# Raport 1

Jakub Kempa, Szymon Stano

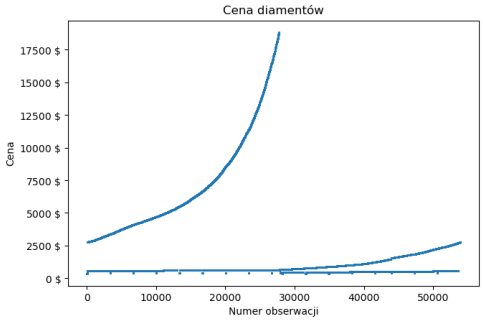
4 lipca 2023

## 1 Wstęp

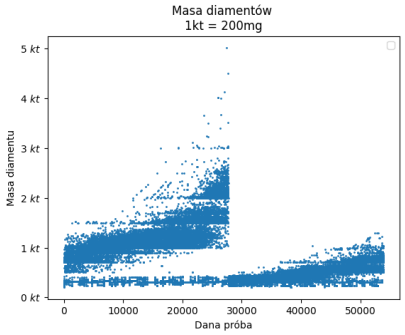
W niniejszym raporcie analizie poddane zostaną dane dotyczące diamentów, które pochodzą ze strony:
https://vincentarelbundock.github.io/Rdatasets/doc/ggplot2/diamonds.html. Jest to artukuł pt. *Prices of over 50,000 round cut diamonds*. Analizowana przez nas próba zawiera 53 940 rzędów, w których znajdują się informacje dotyczące atrybutów poszczególnego diamentu, takie jak:

- cena
- masa w karatach
- jakość oszlifowania
- kolor
- wymiary

Przez nas analizowane będą dwa atrybuty diamentów: cena oraz masa. Celem będzie analiza statystyczna wartości obu parametrów z osobna oraz porównanie dwóch zestawów razem. Poniżej przedstawione są analizowane przez nas dane.




(a) Wykres ceny




(b) Wykres masy

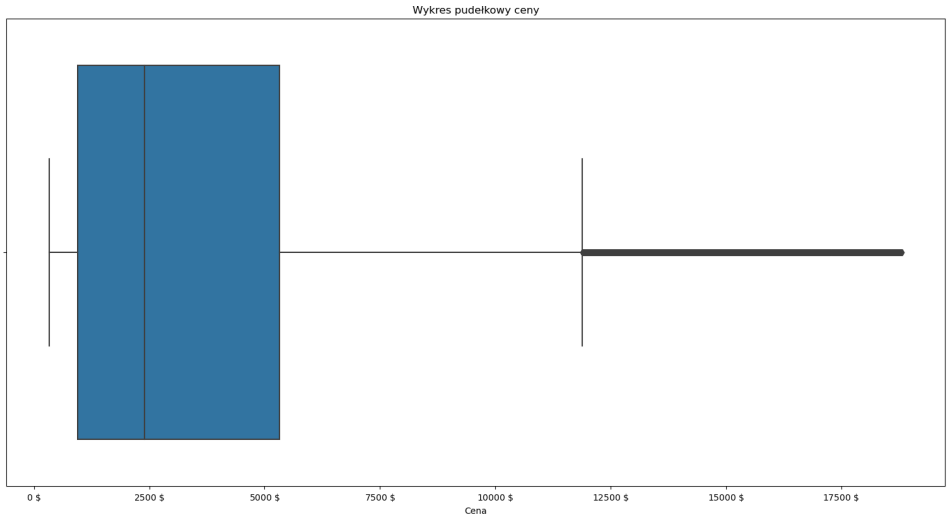
Łatwo zauważyć, że dane cen diamentów wyglądają jak trzy ciągłe funkcje. Może to wynikać z wcześniejszego, specyficznego posortowania próbek i wprowadzenia ich w takim posortowanym wg ceny stanie. Ze względu na to podobnie zachowuje się wykres mas diamentów. Masa diamentu na rynku kamieni szlachetnych jest silnie skorelowana z ceną. Logicznym założeniem jest, że wraz z masą diamentu rośnie jego cena. Większe rozproszenie danych masy wynika pewnie z wpływu innych czynników, takich jak wymiary diamentu, jakość oszlifowania, kolor czy przejrzystość.

# 2 Analiza danych dotyczących ceny

## 2.1 Podstawowe statystyki i wykres pudełkowy

Dla obu zestawów danych możemy obliczyć podstawowe statystki. Wszystkie przedstawione dla ceny są w tabeli poniżej oraz na wykresie pudełkowym:

| Statystyka | Wartość [$] |
|---|---|
| Średnia arytmetyczna | 3932.80 |
| Odchylenie standardowe | 3989.44 |
| Wartość minimalna | 326 |
| Wartość maksymalna | 18823 |
| Q1 | 950 |
| Q2 (mediana) | 2401 |
| Q3 | 5324.25 |
| Rozstaw międzykwartylowy (IQR) | 4374.25 |
| Średnia harmoniczna | 1524.87 |
| Średnia geometryczna | 2408.52 |

