# ONLINE GAMING BEHAVIOR

*Explore Player Engagement Patterns in Online Gaming*

**PROF. DELINA IVANOVA**
Lecturer, Schulich School of Business - York University
Director of Analytics, Mistplay

**Prepared By:**

**Group 8**

Dazhen Yu (218078444)
Mansi Patel (221688007)
Prerana Bhattarai (221733779)
Priya Chaudhuri (221060736)
Wania Syed (221149380)

July 27th, 2024

# TABLE OF CONTENT



# LIST OF TABLES



# LIST OF FIGURES

# Chapter I: Introduction

## 1.1 Introduction to Online Gaming Behaviour

In the dynamic and rapidly evolving landscape of online gaming, deep understanding of the behaviour of players and their engagement is crucial for game developers, marketers, and researchers. This dataset offers a comprehensive collection of metrics and demographics related to player activities in various gaming situations. The dataset we have selected deals with several such features that we will leverage to study player behaviour in a different way.

## 1.2 Challenge Statement

We aim to find out key factors influencing the retention of players in online gaming environments and to develop a model that predicts whether a player falls into one of the categories such as high, medium or low in terms of engagements which will help us to place all our focus on the right player to maximize high retention and in-app purchases to optimize revenue and costs.

We will also look at what factors including the game design and demographics have the most impact on players' behaviour and engagement levels.

## 1.3 Objective

The gaming industry is highly competitive, with players exhibiting diverse preferences and behaviours across various game genres. This dataset includes detailed information on player demographics, game preferences, playtime, session frequency, and in-game purchases. By analyzing this data, we aim to identify actionable insights that can enhance player retention, engagement, and monetization. We aim to use the results of our analysis to give recommendations to game developers to increase revenue across all different genres of games and enhance engagement.

i. Players with greater play times make more in-app purchases and are more likely to have higher engagement levels
ii. Various factors such as player demographics, game-specific details, and engagement metrics significantly influence player behaviour and game performance.
iii. How does the average session duration vary across different game genres, and what does this imply for player engagement strategies?
iv. Are there specific regions or demographics where certain game genres are more popular, and how can targeted marketing strategies be developed based on these insights?

## 1.4 Dataset

This dataset that we have worked on contains exciting information on gaming behaviour, player demographics, and engagement metrics. It consists of comprehensive columns pertain to player behaviour in online gaming environments. Many variables such as GameGenre, InGamePurchases and Playtime Hours are crucial metrics in determining player behaviour across all kinds of games. Whereas the target variable – Engagement Level, aptly represents player attrition rates.

The extensive dataset spans to over 40034 entries, it thoroughly captures all aspect required to study player behaviour and gaming environment. The summary of all columns are as follows:

a. **Numerical Columns:** Age, Playtime Hours, InGamePurchases, SessionsPerWeek, AvgSessionDurationMinutes, PlayerLevel, Achievements Unlocked

b. **Categorical Columns:** Gender, Location, GameGenre, Game Difficulty, Engagement Level

Player demographics are captured by columns like, Age, Gender and Location. Game details can be found in GameGenre and GameDifficulty columns. Whereas engagements metrics are comprehensively captured by PlaytimeHours, SessionsPerWeek, AvgSessionDurationMinutes, PlayerLevel, AchievementsUnlocked and EngagementLevel. Lastly, monetization can be found in InGamePurchases to see when and how much players spend while playing the games.

# Chapter II: Exploratory Data Analysis

## 2.1 Data Cleaning

The first important step in the EDA is to study data and explore in depth. The primary focus was to handle invalid observations and treat missing values appropriately.

Firstly, we checked the head() function to see what columns existed in our data to give a clear picture of what we are dealing with and what metrics can we use for our analysis. Next, we applied the info() function, which gave us a complete description of each column in the Data Frame, its count values and the type it belongs to. It effectively summarises the dataset for us, so that we do not make any mistakes in the analysis and treat data according to their correct types. For example, if there was a date in the data, we would convert it into date time format before working with it.

Using info() we found that we have three datatypes, namely, integer, float and object. The categorical columns like Gender, Location, GameGenre, GameDifficulty and EngagementLevels are objects. There were a total of 13 columns and 40034 rows.

## 2.2 Exploring Data

We used the describe() function to generate quick overview of the numerical columns in the Data Frame. This indicates the count (total number) in each column, the mean, standard deviation, minimum, maximum and the quartiles of each column. We found out there were no missing values in the dataset hence there was no need to handle missing values with filling in techniques.

Next, we changed the values of InGamePurchases from binary output (1 and 0) to ‘Yes’ or ‘No’. This allowed better readability and understanding for the users.

The output [False 40034] indicated that when checking for duplicates using the duplicated() method, no duplicate rows were found in the dataset. This means all 40,034 rows in the dataset are unique. Since there are no duplicate rows, we proceeded with the exploratory data analysis (EDA) without needing to remove duplicates.

## 2.3 Visualising Data

### 2.3.1 Age and Gender

We created a pie chart to show the distribution of genders in our data and the results show that it is mostly equally distributed with slight male domination as seen in the picture below. The chart segments and percentages allow us to quickly see the proportion of male and female players.

