Market Analysis - R

GAMBLING COMPANY

Done by

RAJI Hind
FAN Fangda
NITHARWAL Ashwani

SUMMARY

The data and programming tools are used in so many fields in order to help the decision makers to get the right insights in a timely manner. The goal is to be able to fully understand the customer's profile and persona to help the company come up with the right marketing approach. Today we will focus on the use of R in the Marketing field for a gambling company. Our client is a Gambling company, the managers want to better understand their clients and categorize them to decide what category of clients is worth investing in and how. Us, as data analysts will be helping the business managers get answers on how to classify their clients by studying the dataset given by the proper department. We will first get familiarized with the data, clean the data in all the tables that will be used for the purpose of this study and then the goal is to merge everything in one big Bastable as a final output. Finally, a Marketing analysis will be run to present to the managers with the help of some interpreted and explained graphs.

INTRODUCTION

In order to create our Datamart to generate the Marketing insights that would help our managers in better understanding their customers, we will first start by reading the data, getting familiarized with it. Cleaning the tables from missing values and correcting the right format for each and every variable to help us for the analysis later on. After all the tables are cleaned and all the important variables are freshly created, we will proceed to merging all the tables to create what we call a Datamart. The goal of this Datamart is to indeed have a single massive table that groups all the clients' information as a client per row. Finally, from this finale basetable; Datamart, we will proceed to generate some analytical graphs focused on the customer profile and persona to help better understand, profile and even segment them for the future.

GOALS OF THE STUDY

1. Data Description - Exploration

The goal of this part is very important as it is the most crucial step in any data manipulation study. First, we have to get ourselves used to the datasets and the variables. Understand the connections between the different variables and tables even. Once that is done, we can see the bigger picture behind the simple numbers shown in our datasets and therefore we can come up with new ways to create new variables and help the business force to take the managerial decision thanks to the insights generated from this study.

2. Data Cleaning – Variable Creation

There is an expression that fits perfectly in this context which is garbage in garbage out. We obviously wouldn't want to generate insights that won't be useful for our managers so after getting familiarized with the data. We have to clean it, and that means getting rid or modifying the missing values and the NAs, making sure that every variable has the right variable type for the analysis.

3. Market Analysis

After understanding all the tables, understanding the data, cleaning all the tables, and creating new variables that will be useful for our analysis. It is time to merge all this information in one basetable that we call a Datamart. The purpose of creating this Datamart is to have one table that groups all the variables that could be used to better describe the client and their behavior. As a final output, we are looking for a table of many variables as columns and every row will describe one specific and unique client.

DATA CLEANING

Data cleaning is one of the most important and crucial steps of any data analysis. The goal of any study is to convert a huge amount of raw data into useful business insights to help in the decision making by the managers in any company. Thus, only clean data would help us reach this goal. In order to that we have to take care of the inaccurate or inexact data and the missing values otherwise the study result will be biased.

Demographics

The Demographics dataset is the primary table we have because it contains all of the most important information related to the client's social status etc. The variables in this dataset are a mix of Numeric and text variables like the UserId to distinguish uniquely every client, the country of residence of the client, their primary language and their gender. Also, we get data of their Registration date, first pay-in date, first active date first sports book play date, first casino play date first games play date first poker play date and betting application. As previously mentioned, this table is very important, still it required lot of data cleaning. Therefor, we proceeded to clean the data as follow.

Table 1: Demographics Dataset

UserID	Country	Language	RegDate	FirstPay	FirstAct	FirstSp
1324354	276	2	2005-02-01	20050224	20050224	20050224
1324355	300	8	2005-02-01	20050201	20050201	20050201
1324356	276	2	2005-02-01	20050201	20050202	20050202
1324358	752	1	2005-02-01	20050201	20050201	20050201
1324360	792	7	2005-02-01	20050202	20050202	20050202
1324362	276	2	2005-02-01	20050211	20050211	20050211

First thing, we checked the rows with the missing value in the variable gender. After finding out the row with the missing value we decided to generate a gender value to it based on the gender distribution of the people who have the language 4 which is Spanish as a primary language. Therefore the missing value was replaced by which is Male. Second, in order to get an easier to read data. We have decided to change all the encoders to actual text values to help us in the analysis later. And that is by switching the gender variable values 1 and 0 to Male and Female respectively. Also, the values of the language variable, the country variable and the application variable were switched from numeric codes to readable values giving in the other sheets. Finally, new variables were created to dictate if the client has already played in casino, sport book, poker and games by yes or no. Moreover, a time difference variable was created between the registration date and the first active date to detect who are the active clients and when on average they become active. The table above is a part of the demographics table that gathers all the changes made on the original demographics' dataset.

User Daily Aggregation

Table 2: User Daily Aggregation Dataset

UserID	Date	ProductID	Stakes	Winnings	Bets
1324354	20050224	1	20	0	2
1324354	20050225	1	0	0	0

UserID	Date	ProductID	Stakes	Winnings	Bets
1324354	20050227	1	20	0	2
1324354	20050303	1	10	0	1
1324354	20050304	1	10	0	1
1324354	20050305	1	10	0	1

The User Daily Aggregation dataset is the dataset that contains all the information related to the betting of each product by each participant for each calendar day with at least one transaction from the 1st to the 25th of February 2005. This table contains also a mix of text and numeric data like the User Id that is assigned to each participant at the time of registration, date of betting activity aggregation, betting product id, total betting money in euros, total winnings credited to participant in euros and the total number of bets. For this table, we started by checking the missing values and addressing the data type. Then we merged the product table with the user daily aggregation dataset to generate a column for the product description. After this, we proceeded to create new variables that would help us later in the analysis like balance, which is the total winnings on a giving day, first pay, the month extracted from the date of the first pay and others. Other tables were created in the process like monthwise and product wise to help us extract new variables that will be merged later on with the user daily aggregation called Useraggbasetable. At the end we had to check again for the missing values before jumping to the other table.

