# Visualizing Future of Crowd Funding with Yu'e Bao

SONG Chenxi, WONG Yam Yip, WU Jinglong

**Abstract**— Using various data visualization methodologies and techniques, coupled with user transaction level survival analysis and time-series clustering, this project aims to build an interactive tool on R Shiny framework, so as to unearth the underlying treasures of associations between Yu'e Bao's user profiles, behavior, time and other financial factors. This will let us understand more about how People in China invest their money through Yu'e Bao and the generated insights will be valuable for internet money market fund industry.

**Index Terms**—Yu'e Bao, Interactive Data Visualization, Survivial Analysis, Dynamic Time Warping Clustering , R Shiny.

## 1 Introduction

Yu'e Bao (余额宝) is an investment product offered by Alipay (支付宝), a mobile and online payment platform established by China's multinational conglomerate Alibaba Group. In June 2013, Alibaba Group launched Yu'e Bao, in collaboration with Tianhong Asset Management Co., Ltd., to form the first internet fund in China. Since then, Yu'e Bao has become the nation's largest money market fund and, by Feb 2018, has US$251 billion under its management. In Chinese, Yu'e Bao represents "Leftover Treasure". Alipay users can deposit their extra cash, for example, leftover from online shopping, into this investment product. The money will be invested via a money market fund with no minimum amount or exit charges, with interest paid on a daily basis. While major banks offer 0.35% annual interest on deposits, Yu'e Bao may offer user 6% interest with the convenience and freedom to deposit and withdraw anytime via Alipay mobile app. Thus, Yu'e Bao became extremely popular in China.

## 2 Motivation and Objective

The dataset used in this project is released in by a competition organized by Alibaba Cloud, TIANCHI Aliyun. The competition challenges its participants to train models to predict future cash flow of Yu'e Bao users, based on historical financial data from the government, Yu'e Bao and its user, as well as their profiles. The results can aid Ant Financial Services Group, Alibaba Group's affiliate company operating Alipay, in its business of processing cash inflow and outflow. Hence, most of the works done on this dataset are focused only on achieving the best score for predictive modelling. There is no works published at the time of this project with other data analysis or insights.

In view of this, we have chosen to provide an alternate analytical approach to the dataset by building a Shiny App with interactive features, and employing the data visualization methodologies, to visualize the data and its insights interactively. We also want to perform additional analysis of survival analysis and time-series clustering, and generate dynamically visualizations of the analytical results. This visualization platform is built with RStudio, R programming language with rich libraries, Objectives aim to:

- Provide interactive visualization to users to enable them to explore the various dataset dimensions by different chart type and get corresponding insights
- Dynamically generate different customer segmentation deposit and withdraw behavior and enable users to explore and visualize the Yu'e Bao user behavior difference between customer segmentations
- Provide interactive visualization for time clustering and survival analysis, enable users to perform the analysis with different input parameters

## 3 Review and Critique of Past Work

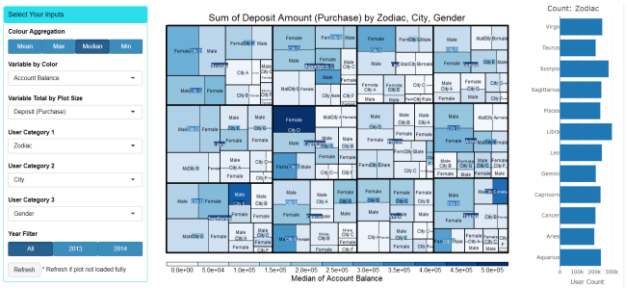
Despite the popularity of internet crowdfunding in China, there is little scholarly research in this area. All the analysis and visualization mentioned above is not interactive, though they provided summary of Yu'e Bao customer behaviour there was little or no detailed analysis on the relationship between customer profile and their behaviour. With larger proportion of funds flowing to such investment tool, a centralize and interactive data visualization platform to analysis Yu'e Bao customer segmentation and behaviour will be very helpful for the healthy growth of its ecosystem.

## 4 Dataset and Data Preparation

The source of data is Alibaba Cloud, TIANCHI, Competition: The Purchase and Redemption Forecasts - Challenge the Baseline. The dataset from this competition comprises of Yu'e Bao user's profiles, transaction behaviour, and financial interest rates over time. Data preparation was done on the original dataset to standardize and cleanse the data. Expectations and missing data point were fixed as well.

