

LAB : CLUSTERING

Clustering in Tableau

- Choose a dataset
- Choose the attributes for clustering
- Choose number of clusters

(Uses K-Means Clustering)

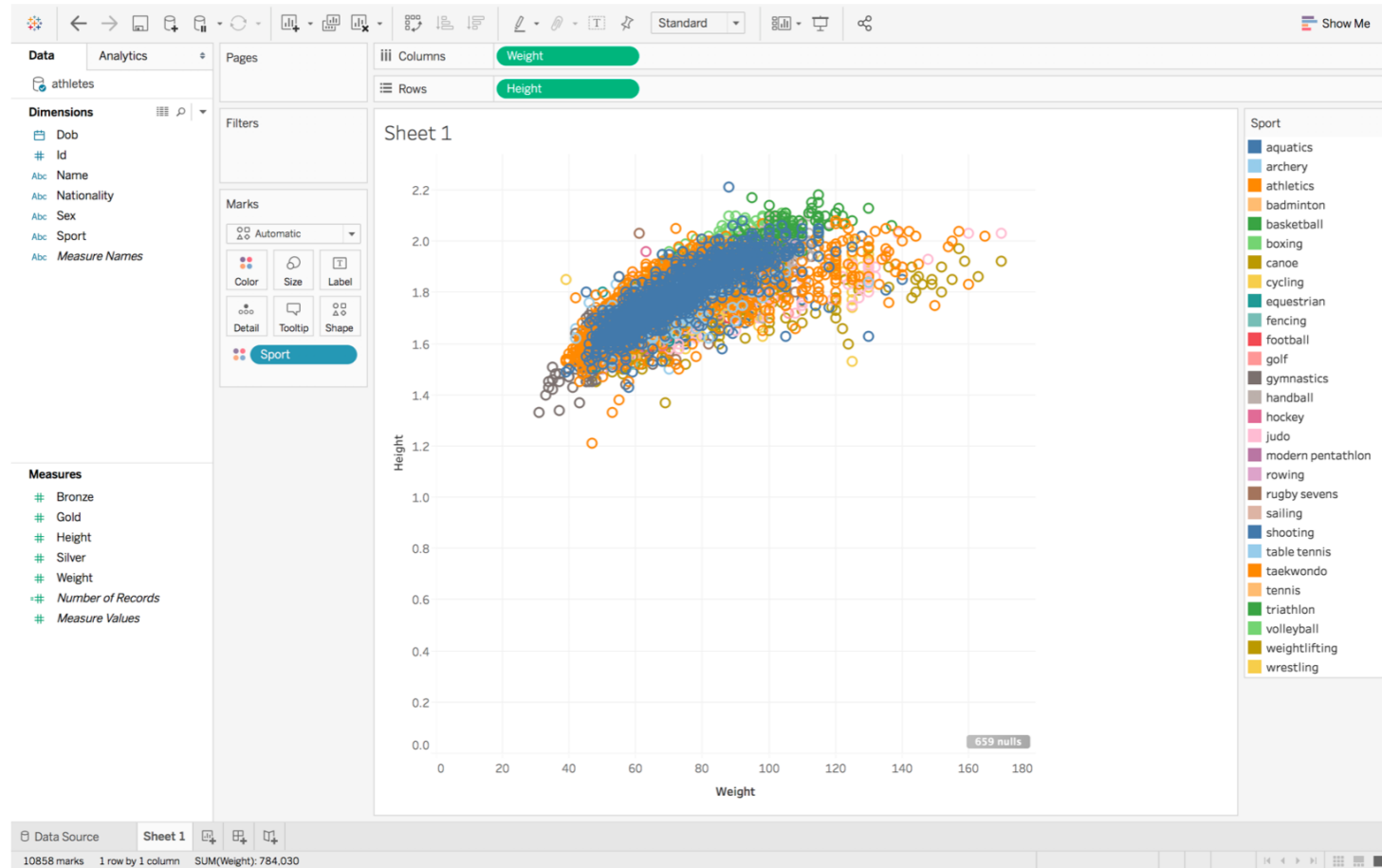
Clustering in Tableau

(using athletes.csv file)

The screenshot shows the Tableau interface with a data source named 'breastcancer' connected. The 'Files' pane on the left lists 'athletes.csv' and 'breastcancer.csv'. The main view displays a table with 683 rows and 11 columns. The columns are: #, Id, Cl.thickness, Cell.size, Cell.shape, Marg.adhesion, Epith.c.size, Bare.nuclei, Bl.cromatin, Normal.nucleoli, Mitoses, and Class. The 'Class' column contains values 'benign' and 'malignant'.

