DATA CLUSTERING

DOMINIC NEWTON

Plan

- Define Clustering
- What is Data Clustering?
  - List different types
- Practical applications for clustering
- What is Kmeans/ the Kmeans algorithm?
- R studios implementation
  - Unknown
  - Auto MPG
- Analysis & Results
- Conclusion & Summary

WHAT IS DATA CLUSTERING?

- IT MAY BE DEFINED AS A GROUPING OF SIMILAR OBJECTS IN A DATASET.
- NO FREE LUNCH THEOREM: NO ONE METHOD OF CLUSTERING WILL WORK FOR EVERY CASE.

Different types of data clustering

- Connectivity-based clustering (hierarchical clustering)
- Centroid-based clustering
- Distribution-based clustering
- Density-based clustering

Data clustering in the real world

- Biology
- City planning
- Earthquake studies
- Insurance
- Libraries
- Marketing
- The world wide web

The kmeans algorithm

- Voronoi cells
- Iterative refinement technique

Two steps of kmeans

- The assignment step
- The update step

THE ASSIGNMENT STEP

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\},$$

THE UPDATE STEP

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

The local optimum

- Does using the kmeans algorithm guarantee that an optimum will be found?

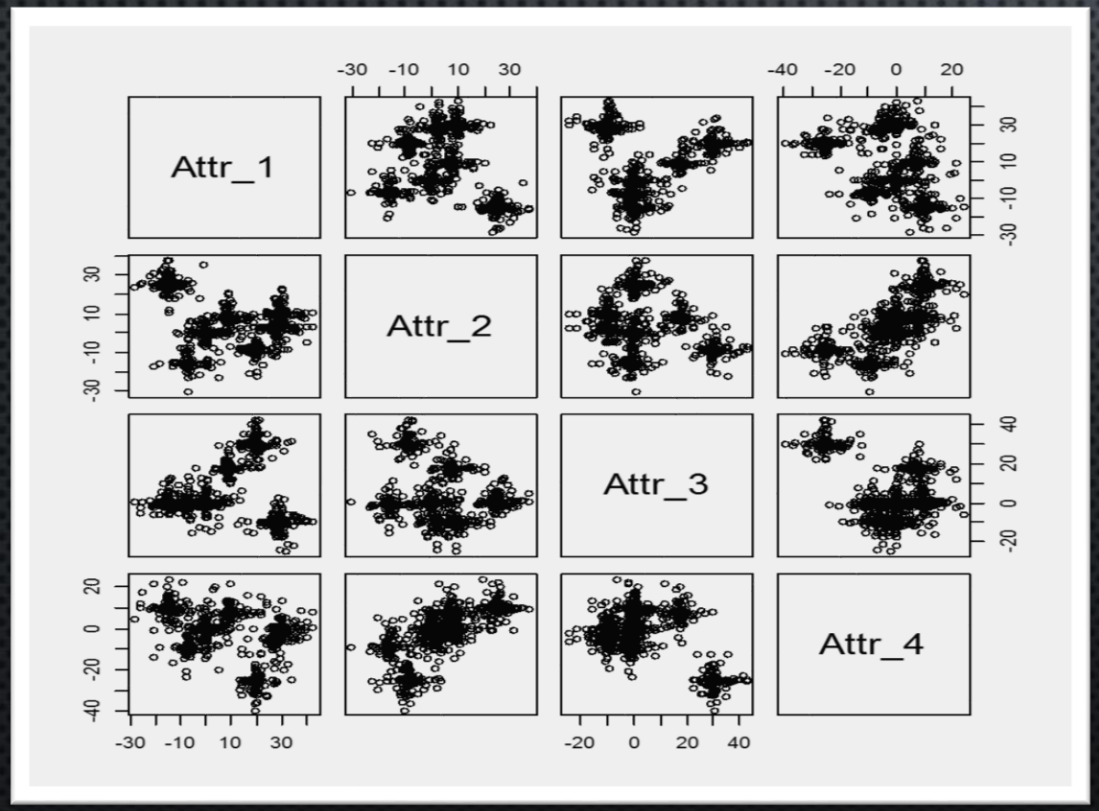
R STUDIOS AND KMEANS IMPLEMENTATION

- R IS A VERY USEFUL DATA MINING TOOL.

