Reducing Income Equality through Elite Education

Michael L. Thompson

February 25, 2017

Contents

Introduction

This is an exploratory analysis with the College Scorecard dataset to see if the academic elite colleges improve income mobility of lower-to-middle-income students.

Universe of Discourse

We'll limit the analysis to just the colleges meeting these criteria, which divide the colleges, and the students that attend them, into two groups – those at large public universities and those at high-median-SAT private colleges:

Load U.S. Department of Education College Scorecard Data

The data are split into 2013 data describing the colleges and into 2005 Treasury data describing the financial status of the students' households. The 2013 data describe the student body of the college at that time as well as the earnings and debt situations of the previous graduates of the college.

Prep the Data: Limit it to our Universe of Discourse

We'll strip out the colleges that are NOT 4-year Bachelor's degree-granting colleges, the private for-profits, the private non-profits having average SAT less than 1440 with at least 500 undergraduates, and the publics with less than 20,000 students.

We're left with just 115 colleges: 86 typical large public colleges, and 29 "elite" private colleges.

```
UGDS_threshold <- 20000
UGDS_threshold_private <- 500
SAT_threshold  <- 1400
student <- student2013 %>% tbl_df() %>%
  select(-contains('2005')) %>%
  filter(
    currop == 'Currently certified as operating',
    CollegeType != 'Private for-profit',
    degree == "Predominantly bachelor's-degree granting",
    (CollegeType == 'Public' & UGDS >= UGDS_threshold) | (CollegeType == 'Private nonprofit' & SAT_AVG 
  ) %>%
  left_join( student2005 %>% select(unitID,contains('2005')) ) %>%
  mutate( Category = ifelse( CollegeType == 'Public', "Large Public Universities", "Elite Private Unive
  select(
    unitID,
    College,
    CollegeType,
    Category,
    UGDS,
    SAT_AVG,
    first_gen_2005,
    contains( "INC_DEBT" ),
    contains( "INC_PCT" ),
    TUITIONFEE_IN,
    matches( "^NPT4[1-5]" ),
    contains( "mn_earn_wne_inc" )
  ) %>%
  mutate_each( funs( ifelse(. == 'PrivacySuppressed',NA, . ) ) , contains("INC_PCT", ignore.case = FALS
  mutate_each( funs( as.numeric ) , contains("INC_PCT", ignore.case = FALSE ) )%>%
  mutate(
    NPT41 = ifelse( CollegeType == 'Public', NPT41_PUB, NPT41_PRIV ),
    NPT42 = ifelse( CollegeType == 'Public', NPT42_PUB, NPT42_PRIV ),
    NPT43 = ifelse( CollegeType == 'Public', NPT43_PUB, NPT43_PRIV ),
    NPT44 = ifelse( CollegeType == 'Public', NPT44_PUB, NPT44_PRIV ),
    NPT45 = ifelse( CollegeType == 'Public', NPT45_PUB, NPT45_PRIV ),
    suspect_debt_number = (LO_INC_DEBT_MDN == 0) & (is.na(mn_earn_wne_inc1_p6_2005) | (mn_earn_wne_inc1_
    LO_INC_DEBT_MDN = ifelse( suspect_debt_number, NA, LO_INC_DEBT_MDN ),
    MD_INC_DEBT_MDN = ifelse( suspect_debt_number, NA, MD_INC_DEBT_MDN ),
    HI_INC_DEBT_MDN = ifelse( suspect_debt_number, NA, HI_INC_DEBT_MDN )
  ) %>%
  select( -matches('_(PUB|PRIV)'), -suspect_debt_number ) %>%
  arrange( desc( SAT_AVG ), unitID )

student %>% group_by(Category) %>% tally() %>% formattable()
```

