

Final Project

Xi Cao

4/29/2020

Project description:

In this project, we will deal with some data related to America tuition costs in each state and then find out some correlation between the tuition costs and other relative variables.

Github repository: <https://github.com/xicao143/data-wrangling> (<https://github.com/xicao143/data-wrangling>)

1. Data import and data scraped

1.(a) Read data from the first data source. As data in this source has already been cleaned, so we do not have to clean them any more. we import these data from

<https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-03-10>

(<https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-03-10>), also, I export them as csv files and upload them to my repository.

1.(b) Read data from the second source: <https://www.collegetuitioncompare.com/state/>

(<https://www.collegetuitioncompare.com/state/>). Th data read from this source is the tuition fees and living costs by state. This data source need to do some cleaning and change some colnames. We deleted some rows and colnames and renamed the colname names. Export it as a csv file.

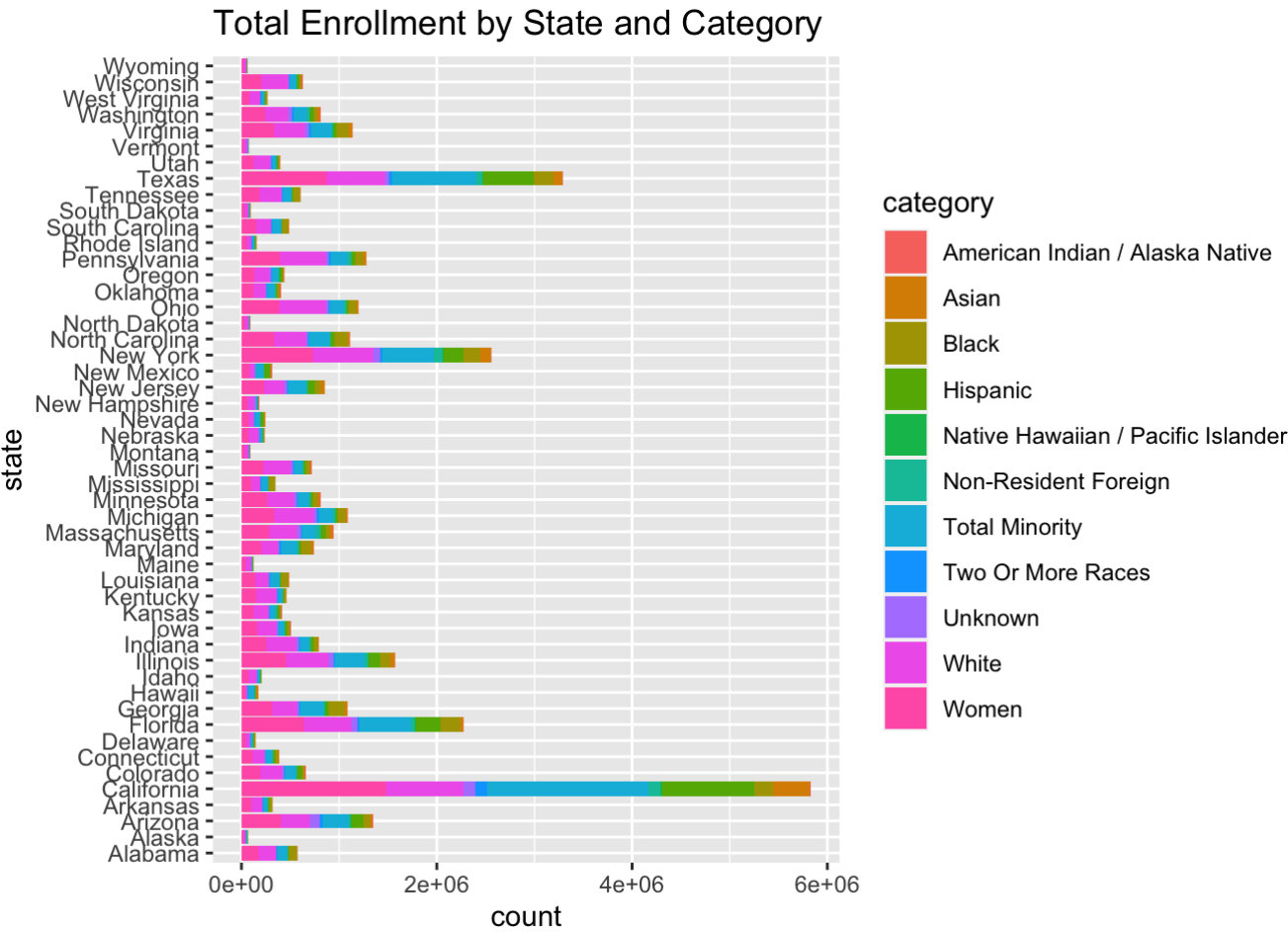
1.(c) Read more specific tuition costs from the website of each state by a loop and store these table into a list.

1.(d) Changed the specific data collected from each state to dataframes and select data according to different tuition costs types.

