Machine Learning Final Report

Jack de Bruyn, Spencer Au, John Mulhern

Background/Introduction

The Dataset

Our dataset consists of data scraped from SteamSpy, related to the popular video game retailer, Steam. It contains data on variables such as peak concurrent players, review scores, and playtime of thousands of games sold on the platform. Steam is one of, if not the most popular distribution platforms for games on the market today. We hope that the wide range of this data will allow us to draw interesting and accurate conclusions for our questions listed below.

Link to Dataset

Our Questions

1. Does increasing the total average playtime lead to more positive reviews for a given game?
2. How does the availability on specific platforms (Windows, Mac, Linux) impact the estimated number of owners?
3. What is the direct causal effect of additional DLC on the average recent playtime (past 2 weeks)?
4. How do metacritic scores compare to the percentage of total positive reviews?

Methods of Analysis

Question #1

We used parameter estimation to run a Bayesian regression model with total number of positive reviews as our parameter to estimate and the total average playtime as our coefficient. We used priors that suggest some positive effect, as we believe that having more positive reviews should lead to a higher average playtime. We added Peak CCU as our covariate, as we believe this has a confounding effect on both total average playtime and the number of positive reviews.

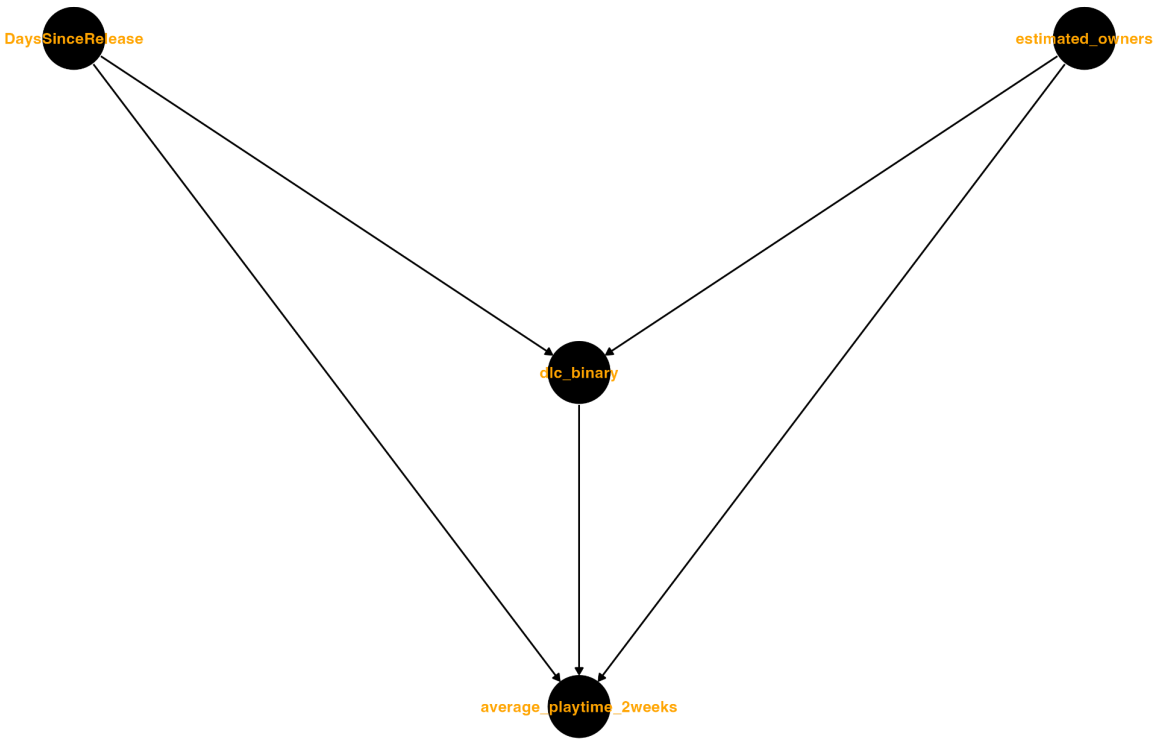
Question #2

We used both a Frequentist and Bayesian approach: for the Frequentist approach, we used a linear regression model with estimated_owners as the variable to predict and windows,

mac, and linux as predictors. For the Bayesian approach, we used a Bayesian linear regression model with estimated_owners as the variable to predict and windows, mac, and linux as predictors. Priors for the coefficients are weakly informative normal distributions, using a weak assumption that having more platforms will result in greater number of owners.

