Politechnika Wrocławska

Raport 1

Pakiety Statystyczne

Prowadzący kurs: dr inż. Agnieszka Kamińska

Aleksander Rzyhak

nr indeksu: 268766

Michał Tokarski

nr indeksu: 268747

17 stycznia 2024

1 Wstęp i opis danych

Celem sprawozdania była analiza wybranych danych na podstawie wiedzy zdobytej na laboratoriach i wykładzie z Pakietów statystycznych. Badanie polegało na próbie odpowiedzenia na przygotowane pytania badawcze. Raport napisano korzystając z języka programowania R, przy pomocy pakietu KnitR.

Dane zaczerpnięte zostały ze strony Kaggle[1], a dotyczą sprzedaży oraz ocen gier komputerowych. Wartości dotyczące sprzedaży zostały stworzone na bazie strony VGChartz, natomiast oceny gier zostały wzięte ze strony Metacritic. Całkowita ilość badanych gier to 16719.

Dane określone są następującymi kolumnami:

- **Name** - nazwa gry (zmienna tekstowa, 11563 różnych wartości),
- **Platfrom** - nazwa platformy (zmienna tekstowa, 31 różnych wartości),
- **Year_of_Release** - rok wydania (zmienna tekstowa, 40 różnych wartości),
- **Genre** - gatunek gry (zmienna tekstowa, 13 różnych wartości),
- **Publisher** - wydawca gry (zmienna tekstowa, 581 różnych wartości),
- **Na_Sales** - ilość sprzedanych kopii gry (w milionach) w Ameryce Północnej (zmienna liczbowa, wartości od 0 do 41.36),
- **Eu_Sales** - ilość sprzedanych kopii gry (w milionach) w Europie (zmienna liczbowa, wartości od 0 do 28.96),
- **JP_Sales** - ilość sprzedanych kopii gry (w milionach) w Japonii (zmienna liczbowa, wartości od 0 do 10.22),
- **Other_Sales** - ilość sprzedanych kopii gry (w milionach) w pozostałej części świata (zmienna liczbowa, wartości od 0 do 10.57),
- **Global_Sales** - całkowita ilość sprzedanych kopii gry (w milionach) (zmienna liczbowa, wartości od 0.01 do 82.53),
- **Critic_Score** - średnia ocena gry przez krytyków w skali od 0 do 100 (zmienna liczbowa, wartości od 13 do 98),
- **Critic_Count** - ilość krytyków, którzy oceniili daną grę (zmienna liczbowa, wartości od 3 do 113),
- **User_Score** - średnia ocena gry przez graczy w skali od 0 do 10 (zmienna liczbowa, wartości od 0 do tbd),
- **User_Count** - ilość graczy, którzy oceniili daną grę (zmienna liczbowa, wartości od 4 do 10665),
- **Developer** - studio, które stworzyło daną grę (zmienna tekstowa, 1697 różnych wartości),
- **Rating** - kategoria wiekowa wyznaczona przez ESRB. (zmienna tekstowa, 9 różnych wartości).

[1] https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

Analiza danych dotyczyła następujących pytań badawczych:

1. Czy sprzedaż gier rozkłada się podobnie dla różnych rynków?
2. Na jakich platformach sprzedaje się najwięcej gier oraz jacy wydawcy sprzedają ich najwięcej?
3. Czy ocena graczy i krytyków wpływa na sprzedaż (i potencjalnie jak)?

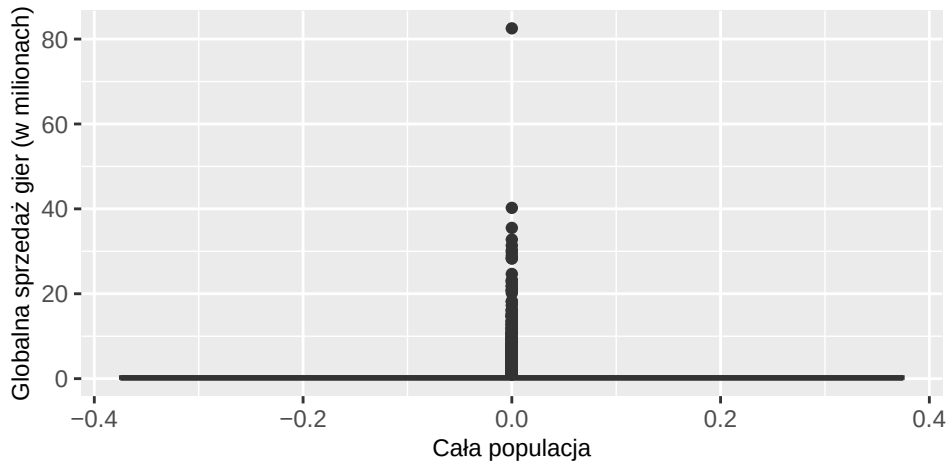
2 Wczytanie danych i obsługa ich braków

Dane zostały wczytane z pliku CSV (ang. *comma-separated values*, wartości oddzielone przecinkiem) w języku programowanie R, korzystając z funkcji read_csv(). Poza jedną kolumną, **User_Score**, nie było problemów z etykietami i typami kolumn w badanym zbiorze. Wartości kolumny **User_score** były zapisane jako tekstowy typ danych. Aby przeprowadzić dalszą analizę, zmieniliśmy ten tekstowy na typ liczbowy. Braki danych (których w pliku jest 46393) nie są z góry usuwane, ponieważ usuwając taki brak, usuwany jest cały wiersz i potencjalnie można przez to utracić przydatne dane. Obsługa tych braków większość czasu wykonywana jest automatycznie przez biblioteki tworzące wykresy. Natomiast w przypadku pewnych funkcji takich jak mean (średnia próbkowa) dodaje się dodatkowy argument na.rm=TRUE, dzięki któremu braki danych są pomijane w obliczeniach.