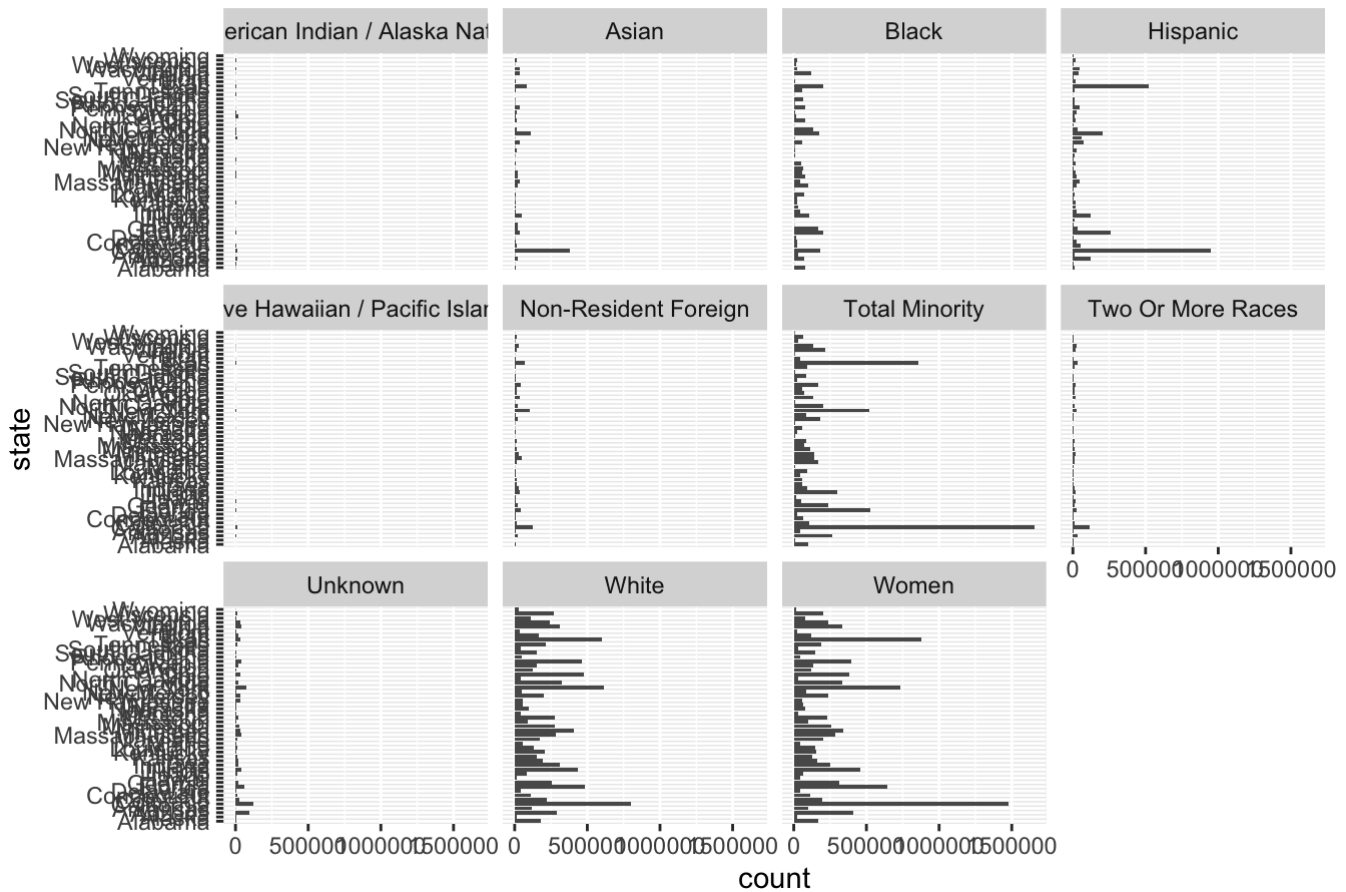
2. Diversity_school wrangling

In the part, we dealt with the diversity_school table to study the difference of enrollment between different states and categories. First, we group this table by state and category and count the total enrollment of different states and categories. Then we make a barplot and a state_choropleth to show it more specific.

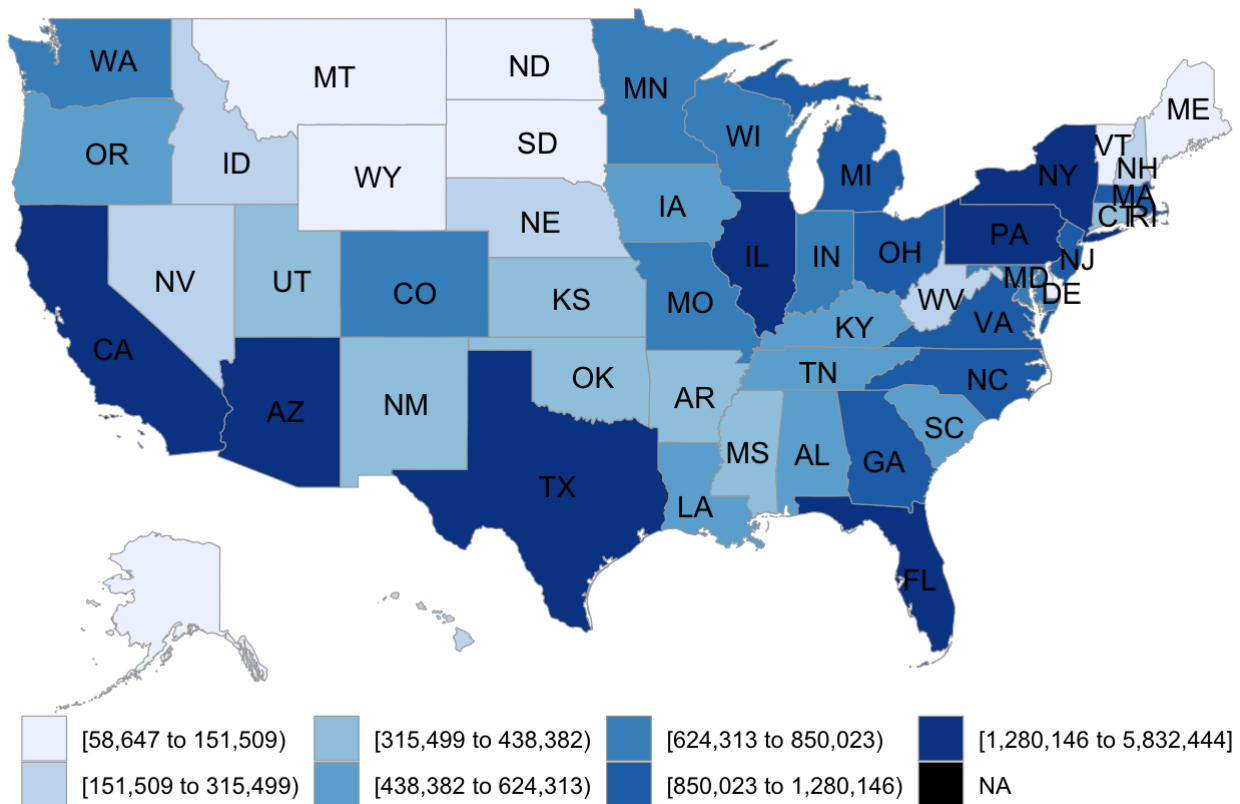
```
## # A tibble: 10 x 3
## # Groups:   state [4]
##   state      category      count
##   <chr>      <chr>      <dbl>
## 1 California Total Minority 1654432
## 2 California Women          1476438
## 3 California Hispanic        949890
## 4 Texas      Women          876547
## 5 Texas      Total Minority 857045
## 6 California White          799516
## 7 New York   Women          731847
## 8 Florida    Women          642964
## 9 New York   White          613885
## 10 Texas     White          598209
```



