Midterm Project - College Scorecard

Luke Clement, JuJuan Brown, Steven Hanaway

Due March 28, 2022

Question of Interest: "What institutional characteristics affect student debt?"

We looked through the data and decided to refine the characteristics down into two separate questions. "Is institution location a contributing factor to student debt?" and "Is the degree of education a major contributing factor to student debt?" The second question was further subdivided to find if the type of institution affected student debt. One question would allow us to answer an interesting physical geographical question, while other one would allow for us to examine more traditional factors in student debt. Both questions lead us into our second major question: "Does the type of institution attended affect later earnings?" This information was then used to have an approximate idea of what the most economically efficient way to earn a degree is while still having good earning prospects.

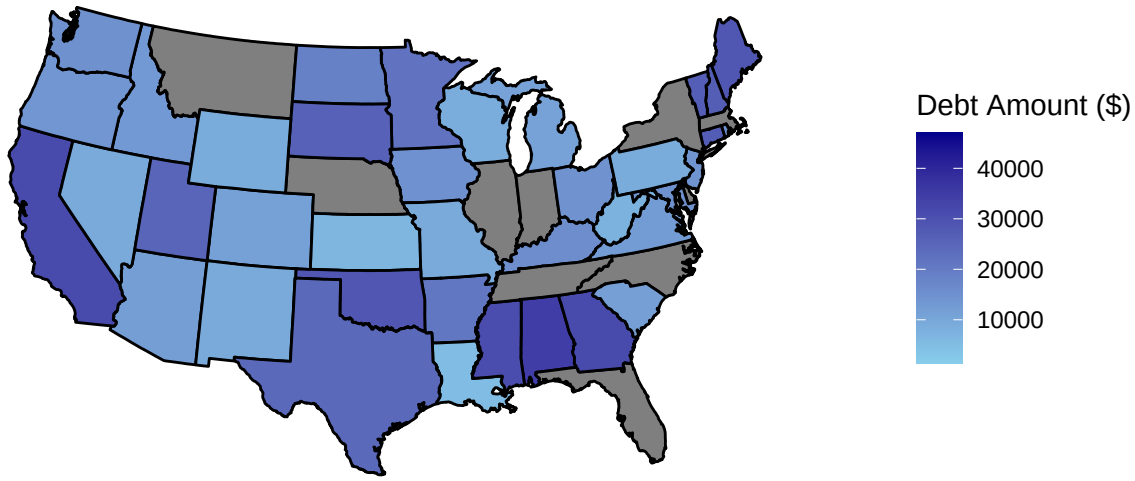
Sub Question: "Is institution location a contributing factor in student debt?"

Showing Debt Distribution Over the United States:

The original inspiration for this question was that we thought schools in areas with more pleasing weather like Florida or more famous attractions like California might have schools with higher tuition because colleges could use those attributes to justify their tuition costs, thus increasing student debt. Since our original data set had information from schools all over the country, we thought doing a graph on a map of the United States would be the best way to visualize our findings. The first challenge was to show how different states varied in the amount of student debt they had, so we included a scale that allowed for each state to have a different level of "darkness" depending on how much debt its students had. We wanted to consider how long students stayed in the school, so we simply created two graphs: one for the students that graduated and a second for the students that did not.

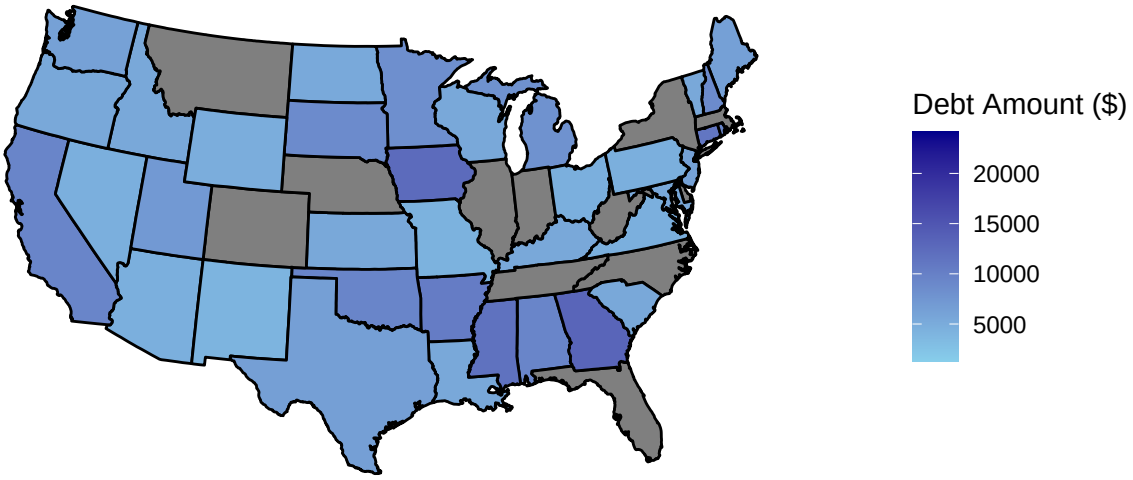
```r
#Selecting Data for Map
data<-read_csv("Most-Recent-Cohorts-All-Data-Elements.csv",
               na=c("NA", "NULL", "PrivacySuppressed")) %>%
  select(STABBR, LATITUDE, LONGITUDE, TUITIONFEE_IN,
         TUITIONFEE_OUT, GRAD_DEBT_MDN, WDRAW_DEBT_MDN) %>%
  rename(region=STABBR)
vec<-data$region
vec<-state.name[match(vec, state.abb)]
vec<-str_to_lower(vec)
data<-data %>%
  mutate(region=vec)
US<-map_data("state")


#Graphing Graduating Student Debt

grad_graph<- ggplot(data, aes(map_id=region, fill= GRAD_DEBT_MDN))+

  geom_map(map = US, color="black") +
  coord_map("ortho", orientation=c(39,-98,0)) +
  xlim(-125,-68)+
  ylim(25,50)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Debt Amount ($)")+
  theme_void()+
  theme(plot.title = element_text(hjust = 0.5, face='bold'),
        plot.subtitle = element_text(hjust = 0.5, face='italic'))+
  labs(title = "National Debt Distribution of Students that Graduated")

grad_graph
```

National Debt Distribution of Students that Graduated



```
#Graphing Withdrawn student Debt

wdraw_graph<- ggplot(data, aes(map_id=data$region, fill= data$WDRAW_DEBT_MDN))+
  coord_map("ortho", orientation=c(39,-98,0))+
  geom_map(map = US, color="black") +
  xlim(-125,-68)+
  ylim(25,50)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Debt Amount ($)")+
  theme_void()+
  theme(plot.title = element_text(hjust = 0.5, face='bold'),
        plot.subtitle = element_text(hjust = 0.5, face='italic'))+
  labs(title = "National Debt Distribution of Students that Withdrew")

wdraw_graph
```

