Expository Clustering Visualizations: Keeping it Simple

Greg Page(✉)

Boston University, Boston, USA
gpage@bu.edu

Abstract. In this paper, the authors present a very basic overview of k-means clustering, before using statistical summaries to demonstrate how such a model separates records in a dataset into distinct groups.

The author then shows how simple visualizations can effectively "tell the story" behind a clustering model, to include the key distinctions that tend to differentiate one group from another.

The author then explores Principal Component (PC) plots, a tool often misused by analysts seeking to convey information about the clusters identified by their models. Such plots are based not on original variables from the data, but upon linear combinations of those variables. While PC plots are colorful and impressive-looking, their meaning often eludes the students who use them in end-of-semester project presentations.

PC plots serve some value as a diagnostic tool for kmeans modelers; however, these plots should not be used in an expository way by someone who wishes to convey the main findings of a clustering model. Instead, boxplots, scatterplots, barplots, and histograms can much more effectively convey the major takeaways for such a model.

Keywords: data mining · k-means clustering · computer science education

1 Introduction

Clustering, a form of unsupervised learning, is commonly taught in data mining and marketing analytics courses. While clustering can be done with many distinct methods, each approach boils down to the same basic principle – placing the observations into distinctive groups, in a way that maximizes within-group similarity as well as between-group difference. [1] In marketing analytics, clustering models identify specific customer personas, which are often characterized by pithy descriptions such as "Single and Carefree" or "Cruising through the Golden Years."

To communicate the results of a clustering model, the modeler may rely on several methods. Among these are: Presenting the groups with descriptive labels, backed by qualitative statements; delivering per-group summary statistics, often starting with the group means for each of the variables used in the model; and using visualizations to offer insights about the model and its clusters.

T. Zlateva and G. Tuparov (Eds.): CSECS 2023, LNICST 514, pp. 48–57, 2023.
https://doi.org/10.1007/978-3-031-44668-9_4

In our experience, students are too quickly drawn to Principal Component (PC)-based plots, which can serve a diagnostic purpose during model-building time, but are simply ineffective for explaining the meaningful differences from cluster to cluster. Students may encounter such plots in course material, or through online searches, and feel that these must somehow be the right way to "show" their clustering model.

In order to explain the key distinctions among clusters, analysts should instead rely on simple visualizations, based on original variables from the dataset.

In this paper, we will first build a clustering model, using the kmeans() function from the R language.

2 Building a k-means Clustering Model in R

The dataset *portland_families.csv* contains simulated info about 15,000 households in the vicinity of Portland, Maine.

After isolating the dataset's numeric variables, we will be build a model using the following variables: total_ppl (total people per household); square_foot (square footage of primary household residence); household_income (estimated household income from previous year); number_pets (estimate of number of household pets); entertainment_spend_est (estimated total household entertainment spending from previous year); travel_spend_est (estimated total household travel spending from previous year); and under_12 (number of household members under the age of 12).

Given the different scales, and the different units of measurement associated with these variables, we will scale the original values into z-scores before building the model with the kmeans() function from R. [2] After some iteration through various possible k-values, we decided to build this model with six clusters.

After building the model, the questions that naturally arise include, "What does this really mean? What can we learn from this model?" As we show immediately below, a segmentation model can be assessed with per-cluster summary stats (Fig. 1).

```
> model$centers
    total_ppl square_foot household_income number_pets entertainment_spend_est travel_spend_est   under_12
1  1.60477638  -0.2408347     0.0007955225  0.04655520              0.06949739       -0.0647511 -0.3433841
2 -0.02651803   0.8298155     0.1096830284 -0.96622633              0.24382259        0.6811863  0.2695893
3 -0.90492036  -0.2672496     0.0913109899 -0.06254586             -0.24377267       -0.1216444 -1.2068592
4  0.05062574  -0.6012920     0.5479376714  0.02546871             -0.70504780       -0.4143889  0.7417900
5 -0.14956028  -0.6940049    -0.8784297372 -0.02902584              0.57304837       -0.8002412  0.2774678
6 -0.10135708   0.8720494     0.0487035289  0.99291182              0.14749940        0.6530828  0.2011637
```

Fig. 1. Per-Cluster Summary Stats, as Centroid Values

This table with per-cluster centroid values helps us to identify distinguishing features from group to group [3].

For instance, the output above shows us that Cluster 1 stands out for having the largest number of people per household. Their homes are slightly smaller than the dataset average, and they have the second-lowest average number of children under 12. Perhaps many of its members are college students, or recent graduates who share apartments with other twentysomethings.

