Analysis of UC Compensation Data, 2012-2014

Angela Dai, Abizer Lokhandwala, Michael Jetsupphasuk, Harry Sutcliffe

Introduction

In 2014, the largest database of California government employee compensation was released: Transparent California. It is searchable by name and job titles on TransparentCalifornia.com and includes over 2 million salary records. Being University of California (UC) students, we will investigate UC employee compensation data. With tuition and student debt at all-time highs, university spending is under scrutiny. Since the delivery of education, research, and health care is labor intensive, payroll costs account for about half of the UC's annual operating budget. Thus, we thought it would be both interesting and important to analyze this data.

Description of Problems

Our main research question is: How has the University of California pay distribution changed over time? To answer this, we will answer these questions: What are the disparities between academic vs. non-academic pay over time? How much does the average lecturer/professor/etc (academic positions only) make? What does the distribution of pay look like? How has total staff grown in relation to enrollment?

Description of Data

Our three primary raw data sources were from TransparentCalifornia.com for 2012, 2013, and 2014 UC compensation data. Each of the data sets had over 260,000 lines of information covering employee name, job title, base pay, overtime pay, total benefits, total pay, total pay with benefits, year, and agency. While the availability of this information online was intended to show "transparency and public accountability," the format of the data hindered analysis since it was incredibly inconsistent within and between years. To help navigate the title coding system, we used the UCOP Academic Title Code document.

Set-up

Set-up for this analysis was nontrivial due to the sheer amount of variation in the underlying dataset. For example, in one year, the employee names were capitalized, whereas in the others, they were not, and in another year, names were in the form of "fname lname" but in the other years they were in the form of "lname, fname." In one year's data, the titles contained full words - e.g. "PROFESSOR-HCOMP," while the other years used abbreviated title names such as "PROF-HCOMP." Also, the raw data files were not properly encoded in some cases - the Title dataset, for example, had strange Unicode characters in a number of titles which we could not correct for in R itself. The Title dataset also had another major problem: in some cases, the words in the titles where separated by a space and/or a hyphen, with no consistency in style, but in the yearly compensation CSVs, the words were separated with different delimiters or number of delimiters, such as two (or more) spaces/hyphens instead of one, or spaces/hyphens switched in with respect to their locations in the Title dataset, and several other tedious problems like this. Cleaning the data was a major hassle and frustration. Data inconsistency handicapped the confidence with which we were able to make any analyses on the dataset at large.

With regards to cleaning, R's tools where unfortunately insufficient for our purposes. We had to resort to command like utilities (cut, sort, sed) and manual munging in a text editor (for cleaning up Unicode, figuring out the permutations of spaces and hyphens necessary to match all the cases we had the patience to find) in order to get a dataset R (and our group) was happy with. Nonetheless, some of R's tools where far superior to

any external ones, especially with respect to transforming CSVs - write.csv in particular generates beautiful output that lends itself very well to replicability.

We begin by loading the data as downloaded from the Transparent California website:

We do an initial filter to only get employees with a total compensation of > $1,000 in order to slightly simplify the analysis.

General housekeeping:

Prior to finding the UC Titles dataset, we tried manually grouping the ~2900 unique titles in the datasets via regex. We set up a hash with associated titles in order to `sapply()` a regex across the list of titles to get them matched to groups.

Using this method, we eventually tagged about 91.5% of the ~250,000 employees in the dataset with a department. However, this method was problematic, as it did not lend itself well to separating employees that fell into multiple categories, such as assistant/associate/visiting/research professors, and still left us with a large number of departments to group further. The code to make this work was also long, slow, and complicated, so we abandoned this thread.

We then found and began cleaning the Title dataset by copy-pasting the table from a PDF provided by the UCOP website into Microsoft Excel and using Excel's CSV output tools to generate a initial, raw CSV datafile. We then read this data into R to begin exploratory analysis. This was a deadend, as we had to use `grep()` and regular expressions to manually correct individual problems in the CSV, and this was ugly, confusing and difficult to maintain. So, we instead used dplyr to extract and sort the columns we wanted and used `write.csv()` to generate a second-stage CSV, which we then piped through a text editor, where we manually made corrections and then used this final CSV as the basis for the rest of our analysis. This was an improvement over our previous attempt, in that it gave us an easy way to separate academic from staff/non-academic titles, but it was not as granular as our manual, regular-expressions based attempt. Nonetheless, it was the schema we went with for our analysis.

