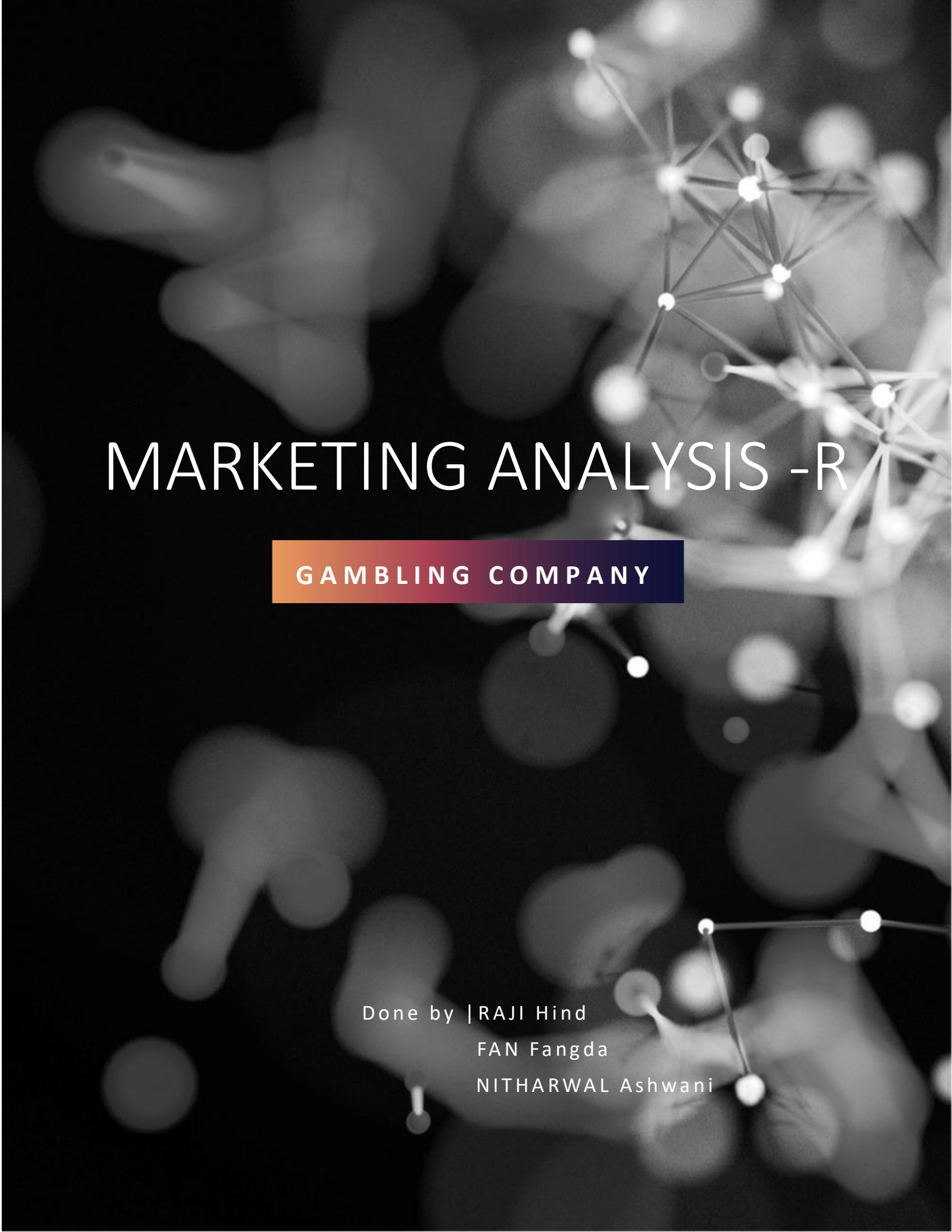


MARKETING ANALYSIS -R

GAMBLING COMPANY

Done by |RAJI Hind
FAN Fangda
NITHARWAL Ashwani

Table of Contents

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.

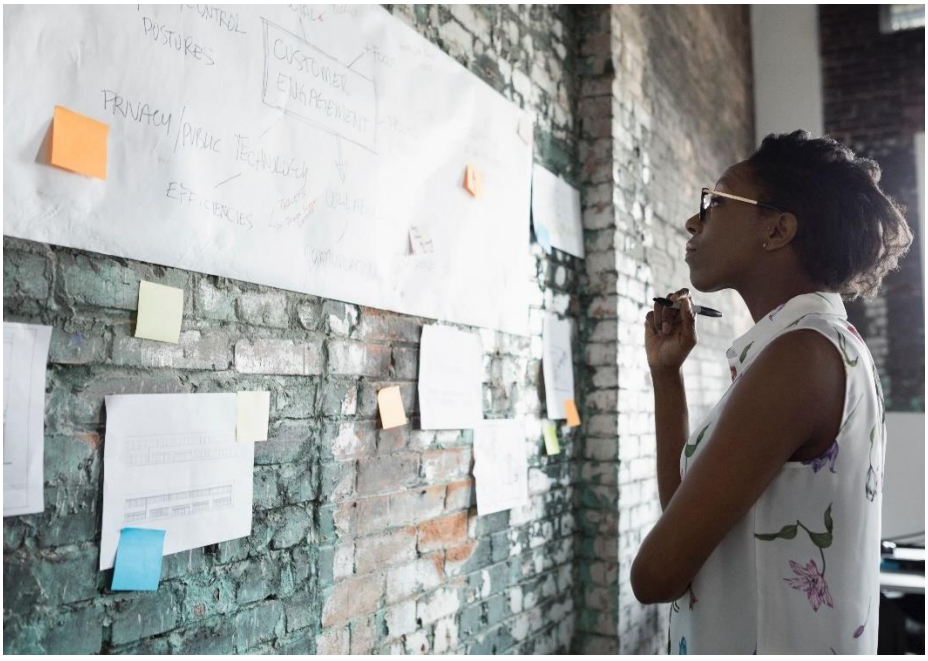
Us, as data analysts will be helping the business managers get answers on how to classify their clients by studying the dataset given by the proper department. We will first get familiarized with the data, clean the data in all the tables that will be used for the purpose of this study and then the goal 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 THIS STUDY

Data description and exploration

- Reading the tables and describing the variables.

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.

