Exploring Video Games

April 27, 2022

The dataset used for this challenge is titled 'videogames.csv'.

What topic does the dataset cover?

The dataset covers video game sales across several platforms and regions.

Acknowledgements

The dataset was pulled from Kaggle and can be found here.

Variables

- Name
- Platform
- Year_of_release
- Genre
- Publisher
- NA_Sales (North American sales in millions of units. Note that all sales variables share convention.)
- EU_Sales
- JP_Sales
- Other_Sales
- Global_Sales

Assignment

Complete each of the following questions using R Studio and submit your answers as a detailed report.

1. Describe and visualize the distribution for global sales of video games.
2. Summarize the various platforms of games by average global sales.
   (a) Which 5 platforms have the highest average global sales?
   (b) Which 5 platforms have the least average global sales.
3. Summarize the various platforms by the number of games each has made.
   (a) Which 5 platforms have made the most games?
   (b) Which 5 platforms have made the least games?

   (Make sure you write out full names of platforms for this question, and research platforms you haven't heard of).

4. Summarize the various genres by how many games of each are made. Which 5 genres have made the most number of games?
5. How many games have critic scores? (You will need to us is.na() here)
6. How many games have more than 20 critic scores?
7. Visualize and describe the distribution of critic scores for games with more than 20 critic scores.
8. Compare critic scores across genre for games with more than 20 critic scores. Is there a significant difference? Use pairwise boxplots to support your answer.
9. Compare critic scores across the platforms with the most games (top 5). Only consider games with more than 20 critic scores. Is there a significant difference? Use pairwise boxplots to support your answer.
10. Is there a linear relationship between critic score and global sales?
11. What are the top 5 video games by critic score?
12. What are the top 5 video games by global sales? Which games of these would you be most likely to play?

Solution:

```
### Load packages and csv
library(ggplot2)
library(tidyverse)
vid <- read.csv('/data/datasets/videogames.csv')

### Take a look at the data
glimpse(vid)

### Question 1: Describe the distribution of Global Sales
vid %>% ggplot(aes(x=Global_Sales)) +
  geom_histogram(color='black', fill='green') +
  labs(x="Global Sales in millions of units", title="Distribution for global video
    game sales") +
  theme_classic()

### Question 2A: Platforms with most global sales
vid %>% group_by(Platform) %>%
  summarise(mean=mean(Global_Sales)) %>%
  arrange(desc(mean))

### Question 2B: Platforms with least global sales
vid %>% group_by(Platform) %>%
  summarise(mean=mean(Global_Sales)) %>%
  arrange(mean)

### Question 3A: Platforms that have made the least games
vid %>% group_by(Platform) %>%
  summarise(count=n()) %>%
  arrange(count)

### Question 3B: Platforms that have made the most games
vid %>% group_by(Platform) %>%
  summarise(count=n()) %>%
  arrange(desc(count))

### Question 4: 3 Genres that make the most games
```

```r
vid %>% group_by(Genre) %>%
  summarise(count=n()) %>%
  arrange(desc(count))

### Question 5: How many games have critic scores?
sum(!is.na(vid$Critic_Score))

### Question 6: How many games have more than 20 critic scores?
sum(vid$Critic_Count>20, na.rm=TRUE)

### Question 7: Visualize the distribution for games with more than 20 critic scores
vid %>% filter(Critic_Count>20) %>%
  ggplot(aes(x=Critic_Score)) +
  geom_histogram(color='black', fill='purple') +
  labs(x="Critic Score", title='Distribution for critic scores')+
  theme_classic()

### Question 8: Are critic scores different across genres?
vid %>% filter(Critic_Count>20) %>%
  ggplot(aes(y=Critic_Score, fill=Genre))+
  geom_boxplot() +
  theme_classic()

### Question 9: Are critic scores different across platforms (comparing top 5
    platforms)
top <- c("PS2", "DS", "PS3", "Wii", "X360")
vid %>% filter(Critic_Count>20) %>%
  filter(Platform %in% top) %>%
  ggplot(aes(y=Critic_Score, fill=Platform))+
  geom_boxplot() +
  theme_classic()

### Question 10: Is there a linear relationship between critic score and global sales
    ?
vid %>% filter(Critic_Count>20) %>%
  ggplot(aes(x=Critic_Score, y=NA_Sales, col=Platform))+
  geom_point(alpha=0.3) +
  theme_classic()

### Question 11: Top 5 games by critic score
vid %>% arrange(desc(Critic_Score)) %>% select(Name) %>%
  head(10)

### Question 12: Top 5 games by global sales
vid %>% arrange(desc(Global_Sales)) %>% select(Name) %>%
  head(10)

```