We are defining the classification as follows:

- Academic employees: those directly engaged in the academic mission - professors, clinical professors, other teaching faculty, researchers, and other academic titles
- Staff/non-academic employees: those who support academic departments, student services, patient care and other university functions

Other Resources: * Desrochers, D. M., & Kirshstein, R. J. (2014). Labor intensive or labor expensive? Changing staffing and compensation patterns in higher education (Issue Brief). Washington, DC: American Institutes for Research. * UC Student Enrollment

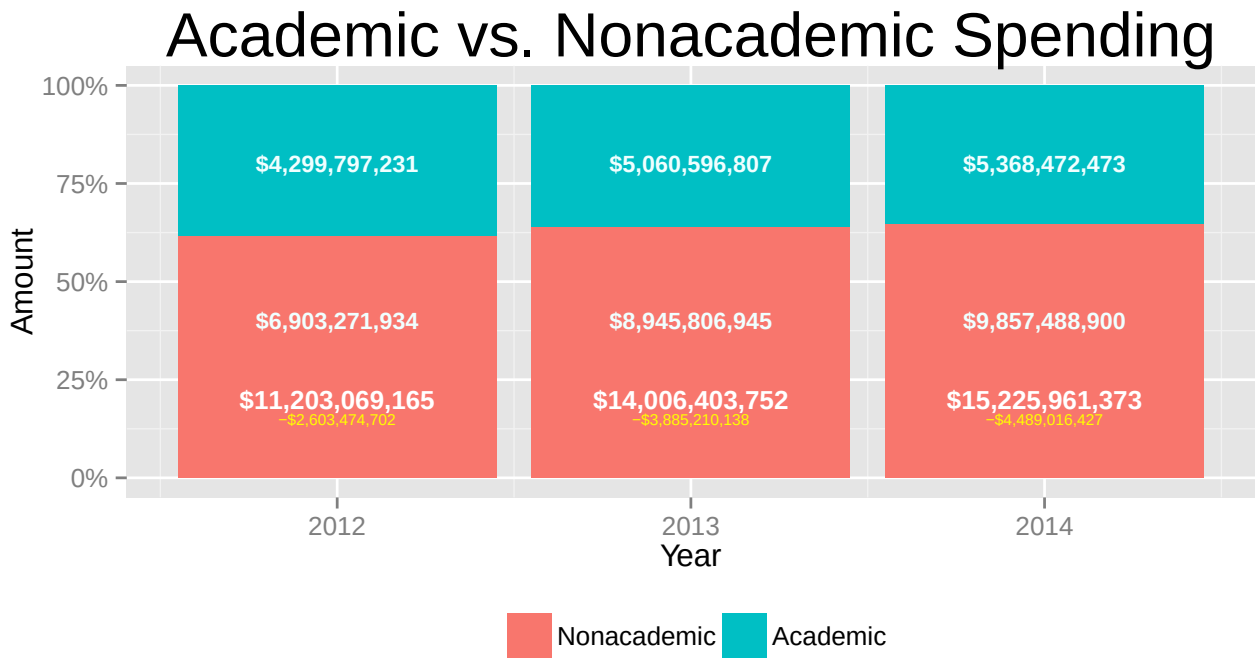
Analysis Approach

Pie chart of Distribution of Total Compensation - executive pay, academic, health centers

Academic vs. Staff/Non-Academic Workforce Distribution of Total Compensation, 2012-2014

```
gp
```

```
## ymax not defined: adjusting position using y instead
## ymax not defined: adjusting position using y instead
```

