Final Project Milestone 1

R Output.R

Student Name: Zihan Ma

Class NameALY6010: Probability Theory and Introductory Statistics

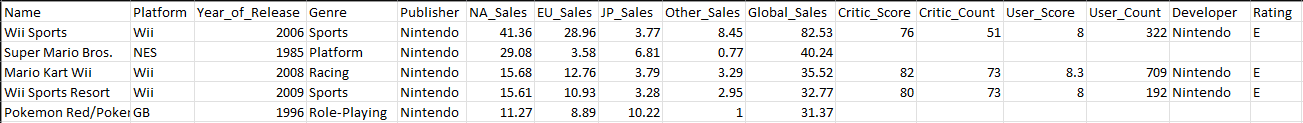
Instructor: Tom Breur

03/05/2023

Introduction:

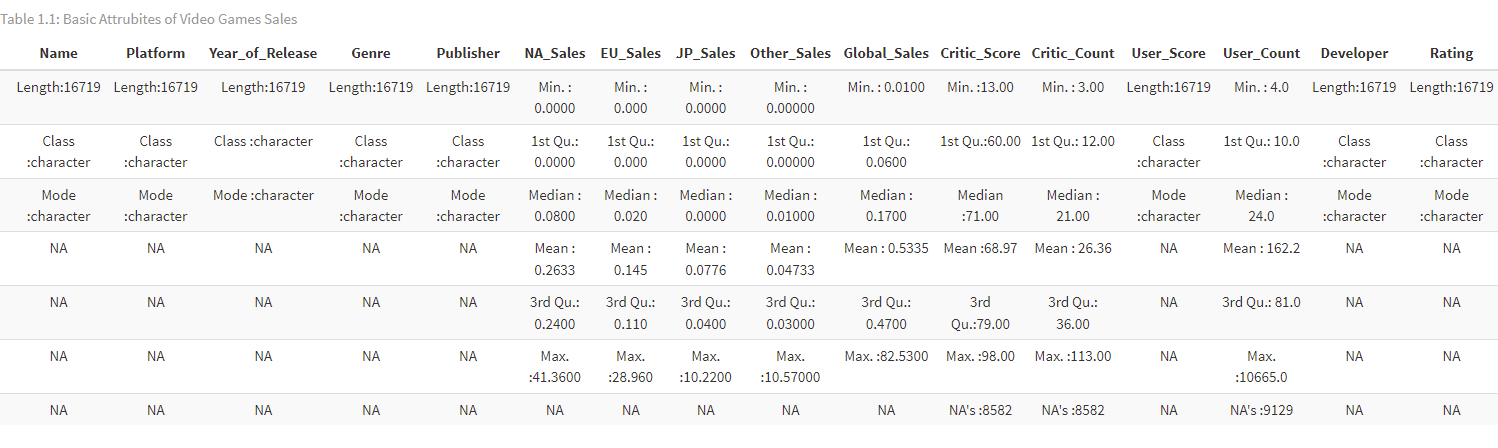
The dataset analyzed in this project is the "Video Games Sales" dataset, which contains information about video games released between 1980 and 2016. The purpose of this dataset is to provide an exploratory data analysis into the video game industry, such as trends in sales, genre popularity, and platform usage, for deciding the future focus of this project. The dataset was sourced from data.world(<https://data.world/sumitrock/video-games-sales>).

