MARKETING DATA MART

GROUP 4: Manju Nagaraj Rudrappa, Lucas Bonett, Dafni Krystallidou

Contents

---

1. Raw Data

This markdown will show the different steps we have gone through in order to draw insights from the raw data on bwin Internet betting datasets. These datasets provide evidence for an eight month period of betting behavioral habits from different users.

- **First dataset** - Includes the `demographic`information of the users using several variables.
- **Second dataset** - Shows `UserDailyAggregation` and contains the actual betting information for each user per day.
- **Third dataset** - Contains information on the `pokerchip`, summarizing poker play through amounts and frequency of transactions to and from the poker site.
- **Last dataset** - Is the `analytic dataset`, gathering all information on live and fixed-odds sport games over an eight month period.

---

2. Approach

The implementation has been done primarily using the dplyr package along with a little use of tidyr package, readxl, sas7bdat and ggplot2 packages which have been used to render this report manual.

We started with reading in all the four SAS files, and cleaning them.

The following steps were performed while cleaning the Demographic dataset:

1. Rows (corresponding to customer IDs) with First pay date outside the period from feb/1/2005 to sept/30/2005 were removed.

2. Country, application, gender and language ids were replaced by their description names.

3. All date columns were converted to date format.

The following steps were performed while Cleaning the Analytical dataset:

1. All date columns were converted to date format.

2. All numeric NA values were converted to 0.

The Following steps were performed while Cleaning the Pokerchip dataset:

1. All poker transactions of the customers with transaction date earlier than First pay date were removed.

2. The TransDateTime variable was splitted as seperate columns for date and time.

3. As user_ids were not unique, we created one row per customer and derived some new variables.

4. All numeric NA values were converted to 0.

5. All date columns were converted to date format.

The Following steps were performed while Cleaning the user Daily Aggregation dataset:

1. All sports transactions of customers with transaction date earlier than First pay date were removed.

2. Product ids were replaced by product description.

3. Stakes, winnings and bets with negative values were converted to 0,because the negative numbers were due to accounting correction.

4. As user_ids were not unique, we created one row per customer and derived some new variables

5. All numeric NA values were converted to 0.

6. All date columns were converted to date format.

Finally all 4 datasets were merged into the Final base table:

1. The Demographic dataset was the main table, hence all 3 tables were left joined with demographic dataset using the key user_id.

2. All numeric column NA values were replaced with 0.

3. All date column NA values were kept as it is.

4. All character column NA values were kept as it is.

5. Some interesting marketing metrics were derived from the final base table.

All the temporary tables and variables were removed throughout the code to keep the memory footprint to the minimum

3. Metric Description

1. **Overall active days** - 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. **First activity lag** - This variable represents the time lag each user exhibited when they started to play on bwin after they registered, and is essentially a potential 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.
   The precise calculation is as follows:
   `FirstActDate_Overall - RegDate`

3. **Overall playing frequency** - This metric indicates how much the user has actually been playing (for the sports games), winning or losing money (for the pokerchip game). Higher ratios might potentially indicate users potentially prone to online gambling addiction. This target group might turn into loyal customers.
   This has been calculated as follows:
   `TotalActDays_overall / (LastActDate_Overall - FirstActDate_Overall + 1)`

4. **Favorite product (max_sports_product_played)** - This measure 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 provide additional incentives to make the user try other products as well.

5. **Overall stakes** - This represents the total revenue from each user during his/her entire duration of activity, and is calculated as:
   `Sum(Stakes)`

6. **Overall winnings** - This repesents the total cost for bwin of maintaining each user,and is calculated as:
   `Sum(Winnings)`