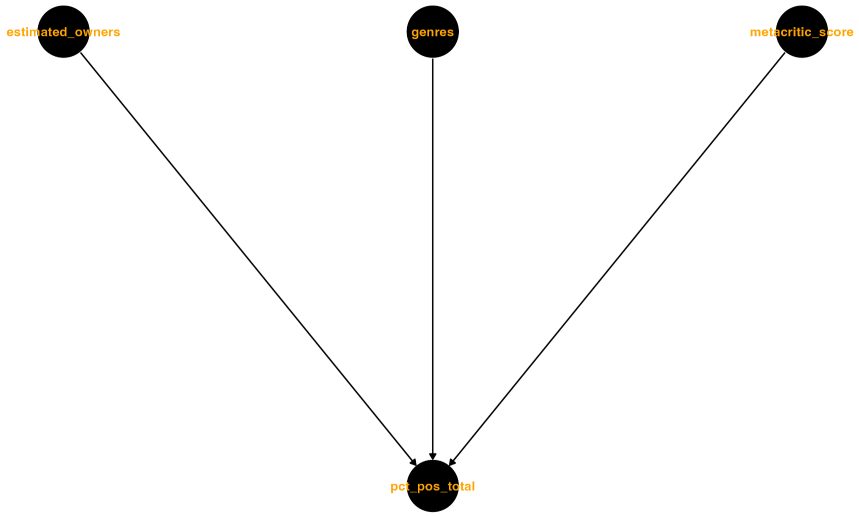
Question #3

To gauge the direct causal effect of the introduction of dlc on player retention, we leveraged linear regression after propensity score matching (PSM) to control for the confounding variables of DaysSinceRelease and estimated_owners. The dataset is filtered to include key variables such as playtime, DLC count, days since release, estimated owners, and the midpoint of ownership ranges is calculated for the estimated_owners column. We dropped rows with zero estimated owners. In addition, for PSM, we utilized a binary variable, dlc_binary, that indicates the presence of DLC (1 if DLC count > 0, 0 otherwise). PSM is applied to match games with and without DLC based on DaysSinceRelease and estimated_owners, using logistic regression to estimate propensity scores and a 3:1 nearest-neighbor matching ratio (in order to ensure better covariate balance between groups). Finally, we use a linear regression model to estimate the effect of DLC count on recent playtime within the last 2 weeks, controlling for dlc_count, DaysSinceRelease, and estimated_owners. This approach isolates the direct impact of DLC on recent playtime while minimizing bias. Below is the DAG we used to justify our covariates.

