MARKETING DATAMART - Manual

Group 6 : Rahul MAHESHWARI & DU Kang

Contents

0.1 Approach

The implementation has been done primarily using the `data.table` package along with a little use of `dplyr` package. `readxl` and `ggplot2` packages have been used to make this report manual. We started off with reading all the four csv files, and cleaning them. Only a couple of `Gender` values were missing which were assigned 0 (female). All the date columns were converted into R `Date` objects to allow calculations on dates. From the analytic dataset, only the `Age` column was taken since this information is not present in Demographics dataset. The first step was to combine the Poker transaction and User Agrregation datasets to facilitate a uniform analysis across all products. For this purpose, the poker transactions were aggregated per day after stripping off the time portion and the table was restructured to make it similar to the User Agrregation table. All the buy transactions per day were summed as stakes and the difference between buy and sell amounts per day were computed as the winnings for that day. The number of buy transactions per day were counted as the number of bets for that day, and a new column was added called `ProductID` with a constant value of 3. Then the poker table was combined with the User Aggregation table using `rbind`, and this combined table was joined with the Demographics table on `UserID`. As per the methodology in the codebook, only the users with registration date between 1 and 27 february, and transactions done after `FirstPay` date were retained. Total active days per product as well as overall, first and last active dates per product as well as overall were calculated in this table, along with the total stakes and winnings per product as aggregations of individual transactions amounts. This table was joined with the Age data on the `UserID` to create a Datacube. User segments were created and assigned based on the gender and age of the user. The segments were ranked based on the aggregated profits made over all the users in that segment and the rank was inserted back into the Datcube. The Datacube at this point had one row per user per product he/she played with all the information needed to create a Datamart. Global metrics were calculated from the Datacube and put in the new Datamart table. Product-wise metrics were calculated and put as individual columns in the Datamart instead of rows, such that each product has a dedicated column per product metric and each user has just one row in the Datamart. This Datamart table was then written into a CSV file with numerical figures formatted as euros or % as applicable, and all `NA` values replaced with 0s representing missing data. The User Segment matrix was written into a CSV file as well for reference. All the temporary tables were removed throughout the code to keep the memory footprint to a minimum.

0.2 Structure

The Datamart consists of the following main columns :

- **UserID** - *[From Demographics table]* Unique identifier of the user assigned at the time of registration.

- **Country** - *[From Demographics table]* User's country of residence.

- **Language** - *[From Demographics table]* Primary language of the user.

- **RegDate** - *[From Demographics table]* Date of registration of the user in the mm/dd/yyyy format.

- **FirstActDate_Overall** - *[From Transaction tables]* User's first active play date.

- **LastActDate_Overall** - *[From Transaction tables]* User's last active play date.

- **TotalActDays_Overall** - *[Calculated]* User's total active days of play between the first and last active date.

- **Ranked Segment** - *[Calculated]* Profitability rank of the segment to which the user belongs. See *Metrics* section for details.

- **First Activity Lag (days)** - *[Calculated]* Gap between the date of registration and first active play date of the user.

- **Total Products Played** - *[Calculated]* Count of the number of different types of products the user has played.

- **Overall Playing Frequency** - *[Calculated]* Ratio representing how much the user has been active in between his first and last active play date.

- **Favorite Product** - *[Calculated]* ID of the product which the user has played most frequently.

- **Overall Stakes** - *[Calculated]* Total stakes the user betted on across products. Also the total revenues for bWin from the user.

- **Overall Winnings** - *[Calculated]* Total winnings the user received on all products. Also the total cost of user for bwin.

- **Overall Profit Margin** - *[Calculated]* Total profit margin from the user for bwin.

- **Lifetime Value (Indicative)** - *[Calculated]* Descriptive approximate lifeime value of the user for bwin over his active play period.

- **P(i) B/D Ratio** - *[Calculated]* Bet to Deposits ratio of the user for a particular product. There are 8 columns for each of the 8 products. If the user has not played the product, the value for the column is 0.

- **P(i) Profit Margin** - *[Calculated]* Proft Margin from the user on a particular product. There are 8 columns for each of the 8 products. If the user has not played the product, the value for the column is 0.

- **P(i) Playing Frequency** - *[Calculated]* Ratio representing how much the user has been active on a particular product in between his first and last active play date for that particular product. There are 8 columns for each of the 8 products. If the user has not played the product, the value for the column is 0.

*Note - If a user has not played a particular product, it means there is no data for that user on that particular product and all the last three P(i) column have 0 in such cases.

0.3 Metrics Description

Two types of marketing metrics have been provided in the Datamart : *Global* (per user across all products) and *Product* (per user per product). This section describes their meaning and how they have been calculated in the order they appear in the Datamart :

1. **Total active days overall** - This gives a general idea about the duration for which different users have been active on bwin over the period of study. This has been calculated as the count of unique dates in the Poker transaction and User aggregation tables combined.

