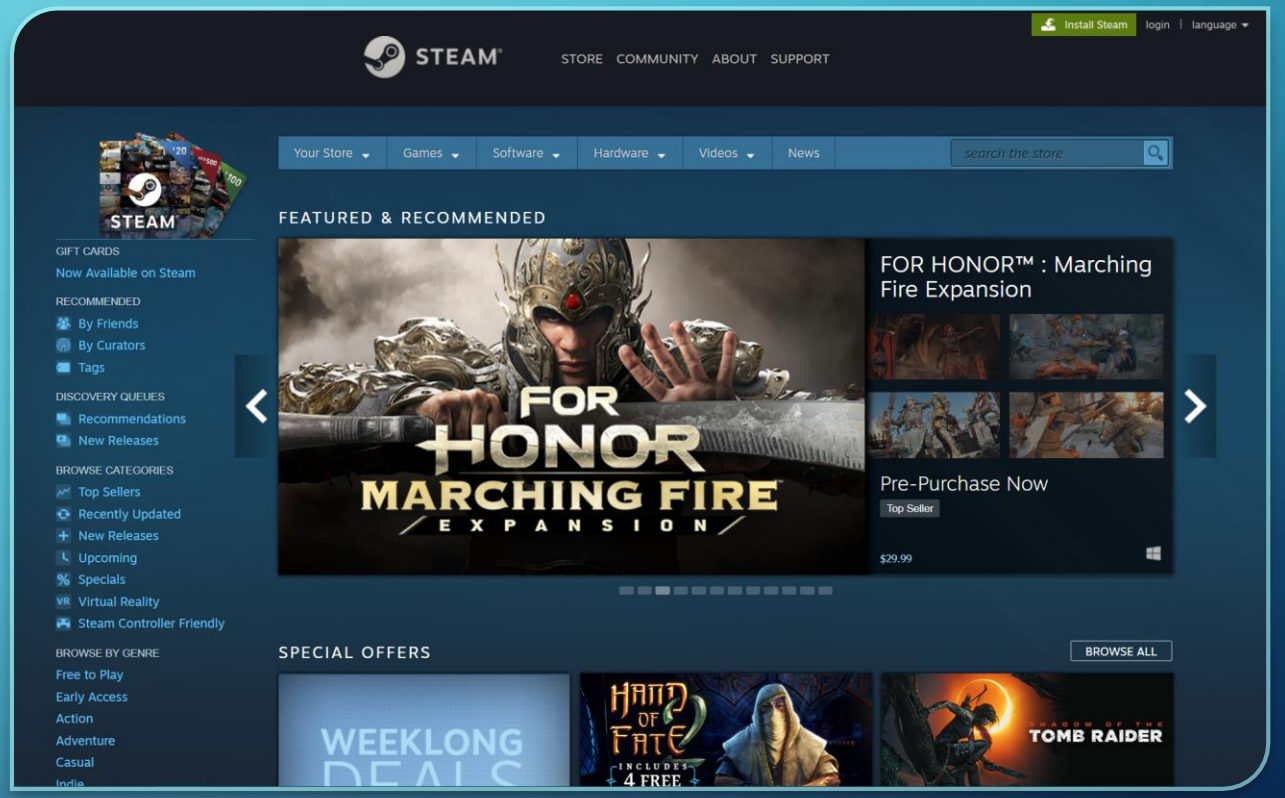
WEB SCRAPING STEAM

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

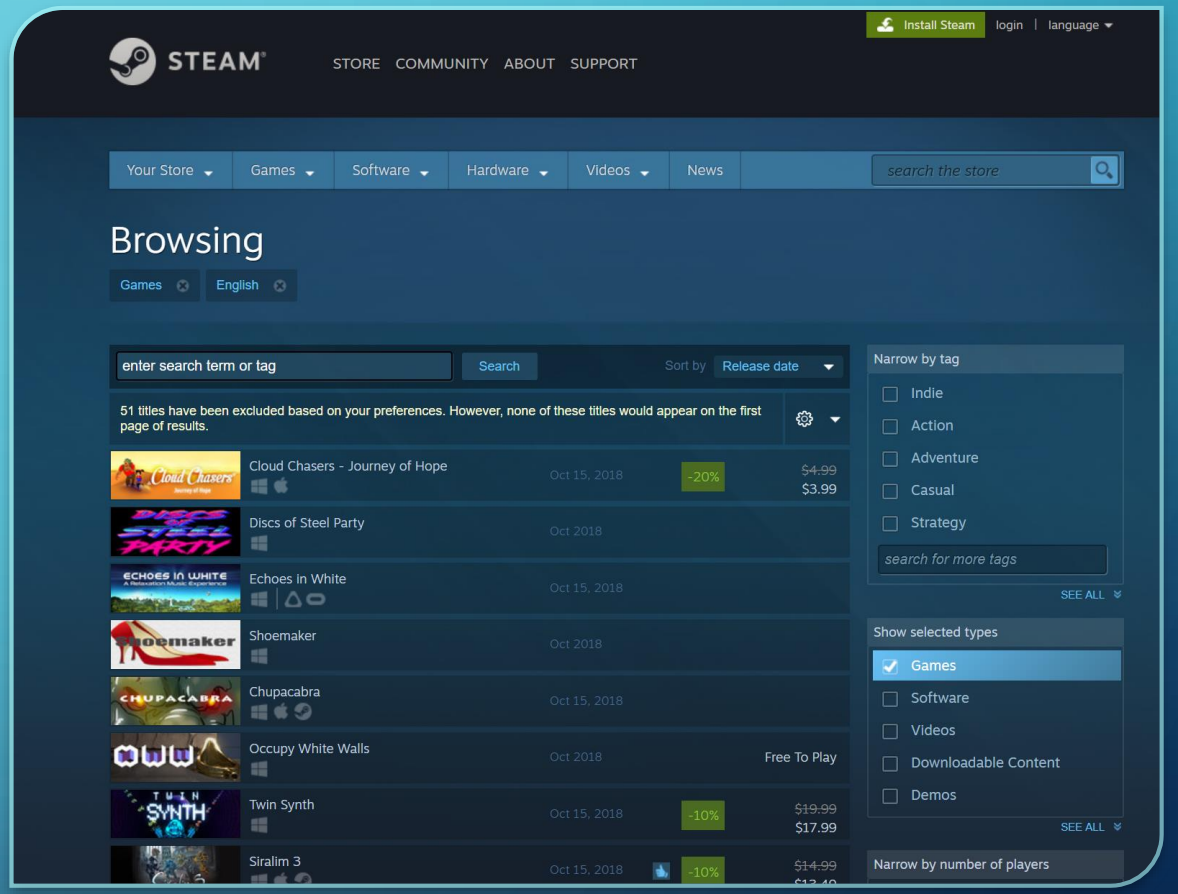


DATA SCRAPED