3 Analiza danych

3.1 Czy sprzedaż gier rozkłada się podobnie na różnych rynkach?

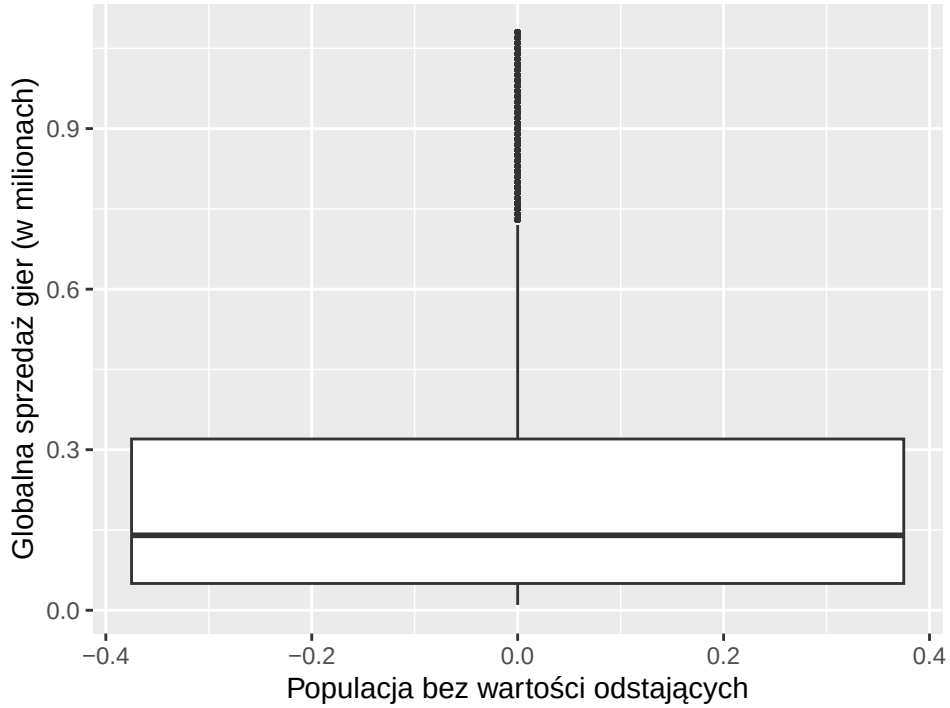
Podczas początkowej wizualizacji danych dotyczących sprzedaży zauważono dużą ilość obserwacji odstających wartości **Global_Sales**. Można to zauważyć na wykresie pudełkowym (Rys. 1), którego część pudełkowa jest tak spłaszczona, że nie widać jak jest szeroka.



Rysunek 1: Wykres pudełkowy bez odrzucenia wartości odstających

Takie same wnioski dotyczące obserwacji odstających wyciągnięto dla wartości **NA_Sales**, **EU_Sales**, **JP_Sales** i **Other_Sales**. W celu przeprowadzenia analizy dotyczącej podobieństwa rozkładów sprzedaży postanowiono odrzucić wartości odstające i silnie wpływowe. Aby to zrobić początkowo odrzucono wartości równe 0, których duża ilość silnie wpływają na rozkład, ale utrudniają analizę. Następnie wyeliminowano wartości będące poza wąsami, czyli te będące poza przedziałem $[Q_1 - 1{,}5 \cdot IQR; Q\_3 + 1{,}5 \cdot IQR]$, gdzie Q_1 i Q_3 to pierwszy i trzeci kwartyl badanych wartości, a IQR to rozstrzał międzykwartylowy.