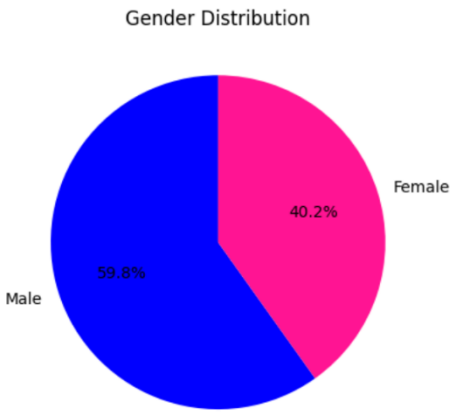


***Figure 2. 1: Gender Distribution***

We also plotted a boxplot of playtime in hours by age and gender as shown below

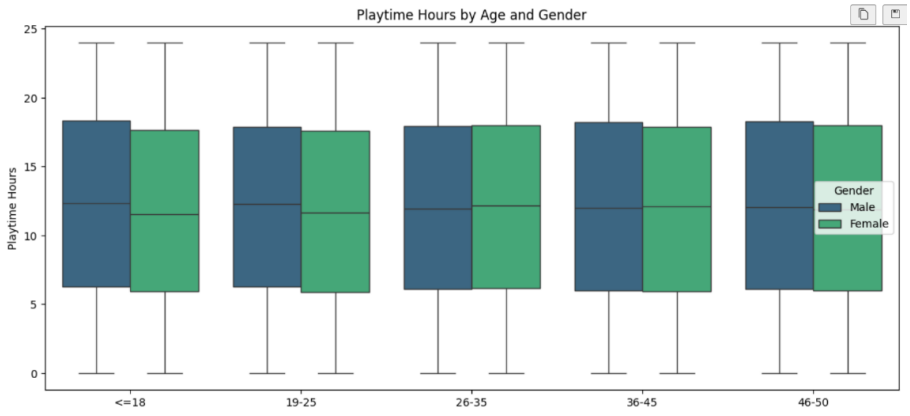


***Figure 2. 2: Playtime Hours by Age and Gender***

This plot focuses on understanding the median playtime in hours per week across different age groups and comparing these metrics between male and female players.

Median Playtime Hours by Age Group:

- Across various age groups, the median playtime typically ranged between 10 to 12 hours per week.
- This indicates that regardless of the age group, players generally spend a similar amount of time playing games weekly.

Comparison of Playtime Hours by Gender:

- When comparing male and female players, there were no significant differences in playtime hours across different age groups.
- This suggests that both male and female players were equally engaged in terms of the time they dedicate to gaming each week.

We further did similar boxplots for ‘Average Session Duration’ by age and gender and got the following results.

1. For most age groups, the median session duration was approximately 90 to 110 minutes.
2. This consistency indicates that regardless of age, players tend to have similar session lengths.
3. When comparing session durations between male and female players, the variability is similar across different age groups.
4. Players of all ages show similar engagement patterns with the game, as evidenced by the consistent session duration

In addition to this, a boxplot showing sessions per week between age and gender showed that the median sessions per week for most age groups hover around 8-10 sessions. In the 19-25 age group, female players exhibited a higher mean session duration compared to male players. This indicates that within this age group, female players tend to spend more time per gaming session than their male counterparts. Moreover, both male and female players in the 46-50 age group show slightly higher mean sessions per week compared to other age groups. This suggests that players within this age range were more engaged, having more frequent gaming sessions per week.

The bar chart and boxplot we plotted of with Genders and Age respectively gives a clear insight on what ages preferred what genres and people of what age groups were inclined towards which games.

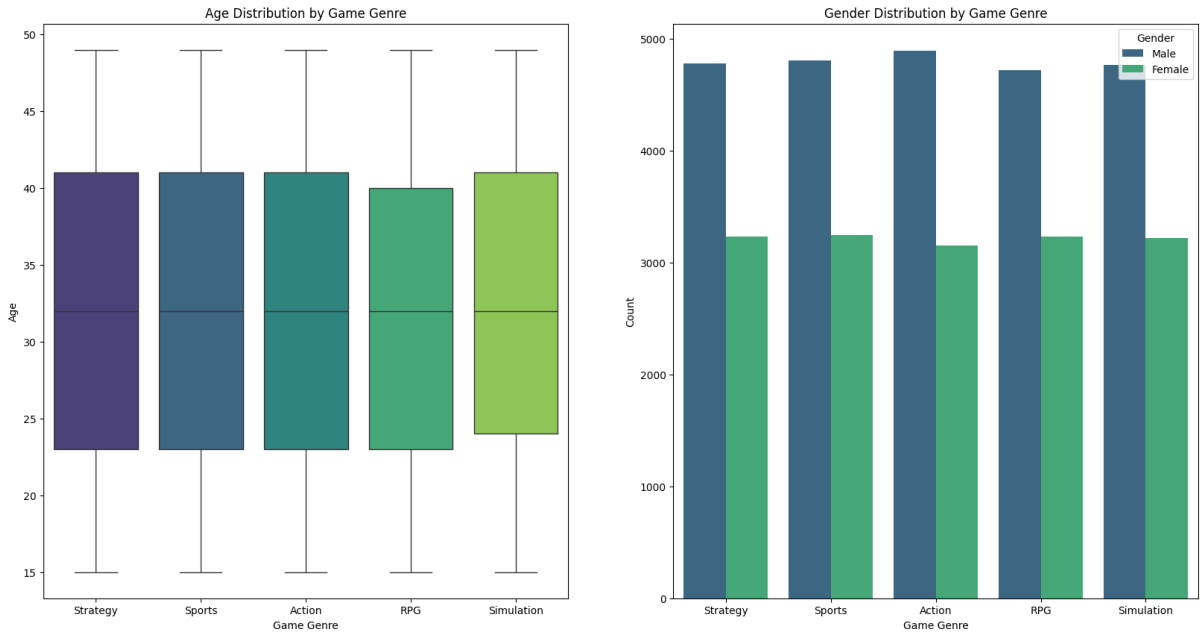


*Figure 2. 3: Age and Gender Distribution by Game Genre*

These plots as well as a count we did of the same columns aims to understand the demographic trends in game genre preferences, focusing on age and gender distributions across different game genres. Our findings show that:

**With Age:**

**RPG and Strategy Games**: These genres tend to attract slightly older players. The average age of players engaging in RPG (Role-Playing Games) and Strategy games is higher compared to other genres.

