Steam Playtime Analysis

Data Science · Mr. Merrick · October 22, 2025

You've been hired by Valve as a junior data scientist to explore trends in Steam game playtime data. Your mission: ask clear questions, explore patterns, and use R to uncover simple, data-driven insights.

Overview

Using the **Steam Playtime Dataset**, conduct open-ended exploratory analysis. Use R to summarize, visualize, and describe patterns in how players spend their time across games.

Guiding Research Questions (Choose, Combine, or Extend)

Pick at least **three** prompts—or design your own.

Q1. Dataset Overview.
How many rows (observations) and variables (columns) are there? What are the mean, median, spread, and range of playtimes?

Q2. Overall Playtime Distribution.
Visualize and describe the overall distribution of `Hours`. Is it symmetric or skewed? Do a few games dominate total time?

Q3. Specific Game Distribution (Skyrim).
Focus on *The Elder Scrolls V: Skyrim* (or a game of your choosing). Visualize the distribution of `Hours` for Skyrim and describe the pattern.

Q4. Playtime Across a Franchise (Call of Duty).
How do *Call of Duty* games differ in playtime distributions? Use string filtering (e.g., `grepl`) to gather all *Call of Duty* titles. Compare the distributions of `Hours` across those games.

Q5. Top Games by Median Playtime.
Which games have the highest *median* playtime? Show the top 10–15 and include a distribution plot so variation is visible.

Q6. Most Popular Games (by Player Count).
Which games appear most often in the dataset (i.e., have the most rows/users)? Plot the distribution of player counts for all games.

Q7. Popularity vs. Playtime.
Create a scatter plot of player counts vs. median playtime. Does there appear to be an association between the variables?

Q8. Extending Further. Since there are only two variables in the dataset we can only explore how 'name' explains variation in playtime. Can you think of some other measurable variables that might explain the variation in playtime?

Dataset Description and Context

The dataset you'll use summarizes how long Steam users have played different games. Each row represents a player–game pair, meaning a single user's playtime for one game.

- **Name:** Title of the video game (e.g., *The Elder Scrolls V: Skyrim*)
- **Hours:** Number of hours the player has spent on that game

This dataset is adapted from the public Steam Playtime Dataset on Kaggle, simplified for beginner analysis. It's designed to let you explore patterns of popularity, engagement, and variation in playtime across titles.

Solution Key (R Script)

```r
# ========================
# Steam Playtime Analysis - Solution Script
# Dataset columns: Name, Hours
# ========================

# --- Libraries ---
library(readr)
library(dplyr)
library(ggplot2)
library(janitor)
library(scales)
library(stringr)

# --- 1) Load & clean ---
# Put the CSV in your working directory. Rename if needed.
st <- read_csv("steam.csv", show_col_types = FALSE)
st <- clean_names(st) # -> name, hours

# ----------------------------
# Q1. Dataset Overview
# ----------------------------
print(dim(st))
print(names(st))

summary_table <- st %>%
  summarise(
    n_rows = n(),
    n_variables = ncol(st),
    mean_hours = mean(hours),
    median_hours = median(hours),
    sd_hours = sd(hours),
    min_hours = min(hours),
    p25_hours = quantile(hours, 0.25),
    p75_hours = quantile(hours, 0.75),
    max_hours = max(hours)
  )
print(summary_table)

# ----------------------------
# Q2. Overall Playtime Distribution
# ----------------------------
# Linear-scale histogram
ggplot(st, aes(x = hours)) +
  geom_histogram(bins = 30, color = "white", fill = "skyblue") +
  scale_x_continuous(labels = comma) +
  labs(
    title = "Distribution of Game Playtime (Hours)",
    x = "Hours Played",
    y = "Number of Records"
  )

# Log-scale histogram (skew-friendly)
ggplot(st, aes(x = hours)) +
  geom_histogram(bins = 30, color = "white", fill = "orange") +
```

```r
  scale_x_log10(labels = comma) +
  labs(
    title = "Distribution of Game Playtime (Log Scale)",
    x = "Hours Played (log10)",
    y = "Number of Records"
  )

# ------------------------------
# Q3. Skyrim: Specific Game Distribution
# ------------------------------
st_skyrim <- st %>%
  filter(grepl("skyrim", name, ignore.case = TRUE))

if (nrow(st_skyrim) > 0) {
  # Summary
  print(st_skyrim %>% summarise(
    n_players = n(),
    med_hours = median(hours),
    mean_hours = mean(hours),
    p90_hours = quantile(hours, 0.90)
  ))

  # Density + rug
  ggplot(st_skyrim, aes(x = hours)) +
    geom_histogram(bins = 30, color = "white", fill = "mediumseagreen") +
    scale_x_continuous(labels = comma) +
    labs(
      title = "Playtime Distribution - The Elder Scrolls V: Skyrim",
      x = "Hours Played",
      y = "Number of Players"
    )

} else {
  message("Skyrim not found in dataset by the simple filter.")
}

# ------------------------------
# Q4. Franchise Comparison - Call of Duty
# ------------------------------
# Use grepl / base R style as requested
cod_rows <- st %>% filter(grepl("call of duty", name, ignore.case=TRUE))

if (nrow(cod_rows) > 0) {
  # Order titles by median hours to keep plots tidy
  cod_summary <- cod_rows %>%
    group_by(name) %>%
    summarise(
      n_players = n(),
      median_hours = median(hours),
      mean_hours = mean(hours)
    ) %>%
    arrange(desc(median_hours))

  print(cod_summary)

  # Boxplots of hours per title (log scale to show spread)
```

```r
  ggplot(cod_rows,
         aes(x = reorder(name, hours, FUN = median), y = hours)) +
    geom_boxplot(fill = "lightblue", outlier.alpha = 0.3) +
    coord_flip() +
    labs(
      title = "Call of Duty: Playtime Distributions by Title",
      x = "Title",
      y = "Hours Played"
    )
} else {
  message("No Call of Duty titles matched with grepl('call of duty', ...).")
}

# ------------------------------
# Q5. Top Games by Median Playtime (+ variation)
# ------------------------------
top_n <- 15
game_summary <- st %>%
  group_by(name) %>%
  summarise(
    n_players = n(),
    median_hours = median(hours),
    mean_hours = mean(hours)
  ) %>%
  filter(n_players>20) %>%
  arrange(desc(median_hours))

top_games <- game_summary %>% slice_head(n = top_n)
print(top_games)

# Bar chart of medians
top_games %>% ggplot(aes(x = reorder(name, median_hours), y = median_hours)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  scale_y_continuous(labels = comma) +
  labs(
    title = paste0("Top ", top_n, " Games by Median Playtime"),
    x = "Game",
    y = "Median Hours"
  )

# Boxplots for variation among the same top games
st_top <- st %>% filter(name %in% top_games$name)
ggplot(st_top, aes(x = reorder(name, hours, FUN = median), y = hours)) +
  geom_boxplot(fill = "lightgreen", outlier.alpha = 0.3) +
  coord_flip() +
  scale_y_log10(labels = comma) +
  labs(
    title = paste0("Playtime Distributions for Top ", top_n, " Games"),
    x = "Game",
    y = "Hours Played (log scale)"
  )

# ------------------------------
# Q6. Most Popular Games (by player count)
# ------------------------------
```

```r
popularity <- st %>%
  group_by(name) %>%
  summarize(player_count = n()) %>%
  filter(player_count>20)

print(head(popularity, 15))

# Distribution of player counts across all games
ggplot(popularity, aes(x = player_count)) +
  geom_histogram(bins = 30, color = "white", fill = "plum") +
  scale_x_continuous(labels = comma) +
  labs(
    title = "Distribution of Player Counts Across Games",
    x = "Player Count (rows per game)",
    y = "Number of Games"
  )

# ----------------------------
# Q7. Popularity vs Median Playtime (scatter)
# ----------------------------
pop_vs_play <- st %>%
  group_by(name) %>%
  summarise(
    player_count = n(),
    median_hours = median(hours)
  )

ggplot(pop_vs_play, aes(x = player_count, y = median_hours)) +
  geom_point(alpha = 0.3) +
  labs(
    title = "Popularity vs Median Playtime (per Game)",
    x = "Player Count (log10)",
    y = "Median Hours (log10)"
  )

# End of script
```