Assignment#3

Graham Tanner Robart

02/22/2016

Exploring Public Florida SUS Salary Data

Open records law in the state of Florida mandate that salaries paid to all state employees be a matter of public record. In particular, salaries for all employees of the State University System can be found here. (Downloaded: 02/20/2016)

In this Document, we will examine the salary data of New College compared to other universities, as well as the differences between administration and faculty salaries, and provide supporting visualizations

Preliminary Findings

We have observed that new college faculty make less than their SUS counterparts for median and mean, and have a much lower upper bound of possible salary. Secondly, administration for most universities earn less than their faculty, and New College administrators make signifcantly less on average than admins at other schools. The data overall contains numerous upper bound outliers.

Data Prep

First we make sure we have all the libraries and packages we need are installed, and set our working directory.

```
library(gridExtra)
library(data.table)
library(ggplot2)
library(knitr)
BASE_DIR <- "C:/Users/Tanner/Desktop/Data Sciences/Visualization/HW/Assignment#3/data"
setwd(BASE_DIR) #Replace BASE_DIR with your own filepath to the data
```

Now lets load the data into a data.table object and take a look at its structure.

```
setwd(BASE_DIR)
data <- fread("emp.csv")
code.types <- fread("code_types.txt")
kable(head(data))
```

University	Budget Entity	Position Number	Last Name	First Name	MI	Employee Type	FTE
FAMU	Educational & General	18503000	ABATE	RANDALL	S	SALARIED	.77
FAMU	Educational & General	18703000	ABAZINGE	MICHAEL	D	SALARIED	.7
FAMU	Contracts & Grants	18703000	ABAZINGE	MICHAEL	D	SALARIED	.3
FAMU	Educational & General	18532000	ABDELRAZIG	YASSIR	.	SALARIED	.75
FAMU	Contracts & Grants	18121000	ABLORDEPPEY	SETH	Y	SALARIED	.1
FAMU	Educational & General	18121000	ABLORDEPPEY	SETH	Y	SALARIED	.9

There do not seem to be many meaningful variables, primarily '`Annual Salary`', which is the salary in USD

for each person. 'University' is a class, one of twelve universities in the SUS. And Class.Title is the job title given Lets get rid of some of the data we are not interested in to make our table smaller to work with. (We examine only Salaried employees)

```
all.data <- data[`Employee Type` == 'SALARIED' ]
all.data[, `FTE` := as.numeric(`FTE`)]
all.data[, sum.fte := sum(FTE), by = list(`First Name`, `Last Name`, MI)]
all.data[, `Annual Salary` := as.numeric(`Annual Salary`)]
```

We wish to compare Faculty and Administration Salary distributions, So we will add a variable on whether Class.Title is an admin or faculty job using a reference list. This list was coded by hand using the top 100 most common job titles (available in code_types.txt). So we merge in our Class titles and job types.

```
all.data <- merge(all.data, code.types, by.x = 'Class Title', by.y = 'code')
kable(head(all.data))
```

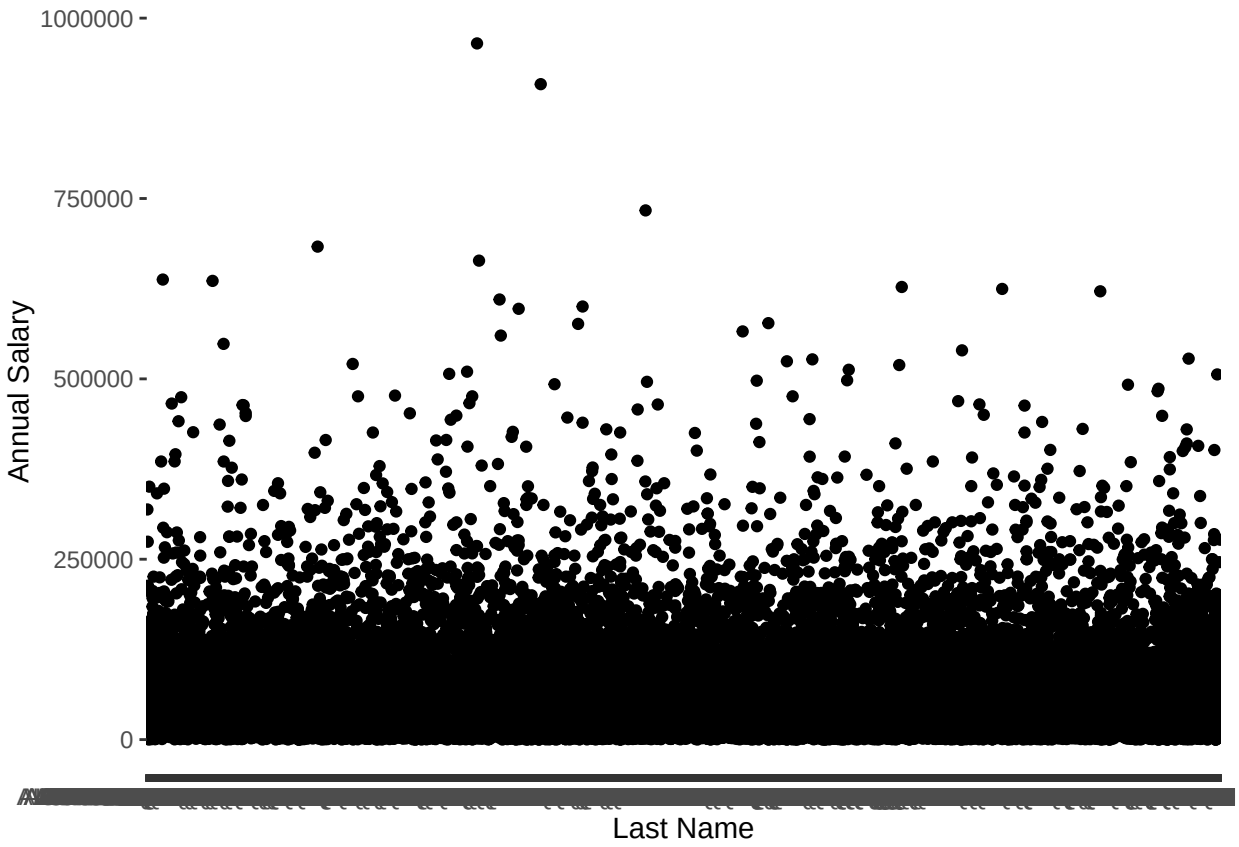
Class Title	University	Budget Entity	Position Number	Last Name	First Name	MI
ACADEMIC ADVISOR	FAU	Educational & General	980225	ANOUFRIEVA	ANNA	A
ACADEMIC ADVISOR	FAU	Educational & General	979682	CARNOT	CHRISTIANA	I
ACADEMIC ADVISOR	FAU	Educational & General	979997	CASTILLO	ELISSA	.
ACADEMIC ADVISOR	FAU	Educational & General	980373	FERGUSON	REUBEN	D
ACADEMIC ADVISOR	FAU	Educational & General	980241	GLASSER	DORRAN	E
ACADEMIC ADVISOR	FAU	Educational & General	980018	GREEN	ERIKA	R

