[anim@pkg]pdftex.code:n=,pdftex.value_forbidden:n=true,luatex.code:n=,luatex.value_forbidden:n=tru
[anim@pkg]
[pdfbase]
[ocgbase]
[anim@user]label.code:n=,label.value_required:n=true,width.code:n=,width.value_required:n=true,heig
[anim@pkg]width.code:n=,width.value_required:n=true,height.code:n=,height.value_required:n=true,to
,buttonfg.value_required:n=true,buttonalpha.tl_gset_x:N=,buttonalpha.value_required:n=true,alttext
[anim@pkg]

Wykorzystanie poznanych metod dotyczących analizy zależności liniowej do wybranych danych rzeczywistych

Autorzy

Kacper Budnik 262286
Maciej Karczewski 262282

Politechnika Wrocławska

Wydział Matematyki

22 grudnia 2022

Spis treści

1 Wprowadzenie

Diamenty są krystalicznym węglem, najtwardszym znanym materiałem na Ziemi. Są one tworzone w głębi Ziemi, w wysokich temperaturach i ciśnieniach, a ich wydobycie wymaga specjalistycznej wiedzy i wielu lat doświadczenia. Diamenty są najcenniejszymi kamieniami szlachetnymi na świecie i są cenione zarówno przez osoby prywatne, jak i przez przemysł jubilerski. Blask i niezwykła trwałość diamentów sprawiają, że są one bardzo cenione, a ich cena zależy od wielu czynników, takich jak kolor, czystość, kształt i wielkość. Najcenniejsze diamenty są przezroczyste i bezbarwne, choć mogą też być innego koloru np. żółte, niebieskie, czerwone czy zielone.

Diamenty są często wykorzystywane do produkcji biżuterii. Są również używane w narzędziach do cięcia i szlifowania oraz w elektronice. W ostatnich latach diamenty zaczęły być również używane jako narzędzia do badań naukowych, ponieważ ich niezwykła trwałość i twardość sprawiają, że są one idealnymi materiałami do wielu zastosowań.

Dane o diamentach pozyskujemy z platformy Kaggle. Nasze dane zawierają 53940 rekordów i pierwsze 7 wierszy wygląda następująco:

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47

Tabela 1: Oryginalne dane

Dla naszych danych sprawdzimy liniową zależność ceny ("price") od iloczynu wymiarów diamentu.

2 Transformacja danych

2.1 Dodanie nowej kolumny

Do naszych danych dodamy nową kolumną i nazwiemy ją "V". Nowa kolumna będzie iloczynem wymiarów diamentu (kolumny "x", "y" oraz "z"). Ten iloczyn możemy interpretować jako objętość diamentu z dokładnością do przemnożonej stałej. Nową zmienną będziemy wyjaśniać cenę diamentu. W dalszej części raportu tą zmienną będziemy nazywać objętością lub iloczynem wymiarów.

