Analysis of Steam Games Dataset

Vasant Kumar Mogia

2024-12-18

About this Dataset

The dataset **All Steam Spiele und deren Metadaten** is a comprehensive collection of data encompassing all games available on the Steam platform, along with their corresponding metadata. It serves as a valuable resource for researchers, developers, and gaming enthusiasts interested in exploring and analyzing the vast Steam gaming ecosystem.

Motivations for Using the Steam Games Dataset

the Steam Games Dataset provides valuable insights into gaming trends, consumer sentiment, and game performance. It helps analyze factors like game popularity, pricing, reviews, and features, enabling better predictions, game development, and targeted marketing strategies. This dataset is essential for understanding the gaming market and improving decision-making within the industry.

Key Features

This dataset includes the following information for each game:

- Title
- Release date
- Developer and publisher details
- Original.Price & Discounted Price
- All.Reviews.Summary
- Popular.Tags
- Supported languages
- Minimum.Requirements and more!

Acknowledgment

Special thanks to the owner of this GitHub repository for compiling and sharing the original dataset.

Objectives

This analysis aims to:

1. Explore and visualize trends in game genres, ratings, and more.
2. Identify key factors that influence game market.
3. Examine the evolution of game releases over time.

Basic Dataset Overview

```
# Load the dataset
steam_data <- read.csv('archive/dataset.csv')

# Take only the first 10,000 rows
steam_data_subset <- head(steam_data, 10000)

# Check the summary of the subset
summary(steam_data_subset)
```

```
##     Title           Original.Price     Discounted.Price   Release.Date
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      Link           Game.Description   Recent.Reviews.Summary
##  Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  All.Reviews.Summary Recent.Reviews.Number All.Reviews.Number
##  Length:10000        Length:10000          Length:10000
##  Class :character    Class :character      Class :character
##  Mode  :character    Mode  :character      Mode  :character
##   Developer           Publisher          Supported.Languages Popular.Tags
##  Length:10000       Length:10000       Length:10000        Length:10000
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##  Game.Features      Minimum.Requirements
##  Length:10000       Length:10000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
# View column names
colnames(steam_data_subset)
```

```
##  [1] "Title"                  "Original.Price"         "Discounted.Price"
##  [4] "Release.Date"           "Link"                   "Game.Description"
##  [7] "Recent.Reviews.Summary" "All.Reviews.Summary"    "Recent.Reviews.Number"
## [10] "All.Reviews.Number"     "Developer"              "Publisher"
## [13] "Supported.Languages"    "Popular.Tags"           "Game.Features"
## [16] "Minimum.Requirements"
```

Load packages

```
# Load all necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(stringr)
library(summarytools)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

Preprocessing:

Handle missing values and incorrect data entries.

```
steam_data_clean <- steam_data_subset %>%
  drop_na() %>%
  distinct() %>%
  mutate(
    Release.Date = ymd(Release.Date)
  ) %>%
  mutate(
    ReleaseYear = year(Release.Date)
  ) %>%
  filter(`Discounted.Price` > 0)
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `Release.Date = ymd(Release.Date)`.
## Caused by warning:
## ! All formats failed to parse. No formats found.
```

```
summary(steam_data_clean)
```

```
##     Title           Original.Price     Discounted.Price    Release.Date
##  Length:587         Length:587         Length:587         Min.   :NA
##  Class :character   Class :character   Class :character   1st Qu.:NA
##  Mode  :character   Mode  :character   Mode  :character   Median :NA
##                                                           Mean   :NaN
##                                                           3rd Qu.:NA
##                                                           Max.   :NA
##                                                           NA's   :587
##      Link           Game.Description   Recent.Reviews.Summary
##  Length:587         Length:587         Length:587
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  All.Reviews.Summary Recent.Reviews.Number All.Reviews.Number
##  Length:587          Length:587            Length:587
##  Class :character    Class :character      Class :character
##  Mode  :character    Mode  :character      Mode  :character
##
##
##
##
##   Developer           Publisher          Supported.Languages Popular.Tags
##  Length:587         Length:587         Length:587          Length:587
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##
##
##
##
##  Game.Features      Minimum.Requirements  ReleaseYear
##  Length:587         Length:587           Min.   : NA
##  Class :character   Class :character     1st Qu.: NA
##  Mode  :character   Mode  :character     Median : NA
##                                          Mean   :NaN
##                                          3rd Qu.: NA
##                                          Max.   : NA
##                                          NA's   :587
```

```
colnames(steam_data_clean)
```

```
##  [1] "Title"                  "Original.Price"         "Discounted.Price"
##  [4] "Release.Date"           "Link"                   "Game.Description"
##  [7] "Recent.Reviews.Summary" "All.Reviews.Summary"    "Recent.Reviews.Number"
## [10] "All.Reviews.Number"     "Developer"              "Publisher"
## [13] "Supported.Languages"    "Popular.Tags"           "Game.Features"
## [16] "Minimum.Requirements"   "ReleaseYear"
```

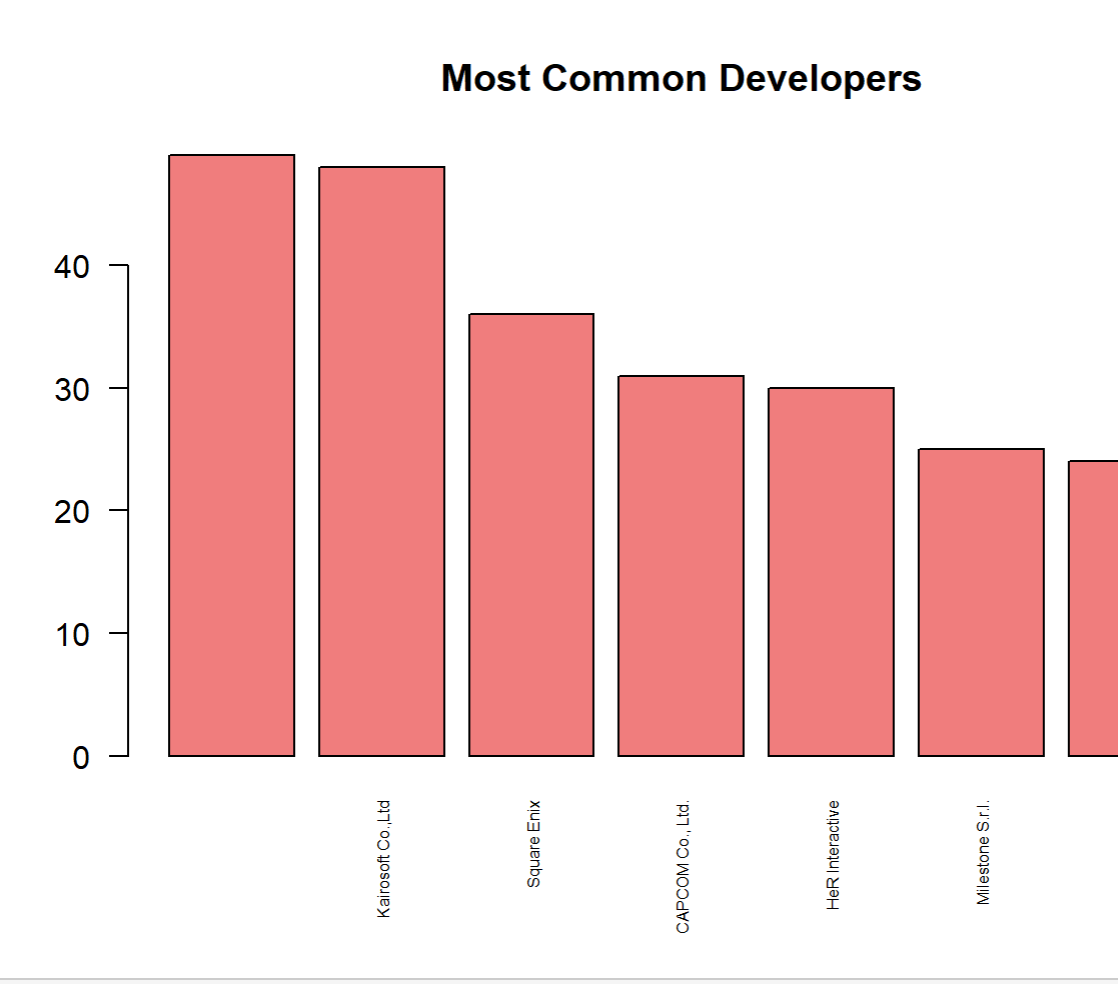
Descriptive Analysis:

```
summary(steam_data_clean)
```

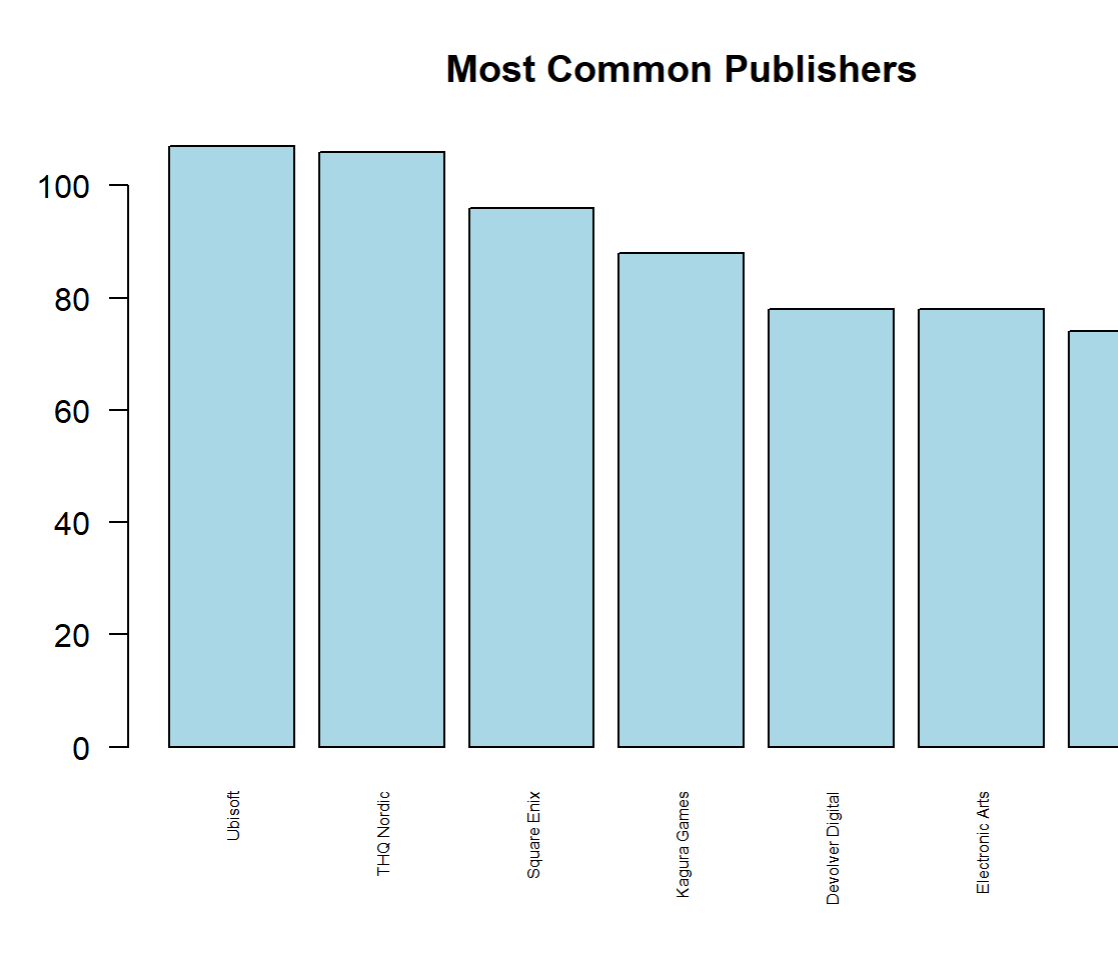
```
##     Title           Original.Price     Discounted.Price    Release.Date
##  Length:587         Length:587         Length:587         Min.   :NA
##  Class :character   Class :character   Class :character   1st Qu.:NA
##  Mode  :character   Mode  :character   Mode  :character   Median :NA
##                                                           Mean   :NaN
##                                                           3rd Qu.:NA
##                                                           Max.   :NA
##                                                           NA's   :587
##      Link           Game.Description   Recent.Reviews.Summary
##  Length:587         Length:587         Length:587
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  All.Reviews.Summary Recent.Reviews.Number All.Reviews.Number
##  Length:587          Length:587            Length:587
##  Class :character    Class :character      Class :character
##  Mode  :character    Mode  :character      Mode  :character
##
##
##
##
##   Developer           Publisher          Supported.Languages Popular.Tags
##  Length:587         Length:587         Length:587          Length:587
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##
##
##
##
##  Game.Features      Minimum.Requirements  ReleaseYear
##  Length:587         Length:587           Min.   : NA
##  Class :character   Class :character     1st Qu.: NA
##  Mode  :character   Mode  :character     Median : NA
##                                          Mean   :NaN
##                                          3rd Qu.: NA
##                                          Max.   : NA
##                                          NA's   :587
```

Most Common Developers and Publishers

```
# Top 10 developers
top_developers <- sort(table(steam_data_subset$Developer), decreasing=TRUE)[2:8]
barplot(top_developers, main="Most Common Developers", col="lightcoral", las=2, cex.names=0.5)
```




```
# Top 10 publishers
top_publishers <- sort(table(steam_data_subset$Publisher), decreasing=TRUE)[2:8]
barplot(top_publishers, main="Most Common Publishers", col="lightblue", las=2, cex.names=0.5)
```

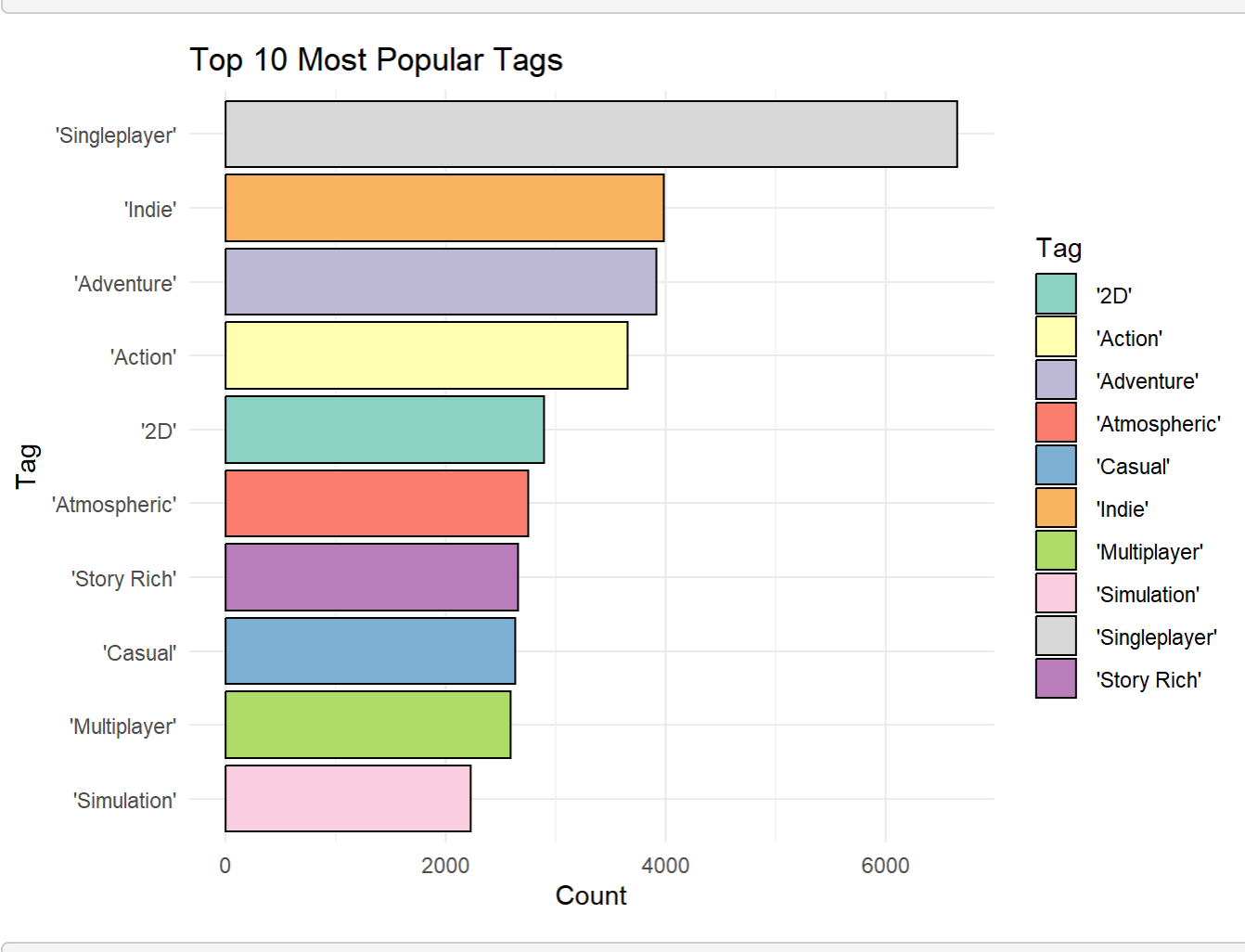


Popular Tags and Game Features

```
# Most popular tags
popular_tags <- unlist(strsplit(as.character(steam_data_subset$Popular.Tags), ","))

popular_tag_table <- table(popular_tags)

top_10_tags <- head(sort(popular_tag_table, decreasing = TRUE), 10)

top_10_tags
```

```
## popular_tags
##  'Singleplayer'         'Indie'     'Adventure'        'Action'            '2D'
##            6654            3985            3916            3658            2898
##   'Atmospheric'    'Story Rich'        'Casual'   'Multiplayer'    'Simulation'
##            2752            2658            2631            2591            2229
```

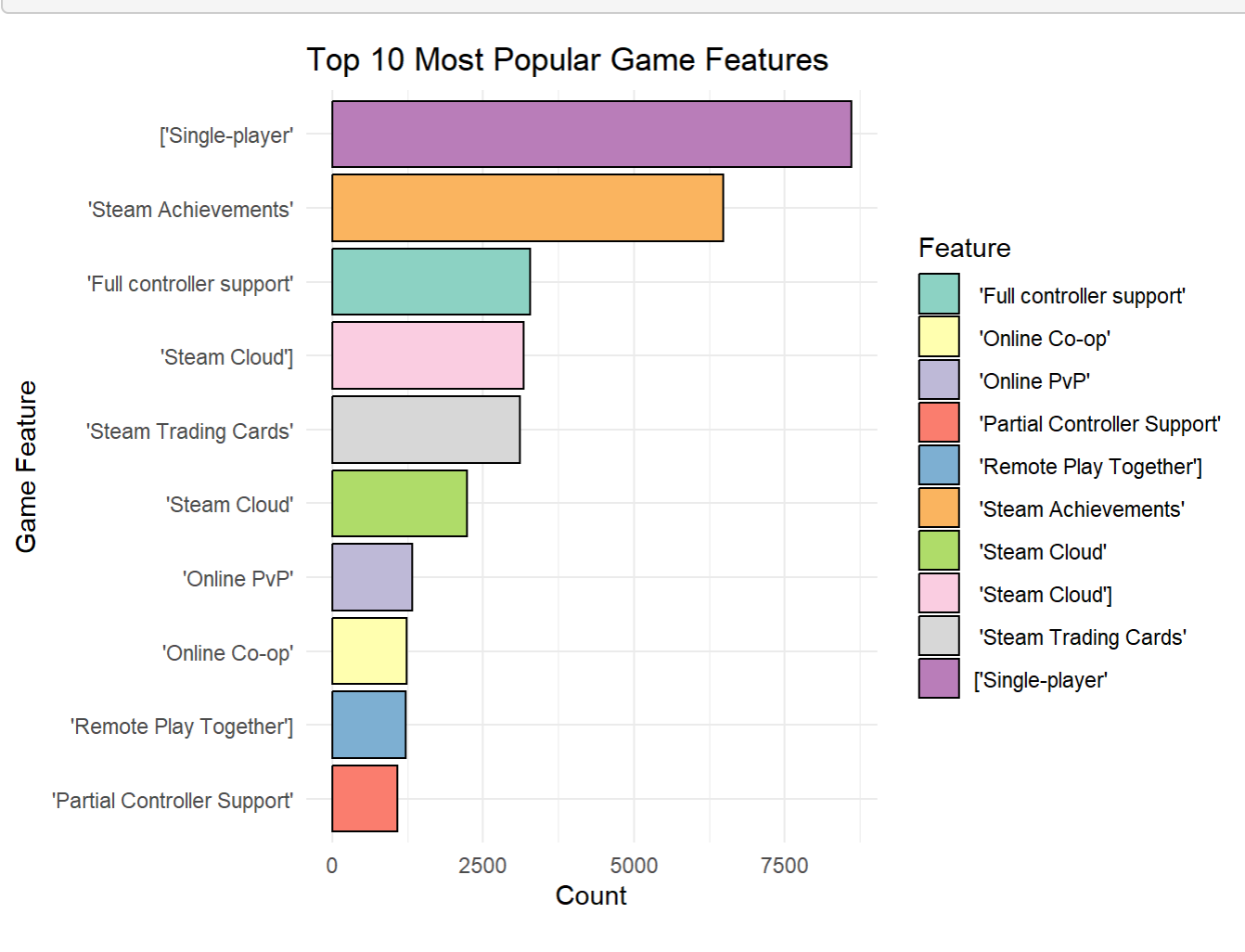
```
library(ggplot2)

# Prepare the top 10 most popular tags
top_10_tags_df <- data.frame(Tag = names(top_10_tags), Count = as.numeric(top_10_tags))

# Bar plot for top 10 most popular tags
ggplot(top_10_tags_df, aes(x = reorder(Tag, Count), y = Count, fill = Tag)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Top 10 Most Popular Tags", x = "Tag", y = "Count") +
  coord_flip() +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal()
```




```
# Most popular game features
game_features <- unlist(strsplit(as.character(steam_data_subset$Game.Features), ","))

game_feature_table <- table(game_features)

top_10_features <- head(sort(game_feature_table, decreasing = TRUE), 10)

top_10_features
```

```
## game_features
##             ['Single-player'          'Steam Achievements'
##                         8602                          6485
##     'Full controller support'                'Steam Cloud']
##                         3281                          3174
##         'Steam Trading Cards'                 'Steam Cloud'
##                         3112                          2237
##                  'Online PvP'                'Online Co-op'
##                         1326                          1237
##       'Remote Play Together']  'Partial Controller Support'
##                         1222                          1081
```

```
# Prepare the top 10 most popular game features
top_10_features_df <- data.frame(Feature = names(top_10_features), Count = as.numeric(top_10_features))

# Bar plot for top 10 most popular game features
ggplot(top_10_features_df, aes(x = reorder(Feature, Count), y = Count, fill = Feature)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Top 10 Most Popular Game Features", x = "Game Feature", y = "Count") +
  coord_flip() +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal()
```

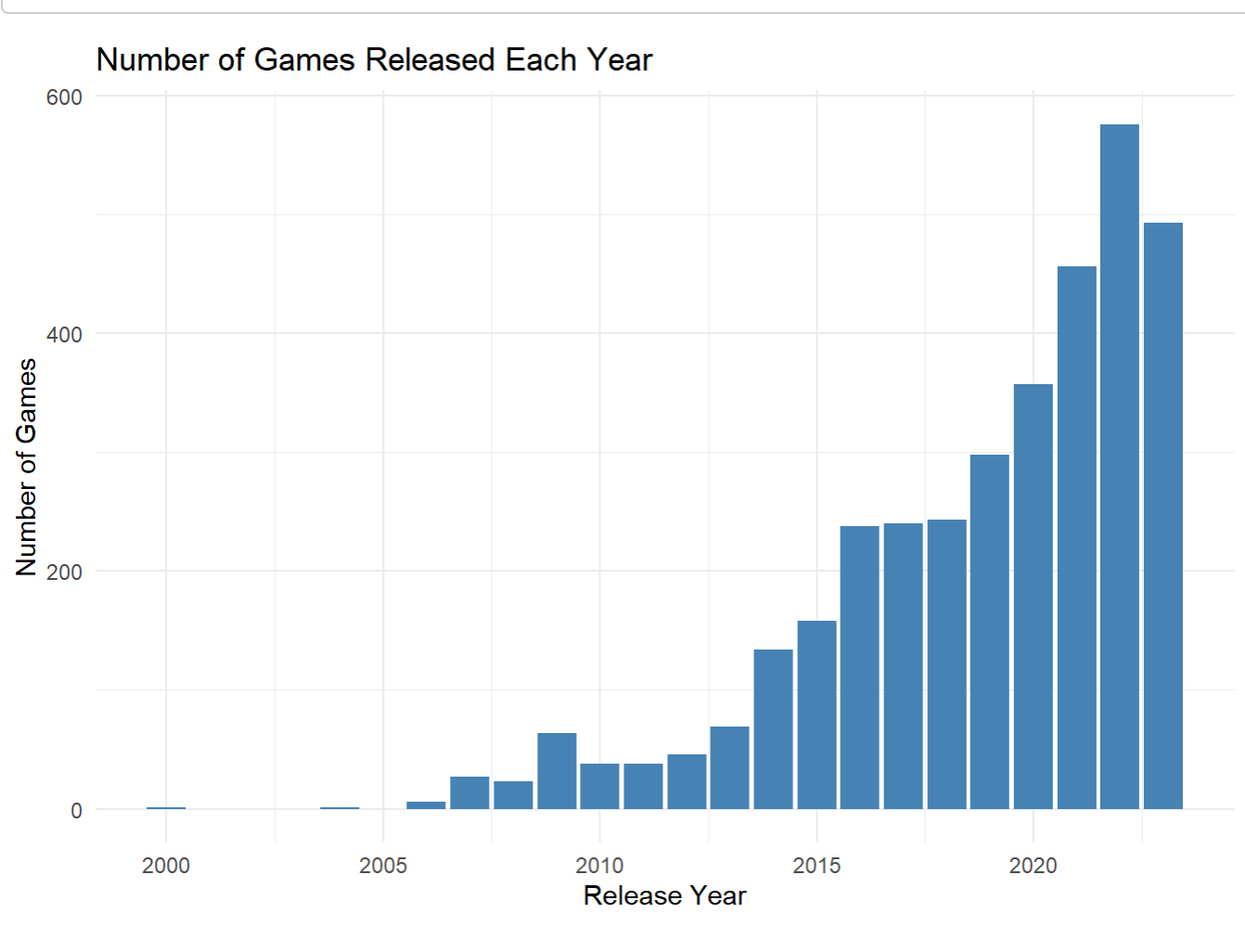


Release Date Analysis

```
library(ggplot2)
library(lubridate)

steam_data_subset %>%
  mutate(
    Release.Date = mdy(Release.Date),
    ReleaseYear = year(Release.Date)
  ) %>%
  ggplot(aes(x = ReleaseYear)) +
  geom_bar(fill = "steelblue") +
  labs(
    title = "Number of Games Released Each Year",
    x = "Release Year",
    y = "Number of Games"
  ) +
  theme_minimal()
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `Release.Date = mdy(Release.Date)`.
## Caused by warning:
## !  6476 failed to parse.
```

```
## Warning: Removed 6494 rows containing non-finite outside the scale range
## (`stat_count()`).
```



Games with Highest Reviews

```
overwhelmingly_positive_games <- subset(steam_data_subset, All.Reviews.Summary == "Overwhelmingly Positive")

library(dplyr)

selected_columns <- overwhelmingly_positive_games %>%
  select(Title, Original.Price, Discounted.Price, Release.Date, Recent.Reviews.Summary)

head(selected_columns, 10)
```

```
##                       Title Original.Price Discounted.Price Release.Date
## 22              Phasmophobia          $7.99            $7.99 18 Sep, 2020
## 24             DAVE THE DIVER         $10.49           $10.49 28 Jun, 2023
## 42   Euro Truck Simulator 2         $10.99           $10.99 12 Oct, 2012
## 47            Stardew Valley          $9.99            $9.99 26 Feb, 2016
## 49 American Truck Simulator         $10.99           $10.99  2 Feb, 2016
## 59                  RimWorld         $16.49           $16.49 17 Oct, 2018
## 61                  Terraria          $6.99            $6.99 16 May, 2011
## 64        Deep Rock Galactic         $14.49           $14.49 13 May, 2020
## 65               BeamNG.drive         $12.49           $12.49 29 May, 2015
## 66                Dead Cells         $12.49            $7.49  6 Aug, 2018
##     Recent.Reviews.Summary
## 22           Very Positive
## 24 Overwhelmingly Positive
## 42 Overwhelmingly Positive
## 47 Overwhelmingly Positive
## 49 Overwhelmingly Positive
## 59 Overwhelmingly Positive
## 61 Overwhelmingly Positive
## 64 Overwhelmingly Positive
## 65 Overwhelmingly Positive
## 66 Overwhelmingly Positive
```

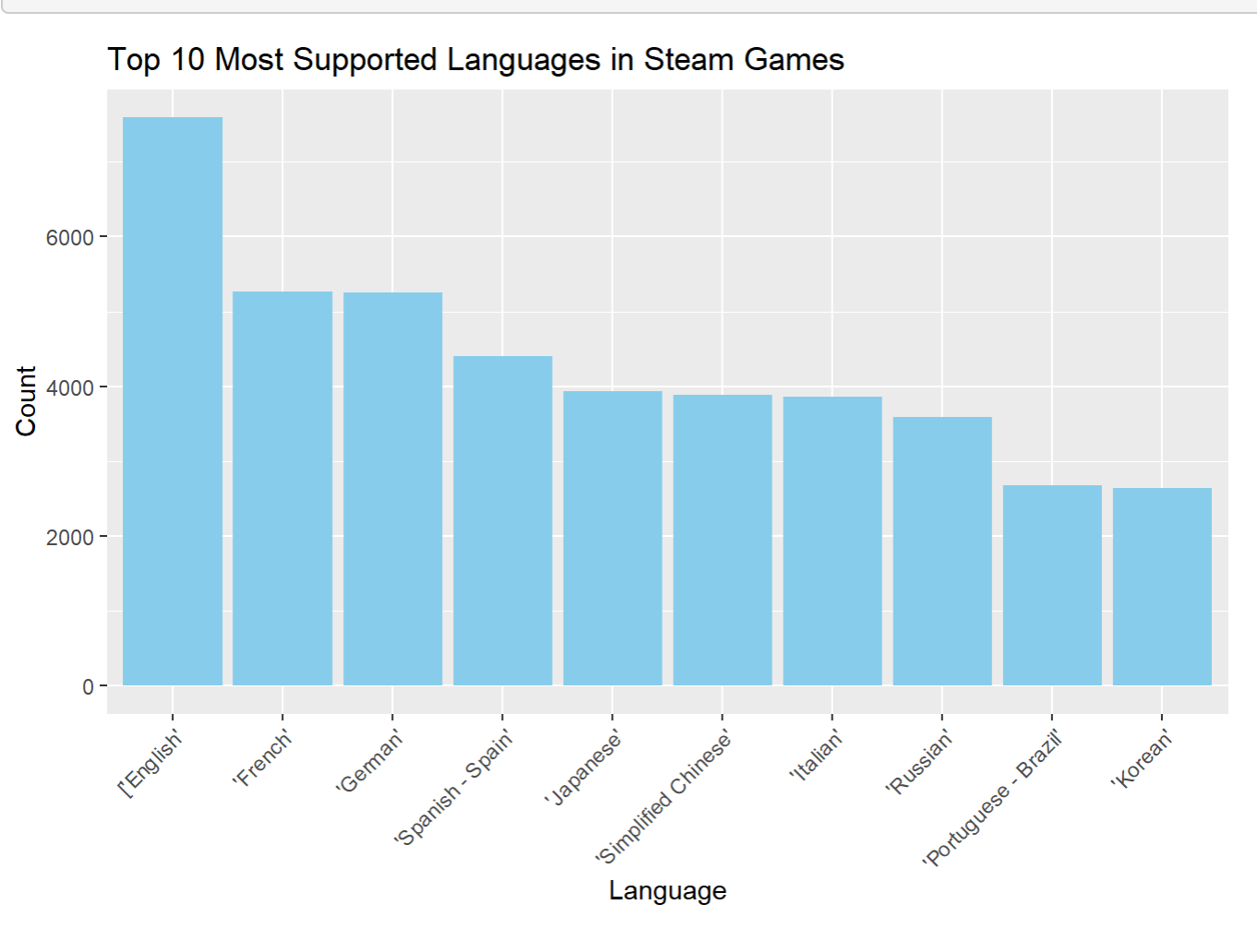
```
library(ggplot2)
```

Games with Most Supported Languages

```
languages <- unlist(strsplit(steam_data_subset$Supported.Languages, ","))

languages <- trimws(languages)

language_counts <- table(languages)

sorted_languages <- sort(language_counts, decreasing = TRUE)

head(sorted_languages, 10)
```

```
## languages
##            ['English'              'French'              'German'
##                  7591                  5262                  5249
##     'Spanish - Spain'            'Japanese'  'Simplified Chinese'
##                  4397                  3936                  3885
##             'Italian'             'Russian' 'Portuguese - Brazil'
##                  3856                  3583                  2678
##              'Korean'
##                  2639
```

```
library(ggplot2)

language_df <- data.frame(Language = names(sorted_languages), Count = as.integer(sorted_languages))

ggplot(language_df[1:10, ], aes(x = reorder(Language, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  xlab("Language") +
  ylab("Count") +
  ggtitle("Top 10 Most Supported Languages in Steam Games") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Games with Most asked Minimum.Requirements

```
min_requirements <- unlist(strsplit(as.character(steam_data_subset$Minimum.Requirements), "[,\\s]+"))

min_requirements <- trimws(min_requirements)

min_requirements_table <- table(min_requirements)

top_min_requirements <- head(sort(min_requirements_table, decreasing = TRUE), 10)

top_min_requirements
```

```
## min_requirements
##          OS: |  Window                   pace                      y
##                   5063                   4758                   3737
##        or and operating         a 64-bit proce                Require
##                   3610                   3580                   3490
##     tem | OS: |  Window            10 | Proce pace | Additional Note
##                   3043                   1367                   1264
##              7 | Proce
##                   1090
```

Review analysis

```
unique_review_categories <- unique(steam_data_subset$All.Reviews.Summary)
print(unique_review_categories)
```

```
## [1] "Very Positive"           "Mixed"
## [3] ""                        "Mostly Positive"
## [5] "Overwhelmingly Positive" "Mostly Negative"
## [7] "Positive"
```

```
library(ggplot2)

steam_data_filtered <- steam_data_subset[steam_data_subset$All.Reviews.Summary != "", ]

review_counts <- as.data.frame(table(steam_data_filtered$All.Reviews.Summary))
colnames(review_counts) <- c("ReviewCategory", "Count")

# Create a pie chart using ggplot2
ggplot(review_counts, aes(x = "", y = Count, fill = ReviewCategory)) +
  geom_bar(stat = "identity", color = "black", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Review Summary Distribution") +
  scale_fill_brewer(palette = "Set3") +
  theme_void() +
  theme(legend.title = element_blank())
```