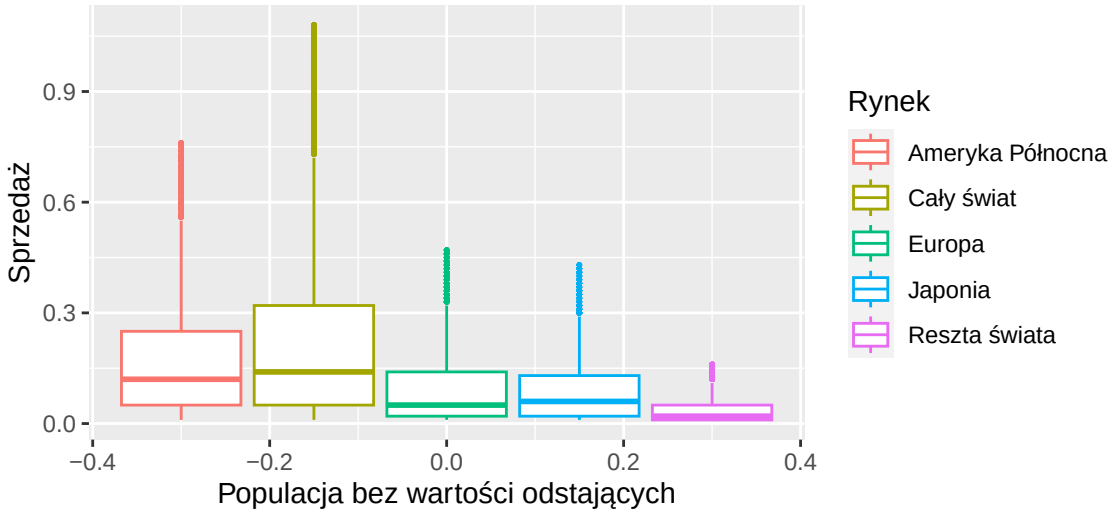
Po eliminacji niechcianych wartości, dla kolumy **Global_Sales** (których było 1892) otrzymano dane, które przedstawiono na poniższym wykresie pudełkowym (Rys. 2). Widać na nim, że wyeliminowanie wartości odstających spowodowało, że rozkład jest mniej zdegenerowany.



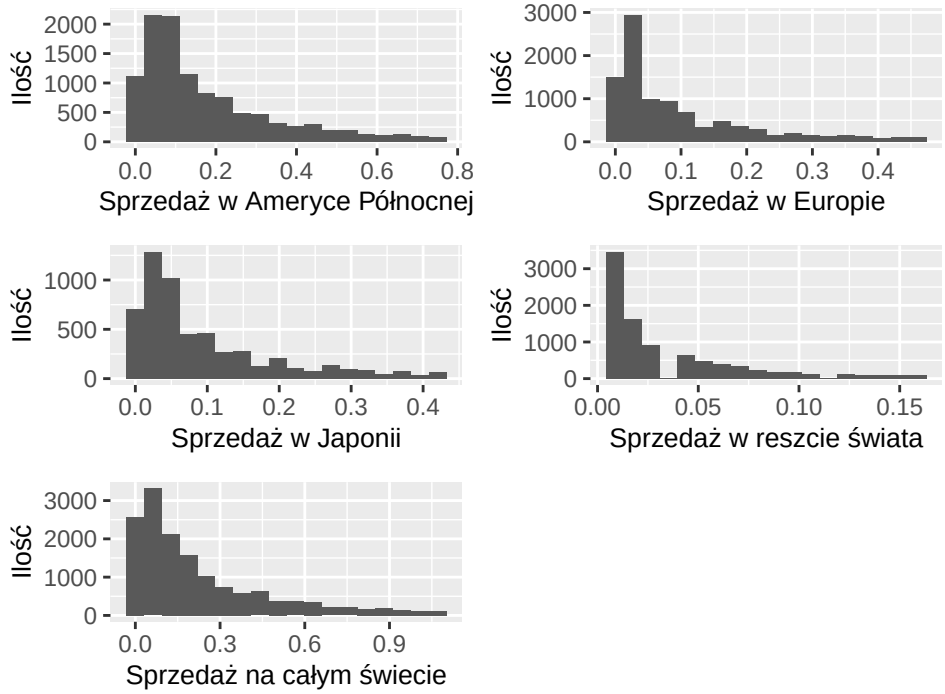
Rysunek 2: Wykres pudełkowy z odrzuconymi wartościami odstającymi

Dla wszystkich pozostałych zmiennych dotyczących sprzedaży gier wykonano podobną procedurę. Dla **NA_Sales**, **EU_Sales**, **JP_Sales** i **Other_Sales** usunięto kolejno: 5799, 7034, 11191 i 7633 wartości. Wykresy pudełkowe przedstawiające te dane narysowano na wykresie (Rys. 3). Z tak zmienionymi danymi można rozpocząć odpowiadanie na zadane pytanie badawcze.

Początkowo narysowano histogramy ilościowe danych dotyczących sprzedaży (Rys. 4), aby wizualnie sprawdzić czy wartości mają podobny rozkład. Patrząc na wykresy można stwierdzić, że prawie wszystkie są dosyć podobne. Jedynym znacznie różniącym się jest ten dotyczący kolumny **Other_Sales**.



Rysunek 3: Wykresy pudełkowe danych dotyczących sprzedaży



Rysunek 4: Histogramy ilościowe danych dotyczących sprzedaży gier

Kolejnym problemem w odpowiedzi na zadane pytanie, jest fakt, że z uwagi na różną wielkość badanych rynków dane ich dotyczące przyjmują wartości na różnych przedziałach. W celu dokładnej analizy podobieństwa unormowano tak, aby zawierały wartości niewiększe niż 1. Wykonano to dzieląc wartości w kolumnach dotyczących sprzedaży przez wartości największe w danych kolumnach.