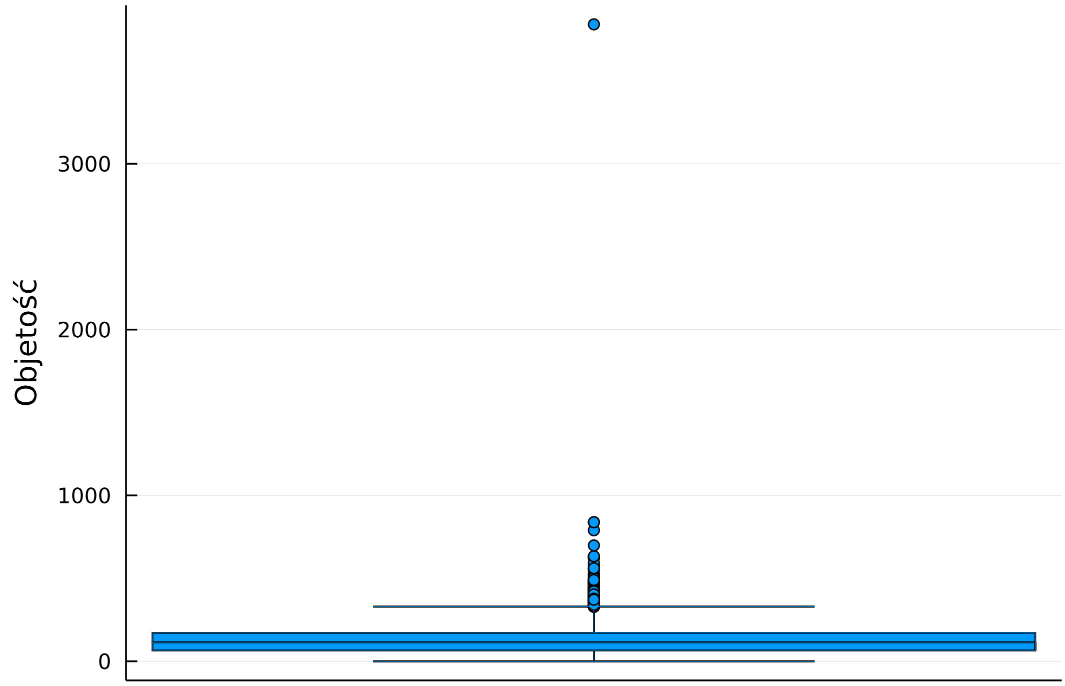
2.2 Usunięcie wartości odstających

W następnym kroku usuniemy 1% najmniejszych cen oraz 10% największych cen. Będziemy usuwać też 1% najmniejszych objętości oraz 5% największych. Ostatecznie usuniemy około 12% danych.

3 Jednowymiarowa analiza

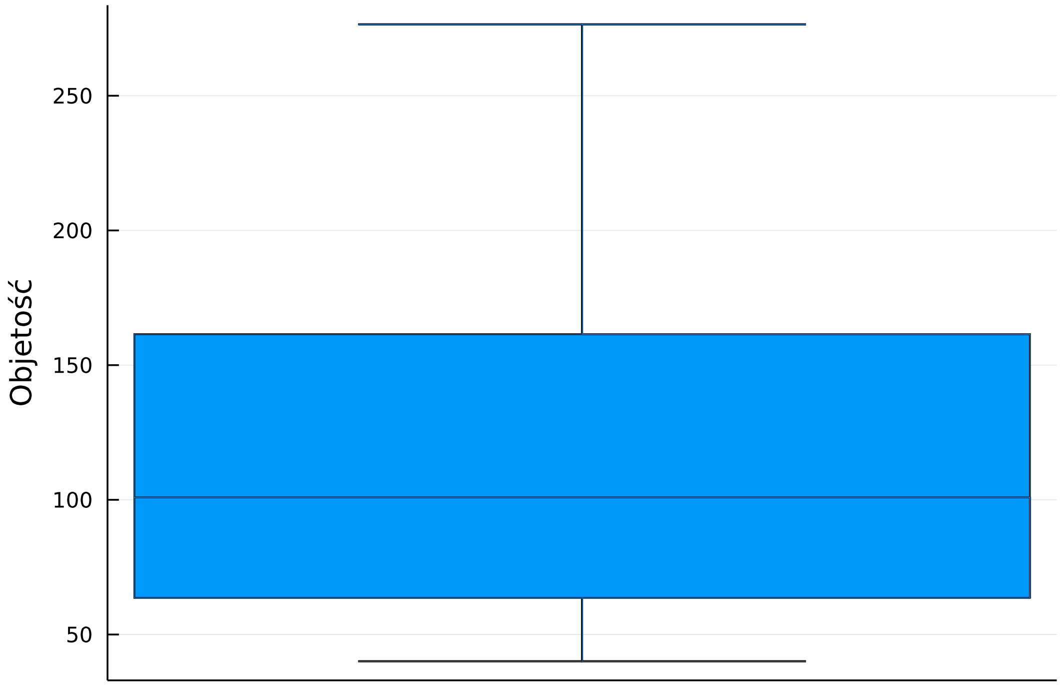
3.1 Objętość

Analizę danych zaczniemy od analizy objętości. Box plot objętości wygląda następująco.



