An Analysis of Sports Betting Behaviour in Singapore

Prepared for Professor Kam Tin Seong

Group 20 – SMU Analytical Gambling Unit

Ng Ngee Heng, Eugene

Pham Minh Khoa

Wilson Wong

TABLE OF CONTENTS

ABSTRACT

In today's interconnected world, the gambling environment has transformed into a multifaceted playing field without boundaries, exposing more people and younger people to the games, and too creates loop holes for illegal gambling operators to enter the market. The result is greater public worry about the social ills of irresponsible gambling.

Our sponsor, Singapore Pools, takes a strong stand in responsible gaming, wanting to offer a safer outlet for the public to play. This paper will explore existing gambling transaction data (n=930,000) to identify and better understand betting patterns that would eventually allow us to flag out players who engage in or is susceptible to irresponsible gambling; in turn suggesting ways to promote responsible gambling. This paper would also consult past literatures to guide our methodological approach and cross compare hypotheses and findings.

The methodological flow of this project begins with exploratory data analysis where the dataset would cleaned and transformed for further modelling. The large set of transaction will be aggregated into a list of user data. We then proceeded with relationship analysis of the parameters and bet preferences of players of different demographics, testing for collinearity between parameters. Using a clustering analysis, we will profiled players into four main segments: (1) Masses (2) High-rollers (3) Players-at-risk (4) Habituals

This unique segmentation would allow our sponsor to identify players who are at risk of irresponsible gambling, and suggest strategies to reach out to these segments and alert them of their betting behaviour whilst educating them about responsible betting. To ensure project continuity and future analyses, our team has created a dynamic dashboard to visualise monthly transaction trends, highlight popular events, flag out players who are at risk, and allow exploration of each individual player's profile and betting patterns (i.e. their betting intensity, transaction history).

INTRODUCTION

Gambling is often seen as a problem in society, no doubt gambling addiction poses a grave societal problem, however banning gambling is not a viable solution, for it would simply drive these activities underground. Our sponsor, Singapore Pools was set up by the Singapore government in 1968 to place gambling on legal grounds and to deal with the social ills tied to gambling. Ever since, Singapore Pools has been the sole legalized operator to run lotteries and sports betting in Singapore.

Unlike in most countries where gambling houses are privately owned organizations, Singapore Pools is a stated-owned organization, registered under Singapore's Ministry of Finance. Singapore Pools offers four main products to the public (TOTO, Singapore Sweep, 4D, Sports Betting) all of which –operations and product configurations – are regulated by Singapore's Ministry of Home Affairs, Ministry of Finance, Ministry of Social & Family Development.

Our sponsor takes a strong stand in responsible gaming, by offering a safer outlet where players can bet responsibly within their financial means. Attrition rates have be raising over the years, and this could meant that Singapore Pools' customers are seeking other avenues to participate in gambling activities such as illegal online-gambling sites, which may lead to irresponsible betting. Therefore, within the next few years, our sponsor seeks to undertake a data-driven approach to promote responsible gambling by monitoring the player's' betting behaviour and performance, in hopes of highlighting alarming patterns that could indicate signs of irresponsible gambling, and too use this data to help usher in their online betting platform that is scheduled to launch in the upcoming year.

Thankfully, our sponsor has been collecting user and transaction data for the past several years, but has yet put it to good use. Singapore Pools too set up a customer insights division about a year ago to better understand their customers through the analysis of these user data. This is their first step towards a data-driven approach to promote responsible gambling and to understand the gambling behavioural patterns of their customers; and thus this is where our team comes in.

PROJECT OBJECTIVES

The aim of this project is to provide Singapore Pools with a better understanding of the gambling behaviours of their customers through the identification of betting preferences and patterns. Clusters of players may be identified base on their betting behaviour – ways of splitting their bets, preference for a league, different decision making process, and ways of selecting their bet selections. Such behavioural patterns could possibly be linked back to certain demographics pertaining to the cluster, allowing us to further infer reasons behind their gambling habits, and hopefully could help us identify those irresponsible gamblers too.

The scope of our project is limited to customers belonging to the Sports Betting segment of Singapore Pools who have opened betting accounts with Singapore Pools. The data provide are confined to line betting transactions made by their 'Gold' and 'Platinum' members.

The overall objectives of our project are as stated:

(1) Provide insights with regards to gambling behavioural patterns

(2) Profile their existing pool of customers into meaningful segments

(3) Build a dashboard to visualize betting patterns and trends on a macro and individual level

And based on the characteristics of each cluster, the sponsor's end objective is to (1) flag out players who display alarming patterns that could lead to irresponsible betting, and (2) on their end to tailor business actions that targets the derived clusters of players to enhance their gambling experience while ensuring that they make bets in a responsible fashion.

LITERATURE REVIEW

Gambling is one topic that is widely researched across the world, from survey polls of gambling participation and perception, gambling risk and pathology, to thorough statistical analysis on gambling behaviours.

According to a survey done by Singapore's Ministry of Community Development, Youth and Sports (MCYS), within a year's period, 58% of Singaporeans over 18 years of age have participated in at least one gambling activity. Further study on pathological gamblers by the MCYS found that players at risk to developing a gambling addiction would gamble at least once a week, and this pool of susceptible players made up 70% of the sample population involved in the study (2005).

Behavioural or betting patterns is another popular area studied across most papers – for they provide cues to possible pathological gambling behaviours; difference between betting behaviour of regular players and players at risk (problem gamblers) is evident, a common finding in most studies. One study revealed that gamblers at risk are more likely to bet more frequently coupled with increasing bet amounts, regardless of their bet outcome (Mizerski, 2011). And that less frequent players are more likely to put more effort into decision-making when making bets to allow for future betting possibilities, as compared to regular or frequent players. Evidently they also found that certain betting games and game arrangements may actually prompt reckless betting that could like to irresponsible gambling.

Several other papers provided insights to a more analytical approach to segment gamblers and identify those at risk. A study by Faregh and Leth-Steensen (2011) discovered clusters of players with variations in terms of their bet activity level (frequency), bet variability (spread of stakes and odds), time spent on making the bets, and the games played. Relationship and predictive analysis between selected parameters may reveal variables that best predicted returns, and reflect bet strategies that are less sophisticated (Gainsbury & Russell, 2013). Suggestions on data collection procedures, selection of metrics and parameters for clustering players in these papers are just some of the secondary insights that have aided our choice of methodology and analysis – determining ways of profiling our result clusters and creation of new behavioural parameters – that will be elaborated on later in this report.

Besides researching the field of gambling and the analytical methodologies, we too examined past data visualization papers to learn about the pit falls and best practices of data visualization. "Different types of graphs are designed to communicate different types of messages" quote data visualization expert, Stephen Few, as he demonstrated in his papers regarding the effective use of points and lines to shape data trends, to the principles of colour selection for data visualization – use of contrasting or analogous colours for varying purposes (Few, Stone, 2004; 2006; 2007). Meanwhile some graphs are best to avoid, such as alluring 3-D graphs or pie-charts which can be rendered better in a two-dimensional plane, for the added depth and angle makes interpretation more difficult (Few, 2005). Returning to the dashboard, two guiding principle in designing the dashboard layout that we took from Few's recommendations was to (1) find balance between being information rich and not oversimplifying and (2) to remove clutter or any distractions that do not add value (Few, 2005).

Leveraging on these prior knowledge, our team hopes to deliver actionable insights with regards to the betting behaviour of our sponsor's pool of customers, and present the findings on a dashboard that is all visually appealing, intuitive, practical, and accurately data driven.

DATA CLEANING & TRANSFORMATION

The original sample data set that was given to us was a comma separated values (CSV) file containing over 930,000 unique observations (transactions). The time period spans from January to March 2015, which coincides with the peak period of the different international soccer leagues, so as to ensure the highest volume of betting activity for analysis.

