# Capstone Final Paper: The Price of Winning

*Grant Cox*

*January 22, 2019*

## The Problem

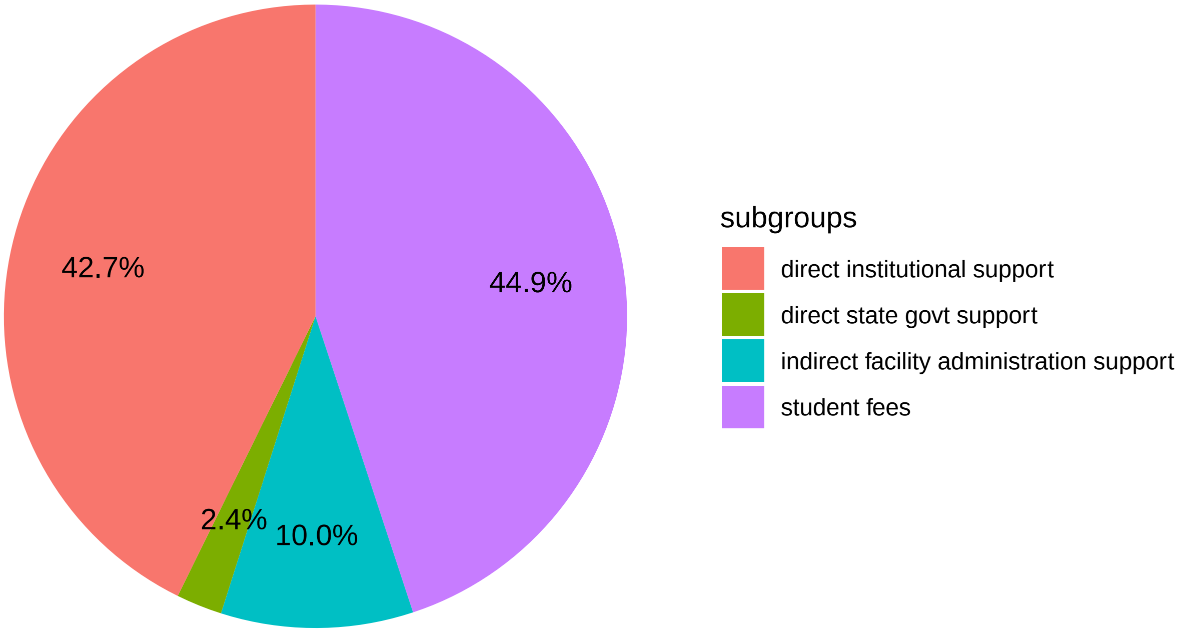
When Ohio State University squares off against it's biggest rival, the University of Michigan, "The Game" garners somewhere north of 10 million sets of eyeballs every year. When Alabama squared off against the University of Georgia this past season (2018-2019) in the SEC title game, ESPN reported an average of 28 million plus viewers across its networks. Notre Dame has secured an exclusive contract with NBC and reaps financial rewards of having one of the NCAA's largest fanbases year in and year out. As a result, everybody is getting paid. USA Today reported 2018 coaching salaries and it would take you scrolling down to number 83 before finding a head coach that isn't paid at least 1 million dollars annually (Turner Gill at Liberty university brings home a measly $950k per year). As far as the average fan can see, football is a cash cow. A university would be foolish not to dip into the war cheset that is collegiate football, right?

A deeper look into publicly available NCAA athletic revenue and expense forms shows us otherwise, though perhaps it isn't immediately apparent. The main reason it might seem like most athletic programs are finishing their years in the black is because of *subsidies*. Most public universities, it turns out, are heavily subsidizing their own athletic ventures in an effort to pay for the initiatives they **must** consider valuable in the long run.

---

We'll be looking intently at this concept of the subsidy, so it's important to see first how this number is derived. A public university's subsidy amount is a combination of the following contributions: student fees, direct institutional support, direct state government support, and indirect facilities and administrative support, all of which are variables detailed out in the README.

On average, subsidies were divided up as displayed by this pie chart. The "student fees" portion is of course coming from the students, while the "direct institutional support" portion is being pumped in directly from the university.



Universities are deceiving themselves in they continue to operate as if tens of millions of dollars isn't considered too much help. Not only do we see universitie losing millions and millions each year, but many are losing that on top of their own investments into their respective programs. What's worse, though, is these universities are deceiving other institutions that see the "success" and then consider it a viable financial move.

Over the last few weeks, I've dived into the data to show just how severe the bleeding is across the NCAA. Hopefully, this information paints a clear enough picture for universities on the fence.

