A7 Clustering

Sharad, Shabbir

Sub-Part 1

Design and Implementaion

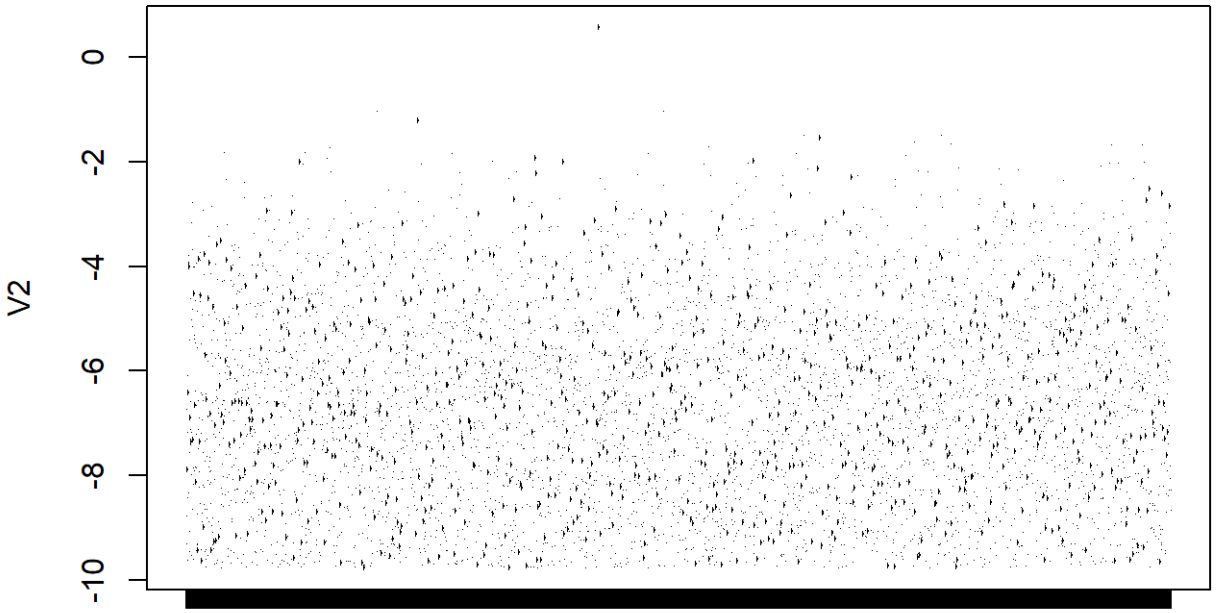
- First we clean the data which has been given to us.
- The cleaned data is stored in a Custom SongRecord Format.
- For each clustering criteria(eg. on the basis of loudness,tempo etc.) we extract the important fields and make a rdd with point object, which is a container class for the cluster points.
- A recursive way is applied on the on the gathered points to compute k-means.
- For k-means of points with 1-dimensions manhattan distance is applied to calculate closeness whereas for 2-dimensional points euclid distance is applied.
- Iterations were performed 10 times.

