Statistics 133 Final Report:
Salary Discrepancy Amongst University of California

Group Name: IRS (Internal Revenue Students)
Members: Omar Buenrosto, Franklin Chan, Stephen Hsu

1. Introduction

On April 11, 2016, Chancellor Nicholas Dirks sent a memo to employees informing them of job reductions and said the reductions will amount to approximately 6 percent of current staff workforce over the course of two years. With 8,500 employees, a 6 percent staff reduction equates to around 500 staff jobs. (Asimov) Berkeley plans to eliminate these 500 staff jobs over the course of two years to help balance its budgets by 2019. Combined with ever-increasing tuition fees, the question arises to how much are university employees actually being paid? Have the employee pay increased dramatically in order to warrant these layoffs or have they remained at a similar pay level? Furthermore, why are none of the other universities announcing mass layoffs? Throughout life, statistical analysis often have real life implications, and by looking into these issues, we look to get an insight into real world data involving real people and real salaries.

In this project, we analyze data collected public information provided by websites such as identical public record requests by the San Francisco Chronicle in UC Pay global (http://ucpay.globl.org/) to examine (1) What is the discrepancy between the UC system pays for employees (2) What is the discrepancy between different job titles within the UC system? (3) Where does the discrepancy between all of these lie - base pay, benefit, or overtime? In order to answer these questions, we will be looking at job title, base pay, total pay, total pay and benefits, and agency provided by the given data sets.

To answer these questions, we first analyzed a sample aggregate amount of spending from each college in respect to employee pay. Second, we explored how the pay was divided amongst different jobs and roles within that respective college. Afterwards, we compared UC Berkeley with other colleges to clearly and concisely see the breakdown of job pay.

2. Data Collection

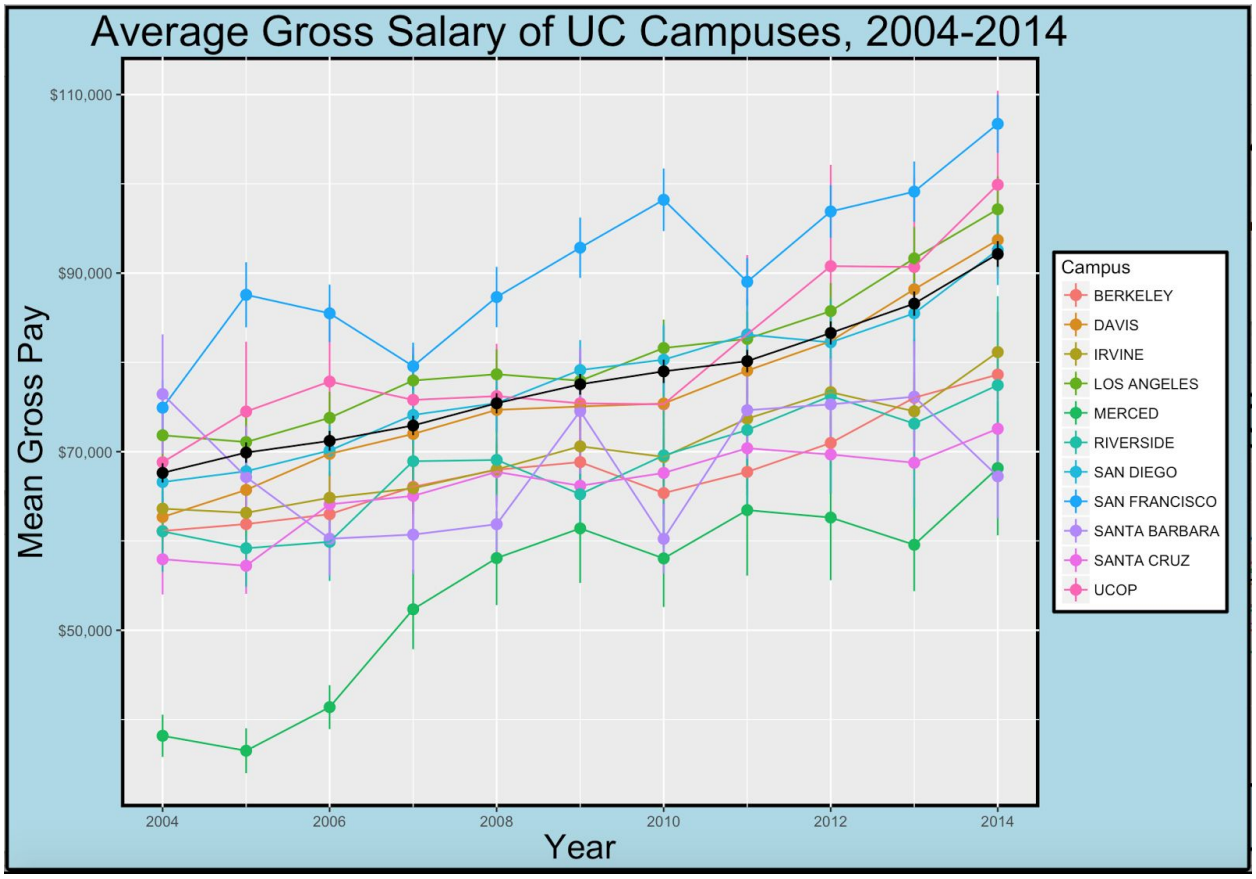
In order to present in a clear and coherent presentation, we first did web scraping from the UC Pay website mentioned, which presents the information as a table. Using what we learned in class about scraping from multiple pages, we were able to get the needed variables: Year, Campus, Name, Title, Base Pay, Overtime Pay, Extra Pay, and Gross Pay. There were several thousand pages of data, so we decided to sample proportionally by campus from the table online. For example, with Berkeley, there were over 7 thousand pages with 350,000 entries; so we sampled 2 percent of that. This was achieved by first scraping the number of pages to indirectly get the size of each table corresponding to a campus. Then we scraped 2% of each respective table and compiled the result. Something to note in this step is that we used a parallelized version of lapply to speed up the scraping and compiling of the data tables. Through parallelization, we were able to achieve our results within a more efficient manner.

