1. Wynik Scrappingu:

Program scrapuje dane z serwisu Otomoto. Dane te są pobierane tylko dla aut osobowych.

Zmienne jakie program pobiera to:  
- Brand – marka samochodu. Jeśli marka nie zostaje dopasowana do bazy nazw marek, dostaje kategorie „Inny”,

- Model – model samochodu podany przez użytkownika w ogłoszeniu (bez nazwy marki). Jeżeli marka samochodu nie została dopasowana do klucza to nazwa modelu pozostaje wraz z nazwą marki,

- Price – cena samochodu wyrażona w złotówkach bez części groszowej,

- Engine – pojemność silnika,

- KM – Konie mechaniczne; moc silnika,

- Mileage – przebieg samochodu wyrażony w km,

- Fuel – zastosowany sposób zasilania samochodu,

- Gearbox – Zastosowana skrzynia w samochodzie. Rozróżnialna jest tylko skrzynia manualna lub automatyczna,

- Year – rok produkcji

- Place – lokalizacja samochodu

- Voivode – województwo w którym jest ogłoszony samochód

Wynik jest zapisywany w pliku json i pliku csv, który jest używany do analizy danych.

1. Założenia jakie program ma osiągnąć w analizie danych.

- podstawowe obliczenia na cenach czy województwach.

- korelacja pomiędzy ceną a zmiennymi technicznymi pojazdu.

- popularność danego rozwiązania technicznego

- wykres rozkładu cen

- tabela krzyżowa

- regresja liniowa ceny

3. Machine Learning ma na celu ukazać, jak model zinterpretuje zmienne i jakie jest prawdopodobieństwo przewidzenia cen pojazdów na podstawie danych technicznych oraz ceny.

4. Przed każdą analizą czy predykcją dane należało oczyścić i ustandaryzować:

- województwa Polski zostały odsiane od innych pozostałych miejsc w których ogłoszenie zostało dodane. Inne miejsca, które nie są województwem zostały podpisane jako „Inny kraj”.

- wszystkie dane liczbowe zostały przeparsowane na Integer. Dodatkowo jeżeli dany rekord nie posiadał wartości liczbowej, to zamiast Nan zostało wpisane 0.

- dane nie posiadające wartości zostały usunięte

- jeżeli do wykresu potrzebne były odpowiednie dla wykresu limitery lub inne działania na danych w celu lepszego przedstawienia wykresu, dane zostały odpowiednie zmodyfikowane.

- marka została wyciągnięta ze stringa podanego w modelu poprzez zastosowanie dopasowania z filtrem nazwy z pliku yaml.

- w przypadku predykcji dane kategoryczne zostały zamienione przy pomocy funkcji get\_dummies.

5. Wyniki podstawowych obliczeń danych.

a) Średnia cena samochodu jest równa 78247 zł. Możliwe że średnia została mocno wyciągnięta do góry przez drogie nowe samochody z marek premium.

b) Maksymalna cena: Mercedes-Benz Klasa G AMG 63 za prawie 5 milionów jest przykładem dlaczego średnia jest aż tak wysoka.

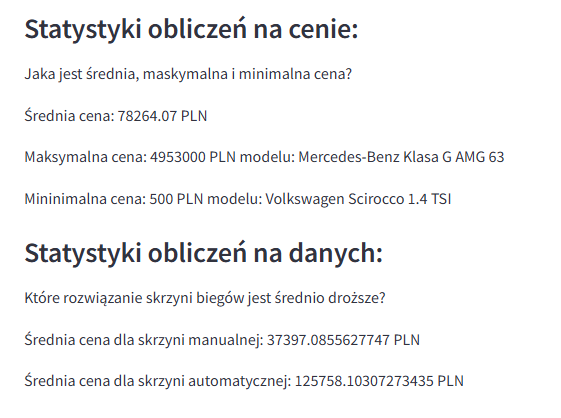
c) Minimalna cena: 500 dla Volkswagena Scirocco 1.4 TSI. Nie było możliwe od sortować samochodów nie działających, po wypadkach, itp. Dlatego nie wiadomo czy ten samochód za 500 zł to rzeczywiście najtańszy dostępny jeżdżący samochód.