2. **Ranked segment** - User segments have been created by a combination of gender and usual phases of life, and assigned to each user. The segments were ranked as per their profitability (Total Stakes - Total Winnings of all users under that segment) and the ranks were directly put in the Datamart. The User segment matrix csv should be referred to map the segments and their ranks. Grouping by profitable segments would allow the marketing to focus on and target each group of users in a customized way. The user segment matrix created is as follows :

User Segment	Profitability(euros)	Rank
<=24 M	2,703,482	3
<=24 F	301,999	6
25-39 M	5,548,491	1
25-39 F	328,047	5
40-64 M	2,205,701	4
40-64 F	160,852	7
>=65 M	47,671	9
>= 65 F	14,344	10
Unknown_Age M	3,716,525	2
Unknown_Age F	76,777	8

M - Male, F - Female.
Unknown_Age - Age data not available from analytic dataset.

3. **First activity lag** - This is the lag each user exhibited in starting to play on bwin after they registered, and is essentially represents opportunity cost for bwin. The marketing should look at users with high lag and communicate with them to try make them play and turn profitable as soon as they have registered. This has been calculated as follows:

   ```
   FirstActDate_Overall - RegDate
   ```

4. **Total products played** - This gives a quick insight into the range of products each user played. More the number of different products tried, less skilled the user is in any

one particular product potentially or not winning much in any particular product. This can be analyzed together with the profits made by the user in each of the products to find a possible correlation.

5. **Overall playing frequency** - This metric indicates how much the user has actually been playing and winning or losing money. Higher the ratio, the more addicted the user is to online gambling and has better prospect of being a loyal customer. This has been calculated as follows:

   `TotalActDays_overall / (LastActDate_Overall – FirstActDate_Overall + 1)`

6. **Favorite product** - This indicates the product which the user has played the most, and represents a level of addiction of the user to the particular product. The marketing can customize offers and promotions related to the most favorite product for that user if the user is profitable for bwin in that particular product or try to incentivize to make the user try other products as well. This has been calculated as:

   `P(i) with max(Playing Frequency of the user)`

7. **Overall stakes** - This represents the total revenue from each user during his/her entire duration of activity, and is calculated as:

   `Sum(Total Stakes on all P(i))`

8. **Overall winnings** - This repesents the total cost for bwin of maintaining each user, and is calculated as:

   `Sum(Total Winnings on all P(i))`

Note - The overall cost of a user also includes administrative and logistics cost but that data is not available.

9. **Overall profit margin** - This is the most direct indicator of how profitable a user has been so far for bwin. Postive margin indicates bwin is making money from the user whereas negative means bwin is losing money to the user. Lower margin indicates the cost to serve that user is high. The marketing can focus only on the most profitable users, and in conjunction with the other metrics, try to encourage users with low margin to play more. This has been calculated as:

   `[(Overall Stakes – Overall Winnings) / Overall Stakes] * 100`

Note - Users with zero stakes but non-zero winnings indicate they have used only the promotional money offered by bwin and won on that. In such cases margin is not possible to calculate and has been indicated as 0 in the datamart.

10. **Lifetime value (Indicative)** - This is a descriptive metric calculated using historical data on the user's activity on bwin and indicates the approximate cash flow bwin can expect from the user over each period of similar duration for which the user was active on bwin, if he/she were to continue playing. Marketing can project this cash flow into future, diminishing it over time based on some parameters, and apply the weighted average cost of capital and the retention probability (out of scope of the Datamart) to get the NPV of the user over the desired projected period. Only the users with postive lifetime are valuable and need to be focused on. Bwin can also try to devise strategy to extract additional value from lower valued user, based on the analysis on whether they are more loyal or cost less to serve. The metric has been calculated as follows:

```
[(Overall Stakes - Overall Winnings) / TotalActDays_Overall]
*
[LastActDate_Overall - FirstActDate_Overall + 1]
```

11. **Product Bet / Deposit ratio** - This is a product metric and shows how many times the user circulate each euro/dollar/pound they have deposited. Higher ratio means that it takes more time to lose whilst lower value indicates that bwin is drying them at a faster pace. Deviation from the standard might also mean that the games don't function as they should (RNG problems, etc.), and there might be a need to check the software functionality. This metric can be aggregated over all the users to compare performance of the different products. As the name indicates, this has been calculated for each product as:

```
Total bets on P(i) / Total stakes on P(i)
```

Note - Users with zero stakes but non-zero bets indicate they have placed bets only using the promotional money offered by bwin. In such cases the ratio is not possible to calculate and has been indicated as 0 in the datamart.

12. **Product profit margin** - This is a prodcut metric and is similar to the overall profit margin, except that it is calculated for each product. This metric indicates the margin bwin is getting from the user on each product, and it can be aggregated over all the users to compare performance of the different products. It has been calculated as:

```
[(Total stakes on P(i) - Total winnings on P(i))
/ Total stakes on the product] * 100
```

Note - Users with zero stakes but non-zero winnings on the product indicate they have used only the promotional money offered by bwin and won on that. In such cases margin is not possible to calculate and has been indicated as 0 in the datamart for that product.

