PROJEKT JĘZYK R

zbiór danych: Video Game Sales [1]
autor: Marcin Belicki
numer indeksu: 273417

1. Opis zbioru danych

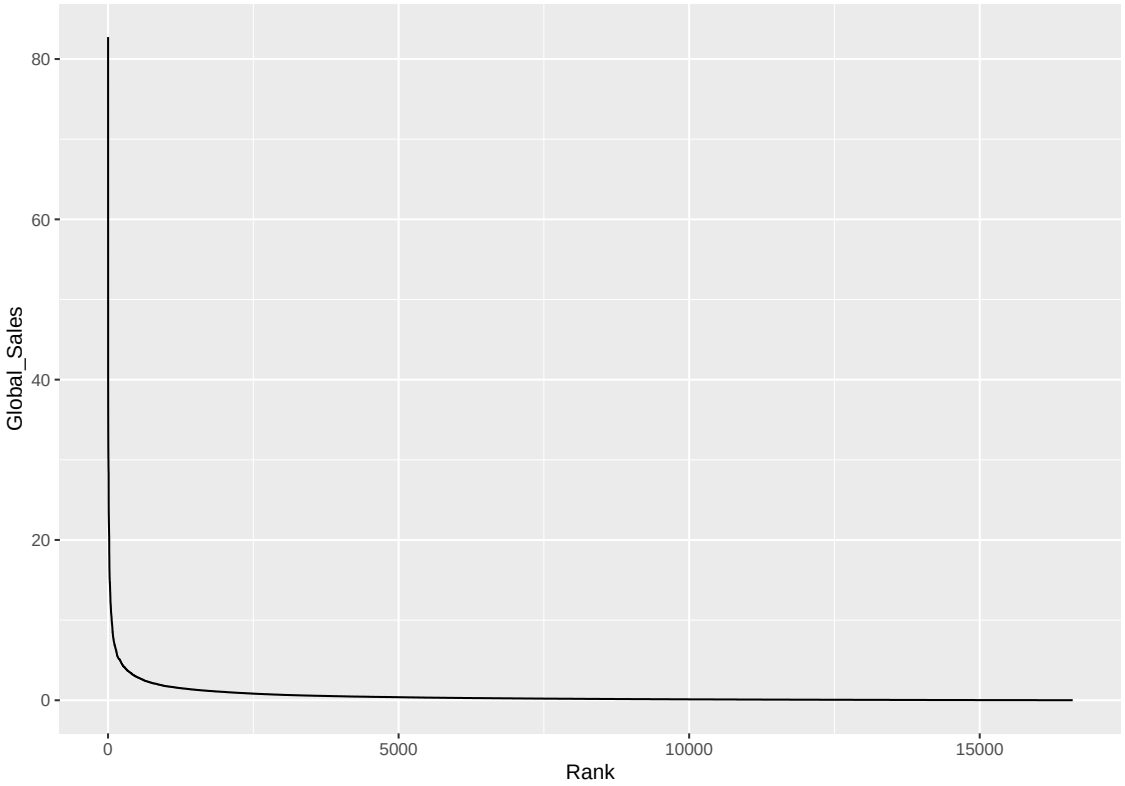
Zbiór danych zawiera listę gier ze sprzedażą przekraczającą 100 000 kopii. Dane zostały pozyskane ze strony vgchartz.com.
Opisy poszczególnych pól:

`Rank`	ranga gry pod względem sprzedaży ogółem
`Name`	nazwa gry
`Platform`	platforma, na którą gra została wydana (np. PC, PS4 itp.)
`Year`	rok wydania gry
`Genre`	gatunek gry
`Publisher`	wydawca gry
`NA_Sales`	liczba sprzedanych kopii w Ameryce Północnej (w milionach)
`EU_Sales`	liczba sprzedanych kopii w Europie (w milionach)
`JP_Sales`	liczba sprzedanych kopii w Japonii (w milionach)
`Other_Sales`	liczba sprzedanych kopii w reszcie świata (w milionach)
`Global_Sales`	liczba sprzedanych kopii na całym świecie (w milionach) - zmienna główna

Import danych został do środowiska R zostł wykonany za pomocą kodu:

```
1  data = read.csv("vgsales.csv", sep = ",")
2
3  list2env(data ,.GlobalEnv)
```

