Steam Games Analysis

SC1015 Mini-Project FCSD Group 6

Glyn Jong, Brian Goh, Muhd Alfiq

START MENU

What is Steam?

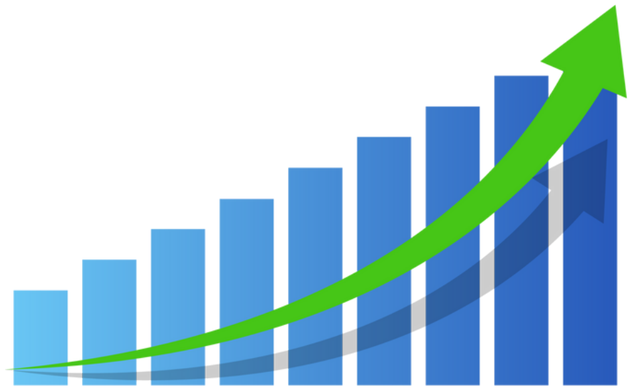
Context

Steam Logo

- A video game digital distribution service and storefront developed by Valve Corporation.
- Launched in 2003, largest platforms for PC gaming

Features

- Purchase, download, install games and software titles from various developers and publishers.
- Users can add and play with other steam users in their friend list.



Rapid growth of steam games

- In 2023, around 33 million peak concurrent Steam users worldwide, an increase from 27.4 million in just two years.
- Over 73 thousand games available on Steam.

What can businesses and game companies do to leverage on this platform by learning about the trends in gaming?

Multiplayer Games?

RPG Games?

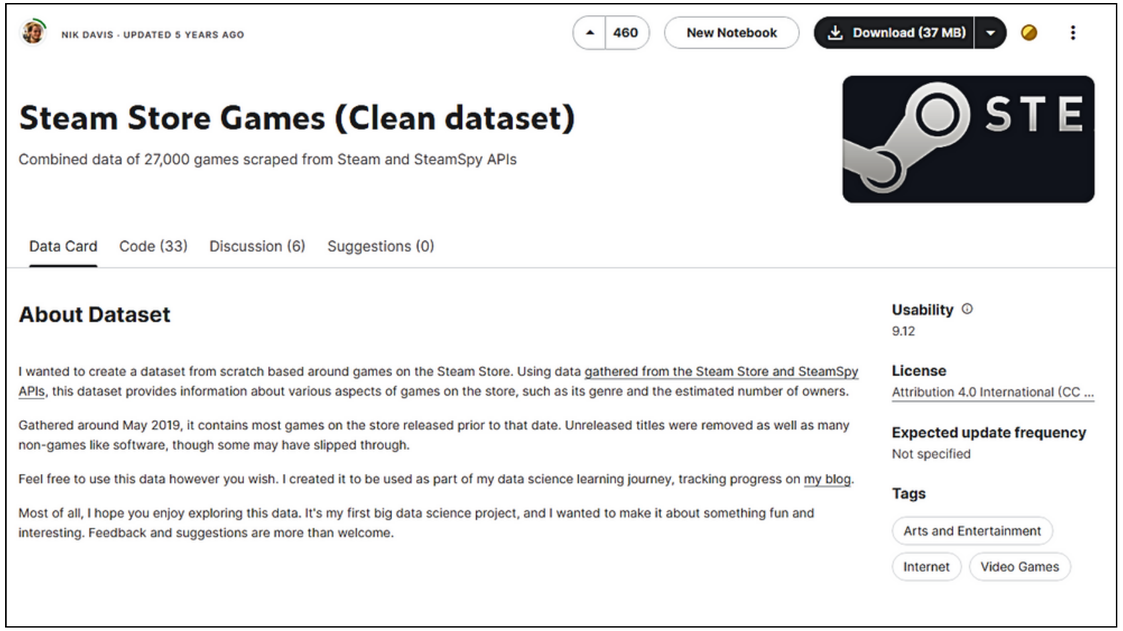
Sandbox Games?

Practical Motivation

Problem Definition

- As an avid fan of video games, we would to like to find out **what makes a video game (on steam) popular**

kaggle



Features

- Dataset of over 27,000 steam games
- Genre of each game
- Estimated number of owners
- Name of developer(s).
- Each games release date
- Average user playtime based on SteamSpy

Exploratory Data Analysis

Data Cleaning

```
# Delete rows where irrelavent
cleaned_df = cleaned_df.drop(columns=["developer", "publisher", "required_age", "platforms"])

cleaned_df.head()
```

	appid	name	release_date	english	categories	genres	steamspy_tags	achievements	positive_ratings	negative_ratings
0	10	Counter-Strike	2000-11-01	1	Multi-player;Online Multi-Player;Local Multi-P...	Action	Action;FPS;Multiplayer	0	124534	3339
1	20	Team Fortress Classic	1999-04-01	1	Multi-player;Online Multi-Player;Local Multi-P...	Action	Action;FPS;Multiplayer	0	3318	633
2	30	Day of Defeat	2003-05-01	1	Multi-player;Valve Anti-Cheat enabled	Action	FPS;World War II;Multiplayer	0	3416	398

Data Cleaning

```
new_columns = []

for i in ["categories", "steamspy_tags", "genres"]:
    possible = cleaned_df[i].str.split(";").explode().unique()
    for p in possible:
        new_column = cleaned_df[i].str.contains(p, regex=False).astype(int)
        new_columns.append(pd.Series(new_column, name=i + "_" + p))

cleaned_df = pd.concat([cleaned_df] + new_columns, axis=1)
cleaned_df = cleaned_df.drop(columns=["categories", "steamspy_tags", "genres"])

cleaned_df.head()
```

