Capstone Statistical Analysis

Grant Cox

January 2, 2019

Football Business by the Numbers

Roughly three years ago, Huffington Post dug deep into every available NCAA Revenue and Expense Report and posted their data in an article here: http://projects.huffingtonpost.com/ncaa/sports-at-any-cost.

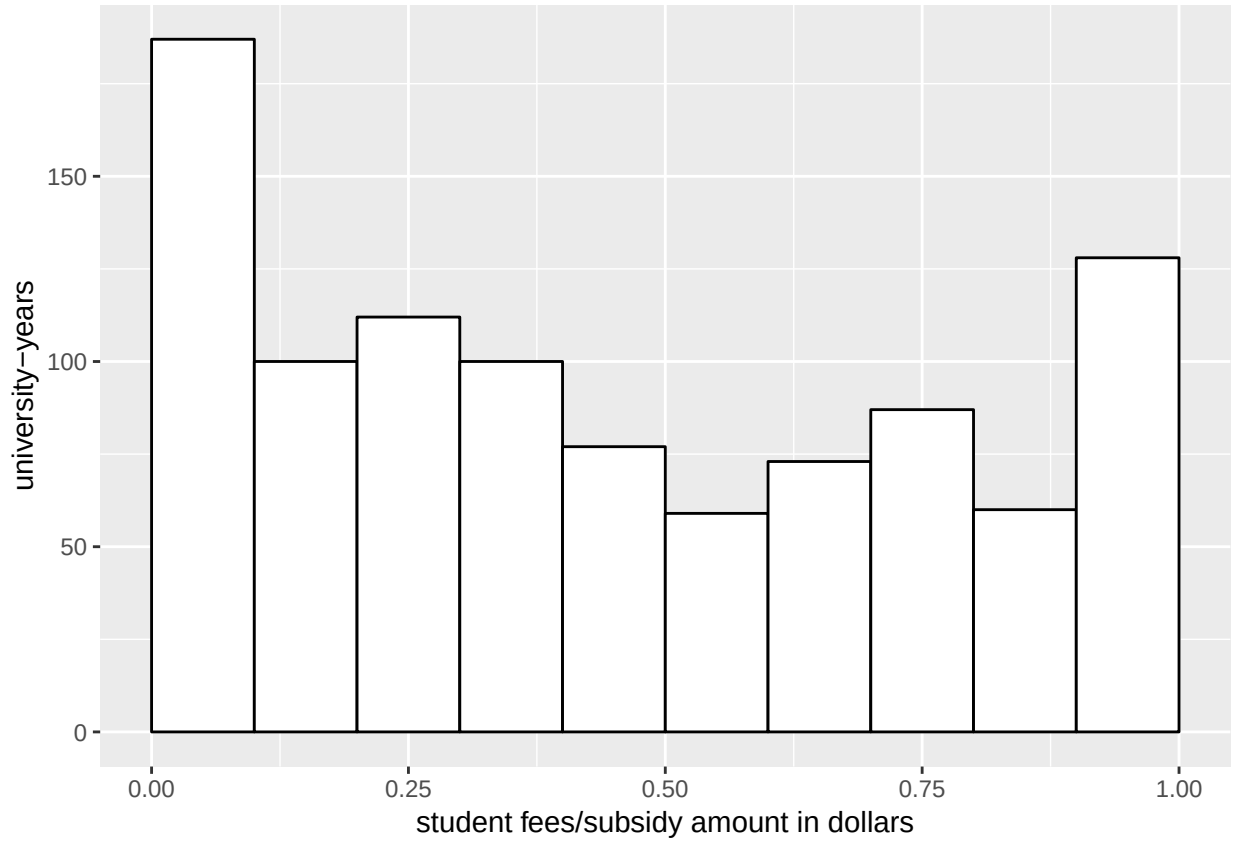
```
library(readxl)
combined_df <- read.csv("C:/Users/gcox3/Desktop/capstone/combined_df.csv")
# View(combined_df)

library(ggplot2)
```

Can you count something interesting?