#	Year	Campus	Name	Title	Base Pay	Overtime Pay	Extra Pay	Gross Pay
1.	2014	LOS ANGELES	MORA , JAMES LAWRENCE	INTERCOL ATH HEAD COACH EX	$300,000.00	$0.00	$3,176,127.00	**$3,476,127.00**
2.	2011	BERKELEY	TEDFORD , JEFF	HEAD COACH 5	$225,000.00	$0.00	$2,659,880.25	**$2,884,880.25**
3.	2007	BERKELEY	TEDFORD , JEFF	HEAD COACH-INTERCOLG ATHLETICS	$225,000.00	$0.00	$2,606,653.50	**$2,831,653.50**
4.	2014	LOS ANGELES	ALFORD , STEPHEN TODD	INTERCOL ATH HEAD COACH EX	$300,000.00	$0.00	$2,445,341.00	**$2,745,341.00**
5.	2013	LOS ANGELES	ALFORD , STEPHEN TODD	INTERCOL ATH HEAD COACH EX	$200,000.00	$0.00	$2,439,609.00	**$2,639,609.00**
6.	2013	BERKELEY	TEDFORD , JEFF	HEAD COACH 5	$0.00	$0.00	$2,442,860.00	**$2,442,860.00**
7.	2013	LOS ANGELES	MORA , JAMES LAWRENCE	INTERCOL ATH HEAD COACH EX	$300,000.00	$0.00	$2,110,128.00	**$2,410,128.00**
8.	2013	BERKELEY	DYKES , DANIEL	HEAD COACH 5	$246,031.00	$0.00	$2,124,708.00	**$2,370,739.00**
9.	2010	BERKELEY	TEDFORD , JEFF	HEAD COACH 5	$225,000.00	$0.00	$2,124,037.96	**$2,349,037.96**
10.	2008	BERKELEY	TEDFORD , JEFF	HEAD COACH-INTERCOLG ATHLETICS	$225,000.02	$0.00	$2,117,314.50	**$2,342,314.52**
11.	2009	BERKELEY	TEDFORD , JEFF	HEAD COACH 5	$225,000.00	$0.00	$2,113,409.39	**$2,338,409.39**
12.	2013	LOS ANGELES	HOWLAND , BENJAMIN CLARK	INTERCOL ATH HEAD COACH EX	$300,000.00	$0.00	$2,015,078.00	**$2,315,078.00**
13.	2014	LOS ANGELES	TABSH , KHALIL M	HS CLIN PROF-HCOMP	$179,402.00	$0.00	$2,123,925.00	**$2,303,327.00**
14.	2012	LOS ANGELES	HOWLAND , BENJAMIN CLARK	INTERCOL ATH HEAD COACH EX	$300,000.00	$0.00	$1,934,191.48	**$2,234,191.48**

After obtaining the mass public record of salaries and department employee names, the data was now ready to be analyzed and subsequently presented in a readable graph.