The first phase of our data cleaning process would be to remove the noise, irrelevant fills and outliers in our data set – identified from our early data exploration.

Irrelevant Fills		Outliers	
F1 race bets	**350 observations**	**Stake Amount' greater than $8000**	**379** observations
Event Type: Championship Winner	**591 observations**	**'Odds' greater than 135**	**4889** observations
Blank fills	**6355 observations**	**'Time before Match' > than 6700 mins**	**2640** observations
Foreign players	**85 observations**		

There were two types of noise in our data set: (1) **irrelevant fills** – some of which are variables or observations that do not pertain to the soccer betting products, rejected bets which have no odds indicated, and 'Championship Winner' bet types for which is considered atypical to regular bet patterns; (2) **outliers** – these observations lie beyond the 99th percentile of selected parameters, the thresholds were identify based on scatterplots of the data distribution on Tableau.

The dilemma we faced then was that if we removed the outlier transactions, we would be artificially changing the users' bet behavior and preference when we aggregate these transaction data into one user's overall bet pattern. The other option was to aggregate all transactions (including extreme observations) into a user's overall bets, and then to filter the users from the population. Given that the outlier transactions would impede of analysis of transaction data, we had no choice but to remove these outliers. And in effect, we had to remove the affected users (users who made those extreme transactions) to maintain the integrity of the user data. Therefore, after data cleaning, we are left with 529,678 unique observations (transactions).

To gather deeper insights in later data exploration analyses, our team created new metrics that reflect certain attributes of bet behaviours. This then allowed us to test some of the hypotheses about betting patterns that our sponsor highlighted to us. The new metrics would be used to test difference in betting preference between gender, age, account types, players of different risk profiles. If the metrics do not differ between the segments of players, or does not significantly affect one's bet placement – as we were to discover during the data exploration phase – they would then be removed; retaining only the useful newly created parameters.

TRANSACTION METRICS	**USER METRICS**
• Profit / Loss • Bet Risk Type (High/Medium/Low Odds) • Bet Time Before Match • Bet Day • Bet Period (Categorical Data)	• Probability of **Bet Day** • Probability of **Bet Period** • (Mean/Median/SD/Z-score) **Stake Amount** • (Mean/Median/SD/Z-score) **Odds** • Probability of Each **Odds Type** (High/Medium/Low) • (Mean/Median/SD/Z-score) **Bet Time Before Match** • (Total/Mean/Median/SD/Z-score) **Returns** • (Total/Mean/Median/SD/Z-score) **Profits / Loss** • Number of Wins & Loss • **Win / Loss Ratio** • **Bet Intensity** • (Count / Probability of) League, Event Type, Market Name

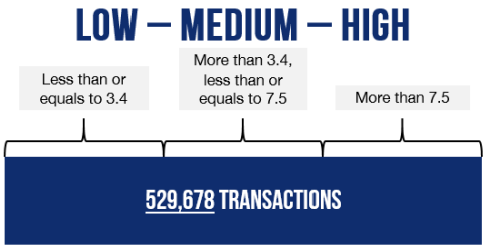
After the data cleaning and data transformation procedures, we moved on to consolidate all the transactions in the "TransactionList" to form list of unique users in another worksheet also known as "UserList". Observations in the transaction data were aggregated into meaningful players' parameters such as profit/loss, returns, transaction count, and number of bets of each league or market type just to name a few. We also calculated the players' individual mean, median and standard deviation for variables such as odds, stake amounts and returns – such statistics allows us to make further inference on the bet patterns, for example the standard deviation or spread of stake amounts offers insights to whether the player is a stable bettor who places consistent amounts for each bets or is an erratic strategist who alters his or her stake based on the games. The final data consolidation resulted in a total of 5,562 unique players (user accounts) for further analysis.

[Note: greater details of the fields of the created parameters can be found in the analytical data cube]

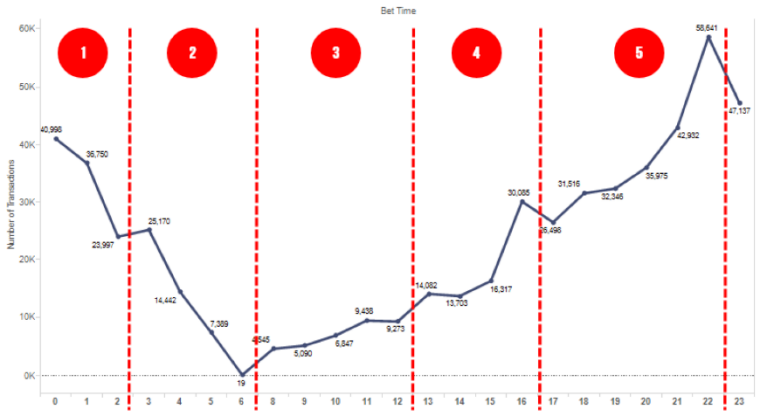
EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) was a frequent procedure that we deployed through our entire project methodology. First used to identify outliers and irrelevant fills, and then to identify relationships and patterns in the existing dataset, which resulted in the creation of new parameters (as mentioned about) that allowed further data exploration and other analytical techniques. Our team also revisited our EDA procedure in later phases of our project to better tune subsequent analytical techniques and analysis, as to be mention in the later sections of this report.

[Note: *full details of all EDA findings may be found in our project's interim report, the following concise findings will be focus on EDA as a procedure to aid creation of new metrics and on additional collinearity data analysis that was not previously mentioned.*]

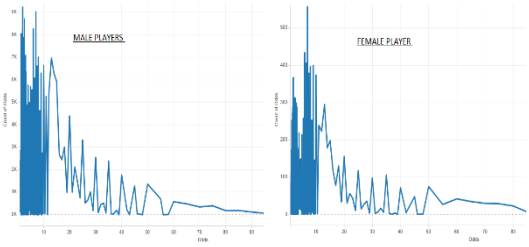


After initial data exploration, our team realized that due to massive amalgamation of transaction observations of the many different users, there were a lot of noise in the data; initial data exploration showed plenty of insignificant relationships. Therefore, we created categorical metrics to segment transactions and users into smaller subsets to reduce variation within these groups, thus allowing more significant observations. For data regarding transactions, the segmentation is based on the bet odds of individual transactions, whereas for data regarding users, the segmentation is based on the total number of betting transactions for the individual player. Classes were determined using the lower and upper quartiles of the distribution of those parameters. We also reaffirmed these classifications with our sponsor in order to verify the fittingness of the proxies.



Other categorical variables that were created included the bet time periods. By plotting the total number of transactions for each hour in a day, we can identify moments of intense betting and absence. Using these high and low peaks, we have drawn up five different segments across the 24 hours in a day as shown on the left.

One interesting demographic related bet pattern would be the relationship of odds and stake amount for gender. Looking at the distribution of odds across the two gender, we see that female players have a higher spike in counts of odds for higher odds (between the 6 to 10 range) relative to their size, compared to male players where their spike for lower odds seemed to be higher than their demand for higher odds. This pattern both agrees with our sponsor's hypothesize that male players would place larger bets due to their possibility greater purchasing power, but it too disagrees with their other hypothesis that male players have preference to make risker bets or bets with higher odds, while female players are more risk adverse.

