Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

Jakub Wujec

Nr albumu: 420463

# Determinanty cen aut w Polsce

Praca zaliczeniowa
na ćwiczenia z Ekonometrii
prowadzone przez dr Olgę Zajkowską

# 1. Wprowadzenie

Zasady działania mechanizmów rynku są jedną z ważnych dziedzin współczesnej ekonomii. Rynek aut osobowych w Polsce w ciągu ostatnich kilkunastu lat nieustannie się rozwija, o czym świadczą statystyki przedstawiające roczną liczbę rejestrowanych aut. W związku z tym coraz istotniejsze staje się zrozumienie tego coraz większego rynku. Znajomość determinantów cen aut pozwoli na lepsze zrozumienie rynku producentom, a co za tym idzie optymalniejszy dobór taryfy cenowej. Konsumentom za to, znając specyfikę rynku, łatwiej będzie znaleźć auto posiadające pożądane przez nich cechy a zarazem mieszczące się w ich ograniczeniu budżetowym.

# 2. Przegląd literatury

Ze względu na braki w literaturze opisującej użycie metod ekonometrycznych w celu analizy cen aut, w niniejszym badania uwaga zostanie skupiona na badaniach używających technik uczenia maszynowego do predykcji cen samochodów. O tym, że bardzo istotny wpływ na cenę auta posiada rok, marka oraz moc silnika piszą Nabarun Pal oraz współtwórcy w How much is my car worth? A methodology for predicting used cars prices using Random Forest. Zauważają oni znaczną różnice w wycenie aut poszczególnych marek, a także typu samochodu. Enis Gegic oraz współtwórcy w pracy Car Price Prediction using Machine Learning Techniques zmieniają podejście do predykcji ceny i zamieniają problem regresyjny w problem klasyfikacyjny, zamieniając ceny na kategorie 500-2000, 2000-3500 itd. Co więcej, oprócz standardowych charakterystyk auta używają oni również rodzaju tapicerki oraz występowania tempomatu.

# 3. Hipotezy badawcze

Zgodnie z literaturą, pierwszą hipotezą postawioną w badaniu jest dodatni wpływ roku produkcji, koni mechanicznych oraz marki danego samochodu. Co więcej, podstawną wydaje się być hipoteza, iż cena auta jest ujemnie skorelowana z jego przebiegiem. Innym przypuszczeniem może być również istotność interakcji zmiennej aso oraz mileage. Ostatnią hipotezą jest dodatni wpływ kwadratu koni mechanicznych na cenę auta.

# 4. Opis zbioru danych oraz definicje zmiennych

## 4.1 Zbiór danych

Ze względu na ciężki dostęp do zbiorów danych z aktualnymi cenami aut w Polsce, podjęta została decyzja o stworzeniu własnego zbioru danych. W tym celu została zescrapowana największa w Polsce platforma pośrednicząca w sprzedaży aut - otomoto. Napisany w tym celu crawler w języku Python pobrał ponad 104 000 wierszy zawierające po 21 kolumn. Ze względu na zanieczyszczony charakter danych pochodzących ze scrapowania, nasz zbiór danych musiał zostać poddany dokładnej obróbce.

## 4.2 Analiza zmiennych

Zmienną objaśniana:

- *price*

  Zmienna ta mówi o cenie za którą auto było wystawione na serwisie otomoto. Jest to cena brutto wyrażona w PLN.

Zmienne objaśniające:

- *aso*

  Zmienna ta mówi o tym czy dane auto było od nowości serwisowane w autoryzowanych serwisach obsługi (ASO). Są to wyspecjalizowane w danej marce serwisy posiadające akredytację danego producenta. Auta serwisowane w takich miejscach posiadają mniejszą szansę na doznanie usterki mechanicznej.

  *aso = 1* → auto było serwisowane w ASO

  *aso = 0* → auto nie było serwisowane w ASO

- *capacity*

  Zmienna ta mówi o pojemności skokowej. Jest to jeden z podstawowych charakterystyk opisujących silnik danego auta. W badaniu została ona wyrażona w centymetrach sześciennych. Jest to zmienna ciągła.

- *new*

  Zmienna ta mówi o tym czy auto jest nowe czy używane

  *new = 1* → auto jest nowe

  *new = 0* → auto jest używane

- *first_owner*

  Zmienna ta mówi o tym czy auto jest sprzedawane przez pierwszego właściciela

  *first_owner = 1* → auto jest sprzedawane przez pierwszego właściciela

  *first_owner = 0* → auto nie jest sprzedawane przez pierwszego właściciela

- *horse_power*

  Zmienna ta opisuje moc silnika w danym aucie. Wraz z pojemnością silnika jest to jedna z podstawowych charakterystyk opisujących silnik danego auta. Moc wpływa na dynamikę, przyspieszenie oraz prędkość maksymalną danego auta, co sprawia, iż jest jednym z czynników wpływających najmocniej na cenę auta. Jest to zmienna ciągła wyrażona w koniach mechanicznych.

- *mileage*

  Zmienna ta opisuje przebieg danego samochodu. Jest to liczba kilometrów jakie dany pojazd przejechał od wyjazdu z fabryki. Jest wyrażona w kilometrach.

- *no_accidents*

  Zmienna ta opisuje bezwypadkowość danego pojazdu.

  *no_accidents = 1* → auto nie uczestniczyło nigdy w wypadku.

  *no_accidents = 0* → auto uczestniczyło w wypadku.

- *number_of_doors*

  Zmienna ta opisuje liczbę drzwi w aucie. Wliczane są wszystkie drzwi z szybą którymi można wejść (także bagażnik). Dlatego sedan posiada 4 drzwi (otwierana część bagażnika bez okien), a auta kombi posiadające otwierany bagażnik - 5. Jest to zmienna ciągła.

- *automatyczna*

  Zmiana ta informuje o tym jaki typ skrzyni biegów dane auto posiada.
  *automatyczna = 1* → auto posiada automatyczną skrzynie biegów.
  *automatyczna = 0* → auto posiada manualną skrzynie biegów.

- *year*

  Zmienna ta opisuję rok produkcji danego auta. Jest to zmienna ciągła.

- *brand*

  Zmienna model opisuję markę danego samochodu. Dane zawierały informację na temat następujących 29 marek samochodów:

  Alfa Romeo, Audi, BMW, Bentley, Chevrolet, Citroën, Dacia, Ferrari, Fiat, Kia, Lamborghini, Land Rover, Lexus, MINI, Maserati, Mazda, McLaren, Mercedes-Benz, Mitsubishi, Porsche, Renault, Rolls-Royce, Saab, Seat, Suzuki, Toyota, Volvo, brand, Škoda

  Zmienna została rozkodowana za zmienne binarne w wyniku czego powstało 28 zmiennych (jedna z nich została usunięta, żeby uniknąć współliniowości). Współczynniki przy tych zmiennych będą oznaczać jak przynależność do danej marki auta wpływa na cenę.