3. Data Visualizations and Analysis

Graph 1: The average gross salary of UC campuses throughout 2004 to 2014

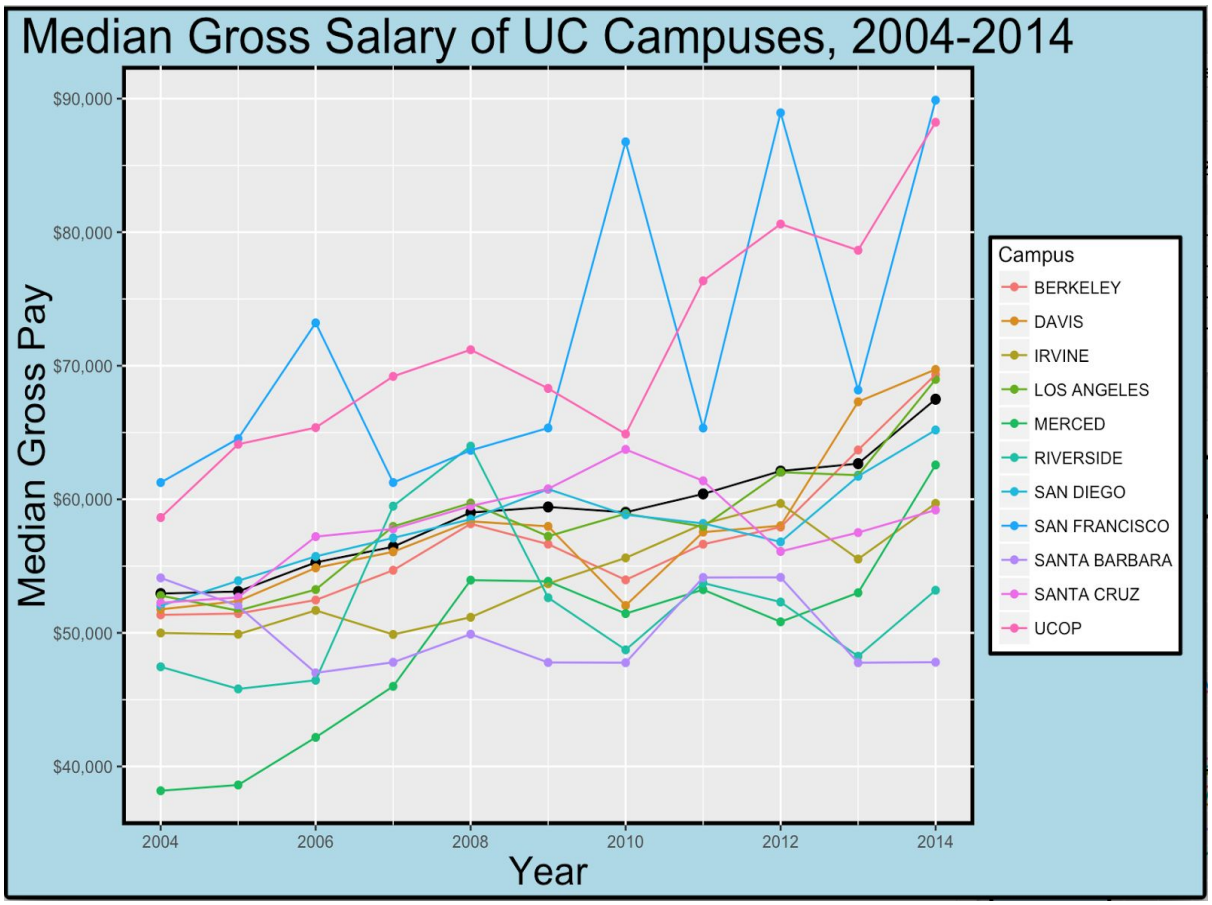


Our first step was to understand how much universities are actually paying in aggregate for their employees (Graph 1). We used summarise to find means of the various salary data based on campus and year. Graphing this leads to Graph 1. Some notes: the vertical bars are the standard error on our averages, which are taken per campus and per year. This is to make sure our sampling is not the reason why we have discrepancies when we compare campuses.

From the first graph, we can note that the University of San Francisco and UCOP (University of California, Office of the President) are the top spenders while Merced was clearly the lowest spender. The other universities tended to stay in two groups, one higher (Los Angeles, Davis, San Diego) and one lower (Berkeley, Irvine, Riverside, Santa Barbara, Santa Cruz). While UCOP and UCSF are not colleges with undergraduate enrollment, the rest are; interestingly all the better ranked universities are grouped together in the higher group except for

Berkeley. This was to our surprise as our initial hypothesis would be that Berkeley would be one of the larger spenders on employee salaries.