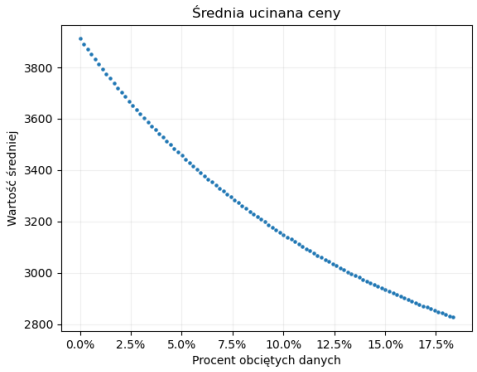



Rysunek 1: Wykres pudełkowy wartości cen

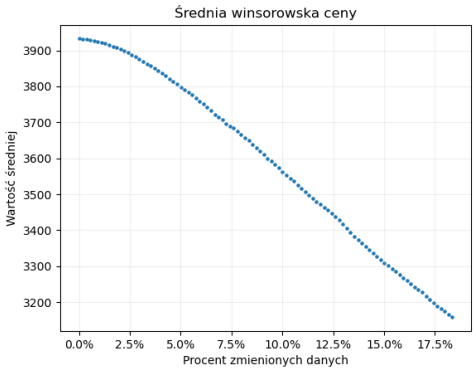
Widzimy że dane w dużej mierze skondensowane są wokół mniejszych wartości, lecz występuje także bardzo dużo wartości skrajnych. Wskazuje na to odchylenie standardowe, które jest zbliżone do średniej. Rozstaw wartości cen diamentów jest szeroki, dlatego widocznych jest wiele wartości odstających, które nie mieszczą się w skrajnych wąsach wykresu (granice te definiowane są jako $Q_1 - 1.5IQR$ oraz $Q_3 + 1.5IQR$). W przypadku niskich wartości jest brak wartości odstających, ponieważ rozstaw międzykwartylowy jest na tyle duży. Ma to związek z częstością sprzedawania diamentów bardziej pospolitych oraz ogromną ceną tych rzadszych. Dokładniejsza analiza dlaczego tak się dzieje zostanie przeprowadzona w dalszej części raportu.

## 2.2 Średnia winsorowska i ucinana

Kolejnym krokiem w analizie danych jest ukazanie zmieniających się wartości średniej winsorowskiej oraz ucinanej. W przypadku analizowanych danych, które mają miedzy sobą duży rozrzut, będą one przydatne w rzeczywistej analizie wartości średnich. Średnia arytmetyczna poprzez duży rozrzut danych jest bardziej podatna na wpływ jej wartości przez skrajności, dlatego średnia winsorowska i ucinana pozwalają nam bardziej realistycznie określić średnią cenę sprzedanego diamentu, wyrzucając te skrajne transakcje.




(a) Zmienność średniej ucinanej ceny

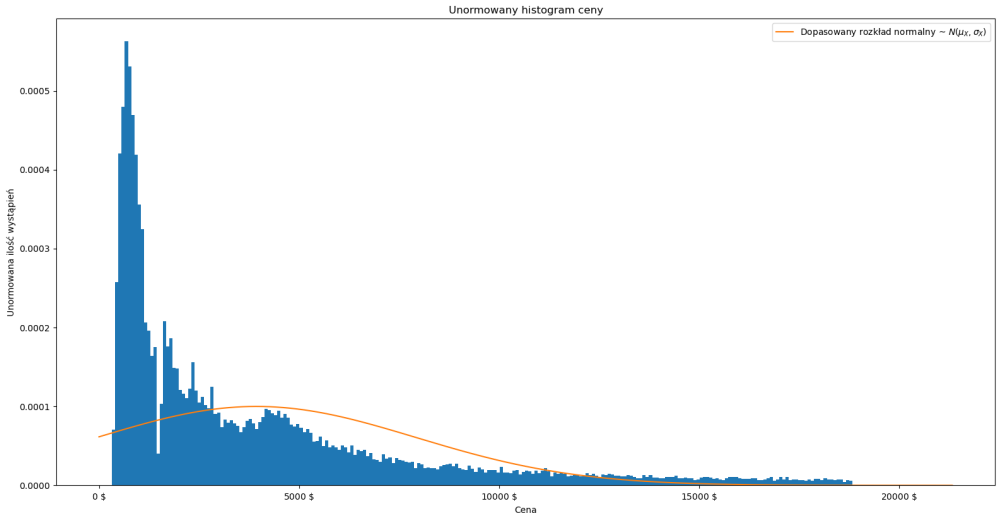



(b) Zmienność średniej winsorowskiej ceny

Na powyższych wykresach ukazana jest tendencja spadkowa średnich, przy odcinaniu lub zmienianiu danych. Ingerencja już w 10% wartości skrajnych zmniejsza średnią ucinaną do ok. 3100$, co jest spadkiem o prawie 30% względem średniej arytmetycznej. Dla jednostajnie rozłożonej próby zależność ucinania skrajnych wartości byłaby stała o wartości mediany. Kształt, który przyjmują wartości średnich ucinanych, sugeruje, dla coraz większych, ucięć wyprostowanie wykresu; rozbieżność wartości cen sprawia, że wykresy średnich nie są funkcjami stałymi.

