**Cover Page**

**Project Title: Steam Games: GTA V Analysis**  
**Group Number: 3**  
**Submission Date: 22 December 2023**

**Group Members: Kindyl Alfonso, Dylan Burbank, Samuel Volpone, Yujie Zheng**  
**Signatures of Group Members:**

**Kindyl, Dylan, Sam, Yujie**

**Table of Contents**

1. Executive summary
2. Introduction
3. Dataset Selection
4. Data Management
5. Data Exploration and Analysis
6. Text Analysis
7. Statistical Modeling
8. Data Visualization
9. Conclusion
10. References
11. Appendix

**Executive Summary**

As a multi-billion-dollar industry, video games have the potential to significantly impact markets, generations, and have a large cultural influence in the present. Analyzing the trends in this industry can lead to a significant amount of influence amongst game developers and marketers alike. This report focuses on the highly successful game, Grand Theft Auto V (GTA V), and games of similar success, Player Unknown’s Battlegrounds and Rocket League. By analyzing these three games, we can provide insights into the trends that successful games follow. By utilizing a dataset with information on over 40,000 Steam games, we aim to pinpoint patterns that lead to the success of the games mentioned.

Using well-known processes of data cleaning and management, we were able to successfully wrangle and clean the dataset for the desired intention of exploratory data analysis and partial statistical analysis. The data cleansing process took place via Power Query and Excel.

We utilized software such as Tableau, R Studio, and PowerBI to provide exploratory data analysis and statistical analysis. The analysis across GTA V, PUBG, and Rocket League revealed that while positive sentiment dominated, it was game mechanics that influenced player emotions. GTA V and PUBG, with violent gameplay, lead to more negative sentiments. Rocket League, a car-soccer game for kids, received overwhelmingly positive sentiments. These insights show the impact of gameplay on player reviews and game success within different genres.

Understanding the relationship between game mechanics, player sentiments, and success paves the way for informed decisions in all areas of game development. Potential future research aims at predictive modeling to offer more thought-out recommendations for game developers and further our understanding of the nuances of gaming preferences.

**Introduction**

For our final project, we decided to analyze the Steam games dataset. All members of our group have grown up with experiences playing video games, and we thought doing a presentation that focused on video game data would help engage classmates and add a connection for those who had grown up gaming. The Steam dataset was filled with over 40,000 games, so we decided it would be best to narrow our focus to a specific game: Grand Theft Auto V. We chose GTA V because of its widespread popularity and the upcoming highly anticipated release of GTA VI.

Our data management process began with cleaning the data. Although we did not have very many columns to deal with in our data, this still became a very time-consuming and irritating process. We tried to create dummy variables for some columns which made our data too large to work with, which slowed down our working process. We decided to analyze the game GTA V because of its popularity and the ability to relate with others. We decided to analyze the game against other popular games to see how GTA V performed against other top games. Finally, we created visuals from these analyses to present to our classmates and to help them visualize how Grand Theft Auto V performed against other games and conducted a text and sentiment analysis on GTA to visualize user feedback of the game.

The goal for our project was to present to our viewers a relatable analysis of a game that many of us grew up playing, and relate that analysis to other games that people may have a mutual interest in. We did this through using visuals from Tableau and PowerBI to show our classmates how these games performed in categories such as price and review score. We also made visuals from our text and sentiment analysis to show our classmates how Grand Theft Auto V was rated by users who had played the game.

**Dataset Selection**

Dataset Choice

The dataset that we chose to use for this analysis was found on Kaggle and is called Steam games complete dataset. The dataset can be found here <https://www.kaggle.com/datasets/trolukovich/steam-games-complete-dataset/>. We chose this dataset because although we are a group of people with different hobbies and backgrounds, we shared a mutual interest in video games and thought this would be a relatable interest to our classmates. This data set gave us plenty of data to analyze, over 40,000 games, which made it easy to create a scope for our project and create analyses and visuals that compared different games. Here is a screenshot confirming our selection in our classes’ discussion board:

A screenshot of a chat

Description automatically generated

We made sure to choose our dataset early, as we were aware that there was a penalty for duplicating another group's dataset, so we made our choice as early in the term as we could.

