Analyzing Gaming Trends: Insights into Game Quality, User Experience, and Pricing Overtime

Mark Sverdlov - 323454710 Or Yacobovich - 325166940

September 2024

Abstract

This research explores the relationships between pricing, user experience, game quality in video games on the Steam platform, using a dataset from Kaggle. By utilizing different statistical tools, the article procures results that are able to offer insight about different facets of the gaming experience, including the grading models of professional critics, the history of commercialized gaming in recent years, and biases across different game genres. Finally, the article uses these many insights to develop a linear model that outperforms other naive linear models.

1 Introduction

Steam is a video game digital distribution service and storefront managed by Valve. Due to its enormous size and unique business model, it has access to rich data both about video games and user experience. A user of a game may recommend or derecommend the game. Some of the games in Steam have a Metacritic score which is the grade that game critics gave them. We aim to use the rich data to learn general phenomena about the connection between user experience and other features. In this research we explored the relationships between pricing, user experience, game quality and actual use. We analyzed the data set about Steam that was in Kaggle and we got some interesting results. In addition, we built a linear model that predict the Metacritic score.

2 Results

2.1 Correlations Between Key Statistics

We calculated the correlations between the statistics 'metacritic_score', 'user_ratio', 'age', 'median_playtime', and 'price' and arrived at the following results:

Table 1: Correlations Table

	metacritic_score	user_ratio	median_playtime	age	price
metacritic_score	1.00, p=0.0%	0.60, p=0.0%	0.07, p=0.0%	-0.10, p=0.0%	0.19, p=0.0%
user_ratio	0.60, p=0.0%	1.00, p=0.0%	0.05, p=0.2%	-0.14, p=0.0%	0.11, p=0.0%
median_playtime	0.07, p=0.0%	0.05, p=0.2%	1.00, p=0.0%	-0.04, p=0.6%	0.07, p=0.0%
age	-0.10, p=0.0%	-0.14, p=0.0%	-0.04, p=0.6%	1.00, p=0.0%	-0.39, p=0.0%
price	0.19, p=0.0%	0.11, p=0.0%	0.07, p=0.0%	-0.39, p=0.0%	1.00, p=0.0%

Blue colors suggest positive correlations, red color suggest negative correlations, and stronger colors suggest stronger correlations (in absolute value). The Dunn-Sidàk correction obtained our

chosen threshold for statistical significance applied on the threshold $p = 0.05$, as we hadn't had prior hypotheses about correlations.

Thus, our main findings are:

1. There is a strong correlation between 'metacritic_score' and 'user_ratio' statistics.
2. There is a strong negative correlation between 'age' and 'price'.
3. There is some correlation between 'price' and both 'metacritic_score' and 'user_ratio'.
4. There is some negative correlation between 'age' and both 'metacritic_score' and 'user_ratio'.
5. 'median_playtime' is weakly correlated with the other positive indicators: 'metacritic_score', 'user_ratio', and 'price'. We aren't however able to conclude anything statistically significant about the relationship between 'age' and 'median_playtime'.

In addition to that, we calculated the correlations restricted for the subset of large games, defined as games with more than a million estimated owners, and arrived at the following results:

Table 2: Correlation Table Restricted On Large Games

	metacritic_score	user_ratio	median_playtime	age	price
metacritic_score	1.00, p=0.0%	0.43, p=0.0%	0.05, p=22.1%	-0.01, p=75.9%	0.04, p=27.6%
user_ratio	0.43, p=0.0%	1.00, p=0.0%	-0.02, p=67.1%	0.08, p=4.6%	-0.07, p=6.4%
median_playtime	0.05, p=22.1%	-0.02, p=67.1%	1.00, p=0.0%	-0.12, p=0.2%	0.00, p=91.4%
age	-0.01, p=75.9%	0.08, p=4.6%	-0.12, p=0.2%	1.00, p=0.0%	-0.43, p=0.0%
price	0.04, p=27.6%	-0.07, p=6.4%	0.00, p=91.4%	-0.43, p=0.0%	1.00, p=0.0%

These results are much sharper than the general results in table 1 and also have some telling differences.

