Data Science Project Presentation

PREDICTING THE SUCCESS OF A VIDEO GAME.

Students:

אדם קרפוביץ' 314080383
סרגיי גרשוב 327232450

About Our Project

The video game industry continues to expand year after year. Every studio/developer wants to grow revenue and reach out to more players and gaming communities.

The developers depend primarily on websites that aggregate reviews of video games (in terms of success).

The developers may see what gamers think about the next game by examining and evaluating this data and acting accordingly.

Research questions:

Can we forecast the success or failure of a released video game based on user and professional factors?

EDA:

What factors influence a video game's success?

Can we find unexpected results? (not logical)

The Process

01	02	03	04
• Data acquisition	• Data cleaning	• EDA	• Machine Learning

Step 1 - Data acquisition (API and crawling)

Step 2 – Data Handling

Step 3 – EDA (visualization and statistical tests)

Step 4 – Machine learning

Step 5 – Conclusion

Data Sources

https://www.metacritic.com

Crawling 01

https://steamspy.com

API 02

Metacritic - Selenium, BeautifulSoup, Pandas
Raw data size: 32K

SteamSpy - API, Pandas, Json, requests
Raw data size : 85K

Metacritic

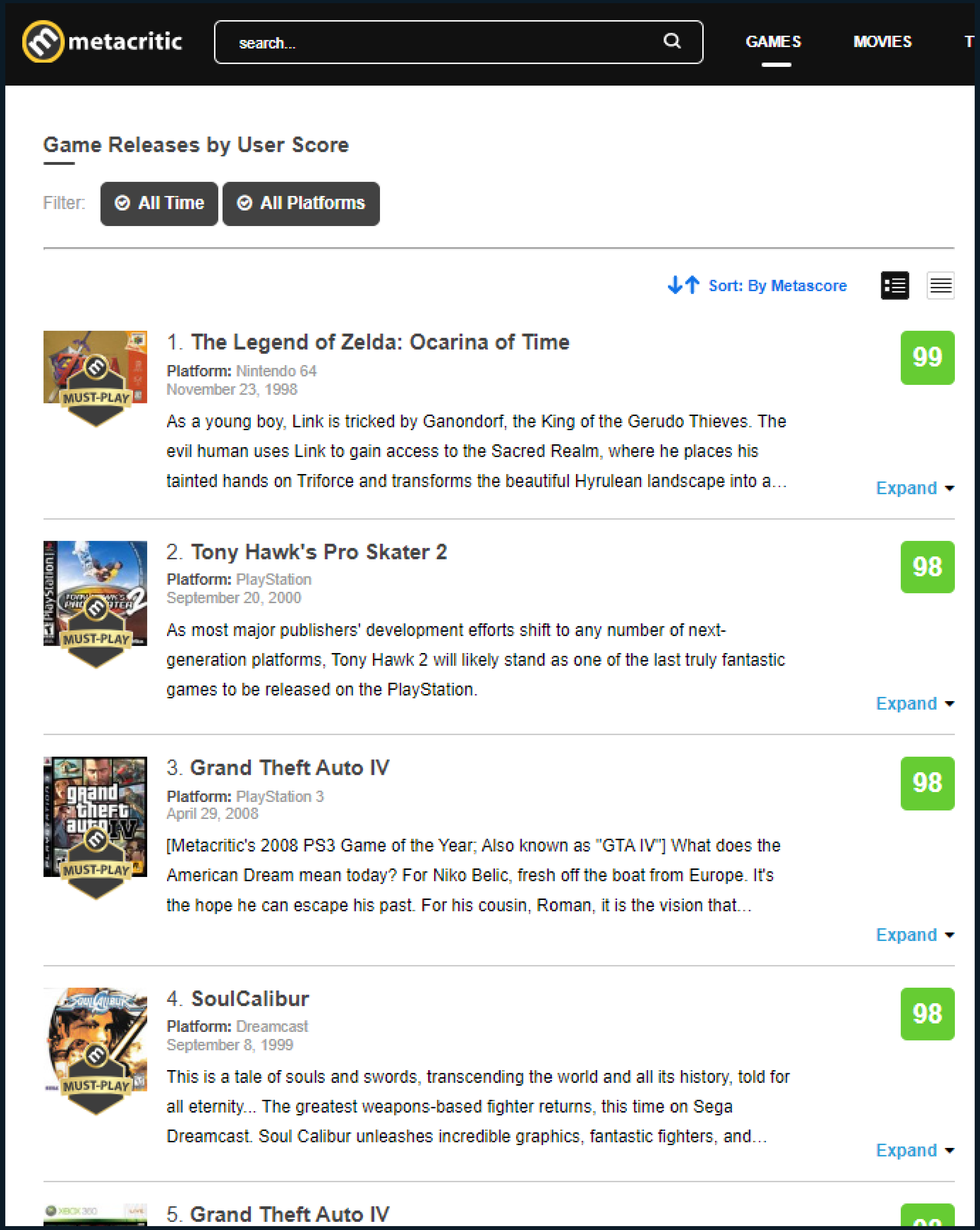
We compiled a list of all game URLs on the website from 2005 through the end of 2021.