7. **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]
   ```

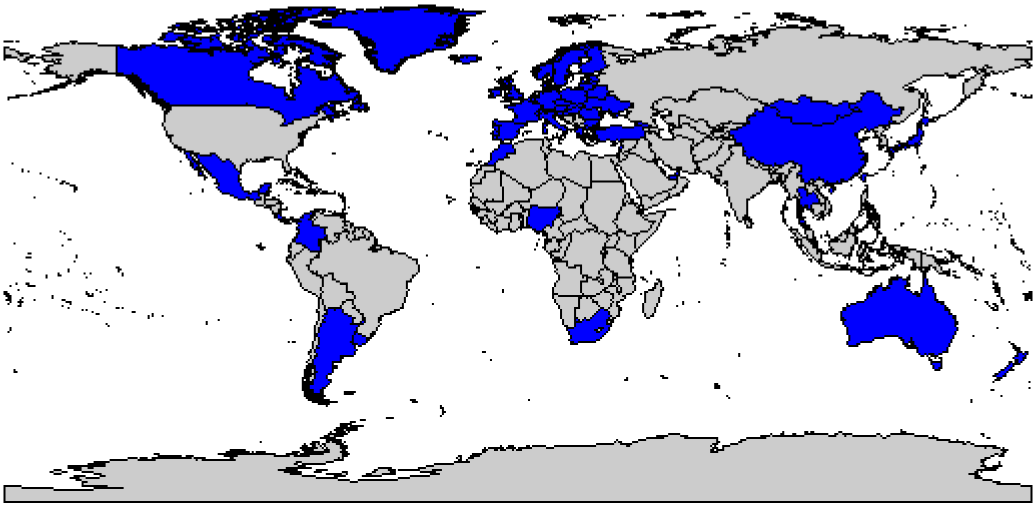
8. **Gross Gaming Revenue** - Gross Gaming Revenue (GGR) is the amount wagered minus the winnings returned to players, GGR is the figure used to determine what the gambling operation earns before taxes, salaries and other expenses are paid.Higher the value, the customer is generating more revenue to the company.
   This is calculated as follows:
   `[Overall stakes - Overall winnings]`

9. **Length of Relationship**
   This variable was calculated by subsrtracting the User's FirstPay date from Last Active Date. This metric provides important insights as long-term loyal cutomers are usually easier and less expensive for bwin to retain than acquiring new customers.

10. **Loyalty** Loyalty was calculated by substracting each customer's last Active date from 2005-09-30 (the last day of our observations).If the difference was less than 40, then the customer was classified as 'loyal'. Otherwise, customers (that had not played for more than 1 month) were classified as non-loyal. In this way, we created a variable that was relevant up until the end of our period of interest.

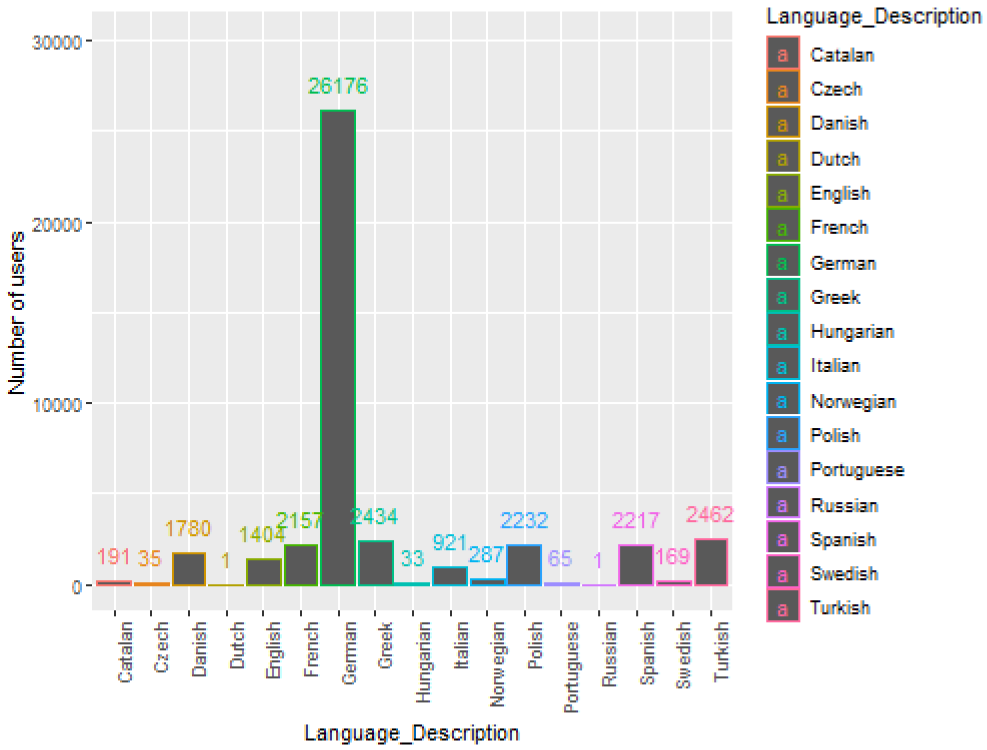
4. Summary Statistics & Insights with Shiny

- **Geographical View**



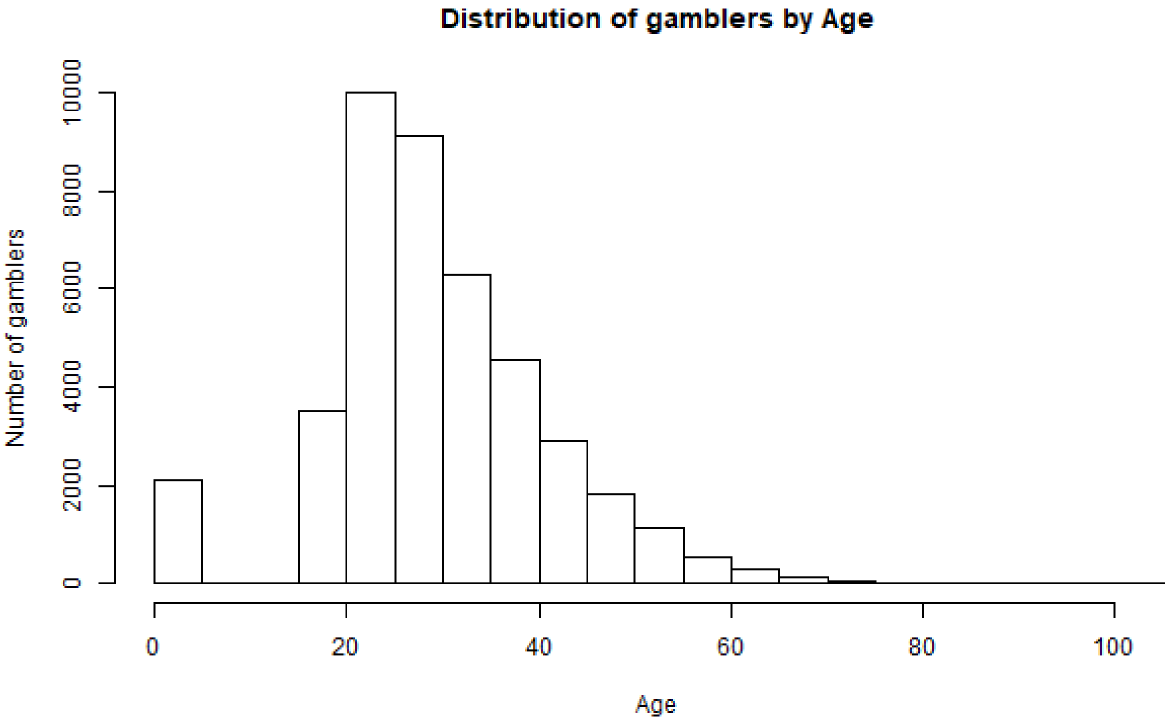