1. There is less correlation between the 'metacritic_score' and the 'user_ratio' (This is shown to be significant by Fisher transformation method).
2. The 'metacritic_score' and the 'user_ratio' lose their correlation with 'price', as well as the negative correlation with 'age'.
3. The 'median_playtime' loses all its weak correlations with the positive indicators 'metacritic_score', 'user_ratio', and 'price', but gains a negative correlation with 'age'.

The comparison between the two correlation tables leads us to interesting scientific conclusions which are further discussed in the 'discussion' section.

2.2 'metacritic_score' Distribution

We plot the distribution of the 'metacritic_score' in figure 1. We first checked whether the distribution is normal using the Kolmogorov-Smirnov test, and arrived at the statistically significant conclusion that its source distribution isn't normal (p=1.314e-23). We also explored the possibility its distribution is some variant of normal distribution, like log-normal, but we have also shown it's very unlikely (p=4.715e-47). Finally, we applied Fisher transformation to the 'metacritic_score' data - the rationale being that like a correlation score, it is also a bounded distribution whose mean is far from the middle point of the bounds. Regardless, even after Fisher transformation the data (depicted in figure 2) still fails to be normal (p=1.11e-6).

While the data is not normal, after inspecting it closely (see table 3) and noticing that 50% of the data is between 68 to 80, we conclude that it's morally correct to partition the total population of the games into three parts, per table 4. We further discuss the meaning of this partition in the 'discussion' section.

Figure 1: The Distribution of 'metacritic_score'

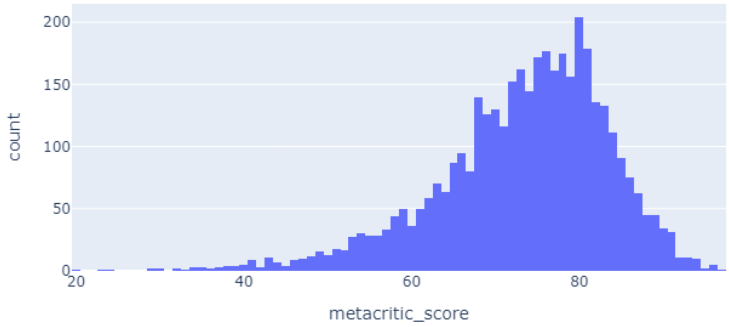


Figure 2: The Distribution of 'metacritic_score' After Applying Fisher Transformation

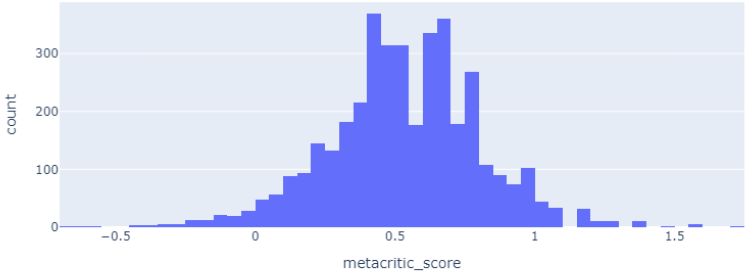


Table 3: Statistical Properties of The Distribution of 'metacritic˙score'

Statistic	Value
mean	73.27
min	20
25%	68
50%	75
75%	80
max	97

Table 4: Partition of The Population of Game Into 'Bad', 'Medium' and 'Good' Games according to 'metacritic_score'

Property	'Bad' Games (70-)	'Medium' Games (70-80)	'Good' Games (80+)
Percent of Total Games	30.58%	44.56%	24.86%
Mean Score	60.92	75.36	84.69
Minimum Score	20	70	81
25%	57	73	82
50%	63	75	84
75%	67	78	87
Maximum Score	69	80	97
Distribution	Concentrated near the maximum.	A uniform distribution along the 70-80 interval.	Concentrated near the minimum.