13. **Product playing frequency** - This is a product metric and is similar to the overall playing frequency, except that it is calculated for each product. It indicates how much the user has actually played each product, and has been calculated as:

TotalActDays on P(i) / (LastActDate on P(i) - FirstActDate on P(i) + 1)
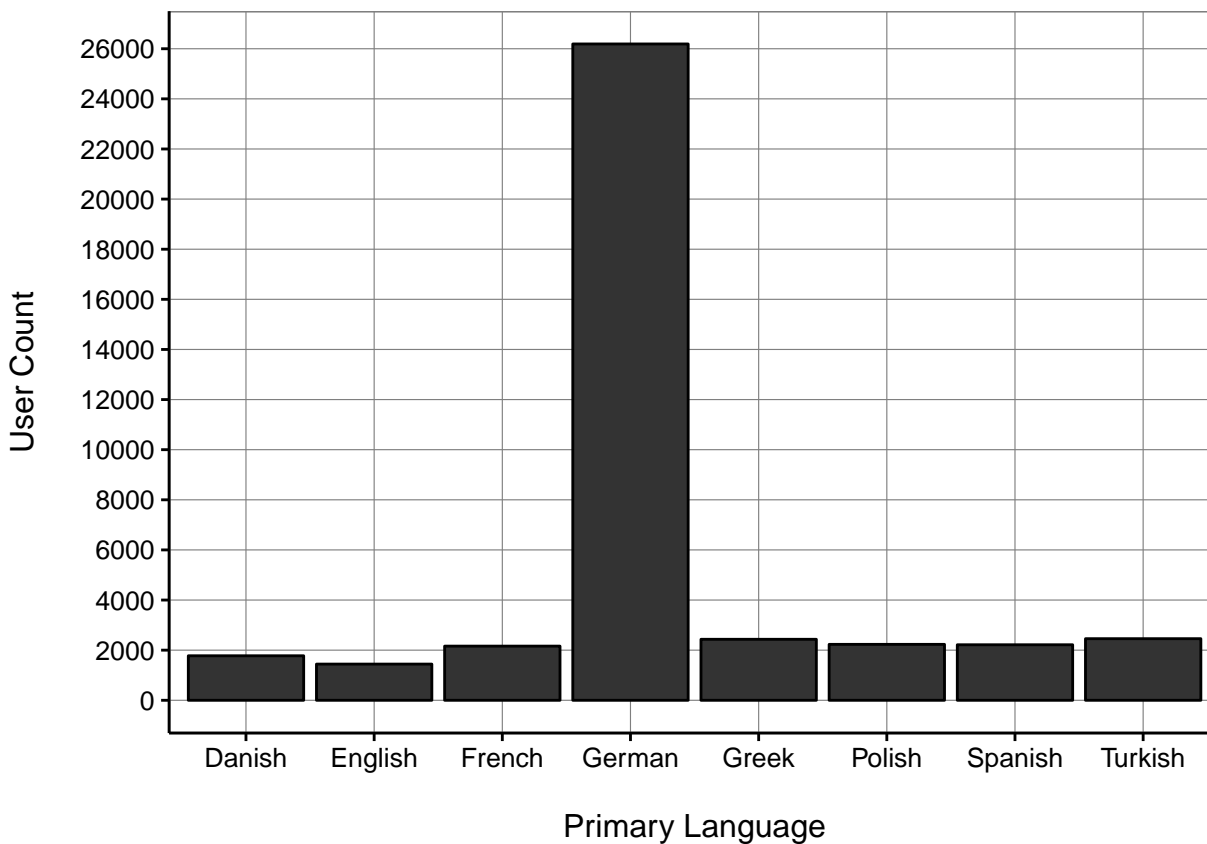

0.4 Summary Statistics

Total number of users : **42612**.

