \in B$

- Podobieństwo $X(a)$ i $Y(b)$ jest opisane poprzez wektor porównań

$\gamma[X(a), Y(b)] = \big(\gamma^1[X(a), Y(b)], \dots, \gamma^k[X(a), Y(b)]\big)$, $\quad \gamma \in \Gamma$

Probabilistyczne łączenie zbiorów danych

- Poszukujemy reguły decyzyjnej $d(\gamma)$, która każdej parze (a, b) przypisze prawdopodobieństwo jednego ze zdarzeń: *link*, *possible link*, *non-link*.
- Oznaczamy prawdopodobieństwa warunkowe

$$m(\gamma) \coloneqq P(\gamma[X(a), Y(b)]|(a, b) \in M)$$
$$u(\gamma) \coloneqq P(\gamma[X(a), Y(b)]|(a, b) \in U)$$

- W procesie łączenia występują dwa rodzaje błędów

$$P(link|U) = \sum_{\gamma} u(\gamma)P(link|\gamma)\,,$$

$$P(non-link|M) = \sum_{\gamma} m(\gamma)P(non-link|\gamma)$$

Model Fellegi-Sunter

- Łączna niezależność rozkładów warunkowych

$$m(\gamma) = \prod_{i}^{k} m_i(\gamma^i) = \prod_{i}^{k} P(\gamma^i | (a,b) \in M)$$

$$u(\gamma) = \prod_{i}^{k} u_i(\gamma^i) = \prod_{i}^{k} P(\gamma^i | (a,b) \in U)$$

- Definiowanie wag służących tworzeniu reguły decyzyjnej

$$w_i(\gamma^i) = \log m_i(\gamma^i) - \log m_i(\gamma^i), \qquad w(\gamma) = \sum_{i}^{k} w_i(\gamma^i)$$

- Model jest zaimplementowany w języku R m.in. w pakiecie *reclin2*.

Zastosowanie modelu

- Przed połączeniem zbiorów zharmonizowano zmienną „rodzaj paliwa”
- Pierwszy etap łączenia CEPIK i ofert sprzedaży samochodów:
  - model i marka jako zmienne blokujące
  - rodzaj paliwa z komparatorem Jaro-Winklera
  - pojemność skokową i rok produkcji z metryką euklidesową

$$sim_{JW} = sim_J + lp(1 - sim_J)$$

gdzie

$$sim_J = \begin{cases} 0 & dla\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & dla\ m > 0 \end{cases}$$

