Data Mart Manual

Mohammad Hadi Alipour Motlagh & Omar Mantilla

12/21/2021

BUSINESS REPORTING – OPEN SOURCE

GROUP PROJECT – MARKETING DATA MART

By Mohammad Hadi Alipour Motlagh & Omar Mantilla

INTRODUCTION

The following data mart was conducted to obtain relevant information about the behavior of users on different betting platforms. For this, we have gotten data previously collected during the research project on internet gambling between the division on Addictions (DOA) and BWIN Interactive Entertainment.

To acquire this significant insights we have created a base table which allows us to have a global overview of users and their metrics. This dataset is basically made up of information from different datasets and after a cleaning and transformation process, it was consolidated in one dataset to facilitate calculations and data extraction.

DATASETS

The datasets provided basically consist of the following files:

An .R file that contains three datasets. Those datasets were separated into three csv different files for ease of use.

1. *Demographics_raw*: This dataset contains information for each user such as User ID, Country of residence (Code), Primary Language, Gender, 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.
2. *PokerChipConversions_raw*: This dataset contains information for each user such as User ID, Poker Chip Transaction Date and Time, Poker Chip Transaction Type and Poker Chip Transaction Amount (Euros).
3. *UserDailyAggregation_raw*: This dataset contains information for each user such as User ID, Date of Betting Activities Aggregation, Betting product ID, Total Betting Money (euro), Total Winnings Credited to Participant (Euro) and Total Number of bets.

In addition to the datasets mentioned above, we also have an excel file with data that complements some of the information in the previous datasets that is in code form. This excel file contains four sheets in which we find information such as ProductID, Country Name, Language Description and Application Description. This file was also separated into four different .csv files to facilitate data processing.

The new four .csv files were named in the same way as the information they contain.

- *ProductDescription*
- *CountryName*
- *LanguageDescription*
- *ApplicationDescription*

DATA CLEANING & DATA MERGING

In this data cleaning process, we start with the dataset *Demographics_raw*

- We removed the only two empty cells from the FirstAct variable during the month of February 2005.
- Since the dates are not in an easy way to manipulate, we changed them to a date format which is more user-friendly. This was done for the values of the following variables: FirstPay, FirstAct, FirstSp, FirstCa, FirstGa and FirstPo.
- Continuing with the data treatment and looking for the best visualization and identification of our data, we decode the gender, where 1 was assigned for male and 0 for female. We identify that there is a missing value in this variable. Making use of the mode, we have decided to replace this value with a male one since this is the dominant value in this variable.
- Because the values of the country variable are codes and we want to see this type of information in a more significant way, we have merged the Demographics_raw and CountryName datasets, that way we can have the name of each country.
- In the same way, and to be consistent about the information we want for our base table, we have also merged to this one the, LanguageDescription and ApplicationDescription datasets.
- After renaming the variables Gender_type, Country Name, Language Description and Application Description to Gender, Country, Language and Application respectively, we finally have our *Demographics* dataset ready.

Continuing with the following dataset *PokerChipConversions_raw*

- We use the *PokerChipConversions_raw* dataset to calculate the Total Transaction Frequency, Total Transaction Amount per customer in a new subset dataset called *trans_agg*
- We have created a subset of the data set to identify the sales which are encoded in the TransType variable under the value of 24.
- Once this is done, we can get metrics like, Frequency and Total Amount for each month from February to September in 2005.
- Once we have these new and important variables, we proceed to merge them into a single dataset called *PokerChipConversions_Sell_agg* where our key is the UserID variable.
- In the same way we proceed to calculate these same metrics of Frequency and Total Amount per month but this time taking into account only transactions referring to purchases, which are identified in the TransType variable with the value 124 and merged all the variables in a single dataset called *PokerChipConversions_Buy_agg* where our key is the UserID variable.

Continuing with the following dataset *UserDailyAggregation_raw*

- This dataset also has a Date variable which we convert to a date type to facilitate its processing.
- Having our dataset ready, we proceed to merge it with the *ProductDescription* dataset, which will allow us to have the information about the Product more clearly in a new dataset called *games_raw.* Using this new dataset, we where able to calculate metrics like: all_First_Date, all_Last_Date, Total Duration, Recent_Play, Total Retantion, Recency, Total Game Frequency, Total Game Stakes, Total Game Winnings and Total Game Bets

- Taking into account that we want to obtain significant information about the customers, we decided to segment our population into quantiles and that way we can locate each user in one of them. During this process we acquire the following variables.
- Continuing with the dataset subset to create useful variables. We decided to analyze the behavior of each user against each of the Product IDs and month, which allows us to calculate the metrics described in the metrics calculations section for this dataset.
- When finishing obtaining the new variables, they were created in different datasets, which are merged to create a single dataset per product with monthly information from the users and ProductID.

BASE TABLE MERGING

At the end of our variables creation by ProductID and Month where now we have meaningful information per user, we proceed to make the merge of the different dataset to create our base table. This merge process is done by UserID.

Base_Table *= Demographics + PokerChipConversions_Sell_agg + PokerChipConversions_Buy_agg + UserDailyAggregation_SBFO_agg + UserDailyAggregation_SBLA_agg + UserDailyAggregation_CBM_agg + UserDailyAggregation_Supertoto_agg + UserDailyAggregation_GVS_agg + UserDailyAggregation_GB_agg + UserDailyAggregation_CC_agg.*

METRIC CALCULATIONS

(Bold Letters are meant to be replaced by the values required to calculate the metric)

PokerChipConversions_raw DataSet

- Total Transaction Frequency

trans_agg <- trans_raw [,“Total Transaction Frequency” : = .N, by = UserID]

- Total Transaction Amount

trans_agg <- trans_raw [, “Total Transaction Amount” := sum (TransAmount), by = UserID]

- Time Frame Definition

PokerChipConversions_raw_Sell_**Month** <- PokerChipConversions_raw_Sell [TransDateTime % between % c (“2005-**MM-DD**”,“2005-**MM-DD**”)]

- Sell Frequency per Month (Sell_Freq_**Month**)

PokerChipConversions_Sell_agg_**Month** <- PokerChipConversions_raw_Sell_**Month**[,Sell_Freq_**Month** : = .N , by = UserID]

- Total Amount per Month (Sell_amount_**Month**)

PokerChipConversions_Sell_agg_**Month** <- PokerChipConversions_raw_Sell_**Month** [,Sell_amount_**Month** : = sum (TransAmount) , by = UserID]

- Dropping the duplicate Rows

PokerChipConversions_Sell_agg_**Month** <- unique (PokerChipConversions_Sell_agg_**Month**, by = “UserID”)

UserDailyAggregation_raw DataSet

- all_First_Date

```
games_agg<-games_raw[,all_Fisrt_Date:=Date[1],by=UserID]
```

- all_Last_Date

```
games_agg<-games_raw[,all_Last_Date:=Date[.N],by=UserID]
```

- Total Duration

```
games_agg <- games_raw [, “Total Duration” : = difftime ( all_Last_Date, all_Fisrt_Date, units = “days”), by = UserID ]
```

- Recennt_Play

```
games_agg <- games_raw [, “Recennt_Play” : = as.integer ( difftime ( “2005-09-30”, all_Last_Date, units =“days”) ), by = UserID ]
```

- Total Retantion

```
games_agg<-games_raw [, “Total Retantion” : =. ( ifelse ( all_Second_Date == all_Fisrt_Date + days (1), TRUE, FALSE)), ]
```

- Recency

```
games_agg <- games_raw [, “Recency”: = .( ifelse ( all_Last_Date >= “2005-09-23”, TRUE, FALSE)), ]
```

- Total Game Frequency

```
games_agg<-games_raw[,“Total Game Frequency”:=.N,by=UserID]
```

- Total Game Stakes

```
games_agg<-games_raw[,“Total Game Stakes”:=sum(Stakes),by=UserID]
```

- Total Game Winnings

```
games_agg<-games_raw[,“Total Game Winnings”:=sum(Winnings),by=UserID]
```

- Total Game Bets

```
games_agg<-games_raw[,“Total Game Bets”:=sum(Bets),by=UserID]
```

- Quantile Calculations

```
dataset$variable<- as.integer( dataset$variable ) quantile ( dataset$variable, probs = seq( 0, 1, 0.25 ), na.rm = TRUE )
```

- Defining the ProductID subset

UserDailyAggregation_raw_**Product_Name** <- UserDailyAggregation [ProductID==**Product_Code_number**„]

- First Play Day (**Product_Name_**Fisrt_Date)

UserDailyAggregation_**Product_Name**_agg<-UserDailyAggregation_raw_**Product_Name** [, **Product_Name**_Fisrt_Date:=Date[1],by=UserID]

- Last Play Day (**Product_Name**_Last_Date)

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_ **Product_Name** [, **Product_Name**_Last_Date:=Date[.N],by=UserID]

- Duration between First and Last play day

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [,**Product_Name**_Dur: = difftime (**Product_Name**_Last_Date, Product_Name_Fisrt_Date,units = "days") , by = UserID]

- One Day Retention Calculation

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_ **Product_Name** [,**Product_Name**_Second_Date:=Date[2],by=UserID]

- Customer Return after One Day

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [, **Product_Name**_1Day_Retention: = .(ifelse (**Product_Name**_Second_Date == **Product_Name**_Fisrt_Date+days (1) ,"Returned","Not_Rerturned")),]

- Customer product usage in last week

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [, **Product_Name**_Recent: = .(ifelse (**Product_Name**_Last_Date > = "2005-09-23"," Recent "," Not_Recent ")) ,]

- Play Frequency

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [,**Product_Name**_Freq_Total:=.N,by=UserID]

- Total Stakes

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [, **Product_Name**_Stakes_Total : = sum(Stakes) , by = UserID]

- Total Winning

UserDailyAggregation_**Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [, **Product_Name**_Winnings_Total : = sum (Winnings), by = UserID]

- Total Bets

UserDailyAggregation_ **Product_Name**_agg <- UserDailyAggregation_raw_**Product_Name** [, **Product_Name**_Bets_Total : = sum (Bets), by = UserID]

- **Monthly Calculations:** To make the monthly calculations we add a line of code to define the period of time.

