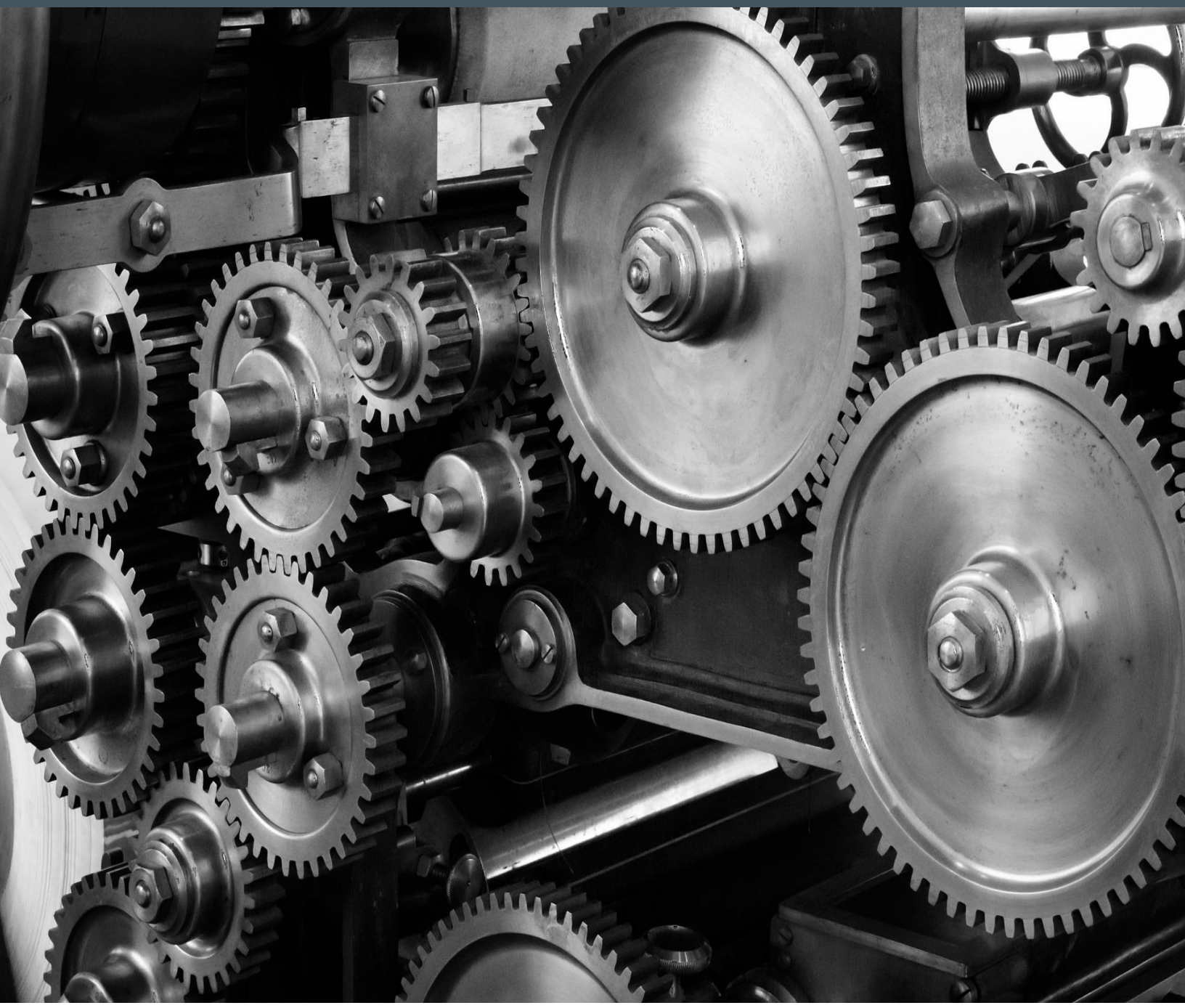
Vol. 1, No. 4 June 2018

FINANCIAL RISK GROUP

New Machinist Journal

Heatmaps – Part 1

Jonathan Leonardelli



Contents

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

Part of this journal's initiative is to provide an overview, description, and usage of common (and, perhaps, not so common) tools available to the new machinist. To that end, we will occasionally highlight a tool that we feel is a valuable addition to the new machinist's toolkit.