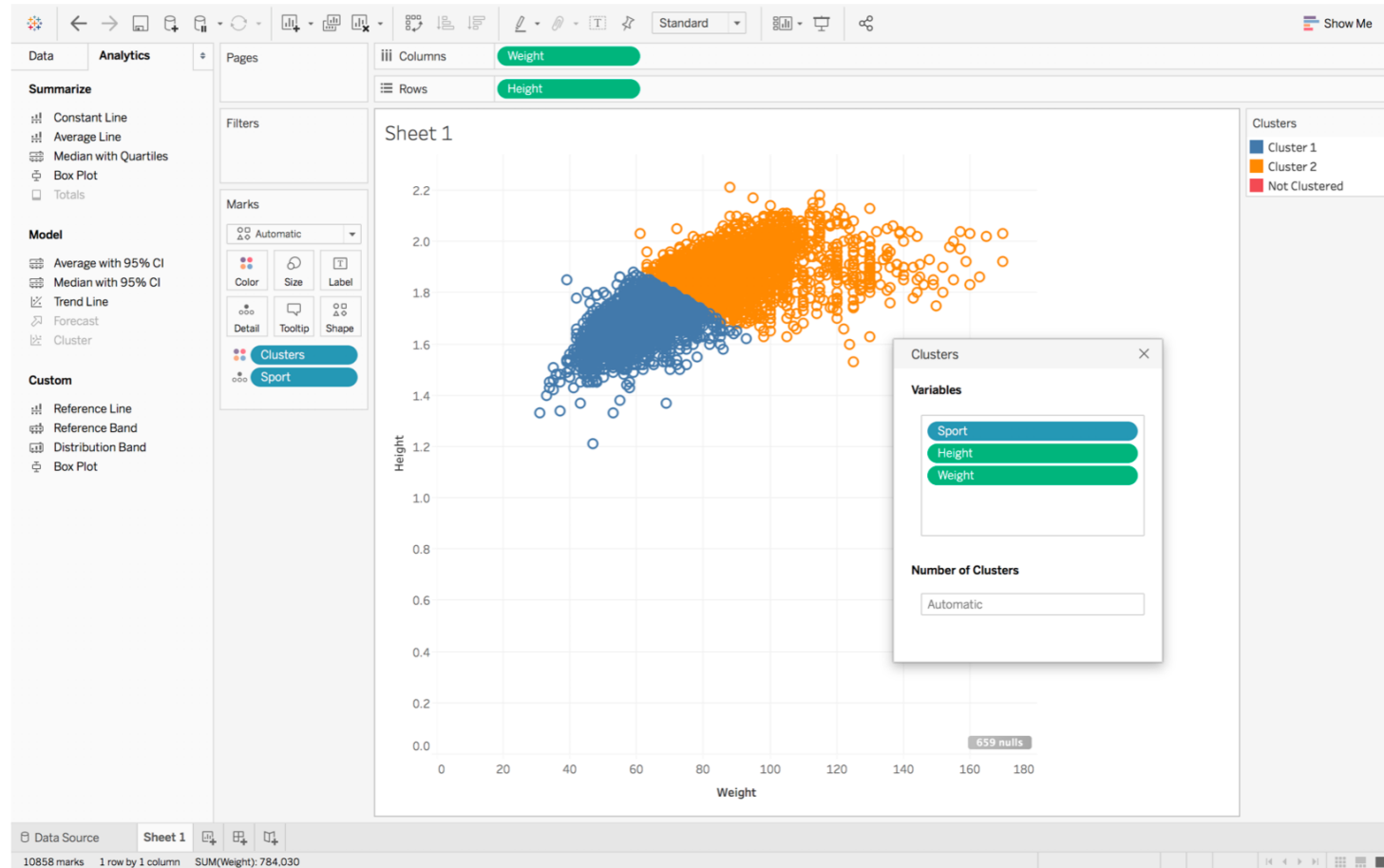
#	breastcancer.csv Id	breastcancer.csv Cl.thickness	breastcancer.csv Cell.size	breastcancer.csv Cell.shape	breastcancer.csv Marg.adhesion	breastcancer.csv Epith.c.size	breastcancer.csv Bare.nuclei	breastcancer.csv Bl.cromatin	breastcancer.csv Normal.nucleoli	breastcancer.csv Mitoses	breastcancer.csv Class
1000025		5	1	1	1	2	1	3	1	1	benign
1002945		5	4	4	5	7	10	3	2	1	benign
1015425		3	1	1	1	2	2	3	1	1	benign
1016277		6	8	8	1	3	4	3	7	1	benign
1017023		4	1	1	3	2	1	3	1	1	benign
1017122		8	10	10	8	7	10	9	7	1	malignant
1018099		1	1	1	1	2	10	3	1	1	benign
1018561		2	1	2	1	2	1	3	1	1	benign
1033078		2	1	1	1	2	1	1	1	5	benign
1033078		4	2	1	1	2	1	2	1	1	benign
1035283		1	1	1	1	1	1	3	1	1	benign
1036172		2	1	1	1	2	1	2	1	1	benign
1041801		5	3	3	3	2	3	4	4	1	malignant