Created a Selenium driver that iterated over the URLs and crawled across the sites. Downloaded the page source for each page.

Created a beautiful soup object that holds the page source and scraped the data of all games on the page, saving it in a dictionary that is needed to build a data frame.

We added a new column to the main data frame called "score" and iterating through both the final and Metacritic data frames to match the game's name and apply the appropriate score to that game.

MetaCritic




```
#Setting up selenium

#install service for selenium access
s=Service(ChromeDriverManager().install())

#toggle needed options
options = Options()
options.headless = True #open browser unseen

#create driver
driver = webdriver.Chrome(service=s, options=options)
#---------------------------------------------------------------------------

for mc_years, mc_details in zip(meta_critic_list_years, meta_critic_detail_list):
    driver.get(mc_years)
    ps = driver.page_source
    soup = BeautifulSoup(ps, 'html.parser')
    page_numbers = int(soup.find("li", class_ = "page last_page").find("a").string)
    for i in range (0, page_numbers):
        driver.get(mc_years + '&page=' + "{0}".format(i))
        ps = driver.page_source
        soup = BeautifulSoup(ps, 'html.parser')
        meta_critic_detail_list[mc_details].append(soup.find_all(class_ = "clamp-summary-wrap"))

driver.quit()
```

SteamSpy

Getting the data from SteamSpy was made by a API call from the website. Some columns, however, were unrelated to our study.

We used the API to gather almost 10 000 rows and saved them as a data frame.

SteamSpy

Return format for an app:

- appid - Steam Application ID. If it's 999999, then data for this application is hidden on developer's request, sorry.
- name - game's name
- developer - comma separated list of the developers of the game
- publisher - comma separated list of the publishers of the game
- score_rank - score rank of the game based on user reviews
- owners - owners of this application on Steam as a range.
- average_forever - average playtime since March 2009. In minutes.
- average_2weeks - average playtime in the last two weeks. In minutes.
- median_forever - median playtime since March 2009. In minutes.
- median_2weeks - median playtime in the last two weeks. In minutes.
- ccu - peak CCU yesterday.
- price - current US price in cents.
- initialprice - original US price in cents.
- discount - current discount in percents.
- tags - game's tags with votes in JSON array.
- languages - list of supported languages.
- genre - list of genres.

	name	developer	publisher	positive	negative	owners	average_forever	median_forever	price
0	Counter-Strike: Global Offensive	Valve, Hidden Path Entertainment	Valve	5718191	761211	50,000,000 .. 100,000,000	28846	6493	0
1	Dota 2	Valve	Valve	1467658	297030	100,000,000 .. 200,000,000	38219	964	0
2	Grand Theft Auto V	Rockstar North	Rockstar Games	1153983	208800	20,000,000 .. 50,000,000	12460	6441	1480
3	PUBG: BATTLEGROUNDS	KRAFTON, Inc.	KRAFTON, Inc.	1146769	892200	50,000,000 .. 100,000,000	21155	6620	0
4	Terraria	Re-Logic	Re-Logic	951689	20646	20,000,000 .. 50,000,000	6126	1739	499
...	...	...	...	...	...	...	...	...	...
9634	HEDE Game Engine	Hede Games	Hede Games	0	7	50,000 .. 100,000	0	0	99
9635	League Space	EURO GAMES STUDIO	EURO GAMES STUDIO (ESP)	0	1	50,000 .. 100,000	0	0	1199
9636	RagDoll MadDoll	Team Booky	Dystopian Edge Publishing	0	1	50,000 .. 100,000	0	0	124
9637	Operation: VICUS	Ilja Soutchilin, Tom Berger	Ilja Soutchilin, Tom Berger	0	2	50,000 .. 100,000	0	0	799
9638	RACE On	SimBin	SimBin	0	0	100,000 .. 200,000	0	0	799

