Gaming Data Analysis Report

1. Introduction

Dataset Overview:

The dataset analyzed in this project was sourced from a fictional gaming company om kaggle. It contains comprehensive information on player demographics, gaming habits, engagement levels, and in-game purchases. The dataset consists of 40,043 rows and 15 features, including variables such as age, gender, game genre, engagement level, and session duration.

Objective:

The goal of this analysis is to:

- Identify key demographics most likely to make in-game purchases.
- Understand factors contributing to high engagement levels.
- Explore player behaviors across game genres, engagement levels, and locations.
- Build a predictive machine learning model to classify player engagement levels based on various features.

How the Analysis Was Conducted:

To achieve the project's goals, a combination of tools, techniques, and methodologies was employed:

1. **Dataset Management**:
    - The dataset was stored and queried using **SQLite**, a lightweight relational database.
    - SQL queries were extensively used to perform data extraction, transformation, and summary analysis.
2. **Data Analysis and Visualization**:
    - Data was processed and analyzed in **Python** using libraries such as:
        - **pandas**: For data manipulation and exploratory analysis.
        - **matplotlib** and **seaborn**: For creating visualizations to represent insights derived from the data.
    - SQL queries were embedded within Python scripts to enable efficient analysis.
3. **Machine Learning**:
    - Two approaches were implemented for predictive analysis:
        - **Existing Model**: A pre-built **Random Forest Classifier** from the Scikit-learn library was utilized to predict engagement levels.
        - **Custom-Built Model**: A manually implemented **Logistic Regression** model was developed to predict engagement levels, offering deeper insights into the modeling process.
4. **Key Steps**:
    - **Data Preprocessing**: Standardizing numerical features and encoding categorical variables for both SQL-based and machine learning analyses.
    - **Exploratory Analysis**: Statistical summaries and SQL-based groupings to explore patterns in demographics, game genres, and player behaviors.
    - **Visualization**: Bar charts, line plots, and feature importance graphs were used to present results effectively.

Tools and Libraries Used

- **Programming Language**: Python
- **Libraries**:
    - **pandas**: For data manipulation and analysis.
    - **matplotlib** and **seaborn**: For visualizations.
    - **Scikit-learn**: For implementing the Random Forest model.
    - **NumPy**: For mathematical operations and data standardization.
- **Database Management**: SQLite for storing, querying, and managing the dataset.

---

2. Data Summary

Dataset Details:

- **Number of rows:** 40,043
- **Number of columns:** 15
- **Key Features:**
    - `Age`, `Gender`, `Location`
    - `EngagementLevel` (High, Medium, Low)
    - `PlayTimeHours`, `SessionsPerWeek`, `AchievementsUnlocked`
    - `InGamePurchases`, `GameGenre`, `GameDifficulty`

Preprocessing Steps:

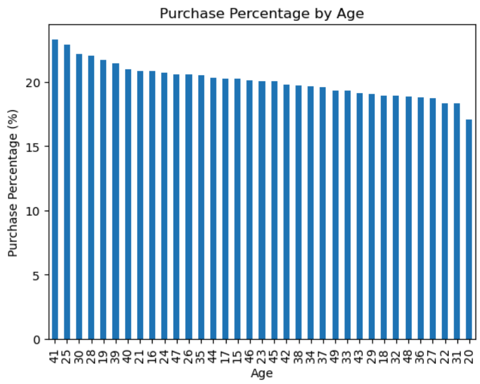
- Encoded categorical variables such as `EngagementLevel`, `GameGenre`, and `GameDifficulty` using one-hot encoding and label encoding.
- Handled missing or infinite values by imputation and standardization for numerical features.

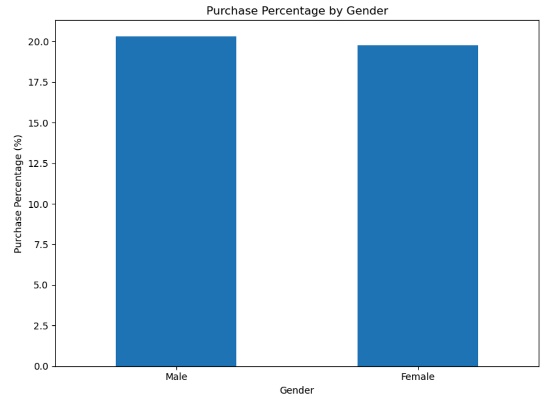
---

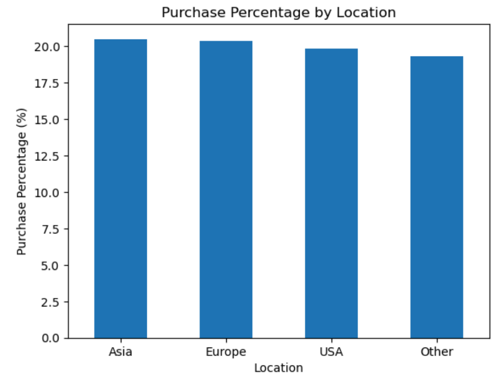
3. Analyses and Insights

3.1. Which demographics are most likely to make in-game purchases?

- **Analysis**:
    - Age: Players aged 41 and 25 had the highest purchase percentages.
    - Gender: Males and females showed comparable purchase percentages, with a slight edge for males.
    - Location: Players from Asia and Europe were more likely to make purchases compared to other regions.
- **Plots**:

