EC 6062 Applied Econometrics for business Project

21041725

29 Jan 2022

Introduction

The concept of a "higher education" which used to be a rare sight back in the day has now become a bare minimum requirement, even for entry level jobs. Indeed, in today's job market, a college education is considered an essential part of schooling. In the United States of America, the percentage of college graduates has risen from 6 percent in 1950 to 21 percent in 1992 (Elfin 1993).

As an increasing number of people achieve higher levels of education, it continuously raises the bar for what counts as a good education. In between government sponsored education schemes and corporations demanding the best (most qualified) candidates, why shouldn't you get a college degree? College degrees are more desirable now than ever before. In 1990, college graduates earned 50 percent more than high school graduates (Lazear 2006).

But with so many choices, so many degrees, so many universities flooding the market, what separates the good from the best colleges? Where does one study to get the best bang for their buck? Is there a "best return on investment"? This project aims to identify the factors which influence the rate of pay after obtaining an undergraduate degree and examine the correlation between them. Put simply, the key question here is **"Does job salary depend on university costs?"** Some of the additional questions examined will be:

1. Are public schools more expensive than private schools?
2. Which states are the most expensive to study in? Which ones pay the most?
3. Does studying in a STEM field pay more?
4. Has university tuition gone up in the last few years? Why?

Literature Review

The cost of college tuition has been steadily increasing every year (Gose 1996). Tuition rates have grown drastically, much faster than median household income rates in the past years (Basch 1997; Heller 1997; Joyner 1996) speculates the reason for this is increase in expenditure, and an increased dependence on tuition for funds for the colleges.Tang (2004) speculates this is due to an increase in government aid. As increased governments introduce programs to help students apply to universities, they in turn ramp up their tuition fees in response. Interestingly enough, Kantrowitz (2002) found no relation between financial aid and tuition costs. Kantrowitz (ibid.) compares changes in tuition costs with the Higher Education Price Index (HEPI) which measures average cost of products purchased by universities. University tuition costs increase at a faster rate than that of the HEPI. The increase from 1997 to 1998 was 8.4 for tuition, 4.7 for HEPI, and 3.8 for the CPI (consumer price index).

Tang (2004) determined that the biggest factors affecting tuition price was the existence of professional schools, ranking, size of student and staff bodies, university location (specifically, in the northeast region of the US). The tuition also reflects the cost of living and expected pay in geographical regions. Kantrowitz (2002) investigated the sensitivity of tuition costs in public and private universities. Tuition costs are much more sensitive to donations for private universities than they are for public ones. Tuition rates are also sensitive to university spending and sources of funding, which is much more volatile for public universities than it is for public ones. The higher the tuition costs, the less sensitive it is to sources of funding.

Intuition

As any student applying to college will confirm, private universities are much more expensive than public universities owing to better facilities. Tuition has definitely been increasing every year, with private universities usually costing more at face value (before scholarships and grants). Intuitively, one would also think that universities with a significant percentage of students in STEM and larger, more diverse student bodies incur more costs due to facilities and equipment, leading to higher fees.

Data

The complete dataset consists of five datasets compiled from multiple sources (*U.S. Department of Education* 2022; *Payscale* 2022; *Tuitiontracker* 2022; *Priceonomics* 2022; *The Chronicle of Higher Education* 2022; *National Center for Education Statistics* 2021), containing:

1. Name of state, reduced to four major regions specified by the U.S. Census Bureau(*Census Regions and Divisions of the United States* 2022) to prevent model inflation.
2. Type of university, public or private
3. Net cost of attending university, the average of in state and out state tuition fees, plus living expenses collected from 3400 colleges, across 1998-2021.
4. In state university ranking
5. Total number of students
6. Percentage of students belonging to minority groups
7. Percentage of alumni who believe they are making the world a better place
8. Percentage of student body in STEM (Science, Technology, engineering, and Mathematics) courses
9. Estimated average career pay after graduation

All data is collected from 4605 universities across the US, during the period of fall 2014 unless stated otherwise. All monetary variables are in USD, accounting for inflation. The data used for the model does not have any missing values.