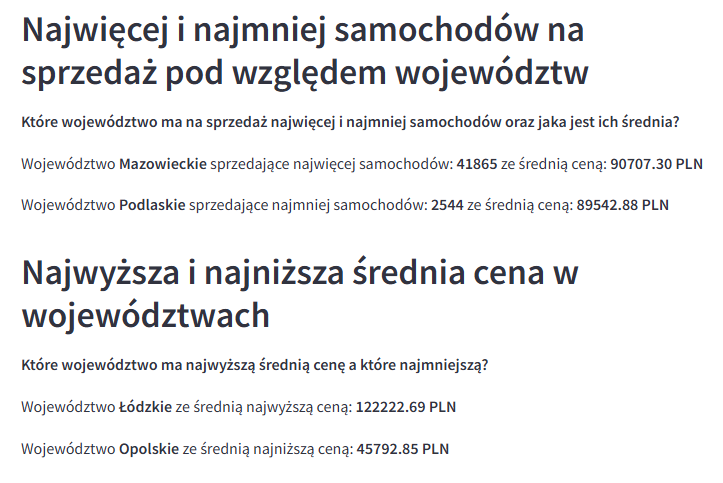
d) Które samochody są średnio droższe ze skrzynią biegów manualną czy automatyczną?

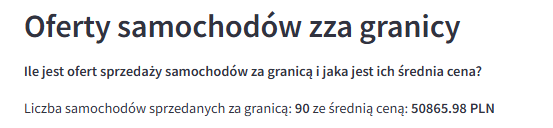
37397 < 125758  
To bardzo duża różnica, więc z całą pewnością można pwoeidzieć, że samochody ze skrzynią biegów automatyczną jest znacznie droższy niż manualny odpowiednik.

e) Województwo Mazowieckie sprzedaje najwięcej aut bo aż 42000 w przybliżeniu, natomiast województwo Podlaskie sprzedaje najmniej bo tylko 2500. To aż 95% mniej niż w Mazowszu.

f) na polskim serwisie ogłoszeniowym Otomoto, jest 90 samochodów sprzedawanych za granicą z różnych krajów.

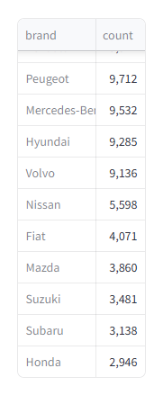






6. Popularność samochodów





Pierwsze 6 pozycji dominuje przez niemieckie koncerny samochodowe. Pokazuje to, że najczęściej sprzedawane są niemieckie maszyny.

Sekcja Inne jest na 7 pozycji, to wskazuje że filtr marek samochodów jest zbyt ogólny i pomimo że marek zostało wypisanych około 80. To ponad 10000 aut nie zostało przypisane do żadnej marki.

Ciekawe też jest obserwacja że numer 19 to Subaru. Wydaje się że ta japońska marka nie jest zbyt popularna w Polsce a jednak 3000 aut jest na sprzedaż. Pytanie czy to z powodu, że się nie sprzedają czy jednak Subaru wcale nie jest tak niepopularną marką.



Najpopularniejsze typy zasilania aut to benzyna. Aut elektrycznych jest 2000. Przy trochę ponad 200000 samochodów spalinowych to przepaść i jednak nowe elektryczne auta nie są aż tak popularne w sprzedaży.

Ciekawostką są ostatnie dwie pozycje, czyli samochód na Etanol i wodór, czyli auta które są bardzo niszowe to 11 samochodów jest na sprzedaż na Otomoto.

Z poprzedniego punktu wiemy, że średnio skrzynie automatyczne są droższe od manualnych. Jak widać na screenshocie, skrzynia manualna jest popularniejsza. Czy to z powodu, ceny? Niestety na tych wykresach nie jest możliwe wykazanie tego.

