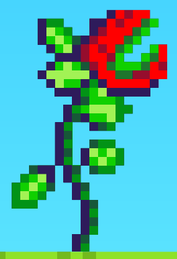
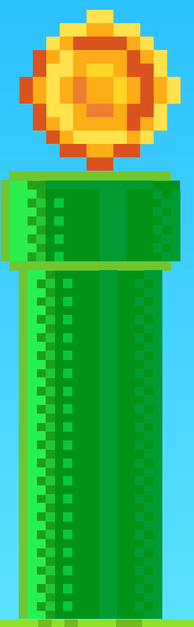
EXPLORATORY DATA ANALYSIS IN R

VIDEO GAME SALES DATA





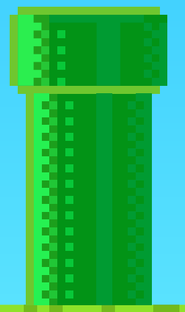


TABLE OF CONTENT

INTRODUCTION

This analysis explores a dataset of video games with sales exceeding 100,000 copies, sourced from vgchartz.com. The dataset, comprising 16,598 records, includes key details such as game rank, name, platform, release year, genre, publisher, and sales figures for different regions. It was cleaned to remove 2 incomplete records.

OBJECTIVES

The primary objectives of this analysis are to prepare the dataset by ensuring its accuracy and completeness, generate descriptive statistics to understand key metrics, and create visualizations to reveal trends and patterns in the data. We aim to analyse changes in sales over time, compare sales performance across different regions and platforms, and provide actionable insights and recommendations based on these findings.

DATA OVERVIEW

The dataset consists of video game sales data, focusing on games that have sold more than 100,000 copies. It includes a total of 16,598 records, with fields capturing various aspects of each game. The dataset features columns for the game's rank, name, platform, release year, genre, publisher, and sales figures in different regions (North America, Europe, Japan, and other parts of the world), as well as total global sales. Two records were removed due to incomplete information.

DATA SOURCE AND FORMAT:

The dataset, scraped from vgchartz.com using a BeautifulSoup-based Python script available on GitHub, is in CSV format. This format is ideal for analysis and visualization with tools like R.

Columns

- Rank
- Name
- Platform
- Year
- Genre
- Publisher
- NA_Sales
- EU_Sales
- JP_Sales
- Other_Sales
- Global_Sales

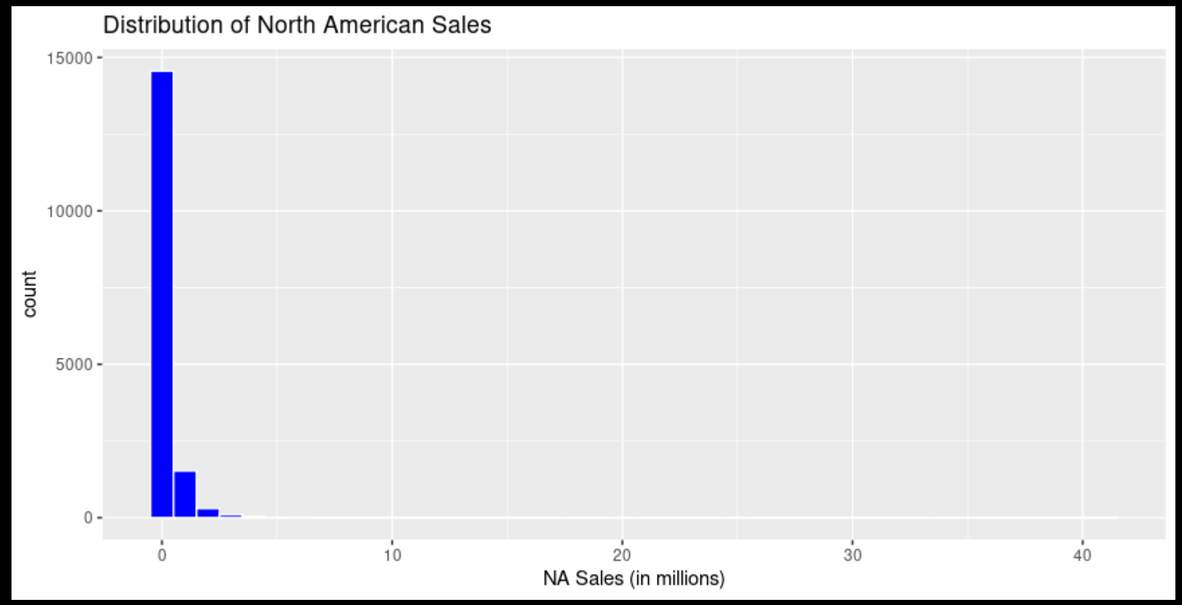
Data Loading and Cleaning

Import the CSV file containing the video game sales data into R.Handle missing values and remove incomplete records.Convert data types as needed (e.g., dates, factors).Normalize or standardize data if required.