Pokerchips

The Poker Chip Conversions dataset contains all the information about the poker chip transactions from the date February 1st to September the 30th 2005. This table contains again the user Id, the poker chip transaction type 124 refers to buy and 24 refers to sell, poker chip transaction date and time and the poker chip transaction amount in euro. For cleaning the data of this specific dataset, started by checking for the missing values and splitting the date time column into date and time separately. Then we created new variables of the total, maximum, minimum, average and count of the poker transactions by transaction type sell and buy and the maximum and minimum date. Then the missing values were encoded to a binary variable of 0 when the value is missing and 1 when it exists. Also for the column we recently created we decided to fill in the missing values by the respective logic to not bias the analysis. For instance, for the column average poker transaction buy and sell's missing values we calculated the average and we assigned it to the missing values, for the inf values we assigned the minimum and the negative inf the maximum was assigned etc. Finally, the marker columns were removed, and we doubled checked for the missing values again.

Table 3: User Daily Aggregation Dataset

UserID	TransDateTime	TransType	TransAmount
1324355	2005-06-12 00:37:00	124	8.9999
1324355	2005-06-12 00:51:00	124	1.9999
1324355	2005-06-12 01:14:00	124	4.9999
1324355	2005-06-12 02:01:00	24	1.8069
1324355	2005-06-14 23:35:00	124	4.9999
1324355	2005-06-14 23:45:00	124	4.9999

DATAMART

After cleaning all the datasets from the missing values, mismatching data type, creating all the necessary variables for the study and grouping everything by user Id in order to have every row represented by a client in all our tables. It is now time to merge all the treated dataset into one massive base table that we call a data mart. The merge was easily done using the merge function by the user Id. Finally, we decided to replace any occurring missing values by the most appropriate value such as median, mode or 0 for no entry or record of activity.

Table 4: DataMart

UserID	RegDate	FirstPay	FirstAct	FirstSp	FirstCa
1324354	2005-02-01	2005-02-24	2005-02-24	2005-02-24	NA
1324355	2005-02-01	2005-02-01	2005-02-01	2005-02-01	NA
1324356	2005-02-01	2005-02-01	2005-02-02	2005-02-02	NA
1324358	2005-02-01	2005-02-01	2005-02-01	2005-02-01	NA
1324360	2005-02-01	2005-02-02	2005-02-02	2005-02-02	2005-02-03
1324362	2005-02-01	2005-02-11	2005-02-11	2005-02-11	NA

1. UserID Identifier for the user
2. RegDate Registeration date for a user
3. FirstPay Participant's first betting money deposits date
4. FirstAct Participant's first active play date
5. FirstSp Participant's first sports book play date
6. FirstCa Participant's first casino play date

Table 5: DataMart

FirstGa	FirstPo	Gender	Country	Language	Application
NA	NA	Male	Germany	German	BETANDWIN.DE
NA	2005-06-11	Male	Greece	Greek	BETANDWIN.COM
NA	NA	Male	Germany	German	BETANDWIN.DE
NA	NA	Male	Sweden	English	BETANDWIN.COM
NA	NA	Male	Turkey	Turkish	BETEUROPE.COM
NA	NA	Male	Germany	German	BETANDWIN.DE

7. FirstGa Participant's first games play date
8. FirstPo Participant's first poker play date
9. Gender Gender of the user
10. Country Country of residence of user
11. Language Primary language
12. Application Betting application

Table 6: DataMart

RegDate_FirstAct	PlaysSportsbook	PlaysCasino	PlaysGames	PlaysPoker
23	Yes	No	No	No
0	Yes	No	No	Yes
1	Yes	No	No	No
0	Yes	No	No	No
1	Yes	Yes	No	No
10	Yes	No	No	No

13. RegDate_FirstAct Time difference between first active date and registration date
14. PlaysSportsbook Whether user plays the sports book category
15. PlaysCasino Whether user plays the casino category
16. PlaysGames Whether user plays the games category
17. PlaysPoker Whether user plays the poker category

Table 7: DataMart

win_vs_stake_product_Sports.book.fixed.odd	count_product_Sports.book.fixed.odd
1.0085614	117
1.1308187	99
0.4160801	51
0.6212249	8
0.6659478	29
0.0000000	7

18. win_vs_stake_product_Sports.book.fixed.odd Ratio of winnings to stake amount per product by product name
19. count_product_Sports.book.fixed.odd Count of presence per product by product name

Table 8: DataMart

total_bets_product_Sports.book.fixed.odd	balance_product_Sports.book.fixed.odd
236	86.7900
231	52.4400
98	-400.6800
7	-93.8215
40	-20.0429
7	-22.0000

20. total_bets_product_Sports.book.fixed.odd Total number of bets placed on a product by product name
21. balance_product_Sports.book.fixed.odd Balance id is the difference between winnings and stakes for that product by product name

Similarly for all products. . . . . .

Table 9: DataMart

win_vs_stake_month_2	count_month_2	total_bets_month_2	balance_month_2
0.0000000	3	4	-40.0000
0.9138909	28	78	-6.9800
0.7137483	27	49	-33.4800
0.7456107	3	4	-30.7130
0.7965358	15	23	-4.4438
0.0000000	3	2	-10.0000

50. win_vs_stake_product_Sports.book.fixed.odd Ratio of winnings to stake amount per month
51. count_product_Sports.book.fixed.odd Count of presence per month
52. total_bets_product_Sports.book.fixed.odd Total number of bets placed on a month
53. balance_product_Sports.book.fixed.odd Balance id is the difference between winnings and stakes for that month

Similarly for all months. . . . . .