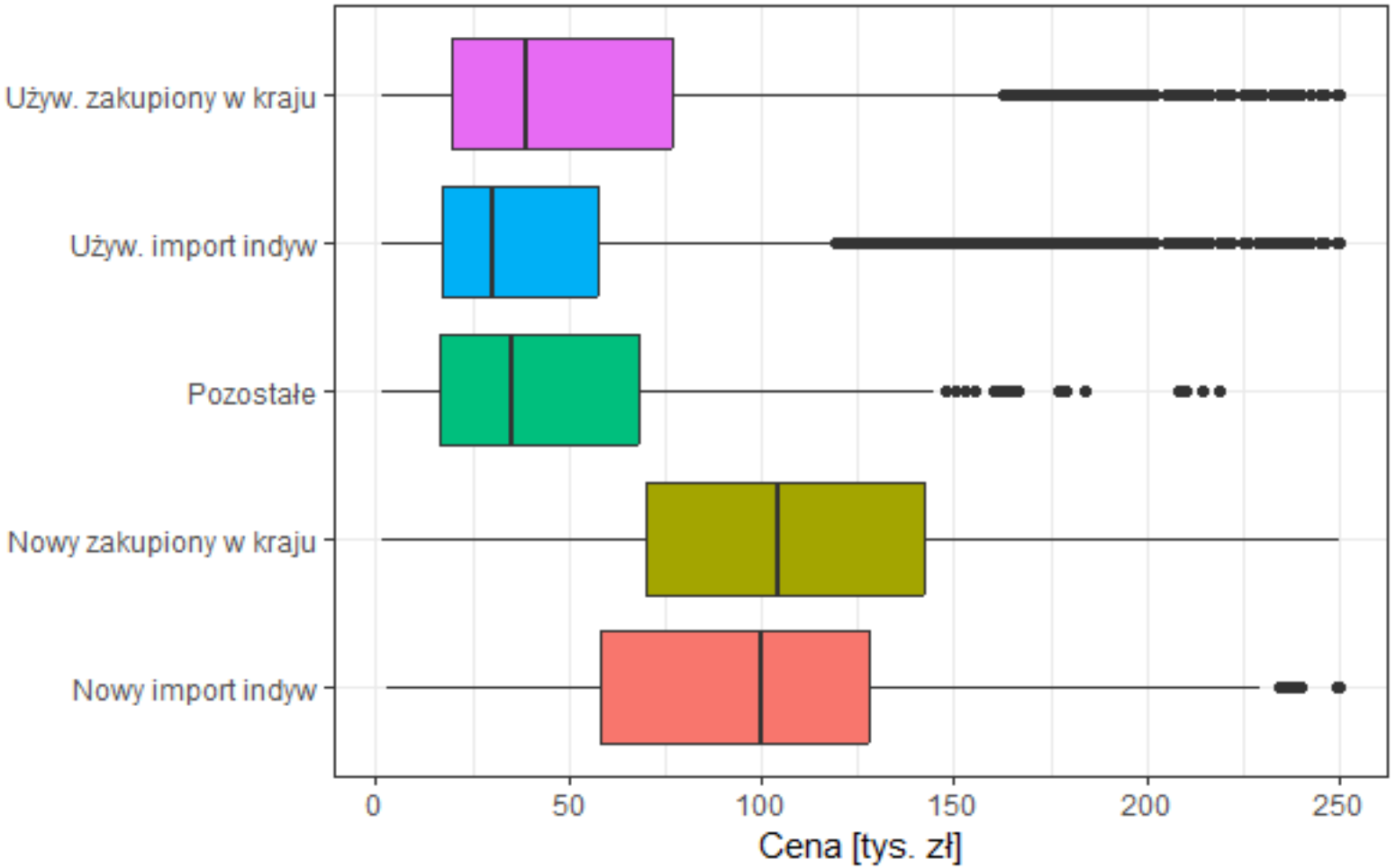
- $sim_{JW}(\text{'Toyota C—HR'},\text{'Toyota CHR'})=0.969$
- Na pierwszym etapie sparowano 78,2 % rekordów

Problemy po drodze

Poprawa dopasowań

- W celu poprawy jakości zdiagnozowaliśmy różnice w nazwie marki i modeli w obu zbiorach:
  - 211 przypadków: marka (?), model 1500
  - 9408 przypadków: marka (?), model Ateca / Cupra Leon / Formentor
  - 48407 przypadków: marka Toyota, model Toyota …
- W drugim etapie model nie był zmienną blokującą a porównywaną komparatorem tekstowym
  - Portal: model XC 40, CEPIK: model XC40, XC 90
- Na drugim etapie sparowano 92,98 %