All the users belong to one of the **81** countries with a maximum of **24155** users from **Germany**. The following plot shows the top 10 countries :
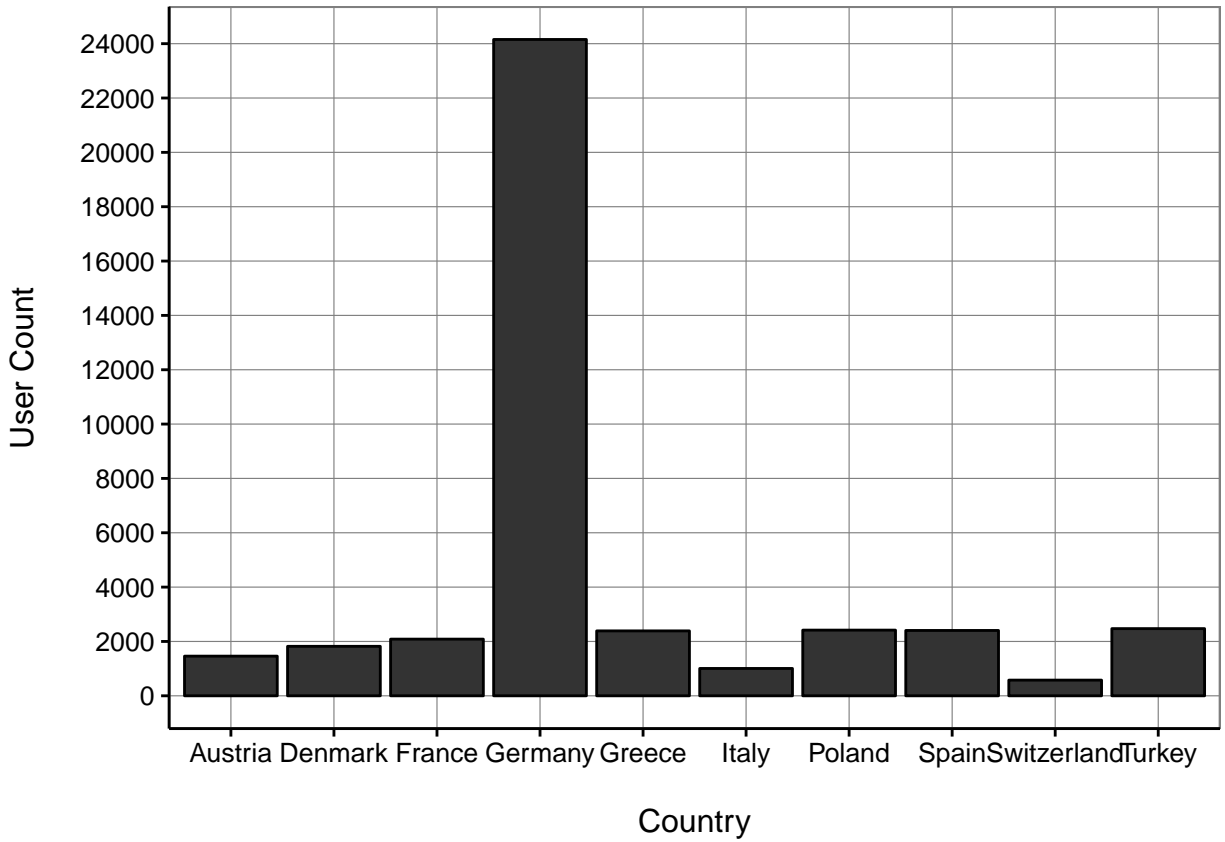
All the users speak one of the **17** languages with a maximum of **26191** users speaking **German**. The following plot shows the top 9 languages :

The users were grouped into 10 segments based on their age and gender and the segments were ranked according to their profitability. The segment with highest number of users is **25-39 M** with **19714** users.

The following plot compares the user count and median playing frequency of the top 5 profitable user segments :
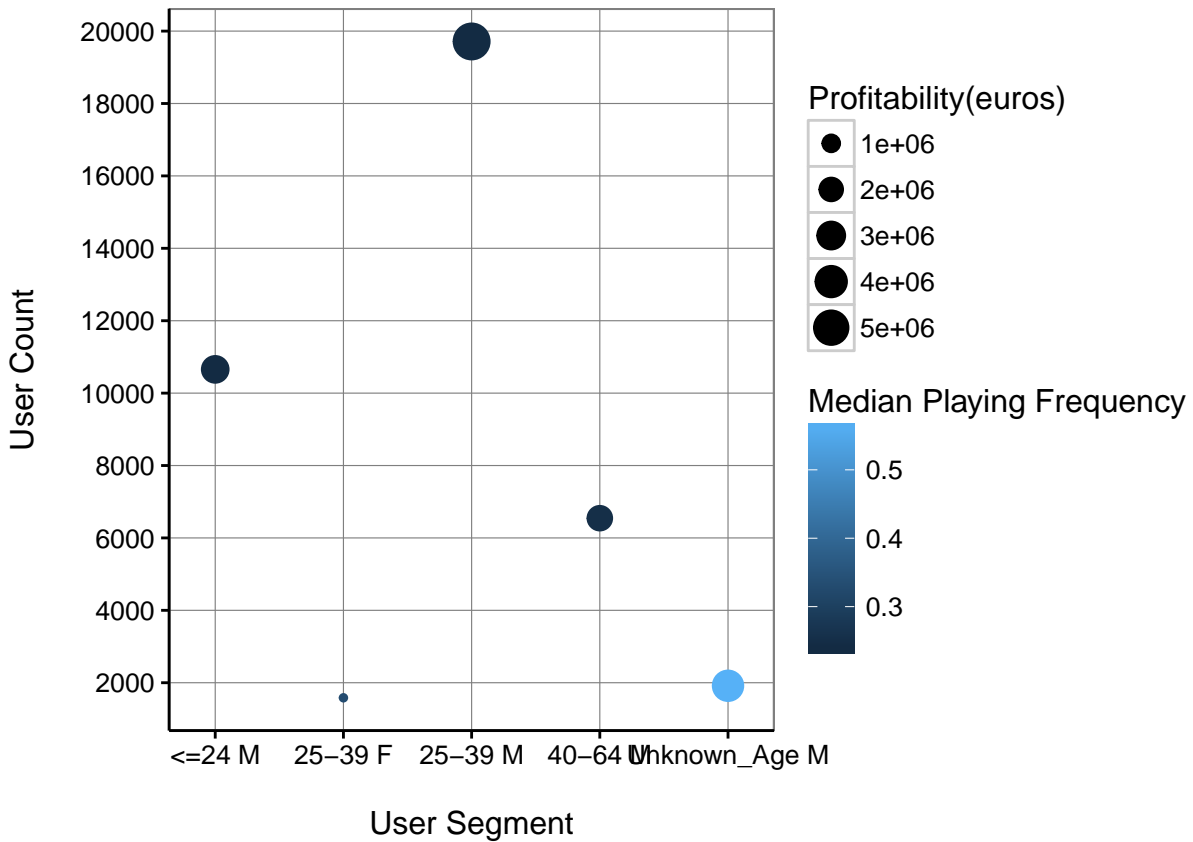
Apparently the segment which played most actively are the males whose ages are not known.