Now to get an idea of the data, lets take a first look at the salaries for each row, without any particular order. (only run the following plot if you don't mind how long it takes to run.)

```
ggplot(data = all.data, aes(x = `Position Number`, y= `Annual Salary` )) + geom_point()
```

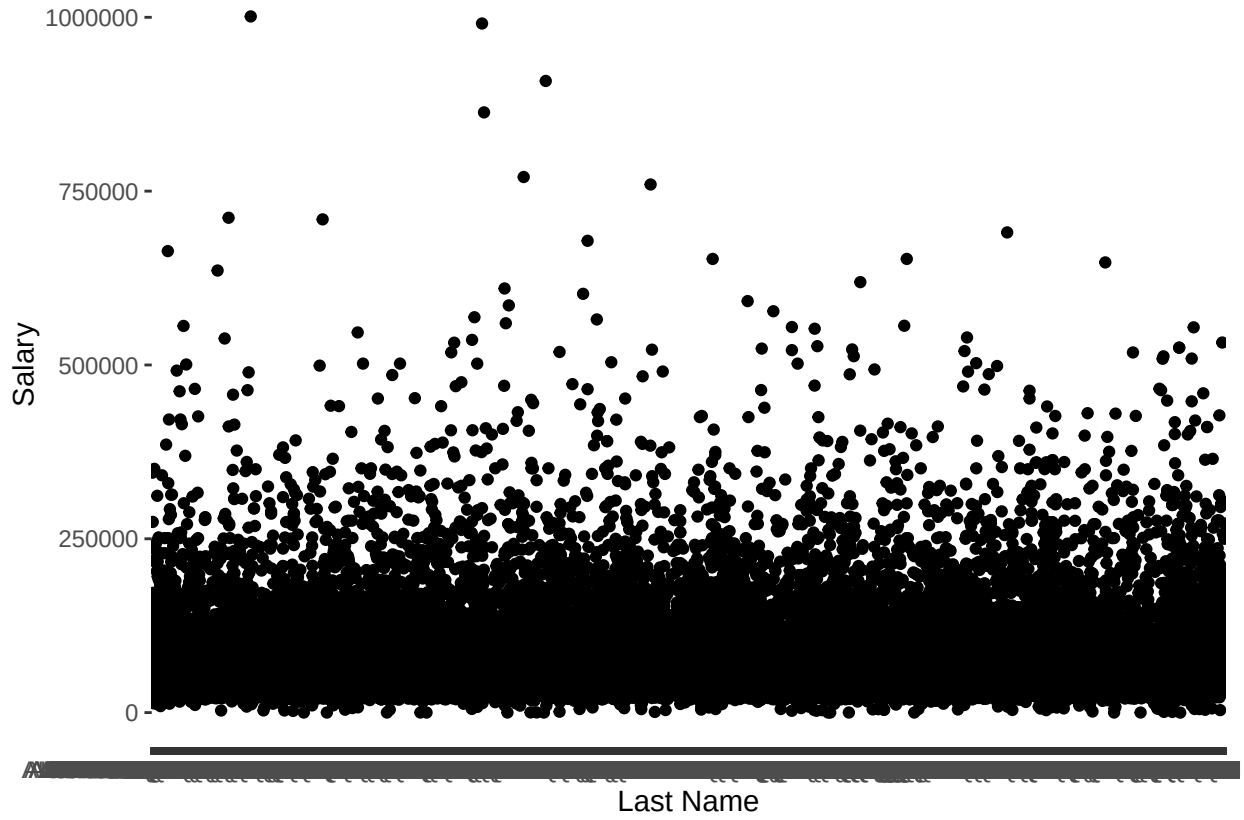
This is hard to read, so lets make it looks nicer, using labels and some empty points, so we can make out some of the overlapping data, also lets look at last Name instead of position number, since many entries have a position number of 0.

```
ggplot(data = all.data, aes(x= `Last Name` , y=`Annual Salary` , ylab= "Salary", xlab= "Employees" )) +
```



Because there are sometimes multiple entries of per person, we sum `Annual Salary` entries by position number in order to derive an overall salary per person. We are also interested in their Full Time Equivalent (FTE), usually a number between 0-1 that represents how many hours out of a full 40 are required of an employee, and which may not sum to 1.0 for all people.