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(readr)
library(ggcorrplot)

# Load the dataset
vgsales <- read_csv("/home/ansio-user/Downloads/vgsales.csv")

# View the structure of the dataset
str(vgsales)

# Summary statistics
summary(vgsales)

# Check for missing values
colSums(is.na(vgsales))

# Data Cleaning: Remove rows with missing values
vgsales <- na.omit(vgsales)
```

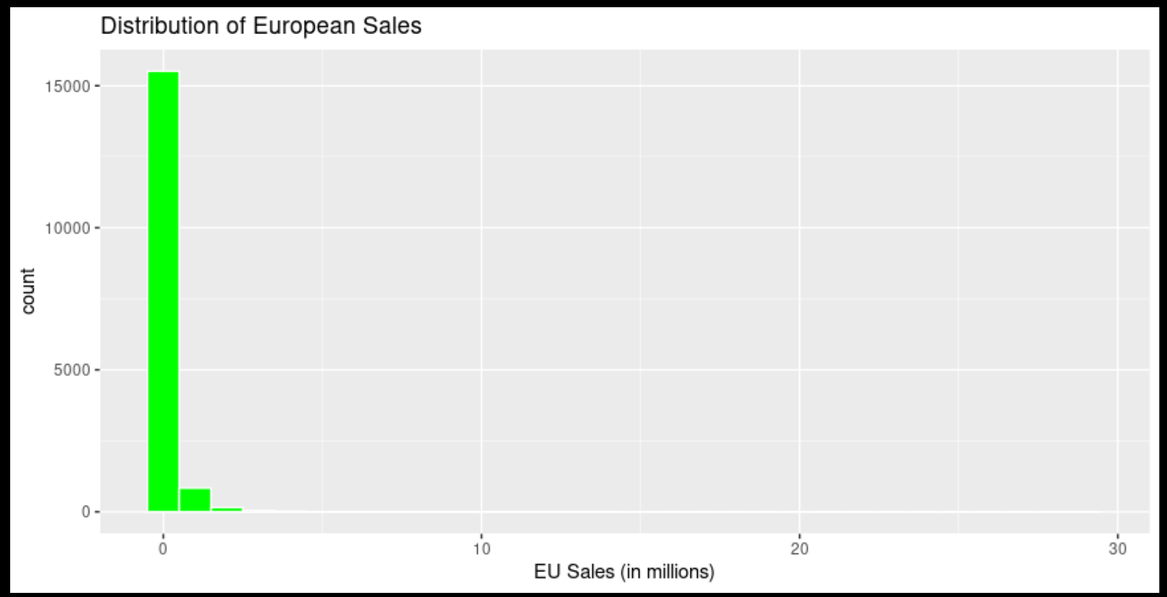
Data Analysis & visualisation

```
# Distribution of Global Sales by Region
ggplot(vgsales, aes(x = NA_Sales)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  labs(title = "Distribution of North American Sales", x = "NA Sales (in
millions)")
```



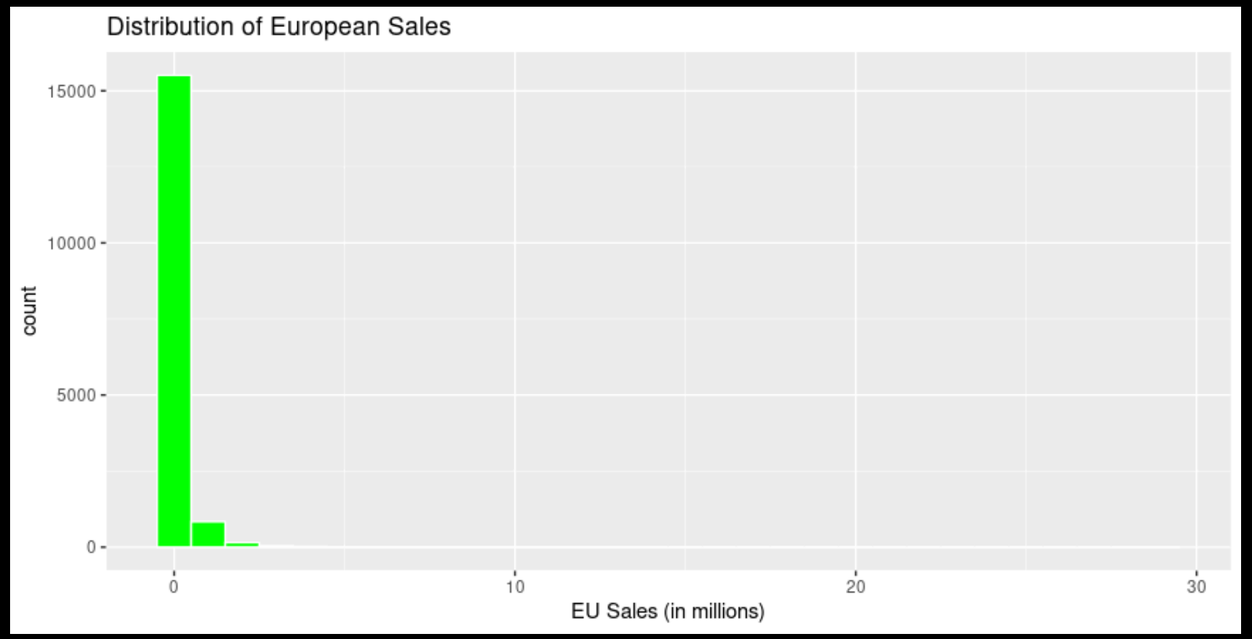
Data Analysis & visualisation