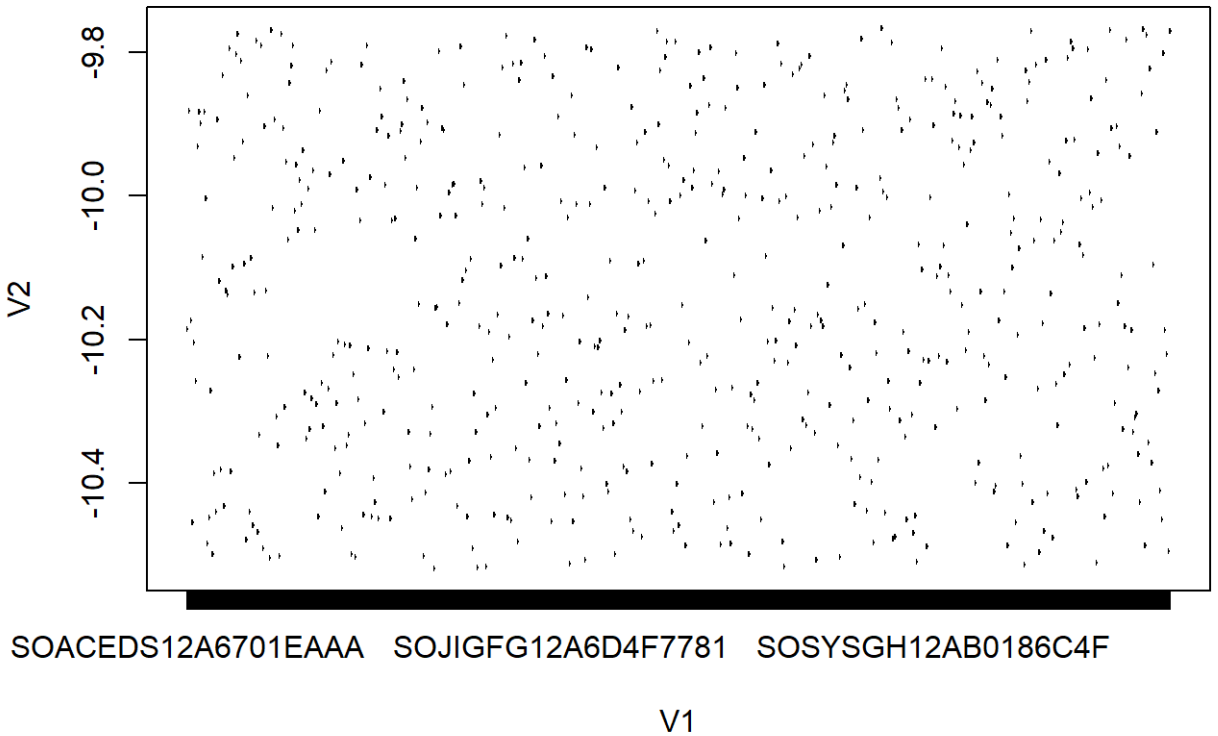
Execution Environment

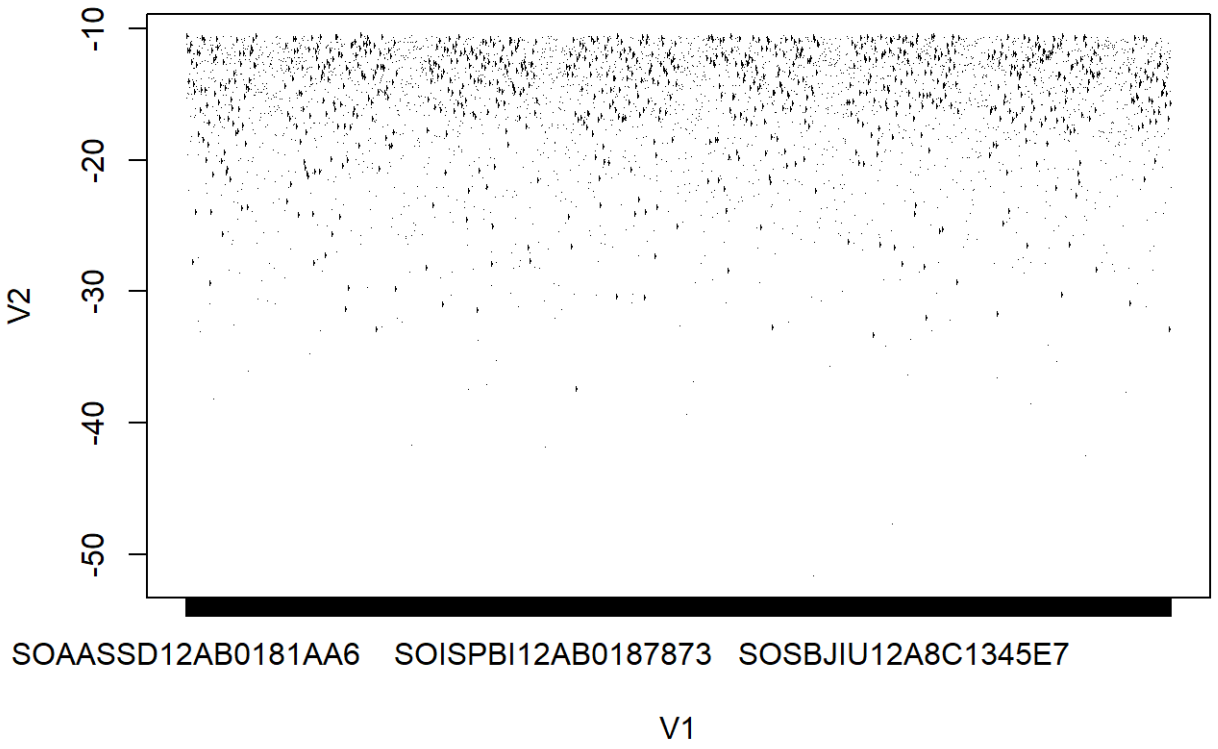
```
OS: Ubuntu 16.04 VM
Processor Name: Intel(R) Core(TM) i7
Processor Speed:    2.70GHz
Number of Processors:   2
Total Number of Cores:  2
L1d cache:  32K
L1i cache:  32K
L2 cache: 256K
L3 cache: 4096K
Memory: 4 GB
SSD: 1 TB
```

