Video games sales data set

2023-02-21

**DATASET DESCRIPTION:**

I collected this dataset through data.gov.in repacitory. It contains numerical as well as categorical datas. And this dataset contains a list of video games with their sales which is greater than 100,000 copies.

**FIELDS INCLUDE:**

**RANK** – Video game rankings on overall sales

**NAME** – Name of the games

**PLATFORM** – Platform of the video game released

**YEAR** – Year of the game release

**GENRE** – Genre of the game

**PUBLISHER** – This attributes contains who will publish that game

**NA\_SALES** – Video game sales in North America in millions

**JP\_SALES** – Video game sales in Japan in millions

**EU\_SALES** – Video game sales in Europe in millions

**OTHER\_SALES** – Sales in rest of the world in millions

**GLOBAL\_SALES** – Sales in worldwide in millions

**ASSUMPTION:**

In this dataset, the rank column is based on rank wise dataset which has been arranged. The correlation between the name and platform column is going to see that most of the video game is released on which platform. So this columns can be correlated. And then the correlation between names and year columns describes which year the game should be released. The Genre and the name columns describes what type the game is so they were correlated. NA\_Sales column describes Sales in North America and they will linked to genre because I could like to see what type of video game is mostly sold in north America and then the same process is done on the balance attributes. I can assumed that Japan country has most number of sales strategy while compared to other countries because japan country people bought most of the video games while compare to other countries

library(plyr)  
library(lattice)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readr)  
df = read\_csv("vgsales.csv")

## Rows: 16598 Columns: 11

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Name, Platform, Year, Genre, Publisher  
## dbl (6): Rank, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Global\_Sales  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

summary(df)

## Rank Name Platform Year   
## Min. : 1 Length:16598 Length:16598 Length:16598   
## 1st Qu.: 4151 Class :character Class :character Class :character   
## Median : 8300 Mode :character Mode :character Mode :character   
## Mean : 8301   
## 3rd Qu.:12450   
## Max. :16600   
## Genre Publisher NA\_Sales EU\_Sales   
## Length:16598 Length:16598 Min. : 0.0000 Min. : 0.0000   
## Class :character Class :character 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Mode :character Mode :character Median : 0.0800 Median : 0.0200   
## Mean : 0.2647 Mean : 0.1467   
## 3rd Qu.: 0.2400 3rd Qu.: 0.1100   
## Max. :41.4900 Max. :29.0200   
## JP\_Sales Other\_Sales Global\_Sales   
## Min. : 0.00000 Min. : 0.00000 Min. : 0.0100   
## 1st Qu.: 0.00000 1st Qu.: 0.00000 1st Qu.: 0.0600   
## Median : 0.00000 Median : 0.01000 Median : 0.1700   
## Mean : 0.07778 Mean : 0.04806 Mean : 0.5374   
## 3rd Qu.: 0.04000 3rd Qu.: 0.04000 3rd Qu.: 0.4700   
## Max. :10.22000 Max. :10.57000 Max. :82.7400

## HYPOTHESIS:

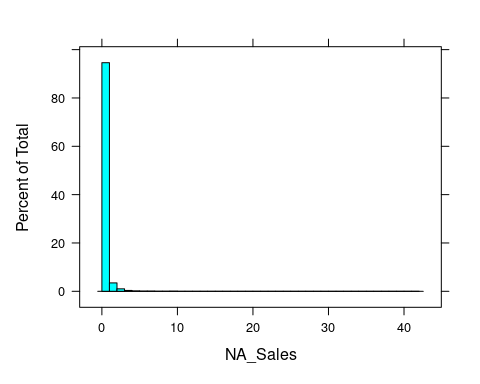
The analyzation of Sales by using histogram which is kept on the dataset seems like North America country has highly sold video games by compared with other countries which were kept in the dataset. I assumed that Japan contains more number of sales but in the histogram, North America contains more number of sales. So my assumption was wrong and then further I can analyzed that what type of video game was mostly sold in that particular country by using histogram. Then in that particular country which year the game was released and how the sales was going on. This analyzation is done by using histogram.

In boxplot, the analyzation is between the country sales and it shows the outlayers which is extreme values are different from the given dataset values. And then the comparison may be happen between the two numerical dataset that is sales between North America and Japan and then sales between Japan and Europe respectively.

