

Mahidol University International College

ICBE484: Data Analytics for Business Strategy

Trimester 3 Year 2023

**The application of supervised machine learning techniques to predict concurrent player (CCU) on Steam**

Submitted to

Dr. Wasin Siwasarit

Mahidol University International College

Submitted by

Stephen Eberlein 6181254

Tharadorn Pornkarnchananun 6381140

Timothy Tosukhowong 6580219

Chinaphat Singyoo 6380905

### 

### 

# Table of Contents

[**Table of Contents**](#_heading=h.p9iiul4fups) **1**

[**Introduction**](#_heading=h.h5jvgo4m18cj) **2**

[Dataset Description](#_heading=h.tj9kornevhum) 3

[Importance of Predicting Concurrent Player Counts](#_heading=h.t7c9gjxob4fs) 4

[Dataset Features](#_heading=h.wiywbuqfn87y) 4

[Analytical Approach:](#_heading=h.wdwwsdgg823) 7

[Data Cleaning](#_heading=h.bjlfbh9n4fes) 7

[Feature Variables Explanation](#_heading=h.shdbfjnx9orj) 8

[Limitations of Converting Categorical Variables to Dummy Variables](#_heading=h.r52u2x1lwgtr) 9

[Correlation Map](#_heading=h.562v8ybg0agq) 11

[Regressions](#_heading=h.6o3ny1ad9px0) 11

[Inner Workings](#_heading=h.btbuz6qsbr7k) 12

[Implementation and Interpretation](#_heading=h.ibur8kg7cl67) 12

[**Experimental methodology:**](#_heading=h.9jy9489dcn6n) **13**

[**Results:**](#_heading=h.xkysftjufkon) **16**

[Performance Metrics:](#_heading=h.uboharo4guvx) 16

[Model Comparisons:](#_heading=h.ysokgppemsou) 17

[Overfitting and Underfitting:](#_heading=h.jwzau4v7rzjv) 17

[Model Simplicity vs. Regularization:](#_heading=h.nv700pgjmhvt) 17

[K-Fold Analysis:](#_heading=h.gz6erw4mqpw2) 18

[OLS Regression Results:](#_heading=h.wnaiapx62hwp) 19

[Model Summary:](#_heading=h.ioa4rzb6bsti) 20

[Coefficients and Significance:](#_heading=h.4ocpa1uca16g) 20

[**Discussion:**](#_heading=h.fsnzj3nm0xk2) **21**

[Data Characteristics:](#_heading=h.x8rnyoyur70i) 21

[Data Quality and Feature Relevance:](#_heading=h.rubq6o7bnafy) 21

[Data Size:](#_heading=h.nigdrq8lai65) 22

[Methodology:](#_heading=h.q02n61aug60h) 22

[Parameter Tuning:](#_heading=h.f1ocujloevmr) 22

[Model Assumptions:](#_heading=h.kb862h6gq1t2) 22

[Regression Discussion:](#_heading=h.5b388b9n1tg) 22

[**Conclusion:**](#_heading=h.3vsbestny4rx) **24**

[**References**](#_heading=h.ao6xyifrpy4m) **25**

# 

# Introduction

The application of supervised machine learning techniques to predict concurrent player counts (CCU) on Steam, a premier digital distribution platform for video games, presents a fascinating and complex challenge. This research focuses on leveraging linear regression, lasso regression, and ridge regression models to analyze and forecast CCU. These methods are chosen for their capability to handle varying data complexities, perform effective feature selection, and provide robust predictions.

Predicting CCU is an intriguing and significant problem for several reasons. The gaming industry has seen rapid growth, with millions of players engaging with a vast array of games on platforms like Steam. The number of concurrent players at any given time is a critical metric that reflects a game's popularity and user engagement. This metric is essential not only for game developers but also for server management teams and marketing strategists. High CCU can indicate a thriving game community, but it also poses challenges such as server overload, which can lead to a poor user experience. Conversely, understanding low CCU trends can help identify issues within the game or broader market trends affecting player engagement (Hastie, Tibshirani, & Friedman, 2009).

Accurately predicting CCU allows developers to optimize server resources, ensuring that infrastructure can handle peak loads without compromising performance. This optimization is crucial for maintaining a smooth gaming experience, which in turn enhances player retention and satisfaction. Additionally, insights into player behavior can inform targeted marketing campaigns, helping to maximize engagement and revenue by reaching potential players more effectively (James et al., 2013).

The experiments conducted in this research are compelling due to the advanced nature of the machine learning techniques applied and their potential impact on real-world applications. Linear regression serves as a foundational model that provides a straightforward approach to understanding relationships between various game features and CCU. However, it often struggles with multicollinearity and overfitting, issues that lasso and ridge regression address effectively (Hastie et al., 2009).

Lasso regression introduces a penalty term for the absolute size of the regression coefficients, promoting sparsity in the model by effectively performing variable selection. This characteristic is particularly useful in high-dimensional datasets common in gaming analytics, where numerous features might be considered. By selecting only the most relevant features, lasso regression helps in creating a more interpretable and robust model (Tibshirani, 1996).

Ridge regression, on the other hand, adds a penalty term for the square of the coefficients. This method is adept at handling multicollinearity among predictors, leading to more stable and reliable estimates. Ridge regression's ability to distribute the coefficient values more evenly prevents any single predictor from dominating the model, thereby enhancing its generalization capabilities (Hoerl & Kennard, 1970).

Our research not only implements these techniques but also conducts a comparative analysis to highlight their respective strengths and limitations in predicting CCU. The results from these experiments are expected to provide valuable insights into which models perform best under different conditions, guiding future research and practical applications in the gaming industry.