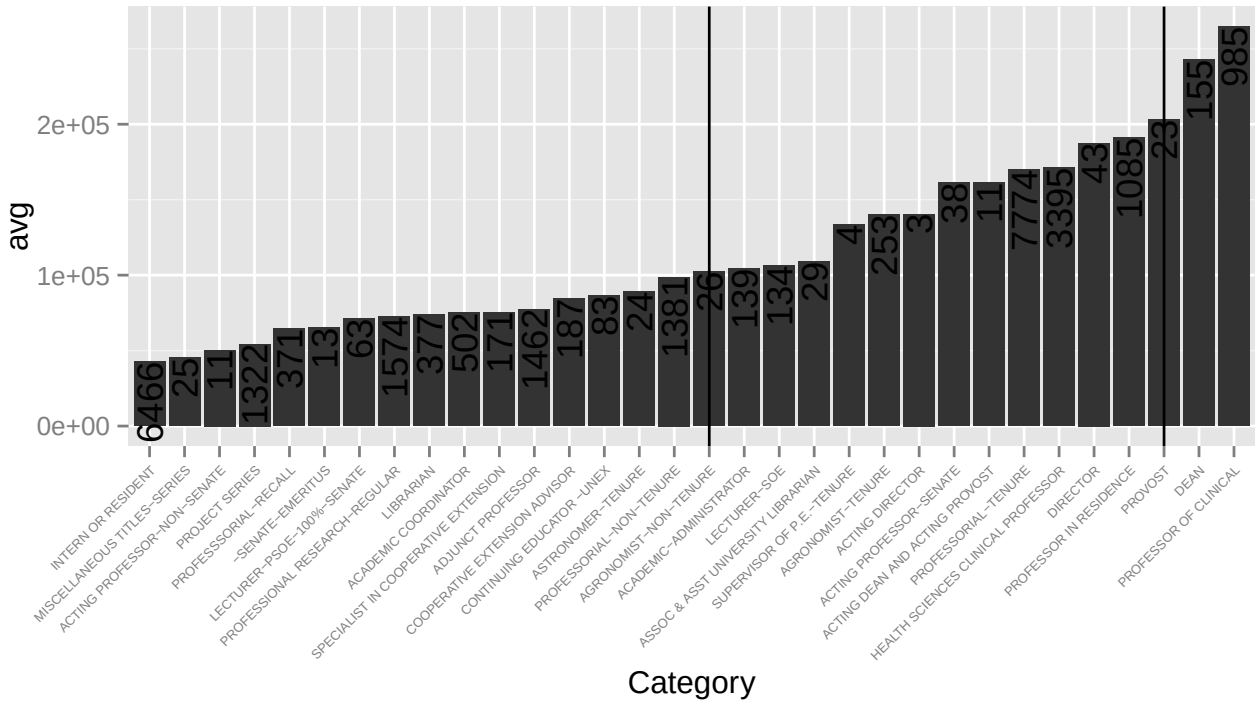


This could be a bar chart showing changes in breakdown of academic vs staff/non-academic % compensation and changes in overall compensation, with: x-axis: $30k groupings of total compensation y-axis: # of employees We could conclude: The largest % of academic employees fall in the $x-$x range