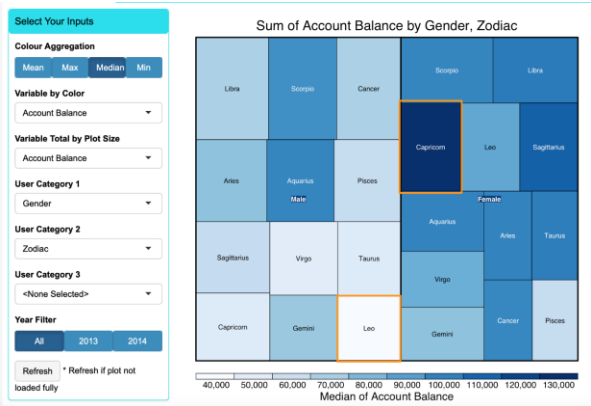
## 5 Insights and Implication

### 5.1 Hierarchical Visualization

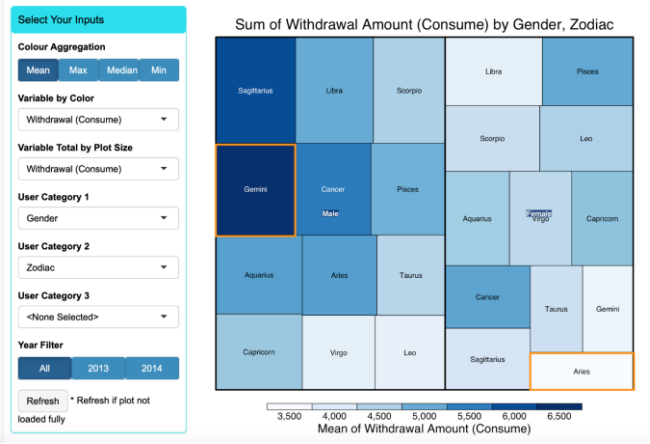



In the R shiny app, a treemap is incorporated to assist user to conduct hierarchical data analysis on user cash flow data with categorial data like user profile and time categories. Users can choose the target cash flow of their interest (account balance, deposits or withdrawals) to be represented by the colour and plot size of the treemap, via Select Input dropdown list in the input controls. For colour, users can select to aggregate the data by 4 methods (mean, median, max, min) from a Group Radio Button. For plot size, the app will perpetually derive the sum of the selected cash flow data. As the data spans across different months for 2013 and 2014, using another Group Radio Button, users can filter the data based on the year of records. Finally, to display the treemap in hierarchical structure of categorial data, users can select up to 3 levels of categorial variable from User

Category dropdown list (zodiac, city, gender, year, month, day of week). It is compulsory to select a category for the first level but users can choose omit the other 2 levels by selecting "None Selected". Based on these selections, the data will be filtered and aggregated using group_by and summarize functions of dplyr R library. Finally, the transformed data will be plot in treemap function of the treemap R library. The count of each class of the first categorial level will also be shown to the right of the treemap to give users a sense of the distribution of Yu'e Bao users across the category.

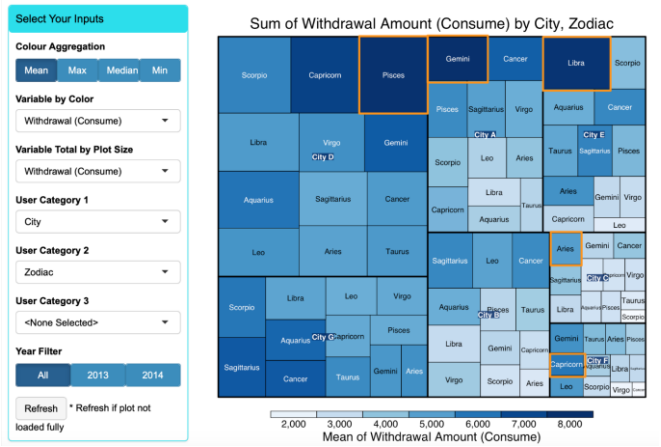



When exploring the Hierarchical visualization and by setting user category input parameters by the three levels: city, zodiac and gender, we found that female Capricorn users has the highest median account balance of 130K yuan, while male Leo users has the lowest median balance of only 40k yuan.
When changing the cash flow variable to mean withdraw amount, we can see that male Gemini users has the highest mean shopping withdraw of around 6.5k, while female Aries has lowest mean shopping withdraw of around 3.5k.
When exploring the Hierarchical visualization and set the user category input parameters by the three layers: city, zodiac and gender, we found that female Capricorn users has the highest median account balance of 130K yuan, while male Leo users has the lowest