---

## The Approach

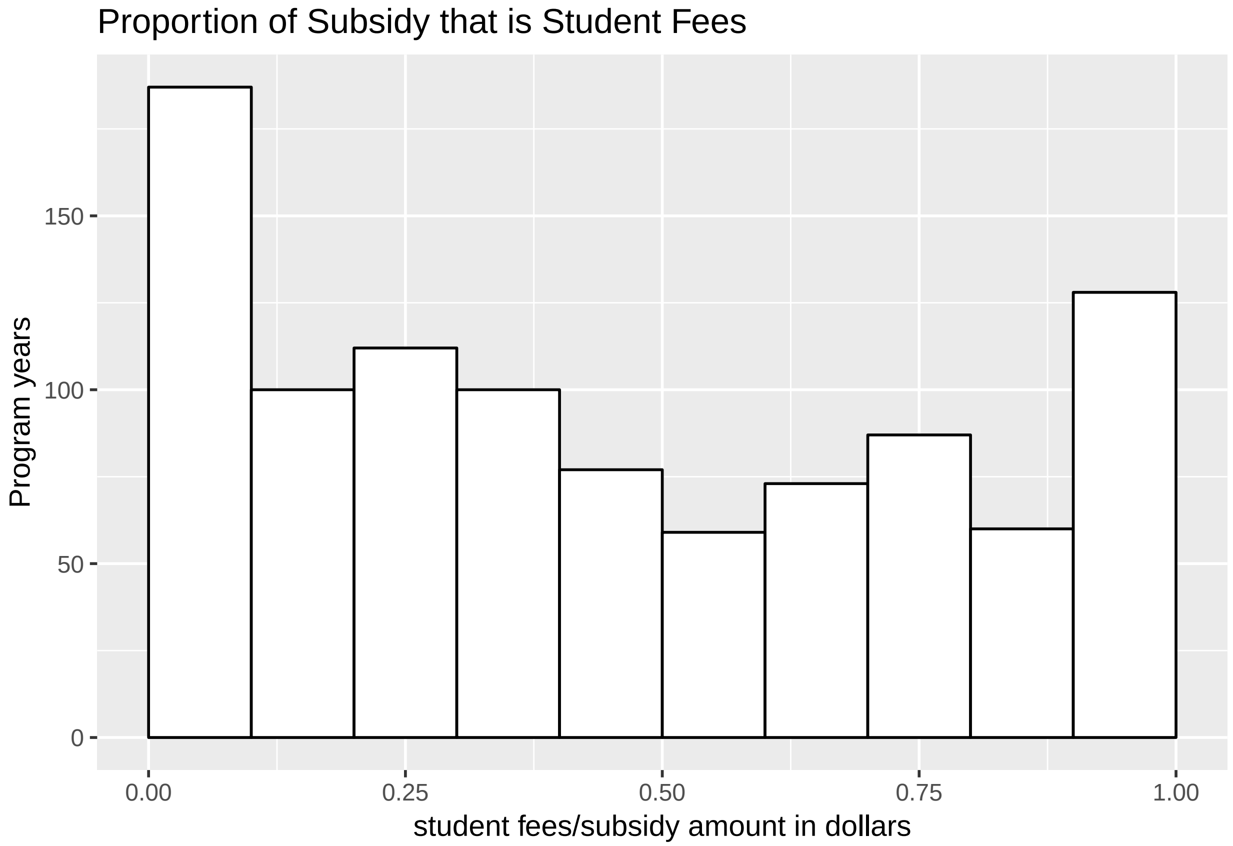
The dataset with win-loss records is pulled from 4 different year specific tables via teamrankings.com and the revenue and subsidy dataset via Huffington Post. The former set holds 5 tables (1 per year) and has roughly 130 records each. The latter dataset holds 1,015 records with 49 variables each. Both sets are in ".xlsx" formats.

This project began without a specific hypothesis apart from "how rampant is university spending in the face of loss?" or perhaps "how much are universities subsidizing their own programs despite losing money?" Because of this more general interest, rather than any particular question to answer, I approached this with a lot of plotting in mind. I wanted to see how these numbers related to each other. What would the data show us when we plotted expenses and revenues on the same graph? Revenue and subsidy? What about net revenue ordered by enrollment size? The plan was to let the data reveal itself through plots, and it certainly ended up showing us some fascinating trends.