Cluster 5's members have the lowest incomes, yet their entertainment spending ranks the highest among all groups. Meanwhile, they have the smallest residences and the lowest travel spending. Perhaps this cluster includes bored retirees, or maybe just some die-hard entertainment fans. Either way, marketers can use the centroid values shown above to assist with the tailored marketing approaches that they may wish to use for each segment.

The centroid values may not appeal to all audiences, though. Visualizations tend to be more eye-catching and memorable than descriptive summary stats for most audiences and in most contexts – and clustering is no exception.

While visualizations are both effective and appropriate for expressing the key differences among groups in a clustering model, some types are more effective than others. Before demonstrating the superiority of simple visualizations for conveying group differences in clustering models, we will first show an alternative plot type.

3 So What are those Principal Component Plots Showing, Exactly?

To generate PC plots for a k-means clustering model, we will use the fviz-cluster() function from the factoextra package [4].

For a k = 3 solution for the *portland_families* dataset, this function renders the image shown below.

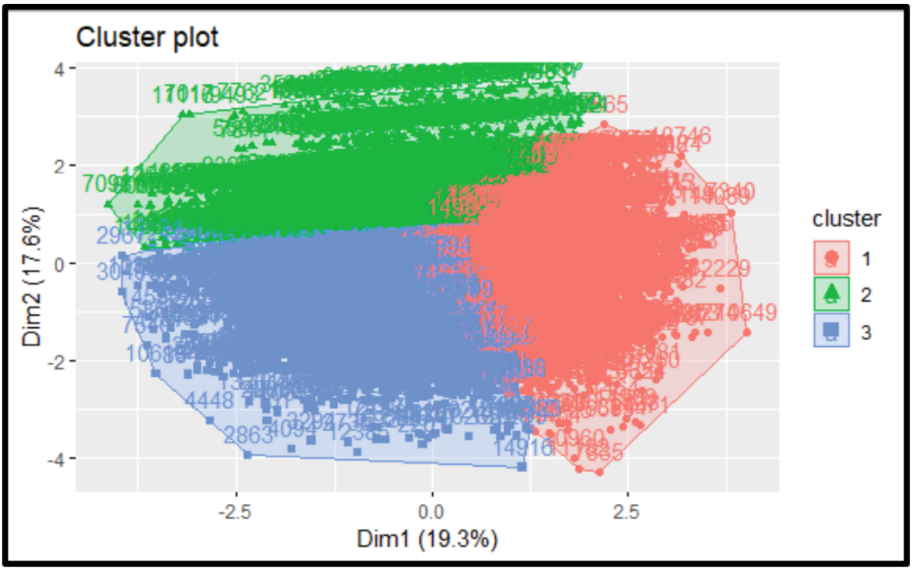

Fig. 2. Principal Component Plot for k = 3 Solution

In the plot shown above, the x-axis is labeled "Dim1" while the y-axis is labeled "Dim2." Dim1 refers to the first PC, which explains 19.3% of the variation among the

values in the dataset. "Dim2" refers to the second PC, which captures a further 17.6% of the variation among the input variables.

This can be verified by calling the prcomp() function on the matrix of standardized values that went into this clustering model. The results shown below indicate that seven PCs are needed to capture all of the variation among the variables in this data. Note also that collectively, the first two PCs based on this data explain less than 40% of the total variation (Fig. 3).

```
> summary(pc)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6     PC7
Standard deviation     1.1627 1.1114 1.0039 0.9994 0.9877 0.8734 0.81729
Proportion of Variance 0.1931 0.1764 0.1440 0.1427 0.1394 0.1090 0.09542
Cumulative Proportion  0.1931 0.3696 0.5135 0.6562 0.7956 0.9046 1.00000
```

Fig. 3. Principal Component Stats for the Portland families dataset

As for the PCs themselves, each one can be viewed as a 7×1 array. In the plot shown above, the labeled points correspond to the product of that record's standardized values, multiplied by each of those first two PCs.