UserDailyAggregation_raw_**Product_Name_Month** <- UserDailyAggregation_raw_**Product_Name**
[Date % between % c (“2005-**MM-DD**” , “2005-**MM-DD**”)]

- Monthly Play Frequency per User

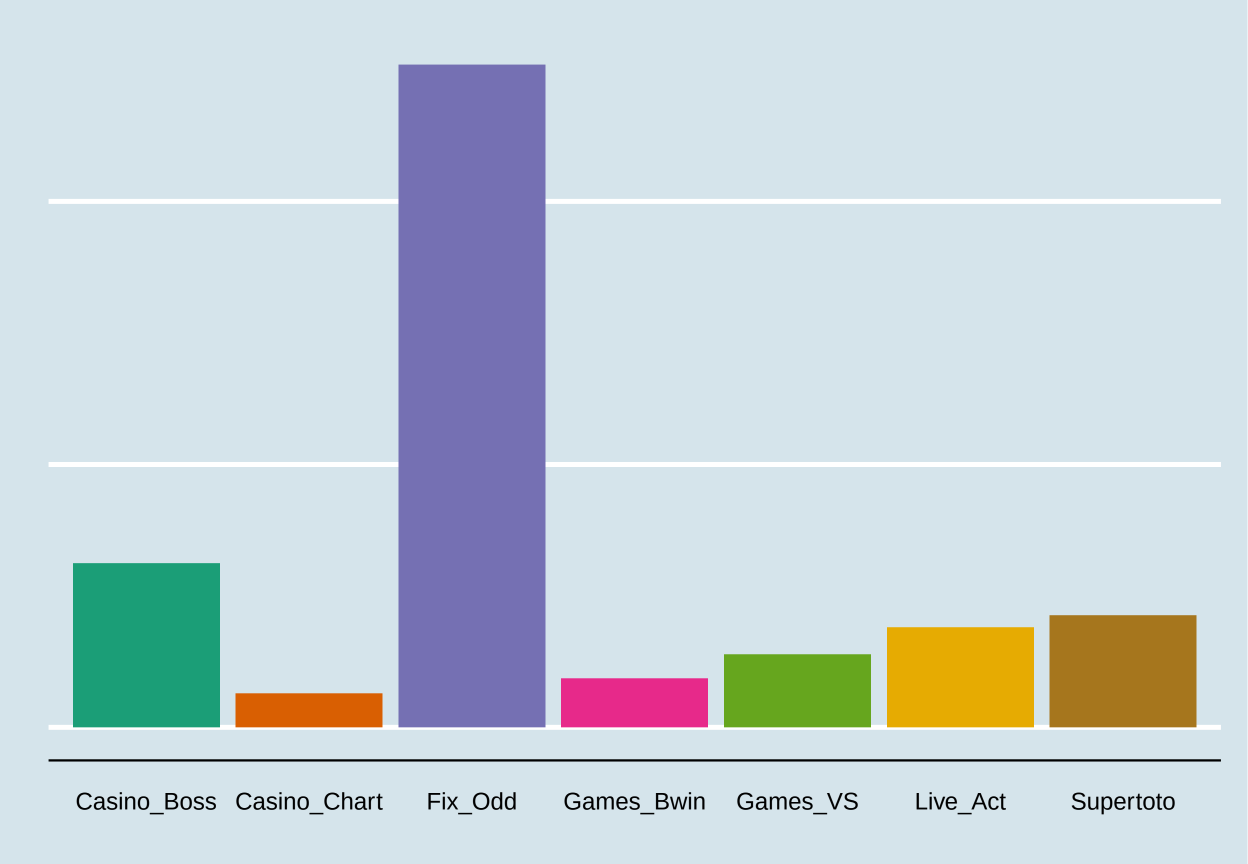
UserDailyAggregation_**Product_Name**_agg_**Month** <- UserDailyAggregation_raw_**Product_Name_Month**
[, **Product_Name**_Freq_**Month** : = .N , by = UserID]

- Monthly Total Stakes per User

UserDailyAggregation_**Product_Name**_agg_**Month** <-UserDailyAggregation_raw_**Product_Name_Month**
[, **Product_Name**_Stakes_**Month** : = sum (Stakes), by = UserID]

CUSTOMER BEHAVIOUR AND MARKETING INSIGHTS

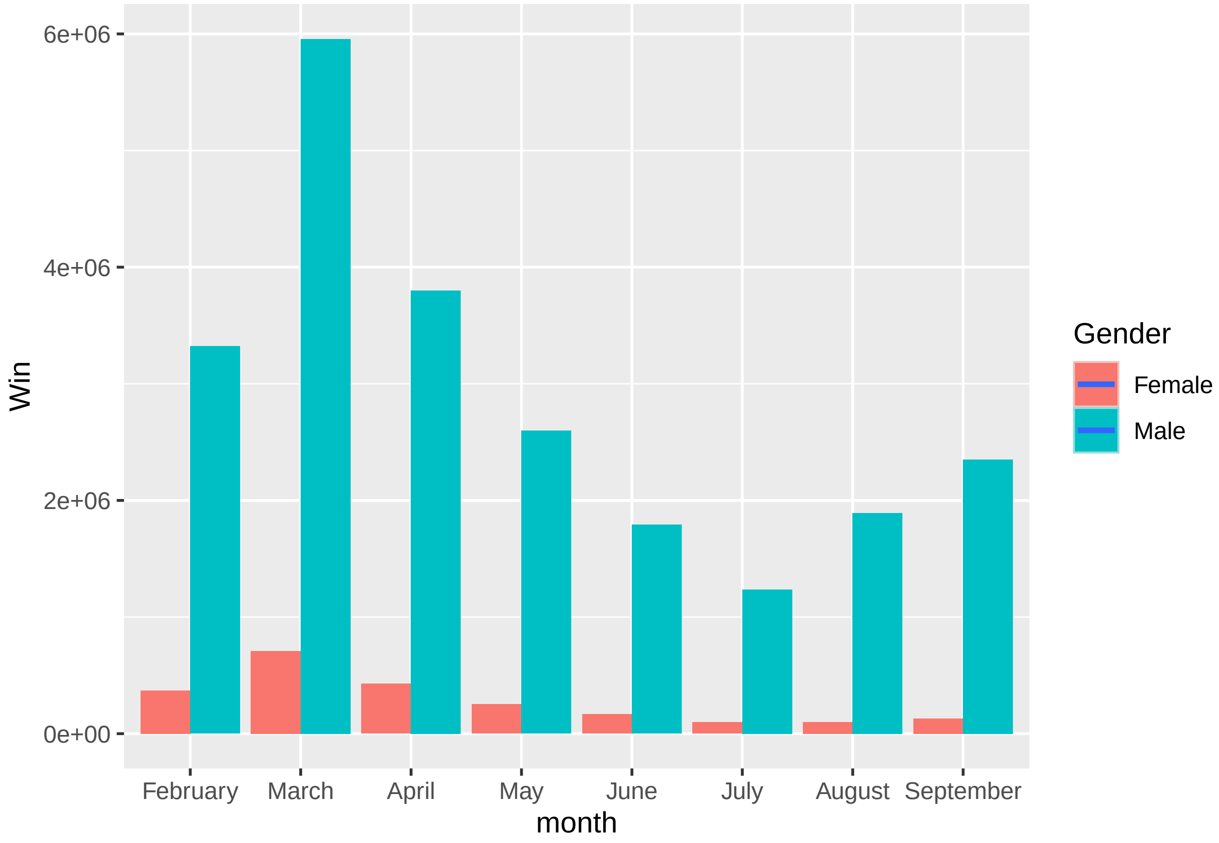
Most Popular Game



One of the interests of the business is to identify which product attracts the largest number of users. In this case we can identify that the Top 3 are:

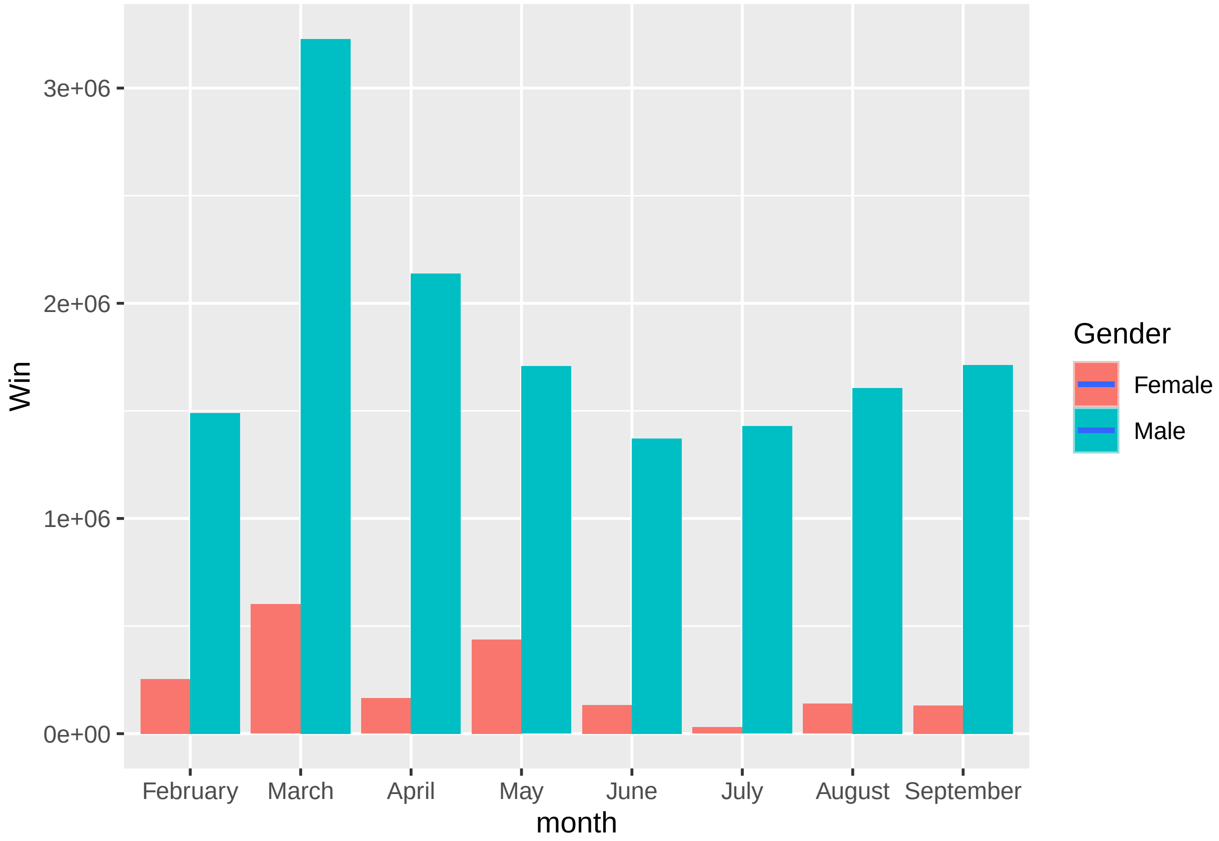
1. Fix Odd.
2. Casino Boss.
3. Superloto.

Fix Odd Wins per Gender



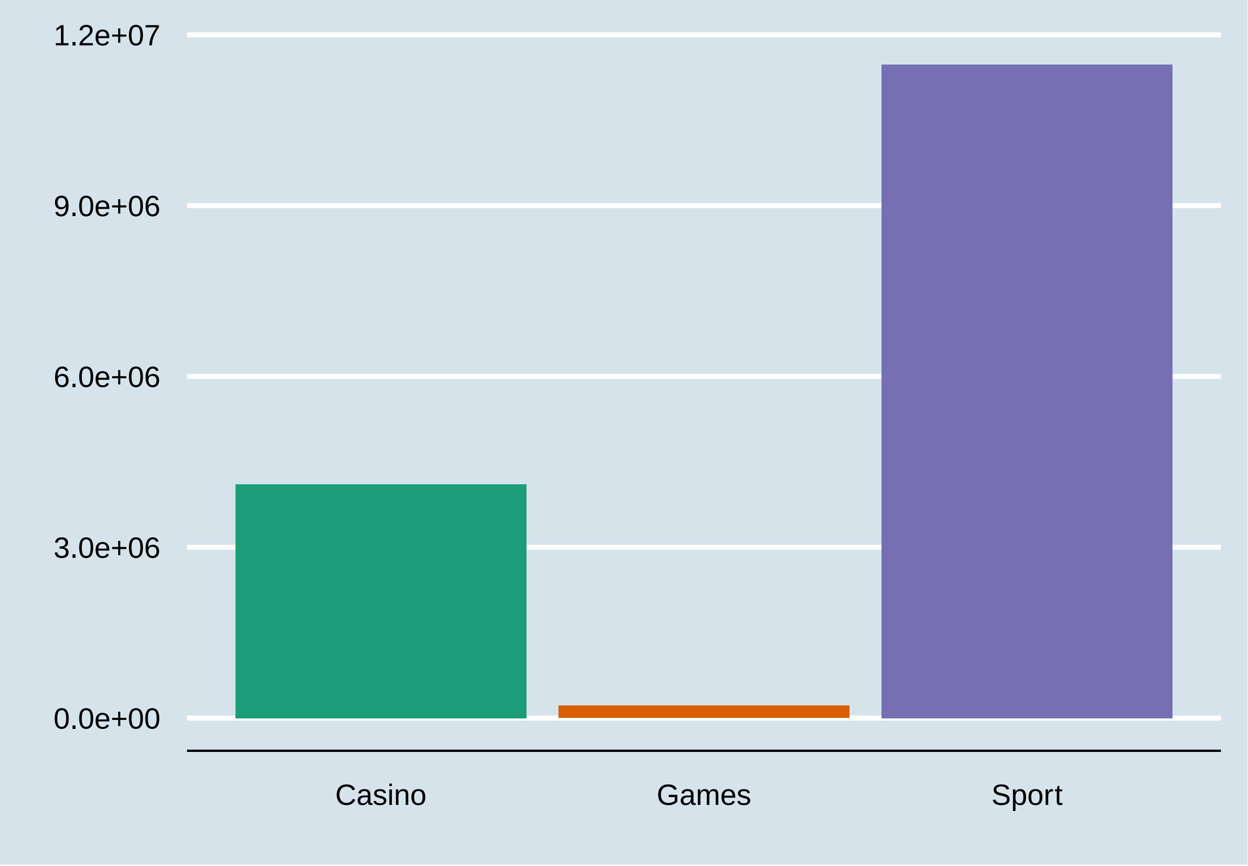
In the case of Fix Odd, we can segment the populations by gender, where we see that the male gender is the most dominant group in this product in terms of victories. We can also appreciate that the months where this behavior occur the most are the months of March, April and February. While the female gender has a considerably reduced participation in general and also makes a greater use of this product in the same months as the male gender

Casino Boss Wins per Gender



Analyzing our next product in our Top 2, we can see that Casino Boss has a similar behavior, with differences in the population of the female gender, where the months where the victories are most presented are in the months of March, May and February. In the case of the male population, this behavior occurs in the months of March, April and May.