Category

n

Elite Private Universities

29

Large Public Universities

86

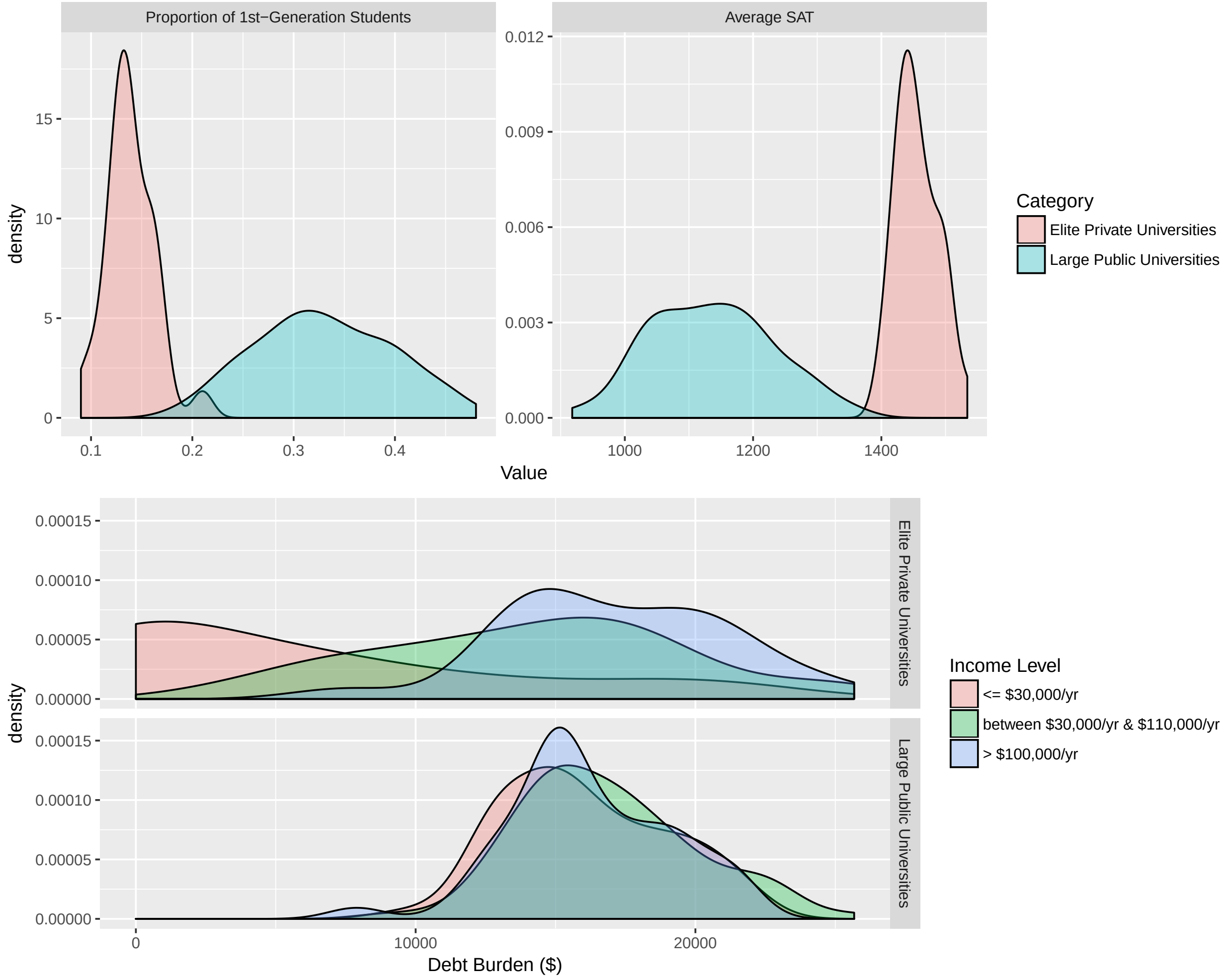
```
hidden_list <- c(
  "unitID",
  "CollegeType",
  "MD_INC_DEBT_MDN", "HI_INC_DEBT_MDN",
  "INC_PCT_M1", "INC_PCT_M2", "INC_PCT_H1", "INC_PCT_H2",
  "NPT42", "NPT43", "NPT44", "NPT45",
  "mn_earn_wne_inc1_p6_2005",
  "mn_earn_wne_inc2_p6_2005",
  "mn_earn_wne_inc3_p6_2005"
  ) %>%
{
  setNames(as.list(rep(FALSE,length(.))),.)
}
student %>%
  mutate(Earnings_6yr = ifelse(mn_earn_wne_inc1_p6_2005==0,NA,mn_earn_wne_inc1_p6_2005)) %>%
  formattable( hidden_list ) %>%
  as.datatable()
```