**Data Management**

**Process Detailing**

**Data Cleaning:**

The original dataset retrieved from Kaggle had plenty of cleaning needs. The first issue was the dataset had no primary key. To create a primary key, the URL column contained numbers within each instance that would correspond to a proper primary key. We extracted the numerical portion and double-checked that each instance was unique. There were also three types of game types: regular, bundle and sub-versions. We decided to only focus on the regular versions of the game for consistency and ease during analysis. Therefore, instances other than regular were removed during the cleaning process. The dataset also contained many columns with cells that contained multiple pieces of information. These columns included: languages, popular tag, game details, minimum technology requirements for gameplay, recommended technology requirements for gameplay, and genre. These columns were split into new columns by the comma delimiter and then factored as dummy variables. The release date column was separated into columns for day, month, year to account for future analysis. Game description and mature content description included text in the beginning for all values so that portion of the string was removed as well.

The more difficult columns were recent reviews, all reviews and mature content. The two review columns included information on the sentimental nature of reviews (mostly positive, very positive, mostly negative etc.), the number of reviews, and the percentage of positive reviews. Splitting these columns proved to be difficult since splitting by delimiter would not work as there were commas for the number of reviews. To account for this, certain parts of the cells were selected via left and right functions in Excel. Also, using the search function to extract the number within the parentheses proved to be helpful. After meticulously extracting these bits of information from the columns, the review column was able to be formed into the columns of sentiment, number of reviews and number of positive reviews (formed from the multiplication of percentage of positive reviews and number of reviews), and percentage of positive reviews.

The next column that was difficult to clean was mature content description. This column included a general description of the type of mature content but was not standardized throughout. To approach a standardized dummy variable column, there was a need to be a method to extract which type of mature content the game contained. To do this, it was necessary to compare to the content to the Entertainment Software Review Board (ESRB) content descriptors, which is the organization that reviews mature content in games for the United States and assigns them levels of appropriateness for age-groups. These content descriptors included substances, blood/gore, humor, violence, gambling, language, nudity, and sexuality (cite). There were key words in the mature content descriptor that would be able to account the related content descriptors. Since there were multiple terms for each content descriptor, there needed to be a large amount of code accounting for each possible phrase used to describe these ratings. Once these phrases were extracted, separated by columns, into a new column, then the process of splitting and creating a dummy variable for each content descriptor was uncomplicated.

For the dataset containing information on reviews, the comments section was cleaned for analysis on sentiment. Other parts of the dataset were not deemed necessary to clean.

**Data Preparation:**

The formatting of the columns was relatively easy as the structure was not compromised when we obtained the original dataset. The only process for formatting we used was checking to make sure the datatypes were inputted as expected.

**Transformation and Integration**

**Data Migration Techniques:**

The data was already in a normalized format before entering the data into PowerBI and Tableau so there was nothing needed to prepare it further than the cleaning format.

**Data Integration Techniques:**

To connect the two datasets used, the datasets were joined on the numerical primary key. The reviews dataset did not have the same primary key, so it was necessary to find and search for the game title in the games dataset and extract the primary key and place it into the review's dataset.

**Rationale for Techniques Used:**

Every technique utilized by the group was to ensure the integrity of our dataset. Making sure the data was clear, concise, and structured was the main priority during the data cleaning process. Splitting multi-valued columns, standardizing columns, and creating a primary key were the main techniques used to ensure the cohesion of our dataset.

**Data Exploration and Analysis**

After the initial cleaning of the data, we started our first analysis of the datasets. One of the first things we found was when most games were released by day of week. It turns out that most games are released around midweek to drum up the most hype for them, as people often have plans on the weekends. Most of the games were released on Thursday. We then found the average price of all games on steam, as well as the average review score. The average review score was around .77 with 1 being all positive reviews. Of the three games we looked at, only PUBG had a review score lower, and it was much lower at .49. We also saw that the average price per game was $15, and our focal games were priced above $30 for GTA and PUBG, and $20 for Rocket League. The more popular games were all above this $15 average.

