Practice1

Dasha Asienga

Due by midnight, Wednesday, Sept. 14

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook(s), course materials in Moodle, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

- I used my map knowledge from STAT 231 to create the map.

Prompt

Explore the data set IPEDS (described below), and pick something you find interesting that you want to illustrate to others. In other words, pick something you find out about the dataset that you want to illustrate with a story, and make a few well-chosen graphics (minimum of 3, expected/reasonable between 3 and 6, but you could do more) to do that, along with appropriate text description/explanation to support your story. You must use at least four variables in some fashion in your story, but don't need to use them all at once. However, you cannot use just a series of 2-variable scatterplots from intro stats to satisfy the requirements for this assignment. Be more creative than that.

You can tell a story with visuals without applying any statistical techniques, though if your story involves regression, a t-test, or ANOVA, that's fine. It is not necessary to use anything other than graphics for this assignment. When describing "Methods" for this assignment, you should be discussing the visuals you are using, and the choices you are making when using them.

The data set (IPEDS) is a conglomeration of a few different data sets, all of which deal with evaluating colleges via self-reported values. This version is from 2012-2013 - though the data set is available in more updated forms if you are interested for other projects. There are 27 variables in the data set. Some are categorical and some are quantitative. The variables included in this data set are:

instnm: Institution (entity) name
city: City location of institution
state: State abbreviation
control: Control of institution
gradoffer: Graduate offering
longitud: Longitude location of institution
latitude: Latitude location of institution
retrate: Full-time retention rate, 2012
stufacr: Student-to-faculty ratio, 2012
ret_pcp: Part-time retention rate, 2012
admitp: Percentage of students admitted
admity: Percentage of admitted students who accept
tufeyr3: Tuition and fees, 2012-13
cost: Total price for out-of-state students living on campus 2012-13
obereg: Geographic region
sector: Sector of institution
enrtot: Total enrollment
efug: Undergraduate enrollment
uagrntp: Percent of undergraduate students receiving federal, state, local, institutional
uagrnta: Average amount of federal, state, local, institutional or other sources of grant
f1endmft: Endowment assets (year end) per FTE enrollment (GASB)
f2endmft: Endowment assets (year end) per FTE enrollment (FASB)
saltotl: Average salary equated to 9 months of full-time instructional staff - all ranks
salprof: Average salary equated to 9 months of full-time instructional staff - (full) professors
sanin01: Number of Full-time non-instructional staff
npgrn2: Average net price-students receiving grant or scholarship aid, 2011-12
npgrn3: Average net price-students receiving grant or scholarship aid, 2012-13

Due to the variable names, you will likely want to strongly consider using your own labels on graphics, or just relabel the variables using the rename command (create a new variable equal to the old one). Similarly, do not forget simple things like titles that can be useful for displays as well.

Note: If you have had Stat 231 or know how to make maps with R, you can include one map among your visuals.

Introduction

Purpose of the Analysis

The aim of this analysis is to explore the IPEDS data set, which is a conglomeration of a few different data sets, all of which deal with evaluating colleges via self-reported values. This data set is really interesting because it combines a lot of useful information on institutions across the US into one robust set and allows us to explore a lot of different interesting variables and uncover some useful insights.

The Story

The variable that interests me most from this data set is `salprof`, which stores values for the average salary equated to 9 months of full-time instructional staff - (full) professors. I'm interested in seeing whether there any variations nationwide, as well as whether there are any visible relationships between this variable and the other variables in the data set.

Variables of Interest

Given my main variable of interest, I will be focusing on `salprof` (professor salary). To craft the map, I will particularly be interested in the `state` variable, and perhaps consider breaking it down to the `city` level if there are any distinct patterns of note. However, I am also interested in exploring any existent patterns with some of the other quantitative variables that I think would be worthwhile investigating. These primarily include `retrate` (full-time retention rate), `stufacr` and (student-to-faculty ratio), `admitp` (percentage of students admitted), `tufeyr3` (tuition and fees), `uagranta` (average amount of federal, state, local, institutional or other sources of grant). I am interested in exploring how these break down across some of the categorical variables included in the data set, such as `control`, which defines whether the institution is public, private for-profit, or private not-for-profit.

Ultimately, the main aim of this analysis is to holistically explore the data set, primarily in looking at professor salary, and to tell a story with my visualizations.

Preliminary Analysis

```
IPEDS <- read.csv("https://awagaman.people.amherst.edu/stat240/IPEDS.csv")
```

Exploring the Data Set

These are the variables that I will begin with. Refer to the front page for a detailed explanation on the variables.

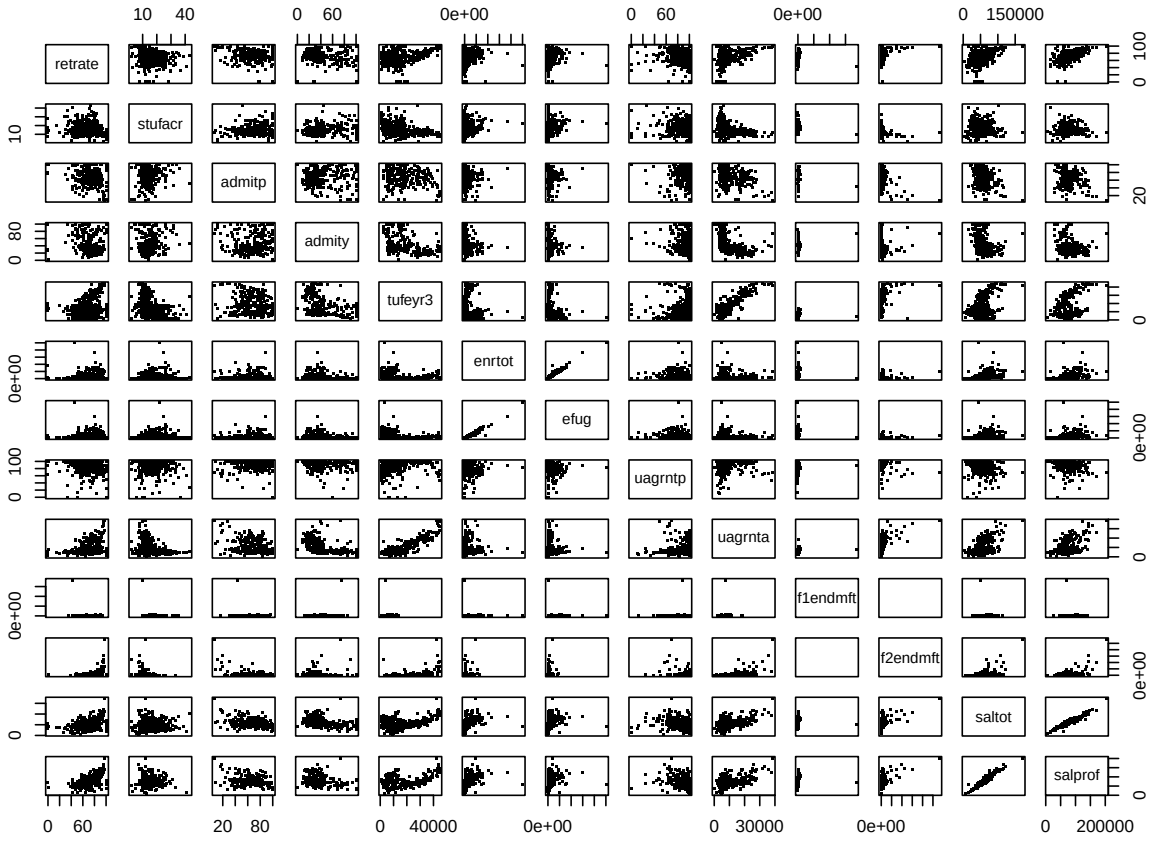
```
IPEDSsub <- IPEDS[ , c(3,4,6,7,8,9,10,11,12,14,16,17,19,20,21,22,23,24)]
names(IPEDSsub)
```

```
##  [1] "state"    "control"  "longitud" "latitude" "retrate"  "stufacr"
##  [7] "admitp"   "admity"   "tufeyr3"  "obereg"   "enrtot"   "efug"
## [13] "uagrntp"  "uagrnta"  "f1endmft" "f2endmft" "saltot"   "salprof"
```

