Geographic Patterns in U.S. Higher Education Enrollment

Ibraheem Saqib Ellahi

August, 2025

Contents

1 Introduction

The choice of where to attend college is a significant economic decision for students, involving trade-offs between institutional quality, cost, and distance from home. This report analyzes the geographic patterns of first-year undergraduate enrollment in the United States for the year 2018. My findings shed light on the determinants of education choice.

2 Data and Methodology

The primary data for this analysis are from the 2018 IPEDS survey. I use the Directory Information file for institutional characteristics, such as location (latitude and longitude) and type (public, private, 4-year, 2-year), and the First-Year Migration file to track enrollment flows from a student's state of origin to their destination institution. Data for in-state tuition in 2018 was sourced from the `ic2018_cy.csv`, also in the 2018 IPEDS survey.

I approximate a student's origin location as the geographic center of their home state, calculated as the mean latitude and longitude of all postsecondary institutions within that state. The travel distance is then calculated as the Haversine distance between the origin state's center and the precise coordinates of the destination school. All data processing and analysis is performed in R.

3 Results

3.1 Objective 1: State-Level Student Migration

I first analyze the inflow and outflow of students for each state. The outflow share is the percentage of students from a given state who attend college out-of-state, while the inflow share is the percentage of students enrolled in a state who are from out-of-state.

The results, shown in Figure 1 and Table 1, indicate that small, densely populated states in the Northeast, such as Vermont and New Jersey, have the highest outflow shares. This is likely due to the close proximity of numerous high-quality institutions in neighboring states.

Table 1: Top 10 States by Student Outflow Share (2018)

State	Share Leaving
District of Columbia	72.1%
Vermont	48.9%
New Hampshire	44.9%
Connecticut	40.2%
Hawaii	39.2%
Alaska	37.9%
New Jersey	35.5%
Massachusetts	33.4%
Maryland	33%
Illinois	31.8%

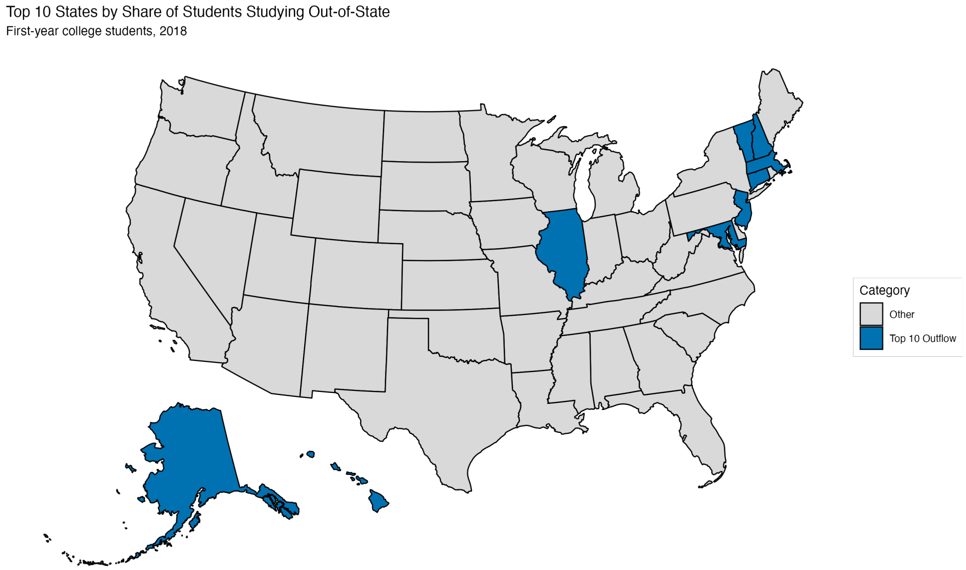
Top 10 States by Share of Students Studying Out-of-State
First-year college students, 2018

Figure 1: Top 10 States by Student Outflow Share (2018)

Conversely, Figure 2 and Table 2 show that the District of Columbia is the largest net importer of students, with over 85% of its student body from elsewhere. States like Rhode Island and Vermont also have high inflow shares, driven by popular private universities that attract a national student body.

Table 2: Top 10 States by Student Inflow Share (2018)

State	Share Arriving
District of Columbia	89.4%
New Hampshire	68.8%
Vermont	67%
Rhode Island	56.3%
North Dakota	47.8%
Delaware	38.8%
Idaho	38.3%
Utah	37.2%
South Dakota	37.1%
West Virginia	36.6%