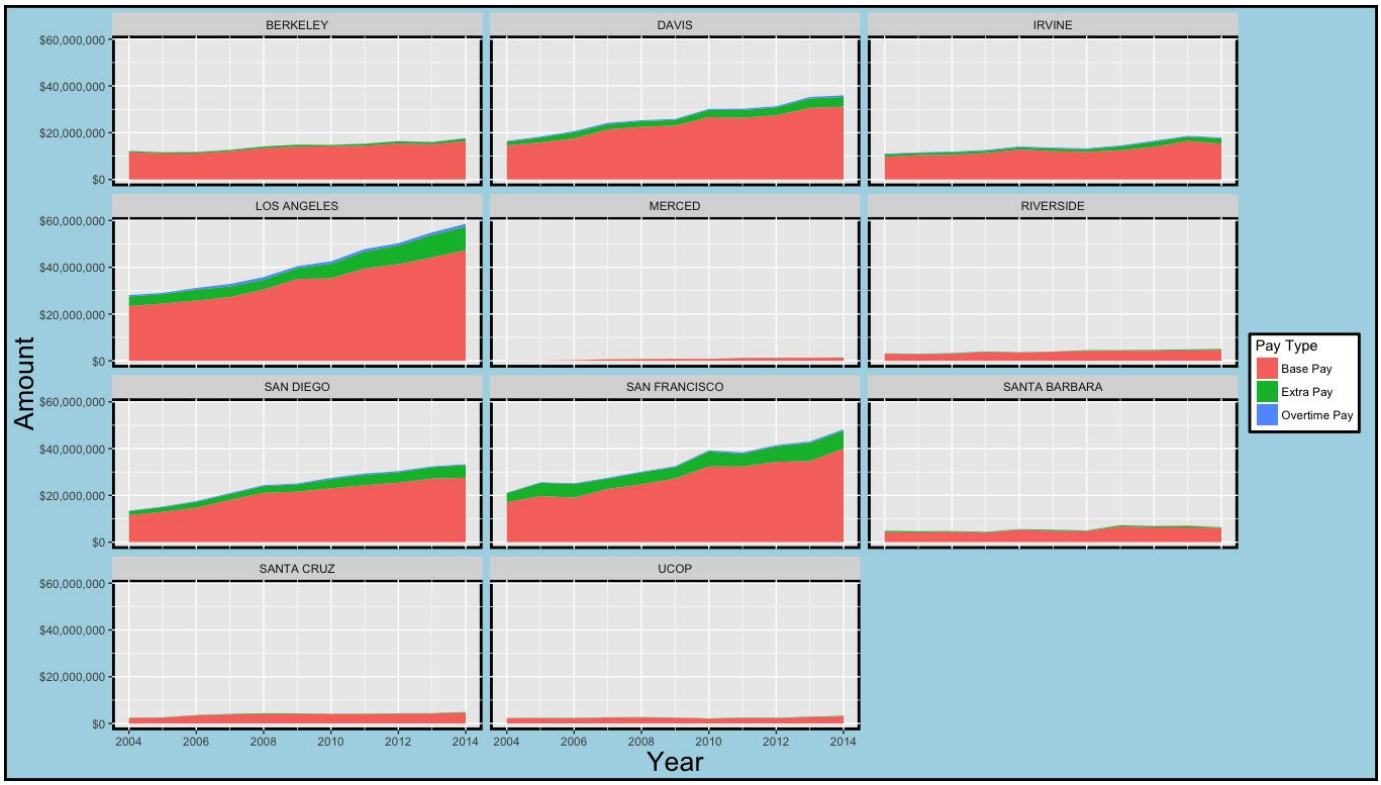
Graph 2: Median Salary



Surprising to us was that Berkeley was relatively low in terms of average salary. To investigate, we decided to look at the medians as well to make sure there were no hidden variables like an unproportional amount of outliers sampled. Our result, Graph 2, was done similarly to Graph 1, but with the median pay instead of the mean pay. Furthermore, there was no standard error included since the standard error is inapplicable to medians. In particular we saw that Berkeley hovers somewhere about the middle of all the universities. In this graph, Berkeley was the fourth highest median gross pay, compared to seventh highest average pay. This could imply that Berkeley employees are paid more but the higher paid employees are not paid as much. Also to note is that yet again, UCSF and UCOP seem to have the highest mean pay, followed by the other higher ranked colleges (Davis, UCLA, San Diego). This is potentially

because UCSF enrolls no undergraduates and is primarily a graduate school focused on medical practices.

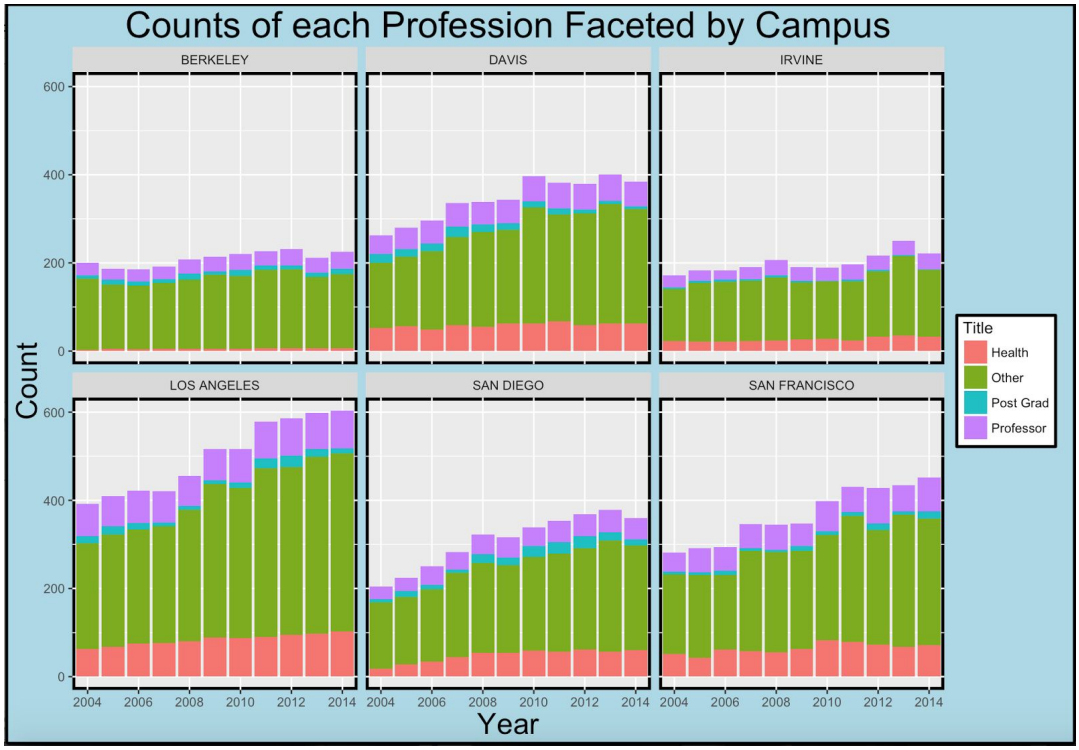
Graph 3: Amount Spent on Employee Gross, Overtime, and Extra Pay