Figure 3: The Distribution of 'metacritic_score' For 'Bad' Games

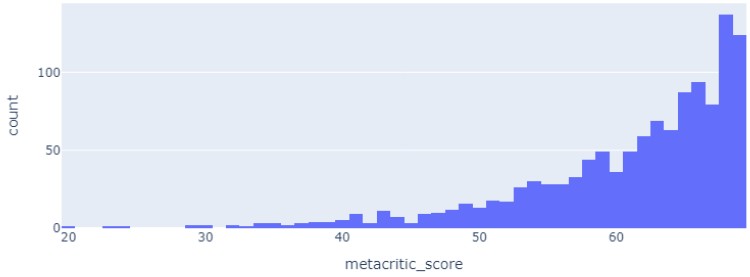


Figure 4: The Distribution of 'metacritic_score' For 'Medium' Games

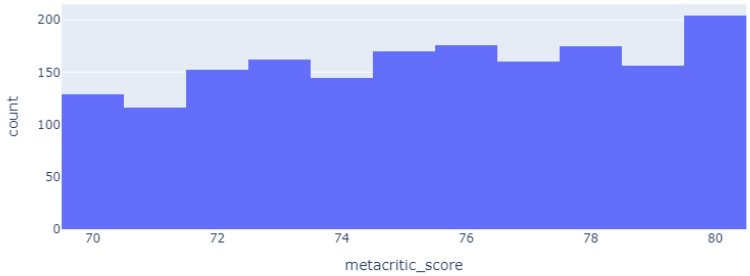
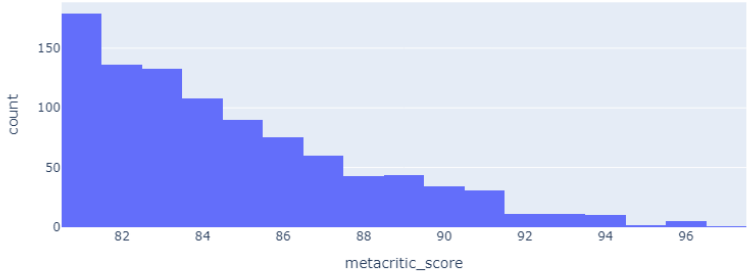


Figure 5: The Distribution of 'metacritic_score' For 'Good' Games



2.3 Dependencies of Statistics with Time

We investigated some questions about the dependency between different statistics. Our results in this regard are detailed here:

2.3.1 Dependency Between 'metacritic_score' and 'release_date'

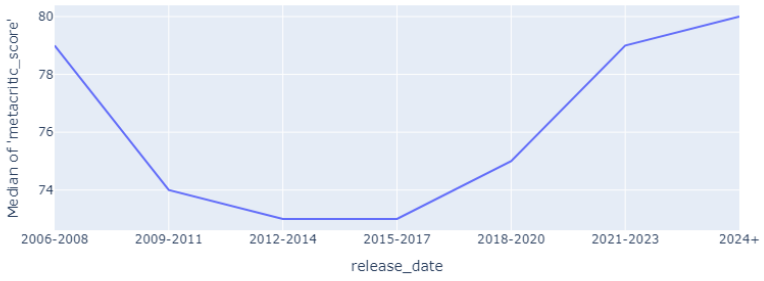
We used the Mann-Whitney U-test to find variances of the score given by critics to games over time. We believe there shouldn't be any such variance, i.e., the critics score games consistently overtime, however inspecting the results after using the Dunn-Sidàk correction to compensate for the large number of comparisons, we find that with very high statistical significance, there is some 'dent' in the score graph in the years 2009-2020 that amount to as much as five points difference.

We depict this phenomenon in table 5 and in figure 6 below:

Table 5: Comparison of 'metacritic_score' Overtime

	2006-2008	2009-2011	2012-2014	2015-2017	2018-2020	2021-2023	2024+
2006-2008	diff=0.00, p=50.0%	diff=5.00, p=100.0%	diff=6.00, p=100.0%	diff=6.00, p=100.0%	diff=4.00, p=100.0%	diff=0.00, p=51.9%	diff=-1.00, p=7.2%
2009-2011	diff=-5.00, p=0.0%	diff=0.00, p=50.0%	diff=1.00, p=99.8%	diff=1.00, p=98.9%	diff=-1.00, p=23.4%	diff=-5.00, p=0.0%	diff=-6.00, p=0.0%
2012-2014	diff=-6.00, p=0.0%	diff=-1.00, p=0.2%	diff=0.00, p=50.0%	diff=0.00, p=18.4%	diff=-2.00, p=0.0%	diff=-6.00, p=0.0%	diff=-7.00, p=0.0%
2015-2017	diff=-6.00, p=0.0%	diff=-1.00, p=1.1%	diff=0.00, p=81.6%	diff=0.00, p=50.0%	diff=-2.00, p=0.0%	diff=-6.00, p=0.0%	diff=-7.00, p=0.0%
2018-2020	diff=-4.00, p=0.0%	diff=1.00, p=76.6%	diff=2.00, p=100.0%	diff=2.00, p=100.0%	diff=0.00, p=50.0%	diff=-4.00, p=0.0%	diff=-5.00, p=0.0%
2021-2023	diff=0.00, p=48.1%	diff=5.00, p=100.0%	diff=6.00, p=100.0%	diff=6.00, p=100.0%	diff=4.00, p=100.0%	diff=0.00, p=50.0%	diff=-1.00, p=2.4%
2024+	diff=1.00, p=92.8%	diff=6.00, p=100.0%	diff=7.00, p=100.0%	diff=7.00, p=100.0%	diff=5.00, p=100.0%	diff=1.00, p=97.6%	diff=0.00, p=50.1%