```
ggplot(vgsales, aes(x = EU_Sales)) +
  geom_histogram(binwidth = 1, fill = "green", color = "white") +
  labs(title = "Distribution of European Sales", x = "EU Sales (in
millions)")
```



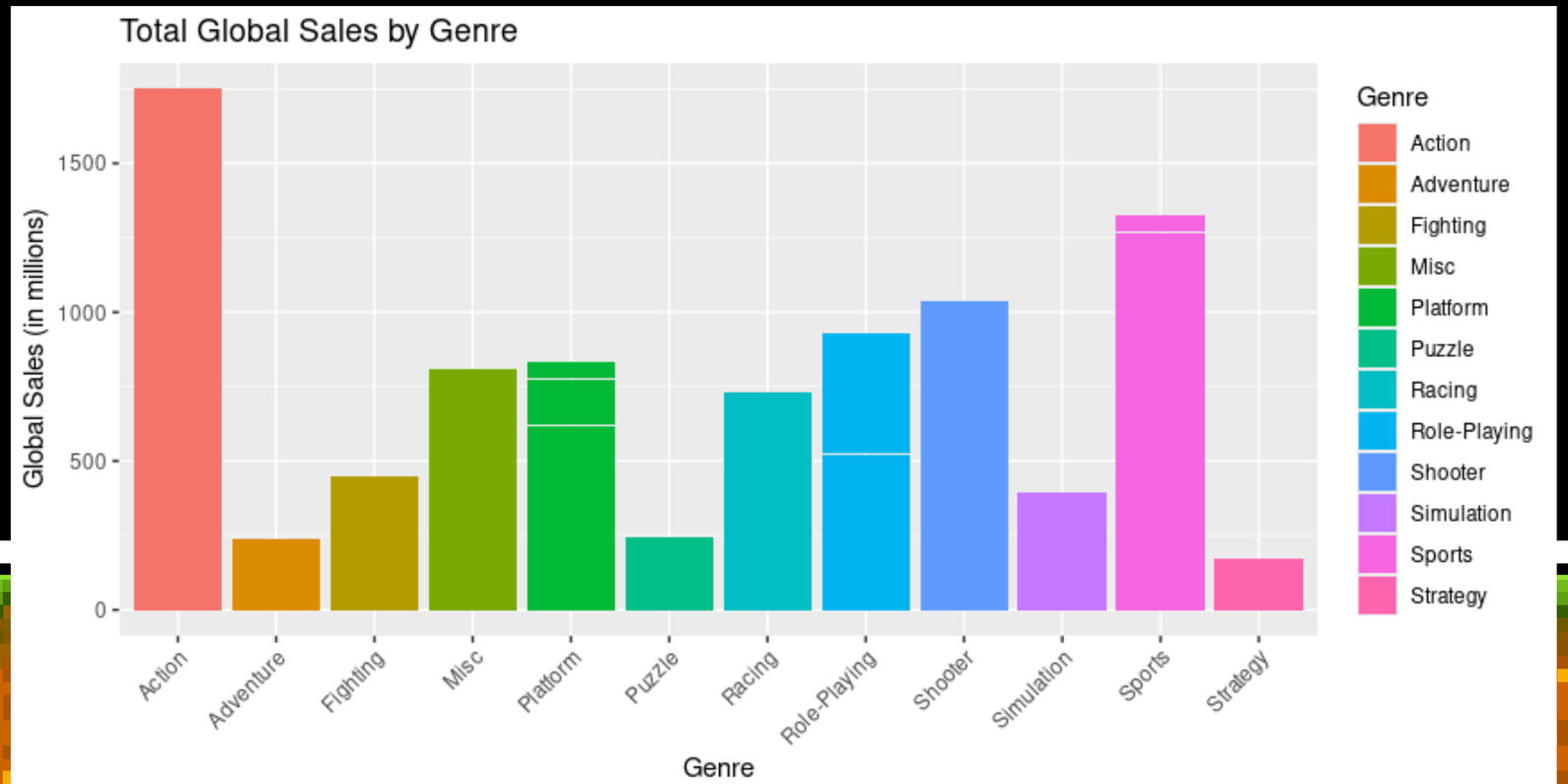
Data Analysis & visualisation