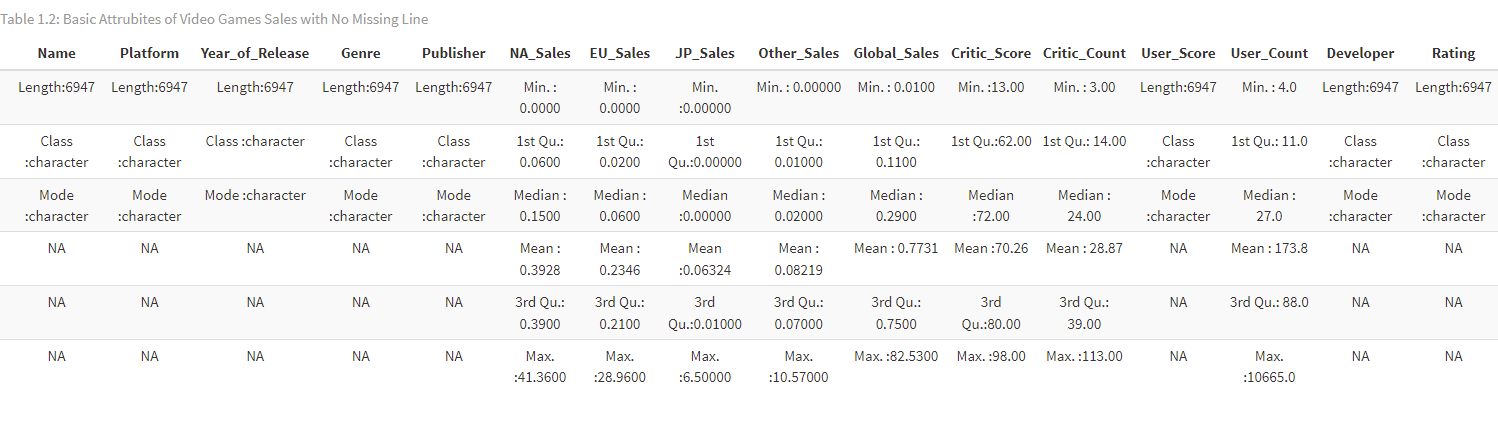
The dataset contains 16,719 rows and 16 fields. Below is the first 5 row of the dataset, including headings.



The dataset includes information about the video game's name, release platform, year of release, genre, publisher, sales in different regions (North America, Europe, Japan, and others), global sales, critic score, and count, user score and count, developer, and rating. The data combines text and numerical data, with some fields being categorical and others being continuous.

Before conducting the exploratory data analysis, the dataset was cleaned to remove missing values and standardize some fields' format. Specifically, rows with missing values were removed, and the "Year\_of\_Release" will be converted to a numeric format after clearing. These cleaning methods were used to ensure that the data was consistent and could be analyzed effectively.





The top two tables show the differences between cleaned data and uncleaned data. Depending on what field I need to focus on, This step will be redone to fit the requirements. The following would be heavily affected by this cleaning process and have serious bias.

Data Analysis:

Chart, scatter chart

Description automatically generated



From this relationship comparison, there are two facts we can quickly notice:

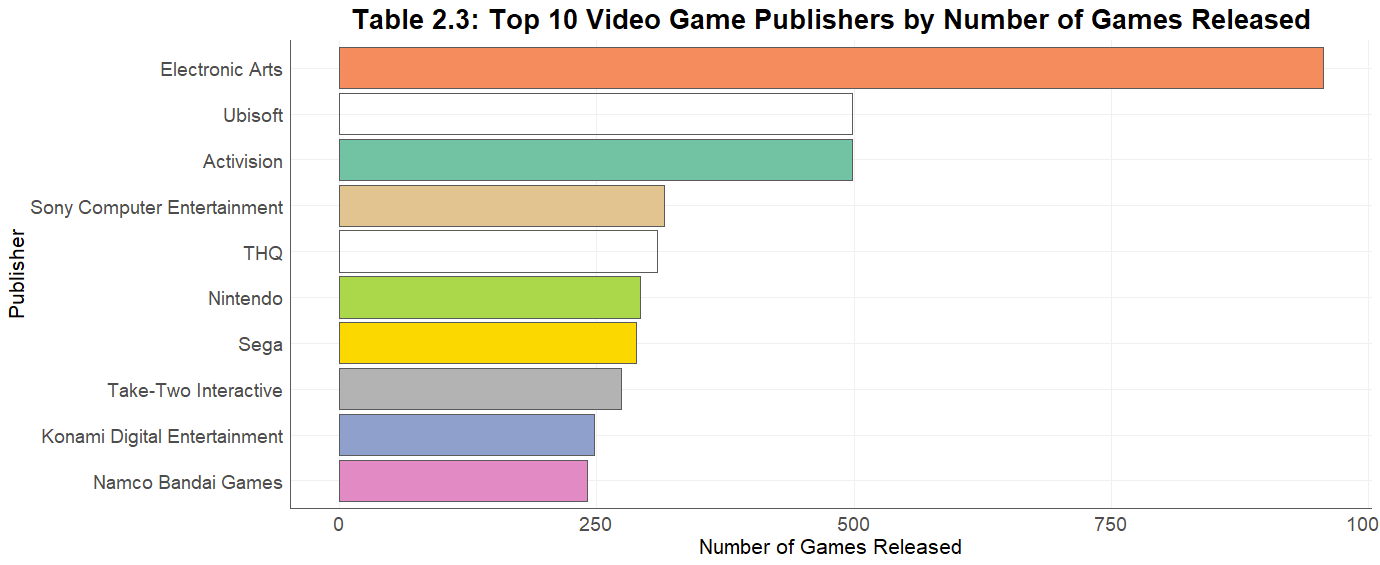
First, there is an outliner on the far end of the corner. This means a game sells a lot across the NA and European; we could focus on that by analyzing the score of games for hot sell games and regular games.

