Small Bank Cash transactions

Table of Contents

Small Bank Cash transactions **1**

Clustering of Time series **2**

Introduction - What is this about? *2*

Objective *2*

Characteristics of data **3**

Data exploration **4**

Unique combinations in data *4*

Number of data records with above combinations including branches *5*

Analysis by Type of Account *9*

Segment level exploration. *10*

Outlier corrections - Deposits and Withdrawals *11*

Analysis for Branches by their business volume. *16*

Clustering **19**

Clustering deposits *19*

Heirarchial clustering - SBD *21*

Plot of clusters - SBD 25

Hierarchical clustering - DTW *28*

Plot of clusters - DTW 32

Clustering for withdrawals CWA **36**

Heirarchial clustering - SBD *38*

Plot of clusters 38

Hierarchical clustering - DTW - 6 clusters *41*

Heirarchial clustering - DTW - 7 clusters *42*

Plot of branches by clusters - SBD 43

Conclusions **48**

Next Steps **48**

Clustering of Time series

Introduction - What is this about?

Time series forms one of the aspects of looking at the past and predicting future. It reflects the resultant of numerous (or few) factors affecting over time to influence certain behaviour. Time series analysis is usually undertaken when one cannot articulate all the underlying variables affecting outcome. Examples of time series are - sales over past few years, stock price over a day or month or year. There are many variables and not all can be articulated.

And why would past reflect future? Good question. On a going basis, the assumption is that the outcome at a particular instance in past was sum total of numerous variables and going forward in a similar instance of time, the influence would be corelated. e.g. sales showing seasonality at a week/month/yearly levels.

In a typical business processes, time series analysis and forecasting will be part of periodic planning cycle and would be a good starting point for superimposing business strategies in future time periods. In an organizational context, there would be significant number of products or SKUs (Stock keeping Units) that would be planned across various distribution centres.

So where does clustering fit in? Clustering in theoretical sense is a grouping of entities which display similar behavioural characteristics. It can be of one's interest to find out which entities exhibiting common characteristics to articulate strategies to a specific cluster to gain control over them.

Objective

This dataset is of a small financial institution with transactions of cash deposit and withdrawals over a period of time. These are at the day level. There is a geographical spread of various branches. Each branch will have its own characteristics of deposits and withdrawals based on the type of catchment area it caters to.

Objective here would be to cluster various branches based on behaviour of deposits and withdrawals. Overarching objective (beyond scope of this exercise) which such study would be to use this cluster information to develop forecasting model with its parameters tuned to a certain cluster.

With this objective in mind, we will proceed with exploration of data, cleaning it and then clustering it.

Characteristics of data

Column descriptions Branchid - Total number of branches `128`
Type of Account - Different types of account `CA, SA, RD, NA, FD_QO, PPF`
Segments - Different segments as part of accounts `GPCD, HH, AAA, NON_RTL, SAL, SEG, TAS, WEALTH, NRI RETAIL, DUMMY, IPB, RFIG`
date - Earliest start date - `2010-10-01` - Last date in data - `2012-09-28`
Total count of transactions
Count of withdrawals
Count of deposits
Cash withdrwal amount
Cash deposit amount

Total number of records in data (including NAs) `390755`
Number of unique combinations for type of accounts and segments for each account `37`

```
##      SOLID          TRAN_START_DATE   TYPE_OF_ACCT      SEGMENT
##  Min.   :  1.00   2012-08-21:   852   CA   :183567   HH     :103040
##  1st Qu.: 21.00   2012-09-10:   846   FD_QO:    58   SAL    : 63372
##  Median : 48.00   2012-04-03:   837   PPF  :   146   SEG    : 63033
##  Mean   : 52.84   2012-08-13:   836   RD   :  3904   NON_RTL: 57895
##  3rd Qu.: 79.00   2012-07-09:   835   SA   :202826   WEALTH : 33957
##  Max.   :128.00   2012-08-16:   831   NA's :   254   TAS    : 26632
##                   (Other)   :385718                  (Other): 42826
##   COUNT_ACCTS    CASH_WITHDRAWAL_CNT CASH_DEPOSIT_CNT CASH_WITHDRAWAL_AM
T
##  Min.   :  1.0   Min.   :  0.00      Min.   :  0.00   Min.   :0.000e+00
##  1st Qu.:  1.0   1st Qu.:  0.00      1st Qu.:  1.00   1st Qu.:0.000e+00
##  Median :  4.0   Median :  1.00      Median :  4.00   Median :1.770e+04
##  Mean   : 14.4   Mean   :  3.51      Mean   : 12.27   Mean   :1.081e+06
##  3rd Qu.: 18.0   3rd Qu.:  4.00      3rd Qu.: 16.00   3rd Qu.:1.923e+05
##  Max.   :306.0   Max.   :218.00      Max.   :298.00   Max.   :4.810e+09
##
##  CASH_DEPOSIT_AMT         date
##  Min.   :0.000e+00   Min.   :2010-10-01
##  1st Qu.:1.747e+04   1st Qu.:2011-05-24
##  Median :1.148e+05   Median :2011-11-15
##  Mean   :3.683e+06   Mean   :2011-11-01
##  3rd Qu.:4.583e+05   3rd Qu.:2012-04-28
##  Max.   :6.076e+09   Max.   :2012-09-28
```

Data exploration

Unique combinations in data

This list signifies the various combinations of type of account and segment contained in data set for each branch. This is extended to each branch and hence there many time series' as seen later.

```
##        TYPE_OF_ACCT    SEGMENT
## 1                CA       GPCD
## 2                CA         HH
## 3                CA        AAA
## 4                CA    NON_RTL
## 5                CA        SAL
## 6                CA        SEG
## 7                CA        TAS
## 8                CA     WEALTH
## 9                SA       GPCD
## 10               SA         HH
## 11               SA    NON_RTL
## 12               SA NRI RETAIL
## 13               SA        SAL
## 14               SA        SEG
## 15               SA        TAS
## 16               SA     WEALTH
## 25               SA      DUMMY
## 53               SA        IPB
## 76               RD         HH
## 140            <NA>         HH
## 232            <NA>     WEALTH
## 472              RD        SAL
## 799              CA NRI RETAIL
## 2449           <NA>        SEG
## 5565             CA       RFIG
## 15021         FD_QO         HH
## 15502         FD_QO        SAL
```

```
## 28970            RD        SEG
## 55064            RD     WEALTH
## 77934            CA        IPB
## 79050           PPF        SAL
## 79099           PPF         HH
## 87291        FD_QO NRI  RETAIL
## 95230            CA      DUMMY
## 130980          PPF     WEALTH
## 141466           RD        TAS
## 327586         <NA>        SAL
```

Number of data records with above combinations including branches

Show 5102550100 entries

Search:

	SOLID	TYPE_OF_ACCT	SEGM
1	1	CA	AAA
2	1	CA	GPCD
3	1	CA	HH
4	1	CA	IPB
5	1	CA	NON_RTL

Showing 1 to 5 of 2,214 entries

Previous12345…443Next

Number of time series in dataset - No. of branches X type of accounts X Segment = `2214`

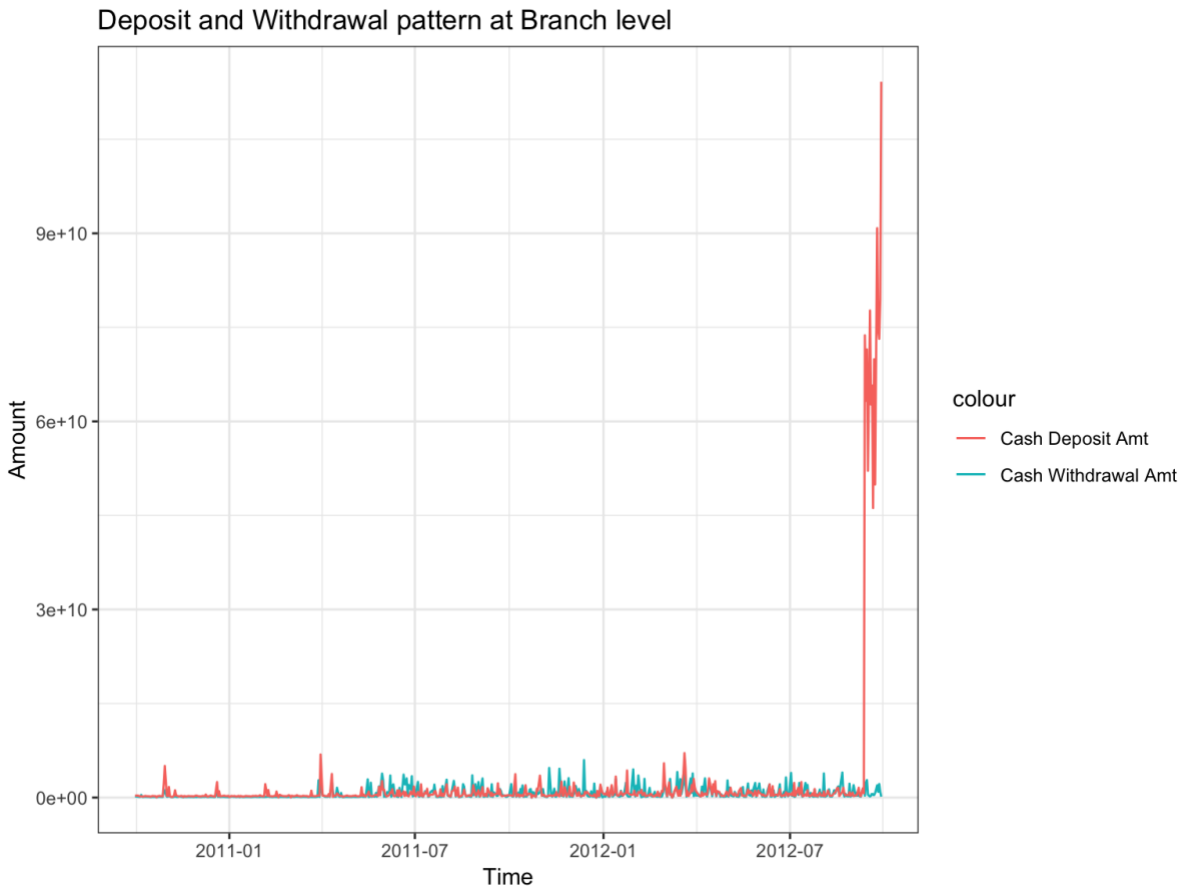
Above table shows that there are NA in Type of Account. Imputing values would not be logical as the segment belong to different type of account. Hence these have to be removed.

Number of NAs in the data - `254`

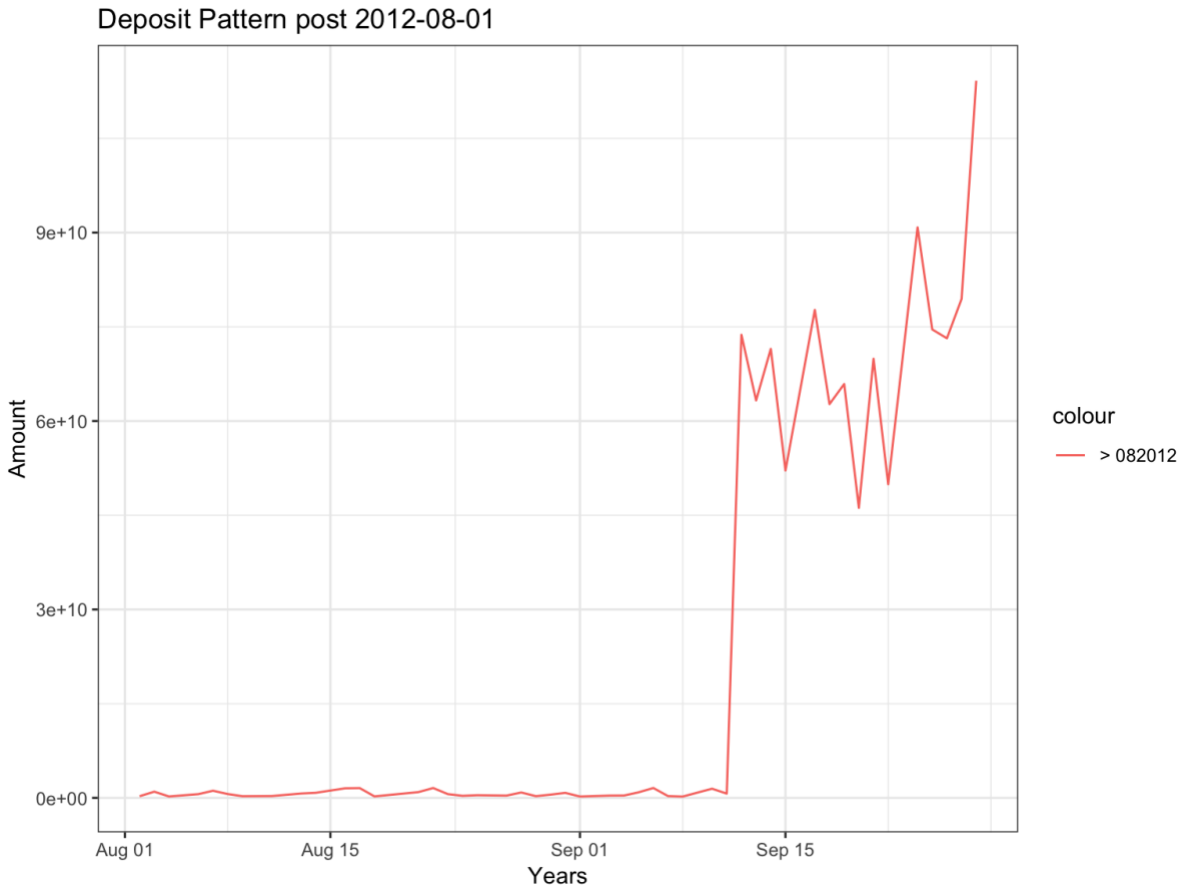
No of rows of data after the NAs are removed - `390501`

Aggregation by date

We will take a top down approach. By aggregating the data and checking at high level may highligt if there are any major descrepancies. Here we have aggregated data by date for all branches, account type and segment. We will plot cumulative withdrawal and deposit amounts by date for complete time period.



Plot shows that there is a major spike seen in the data for cash deposit amount(CDA) in last couple of month of data. Let us magnifiy that period to take a closer look.



Are these ouliers?
Relative to complete period, there is manyfold jump in the cumulative amount only in last couple of days of data period. It does seem they are outliers but it would warrent further analysis at a detailed level since current data is aggregated data. We need to find few reasons before we term it as outlier.

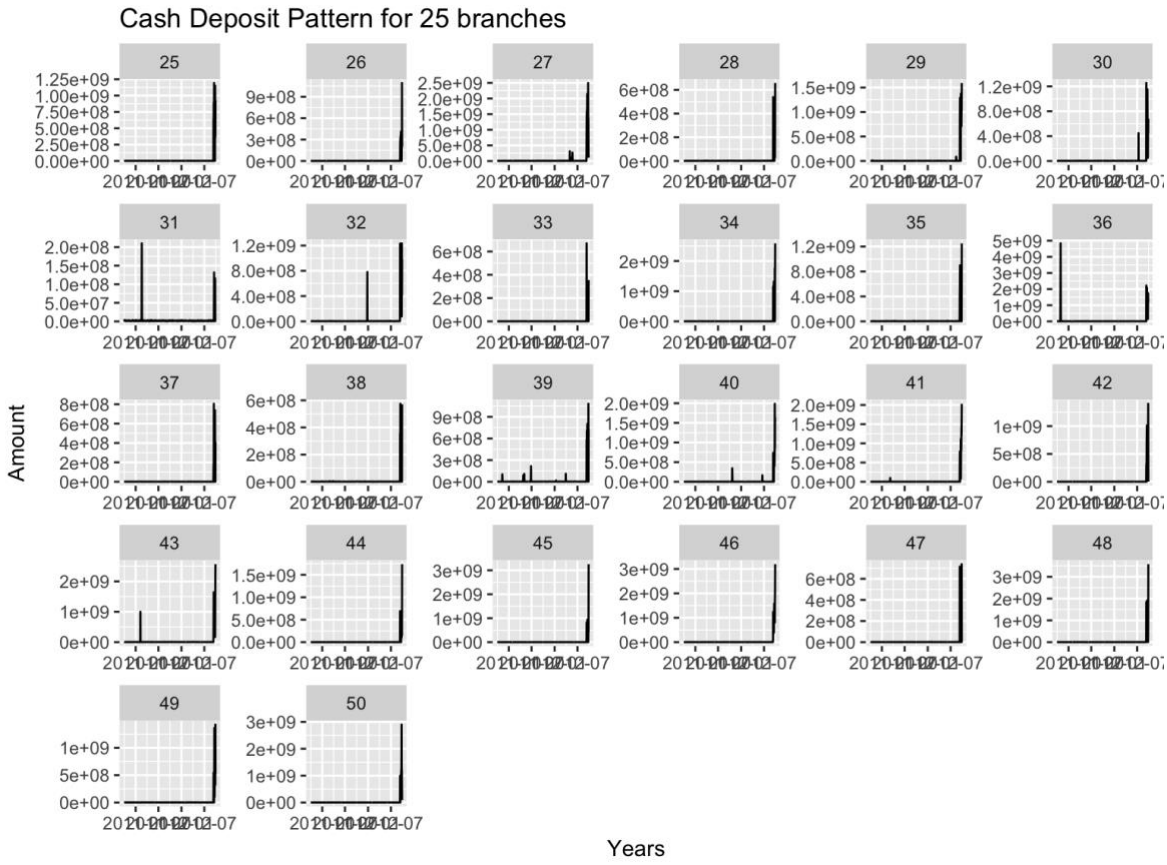
a. Is it reflected in a spike in the transaction counts for deposits too? More number of deposits can be a reason of spike.

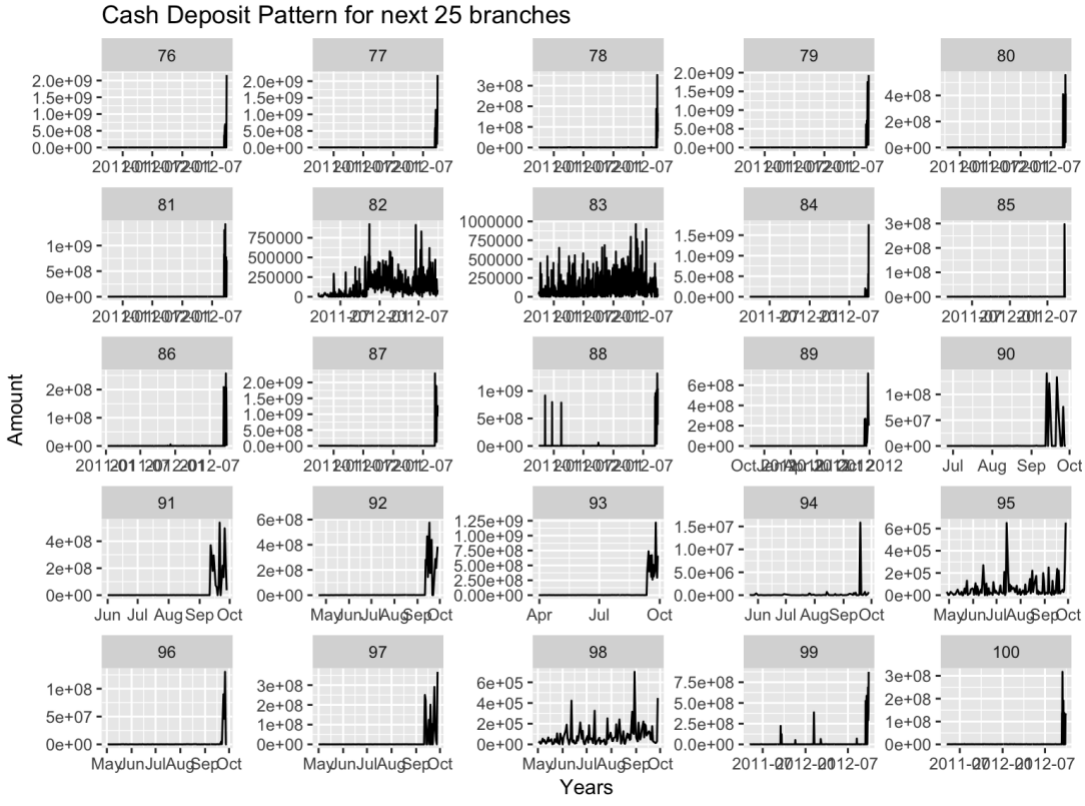
Both time series show stable pattern with no major spikes as seen in the later dates for Cash Deposit amount.

b. Is spiked pattern reflected across all the branches or only few?

Lets plot few branches for deposit patterns.

