

# Final Project: Is College Worth It? (Tuition VS Salary)

*Junjie Yang*

*5/2/2020*

## I. Introduction

Is college worth it? Is college a good investment for your future? If it is, what kind of factors in college would have an impact on career performance?

On one hand, college could be worth it by leading to higher employment rates and higher career performance, in terms of various financial measurements, than people who do not go to college. On the other hand, college tuition is constantly rising and is the same for student loan debt.

In this project, four data sources are acquired from the US Department of Education, the Chronicle of Higher Education, the National Center for Education Statistics, and payscale.com. A final dataset in tidy version is created by conducting a significant amount of data cleansing and data wrangling techniques, so as to retrieve insightful information regarding the relationship between tuition or other factors in college and future career performance of college graduates.

## Github Link

(Include several Data in tidy version, Rmd File, Report in PDF File and HTML File)

(<https://github.com/Junjie-Dylan-Yang/Data-Wrangling-Project>)

## II. ETL process: Data Import and Data Cleansing

1,

### Import first data: tuition\_cost

First data source, tuition\_cost, is from “College tuition, Diversity, and Pay” in rfordatascience/tidetuesday/2020-03-10, which is originally acquired from the US Department of Education and the Chronicle of Higher Education.

### Data Cleaning for tuition\_cost data

In the tuition\_cost data, relevant columns are selected (name of the school, state, state code, type of the school, length of the degree). Also, room and board fee and tuition are combined as total tuition and fee.

### Below is the snippet of tuition\_cost data

```
## # A tibble: 10 x 7
##   name      state state_code type degree_length in_state_tuitio~ out_of_state_tu~
##   <chr>   <chr> <chr>      <chr> <chr>                <dbl>          <dbl>
## 1 Aanii~ Mont~ MT          Publ~ 2 Year                2380           2380
## 2 Abile~ Texas TX          Priv~ 4 Year                45200          45200
```

##	3	Abrah~ Geor~ GA	Publ~ 2 Year	12602	21024
##	4	Acade~ Minn~ MN	For ~ 2 Year	17661	17661
##	5	Acade~ Cali~ CA	For ~ 4 Year	44458	44458
##	6	Adams~ Colo~ CO	Publ~ 4 Year	18222	29238
##	7	Adelp~ New ~ NY	Priv~ 4 Year	54690	54690
##	8	Adiro~ New ~ NY	Publ~ 2 Year	17035	21595
##	9	Adria~ Mich~ MI	Priv~ 4 Year	48405	48405
##	10	Advan~ Virg~ VA	For ~ 2 Year	13680	13680

2,

### Import second data: student\_diversity

Second data source, student\_diversity by college/university, along with school type, degree length, state, in-state vs out-of-state is from the Chronicle of Higher Education.

### Data Cleaning for student\_diversity data

In the student\_diversity data, the main data cleansing task is to modify name of institution to match the “name” column and “state” column in the tuition\_cost data, in order to combine dataset. Several data wrangling steps were applied. First is to change the column name “INSTITUTION” to “name”. After that, convert any abbreviation of University from “U.” to “University”. From the first glance, the name of state is located at the very end of the name of institution. The next step is to extract state from school name with the help of state.name which contains the list of all the state name and column “state” is created. Last but not least, state name inside the name of institution needed to remove. Using str\_count to count the letters within state in each observation and str\_sub help to keep the name of school only in the “name” column. Str\_trim and str\_squish are used to remove unnecessary spaces in “name”.