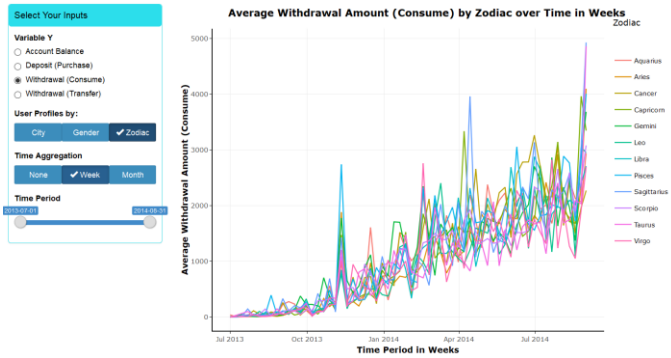



median balance of only 40k yuan.
When change the observe variable to mean withdraw amount, we can see that male Gemini users has the highest mean shopping withdraw of around 6.5k, while female Aries has lowest mean shopping withdraw of around 3.5k.

Next, by changing the user category to City and Zodiac, we found that different city has different top zodiac withdrawal (consume) group. In city D, Pisces users has the highest average withdrawal by consumption of around 8,000 yuan whereas Gemini users has the highest of 8,000 yuan in city A. The user groups by zodiac with highest withdrawal by consumption in city C, F are Aries and Capricorn respectively. We do not see significant top zodiac users' group with highest withdrawal by consumption activity in city B and G.

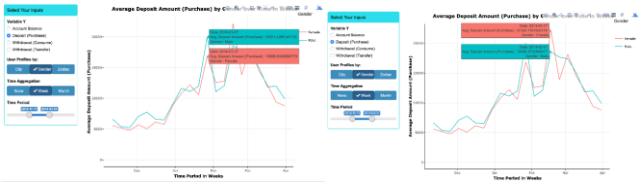



## 5.2 Time series analysis



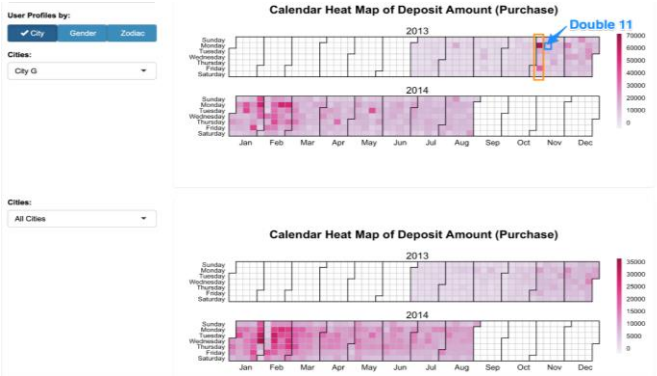

The time series line graph plots the trend of Yu'e Bao users' cash flow over time. App users can select the cash flow variable (Variable Y) from a list of radio buttons. The time series line will be subdivided into classes of user profile categories, which users can select from the User Profile radio group buttons. Additionally, app users can choose to aggregate the time series cash flow data to weeks or months by selecting from the Time Aggregation radio group buttons. To allow users to filter and view particular time period of interest, a 2-selection slider bar is added to filter according to values of time. As the values of time change depending on the selection of time aggregation, the sliderbar is set to be reactive to changes in time aggregation and will automatically update itself accordingly. Finally, all selections will be filter, grouped and summarize using the corresponding functions in the dplyr library, plot using ggplot, and wrapped by ggplotly function of plotly R library, to generate an interactive time series line graph.

To further understand the Yu'e Bao's user behaviour, it's essential to look into different user groups' behaviour over time. Cash flow time series data like account balance or deposits can be grouped by their user profiles like gender or zodiac sign to visualize their individual trends over time. Data can also be aggregate across time by weeks or months. When we plot average deposit by gender over time and zoom into 2014 Chinese New Year period (early February), we found that on the week before CNY average deposit by male is ~20% higher than that of the female; conversely, female deposits where ~20% higher than that of male the week after CNY.

