**Tuition Hikes versus Inflation Rates,**

**Enrollment of Underrepresented Minorities at Top-Ranked Universities,**

**and Other Investigations**

**An Exploration and Analysis of U.S. Higher Education in Python**

Chart, line chart

Description automatically generated

Elizabeth Axworthy

Samuel Deery-Schmitt

Syracuse University

M.S., Applied Data Science Program

**I. INTRODUCTION**

**High Level Summary**

A college education is associated with higher lifetime earnings. However, the cost of attending college has risen dramatically over the last several decades. Before financial aid, the cost of attending top-ranked private research universities approaches $80,000 – a year at Stanford costs approximately $74,000, and a year at Harvard is approximately $78,000- and the cost of tuition alone exceeds $50,000 annually at such institutions.

At the same time, colleges and universities have become more selective over time, in part because the internet has made it far easier to apply to college than in the days of paper applications, so schools are receiving more applications than ever. This is especially true at the top-ranked schools, where admissions rates below 10% are not uncommon (Stanford and Harvard each admitted less than 5% of applicants in 2020). Given the importance of a college education, and especially given the career-earning potential of attending a top-ranked university, it is important that these institutions offer opportunities to underrepresented minorities and women.

**Context and Scope of Analysis**

The goal of this project is two-fold: analyze diversity at colleges and universities in the United States, and investigate rising tuition costs. We will look at demographic data at several different levels, and compare tuition cost to the rate of inflation. We will limit our analysis to relatively few variables. Our demographic dataset is limited to 2014, while the rest of the datasets generally cover several decades.

**Research Questions**

1. Lizzy: **What was average demographic university makeup, grouped by state, in 2014?** I will use the diversity\_school.csv file from kaggle for this, and the fields [‘state’, ‘category’, ‘enrollment’, and ‘total enrollment’].
2. Lizzy**: How has the nationwide average rate of tuition increased compared with the inflation rate?** Schools in the category of 2-Year, 4-Year and All. Inflation rate being both measured in the difference from the first measured year (1985) and difference from the preceding year. From the original data sources, I use historical\_tuition.csv with all the fields [‘type’, ‘year’, ‘tuition\_type’, and ‘tuition\_cost’]. Type referred to whether the school is private, public, or the combined average as ‘All Institutions’, which tuition\_type is whether the school is a 2-year, 4-year, or combined average of all schools, and whether or not the dollar value is given in a constant of today’s value, or in the contemporary value for that year. I also use cpi.csv with the fields [‘year’, ‘annual’], and table.csv which I reformat to follow historical\_tuition’s format, ultimately ending with fields [‘year’, ‘type’, ‘tuition\_2’, ‘tuition\_type’].
3. Sam: **What does the proportion of minority students and the proportion of female students in 2014 look like at the regional level, and at the top-25 ranked schools?** Minority students are defined in the dataset diversity\_school.csv from Kaggle as students who identify as the following: American Indian/ Alaskan Native, Asian, Black, Hispanic, Native Hawaiian / Pacific Islander, and two or more races. To perform this analysis, I used the fields ‘name’, ‘state’, ‘category’, ‘enrollment’, ‘total enrollment’, and created calculated fields for ‘region’, ‘ranking’, ‘proportion minority’, and ‘proportion women’. I used top\_25.csv for the rankings, which I created by modifying another dataset, which I will discuss in more detail in the next section.
4. Sam: **Are there significant demographic differences when we group schools by region or by ranking?** I used the same datasets I cleaned in question 3 for this analysis.
   1. Is there a significant difference between:
      * 1. the mean proportion of minority students at top-25 nationally ranked research universities and at all schools?
        2. the mean proportion of female students at top-25 nationally ranked research universities and at all schools?
      1. To answer these questions, I performed Welch’s two-sample t-test for samples of unequal sizes.
   2. Is the relationship between region and diversity for the top-25 nationally ranked research universities independent?
      1. To answer this question, I performed two chi-square tests of independence, one for minority enrollment and one for female enrollment. I calculated my expected frequencies by using the enrollment data for all schools.