Statistics on the total days for which the users have been active :

Mean	Median	Minimum	Maximum
27.04	15	1	234

Statistics on the gap (in days) between the registration day of the users and the day they first started playing :

Mean	Median	Minimum	Maximum
14.4	2	0	239

Statistics on the total number of different products (out of 8) played by the users :

Mean	Median	Minimum	Maximum
1.88	2	1	8

Statistics on the overall playing frequency of the users which represents how many days out of their range of active dates they have been actually played :

Mean	Median	Minimum	Maximum
0.35	0.25	0	1

The most favorite product of a user is the one having highest playing frequency. The following plot shows all the products and the count of users who played them most frequently :
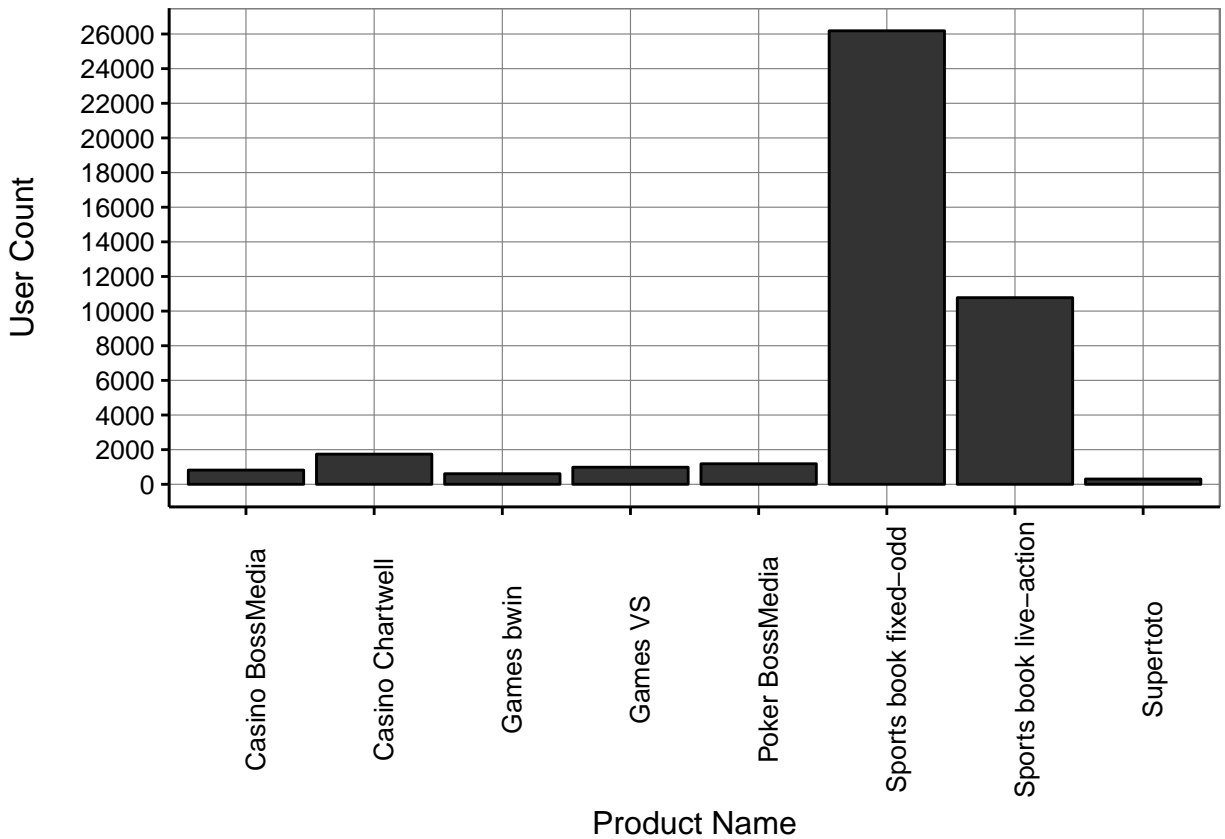
The most favorite product among all the users is apparently **Sports book fixed-odd** with **26185** users playing it most frequently, followed by **Sports book live-action** with **10776** users. This could be because Sports book products are low risk or do not need much skill to play.