- *features*

  W każdej obserwacji zmienna features zawiera listę wszystkich elementów wyposażenia.
  Łącznie możliwych elementów wyposażenia jest 71:
  ABS, ASR (kontrola trakcji), Alarm, Alufelgi, Asystent parkowania, Asystent pasa ruchu, Bluetooth, CD, Centralny zamek, Czujnik deszczu, Czujnik martwego pola, Czujnik zmierzchu, Czujniki parkowania przednie, Czujniki parkowania tylne, Dach panoramiczny, ESP (stabilizacja toru jazdy), Elektrochromatyczne lusterka boczne, Elektrochromatyczne lusterko wsteczne, Elektryczne szyby przednie, Elektryczne szyby tylne, Elektrycznie ustawiane fotele, Elektrycznie ustawiane lusterka, Gniazdo AUX, Gniazdo SD, Gniazdo USB, HUD (wyświetlacz przezierny), Hak, Immobilizer, Isofix, Kamera cofania, Klimatyzacja automatyczna, Klimatyzacja czterostrefowa, Klimatyzacja dwustrefowa, Klimatyzacja manualna, Komputer pokładowy, Kurtyny powietrzne, MP3, Nawigacja GPS, Odtwarzacz DVD, Ogranicznik prędkości, Ogrzewanie postojowe, Podgrzewana przednia szyba, Podgrzewane lusterka boczne, Podgrzewane przednie siedzenia, Podgrzewane tylne siedzenia, Poduszka powietrzna chroniąca kolana, Poduszka powietrzna kierowcy, Poduszka powietrzna pasażera, Poduszki boczne przednie, Poduszki boczne tylne, Przyciemniane szyby, Radio fabryczne, Radio niefabryczne, Regulowane zawieszenie, Relingi dachowe, System Start-Stop, Szyberdach, Tapicerka skórzana, Tapicerka welurowa, Tempomat, Tempomat aktywny, Tuner TV, Wielofunkcyjna kierownica, Wspomaganie kierownicy, Zmieniarka CD, features, Łopatki zmiany biegów, Światła LED, Światła Xenonowe, Światła do jazdy dziennej, Światła przeciwmgielne

Zmienna ta nie zostanie uwzględniona w modelu z powodu zbyt wielu parametrów oraz brakach w danych. W serwisie otomoto użytkownicy bardzo często wpisują wyposażenie do opisu zamiast pól z wybranymi dodatkami co sprawia, że dane są niepełne.

# 5. Wstępna analiza danych

## 5.1 Statystyka opisowa zmiennych

Poniższa tabela przedstawia statystyki takie jak średnia, odchylenie standardowe, wartość minimalna, wartość maksymalna oraz kwantyle zmiennych objaśniających z wykluczeniem *brand* oraz *features*.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **aso** | 80165.0 | 0.522036 | 0.499517 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| **capacity** | 80165.0 | 1936.723807 | 774.353906 | 647.0 | 1498.0 | 1896.0 | 1998.0 | 7011.0 |
| **new** | 80165.0 | 0.125816 | 0.331643 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| **first_owner** | 80165.0 | 0.372993 | 0.483603 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| **horse_power** | 80165.0 | 165.771933 | 89.393867 | 1.0 | 110.0 | 143.0 | 190.0 | 1321.0 |
| **mileage** | 80165.0 | 122414.565833 | 93871.439095 | 1.0 | 38000.0 | 118140.0 | 189716.0 | 999999.0 |
| **no_accidents** | 80165.0 | 0.630599 | 0.482646 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| **number_of_doors** | 80165.0 | 4.618512 | 0.799441 | 0.0 | 5.0 | 5.0 | 5.0 | 6.0 |
| **price** | 80165.0 | 94327.677640 | 130995.259228 | 690.0 | 23990.0 | 53900.0 | 115900.0 | 3500000.0 |
| **automatyczna** | 80165.0 | 0.451731 | 0.497668 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| **year** | 80165.0 | 2013.711732 | 6.226966 | 1951.0 | 2009.0 | 2015.0 | 2019.0 | 2022.0 |
| **Benzyna** | 80165.0 | 0.545812 | 0.497900 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| **Benzyna+LPG** | 80165.0 | 0.037323 | 0.189553 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| **Diesel** | 80165.0 | 0.371409 | 0.483184 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| **Hybryda** | 80165.0 | 0.045456 | 0.208304 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

*Tabela 1. Statystyka opisowa wybranych zmiennych*

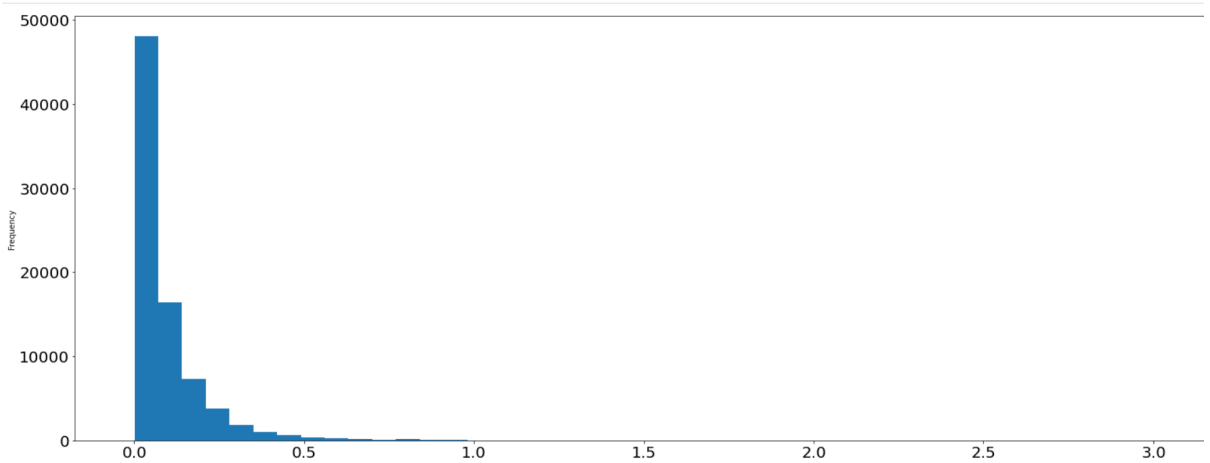
Ciekawym wydaje się fakt średniego oraz mediany wieku aut - kolejno 2013.7 oraz 2015. Oznacza to, że auta oferowane na serwisie otomoto są stosunkowo młode. Co więcej średnia moc na poziomie 165 km oraz średnia pojemność na poziomie 1937 mogą wskazywać na częste użycie mniejszych ale stosunkowo mocnych silników. Można również wywnioskować, iż najwięcej aut napędzanych jest na benzynę.