_playtime	median_playtime	owners	...	genres_Web Publishing	genres_Education	genres_Software Training	genres_Sexual Content	genres_Audio Production	genres_Game Development
17612	317	10000000-20000000	...	0	0	0	0	0	0
277	62	5000000-10000000	...	0	0	0	0	0	0
187	34	5000000-10000000	...	0	0	0	0	0	0
258	184	5000000-10000000	...	0	0	0	0	0	0
624	415	5000000-10000000	...	0	0	0	0	0	0

Data Cleaning

Use regex

Learned how to use regex to extract the values we want in the data

Do one-hot encoding

Split the columns to be used for analysis into individual variables

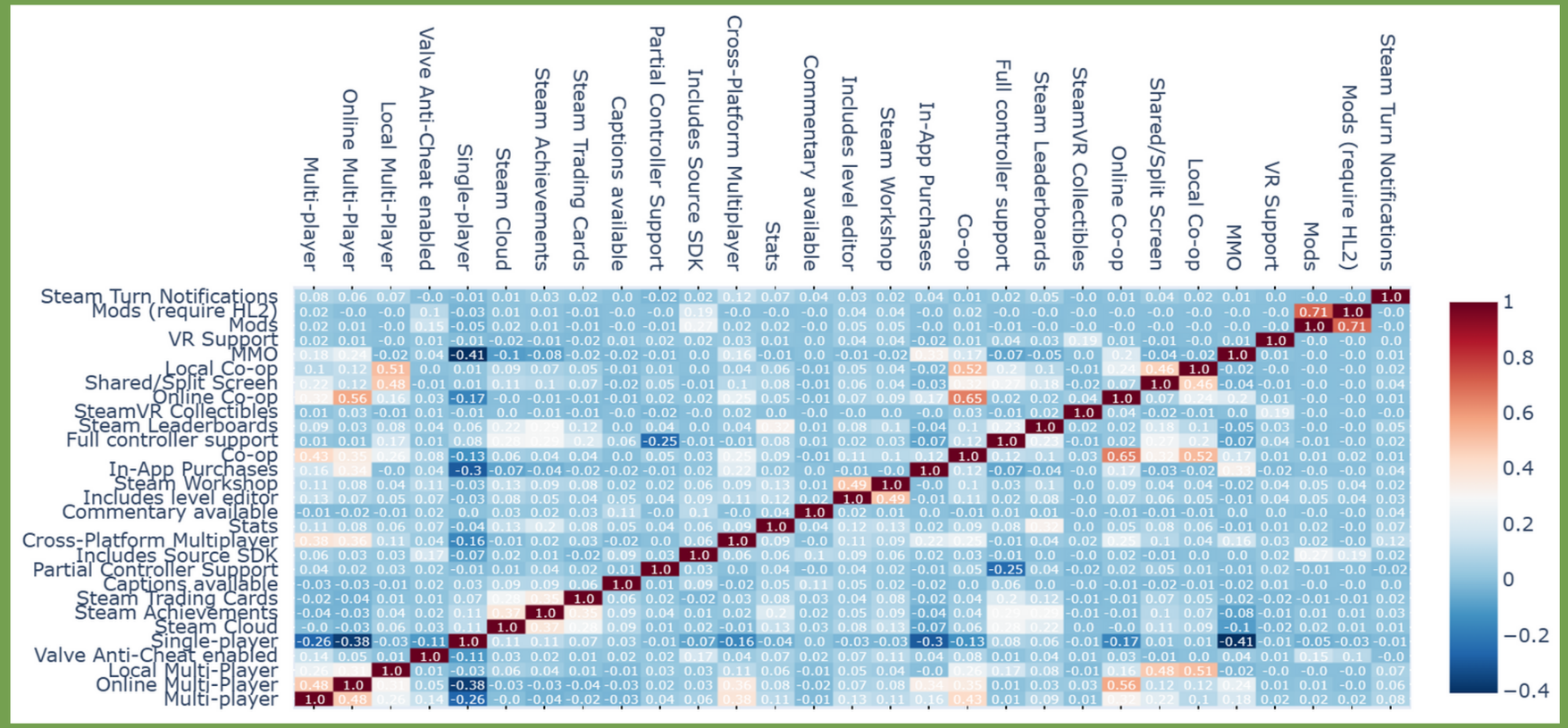
Create new “age” column

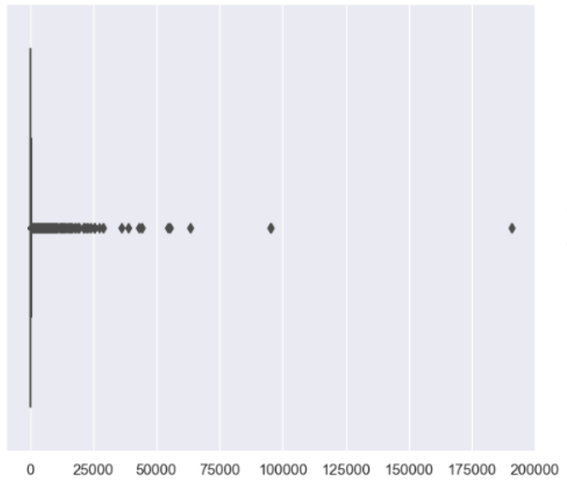
Easier to compare objectively compared to raw date, drop the raw date column