**Sports Games**: Players of Sports games generally belong to a younger demographic, indicating a preference among younger players.

**With Gender:**

**Sports and Action Games**: These genres have a higher proportion of male players. The majority of players in these categories are male, indicating a strong preference among male gamers.

**RPG and Strategy Games**: These genres show a more balanced gender distribution, with both male and female players engaging in these games in roughly equal proportions.

### 2.3.2 Location

Looking at the distribution of players by location, we can see that most players are from the USA, Europe having the second largest share and Asia stands on third.

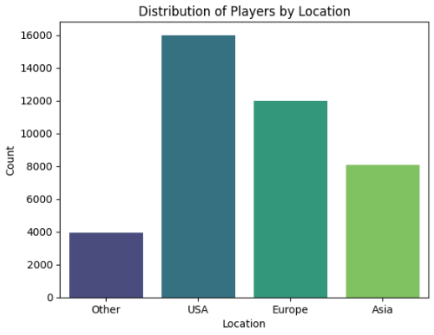


*Figure 2. 4: Distribution of Players by Location*

The player engagement metrics by location is measured in different ways and the results showed the following findings:

The mean and median number of sessions per week were highly consistent across all locations. All medians were at 9 sessions per week, indicating that players from different regions engage in a similar number of gaming sessions each week.

The session durations were similar across different locations, suggesting that players, regardless of where they were, tend to have comparable gaming habits in terms of how long they played in a single session. Players from "Other" regions and the USA exhibit similar mean and median playtime hours. The mean playtime hours for these regions were slightly over 12 hours, showing a consistent level of engagement.

### 2.3.3 Genre

A simple plot of a bar chart of GameGenre reveals that all genres were quite popular and did not give many insights. We then did made boxplots of GameGenre with playtime hours, average session per week and average session duration and the values were quite close to each other. This led us to do counts for greater insights and a better picture. Our detailed analysis using counts focuses on understanding how different game genres influence player engagement metrics, such as playtime hours, session frequency, and session duration. The key genres examined are Action, Strategy, and RPG (Role-Playing Games).

**Playtime with Genre**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **GameGenre** | | | | | | | | |
| Action | 8039.0 | 12.164645 | 6.879331 | 0.000630 | 6.268037 | 12.269403 | 18.078077 | 23.995739 |
| RPG | 7952.0 | 12.008113 | 6.915510 | 0.000158 | 6.106444 | 11.979067 | 17.987556 | 23.991246 |
| Simulation | 7983.0 | 11.898085 | 6.890893 | 0.000950 | 5.911342 | 11.897136 | 17.792838 | 23.999592 |
| Sports | 8048.0 | 11.968329 | 6.944544 | 0.003188 | 5.961856 | 11.909736 | 17.974933 | 23.997838 |
| Strategy | 8012.0 | 12.081855 | 6.941233 | 0.000115 | 6.126750 | 12.079734 | 18.041660 | 23.991985 |

***Table 2. 1: Descriptive Statistics of Game Genre***

Players of Action and Strategy games exhibited higher median playtime hours compared to other genres. This indicated that these genres were more engaging or time-consuming for players, leading to longer cumulative playtime.

Strategy games lead to more frequent gaming sessions per week compared to other genres.

Conversely, RPG games had slightly fewer sessions per week, suggesting that even though RPG players might play for long durations, they played less frequently.

Strategy games had the highest median session duration, indicating that players tend to engage in longer gaming sessions. RPG games, while still having significant session durations, showed slightly lower median session durations compared to other genres.

### 2.3.4 Player Level & Achievement Unlocked

Next, we categorized the ‘Player Level’ into discrete groups or bins. This is helpful for analysing data by grouped levels rather than individual level values. The new column created contained the bin labels corresponding to the PlayerLevel values. Each player's level was grouped into a category based on the defined ranges (0-10, 11-20, etc.). We divided them into 10 bins, and then plot the distribution on a histogram. Similarly, for achievement unlocked we created 10 bins and grouped the date together to plot it in a histogram plot. The results showed nothing significant as all groups has almost the same frequency.

### 2.3.5 Engagement Level

After this, our next step was to analyse engagement levels in relation to other parameters. The pie chart shows that medium level of engagement was the most common boasting almost 50% of the entries, whereas high and low stood at exactly the same levels of 25.8%

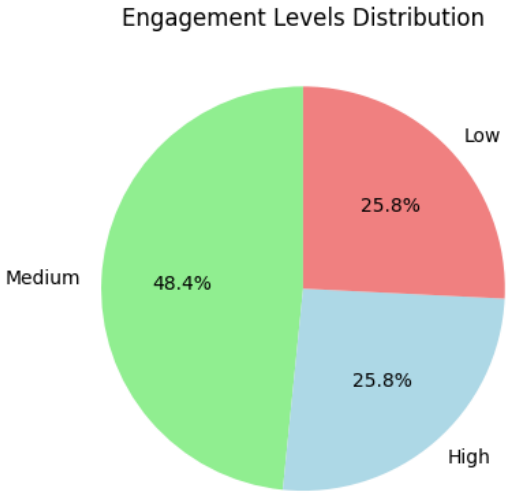


***Table 2. 2: Engagement Level***

We did engagement level analysis by plotting boxplots of engagement level with playtime hours, average session duration and sessions per week.