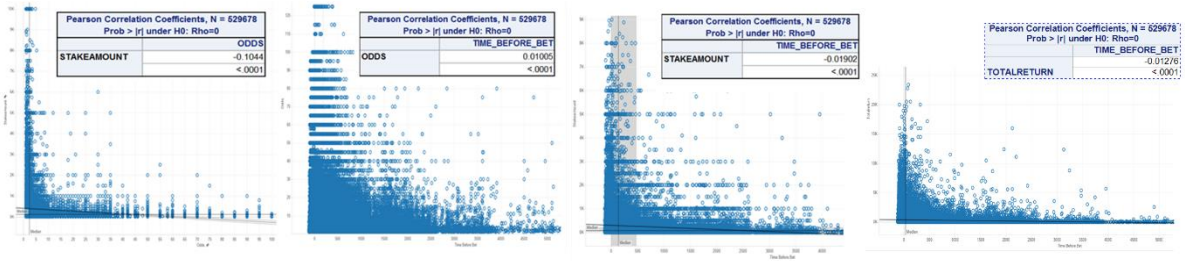


	ODDS & STAKE	ODDS & RETURNS	STAKE & RETURNS	TIME & RETURNS	TIME & STAKE	TIME & ODDS
A-League	-0.12222	-0.08353	0.57547	NOT SIG.	0.01422	-0.03574
African Nations Cup	-0.10868	-0.07409	0.52697	0.02051	NOT SIG.	-0.02833
Asian Champ	-0.10825	-0.07499	0.74874	0.0643	0.08656	-0.03779
Asian Cup	-0.10829	-0.07978	0.68992	0.02786	0.02032	-0.04037
Asian FC Cup	-0.08525	-0.10426	0.53362	0.12668	NOT SIG.	NOT SIG.
Dutch Cup	-0.12226	-0.08802	0.77098	NOT SIG.	-0.06141	NOT SIG.
Dutch League	-0.13514	-0.07458	0.7556	NOT SIG.	NOT SIG.	-0.06258
English Cup	-0.12752	-0.07956	0.61384	0.01511	0.01522	-0.01775
English League Champ	-0.10274	-0.05677	0.73577	NOT SIG.	0.04589	-0.04174
English League Cup	-0.12943	-0.07691	0.50384	NOT SIG.	NOT SIG.	-0.0601
English Premier	-0.11217	-0.07554	0.64171	NOT SIG.	NOT SIG.	-0.03469
Football Singapore	-0.14253	NOT SIG.	0.94974	0.13898	NOT SIG.	NOT SIG.
French Cup	-0.12518	-0.08888	0.81266	NOT SIG.	NOT SIG.	NOT SIG.
French League	-0.10597	-0.07635	0.44826	0.03391	0.0313	-0.05434
French League Cup	-0.14443	NOT SIG.	0.68654	NOT SIG.	NOT SIG.	NOT SIG.
German Cup	-0.13034	-0.10874	0.81745	0.0701	0.1741	NOT SIG.
German League	-0.13173	-0.09887	0.62257	0.02768	0.0465	-0.0576
Italian Cup	-0.12835	-0.08216	0.63248	NOT SIG.	NOT SIG.	-0.03301
Italian League	-0.13415	-0.08837	0.59065	NOT SIG.	NOT SIG.	-0.02533
J League	-0.10766	-0.10493	0.38103	NOT SIG.	NOT SIG.	-0.09127
M FA Cup	-0.10102	-0.07079	0.82028	0.10122	0.13437	NOT SIG.
M League	-0.12331	-0.07773	0.46336	0.12988	0.08943	NOT SIG.
S League	-0.10635	-0.07916	0.83905	NOT SIG.	NOT SIG.	NOT SIG.
Spanish Cup	-0.13509	-0.10376	0.67648	NOT SIG.	NOT SIG.	NOT SIG.
Spanish League	-0.11543	-0.06472	0.70365	NOT SIG.	NOT SIG.	-0.03209
UE Champions	-0.1083	-0.07256	0.63484	NOT SIG.	NOT SIG.	NOT SIG.
UE Europe	-0.09236	-0.07404	0.68789	NOT SIG.	NOT SIG.	NOT SIG.
US Soccer League	-0.1095	-0.09523	0.36419	0.08998	NOT SIG.	NOT SIG.

As requested from our sponsors, we also explored the relationship between user betting parameters for each leagues to find any possible influence of the type of leagues on one's betting behaviour.

We conducted a multi-collinearity test, and these were the findings for the leagues observations on the left

Across all leagues, high stake is placed for matches with low odds, and returns are also lower for matches with higher odds. Across half of the leagues, returns were better when longer time was taken before bets were placed. Higher odds also tend to lead to later bets being made (shorter duration between bets placed and actual match), and higher stakes tend to be placed earlier (though this does not apply to the "Dutch League"). Notably, the effects of timing (correlation with the other parameters) appears to be not significant across most leagues. Meanwhile, although the relationships between odds, stake amount and returns across the leagues were significant, the magnitude or strength of the relationship is fairly small with the regression line almost being flat / horizontal.



When these relationships between the parameters are test on the aggregated dataset, the same findings were found – with higher odds, stake amounts are less, and bets are placed much earlier; and returns tend to be higher when bets are placed later. Indeed relationship were significant, but they still yielded very low magnitude.

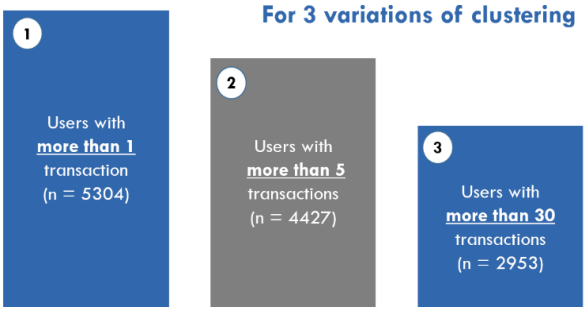
METHODOLOGY FOR CLUSTER ANALYSIS

The two tools that we have used for the clustering analysis phase were SAS Enterprise Guide and JMP Pro 12. SAS Enterprise Guide's process map offers a great way to organize and track our procedures from early data exploration tasks to the subsequent multivariate analysis; it too allows quick modification of the tasks of each modules, making duplication of analysis of slight input variations a lot easier. The main drawback of the software would be the lack of (or poor) data visualization of the analysis results, therefore we had complemented this using JMP Pro 12 to provide the necessary data plots and charts.

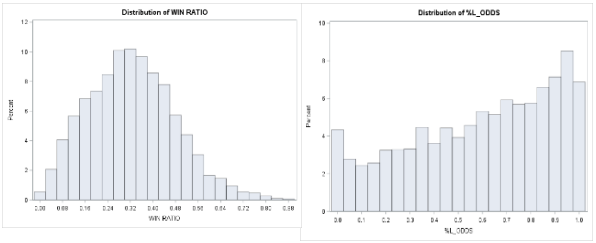
Our data preparation for clustering analysis involved standardizing and normalizing the input parameters. Two methods that were used for the standardization process: (1) **Z-scores** of the variables (2) And transforming the parameters into a **probability** of it being selected as an event.

To identify the need to normalize the data variables, we plotted out the distribution on JMP to check for the presence of skewness in the spread. Instead of carrying the commonly used statistical methods – such as using Windsorizing the variables or using taking the log of the variables – to transform input variables to remove the skewness, we decided to take a different approach. Normalizing the data does not always improve the results, it may pose problems for clustering algorithms by transforming spherical clusters into elliptical clusters affecting the density zoning. Returning back to our EDA step, we figured we could remove the long tails of our input variable by filtering out certain "volatile" users who actually made up the outliers.

Our aggregated dataset retains users of few transaction records, they may comprise of new customers or one-time-off bettors. As such, they do not have a stable betting pattern, addition to that, those single (or few) bets alone would result in misrepresentation of certain parameters. For example, given a player with one transaction count, a single bet won would mean that his or her winning ratio is at a 100%.