**Group Roles**

Both members of the group participated equally. We selected our initial set of data sources together, then found additional sources to aid in our research questions independently. We answered own research questions on our own, then worked together on compiling the report and presentation.

**II. DATA**

**Sources:**

1. The College Tuition, Diversity, and Pay dataset, found on Kaggle. It has 5 .csvs containing information about tuition costs, historical tuition rates, salary potentials, student income brackets, and diversity of student populations. There are a total of a total of 33 columns and 263,845 rows; we are particularly interested in the following variables: minority enrollment, tuition cost by year, total in-state and out-of-state cost, in-state and out-of-state tuition, income level, and mid-career pay.

[<https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay>]

2. To examine the tuition costs and inflation, we will use Tuition Costs of Colleges and Universities from the U.S. Department of Education, National Center for Education Statistics Tuition Costs (62 observations of 6 variables) which is itself derived from the much more extensive Average undergraduate tuition and fees and room and board rates charged for full-time students in degree-granting postsecondary institutions, by level and control of institution: Selected years, 1963-64 through 2018-19 (200 observations of 25 variables), also from the U.S. Department of Education, National Center for Education Statistics Tuition Costs.

[<https://nces.ed.gov/fastfacts/display.asp?id=76>]

[<https://nces.ed.gov/programs/digest/d19/tables/dt19_330.10.asp>]

3. For income data, we will use U.S. Census data from the following source: Information on income at a state level from the U.S. census: Median Household Income by State: 1984 - 2018 (102 observations of 68 variables) and supplement with additional data from the U.S. Census as necessary. The results of [this search](https://www.census.gov/search-results.html?q=median+household+income+by+state&page=1&stateGeo=none&searchtype=web&cssp=SERP) will provide any additional datasets we need.

[<https://www2.census.gov/programs-surveys/cps/tables/time-series/historical-income-households/h08.xls>]

[<https://www.census.gov/search-results.html?q=median+household+income+by+state&page=1&stateGeo=none&searchtype=web&cssp=SERP>]

4. We will also add inflation data to our analysis, as the previous dataset only captured tuition adjusted for inflation, not inflation itself. The Consumer Price Index Data from 1913 to 2021 (108 observations of 15 variables) is derived from the Bureau of Labor Statistics dataset (108 observations of 15 variables) and we will incorporate the source that we deem the most user-friendly.

[<https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/>]

[<https://data.bls.gov/timeseries/CUUR0000SA0>]

5. We will also use the US News & World Report’s 2021 rankings for the top 25 national research universities. While quantifying if one elite school is “better” may have subjective components, certain objective benchmarks such as student loan debt after graduation and student-to-faculty ratio can be used to rank one school over another, and the US News & World Report has been doing this for 37 years. The 2022 rankings were only recently reported, which is why we use the 2021 rankings. I (Sam) created a .csv called top\_25.csv that contains the top 25 nationally ranked research universities, their state, region, and ranking, by using yet another dataset from the U.S. Department of Education, the Database of Postsecondary Institutions and Programs, a massive dataset with more than 30,000 observations (one for each accredited postsecondary school or program in the U.S.). Cleaning method will be discussed in preprocessing section.

[<https://www.usnews.com/best-colleges/rankings/national-universities>]

[<https://ope.ed.gov/dapip/#/home>]

**Preprocessing**

Lizzy: I turned the cpi data I would need into a csv for ease of work. The results of this can be found in the Data folder, under cpi.csv, which hold the monthly, half yearly, and annual cpi for all years from 1985-2021. I also scraped the table from the page listed as source 2, Tuition Costs of Colleges and Universities from the U.S. Department of Education, National Center for Education Statistics Tuition Costs and turned that into a csv. The process of doing that is contained in ‘table\_scrape.ipynb’ in the main folder and involved using beautiful soup to grab the table, then the contained table row elements. Csv construction was done using string methods to format the data. Output saved to ‘table.csv’, found in the Data folder.

