Exploring the Video Game Sales with Ratings dataset

Rodrigo N. M. da Silva

December 2017

1 Introduction

The video games industry is in constant growing and the companies investment are bigger year after year. Hundreds of games are launched by year and them are evaluated by several media specifically segments. Metacritic is an important tool that a that aggregates reviews of games. For each product, the scores from each review are averaged (weighted average). So the metacritic score is a valuable reference to measure the quality of the games. In this work we will explore a dataset with a large amount of games with its metacritic score and number of sales. The dataset is available in the following website:

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

2 Goal

We want to understand the trends in video games area, analyzing its sales around the world and its critic scores. Our objective is uncover the bestsellers considering the metacritic score, sales region and game genre. To achieve this we defined a set of questions which will be answered along this work. We will use RStudio, manipulating well-known verbs of the R programming language, as well as specific libraries of tidyverse metapackage (readr, tidyr, dplyr) to create a report using Literate Programming.

3 Questions

1. There is linear correlation between critic score and number of global sales? If yes, are the games with the highest critic score the most sold?
2. Are the games sold equally in NA, EU?

```
library(readr);
library(dplyr);
```

```
##
## Attaching package: 'dplyr'

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

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

```
library(magrittr);
library(ggplot2);
library(nortest);
library(knitr)
URL <- "https://github.com/rodrimoni/lps/raw/master/FinalProject/Video_Games_Sales_as_at_22_Dec_2
```