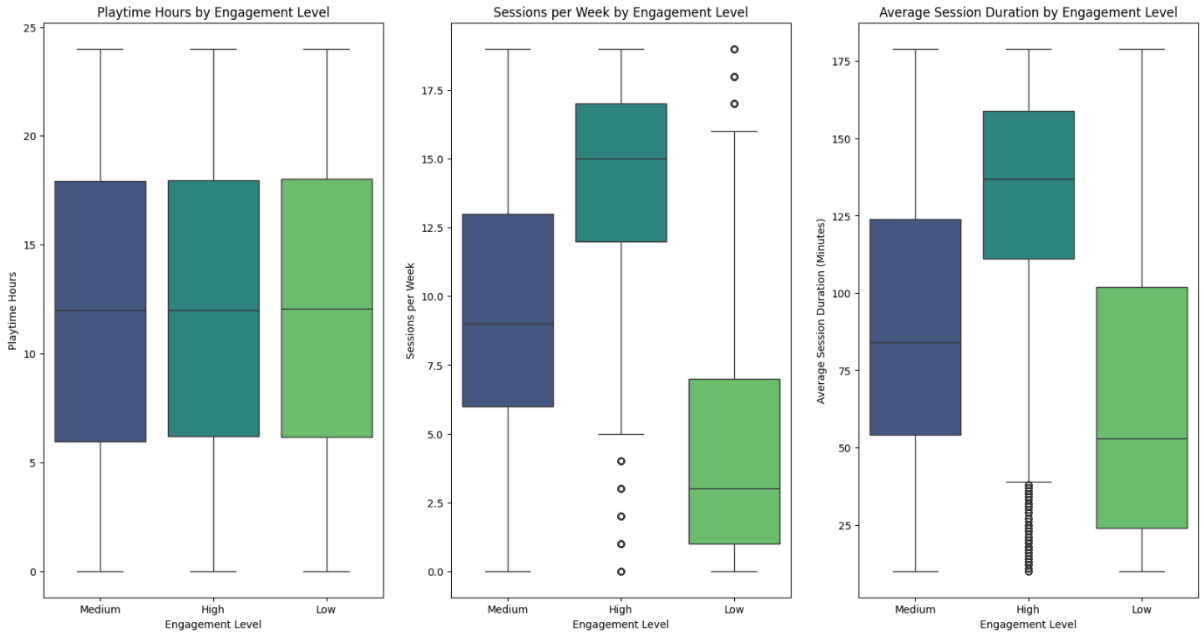


***Figure 2. 5: Engagement Level by Playtime Hours, Sessions per Week and Average Session Duration***

Analyzing these box plots, we can infer that the median playtime hours for Medium and High engagement levels are slightly higher than Low engagement level, and all three engagement levels have median playtime hours around 10-15 hours.

The IQR (the range between the first quartile Q1 and the third quartile Q3) is similar across all engagement levels. This indicates that the spread of playtime hours is consistent. The whiskers and the absence of outliers show that most players' playtime hours fall within a similar range across all engagement levels.

Players with High engagement levels had a higher median number of sessions per week compared to Medium and Low engagement levels. The median sessions per week for High engagement was around 12-13, for Medium it was around 8-9, and for Low around 5. Moreover, The IQR was widest for High engagement levels, showing that there was more variability in session duration for highly engaged players.

Then we checked engagement levels by location and did a bar plot of it to show which continents had high, medium or low engagement levels.

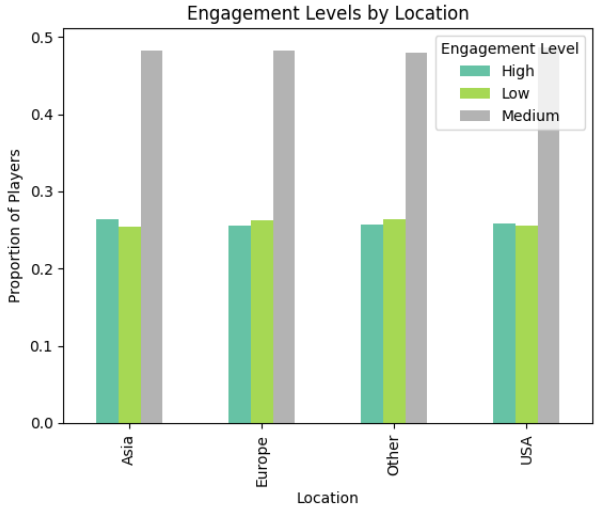


***Figure 2. 6: Engagement Levels by Location***

Asia: In Asia, approximately 50% of players were in the Medium engagement category, indicating moderate interaction with the game. High and Low engagement levels were nearly equal, each around 25%. This distribution suggests potential for increasing player engagement. Efforts could focus on converting medium engagement players to High engagement.

**Europe:** Europe also showed around 50% of players with medium engagement. High engagement levels slightly surpass Low engagement, each at approximately 25-30%. This indicated a somewhat more dedicated player base. Strategies could aim to elevate medium engagement players to higher levels of involvement.

**USA:** The USA mirrors the global trend, with 50% of players in the Medium engagement category. High and Low engagement levels were balanced at about 25% each. This suggests a stable base of regular users. Efforts could focus on boosting medium engagement players' involvement and addressing less engaged players' needs.

Across all regions, medium engagement was the most common, showing global consistency in player behaviour. High and Low engagement levels were similarly distributed, indicating a balanced mix of player dedication.