Public vs. Elite Private College Differences

Now, let's take a look at some of the differences between the two categories of colleges we've focused upon, as illustrated by the plots below:

- The public colleges enroll considerably more first-generation-college students than the elite private colleges.
- However, considering that the average SAT scores of students at the elite private colleges are considerably higher than those of the students at the public colleges, the disparity in first-generation students may reflect more the differences in early education opportunities offered to students whose parents attended college relative to those for students whose parents did not attend college.
- Lower-to-middle-income students graduating from the elite private colleges leave with considerably less debt than do those of the same income cohort groups graduating from the public colleges.

```
# First Generation College: first_gen
# SAT distributions: SAT_AVG
var_label <- c( first_gen_2005 = 'Proportion of 1st-Generation Students', SAT_AVG = 'Average SAT' )
gplt_1st_gen <- student %>%
  select( College, Category, first_gen_2005, SAT_AVG ) %>%
  gather( key = Characteristic, value = Value , -College, -Category ) %>%
  mutate( Characteristic = factor( var_label[Characteristic], levels = var_label ) ) %>%
  {
    ggplot( ., aes( x = Value, fill = Category ) ) +
      geom_density( alpha = 0.3 ) +
      facet_wrap( ~ Characteristic , scales = 'free' )
  }

# Debt Burden by Income: LO_INC_DEBT_MDN, MD_INC_DEBT_MDN, HI_INC_DEBT_MDN
debt_label <- c(
  LO_INC_DEBT_MDN = '<= $30,000/yr',
  MD_INC_DEBT_MDN = 'between $30,000/yr & $110,000/yr',
  HI_INC_DEBT_MDN = '> $100,000/yr'
)
gplt_debt_burden <- student %>%
  select( College, Category, LO_INC_DEBT_MDN, MD_INC_DEBT_MDN, HI_INC_DEBT_MDN ) %>%
  gather( key = `Income Level`, value = `Debt Burden ($)` , -College, -Category ) %>%
  mutate( `Income Level` = factor( debt_label[`Income Level`], levels = debt_label ) ) %>%
  {
    ggplot( ., aes( x = `Debt Burden ($)`, fill = `Income Level` ) ) +
      geom_density( alpha = 0.3 ) +
      facet_grid( Category ~ . )
  }
income_cohort <- c(
  DEP_INC_PCT_LO = '< $30,000/yr',
  DEP_INC_PCT_M1 = 'between $30K/yr & $48K/yr',
  DEP_INC_PCT_M2 = 'between $48K/yr & $75K/yr',
  DEP_INC_PCT_H1 = 'between $75K/yr & $110K/yr',
  DEP_INC_PCT_H2 = '> $110,000/yr',
  INC_PCT_LO = '< $30,000/yr',
  INC_PCT_M1 = 'between $30K/yr & $48K/yr',
  INC_PCT_M2 = 'between $48K/yr & $75K/yr',
  INC_PCT_H1 = 'between $75K/yr & $110K/yr',
  INC_PCT_H2 = '> $110,000/yr'
)
grid.arrange( gplt_1st_gen, gplt_debt_burden, nrow = 2 )
```



