Data Science Final Project

Doron Firman

Motivation

- Enjoyed playing video games since a young age.
- Video games provided a distraction and encouraged unique thinking.
- Introduced to the Steam platform after getting a computer.
- Used Steam as primary gaming platform for around 12 years.
- Observed growth and evolution of Steam, including new games and software.
- Developed interest in understanding the factors behind game popularity.

Business Problem:

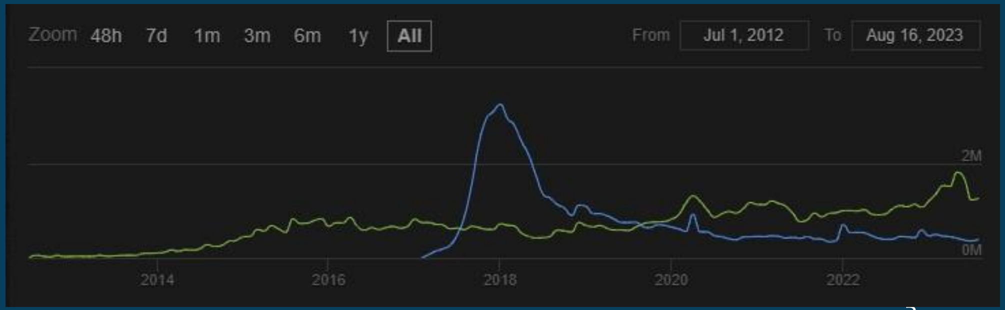
What factors influence the peak concurrent users for games or software on the Steam platform?

Introducing Steam

- Active Users: Steam has over 120 million monthly active users.
- Games Available: The platform offers more than 30,000 games across various genres.
- Peak Concurrent Users: The record peak concurrent user count was around 26 million.
- Average Playtime: The average playtime for Steam users is approximately 21 hours per week.
- Number of Developers: Steam has over 100,000 active developers.
- Market Share: Steam holds a significant market share in the PC gaming distribution space, estimated to be over 70%.
- Regional Usage: Steam is used by gamers around the world, with a strong presence in North America, Europe, and Asia.

Why Peak Concurrent Users?

- Peak concurrent users on Steam = highest users playing a game/software simultaneously
- Metric to gauge game/software success/popularity in Steam community
- Higher peak users = more successful; larger user base, active engagement
- High count shows thriving multiplayer, user retention, visibility
- Leads to positive reviews, potential sales growth



*The data was taken from https://steamcharts.com/

Data Review

Data Review

About the dataset:

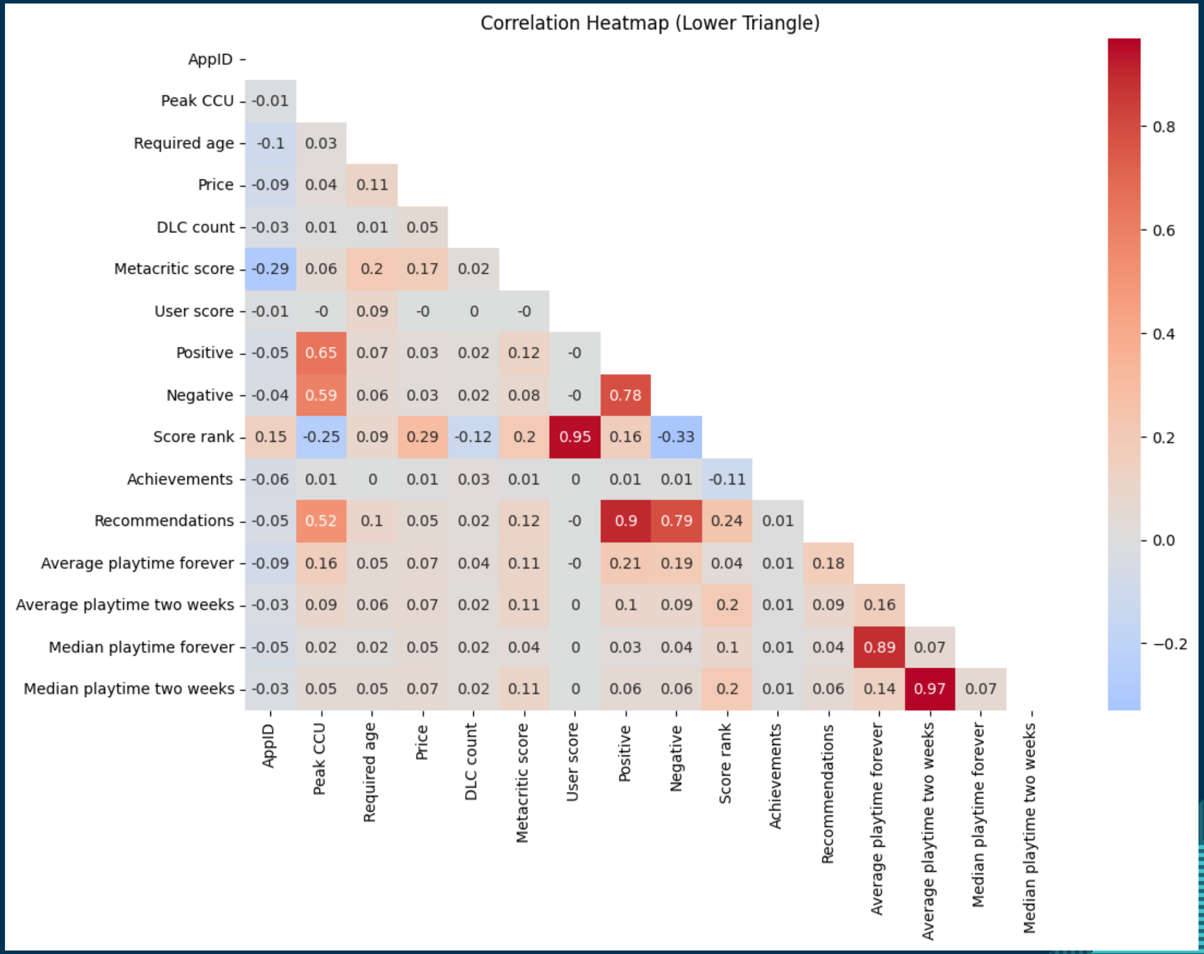
- Kaggle dataset*
- Approximately 70k rows and 39 columns
- Games and software released between the years 1997 and 2023
- Includes English and non-English games and software

*The data was taken from https://www.kaggle.com/datasets/mexwell/steamgames

Data Review

Correlation Heatmap:

- Higher **peak CCU** tend to correlate with more **positive and negative reviews** from users.
- A higher **Metacritic score** is moderately correlated with more **positive reviews** and less **negative reviews**.
- **User scores** have a weak correlation with **positive and negative reviews**.
- Higher-**priced** games and software might correlate with better **Metacritic score** and **positive reviews**.
- More **recommendations** are strongly linked with higher **positive reviews** but also with higher **negative reviews**.
- **User scores** significantly influence the games/software **ranking**.
- Longer **average playtime** relates to better **positive reviews**.
- More **DLC's** might lead to higher **peak CCU**.



