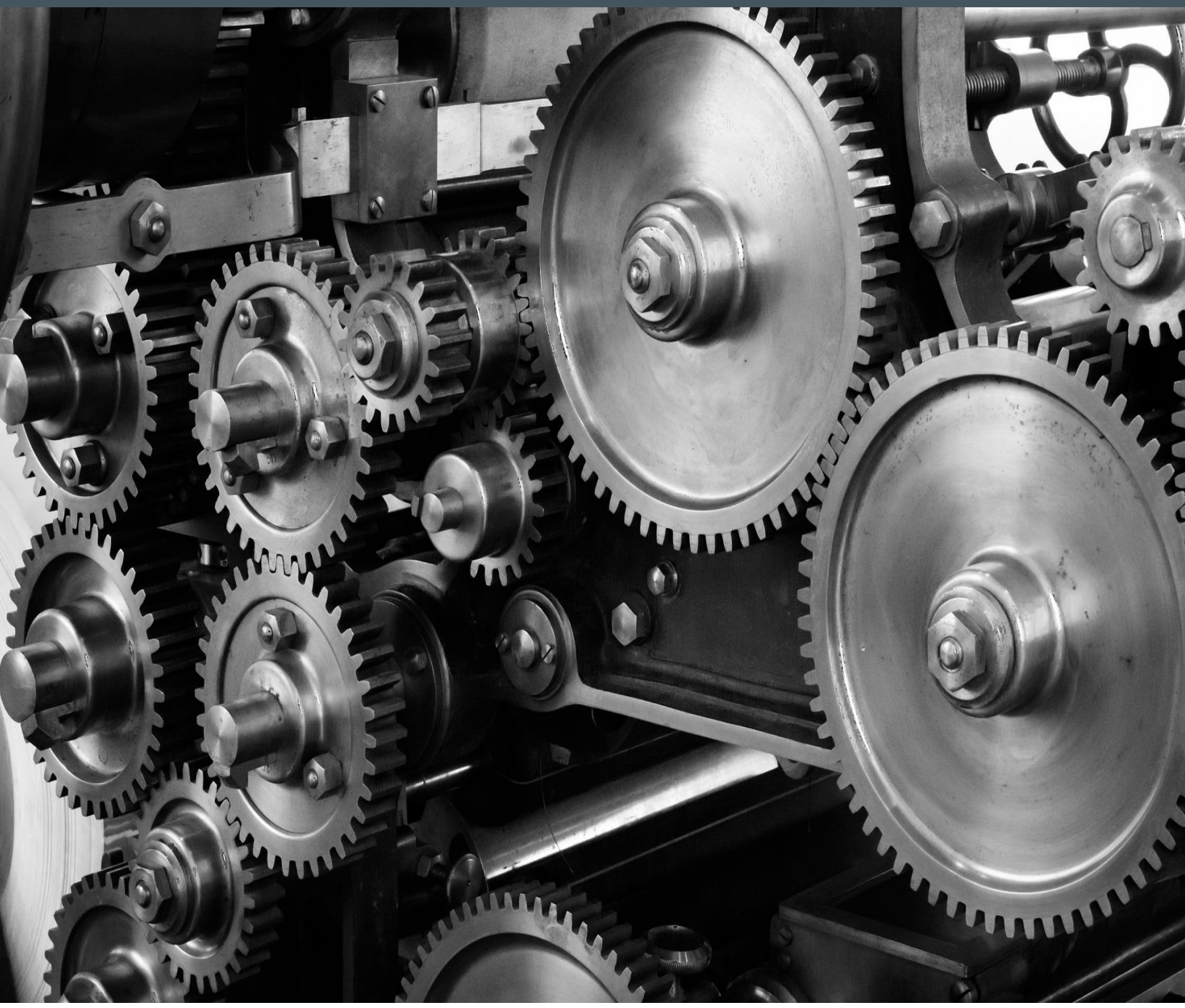
Vol. 1, No. 5 August 2018

FINANCIAL RISK GROUP

New Machinist Journal

The New Machinist's Toolkit: Heatmaps – Part II

Jonathan Leonardelli



Contents

What is Hierarchical Clustering? 3
Why Use a Dendrogram? 4
Determining Clusters 4
Example – Behavior Among Macroeconomic Variables 4
Final Thoughts 6

The New Machinist's Toolkit: Heatmaps – Part II

Machinist (noun): A person who operates a machine, especially a machine tool

In the previous volume of this journal we discussed heatmaps. We are now going to continue that discussion by addressing a common addition to heatmaps: dendrograms.

Dendrograms are a visualization tool that allows one to view the similarities among objects in addition to the "strength" of the similarity. They are typically represented as an upside-down treelike structure and are produced from the *hierarchical clustering* process. Objects that join together (i.e., cluster) lower in the dendrogram have stronger similarities than those that join together higher in the dendrogram (which likely means they are fairly dissimilar).