One of the key steps in building insight from the available data is *exploratory analysis*. There are many tools that one can employ – from visualization techniques to statistics – to help build a hypothesis in which to ultimately test. Today's volume will address just one of those visualization techniques: the heatmap.

What Is A Heatmap?

A heatmap is a two-dimensional depiction of *numeric* data with colors to represent values. The below image is a barebones representation of a heatmap built on data that has 18 rows and 8 columns. Each of the 144 cells is assigned a color.

Barebones Heatmap

Before we discuss how one might interpret a heatmap we will address the question: W*hy should I use a heatmap*?

Why Use a Heatmap?

There are many different uses of heatmaps. For example, heatmaps can be used to visualize website behavior or provide weather severity on maps. In general, heatmaps offer the following.

- They illustrate relationships across variables (columns) or observations (rows).
- They easily depict outliers.
- They show patterns.

- They indicate intensity of values in relation to others.

Succinctly stated: heatmaps are used to visualize data in order to quickly pick out patterns or behavior.

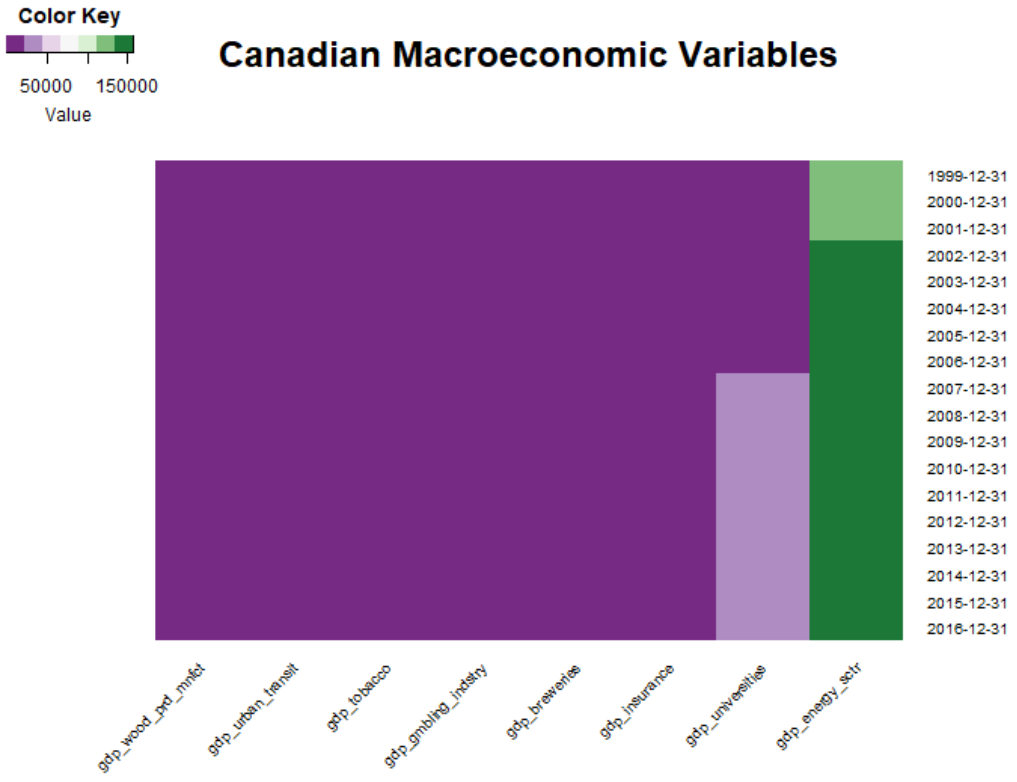
It should be noted that the process of representing information in a two-dimensional view can remove useful information. As a result, consider this as one tool of a few (or many) to use when doing exploratory analysis.

Now that some of the basics of heatmaps are out of the way, let's consider an example.

Example – Behavior Among Macroeconomic Variables

Suppose we want to review the behavior of Canadian macroeconomic variables (MEVs) as part of our model building process. We are curious in identifying which variables warrant further investigation and should, perhaps, be considered as part of our final model.

