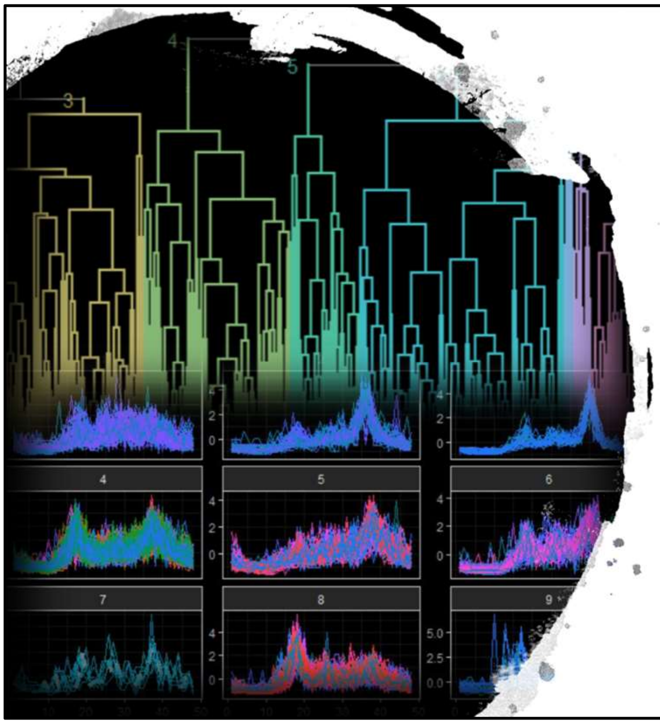


VizTS:

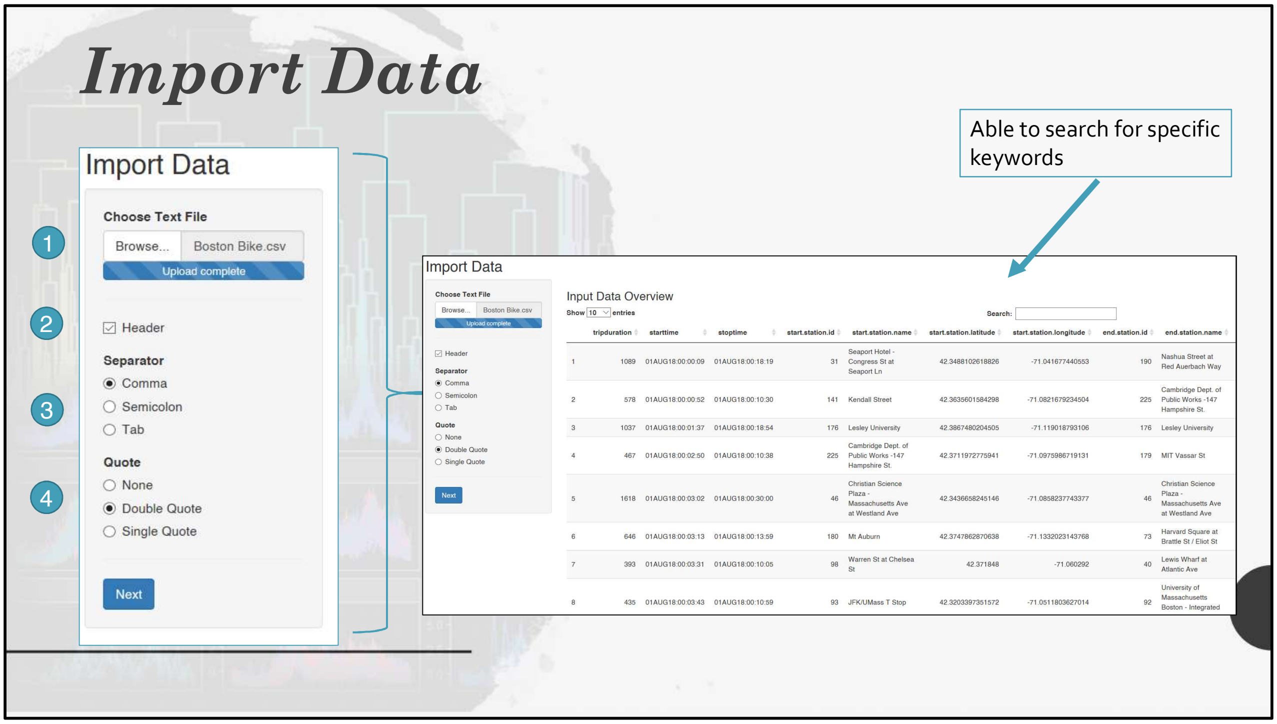
Clustering Edition

The User Guide

Introduction

VizTS: Clustering Edition, The User Guide

- Welcome to the VizTS: Clustering Edition. The application that aims to;
  - Empower user to perform Time Series Clustering without the need to code,
  - Enhance user experience by exposing the Cluster Evaluation and recommending optimal cluster,
  - Enable user to interactively visualize the output from Time Series Clustering,
  - Engage user in data discovery and in exploring the characteristics and idiosyncrasies of each cluster.
- The User Guide provides step-by-step instruction to assist user in leveraging the VizTS: Clustering Edition application



Import Data

Able to search for specific keywords

Import Data

Choose Text File

1 Browse... Boston Bike.csv

Upload complete

2 ☑ Header

3 Separator

● Comma
○ Semicolon
○ Tab

4 Quote

○ None
● Double Quote
○ Single Quote

Next

Import Data

Choose Text File

Browse... Boston Bike.csv

Upload complete

☑ Header

Separator

● Comma
○ Semicolon
○ Tab

Quote

○ None
● Double Quote
○ Single Quote

Next

Input Data Overview

Show 10 entries

Search:

	tripduration	starttime	stoptime	start.station.id	start.station.name	start.station.latitude	start.station.longitude	end.station.id	end.station.name
1	1089	01AUG18:00:00:09	01AUG18:00:18:19	31	Seaport Hotel - Congress St at Seaport Ln	42.3488102618826	-71.041677440553	190	Nashua Street at Red Auerbach Way
2	578	01AUG18:00:00:52	01AUG18:00:10:30	141	Kendall Street	42.3635601584298	-71.0821679234504	225	Cambridge Dept. of Public Works -147 Hampshire St.
3	1037	01AUG18:00:01:37	01AUG18:00:18:54	176	Lesley University	42.3867480204505	-71.119018793106	176	Lesley University
4	467	01AUG18:00:02:50	01AUG18:00:10:38	225	Cambridge Dept. of Public Works -147 Hampshire St.	42.3711972775941	-71.0975986719131	179	MIT Vassar St
5	1618	01AUG18:00:03:02	01AUG18:00:30:00	46	Christian Science Plaza - Massachusetts Ave at Westland Ave	42.3436658245146	-71.0858237743377	46	Christian Science Plaza - Massachusetts Ave at Westland Ave
6	646	01AUG18:00:03:13	01AUG18:00:13:59	180	Mt Auburn	42.3747862870638	-71.1332023143768	73	Harvard Square at Brattle St / Eliot St
7	393	01AUG18:00:03:31	01AUG18:00:10:05	98	Warren St at Chelsea St	42.371848	-71.060292	40	Lewis Wharf at Atlantic Ave
8	435	01AUG18:00:03:43	01AUG18:00:10:59	93	JFK/UMass T Stop	42.3203397351572	-71.0511803627014	92	University of Massachusetts Boston - Integrated

The Import Data Tab

In the import data tab, user can upload the text file of interest, have a quick overview of their data and make use of the Search function to search for specific data of interest.

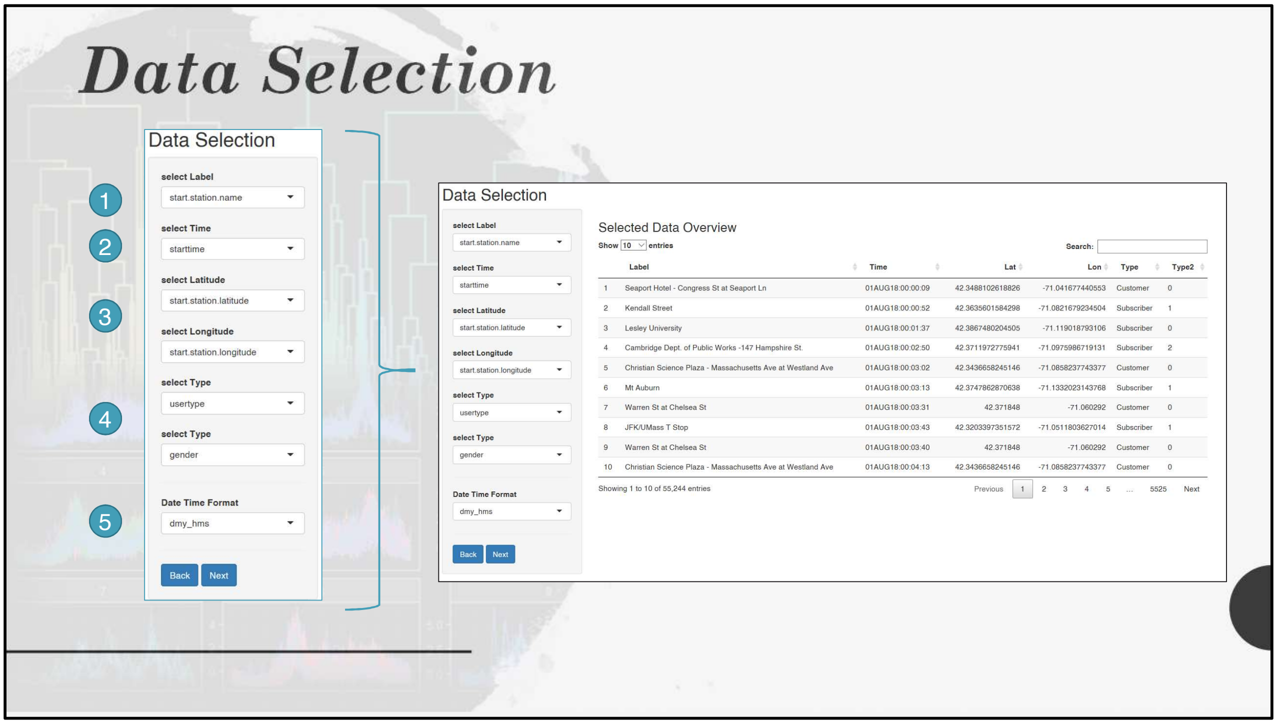
1) Upload the dataset of interest by clicking on the "Browse" button

2) User have the option to upload with or without Header just by checking the check box.

3) Depending on the raw dataset, user can have the option of Comma, Semicolon or Tab separator by selecting the different radio button.

4) Depending on the raw dataset, user can have the option of None, Double Quote or Single Quote text qualifier by selecting the different radio button.

Once the dataset is uploaded as desired, click on the "Next" button.



Data Selection

Data Selection

select Label: start.station.name

select Time: starttime

select Latitude: start.station.latitude

select Longitude: start.station.longitude

select Type: usertype

select Type: gender

Date Time Format: dmy_hms

Data Selection

select Label: start.station.name

select Time: starttime

select Latitude: start.station.latitude

select Longitude: start.station.longitude

select Type: usertype

select Type: gender

Date Time Format: dmy_hms

Selected Data Overview

Show 10 entries   Search:

	Label	Time	Lat	Lon	Type	Type2
1	Seaport Hotel - Congress St at Seaport Ln	01AUG18:00:00:09	42.3488102618826	-71.041677440553	Customer	0
2	Kendall Street	01AUG18:00:00:52	42.3635601584298	-71.0821679234504	Subscriber	1
3	Lesley University	01AUG18:00:01:37	42.3867480204505	-71.119018793106	Subscriber	0
4	Cambridge Dept. of Public Works -147 Hampshire St.	01AUG18:00:02:50	42.3711972775941	-71.0975986719131	Subscriber	2
5	Christian Science Plaza - Massachusetts Ave at Westland Ave	01AUG18:00:03:02	42.3436658245146	-71.0858237743377	Customer	0
6	Mt Auburn	01AUG18:00:03:13	42.3747862870638	-71.1332023143768	Subscriber	1
7	Warren St at Chelsea St	01AUG18:00:03:31	42.371848	-71.060292	Customer	0
8	JFK/UMass T Stop	01AUG18:00:03:43	42.3203397351572	-71.0511803627014	Subscriber	1
9	Warren St at Chelsea St	01AUG18:00:03:40	42.371848	-71.060292	Customer	0
10	Christian Science Plaza - Massachusetts Ave at Westland Ave	01AUG18:00:04:13	42.3436658245146	-71.0858237743377	Customer	0

Showing 1 to 10 of 55,244 entries

The Data Selection Tab

As the raw dataset contains many columns, the application provides a data selection function, where users can select the data column needed to perform Time Series Clustering.