Data Visualization

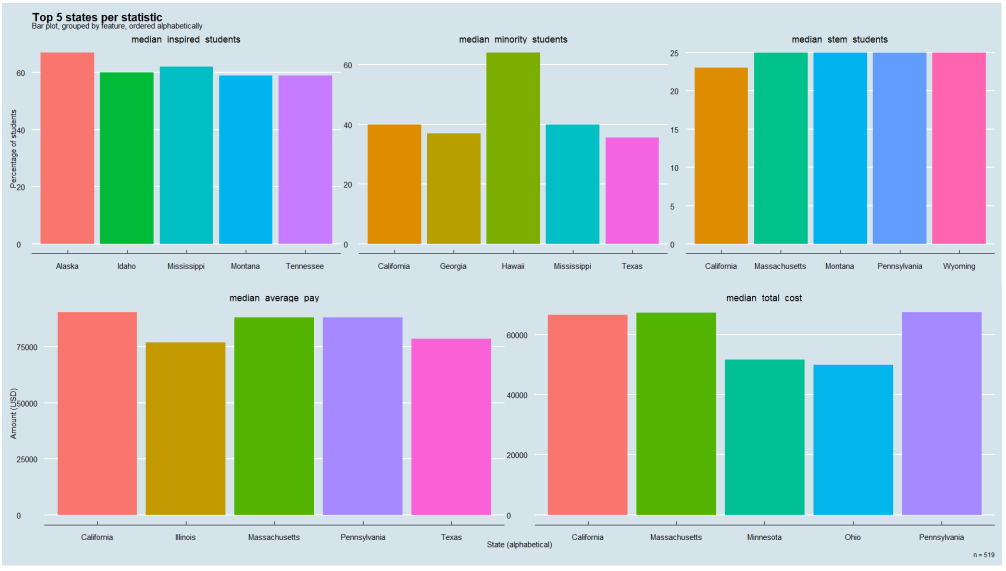

Figure 1: Top 5 states for specific statistics

Alaska has the highest percentage of students who believe their work will make a difference in the world. Hawaii has over 60 percent minority students, which is quite the jump from California at 40. Multiple states tie for highest percentage of STEM students, but none of them dominate. Universities from California, Massachusetts and Pennsylvania cost the most, but also present students with the highest paying jobs after graduation.

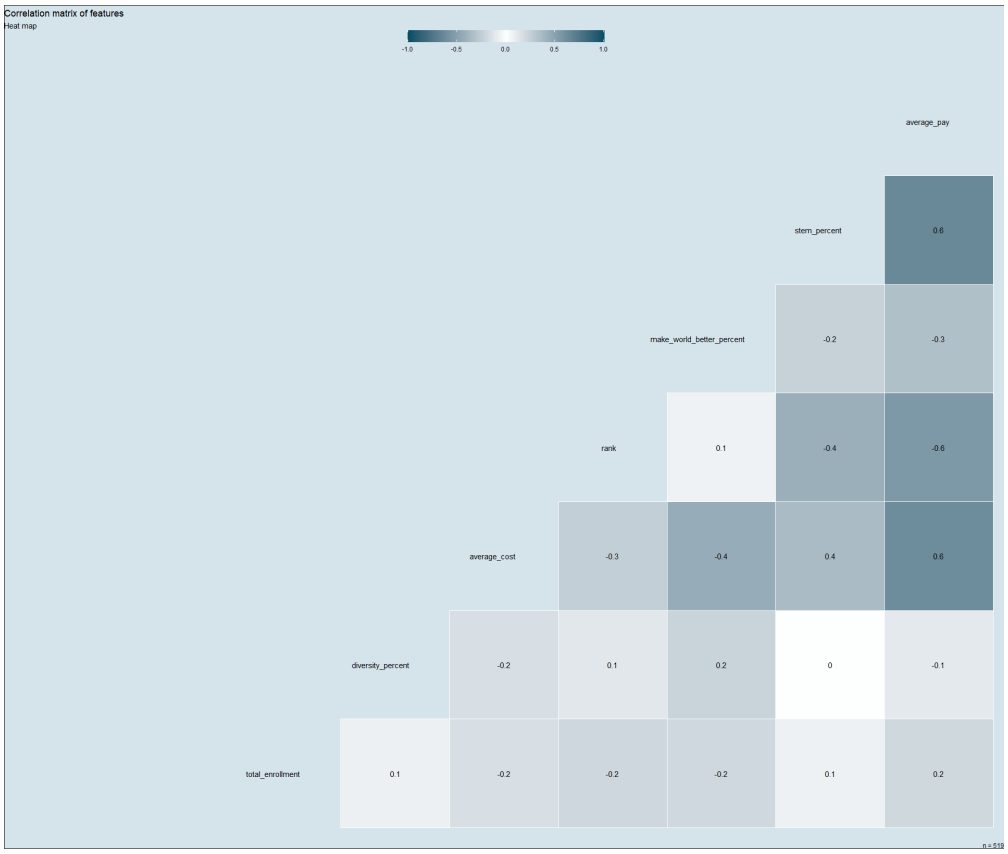

Figure 2: Correlation matrix

A correlation matrix is plotted to observe the correlation of all variables. There is a moderate correlation between post-graduation pay and percentage of stem students, university ranking, and average cost of studying. These variables are also correlated with each other, which is noted for further investigation. Categorical variables such as type of university and location are excluded.