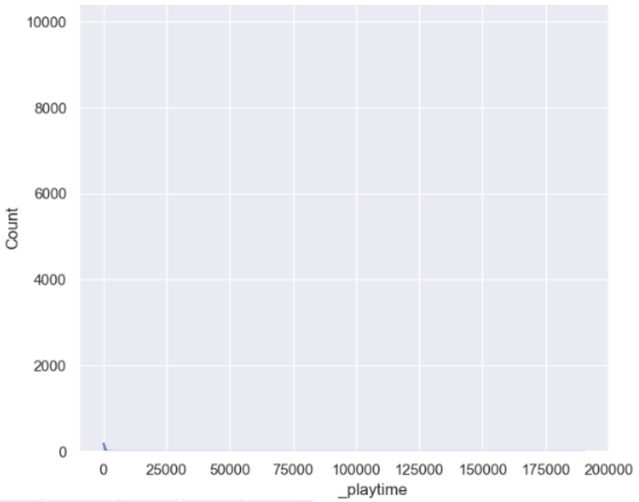
Use only games with >25 ratings

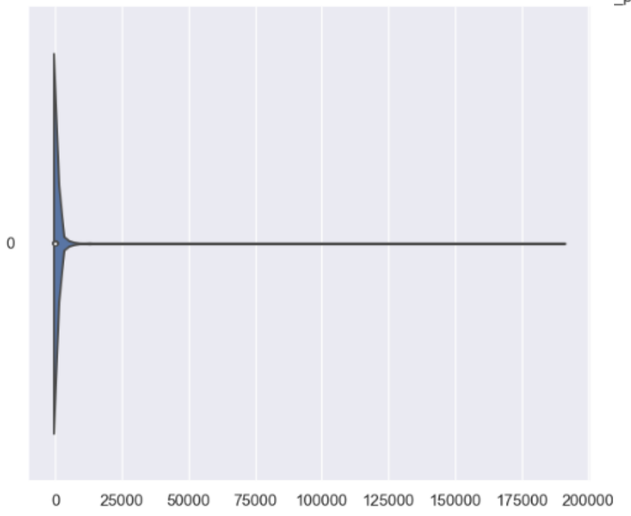
Create new metric “positive_ratio”, then drop the ratings columns

Categories


Steam Turn Notifications
Mods (require HL2)
Mods
VR Support
MMO
Local Co-op
Shared/Split Screen
Online Co-op
SteamVR Collectibles
Full controller support
Steam Leaderboards
In-App Purchases
Steam Workshop
Includes level editor
Commentary available
Stats
Cross-Platform Multiplayer
Includes Source SDK
Partial Controller Support
Captions available
Steam Trading Cards
Steam Achievements
Steam Cloud
Single-player
Valve Anti-Cheat enabled
Local Multi-Player
Online Multi-Player
Multi-player
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4







average_playtime data

- Very heavily skewed
- 1506 outliers
  - remove all outliers
- Many games with 0 hours of playtime
  - remove all 9000 games
  - left 4412 games in dataset

Compare categories to average_playtime

- Create dataframe containing ONLY the categories variables
- Append cleaned average_playtime column to the dataframe
- Drop categories where >95% of games are in either 0 or 1 (heavily unbalanced/skewed distribution)
- Plot the box plots of individual categories against average_playtime

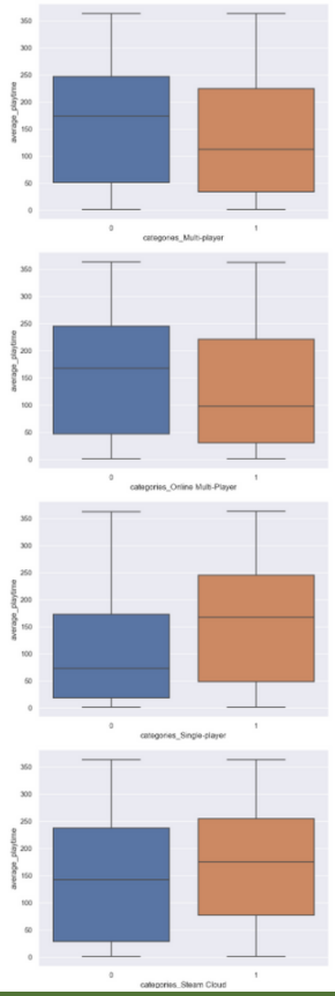
```
categories_Multi-player
0    0.792384
1    0.207616
Name: proportion, dtype: float64

categories_Online Multi-Player
0    0.897552
1    0.102448
Name: proportion, dtype: float64

categories_Local Multi-Player
0    0.955802
1    0.044198
Name: proportion, dtype: float64

categories_Valve Anti-Cheat enabled
0    0.992294
1    0.007706
Name: proportion, dtype: float64

categories_Single-player
1    0.93563
0    0.06437
Name: proportion, dtype: float64

categories_Steam Cloud
0    0.63282
1    0.36718
Name: proportion, dtype: float64
```

Compare categories to average_playtime

- Plot the box plots of individual categories against average_playtime

CONCLUSION:

Not viable



Compare genres to average_playtime

- Create dataframe containing ONLY the genres variables
- Append average_playtime column to the dataframe
- Drop genres where >95% of games are in either 0 or 1 (heavily unbalanced/skewed distribution)
- Plot the box plots of individual genres against average_playtime