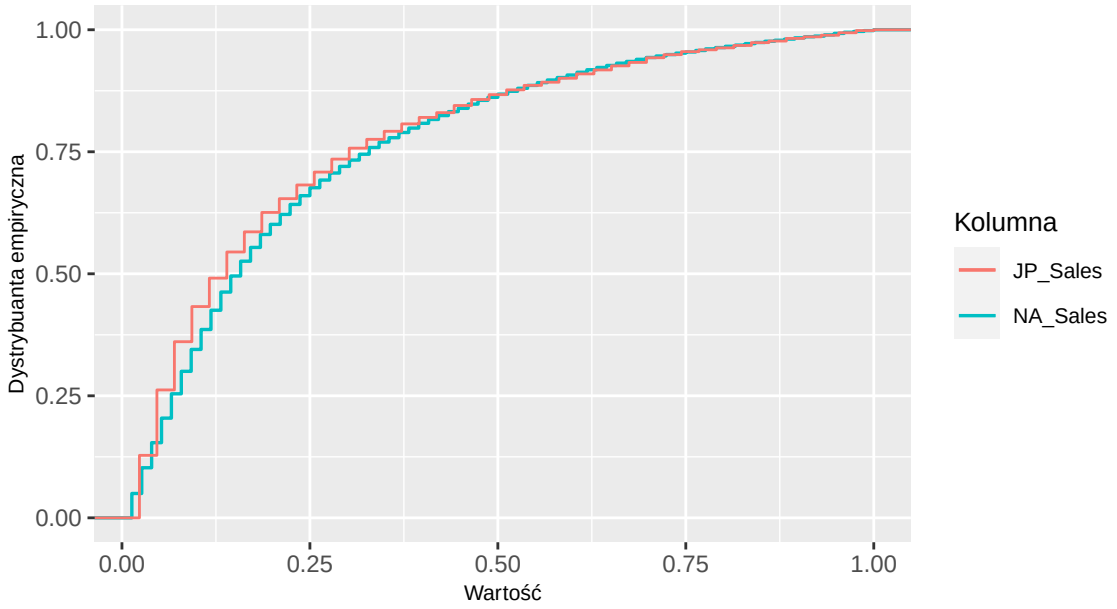
	NA_Sales	**EU_Sales**	**JP_Sales**	**Other_Sales**	**Global_Sales**
średnia	0.2345	0.2098	0.2223	0.2248	0.2123
wariancja	0.0498	0.0527	0.0519	0.0478	0.0499
skośność	1.4182	1.5862	1.4906	1.6743	1.5181
kurtoza	4.3882	4.8042	4.4873	5.2227	4.7018

Tabela 1: Podstawowe statystyki danych dotyczących sprzedaży (do 4 miejsc po przecinku)

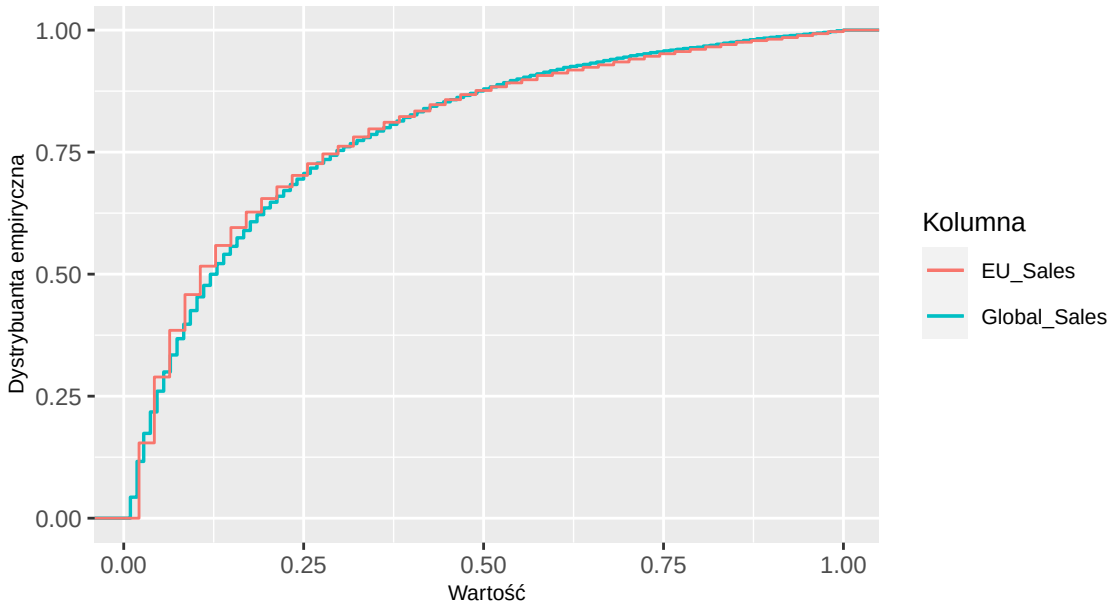
Dla tak unormowanych danych stworzono tabelkę (Tab. 1) z podstawowymi statystykami rozkładu. Widać, że wartości średniej, wariancji i skośności są dosyć mocno zbliżone między kolumnami (poza skośnością **Other_Sales**). Kurtoza jednak ma wyraźniejsze różnice, ale jest blisko (różnica mniejsza niż 0.1) między **NA_Sales** i **JP_Sales** oraz między **EU_Sales** i **Global_Sales**.