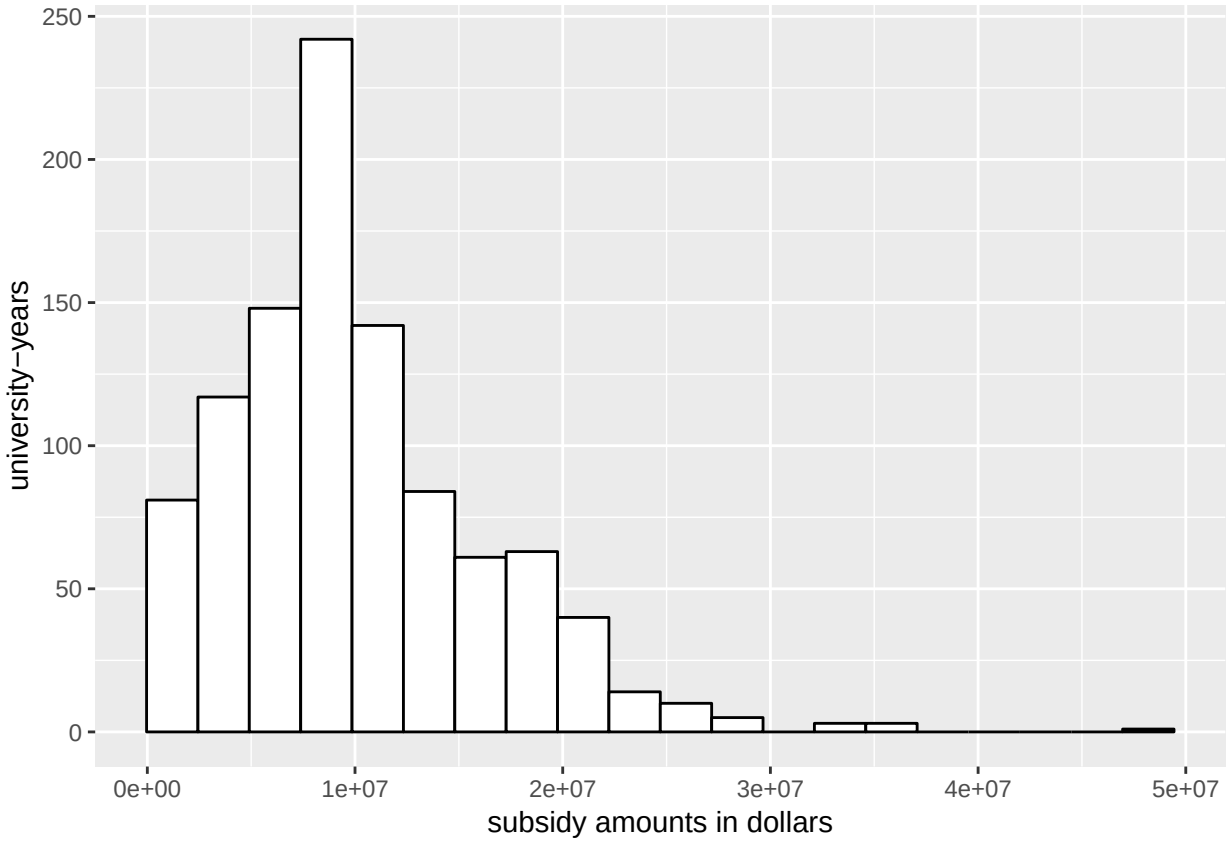
When looking at what we might want to count, there are a few considerations to take, as certain questions will want to count certain entities differently. For example, we could of course count *universities* when looking specific revenues or enrollments, but the data available gives 5 years to every university. We would count each data point that includes both an institution name and a correlate year as a *univeristy-year*. Because we don't necessarily need 5 data points for each university to answer some questions, we'll ultimately create a composite of all the years available and just call it university 5-year average. ## Can you find some trends (high, low, increase, decrease, anomalies)? We could probably intuit some trends, such as revenue and expenses, but it's likely much more involved. ## Can you make a bar plot or a histogram? The percentage of the subsidy that is actually taken directly from students ("student fees") varied from university to university (and even year to year). We can use a histogram to establish some buckets and see just how schools rely on student fees to fund their subsidy.

```
ggplot(combined_df, aes(x = student_fees/subsidy),stat_bin()) +
  geom_histogram(binwidth = 0.1, color = "black", fill = "white", center = 0.05, na.rm = TRUE) +
  labs(x = "student fees/subsidy amount in dollars", y = "university-years")
```



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:

```
ggplot(combined_df, aes(x = subsidy)) +
    geom_histogram(bins = 20, color = "black", fill = "white", na.rm = TRUE, center = 1200000) +
    labs(x = "subsidy amounts in dollars", y = "university-years")
```

