



Steam Data Explorer: What Makes a Popular Game?

A Data Analysis Project

By Ahmad Azeez

Overview

This is a documentation of my findings from the data analysis project “**Steam Data Explorer: What Makes a Popular**”.

It tries to answer the question:

- Which tags appear most often in popular games?

Tools Used

RStudio and the **R (Python)** language were used for this project.

Libraries Used

- tidyverse
- stringr
- tidyr

Important Note

All the information provided in this documentation is based on the data obtained from Aleksandr Antonov’s [“Steam games complete dataset”](#) on [Kaggle](#).

Table Of Contents

Overview.....	1
Tools Used.....	1
Libraries Used.....	1
Important Note.....	1
Table Of Contents.....	2
Data Preparation.....	3
What Are The Most Common Genres Among Top-Rated Games?.....	4
Does Price Affect Popularity?.....	7
Are Newer Games More Popular, or Do Older Games Still Dominate?.....	8
Which Publishers and Devs Release The Most High-Rated Games?.....	9
Which Tags and Features Appear Most Often in Popular Games?.....	10
Conclusion.....	11
References.....	12

Data Preparation

Because any information could be extracted from the dataset, it had to be loaded in RStudio first.

```
# Importing the csv file
steam_games <- read.csv("steam_games.csv", header = TRUE, sep=",")
steam_games
```

This code reads the csv file (dataset) and loads it in RStudio.

```
# Installing libraries I will use
install.packages("tidyverse")
install.packages("stringr")
install.packages("tidyr")

# Loading the libraries
library(tidyverse)
library(stringr)
library(tidyr)
```

This installs and loads the libraries that will be used.

- **tidyverse:** For data wrangling and plotting
- **stringr:** Part of the tidyverse library, for working with strings
- **tidyr:** For tidying data, also part of the tidyverse library

Which tags appear most often in popular games?

```
# Cleaning and extracting parts
cleaned_reviews <- steam_games %>%
  filter(!is.na(all_reviews)) %>%
  mutate(
    rating_label = str_extract(all_reviews, "^[^,]+"), # Gets
    "Positive" for example
    percent_positive = str_extract(all_reviews, "\\d+%") %>%
    str_remove("%") %>%
    as.numeric() / 100) # 92% to 0.92 for example
```

This extracts the reviews into **2** separate columns “**rating_label**” and “**percent_positive**”. This makes it easier to find the highly rated games.

“%>%” is from the “**tidyverse**” library and enables function chaining.

For example, “**x %>% f()**” is the same thing as **f(x)**. In this case, it is like saying “Take **steam_games**, then filter out any rows where **reviews** are empty. Next, create (or overwrite) the **rating_label** column by extracting the first chunk of text before the first comma using the regex **^[^,]+**. Do the same for **percent_positive**, but this time extract the percentage value (like '92%'), remove the % symbol, and divide the number by **100** to turn it into a decimal.”

```
# For top rated games
top_rated <- cleaned_reviews %>% filter(percent_positive >= 0.85)
```

This line of code creates a new list called **top_rated** that only includes games where **85%** or more of the reviews are positive.

```
# For the tags, I'll separate the tags into rows
tags <- top_rated %>%
  separate_rows(popular_tags, sep = ",") %>%
  mutate(popular_tags = str_trim(popular_tags)) %>%
  filter(popular_tags != "")
```

```
# Most common tags
tag_counts <- tags %>%
  count(popular_tags, sort = TRUE)
```

This takes the list of top-rated games and breaks up the tags (like "**Multiplayer**, **Indie**, **Action**") so that each tag gets its own row. Then it cleans up any extra spaces and removes empty tags.