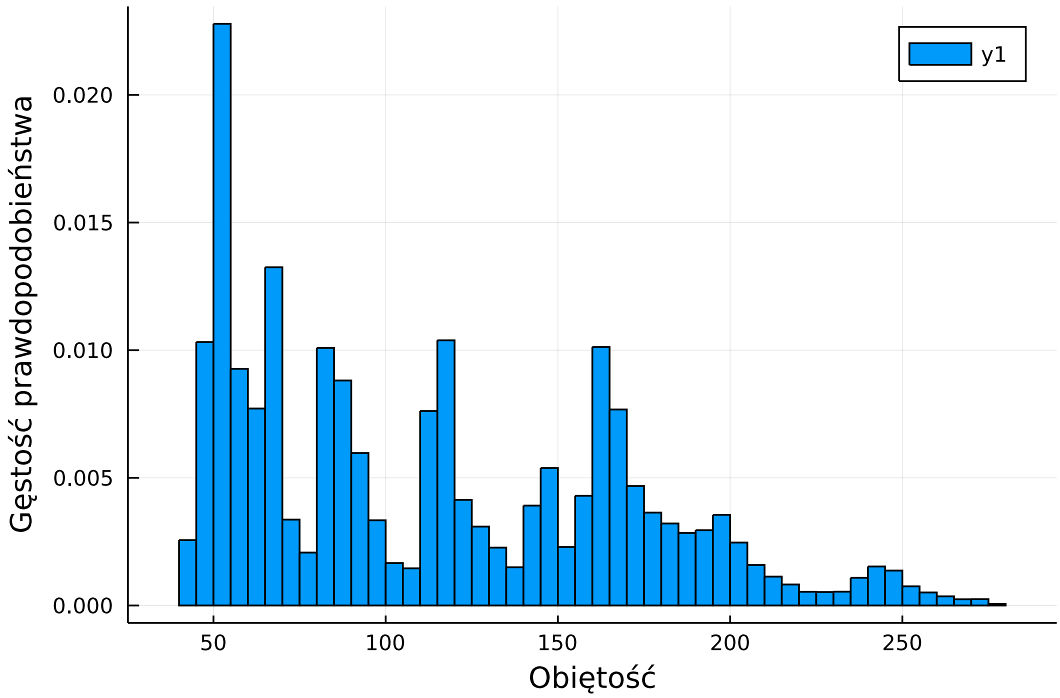
Rysunek 1: Boxplot objętości dla podstawowych danych

Jak możemy zobaczyć na box plocie nasze dane mają dużo wartość odstających. Z tego powodu ciężko analizować wykres 1, a także stworzyć poprawny model, dlatego pozbędziemy się ich. Gdy usuniemy 1% najmniejszych objętości oraz 5% największych otrzymamy czytelniejszy wykres.



Rysunek 2: Box plot objętości dla oczyszczonych danych

Tutaj nasze dane są o wiele bardziej czytelne. Z wykresu 2 możemy odczytać, że mediana naszych danych wynosi około 100 oraz, że mamy do czynienia z rozkładem prawostronnie skośnym. Widzimy też, że pierwszy kwantyl wynosi około 60 a trzeci około 160. Zobaczymy teraz jak wygląda rozkład objętości na histogramie.



Rysunek 3: Histogram objętości dla oczyszczonych danych

Jak widzimy na histogramie gęstość na histogramie nie zachowuję się monotoniczne, ciężko stwierdzić jaki rozkład ma objętość. Pomimo tego faktu możemy zauważyć, że diamenty dzielą się głównie na

4 grupy. Pierwszą najbardziej liczą jest grupa o iloczynie wymiarów należących do przedziału $[45; 70]$, następną grupą jest o iloczynie należącym do $[80; 100]$. Pozostałymi grupami są iloczyny należące do przedziałów $[110; 120]$ oraz $[160; 170]$.

Zobaczymy teraz jak wyglądają statystyki opisowe dla naszych oczyszczonych danych.