Whoo!.. All the branches have some spikes in the later period.. Lets confirm with another rest of branches if there is deviation. This is leading to a pattern…


Cash Deposit Pattern for 25 branches
Amount
Years


Cash Deposit Pattern for next 25 branches
Amount
Years

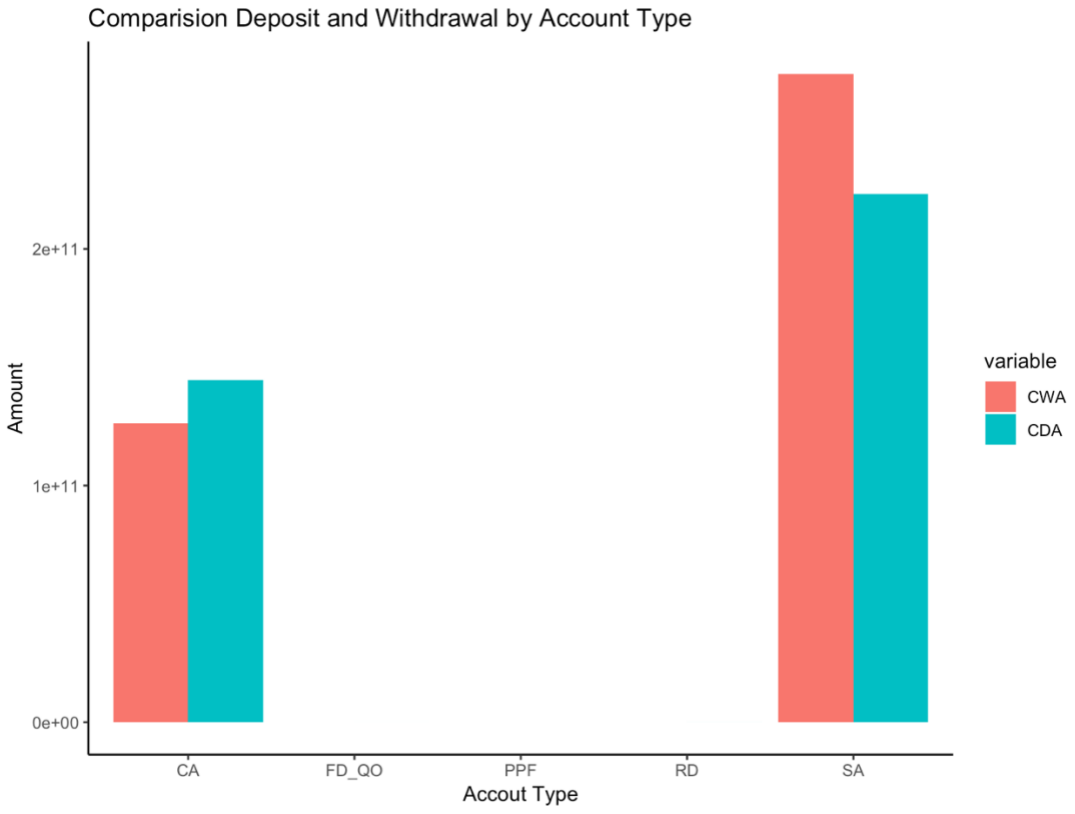
Confirmed!! - There seems to be some data issue as all the branches exhibit similar behaviour. We have option of either using outlier correction or ignoring this part of data. The better option here seems to be ignoring the data as outlier would also either ignore the data point or we would need to come up with a appropriate logic to impute data e.g. mean, median etc. since there is no significant loss of data due to this (last 12 days) it would convenient to ignore.

Now lets check how withdrawals look like for few branches.

Data looks much cleaner now reveling few more actual outliers and steady pattern in many cases. Before we are happy, lets confirm for few more branches..

Analysis by Type of Account

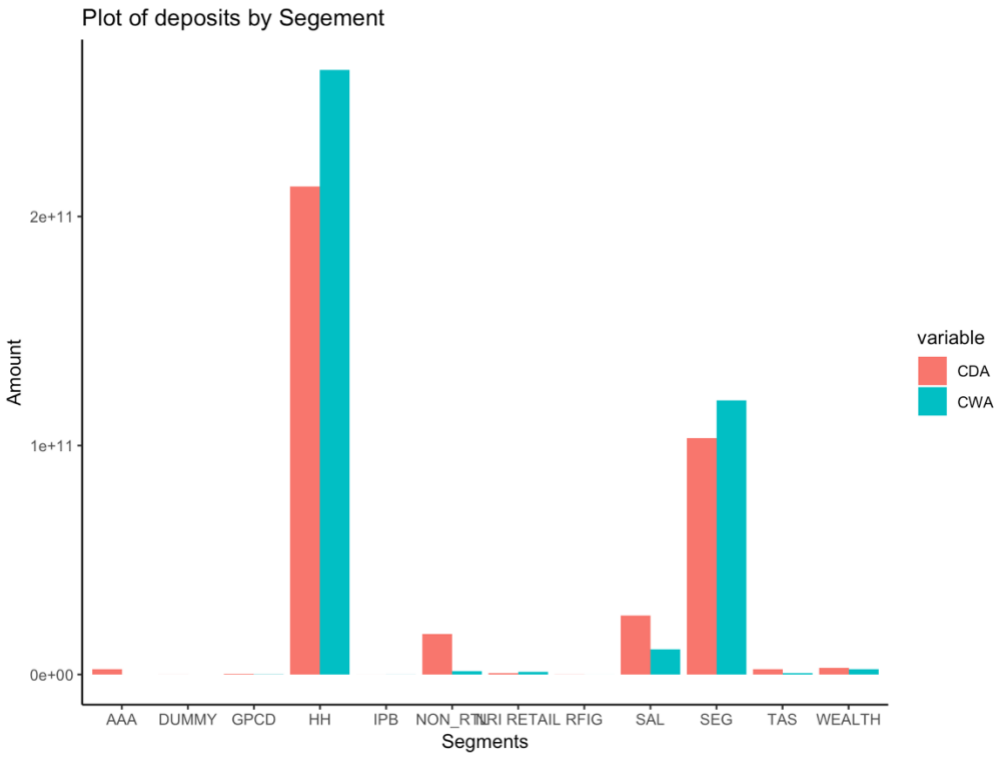
With data looking good at aggregate level for both deposits and withdrawals, lets check at next characteristics of type of account. A below bar plot will show how the deposits and withdrawals fair at account levels.

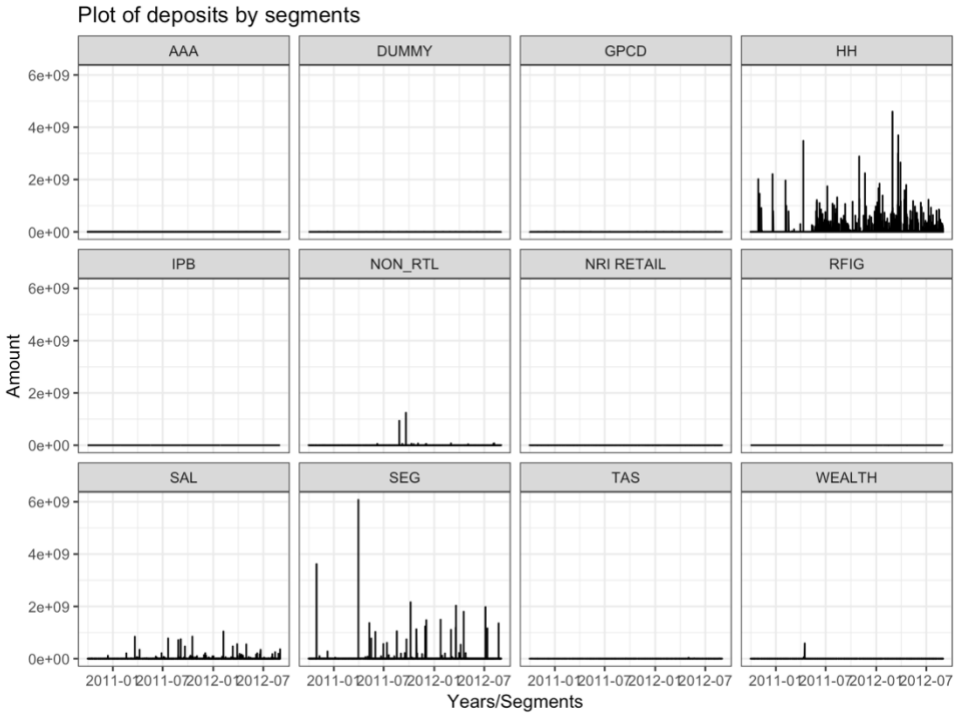


What is seen here is that there are two dominant accounts current and salary. The volume of CA and SA are comparable. Do we need a separate time series analysis for this? Lets come back to this when we check the data at segement level too.

Segment level exploration.

Segment is a level deeper than type of account. There are many segements and our objective of analysis would be if we have a good continious data for all segments for time series analysis. Usually in such cases, we have gaps and erratic data leading to lot of noise. This noise gets evened out when some level of aggregation is applied.





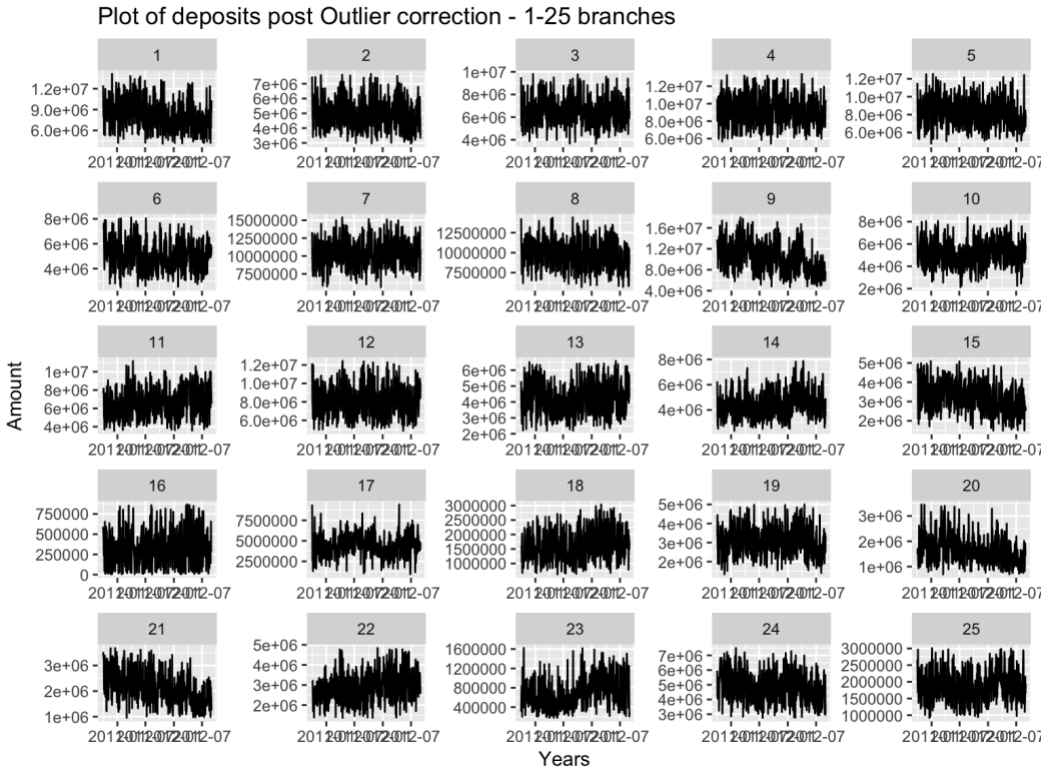
Only two segments are dominant i.e. HH and SEG out of twelve. Hence the data would be sparse for good time series. Also, a business consideration comes into play if data is being used for which purpose.

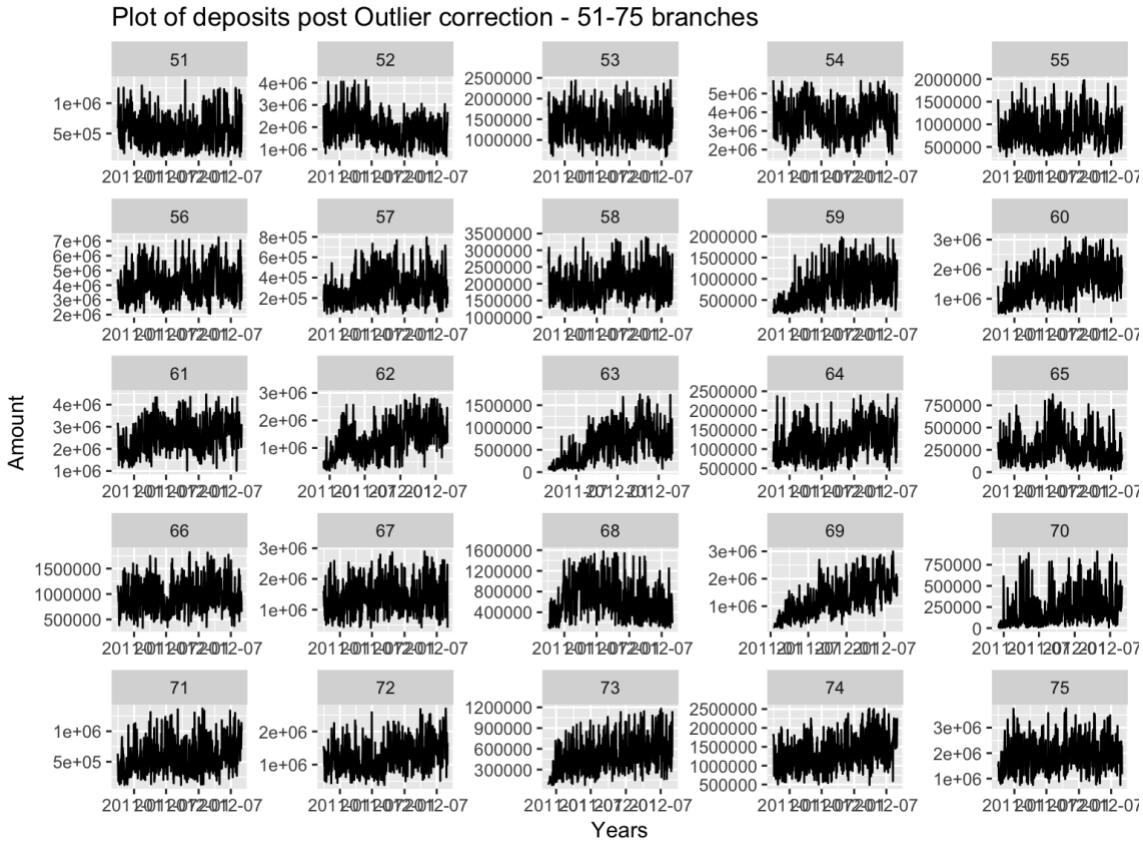
Our current objective being a clustering, it would not make sense to cluster the data which belongs to two different branches.

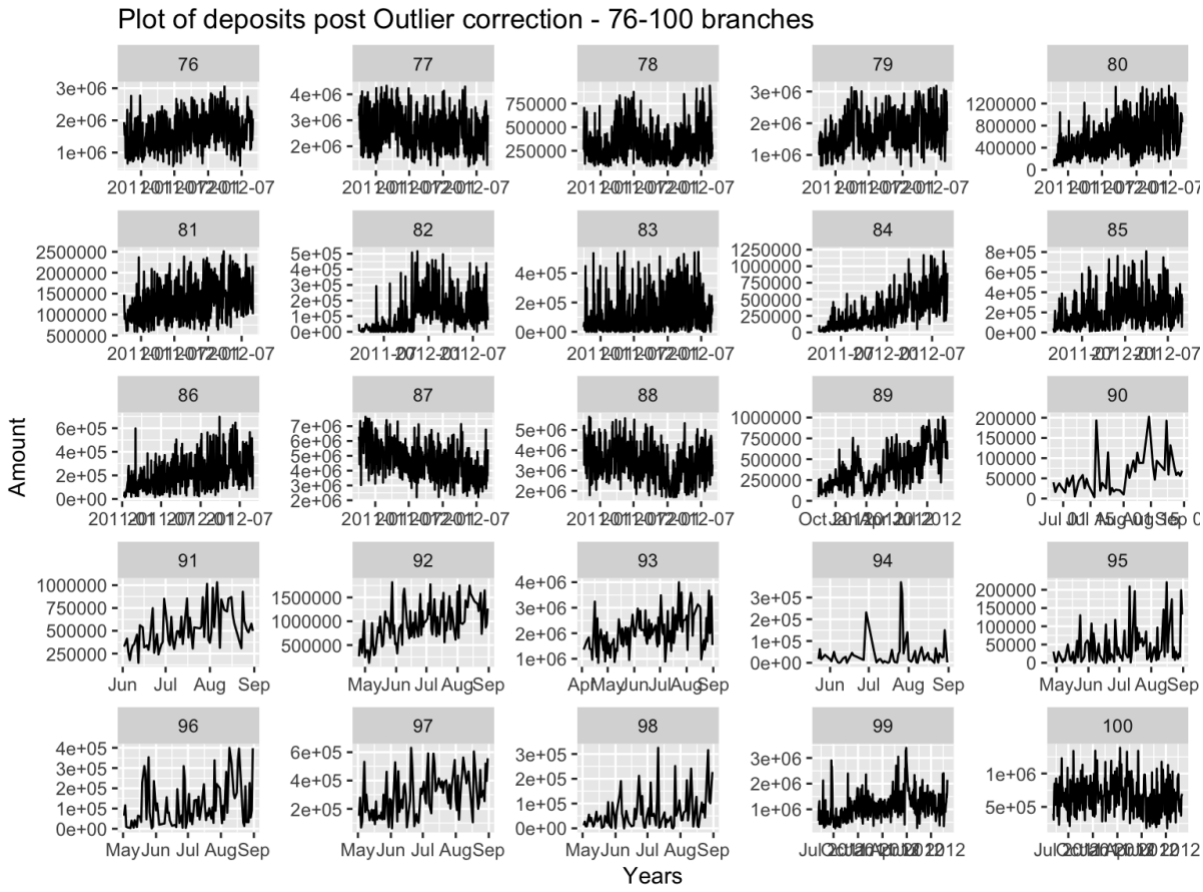
Sub Conclusion: Further analysis can be carried out at the branch id level. Hence we need to aggregate data at the branch level for account and segments at daily level and check for data sanctity.

Outlier corrections - Deposits and Withdrawals

There are lot of outliers which are seen at the branch level the data. We will try to remove them.
Since we want to remove only extreme values, the boundry that has been retained is 0.02 to 0.98 quantile.
outliers are present for both deposit and withdrawals.

