A Report of College Students Economic Situations

2024-12-10

Contents

1 Introduction

The education situation of college students, their consumption during school and the salary level of graduates have always been issues that have attracted widespread social attention. In this report, I analyze the cost and debt situation of college students during their schooling, as well as the year-by-year trend of graduates' salary changes and the correlation between them. I first analyzed the year-to-year trends and correlations between students' family income and school consumption, and used year-to-year Consumer Price Index (CPI) data to exclude the impact of inflation. Then I explore the main reasons for student loan defaults, use Cohort Default Rate (CDR) as an indicator of the situation, explore the correlation between defaults and graduates' salary, family income and debt situation. Based on the results of the analysis, I give some corresponding suggestions from the perspective of policy making.

2 Data Overview

Education data is a set of biennial data recorded with student debts from major US universities from 2010 to 2020 (Dickson et al., 2023).

Cost data provides information on the net out-of-pocket costs that families pay for the university the student enrolled in (Radwin & Wei, 2015).

CPI data from the Federal Reserve of Minneapolis between 2000 and 2023. It shows CPI in each year and the annual rate of change.

Graduates income data is a set of data on the average salary of graduates from major US universities from 2010 to 2022. Published by Statista Research Department

The four data chosen for this project are all in a manipulable csv format. Importantly, the data are all open-source, available from papers or open-source data sites, and the data sources and references have been signed in this project. In summary, this data satisfies the FAIR and/or CARE principles.

3 Debt and Cost Analysis

3.1 Data Description and Preprocessing

Data Description

In this section, I use three sets of data, education data, cost data, and CPI data. I will heavily rely on the five numeric variables in education data. They are the median debt of low income families (with income less than $30,000), median debt of median income family (with income between $30,001 and $75,000), median debt of high income family (with income more than $75,001), three-year cohort default rate, and average family income. Cost data provides information on the net out-of-pocket costs that families pay for the university the student enrolled in. It includes mean costs of low, median, and high income families grouped by public and private institutions. Moreover, CPI data comes from Federal Reserve of Minneapolis. It shows CPI in each year and the annual rate of change.

Preprocessing

For the purpose of data analysis, I first conduct the data cleaning on education data and cost data. Variables are renamed to convey more accurate meanings. All states are transferred into lower case so that they can match in the subsequent data consolidation. In the education data, an important issue that needs to be alert is that there is quite a proportion of missing values. Particularly, I found that the default rate is missing for all observations in 2010 and the values of average family income are missing in 2018 and 2020. For the default rate data loss throughout the 2010, I applied random sampling to fill the data based on the parallel values in other years. For that of average family income, I also applied random sampling, but this time

multiplied by a CPI annual rate factor to correct the deviation. For a small amount of discrete missing data, I only interpolate based on the previous row of data. For the purpose of comparison, all dollars are transformed to the real 2018 dollar so that the impact of inflation can be reduced to minimum. An easy approach to achieve this is using the formula:

$$P_{2018} = P_n \times \left(\frac{CPI_{2018}}{CPI_n} \right)$$

where P_n implies the actual price in that year and P_{2018} is the real 2018 dollar. In the cost data, there is no situation where the overall year data is missing. So, I interpolate the missing values within each year by multiple interpolation. At last, the education data and the cost data are left joined, which means that the values in education data are kept.

3.2 Data and Cost Analysis

Median debt is used to represent the overall of debt. As the Figure 1 showed that the overall trend of student debt by median debt. It shows that the debt reaches the minimum in 2012 and the peak in 2016. The following years follow a decreasing trend.

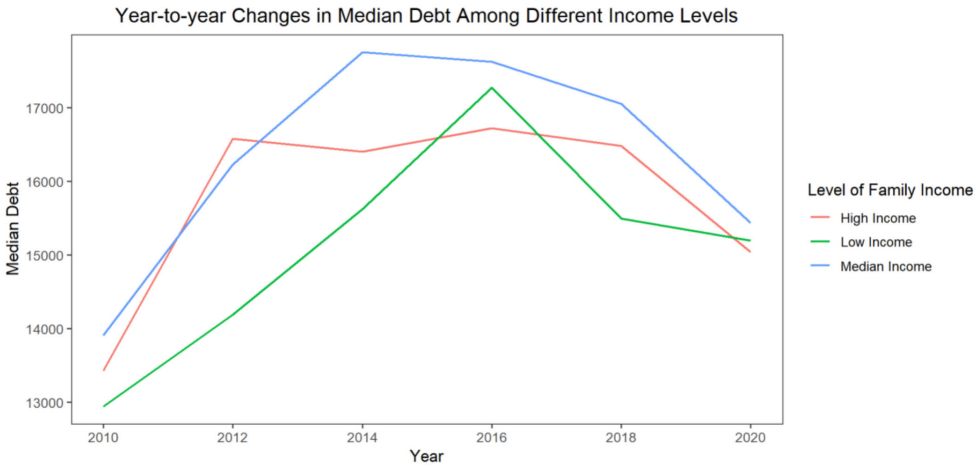

Figure 1. Year-to-year Changes in Median Debt among Different Income Levels

In Figure 2, I divide this overall trend into two parts concerning the institution type. The changing patterns are similar to the overall trend but we can observe that debts in private schools is clearly more than that in public schools.