Korzystając z zauważonej zależności wykonano wykresy porównujące dystrybuanty empiryczne między wskazanymi kolumnami o najbardziej podobnych statystykach (Rys. 5 i 6). Widać na nich duże podobieństwo badanych rozkładów. Wykonano również test Kołmogorowa-Smirnowa dla dwóch prób, którego hipotezą zerową jest równość rozkładów, a alternatywną nierówność. Statystyka testowa to $D = \sup |F_{1,n}(x) - F_{2,n}(x)|$, gdzie $F_{1,n}(x)$ i $F_{2,n}(x)$ to dystrybuanty emipryczne pierwszej i drugiej próby. Test w obu przypadkach odrzucił hipotezę zerową z p-wartością rzędu 10^{-16}. Dla danych dotyczących sprzedaży w Ameryce Północnej i w Japonii wartość statystyki testowej wyniosła: $D = 0.1080908$, natomiast dla danych dotyczących sprzedaży w Europie i na całym świecie wyniosła: $D = 0.1163418$.

Dzięki otrzymanym wynikom można stwierdzić, że rozkład sprzedaży gier na różnych rynkach nie jest taki sam, ale jest między nimi dosyć duże podobieństwo szczególnie wizualne, ale też jeśli chodzi o podstawowe statystyki rozkładu. Oczywiście należy pamiętać, że dane przed analizą zostały odpowiednio obrobione, ponieważ ciężko przeprowadzić analizę, z tak mocno odstającymi obserwacjami, oraz z tak dużą ilością zer.



Rysunek 5: Porównanie dystrybuant empirycznych danych sprzeday w Ameryce Północnej i Japonii



Rysunek 6: Porównanie dystrybuant empirycznych danych sprzedaży w Europie i na całym świecie

3.2 Na jakich platformach sprzedaje się najwięcej gier oraz jacy wydawcy to robią?

Rozpoczynając analizę znaleziono wartości, a następnie stworzono tabelkę (Tab. 2) zawierającą 10 Platform dla których suma sprzedaży wszystkich gier na całym świecie była największa. Wynik nie jest zaskakujący, gdyż nie uwzględniając komputerów osobistych (PC) ta mocono pokrywa się z rankingiem dziesięciu najlepiej sprzedających się konsol[2] będącym w kolejności:

1. PS2 (PlayStation 2),
2. DS (Nintendo DS),
3. NS (Nintendo Switch),
4. GB (Game Boy),
5. PS4 (PlayStation 4),
6. PS (PlayStation),
7. Wii,
8. PS3 (PlayStation 3),
9. X360 (Xbox 360),
10. PSP (PlayStation Portable).

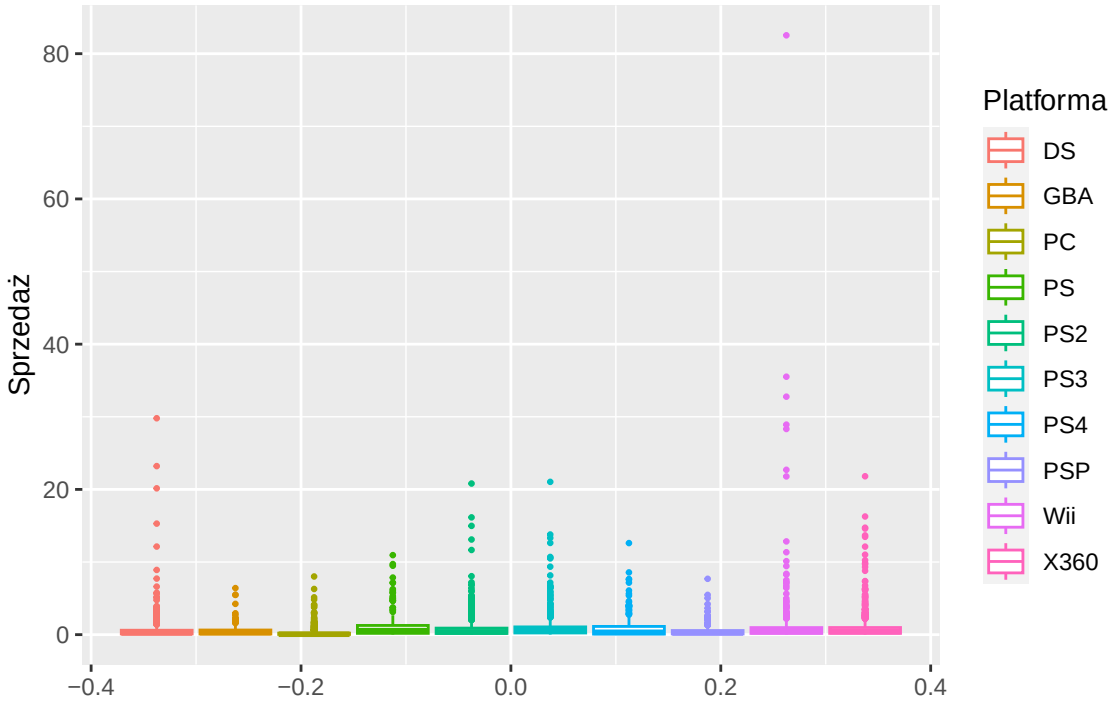
Jedyne platfromy jakie się nie pojawiają w tabelce to NS oraz GB, gdzie NS nie pojawia się w badanym zbiorze (najpewniej z powodu, że to stosunkowo nowa konsola i nie było jej na rynku w czasie tworzenia zbioru danych). Dodatkowo GBA (Game Boy Advanced) nie znajduję się wśród najlepiej sprzedających się konsol, ale znajduje się w stworzonej tabelce.