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

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

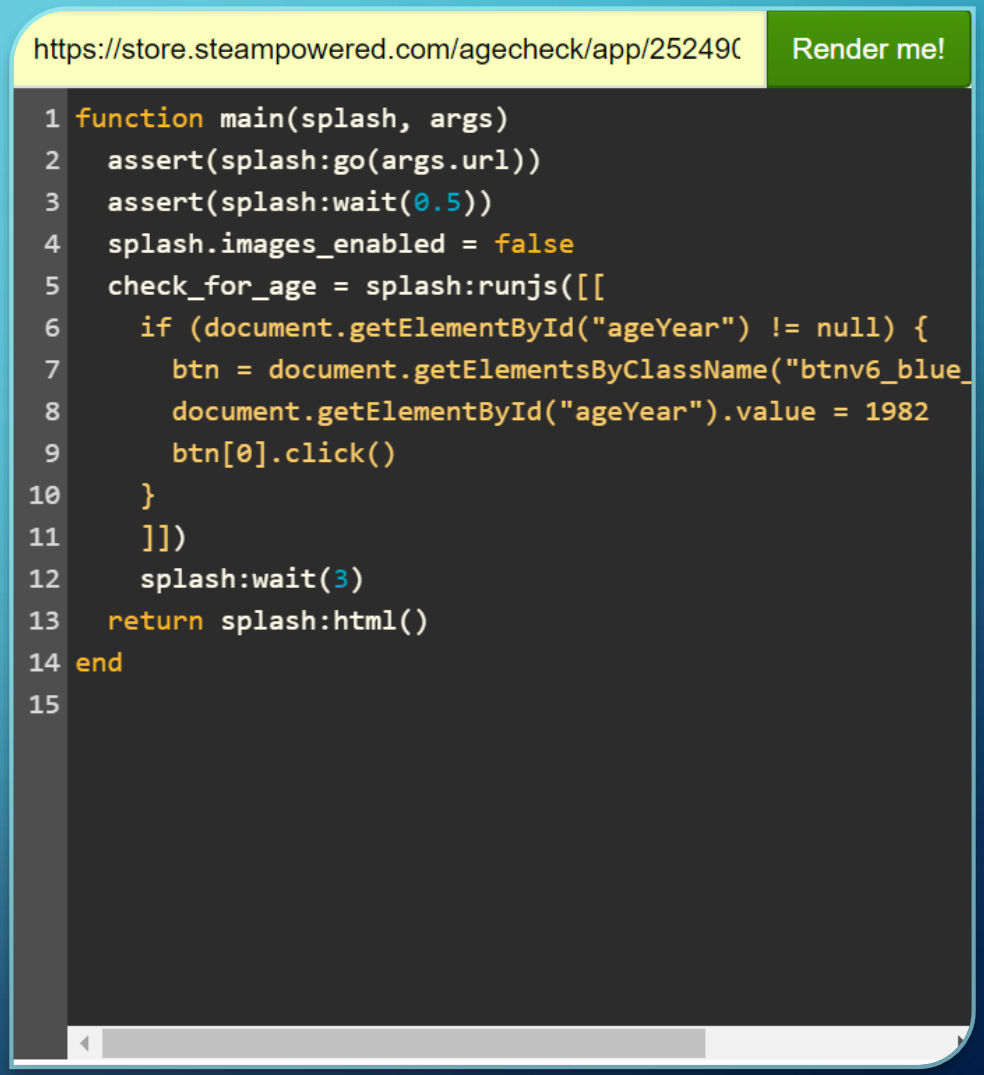


BIGGEST CHALLENGE