Graphical Representation

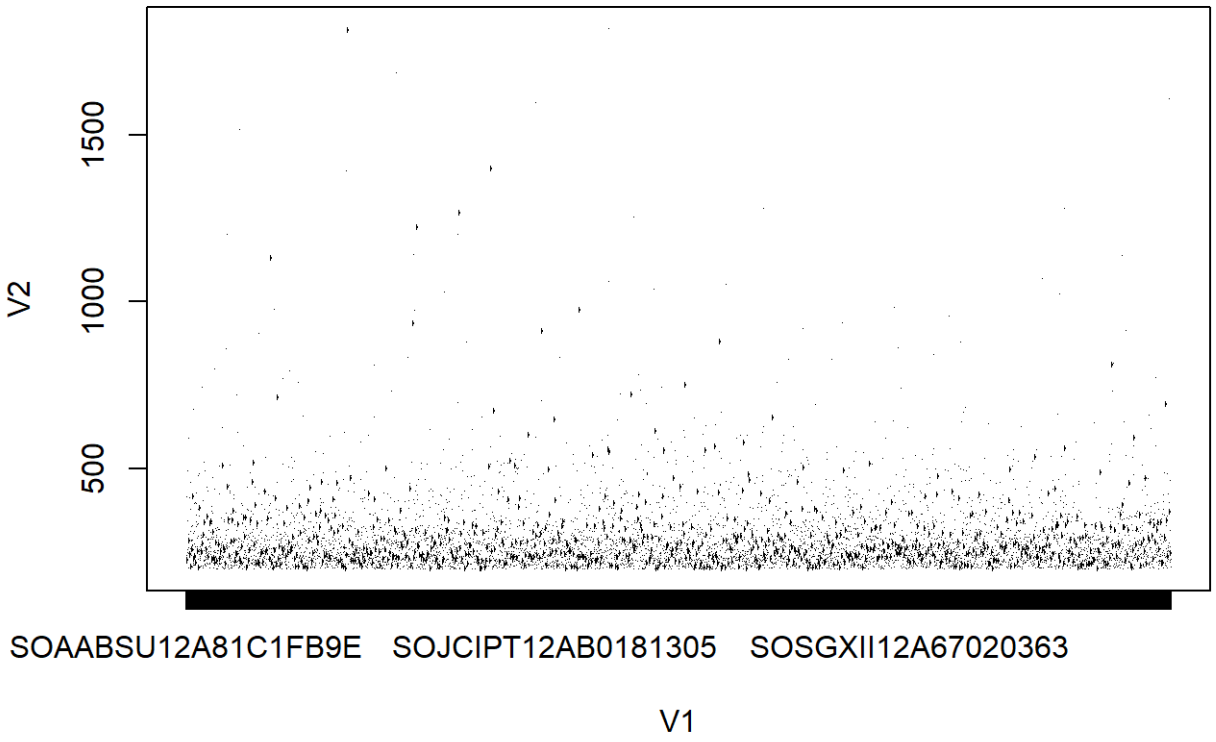
- Frequeancy graph for clusters with Loudness:

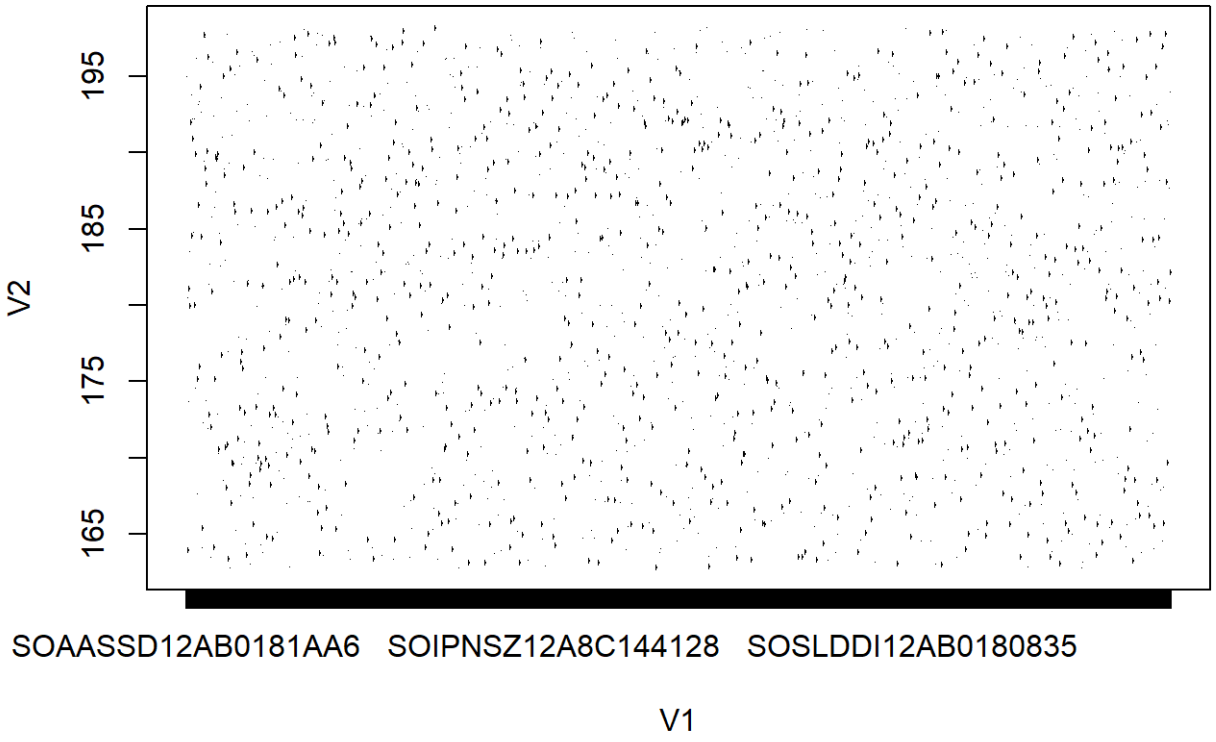

V2
0
-2
-4
-6
-8
-10
SOAAAQN12AB01856D3
SOJCIPT12AB0181305
SOSGXII12A67020363
V1

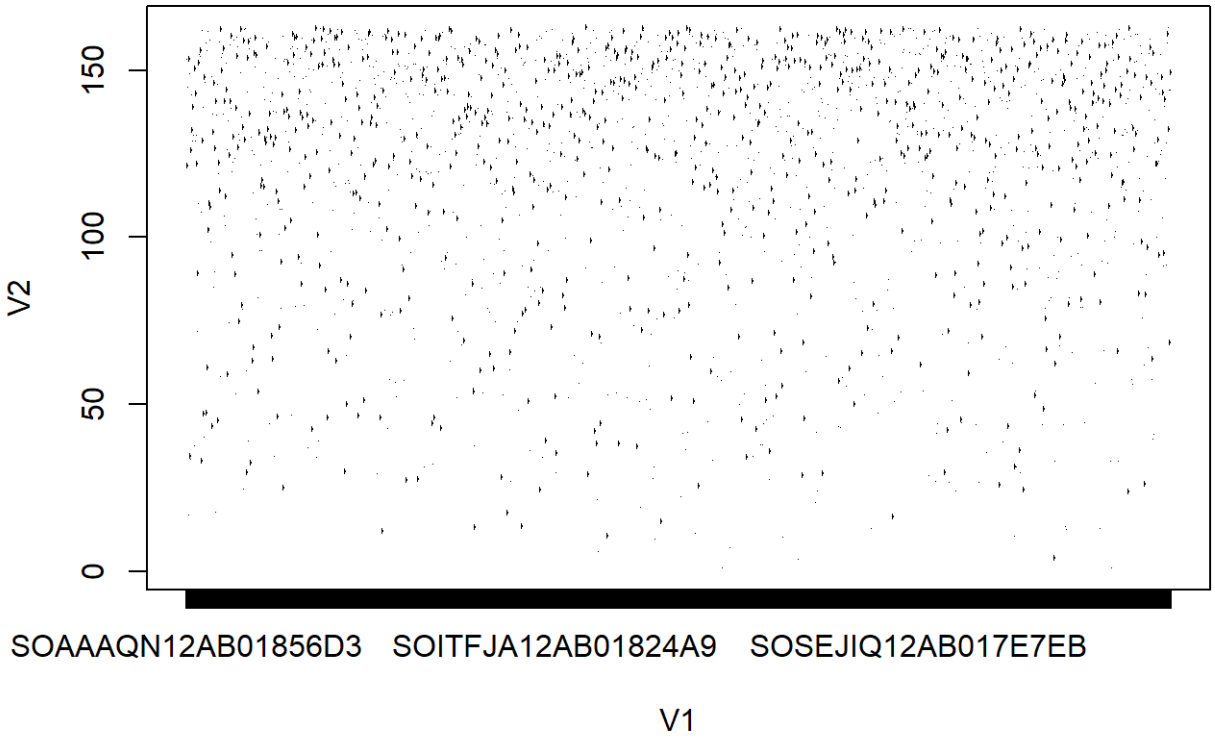

V2
-9.8
-10.0
-10.2
-10.4
SOACEDS12A6701EAAA
SOJIGFG12A6D4F7781
SOSYSGH12AB0186C4F
V1


V2
-10
-20
-30
-40
-50
SOAASSD12AB0181AA6
SOISPBI12AB0187873
SOSBJIU12A8C1345E7
V1

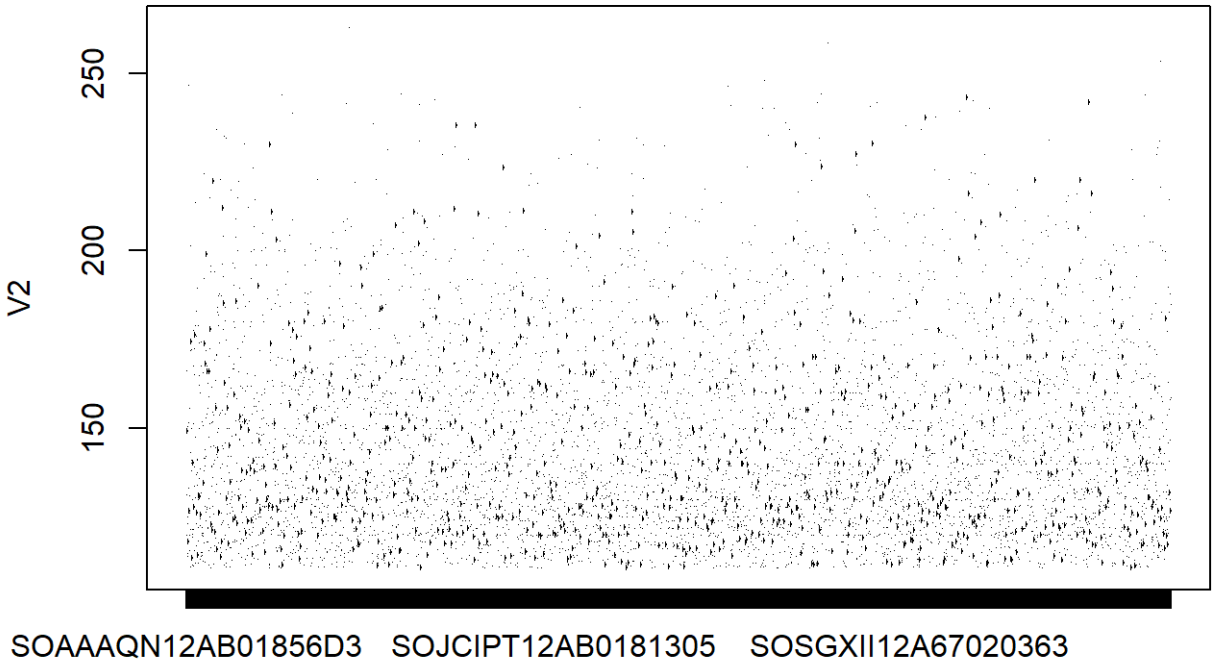
- Frequeancy graph for clusters with Length:

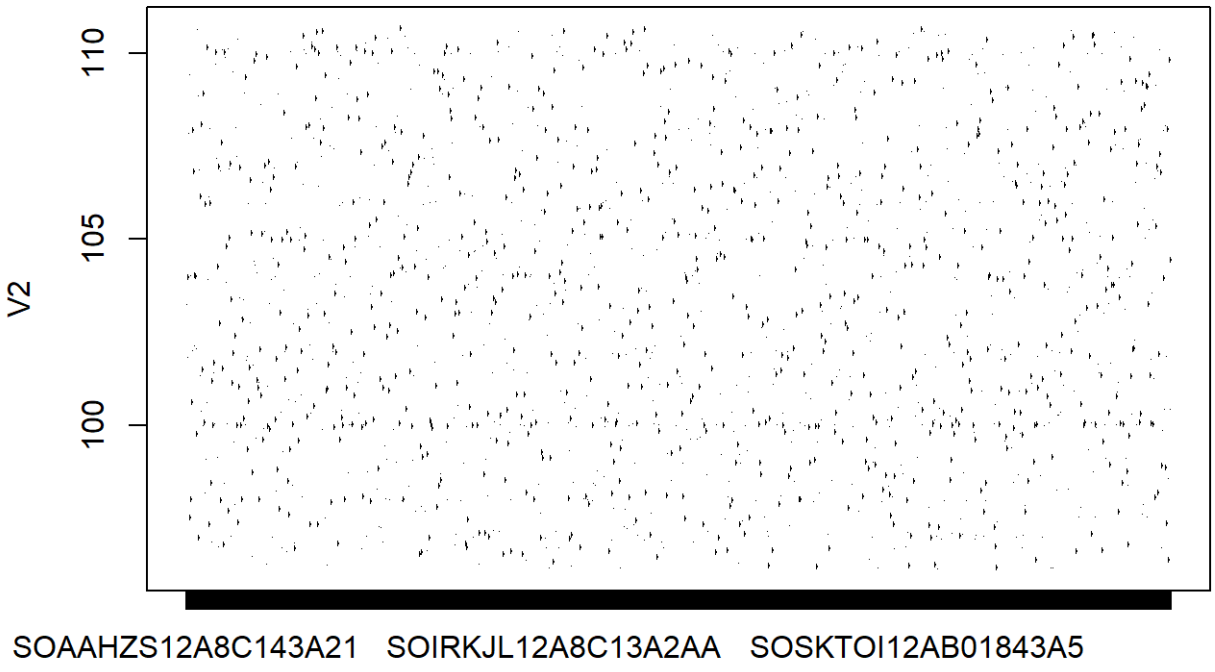


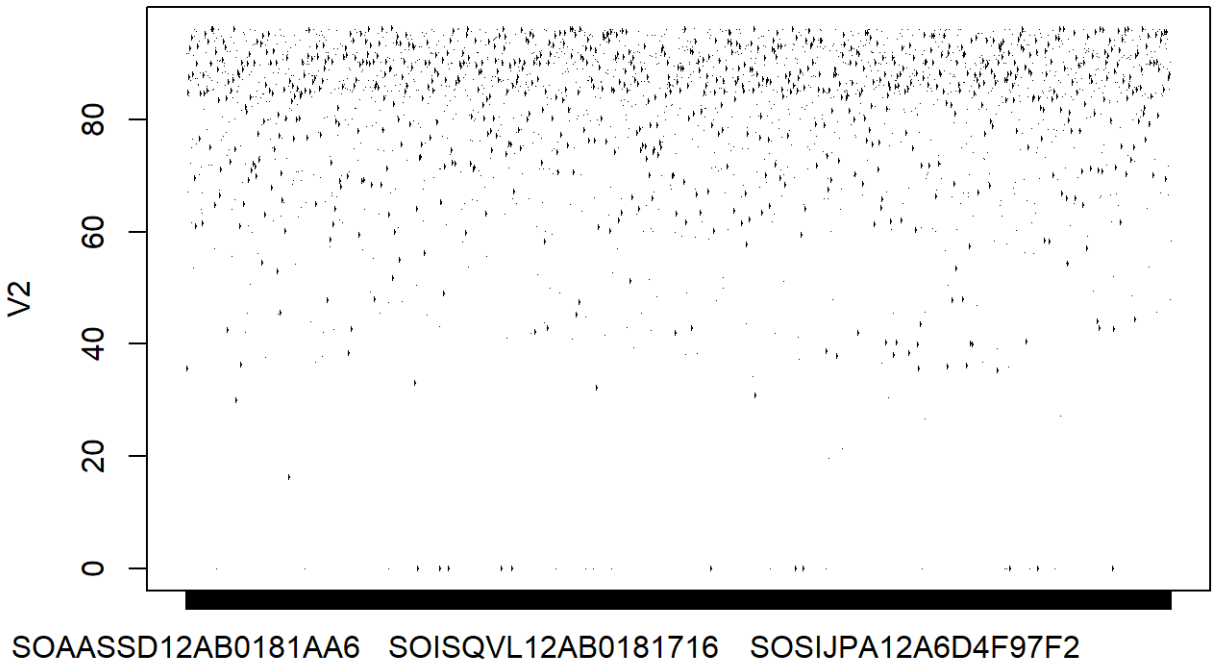




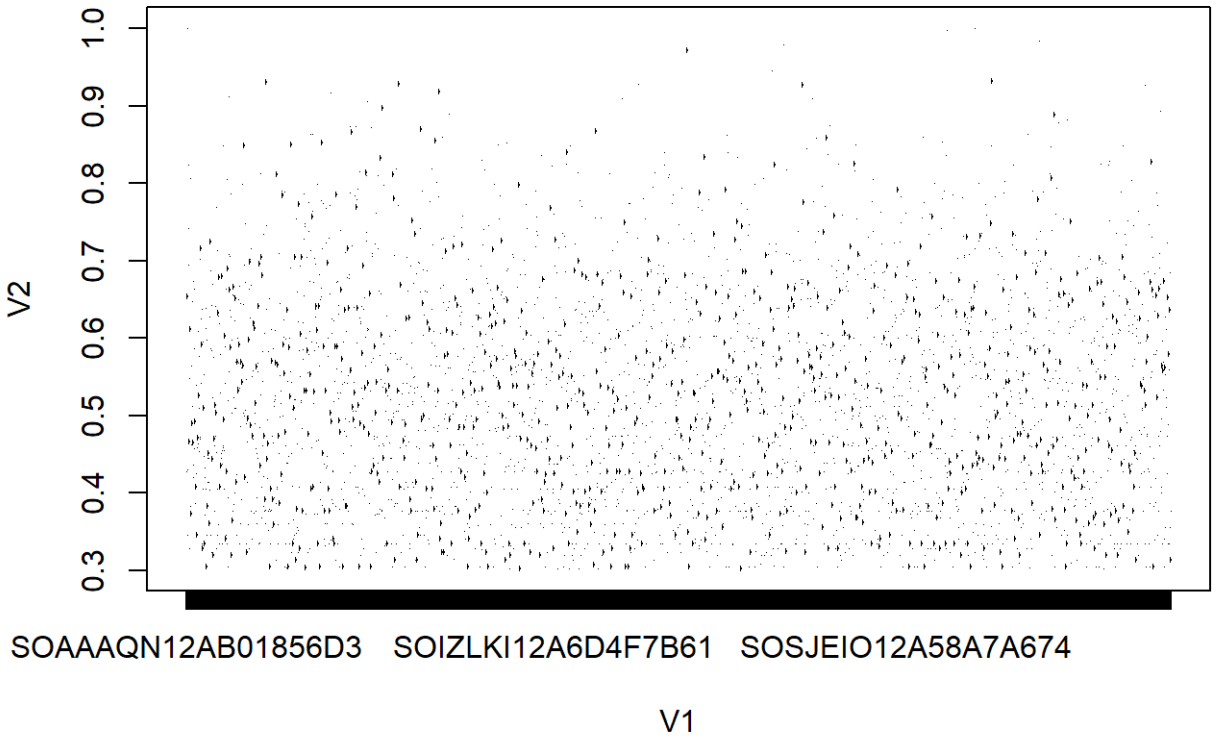
- Frequeancy graph for clusters with Tempo:

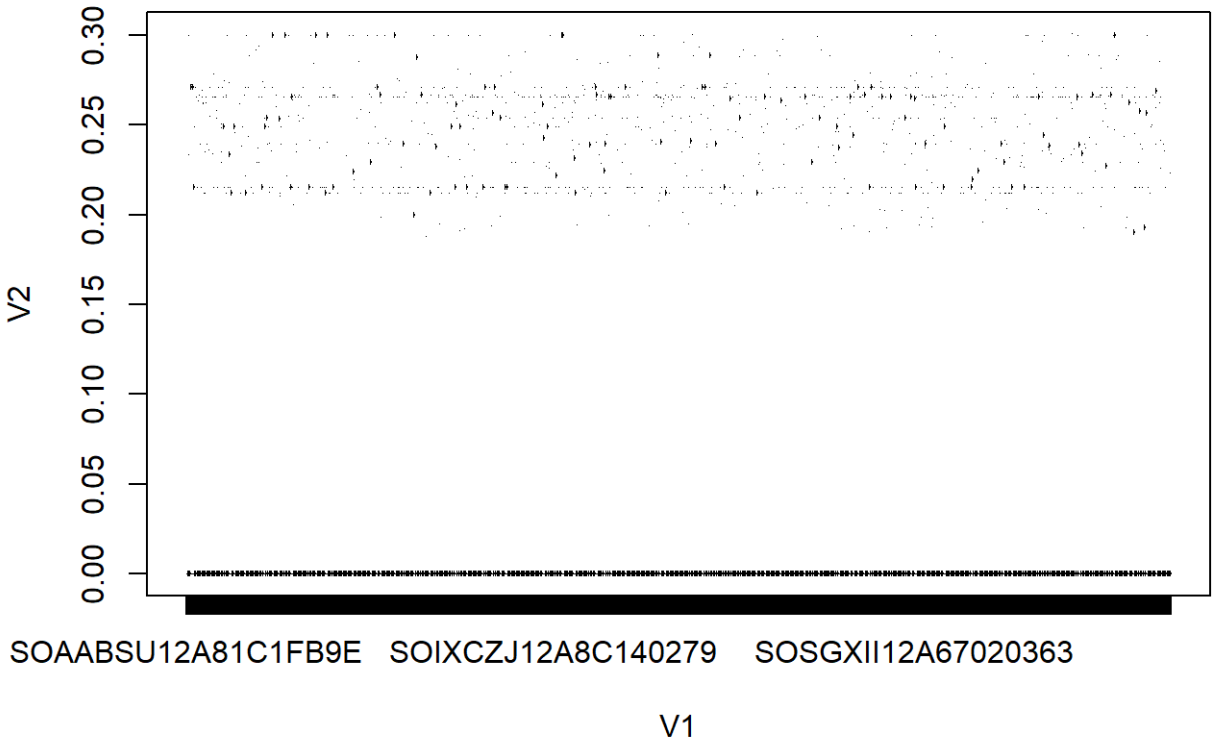

V2
250
200
150
SOAAAQN12AB01856D3
SOJCIPT12AB0181305
SOSGXII12A67020363
V1


V2
110
105
100
SOAAHZS12A8C143A21
SOIRKJL12A8C13A2AA