Table 10: DataMart

total_win_amount	total_stake_amount	final_balance	total_bets
11736.6100	11976.6100	-240.0000	279

total_win_amount	total_stake_amount	final_balance	total_bets
464.5000	425.5600	38.9400	252
910.6600	1365.2600	-454.6000	214
209.8575	336.2898	-126.4323	11
43.1573	65.7427	-22.5854	47
0.0000	22.0000	-22.0000	7

82. total_win_amount The total amount of money won by a client
83. total_stake_amount The total amount of money bet by a client
84. final_balance The difference between the total winnings and bets per customer. A negative balance reffers to losses and a positive one reffers to winnings
85. total_bets The total number of bets per users

Table 11: DataMart

total_count_of_presence	bets_per_presence	final_win_vs_stake	status
136	2.051471	0.9799609	Loss
106	2.377359	1.0915030	Profit
75	2.853333	0.6670231	Loss
9	1.222222	0.6240377	Loss
32	1.468750	0.6564577	Loss
7	1.000000	0.0000000	Loss

86. total_count_of_presence The total number of times the user was present in the casino
87. bets_per_presence Total number of bets per casino visit
88. final_win_vs_stake Overall ratio of winnings vs stakes per user
89. status User's status. If the final_win_vs_stake <1 then the status is Loss and if it is higher than 1 then the status is Profit

Table 12: DataMart

MaxPokerTranDate	MinPokerTranDate	TotalPokerTranSell	TotalPokerTranBuy
NA	NA	NA	NA
2005-06-15	2005-06-12	8.2593	30.8295
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA
NA	NA	NA	NA

90. MaxPokerTranDate The maximum pokerchip transaction date for a user
91. MinPokerTranDate The minimum pokerchip transaction date for a user
92. TotalPokerTranSell The total sell type pokerchip transaction amount for a user
93. TotalPokerTranBuy The total buy type pokerchip transaction amount for a user

Table 13: DataMart

MaxPokerTranSell	MaxPokerTranBuy	MinPokerTranSell	MinPokerTranBuy	AvgPokerTranSell	AvgPokerTranBuy
NA	NA	NA	NA	NA	NA
6.4524	8.9999	1.8069	1.9999	4.12965	5.13825
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA

MaxPokerTranSell	MaxPokerTranBuy	MinPokerTranSell	MinPokerTranBuy	AvgPokerTranSell	AvgPokerTranBuy
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA

94. MaxPokerTranSell The maximum sell type pokerchip transaction amount for a user
95. MaxPokerTranBuy The maximum buy type pokerchip transaction amount for a user
96. MinPokerTranSell The minimum sell type pokerchip transaction amount for a user
97. MinPokerTranBuy The minimum buy type pokerchip transaction amount for a user
98. AvgPokerTranSell The mean sell type pokerchip transaction amount for a user
99. AvgPokerTranBuy The mean buy type pokerchip transaction amount for a user

Table 14: DataMart

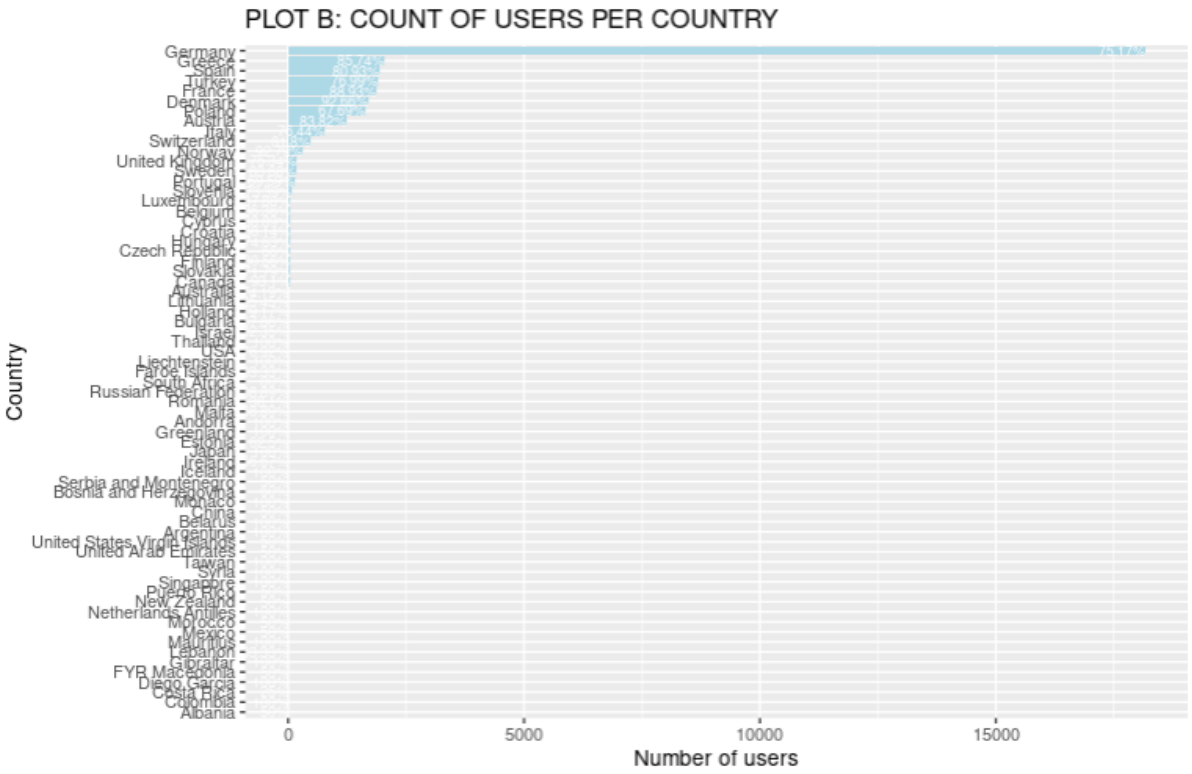
CountPokerTranSell	CountPokerTranBuy	PokerTranDays	FrequencyPokerSell	FrequencyPokerBuy	PokerBalance
NA	NA	NA	NA	NA	NA
2	6	3	0.6666667	2	-22.5702
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA

100. CountPokerTranSell Count of sell type pokerchip transactions for a user
101. CountPokerTranBuy Count of buy type pokerchip transactions for a user
102. PokerTranDays Total number pokerchip transaction days per user
103. FrequencyPokerSell Count of sell type pokerchip transactions for a user divided by the total number pokerchip transaction days per user
104. FrequencyPokerBuy Count of buy type pokerchip transactions for a user divided by the total number pokerchip transaction days per user

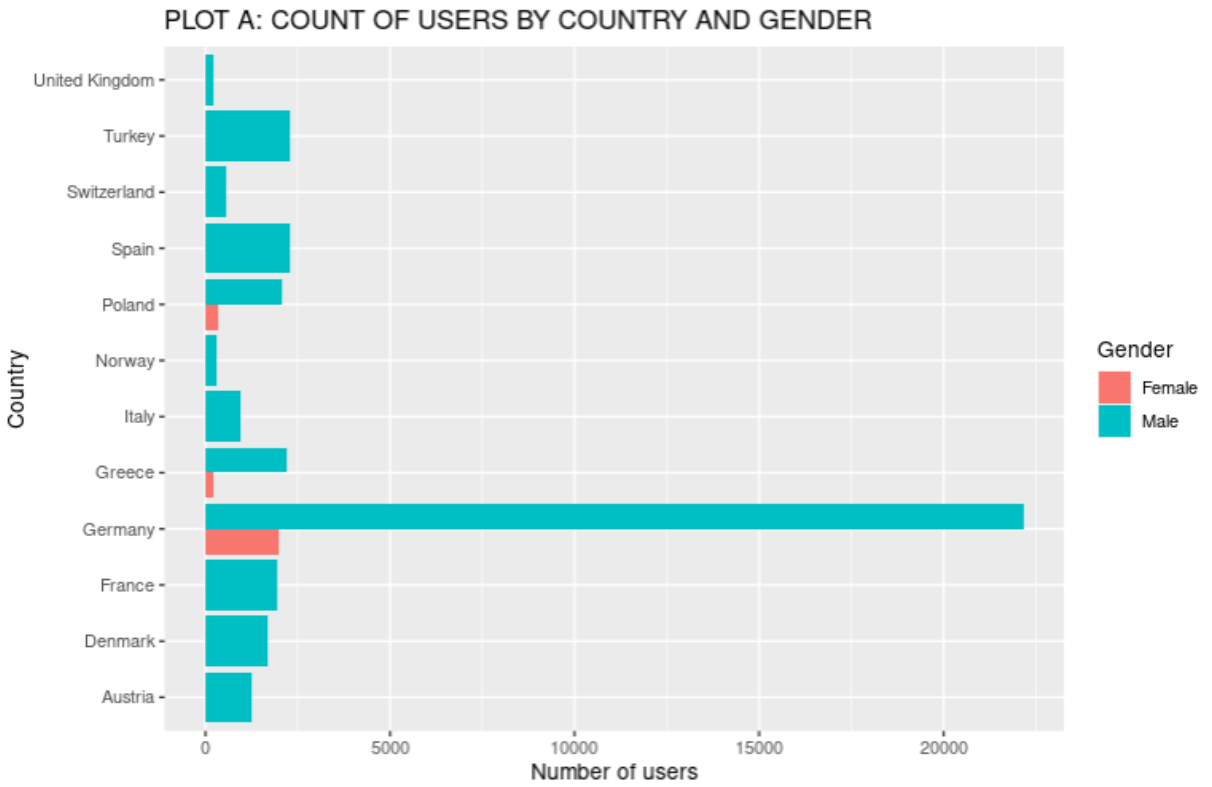
105. PokerBalance The total sell type pokerchip transaction amount for a user minus the total buy type pokerchip transaction amount for a user