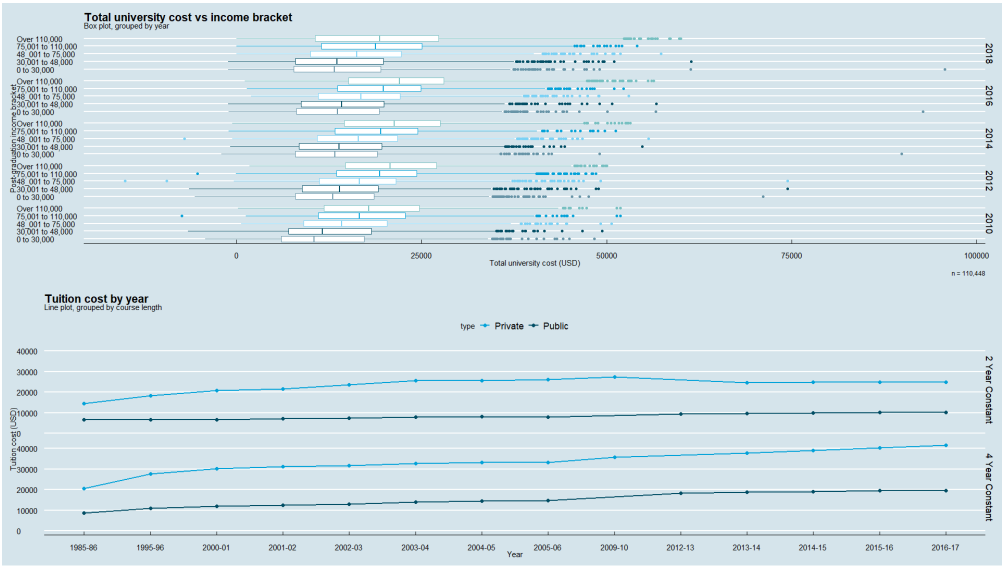

Figure 3: Tuition costs by year

The plots show that through the years, the cost of university for each income group has been increasing steadily. An interesting point to note here is the presence of a few data points where cost is negative. These have been identified as students who have been granted "full ride" scholarships and allowances to study, leading to a negative university cost.

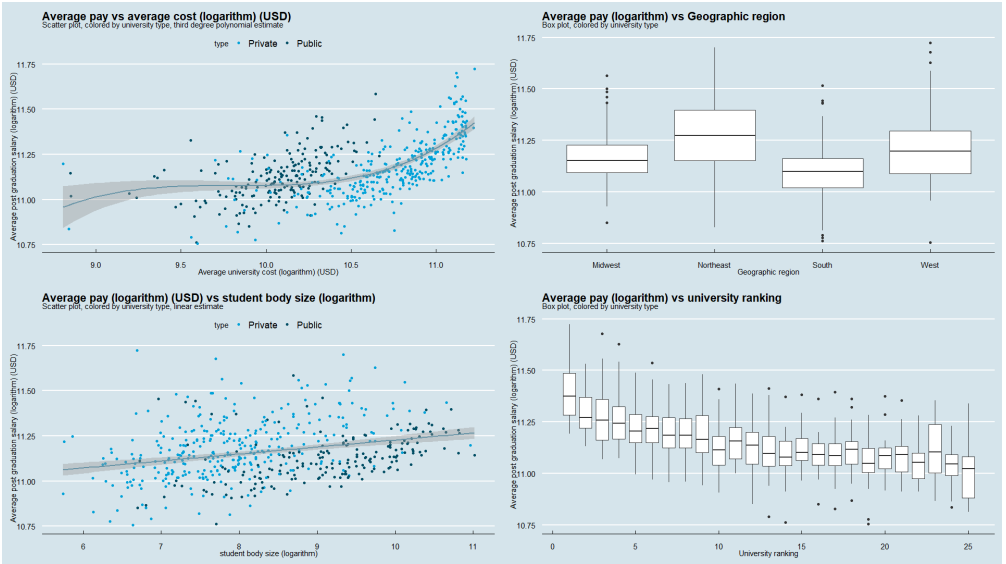

Figure 4: Individual relationships between dependent and independent variables

Plotting the key variables used in the final model in order to better observe the relationships between them. Average pay follows a third-degree polynomial relationship with average cost, with private universities at the latter end of the graph. Median tuition is highest for private universities in the northeast region. There is a linear relationship between average pay and student body size. This can be explained by larger universities requiring more funds to accommodate a higher number of students. Average pay is also inversely proportional to ranking, which makes sense as a lower numerical ranking is better.