Sam: I added rankings and regions to the Database of Postsecondary Institutions and Programs, then removed fields and observations so it only contained the top-25 ranked national research universities along with their state, region, and ranking prior to loading it into Python as a pandas DataFrame. I merged this with the diversity\_school dataset, which I also loaded in as a pandas Dataframe, in order to make a subset that included only the top-25 ranked schools. Before loading this dataset, I checked that university names were spelled consistently, as I encountered numerous inconsistencies between the two: for example, University of California, Berkeley, a top-25 ranked school, can also be represented as “University of California – Berkeley” or “University of California: Berkeley”. I also added calculated fields to the diversity\_school dataset for proportion minority and proportion women enrollment, and transformed the dataset so it contained one observation per school with only target demographic data, instead of 11 observations for each school with one piece of demographic data in each. I did the same for the subset. I also removed unnecessary columns, such as ‘category’.

**III. ANALYSIS**

**Program Description:**

Lizzy - Questions 1 & 2 - Both questions 1 & 2 are explored in the notebook, ‘notebook\_lizzy.ipynb’.

* I start with imports and loading the datasets I’ll be using (the 5 kaggle csvs, table.csv, and cpi.csv) into a dictionary of dataframes called datasets.
* I then do some further data cleaning by changing an incorrect year in table and replacing an odd character in the year column with a standard ‘-’. I then extract tuition cost by tuition type and institution type into individual series before combining the series into one large dataframe that matches the format in historical\_tuition and confirm no NAs are present. I name the tuition costs given here ‘tuition\_2’ and combine the results with historical\_tuition to form the hist\_tuition dataframe. Lastly I make a column called avg\_tuition from the average of the two tuition values, as the difference between them is too large to justify throwing one out as redundant, and both data sources are reliable.
* Question 1, Gender Diversity by State Analysis: I grab the fields I need from kaggle’s diversity\_school, [total\_enrollment, state, category, and enrollment] and save as a dataframe called div\_df. I then group div\_df by state and the category of the student population they’re measuring so that for each state I have the category, that category’s enrollment, and the total enrollment. Then I create a new column called percentage to hold the percent of the total population that category forms. I grab the unique categories and create a dictionary of dateframes, one for each category that isn’t gender based, and populate it with the output of the describe function for that category. This gives me the average enrollment for that category across all states, as well as the other basic descriptive stats. I then graph the data and save the output into the images folder, under ‘racial\_diversity\_2014’. I then graph gender distribution by state, with gender counted as either Women or Other, as only the number of women are included in the dataset and I don’t believe they were measuring non-binary answers in 2014, so I can’t be sure that the remaining percent were all men. The graph is saved to the images folder, under gender\_diversity\_2014.
* Question 2, Tuition and Inflation Analysis: I copy the year and annual cpi fields from the cpi dataframe into a smaller frame called cpi\_df and rename the columns so that year can be joined with hist\_tuition and change ‘Annual’ to cpi. Next I create a column called 1\_year\_change to hold the percent different between one years cpi and the next and a total\_change column to hold the different in total values since 1985, the first year measured. I set the values manually for the first row as they don’t have previous values to compare with, drop rows for years not present in hist\_tuition, and rename the years to follow the same format. I have to break hist\_tuition back into series to calculate the percentage changes and total for each tuition and institution type based on the avg\_tuition value. I then join cpi\_df and hist\_df into tuition\_vs\_inflation, my final data frame organized by year, tuition type, and institution type, which I save as a csv ‘tuition\_vs\_inflation.csv’. All further calculations are done for the averaged ‘All Institutions’ and ‘Constant All’ tuition values. I then graph the value of cpi vs the cost of tuition, using log to normalize the values and make for meaningful comparison given the different in units. I chart the percent differences for CPI and Tuition against each other and run some descriptive stats on the one year change fields. I then chart the difference in the percentage of the change in both categories against each other to get an idea of their correlation, and check descriptive stats. Lastly, I take a look at the first and last measured years against each other.