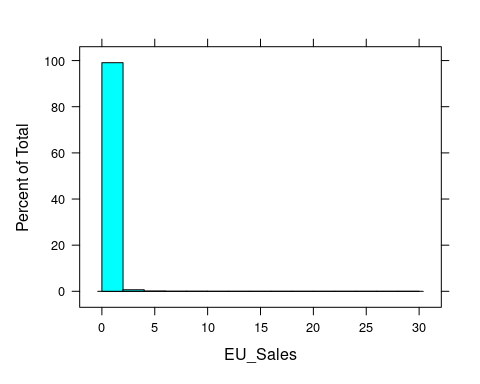
In scatterplot, the analyzation is done by using two numerical datasets.

## Histogram

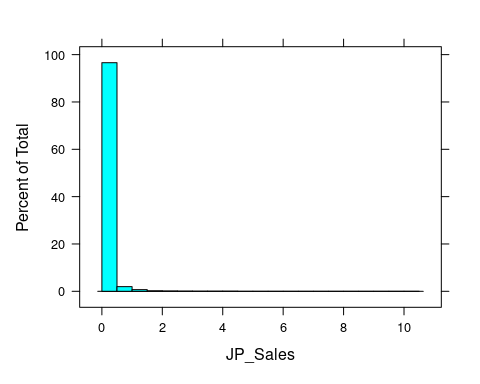
histogram(~NA\_Sales, data=df, breaks = 50)



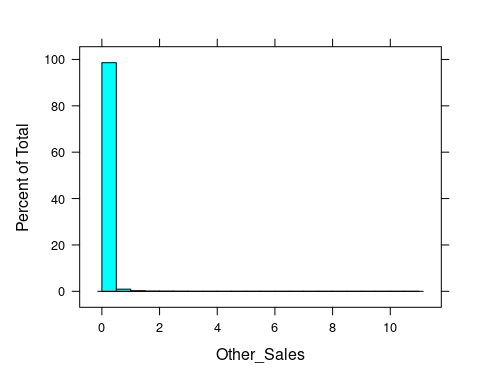
histogram(~EU\_Sales, data=df, breaks = 20)



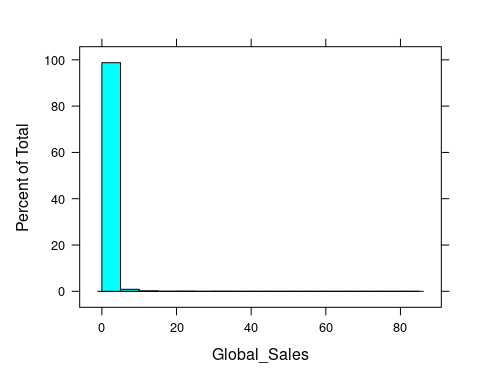
histogram(~JP\_Sales, data=df, breaks = 30)



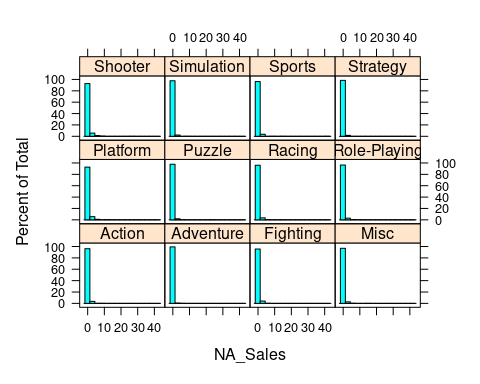
histogram(~Other\_Sales, data=df, breaks = 20)



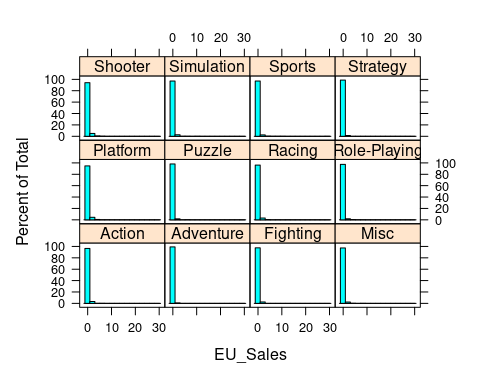
histogram(~Global\_Sales, data=df, breaks = 20)