Choosing Variables of Interest

In order to determine what potential relationships may be of interested, I began by looking at a high-level overview of the bivariate relationships between the quantitative variables in the data set. This would allow me to have a better understanding on what variables and relationships may be worth exploring and looking deeper into.

```
IPEDSsubsample <- sample(IPEDSsub, 500) %>%
  select(-orig.id)
pairs(IPEDSsubsample[, -c(1, 2, 3, 4, 10)], pch = ".", cex = 1.5)
```



In looking at the scatterplot matrix above, we can begin to see some key variables of interest that may have strong relationships with each other and that would be worth exploring further.

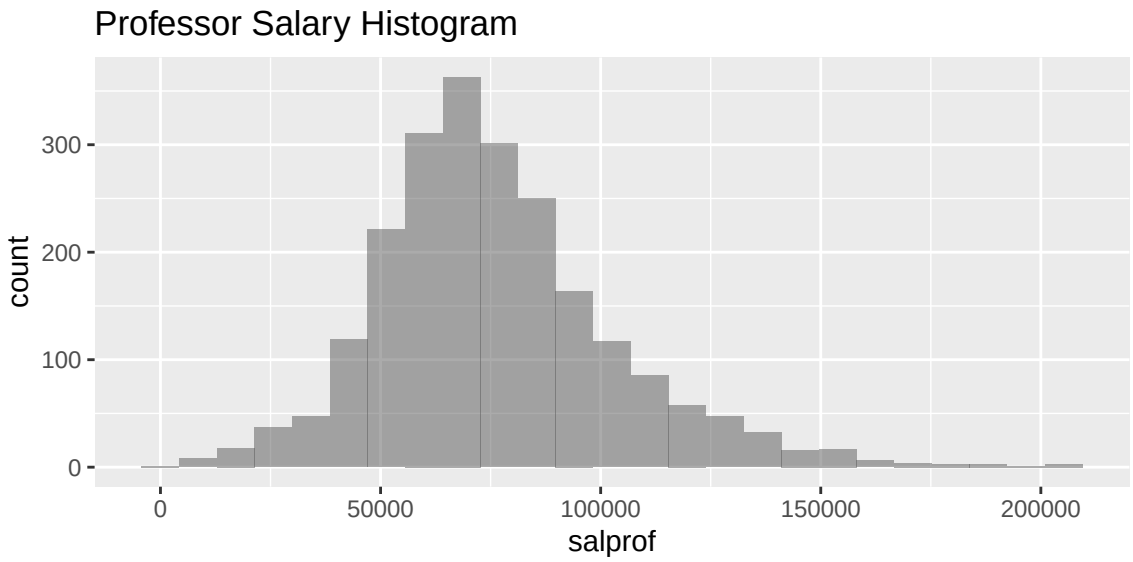
Some notable relationships include those between retention rate and tuition/ fees, retention rate and the average amount of grant received, retention rate and professor salary, student faculty ratio and professor salary, tuition/ fees and the average amount of grant, tuition/ fees and professor salary, professor salary and average amount of grant, to name a few. Particularly, we see that professor salary has some relationship with a lot of our variables of interest. We would expect `salprof` and `saltot` (average salary equated to 9 months of full-time instructional staff - all ranks) to have a very strong relationship, and professor salary data is included in `saltot`. Therefore, I will not consider `saltot` and mostly focus on full professor salary.

Based on this, I am interested in proceeding with the following variables and performing univariate, bivariate, and multivariate analyses to uncover any useful insights: `state`, `control`, `longitud`, `latitude`, `obereg`, and primarily, `retrate`, `stufacr`, `admitp`, `tufey3`, `enrtot`, `uagranta`, `f2endmft` and `salprof`. I would be interested in exploring what types of relationships exist between these various variables as well as faceting along some of the categorical variables included.

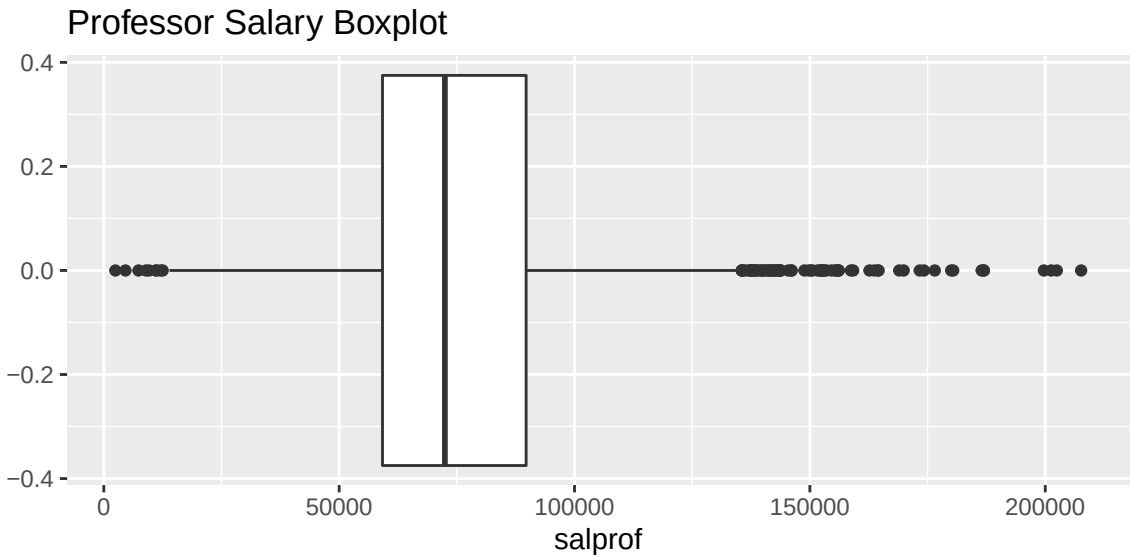
Univariate Analysis

The distribution of `salprof` was very interesting to me. We see in the visuals below a relatively normal distribution with a right skew. However, the boxplot is quite interesting because we see that there are outliers on both ends of the whiskers. Most schools pay their professors between 60000 and 90000. However, it appears that some pay them very well, and that there are some points that are on the extreme left. I wonder if those on the left are data entry errors or indicators of something worth noting.

```
gf_histogram(~ salprof, data = IPEDSdata) %>% gf_labs(title = "Professor Salary Histogram")
```




```
gf_boxplot(~ salprof, data = IPEDSdata) %>% gf_labs(title = "Professor Salary Boxplot")
```

