



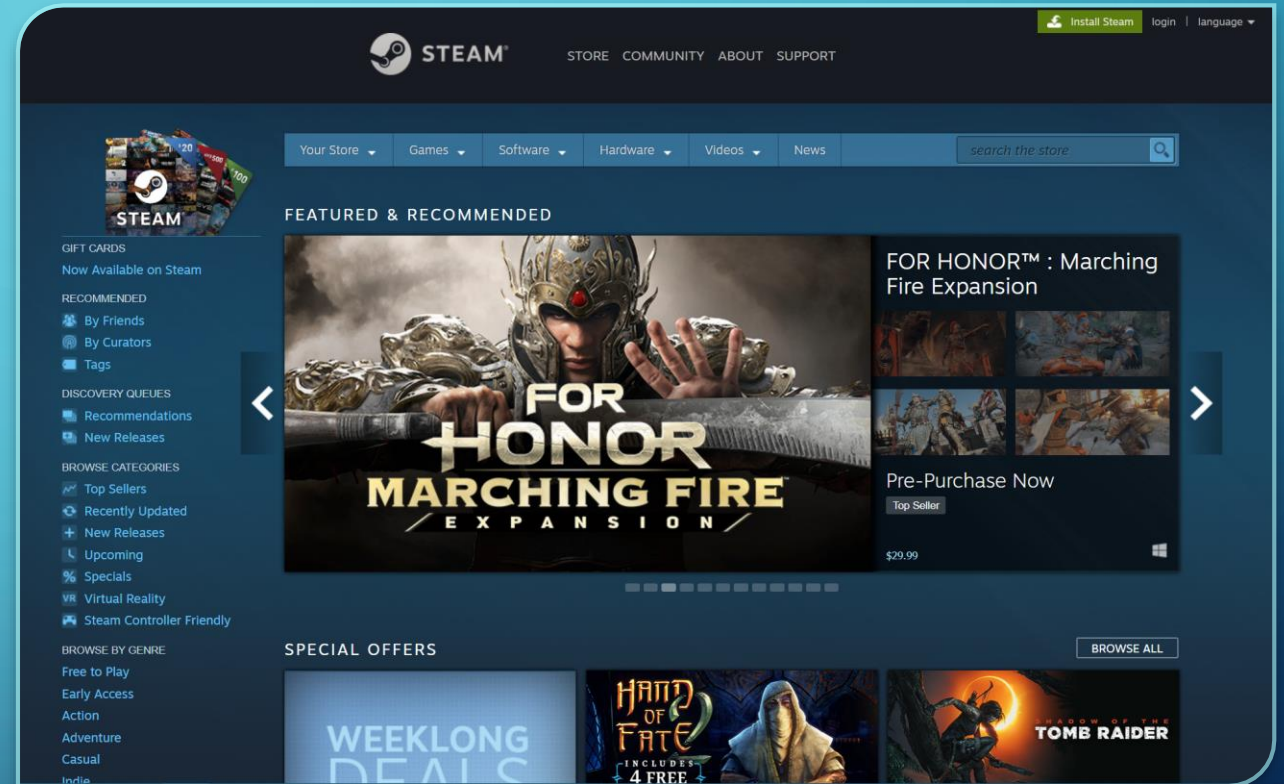
# WEB SCRAPING STEAM

OCTOBER 17, 2018

SEAN JUSTICE

# WHAT IS STEAM

- Steam is a digital distribution platform for PC gaming
  - Created by Valve Software in 2003
  - 150 million registered users
  - 18.5 million concurrent users



# DATA SCRAPED

- Game item contains
  - Title
  - Developer
  - Price
  - Description
  - Total # reviews
    - Percentage of reviews positive
  - Release date
  - Category of game – Defined by users



ASSASSIN'S  
CREED  
ODYSSEY

Choose your fate in Assassin's Creed® Odyssey. From outcast to living legend, embark on an odyssey to uncover the secrets of your past and change the fate of Ancient Greece.