Sam – Questions 3 & 4 – Both questions 3 & 4 are explored in the notebook ‘sds\_demographic\_analysis.ipynb’

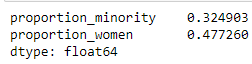
* First I import the datasets as pandas DataFrames. These datasets include all 5 Kaggle .csvs, and my top\_25.csv. At this point, I have already changed the name of the top 25 schools in every dataset to a uniform spelling directly in Excel (to match the U.S. Department of Education’s school names in their database). This applies only to a handful of schools, like Berkeley, UCLA, and Washington University in St. Louis. Attempting to change the name of every school in every dataset instead of just the 6 mismatches in top-25 could potentially be a project in and of itself unless there were another common field that could be used to match schools.
* After viewing all the datasets to better understand them, I decide that I will focus on the diversity\_school for my demographic analysis, so I merge the top\_25 dataset with the diversity\_school dataset (which I have renamed diversity\_school\_2014 to reflect that it contains data only for that year. Upon completing this merge (the equivalent of an inner JOIN in SQL), I discover that, for whatever, reason, Georgetown, a top-25 ranked school, is not present in the diversity\_school\_2014 data. I make note of this and continue on, adding the calculated field ‘proportion’ by dividing ‘enrollment’ by ‘total\_enrollment’. This way give the percentage of students at a given school who identify as female, or who identify as racial minorities as defined in the dataset.
* After viewing the merged dataset, I delete unnecessary duplicate columns and then do additional cleaning. As it stands, there are 11 rows for each school, and I want one row. This table is actually quite well-normalized, but I don’t need most of that information. I am only interested in two out of the eleven observations for each school (‘Women’ and ‘Total Minority’ so I opt to collapse the table. After reading pandas documentation, Stack Overflow, and more, I could not find an easy way to use pandas methods to do this, although I am sure there is one , so I created a few ‘for:’ loops to do it for me that conditionally iterated over the table by specific index and created a list of total minority proportions and total women proportions.
* Next I created DataFrames from these lists, and merged them with the master dataset in such a way (inner JOIN) as to only capture a single row for each school. Following that, I deleted the now unnecessary ‘proportion’ column as well as ‘category’ and ‘enrollment’. As it turns out, ‘enrollment’ would be necessary to do chi-square tests later, so this was a mistake, and I would have some time if I had avoided deleting this column because I ended up having to recalculate it.

The final table for top-25 ranked national research institutions is displayed on the next page:

Table

Description automatically generated

* + Regions: M = Midwest, N = Northeast, S = South, W = West, O = Other (Hawaii and Alaska mainly, which do not appear in the top 25)
* Next, I looked at the mean values for the two fields of interest, yielding these results:



* After that, I wanted to aggregate by region and include the mean proportion minority and mean proportion women for each region. As it turns out, this is straightforward, but I did not know the simple solution until after I had already written a lengthy series of steps to conditionally iterate over the table. I included the simple solution later on as I had already written this code. First I made this table by grouping by region.

Graphical user interface, application

Description automatically generated

* + Then I created a list for each cell I wanted to fill in (the proportion minority and proportion women for each region) and conditionally iterated over the DataFrame to fill the lists with the appropriate values.
  + Then I made a function to take the mean of each of those lists and return it as another list. I ran this twice to get a list of regional means for proportion minority and proportion women.
  + Finally, I added these lists to the DataFrame, to yield this table, displayed on the next page:

Table

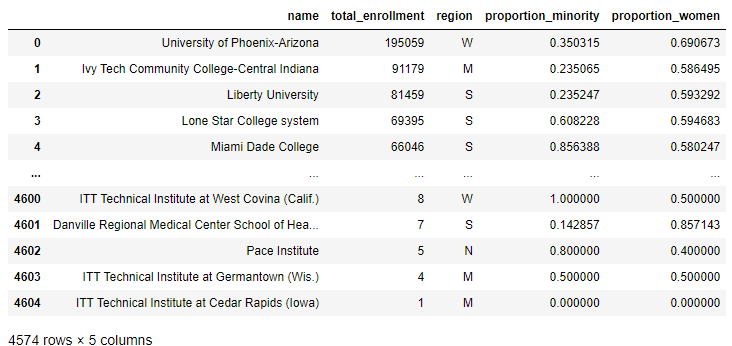
Description automatically generated

* + As much of an adventure as it was to figure out how to do this creatively, I later discovered that DataFrame.groupby([‘region’]).mean() yields a nearly identical result:

Table

Description automatically generated

* + The unnecessary columns can be dropped or deleted, and the counts can be separately calculated and added in. Live and learn.
* At this point I created a histogram of both proportion\_minority and proportion\_women; these outputs will be displayed and discussed in the conclusions section.
* In order to compare findings about nationally-ranked universities to all schools, and do a regional analysis, I needed to take similar steps to modify the diversity\_school\_2014 dataset, collapsing it to include one observation per school, added calculated fields, adding regions, and deleting unnecessary columns.
  + I used the same code as before, with one exception: I had to add regions to every school, and I had regions for the top-25 schools, which I had previously added to the top\_25.csv.
  + To do this, I had to use two dictionaries, one with ‘State’ : Abbrevation key : value pairs and one with ‘Abbreviation’ : ‘ Region’ key : value pairs. I could have switched the keys or values around in either dictionary, but it was just as easy to just use both and iterate over them.
  + I iterated over the state names in the DataFrame, storing the corresponding state abbreviation in a list. Then I iterated over the list to retrieve the region for each state abbreviation, storing those results in list, which I added to the DataFrame as region.
  + I noticed there were some regions that came up as unknown (‘U’), so I investigated and found that there were 31 observations with numerous NAs in fields including state and name, so I opted to drop all the observations with NAs as it was a relatively small number (31 out of about 4,600].
  + Here is the final table that I used for analysis, presented on the next page:



* Now I created histograms (to be displayed and discussed in the conclusions) and calculated means:

Text

Description automatically generated

* And finally I grouped by region, using my lengthy aggregation and retrieval method that I already documented. Here is the result:

Table

Description automatically generated

* I could effectively answer research question 1 at this point and now moved onto statistical analysis. I performed two chi-square tests of independence, and two Welch’s t-tests. I had take additional steps to get the data prepared to run the tests, but once I had the data in the right form, executing the tests was straightforward.
* For the chi-square tests, I needed to create contingency tables with observed frequencies – the actual enrollment numbers for women and minorities at the top-25 schools by region – and expected frequencies – the enrollment we would expect if there were no difference in diversity at top-25 schools vs all schools, which is calculated by taking the total minority enrollment at all schools by region and dividing by the total enrollment at all schools by region, then multiplying this value by the total enrollment by region at the top-25 schools. This calculation is repeated for female enrollment.
  + Note that this is different than multiplying by the mean proportion minority and mean proportion women; doing that does not yield the correct expected counts, as these values are means of the proportion admitted and do not reflect the overall proportion admitted for each category.
* Unfortunately, I had removed the counts of enrollment from my DataFrames, so I had to make a new DataFrame, ‘expected’, where I added the counts back in and used those to generate the expected frequencies. I similarly accessed the observed enrollment data, stored it in a DataFrame, and added the expected frequencies to it. Here’s the final table with the frequencies (note that this is not a true contingency table, but all we need for the chi-square test are observed and expected values; scipi.stats takes care of the rest):

