FP_Stat184

Bryan Xiao, Daiwik Kashyap

Table of contents

```r
# Load necessary libraries
library(tidytext)
```

```
Warning: package 'tidytext' was built under R version 4.4.3
```

```r
library(tidyverse)
library(googlesheets4)
library(ggplot2)
library(dcData)
library(knitr)
```

```
Warning: package 'knitr' was built under R version 4.4.3
```

```r
library(tinytex)
library(stringr)
library(scales)

# Load datasets
MostPlayedDataset <- read.csv("~/GitHub/Sec4_FP_BryanXiao_DaiwikKashyap/data/data.csv",
                              header=TRUE, row.names=1)
SteamStoreDataset <- read.csv("~/GitHub/Sec4_FP_BryanXiao_DaiwikKashyap/data/steam.csv")
```

Data Wrangling

We cleaned and merged the datasets Steam Store Games and Most Played Games of All Time, and isolated the top 200 games based on peak players.

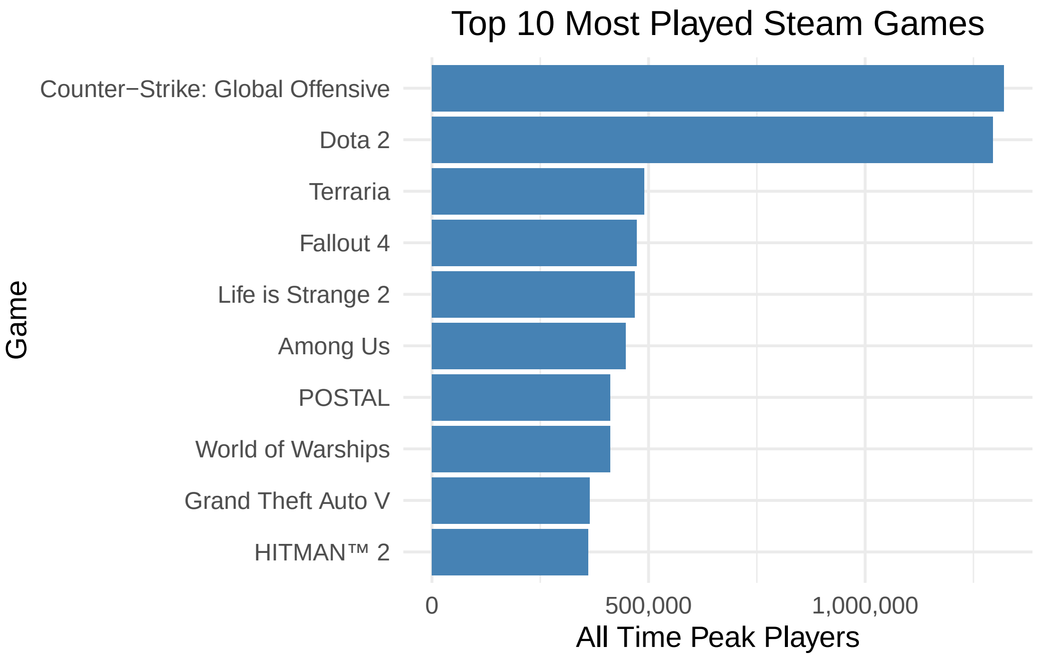
```
# Wrangle most played dataset
MostPlayedDataset$All_time.peak <- str_replace_all(MostPlayedDataset$All_time.peak, ",", "")
MostPlayedDataset$All_time.peak <- as.numeric(as.character(MostPlayedDataset$All_time.peak))

# Clean Steam store data
SteamStoreDataset <- SteamStoreDataset %>% rename(Name = name)

# Merge datasets and tidy
MergedData <- merge(SteamStoreDataset, MostPlayedDataset)

MergedDataTidy <- MergedData %>%
  arrange(desc(All_time.peak), .by_group = TRUE) %>%
  select("Name", "genres", "All_time.peak") %>%
  rename(
    Genres = genres,
    All_Time_Peak = All_time.peak
  ) %>%
  slice(1:400) %>%
  separate(
    col = "Genres",
    sep = ";",
    into = c("Genre 1", "Genre 2", "Genre 3"),
    fill = "right"
  )
```

```
Warning: Expected 3 pieces. Additional pieces discarded in 106 rows [3, 8, 14, 16, 18,
23, 29, 30, 31, 32, 39, 43, 47, 48, 51, 54, 61, 64, 65, 68, ...].
```