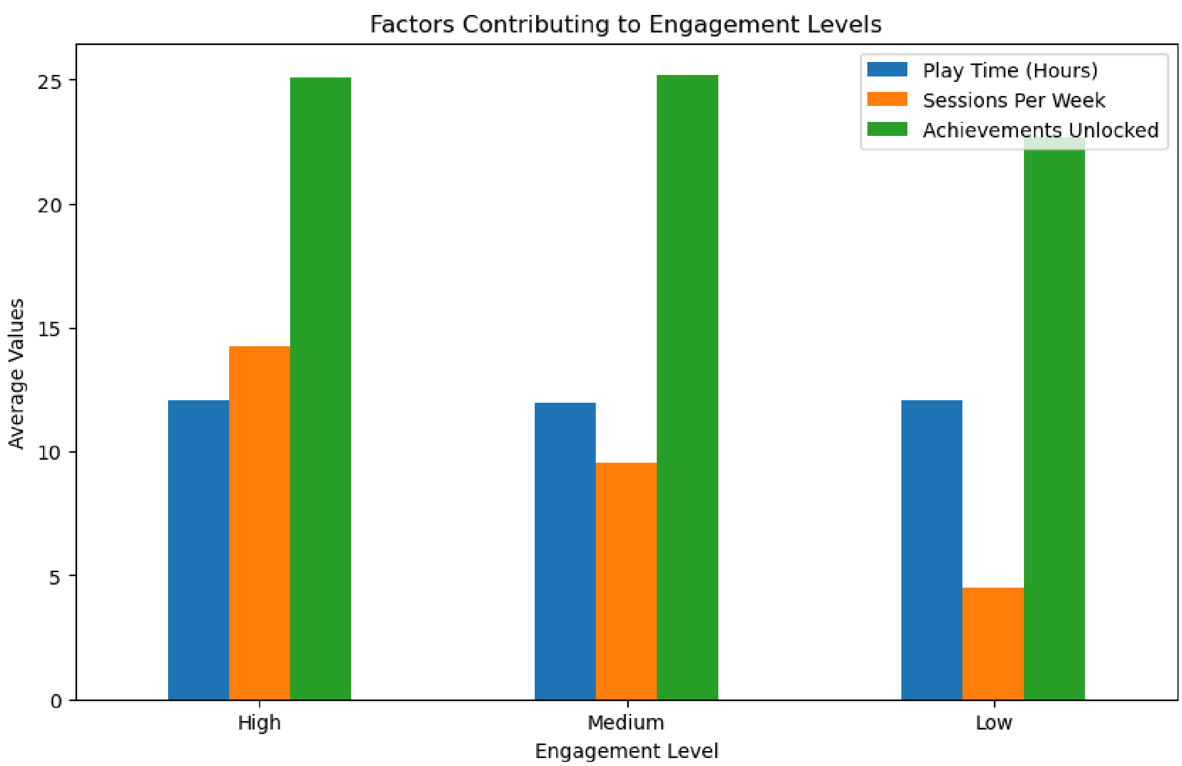






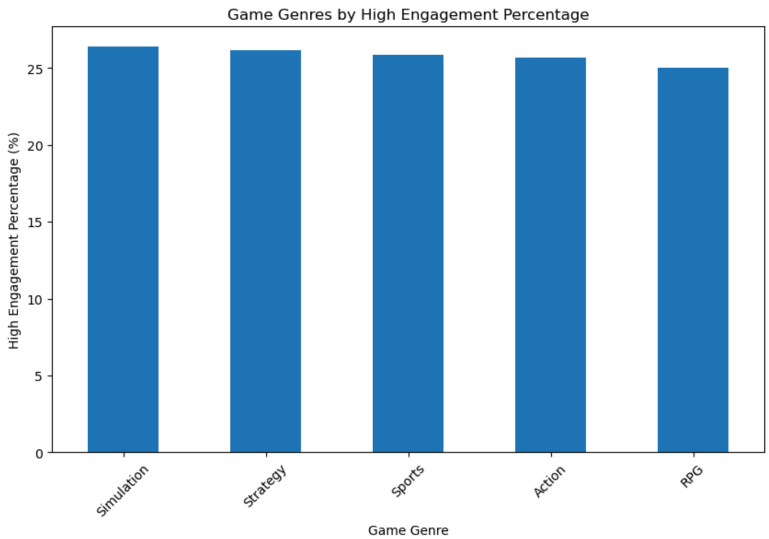
3.2. What factors contribute most to high engagement levels?

- **Analysis**:
  - Sessions per week was the most significant factor that contributed to high engagement levels as an increase in sessions per week led to an increase in high engagement levels.
- **Plot**:



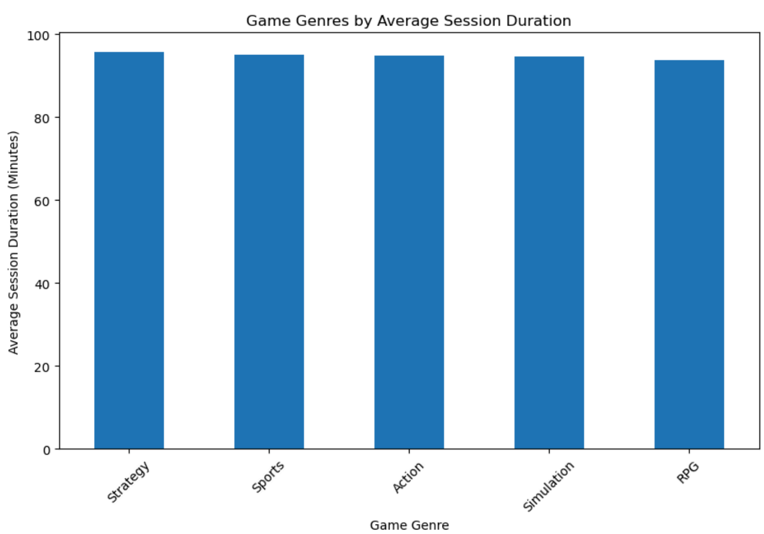
3.3. Which game genres have the highest engagement?