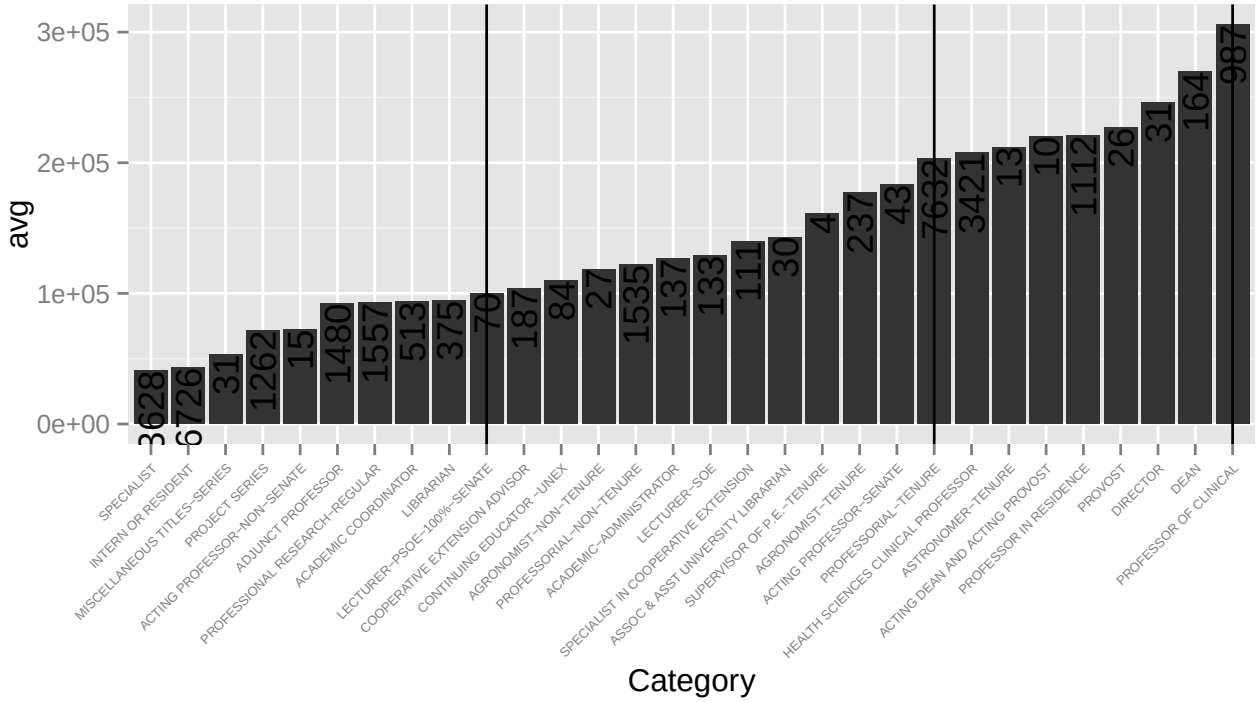
Average Total Compensation for Academic Positions, 2012-2014

```
uc2012.by_department.plot
```

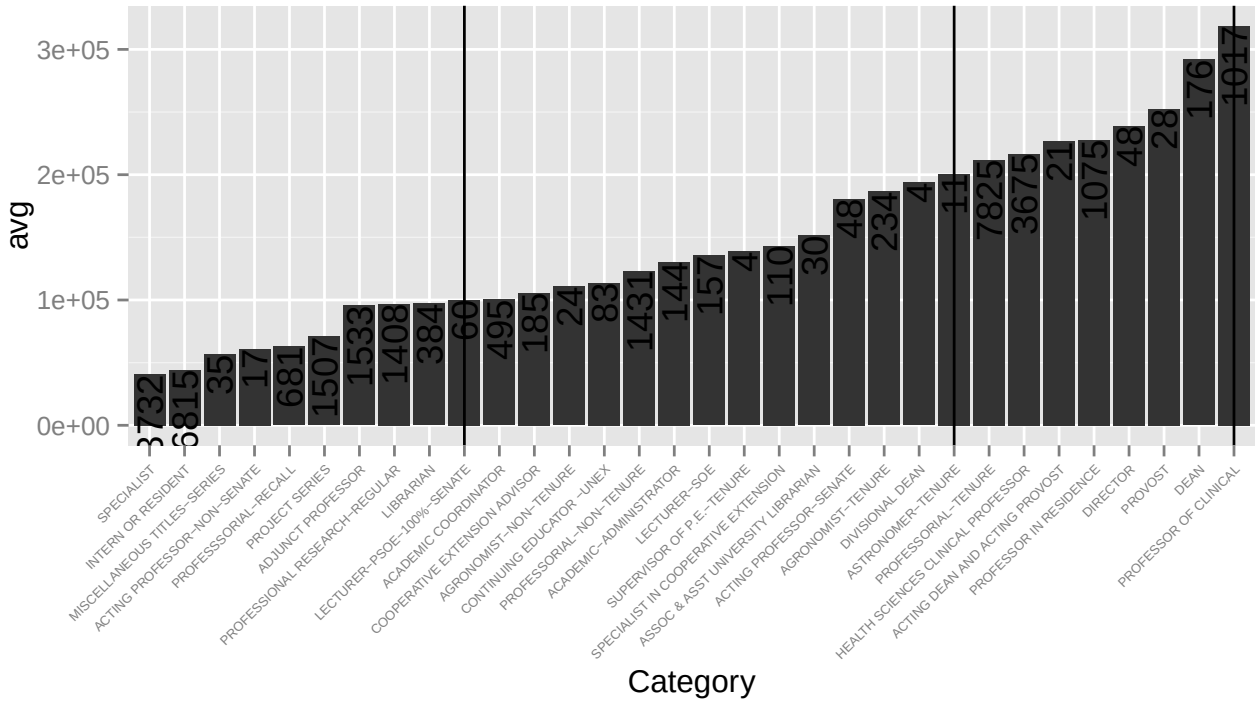
```
## Warning: Removed 1 rows containing missing values (geom_segment).
```



```
uc2013.by_department.plot
```