In order to gain a general overview of the gamblers, an interactive geographical view was created depicting the distribution of betting customers worldwide for the different application ids. Betandwin.com is the most popular route of access to bwin, bringing customers from different countries.

- **Distribution of Players by Language**



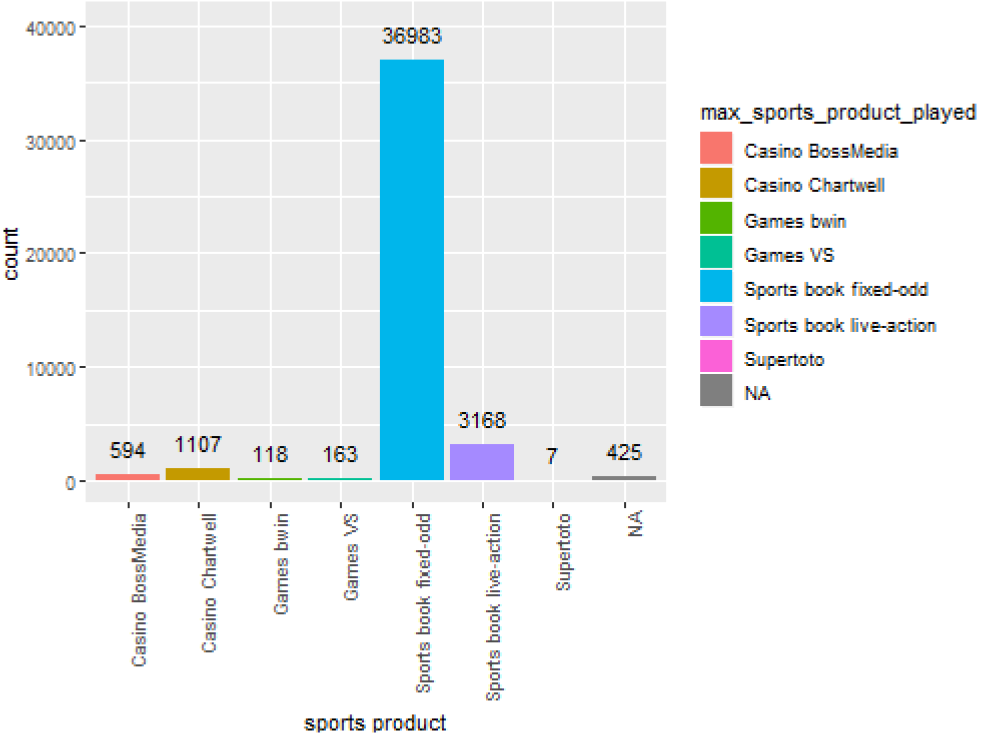
A barplot showing the distribution of gamblers by language for the different application ids and continents. Most customers are German speaking.

- **Distribution of players by Age**



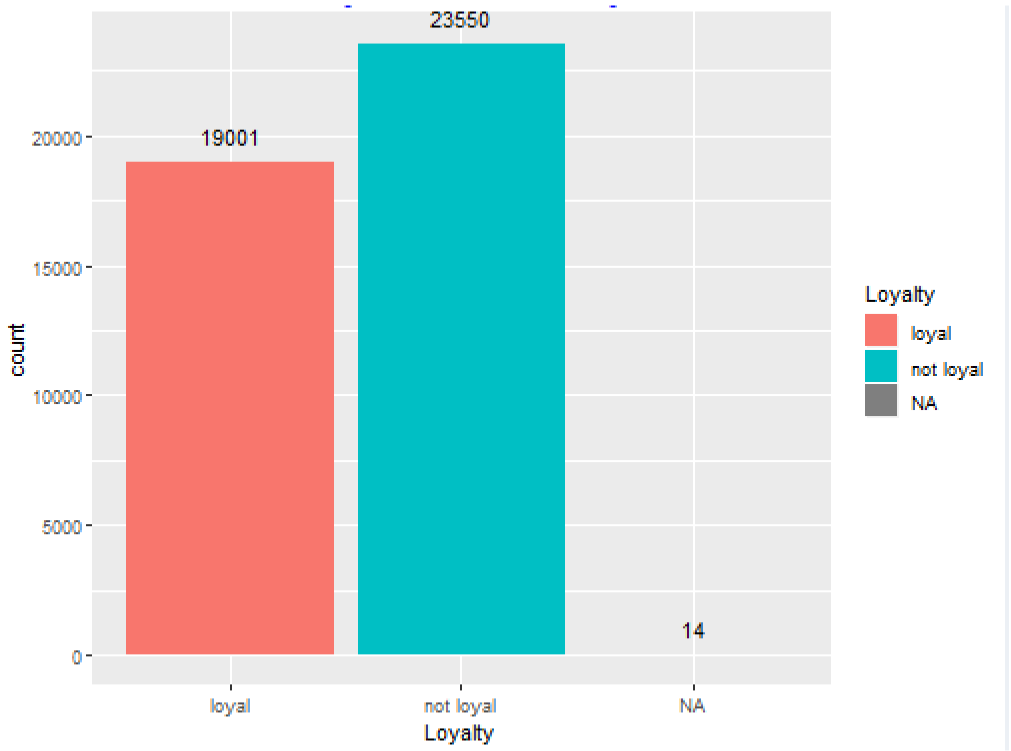
A third barplot was created depicting the distribution of gamblers by age. The worldwide distribution is right-skewed with most customers in the 20-30 age group. Interestingly, certain continents display location-specific trends. For instance, the most prevalent age group in Oceania is gamblers aged 40-50 years-old.

- **The most preffered Sports Game**



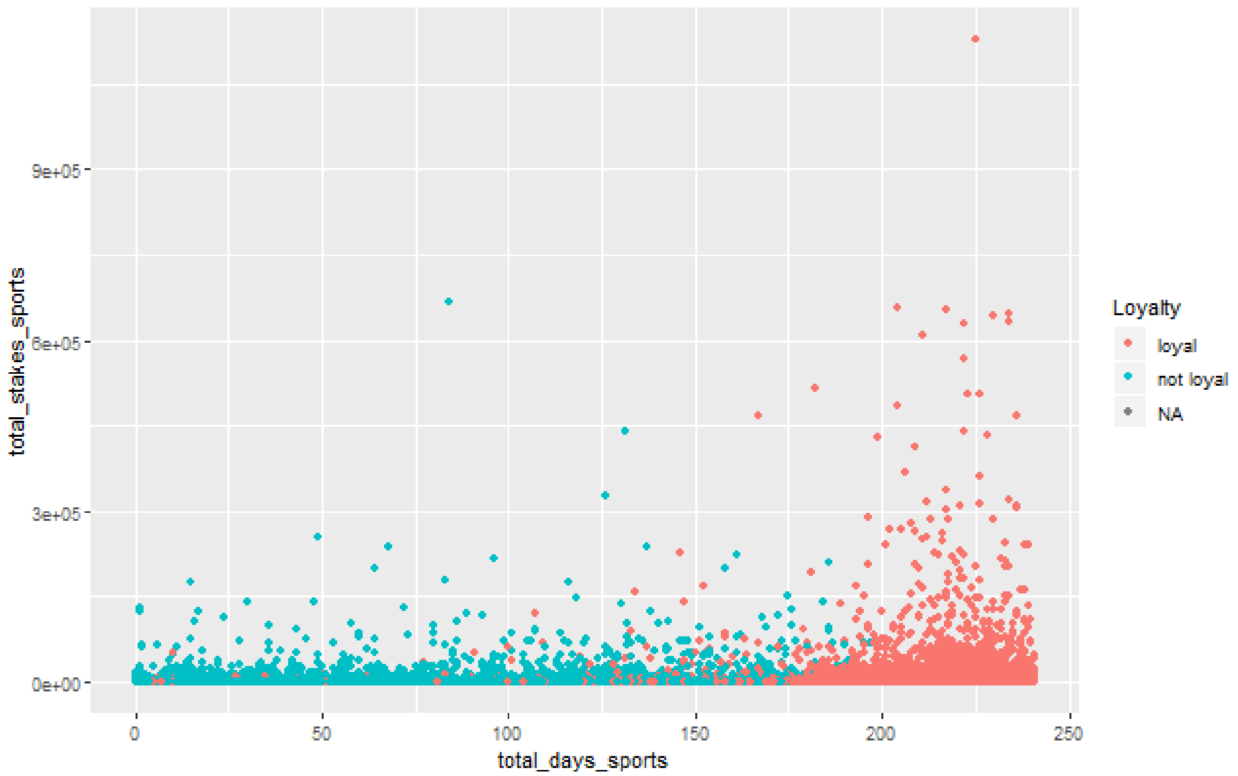
This barplot was designed to identify the preferred game across all sports game. The most played game is Sports Book Fixed odds with the total number of users exceeding 36000.

- **Number of Loyal & not Loyal Customers**



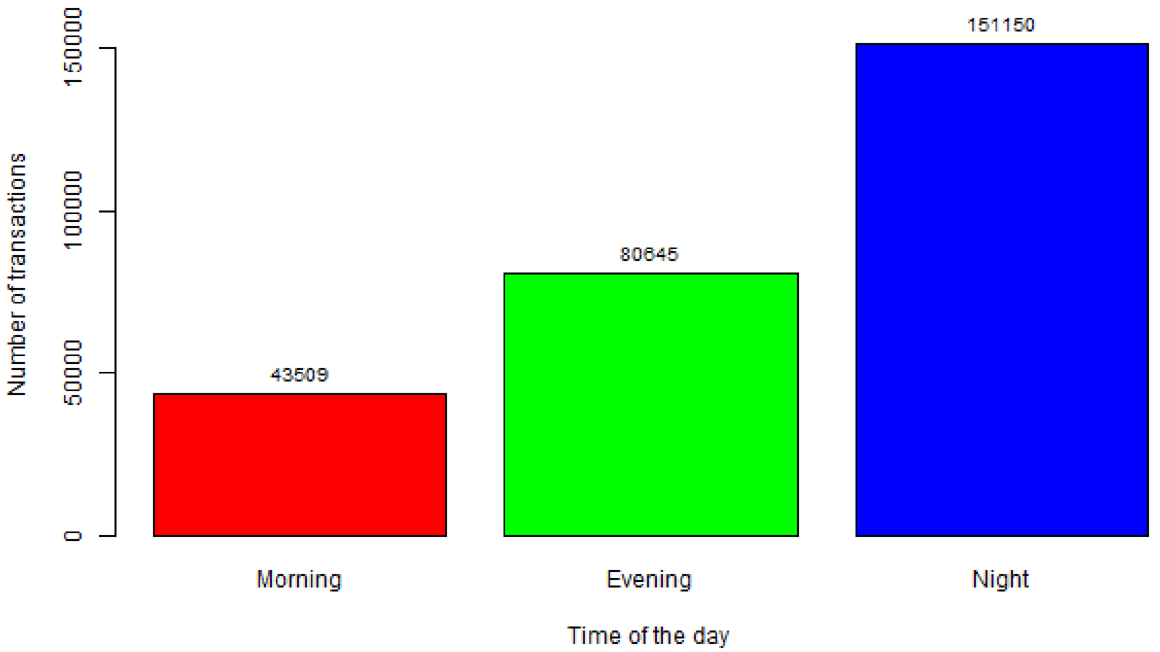