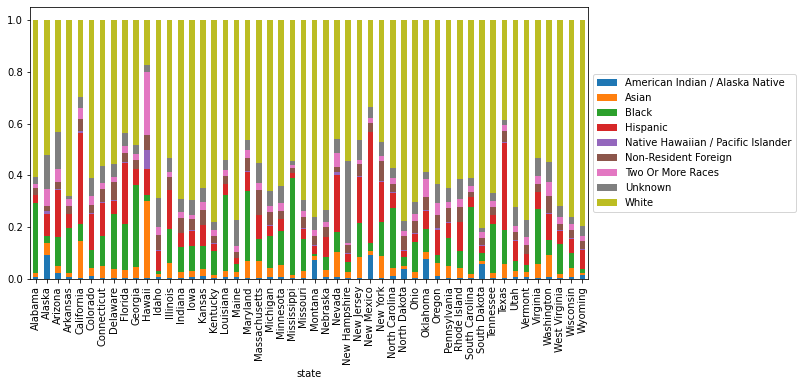
Table

Description automatically generated with low confidence

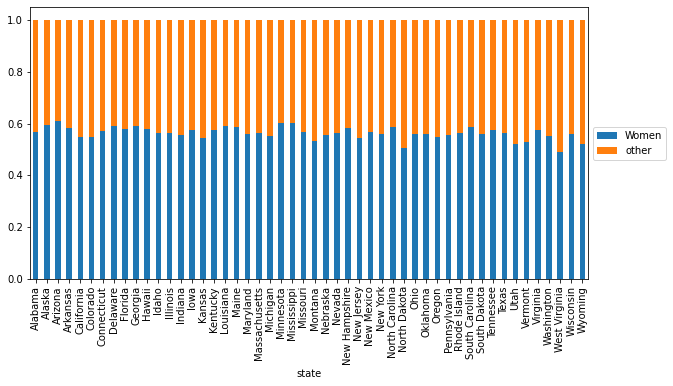
* From here, I was able to perform my chi-square tests. Both yielded statistically significant results, and I will discuss them in the conclusions section.
* The final analyses were running two Welch’s t-tests, which did not require any additional data manipulation. Both yielded statistically significant results and will be discussed in conclusions.

**IV. CONCLUSIONS**

**Research Q1: How Does Diversity on College Campuses in 2014 Vary By State?**

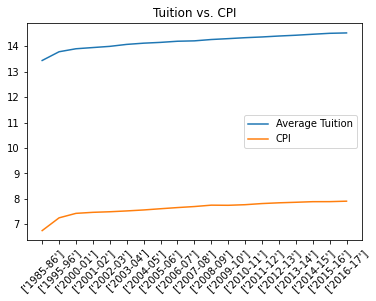
For most states, the majority of students were white, with some exceptions being Arizona, California, Florida, Hawaii, Maryland, Nevada, New Mexico, New York, and Texas. The largest proportions of Hispanic students were found in states that were historically part of Mexico, like California, New Mexico, and Texas. Asian students are found in the greatest proportion in Hawaii, and also the greatest proportion of Native Hawaiian students, who are the least represented overall. Black students are more prominent in southern schools, such as Mississippi, but also Maryland. Native American students are most prominent in Alaska, Montana, New Mexico, Oklahoma, and South Dakota.

Further research could take advantage of the dictionary of dataframe I created for combined state demographic stats for each demographic.

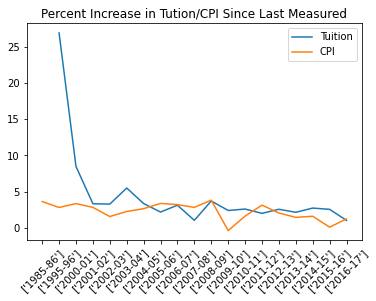


Across all states, approximately 50% of students identified as women, with the smallest number being found in West Virginia and the greatest number being found in Minnesota and Mississippi. There seem to be no geographic trends for gender-based enrollment in 2014.