ANALYSIS

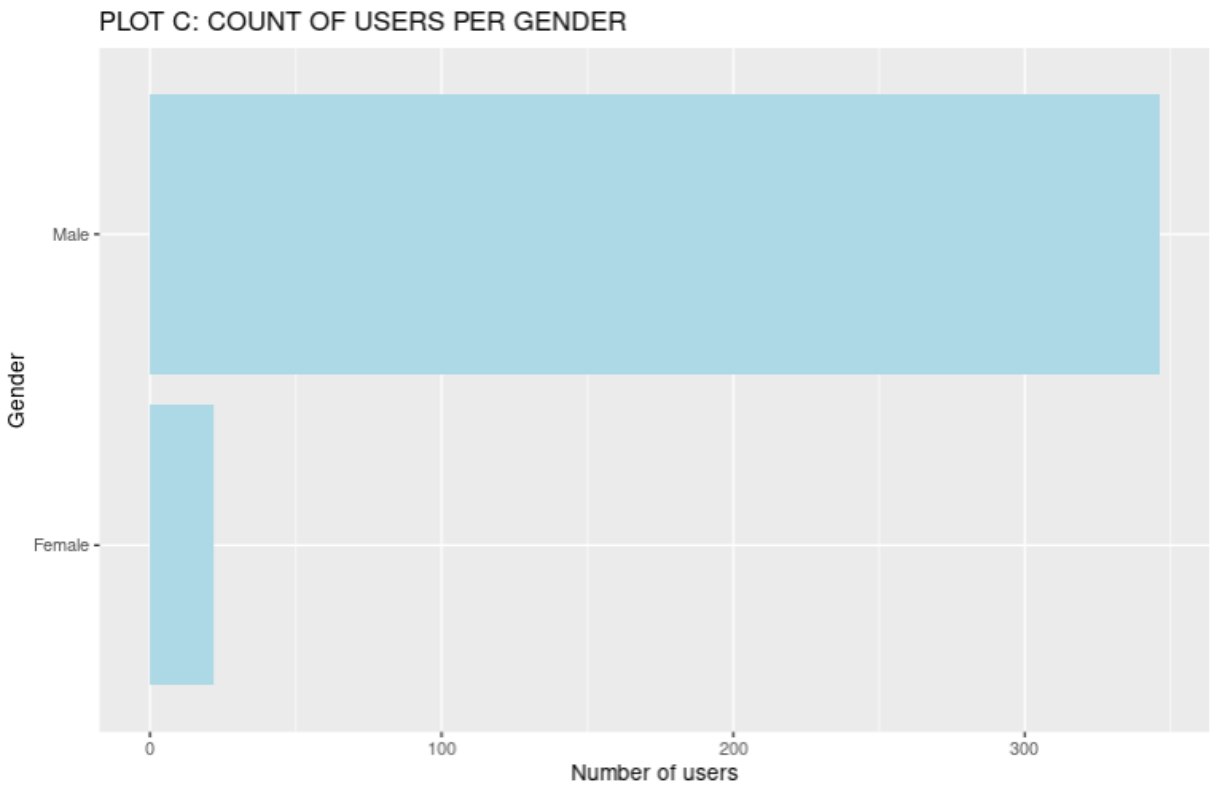
All the tables have been cleaned and merged now. All the valuable variables were created, and we are ready to generate some plots and graphs to better describe the client and could even categorize them in different personas to help the managers and/ or owners better understand their own clients to help them in the strategical decision making in the future. For instance, thanks to this marketing analysis we can easily define what kind of publicity we should go for, how to impact and influence the target and make them use more the company's products.



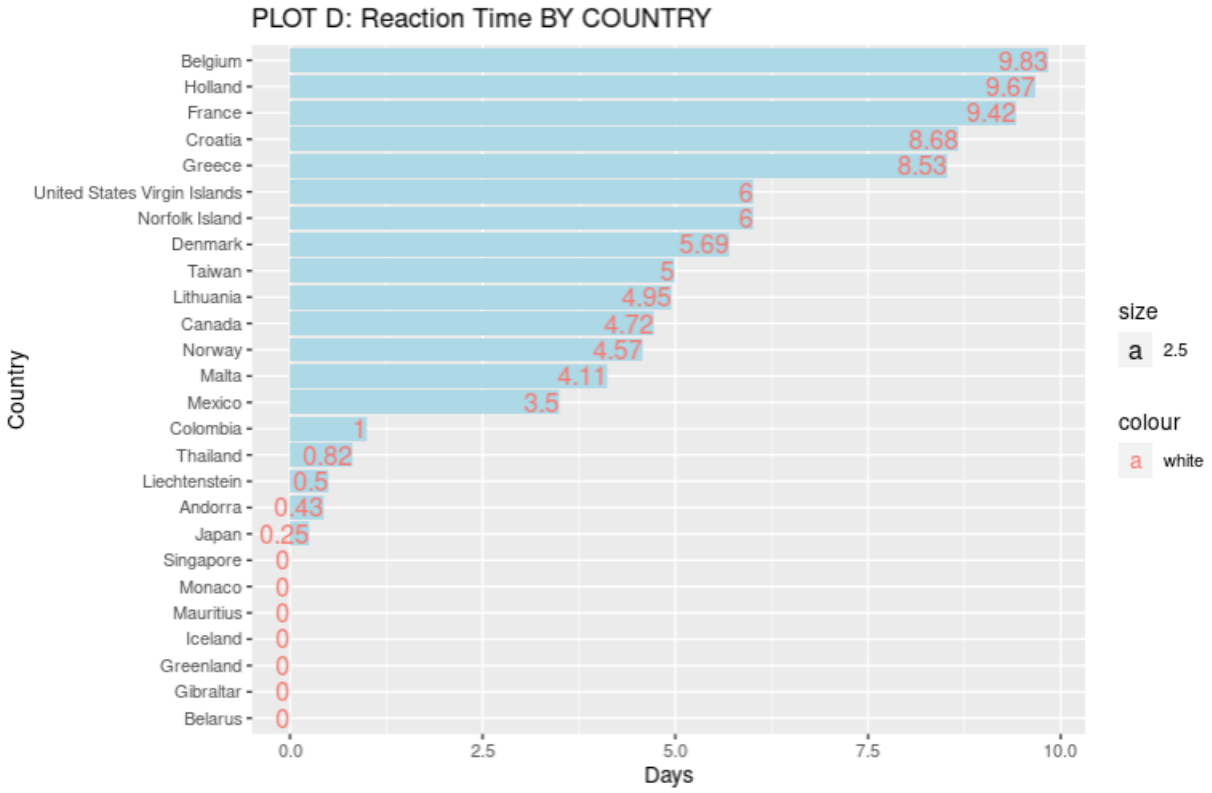
This graph shows the users distribution per country and we can easily see that the country which is Germany is leader here by having the highest number of users represented by more than 75.17% of the total clientele of the company. We can say Germany is a really good country for this business and we can there make a lot of profit.