2. Ilustracja graficzna zmiennej głównej



Rysunek 1: Wykres zależności zmiennej głównej od rangi

Wykres został wygenerowany za pomocą kodu:

```
ggplot(
  data = data,
  aes(
    y = Global_Sales,
    x = Rank
  )
)+
  geom_line()
```

3. Obliczenie podstawowych statystyk opisowych zmiennej głównej

Obliczenia zostały wykonane za pomocą kodu:

```
description_statistics = function (data) {

  number_of_records = length(data)
  number_of_records_inverted = 1/number_of_records

  mean = mean(data)

  variance = var(data) * (number_of_records - 1) * number_of_records_
   inverted

  standard_deviation = variance ^ .5
  inside = (data - mean)/standard_deviation

  assymetry = sum(inside^3)*number_of_records_inverted
  curtosis  = sum(inside^4)*number_of_records_inverted - 3

  c(min, first_quartil, median, third_quartil, max) %<-% fivenum (data)

  interquartile_range = third_quartil - first_quartil

  quartil_deviation = interquartile_range * .5

  mean_quartil = (first_quartil + third_quartil) * .5

  quartil_assymetry_coefficient = (mean_quartil - median) / quartil_
   deviation

  quartil_variance_coefficient = quartil_deviation / median * 100

  list(
    number_of_records = number_of_records,
    mean = mean,
    standard_deviation = standard_deviation,

    assymetry = assymetry,
    curtosis = curtosis,

    min = min,
    first_quartil = first_quartil,
    median = median,
    third_quartil = third_quartil,
    max = max,

    interquartile_range = interquartile_range,
    quartil_deviation = quartil_deviation,
    mean_quartil = mean_quartil,
```

```
45
46      quartil_assymetry_coefficient = quartil_assymetry_coefficient,
47      quartil_variance_coefficient = quartil_variance_coefficient
48    )
49  }
50
51  Global_Sales_stats = description_statistics(Global_Sales)
52
53  Global_Sales_stats
```

Uzyskane wyniki

Liczba prób

$$n(\texttt{Global\_Sales}) = 16598$$

Średnia

$$mean(\texttt{Global\_Sales}) = 0.5374407$$

Odchylenie standardowe

$$\sigma(\texttt{Global\_Sales}) = 1.554981$$

Współczynnik asymetrii

$$A(\texttt{Global\_Sales}) = 17.39907$$

Kurtoza

$$K(\texttt{Global\_Sales}) = 603.7501$$

Wartość minimalna

$$min(\texttt{Global\_Sales}) = 0.01$$

Pierwszy kwartyl

$$Q_1(\texttt{Global\_Sales}) = 0.06$$

Mediana

$$M_e(\texttt{Global\_Sales}) = 0.17$$

Trzeci kwartyl

$$Q_3(\texttt{Global\_Sales}) = 0.47$$

Wartość maksymalna

$$max(\texttt{Global\_Sales}) = 82.74$$

Rozstęp między kwartylowy

$$IQR(\texttt{Global\_Sales}) = 0.41$$

Odchylenie ćwiartkowe

$$Q(\texttt{Global\_Sales}) = 0.205$$

Kwartyl średni

$$\bar{Q}(\texttt{Global\_Sales}) = 0.265$$

Kwartylny współczynnik asymetrii

$$A_k(\texttt{Global\_Sales}) = 0.4634146$$

Kwartylny współczynnik zmienności

$$V_k(\texttt{Global\_Sales}) = 120.5882$$

4. Dobór zmiennych (przynajmniej trzech), które mogą mieć wpływ na zmienną główną

Do analizy zostały wybrane 3 następujące zmienne.

