TEB2164

INTRODUCTION TO DATA SCIENCE

FINAL PROJECT

"ANALYSING PERFORMANCE OF VIDEO GAME SALES BASED ON REGION"

Prepared for:

Ts Ahmad Izuddin B Zainal Abidin

Dr. Said Jadid Abdulkadir

Prepared by:

NAME	**ID**
ENGKU MUHAMMAD ZAHIMI BIN CHE ENGKU IBRAHIM	16003930
IQBAL NURIL ANWAR BIN TORMIZI	17004961
KHAIRUL ILHAM	17004942
AHMAD ALIF BIN MOHAMAD AFIFI	17004106
NATHANEEL NAAVIN AISU	16004681
TIA FARISHA BINTI SHAHARUL MIZA	17002588

TABLE OF CONTENTS

1.0 EXECUTIVE SUMMARY

Topic :Analysing Performance of Video Games Based On Region

Category :Sales

Sources :

- Original Owners : Maksim Klimentyev
- Donor of database : Kaggle.com

Problem Description : Annual sales of video games from various genres, developers and even various years based on each country. In this research, our task with the analysis is to examine the success of the year-to-year sales of video games.1.54mb/1.01mb

Data Description : The annual performance of each developer around the country is obtained from Kaggle, while the year begins from 1985 to 2020. The size of the raw database is 1.54 MB, which includes 16 row numbers and 16720 column numbers.

Data Preparation : As there are so many redundant quantities of data in the raw data collection, we continue our analysis by using data preparation measures such as data cleaning, removing and replacing null values for standardized purposes. The technique for preparing the data is CRISP-DM.

Solution and Finding : The elements that can be obtained from the data set include Global Sales, Game Produced, and Distribution of Release Year by Year. In the final section, the identification of the Top values in the data set and game count by Developer is created according to the results of "Platforms with most dataset games."

2.0 PROBLEM DESCRIPTION

2.1 Business Goal

The Data Science Concepts are designed to incorporate mathematics, programming, and business analysis to generate consistent trends within the data itself. Knowledge is power in business, and data is the fuel that generates the power. It is extremely necessary to be able to harness the power of this data through data science.Through drawing numbers and statistics from **data analysis**, an organization will create statistical models to simulate a variety of choices. As a result, companies can learn which path to take to the best possible outcome of how to better market their games on the basis of derived data analysis.

As an example,valuable trends can help businesses accomplish their revenues appropriately. Many raw data can be obtained from the sales of video games produced annually by the developers. This can be used to adapt or customize programs and goods to different markets, the producer can concentrate more on what kind of games to develop and distribute according to the area they want to increase their sales in. It means that organizations may tailor games and other products to cater to specific audiences, for example, by knowing which genres have been most played in one area, will help the company develop new promotions or deals for audiences that may not have been available before.

2.2 Technical Goal

The main aim of this project is to evaluate and determine the annual revenue of each developer for each given year. Nevertheless, in the process of analyzing these results, there have been few difficulties in achieving the objectives of the project.

Below is a list of problems that we encountered :-

- **Datasets are incomplete (null values)**

  Unfilled data tables (null values) causes errors when visualising said databases. The graphs would become inconsistent and drastically askew in various phases. This can be seen when null values are kept and the graph spikes in random phases or that the average is unrealistically lower than predicted.

- **Datasets that are irrelevant.**

  Raw databases had many categories of data, some of these categories are unrelated to our project goals. Including said data categories would be redundant as they do not provide the necessary relations that we aim for in our goals, hence the data must be removed to streamline our data analysis.

- **Wrong Data Visualisation Method.**

  Many methods of visualization can misrepresent data, as can be clearly seen in pie chart visualization where the partition is a misrepresentation of facts and the consequence is that an incorrect conclusion can be drawn from it.It the happen if it unintentionally obstructs the correct understanding of the data or is mistakenly due to a lack of familiarity with the graphics program, a misinterpretation of the data, or because the data can not be properly conveyed.

3.0 DATA DESCRIPTION

3.1 Data Quantity

The format of the data is Comma Separated Values (CSV) format.

The method used to capture the data is Web Scraping. Web scraping is a technique used to extract a large amount of the data from websites where data is exported to be extracted and saved to a local file or to a database in a table/spreadsheet.This data is scraped from Metacritic.com, a website where critics leaves rating on Movies and games.

The size of the database is 1.54MB which is the number of rows is 17 and the number of columns is 16720.

3.2 Data Quality

This data is about the performance of games sales performance in different regions around the world . This data shows the figure of year released, sales in Europe, Japan, North America and Global region, the platform developed, the publisher, the developer and critics score. This data is mostly relevant to conclude key sales performances for further analysis. Our main purpose for using this data is to identify the performance of games based on regions and platforms.The data types are present in numbering and text. Data will be computed and visualized to give a better picture of the performance and where each developer/publisher can focus on.

4.0 DATA PREPARATION

The data that was collected could not be used directly as there are flaws that must be corrected. In order for the dataset to be usable for our project, one that provides accurate and consistent data, certain steps must be taken to clean and prepare the database. Selection and cleansing of data was done using R.studio. We then saved the new dataset and used Power BI for our visualisation as we found it easier and simpler to use. Mistakes in coding would affect our overall analysis, hence to minimize that, we opted for Power Bi.

4.1 Selecting Data

The database that we used consists of several categories. However, not all categories are relevant to our goal. Because our main focus is on the global sales of video games, we decided that ratings and critic score is irrelevant due to those being of a qualitative nature and does not relate with the quantity aspect. Hence, we removed those two categories in order to streamline our database and provide a more consistent dataset. Removing those two categories also reduced the size of our database for further efficiency.

4.2 Cleaning Data

Cleaning data is necessary to provide an accurate data analysis. Some of the datasets had null values. If no cleansing was done, the analysis would provide inconsistent and flawed results as certain values would manipulate the results and cause inaccurate analysis. The dataset had 6815 rows of data that was incomplete. From the original 16720 data, 9905 data was left after we removed those with null values. Even though a large number of data was removed, this cleansed dataset will provide a more consistent relationship when processed visually.