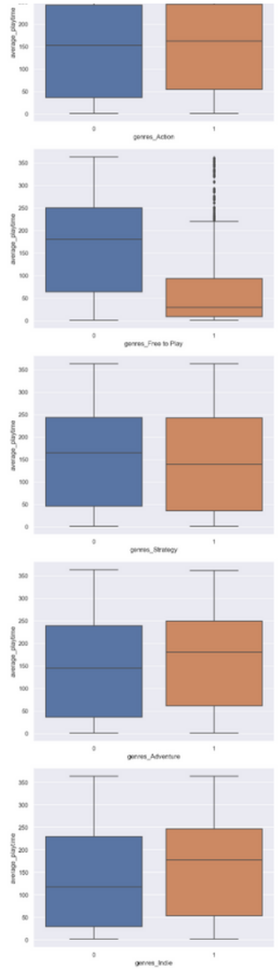
```
genres_Strategy
0    0.801677
1    0.198323
Name: proportion, dtype: float64

genres_Adventure
0    0.625567
1    0.374433
Name: proportion, dtype: float64

genres_Indie
1    0.703762
0    0.296238
Name: proportion, dtype: float64

genres_RPG
0    0.838395
1    0.161605
Name: proportion, dtype: float64

genres_Animation & Modeling
0    0.998413
1    0.001587
Name: proportion, dtype: float64

genres_Video Production
0    0.999093
1    0.000907
Name: proportion, dtype: float64

genres_Casual
0    0.68835
1    0.31165
```

Compare genres to average_playtime

- Plot the box plots of individual categories against average_playtime

CONCLUSION:

Not viable



Using steamspytag

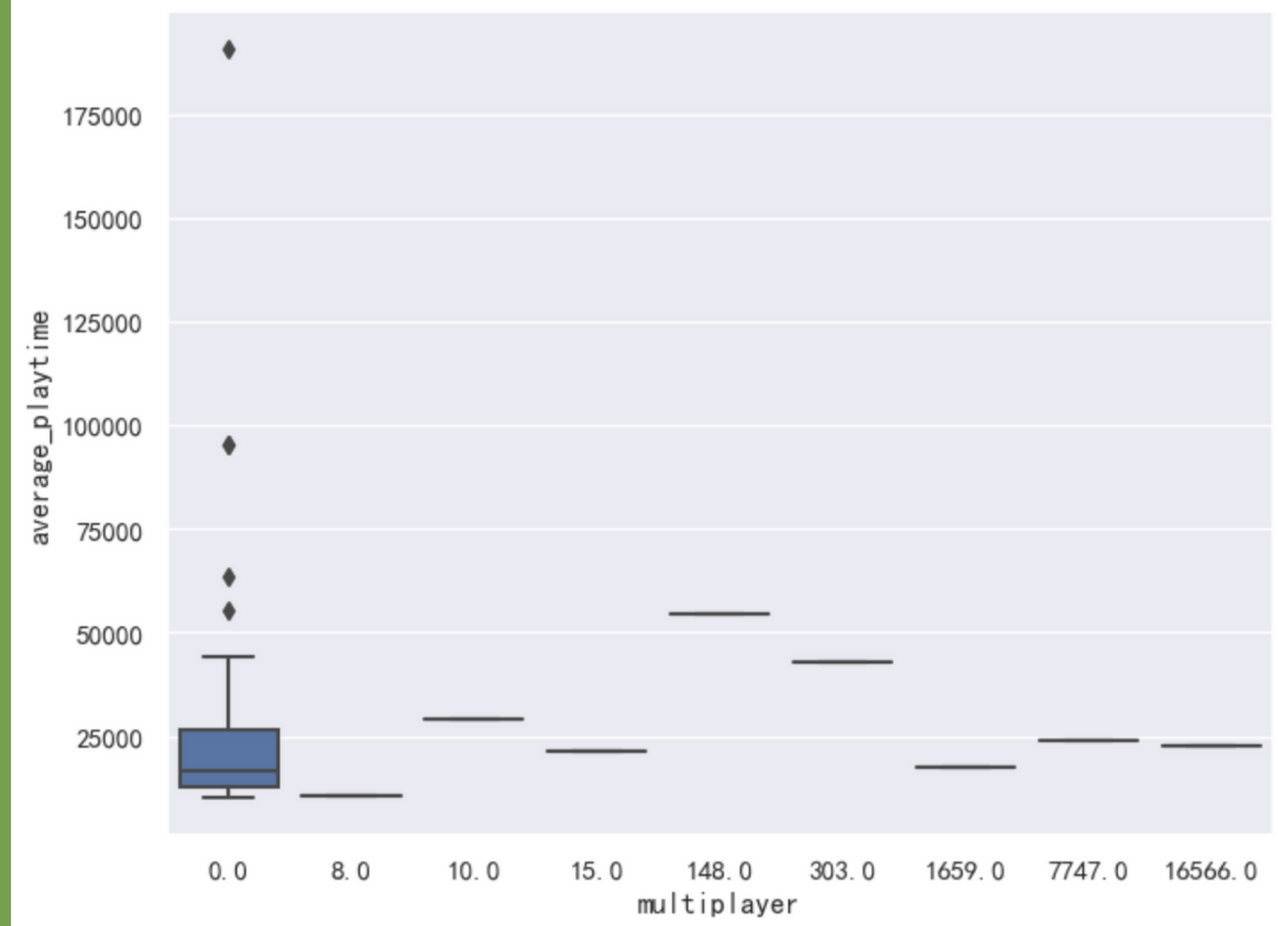
Drop games not in main steam file

Use appid to compare as the primary key, ensured data tallies, removed 14000 rows
Check:

- same number of games
- same games included
- games ordered in the same manner

Streamline data

After add average_playtime column, remove games <10000 hours playtime



Compare steamspytags to average_playtime

- Plot the box plots of individual steamspytag against average_playtime

CONCLUSION:

Better,

but still not viable

Machine Learning

Results

- Here, we have the results from running our dataset through the different models
- Both Random forest and Boosting regressor has the highest R^2 value of 0.57
- Although decision tree R^2 value is low, it has a good p-value (below 0.05)

```
=== Dummy Regressor (Baseline) ===
Score: -0.00 (p-value 0.952)
                Best: 0.048

=== Linear Regression ===
Score: -671100763278016944516628 48.00 (p-value 0.238
                Best: 0.048

=== Decision Tree ===
Score: 0.17 (p-value 0.048)
                Best: 0.048

=== Boosting Regressor ===
Score: 0.57 (p-value 0.167)
                Best: 0.167

=== Multi-layer Perceptron ===
Score: 0.21 (p-value 0.250)
                Best: 0.250

=== Random Forest ===
Score: 0.57 (p-value 0.091)
                Best: 0.091
```

Machine Learning

Random forest

- used for both classification and regression

HistGradient Boosting

- fast, scalable version of gradient boosting
- effective for large datasets.

XGBoost

- Used for regression tasks

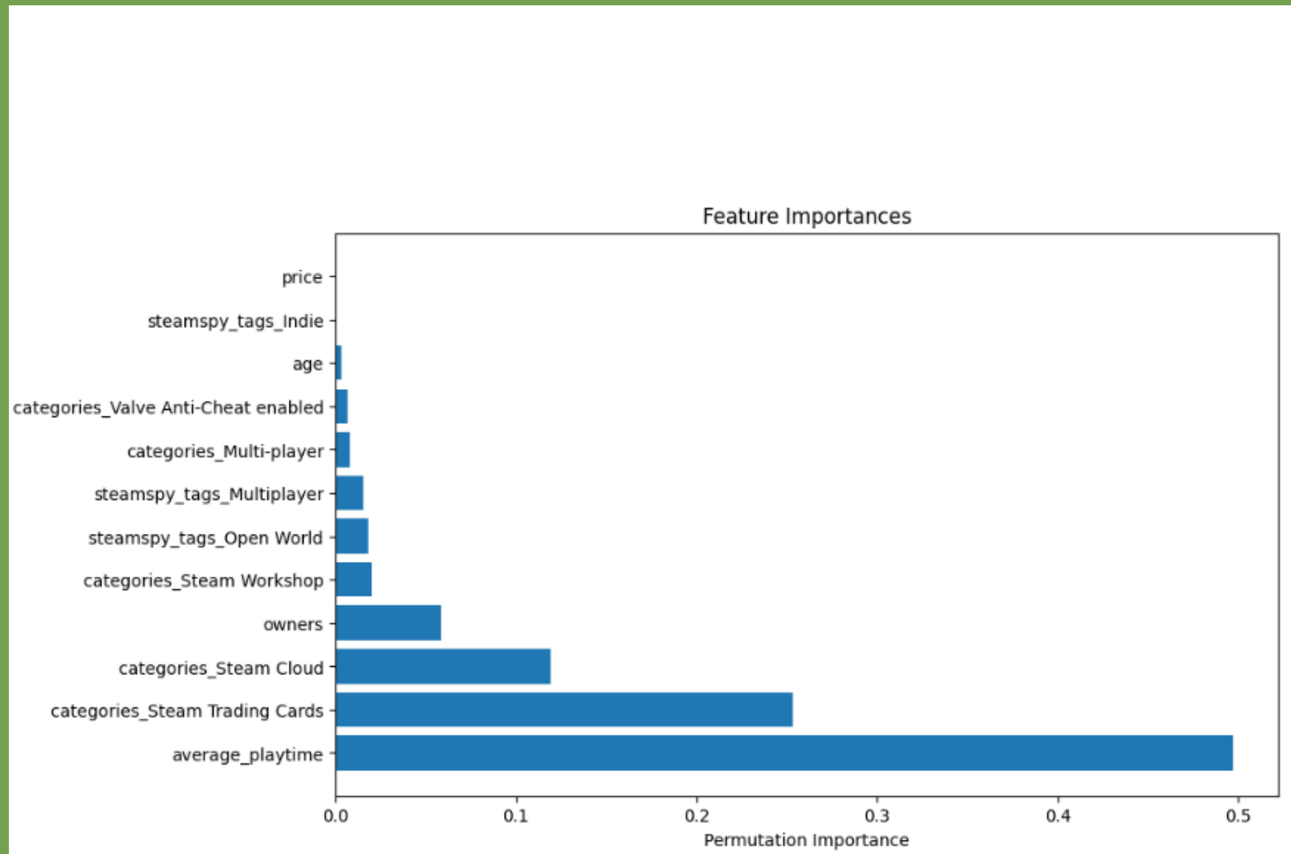
AdaBoost

- useful when dealing with complex datasets where the relationships between variables are non-linear