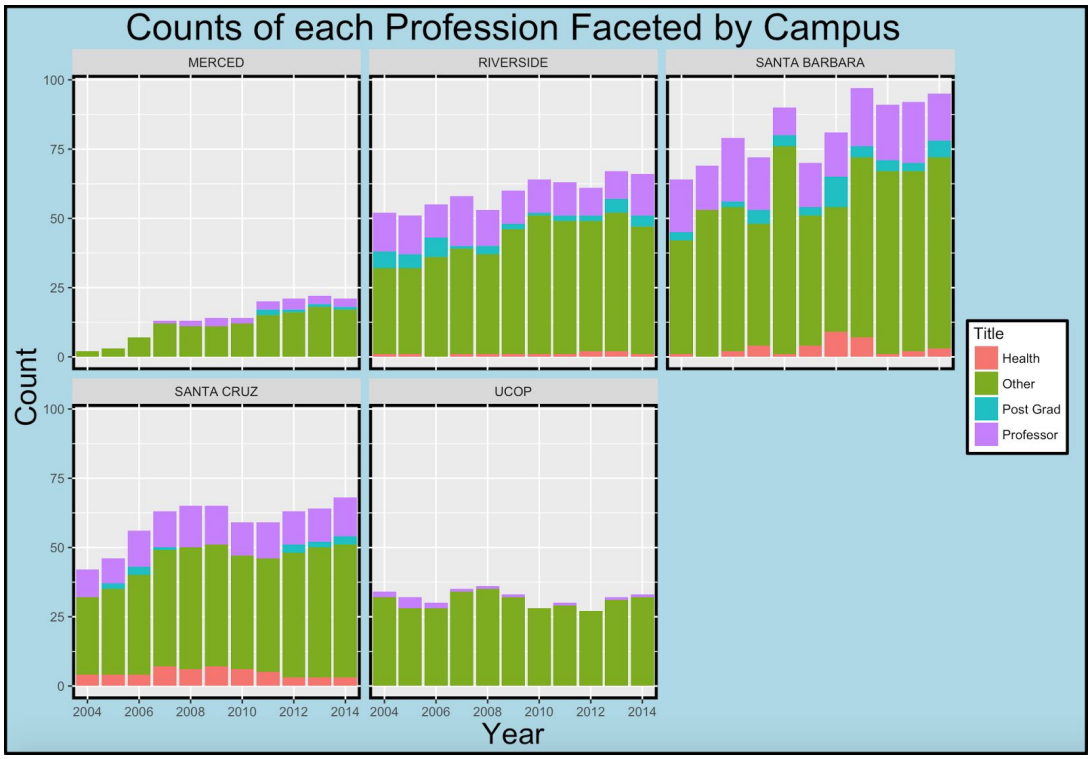
To create Graph 3 we simply summed the gross pay by campus and by year. The purposes of this graph was to see an overall breakdown of how salaries were allocated. In particular it became evident that of the campuses with the highest aggregate salaries, extra pay was making up a significant chunk of of it.

Besides UCSF, Davis, and San Diego Los Angeles, extra pay was almost unnoticeable in the amount spent. Furthermore, this graph makes it easier to see the trends of base and extra pay over the years. Berkeley has remained at a straighter line than the other top ranked universities, showing that its pay has not changed over the years, even though it would be natural for it to curve up due to inflation. Also to note is that overtime pay made up a tiny percentage of the total amounts.

Graph 4: Proportion of Job Titles within Sample Size Collected



Graph 4 (cont.):