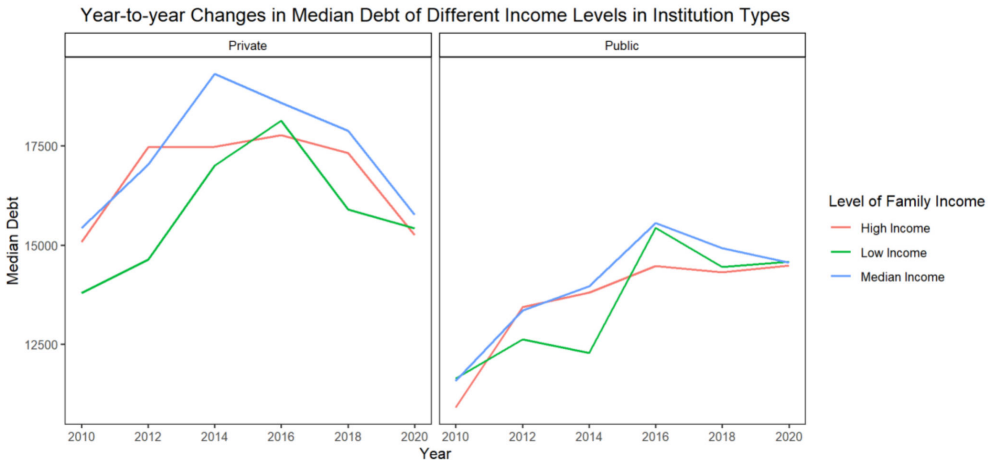

Figure 2. Year-to-year Changes in Median Debt of Different Income Levels in Institution Types

In Figure 3, I compare the situation between my state (New York) and other states. Overall, the median debt in my state is higher than the same level in other states. Comparing my state to other states, average private debt is higher than public debt across incomes. Similarly, under the same conditions, the trends in income and debt of all households in private institutions in our state and other states are similar to Figure 1, and the average loans of private institutions in our state reached the highest value in 2016. The average loan value of our state's public institutions reached its lowest level in 2014.

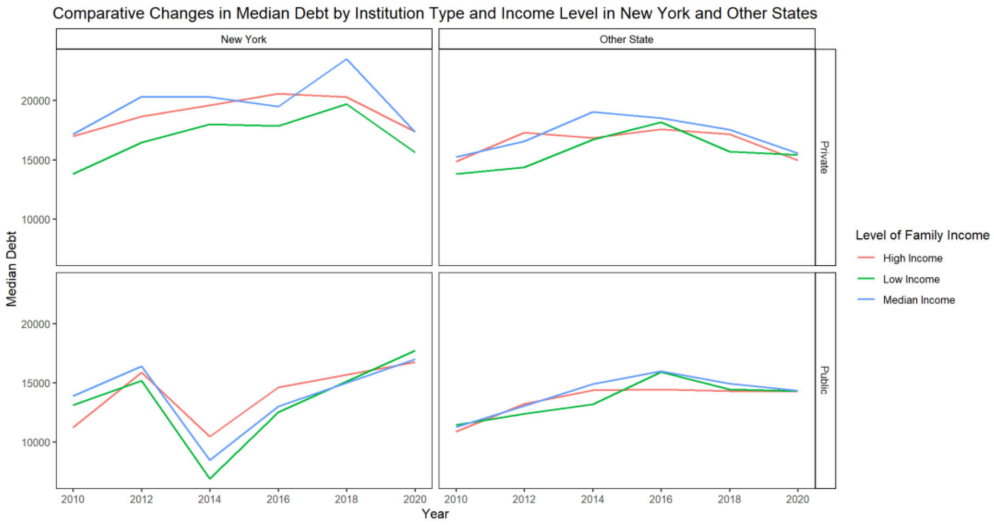

Figure 3. Comparative Changes in Mean Debt by Institution Type and Income Level in New York and Other States

To find the relationship between debts and costs, I established the Figure 4. It displays the costs and debts for different income families by years. One noteworthy feature is that high-income families spend significantly more than debt, while other income families have no such significant difference. High income families also have a larger discrepancy in trends of costs and debts. In addition, after 2016, the costs for high-income families in Illinois have increased significantly.

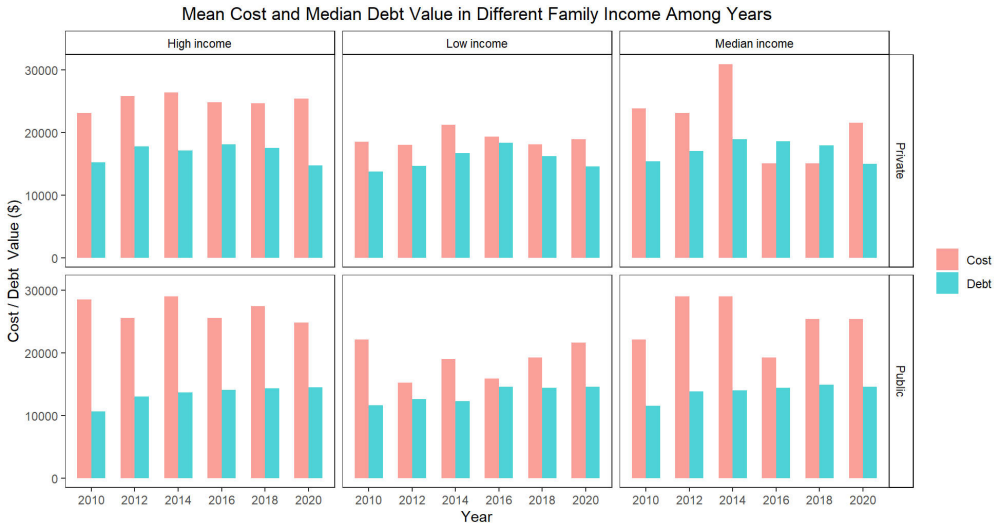

Figure 4. Mean Cost and Median Debt Value in Different Family Income among Years

In Figure 5, I analyze the correlation between debt and default rates. To better observe the relationship between debt and rate on the same graph across year, I multiply rate by a coefficient. The same trend is observed in both private and public institutions. Even though the debt amount of each family is increasing year by year, the average default rate is decreasing year by year, indicating that the family income is able to repay the debt.