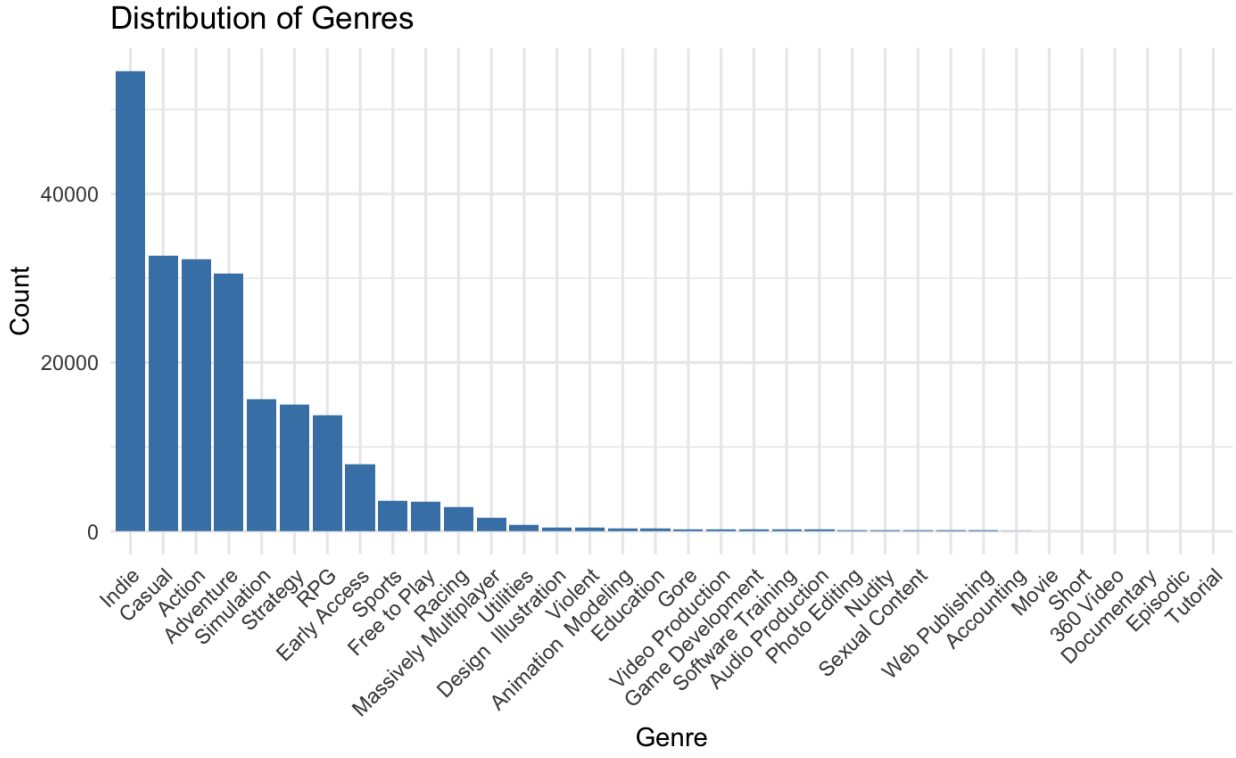


Question #4

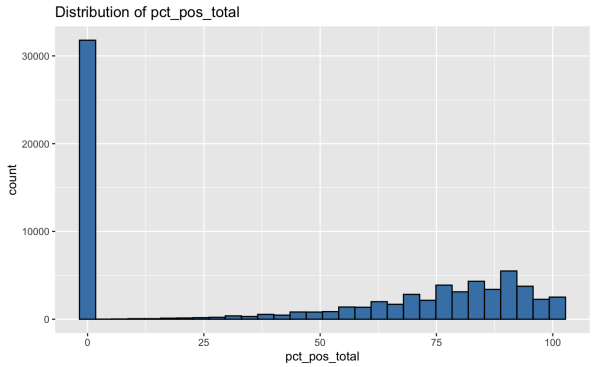
In order to determine if there is a relationship between metacritic scores and pct_pos_total, which is the percentage of positive reviews over total reviews on steam, we decided to utilize a Bayesian linear model, while controlling for estimated_owners and genres as covariates. Below is the DAG we used to justify the covariates.

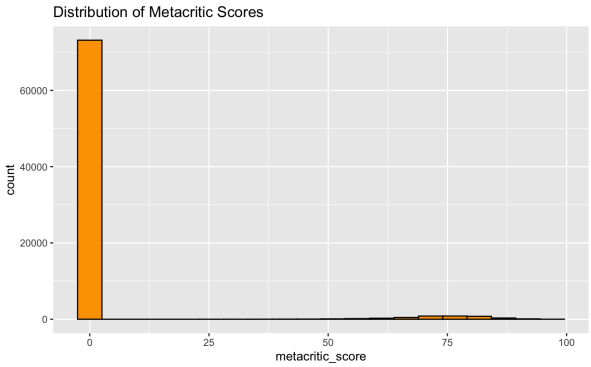


Because the genres column had a list like structure, where game 1 could be “Action, Adventure”, and game 2 could be “Action, etc”, we decided to take the most relevant and popular genre columns and reconstruct them as binary columns for each respective genre. From our EDA, this was what distribution of genres looked like initially:



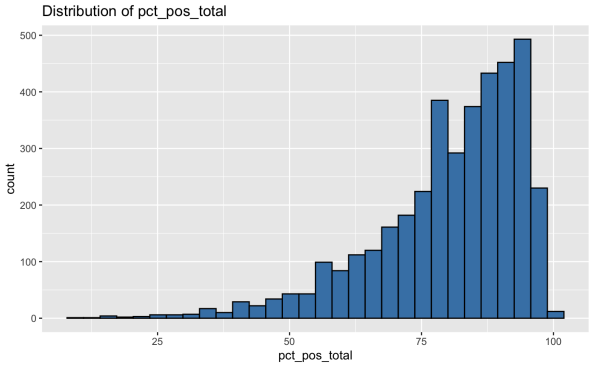
Massively_Multiplayer genres, and use each genre as a covariate. In addition, upon inspect of the score distribution, it seemed that the majority of the dataset had scores of 0s, as indicated by the EDA graphs below:

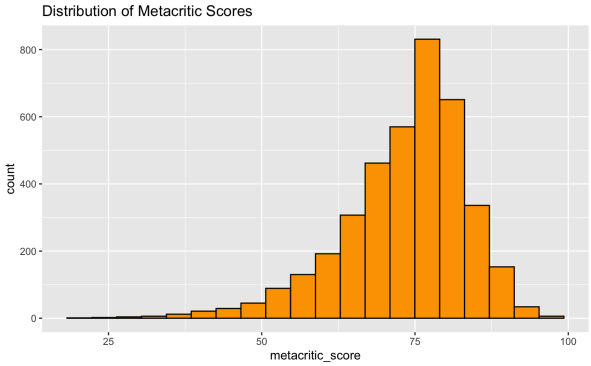




As a result, we decided to drop any row that had a zero in either category, resulting in a subset that included only some 3,500 games. The new distribution of games in the subset is indicated below, alongside relevant basic statistical metrics.

```
 pct_pos_total    metacritic_score
 Min.   :  9.00   Min.   :20.00
 1st Qu.: 73.00   1st Qu.:68.00
 Median : 84.00   Median :75.00
 Mean   : 80.25   Mean   :73.36
 3rd Qu.: 91.00   3rd Qu.:81.00
 Max.   :100.00   Max.   :97.00
[1] "Standard Deviation of pct_pos_total: 14.06"
```





The predictors metacritic_score and estimated_owners are scaled to standardize their values, improving model interpretability. A Bayesian regression model is then fitted using the brms package to analyze the relationship between pct_pos_total and the predictors [metacritic_score, estimated_owners, and genres (as a group of binary variables)]. The model uses a Gaussian family and includes priors that are informed but weakly regularizing, ensuring the model remains flexible. Notably, the intercept prior normal(80, 15) is based on the observed mean and standard deviation of pct_pos_total, reflecting the calculated mean and standard deviation. The model runs with 9,000 iterations, and a 3,000-iteration warmup phase to optimize sampling and convergence.

