Deliverable 1

Abdulshaheed Alqunber, Rami Bassil, Ekaterina Gorbunova, Khalid Khumayis

Abstract

One of the main branches of the entertainment industry is video games. It was then decided to investigate this area to look for exciting correlations and trends in that industry. The proposed research is based on a dataset of 167,000 video games with a detailed description of each. The purpose of this study is to find a pattern and correlation between ratings of the games, popularity, platforms, genres, and developers. R Studio will be used to analyze the data.

The dataset contains both qualitative and quantitative data. However, not all variables will be used in the research since the focus will be on game genres and how they affect gaming consumption. In terms of cleaning the data, outliers or missing data will be first observed in order to omit unnecessary information. Ultimately, the objective is to find a meaningful correlation between genres of listed games and other variables to make predictions about future releases and games' popularity.

Data Description

The `video-game.csv` data set was found on kaggle.com and was created by a user named `Juttuga Rakesh`. The dataset includes 167,000 observations and 17 columns, i.e. variables, which consist of both categorical and numerical information about different video games such as platform, developer, genre, and year of release. Furthermore, it includes a number of players from different regions, critic scores, and user scores.

The size of the `.csv` file is roughly around 2.1MB. It includes categorical variables such as platform, year of release, genre, publisher, developer, and rating. There are also continuous variables such as the number of players in different regions (North America, Europe, Japan, Others, and Global), Critic and User scores and Critic and User counts, which indicates the number of players and critics. We will also use 'User count' as an indication of sales in units value.

```
After investigating the dataset, we found a lot of missing values for recent games. As seen in the below gra
ph, the number of games released decreases after 2009, which does not make sense. Indeed, releases increase
over time, especially with a rise of game consumption in Asia. This is probably due to the fact that data co
llection was done improperly. For the scope of this deliverable, proportions will be used instead of nominal
values to portray trends. However, we are backtracking the data collection and will update it so that our fi
nal deliverable includes the most recent data.
```

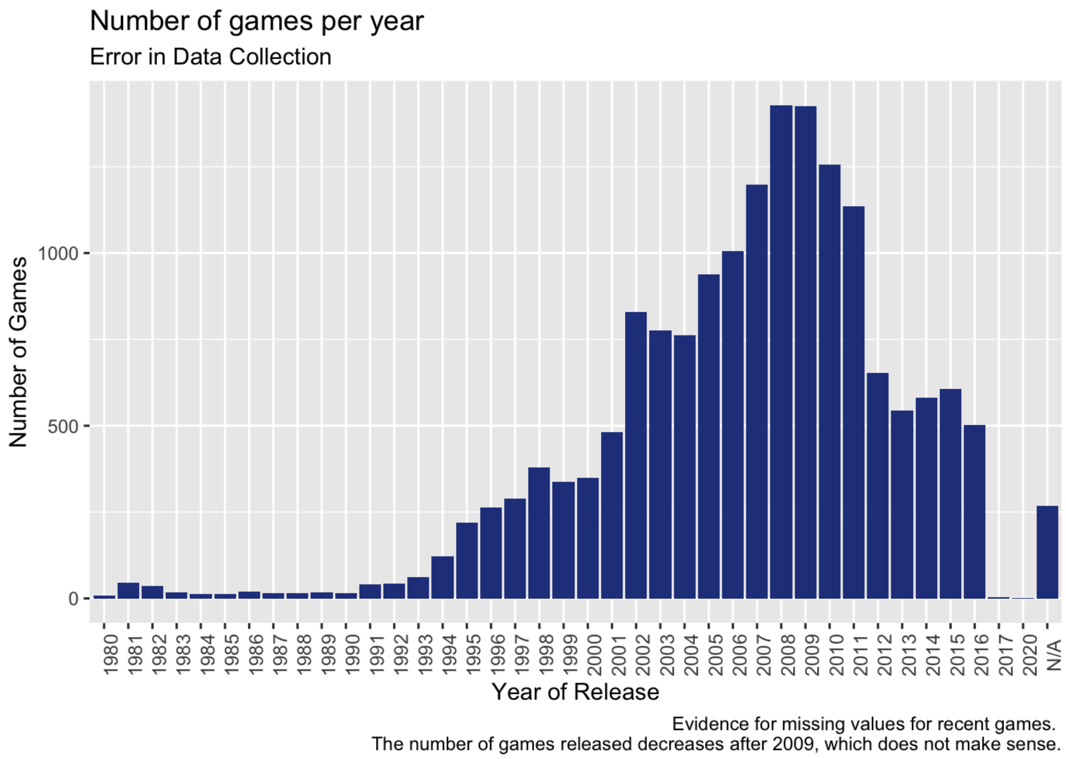
```
#This chunk is only to portray the error in data collection. It will not be used in the final deliverable.

library(tidyverse)
```

```
## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────
───────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.1.0       ✔ purrr   0.3.0
## ✔ tibble  2.1.1       ✔ dplyr   0.8.0.1
## ✔ tidyr   0.8.2       ✔ stringr 1.3.1
## ✔ readr   1.3.1       ✔ forcats 0.3.0
```

```
## ── Conflicts ─────────────────────────────────────────────────────────────────────────────────────────────
─────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
data <- read.csv("video_game.csv")
data %>% group_by (Year_of_Release, Name) %>% ggplot() + geom_bar(aes(Year_of_Release), fill = 'royalblue4')
+
  labs(title = 'Number of games per year', subtitle = 'Error in Data Collection',
       caption = 'Evidence for missing values for recent games.
       The number of games released decreases after 2009, which does not make sense.',
       x = 'Year of Release', y = 'Number of Games') + theme(axis.text.x = element_text(angle = 90, size=9)
)
```