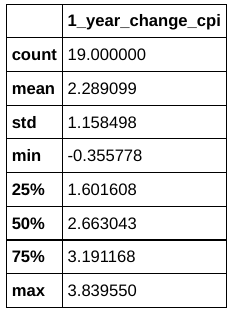
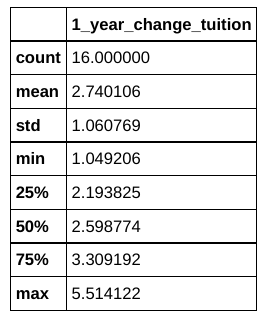
**Research Q2: Is Tuition Changing at the Same Rate as Inflation?**

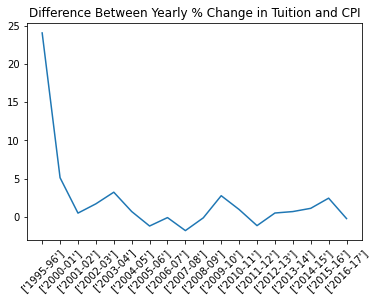


Comparing the logged value of CPI against tuition, their growth rates appear about the same, linear with a comparable slight slope. The hook at the beginning is caused by compressed time period.

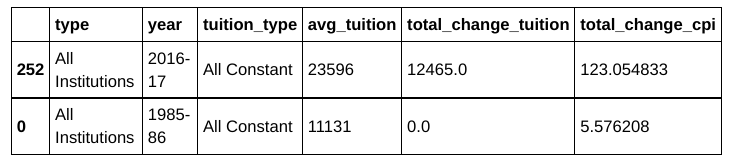


Examining the percentage increases of cpi and tuition against each other, we find the tuition always increases by a positive percent, while CPI can decrease some years. The magnitude of the changes do not appear to be tightly related. The median change for both measures was 2.6% per year, with tuition having a standard deviation of 1.1%, and cpi a standard deviation of 1.26%. The maximum change for tuition, discounting outliers, was 5.5%, and for CPI it was 3.84%. Minimums were 1% for tuition and –0.36% for CPI.





A positive value here indicates the tuition changed by a greater percent than CPI. Neagtive represents the inverse. The median value of the difference was 0.6 with a standard deviation of 1.4, so we can say that tuition usually increases slightly more than CPI, though this is not always true.



I had though that this meant I could conclude with tuition increasing at a rate higher than inflation, but looking at the final year measured, 2016-17, tuition costs had gone up by $12,465 (constant) from an original value of $11,131, or by 89% since 1985 while CPI has increased by 123%. If tuition had followed at the inflation rate it would be $24,822. This would indicate the tuition prices are increasing slower than inflation. Reasons for the differences between what the graphs seem to indicate and this finding are that I was running statistics for the average of all schools across private and public colleges and 2 year and 4 year colleges. Looking at the categories individually could reveal different growth rates, as how those different type of institutions handle their tuition is varied. There is also gaps in my knowledge about the specifics of inflation and it’s effects. Further research would include exploring the different categories and doing more background research to make sure I fully understand the context of what I’m seeing.

**Research Q3: What does the proportion of minority students and the proportion of female students in 2014 look like at the regional level, and at the top-25 ranked schools?**

Text

Description automatically generated

Text

Description automatically generated

*Top 25 Nationally Ranked Research Universities\**  *All Postsecondary Institutions\*\**

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

\* Rankings based off 2021 U.S. News & World Report National University Rankings;

Georgetown University (#23) ommited from dataset

\*\* Dataset includes 4,574 postsecondary institutions and programs in the United States

**Python:**

Text

Description automatically generated

**Interpretations:**

**Mean Proportions**

On average, a national research university ranked in the top-25 in 2021 enrolled 32.49% minority students and 47.72% female students in 2014.

On average, a U.S. postsecondary institution in this sample enrolled 37.36% minority students and 58.29% female students in 2014.

It appears that the top-ranked schools enroll approximately 5% less underrepresented minority students and 10% less female students than the national averages. Welch’s two-sample t-test will help determine if this is a significant difference.

One potential issue could be the weight that all-women colleges place on the national average. A subsequent analysis that segments schools using dummy variables for all-women, all-men, or both would help determine if this is a factor. Additionally, it would be interesting to look at only the 4-year universities in the sample, as the sample also includes community colleges, technical schools, liberal arts colleges, and other schools that may have significantly different demographics than nationally ranked research institutions, and research institutions in general. Nonetheless, these differences are intriguing. The histograms help us understand the differences in more detail.