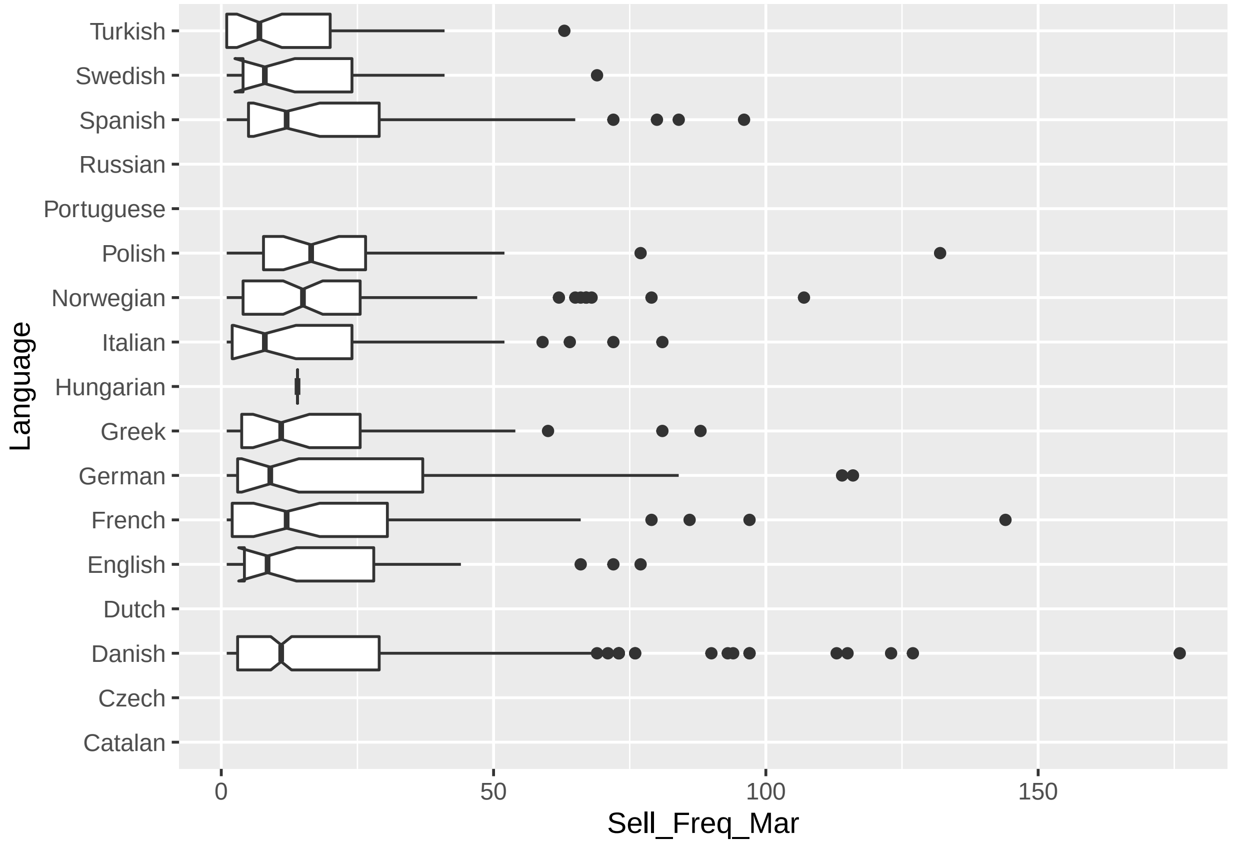
March Chosen Platforms



Reviewing the month in which more victories were presented. We see that the products in which these occur in order are:

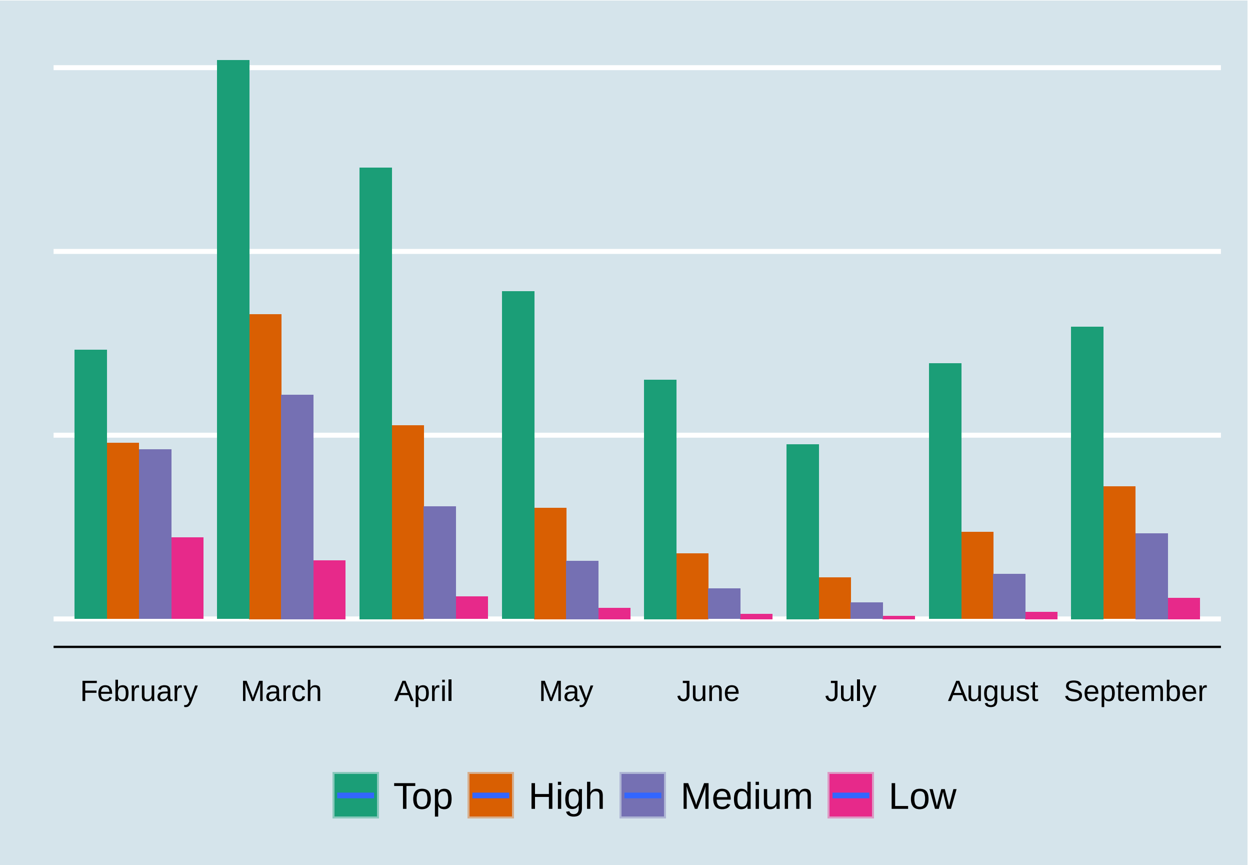
1. Sports.
2. Casino.
3. Games

March Customer Sell Frequency per Language Chosen



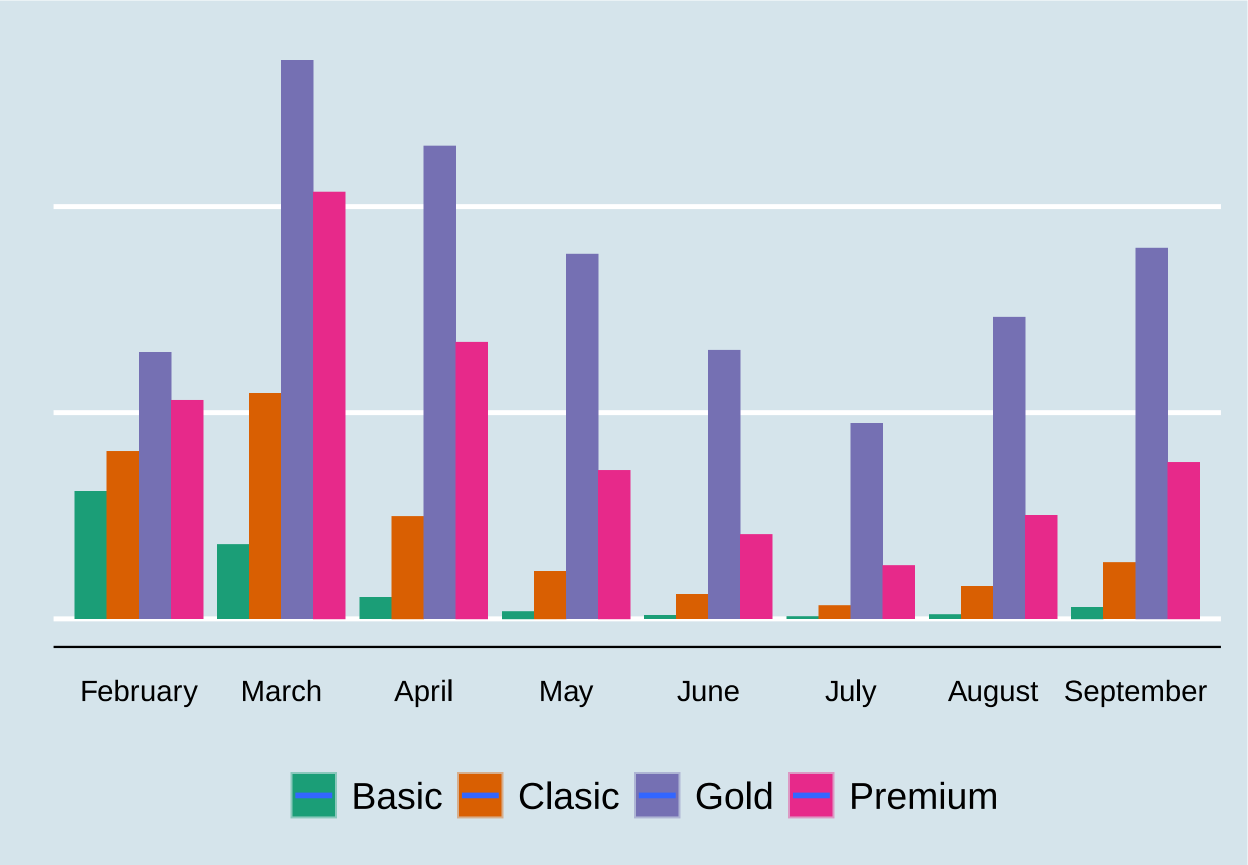
Analyzing at the language level, we can see that the users who are German-speaking were the ones who collected the most prizes in March, followed by French, Danish and Spanish.