```
ggplot(vgsales, aes(x = JP_Sales)) +
  geom_histogram(binwidth = 1, fill = "red", color = "white") +
  labs(title = "Distribution of Japanese Sales", x = "JP Sales (in millions)")
```



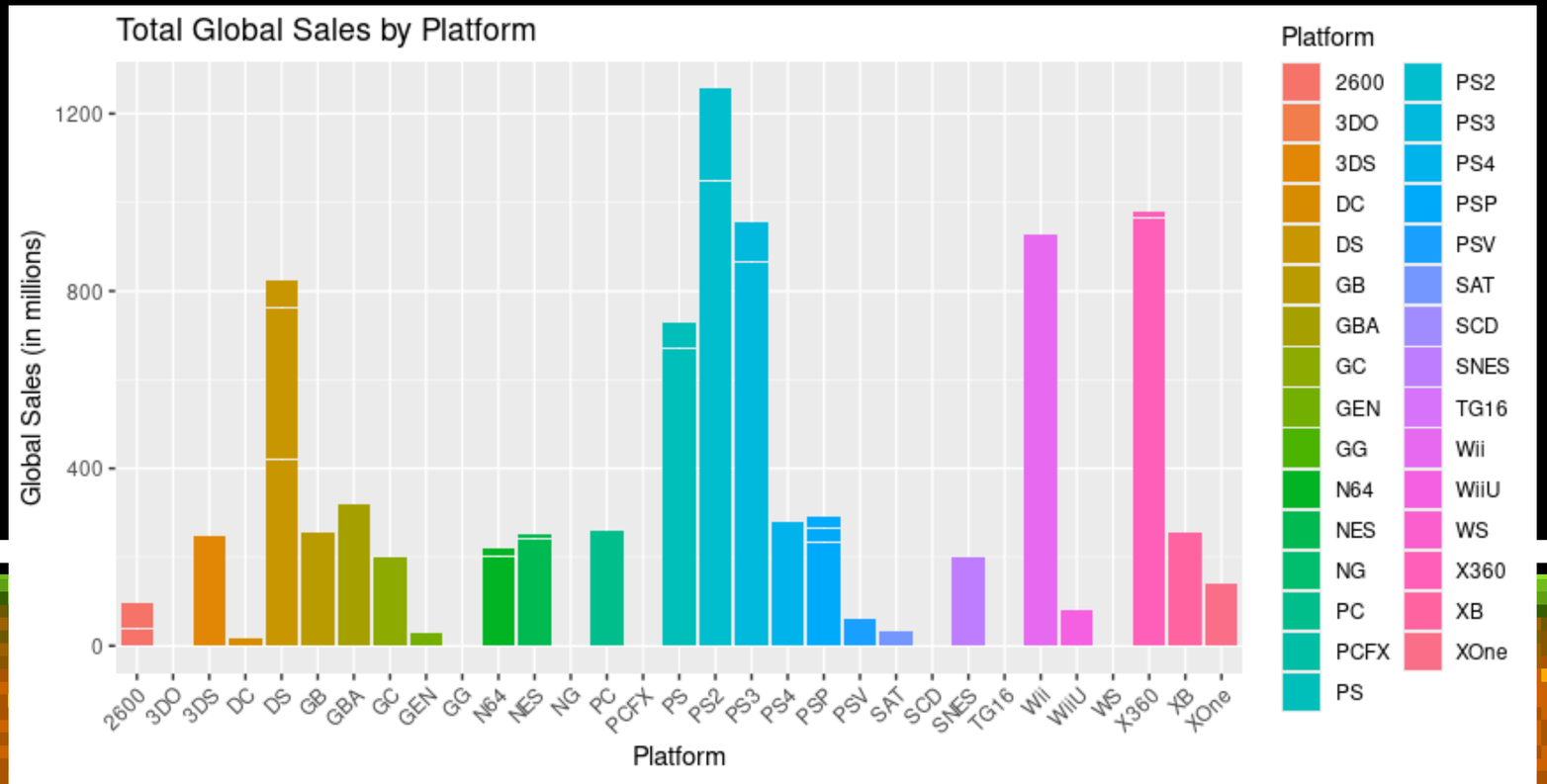
Data Analysis & visualisation

```
# Sales by Genre - Bar Plot
ggplot(vgsales, aes(x = Genre, y = Global_Sales, fill = Genre)) +
 geom_bar(stat = "identity") +
 labs(title = "Total Global Sales by Genre", x = "Genre", y = "Global Sales (in
millions)") +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