This graph shows the distribution of loyal and not loyal customers. In total, the proportion of loyal customers is higher than that of non-loyal customers. This trend holds across all the continents. The difference between the two groups is more evident in Africa, Oceania, Asia. These continents have the highest proportions of loyal customers.

- **Stakes & Winnings as a Function of Betting Days**



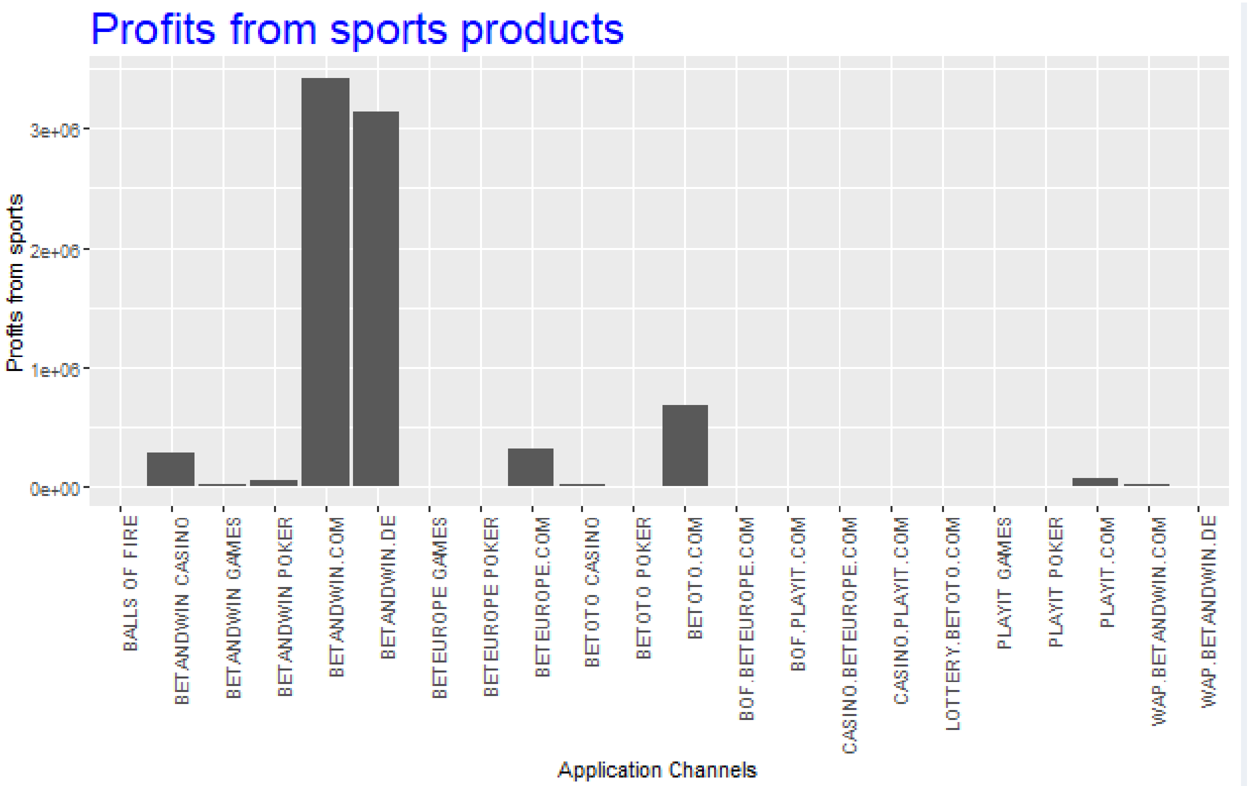
This scatterplot illustrates the relationship between stakes/ winnings and measures of betting frequency (frequency and total betting days) for the different type of games. Gender, and loyalty filtering are applied. Interestingly, for sports bets, customers that play for longer periods (more total days) tend to play at higher stakes. This is particularly true for the loyal group. These customers also are associated with higher sports winnings. Moreover, loyal and non-loyal customers seem to form distinct clusters. A similar trend is observed for the Live Action (LA) games. Gamblers playing for more total days tend to play at higher stakes and receive more winnings. Regarding poker gambling, the more days a user plays the greater the total poker amount sold or bought. Finally, a general insight that can be observed is that males gamble a lot more than females.

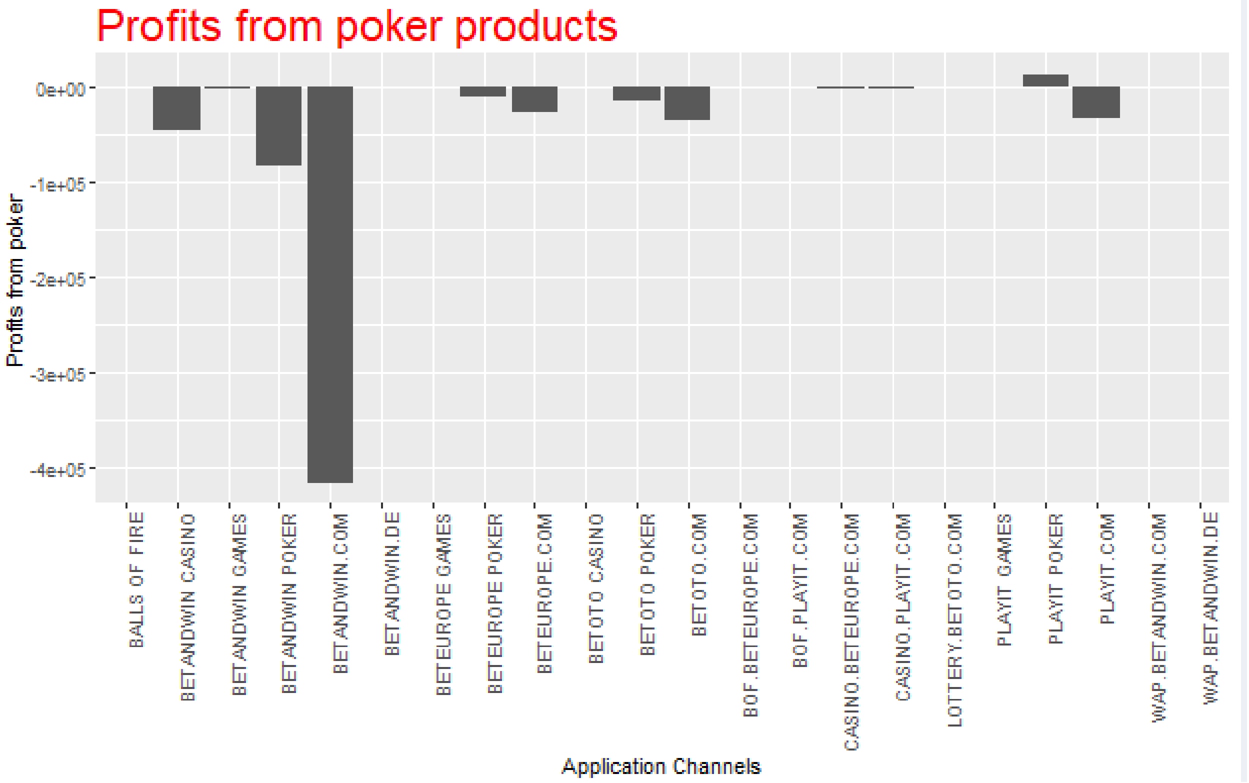
- **Poker Chip Transaction Time**



In the ‘Poker Chip Transaction Time’ graph, the number of transactions is shown for the different times of day. The graph illustrates that more poker chip transactions take place at night, while the least pokerchip transactions take place in the morning.