9639 rows × 9 columns

steamspy

Search:

#	GAME	RELEASE DATE	PRICE	SCORE RANK (USERSCORE / METASCORE)	OWNERS
1	Tiny Tina's Wonderlands	Jun 23, 2022	$47.99	N/A (N/A)	200,000 .. 500,000
	Sonic Origins	Jun 22, 2022	$39.99	N/A (N/A)	0 .. 20,000
	Raft	Jun 20, 2022	$16.99	N/A (N/A)	5,000,000 .. 10,000,000
	Capcom Fighting Collection	Jun 23, 2022	$39.99	N/A (N/A)	0 .. 20,000
	Good Company	Jun 21, 2022	$19.99	N/A (N/A)	200,000 .. 500,000
	FINAL	Jun 17, 2022	$49.69	N/A (N/A)	100,000 .. 200,000

Data handling

In the 'publisher' and 'developer' columns, there were some NaN values. We duplicated the 'developer' name into the 'publisher' and vice versa instead of dropping those rows.

We saved alot of rows and data by doing so. We also removed duplicates and limited edition games.

The values of the 'price' were changed from cents to dollars.

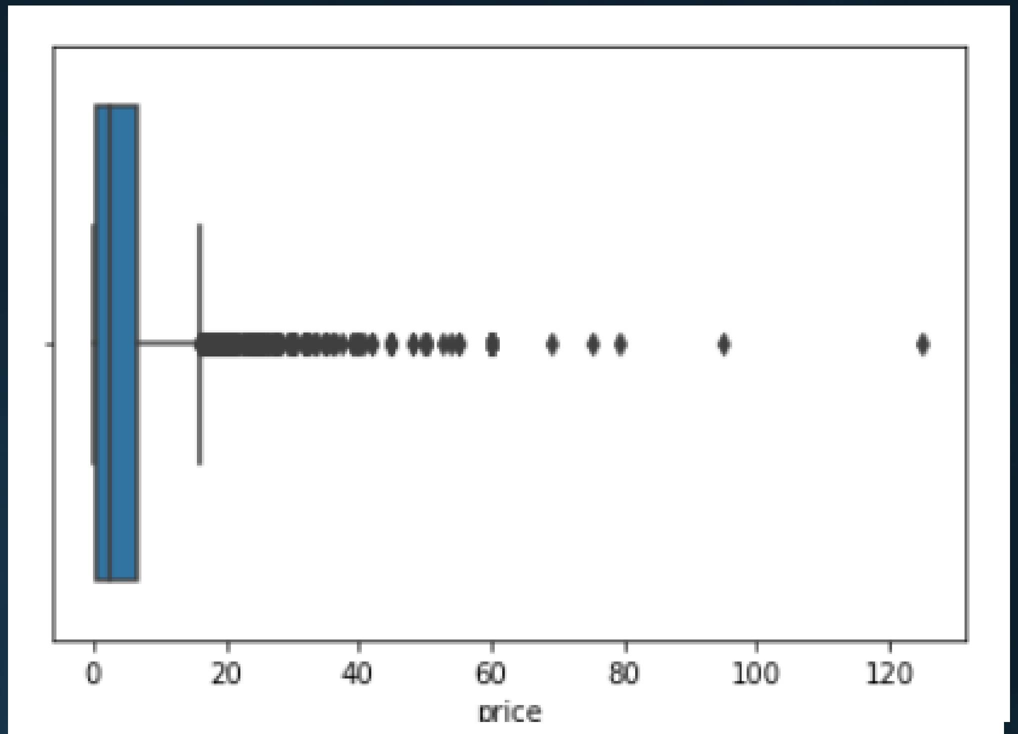
The values for 'owners' were a range of owners, we replaced them with the average of that range and renamed the column 'owners approx.'

Some titles were not listed on Metacritic. We used the 'positive' and 'negative' values, which describe the quantity of positive and negative input, to handle their'score' value from SteamSpy.

