Online Gaming Behavior Analysis

Gregory A. Martin

Western Governors University

Table of Contents

A. Project Highlights

- This project aimed to answer a business question for a fictitious video game developer. The business need was looking for a correlation between higher levels of player engagement and in-game purchases. However, a correlation could not be found, but I was able to provide some other relevant insights derived from the data that could be used by video game developers to draw conclusions necessary to increase profitability. The overall goal was to find the best way to increase profitability based on what the data said. This was accomplished by reading a data set of player-generated data, performed cleaning steps, analysed the data using exploratory analysis, looked for correlations, and finally detailed / visualized the results. The report concluded that a correlation could not be found, however valuable insights were provided which included top selling genres and the most profitable countries.
- The scope of this project is a single python script. The script read in the dataset, performed cleaning steps, and provided formatted data and several visualizations using a number of python libraries. Since there was no need for machine learning algorithms, we did not use any libraries for that, because that is considered to be out of the scope of this project.
- This project used the Agile methodology. Agile proved to fit this project well due to the exploratory nature of the project. This helped the developer make informed decisions as more information became available during analysis, and allowed me to make adjustments along the way. One of the adjustments that was made was the type of visualizations used. Initially, I had planned on using a catter plot with a trend line. But since I could not find a correlation in the data, I opted for bar charts and histograms. Python and Jupyter Notebooks were the tools used for this project. These tools provided a great environment for the required analysis.

B. Project Execution

- The goal of this project was to create a Python application that will give insight into player behavior for the purpose of maximizing profit and revenue.

The objective was to perform exploratory data analysis and visualizing the results.

The deliverables were the resulting data output and visualizations.

  - Goal: create a Python application that will give insight into player behavior for the purpose of maximizing profit and revenue.
    - Objective: perform exploratory data analysis and visualizing the results.
      - Deliverable 1: Data output
      - Deliverable 2: Visualizations

- This project originally stated that it would use the Agile methodology. It was believed at the time that agile would fit this project well due to the exploratory nature of the project. However, due to the small variance in the data, waterfall methodology was found to be more efficient for this project. This was a deviation from the original plan stated in my earlier proposal.

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Task 1	1 day	*July 1st*	*July 2nd*
Task 2	3 days	*August 1st*	*August 4th*
Task 3	3 days	*August 5th*	*August 8th*

C. Data Collection Process

- The data was collected by downloading the .csv file from https://www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset. The collection process did not deviate from the original plan.
- The quality and completeness of the data was good. There was little to no cleaning required which means more time was spent on doing the analysis. However, one issue to note with the data is that the *InGamePurchases* column is of a binary data type. The data is either a 0 or a 1, which means either the player has spent money, or has not. This provided a challenge when attempting to plot the data related to purchases. But we were still able to derive meaningful insights and had little impact on the outcome and overall success of our analysis.
- The data is public and freely available to anyone without any restrictions. There is no concern for governance, privacy, security, ethical, legal or regulatory issues. There is some demographic information included such as Gender, Age and Location (country), but there is no concern as these are not considered PHI or PII, so each player's privacy is protected.

C.1 Advantages and Limitations of Data Set

- One advantage of the data set is that it appears mostly clean. There was not really any cleaning steps required. For example, when checking for null values, there were none.
- One major disadvantage of the data set is the InGamePurchases column is of a binary data type rather than a cumulative total. For example, the column contains either a 0 or 1, which either means the player made a purchase or did not. This was overcome by splitting the data set into two subsets – those with purchases and this without purchases. This allowed us to more easily assess the subset of data with purchases.

D. Data Extraction and Preparation

The data was collected by downloading the .csv file from https://www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset by clicking on the link then clicking the download button. This required a Google account, which I signed in with my personal Google account. Once the data was downloaded it was then extracted from the zip file, renamed to "dataset.csv", then moved to the project folder on my personal laptop. Then I used df = pd.read_csv('dataset.csv') to load the data into the python script and assign it to the df variable. This extraction process was appropriate for the project due to the nature of downloading files from a place such as Kaggle.

E. Data Analysis Process

E.1 Data Analysis Methods

To start, I used df.corr to see what the correlation score is for each column compared to every other column. There was not a strong correlation score for any two data points. Then I split the data set into two subsets – one containing purchases and one with no purchases, so that further assessment and comparison could be done. I used a variety of tools to look at player statistics and averages, comparing them between the two subsets. I used tools such as df.value_counts(normalize=True) and df.mean() to measure player EngagementLevel and PlayTimeHours as a percentage and an average, respectively. I was able to determine probability of purchases of each gender by dividing the number of purchases by their respective gender total players. For example, divide the number of males who made purchases by the total number of male players. And of course the same for female players. I also used tools such as df.groupby() in order to group data together. I did this in cases such as Age, and also for location. This was especially helpful for location because it gave us a great insight into which countries were the most profitable in terms of having the most in-game purchases. These were appropriate for the project because we were looking for data trends and ways to maximize profitability.

E.2 Advantages and Limitations of Tools and Techniques

Python and Jupyter Notebooks will be the tools used for this project. These tools provide a great environment for the required analysis. One advantage is the flexibility of the jupyter notebooks code blocks with the formatting capabilities make for a presentable report. One disadvantage is everything is stored on the local machine, so manual backups must be made. As for the analytical tools used, these were necessary as part of exploratory analysis is utilizing a number of various libraries and function calls in order to thoroughly examine the data. The advantages of using these various tools is there was a number of interesting and helpful insights that were discovered. A limitation of the chosen tools is there is no predictions being made, since we are not using predictive analysis or machine learning algorithms.