## 2.3 Porównanie unormowanego histogramu danych z gęstością rozkładu normalnego

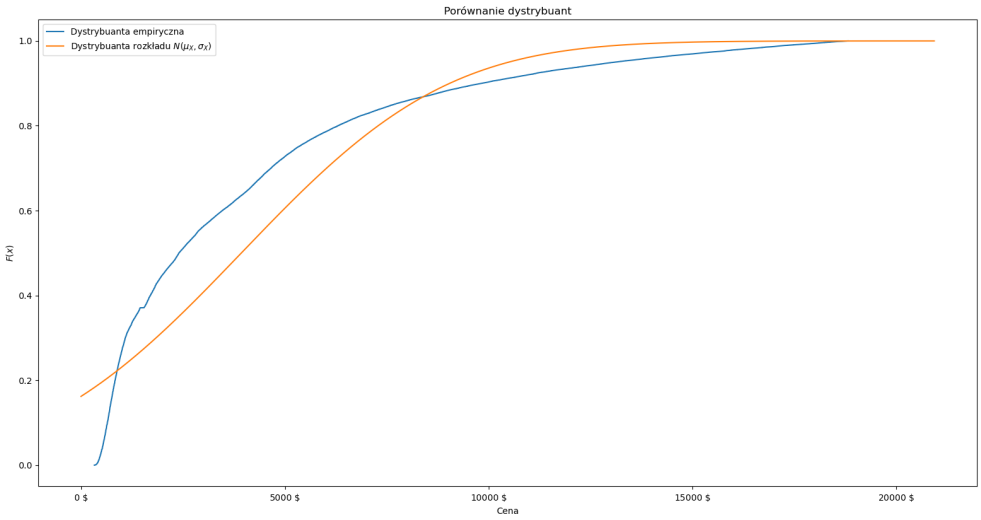
Dane dotyczące ceny diamentów można przedstawić na unormowanym histogramie. Dodatkowo, dzięki wcześniej policzonym wartościom, możemy dopasować do histogramu wykres funkcji gęstości o odpowiednich parametrach. Dla rozkładu $N(\mu, \sigma^2)$ otrzymujemy gęstość $f_X(x) \sim N(3932.8, (3989.44)^2)$. Na poniższym wykresie możemy zobaczyć zarówno histogram, jak i dopasowaną gęstość.




Rysunek 2: Histogram danych oraz dopasowana gęstość rozkładu normalnego

Wyniki są jednoznaczne: analizowane dane są mocno rozbieżne z wartościami rozkładu normalnego. Wynika to ponownie z dużej rozbieżności danych, które zawyżają średnią oraz odchylenie standardowe. Duże odchylenie sprawia, że wzrost i spadek $f_X(x)$ są łagodne, natomiast średnia arytmetyczna przesuwa środek funkcji. Gdyby średnia była niższa oraz dane nie były tak skrajnie rozstawione, to rozkład danych bardziej przypominałby rozkład normalny.
Poniżej widoczne są również porównane dystrybuanty obu rozkładów. Dystrybuanta empiryczna utworzona na podstawie analizowanych przez nas danych znacznie szybciej osiąga wysokie wartości; znacza większość wartości cen to niskie ceny, dlatego dystrybuanta empiryczna szybciej osiąga wysokie wartości, natomiast przez duży rozstaw wysokich cen i małą ilość tych cen maksymalne wartości dystrybuanty szybciej przyjmuje dystrybuanta rozkładu normalnego.

Błędnym byłaby zatem analiza danych, jako pochodzących z rozkładu normalnego, ze względu na zbyt duże rozbieżności obu rozkładów.




Rysunek 3: Porównanie empirycznej dystrybuanty danych z dystrybuantą rozkładu normalnego

## 2.4 Współczynniki skośności, zmienności oraz kurtoza

Aby szerzej udowonić rozbieżność danych możemy również policzyć współczynniki skośności i zmienności oraz kurtozę dla obu rozkładów. Wyniki tych analiz przedstawione są w poniższej tabeli:

| Statystyka | Wartość dla analizowanych danych | Wartość dla rozkładu normalnego |
|---|---|---|
| Współczynnik skośności | 1.62 | 0 |
| Współczynnik zmienności | 101.44 % | średnio przedział $[15\%, 30\%]$ |
| Kurtoza | 5.17 | 3 |

Oczywiste różnice w danych są widoczne przy każdej statystyce.

- Współczynnik skośności informuje nas o asymetrii naszych danych, która manifestuje się poprzez wydłużone prawe ramię rozkładu (prawostronnie skośny rozkład). Odbiega to od symetrii rozkładu normalnego.
- Współczynnik zmienności informuje nas o zmienności ("stromości") rozkładu. Tutaj widać największą rozbieżność z rozkładem normalnym, którego wartości przeważnie mieszczą się w przedziale od 15% do 30%. Sugeruje to większą zmienność rozkładu dla analizowanych danych.
- Kurtoza, czyli miara kształtu rozkładu, analizuje co się dzieje w ogonach rozkładu. W naszym przypadku, ze względu na dużą rozbieżność wartości skrajnych, kurtoza jest wyższa, niż dla rozkładu normalnego.