Macierz korelacji między zmiennymi prezentuję się w następujący sposób:

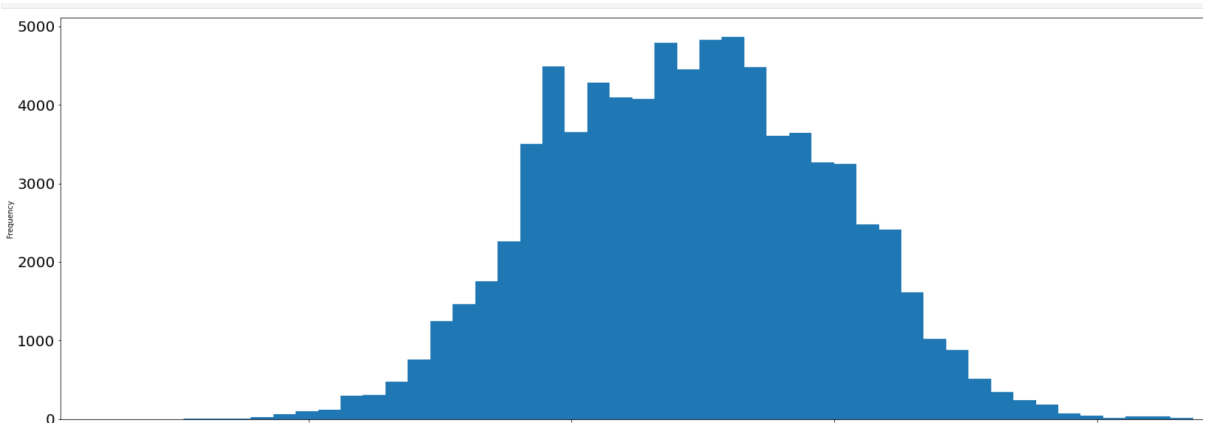
| | aso | capacity | new | first_owner | horse_power | mileage | no_accidents | number_of_doors | automatyczna | year |
|---|---|---|---|---|---|---|---|---|---|---|
| **aso** | 1.000000 | -0.039273 | -0.367110 | 0.513223 | -0.004226 | -0.024231 | 0.254934 | 0.065847 | 0.009661 | 0.108462 |
| **capacity** | -0.039273 | 1.000000 | -0.019230 | -0.063047 | 0.832994 | 0.105371 | -0.035211 | -0.243028 | 0.461263 | -0.142124 |
| **new** | -0.367110 | -0.019230 | 1.000000 | -0.243603 | 0.131864 | -0.494277 | 0.290361 | 0.036967 | 0.231869 | 0.454689 |
| **first_owner** | 0.513223 | -0.063047 | -0.243603 | 1.000000 | 0.000742 | -0.147024 | 0.205037 | 0.066754 | 0.056123 | 0.227053 |
| **horse_power** | -0.004226 | 0.832994 | 0.131864 | 0.000742 | 1.000000 | -0.157548 | 0.054841 | -0.219550 | 0.552674 | 0.163970 |
| **mileage** | -0.024231 | 0.105371 | -0.494277 | -0.147024 | -0.157548 | 1.000000 | -0.212445 | -0.010195 | -0.275358 | -0.735862 |
| **no_accidents** | 0.254934 | -0.035211 | 0.290361 | 0.205037 | 0.054841 | -0.212445 | 1.000000 | 0.060554 | 0.104653 | 0.230649 |
| **number_of_doors** | 0.065847 | -0.243028 | 0.036967 | 0.066754 | -0.219550 | -0.010195 | 0.060554 | 1.000000 | -0.065724 | 0.174366 |
| **automatyczna** | 0.009661 | 0.461263 | 0.231869 | 0.056123 | 0.552674 | -0.275358 | 0.104653 | -0.065724 | 1.000000 | 0.327302 |
| **year** | 0.108462 | -0.142124 | 0.454689 | 0.227053 | 0.163970 | -0.735862 | 0.230649 | 0.174366 | 0.327302 | 1.000000 |

*Tabela 2: Macierz korelacji wybranych zmiennych*

## 5.2 Analiza zmiennych ciągłych



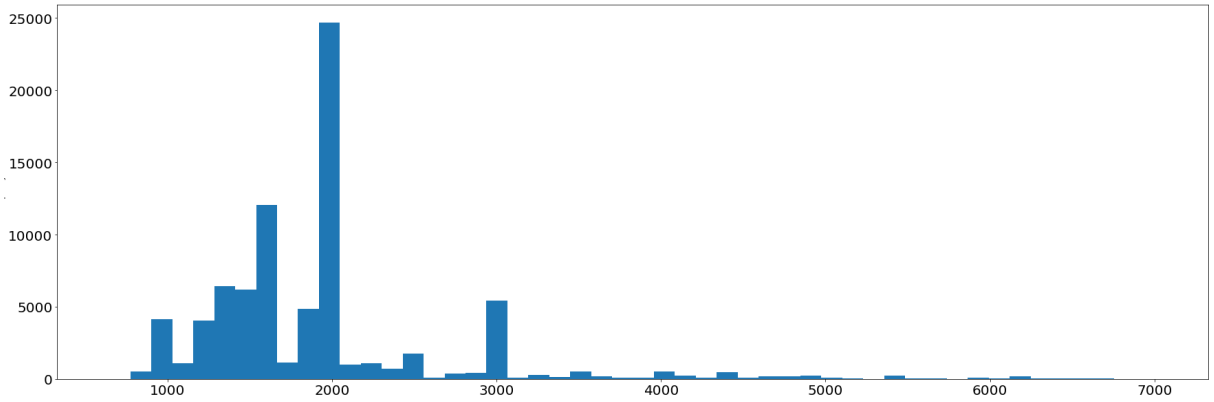

*Wykres 1. Histogram cen samochodów (cena w 1e6)*



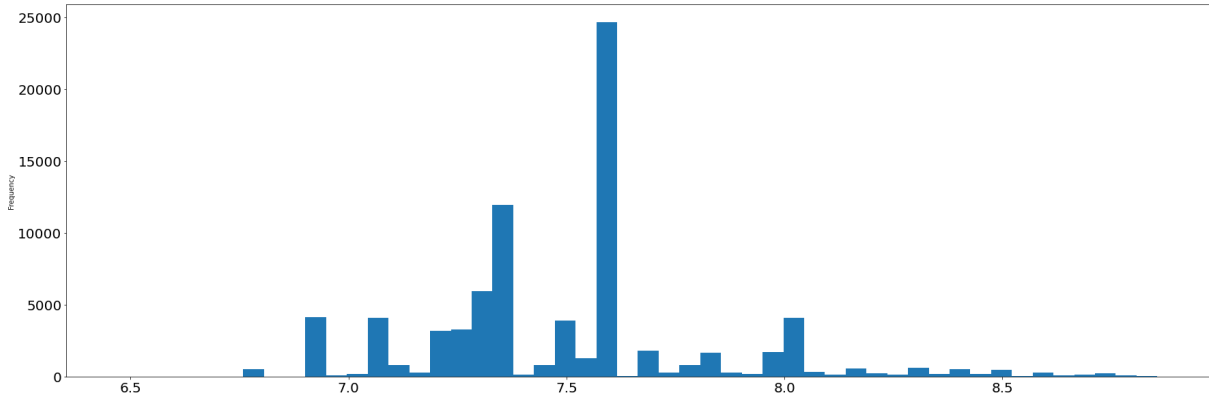

*Wykres 2. Histogram logarytmu cen samochodów (cena w 1e6)*

Wykres 1. przedstawia histogram cen samochodów oferowanych na serwisie otomoto. Jak można się było spodziewać, najwięcej jest aut kosztujących stosunkowo mało. Jak można zauważyć gołym