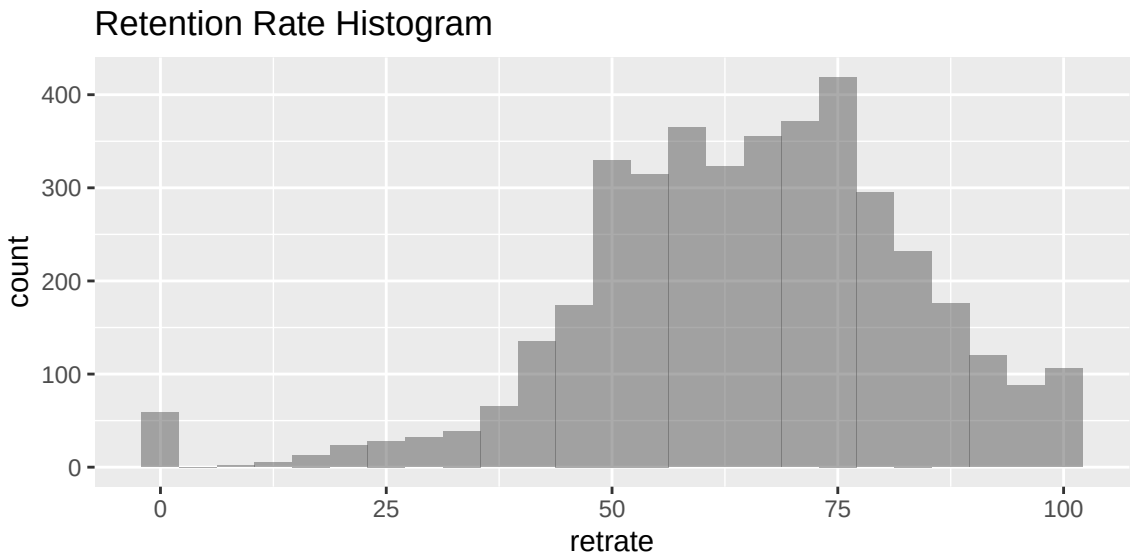



```
favstats(~ salprof, data = IPEDSdata)
```

```
##   min      Q1 median      Q3    max    mean      sd    n missing
##  2466 59206.5  72459 89716.5 207630 76198.8 27136.2 2232    2179
```

In looking at the distribution of some of the other variables, we also see that some schools self-reported values of 0 for `retention rate` and I am curious as to why that is the case. There is a notable left skew, suggesting that some schools report very low retention rate. While the focus on my analysis is `salprof`, this could be a secondary point of exploration for this, or a future, analysis.

```
gf_histogram(~ retrate, data = IPEDSdata) %>% gf_labs(title = "Retention Rate Histogram")
```

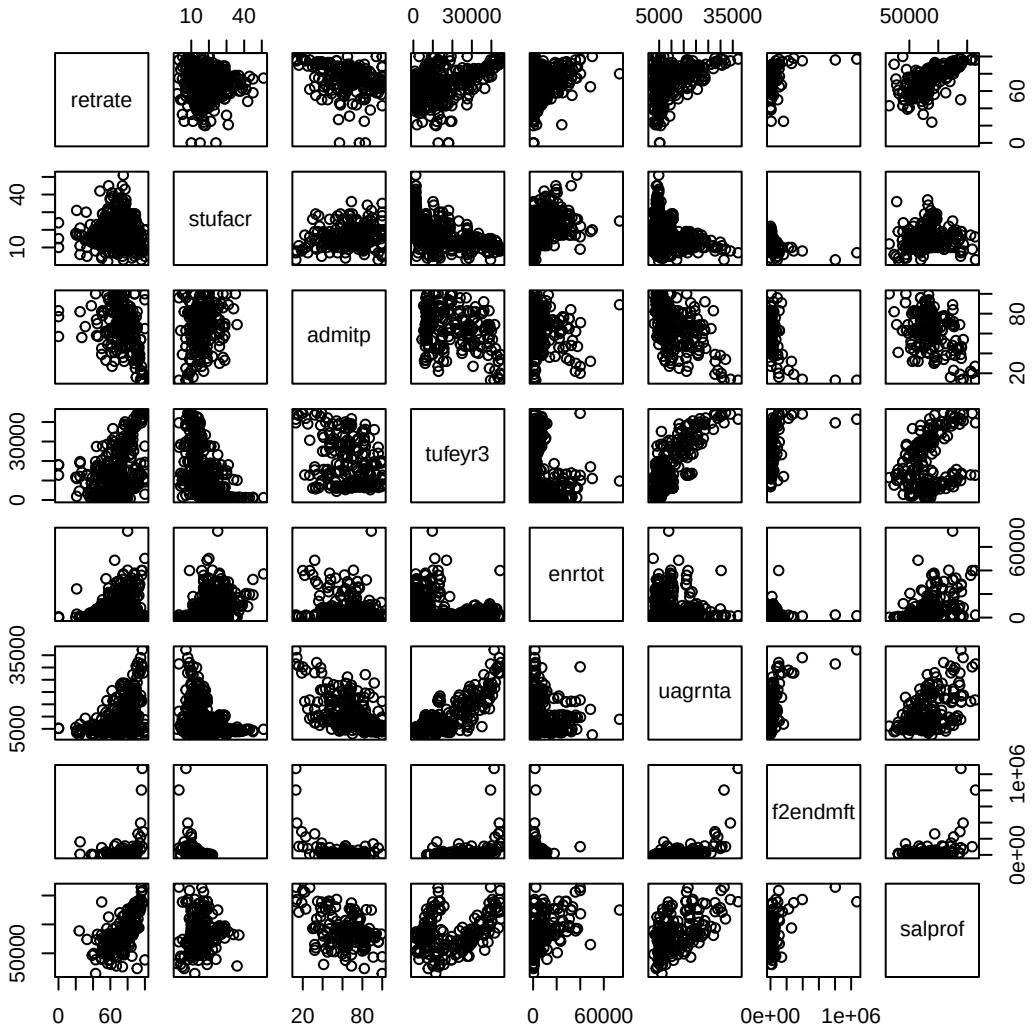


Based on my univariate analysis, there is an extreme right skew in the distributions for tuition/ fees, average grant received, total enrollment, student-faculty ratio, and endowment. This suggests that a few schools in the data set have very high tuition/ fees, receive substantially more grant, and also have substantially larger endowment. However, on the flip side, we see that some few schools (likely a different set of schools) have high enrollement and a high student-to-faculty ratio, suggesting that these schools have larger classes and perhaps less professors than ideal. We also see a slight left skew for the admitted percentage, suggesting that a few schools in the data set are highly competitive schools. I would be interested in exploring whether any of these skewed variables had relationships with each other.