4.3 Visualisation

Data in its table form is hard to analyse and draw relations from. Because of this, we need to convert this into something whereby patterns can be concluded from. With the available data, we picked two categories and produced graphs that showed the patterns between them. We took the Game genre and compared it to Country sales to find out which games were more popular regionally. We did this with several pairs of categories to provide as much relevant information from the data that we had.

5.0 SOLUTION

In this sales analysis report, it shows an overview of performance of video game sales based on region. It shows the different trends happening in the production of video games over the year until today. As we can conclude that in early 1985 until 2000, there was very low production of game consoles. But the starting peak of rising in the production of console games comes in the year of 2008 where many people were introduced to console games. As for today, not so many developers produce console games because people nowadays are more likely to use social media and do other activities compared to playing games.

Next, the sales for countries like Europe (EU), North America (NA), and Japan (JP) can be shown in this report. As we can see in each country, Nintendo's developer has the highest sales in EU, NA and even JP and then followed by games developed from EU and NA itself. To conclude, from Japanese sales by developer, it is hard for developers from EU and NA to infiltrate the market in Japan as the people in that regional area are more interested in playing games from the developer of its country. For EU and NA, maybe the developer can focus more on selling their console games in other parts of the world except Japan.

Besides, in this report, it also shows the sales for each country by genre. It helps developers to decide which type of genre they want to develop for that particular region. For example, EU and NA have the most sales for the action genre. But for JP, the role-playing genre has the highest sales. By doing this, it will help to reduce the cost of developing unused console games by using target segmentation.

Sales reports help to identify possible new business markets where results may be enhanced. So, this report's aim is to make it crystal clear to the sales team as to what is important and what they should be focusing on.

As stated, everyday sales is different and there are many things that are simply uncontrollable. Now we will concentrate on some of the problems or concerns that you can monitor by looking at the report on a regular basis.

Identified problem:

- **What can we understand from this data set?**

1.The changing trend in using technology and video games.

2. Different preferred developers game based on region.

3.Different preferred genre based on region.

6.0 CONCLUSION

In conclusion, video game sales can help keep the above parameters in mind and a well-planned game launch could be successful. Knowing the trends in genre, regions preferences, popularity of games and presently used platforms can help in predicting how successful the game could be. As for the data preparation, we defined the data using a specific method which is CRISP- DM. Here, we gather the data, discover and assess data, cleanse and validate the data, transform and enrich data and finally restore the data. In short, we do think that if we concentrate on certain issues or difficulties we can increase the performance of video game sales.