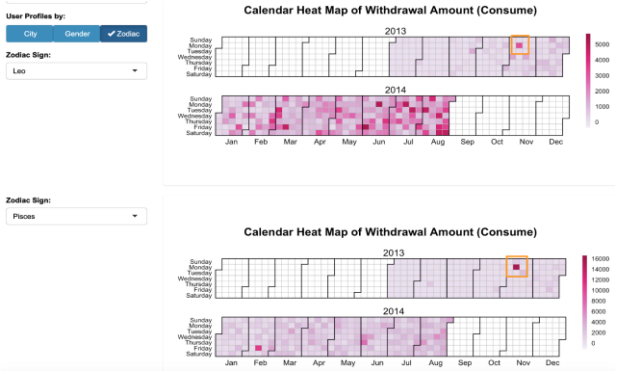


## 5.3 Calendar heatmap comparison

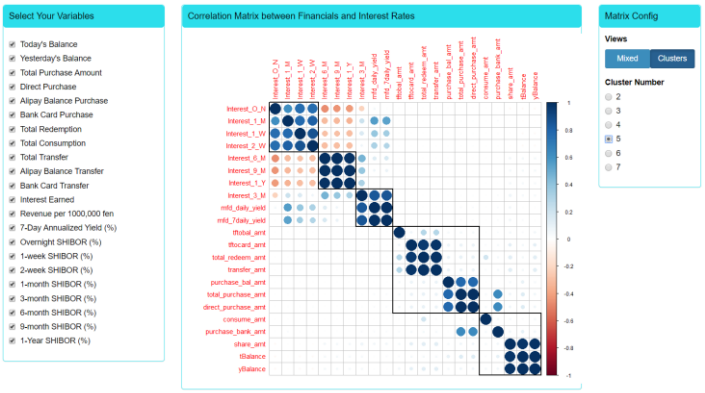
Calendar heatmap allows users to visualizes time series data attributes over days in a calendar-like view, making it easy for them to identify daily patterns or anomalies. In our implementation, we want to allow users to compare 2 calendar heatmaps of different classes within the same user profile categories, which can be selected via the User Profile radio group buttons (labels will also be updated accordingly). Once selected, the dropdown list, for each heatmap, will automatically update to the unique classes of the selected user profile, from which users can select 2 to compare. Upon these selections, the corresponding calendar heatmap will be plot using the calendarHeat.R function created by Paul Bleicher which uses R libraries like plyr, chron, grid and lattice.
From the calendar heatmap we can see that during 2013 double 11 (Nov 11th, biggest online shopping festival in China), only city G has significant higher deposit on Monday and Friday the week before the mega sales day compare with other cities. Our guess is that before double 11, the online sellers needs to stock up to prepare for the mega sales and they will be paying manufactures who could possibly be found mostly in city G. Hence, a significant amount of money was deposit to Yu'e Bao during that time.



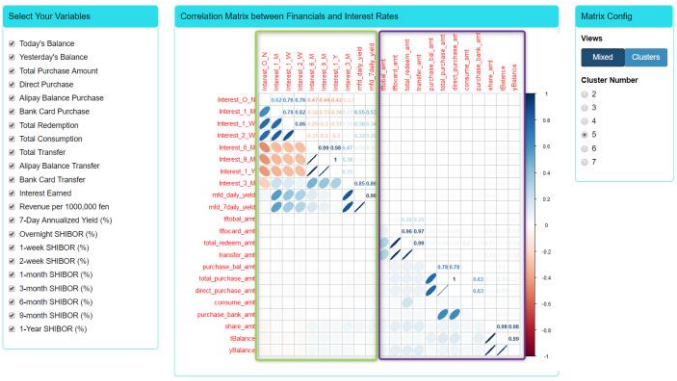

We set the parameters to observe the withdrawal (consume) amount between users of different Zodiac signs. The result shows on Double 11 of 2013, Pisces users spent around 16,000 yuan while Leo users spent only around 5,000 yuan averagely.




## 5.4 Correlation matrix




The correlation matrix is designed to display the correlation between continuous variables of user cash flow and financial interest rates, and confirm if there are any correlations between any of these 2. The continuous variables to be included in the matrix can be selected from the checkbox group input. By default, all variables are selected and users can choose to include or exclude any of them. Users can also choose to view the matrix in Mixed or Clusters view. For the cluster view, as shown above, a hierarchal clustering will be performed for the variables and users can select the number clusters to create. With these selections, the data will be fileted based on the selected variables and the resulting data will be visualized using the corrplot R library.
As shown below, for the mixed view, the matrix is sub divided into 2 triangles where one will show the correlation coefficient between variables and the other will show the ellipse, whose shape and colour will change according to the correlation coefficient. A rounder ellipse represents higher magnitude of correlation while the slant, to the right or left, represents a negative or positive correlation (like gradient of line plot). Similarly, a darker colour will represent higher magnitude of correlation, while blue and red colour will represent positive or negative correlation respectively. From the diagram below, we can see that there is little or no correlation between interest rates (green box) and user cash flow data (purple box).