Visualizations

Top 10 Most Played Steam Games

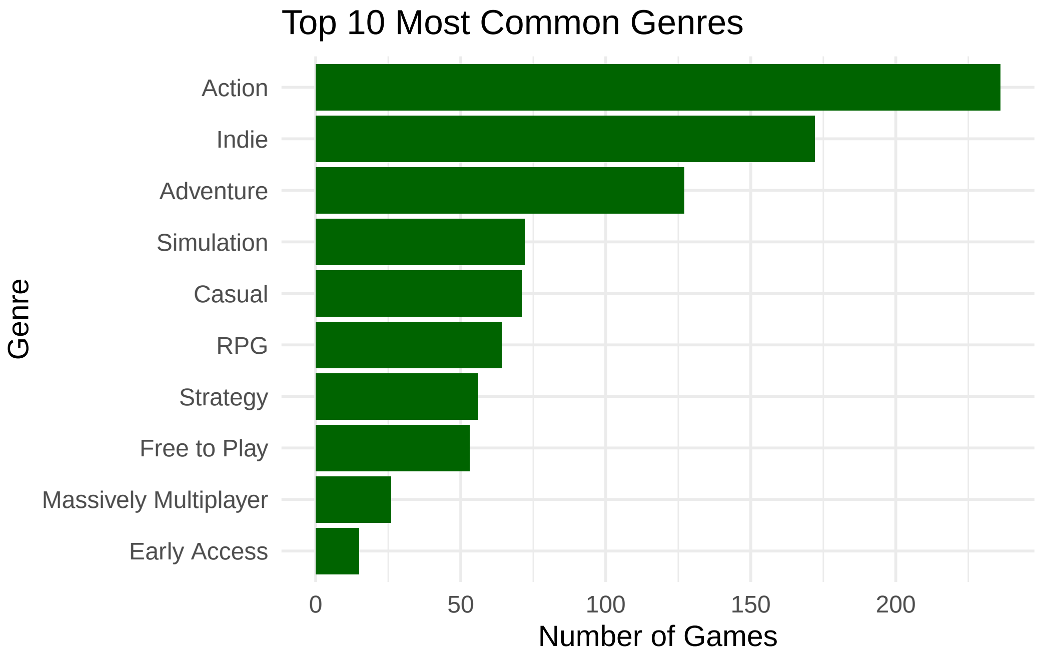
```
Top10Games <- MergedDataTidy %>% arrange(desc(All_Time_Peak)) %>% slice(1:10)

ggplot(Top10Games, aes(x = reorder(Name, All_Time_Peak), y = All_Time_Peak)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 Most Played Steam Games",
       x = "Game",
       y = "All Time Peak Players") +
  scale_y_continuous(labels = comma) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



Count of Games by Primary Genre

```
Top10Genres <- MergedDataTidy %>%
  pivot_longer(cols = starts_with("Genre"), names_to = "GenreType", values_to = "Genre") %>%
  filter(!is.na(Genre)) %>%
  count(Genre, sort = TRUE) %>%
  slice(1:10)
```