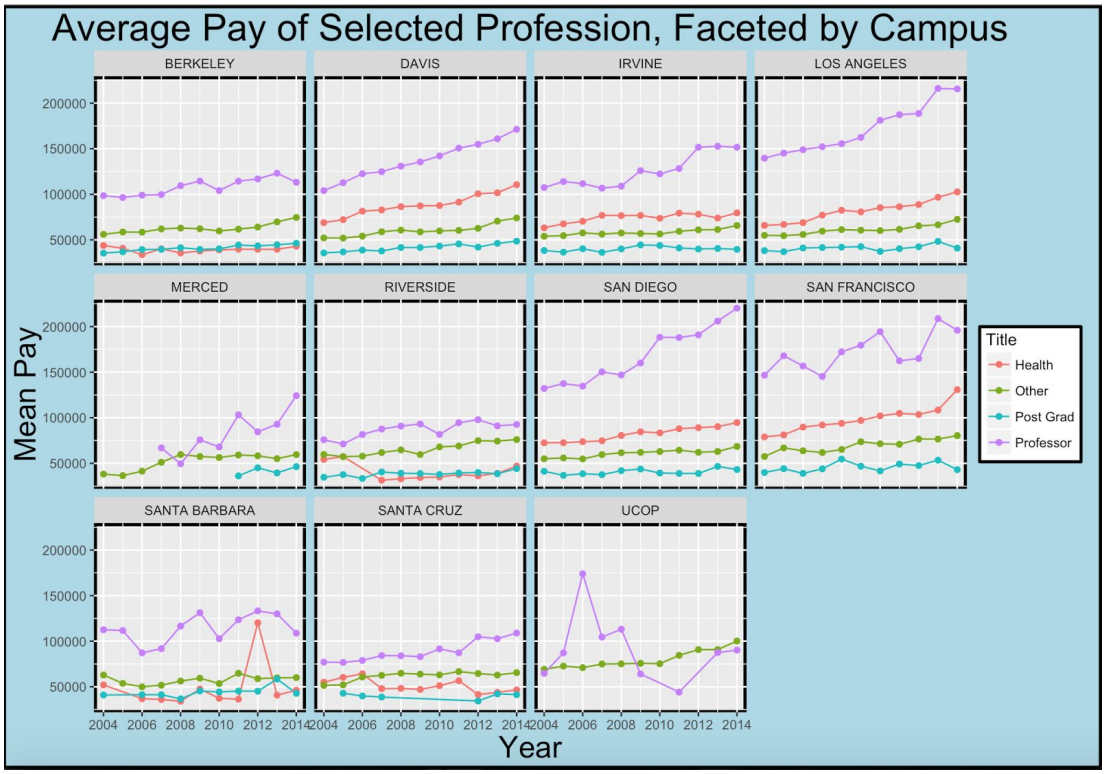


After developing an intuition as to how much each university is paying and realizing that Berkeley is neither the top nor the bottom spender towards employees, we decided to go further into the differences between job title pays. By further researching into the job title pay differences, we intended to see if there was any exorbitant salaries amongst either administration, graduate students, health workers, or other types of workers.

Because the universities all had similar spendings except for UCSF and UCOP (which are both non-undergraduate infrastructures), we wanted to compare side by side all universities pay towards respective titles. Because we sampled the universities, we found it necessary to see how many individuals with certain titles were selected. Therefore, Graph 4 was created to show the percentage of different employees per university. We sorted job titles into the Categories Health, Other, Post Grad, and Professor, and then totaled each amount to create a breakdown of professions.

To our surprise, there was the lack of Berkeley employees who were in health. Since health jobs such as nurses and practitioners tend to have a higher salary, this could be a reason why Berkeley's employee spendings appear lower. To evaluate if this is true, Graph 5 reflects the average salaries of each profession, where again Berkeley seems to be at the bottom.

Graph 5: Pay Broken Down by Profession and Campus



4. Concluding Remarks

Using year, gross pay, and campus variables collected from http://ucpay.globl.org/, we explored the determinants of money spent per year per campus. Our analyses suggests that the University of California, Berkeley does not spend the most nor the least compared to the universities. We can conclude this because from the average and median graph, Berkeley is not the top spender, and in fact, has standard spending compared to other universities. Potential outliers would not have a large role since the median is less affected by such points. And when broken down, the delegation of resources towards Berkeley faculty's base pay, extra pay, and overtime pay has been relatively linear throughout the past decade. While other colleges have all seen an increase in average pay per title, especially in the professor department, Berkeley has remained stagnant. Such results are alarming because of news of cutting faculty down 6 percent. The results found in this report suggest Berkeley is an average spender and has not spent any superfluous amounts on its professors, post graduates, health workers, or other workers; has not increased its pay substantially over the past decade. Therefore, from this report, Berkeley should look into other ways to balance its budget before cutting jobs that have cost the school as much as it has historically.

5. Appendix and Code

Table 1: Mean and median salaries of all campuses

	Year	mean	mid
1	2004	67631.52	52944.48
2	2005	69899.74	53098.08
3	2006	71214.41	55273.75
4	2007	72930.47	56450.01
5	2008	75424.81	59012.29
6	2009	77550.63	59428.53
7	2010	79001.08	59030.61
8	2011	80143.76	60403.10
9	2012	83295.98	62118.00
10	2013	86584.48	62670.50
11	2014	92139.43	67495.00

Table 2: mean and median of individual colleges per year

	Campus	Year	mean	mid
1	BERKELEY	2004	61121.30	51351.54
2	BERKELEY	2005	61893.95	51446.16
3	BERKELEY	2006	62997.91	52458.07
4	BERKELEY	2007	66060.54	54689.35
5	BERKELEY	2008	67972.69	58168.14
6	BERKELEY	2009	68833.70	56643.67
7	BERKELEY	2010	65360.15	53966.44
8	BERKELEY	2011	67734.34	56638.01
9	BERKELEY	2012	70978.65	57918.96
10	BERKELEY	2013	76047.06	63694.00
11	BERKELEY	2014	78633.55	69343.00
12	DAVIS	2004	62682.28	51765.00
13	DAVIS	2005	65719.10	52375.98
14	DAVIS	2006	69764.23	54862.47

Showing 1 to 15 of 121 entries

Code:

#load necessary tables

```{r}
library(XML)
library(rvest)
library(parallel)
library(plyr)
library(ggplot2)
```