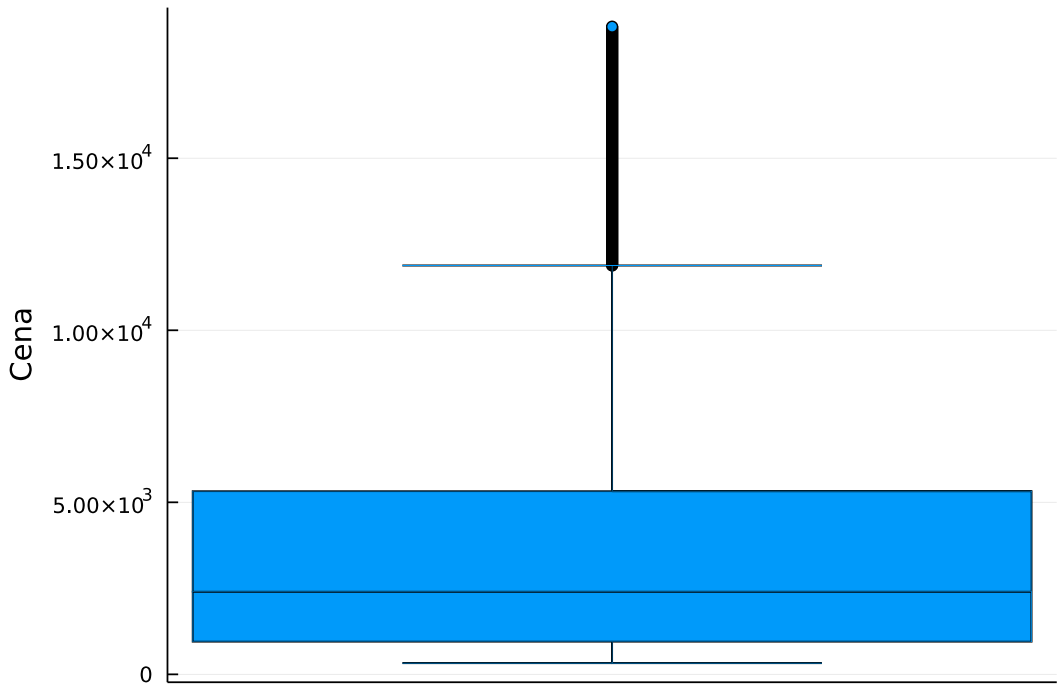
	Średnia	Mediana	Wariancja	Skośność	Kurtoza
V	113.48	100.87	3057.10	0.60	-0.63

Tabela 2: Podstawowe statystyki opisowe dla objętości

Możemy zobaczyć, że średnia jest większa od mediany, oraz że skośność jest dodatnia co sugeruje nam, że rozkład objętości jest prawostronnie skośny co się pokrywa za informacją zawartą na box plocie 2. Kurtoza jest ujemna to znaczy, że rozkład ma ogony węższe niż rozkład normalny czyli jest rozkładem platykurtycznym.