We further checked average engagement level with gender and age group and got the following results:

**Females:** Female players aged 19-25 and 26-35 show higher engagement compared to other female age groups, with a noticeable proportion of Medium engagement. Younger females (under 18) and older females (46-50) had lower engagement levels, indicating less interaction with the game.

**Males:** Male players exhibit higher engagement levels overall compared to females, particularly in the 26-35 and 36-45 age groups, where Medium engagement is predominant. This suggests that males in these age ranges were more actively engaged with the game.

The chart highlights that the highest engagement levels are found among males aged 26-45 and females aged 19-35, with Medium engagement being the most common across all groups. This data suggests targeted strategies could be developed to enhance engagement further in these high-engagement demographics.

By plotting engagement level with difficulty level of the game we found that the games with Easy difficulty had the highest average engagement level overall. Medium engagement dominates this category, significantly surpassing both High and Low engagement levels. This suggests that players were more consistently engaged in games that are easier, potentially due to the less challenging nature making them more accessible and enjoyable for a broad audience. Games which are harder have noticeable lower engagement levels as it is understandable, and games with medium difficulty has medium engagement the most common level.

This indicates that easy games are most preferred among players and thus is keeps players constantly engaged.

### 2.3.6 In-Game Purchases

To analyze In-game purchases, we again plotted it with Genre, playtime and player level with the following figures

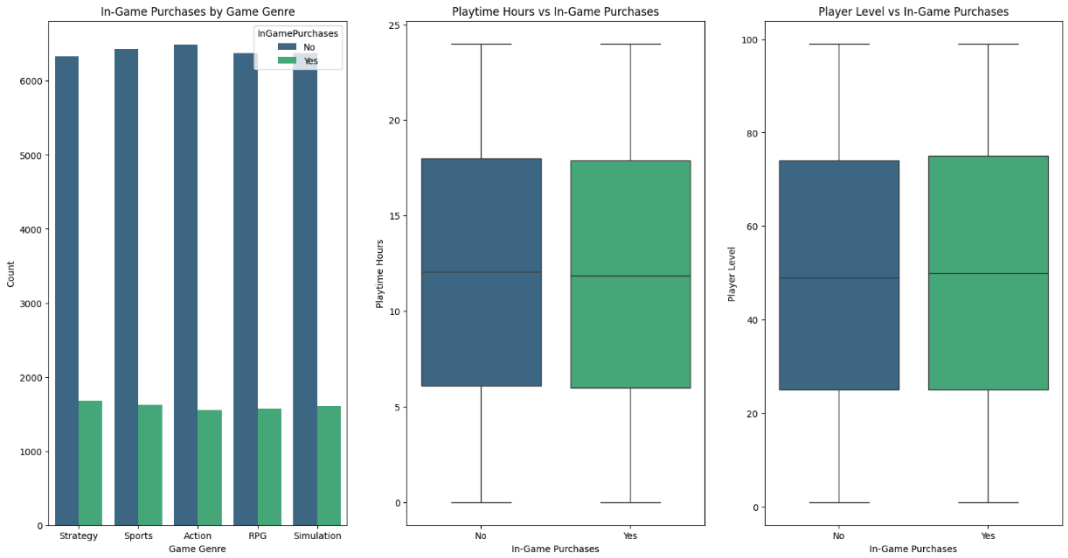


***Figure 2. 7: In-Game Purchase by Game Genre, Playtime Hours and Player Level***

**Genre:** Across all game genres (Strategy, Sports, Action, RPG, and Simulation), a smaller proportion of players made in-game purchases compared to those who do not. This trend is consistent, showing that many players in each genre do not engage in purchasing in-game items.

**Playtime hours:** There is no significant difference in the playtime hours between players who make in-game purchases and those who do not. Both groups exhibit similar median playtime hours, indicating that purchasing behaviour does not strongly correlate with the amount of time spent playing the game.

**Player Level:** The median player level is comparable for both players who make in-game purchases and those who do not. This suggests that in-game purchases do not significantly influence player progression in terms of achieving higher levels.

**Correlation:** The correlation for Engagement Level was verified after one-hot coding the variables in the dataset. The results of the heat map showed that there was no strong correlation between the original variables. The numerical indicators did not exceed 0.5. Therefore, feature engineering is necessary for accurate predictive modelling

# Chapter III: Feature Engineering

Purpose: The goal of feature engineering is to create new features that can improve the performance of machine learning models. By generating new features, we aim to provide the model with more relevant information that can help in better understanding the underlying patterns in the data.

### 3.1 New feature variable

#### 3.1.1. Total Session Duration

The total session duration for each player was calculated by multiplying the average session duration by the number of sessions per week. This provides a comprehensive metric of player engagement over time.

#### 3.1.2. Level-PlayTime Interaction

The interaction between a player's level and their playtime hours was calculated. This feature captures the relationship between the amount of time a player spends in the game and their progression level, which can be an indicator of player commitment and experience.

#### 3.1.3. Achievement Rate

The achievement rate was calculated by dividing the number of achievements unlocked by the player's level. This feature provides insight into how efficiently players are unlocking achievements relative to their level, indicating their skill or dedication.

### 3.2 Polynomial Features

To further enhance the dataset, we generated polynomial features for key numerical columns to capture non-linear relationships. This was followed by applying a log transformation to the PlayTimeHours column to normalize its distribution. For the Location feature, we used frequency encoding to represent each location based on its occurrence in the dataset, which can be more informative than one-hot encoding.