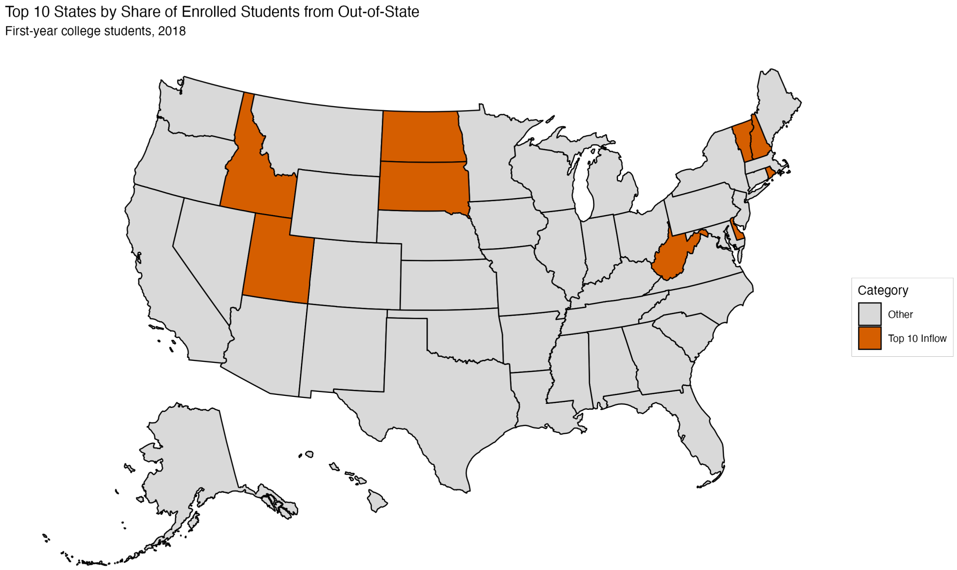
Top 10 States by Share of Enrolled Students from Out-of-State
First-year college students, 2018
Category
Other
Top 10 Inflow

Figure 2: Top 10 States by Student Inflow Share (2018)

3.2 Objective 2: Distribution of Travel Distances

Figure 3 displays the distribution of travel distances for students attending different types of postsecondary institutions. The distributions are heavily right-skewed, necessitating a log scale on the x-axis for clear visualization.

A clear pattern emerges: students attending 2-year institutions, both public and private, travel overwhelmingly short distances, consistent with their role as local community colleges. In contrast, students at 4-year institutions are willing to travel much farther.

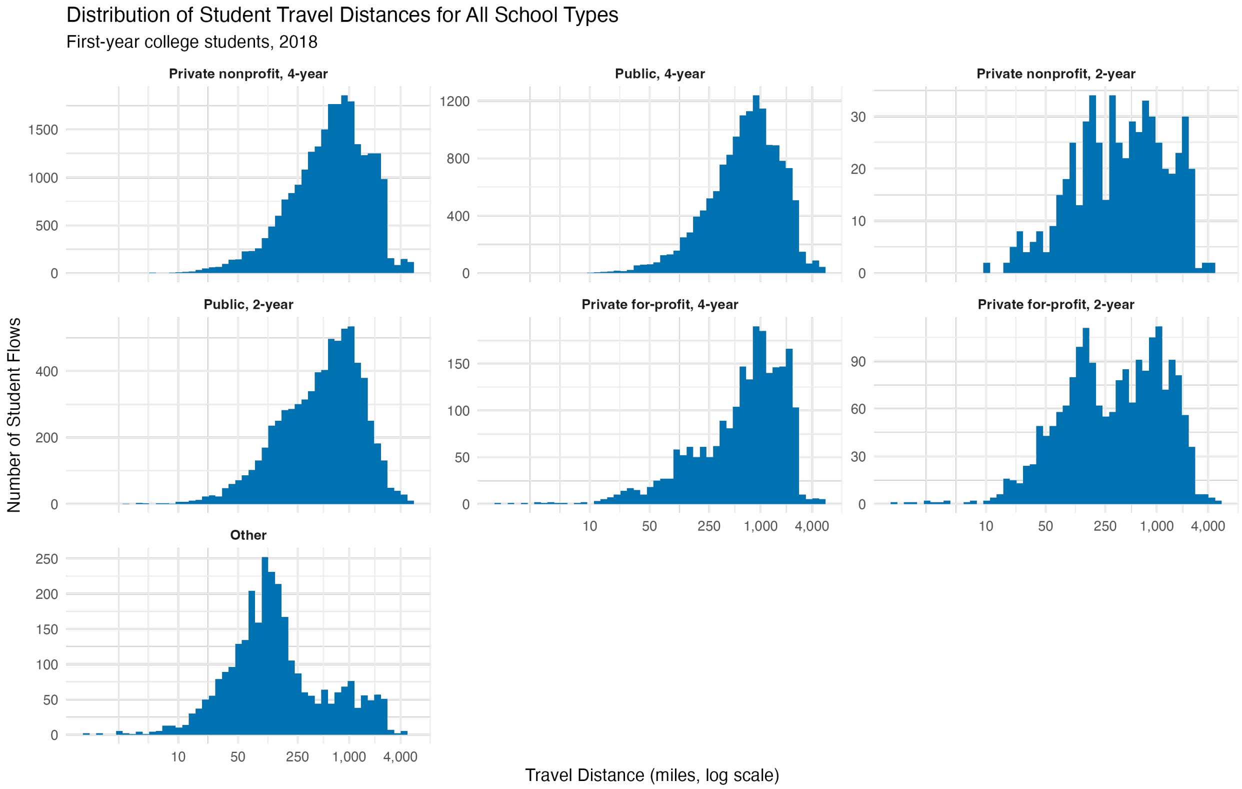

Figure 3: Distribution of Student Travel Distances for All School Types (2018)