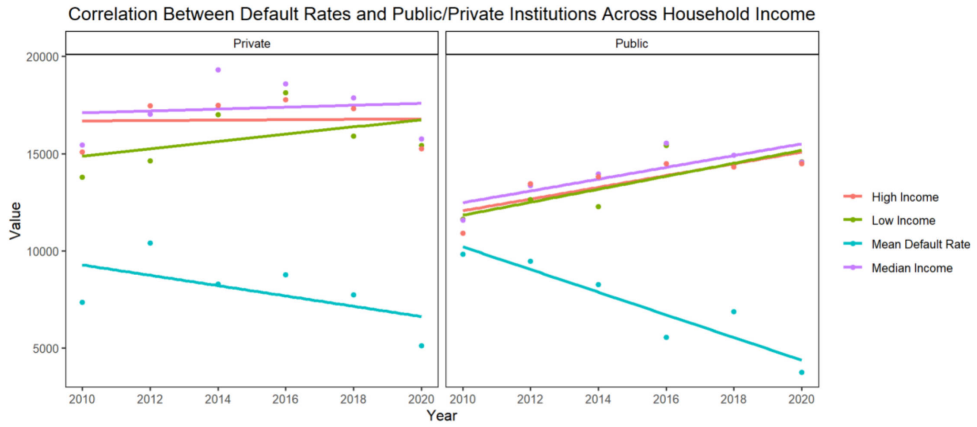

Figure 5. Correlation of Default Rates and Public/Private Institutions cross Family Income

In my opinion, the education loans in my state are significantly higher than those in other states. It is recommended to appropriately reduce students' consumption in school, such as lowering tuition fees and other measures. In addition, for low-income families, appropriate living subsidies can be provided to students based on the school's consumption.

4 Debt, Cost, and Earnings after Graduation Analysis

4.1 Data Description and Preprocessing

Data Description

Since the data of debt and cost have been cleaned in above chapter, I only need to focus on the graduation income data. The transformation of previous data into Real 2018 Dollar provides convenience for the comparison to the graduate income because the values in graduate income are all in Real 2018 dollar. The variables in graduation income data includes nine categories. The first type is the median earnings of students working and not enrolled 6 years after entry in the low income tercile $0 - $30,000 for each university. The other eight types change in the working years to 8 years and 10 years, and at the same time considering the other two levels of income families. However, in my analysis, I only use the median income data, as a representative index for the graduate incomes.

Preprocessing

The data cleaning includes renaming, transforming variables to numeric and joining all tables by school id as what I did in the previous section. I also used random interpolation to replace the missing values. I use this method because the data is not related to the time series. After finishing all preparations, I decide to look into the relationships between education cost and graduate incomes, between debt and graduate incomes, and between default rate and graduate incomes. These relationships are between our key factors and the graduate incomes. I expect the outcome to be enlightening towards the recommendations on policies.

4.2 Cost, and Earnings Analysis

In Figure 6, I analyzed the correlation between post-graduation salaries in New York and other states, 6, 8, and 10 years after graduation, and students' college cost. Low-income graduates from other states exhibit a clear positive correlation, while other groups lack consistent significance. As Figure 7 shows, I analyze the correlation between wage levels and loan conditions in our state and other states, using the same grouping method as Figure 6. Similarly, the results are consistent with Figure 6. Again, low-income graduates in other states show a positive correlation with debt, while New York only shows this among high-income graduates. Surprisingly, in the last group, the correlation between wages and family income is positive for both New York and other states. Despite this, Figures 6-8 consistently demonstrate that longer-graduated individuals earn higher wages.

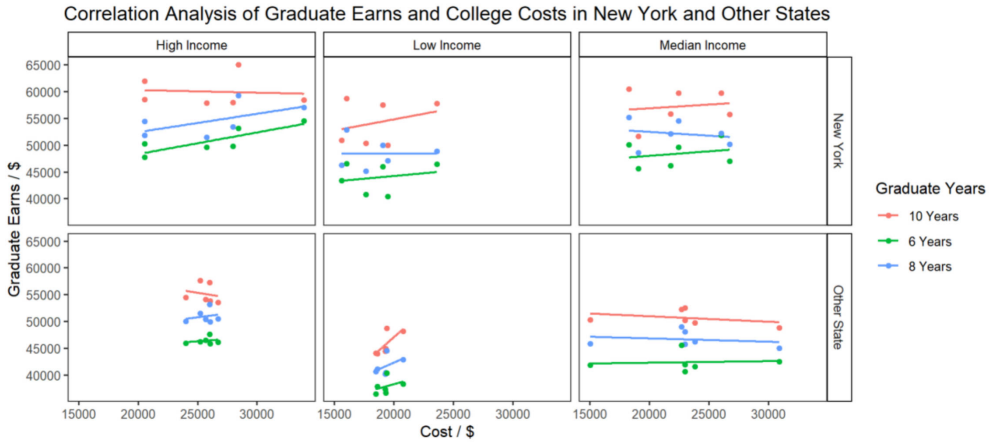

Figure 6. Correlation Analysis of Graduate Earns and College Costs in New York and Other States