Fix Odd Engagement



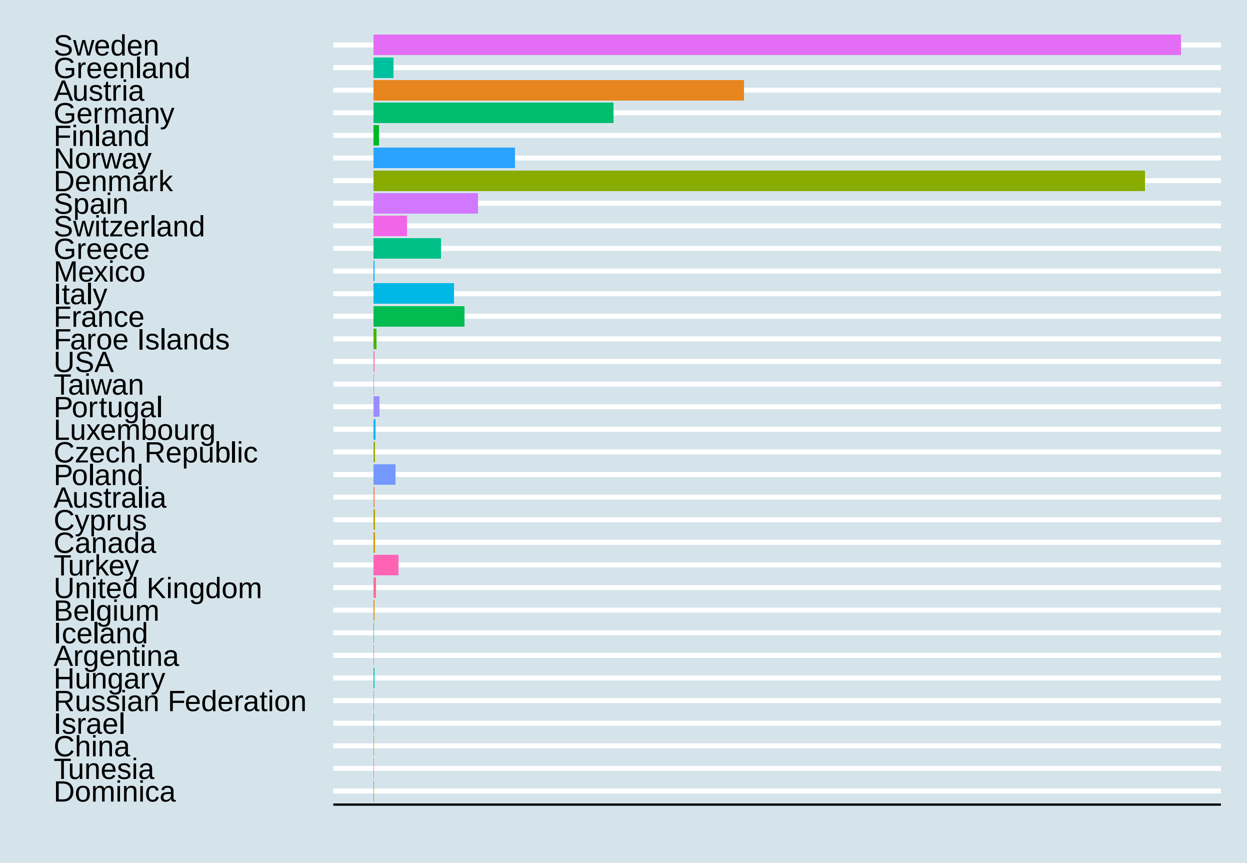
Taking into account that our most widely accepted product is Fix Odd, we wanted to analyze the behavior of our users and identify the level of engagement. Where we can highlight that during the months of March, April and May is where this behavior occurs most and on the contrary, the months of July and June, are the months in which this behavior is reduced almost by half.

Loyalty Frequency



Wanting to identify the behavior of the customers, we decided to analyze, in which months they make the most usege of the products. In this way, we can see that in all the months the group of Loyalty Gold customers makes greater usage of our products followed by the Premium, Classic and Basic Loyalty groups, with a minimal difference in the month of February where the clients of the Gold and Premium loyalty group will become quite the same. Something to rule out is that the basic loyalty group shows low indicators in the months of April to September, but in February this group increases considerably compared to the other months.

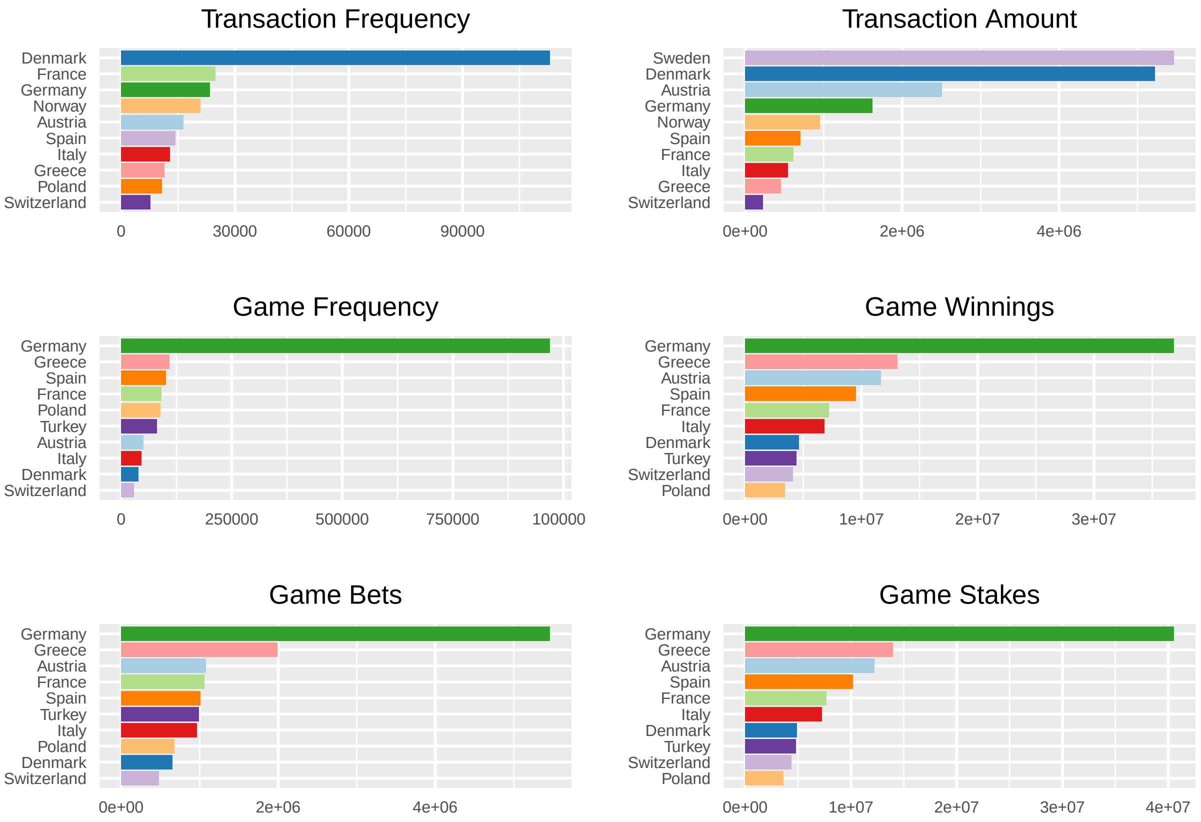
Country Visitis



Having analyzed the different types of users, we wanted to have a clearer picture of the countries from where our products are most consumed. In this way, we were able to detect that users located in Sweden and Denmark represent the majority of customers, followed by customers located in Austria and Germany.