In light of this finding, we conducted 3 variations of clustering analysis. For the first, we filtered out single-transaction users; the second included only users with more than 5 transactions (which was determined by using the 10th quartile of the spread of transaction count); and lastly a clustering analysis for users with more than 30 transactions (a statistics rule of thumb to accommodate the t-test). And indeed, users with more than 30 transactions showed a more stable betting pattern. The spread of parameters followed more closely to a normal distribution, and the hierarchical clustering analysis returned a favourable result for selecting our k-value for the k-means clustering that is to determine our final cluster results.



Selection of input variables for clustering analysis is a highly judicious job for analyst, we risk missing distinctions that ought to be highlighted if we select too few variables, and have too many variables will only blur those distinctions. One innovative way to determine the input variables for clustering was demonstrated by Raftery and Dean (2006), using a model comparison criterion to compare two clusters with varying inputs.

Delving into the workings behind this model, their method is analogous to the stepwise regression analysis. By comparing the predictive strength of the input variable to a guiding distinction or rule, one may determine the influencing factors that would be key to be highlighted in the cluster profile.

Using the scope of this project – which prioritizes the players' profits and loses which ties back to responsible gambling – to guide our clustering, we carried out a stepwise regression analysis using the players' standardized (z-score) profit or loss parameter as the dependent variable for the model; and inputted a bunch of other parameters to test its significance as a predictor variable.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	386.97192	38.69719	44.37	<.0001
Error	2942	2566.02808	0.87221		
Corrected Total	2952	2953.00000			

	Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
	Intercept	-0.28244	0.14458	3.32841	3.82	0.0509
*	WIN RATIO	2.60388	0.22596	115.82143	132.79	<.0001
*	Z_STAKE	0.23297	0.04095	28.22921	32.37	<.0001
*	Z_SD_STAKE	-0.49960	0.04083	130.57530	149.71	<.0001
*	%L_ODDS	-1.08955	0.13899	53.59511	61.45	<.0001
*	%H_ODDS	0.01531	0.14147	0.01021	0.01	0.9138
*	Z_SCORE_BEFORE_MATCH	0.09287	0.01949	19.79962	22.70	<.0001
	BET_TIME_1700_TO_2200_%	0.40062	0.13486	7.69637	8.82	0.0030
	BET_TIME_2300_TO_0200_%	0.28911	0.14395	3.51830	4.03	0.0447
	BET_DAY_FRI_%	-0.65569	0.32269	3.60124	4.13	0.0422
	BET_DAY_SUN_%	-0.48904	0.20051	5.18840	5.95	0.0148

* Forced into the model by the INCLUDE= option

Bounds on condition number: 5.8205, 316.55

All variables left in the model are required or significant at the 0.1500 level.

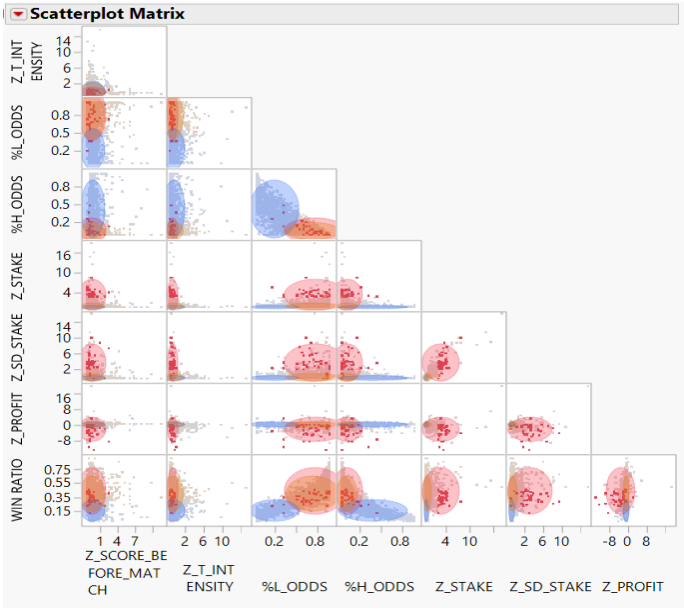
No other variable met the 0.1500 significance level for entry into the model.

On top of that, we had other clustering criterions or caveats that we had to abide (which were sponsor requests). This included: (1) that the input variables had to be actionable by sponsors, and (2) that the number of resultant clusters was to be limited to 5. As such, our final input variables were narrowed down to the following (highlighted in the red box):

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	0.0644	0.02634	2.44	0.0146	0
Z_SCORE_BEFORE_MATCH	1	-0.01	0.01093	-0.91	0.0103	-0.01
Z_T_INTENSITY	1	0.27966	0.01128	24.8	<.0001	0.27966
%L_ODDS	1	-0.23069	0.07406	-3.12	0.0019	-0.06882
Z_STAKE	1	0.5597	0.02596	21.56	<.0001	0.5597
Z_SD_STAKE	1	0.16639	0.02636	6.31	<.0001	0.16639
WIN RATIO	1	0.73687	0.14511	5.08	<.0001	0.11358
Z_PROFIT	1	-0.12401	0.012	-10.33	<.0001	-0.12401

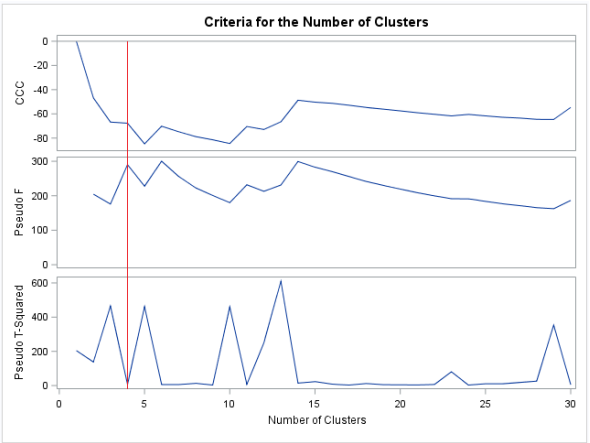
Collinearity is another worry when it comes to clustering; it is better to input variables that do not correlate with one another. Such variables behave in a similar manner, adjoining these correlated variable increases their contribution (a higher weight compared to the other variables) thus, skewing and distorting the results.



With reference to the 95% bivariate normal density ellipse (the coloured ovals) in each scatterplot, we can identify the variables that are correlated – represented by a narrow and diagonally oriented oval. From the scatterplot results, there is not much collinearity between our input variables, as the density ellipse are either rounded or has a vertical or horizontal orientation.

Given the lack of correlation, we thus do not need to carry out a principal component analysis to de-correlate the data before moving on to our cluster analysis.

The next phase was for us to carry out a hierarchical clustering to determine number of cluster to be used for our K-means clustering analysis. To select the K-value we used the elbow criterion, comparing for the percentage of variance given each addition cluster. We have identified several potential peaks (K = 4, 6) points where marginal gain drops, whereby adding another cluster would not improve the modelling.



Given the limit of number of clusters given by our sponsor, our team selected a K-value of 4 for the subsequent K-means clustering.