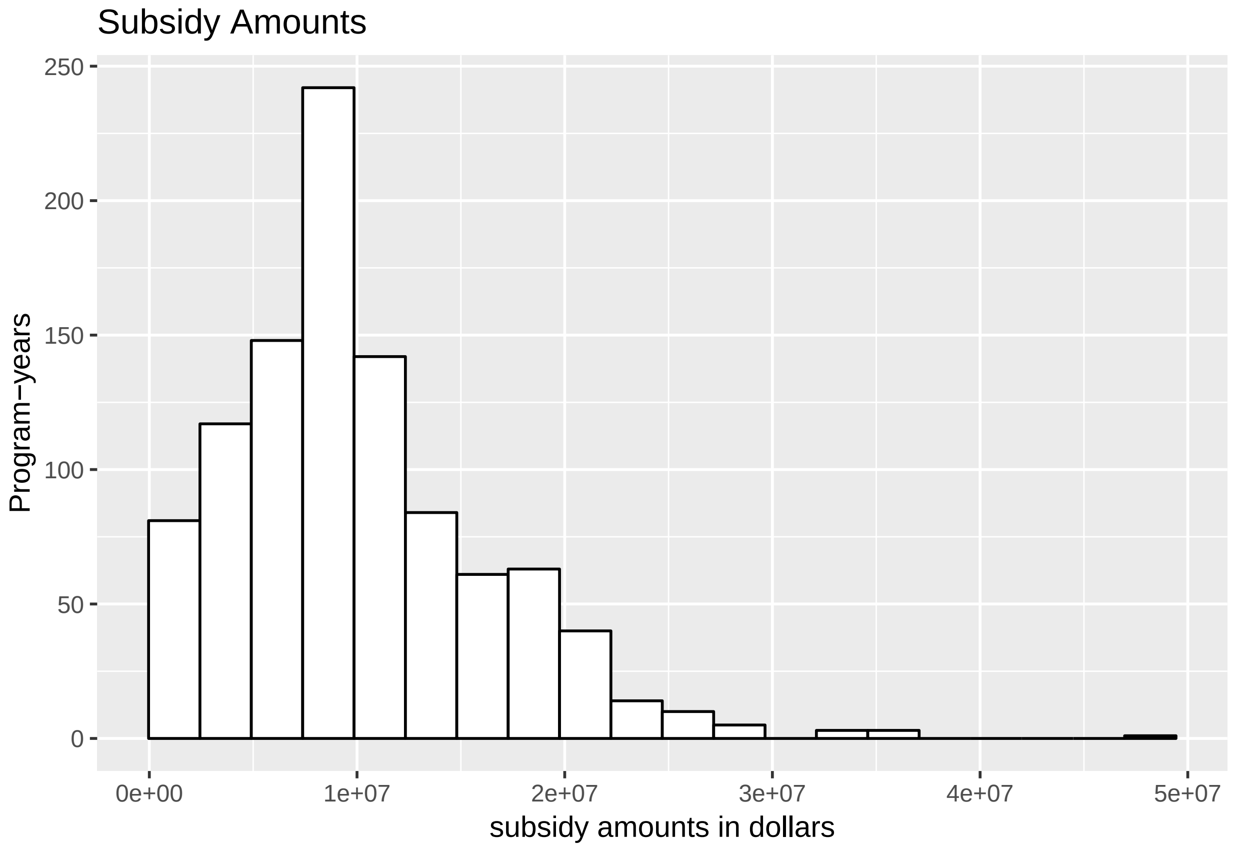
## The Findings

### Subsidies

While, on average, we see a 44% chunk of the subsidy be covered by student fees, the truth is it can vary greatly just how much a university by rely on the student fees to cover their subsidy. The histogram below shows that this proportion is hardly the most common arrangement.



Now, just because we see this proportion–plenty of universities accounting for nearly their whole subsidy by adding the fee to tuition–what sort of absolute numbers are we actually looking at? The histogram below displays where these public universities fall as far as what amounts of money they have deemed necessary to subsidize their athletic programs:



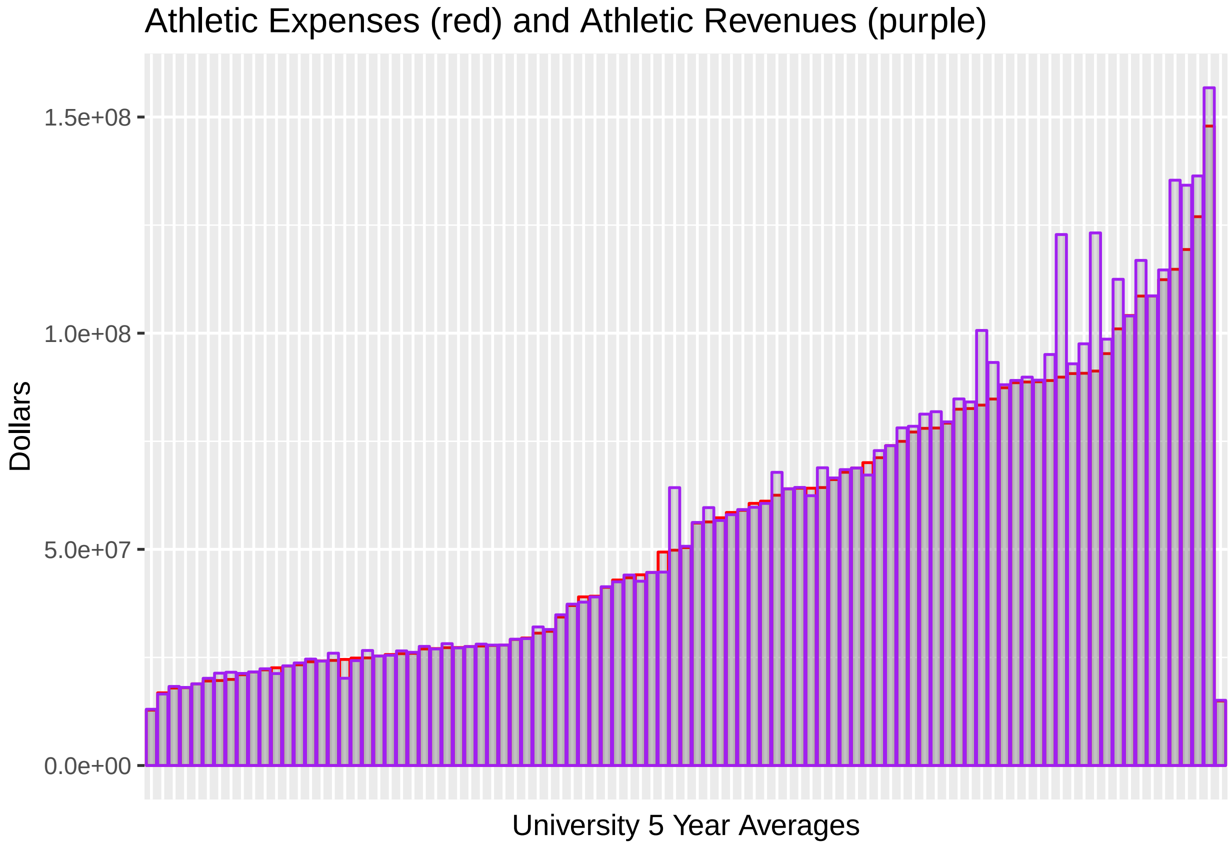
We see some serious density in the 5 million to 12.5 million range, with the greatest frequency in the 7.5 to 10 million dollar range.

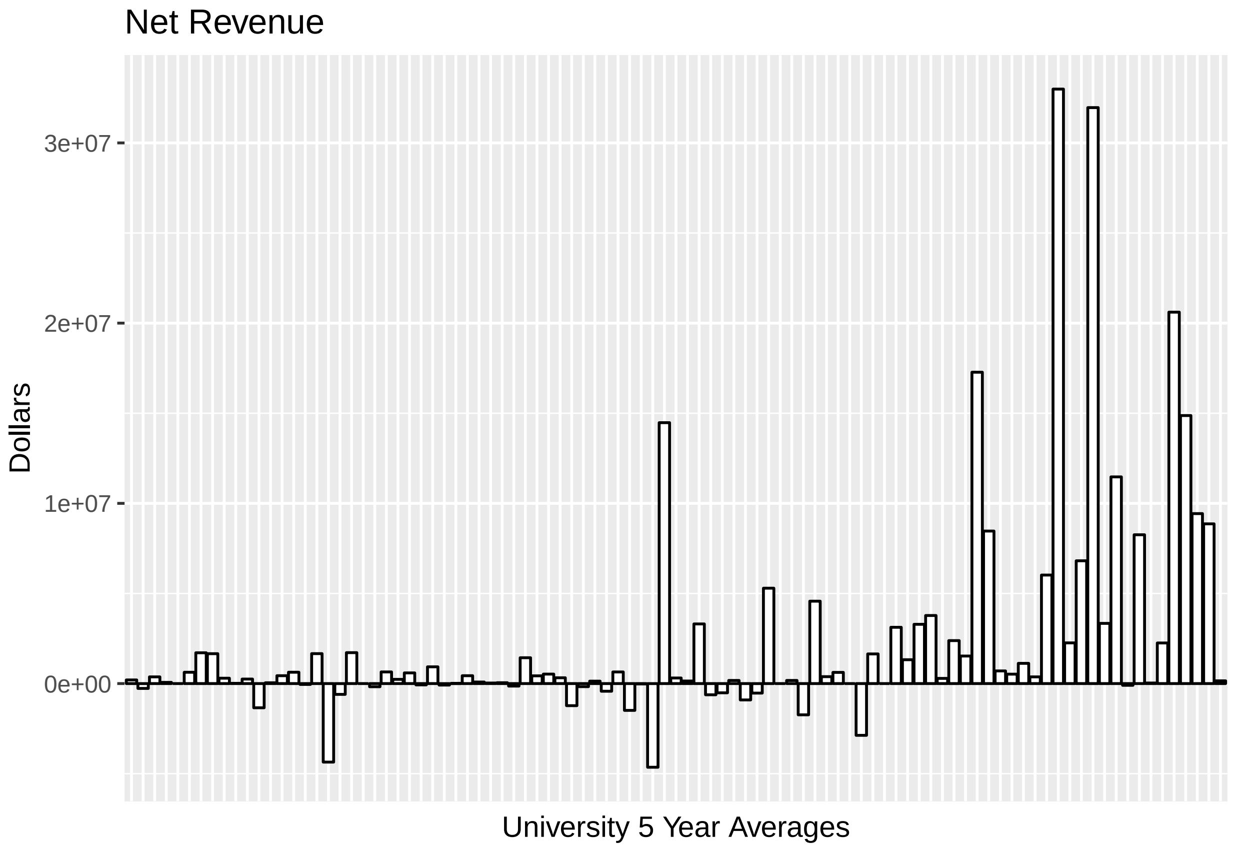
Now that we see how the schools are subsidizing their athletic programs, let's ask the bigger question: how important is this to their bottom line?