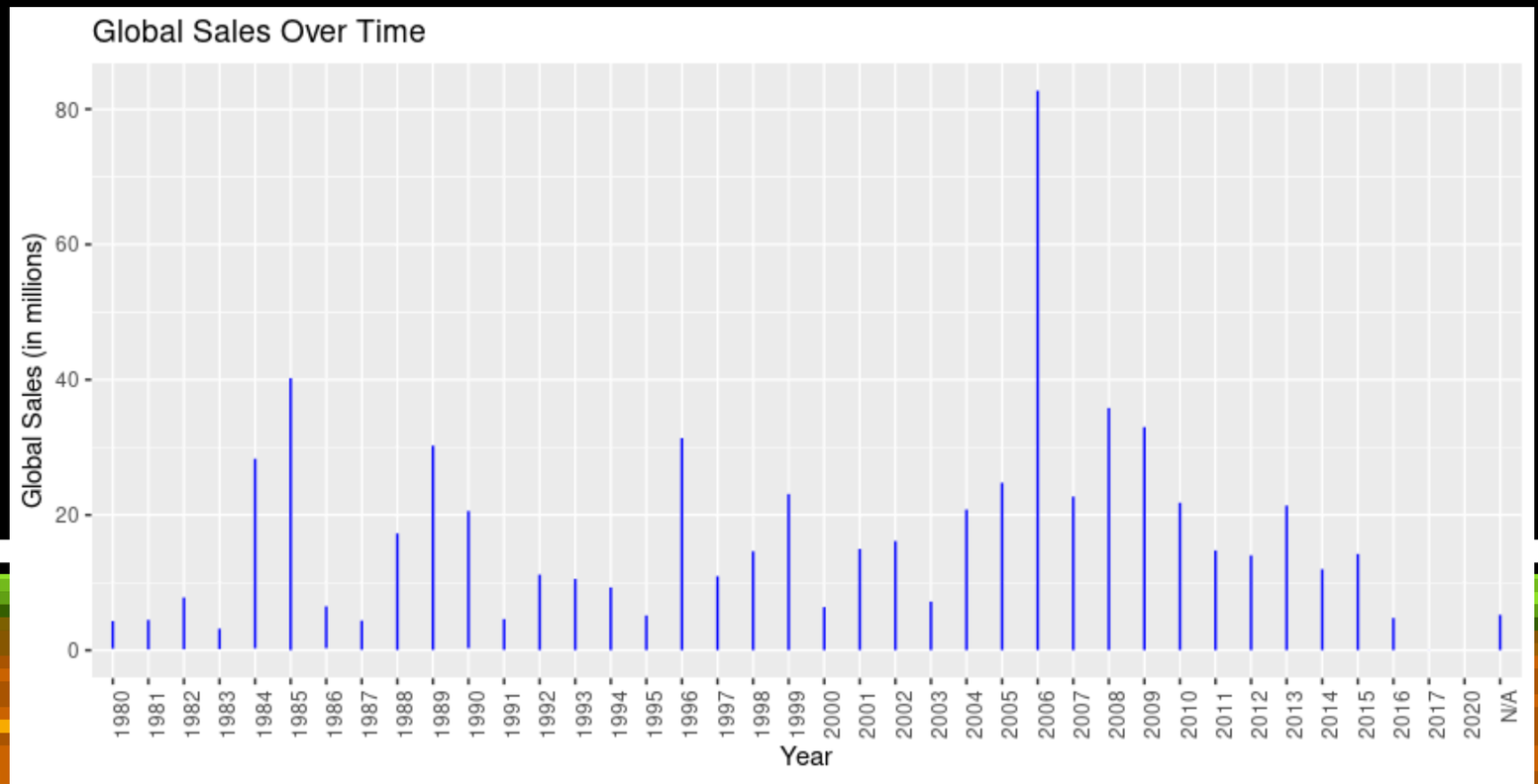
Data Analysis & visualisation

```
# Sales by Platform - Bar Plot
ggplot(vgsales, aes(x = Platform, y = Global_Sales, fill = Platform)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Global Sales by Platform", x = "Platform", y = "Global
Sales (in millions)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



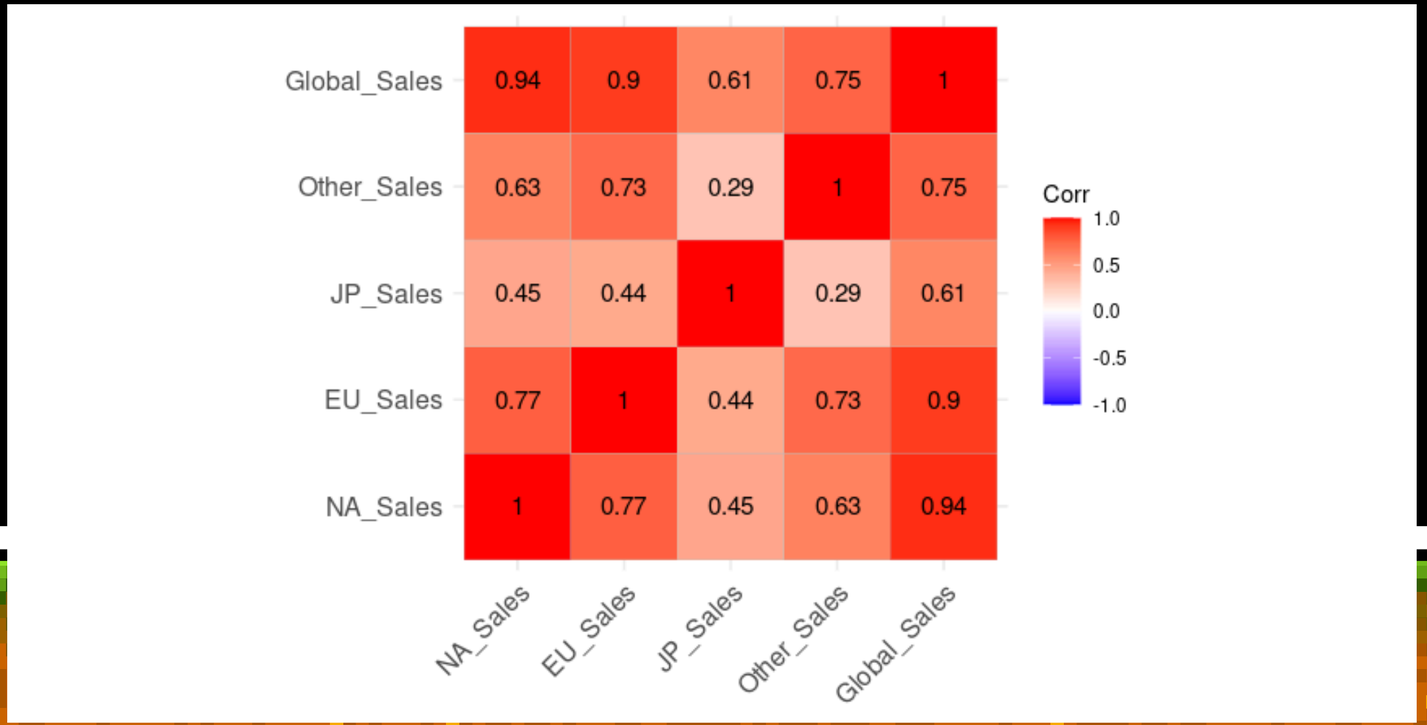
Data Analysis & visualisation