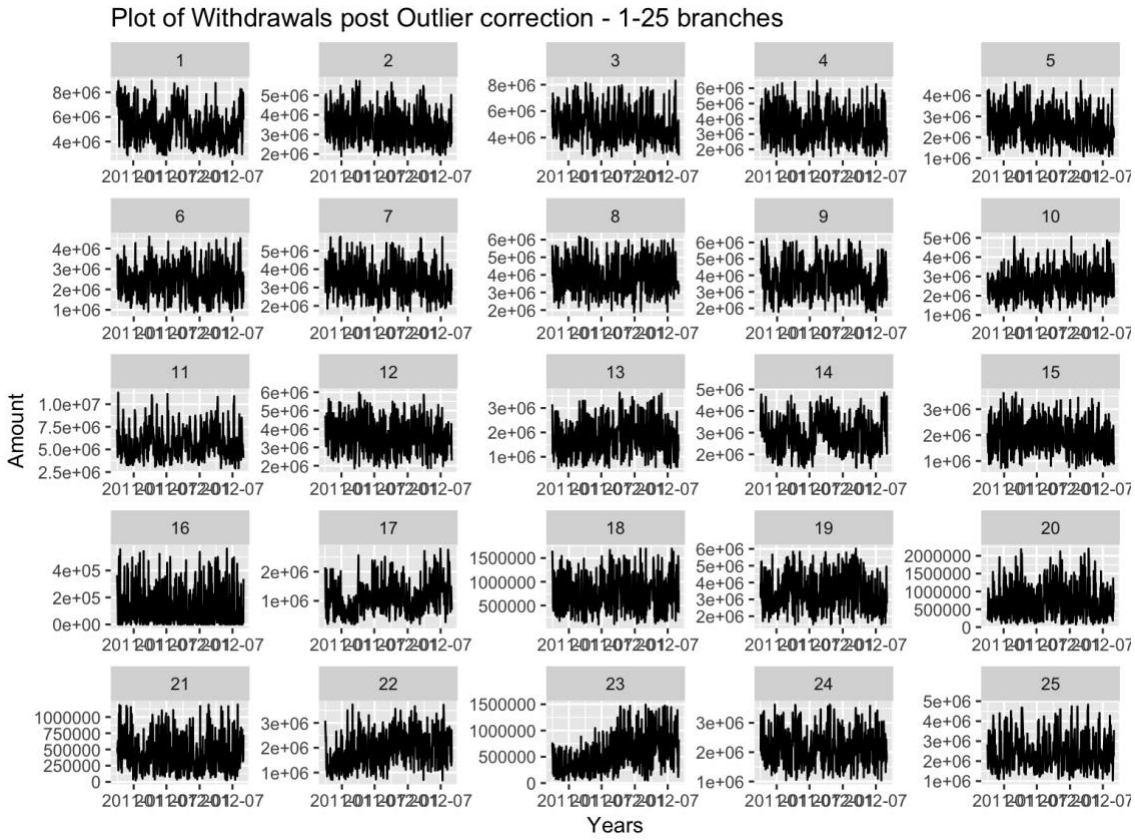

Plot of deposits post Outlier correction - 1-25 branches
Amount
Years

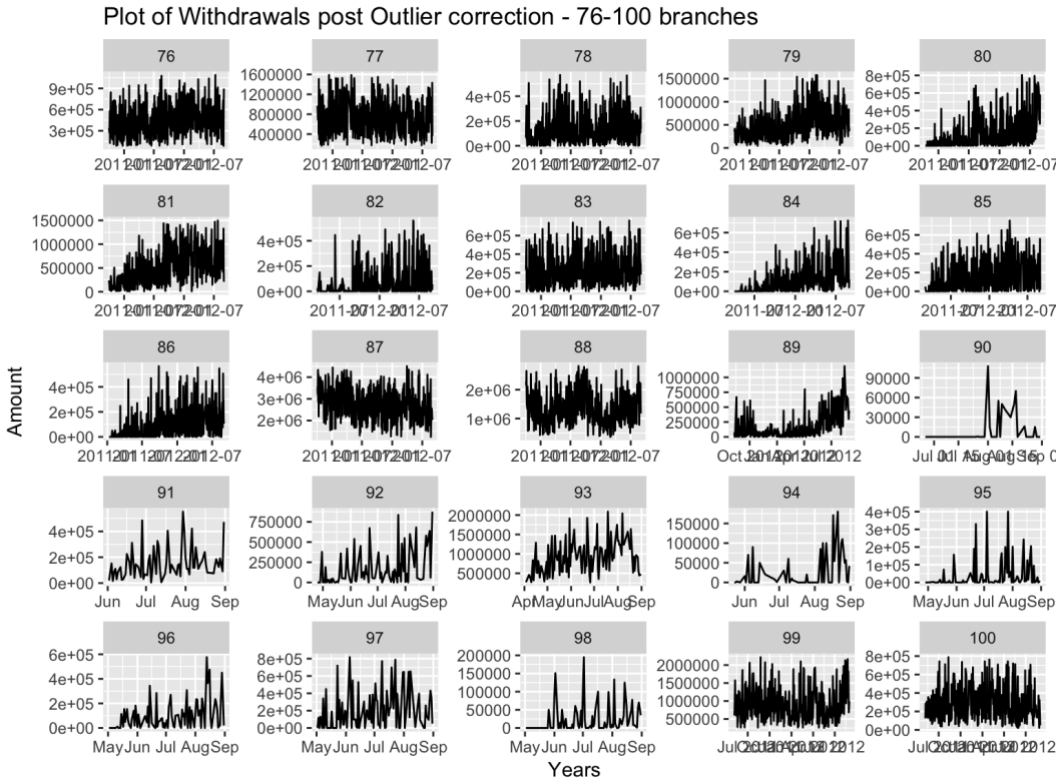

Plot of deposits post Outlier correction - 51-75 branches
Amount
Years

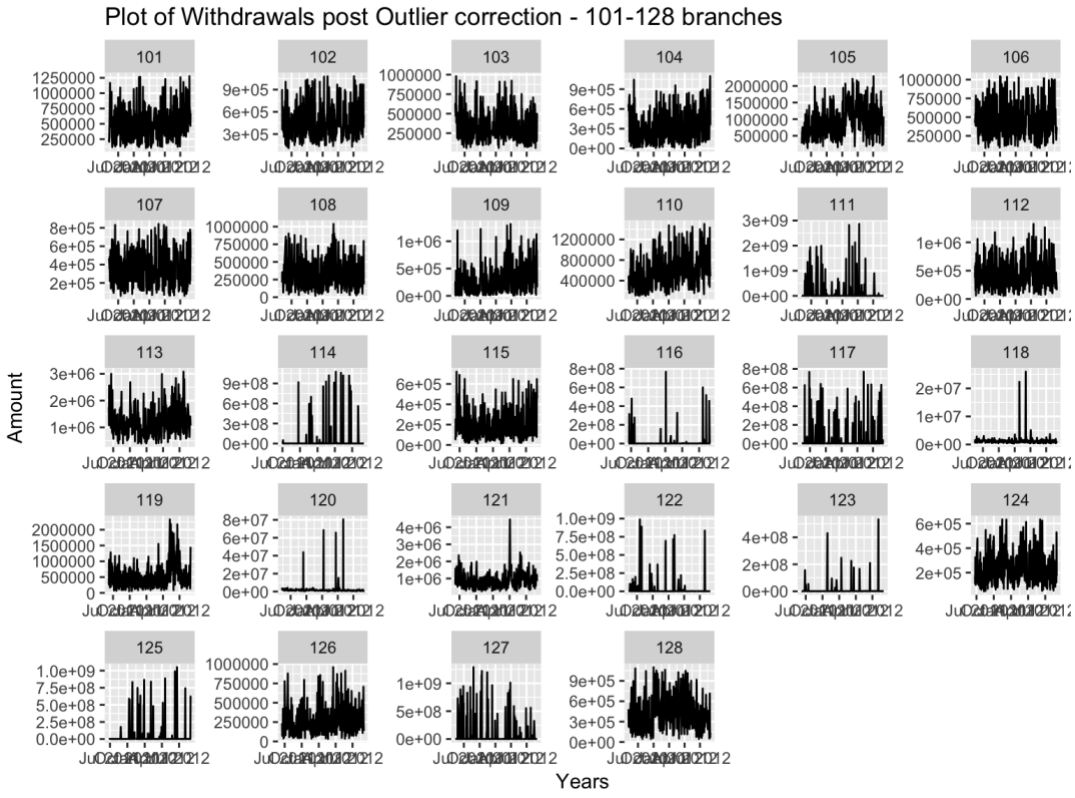


In the later part of the plots with branches do still have outliers but largely some significant transactions as the base level of other data is quite low. For now we will live with these outliers.

Now lets look at the Withdrawals.


Plot of Withdrawals post Outlier correction - 1-25 branches
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Amount
Years





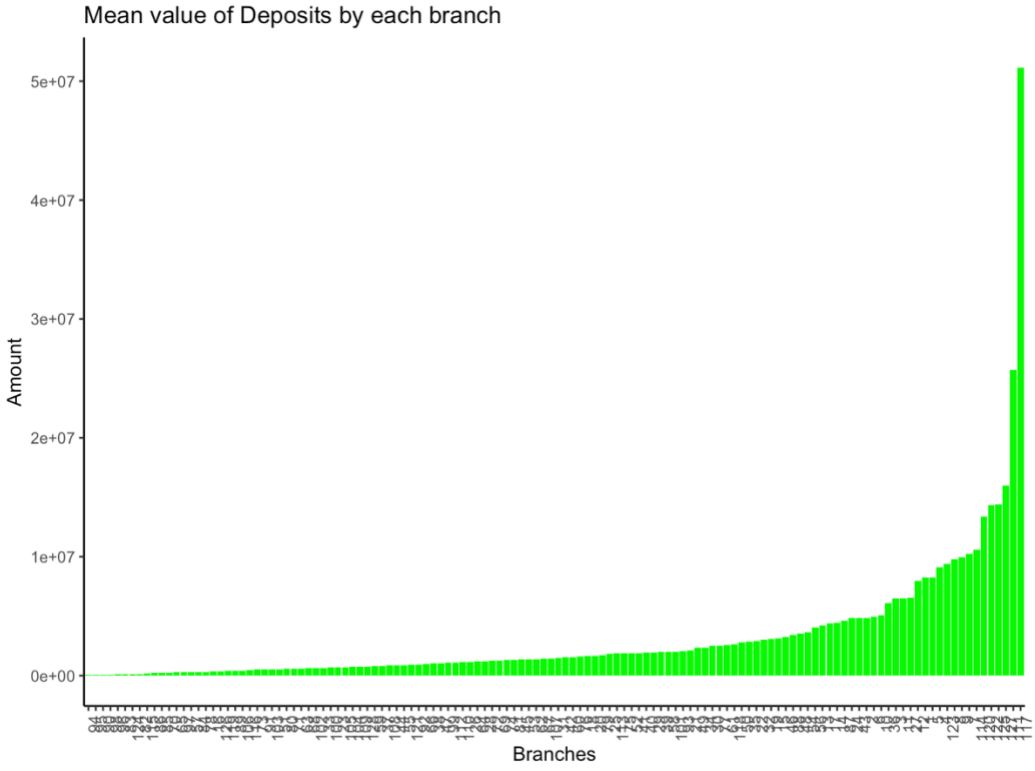
Plotting these distributions we see that they appear quite regular with no unusual peaks or drops. There are few branches which do show intermittent peaks. The base level of data is also low for these branches and few bigger transactions are seen as outliers. From our objective of clustering we can live with same.

Let us evaluate the potential data loss with outlier corrections.
- Number of records lost due to outliers - `4808`
- % of data loss - `7.7882528`

Analysis for Branches by their business volume.

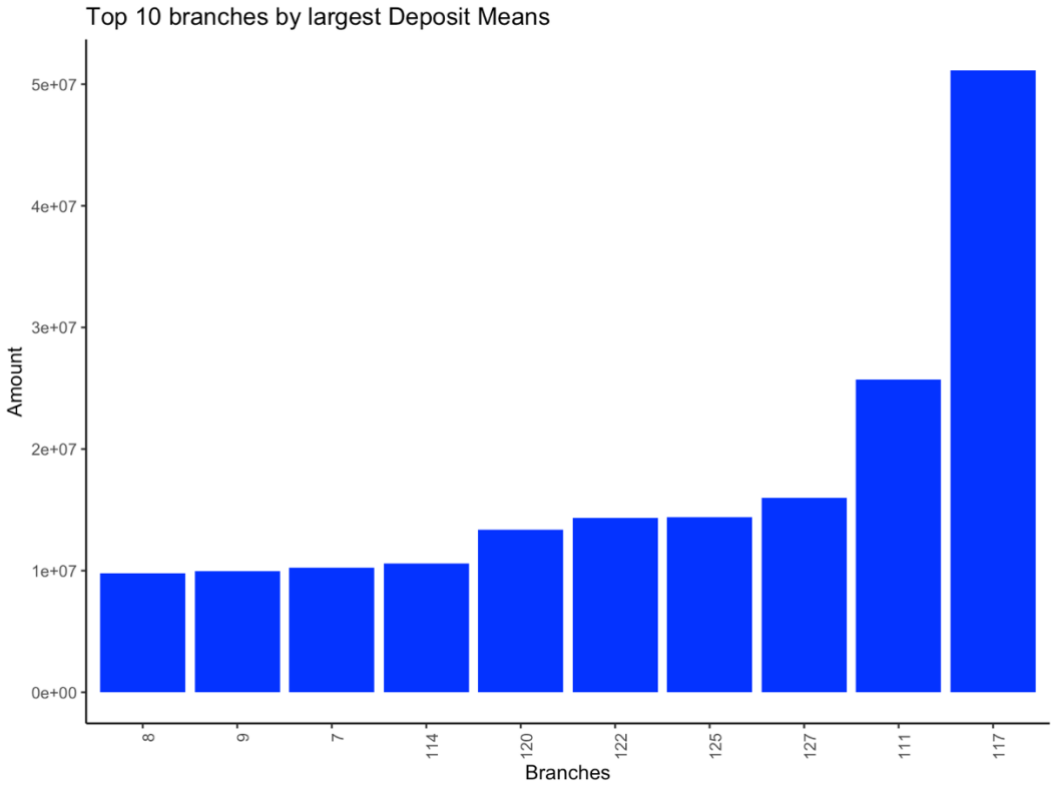
Let us analyze a bit more on the branches by their volume of business and check if there are any significant patterns. One approach is to check the mean(average) transaction sizes over the period of time and rank them. Cumulative might also work but should work but would be similar.

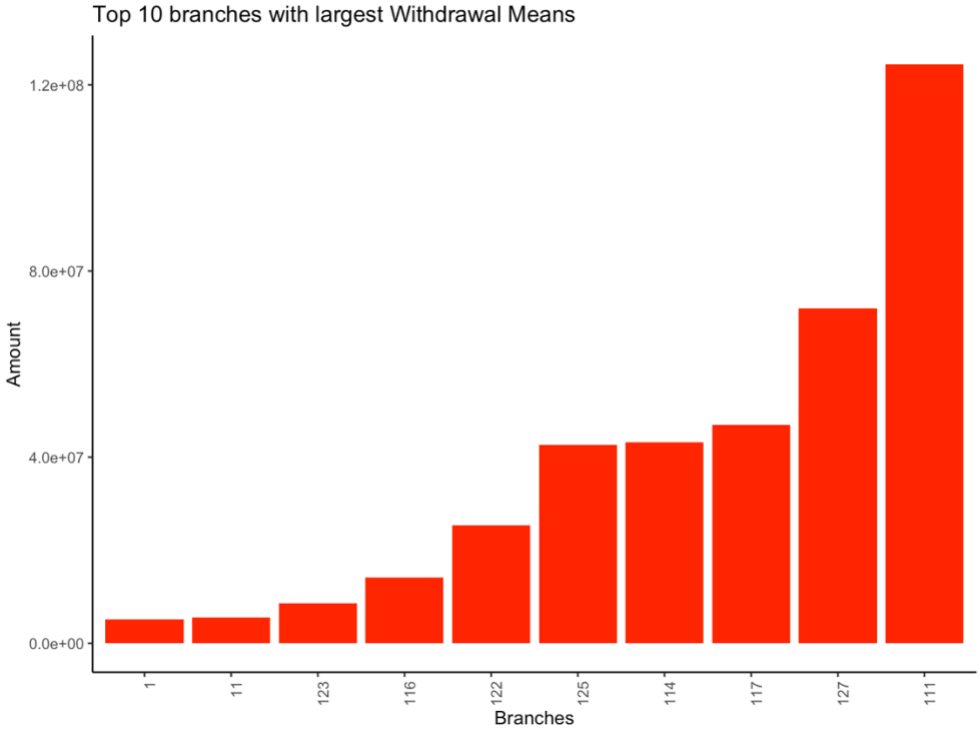
Means of both deposits and withdrawals would provide delta between two at branch level.



The intent of above plots is to show the Pareto effect where few branches are significantly higher than others. Withdrawal has even steeper pattern than deposits. Business significance of this information is very high which will form input to designing of service levels and cash handling capacities.

To have an ease of reading, below we will plot top 10 by means for deposits and withdrawals.





Ok, we got the top then branches for each. However, we notice that **scales** of these two plots is different. Lets get the deposit and withdrawals in one image to check magnitude of difference between two. Since top 10 branches are different for each, we take a union of same.

For few branches, the difference is significantly higher. These branches would be always running short of cash. This information has a logistical value and operations can plan for this behaviour.

Below table shows all the branches where withdrawals are more than deposits. There are 15 such branches.

```
##     SOLID    CDAMean     CWAMean         diff
## 19     19  3064296.7   3407789.3   -343492.57
## 25     25  1852136.4   2372186.2   -520049.87
## 44     44   865477.4   1069025.3   -203547.87
## 55     55   868951.8    988373.1   -119421.34
## 83     83   120661.0    218031.7    -97370.73
## 105   105   709236.8    926617.3   -217380.51
## 106   106   429989.3    463300.1    -33310.84
## 111   111 25712531.4 124389201.9 -98676670.53
## 114   114 10608749.2  43147085.3 -32538336.07
## 116   116   465299.3  14095110.9 -13629811.62
## 119   119   387886.5    521021.0   -133134.50
## 122   122 14347768.2  25280322.4 -10932554.26
## 124   124   121641.4    236580.6   -114939.20
## 125   125 14378797.8  42653101.4 -28274303.68
## 127   127 15972091.6  71935656.3 -55963564.68
```

This concludes the data exploratory part of the project. Following information would be key to further analysis

a. Aggregate data at branch level is best suited for clustering analysis as account and segment has lot of noise.
b. There are significant differences in the deposits and withdrawals at branch level.

Clustering

With data **cleaned** with outliers we can proceed with clustering. We will try to find clusters of branches based on deposit and withdrawals.

since we do not have adequate information on potential number of clusters, we will need to find good number to start with and then use it with visual inspection to ascertain it.

Clustering deposits

Lets look at clustering based on deposits made in branches. To determine *number of clusters* we will deploy 3 difference methods

a. best number of clusters across methods offered by package clValid
b. *Scree plot* based on the package TSrepr
c. plot *hierarchical cluster* based on different distance types
d. **clValid** - Here we will check with 3 to 8 clusters as a range for input along with hierarchical and partitional clustering technique. This function provides score with best number of cluster to go with each clustering method. One of the key requirement of this package is that timeseries should be in a row. Hence we use function "spread" to transpose it. This function also fill the value as NA for missing date. It also makes each time series of uniform length. it will take the max/min date in the data frame and use it for all-time series.

Parameters - data is in the matrix form. - range of clusters - Methods of clustering - Validation measures - check various options (internal, stability, biological) - stability is chosen here as it seemed to have better response to data. It removes and replaces data to form various clusters. - method - it is the distance calculated between clusters. "ward" seems to provide with distinct clusters This function takes significant time. One would require to run through different iterations to arrive at stable numbers

```
##
## Clustering Methods:
##  hierarchical pam
##
## Cluster sizes:
##  3 4 5 6 7 8
##
## Validation Measures:
##                                 3            4            5            6
7            8
##
## hierarchical APN        0.1375       0.1601       0.2073       0.2318
0.2432       0.2646
```

```
##              AD         0.7803       0.7611       0.7528       0.7450
0.7388       0.7307

##              ADM   964017.9277  931220.9618 1417287.9354 2672248.0388 30
45588.9263 3006169.8670

##              FOM  8435784.1297 8406145.7708 8355071.6170 8291236.3697 81
01489.7375 8015002.3959

## pam          APN        0.0221       0.0129       0.0156       0.0326
0.0271       0.0156

##              AD         0.7847       0.7621       0.7498       0.7406
0.7270       0.7138

##              ADM   154664.4945  103043.4695  122768.5668  258808.4011  2
52527.5262  108097.0281

##              FOM  8689851.5160 8608095.3512 8590544.5690 8571521.6537 85
27764.1662 8478047.9033

##

## Optimal Scores:

##

##     Score        Method       Clusters

## APN       0.0129 pam          4

## AD        0.7138 pam          8

## ADM  103043.4695 pam          4

## FOM 8015002.3959 hierarchical 8
```

The choice of clusters that we get here are **4 and 8**. If hierarchical is preference then it shows only 8 clusters. To validate further we will use another method if the choice of number of clusters is appropriate.

TSrepr - this library focuses on specifically time series and its representations. Function used is repr_matrix which will normalize the data and one can provide the frequency of the data. Here the data is at daily level and we would like to take complete range in evaluation.

Once data is converted using this function, we can use *K - mediods* algorithm to extract typical profiles. Here we provide range of 3-10 clusters since we had a been suggested 8 clusters in our earlier analysis. Internally the cluster have to be evaluated using some index and we use Davies-Bouldin index. Other evaluation measures are Dunn or Silhouette. *https://en.wikipedia.org/wiki/Davies–Bouldin_index*

we will use two methods DB index and Silhouette to check number of clusters.

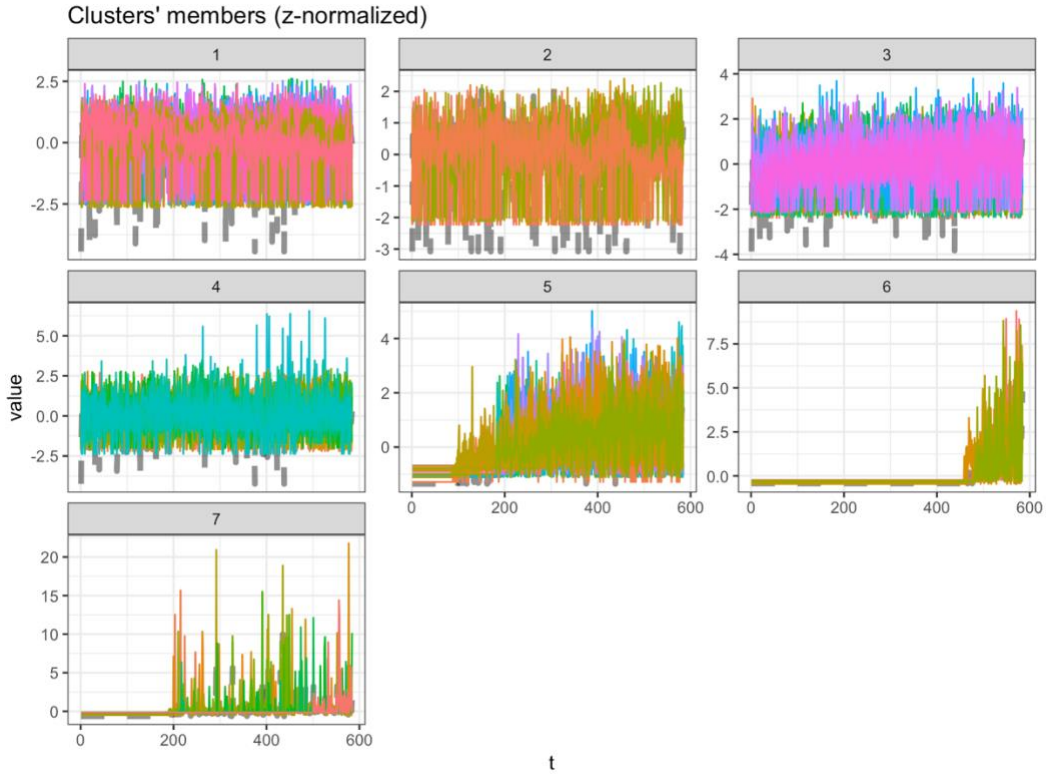
With Davis Boulding index, we can choose between 6, 7 and 8. Though slope remains constant post 6 clusters, the value of index is quite significant till 8 clusters. Hence with this 8 can be chosen.

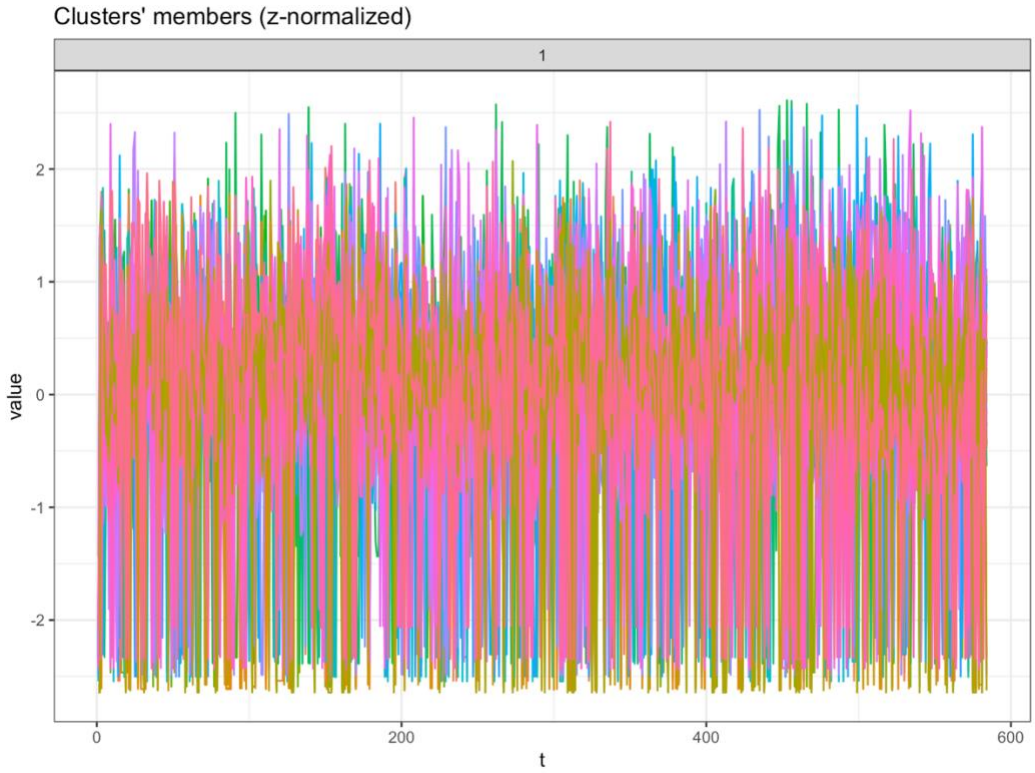
With Silhouette too best value seen in 7. Hence with this approach we can keep 7 and proceed to check how the clusters look like in heiarchical plot.

