Joining REF2014 and HESA data

Christophe Rhodes

December 28, 2014

This vignette demonstrates how to merge ('join', in relational database terms) the datasets released by REF itself and by HESA in the REF2014 research assessment.

The data itself

```
> library(ref2014)
> data(ref2014)
> data(hesa2014)
```

is coded such that a `merge` should do The Right Thing:

```
> hesa2014.columns <- c("INSTID","UKPRN","Region","UOA","msubId","EligibleFte")
> join2014 <- merge(ref2014, hesa2014[,hesa2014.columns], all=TRUE)
```

though the duplicated character metadata columns (*e.g.* `Institution`) have not been audited for consistency, and so are removed from one of the datasets before the merge.

The `join2014` data frame needs analysing with care: the HESA data contains rows where the `UOA` is unspecified, which are preserved in the joined data (since we have specified `all=TRUE`). To illustrate, we will generate a Tufte-style slopegraph plot of QR-related scores, contrasting the score in absolute terms with the 'intensity'-scaled score (scaled by the proportion of eligible staff who were included in an institution's REF2014 submission).

First, we include some needed libraries:

```
> library(dplyr)
> library(ggplot2)
> library(scales)
```

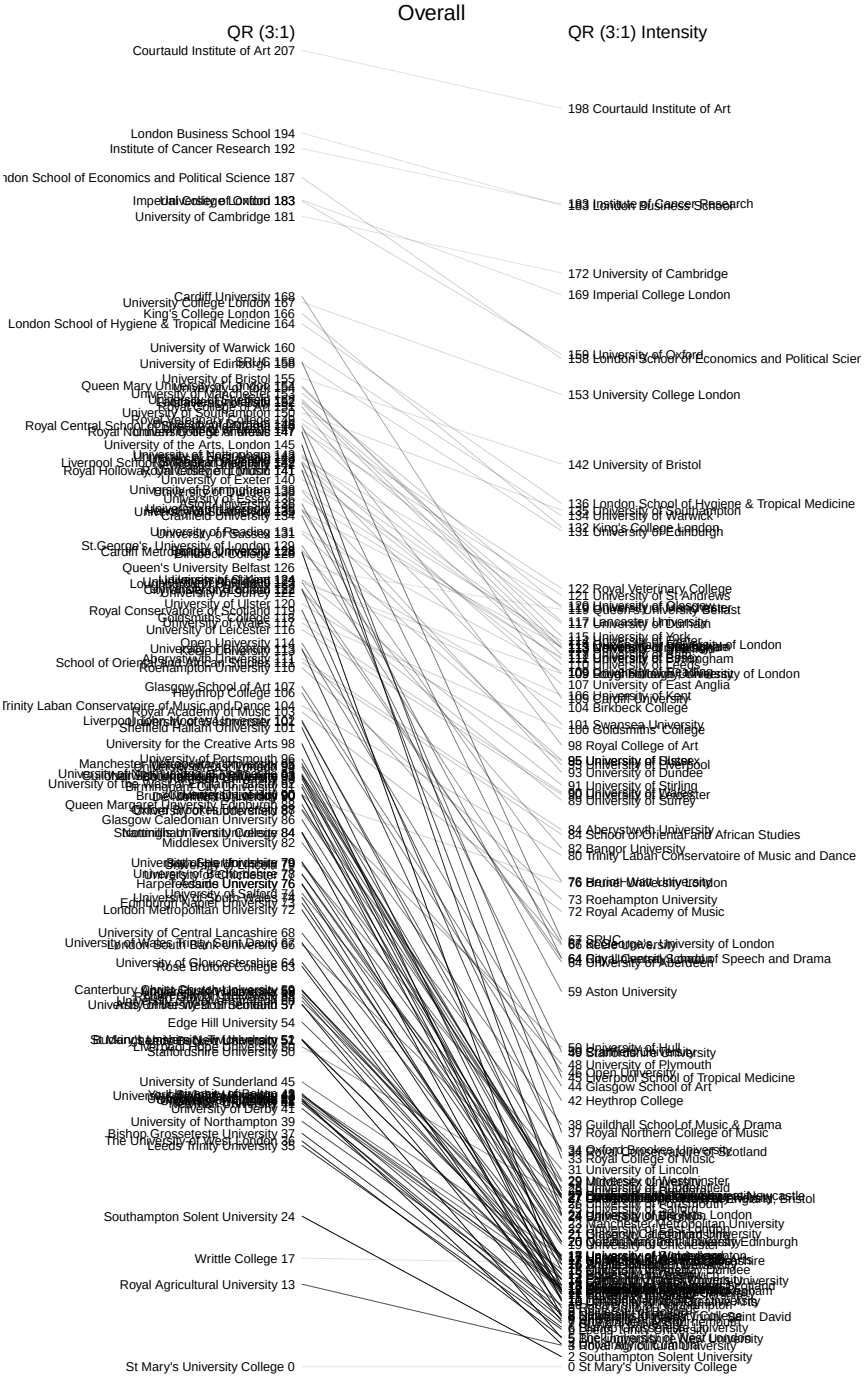
Then, we generate a summary table, keeping the `Overall` REF profile for the scores, as well as the rows where `Profile` is `NA` (from the HESA data) in order to be able to correct for the submission proportions. We group by the `UKPRN` (the provider reference number, identifying an institution), and compute the weighted mean of the REF scores. We also filter out institutions where the number of submitted staff is greater than the number of eligible staff (according to HESA), and where the number of staff submitted is nonpositive. Finally, we add columns for the two QR scores we will be plotting.

```
> overall2014 <- join2014 %>%
+     filter(is.na(Profile) | Profile == "Overall") %>%
+     group_by(UKPRN) %>%
+     summarise(FourStar=weighted.mean(FourStar, StaffFte, na.rm=TRUE),
```

```
+                ThreeStar=weighted.mean(ThreeStar, StaffFte, na.rm=TRUE),
+                TwoStar=weighted.mean(TwoStar, StaffFte, na.rm=TRUE),
+                OneStar=weighted.mean(OneStar, StaffFte, na.rm=TRUE),
+                Unclassified=weighted.mean(Unclassified, StaffFte, na.rm=TRUE),
+                Institution=first(Institution[!is.na(Institution)]),
+                nUOA=n(),
+                StaffFte=sum(StaffFte, na.rm=TRUE),
+                EligibleFte=sum(EligibleFte, na.rm=TRUE)) %>%
+     filter(StaffFte <= EligibleFte, StaffFte > 0) %>%
+     mutate(QR=3*FourStar+ThreeStar) %>%
+     mutate(QR.I=QR*StaffFte/EligibleFte)
```