Dla znalezionych platform wykonano wykresy pudełkowe (Rys. 7) rozkładu sprzedaży na całym świecie. Postanowiono zostawić wartości odstające, gdyż uwzględniane są w zestawieniu ilości sprzedanych gier. Warto zwrócić uwagę na jedną obserwację, która jest najbardziej odstająca (82.53 miliony sprzedanych kopii). Jest to gra *Wii Sports*, której wysoka sprzedaż spowodowana jest tym, że była dodawana do prawie każdej zakupionej konsoli Wii.

	Platforma	Suma sprzedanych gier (w milionach)
1	PS2	1255.64
2	X360	971.63
3	PS3	939.43
4	Wii	908.13
5	DS	807.1
6	PS	730.68
7	GBA	318.5
8	PS4	314.23
9	PSP	294.3
10	PC	260.3

Tabela 2: Ranking dziesięciu platfrom z największą ilością sprzedanych gier.

[2]https://www.statista.com/statistics/1101872/unit-sales-video-game-consoles/

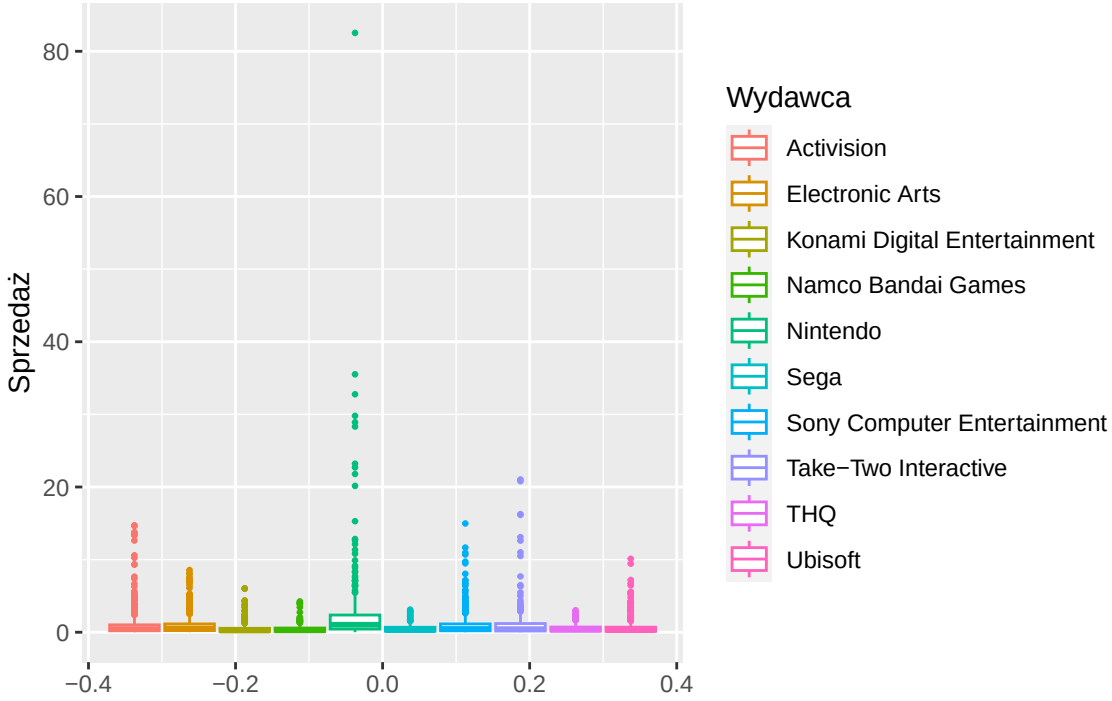


Rysunek 7: Wykresy pudełkowe sprzedaży gier na całym świecie ze względu na platforme

Podobną tabelkę wykonano dla wydawców, którzy sprzedali największą ilość gier (Tab. 3). Ponownie nie jest to zaskakujący wynik, ponieważ są to jedni z największych i najbardziej znanych wydawców gier komputerowych. Wykonano również wykresy pudełkowe (Rys. 8) rozkładu sprzedaży na całym świecie. Ponownie widać najbardziej odstającą wartość będącą grą *Wii Sport*, ale można również zauważyć, że firma Nintendo dominuje jeśli chodzi o najlepiej sprzedające się poszczególne gry. Jest to ciekawa obserwacja, ponieważ w przeciwieństwie do reszty wydawców z zestawieniu (poza Sony Computer Entertainment), Nindendo wydaje gry wyłącznie na swoje konsole, więc działają na mniejszym rynku niż reszta firm.