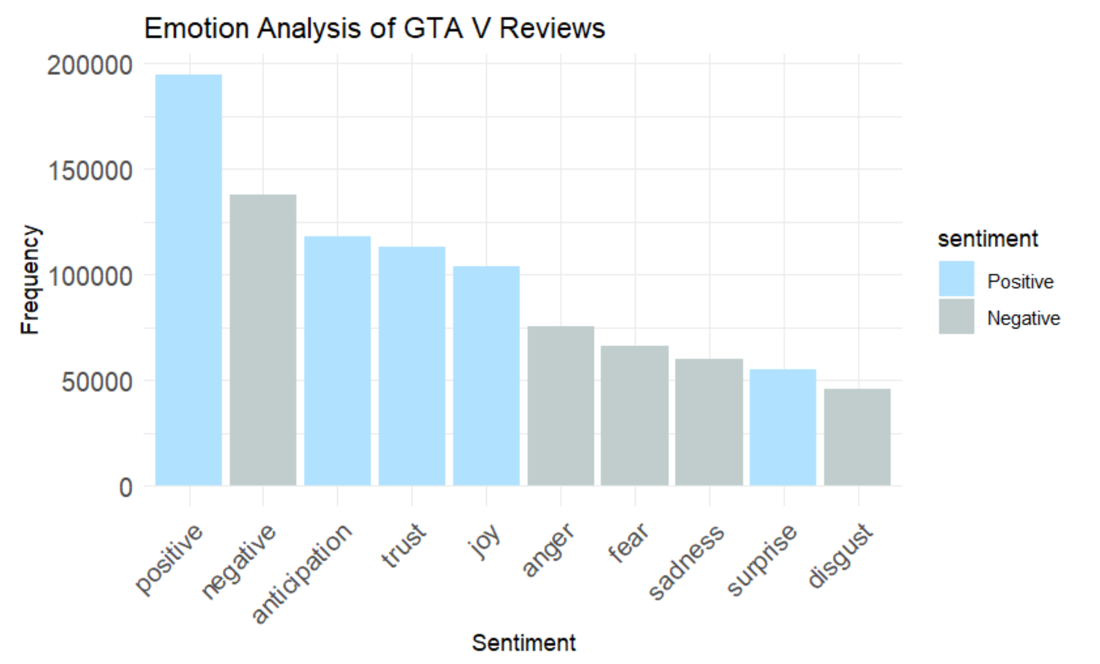
**Text Analysis**

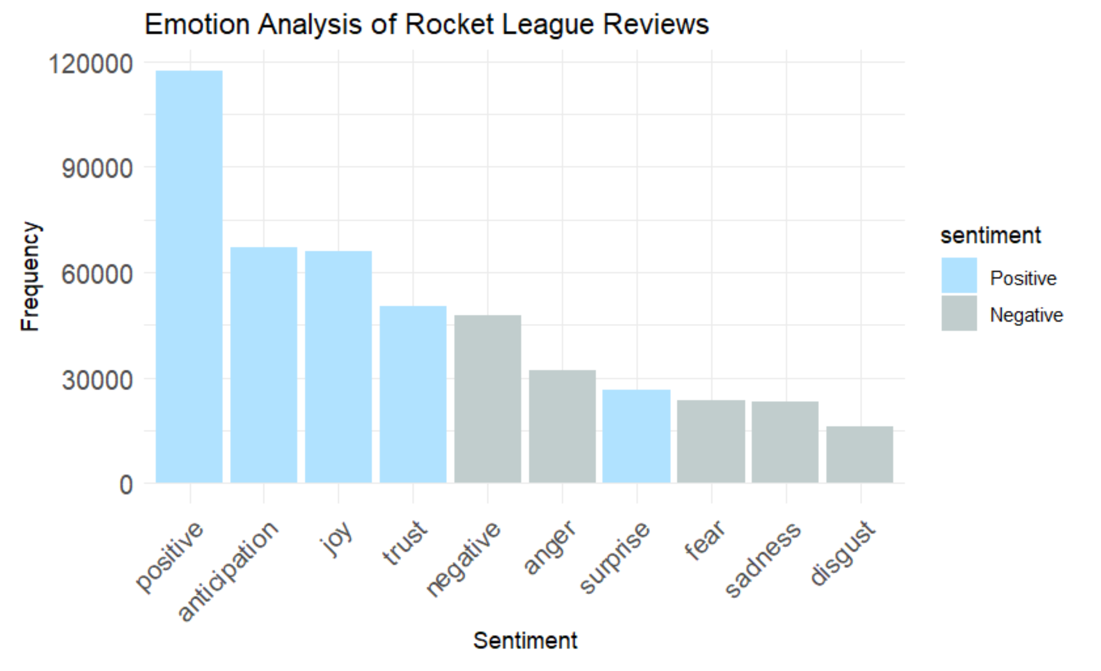
Text analysis proved invaluable with the Steam reviews dataset, complemented by a supplemental Kaggle dataset comprising individual player reviews for about 20 games on Steam. Our focus narrowed to three key games: Grand Theft Auto V, Rocket League, and PlayerUnknown Battlegrounds. To harness this data for text analysis, we structured three distinct data frames, each derived by filtering the original dataset to include entries solely aligned with one of these game titles.