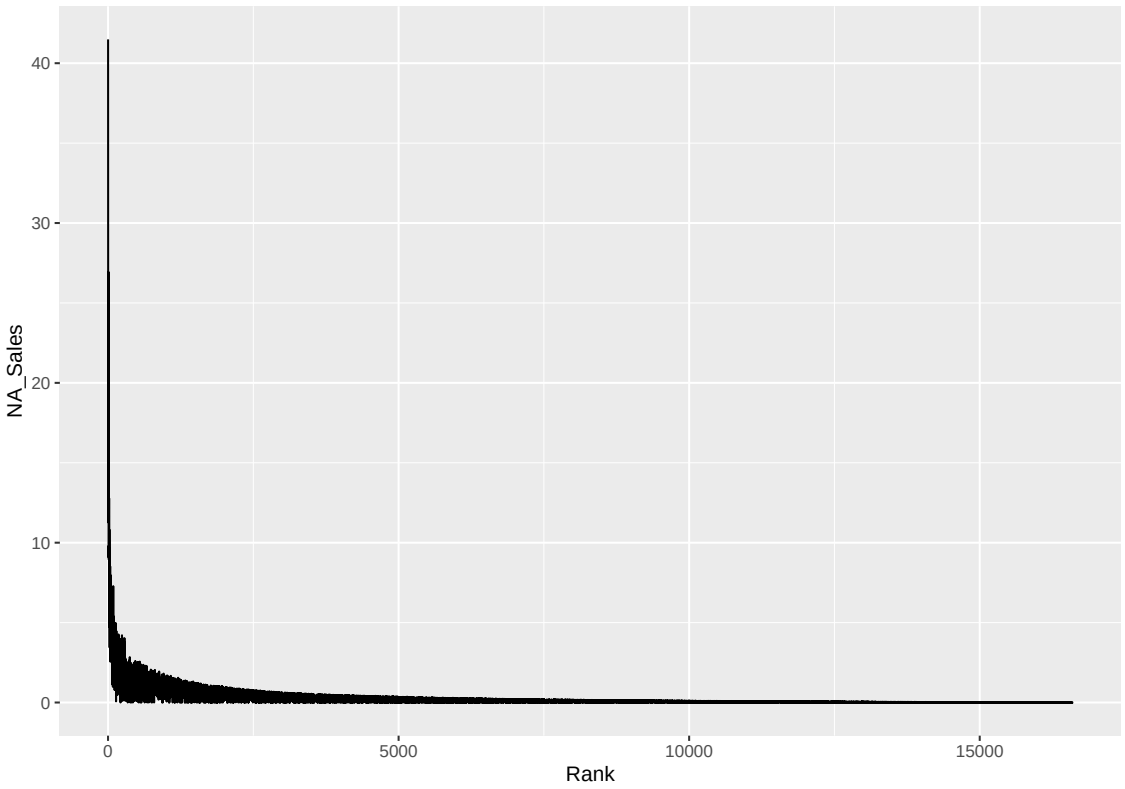
```
NA_Sales
EU_Sales
JP_Sales
```

Te zmienne mogą pozwolić ustalić który region ma najważniejszy wpływ na globalną sprzedaż.

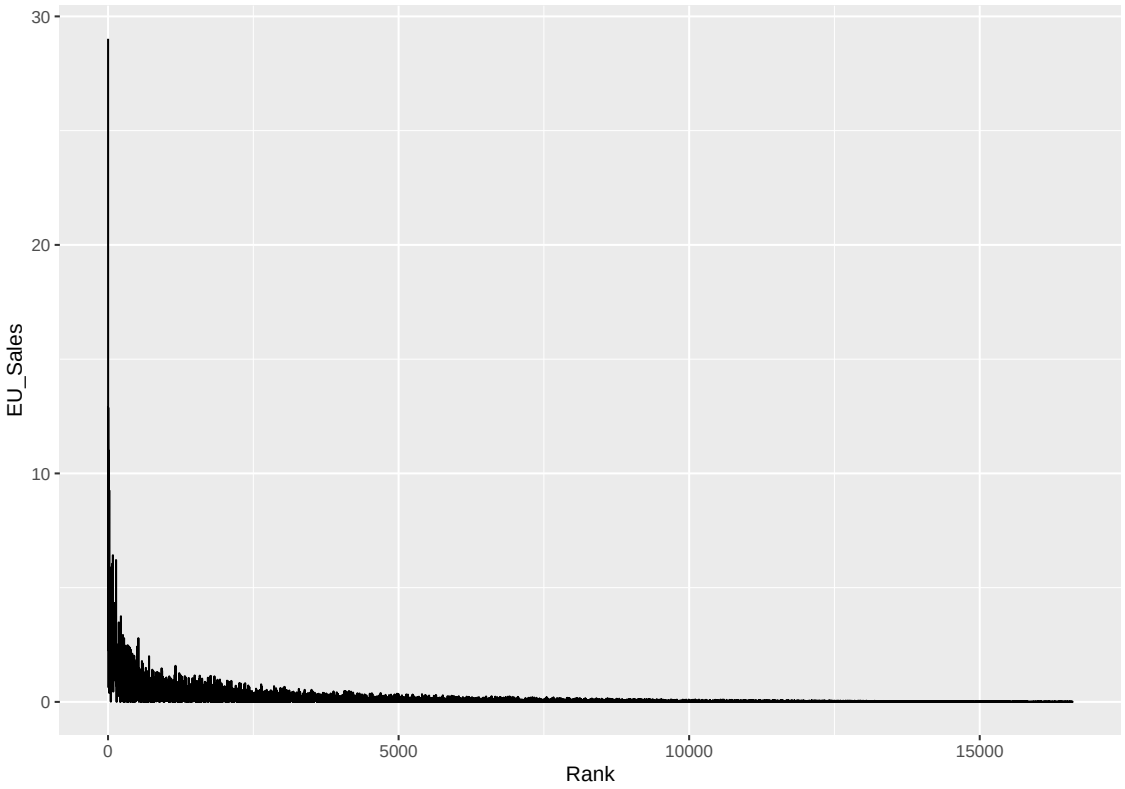
5. Graficzna prezentacja wybranych w poprzednim kroku zmiennych