We can see the values for each of the seven PCs below (Fig. 4).

```
> pc
Standard deviations (1, .., p=7):
[1] 1.1626579 1.1113566 1.0039463 0.9994144 0.9876935 0.8734305 0.8172860

Rotation (n x k) = (7 x 7):
                              PC1         PC2         PC3         PC4         PC5         PC6         PC7
total_ppl              0.086230040 -0.70097276  0.031685065  0.007936918  0.04868452 -0.70506029 -0.025582015
square_foot            0.686941731  0.08043302 -0.007052183 -0.060218475  0.11881583  0.03697669 -0.708859021
household_income       0.195923697  0.01223611 -0.181059808  0.633001660 -0.72536683 -0.03986776  0.015618877
number_pets            0.009382622 -0.01082705 -0.710073168 -0.609170398 -0.34907700 -0.05115935  0.005498643
entertainment_spend_est 0.119523481 -0.04343833  0.674296162 -0.471601831 -0.54949695  0.04288193  0.054386511
travel_spend_est       0.679139790  0.09624475 -0.044769462 -0.034915535  0.18240662 -0.02789282  0.701592837
under_12               0.081759355 -0.70053611 -0.072664619  0.030061320  0.01678559  0.70334851  0.037414735
```

Fig. 4. The vectors associated with PC1 through PC7

Towards the bottom left corner of the plot, we can see a clearly labeled point for observation 2863. Why does that point land in that particular spot on the graph? It's because observations 2863's standardized values, multiplied by the first two PCs, yield values of -2.43 for PC1 and -3.93 for PC2. Along with many other observations whose standardized values yield negative results when multiplied by PC 1 and PC2, this observation lands in Cluster 3 (Fig. 5).

Such plots can serve a valuable diagnostic purpose for a modeler. The plot shown in Fig. 2 indicates that when this dataset's standardized values are multiplied by the first two principal components, a three-cluster solution very neatly cleaves the observations into distinct groups, with little overlap.

A four-cluster solution, by contrast, does not separate the data in such a way. While the plot below shows a strong separation between clusters 1 and 3, many of the observations in clusters 2 and 4 overlap, in terms of where they fall along the data's first two PCs (Fig. 6).

```
> pc1 <- c(0.086230040, 0.686941731, 0.195923697, 0.009382622, 0.119523481, 0.679139790,0.081759355)
> pc2 <- c(-0.70097276, 0.08043302,  0.01223611, -0.01082705,  -0.04343833,  0.09624475, -0.70053611)
> first_two <- data.frame(pc1, pc2)
> new_matty <- as.matrix(first_two)
>
> combo <- port_scaled %*% new_matty
> combo[2863,]
      pc1       pc2
-2.343664 -3.932016
```

Fig. 5. Connecting the Dots between an observation from the the dataset and its position on the PC plot

Fig. 6. PC plot for a k = 4 solution

However, some important limitations of such plots should be noted here. First, this plot is limited to just two dimensions, and only shows Principal Components 1 and 2. For some datasets, the first two principal components may account for the overwhelming proportion of overall variance, but here, these two only explain 36.96% of the variance; therefore, PC analysis may not prove to be a particularly valuable aid when it comes to this particular dataset. Second, and perhaps even more important, a purely statistical answer to the "How many clusters?" question may not align with the associated business goals. Clustering is an unsupervised learning task, for which there is no solution. Furthermore, there is no threshold for statistical significance when it comes such models.[5] In fact, a company may choose the number of clusters to use in a segmentation model before any analysis of the data has even begun.[6].

In student-led presentations, such plots are frequently misused. Students often present a graph such as the one in Fig. 2, perhaps because they have been told that such plots are effective for demonstrating clustering results. When asked simple, straightforward questions such as "What does this graph show you about your model?" students

tend to stammer and struggle at first, before ultimately responding with something along the lines of "one of the clusters is red, another one of them is green, and yet a third one is blue."

Even the most data-savvy presenter, speaking to a completely data-savvy audience, cannot use such a plot to effectively explain the key distinctions from group to group. At best, a plot such as this could be used to indicate the level of differentiation among clusters for those first two PCs. However, such a plot does not enable the presenter to make statements such as "This cluster stands out for having more people per household, but that other cluster stands out for the way its members spend money on travel." For this reason, alternative visualization methods should be employed instead.

4 Simple Visualizations as a Far More Effective Option

Simple visualizations can convey the essential information about a clustering model.

1. *Such visualizations should depict original variables from the dataset, along with information about the model's cluster assignments.*
2. *Such visualizations do not need to depict all of the model's input variables, or all of the model's clusters.*
3. *Such visualizations can include variables that were not used as inputs in the original model.*

To demonstrate the power of a simple illustrative visualization, we will start with a boxplot, built with R's ggplot package [7].

