Final Capstone Project Draft

The objective of my data analysis is to find universities across the US that are most catering to students coming from low income backgrounds and in addition prepare the underserved for success post grad rather than lead into a financially burdensome rut. My hope is that the set of schools would show a general trend or be schools with specific factors (private or public, religious or non-religious, etc).

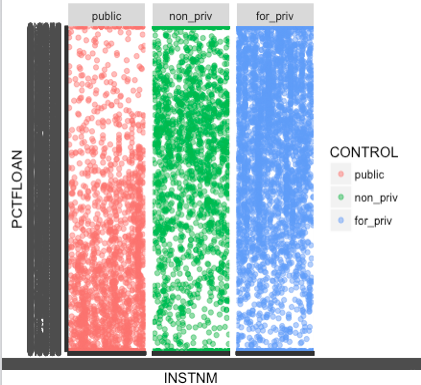
The first thing I did is divide up the data set. The initial data frame is 7703 observations of 122 variables. I cut it down to 10 variables: INSTNM (institution name), STABBR (state), CONTROL (public or private), PCTFLOAN (% of undergraduate students having financial aid), RELAFFIL (where school ranks on religious affiliation scale), GT\_25K\_P6 (fraction of former students earning above $25,000), NPT4\_PUB (net price for public university), NPT41\_PUB (net price of public university from income 0-$30,000), NPT4\_PRIV (net price, private), NPT41\_PRIV (net price of private university from income 0-$30,000). I assigned this data frame as *college\_edit.*

***1. Initial exploratory analysis:***

**a. Loans:**

My initial exploratory analysis consisted of doing basic scatter plots and frequency distributions. The first one is a scatter plot between schools (INSTNM) and the percentage of undergraduate students having loans (PCTFLOAN) partitioned by whether the school is public, private for-profit or private non-profit (CONTOL which is factored into levels *“public”, “for\_priv”, and “non\_priv”*):

ggplot(college\_edit, aes(x = INSTNM, y = PCTFLOAN, col = CONTROL)) + geom\_jitter(alpha = 0.5) + facet\_grid(.~CONTROL)



The result is a very over plotted scatter plot since there are 7703 observations. Nevertheless, we see that at public (*public*) universities there are many occurrences where the percentage of students having loans is low and many occurrences where percentage of students having loans is high at private for-profit (*for\_priv*) universities. The average for private nonprofit universities is right between these two. My initial thought here was that **private for-public universities are more expensive than those that are private nonprofit or public and hence results in more students taking loans.** I decided to then take the average percentage of undergraduates taking loans for each of the CONTROL variable. I divided the data set college\_edit to three different data sets and took the average loan percentage for each of them:

college\_public <- subset(college\_edit, CONTROL == "public")

loan <- college\_public$PCTFLOAN

loan <- as.numeric(as.character(loan))

avgLoan\_public <- mean(loan, na.rm = TRUE)

college\_forPriv <- subset(college\_edit, CONTROL == "for\_priv")

loanThree <- college\_nonPriv$PCTFLOAN

loanThree <- as.numeric(as.character(loanThree))

avgLoan\_non\_priv <- mean(loanThree, na.rm = TRUE)

college\_nonPriv <- subset(college\_edit, CONTROL == "non\_priv")

loanTwo <- college\_forPriv$PCTFLOAN

loanTwo <- as.numeric(as.character(loanTwo))

avgLoan\_for\_priv <- mean(loanTwo, na.rm = TRUE)

The results were the following: *avgLoan\_public = 0.32 (32%)*, *avgLoan\_non\_pub = 0.56 (56%)*, and *avgLoan\_for\_priv = 0.62 (62%)*. It’s plausible to infer that private for-profit universities generally require more loans, maybe because it is more expensive.