Figure 4 contrasts the travel distance distributions for first-year students at 4-year public and private nonprofit universities in 2018 using an overlapping density plot to allow for fairer comparison. The plot reveals two key distinctions. First, the public university distribution is taller and narrower, indicating that its students are highly concentrated within a specific range of travel distances. In contrast, private university students are drawn from a wider, national geographic area. Second, the private university distribution is flatter and has a ”fatter tail” on the extreme right, indicating a slightly higher willingness to travel extreme distances for private universities.

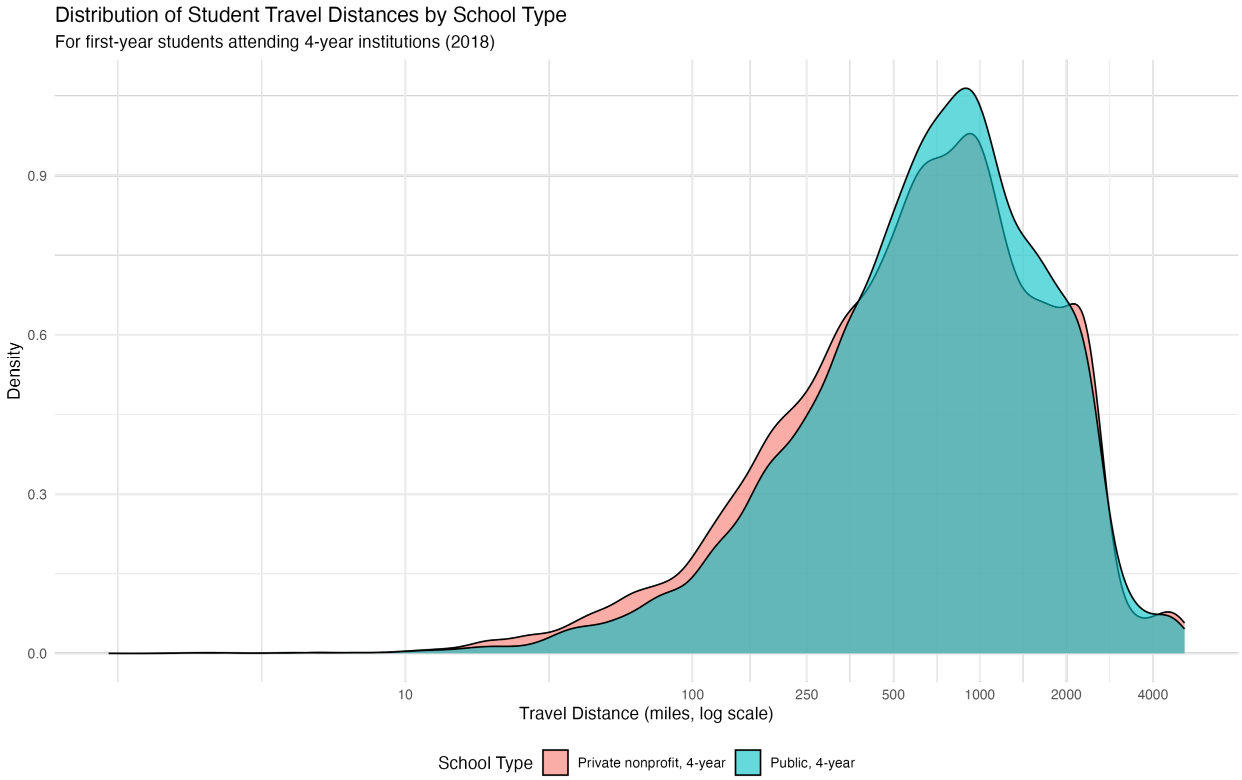

Figure 4: Distribution of Student Travel Distances by School Type (2018)

3.3 Objective 3: Economic Forces and Student Choice

The observed patterns can be understood through a simple utility maximization framework. A prospective student, i, chooses a college, j, to maximize their utility, which can be modeled as:

$$U_{ij} = V_j - P_j - C(D_{ij}) + \epsilon_{ij}$$

where V_j is the value or quality of school j, P_j is the net price, and $C(D_{ij})$ is the cost associated with the distance D_{ij} between the student's home and the school. This cost includes not only travel expenses but also the psychic cost of being far from home.

This model predicts that students are only willing to incur the high cost of traveling a long distance if it is compensated by a significant increase in school quality (V_j) or a lower net price (P_j). The distributions in Figure 3 support this: public 2-year colleges offer low distance cost but perhaps lower perceived V_j for many students, while elite private universities may offer a high enough V_j to justify a very large distance cost.