Data Review

Important columns to note:

- Name (str): The name of the game or software
- Release date (date): The release date of the game or software
- Estimated owners (categorical): A range with the number of owners of a game or platform
- Peak CCU (int): The maximal number of users who played the game at the same time
- Price (int): The original price of the game or software (in US dollars)
- Platform (bool): (Windows\Linux\Mac): Dummy variables that describe which platforms the game or software supports
- Scores (int): (Metacritic\Users): The score assigned in a scale of 0-100
- Playtime (int): (Average\Median): The number of hours played per user per period
- Developers (str): The name of the game studio\s or people who developed the game
- Publishers (str): The name of the publisher\s of the game or software
- Categories (str): Steam assigned categories
- Genres (str): Developer assigned genres
- Tags (str): User assigned tags for the game or software

Feature Engineering

New important features to note:

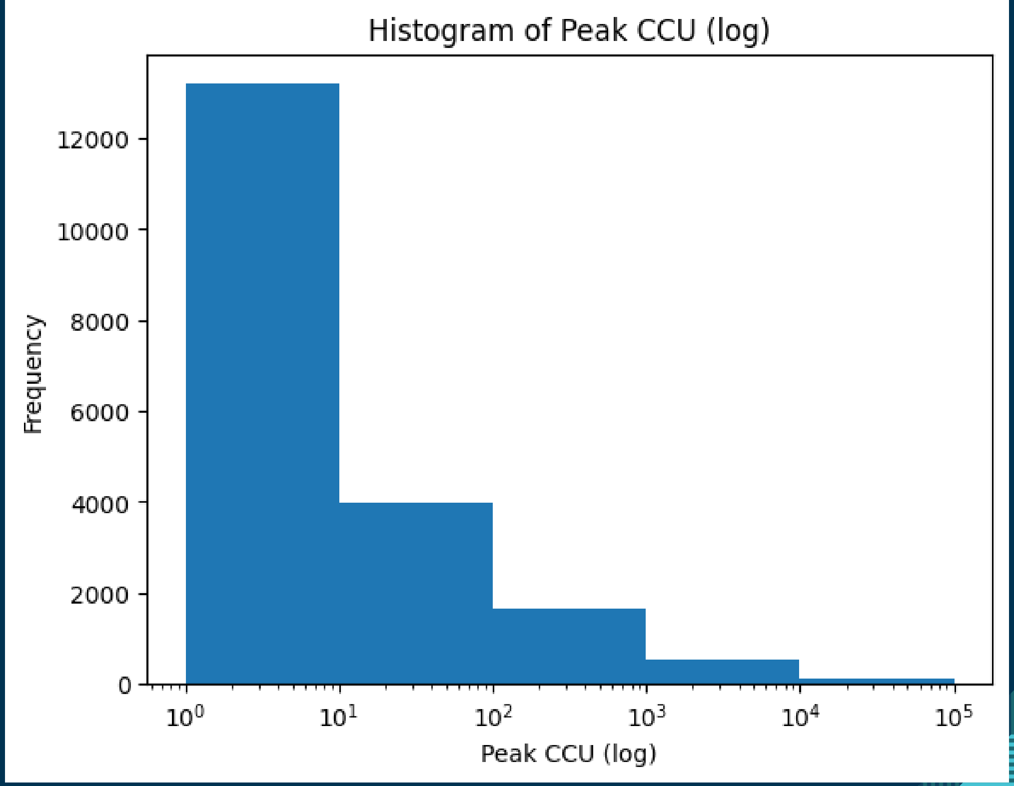
- Price range (categorical): The price range of the game
- Release by Quarter (categorical): The quarter in which the game or software was released
- (#)_Publishers (int): Number of publishers
- (#)_Categories (int): Number of categories
- (#)_Genres (int): Number of Genres
- (#)_Tags (int): Number of tags
- Review Ratio (int): The ratio of positive review out of all reviews
- Price per (x) playtime (int): The price per hour of use

Label Analysis

Label Analysis

Peak CCU histogram:

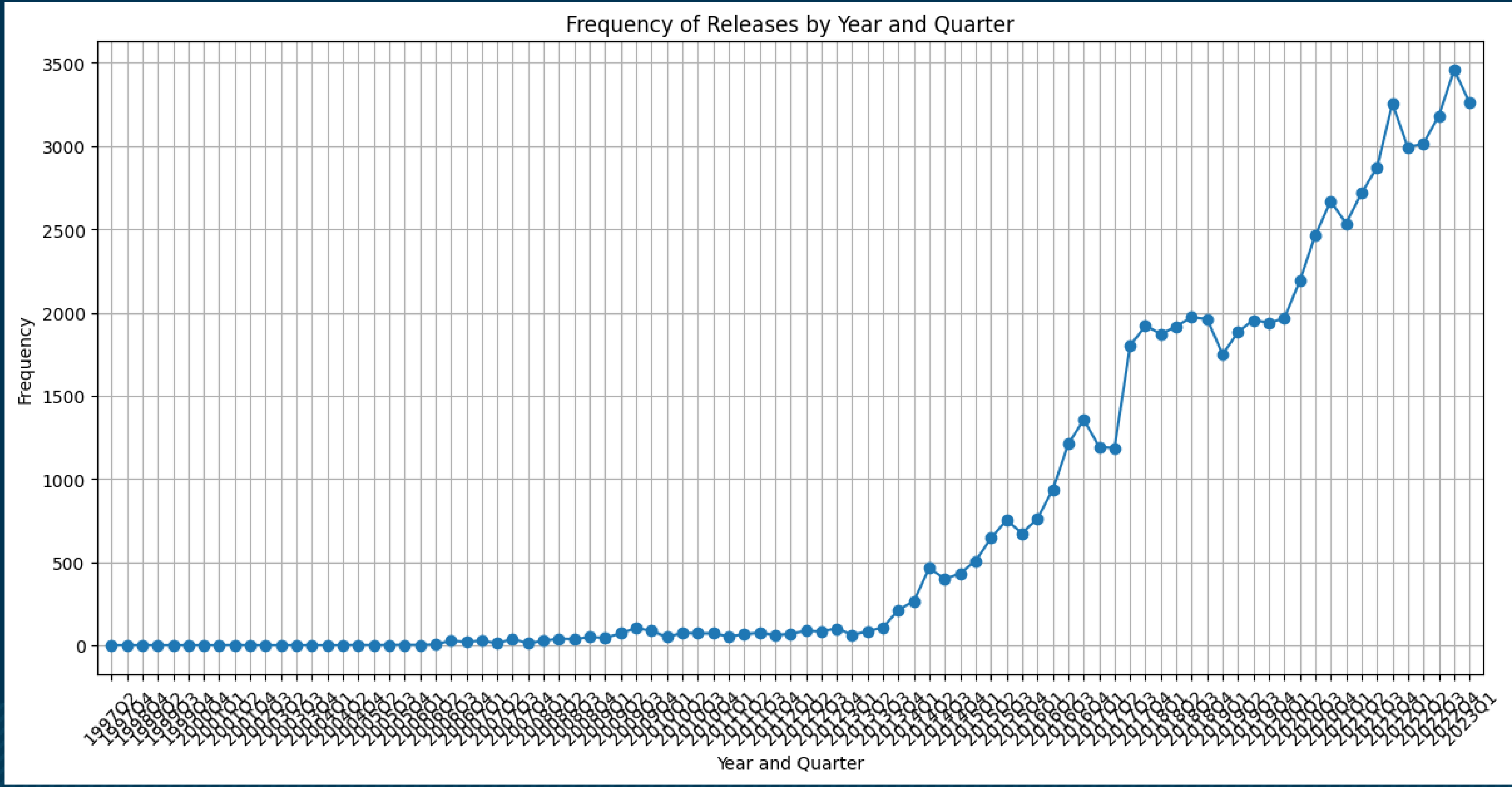
- We can see that most gams have a low number of peak CCU
- We can also see that it is very rare to have more than 10k peak CCU