## Dataset Description

The dataset used in this study is sourced from Kaggle, specifically the "Steam Games Dataset" (<https://www.kaggle.com/datasets/fronkongames/steam-games-dataset>). This dataset comprises detailed information about games available on the Steam platform, including attributes such as game titles, release dates, developers, publishers, genres, tags, and various player metrics. Notably, it includes historical data on CCU, user ratings, and pricing information. The richness and granularity of this dataset make it an ideal candidate for applying machine learning techniques to predict player engagement metrics. The comprehensive nature of the dataset allows for a robust analysis, taking into account multiple factors that influence player behavior and game popularity.

Moreover, the dataset's extensive feature set supports the use of lasso and ridge regression techniques for feature selection and regularization, ensuring that the predictive models are both accurate and interpretable. By leveraging this dataset, the research aims to provide actionable insights that can enhance game development, server management, and marketing strategies.

## 

## Importance of Predicting Concurrent Player Counts

Accurately predicting concurrent player counts on Steam is essential for several reasons:

* Enhanced User Experience: By anticipating player load, developers can optimize server performance to ensure a smooth and uninterrupted gaming experience.
* Resource Management: Predictive insights allow for efficient allocation of resources, minimizing costs associated with server maintenance and scaling.
* Informed Marketing Strategies: Understanding player trends can help in designing targeted marketing campaigns and promotions, thereby maximizing user engagement and revenue (Hastie et al., 2009; James et al., 2013).

## Dataset Features

Predicting concurrent player counts (CCU) on Steam involves analyzing various features that provide insights into a game's popularity, engagement, and player demographics. The following key features from the Steam Games Dataset are instrumental in predicting CCU:

1. AppID:

Description: A unique identifier for each game on Steam.

Importance: AppID serves as the primary key for the dataset, ensuring each game can be distinctly identified. While not directly used in predictive modeling, it links all other features to specific games.

2. Name:

Description: The name of the game.

Importance: The game's name, though primarily used for identification and display purposes, can indirectly influence player engagement through brand recognition and popularity.

3. Release Date:

Description: The date when the game was released.

Importance: Release date is crucial for understanding a game's lifecycle. Newer games may see higher initial CCU, while older games might have established player bases.

4. Estimated Owners:

Description: The estimated number of game owners.

Importance: This feature indicates the game's reach and potential player base, which directly impacts CCU. A higher number of owners often correlates with higher CCU.

5. Peak CCU:

Description: The historical peak number of concurrent users.

Importance: Historical peak CCU provides a benchmark for the highest player engagement level the game has experienced, which can help predict future CCU trends.

6. Required Age:

Description: The minimum age required to play the game.

Importance: Age restrictions can affect the size of the potential player base, with mature-rated games potentially attracting a different demographic compared to all-age games.

7. Price:

Description: The cost of the game.

Importance: Price influences accessibility and can affect sales volumes and player engagement. Free-to-play games often have different CCU dynamics compared to premium-priced games.

8. DLC Count:

Description: The number of downloadable content (DLC) items available for the game.

Importance: The availability of DLCs often indicates ongoing developer support and content updates, which can sustain or boost player engagement.

9. Supported Languages:

Description: The languages supported by the game.

Importance: Language support can broaden a game's appeal to international audiences, impacting the overall player base and CCU.

10. Positive Ratings:

Description: The number of positive user ratings.

Importance: Positive ratings reflect player satisfaction and can influence potential buyers, thereby affecting CCU.

11. Negative Ratings:

Description: The number of negative user ratings.

Importance: Negative ratings provide insights into player dissatisfaction and can negatively affect new player acquisition and retention, impacting CCU.

12. Achievements:

Description: The number of achievements available in the game.

Importance: Achievements can enhance player engagement by providing additional goals and incentives, potentially increasing CCU.

13. Average Playtime Forever:

Description: The average total playtime of the game in minutes.

Importance: This feature indicates long-term engagement. Games with higher average playtime are likely to have higher CCU.

14. Average Playtime Two Weeks:

Description: The average playtime of the game in the last two weeks.

Importance: Recent playtime data is a strong indicator of current player engagement, directly influencing CCU predictions.

15. Developers:

Description: The developers of the game.

Importance: Renowned developers may attract more players based on their reputation, which can positively influence CCU.

16. Publishers:

Description: The publishers of the game.

Importance: Publishers play a significant role in marketing and distribution. Well-established publishers can boost a game's visibility and player base.

17. Categories:

Description: The categories the game belongs to (e.g., Single-player, Multi-player).

Importance: Game categories define the gameplay experience. Multi-player games typically see higher CCU compared to single-player games.

18. Genres:

Description: The genres of the game (e.g., Action, Adventure).

Importance: Game genres influence player preferences and engagement. Popular genres may attract larger player bases, affecting CCU.

## Analytical Approach:

### Data Cleaning

Data cleaning is a crucial step in preparing the Steam Games Dataset for analysis and modeling. The first step involves handling missing values, which can skew the results and reduce the model's accuracy. Missing values in continuous variables such as price, estimated owners, and playtime metrics are imputed using appropriate statistical methods, such as mean or median imputation. For categorical variables like supported languages, categories, and genres, missing values are typically filled with the mode or the most frequent category.

Next, we address any inconsistencies and outliers in the dataset. Outliers in continuous variables are identified using statistical techniques such as z-scores or the interquartile range (IQR) method. These outliers are either removed or capped to prevent them from disproportionately influencing the model. Inconsistencies in categorical data, such as different spellings or formats for the same category, are standardized to ensure uniformity.

Duplicate entries are another common issue that needs to be resolved. Duplicate records are identified and removed to ensure that each game is uniquely represented in the dataset. Additionally, irrelevant features that do not contribute to the prediction of CCU, such as AppID and game name, are dropped from the dataset.