```
# Global Sales Over Time - Line Plot
ggplot(vgsales, aes(x = Year, y = Global_Sales)) +
  geom_line(color = "blue") +
  labs(title = "Global Sales Over Time", x = "Year", y = "Global Sales (in
millions)")+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



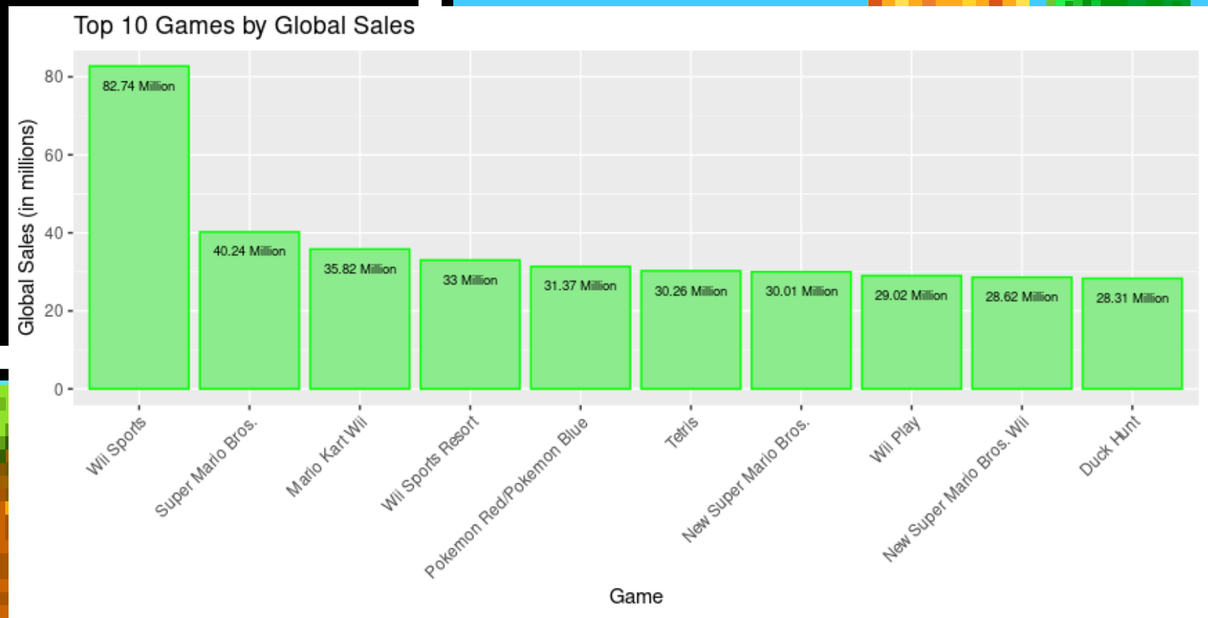
Data Analysis & visualisation

```
# Correlation Matrix
sales_data <- vgsales %>% select(NA_Sales, EU_Sales, JP_Sales, Other_Sales,
Global_Sales)
cor_matrix <- cor(sales_data)
ggcorrplot(cor_matrix, lab = TRUE)
```