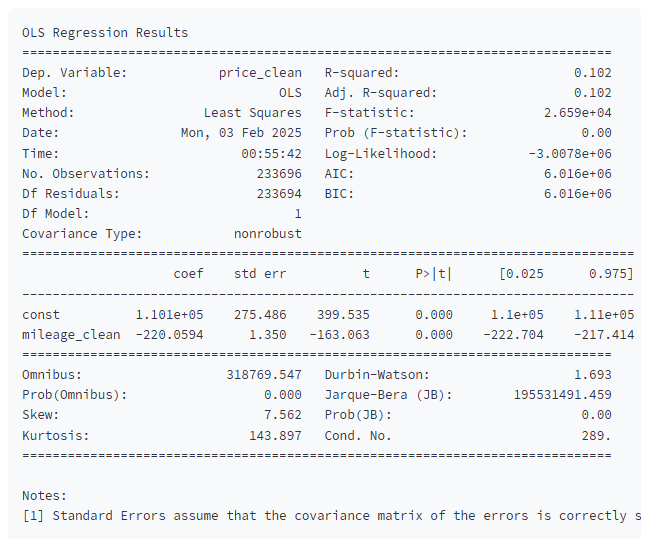
7. Została wykonana statystyka chi-kwadratu.

Chi2: 183994, natomiast p-value: 0.0

Wysoki wynik chi2, sugeruje, że różnice między obserwowanymi a oczekiwanymi wartościami są znaczne. Wskazuje to na potencjalną zależność między badanymi zmiennymi.

p-value wynosi 0.0 a to oznacza, że prawdopodobieństwo uzyskania takiego wyniku, jeśli nie ma zależności między zmiennymi jest praktycznie zerowe. P-value jest mniejsze niż 0,05, to oznacza że można odrzucić hipotezę zerową i przyjąć, że istnieje istotna zależność między zmiennymi.

8. Została wykonana regresja liniowa na danych



Regresja liniowa potrafi przewidzieć jedną zmienną na podstawie innych. W projekcie został zbadany czy przebieg wpływa na cenę auta.

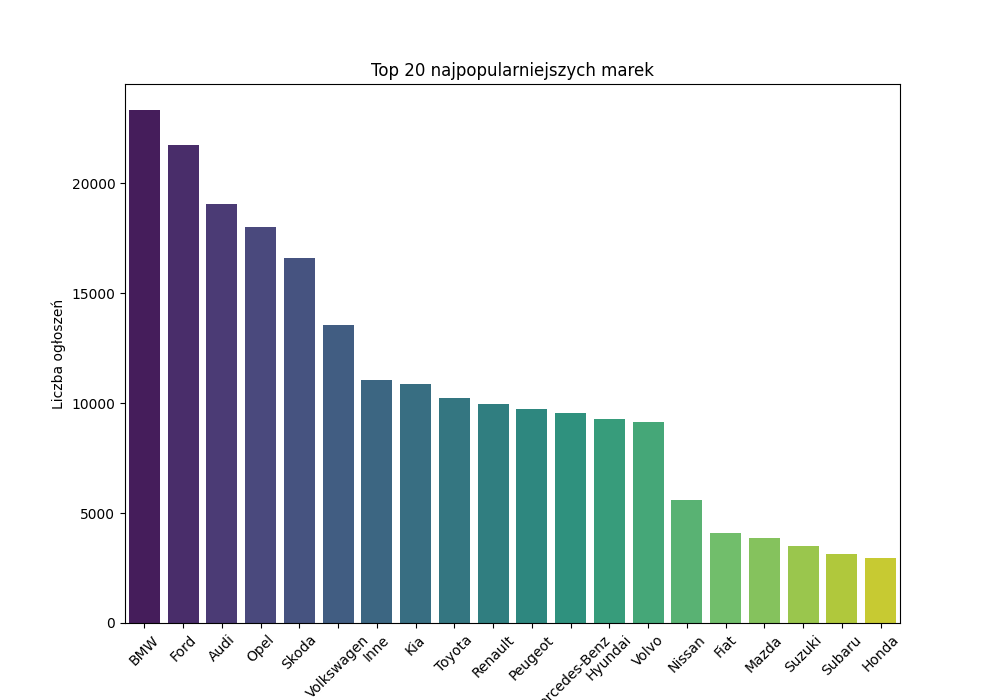
R-squared – współczynnik determinacji wynosi 0,102 czyli 10,2% zmienności ceny, można wyjaśnić przebiegiem. Sugeruje to, że cena zmienia się również przez inne czynniki.

P>|t| - wynosi 0 co oznacza, że zmienna przebieg jest istotna statystycznie. Nie jest to duży wpływ jak pokazuje R^2 ale nie znikomy.