### Below is the snippet of student\_diversity\_cleaned data

```
## # A tibble: 10 x 11
##   name ENROLLMENT WOMEN `AMERICAN INDIA~ ASIAN BLACK HISPANIC
##   <chr>      <dbl>  <dbl>      <dbl> <dbl> <dbl> <dbl>
## 1 Univ~    195059 134722      876  1959 31455 13984
## 2 Ivy ~     91179  53476      357  1369 12370  5533
## 3 Libe~     81459  48329      447   856 14751  1186
## 4 Lone~     69395  41268      168  4198 12094 23751
## 5 Miam~     66046  38323       47   655 10722 44870
## 6 Gran~     62304  46647      586  2446 13856  8933
## 7 Texa~     61642  29277      173  3545  1879 11256
## 8 Univ~     60767  33482      120  3343  6400 13108
## 9 Ohio~     58322  28658       76  3339  3108  2049
## 10 Hous~     58276  34007      116  5391 18520 18411
## # ... with 4 more variables: `NATIVE HAWAIIAN / PACIFIC ISLANDER` <dbl>,
## #   WHITE <dbl>, `TOTAL MINORITY` <dbl>, state <chr>
```

### Combine tuition\_cost and student\_diversity data based on “name” and “state”

So far, student\_diversity and tuition\_cost are modified to share two common column, “name” – name of the school and “state” – the state that the school is located. Thus, student\_diversity and tuition\_cost datasets are merged for late development. There are a few schools appears in the tuition\_cost dataset but not in the

student\_diversity and “NA” value appear. It is reasonable and schools with “NA” value are removed from the combined dataset. The combined dataset is arranged by state and the name of the school.

Below is the snippet of the combined dataset, `tuition_with_diversity`

```
## # A tibble: 10 x 16
##   name state state_code type degree_length in_state_tuitio~ out_of_state_tu~
##   <chr> <chr> <chr>    <chr> <chr>                <dbl>                <dbl>
## 1 Alab~ Alab~ AL      Publ~ 2 Year                4440                8880
## 2 Alab~ Alab~ AL      Publ~ 4 Year                16490               24818
## 3 Amri~ Alab~ AL      Priv~ 4 Year                6900                6900
## 4 Athe~ Alab~ AL      Publ~ 4 Year                6810               12870
## 5 Aubu~ Alab~ AL      Publ~ 4 Year                24608               43856
## 6 Aubu~ Alab~ AL      Publ~ 4 Year                17268               29028
## 7 Bevi~ Alab~ AL      Publ~ 2 Year                6070                9940
## 8 Birm~ Alab~ AL      Priv~ 4 Year                30000               30000
## 9 Bish~ Alab~ AL      Publ~ 2 Year                4740                8610
## 10 Calh~ Alab~ AL      Publ~ 2 Year                4840                8690
## # ... with 9 more variables: ENROLLMENT <dbl>, WOMEN <dbl>, `AMERICAN INDIAN /
## # ALASKA NATIVE` <dbl>, ASIAN <dbl>, BLACK <dbl>, HISPANIC <dbl>, `NATIVE
## # HAWAIIAN / PACIFIC ISLANDER` <dbl>, WHITE <dbl>, `TOTAL MINORITY` <dbl>
```

3,

### Import third data: `Best_School`

Third data source, `best_school` is html data, acquired from the from the [payscale.com](https://www.payscale.com). It contains all the schools in United States that are arranged by various measurement of career performance, such as “Early Career Pay” and “Mid Career Pay”.

**Problem encountered** When importing html data from <https://www.payscale.com/college-salary-report/bachelors>, I realized the table only include the data with the top 25 schools in the United States, descending by measurement of career performance. That’s the issue that I am not expecting. Moreover, this is the first page in the web and there are 63 pages in total, which consists all the school data.

**Problem resolved** Instead of importing data 63 times to get the entire dataset, one alternative webpage is found by navigating the [payscale.com](https://www.payscale.com/college-salary-report/best-schools-by-state). The page “Best Schools By State” (<https://www.payscale.com/college-salary-report/best-schools-by-state>) outlays all the best schools ranked by measurement of career performance of all 50 states. Clicking on each state would direct to the schools data within that particular state. In order to import the entire data, I first convert the string format in the list of state.name to match the url (“New York” to “New-York”). Then, a data frame is created. For-Loop is implemented to import all 50 states data to the R environment and to keep loading data into the data frame to form a complete dataset, “`Best_School`”, for data cleansing.