Number of games per year

Error in Data Collection

Evidence for missing values for recent games.
The number of games released decreases after 2009, which does not make sense.

Research Question

General Question:

How do game genres affect gaming industry and popularity (Global players and Critic scores)?

More specific questions to ask:

What genre of games receives the highest proportion of players and how does it relate to the critic score of the games? What is the most played genre for each platform? How do demographics affect game genre consumption? What is the trend for genres for future years?

Motivation:

The video games industry was valued at $78.61 billion in 2017. With so many different genres and platforms to choose from, consumers have leverage over big gaming companies. Therefore, it is essential for those companies to analyze key trends and understand what types of games are the most popular. By analyzing past trends in the industry, especially how genre affects popularity, a prediction on where the industry is heading will help game developers create games that will generate high sale volumes.

Data Import & Cleaning

There were no problems importing the data. However, there were a lot of missing values found in specific fields, especially critic scores and years. This is understandable because not all games in the dataset were critically acclaimed, especially smaller ones. Therefore, it was decided to eliminate games that don't have a year record but keep the rest. More specifically, columns `Critic_Score`, `Critic_Count`, `User_Score`, and `User_Count` got almost 50% of null values in them, that will be filtered out for some of the analysis. However, other studies will require us to keep those values.

Furthermore, some observations were deleted in order for the analysis to make sense. For example, observations between the years 2017 and 2020 were eliminated because the dataset was last updated 2 years ago so it doesn't make sense to include games that were not released back then.

Another observation is that there were repeating entries for many games due to different platforms used for one game, so when grouping by platforms or genres, the data was also grouped by name in order to get rid of repetitive data.

```r
library(tidyverse)
data <- read.csv("video_game.csv")
problems(data)
```

```
0 rows | 1-1 of 4 columns
```

