

# DSC Competition

Asia Sowinska & Kinga Marszałkowska

# INTRO

Temat: Predykcja cen  
samochodów

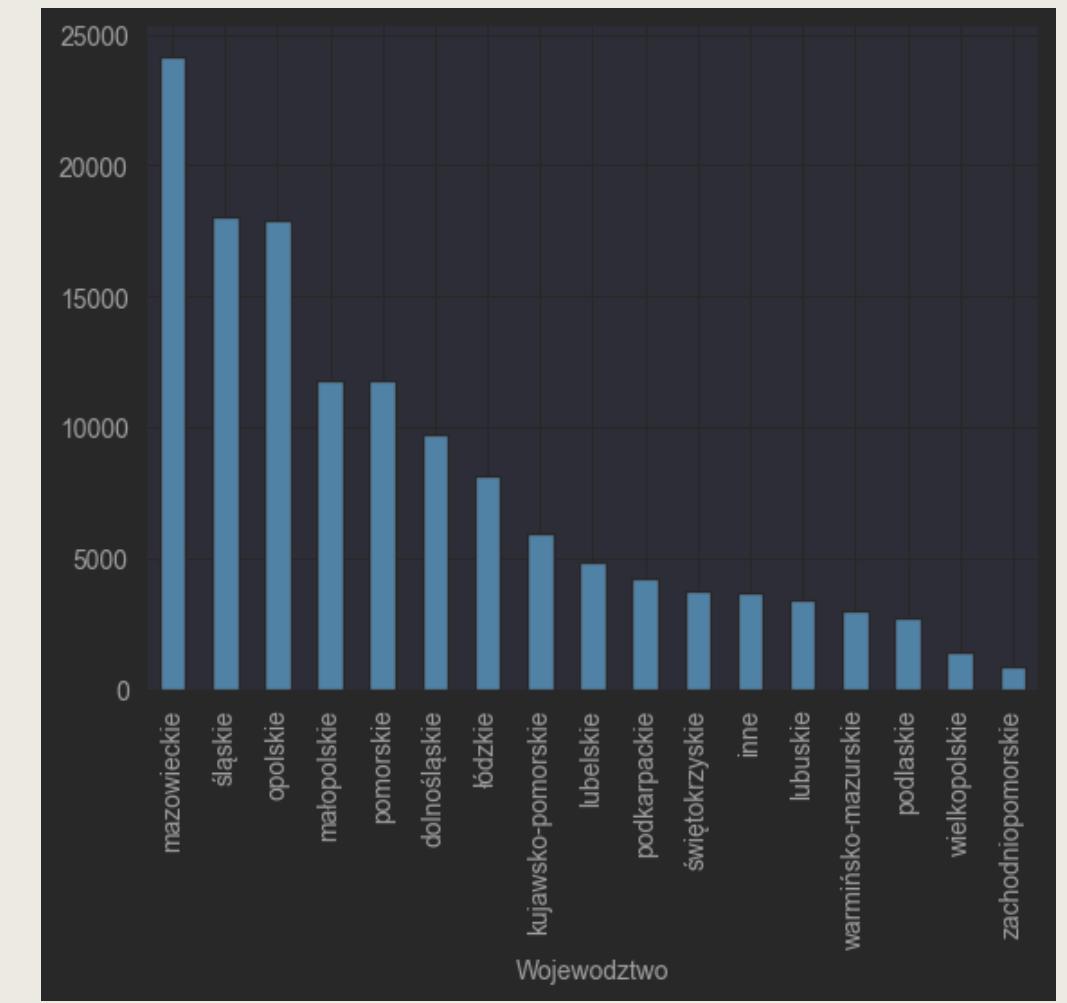
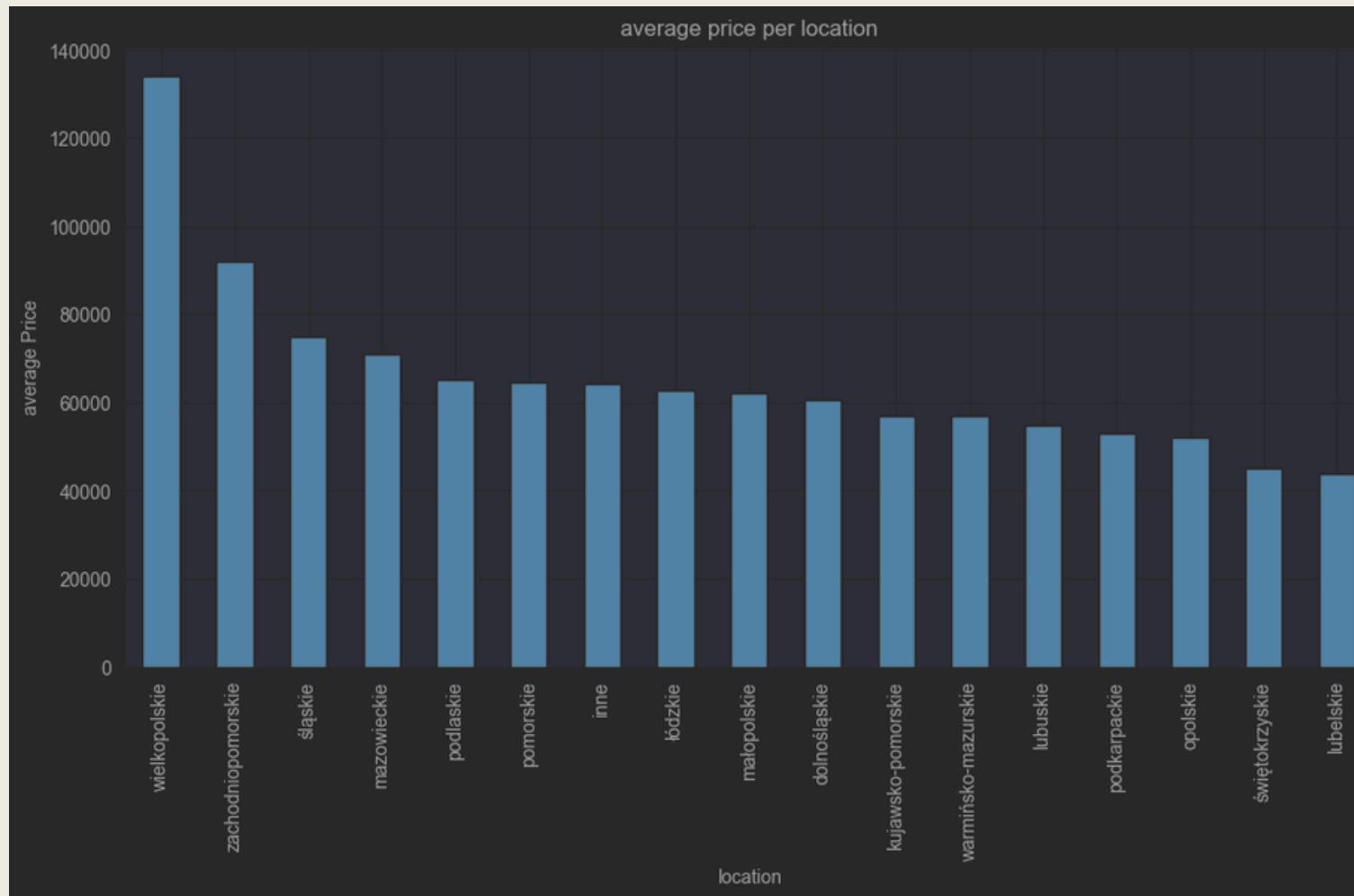
– – w tej prezentacji opowiadamy  
krótko o obserwacjach i  
wyzwaniach przy tworzeniu  
rozwiązania na organizowany  
przez Was (DSC) konkurs



