Data Set Title
Exploratory Analysis

Name, Kadin McWilliams
Name, Zoe Mecklenburg

I. INTRODUCTION

We decided to use the steam games data set found on Kaggle that has over 85,000 steam games. We decided to use this data set because it had many categorical and numerical data columns that we would be able to choose from and do analysis on. We also both enjoy playing video games, so the data set lined up with our interest. The data set shows information like number of achievements, estimated ownership of games, name of the games and other general information on the games as well as critic ratings.

II. DATA SET DESCRIPTION

The data set contains 85,103 samples and 39 columns with various types of data (bool (3), int64 (14), object (20), float64 (2).

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Data Type(Pandas)*	*Missing Data (%)*
V1 AppID	interval	int64	0
V2 Name	nominal	object	>0.01
V3 Release Date	ordinal	object	0
V4 Estimated owners	ratio	object	0
V5 Peak CCU	ratio	int64	0
V6 Required age	interval	int64	0
V7 Price	ratio	float64	0
V8 DLC count	ratio	int64	0
V9 About the game	nominal	object	4.19
V10 Supported languages	ordinal	object	0
V11 Full audio languages	ordinal	object	0
V12 Reviews	nominal	object	88.55
V13 Header image	nominal	object	0
V14 Website	nominal	object	53.64
V15 Support url	nominal	object	51.12
V16 Support email	nominal	object	15.97
V17 Windows	ordinal	bool	0
V18 Mac	ordinal	bool	0
V19 Linux	ordinal	bool	0
V20 Metacritic score	interval	int64	0
V21 Metacritic url	nominal	object	95.4
V22 User score	interval	int64	0
V23 Positive	interval	int64	0
V24 Negative	interval	int64	0

V25 Score rank	interval	float64	99.95
V26 Achievements	ratio	int64	0
V27 Recommendations	interval	int64	0
V28 Notes	nominal	object	84.7
V29 Average playtime forever	ratio	int64	0
V30 Average playtime two weeks	ratio	int64	0
V31 Median playtime forever	ratio	int64	0
V32 Median playtime two weeks	ratio	int64	0
V33 Developers	nominal	object	4.21
V34 Publishers	nominal	object	4.51
V35 Categories	ordinal	object	5.4
V36 Genres	ordinal	object	4.18
V37 Tags	ordinal	object	24.79
V38 Screenshots	nominal	object	2.36
V39 Movies	nominal	object	7.58

III. Data Set Summary Statistics

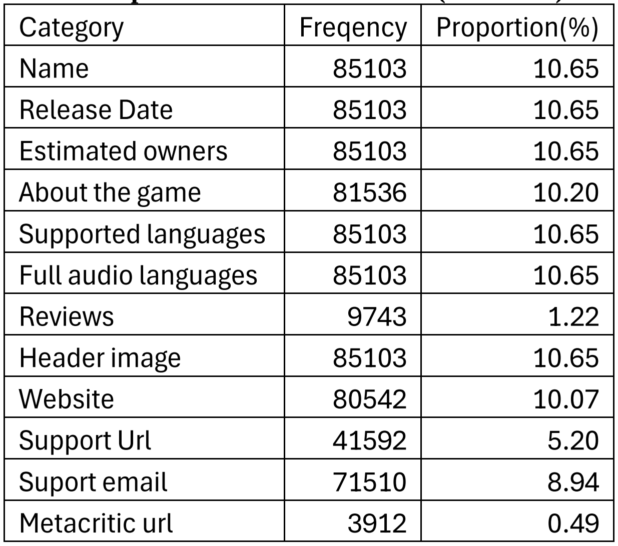
Since some of these variables are almost completely empty or were useless to the analysis, we decided to drop them. The list of variables we dropped are as follows: About the game, Median playtime two weeks, Categories, Screenshots, Tags, Average playtime two weeks, Score rank, App ID, Reviews, Header image, Movies. Many of these categories had links to things which would not be useful for analysis. For the rest of the tables there will only be the values that are not dropped since adding summary statistics will not provide anything meaningful.

Table 2: Summary Statistics for Steam Games

Variable Name	Count	Mean	Standard Deviation	Min	25th	50th	75th	Max
Peak CCU	85103	132.9	5403.5	0	0	0	1	872138
Required age	85103	0.313	2.25	0	0	0	0	21
Price	85103	7.19	12.36	0	0.99	4.49	9.99	999.98
DLC Count	85103	0.54	13.72	0	0	0	0	2366
Metacritic Score	85103	3.35	15.42	0	0	0	0	97
User Score	85103	0.03	1.79	0	0	0	0	100
Positive	85103	958.56	24359.2	0	0	7	45	5764420
Negative	85103	159.77	4574.58	0	0	2	14	895978
Achievements	85103	19.86	171.45	0	0	0	18	9821
Recommendations	85103	775.51	17893.38	0	0	0	0	3441592
Average Playtime Forever	85103	104.73	1142.45	0	0	0	0	145727
Median Playtime Forever	85103	93.32	1510.73	0	0	0	0	145727
Month	85103	6.79	3.42	1	4	7	10	12
Year	85103	2019.8	2.89	1997	2018	2020	2022	2025

Looking at the summary statistics of the above selection there are a few things that stick out in the data set. There tends to be a significantly greater number of positive reviews than there are negative reviews on Steam. This can be found from looking at the maxes of the summary statistics table. The median for the average play time and median play time are both 0 because a majority of the games have such a low player base that the extremely popular games are carrying the mean average for each game. I also noticed that the mean for DLC count being 20 made sense to me logically. Most of the games I play tend to have about, more or less, 20 achievements.