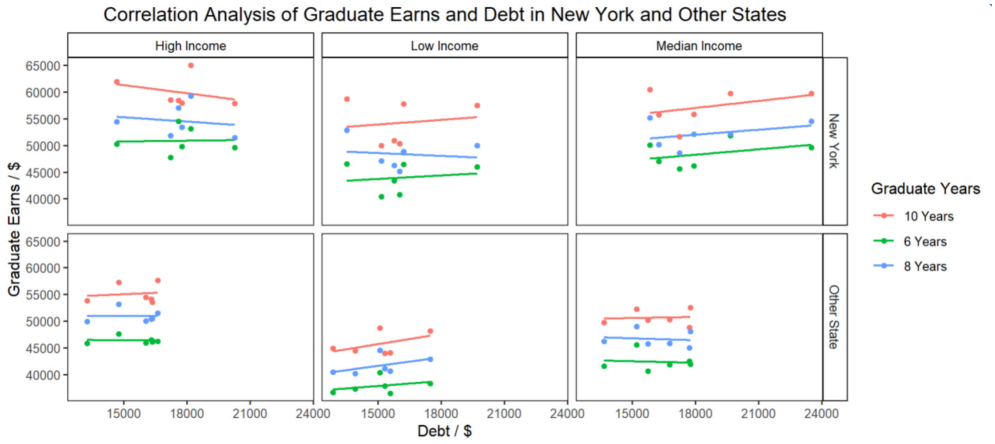

Figure 7. Correlation Analysis of Graduate Earns and Debt in New York and Other States

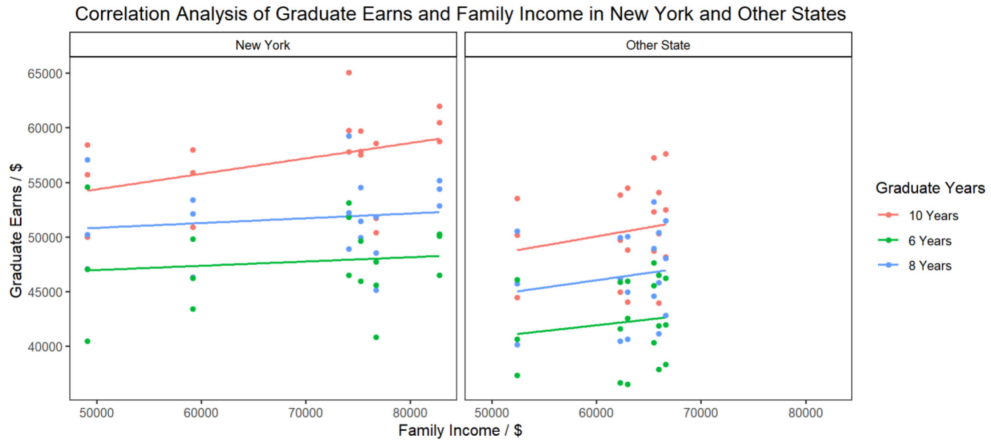

Figure 8. Correlation Analysis of Graduate Earns and Family Income in New York and Other States

Overall, New York's graduates exhibit a positive correlation between salary and family income, with no correlation to education costs or debt among low-income graduates. This may exacerbate social inequality. Therefore, I recommend introducing policies to uplift the salaries of low-income graduates in New York, so that students' cost and debt between universities are more meaningful, and more people are willing to accept a university education.

5 Conclusion

This report discussed the relationship and trends of education cost, debt, and graduate incomes. The overall situation is positive because we saw that the default rate was decreasing and the graduate incomes are permanently increasing. However, we also witness students from low- and median- income families are still suffering from their family financial conditions when enrolling a university. So, I provided the recommendation to relax the debt policies for students in need. we also noticed that the risk of debt is reducing over time. So, I believe students with worse financial conditions should be encouraged to attend to universities and the government should have enough confidence for them to repay the debts.

6 References

Dickson, T., Mulligan, E. P., & Hegedus, E. J. (2023). Impacts of educational debt on physical therapist Employment Trends. *BMC Medical Education*, 23(1). https://doi.org/10.1186/s12909-023-04454-3

Radwin, D., & Wei, C. C. (2015). What is the price of college? Total, Net, and Out-of-Pocket prices by type of institution in 2011-12. NCES 2015-165. Stats in brief. National Center for Education Statistics. http://files.eric.ed.gov/fulltext/ED555646.pdf

Federal Reserve Bank of Minneapolis. Consumer price index: 1913–present. Retrieved December 10, 2024, from https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-

U.S. average wages of college graduates 1990-2022.Published by Statista Research Department, Jul 5, 2024. https://www.statista.com/statistics/642041/average-wages-of-us-college-graduates/

Dynarski, S., Page, L., & Scott-Clayton, J. (2022). College costs, financial aid, and student decisions. https://doi.org/10.3386/w30275