Input variables for the K-means clustering were the same used for the hierarchical clustering:

- *Bet Time Before Match (Z-score)*
- *Odds Type H-M-L (%)*
- *Stake Amount (Z-score)*
- *Spread of Stake Amount (SD Z-score)*
- *Profits or Losses (Z score)*
- *Win-Loss Ratio (%)*

RESULTS & FINDINGS

The results of our K-means clustering are as follows:

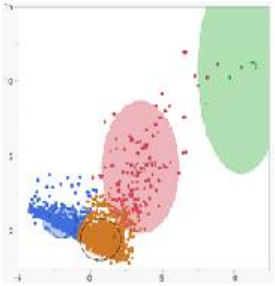
	Cluster Means										Demographics			
Cluster	FREQUENCY	Z_SCORE_BEFORE_MATCH	Z_T_INTENSITY	%L_ODDS	%M_ODDS	%H_ODDS	Z_STAKE	Z_SD_STAKE	Z_PROFIT	WIN RATIO	Gender (M)	Gender (F)	Acct Type (Plat)	Acct Type (Gold)
1	2652	0.01	-0.17	0.59	0.23	0.18	-0.16	-0.17	0.16	0.32	2553	98	747	1905
2	3	-0.10	-0.29	0.88	0.11	0.01	16.38	11.97	8.36	0.52	3	0	3	0
3	143	-0.04	-0.19	0.77	0.14	0.08	2.81	3.11	-2.19	0.41	139	4	73	70
4	155	-0.10	3.12	0.52	0.25	0.23	-0.22	-0.18	-0.84	0.29	147	8	56	99

To profile the 4 clusters, we have applied the profiling by centroids method for a meaningful interpretation of the target segments or clusters. We compared the z-scores between our clusters' means and that of the population mean. The pooled standard deviation (RMS STD) and variables variation within the cluster (Within STD) are within acceptable standards never going beyond 1 standard deviation from the mean – a check for homogeneity within. Meanwhile the RSQ measure is higher, assuring us that the clusters created are different from one another (with the exception of the "Time-Before-Match" parameter).

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
Z_PROFIT	1.00017	0.79914	0.362236	0.567979
WIN RATIO	0.15416	0.15264	0.020616	0.021050
Z_T_INTENSITY	1.00017	0.67861	0.540117	1.174467
Z_SCORE_BEFORE_MATCH	1.00017	1.00031	0.000736	0.000737
%L_ODDS	0.29836	0.29519	0.022128	0.022629
%H_ODDS	0.21782	0.21654	0.012766	0.012932
Z_STAKE	1.00017	0.56682	0.679148	2.116700
Z_SD_STAKE	1.00017	0.59984	0.640674	1.782987
OVER-ALL	0.80327	0.60603	0.431376	0.758631

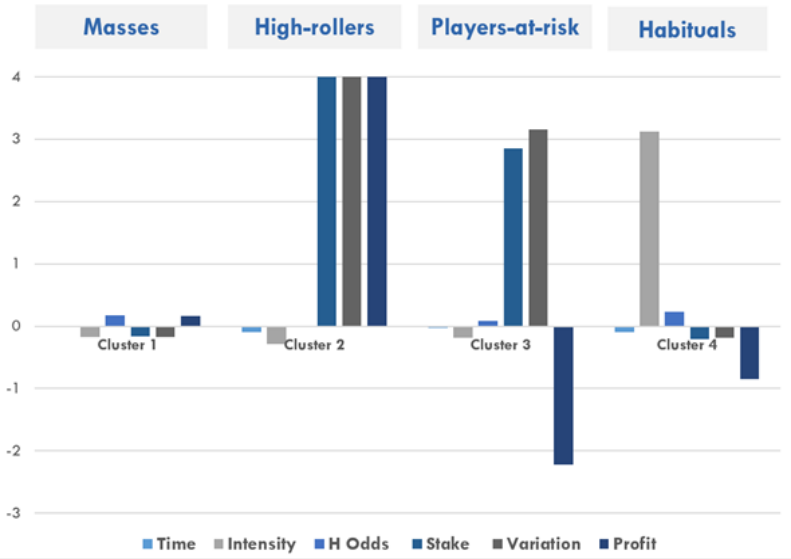
Variable	Mean
Z_SCORE_BEFORE_MATCH	0.00
Z_T_INTENSITY	0.00
%L_ODDS	0.59
%M_ODDS	0.26
%H_ODDS	0.18
Z_STAKE	0.00
Z_SD_STAKE	0.00
Z_PROFIT	0.00
WIN RATIO	0.33

Looking at the Biplots of the clustering analysis, we too can identify parameters which make certain clusters distinct. Although there are no visible overlaps (of the density ellipse) on this two-dimensional plane – which is ideal – our team still explored further using the three-dimensional score plot (3D rotation of the ellipse) to find any overlapping clusters. With confirmation that the result clusters and distinct and significant, we proceeded to profile the clusters.



Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.9141	36.426		36.426
2	1.7524	21.906		58.332
3	1.2006	15.007		73.339
4	0.9764	12.205		85.544
5	0.7028	8.785		94.329
6	0.2533	3.166		97.495
7	0.1114	1.393		98.888
8	0.0889	1.112		100.000



The profile of our four clusters are as follows:

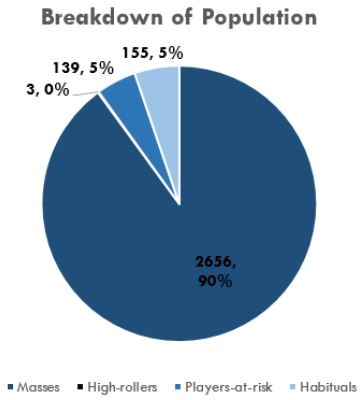
1. **Masses** – Conservative players with preference for low-odds and low-stakes that make up the bulk of the market. Their betting parameters all fall around the population mean, which is what we expect to see since they are the majority.

2. **High-rollers** – One-off low-odds high rollers, with a low bet intensity (below that of the population mean) but extremely high stake amounts of over 16 standard deviations more than the population mean. Their strategy to bet on low odds (safer bets) with high leverage seem to have paid off given their high profits (8 standard deviations higher than the population mean).

3. **Players-at-risk** – These are our less strategic players who have a preference for higher-stakes and despite playing it safe (having a preference for lower odds), these group of players lost the most. Our team would thus flag out such players for their betting patterns makes them more susceptible to irresponsible gambling.

4. **Habituals** – The most regular players in our sample (highest bet intensity, 3 standard deviations about the population mean), with greater preference for higher odds compared to the rest. And have yet made much profits. This segment too deserves greater monitoring, and Singapore Pools should be alerted if the players continue to amass larger losses.

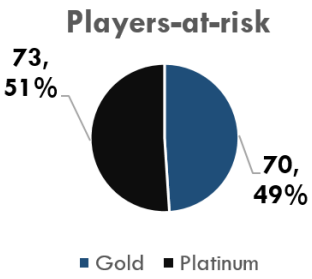
RECOMMENDATIONS



Looking into the cluster sizes and share breakdown, the "Players-at-risk" and "Habituals" make up a sizeable total share of 10% of the population (entire sample data). Meanwhile, the masses takes the majority, with approximately 90% of the players falling in that segment.

Given this finding, it would be in Singapore Pools best interest to monitor and perhaps work on strategies to promote responsible gambling in these two segments. **We would recommend simple strategies such as sending out informative emailers and text messages concerning responsible gaming to these players may generate positive returns.**

In search for a reason behind the riskier gambling patters of the "Players-at-risk" segment, we explored the demographic breakdown of the clusters and found an interesting finding. Unlike the other clusters which had a lower proportion of Platinum members (~30%), the "Players-at-risk" segment had a larger share of Platinum members (approximately 20% more).