3.4 Objective 4: Determinants of Travel Distance

Drawing from the utility model in the previous section, I investigate the link between in-state tuition and travel distance. I hypothesized that a high P_j (in-state tuition) for a student's local options would decrease their utility, making them more willing to incur a higher $C(D_{ij})$ (distance cost) to attend an out-of-state school. This "push factor" hypothesis predicts a positive correlation between in-state tuition and travel distance. However, the regression analysis in Table 3 shows a statistically significant negative relationship. This model is also limited, with an R-squared of only 4.0%, indicating it explains very little of the overall variation. This counter-intuitive result is likely driven by omitted variable bias. For instance, high in-state tuition may be correlated with an unobserved factor like higher institutional quality, which acts as a powerful "pull factor" for students to stay in-state, thus reducing average travel distances.

Table 3: Determinants of Student Travel Distance

	Dependent variable:
	Log(Travel Distance in Miles)
Avg. Home-State In-State Tuition	-0.0001^{***}
	(0.00000)
Constant	7.046^{***}
	(0.017)
Observations	38,796
R^2	0.040
Adjusted R^2	0.040
Residual Std. Error	1.001 (df = 38794)
F Statistic	$1{,}636.859^{***}$ (df = 1; 38794)
Note:	*p<0.1; **p<0.05; ***p<0.01

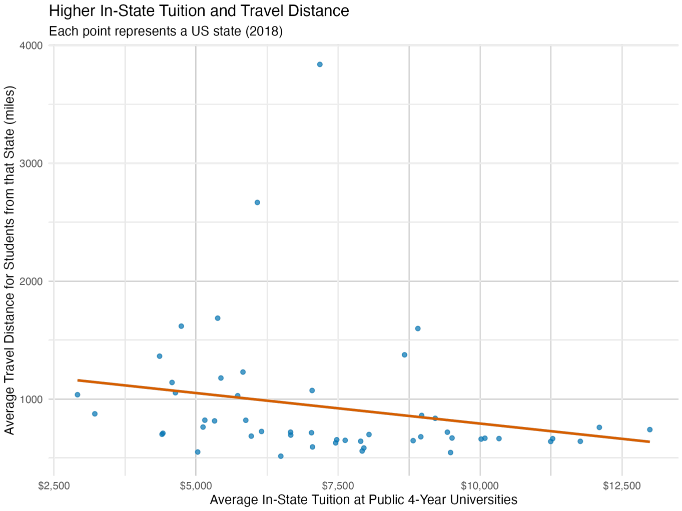

Figure 5: Distribution of Student Travel Distances for All School Types (2018)

To account for the possibility that high in-state tuition is simply correlated with higher state-level income (which enables travel), I introduced state median income as a control variable. Despite the statistical significance, the model's explanatory power is still limited with a 4.4% R-squared value as seen in Table 4. This suggests that the primary drivers of this decision are likely other factors, such as school prestige, student academic profile, and personal preferences.

Table 4: Impact of Tuition on Travel Distance, with and without Income Control

	Dependent variable:	
	Log(Travel Distance in Miles)	
	Simple Model	With Income Control
	(1)	(2)
Avg. Home-State Tuition	−0.0001***	−0.0001***
	(0.00000)	(0.00000)
State Median Income		0.00001***
		(0.00000)
Constant	7.046***	6.715***
	(0.017)	(0.033)
Observations	38,796	38,796
R^2	0.040	0.044
Adjusted R^2	0.040	0.044
Residual Std. Error	1.001 (df = 38794)	0.999 (df = 38793)
F Statistic	1,636.859*** (df = 1; 38794)	886.772*** (df = 2; 38793)
Note:		*p<0.1; **p<0.05; ***p<0.01

This is a cross-sectional analysis based on 2018 data. Further research could use a time-series component to build a panel data model. Such a model, using state-level fixed effects, would be able to control for unobserved, time-invariant state characteristics (like institutional quality) and could more accurately isolate the true causal effect of a change in tuition on student travel distance.

4 Conclusion

This analysis of 2018 IPEDS data reveals distinct geographic patterns in U.S. higher education enrollment. Key findings show that student migration is highly concentrated in the Northeast and that private nonprofit institutions draw students from a much wider geographic area than their public counterparts. These patterns align with an economic model of student choice, where individuals trade off distance costs against perceived institutional quality. However, econometric analysis provided a counter-intuitive result: higher in-state tuition is negatively correlated with travel distance, refuting my initial hypothesis. This suggests that the model is likely influenced by unobserved variables, such as institutional quality, which acts as a ”pull factor” for students to stay in-state.