## 5.5 Interactive Animated Bubble Plot

A scatter plot can help users visualize the relationship between 2 cash flow (continuous) variables. The X and Y dimensions of the scatter plot can be selected from the dropdown list of cash flow variables. The scatter plot is upgraded to a bubble plot by adding

additional dimensions of colour and size to the scatter markers. The marker size takes another cash flow variable from the dropdown list and the marker colour will represent a user profile or time categorial variable selected from the Marker Colour dropdown list. The user can also choose to split the colour marker by a secondary variable selected from the Marker Secondary dropdown list. By selecting the Animation toggle button, it will add a 5th dimension of time (in months) to form an interactive animated bubble plot. Based on these selections, the dates will be transformed to the format of just year and month using mutate function of dplyr R library. This is then again aggregated by the group_by and summarize functions. Next, the (animated) bubble plot can be generated via plotly. By running the bubble plot in animation mode, users can observe how the relationship of these 5 dimensions change over time, in months, and some trend may be observed.

### 5.6 Survival Analysis and Insights

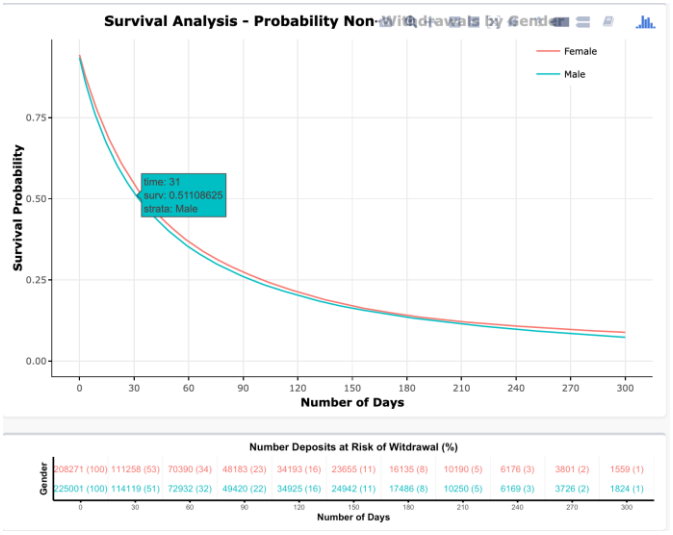

Survival Analysis is used to explore the impact of different factor on expected time difference between each deposit/withdraw activity. Raw dataset transactions consist of customer deposit and withdraw activities, to conduct survival analysis we need to transform the data so that it contains; (1) Duration of observation and; (2) Status of observation. In our project context, individual withdraw records need to be paired with one or more deposit records. We conducted this mapping based on FIFO (First in first out) and LIFO (Last in first out) assumption. In FIFO approach, we map always the first deposit transaction with first withdraw transaction, then calculate use the difference of deposit and withdraw date as the survival duration. In the LIFO approach, we map the last deposit transaction with first withdraw activity. In case the deposit amount is different from withdraw amount, we map the transaction with smaller amount first and then map the leftover transaction with bigger amount.

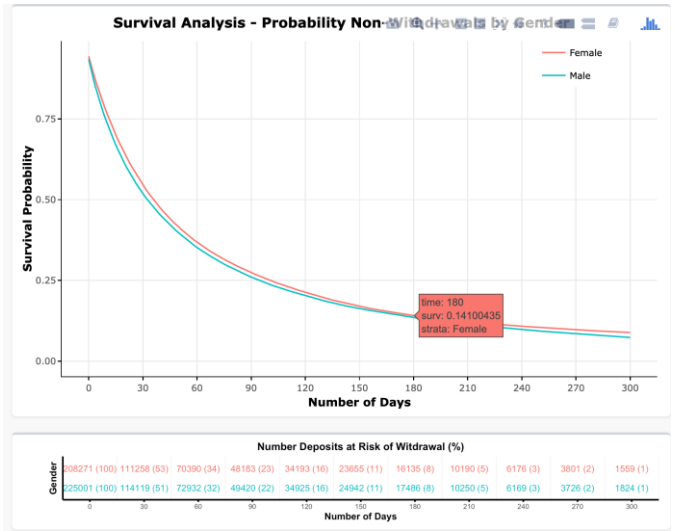
In the Shiny app, users are able to conduct the survival analysis interactively. Options are provided for user to choose transaction mapping logic (LIFO/FIFO) and variables to conduct survival analysis (Gender, City, Zodiac etc.).

