Worldwide Gaming

Tucker Mackie

Fig. 1: Gaming across the world.

Abstract—This project breaks down how worldwide gaming trends breakdown by using datasets based on Steam and PlayStation users. This dataset was sourced from Kaggle. By seeing how different locations have a preference of games or a specific genre, you can build knowledge by targeting specific demographics. I built an interactive map to breakdown how gaming is portrayed across the world along with having a bar plot for the ten most popular games. There is a given data table that breakdown global data along with country specific data as well. This application works to show how regional difference impact gaming to understand market niches how gaming is impacted at a global level. A free copy of this paper and all supplemental materials are available at https://github.com/TMackie1116/Worldwide_Gaming.

1 Introduction

Gaming has evolved into a global sensation with different preferences in every corner of the globe. Not only are there different preferences but the behaviors and tendencies are different depending on the location you are looking at. Using data from both Steam and PlayStation to understand the location breakdowns. There are given gaming sites and developers that do not release their data as they protect their data. Major games and develops such as League of Legends and Fortnite being published by Riot Games and Epic will not be included in this. The data for this project was obtained from Kaggle [1].

2 Author Details

Tucker Mackie
Masters of Applied Data Science
University of North Carolina Chapel Hill
tmackie@unc.edu

3 Methods

3.1 Data Collection

Using Kaggle [1], I was able to compile the given data for my project. Using this dataset, I was given three scopes of data based on the given platforms: Xbox, PlayStation, and Steam. For my project, I had to omit the Xbox data as they did not have countries listed, and without that, it would have been deleted anyway. Instead of going through the process of cleaning it for it to eventually be deleted, I chose to leave it out altogether.

The raw data was comprised of multiple .csv files, corresponding to players, games, achievements, purchases, and gameplay history. Using these different metrics, I later combine all the .csv files into one master dataset.

3.2 Data Loading and Processing

As taught in this class, along with other classes, managing larger datasets can be very tricky. The key is to manage them efficiently and ensure that you are not doing extra work that causes you to add more load to your computer. In doing this, I found a method that loads all of the .csv files by using a loop method to load each given file. I had this broken down into chunks of 100,000 rows each, as some files had hundreds of thousands of rows, and upwards of several million rows. With data this size, I wanted to avoid overloading a computer and making sure it was loaded seamlessly and efficiently.

With the datasets being broken up by platform, this step made it easier to keep the data separated by their respective platforms. I used an if-else loop to look in the file name and see if "PlayStation" or "Steam was listed to add in that respective column into the data. This allowed me to split the data up into the platforms when cleaning the data, as well as building my visuals. This then allowed me to have my files split into five categories, along with their respective datasets: `games`, `players`, `achievements`, `purchased`, and `history`.

As I had these five datasets for both systems, I followed the same process across both platforms to then combine them into one master dataset. Finally, when processing the data, I worked on exploding the `library` column for the `purchased` datasets. The reasoning for this was that one player would have multiple games within their library, and this allowed them to view games as individuals instead of looking at a whole library. Everything was stored within data frames to clean the

• *Tucker Mackie. E-mail: tmackie@unc.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

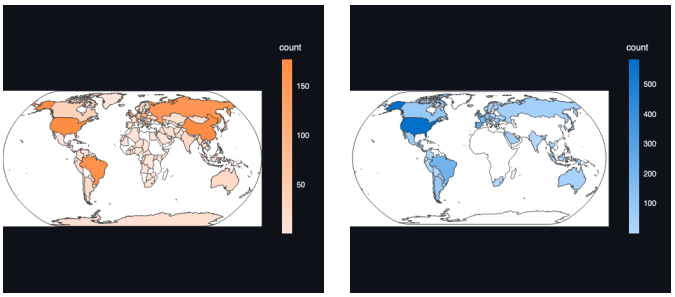

Fig. 2: Comparison of player counts per country by platform: (left) Steam, (right) PlayStation.

data.

3.3 Data Cleaning

I had to take several given steps to make sure my data was ready to use within a Streamlit application. To start with, I dropped any columns that were not needed, one of these being `developers`. The columns I dropped were not going to be used, and keeping them in would cause more data to be parsed through to make the given datasets larger. The next step I had was creating a filtered dataset, so my main dashboard would not be bogged down to cause the application to not run. Streamlit has limitations, and the more data it has, the slower it can run, eventually not being able to run at altogether. To prevent getting to this point, my main dataset for the world map focuses on the trailing year.