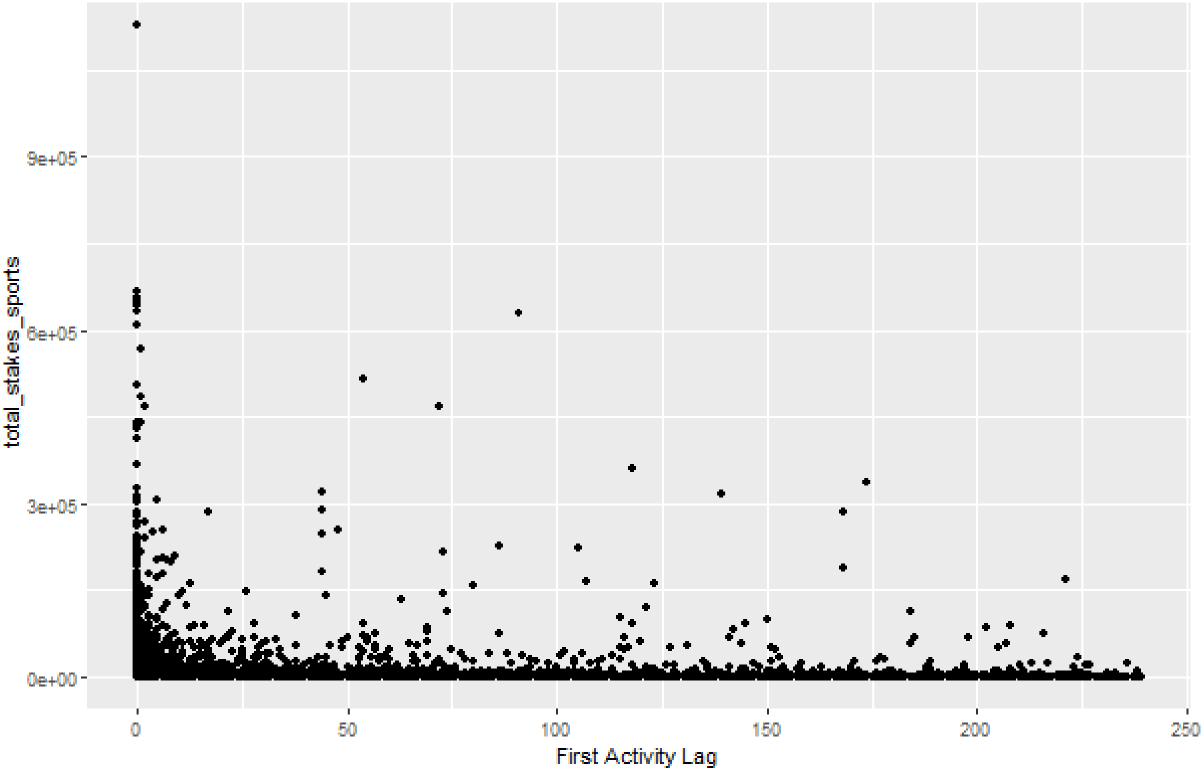
- **Gross Game Revenue Per Application Description**





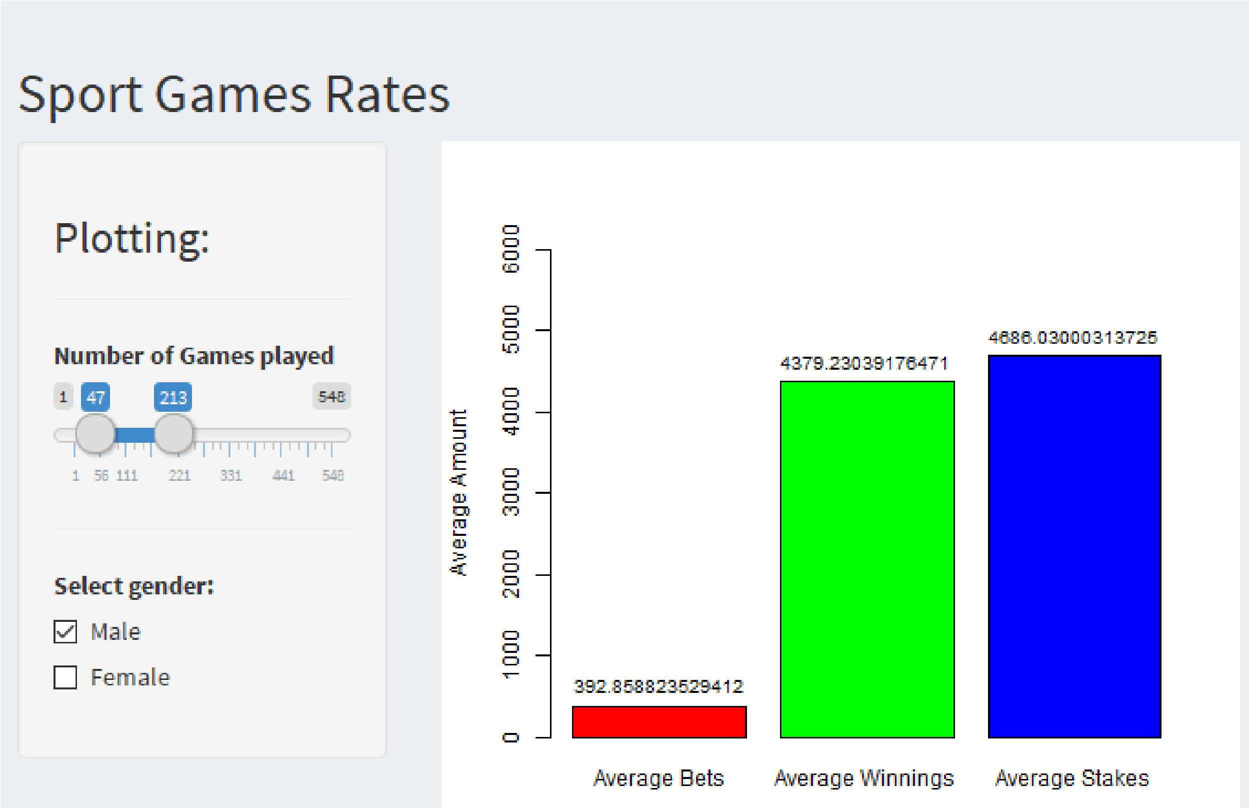
A barplot was created displaying the profit of the company plotted against the different Application IDs for both the sports and pokerchip games. Interestingly, sports games lead to profits while pokerchip games lead to losses for the company. Greater profits come from ‘Betandwin.com’ channels and ‘Betandwin.de’ channels (which also reflects the fact that most players are Germans). Higher losses result from ‘Betandwin.com’, which is also the most common application channel.

- **Bets against First Activity Lag**



A final scatterplot was created depicting the relationship between betting behaviour indicators and a customer’s First Activity Lag (i.e. the amount elapsed between a user’s registration date and first actuve date). The results indicated that the shorter the First Activity Lag, the higher the customer’s total bets, winnings and stakes. The effect is more pronounced for the sports and particularly for the live action (LA) games as opposed to the fixed-odds (FO) games.

- **Bets, winnings and stakes per number of sports games played**



Sport Games Rates

Plotting:

Number of Games played

1 47 213 548

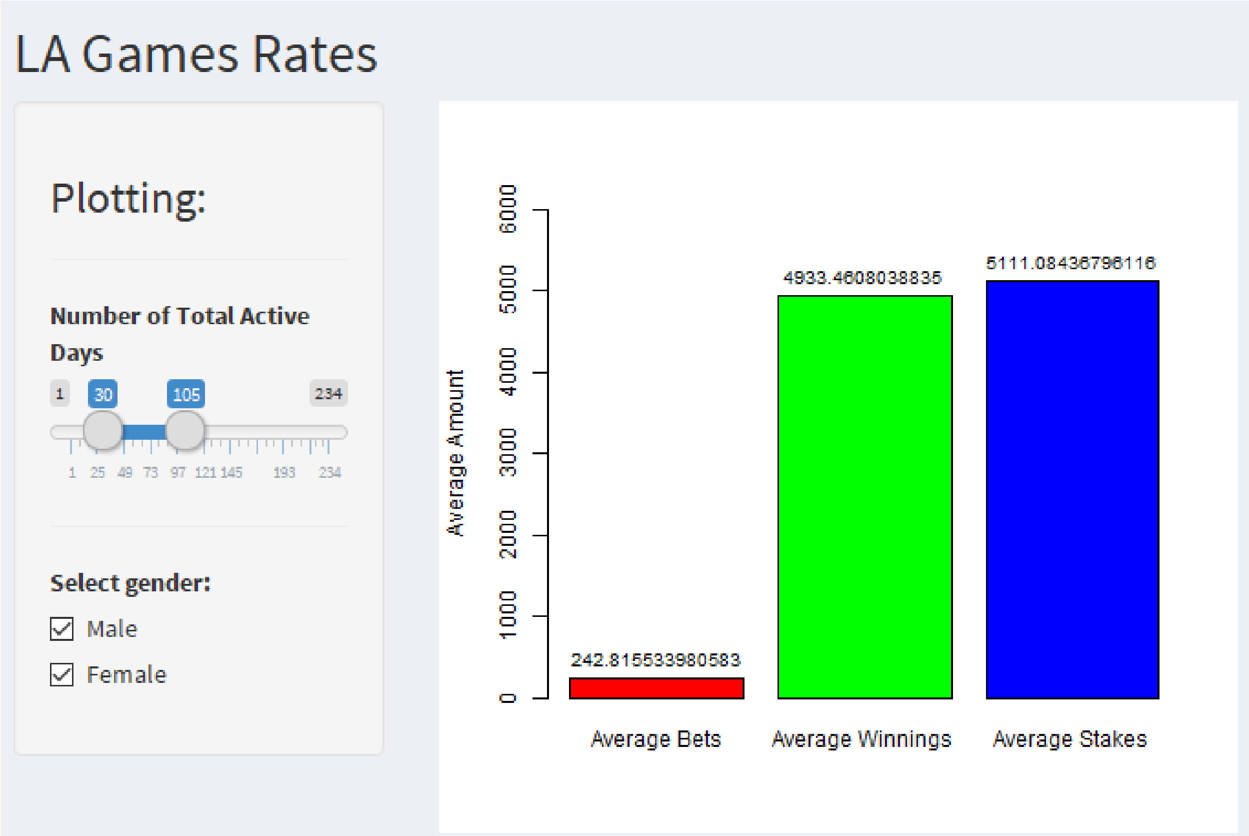
