Wykorzystanie poznanych metod dotyczących analizy zależności liniowej do wybranych danych rzeczywistych

Autorzy

Mateusz Stasiak 262339
Karolina Wypych 262333

Politechnika Wrocławska

Wydział Matematyki

24 grudnia 2022r.

Spis treści

1 Wprowadzenie

W analizie danych przeważnie ma się do czynienia z dużymi zbiorami. Wybór odpowiednich metod do ich przetwarzania jest kluczowy. Zdarza się, że po przedstawieniu zmiennych na wykresie punktowym, można dostrzec, że w przybliżeniu układają się w linii prostej. Wówczas pod pewnymi warunkami, jest to dobry moment na zastosowanie metody regresji liniowej. Te warunki to:

- Zmienne powinny mieć charakter ciągły.
- Obserwacje powinny być od siebie niezależne (tzn. nie mogą występować żadne zależności).
- W danych nie powinny występować żadne istotne elementy odstające.
- Błędy (residua, wartości resztkowe) linii najlepszego dopasowania powinny mieć rozkład normalny.

Celem niniejszego raportu jest sprawdzanie tych warunków i próba przewidzenia cen domów w zależności od średnich przychodów w ich okolicy. Wszystkie analizy przeprowadzone zostały za pomocą języka Python i jego bibliotek.

2 Dane

Dane użyte w raporcie zostały pobrane ze strony nickmccullum.com. Zawierają informacje na temat m.in. przeciętnych przychodów i liczby pokoi w okolicy danego domu, ceny sprzedaży i adresu.

2.1 Wczytanie i przygotowanie danych

Dane zostały wczytane z pliku .csv za pomocą funkcji `pd.read_csv()`. Następnie wybrano dwie kolumny - opisujące przeciętny dochód oraz cenę domu, które są zmiennymi ciągłymi, a poszczególne obserwacje są od siebie niezależne (np. cena domu w jednej części kraju nie zależy od ceny w innej). Na ich podstawie zostaną przeprowadzone wszystkie dalsze analizy mające na celu przewidzenie cen mieszkań. W rozważanym modelu zmienną niezależną będzie przeciętny dochód a zależną cena domu.

nazwa zmiennej	status	opis zmiennej	rodzaj zmiennej	typ zmiennej
przeciętny dochód	zmienna zależna	przeciętny dochód osób zamieszkujących okolice danego domu	ciągła	float
cena domu	zmienna niezależna	kwota za jaką sprzedano dany dom	ciągła	float

Tabela 1: Charakterystyka zmiennych

3 Analiza jednowymiarowa zmiennych

Przed przystąpieniem do budowania modelu regresji liniowej należy przyjrzeć się każdej zmiennej z osobna. Pozwala to w łatwy sposób wychwycić obserwacje odstające i najczęściej przyjmowane wartości.

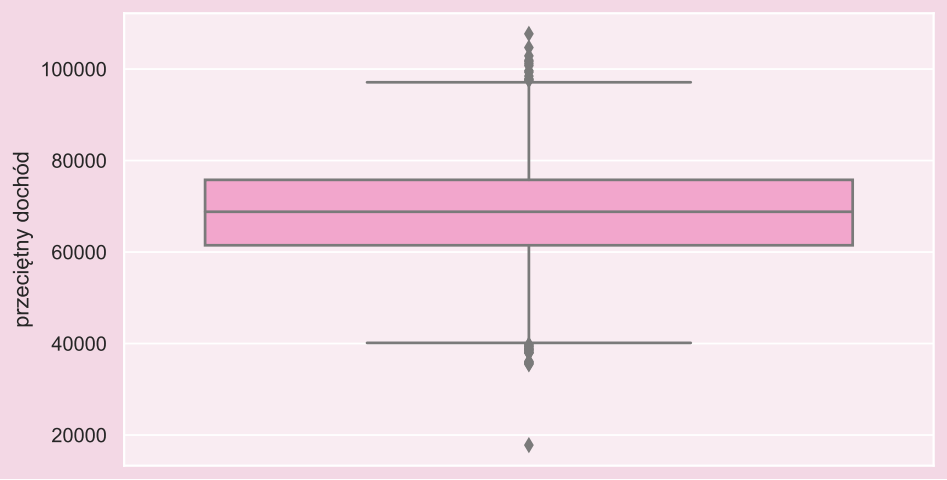
3.1 Zmienna niezależna

W pierwszej kolejności rozważaniom poddano ceny domów, będące zmienną niezależną. Zwizualizowano je na wykresach, a także obliczono podstawowe miary miary położenia, rozproszenia, skośności i spłaszczenia.

3.1.1 Wizualizacja danych

Wizualizacja danych jest niezwykle istotna przy ich analizowaniu. By lepiej poznać rozważaną zmienną, warto wykonać kilka wykresów. Ułatwią one zidentyfikowanie najczęściej pojawiających się wartości oraz rozkładu, z którego pochodzą dane.

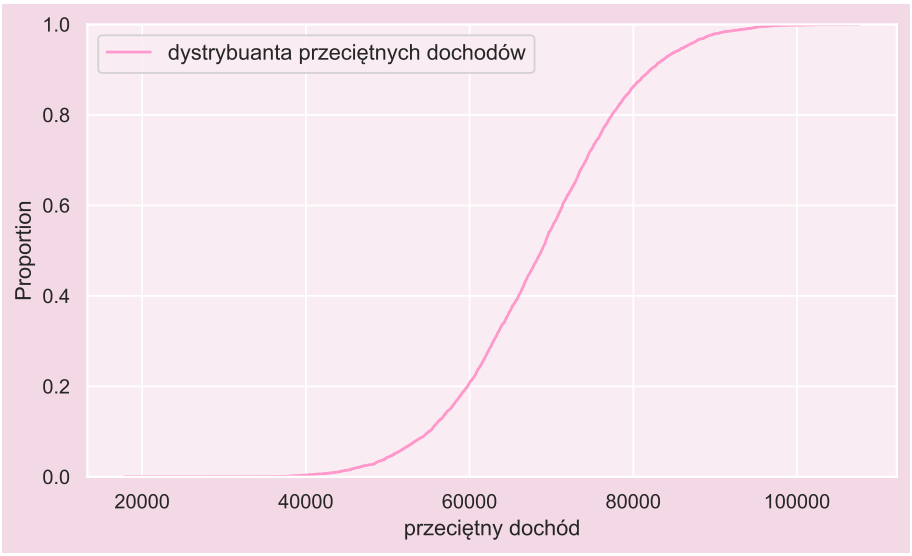
Jako pierwszy wygenerowano wykres pudełkowy i to od niego rozpoczęto analizę obserwacji.