Total Enrollment by State and Category



Total Enrollment by State



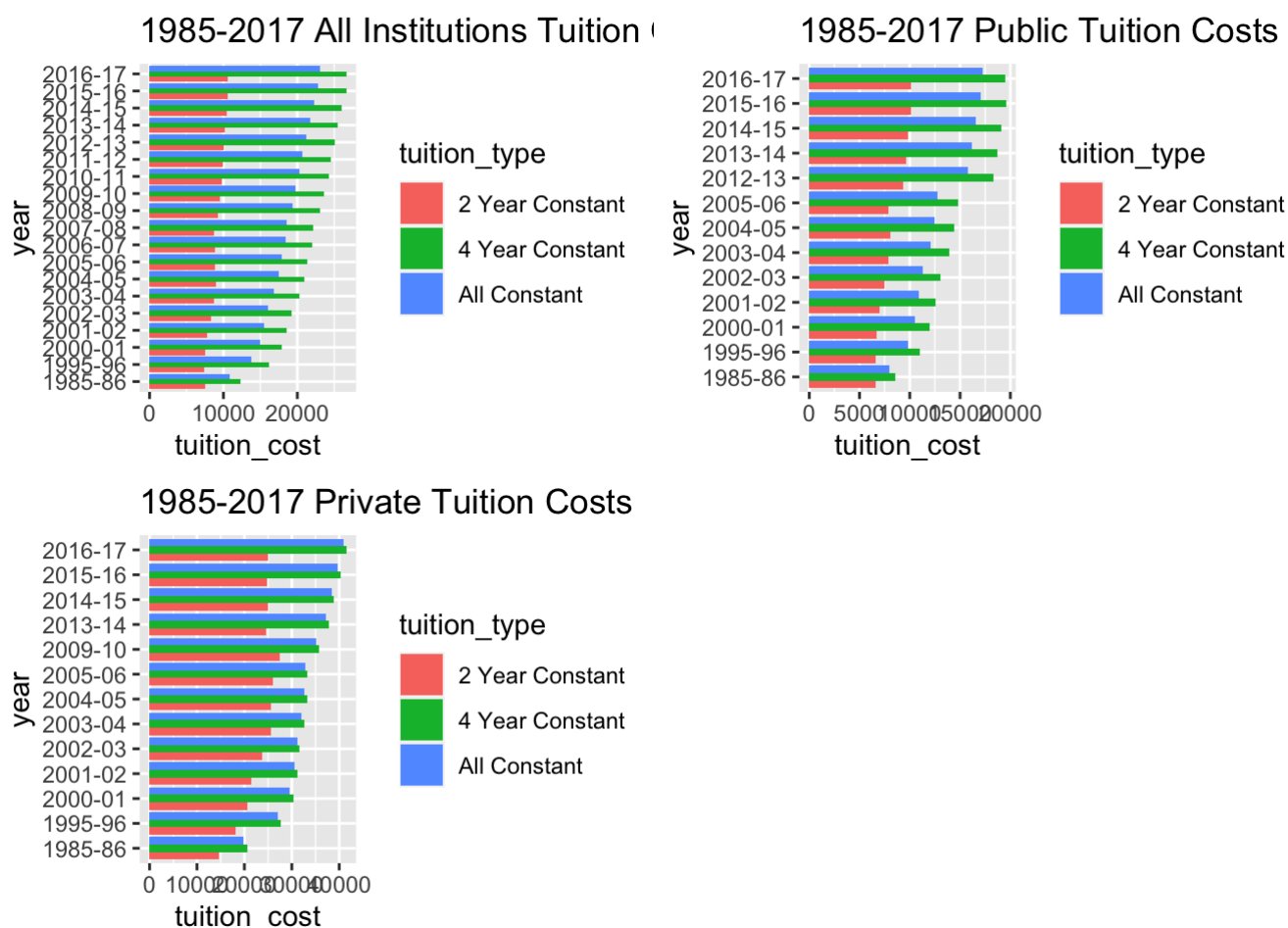
From the table, we can see that, the category of “Total Minority” in California has the highest enrollment, and the second and third rank also come from California, that is “Women” and “Hispanic”. And the fourth and fifth come from Texas which is “Women” and “Total Minority”. Besides, we can learn that from the best 10 enrollment, almost half of them come from the category of “Women”.

From the first bar plot, we can see that the best 5 high enrollment states are California, Texas, New York, Florida, Illinois and they almost have all categories which means these states have more comprehensive universities. And combined with the third plot, we can see that those states with higher total enrollment are mostly spread at the west and east coast of America where economic development is faster.

From the second bar plot, we can see that Women, White, Total Minority, Hispanic, Black have more enrollment than other categories and Women has the highest proportion in all of these categories which is same as we have learned from the table.

3. Historical_tuition wrangling

This table show the historical tuition costs from 1985 to 2017. At this table, we only use constant dollars which is more stable with economic status. Then, we make 3 barplot to show the variation of tuition costs between different university types and tuition types.



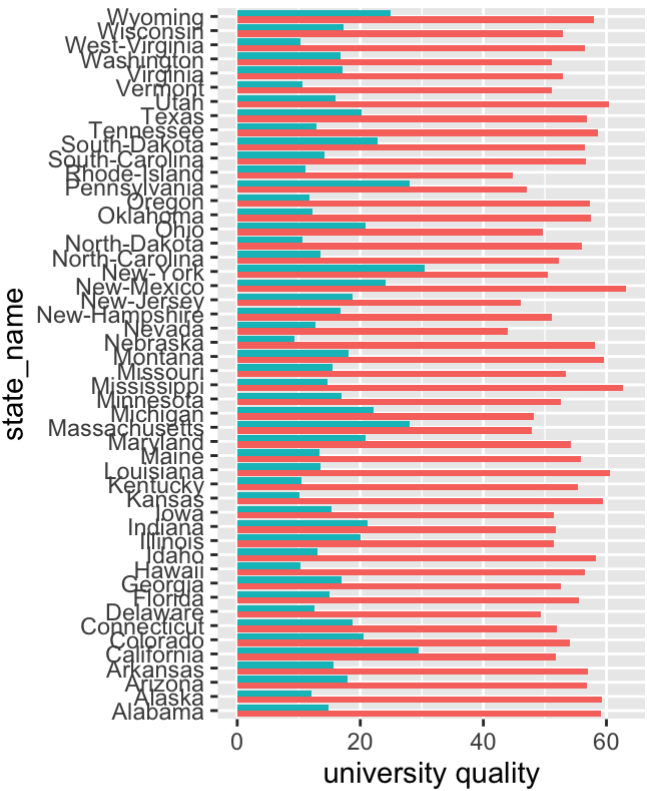
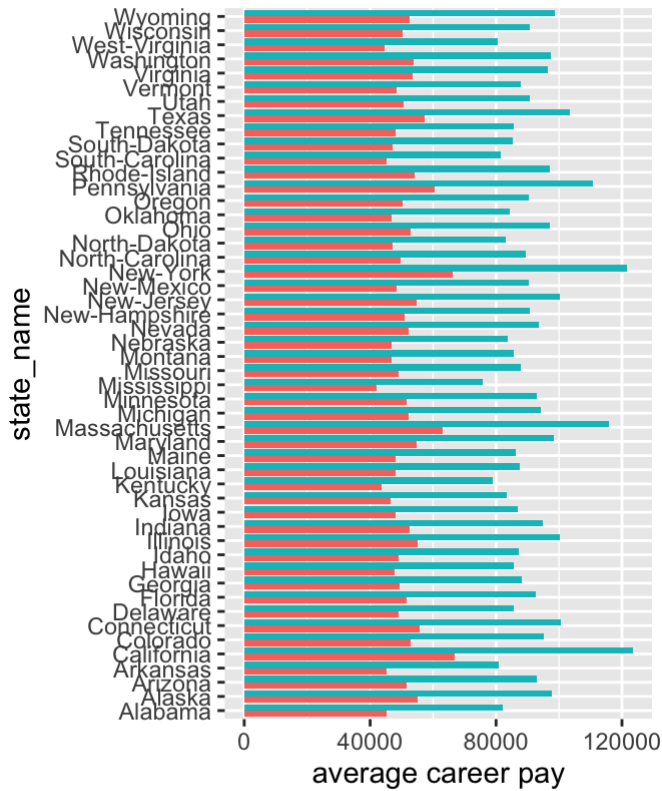
From the plot, although there are some missing values in time in public and private part, basically, the tuition costs has been increasing continuously from 1985 to 2017 no matter which tuition type is. Among these, private universities have the highest tuition cost, and public universities have the lowest tuition cost. Besides, all these three tuition type have increased through years, but “2 year constant” did not increased much while other two types increased over 10000 dollars through these years.