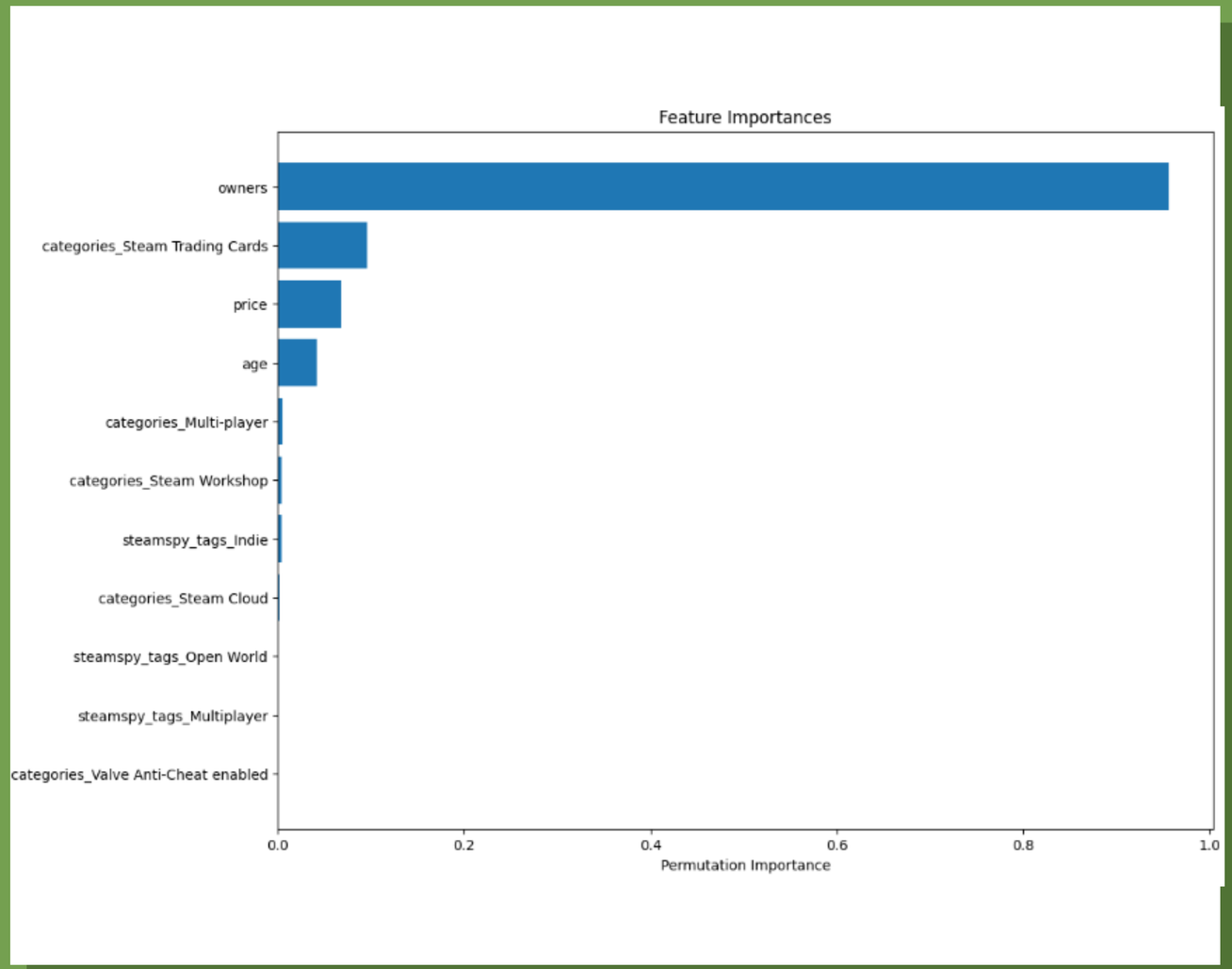
Correlation

- Here, we pick the top 12 variables that have the highest correlation with average_playtime
- These features will be used as the variables to try to predict average_playtime

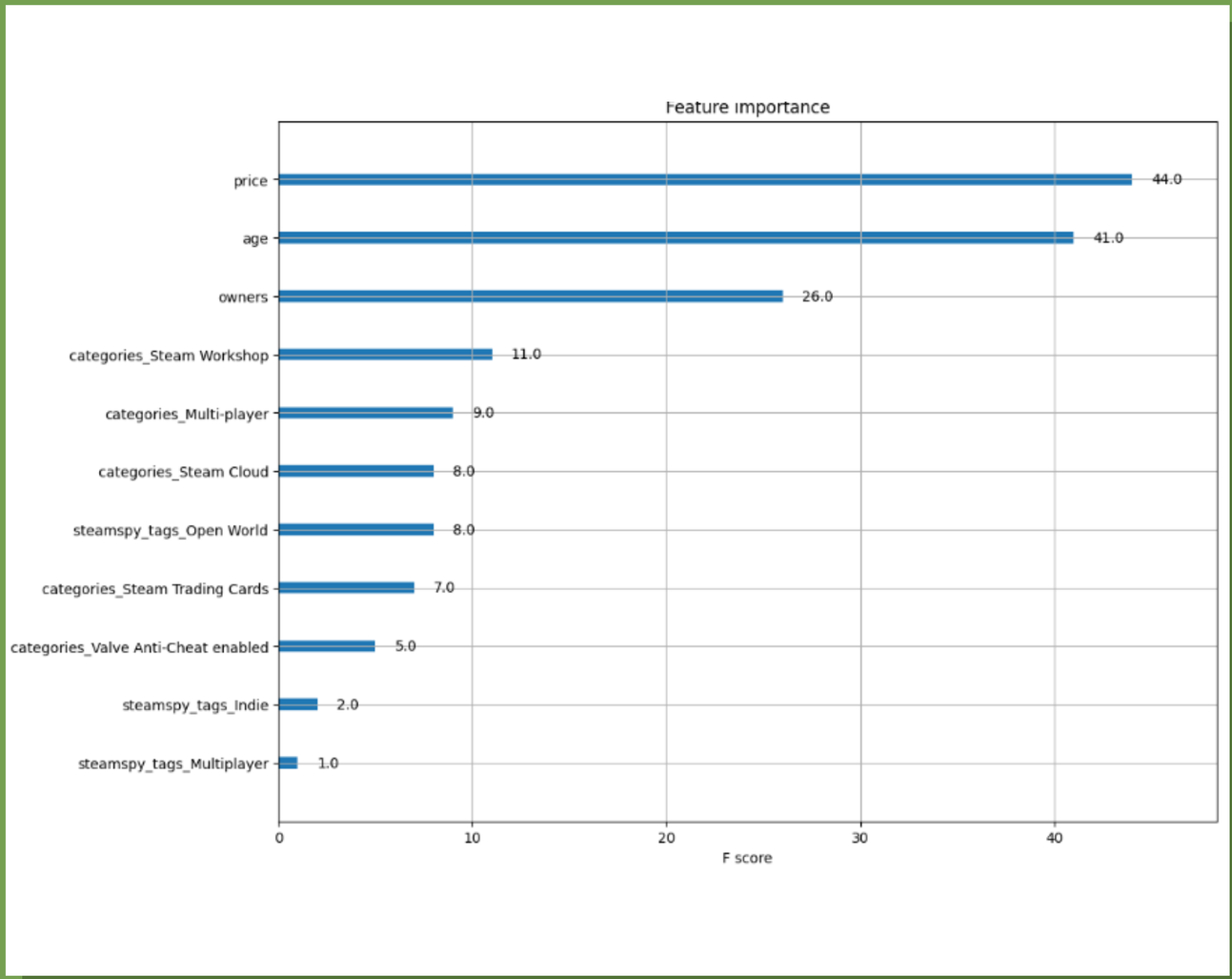
```
average_playtime                          1.000000
owners                                    0.695850
categories_Steam Trading Cards            0.338400
age                                       0.261525
steamspy_tags_Open World                  0.161696
steamspy_tags_Indie                       0.157391
categories_Valve Anti-Cheat enabled       0.125558
steamspy_tags_Multiplayer                 0.124886
categories_Multi-player                   0.122959
categories_Steam Cloud                    0.121746
categories_Steam Workshop                 0.121129
price                                     0.118525
categories_Co-op                          0.108482
steamspy_tags_Adventure                   0.107562
steamspy_tags_FPS                         0.105334
Name: average_playtime, dtype: float64
```