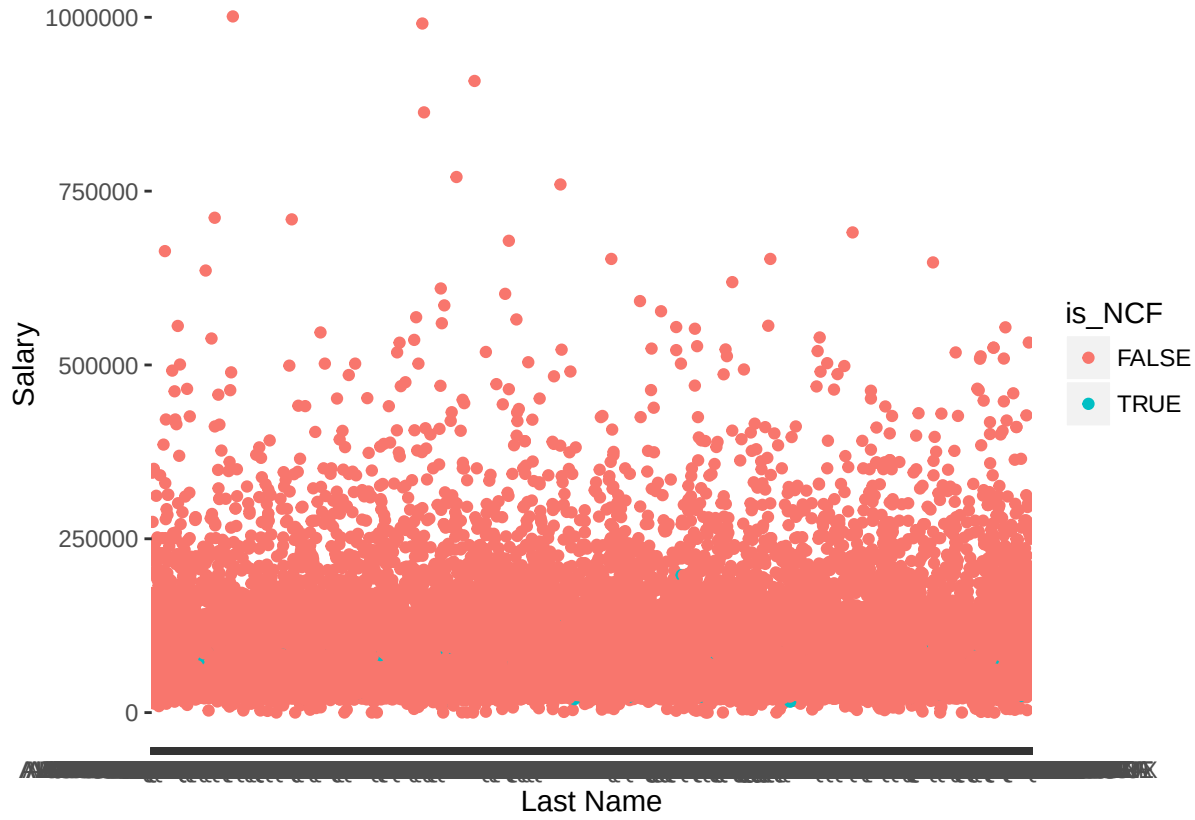
```
all.data[,Salary := (sum(`Annual Salary`)), by = list(`First Name`, `Last Name`, MI)]
all.data <- unique(all.data[, list(University, `Last Name`, `First Name`, MI, sum.fte, Salary, type)])
all.data[sum.fte != 0.00, fte.salary := Salary / sum.fte]
ggplot(data = all.data, aes(x= `Last Name`, y=Salary , ylab= "Salary", xlab= "Employees" )) + geom_poin
```