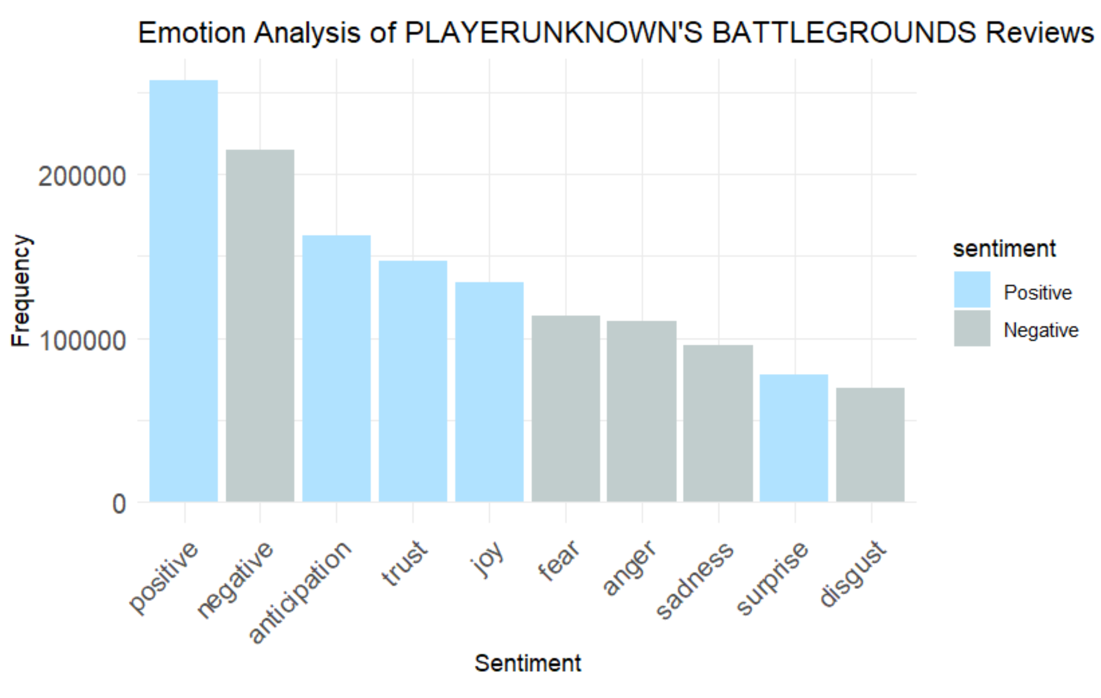
In this analysis we used libraries such as ‘tm,’ ‘wordcloud,’ and ‘rcolorbrewer,’ we transformed the reviews column for each game into captivating word clouds. Taking Grand Theft Auto V as an example, we began by creating a corpus from the review column. Standardizing the text was crucial to eliminate redundancies. Lowercasing, eliminating punctuation and numeric characters, and stripping out stop words those devoid of meaningful insight like 'the,' 'and,' and 'like'—primed the data for word cloud generation. After an initial output, recognizing recurring words like ‘game,’ ‘gta’, and ‘just,’ we further refined our stop words list to distill more relevant insights from the GTA V reviews. This process was repeated for Rocket League and PlayerUnknown Battlegrounds, resulting in three visually compelling word clouds, each color-coordinated to highlight the top words within their respective game reviews.

Our sentiment analysis mirrored this approach. Utilizing the three separate data frames for each game, we commenced with GTA V by isolating the review column into a new data frame. Employing `unnest\_tokens()` dissected the reviews into individual words for further analysis. Instead of a traditional sentiment analysis with the ‘bing’ lexicon, we delved into both emotions and sentiments. Employing the ‘nrc’ lexicon which maps over 14,000 words to eight distinct emotions (anticipation, trust, joy, anger, fear, sadness, surprise, and disgust) alongside positive and negative sentiments, we got the word count associated with each emotion and sentiment. After getting the word count for each emotion and sentiment, we represented the data on a graph. We chose bar charts and had ten columns showing eight emotions and two sentiments. The bars were color coded to positive and negative sentiment as well.

The results for our emotion/analysis are shown below:







For each game positive sentiment was higher than negative sentiment which is a great for the game, especially in the context of review data. But Rocket League stood out in terms of its positiveness. In both PUBG and GTA, the top two bars were positive sentiment and negative sentiment, respectively. For Rocket league we see a massive bar relative to each other emotion on the x axis. Initially, we thought this meant that the game has more positive reviews. While this could be true it is important to understand the core gameplay in each of these games. For GTA and PUBG, the gameplay is violent. These games are rated for 17+ and 16+ year olds, respectively. In the case of Rocket League, the official rating is 8+ years old. So, the reviewers describing their experience with the game might be inherently more negative based on the word association within the ‘nrc’ lexicon. Rocket League is a game like soccer; expect you control a rocket-powered car. This might lend itself to not receiving the same amount of negative sentiment as seen in the other two games. Overall, with these three successful games, it's important to see the highest occurring emotions within each.