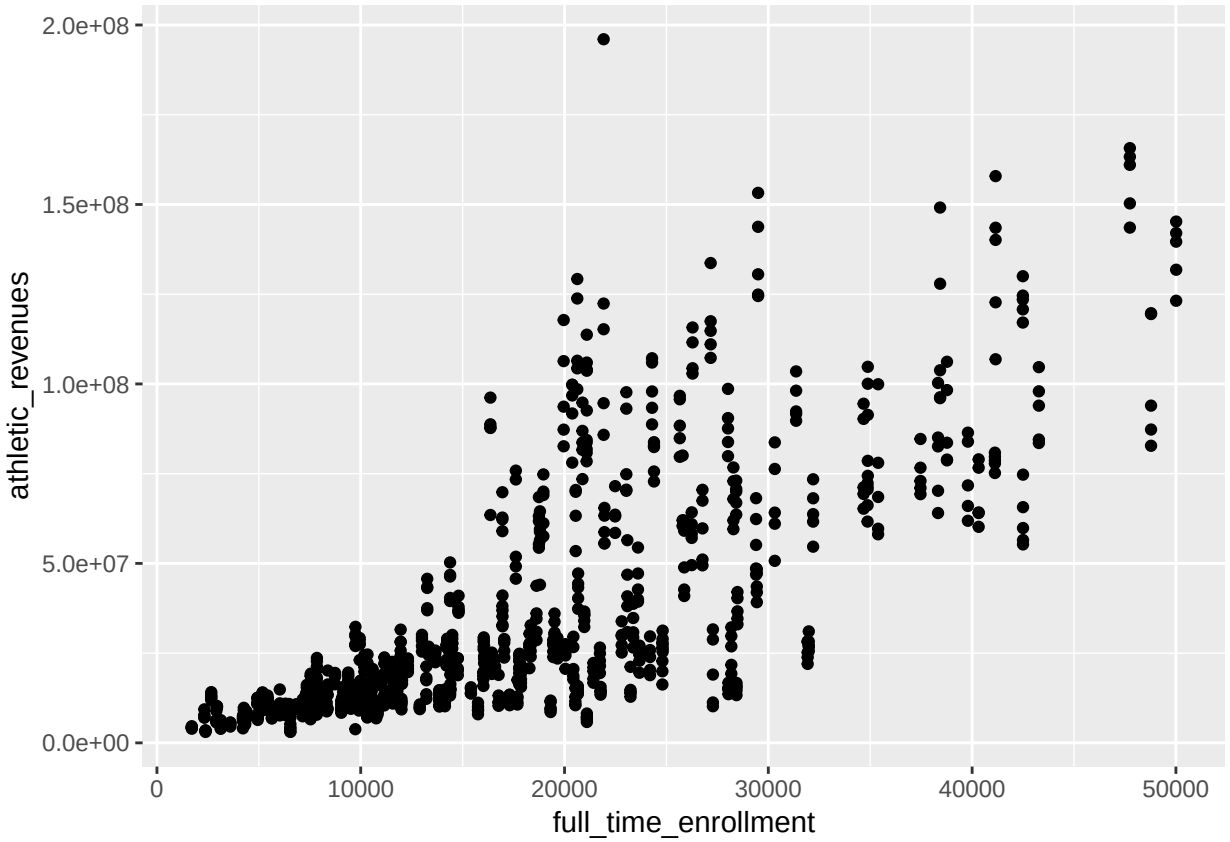


Right skewed ## Can you compare two related quantities? There are two "types" of quantities that are worth comparing considering their intuitive relations. The first type is "revenues" and includes the variables Athletic Revenues, Net Revenue, and Net Revenue before Subsidy. These all relate uniquely to Athletic Expenses. ## Can you make a scatterplot? We can make a few different scatter plots to check for trends. The most interesting might be how certain variables relate to the dependent variable of Win Percentage. In essence, which inputs from the school correlate positively with increased winning? A scatterplot may also give us a chance to see how all these data points trend when observing how the sheer size of the university (enrollment) affects revenue.

```
known_exp <- combined_df[which(combined_df$athletic_expenses!=0),]

# revenues v university size

point3 <- ggplot(known_exp, aes(x = full_time_enrollment, y = athletic_revenues))
point3 +
  geom_point()
```



*** Insights and Next Steps

Having made these plots, what are some insights you get from them? Do you see any correlations? Is there a hypothesis you would like to investigate further? What other questions do they lead you to ask?

As predicted, there are some strong correlations between revenues and expenses, but we remove the universities' subsidy contributions, that correlation weakens. There is more variation from university to university when looking at net revenue before subsidy and the correlating expenses. This shows us that some universities are managing their budgets (at least in this specific regard) better than others.

What we undoubtedly want to investigate further is the financial variables that might best predict winning. At this point, statistics and probability might not point to that answer, perhaps a linear or logistic regression might?