National Debt Distribution of Students that Withdrew

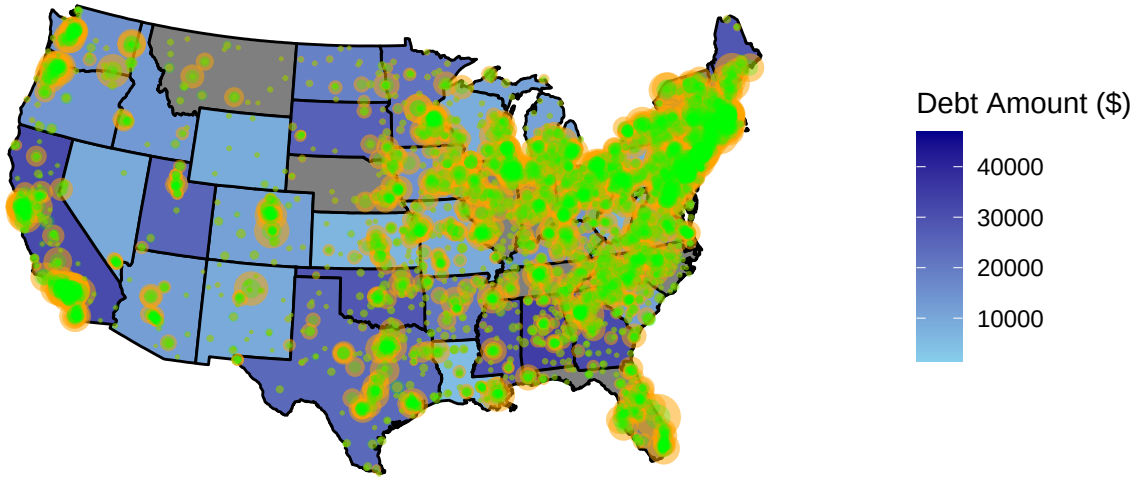


Showing Tuition Fees in the U.S:

The second major factor that was much harder to account for was whether a student was paying in-state or out-of-state tuition. Since we were essentially bringing tourist attractions into question, we needed a way to account for the costs that different students were paying. We originally considered overlaying a scatter plot on top of the map with two different colors: one for the in-state tuition at each school and the one for the out-of-state tuition for each school. This was the result:

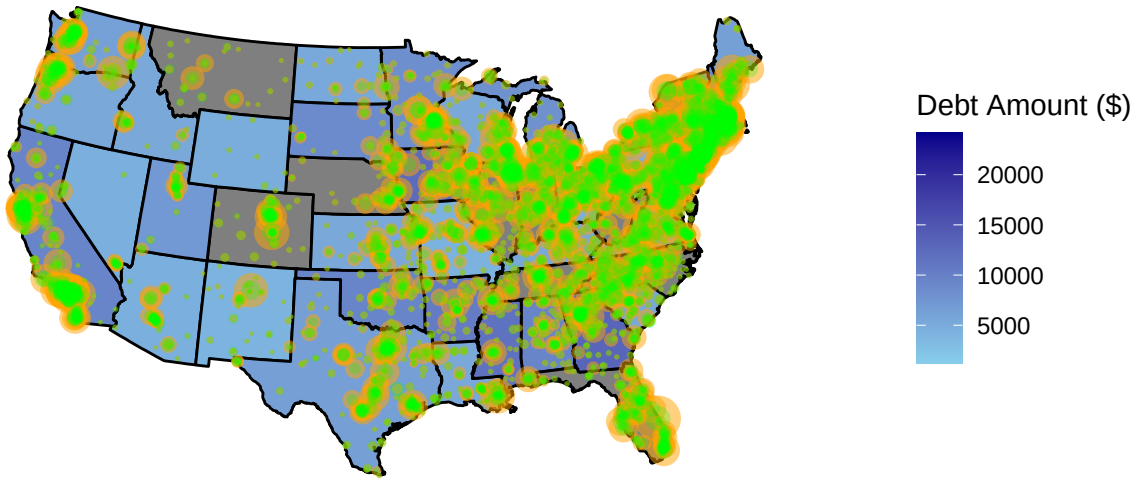
```
# Accounting for in-state and out-of-state tuition for graduating students
grad_graph +

  geom_point(aes(LONGITUDE, LATITUDE), alpha=.5, color="orange",
             size= (data$TUITIONFEE_IN)/10000, na.rm = FALSE)+

  geom_point(aes(LONGITUDE, LATITUDE), alpha=.25, color="green",
             size= (data$TUITIONFEE_OUT)/20000, na.rm = FALSE)
```

National Debt Distribution of Students that Graduated



```
# Accounting for in-state and out-of-state tuition for graduating students

wdraw_graph +

  geom_point(aes(LONGITUDE, LATITUDE), alpha=.5, color="orange",
             size= (data$TUITIONFEE_IN)/10000, na.rm = FALSE)+

geom_point(aes(LONGITUDE, LATITUDE), alpha=.25, color="green",
           size= (data$TUITIONFEE_OUT)/20000, na.rm = FALSE)
```