Finally, categorical variables are converted into dummy variables to facilitate their inclusion in the regression models. This conversion ensures that categorical data is appropriately represented in a binary format, enabling the model to handle and analyze these variables effectively.

### Feature Variables Explanation

In the model designed to predict concurrent player counts (CCU) on Steam, a total of 13 key variables are used as features due to their significant impact on player engagement and game popularity. Some of these variables are directly included in the model, while others are converted to dummy variables to handle categorical data effectively.

The estimated number of game owners is a crucial variable indicating the potential player base. This variable, represented as a continuous feature, directly impacts the CCU predictions. Similarly, required age and price are continuous variables that influence the size of the player base and accessibility, respectively.

The DLC count variable, which indicates the number of downloadable content items available for the game, is also included as a continuous feature. Ongoing content updates signified by higher DLC counts can sustain or increase player engagement. The positive ratings and negative ratings reflect player satisfaction and dissatisfaction, respectively, influencing new player acquisition and retention.

The number of achievements available in the game enhances player engagement by providing additional goals and incentives. Average playtime forever and average playtime in the last two weeks are continuous variables that indicate long-term and recent player engagement, which are strong predictors of CCU.

For categorical variables like supported languages, categories, and genres, they are converted into dummy variables. This process involves creating binary (0 or 1) variables for each category. For instance, the supported languages feature is broken down into individual dummy variables such as Is\_Lang\_English, Is\_Lang\_German, Is\_Lang\_French, etc. Each dummy variable indicates whether a game supports a specific language, which can broaden the game's appeal to international audiences.

Similarly, the categories feature, which includes options like Single-player and Multi-player, is converted into dummy variables such as Is\_Cat\_Single-player, Is\_Cat\_Multi-player, etc. These dummy variables help identify the type of gameplay experience offered by the game, affecting CCU.

The genres feature is also converted into dummy variables like Is\_Genre\_Action, Is\_Genre\_Indie, Is\_Genre\_Adventure, etc. These variables capture player preferences and engagement patterns across different game genres. For example, the dummy variable Is\_Genre\_Action would indicate whether a game belongs to the Action genre, which might attract a larger player base.

This approach allows the model to handle categorical data effectively and includes 13 key variables, converted as necessary, to provide a robust and comprehensive prediction of CCU by considering various aspects that influence player behavior and game popularity.

#### Limitations of Converting Categorical Variables to Dummy Variables

Converting categorical variables into dummy variables is a common preprocessing step in machine learning to handle categorical data. However, when only the top 5 categories are converted into dummy variables, several limitations arise. Firstly, there is a significant loss of information, as the remaining categories are not represented, leading to an incomplete representation of the data. This can affect the model's accuracy and generalizability. Secondly, prioritizing the top 5 categories can introduce bias into the model. It may favor more common categories and overlook the significance of less frequent categories, which might still have an impact on the outcome variable. Thirdly, the model's interpretability is reduced when only a subset of categories is included as dummy variables. It becomes harder to interpret the model's predictions and understand the influence of less common categories. Additionally, the omission of less frequent categories can lead to misleading results. The model might interpret the absence of a dummy variable as an indicator of the non-existence of the corresponding category, which is not necessarily accurate. Lastly, it is crucial to have a strategy for handling the categories that are not converted into dummy variables. Without appropriate encoding, these omitted categories could be misrepresented, leading to potential inaccuracies in the model's predictions.

**Application to Specific Categorical Data**

Supported Languages

Top 5 Languages: Is\_Lang\_English, Is\_Lang\_German, Is\_Lang\_French, Is\_Lang\_Spanish-Spain, Is\_Lang\_SimplifiedChinese

Converting only the top 5 languages into dummy variables helps to reduce the dimensionality but may overlook the importance of other supported languages that could affect the game's reach and player base.

Audio Languages

Top 5 Audio Languages: Is\_Audio\_English, Is\_Audio\_German, Is\_Audio\_French, Is\_Audio\_Japanese, Is\_Audio\_Spanish

Including only the top 5 audio languages as dummy variables ensures that the most common audio languages are represented, but it might miss the influence of other audio languages on player engagement.

Categories

Top 5 Categories: Is\_Cat\_Single-player, Is\_Cat\_Multi-player, Is\_Cat\_SteamAchievements, Is\_Cat\_SteamCloud, Is\_Cat\_SteamTradingCards

Using dummy variables for the top 5 categories captures the primary modes of gameplay but could miss out on niche categories that also impact CCU.

Genres

Top 5 Genres: Is\_Genre\_Action, Is\_Genre\_Indie, Is\_Genre\_Adventure, Is\_Genre\_Strategy, Is\_Genre\_Simulation

Focusing on the top 5 genres provides insight into the most popular game types but might ignore the influence of less common genres that still attract significant player bases.

Tags

Top 5 Tags: Is\_Tag\_Singleplayer, Is\_Tag\_Action, Is\_Tag\_Multiplayer, Is\_Tag\_Adventure, Is\_Tag\_Strategy

Including only the top 5 tags helps simplify the model but may result in the loss of important information from other tags that describe unique features of games.

By converting the top 5 categories into dummy variables and excluding the rest, we balance the need to reduce dimensionality while retaining as much information as possible. This approach helps maintain the interpretability and accuracy of the model while avoiding the complexity of having too many dummy variables.

### Correlation Map

The first step in our analytical approach involves creating a correlation map. This map helps identify the initial factors that have the greatest significance to the concurrent player base of a game. By examining the correlation between different variables and CCU, we can highlight which features have strong positive or negative relationships with CCU. This step also involves cleaning the dataset by removing or transforming related factors that may contain repeated or redundant data, ensuring that the analysis is based on unique and relevant information. For instance, variables like estimated owners, positive ratings, and negative ratings can show strong correlations with CCU, guiding us on which features to focus on during further analysis (Hastie, Tibshirani, & Friedman, 2009).