Na poniższych wykresach zostały przedstawione zmienne (wybrane w poprzednim punkcie) w zależności od rangi wiodącej.
Wykresy zostały uzyskane za pomocą poniższego kodu:

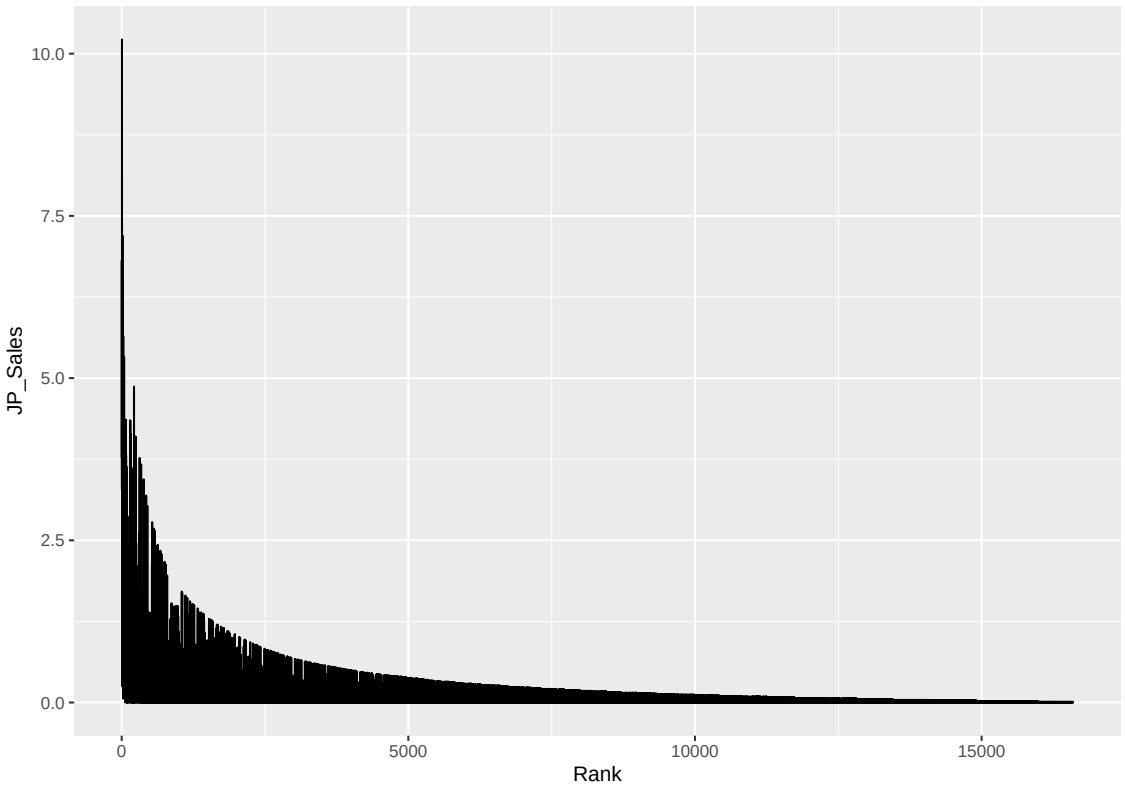
```
ggplot(
  data = data,
  aes(
    y = NA_Sales,
    x = Rank
  )
)+
  geom_line()

ggplot(
  data = data,
  aes(
    y = EU_Sales,
    x = Rank
  )
)+
  geom_line()

ggplot(
  data = data,
  aes(
    y = JP_Sales,
    x = Rank
  )
)+
  geom_line()
```



