A Graph for the Economist

Data Computing

Computing project

The *Economist* is a well-regarded weekly news magazine. The following graphic accompanied their article about the release of the "College Scorecard" data in Sept. 2015.

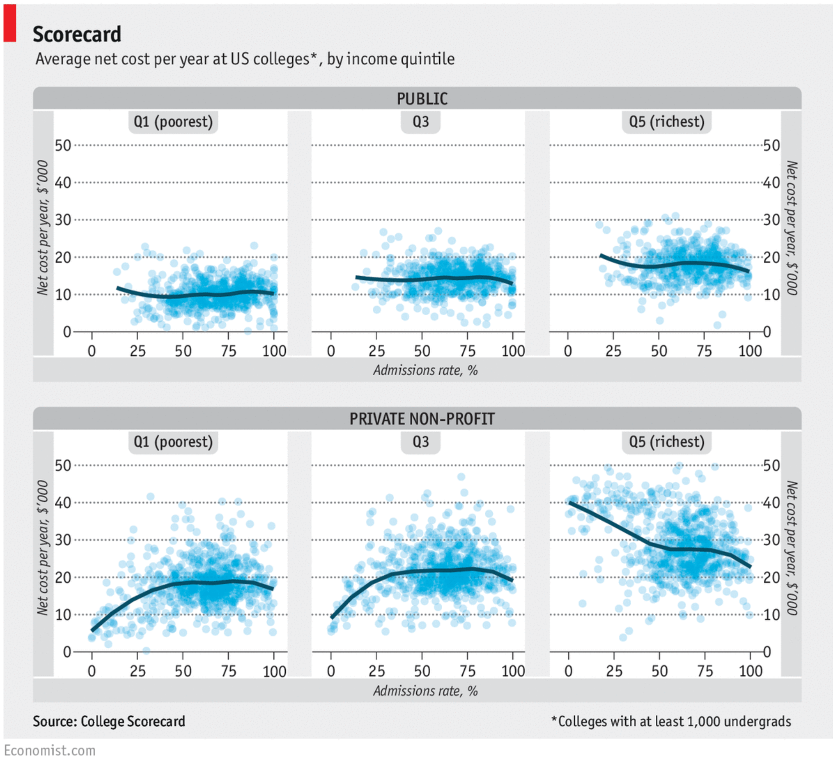

Figure 1: Yearly cost of attending college versus college selectivity. Each dot is one college or university.

Your task is to reproduce this graph from the College Scorecard data, and perhaps enhance it.

The data

The Scorecard data is too voluminous to work with conveniently in class; it takes too long to download. You'll be working with a subset available at `tiny.cc/dcf/ScorecardSmall.Rda` which contains a single object, the data table `ScorecardSmall`.

```r
download.file("http://tiny.cc/dcf/ScorecardSmall.Rda", destfile = "ScorecardSmall.Rda")
load("ScorecardSmall.Rda")
```

The subset includes all 7804 institutions in the original 2013 Scorecard file, but just 54 variables. Some that you may be interested in are:

1. `CONTROL`: public (1) or private (2) institution. (You can discard cases with `CONTROL == 3`. They are not in the Economist's graphic.)
2. `INSTNM`: name of the institution
3. `ADM_RATE`: admissions rate in percent
4. `CCSIZSET`: Carnegie size classification of the institution. Values 1, 6, 7, 8 correspond to schools with fewer than 1000 students.
5. `AVGFACSAL`: Average faculty salary per month
6. `TUITFTE`: Tuition revenue received by the institution per student full-time-equivalent.
7. `NPT4_PUB`: average net cost for students in public institutions
8. `NPT4_PRIV`: average net cost for students in private institutions
9. `NPT41_PUB` : average net cost for students at public institutions whose families are in the lowest of five income groups. Similarly, `NPT42_PUB` is for students whose family income is in the 2nd group, and so on up to the 5th group. The groups are defined as $0 to $30K per year, $30-48K, $48-75K, $75-110K, $110K or more. There is also `NPT41_PRIV`, and so on, for private institutions.

All of the `NPT4` variables are for students receiving aid from the federal government under Title IV.

What's the case?

The case in the Scorecard data is an institution. In the *Economist* graphic, however, the case is a level of family income (as in `NPT4`) at an institution. That is, from the perspective of the graphic, the Scorecard data is in wide form. You'll have to convert it to narrow form to make the graph.

1. Select just the variables you need from the Scorecard data.
2. Use `gather()` to convert from wide to narrow format.
3. After (2) you will have a variable with levels like `NPT43_PUB`, `NPT45_PRIV`, etc. You will want to translate these to `Q3`, `Q5`, etc. For your convenience, the file `http://tiny.cc/dcf/NPT4-names.csv` contains a table with the appropriate translations. You can use a join of the narrow-format Scorecard data with this table to perform the translations.