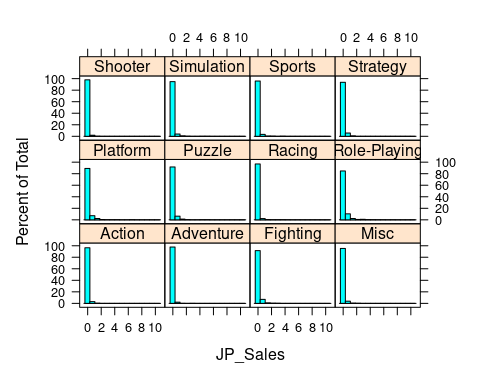
#Subset   
histogram(~NA\_Sales|Genre, data=df)



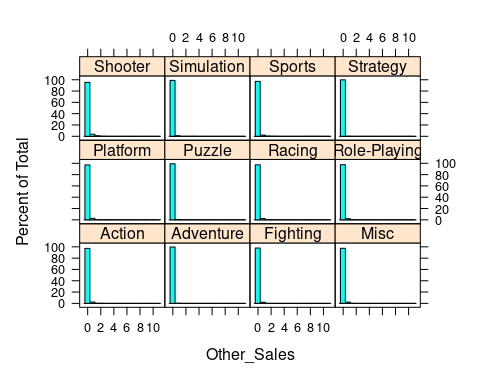
histogram(~EU\_Sales|Genre, data=df)



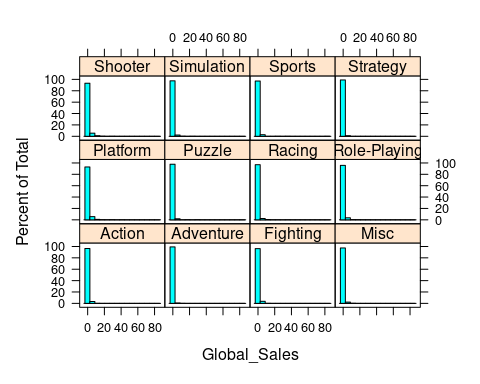
histogram(~JP\_Sales|Genre, data=df)



histogram(~Other\_Sales|Genre, data=df)



histogram(~Global\_Sales|Genre, data=df)

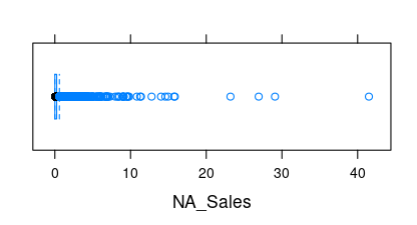


## Boxplot

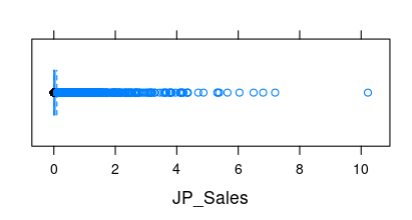
#Boxplot  
library(ggplot2)  
input<-df[,c('NA\_Sales','Year')]  
print(head(input))

## # A tibble: 6 × 2  
## NA\_Sales Year   
## <dbl> <chr>  
## 1 41.5 2006   
## 2 29.1 1985   
## 3 15.8 2008   
## 4 15.8 2009   
## 5 11.3 1996   
## 6 23.2 1989

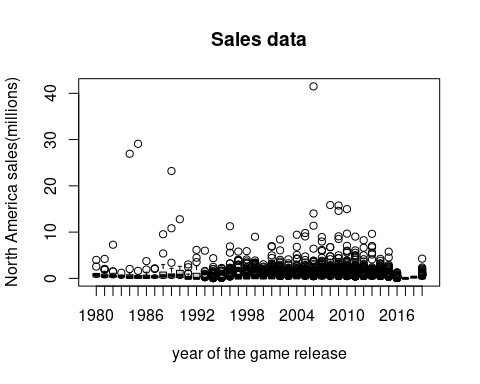
#Plot the chart  
bwplot(~NA\_Sales,data=vgsales)



Bwplot(~JP\_Sales,data=vgsales)



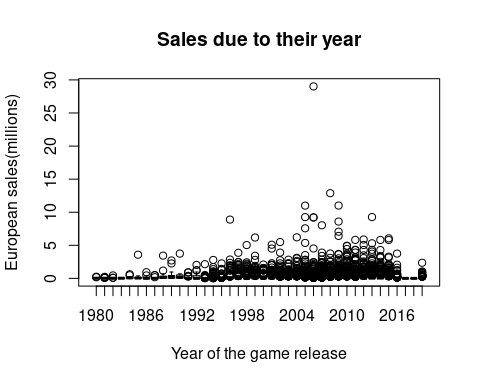
boxplot(NA\_Sales ~ Year, data=df, xlab = "year of the game release", ylab = "North America sales(millions)", main = "Sales data")



input<-df[,c('EU\_Sales','Year')]  
print(head(input))