Rozkład cen według pochodzenia pojazdu



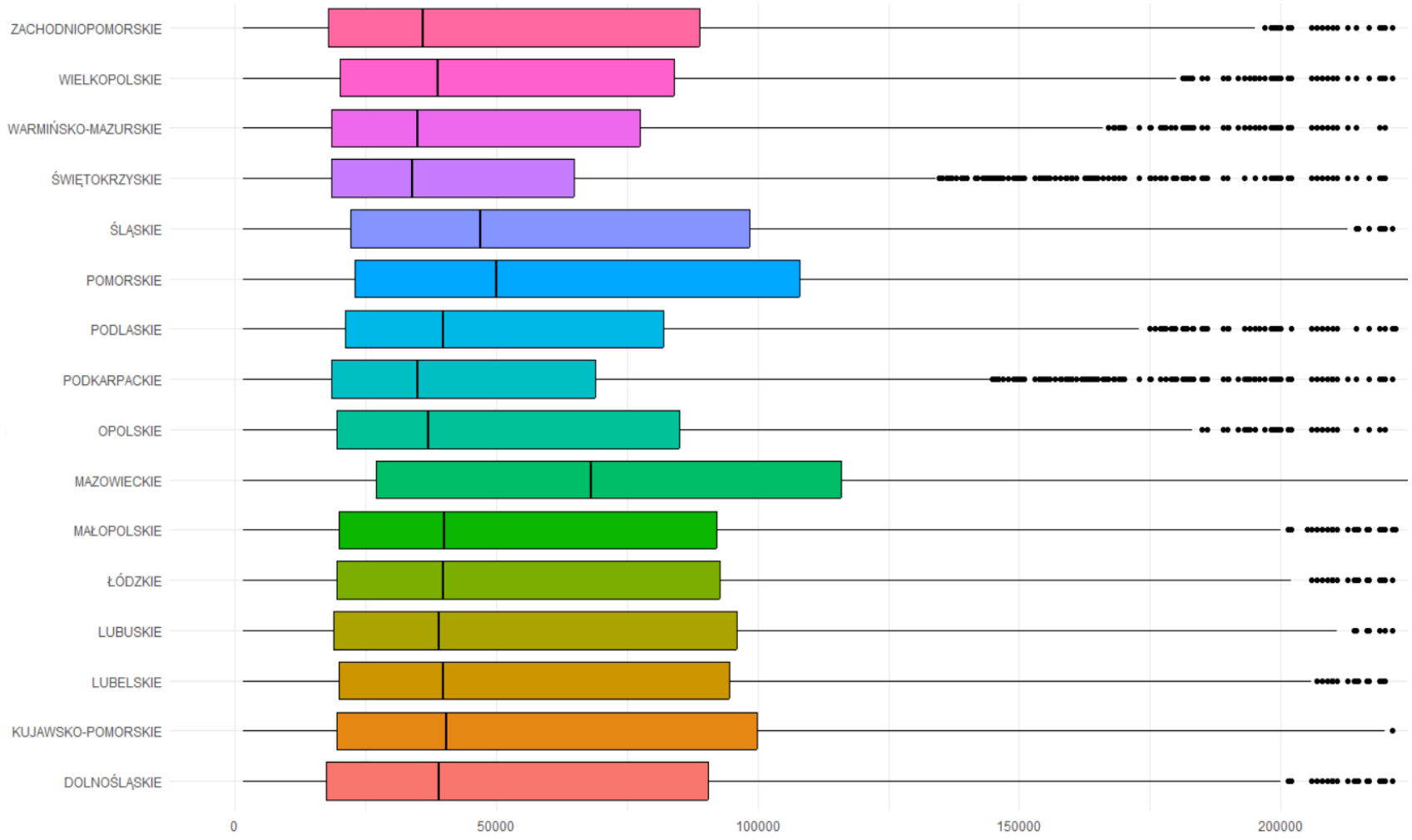
Pozostałe: z przepadku na rzecz Skarb Państwa, odzyskany po kradzieży, ponowna rejestracja, zakupiony od służb, powierzony przez podmiot zagraniczny