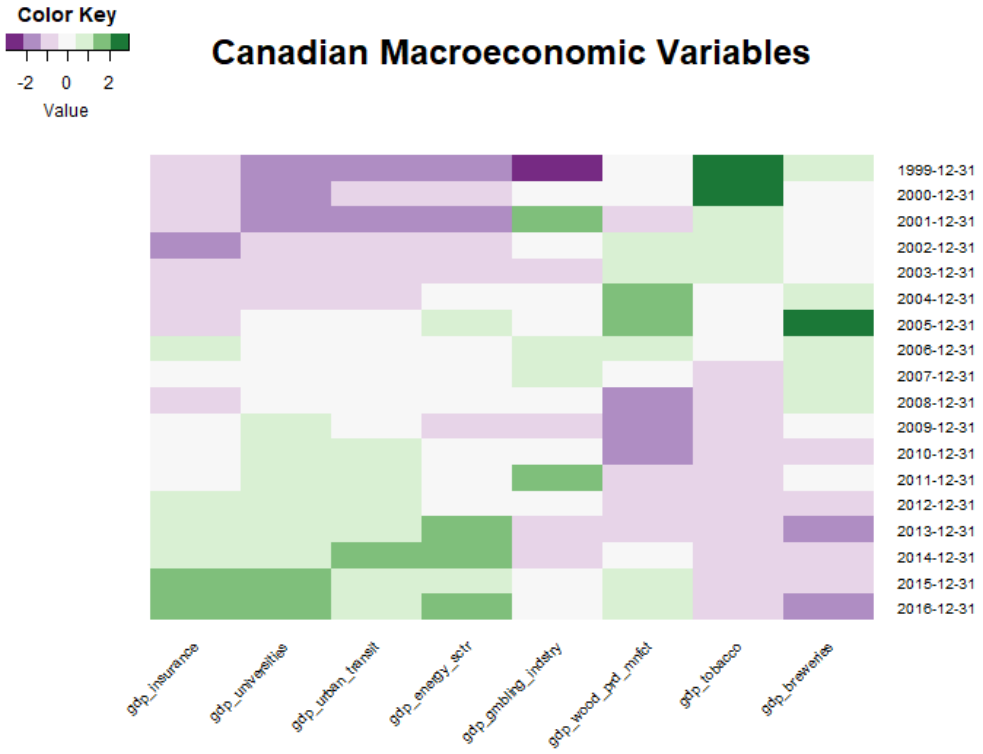
The first heatmap created looks like this:



What went wrong? This looks nothing like the one shown at the beginning of the volume!

In this data, the values for GDP in the Energy sector are about 8-40 times greater than the other variables. As a result, variations of green (large values) are given to that variable and variations of purple to all others. This raises an important point: depending on the type of analysis one is doing, it might be best to transform the data to ensure the scale of variables do not inadvertently impact the results.

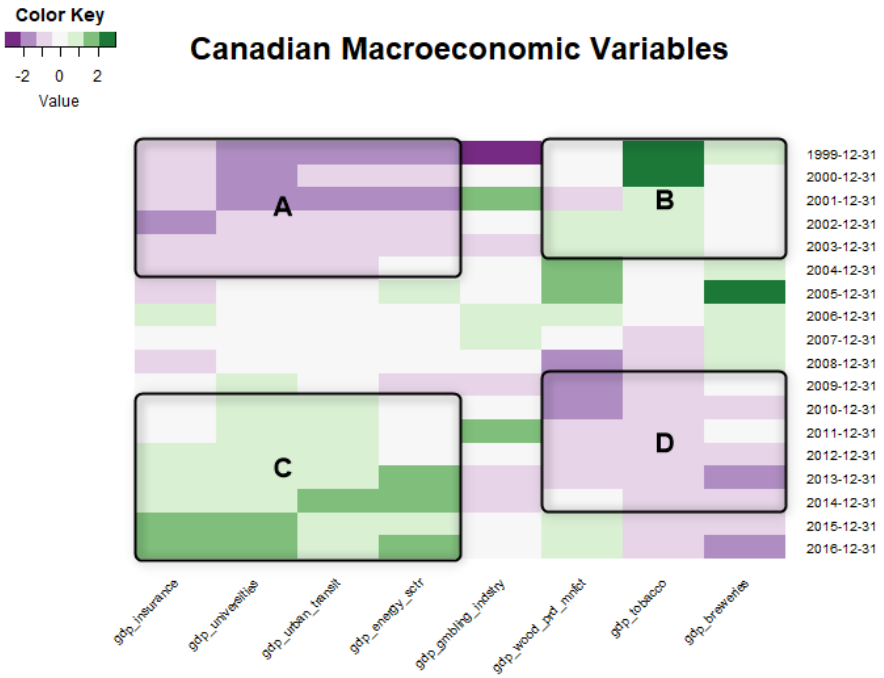
One way we can make the data in this heatmap comparable is to normalize the values of each column. After scaling the data in this regard, we end up with this heatmap:



The color key for this heatmap provides guidance on how to interpret the image. There are seven groups of values:

- three that are shades of purple and indicate values below the mean of the time series
- one that is white and indicates values near the mean
- three that are shades of green and indicate values above the mean

What do we see? Consider the same heatmap with annotations for discussion purposes:



The boxes are around cells that haver similar, scaled values. The variables on the left-hand side of the matrix, encompassed by boxes A and C, have similar patterns in the distant past (top rows) and near past (bottom rows). The same can be said for those variables on the right-hand side of the matrix, encompassed by boxes B and D.

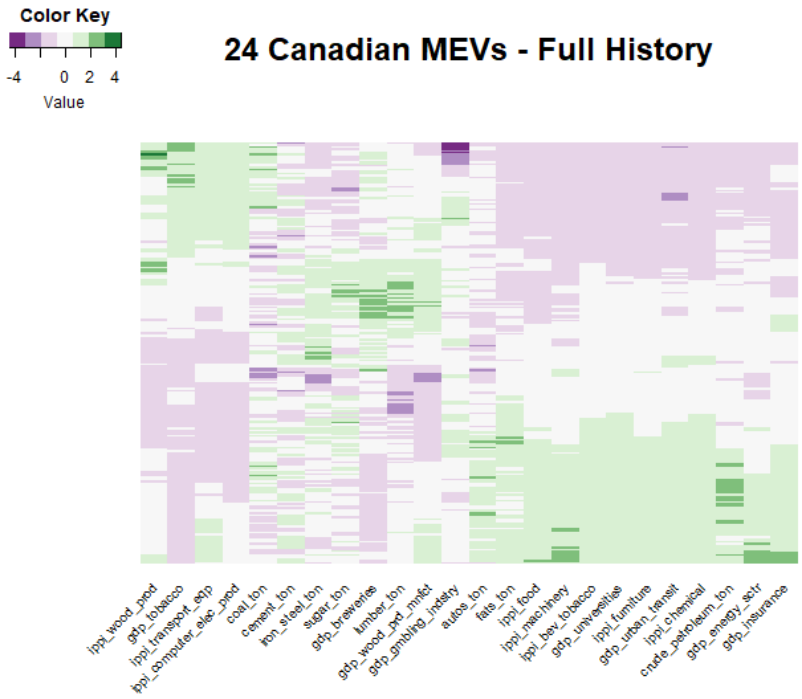
After analyzing the heatmap we may reach the following conclusions:

1. All four variables on the left-hand side might not be needed. They exhibit similar patterns.
2. All three variables on the right-hand side might not be needed. They, too, exhibit similar patterns.
3. The behavior of the variable on the left start low and trend high; those on the right start high and then trend low. That suggests an inverse relationship between the two groups.
4. The middle column – GDP for the Gambling Industry – seems to be an oddball. Its pattern is not like the others.
5. GDP values in the Tobacco industry (second to last column) have a couple of values that may be outliers (dark green). Same can be said about the middle column – GDP for the Gambling Industry (dark purple).

Final Thoughts

What we did here is the first of many steps in exploratory analysis. The conclusions we reached from reviewing the heatmap could be validated or refuted by looking at a correlation matrix, line plots, and t-statistics. Arguably, one might ask: *Why even use a heatmap if I'm going to still use statistics?*

Because, it provides a quick visualization of the behavior of variables. The heatmaps shown in this article are simple for pedagogical purposes. However, consider this one:



Do any patterns, behaviors, or oddities immediately jump out?

Next up – taking heatmaps one step further: the inclusion of dendrograms.