## # A tibble: 6 × 2  
## EU\_Sales Year   
## <dbl> <chr>  
## 1 29.0 2006   
## 2 3.58 1985   
## 3 12.9 2008   
## 4 11.0 2009   
## 5 8.89 1996   
## 6 2.26 1989

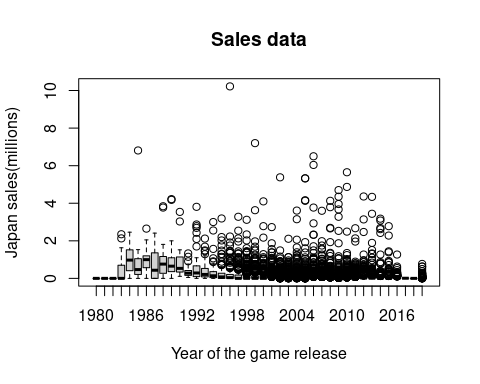
boxplot(EU\_Sales ~ Year, data=df, xlab = "Year of the game release", ylab = "European sales(millions)", main="Sales due to their year")



input<-df[,c('JP\_Sales','Year')]  
print(head(input))

## # A tibble: 6 × 2  
## JP\_Sales Year   
## <dbl> <chr>  
## 1 3.77 2006   
## 2 6.81 1985   
## 3 3.79 2008   
## 4 3.28 2009   
## 5 10.2 1996   
## 6 4.22 1989

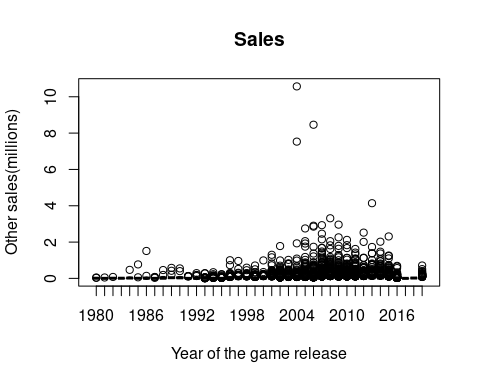
boxplot(JP\_Sales ~ Year, data=df, xlab = "Year of the game release", ylab = "Japan sales(millions)", main="Sales data")



input<-df[,c('Other\_Sales','Year')]  
print(head(input))

## # A tibble: 6 × 2  
## Other\_Sales Year   
## <dbl> <chr>  
## 1 8.46 2006   
## 2 0.77 1985   
## 3 3.31 2008   
## 4 2.96 2009   
## 5 1 1996   
## 6 0.58 1989

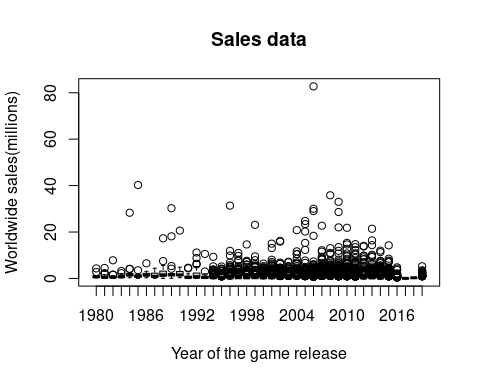
boxplot(Other\_Sales ~ Year, data=df, xlab = "Year of the game release", ylab = "Other sales(millions)", main="Sales")



input<-df[,c('Global\_Sales','Year')]  
print(head(input))

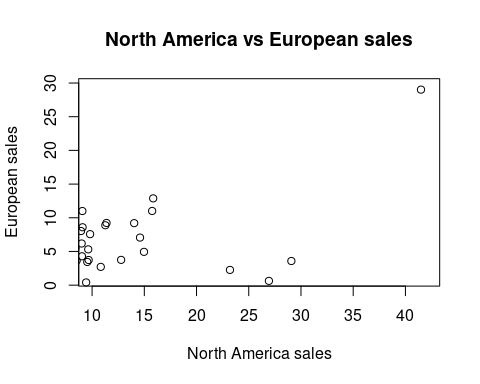
## # A tibble: 6 × 2  
## Global\_Sales Year   
## <dbl> <chr>  
## 1 82.7 2006   
## 2 40.2 1985   
## 3 35.8 2008   
## 4 33 2009   
## 5 31.4 1996   
## 6 30.3 1989

boxplot(Global\_Sales ~ Year, data=df, xlab = "Year of the game release", ylab = "Worldwide sales(millions)", main="Sales data")