### 3.3 Converting categorical variables to continuous variables

This is the last step before building a machine learning model, where we performed the necessary coding and normalization steps in order to prepare the dataset for the mod. First, we applied label encoding to the categorical features to convert them into numerical values. This transformation was necessary for features such as Gender, Location, GameGenre, GameDifficulty, EngagementLevel, and InGamePurchases, allowing the machine learning algorithms to process these categories effectively. By using LabelEncoder, we ensured that each category within these features was assigned a unique numerical label.

Next, we focused on standardizing the numerical features. Standardization is a crucial preprocessing step that scales numerical features to have a mean of 0 and a standard deviation of 1. This process ensures that the numerical features are on a similar scale, which is important for the optimal performance of many machine learning algorithms. We applied this transformation to key numerical features such as Age, PlayTimeHours, SessionsPerWeek, AvgSessionDurationMinutes, PlayerLevel, and AchievementsUnlocked using StandardScaler.

# Chapter IV: Model Development

## 4.1. Predicting Engagement Level

In our approach, we aimed to predict the player's EngagementLevel based on a set of input features obtained from the dataset. These features included both numerical and categorical variables such as PlayTimeHours, SessionsPerWeek, AvgSessionDurationMinutes, PlayerLevel, AchievementsUnlocked, Gender, Location, GameGenre, GameDifficulty, and InGamePurchases. In our initial analysis, we explored several machine learning models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting Classifier, and Random Forest. Each model was evaluated based on its performance metrics, including accuracy, precision, recall, F1 score, and the confusion matrix.

### 4.1.2 Splitting the Dataset and baseline model

The dataset was first split into features (X) and the target variable (y). The features included all columns except EngagementLevel. The data was then divided into training and testing sets with an 80-20 split to ensure that the models were trained on a majority of the data while being evaluated on a separate, unseen portion to test generalization. In addition, A Random Forest Classifier was trained on the training set to establish a baseline model.

**Logistic Regression Model**
Logistic Regression was used to predict the EngagementLevel of players based on various input features. This model is known for its simplicity and interpretability, making it useful for understanding the impact of features on the predicted outcomes. The Logistic Regression model achieved an accuracy of 86%, with precision scores of 90% for the "High" class, 85% for the "Low" class, and 85% for the "Medium" class. Recall scores were 83%, 81%, and 91% for the "High", "Low", and "Medium" classes, respectively. The F1 scores were 87% for "High", 83% for "Low", and 88% for "Medium". These results indicate that the model performed well, especially for the "Medium" class, and achieved balanced performance across all classes, making it effective at predicting EngagementLevel with reasonable accuracy.

**Support Vector Machine**
Support Vector Machine (SVM) was employed to classify EngagementLevel due to its effectiveness in high-dimensional spaces and versatility with different kernel functions. The SVM model achieved an accuracy of 89%, with precision scores of 92% for the "High" class, 87% for the "Low" class, and 88% for the "Medium" class. Recall scores were 87%, 83%, and 93% for the "High", "Low", and "Medium" classes, respectively. The F1 scores were 89% for "High", 85% for "Low", and 90% for "Medium". These results demonstrate that the SVM model performed excellently, particularly in predicting the "Medium" and "High" classes, indicating its robustness in handling complex data structures.

**K-Nearest Neighbors**
The K-Nearest Neighbors (KNN) algorithm was chosen for its simplicity and effectiveness in handling multi-class classification problems. KNN classifies samples based on the majority class among their k-nearest neighbors in the feature space. This model achieved an accuracy of 88%, with precision scores of 91% for the "High" class, 85% for the "Low" class, and 89% for the "Medium" class. Recall scores were 87%, 85%, and 90% for the "High", "Low", and "Medium" classes, respectively. The F1 scores were 89% for "High", 85% for "Low", and 90% for "Medium". These results indicate that KNN is highly effective, especially in predicting the "Medium" and "High" classes, demonstrating its ability to handle the dataset well.

**Gradient Boosting**
Gradient boosting classifiers are included in the evaluation for their flexibility and ability to handle complex datasets. This ensemble learning method builds models sequentially, with each new model correcting the errors of the previous one. The accuracy of the gradient boosting model was 92% for the "high" category, 93% for the "low" category and 92% for the "medium" category. The recall rates for the "high", "low" and "medium" categories were 89%, 90% and 95% respectively. The F1 scores for the 'high', 'low' and 'medium' categories are 91%, 91% and 94% respectively. These results highlight the model's ability to improve its predictive accuracy through iterative improvement, making it the most accurate of the models evaluated. But it also raises concerns about overfitting.

### 4.2 Classification Report

This evaluation step aims to compare and summarize the performance of different models on the test dataset. Detailed metrics such as accuracy, precision, recall, F1 score, and confusion matrix are calculated and analyzed to provide a comprehensive understanding of the strengths and weaknesses of each model. Visualizing the confusion matrix helps identify specific types of errors and guides potential model improvements. This will help in selecting the most appropriate model.

### 4.3 Feature_importances

The bar chart below illustrates the importance of the features of the model, highlighting which features contribute most to predicting player engagement. This analysis helps to understand the relative importance of each feature in the model.

The feature importance analysis reveals that the TotalSessionDuration is the most significant predictor of player engagement levels, contributing nearly 50% to the model's decisions. This indicates that the total time players spend in sessions is crucial in determining their engagement. Following this, the SessionsPerWeek feature, which reflects the frequency of player engagement, is also highly influential. The average duration of each session, represented by AvgSessionDurationMinutes, further highlights that not just the frequency but also the length of individual sessions plays an important role in predicting engagement.