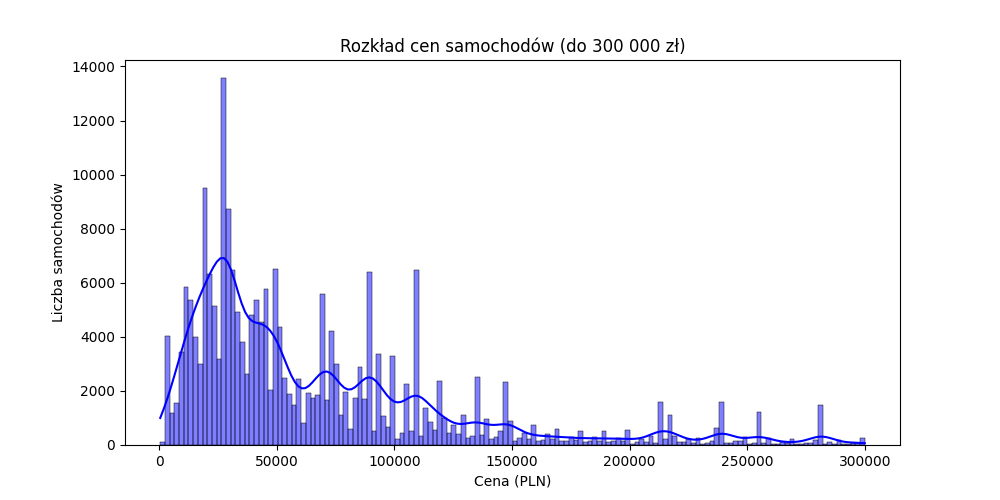
Coef – zapis 1.101e+05 przy const oznacza 110 100 zł. Oznacza że nowe auto o przebiegu 0 ma średnio wartość 110 100 zł. Mileage\_clean = -220 oznacza, że każdy dodatkowy 1 km przebiegu obniża cenę auta o średnio 220 zł. Jeśli auto ma 100 000 km to jego wartość spada o 22 mln zł co jest nielogiczne. Prawdopodobnie przebieg powinien być podany w tysiącach km zamiast pojedynczych kilometrów.

9. Wykresy analizy danych

Wracając do top 20 najpopularniejszych marek.

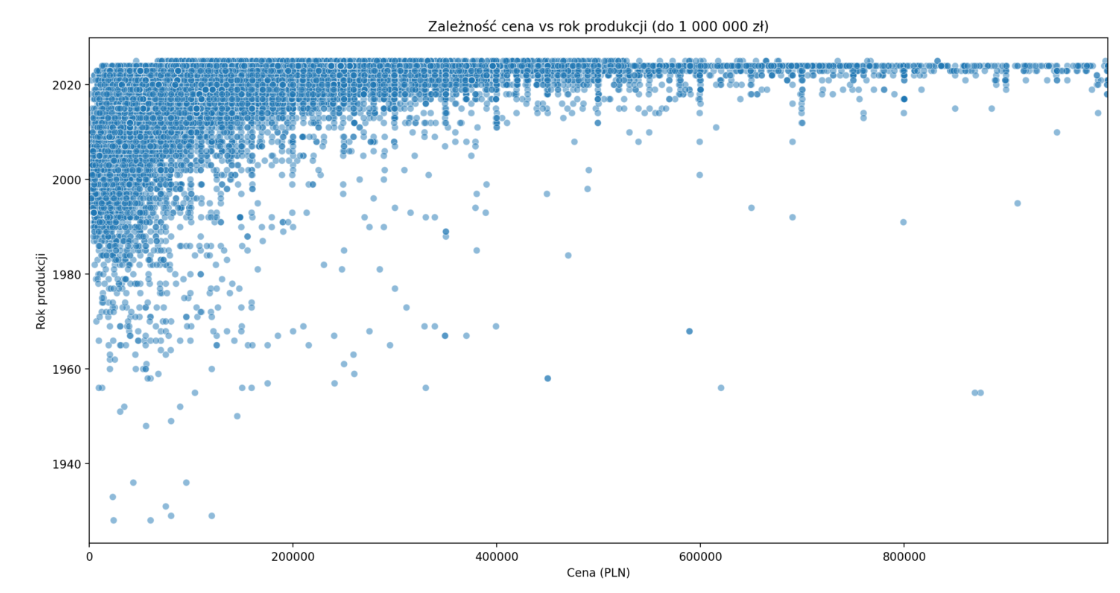
 Wykres ten pokazuje ciekawą zależność. Pierwsze 6 najpopularniejszych marek wzrasta w rownym stopniu. Następne 8 marek ma bardzo przybliżoną ilość samochodów na sprzedaż na poziomie 11000. Następne 6 marek również jest blisko tego samego poziomu pomiędzy sobą czyli około 5000 sztuk.