Rysunek 1: Wykres pudełkowy przeciętnych przychodów

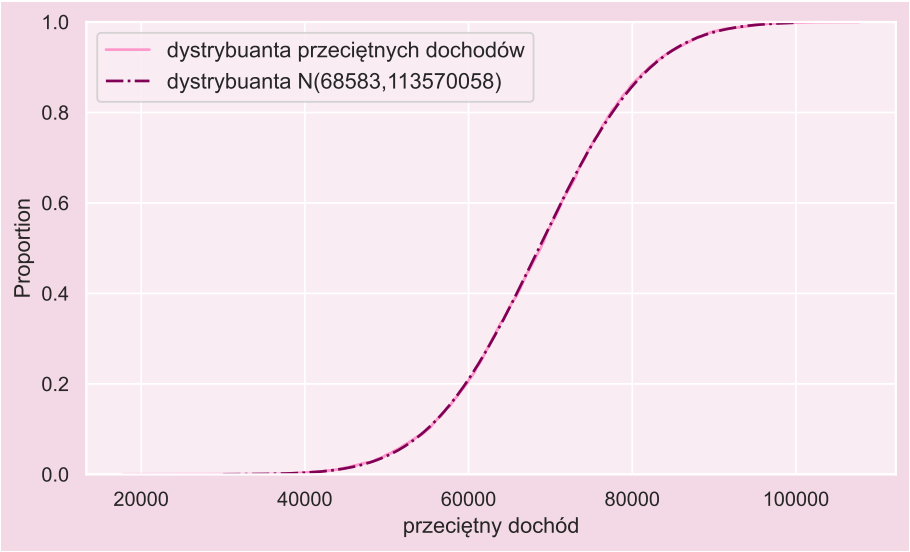
Powyższy wykres dostarcza kilku istotnych informacji. 50% wszystkich obserwacji znajduje się wewnątrz pudełka, a przecinająca je pionowa linia jest medianą rozważanego zbioru danych, która dzieli go na połowy. Stąd widać, że ponad połowa osób zarabia między 60000 a 80000. Wąsy natomiast łączą pudełko z największą i najmniejszą wartością badanej zmiennej odpowiednio z przedziału

$(Q1 - 1,5 \cdot IQR; Q1)$ oraz $(Q3; Q3 + 1,5 \cdot IQR)$. Zatem w pierwszym z nich mieści się 25% obserwacji o wartościach niższych od dolnego kwartyla a w drugim wyższych od górnego. Pozwala to wywnioskować, że zdecydowana większość społeczeństwa zarabia od 40000 do niemal 100000. Na wykresie nie brak także wartości odstających, które oznaczają zarobki znacznie odbiegające od reszty. Można także spróbować wyznaczyć rozklad z jakiego pochodzą omawiane obserwacje. W tym celu należy narysować jej dystrybuantę empiryczną i wykorzystać fakt, że dystrybuanta jednoznacznie określa rozkład.



Rysunek 2: Dystrybuanta empiryczna przeciętnych przychodów.

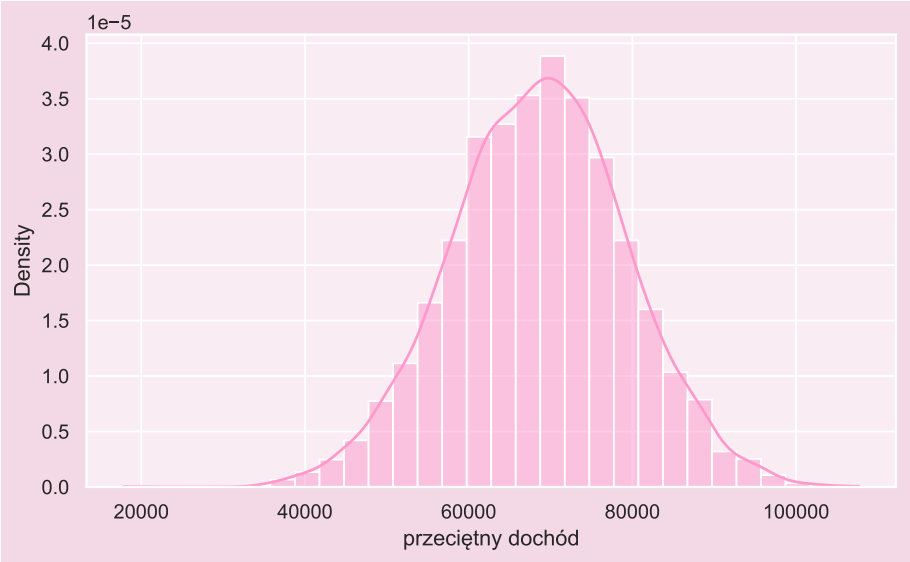
Wykres kształtem przypomina rozkład normalny, dlatego bardzo prawdopodobne wydaje się, że szukanym rozkładem będzie właśnie ten rozkład. By potwierdzić lub odrzucić tę tezę obliczono średnią i wariancję danych, które wyniosły kolejno $\mu \approx 68583$, $\sigma^2 \approx 113570058$, a następnie dorysowano dystrybuantę teoretyczną $\mathcal{N}(\mu, \sigma^2)$.



Rysunek 3: Zestawienie dystrybuanty empirycznej i teoretycznej

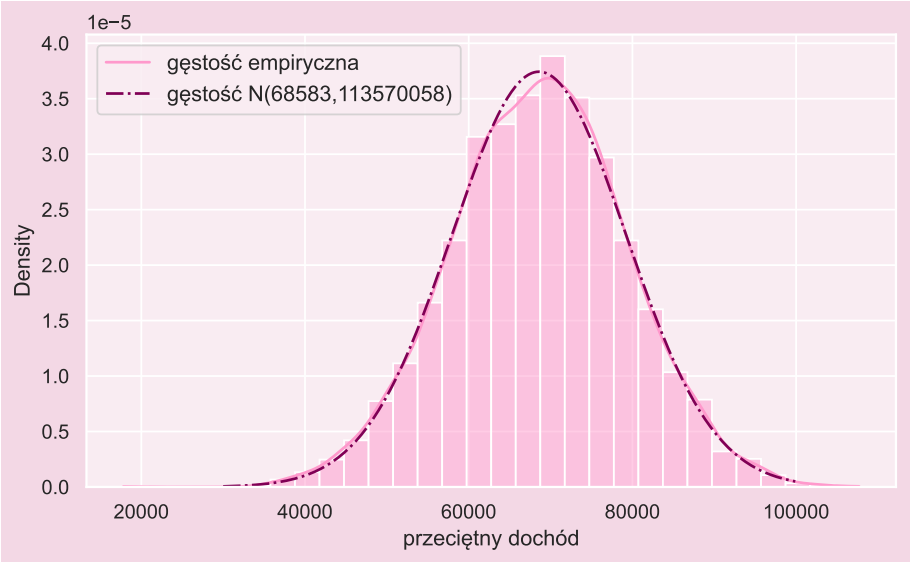
Z wykresu 3. widać, że obie dystrybuanty pokrywają się ze sobą na całej długości, co świadczy o tym, że rozkład i jego parametry zostały prawidłowo dobrane. Na pierwszy rzut oka tak duże wartości mogą się wydawać podejrzane, jednak patrząc na to, że rozpatrywane obserwacje przekraczają nawet 100000 trzeba przyznać, że są one prawidłowe.