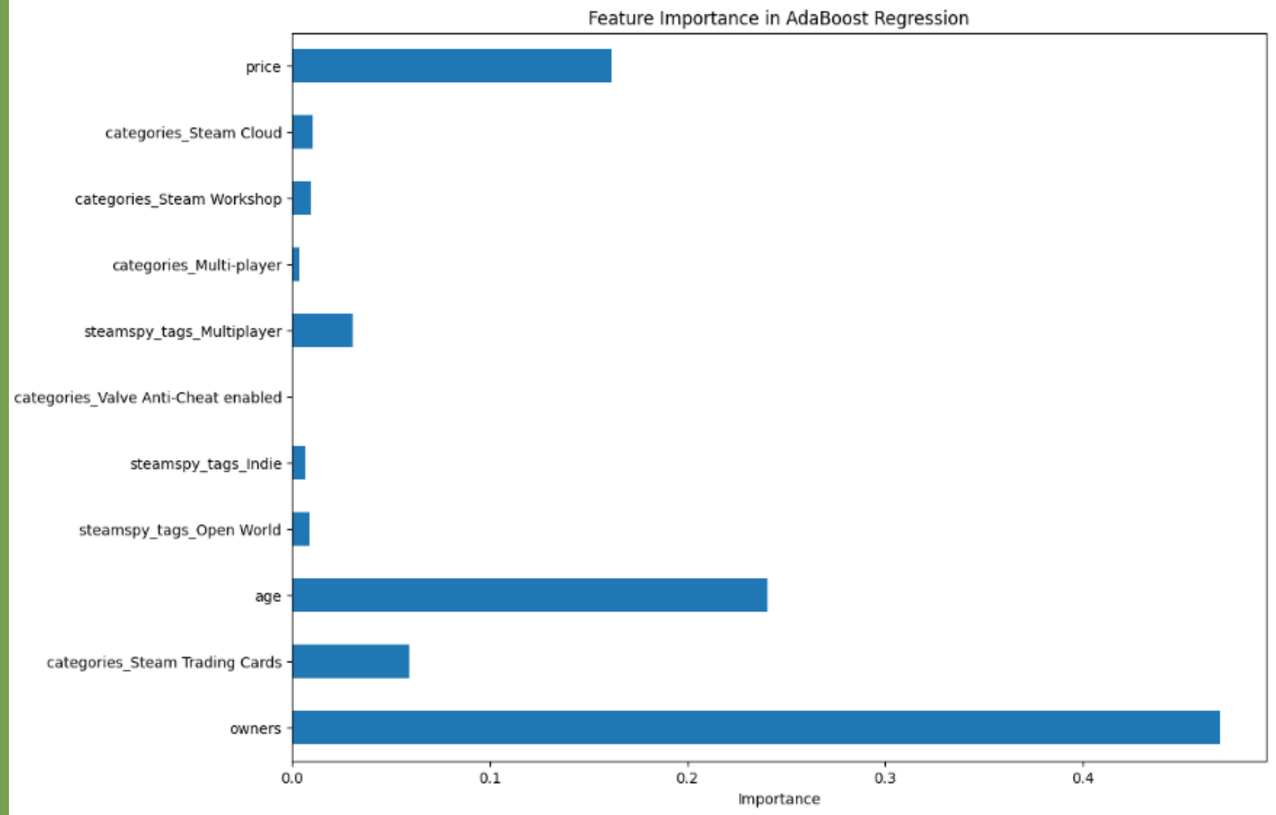
Random Forest



HistGradient Boosting Regressor


Feature Importance
price
age
owners
categories_Steam Workshop
categories_Multi-player
categories_Steam Cloud
steamspy_tags_Open World
categories_Steam Trading Cards
categories_Valve Anti-Cheat enabled
steamspy_tags_Indie
steamspy_tags_Multiplayer
44.0
41.0
26.0
11.0
9.0
8.0
8.0
7.0
5.0
2.0
1.0
0
10
20
30
40
F score

