Gambeling Analysis Manual

Dimitri Kestenbaum, Nixia Sancy, and Inder Rana

12/17/2021

Figure 1: IESEG Logo

Analysis Overview

The purpose of this manual is to provide our marketing analysis team with a high-level overview of the relevant statistical summaries of this descriptive analysis project. Provided additionally, are explanations for the different features or variables which we engineered as a team in order to extract additional insight from the dataset, which can be used in future analysis such as segmentation or predictive analytics.

Basetable Variable Summary Statistics:

Total Bets Stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	0.0	16.0	59.0	359.4	196.0	193442.0

Total Stakes Stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	0	74	227	2744	807	1127196

Total Wins Stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	0	34	170	2555	699	1093423

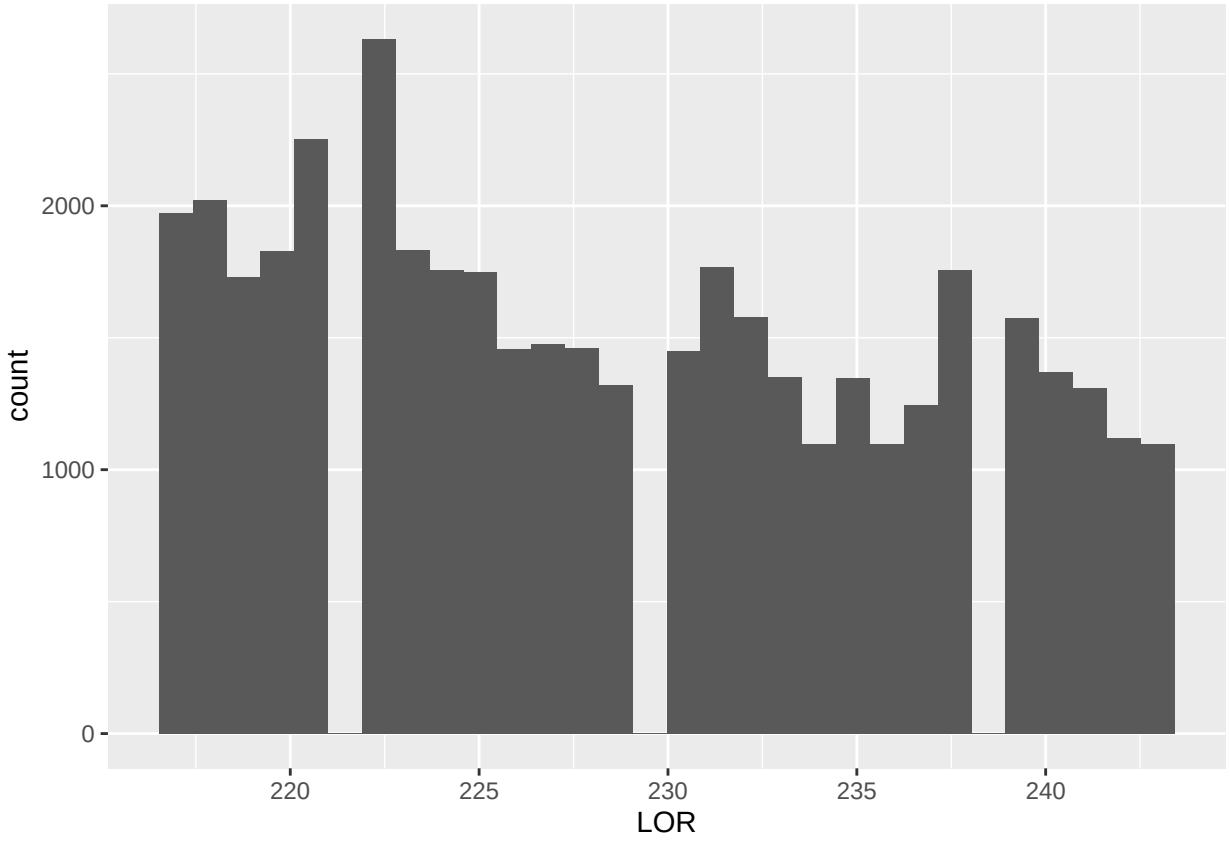
Total Profit Stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
FALSE	-37038.0	15.0	47.0	188.4	120.0	76165.0

Created Metrics:

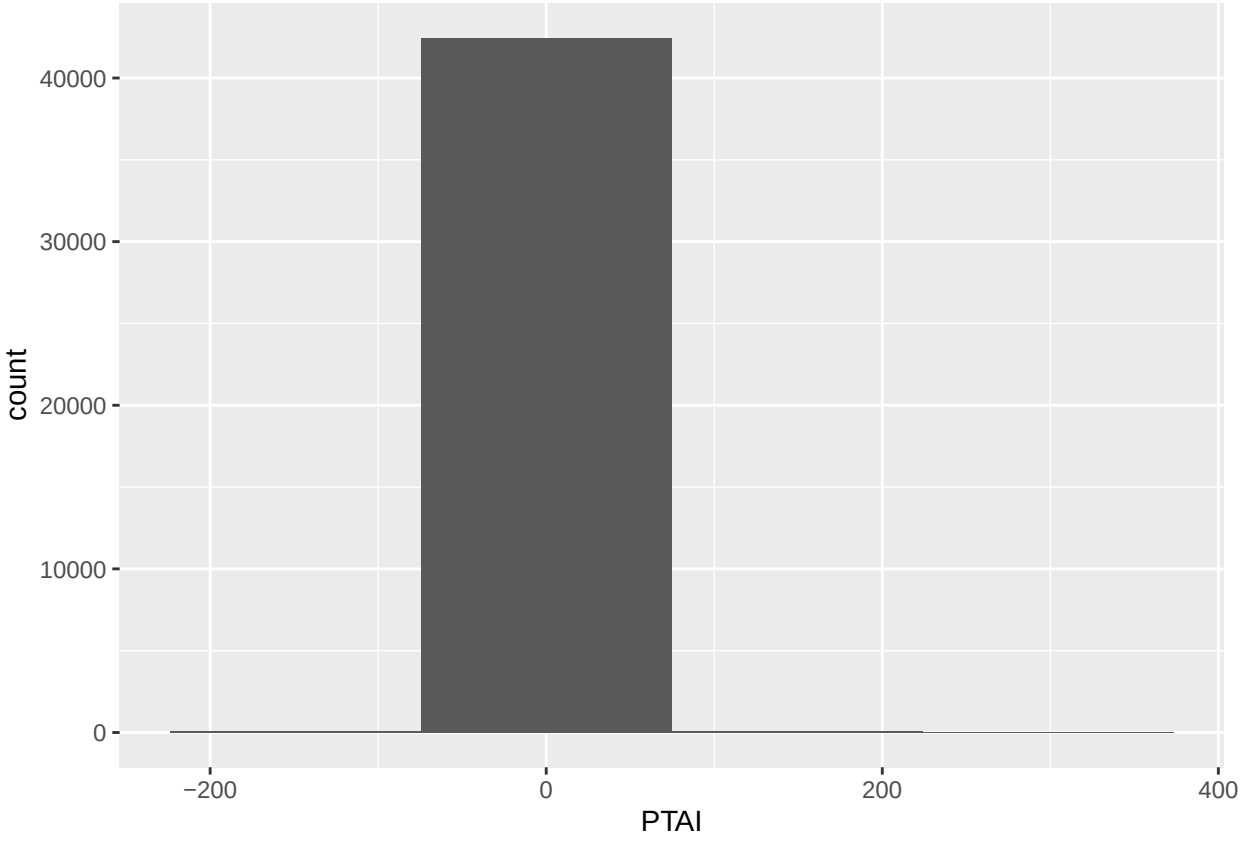
LOR (Length of Relationship):