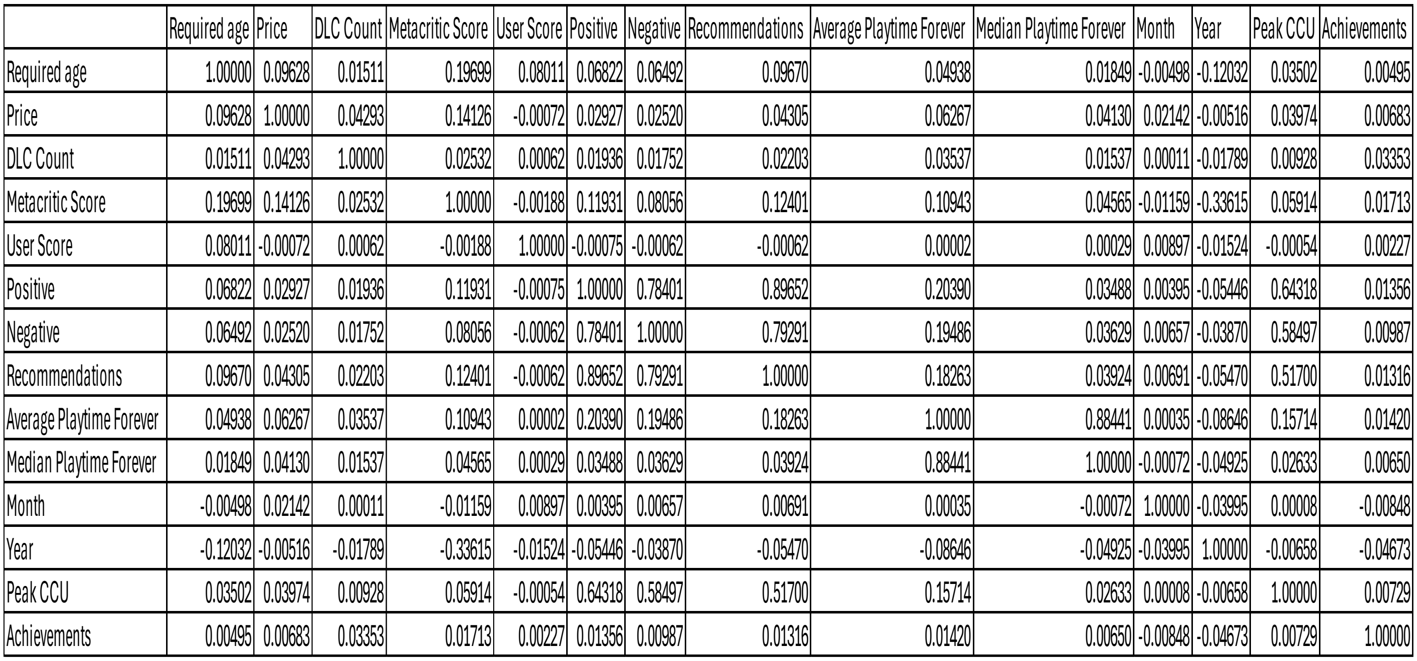
Table 3: Proportions for Steam Games (n=799453)



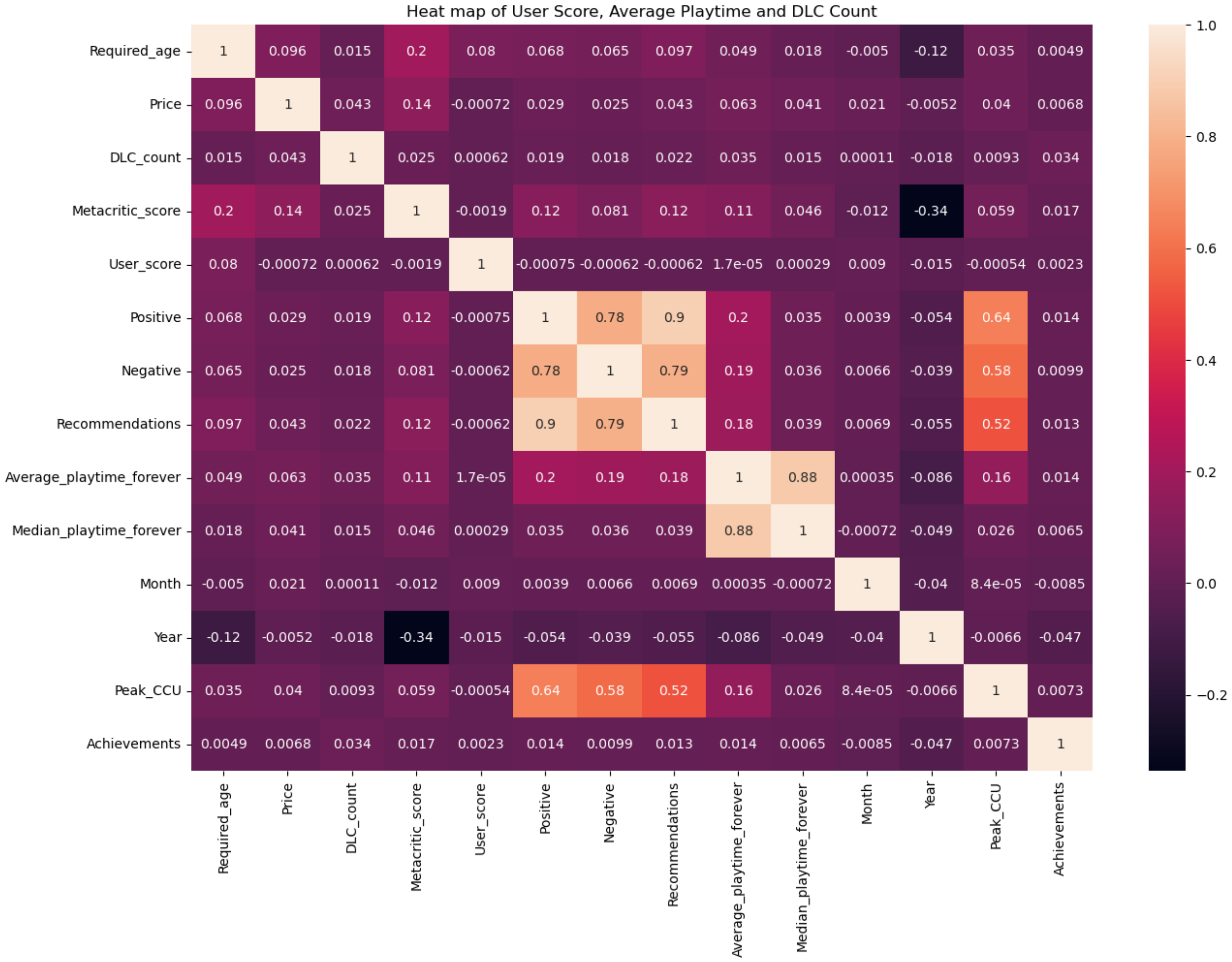
Category	Freqency	Proportion(%)
Name	85103	10.65
Release Date	85103	10.65
Estimated owners	85103	10.65
About the game	81536	10.20
Supported languages	85103	10.65
Full audio languages	85103	10.65
Reviews	9743	1.22
Header image	85103	10.65
Website	80542	10.07
Support Url	41592	5.20
Suport email	71510	8.94
Metacritic url	3912	0.49