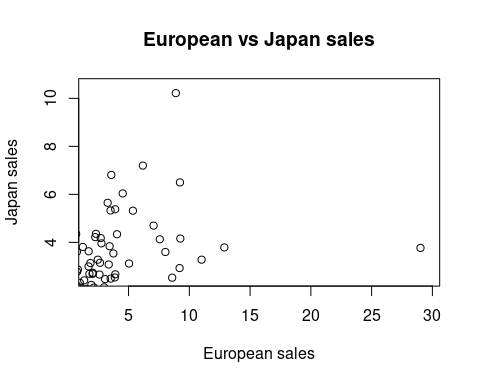


## Scatterplot

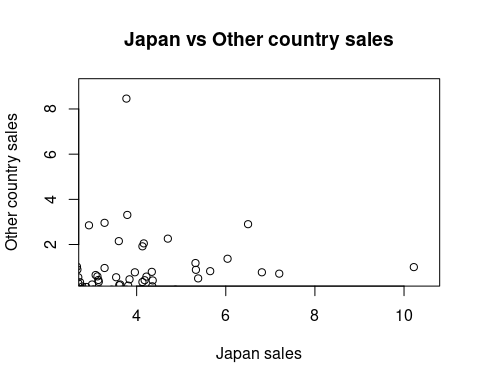
## # A tibble: 6 × 2  
## NA\_Sales EU\_Sales  
## <dbl> <dbl>  
## 1 41.5 29.0   
## 2 29.1 3.58  
## 3 15.8 12.9   
## 4 15.8 11.0   
## 5 11.3 8.89  
## 6 23.2 2.26



## # A tibble: 6 × 2  
## EU\_Sales JP\_Sales  
## <dbl> <dbl>  
## 1 29.0 3.77  
## 2 3.58 6.81  
## 3 12.9 3.79  
## 4 11.0 3.28  
## 5 8.89 10.2   
## 6 2.26 4.22



## # A tibble: 6 × 2  
## JP\_Sales Other\_Sales  
## <dbl> <dbl>  
## 1 3.77 8.46  
## 2 6.81 0.77  
## 3 3.79 3.31  
## 4 3.28 2.96  
## 5 10.2 1   
## 6 4.22 0.58



**INFERENCE:**

**#Histogram:**

The visualization of North America sales shows that the video game sales ranges between 80 to 100. And then the another bar shows that the continuous sales is going on in that country and it ranges between 0 to 20. This shows that video game is highly sold in North America.

The visualization of Europe sales shows that the video game ranges from 90 to 100 and then the other country shows that the same sales range they will be visualized.

The comparison of sales country which is kept in the dataset and video game genre shows that the ranges between 80 and 100 which is visualized. In that genre, The shooter game is highly sold by comparison of other genre game and then the adventure game is less sold by the comparison of other game genre.

**#Boxplot:**

The visualization of Japan sales shows that outlayers which is ranges above 10 it shows that highly extreme value found in that dataset.

The comparison of North America sales and which year the game was released shows that outlayers and it ranges above 40

The comparison of Europe sales and year column shows that the outlayers is formed and the extreme outlayer range above 30

**#Scatterplot:**

The comparison is done by using two numerical dataset which is North America sales and then Europe sales shows that highly correlated in the range between 10 to15

The another comparison is done by using Europe and Japan sales which also highly correlated and in the range between 0 to 10

The columns contains Other sales and Worldwide sales which is not highly correlated.

**INSIGHTS:**

**#Histogram:**

The histogram visualization shows that North America country has bought so many video games because that country shows that high video game sales.

The comparison of video game sales country and then video game genre shows that the shooter game sold many times while compared with other genre game because the youth generation mostly liked that shooter game which may be gun shooting or other weapon shooting like that.

**#Boxplot:**

The comparison of two numerical data is done by using boxplot.

It shows that outlayers is formed which is highly extreme values are found in that dataset

**#Scatterplot:**

The relation between two numerical attributes which is Sales between two countries is analyzed by using scatterplot.

It shows that the values are highly correlated that is the closer the data which is formed by putting the straight line when plotted it shows that the variables are stronger together, that’s why the data’s are highly correlated.