Model estimation methodology

Table 1:

	Dependent variable:		
	average pay (logarithm)		
	M1	M2	M3
	(1)	(2)	(3)
poly(log(average_cost), 3, raw = T)1	20.694^{***}	19.612^{***}	
	(6.071)	(6.786)	
poly(log(average_cost), 3, raw = T)2	-2.133^{***}	-2.033^{***}	
	(0.599)	(0.666)	
poly(log(average_cost), 3, raw = T)3	0.073^{***}	0.070^{***}	
	(0.020)	(0.022)	
I(average_cost^3)			$7.9779\text{e-}16^{***}$
			(5.3435e-17)
log(total_enrollment)	0.039^{***}	0.039^{***}	0.037^{***}
	(0.004)	(0.004)	(0.004)
poly(rank, 2)1	-1.231^{***}	-1.202^{***}	
	(0.094)	(0.095)	
poly(rank, 2)2	0.435^{***}	0.474^{***}	
	(0.089)	(0.090)	
rank			-0.008^{***}
			(0.001)
RegionNortheast	0.018	0.021	0.004
	(0.013)	(0.013)	(0.013)
RegionSouth	-0.046^{***}	-0.044^{***}	-0.049^{***}
	(0.009)	(0.010)	(0.010)
RegionWest	0.007	0.013	0.005
	(0.012)	(0.012)	(0.012)
make_world_better_percent		0.0003	
		(0.001)	
Constant	-56.179^{***}	-52.303^{**}	10.884^{***}
	(20.483)	(23.028)	(0.034)
Observations	519	504	519
R^2	0.688	0.686	0.674
Adjusted R^2	0.682	0.680	0.670
Residual Std. Error	0.088 (df = 509)	0.086 (df = 493)	0.089 (df = 512)
F Statistic	124.468^{***} (df = 9; 509)	107.771^{***} (df = 10; 493)	176.653^{***} (df = 6; 512)
Note:			*p<0.1; **p<0.05; ***p<0.01

The model was initially estimated using the ordinary least squares (OLS) estimator, with a linear equation form. Ramsey's RESET test (Ramsey 1969) was carried out to determine correctness of functional form, after which the functional form was modified to include logarithm and polynomial terms. The Variance Inflation Factor (VIF) for each variable was calculated to investigate multicollinearity, and those with VIF < 4 were removed. Finally, the Breusch-Pagan (Breusch and Pagan 1979), and White's (White 1980) tests for heteroskedasticity were performed, both of which presented strong evidence of heteroskedasticity. The covariance matrix of the coefficients was re-estimated using heteroskedasticity robust estimation techniques.

Regression metrics and Inference

The final model equation is given by:

$$ln(Average\ pay) = \alpha + \beta_1(ln(Average\ Cost)^3) + \beta_2(ln(Total\ Enrollment)) + \beta_3(Rank) + \beta_4(Region,\ Northeast) + \beta_5(Region,\ South) + \beta_6(Region,\ West) + \epsilon$$

Table 2:

Variable	Estimate	Std. Error	t value	Pr(> |t|)
α	1.0884e+01	4.0033e-02	271.8727	< 2.2e-16
β_1	7.9779e-16	5.3435e-17	14.9300	< 2.2e-16
β_2	3.7263e-02	3.9550e-03	9.4219	< 2.2e-16
β_3	-7.7146e-03	6.6544e-04	-11.5931	< 2.2e-16
β_4	4.3399e-03	1.5777e-02	0.2751	0.7834
β_5	-4.8949e-02	8.3941e-03	-5.8313	9.749e-09
β_6	4.9729e-03	1.3768e-02	0.3612	0.7181

Table 3:

Variable	GVIF	Df	GVIF^(1/(2*Df))
Average Cost (logarithm)	1.312549	1	1.145665
Total Enrollment (logarithm)	1.069954	1	1.034386
Rank	1.147211	1	1.071079
Region	1.241954	3	1.036774

The RESET test conducted provided a test statistic of 0.50018, with a p-value of 0.6067. The VIF results of the model variables are given in table 3. All VIF values fall below the threshold for multicollinearity. The Breusch-Pagan Test for heteroskedasticity conducted resulted in a test statistic of 45.4, with a p-value of 3.94e-8. White's Test for heteroskedasticity returned a test statistic of 62.5, with a p-value of 7.86e-9. Both tests indicated the presence of heteroskedasticity, and as such, the standard errors for the model were recalculated and presented in table 2.

The variables included in the data but excluded from the chosen model are not significantly correlated with the target variable, eliminating the chance of known omitted variable bias. For the given regression model, simultaneous causality is a rare possibility. Certain fields might be popular and highly lucrative, leading to a higher influx of students, encouraging universities to increase tuition costs for that specific course.