4. salary_potential wrangling

This table shows the estimated salary according to different level of career and we will show that the difference of potential salary and university quality(make_world_better_percent: Percent of alumni who think they are making the world a better place, stem_percent: Percent of student body in STEM) between states. First, do some summarise and cleaning on the table and create two barplots and a state_choropleth which can be used to make a comparison with the plot in step 2.

Average Career Pay By Stat

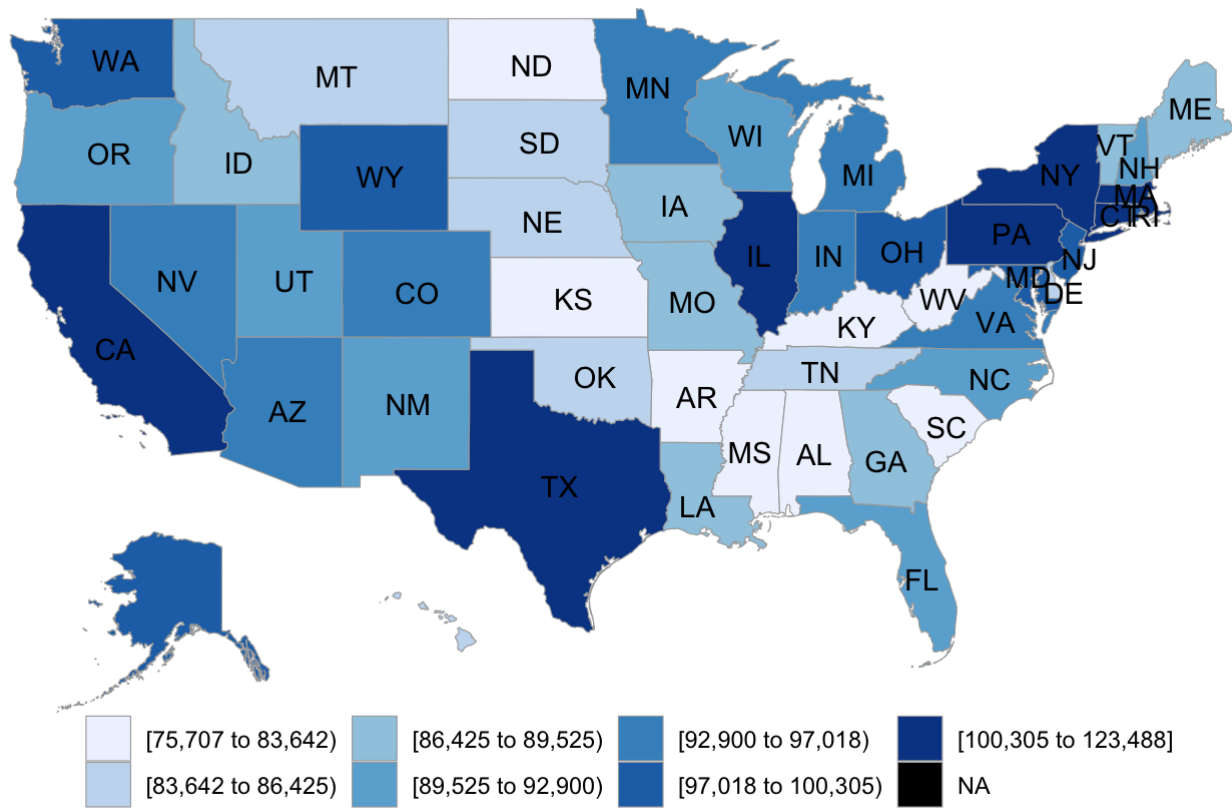
University Quality By States



variable avg_early_career_pay avg_mid_career_pay

variable avg_make_world_better_perc avg_mid_career_pay

Average Mid-Career Pay by State



First, from the two barplot, we can see that the estimated early career pay does not differ greatly among states, it is about 40000 dollars, and the difference of `make_world_better_percent` between states is also not big, around 55%. But the other two variable have large difference between different states: California, Massachusetts, New York and Pennsylvania has the highest estimated mid-career pay, and correspondingly, these four states also have higher stem percentage compared with other states. This may mean that stem correlated majors need more professors for a university. Besides, combined the next plot with the plot "Total Enrollment by State" in part2, we also know that California and New York have more enrollment than other states. This means enrollment and salary may have some connections, but this table only shows the estimated salary, so in the next part, we will show the relationship between enrollment and real tuition income.