1) Select Label: Select the data column which user wants to perform their time series clustering on. For example, Station Name, Country, etc.

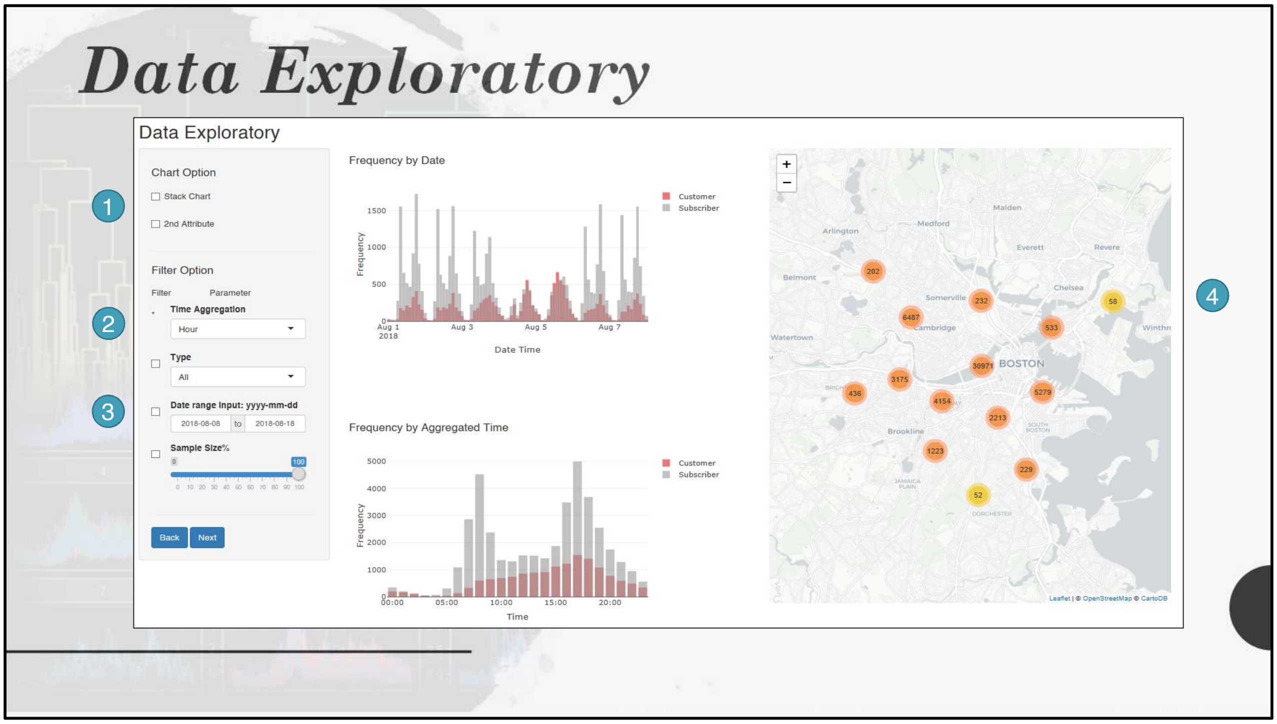
2) Select Time: Select the data column with time data. For example, Start Time, End Time, etc.

3) Select Latitude, select Longitude: Select the data column with latitude and longitude data (if applicable), otherwise, option to leave it "Blank".

4) Select Type: Select the data column with variable of interest. The application allows users to select up to two different variable. For example, usertype, gender, continent, etc.

5) Select Date Time Format: Select the appropriate date time format as per the dataset uploaded. Option of dmy_hms or ymd_hms.

Once the data is selected, click on the "Next" button.



Data Exploratory

Time Aggregation: Hour

Type: All

Date range input: yyyy-mm-dd 2018-08-08 to 2018-08-18

Sample Size% 100

The Data Exploratory Tab

The application allows user to further explore their dataset and to employ the various filter option to prepare their data for time series clustering.

1) Chart Option

- The user have an option to view their data as a stack chart or by default the histogram is overlay.
- The user have an option to view their data as Type 2 or by default as Type 1. Type 1 and Type 2 is referring to the variable of interest as selected in the Data Selection tab.

2) Filter Option (Time Aggregation – User can select the time aggregation desired or default will be Hour)

- The Time Aggregation option allows user to have a quick overview of their time series data at different time aggregation. Option of Day_Time, Day, 6 Hours, Hour, 30 minutes and 15 minutes.
- The Time Aggregation is also a data preparation step, where user can choose the desired time aggregation to perform time series clustering on. This will be illustrated in subsequent page.

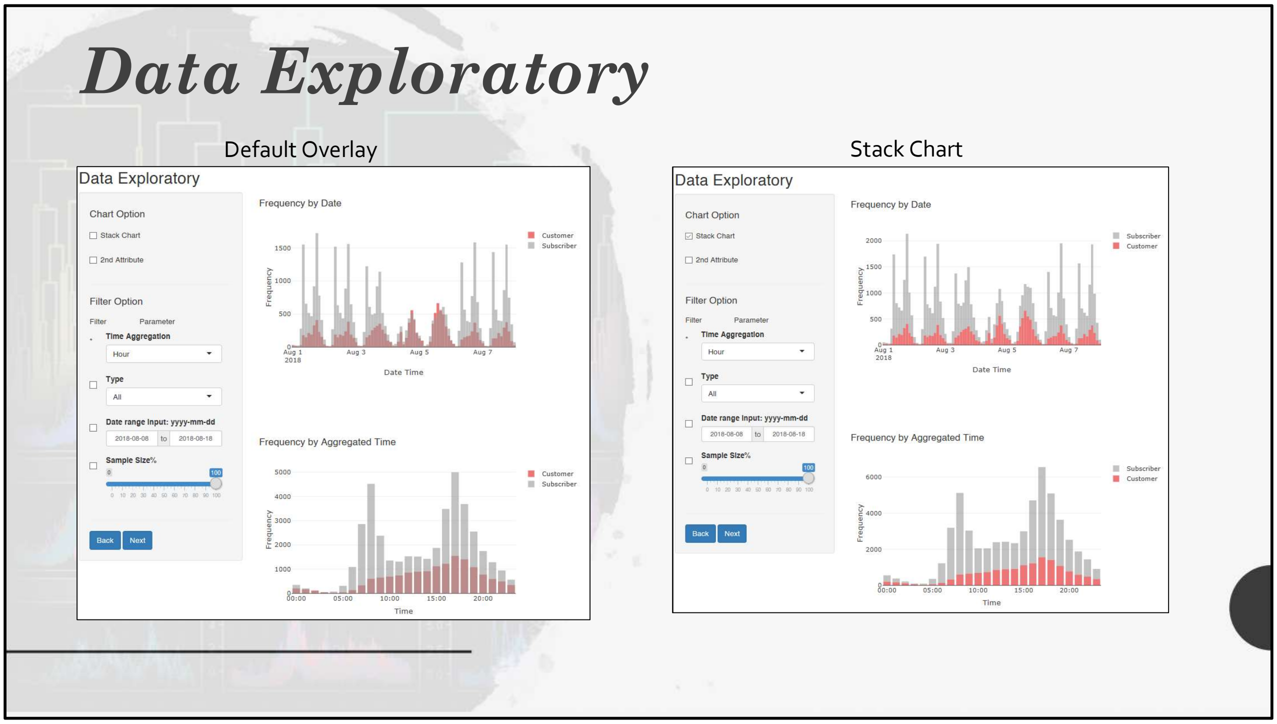
3) Filter Option (Type, Date range input, Sample Size – User can leave the check box uncheck if no further data filtering is needed to perform time series clustering)