SOSKTOI12AB01843A5
V1

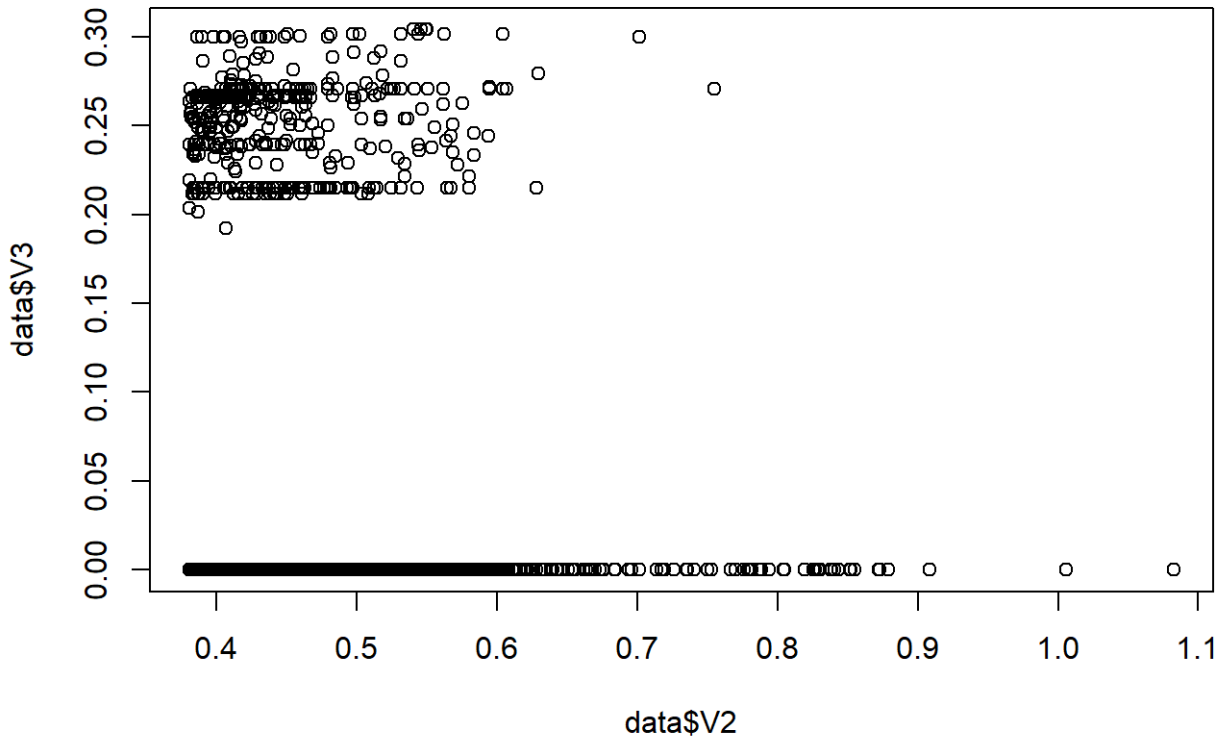


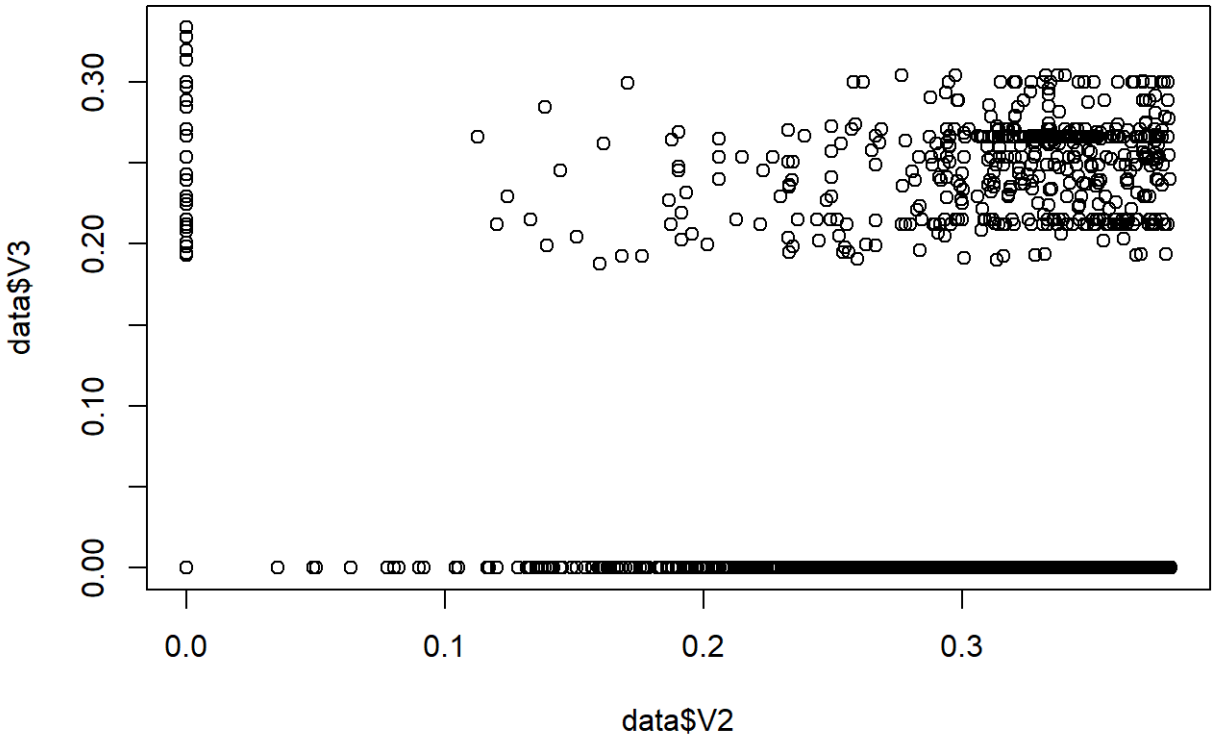
- Frequeancy graph for clusters with Hotness:

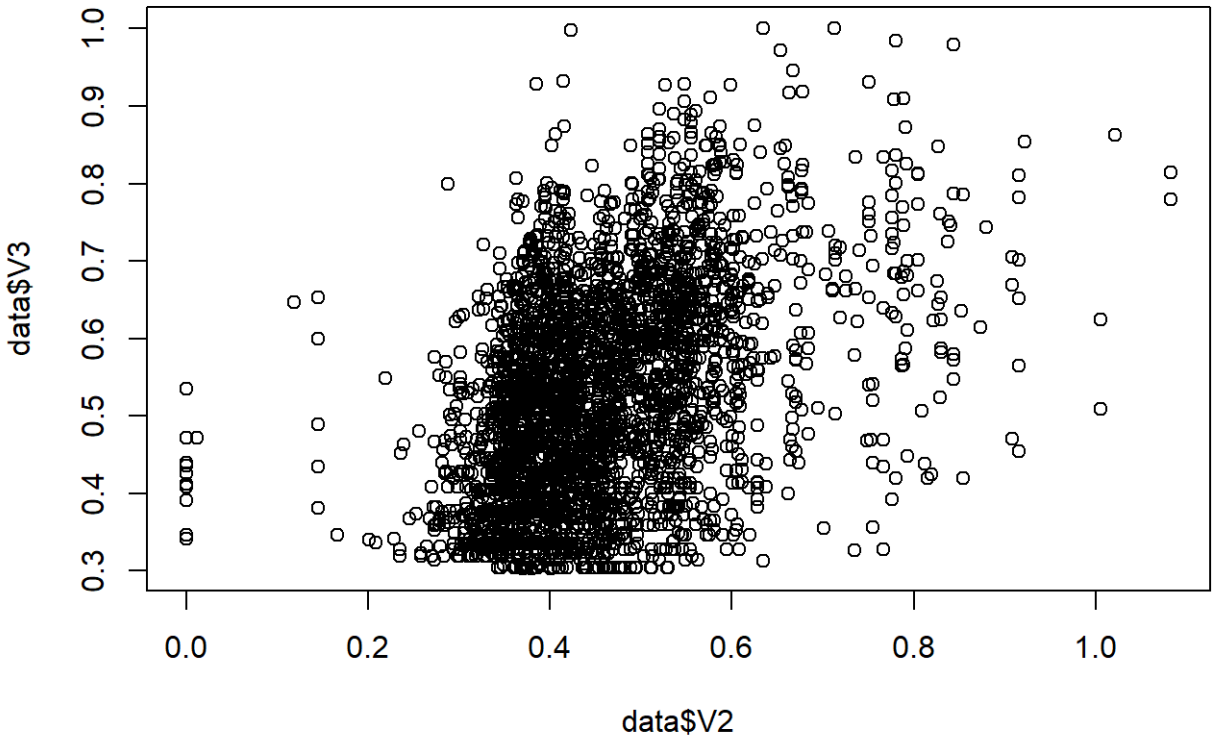




- Frequeancy graph for clusters with Combined Hotness:







Data Observations

Doing an inner join on the supplied clusters of the respective criteria show that Loudness, Length and Tempo decide if the songs combined hotness would be clustered in a particular supplied cluster category. There were 966 obersevartions of the Song ID which depicted this claim.

Also doing the same thing with addition of the song hotness criteria again gave 966 observations.

But if the particular 1 criteria is seen then the combined hotness depends on the following criteria: * Song Hotness with 3219 observations * Song Tempo with 2024 observations * Song Length with 2210 observations * Song Loudness with 20160 observations

Conclusion

Clustering algorithm is quite handy when unlabled data is to be grouped. The clusters formed with various characteristics of the song data were indicative of which songs would be clustered in the combined hotness criteria when done an inner join .As well as there were some indications pertaining to the individual data as well which stated that each criteria contributed to a certain dgree for the clustering in the combined hotness criteria.