Rysunek 2: Wykres zależności zmiennej `NA_Sales` od rangi



Rysunek 3: Wykres zależności zmiennej `EU_Sales` od rangi



Rysunek 4: Wykres zależności zmiennej `JP_Sales` od rangi

6. Charakterystyka powyższych zmiennych (statystyki opisowe lub rozkłady liczebności, w zależności od klasy zmiennych)

Wyniki przedstawione w tabeli zostały uzyskane za pomocą poniższego kodu:

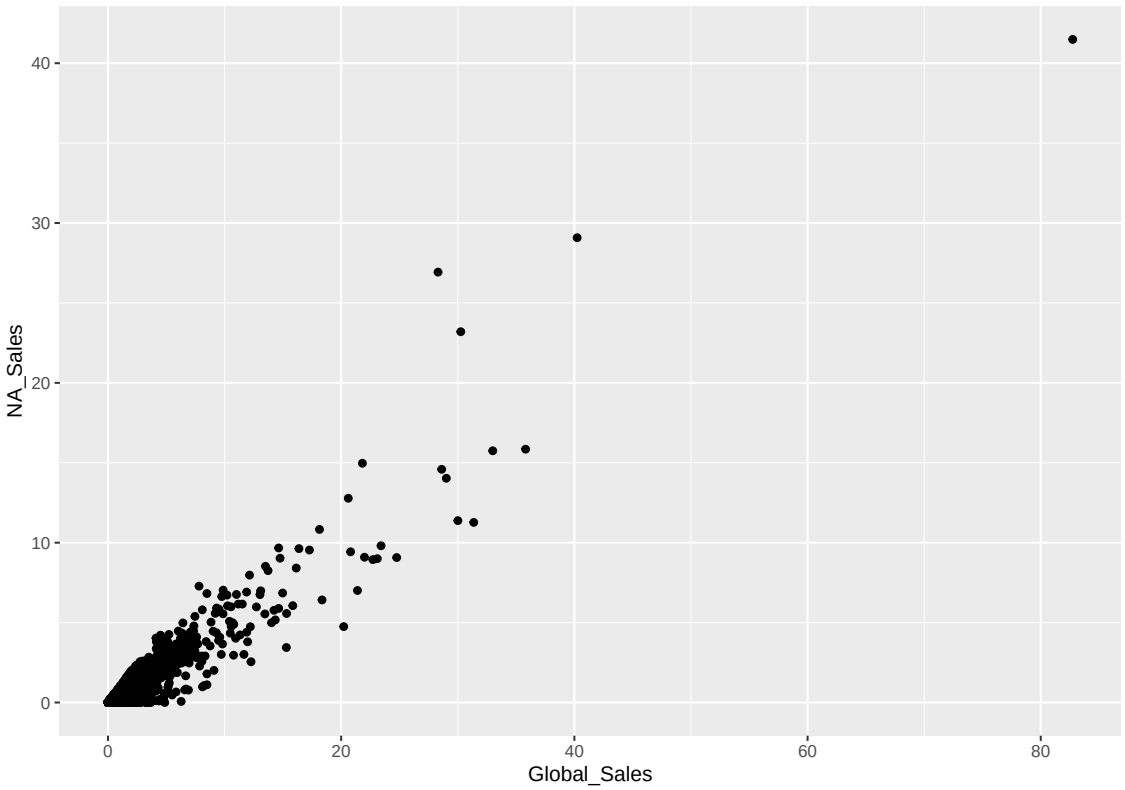
```
NA_Sales_stats = description_statistics(NA_Sales)

EU_Sales_stats = description_statistics(EU_Sales)

JP_Sales_stats = description_statistics(JP_Sales)

NA_Sales_stats

EU_Sales_stats

JP_Sales_stats
```

data	NA_Sales	EU_Sales	JP_Sales
n(data)	16598	16598	16598
$mean$(data)	0.2646674	0.146652	0.07778166
σ(data)	0.8166584	0.505336	0.3092813
A(data)	18.79793	18.87383	11.20545
K(data)	648.9344	755.7997	194.1751
min(data)	0	0	0
Q_1(data)	0	0	0
M_e(dat)	0.08	0.02	0
Q_3(data)	0.24	0.11	0.04
max(data)	41.49	29.02	10.22
IQR(data)	0.24	0.11	0.04
Q(data)	0.12	0.055	0.02
$\bar{Q}$(data)	0.12	0.055	0.02
A_k(data)	0.3333333	0.6363636	1
V_k(data)	150	275	Inf

