UpSet Plots in R

Center for Peace and Security Studies (cPASS)

Thomas Brailey

October 2019

UpSet Plots in R: A Preferred Alternative to Venn Diagrams

UpSet plots are a concise and easy-to-understand version of a Venn Diagram. Rather than looking at several proportional overlapping circles, one can view the frequencies of interactions of given variables in bar-plot form. They are very easy to plot in R, are incredibly useful for conveying multiple variables and their relationship with one another, and there exists several useful resources online (see the final section of this R Markdown file).

This document provides a basic but thorough workflow for formatting and plotting data using UpSetR.

UpSet with MSSL Nation Q-Code Data

First we need to set up the workspace and install the data. UpSet works best with long-form data, so time-series and wide data should be melted accordingly.

```
# Load packages
library(magrittr)
library(ggplot2)
library(kableExtra)

# Install data
entities <- rio::import(paste0(here::here(), '/data/entities.RData'))
```

We want to look at which nation q-codes are present in the Machine Learning for Social Science Lab (MSSL) dataset compared with those in existing datasets (COW, GW, IMI, UCDP, and ISD). Note that, for the purposes of visualization, NA values need to be set to 0.

```
# Subset by relevant variables (if necessary) and clean data
entities_clean <- entities %>%
  dplyr::rename("In MSSL" = in_MSSL,
                "In COW" = in_COW_nat,
                "In GW" = in_GW,
                "In IMI" = in_IMI,
                "In UCDP" = in_UCDP,
                "In ISD" = in_ISD)
entities_clean[is.na(entities_clean)] <- 0
entities_clean <- as.data.frame(entities_clean)
```

Here is a glimpse of our cleaned data. Our unit of analysis is “nation_wikidata_id”.

```
print(xtable::xtable(head(entities_clean, n = 10), type = "latex"), include.rownames = F)
```

nation_wikidata_id	In MSSL	In COW	In GW	In IMI	In UCDP	In ISD
Q1000	1.00	1.00	1.00	1.00	1.00	1.00
Q1005	1.00	1.00	1.00	1.00	1.00	1.00
Q1006	1.00	1.00	0.00	1.00	1.00	0.00
Q1007	1.00	1.00	1.00	1.00	1.00	1.00
Q1008	1.00	1.00	1.00	1.00	1.00	1.00
Q1009	1.00	1.00	1.00	1.00	1.00	1.00
Q1011	1.00	1.00	1.00	0.00	0.00	1.00
Q1013	1.00	1.00	1.00	1.00	1.00	1.00
Q1013421	1.00	0.00	0.00	0.00	0.00	0.00
Q1014	1.00	1.00	1.00	1.00	1.00	1.00