7.0 APPENDIX

PROJECT SOURCE CODE: https://github.com/engkuzahimi/introds-jan20

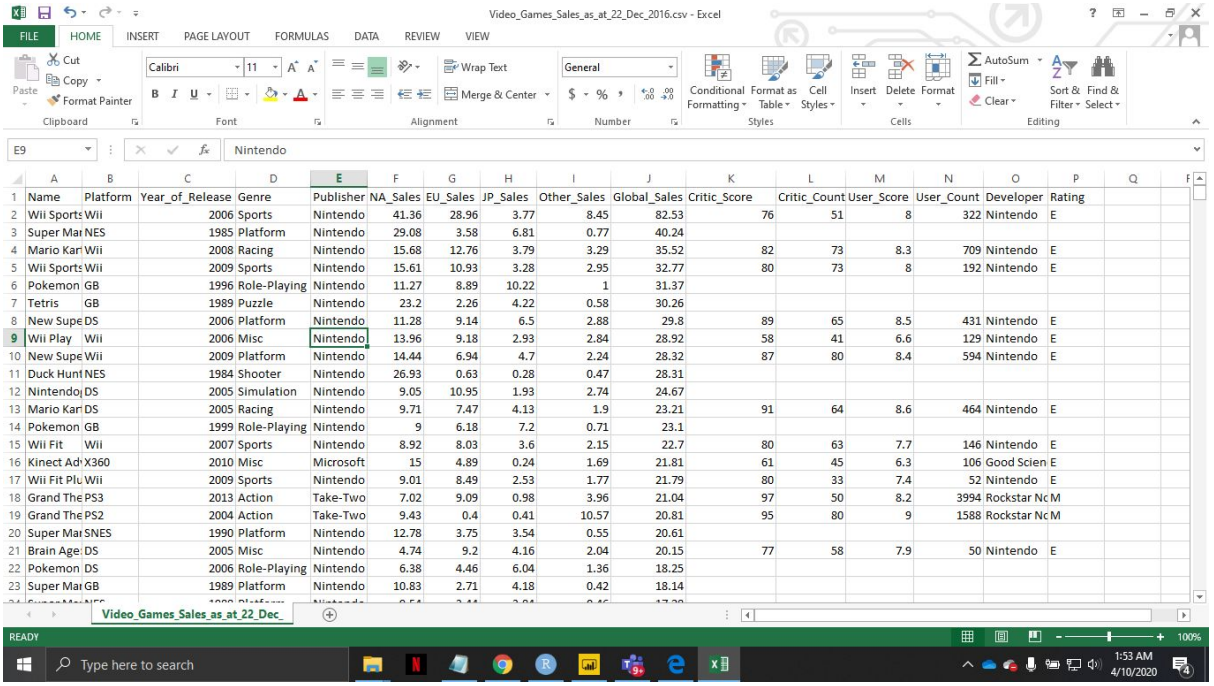


Figure 1.0 Original Data Before Cleansing

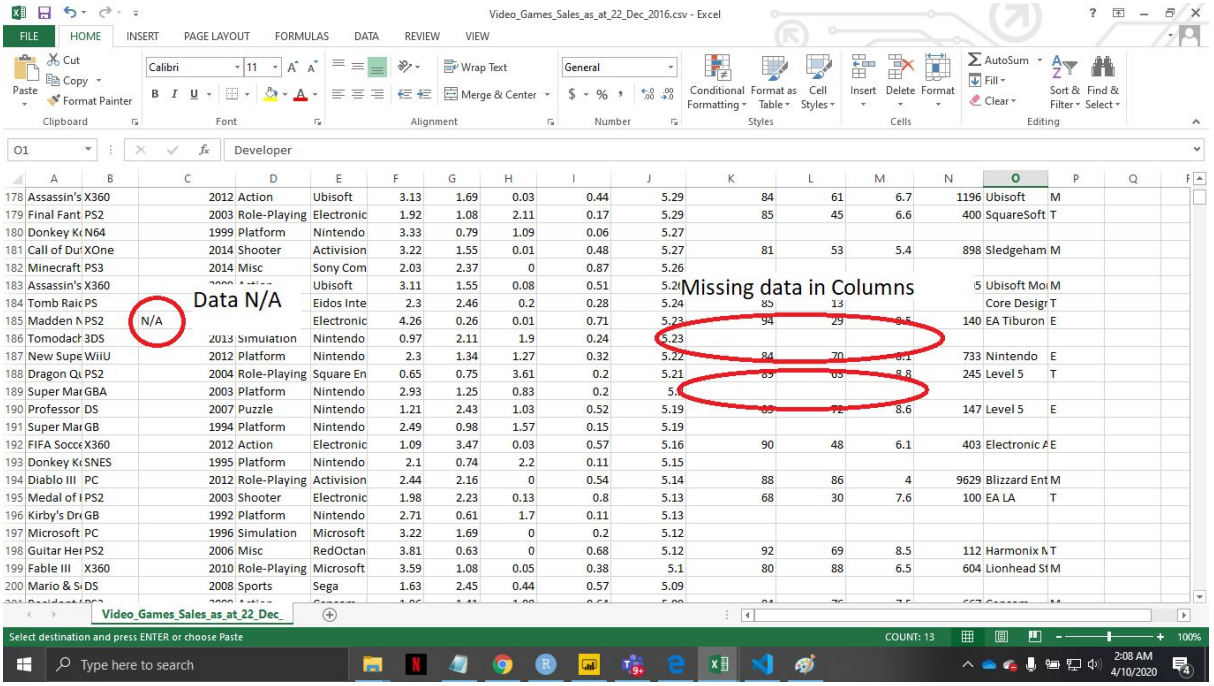


Figure 2.0 Data Problem

```
server <- function(input, output) {
  # Reading the dataset into data and then omit rows with empty cells
  data <- na.omit(read.csv('s.csv', stringsAsFactors = FALSE, na.strings=c("N/A")))

  #
  # # Removing rows with any empty cells inside it
  data <- data[!(is.na(data$Name) | data$Name==""), ]
  data <- data[!(is.na(data$Platform) | data$Platform==""), ]
  data <- data[!(is.na(data$Year) | data$Year==""), ]
  data <- data[!(is.na(data$Genre) | data$Genre==""), ]
  data <- data[!(is.na(data$Publisher) | data$Publisher==""), ]
  data <- data[!(is.na(data$NA_Sales) | data$NA_Sales==""), ]
  data <- data[!(is.na(data$EU_Sales) | data$EU_Sales=="") , ]
  data <- data[!(is.na(data$JP_Sales) | data$JP_Sales==""), ]
  data <- data[!(is.na(data$Other_Sales) | data$Other_Sales==""), ]
  data <- data[!(is.na(data$Global_Sales) | data$Global_Sales==""), ]
  data <- data[!(is.na(data$Developer) | data$Developer==""), ]

  #reexport cleansed data
  write.csv(data,"C:\\Users\\Jim\\Desktop\\CleanedData.csv", row.names = TRUE)
```

Figure 3.0 Cleansing Code

	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Developer
1	Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	Nintendo
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	Nintendo
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	Nintendo
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.14	6.50	2.88	29.80	Nintendo
8	Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92	Nintendo
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.44	6.94	4.70	2.24	28.32	Nintendo
12	Mario Kart DS	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.90	23.21	Nintendo
14	Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.60	2.15	22.70	Nintendo
15	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	15.00	4.89	0.24	1.69	21.81	Good Science Studio