Debt, net price, and earnings

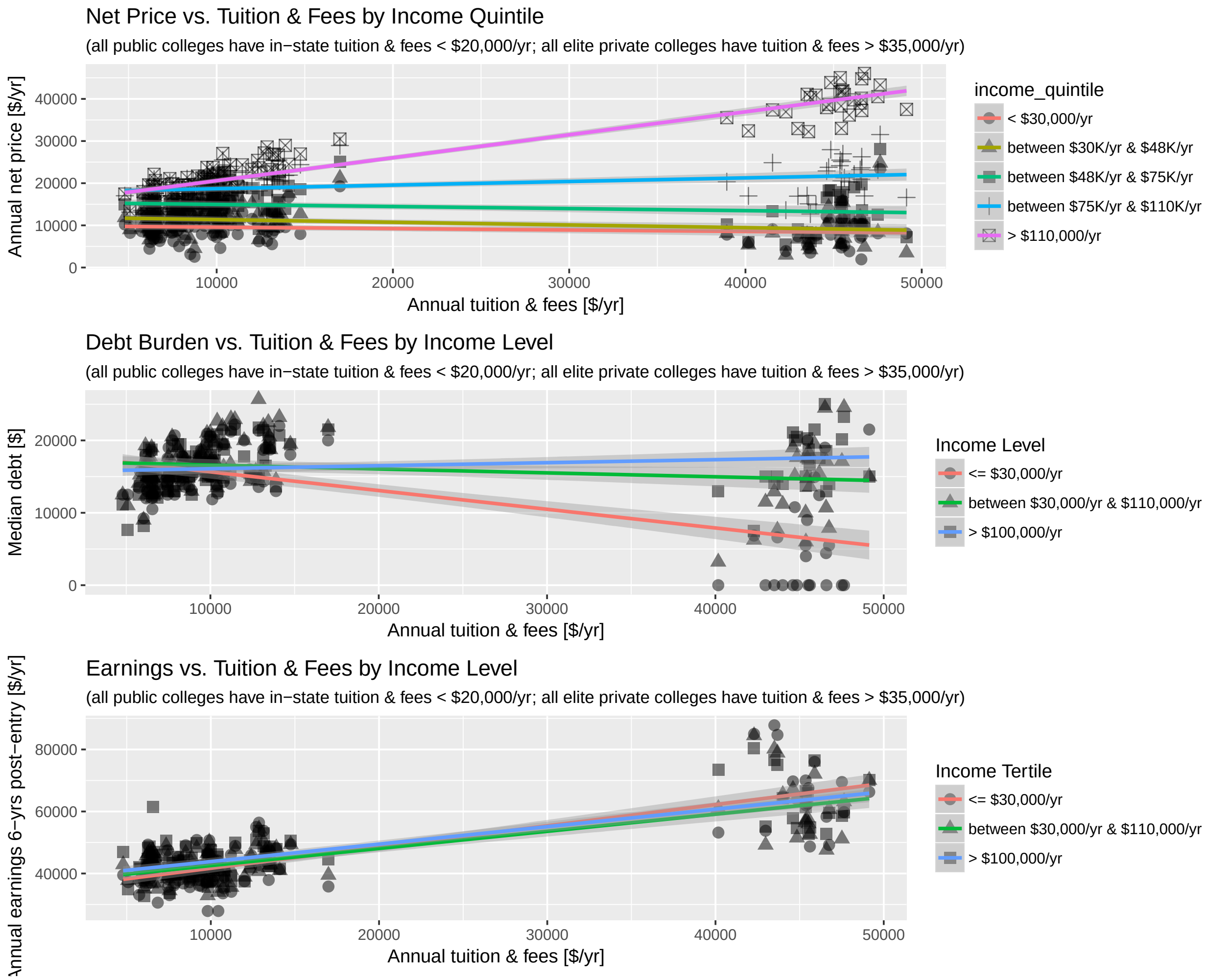
In the next plots, we show lines extending from the cluster of points at the lower annual tuition and fees, which correspond to the public colleges, to the cluster at the higher annual tuition and fees, which correspond to the elite private colleges. These lines are not intended to imply a linear relationship between either of the variables plotted on the vertical (y-axis) vs. annual tuition and fees on the horizontal (x-axis). (After all, we've intentionally omitted all of the colleges that would fill in the full continuous range of annual tuition and fees.) Instead, the lines are there to simply make it easier to see the differences, or lack thereof, between the y-variables of the two different clusters at the different income levels. In other words, amongst these colleges we're comparing, annual tuition and fees just serves as a continuous proxy metric for the categories "public college" and "elite private college". And we can see from these plots that the annual tuition and fees at the elite private colleges are more than double those of the public colleges.