- Type:
  - User can select the variable to explore. From the dropdown menu, select type of interest and then check the check box for the application to filter the data accordingly.
  - Option of All (default) and list of Type (depending on the list of type). For example, Type 1 selected from the Date Selection tab is usertype, the list of type will be Customer or Subscriber. This will be illustrated in subsequent page.
- Date range input:
  - User can select the time frame of interest by making use of the calendar. Once the range is selected, check the check box for the application to filter the data accordingly
- Sample Size
  - User can choose to explore a sample size of their data by using the slider bar. Once the sample size is selected, check the check box for the application to filter the data accordingly

4) Exploratory Map – For exploratory purpose only

- The Map will be available for user to view their data if latitude and longitude data is provided. If none are provided, this space will be blank.
- The numbers represents the frequency at the geographical location. By zooming into the map, the geographical location will spread out and the number will change accordingly.

Once the desired filter is selected, click on the “Next” button.



Data Exploratory

Default Overlay

Data Exploratory

Chart Option

□ Stack Chart

□ 2nd Attribute

Filter Option

Filter Parameter

Time Aggregation: Hour

□ Type: All

□ Date range input: yyyy-mm-dd 2018-08-08 to 2018-08-18

□ Sample Size%: 0 – 100

Back Next

Stack Chart

Data Exploratory

Chart Option

☑ Stack Chart

□ 2nd Attribute

Filter Option

Filter Parameter

Time Aggregation: Hour

□ Type: All

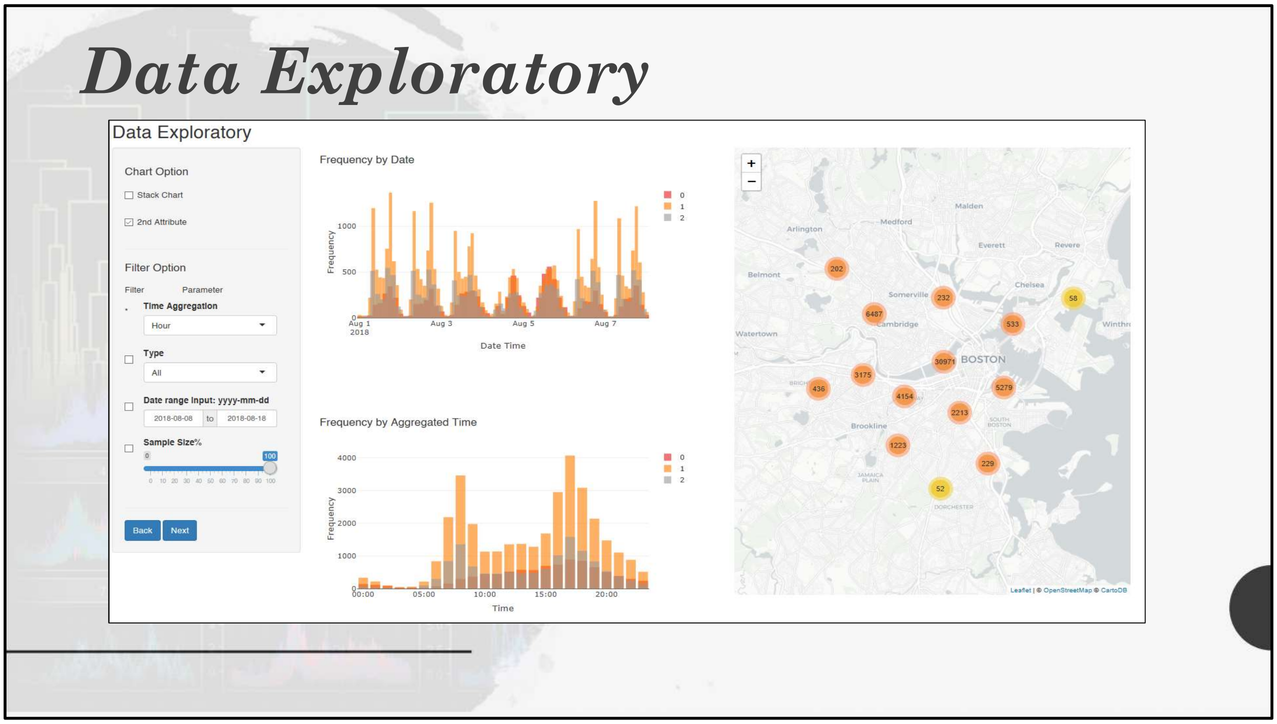
□ Date range input: yyyy-mm-dd 2018-08-08 to 2018-08-18

□ Sample Size%: 0 – 100

Back Next

Illustration

To note that the stack chart can also be applied when 2nd attribute is selected. This option is mainly for data exploratory purpose.