Figure 4.0 Data After Cleaning

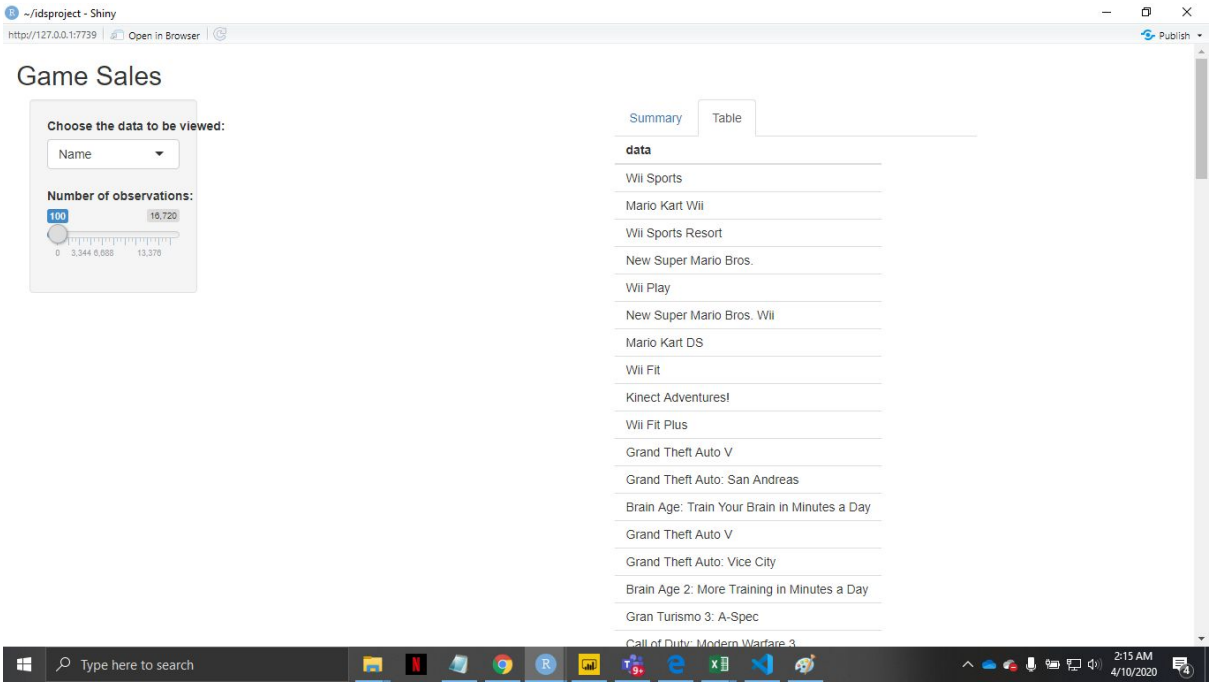


Figure 5.0 Data Presentation Using Shiny - Table View

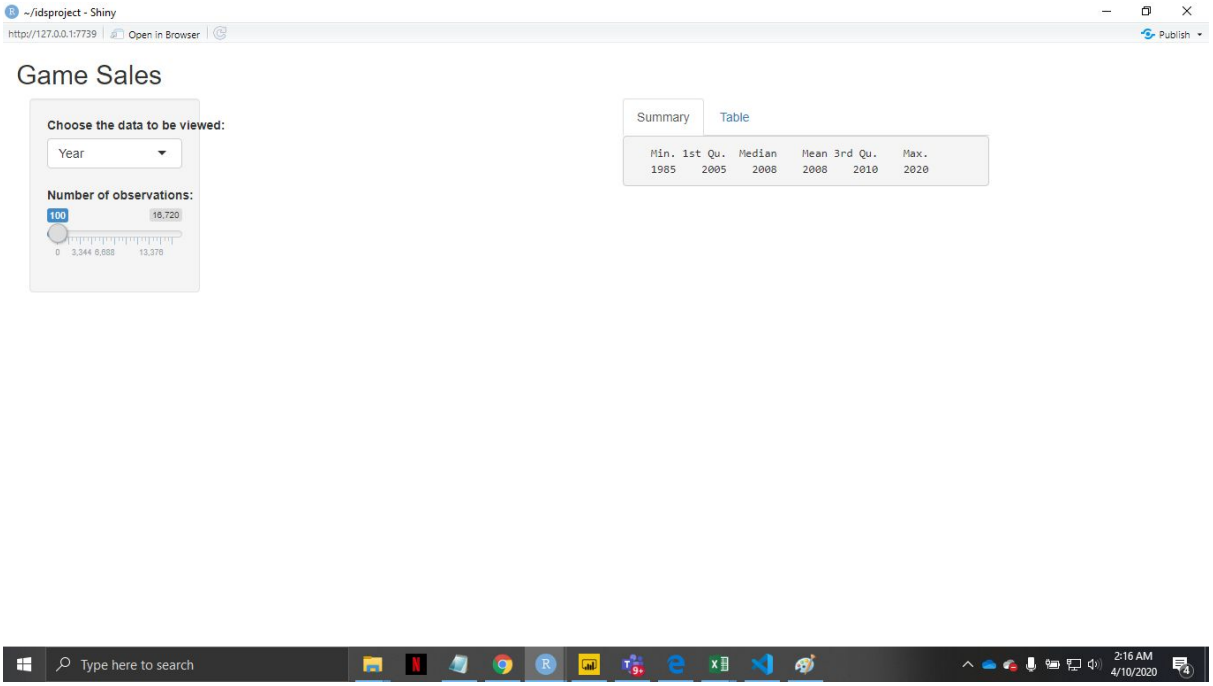


Figure 6.0 Data Presentation Using Shiny - Summary

```r
ui <- fluidPage(
  titlePanel('Game Sales'),
  splitLayout(
    sidebarPanel(
      selectInput(
        inputId = 'dataset',
        label = 'Choose the data to be viewed:',
        choices = c('Name', 'Platform', 'Year', 'Genre', 'Publisher', 'SalesNorthAmerica', 'SalesEurope', 'SalesJapan', 'SalesOther', 'SalesGloba
      ),
      sliderInput(
        inputId = 'obs',
        label = 'Number of observations: ',
        min = 0,
        max = 16720,
        value = 100
      )
    ),

    mainPanel(
      tabsetPanel(type = "tabs",
                  #tabPanel("Plot", plotOutput("plot")),
                  tabPanel("Summary", verbatimTextOutput("summary")),
                  tabPanel("Table", tableOutput("view"))
      )
    )
  )
)
```

Figure 7.0 Shiny Code - UI

```r
# Getting the input from user on which column of table to display
  dataInput <- reactive({
    switch(input$dataset,
           'Name' = data$Name,
           'Platform' = data$Platform,
           'Year' = data$Year,
           'Genre' = data$Genre,
           'Publisher' = data$Publisher,
           'SalesNorthAmerica' = data$NA_Sales,
           'SalesEurope' = data$EU_Sales,
           'SalesJapan' = data$JP_Sales,
           'SalesOther' = data$Other_Sales,
           'SalesGlobal' = data$Global_Sales,
           'Developer' = data$Developer
    )
  })

  output$summary <- renderPrint({
    dataColumn <- dataInput()
    summary(dataColumn)
  })

  output$view <- renderTable({
    dataColumn <- dataInput()
    head(dataColumn, n = input$obs)
  })