The chief difference between a Gold and Platinum member lies in the way payment for a bet is deducted; there is an e-wallet payment system in place for Gold members where they would need to make top-ups through money teller machines before placing a bet, whereas platinum members have their gaming account linked to their bank accounts, and deduction for bets are instantly made through GIRO. Such payment arrangement makes winnings and losses less salient, for there isn't a regular need to monitor the cash flows of bet transactions (unlike the Gold members). **Therefore, we recommend strategies such as sending Platinum players reminders or transaction updates via their mobiles to inform them about their bets results and remaining value in their bank accounts, would increase their awareness of their own betting behaviour.**

DASHBOARD

Sponsor's Requirements for Dashboard:

- Ability to view latest performance updates
- Ability to identify top performing products
- Ability to highlight new overall monthly transaction trends
- Ability to flag out players who are at risk of irresponsible gambling
- Ability to explore betting patterns and preference of specific players
- Enables easy switching or input of new datasets

Objectives

✓ To highlight monthly figures updates
✓ To highlight shares of sales between player segments
✓ To explore month-on-month sales value and frequency trend
✓ To identify top (most popular) events & leagues for the month

Objectives

✓ To filter / drill down to specific player profiles
✓ To provide an overview of an individual player's bet activity / history
✓ To provide player-population comparison
✓ To highlight periods of intense betting activity of a player across a year

Automated Data Cleaning

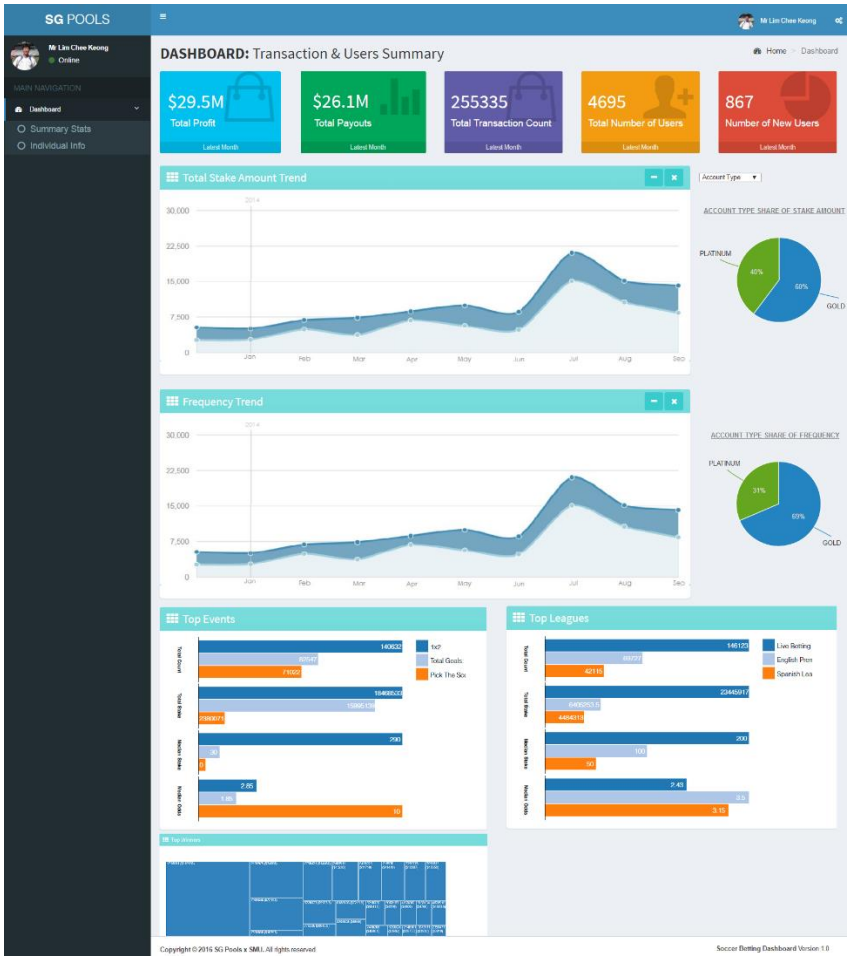
Our dashboard is designed to provide continued monitoring capabilities by accommodating future data inputs. To facilitate that feature and for the convenience of our sponsor, we had to automate the data cleaning process such that the format can be read and displayed by the dashboard. Given the large dataset Singapore Pools has each year, using Microsoft excel formula and VBA scripts to clean and transform the data would be inefficient.

There are many languages for the job (such as Python or MATLAB) but we decided to use R due to its open-source nature, having strong community support with an enormous number of different libraries/packages. The steps implemented in our R codes are similar to that described in our manual data cleaning and transformation procedures, with the final parameters as listed in our analytical data cubes, but in a JSON format – to reduce load times.

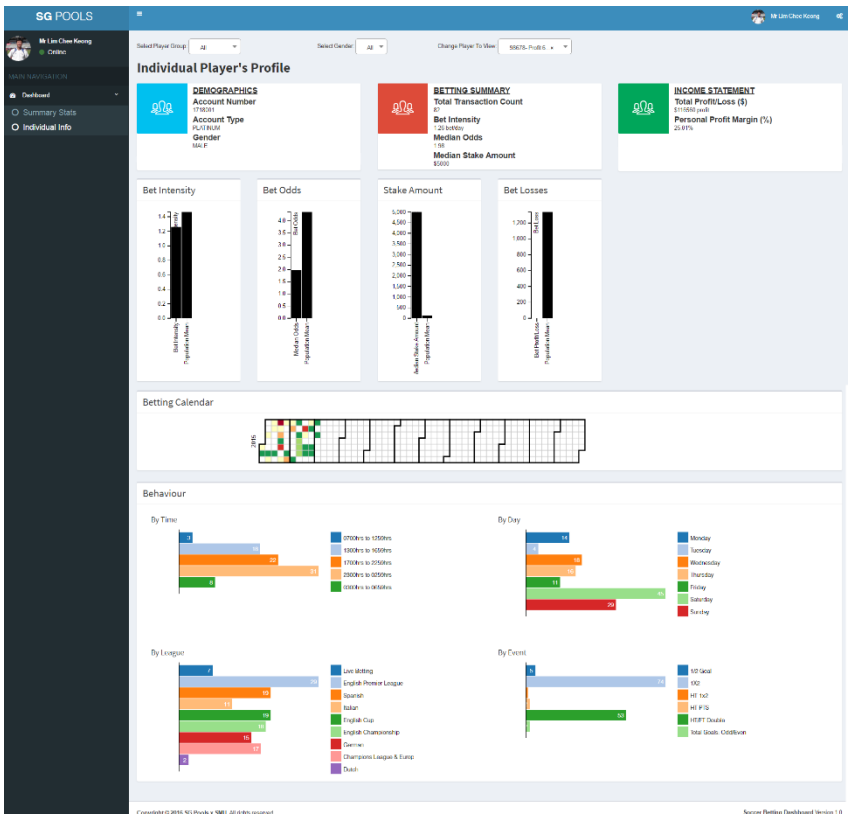
To "integrate" R into the workflow of a web application, "Full stack" solutions such as Shiny framework – for both frontend and backend – would not be feasible due to performance speed given the large dataset.

To resolve that issue we decided to keep our frontend purely HTML/CSS/Javascript and build only our backend side by R by using rApache. Our backend side will generate APIs (in which parameters are basically user inputs) which will be consumed by our frontend side. We will cater the format of the response of our APIs to be "friendly" to D3.js so that it can directly visualize the response without doing much data transformation.

With rApache, it is also possible to host our web application on a standard Ubuntu server or cloud instance so it also offer flexibility on deployment which will be helpful in case if our sponsor needs more processing power to process bigger datasets in future.