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

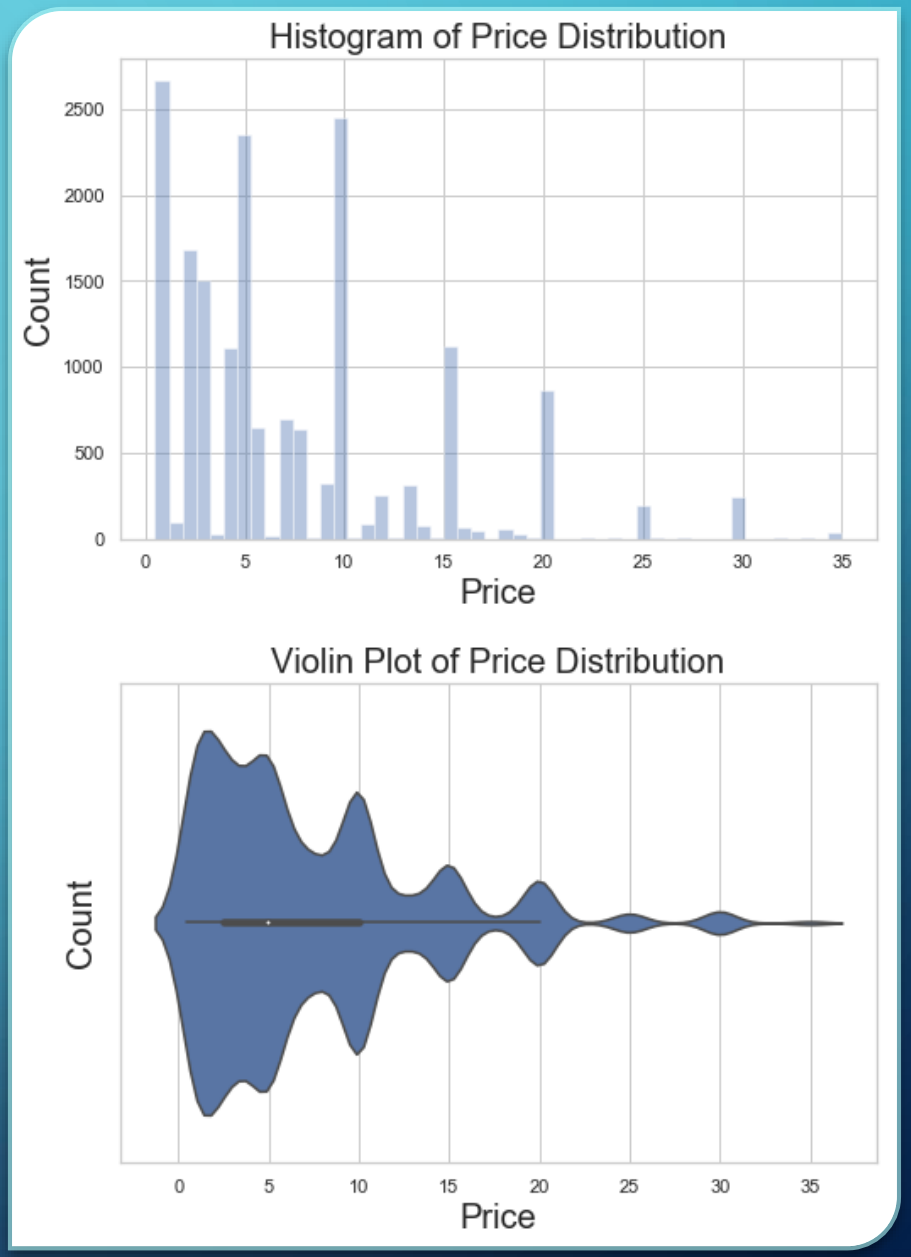
BIGGEST CHALLENGE

- 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