ALL REVIEWS: **Mostly Positive** (4,707)

RELEASE DATE: Oct 5, 2018

DEVELOPER: **Ubisoft Quebec, Ubisoft Montreal, ...** +

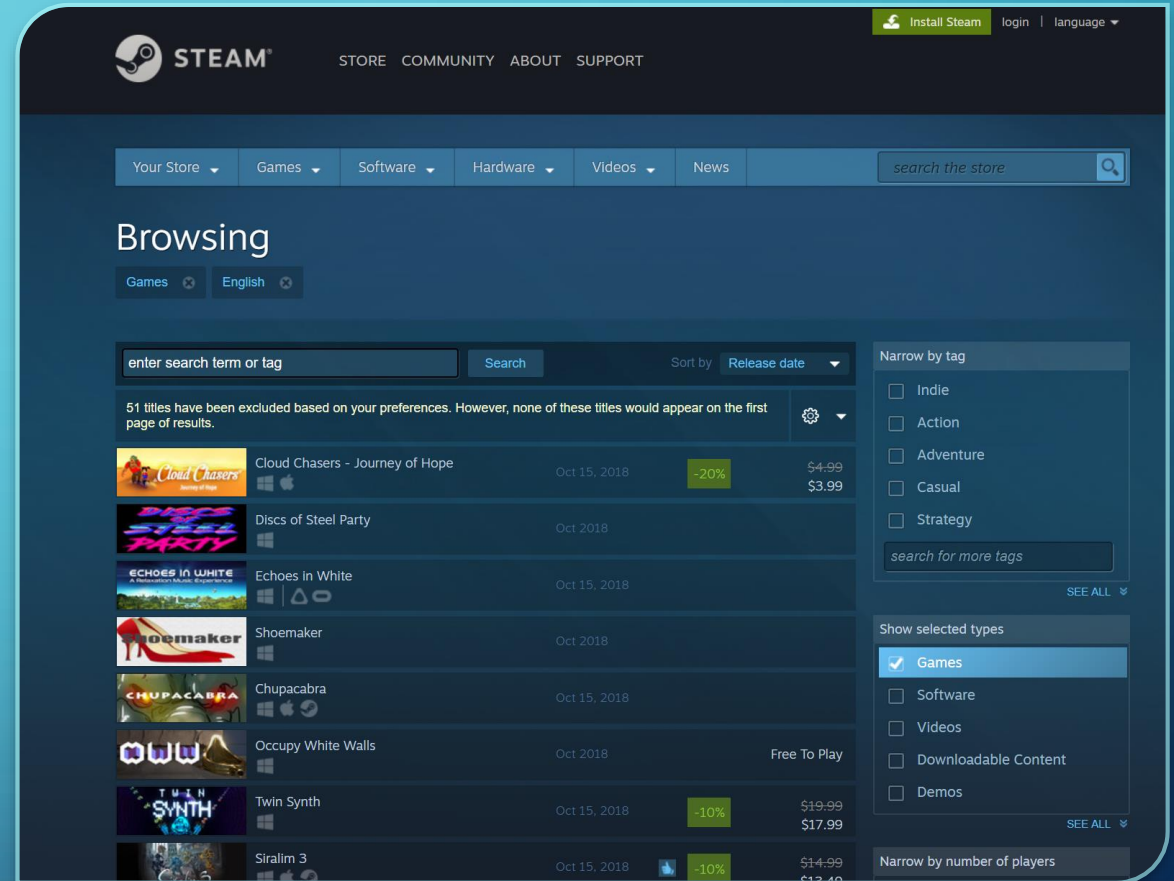
PUBLISHER: **Ubisoft**

Popular user-defined tags for this product:

**Open World** **Action** **RPG** **Adventure** **Assassin** +

# CHALLENGES

- Games not yet available
  - Skipped those games
- Games on sale
  - Save both the sale and original price
- Games without enough reviews
  - Set percent positive to N/A





# BIGGEST CHALLENGE

- Age check for some games
  - Check uses javascript
  - Scrapable data not generated until after clicking view page button



# BIGGEST CHALLENGE

- Used scrapy-splash to render page
  - Splash is a low level browser
  - Renders page after running a lua script
    - Never used lua before
- Documentation isn't great so very steep learning curve
  - Benefit is that it's a very robust tool once learned