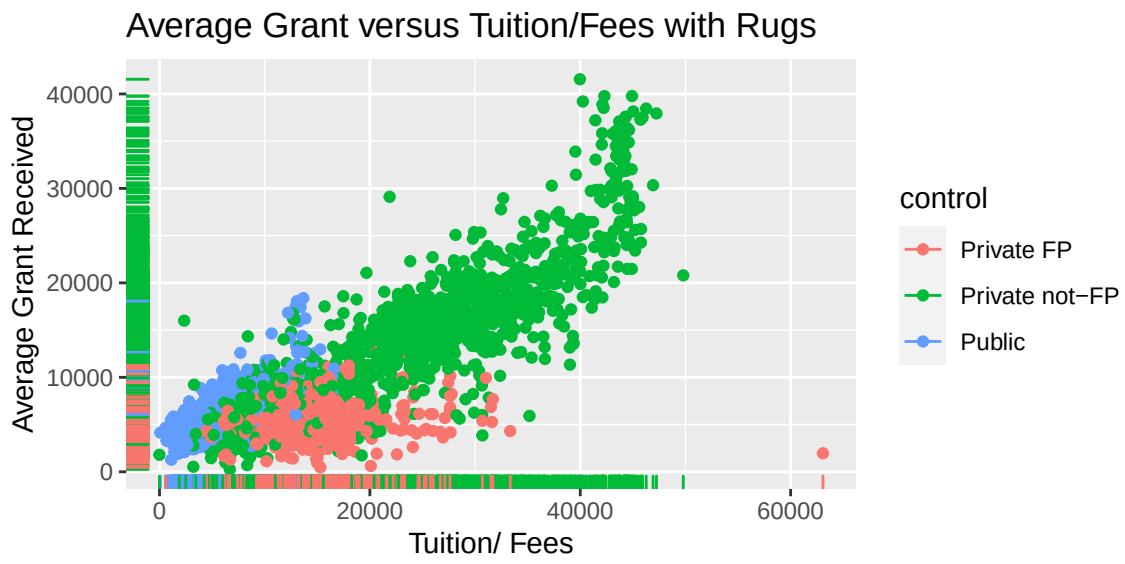
Bivariate Analysis

Looking closer at some of my final variables, we can begin to explore some bivariate relationships that will inform the multivariate analysis that we perform.

```
pairs(IPEDSdata[1:500, -c(1,2,3,4,9)])
```

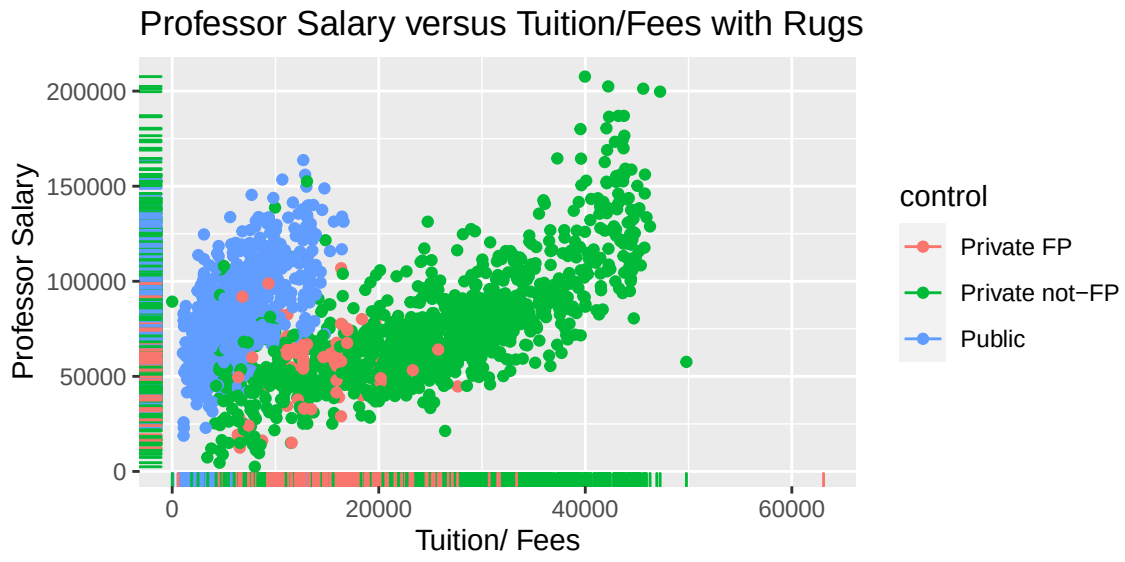



```
ggplot(IPEDSdata, aes(x = tufeyr3, y = uagrnta, color = control)) +
  geom_point() +
  geom_rug() +
  labs(title = "Average Grant versus Tuition/Fees with Rugs",
       x = "Tuition/ Fees",
       y = "Average Grant Received")
```



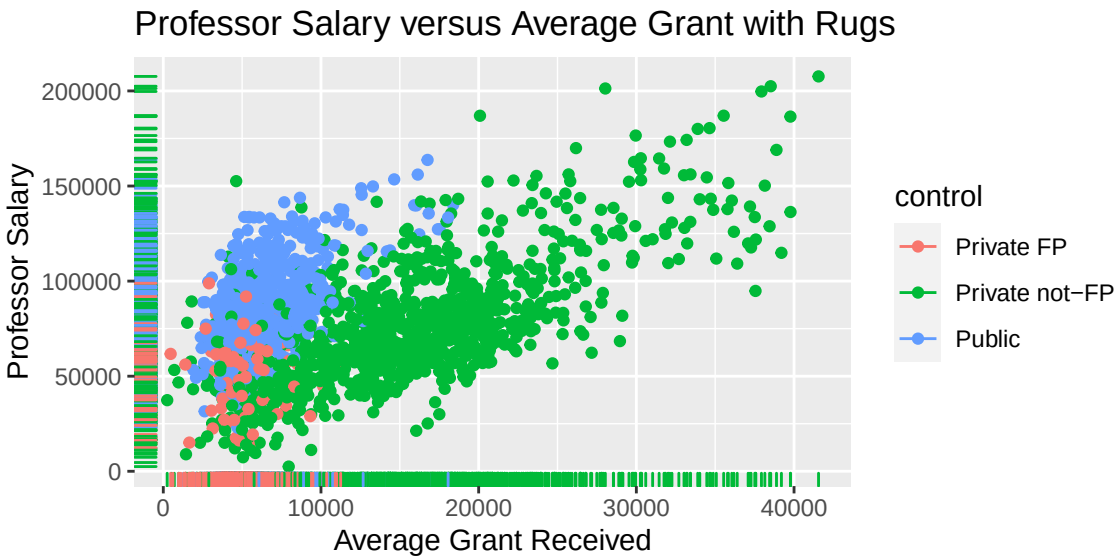
We see that tuition and average grant have a positive linear relationship. However, this is largely pronounced for private, not-FP institutions. Some of these institutions have higher tuition/ fees and higher average grant received than the other types of institutions.

```
ggplot(IPEDSdata, aes(x = tufeyr3, y = salprof, color = control)) +
  geom_point() +
  geom_rug() +
  labs(title = "Professor Salary versus Tuition/Fees with Rugs",
       x = "Tuition/ Fees",
       y = "Professor Salary")
```