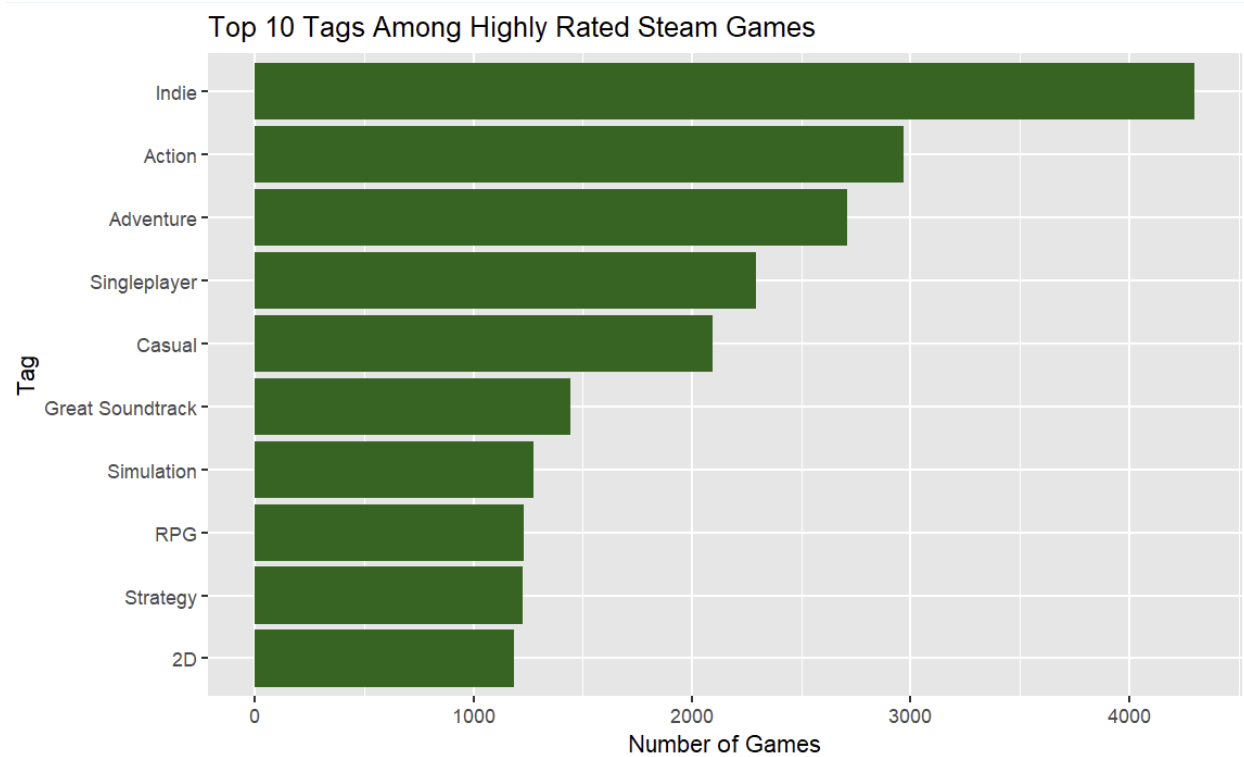
After that, it counts how often each tag appears, and sorts the list so the most common tags are at the top.

```
# Plotting the top 10
tag_counts %>%
  top_n(10, n) %>%
  ggplot(aes(x = reorder(popular_tags, n), y = n)) +
  geom_col(fill = "darkgreen") +
  coord_flip() +
  labs(
    title = "Top 10 Tags Among Highly Rated Steam Games",
    x = "Tag",
    y = "Number of Games"
  )
```

This then creates a **bar chart** that shows the **top 10** most common **tags** from the highest-rated Steam games.

In more detail, it:

- Picks the top 10 tags.
- Plots them as bars, with the most frequent tags at the top.
- Flips the chart sideways so it's easier to read.
- Adds a title and labels to make the chart clear.



According to the dataset used, Indie, Action, Adventure, Singleplayer, Casual, Great Soundtrack, Simulation, RPG, Strategy, and 2D are the top 10 tags that appear most often in popular games.

References

Dataset:

<https://www.kaggle.com/datasets/fronkongames/steam-games-dataset?resource=download>