Also, we can see that games NA and European have the same selling amounts, and we can try to find out the factor behind that.

Chart, bar chart

Description automatically generated

This plot shows the games before 2000 have much lower production amounts, which we could avoid making analytics about these games because they might have lower amount, or it could cause games before 2000 to have more NA lines.



This table shows the number of games they released; we could try to find the relationship between the game's scores and the amount of game that company produced.

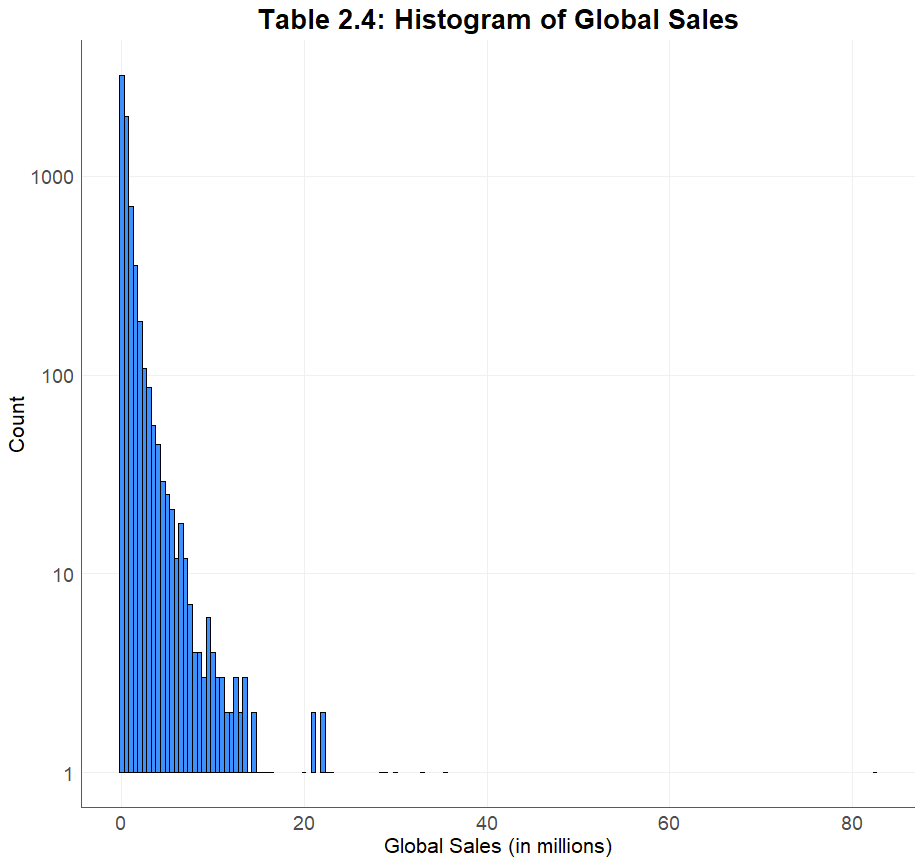


Table 2.4 shows the relationships between player count seals and global sales. It could helps use find out what's the sales' expeditions.

Summary:

The data analysis included a comparison of sales in different regions, revealing an outliner and equal sales in NA and Europe. Additionally, a plot of games by year of release showed that games before 2000 had lower production amounts. A table of the number of games released by companies was also analyzed to find a relationship between game scores and the number of games produced. Finally, a table of player count sales and global sales was examined to determine sales expectations.

Future exploratory data analysis could investigate the relationship between game scores, the number of games produced, and the factors behind the high sales of certain games in NA and Europe. Additionally, further analysis could be done on the popularity of different genres and platforms over time. Overall, this dataset provides valuable insights into the video game industry and has the potential for further analysis.

Data Sources:

Sumit Kumar Shukla (2023). Video Games Sales [Dataset]. https://data.world/sumitrock/video-games-sales