5.(a) tuition_cost wrangling

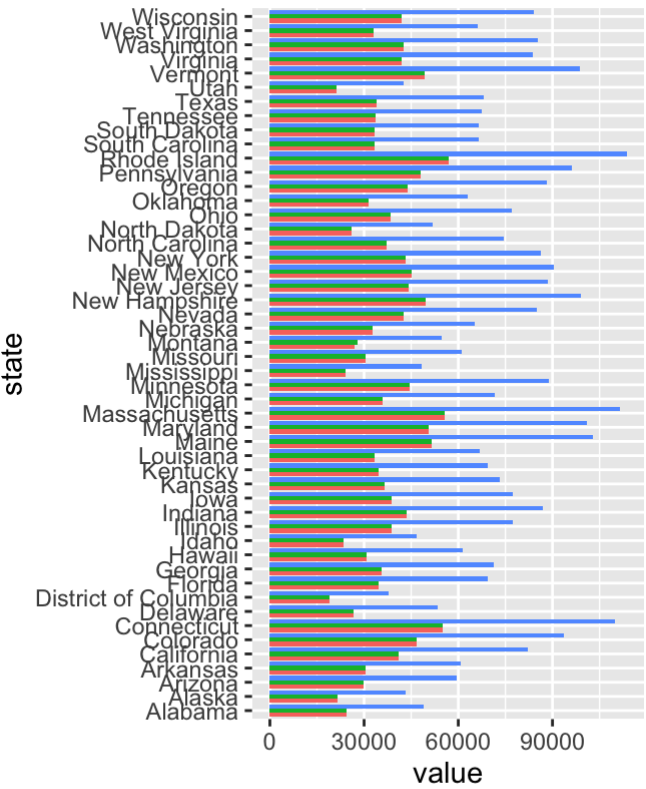
This table shows the tuition cost for 2018-2019. First, we change missing values to zero. And then we analyze the tuition cost according to different classification criterion.

```
## # A tibble: 5 x 5
## # Groups:   state [5]
##   state      degree_length avg_in_state_total avg_out_of_state_total avg_total
##   <chr>      <chr>          <dbl>          <dbl>          <dbl>
## 1 Virginia      4 Year          24331.          42394.          66726.
## 2 Vermont        4 Year          25012           40924           65936
## 3 Arizona        4 Year          23451           42010           65461
## 4 California     4 Year          23592.          39720.          63312.
## 5 Rhode Island  4 Year          23923           38666.          62590.
```

```
## # A tibble: 5 x 5
## # Groups:   state [5]
##   state      degree_length avg_in_state_total avg_out_of_state_tot... avg_total
##   <chr>      <chr>          <dbl>          <dbl>          <dbl>
## 1 Vermont        2 Year          68640           68640          137280
## 2 Rhode Island   4 Year          56793.          56793.          113586.
## 3 Massachusetts  4 Year          55763           55763           111526
## 4 Connecticut    4 Year          55008.          55008.           110015
## 5 Maine           4 Year          51419.          51419.           102837.
```

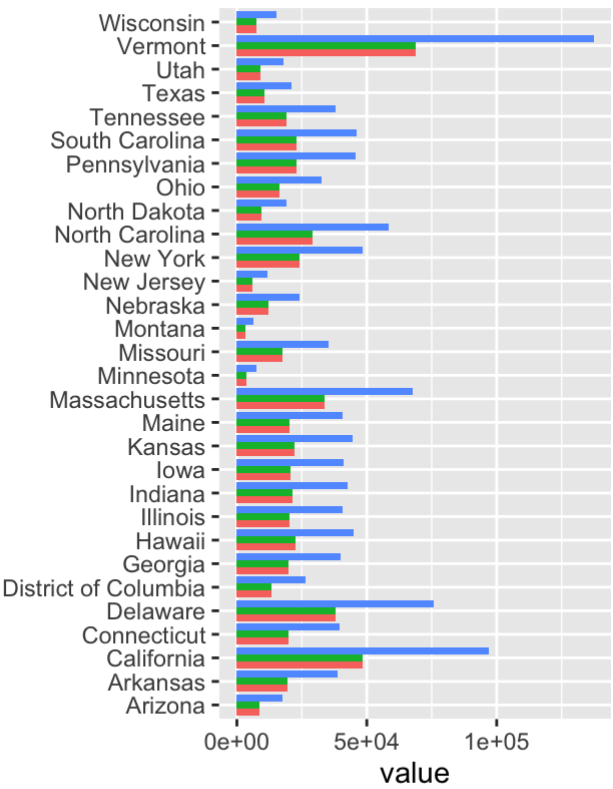
```
## # A tibble: 5 x 5
## # Groups:   state [5]
##   state      degree_length avg_in_state_total avg_out_of_state_tot... avg_total
##   <chr>      <chr>          <dbl>          <dbl>          <dbl>
## 1 Missouri       4 Year          47570           47570           95140
## 2 New York        4 Year          41380.          41380.           82759
## 3 Massachusetts  2 Year          41050           41050           82100
## 4 Texas           4 Year          33624           33624           67248
## 5 Vermont         2 Year          32676           32676           65352
```

4 Year Tuition Cost Private



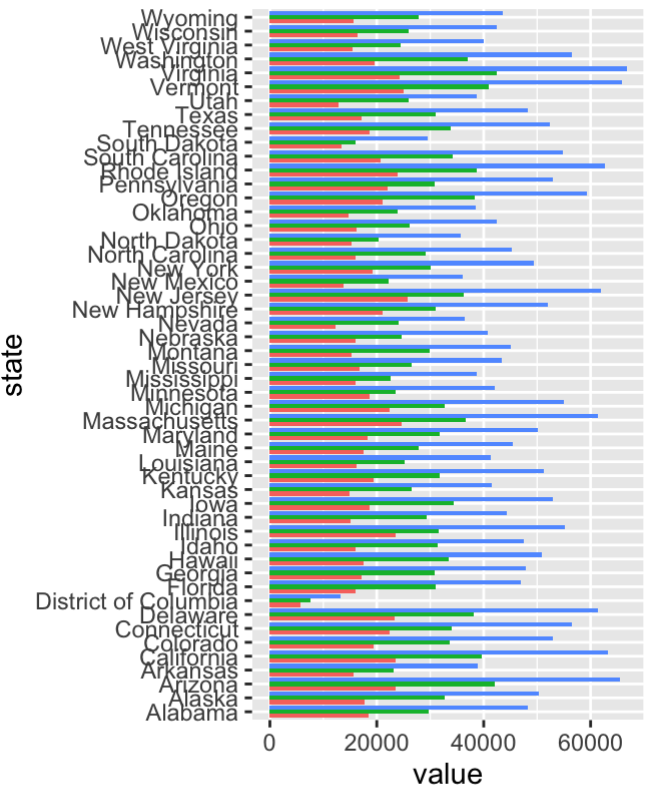
variable avg_in_state_total avg_out_of_state_