Splitting data by University

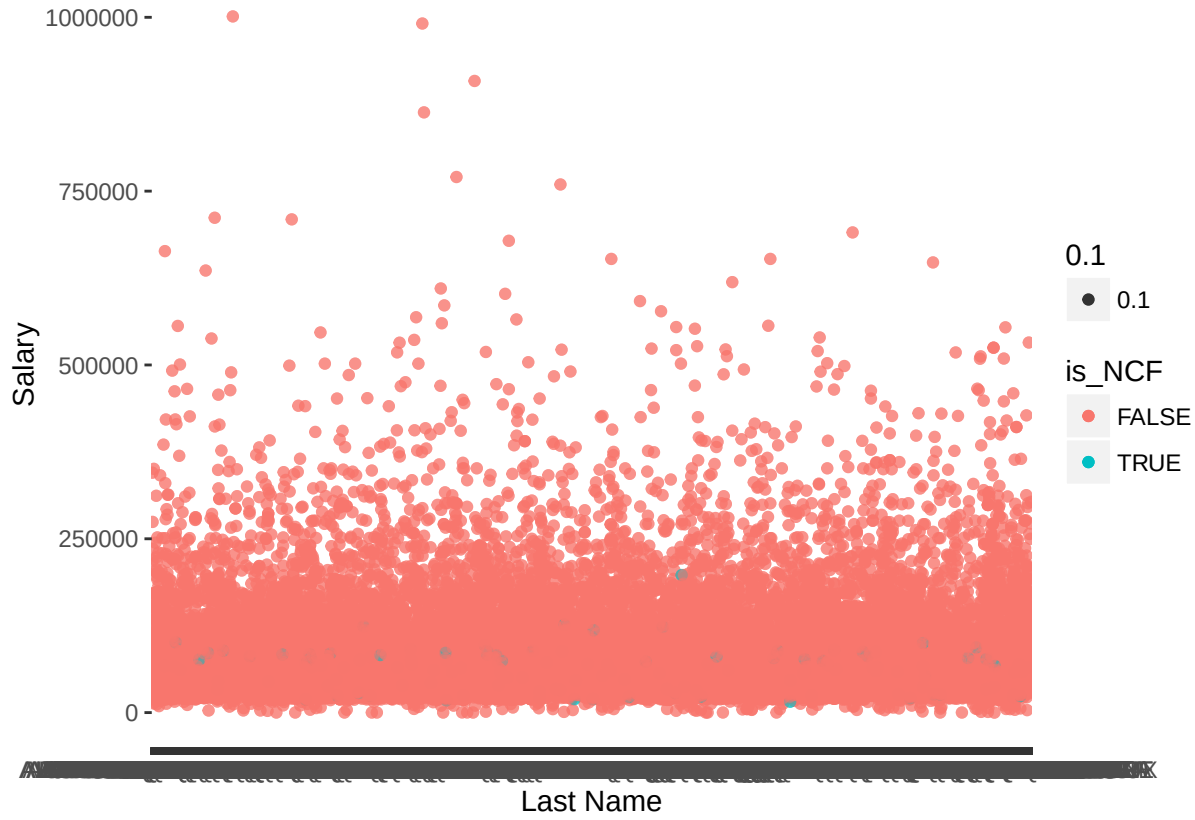
Now we want to organize the data to be comparable between New College and other Florida Universities, so we chunk the data into those classes.

```
all.data[, is_NCF := (University=='NCF')]
ggplot(data = all.data, aes(xlab = "Employees", x= `Last Name`, y=Salary ))  + geom_point(aes(colour= is
```



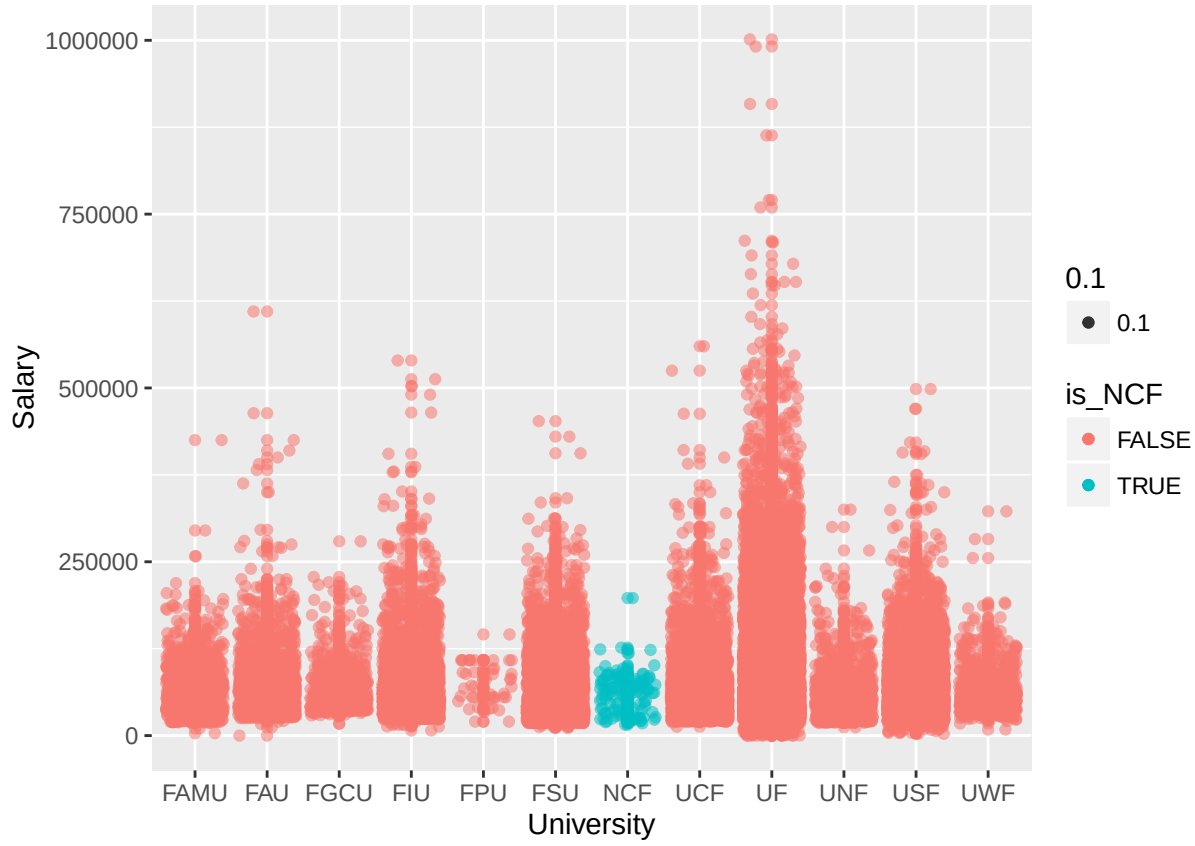
Although we have colored the points based on whether or not they are New College, it is still impossible to see the points. Perhaps if we add some opacity and jitter to the plot.

```
ggplot(data = all.data, aes(xlab = "Employees", x= `Last Name`, y=Salary )) + geom_point(aes(colour=is_
```