The proportions above for categorical data show the frequency and proportion of the data compared to other categorical data. In my calculations I did not take it against every single variable in the sample size and only put it against the whole sample size of categorical data. When looking at the data you can see the reviews, support url, support email, and Metacritic URL are all low on the proportion percentage. I honestly don't know why the author of this data set decided to include them as there is no significant or meaningful way to interpret these. I think it does add validity to the data set but is not useful for analytics otherwise.

Table 4: Correlation Tables/Raw Correlation Data



	Required age	Price	DLC Count	Metacritic Score	User Score	Positive	Negative	Recommendations	Average Playtime Forever	Median Playtime Forever	Month	Year	Peak CCU	Achievements
Required age	1.00000	0.09628	0.01511	0.19699	0.08011	0.06822	0.06492	0.09670	0.04938	0.01849	-0.00498	-0.12032	0.03502	0.00495
Price	0.09628	1.00000	0.04293	0.14126	-0.00072	0.02927	0.02520	0.04305	0.06267	0.04130	0.02142	-0.00516	0.03974	0.00683
DLC Count	0.01511	0.04293	1.00000	0.02532	0.00062	0.01936	0.01752	0.02203	0.03537	0.01537	0.00011	-0.01789	0.00928	0.03353
Metacritic Score	0.19699	0.14126	0.02532	1.00000	-0.00188	0.11931	0.08056	0.12401	0.10943	0.04565	-0.01159	-0.33615	0.05914	0.01713
User Score	0.08011	-0.00072	0.00062	-0.00188	1.00000	-0.00075	-0.00062	-0.00062	0.00002	0.00029	0.00897	-0.01524	-0.00054	0.00227
Positive	0.06822	0.02927	0.01936	0.11931	-0.00075	1.00000	0.78401	0.89652	0.20390	0.03488	0.00395	-0.05446	0.64318	0.01356
Negative	0.06492	0.02520	0.01752	0.08056	-0.00062	0.78401	1.00000	0.79291	0.19486	0.03629	0.00657	-0.03870	0.58497	0.00987
Recommendations	0.09670	0.04305	0.02203	0.12401	-0.00062	0.89652	0.79291	1.00000	0.18263	0.03924	0.00691	-0.05470	0.51700	0.01316
Average Playtime Forever	0.04938	0.06267	0.03537	0.10943	0.00002	0.20390	0.19486	0.18263	1.00000	0.88441	0.00035	-0.08646	0.15714	0.01420
Median Playtime Forever	0.01849	0.04130	0.01537	0.04565	0.00029	0.03488	0.03629	0.03924	0.88441	1.00000	-0.00072	-0.04925	0.02633	0.00650
Month	-0.00498	0.02142	0.00011	-0.01159	0.00897	0.00395	0.00657	0.00691	0.00035	-0.00072	1.00000	-0.03995	0.00008	-0.00848
Year	-0.12032	-0.00516	-0.01789	-0.33615	-0.01524	-0.05446	-0.03870	-0.05470	-0.08646	-0.04925	-0.03995	1.00000	-0.00658	-0.04673
Peak CCU	0.03502	0.03974	0.00928	0.05914	-0.00054	0.64318	0.58497	0.51700	0.15714	0.02633	0.00008	-0.00658	1.00000	0.00729
Achievements	0.00495	0.00683	0.03353	0.01713	0.00227	0.01356	0.00987	0.01316	0.01420	0.00650	-0.00848	-0.04673	0.00729	1.00000



After the table with the raw data, include a heatmap of the correlation matrix as a figure.