```
df <- read_delim(URL, delim=",");
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Platform = col_character(),
##   Year_of_Release = col_character(),
##   Genre = col_character(),
##   Publisher = col_character(),
##   NA_Sales = col_double(),
##   EU_Sales = col_double(),
##   JP_Sales = col_double(),
##   Other_Sales = col_double(),
##   Global_Sales = col_double(),
##   Critic_Score = col_integer(),
##   Critic_Count = col_integer(),
##   User_Score = col_character(),
##   User_Count = col_integer(),
##   Developer = col_character(),
##   Rating = col_character()
## )
```

4 Semantics of the Data

In this analysis the following variables were used:

1. Name: Variable of type "chr" that identifies the name of the game.
2. Year_of_Release: Variable of type "char" that identifies the year of the launch and of the game.
3. NA_Sales: Variable of type "double" that identifies the number of the sales in North America in millions of units.
4. EU_Sales: Variable of type "double" that identifies the number of the sales in European Union in millions of units.
5. Global_Sales: Variable of type "double" that identifies the number of the sales in the whole World in millions of units.
6. Critic_Score: variable of type "int" that identifies the aggregate score compiled by Metacritic staff.

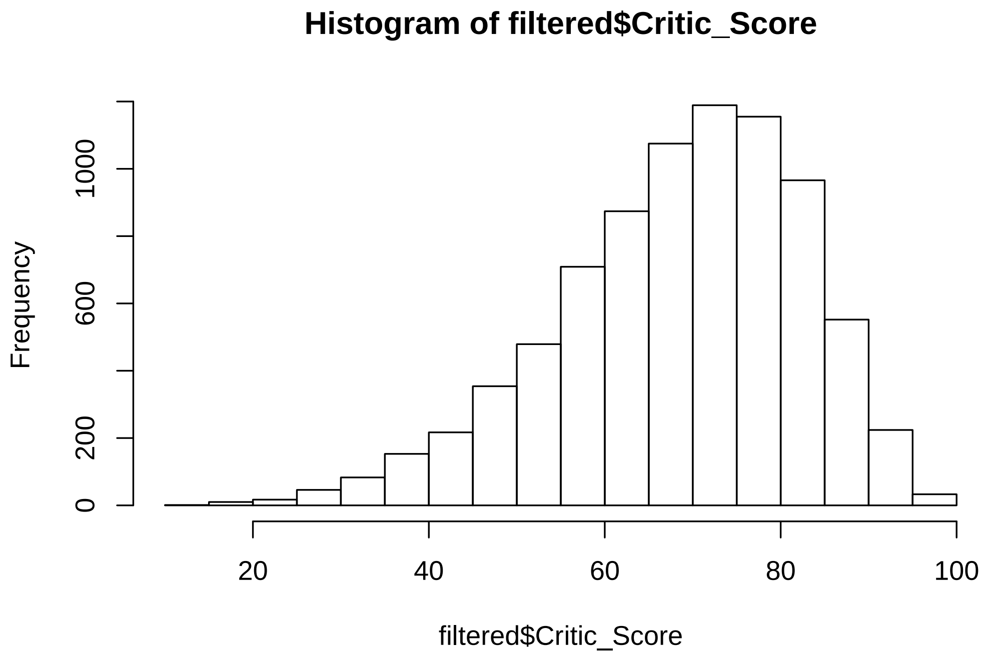
5 There is linear correlation between critic score and number of global sales? If yes, are the games with the highest critic score the most sold?

To answer this question we need to explore the following variables: Critic Score and Global Sales.

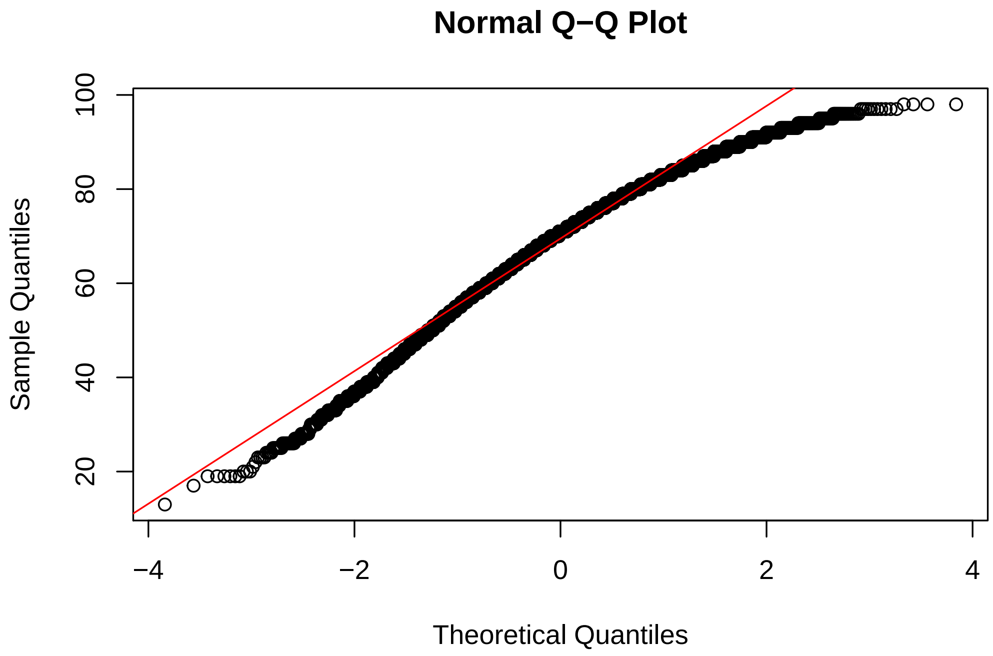
A well-know method to do the correlation analysis between variables is the Pearson Correlation, which measures a linear dependence between two variables (x and y). It's also known as a parametric correlation test because it depends to the distribution of the data. However, it can be used only when x and y are from normal distribution. So, the first step is check if Critic_Score and Global_Sales are normal distributed. We start plotting the histogram and the Quantile-Quantile plot to have an data overview.