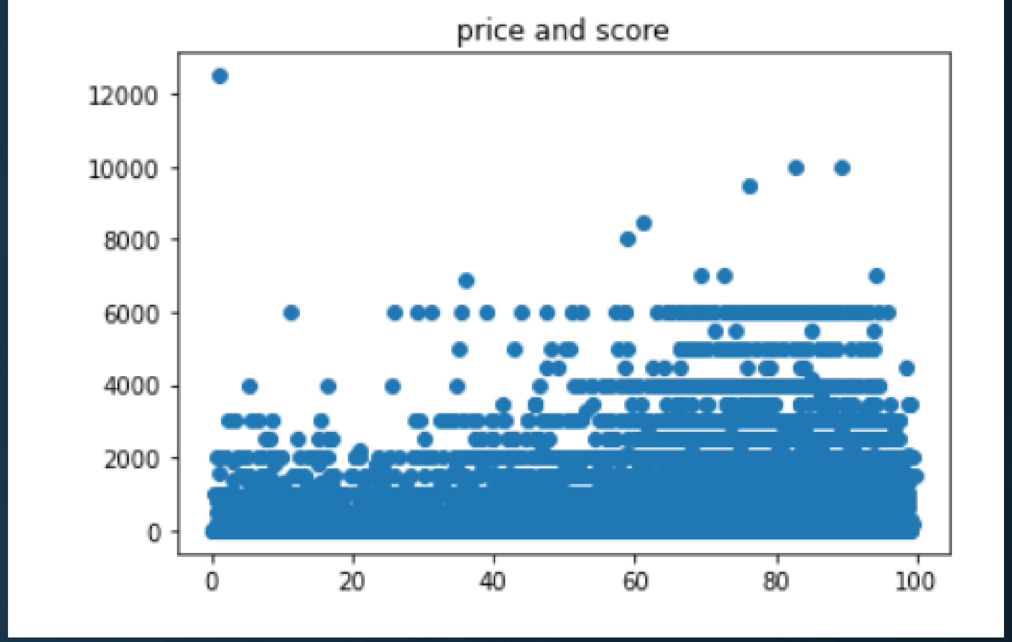
EDA and statistical tests

The first step was to use a bar plot to represent the numeric columns. To better understand the distribution, we scaled each column.
The second step was to examine the relationships between the 'score' column and the other numeric columns. We utilized the Spearman approach because none of the columns had a normal distribution. The scatter plot between score and other columns, as well as the actual correlation value between said columns, were also shown.