Clustering in Tableau



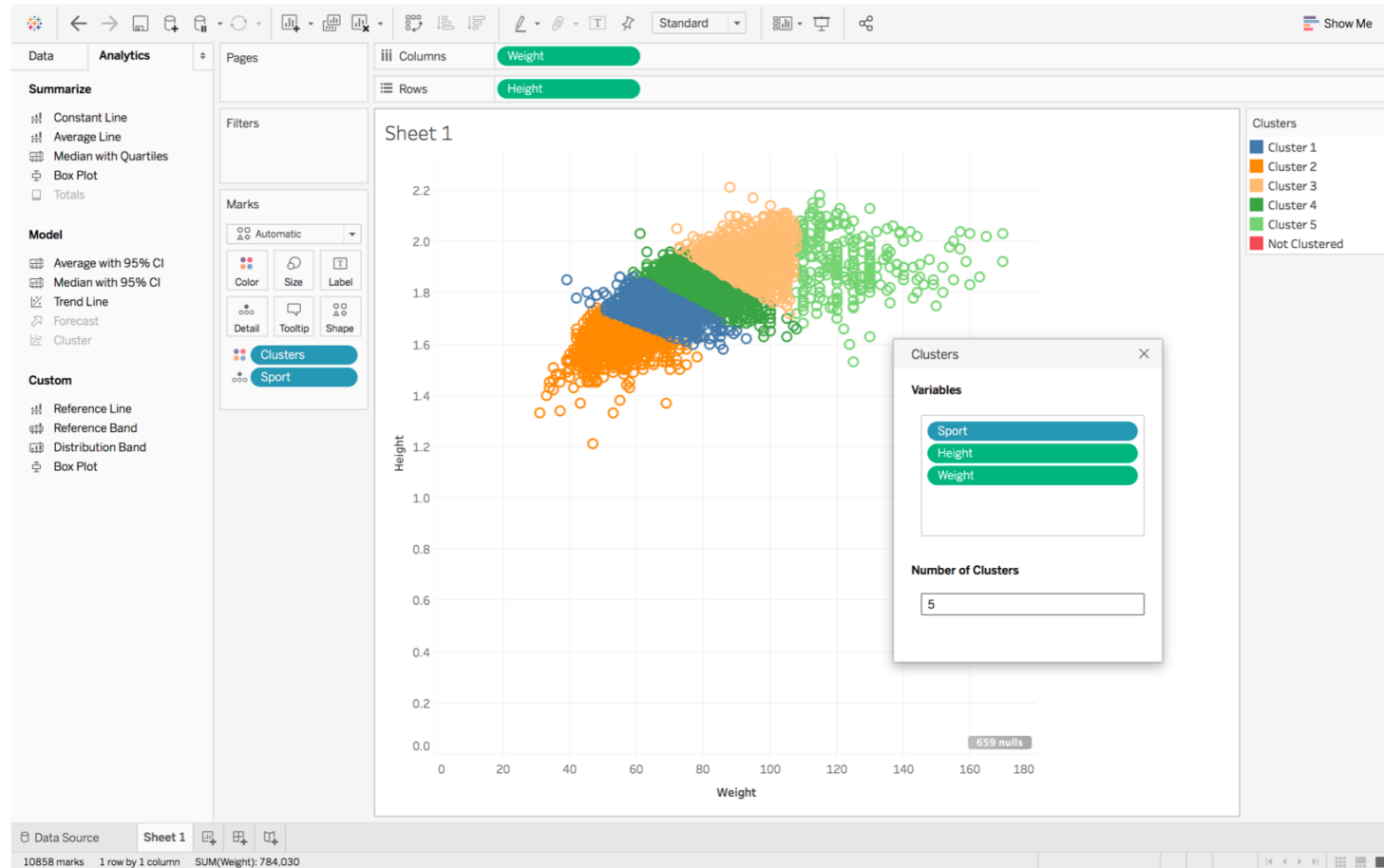
Clustering in Tableau

(Double click on clusters in Analytics tab)