DATA CLEANING AND VARIABLE CREATION

- Getting rid of the NA and the missing values.
- Creating new variables for the analysis purposes.
- Creating the Datamart.

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.

MARKETING ANALYSIS

- Analyzing the Marketing metrics and presenting them in graphs.

MARKETING 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

Say NO to: "Garbage in, Garbage out !! "

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.

Demographics

Gender	Country	Language	Application	RegDate_FirstAct	PlaysSportsbook	PlaysCasino	Play
Male	Poland	English	BETANDWIN.COM	0 days	Yes	No	No
Male	Japan	English	BETANDWIN.COM	0 days	Yes	No	No
Male	United Kingdom	English	BETANDWIN.COM	0 days	No	No	Yes
Male	Slovenia	English	BETANDWIN.COM	18 days	Yes	No	No
Male	Austria	German	BETANDWIN.COM	0 days	Yes	Yes	No
Male	Belgium	English	BETANDWIN.COM	0 days	Yes	No	No
Male	Austria	German	BETANDWIN.COM	0 days	Yes	Yes	No
Male	France	French	BETANDWIN.COM	0 days	Yes	No	No
Male	Italy	English	BETANDWIN.COM	0 days	Yes	No	No
Male	Slovakia	English	BETANDWIN.COM	50 days	Yes	No	No
Male	Spain	English	BETANDWIN.COM	47 days	Yes	Yes	No
Male	Estonia	English	BETANDWIN.COM	1 days	Yes	No	No
Male	Spain	Spanish	BETANDWIN.COM	0 days	Yes	Yes	No

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

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.

UserID	win_vs_stake_product_Sports book fixed-odd	count_product_Sports book fixed-odd	total_bets_product_Sports book fixed-odd	balance_product_Sports book fixed-odd	win_vs_stake_product_Sports book live-action	count_product_Sports book live-action	total_bets_product_Sports book live-action	balance_product_Sports book live-action	win_vs_stake_product_Poker BossMedia	count_product_Poker BossMedia	total_bets_product_Poker BossMedia	b E
1324354	1.00856143	117	236	86.7900	0.8223272	19	43	-326.7900	0	0	0	
1324355	1.13081874	99	231	52.4400	0.4534413	7	21	-13.5000	0	0	0	
1324356	0.41608009	51	98	-400.6800	0.9205973	24	116	-53.9200	0	0	0	
1324358	0.62122488	8	7	-93.8215	0.6319020	1	4	-32.6108	0	0	0	
1324360	0.66594777	29	40	-20.0429	0.6888264	2	3	-0.5425	0	0	0	
1324362	0.00000000	7	7	-22.0000	0.0000000	0	0	0.0000	0	0	0	
1324363	0.00000000	3	3	-41.5300	0.0000000	0	0	0.0000	0	0	0	
1324364	0.03529412	10	11	-82.0000	0.1911905	12	37	-169.8500	0	0	0	
1324368	0.94979571	73	149	-102.1100	0.9981204	60	328	-11.4400	0	0	0	
1324369	0.44442358	40	91	-60.0682	0.8096041	38	112	-28.0576	0	0	0	
1324372	0.98132297	45	198	-7.4814	0.8603930	28	197	-28.2703	0	0	0	
1324377	0.66266208	3	8	-9.4004	0.0000000	1	4	-6.7201	0	0	0	
1324378	0.00000000	7	8	-69.0000	0.0000000	2	4	-11.0000	0	0	0	
1324379	0.74322951	126	440	-737.8300	0.9749126	89	868	-97.5829	0	0	0	
1324383	2.54873045	4	4	1841.9711	0.0000000	0	0	0.0000	0	0	0	
1324385	1.33959111	17	22	1603.4100	0.0000000	0	0	0.0000	0	0	0	
1324386	1.08616083	64	171	314.6729	0.9584620	52	265	-304.3832	0	0	0	
1324393	0.82819731	55	254	-65.0172	0.0000000	0	0	0.0000	0	0	0	
1324396	0.92325403	16	42	-25.0000	0.0000000	0	0	0.0000	0	0	0	
1324400	0.93632601	39	63	-78.0000	0.0000000	0	0	0.0000	0	0	0	

The table above is a glimpse of the final version of the UserDailyAggregation table called Useraggbasetable since it contains more than 50 columns.

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.