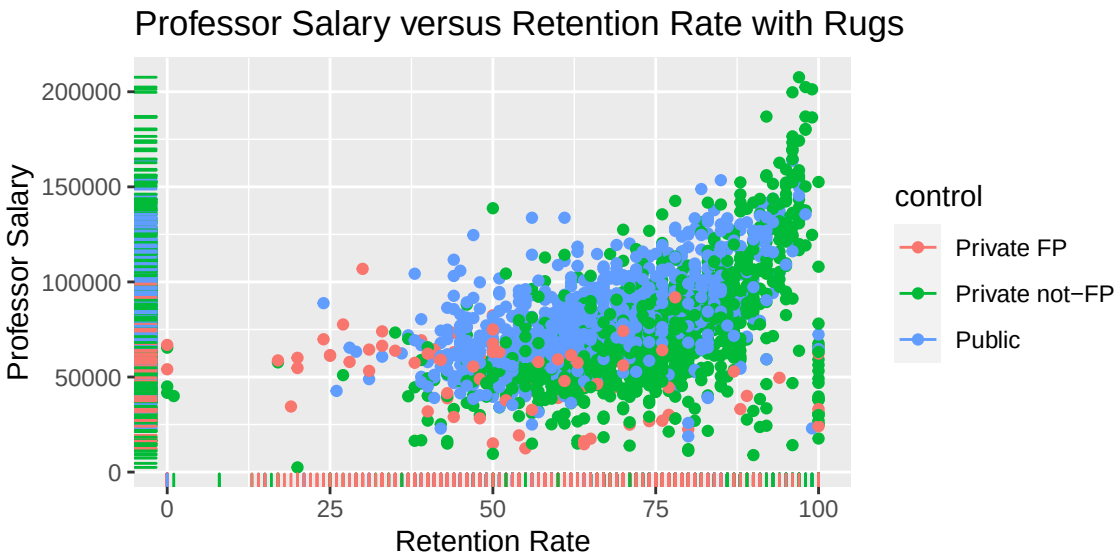
We see that professor salary increases with tuition/ fees. However, this increase is steeper for public institutions. Private not-FP institutions have a large range of professor salary. This will be important to consider in our multivatiate analysis. Given the strong relationships between tuition and average grant received, let's consider that next.

```
ggplot(IPEDSdata, aes(x = uagrnta, y = salprof, color = control)) +
  geom_point() +
  geom_rug() +
  labs(title = "Professor Salary versus Average Grant with Rugs",
       x = "Average Grant Received",
       y = "Professor Salary")
```



We see the same trend when we examine the relationship between average grant and professor salary. In my multivariate analysis, I am interested in looking at these 4 variables together: average grant received, tuition/ fees, professor salary, and control.

```
ggplot(IPEDSdata, aes(x = retrate, y = salprof, color = control)) +
  geom_point() +
  geom_rug() +
  labs(title = "Professor Salary versus Retention Rate with Rugs",
       x = "Retention Rate",
       y = "Professor Salary")
```



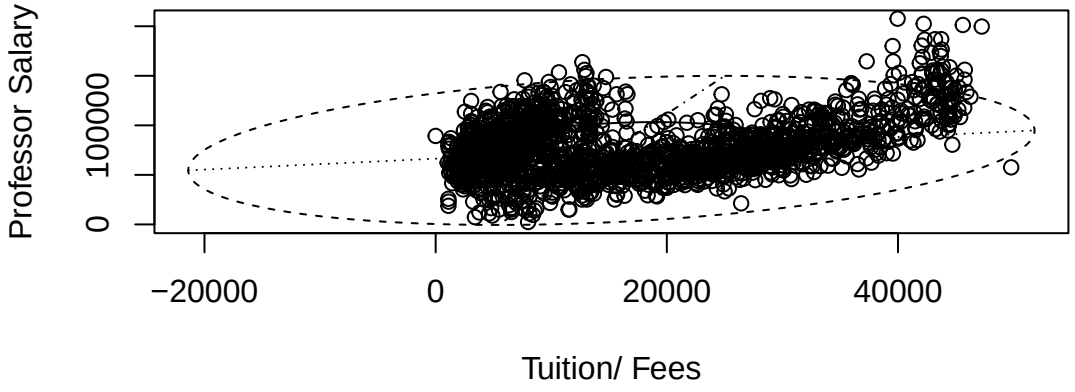
We also see that retention rate and professor salary have a positive, linear relationship. However, we don't see much of a difference when we differentiate by control.

From my univariate analysis and bivariate analysis, my key variables of interest are: professor salary, tuition/ fees, average grant received, control (type of institution), and retention rate. We see that control is a key categorical variable to consider.

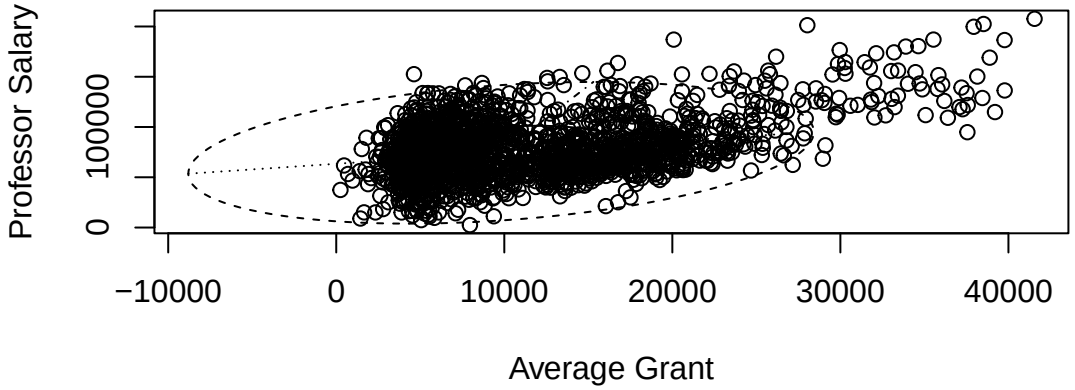
Outliers

```
IPEDSdata_bvbox <- IPEDSdata %>%
  filter(!is.na(tufeyr3),
         !is.na(uagrnta),
         !is.na(salprof))

twoVars <- IPEDSdata_bvbox[, c("tufeyr3", "salprof")]
bvbox(twoVars, mtitle = "Bivariate Boxplot of Tuition v Professor Salary",
      xlab = "Tuition/ Fees", ylab = "Professor Salary")
```




```
twoVars2 <- IPEDSdata_bvbox[, c("uagrnta", "salprof")]
bvbox(twoVars2, mtitle = "Bivariate Boxplot of Average Grant v Professor Salary",
      xlab = "Average Grant", ylab = "Professor Salary")
```