**Expenses and Revenues**

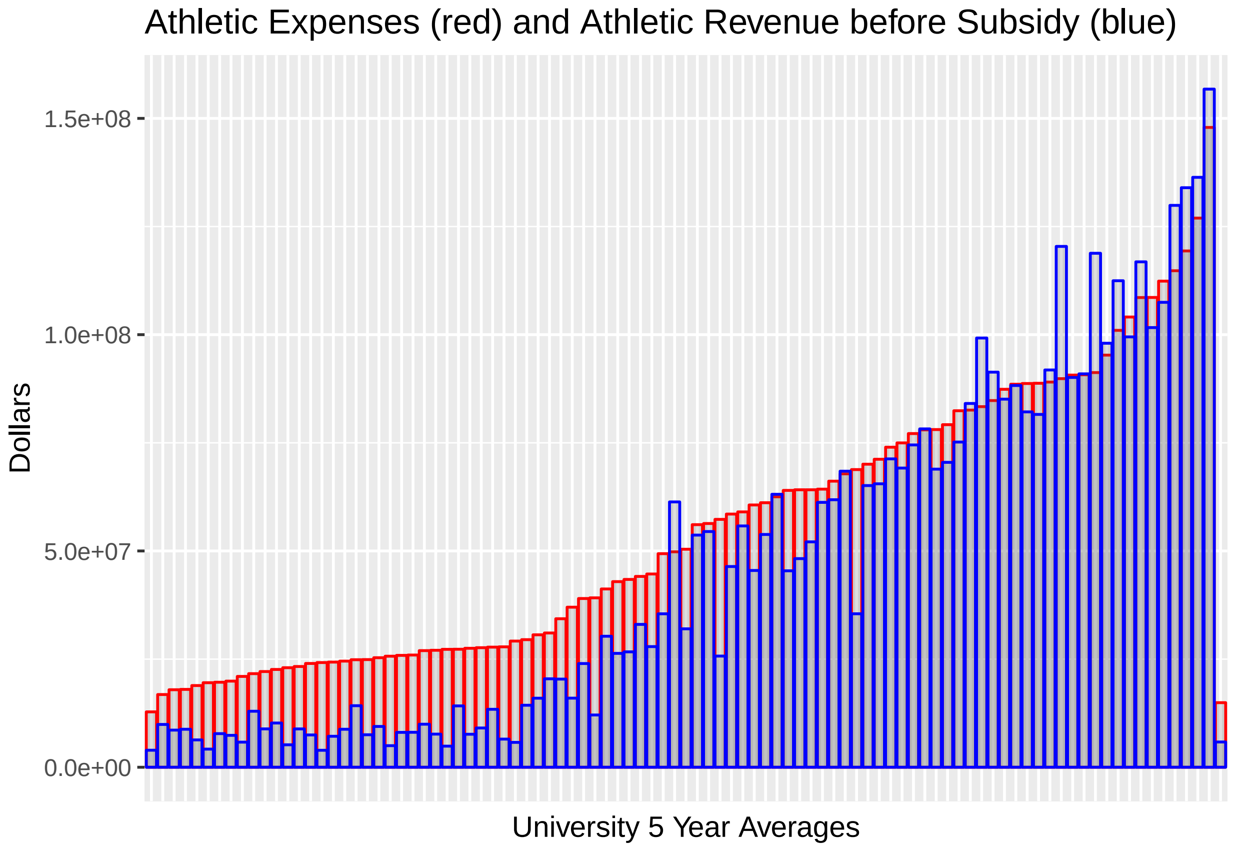
Now what these universities ultimately care about (before they have and during their time with a football program) is the money they make or can/will make. Graphing a few iterations of athletic revenue against athletic expenses ought to give us at least an idea of what these school might expect based on their expense level. First we see a simple Expense bar plot where I've taken 5 year averages from 130 universities.

From there we map Athletic Revenues directly on top and see that, in general, universities do an okay job of breaking even–if that is a worthy goal.