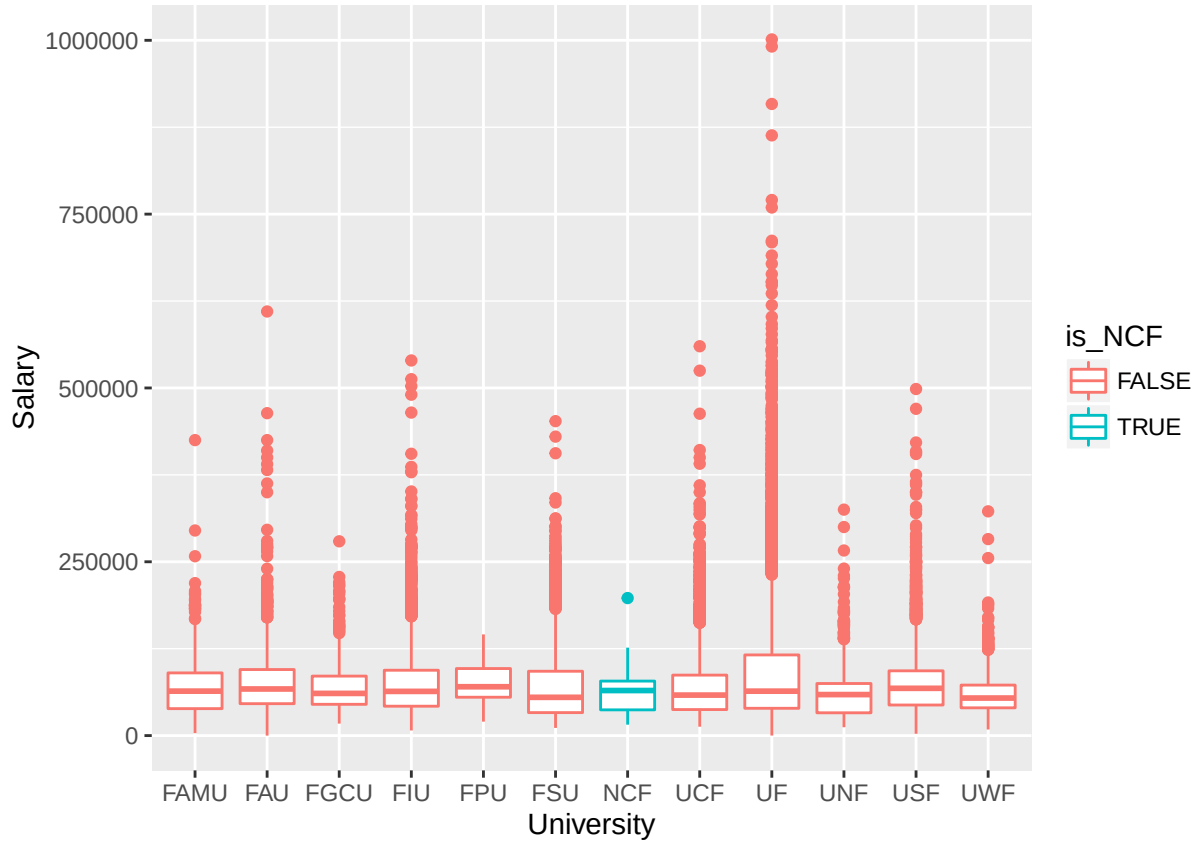
Opacity and jitter do not help us see the overall distribution compared to NCF very well, so lets try to split it by university.

```
ggplot(data = all.data, aes(xlab = "University", x= University, y=Salary )) + geom_point(aes(colour=is_
```



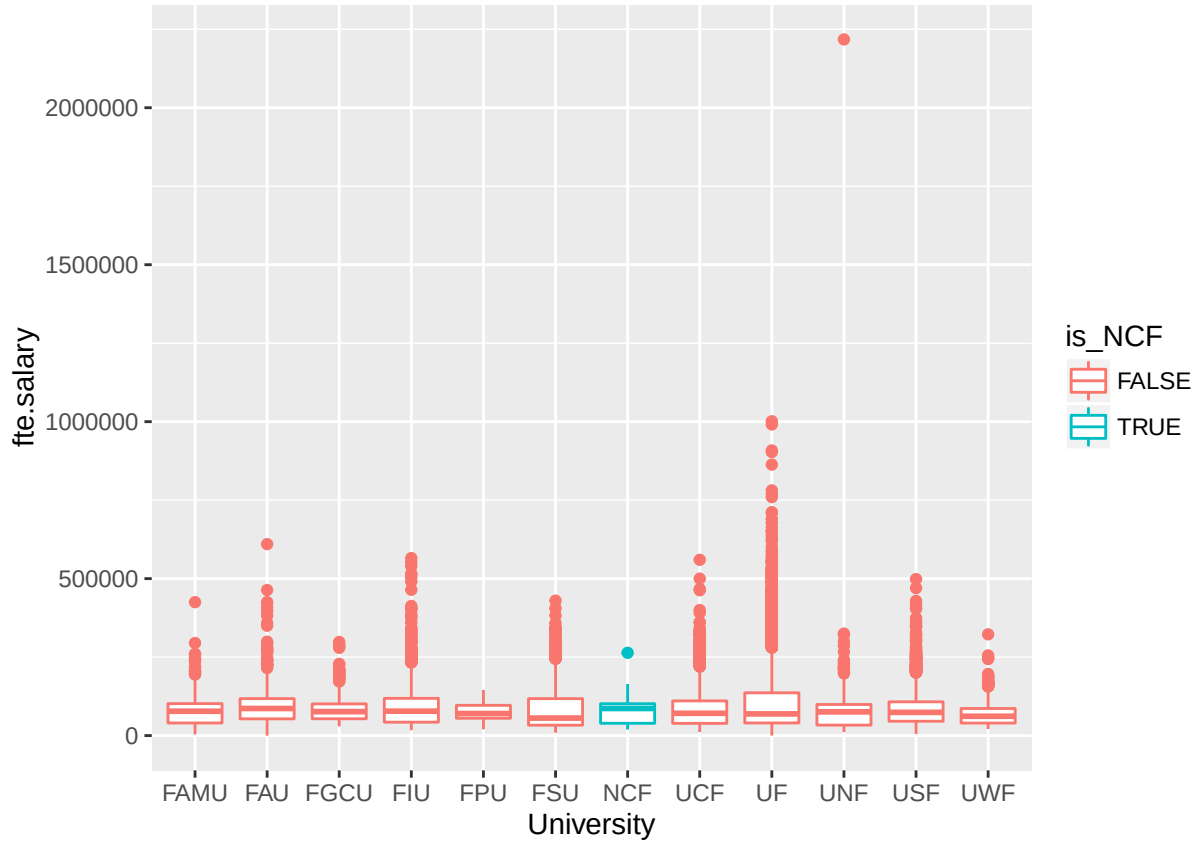
We can immediately see the drastic difference in range for New College's salaries compared to other universities and get some idea of the maginitudes and distributions However, the overall shape of the distributions are still hard to see like this, so lets see it before and after as a boxplots

```
ggplot(data = all.data, aes( x= University, y=Salary)) + geom_boxplot(aes( ylab = "Salary (in USD $)",
```