Ostatnim etapem wizualizacji zmiennej niezależnej jest narysowanie histogramu prawdopodobieństwa i gęstości empirycznej.



Rysunek 4: Histogram przeciętnych dochodów wraz z gęstością empiryczną.

Wykres ten potwierdza informacje, które wyczytano z boxplotu na wykresie 1. Osoby zarabiające około 70000 stanowią najliczniejszą grupę. Większość skupiona jest pomiędzy 50000 i 90000, a najniższe słupki odnotowuje się poniżej 20000 i powyżej 100000. Histogram i dorysowana gęstość empiryczna swoim kształem jedynie potwierdzają wyznaczony wcześniej rozkład normalny. Dla pewności dorysowano jednak gęstość teoretyczną o wyliczonych wcześniej parametrach, która ponownie pokryła wykres empiryczny.



Rysunek 5: Histogram przeciętnych dochodów wraz z empiryczną i teoretyczną gęstością.

3.1.2 Podstawowe miary

Dodatkowych informacji o danych dostarcza także obliczenie podstawowych miar. Dla zwiększenia przejrzystości wyniki zaprezentowano w tabeli.

miary rozproszenia					
Q1	**Q3**	**IQR**	**wariancja**	**std**	**wsp. zmienności**
61480.562388	7.578334e+04	14302.776278	1.135701e+08	10656.925361	15.540257
miary położenia					
śr. arytm.		**śr.harm.**	**śr. geom.**	**mediana**	
6.858311e+04		6.681637e+04	6.772310e+04	6.880429e+04	
miary skośności i spłaszczenia					
wsp. skośności			kurtoza		
-0.033710			0.044329		

Tabela 2: Miary dla przeciętnych przychodów

Część z przedstawionych miar została już omówiona w paragrafie 3.1.1, ale to na co należy szczególnie zwrócic uwagę to wartość kurtozy. Ponieważ biblioteka *scipy* w Pythonie przyjmuje za kurtozę rozkładu normalnego 0, otrzymana $K = 0.044329$ stanowi kolejne potwierdzenie prawidłowego wyznaczenia rozkładu zmiennej.

3.1.3 Wnioski

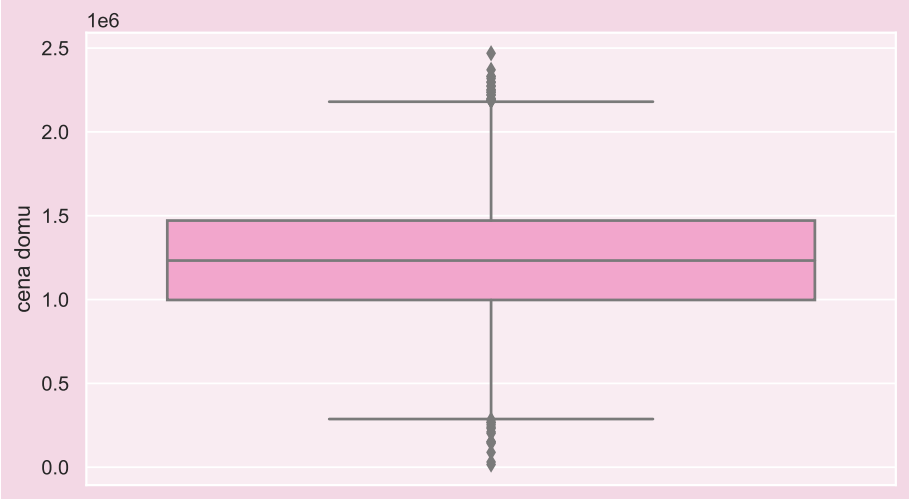
Podsumowując cały rozdział poświęcony zmiennej zależnej można stwierdzić, że zrealizowano pierwsze cele raportu. Wnikliwe przeanalizowanie zmiennej za pomocą wykresów i miar rozproszenia pozwoliło zobaczyć jak rozmieszczone są dane i zidentyfikowac ich rozkład jako $\mathcal{N}(\mu, \sigma^2)$. Ponadto pokazano, że nie zawierają żadnych istonie odstających wartości, co czyni je dobrą podstawą budowy modelu regresji liniowej.

3.2 Zmienna zależna

W drugiej kolejności przeanalizowano ceny domów, będące zmienną niezależną. Podobnie jak dla przeciętnych dochodów zwizualizowano je na wykresach i obliczono podstawowe miary.

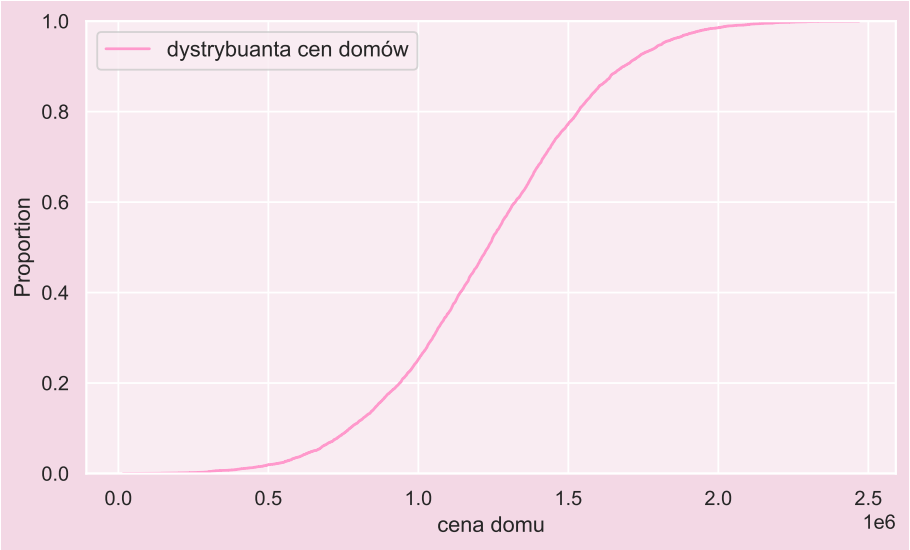
3.2.1 Wizualizacja danych

Poprzedni podrozdział udowodnił, jak potężnym narzędziem jest wizualizacja danych i ilu informacji można doszukać się poprzez wygenerowanie zaledwie kilku wykresów. Dlatego przy analizowaniu zmiennej zależnej nie sposób pominąć tego kroku. Najpierw wygenerowano boxplot, z którego odczytano, że najpopularniejsze ceny mieszkań mieszczą się w przedziale $1 - 1.5$ miliona. Tym razem ponownie nie brakuje wartości odbiegających od pozostałych. Z wykresu można odczytać, że są to wartości, które przekroczyły około 2.2 miliona lub nie dosięgnęły 300 tysięcy.