Figure 6: Median of 'metacritic_score' Among Different 'release_date'



2.3.2 Dependency between and 'estimated_owners' and 'release_date'

We explored the relationship between the statistics 'estimated_owners' and 'release_date'. We used χ^2-test and we conclude that these statistics are dependent ($p = 0$).

We estimate the strength of this dependency using Cramer's coefficient formula and find out that:

$$r_c = 0.113$$

so that overall there is a weak link between the statistics.

Finally, we analyze the residues of the test and arrive at the results shown in table 6. We drop cells with too few observations. We color the cells blue if there are more games than expected in the cell (with statistical significance) and red if there are fewer games than expected.

Table 6: Amount of Games Among Different Release Dates and Popularity

	0-20,000	20,000-50,000	50,000-100,000	100,000-200,000	200,000-500,000	500,000-1,000,000
2006-2008	4, p=0.0%	18, p=15.7%	29, p=39.9%	26, p=64.6%	46, p=5.4%	33, p=1.2%
2009-2011	17, p=0.0%	47, p=15.7%	67, p=18.9%	79, p=9.4%	91, p=21.0%	58, p=20.6%
2012-2014	32, p=0.0%	74, p=1.5%	101, p=67.1%	134, p=2.9%	177, p=0.0%	91, p=40.1%
2015-2017	129, p=44.9%	187, p=0.0%	129, p=42.9%	149, p=44.4%	175, p=18.7%	101, p=9.8%
2018-2020	150, p=0.0%	119, p=89.6%	113, p=76.6%	120, p=24.9%	129, p=0.9%	100, p=99.7%
2021-2023	113, p=0.0%	71, p=44.8%	71, p=50.9%	77, p=23.8%	95, p=23.3%	56, p=21.4%

The results suggest that most of the changes over time occur in the smallest category (0-20,000 owners), namely, that in the beginning there were fewer small games, in the present there are many small games (as a percentage of the total amount of games).

2.4 Comparison between Generes

We used a one-sided Mann-Whitney U-test to compare various statistics among different genres. Since we compare 7 genres (total 21 comparisons) without prior hypotheses about the differences between genres, we employ the Dunn-Sidàk correction and we look for comparisons with p-values smaller than 0.244%.

In the results below, we color cells representing a statistically significant finding in red. We color cells representing less statically significant findings ($0.244\% \leq p < 5\%$ in light red.

2.4.1 Differences of 'metacritic_score' Between Genres

We applied the Mann-Whitney U test for 'metacritic_score'. Our results suggest that professional critics don't have any statistically significant bias between genres, except a slight bias against the simulation genre.