Comparing Diferent groups of customers

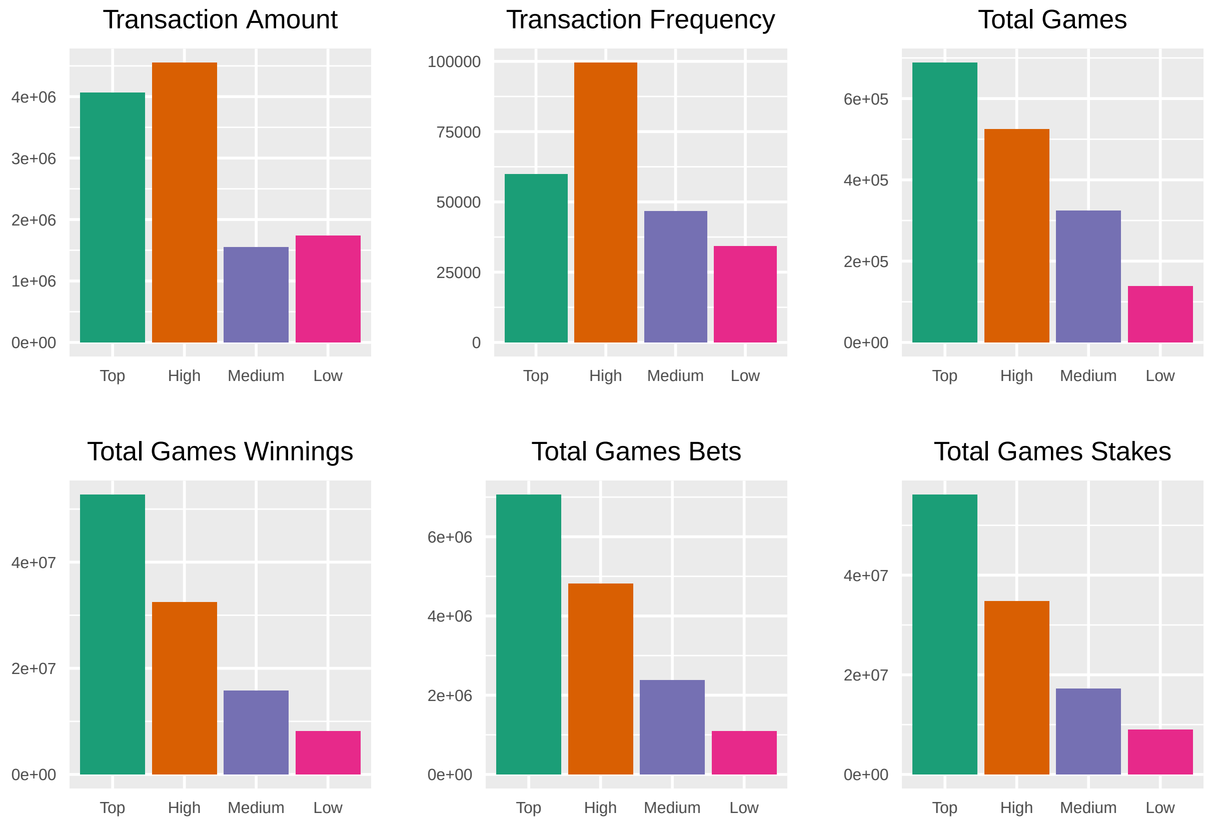
Comparing Countries



KPIs by country

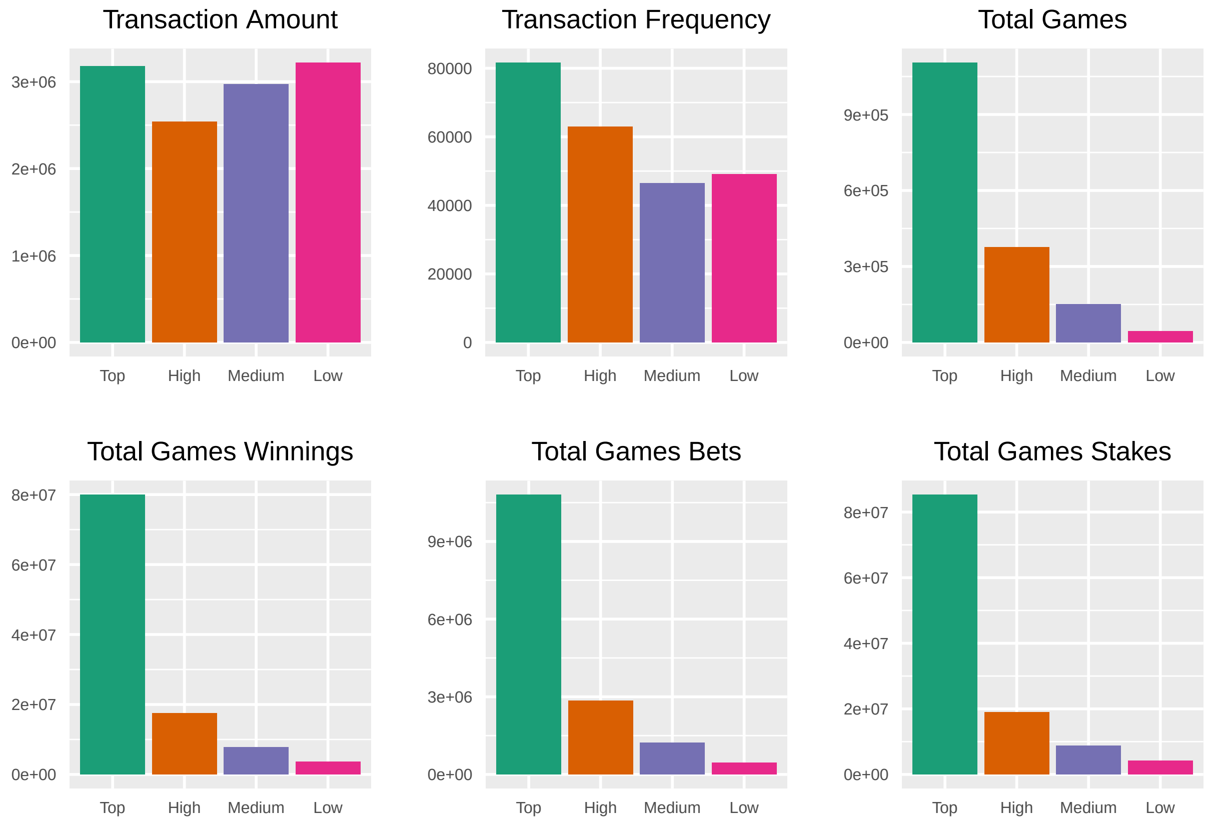
As we can see the most active users of the games is in Germany and Greece, however when we look at the transaction amounts we can see that Sweden is the first country with the most transaction amount, and for the transaction frequency Denmark is in the top, we can say that German users have more engagement and they don't want to put their money out of the game, they use their earnings to play more and they have more bets and more winnings. We don't have good knowledge about the business model of the company, but as we know there are so many types of revenue models and cooperation with third-party entities, we think that if the business model of the company is based on the transaction amount and number of transaction, then its better to focus on the countries like Sweden and profile these customers and use the profiling results in next marketing campaigns to attract more customers from Sweden and more customers from another country which have the same profile like Swedish clients. if the company's business model is based on engagement and for example, the main revenue stream is from advertising then the company needs more active users, then the profiling and segmentation and the positioning of marketing campaign must be designed based on German and german client profile. however, the third thing that comes to mind is this company can use mixed kinds of client profiling for maximizing profit and boosting different income streams.

Recency Groups



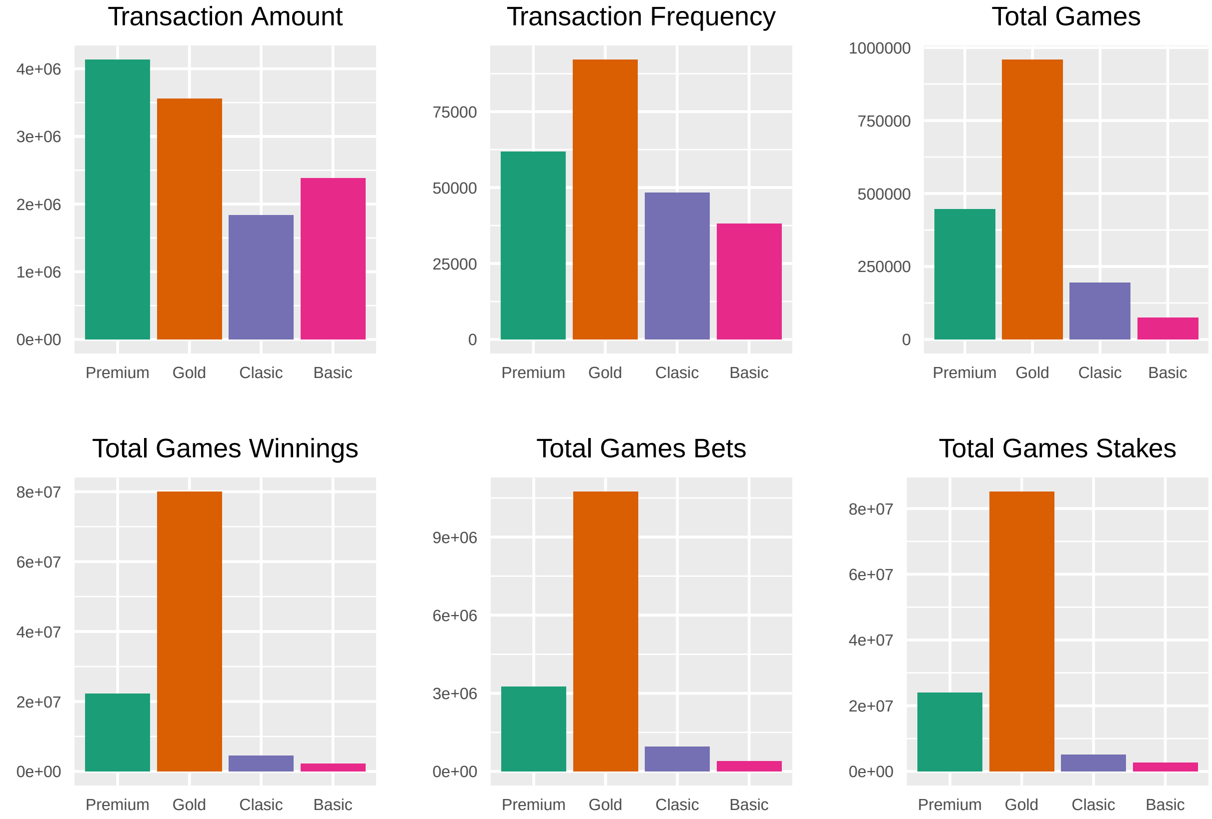
As we can see the group of customers with most recent activities have played more than other customers. But the transaction frequency and amount of transaction of second group of customers are more than the most recent ones. We can say that customers with which probably are not from Germany and maybe are from Sweden or Denmark are amount the users who came less frequently to the website and they don't want to have their money in the website. We can say that maybe these are not our loyal customers and they came to try playing some games from time to time and after they will come back in longer period than the loyal customers with high engagement rate.