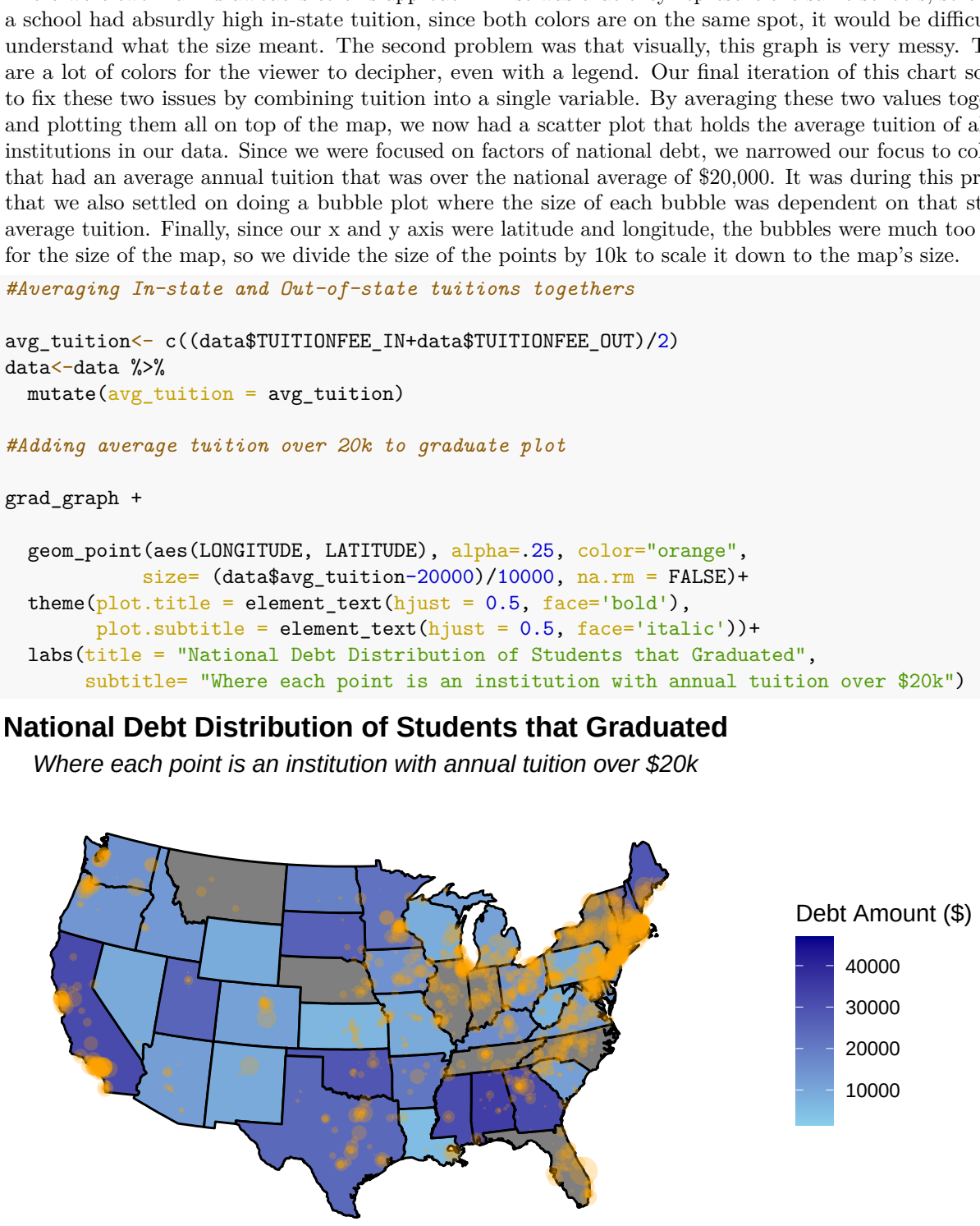
National Debt Distribution of Students that Withdrew



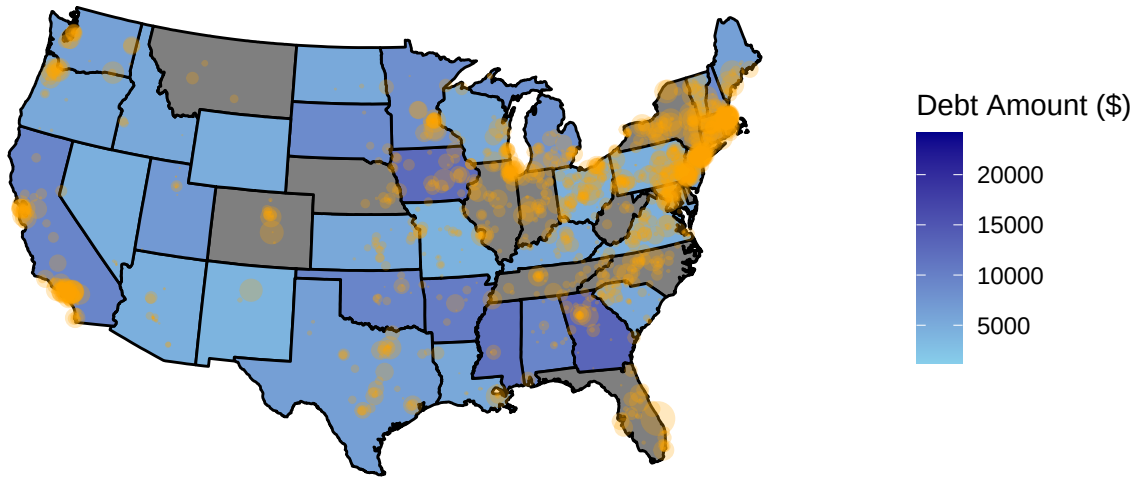
Fixing and Refining the Tuition the Graph:

There were two main drawbacks to this approach. First was that they represent the same schools, so even if a school had absurdly high in-state tuition, since both colors are on the same spot, it would be difficult to understand what the size meant. The second problem was that visually, this graph is very messy. There are a lot of colors for the viewer to decipher, even with a legend. Our final iteration of this chart sought to fix these two issues by combining tuition into a single variable. By averaging these two values together and plotting them all on top of the map, we now had a scatter plot that holds the average tuition of all the institutions in our data. Since we were focused on factors of national debt, we narrowed our focus to colleges that had an average annual tuition that was over the national average of $20,000. It was during this process that we also settled on doing a bubble plot where the size of each bubble was dependent on that state's average tuition. Finally, since our x and y axis were latitude and longitude, the bubbles were much too large for the size of the map, so we divide the size of the points by 10k to scale it down to the map's size.

```
#Averaging In-state and Out-of-state tuitions togethers

avg_tuition<- c((data$TUITIONFEE_IN+data$TUITIONFEE_OUT)/2)
data<-data %>%
  mutate(avg_tuition = avg_tuition)

#Adding average tuition over 20k to graduate plot

grad_graph +

  geom_point(aes(LONGITUDE, LATITUDE), alpha=.25, color="orange",
             size= (data$avg_tuition-20000)/10000, na.rm = FALSE)+
  theme(plot.title = element_text(hjust = 0.5, face='bold'),
        plot.subtitle = element_text(hjust = 0.5, face='italic'))+
  labs(title = "National Debt Distribution of Students that Graduated",
       subtitle= "Where each point is an institution with annual tuition over $20k")
```



```
#Adding average tuition over 20k to withdraw plot

 wdraw_graph + geom_point(aes(LONGITUDE, LATITUDE), alpha=.25, color="orange",
             size= (data$avg_tuition-20000)/10000, na.rm = FALSE)+
     theme(plot.title = element_text(hjust = 0.5, face='bold'),
        plot.subtitle = element_text(hjust = 0.5, face='italic'))+
  labs(title = "National Debt Distribution of Students that Withdrew",
       subtitle= "Where each point is an institution with annual tuition over $20k")
```



Summary

Before the graph, we assumed that students who were paying out-of-state tuition would have a greater amount of debt and that more "popular" states like California, Florida, and New York would have both a higher average tuition and student debt, but only a few of those were shown true in our maps. Our maps revealed that the vast majority of schools with a tuition over the national average of $20,000 were concentrated in North Eastern United States. Outside of this main cluster, the states that appeared to have higher tuition did appear to be the more "popular" ones like California, Florida, Texas, and especially New York. As for student debt, the states with the highest student debt for both those who graduated and those who withdrew were located in the Southern United States. We are very pleased with the maps though one thing that was slightly disappointing was that there is data for grey states, though admittedly most of it was NA, so the only way we could find to include it was to remove most of the other data in our graphs. Aside from this, we can conclude that there is a correlation between the location of a school and the amount of debt a student incurs because the only outliers to the trend of high tuition schools being in the North East are the states that are most popular for their tourist attractions or better weather such as California and Texas.