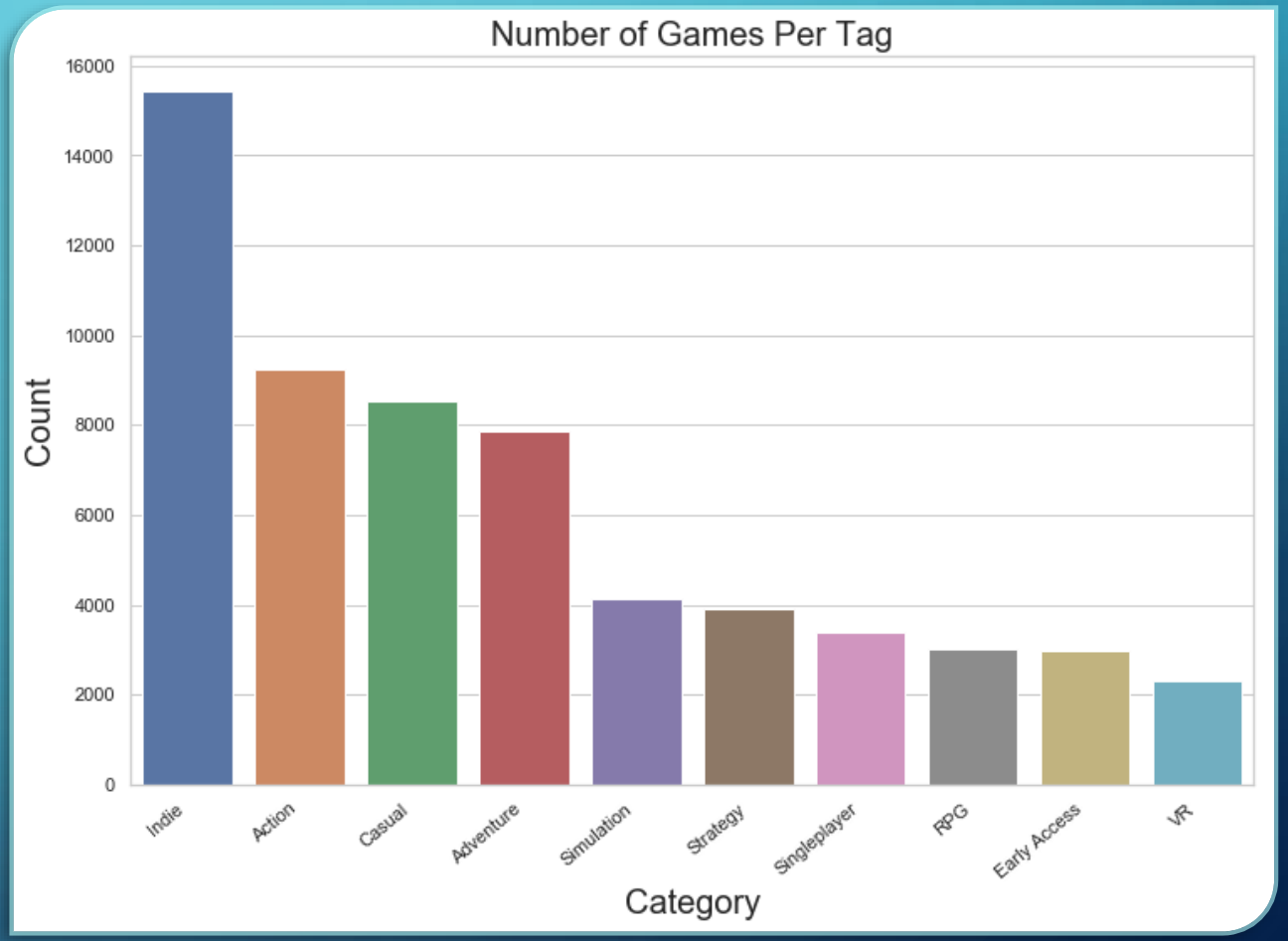
PRICE ANALYSIS

- 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



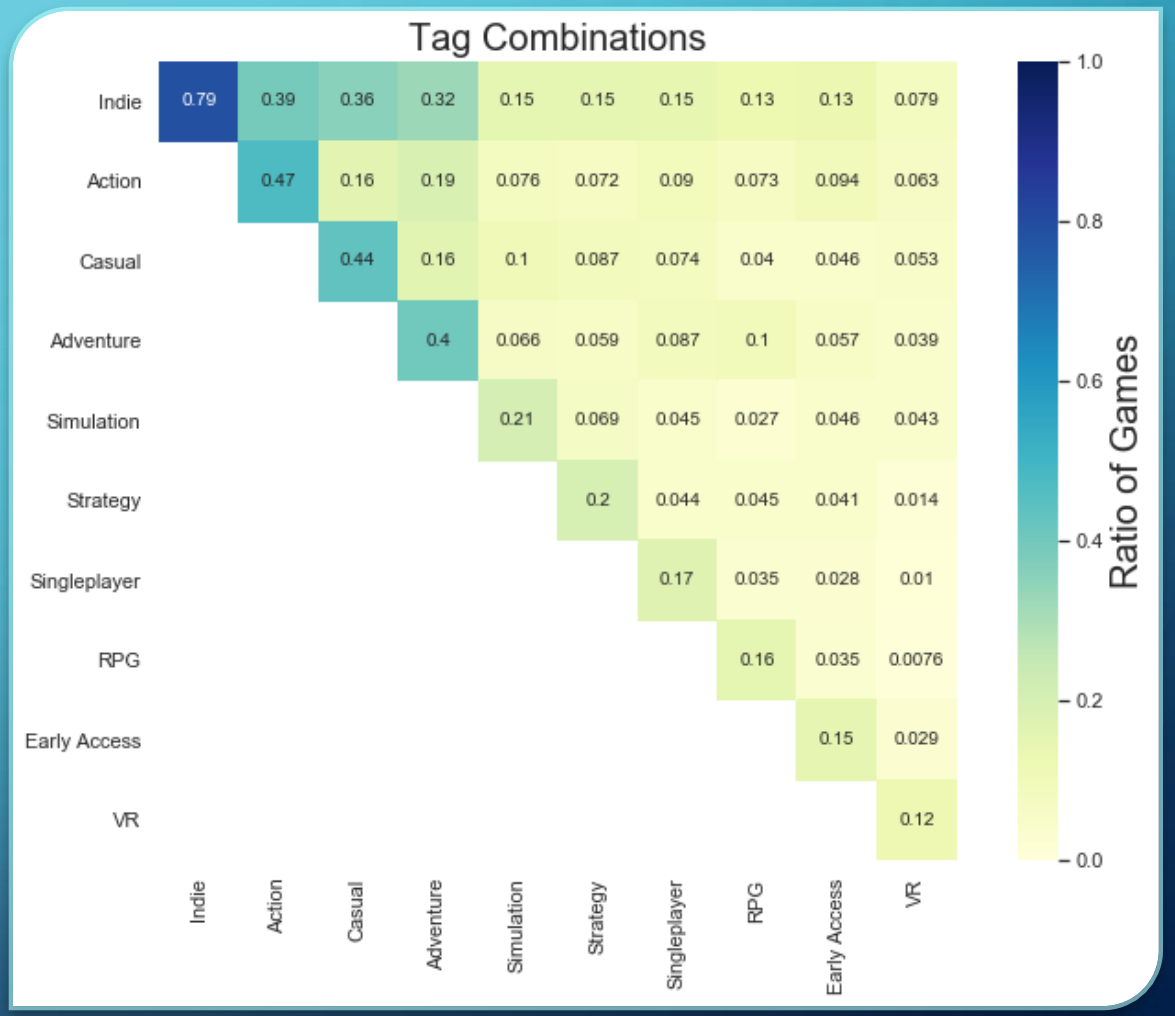
CATEGORY ANALYSIS

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



CATEGORY COMBINATIONS

- Indie and Action are the most used tag combination



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