Pełna analiza sugeruje zatem, że błędnym byłoby przypasowanie do danych rozkładu normalnego, ze względu na duże różnice obu rozkładów.
Dodatkowo odrzucenie tej hipotezy potwierdzają testy statystyczne: Kołmogorowa-Smirnowa - służący do porównywania rozkładów jednowymiarowych oraz Jarque-Bera - badający bezpośrednio normalność rozkładu:

- KstestResult(statistic=0.9999999056036948, pvalue=0.0)
- Jarque_beraResult(statistic=34200.714921690815, pvalue=0.0)

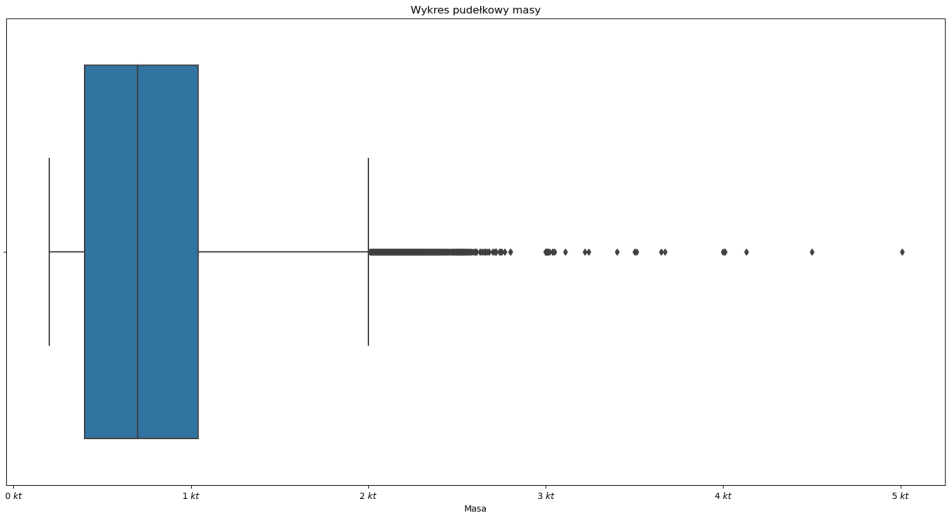
Parametry *statistics* określają odległość dystrybuanty empirycznej próby od dystrybuanty teoretycznej. Im większa wartość, tym rozkłady te są bardziej oddalone, co w skali dla danych testów kolejno do 1 oraz do $\infty$, bezpośrednio stwierdza sprzeczność założeń. Dodatkowo, obydwie wartości *pvalue*, będące prawdopodobieństwem z jakim możemy przyjąć hipotezę zgodności, wynoszą 0, nie pozwalając na przyjęcie jakiegokolwiek poziomu istotności, tym samym potwierdzając wcześniejsze wnioski.

# 3 Analiza danych dotyczących masy

## 3.1 Podstawowe statystyki i wykres pudełkowy

Dla obu zestawów danych możemy obliczyć podstawowe statystki. Wszystkie przedstawione dla masy są w tabeli poniżej oraz na wykresie pudełkowym:

| Statystyka | Wartość $[kt]$ |
|---|---|
| Średnia arytmetyczna | 0.798 |
| Odchylenie standardowe | 0.2247 |
| Wartość minimalna | 0.2 |
| Wartość maksymalna | 5.01 |
| Q1 | 0.4 |
| Q2 (mediana) | 0.7 |
| Q3 | 1.04 |
| Rozstaw międzykwartylowy (IQR) | 0.64 |
| Średnia harmoniczna | 0.5724 |
| Średnia geometryczna | 0.6737 |



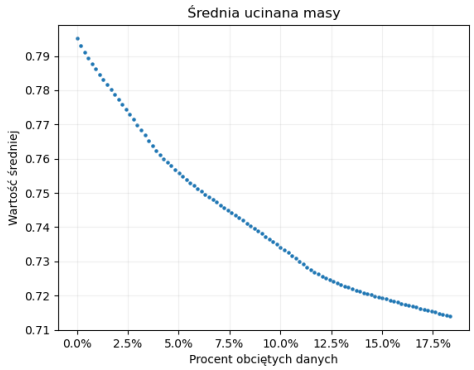

Rysunek 4: Wykres pudełkowy wartości mas diamentów

Podobnie, jak w przypadku cen, można zauważyć dużą rozbieżność danych. Wartość średniej arytmetycznej jest jednak w tym przypadku bardziej zbliżona do mediany oraz odchylenie stan-

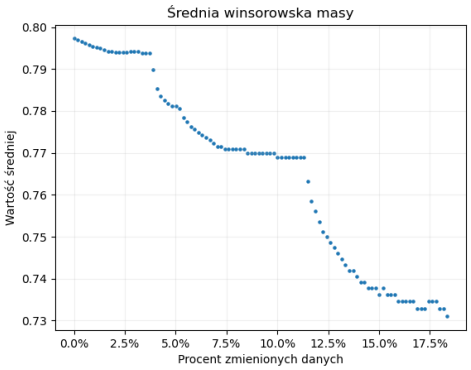
dardowe danych jest bardziej oddalone od średniej arytmetycznej (wciąż jednak jest duże). Po dokładniejszej analizie widać znaczącą różnicę pomiędzy trzecim kwartylem, a maksymalną wartością, którą dobrze widać na wykresie pudełkowym w postaci wielu wartości odstających.
Różnicą statystykach cen i mas wydaje się być nieliniowa zależność jednego od drugiego; choć wartości masy są wciąż rozbieżne, to nie są aż tak rozbieżne, jak wartości cen.