Examining the p-values of the coefficients (adjusted for heteroskedasticity), average net cost, total enrollment and university ranking are all statistically significant (p-value < 2.2e-16). The location of the school in the south is statistically significant (p-value = 9.749e-09), but not for other locations. Finally, the p-values for all the other terms indicate that they are

The Adjusted R-squared value of 0.6705 indicated that 67.05 percent of the variance in the model is explained by the chosen variables, signifying a moderate

correlation. The Residual standard error is 0.08913 on 512 degrees of freedom, and the F-statistic is 176.7 on 6 and 512 DF, with a p-value less than 2.2e-16.

Conclusion

Through this experiment, the relationship between multiple variables and the average rate of pay was investigated. The average job salary after obtaining a university degree was found to be dependent on the cost of tuition of the university attended, the size of the student body, the university ranking, and the location. This is backed up with the literature surveyed, which included additional factors which are indirectly correlated with our target variable. Multiple statistical tests were applied in order to estimate an accurate model to determine the statistical significance of these variables. Observing the final model confirmed that these variables are statistically significant.

The findings from this project also reaffirm the intuitive answers to the questions posed earlier. Further research possibilities involve re-estimating the model with additional data revolving around the education industry such as online certification courses, university placement programs, and also examining data for other countries.

References

Basch, D L (1997). "Private colleges' pricing experience in the early 1990s: the impact of rapidly increasing college-funded grants". In: *Research in Higher Education* 38.3, pp. 271–296.

Breusch, T. S. and A. R. Pagan (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Variation". In: *Econometrica* 47.5, pp. 1287–1294. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1911963.

Census Regions and Divisions of the United States (2022). URL: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.

Chu, Kathy (2009). "Average college credit card debt rises with fees, tuition". In: *USA Today. Retrieved from http://www. usatoday. com.*

Couch, Andrew (2020). *TidyTuesday: Estimating University ROI in R*. YouTube. URL: https://www.youtube.com/watch?v=ZUweG_URClw.

Craney, Trevor A and James G Surles (2002). "Model-dependent variance inflation factor cutoff values". In: *Quality engineering* 14.3, pp. 391–403.

Elfin, M (1993). "Does College Still Pay? America's Best Colleges: 1994 College Guide". In: *US News and World Report*.

Gose, B (1996). "Undergraduate tuition rises by an average of 5 percent (and) fact file: tuition and fees at more than 3,000 colleges and universities". In: *The Chronicle of Higher Education* 43.6, A38–A45.

Heller, D E (1997). "Student price response in higher education: an update to Leslie and Brinkman". In: *Journal of Higher Education* 68.6, pp. 624–659.

Ihlanfeldt, William (1980). "Achieving Optimal Enrollments and Tuition Revenues." In.

Joyner, C C (1996). *Higher Education: Tuition Increasing Faster than Household Income and Public Colleges' Costs. Report to Congressional Requesters, General Accounting Office, Health, Education, and Human Services Division*. Washington, DC.

Kantrowitz, Mark (2002). "Causes of faster-than-inflation increases in college tuition". In: *College and University* 78.2, pp. 3–10. URL: https://login.proxy.lib.ul.ie/login?url=https://www.proquest.com/scholarly-journals/causes-faster-than-inflation-increases-college/docview/225612640/se-2?accountid=14564.

Lazear, Edward P (2006). *Personnel Economics for Managers*. Nashville, TN: John Wiley & Sons.

Marek, Hlavac (2018). *Stargazer: Well-Formatted Regression and Summary. Statistics Tables. R package version 5.2. 1*.

National Center for Education Statistics (2021). URL: https://nces.ed.gov/programs/coe/indicator/cua.

National Center for Education Statistics (2022). URL: https://nces.ed.gov/fastfacts/display.asp?id=76.

Payscale (2022). URL: https://www.payscale.com/college-salary-report/best-schools-by-state/bachelors/new-hampshire.

Priceonomics (2022). URL: https://priceonomics.com/ranking-the-most-and-least-diverse-colleges-in/.

Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis". In: *Journal of the Royal Statistical Society.*

Series B (Methodological) 31.2, pp. 350–371. ISSN: 00359246. URL: http://www.jstor.org/stable/2984219.

Robinson, David (2018). *Tidy Tuesday Screencast: analyzing college major & income data in R*. YouTube. URL: https://www.youtube.com/watch?v=nx5yhXAQLxw.

Rodrigues, Bruno (2018). *Dealing with heteroskedasticity; regression with robust standard errors using R*. URL: https://www.brodrigues.co/blog/2018-07-08-rob_stderr/.