Frequency Analysis

Frequency Analysis

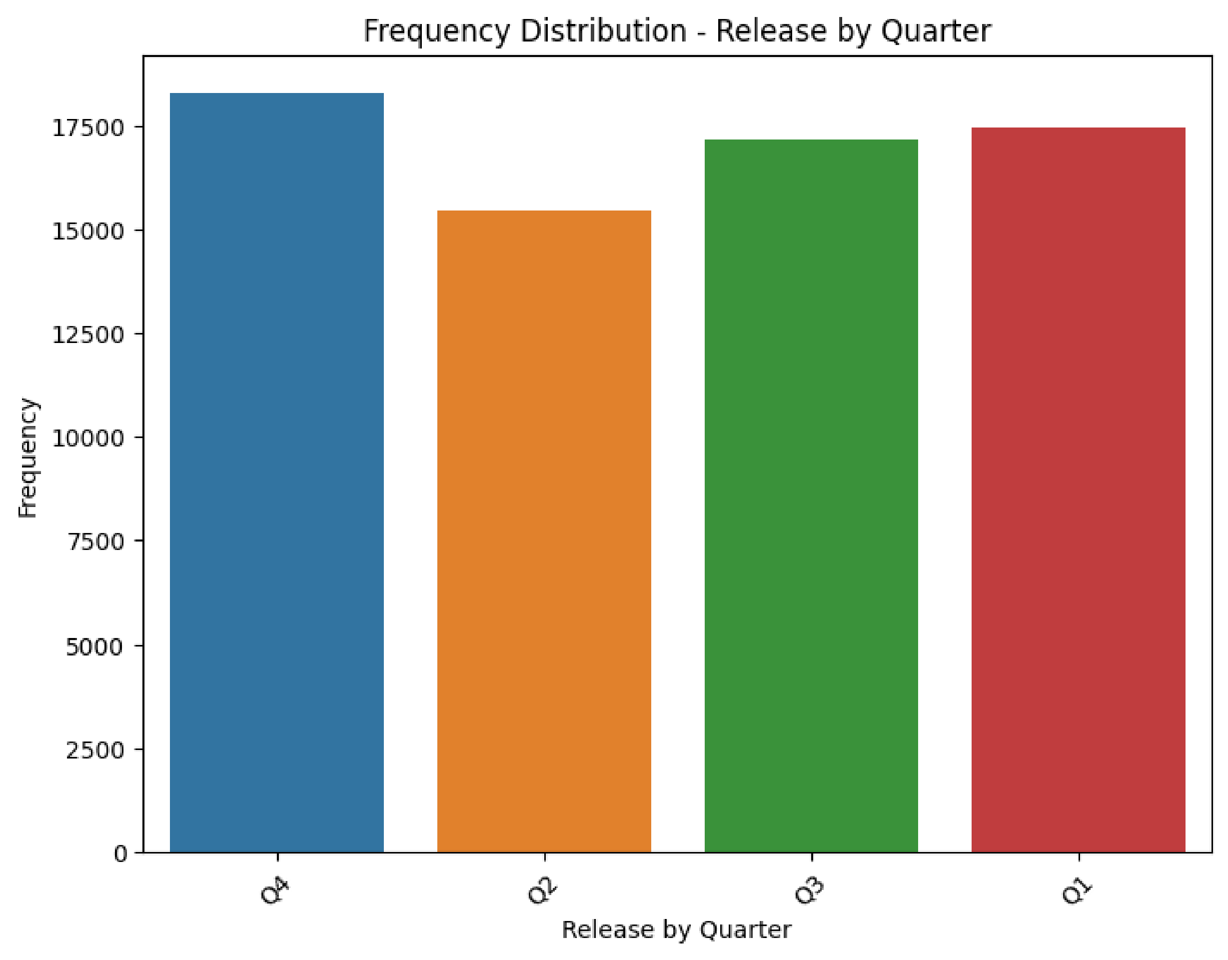
Year and quarter frequency:



Frequency Analysis

Quarter frequency:

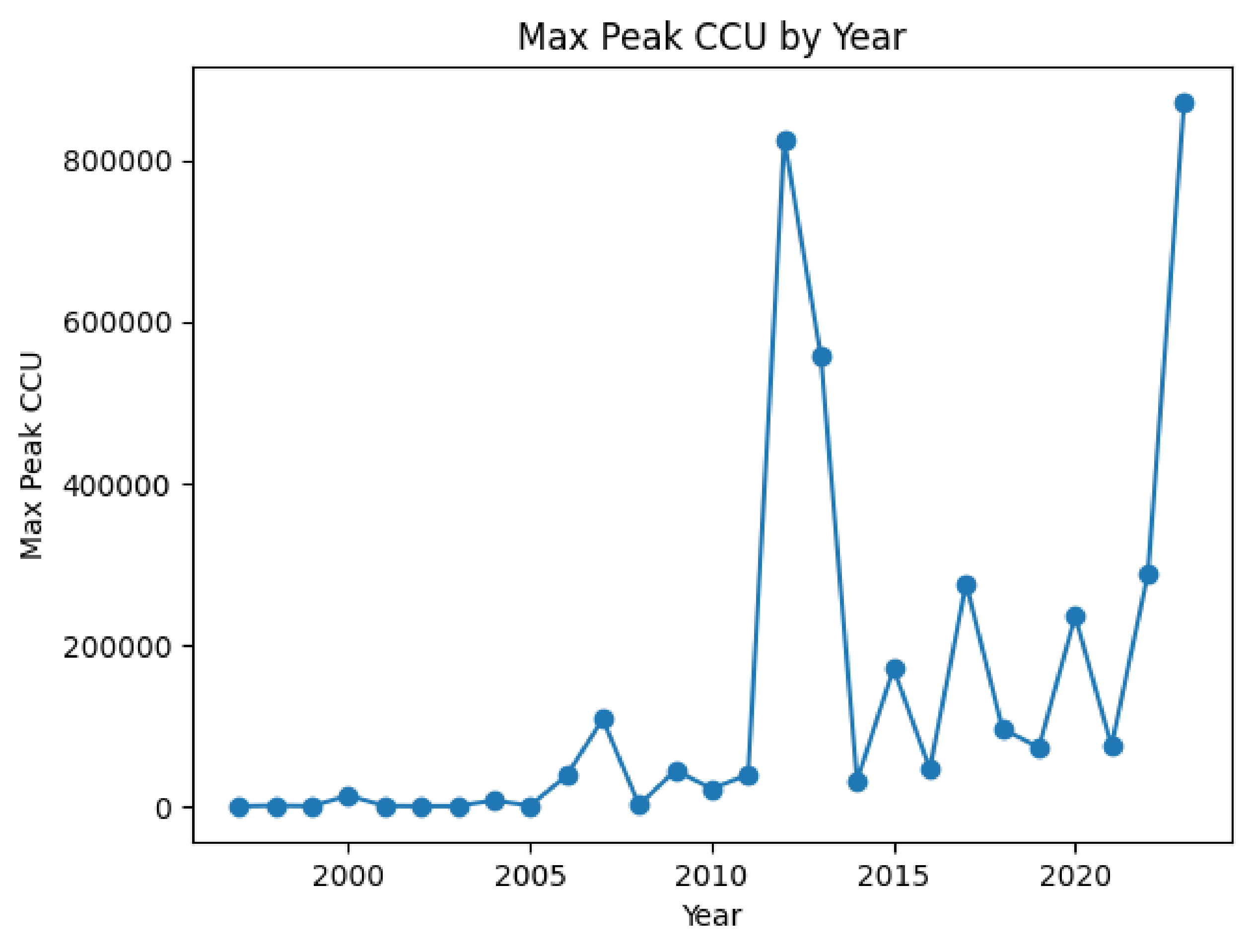
- Evenly distributed
- Q4 has the most releases
- Q2 has the least releases (although 2023 Q2 and onwards is not in the data)



Frequency Analysis

Maximal peak CCU for a game by year:

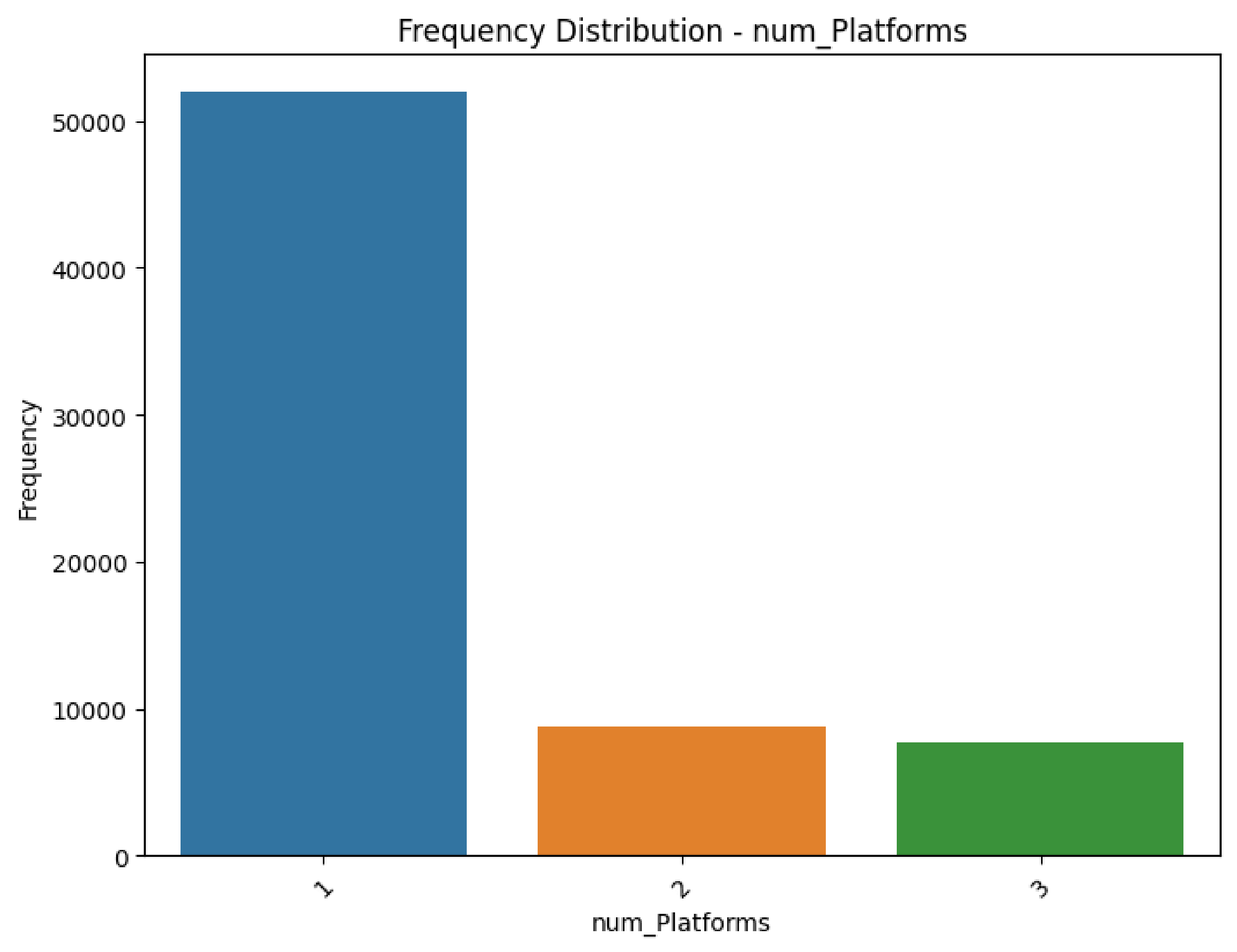
- We can see a gradual increase over the years
- On average excluding 2012, 2013 and 2023 we can expect the maximal peak CCU to be at around 200K-300K



Frequency Analysis

Frequency distribution of (#) of platforms

- We can see most games support a single operating system
- We can also see that the number of games supporting 2 and 3 operating systems are relatively equal



Frequency Analysis

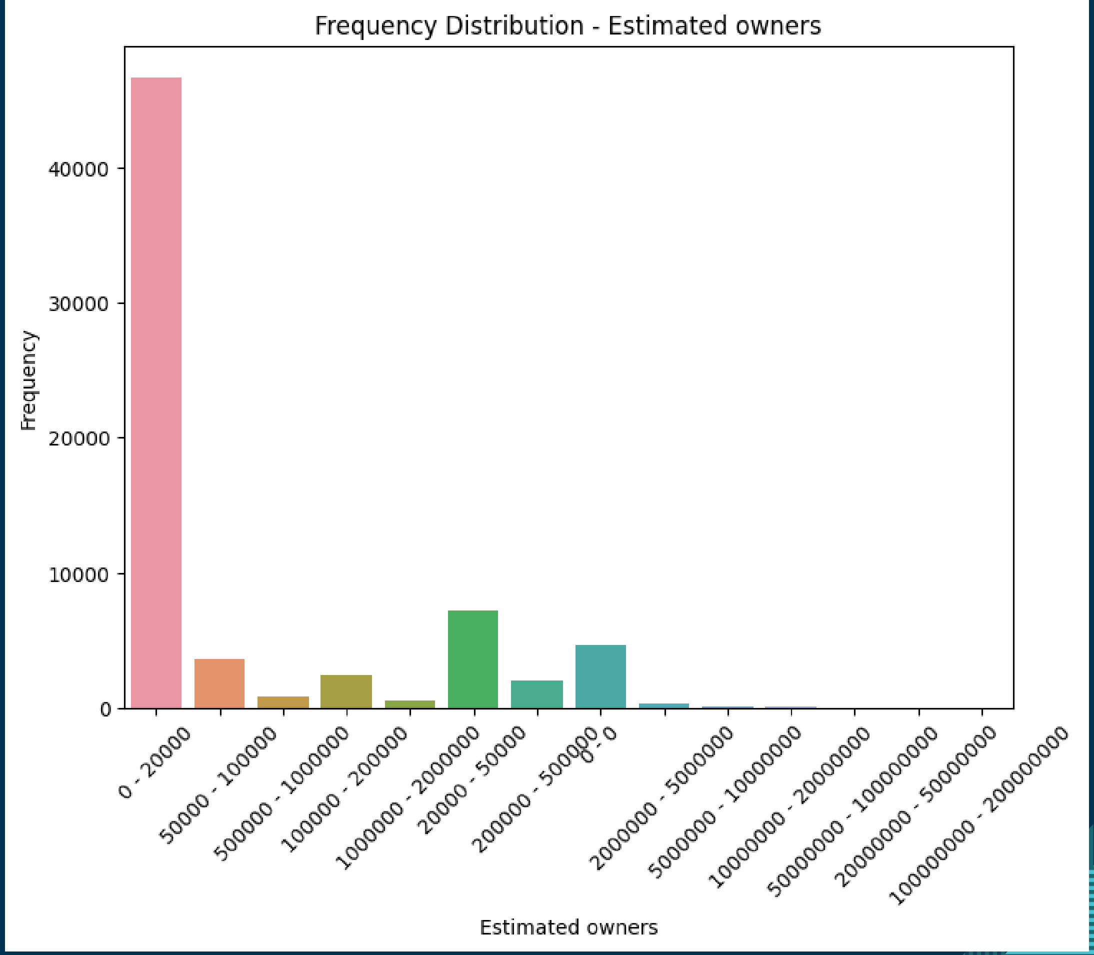
Frequency distribution of estimated owners