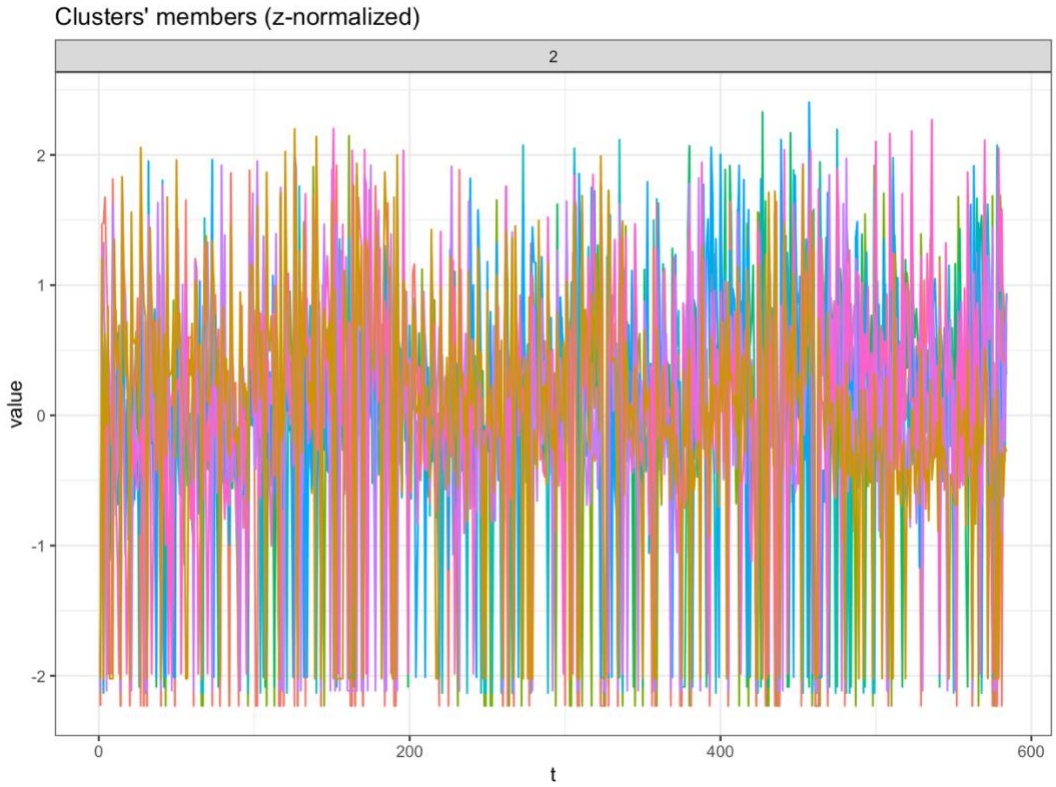
Hierarchical clustering - SBD

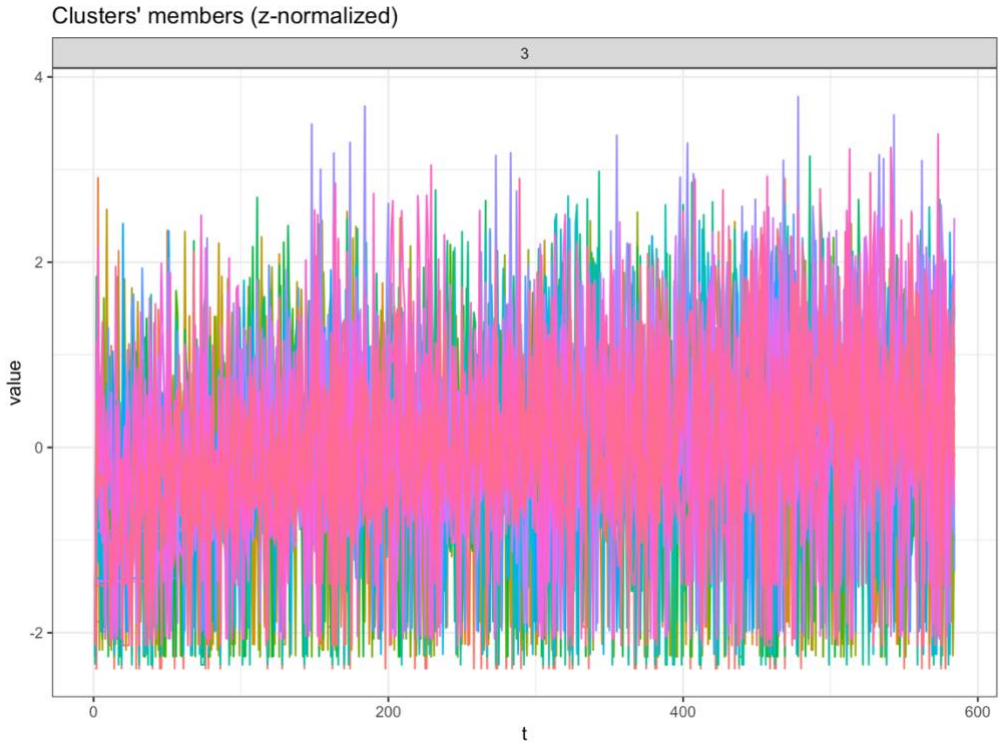
For hierarchical clustering we use **dtwclust** library with function tsclust. There are many distance measures available for parameter distance. Here we use **SBD(shape based distance) and DTW (Dynamic Time Wrap)** for trying different output that suits data.
DTW is Euclidean squared distance between two given vectors.

From the plot of cluster 7 clusters do seem logical here visually. Let us try to see how each cluster looks like..

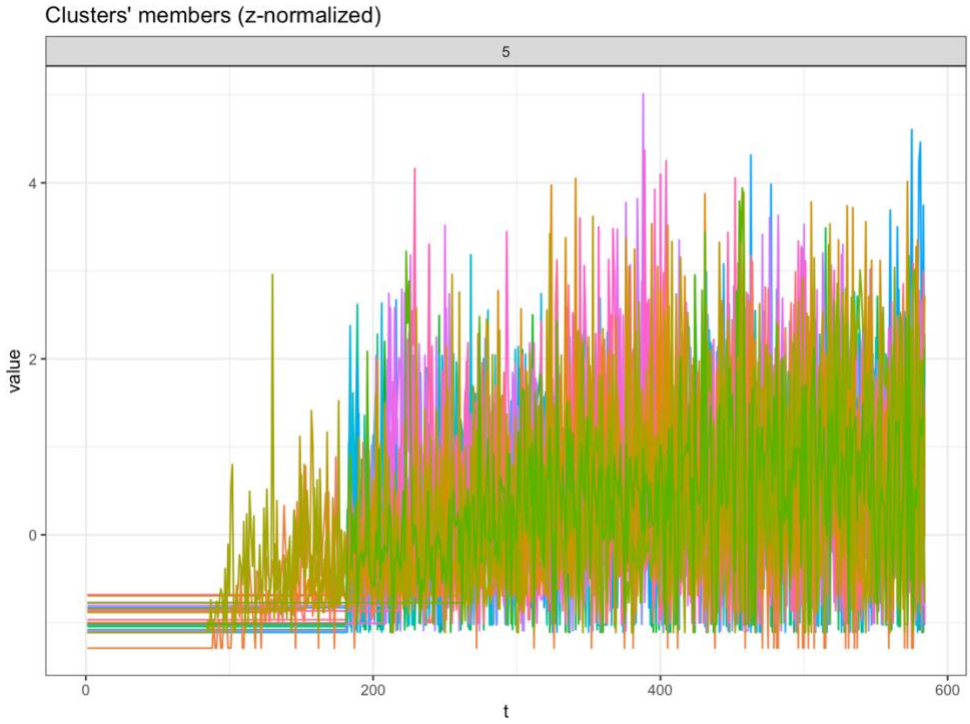


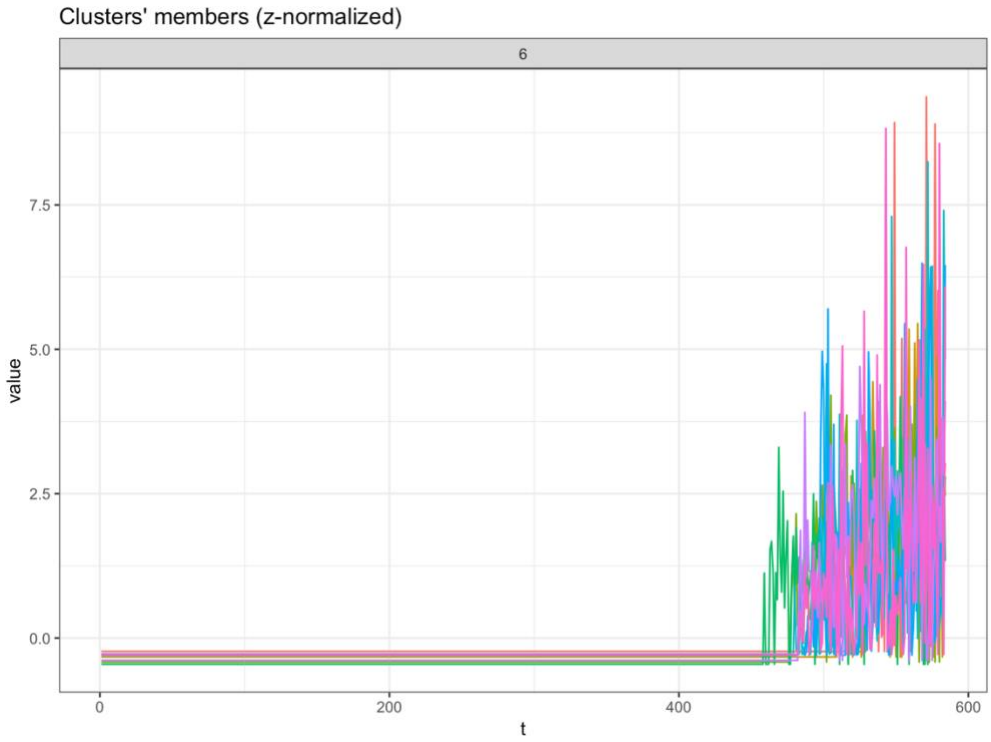

Clusters' members (z-normalized)
1
value
t
2
1
0
-1
-2
0
200
400
600


Clusters' members (z-normalized)
2
value
t
2
1
0
-1
-2
0
200
400
600


Clusters' members (z-normalized)
3
4
2
0
-2
value
0
200
400
600
t


Clusters' members (z-normalized)
4
5.0
2.5
0.0
-2.5
value
0
200
400
600
t





Clustering has largely happend based on shapes of series. There are different patterns in each cluster.

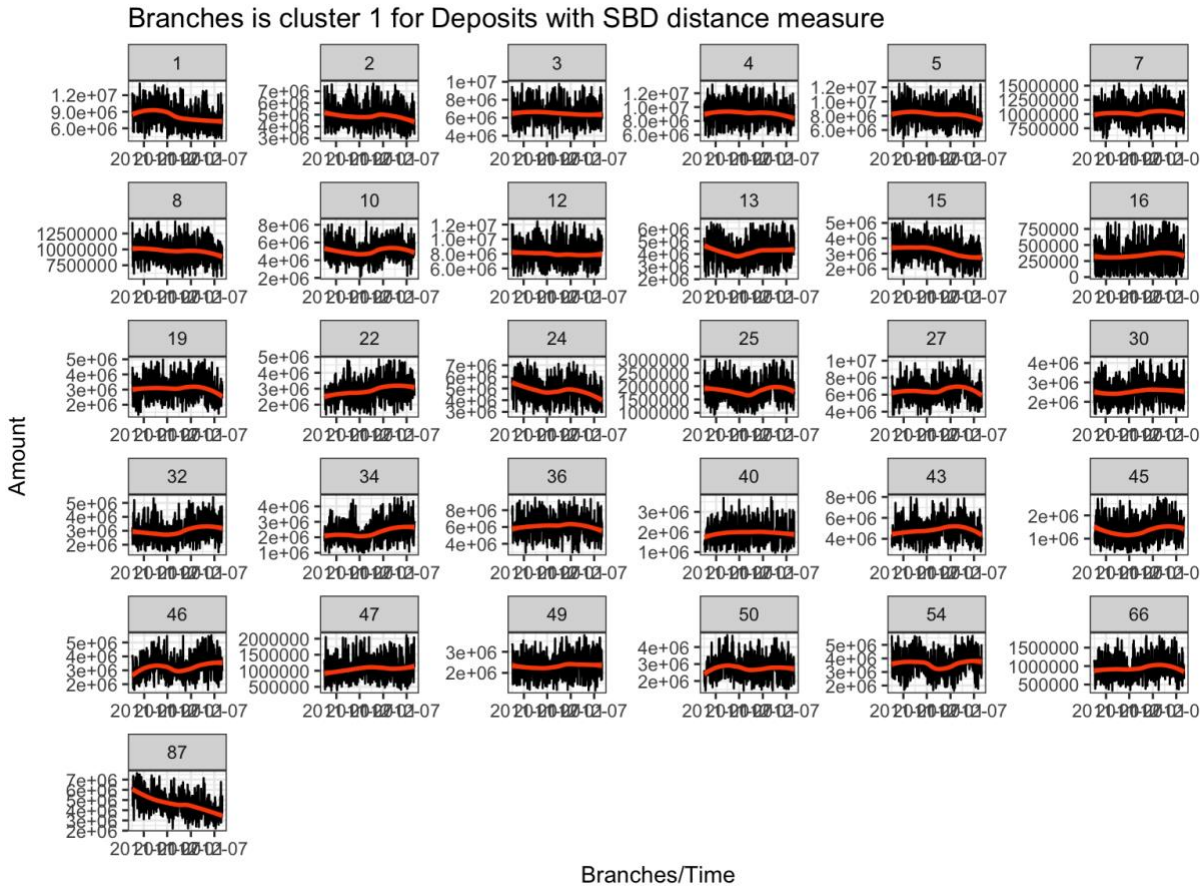
```
##   Var1 Freq
```

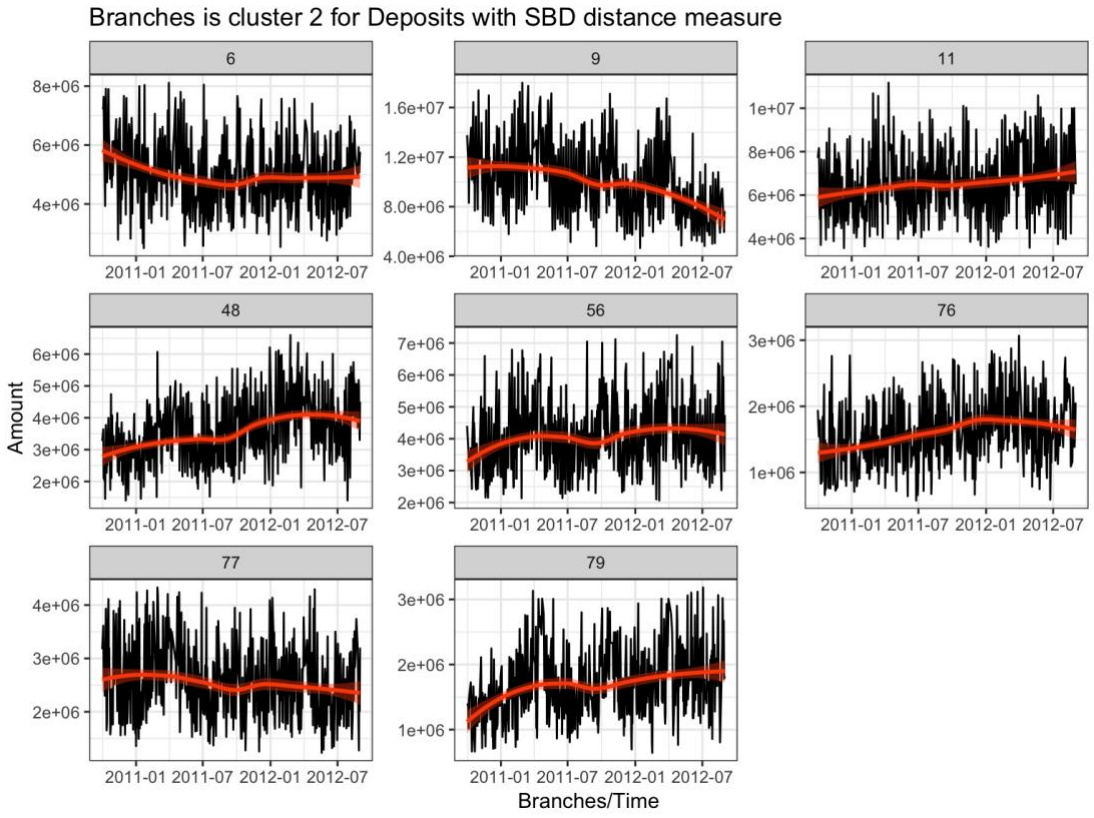
```
## 1     1    31
## 2     2     8
## 3     3    27
## 4     4    17
## 5     5    25
## 6     6     8
## 7     7    12
```

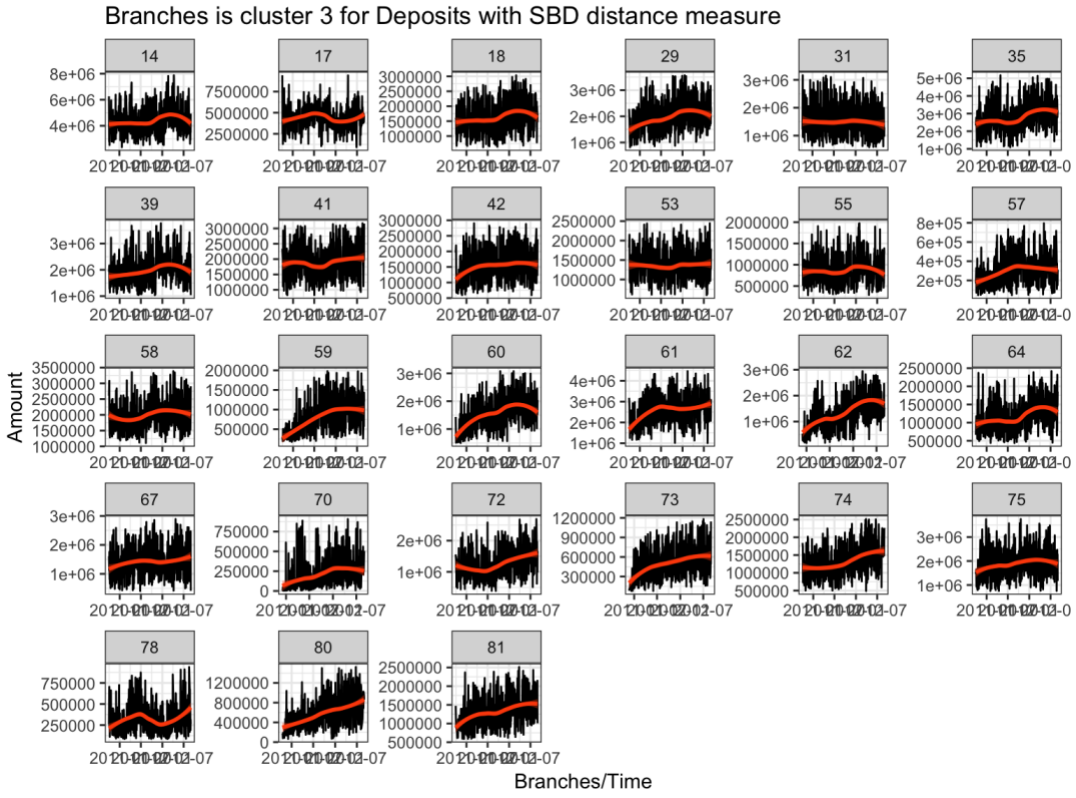
Above table shows the number of branches in different clusters. lets us plot all the clusters with their branches to check if the shapes resemble of cluster members