7 Code Appendix

```
knitr::include_graphics("F1.jpg")
knitr::include_graphics("F2.jpg")
knitr::include_graphics("F3.jpg")
knitr::include_graphics("F4.jpg")
knitr::include_graphics("F5.jpg")
knitr::include_graphics("F6.jpg")
knitr::include_graphics("F7.jpg")
knitr::include_graphics("F8.jpg")
rm(list=ls()) # clear the environment
setwd("D:/02Work/20241210_final_r")

#-------Import necessary packages here--------------------#
library(tidyverse)
library(janitor)
library(mice)
library(ggthemes)
library(Hmisc)


#------ Debt and Cost Analysis --------------------------#

##Cleaning Education data
# Read CPI data and select only the first two columns
cpi_data <- read_csv("CPI_data.csv") %>% select(1,2)

# Read education data, rename columns for clarity,
# convert school names to lowercase, and convert columns 7 to 11 to numeric
education_data <- read_csv("education_data.csv") %>%
  rename(year = 'YEAR',
         school_id = 'UNITID',
         school_name = 'INSTNM',
         state_id = 'STABBR',
         predominant_degree = 'PREDDEG',
         institution_type = 'CONTROL',
         median_debt_low_income = 'LO_INC_DEBT_MDN',
         median_debt_med_income = 'MD_INC_DEBT_MDN',
         median_debt_high_income = 'HI_INC_DEBT_MDN',
         default_rate = 'CDR3',
         avg_family_income = 'FAMINC') %>%
  mutate(school_name = tolower(school_name)) %>%
  mutate(across(.cols=7:11, .fns=as.numeric)) #

# Convert numeric institution_type to descriptive categories:
# 1 = public, otherwise = private
education_data_clean <- education_data %>%
  mutate(institution_type = ifelse(institution_type == 1, "public", "private"))

# Filter data to include only institutions awarding bachelor's degrees
education_data_BA1 <- education_data_clean %>%
  filter(predominant_degree == 3)
```

```r
# Merge with CPI data, calculate real debts and family income adjusted by CPI,
# and remove unnecessary columns (7-9, 11-12)
education_data_BA <- education_data_BA1 %>%
  left_join(cpi_data) %>%
  mutate(real_debt_low_income = median_debt_low_income * 251.1/CPI,
         real_debt_med_income = median_debt_med_income * 251.1/CPI,
         real_debt_high_income = median_debt_high_income * 251.1/CPI,
         real_family_income = avg_family_income * 251.1/CPI) %>%
  select(-7:-9,-11:-12)

# Impute missing values for each year using random forest method
add_education_data_BA <- education_data_BA %>%
  group_by(year) %>%
  mice(method = "rf", m=5, printFlag = FALSE)

# Complete the imputation and update the main dataset
education_data_BA <- complete(add_education_data_BA)

## Cleaning Cost Data
# Read cost data, select relevant columns only
cost_data1 <- read_csv("cost_data.csv") %>%
  select(UNITID, INSTNM, YEAR, NPT41_PUB, NPT43_PUB, NPT45_PUB, NPT41_PRIV, NPT43_PRIV, NPT45_PRIV)

# Rename columns, convert school names to lowercase, and ensure columns 4 to 9 are numeric
cost_data2 <- cost_data1 %>%
  rename(year = 'YEAR',
         school_id = 'UNITID',
         school_name = 'INSTNM',
         mean_cost_low_income_public = 'NPT41_PUB',
         mean_cost_med_income_public = 'NPT43_PUB',
         mean_cost_high_income_public = 'NPT45_PUB',
         mean_cost_low_income_private = 'NPT41_PRIV',
         mean_cost_med_income_private = 'NPT43_PRIV',
         mean_cost_high_income_private = 'NPT45_PRIV') %>%
  mutate(school_name = tolower(school_name)) %>%
  mutate(across(.cols=4:9, .fns=as.numeric)) #

# 2-3 If public data is missing, use private data for that income category.Then remove the original cos
cost_data3 <- cost_data2 %>%
  mutate(mean_cost_low_income = ifelse(is.na(mean_cost_low_income_public),mean_cost_low_income_private,m
         mean_cost_med_income = ifelse(is.na(mean_cost_med_income_public),mean_cost_med_income_private,m
         mean_cost_high_income = ifelse(is.na(mean_cost_high_income_public),mean_cost_high_income_privat
  select(-4:-9) #

# Merge with CPI data and calculate real costs adjusted by CPI
cost_data4 <- cost_data3 %>%
  left_join(cpi_data) %>%
  mutate(real_cost_low_income = mean_cost_low_income * 251.1/CPI,
         real_cost_med_income = mean_cost_med_income * 251.1/CPI,
         real_cost_high_income = mean_cost_high_income * 251.1/CPI)

# Select only the necessary columns and use mice to impute missing values by year
cost_data <- cost_data4 %>%
```

```r
  select(-4:-7) #

add_cost_data <- cost_data %>%
  group_by(year) %>%
  mice(method = "rf", m=5,printFlag = FALSE)

cost_data <- complete(add_cost_data)

## Merging debt and cost data
# Merge education and cost datasets by year and school_id,then remove the redundant school_name.y colum
education_data_BA_cost <-  education_data_BA %>%
  left_join(cost_data, by = c("year","school_id")) %>%
  select(-school_name.y)

# Use mice to impute missing values.
add_education_data_BA_cost <- education_data_BA_cost %>%
  mice(method = "rf", m=5,printFlag = FALSE)
education_data_BA_cost <- complete(add_education_data_BA_cost)


# Group the merged data by year and institution_type,and calculate the mean debt and cost for different
debt_cost_sumstat_year <- education_data_BA_cost %>%
  group_by(year, institution_type) %>%
    summarise(
      mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
      mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
      mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
      mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
      mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
      mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T)
    ) %>%
  ungroup()

# Reshape data from wide to long format
debt <- debt_cost_sumstat_year %>%
  select(-c(6:8)) %>%
  pivot_longer(cols = c(3:5), names_to = c("income_category"), values_to = c("debt")) %>%
  mutate(income_category = case_when(
    str_detect(income_category, "low_income") ~ "low income",
    str_detect(income_category, "median_income") ~ "median income",
    str_detect(income_category, "high_income") ~ "high income"))

cost <- debt_cost_sumstat_year %>%
  select(-c(3:5)) %>%
  pivot_longer(cols = c(3:5), names_to = c("income_category"), values_to = c("cost")) %>%
  mutate(income_category = case_when(
    str_detect(income_category, "low_income") ~ "low income",
    str_detect(income_category, "median_income") ~ "median income",
    str_detect(income_category, "high_income") ~ "high income"))

debt_cost_data_by_year <- debt %>%
  inner_join(cost)
```

```r
# Summarize debt and cost statistics by institution type and by year.
debt_sumstat_school_type <- education_data_BA_cost %>%
  group_by(institution_type) %>%
  summarise(
    mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
    mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
    mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
    mean_family_income = mean(real_family_income, na.rm = T)
  ) %>%
  ungroup()


debt_sumstat_year <- education_data_BA_cost %>%
  group_by(year) %>%
  summarise(
    mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
    mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
    mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
    mean_family_income = mean(real_family_income, na.rm = T)
  ) %>%
  ungroup()

cost_sumstat_school_type <- education_data_BA_cost %>%
  group_by(institution_type) %>%
  summarise(
    mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
    mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
    mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T)
  ) %>%
  ungroup()

cost_sumstat_year <- education_data_BA_cost %>%
  group_by(year) %>%
  summarise(
    mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
    mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
    mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T)
  ) %>%
  ungroup()


# Plot: Debt and Cost Analysis

# F1
# Year-to-year Changes in Median Debt Among Different Income Levels
d1 <- debt_sumstat_year %>%
  select(1:4)
colnames(d1) <- c("Year", "Low Income","Median Income","High Income")
d1 <- d1 %>% pivot_longer(cols = !Year,
                     names_to = "Type",
                     values_to = "Debt")

p1 <- d1 %>%
```

```r
  ggplot(aes(Year, Debt, fill=Type, color=Type), group=1)+
  geom_line(size = 0.7) +
  #scale_y_continuous(limits = c(11000, 19000)) +
  labs(x="Year",
       y="Median Debt",
       title="Year-to-year Changes in Median Debt Among Different Income Levels",
       color = "Level of Family Income")
p1 + theme_few() + theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks = c(2010, 2012, 2014, 2016, 2018, 2020))

# F2
# Year-to-year Changes in Median Debt of Different Income Levels in Institution Types
d2 <- debt_cost_sumstat_year %>%
  select(1:5)
colnames(d2) <- c("year", "institution_type","Low Income","Median Income","High Income")
d2 <- d2 %>%
  pivot_longer(cols = !year & !institution_type,
               names_to = "Type",
               values_to = "Debt") %>%
  mutate(institution_type = capitalize(institution_type))


p2 <- d2 %>%
  ggplot(aes(year, Debt, fill=Type, color=Type), group=1)+
  geom_line(size = 0.7) +
  labs(x="Year",
       y="Median Debt",
       title="Year-to-year Changes in Median Debt of Different Income Levels in Institution Types",
       color = "Level of Family Income")
p2  + facet_wrap(institution_type~.) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks = c(2010, 2012, 2014, 2016, 2018,2020))

# F3
# Comparative Changes in Median Debt by Institution Type and Income Level in New York and Other States
debt_cost_sumstat_year_3 <- education_data_BA_cost %>%
  mutate(state = ifelse(state_id == "NY", "New York","Other State")) %>%
  group_by(state, year, institution_type) %>%
  summarise(
    mean_default_rate = mean(default_rate, na.rm = T),
    mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
    mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
    mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
    mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
    mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
    mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T)
  ) %>%
  ungroup()
add3 <- debt_cost_sumstat_year_3 %>%
  mice(method = "rf", m=5,printFlag = FALSE)
```

```r
d3 <- complete(add3) %>%
  select(1:3,5:7)
colnames(d3) <- c("state","year", "institution_type","Low Income","Median Income","High Income")
d3 <- d3 %>%
  pivot_longer(cols = c(4:6),
               names_to = "Type",
               values_to = "Debt") %>%
  mutate(institution_type = capitalize(institution_type))

p3 <- d3 %>%
  ggplot(aes(year, Debt, fill=Type, color=Type), group=1)+
  geom_line(size = 0.7) +
  labs(x="Year",
       y="Median Debt",
       title="Comparative Changes in Median Debt by Institution Type and Income Level in New York and O
       color = "Level of Family Income")
p3 + facet_grid(institution_type~state) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks = c(2010, 2012, 2014, 2016, 2018,2020))