Wartość rynku

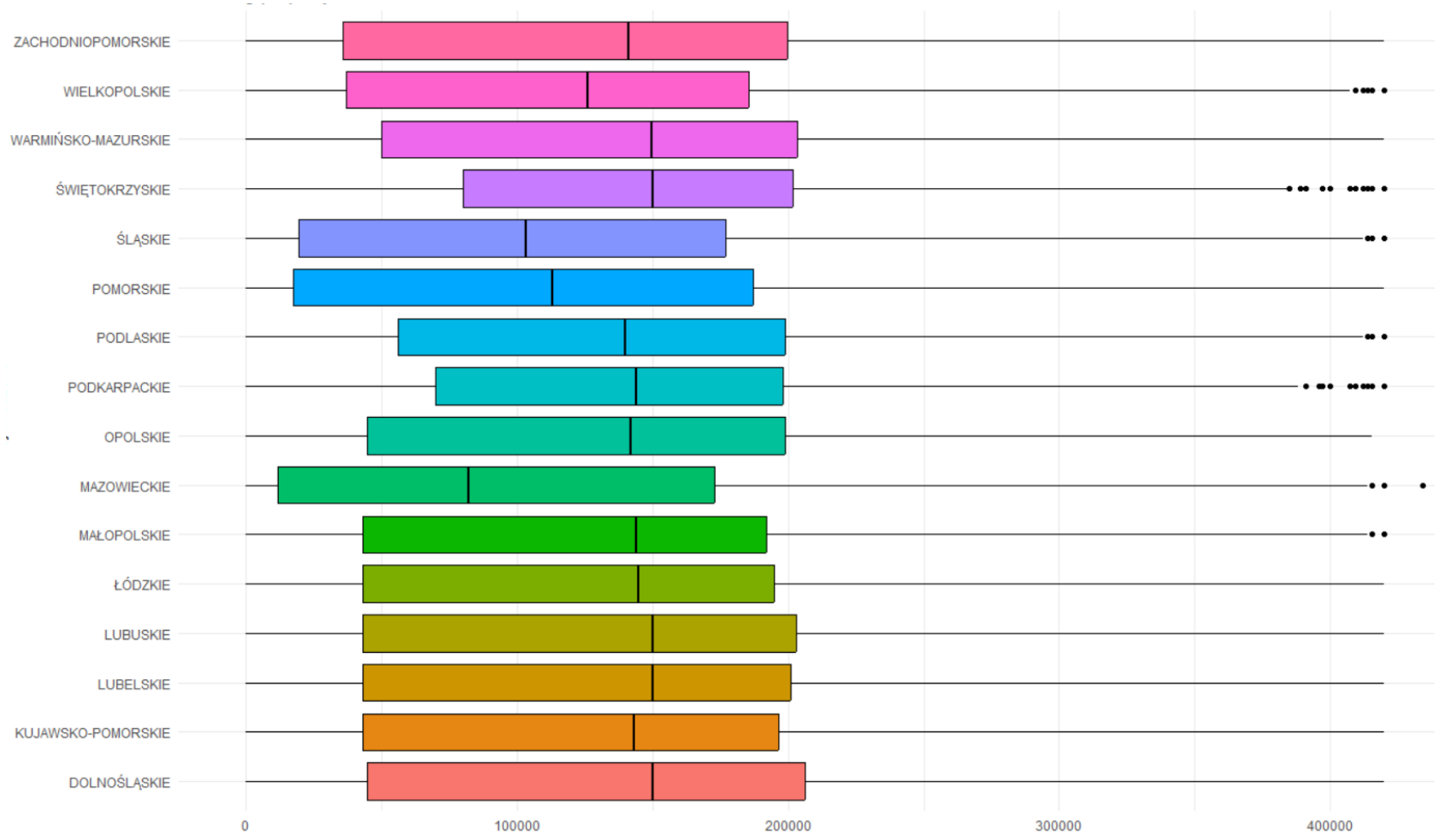
Pochodzenie pojazdu	Liczba pojazdów	Średnia cena [tys. zł]	Wartość ogółem [tys. zł]
Nowy import indywidualny	2 456	108,16	265 658
Nowy zakupiony w kraju	339 849	112,29	38 162 531
Pozostałe*	724	50,49	36 561
Używ. import indywidualny	701 901	46,86	32 897 853
Używ. zakupiony w kraju	141 650	57,16	8 096 775

*z przepadku na rzecz Skarb Państwa, odzyskany po kradzieży, ponowna rejestracja, zakupiony od służb, powierzony przez podmiot zagraniczny

Rozkład cen według województw



Rozkład przebiegów według województw



Co dalej?

- W kolejnym etapach prac:
  - oszacujemy emisję zanieczyszczeń (w tym CO_2) na postawie takich zmiennych jak rok produkcji, przebieg , rodzaj paliwa oraz standardów emisji
    (w tej chwili 93% rejestrowanych w CEPIK pojazdów nie ma informacji o poziomie emisji)
  - określimy profile demograficzno-ekonomiczne użytkowników samochodów na podstawie badania *Budżetów Gospodarstw Domowych* oraz badania *Zużycia Paliw i Energii w Gospodarstw Domowych*

Dziękujemy za uwagę

Agnieszka Giemza

Specjalista

a.giemza@stat.gov.pl

dr Sebastian Wójcik

kierownik działu

s.wojcik@stat.gov.pl

Warszawa, 2-4.07.2024 r.