```

Figure 8.0 Shiny Code - Server

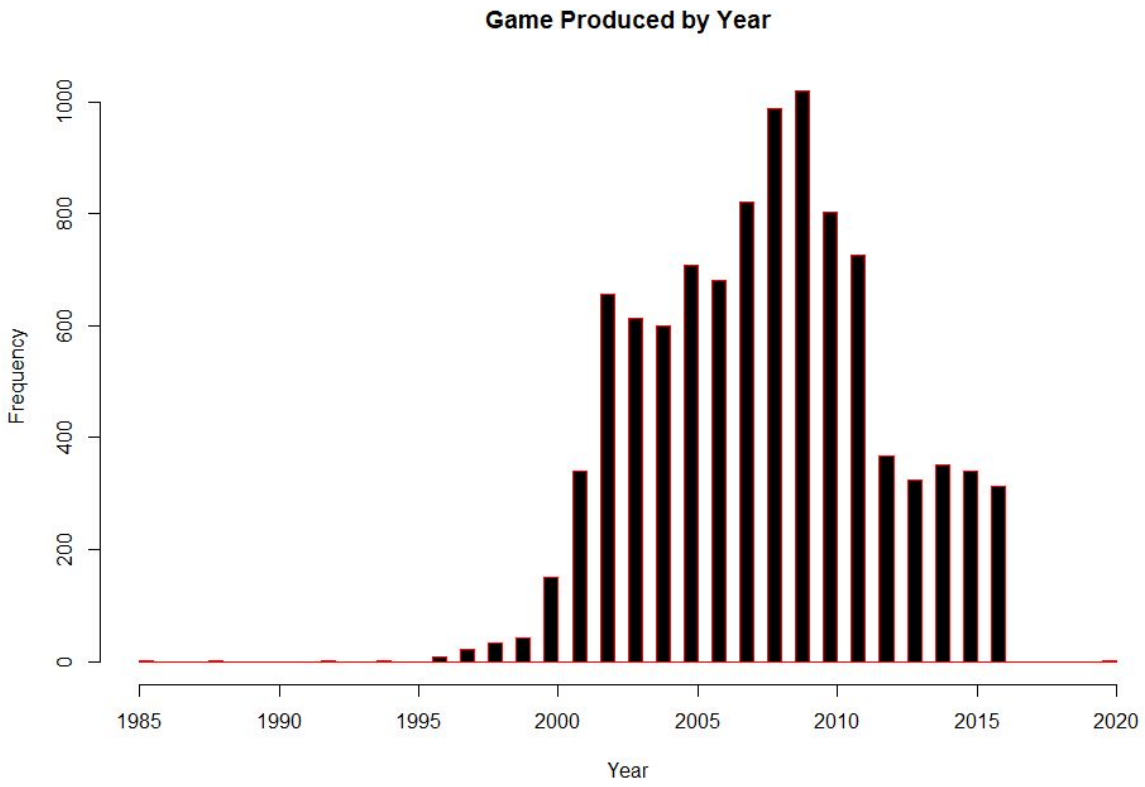


Figure 9.0 Graph made using R - Game Produced by Year

```r
hist(data$Year,
     main="Game Produced by Year",
     xlab="Year",
     border="red",
     col="black",
     xlim=c(1985,2020),
     breaks=55)