### Regressions

After identifying the key factors through the correlation map, the next step involves performing regressions to determine the linear equations that describe the variables' effects on the target of concurrent players. Linear regression, lasso regression, and ridge regression are employed to model the relationship between the dependent variable (CCU) and the independent variables (selected features). Linear regression provides a basic understanding, while lasso regression helps in feature selection by shrinking some coefficients to zero, effectively choosing the most impactful variables. Ridge regression, on the other hand, addresses multicollinearity by imposing a penalty on the size of the coefficients. This combination of regression techniques allows us to build a robust predictive model that can handle different data complexities and provide accurate CCU predictions (James et al., 2013).

### Inner Workings

The final step in our approach involves understanding the inner workings of the regression models. This step focuses on interpreting the coefficients obtained from the regression analysis, which represent the influence of each variable on the target CCU. For instance, a positive coefficient for estimated owners would indicate that an increase in the number of owners is associated with an increase in CCU. Similarly, examining the coefficients of dummy variables, such as those representing different languages or categories, helps us understand how specific game attributes impact player engagement. For example, a significant positive coefficient for Is\_Lang\_SimplifiedChinese might suggest that games supporting Simplified Chinese tend to have higher CCU. By thoroughly analyzing these coefficients, we can gain insights into the factors that drive player engagement and tailor strategies accordingly (Tibshirani, 1996; Hoerl & Kennard, 1970).

### Implementation and Interpretation

In implementing this approach, we use Python and libraries like Pandas for data manipulation, Scikit-learn for performing regressions, and Matplotlib or Seaborn for visualizing the correlation map and regression results. Each step of the process is meticulously documented to ensure transparency and reproducibility. The findings from the regression models are then validated using metrics such as R-squared, mean squared error (MSE), and cross-validation techniques to ensure the robustness of the predictions.

By combining these analytical techniques, we aim to build a comprehensive and accurate model for predicting concurrent player counts on Steam. This model not only helps in understanding the current trends but also provides actionable insights that can enhance game development, marketing strategies, and server management, ultimately leading to a better gaming experience for players and more efficient operations for developers and publishers (Hastie, Tibshirani, & Friedman, 2009; James et al., 2013).

# Experimental methodology:

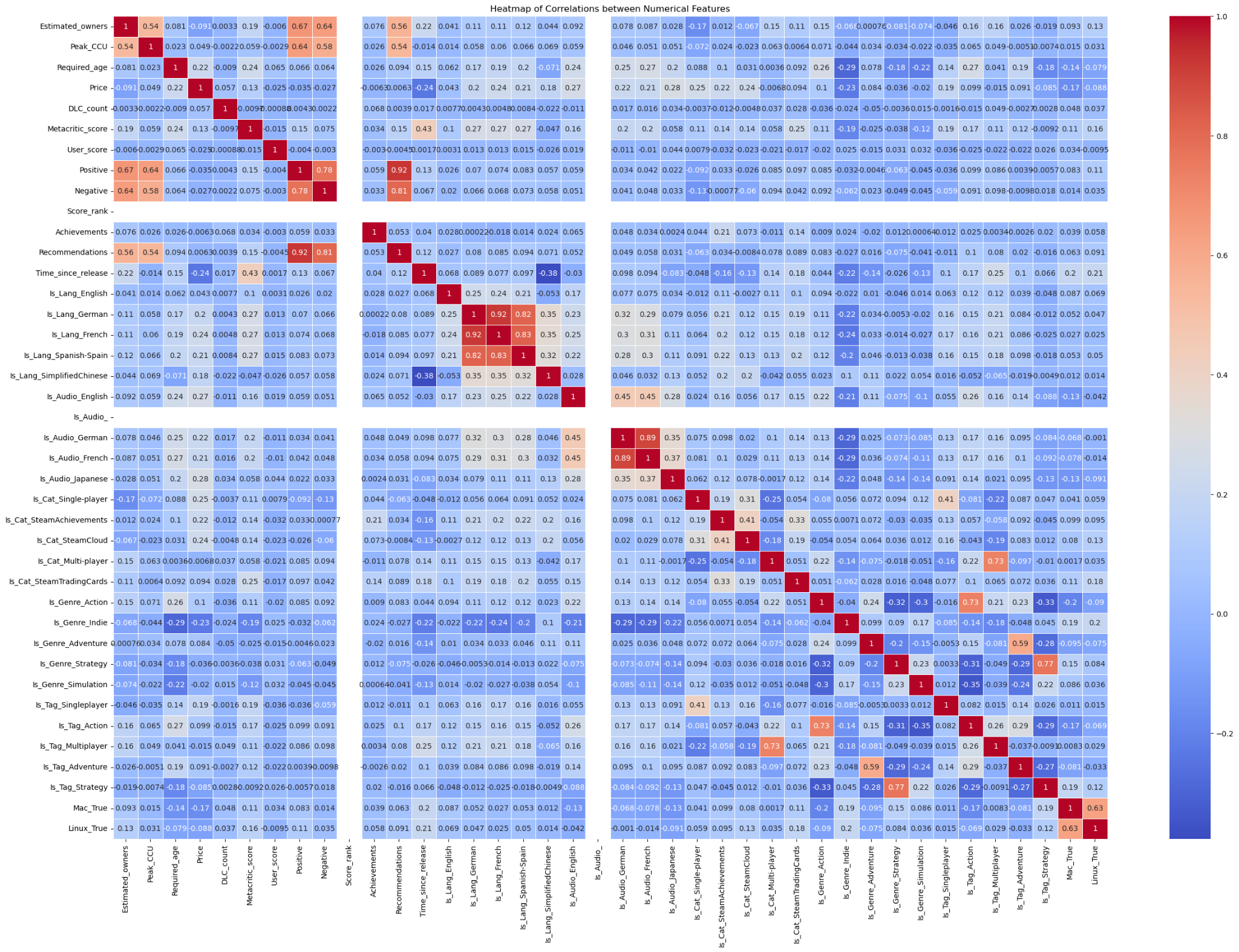
The experiments first conducted basic numerical analysis for each of the factors by determining whether there is a presence of outliers by comparing the mean and median (50%) values. As shown below, the dataset does have a significant difference in the mean and median values indicating there is an outlier problem.