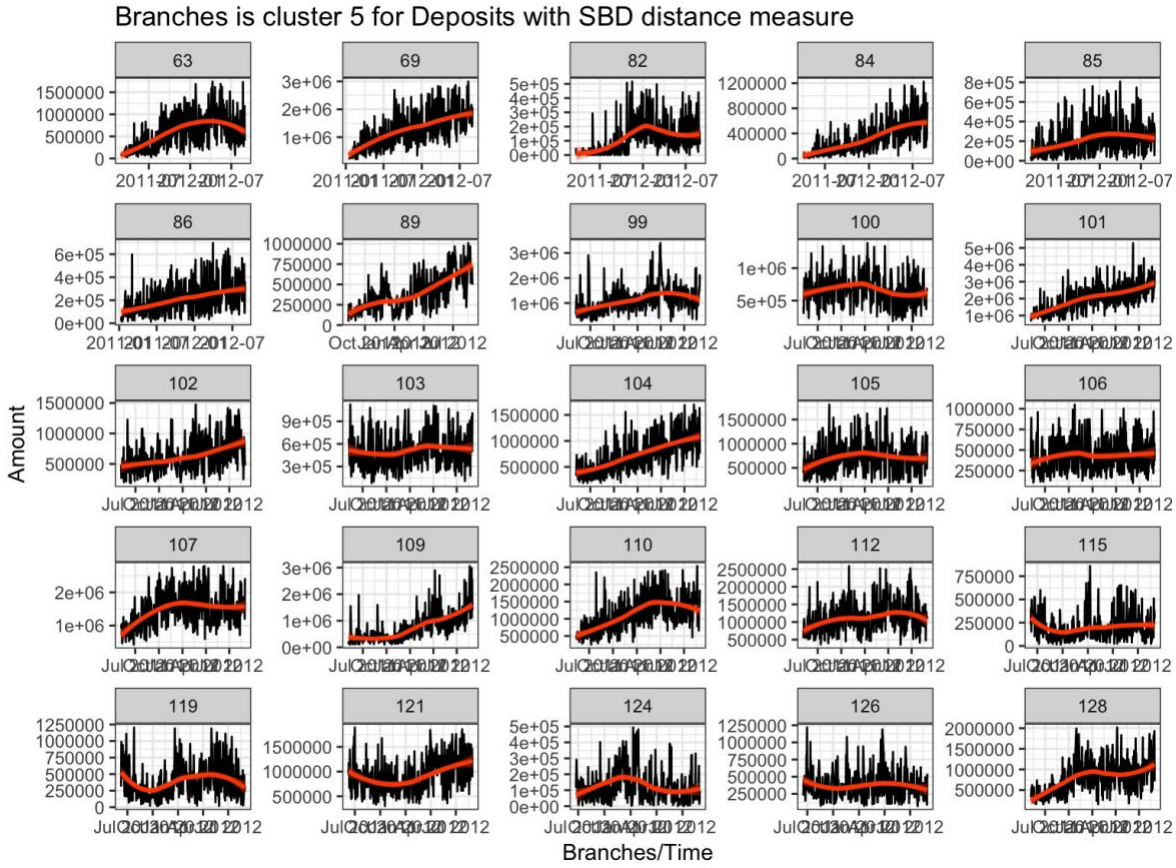
Plot of clusters - SBD

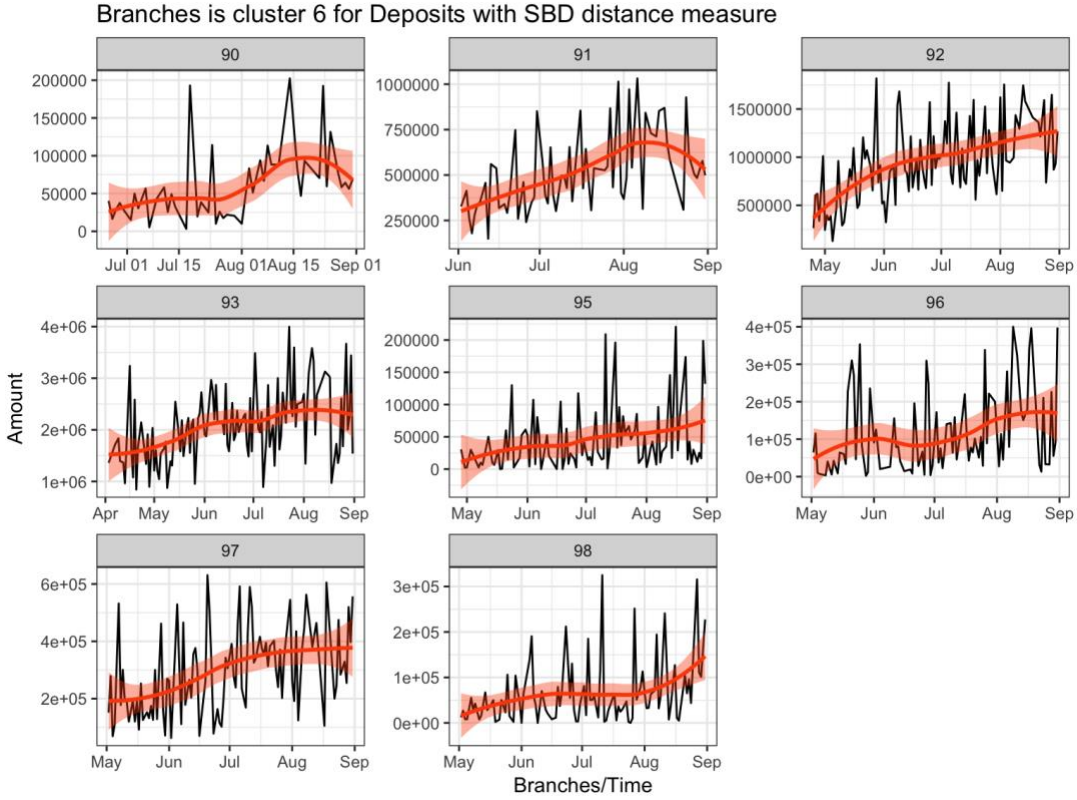
Based on the consolidated graphs seen earlier for all the clusters, we will plot the distribution of each branch for deposits based on the SBD distance matrix.

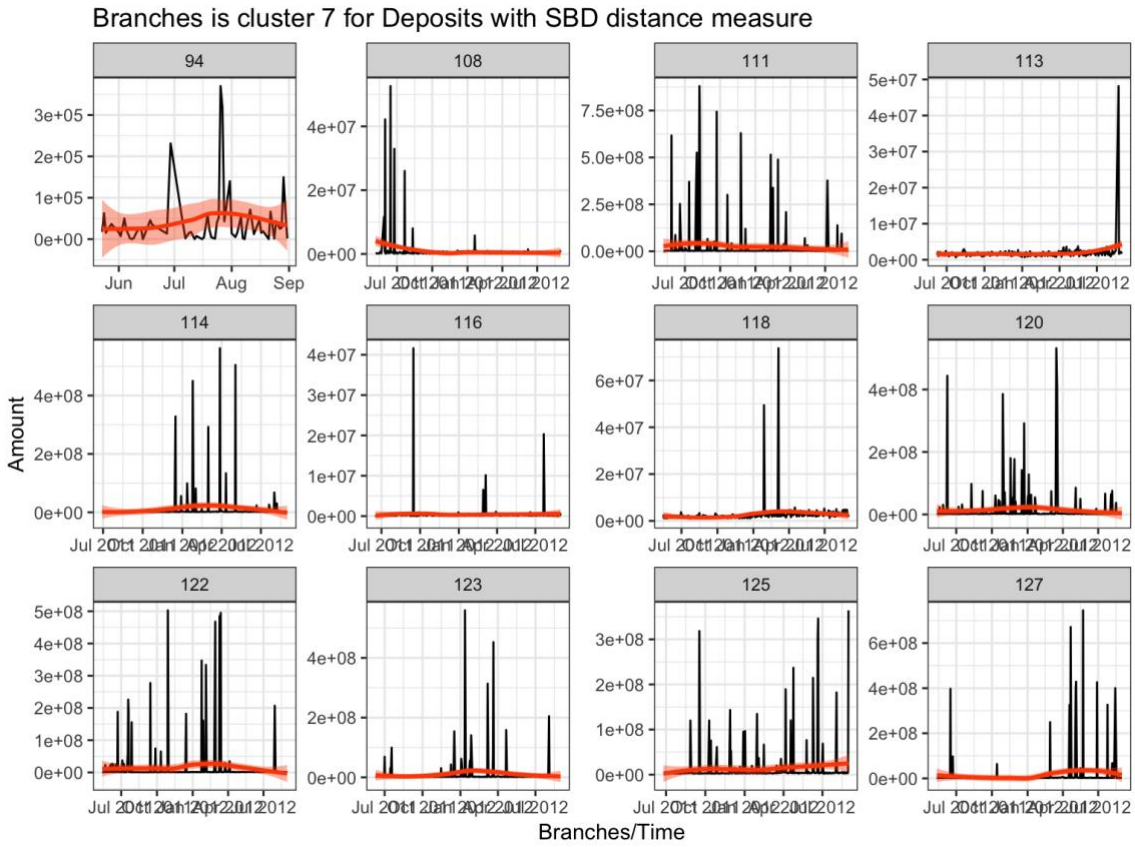



Branches is cluster 2 for Deposits with SBD distance measure
6
9
11
48
56
76
77
79
Amount
2011-01 2011-07 2012-01 2012-07
Branches/Time


Branches is cluster 3 for Deposits with SBD distance measure
14
17
18
29
31
35
39
41
42
53
55
57
58
59
60
61
62
64
67
70
72
73
74
75
78
80
81
Amount
Branches/Time


Branches is cluster 5 for Deposits with SBD distance measure
63
69
82
84
85
86
89
99
100
101
102
103
104
105
106
107
109
110
112
115
119
121
124
126
128
Amount
Branches/Time

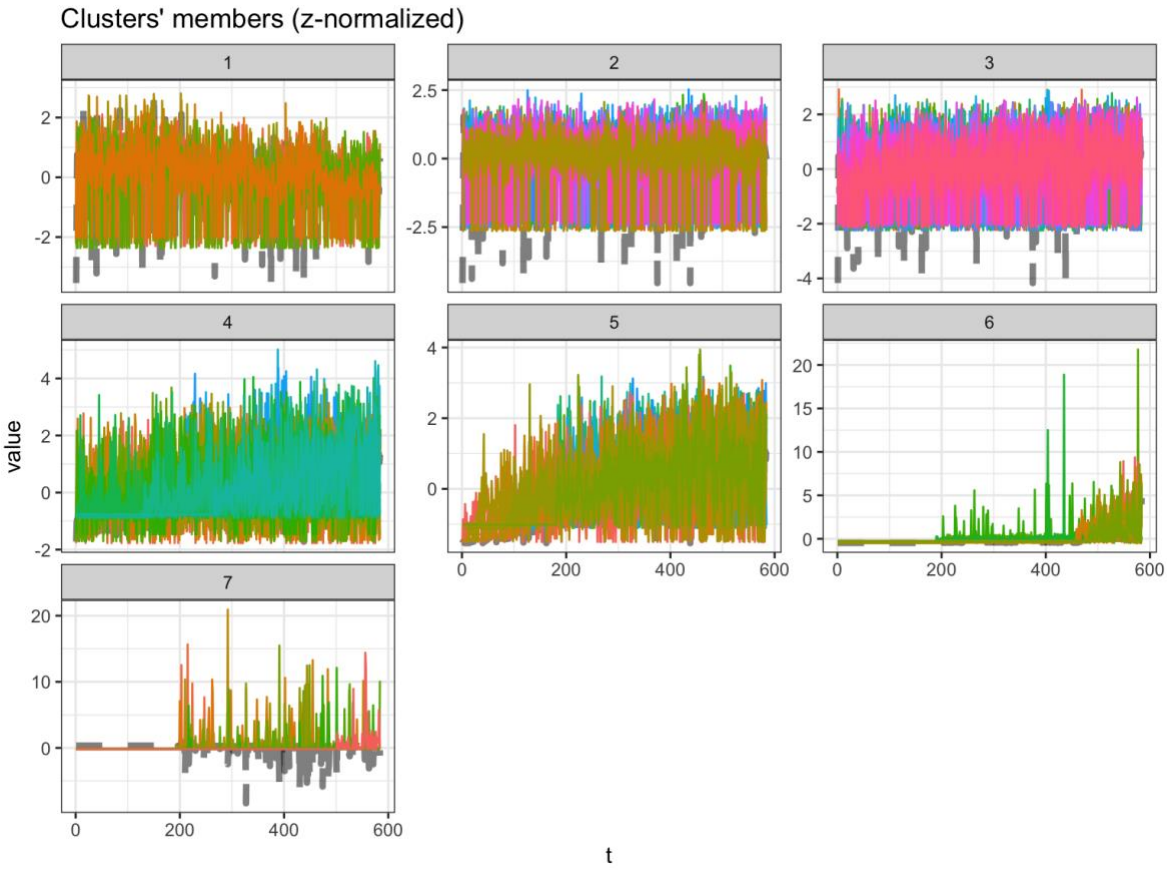

Branches is cluster 6 for Deposits with SBD distance measure
90
91
92
93
95
96
97
98
Amount
Jul 01
Jul 15
Aug 01
Aug 15
Sep 01
Jun
Jul
Aug
Sep
May
Apr
Branches/Time

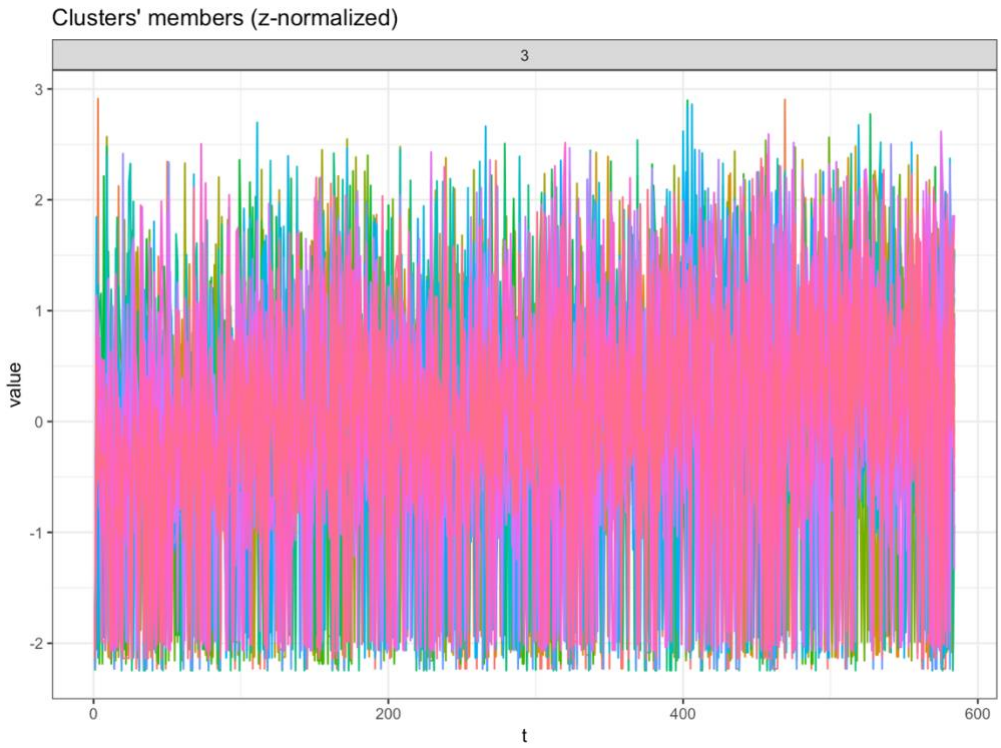


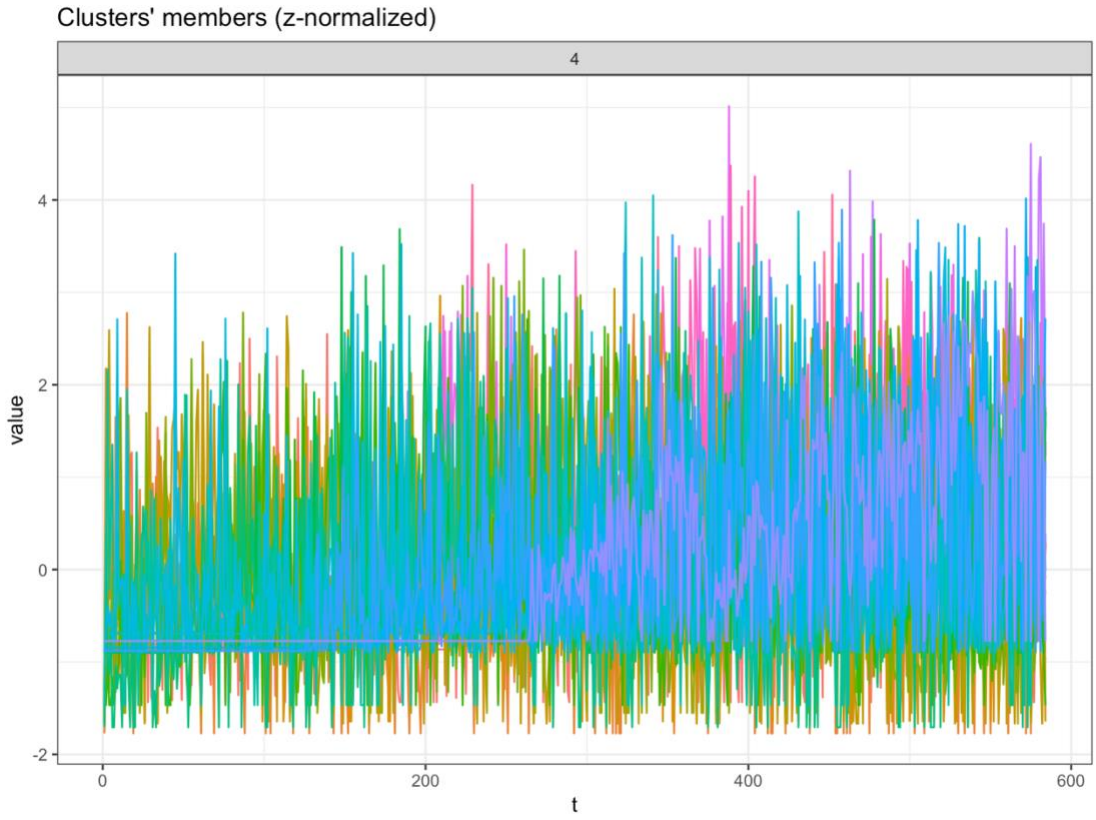
Hierarchical clustering - DTW

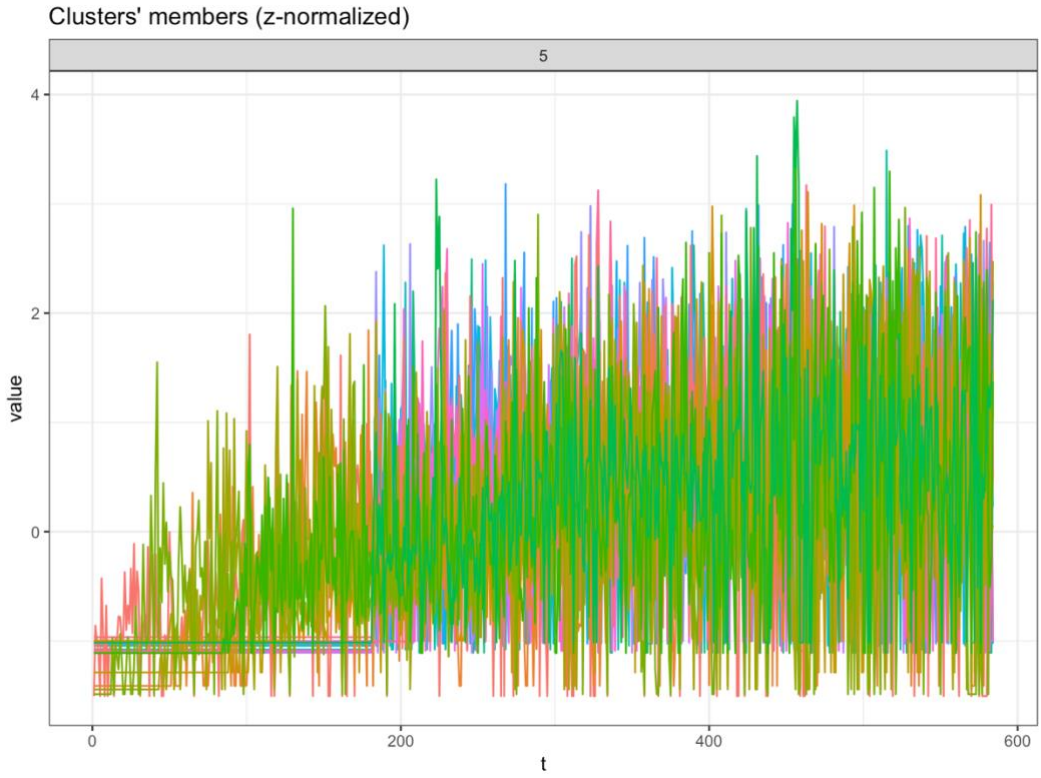
Let us use the DTW distance metric and plot cluster and try to see if we get something different than SBD plot.

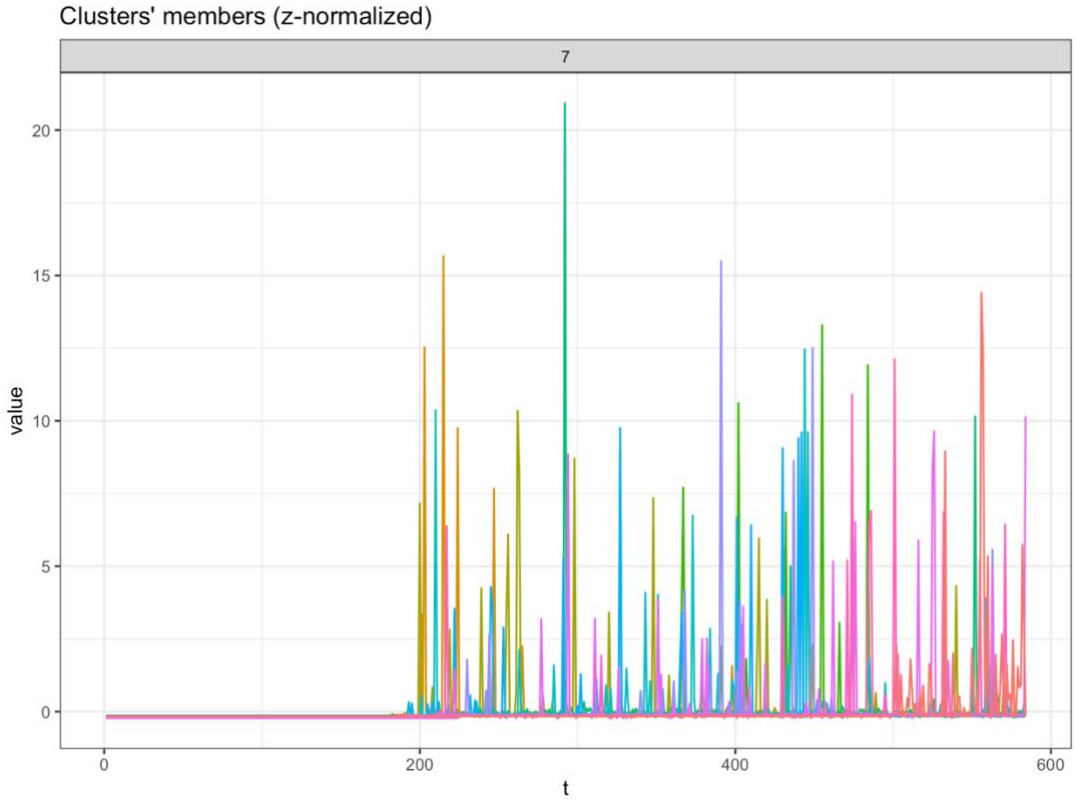
From above plot we can see that 7 clusters seem logical based on visual cutoff. The rectangles would usually try to capture the branches and the height combination for segregation. Clusters 4th and 5th though can be clubed into one, it would be interesting to observe what are the characteristics difference between two.