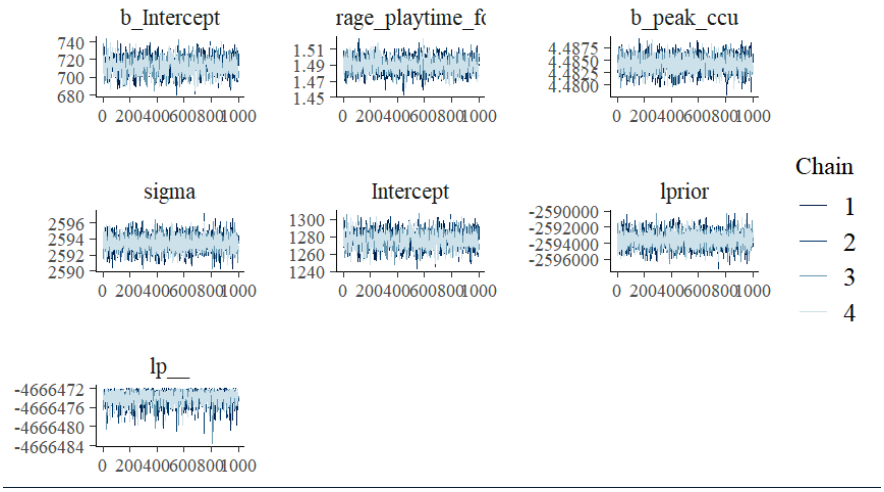
Results

Question #1

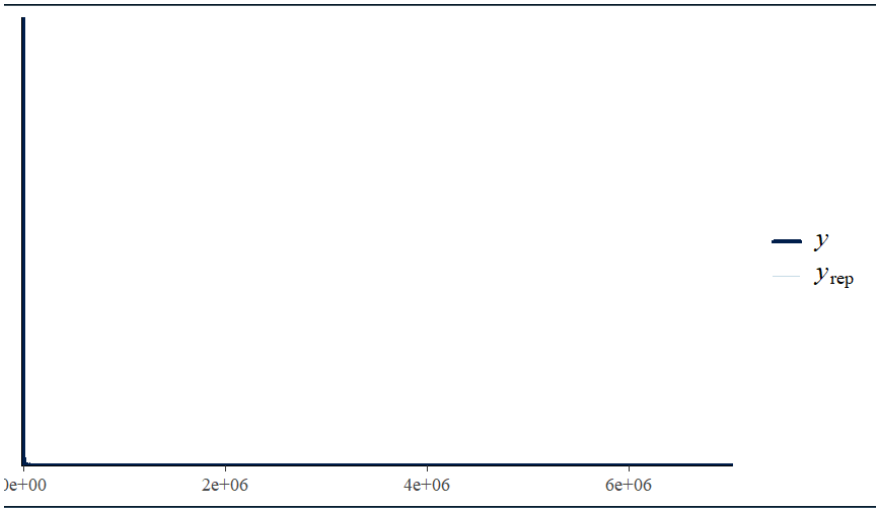
```
Regression Coefficients:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept               713.17      9.13   694.20   730.34 1.00     1452
average_playtime_forever  1.49      0.01     1.47     1.51 1.00     1846
peak_ccu                  4.48      0.00     4.48     4.49 1.00     3995
                      Tail_ESS
Intercept                 1628
average_playtime_forever  1814
peak_ccu                  2571

Further Distributional Parameters:
```

As we can see from the data above, average playtime has a very slight positive effect on the number of positive reviews. Our interval estimate would place this positive effect as being somewhere between 1.47 and 1.51.



These are the trace plots for the BRM. Each of these plots look like the requisite ‘fuzzy caterpillars’, implying that our MC simulations have converged.



This is the graph of our posterior check. It appears as though our expectations were in line with reality, however the graph itself seems somewhat unusual. This may warrant future investigation.

Question #2

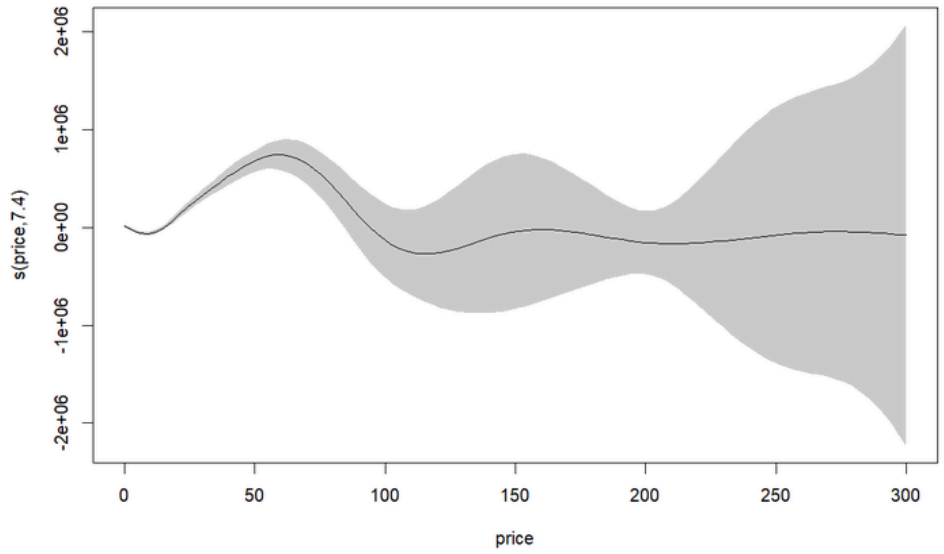
```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: estimated_owners_numeric ~ windows + mac + linux + price
   Data: steam_data (Number of observations: 83646)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
           Estimate Est.Error  l-95% CI  u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept   44888.36  50227.59 -54055.87 142159.18 1.00     4497     2495
windowsTrue 13915.07  49508.16 -80261.72 110913.65 1.00     4513     2551
macTrue     64799.98  16916.40  31639.63  98313.21 1.00     3332     2742
linuxTrue   88193.20  18763.43  51273.40 124763.82 1.00     3265     3216
price        3502.45    451.47   2627.17   4393.12 1.00     4326     2960

Further Distributional Parameters:
      Estimate Est.Error   l-95% CI   u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 1734233.58   4207.74 1726114.84 1742526.54 1.00     7802     2636

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
```

Observing our data, the credible interval for Mac and Linux operating systems indicates an increase in estimated player counts while the credible interval for Windows is incredibly sporadic. This is to be expected as Windows has nearly 100% coverage in terms of the games offered in the dataset. With reference to the dataset, there are 83,646 listed as being available on steam and of those, 83,616 of them are available on Windows. The near total availability on that platform would make it not statistically significant in terms of answering the question. In addition, when analyzing price as a confounding variable, there is also a positive effect in terms of increasing player counts as price increases. Referencing the graph below, the confidence intervals are easily measurable until price escapes roughly one-hundred dollars. The price range for video games on Steam ranges from zero dollars to ninehundred and ninety-nine dollars.