Table 7: Comparison of 'metacritic_score' Between Different Genres

	action	adventure	sport	indie	massively multiplayer	strategy	simulation
action	diff=0.0, p=50.0%	diff=0.0, p=70.9%	diff=1.0, p=77.7%	diff=1.0, p=98.2%	diff=-1.0, p=26.8%	diff=0.0, p=57.8%	diff=2.0, p=99.9%
adventure	diff=0.0, p=29.1%	diff=0.0, p=50.0%	diff=1.0, p=72.3%	diff=1.0, p=94.0%	diff=-1.0, p=20.8%	diff=0.0, p=40.2%	diff=2.0, p=99.9%
sport	diff=-1.0, p=22.3%	diff=-1.0, p=27.7%	diff=0.0, p=50.0%	diff=0.0, p=46.2%	diff=-2.0, p=15.4%	diff=-1.0, p=23.7%	diff=1.0, p=78.0%
indie	diff=-1.0, p=1.8%	diff=-1.0, p=6.0%	diff=0.0, p=53.8%	diff=0.0, p=50.0%	diff=-2.0, p=10.4%	diff=-1.0, p=5.8%	diff=1.0, p=97.9%
massively multiplayer	diff=1.0, p=73.2%	diff=1.0, p=79.2%	diff=2.0, p=84.7%	diff=2.0, p=89.6%	diff=0.0, p=50.1%	diff=1.0, p=75.7%	diff=3.0, p=97.5%
strategy	diff=0.0, p=42.2%	diff=0.0, p=59.8%	diff=1.0, p=76.3%	diff=1.0, p=94.2%	diff=-1.0, p=24.3%	diff=0.0, p=50.0%	diff=2.0, p=99.9%
simulation	diff=-2.0, p=0.1%	diff=-2.0, p=0.1%	diff=-1.0, p=22.0%	diff=-1.0, p=2.1%	diff=-3.0, p=2.5%	diff=-2.0, p=0.1%	diff=0.0, p=50.0%

2.4.2 Differences of 'median_playtime' Between Genres

We applied the Mann-Whitney U test for 'metacritic_score'. Our results suggest that 'indie' is a significantly shorter game genre than the others. If we allow a deduction from relatively large p-values, we may deduce that 'simulation', 'strategy', 'sport', and 'massively multiplayer' are long genres and 'action', 'adventure', and 'indie' are short genres.

Table 8: Comparison of 'median_playtime' Between Different Genres

	action	adventure	sport	indie	massively multiplayer	strategy	simulation
action	diff=0.0, p=50.0%	diff=37.0, p=98.3%	diff=-26.0, p=3.0%	diff=42.5, p=100.0%	diff=-66.0, p=2.7%	diff=-9.5, p=1.6%	diff=-65.0, p=0.0%
adventure	diff=-37.0, p=1.7%	diff=0.0, p=50.0%	diff=-63.0, p=0.5%	diff=5.5, p=90.2%	diff=-103.0, p=0.6%	diff=-46.5, p=0.0%	diff=-102.0, p=0.0%
sport	diff=26.0, p=97.0%	diff=63.0, p=99.5%	diff=0.0, p=50.0%	diff=68.5, p=99.9%	diff=-40.0, p=40.9%	diff=16.5, p=85.9%	diff=-39.0, p=55.0%
indie	diff=-42.5, p=0.0%	diff=-5.5, p=9.8%	diff=-68.5, p=0.1%	diff=0.0, p=50.0%	diff=-108.5, p=0.2%	diff=-52.0, p=0.0%	diff=-107.5, p=0.0%
massively multiplayer	diff=66.0, p=97.3%	diff=103.0, p=99.4%	diff=40.0, p=59.2%	diff=108.5, p=99.8%	diff=0.0, p=50.1%	diff=56.5, p=88.0%	diff=1.0, p=63.4%
strategy	diff=9.5, p=98.4%	diff=46.5, p=100.0%	diff=-16.5, p=14.1%	diff=52.0, p=100.0%	diff=-56.5, p=12.0%	diff=0.0, p=50.0%	diff=-55.5, p=2.8%
simulation	diff=65.0, p=100.0%	diff=102.0, p=100.0%	diff=39.0, p=45.0%	diff=107.5, p=100.0%	diff=-1.0, p=36.6%	diff=55.5, p=97.2%	diff=0.0, p=50.0%

2.4.3 Differences of 'user_ratio' Between Genres

We applied the Mann-Whitney U test for 'metacritic_score'. Our results suggest that, unlike professional critics, users have various biases towards different genres. For example, users give 'indie' games better scores than almost any other genre, while 'massively multiplayer' games get

worse scores from users.