- We can see that most of the game have

  between 1-20k owners

- We can also see that having above 500k

  owners is rare with less than 5% of the games



Frequency Analysis

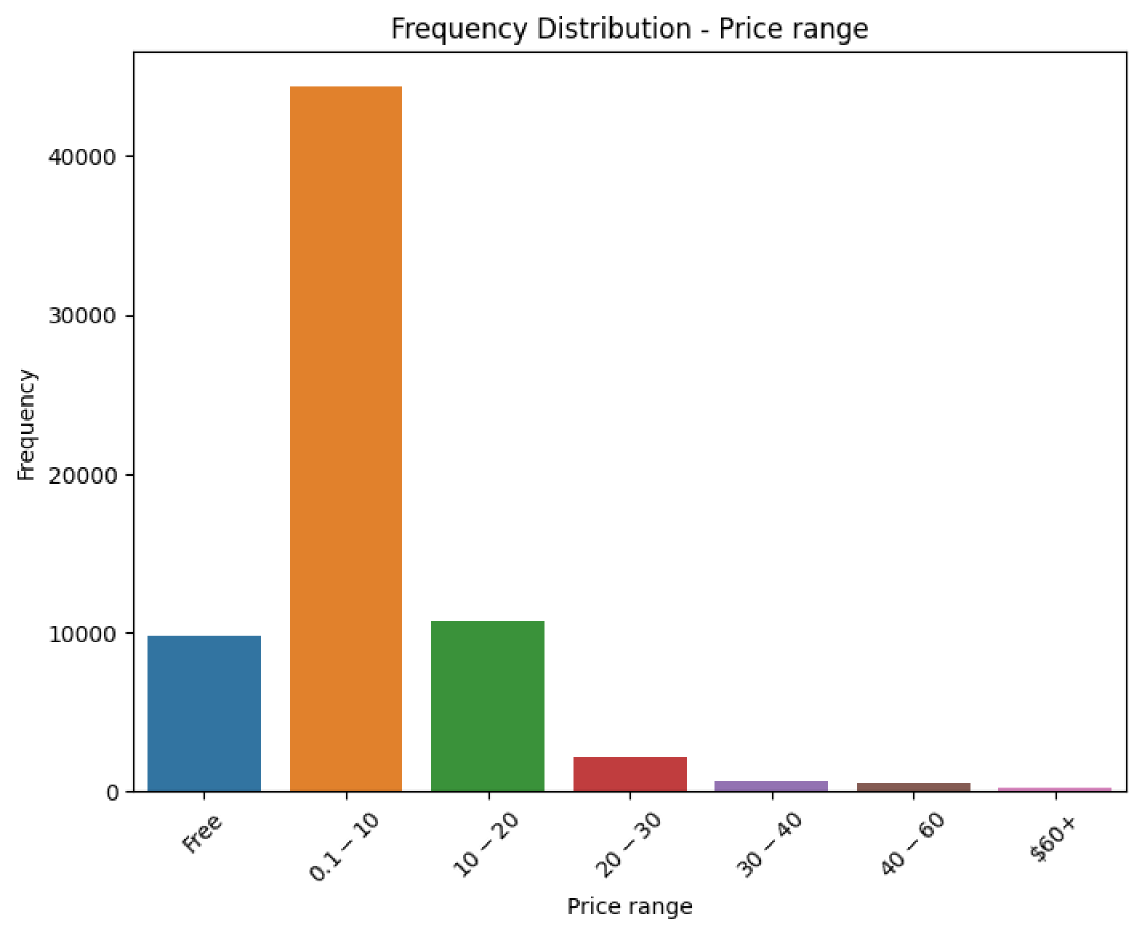
Frequency distribution of price range

- We can see that most of the game are priced

  between 0.1-10 USD

- We can also see that having a price above 20

  USD is rare with less than 5% of the games

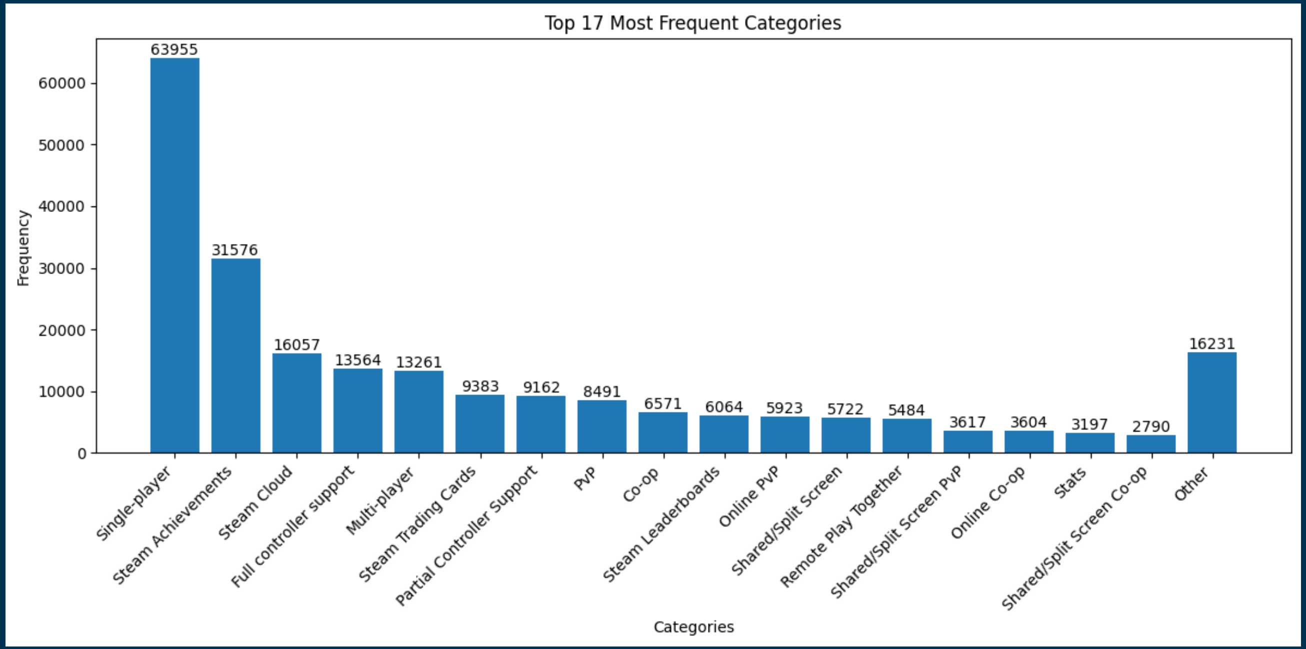


Dummy Variables

Dummy Variables

Top Categories

- We can see that “single-player” is the biggest category
- Every category which appeared less than roughly 5% was joined together into the ‘Other’ category



Outlier Detection

Outlier Detection

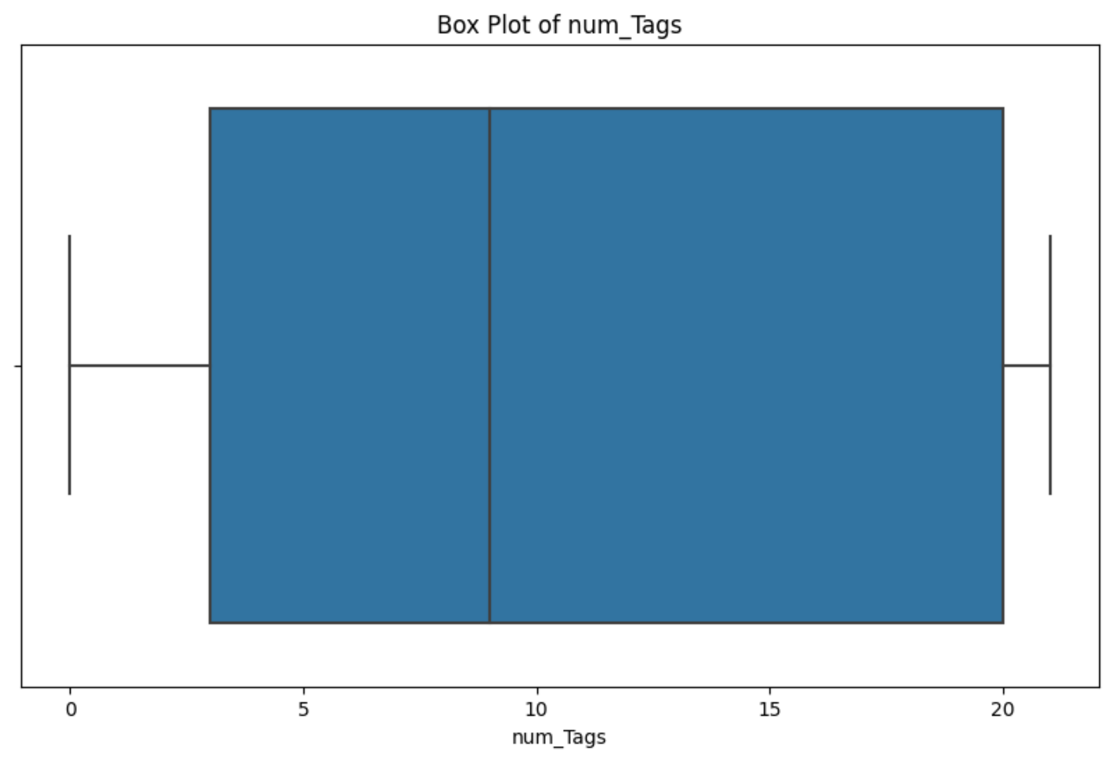
Assumptions:

- For this part I used IQR Visualization which means Box Plot with IQR-based Thresholds
- Assumed that the data has a skewed distribution.
- Assumed that the interquartile range (IQR) is a robust measure of spread.
- Assumed that outliers can be identified based on being below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.