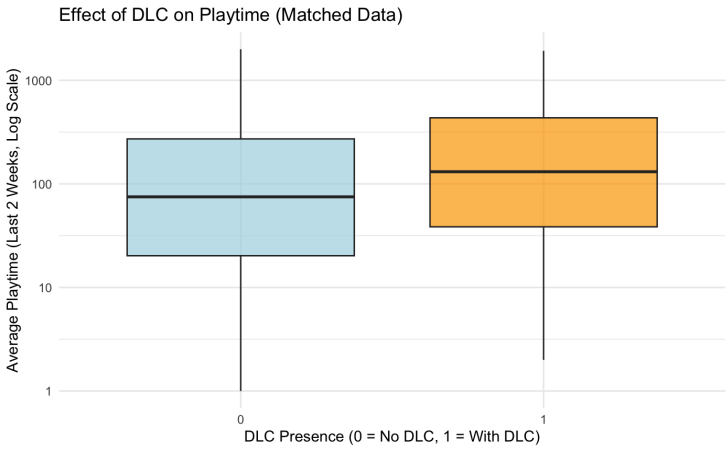
Question #3

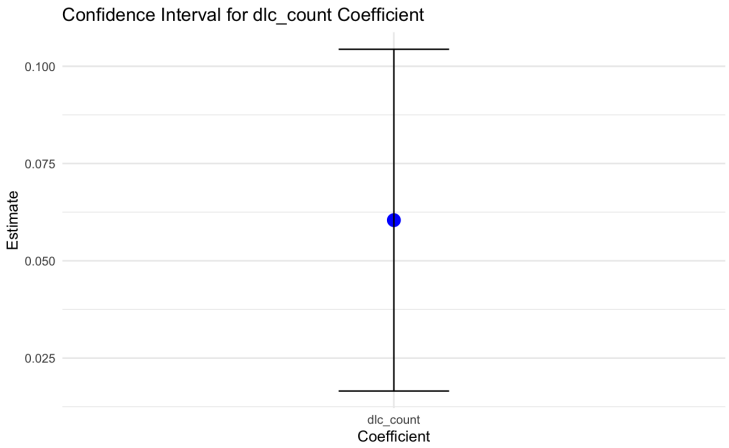
```
Residuals:
    Min      1Q  Median      3Q     Max
-1564.9    -4.5    -4.3    -4.3  6803.1

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.273e+00  8.029e-01   5.322 1.03e-07 ***
dlc_count         6.045e-02  2.241e-02   2.698  0.00697 **
DaysSinceRelease -4.514e-05  3.526e-04  -0.128  0.89812
estimated_owners  8.251e-06  1.832e-07  45.025  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.69 on 50576 degrees of freedom
Multiple R-squared:  0.0389,    Adjusted R-squared:  0.03884
F-statistic: 682.4 on 3 and 50576 DF,  p-value: < 2.2e-16
```

The analysis indicates that the effect of dlc_count on average_playtime_2weeks is statistically significant (p = 0.007), with an estimate of 0.060, suggesting that each additional DLC is associated with an expected increase of **0.060 minutes** (3.6 seconds) in the average playtime over the last two weeks, holding other variables constant. The intercept estimate of 4.273 (p < 0.001) represents the predicted playtime for games with no DLCs, assuming all other predictors are at baseline. Other predictors, such as DaysSinceRelease, were not significant (p = 0.898), while estimated_owners was highly significant (p < 2e-16), indicating its strong contribution to the model.