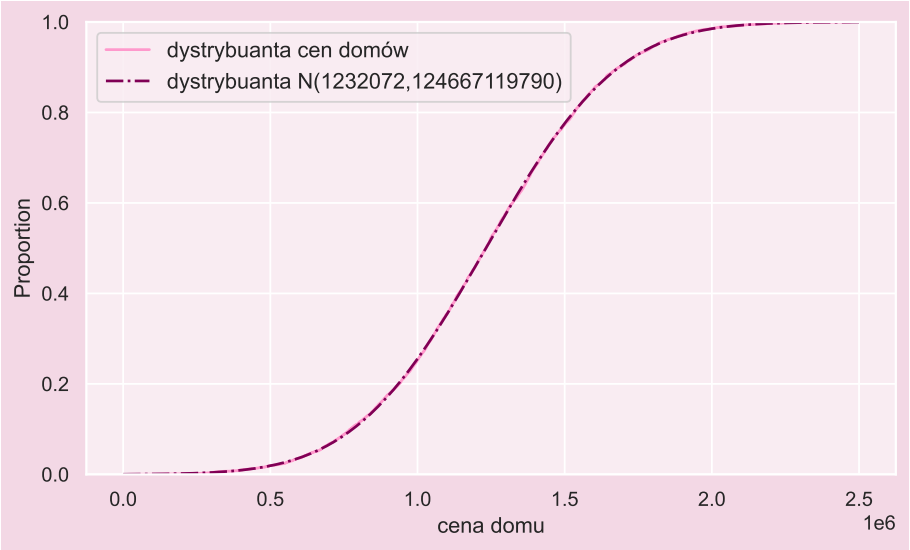
Rysunek 6: Wykres pudełkowy cen domów

Następnie narysowano dystrybuantę empiryczną, która tym razem ponownie bardzo przypomina kształtem rozkład normalny.



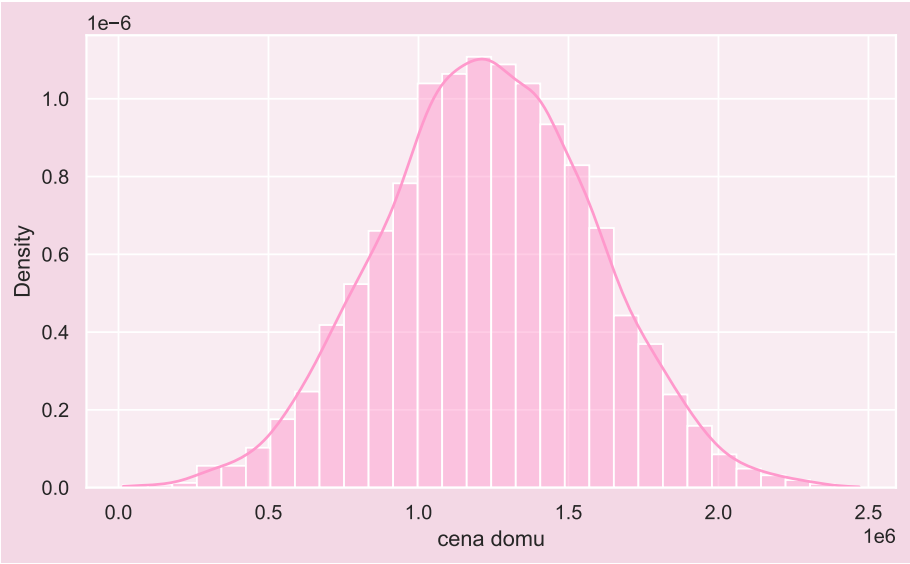
Rysunek 7: Dystrybuanta empiryczna cen domów

Z tego powodu tak samo jak przy zmiennej niezależnej wyznaczono parametry i dorysowano wykres dystrybuanty teoretycznej tym razem dla $\mu \approx 1232072$ i $\sigma^2 \approx 124667119790$. Kolejny raz otrzymano bardzo duże wartości, ale patrząc na to, że w tym raporcie pracuje się na cenach domów wartych miliony, takie rezultaty są jak najbardziej prawdopodobne.



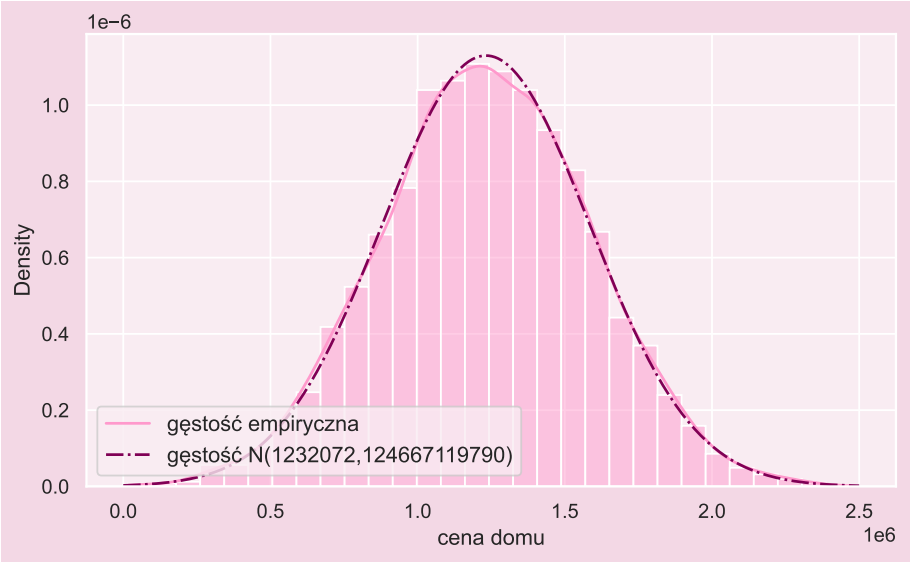
Rysunek 8: Dystrybuanta empiryczna cen mieszkań i dystrybuanta teoretyczna rozkładu normalnego z parametrami $\mu \approx 1232072$ i $\sigma^2 \approx 124667119790$

Na ostatnim etapie wizualnej analizy zmiennej zależnej skupiono się na histogramie i informacjach, które można z niego wyciągnąć.



Rysunek 9: Histogram cen domów wraz z gęstością empiryczną

Pokazuje on, że najbardziej popularną ceną jest 1.2 miliona, a niewiele mniej domów zostaje sprzedane za kwotę z przedziału $1 - 1.5$ miliona. Symetryczność wykresu wskazuje, że ceny sa podobnie zróżnicowane powyżej i poniżej średniej. Najniższe słupki w okolicach 300 tysięcy i powyżej 2.2 miliona odpowiadają outlayersom z wykresu pudełkowego. Taki kształt wykresu od razu przysuwa na myśl rozkład normalny, a dorysowanie gęstości teoretycznej jedynie potwierdza ten fakt.