IV. DATA SET GRAPHICAL EXPLORATION

Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

When looking at our findings as a whole and our graphs that we have found, there are a few things that stick out from our findings. An overwhelming majority of users are windows users and Linux and Mac were almost tied for number of users. There seems to be a minor correlation between having positive reviews and achieving a high peak CCU. It was found that the top 5 genres were: Indie, Casual, Action, Adventure, Strategy

A. Distributions

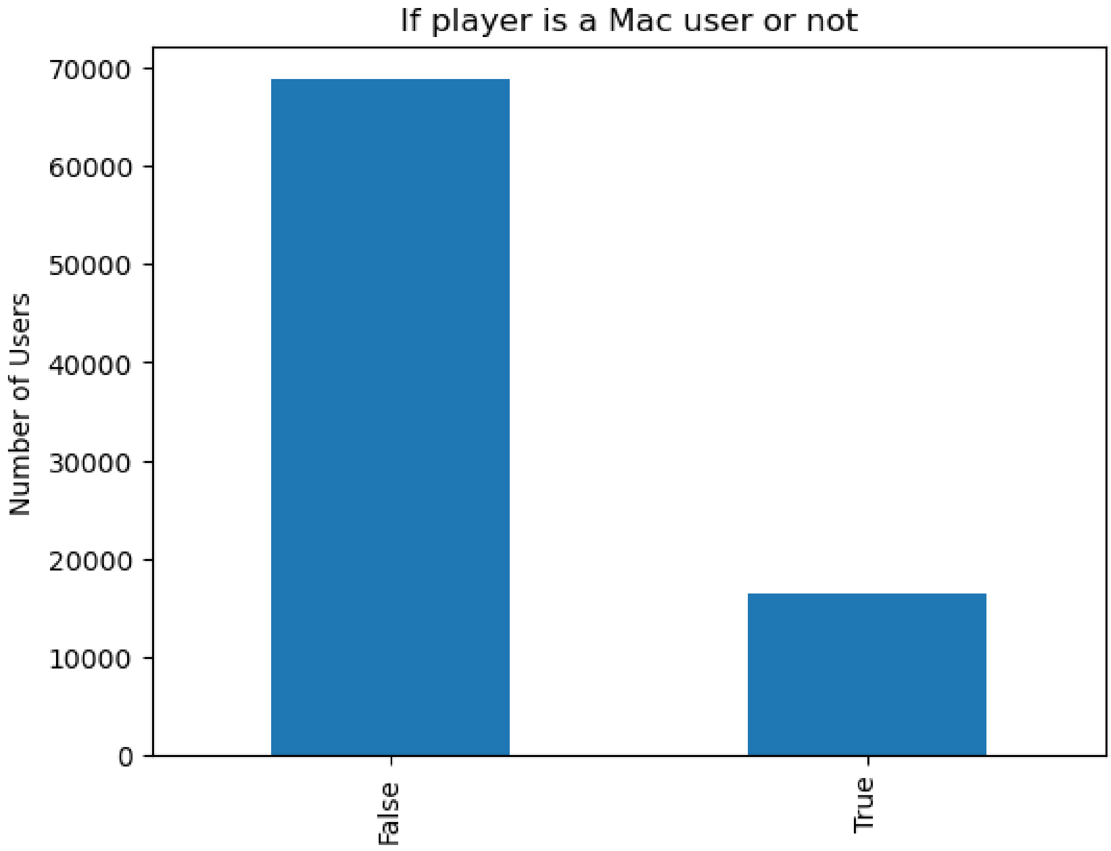


Figure 1: Distribution of Mac vs Non-Mac users

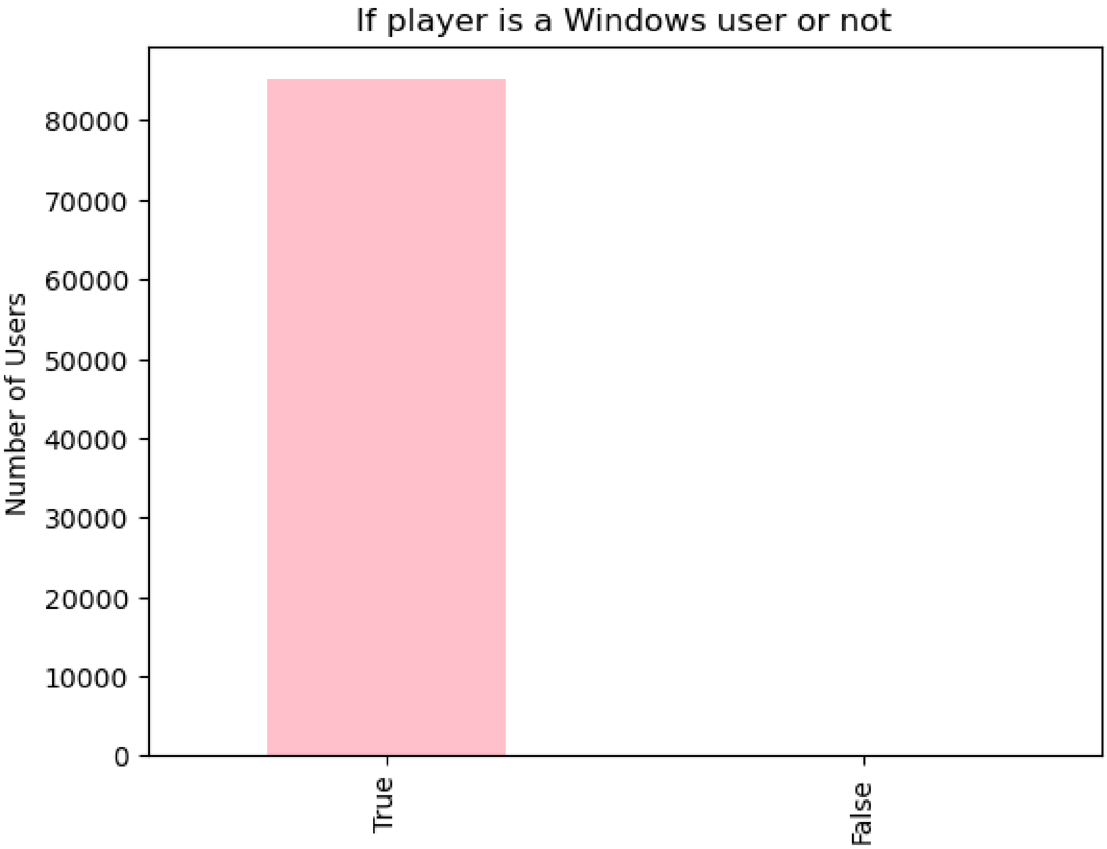


Figure 2: Distribution of Windows vs Non-Windows users

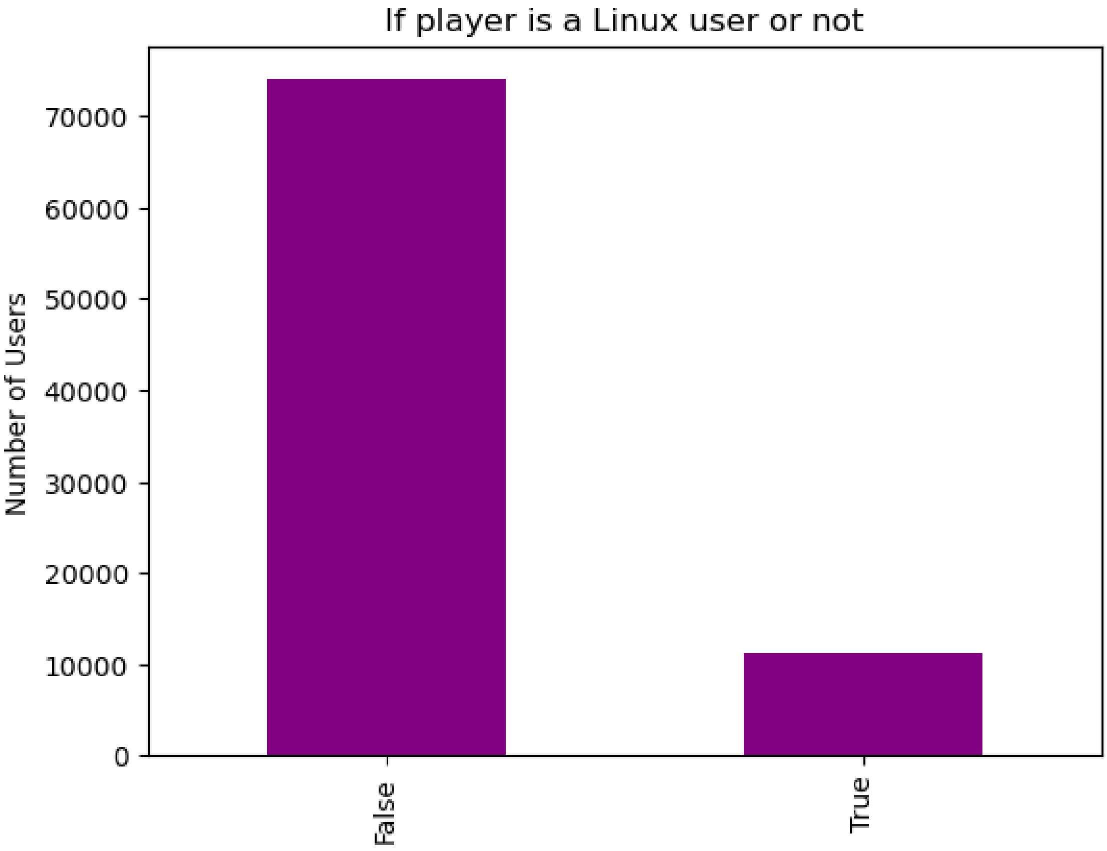


Figure 3: Distribution of Linux users vs Non-Linux users

B. ScatterPlots / Pairwise Plots (continuous variables)

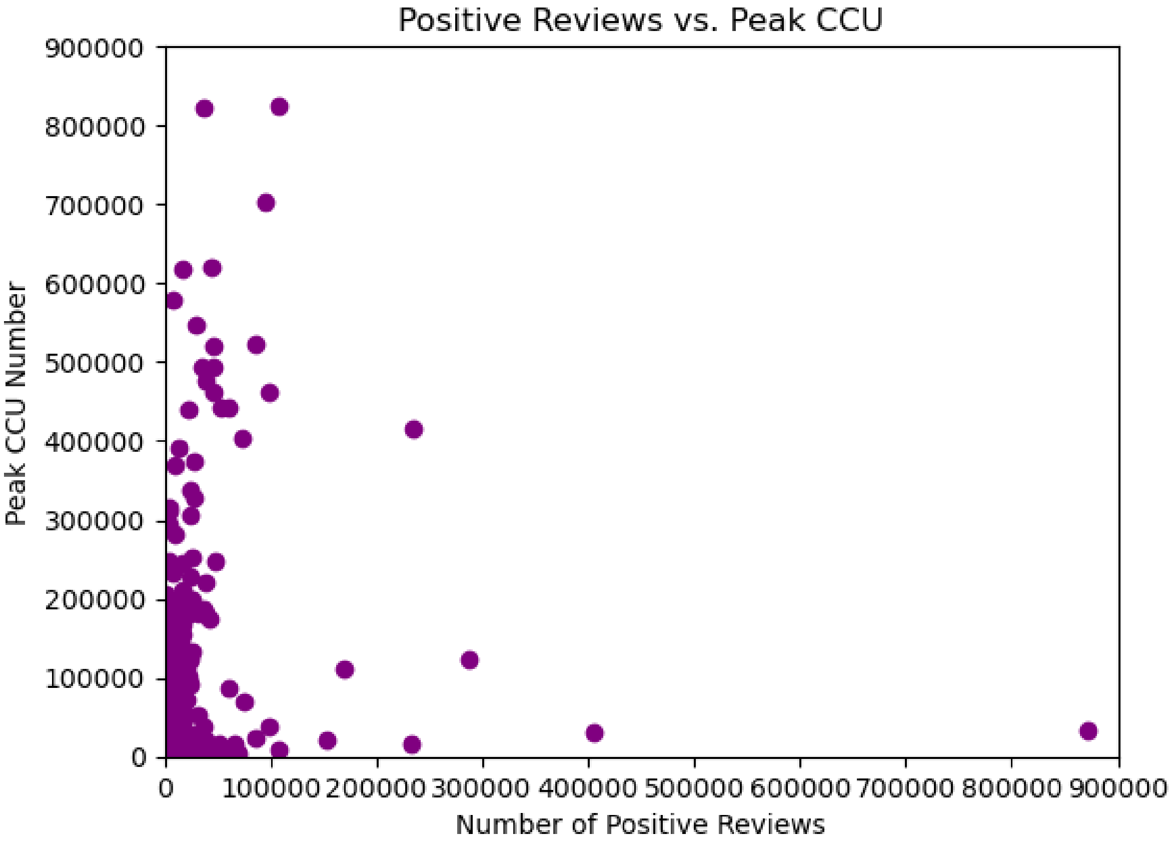


Figure 4: Scatter plot of Positive Reviews cs Peak CCU