For Streamlit, when parsing NAs through, you have difficulties, as the application does not translate NAs very well. To counteract this issue, I worked through each layer to remove NAs. In doing so, I dropped all missing values across every data frame. To create a master data file, I had to merge these, and in doing so, after each merge, I would drop the missing values. This not only got rid of the rows with an NA, but it also eliminated some of the run time, as I did not have to run merges that were unnecessary. Along with getting rid of the missing values, I also made sure to get rid of duplicated rows. Some of the reasons for the duplicated rows are from the achievements data, as you will have multiple achievements for that given game. Since I am using this as a basis for player count, these duplicates would inflate the numbers for how many players are on a given game.

3.4 Data Merging

One of the major steps for this project was getting a usable final dataset that was not broken into pieces. In doing so, I had to merge each dataset. I used merge coding within Python to handle this, and broke it into four steps. I started by merging `purchased` with `players` so I could see all of the games purchased by each player. This helps to understand a player's full library, and this was stored as `player_purchase`. My next merge combined `achievements` with `history`, and this is where there is a lot of data drop off. There were four million rows after this, and once cleaning out the NAs after this merge, it dropped to two million rows. A lot of player IDs did not match, as the data is off. Some of this is due to how both systems store their player IDs differently, and in doing so, I matched accordingly by using the given platform and achievement IDs to connect the `history` of players with their actual achievements that they earned. The next merge that I had to do was to connect `achievement_history` with `games`. This allows me to see what players and given achievements are connected to all players and games alike. Once I cleaned the NAs for this section, only 300,000 rows were lost, showing that most of the data aligned with itself and did not have missing values.

The last merge combined `game_info` and `player_purchase`, forming the final master dataset. After cleaning out NA values, around 190,000 rows were removed. Removing the NAs between each step prevented the data from exploding, and there were significantly more rows than what there needed. This helped me reduce run times, along with not running unnecessary merges.

3.5 Geocoding

This was one of the trickiest areas for me to tackle outside of obtaining the data. With my data, due to privacy restrictions, I was only given the countries for each player. This caused many issues, as for my graph, I originally wanted to have a scatterplot to show the distribution within a given country. The issue with this was that the two best options were adjusting the jitter strength and assigning random city-state locations per unique player. For the jitter adjustment, even with trying to have a weighting scale, the plot was very sporadic and did not look clean. Two issues kept happening to where small countries like the Caribbean had a lot of their players in open water, or for countries the size of Russia, they would still be clustered in one spot. Since this did not work, I attempted to create a unique city-state combination for each player, but by running this, the run times for creating this were too long and provided unclean results. This method still yielded a plot that was clustered in one spot or looked off compared to my final map.

In my third attempt, I attached each unique country to a given longitude and latitude and switched to a heat map with countries being the identifier. This allowed for the locations to be cleaner and prevent colluded data, which in turn caused the map to look cluttered.

3.6 Aggregation

For my final step of preparing my data to make a visual, I needed to have a final aggregation. As Streamlit struggles as the data gets larger and larger, I needed to provide a safe ground for the world map. I filtered the data to be just the top 20 games for every country. This allowed for 20 lines per country, with the player counts summed together. This way, I am passing through the top portion, versus some countries having upwards of 8000 individual games. My goal is that as I expand further on this project that I scale this back to keep as much data as I can without crashing the application. In doing so, I will provide a larger picture of player counts per country, but with limitations, this was a safer option.

Once my final aggregation was done, I was left with my two datasets: `test_data` and `world_map_data`. These two tables were exported into .csv files to provide a clean upload for my visuals.

4 Results

For my results, I chose to create a dashboard with multiple infographics, one already being shown, the world map. (see Figure 2 for the Steam and PlayStation world maps). I also wanted a plot that you can see the top 10 games for a given country, along with data tables depicting some of the raw data.

