Official statistical classifications can be big, complex things. For example, the complete version of the “International Standard Industrial Classification of All Economic Activities, Rev.4” comes as [either a 300 page PDF or a 1.3GB Microsoft Access database](https://unstats.un.org/unsd/cr/registry/regdnld.asp?Lg=1). Classifications typically have a hierarchy of four or more levels and hundreds or thousands of categories, each with a code, a name and a detailed definition. One feature that causes endless definition is that the *names* aren’t necessarily unique identifiers; it’s quite common for the same name to be used for nested categories at different levels of the hierarchy, particularly where going down a level in that particular direction gets you a category that has only one sub-category.

Statistics New Zealand’s [ethnicity classification](http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/ethnicity.aspx) is a more or less typical example of a relatively straightforward classification. “Ethnicity” is a key but very artificial and constructed social variable, that is essential for many analyses. It appears to have a simple “common knowledge” definition but that breaks down very quickly when you start to think hard about it. It’s perhaps related to but very different from the similarly human-constructed categories of “citizenship” and “race”, with much potential for confusion. Ultimately, it comes down to self-identification (which is part of the standard for measuring it) – you are the ethnicity that you think you are, within the disciplines, constraints and opportunities created by social norms of categorisation. Nothing makes this more explicit and noticeable than the brutal, hard nosed black and white compulsion of a statistical standard! Recommended reading – [Foucault’s The Order of Things](https://www.goodreads.com/book/show/119561.The_Order_of_Things).

The ethnicity classification has four levels of hierarchy, with 1, 2, 3 or 5 digits in the code (there are no four digit codes because there can be more than 10 sub-categories per grouping in the fourth level, so it needs two digits). For example, code “1” means “European”, “11” means “New Zealand European”, and “111” and “11111” also mean “New Zealand European” – there are no subcategories of “New Zealand European” at the 3rd and 4th level of the hierarchy, so this forms an example of a non-unique category label. If you have data that includes the observation “New Zealand European” you don’t know which level of the classification hierarchy it is at (and hence what to compare with) unless you also have the code.

The term “New Zealand European” is part of both the everyday language and the official classification, but it’s [potentially divisive](http://salient.org.nz/2013/04/pakeha-or-nz-european-the-white-choice/). Certainly, it confuses me as an Australian of European descent who happens to be living in New Zealand. With the [New Zealand census](https://www.census.govt.nz/) coming up I was reminded of a long outstanding “must do sometime” note to myself to better understand the official ethnicity classification, so I set out to visualise it. A network graph is ideal, and here’s what I came up with:

* Zoom in to read it properly – with a mouse, just hover over the frame and use the mouse scroll wheel.
* The four levels in the hierarchy are colour-coded, with the larger text in the red boxes indicating the top level categories of five ethnic groups plus “other” and “residual categories”.
* Move a node around to see dynamically everything it’s connected to.
* Move the whole diagram around by clicking and dragging on some white space.
* Hover over a category to see its code.
* Click on an individual category to highlight what it is linked to.
* Double-click to collapse or expand a category’s sub-categories.

I quite like this as a way of exploring a hierarchical classification and think I’ll use it for others that I need to understand. In this case, it makes it easier for me to self-identify my own ethnicity as “European”, “Other European”, “Australian”, “Australian” (code 12811). That’s not how I spontaneously think of myself – but it’s definitely where I fit in this classification. Note that, importantly, the New Zealand standard requires users to self-identify with at least three different ethnicities (ie “leafs” in this tree) if they want.

There’s also a [larger version](http://ellisp.github.io/img/0117-ethnicity-large.html) available for using a full desktop sized screen. Probably neither of them will look great on a mobile device; due to the way I’ve set up my website I’m avoiding using knitr directly to generate webpages, and this is one instance where it makes it a bit fiddlier to include a nice object like this graph; so they are just showing in iframes. There is actually a method to [properly embed an htmlwidget into an existing webpage](https://stackoverflow.com/questions/34439928/embedding-an-r-htmlwidget-into-existing-webpage), but it’s a bit of a fiddle

Here’s the R code that drew this. It’s straightforward, other than perhaps a couple of tricks used to identify all the “edges” connecting categories with their sub-categories. Of the various options out there for creating JavaScript network diagrams from R, I find visNetwork the most intuitive to use and easiest to customise.

library(tidyverse)

library(openxlsx)

library(htmlwidgets)

library(stringr)

library(visNetwork)

library(RColorBrewer)

download.file("<http://archive.stats.govt.nz/~/media/Statistics/surveys-and-methods/methods/class-stnd/ethnicity/ETHNIC05-v2-classification-all.xlsx>",

mode = "wb", destfile = "ethnic05.xlsx")

nodes <- read.xlsx("ethnic05.xlsx")

names(nodes) <- c("id", "label")

# clean up:

unlink("ethnic05.xlsx")

# edges ie connection from one node to another

edges <- expand.grid(nodes$id, nodes$id, stringsAsFactors = FALSE) %>%

rename(from = Var1, to = Var2)%>%

mutate(f\_length = str\_length(from)) %>%

filter((str\_length(to) == f\_length + 1) | str\_length(to) == 5 & f\_length == 3) %>%

filter(str\_sub(to, 1, f\_length) == from) %>%

select(from, to)

# extra columns in the nodes data frame for the visualisation

nodes <- nodes %>%

mutate(

# `title` is for tooltips:

title = id,

# `group` is used for colour, so we use it to indicate the level of the classification:

group = str\_length(id))

# a nice, ordered colour palette:

pal <- brewer.pal(5, "YlOrRd")[4:1]

# draw diagram and save as a web page:

visNetwork(nodes, edges, width = "1700px", height = "900px") %>%

visPhysics(stabilization = FALSE, timestep = .3,

barnesHut = list(centralGravity = 0.6,

gravitationalConstant = -4000,

springLength = 80,

avoidOverlap = 0)) %>%

visOptions(highlightNearest = TRUE, collapse = TRUE) %>%

visGroups(groupname = "1", color = pal[1], shape = "box", font = list(color = "white", size = 23)) %>%

visGroups(groupname = "2", color = pal[2], shape = "ellipse") %>%

visGroups(groupname = "3", color = pal[3], shape = "ellipse") %>%

visGroups(groupname = "5", color = pal[4], shape = "ellipse") %>%

visEdges(smooth = FALSE, arrows = "end") %>%

saveWidget(file = "0117-ethnicity-large.html", selfcontained = TRUE, title = "Stats NZ ethnicity classification")