Data Exploratory

Data Exploratory

Chart Option

☐ Stack Chart

☑ 2nd Attribute

Filter Option

Filter Parameter

Time Aggregation: Hour

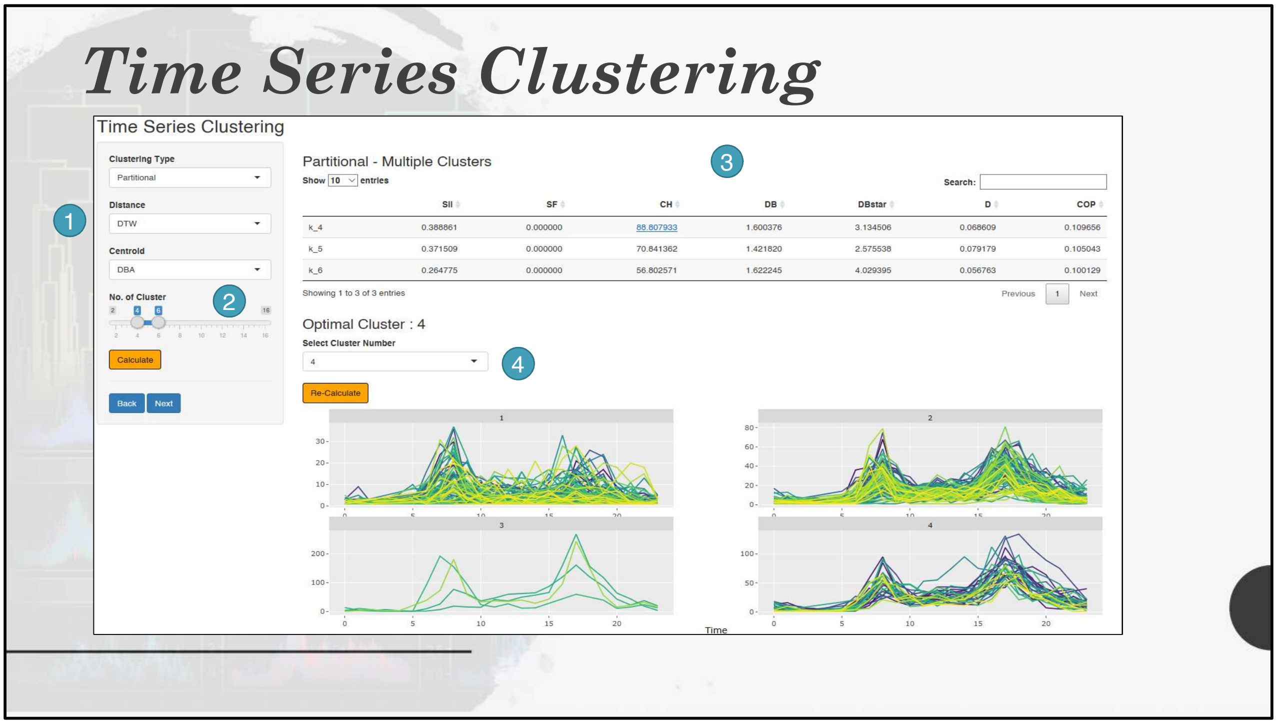
☐ Type: All

☐ Date range input: yyyy-mm-dd: 2018-08-08 to 2018-08-18

☐ Sample Size%: 0 – 100

Illustration

Able to perform time series clustering on the Type 2 (2nd Attribute) selected. In this example, the Type 2 selected was Gender which have 3 different parameter 0, 1, and 2.



Time Series Clustering

Time Series Clustering

Clustering Type: Partitional

Distance: DTW

Centroid: DBA

No. of Cluster: 4 – 6

Partitional - Multiple Clusters

Show 10 entries

Search:

	SII	SF	CH	DB	DBstar	D	COP
k_4	0.388861	0.000000	88.807933	1.600376	3.134506	0.068609	0.109656
k_5	0.371509	0.000000	70.841362	1.421820	2.575538	0.079179	0.105043
k_6	0.264775	0.000000	56.802571	1.622245	4.029395	0.056763	0.100129

Showing 1 to 3 of 3 entries

Optimal Cluster : 4

Select Cluster Number: 4

The Time Series Clustering Tab

The main data analysis tab where user can perform time series clustering on data they have selected and prepared from the previous tabs.

1) Clustering Options

- Clustering Type:
  - User can select between Hierarchical Type or Partitional Type. The application provides these two type of clustering at the moment as they are the more common type.
  - If Hierarchical Type is selected. Additional option of Method will be available. Choices of ward.D, ward.D2, single, complete, average, mcquitty, median and centroid. This will be illustrated at subsequent page.
- Distance:
  - User can choose the different distance measure. Choices of DTW, GAK and SBD. If SBD is selected the centroid algorithm selected will be Shape. As this is the recommended centroid. SBD and Shape is the k-shape clustering algorithm.

- Centroid:
    - User can choose the different centroid algorithm. Choices of DBA and PAM.

2) No of Cluster

- The application provides user to select a range of cluster to be calculated by using the slider bar. For cases of partitional clustering, there is a need to choose a cluster first, therefore, by selecting a range of cluster, user does not need to compute it one by one.
- Once the desired range of cluster is chosen, click on “Calculate”.
- User can also select a single cluster instead. This will be further illustrated at subsequent page.