Statistics on the revenue from the users (in euros) which is the stake deposited by them :

Mean	Median	Minimum	Maximum
2960.17	241	0	2259812

Statistics on the cost of users (in euros) for the company which is the user winnings :

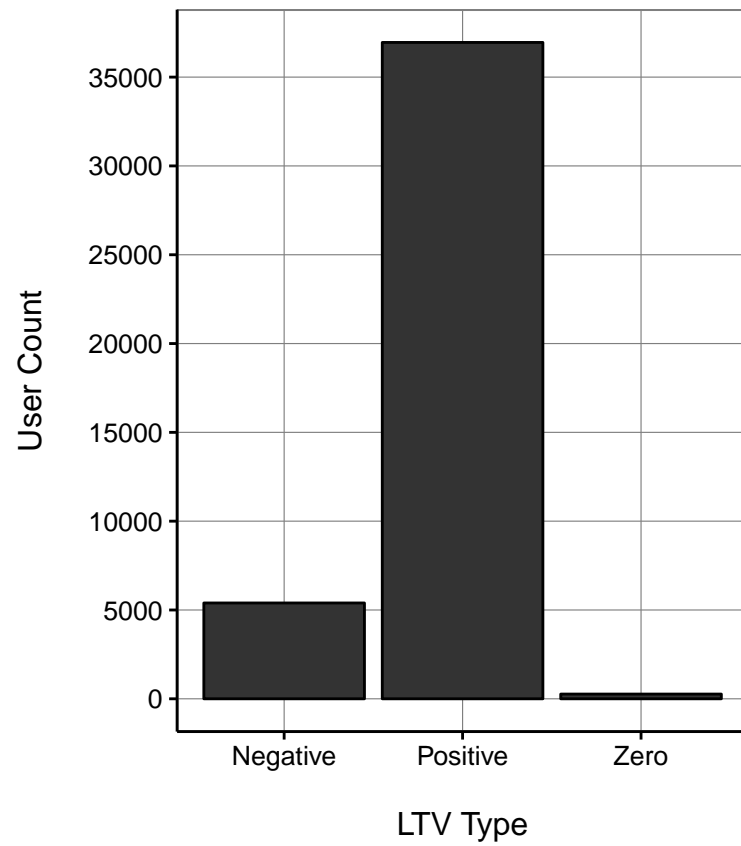| Mean	Median	Minimum	Maximum
2605.72	181	0	1093423

Statistics on the profit margin (in %) from the users for the company :

Mean	Median	Minimum	Maximum
21.72	22.7	-212700	100

Statistics on the approximate descriptive lifetime Value (in euros) of the users :

Mean	Median	Minimum	Maximum
1169.29	205	-315250	2997701

The following plot shows the count of users with positive, negative and zero lifetime values :

Statistics on **Product 1 (Sports book fixed-odd)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	0.38	0.23	0.0	42.00
Profit Margin (%)	32.18	29.20	-2834.9	100.00
Playing Frequency	0.32	0.23	0.0	1.35

Statistics on **Product 2 (Sports book live-action)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	0.42	0.26	0.00	27.15
Profit Margin (%)	23.32	17.80	-2140.00	100.00
Playing Frequency	0.43	0.28	0.01	1.80

Statistics on **Product 3 (Poker BossMedia)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	0.17	0.08	0.00	4
Profit Margin (%)	-226.72	56.80	-212700.00	100
Playing Frequency	0.51	0.43	0.01	1

Statistics on **Product 4 (Casino BossMedia)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	0.14	0.030	0e+00	11
Profit Margin (%)	12.67	8.300	-4e+02	100
Playing Frequency	0.55	0.475	1e-02	1

Statistics on **Product 5 (Supertoto)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	1.32	0.73	0.01	10.5
Profit Margin (%)	17.26	16.70	0.00	100.0
Playing Frequency	0.53	0.35	0.01	1.0

Statistics on **Product 6 (Games VS)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	0.82	0.75	0.00	10

Metric	Mean	Median	Minimum	Maximum
Profit Margin (%)	29.64	25.00	-1982.40	100
Playing Frequency	0.61	0.73	0.01	1

Statistics on **Product 7 (Games bwin)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	1.06	1.00	0.05	2.5
Profit Margin (%)	31.19	27.05	-575.00	100.0
Playing Frequency	0.66	1.00	0.01	1.0

Statistics on **Product 8 (Casino Chartwell)** :

Metric	Mean	Median	Minimum	Maximum
B/D Ratio	0.84	0.5	0.00	54.75
Profit Margin (%)	19.74	11.0	-780.00	100.00
Playing Frequency	0.56	0.5	0.01	1.00