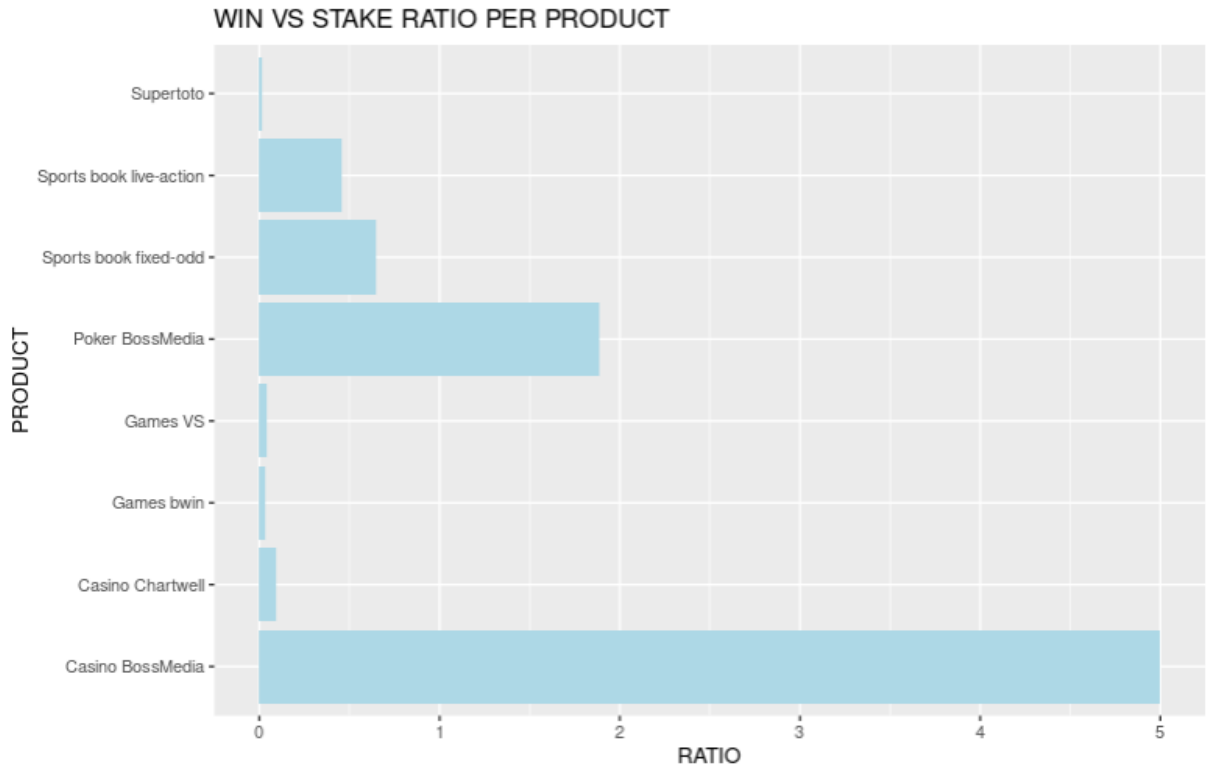
This graph shows us the comparison between the gender in different countries. We can see that most of the clients are male and hence can be said that males are more inclined towards this activity. Now we saw if the first graph that Germany has the most market shares in terms of active users by 56% of the entire company's clientele and we also saw that male users are reigning by more than 90%. Let's see now the distribution of the users by country and by gender at the same time. The previous results are once more here confirmed. We see that Germany is again having the highest market share by almost 25000 male user and around 2000 female user.