Table 9: Comparison of 'user_ratio' Between Different Genres

	action	adventure	sport	indie	massively multiplayer	strategy	simulation
action	diff=0.00, p=50.0%	diff=-0.00, p=1.6%	diff=0.03, p=99.5%	diff=-0.00, p=1.6%	diff=0.07, p=100.0%	diff=0.03, p=100.0%	diff=0.01, p=99.0%
adventure	diff=0.00, p=98.4%	diff=0.00, p=50.0%	diff=0.03, p=100.0%	diff=0.00, p=54.6%	diff=0.07, p=100.0%	diff=0.03, p=100.0%	diff=0.02, p=100.0%
sport	diff=-0.03, p=0.5%	diff=-0.03, p=0.0%	diff=0.00, p=50.0%	diff=-0.03, p=0.1%	diff=0.04, p=98.5%	diff=0.00, p=41.3%	diff=-0.01, p=10.0%
indie	diff=0.00, p=98.4%	diff=-0.00, p=45.4%	diff=0.03, p=99.9%	diff=0.00, p=50.0%	diff=0.07, p=100.0%	diff=0.03, p=100.0%	diff=0.02, p=100.0%
massively multiplayer	diff=-0.07, p=0.0%	diff=-0.07, p=0.0%	diff=-0.04, p=1.5%	diff=-0.07, p=0.0%	diff=0.00, p=50.1%	diff=-0.04, p=0.6%	diff=-0.05, p=0.1%
strategy	diff=-0.03, p=0.0%	diff=-0.03, p=0.0%	diff=-0.00, p=58.7%	diff=-0.03, p=0.0%	diff=0.04, p=99.4%	diff=0.00, p=50.0%	diff=-0.02, p=3.2%
simulation	diff=-0.01, p=1.0%	diff=-0.02, p=0.0%	diff=0.01, p=90.0%	diff=-0.02, p=0.0%	diff=0.05, p=99.9%	diff=0.02, p=96.8%	diff=0.00, p=50.0%

2.5 Linear Models

We use linear regression to predict the 'metacritic_score' of a game by the other statistics. We utilize a naive model (predicting via 'user_ratio' and 'age') alone, a naive but slightly larger model (using 'user_ratio', 'age', as well as, 'median_playtime' and 'price'). Finally, we use the rest of our results to devise interaction terms for 'age' and 'smallness' (having fewer than a million estimated owners) and interaction terms for 'user_ratio' and the different genres. Our results are depicted in table 10.

Table 10: Different Linear Models That Predict 'metacritic_score'

Model	Statistics	r^2	Adjusted r^2
Naive	'user_ratio', 'age'	35.79%	35.76%
Larger Naive	'user_ratio', 'age', 'median_playtime', 'price'	37.61%	37.55 %
Genre Aware Model	'user_ratio', 'age'-'smallness', 'user_ratio'-genre for each genre	39.38%	39.24%

3 Methods

3.1 Pre-Processing

Before we analyzed the data, the dataset included over 80,000 games and 46 columns of information about every game in the Steam database[1]. But, in this research, we used only the games with a Metacritic score (where the 'metacritic_url' column in the data set is defined) and we got that there are 3940 games that have a Metacritic score.

[1]The original database can be found on Kaggle in https://www.kaggle.com/datasets/artermiloff/steam-games-dataset/data

In total, after pre-processing we have 16 statistics (columns in addition of AppID and name). We elaborate on the statistics on table 11 below.

Table 11: The Columns in our Data After Pre-Processing

Statistic	Definition	Comment
'user_ratio'	$\frac{\text{'positive'}}{\text{'positive'+'negative'}}$	Every user that owns a game may rate the game either as positive or negative.
'metacritic_score'	NA	The Metacritic score represents how well received the game is among professional game critics. It is a number between 0 to 100.
'age'	Number of days from the day that the game was published until the date 26.5.2024 (this is the last date that the data was updated).	
'action'	1- if the genre is action, 0-else.	
'adventure'	1- if the genre is adventure, 0-else.	
'sport'	1- if the genre is sport, 0-else.	
'indie'	1- if the genre is indie, 0-else.	
'massively_multiplayer'	1- if the genre is massively multiplayer, 0-else.	
'strategy'	1- if the genre is strategy, 0-else.	
'simulation'	1- if the genre is simulation, 0-else.	
'price'	NA	Every game have a price. Some of the games are free (0 dollars) and some of them cost a money (in dollars).
'release_date'	The categorical variable that represent the period that the game was released.	The periods are three years periods between 1997-2024.
'estimated_owner'	NA	The amount of the game downloads. (Defined by the data).
'total_reviews'	'positive' + 'negative')	Every user that owns a game may rate the game either as positive or negative.
'median_playtime'	NA	The time in minutes of the median user played in the game.
'peak_ccu'	NA	The number of players that played the game concourtly.