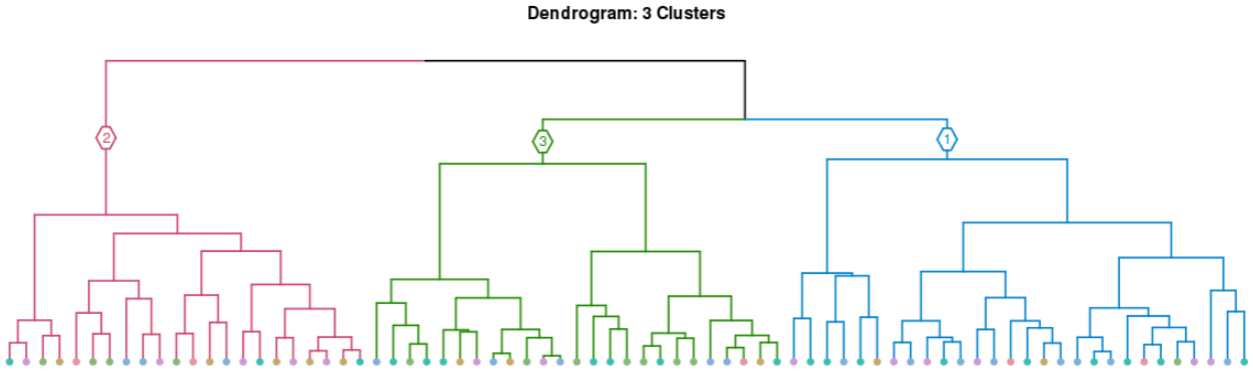
What is Hierarchical Clustering?

There are two common approaches to clustering – K-means and hierarchical. The K-means clustering approach groups observations into a predefined number (i.e., K) of clusters. Hierarchical clustering, on the other hand, does not result in a set number of clusters. Instead, it produces a dendrogram that depicts how observations are grouped and allows the end user to determine the number of clusters.

When setting up a hierarchical clustering algorithm the user must specify two measures. The first is the measure of dissimilarity between pairs of *observations*; the second is the measure of dissimilarity between *groups of observations* (i.e., clusters)[1]. Once defined the algorithm follows these steps:

1. Start by computing the pairwise dissimilarities of all single-observation "clusters"
2. Compute the dissimilarities among the clusters and group the two that are least dissimilar
3. Repeat Step 2 until one cluster remains

Note: the process of grouping objects together can be done either using a bottom-up (i.e. agglomerative) approach or top-down (i.e. divisive) approach. The image below shows an example of a dendrogram built from an agglomerative approach.



If we continue with the tree analogy, the bottom most nodes can be viewed as the leaves. As one moves from the bottom upwards the leaves come together to form branches, the branches merge with other branches or leaves to form new branches, and so on.

[1] A full explanation of dissimilarity measures is out of the scope of this volume but may be addressed in future one. For reference, common dissimilarity measures: for *observations* are Euclidean distance and correlation-based distance); for *groups of observations* are average, complete, single, and centroid linkage.

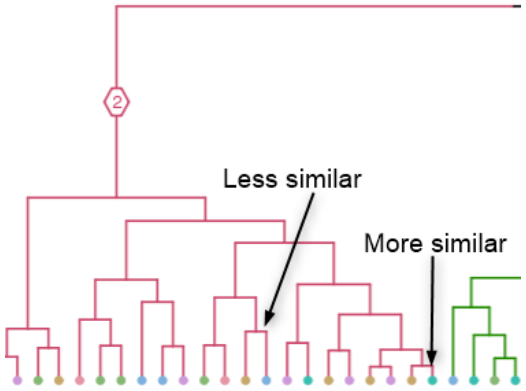
Why Use a Dendrogram?

Dendrograms provide a visualization of how objects are grouped together. As a result, they are very useful when looking at data and trying to find similarities across variables or observations. While a visual, intuition-based assessment could be a good first pass, adding some type of quantification of the similarities is a critical part of rigorous analysis.

Determining Clusters

How does one determine the number of clusters? Unfortunately, there is no hard and fast rule for this. One merely draws (or visualizes) a horizontal line at some point on the dendrogram and the number of branches that the lines passes through determines the number of clusters. Typically, this line occurs at a location that seems to have a reasonable number of clusters given the height of the joins along with a preconceived idea about how many clusters one wanted.