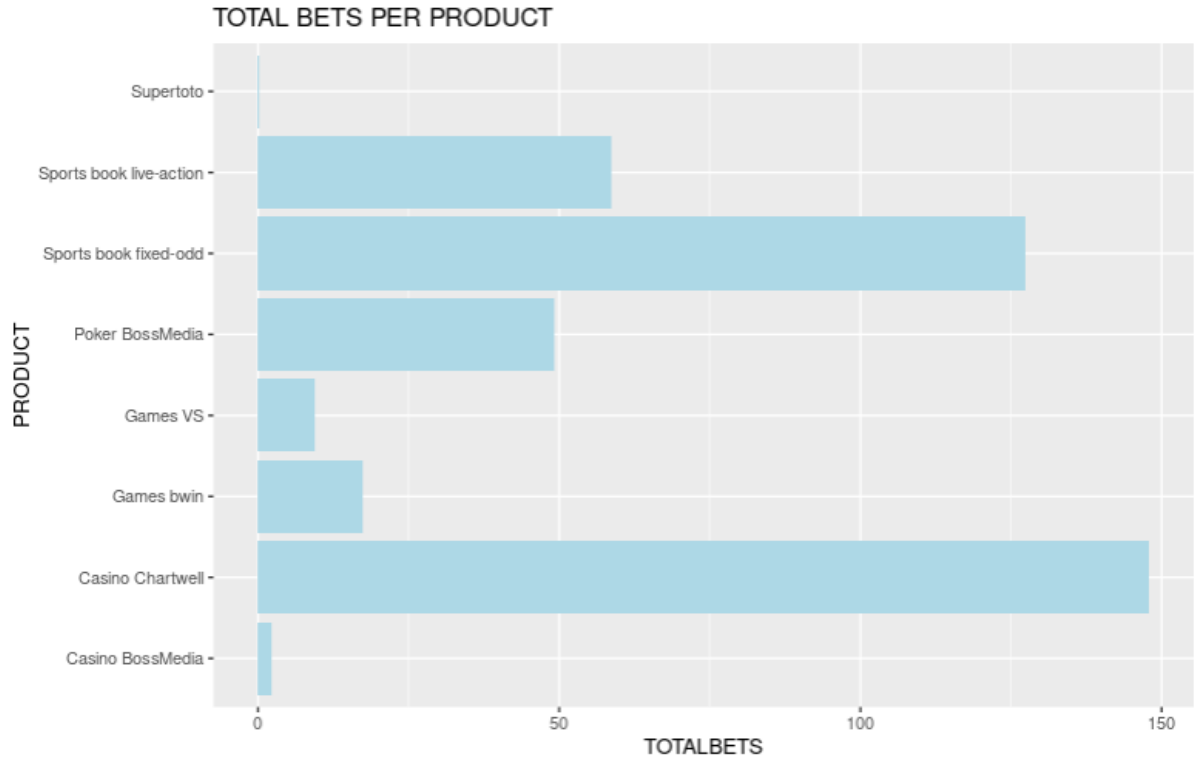
It is obvious that the company would want to know how many female and male clients they have, therefore we plotted the count of users per gender. It is clear that most of the users as more than 90% are male clients and the rest which is the smallest minority are female users. We can conclude that this gambling company attracts more male clients than female ones or men are in general more interested in gambling than women.



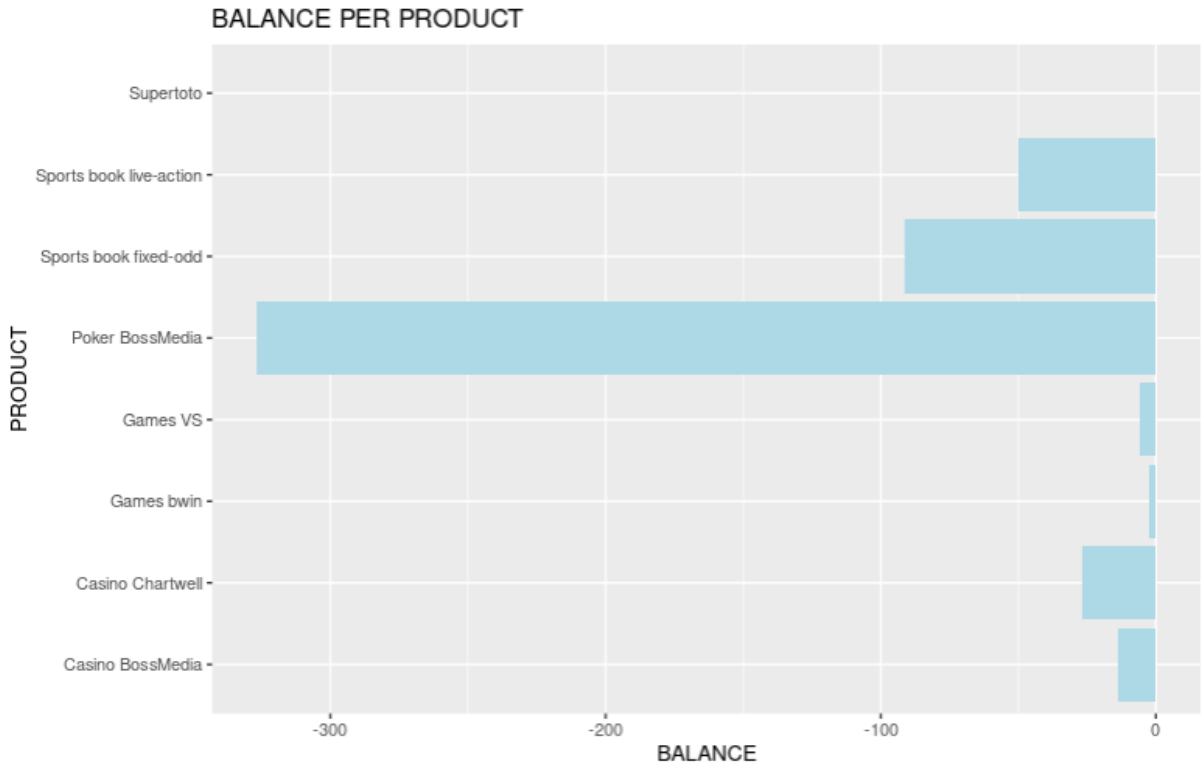
It is a good practice to also see the distribution of when the clients switch to active users per country. Therefore, the above graph was plotted of the time from registration to first active date per country. We can see that Belgium, Holland, France, Croatia and Greece are the countries that have the clients they take the most time to switch to active users. As a business recommendation we can propose to invest more in these countries in terms of marketing to better promote the company and its presence.