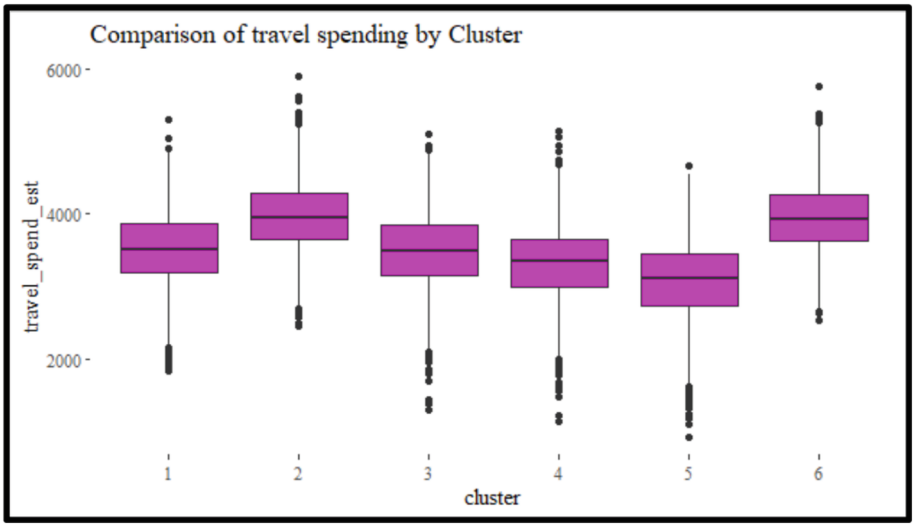

Fig. 7. Boxplot comparing travel spending among the six clusters

The plot shown in Fig. 7 clearly identifies a distinction among clusters – from this plot, we know that Cluster 2 and Cluster 6's households tend to be bigger travel spenders,

compared to households in the other clusters. We also know that Cluster 5's households have the lowest average travel spending, and that this cluster contains several outliers with very low travel spending amounts.

Effective cluster model visualizations do not need to depict every single segment. Reviewing subsets of the original data may be helpful for marketers who wish to zero in on particular groups. In the visualization below, we can come away with a clear takeaway regarding Clusters 5 and 6 – for the most part, these groups are well-separated in terms of their members' estimated household incomes, as well as their annual spending on travel (Fig. 8).

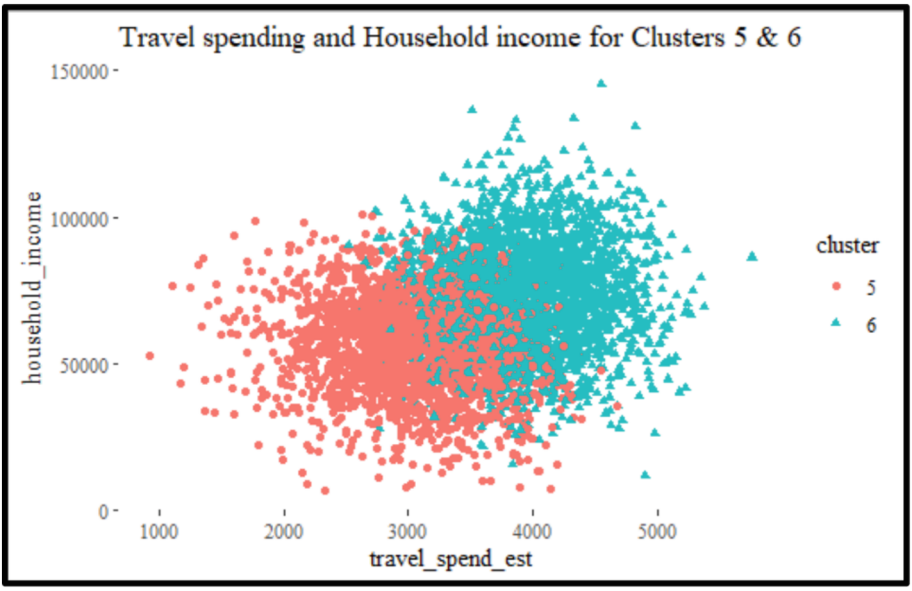

Fig. 8. A Look at Travel Spending and Household Income for a two-cluster subset

In a similar spirit, we could use a "one versus the rest" approach to a cluster visualization, in order to emphasize a point about a particular cluster and the way it stands out among the others. From the per-cluster summary stats, we know that Cluster 6 stands for its high rate of pet ownership. In the graph below, we compare the average number of pets owned by members of Cluster 6, compared with the overall average among members of the other five clusters (Fig. 9).