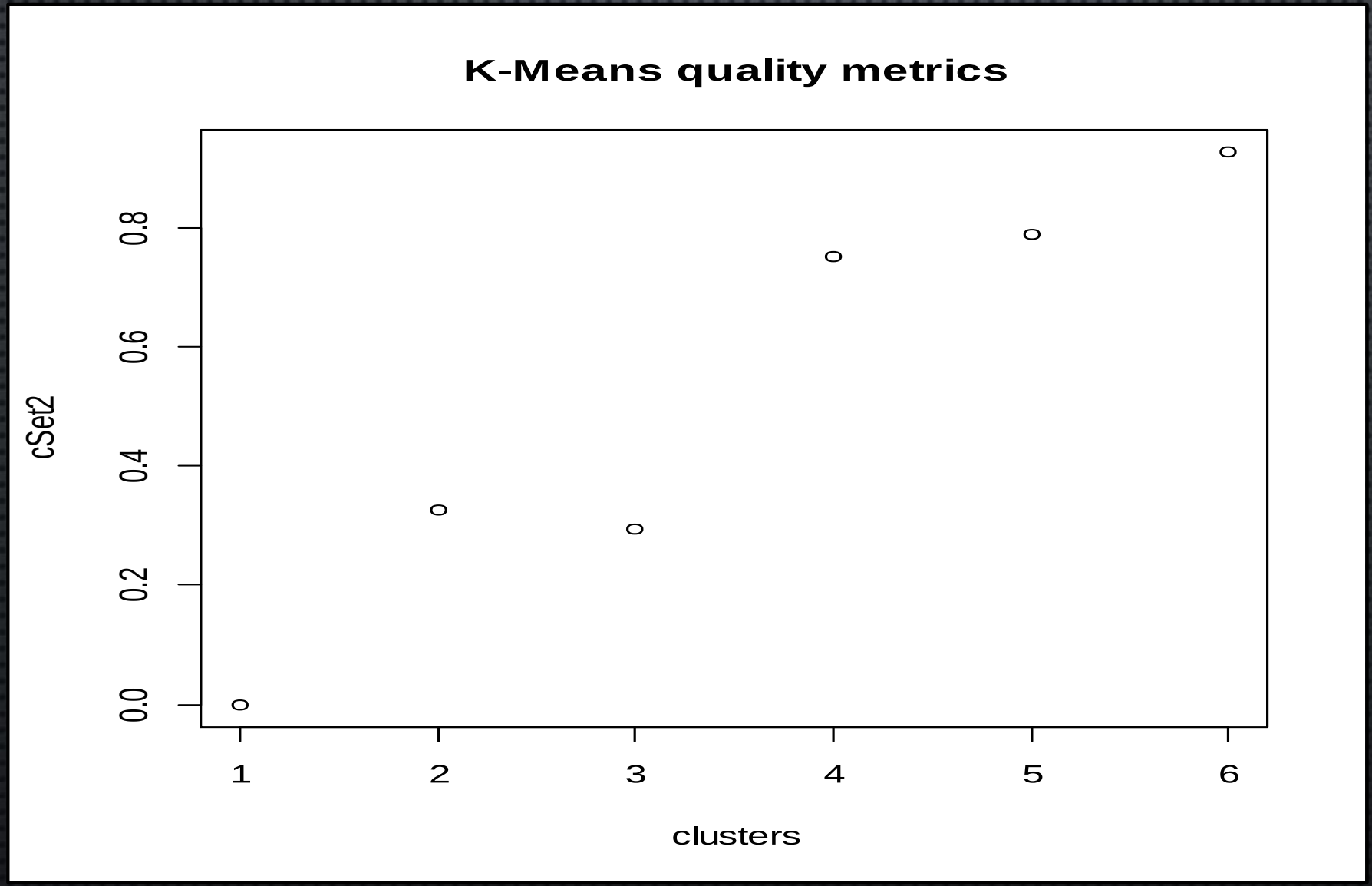
VIEW THE DATA IN R

- > str(clusterdata)
-
- 'data.frame': 827 obs. of 4 variables:
- $ Attr_1: num 11.158 27.671 -0.931 35.427 -15.052 ...
- $ Attr_2: num 6.4979 3.1331 0.0299 1.777 24.0318 ...
- $ Attr_3: num 17.706 -11.325 -0.183 -7.041 -0.968 ...
- $ Attr_4: num 7.21 -5.57 -1.42 3.72 10.03 ...