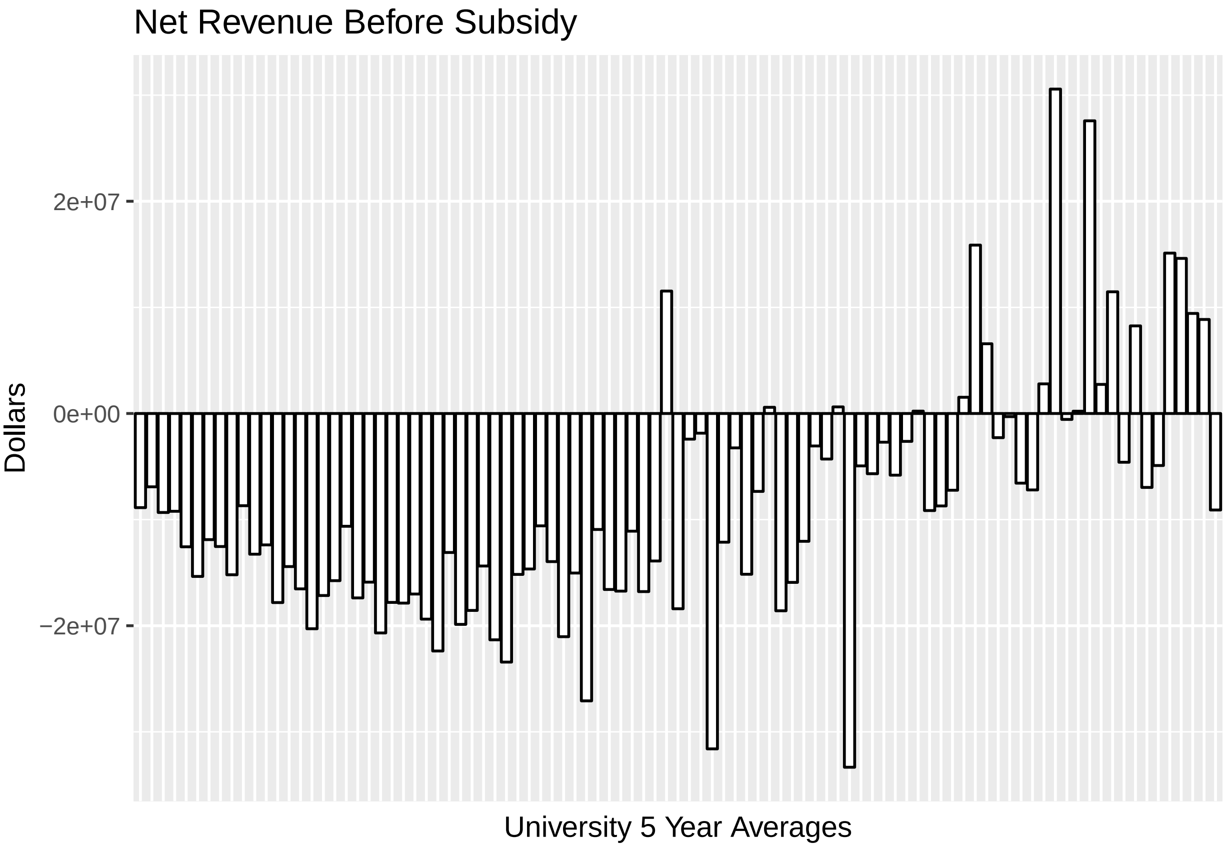




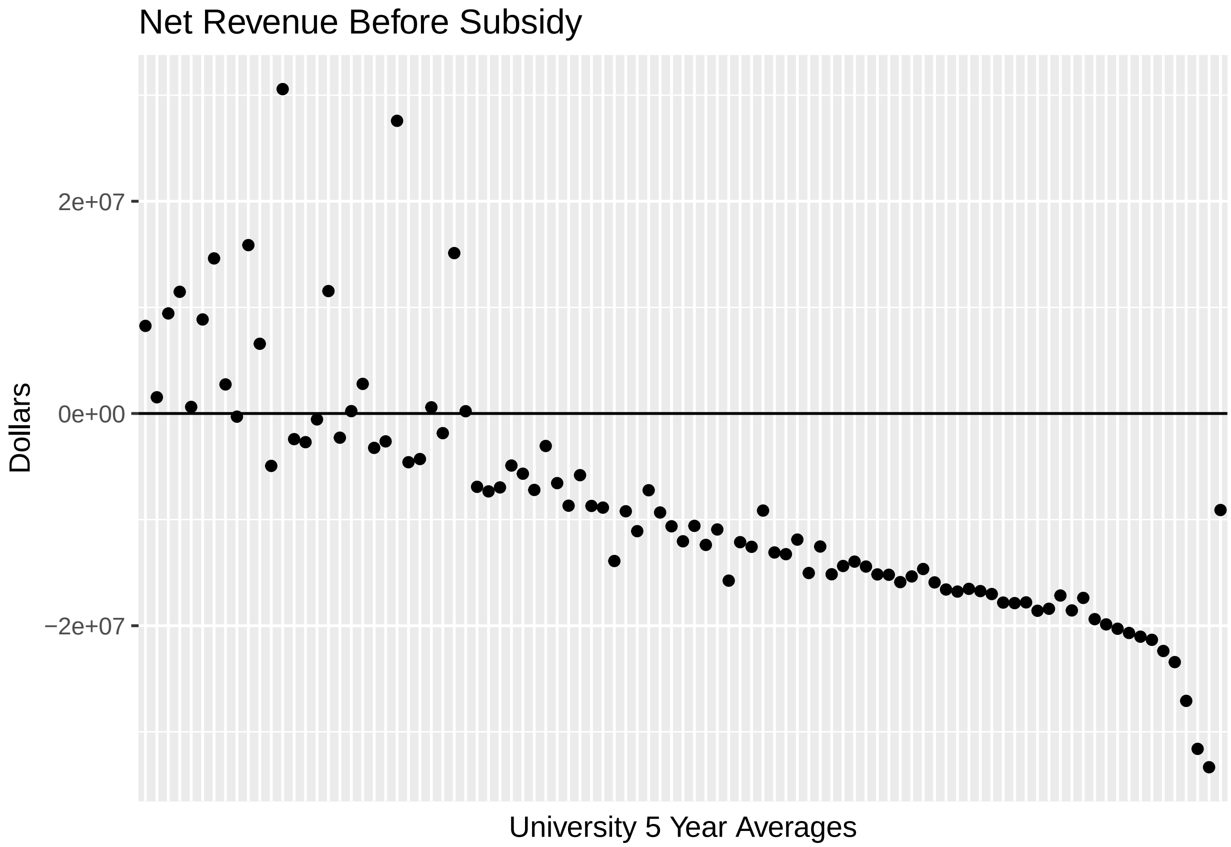
When we plot Athletic Revenues on top of this Expense chart, we see two things immediately: 1) the generals shapes of the plots are very similar, and 2) not every school is making money (at least, based on these plots). This plot above is titled “Net Revenue” and is a function of Athletic Revenues minus Athletic Expenses. Universites are still ordered on this plot ascending by their Expense amount.



Suppose we were to take out all the money these universities are pumping directly into their own programs. Ought we really consider that as a source of revenue anyway? That chart looks like this. You'll see here that the number of schools actually turning a profit are much fewer and farther in between. In fact, the subsidies in a lot of these cases has masked some fairly substantial losses.



This plot is, ironically, the “jackpot” moment of this deep dive. On average, we see that it was more likely than not that a university was losing and losing big each year. Universities, if they’re able to adequately deceive themselves, are (at least across these 5 years, 2010-2014), are losing crazy money, with few exceptions. When testing reordering these universities by different independent variables, I eventually landed on ordering the schools by amount pumped in from the school itself–by subsidy, that is. When we ordered these schools by subsidy and then plotted net revenue before subsidy (i.e. money in the school’s pocket), we saw a very interesting and very strong trend.



**As universities increased their own subsidy amounts, we saw a very clear negative relationship with that magic number, net revenue before subsidy.**

In other words, the more desperate a university is to "right the ship" the less money they were making. Naturally, such a nice scatter plot led nicely into my attempts to predict net revenue before subsidy. Those machine learning efforts are detailed next.