Calculated by subtracting registration date (RegDate) from the latest date in the dataset October, 2nd 2005. The unit of this measure is days since registration date. As the histogram below displays, the highest count of in the distribution is found in the interval from 220 days to 225 days.



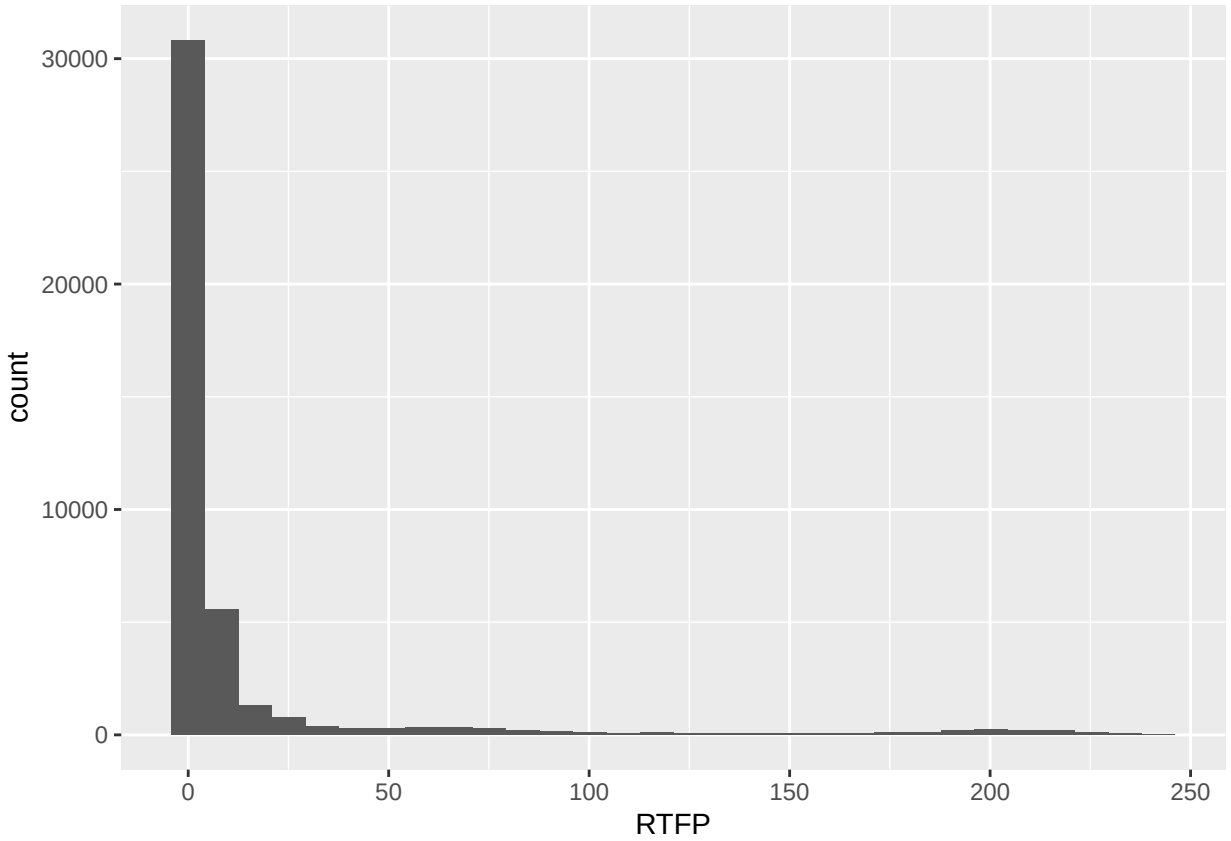
PTAI (Pay to Active Interval):

This time interval is found by subtracting the time from the first deposit of funds for betting to the date where the gambler is first actively participating in play. The unit of this measure is days. The histogram below shows that the dataset's PTAI is uniformly distributed with very little variance. However, there are a few outliers in both the positive and negative directions. This means in a few cases gamblers were actively playing prior to their first date due to promotional campaigns.



RTFP (Registration to First Payment):

This metric is a time interval from a gambler registration date to the first deposit of funds for betting. It is calculated by subtracting the registration date from the first pay date. It's unit of measure is days.



The distribution of RTFP above demonstrates that the vast majority of gamblers made their first deposit of gambling funds the same day they were registered. However, some outliers on the far right of the horizontal axis indicate anomalous behavior paying as much as 242 days after registering. There are no gamblers with a negative RTFP, as all gamblers with a first pay date before registration were filtered from the dataset.

Maximum RTFP:

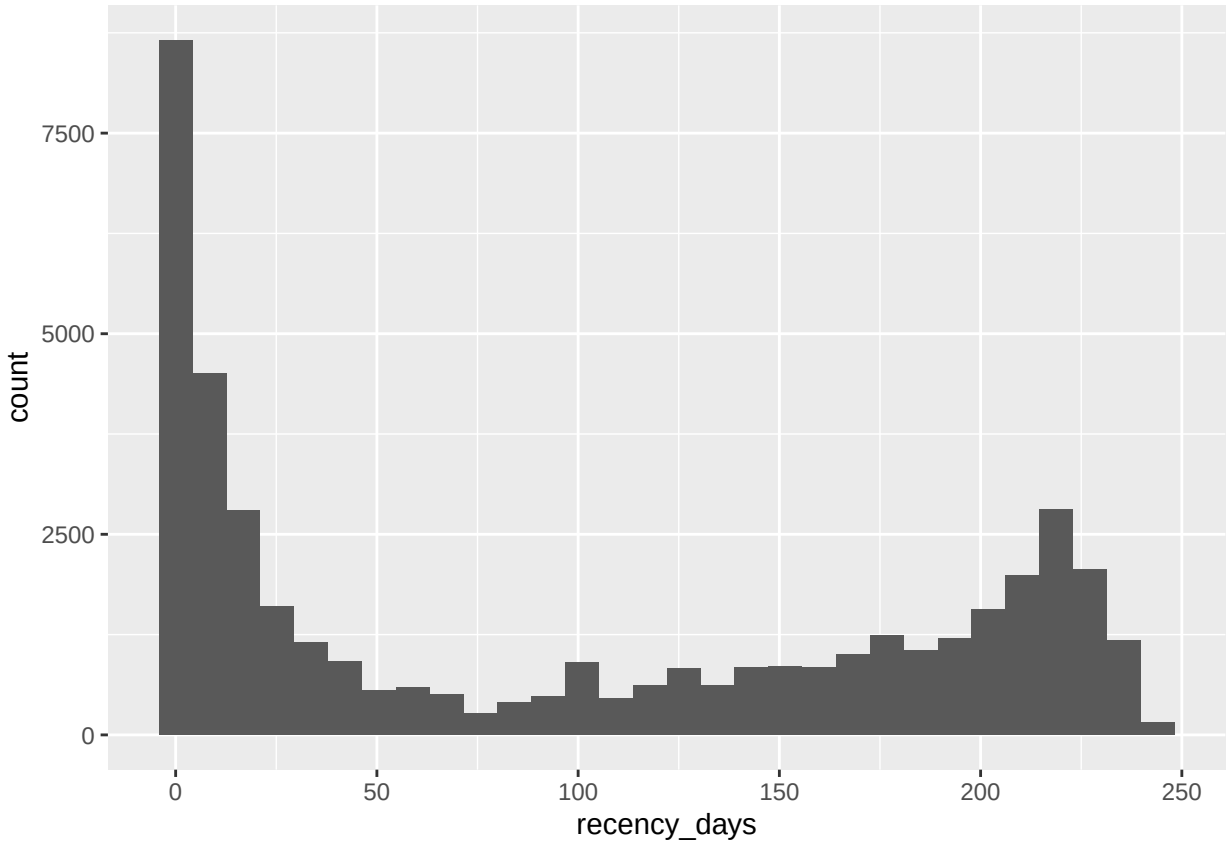
```
FALSE Time difference of 242 days
```

Minimum RTFP:

```
FALSE Time difference of 0 days
```

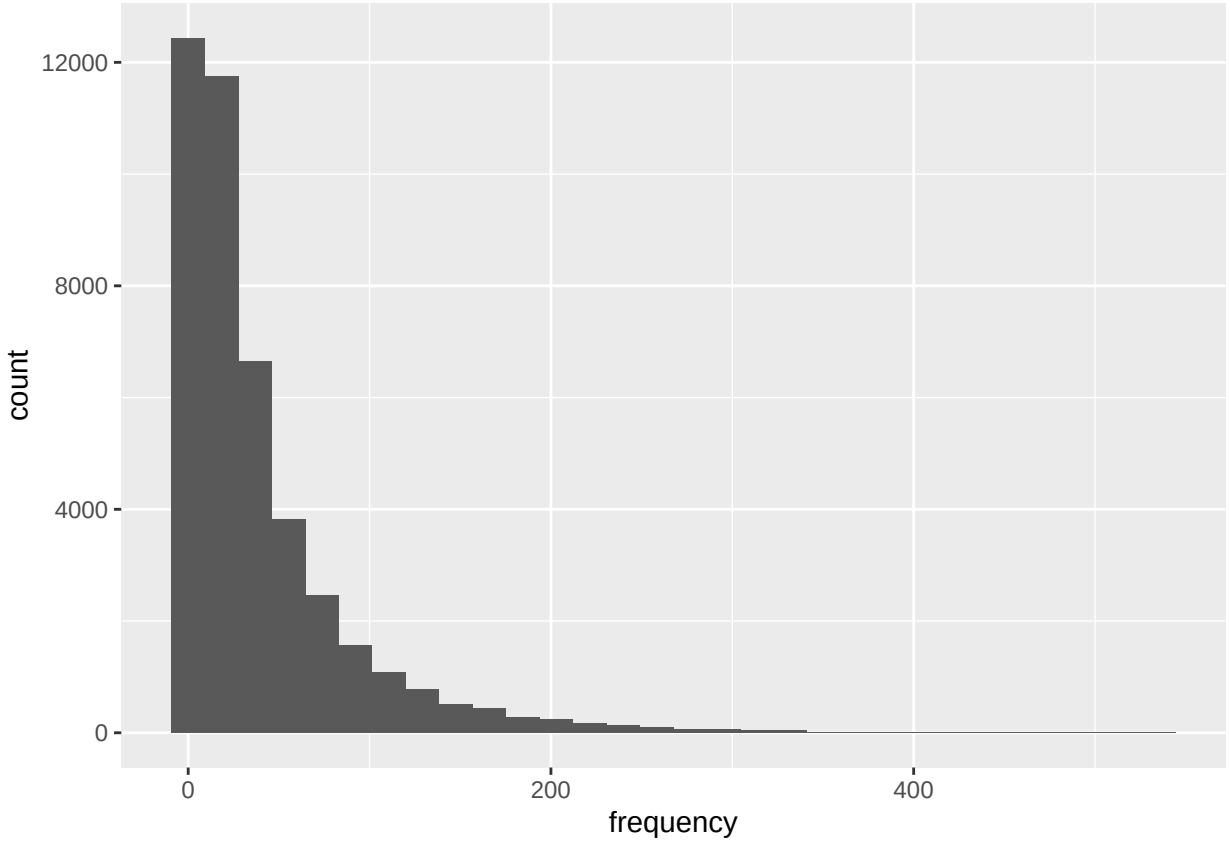
Recency:

The `recency_days` variable indicates the number of days elapsed ranging from the gambler's most recent transaction date to the last date featured in the dataset; October 10th, 2005.



Frequency:

The metric frequency represents the number of dates which a gambler in the dataset made a transaction of any form. It's visible in the histogram figure below that the majority gamblers had a frequency somewhere between 0-100 days or transactions during the aforementioned period.