UserID	MaxPokerTranDate	MinPokerTranDate	TotalPokerTranSell	TotalPokerTranBuy	MaxPokerTranSell	MaxPokerTranBuy	MinPokerTranSell	MinPokerTranBuy	AvgPokerTranSell	AvgPokerTranBuy	CountPokerTranSell	CountPokerTranBuy	PokerTranDays	FrequencyPokerSell	FrequencyPokerI
1324355	2005-06-15	2005-06-12	8.259300e+00	3.082950e+01	6.452400e+00	8.99990000	1.806900e+00	1.999000e+00	4.129650e+00	5.13825000	2	6	3	0.66666667	2.00
1324368	2005-05-30	2005-04-05	9.620815e+03	9.161658e+03	8.052394e+02	494.99990000	8.000000e-04	7.717000e+01	3.006505e+02	136.74116716	32	67	55	0.58181818	1.21
1324369	2005-03-03	2005-03-03	2.949455e-01	2.949455e-01	2.949455e-01	0.2949452	2.949455e-01	2.949455e-01	2.949455e-01	0.2949452	1	1	0	0.00000000	0.00
1324372	2005-10-02	2005-09-02	9.184137e+02	8.607840e+02	2.525185e+01	13.81175762	2.661912e-03	1.431154e-01	7.917359e+00	7.05560642	116	122	30	3.86666667	4.06
1324379	2005-04-20	2005-04-20	8.828000e-01	9.600000e-01	8.828000e-01	0.96000000	8.828000e-01	9.600000e-01	8.828000e-01	0.96000000	1	1	0	0.00000000	0.00
1324418	2005-04-11	2005-04-11	0.000000e+00	4.349873e-01	6.120000e-03	0.43498727	9.800000e+02	4.349873e-01	3.840000e+01	0.43498727	0	1	0	0.00000000	0.00
1324437	2005-09-19	2005-08-27	2.212834e+01	2.111274e+01	1.200969e+01	6.00484340	3.098208e+00	3.098208e+00	7.376079e+00	5.27818447	3	4	23	0.13043478	0.17
1324573	2005-06-14	2005-06-13	2.391283e-01	1.056917e+00	2.391283e-01	0.69667484	2.391283e-01	1.200567e-01	2.391283e-01	0.35230567	1	3	1	1.00000000	3.00
1324617	2005-02-02	2005-02-02	2.487108e-01	1.099562e+01	2.487108e-01	10.99562373	2.487108e-01	1.099562e+01	2.487108e-01	10.99562373	1	1	0	0.00000000	0.00
1324820	2005-09-24	2005-07-21	5.767792e+03	9.393168e+03	5.759163e+02	400.97000000	3.000000e-04	4.500000e-01	9.612987e+01	123.59431842	60	76	65	0.92307692	1.16
1324851	2005-03-09	2005-03-08	2.423199e+01	4.541783e+01	1.258929e+01	20.15587208	6.653319e-01	6.718597e+00	8.077130e+00	15.13927708	3	3	1	3.00000000	3.00
1324861	2005-07-12	2005-03-14	8.704384e+02	1.593979e+03	9.054703e+01	58.99704903	1.654670e-03	6.195086e-01	1.892257e+01	18.97594404	46	84	120	0.38333333	0.70
1325126	2005-07-24	2005-07-05	1.157401e+02	1.889299e+02	2.000000e+01	20.00000000	1.230000e-02	4.820000e+00	8.903085e+00	11.80811875	13	16	19	0.68421053	0.84
1325189	2005-06-25	2005-02-19	3.059320e+04	3.200045e+04	6.284949e+02	695.43990000	1.296000e-01	1.624990e+01	1.112652e+02	105.61203795	275	303	126	2.18253968	2.40
1325199	2005-02-26	2005-02-03	3.266691e+01	5.953418e+01	1.536470e+01	15.36459813	1.197516e-02	1.344008e+01	8.166728e+00	14.88354457	4	4	23	0.17391304	0.17
1325200	2005-07-25	2005-07-25	8.674030e+00	8.674030e+00	8.674030e+00	8.67402995	8.674030e+00	8.674030e+00	8.674030e+00	8.67402995	1	1	0	0.00000000	0.00
1325318	2005-08-31	2005-08-30	5.485804e+01	7.557787e+01	3.188845e+01	34.17996113	4.422626e-01	4.422626e-01	1.371401e+01	18.89446753	4	4	1	4.00000000	4.00
1325367	2005-04-03	2005-04-03	1.602747e-01	1.602747e-01	1.602747e-01	0.16027471	1.602747e-01	1.602747e-01	1.602747e-01	0.16027471	1	1	0	0.00000000	0.00

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.

DATA MART

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 a 0.

UserID	RegDate	FirstPay	FirstAct	FirstSp	FirstCa	FirstGa	FirstPo	Gender	Country	Language	Application	RegDate_FirstAct	PlaysSportsbook	PlaysCasino	PlaysGames	PlaysPoker	win_vs_stake_product_Sports book fixed-odd	count_product_Sp book fixed-odd
1324354	2005-02-01	2005-02-24	2005-02-24	2005-02-24	NA	NA	NA	Male	Germany	German	BETANDWIN.DE	23 days	Yes	No	No	No	1.00856143	
1324355	2005-02-01	2005-02-01	2005-02-01	2005-02-01	NA	NA	2005-06-11	Male	Greece	Greek	BETANDWIN.COM	0 days	Yes	No	No	Yes	1.13081874	
1324356	2005-02-01	2005-02-01	2005-02-02	2005-02-02	NA	NA	NA	Male	Germany	German	BETANDWIN.DE	1 days	Yes	No	No	No	0.41608009	
1324358	2005-02-01	2005-02-01	2005-02-01	2005-02-01	NA	NA	NA	Male	Sweden	English	BETANDWIN.COM	0 days	Yes	No	No	No	0.62122488	
1324360	2005-02-01	2005-02-02	2005-02-02	2005-02-02	2005-02-03	NA	NA	Male	Turkey	Turkish	BETEUROPE.COM	1 days	Yes	Yes	No	No	0.66594777	

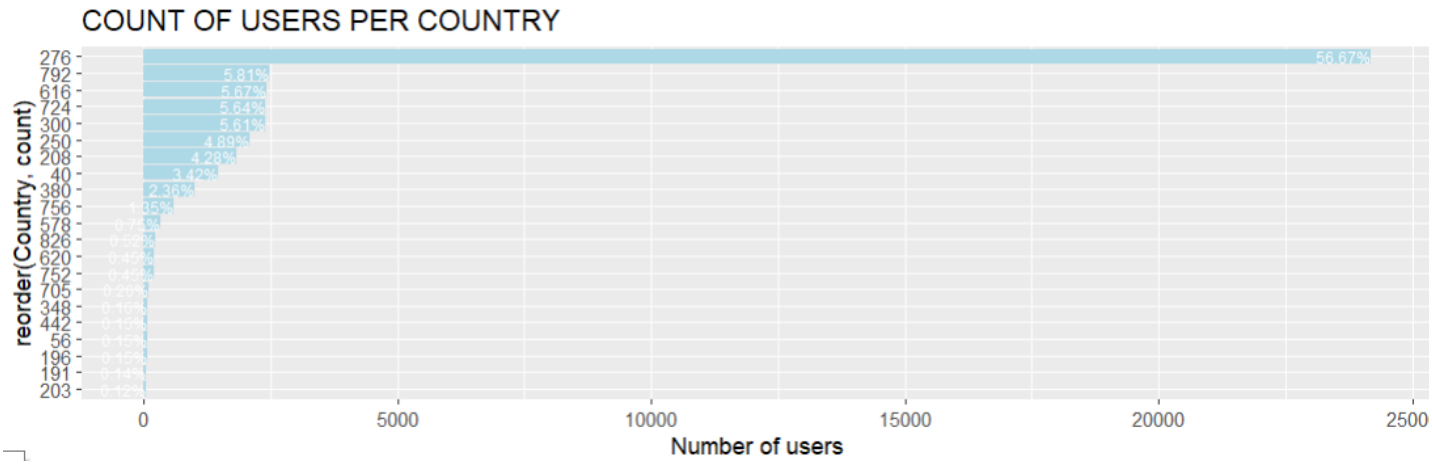
win_vs_stake_product_Sports book live-action	count_product_Sports book live-action	total_bets_product_Sports book live-action	balance_product_Sports book live-action	win_vs_stake_product_Poker BossMedia	count_product_Poker BossMedia	total_bets_product_Poker BossMedia	balance_product_Poker BossMedia	win_vs_stake_product_Casino BossMedia	count_product_Casino BossMedia	total_bets_product_Casino BossMedia
0.8223272	19	43	-326.7900	0	0	0	0	0.0000000	0	
0.4534413	7	21	-13.5000	0	0	0	0	0.0000000	0	
0.9205973	24	116	-53.9200	0	0	0	0	0.0000000	0	
0.6319020	1	4	-32.6108	0	0	0	0	0.0000000	0	
0.6888264	2	3	-0.5425	0	0	0	0	0.0000000	0	
0.0000000	0	0	0.0000	0	0	0	0	0.0000000	0	