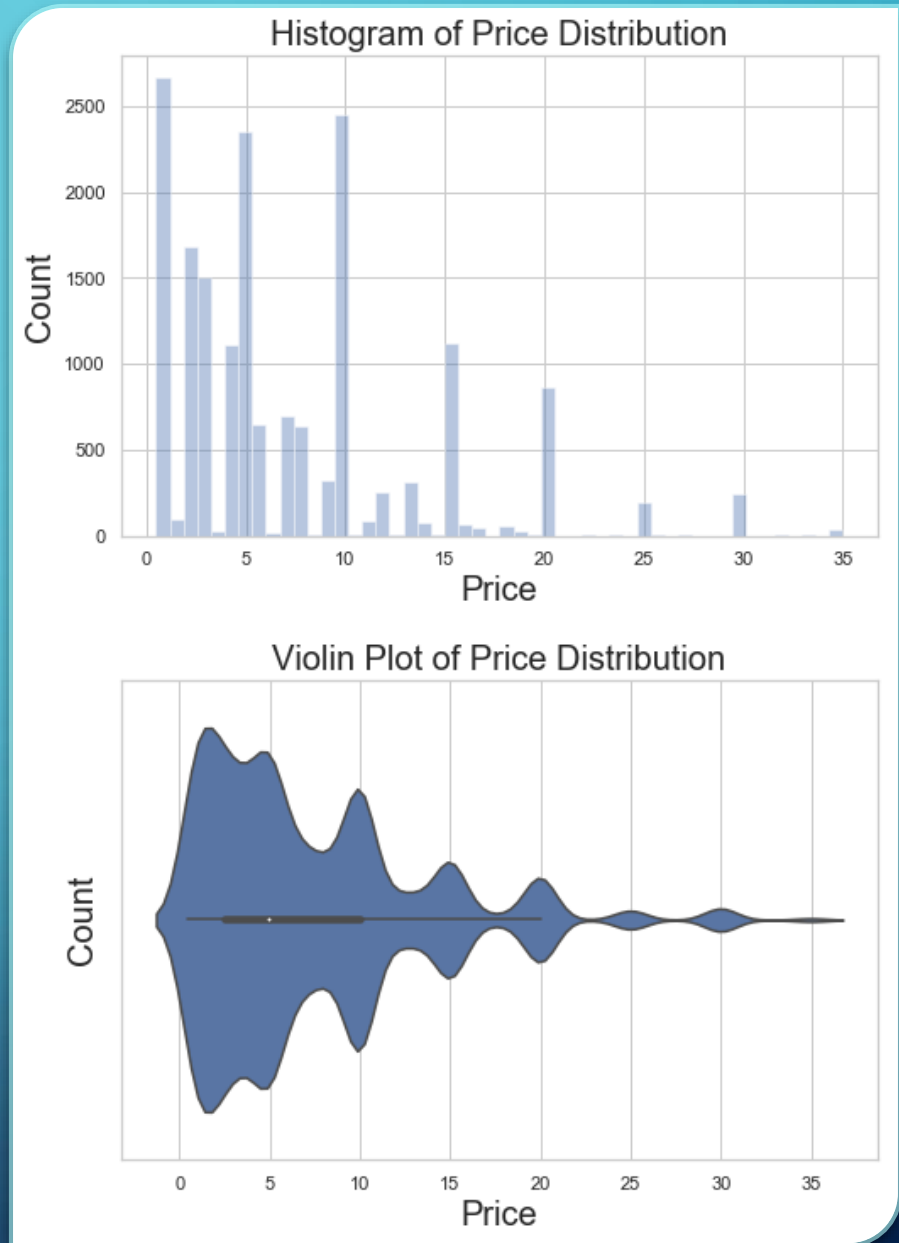
<https://store.steampowered.com/agecheck/app/252490>

Render me!

```
1 function main(splash, args)
2   assert(splash:go(args.url))
3   assert(splash:wait(0.5))
4   splash.images_enabled = false
5   check_for_age = splash:runjs([[
6     if (document.getElementById("ageYear") != null) {
7       btn = document.getElementsByClassName("btnv6_blue_
8       document.getElementById("ageYear").value = 1982
9       btn[0].click()
10    }
11  ]])
12   splash:wait(3)
13   return splash:html()
14 end
15
```