---

## Predictions using Maching Learning

### Checking for co-linearity

When beginning to examine how machine learning might play into this data set, I had a strong hunch that many variables would show strong co-linearity for two reasons: 1) many of them are functions of each other (e.g. "subsidy" is the sum of four other variables; "athletic revenues" is the sum of "subsidy" and "external revenue") and 2) the variable of school size (quantified by "full time enrollment") will intuitively correlate strongly with budget related variables (e.g. athletic expenses). To elaborate on the second point, we could intuit a few things about a larger school compared to a smaller school. Ticket sales would intuitively be higher, the expense budget would intuitively be greater, and so on. With these considerations in mind, selecting predictors for a regression wasn't as straightforward as it initially may have seemed.

```
cor(football$full_time_enrollment, football$athletic_expenses, use = "complete.obs")
fbsub <- football[c(7,13:33)]
cor(fbsub, use = "complete.obs") # for some universities, I only have win/loss data and not financial d
```

### Building Linear Models

```
model1 = lm(athletic_revenues ~ full_time_enrollment, data = football)
summary(model1)

model2 = lm(athletic_revenues ~ athletic_expenses, data = football) # Rsq is 0.96; co-linearity; spend 
summary(model2)

model2b = lm(athletic_revenues ~ athletic_expenses + Win_percentage, data = football)
summary(model2b) # Rsq actually decreases by adding win percentage

model3 = lm(net_revenue_before_subsidy ~ athletic_expenses, data = football) # Rsq is 0.12; very tellin
summary(model3)

model4 = lm(net_revenue_before_subsidy ~ Win_percentage, data = football)
summary(model4)

model5 = lm(net_revenue_before_subsidy ~ full_time_enrollment + Win_percentage + external_revenue,
            data = football)
summary(model5) #Rsq .5499 AdRsq .5469
```