Rozkład cen samochodów



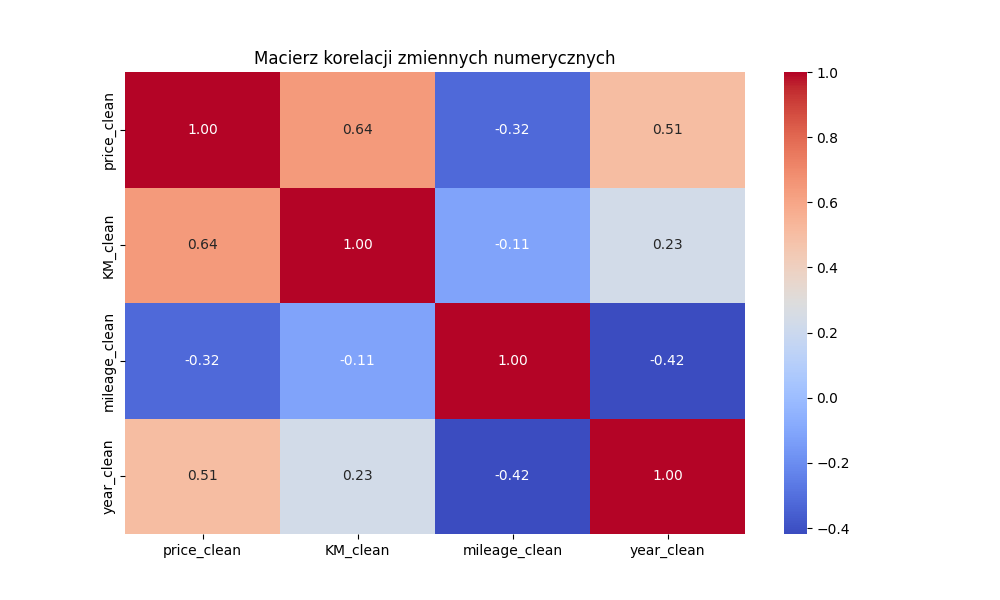
W celu lepszego przedstawienia danych, ceny bardzo wysokie czyli od 300 000 zł zostały pominięte. Dzięki temu widać które samochody mają najczęściej cenę. Jest to około 40 000 zł. Inna ciekawa obserwacja to są zwyżki ilości aut co około 10 tyś. aut. Nie jestem wstanie powiedzieć czemu tak się dzieje ale obserwacja jest ciekawa.

Zależność ceny od roku produkcji



Można z tego wykresu odczytać wyraźny wzrost ceny wraz z nowszym samochodem.

Macierz korelacji zmiennych numerycznych tj. rok produkcji, przebieg, KM i cena.



Korelacja ta pokazuje, że największy wpływ pomiędzy zależnościami to moc silnika wykazana w koniach mechanicznych a ceną, bo to aż 0,64.   
Najmniejszy wpływ, bo tylko -0,11 to wpływ pomiędzy przebiegiem a końmi mechanicznymi, co ma odzwierciedlenie w rzeczywistości, bo fakt te dwie zmienne nie mają ze sobą dużo wspólnego.  
Wysokimi wskaźnikami jest 0,51 oraz -0,42. 0,51 dla linii rok – cena oznacza, że jest spora korelacja pomiędzy tymi wartościami, im większy rok tym większa cena, co ma przełożenie na życie, nowy samochód z 2025 jest zwykle droższy niż samochód z 2010. Przebieg i rok też posiada wysoką korelację na poziomie -0,42. Wartość ujemna oznacza, że jest duża zależność tylko odwrotna, tzn. im mniejszy przebieg tym większy rok. To też jest logiczne, nowy samochód z tego roku ma zerowy przebieg stąd takie wyniki.