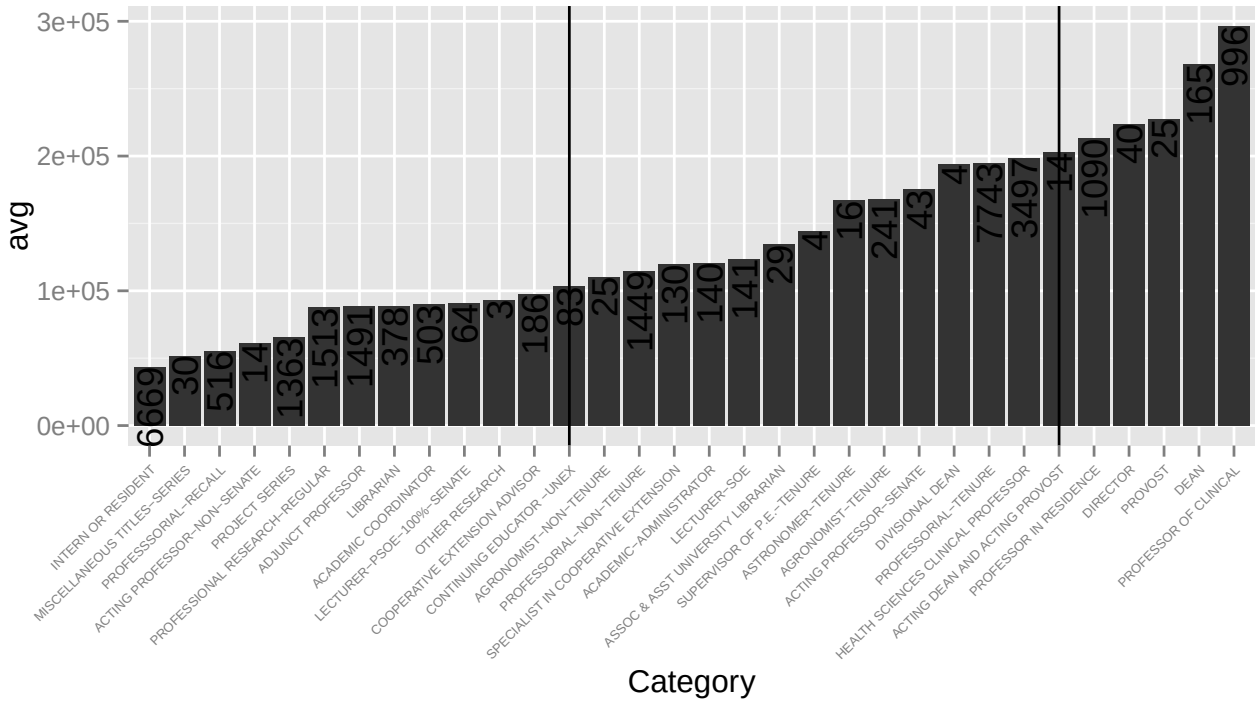



```
uc2014.by_department.plot
```




```
uc12_14.plot
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```



Conclude: the average pay of academic employees has been increasing/decreasing

Workforce Headcount vs. Total Compensation Trends, 2012-2014

Hypothesize that as headcount rises, compensation expected to rise. Is this tied to changes in student enrollment?