It seems that there are quite a few outliers to be careful about. Given that we used 2,187 observations to create this bivariate boxplot, I'm not too concerned about the outliers. However, it will be important to consider them in future for inferential analysis.

Methods

As mentioned above, the key variables involved in my story will be professor salary, tuition/ fees, average grant received, control (type of institution), state, and retention rate. My story centers around professor salary and I am curious to understand the relationships present with other variables, especially because of the presence of outliers on both whiskers of the univariate boxplot.

Map

I will begin by plotting a map that details the distribution of professor salary nationwide to see if state, city, or region can explain some of the variation. This will allow us to have a deeper look at how professor salary ranges nationwide and perhaps explore whether geographic region can explain it. I chose to use a continuous blue scale, with a lighter blue indicating higher salary and a darker blue indicating lower salary.

Bubble Plot

I will then plot a bubble plot with the 3 quantitative variables of interest: average grant received, tuition/ fees, and salary. My choice of a bubble plot is because all variables are quantitative, but also because it'll be easy to see if there is any relationships between these 3 variables. It's easier to differentiate circles of different radii than the length of the edges of a star.

3D Scatterplot

Next, I will use a 3D scatterplot to view these 3 quantitative variables. However, I will use symbols (cross, circles, and triangles) to differentiate between institutions of different control, that is, public institutions, private not-FP institutions, and private FP-institutions. We found in the preliminary analysis that `control` is a key categorical variable in our data set.

Trellis Graphic

If there is a visible difference across the 3 types of control, I will then use a trellis graphic to factor across the 3 types of control and view the 3 separate 3D scatterplots. This will allow us to see whether the relationship is consistent across all 3 controls or it is different for different types of institutions.

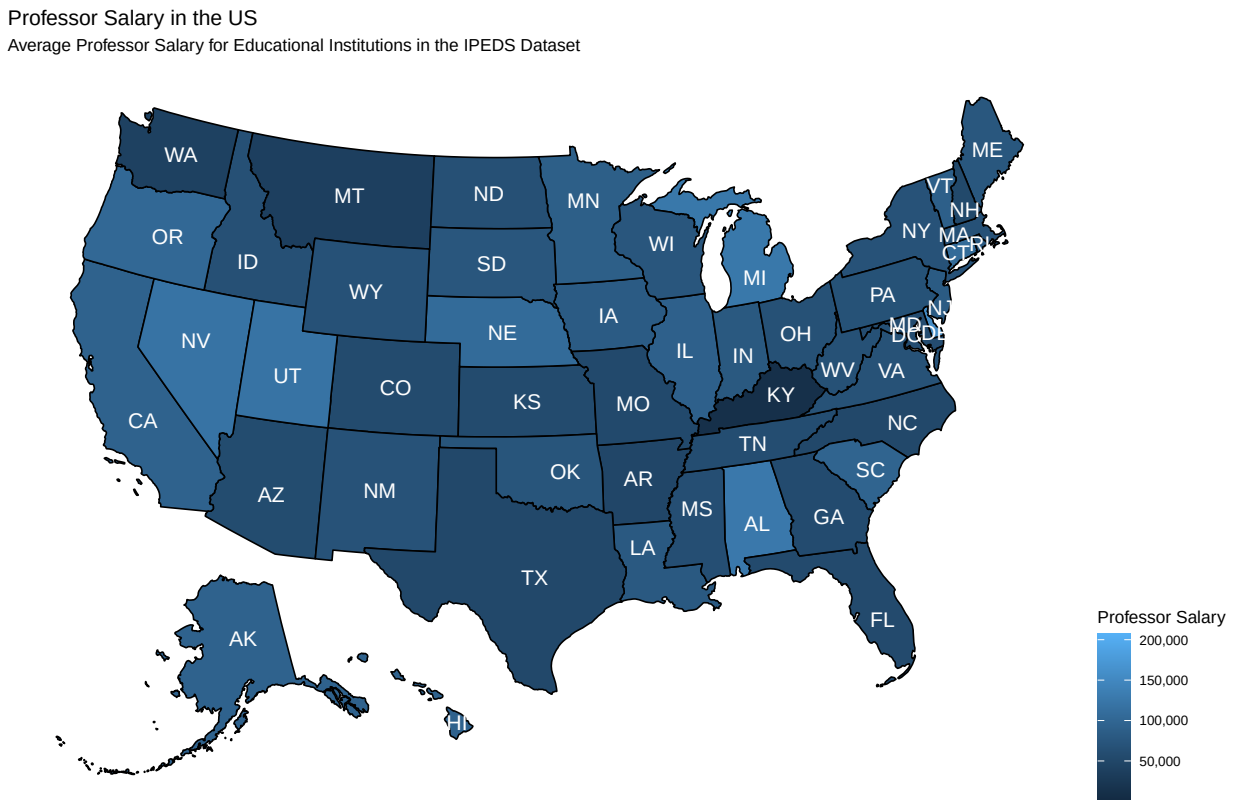
Parallel Coordinates Plot

Finally, I will use a parallel coordinates plot to further illustrate and conclude my story. The observations will be factored by control, and this will allow us to observe the relationship between the various variables for the observations on the data set.