count_product_Supertoto	total_bets_product_Supertoto	balance_product_Supertoto	win_vs_stake_product_Games VS	count_product_Games VS	total_bets_product_Games VS	balance_product_Games VS	win_vs_stake_product_Games bwin	count_product_Games bwin	total_bets_product_Games bwin	balance_product_G bwin
0	0	0.0000	0.00000000	0	0	0.0000	0.0000000	0	0	
0	0	0.0000	0.00000000	0	0	0.0000	0.0000000	0	0	
0	0	0.0000	0.00000000	0	0	0.0000	0.0000000	0	0	
0	0	0.0000	0.00000000	0	0	0.0000	0.0000000	0	0	
0	0	0.0000	0.00000000	0	0	0.0000	0.0000000	0	0	
0	0	0.0000	0.00000000	0	0	0.0000	0.0000000	0	0	

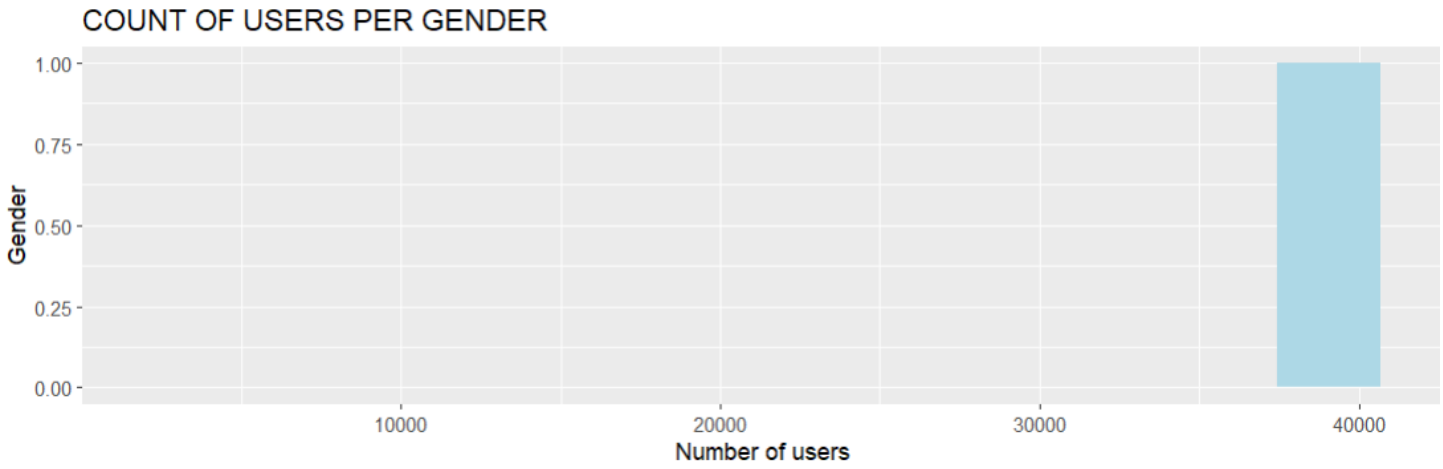
total_bets_product_Casino Chartwell	balance_product_Casino Chartwell	win_vs_stake_month_2
0	0.00	0.00000000
0	0.00	0.91389094
0	0.00	0.71374829
0	0.00	0.74561074
4	-2.00	0.79653583
0	0.00	0.00000000

MARKETING ANALYSIS

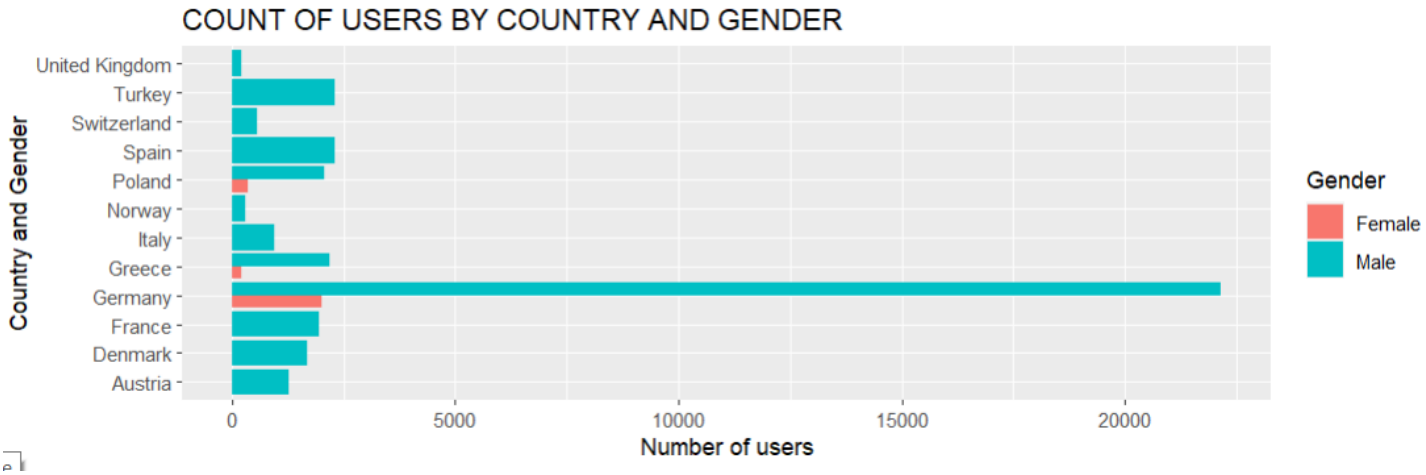
All the tables have been cleaned and merged now. All the valuable variables were created, and we are ready to generate some plots and graphs to better describe the client and could even categorize them in different personas to help the managers and/ or owners better understand their own clients to help 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.