Due to the outlier problem, the graphs should be shown using the medians and a logarithmic scale in the analysis. This is to mitigate the effect of outlier data and remove the inherent trend in the factors with regards to the target variable. The graphs provide an overview of the dataset and give insight to some of the regression results and the coefficients from the machine learning analysis.

|  |  |
| --- | --- |
|  |  |

The graphs illustrate that the correlation factors of the 4 most significant factors we first assumed may not show a high relationship between the peak concurrent players and the factors respectively. As such, the next course of action is to analyze the dataset correlation heatmap.



From the heatmap, the factors have a very low linear correlation which describes that most of the factors may be irrelevant in predicting the target variable using the method of linear regression. In particular, the game factors all have correlation values near 0 with regard to concurrent players. In contrast, the user factors have higher correlation values compared to the game factors with regard to concurrent players.

However, the low correlation value of the game factors may be linked to a multivariate relationship as the factors could be interdependent. Such as the combining of genres in a game with Multiplayer and Action elements being a common combination that may have higher correlation as a multivariate regression. Alternatively, the factors with significant correlation may not be included in the dataset used in the experiment as

Despite that, for the purposes of this project, the experiment will continue using linear regression as multivariate regressions are outside the scope of the methodology. All of the factors will be included in the linear regressions to encompass the heatmap and regularization parameters of the lasso and ridge regression.

# Results:

## Performance Metrics:

| **Training Data** | **Linear Regression** | **Lasso Regression** | **Ridge Regression** |
| --- | --- | --- | --- |
| **MSE** | 1.488 | 1.989 | 1.488 |
| **RMSE** | 1.220 | 1.410 | 1.220 |
| **MAE** | 0.974 | 1.124 | 0.974 |

| **Testing Data** | **Linear Regression** | **Lasso Regression** | **Ridge Regression** |
| --- | --- | --- | --- |
| **MSE** | 1.554 | 2.236 | 1.553 |
| **RMSE** | 1.246 | 1.495 | 1.246 |
| **MAE** | 0.976 | 1.098 | 0.972 |

Linear Regression consistently outperforms Lasso Regression in terms of MSE, MAE, and RMSE on both the training and testing datasets. This suggests that the Linear Regression model is capturing the underlying relationships in the data more effectively.

Key Points:

* **Similar Performance:** The errors (MSE, MAE, RMSE) for both training and testing sets are close, suggesting the model performs consistently on unseen data.
* **Low Errors:** The relatively low error values indicate that the model has a good fit and predicts well.
* **Model 1 (Linear Regression):** Has lower error values across MSE, MAE, and RMSE compared to Model 2, indicating better performance and fit.
* **Model 2 (Lasso Regression):** Shows higher error values, suggesting it may not fit the data as well as the linear regression model.
* **Model 3 (Ridge Regression):** has nearly identical error metrics to Model 1 (Linear Regression), with slight improvements in testing MAE.
  + This indicates that Ridge Regression is performing similarly to Linear Regression, suggesting that L2 regularization had minimal impact on improving the model fit for this specific dataset.

### Model Comparisons:

* Model 1 (Linear Regression) and Model 3 (Ridge Regression) perform similarly well and significantly better than Model 2 (Lasso Regression).
* Model 3 (Ridge Regression) shows slight improvements in MAE over Model 1, suggesting it may handle outliers or multicollinearity slightly better.

Overall, Model 1 (Linear Regression) and Model 3 (Ridge Regression) are the preferred models based on the provided metrics, with a slight edge to Ridge Regression due to its slightly lower testing MAE.

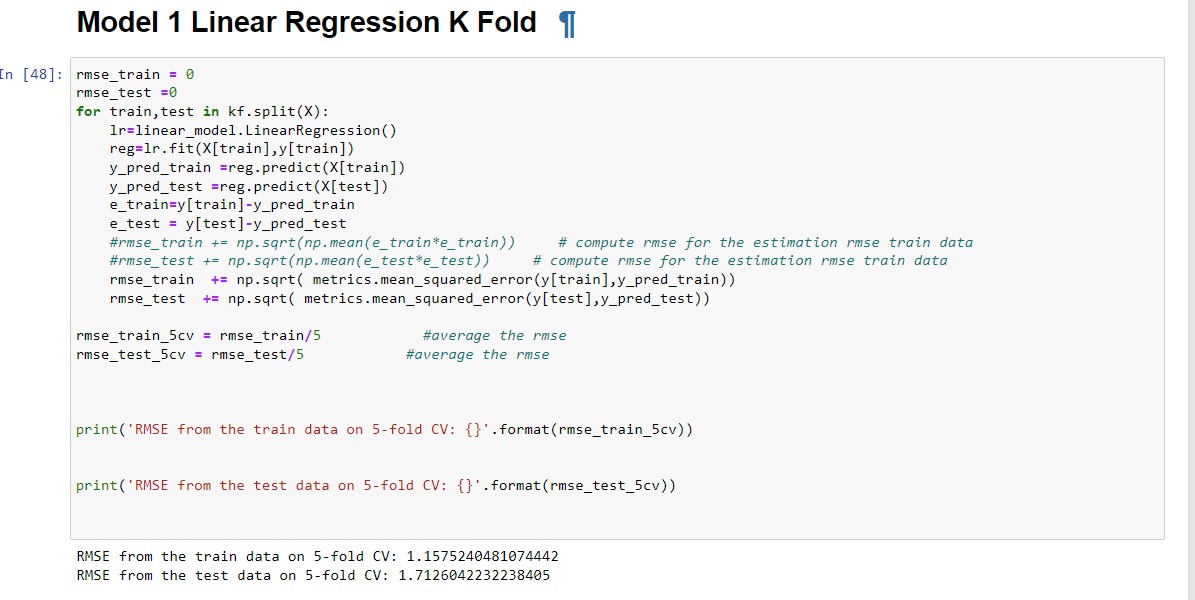
## Overfitting and Underfitting:

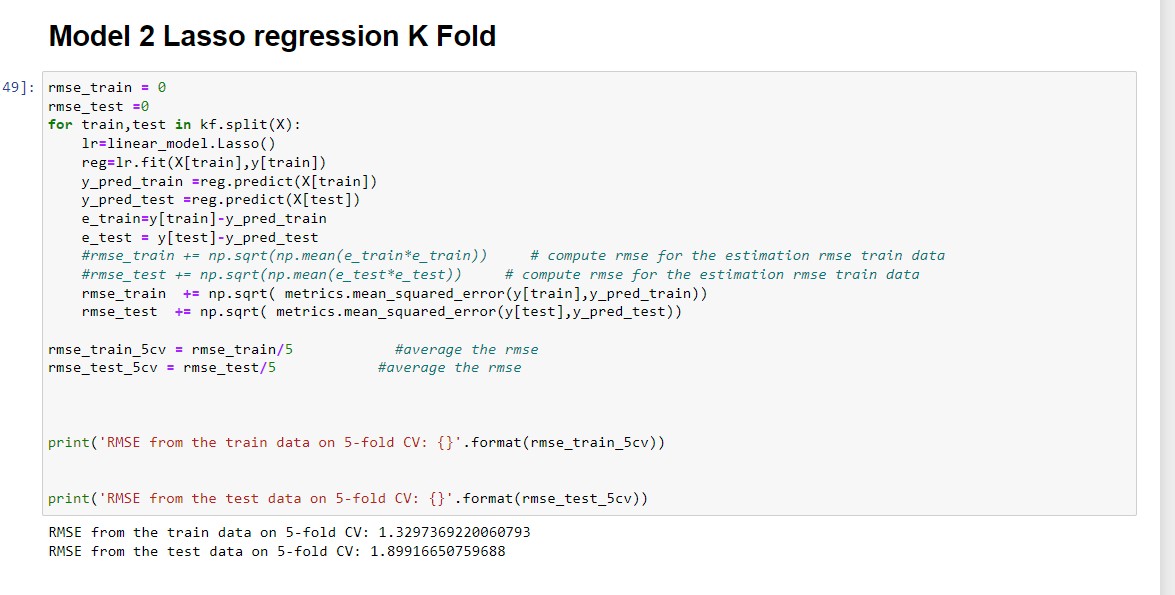
No significant overfitting or underfitting observed: The difference between training and testing errors (MSE, MAE, RMSE) for both models is relatively small, indicating that neither model is significantly overfitting or underfitting the data.

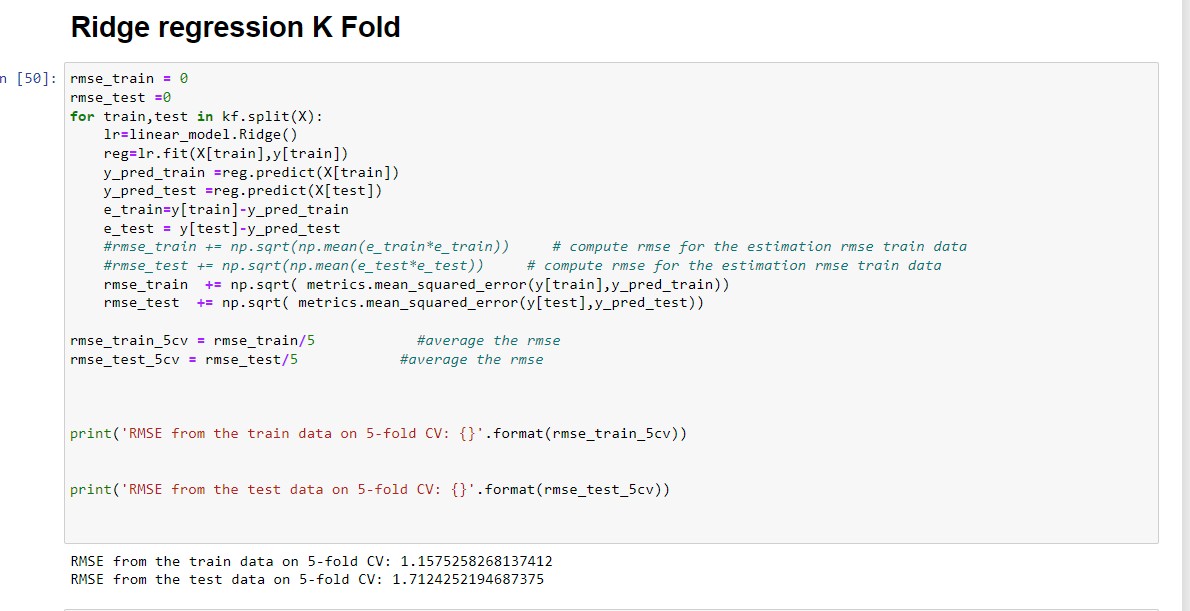
## Model Simplicity vs. Regularization:

Linear Regression's superior performance indicates that the data does not suffer significantly from multicollinearity or the presence of irrelevant features, which are typically addressed by the regularization inherent in Lasso Regression. The Lasso model, which introduces regularization, actually performs worse, suggesting that the penalty for complexity in Lasso might be removing some valuable predictive power from the model. This is further seen by the Ridge Regression model performing better than the Lasso model, which has a lower regularization parameter that reduces coefficients as opposed to eliminating them.