Results / Story

```
IPEDSdata_salprof <- IPEDSdata %>%
  filter(!is.na(salprof))

plot_usmap(data = IPEDSdata_salprof, values = "salprof", regions = "state",
           labels = TRUE, label_color = "white") +
  scale_fill_continuous(name = "Professor Salary", label = scales::comma) +
  labs(title = "Professor Salary in the US",
       subtitle = "Average Professor Salary for Educational Institutions in the IPEDS Dataset") +
  theme(legend.position = "right")
```

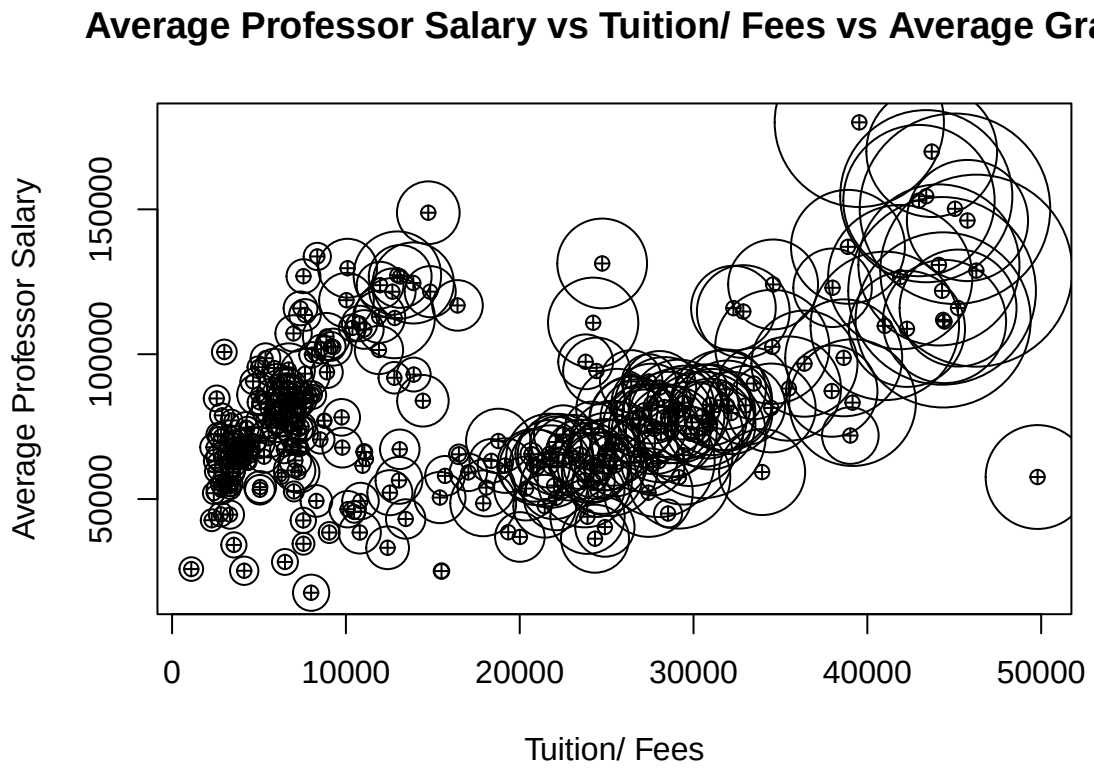


From the above map, we see that there are some nationwide differences in average professor salary. In particular, Kentucky has reportedly low values reported for professor salary. Other states on the lower end include Montana, Washington, and New Hampshire. On the other end, some states with reportedly high values reported for professor salary include Michigan, Atlanta, Nevada, Utah, and Delaware. This raises the question of what factors could be driving this difference. Let's take a look at some of our key variables in a bubble plot in order to evaluate whether some factors can help explain professor salaries.

```
set.seed(2022)

IPEDSdata_sub <- sample(IPEDSdata, 500) %>%
  select(-orig.id) %>%
  filter(!is.na(salprof))
```

```
ylim <- with(IPEDSdata_sub, range(salprof)) * c(0.95, 1)
plot(salprof ~ tufeyr3, data = IPEDSdata_sub, xlab = "Tuition/ Fees",
     ylab = "Average Professor Salary", pch = 10, ylim = ylim,
     main = "Average Professor Salary vs Tuition/ Fees vs Average Grant")
with(IPEDSdata_sub, symbols(tufeyr3, salprof, circles = uagrnta, inches = 0.5,
    add = TRUE))
```



The diameter of the bubble represents the average grant an institution received.

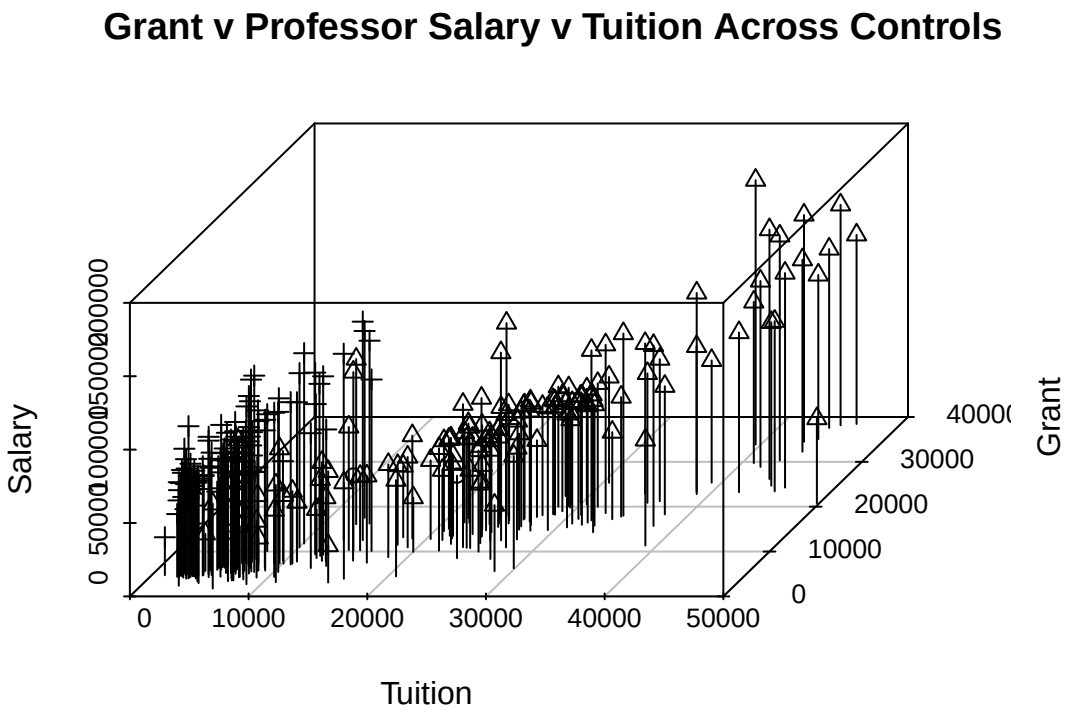
In general, we observe that there is a positive relationship between the 3 variables. Professor salary seems to increase as tuition/ fees increases, and the diameter of the bubble seems to get larger as well, and therefore, average grant increases as these 2 variables increase. This is important in thinking of professor salary distribution nationwide and whether such differences in tuition and average grant received are present as well.