Moderately important features such as AchievementRate, PlayerLevel, and AchievementsUnlocked suggest that a player's progress and achievements in the game are relevant factors in their engagement. Additionally, features like LevelPlayTimeInteraction, Log_PlayTimeHours, and PlayTimeHours indicate that various aspects of playtime and interaction between different levels were considered by the model.

Less impactful features, including Age, GameGenre, Location_freq, Location, GameDifficulty, Gender, and InGamePurchases, contribute less to the model's predictions. While these features have some influence, their lower importance suggests they have a smaller effect on determining engagement levels compared to the top features.

Overall, the analysis underscores the importance of session duration and frequency in predicting player engagement. By focusing on these key factors, game developers and marketers can design strategies to enhance player retention and engagement, ultimately contributing to the success of the game.

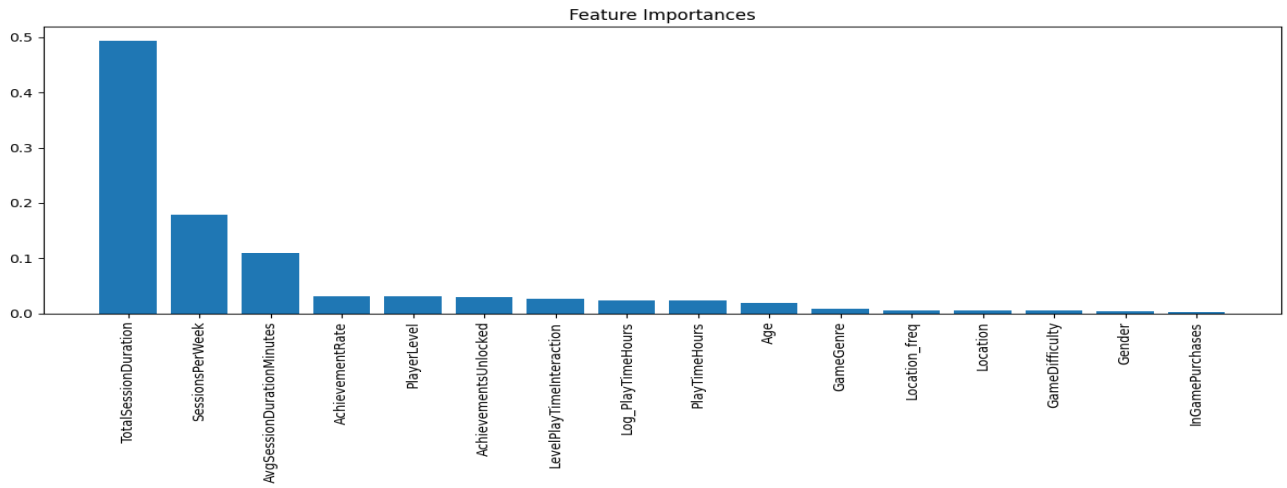


***Figure 3. 1: Feature Importance***

### Hyperparameter

In this step, we aimed to optimize the hyperparameters of the Random Forest classifier to achieve the best possible performance in predicting EngagementLevel. We began by defining a parameter grid that specified the hyperparameters and their possible values to be evaluated. The grid included the number of trees in the forest (n_estimators), with values of 100, 200, and 300; the maximum depth of the trees (max_depth), with options of None, 10, 20, and 30; the minimum

number of samples required to split an internal node (min_samples_split), set to 2, 5, and 10; and the minimum number of samples required to be at a leaf node (min_samples_leaf), with values of 1, 2, and 4.

Next, we initialized the GridSearchCV with the Random Forest model as the estimator, the defined parameter grid, and additional settings such as 3-fold cross-validation (cv=3), utilizing all available cores for parallel computation (n_jobs=-1), and detailed logging (verbose=2). This setup allowed us to perform an exhaustive search over 108 different combinations of parameters, resulting in a total of 324 fits (3 folds for each combination).

We then fitted the GridSearchCV to the training data (X_train and y_train). This process evaluated each combination of hyperparameters, training and validating the model multiple times to ensure robust performance. After fitting, we retrieved the best estimator—the model with the optimal hyperparameters—using grid_search.best_estimator_.

The optimized Random Forest model was then used to make predictions on the test set (X_test), and its performance was evaluated using accuracy, precision, recall, and F1-score. The best model achieved an accuracy of 92.1%, with precision, recall, and F1 scores for each class (0, 1, 2) being approximately 91-95%, indicating a well-performing model across all classes.

This step is crucial because it systematically searches for the best combination of hyperparameters, ensuring that the Random Forest model is optimized for the dataset. By performing cross-validation, it ensures the model's performance is robust and not overly dependent on a particular split of the data. The result is a highly tuned model that generalizes well to new, unseen data, providing accurate and reliable predictions. GridSearchCV's exhaustive search and validation make this process critical in achieving a finely tuned and high-performing model.

**Limitation of Random Forest Model**

The Random Forest model works by constructing a multitude of decision trees during training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It reduces overfitting and improves accuracy by averaging multiple decision trees, thus leveraging the "wisdom of the crowd".

- **Data Quality:**
  The accuracy and insights of the model are highly dependent on the quality and completeness of the data. Missing or inaccurate data can significantly impact model performance.
- **Overfitting:**
  While Random Forest mitigates overfitting better than a single decision tree, it can still be prone to overfitting if the trees are not pruned, or the dataset is small.
- **Interpretability:**
  Although Random Forest models provide high accuracy, they are often less interpretable than simpler models like logistic regression, making it harder to understand individual decision paths.