## 3.2 Średnia winsorowska i ucinana

Kolejnym krokiem w analizie danych jest ukazanie zmieniających się wartości średniej winsorowskiej oraz ucinanej. W przypadku analizowanych danych, które mają miedzy sobą duży rozrzut, będą one przydatne w rzeczywistej analizie wartości średnich. Średnia arytmetyczna poprzez duży rozrzut danych jest bardziej podatna na wpływ jej wartości przez skrajności, dlatego średnia winsorowska i ucinana pozwalają nam bardziej realistycznie określić średnią masę sprzedanego diamentu, wyrzucając te skrajne.




(a) Zmienność średniej ucinanej masy

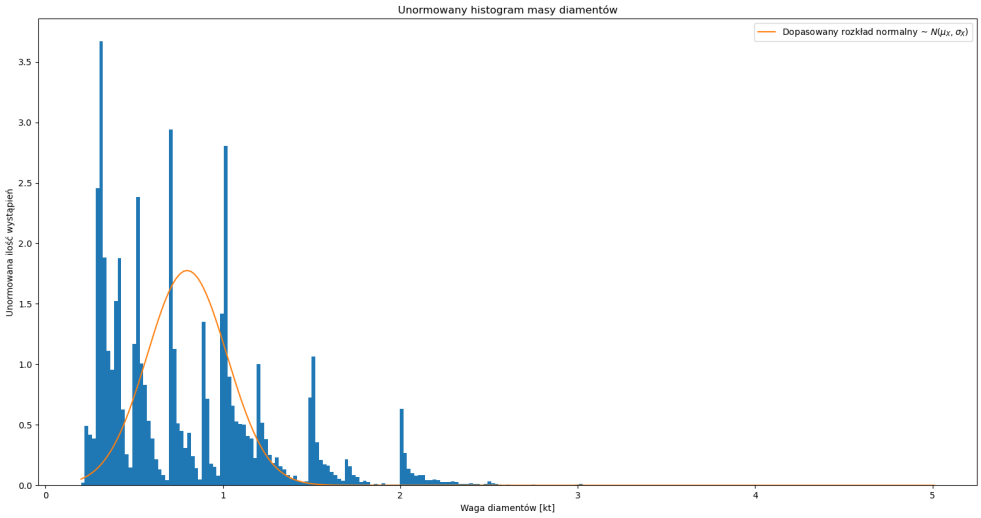



(b) Zmienność średniej winsorowskiej masy

Na powyższych wykresach ukazana jest tendencja spadkowa średnich, przy odcinaniu lub zmienianiu danych. Szczególnie ciekawy jest wykres zmieniającej się średniej winsorowskiej. Gładkie, poziome odcinki wartości sugerują symetryczne nagromadzenie dużych wartości z małymi wartościami, co wpływa na średnią, stabilizując ją - natomiast zamieniając coraz to większy procent skrajnych danych wykres ponownie zaczyna gwałtowniej spadać. Można więc wywnioskować, że rozbieżność mas jest duża, jednak silnie skupiona w kilku obszarach.

## 3.3 Porównanie unormowanego histogramu danych z gęstością rozkładu normalnego

Dane dotyczące ceny diamentów można przedstawić na unormowanym histogramie. Dodatkowo, dzięki wcześniej policzonym wartościom, możemy dopasować do histogramu wykres funkcji gęstości o odpowiednich parametrach. Dla rozkładu $N(\mu, \sigma^2)$ otrzymujemy gęstość $f_X(x) \sim N(0.798, (0.2247)^2)$. Na poniższym wykresie możemy zobaczyć zarówno histogram, jak i dopasowaną gęstość.

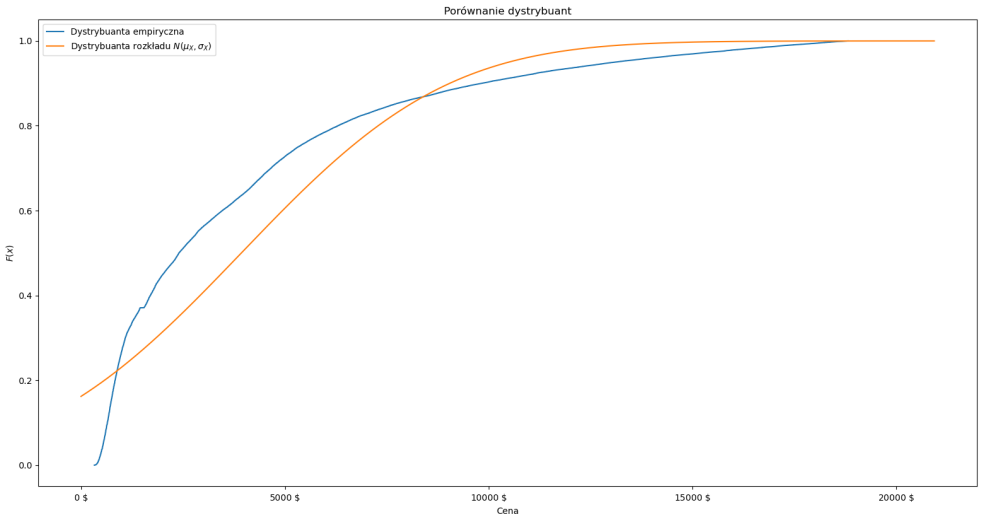



Rysunek 5: Histogram danych oraz dopasowana gęstość rozkładu normalnego

Histogram analizowanych danych potwierdza wnioski wyciągnięte z wykresu zmieniających się wartości średniej winsorowskiej. W kilku miejscach zauważalne są podskoki wartości ilości mas diamentów. Najprawdopodobniej bierze się to z tendencji ludzi do wybierania bardziej okrągłych liczb. Niezależnie czy masa zależy od kupującego, czy sprzedawcy, zdecydowana większość mas diamentów ma "okrągłą liczbę widać skoki przy wartościach karata 0.5, 1, 1.5, 2, 2.5, 3 itd.