A analysis.R code

```r
# ###########################################################################
# Description:
# This script loads and cleans 2018 IPEDS data to:
# 1. Analyze state-level student inflow and outflow.
# 2. Plot travel distance distributions for different school types.
# 3. Run regression models on the determinants of travel distance.
# All plots and tables are saved to the ’/output’ folder.
# This file is contained in the ’/code’ folder
# ###########################################################################

# Load required packages, installing them if they are not already present
if (!require("pacman")) install.packages("pacman")
pacman::p_load(
  tidyverse, # For data manipulation (dplyr, ggplot2, etc.)
  haven,     # For reading Stata, SAS, and SPSS files
  janitor,   # For cleaning data frames
  usmap,     # For creating US maps
  geosphere,  # For calculating geographic distances
  gridExtra,  # For arranging multiple grid-based plots
  stargazer,   # For table output to Latex
  dplyr,       # For data frames
  readr
)

# Define a relative path to the data folder.
data_path <- "./data/"


# ###########################################################################
# Objective 1: Top 10 Outflow and Inflow States in 2018
# ###########################################################################

# 1.Import Data
# ---------------------------------------------------------------------------

# IPEDS Data (Stata .dta files)
schools_2018 <- read_dta(file.path(data_path, "ipeds_directory_info_2018.dta"))
migration_2018 <- read_dta(file.path(data_path, "ipeds_fy_migration_2018.dta"))

# ACS Data (Stata .dta files)
acs_county_2018 <- read_dta(file.path(data_path, "traits_county_ACS_20132017.dta"))
acs_cbsa_2018 <- read_dta(file.path(data_path, "traits_cbsa_ACS_20132017.dta"))

# Geographic Data
# tab-separated text file, so read_tsv() from 'readr'.
counties_geo <- read_tsv(file.path(data_path, "2017_Gaz_counties_national.txt"))

# Cleaning up potential messy column names
counties_geo <- counties_geo %>%
  clean_names()

# Additional Data from IPEDS for tuition for Objective 4
tuition_data <- read_csv(file.path(data_path, "ic2018_ay.csv"))

# 2. Data Exploration
# ---------------------------------------------------------------------------
```

```

cat("--- School Directory Info ---\n")
glimpse(schools_2018)

cat("\n--- Student Migration Info ---\n")
glimpse(migration_2018)

cat("\n--- County Demographics (ACS) ---\n")
glimpse(acs_county_2018)

cat("\n--- County Geographic Info ---\n")
glimpse(counties_geo)

# 3. Data Preparation
# ---------------------------------------------------------------------------

# Merge school location and student migration data
student_flows <- left_join(migration_2018,
                           schools_2018 %>% select(unitid, fips, instnm),
                           by = "unitid")

# Create clean dataset for analysis
state_flows <- student_flows %>%
  rename(fips_origin = state_consumption,
         fips_school = fips,
         num_students = efres01) %>%
  # Convert FIPS codes from character to integer for clean joins
  mutate(fips_origin = as.integer(fips_origin),
         fips_school = as.integer(fips_school)) %>%
  # Filter out non-state territories, missing data, etc.
  filter(!is.na(fips_origin), !is.na(fips_school), fips_origin <= 56) %>%
  # Create the out-of-state identifier
  mutate(is_out_of_state = (fips_origin != fips_school))

```

```r
# State FIPS to name mapping
state_fips_map <- read_csv(
  "state_fips,state_name,state_abb
01,Alabama,AL
02,Alaska,AK
04,Arizona,AZ
05,Arkansas,AR
06,California,CA
08,Colorado,CO
09,Connecticut,CT
10,Delaware,DE
11,District of Columbia,DC
12,Florida,FL
13,Georgia,GA
15,Hawaii,HI
16,Idaho,ID
17,Illinois,IL
18,Indiana,IN
19,Iowa,IA
20,Kansas,KS
21,Kentucky,KY
22,Louisiana,LA
23,Maine,ME
24,Maryland,MD
25,Massachusetts,MA
26,Michigan,MI
27,Minnesota,MN
28,Mississippi,MS
29,Missouri,MO
30,Montana,MT
31,Nebraska,NE
32,Nevada,NV
33,New Hampshire,NH
34,New Jersey,NJ
35,New Mexico,NM
36,New York,NY
37,North Carolina,NC
38,North Dakota,ND
39,Ohio,OH
40,Oklahoma,OK
41,Oregon,OR
42,Pennsylvania,PA
44,Rhode Island,RI
45,South Carolina,SC
46,South Dakota,SD
47,Tennessee,TN
48,Texas,TX
49,Utah,UT
50,Vermont,VT
51,Virginia,VA
53,Washington,WA
54,West Virginia,WV
55,Wisconsin,WI
56,Wyoming,WY
", col_types = "ic" # ’i’ for integer, ’c’ for character
)

# 4. Calculate Migration Shares
# ---------------------------------------------------------------------------

# Outflow
state_outflow_shares <- state_flows %>%
  group_by(fips_origin) %>%
  summarize(
    total_students_from_state = sum(num_students, na.rm = TRUE),
    students_leaving_state = sum(num_students[is_out_of_state], na.rm = TRUE)
  ) %>%
  mutate(share_outflow = students_leaving_state / total_students_from_state) %>%
  left_join(state_fips_map, by = c("fips_origin" = "state_fips")) %>%
  arrange(desc(share_outflow))