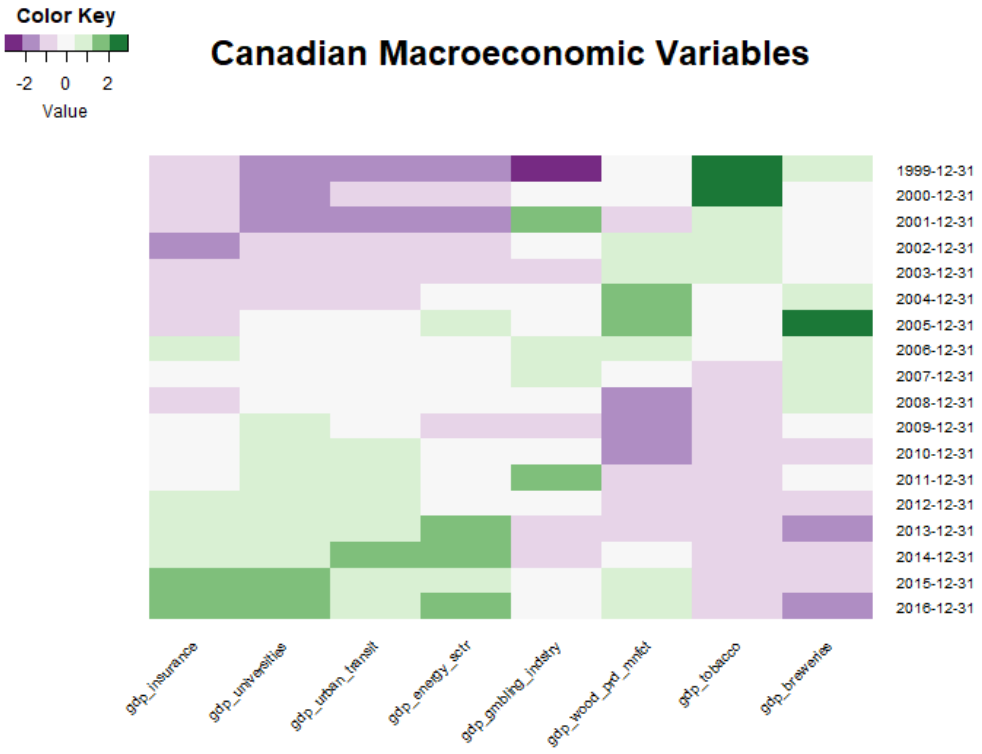
A word of caution: conclusions about similarity of observations should be based ***only*** on where they join vertically, not where they are located horizontally. In the image below, the joining of the two objects on the right indicate stronger similarity than those on the left because they occur closer to bottom.



Now that the overview is out of the way, let's return to the Canadian macroeconomic variable example started in the last volume.

Example – Behavior Among Macroeconomic Variables

In the last volume's example, we created the following heatmap and reached a couple of conclusions:

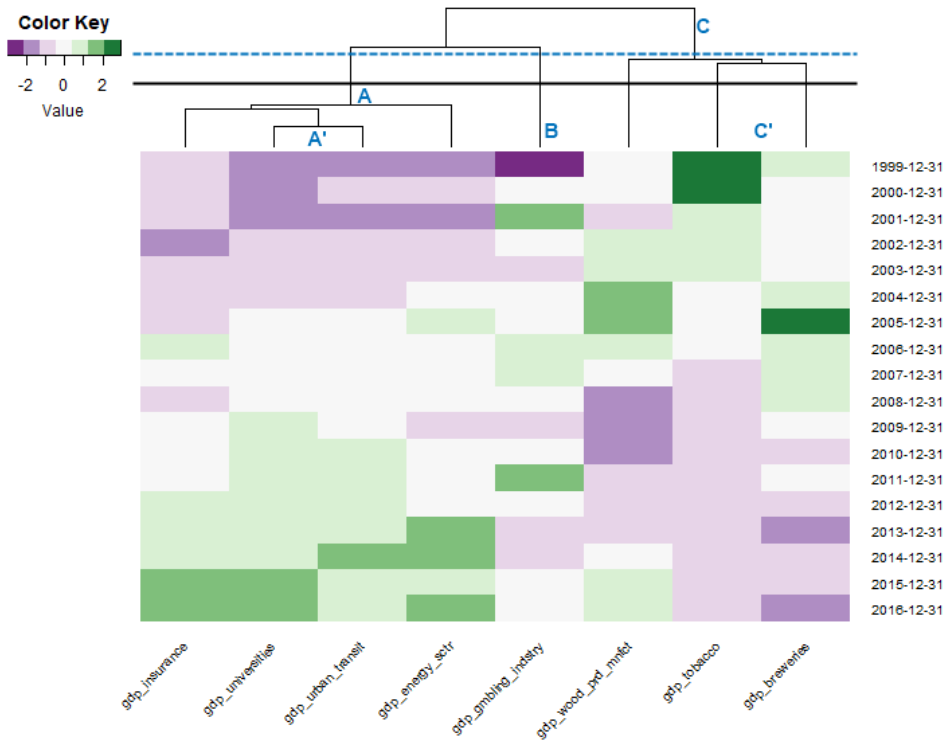


Conclusions (those relevant to this discussion):

1. The four left most variables seem to exhibit similar behavior.
2. The three right most variables seem to exhibit similar behavior.
3. The center variable – GDP for Gambling Industry – appears to behave unlike the others.