Clustering in Tableau

(Changing number of clusters)



Hierarchical clustering

- Look into the Jupyter notebook: 'Lab Clustering.ipynb'

Tabpy and DB Scan

<https://github.com/tableau/TabPy>

Tabpy Installation (Mac/Linux)

- Create a new conda environment

```
>>conda create --name Tableau-Python-Server python=3.5 anaconda
```

```
>>source activate Tableau-Python-Server
```

```
>>pip install tabpy-server
```

```
>>pip install --upgrade tabpy-server
```

Go to this directory: (For example) /Users/username/anaconda/envs/Tableau-Python-Server/lib/python3.5/site-packages/tabpy_server. You should find setup.sh file, just run it

```
>>./setup.sh
```

let this run..... It is just that you are running your server, don't stop it. (You can stop it by using control+C)

Execute below command in another terminal

```
>>pip install tabpy-client
```


Tabpy

- Look into 'tabpy_clustering.ipynb'

Using breastcancer.csv file

Connections [Add](#)

breastcancer
Text File

Files

☐ Use Data Interpreter
Data Interpreter might be able to clean your Text File workbook.

athletes.csv
breastcancer.csv

[New Union](#)

breastcancer

breastcancer.csv

Connection: ☒ Live ☐ Extract

Filters: 0 [Add](#)