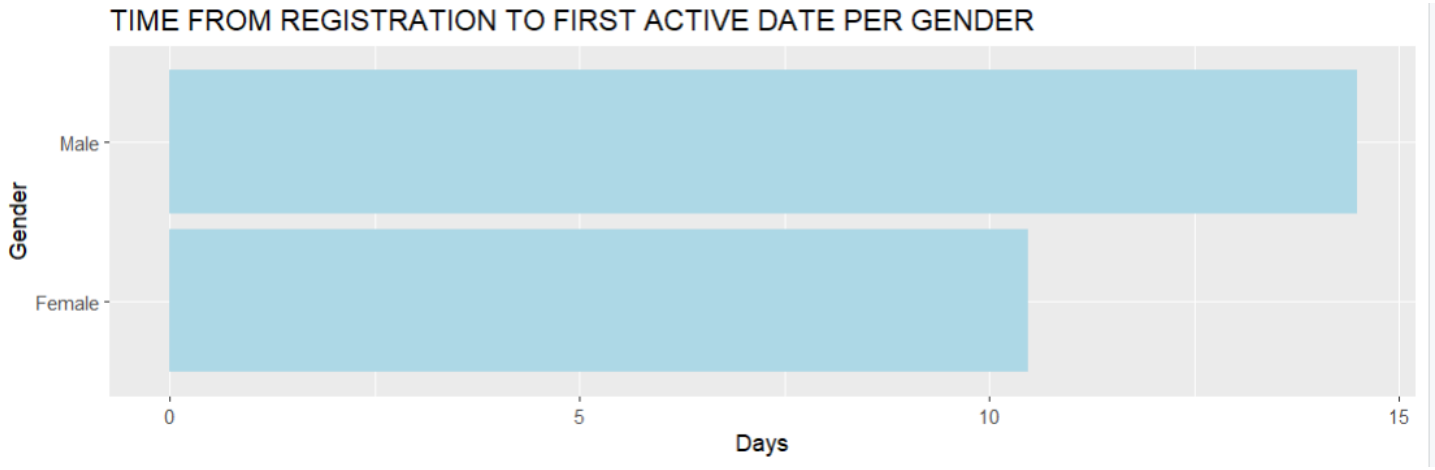
This first graph shows the users distribution per country and we can easily see that the country number 276 which is Germany is leader here by having the highest number of users represented by more than 56% of the total clientele of the company. We can say Germany is a really good country for this business and we can there make a lot of profit.



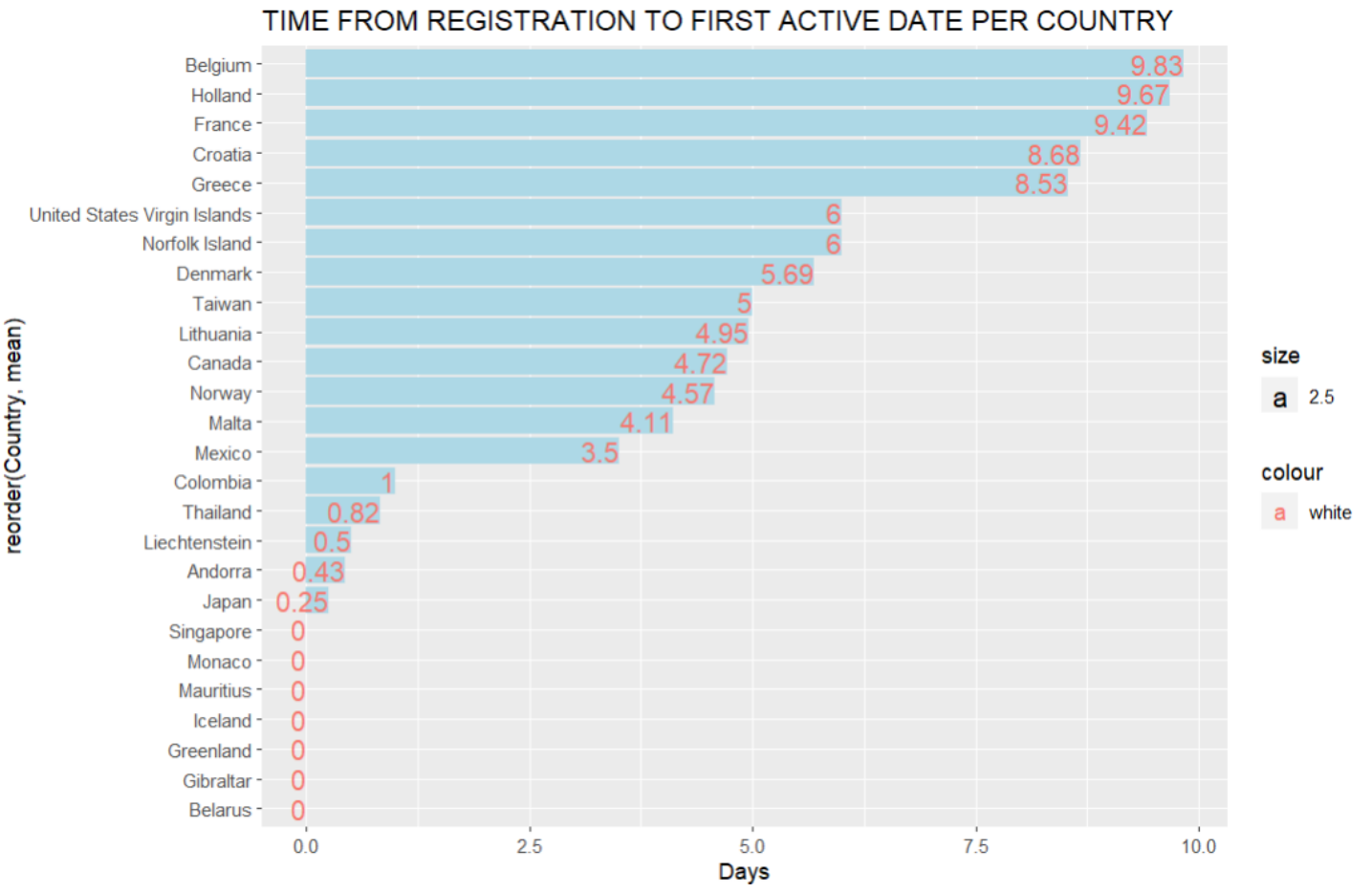
It is obvious that the company would want to know how many female and male clients they have, therefore we plotted the count of users per gender. It is clear that most of the users as more than 90% are male clients and the rest which is the smallest minority are female users. We can conclude that this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