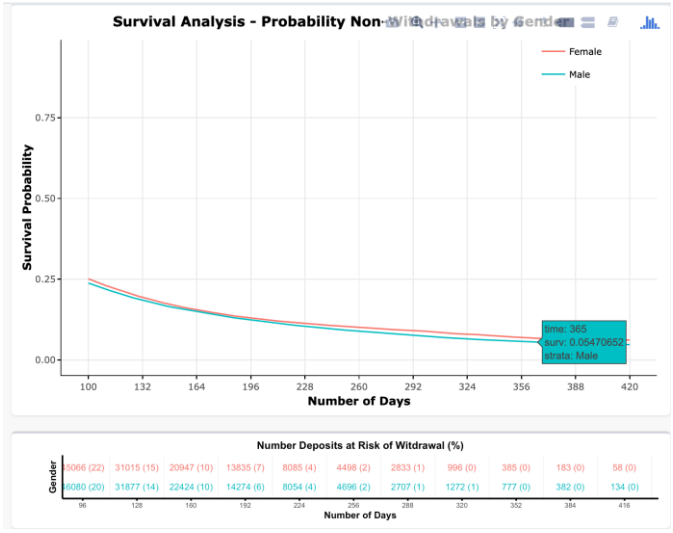
Kaplan–Meier curves will be generated to show the survival possibility (users keep their money in Yu'e Bao) among duration (number of days). A risk table is shown to display the number and percentage of observations at risk. Censor plot shows the distribution of censors activity by different variable (in this case people don't withdraw and no further data is provided).

We found that the survival analysis result does not differ much between FIFO and LIFO dataset, this is probably due to most of the transactions in Yu'e Bao follows the single deposit and withdraw pattern, the mapping result in that case we be the same.

Overall, 50% of users will withdraw their money within 1 month after deposit. 86% of user decided to withdraw their money before 6 months, only 6% of users deposit their money without withdraw for more than 1 year.

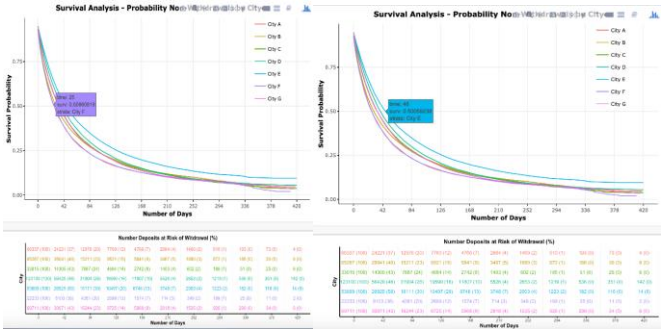


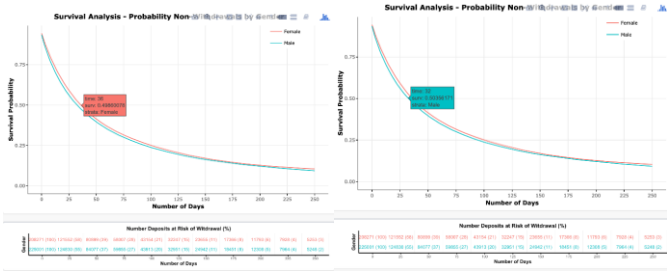




Probability of non-withdrawals by different city differs, City F users has higher probability of withdrawal compare with other cities, City E users has low probability to withdrawal their money in Yu'e Bao for the same amount of time duration. 50% of City F user

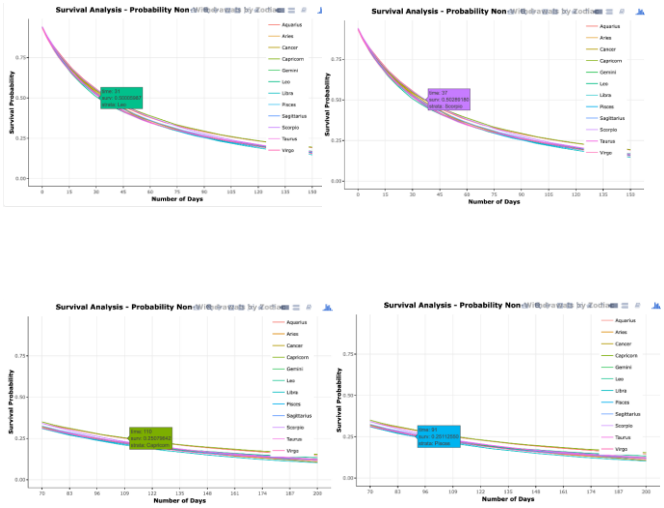
transactions will be withdrawn at 25th day while for City E user transactions it is 46 days. In Alipay apart from Yu'e Bao, there are also other investment tools to gain better interest rate if user place fix deposit for 30 days or 45 days, if Alipay would like to promote these tools City E seems a good place to start with.