# OBSERWACJE:

1

lokalizacja lokalizacji  
nierówna:

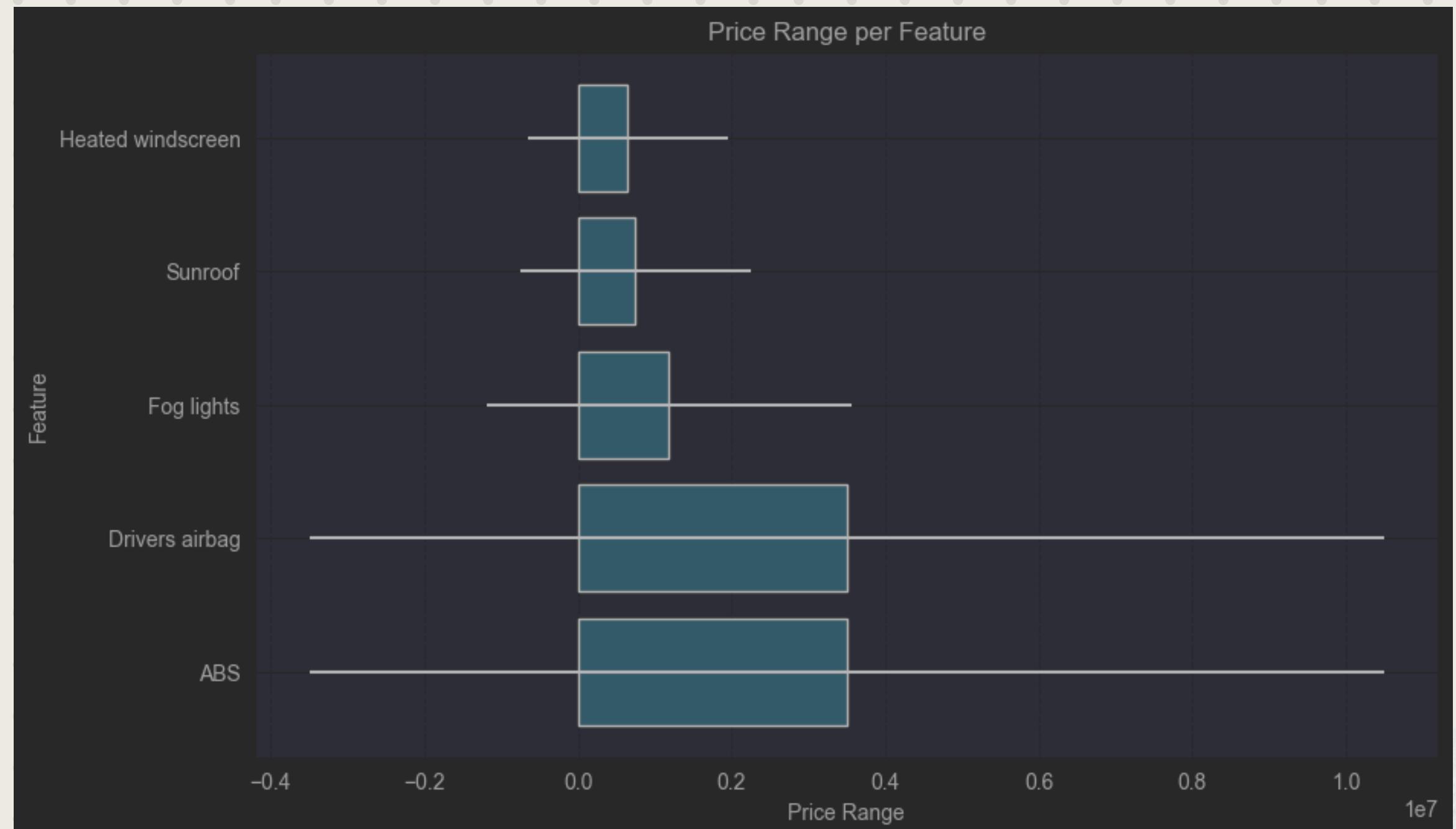


najwięcej samochodów  
sprzedawano w mazowieckim, ale  
najdrożej było w wielkopolskim