```

Figure 10.0 Graph Code

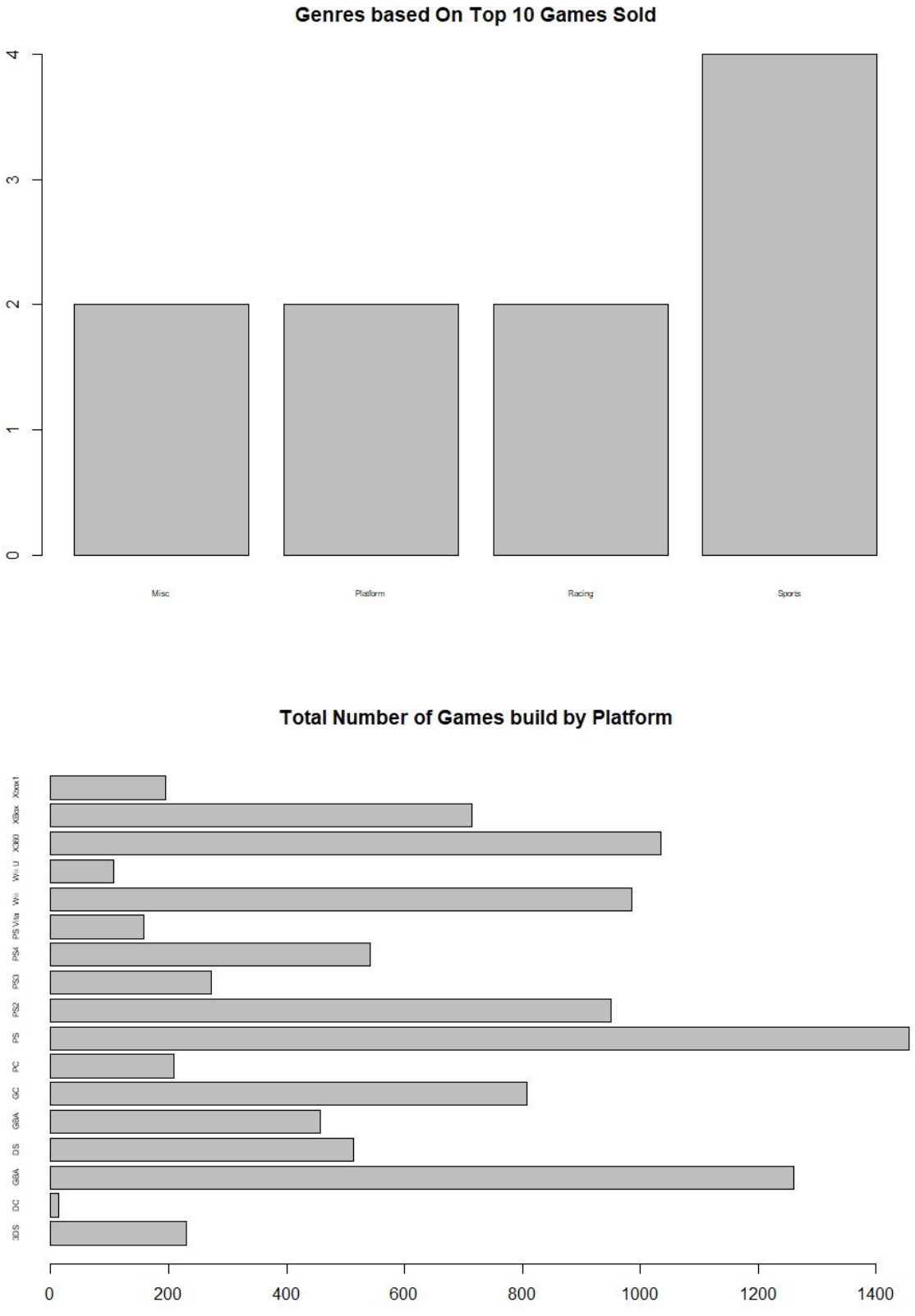
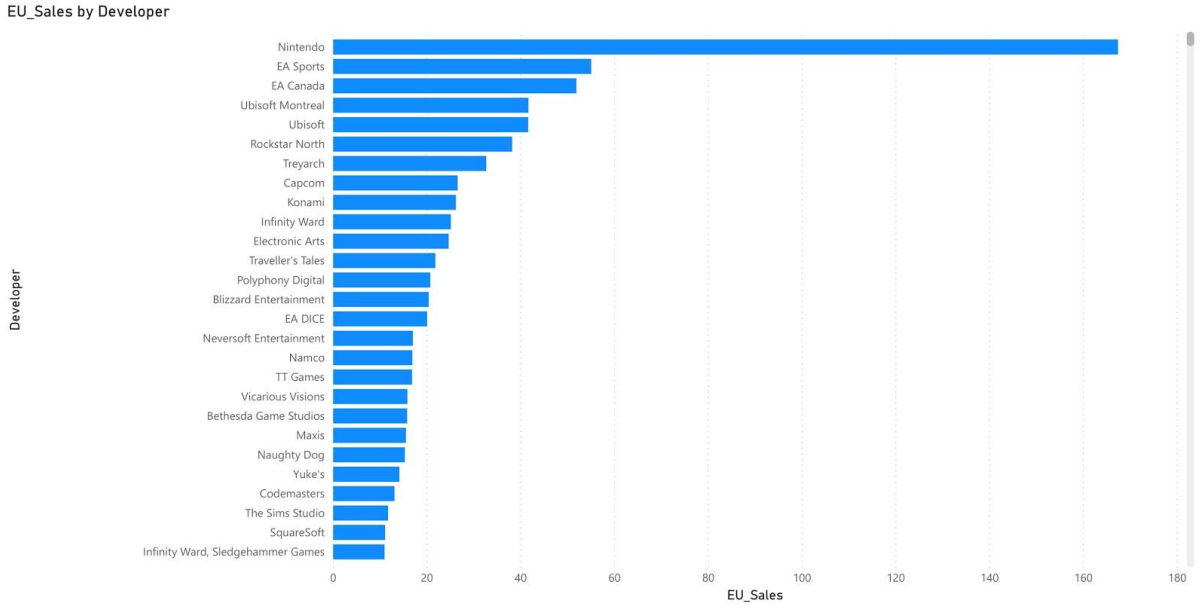
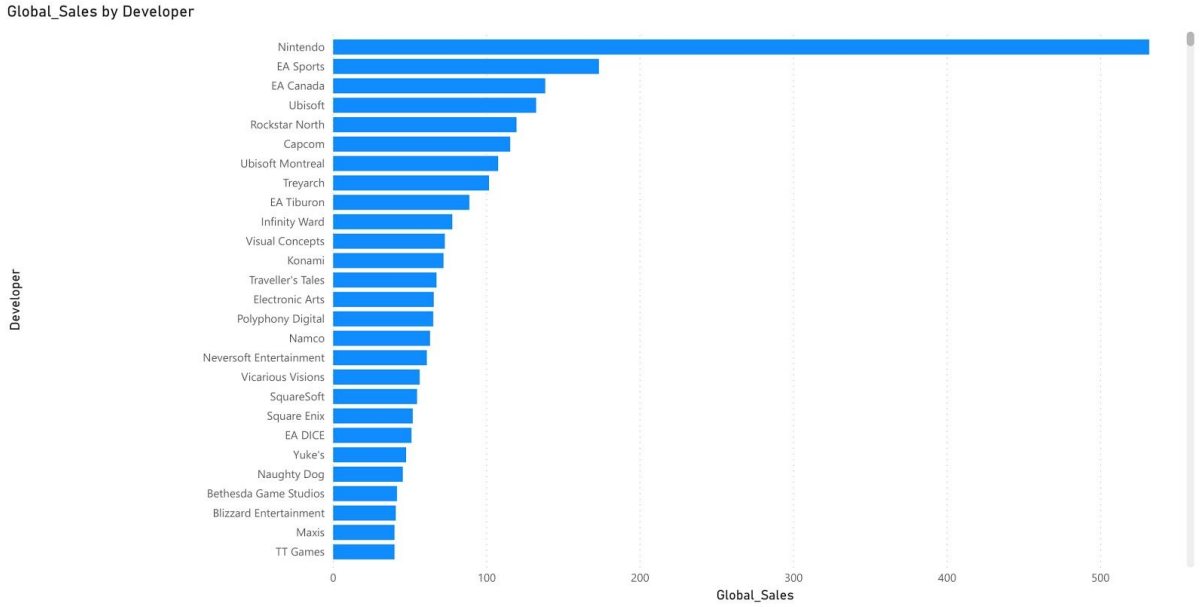


Figure 11.0 Graph made using R - Genres based on Top 10 Games & Total Number of Games by Platform