# F4 debt-cost across years
# F4 (my/other states)
debt_cost_sumstat_year_4 <- education_data_BA_cost %>%
  mutate(state = ifelse(state_id == "NY", "My State","Other State")) %>%
  group_by(state, year) %>%
  summarise(
    mean_default_rate = mean(default_rate, na.rm = T),
    mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
    mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
    mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
    mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
    mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
    mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T)
  ) %>%
  ungroup()

debt4 <- debt_cost_sumstat_year_4 %>%
  select(-c(7:9)) %>%
  pivot_longer(cols = c(4:6), names_to = c("income_category"), values_to = c("debt")) %>%
  mutate(income_category = case_when(
    str_detect(income_category, "low_income") ~ "Low Income",
    str_detect(income_category, "median_income") ~ "Median Income",
    str_detect(income_category, "high_income") ~ "High Income"))

cost4 <- debt_cost_sumstat_year_4 %>%
  select(-c(4:6)) %>%
  pivot_longer(cols = c(4:6), names_to = c("income_category"), values_to = c("cost")) %>%
  mutate(income_category = case_when(
```

```
    str_detect(income_category, "low_income") ~ "Low Income",
    str_detect(income_category, "median_income") ~ "Median Income",
    str_detect(income_category, "high_income") ~ "High Income"))

debt_cost_data_by_year_4 <- debt4 %>%
  inner_join(cost4)

add4 <- debt_cost_data_by_year_4 %>%
  mice(method = "rf", m=5,printFlag = FALSE)
d4 <- complete(add4) %>%
  pivot_longer(cols = c(5:6),
               names_to = "Type",
               values_to = "Value")

p4 <- d4 %>%
  ggplot(aes(x=year, y=Value,fill=Type))+
  geom_bar(position = "dodge", stat = "identity", alpha = 0.7, width = 1.3)+
  labs(x="Year",
       y="Cost / Debt  Value ($)",
       title="Mean Cost and Median Debt Value in Different Family Income Among Years") +
  guides(fill = guide_legend(title = NULL))

p4 + facet_grid(income_category ~ state) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5))

# F4  Public/private
add42 <- debt_cost_data_by_year %>%
  group_by(year) %>%
  mice(method = "rf", m=5,printFlag = FALSE)
d42 <- complete(add42) %>%
  pivot_longer(cols = !c(1:3),
               names_to = "Type",
               values_to = "Value")  %>%
  mutate(institution_type = capitalize(institution_type),
         income_category  = capitalize(income_category),
         Type    = capitalize(Type))

p42 <- d42 %>%
  ggplot(aes(x=year, y=Value,fill=Type))+
  geom_bar(position = "dodge", stat = "identity", alpha = 0.7, width = 1.3)+
  labs(x="Year",
       y="Cost / Debt  Value ($)",
       title="Mean Cost and Median Debt Value in Different Family Income Among Years") +
  guides(fill = guide_legend(title = NULL))

p42 + facet_grid(institution_type~income_category) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
```

```
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks = c(2010, 2012, 2014, 2016, 2018,2020))

# F5 debt/cost - default rate (public/pri)
# low high median together

debt_cost_sumstat_year_5 <- education_data_BA_cost %>%
  group_by(institution_type, year) %>%
  summarise(
    mean_default_rate = mean(default_rate, na.rm = T),
    mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
    mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
    mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
    mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
    mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
    mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T)
  ) %>%
  ungroup()

debt5 <- debt_cost_sumstat_year_5 %>%
  select(-c(7:9)) %>%
  pivot_longer(cols = c(4:6), names_to = c("income_category"), values_to = c("debt")) %>%
  mutate(income_category = case_when(
    str_detect(income_category, "low_income") ~ "Low Income",
    str_detect(income_category, "median_income") ~ "Median Income",
    str_detect(income_category, "high_income") ~ "High Income"))

cost5 <- debt_cost_sumstat_year_5 %>%
  select(-c(4:6)) %>%
  pivot_longer(cols = c(4:6), names_to = c("income_category"), values_to = c("cost")) %>%
  mutate(income_category = case_when(
    str_detect(income_category, "low_income") ~ "Low Income",
    str_detect(income_category, "median_income") ~ "Median Income",
    str_detect(income_category, "high_income") ~ "High Income"))

debt_cost_data_by_year_5 <- debt5 %>%
  inner_join(cost5)

add5 <- debt_cost_data_by_year_5 %>%
  mice(method = "rf", m=5,printFlag = FALSE)

d5 <- complete(add5) %>%
  mutate(institution_type = capitalize(institution_type))

p5_debt <- d5 %>%
  ggplot(aes(x=year, y=debt, group=income_category, color=income_category)) +
  geom_smooth(method="lm",se = FALSE)  +
  geom_point() +
  geom_point(aes(y = mean_default_rate*100000, col = "Mean Default Rate")) +
  geom_smooth(aes(y = mean_default_rate*100000, col = "Mean Default Rate"), method = "lm", se = FALSE) +
  labs(x="Year",
       y="Value",
```

```
      title="Correlation Between Default Rates and Public/Private Institutions Across Household Income
  labs(col=" ")
p5_debt  + facet_wrap(institution_type~.) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks = c(2010, 2012, 2014, 2016, 2018,2020))


#------ Debt, cost, and earnings after graduation ----------#
# Data Preprocessing
graduates_income_2018 <- read_csv("graduates_income_2018.csv") %>%
  rename(year = 'YEAR', school_id = 'UNITID', school_name = 'INSTNM')%>%