E.3 Application of Analytical Methods

I used descriptive analysis as the analytical method in this project. I used python and jupyter notebooks, and various different python libraries. I used exploratory data analysis to work through the data set and describing what the data had to say in the form of various data output, visualizations and data inisghts. Here are the steps I used to perform the analytical analysis:

- Import the data
- df.head() to preview the data
- df.isnull().sum() to view any null data to see if cleaning is required

- I then proceeded to use a number of libraries and function calls as part of exploratory analysis. These include df.value_counts() and df.corr().
- I then split up the data into two subsets using the following code:
    - df_nop = df[df['InGamePurchases'] == 0] #No purchases
    - df_p = df[df['InGamePurchases'] == 1] #Purchases
- I used df_p.head() and df_nop.head() to preview the two subsets
- I used the following code to view counts of purchases by each genre:
    - genre_counts = df_p.groupby('GameGenre').size()
    - print(genre_counts)
- Then I used df.value_counts(normalize=True) on each subset
- Then I used plt.bar to plot the engagement level as a bar chart
- Then I used a number of mathematical function calls to list player statistics without purchases, including df.mean(), df.max(), and df.value_counts()
- Then I used plt.hist() to view average playtime distribution
- Then I used statistics.stdev() to calculate the standard deviation
- I repeated the above steps again for players with purchases
- Then I used division to calculate purchase probability of each gender
- Next I used plt.bar() to plot the purchase probability of each gender
- Then I used the following code to group the data by Age and df.mean() to view multiple player playtime averages at once:
    - df_p_grp = df[df['InGamePurchases'] == 1]
    - df_p_grp = df_p_grp.groupby('Age').mean(numeric_only=True)

    - df_p_grp = df_p_grp[['PlayTimeHours', 'SessionsPerWeek', 'AvgSessionDurationMinutes']]
- Then I used plt.bar() to plot the data
- Next I used the following code to group the data by Location, sum the values, sort the values, and plot the results in a horizontal bar chart:
    - df_p_grp_loc = df_p.groupby("Location").sum(numeric_only=True).reset_index()
    - df_p_grp_loc = df_p_grp_loc[['Location','InGamePurchases']]
    - df_p_grp_loc.sort_values(['InGamePurchases'], ascending=[True], inplace=True)
    - df_p_grp_loc.plot(kind='barh', x='Location', y='InGamePurchases')
- Finally I carried out two Hypothesis tests using stats.pearsonr()

F Data Analysis Results

F.1 Statistical Significance

- My null hypothesis is that there is no correlation between in-game purchases and player engagement.
- The statistical test that will be used is a Pearson correlation test.
- The test will return a p-value score and correlation coefficient.
- If p-value is greater than 0.05, then there is statistical significance, otherwise if the p-value is less than 0.05, then there is no statistical significance.
- The *alpha* value will be set at 5% or 0.05
- The first hypothesis test was done to look for correlation between PlayTimeHours and InGamePurchases. However, due to the nature of InGamePurchases being a binary data

type, the results were inconclusive as both the p-value and correlation coefficient were 'nan'. Therefore, there is not enough evidence to support there being a correlation between these two data points.

- An alternatie hypothesis test was done to look for correlation between PlayTimeHours and Age. Even though the p-value was 0.6223, we can see a very small correlation coefficient of 0.0025. This means there is a very weak positive correlation between PlayTimeHours and Age.

F.2 Practical Significance

The practical significance of these results will create real world benefit to the company in the form of revenue increase. For example, we see from our analysis that the most in game purchases were from strategy games in the US and Europe markets. Therefore, we can conclude that the developers can increase revenue the most by releasing Strategy games in the US and Europe video game markets.

F.3 Overall Success

Based on the results above, we can conclude that the company would be able to use these results to increase revenue. By creating strategy games for the USA and Europe locations, we can maximize profits for the developers. Therefore, we can conclude the task was a success.

G. Conclusion

G.1 Summary of Conclusions

Through exploratory analysis of the chosen data set, here is what we learned:

- We could not find a positive correlation in the data.

- The hypothesis and alternate hypothesis were rejected. However, this is not a bad thing and just reinforces the importance and benefits of thorough analysis.
- We still gained actionable inights that have a real world benefit to the company in the form of revenue increase.

G.2 Effective Storytelling

The first visualization shows engagement levels of players with purchases. This is impactful because it answers an important question of whether players who make purchases are engaged at a higher level. It is a bar chart.

The next two visualizations show distribution of play time between players with purchases and players without purchases. Similar to engagement level, however it is simply another metric to look at to see if there is much of a difference between the two types of players. They are histograms.

The next visualization shows the probability of each gender (male and female genders only) making a purchase. This was an important metric to determine because it was an alternate hypothesis. It is a bar chart.

The next visualization is showing average play time by age. It is another demographic metric to consider the age of your players and how much time they spend playing. It is a bar chart.

The last visualization shows purchases by location. This is impactful because we can clearly see where majority of the purchases are coming from. This was one of the important insights gained from the analysis. It is a horizontal bar chart.

G.3 Recommended Courses of Action

Based on the analysis, I have two recommendations:

1. Develop strategy games for the USA and Europe video game market

2. Create them to appeal to all age groups and genders

H Panopto Presentation

Please see my Panopto video using the link below:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b2dde372-8fb2-4849-8ee4-b1cb00441a39

Appendix A

Python code will be attached with the submission of this file