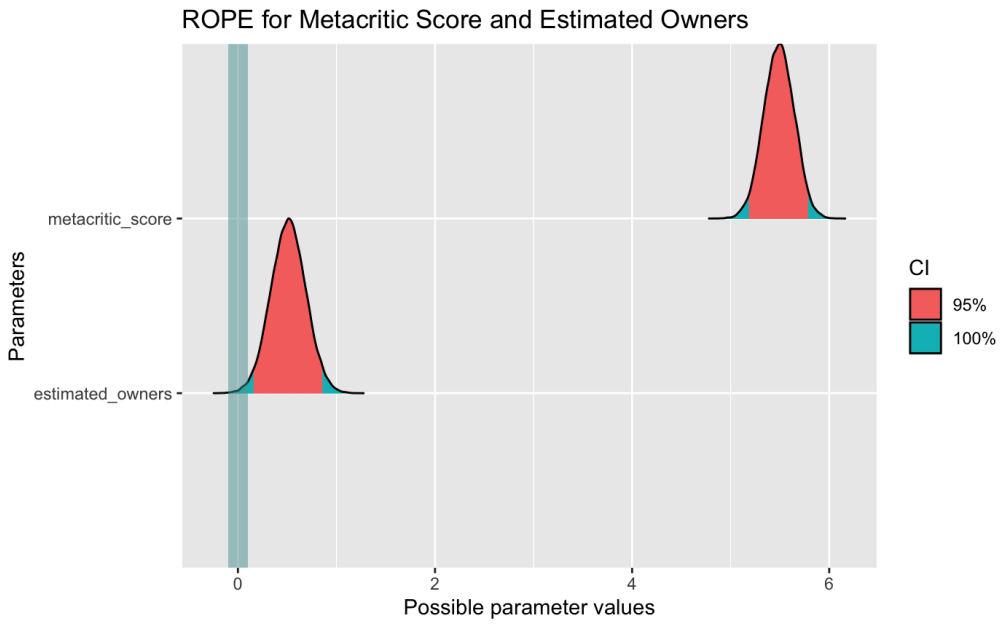
Overall, while it does seem like there is a positive correlation between additional DLC and recent playtime, the effect is fairly minimal, especially when we consider the fact that the scale of playtime is in minutes. In addition, while the 95% confidence interval for dlc_count does not include 0, indicating that there is an effect of dlc_count on playtime, the effect is simply too meager to be of any real consequence.

Question #4

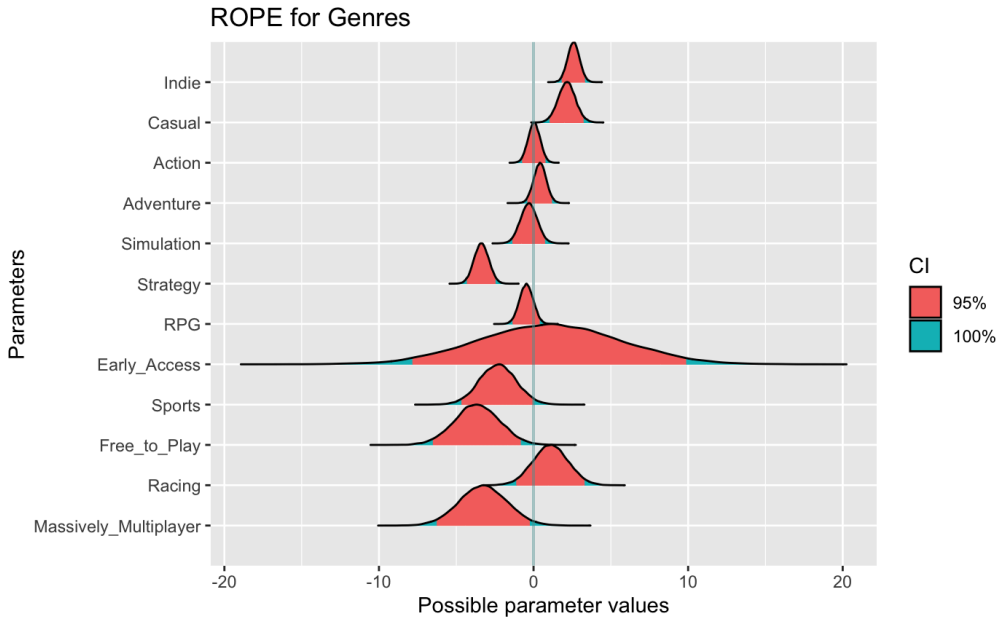
```
Regression Coefficients:
                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept            79.49      0.46    78.58    80.40 1.00    43106    20148
metacritic_score      5.49      0.15     5.19     5.79 1.00    44575    20212
estimated_owners      0.51      0.18     0.16     0.86 1.00    52385    19031
Indie                 2.59      0.38     1.84     3.34 1.00    46491    17406
Casual                2.15      0.56     1.06     3.26 1.00    50979    19335
Action                0.04      0.39    -0.72     0.79 1.00    41537    19193
Adventure             0.42      0.41    -0.38     1.22 1.00    39449    19931
Simulation           -0.30      0.54    -1.35     0.77 1.00    44876    19067
Strategy             -3.37      0.47    -4.30    -2.44 1.00    39121    20314
RPG                  -0.46      0.45    -1.34     0.42 1.00    48955    19169
Early_Access          1.02      4.56    -7.82     9.86 1.00    52169    17716
Sports               -2.31      1.18    -4.64    -0.01 1.00    43053    19373
Free_to_Play         -3.66      1.45    -6.51    -0.80 1.00    35313    21150
Racing                1.10      1.13    -1.10     3.33 1.00    45459    20192
Massively_Multiplayer -3.25     1.54    -6.25    -0.24 1.00    35098    21100

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    11.37      0.14    11.11    11.64 1.00    46766    18650
```