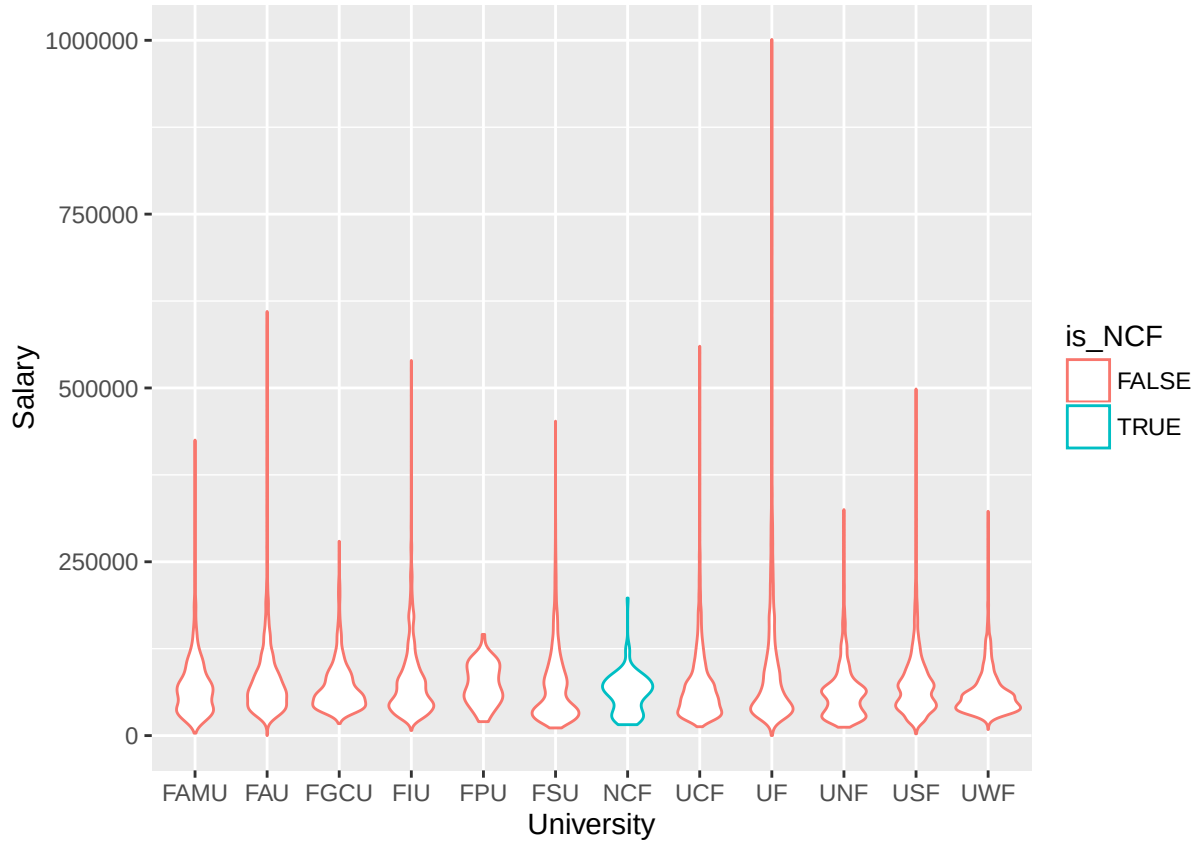

```
ggplot(data = all.data, aes( x= University, y=fte.salary)) + geom_boxplot(aes( ylab = "FTE adjusted Sala
```

```
## Warning: Removed 28 rows containing non-finite values (stat_boxplot).
```