2 Year Tuition Cost Private



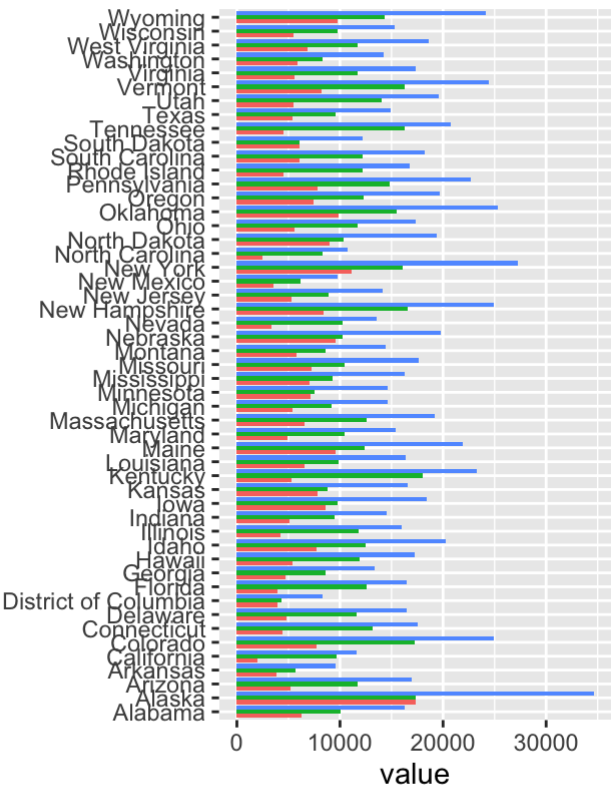
variable avg_in_state_total avg_out_of_state_

4 Year Tuition Cost Public

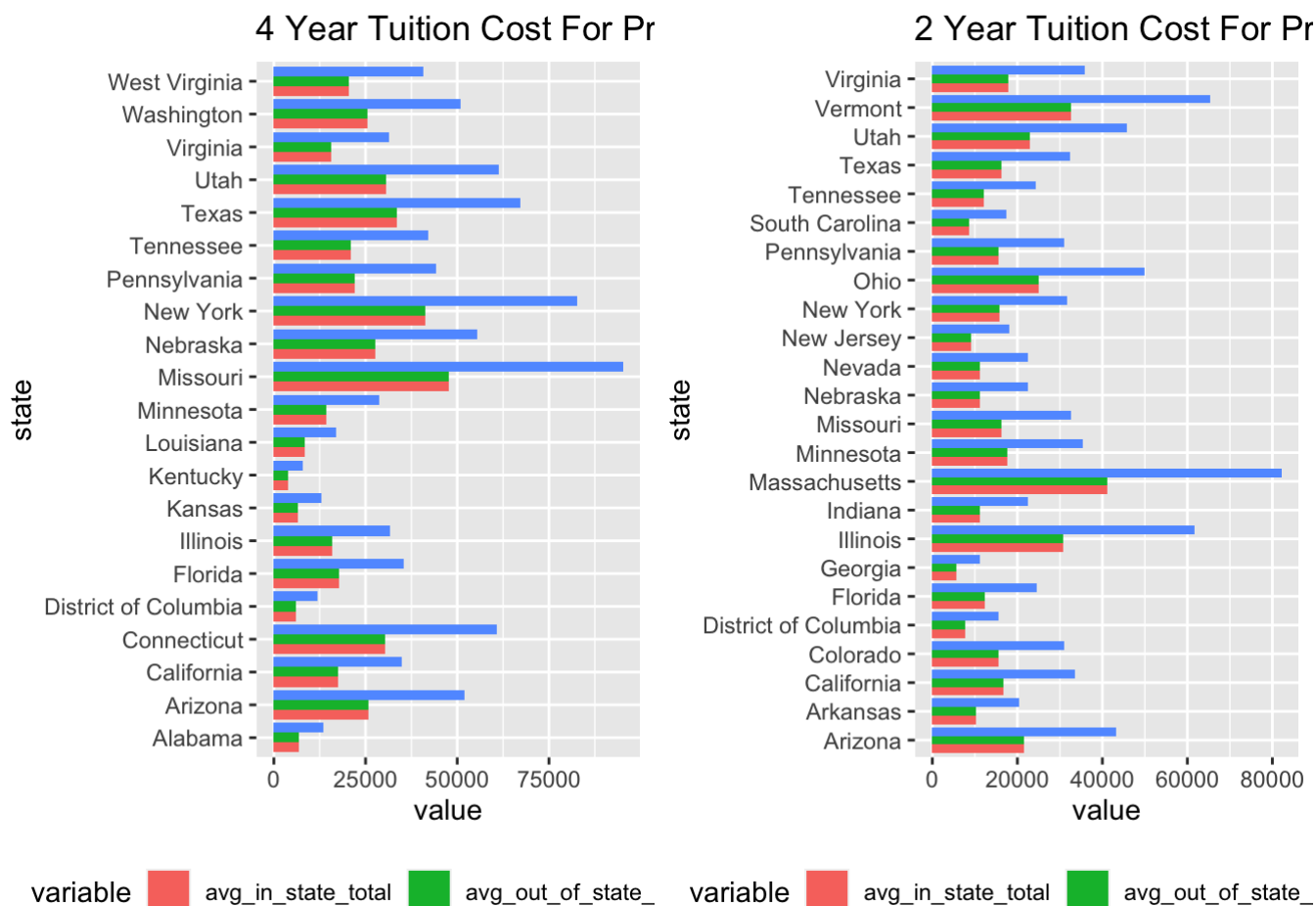


variable avg_in_state_total avg_out_of_state_

2 Year Tuition Cost Public



variable avg_in_state_total avg_out_of_state_



The three tables show the best 5 high tuition cost according to different tuition type. we can see that Vermont always has higher tuition cost in all these three tuition types, and Rhode Island and Massachusetts is close to Vermont in private tuition type. From part4, we also know that Massachusetts has almost highest estimated potential. Besides, we can see that the tuition cost has increased compared with the historical tuition cost in part3.

In this part, we have also created 6 bar plots to compare the difference of in-state and out-of-state under different situations which is similar to what we have done in part3. Overall, the in-state tuition cost is much small than out-of-state tuition cost.

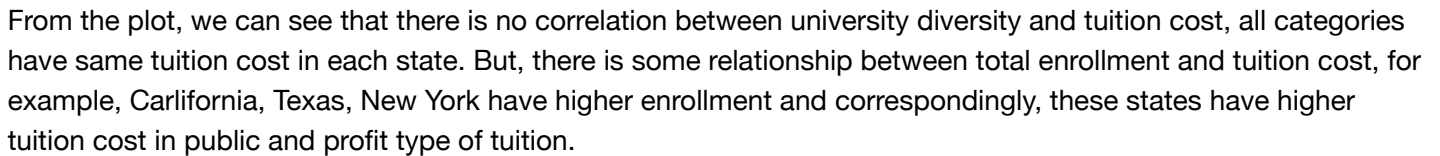
In private type, in-state tuition cost is the same as out-of-state tuition cost which is quite different from the public universities no matter 4-year or 2-year, but the 4-year tuition cost is still about twice that of 2-year. And in some states, they do not provide 2-year education.