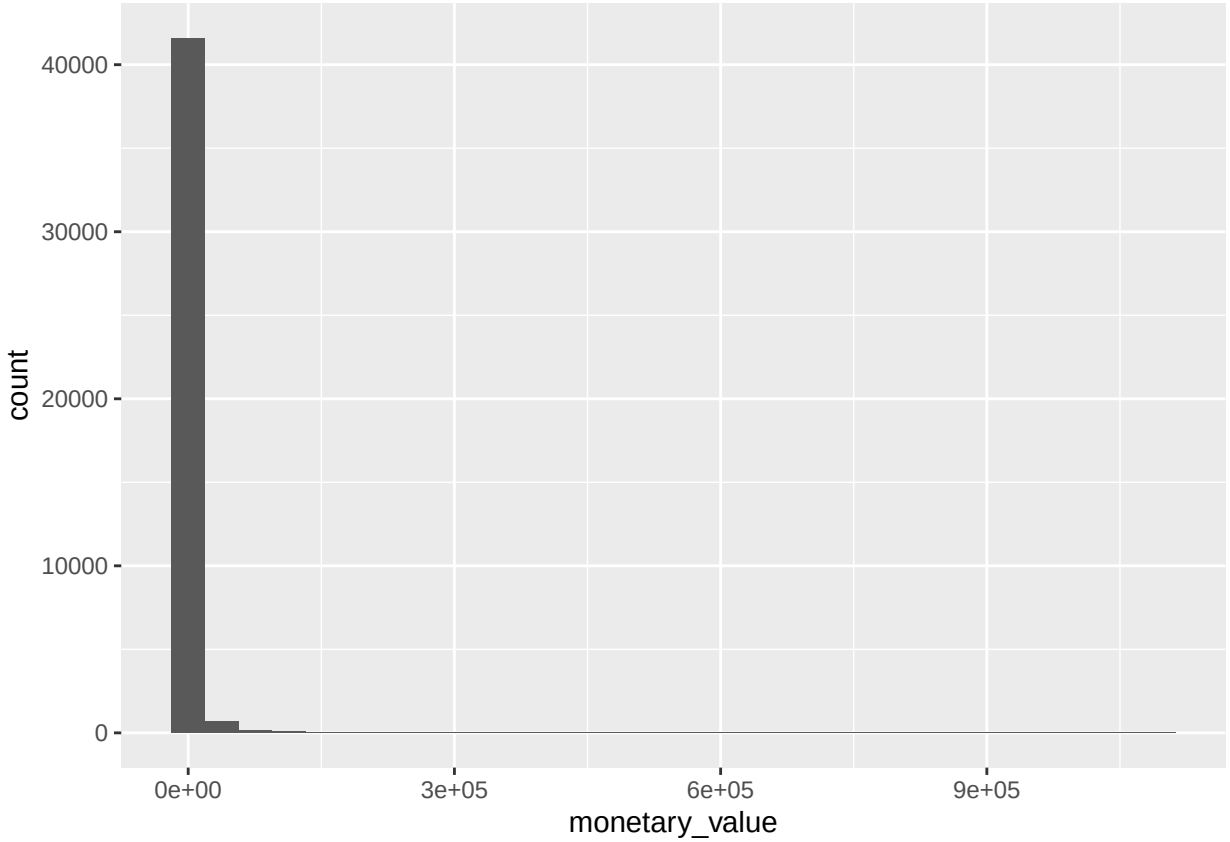
The median frequency displays a point of central tendency for gambler frequency without being distorted by extreme outliers on the far right of the distribution.

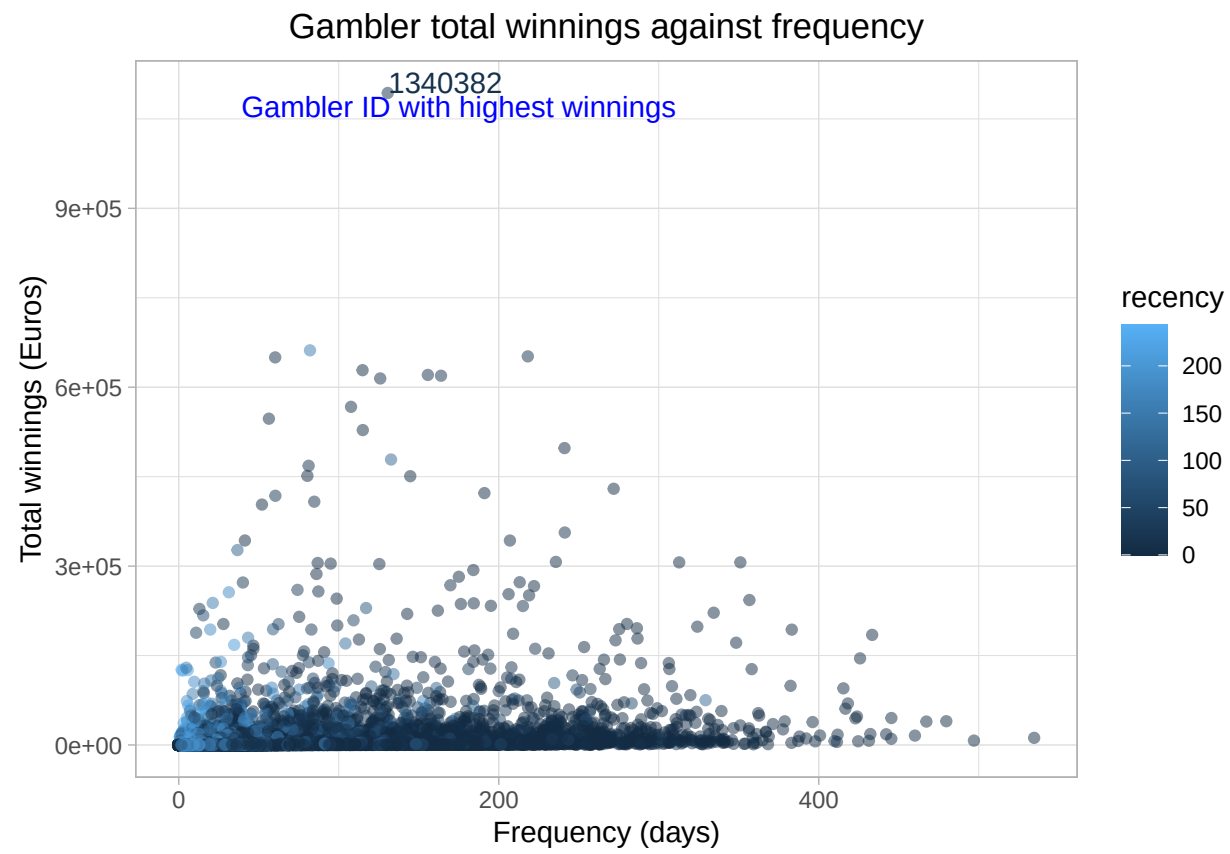
Median Gambler Frequency:

```
FALSE [1] 22
```

Monetary Value:

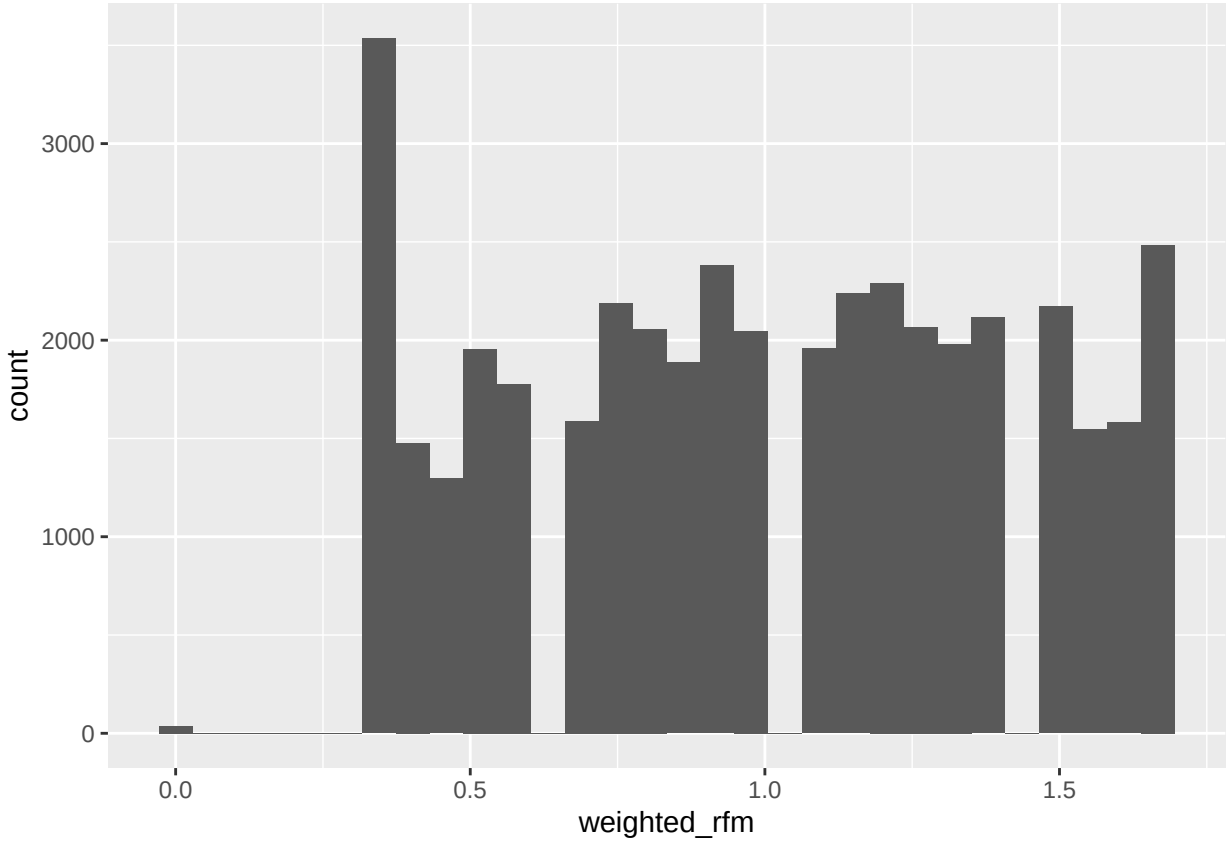
The metric `monetary_value` represents the total amount of money won by each individual gambler for the relevant time period of February 2nd, 2005 to February 27th, 2005. The second figure below shows the relationship between frequency, recency, and monetary value as well as an annotation displaying the gambler ID number with the highest total winnings or monetary value.


40000
30000
20000
10000
0
count
0e+00
3e+05
6e+05
9e+05
monetary_value

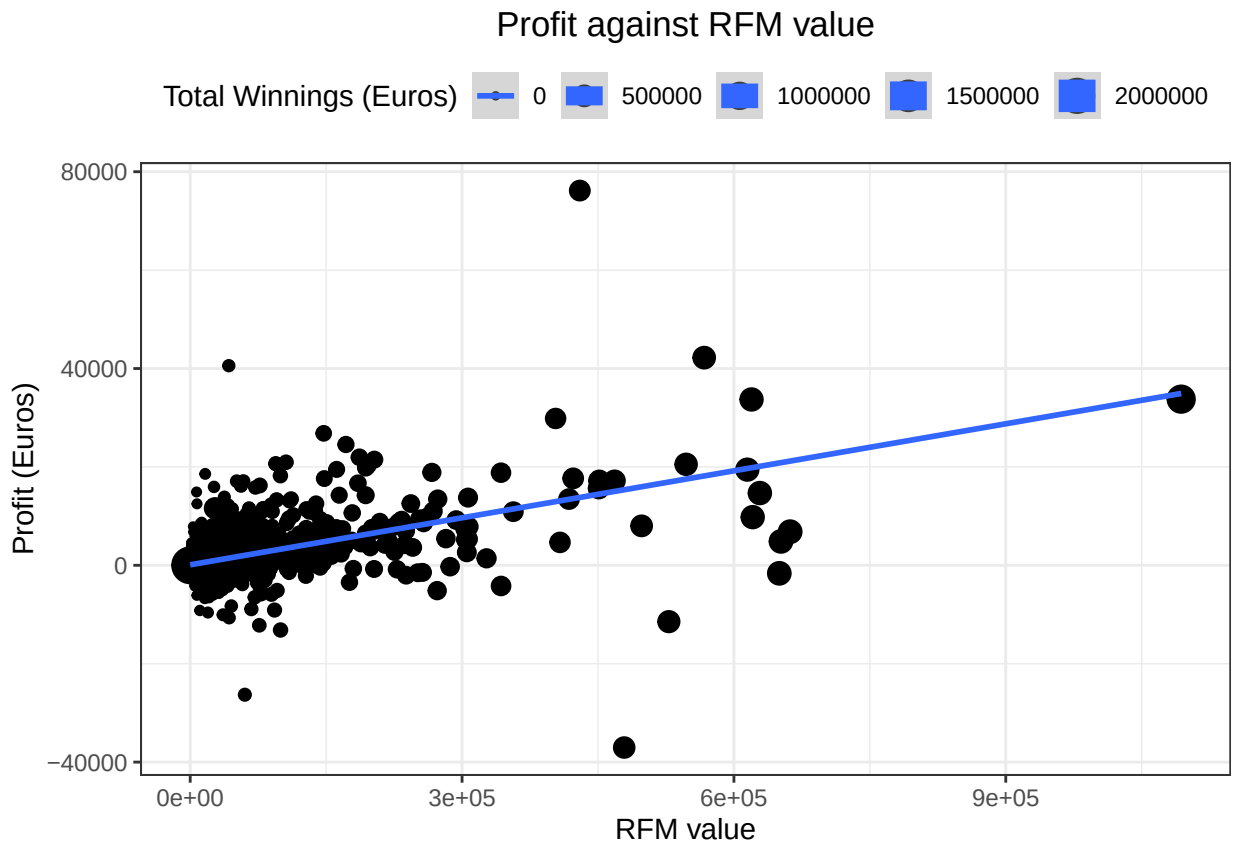

Gambler total winnings against frequency
1340382
Gambler ID with highest winnings
9e+05
6e+05
3e+05
0e+00
Total winnings (Euros)
0
200
400
Frequency (days)
recency
200
150
100
50
0

Weigthed RFM (Recency Frequency and Monetary Value) Score:

The Weighted RFM metric encapsulates information from the three previous metrics which are weighted and represented in one coefficient, where 1.6 represents the first percentile of gamblers for the metric, i.e. those with the highest score or summation of the metrics (`recency_score`,`frequency_score`, and `monetary_value_score`), which is itself represented in its full numerical value as `rfm_value`. The three metrics are then transformed into numeric values respectively, with the same ranking system as weighted RFM.`weighted_rfm` is an aggregate weighted value of these three intermediary metrics to give a comprehensive measure of each gambler's RFM value.