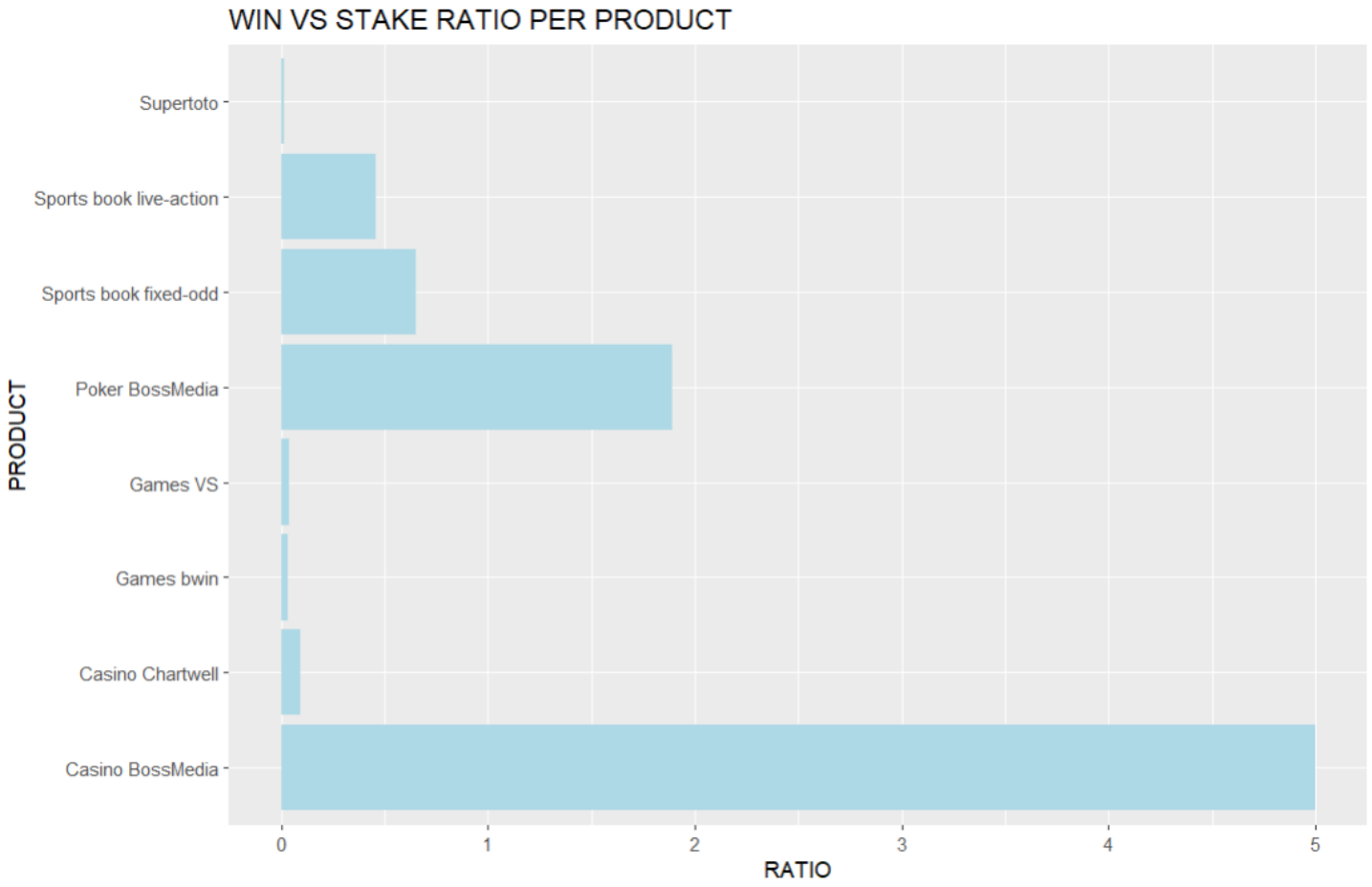
Now we saw if the first graph that Germany has the most market shares in terms of active users by 56% of the entire company's clientele and we also saw that male users are reigning by more than 90%. Let's see now the distribution of the users by country and by gender at the same time. The previous results 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.