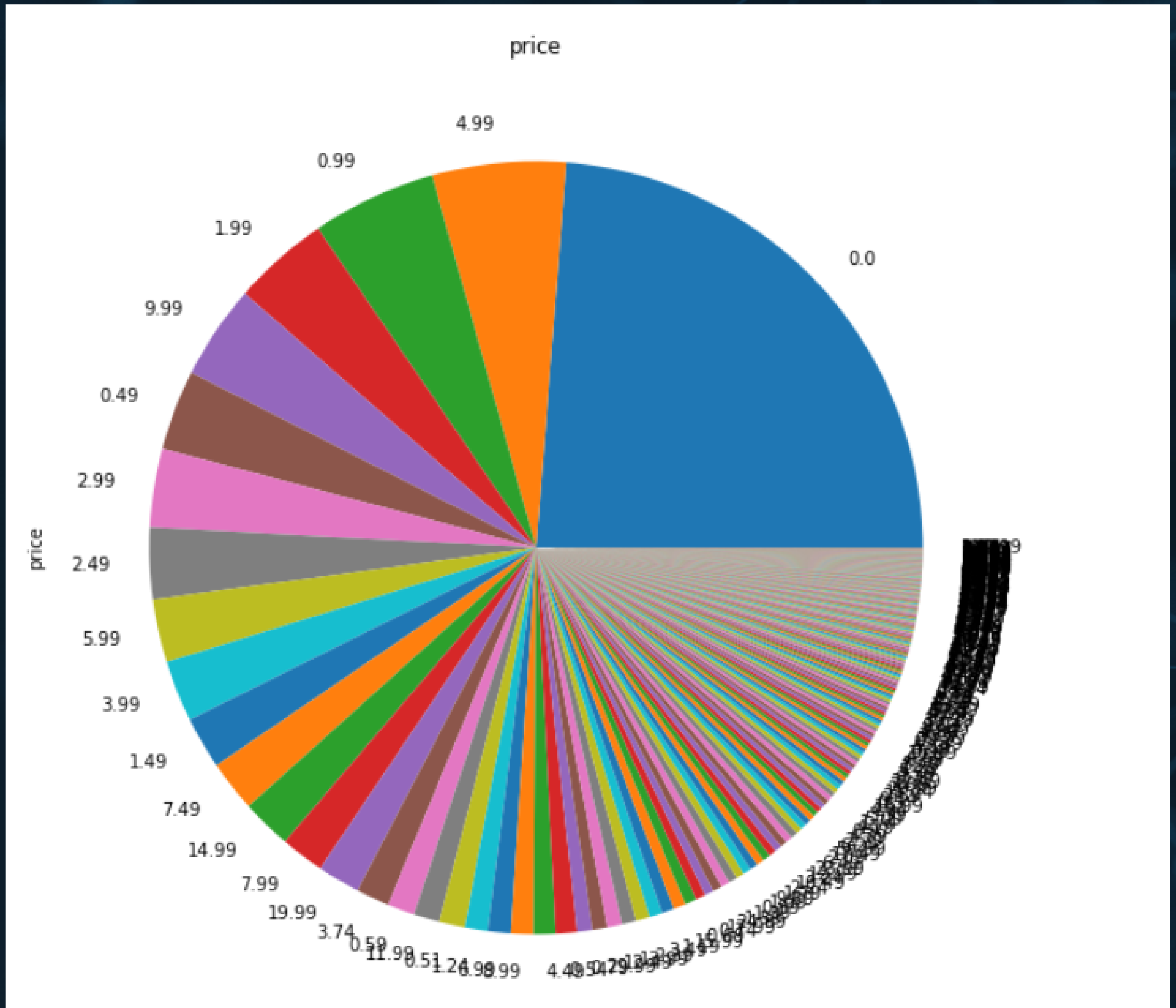
Eda



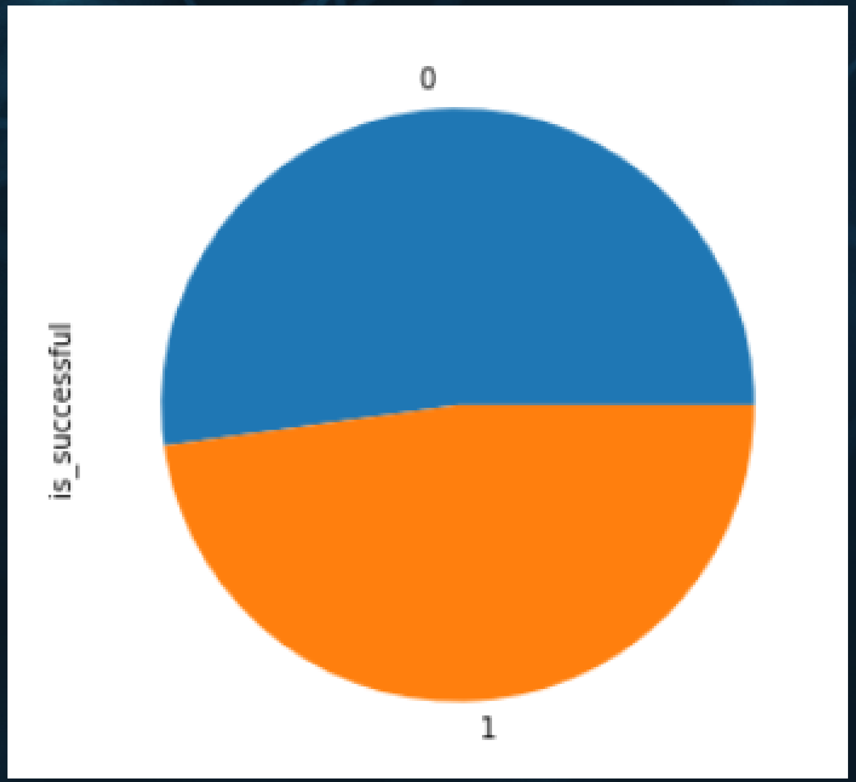
0
20
40
60
80
100
120
price



price and score
12000
10000
8000
6000
4000