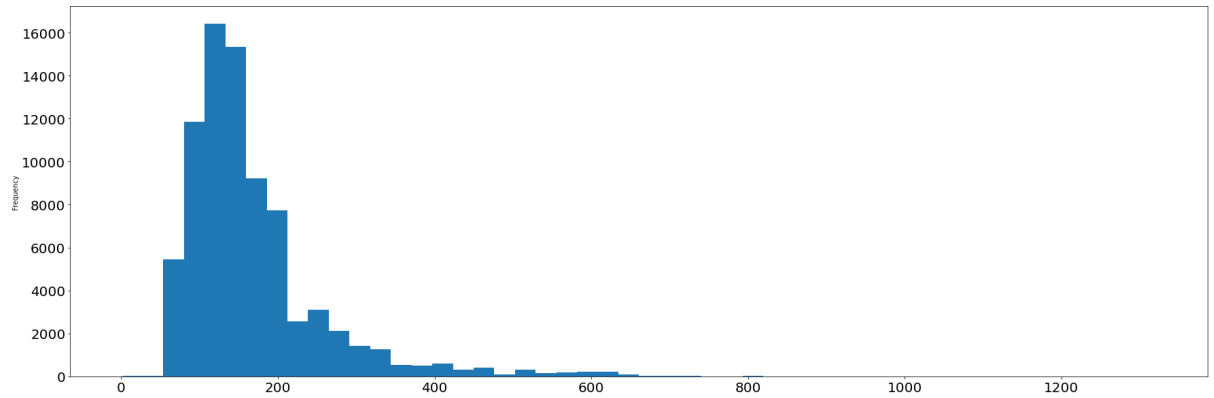
okiem, rozkład oryginalnych danych daleko jest od rozkładu normalnego. W tym celu cena została zlogarytmowana. Niestety w obydwu przypadkach test Jarque-Bera zwrócił p-value poniżej wybranego poziomu istotności = 0.05, a więc w obydwu przypadkach rozkład nie jest normalny.