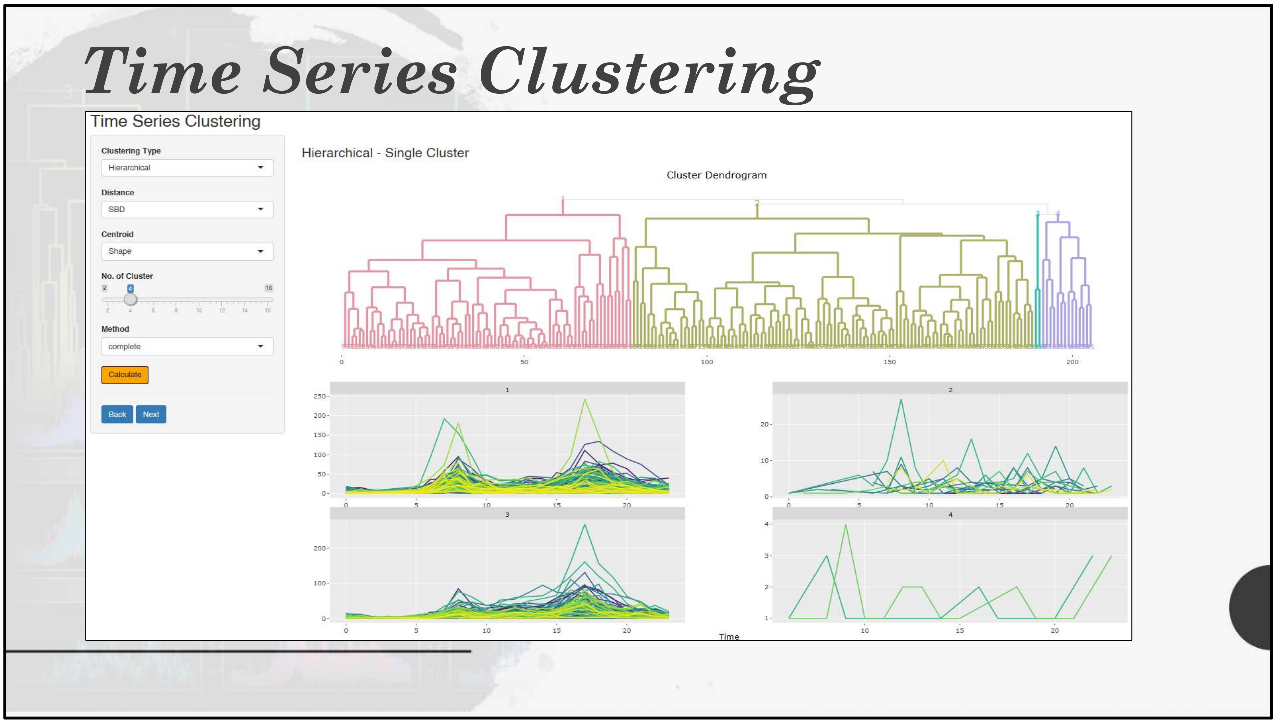
3) Cluster Evaluation

- One of the key feature of this application is the Cluster Evaluation. From the range of the cluster selected, the application will evaluate the cluster by employing the various Cluster Validity Indices (CVI).
- The table shows the computed CVIs and user are able to sort the CVI values for further understanding.

4) Cluster Recommendation

- From the CVIs calculated, the application will recommend an optimal cluster. Optimal cluster is recommended based on majority vote from the various CVI. Generally, the cluster with the majority highest CVIs is selected.
- User can then select the optimal cluster or if desired, the other cluster from the range selected from the dropdown menu.
- With the cluster selected, click “Re-Calculate” to plot the time series clustering.

Once the desired cluster is selected, click on the “Next” button.



Time Series Clustering

Time Series Clustering

Clustering Type

Hierarchical

Distance

SBD

Centroid

Shape

No. of Cluster

4

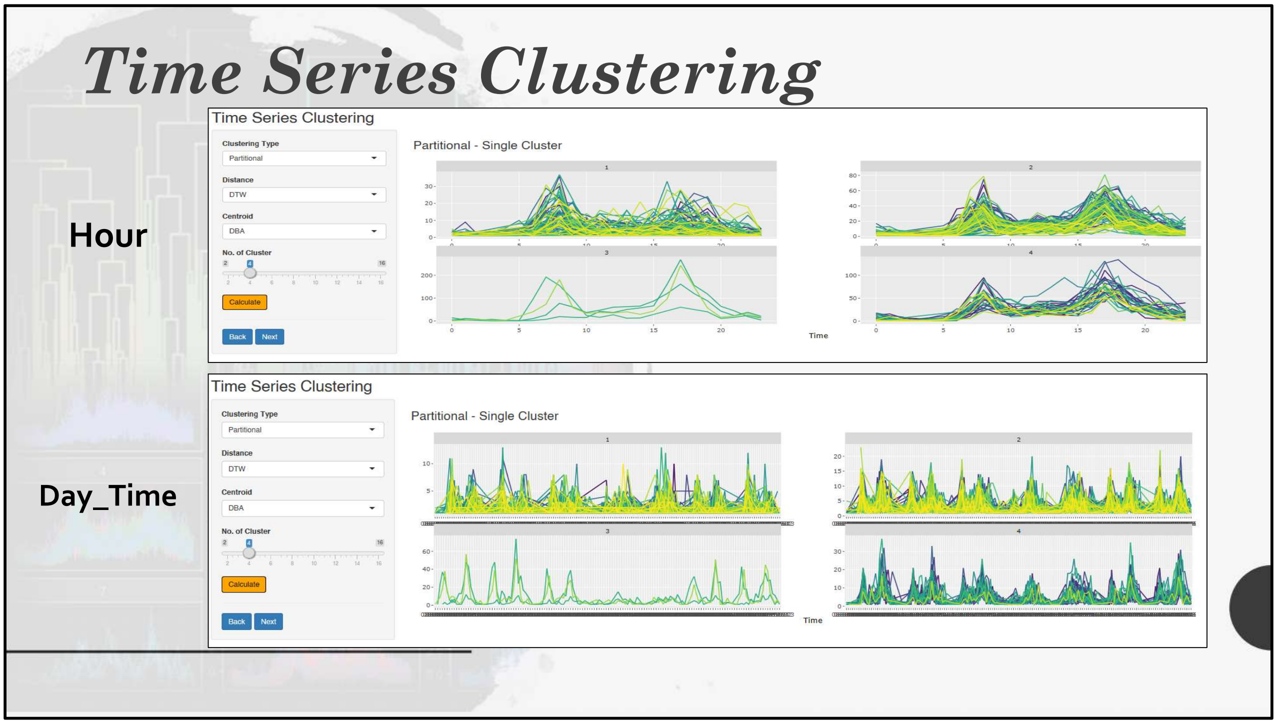
Method

complete

Hierarchical - Single Cluster

Illustration

When Hierarchical Clustering and a Single Cluster is selected, a Cluster Dendrogram and the Time Series line graph will be plotted. The Method option is also available when Hierarchical is selected. Note that the dendrogram is for exploratory purpose.