The Bayesian regression results reveal that the Intercept estimate is 79.49, indicating that when all predictors are at their baseline, the average percentage of positive reviews is approximately 79.5%. For the metacritic_score predictor, there is a strong positive association with pct_pos_total, with an estimate of 5.49. This suggests that a one-standard-deviation increase in metacritic_score is associated with an 5.49-point increase in positive reviews, highlighting its significant influence on review outcome.



This graph demonstrates the posterior distributions for the metacritic_score and estimated_owners predictors in relation to the percentage of positive reviews. The 95% credible intervals (CIs) for these predictors do not overlap with the Region of Practical Equivalence (ROPE) around zero, indicating a significant positive association. Specifically, the distribution for metacritic_score suggests that higher scores have a substantial positive impact on review positivity.



This shows the posterior distributions for the genre-related predictors. Several genres (Sports, Free_to_Play, Massively_Multiplayer) exhibit distributions that extend beyond the ROPE, suggesting negative associations with review positivity. Conversely, for genres such as Indie and Casual games, they have distributions that extend positively beyond the ROPE. Interestingly enough, Early Access games seem to have a very wide distribution for its CI, indicating that a game in early access does not have a clear directional effect on review positivity.

Discussion

Question #1

Based on our data, we can determine that increasing the total average playtime will in turn increase the total number of positive reviews for a game on Steam. This could be very useful information for video game development companies hoping to launch their products on Steam, as a positive review score serves as good branding and can help drive profits. A company might use this information to prioritize increasing the amount of time their game can be played for.

For next steps, I would like to potentially rerun this experiment with different updated priors, because I believe the priors used for this experiment were not as accurate as they could be. Additionally, I would also like to try searching for other potentially confounding variables within this dataset or others.

Question #2

The data gained from our linear regression models provided clear, direct answers to our initial question. For example, regression models showed that offering a game on multiple platforms had a statistically significant impact on player counts for the Mac and Linux operating systems, but not for Windows, despite the fact that the Windows operating system accounts for 99.96% of all games available in the dataset. Additionally, when looking at confounding variables and their relationship to estimated player counts, an interesting discovery was the relationship between price and estimated player counts. Looking in the graph below, for games priced between $25 and $60, there is a clear increase in player counts until a peak is reached. Then between roughly $60 and $100, the estimated player counts decline until the confidence intervals begin to distort wildly as prices increase into the hundreds. Overall, it is clear that offering a game on an increased number of operating systems has a clear effect on estimated player counts, but it is clearly not the only factor that affects this value. Price, developers, publishers, and user scores are all examples of other variables that affect estimated player counts, but they aren't relevant to answering the specific outlined question.

Question #3

The analysis reveals a positive but minimal correlation between additional DLC and recent playtime, with each additional DLC corresponding to only 3.6 seconds of increased playtime. This suggests that the impact of DLC on playtime may not justify the significant resources and development effort required for its creation. However, it's important to acknowledge the shift in the gaming industry towards both the Software as a Service (SaaS) and live-service models, which often replace the more traditional DLC practices. These evolving trends are challenging to

capture within the dataset, highlighting a limitation of the analysis and the need for more comprehensive data to assess the influence of modern “new” content.

Question #4

The clear positive correlation between metacritic scores and the percentage of positive user reviews reinforces the idea that professional critic scores can significantly influence user sentiment. However, it’s important to consider potential nuances, such as whether highly rated games are inherently better or if high scores drive positive perception. These findings highlight the value of Metacritic scores for developers and publishers as a metric that aligns with user feedback, while also underscoring the need to understand other contributing factors to user reviews, such as gameplay experience, community support, and game genres.