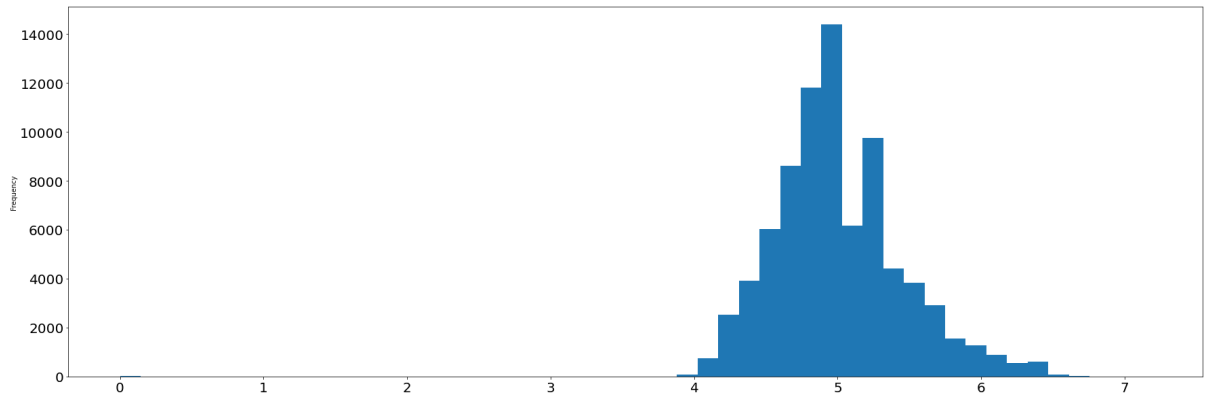

*Wykres 3. Histogram pojemności skokowej silnika.*



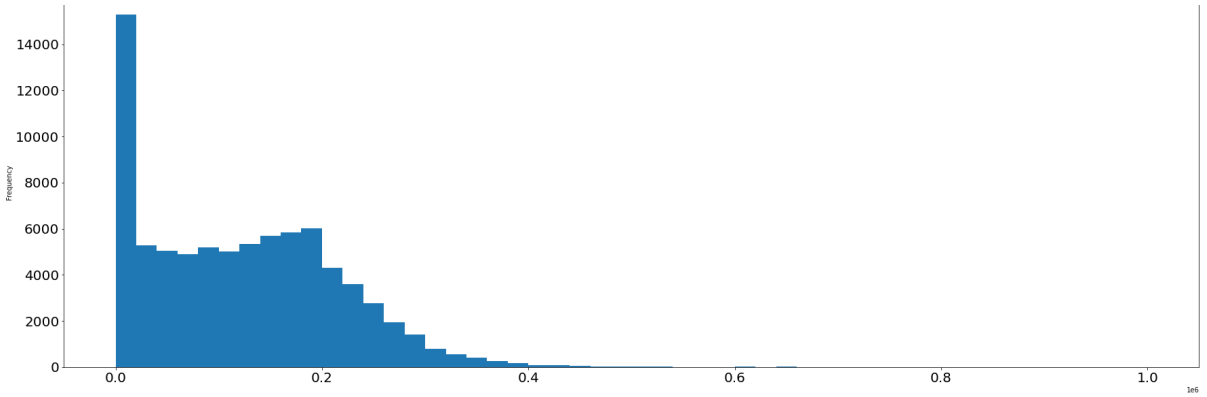

*Wykres 4. Histogram logarytmu pojemności skokowej silnika*



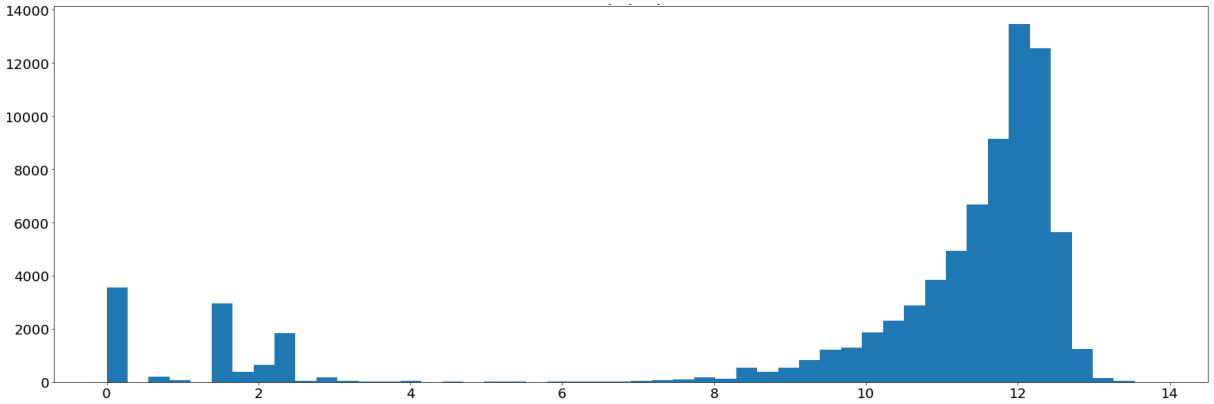

*Wykres 5. Histogram mocy silnika*



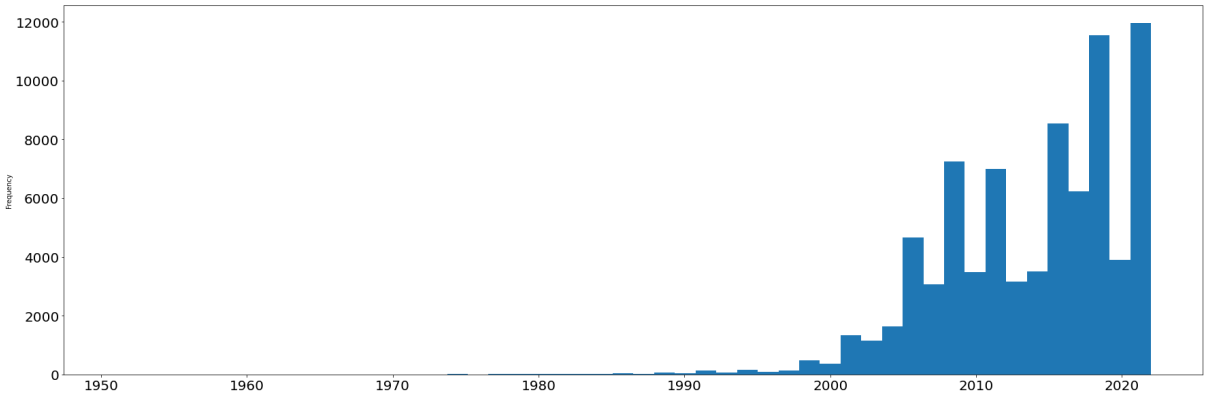

*Wykres 6. Histogram logarytmu mocy silnika*



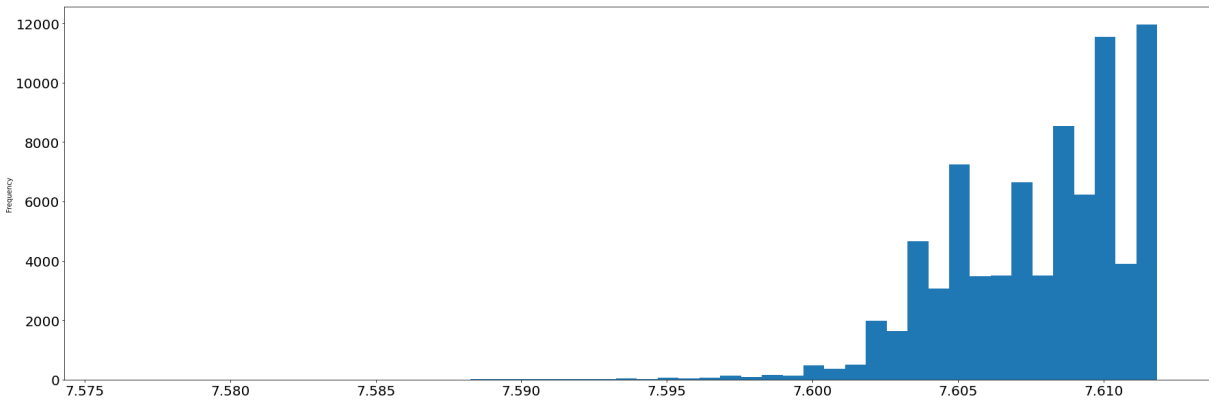

*Wykres 6. Histogram przebiegu*




*Wykres 7. Histogram logarytmu przebiegu*




*Wykres 8. Histogram roku produkcji*




*Wykres 9. Histogram logarytmu roku produkcji*

Na wykresach 3 - 9 przedstawiono histogramy oraz histogramy logarytmów pozostałych zmiennych ciągłych. Widać niestety, co również potwierdzają testy Jarque-Bera, że zarówno zmienne jak i ich logarytmy nie posiadają rozkładu normalnego.

## 5.3 Analiza współliniowości

W celu analizy w współliniowości został użyty parametr VIF. Według literatury przyjmuje się, że zmienne są skorelowane, jeżeli parametr VIF jest większy od 10. W naszym przypadku parametrami współliniowymi okazały się *capacity, horse_power, number_of_doors,* oraz *year.* Z tego względu do finałowego modelu zostały usunięte zmienne *number_of_doors oraz capacity*.
Zmienne *year* oraz *horse_power* zostały, gdyż wydają się one zbyt istotne w predykcji ceny auta by je usunąć.

| | feature | VIF |
|---|---|---|
| **0** | aso | 3.548800 |
| **1** | capacity | 31.055191 |
| **2** | new | 2.383357 |
| **3** | first_owner | 2.372844 |
| **4** | horse_power | 20.058373 |
| **5** | mileage | 5.125307 |
| **6** | no_accidents | 3.618335 |
| **7** | number_of_doors | 38.116451 |
| **8** | automatyczna | 3.098476 |
| **9** | year | 55.396071 |
| **10** | Benzyna+LPG | 1.111113 |
| **11** | Diesel | 1.997970 |
| **12** | Hybryda | 1.200736 |

*Tabela 3. Parametr VIF dla zmiennych objaśnianych*

# 6. Wyniki modelu

## 6.1 Estymacja parametrów modelu

W pierwszej iteracji modelu zostały użyte wszystkie zmienne oprócz *features.* Wyniki prezentują się w następujący sposób:

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.746 |
| Model: | OLS | Adj. R-squared: | 0.746 |
| Method: | Least Squares | F-statistic: | 5886. |
| Date: | Tue, 25 Jan 2022 | Prob (F-statistic): | 0.00 |
| Time: | 22:51:38 | Log-Likelihood: | -1.0034e+06 |
| No. Observations: | 80165 | AIC: | 2.007e+06 |
| Df Residuals: | 80124 | BIC: | 2.007e+06 |
| Df Model: | 40 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -3.97e+06 | 1.41e+05 | -28.133 | 0.000 | -4.25e+06 | -3.69e+06 |
| aso | 1200.4872 | 613.694 | 1.956 | 0.050 | -2.348 | 2403.323 |
| capacity | -25.5909 | 0.720 | -35.545 | 0.000 | -27.002 | -24.180 |
| new | 5.741e+04 | 1047.127 | 54.821 | 0.000 | 5.54e+04 | 5.95e+04 |
| first_owner | 1.366e+04 | 596.931 | 22.884 | 0.000 | 1.25e+04 | 1.48e+04 |
| horse_power | 911.9650 | 6.486 | 140.601 | 0.000 | 899.252 | 924.678 |
| mileage | -0.3001 | 0.004 | -69.977 | 0.000 | -0.308 | -0.292 |
| no_accidents | 1.274e+04 | 561.431 | 22.700 | 0.000 | 1.16e+04 | 1.38e+04 |
| number_of_doors | -4653.6554 | 325.183 | -14.311 | 0.000 | -5291.012 | -4016.299 |
| automatyczna | -6817.6556 | 661.931 | -10.300 | 0.000 | -8115.035 | -5520.276 |
| year | 1964.0208 | 70.102 | 28.017 | 0.000 | 1826.621 | 2101.420 |
| Audi | 4.375e+04 | 2110.003 | 20.732 | 0.000 | 3.96e+04 | 4.79e+04 |
| BMW | 3.927e+04 | 2111.404 | 18.600 | 0.000 | 3.51e+04 | 4.34e+04 |
| Bentley | 1.894e+05 | 8011.116 | 23.641 | 0.000 | 1.74e+05 | 2.05e+05 |
| Chevrolet | 2.129e+04 | 2739.667 | 7.772 | 0.000 | 1.59e+04 | 2.67e+04 |
| Citroën | 2.389e+04 | 2208.641 | 10.814 | 0.000 | 1.96e+04 | 2.82e+04 |
| Dacia | -297.3549 | 3223.145 | -0.092 | 0.926 | -6614.699 | 6019.989 |
| Ferrari | 4.742e+05 | 6767.570 | 70.070 | 0.000 | 4.61e+05 | 4.87e+05 |
| Fiat | 2.335e+04 | 4629.283 | 5.044 | 0.000 | 1.43e+04 | 3.24e+04 |
| Kia | 1.086e+04 | 2207.075 | 4.923 | 0.000 | 6539.027 | 1.52e+04 |
| Lamborghini | 8.277e+05 | 1.01e+04 | 82.321 | 0.000 | 8.08e+05 | 8.47e+05 |
| Land Rover | 8.813e+04 | 2880.887 | 30.591 | 0.000 | 8.25e+04 | 9.38e+04 |
| Lexus | 3.556e+04 | 2912.568 | 12.209 | 0.000 | 2.98e+04 | 4.13e+04 |
| MINI | 1.769e+04 | 2358.554 | 7.501 | 0.000 | 1.31e+04 | 2.23e+04 |
| Maserati | -1.72e+04 | 5475.883 | -3.141 | 0.002 | -2.79e+04 | -6468.457 |
| Mazda | 2.936e+04 | 2336.764 | 12.563 | 0.000 | 2.48e+04 | 3.39e+04 |
| McLaren | 4.453e+05 | 1.67e+04 | 26.635 | 0.000 | 4.13e+05 | 4.78e+05 |
| Mercedes-Benz | 5.436e+04 | 2122.796 | 25.610 | 0.000 | 5.02e+04 | 5.85e+04 |
| Mitsubishi | 2.985e+04 | 2554.151 | 11.687 | 0.000 | 2.48e+04 | 3.49e+04 |
| Porsche | 8.053e+04 | 3104.318 | 25.941 | 0.000 | 7.44e+04 | 8.66e+04 |
| Renault | 1.992e+04 | 2133.965 | 9.336 | 0.000 | 1.57e+04 | 2.41e+04 |
| Rolls-Royce | 1.151e+06 | 8064.302 | 142.675 | 0.000 | 1.13e+06 | 1.17e+06 |
| Saab | 2.204e+04 | 3917.500 | 5.627 | 0.000 | 1.44e+04 | 2.97e+04 |
| Seat | 2.733e+04 | 2294.211 | 11.911 | 0.000 | 2.28e+04 | 3.18e+04 |
| Suzuki | 3.174e+04 | 2481.903 | 12.787 | 0.000 | 2.69e+04 | 3.66e+04 |
| Toyota | 3.241e+04 | 2167.011 | 14.954 | 0.000 | 2.82e+04 | 3.67e+04 |
| Volvo | 2.989e+04 | 2367.788 | 12.625 | 0.000 | 2.53e+04 | 3.45e+04 |
| Škoda | 2.285e+04 | 2155.925 | 10.600 | 0.000 | 1.86e+04 | 2.71e+04 |
| Benzyna+LPG | 1.332e+04 | 1295.088 | 10.286 | 0.000 | 1.08e+04 | 1.59e+04 |
| Diesel | 3.129e+04 | 600.744 | 52.088 | 0.000 | 3.01e+04 | 3.25e+04 |
| Hybryda | 3.218e+04 | 1285.149 | 25.042 | 0.000 | 2.97e+04 | 3.47e+04 |

| | | | |
|---|---|---|---|
| Omnibus: | 86851.565 | Durbin-Watson: | 1.614 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 77643218.632 |
| Skew: | 4.653 | Prob(JB): | 0.00 |
| Kurtosis: | 155.179 | Cond. No. | 9.34e+07 |

*Tabela 4: Wynik pierwszej iteracji regresji*

W pierwszej iteracji z wydruku można zauważyć, iż prawie wszystkie zmienne są istotne statystycznie. Względnie nieprawdopodobna wydaje się stała będąca liczbą ujemną. Można ją jednak wytłumaczyć pozytywnymi współczynnikami przy wszystkich statystycznie istotnych markach oraz koniami mechanicznymi które zawszę są większe od zera. Zmienne aso, capacity oraz horse_power zwracają oczekiwane rezultaty, zmiana każdej z tych jednostek wpływa pozytywnie na cenę auta. Wzrost mocy o 1 koń mechaniczny zwiększa cenę o 911 zł. Ciekawą anomalią wydaje się być negatywny wpływ marki maseratti, która zmniejsza cenę auta o 17000 zł. Warto zauważyć, że auta bardzo luksusowe zauważalnie zwiększają cenę o cały rząd wielkości, co jest oczekiwanym i logicznym wynikiem. Również interesującym wydaje się być fakt, iż Dacia jako jedyna marka samochodu jest nieistotna statystycznie. Co ciekawe, *number of doors* wpływa negatywnie na cenę

auta, zmniejszając ją o 4653 zł na każde jedne drzwi. Może to być związane z faktem, iż drogie auta sportowe posiadają jedynie 3 drzwi. Logicznym wydaje się również wniosek płynący z współczynnika przy zmiennej *mileage*. Oznacza on, iż wzrost przebiegu o 1 km zmniejsza jego cenę o 30 groszy. Wydaje się to nie być dużo, jednak biorąc pod uwagę fakt, iż przebiegi potrafią osiągać wartości rzędu kilkuset tysięcy kilometrów, parametr jest dosyć wysoki. Wnioski wyniesione z modelu wydają się być zgodne z przedstawionymi wcześniej hipotezami oraz z podstawami teoretycznymi, jednak ze względu na test RESET wskazujący na nieprawidłową formę funkcyjną, test Breusch–Pagana wskazujący na homoskedastyczność oraz test Jarque-Bera wskazujący na nieistotność reszt, interpretacja wskaźników modelu nie ma sensu. Ze względu na test RESET nie możemy wierzyć w oszacowania współczynników przy zmiennych objaśniających, a przez homoskedastyczność oraz brak normalności reszt nie możemy interpretować zarówno odchyleń standardowych jak i p-value z wydruku.