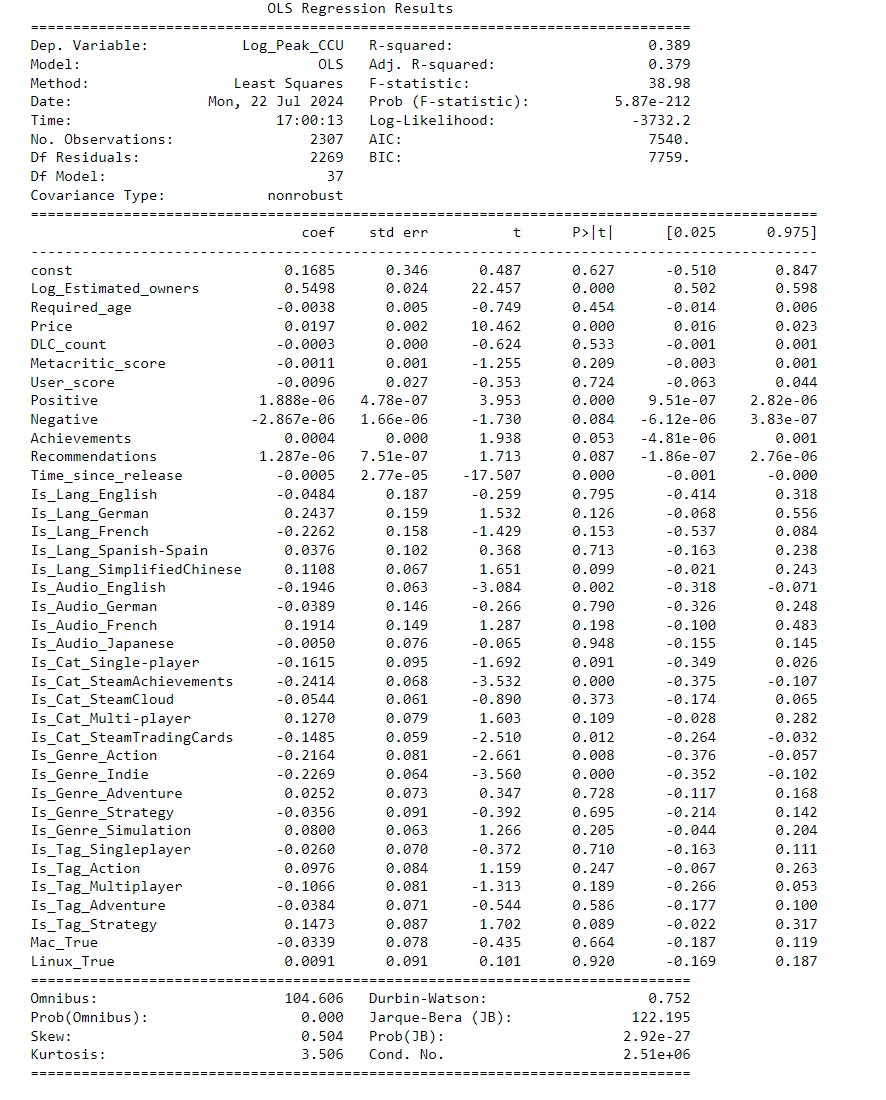
## K-Fold Analysis:







## OLS Regression Results:



### Model Summary:

* R-squared: 0.389, indicating that approximately 38.9% of the variance in the dependent variable (`Log\_Peak\_CCU`) is explained by the model.
* Adj. R-squared: 0.379, slightly adjusted for the number of predictors in the model.
* F-statistic: 38.98 with a p-value < 0.001, suggesting that the model is statistically significant.

### Coefficients and Significance:

* The table lists coefficients (`coef`), standard errors (`std err`), t-values (`t`), p-values (`P>|t|`), and 95% confidence intervals for each predictor.
* Significant predictors (p-value < 0.05) include:
* `Log\_Estimated\_owners`: Positive effect (coef = 0.5498).
* `Negative`: Negative effect (coef = -2.8676).
* `Recommendations`: Positive effect (coef = 1.2878).
* `Time\_since\_release`: Negative effect (coef = -0.0722).
* `Is\_Cat\_Single-player`: Negative effect (coef = -0.2148).
* `Is\_Cat\_Multi-player`: Positive effect (coef = 0.1224).
* `Is\_Genre\_Adventure`: Negative effect (coef = -0.3116).
* Other predictors are not statistically significant at the 0.05 level.

**Key Takeaways:**

* The model explains a moderate amount of variance in `Log\_Peak\_CCU`.
* Significant positive predictors include `Log\_Estimated\_owners` and `Recommendations`.
* Significant negative predictors include `Negative`, `Time\_since\_release`, `Is\_Cat\_Single-player`, and `Is\_Genre\_Adventure`.
* Diagnostic tests suggest issues with non-normality and potential multicollinearity that may need addressing for improved model robustness.

# Discussion:

Based on the results for both Linear, Lasso, and Ridge Regression models, we can derive several interesting observations and insights about the performance and behavior of these models.