Select gender:

☑ Male

☐ Female

This graph shows the amounts of total bets, winnings and stakes for all customers in our final table. Some more filters have been applied in order to see some differences, for instance between gender, and also the number of games played by users. For the company, knowing the amount of stakes implied depending on the number of sports games played is important in order to design promotional offers for instances. The results shows big differences in stakes, bets and winnings between gender as usual. But also differences between the total number of games played. Indeed we can identify several categories in terms of profitability for the company. From 1 to 100 games played, the company is not making huge profits as bets are really low, while winnings and stakes are almost equal. Then from 200 to 300, the company is making a little more profits as average bets start to represent a greater proportion compared to the average winnings. Then comes the sweet spot when players play over 300 sports games as average bets become higher than both average stakes and average winnings. So the company should make promotion to push people to play to a high variety of games rather than focusing on a few.

- **Bets, Winnings and Stakes for Live Action Games**



LA Games Rates

Plotting:

Number of Total Active Days

1 30 105 234

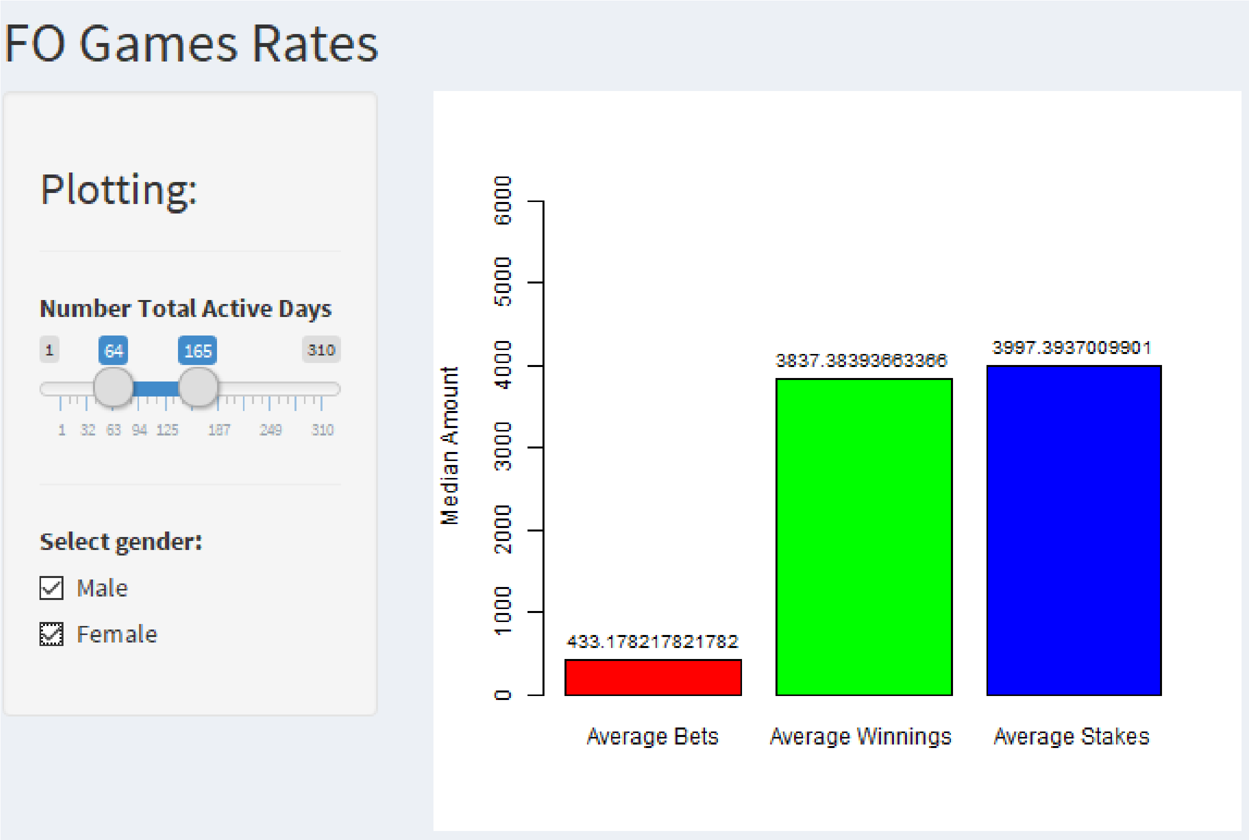
Select gender:

☑ Male

☑ Female

This graph shows the amounts of total bets, winnings and stakes for all customers. Some more filters have been applied in order to identify some patterns. For instance, the user can choose to display the total amount per gender and also uses a slider to select the total number of active days for Live Action games. The results for the Live Action Games shows less potential profits for the company as no matter the amount of total days of activity, the bets always remain lower than the winnings. However, the greater the amount of total days of activity, the greater the stakes and the winnings, meaning that the users are spending more and thus are generating direct revenues. Therefore if the marketing budget is limited, the company should not focus on retaining or gaining new customers for those specific games.

- **Bets, Winnings and Stakes for Fixed Ods Games**



FO Games Rates

Plotting:

Number Total Active Days: 64 – 165

Select gender:

☑ Male

☑ Female

This graph shows the amounts of total bets, winnings and stakes for all customers. Some more filters have been applied in order to identify some patterns. For instance, the user can choose to display the total amount per gender and also uses a slider to select the total number of active days for Fixed Ods games.

The results show no particular effects of number of active days on the average returns as it is proportionnal, so the more the active days, the more the average returns. However, we could identify a sweet spot from 215 to 310 active days where bets are almost equals to the winnings, thus, adding the stakes, the company is making a huge profit there.

Overall Insights & Points of Action:

- **German speaking is the most prevalent group.**
- **Most common age group is 20-40.**
- **Pokerchip games lead to company losses.**
- **Sports games lead to company profits.**
- **Most popular sports game is Sports Book Fixed odds.**
- **Most pokerchip transactions occur at night.**
- **There are more loyal than no loyal customers, particularly in Africa, Oceania, Asia.**
- **Gamblers that play more days also play at higher stakes.**
- **Gamblers that play for the first time immediately after their registration date tend to play at higher stakes and place more bets in total.**
- **Company should make incentive to push people towards playing a high number of different sports games as they make huge profits over 300 different games.**
- **Live Action Total Active days does not have a clear influence on the average returns.**
- **For Fixed-Ods, the greater the amount of active days, the greater the returns with a special profit made after 215 days of activity.**
- **Customers with high fidelity are the most profitable ones.**