Clusters' members (z-normalized)
1
2
3
4
5
6
7
value
t
0
200
400
600


Clusters' members (z-normalized)
3
value
t
3
2
1
0
-1
-2
0
200
400
600


Clusters' members (z-normalized)
4
value
t
4
2
0
-2
0
200
400
600


Clusters' members (z-normalized)
5
4
2
0
value
0
200
400
600
t


Clusters' members (z-normalized)
7
20
15
10
5
0
value
0
200
400
600
t

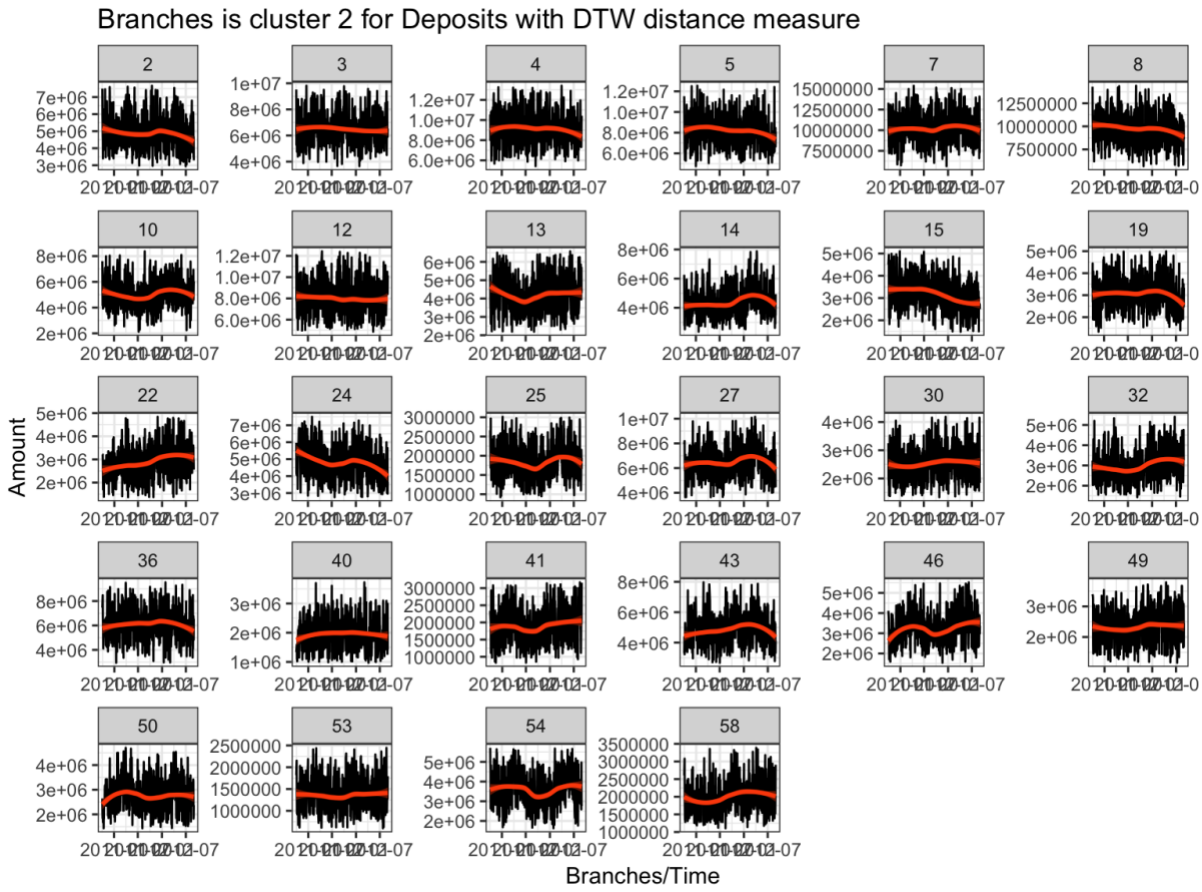
```
##   Var1 Freq
## 1    1   31
## 2    2    8
## 3    3   27
## 4    4   17
## 5    5   25
## 6    6    8
## 7    7   12
```

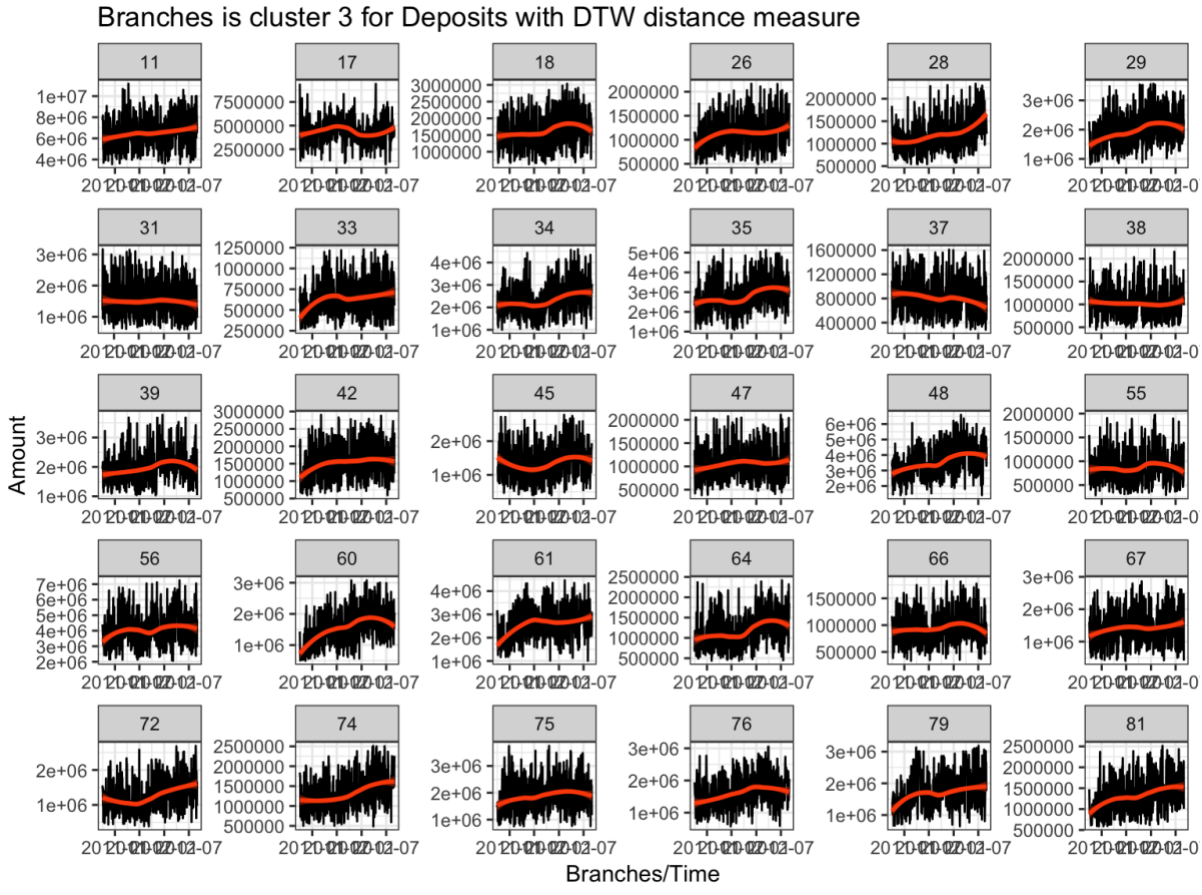
```
##   Var1 Freq
## 1    1    9
## 2    2   28
## 3    3   30
## 4    4   20
## 5    5   20
## 6    6   11
## 7    7   10
```

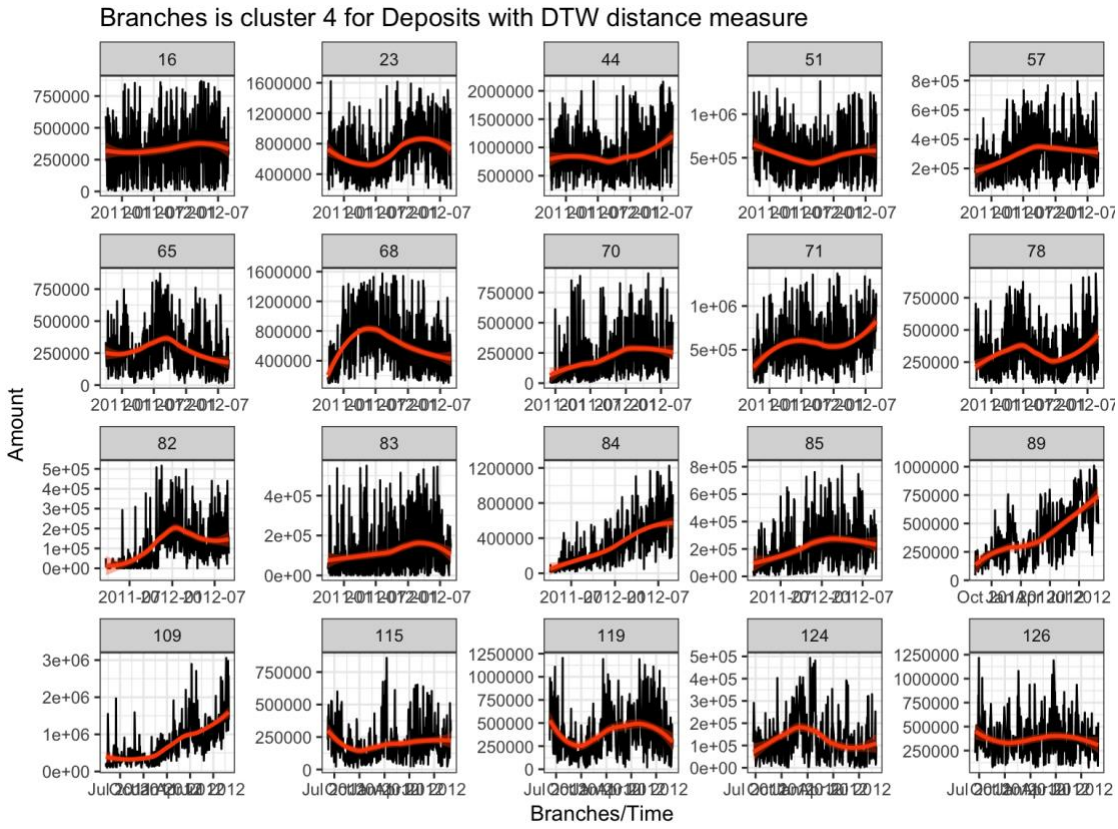
Above tables shows the distribution of branches amongst different clusters

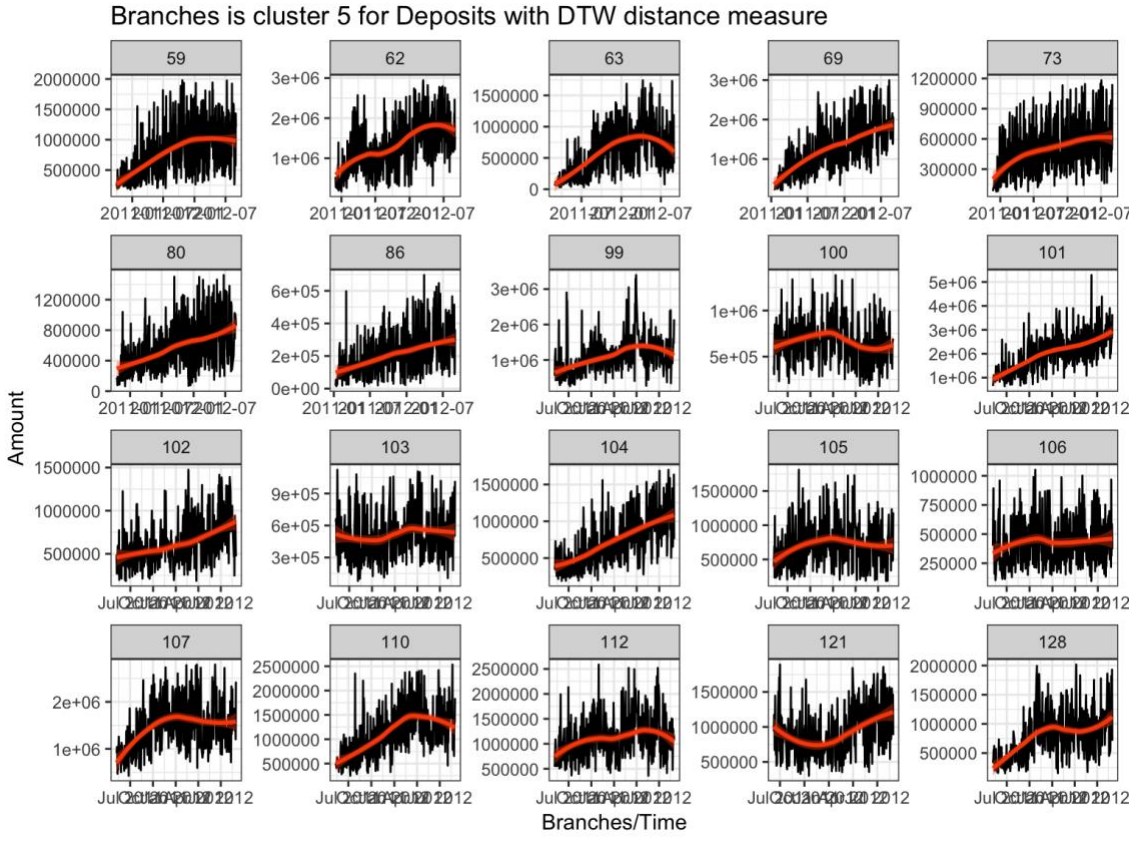
Plot of clusters - DTW

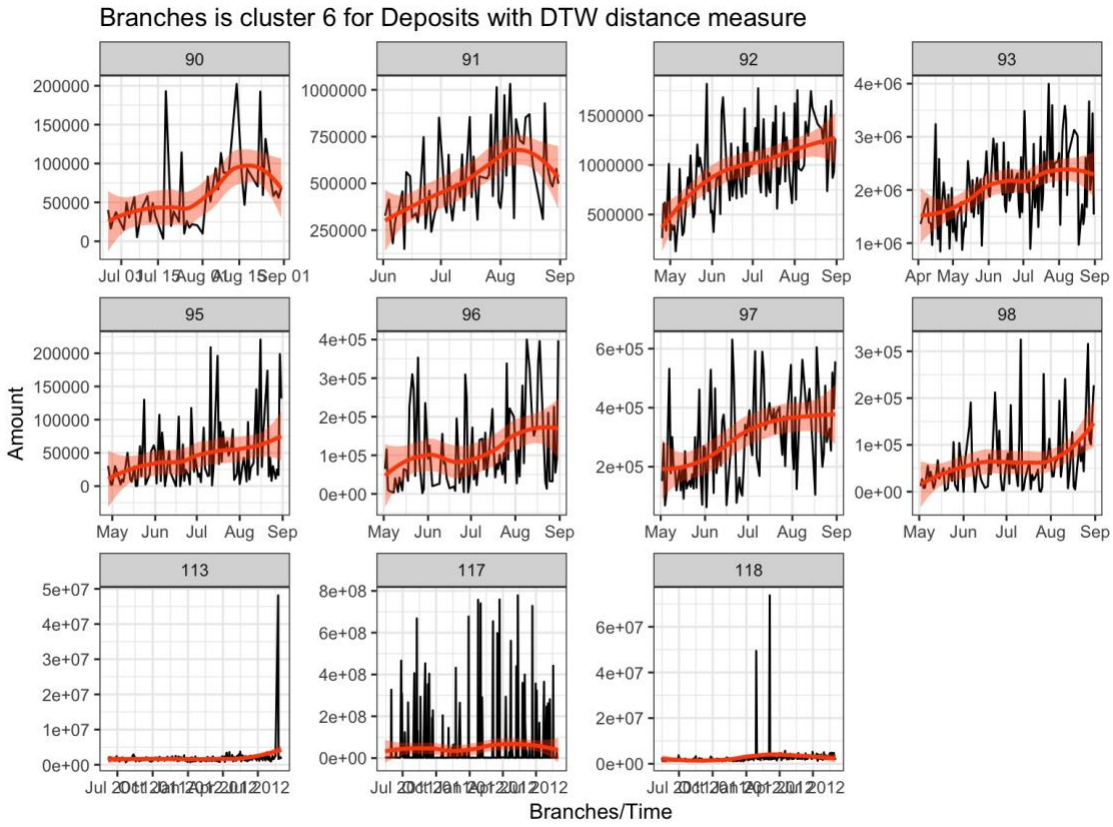
Let’s plot the individual branches shown in the clusters and plot their behaviour.


Branches is cluster 2 for Deposits with DTW distance measure
Amount
Branches/Time


Branches is cluster 3 for Deposits with DTW distance measure
Amount
Branches/Time


Branches is cluster 4 for Deposits with DTW distance measure
16
23
44
51
57
65
68
70
71
78
82
83
84
85
89
109
115
119
124
126
Amount
Branches/Time


Branches is cluster 5 for Deposits with DTW distance measure
59
62
63
69
73
80
86
99
100
101
102
103
104
105
106
107
110
112
121
128
Amount
Branches/Time



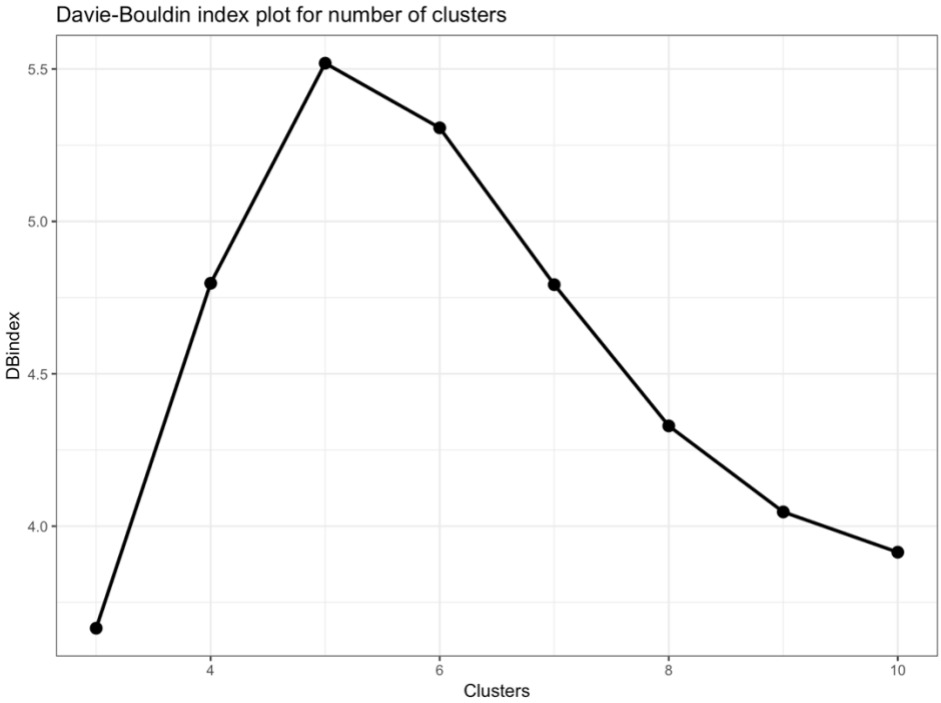
Above plots are for DTW distance metric. Visually inspecting, we get better segregation of branches based on their behaviour. We will go with DTW metric for cluster distribution

Clustering for withdrawals CWA

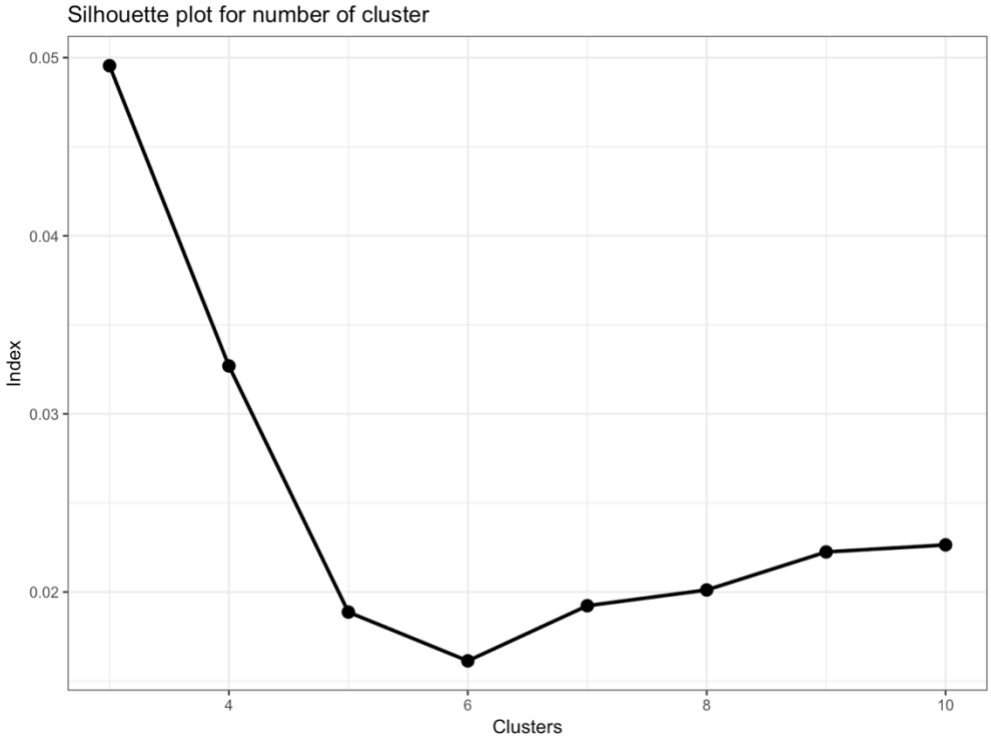
Similar to deposits, we will explore the plots for withdrawals. The steps would be repeated as those of deposits analysis.

```
##
## Clustering Methods:
##  hierarchical pam
##
## Cluster sizes:
##  4 5 6 7 8
##
## Validation Measures:
##                                  4        5        6        7        8
##
## hierarchical Connectivity  109.5270 114.6810 129.4567 129.7484 150.0262
##              Dunn            0.5542   0.5542   0.5542   0.5542   0.5542
##              Silhouette      0.0678   0.0497   0.0486   0.0505   0.0501
## pam          Connectivity   96.8298 148.0147 165.1591 167.4060 166.6821
##              Dunn            0.5108   0.4479   0.4479   0.4479   0.4479
##              Silhouette      0.0553   0.0241   0.0221   0.0228   0.0256
##
## Optimal Scores:
##
##              Score   Method       Clusters
## Connectivity 96.8298 pam          4
## Dunn          0.5542 hierarchical 4
## Silhouette    0.0678 hierarchical 4
```

Optimal Scores section is a summary of min/max scores. Cluster we have been suggested either **4 or 8** clusters with method PAM.