Taking a random sample of size of 50 pages of data from each UC. Each page has 50 cases (except the last).

```{r}
campuses <- c("BERKELEY", "LOS+ANGELES", "SAN+DIEGO", "DAVIS", "IRVINE", "MERCED", "RIVERSIDE", "SAN+FRANCISCO", "UCOP", "SANTA+CRUZ", "SANTA+BARBARA")

names(campuses) <- campuses

campusURL <- function(campus){
```

```
  url1 <- "http://ucpay.globl.org/index.php?campus="
  url2 <- "&name=&title=&base=&overtime=&extra=&gross=%3E30000&year=&s=gross"
  paste(url1,campus,url2,sep="")
}

setSample <-function(campus,seed=1234){
  set.seed(seed)
  pages <- campusURL(campus) %>% htmlParse() %>%
getNodeSet('//p[@class="pagelinks"]/a') %>% sapply(xmlValue)
  sample(as.numeric(pages[10]), floor(as.numeric(pages[10]) * 0.02))
}

samples<-lapply(campuses,setSample)
```

Generates a List of Data Tables
```{r}
# Function that scrapes pages from UCPay
#http://ucpay.globl.org/index.php?campus=&name=&title=&base=&overtime=&extra=&gross=
>30000&year=&s=gross
gettable <- function(pagenum,campus){
  tables<-read_html(paste(campusURL(campus),"&p=",pagenum, sep="")) %>%
html_nodes(xpath="//table") %>% html_table(fill=TRUE)
  tables[[4]]
}

UCtable <- function(campus,samples){
  # A random Sample
  randomPages <- samples[[campus]]
  lapply(randomPages,gettable,campus) %>% rbind.fill()
}

```

Compute The List of Wage Data Frames from each UC campus in parallel
```{r}
UCtableList <- campuses %>%
  mclapply(UCtable,samples, mc.preschedule = FALSE)
```

##turn the UCtableList into a dataframe
```{r}
UCdataFrame<-rbind.fill(UCtableList) %>%
  mutate(`Base Pay`=as.numeric(gsub("[$,]","",`Base Pay`)),
         `Overtime Pay`=as.numeric(gsub("[$,]","",`Overtime Pay`)),
         `Extra Pay`=as.numeric(gsub("[$,]","",`Extra Pay`)),
         `Gross Pay`=as.numeric(gsub("[$,]","",`Gross Pay`)))
```

###clean the table names by getting rid of commas and middle names
```{r}
UCdataFrame$Name <- UCdataFrame$Name %>% gsub(",","", .) %>% gsub("^([^ ]+) +([^
]+).*", "\\2 \\1", .)
```

Line Plots
```{r}
totalsalaries <- UCdataFrame %>%
     group_by(Year) %>%
     summarize(mean = mean(`Gross Pay`), mid = median(`Gross Pay`))

totalsalariesYearly <- UCdataFrame %>%
     group_by(Campus, Year) %>%
     summarize(mean = mean(`Gross Pay`), mid = median(`Gross Pay`))

#create a new theme
abluetheme <- theme(plot.background = element_rect(fill = "lightblue", colour = "black", size =
2, linetype = "solid"), legend.background=element_rect(colour = "black", size = 1, linetype =
"solid"), panel.background=element_rect(colour = "black", size = 2, linetype = "solid"),
plot.title=element_text(size=15))

#create average gross salary of UC campuses from 2004-2014
avglplot <- UCdataFrame %>%
  group_by(Year) %>%
  summarize(mean = mean(`Gross Pay`), mid = median(`Gross Pay`), SE = sd(`Gross Pay`))
%>%
  ggplot(aes(x=Year, y=mean)) + geom_line() + geom_point(size=2) +
```

```
  geom_smooth(linetype=0, fill="light blue")+ geom_point(data=totalsalariesYearly, aes(x=Year,
y=mean, col=Campus)) +
  stat_summary() +
  geom_line(data=totalsalariesYearly, aes(col=Campus, group=Campus)) +
  labs(title="Average Gross Salary of UC Campuses, 2004-2014") + ylab("Mean Gross Pay") +
scale_x_continuous(breaks=c(2004, 2006, 2008, 2010, 2012, 2014), labels=c(2004, 2006, 2008,
2010, 2012, 2014)) +
  scale_y_continuous(labels = scales::dollar) +
  abluetheme

#create average gross salary of UC campuses from 2004-2014 with standard errors
avglplot <- UCdataFrame %>%
  ggplot(aes(x=Year, y=`Gross Pay`))  +
  stat_summary(aes(group=Campus, col=Campus)) +
  stat_summary(geom="line", aes(group=Campus, col=Campus)) +
  stat_summary() +
  stat_summary(geom="line") +
  labs(title="Average Gross Salary of UC Campuses, 2004-2014") + ylab("Mean Gross Pay") +
scale_x_continuous(breaks=c(2004, 2006, 2008, 2010, 2012, 2014), labels=c(2004, 2006, 2008,
2010, 2012, 2014)) +
  scale_y_continuous(labels = scales::dollar) +
  abluetheme

#create mean salary plot
avglplotYearly <- totalsalariesYearly %>% ggplot(aes(x=Year, y=mean)) +
geom_line(aes(col=Campus, group=Campus)) + geom_point()  + labs(title="Mean Salary") +
ylab("Mean") + abluetheme

#create median plot
midplot <- UCdataFrame %>%
  group_by(Year) %>%
  summarize(mean = mean(`Gross Pay`), mid = median(`Gross Pay`), SE = sd(`Gross Pay`))
%>%
  ggplot(aes(x=Year, y=mid)) + geom_line() + geom_point(size=2) +
  geom_smooth(linetype=0, fill="light blue")+ geom_point(data=totalsalariesYearly, aes(x=Year,
y=mid, col=Campus)) +
  stat_summary() +
  geom_line(data=totalsalariesYearly, aes(col=Campus, group=Campus)) +
```

```
  labs(title="Median Salary of UC Campuses, 2004-2014") + ylab("Median Gross Pay") +
scale_x_continuous(breaks=c(2004, 2006, 2008, 2010, 2012, 2014), labels=c(2004, 2006, 2008,
2010, 2012, 2014)) +
  scale_y_continuous(labels = scales::dollar) +
  abluetheme