```r
filtered <- df %>% select(Critic_Score, Global_Sales, Year_of_Release) %>% filter (!(is.na(Critic

hist(filtered$Critic_Score)
```



```r
qqnorm(filtered$Critic_Score)
qqline(filtered$Critic_Score, col = 2)
```



```r
ad.test(filtered$Critic_Score)
```

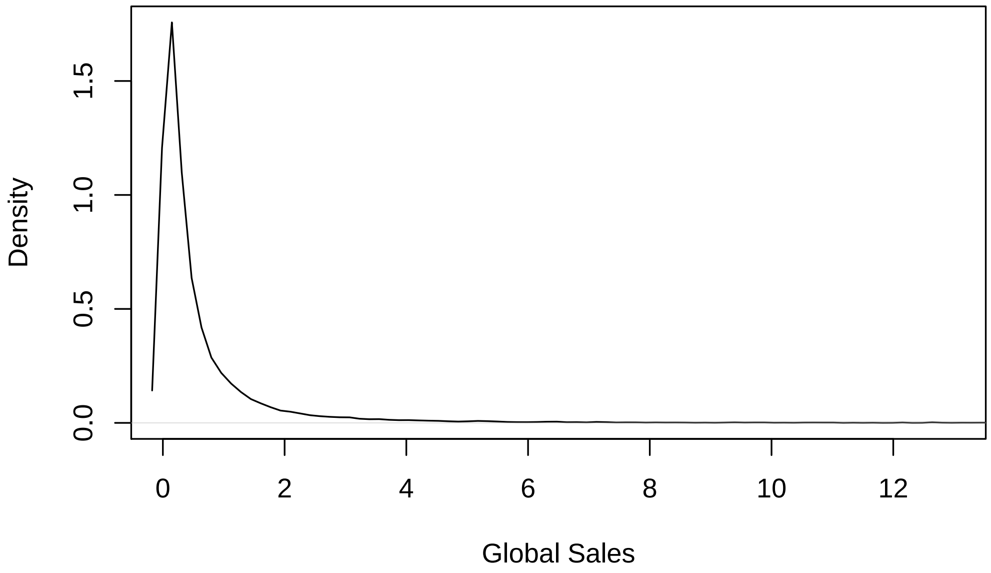
```
##
##  Anderson-Darling normality test
##
## data:  filtered$Critic_Score
## A = 51.695, p-value < 2.2e-16
```

Analyzing the histogram the variable seems to be a normal distribution, but when we check the Q-Q Plot with support of the QQ Line that sets a drawn line in the first and third quartiles gives to us a robust approach for estimating the parameters of the normal distribution. So, departures from the line (except in the tails) are indicative of a lack of normality as we can see in the plot.

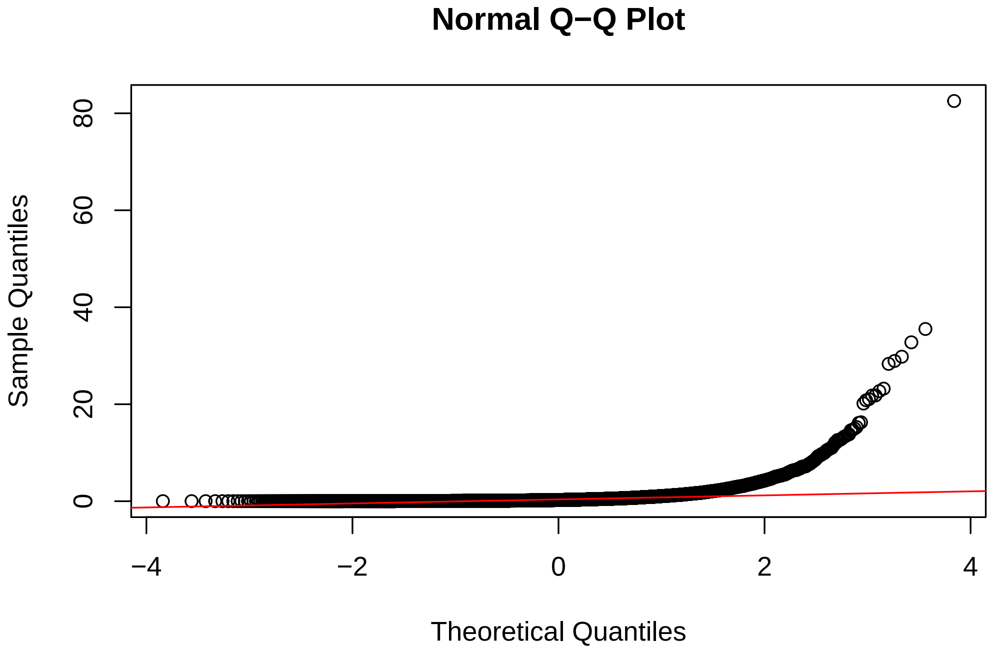
To confirm we run a Anderson-Darling test to check if is a normal distribution and then we get as a result a p-value < 0.01 indicatting that is not a normal distribution.

Checking our second variable, Global Sales, we get:

```
plot(density(filtered$Global_Sales), main="", xlab="Global Sales", xlim=c(0,13))
```



```
qqnorm(filtered$Global_Sales)
qqline(filtered$Global_Sales, col = 2)
```



Plotting the Density of Global_Sales and the Q-Q Plot we clearly see that distribuition is not normal.

Then we have to use a different method to analyze the correlation between sales and critic scores. Kendall method is indicated in this case, which are a rank-based correlation coefficient (non-parametric).

```
kendallTest <-cor.test(filtered$Critic_Score, filtered$Global_Sales,  method = "kendall")
kendallTest
```

```
##
##  Kendall's rank correlation tau
##
## data:  filtered$Critic_Score and filtered$Global_Sales
## z = 36.563, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2749158
```

It estimated that the linear correlation between the Critic_Score and Global_Sales is 0.2749, accordingly the Kendall's test. This correlation is positive and significative ($p < 0.01$) indicating that sales grow linearly as the critic score.

6 Are the games sold equally in NA, EU?

Now we want to analyze if the sales are equally distributed in the following regions: 1. North America (NA) 2. European Union (EU)

The main market of the games lays on these regions, it is where the most games are created and distributed.

Below we plotted the percentage of the sales by year, to analyze which region sold more.

```
a <- df %>% group_by(Year_of_Release) %>% summarize(glboal = sum(Global_Sales),na = sum(NA_Sales)

northAmerica <- data.frame(Year = as.numeric(a$Year_of_Release),'Region' = 'North America', 'Regi
```

```
## Warning in data.frame(Year = as.numeric(a$Year_of_Release), Region = "North
## America", : NAs introduzidos por coerção
```

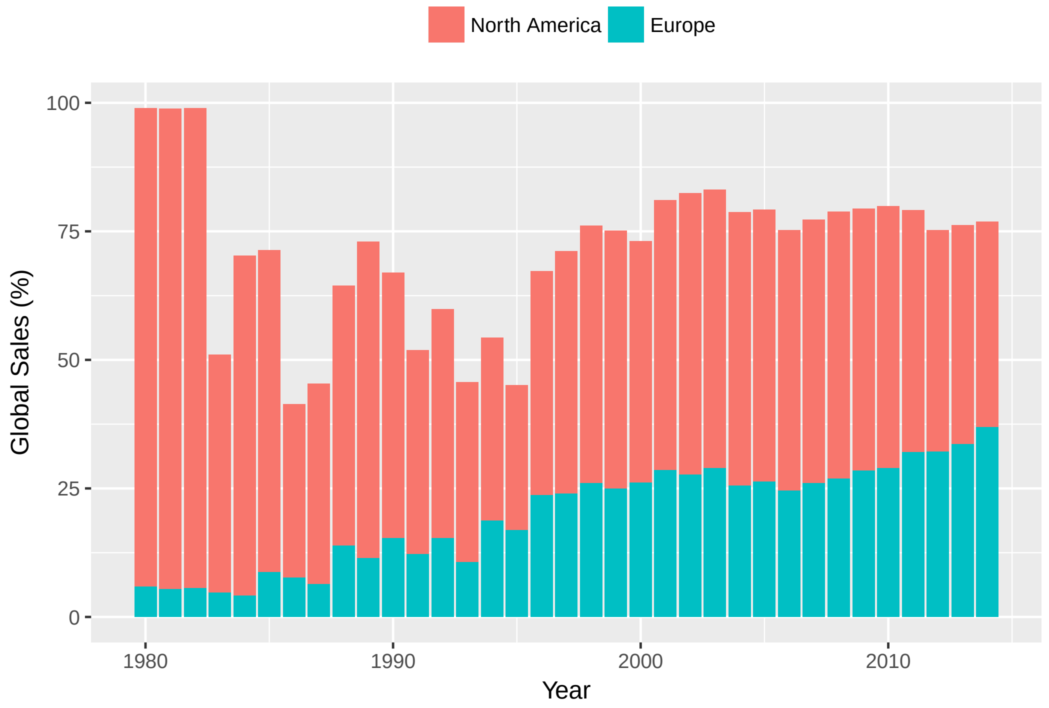
```
europe <- data.frame(Year = as.numeric(a$Year_of_Release), 'Region' = 'Europe', 'Region Sales' =
```