- **Analysis**:
  - Simulation games had the highest engagement but not by a significant amount.
- **Plot**:



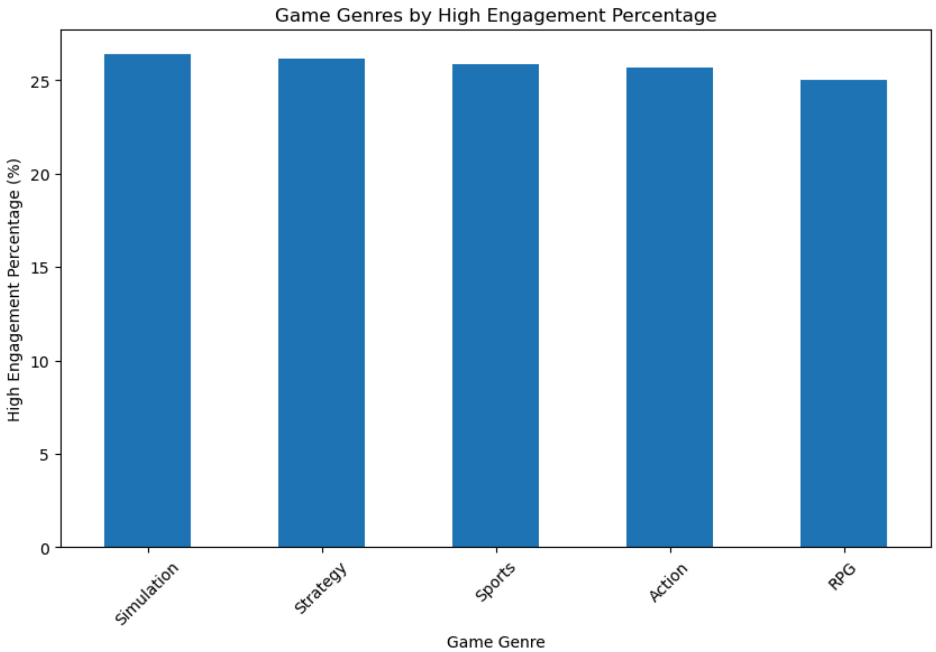
3.4. Which game genres have the longest average session durations?

- **Analysis**:
    - No significant difference in Average Session Duration between different Game Genres.
- **Plot**:



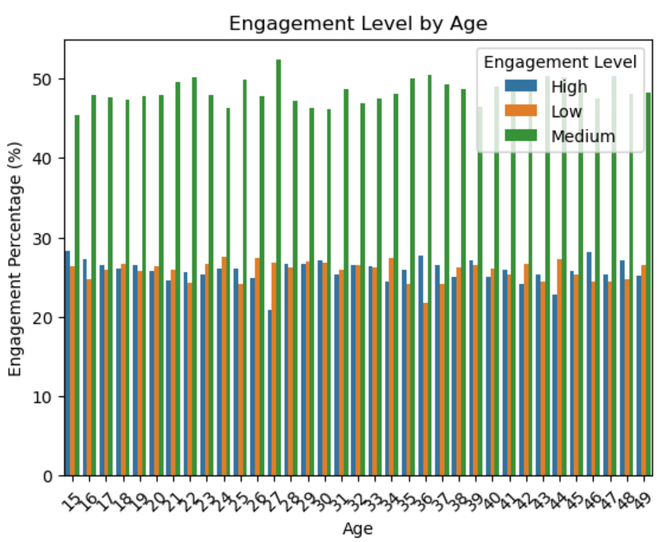
3.5. Which genres are most popular among high engagement players?

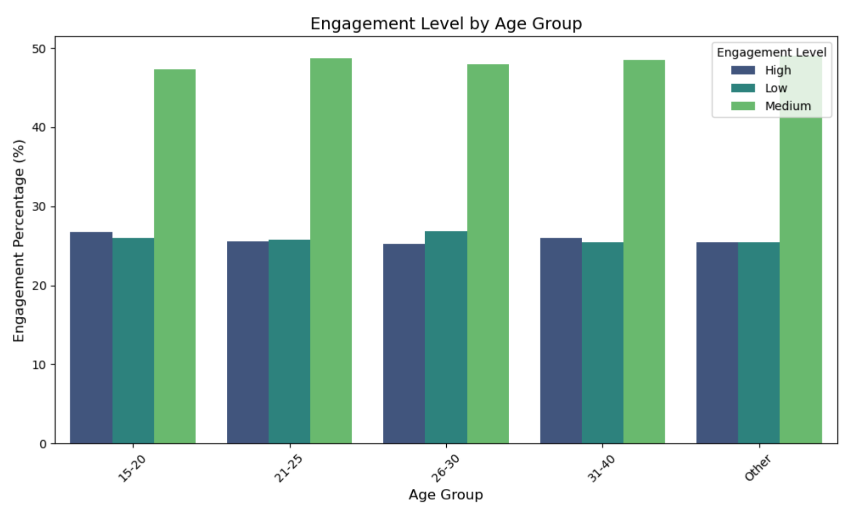
- **Analysis**:
    - Simulation and Strategy games are the most popular among players with high engagement levels but not by a significant amount.
- **Plot**:



3.6. How does engagement level vary by age?

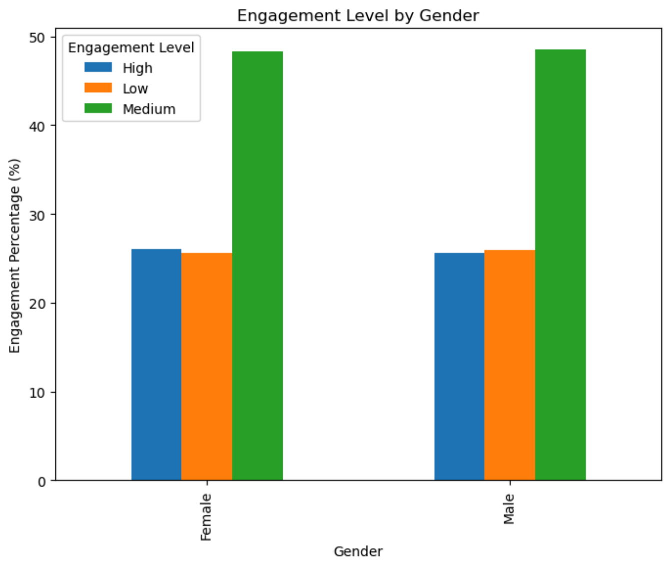
- **Analysis**:
  - Engagement levels were distributed almost evenly across age, with no significant differences.
- **Plots**:





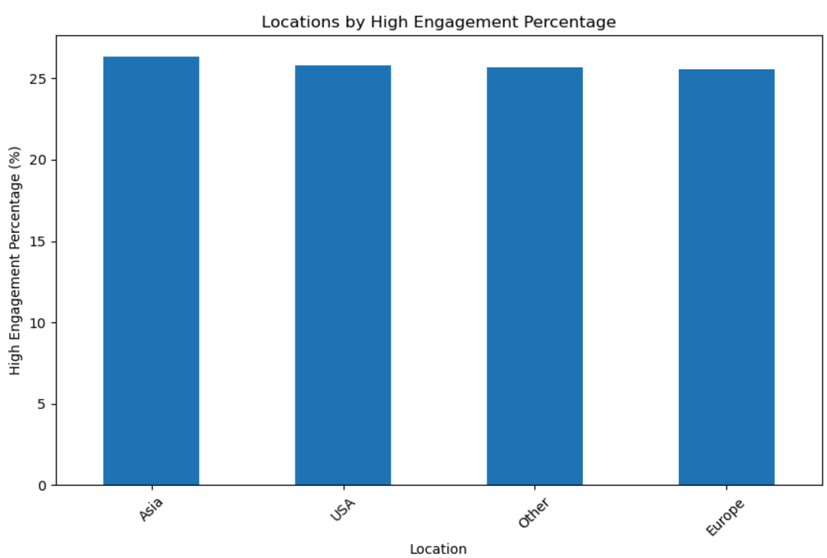
3.7. How does engagement level vary by gender?

- **Analysis**:
    - Engagement levels were distributed almost evenly across genders, with no significant differences.
- **Plot**:



3.8. Which locations have the highest proportion of high engagement players?

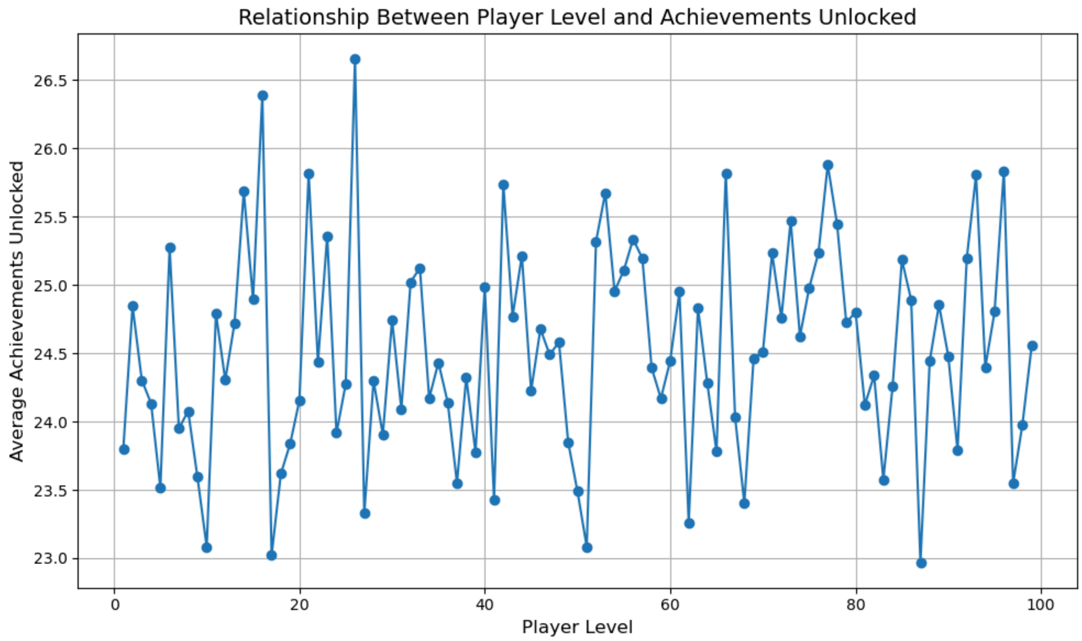
- **Analysis**:
    - Players from Asia exhibited the highest proportion of high engagement levels but not by a significant amount.
- **Plot**:



3.9. Is there a significant relationship between player level and achievements unlocked?

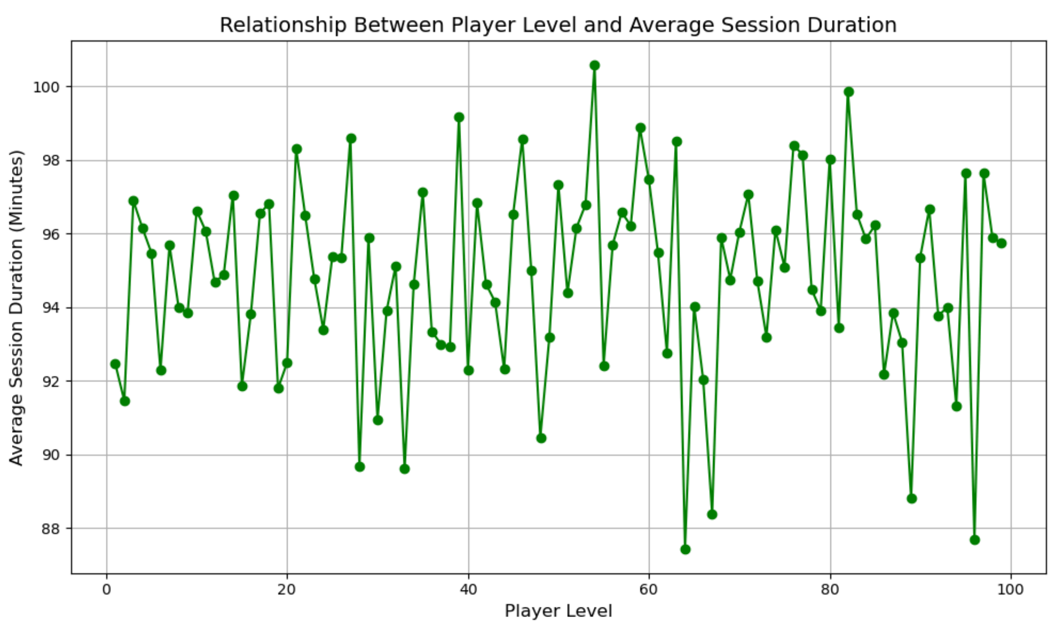
- **Analysis**:

    - There is no relationship between player level and achievements unlocked.
- **Plot**:



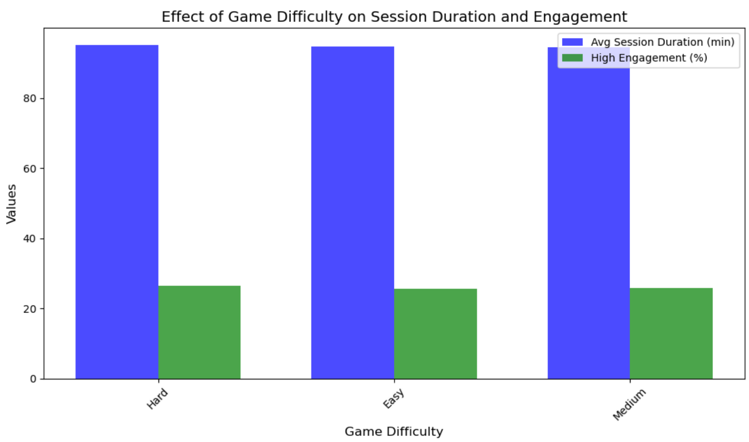
3.10. Do higher levels correlate with longer average session durations?

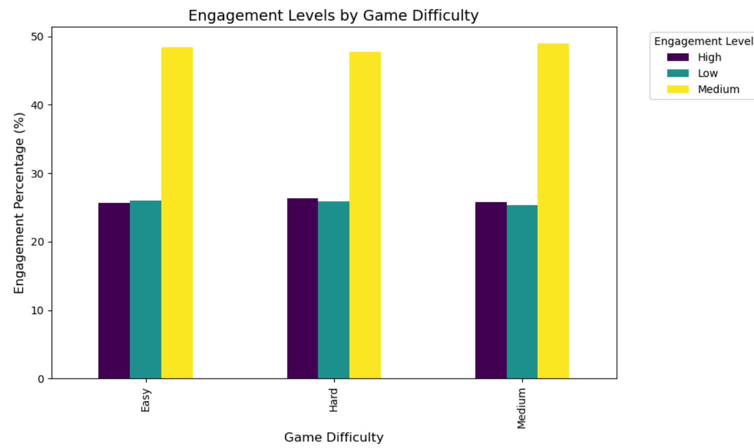
- **Analysis**:
    - Higher levels do not correlate with longer average sessions durations.
- **Plot**:



3.11. How does game difficulty affect session duration, engagement, and retention?

- **Analysis**:
    - Game difficulty does not seem to affect session duration, engagement, and retention.
- **Plots**:





---

4. Machine Learning Model

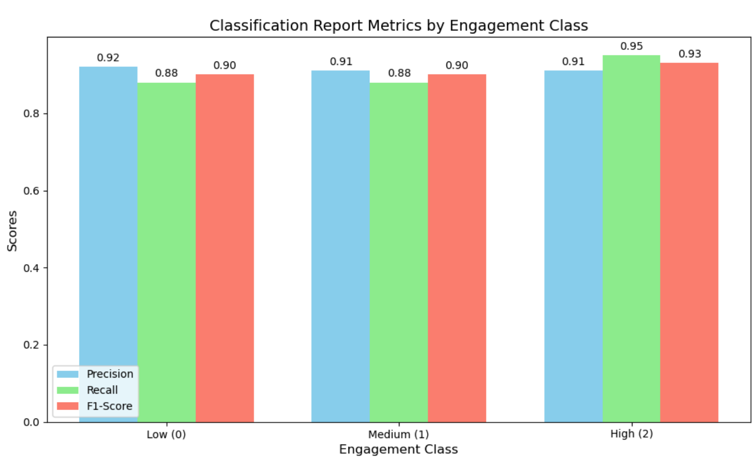
Objective:

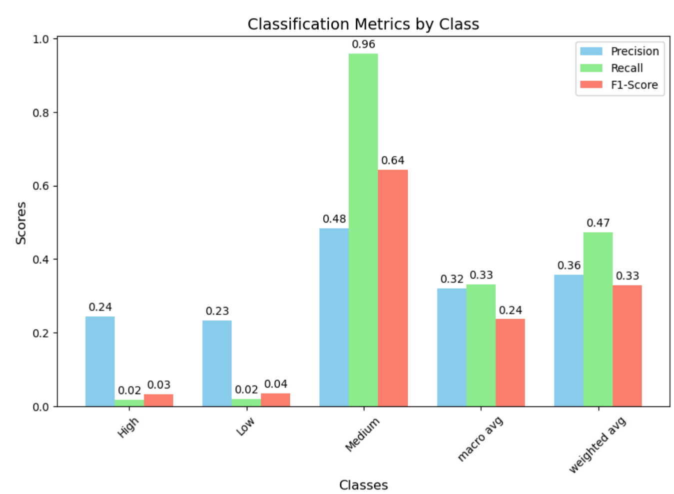
To predict player engagement levels (high, medium, low) using features such as PlayTimeHours, SessionsPerWeek, AchievementsUnlocked, and more.