**Statistical Modeling**

**Model Application**

First, in our project, we analyzed the number of games released by week using descriptive statistics and data visualization with the aim of identifying the days when games were released most frequently. Second, we used sentiment analysis and text mining techniques, focusing on identifying the most frequent words in the game reviews of GTA and comparing them to the reviews of PUBG and Rocket League. In addition, we analyzed the sentiments expressed in these reviews to understand players' emotions. Our study also used correlation analysis to examine the relationship between various variables, including the price of the game, the percentage of positive reviews, the number of descriptions of the game, and the presence of themes such as sex or violence in the positive reviews. This multifaceted approach allowed us to cover a wide range of analytical objectives from understanding market trends to player psychology.

**Assumptions Made**

In our statistical analysis, we've made several key assumptions and formulated hypotheses to guide our exploration of game data, particularly focusing on how Grand Theft Auto (GTA) compares to other games, the emotions expressed in successful games' reviews, and the relationship between various game metrics.

For comparing GTA to other games, our null hypothesis posits no significant difference between GTA and other games in the dataset across chosen metrics, while the alternative hypothesis suggests significant disparities, perhaps in aspects like reviews or pricing. When examining the emotions in successful games' reviews, we assume that if there is a distinct emotional pattern, it should be significantly different from a neutral or random distribution of emotions, with the alternative hypothesis expecting a higher prevalence of specific emotions, such as joy in positive reviews.

In employing visual tools like bar charts for comparing three games by price and positive reviews, we assume these metrics are representative and revealing of broader game performance and player satisfaction. Similarly, our word cloud analysis presupposes that the frequency of words in reviews mirrors the most critical and recurrent themes or sentiments. For the emotion analysis of three specific games' reviews, we hypothesize that each game might elicit a distinctly different emotional response, which should be evident in the comparative review analysis.

Lastly, our correlation analysis includes the null hypothesis of no significant correlation between studied variables like game prices and positive reviews against the alternative of a meaningful relationship. These hypotheses and assumptions are foundational in guiding our statistical inquiry, shaping our methods, and interpreting the data to glean insights about game trends, player perceptions, and market dynamics. Each is carefully considered to ensure a robust and credible analysis, providing a structured approach to understanding the complex dynamics of game success and player engagement.

**Model Results**

Our statistical analysis, underpinned by the previously stated assumptions and hypotheses, has yielded several notable findings across different aspects of game release patterns, text analysis, emotion analysis, and correlation analysis.

Firstly, in our examination of game release patterns using descriptive statistics, we found that games are most frequently released on Thursdays. This finding rejects our null hypothesis, which assumed no significant difference in the number of games released on any day of the week and suggests a strategic preference for Thursdays in the gaming industry.

In our text analysis comparing GTA with the other two games, the results indicated no significant differences in the textual content of the reviews. This aligns with our null hypothesis, suggesting that, at least in terms of the frequency of words, GTA does not stand out distinctly from other games in the dataset.

However, the emotion analysis provided more nuanced insights. While GTA showed similar levels of positive and anticipation words compared to the other two games, it notably had fewer negative words and a higher presence of trust and joy words. This indicates a generally more positive and trustful emotional response to GTA, differing from our null hypothesis, which suggested no significant emotional difference across games. The result underscores GTA's unique position in eliciting a more positive and joyous emotional reaction from players.

Lastly, our correlation analysis, which looked at the relationship between game prices, positive reviews, and game description length, found no significant connections. This outcome supports our null hypothesis of no correlation between these variables, suggesting that these factors might independently affect game success or player perception without a discernible linear relationship.