The following plot displays the relationship of RFM value (unweighted) to profit (for Bwin).



It is visible that there's actually a somewhat weak correlation between profit for the brand and RFM value of the gambler. Additionally, the gamblers with a mid-range RFM value are the most profitable gamblers for Bwin. The pearson correlation coefficient can be found below.

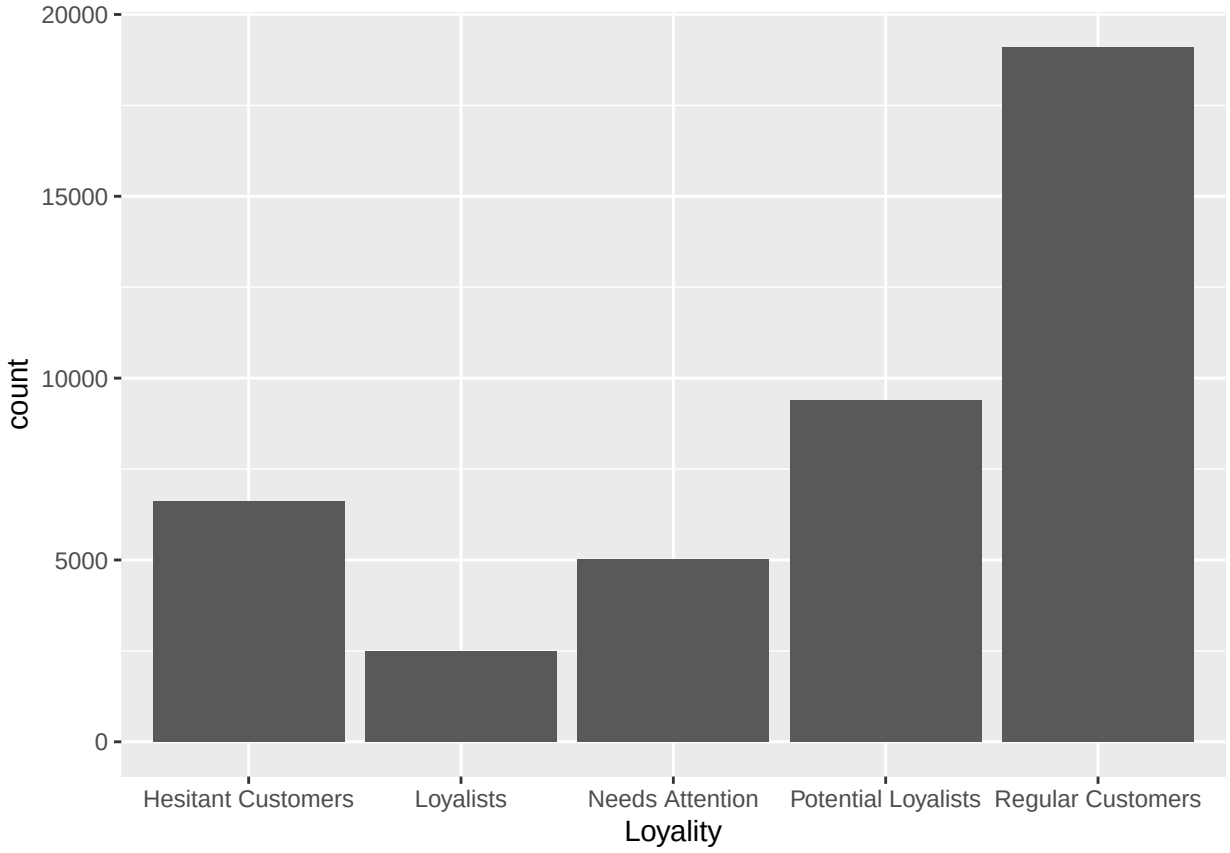
```
FALSE [1] 0.5367227
```

Loyality:

The loyality metric categorizes each gambler in the dataset in one of the following bins: Regular Customers, Hesitant Customers, Needs Attention, Potential Loyalists, and Loyalists. The tier in which a gambler falls within is dictated by their weighted RFM scored classified as follows:

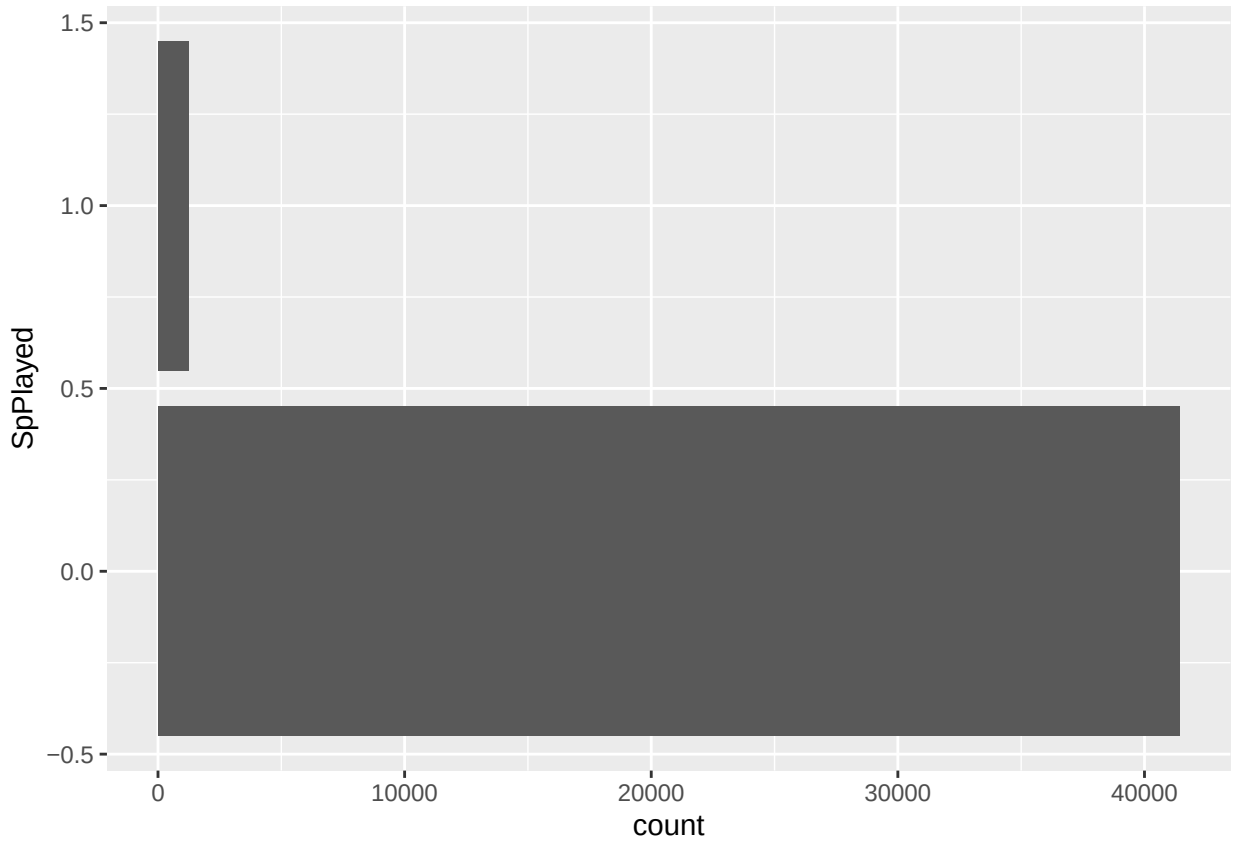
- Weighted RFM score greater than 1.6: Loyalists
- Weighted RFM between 1.6 and 1.3: Potential Loyalists
- Weighted RFM between 1.3 and 0.7: Regular Customers
- Weighted RFM between 0.7 and 0.4: Hesistant customers
- Finally, a weighted RFM below 0.4: Needs Attention