Frequencey Groups



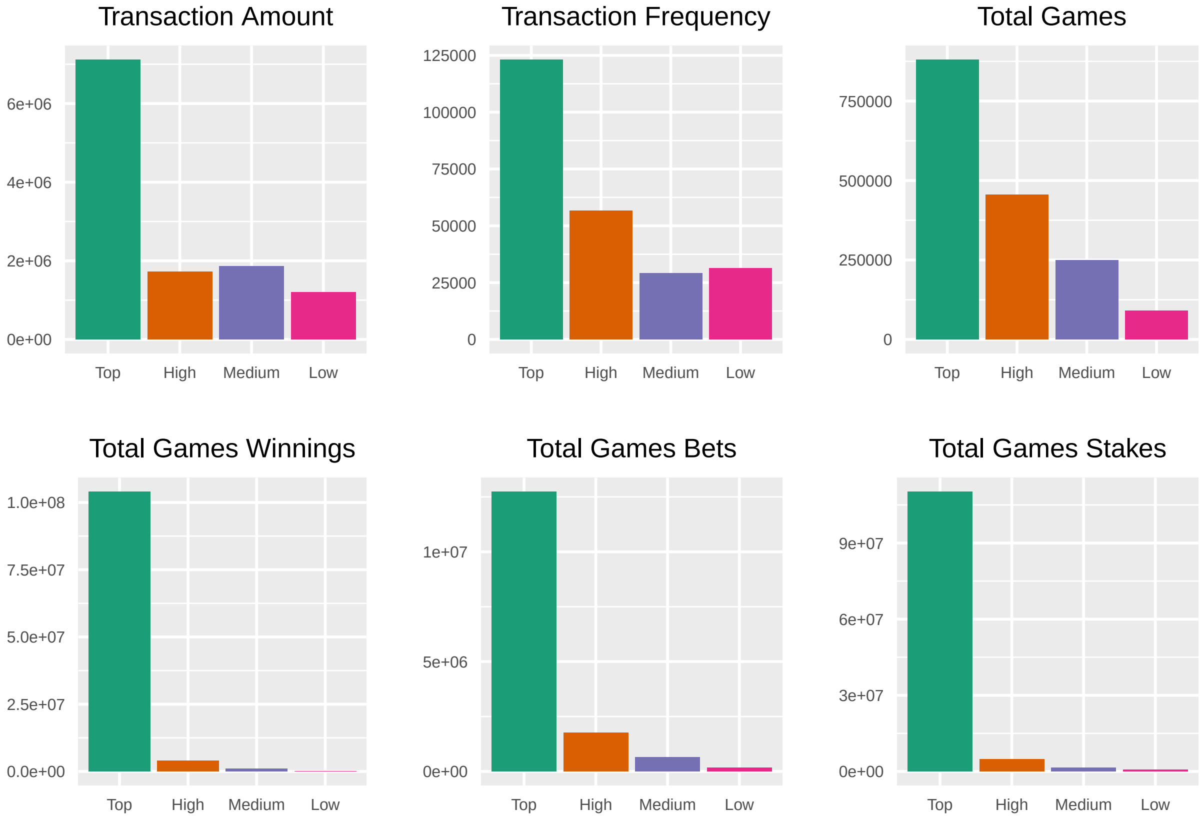
We can see that the total amount of transactions of all different groups of customers based on their frequency of playing the games have no meaningful difference, but as we expect, the more the customer came to the site, the more they played.

Loyalty Groups



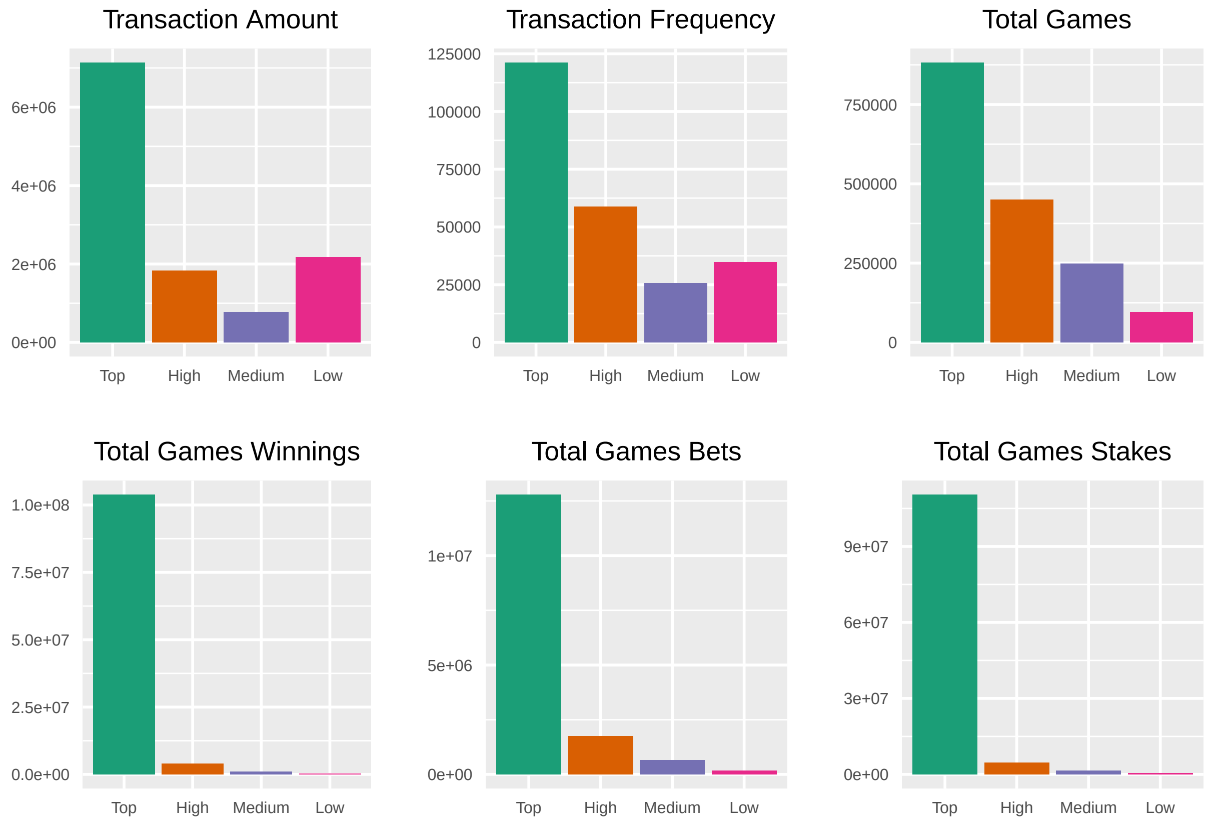
The loyalty factor is based on the RFM score and the calculation of the RFM was based on the scoring the different quantiles of the Recency, Frequency and Monitory and combining them together. Categorizing the customers based on these factor to 4 loyalty groups. We can see that the company earned more amount of transactions from loyal customers but the level of engagement of the Gold customers is higher that the Premium (most loyal) customers. If we had the age of these customers we could guess better but we think that the gold customers are younger that the premium ones and based on their age, their average salary must be lower than the premium group. Therefore the transaction frequency and level of activity of these customers are the heist, however because of the salary average, they spend less amount and have smaller bets comparing to the premium customers.