```
## Warning in data.frame(Year = as.numeric(a$Year_of_Release), Region =
## "Europe", : NAs introduzidos por coerção
```

```
all <- rbind(northAmerica, europe)

all <- all %>% filter (Year < 2015)

p4 <- ggplot() + geom_bar(aes(y = Region.Sales, x = Year, fill = Region), data = all, stat="ident
p4
```



In the plot we can see that North America dominated the scenario for most part of the time. In the first years of the video game era Europe had significance in the market, most of the sales were made in NA. However, in last 20 years Europe sales are growing considerably and in the last 5 years are getting close of North American Sales. We can see in the plot that the sales are not equally distributed.

So to validate our answer we will test the normality of the data using Anderson-Darling in order to use the T-test for the our two independent samples, NA_Sales and EU_Sales:

```
ad.test(df$NA_Sales)
```

```
##
##  Anderson-Darling normality test
##
## data:  df$NA_Sales
## A = 3281.3, p-value < 2.2e-16
```

```
ad.test(df$EU_Sales)
```

```
##
##  Anderson-Darling normality test
##
## data:  df$EU_Sales
## A = 3615.8, p-value < 2.2e-16
```

With the result of both tests we reject the null hypothesis because the p-value is not significative ($p < 0.01$), therefore the samples did not present normality.

Then, we will use the non-parametric test U of Wilcoxon to two independent medians, analogue to parametric T-test to two independent samples.

```
summary (df$NA_Sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0800  0.2633  0.2400 41.3600
```

```
summary (df$EU_Sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.020   0.145   0.110  28.960
```

The research question that we must do is: “The median of North America sales (median = 0.08), observed in first sample, it is significantly bigger than the median of European Sales (median = 0.02), observed in the second sample?”

```
wilcox.test(df$NA_Sales, df$EU_Sales)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df$NA_Sales and df$EU_Sales
## W = 169580000, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

With this result we can estimate that the median of the sales is 0.08 millions to North America and 0.02 millions to Europe. This values are significantly differents ($p < 0.01$), therefore the sales of games in North America should be greater than in Europe.

The data is summarized in the following table:

```
resume <- matrix(c("0.08 (0;0.24)","0.02(0;0.11)", "2.2e-16"), ncol=3,byrow=TRUE)
colnames(resume) <- c("North America", "Europe", "p-value")
rownames(resume) <- c("Sales in Millions")
resume <- as.table(resume)
resume
```

```
##                   North America Europe       p-value
## Sales in Millions 0.08 (0;0.24) 0.02(0;0.11) 2.2e-16
```

This data are presented by medians (quartil 25% - quartil 75%)