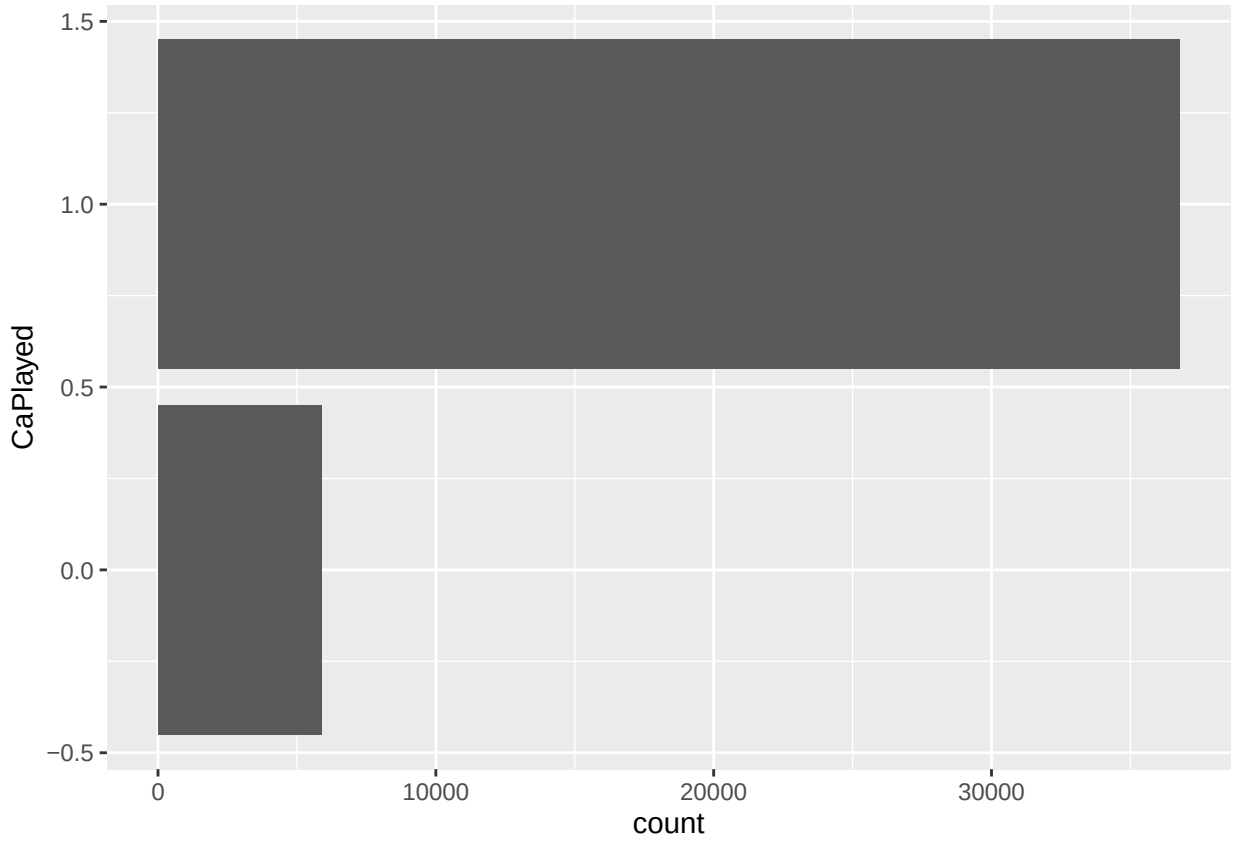
Below is a bar chart of the frequency count of gambler loyality segments:

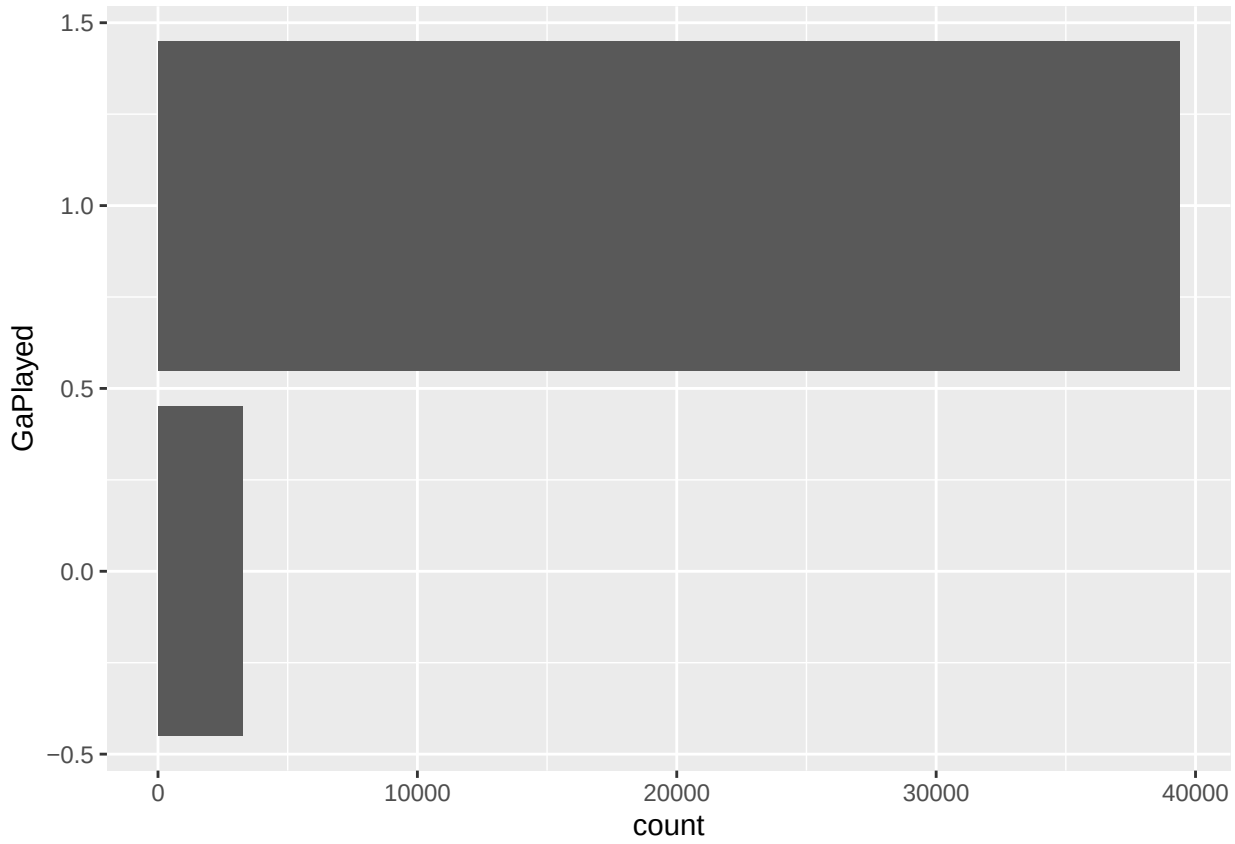


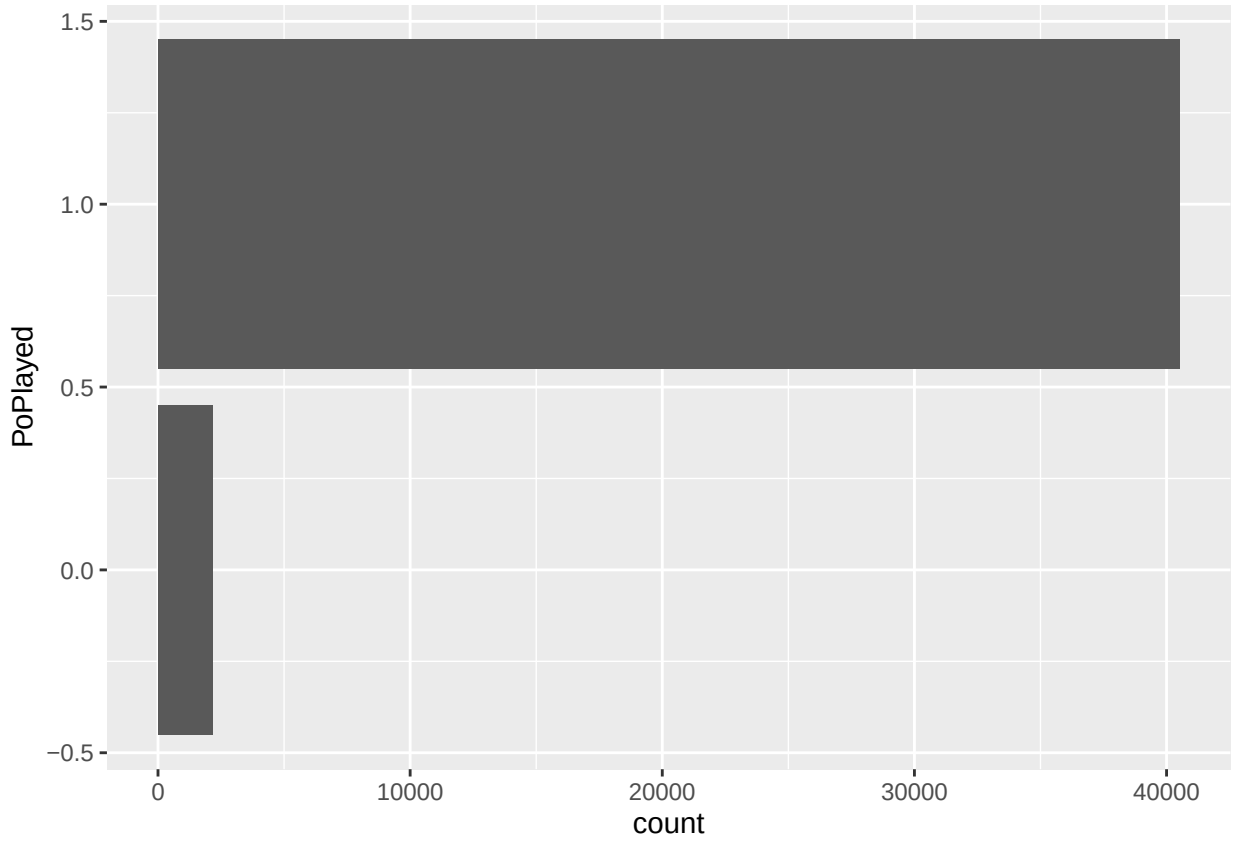
Gambler Betting Behavior

The group of metrics with the suffix 'Played'(`SpPlayed`,`CaPlayed`,`GaPlayed`,and `PoPlayed`) are binary variables which represent if a gambler has participated in a betting category (Sports Book, Casino, Games, or Poker), with 1 representing that there is a first play date for the category and 0 representing that there is no play date for the category during the relevant time period.


1.5
1.0
0.5
0.0
−0.5
SpPlayed
0
10000
20000
30000
40000
count


1.5
1.0
0.5
0.0
−0.5
CaPlayed
0
10000
20000
30000
count


1.5
1.0
0.5
0.0
−0.5
GaPlayed
0
10000
20000
30000
40000
count


1.5
1.0
0.5
0.0
−0.5
PoPlayed
0
10000
20000
30000
40000
count

Final Takeaways:

The most relevant insights from a business perspective that our analysis brought to the forefront are as follow:

- We have a small population of loyalists with a mid-range betting frequency, who seldomly win big, but bet a lot of money, making them our most profitable gamblers.
- We have large population of passive or casual betters with low-frequency betting, whom are winning at a larger rate substantially than our loyalists.
- RFM is weakly correlated with profit for Bwin, and gamblers with a mid-range RFM value are our most profitable segment of gamblers.
- Finally, we know that the gamblers are segmented into five different segments (Need attention, Hesitant Customers, Regular Customers, Potential Loyalists, and Loyalists). These clusters have been validated using a K-means clustering algorithm in addition to our RFM scoring system. Furthermore, we know that the quantity of gamblers in each of these bins which will be useful when targeting them in marketing campaigns.
- To explore the dataset and view other relevant plots via a Shiny R app, click here