PLOT THE DATA



KMEANS QUALITY METRIC



FIND THE CENTERS

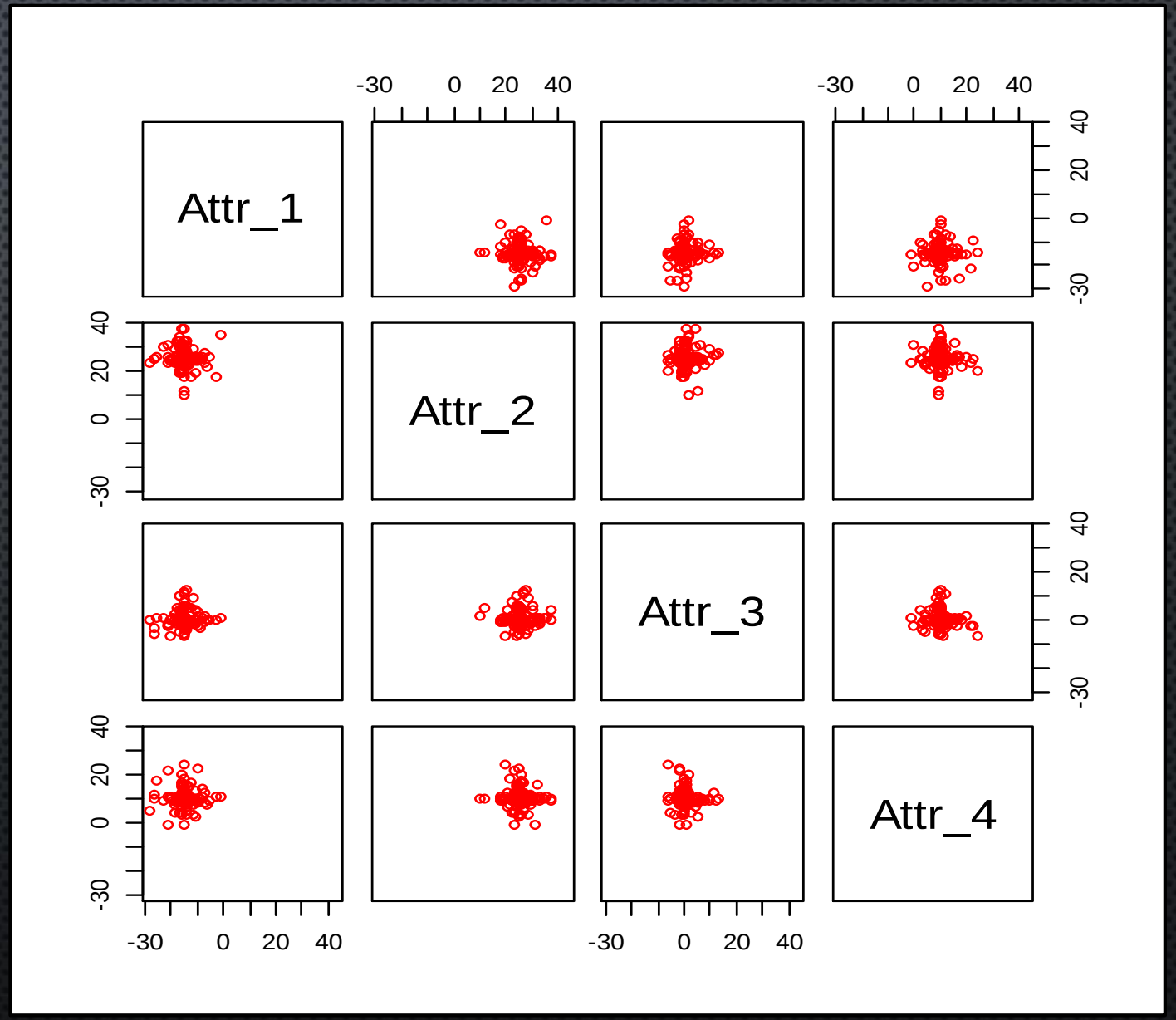
- PLOT(CLUSTERDATA)
-
- > CLUSTERDATA.6MEANS <- KMEANS(CLUSTERDATA, CENTERS = 6)
-
- > #K-MEANS CLUSTERING ABOVE
-
- > #SHOW THE CENTERS
- >
- > CLUSTERDATA.6MEANS$CENTERS
-
-
-
-

	ATTR_1	ATTR_2	ATTR_3	ATTR_4
1	-14.924504	25.369178	0.5038865	9.995796
2	-3.297940	-7.638349	-0.8349637	-4.348270
3	23.865671	-7.724028	34.2889051	-23.681296
4	19.079821	-9.144853	29.5052470	-25.695706
5	9.112023	7.603083	16.9043424	7.455988
6	29.209084	6.257228	-10.2067502	-2.278471