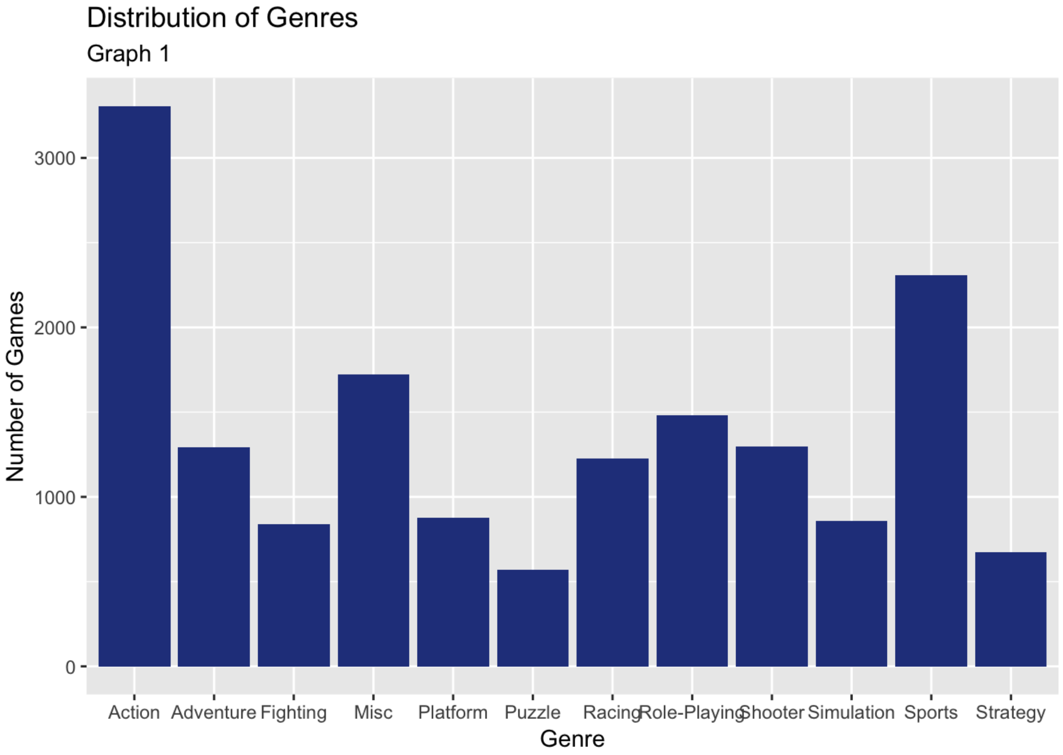
```r
data <- data %>%
  filter(Genre != '', Year_of_Release != 'N/A', Year_of_Release != '2020', Year_of_Release !='2017')
```

Variation of Single Variables

First, a quick study of the variation of single variables was conducted to see their own patterns and distributions. Plots constructed below give information about counts of genres and platforms.

The first figure takes unique values of genres and creates a plot with a count for each on Y-axis and genres themselves on X-axis. Based on this graph, we see that the most popular genre overall is action with 3307 entries in our dataset, followed by ‘Sports’ category with 2306 records. The least frequent appered to be Puzzle and Fighting.

```
data %>% group_by(Genre, Name) %>% ggplot() + geom_bar(aes(Genre), fill = 'royalblue4') + labs(title = 'Dist
ribution of Genres', x = 'Genre', y = 'Number of Games', subtitle = 'Graph 1')
```