Despite the large difference in "list price" between elite private colleges and public colleges, the students in the lower two income quintiles who attended the elite private colleges still ended up having considerably less debt than the students from families with similar incomes but who attended the public colleges. Financial aid offered at the elite private colleges ends up making the "net price", which is "list price" (tuition + fees + living expenses) less "aid", basically the same for all but the students from the highest income families.

Considering the academic abilities (as judged by average SAT scores, shown above) of the students at elite private colleges, those of them in the lower two income quintiles might more easily find summer employment within their fields or at least summer employment paying them more than students with similar family incomes but having lower SAT and attending public colleges. Therefore, despite having similar net prices each year whether they attend an elite private college or a public college, lower-to-middle income students who attend the elite private colleges may earn considerably more money than those at public colleges and end up having less debt.

We can also see that the average earnings (6 years after entry into the college) of students from the elite private colleges are considerably higher than that of the students from the public colleges. This also suggests, as mentioned above, that even during their time at the college, the students who attended the elite private colleges, may have been able to earn more of their education costs and thus incur less debt than students of similar family income who attended public colleges.

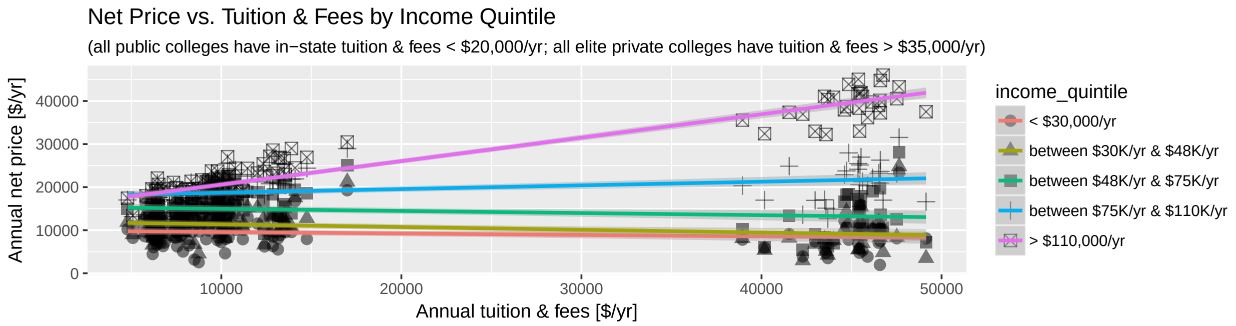
```
# Net price

df_price <- student %>%
  select(College,CollegeType,Category,matches("^NPT4[1-5]"),TUITIONFEE_IN) %>%
  gather(key=income_quintile,value=net_price,starts_with('NPT'),na.rm = TRUE) %>%
  mutate(income_quintile=factor( income_cohort[as.integer(gsub('NPT4([1-5]).*','\\1',income_quintile))]
  filter(complete.cases(.)) %>%
  arrange(CollegeType,College) %>%
  mutate( net_price_pct = 100 * net_price / TUITIONFEE_IN )

gplt_price <- df_price %>%
{
  ggplot(., aes( x = TUITIONFEE_IN, y = net_price, shape = income_quintile, group = income_quintile  ))
    geom_point( size = 3, alpha = 0.5 ) +