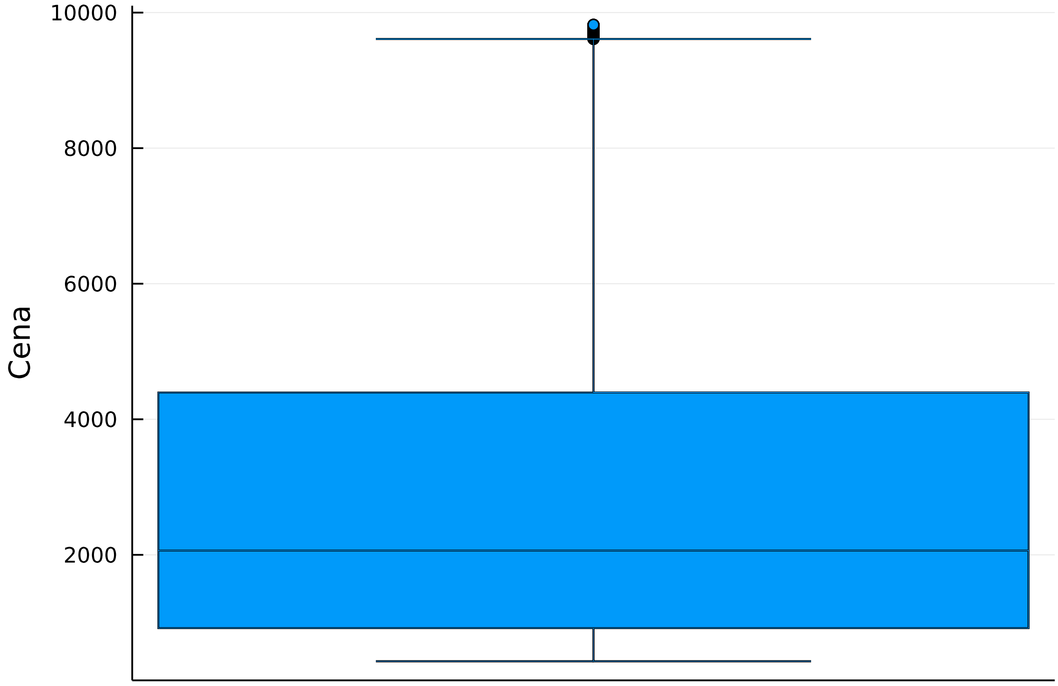
3.2 Cena

Sprawdziliśmy jak wygląda rozkład objętości to teraz sprawdzimy rozkład ceny diamentów. Box Plot dla ceny diamentów bez usunięcia wartości skrajnych wygląda następująco.



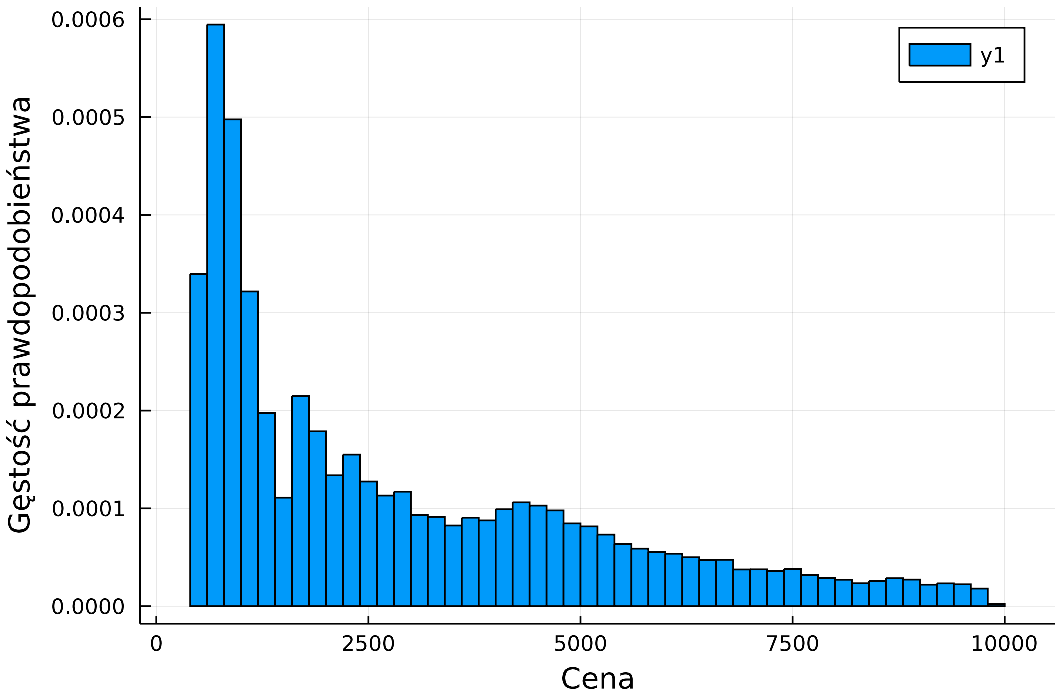
Rysunek 4: Boxplot ceny dla podstawowych danych

Z box plotu możemy odczytać, że ponownie mamy dużo wartości odstających, które przeszkadzają w analizie danych. Właśnie z tego powodu usuniemy 1% najmniejszych cen oraz 10% największych cen. Po usunięciu wartości nasz Box plot wygląda następująco



Rysunek 5: Box plot ceny dla oczyszczonych danych

Po usunięciu danych skrajnych wykres pudełkowy staje się czytelniejszy. Z wykresu możemy odczytać, że mediana cen jest w okolicy 2000 natomiast pierwszy w okolicach 1000 a trzeci w okolicach 4300. Możemy sądzić, że rozkład ceny jest prawostronnie skośny.