W celu poprawy charakterystyk modelu następnej iteracji został wzięty logarytm zmiennej objaśnianej *price*.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.879 |
| Model: | OLS | Adj. R-squared: | 0.879 |
| Method: | Least Squares | F-statistic: | 1.462e+04 |
| Date: | Tue, 18 Jan 2022 | Prob (F-statistic): | 0.00 |
| Time: | 22:30:21 | Log-Likelihood: | -34355. |
| No. Observations: | 80165 | AIC: | 6.879e+04 |
| Df Residuals: | 80124 | BIC: | 6.917e+04 |
| Df Model: | 40 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -146.1070 | 0.794 | -184.013 | 0.000 | -147.663 | -144.551 |
| aso | 0.0832 | 0.003 | 24.080 | 0.000 | 0.076 | 0.090 |
| capacity | 0.0002 | 4.05e-06 | 39.680 | 0.000 | 0.000 | 0.000 |
| new | 0.1765 | 0.006 | 29.961 | 0.000 | 0.165 | 0.188 |
| first_owner | 0.0931 | 0.003 | 27.704 | 0.000 | 0.086 | 0.100 |
| horse_power | 0.0029 | 3.65e-05 | 78.461 | 0.000 | 0.003 | 0.003 |
| mileage | -2.765e-06 | 2.41e-08 | -114.585 | 0.000 | -2.81e-06 | -2.72e-06 |
| no_accidents | 0.1317 | 0.003 | 41.677 | 0.000 | 0.125 | 0.138 |
| number_of_doors | -0.0161 | 0.002 | -8.783 | 0.000 | -0.020 | -0.012 |
| automatyczna | 0.2285 | 0.004 | 61.358 | 0.000 | 0.221 | 0.236 |
| year | 0.0775 | 0.000 | 196.536 | 0.000 | 0.077 | 0.078 |
| Audi | 0.3199 | 0.012 | 26.945 | 0.000 | 0.297 | 0.343 |
| BMW | 0.2867 | 0.012 | 24.130 | 0.000 | 0.263 | 0.310 |
| Bentley | 0.3887 | 0.045 | 8.623 | 0.000 | 0.300 | 0.477 |
| Chevrolet | -0.1778 | 0.015 | -11.533 | 0.000 | -0.208 | -0.148 |
| Citroën | -0.1338 | 0.012 | -10.765 | 0.000 | -0.158 | -0.109 |
| Dacia | -0.0929 | 0.018 | -5.122 | 0.000 | -0.128 | -0.057 |
| Ferrari | 0.6382 | 0.038 | 16.759 | 0.000 | 0.564 | 0.713 |
| Fiat | -0.0180 | 0.026 | -0.692 | 0.489 | -0.069 | 0.033 |
| Kia | 0.0533 | 0.012 | 4.292 | 0.000 | 0.029 | 0.078 |
| Lamborghini | 0.7069 | 0.057 | 12.494 | 0.000 | 0.596 | 0.818 |
| Land Rover | 0.4171 | 0.016 | 25.730 | 0.000 | 0.385 | 0.449 |
| Lexus | 0.3152 | 0.016 | 19.229 | 0.000 | 0.283 | 0.347 |
| MINI | 0.1154 | 0.013 | 8.691 | 0.000 | 0.089 | 0.141 |
| Maserati | 0.1793 | 0.031 | 5.820 | 0.000 | 0.119 | 0.240 |
| Mazda | 0.0567 | 0.013 | 4.312 | 0.000 | 0.031 | 0.082 |
| McLaren | 0.7114 | 0.094 | 7.562 | 0.000 | 0.527 | 0.896 |
| Mercedes-Benz | 0.3845 | 0.012 | 32.191 | 0.000 | 0.361 | 0.408 |
| Mitsubishi | 0.0788 | 0.014 | 5.480 | 0.000 | 0.051 | 0.107 |
| Porsche | 0.5794 | 0.017 | 33.165 | 0.000 | 0.545 | 0.614 |
| Renault | -0.0781 | 0.012 | -6.507 | 0.000 | -0.102 | -0.055 |
| Rolls-Royce | 1.0276 | 0.045 | 22.644 | 0.000 | 0.939 | 1.117 |
| Saab | -0.0831 | 0.022 | -3.769 | 0.000 | -0.126 | -0.040 |
| Seat | 0.0146 | 0.013 | 1.132 | 0.258 | -0.011 | 0.040 |
| Suzuki | 0.1024 | 0.014 | 7.331 | 0.000 | 0.075 | 0.130 |
| Toyota | 0.1567 | 0.012 | 12.849 | 0.000 | 0.133 | 0.181 |
| Volvo | 0.2710 | 0.013 | 20.341 | 0.000 | 0.245 | 0.297 |
| Škoda | 0.0881 | 0.012 | 7.260 | 0.000 | 0.064 | 0.112 |
| Benzyna+LPG | 0.0150 | 0.007 | 2.053 | 0.040 | 0.001 | 0.029 |
| Diesel | 0.1659 | 0.003 | 49.087 | 0.000 | 0.159 | 0.173 |
| Hybryda | 0.1059 | 0.007 | 14.639 | 0.000 | 0.092 | 0.120 |

| | | | |
|---|---|---|---|
| Omnibus: | 26665.178 | Durbin-Watson: | 1.806 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 1082018.936 |
| Skew: | 0.901 | Prob(JB): | 0.00 |
| Kurtosis: | 20.908 | Cond. No. | 9.34e+07 |

*Tabela 5. Wyniki drugiej iteracji regresji*

W przypadku tego modelu, ze względu na zlogarytmowaną cenę, parametry w modelu powinny być interpretowane w inny sposób. W tym przypadku fakt iż auto jest nowe, zwiększa jego cenę o około 176%. Co ciekawe, w tym modelu jedyną marką auta nieistotną statystycznie okazał się Fiat. W przypadku tego modelu, skrzynia automatyczna pozytywnie wpływa na cenę auta, zwiększając ją o 23%. W tym przypadku również można zauważyć, iż *number_of_doors* wpływa negatywnie na cenę auta. Niestety również w tym przypadku testy pokazały, iż model ten nie opisuje w dobry sposób rzeczywistości.