Time Series Clustering

Hour

Time Series Clustering

Clustering Type: Partitional

Distance: DTW

Centroid: DBA

No. of Cluster: 4

Partitional - Single Cluster

Day_Time

Time Series Clustering

Clustering Type: Partitional

Distance: DTW

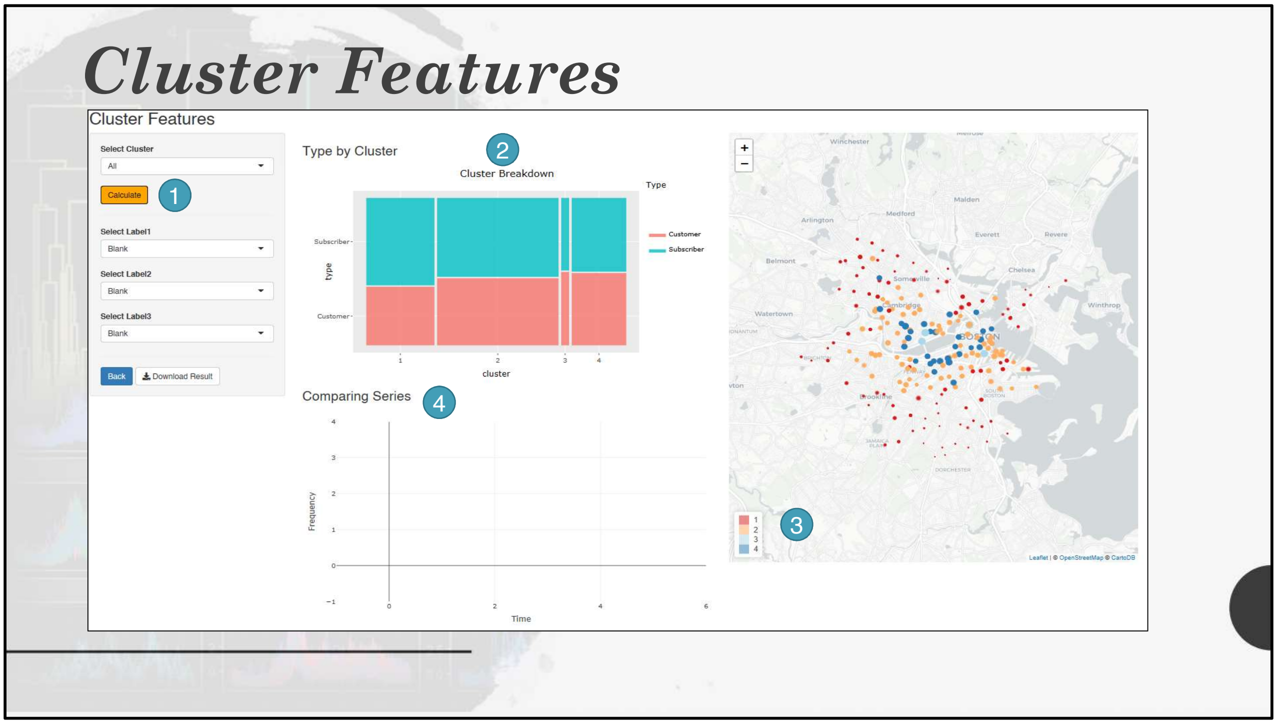
Centroid: DBA

No. of Cluster: 4

Partitional - Single Cluster

Illustration

When different time aggregation is selected, the time series clustering will change accordingly.



Cluster Features

Cluster Features

Select Cluster: All

Select Label1: Blank

Select Label2: Blank

Select Label3: Blank

The Cluster Feature Tab

From the clustering analysis performed, the Cluster Feature tab allows user to explore their times series clustering results and download their results.

1) Select Cluster

- From the previous Time Series Clustering tab, user selected the desired cluster. In this Cluster Feature tab, user can view how each of their cluster results.
- In this example, the Time Series Clustering was perform using the optimal cluster of 4. In the Select Cluster, the dropdown menu provides user with the option to view the clustering results of each of the cluster. Select the desired cluster to view and click on Calculate. In this example, All was selected.
- When a specific cluster is chosen, the application will display the results for that cluster. This will be illustrated in subsequent page.

2) Type of Cluster

- From the Data Exploratory Tab, user are able to select the different variable of interest. In this example, usertype was chosen and usertype consists of Customer and Subscriber.

- The mosaic plot allows user to visualize the proportion of the type in each cluster.
- If a single cluster is selected, only the proportion of that cluster will be display.

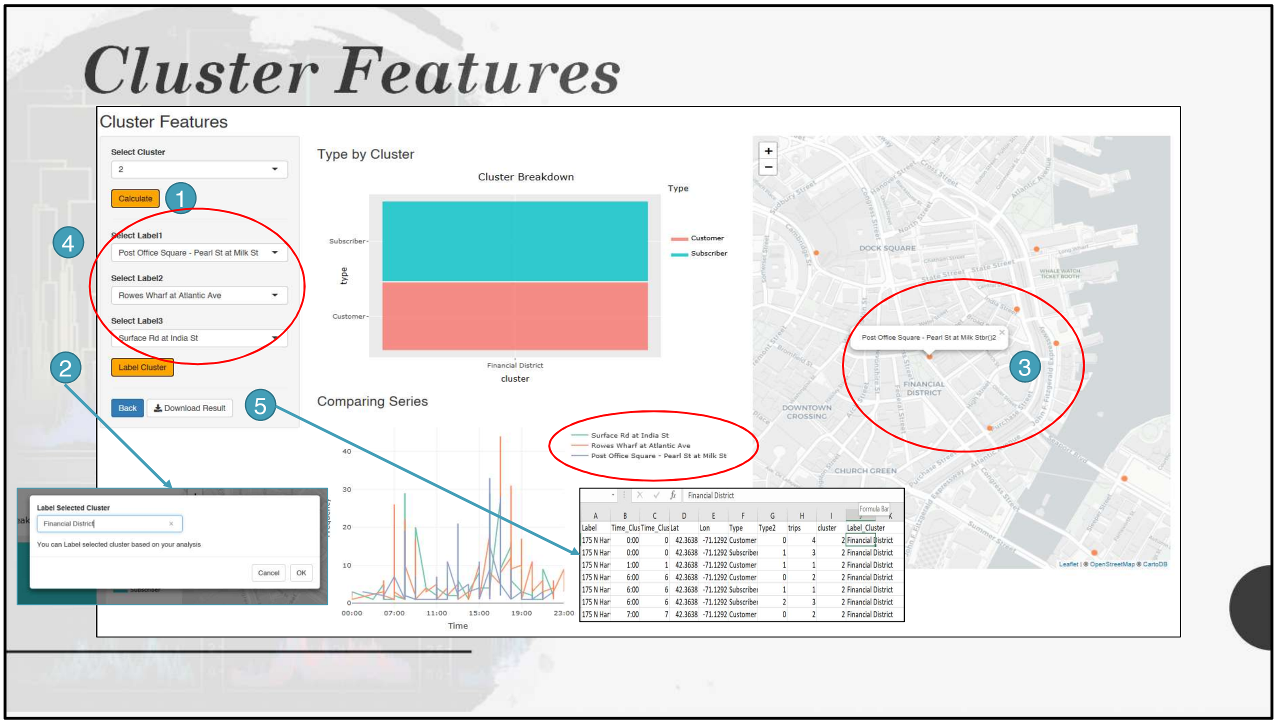
3) Map

- One of the unique feature of this application is that it allows user to further view their time series clustering results on the map when the latitude and longitude data is available.
- From the map, users can easily visualize their resulting time series cluster. Each different cluster is a different colour.
- User are able to zoom in and out of the map for further exploration on where are these clusters.