Outlier Detection

Outlier detection of number of tags

- We can see that most of the values are between 3-20 tags
- We can also see that the average value is about 9 tags per game or software

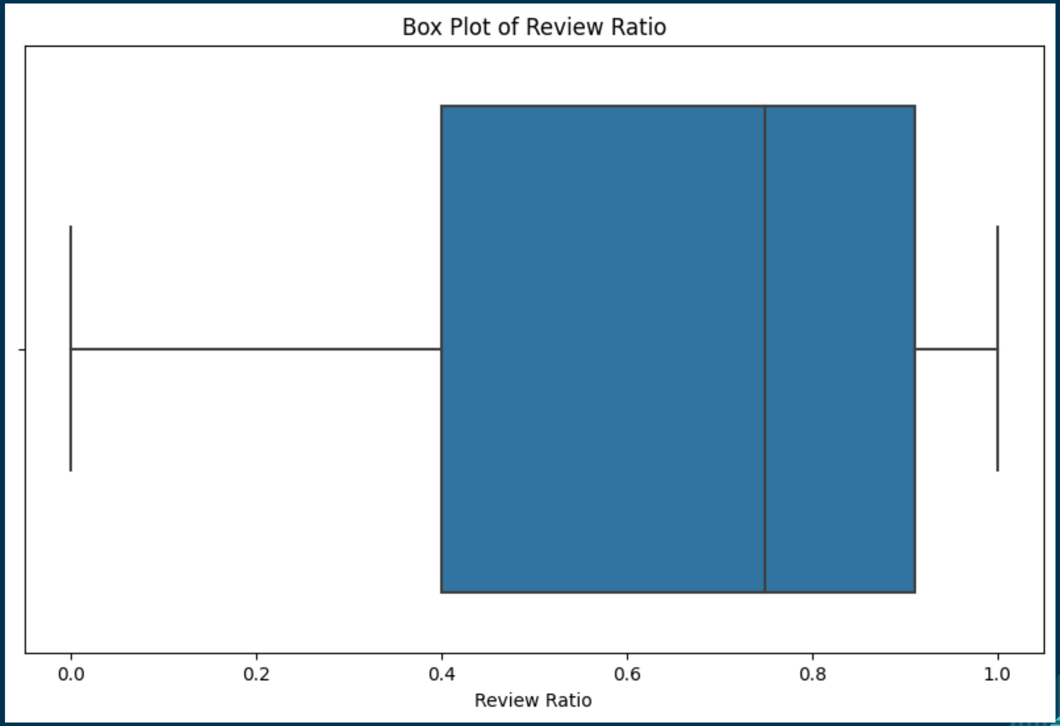


Outlier Detection

Outlier detection of number of tags

- We can that most of the ratios lie between 0.4-0.9 with the average ratio at about 0.7 per game or software

- This tells us that in general most games have more positive reviews than negative



Model Training

Testing and Training

The Models:

For this part I chose to use 4 models which are:

- Linear Regression
- Lasso
- Random Forest
- xGradiant Boosting

I measured the performance of each model based on the following criteria: R- square, MSE, R-MSE and MAE

Testing and Training

Model Performance Results:

- We can see from the results that RF and xGB performed quite well with xGB being a bit better.

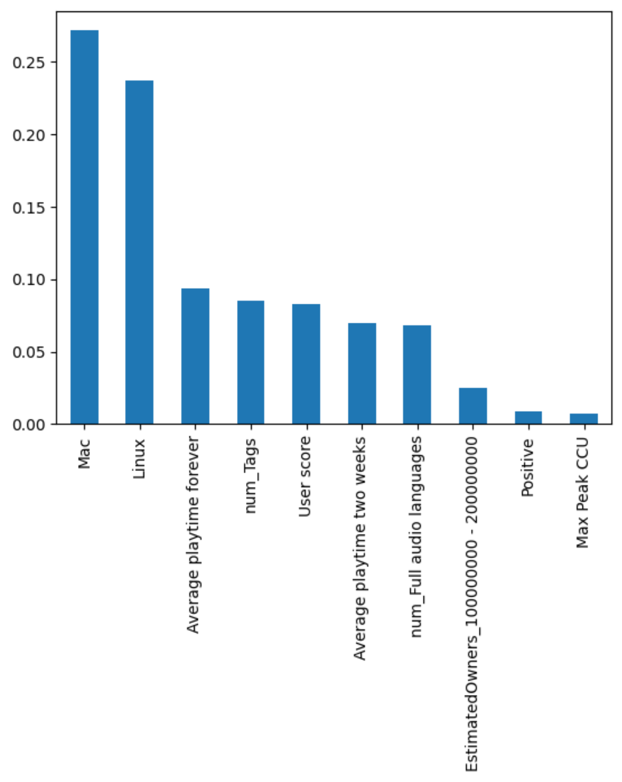
Model	R-Square	MSE	R-MSE	MAE
Random Forest	0.36	6824535.6	2612.38	110.73
Linear Regression	0.43	6063429.01	2462.4	397.4
Lasso	0.37	6676972.77	2583.98	353.16
XGB	0.28	7641913.73	2764.4	107.44

Testing and Training

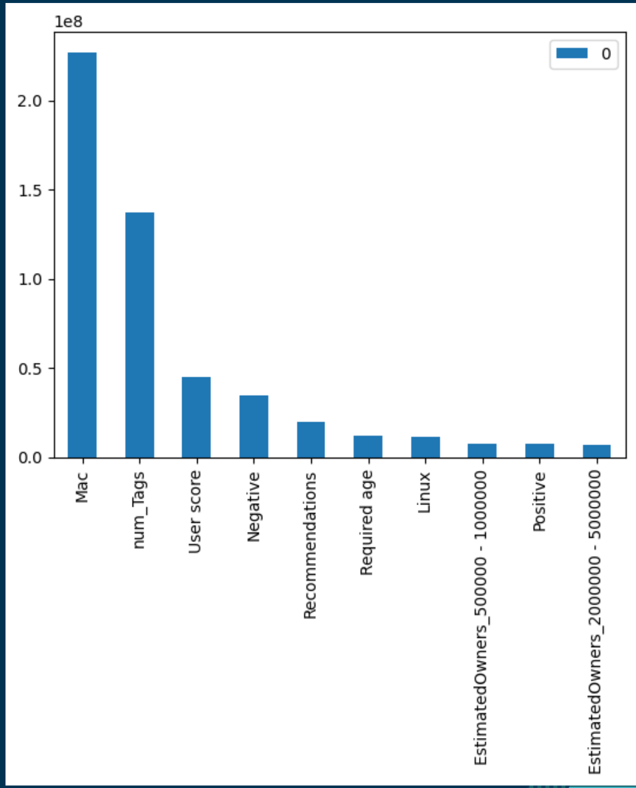
Feature Importances:

- We can see that in both models we get mostly the same features
- For both cases it seems that supporting Mac increases the prediction rate
- All in all, I believe that the features appearing here are very logical because implementing these features allows for a greater target market or imply a high number of users

Random Forest – Feature Importances



xGradiant Boosting– Feature Importances



Testing and Training

Model Improvement:

For this part I chose to take the RF model (due to time concerns) and try to improve its results using grid search.

I then used the same metrics and compared both models.