Silge, Julia (2021). *Model student debt inequality with tidymodels*. YouTube. URL: https://www.youtube.com/watch?v=4ayOjlRv8bA.

Stine, Robert A (1995). "Graphical interpretation of variance inflation factors". In: *The American Statistician* 49.1, pp. 53–56.

Stobierski, Tim (2022). *Average Salary by Education Level: Value of a College Degree*. URL: https://www.northeastern.edu/bachelors-completion/news/average-salary-by-education-level/.

Syll, Lars (2022). *Overcontrolling in econometrics — a wasteful practice ridden with errors*. URL: https://larspsyll.wordpress.com/2020/02/05/overcontrolling-in-econometrics-a-wasteful-practice-ridden-with-errors/.

Tang (2004). "College tuition and perceptions of private university quality". In: *International Journal of Educational Management*.

The Chronicle of Higher Education (2022). URL: https://www.chronicle.com/article/tuition-and-fees-1998-99-through-2018-19/.

The Chronicle of Higher Education (2022). URL: https://www.chronicle.com/article/student-diversity-at-4-725-institutions/.

Tuitiontracker (2022). URL: https://www.tuitiontracker.org/.

U.S. Department of Education (2022). URL: https://www.ed.gov/.

White, Halbert (1980). *A Heteroskedasticity-Consistent Covariance Matrix Estimator And A Direct Test For Heteroskedasticity*.

Wooldridge, Jeffrey M (2005). "Violating ignorability of treatment by controlling for too many factors". In: *Econometric Theory* 21.5, pp. 1026–1028.

Zarembka, Paul (1974). *Frontiers in econometrics*.

Zeileis, Achim (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators". In: *Journal of Statistical Software* 11.10, pp. 1–17. DOI: 10.18637/jss.v011.i10. URL: https://www.jstatsoft.org/index.php/jss/article/view/v011i10.