```

Stacked plot plot
```{r}
wideUCdataFrame <- UCdataFrame %>%
  gather(`Pay Type`, Amount, `Base Pay`, `Overtime Pay`, `Extra Pay`) %>%
  group_by(Campus, Year, `Pay Type`) %>%
  summarise(Amount = sum(Amount))

df <- wideUCdataFrame %>%
  ungroup() %>%
  unite(CampusPay, Campus, `Pay Type`, sep=" ")

stackedWide <- wideUCdataFrame  %>%
  filter(Campus=="BERKELEY") %>%
  ggplot(aes(x=Year, y=Amount, fill=`Pay Type`)) + geom_area(position = "stack") +
  abluetheme + scale_y_continuous(labels = scales::dollar)

stackedPlotCampus <- df %>%
  ggplot(aes(x=Year, y=Amount, fill=CampusPay)) + geom_area(position = "stack") +
abluetheme + scale_y_continuous(labels = scales::dollar)
```
#Profession wages between different universities
```{r}
patternprof <- "PROF(?!ESSIONAL)|LECT"
patternpostgrad <- "POST"
patternhealth <- "(NURSE|CUSTODIAN|MECHANIC)"

searchPattern <- function(str){
    if (grepl(patternprof, str, perl=TRUE)) {
      return("Professor")
    } else if (grepl(patternpostgrad, str, perl=TRUE)) {
      return("Post Grad")
```

```
    } else if (grepl(patternhealth, str, perl=TRUE)) {
      return("Health")
    } else {
      return("Other")
    }
}

searchPatternVect <- function(vec) {
  return(sapply(X=vec, FUN=searchPattern))
}

UCLA_BerkDF <- UCdataFrame %>%
  filter(Campus=="LOS ANGELES" | Campus == "BERKELEY") %>%
  mutate(Title=searchPatternVect(Title))

BerkvsLA <- UCLA_BerkDF %>%
  group_by(Campus, Year, Title) %>%
  summarise(Average=mean(`Gross Pay`)) %>%
  ggplot(aes(x=Year, y=Average)) +
  geom_line(aes(col=Title)) +
  geom_point(aes(col=Title)) +
  facet_wrap(~ Campus) + abluetheme +
  labs(title="Average Pay of Selected Profession: UC Berkeley vs. UCLA") + ylab("Mean Pay")

profAvglplot <- UCdataFrame %>%
  mutate(Title=searchPatternVect(Title)) %>%
  group_by(Campus, Year, Title) %>%
  summarise(Average=mean(`Gross Pay`)) %>%
  ggplot(aes(x=Year, y=Average)) +
  geom_line(aes(col=Title)) +
  geom_point(aes(col=Title)) +
  facet_wrap(~ Campus) + abluetheme +
  labs(title="Average Pay of Selected Profession, Faceted by Campus") + ylab("Mean Pay")

bigNames <- c("BERKELEY", "DAVIS", "IRVINE", "SAN DIEGO", "SAN FRANCISCO",
"LOS ANGELES")

professionCountPlot1 <- UCdataFrame %>%
  filter(Campus %in% bigNames) %>%
```

```
  mutate(Title=searchPatternVect(Title)) %>%
  ggplot(aes(x=factor(Year))) + geom_bar(aes(fill=Title)) + scale_x_discrete(breaks=c(2004,
2006, 2008, 2010, 2012, 2014), labels=c(2004, 2006, 2008, 2010, 2012, 2014)) + facet_wrap(~
Campus) + abluetheme + labs(title="Counts of each Profession Faceted by Campus") +
ylab("Count") + xlab("Year")

professionCountPlot2 <- UCdataFrame %>%
  filter(!(Campus %in% bigNames)) %>%
  mutate(Title=searchPatternVect(Title)) %>%
  ggplot(aes(x=factor(Year))) + geom_bar(aes(fill=Title)) + scale_x_discrete(breaks=c(2004,
2006, 2008, 2010, 2012, 2014), labels=c(2004, 2006, 2008, 2010, 2012, 2014)) + facet_wrap(~
Campus) + abluetheme + labs(title="Counts of each Profession Faceted by Campus") +
ylab("Count") + xlab("Year")

professionCountTable <- UCdataFrame %>%
  mutate(Title=searchPatternVect(Title)) %>%
  group_by(Campus, Title) %>%
  summarise(Total=n())

```

6. Bibliography

Asimov, Nanette. “UC Berkeley to Eliminate 500 Staff jobs.” *Sfgate* [San Francisco, CA]. 12 April 2016. Print.