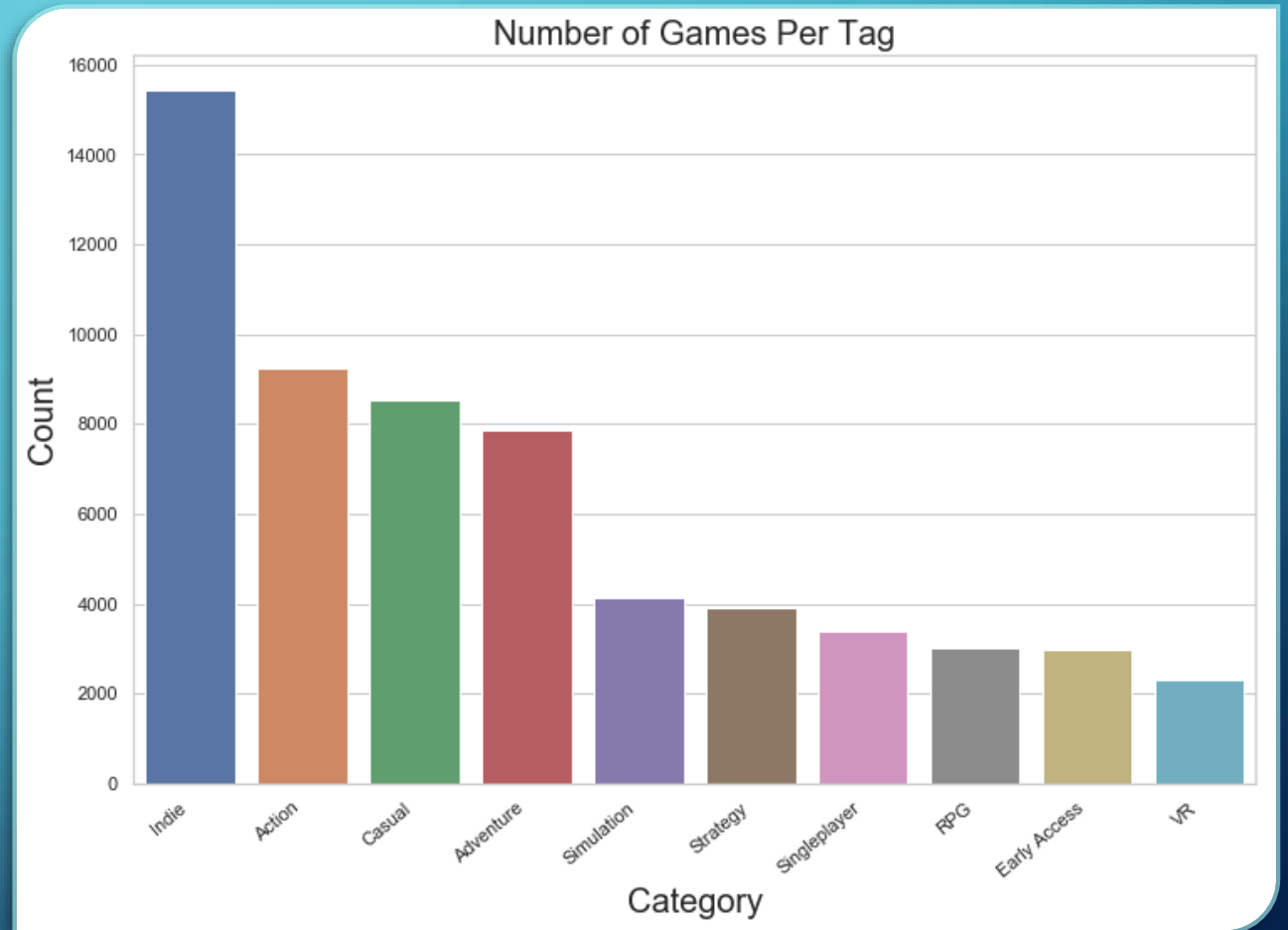
# PRICE ANALYSIS

- Average price of a game is \$8.42
- Median price is \$5.99
- CrisisActionVR is the most expensive game
  - \$199.99
  - Has a 43% rating
  - No one seems to know why it's so expensive



# CATEGORY ANALYSIS

- Each game can have up to 20 tags
- 354 total game categories
- Indie is the most popular tag







## WORD CLOUD OF DESCRIPTION

- Most used words from the description provided by developer



# IDEAS FOR FUTURE DEVELOPMENT

- Scrape the reviews of all of the games
  - Uses an infinite scroll page so will need to use Splash
  - Some games have very large number of reviews ( $>2$  million)
- Map out the relationship between tags
  - Look at trends over the years