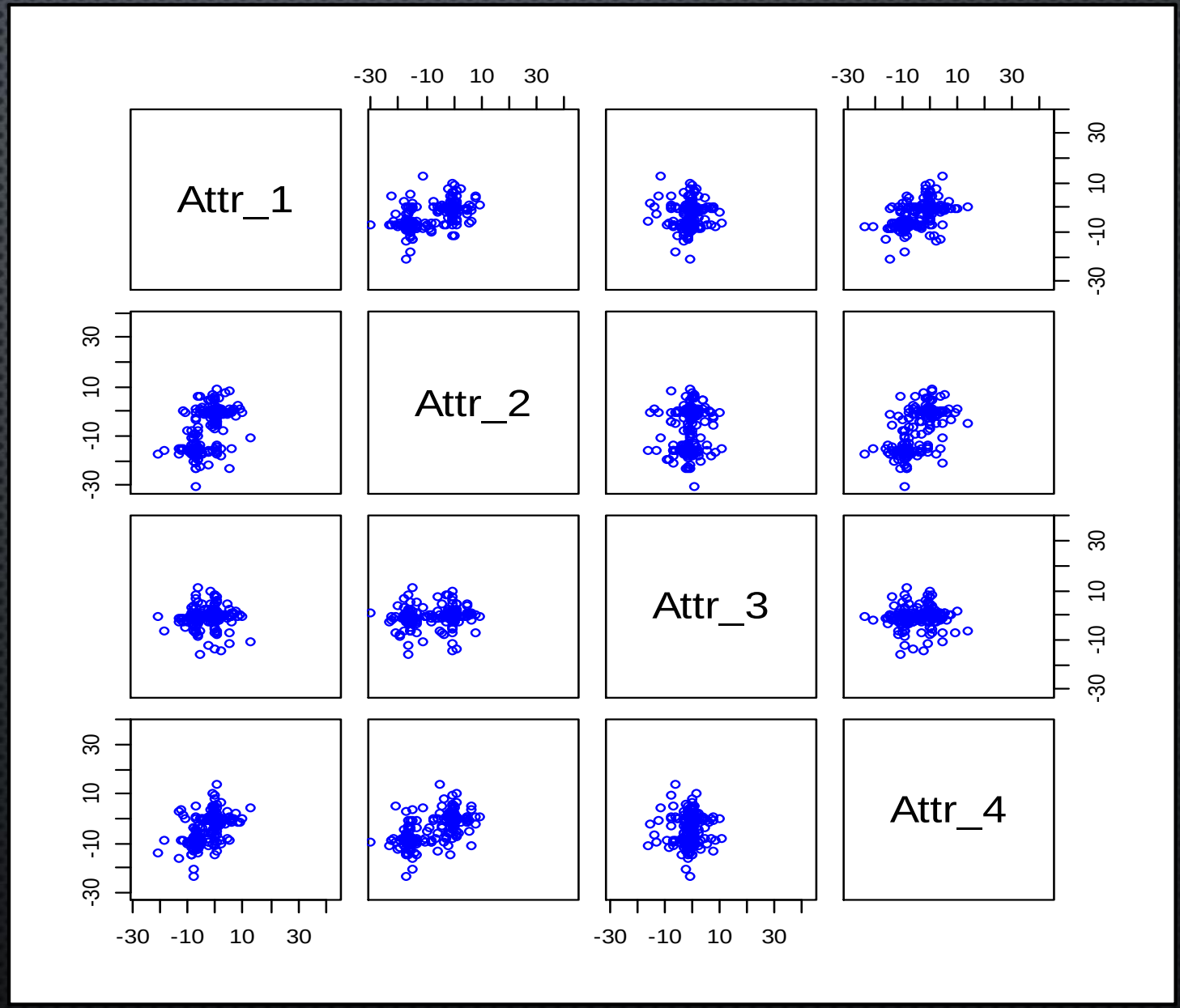
Separate the estimated clusters

- Use the centers and the kmeans algorithm along with the plot function to show the separation in the data