```

```r
# Inflow
state_inflow_shares <- state_flows %>%
  group_by(fips_school) %>%
  summarize(
    total_students_in_state = sum(num_students, na.rm = TRUE),
    students_from_out_of_state = sum(num_students[is_out_of_state], na.rm = TRUE)
  ) %>%
  mutate(share_inflow = students_from_out_of_state / total_students_in_state) %>%
  left_join(state_fips_map, by = c("fips_school" = "state_fips")) %>%
  arrange(desc(share_inflow))

# 5. Print Top 10 Results
# ---------------------------------------------------------------------------


cat("--- Top 10 States by Student Outflow Share ---\n")
knitr::kable(head(state_outflow_shares, 10) %>% select(state_name, share_outflow), digits = 3)

cat("\n--- Top 10 States by Student Inflow Share ---\n")
knitr::kable(head(state_inflow_shares, 10) %>% select(state_name, share_inflow), digits = 3)

# 6. Create and Save Tables .tex
# ---------------------------------------------------------------------------


# Prepare data frame
outflow_table_data <- state_outflow_shares %>%
  head(10) %>%
  select(state_name, share_outflow) %>%
  mutate(share_outflow = paste0(round(share_outflow * 100, 1), "%")) %>%
  rename(State = state_name, ‘Share Leaving‘ = share_outflow)

# Use stargazer to create and save the LaTeX code
stargazer(
  outflow_table_data,
  type = "latex",                             # Specify LaTeX output
  summary = FALSE,                            # Print the data frame as-is
  rownames = FALSE,                           # Remove row numbers
  header = FALSE,                             # Remove the default LaTeX header
  title = "Top 10 States by Student Outflow Share (2018)",
  label = "tab:outflow",                      # LaTeX label for cross-referencing
  out = "./output/outflow_table.tex"          # File to save the code in
)

# Prepare data frame
inflow_table_data <- state_inflow_shares %>%
  head(10) %>%
  select(state_name, share_inflow) %>%
  mutate(share_inflow = paste0(round(share_inflow * 100, 1), "%")) %>%
  rename(State = state_name, ‘Share Arriving‘ = share_inflow)

# Use stargazer to create and save the LaTeX code
stargazer(
  inflow_table_data,
  type = "latex",
  summary = FALSE,
  rownames = FALSE,
  header = FALSE,
  title = "Top 10 States by Student Inflow Share (2018)",
  label = "tab:inflow",
  out = "./output/inflow_table.tex"
)

```

```r
# 7. Create and Save Map PNGs
# ------------------------------------------------------------------------

# rename the ’state_abb’ column to ’state’ for the outflow data
outflow_map_data <- state_outflow_shares %>%
  rename(state = state_abb) %>%
  # Use rank() to find the top 10 states directly
  mutate(category = ifelse(rank(-share_outflow) <= 10, "Top 10 Outflow", "Other"))

glimpse(outflow_map_data)

# do the same as above for the inflow data
inflow_map_data <- state_inflow_shares %>%
  rename(state = state_abb) %>%
  mutate(category = ifelse(rank(-share_inflow) <= 10, "Top 10 Inflow", "Other"))

glimpse(inflow_map_data)

# Outflow Map
outflow_map <- plot_usmap(data = outflow_map_data, values = "category", labels = FALSE) +
  scale_fill_manual(name = "Category", values = c("Top 10 Outflow" = "#0072B2", "Other" = "grey85")) +
  theme(legend.position = "right") +
  labs(title = "Top 10 States by Share of Students Studying Out-of-State",
       subtitle = "First-year college students, 2018")

# Inflow Map
inflow_map <- plot_usmap(data = inflow_map_data, values = "category", labels = FALSE) +
  scale_fill_manual(name = "Category", values = c("Top 10 Inflow" = "#D55E00", "Other" = "grey85")) +
  theme(legend.position = "right") +
  labs(title = "Top 10 States by Share of Enrolled Students from Out-of-State",
       subtitle = "First-year college students, 2018")

ggsave(
  "./output/outflow_map.png",
  plot = outflow_map,
  width = 10,
  height = 6,
  dpi = 300
)

ggsave(
  "./output/inflow_map.png",
  plot = inflow_map,
  width = 10,
  height = 6,
  dpi = 300
)

# ############################################################################
# Objective 2: Graph distributions of student travel distances
# ############################################################################

# 1. Prepare Origin and Destination Coordinates
# ----------------------------------------------------------------------------

# Create a clean dataset of school locations with the correct column names
school_coords <- schools_2018 %>%
  select(unitid, latitude, longitud, sector, instnm)