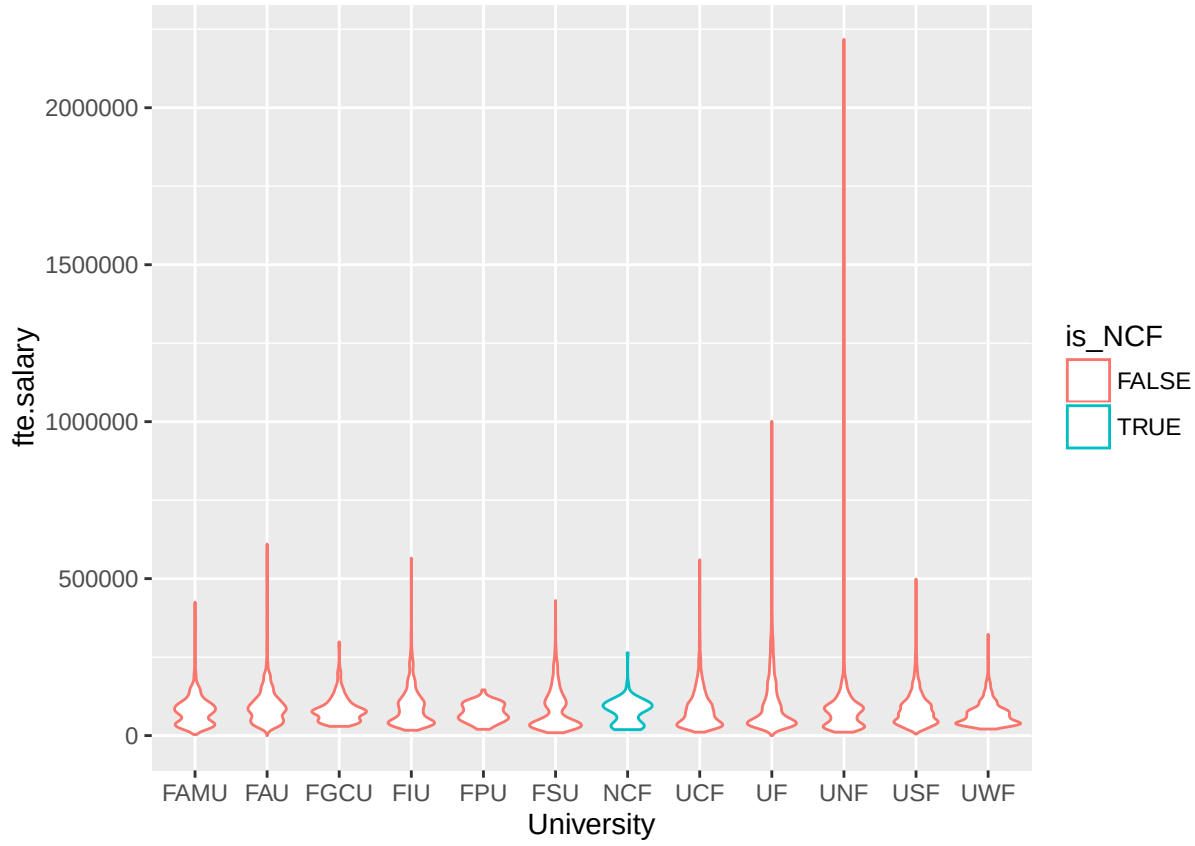
Salary adjusted for time (Salary / FTE) does not seem to show any major difference in distribution besides add more outliers. And y axis values become scaled down and compressed for the most part. Lets take another look, with violin plots this time.

```
ggplot(data = all.data, aes(xlab = "Universities", x= University, y=Salary, ylab = "Salary (in USD $)")
```




```
ggplot(data = all.data, aes(  x= University, y=fte.salary)) + geom_violin(aes(xlab = "Universities", yla
```

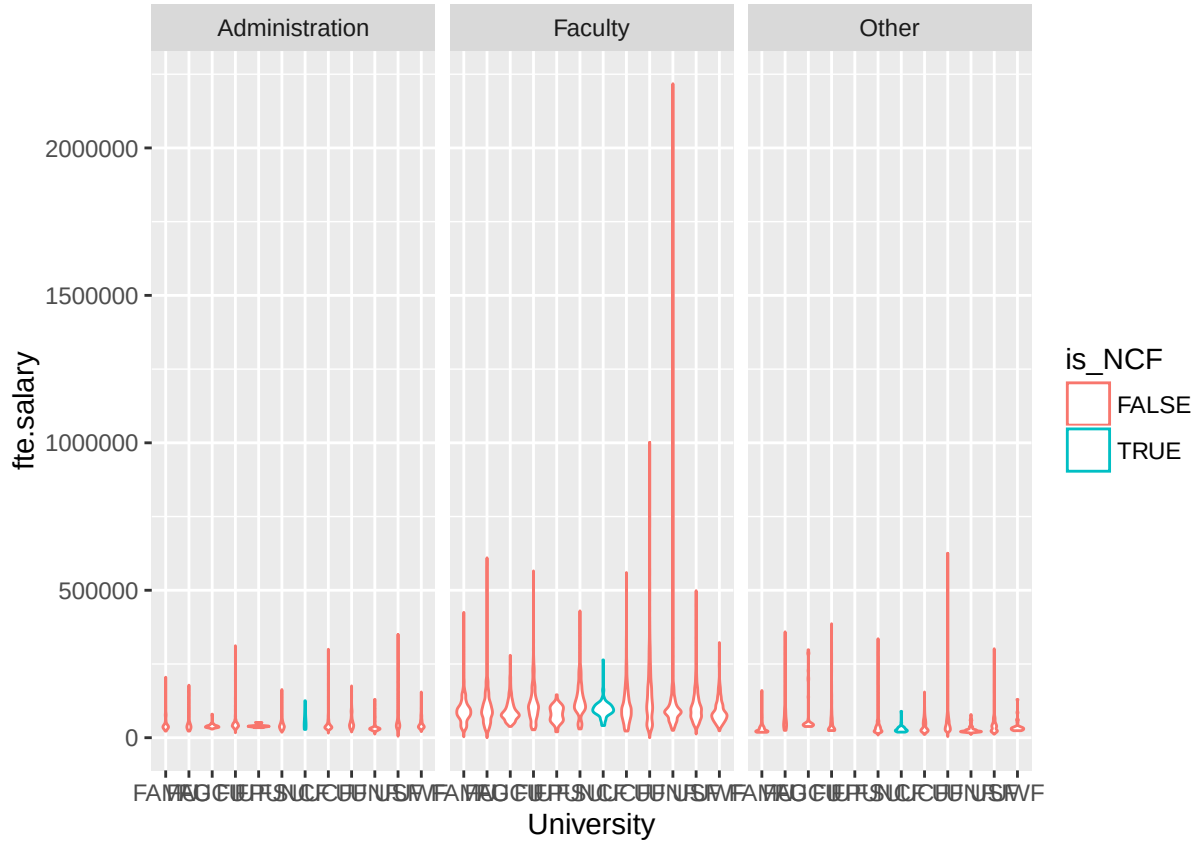
```
## Warning: Removed 28 rows containing non-finite values (stat_ydensity).
```



