|  |
| --- |
| R in a Hurry, Hour 2 |

# Introduction

The second hour of this session is set aside for individual project work. This document describes six possible projects, in order of increasing difficulty. Please pick one and have fun :-)

Please note: the more difficult projects cannot be completed in one hour (unless you are Jeff Dean; google “best programmer dean”).

# Project 1: Run the R code in the slides from hour 1

If you are new to R, you may want to reinforce what you heard in hour 1 by going through the code again, at your own pace. You can get the script will all the code from hour 1 at the following location:

# https://github.com/derrenbarken/data\_science\_fun\_conference\_2015\_R\_in\_a\_hurry

# Project 2: Run the R code to do the data preparation for the hour 1 data

As discussed in hour 1, data preparation is an important school. You can download the raw data that was used to prepare the data in hour 1 at the following URL: http://star.cde.ca.gov/star2012/ResearchFileList.aspx?rf=True&ps=True

# Project 3: Analyze subgroup (covariate) data like gender, ethnicity, etc

There is a treasure trove of additional information breaking down test scores by interesting covariates like gender, ethnicity, parental education, and more.

This data is coded by using an additional column called “subgroup”.

The key for what the subgroup numbers mean (for 2013) appears here:

<http://star.cde.ca.gov/star2013/research_fixfileformat.aspx>

(Scroll down, starting at “Gender”).

You may download a “clean” (pre-processed) file, similar to the data from hour 1 (i.e. at the county level), but with the additional subgroup column. It is in the file named star\_subgroup\_clean.csv

In hour 1, we created a graph showing English scores over time (grades 2-11).

Using the subgroup data, can you make a similar chart, but with two lines, one for boys and one for girls?

What other interesting insights can you glean from the subgroup data?

# Project 4: Analyze data latitudinally

It’s SoCal vs Nocal… Every district has a latitude. Is latitude a useful predictor? Are there any differences between SoCal and NoCal schools? If you find an effect, can you demonstrate that it is valid even when you control for covariates such as income which might be correlated to latitude?

# Project 5: Analyze data at the district or school level

All of the data analyzed in hour 1 was at the county level. However, the actual STAR data is broken down to the district and school levels.

Do counties with smaller schools get better scores than counties with larger schools? Do counties with higher income tend to have smaller or larger schools? How would you measure school size within a county? Average? Median?

Can you find district level data for income or spending online? Can you merge it with the district level data for student scores?

If you can find (and process / merge) district level data for income or spending, when you see a relationship that seems to hold at the county level (e.g. income to test scores), does that hold uniformly within all the counties? Looking at districts within counties, are there some counties where the relationship holds and some where it doesn’t?

# Project 6: Analyze data longitudinally

All of the data analyzed in hour 1 was from 2013. There were test scores from multiple grades, but each of those grades was tested in 2013. However, there is more data available at the STAR website. Specifically, there are 16 years’ worth of data on the STAR website, from 1998-2013. You can access this data at <http://star.cde.ca.gov/>

Because this data is available, if you combine this data appropriately, you can construct true longitudinal data, rather than cross sectional data. For example, the 11th graders of 2013 were the 10th graders of 2012, the 9th graders of 2011, etc.

Let’s look at an example. In hour 1, we made a slide with English scores by grade. That data was cross sectional. However, we could make the same type of plot with multiple lines – one per cohort. For example, there could be a line that connects scores for the 2nd graders of 1998, the 3rd graders of 1999, the 4th graders of 2000, etc. Another line could connect the scores for the 2nd graders of 1999, the 3rd graders of 2000, the 4th graders of 2001, etc.

Looking at the resulting plot, with multiple lines for multiple cohorts, one could get a global picture of whether English education in California, as judged by test scores across all grades, has been getting better or worse over the last 16 years. Are there any identifiable trends? What about other tests?

Bonus question: imagine you had a thousand tests – can you devise an automated way to look for trends in data like this? Does your automated test catch the trends you identified “by inspection”?