To generate the slopegraph, we use `ggplot`, removing most of the chartjunk, and adding in line segments and appropriate labels. The overall picture is still somewhat unclear, but this is hopefully a sufficient illustration to demonstrate use of the data.

```
> step <- 1
> left <- sprintf("%s %.0f", overall2014$Institution, overall2014$QR)
> right <- sprintf("%.0f %s", overall2014$QR.I, overall2014$Institution)
> p <- ggplot(overall2014)
> p <- p + geom_segment(aes(x=0, xend=step, y=QR, yend=QR.I,
+                           alpha=0.5+ifelse(QR==0, 0, 0.5*abs(QR-QR.I)/(QR+QR.I))),
+                       size=0.1)
> p <- p + theme(panel.background=element_blank(), panel.grid=element_blank(),
+                axis.ticks=element_blank(), axis.text=element_blank(),
+                panel.border=element_blank(), legend.position="none",
+                axis.line=element_blank())
> p <- p + xlab("") + ylab("") + xlim((0-1), (step+1)) + ylim(0, 214)
> p <- p + geom_text(label=right, y=overall2014$QR.I, x=step+0.025, hjust=0, size=3.5)
> p <- p + geom_text(label=left, y=overall2014$QR, x=-0.025, hjust=1, size=3.5)
> p <- p + geom_text(data=data.frame(x=1), label="QR (3:1)", x=-0.025, y=210, hjust=1, size=5) +
+     geom_text(data=data.frame(x=1), label="QR (3:1) Intensity",
+               x=step+0.025, y=210, hjust=0, size=5) +
+     geom_text(data=data.frame(x=1), label="Overall", x=0.5, y=213, size=6)
> print(p)
```


Overall
QR (3:1)
QR (3:1) Intensity
Courtauld Institute of Art 207
198 Courtauld Institute of Art
London Business School 194
Institute of Cancer Research 192
ndon School of Economics and Political Science 187
Imperial College London 183
University of Cambridge 181
172 University of Cambridge
169 Imperial College London
158 University of Oxford
153 University College London
142 University of Bristol
136 London School of Hygiene & Tropical Medicine
132 King's College London
122 Royal Veterinary College
107 University of East Anglia
104 Birkbeck College
101 Swansea University
98 Royal College of Art
89 University of Surrey
84 School of Oriental and African Studies
80 Trinity Laban Conservatoire of Music and Dance
73 Roehampton University
72 Royal Academy of Music
59 Aston University
48 University of Plymouth
46 Open University
44 Glasgow School of Art
42 Heythrop College
38 Guildhall School of Music & Drama
37 Royal Northern College of Music
33 Royal College of Music
31 University of Lincoln
2 Southampton Solent University
0 St Mary's University College
Cardiff University 169
University College London 167
King's College London 166
London School of Hygiene & Tropical Medicine 164
University of Warwick 160
University of Bristol 155
University of the Arts, London 145
University of Exeter 140
Queen's University Belfast 126
University of Ulster 120
University of Leicester 116
Open University 114
Glasgow School of Art 107
Heythrop College 106
Royal Academy of Music 103
University for the Creative Arts 98
Glasgow Caledonian University 86
Middlesex University 82
University of Central Lancashire 68
University of Gloucestershire 64
Rose Bruford College 63
Edge Hill University 54
University of Sunderland 45
University of Derby 41
University of Northampton 39
Bishop Grosseteste University 32
Southampton Solent University 24
Writtle College 17
Royal Agricultural University 13
St Mary's University College 0