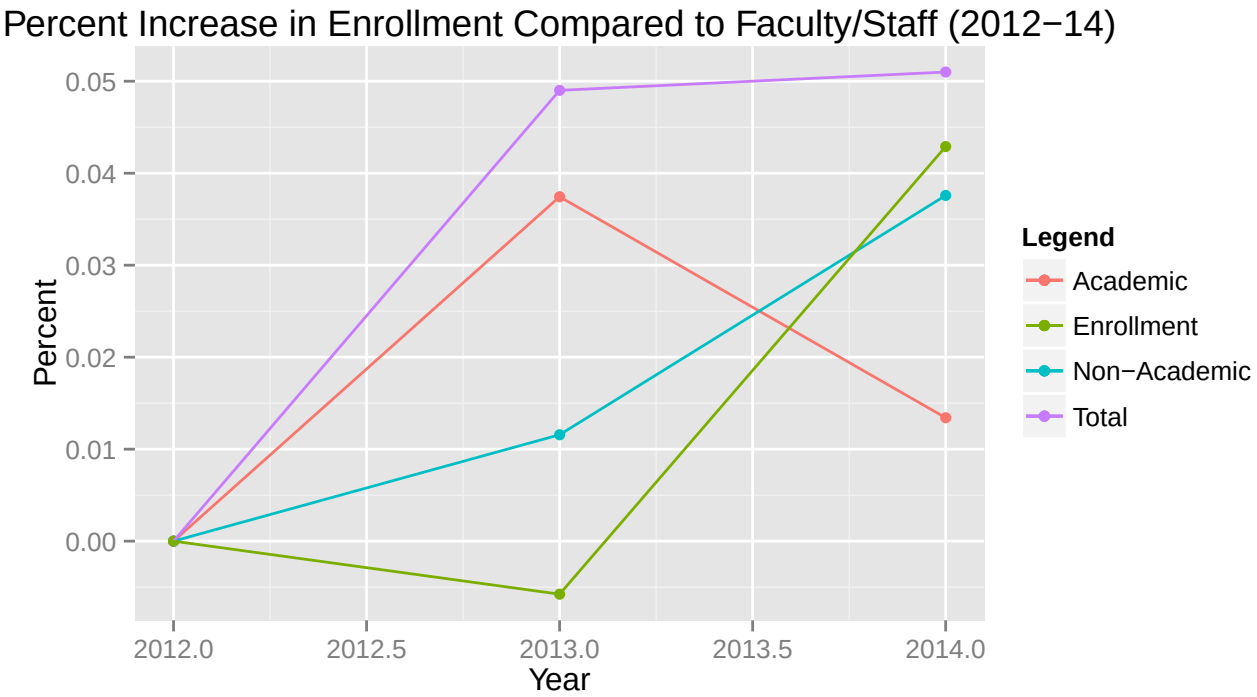
Components of Total Compensation for Academic vs Staff

Hypothesize more benefits for academic positions? E.g. they may have the same salary, but total compensation differs because of benefits.

Student Services Compensation vs. UC Student Enrollment, 2012-2014

Differences in % changes How many student services staff per 1,000 students How much has student services compensation changed per 1,000 students