We used various statistical tests as described in the results section to obtain our results. The code generating the results is available on github[2].

3.2 Tools Used

1. We used Pearson correlation matrix to check the correlations between the statistics 'Metacritic_score', 'user_ratio', 'age', 'median_playtime', 'price'. We also used Fisher transfor-

[2]The code is available on https://github.com/MarkSverdlov/Statistical_Theory_Final_Assignment2024.git

mation method to verify the difference in correlations between large games and the general population. The Pearson correlations were calculated using scipy implementation.

2. We use Kolmogorov-Smirnov test in stats package to check if the variable 'metacritic_score' or its varients are normally distributed. The Kolmogorov-Smirnov test test was calculated using scipy implementation.

3. We used Mann–Whitney U test to differentiate between Metacritic score distribution over-time, and between Metacritic score, user ratio, and median playtime distributions for different genres.

4. We used χ^2-test to understand the dependency between release date and the estimated number of owners of games. We also utilized Cramer's coefficient and residues analysis toward this aim. We calculated the test manually using numpy and pandas features.

5. We used linear regression in order to build the linear models for predicting Metacritic score. The regression were calculated using scikit-learn implementation.

4 Discussion

4.1 Analysis of the Correlation Matrix

We are interested in understanding the correlations between 5 statistics: Metacritic score, user ratio, median playtime (which we believe indicates quality), age, and price. In the general population of games, we observe strong correlations between price and quality indicators, together with a strong negative correlation between age and quality indicators.

The picture changes, however, when we restrict our view only to large games (at least a million estimated owners). There, we lose both the price and age correlations with quality indicators.

We hypothesize that, for the general population, older games (originating in the 'explosion' of the popularity of gaming in 2009-2020, as discussed below) are both of worse quality and cheaper than games from recent years. This creates a correlation between price and quality indicators, that doesn't exist among the large games, whose quality is consistent over time.

Another intriguing phenomenon we observe is that among large games, older games are played for less time. We can think of different possible explanations: perhaps the popularity of gaming that was garnered after 2009 hadn't gone, but instead had channeled into more playtime for larger games? Perhaps in recent years, it's more trendy among large gaming corporations to create larger games that offer more playtime for the user? We don't have a cardinal hypothesis to suggest.

4.2 Understanding the Distribution of Metacritic Score

It is a commonly known fact that many 'natural' random and statistical variables are distributed via some variant of the normal distribution $N\left(\mu, \sigma^2\right)$. The fundamental reason for this is the central limit theorem - a proper normalization of a large amount of i.i.d but otherwise arbitrary random variables is distributed like the normal distribution.

The i.i.d variables symbolize some "story of becoming" of the final statistic common in many natural statistics. The fact that we have shown that the Metacritic score is not distributed normally suggests that this statistic's "story of becoming" is different.

We build upon our further exploration shown in the results section and hypothesize that the distribution of the Metacritic score clues us into a possible psychological model used by critics who score these games. We hypothesize that critics' score is done in two stages: First, they decide whether they consider the game 'bad', 'medium', or 'good'. If they believe the game is

'medium', they would score it between 70 to 80, among which the score is uniform. If, however, they believe the game is good or bad, they will score it slightly more than 80 or somewhat less than 70, giving more extreme scores only for truly remarkable games in this way or another.

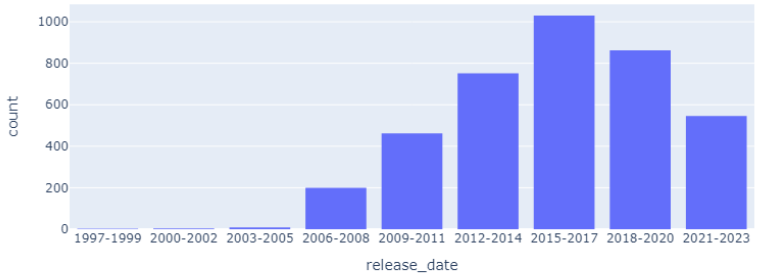
4.3 History of Popularity of Gaming and the 'Explosion' in Popularity of 2009-2020

We tried to use various results that explore the dependency between the release date and other statistics, to understand the history of the games. We first contemplate the number of games released on each date (see figure 7). It's interesting to see that the number of games started to rise in 2006, then peaked in 2015-2017, and finally dwindled in recent years. This behavior in the years 2009-2020 is mirrored in the Metacritic score graph (see figure 6). i.e., from 2009 to 2020, the number of games swelled, but their quality deteriorated.