	Wydawca	Suma sprzedanych gier (w milionach)
1	Nintendo	1788.81
2	Electronic Arts	1116.96
3	Activision	731.16
4	Sony Computer Entertainment	606.48
5	Ubisoft	471.61
6	Take-Two Interactive	403.82
7	THQ	338.44
8	Konami Digital Entertainment	282.39
9	Sega	270.35
10	Namco Bandai Games	254.62

Tabela 3: Ranking dziesięciu wydawców z największą ilością sprzedanych gier.



Rysunek 8: Wykresy pudełkowe sprzedaży gier na całym świecie ze względu na wydawcę

Ostatnią częścią odpowiedzi na zadane pytanie badawcze, było znalezienie jednocześnie jacy wydawcy i na jakich platformach sprzedali największą ilość gier. Wykonano tabelką (Tab. 4) ze znalezionymi wynikami. Można zauważyć, że tylko trzech różnych wydawców znalazło się w zestawieniu. Zarówno Nintendo jak i Sony Computer Enternainment są jednocześnie producentami konsol, na których wydają gry (tzw. wydawcy first party), więc ich pojawienie się było spodziewane. Wyjątkiem jest Electronic Arts, którego pojawienie się jest zastanawiające i prowadzi do wniosków,

że jest jest wyjątkowo dużym wydawcą nie będącym jednocześnie producentem kosol. Warto również zauważyć, że w zestawieniu pojawiła się platforma, której nie było w poprzednich rankingach i jest to kosola NES (Nintendo Entertainmetn System), która jest jedną ze pierwszych globalnie dostępnych konsol.

	Wydawca	Platforma	Suma sprzedanych gier (w milionach)
1	Nintendo	Wii	386.25
2	Nintendo	DS	345.71
3	Electronic Arts	PS2	255.79
4	Nintendo	GB	230.09
5	Sony Computer Entertainment	PS	193.73
6	Nintendo	NES	183.97
7	Electronic Arts	X360	179.75
8	Sony Computer Entertainment	PS2	172.8
9	Nintendo	3DS	166.76
10	Electronic Arts	PS3	165.59

Tabela 4: Ranking dziesięciu wydawców i platform z największą ilością sprzedanych gier.

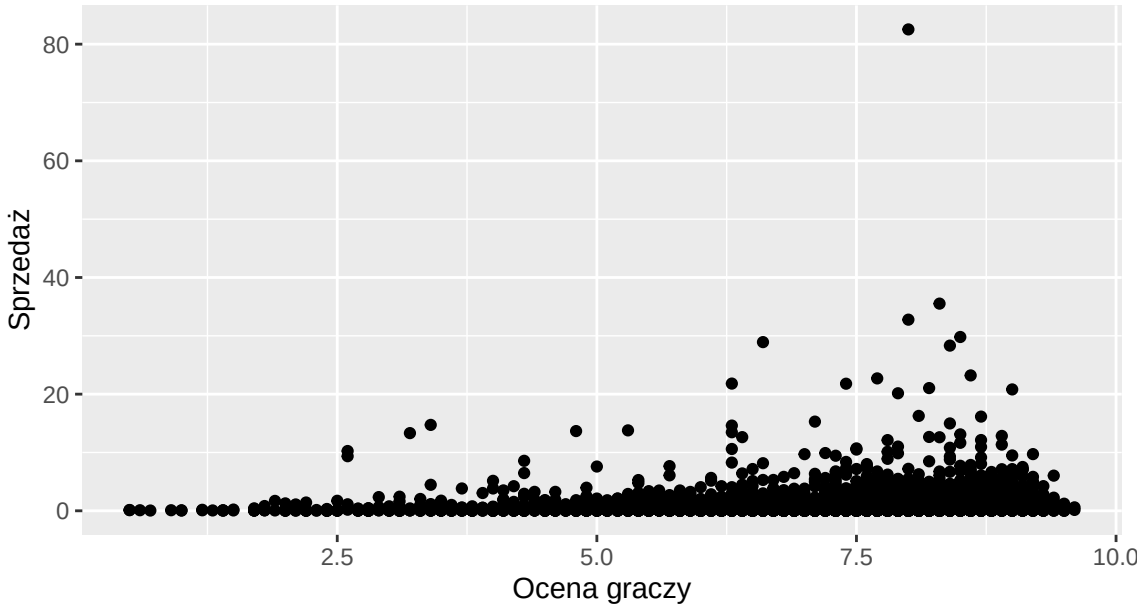
3.3 Czy ocena graczy i krytyków wpływa na sprzedaż?