Sub Question: "Is the degree of education a major contributing factor in student debt?"

We wanted to see how well the degree of education available at institutions correlated with student debt. We also wanted to examine how the most common degree at institutions affected debt. We then wanted to see how subdividing the data by the type of educational institution (public, private nonprofit, private for-profit) affected student debt. We took into account four variables for this portion of our analysis on student debt: highest degree offered at an institution, most common degree offered at an institution, the type of school (public, private nonprofit, private for-profit), and median student debt.

Before doing our analysis we thought that as the level of highest degree and most common degree increased, then student debt would also go up. The more education offered, the more debt. In regards to the type of school, we thought that either public school would be cheaper because they receive state funding, or that there would be little difference between the types of schools.

HIGHDEG = Highest degree
PREDDEG = Most Common Degree
CONTROL = Type of institution. (public, private nonprofit, private for-profit)
DEBT_MDN = Median debt

This portion of code loads the college scorecard into R. We then create two tables. One stores data on the highest degree while one stores data on the most common degree.

```
college <- read_csv("Most-Recent-Cohorts-All-Data-Elements.csv", na = c("NA", "NULL"))

# Stores HIGHDEG, CONTROL, AND DEBT_MDN variables from college
H_deg<- college %>%
        select(HIGHDEG,CONTROL,DEBT_MDN) %>%
        mutate(DEBT_MDN = as.numeric(DEBT_MDN)) %>%
        drop_na()

# Stores PREDDEG, CONTROL, AND DEBT_MDN variables from college
MC_deg<- college %>%
        select(PREDDEG,CONTROL,DEBT_MDN) %>%
        mutate(DEBT_MDN = as.numeric(DEBT_MDN)) %>%
        drop_na()
```

This Section builds and refines the data down into usable tibbles

```
# Stores overall median debt by highest degree
Avg_H_deg_row <- H_deg %>%
              group_by(HIGHDEG) %>%
              summarize(Avg_Debt = mean(DEBT_MDN)) %>%
              mutate(CONTROL = 4)

# Stores median debt by highest degree for each type of school
Avg_Debt_H_deg <- H_deg %>%
              group_by(HIGHDEG,CONTROL) %>%
              summarize(Avg_Debt = mean(DEBT_MDN))

# Merges values into final tibble for median debt based on highest degree awarded
Avg_Debt_H_deg <- merge(Avg_Debt_H_deg, Avg_H_deg_row, all = TRUE)

# Stores median debt by most common degree
Avg_MC_deg_row <- MC_deg %>%
              group_by(PREDDEG) %>%
              summarize(Avg_Debt = mean(DEBT_MDN)) %>%
              mutate(CONTROL = 4)

# Stores median debt by most common degree for each school
Avg_Debt_MC_deg <- MC_deg %>%
              group_by(PREDDEG,CONTROL) %>%
              summarize(Avg_Debt = mean(DEBT_MDN))

# Merges values to final table for median debt based on most common degree awarded
Avg_Debt_MC_deg <- merge(Avg_Debt_MC_deg, Avg_MC_deg_row, all = TRUE)
```

We decided a bar graph was the most compelling tool to visualize this kind of information since it was easy to view and break down each data point. It was also easier to see comparative amounts of debt between institutions and degree levels. We made two graphs with one plotting Highest Degree against median student debt and another plotting Most Common Degree against median student debt.

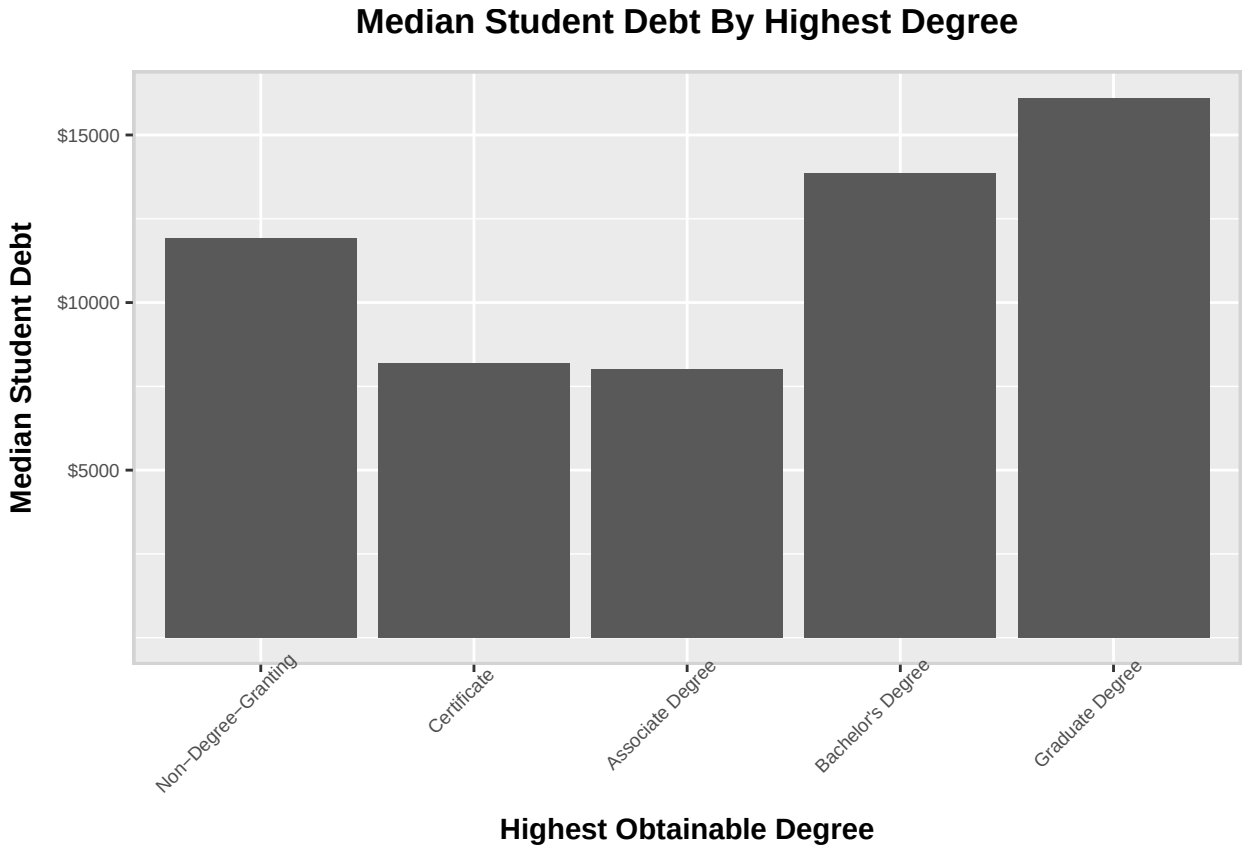
Function that builds bar graph

```
Debt_graph <- function(a,b,c,d = NULL,e = FALSE,f,g) {

  if(e == FALSE)
    {
  ggplot(data = a,aes(x = as.factor(b),
                      y = c)) +
         geom_bar(position="dodge", stat="identity") +
         theme(axis.text.x = element_text(angle=45, hjust=0.8),
               panel.border = element_rect(color = "light gray",
                                           fill = NA,
                                           size = 1)) +
         labs(x = g,
              y = "Median Student Debt",
              title = f) +
         scale_x_discrete(labels = c("0" = "Non-Degree-Granting",
                                     "1" = "Certificate",
                                     "2" ="Associate Degree",
                                     "3" ="Bachelor's Degree",
                                     "4" ="Graduate Degree")) +
         scale_y_continuous(breaks = c(5000,10000,15000),
                            labels = c("$5000","$10000","$15000")) +
         scale_fill_continuous(labels = c("Public", "Private Nonprofit",
                                          "Private For-profit",
                                          "All")) +
           theme( axis.text=element_text(size=7),
                axis.title=element_text(size=11,vjust = 0.5,face="bold"),
                axis.title.y=element_text(vjust = 3),
                axis.title.x=element_text(vjust = 3),
                plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
  }

  else if(e == TRUE)
    {
  ggplot(data = a,aes(x = as.factor(b),
                      y = c,
                      fill = as.factor(d))) +
         geom_bar(position="dodge", stat="identity") +
         theme(axis.text.x = element_text(angle=45, hjust=0.8),
               panel.border = element_rect(color = "light gray",
                                           fill = NA,
                                           size = 1)) +
         labs(x = g,
              y = "Median Student Debt",
              fill = "School Type",
              title = f) +
         scale_x_discrete(labels = c("0" = "Non-Degree-Granting",
                                     "1" = "Certificate",
                                     "2" ="Associate Degree",
                                     "3" ="Bachelor's Degree",
                                     "4" ="Graduate Degree")) +
         scale_y_continuous(breaks = c(5000,10000,15000),
                            labels = c("$5000","$10000","$15000")) +
```