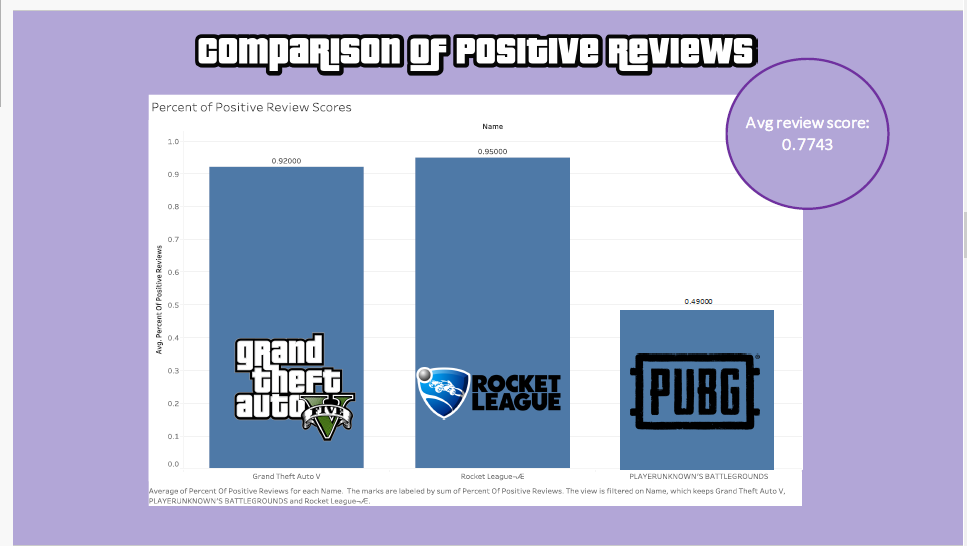
**Implications of Model Results**

These results, taken together, provide a comprehensive view of the gaming landscape, particularly highlighting how specific days might be favored for releases, how GTA maintains a unique emotional profile among players despite similar textual patterns, and the complexities of correlating game features with player reception. Our findings not only respond to the initial assumptions and hypotheses but also offer a deeper understanding of the nuanced dynamics at play in game releases and reception. This analysis serves as a valuable base for further investigation and strategic decision-making in the gaming industry.

**Data Visualization**

We used Excel, Power Query to clean the data, and used Power BI, Tableau to do the visualization.

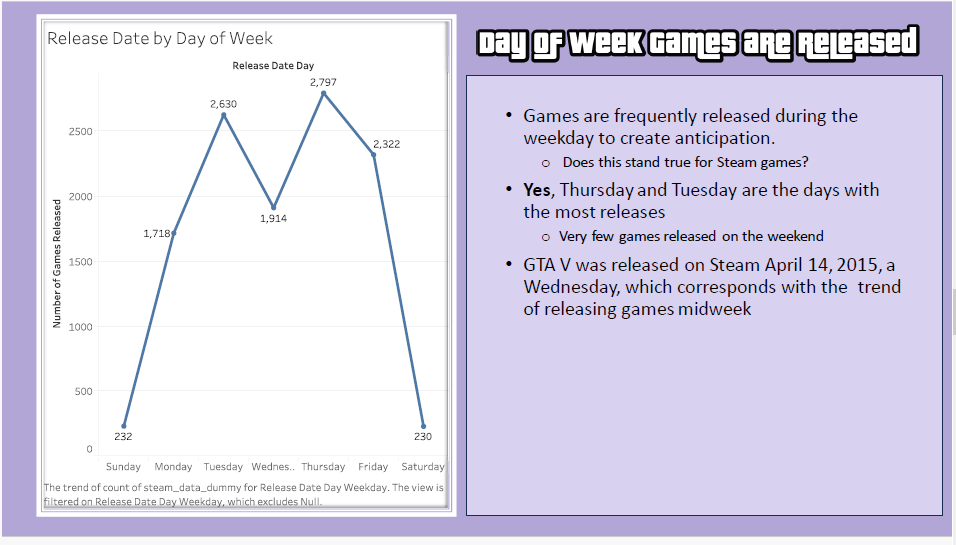
**Description of Visualizations**



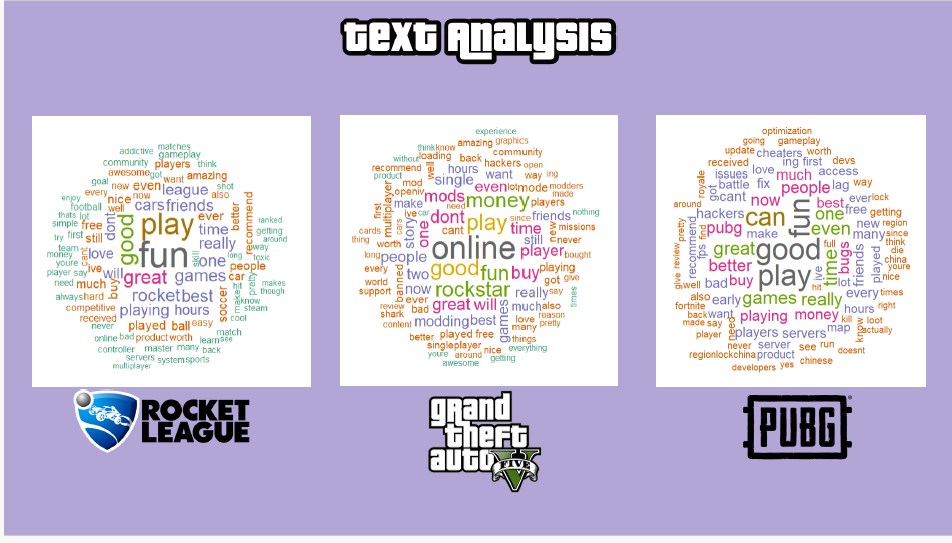
The bar chart visualization comparing positive review scores for 'Grand Theft Auto V' (GTA V), 'Rocket League', and 'PLAYERUNKNOWN'S BATTLEGROUNDS' (PUBG) depicts a clear disparity in player satisfaction across these games. GTA V boasts a high positive review score of 92%, which is impressive, yet slightly lower than Rocket League's 95%, indicating a very positive reception among players for both games. PUBG, on the other hand, shows a significantly lower score of 49%, suggesting mixed reactions from its player base. The average review score across these games stands at 77.43%, reflecting a generally positive trend but also highlighting the variance in player experience and perception among these popular titles.