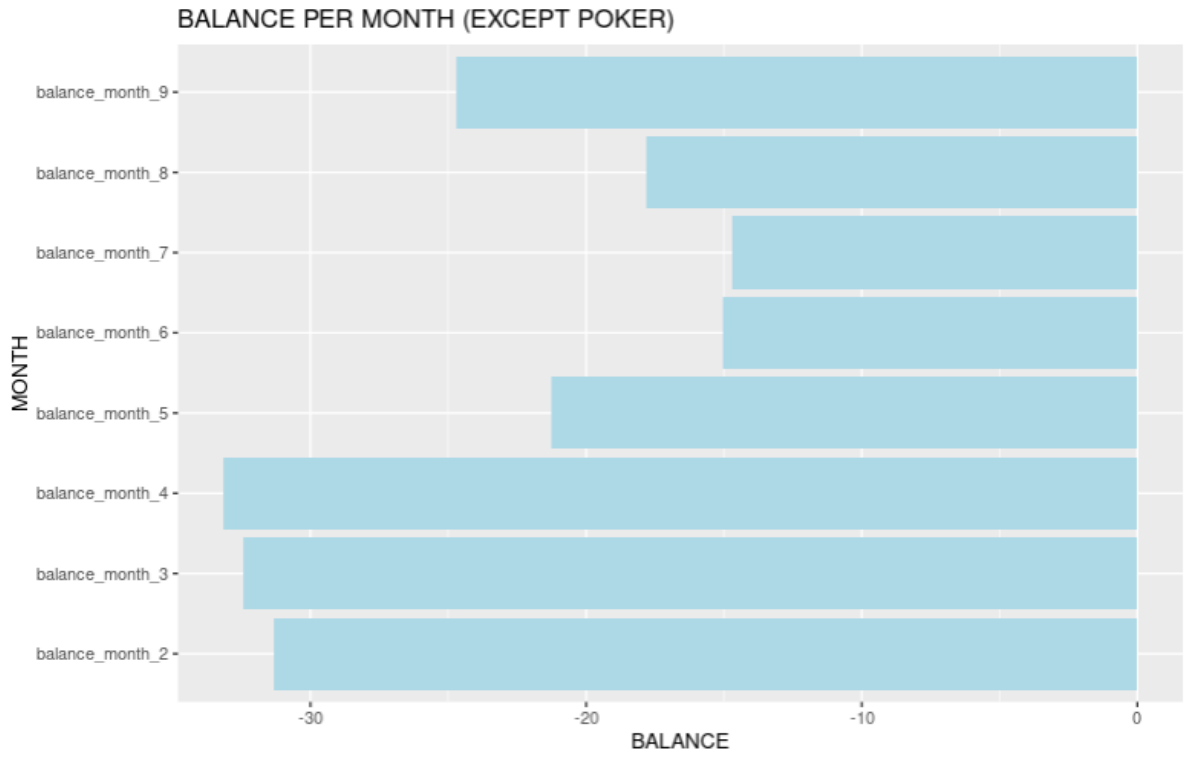
This is a win Vs stake ration per product graph which basically illustrates the distribution of the products per ratios (sell vs buy). We can see that Casino BossMedia has the highest ration of 5, just behind it Poker BossMedia which is the same chain has a ratio of almost 2 and finally Sports book fixed-odd with a ration 0.6.



In order to determine which product performs better than the other we decided to plot total bets per products. We can see in this graph that Casino Chatwell is the one performing well with almost 150 bets and just behind it Sports book fixed-odd with more than 125 bets. We can conclude from this that these two later are performing very well in generating money but other products are still far away like Games VS and Games bwin which need definitely more work in terms of product placement and positioning, also the choice of the channels is super crucial to make a product pop using a marketing strategy of the 4Ps.



The balance represents the difference between the amount of money that the user used in betting and the amount of money won. If the difference is positive than the client has made a profit and won more than what he bet. We can see from the graph above of the Balance per product that people who bet on Poker BossMedia usually lose more money than what they win as represented by the graph to -320. Maybe the game is harder to understand or to win or simply the company is doing an amazing job promoting the Poker BossMedia and attracting lot of new customers. Either ways obviously the company is making lot of money from this product in particular.



As we have previously explained the balance is the difference between the money the client used to bet and the money they won at the end. And a negative balance for the client represents a positive balance for the company. We can see that the client are more active during the months of February, march, April and September; they are losing money more than what they are betting which makes the company make more profit in these months respectively.

PROFILING

As we previously mentioned, the goal of this study is to be able to understand the customers of the gambling company and categorize them into personas if possible. After cleaning all the tables, studying them, merging everything into a datamart and analyzing the variables with the previous graphs.

Now we want to build a client persona, we will only go with the extreme values to generate this persona.

- Almost 90% of the clients are male
- Clients take up to 15 days to become active
- Casino Boss Media has the highest ratio of 5
- Casino Chartwell has the highest total bets of 150
- 56% of the clients come from Germany, and speak German

CONCLUSION

In conclusion, we wanted to analyze the datasets provided by the gambling company to help us in profiling the customers they have to help them run strategically their marketing campaigns in the future. In order to do this, we used one of the well know programming languages which is R in reading in the data, treating the missing values, creating new variables for the sake of the study and generating the graphs that will be easily understood by the business capacity of the gambling company.

After plotting all the relevant variables, we interpreted every single graph as to transform the hard codes that won't necessarily be understood by a person with a limited technical background to easy-to-understand

business insights. For every graph we made sure to first analyze the plot and then come up with a business explanation and or recommendation.

Lastly and definitely not least we made a concise profiling of the customers, keep in mind that we only took the highest values to create 1 most probable persona. Therefore, we could see that the usual or most probable client at this gambling company is referred to as male, lives in Germany and speaks German as primary language and it takes him around 15 days as a maximum to switch to an active customer. Also, the Chartwell casino is the won scoring the highest bets by 150 in total.

RESOURCES

- Professor's course material.
- Professor's group assignment PDF document.
- R Community
- Google resources for knowledge gain and debugging