Rysunek 10: Histogram cen domów wraz z gęstością empiryczną i teoretyczną

3.2.2 Podstawowe miary

Dla rozważanych danych obliczono także rózne miary i pogrupowano je w tabelach.

miary rozproszenia					
Q1	**Q3**	**IQR**	**wariancja**	**std**	**wsp. zmienności**
997577.135049	1.471210e+06	473633.069163	1.246671e+11	353082.313053	28.660455

miary położenia			
śr. arytm.	**śr.harm.**	**śr. geom.**	**mediana**
1.232073e+06	1.081459e+06	1.173387e+06	1.232669e+06

miary skośności i spłaszczenia	
wsp. skośności	**kurtoza**
-0.002717	0.056063

Tabela 3: Miary dla cen domów

Tym razem również wartość kurtozy -0.056063 stanowi kolejne potwierdzenie,

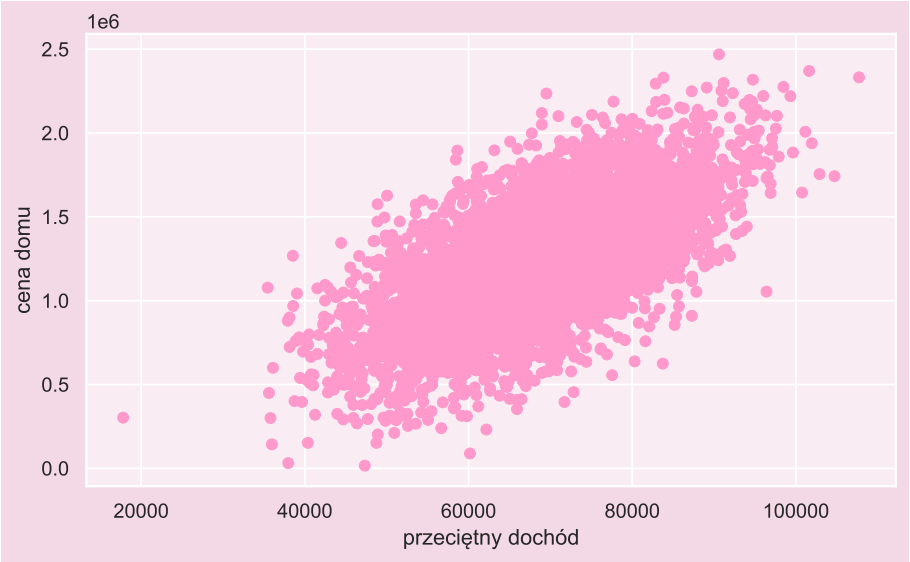
że prawidłowo zidentyfikowano rozkład omawianej zmiennej.

3.2.3 Wnioski

W powyższym podrozdziale rozłożono na czynniki pierwsze zmienną zależną, co zaowocowało znalezieniem jej rozkładu i poznaniem chrakterystyki występujących wartości. Dodatkowo nie znaleziono istotnie odstających wartości. Wszystko to sprawia, że przeanalizowane dane doskonale nadają się do zbudowania modelu regresji liniowej.

4 Analiza zależności liniowej pomiędzy zmienną objaśniającą a zmienną objaśnianą

4.1 Prezentacja danych



Rysunek 11: Wysokość ceny domu w zależności od przeciętnego dochodu w jego okolicy

Na wykresie rozproszenia danych (Rysunek 11) widać wyraźną zależność liniową. Ponieważ obserwacje układają się wzdłuż jednej prostej, ale nie są wokół niej ściśle skumulowane, możemy wyciągnąć wniosek, że powyższa korelacja nie jest bardzo silna. Potwierdza to współczynnik Pearsona, który wynosi 0.64.

4.2 Estymacja współczynników w klasycznym modelu regresji liniowej

Podczas analizy danych wykorzystano klasyczny model regresji liniowej. Ma on następujące założenia:

- $\mathbb{E}\varepsilon_i = 0 \quad \forall i = 1, 2, ..., n$
- $Var\varepsilon_i = \sigma^2 \quad \forall i = 1, 2, ..., n$
- $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ - niezależne zmienne losowe
- $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2) \quad \forall i = 1, 2, ..., n$

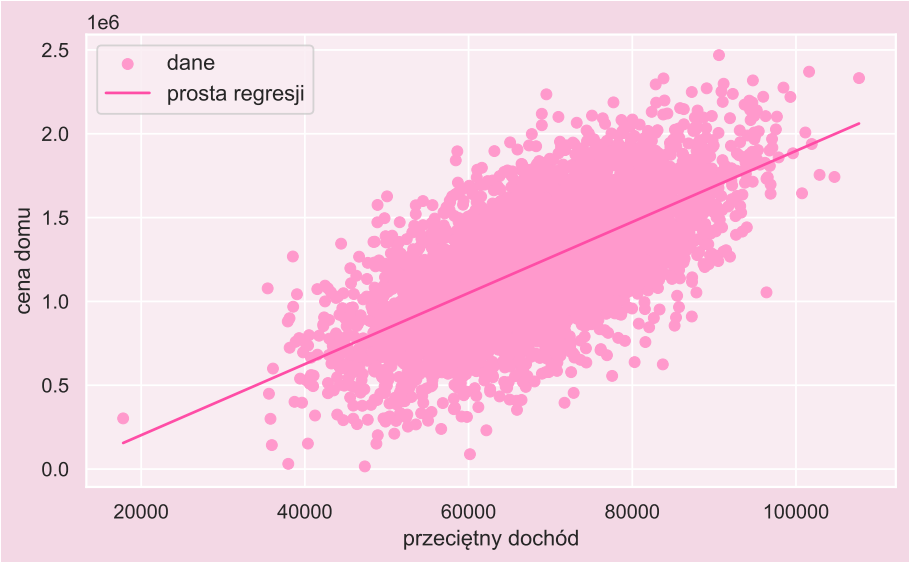
Proces estymacji współczynników można przeprowadzić na dwa sposoby. Estymacja punktowa korzysta z metody najmniejszych kwadratów lub metody największej wiarygodności. Znając teorię, można wyprowadzić wzory na estymatory współczynników. Natomiast estymacja przedziałowa zakłada skonstruowanie przedziałów ufności dla B_0 i B_1. Korzystając z tego sposobu, można stwierdzić, że na wybranym pozomie istotności α, współczynniki B_0, B_1 mieszczą się w uzyskanym przedziale z prawdopodobieństwem $1 - \alpha$.