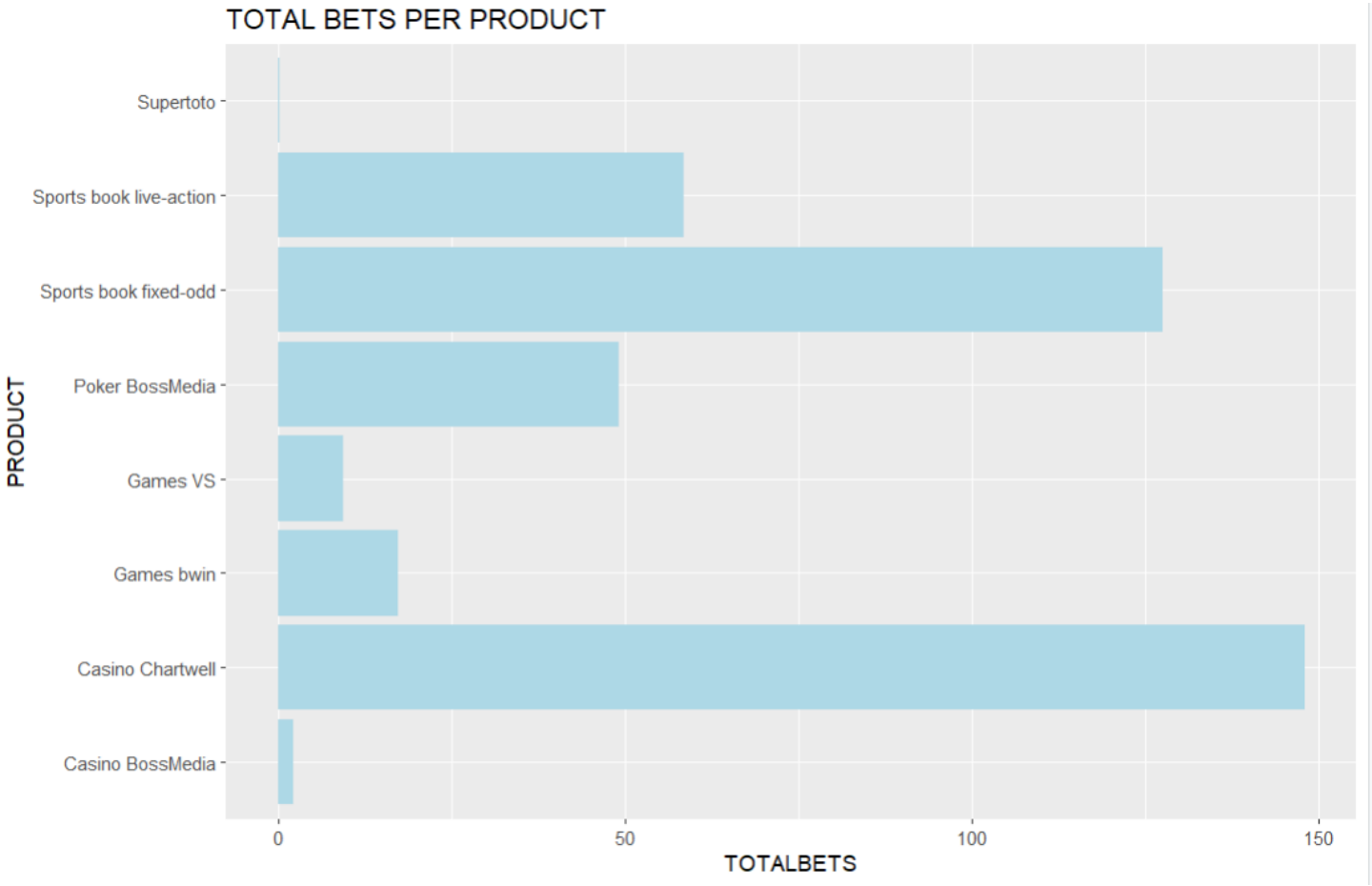
Now that we got the most probable location of the user and their gender, we would like to know at what time they become active. From this graph we see that the male user can take up to 15 days to become active and start playing and the female users only take around 10 days to activate their presence. 10 to 15 days are still a long time to make the client forget about the company and the risk of not playing or being an active user becomes higher and higher, so maybe a good business recommendation would be to use email marketing, SMS marketing and social media marketing to remind the new clients of the company so this time last will shorten by the time.



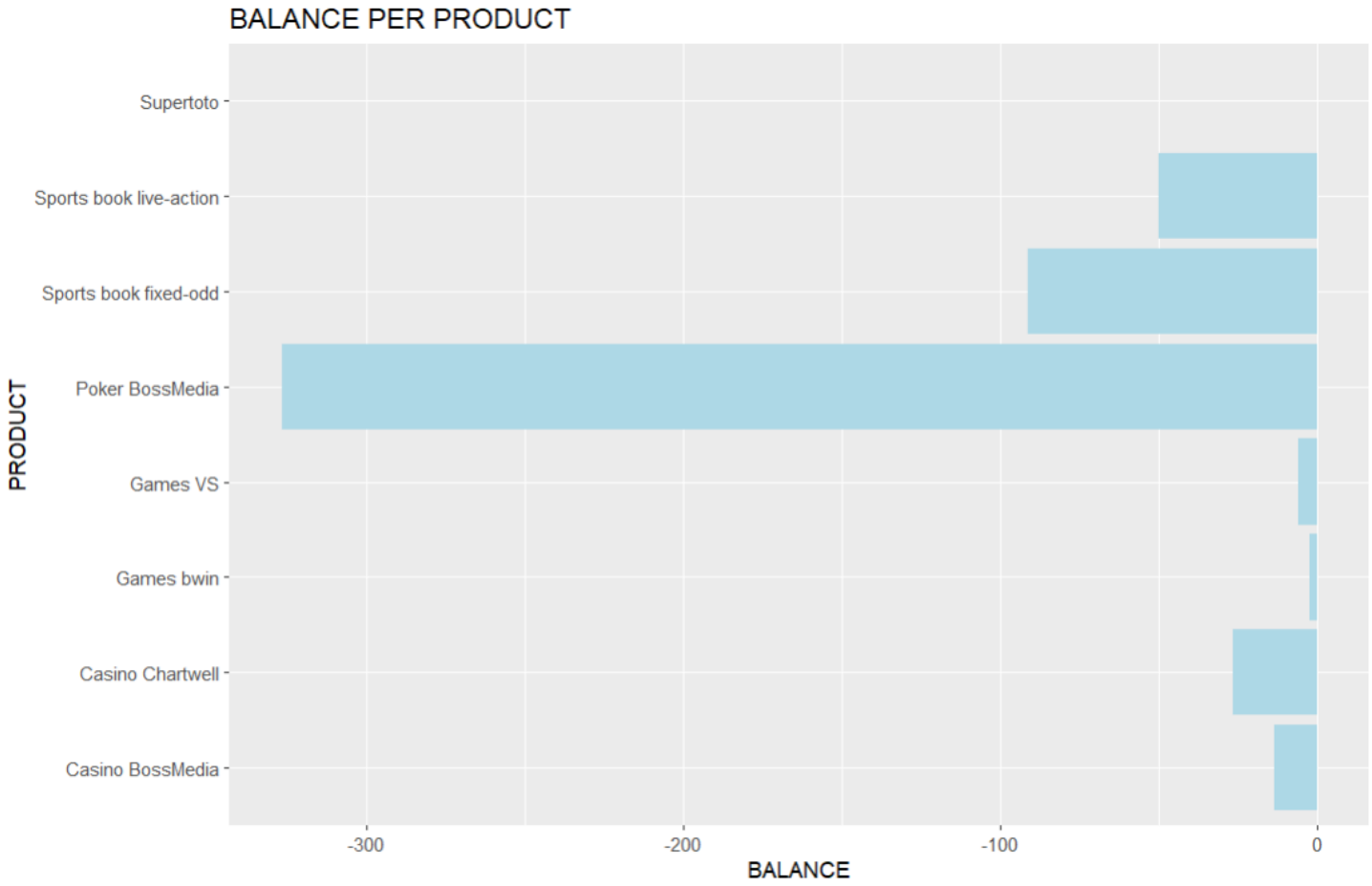
It is a good practice to also see the distribution of when the clients switch to active users per country. Therefore, the above graph was plotted of the time from registration to first active date per country. We can see that Belgium, Holland, France, Croatia and Greece are the countries that have the 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.