Porównując dany histogram do rozkładu normalnego ponownie widać dużą rozbieżność obu gęstości.
Poniżej widoczne są również porównane dystrybuanty obu rozkładów. Dystrybuanta empiryczna masy zachowuje się podobnie, co dystrybuanta empiryczna cen, i, podobnie jak poprzednia, nie pokrywa się z dystrybuantą rozkładu normalnego. Masy również bardzo szybko (choć łagodniej, niż w przypadku cen) osiągają niskie wartości. Podobnie - maksymalne wartości dystrybuanty szybciej przyjmuje dystrybuanta rozkładu normalnego.

Błędnym byłaby zatem analiza danych, jako pochodzących z rozkładu normalnego, ze względu na zbyt duże rozbieżności obu rozkładów.




Rysunek 6: Porównanie empirycznej dystrybuanty danych z dystrybuantą rozkładu normalnego

## 3.4 Współczynniki skośności, zmienności oraz kurtoza

Aby szerzej udowonić rozbieżność danych możemy również policzyć współczynniki skośności i zmienności oraz kurtozę dla obu rozkładów. Wyniki tych analiz przedstawione są w poniższej tabeli:

| Statystyka | Wartość dla analizowanych danych | Wartość dla rozkładu normalnego |
|---|---|---|
| Współczynnik skośności | 1.12 | 0 |
| Współczynnik zmienności | 59.4 % | średnio przedział $[15\%, 30\%]$ |
| Kurtoza | 4.26 | 3 |

Oczywiste różnice w danych ponownie są widoczne przy każdej statystyce.

- Współczynnik skośności informuje nas o asymetrii naszych danych, która manifestuje się poprzez wydłużone prawe ramię rozkładu (prawostronnie skośny rozkład). Odbiega to od symetrii rozkładu normalnego.
- Współczynnik zmienności informuje nas o zmienności (”stromości”) rozkładu. Procent współczynnika zmienności jest mniejszy, niż dla cen, jednak wciąż odbiega od odpowiednich procentów współczynnika zmienności dla wykładu normalnego, które przeważnie mieszczą się w przedziale od 15% do 30%. Sugeruje to większą zmienność rozkładu dla analizowanych danych.
- Kurtoza, czyli miara kształtu rozkładu, analizuje co się dzieje w ogonach rozkładu. W naszym przypadku, ze względu na dużą rozbieżność wartości skrajnych, kurtoza jest wyższa, niż dla rozkładu normalnego.

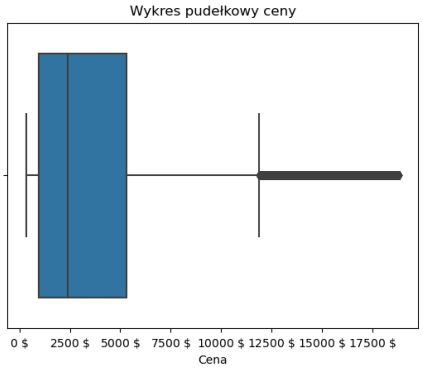
Pełna analiza sugeruje zatem, że, podobnie jak dla cen, błędnym byłoby przypasowanie do danych dla masy diamentów rozkładu normalnego, ze względu na duże różnice obu rozkładów. Tak samo w tym przypadku testy statystyczne potwierdzają nasze obserwacje:

- KstestResult(statistic=1.0, pvalue=0.0)
- Jarque_beraResult(statistic=14756.80441887217, pvalue=0.0)

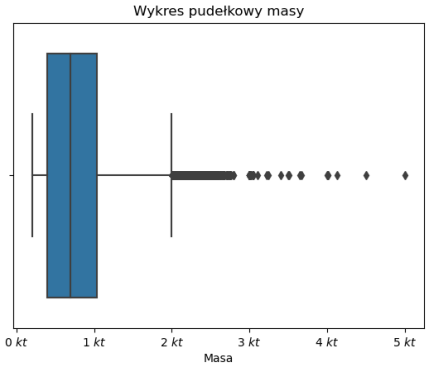
# 4 Analiza porównawcza

## 4.1 Wykresy pudełkowe

W następującej części raportu analizowane będą dwa zbiory danych względem siebie, zależności między nimi, charakterystyki. Poniżej przedstawione są ponownie wykresy pudełkowe obu zbiorów danych oraz wykresy pudełkowe ceny w zależności od masy.