# OBSERWACJE:

2

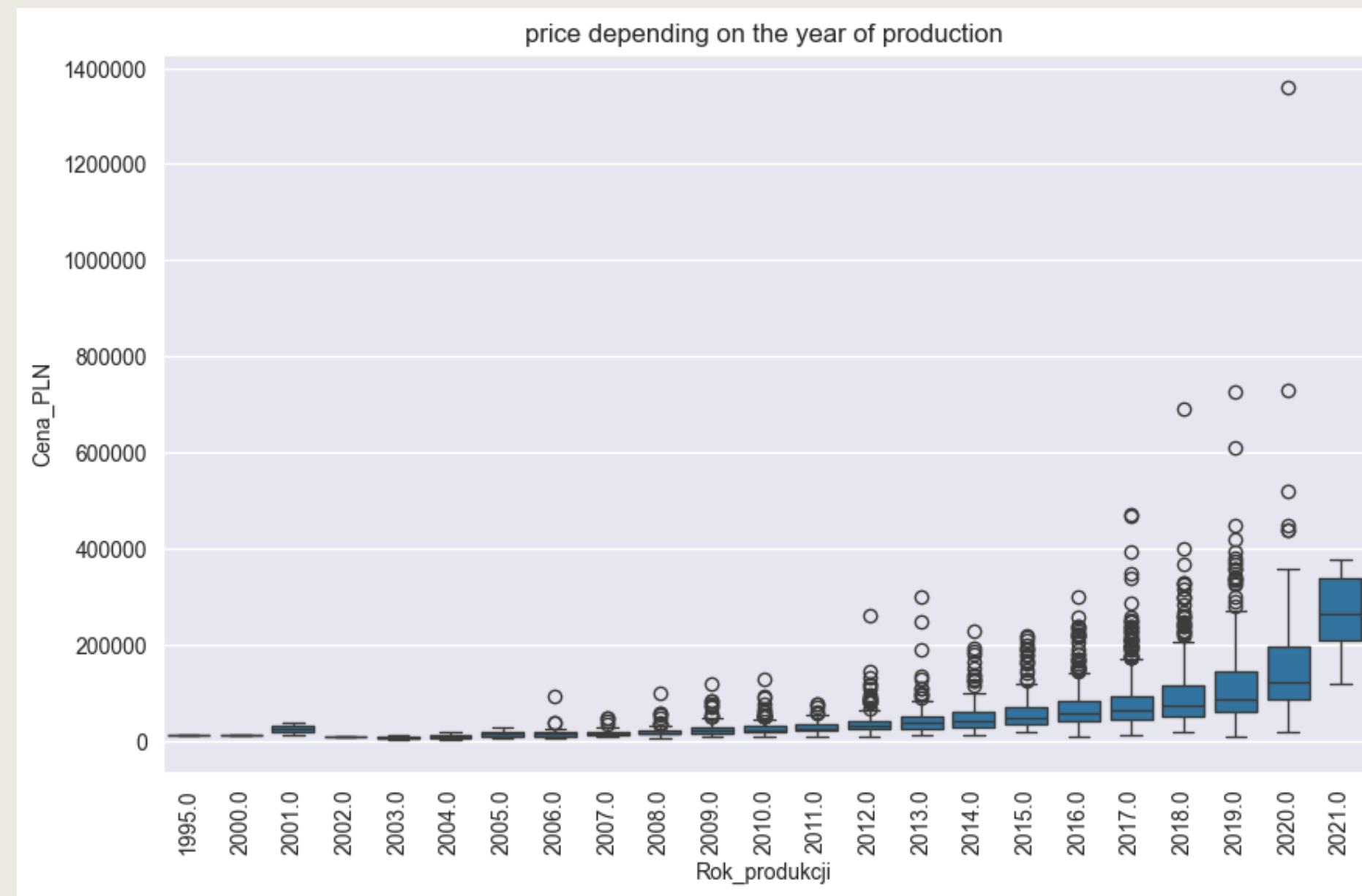
nie liczy się ile samochód ma