4.2.1 Estymacja punktowa

Współczynniki prostej regresji wyznaczono z wykorzystaniem metody najmniejszych kwadratów. Estymatory wynoszą

$$\hat{B_1} = \frac{\sum_i^n x_i(y_i - \overline{y})^2}{\sum_i^n (x_i - \overline{x})^2} = 21.195$$

$$\hat{B_0} = \overline{y} - \hat{B_1}\overline{x} = -221579.478.$$



Rysunek 12: Dopasowanie prostej regresji o współczynnikach $\hat{B_0} = 21.195$, $\hat{B_1} = -221579.478$

4.2.2 Estymacja przedziałowa

Do obliczeń wykorzystano poziom istotności $\alpha = 0.05$, zatem współczynniki B_0, B_1 znajdują się w poniższych przedziałach z prawdopodobieństwem 95%. Dla B_0 wykorzystano wzór

$$\left[\hat{B_0} - t_{1-\frac{\alpha}{2},n-2} \cdot S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i^n (x_i - \overline{x})^2}}, \quad \hat{B_0} + t_{1-\frac{\alpha}{2},n-2} \cdot S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_i^n (x_i - \overline{x})^2}}\right] =$$

$$[-221579.6587, -221579.2976]\,.$$

Podobnie dla współczynnika B_1:

$$\left[\hat{B_1} - t_{1-\frac{\alpha}{2},n-2} \cdot \frac{S}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}, \quad \hat{B_1} + t_{1-\frac{\alpha}{2},n-2} \cdot \frac{S}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}\right] =$$

$$[20.4893, 21.9016]\,.$$

4.3 Jakość dopasowania w modelu regresji liniowej

Jakość dopasowania prostej regresji można sprawdzić za pomocą współczynnika determinacji $\phi = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$, gdzie:

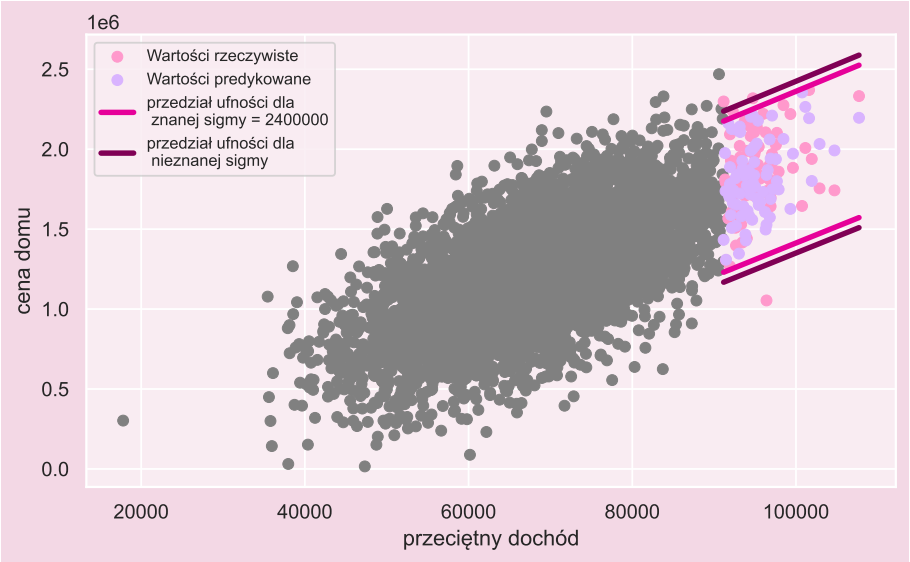
- SSE = $\sum_i^n (y_i - \hat{y}_i)^2$
- SSR = $\sum_i^n (\hat{y}_i - \overline{y})^2$
- SST = SSE + SSR.

Dla analizowanych danych współczynnik ten wynosi 0.409. Oznacza to, że 40.9% zmiennych objaśnianych udało się opisać za pomocą zmiennych objaśniających. Rozpisujac wzory na powyższe wskaźniki, otrzymamy kwadrat współczynnika korelacji Pearsona. Istotnie, korzystając z wcześniejszych obliczeń, $(0.64)^2 = 0.409$.

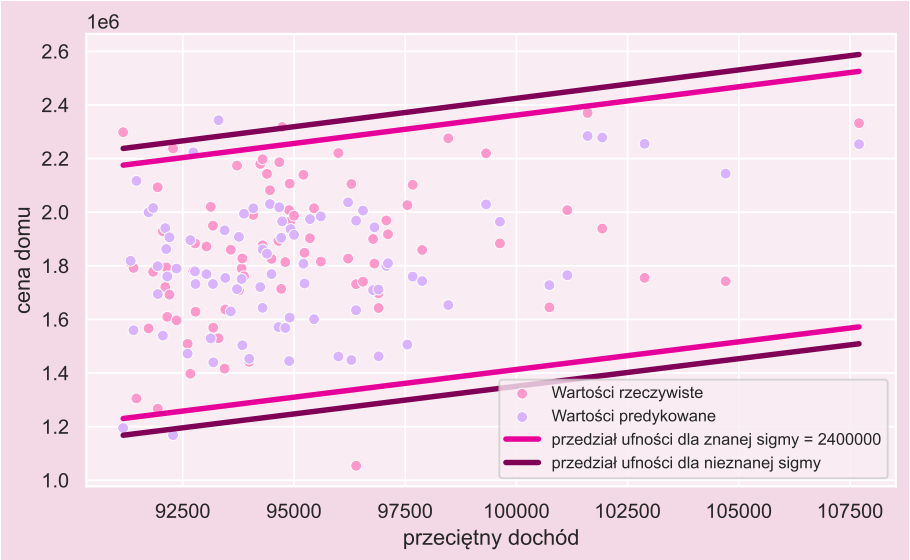
4.4 Predykcja oraz przedziały ufności dla danych testowych

Predykcja oraz ocena jej efektywności przebiega według poniższego algorytmu.

1. Podział danych na zbiór treningowy i zbiór testowy.
2. Wyznaczenie współczynników $\hat{B_0}$ i $\hat{B_1}$ na podstawie zbioru treningowego.
3. Wyznaczenie prostej regresji na zbiorze treningowym. Wyliczenie błędów i znalezienie ich rozkładu.
4. Wyznaczenie predykowanych wartości na zbiorze testowym, uwzględniając błędy jako zmienne z poznanego rozkładu.
5. Wyznaczenie przedziałów ufności dla zbioru testowego na poziomie istotności α.
6. Zliczenie jaka część wartości predykowanych oraz jaka część danych rzeczywistych ze zbioru testowego znalazła się w wyliczonym przedziale ufności. Uzyskane ułamki powinny w przybliżeniu równać się $1 - \alpha$.