The following plot compares the performance of the 8 products in terms of median values of the B/D Ratio observed across the users :
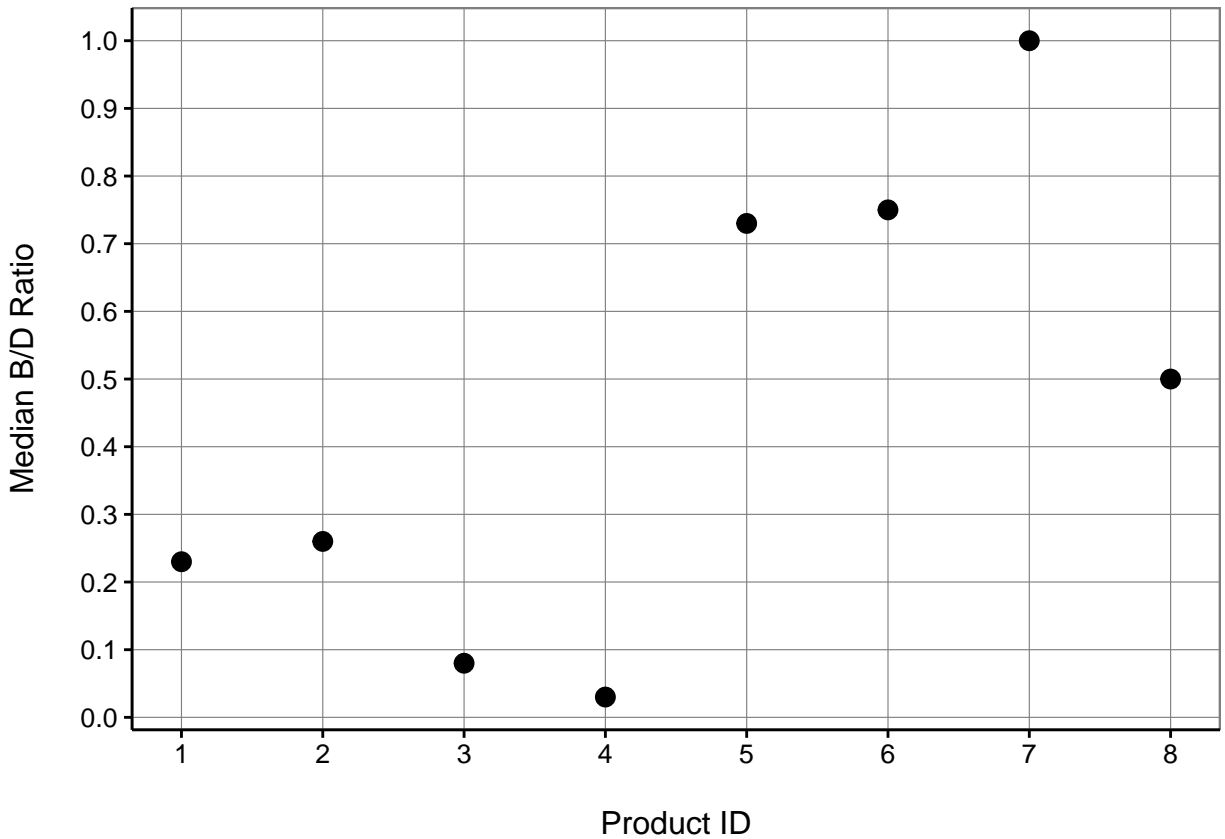

Products 5,6,7 & 8 appear to have shown higher B/D rations than rest of the products, which implies that users take more time to lose on these games. Bwin might need to check these with the standard values to see if the games are functioning properly.

The following plot compares the performance of the 8 products in terms of median values of the playing frequency exhibited by the users :
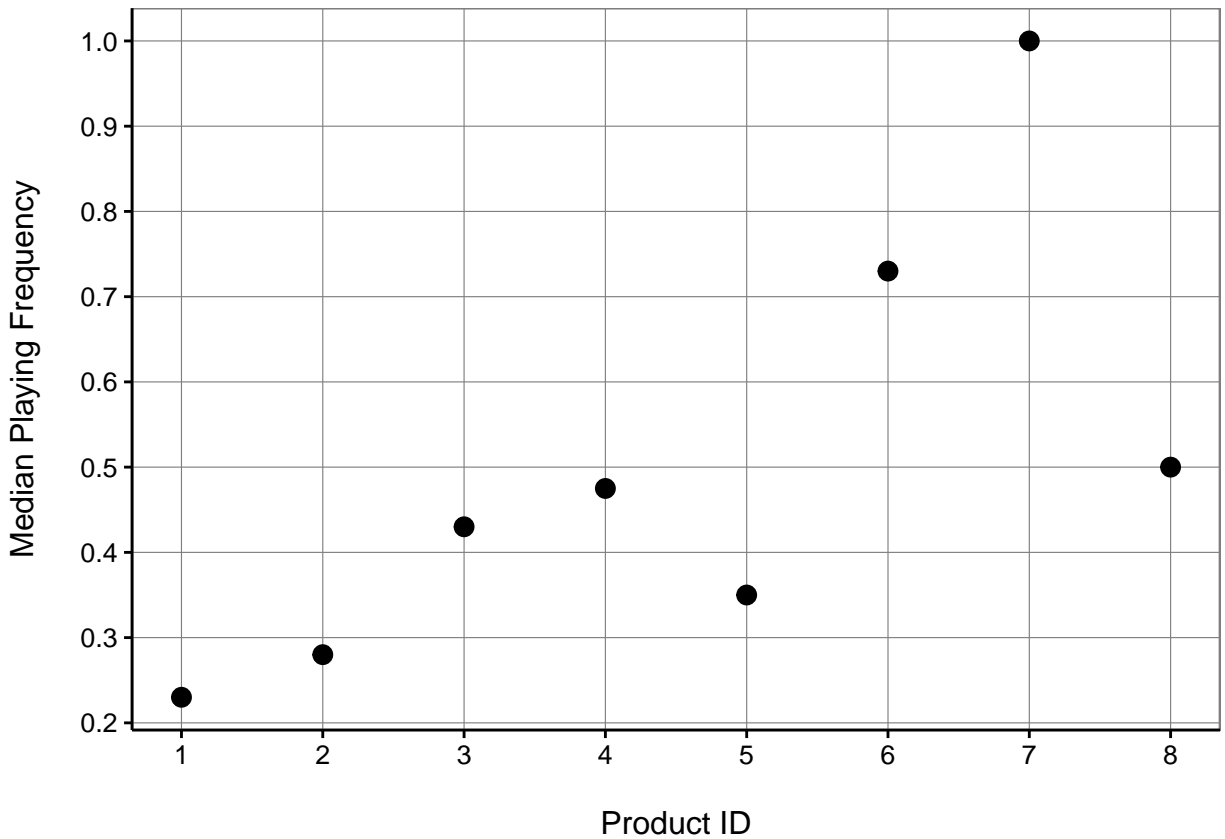
Games (6 & 7) appear to have the highest median playing frequency, followed by Casino (4 & 8) and sports book (1 & 2) neing the lowest. Since the number of users who have played these products is significantly lower than other products, this implies that most of the users who did use these products, played them more actively than they played others.