```
ggplot(Top10Genres, aes(x = reorder(Genre, n), y = n)) +
  geom_col(fill = "darkgreen") +
  coord_flip() +
  labs(
    title = "Top 10 Most Common Genres",
    x = "Genre",
    y = "Number of Games"
  ) +
  theme_minimal()
```

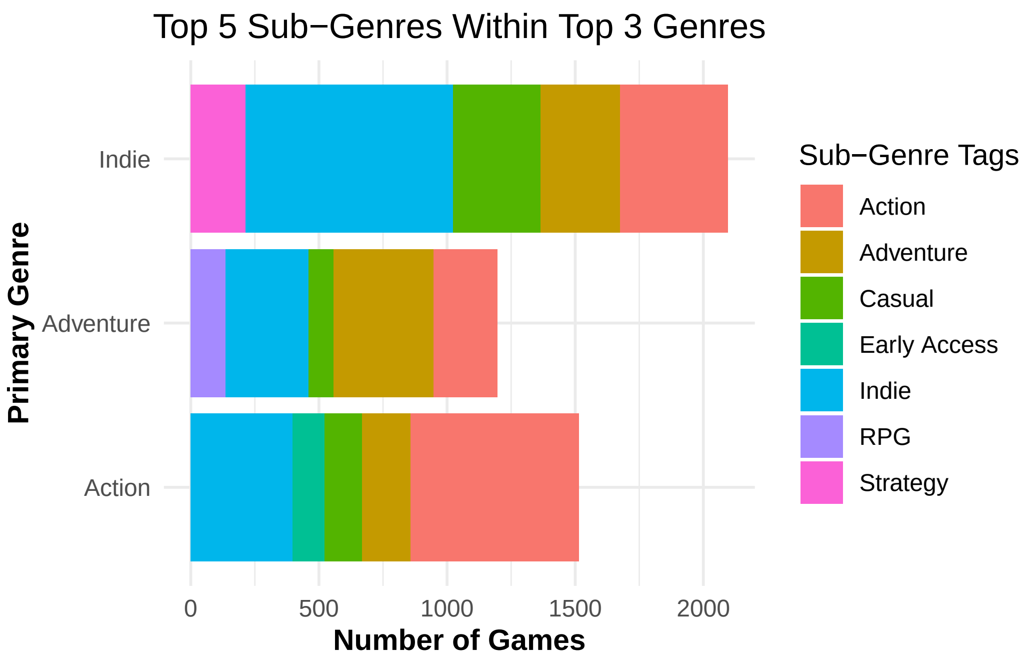


Sub-Genres in Action

Shows the sub genres of action games

```
# Prepare a subset with only top 3 genres
TopGenres <- c("Indie", "Action", "Adventure")

# Filter and explode tags
TopSubGenres <- MergedData %>%
  filter(str_detect(genres, paste(TopGenres, collapse = "|"))) %>%
  select(Name, genres, steamspy_tags) %>%
  separate_rows(genres, sep = ";") %>%
  filter(genres %in% TopGenres) %>%
  separate_rows(steamspy_tags, sep = ";") %>%
  filter(steamspy_tags != "") %>%
  group_by(genres, steamspy_tags) %>%
  summarise(count = n(), .groups = "drop") %>%
  arrange(genres, desc(count)) %>%
  group_by(genres) %>%
  slice_max(count, n = 5)  # top 5 sub-genres for each genre
```

```
ggplot(TopSubGenres, aes(x = genres, y = count, fill = steamspy_tags)) +
  geom_col() +
  coord_flip() +  # ← Flip axes
  labs(
    title = "Top 5 Sub-Genres Within Top 3 Genres",
    x = "Primary Genre",
    y = "Number of Games",
    fill = "Sub-Genre Tags"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  )
```



Conclusion

The Steam gaming reveals multiple intriguing patterns which become visible through visual analysis. Firstly,

The Top 10 Most Played Games chart shows Counter-Strike: Global Offensive and Dota 2 as top games which sustain more than one million peak concurrent players.

The data shows competitive multiplayer games remain popular throughout time according to this information.

The Genre Count chart reveals that Action and Indie genres make up the biggest collection of available Steam titles. The tools which allow indie developers to develop and distribute games in these genres could be the reason for this situation.

The Peak Players by Genre boxplot reveals that user engagement levels differ substantially between different genres. Free to Play and Action games achieve the highest player count peaks compared to all other genres but Nudity and Animation & Modeling genres show very low user engagement.

The data reveals that genre selection acts as a strong indicator to determine both audience size and user participation levels.

This analysis gives us a current view of Steam user preferences together with publishing behavior on the platform. The analysis provides important knowledge about game development and marketing strategies as well as user conduct patterns.

Thank you for taking the time to explore our analysis!