XGBoost


Feature Importance in AdaBoost Regression
price
categories_Steam Cloud
categories_Steam Workshop
categories_Multi-player
steamspy_tags_Multiplayer
categories_Valve Anti-Cheat enabled
steamspy_tags_Indie
steamspy_tags_Open World
age
categories_Steam Trading Cards
owners
0.0
0.1
0.2
0.3
0.4
Importance

AdaBoost

Conclusion

Outcome of the Project

Machine Learning

Owners

- How many people have owned the game based on how long the has been released

Age

- How long the game has been out for since 1st January 2024

Multiplayer

- Whether a game can be played with more than 1 human player

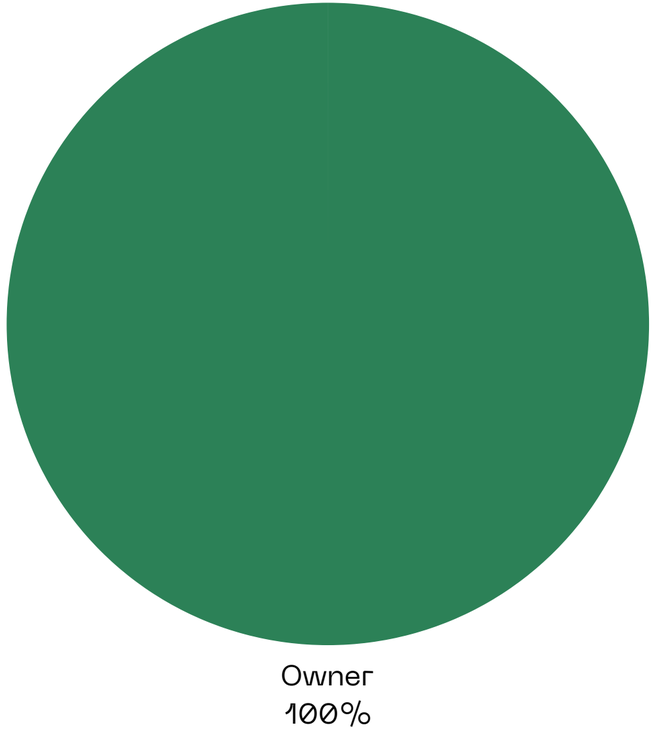
Trading Cards

- Whether the game gives you trading cards after you play the game for a certain number of hours

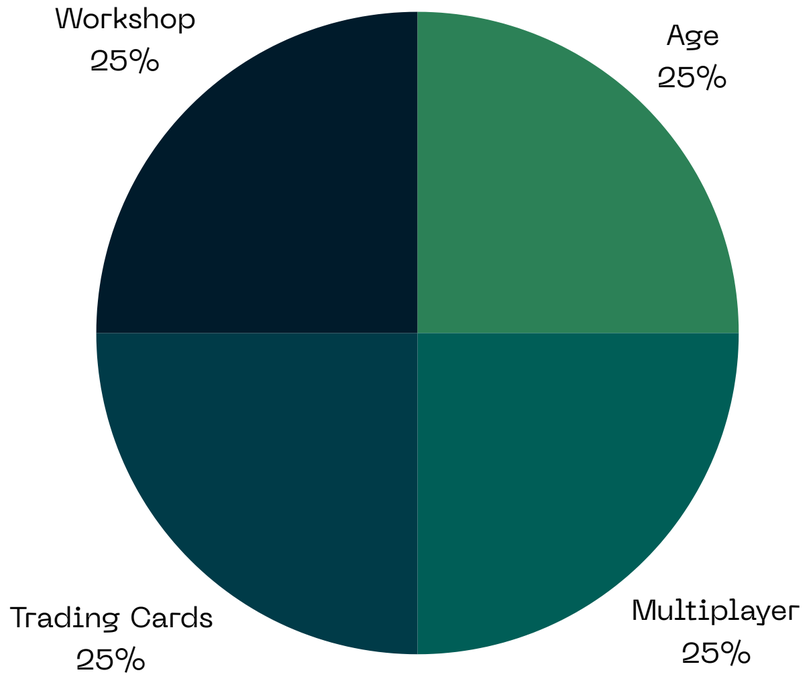
Workshop

- Whether the game is allowed and able to be modded (modifiable) from the community workshop

Factors that are on 4 out of 4 ML models



Factors that are on 3 out of 4 ML models



- **Owners have strongest correlation**
- **Others (Workshop, Age, Trading Cards, Multiplayer) slightly weaker correlation**

Insights

Popular

Multiplayer

- Play games when they are able to play it with their friends or online players

Trading Cards

- Value getting rewards as gamers would want to be able to earn money from the selling of cards and also feel rewarded

Workshop

- Customisation as gamers would be able to modify and change the game which adds to the replayability of the game

Popular

- Multiplayer
- Trading Cards
- Workshop

YOU WIN!

Thank you!

Hope you enjoy
our presentation.

YES

NO