bajerów, ale jakie ma bajery (!)



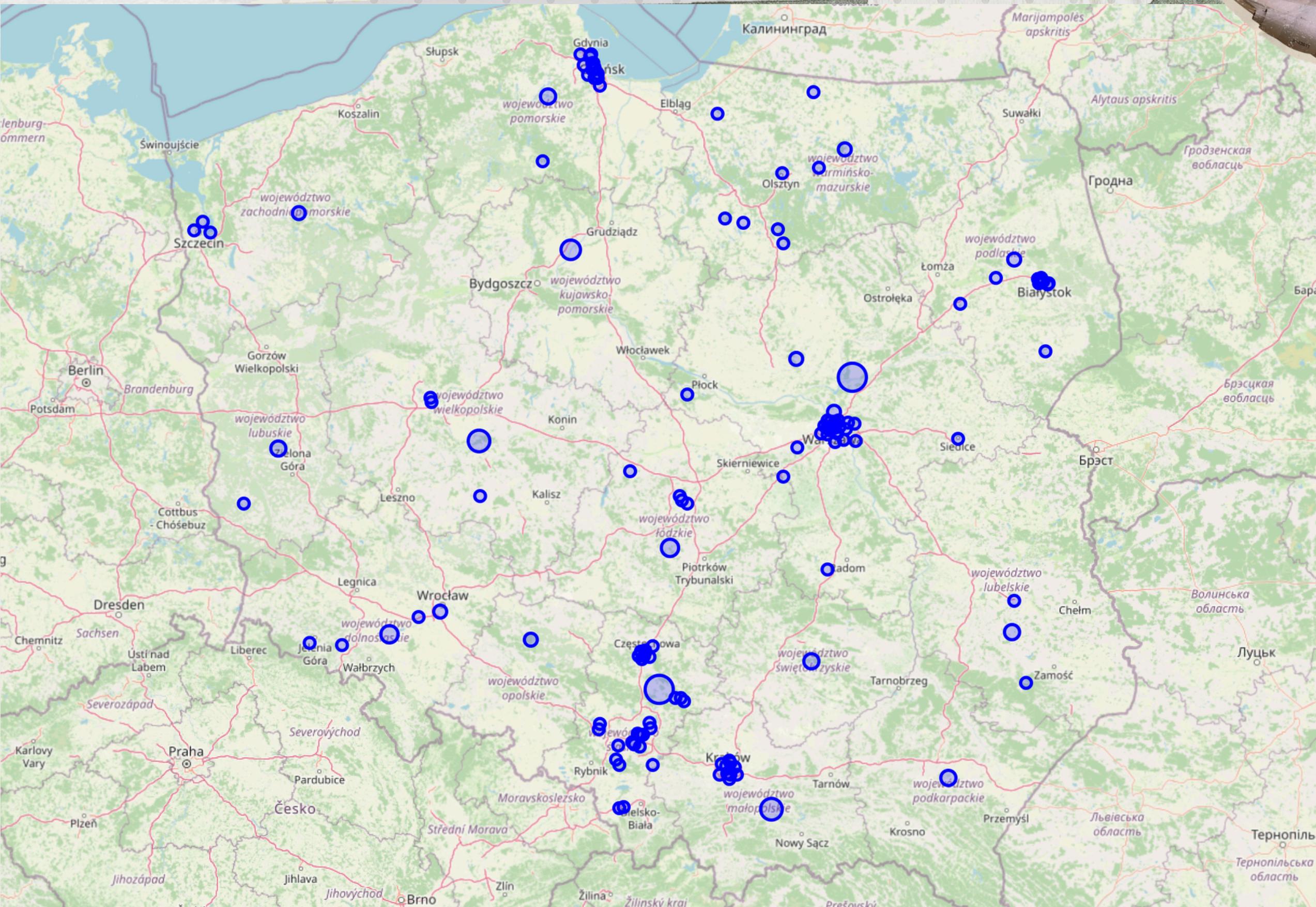
# OBSERWACJE:

3

całkiem spodziewane, ale rok produkcji był  
najważniejszym czynnikiem/ficzerem w tej predykcji



# Mapka ogłoszeń



# MODELOWANIE

- ☒ regresja, LightGBM
- ☒ nowe ficzery
- ☒ optuna do hiperparametrów
- ☒ encoding: OHE
- ☒ feature importance
- ☒ wyrzucamy kolumny datowe

# ANALIZA

największy wpływ na ocenę mają: rok ⓘ|

produkcji, moc i przebieg

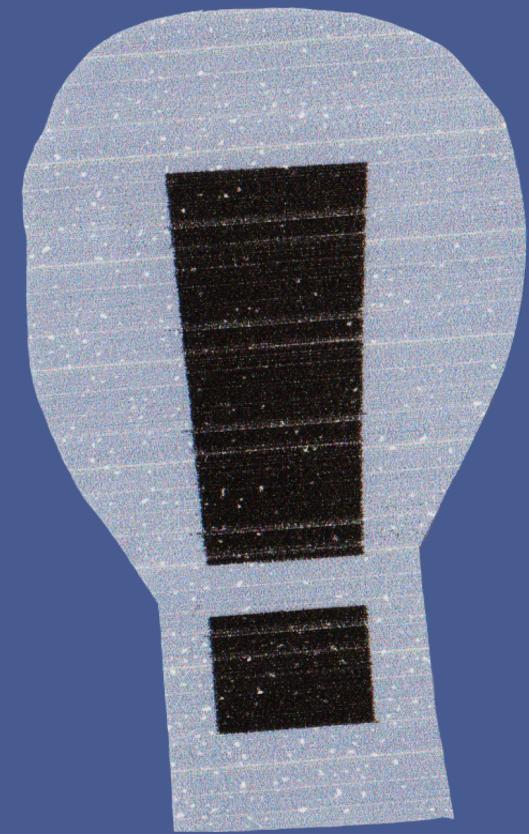
syntetyczne dane są słabe ⓘ|

cena występuje w różnych walutach ⓘ|

logarytmiczne skalowanie ceny (!) ⓘ|



# Wyzwania



CENA –  
SKALOWANIE  
LOGARYTMICZNE  
(KTO BY POMYŚLAŁ...)



PRZEBICIE  
BARIERY 20K  
RMSE  
(DOSYĆ DŁUGO NAM TO  
ZAJĘŁO)



KOTY  
SPACERUJĄCE PO  
NASZYCH  
KLAWIATURACH  
(TOP 1 WYZWANIE)

# WYNIKI

Po porównaniu różnych modeli jak dummy regressory, XGBoost, CatBoost, zdecydowałyśmy się postawić na **LightGBM**, którego predykcje w połączeniu ze wszystkimi innymi użytymi przez nas technikami były najbardziej trafne.

Final RMSE (source: Kaggle): 19339.342833

Dziękujemy za możliwość wzięcia udziału oraz za wszystko, czego się nauczyłyśmy - jak się okazuje, "Warszawa" w markach pojazdu to nic martwiącego...