And compared within "Public" type, we can see that 4-year tuition cost is about twice that of 2-year, 4-year tuition cost is about 35000 dollars, but 2-year tuition cost is only about 12500, much smaller than 4-year. Besides, every state have public universities.

For profit type, it's the same with private type, that is, in-state tuition cost is the same as out-of-state tuition cost no matter what the degree length is. Also, there are some states don't have universities for profit type.

5.(b) tuition_cost + diversity_school

In this part, we join these two tables to find out whether there are some correlations between university diversity and tuition cost.



6.(a) In this part, we will deal with the specific data we collected from each state and combine them together. We also turn costs to numeric type for easy calculation. And also, we output the tidy version of our data as csv files which can be found in my github repository.

##	State	Number of Schools	In-State Public Tuition Fees
## 3	Alaska	10	7293
## 4	Alabama	97	6931
## 5	Arkansas	90	4877
## 6	American Samoa	1	3950
## 7	Arizona	140	4667

##	Out-State Public Tuition Fees	Private Tuition Fees	On-Campus Living Costs
## 3	18608	12891	13454
## 4	13348	16852	12115
## 5	7743	18518	12023
## 6	4250	0	0
## 7	11342	17964	13820

##	Off-Campus Living Costs
## 3	16356
## 4	12092
## 5	12972
## 6	4000
## 7	14597

##	State	In-state Undergraduate Total	Out-state Undergraduate Total
## 1	alaska	21822	33137
## 2	alabama	20328	26222
## 3	arkansas	19003	21868
## 4	american samoa	6550	6850
## 5	arizona	21161	28374

##	In-state Graduate Total	Out-state Graduate Total
## 1	27070	39145
## 2	23198	33250
## 3	21549	27222
## 4	2600	2600
## 5	28452	39308

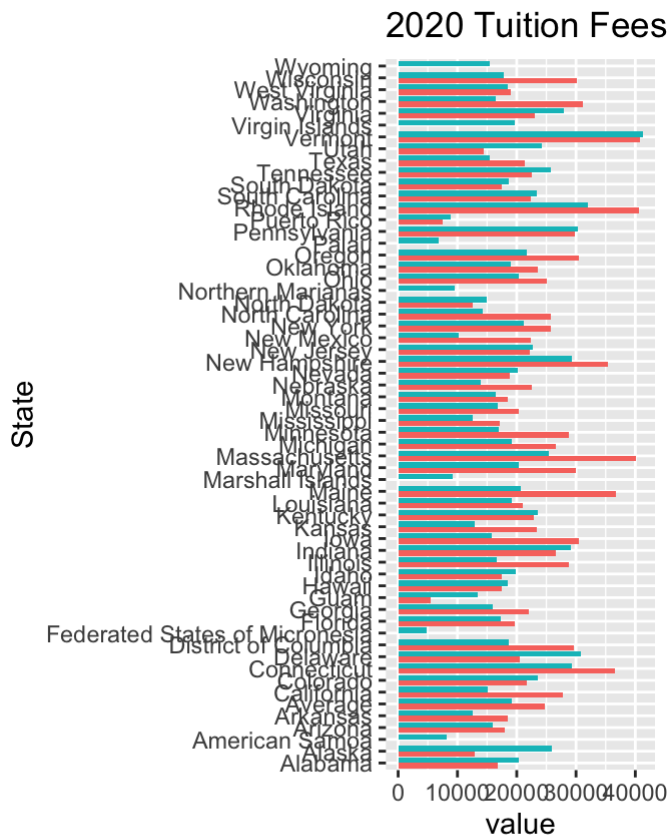
##	In-state Undergraduate Tuition Fee	Out-state Undergraduate Tuition Fee
## 1	7293	18608
## 2	6770	12664
## 3	4877	7743
## 4	3950	4250
## 5	5180	12393


##	In-state Graduate Tuition Fee	Out-state Graduate Tuition Fee
## 1	12541	24616
## 2	9639	19692
## 3	7423	13096
## 4	0	0
## 5	12471	23327

##	Undergraduate-room	Graduate-room	Undergraduate-books	Graduate-books
## 1	13454	13454	1074	1074
## 2	12115	12115	1443	1443
## 3	12023	12023	2102	2102
## 4	0	0	2600	2600
## 5	13820	13820	2161	2161



These two tables are come from the second data source.

6.(b) Do some analysis with the data we collected from the website and find some connection with the previous analysis.