**Distributions of Proportions**

Neither distribution of enrolled minority students is normally distributed; both are skewed to the right. The ‘all schools’ distribution has a wider range that extends from 0% to 100%, and more significantly, the top-25 schools distribution has a much narrower range, from ~20% - ~50%, which indicates that no historically Black universities are in the top-25. This is also true for the distributions of female students – no all-women schools are ranked in the top-25. Perhaps if I expanded the rankings to include top-ranked liberal arts schools, this may change. Of significance is the fact that the mode of the distribution of enrolled minority students at top-25 schools is in the 30%-35% bin. I am unsure why that would be, but with 9 schools in this bin, it is by far the most commonly occurring bin in the histogram. What it does indicate is that a plurality of top-ranked universities enroll approximately 1/3 minority students, which, according to the latest census data (<https://www.census.gov/quickfacts/fact/table/US/PST045219>) appears to be less than the national proportion.

Both distributions of enrolled female students are skewed to the left. It is important to note that the mean enrollment at the top-ranked schools is less than 50%, and is approximately 10% less than the mean enrollment for all schools in the sample.

This has two implications: first of all, less women than men are enrolled, on average, at the top-ranked research intuitions. This is not ideal as the top-ranked universities also have top-ranked science, technology, engineering, and math (STEM) programs, fields where women are traditionally underrepresented (as are minority groups). To change the professional landscape, more women and underrepresented minorities need to have the opportunity to earn STEM degrees, especially from the best programs in the country.

Second, as previously mentioned, there are no all-women schools in the top-25, as the range does not extend past the 60% bin. This may be because there are very few all-women research universities; many all-women schools are four-year colleges. Further analysis could reveal whether this is the case. If that were true, it raises the question: why aren’t there any (or more) all-women research universities? This would be an interesting avenue to pursue in the future, as there may be many factors involved.

I next investigated if these observed differences were statistically significant, and also analyzed the differences on the basis of region of the country.

**Research Q4: Are there significant demographic differences when we group schools by region or by ranking?**

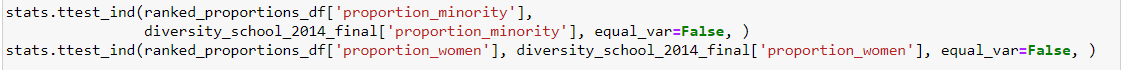
Text

Description automatically generated

A picture containing text

Description automatically generated

**Python**:



Note that scipy.stats was imported as stats.

**Interpretations:**

**Welch’s Two-Sample T-Tests**

I observed that the mean enrollment for both women and underrepresented minorities appeared lower at the top-25 schools than at all schools, and the statistical tests confirmed a statistically significant difference (alpha = 0.05). I used Welch’s 2-sample t-test instead of a normal 2-sample t-test because the samples were drastically different in size (24 vs nearly 5,000) and Welch’s test is for precisely such scenarios when you cannot assume equal variance in samples because of large differences in size. This confirms that there is indeed a discrepancy worth investigating in more detail in the future.

**Chi-Square Tests of Independence**

The chi-square tests of independence revealed more detail about the distribution of the enrollments, which we have identified as significantly lower than the national averages. The chi-square tests reveal that there is a relationship between region and minority enrollment, and region and female enrollment at the top-25 ranked schools. This indicates that top-ranked schools in some regions of the country admit less women and underrepresented minority students than schools in other regions of the country. Schools in all regions except the Midwest have less minority enrollment than the regional averages, which is surprising because all the top-25 Midwest schools admit less than 30% underrepresented minority students. Looking at the regional average for all schools, however, the average Midwest schools enrolls just over 26% minority students, so this finding makes more sense. A natural follow-up question, though, is why is this? This would be a complex, interesting, and hopefully impactful investigation that would certainly be worth pursuing in the future.