(a) Wykres pudełkowy ceny

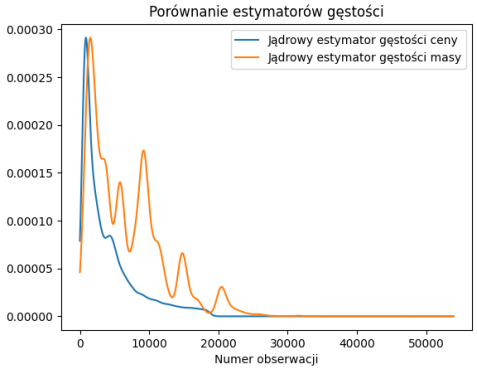



(b) Wykres pudełkowy masy

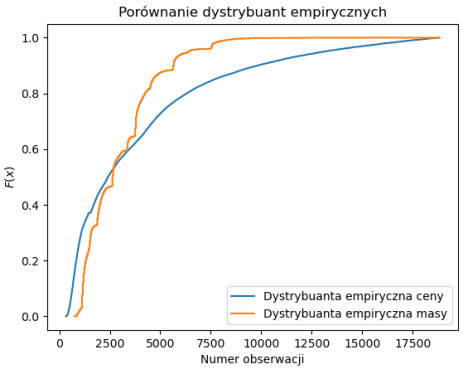
Porównując oba wykresy pudełkowe do siebie widać zauważalne różnice. Mediana cen znajduje się bliżej pierwszego kwartyla, podczas gdy mediana masy jest równomiernie rozłożona pomiędzy swoimi kwartylami. Rozstaw międzykwartylowy w porównaniu do wszystkich wartości jest mniejszy dla masy, niż dla ceny, co wpływa również na rozstaw wąsów wykresu. Mimo, że wąsy wykresu pudełkowego masy są węższe, to ilość wartości odstających wydaje się być mniejsza, niż w przypadku ceny. Mniejszy rozstaw międzykwartylowy oznacza również bardziej skondensowane wartości wokół mediany; duży rozstaw kwartylowy sugeruje dla cen większą rozbieżność wartości.

## 4.2 Estymatory gęstości oraz dystrybuanty empiryczne

Możemy z danych wyznaczyć dystrybuanty empiryczne oraz estymatory gęstości. Przedstawione gęstości i dystrybuanty są przestawione i porównane na wykresach poniżej.




(a) Porównanie estymatorów gęstości




(b) Porównanie empirycznych dystrybuant

Widocznych jest kilka faktów:

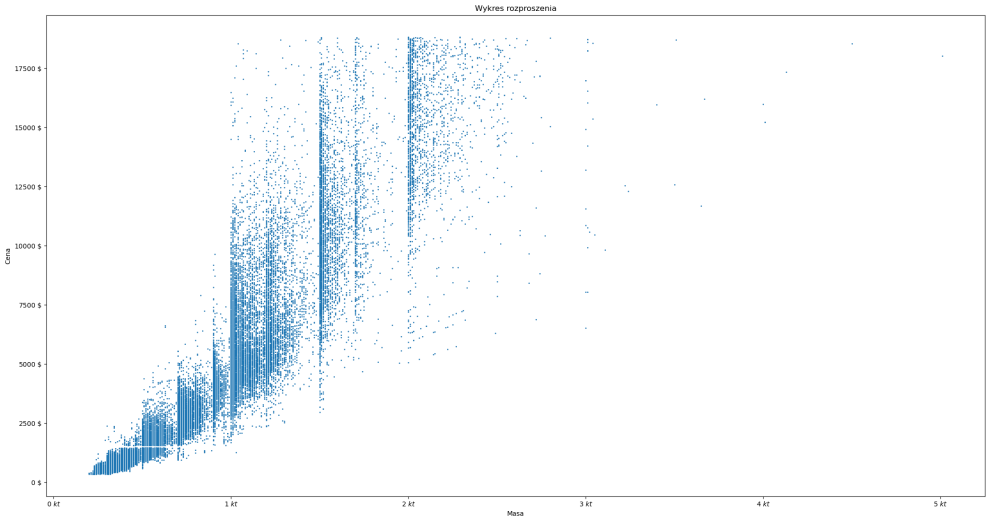
- Wykresy obu estymatorów gęstości pokrywają się z wcześniej wyznaczonymi unormowanymi histogramami tych rozkładów, co potwierdza estymację.

- Wyprostowanie wykresu dla obu funkcji znajduje się w innym miejscu oraz estymator gęstości masy ma dużo spadków i wzrostów w miejscach, w których estymator gęstości ceny jest gładki.
- Dystrybuanta empiryczna ceny jest znacznie bardziej gładka, niż dystrybuanta empiryczna masy, której trajektoria jest znacznie bardziej nieregularna - miejscami przypomina dystrybuantę empiryczną.

Oba estymatory gęstości mają dość dużą rozbieżność względem siebie, co wskazuje na rozbieżność danych. Mimo, że dane dystrybuanty oraz gęstości raczej się rozbiegają, to nie wyklucza to zależności pomiędzy próbkami.

## 4.3 Zależności danych i współczynniki korelacji

Ostatnim krokiem analizy porównawczej będzie przeanalizowanie danych pod kątem zależności tych danych od siebie. Użytecznym będzie wykonanie wykresu wartości cen z próby, w zależności od wartości mas. Poniżej widnieje wykres takiej zależności.