4) Comparing Series

- The Comparing Series visualization allows user to select three Label to compare and a line graph will be plotted for user to visualize the results
- This is mainly use when comparing the series in one cluster to allow user to visualize how similar the time series are within the cluster. This will be further elaborated in subsequent page.

Once user are satisfied with their Time Series Clustering results, click on the Download Results button to export the results out.



Select Cluster: 2

Select Label1: Post Office Square - Pearl St at Milk St

Select Label2: Rowes Wharf at Atlantic Ave

Select Label3: Surface Rd at India St

Label Selected Cluster

Financial District

You can Label selected cluster based on your analysis

Cancel OK

Financial District

A	B	C	D	E	F	G	H	I	J	K
Label	Time_Clus	Time_Clus	Lat	Lon	Type	Type2	trips	cluster	Label_Cluster	
175 N Har	0:00	0	42.3638	-71.1292	Customer	0	4	2	Financial District	
175 N Har	0:00	0	42.3638	-71.1292	Subscriber	1	3	2	Financial District	
175 N Har	1:00	1	42.3638	-71.1292	Customer	1	1	2	Financial District	
175 N Har	6:00	6	42.3638	-71.1292	Customer	0	2	2	Financial District	
175 N Har	6:00	6	42.3638	-71.1292	Subscriber	1	1	2	Financial District	
175 N Har	6:00	6	42.3638	-71.1292	Subscriber	2	3	2	Financial District	
175 N Har	7:00	7	42.3638	-71.1292	Customer	0	2	2	Financial District	

Demostration

Using Cluster 2 as example.

1) Select Cluster

- Select Cluster 2 from the dropdown menu and click on calculate.
- The application will first only display the Type by Cluster and Map of cluster 2 only.
- User can further explore the data by using the Comparing Series function and by zooming into the Map.

2) Label Cluster

- As the main purpose of clustering analysis is to discover and investigate features and characteristic of their data, the application provides user with the capability to label their cluster on the go.
- If user wants to label their cluster, click on the Label Cluster button. A pop-up window will prompt user to type in the Label. In this example, Financial District was typed in. Click on OK.
- Click on the Calculate button again to refresh the tab.

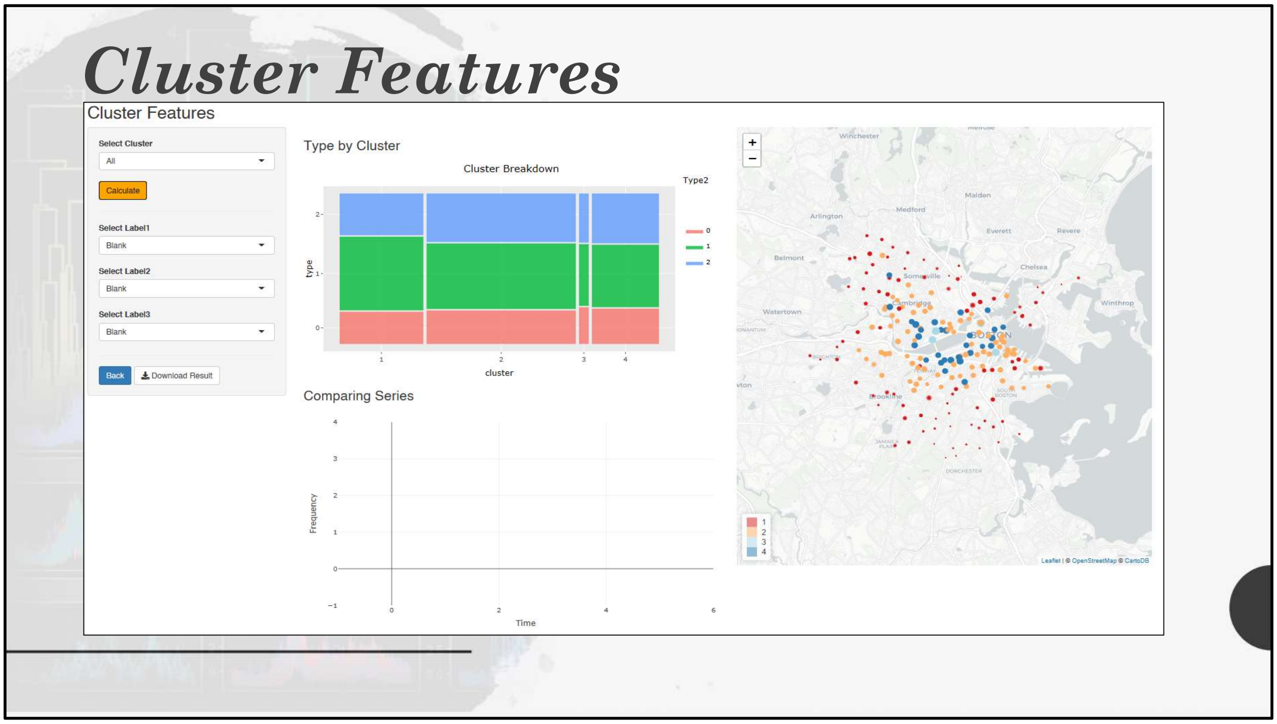
- The Type by Cluster plot will update the mosaic plot’s axis.

3) Map
    - From the Map, user can click on the point to view the Label.
    - In this example, the map was utilize to view the three stations that are at close proximity.

4) Select Label
    - From the Label selection, the three stations was selected from the dropdown menu.
    - The Comparing Series line graph will then plot the selected label time series to enable user to further explore and understand the time series clustering results.

5) Download Results
    - Once satisfied, user can download their results. The result will be saved as a csv file.
    - If user have labelled their cluster. The csv file will also capture the label as part of the clustering results.

User can click on the Back button to explore and investigate the time series clustering using the different clustering selection, option and filters provided by the VizTS: Clustering Edition.



Cluster Features

Cluster Features

Select Cluster: All

Calculate

Select Label1: Blank

Select Label2: Blank

Select Label3: Blank

Back

Download Result

Type by Cluster

Comparing Series

Illustration

Selecting Type 2 (Gender) as a variable instead. The mosaic plot show the proportion of gender within the cluster.