```

```r
# To find the center of each state, I calculate the average lat/lon
# of all schools within that state using the correct 'longitud' column.
# A better approach would involve weighting number of students in school.
state_centers <- schools_2018 %>%
  mutate(fips = as.integer(fips)) %>% # ensure fips is int
  group_by(fips) %>%
  summarize(
    state_lat = mean(latitude, na.rm = TRUE),
    state_lon = mean(longitud, na.rm = TRUE)
  )

# 2. Combine Data and Calculate Distances
# -----------------------------------------------------------------------------

# Use 'state_flows' data frame from Objective 1, merge on origin and dest coords
distance_data <- state_flows %>%
  # Join state center coordinates for the student's origin
  left_join(state_centers, by = c("fips_origin" = "fips")) %>%
  # Join school coordinates for the destination
  left_join(school_coords, by = "unitid") %>%
  # Ensure valid coordinates for the calculation
  filter(!is.na(state_lat) & !is.na(latitude))

# Haversine distance for each flow
# (for calculating shortest distance between spherical coordinates)
distance_data <- distance_data %>%
  mutate(
    distance_miles = distHaversine(
      p1 = cbind(state_lon, state_lat),
      p2 = cbind(longitud, latitude)
    ) / 1609.34 # meters to miles
  )

```

```r
# 3. Overlapping Density Plot for 4-year Institutions
# ---------------------------------------------------------------------------

# Add labels for our plots
# This mapping is standard for IPEDS data
distance_data <- distance_data %>%
  mutate(school_type = case_when(
    sector == 1 ~ "Public, 4-year",
    sector == 2 ~ "Private nonprofit, 4-year",
    sector == 3 ~ "Private for-profit, 4-year",
    sector == 4 ~ "Public, 2-year",
    sector == 5 ~ "Private nonprofit, 2-year",
    sector == 6 ~ "Private for-profit, 2-year",
    TRUE ~ "Other"
  )) %>%
  # Focus on the major 4-year institutions
  filter(sector %in% c(1, 2))

# Create the overlapping density plot
distance_plot <- ggplot(distance_data, aes(x = distance_miles, fill = school_type)) +
  geom_density(alpha = 0.6) +
  # A log scale is crucial for seeing the skewed distance data clearly
  scale_x_log10(breaks = c(10, 100, 250, 500, 1000, 2000, 4000)) +
  labs(
    title = "Distribution of Student Travel Distances by School Type",
    subtitle = "For first-year students attending 4-year institutions (2018)",
    x = "Travel Distance (miles, log scale)",
    y = "Density",
    fill = "School Type"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```r

ggsave(
  filename = "./output/distance_distribution_4year_log.png",
  plot = distance_plot,
  width = 11,    # Inches
  height = 7,    # Inches
  dpi = 300      # Dots per inch (resolution)
)

# 4. Histograms for all School Types
# ---------------------------------------------------------------------------

# Define desired order
school_type_order <- c(
  "Private nonprofit, 4-year",
  "Public, 4-year",
  "Private nonprofit, 2-year",
  "Public, 2-year",
  "Private for-profit, 4-year",
  "Private for-profit, 2-year",
  "Other"
)

# Add labels
distance_data_all_types <- distance_data_all_types %>%
  mutate(school_type = case_when(
    sector == 1 ~ "Public, 4-year",
    sector == 2 ~ "Private nonprofit, 4-year",
    sector == 3 ~ "Private for-profit, 4-year",
    sector == 4 ~ "Public, 2-year",
    sector == 5 ~ "Private nonprofit, 2-year",
    sector == 6 ~ "Private for-profit, 2-year",
    TRUE ~ "Other" # Catches any other sector codes (e.g., 0, 7, 8, 9)
  ),
  school_type = factor(school_type, levels = school_type_order)
  )

# Create histogram
distance_plot_all_types <- ggplot(distance_data_all_types, aes(x = distance_miles)) +
  geom_histogram(bins = 50, fill = "#0072B2") +
  # separate plots for each school type, arranged in 3 columns
  facet_wrap(~ school_type, ncol = 3, scales = "free_y") +
  # log scale for the x-axis for better visibility
  scale_x_log10(breaks = c(10, 50, 250, 1000, 4000), labels = scales::comma) +
  labs(
    title = "Distribution of Student Travel Distances for All School Types",
    subtitle = "First-year college students, 2018",
    x = "Travel Distance (miles, log scale)",
    y = "Number of Student Flows"
  ) +
  theme_minimal() +
  # improve readability
  theme(strip.text = element_text(face = "bold"))

ggsave(
  filename = "./output/distance_distribution_all_schools_log.png",
  plot = distance_plot_all_types,
  width = 11,
  height = 7,
  dpi = 300
)

# ############################################################################
# Objective 4: Determinants of Travel Distance
# ############################################################################

# 1. Data Setup
# ---------------------------------------------------------------------------

# Additional file from IPEDS for tuition rates
glimpse(tuition_data)