Rysunek 13: Predykcja oraz przedziały ufności dla danych testowych



Rysunek 14: Predykcja oraz przedziały ufności dla danych testowych

Zgodnie z algorytmem, analizowane dane zostały podzielone na dwie grupy. Zbiór testowy składa się z 80 obserwacji o największym przeciętnym docho-

dzie. Współczynniki regresji liniowej wyliczone dla zbioru treningowego wynoszą $\hat{B_0} = -202938.6469$ i $\hat{B_1} = 20.905821$, a błędy pochodzą z rozkładu $N(\mu = 0, \sigma = 271377)$. Przedziały ufności zostały skonstruowane ze wzoru

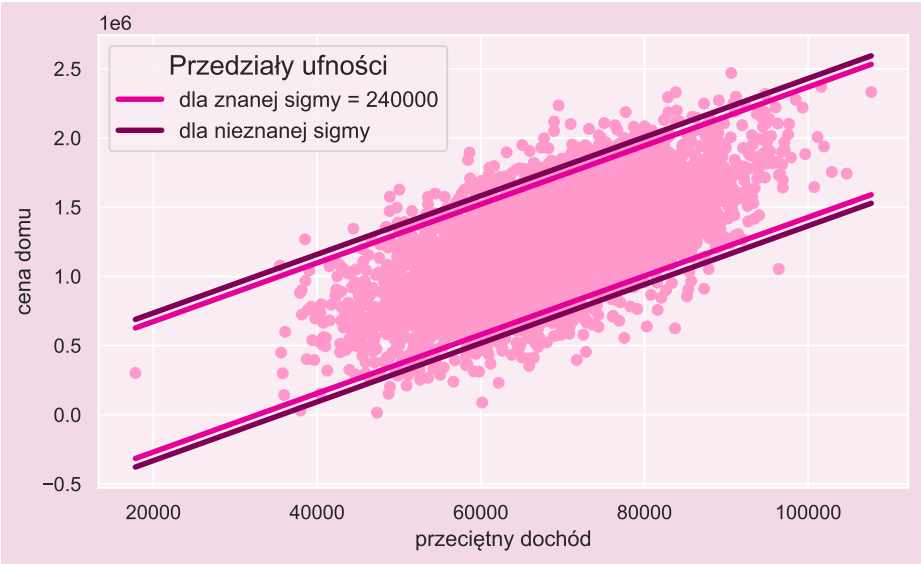
$$\left[\hat{y}\left(x_0\right) - t_{1-\frac{\alpha}{2},n-2} \cdot S\sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum_i \left(x_i - \overline{x}\right)^2}}, \quad \hat{y}\left(x_0\right) + t_{1-\frac{\alpha}{2},n-2} \cdot S\sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}\right)^2}{\sum_i \left(x_i - \overline{x}\right)^2}}\right]$$

W obliczeniach wykorzystano poziom istotności $\alpha = 0.05$ oraz nieobciążony estymator odchylenia standardowego $S = \sqrt{\frac{1}{n-2}\sum_i \left(Y_i - \hat{Y_i}\right)}$. Dla porównania, na wykresach 13 i 14 przedstawiono również przedział ufności wyliczony sposobem dla znanej sigmy, równej 240000. Fakt, że owy przedział jest węższy od nominalnego sugeruje, że prawdziwa wartość σ jest większa niż przyjęto w tej metodzie.

4.5 Ocena jakości dopasowania do danych testowych

Dla poziomu istotności $\alpha = 0.05$ procenty danych predykowanych i danych rzeczywistych zawartych w przedziale ufności wynosiły odpowiednio 96.25% i 95%. Uzyskane liczby w przybliżeniu równają się $1 - \alpha$. Oznacza to, że sposób predykowania danych został zaimplementowany poprawnie.

Otrzymane przedziały ufności można zastosować także dla wszystkich cen domów.



Rysunek 15: Przedziały ufności dla znanej i nieznanej wariancji

Dla empirycznie wyliczonego przedziału ufności (nieznana sigma) 94.82% obserwacji znajduje się między prostymi.

4.6 Wnioski

Wykonanie powyższych analiz doprowadziło do wyznaczenia estymatorów współczynników B_0 i B_1 oraz ich przedziałów ufności na poziomie istotności $\alpha = 0.05$. Oceniono także jakość dopasowania w modelu regresji liniowej za pomocą współczynnika determinacji, który wyniósł 0.49. Otrzymane rezultaty zaprezentowano na wykresie. Na koniec poprawnie zaimplementowano sposób predykcji danych i sprawdzono jego efektywność.

5 Analiza residuów

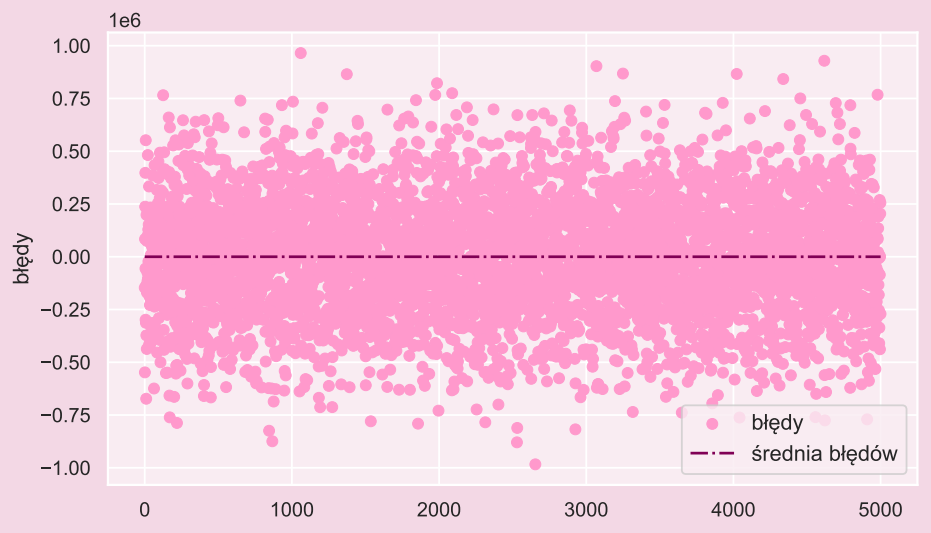
5.1 Sprawdzenie założeń

Chcąc stwierdzic, czy poprawnie wyznaczono współczynniki $\hat{B_0}$ i $\hat{B_1}$ konieczne jest wykonanie analizy residuów, czyli wartości resztkowych w modelu regresji liniowej. Muszą one spełniać następujące warunki:

- $\mathbb{E}\varepsilon_i = 0 \quad \forall i = 1, 2, ..., n$
- $Var\varepsilon_i = \sigma^2 \quad \forall i = 1, 2, ..., n$
- $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ - niezależne zmienne losowe

- $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2) \quad \forall i = 1, 2, ..., n$

By móc powiedzieć coś więcej na temat wartości oczekiwanej i wariancji residuów, sporządzono wykres punktowy przedstawiający wszystkie residua.