Sort fields: Data source order

☐ Show aliases ☐ Show hidden fields 683 rows

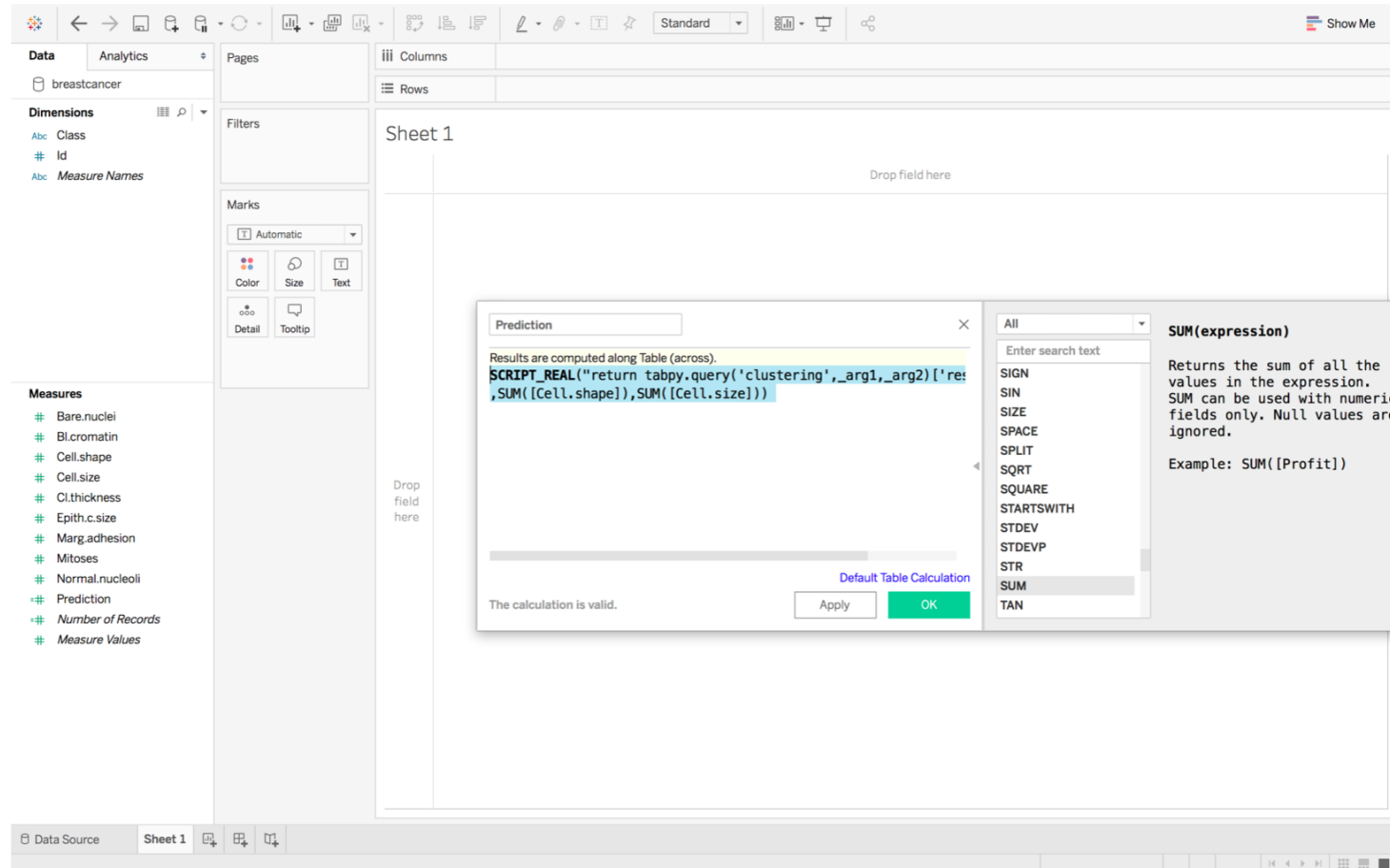
#	#	#	#	#	#	#	#	#	#	#	Abc
breastcanc... Id	breastcancer.csv Cl.thickness	breastcancer.csv Cell.size	breastcancer.csv Cell.shape	breastcancer.csv Marg.adhesion	breastcancer.csv Epith.c.size	breastcancer.csv Bare.nuclei	breastcancer.csv Bl.cromatin	breastcancer.csv Normal.nucleoli	breastcancer.csv Mitoses	breastcancer.csv Class	
1000025	5	1	1	1	2	1	3	1	1	benign	
1002945	5	4	4	5	7	10	3	2	1	benign	
1015425	3	1	1	1	2	2	3	1	1	benign	
1016277	6	8	8	1	3	4	3	7	1	benign	
1017023	4	1	1	3	2	1	3	1	1	benign	
1017122	8	10	10	8	7	10	9	7	1	malignant	
1018099	1	1	1	1	2	10	3	1	1	benign	
1018561	2	1	2	1	2	1	3	1	1	benign	
1033078	2	1	1	1	2	1	1	1	5	benign	
1033078	4	2	1	1	2	1	2	1	1	benign	
1035283	1	1	1	1	1	1	3	1	1	benign	
1036172	2	1	1	1	2	1	2	1	1	benign	
1041801	5	3	3	3	2	3	4	4	1	malignant	

[Go to Worksheet](#)

Data Source **Sheet 1**

Create new calculated field 'Prediction'

```
SCRIPT_REAL("return tabpy.query('clustering',_arg1,_arg2)['response']",SUM([Cell.shape]),SUM([Cell.size]))
```



Prediction field in action

breastcancer

Dimensions

- Class
- Id
- Measure Names

Measures

- Bare.nuclei
- Bl.cromatin
- Cell.shape
- Cell.size
- Cl.thickness
- Epith.c.size
- Marg.adhesion
- Mitoses
- Normal.nucleoli
- Prediction
- Number of Records
- Measure Values

Columns

Rows

Id

Sheet 1

Id	Prediction
63375	-1.000
76389	-1.000
95719	-1.000
128059	-1.000
142932	-1.000
144888	-1.000
145447	-1.000
160296	-1.000
167528	-1.000
183913	-1.000
183936	-1.000
188336	-1.000
191250	-1.000
242970	-1.000
255644	-1.000
263538	-1.000
274137	-1.000
303213	-1.000
314428	-1.000
320675	-1.000
324382	-1.000
324427	-1.000
333093	-1.000
342245	-1.000
352431	-1.000
353098	-1.000
369565	-1.000
378275	-1.000
385103	-1.000
390840	-1.000
411453	-1.000
412300	-1.000
428598	-1.000

Data Source Sheet 1

630 marks 630 rows by 1 column SUM of AGG(Prediction): -630.000