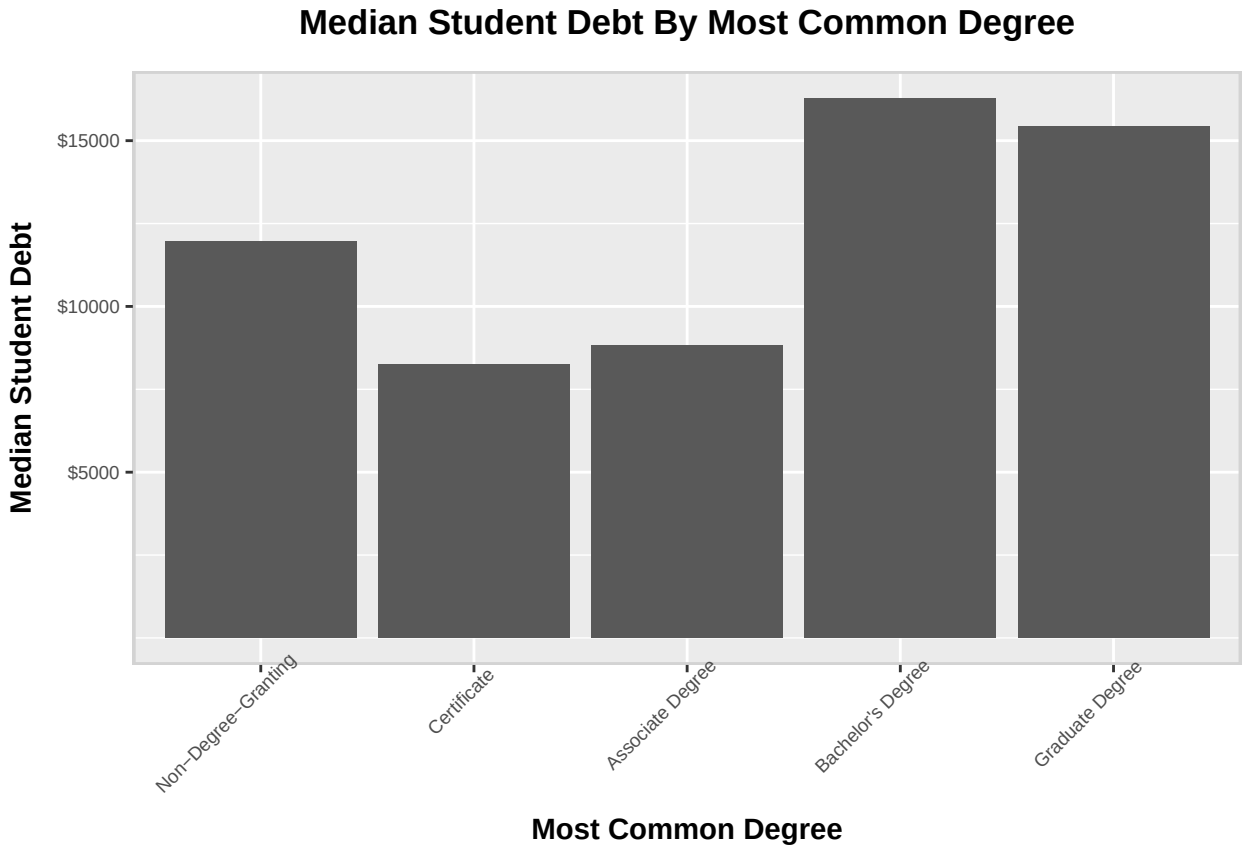
```
        scale_fill_discrete(labels = c("Public", "Private Nonprofit",
                                       "Private For-profit",
                                       "All")) +
           theme( axis.text=element_text(size=7),
                axis.title=element_text(size=11,vjust = 0.5,face="bold"),
                axis.title.y=element_text(vjust = 3),
                axis.title.x=element_text(vjust = 3),
                plot.title=element_text(size=13,hjust = 0.5,
                                       vjust = 3, face="bold"),
                legend.text=element_text(size=7),
                legend.title=element_text(size=10,face="bold"))
  }
}
```

At first we only included the overall median debts instead of subdividing it by type of school. We later felt adding the school type factor would make for a better and more informative visual.

```
Debt_graph(Avg_H_deg_row,Avg_H_deg_row$HIGHDEG,Avg_H_deg_row$Avg_Debt,
           "ALL",FALSE,"Median Student Debt By Highest Degree",
           "Highest Obtainable Degree")
```