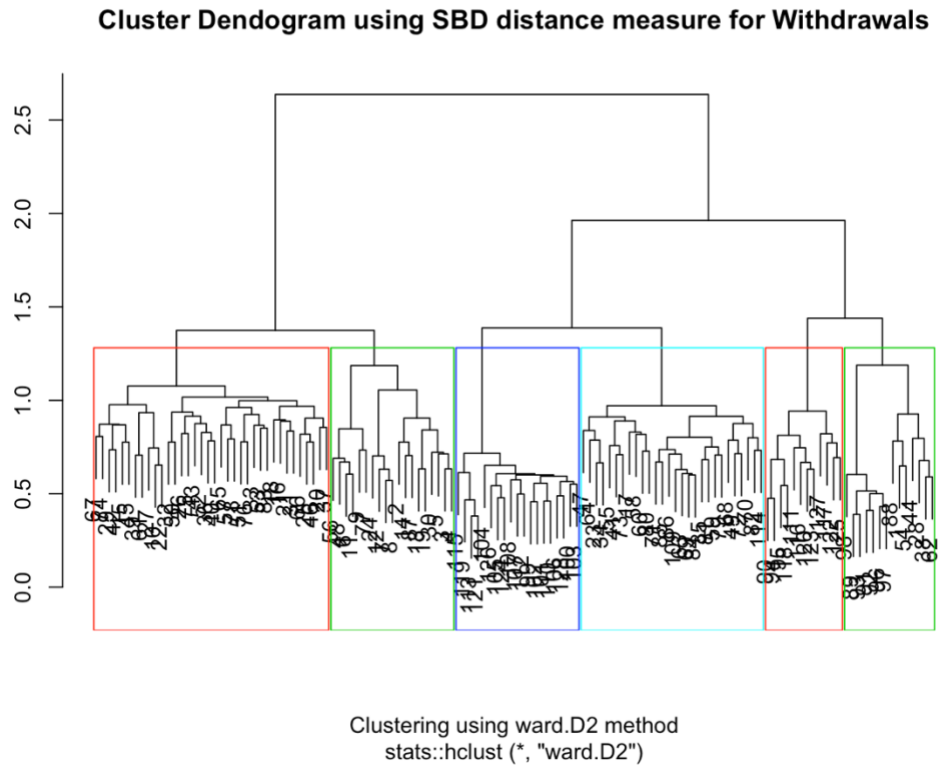
Above plot indicates that any number of clusters can be chosen between **6 annd 8** as till 8 there is significant index value that is carried.



Using Silhouette method there is distinct indication that **6** is the ideal number of clusters to have.

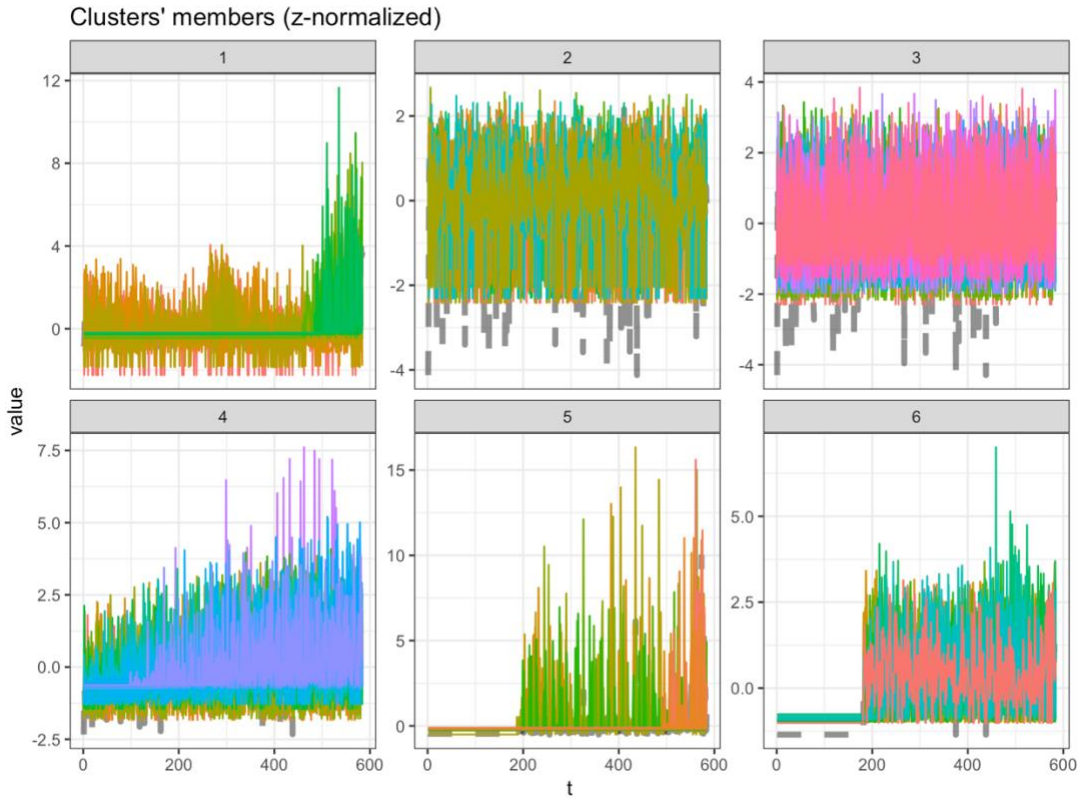
Hierarchical clustering - SBD

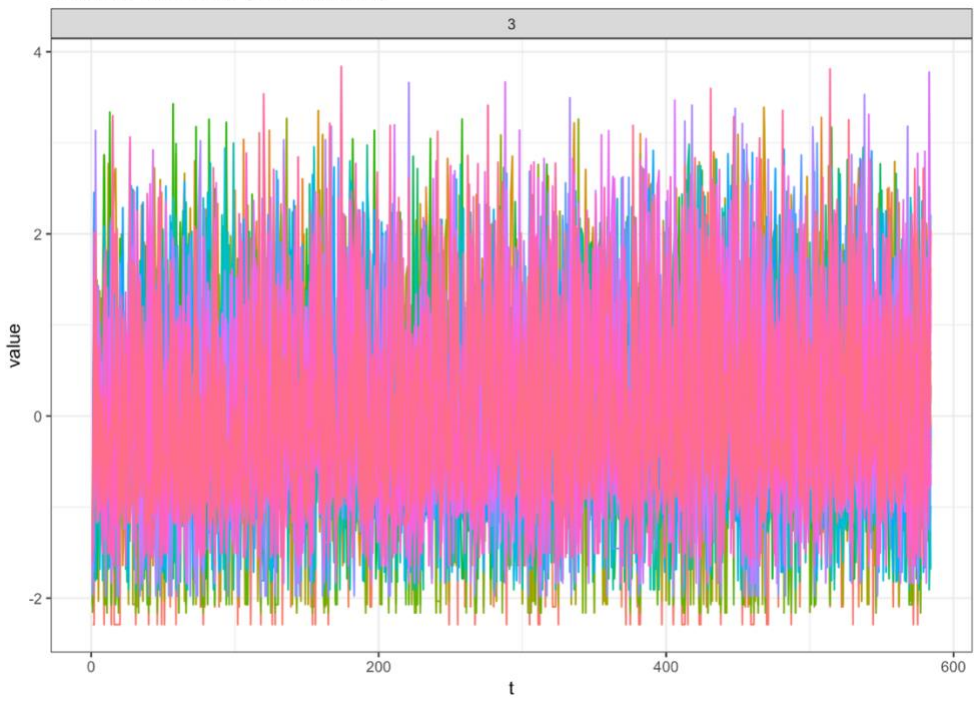
Lets visualize hierarchical plot now for SBD distance matrix with 6 and check if we get any other result for number of clusters.

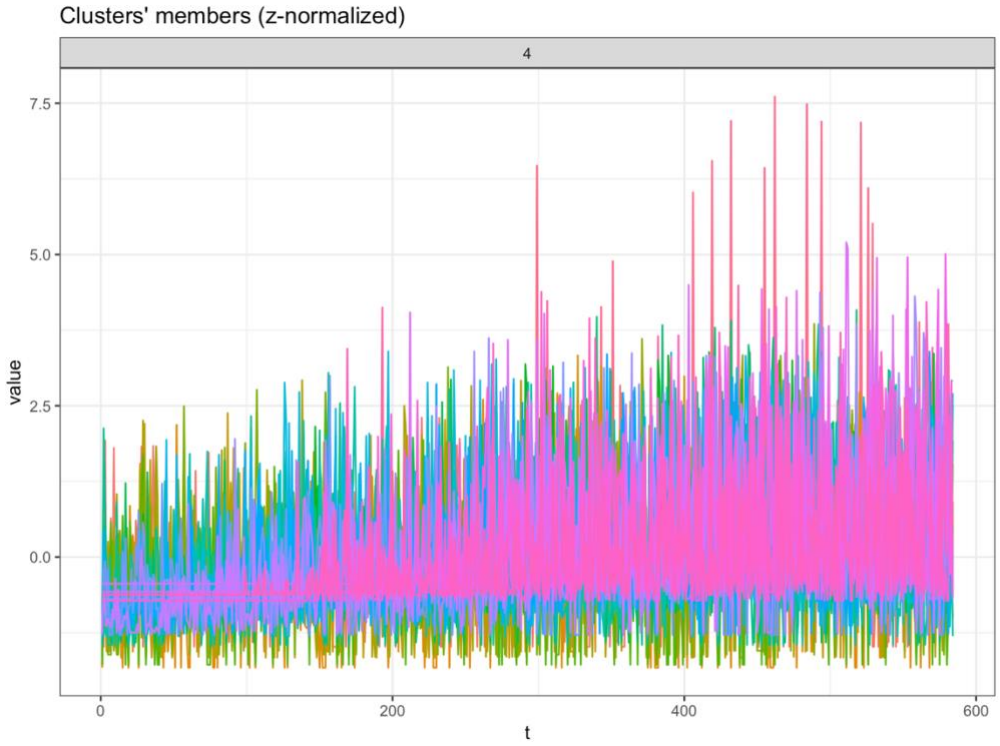


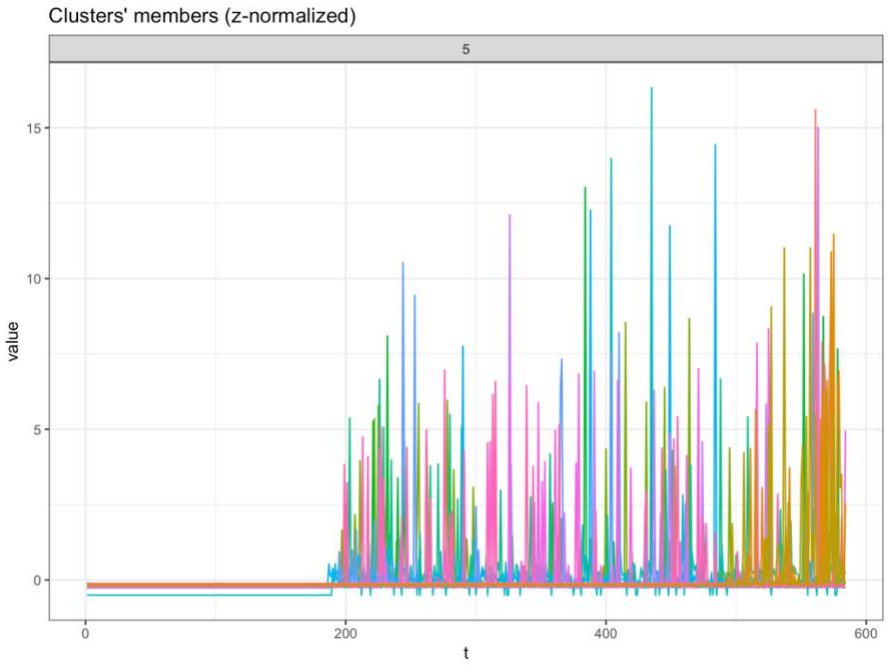
Plotting hierarchical cluster for with SBD and 6 is the good fit.

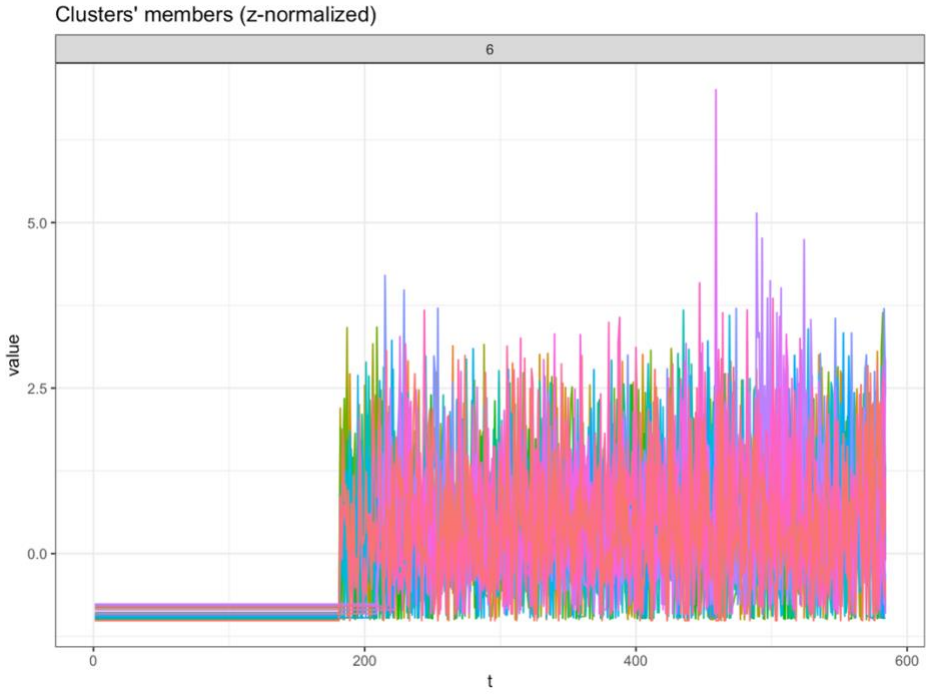
Plot of clusters


Clusters' members (z-normalized)
1
2
3
4
5
6
value
t
0
200
400
600


Clusters' members (z-normalized)
3
value
t
0
200
400
600

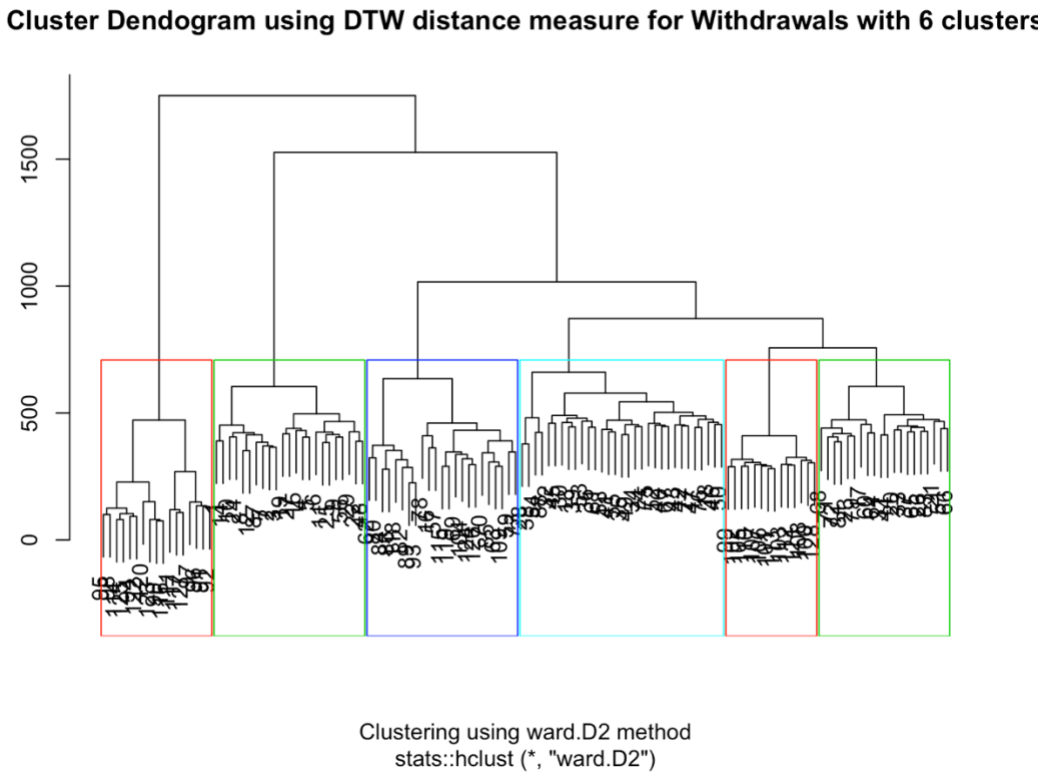

Clusters' members (z-normalized)
4
7.5
5.0
2.5
0.0
value
0
200
400
600
t


Clusters' members (z-normalized)
5
15
10
5
0
value
0
200
400
600
t



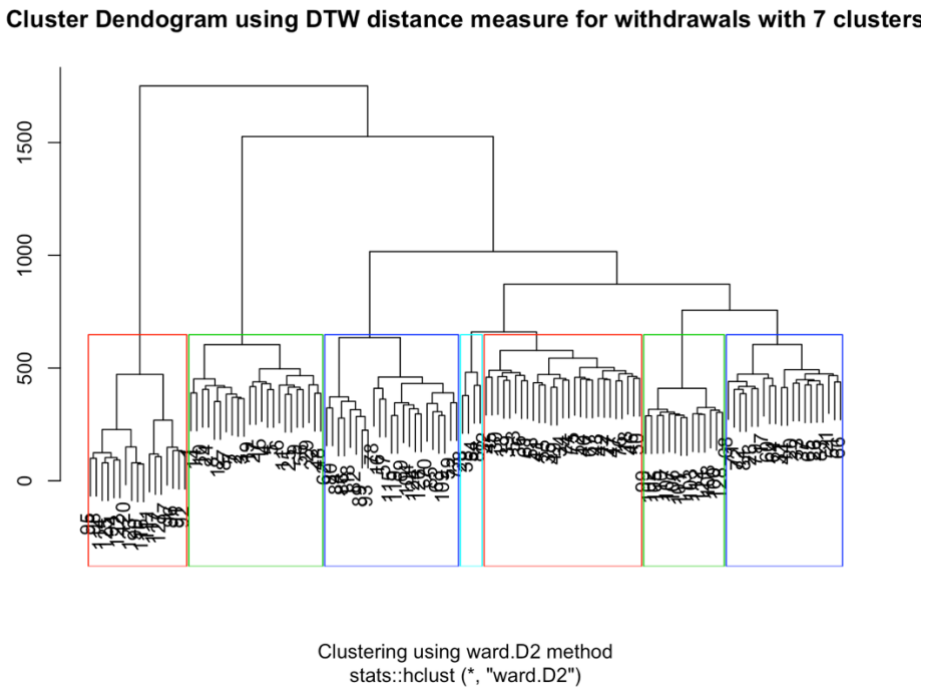
Above plots are with using SBD distance for 6 clusters as suggested. Each clusters seem to be appropriately capturing the behaviour at the aggregate level. we will plot data at branch level and ascertain.

Hierarchical clustering - DTW - 6 clusters



With DTW, we have cluster distribution. Here too we have 6 clusters. There might be a scope of splitting the 6th cluster into two diff clusters. Lets try and plot same.