  mutate(school_name = tolower(school_name)) %>%
  mutate(across(.cols=4:12, .fns=as.numeric)) #

all_data <- education_data_BA_cost %>%
  left_join(graduates_income_2018, by = c("school_id")) %>%
  select(-c(1:2),-5) %>%
  select(-school_name,-year.y)

add_all_data <- all_data %>%
  mice(method = "rf", m=5, printFlag = FALSE)
all_data_clean <- complete(add_all_data)

# mean
mean_all_data <- all_data_clean %>%
  mutate(state = ifelse(state_id == "NY", "New York","Other State")) %>%
  group_by(state, year.x) %>%
  summarise(
    mean_default_rate = mean(default_rate,na.rm = T),
    mean_real_family_income = mean(real_family_income,na.rm = T),
    mean_debt_for_low_income = mean(real_debt_low_income, na.rm = T),
    mean_debt_for_median_income = mean(real_debt_med_income, na.rm = T),
    mean_debt_for_high_income = mean(real_debt_high_income, na.rm = T),
    mean_cost_for_low_income = mean(real_cost_low_income, na.rm = T),
    mean_cost_for_median_income = mean(real_cost_med_income, na.rm = T),
    mean_cost_for_high_income = mean(real_cost_high_income, na.rm = T),
    mean_earn6_for_low_income = mean(MD_EARN_WNE_INC1_P6, na.rm = T),
    mean_earn6_for_median_income = mean(MD_EARN_WNE_INC2_P6, na.rm = T),
    mean_earn6_for_high_income = mean(MD_EARN_WNE_INC3_P6, na.rm = T),
    mean_earn8_for_low_income = mean(MD_EARN_WNE_INC1_P8, na.rm = T),
    mean_earn8_for_median_income = mean(MD_EARN_WNE_INC2_P8, na.rm = T),
    mean_earn8_for_high_income = mean(MD_EARN_WNE_INC3_P8, na.rm = T),
    mean_earn10_for_low_income = mean(MD_EARN_WNE_INC1_P10, na.rm = T),
    mean_earn10_for_median_income = mean(MD_EARN_WNE_INC2_P10, na.rm = T),
    mean_earn10_for_high_income = mean(MD_EARN_WNE_INC3_P10, na.rm = T)
  ) %>%
  ungroup() %>%
  mutate(mean_cost_for_median_income = if_else(is.nan(mean_cost_for_median_income),(mean_cost_for_low_i
```

```r
med_earns <- mean_all_data %>%
  select(c(1:4),mean_earn6_for_median_income,mean_earn8_for_median_income,mean_earn10_for_median_income)
colnames(med_earns) <- c("State","Year","mean_default_rate","mean_real_family_income", "6 Years", "8 Ye
med_earns <- med_earns %>%
  pivot_longer(cols = -c(1:4),
               names_to = "Graduate Years",
               values_to = "Median Income")

low_earns <- mean_all_data %>%
  select(c(1:4),mean_earn6_for_low_income,mean_earn8_for_low_income,mean_earn10_for_low_income)
colnames(low_earns) <- c("State","Year","mean_default_rate","mean_real_family_income", "6 Years", "8 Ye
low_earns <- low_earns %>%
  pivot_longer(cols = -c(1:4),
               names_to = "Graduate Years",
               values_to = "Low Income")

high_earns <- mean_all_data %>%
  select(c(1:4),mean_earn6_for_high_income,mean_earn8_for_high_income,mean_earn10_for_high_income)
colnames(high_earns) <- c("State","Year","mean_default_rate","mean_real_family_income", "6 Years", "8 Y
high_earns <- high_earns %>%
  pivot_longer(cols = -c(1:4),
               names_to = "Graduate Years",
               values_to = "High Income")

earns <- low_earns %>%
  left_join(med_earns) %>%
  left_join(high_earns) %>%
  pivot_longer(cols = -c(1:5),
               names_to = "Family Income Level",
               values_to = "Earns")

debt <- mean_all_data %>%
  select(c(1:4),mean_debt_for_low_income,mean_debt_for_median_income,mean_debt_for_high_income)
colnames(debt) <- c("State","Year","mean_default_rate","mean_real_family_income", "Low Income", "Median
debt <- debt %>%
  pivot_longer(cols = -c(1:4),
               names_to = "Family Income Level",
               values_to = "Debt")

cost <- mean_all_data %>%
  select(c(1:4),mean_cost_for_low_income,mean_cost_for_median_income,mean_cost_for_high_income)
colnames(cost) <- c("State","Year","mean_default_rate","mean_real_family_income", "Low Income", "Median
cost <- cost %>%
  pivot_longer(cols = -c(1:4),
               names_to = "Family Income Level",
               values_to = "Cost")

data <- debt %>%
  inner_join(cost) %>%
  inner_join(earns)


# Plot: Cost, and Earnings Analysis
```

```r
# F6
p6 <- data %>%
  ggplot(aes(x=Cost, y=Earns, group=`Graduate Years`, color=`Graduate Years`)) +
  geom_smooth(method="lm",se = FALSE,size = 0.7)  +
  geom_point() +
  labs(x="Cost / $",
       y="Graduate Earns / $",
       title="Correlation Analysis of Graduate Earns and College Costs in New York and Other States")

p6  + facet_grid(State ~ `Family Income Level`) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5))

# F7
p7 <- data %>%
  ggplot(aes(x=Debt, y=Earns, group=`Graduate Years`, color=`Graduate Years`)) +
  geom_smooth(method="lm",se = FALSE,size = 0.7)  +
  geom_point() +
  labs(x="Debt / $",
       y="Graduate Earns / $",
       title="Correlation Analysis of Graduate Earns and Debt in New York and Other States")


p7  + facet_grid(State ~ `Family Income Level`) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5))


# F8
p8 <- data %>%
  ggplot(aes(x=mean_real_family_income , y=Earns, group=`Graduate Years`, color=`Graduate Years`)) +
  geom_smooth(method="lm",se = FALSE,size = 0.7)  +
  geom_point() +
  labs(x="Family Income / $",
       y="Graduate Earns / $ ",
       title="Correlation Analysis of Graduate Earns and Family Income in New York and Other States")


p8  + facet_grid(~State) + theme_bw() +
  theme(strip.background = element_rect(fill = "white",color = "black"),
        strip.text.x = element_text(color = "black"),
        #strip.background = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5))
```