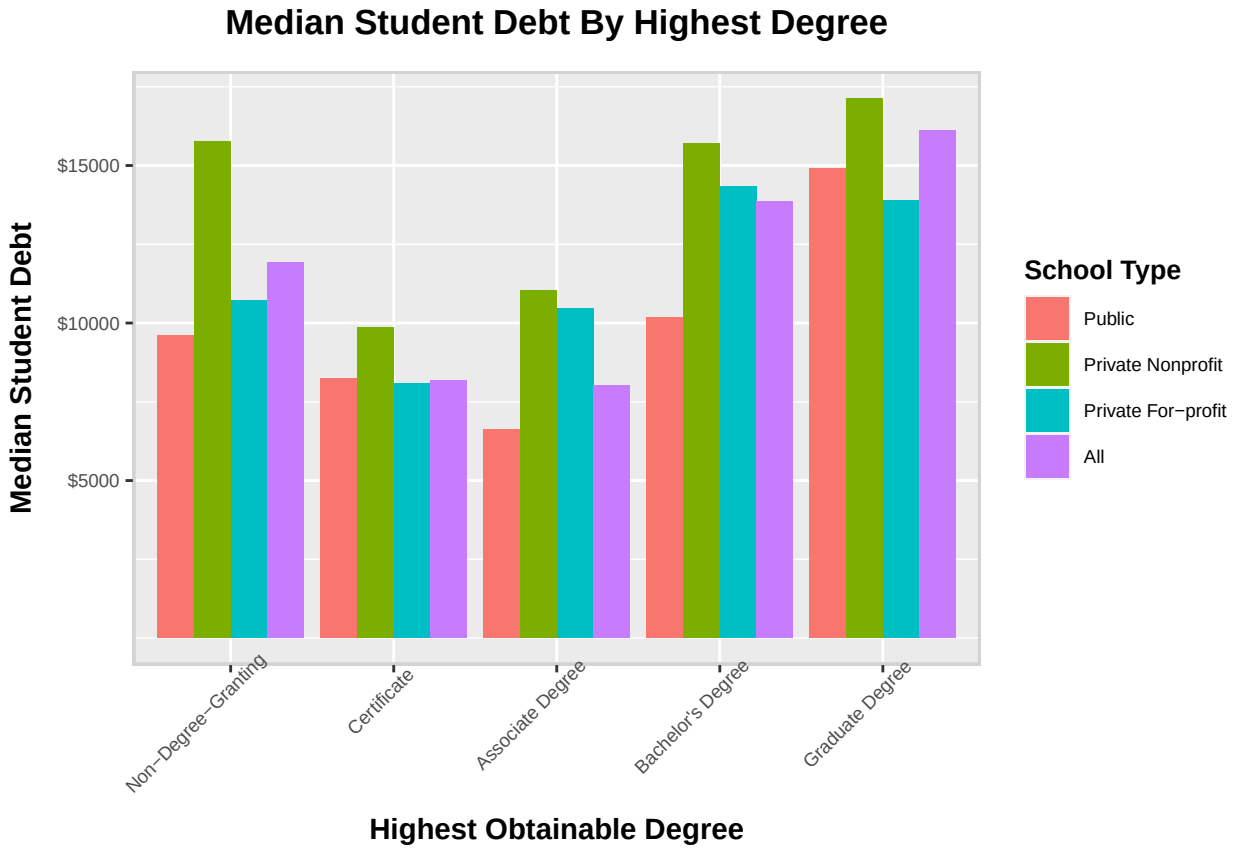


```
Debt_graph(Avg_MC_deg_row,Avg_MC_deg_row$PREDDEG,Avg_MC_deg_row$Avg_Debt,
           "ALL",FALSE,"Median Student Debt By Most Common Degree",
           "Most Common Degree")
```



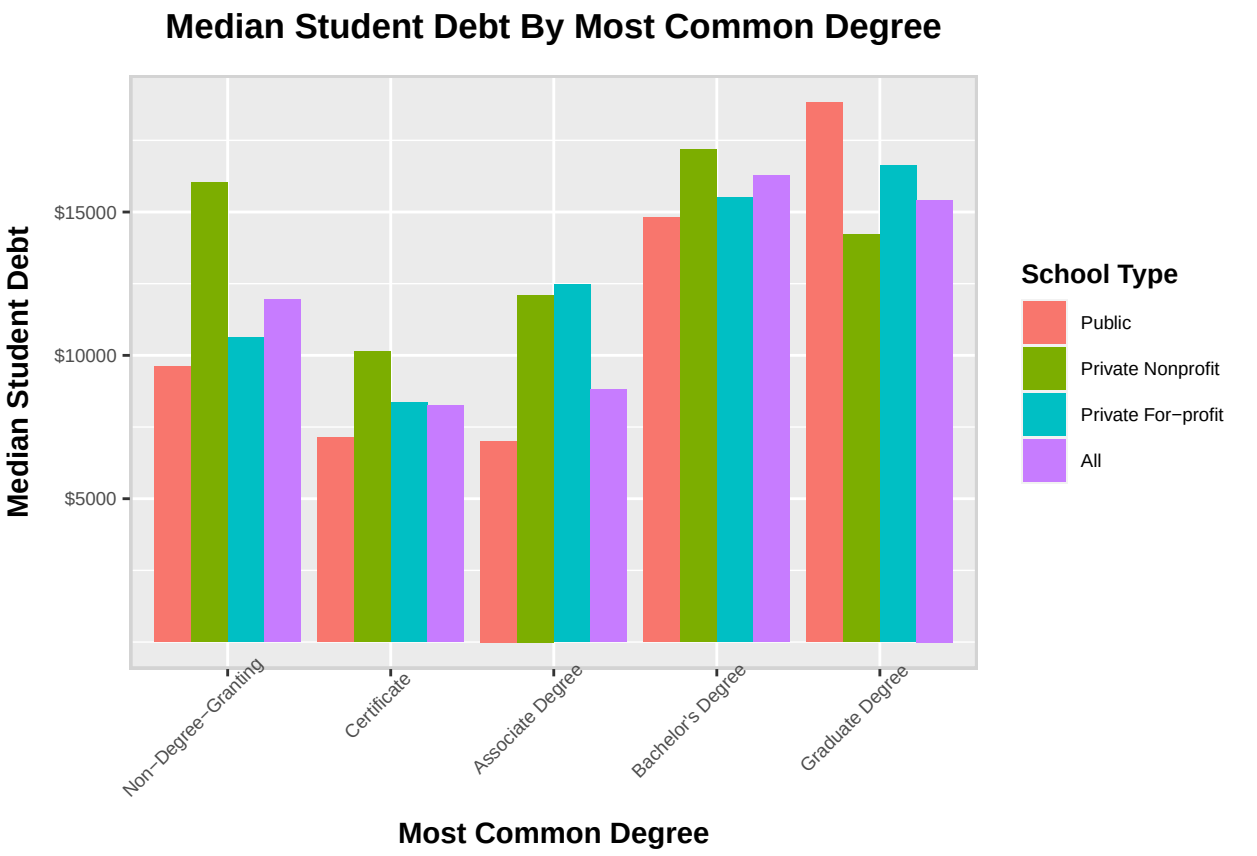
This graph visualizes the median debt based on highest obtainable degree

```
# Makes a bar graph for median student debt based on highest degree
Debt_graph(Avg_Debt_H_deg,Avg_Debt_H_deg$HIGHDEG,Avg_Debt_H_deg$Avg_Debt,
           Avg_Debt_H_deg$CONTROL,TRUE,
           "Median Student Debt By Highest Degree",
           "Highest Obtainable Degree")
```