Heirarchial clustering - DTW - 7 clusters



In this case, 7 seems to be a good indicator visually for the clusters.

```
##   Var1 Freq
## 1    1   14
## 2    2   19
## 3    3   36
## 4    4   28
## 5    5   12
## 6    6   19
```

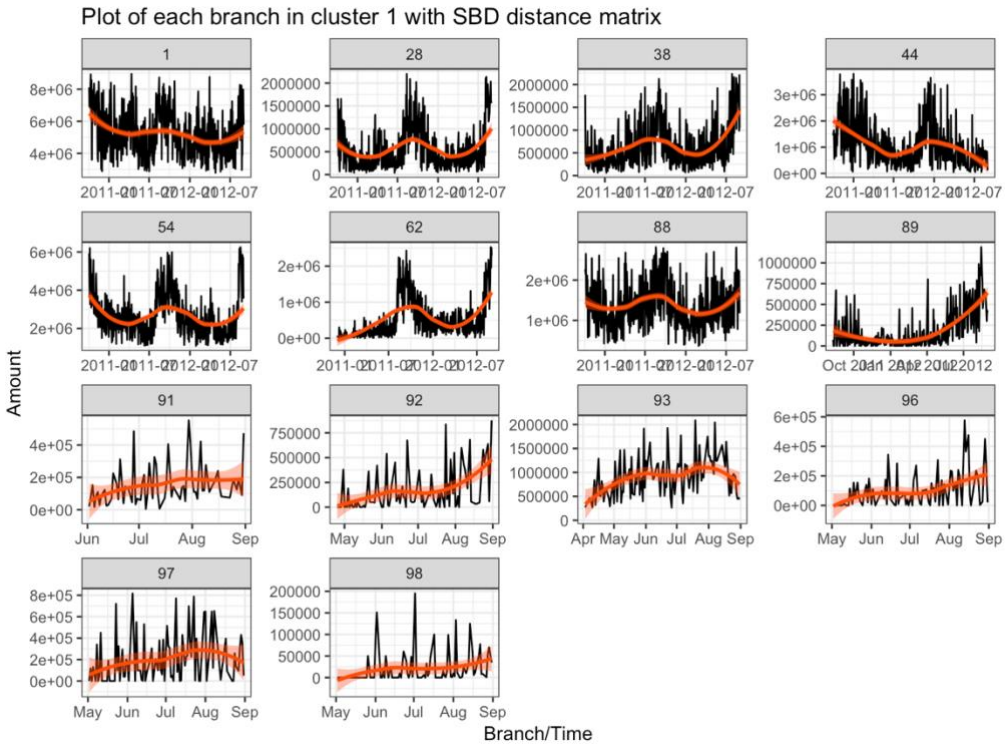
```
##   Var1 Freq
## 1    1   23
## 2    2   27
## 3    3   23
## 4    4   20
## 5    5    4
```

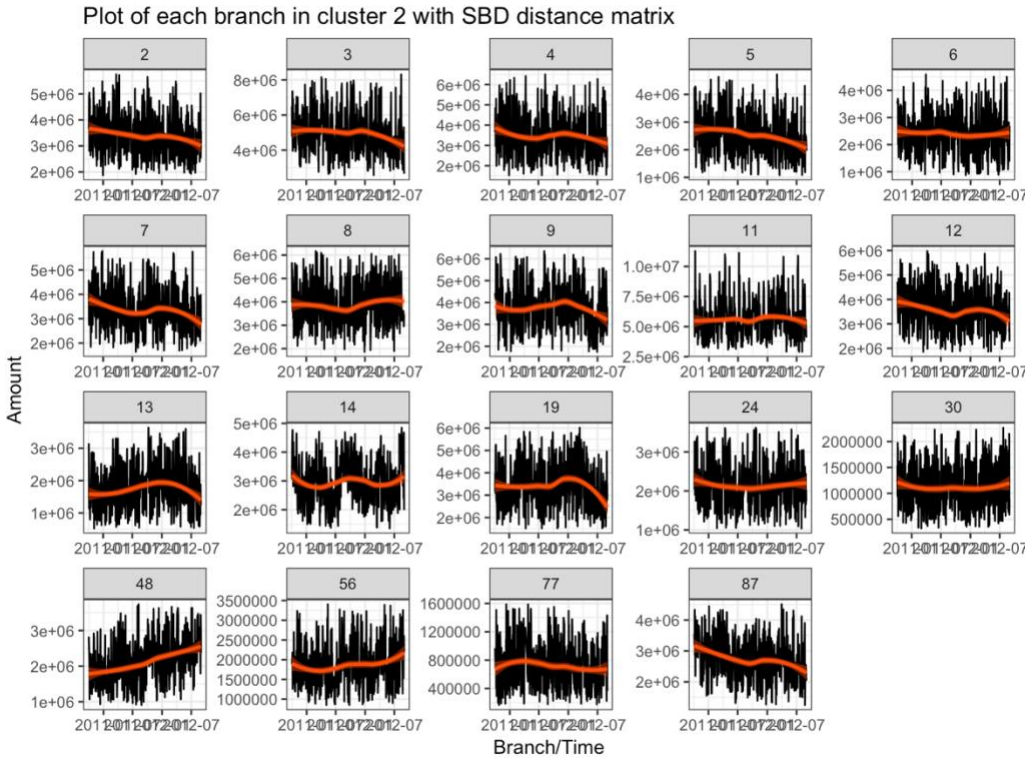
```
## 6     6    17
## 7     7    14
```

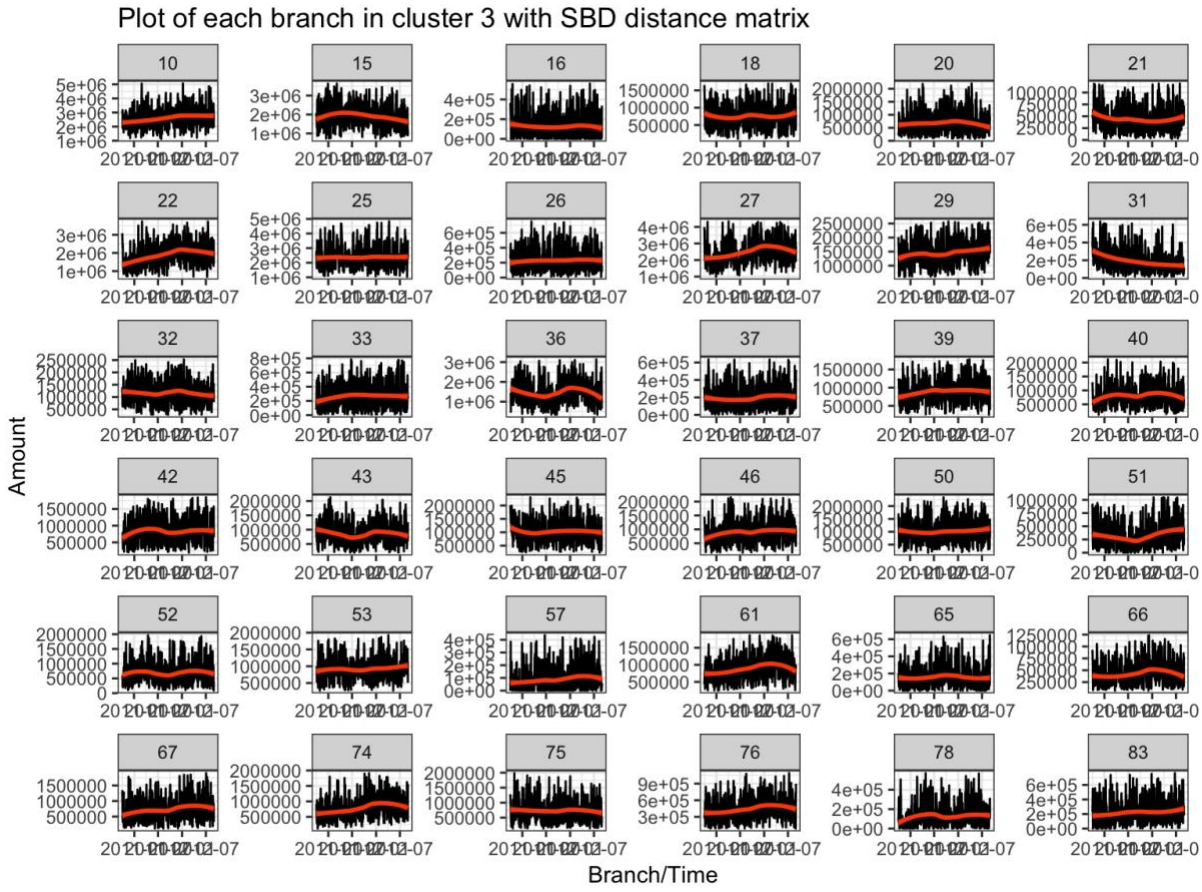
Above distribution is shown for the cluster distribution between SBD and DTW.

Plot of branches by clusters - SBD

We want a visual analysis of how the plots look for each branch by clusters with SBD distance matrix. Each cluster should represent some characteristics behaviour

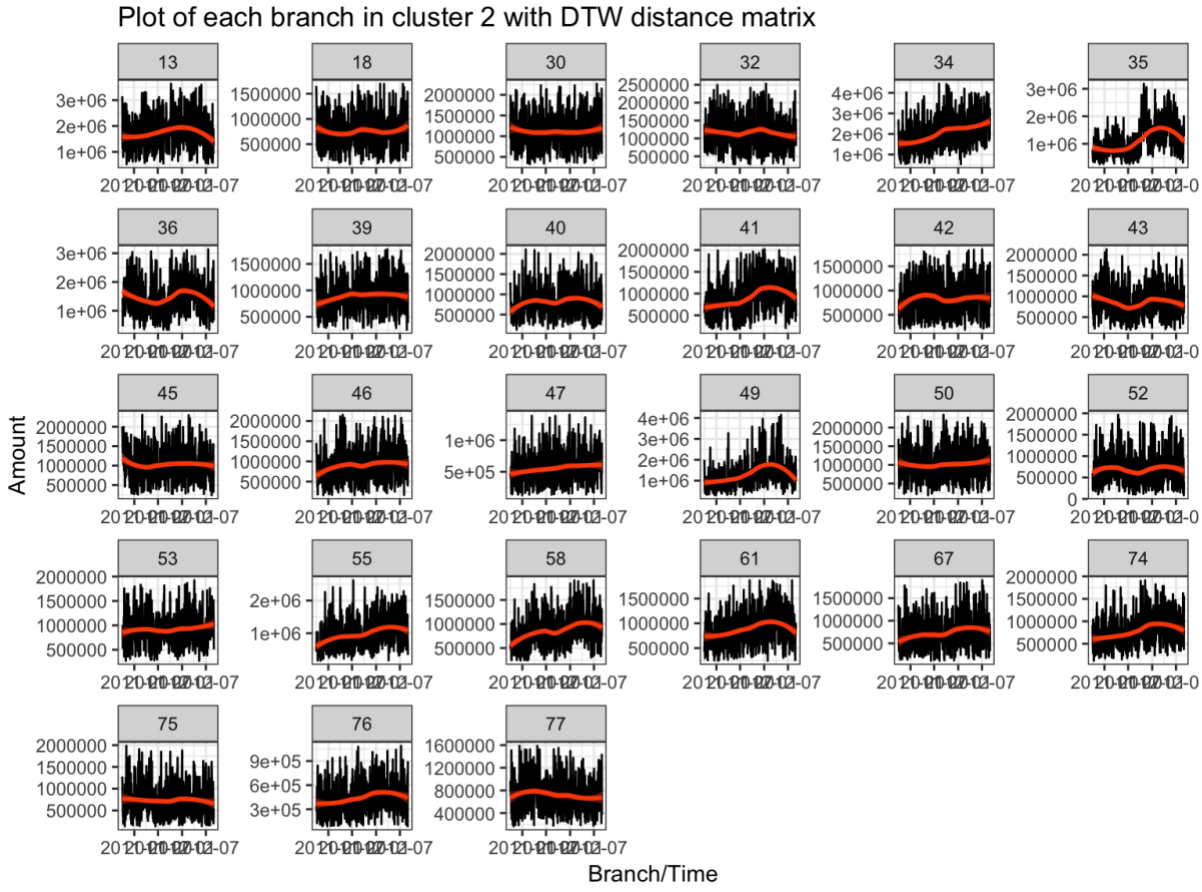


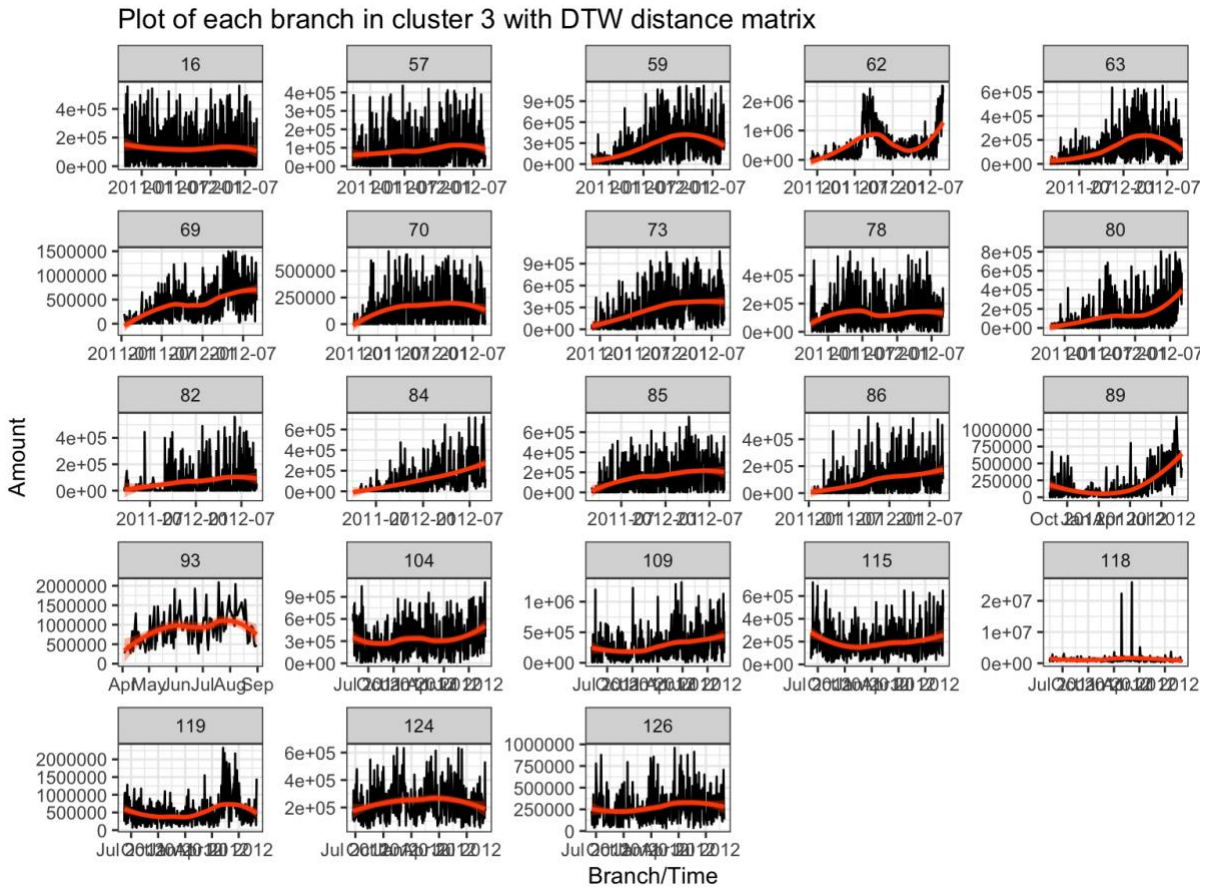


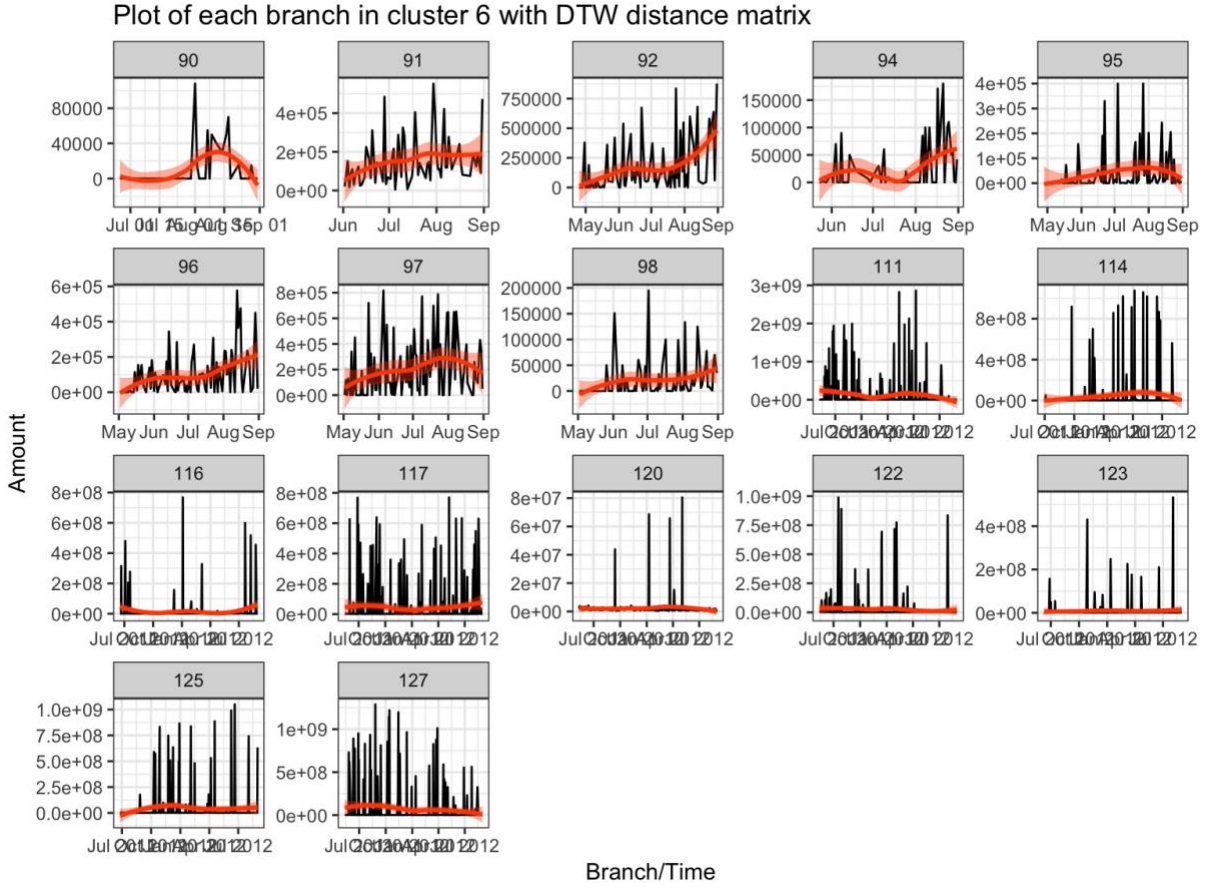


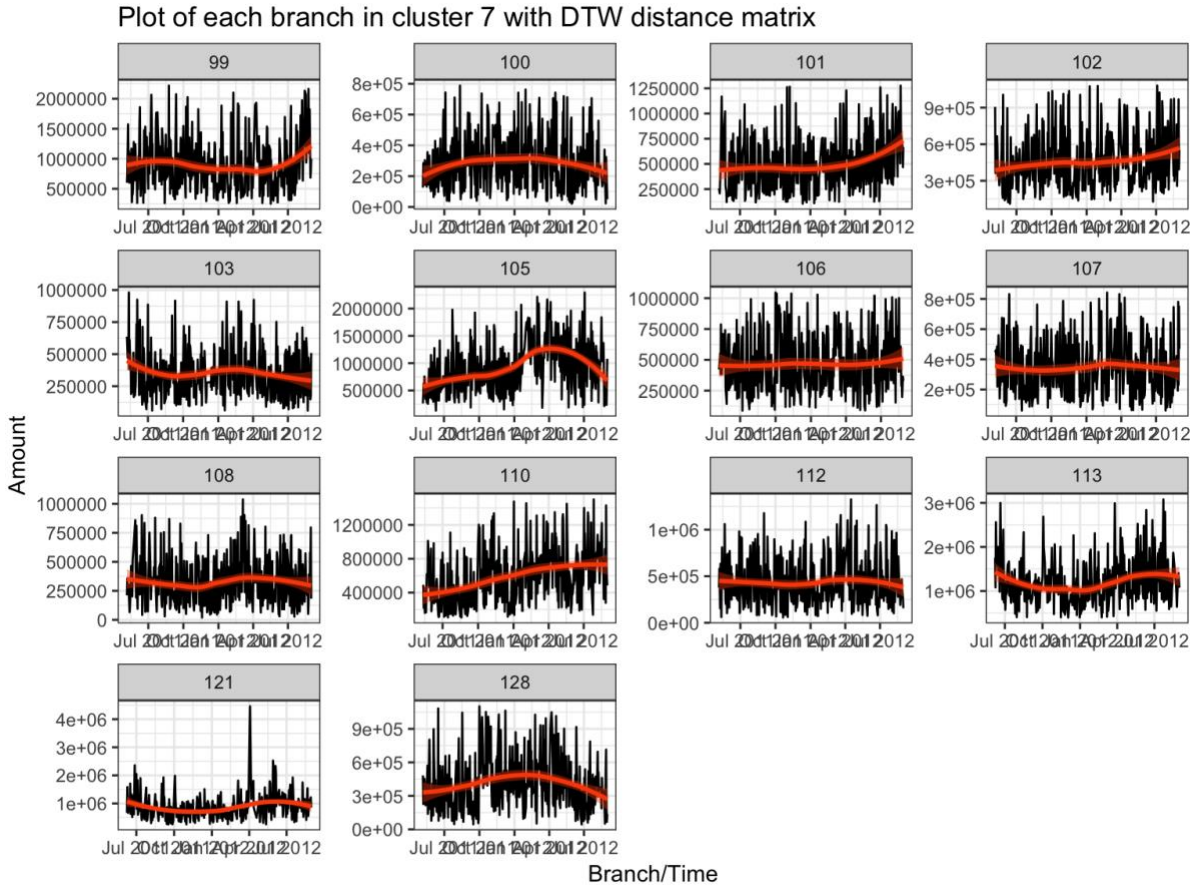
Cluster 1 here has expressed seasonality and volatility.
Cluster 2 looks a more of steady state affair with some deviations.

Cluster 3 has maximum elements and has steady state for complete duration.
Cluster 4 does exhibit some trend component in many branches.
Cluster 5 has collection of branches with sporadic demands.
Cluster 6 has higher degree of volatility compared to other clusters.




Plot of each branch in cluster 3 with DTW distance matrix
16
57
59
62
63
69
70
73
78
80
82
84
85
86
89
93
104
109
115
118
119
124
126
Amount
Branch/Time


Plot of each branch in cluster 6 with DTW distance matrix
90
91
92
94
95
96
97
98
111
114
116
117
120
122
123
125
127
Amount
Branch/Time



Cluster 1 here has some slight seasonality and volatility.
Cluster 2 looks a more of steady state affair with some deviations.
Cluster 3 has high volatility with steady mean for complete duration.
Cluster 4 does exhibit some trend component in many branches.
Cluster 5 has seasonality and trend component.
Cluster 6 has collection of branches with sporadic demands.
Cluster 7 exhibits volatility with steady mean, similar to cluster 3.

Between the plots of clusters with 6 and 7, we find the data with 7 clusters is better segregated. Especially it is able to group the branches which has sporadic rises on the timelines.

Conclusions

In this project, we started with data exploration and then moved on to clustering different branches which was the objective of the project. Following conclusion were drawn from this:

1. Between methods determining number of clusters, clValid, TSrepr and hierarchical plots, we seem to fall back more with visual concurrence of number of clusters. This was typically seen in withdrawals section where elbow plot showed 6 but h plot showed 7 to be better number.
2. Plotting each branch gives us fair idea on the distribution. e.g. sporadic data branches are clubbed together. Rising trend data also are grouped together.
3. Between the distance SBD and DTW methods used, DTW seems to be responding better to this data. Due to scope limitation in representing here, one needs to try different methods of distance measures to check which fits best.

There are many other functions which also provide functionalities of timeseries exploration and clustering. Functions here were chosen based on ease of use of data.

Next Steps

Use the information of clustering to fine tune parameter for forecasting model. Deploy a separate forecasting model for each cluster.