Even though categorical variables are not used as inputs in the k-means model, they can still be used as grouping variables in statistical summaries and visualizations based on the model. The graph below depicts counts of records, per cluster, along with county information as the fill variable (Fig. 10).

From the plot above, we can know that the clusters generally contain similar numbers of records, with Cluster 1 showing slightly fewer than the others. We can also know something interesting about the way the counties are represented among the six clusters – Sagadahoc, the least-represented county in the dataset, makes up a larger

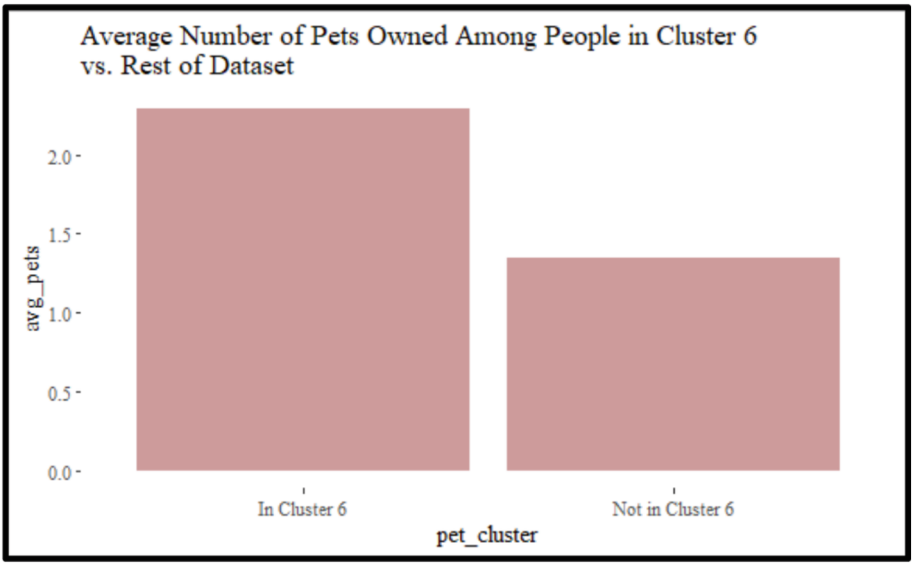

Fig. 9. Comparing Pet Ownership between Cluster 6 and the rest of the data

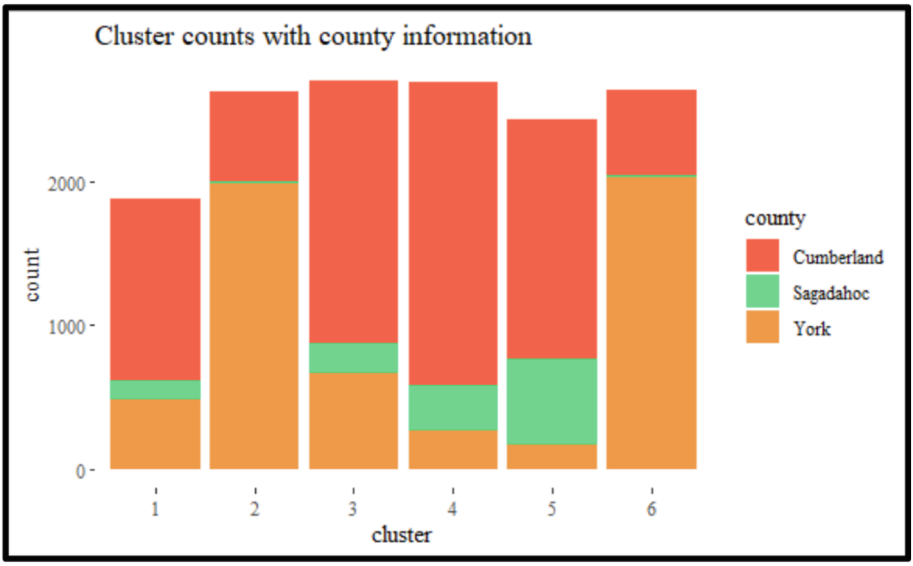

Fig. 10. Count comparison among the six clusters, with categorical info from the counties

proportion of Cluster 5, compared with its proportion of the others. Meanwhile, York County households comprise the overwhelming majority of Clusters 2 and 6, but do not comprise a majority of any of the others.