However, there is a noticeable distinct pattern which begs the question of whether there is a difference that can be explained by one of our categorical variables. Let's look closer at how this may differ across the different institutional controls through a 3D scatterplot with symbols.

```
IPEDSdata_sub2 <- IPEDSdata_sub %>%
  rename("Salary" = salprof,
         "Grant" = uagrnta,
         "Tuition" = tufeyr3)

par(mfrow = c(1,1))
with(IPEDSdata_sub2, scatterplot3d(Tuition, Grant, Salary,
                                   pch = (1:3)[as.factor(control)], type = "h",
```

```
                                    angle = 55,
                                    main = "Grant v Professor Salary v Tuition Across Controls"))
```

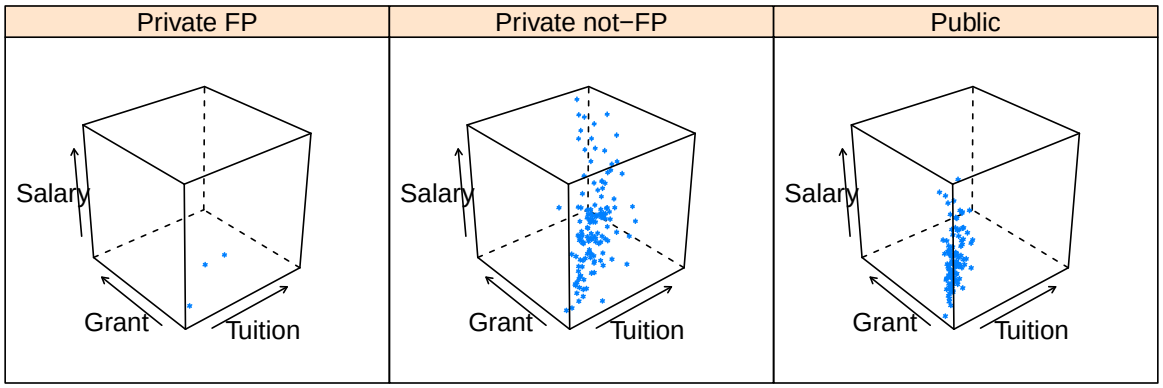


In this visual, a circle represents private-FP institutions, a triangle represents private not-FP institutions, and a cross represents public institutions.

We still see the same positive linear trend from the bubble plot. However, what becomes clearer is that a larger part of that relationship is determined by the control of the institution. We can see a lot of private not-FP institutions concentrated on the far right on the plot, suggesting that there are a few select private not-FP institutions that have really high average grant, tuition, and salary. However, these institutions fall on a large range. In general, we observe a positive trend with salary increasing as grant and tuition increases. We also see a lot of public institutions concentrated on the far left of the plot with lower tuition, salary, and grant, but still that positive relationship nevertheless.

Given this difference in control, I decided to factor the 3D scatterplots by control into a trellis graphics as follows. This will allow us to see this more distinctly.

```
plot(cloud(Salary ~ Tuition * Grant | control, panel.aspect = 0.9,
    data = IPEDSdata_sub2))
```

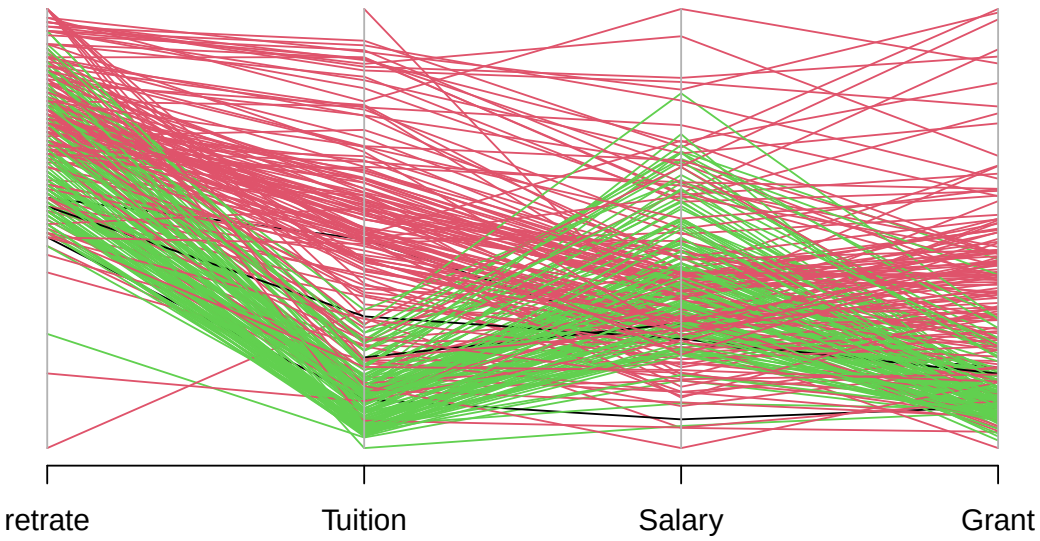


We now see that there aren't a lot of private for-profit institutions in our sample. For public institutions, we see that tuition increases as average grant increases, but it's hard to describe the behavior of professor salary. This is a key insight.

However, what we now see very clearly is that the key relationship we see between tuition, grant, and salary, is largely for the private not-FP institutions in out data set. Because these make up a lot of the institutions in our sample, they influence the general relationship we described earlier.

To illustrate this even further, let's examine a parallel coordinates plot for the observations in our sample and see how these multivariate relationships differ (or do not) across different institution controls.

```
MASS::parcoord(IPEDSdata_sub2[, c(5,8,13,11)], col = factor(IPEDSdata_sub2[ , 2]))
```



In this plot, black represents private FP institutions, green represents private not-FP institutions, and red represents public institutions.

The private not-FP institutions have a very strong relationship across the 4 variables defined. We see that institutions with higher retention rate have higher tuition, higher salaries, and higher average grants received.

This illustrates that the key results from this analysis explain the multivariate relationship between salary, grant, and tuition, primarily and strongest for the private not-FP institutions in our data set.

Conclusion

The aim of this analysis was to explore the IPEDS data set, looking closely at professor salary and investigating any multivariate relationships. I began with a preliminary analysis where I chose the variables of most interest to me and singled out what relationships I was most curious about and wanted to take a deeper look into. I ended up using a map, a bubble plot, a 3D scatter-plot, a trellis graphic, and a parallel coordinates plot to tell my story.

The distribution for professor salary had outliers on both ends, and the map revealed that some of those differences were visible nationwide when we looked into different states across the country. Some states seemed to have higher professor salaries than the median and others seemed to have lower professor salaries. In an aim to understand what some of the factors related to professor salary are and that could potentially be driving this difference, I constructed a bubble plot with 2 additional key variables: tuition/ fees and average grant. We detected a positive relationship between these 3 variables, but it seemed as though there was a categorical variable we needed to consider. Using symbols to represent institutional control and plotting the 3 variables in a 3D scatterplot, we begin to see that the key relationship we observed before is largely for the private not-FP institutions in our data set, and this is more apparent when we facet the 3D scatter-plots by control. The parallel coordinates plot of our sample helps us to see this more clearly and conclude our story.

Ultimately, my visuals depict a strong positive relationship between tuition/ fees, average grant, and professor salary, primarily for the private not-for-profit educational institutions. This allows us to conclude that these 3 variables are strongly related and further inferential analysis can allow us to conclude causal or correlational effects.