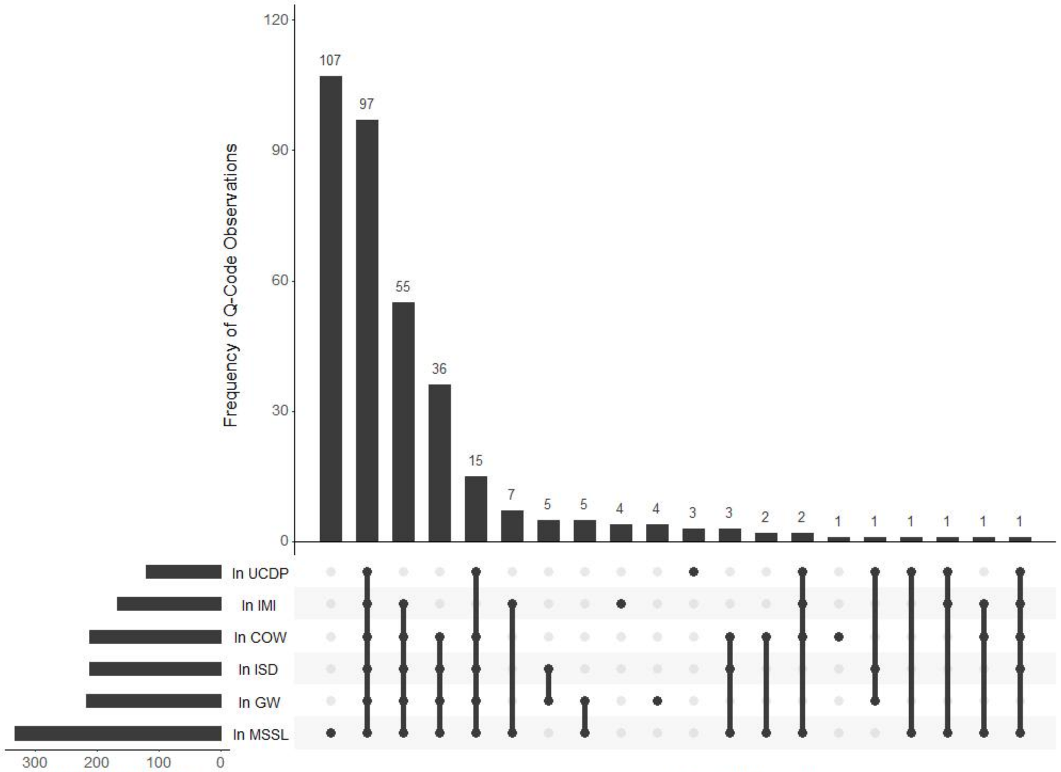
Using UpSetR

Now we generate the UpSet plot. Note that the grid:: commands are for additional annotations and are completely optional. That said, they are helpful in providing information on the number of observations in the dataset.

```
jpeg(filename = paste0(here::here(), "/paper/figure/entities_upset.jpeg"),
     width = 1000,
     height = 750
     )
# UpSetR commands
UpSetR::upset(entities_clean,
      nsets = 6,
      number.angles = 0,
      point.size = 4,
      line.size = 2,
      text.scale = 2,
      mainbar.y.label = "Frequency of Q-Code Observations",
      sets.x.label = "Total Observations",
      order.by = "freq"
      )

grid::grid.text(
  "Source: MSSL (2019)",
  x = 0.70,
  y = 0.02,
  gp = grid::gpar(
    fontsize = 14,
    fontface = 3
  )
)
dev.off()
```

Here is our output. Clear, concise, and able to simplify our large-N q-code entities dataset. Ordering the visualization by descending frequency shows that the MSSL dataset has the single largest number of nation Q-Code IDs. 97 of these Q-Codes can be found in all six of the datasets.



Source: MSSL (2019)

Figure 1: UpSet Plot of Nation Q-Code Observations Across Datasets

Limitations

While UpSet plots are useful for large-N datasets and comparing combinations of variables, the package itself has some limitations. For example, the standard package does not allow for titles and so the grid:: package must be employed. As such, aesthetic tweaks are somewhat more challenging than, say, ggplot2. That means, if we were to create a function that loops across multiple country/years (as is often the case with visualizations in the In-The-Loop Lab (ITL)), we would likely have to resort to manual aesthetic tweaks. Additionally, interpreting UpSet plots can become challenging when there are a large number of sets. In our case, we only have six sets (UCDP, IMI, COW, ISD, GW, and MSSL), but our data would need serious manipulation if we were comparing countries across, say, sixteen sets.

Further Resources

Below are some useful resources that I found particularly helpful for my own understanding of the UpSetR package.

- CRAN - The creators of the UpSetR package have a concise and easy-to-follow vignette on Cran.
- Little Miss Data's Blog - Little Miss Data's blog on the benefits of UpSet plots over the use of Venn diagrams is an excellent starting point for those wanting to implement these visualizations in their own work.
- Little Miss Data's GitHub - Little Miss Data all posts their replication code for their blog post on their GitHub.
- R-bloggers - One of the benefits of UpSetR is that it can be modified with ease. From adding extra labels, titles, and text using the grid:: package, to implementing different color schemes along the observations bar, UpSetR is very versatile. This page assumes a strong understanding of R and writing packages.