The above code gives some of my early attempts at linear regression models. What I realized after the first few was that predicting “athletic revenues” was 1) easy, and 2) not helpful. Athletic revenues proved to essentially be a product of athletic expenses (i.e. the more you spend, the more you make). But as discussed earlier, the variable that means the most to us is the “net revenue before subsidy,” which gives us the amount of revenue pulled after expenses and without factoring in the universities own investements in itself. For all intents and purposes, this number is a universities profits.

```
model6 = lm(net_revenue_before_subsidy ~ subsidy + external_revenue + full_time_enrollment + instate_tu
            data = football) #Rsq .6844 AdRsq .6832
summary(model6)
confint(model6)
hist(residuals(model6))
plot(model6, which = c(1,2))
```

This sixth model gave us what we wanted in a few ways: 1) it predicted the number that actually means the most–“athletic revenues” was almost a 1 for 1 product of “athletic expenses” and this didn’t tell us much–and 2) it gave us the strongest R squared and Adjusted R squared while avoiding predictors with co-linearity, though it should be noted that both of these metrics aren’t quite at 0.7, so “strong” is relative.

### Testing the Model

```
set.seed(88)
trainingRowIndex <- sample(1:nrow(football), 0.8*(nrow(football))) # row indices for training data
trainingData <- football[trainingRowIndex, ] # model training data
testData <- football[-trainingRowIndex, ] # test data

model6test <- lm(net_revenue_before_subsidy ~ subsidy + full_time_enrollment + external_revenue + insta
                 data = testData)

netRevPreds <- predict(model6test, testData)

cor(netRevPreds, testData$net_revenue_before_subsidy) # correlation accuracy of 91.74%
cor(netRevPreds, testData$net_revenue_before_subsidy)**2 # model has an R-squared of 0.8417 on the test
```

A very nice 91.74% correlation accuracy gives us an R-squared score of 0.8417 on the test data–a strong but not over-fitting linear regression. Perhaps we could improve this even further by conducting Random Forests though. That code is below:

```
library(randomForest)

rf <- randomForest(net_revenue_before_subsidy ~ subsidy + full_time_enrollment + external_revenue + ins
                   data = trainingData,
                   importance = TRUE,
                   ntree = 1000)
summary(rf)

which.min(rf$mse) # 160

plot(rf)

imp <- as.data.frame(sort(importance(rf)[,1],decreasing = TRUE), optional = T)
names(imp) <- "% Inc MSE"
imp

netRevPredsRF <- predict(rf, testData)
cor(netRevPredsRF, testData$net_revenue_before_subsidy) # 91.28%
cor(netRevPredsRF, testData$net_revenue_before_subsidy)**2 # 0.8333 is our Random Forests R-squared
```

Random Forests in fact does not improve on the model, though the drop off is hardly substantial. We get a correlation accuracy of 91.28% and an R-squared score of 0.8333

## Recommendations

Based on the analysis of these public universities:

1. If you are a university attracted to the allure of the most prominent programs, know that these are the exception, not the rule. Even for these exceptions, the success is a mirage, it's often measured by the number of positive articles online or the tv coverage of games–not measured in dollars and cents, at least not in the immediate numbers detailed in their NCAA Expenses and Revenues forms. **If this is your situation, be prepared to hemmorage lots and lots of money.**
2. Practice fiscal responsibility before getting caught up in an arms race. Rivalries have driven universities to spend more and more in the past 10-20 years, and the numbers say that it's gotten away from just about everyone. Winning records did not correlate well with any spending-related predictor. An athletic program is possible to sustain, and better yet, the revenue potential is absurd, but so is the spending potential.
3. If you are a university subsidizing your own athletic program, and you find your subsidy contribution is increasing over time, understand that an increase in subsidy amount strongly predicts lower net revenue before subsidy (our "profit" metric). **If you are in this boat, strongly consider cutting your losses and assess the continued future of your football program.**

## Ideas for Further Research

This data set includes Win percentages for football teams for each year that correlate with the financial data examined. I certainly set out to see how winning would/could affect (or be effected by) different financial variables, but nothing immediately appeared to me (and honestly the wrangling required to assess win percentages as a possible variable was miserable).

It would also be productive to examine the impacts of subsidizing and astronomical expenses on the individual students. Because we have the full time enrollment variable, we can find a "per student" amount for a lot of significant information. For example, futher research could display the impact athletic programs have on individual student tuition. How much lower could tuition be? What are students contributing to a program that is losing the university tens of millions of dollars?