Appendix

List of Figures

List of Tables

Code

```
# Setup ----

setwd("D:/RStudio/UL/EC6062 project") # Set working
    directory

library("car") # Calculate Variance Inflation Factor
library("GGally") # Correlation matrix
library("ggplot2") # Graphs
library("ggthemes") # Theme options for graphs
library("gridExtra") # Combine multiple graphs
library("lmtest") # Calculate standard errors adjusted
    for heteroskedasticity
library("sandwich") # Heteroscedasticity-consistent
    covariance matrix estimation
library("skedastic") # Breusch-Pagan and White's tests
    for heteroskedasticity
library("stargazer") # Regression analysis for multiple
    models
library("tidyverse") # Better data processing

# Download data ----

salary_potential <-
    readr::read_csv(
        "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
    )

tuition_cost <-
    readr::read_csv(
        "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
    )

diversity_school <-
    readr::read_csv(
```

```
        "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
    )

tuition_income <-
    readr::read_csv(
        "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
    )

historical_tuition <-
    readr::read_csv(
        "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
    )

state_region <-
    readr::read_csv(
        "https://raw.githubusercontent.com/cphalpert/census-regions/master/us%2
    )

# Cleaning and combining data

salary_potential <- salary_potential %>%
    rename(state = state_name) %>% # Rename "state_name"
       to state
    mutate(# Calculate "average pay" as mean of early
       and mid career pay
        average_pay = (early_career_pay +
           mid_career_pay) / 2,
        .keep = "unused")

tuition_cost <- tuition_cost %>%
    mutate(# Calculate "average cost" of university as
       mean of in state and out state
        average_cost = (in_state_total +
           out_of_state_total) / 2,
        .keep = "unused") %>%
    select("type", "name", "state", "average_cost") #
       Select specific columns

diversity_school <- diversity_school %>%
    filter(category == "Total Minority") %>% # Select
       total number of "minority" students for each
       school
    mutate(# Calculate"diversity_percent" as percentage
       of total students who belong to a "minority" group
        diversity_percent = ceiling(enrollment * 100 /
           total_enrollment)) %>%
    select("name", "total_enrollment",
       "diversity_percent", "state") # Select all
       columns except "category"
```

```
state_region <- state_region %>%
    rename("state" = "State") %>%
    select(-c("State Code"))

# Combine above datasets using inner joins based on
    matching name and state

data <- diversity_school %>%
    inner_join(tuition_cost, by = c("name", "state")) %>%
    inner_join(salary_potential, by = c("name",
        "state")) %>%
    inner_join(state_region, by = "state")

# Data visualization ----

# Plot top 5 states for certain attributes (percentage
    of students)
g1 <- data %>%
    group_by(state) %>% # For each state
    summarize(
        # Calculate median values
        median_inspired_students =
            median(make_world_better_percent),
        median_stem_students = median(stem_percent),
        median_minority_students =
            median(diversity_percent)
    ) %>%
    gather(key = "key", value = "value", -state) %>% #
        Calculate for each statistic
    group_by(key) %>%
    slice_max(value, n = 5) %>% # Take top 5 values
    ggplot(aes(x = state, y = value, fill = state)) + #
        Create plot
    geom_col() + # Bar plot
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    theme(legend.position = "none") + # Remove legend
    facet_wrap(~ key, scales = "free") + # Wrap by
        measured statistic
    labs(
        # Labels
        title = "Top 5 states per statistic",
        subtitle = "Bar plot, grouped by feature,
            ordered alphabetically",
        x = "",
        y = "Percentage of students"
    )

# Plot top 5 states for certain attributes (cost in USD)
g2 <- data %>%
```

```
    group_by(state) %>%
    summarize(
        median_total_cost = median(average_cost),
        median_average_pay = median(average_pay)
    ) %>%
    gather(key = "key", value = "value", -state) %>%
    group_by(key) %>%
    slice_max(value, n = 5) %>%
    ggplot(aes(x = state, y = value, fill = state)) +
    geom_col() +
    theme_economist() +
    scale_color_economist() +
    theme(legend.position = "none") +
    facet_wrap(~ key, scales = "free") +
    labs(caption = "n = 519",
         x = "State (alphabetical)",
         y = "Amount (USD)")

# Combine and display above graphs
grid.arrange(g1, g2)

# Calculate and plot correlation matrix
data %>%
    select(-c("name", "state", "type", "Region",
       "Division")) %>% # De-select categorical variables
    ggcorr(
        # Correlation plot
        hjust = 0.75,
        label = TRUE,
        layout.exp = 2,
        high = "#014d64",
        mid = "white",
        low = "#014d64"
    ) +
    theme_void() + # Colors
    theme(
        legend.position = "top",
        plot.background = element_rect(fill = "#d5e4eb"),
        legend.key.width = unit(2, "cm")
    ) +
    labs(# Labels
        title = "Correlation matrix of features",
        subtitle = "Heat map",
        caption = "n = 519")

# Plot historical tuition cost vs post-graduation income
    bracket data
tut1 <- tuition_income %>%
    filter(year %% 2 == 0) %>% # Only even years,
       prevent crowded graph
```

```
    ggplot(aes(# Plot graph
        y = net_cost,
        x = income_lvl,
        color = income_lvl)) +
    geom_boxplot() + # Specify box plot
    facet_grid(rows = vars(year), as.table = FALSE) + #
       Separate plot for each year
    coord_flip() + # Flip horizontally
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    theme(legend.position = "none") + # Remove legend
    labs(
        # Labels
        title = "Total university cost vs income
           bracket",
        subtitle = "Box plot, grouped by year",
        caption = "n = 110,448",
        y = "Total university cost (USD)",
        x = "Post-graduation income bracket"
    )

# Plot historical tuition data for public and private
   schools, based on course length
tut2 <- historical_tuition %>%
    filter(type != "All Institutions",
           # Keep only public and private schools
           tuition_type %in% c("4 Year Constant", "2
              Year Constant")) %>%
    ggplot(aes(# Plot
        x = year,
        y = tuition_cost,
        group = type)) +
    geom_point(aes(color = type), size = 2) + # Scatter
       plot for points