10. Tabele opisujące zależności

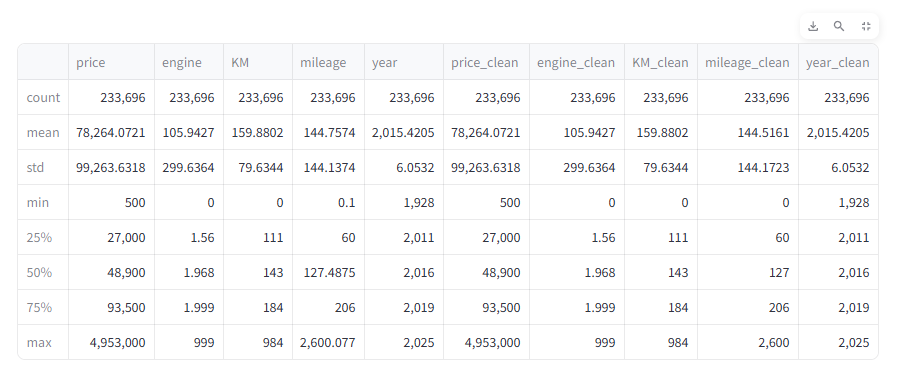


Tabela ta pokazuje różne obliczenia na wszystkich danych liczbowych.

Obserwacje jakie można tutaj znaleźć to m. in. Najwyższy przebieg, który wynosi aż 2 600 000 km, rok produkcji najstarszego samochodu dostępnego na Otomoto, to aż 1928 rok.

Dla innej tabeli jesteśmy w stanie zobaczyć dane w szybki sposób dla każdego województwa z osobna. Dla przykładu poniżej województwo Lubelskie. Z rozwijanej listy można wybrać inne województwa i z nich pobrać interesujące nas dane.

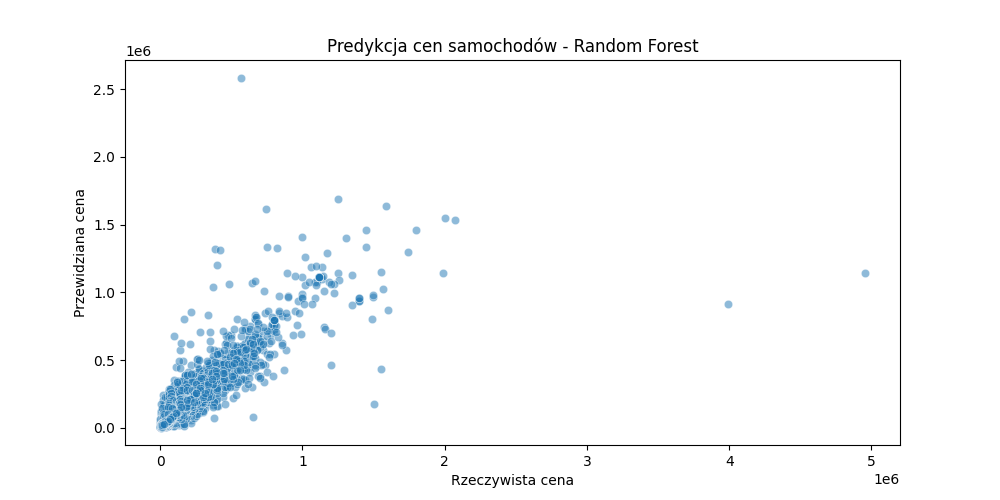


11. Predykcja ceny samochodów w oparciu o uczenie maszynowe.

Do tego celu zostały wykorzystane dwa modele uczenia maszynowego:

- Regresja liniowa R^2. Wynosi ona 0,54 czyli 54% - taka jest zmienność cen aut. Oznacza to że przy wykorzystaniu regresji liniowej model nie radzi sobie idealnie, ale coś już przewiduje.

- Random Forrest R^2 wynosi aż 0,861 i to już jest bardzo dobry wynik. 86% zmienności cen jest przewidywane poprawnie co oznacza, że model bardzo dobrze rozpoznaje zależności.

Wykres który przedstawia te uczenie się poniżej: