UNDERSTANDING ENGAGEMENT ON STEAM:

A STATISTICAL EXPLORATION OF PC GAMING

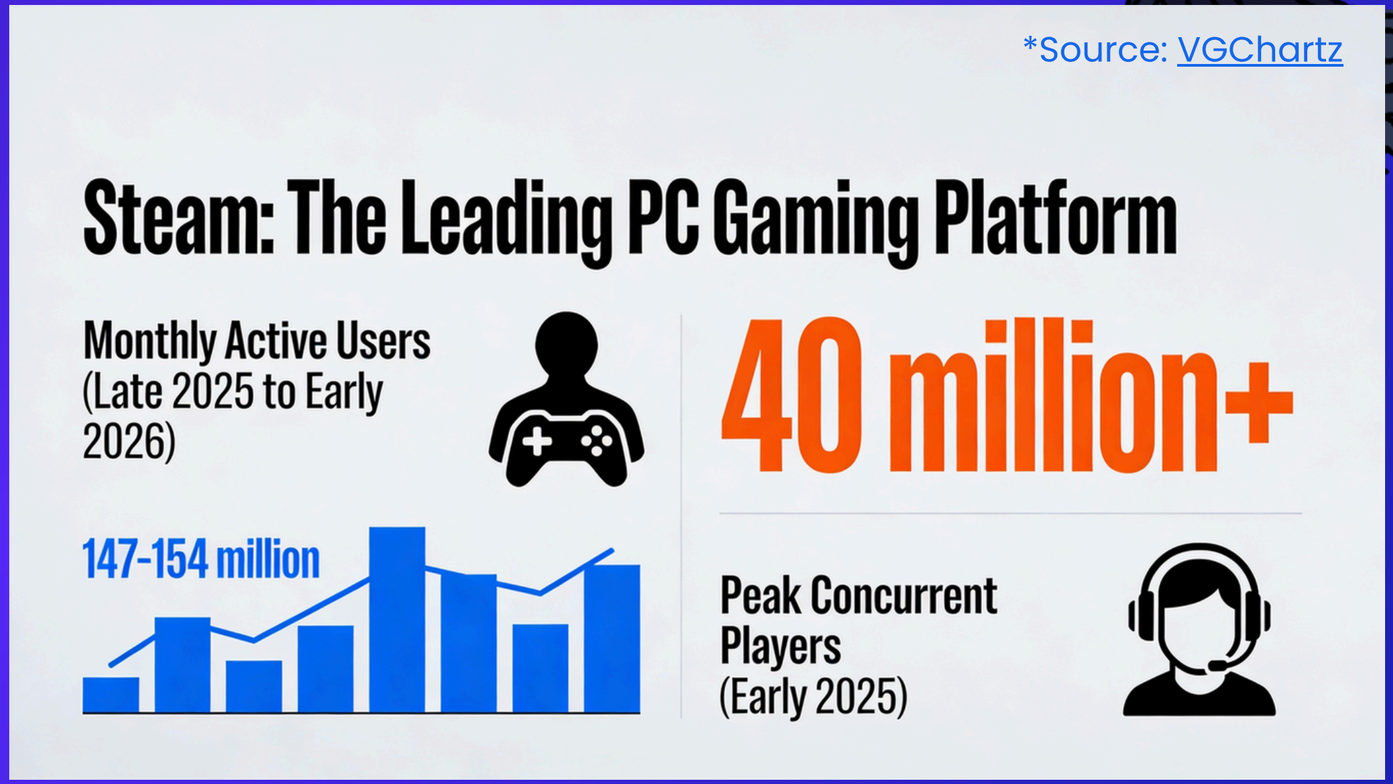
By Rui Parreira
Ironhack - Data Analytics
30 jan 2026

INTRODUCTION

This project explores player behavior and market dynamics on Steam using the **SteamSpy** dataset.

Through a structured exploratory data analysis, it examines how ownership, **engagement**, playtime, and **pricing** interact across hundreds of games.

*Source: VGChartz



DATA SELECTION AND PREPARATION

Data Cleaning

The dataset was cleaned by removing unnecessary columns, handling null values, and preparing the data for analysis.

Exploratory Data Analysis

EDA was conducted to understand gaming patterns, identify outliers, and create key performance indicators (KPIs).

Tests to Validate Hypothesis

The project used Welch's t-test to validate the hypothesis.

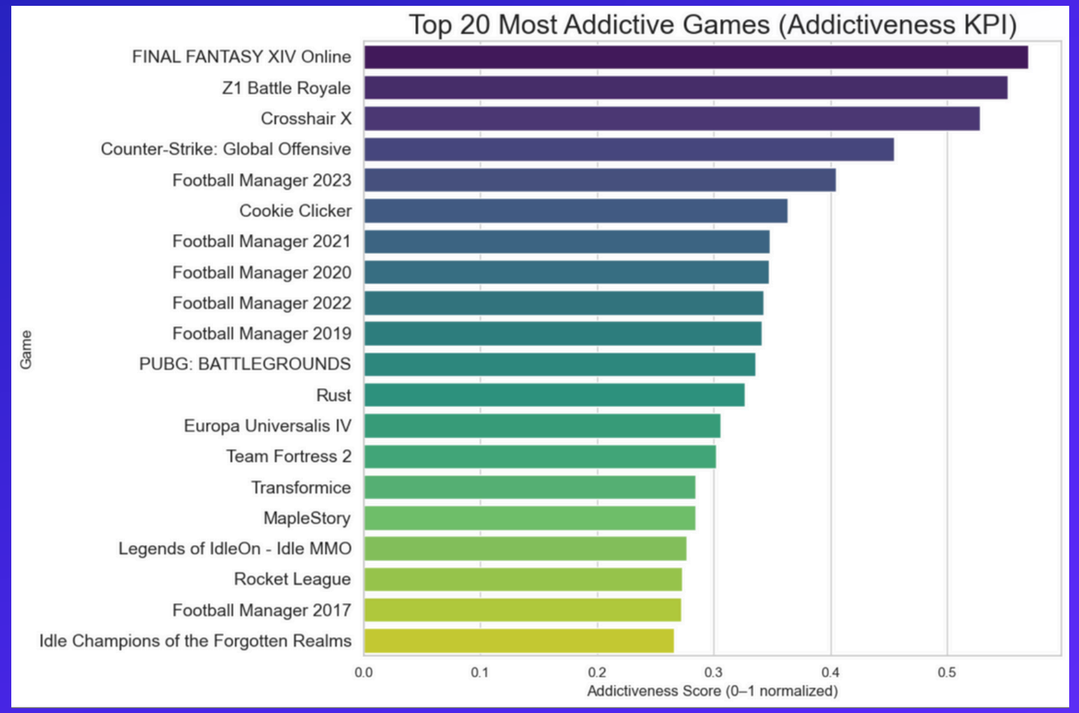
Strategic Takeaways

The data was used to answer business questions and draw conclusions.

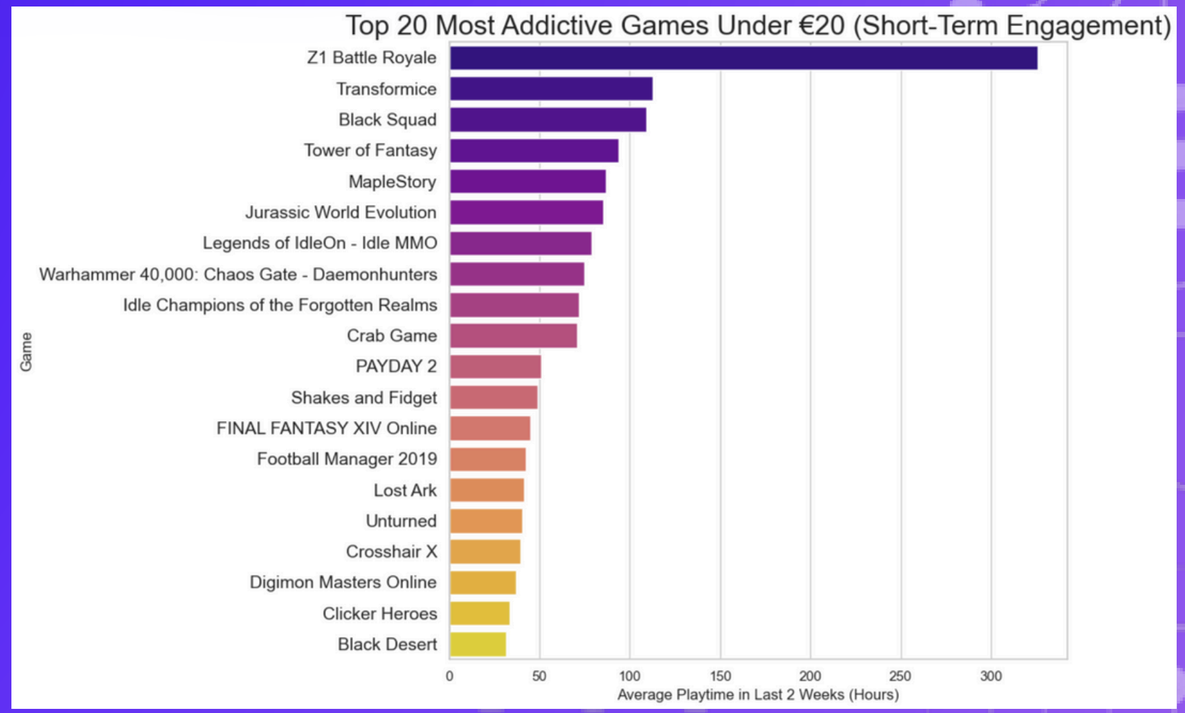
I used a SteamSpy API dataset connected to Steam. The dataset is composed of a sample of the 1,000 (760 paid vs 240 free-to-play) most played games, refreshing every 24 hours.

Link to dataset

MOST ADDICTIVE GAMES IN STEAM



TOP 20 MOST ADDICTIVE GAMES ON STEAM



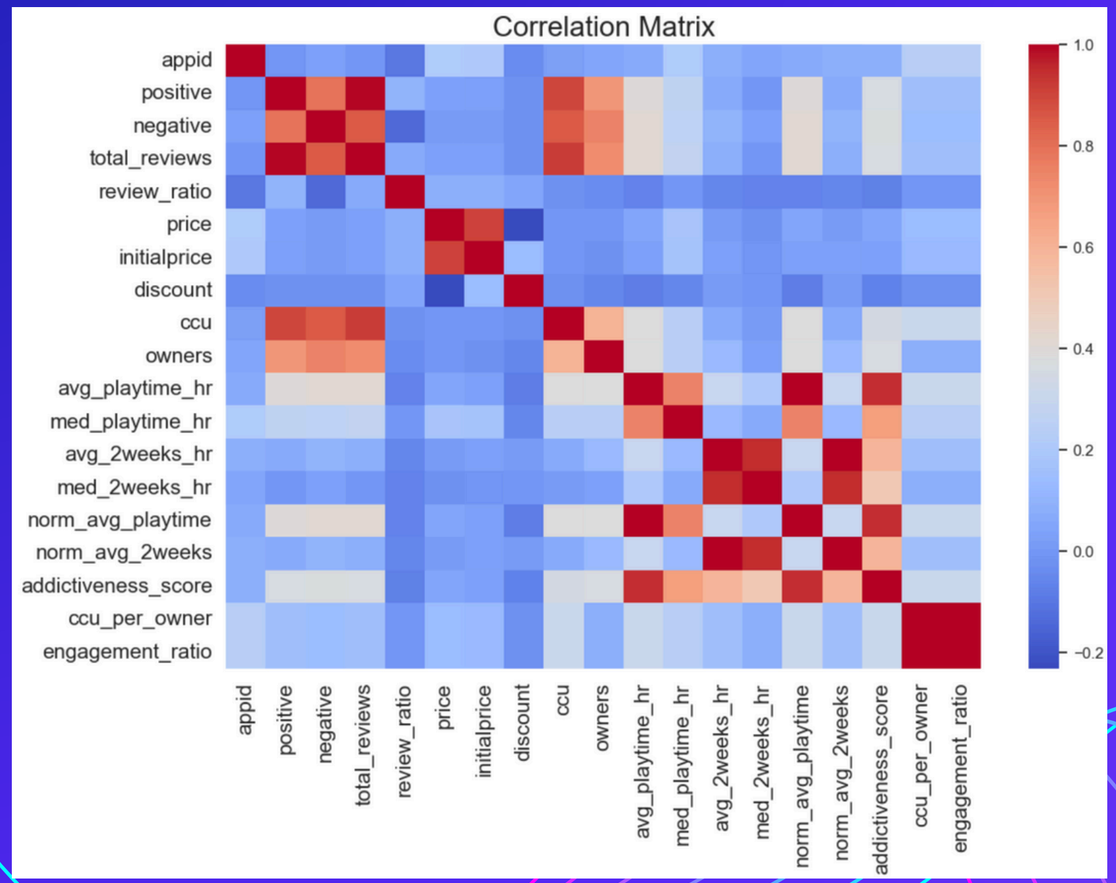
TOP 20 MOST ADDICTIVE GAMES
ON STEAM UNDER $20 LAST 2 WEEKS

HYPOTHESIS

FREE-TO-PLAY GAMES HAVE HIGHER AVERAGE CONCURRENT USERS THAN PAID GAMES

H_0 Free-to-play games have the same/less average concurrent users than paid games

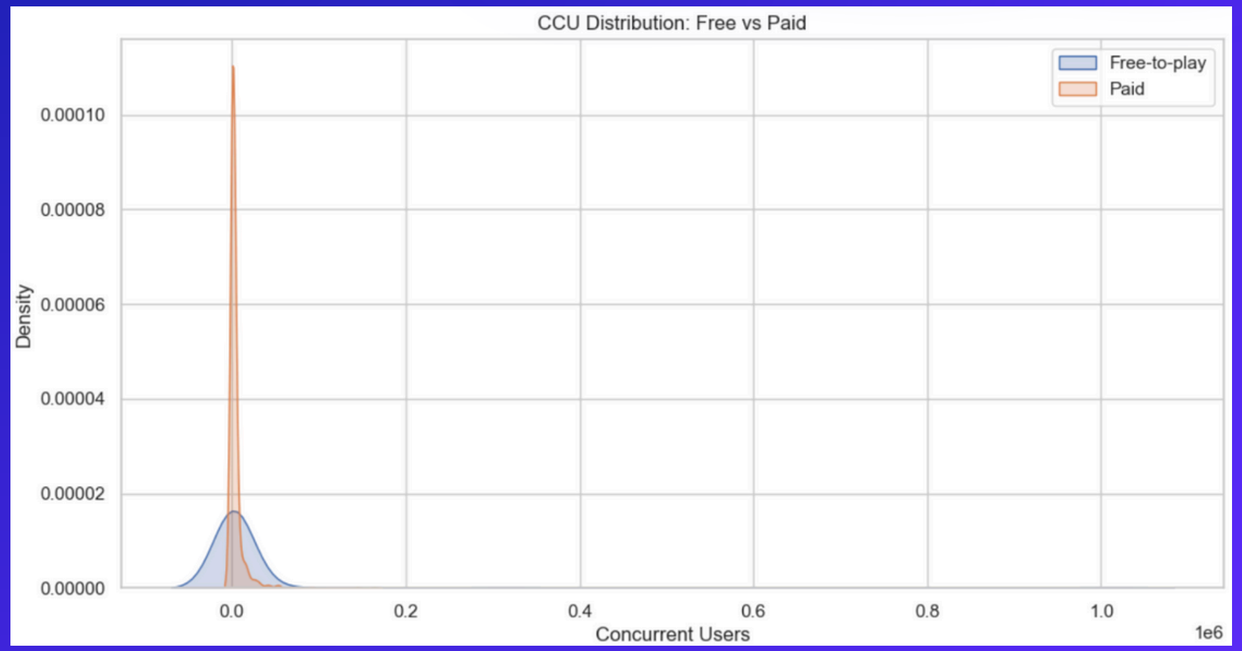
CORRELATION MATRIX

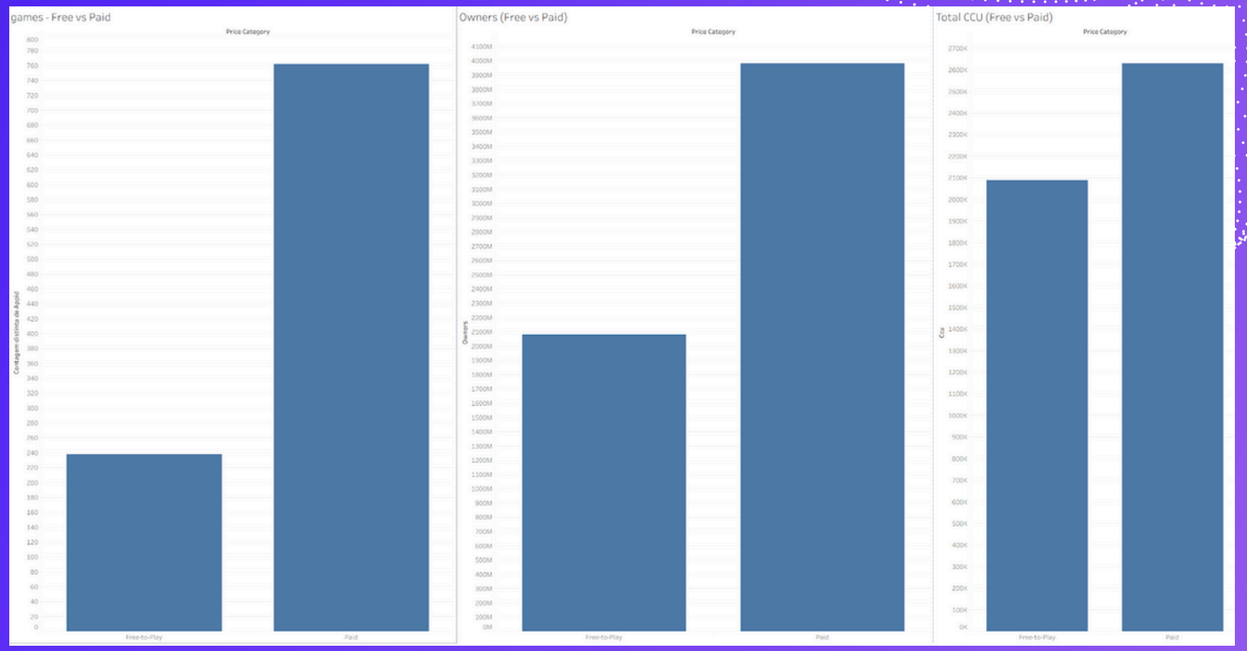


Main indicators

- There is a strong relation between CCU and owners
- Most games don’t change price much post-launch
- Consistency in playtime distribution: median and average

DISTRIBUTION FREE VS PAID



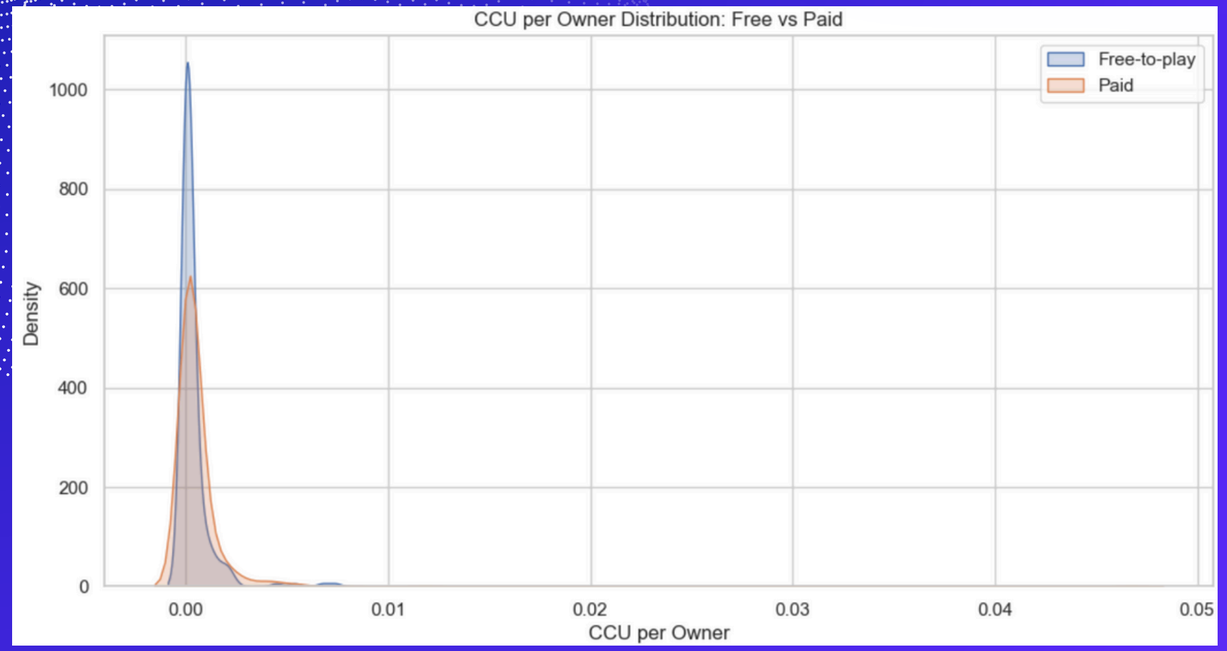


Welch's t-test (one-sided) results:
t-statistic: 1.18
p-value: 0.12
"Not statistically significant at p < 0.05. I fail to reject the null hypothesis."

- The average CCU for free-to-play games is higher than for paid games.

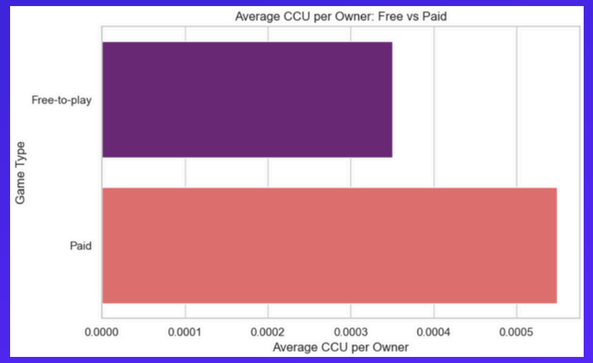
- Paid games vastly outnumber free-to-play games (760 vs 240)
- Most owners in Paid games (Bias)
- More CCU in Paid games

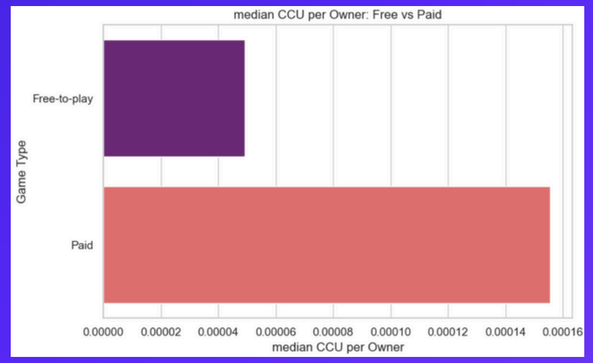
NORMALIZING DATA



Welch's t-test (one-sided) results:
t-statistic: −2.20
p-value: 0.99
Both distributions are right-skewed

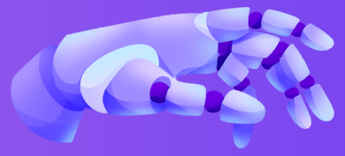
"Of all the people who own the game, how many are playing right now?"





- Paid games have a higher average CCU per Owner (affected by outliers)
- Paid games have a higher median CCU per Owner (resistant to outliers)

CONCLUSIONS



The data strongly contradicts my hypothesis. There is no statistical support for “free > ”paid”. The evidence points in the opposite direction.

TAKEAWAY 01

Free-to-play games attract huge audiences, but a smaller fraction of those players are active at any given moment

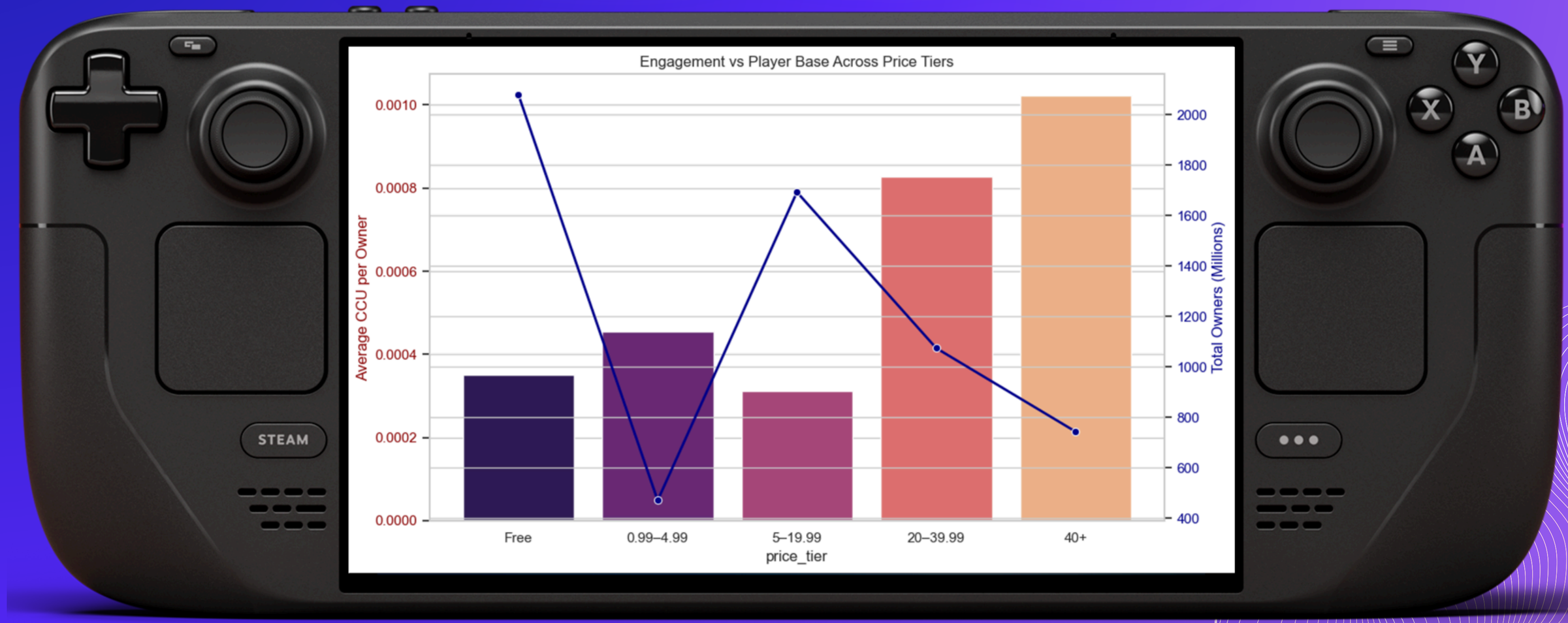
TAKEAWAY 02

Paid players are more likely to be actively playing the games they own

TAKEAWAY 03

Paid games have more loyal or committed players relative to their owner base

CCU PER OWNER ACROSS DIFFERENT PRICE TIERS



STRATEGIC TAKEAWAYS

01

PRICE IS A STRONG PREDICTOR OF ENGAGEMENT INTENSITY

02

FREE GAMES ATTRACT, BUT PAID GAMES RETAIN

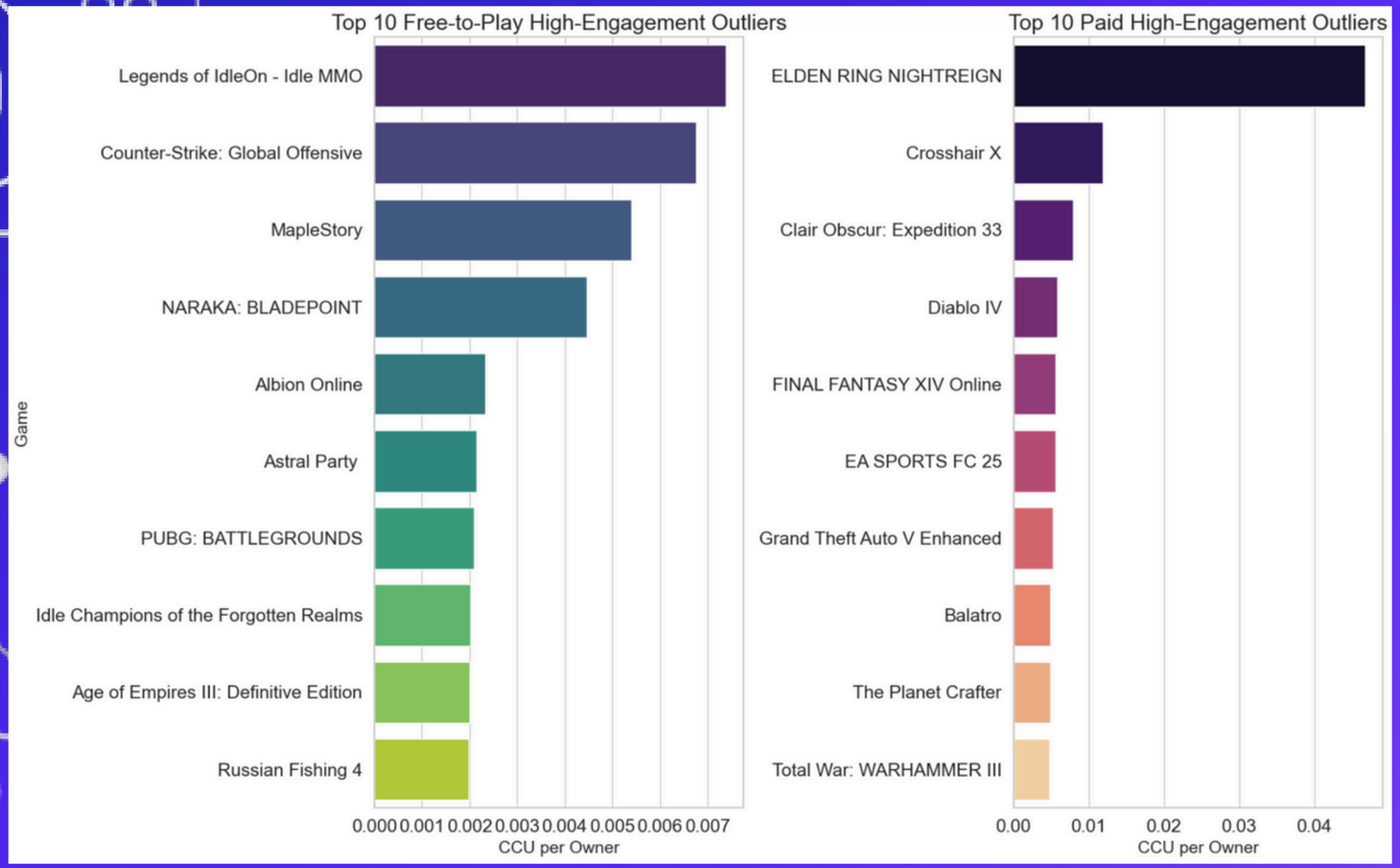
03

MID-PRICE GAMES MAY NEED BETTER RETENTION STRATEGIES

04

HIGH-PRICE GAMES JUSTIFY THEIR COST WITH DEEPER ENGAGEMENT

OUTLIERS: TOP 10 MOST ENGAGEMENT GAMES (FREE AND PAID)



- Z-score and Interquartile Range analysis confirms that the dataset contains extreme outliers across owners
- These outliers are not errors, they represent genuine blockbuster phenomena
- CS:GO, PUBG, GTA V, FC 25 or Elden Ring Nightreign have more than 3 standard deviations above the mean

KPIS

- **Monetization Model KPIs**
(Free vs Paid comparison)

- **Addictiveness score kpi**
(Using average playtime ever and 2 weeks)

- **Concurrent users per owner**
(Of all the people who own the game, how many are playing right now?")

- **Recent engagement**
(Playtime in last 2 weeks)

- **CCU per owner across different price tiers**
(we can see how different price bins differ)

- **Top Outliers - high engagement**
(What are the games with most engagement)

CHALLENGES AND LEARNINGS

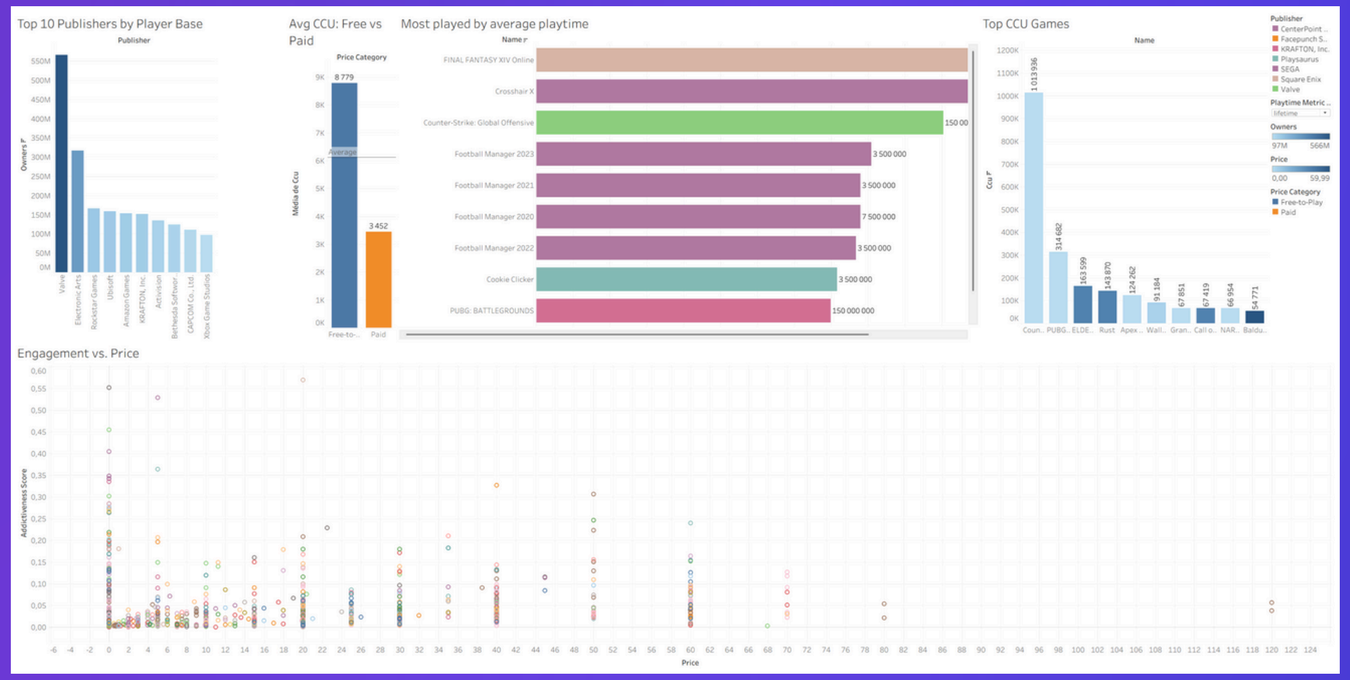
NORMALIZING DATA (FREE VS PAID GAMES)

CHOOSING THE RIGHT METHOD TO ANALYSE HYPOTHESIS

IDENTIFYING OUTLIERS

COMPOSING NARRATIVE

TABLEAU DEMO



[Click to open Tableau Demo]

THANK YOU!

Rui Parreira | Ironhack