Now we wish to compare faculty and administration salaries, we can do this with facet_wrap() over the type coing we made.

```
ggplot(data = all.data, aes( x= University, y=fte.salary)) + geom_violin(aes(xlab = "Universities", ylab
```

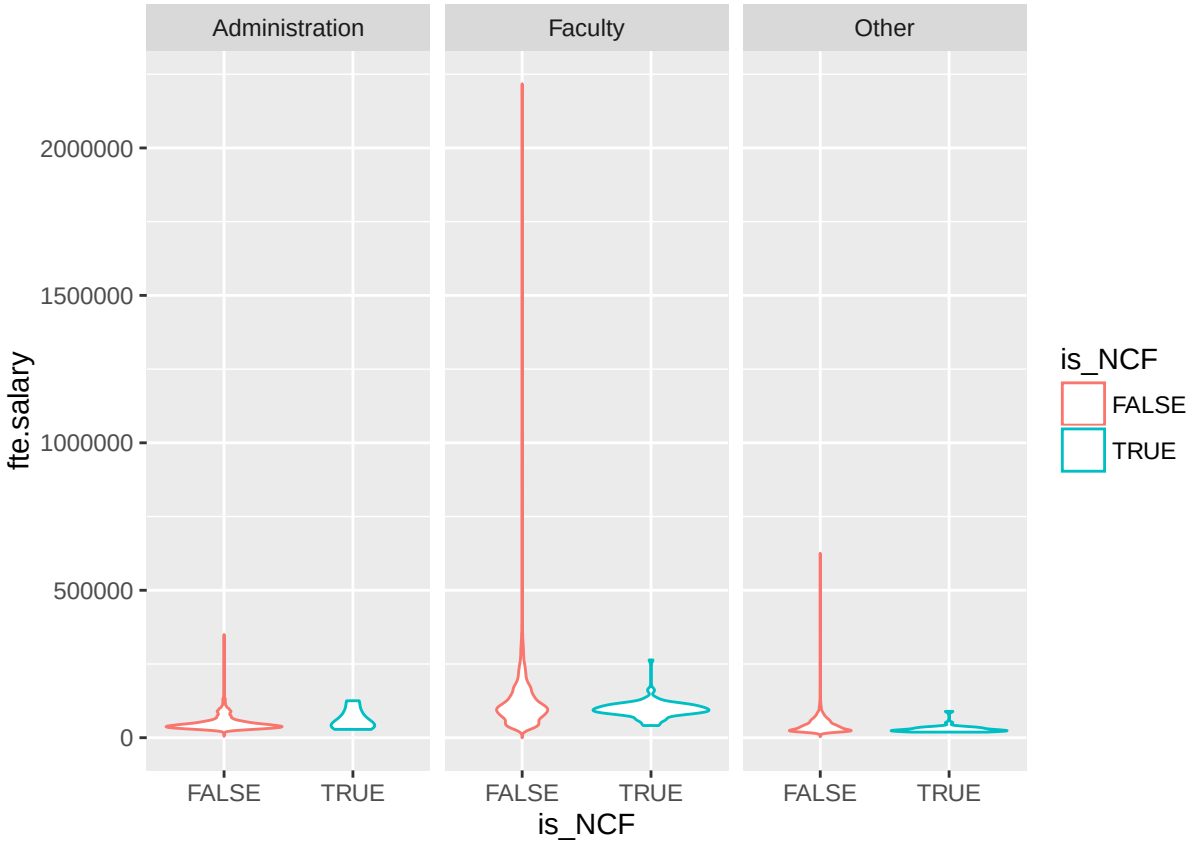
```
## Warning: Removed 28 rows containing non-finite values (stat_ydensity).
```



This is good, we can make out that the NCF distribution has a much lower variance than other universities, most of which have long tails and lots of upper outliers. It appears faculty usually make more than the administration for every school as well. Lets try to compare NCF against all others together as a single group.

```
ggplot(data = all.data, aes(xlab = "Universities", ylab = "FTE adjusted Salary (in USD $)", x= is_NCF, y
```

```
## Warning: Removed 28 rows containing non-finite values (stat_ydensity).
```


Administration
Faculty
Other
2000000
1500000
1000000
500000
0
fte.salary
FALSE
TRUE
FALSE
TRUE
FALSE
TRUE
is_NCF
is_NCF
FALSE
TRUE