2000
0
0
20
40
60
80
100

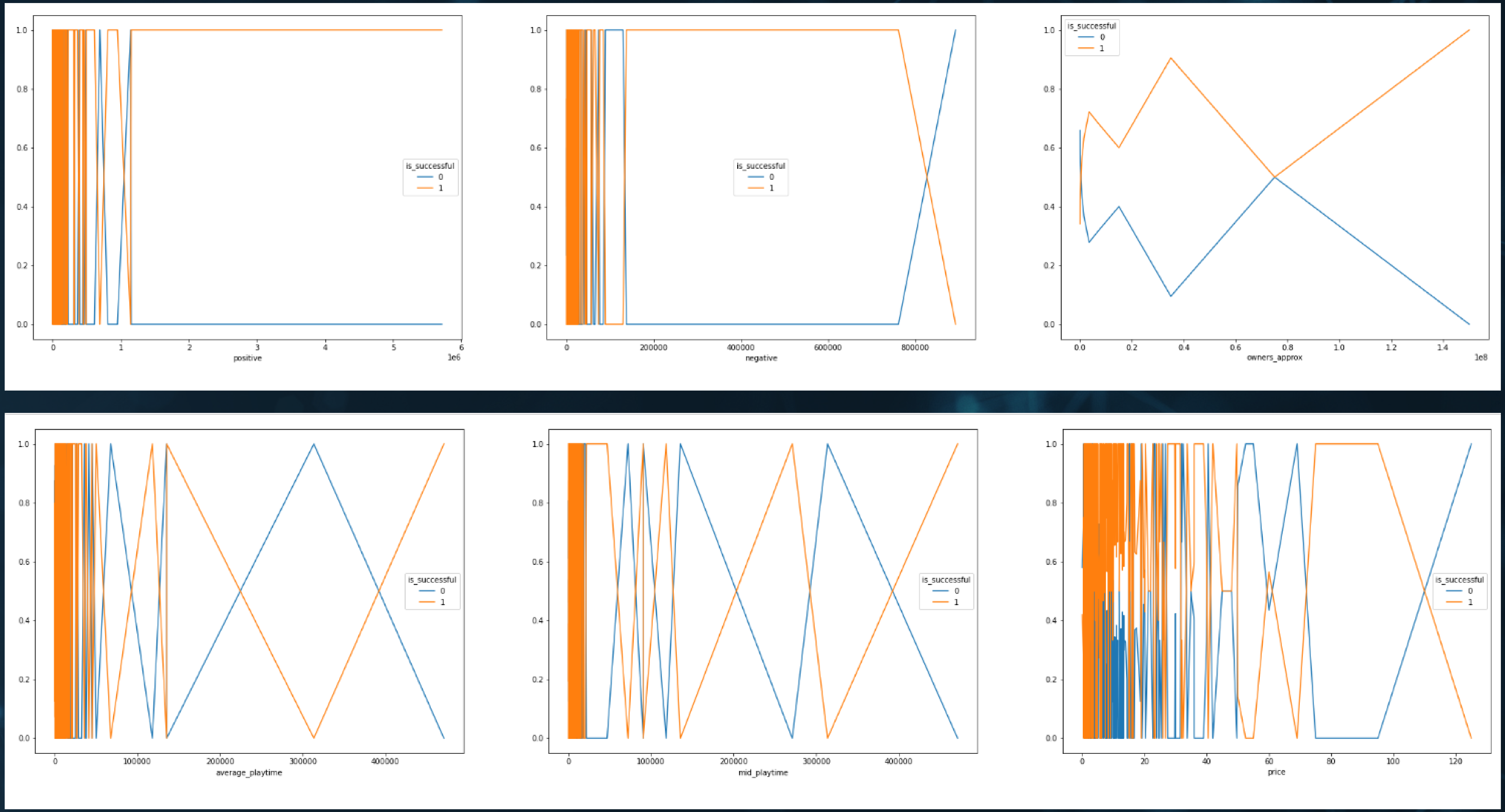


price
price
0.0
4.99
0.99
1.99
9.99
0.49
2.99
2.49
5.99
3.99
1.49
7.49
14.99
7.99
19.99
3.74