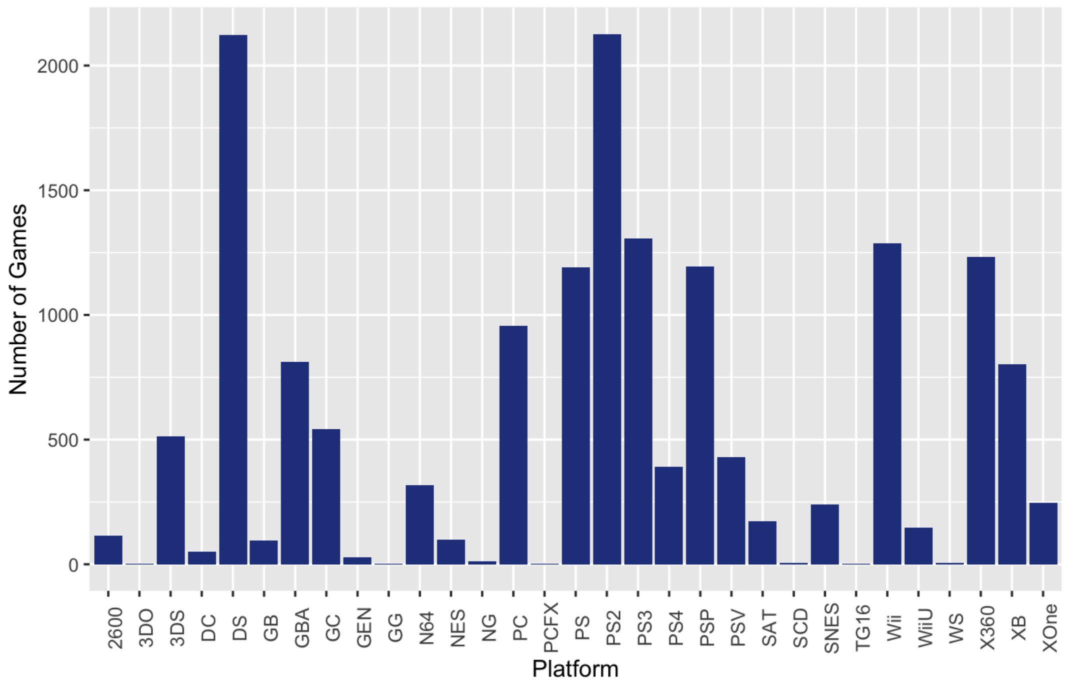


The same analysis was done with platforms to investigate their popularity. Because the data contains information from year 1980, two most popular platforms appeared to be Nintendo DS and PlayStation 2.

```
data %>% group_by(Platform, Name) %>% ggplot() + geom_bar(aes(Platform), fill = 'royalblue4') + labs(title =
'Distribution of Platforms', x = 'Platform', y = 'Number of Games', subtitle = 'Graph 2') + theme(axis.text.
x = element_text(angle = 90, size=9))
```

Distribution of Platforms

Graph 2



Variation between Multiple Variables

Next, a look at how two variables correlate to one another was observed.

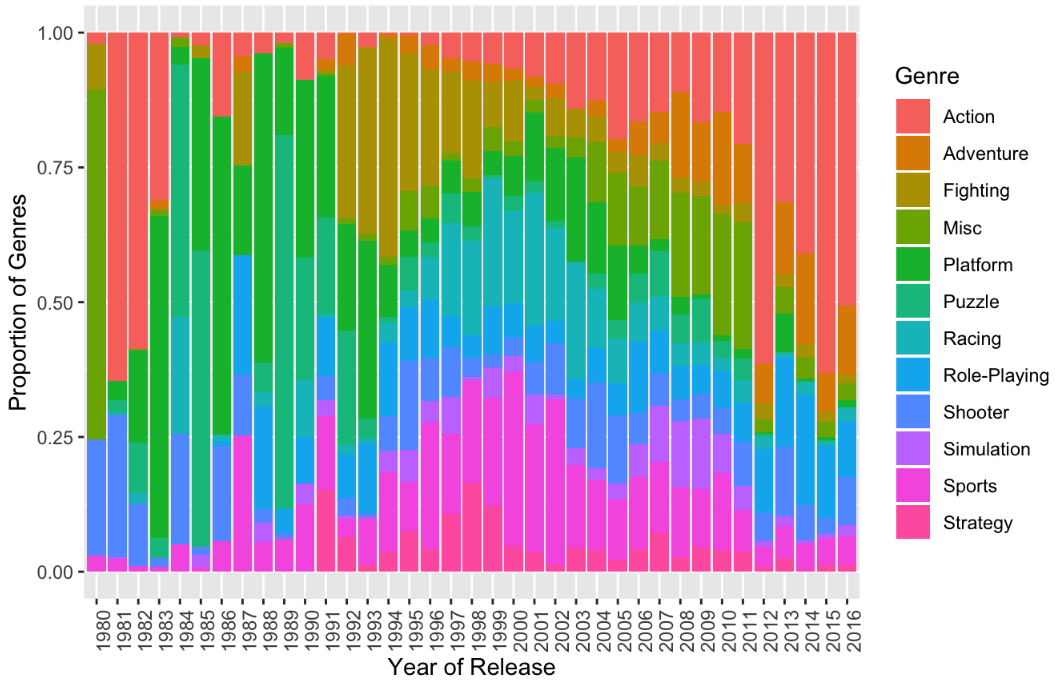
1. Proportion of Genres vs Years

The first plot shows the proportion of genres for each year with years on the x-axis, proportion on the y-axis and colored by genres. The plot tells us a lot about the distribution of genres, especially pointing out the trend of 'Action' inreasing popularity. Since it is the most popular genre overall in our dataset, we decided to look at at individually as well, correlating that one genre with years of release.