This graph visualizes the median debt based on most common degree

```
# Makes a bar graph for median student debt based on most common degree
Debt_graph(Avg_Debt_MC_deg,Avg_Debt_MC_deg$PREDDEG,Avg_Debt_MC_deg$Avg_Debt,
           Avg_Debt_MC_deg$CONTROL,TRUE,
           "Median Student Debt By Most Common Degree",
           "Most Common Degree")
```



Summary

In general, student debt increased as levels of Highest and Most Common degree increased. One exception was non-degree granting school which were a special case as they do not match up to traditional degree offering schools. They generally fit in between associate and bachelor degree schools in terms of debt. We also found that public universities almost always had lower student debt than private universities when their Most Common degree offered was certificates, associate degrees, or bachelor degrees. This was the same for when they were the Highest Degrees offered. When it came to the graduate level, public institutions were either about the same or more expensive. Another exception was Graduate schools grouped by Most Common Degree had slightly lower overall student debt than Bachelor degree schools. Another note is that certificates and associate degrees have near identical debt levels. While certificates usually have shorter programs, there are more scholarships and financial support programs for associate degree seekers. The general explanation for our findings is that as a university offers more and higher levels of education then the average student debt incurred at these institutions will be greater. Public institutions are likely cheaper than private institutions below the graduate level because they receive more state funding that lower the costs of attendance. This also means they can hand out more scholarships and lower the overall cost burden on students. These factors are minimized as more debt is shifted back to the student at the graduate level.

Question of Interest: "Does the type of institution attended affect later earnings?"

After examining the debt accrued from attending university, the next logical step was to examine earnings after college. Data was specifically chosen in the analysis of earnings so as to tie in with the previous question. For instance, as those who attend but do not graduate still incur debt, this process included examining the earnings of individuals ten years after initial enrollment; regardless of their graduation status. Overall, we explored which proportion of attendees at various institution types later entered which income bracket.

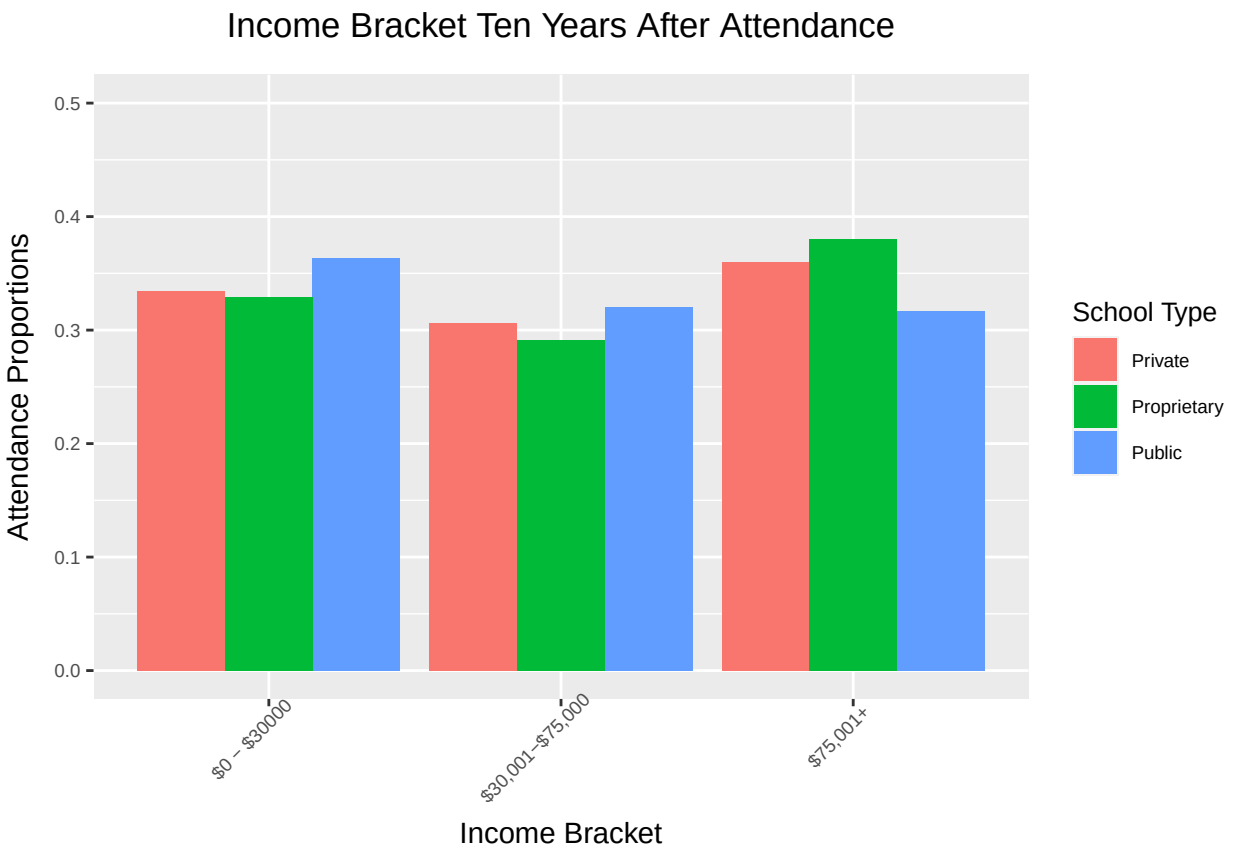
First, a select few variables were chosen to work with from the large "College Scorecard" data set. The variables that would best reflect a promising career after enrolling in higher education were, the mean earnings of students working and not enrolled ten years after entry in each of the three tercile earning brackets, and the control of the institution. These were fruitful choices, as some degrees such as post-graduation degrees take longer to earn. By examining earnings ten years post-enrollment was a precaution taken to avoid bias towards high-earning bachelor's degrees such as those in computer science and engineering. Further, the ten year lag allows those with less scholastic investment on their resume to earn promotions and raises in their field, and thereby reflect the effect of their disproportionate work experience compared to those who have an extensive education background. This had the intention of leveling the playing field in terms of comparing a mechanic or real estate agent with only 2 years of schooling, but 8 years of experience against an individual with a doctorate degree with only a couple of years of work experience. The school type variable was easy to work with directly from the data set as it labeled each institution as "private", "public", or "proprietary".