We can see the big improvement which yields the best overall results compared to all the previous models

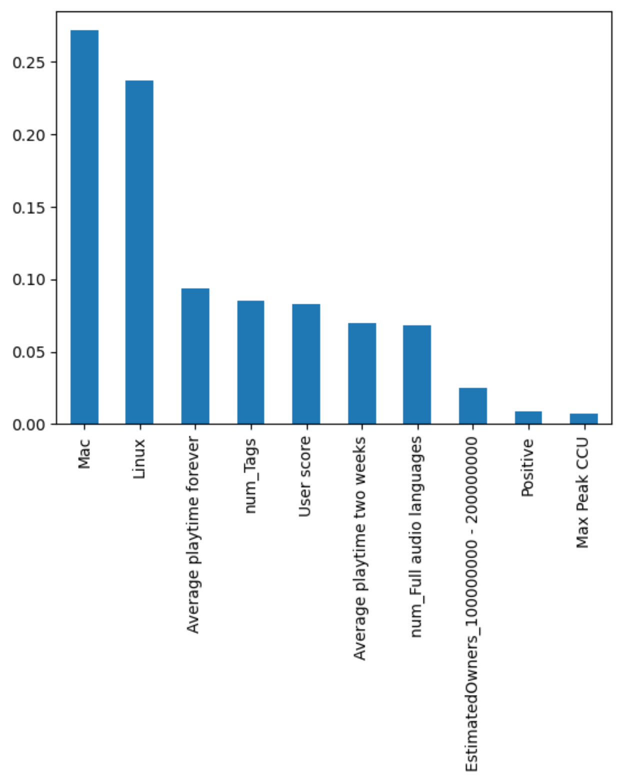
Model	R-Square	MSE	R-MSE	MAE
Random Forest	0.36	6824535.6	2612.38	110.73
RF - Grid Search	0.46	5683429.33	2383.99	101.59

Testing and Training

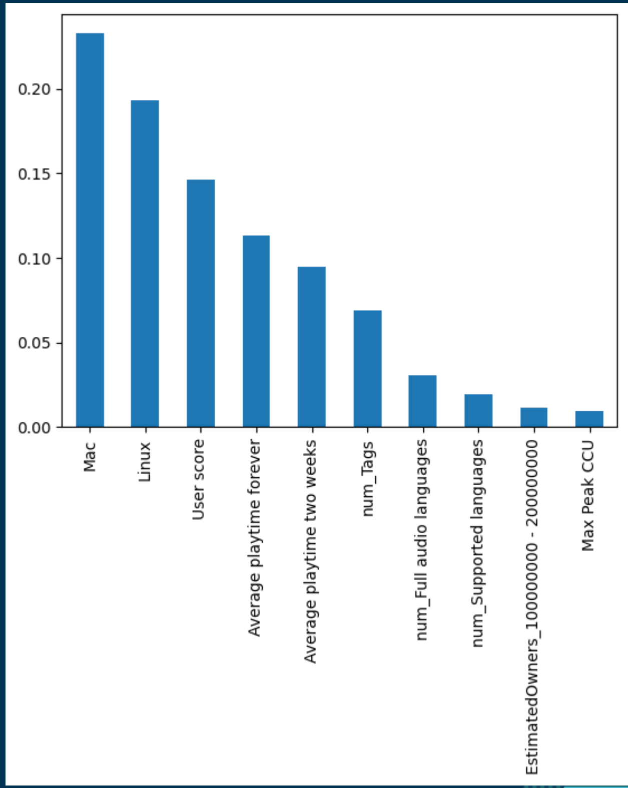
Feature Importances:

- We can see that in both models we get mostly the same features
- The only visible change is between positive and number of supported languages
- What seems surprising to me is that user reviews have much less effect than one might think
- What is unsurprising is that quarter of release is non apparent at all

Random Forest – Feature Importances



RF Grid Search– Feature Importances



PCA and Clustering

PCA and Clustering

The Process:

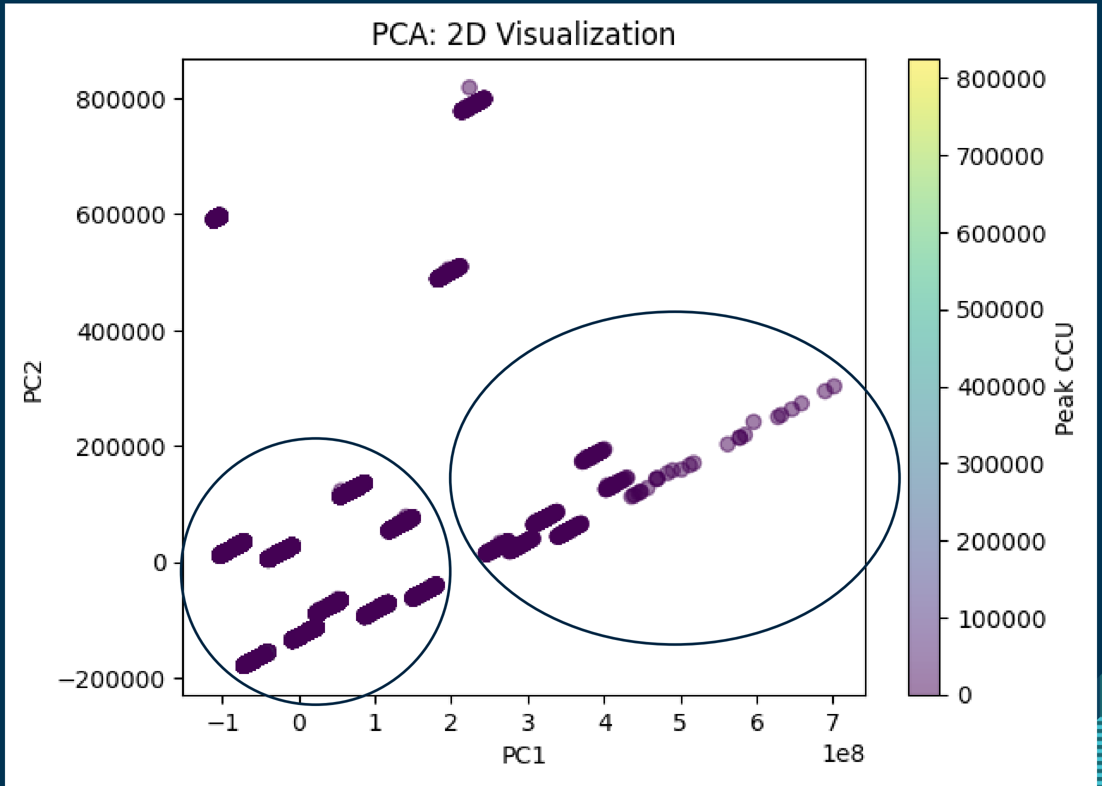
For this part I used the following:

- PCA reduction to 2-dimensions
- Elbow function and silhouette score to determine optimal number of clusters
- Clustering of all the games and platforms

PCA and Clustering

PCA

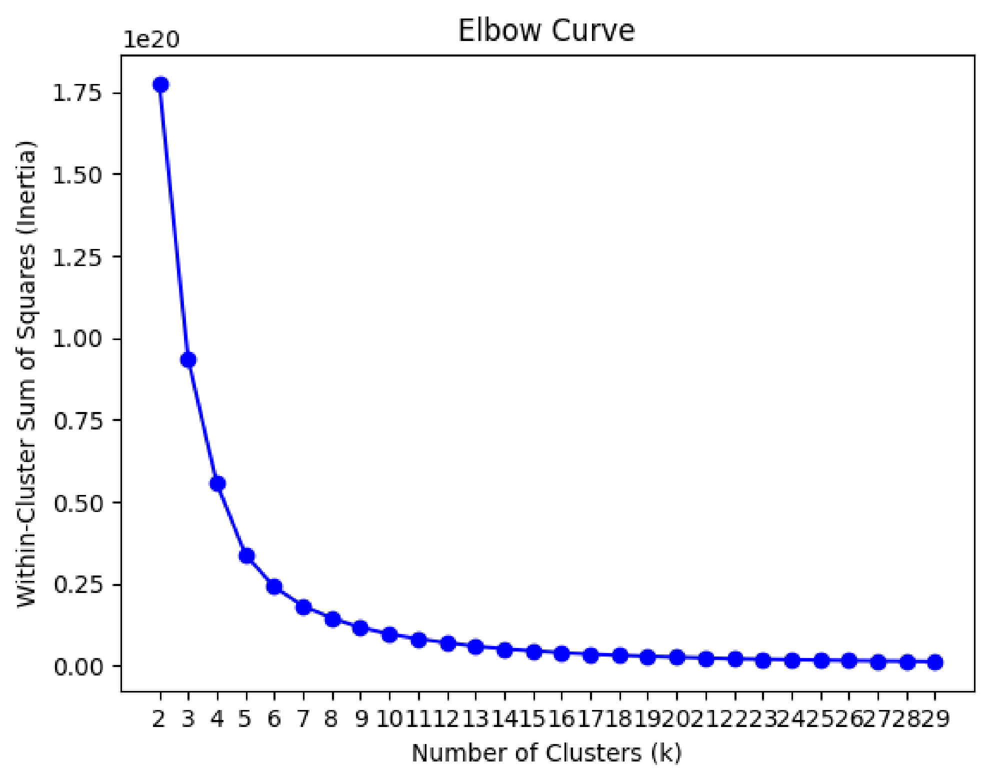
- We can see that there is a positive correlation between PC1 and PC2
- We can also see that there are two clusters of data points, one at the bottom left and one at the middle right