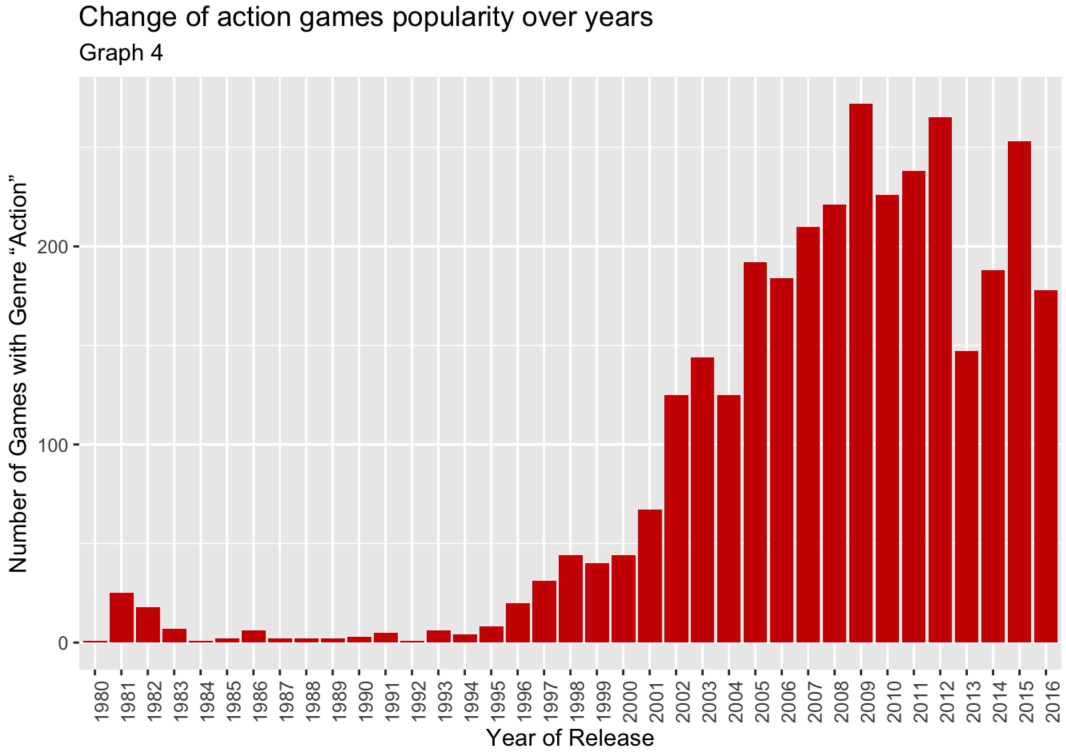
```
data1 <- data %>% group_by(Genre) %>% mutate(sum_each = n()) %>% ungroup() %>% group_by(Year_of_Release, Gen
re) %>% mutate(Proportion_of_Genre = n()/sum_each)

data1 %>% ggplot(aes(Year_of_Release, Proportion_of_Genre, fill = Genre)) + geom_bar(stat="identity", positi
on = 'fill') + labs(title = 'Proportion of Genres for each year of release', subtitle = 'Graph 3', x = 'Year
of Release', y = 'Proportion of Genres') + theme(axis.text.x = element_text(angle = 90, size=9))
```