    geom_line(aes(color = type), size = 1) + # Line plot
    scale_y_continuous(limits = c(0, 42000)) + # Start
       y-axis at zero
    facet_grid(rows = vars(tuition_type)) + # Group
       plots by tuition type
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    labs(
        # Labels
        title = "Tuition cost by year",
        subtitle = "Line plot, grouped by course length",
        x = "Year",
        y = "Tuition cost (USD)"
    )

# Combine and display above graphs
```

```
grid.arrange(tut1, tut2)

# Plot tuition cost vs average post graduation salary

kv1 <- data %>%
    ggplot(aes(# Plot
        x = log(average_cost),
        y = log(average_pay))) +
    geom_point(aes(color = type)) + # Scatter plot for
       points
    geom_smooth(method = "lm",
                formula = y ~ poly(x, 3),
                color = "#6794a7") +
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    labs(
        # Labels
        title = "Average pay vs average cost (logarithm)
           (USD)",
        subtitle = "Scatter plot, colored by university
           type, third degree polynomial estimate",
        x = "Average university cost (logarithm) (USD)",
        y = "Average post graduation salary (logarithm)
           (USD)"
    )

# Plot geographic region vs average post graduation
   salary

kv2 <- data %>%
    ggplot(aes(
        # Plot
        x = Region,
        y = log(average_pay),
        group = Region
    )) +
    geom_boxplot() + # Scatter plot for points
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    labs(
        # Labels
        title = "Average pay (logarithm) vs Geographic
           region",
        subtitle = "Box plot, colored by university
           type",
        x = "Geographic region",
        y = "Average post graduation salary (logarithm)
           (USD)"
    )
```

```
# Plot total enrollment vs average post graduation salary

kv3 <- data %>%
    ggplot(aes(# Plot
        x = log(total_enrollment),
        y = log(average_pay))) +
    geom_point(aes(color = type)) + # Scatter plot for
        points
    geom_smooth(method = "lm",
                formula = y ~ x,
                color = "#6794a7") +
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    labs(
        # Labels
        title = "Average pay (logarithm) (USD) vs
            student body size (logarithm)",
        subtitle = "Scatter plot, colored by university
            type, linear estimate",
        x = "student body size (logarithm)",
        y = "Average post graduation salary (logarithm)
            (USD)"
    )

# Plot rank vs average post graduation salary

kv4 <-
    data %>%
    ggplot(aes(
        # Plot
        x = rank,
        y = log(average_pay),
        group = rank
    )) +
    geom_boxplot() + # Scatter plot for points
    theme_economist() + # Set theme
    scale_color_economist() + # Set colors
    labs(
        # Labels
        title = "Average pay (logarithm) vs university
            ranking",
        subtitle = "Box plot, colored by university
            type",
        x = "University ranking",
        y = "Average post graduation salary (logarithm)
            (USD)"
    )

# Combine and display above graphs
grid.arrange(kv1, kv2, kv3, kv4)
```

```
# Data analysis ----

# Select data for model
model_data <-
    data %>%
    select(-c(name, state, Division)) %>% # Exclude name
       of school and location
    mutate_at("type", as.factor) # Convert "type" to
       dummy variable

# Estimate models
lm_1 <-
    lm(
        log(average_pay) ~ log(average_cost) +
            total_enrollment +
            I(make_world_better_percent^2),
        data = model_data
    )

lm_2 <-
    lm(
        log(average_pay) ~ log(average_cost) +
            total_enrollment +
            I(rank^2),
        data = model_data
    )

lm_3 <-
    lm(
        log(average_pay) ~ log(average_cost) +
            total_enrollment +
            rank,
        data = model_data
    )

lm_4 <-
    lm(
        log(average_pay) ~ poly(log(average_cost), 3,
           raw = T) +
            log(total_enrollment) +
            poly(rank, 2),
        data = model_data
    )

lm_5 <-
    lm(
        log(average_pay) ~ poly(log(average_cost), 3,
           raw = T) +
            log(total_enrollment) +
```

```
            poly(rank, 2) +
            Region,
        data = model_data
    )

lm_6 <-
    lm(
        log(average_pay) ~ poly(log(average_cost), 3,
            raw = T) +
            log(total_enrollment) +
            poly(rank, 2) +
            Region +
            make_world_better_percent,
        data = model_data
    )

# View regression analysis for models

independant_variable_names_1 = c("average cost
    (logarithm)",
                                 "total enrollment",
                                 "make world better \\%",
                                 "rank\\^2",
                                 "rank",
                                 "(Intercept)")

independant_variable_names_2 = c("average cost
    (logarithm)",
                                 "average cost
                                    (logarithm)\\^2",
                                 "average cost
                                    (logarithm)\\^3",
                                 "total enrollment
                                    (logarithm)",
                                 "rank",
                                 "rank\\^2",
                                 "Region, Northeast",
                                 "Region, South",
                                 "Region, West",
                                 "make world better \\%",
                                 "(Intercept)")

stargazer(lm_1,lm_2,lm_3,
          type="latex",
          out="./stargazer1.tex",
          font.size="tiny",
          no.space = TRUE,
          dep.var.labels = c("average pay (logarithm)"),
          covariate.labels =
              independant_variable_names_1,
```

```
            column.labels = c("M1","M2","M3"))

stargazer(lm_4,lm_5,lm_6,
            type="latex",
            out="./stargazer2.tex",
            font.size="tiny",
            no.space = TRUE,
            dep.var.labels = c("average pay (logarithm)"),
            covariate.labels =
                independant_variable_names_2,
            column.labels = c("M4","M5","M6"))

# View model metrics
lm_5 %>% summary()

# Ramsey's RESET test to determine model misspecification
resettest(lm_5)

# Calculate Variance Inflation Factor
vif(lm_5)

# Breusch-Pagan Test for heteroskedasticity
breusch_pagan(lm_5)

# White's Test for heteroskedasticity
white_lm(lm_5)

# Linear model with standard errors adjusted for
    heteroskedasticity
coeftest(lm_5, vcov = vcovHC(lm_5))
```