Workforce Headcount vs. UC Student Enrollment, 2012-2014



Hypothesize more headcount with more enrollment

Workforce Compensation vs. UC Student Enrollment, 2012-2014

Hypothesize more compensation with more enrollment. Since academic pay has decreased from 2013-2014, how are costs being contained? See next graph

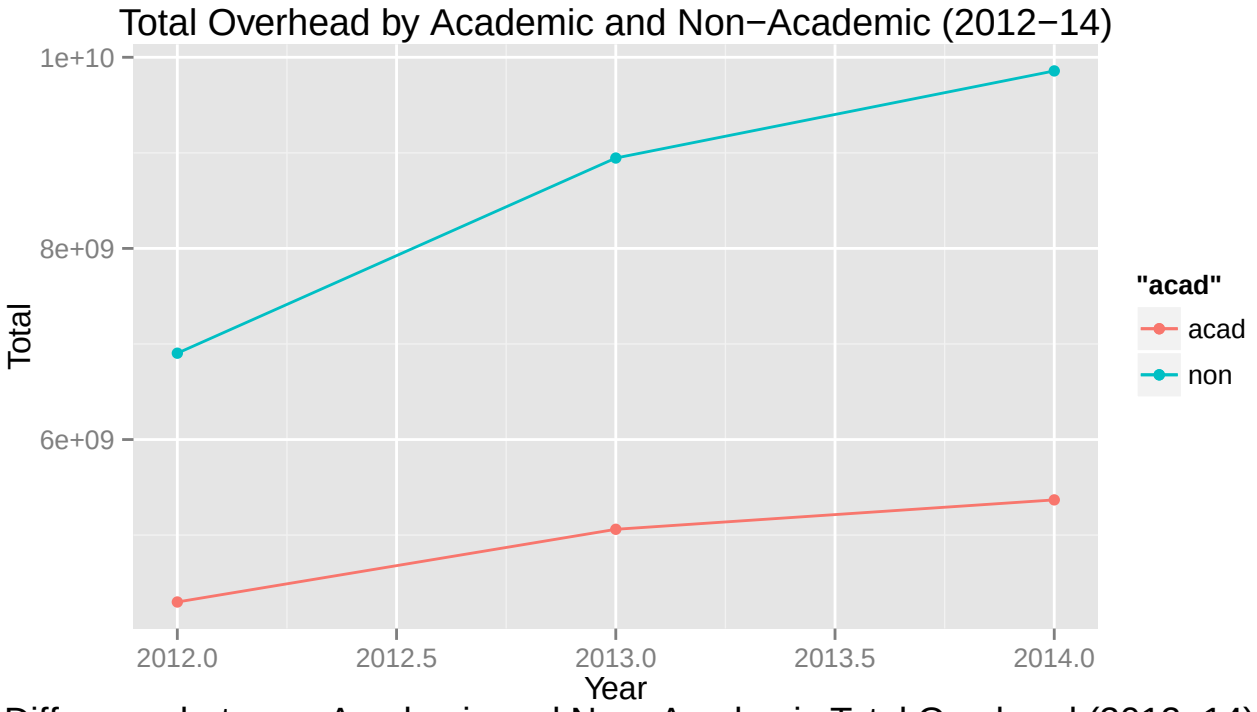
Changes in % Lecturers vs % Tenured Professors

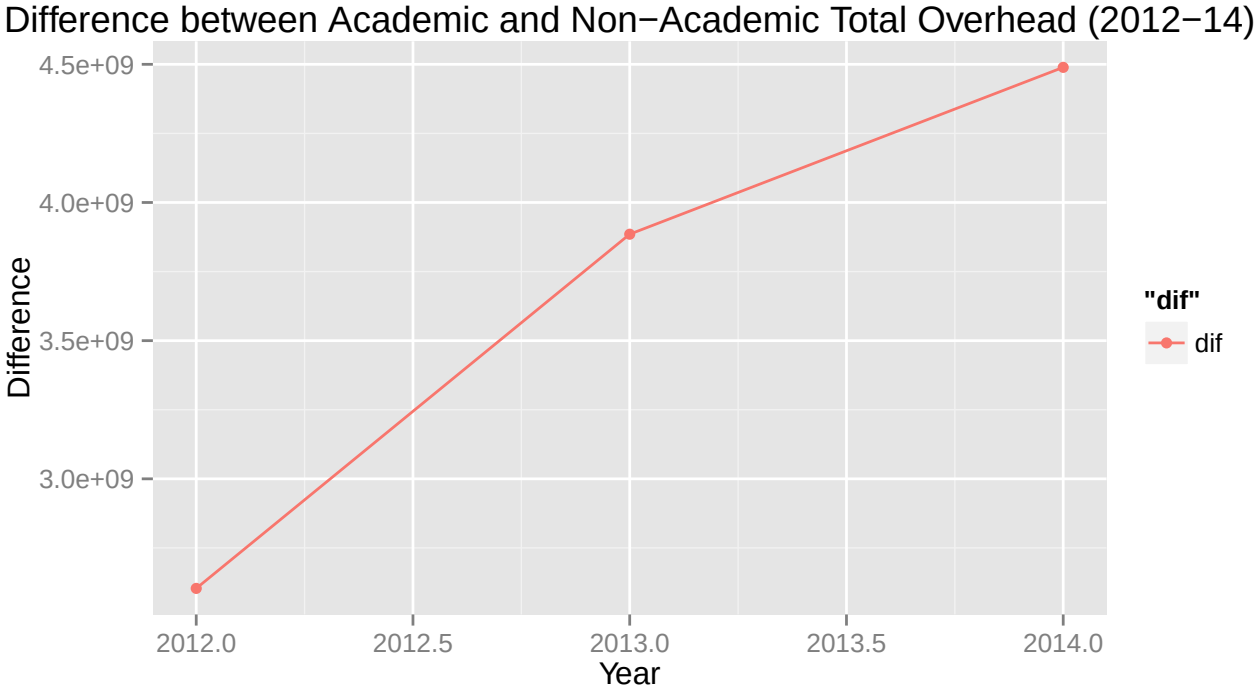