Next, we generated a statistic from the four variables chosen. The intention was to compare the income brackets across school types; however, a statistic was needed to create the y-axis. Ultimately, a simple attendance proportion was used as the number of people enrolled in public school was obscuring interesting results. The proportion was calculated for the type of institution enrolled across each of the three income bracket terciles.

```
# read data in
college_earnings <- read_csv("Most-Recent-Cohorts-All-Data-Elements.csv",
                             na = c("", "NA", "NULL",
                                    "PrivacySurpressed")) %>%
  select(SCHTYPE, MN_EARN_WNE_INC1_P10, MN_EARN_WNE_INC2_P10, MN_EARN_WNE_INC3_P10)
# make data nice
college_earnings$MN_EARN_WNE_INC1_P10 <- as.numeric(college_earnings$MN_EARN_WNE_INC1_P10)
college_earnings$MN_EARN_WNE_INC2_P10 <- as.numeric(college_earnings$MN_EARN_WNE_INC2_P10)
college_earnings$MN_EARN_WNE_INC3_P10 <- as.numeric(college_earnings$MN_EARN_WNE_INC3_P10)


# number of students in each school type
sum_pri <- college_earnings %>%
  pivot_longer(-SCHTYPE,
               names_to = "Income_Bracket",
               values_to = "Number_of_data_points") %>%
  filter(SCHTYPE == 1) %>%
  select(Number_of_data_points) %>%
  sum(na.rm = T)

sum_pub <- college_earnings %>%
  pivot_longer(-SCHTYPE,
               names_to = "Income_Bracket",
               values_to = "Number_of_data_points") %>%
  filter(SCHTYPE == 2) %>%
  select(Number_of_data_points) %>%
  sum(na.rm = T)

sum_prop <- college_earnings %>%
```

```
  pivot_longer(-SCHTYPE,
               names_to = "Income_Bracket",
               values_to = "Number_of_data_points") %>%
  filter(SCHTYPE == 3) %>%
  select(Number_of_data_points) %>%
  sum(na.rm = T)


# number of private school attendees in each of the income brackets
sum_inc_pri <- college_earnings %>%
  pivot_longer(-SCHTYPE,
               names_to = "Income_Bracket",
               values_to = "Number_of_data_points") %>%
  filter(SCHTYPE == 1) %>%
  group_by(Income_Bracket) %>%
  summarize(sum(Number_of_data_points, na.rm = T))
# proportion of private school attendees in each of the income brackets
pri_prop <- sum_inc_pri$`sum(Number_of_data_points, na.rm = T)`/sum_pri
# number of private school attendees in each of the income brackets
sum_inc_pub <- college_earnings %>%
  pivot_longer(-SCHTYPE,
               names_to = "Income_Bracket",
               values_to = "Number_of_data_points") %>%
  filter(SCHTYPE == 2) %>%
  group_by(Income_Bracket) %>%
  summarize(sum(Number_of_data_points, na.rm = T))
# proportion of private school attendees in each of the income brackets
pub_prop <- sum_inc_pub$`sum(Number_of_data_points, na.rm = T)`/sum_pub
# number of private school attendees in each of the income brackets
sum_inc_prop <- college_earnings %>%
  pivot_longer(-SCHTYPE,
               names_to = "Income_Bracket",
               values_to = "Number_of_data_points") %>%
  filter(SCHTYPE == 3) %>%
  group_by(Income_Bracket) %>%
  summarize(sum(Number_of_data_points, na.rm = T))
# proportion of private school attendees in each of the income brackets
prop_prop <- sum_inc_prop$`sum(Number_of_data_points, na.rm = T)`/sum_prop


# create a tibble
inc_brackets_proportions <- as_tibble(pri_prop) %>%
  mutate(pub_prop) %>%
  mutate(prop_prop) %>%
  mutate( c("top", "middle", "bottom")) %>%
  pivot_longer(-`c("top", "middle", "bottom")`,
               names_to = "School Type",
               values_to = "Attendance Proportions")
# make the column names nicer
inc_brackets_proportions <- inc_brackets_proportions %>%
  mutate(`School Type` = replace(`School Type`, `School Type` == "value", "Private")) %>%
  mutate(`School Type` = replace(`School Type`, `School Type` == "pub_prop", "Public")) %>%
  mutate(`School Type` = replace(`School Type`, `School Type` == "prop_prop", "Proprietary"))
```

```r
# plot creation
ggplot(data = inc_brackets_proportions ,aes(x = as.factor(`c("top", "middle", "bottom")`),
                                            y = `Attendance Proportions`,
                                            fill=as.factor(`School Type`))) +
  geom_bar(position="dodge", stat="identity") +
  theme(axis.text.x = element_text(angle=45, hjust=0.8)) +
  labs(x = "Income Bracket",
       y = "Attendance Proportions",
       fill = "School Type",
       title = "Income Bracket Ten Years After Attendance") +
  scale_x_discrete(labels = c("top" = "$75,001+ ",
                              "middle" = "$30,001-$75,000",
                              "bottom" ="$0 - $30000")) +
  scale_y_continuous(breaks = seq(from = 0, to = 1, by = 0.1),
                     limits = c(min = 0, max = .5)) +
  theme( axis.text=element_text(size=7),
         axis.title=element_text(size=11),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 3),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3),
         legend.text=element_text(size=7),
         legend.title=element_text(size=10))
```



Summary

Our comparison was made across income brackets in order to elucidate whether the institution type enrolled in affected earnings. Does rubbing elbows with Ivy League attendees disproportionately affect earnings later in life? From the plot, it is easy to see that the intuition type has little effect on earnings averaged over a large population. Each income bracket is composed almost equally of students from all three institution types. In fact, the largest disparity within any given income bracket was only 6.3% by proportion of attendees of propriety schools versus public school attendees. Proprietary school attendees interestingly had the highest proportion of former students in the highest income bracket. This may be explained with school such as aviation mechanics that have short educational programs. Attendees are able to begin working their way into a competitive position sooner than those with ambitions of a 4-year degree or beyond. Overall, it seems as though institution type attended is a poor predictor of later earnings.

Conclusion

We found that location correlated with student debt and tuition costs and that the level of education affected student debt. We also found that while public institutions were cheaper, the type of institution did not greatly affect earnings 10 years after college. Therefore, it is probably economically most efficient to go to a 2-4 year public university. It is important to note that two different "type of institution" variables were used, but that the data still stands. You could further argue it is best to stay away from the southeast because you are more likely to end up with more debt there and maybe the northeast where there are a lot of universities with high tuition costs.