    geom_smooth( aes(color = income_quintile), method = "lm" ) +
    labs( x = 'Annual tuition & fees [$/yr]', y = 'Annual net price [$/yr]' ) +
    ggtitle(
      label = "Net Price vs. Tuition & Fees by Income Quintile" ,
      subtitle = "(all public colleges have in-state tuition & fees < $20,000/yr; all elite private col
    )
}

# Debt burden
df_debt <- student %>%
  select(College,CollegeType,Category,contains("INC_DEBT"),TUITIONFEE_IN) %>%
  gather( key = `Income Level`, value = `Debt Burden ($)` , contains("INC_DEBT") ) %>%
  mutate( `Income Level` = factor( debt_label[`Income Level`], levels = debt_label ) ) %>%
  filter( complete.cases(.) ) %>%
  arrange( CollegeType, College )

gplt_debt <- df_debt %>%
{
  ggplot(., aes( x = TUITIONFEE_IN, y = `Debt Burden ($)`, shape = `Income Level`, group = `Income Leve
    geom_point( size = 3, alpha = 0.5 ) +
    geom_smooth( aes(color = `Income Level`), method = "lm" ) +
    labs( x = 'Annual tuition & fees [$/yr]', y = 'Median debt [$]' ) +
    ggtitle( label = "Debt Burden vs. Tuition & Fees by Income Level" ,
             subtitle = "(all public colleges have in-state tuition & fees < $20,000/yr; all elite priva
    )
}

# Earnings
inc_tertile_label <- c(
  mn_earn_wne_inc1_p6_2005 = '<= $30,000/yr',
  mn_earn_wne_inc2_p6_2005 = 'between $30,000/yr & $110,000/yr',
  mn_earn_wne_inc3_p6_2005 = '> $100,000/yr'
)
df_earnings <- student %>%
  select(College,CollegeType,Category,contains("mn_earn_wne_inc"),TUITIONFEE_IN) %>%
  gather( key = `Income Tertile`, value = `Earnings ($/yr)` , contains("mn_earn_wne_inc") ) %>%
  mutate( `Income Tertile` = factor( inc_tertile_label[`Income Tertile`], levels = inc_tertile_label )
  filter( complete.cases(.),  `Earnings ($/yr)` > 0) %>%
  arrange( CollegeType, College )

gplt_earnings <- df_earnings %>%
{
  ggplot(., aes( x = TUITIONFEE_IN, y = `Earnings ($/yr)`, shape = `Income Tertile`, group = `Income Ter
    geom_point( size = 3, alpha = 0.5 ) +
    geom_smooth( aes(color = `Income Tertile`), method = "lm" ) +
    labs( x = 'Annual tuition & fees [$/yr]', y = 'Annual earnings 6-yrs post-entry [$/yr]' ) +
    ggtitle( label = "Earnings vs. Tuition & Fees by Income Level" ,
             subtitle = "(all public colleges have in-state tuition & fees < $20,000/yr; all elite priva
    )
}

grid.arrange( gplt_price, gplt_debt, gplt_earnings, ncol = 1 )
```