The image below now contains a dendrogram[2] with annotations.

[2] For this example, we used Euclidean distance on the standardized data.



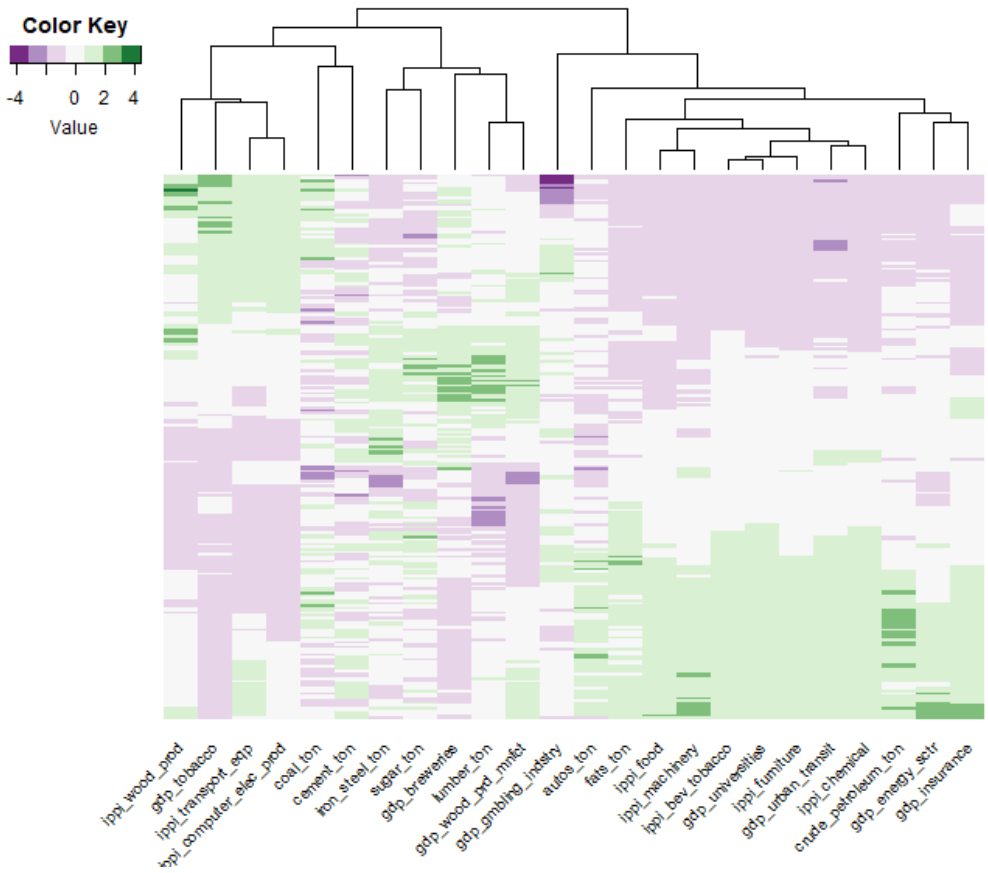
What does the hierarchical clustering tell us?

1. Where we draw the line across the dendrogram there are 5 clusters. The original thinking was there could be 3 clusters. However, to get three clusters the line would have to be drawn within a tight band (indicated by blue-dashed line) which would beg the question: *Why not just move it up slightly to get two clusters?* Where the solid black line *appears reasonable,* given its height within the dendrogram.
2. The four variables associated with **A** *looked* similar and the clustering process finds them similar too. Of all the variables in the table, those associated with **A'** have the most similarity (as indicated by the height at which the leaves form a branch).
    a. In fact, these four variables have the greatest similarity of all variables as indicated by the height where their leaves and branches come together as compared to others.
3. There is no real surprise with the middle variable associated with **B**. It looked like an oddball and the clustering process shows that it is not that similar to other variables. Therefore, it becomes its own cluster.
4. There is a bit of surprise with those variables associated with **C**. The left most variable in this group, upon reflection, looks like the odd one out. The two variables associated with **C** are a bit surprising given the weakness of their similarity (i.e. how high their leaves come together). They appear, visually at least, to be more similar than dissimilar.

Final Thoughts

This volume concludes the discussion of heatmaps and dendrograms. These two tools, used together, can help identify relationships. Given the fact the clustering algorithm can produce different clusters based on the data passed into it (e.g. scaled versus not-scaled) or the process used to determine dissimilarity one should use these tools as a beginning, not an end, to exploratory analysis.

There is one final image we would like to leave you with. This image contains a dendrogram on the full data set of Canadian macroeconomic variables.



How many clusters do you think exist?