The bar chart presenting the price comparison between GTA V, PUBG, and Rocket League reveals the differing pricing strategies and perceived value of these games. GTA V and PUBG are priced equally at $29.99, suggesting a similar market positioning in terms of cost. 'Rocket League', however, is offered at a lower price point of $19.99, which might reflect different content volume, development costs, or strategic pricing decisions aimed at a broader audience.



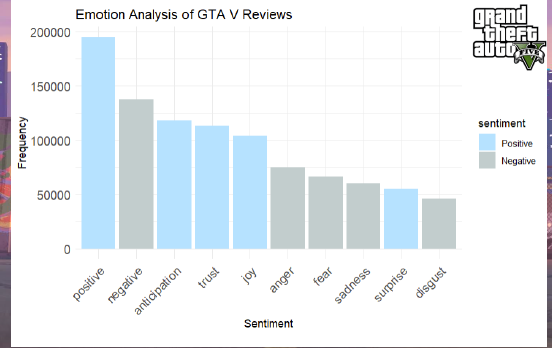
The line graph depicting the release date of games by day of the week clearly illustrates that Thursday is the peak day for game releases with 2,797 games, followed closely by Tuesday with 2,630 releases. Friday also sees a substantial number of releases at 2,322. The beginning of the week, particularly Monday, sees a moderate number of releases at 1,718, while the weekend has the fewest, with Sunday at 232 and Saturday at 230, reinforcing the trend that game releases are predominantly a weekday event.



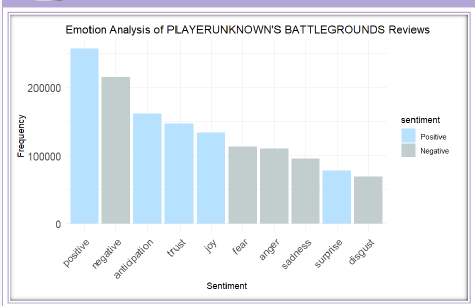
For 'Rocket League', the terms "fun," "play," "good," "great," and "games" are prominent, suggesting a positive and enjoyable gaming experience. The emphasis on "fun" and "play" indicates that the core gameplay loop is well-received and engaging.

In the case of GTA V, "online," "play," "good," "fun," and "money" are the most frequent words. This points to the significant role that the online component and the in-game economy play in the players' experience. The presence of "money" could refer to the game's virtual economy or possibly the cost of the game and in-game transactions.

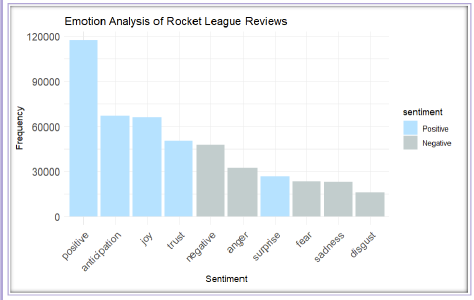
For PUBG, the words "good," "play," "fun," "can," and "games" are most often mentioned. Similar to the other games, the focus is on the quality of gameplay and enjoyment. The word "can" might be associated with players' abilities or options within the game, such as "can play," "can win," or "can do" something, indicating a sense of possibility and empowerment.



The sentiment/emotion analysis bar chart for 'Grand Theft Auto V' (GTA V) reviews quantifies the frequency of various emotions expressed by the players. The chart shows a dominant frequency of positive words at around 200,000 occurrences, underscoring the overall favorable reception of the game. Negative sentiments are also significant but lower, with around 140,000 mentions, indicating that while there are aspects of the game that players critique, these do not overshadow the positive experiences. The emotions of anticipation and trust are expressed with approximately 120,000 and 100,000 mentions respectively, suggesting that players are engaged with the game's content and have confidence in the brand. Joy is reflected with about 90,000 mentions, reinforcing the game's ability to deliver enjoyable experiences to its players.



