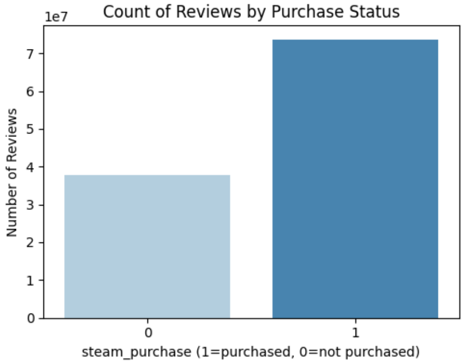
Hypothesis: Reviews from users who have purchased the game (as opposed to obtained it for free) will be rated as more helpful.

1. What's a balance of reviews from users who purchased the game vs obtained for free?

First we need to balance to see how balanced out data, to study our hypothesis:

```
        steam_purchase     review_count
0         0   (False)        37719573
1         1    (True)        73658878
```

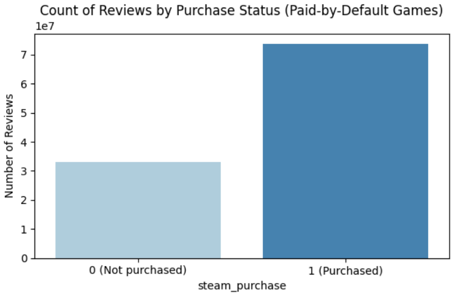


It turns out that most of the users who left the review purchased the game, but we don't know if some of the games are free by default. For example, games like CSGO, PUBG, .. are free, while DOTA2 is paid.

2. What's a balance of reviews from users who purchased the game vs obtained for free among the games that are paid by default?

For this we need to query all the games, where there was at least one review whose author paid for the game.

```
        steam_purchase     review_count
0         0  (False)         33072286
1         1   (True)         73658727
```



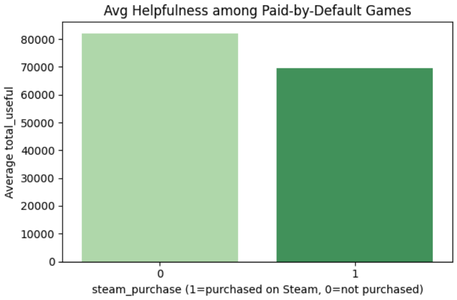
Now, the number of reviews significantly decreased, since we are not considering games that are free, so there cannot be paid reviews.

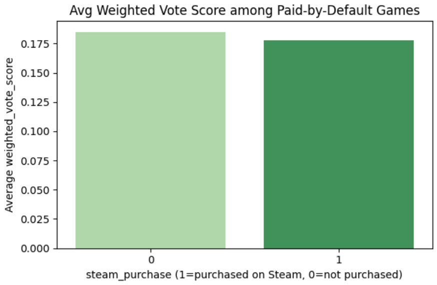
3. What's the average helpfulness of the review for both these groups?

To test this, we need to first understand that dataset provides three attributes:

1. Votes_up: the review that was marked as helpful
2. Votes_funny: the review that's marked as funny
3. Weighted_vote_score: Generated and provided by Steam

Since we don't know how vote scores are weighted by STEAM, we will calculate the average based on the sum of `votes_up` and `votes_funny` for paid and obtained for free reviews, and in the second chart we will use `weighted_vote_score` by STEAM.

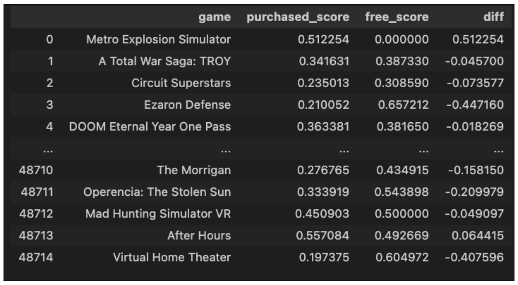




It turns out that games that were obtained for free on average have a better helpfulness for the reader, which questions our initial hypothesis.

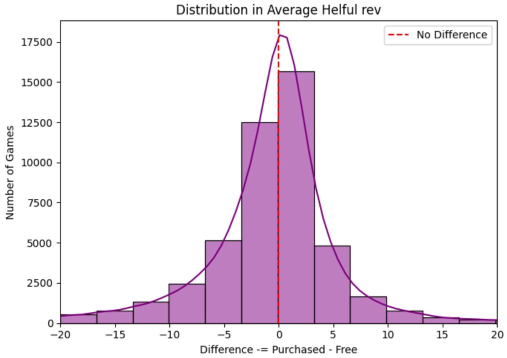
4. What's causing this? Maybe there are games that have extremely hateful or positive reviews?

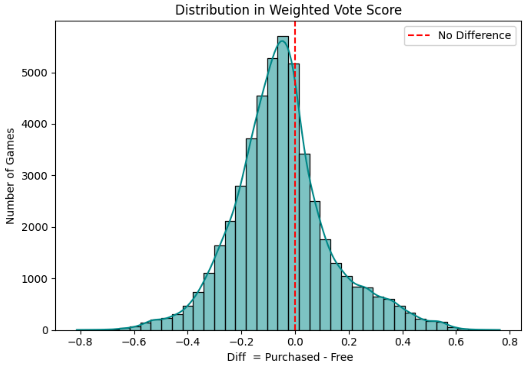
For this, I've reformatted the data frame by games and added another column `diff`, which is a difference between the average purchased score and average obtained for free score.



	game	purchased_score	free_score	diff
0	Metro Explosion Simulator	0.512254	0.000000	0.512254
1	A Total War Saga: TROY	0.341631	0.387330	-0.045700
2	Circuit Superstars	0.235013	0.308590	-0.073577
3	Ezaron Defense	0.210052	0.657212	-0.447160
4	DOOM Eternal Year One Pass	0.363381	0.381650	-0.018269
...	...	...	...	...
48710	The Morrigan	0.276765	0.434915	-0.158150
48711	Operencia: The Stolen Sun	0.333919	0.543898	-0.209979
48712	Mad Hunting Simulator VR	0.450903	0.500000	-0.049097
48713	After Hours	0.557084	0.492669	0.064415
48714	Virtual Home Theater	0.197375	0.604972	-0.407596

Then we will plot them into a distribution graph excluding 1% of games that have extreme differences.





Even excluding 1% of extreme values, both graphs show relatively the same distributions of relations of reviews from users who paid for the game vs obtained for free. Hence, most likely the users who paid for the game do not necessarily leave more constructive/helpful reviews.

5. Why might free reviewers be just as (or even more) helpful than purchasers?

Since our hypothesis was rejected, my next assumption is that most of the reviews from authors who obtained the game for free tend to write a better reviews in general that reaches positively the audience.