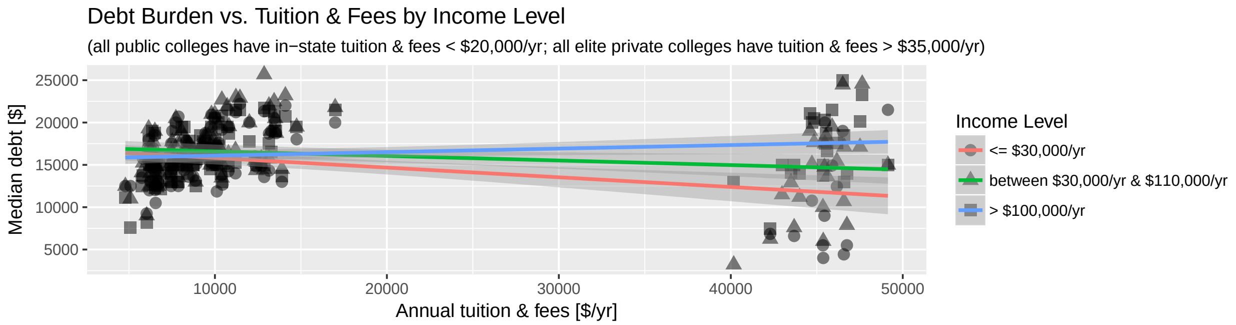


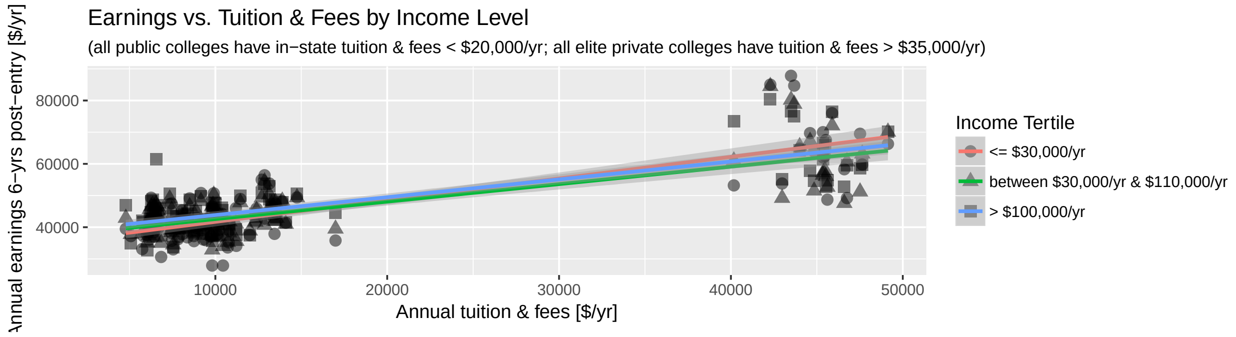
It's unclear whether all the zeros for the debt burden of lower-income students at the elite private colleges are genuine zeros or missing data. So, we redo the last set of plots excluding those zeros to check the robustness of the conclusion that lower-income students graduating from the elite private colleges have less debt than those graduating from the public colleges.

The middle plot below shows that the difference still exists but may not be statistically significant, calling into question our speculation about greater earnings from summer employment by students at the elite private colleges. Of course, rigorous statistical analysis would provide a quantitative result.

```
gplt_debt <- df_debt %>% filter( `Debt Burden ($)` > 0 ) %>%
{
  ggplot(., aes( x = TUITIONFEE_IN, y = `Debt Burden ($)`, shape = `Income Level`, group = `Income Level
    geom_point( size = 3, alpha = 0.5 ) +
    geom_smooth( aes(color = `Income Level`), method = "lm" ) +
    labs( x = 'Annual tuition & fees [$/yr]', y = 'Median debt [$]' ) +
    ggtitle( label = "Debt Burden vs. Tuition & Fees by Income Level" ,
             subtitle = "(all public colleges have in-state tuition & fees < $20,000/yr; all elite priva
    )
}

grid.arrange( gplt_price, gplt_debt, gplt_earnings, ncol = 1 )
```







Summary

It's clear from these data, that the elite private colleges represent a considerable financial opportunity to students from families in the lower two income quintiles, under the big caveat that the students have academic credentials impressive enough to warrent admission to the elites (here only indicated by SAT score).

A stab at an exploratory analysis of both an "earnings premium" and a "value proposition" of each college is available here as one of the Kaggle.com public scripts, which I also created.

Either way, it's important to keep in mind that these visualizations are merely qualitative and "conversational" in nature, in that they don't constitute rigorous statistical, quantitative analyses.

-Michael L. Thompson