Proportion of Genres for each year of release

Graph 3



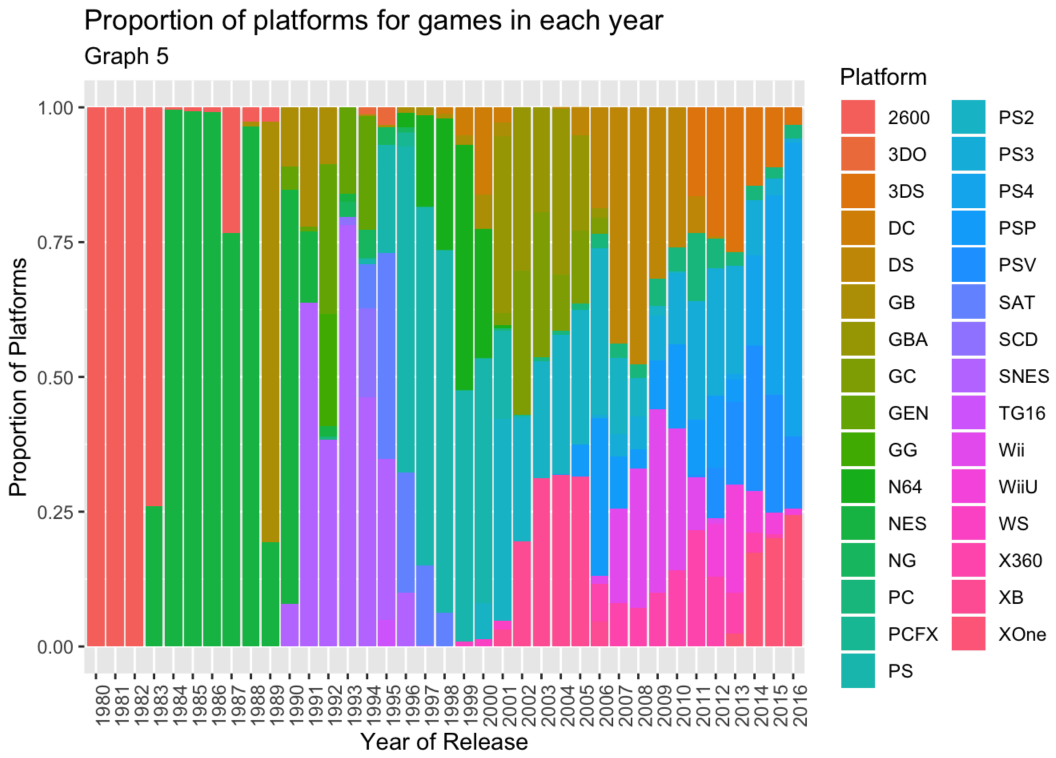
```
#Genre 'Action' over time
data1 %>% filter(Genre == "Action") %>%
  ggplot() + geom_bar(aes(Year_of_Release), fill = 'red3') + labs(title = 'Change of action games popularity
over years', x = 'Year of Release', y = 'Number of Games with Genre “Action”', subtitle = 'Graph 4') + theme
(axis.text.x = element_text(angle = 90, size=9))
```



2. Proportion of Platforms vs Years

The same analysis was done with the proportion of platforms vs years. This plot clearly shows how new platfroms were introduced and some became obsolete, like blue colors indicating PS systems.