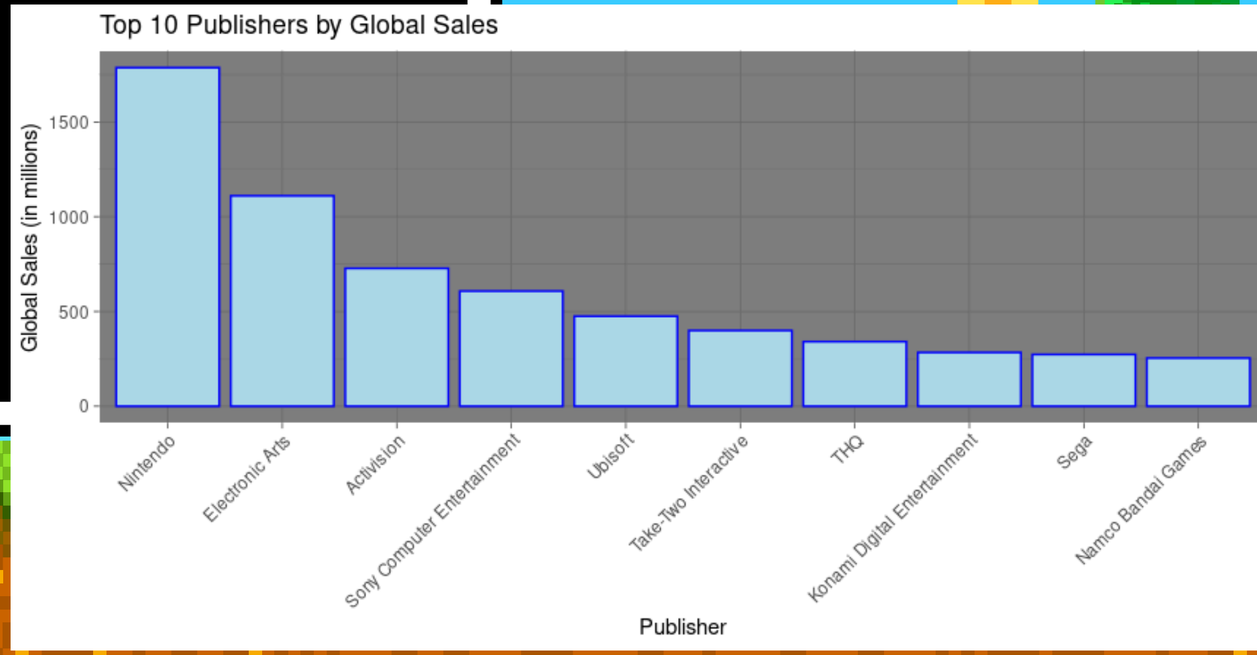
Data Analysis & visualisation

```
# Top 10 Games by Global Sales - Bar Plot
top10_games <- vgsales %>% arrange(desc(Global_Sales)) %>% head(10)

ggplot(top10_games, aes(x = reorder(Name, -Global_Sales), y =
Global_Sales)) +
  geom_bar(stat = "identity", fill = "lightgreen", color = "green") +
  geom_text(aes(label = paste(Global_Sales, "Million")), vjust = 2.5, color =
"black", size = 2.5) +
  labs(title = "Top 10 Games by Global Sales", x = "Game", y = "Global Sales
(in millions)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



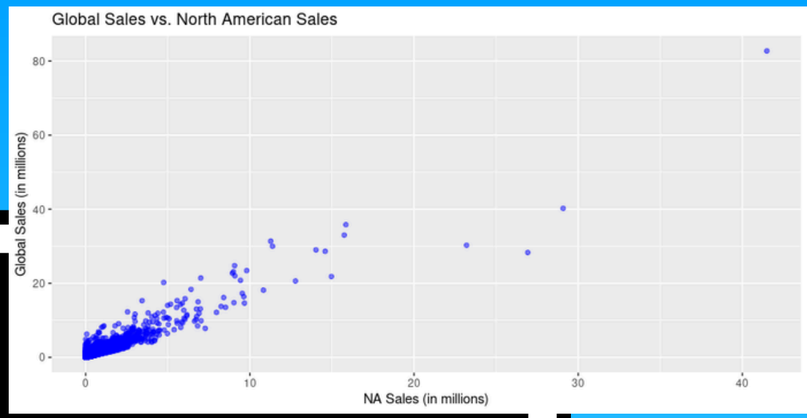
Data Analysis & visualisation

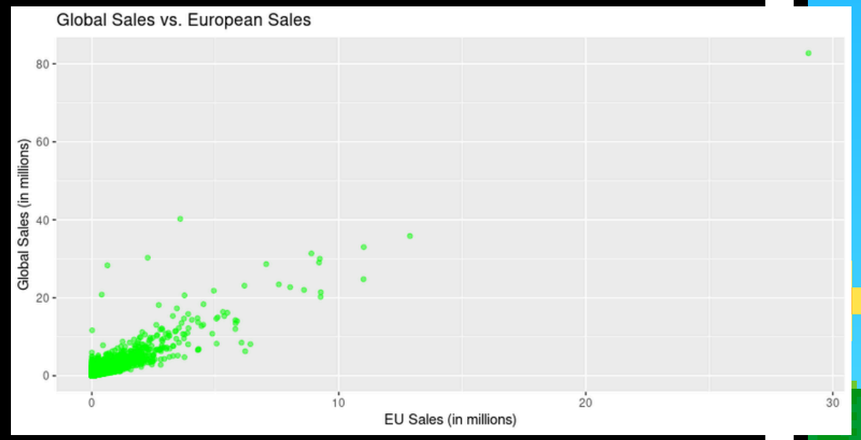
```
# Sales by Publisher - Bar Plot
top_publishers <- vgsales %>% group_by(Publisher) %>%
summarise(Total_Sales = sum(Global_Sales)) %>% arrange(desc(Total_Sales))
%>% head(10)
ggplot(top_publishers, aes(x = reorder(Publisher, -Total_Sales), y =
Total_Sales)) +
  geom_bar(stat = "identity", color='blue',fill='lightblue') +
  labs(title = "Top 10 Publishers by Global Sales", x = "Publisher", y = "Global
Sales (in millions)") +theme_dark()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