Rysunek 6: Histogram ceny diamentów

Z histogramu możemy odczytać, że gęstość prawdopodobieństwa ceny jest prawie monotoniczna. Im większa cena tym na ogół, rzadszy diament. Możemy zobaczyć, że mamy najwięcej diamentów tanich nieprzekraczających 2500 . Możemy także zobaczyć większą liczebność diamentów z ceną w okolicach 4000.

Zobaczymy teraz jak wyglądają statystyki opisowe dla naszych oczyszczonych danych

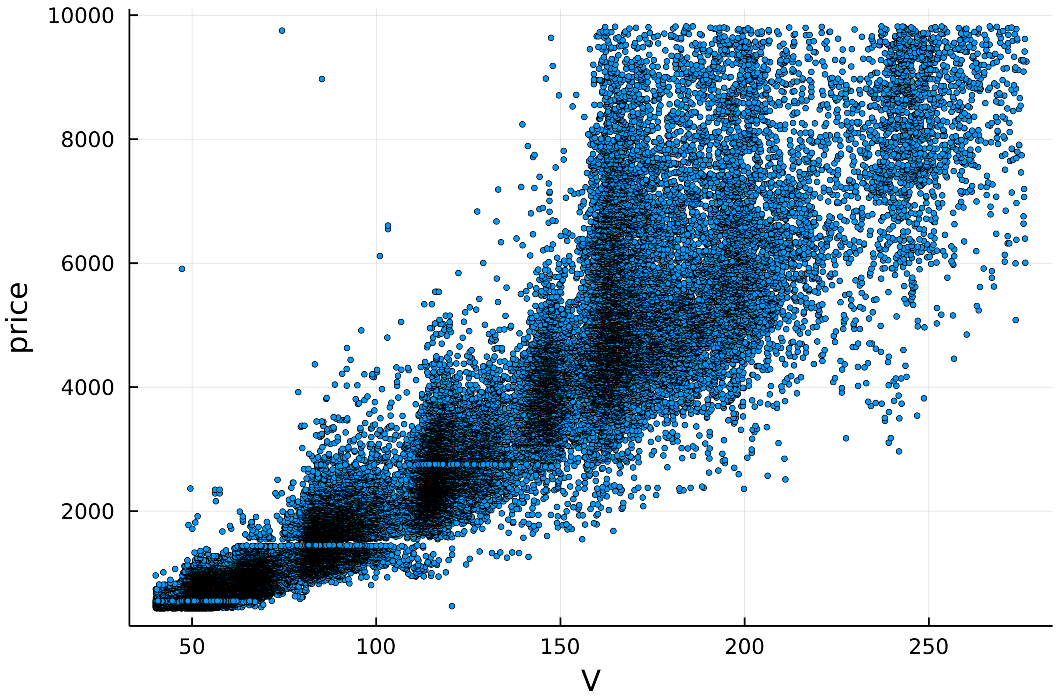
	Średnia	Mediana	Wariancja	Skośność	Kurtoza
Cena	2890.46	2064.0	$5.49 \cdot 10^6$	1.02	0.10

Tabela 3: Podstawowe statystyki opisowe dla objętości

Możemy zobaczyć, że średnia jest większa od mediany, oraz że skośność jest dodatnia co sugeruje nam, że rozkład objętości jest prawostronnie skośny co się pokrywa za informacją zawartą na box plocie 5. Kurtoza jest dodatnia ale bliska 0 to znaczy, że rozkład ma ogony grubsze niż rozkład normalny lub ma porównywalne do rozkładu normalnego. Zatem jest to rozkład leptokurtyczny lub mezokurtyczny. Wariancja ceny jest znacząco większa niż wariancja obiętości 2.