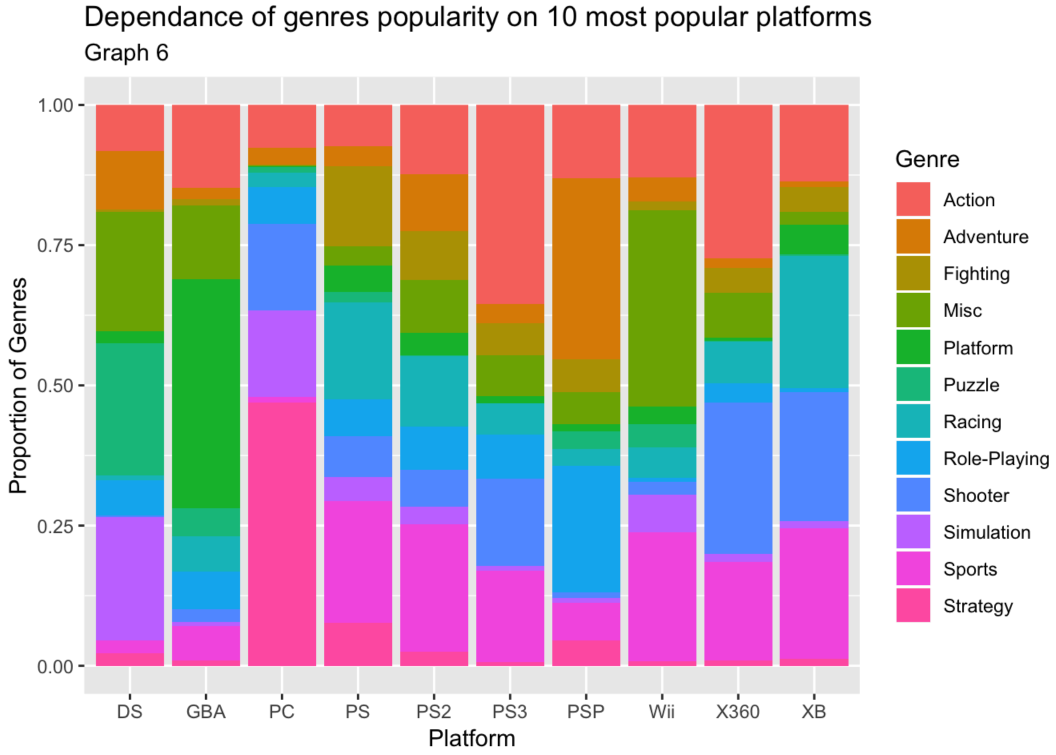
```
data2 <- data %>% group_by(Platform) %>% mutate(sum_each1 = n()) %>% ungroup() %>% group_by(Platform, Year_o
f_Release) %>% mutate(Proportion_of_Plat = n()/sum_each1)

data2 %>% ggplot(aes(Year_of_Release, Proportion_of_Plat, fill = Platform)) + geom_bar(stat="identity", posi
tion = 'fill') + labs(title = 'Proportion of platforms for games in each year', subtitle = 'Graph 5', x = 'Y
ear of Release', y = 'Proportion of Platforms') + theme(axis.text.x = element_text(angle = 90, size=9))
```



3. Proportion of Platforms vs Genres

Finally, we correlated the proportion of genres for 10 most popular platforms, to see if there is any relationship between those two variables and whether a specific platform causes any particular genre to be more popular.

```
data3 <- data %>% group_by(Genre) %>% mutate(sum_each2 = n()) %>% group_by(Genre, Platform) %>%  mutate(prop2 = n()/sum_each2)

data3 %>%  group_by(Platform) %>% filter(n() > 700) %>% ggplot(aes(Platform,prop2, fill=Genre)) + geom_bar(stat = "identity", position = 'fill') + labs(title = 'Dependance of genres popularity on 10 most popular platforms', subtitle = 'Graph 6', x = 'Platform', y = 'Proportion of Genres')
```



Discussion

After analyzing and visualizing the data, we deduced some interesting findings. Variation of single variables gave us information about the most popular genres and platforms in our dataset. As it turns out, ‘Action’ is the most popular genre, as seen in Graph 1. However, it only started to increase in popularity in 1995, when the number of games for this genre released doubled and then started growing rapidly (Graph 4). There is a clear explanation for that as the first PlayStation console, the PS (aka PS1), was released on December 3, 1994. Meanwhile, publishers started creating games for that new console, which started being released in 1995. We can also see statistical evidence of this fact on the graph 5, as the blue color that represents PS starts to appear in the year 1995. This relationship between the introduction of PlayStation and action games cannot be a coincidence and shows that the PS might have ignited the development of a new genre.

That being said, action games were not as high in popularity with the PS as they are with later versions. Instead, the main focus of the first PS platform were sports and racing games, as can be seen on graph 6, but shifted more towards action games with PS2 then PS3. This makes sense because as with every other new trend, it takes some time for them to take off.