4.1 World Map

When compiling the world maps, I decided to go the route of a heat map as previously stated. This allows us to see the impact within a given country of how many players are on a specific game per platform type. As the nature of PlayStation games differs from Steam, there is a clear distinction between the two. Some of the more popular games that we do not have data for are common across both platforms. Since we don't have that, we see some differences in where some games are strictly on Steam. This is due to the ability for small developer teams or solo developers to publish a game for sale. Some of the best Indie games, and games overall, are strictly on PC due to this nature. There are many regulations and hoops you have to jump through for both console types that Steam can get around.

To provide a more digestible view of the world map, we implemented interactive filters that allow users to explore different aspects of the data. A **global time slider** enables selection of a date range within the available year of data, highlighting seasonal peaks and troughs in gaming activity. In addition to this universal filter, world map-specific filters for `Genres` and `Game` allow users to adjust the visualization to examine platform-specific player distributions and game popularity across countries.

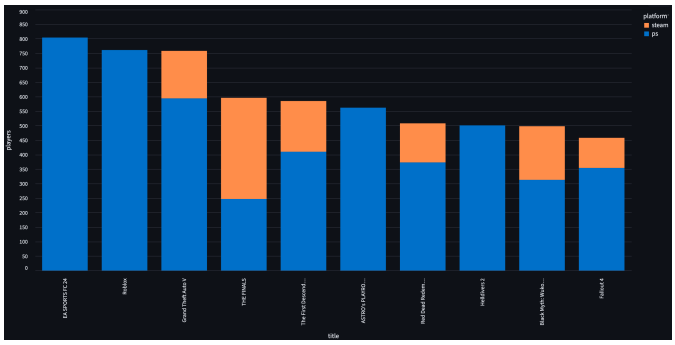

Fig. 3: Top 10 games globally based on player counts across platforms.

Global Game Totals

title	total_players
EA SPORTS FC 24	804
Roblox	761
Grand Theft Auto V	758
THE FINALS	596
The First Descendant	585
ASTRO's PLAYROOM	562
Red Dead Redemption 2	508
Helldivers 2	501
Black Myth: Wukong	498
Fallout 4	458

Fig. 4: Global Game Totals

Country-Level Player Breakdown

country	title	ps	steam	total_players
United States	Helldivers 2	191	0	191
United States	Roblox	149	0	149
United States	ASTRO BOT	143	0	143
United States	FINAL FANTASY VII REBIRTH	141	0	141
United States	THE FINALS	77	63	140
United States	ASTRO's PLAYROOM	135	0	135
United States	Grand Theft Auto V	119	16	135
United States	The First Descendant	106	23	129
Spain	EA SPORTS FC 24	128	0	128
United States	HELLDIVERS™ 2	0	113	113

Fig. 5: Country-Level Player Breakdown

4.2 Global Game popularity

Another visual I wanted to see is how gaming is impacted in the top echelon. (as seen in Figure 3) To explore the global popularity of games, we created a bar chart showing player counts for PlayStation and Steam separately, as well as the combined totals, highlighting the overall top 10 games. Adding filters helps see the breakdown even further, and in doing so, I have multiple filters available to select from. For the barplot, you are able to filter with `Countries`, `Platform`, and `Game`.

4.3 Data Tables

The next infographic that shows a strong correlation with the data is some of the raw data. Being able to see how a game breaks down from a global level helps show where trends are. This will also help decide demographic structures of target audiences. (as seen in Figure 4) You are able to see how the player count is for a given game. Along with this, you have filters for `Countries`, `Game`, `Platform`, and `Genres`.

Being able to see how a given country has a breakdown of players for not just a game, but the split between platforms is very useful. In turn, I created a second data table to depict the country-level data. (as seen in Figure 5)

4.4 Summary of Findings

Within the global map Figure 2, we can see heavy concentrations in the larger countries: USA, Russia, China, and Brazil. As these are the more heavily populated locations, it is to be expected that most of the players reside within these countries. A note for the bar plot Figure 3 is that the PlayStation platform takes over by a significant margin compared to Steam. This is a different view than what I originally anticipated, as Steam has more access to games, but also, their data was the one that was the most unclean. For both Figure 4 and Figure 5, we see that most of the player counts stem from the US for the top global games.