Na wstępie analizę podzielono na dwie kategorie: ocenę graczy oraz ocenę krytyków. Dla każdej z nich stworzono wykres punktowy. Zarówno na wykresie 9 opisującym zależność sprzedaży produktu od oceny graczy, jak i na wykresie 10 w którym zmienną niezależną jest ocena krytyków widać bardzo dużą ilość gier, która ma znikomą sprzedaż, zarówno w części, w której oceny są wysokie, jak i w tej, w które są one niskie. Pomijając gry, których sprzedaż jest bardzo niska, możemy zauważyć pewną ciekawą zależność. Otóż większość gier, która wykazuje całkiem sporą sprzedaż, ma zwykle całkiem dobre oceny (powyżej połowy według skali). Mimo wszystko, aby sprawdzić zależność, postanowiono policzyć współczynnik korelacji Pearsona dla globalnych sprzedaży i oceny graczy, oraz globalnych sprzedaży i oceny krytyków. Współczynnik korelacji Pearsona liczy się ze wzoru:

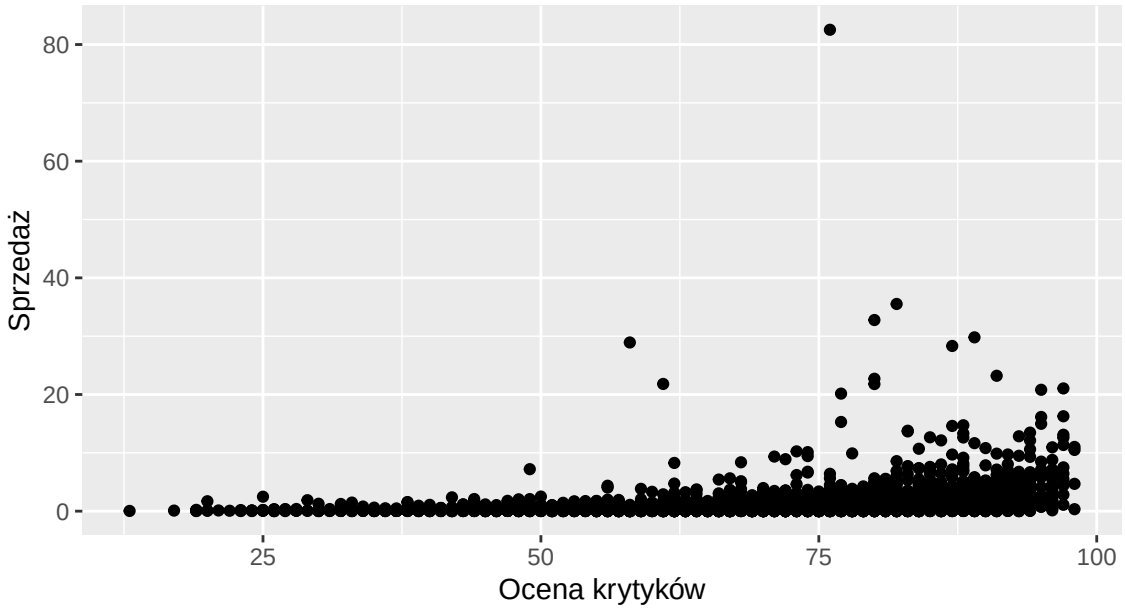
$$r = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2 \sum_{i=0}^{n}(y_i - \bar{y})^2}}$$

W powyższej części sprawozdania aby policzyć współczynnik korelacji Pearsona skorzystano w R z funkcji cor.

Otrzymane wartości współczynniku korelacji Pearsona zawierają się w przedziale $[-1,1]$. Im wartość bezwzględna ze współczynnika jest bliższa 1, tym silniej skorelowane dane, natomiast wartość bliska zero sugeruje, że dane wykazują słabą korelację. Otrzymano wartości kolejno $r = 0.2374$, oraz $r = 0.0886$, co wskazuje na to, że mimo iż oceny krytyków mają mniejsze znaczenie dla sprzedaży niż oceny graczy, to jednak analizowane dane nie wykazują większej korelacji.



Rysunek 9: zależność sprzedaży od oceny graczy



Rysunek 10: zależność sprzedaży od oceny krytykóq

4 Podsumowanie

W powyższym sprawozdaniu zostały przeanalizowanie 3 kwestie:

1. Czy sprzedaż gier rozkłada się podobnie dla różnych rynków?
2. Na jakich platformach sprzedaje się najwięcej gier oraz jacy wydawcy sprzedają ich najwięcej?
3. Czy ocena graczy i krytyków wpływa na sprzedaż (i potencjalnie jak)?

W pierwszej części przeanalizowane zostały statystyki rozkładu takie jak średnia, wariancja, kurtoza i skośność. Dodatkowo zostały porównane wykresy pudełkowe, histogramy będące estymatorami gęstości rozkładów, oraz dystrybuanty empiryczne dla rynków z różnych regionów. Głównym wnioskiem z tamtejszej analizy jest to, że rynki gier komputerowych w różnych częściach świata wykazują podobieństwo.

W drugiej części sporządzono listę platform oraz wydawców, którzy osiągnęli największą sprzedaż gier. W przypadku platformy bezkonkurencyjnym liderem jest konsola PlayStation 2 producenta Sony wydana 4 marca 2000 roku. Natomiast wydawcą, który odniósł największy sukces w łącznej sprzedaży gier jest japońska firma Nintendo.

Na końcu w sprawozdaniu wyliczone zostały współczynniki korelacji Pearsona oraz wywołane wykresy, które miały na celu pomóc odpowiedzieć na pytanie odnośnie wpływu oceny graczy oraz krytyków na sprzedaż gier. Zarówno ze współczynniku jak i z graficznego porównania widać, że takowa zależność jest bardzo słaba.