Finally, we will look at a faceted histogram plot, which shows the distributions of entertainment spending for all observations in the dataset, separated by cluster membership, and including a fill variable that indicates whether the household has a Lobster Land season pass. In this graph, the contrast between Clusters 4 and 5 appears in a particularly stark way – the midpoint of Cluster 5's distribution aligns with right-most tail of Cluster 4's distribution. Interestingly, this also shows a higher proportion of Lobster Land passholders among Cluster 6 and Cluster 2, compared with Clusters 5 and 1 (Fig. 11).

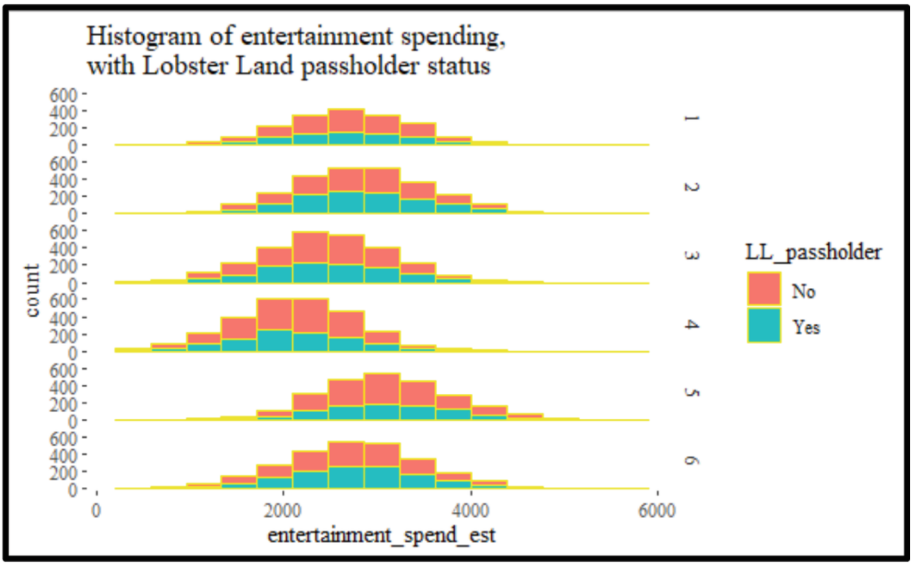

Fig. 11. Comparing entertainment spending across clusters, with an insight into passholder status

There are myriad combinations of variables and visualization types that could be used to convey information about a clustering model. While no single visualization is universally "best" at capturing distinctions among the groups in such a model, a mosaic of different visualizations – such as the ones shown here in this section – can lay a strong foundation for communicating model results.

5 Summary and Conclusion

Clustering is one of the most practically useful skills taught in data mining and marketing analytics courses. It is used every day, by companies large and small, in many different contexts.

Furthermore, data visualization is one of the most fundamental elements of the toolkit of any data analytics professional.

Data visualizations can serve an essential role in the way that a modeler communicates the results of a clustering solution to an audience. However, it is not enough to

"just" have data visualizations for expressing model results – it is also vital that these visualizations are built in a sensible way.

Plots based on Principal Components may appeal to students for their sophisticated look – however, such plots offer very little expository value when it comes to explaining clustering models. While these plots can have diagnostic value during the model-building process, they are not an appropriate choice for situations in which an analyst seeks to explain the key distinctions among the segments formed by such a model.

Instead, simple visualizations should be used for this purpose. These visualizations should depict original variables from the dataset. They do not need to include every single cluster, or every single variable – instead, they might just feature a tiny "slice" of the clusters or the variables to make some particular point.

References

1. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, p. 385. Springer, New York (2013). https://doi.org/10.1007/978-1-0716-1418-1
2. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022). https://www.R-project.org/
3. Ahmed, M., et al.: The k-means algorithm: a comprehensive survey and performance evaluation. Electronics **9**, 1295 (2020)
4. Kassambara, A, Mundt, F.: _factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7 (2020). https://CRAN.R-project.org/package=factoextra
5. Yuan, C., Yang, H.: Research on K-value selection method of K-means clustering algorithm. Multidisc. Sci. J. **2**, 226–235 (2019)
6. Artun, O., Levin, D.: Predictive Marketing: Easy Ways Every Marketer Can Use Customer Analytics and Big Data, p. 94. Wiley, New York (2015). The authors note that "Strategy informs segmentation and not the other way around"
7. Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York (2016). https://doi.org/10.1007/978-0-387-98141-3