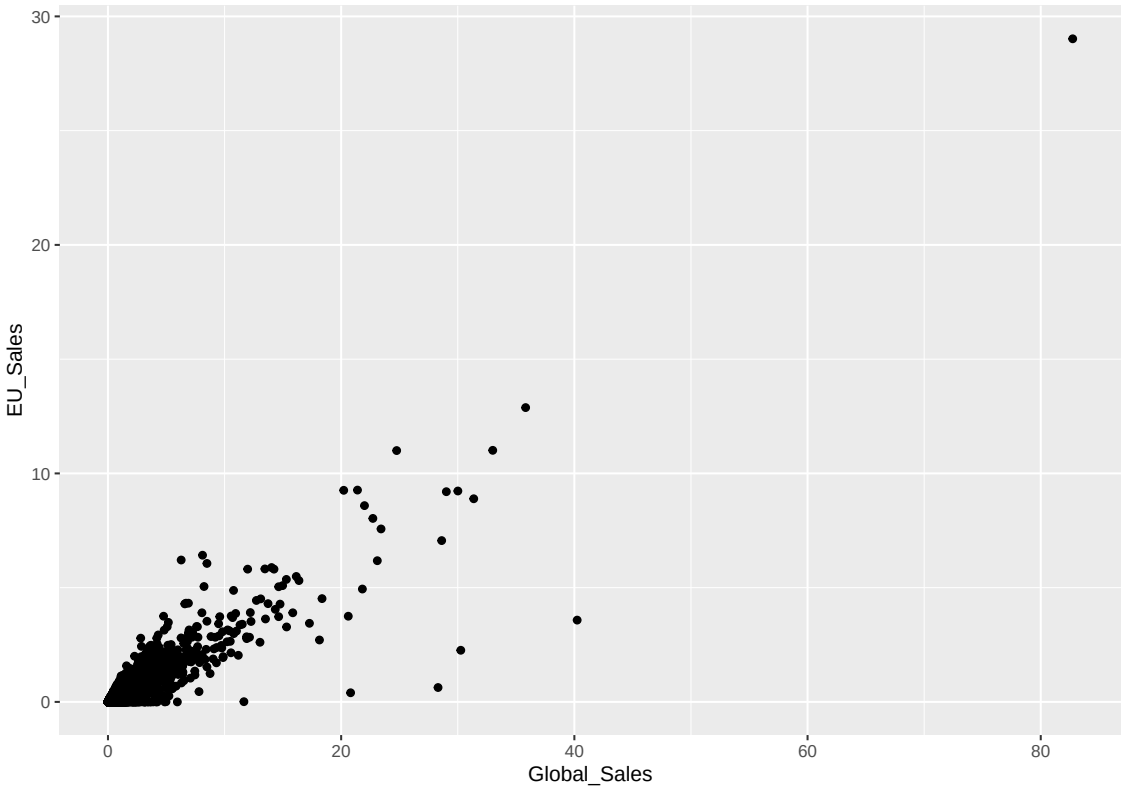
7. Graficzna prezentacja zależności

Wykresy zostały uzyskane za pomocą poniższego kodu:

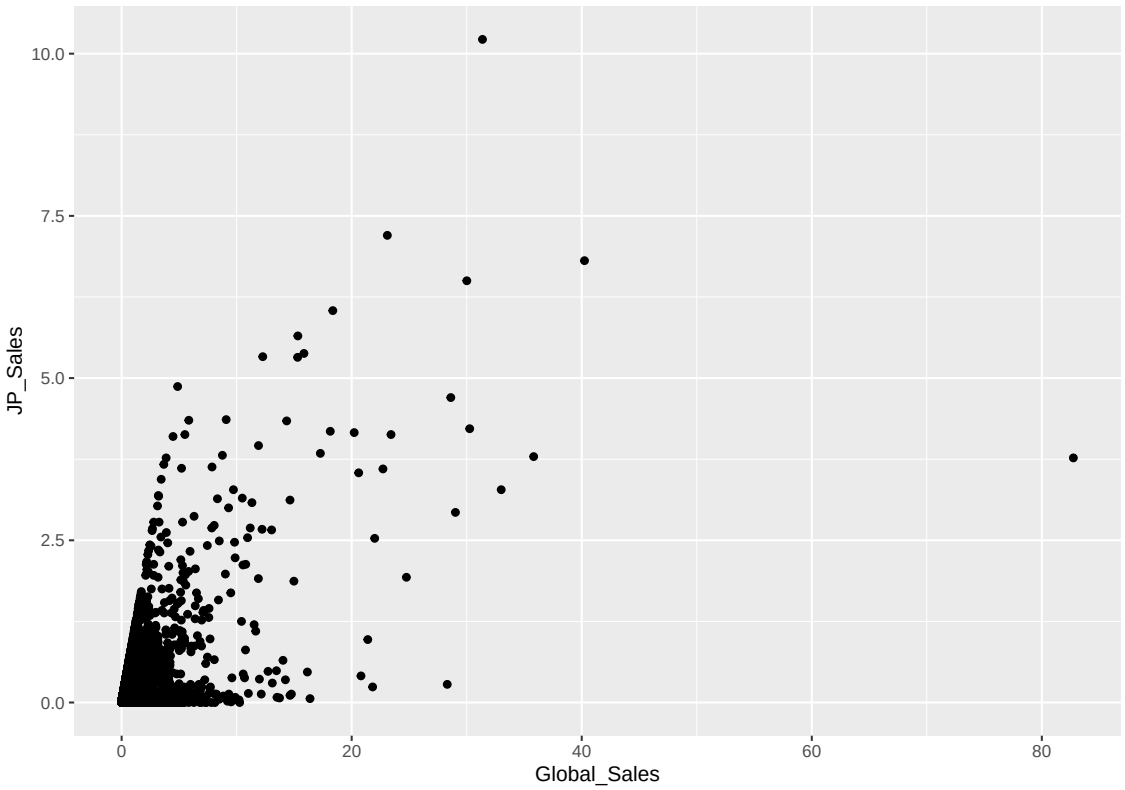
```
ggplot(
  data = data,
  aes(
    y = NA_Sales,
    x = Global_Sales
  )
)+
  geom_point()

ggplot(
  data = data,
  aes(
    y = EU_Sales,
    x = Global_Sales
  )
)+
  geom_point()

ggplot(
  data = data,
  aes(
    y = JP_Sales,
    x = Global_Sales
  )
)+
  geom_point()
```



Rysunek 5: Wykres zależności zmiennej `NA_Sales` od `Global_Sales`



Rysunek 6: Wykres zależności zmiennej `EU_Sales` od `Global_Sales`



Rysunek 7: Wykres zależności zmiennej `JP_Sales` od `Global_Sales`

8. Wykonanie odpowiedniego testu statystycznego, który potwierdzi lub odrzuci hipotezę o zależności

W celu sprawdzenia korelacji między wybranymi zmiennymi a zmienną główną przeprowadzony został test Spearmana za pomocą poniższego kodu.

```
spearman_test = function(data) {
  cor.test(Global_Sales, data, method = "spearman", exact = F)
}
spearman_test(NA_Sales)

spearman_test(EU_Sales)

spearman_test(JP_Sales)
```

Współczynniki dla poszczególnych przypadków przedstawiają się w następujący sposób:

$$\rho(\texttt{Global\_Sales}, \texttt{NA\_Sales}) = 0.7955717$$

$$\rho(\texttt{Global\_Sales}, \texttt{EU\_Sales}) = 0.6968457$$

$$\rho(\texttt{Global\_Sales}, \texttt{JP\_Sales}) = 0.1519311$$

9. Wnioski

Z wyników testu Spearmana wynika, że najbardziej skorelowaną zmienną ze zmienną główną jest `NA_Sales`. Wykres przedstawiony na rysunku 7 również zdaje się potwierdzać tę hipotezę (ze względu na widoczną niemalże liniową zależność). Ze statystyk opisanych w punkcie 6 można wywnioskować, że najmniej ze wszystkich gier w tym zestawieniu sprzedało się na rynku

japońskim. Wykres przedstawiony na rysunku 4 przedstawia najbardziej wahające się wartości ze wszystkich wybranych zmiennych. Może być to spowodowane niedostępnością niektórych z gier na rynku japońskim.
Wszystkie z przedstawionych zmiennych zdają się (w uśrednieniu) rosnąć hiperbolicznie względem rangi.

Bibliografia

[1] https://www.kaggle.com/datasets/gregorut/videogamesales
[dostęp: 13.06.2022 22:37 GMT+2]