5 Discussion

5.1 Platform Distribution

We can see that, based on the global map (Figure 2), it shows that player activity is heavily populated in the more populous countries. This most likely is not just due to the population densities but also access to gaming as well. Steam, or PC, is a more expensive piece of equipment, which is sometimes double to triple the cost of a PlayStation at minimum. With this being the case, that is why a console system is more popular in most of the world. Along with this, it seems that PlayStation has a higher player engagement as well, which backs the notion of availability in gaming. That being said, the marketing strategy is most likely geared to where PlayStation knows they can reach a wider net of people, along with being able to be successful in their current space.

These trends also point to potential disparities in platform reach. Regions with limited console penetration or internet access may be underrepresented in the dataset, which could impact the generalizability of global insights. Future work could involve integrating additional datasets or platform sources to better capture regional differences and provide a more balanced understanding of player engagement.

5.2 Top Games and Genres

The top 10 games bar plot (Figure 3) reveals that PlayStation often surpasses Steam in total player counts for popular titles. This finding contrasts with the initial expectation that Steam, with its broader digital catalog, would dominate. Several factors may contribute to this outcome, including differences in user demographics, regional platform adoption, and data quality issues, as Steam's dataset contained more incomplete entries before cleaning. The interactive genre filters allow further exploration of how specific game types perform in different regions, indicating that certain genres may resonate more strongly with players on one platform compared to another.

These insights suggest that game developers should consider platform-specific preferences when planning releases or marketing campaigns. While globally popular games perform well across both platforms, niche or region-specific titles may see concentrated success on a single platform, emphasizing the utility of targeted strategies informed by data visualization.

5.3 Temporal Patterns

Using the date slider in the dashboard, seasonal trends in player activity become apparent. Peaks and troughs correspond to holidays, new game releases, or promotional events, underscoring the temporal dynamics inherent in gaming behavior. Understanding these patterns is critical for both researchers and industry stakeholders, as they provide context for variations in engagement and revenue generation. For example, game developers could align content updates or marketing campaigns with periods of heightened player activity to maximize impact.

Moreover, the temporal analysis highlights the importance of filtering data to avoid misleading conclusions. Aggregating across the entire year without considering temporal patterns could obscure key trends, such as platform-specific spikes or regional shifts in activity. Incorporating time as an analytical dimension enhances the dashboard's utility for strategic decision-making.

5.4 Implications and Usefulness

The combination of platform, genre, country, and date filters allows users to derive actionable insights from the dashboard. Stakeholders, including game developers and marketers, can identify emerging trends, popular titles, and regional player preferences. For example, understanding that PlayStation dominates in certain countries may inform

decisions about console-focused promotions or content localization. Similarly, the identification of globally popular titles can guide cross-platform marketing strategies.

Additionally, the dashboard demonstrates the value of interactive visualization for large, complex datasets. By enabling dynamic exploration, users can generate custom views that highlight specific patterns or anomalies. This flexibility not only supports research objectives but also facilitates data-driven decision-making in the gaming industry, reinforcing the importance of robust data cleaning, integration, and visualization methods.

6 Conclusion

This study presented an interactive dashboard to explore global gaming trends using data from Steam and PlayStation platforms. By aggregating and cleaning large-scale datasets, the dashboard provides visualizations including world maps, top 10 game bar plots, and detailed data tables, allowing users to investigate player distributions across countries, platforms, and genres.

The results highlight that player activity is concentrated in highly populated countries such as the USA, Russia, China, and Brazil, with PlayStation generally surpassing Steam in total engagement for popular games. The use of interactive filters, including date, platform, genre, and country, enables more granular exploration of regional, temporal, and platform-specific patterns, revealing insights into game popularity and player preferences.

The dashboard demonstrates the value of combining robust data cleaning, aggregation, and interactive visualization for understanding complex datasets. Future work could incorporate additional platforms, longer time spans, or demographic information to provide a more comprehensive view of global gaming behavior, supporting data-driven strategies for developers, marketers, and researchers.

Figure Credits

Fig. 1: Adapted from VANAS Team, "Top 10 Countries That Play the Most Video Games," VANAS Blog, 2025.

Acknowledgments

I wish to thank Artyom Kruglov for access to the dataset. Without this, this project would not be possible.

References

[1] A. Kruglov, "Gaming Profiles 2025 (Steam, PlayStation, Xbox)," https://www.kaggle.com/datasets/artyomkruglov/gaming-profiles-2025-steam-playstation-xbox, 2025, accessed: 2025-11-20. 1