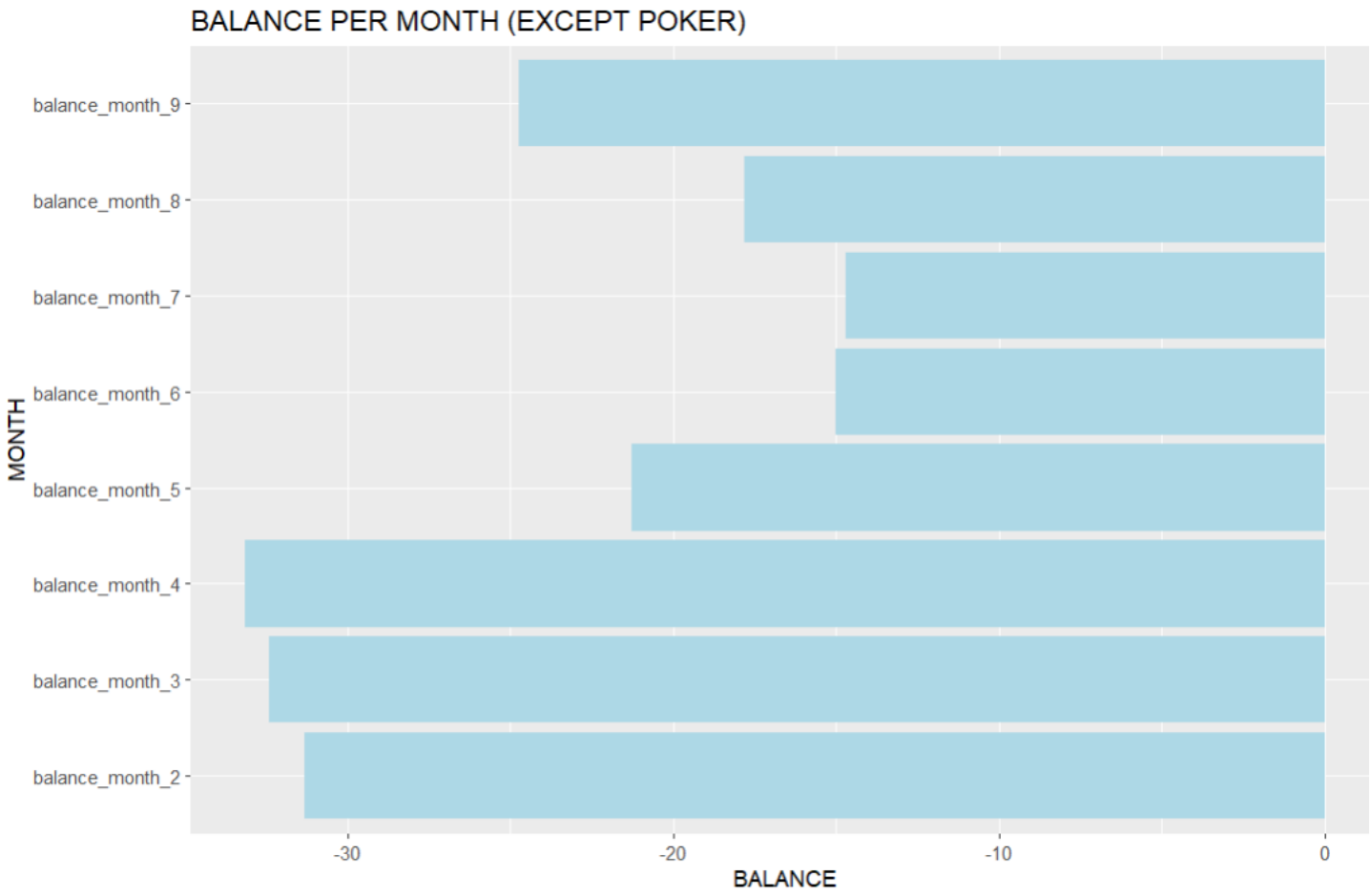
This is a win Vs stake ration per product graph which basically illustrates the distribution of the products per ratios (sell vs buy). We can see that Casino BossMedia has the highest ration of 5, just behind it Poker BossMedia which is the same chain has a ratio of almost 2 and finally Sports book fixed-odd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

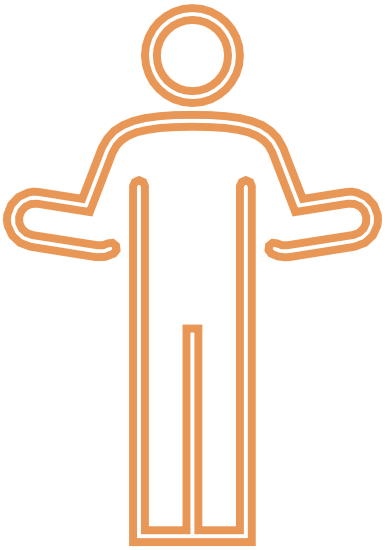
As we previously mentioned, the goal of this study is to be able to understand the customers of the gambling company and categorize them into personas if possible. After cleaning all the tables, studying them, merging everything into a datamart and analyzing the variables with the previous graphs.

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.

56% of the clients come from Germany, and speak German



Casino Boss Media has the highest ratio of 5

Casino Chartwell has the highest total bets of 150.

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.

Hind Raji, Python Project Final Report.