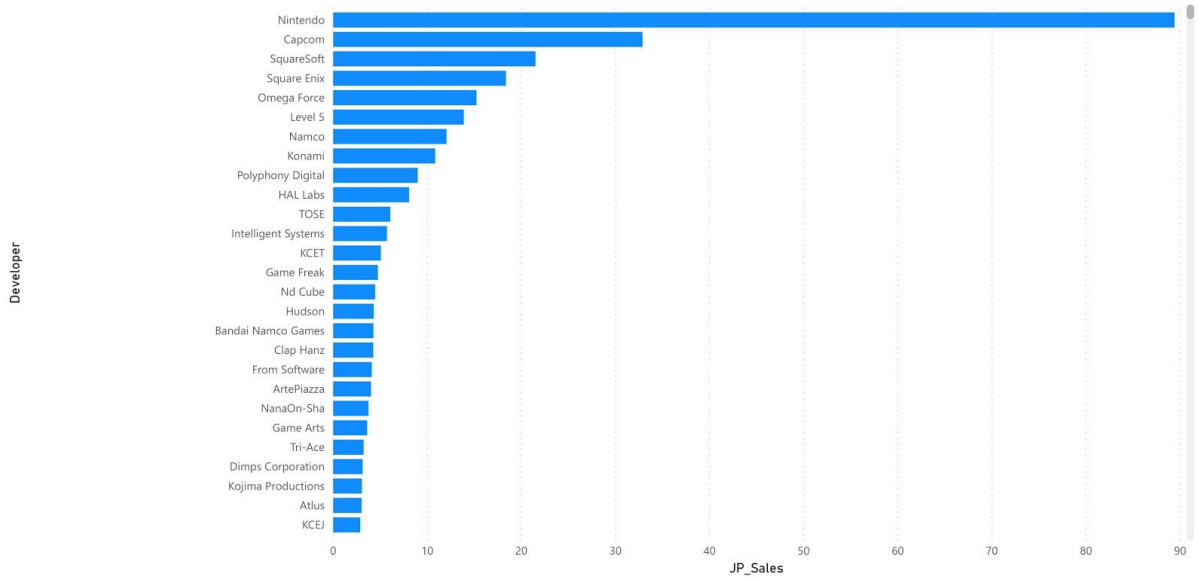
```r
top10 <- head(data, 10)
barplot(table(head(data$Genre,10)),cex.names=0.5, main="Genres based On Top 10 Games Sold",
        names.arg=c("Misc","Platform", "Racing", "Sports"))

table(data$Platform)
barplot(table(data$Platform),cex.names=0.5, main="Total Number of Games build by Platform", horiz=TRUE,
 mes.arg=c("3DS", "DC", "GBA", "DS","GBA","GC","PC","PS","PS2","PS3","PS4","PS Vita","Wii","Wii U","X360","XBox","Xbox1"))
```