The following plot compares the performance of the 8 products in terms of median values of the profit margin obtained on them :
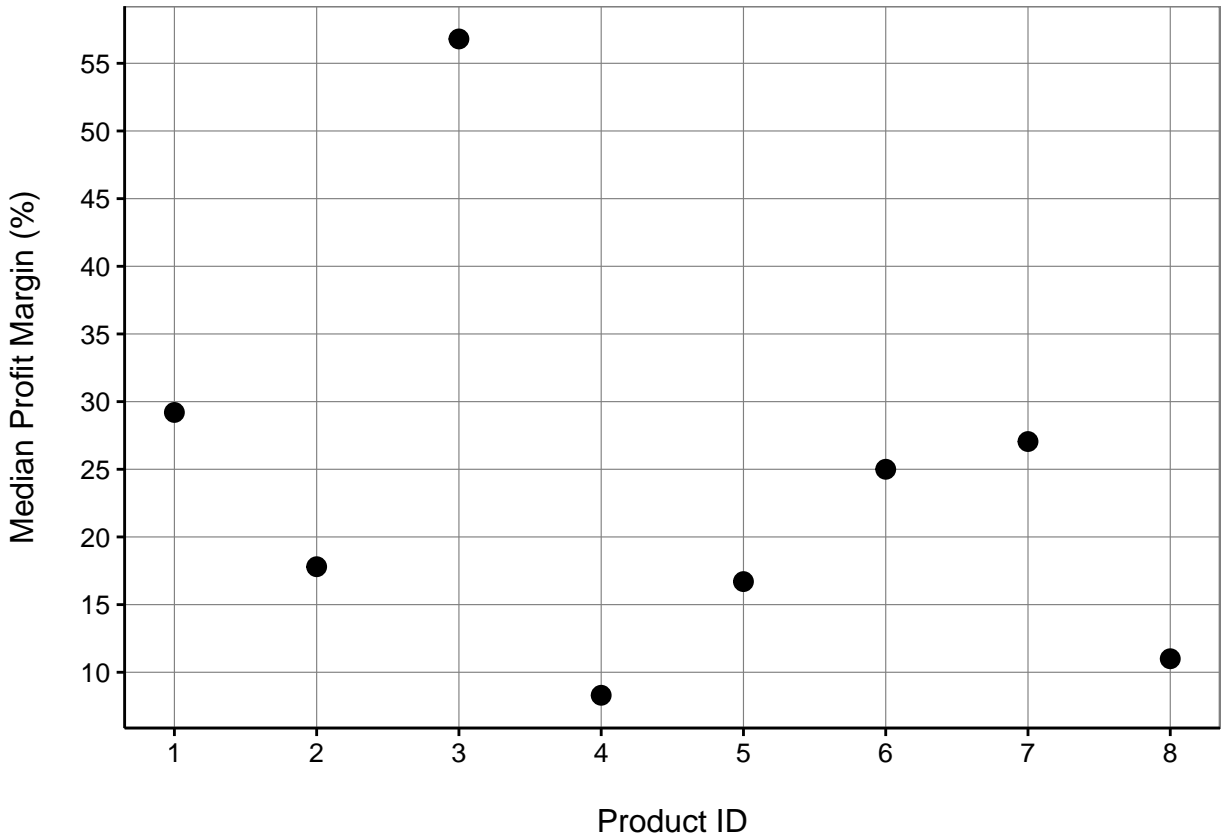
Poker (3) appears to be the highest margin product, followed by Games (6 & 7), Sports book (1 & 2), Supertoto (5) and Casino (4 & 8).