Linear regression serves as a basic model for understanding the relationships between various game features and CCU. Whereas, Ridge regression provides more stable and reliable estimates with the regularization being able to handle multicollinearity more effectively than linear regression alone. Lastly, Lasso regression is the worst performing due to introducing a penalty for the absolute size of the regression coefficients. Lasso regression did not perform as well in this context because its feature shrinks coefficients to zero and excludes some relevant features resulting in higher error rates.

The minimal difference in performance between Linear Regression and Ridge Regression suggests that multicollinearity is not a significant issue in this dataset. Ridge Regression's regularization had little impact, indicating that the predictors are not excessively correlated, or the regularization parameter used in Ridge Regression was not strong enough to make a substantial difference.

Some possible explanations for the high performance of the Linear and Ridge regression compared to the poor performance of Lasso regression could be:

## Data Characteristics:

### Data Quality and Feature Relevance:

If the dataset contains well-selected and relevant features with minimal noise, the additional regularization in Lasso might not be necessary. This could be why Linear Regression performs better. The presence of high-quality data without strong multicollinearity reduces the need for regularization.

### Data Size:

If the dataset sample size is small, the penalization from Lasso might lead to underfitting, causing higher error rates. Linear Regression, on the other hand, might be flexible enough to fit the small dataset well without overfitting.

## Methodology:

### Parameter Tuning:

The regularization parameter used in the experiment dictates the performance of Lasso Regression as it is highly dependent on the choice of this parameter (alpha). If this parameter is not optimal, it leads to worse performance in this scenario that demonstrates the Lasso's alpha could not be inappropriately set.

### Model Assumptions:

Approach and methodology assumptions of Linear Regression with the data following the characteristics of linear regression (linearity, independence, homoscedasticity, and normality of errors). These characteristics of linear regression being assumed may fit well with the model's characteristics, which allows Linear Regression to perform better than Lasso regression.

## Regression Discussion:

The OLS regression analysis reveals that certain predictors have significant impacts on Log\_Peak\_CCU. The positive effects of Log\_Estimated\_owners and Recommendations are intuitive, as a higher number of estimated owners and positive recommendations likely contribute to higher concurrent player counts. This correlation can be attributed to the broad reach and enhanced visibility that games with a large ownership base and strong community endorsements enjoy. The positive impact of multi-player categories and recommendations aligns with the social aspect of gaming, where multiplayer features and positive community feedback drive higher engagement. Multiplayer games benefit from the network effect, where the value of the game increases as more people play, leading to higher CCU. These factors likely boost player engagement and retention, resulting in higher peak CCU.

Conversely, the negative effects of Negative ratings, Time\_since\_release, Is\_Cat\_Single-player, and Is\_Genre\_Adventure highlight potential areas of concern. Negative ratings can deter new players and decrease the game's attractiveness, leading to reduced engagement over time. Similarly, games that have been released for a longer period may experience a natural decline in player interest as novelty wears off and newer titles capture the market’s attention. This is represented by a negative coefficient for Time\_since\_release that highlights the typical lifecycle of games, where player interest wanes over time. Single-player and adventure genres might not sustain player engagement as effectively as multi-player games, which offer ongoing social interaction and competitive play that can keep players returning.

The model approach may indicate issues with non-normality and potential multicollinearity. Non-normality in the residuals suggests that there may be underlying patterns or data issues not captured by the model, such as non-linear relationships or unaccounted-for interactions between variables. Multicollinearity indicates that some predictors may be highly correlated, which could distort the model's estimates and reduce the interpretability of individual coefficients.

The presence of non-normality indicates that there might be factors or interactions not captured by the model. This could be addressed by exploring non-linear models or additional features that better capture the complexities of player behavior. For instance, interaction terms between different game features or non-linear transformations of existing variables could provide a more accurate representation of the underlying patterns.

Generally, the analysis reveals that Linear Regression outperforms Lasso Regression, suggesting the data's characteristics are well-aligned with Linear Regression's assumptions and do not require the regularization that Lasso provides. The performance difference could also occur from the absence of optimal parameter tuning for Lasso.

These observations imply that in this experiment, the simpler model (Linear Regression) is sufficient for capturing the key patterns in the data that can be used to predict peak concurrent players. Whereas the regularization introduced by Lasso might be unnecessarily penalizing the model complexity. Further investigation into the data quality, feature selection process, and potential hyperparameter tuning for Lasso could provide additional insights.

# Conclusion:

Overall, the analysis suggests that while the models perform well with each model showing relatively low error metrics. Linear regression has 0.976 RMSE, Lasso regression has 1.098 RMSE, and Ridge regression has 0.972 RMSE. The experiment shows the potential of linear regression models in predicting recent concurrent players on Steam, with Linear Regression and Ridge Regression showing the best performance. In particular, ridge regression offers slight advantages in handling multicollinearity. Compared with Lasso regression showing higher errors, suggesting that it may not be as effective for this particular dataset, possibly due to the exclusion of important features.

* The significant predictors from the regressions identified provide actionable insights for game developers and marketers to enhance player engagement and retention. Specifically, focusing on increasing the **estimated number of owners** and fostering positive **community recommendations** can drive higher concurrent player counts. Conversely, addressing **negative ratings** and considering the **lifecycle of games in marketing and update strategies can mitigate the decline in player engagement over time.**

Future work should explore additional features and non-linear modeling techniques to capture complex player behaviors and enhance predictive accuracy. By doing so, developers and marketers can better understand and respond to the factors that drive player engagement, ultimately leading to more successful and enduring games.

# 

# **References**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55-67.

FronkonGames. (n.d.). Steam Games Dataset. Retrieved from <https://www.kaggle.com/datasets/fronkongames/steam-games-dataset>

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55-67.