Model Details:

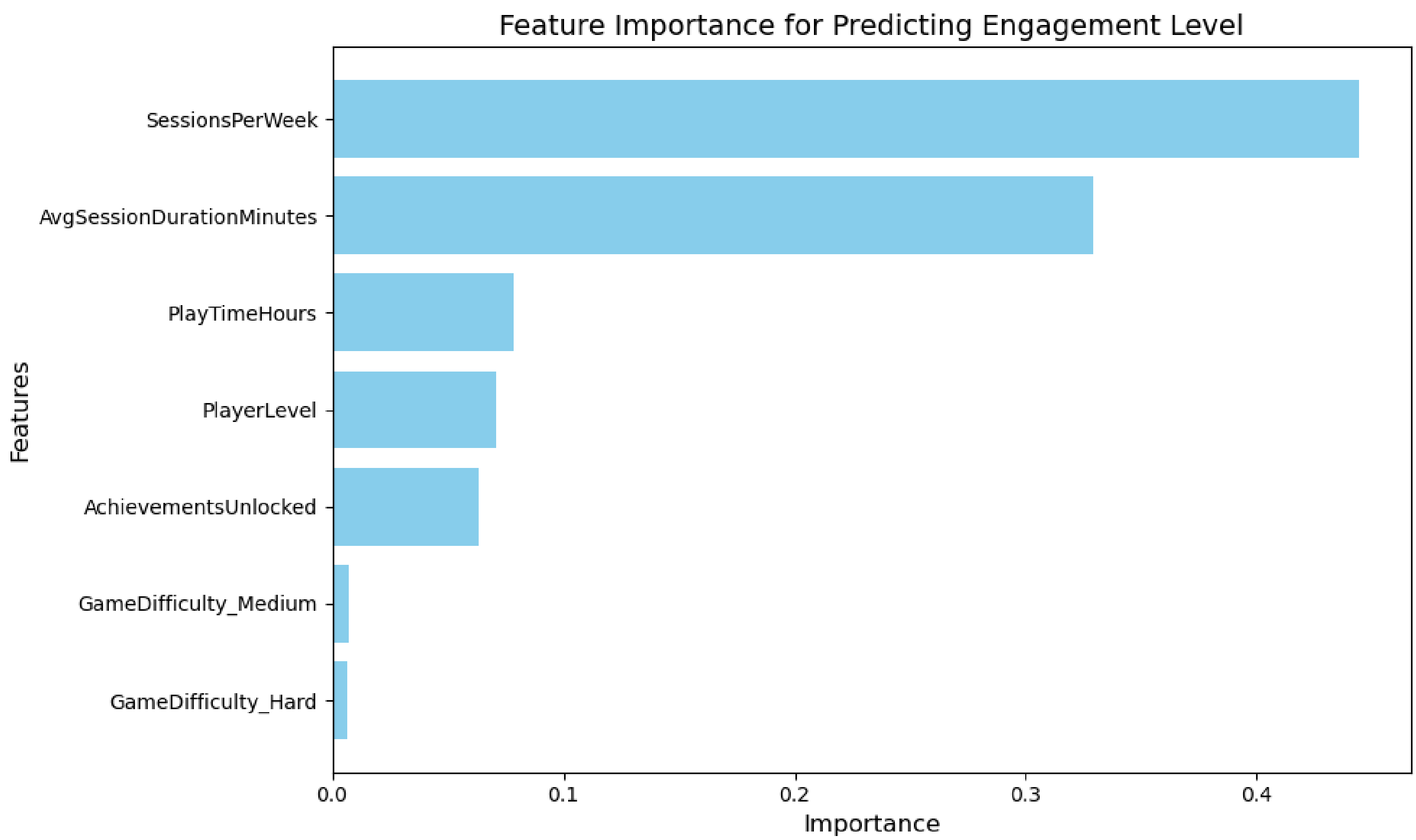
- **Algorithm Used**: Random Forest Classifier.
- **Preprocessing Steps**:
    - Categorical variables were one-hot encoded.
    - Numerical features were standardized.

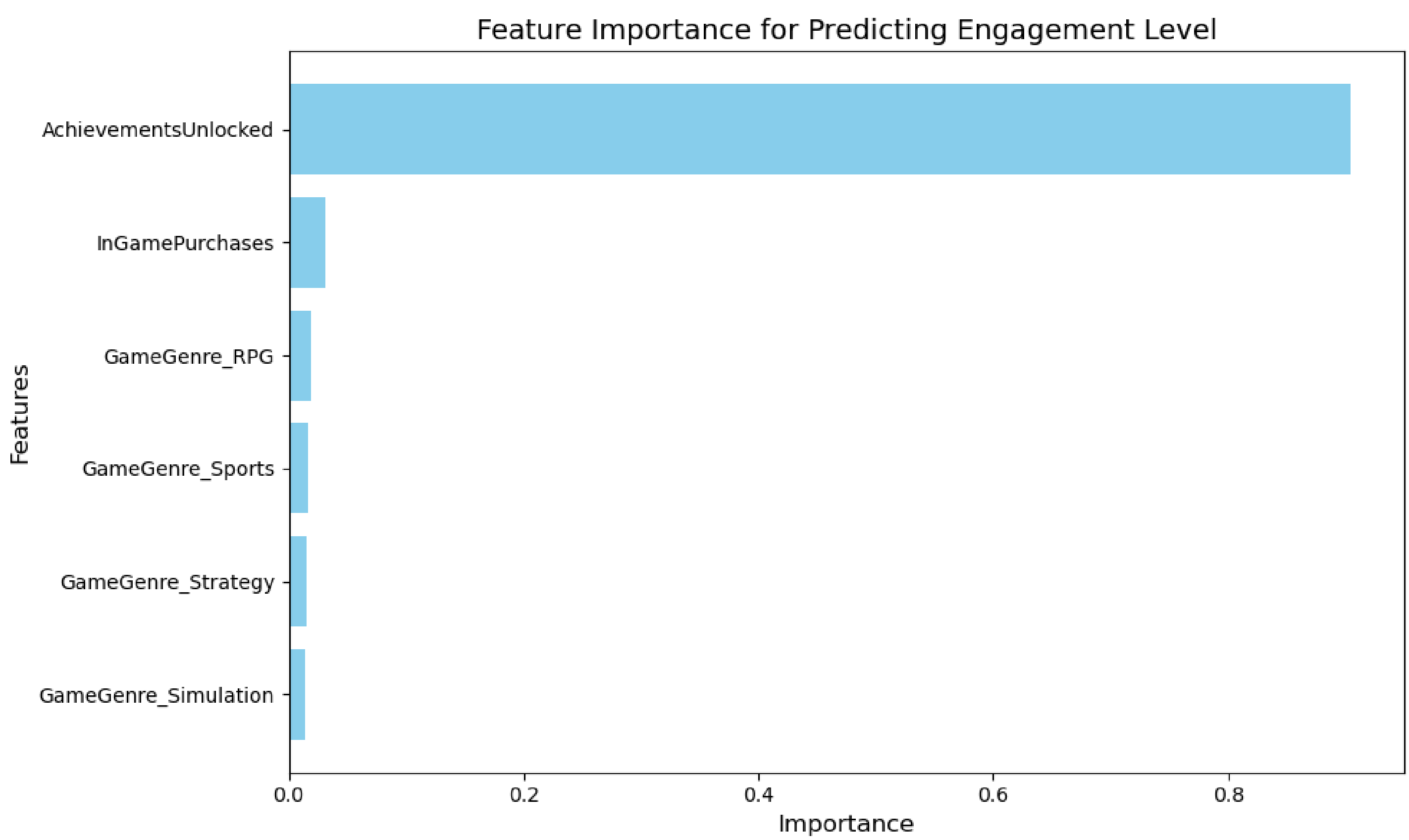
Evaluation Metrics:





Plots:


Feature Importance for Predicting Engagement Level
SessionsPerWeek
AvgSessionDurationMinutes
PlayTimeHours
PlayerLevel
AchievementsUnlocked
GameDifficulty_Medium
GameDifficulty_Hard
Features
0.0
0.1
0.2
0.3
0.4
Importance


Feature Importance for Predicting Engagement Level
AchievementsUnlocked
InGamePurchases
GameGenre_RPG
GameGenre_Sports
GameGenre_Strategy
GameGenre_Simulation
Features
0.0
0.2
0.4
0.6
0.8
Importance

5. Machine Learning Model - Custom-Built

Objective

The goal of this analysis was to develop a machine learning model from scratch to predict player engagement levels (High, Medium, or Low) using features such as SessionsPerWeek, AvgSessionDurationMinutes, AchievementsUnlocked, PlayerLevel, PlayTimeHours, GameDifficulty Hard, GameDifficulty_Medium

---

Model Details

- **Algorithm Used**: Logistic Regression
    - Logistic regression was implemented manually without using pre-built libraries for training and predictions.
    - This approach allowed greater control over the optimization process and a deeper understanding of the algorithm.

Preprocessing Steps

1. **Categorical Variables**:
    - The `GameDifficulty` column was one-hot encoded to convert it into binary columns for each difficulty level.
2. **Numerical Features**:
    - Standardized numerical features by subtracting the mean and dividing by the standard deviation, ensuring all features were on the same scale.
3. **Target Variable**:
    - The `EngagementLevel` column was label-encoded into numeric values for compatibility with the logistic regression algorithm:
        - High = 2
        - Medium = 1
        - Low = 0.
4. **Data Cleaning**:
    - Any `NaN` or infinite values were replaced with zero or appropriately handled during preprocessing.
5. **Dataset Split**:
    - Data was split into training (80%) and test (20%) sets using manual splitting to evaluate the model's performance.

Model Training

- **Gradient Descent Implementation**:
    - The model parameters (weights) were optimized using gradient descent with a sigmoid activation function.
    - Loss Function: Binary Cross-Entropy for each class.
    - Iterations: 1000

    ○ Learning Rate: 0.01
- Separate one-vs-all logistic regression models were trained for each class (High, Medium, Low) to handle the multi-class classification problem.

Evaluation Metrics

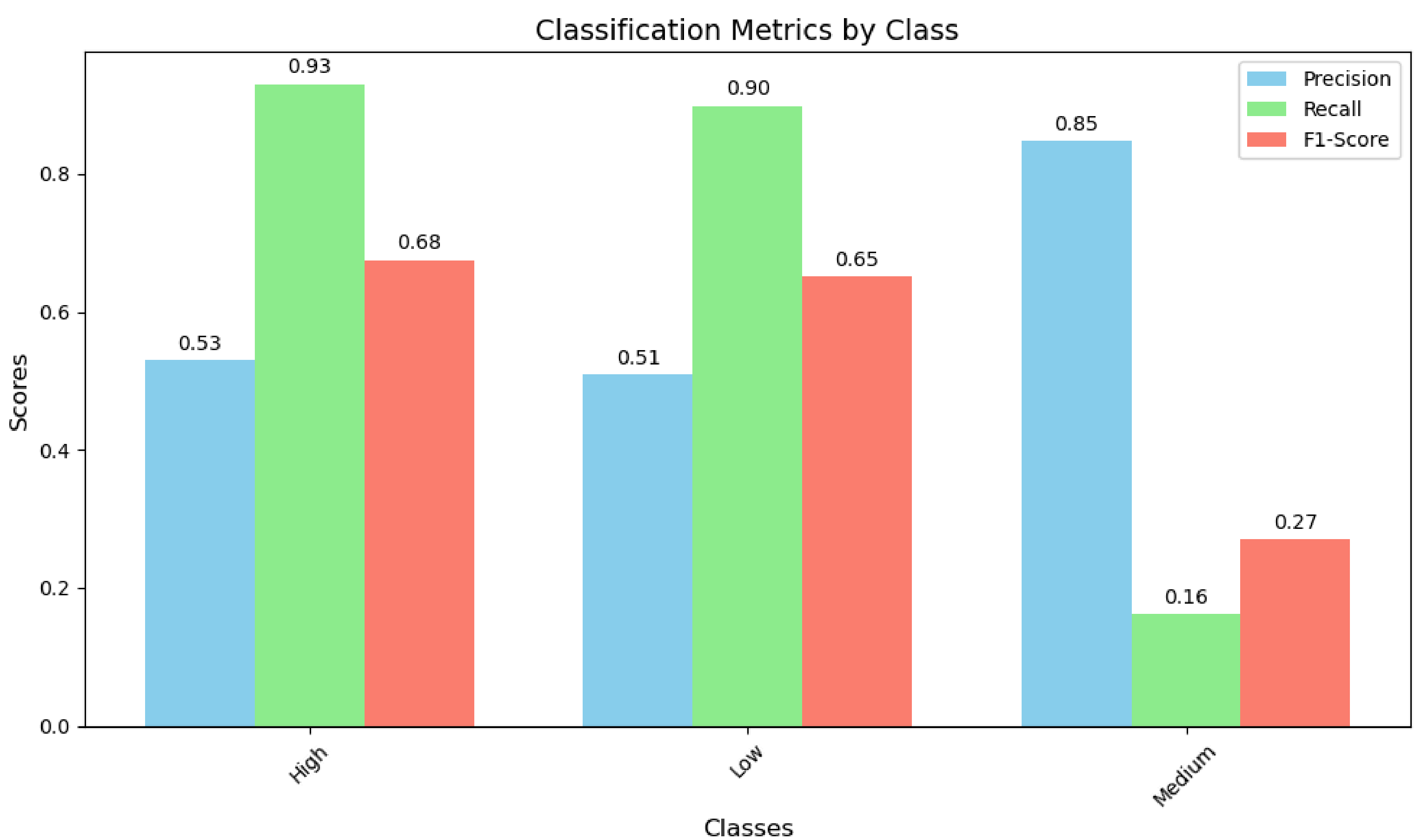
- Predictions were made for each test data point using the sigmoid function and the trained parameters.
- The predicted class was chosen based on the highest probability score from the three logistic regression models.