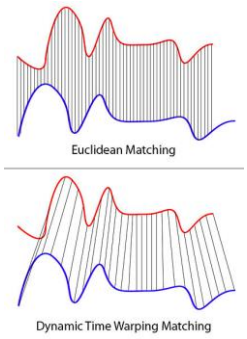
Regarding gender, different from people's stereotype, averagely males withdraw their money earlier than female users. 50% of male users' transactions are withdrawn in 32 days where 50% of female users' transactions are withdrawn in 35 days.



For different Zodiac, 50% of Leo users transactions are withdrawn within 30 days, while for Scorpio users it is 38 days. 25% of Pisces users' transactions are withdrawn after 91 days where 25% of Capricorn and Aquarius users' transaction s are withdrawn after 110 days.



## 5.7 Dynamic Time Warping Clustering and Insights




Time series clustering is the partition of time series data based into segments based on their similarities or distance between each other[1], and one of the algorithms to do this is dynamic time warping (DTW) by measuring similarities between time sequences, which may vary in speed[2]. At each time period of a time sequence, the distances to a varying number of time periods on another time sequence is measured. The number of additional time period on the second sequence is controlled by a configurable windows size that defines the limit before and after the time period of the first sequence. In this way, DTW clustering takes into consideration shifts or distortion in time sequences, when measuring the similarities between given sequences, independent of non-linear variations, to generate desired clusters. In R, this is implemented by TADPole Clustering in dtwclust and TSclust R library.

In this analysis, we are interested in clustering Yu'e Bao users based on their account balance over time. 28,041 Yu'e Bao users' account balance data is transformed into time series using the spread function of tidyr R library. However, there are a large amount of user in this dataset without cash flow activities (zero account balance throughout). These users are removed from our analysis so as not to affect the clustering, leaving behind 14,923 time series. To further reduce the resource requirement of our clustering analysis, we separately aggregate the mean of account balance by time period in weeks and in months, using apply.weekly and apply.monthly functions of xts R library. For each of the 2 new data tables (weekly and monthly aggregated), the data is use to perform DTW clustering using tsclust function of TSclust R library, which is also highly reliant of dtwclust R library. Firstly, clustering type is selected as "tadpole" to define the clustering method as TADPole Clustering. Various clusters sets are generated using different permutations of:

- Number of clusters – number of clusters to generate
- Distance cutoff – distance between time series within this limit is considered a neighbour
- Window size – number of time points to measure distance before and after each time period

In total, 210 different cluster sets are created for all weekly and monthly data. To measure the performance and validity of these clusters sets, 6 cluster validation indicator (CVI) values are generated for each cluster set. The 6 CVIs are as follows:

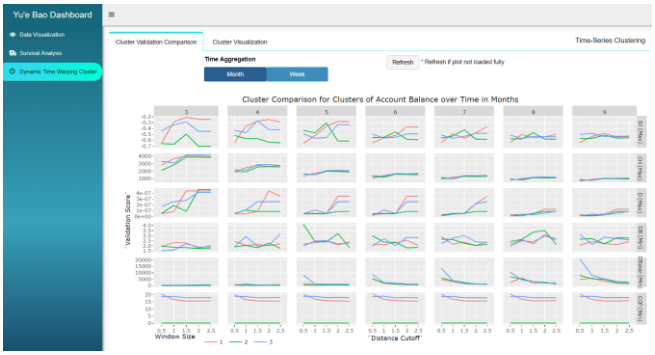
1. Silhouette index (Arbelaitz et al. (2013); to be maximized)
2. Calinski-Harabasz index (Arbelaitz et al. (2013); to be maximized)
3. Dunn index (Arbelaitz et al. (2013); to be maximized)
4. Davies-Bouldin index (Arbelaitz et al. (2013); to be minimized)
5. Modified Davies-Bouldin index (DB*) (Kim and Ramakrishna (2005); to be minimized)
6. COP index (Arbelaitz et al. (2013); to be minimized)

To visualize the comparison of CVI metric for different clusters sets with different permutations, they are plotted using R library ggplot –

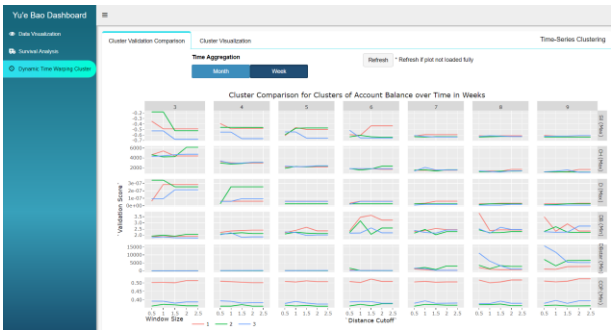
[1] Time Series Clustering and Classification, RDataMining.com, http://www.rdatamining.com/examples/time-series-clustering-classification