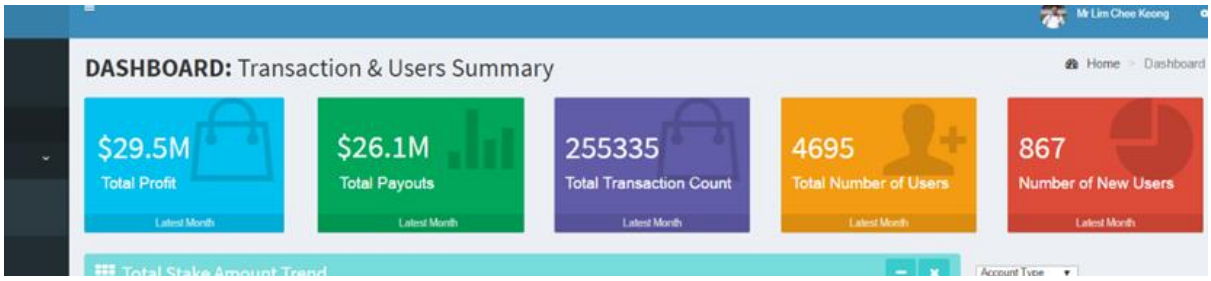
SUMMARY VIEW



INDIVIDUAL PLAYER VIEW

SUMMARY VIEW

The main view of our dashboard would be the "Summary View" page whereby users can evaluate sales trends and have a glance of their latest overall monthly performance. The first feature that the user will notice is the colourful statistics boxes above all the other charts; these boxes display the latest monthly performance figures – which were derived from discussions with our sponsor – that includes the following: (1) Total profits made by Singapore Pools (2) Total payouts (3) Total transaction count (4) Total number of active users for the month (5) Number of new users.

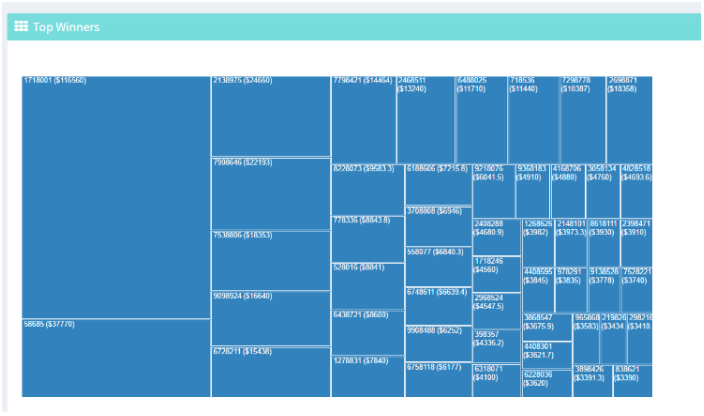


Right underneath the boxes are 2 trend line charts and pie charts. The first trend line chart displays the total stake amount (in dollars) on a month-on-month horizontal axis, while the chart below represents the frequency or count of the transaction across the time range. The pie charts on the right sides of the charts shows the share of value stake amount and count of transaction amongst the different segment categories: (1) Share of gender (2) Share of account type (3) Share of bet odds type (4) Share of bet time (5) Share of bet day. Clicking on the drop down list button above the pie charts allows the user to switch between the segments, and it too transforms the trend line by splitting the trend lines into the segments of the pie (e.g. by selecting the gender share option in the drop down list, the trend lines will split into the stake amount of all male and all female, and the same happens for the frequency trend line chart).

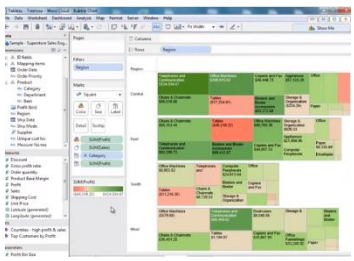
Users can use the two trend line charts to make comparisons and perhaps highlight abnormalities or unique events in certain months. For example, the usual trend or correlation is that one would see synchronize fall or rise in both stake amount value and frequency lines, however when the reverse is seen such as in the case when stake amount value rises while the count of transactions stagnant or falls, it could mean that for the particular period people have been making ridiculously high bets.

The next data visualization chart would be the bar charts underneath the trend line charts. These bar charts feature the top 3 event types (bet game options such as 1x2, or pick the score) and the top 3 leagues for the selected time period. The parameters that are displayed includes the (1) **Total frequency** (2) **Total stake amount** (3) **Median stake amount** (4) **Median odds** of the particular top event or league. The same parameters of each event or league is grouped together (placed side by side) for easy comparison to understand how these top events or leagues differ on those parameters.

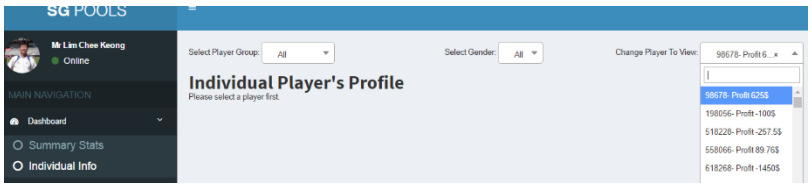
Lastly, the visualization tool at the bottom of the "Summary View" page is the tree-map chart. Tree maps are great for showing the big picture (high level view of the data) by allowing comparison of the different units within a small space, and allows us to add additional labels to the boxes for a more detailed comparison. One downside of tree maps is that it cannot size negative values, which is a huge concern in our case for we are using profits or losses are of main quantity measure, as such we had to split the losses first and later transform the data for sizing.



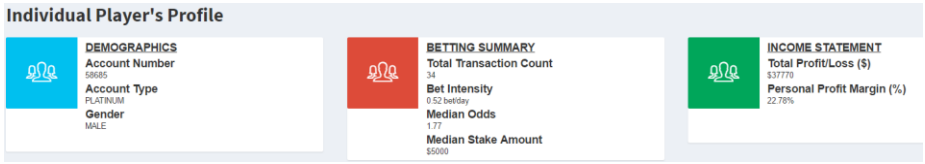
As of currently our tree map only ranks players based on the size of their profits for that is our client's main interest. Clicking on the box (individual player ID) will redirect user to that player's profile view should would provide more in-depth data about his or her bet patterns and transaction history. However, we will offer further customizations to the tree map (i.e. adding colour formatting and secondary dimensions) if our sponsors want more information to be displayed. With the use of Tableau, our team will create different sample variations of tree maps for our sponsor to pick from. For example, using colour as an identifier for gender or account type, or additional labels such as median odds. Likewise, this approach of "pre-visualizing" the charts can be applied for the other charts in our dashboards in coming user testing meetings with our sponsor.



INDIVIDUAL PLAYER VIEW

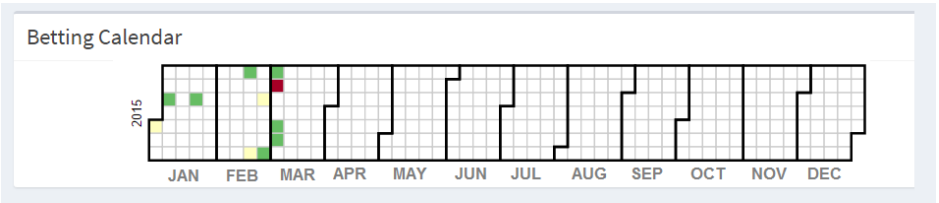


Users can enter the "Individual Player View" via two ways, either by clicking on the player cube in the tree-map on the "Summary View" page which redirects users to that individual player's profile view, or by clicking on the "Individual Info" tab on the left navigation bar. If the user performs the latter, he or she will land on an empty player view page, and in order to select a player to view, the user can use the dropdown list to filter the players to view. The users may filter users based on their "player group" and/or their "gender"; selecting the options in those filters will narrow down the selections in the third dropdown list "Player to view" making it easier for the user to locate a particular user.



Upon selecting a particular player, the user will first get an overview of the player's profile and betting behaviour, as displayed in three info summary boxes: (1) Demographics Data; (2) Betting Summary Statistics; (3) Income Statement / Performance. The selected parameters of the players were derived from our discussions with our sponsors, and aims to deliver immediate insights that would allow the users to quickly infer bet patterns to hypothesize if the player is at risk of irresponsible gambling. Still, it is recommended that the user explores the other data charts before making the claim, for these other charts will allow population comparison.