Metrics:

1. **Accuracy**: Measure of overall correctness of the predictions.
2. **Precision**: The proportion of correctly identified positive instances for each class.
3. **Recall**: The proportion of actual positive instances that were correctly identified.
4. **F1-Score**: The harmonic mean of Precision and Recall for each class.

Evaluation Plots

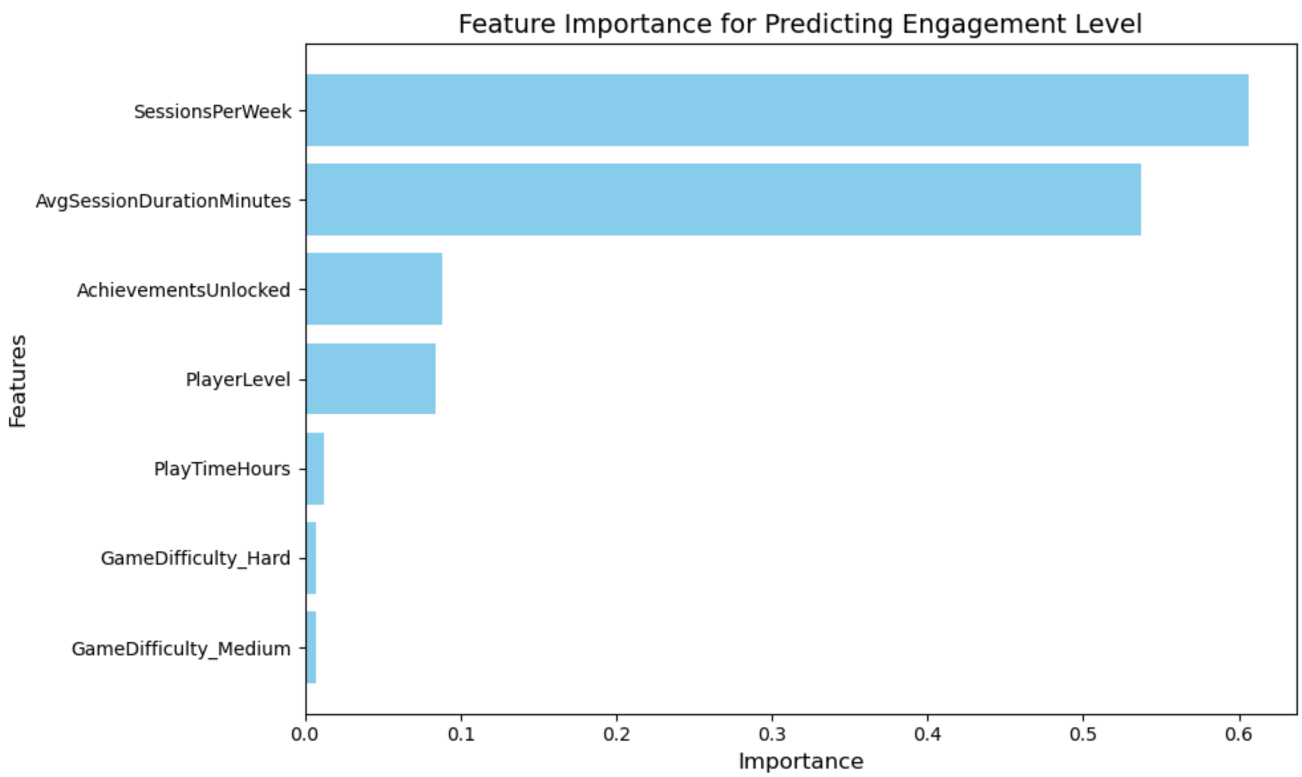
- A bar chart was created to visualize the precision, recall, and F1-score for each engagement class (High, Medium, Low).



Feature Importance

- Feature weights were extracted to analyze their contribution to predicting engagement levels.
- A bar chart was created to display the importance of features such as PlayTimeHours, SessionsPerWeek, AchievementsUnlocked, etc.

Plot:



6. Conclusion

Key Findings:

- **In-Game Purchases**: Players aged 41 and those from Asia/Europe are more likely to make purchases.
- **High Engagement Levels**: Factors such as sessions per week, average session, and achievements unlocked are critical.
- **Game Genres**: Simulation games dominate in terms of retention and popularity among high-engagement players but only slightly

Applications:

- These insights can be used to:
  - Optimize game design and difficulty levels.
  - Develop targeted marketing strategies for specific demographics.
  - Enhance player retention and engagement through personalized recommendations.