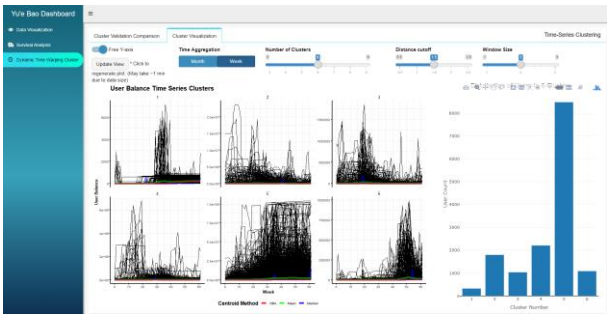
[2] Dynamic time warping, Wikipedia, https://en.wikipedia.org/wiki/Dynamic_time_warping

line graphs with facet-grid and wrapped with R library plotly in the Cluster Validation Comparison sub-function as shown below.



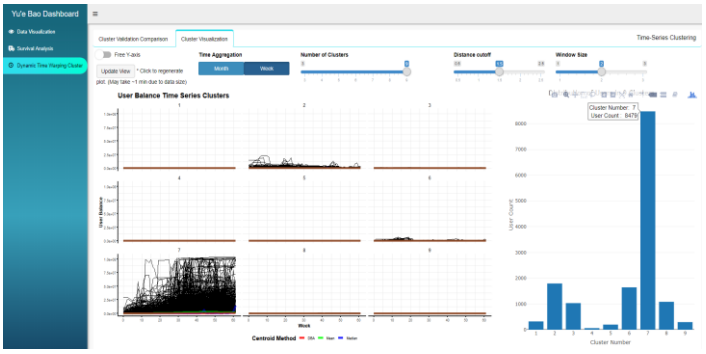

From the comparison, shown above, we can see that cluster set of less clusters generally generate better CVI than cluster sets with more clusters. While changes in window size give mixed results depending on CVI, distance cutoff of 1.5 or 2 give better CVI value for monthly aggregated data. For weekly aggregated data, shown below, distance cutoff of 1 is generated better CIV metrics. More detailed explanation of this function will be explained in later section of application user guide.




With a cluster set of interest selected, we can also visualize the comparison between the clusters using the Cluster Visualization sub-function. For example, the image above shows a set of 6 clusters generated from the weekly data using DTW parameters of distance cutoff 1.5 and window size 2. It is clear that cluster 5 is the dominating cluster with 8,479 users making up ~57% of the total 14k users. Note also that the y-axis scale is not the same for all clusters and Cluster 5 has significantly higher account balance that other clusters. We observe that user of Cluster 5 are early adopters of Yu'e Bao who continue to invest and increase their balance in Yu'e Bao through the period. For the other smaller clusters, we see a cluster of late adopters in Cluster 6. Additionally, Cluster 1 user appear to be mid-term adopters, whose balance spiked initially, but subsequently reduced and stayed relatively constant for the remaining period. Cluster 2, 3 and 4 showed a vary increase in account balance through the time period.

We next turn our focus to the domination cluster. As shown above, even if we increase the number of clusters, the dominating cluster remained at the same user count which means that new clusters are split from the smaller clusters instead of the dominating one. In this case, the dominating cluster could represent most typical Yu'e Bao users, while the other clusters could be abnormally or outliers. This also explain why smaller number of clusters gives better CVI as mentioned in the earlier section.




## 6 Future Work

Additional functions:

- Auto data aggregation and data preparation
- Group users with the deposit and withdraw amount basket, perform more specific user profiling analysis.
- In survival analysis, look into transaction amount, explore the difference between high, medium and low transactions amount survival duration by different data dimension
- Explore more deposit and withdraw mapping rules and observe the difference in data pattern
- Add trace lines while the bubbles are in transition during the animation of bubble plot so users can see the history of transition clearly.

Real world use cases:

- Connect the shiny app with Yu'e Bao data warehouse, so that it can also perform analysis on the data after 2014 with a bigger user group
- Automate the creation of data object for each function so that an data upload function can be created for more internet crowdfunding platform data to be analyzed and visualized in this app
- Use this shiny app as the data dashboard for Yu'e Bao operation team, for performance monitoring, customer engagement and target marketing purpose


## Acknowledgements

The authors greatly thank Dr Tin Seong KAM for his guidance and recommendations.