We further analyzed the dependency between the release date and the size of the games, by considering the estimated number of owners. From our analysis, we concluded that the sizes of the games change over time, especially for smaller games. In the early years (before 2014), the super small games (0-20,000 estimated owners) were under-represented, while they consist of a large share of the games in recent years (after 2018). In the peak years of 2012-2014, medium games (100,000-500,000 estimated owners) are over-represented while in the peak years of 2015-2017, small-ish games (20,000-50,000 estimated owners) are over-represented.

We believe these facts combine into a compelling narrative: up to 2009, there weren't many games - the few games that exist garnered a large amount of users, and there weren't many tiny games. As gaming gained popularity, we see from 2009 to 2014 a swell in the number of games, where a large number of games were medium - not from large and established companies but still had a lot of users. This swell, however, was accompanied by a deterioration in the quality of the games. This deterioration caught up to many games in the 2015-2017 time frame, as there weren't as many medium games, while there were still many games in general. From 2018 onwards, the number of games began shrinking, many of them (up to 20,000 estimated owners). However, the quality of the remaining games rose again, slowly recovering the quality of the games before the 'explosion' of popularity in 2009.

Figure 7: Number of Games Released Overtime



4.4 Differences Between Different Genres

We have shown significant differences in the distributions of three key statistics: Metacritic score, median playtime, and user ratio among different genres. The reason we looked into these statistics is that all three of these statistics indicate the quality of the games in a different way.

By exploring and establishing these differences, we draw hypotheses and conclusions in two directions: first, we understand better the intrinsic differences between the genres. Second,

by comparing how these three different statistics change over the genres, we understand the statistics and the differences between them.

Our first observation is that in the Metacritic score, there is very little difference among different genres. The professional critics that are responsible for this statistic have taken care not to have a bias between the genres. This observation is in sharp contrast to the situation in the user ratio statistics. The common user hasn't tried to neutralize their bias for or against certain genres. Indeed, we can detect that among users, the 'massively multiplayer' and 'strategy' genres are considered worse, while the 'indie' genre consistently gets better user ratios than many other genres, which might suggest that many users 'root for the underdog games'.

Our last interesting observation is that by considering the median playtime, genres that are considered lesser by user ratio, are played more time. The contrast is especially sharp when one considers for example that 'indie' games are considered better, but are played for less time than most other genres, in particular, 'massively multiplayer' games (by almost two hours in median) or 'strategy' games (by almost an hour in median). This observation is consistent with our prior observations that median playtime, in general, doesn't correlate strongly with the other quality indicators, and we believe that this suggests that different genres are intrinsically shorter or longer (in playtime), regardless of their quality.

4.5 Analysis of The Genre Aware Linear Model

As shown in the results sections, we were able to get a better linear model than the naive baseline, using the interaction terms between the 'user_ratio' and the different genres, together with the interaction term between the 'age' and the 'smallness'.

The interaction terms between the 'user_ratio' and the different genres, as can be seen in table 12, represent the corrections done to the predicted Metacritic score when trying to predict it by the 'user_ratio'. For example, a genre like 'indie', which is over-rated by users, gets an interaction term of -2.6, which represents the correction that is done to predict the Metacritic score, while other genres are under-rated and thus get positive interaction terms. These results are consistent with our findings when comparing the genres.

The interaction term between the 'age' and the 'smallness', represents our previous observation that 'age' is an important indicator of Metacritic score for the general population of games, but not for large games (due to the 'explosion' of popularity in 2009-2020) as discussed above.

The success of the interaction terms we devised to predict the Metacritic score better than the naive baseline, further reinforces the insights and hypotheses we discussed before.

Table 12: Coefficients of the Genre Aware Model

Coefficient	Value
user_ratio	44.13
age-large	-0.00082
action-user_ratio	-0.64
adventure-user_ratio	-0.9
sport-user_ratio	0.06
indie-user_ratio	-2.66
massively multiplayer-user_ratio	2.38
strategy-user_ratio	1.67
simulation-user_ratio	-2.53
intercept	41.55