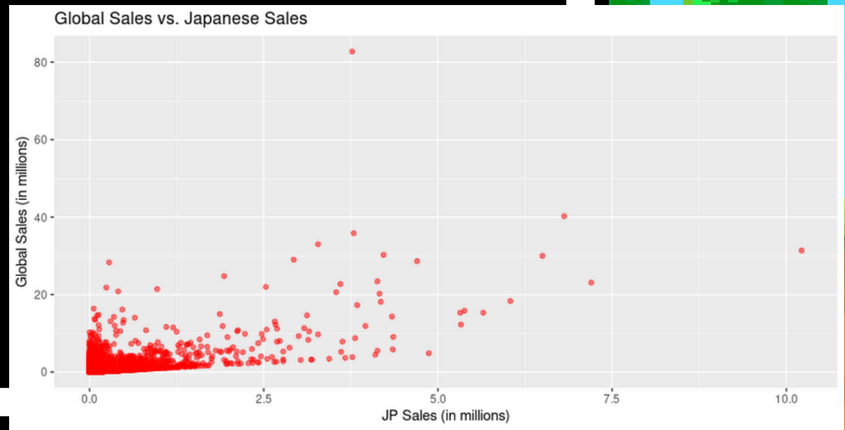


Data Analysis & visualisation

```
# Scatter Plot: Global Sales vs. NA Sales
ggplot(vgsales, aes(x = NA_Sales, y = Global_Sales)) +
  geom_point(color = "blue", alpha = 0.5) +
  labs(title = "Global Sales vs. North American Sales", x = "NA Sales (in millions)", y = "Global
Sales (in millions)")

# Scatter Plot: Global Sales vs. EU Sales
ggplot(vgsales, aes(x = EU_Sales, y = Global_Sales)) +
  geom_point(color = "green", alpha = 0.5) +
  labs(title = "Global Sales vs. European Sales", x =
"EU Sales (in millions)", y = "Global Sales (in millions)")

# Scatter Plot: Global Sales vs. JP Sales
ggplot(vgsales, aes(x = JP_Sales, y = Global_Sales)) +
  geom_point(color = "red", alpha = 0.5) +
  labs(title = "Global Sales vs. Japanese Sales",
 x = "JP Sales (in millions)", y = "Global Sales (in millions)")
```

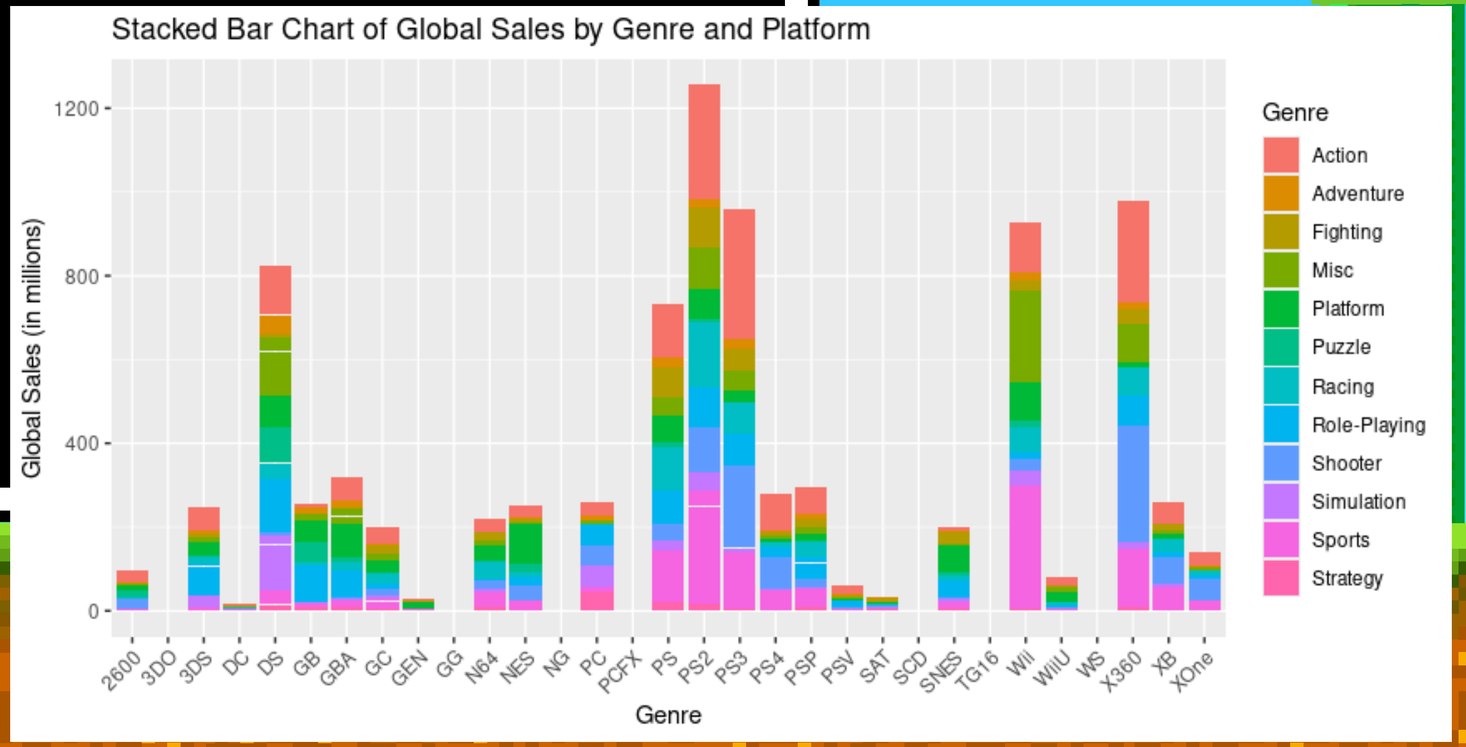






Data Analysis & visualisation

```
# Stacked Bar Chart: Global Sales by Genre and Platform
ggplot(vgsales, aes(x =Platform , y = Global_Sales, fill = Genre)) +
  geom_bar(stat = "identity") +
  labs(title = "Stacked Bar Chart of Global Sales by Genre and Platform",
       x = "Genre",
       y = "Global Sales (in millions)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



THANKYOU

FOR MORE DETAILS

vmgomathisankar@gmail.com

Rajapalayam,TamilNadu 626117