4 Analiza zależności między

Analizujemy zależność między ceną, a objętością. Przyjrzyjmy się zalezności naszych danych na wykresie.



Rysunek 7: Wykres zależności ceny od objętość

Z wykresu możemy zauważyć silną zależność między naszymi danymi. Z powodu rozłożenia naszych danych zależność ta może być linowa. W celu określenia tej zależności obliczymy współczynnik korelacji pearsona

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \approx 0.93,$$

gdzie x_i, y_i to i-te obserwacje odpowiednio objętości oraz ceny, $\overline{x}$ oznacza średnią z x, a n to rozmiar próby, oznaczeniami tymi będziemy posługiwali się w całym raporcie. Wartość ta jest blisko wartości 1, zatem nasze dane są silnie skorelowane dodatnie oraz liniowo. Dlatego model, którym będziemy opisywać dane będzie miał postać

$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, \tag{1}$$

gdzie ξ_i jest losowym błędem pomiarowym. Y_i jest zmienną losową, której realizacją jest zaobserwowana wartość y_i. Zgodnie z wytycznymi dotyczących naszego zadania, zakładamy, że wszystkie zminne losowe ξ_i są idd. o rozkładzie normalny ze średnią 0. Naszym celem będzie estymować wartość $\hat{y}_i$ poprzez estymowanie realizacji zmiennych losowych $\hat{\beta}_0$ oraz $\hat{\beta}_1$.

4.1 Estymacja punktowa

By stworzyć estymator ceny $\hat{y}$ skorzystamy z metody najmniejszych kwadratów. Do estymacji parametrów β_0 oraz β_1, występujące we wzorze (1), wykorzystamy losowo wybrane 80% naszych danych. Pozostałe 20% posłuży nam w celu sprawdzenie poprawności modelu. Dane zostały wylosowane przy użyciu podstawowych funkcji języka `Julia` z "inżynierskim" ziernem, a mianowicie `1410`. Estymowane parametry, w zaobserwowanej realizacji $Y = (Y_1, Y_2, \cdots, Y_n)$, mają wartość

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \approx 39.089 \quad \text{oraz} \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \approx -1548.34\ .$$

Zatem estymowana wartośc ceny $\hat{y}$ modelu (1) będzie miała postać

$$\hat{y}_i = 38.089 \cdot x_i - 1548.34\ .$$

4.2 Estymacja przedziałowa

W tym przypadku będziemy szukać przedziału do w którym będzie należał nasz Y_i z dużym prawdopodobieństwem. Będziemy chcieli by prowdopodobieństow to wynosiło $1 - \alpha$. Znany jest fakt, że w modelu (1) zmienna $\hat{Y}_i$ ma poniższe parametry

$$\mathbb{E}\hat{Y}_i = \beta_0 + \beta_1 x_i \quad \text{oraz} \quad Var\hat{Y}_i = Var(\xi_1)\left(\frac{1}{n} + \frac{(x_0 - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

Jeśli $Var(\xi_1)$ nie jest znana w jej miejsce wstawiamy jej estymator

$$s^2 = \frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)$$

otrzymując estymator $Var\hat{Y}_i$. W tym przypadku zmienna losowa

$$\frac{\hat{Y}_i - \mathbb{E}\hat{Y}_i}{s^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}$$

ma rozkład t-studenta z $n-2$ stopniami swobody. Przez $z_{\alpha/2}$ będziemy oznaczać $1-\alpha/2$ kwantyl z właśnie tego rozkładu. Dla estymowane wartość $\hat{Y}_0$, dla znanej objętości, równej x_0 będzie zachodziła równość