# Chapter V: Conclusion and Recommendations

## 5.1 Conclusion

In the ever-expanding universe of online gaming, understanding the behavior of players and the level of engagement is important for retaining customers as well as maximizing the revenue for the gaming companies. The analysis on the online gaming behavior of the players helped explore the player engagement patterns in online gaming. The study has demonstrated critical insights into the behavior and engagement of the customers in the online gaming environments.

To improve the performance of machine learning models various new features such as total session duration, level-playtime interaction and achievement rate were generated. The analysis demonstrated that total session duration, session frequency, and average session duration are the most significant predictors of player engagement. This indicated that the length and frequency of play are important factors in retaining players which further suggests that if games are designed to encourage longer as well as frequent sessions, then the engagement level and the retention rates improve.

After conducting feature engineering, polynomial features were generated for the key numerical columns to capture non-linear relationships by applying log transformation to PlayTimeHours and frequency encoding for the Location feature. Then, the categorical features such as Gender, Location, GameGenre, GameDifficulty, EngagementLevel, and InGamePurchases were converted into numerical values by applying label encoding. Besides, the dataset was split into features and target variables and then the data was divided into training and testing sets with an 80-20 split. For analysis purposes, various machine learning models such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest and Gradient Boosting Classifier were used. Each of the five models were evaluated based on key performance metrics such as accuracy, precision, recall and F1 score. Among all these models, Random Forest model achieved the highest accuracy of 92.1% after optimizing it through hyperparameter tuning. The highest accuracy demonstrated the robustness of the model in handling complex datasets with multiple features. For the hyperparameter tuning process, GridSearchCV was used which ensured that the parameters of the models were finely tuned to provide a high performing model.

The comprehensive analysis of online gaming behavior has provided an invaluable insight into the player's engagement patterns. Online gaming companies should focus more on designing strategies to improve the playtime of the players along with the frequency of each session to retain players and improve the in-game purchases which ultimately improves the success of online games and company.

## 5.2 Recommendations

For the gaming industry, a moderate level of participation is normal because gaming is just a part of life. However, based on the results of the analysis, it is recommended to stimulate players with low levels of engagement to increase their motivation. From the EDA and Feature Engineering section, the first is that players with higher player level have higher engagement. The game company needs to increase more attractive leveling up rewards. The results of the analysis of game difficulty on engagement show that players are significantly more engaged in games with easy difficulty. The company needs to reduce the overall game difficulty and add more newbie-friendly policies to prevent player motivation from diminishing.

An analysis of engagement levels by gender and age group shows that male players, especially those in the 26-35 and 36-45 age groups, exhibit higher overall engagement levels compared to female players. Companies need to design in-game details and merchandise for this segment. For example, Tencent, the world's largest gaming company, uses a common marketing strategy to incorporate popular movie, anime, and game images from the childhood of its age-segmented customers into its game designs (Grguric, 2024). This has a significant effect on building a community of players while also providing attractive appeal to them. And for female players who are less engaged, companies need to add more feminine designs. For example, introducing more in-game skinning options for female players would significantly increase player engagement.

The last is the significant impact of achievement rate and achievement unlocking on player engagement shown in the model's feature significance analysis. The company needs to upgrade the achievement system by introducing dynamic challenges and seasonal events that provide additional achievements. This can spark interest and encourage players to play longer.

Furthermore, the analysis of the feature importance of the model shows that factors related to Session Duration have the most significant effect on the player's engagement level. This suggests that game companies need to design highly attractive and even addictive game modes and content to increase player session duration.

The first is the ladder or ranking system that inspires a sense of competitive accomplishment in players. Take the example of dataset sports and action games. Such systems have two benefits. A segmented ranking system allows players to compete with opponents of their own level, creating a more balanced competitive environment and making winning much easier (Tam, 2023). It is also fundamental to increasing player engagement as it constantly sets goals for the player and incentivizes them to keep competing.

Then there is the in-game collection system. The Sessions per Week by Engagement Level result that highly engaged players tend to have a higher median number of weekly play counts than average and less engaged players. Companies need to design more weekly check-ins or in-game collection rewards to incentivize engagement. The purpose of this recommendation is the same as the ranking system, which is to drive motivation by constantly setting soft goals for players to increase session length and engagement.

Finally, companies that have succeeded in motivating players need to turn that motivation into demand. A complete and strong online community is necessary for success. The critical function of a community is to bring together people from different countries, cultures and ages (Games, 2023). Players will display their personal achievements and share their experiences within the community. Highly engaged players will drive the interest and enthusiasm of less engaged and newer players. In addition, gaming companies can get feedback directly from players. Gaming communities can provide valuable ideas and suggestions to help improve the gaming experience (Games, 2023). This interaction helps to better identify problems and respond quickly. In short, this kind of ecosystem-like community will upgrade gaming into a part of players' life demands when it's completed.

# REFERENCES


https://scikit-learn.org/stable/

https://seaborn.pydata.org/generated/seaborn.boxplot.html

https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.pie.html

https://seaborn.pydata.org/generated/seaborn.countplot.html#

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.unstack.html

https://seaborn.pydata.org/generated/seaborn.heatmap.html

Tam, H. (2023, July 12). 4 ways to improve player engagement with real-time data. Redis. https://redis.io/blog/real-time-data-for-improving-player-engagement/

Games, B. (2023, August 16). Gaming communities: The rise and impact of gaming communities on the world of video games. Medium. https://bggames.medium.com/gaming-communities-the-rise-and-impact-of-gaming-communities-on-the-world-of-video-games-1fec152f649f

Grguric, M. (2024, June 4). Mobile game session length: How to track & increase it. Udonis Mobile Marketing Agency. https://www.blog.udonis.co/mobile-marketing/mobile-games/session-length