W celu poprawy formy funkcyjnej został stworzony kolejny model. Została tutaj dodana interakcja pomiędzy zmienną *aso* a *mileage*. Dodano także kwadrat zmiennej *horse_power*.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.884 |
| Model: | OLS | Adj. R-squared: | 0.884 |
| Method: | Least Squares | F-statistic: | 1.528e+04 |
| Date: | Tue, 18 Jan 2022 | Prob (F-statistic): | 0.00 |
| Time: | 22:31:43 | Log-Likelihood: | -32779. |
| No. Observations: | 80165 | AIC: | 6.564e+04 |
| Df Residuals: | 80124 | BIC: | 6.602e+04 |
| Df Model: | 40 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -132.5664 | 0.710 | -186.801 | 0.000 | -133.957 | -131.175 |
| aso | 0.1122 | 0.006 | 19.268 | 0.000 | 0.101 | 0.124 |
| new | 0.1808 | 0.006 | 29.050 | 0.000 | 0.169 | 0.193 |
| first_owner | 0.0893 | 0.003 | 26.816 | 0.000 | 0.083 | 0.096 |
| horse_power | 0.0081 | 6.06e-05 | 132.966 | 0.000 | 0.008 | 0.008 |
| mileage | -2.755e-06 | 2.82e-08 | -97.739 | 0.000 | -2.81e-06 | -2.7e-06 |
| no_accidents | 0.1277 | 0.003 | 41.203 | 0.000 | 0.122 | 0.134 |
| automatyczna | 0.1700 | 0.004 | 44.975 | 0.000 | 0.163 | 0.177 |
| year | 0.0706 | 0.000 | 200.356 | 0.000 | 0.070 | 0.071 |
| Audi | 0.3606 | 0.012 | 31.076 | 0.000 | 0.338 | 0.383 |
| BMW | 0.3381 | 0.012 | 29.111 | 0.000 | 0.315 | 0.361 |
| Bentley | 1.0234 | 0.045 | 22.956 | 0.000 | 0.936 | 1.111 |
| Chevrolet | -0.0177 | 0.015 | -1.178 | 0.239 | -0.047 | 0.012 |
| Citroën | -0.0024 | 0.012 | -0.194 | 0.846 | -0.026 | 0.022 |
| Dacia | 0.0562 | 0.018 | 3.161 | 0.002 | 0.021 | 0.091 |
| Ferrari | 1.4100 | 0.039 | 36.558 | 0.000 | 1.334 | 1.486 |
| Fiat | 0.1525 | 0.026 | 5.968 | 0.000 | 0.102 | 0.203 |
| Kia | 0.1565 | 0.012 | 12.921 | 0.000 | 0.133 | 0.180 |
| Lamborghini | 1.4631 | 0.057 | 25.893 | 0.000 | 1.352 | 1.574 |
| Land Rover | 0.4703 | 0.016 | 29.756 | 0.000 | 0.439 | 0.501 |
| Lexus | 0.3846 | 0.016 | 24.176 | 0.000 | 0.353 | 0.416 |
| MINI | 0.2129 | 0.013 | 16.396 | 0.000 | 0.187 | 0.238 |
| Maserati | 0.2201 | 0.030 | 7.282 | 0.000 | 0.161 | 0.279 |
| Mazda | 0.1705 | 0.013 | 13.424 | 0.000 | 0.146 | 0.195 |
| McLaren | 1.2583 | 0.093 | 13.579 | 0.000 | 1.077 | 1.440 |
| Mercedes-Benz | 0.4703 | 0.012 | 40.353 | 0.000 | 0.447 | 0.493 |
| Mitsubishi | 0.2143 | 0.014 | 15.292 | 0.000 | 0.187 | 0.242 |
| Porsche | 0.6560 | 0.017 | 38.444 | 0.000 | 0.623 | 0.689 |
| Renault | 0.0328 | 0.012 | 2.779 | 0.005 | 0.010 | 0.056 |
| Rolls-Royce | 1.8328 | 0.044 | 41.434 | 0.000 | 1.746 | 1.920 |
| Saab | -0.0696 | 0.022 | -3.220 | 0.001 | -0.112 | -0.027 |
| Seat | 0.1273 | 0.013 | 10.060 | 0.000 | 0.102 | 0.152 |
| Suzuki | 0.2365 | 0.014 | 17.258 | 0.000 | 0.210 | 0.263 |
| Toyota | 0.2952 | 0.012 | 24.835 | 0.000 | 0.272 | 0.318 |
| Volvo | 0.2864 | 0.013 | 21.967 | 0.000 | 0.261 | 0.312 |
| Škoda | 0.1995 | 0.012 | 16.799 | 0.000 | 0.176 | 0.223 |
| Benzyna+LPG | 0.0190 | 0.007 | 2.656 | 0.008 | 0.005 | 0.033 |
| Diesel | 0.1830 | 0.003 | 57.607 | 0.000 | 0.177 | 0.189 |
| Hybryda | 0.1447 | 0.007 | 20.511 | 0.000 | 0.131 | 0.158 |
| aso_mileage | -2.425e-07 | 3.23e-08 | -7.498 | 0.000 | -3.06e-07 | -1.79e-07 |
| horse_power_2 | -6.895e-06 | 9.82e-08 | -70.187 | 0.000 | -7.09e-06 | -6.7e-06 |

| | | | |
|---|---|---|---|
| Omnibus: | 28216.430 | Durbin-Watson: | 1.808 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 1133256.790 |
| Skew: | 0.999 | Prob(JB): | 0.00 |
| Kurtosis: | 21.311 | Cond. No. | 9.63e+07 |

*Tabela 6. Wyniki trzeciej iteracji regresji*

Nowo dodane parametry dały ciekawe wyniki. Obydwa są statystycznie istotne i posiadają bardzo małe wartości współczynników. Jest to zrozumiałe, gdyż zarówno *mileage*, jak i kwadrat *horse_power* osiągają bardzo duże wartości. Niestety również tym razem wyniki testów nie były zadowalające. Zarówno w teście RESET, Jarque-Bera i Breusch–Pagana p-value było równe 0. Oznacza to, iż model nie spełnia założeń KRML oraz nie jest w stanie opisywać rzeczywistości.

## 6.2 Diagnostyka modelu

### 6.2.1 Test RESET Ramseya

Test Ramsey Regression Equation Specification Error Test sprawdza poprawność formy funkcyjnej danego modelu. Hipoteza zerowa zakłada liniowość modelu, natomiast hipoteza alternatywna jej brak. Dla wszystkich sprawdzonych modeli p-value wynosi 0.00. Oznacza to, że przy przyjętym poziomie istotności=0.05 musimy odrzucić hipotezę zerową o liniowości modelu.

### 6.2.2 Test Jarque-Bera

Ze względu na dużą ilość obserwacji nie jest konieczne użycie testu Jarque-Bera.

### 6.2.3 Test Breuscha-Pagana

Test Breuscha-Pagana pozwala zbadać homoskedastyczność modelu. Hipotezą zerowa jest wspomniana wcześniej homoskedastyczność, zaś hipotezą alternatywną heteroskedastyczność. Dla wszystkich sprawdzonych modeli p-value wynosi 0.00. Oznacza to, że przy przyjętym poziomie istotności=0.05 musimy odrzucić hipotezę zerową o liniowości modelu.

## 6.3 Weryfikacja hipotez

Podczas badania zostały podjęte dwie próby ulepszenia modelu w celu uzyskania poprawnej formy funkcyjnej. W drugiej iteracji badania został dodany logarytm ceny. W trzeciej iteracji została dodana interakcja oraz kwadrat zmiennej. Niestety żaden z tych zabiegów nie poprawił formy funkcyjnej tego modelu. Z tego względu nie jesteśmy w stanie wyciągać wniosków z modelu a co za tym weryfikować hipotez.

# 7. Zakończenie

Badanie to miało sprawdzić szereg postawionych na początku pracy hipotez. W związku z nie poprawnością formy funkcyjnej modelu nie opisuje on rzeczywistości w sposób dostateczny a co za tym idzie nie jesteśmy w stanie z niego wnioskować. Kontynuując badanie zostanie zwrócona większa uwaga na jakość zbioru danych uzyskany podczas scrapowania serwisu otomoto.pl. Co więcej ciekawym pomysłem wydaje się dodanie jeszcze większej ilości zmiennych, interakcji oraz potęg zmiennych.

# 8. Bibliografia

Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy *How much is my car worth? A methodology for predicting used cars prices using Random Forest*

Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric *Car Price Prediction using Machine Learning Techniques*

Sameerchand Pudaruth *Predicting the Price of Used Cars using Machine Learning Techniques*

Praful Rane, Deep Pandya, Dhawal Kotak *Used Car Price Prediction*