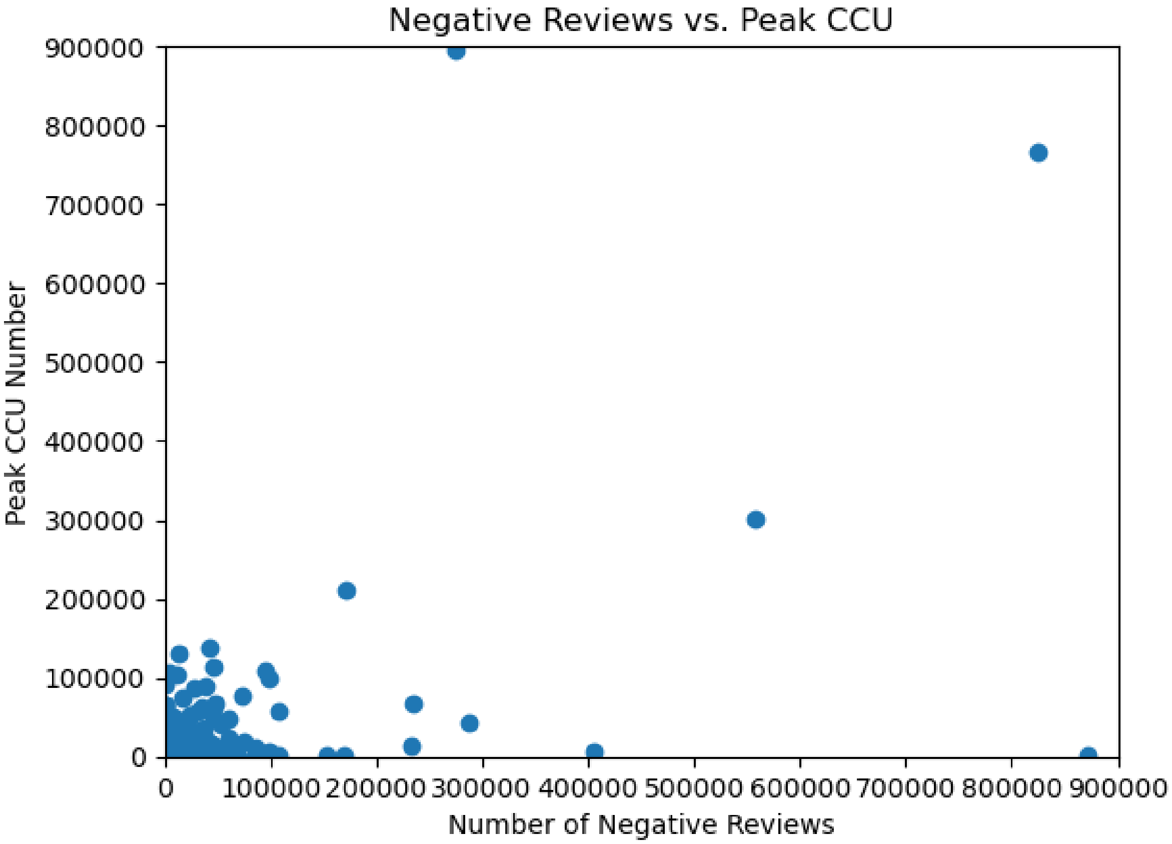


Figure 5: Negative Reviews vs. Peak CCU

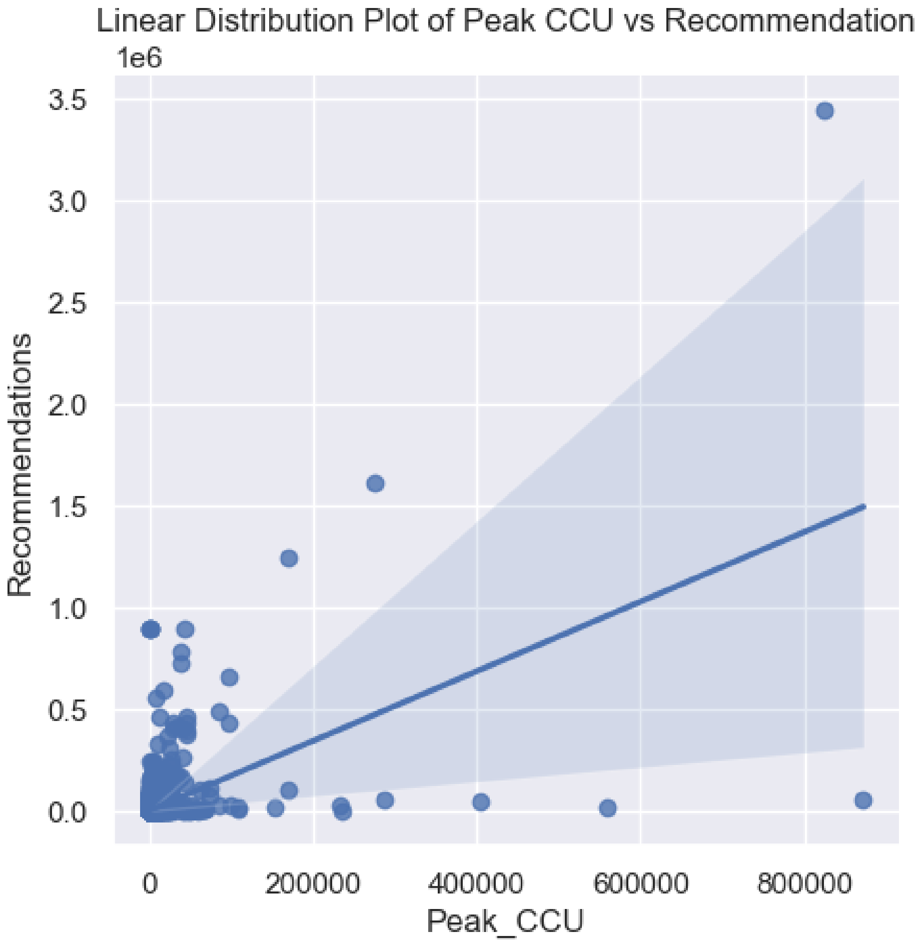


Figure 6: Linear Distribution Plot of Peak CCU vs Recommendation

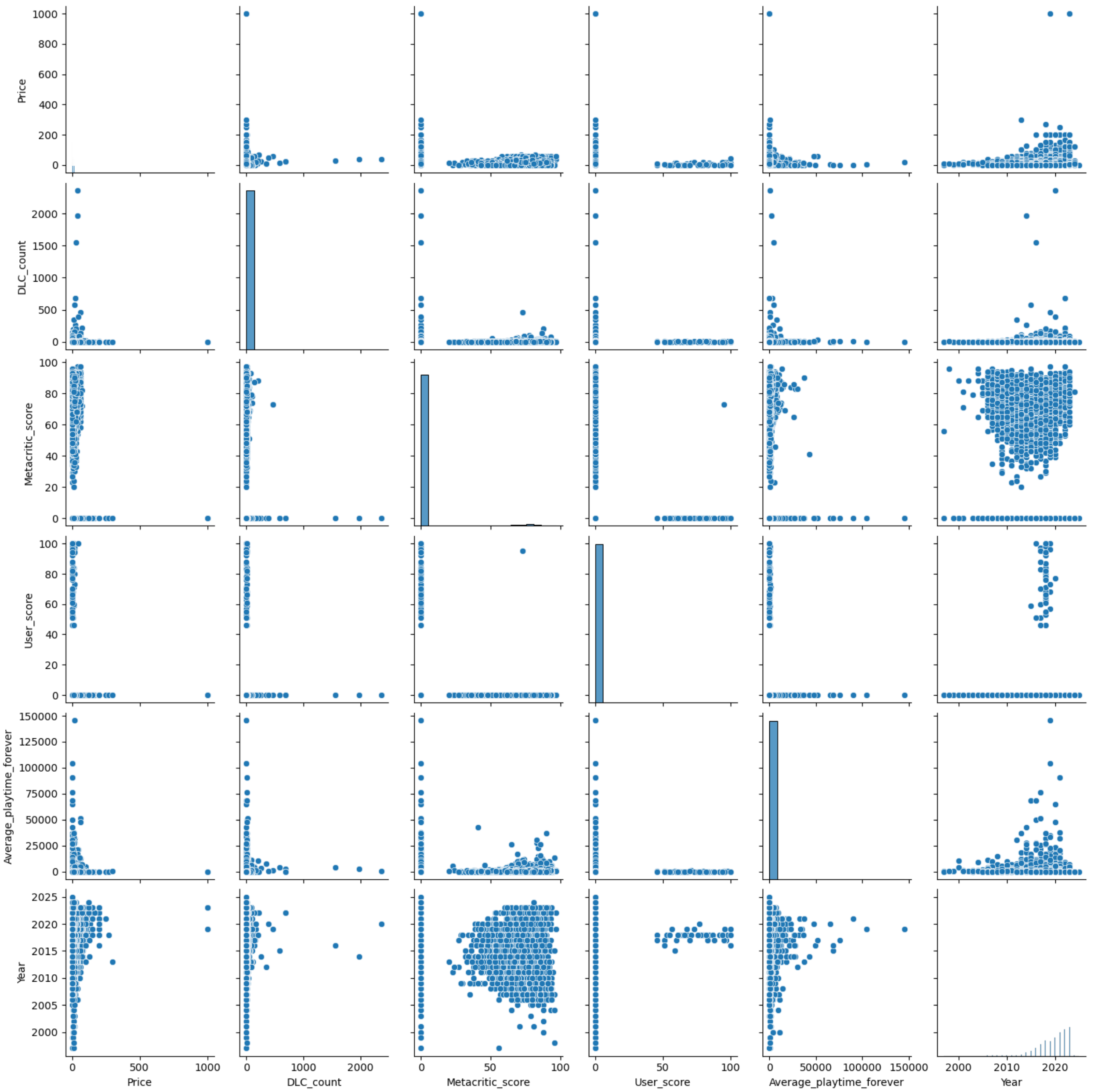


Figure 7: Pair plot of categorical data

C. Bar charts (categorical variables)

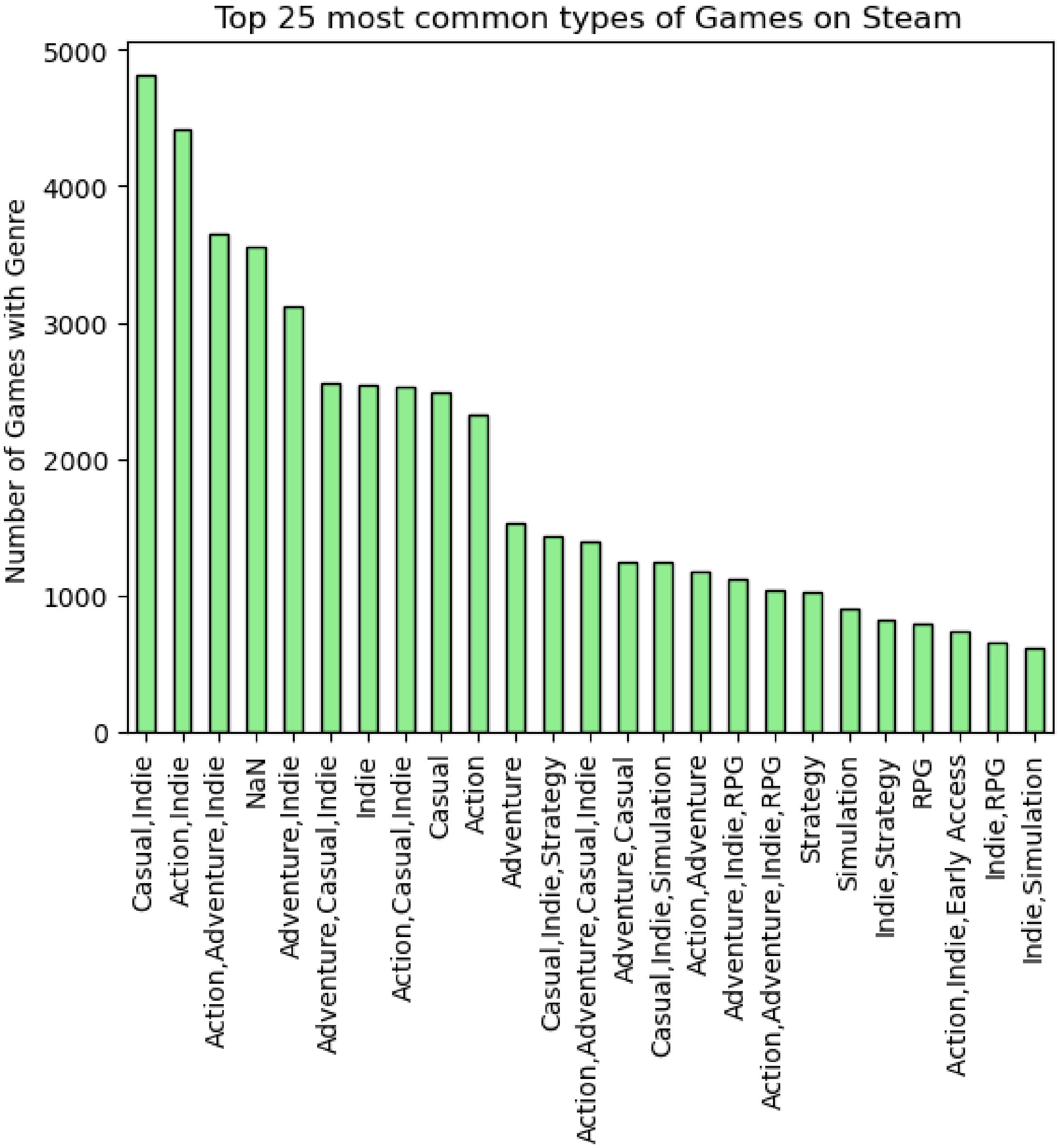


Figure 8: Top 25 most common types of Games on Steam

D. *Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.*

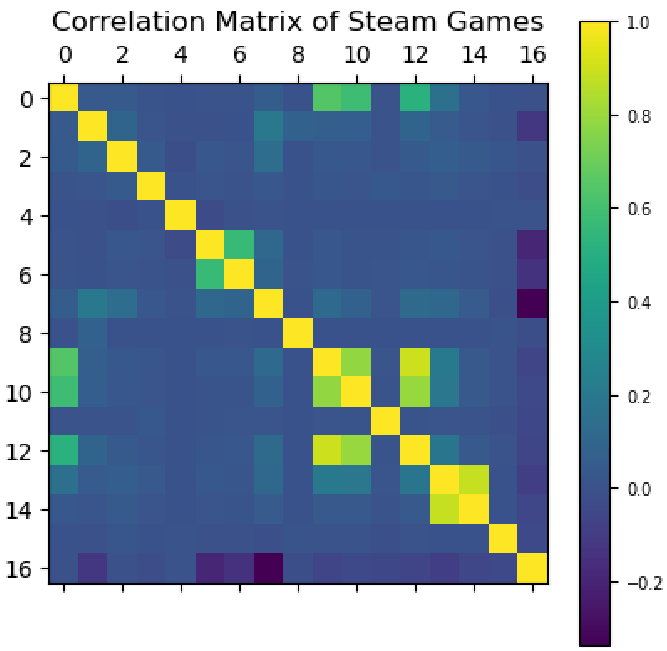


Figure 9: Correlation Matrix of Steam Games

V. SUMMARY OF FINDINGS

Finish up with a paragraph or two of summarizing your findings about this data set.

Summarizing our findings, we found out that there seemed to be minor influence on a games success rate based on reviews or other factors such as language availability and other factors like that. We found that actions games had the highest peak CCU and tended to be the most successful games overall. There are more indie and casual games on the market, but they do not keep a returning player base. We noticed that all the top games were all multiplayer games. There was not a top game that was a single player game. If someone wanted to market a game on the Steam webstore, we would recommend that they make an action and or adventure game and try to get as many positive reviews as possible. We would also tell them to make sure that they game is playable on windows if they have to pick only one system to play it on.