Figure 12.0 Graph Code

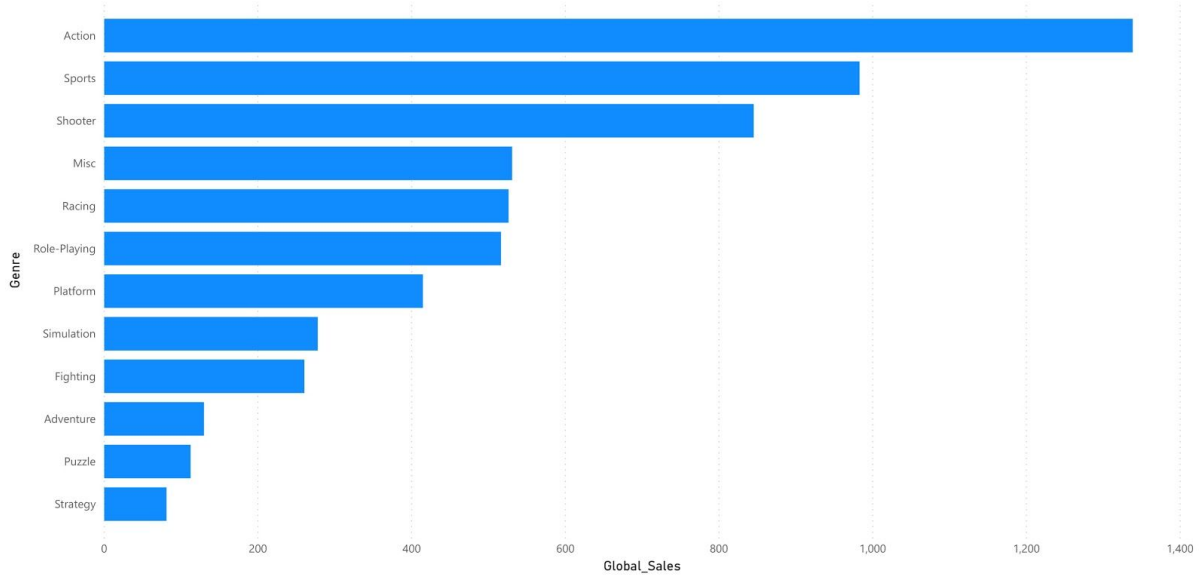


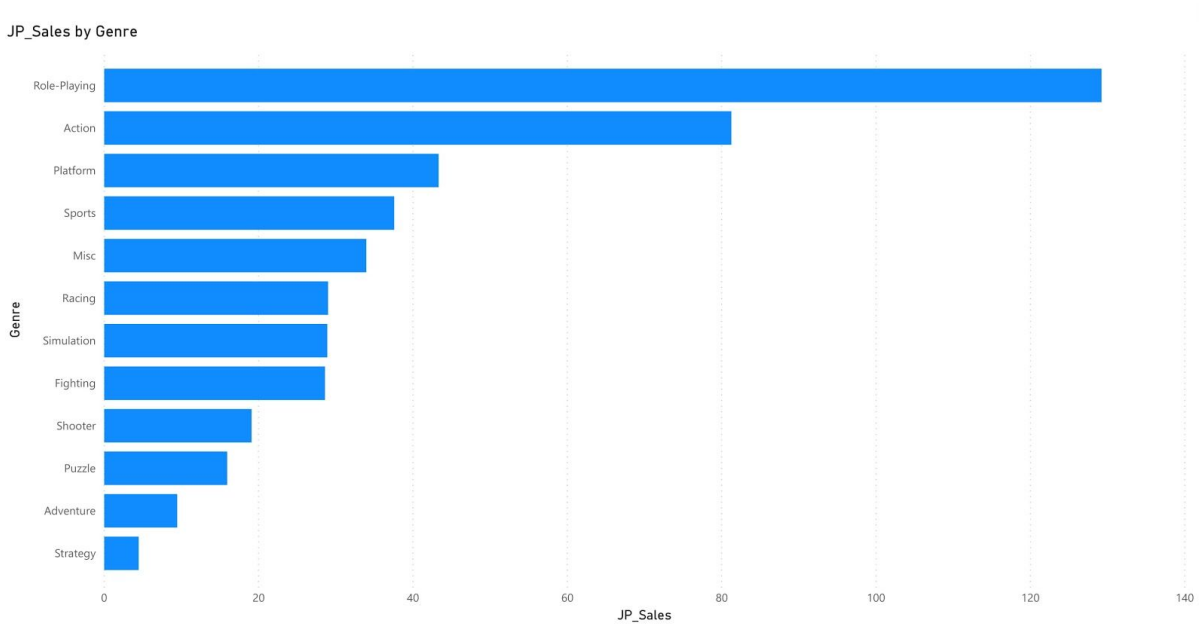


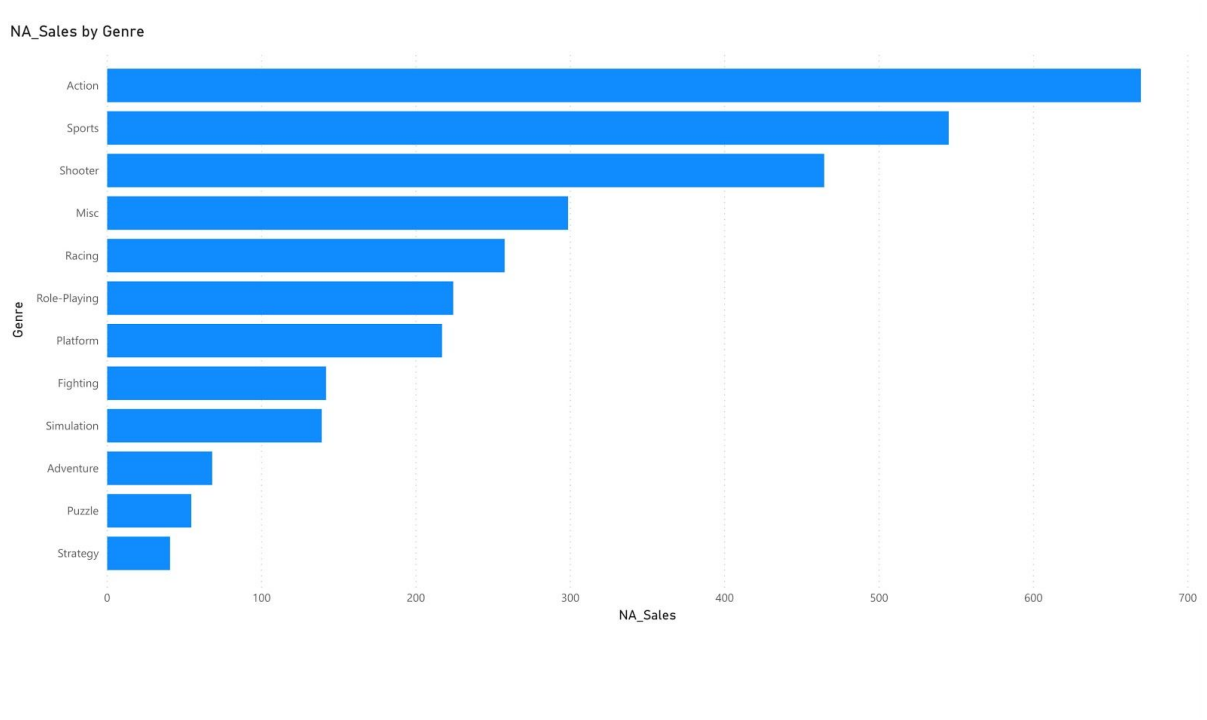
JP_Sales by Developer



Global_Sales by Genre




JP_Sales by Genre
Role-Playing
Action
Platform
Sports
Misc
Racing
Simulation
Fighting
Shooter
Puzzle
Adventure
Strategy
Genre
0
20
40
60
80
100
120
140
JP_Sales



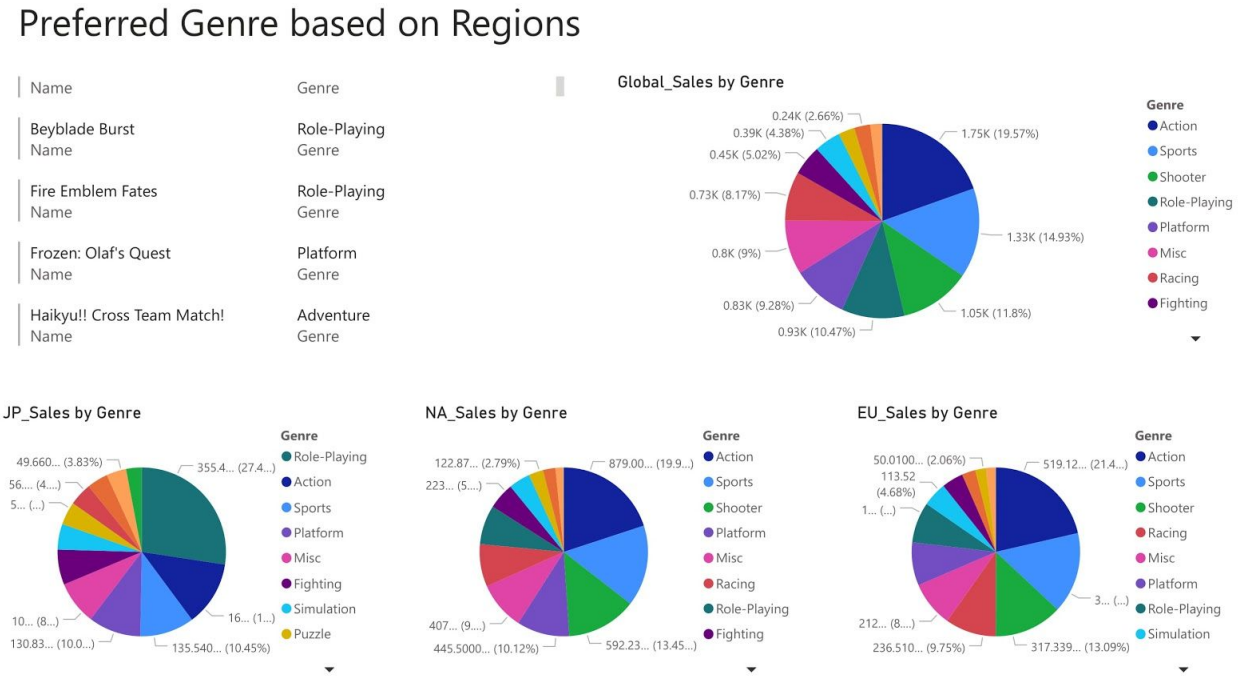


Figure 13.0 Additional Graph made using Power BI