Are lecturers replacing full-time professors (or vice versa) in effort to contain costs?

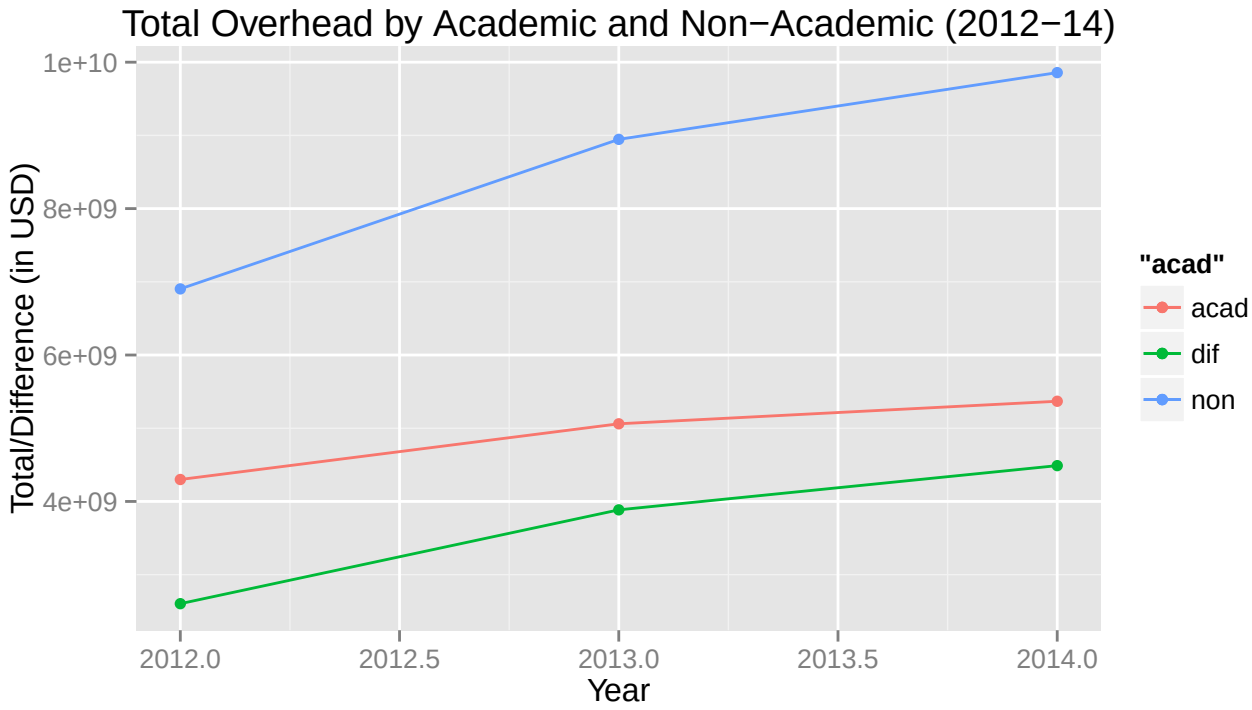
Changes in Tuition vs Academic Spending

It is often perceived that faculty spending is driving tuition increases. Is tuition increasing at the same rate as academic spending?

Total Compensation by Acad/Non-Acad over Time







Comparing Compensation for Three Types of Professor