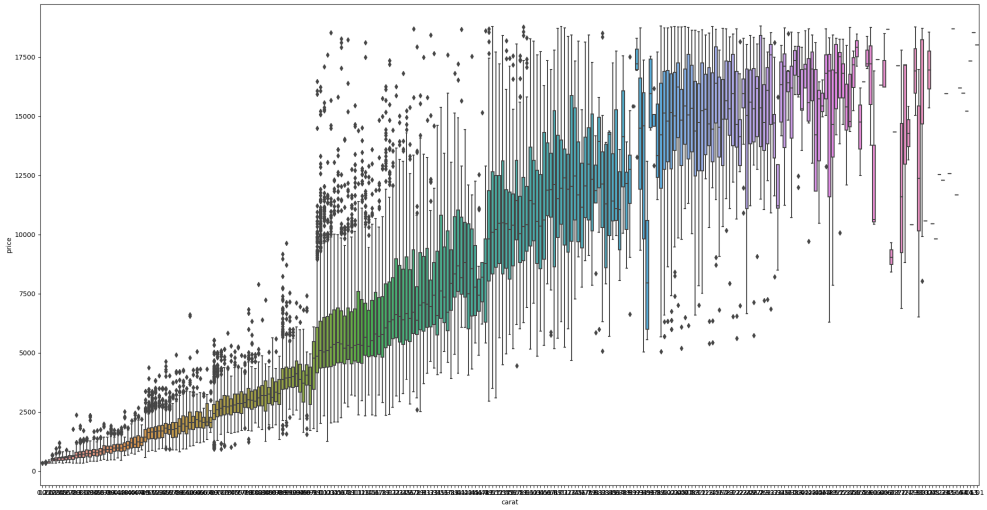

Rysunek 7: Wykres rozproszenia cen w zależności od masy

Powyższy rysunek wskazuje na pewną liniową, bądź potęgową zależność danych od siebie. Logicznym jest, że im większa masa diamentu, tym większa będzie jego cena - dane to również odzwierciedlają. Widoczne są tutaj tendencje wartości masy do osiągania wcześniej wspomnianych *okrągłych* wartości oraz zauważyć można (szczególnie dla niskich wartości) pewne pionowe luki masy, które najprawdopodobniej wskazują na dokładność wartości masy w danym zestawie.
Zależność danych od siebie jest widoczna, jednak ceny za tą samą próbę diamentu potrafią się od siebie bardzo różnić. Wynika to z faktu, że na cenę diamentu wpływa również wiele innych czynników, takich jak jakość oszlifowania, przejrzystość czy wymiary. Przykładowo diament o tej samej masie może kosztować więcej, bo jest przejrzystrzy od standardowego diamentu o takiej masie.

Aby dokładnie przeanalizować podane dane pod kątem statystycznym policzyliśmy współczynnik korelacji Pearsona oraz Spearmana. Otrzymaliśmy wartości kolejno: 0.921591301193477 oraz 0.9628827988813001. Najwyższą wartością koorelacji jest liczba 1, która oznacza zależność dwóch prób losowych i idealną koorelację danych. Oznacza to, że analizowane dwa zestawy danych są ze sobą silnie powiązane.
Współczynnik koorelacji Pearsona jest mniejszy, niż współczynnik koorelacji Spearmana, który nie bada liniowej zależności, tylko dowolną monotoniczną zależność, co sugeruje, że prawdopodobnie bardziej niż funkcja liniowa, będzie do danych pasować funkcja wykładnicza.

Liczbą również podawaną podczas generowania tych współczynników jest wartość *pvalue*, która w tym przypadku odpowiada prawdopodobieństwu tego, że korelacja między tymi zmiennym jest przypadkowa. W obydwu przypadkach wartość wynosiła 0, co sugeruje nam, że powiązanie i zależność obu zestawów danych ze sobą nie było kwestią przypadku.

Na wykresie poniżej przedstawione są wykresy pudełkowe cen dla wielu wybranych wartości masy. Zachowania tego wykresu odpowiadają zachowaniom wykresu rozproszenia cen; dla małych wartości masy diamentów ceny są bardziej skondensowane wokół mediany (niewielka ilość, bądź brak wartości odstających), im większa waga, tym większy rozstaw międzykwartylowy oraz tym więcej wartości odstających. Taki wygląd wykresów pudełkowych zwraca uwagę na fakt rozrzucenia wartości ceny dla coraz to większych mas diamentów.




Rysunek 8: Wykres pudełkowy ceny w zależności od masy

# 5 Podsumowanie i wnioski

Analiza wybranych przez nas danych udowodniła użyteczność statystyk, z których korzystaliśmy. Z wybranej przez nas próby ponad 50 000 danych byliśmy w stanie określić zachowania oraz charakterystyki próby.

Histogram, estymatory gęstości oraz dystrybuanty empiryczne bezpośrednio wskazują na to, że naszych danych nie można analizować pod kątem rozkładu normalnego, ze względu na średnią, odchylenie standardowe, wartości skrajne oraz asymetryczność. Błędnym byłoby zatem analizowanie danych pod kątem rozkładu normalnego zadanego naszymi wartościami statystyk. Ponadto, wykazaliśmy wysoką korelację obu zbiorów danych oraz zależność liniową lub wykładniczą. Wyjaśniona została również duża rozbieżność danych oraz nieregularność prób.

Choć cena diamentu zależy od wielu czynników, takich jak jakość oszlifowania, przejrzystość czy wymiary, to zauważalna jest zależność ceny diamentów od masy. Duża rozbieżność wskazuje na popyt ciężkich diamentów, których cena również wzrasta, a skoki w histogramach biorą się z czynnika ludzkiego i tendencji do faworyzowania *okrągłych* liczb.