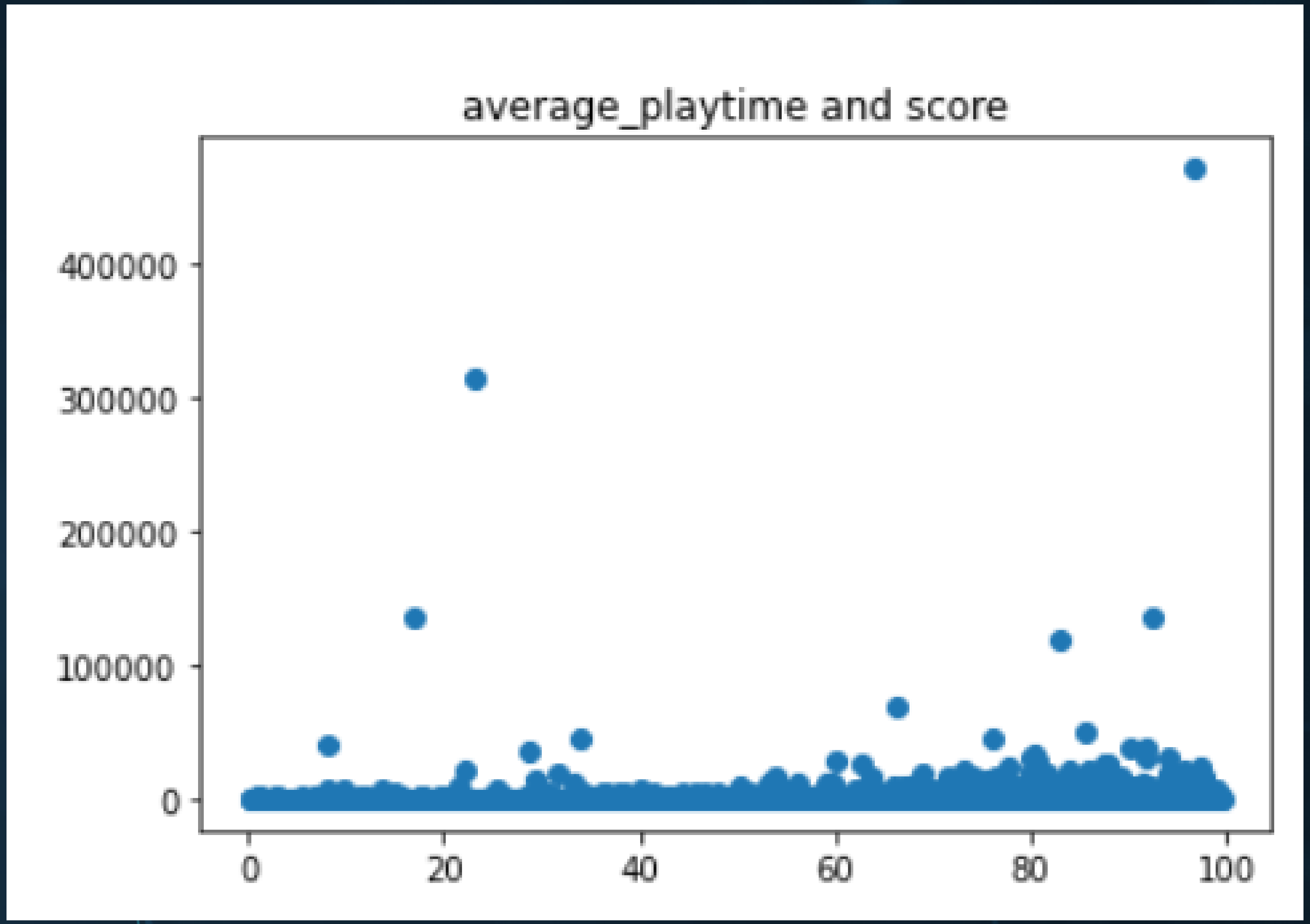


0
1
is_successful

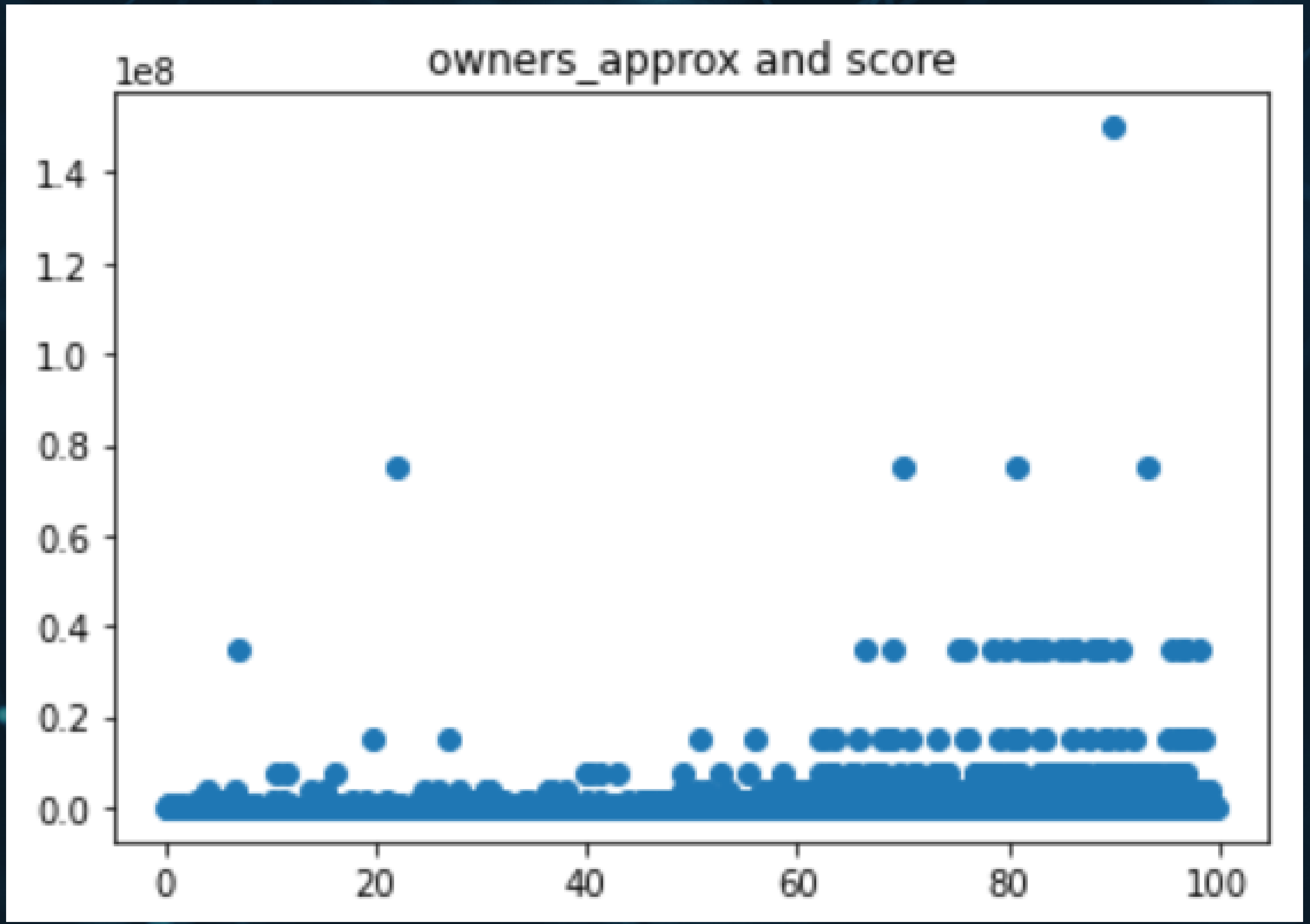
Eda


is_successful
0
1
positive
1e6
negative
owners_approx
1e8
average_playtime
mid_playtime
price

Eda



average_playtime and score
400000
300000
200000
100000
0
0
20
40
60
80
100



owners_approx and score
1e8
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100

Machine learning

We employed a supervised method in the machine learning phase.
All of the numeric columns were included in the feature vector (X), except for the 'is successful' and score columns ('is successful' is the goal column (y)).
We utilized logistic regression to train the model, which is designed for binary problems (in our instance, 0 or 1).

The accuracy of the prediction was impressive: 85%.
We divided the data frame into training and testing groups and compared the measurements of Logistic regression and the random forest model as well.

Our major strategy was to use logistic regression, which yielded an accuracy of 85%.
As well as that we did the ensemble wombo combo just for the fun of it.

Conclusion

1. The unscaled model produced excellent results.

Eda:
2. We discovered that some characteristics can indeed indicate whether a game will succeed.
The price and the postive reviews of a game really matter to sigma gamers.

3. A game can be successful even tho the amount of negative reviews is high and the game can be unsuccessful and have a very high price.

Thank You!

אדם קרפוביץ' 314080383 |
adambordav.k@gmail.com

סרגיי גרשוב 327232450 |
gers7777@gmail.com

מתרגל: בוירסקי תומר