PUBG has positive sentiments dominate with 230,000 mentions, closely followed by negative sentiments at 200,000, suggesting a highly engaged yet divided player response. Anticipation and trust are also prominent at 170,000 and 160,000 mentions respectively, highlighting the game's ability to generate excitement and player loyalty. Joy rounds out the sentiment profile with 120,000 mentions.



'Rocket League', on the other hand, presents a more positive sentiment landscape, with 120,000 mentions of positivity, significantly outweighing the negative sentiments at 39,000. The game also evokes a strong sense of anticipation and joy, both recorded at 60,000 mentions, followed by trust with 40,000 mentions, reflecting a consistent and upbeat player experience.

**Adherence to Best Practices**

In crafting the visualizations for this project, we have adhered to best practices in design and clarity to ensure that the data is communicated effectively. The charts are characterized by their simplicity and focus, using a single color per sentiment category to allow for immediate visual comparison without distraction. Consistency in color coding—blue for positive and grey for negative sentiments—across all visual aids in quick and unambiguous interpretation. The legibility of each chart is maintained with clear, well-sized text and labels, making the data easily readable, and specific figures on the bars provide precise information briefly. The selection of chart types is deliberate, with bar charts for sentiment frequencies and line graphs for release date trends, each chosen for their ability to best represent the respective datasets. Finally, the charts include informative titles and labels, along with helpful annotations where necessary, to offer context and enhance understanding, ensuring that each visualization is as informative as it is accurate.

**Visualization Interpretation**

**Meaning of Visualizations**

The visualization showcases rich data that speaks volumes about player engagement and game launch strategy. From sentiment analysis, we learned that GTA games promote predominantly positive emotional responses, as can be seen from the high frequency of positive terms in player reviews. In contrast, the occurrence of negative emotions, although less frequent, suggests potential to benefit from further development or community involvement. The release pattern visualization highlights strategic preferences for mid-week, with Thursday being a prime time for new game releases, in line with industry trends and likely consumer behavior.

**Relevance to Objectives**

These insights are directly related to the goals of our project, which are to understand player sentiment, determine optimal game release times, and benchmark industry practices. Sentiment analysis directly tells us about the emotional impact of a game and highlights the perceived strengths and weaknesses of a player base. At the same time, release date analysis provides strategic insights into gaming industry timing strategies, providing valuable information for publishers and developers looking to maximize the impact of their game launches. These visualization effects jointly enhance our understanding of the GTA V game and lay a certain foundation for our future analysis of GTA 6.

**Conclusion**

Across the three games, Grand Theft Auto V (GTA), PlayerUnknowns Battlegrounds (PUBG), and Rocket League, positive sentiment dominated over negative sentiment. However, Rocket League notably stood out for its overwhelmingly positive sentiment compared to the other emotions, in stark contrast to GTA and PUBG, where positive sentiment was the top bar, followed by negative sentiment. The stark contrast could be attributed to the core gameplay differences: violent gameplay in GTA and PUBG versus the non-violent, sports-oriented gameplay of Rocket League.

These findings hold significant implications within the gaming landscape. They highlight how gameplay dynamics, and the nature of gaming experiences potentially influence sentiment portrayed in reviews. The inherent violent nature of GTA and PUBG likely triggers more negative sentiments, whereas Rocket League, with its non-violent, soccer-like gameplay, appears to elicit a significantly more positive sentiment overall.

Understanding the interplay between game genres, gameplay mechanics, and the emotional responses reflected in player reviews not only aids in comprehending consumer sentiments but also sheds light on the factors contributing to the success and reception of these games within their respective genres. Such insights can be pivotal for game developers, marketers, and industry analysts aiming to create, promote, or understand games' reception among diverse audiences.

As for future research, a predictive model would aim towards stronger recommendations for game developers. Through these models and regression analyses, further insights into the strengths of certain variables would lead to more insightful recommendations beyond the scope of analyses in this report. The findings from our analyses are far from concrete, and a more statistical approach would allow for a credible interpretation of the dataset.

“ESRB Ratings Guides, Categories, Content Descriptors.” ESRB Ratings, 6 Feb. 2023, [www.esrb.org/ratings-guide/#cont\_desc](http://www.esrb.org/ratings-guide/#cont_desc).

“Games Industry Data and Analysis.” *Video Game Insights*, vginsights.com/game/271590. Accessed 22 Dec. 2023.

**Appendix**

****