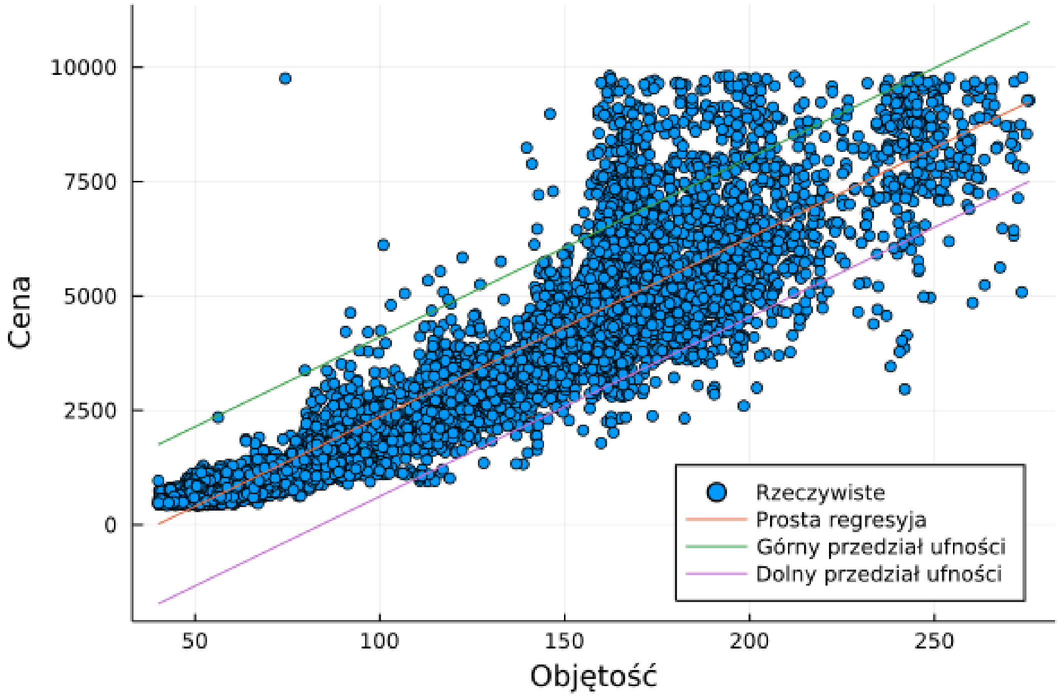
$$\mathbb{P}\left(\hat{Y}_0 - z_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} \leqslant Y_0 \leqslant \hat{Y}_0 + z_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right) = 1 - \alpha.$$

Zatem naszym przedziałem ufności ceny dla objętości x_0 będzie przedział

$$\left(\hat{y}_0 - z_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}, \hat{y}_0 + z_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right).$$

4.3 Predykcja danych

Tak jak wspominaliśmy 20% danych posłuży nam do sprawdzenia modelu. Dla tych danych przyjrzymy się wartością rzeczywistym, estymowanym punktowo oraz przedziałowo.



Rysunek 8: Dane testowe

Analizując wykres możemy zauważyć, że dane są w oklicach prostej regresyjnej 4.1 dla małych objętości. Wraz ze wzrostem iloczynu wymiarów rośnie też odległość danych od prostej. Do podobnych wniosków możemy dojść analizując przedziały ufności dla naszych danych 4.2. Fakt ten może sugerować nam, że objętość nie wpływa liniowo na cenę diamentu, w szczególności cena większych diamentów nie jest liniowa od iloczynu wymiarów diamentu. Może wpływać na to fakt, że diamenty mają jeszcze inne cechy, których tutaj nie bierzemy pod uwagę.

5 Analiza residuów

Podczas tworzenia modelu regresji liniowej oraz dalszych obliczeń zakładaliśmy następujące warunki

1. $\mathbb{E}\xi_i = 0$;
2. $Var\xi < \infty$;
3. ξ_i mają rozkład normalny;
4. $\xi_i \perp\!\!\!\perp \xi_j$ dla $i \neq j$.

Do sprawdzienia tych warunków wykonamy analizę residuów e_i danych wzorem

$$e_i = y_i - \hat{y}_i$$

Zaczniemy od sprawdzenia średniej rozkładu. W naszej próbie wynosi

$$\mathbb{E}\xi_i \approx \overline{e} = 1.18e - 8,$$

zatem pierwsze założenie jest spełnione, co nie jest zaskakujące, uwzględniając, że skorzystaliśmy z metody najmniejszych kwadratów. Do wyestymowania wariancji posłużymy się estymatorem

$$s_k^2 = \frac{1}{k-1} \sum_{i=0}^{k} e_i.$$