Rysunek 16: Wykres punktowy wartości resztkowych

Juz na pierwszy rzut oka widać, że błędy rozkładają się równomiernie po obu stronach osi poziomej. Dodatkowo potwierdza to ich średnia, którą również naniesiono na wykres. Jej wartość policzona przy pomocy funkcji wbudowanej w Pythonie to $1.19791e-9$. Już na tym etapie można także stwierdzić, że residua mają stałą wariancję (według wbudowanej funkcji jest to 73645940735.18942, ponieważ tworzą w przybliżeniu pas jednej grubości.

Następnie sprawdzono, czy rozważane dane są niezależnymi zmiennymi losowymi. W tym celu posłużono się dwiema funkcjami:

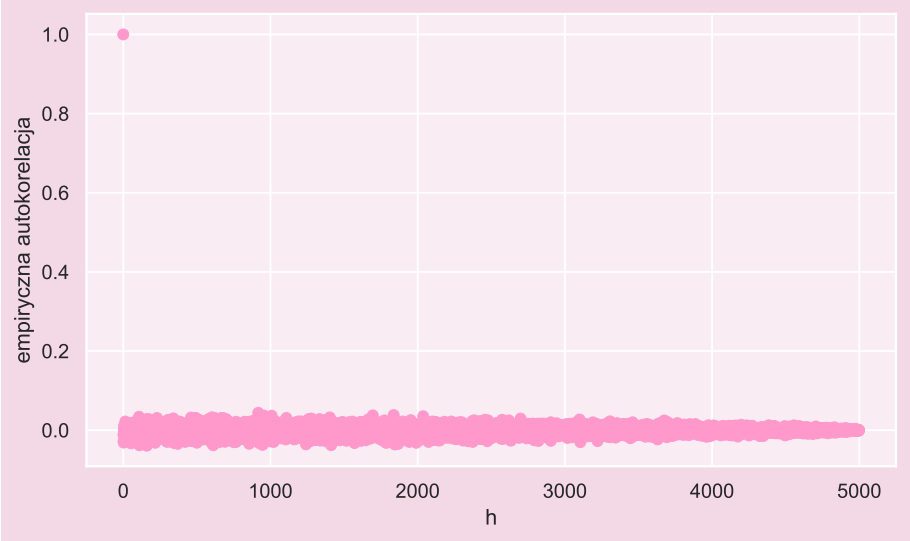
- empirycznej autokowariancji

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (e_{i+|h|} - \bar{e})(e_i - \bar{e}) \quad -n < h < n, \quad h \in Z,$$

- empirycznej autokorelacji

$$\hat{g}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

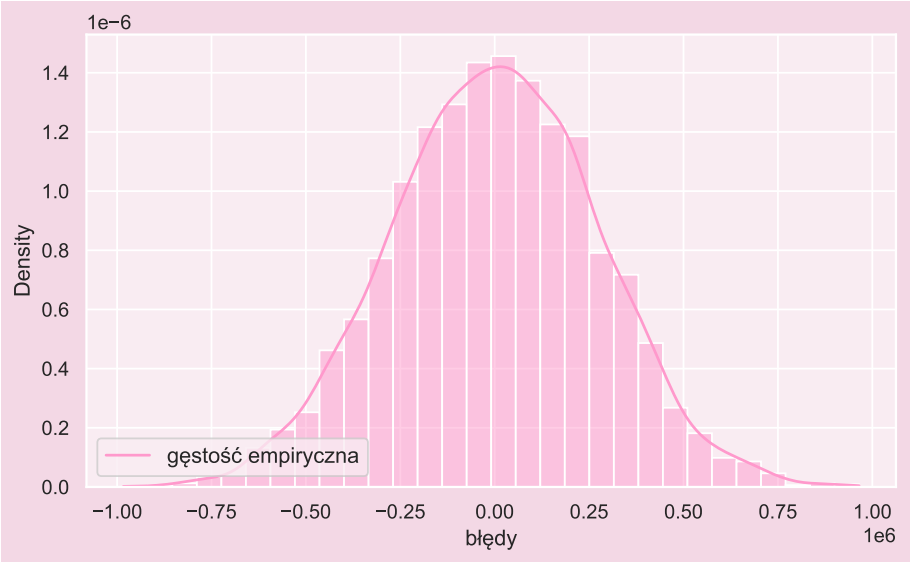
Najpierw policzono funkcję empirycznej autokowariancji dla wszystkich możliwych wartości h, a następnie wyznaczono empiryczną autokorelację i to ją przedstawiono na wykresie 17.



Rysunek 17: Wykres funkcji empirycznej autokorelacji w zależności od h

Funkcja dla $h = 1$ wynosi 1 a poza tym wszędzie jest bliska zeru. Taka postac funkcji autokorelacji empirycznej dowodzi, że badane residua są zmiennymi niezależnymi.

Ostatnim założeniem jakie należy udowodnić jest pochodzenie wartości resztkowych z rozkładu normalnego o zerowej średniej i stałej wariancji. Jednym ze sposobów, by to osiągnąć, jest narysowanie histogramu i próba dopasowania gęstości rozkładu normalnego o odpowiednich parametrach.



Rysunek 18: Histogram wartości resztkowych

Ponieważ histogram kształtem bardzo przypomina rozkład normalny o średniej zero, najrozsądniejszym jest dorysowanie gęstości rozkładu normalnego o parametrach równych wariancji i średniej wyznaczonych na początku tego podrozdziału.

Rysunek 19: Histogram wartości resztkowych wraz z gęstością empiryczną i teoretyczną z rozkładu $\mathcal{N}(0, 73645940735)$

Na wykresie 19. obie gęstości pokrywają się, a ponieważ gęstość jednoznacznie identyfikuje rozkład, jest to dowód na to, że residua pochodzą z rozkładu normalnego o średniej zero i stałej wariancji. Dodatkowo wykonano wykres kwantylowy i jako rozkład teoretyczny przyjęto $\mathcal{N}(0, 73645940735)$. Nachylenie linii kwantylowej do osi poziomej pod kątem 45° ostatecznie potwierdza wysnuty wniosek.

5.2 Wnioski

W podrozdziale tym wzięto pod lupę residua i dokładnie przeanalizowano je pod kątem spełniania założeń niezbędnych w regresji liniowej. Ponieważ w odniesieniu do każdego z nich otrzymano satysfakcjonujące rezultaty, można stwierdzić, że poprawnie zbudowano model regresji liniowej.

6 Podsumowanie

Podsumowując raport można stwierdzić, że zrealizowano wszystkie postawione na początku cele. Skrupulatnie przeanalizowano dane i sprawdzono poprawność założeń. Zdołano zidentyfikować rozkłady z jakich pochodziły zmienne i poprawnie wyznaczyć współczynniki w modelu regresji linowej. Praca z danymi oparta na statystyce i programowaniu zaowocowała w tym przypadku dobraniem właściwego modelu do danych i wykonaniem predykcji dla cen domów w zależności od średnich zarobków w jego okolicy.