Winning Groups



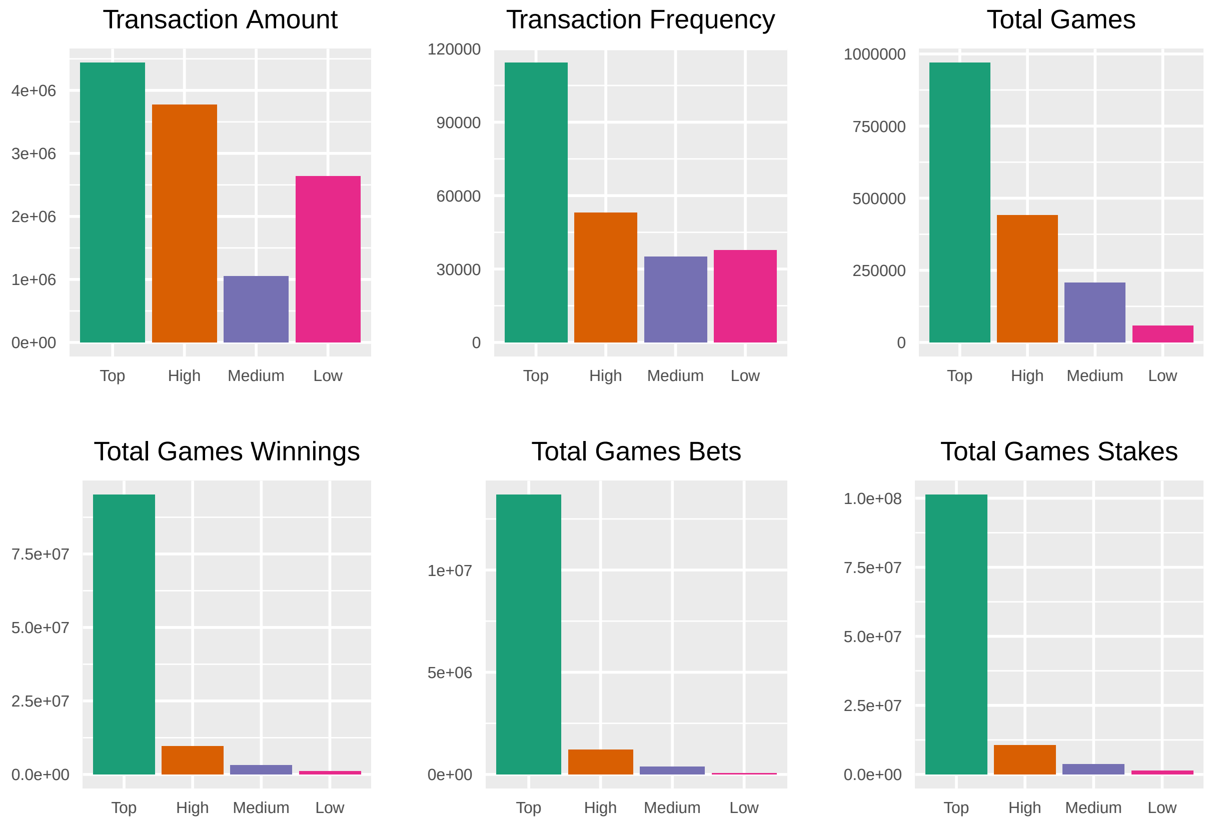
When somebody think about the gambling, maybe it comes to mind that if the clients win more, then the company will loose the money, but in reality it is the cash flow and the number of transactions that matter more than the amount of the transactions, and the main goal of these businesses is to engage more and make customers to come to play games more frequently and selling different products. As we can see here the main transaction and revenue stream comes from the winners and they played more, have more bets, win more and take more risk compare to other groups. As we told earlier, based on the revenue model of these websites that could be on the transaction and the advertisement and also data analysis, we can say that this platforms are win-win platforms that can make good revenue and also have happy customers. We don't have enough time but we propose to cross check these factors together. If we can have a cross tabs based on the loyalty and engagement and more specific demographic factors we can profile our customers better and then based on the business model and strategy plan of the company we can target our customers.

Stakes Group



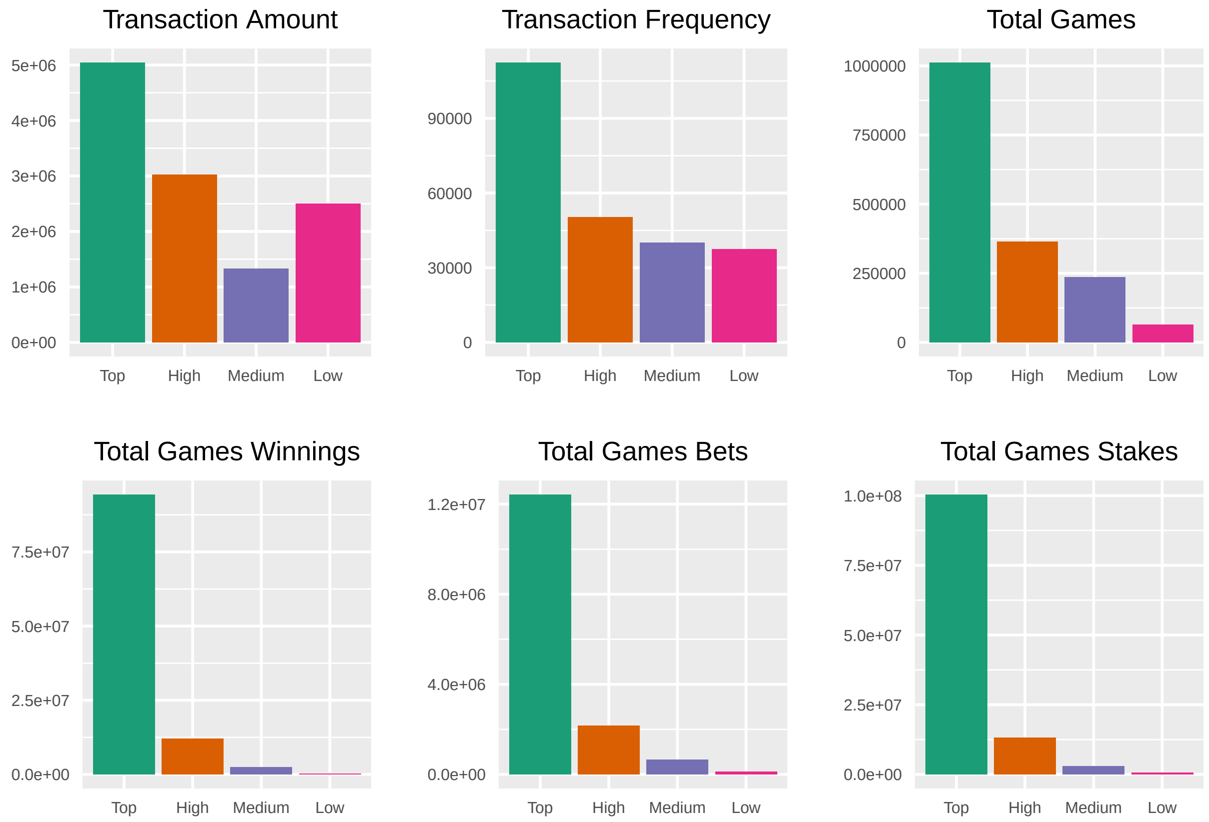
As we discuss in above the frequency of playing is more important than the money that somebody want to put in the game. The evidences in this page also tell us that number of stakes customers have is the most important factor for the busines.

Bets Group



Here again we can see that maybe some people bets more than the other but it doesn't mean that they win more, we had to check the cross values between frequency and bets amount, by having cross table we can say with more confidence that the frequency of bets is more important than the amount of bet, it is important either for the business and also for the gamblers.

Engagement Groups



As we expect the higher the level engagement led to more game winning and playing, however beside this when we look at the graphs of transaction frequency and amount we can see that the less engaged group spend more and played less, this group is probably the clients who are amateur therefore they spend lot at first and get disappointed after loosing their money and haven't come back again.

MARKETING SUMMARY INSIGHTS

When we look at the amount of the money spent by to groups of clients based on the gender at first maybe we think, the most profitable customer group is men, but when we compare the distribution of the frequency and amount spent by these two groups we will see the boxplots of these two groups are same with equal interquartile size. Then the main reason of this huge difference is the number of women clients compare to men. We propose that the company is better to focus on having more women client and this can be achieve through a good strategic plan which consist of marketing and promotions and also focusing on product development to have new products or change some part of existing products to be more convenient for female customers.

Other Items of insights

1. The products most consumed by users are Fix Odd, Casino Boss and Supertoto.
2. The months in which the most victories are presented in the Fix Odd product are the months of March, April and February.
3. The gender that obtains the most victories in the Fix Odd product is the male gender.
4. The months in which there are more victories in our second Casino Boss product are March, April and May.
5. During the month of March, the largest number of transactions are carried out at sporting events followed by casino activities.
6. The highest level of engagement for Fix Odd occurs in the months of March, April and May.
7. The loyalty group that most consumes the products is Gold, followed by Premium, Classic and Basic.
8. The countries that consume the most the products are Switzerland, Denmark, Austria and Germany.