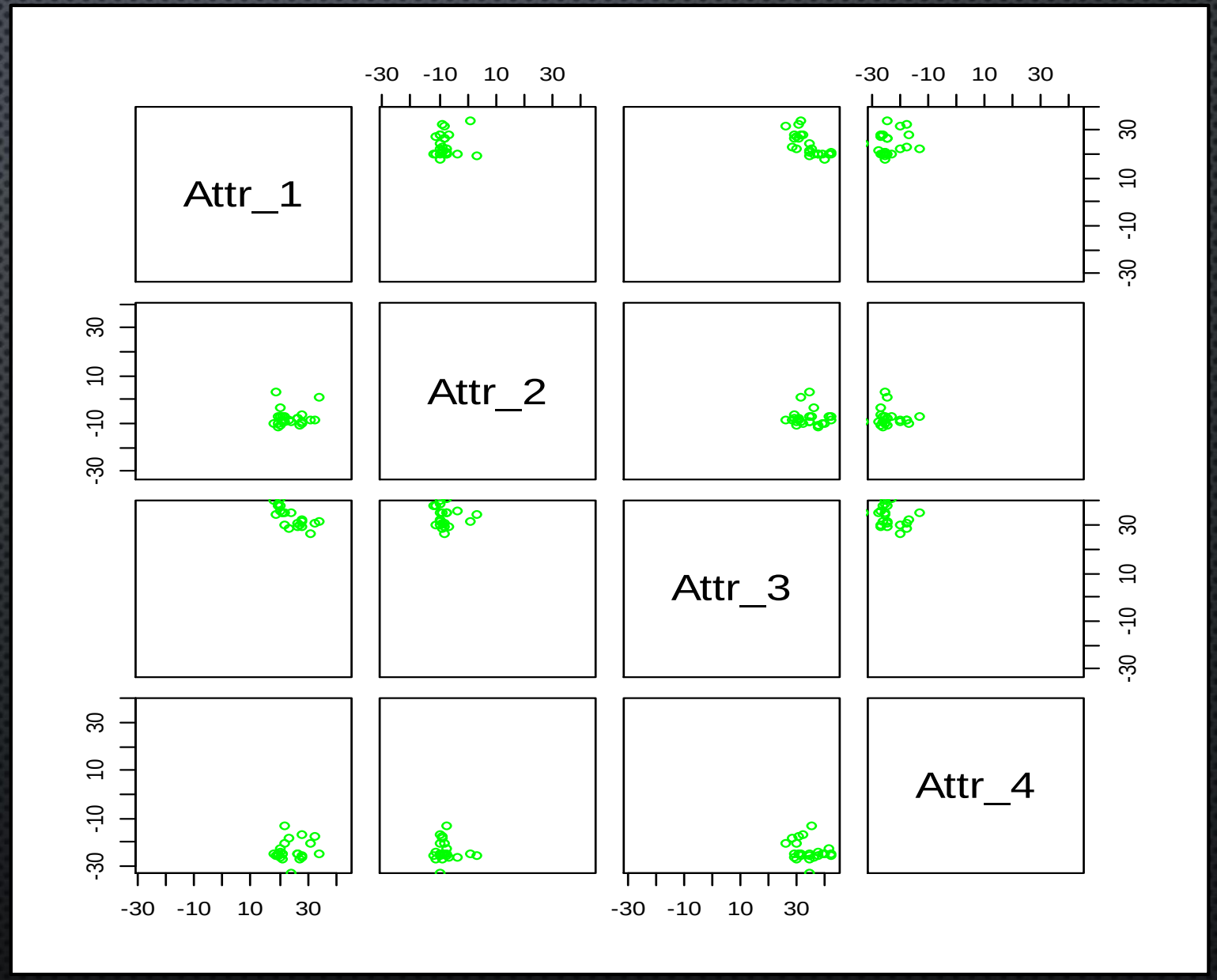
CLUSTER 1



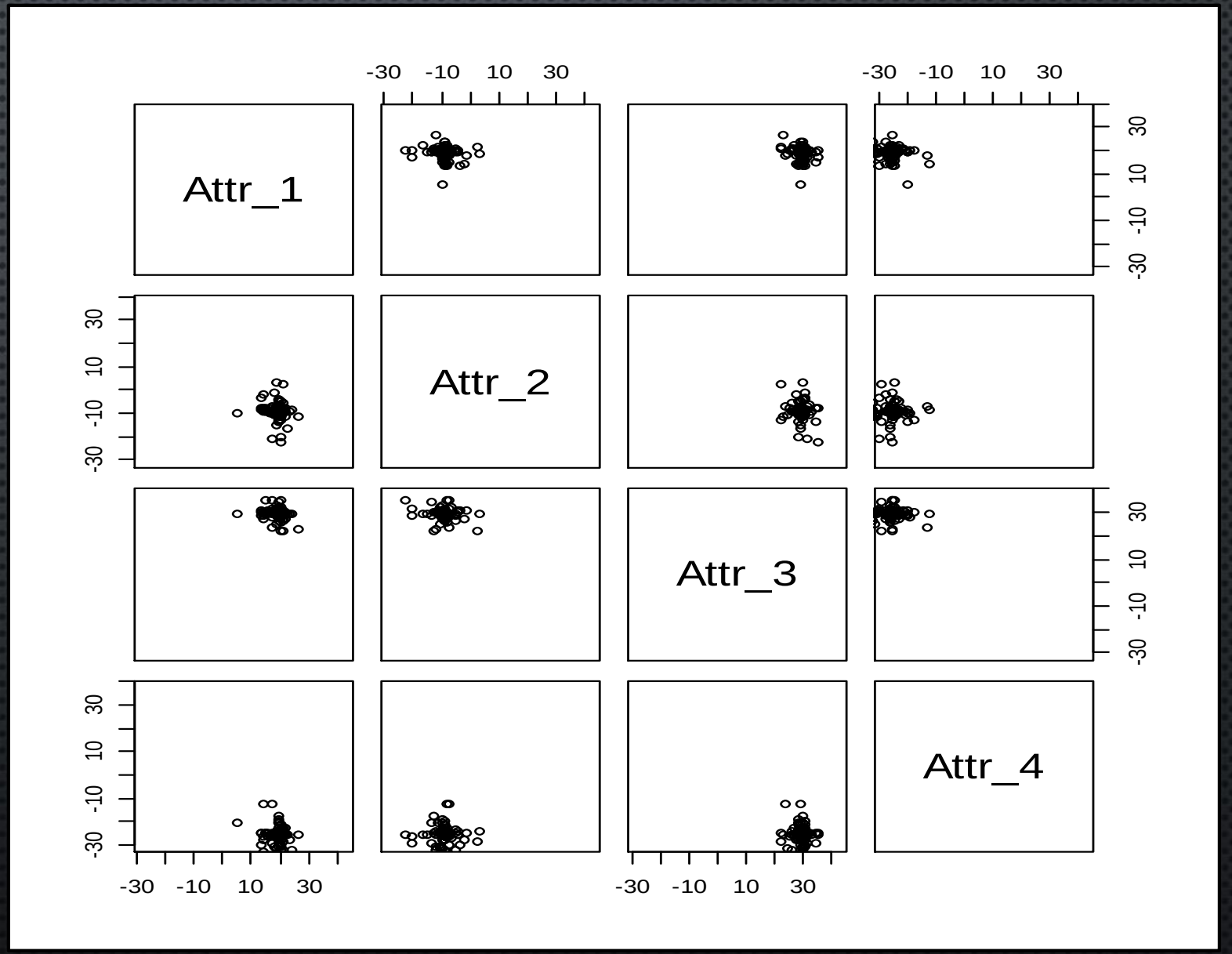
CLUSTER 2



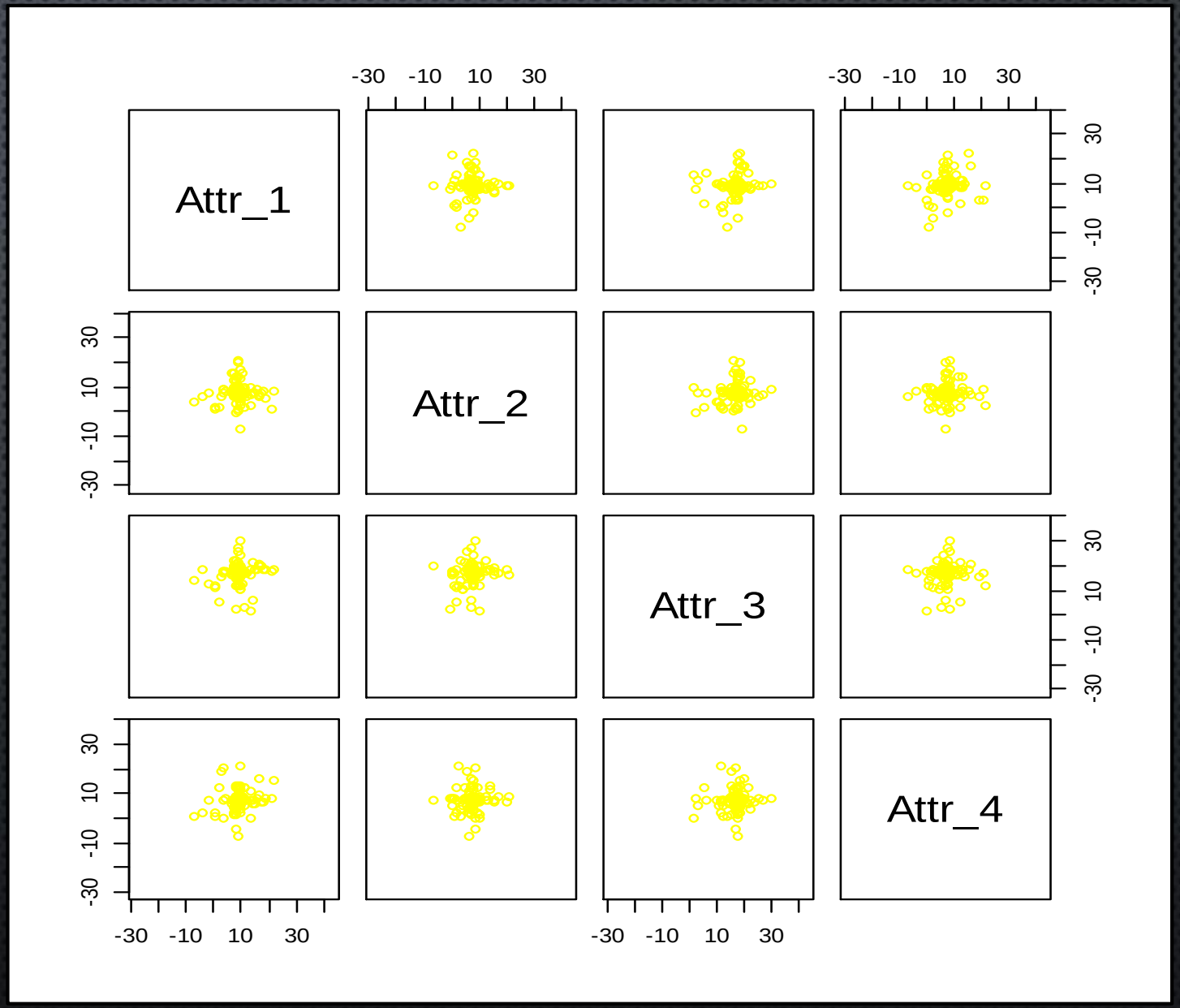
CLUSTER 3



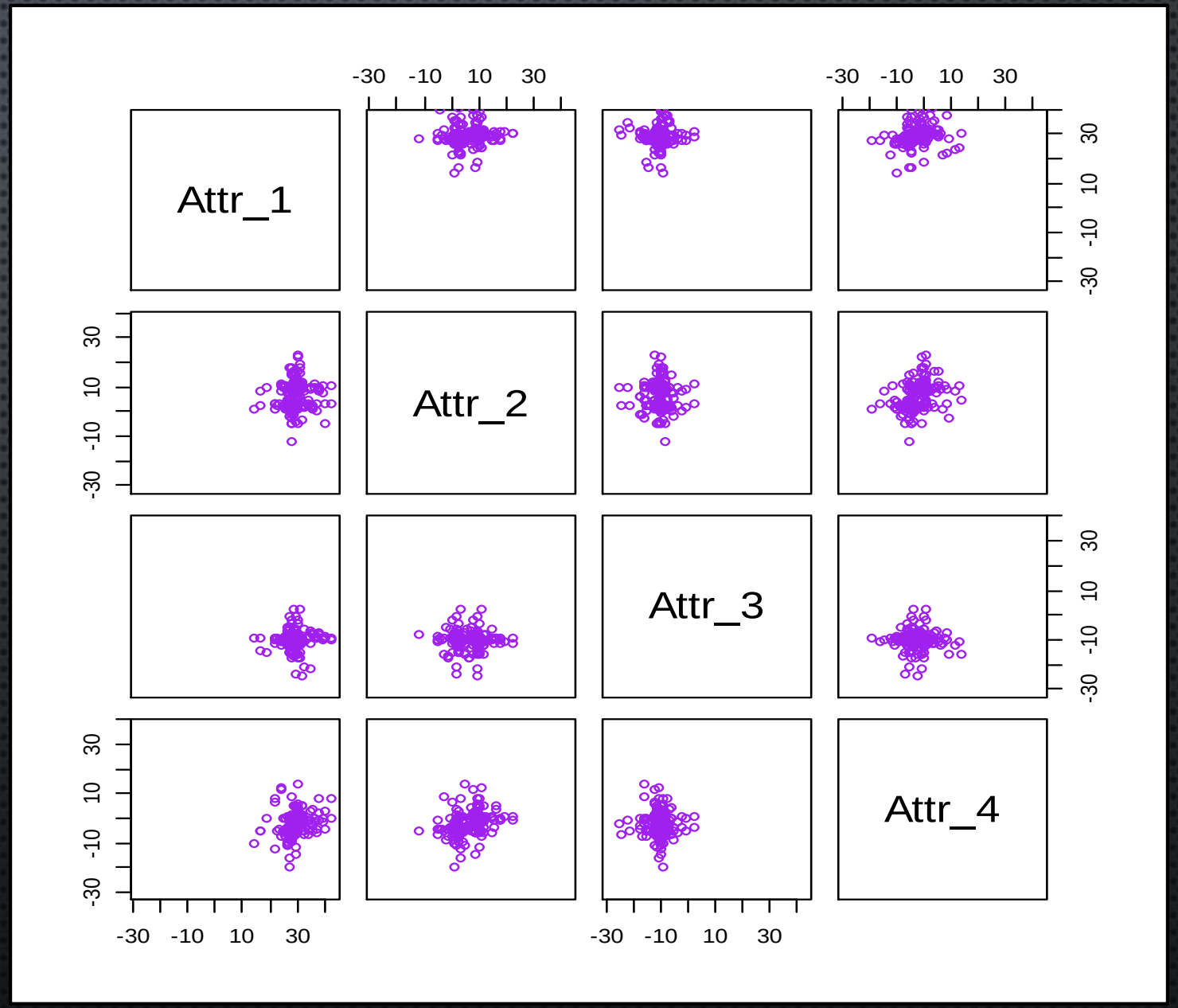
CLUSTER 4



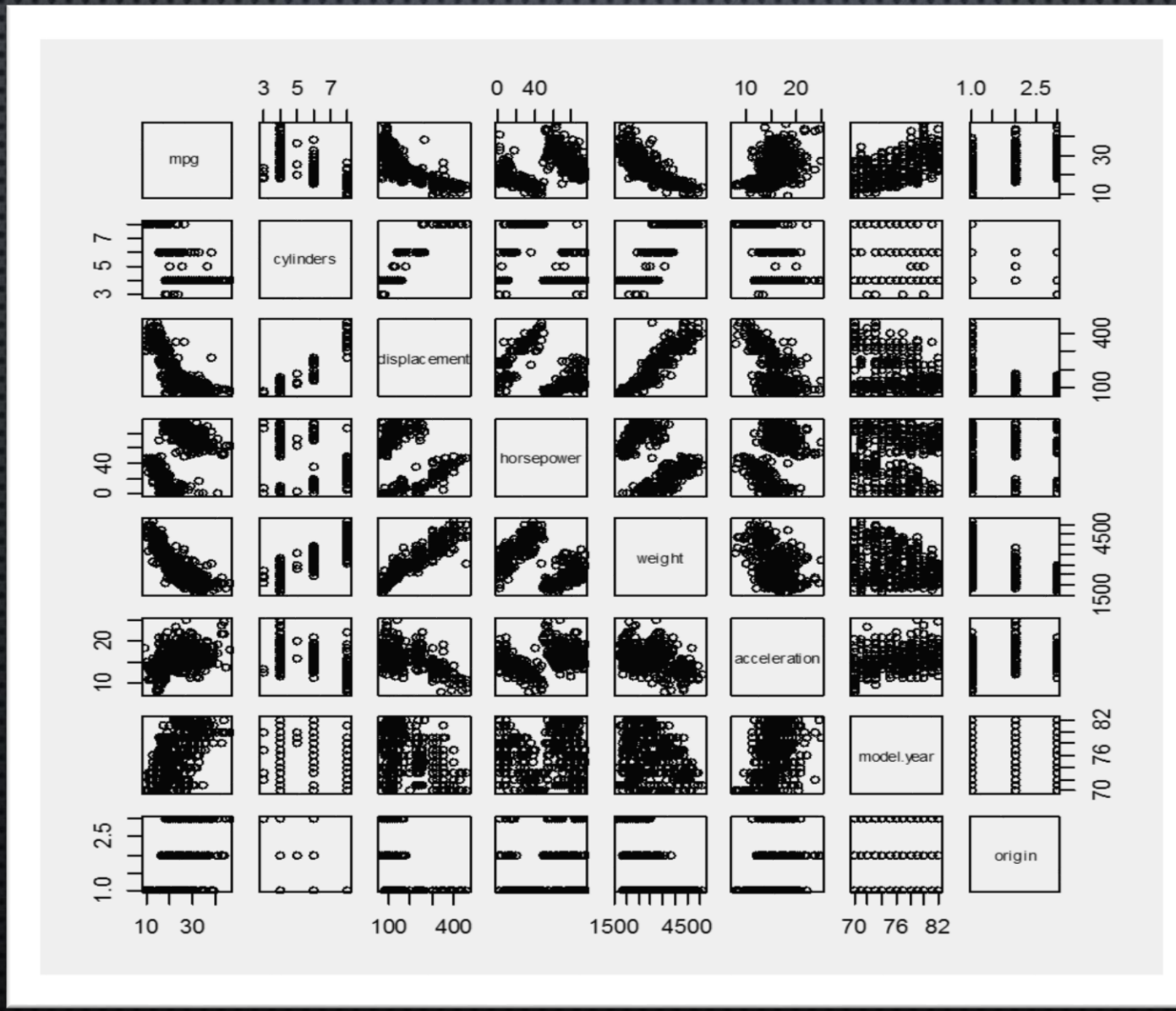
CLUSTER 5



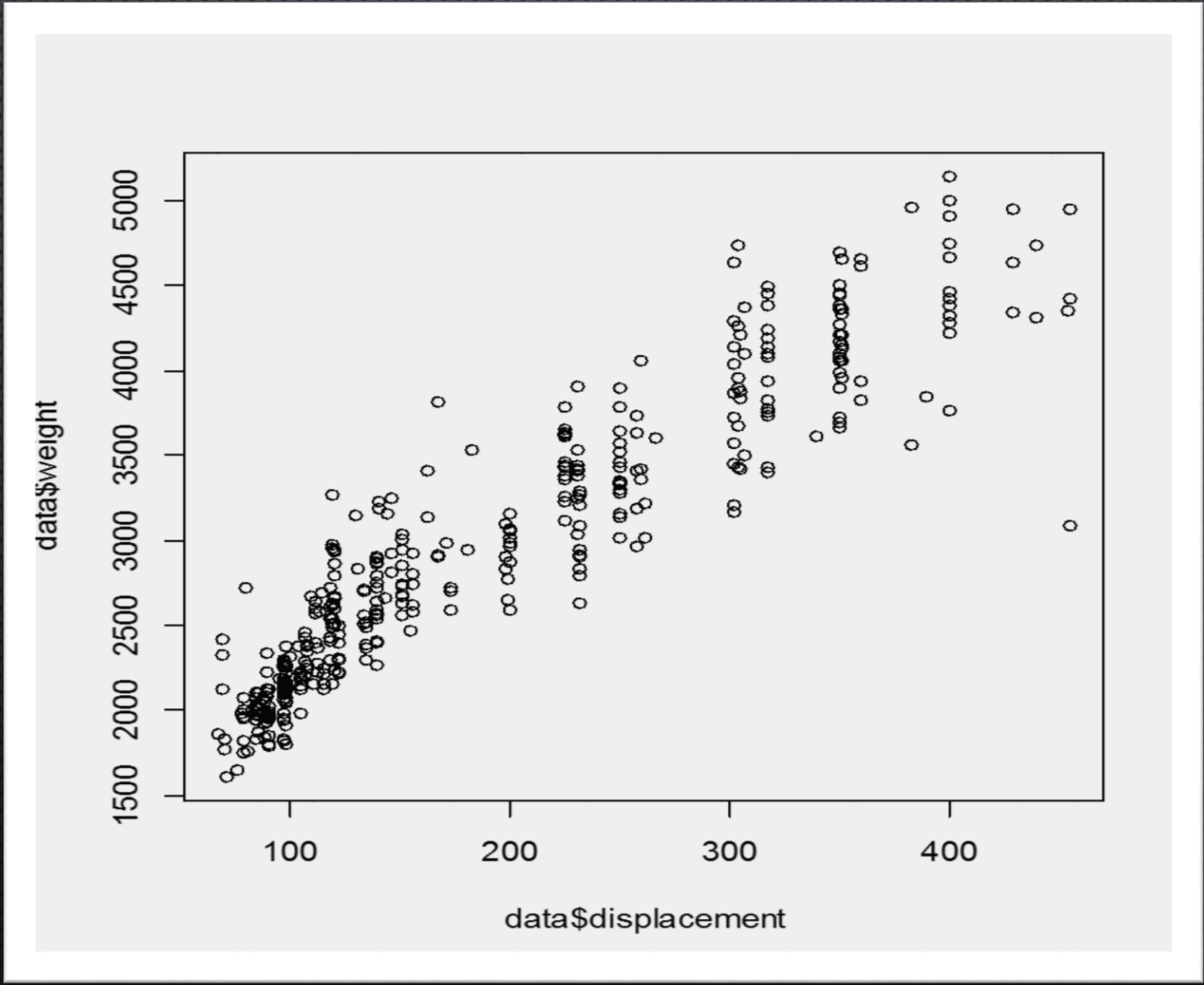
CLUSTER 6



AUTO MPG DATASET



EXAMPLE OF VARIABLE ISOLATION



CONCLUSION

- NO FREE LUNCH!
- THERE MAY BE A BETTER WAY TO WORK WITH THE DATA
- KMEANS IS USEFUL IF YOU ALREADY HAVE AN INCLINATION TO THE NUMBER OF CLUSTERS IN THE DATA SET.

Summary

- Define Clustering
- What is Data Clustering?
  - List different types
- Practical applications for clustering
- What is Kmeans/ the Kmeans algorithm?
- R studios implementation
  - Unknown
  - Auto MPG
- Analysis & Results
- Conclusion

REFERENCES

Deza, E. (2016, October 20). *Euclidean distance.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Euclidean_distance

MacQueen, J. B. (2016, October 20). *k-means clustering.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/K-means_clustering

Stewart, J. (2016, October 20). *Local optimum.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Local_optimum

Tryon, R. C. (2016, October 20). *Cluster analysis.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Cluster_analysis

Voronoi, G. (2016, October 20). *Voronoi diagram.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Voronoi_diagram

Chrysler. (2016, December 7). *Engine Displacement.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Engine_displacement

THANK YOU!

Are there any questions?