PCA and Clustering

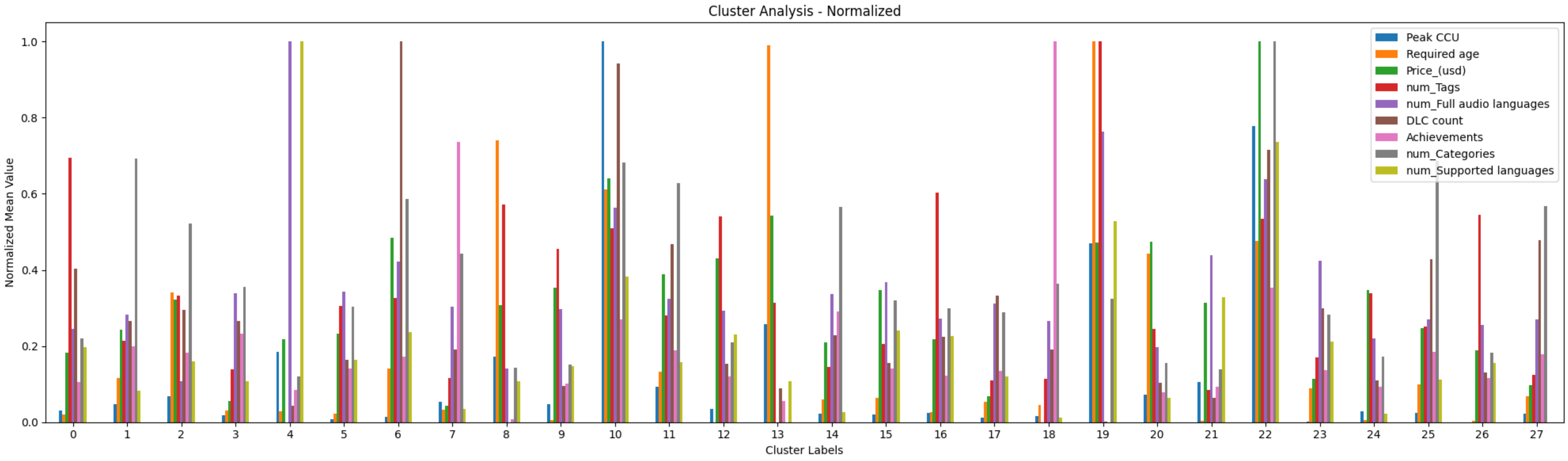
Elbow Function

- We can see that for the following elbow function it is hard to determine what is the optimal number of clusters
- Using silhouette score I found that 28 clusters has the minimal value



PCA and Clustering

Clustering

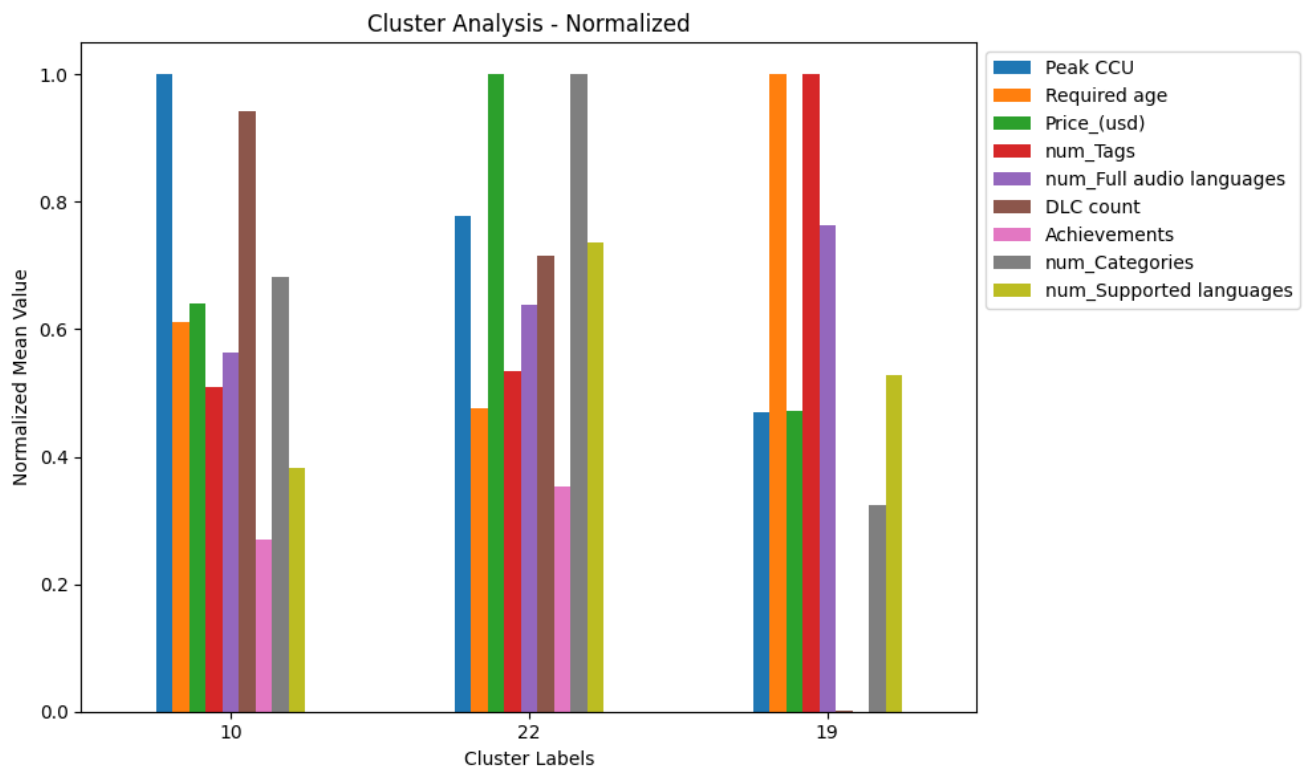


PCA and Clustering

Clustering Motivation:

- For this part I wanted to see how developers can try and manipulate the peak CCU *before* releasing the game

- We can see 3 clusters with the highest peak CCU which are 10, 22 and 19



PCA and Clustering

Clustering Conclusions:

Games and software in these clusters tend to have relatively high values for almost all the chosen features, with the exception being “Achievements”. Specific features that are lower include “num_Categories” and “DLC_count” for cluster 19

Conclusions

Conclusions

Conclusions:

- The project focused on analyzing factors affecting peak CCU on Steam for games/software.
- Aims to guide developers on achieving higher peak CCU.
- Findings from analysis:
  - Publishing for Mac and Linux has significant positive impact.
  - More tags, higher user rating, and more reviews correlate with higher peak CCU.
  - Increasing supported languages and audio has positive influence.
  - Post-release support through DLC contributes to higher peak CCU.
- Supporting additional platforms requires increased investment, impacting costs.
- Suggests that being a AAA game or software could encompass and cover various identified factors.

Further Research

Further Research

Further Research:

- Future projects could involve integrating current data with data from other platforms (e.g., Uplay, Origin, GOG, Epic Games Store). Aims to enhance accuracy and confirm research conclusions across multiple platforms.
- Another approach could be integrating data from console sales for broader insights.
- Potential inclusion of HLTB (How Long To Beat) database, offering "real" game length based on player time. HLTB data might reveal new insights and provide additional information for analysis.

Thank You