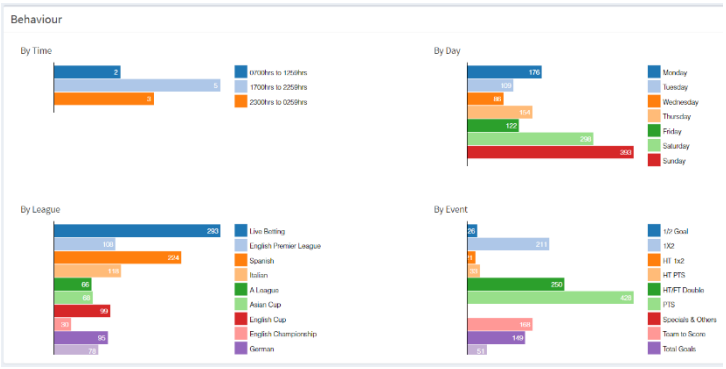
Bar charts are used as it gives us an easy comparison of the player's statistics against the population's mean at a single glance. The length of the player's bar as compared to the population's bar would let us know whether the player is under or above the mean. Other charts such as the line graphs and pie chart are inappropriate and would not allow us to interpret the comparison as easily as the bar charts.



Our team has decided to use a calendar view chart to visualize the bet intensity of a particular player over the year. The calendar view that allows to view a player's daily number of betting transactions over a year, which is quantized into a diverging colour scale. The values are visualized as coloured cells each day with red being highest quantile of transaction count and green being the lowest quantile of transaction count. The days are arranged into columns by weeks, and then grouped by month and years if the transaction spread over a year. This

allows the sponsor to easily identify any possible patterns in the transaction count of the players in this summarized view, instead of viewing these transactions from a list.

The last section of the player view is the behavioral pattern breakdown. Initially, we chose to use pie charts to compare different parts of the player's transaction count in terms of betting day, betting time period, bet leagues and bet events. However to reduce wastage of space and to allow easier segment comparison (given the many breakdowns for the events and leagues, the relatively equal slices of the pie makes comparison difficult), we have decided to use horizontal bar charts to display the breakdown. Bar charts would to provide sufficient depth of data for theses parameters for we are not interested in showing these changes over time but only interested if any of the segments stands out in terms of high frequency.



However, do note that are the graphs are still subjected to changes depending upon further feedback and request from our sponsor.

LIMITATIONS & CHALLENGES

There were four main challenges that our team faced over the course of the project.

1) LARGE DATASET

Load times was one of the key considerations we accommodated when creating the dashboard. Given the large dataset, our team had to veer away from using loops and iterations of arrays to reduce load time. Utilizing most JavaScript's native functions and libraries such as Underscore.js were some ways to optimize performance.

2) RIGID DATA FORMAT

As the result of the automated data cleaning process, the final data format to be bootstrapped into the dashboard would in a JSON format. This dataset is thus read as an object, and not as an array which is the commonly accepted input for most of the D3.js charts that are used for the dashboard. As such we had to iterate through the index array [0,..,length-1].

3) LIMITING SAMPLE DATE RANGE

The sample data that was given to us had a historical 3 month date range. Initially our implementation was coded in a way to complement the sample data range. But to accommodate future datasets of varying date range, our team had to fine tune our codes to support dynamic start and end dates.

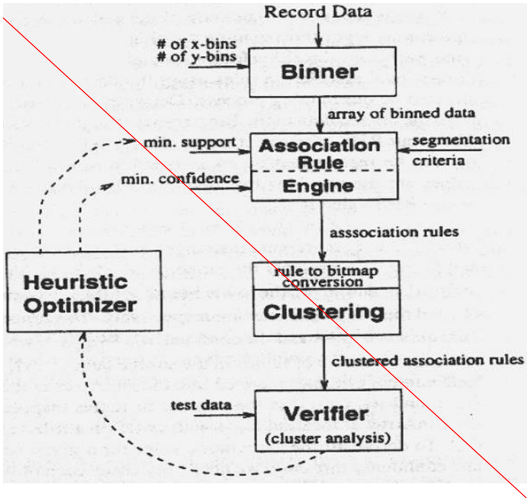
4) VISUALIZATION OF CLUSTERS

Given that k-means clustering analysis is an unsupervised learning method, we are unable to automate the analysis and visualize the clustering results on our dashboard. Interpretation of the cluster profiles would too require some degree of statistics knowledge. In light of this, our team will present the clustering analysis result as a separate deliverable.

FURTHER DEVELOPMENT

Moving forward to gather deeper consumer understanding, our team proposes a more extensive data collection procedure. Singapore Pools have collected basic **demographical data** on their Platinum users, however this was not extended to the Gold users. With greater demographic data – such as customer's salary and occupation, customer's address, customer's family background – we may discover new betting patterns, and new relationship between existing (betting behaviour) parameters and player demographic data; and so offering a better understanding of what responsible gambling should be at an individual level rather than a general take across the entire population.

Besides collecting demographic data, there are other **transaction data** such as account top-ups which could too provide greater insight on one's betting behaviour. Historical trends of one's frequency and amount of top-ups, coupled with his or her past winnings records, could reveal patterns of irresponsible gambling.



There are many papers and research carried out regarding the use rule-based classifiers and pre-learning data clustering such as association rule clustering and automated genetic clustering, but still this process is yet to be viable in the near future. Such machine learning approaches are still unperfected given the vast number of rules that needs be considered, and the incapability of the machine to learn beyond the training data. Coupled with another obvious limitation regarding the interpretation of clusters' profiles that will be left to the dashboard user, and as subjective as it is, the user would require some degree of statistics knowledge to make meaningful inferences.

Our team would therefore suggest that having a **workable methodology or guide** on how to perform the clustering analysis would be more feasible. The results from the current clustering analysis only provides a one-time-off understanding of the current market context, the insights cannot be replicated for future references. There is a need to revise the clustering analysis from time to time, updated with new datasets representative of the latest trends and context. A more sustainable solution would be to have a trained analyst to conduct the clustering analysis each round.

Lastly as mentioned above, the dataset that our team have been working on is merely a three month long dataset, as such we will continue to work with our sponsor to carry out load testing with a larger set of data. Client user testing of the dashboard is also currently on going, and we too will continue to provide support to update the dashboard charts and interactive tools upon further feedback.

REFERENCES

1. Faregh, N., & Leth-Steensen, C. (2011). The gambling profiles of Canadians young and old: Game preferences and play frequencies. International Gambling Studies, 11(1), 23–41.
2. Gainsbury, S., Sadeque, S., Mizerski, R., & Blaszczynski, A. (2012a). Wagering in Australia: A retrospective behavioural analysis of betting patterns based on player account data. Journal of Gambling Business and Economics, 6(2), 50–68.
3. Raftery, A., & Dean, N. (2006). Variable Selection for Model-Based Clustering. Journal of the American Statistical Association, 101(473), 169 –177.
4. Gainsbury, S., & Russell, A. (2013). Betting Patterns for Sports and Races: A Longitudinal Analysis of Online Wagering in Australia. J Gambl Stud (31), 17–32.
5. MCYS (2005). More than Half of Singapore Gambles; But Only 2 in 100 at Risk of Gambling Addiction.
6. Mizerski (2011). Lessons from an Analysis of Online Gambling Behaviour. Achieve Internal Excellence. University of Western Australia.
7. National council On Problem Gambling (2015). Report of survey on participation in gambling activities among Singapore residents 2014. 2–18.
8. Stone, M. (2006). Choosing Colors for Data Visualization. Perceptual Edge.
9. Few, S. (2004). Common Mistakes in Data Presentation. Perceptual Edge.
10. Few, S. (2004). Elegance Through Simplicity. Perceptual Edge.
11. Few, S. (2005). Bad Graphs: The Stealth Virus. Perceptual Edge.
12. Few, S. (2005). Intelligent Dashboard Design. Perceptual Edge.
13. Few, S. (2006). Data Visualization: Rules for Encoding Values in Graph. Perceptual Edge.