tuition_clean <- tuition_data %>%
  select(unitid = UNITID, tuitionfee_in = TUITION2) %>% #TUITION2 is in-state avg
  mutate(unitid = as.character(unitid))

schools_2018 <- schools_2018 %>%
  left_join(tuition_clean, by = "unitid")

# 2. Calculate Average In-State Tuition by State
# ---------------------------------------------------------------------------

home_state_tuition <- schools_2018 %>%
  filter(sector == 1) %>% # Public, 4-year schools
  mutate(fips = as.integer(fips)) %>%
  mutate(tuitionfee_in = na_if(tuitionfee_in, ".")) %>%
  mutate(tuitionfee_in = parse_number(tuitionfee_in)) %>%
  group_by(fips) %>%
  summarize(
    avg_home_tuition = mean(tuitionfee_in, na.rm = TRUE)
  ) %>%
  filter(!is.na(avg_home_tuition))

# 3. Merge Tuition Data into the Main Analysis Frame
# ---------------------------------------------------------------------------

analysis_data <- distance_data %>%
  left_join(home_state_tuition, by = c("fips_origin" = "fips")) %>%
  filter(!is.na(avg_home_tuition))

# 4. Scatter Plot
# ----------------------------------------------------------------------------

state_level_summary <- analysis_data %>%
  group_by(fips_origin, avg_home_tuition) %>%
  summarize(avg_distance = mean(distance_miles, na.rm = TRUE))

ggplot(state_level_summary, aes(x = avg_home_tuition, y = avg_distance)) +
  geom_point(alpha = 0.7, color = "#0072B2") +
  geom_smooth(method = "lm", se = FALSE, color = "#D55E00") +
  scale_x_continuous(labels = scales::dollar) +
  labs(
    title = "Higher Home-State Tuition is Correlated with Farther Travel",
    subtitle = "Each point represents a US state (2018)",
    x = "Average In-State Tuition at Public 4-Year Universities",
    y = "Average Travel Distance for Students from that State (miles)"
  ) +
  theme_minimal()

# Create the plot and assign it to a variable
tuition_plot <- ggplot(state_level_summary, aes(x = avg_home_tuition, y = avg_distance)) +
  geom_point(alpha = 0.7, color = "#0072B2") +
  geom_smooth(method = "lm", se = FALSE, color = "#D55E00") +
  scale_x_continuous(labels = scales::dollar) +
  labs(
    title = "Higher In-State Tuition and Travel Distance",
    subtitle = "Each point represents a US state (2018)",
    x = "Average In-State Tuition at Public 4-Year Universities",
```

```r
    y = "Average Travel Distance for Students from that State (miles)"
  ) +
  theme_minimal()

ggsave(
  filename = "./output/tuition_distance_plot.png",
  plot = tuition_plot,
  width = 8,
  height = 6,
  dpi = 300
)

# 7. Regression Analysis
# ---------------------------------------------------------------------------

tuition_model <- lm(log(distance_miles) ~ avg_home_tuition, data = analysis_data)
summary(tuition_model)

stargazer(
  tuition_model,
  type = "latex",                                 # Specify LaTeX output
  title = "Determinants of Student Travel Distance",
  dep.var.labels = "Log(Travel Distance in Miles)", # Clean name for the dependent variable
  covariate.labels = "Avg. Home-State In-State Tuition", # Clean name for your variable
  header = FALSE,                                 # Removes extra LaTeX preamble
  out = "./output/tuition_regression.tex"        # The output file
)

# 8. Control for state median incomes
# ----------------------------------------------------------------------------

# Data is from the U.S. Census Bureau, American Community Survey (Table S1901)
state_income_data <- tibble(
  fips = c(1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56),
  median_income = c(50536, 77640, 58945, 47597, 75235, 72331, 78833, 68287, 86420, 55660, 58700, 81275, 55785, 65886, 56303, 60523, 59597, 50586, 49468, 56277, 84805, 85843, 57144, 71306, 45081, 55461, 54970, 59970, 58646, 76768, 82545, 49754, 68486, 54602, 64577, 56207, 52919, 62818, 61744, 64996, 53199, 58275, 53320, 61874, 71621, 61965, 74222, 73775, 51340, 61747, 61584)
)

# Merge income data into the main analysis Frame
analysis_data_controlled <- analysis_data %>%
  left_join(state_income_data, by = c("fips_origin" = "fips")) %>%
  filter(!is.na(median_income))

# 9. Run Both Regression Models
# ----------------------------------------------------------------------------

model_simple <- lm(log(distance_miles) ~ avg_home_tuition, data = analysis_data_controlled)
summary(model_simple)

model_controlled <- lm(log(distance_miles) ~ avg_home_tuition + median_income, data = analysis_data_controlled)
summary(model_controlled)

stargazer(
  model_simple, model_controlled,
  type = "latex",
  title = "Impact of Tuition on Travel Distance, with and without Income Control",
  dep.var.labels = "Log(Travel Distance in Miles)",
  covariate.labels = c("Avg. Home-State Tuition", "State Median Income"),
  header = FALSE,
  column.labels = c("Simple Model", "With Income Control"),
  out = "./output/tuition_regression_controlled.tex"
)
```