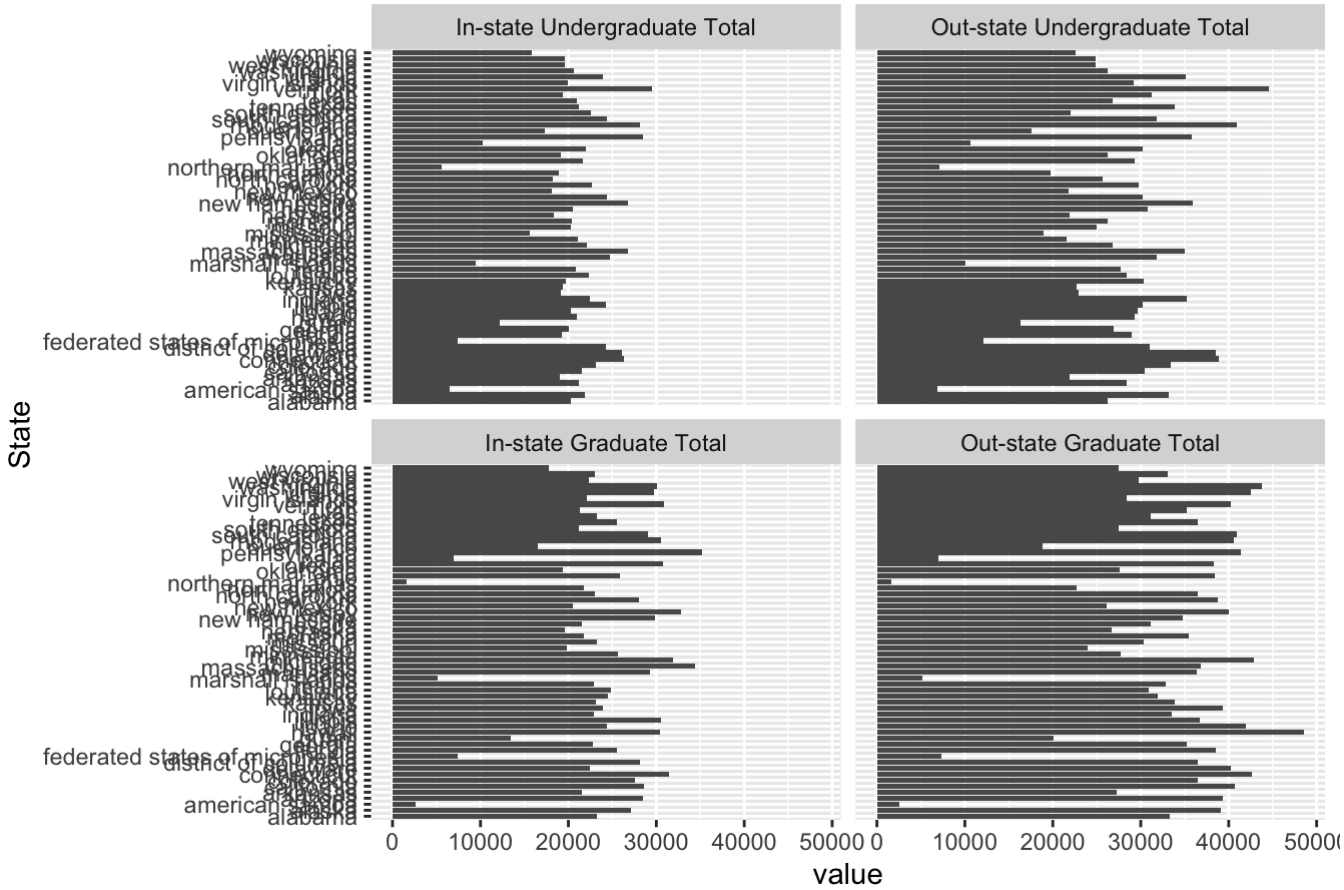


variable  Private Tuition Fees  Private Tuition Fees

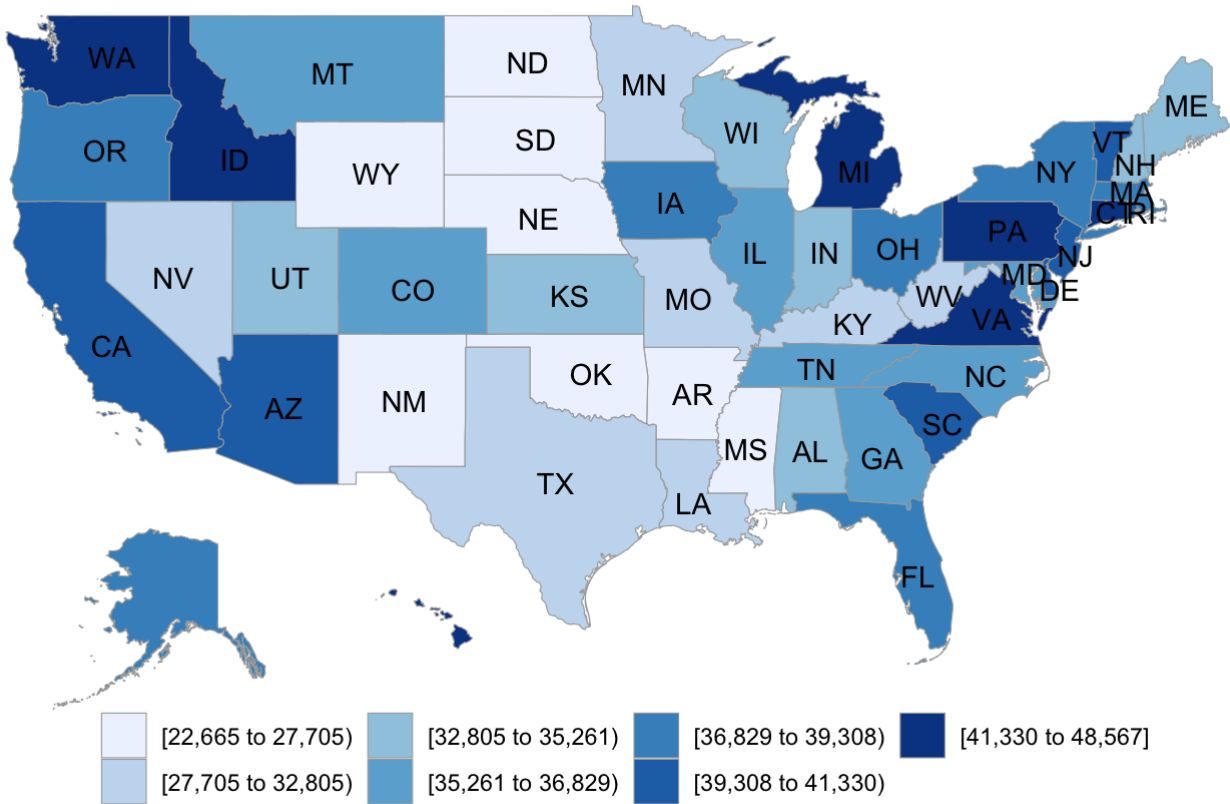


variable  On-Campus Living Costs  Of

2020 Costs of Attendance By States



2020 Costs of Attendance on Graduate By States



From the first two barplot, we can know that, the cost of private type tuition is still higher than that of public type of tuition, but the difference is smaller in some states. And Vermont has the highest tuition costs both in public and private, which is similar as we have learned from part5.(a). As for living costs, there is slightly difference between on-campus living and off-campus living, California and Hawaii have relatively higher off-campus living cost and District of Columbia, Massachusetts and Rhode Island have relatively higher on-campus living. But no matter on-campus or off-campus, it's seems like these are some relationship between tuition costs and living costs: basically, higher tuition costs may lead to higher living costs, and in these states where mostly are coastal states, the potential salary would be higher either according to Part4.

From the third barplot, we can see that the COA of graduate study is more higher than that of undergraduate study, and obviously, students who are not resident of this state need to pay more than those who are resident of this state. And from the last plot, we caan see that, those states with higher COA are basically coastal states where have more international students and the economic development is faster.

7. Conclusion

From this project, we can learn that, basically, coastal states have higher tuition costs in America where are known to have better universities and more international students. But also, the salary in these regions is also higher. And in general, private universities have higher tuition fees than public universities, but the tuition costs have no relationship with the university diversity: tuition costs is same for all kinds of diversities.