1. **Net Price:**

To compare percentage of undergraduates having loans to average net price of schools could help confirm whether for-profit colleges are the most expensive and need more loans. To do this, I took the average net price (NPT4\_PUB for public and NPT4\_PRIV for private nonprofit, private for profit).

netPricePublic <- mean(college\_public$NPT4\_PUB, na.rm = TRUE)

netPriceNonPriv <- mean(college\_nonPriv$NPT4\_PRIV, na.rm = TRUE)

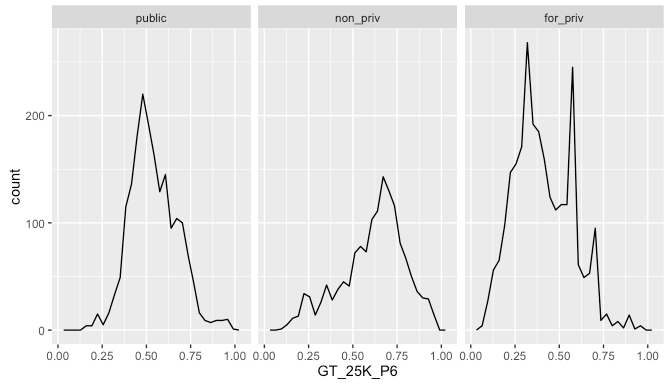
netPriceForPriv <- mean(college\_forPriv$NPT4\_PRIV, na.rm = TRUE)

The results were the following: *netPricePublic = $9,624.65, netPriceNonPriv = $20,266.36 and netPriceForPriv = $17,293.57.* Since average net price for private for profit universities is lower than that of private non-profit universities, **we cannot with certainty see a positive correlation between average net price of a university and loans when partitioned by public, private non-profit, and private for-profit.**

1. **Post Grad Earnings > 25K:**

Then I used a frequency distribution to observe the fraction of former students earning above $25,000 (GT\_25K\_P6) partitioned by public vs. private for-profit vs. private non-profit:

ggplot(college\_edit, aes(x = GT\_25K\_P6)) + geom\_freqpoly() + facet\_grid(.~CONTROL)



The following doesn’t show much. The graphs of public universities and private non-profits show a general peak. There are a lot of public universities where a little less than 50% of undergraduates earn above $25,000 after they graduate, whereas that percentage is around 64% for private non-profit universities. The private for-profit universities show more variation with around 3 local peaks, and higher counts of each of those percentages. We must take into account however that in our list of 7703 universities/colleges, there are 2044 public, 1956 private non-profit universities, and 3703 private for-profit universities, which probably accounts for the higher number of incidents in the graphs for public and private for-public universities.

1. **Religious Affiliation:**

Lastly, I wanted to see if more religiously affiliated (RELAFFIL) universities had any effect on loans and earnings above $25,000 after graduation. All religious universities are listed as private non-profit. So we can perform our analyses on just the college\_nonPriv data frame. The resulting plots did not show much correlation between more religiously affiliated schools and loans/post grad earnings, as the points were scattered all over the graph with no real trend. Therefore, I will not show the graph and temporarily present that religious affiliation does really associate to undergraduate loan percentage or higher percentage of successful graduates.

***2. Next Steps:***

1. **Correlation:**

The immediate step that comes to mind is to start correlations between all the variables. Certain correlations to keep an eye on in particular would be between net price of universities and the percentage of undergraduates having loans in school which I believe could clarify whether higher loan rate could mean higher cost of attending. Another correlation is between what type of school (public, private non-profit, private for-profit) and percentage of undergraduate students earning above $25,000 post graduation.

1. **Linear/Logistic Regression:**

Making average net price and loans independent variables and the percentage of undergraduate students earning above $25,000 after graduation the independent variables, I will do a linear regression, seeing if those variables in particular affect post grad income in any way. The reason for this is because for a college to be affordable and valuable, it should first of all not require a ton of loans and in addition provide a smooth transition into the work world.

1. **Clustering:**

Clustering the universities by net price, post grad income and other variables are all possible depending on any trends in the data set as I delve into more analysis. It’s difficult to know exactly which clustering would be most insightful at this point.

1. **Final word:**

I will expand and reduce the dataset as I go, to see whether variables I didn’t include initially could be serve as a new asset to the problem we are trying to solve. Debt repayment rate is probably one of those variables.