### Data Cleaning for `Best_School` data

First step is to modify the column name “School Name” to “name” and to keep the exact name of school only, in order to match the `tuition_with_diversity` dataset for binding. After that, there are several data cleansing steps that are applied to other columns. Only numeric values are extracted from the column, “Rank”, “Early Career Pay”, “Mid-Career Pay”, “% High Meaning”, “% STEM Degrees”. One lesson learned is that R suggests to use `parse_number()`, instead of `extract_numeric()` for extracting numeric value.

Below is the snippet of Best\_School\_clean data

##	name	Early Career Pay	Mid-Career Pay
## 1	Auburn University	54400	104500
## 2	University of Alabama in Huntsville	57500	103900
## 3	The University of Alabama	52300	97400
## 4	Tuskegee University	54500	93500
## 5	Samford University	48400	90500
## 6	Spring Hill College	46600	89100
## 7	Birmingham Southern College	49100	88300
## 8	University of Alabama at Birmingham	48600	87200
## 9	University of South Alabama	47700	86400
## 10	Alabama A & M University	48700	83500
##	% High Meaning % STEM Degrees		
## 1	51	31	
## 2	59	45	
## 3	50	15	
## 4	61	30	
## 5	52	3	
## 6	53	12	
## 7	48	27	
## 8	57	17	
## 9	56	17	
## 10	58	20	

Combine Best\_School\_clean data and tuition\_with\_diversity to form the final data

Finally, Best\_School\_clean data, which contains different measurements of career performance, merges with tuition\_with\_diversity data, which contains detailed school information including tuition and race. The column both datasets have in common is “name” and left\_join is performed. Similar to the previous merged dataset, schools with “NA” are removed from the dataset.

**Create new variables:**

Mid\_career\_pay\_paidoff: different between median salary for alumni with 10+ years experience and out of state tuition and fee.

Early\_career\_pay\_paidoff: different between median salary for alumni with 0-5 years experience and out of state tuition and fee.

Below is the snippet of the Final\_data

There are 622 observations in all 50 states in United States and each college or university is a unique observation. This is the tidy version of the final data and it will be stored as a csv file.

**Attribute Information**

Below information is from payscale.com:

“Early Career Pay” is defined as median salary for alumni with 0-5 years experience.

“Mid-Career Pay” is defined as Median salary for alumni with 10+ years experience.

“% High Meaning” is defined as the percentage of alumni who say their work makes the world a better place.

“% STEM Degrees” is defined as the percentage of degrees awarded in science, technology, engineering or a math subjects.

##		name	Early Career Pay	Mid-Career Pay
## 1		Auburn University	54400	104500
## 2		Tuskegee University	54500	93500
## 3		Samford University	48400	90500
## 4		Spring Hill College	46600	89100
## 5	University of Alabama at Birmingham		48600	87200
## 6	University of South Alabama		47700	86400
## 7	Troy University		44500	81500
## 8	Jacksonville State University		43800	80000
## 9	Auburn University at Montgomery		45000	79600
## 10	Huntingdon College		42400	78900

##	% High Meaning	% STEM Degrees	state	state_code	type	degree_length
## 1	51	31	Alabama	AL	Public	4 Year
## 2	61	30	Alabama	AL	Private	4 Year
## 3	52	3	Alabama	AL	Private	4 Year
## 4	53	12	Alabama	AL	Private	4 Year
## 5	57	17	Alabama	AL	Public	4 Year
## 6	56	17	Alabama	AL	Public	4 Year
## 7	60	8	Alabama	AL	Public	4 Year
## 8	61	7	Alabama	AL	Public	4 Year
## 9	61	12	Alabama	AL	Public	4 Year
## 10	69	14	Alabama	AL	Private	4 Year

##	in_state_tuition_and_fee	out_of_state_tuition_and_fee	ENROLLMENT	WOMEN
## 1	24608	43856	25912	12798
## 2	31820	31820	3103	1855
## 3	42200	42200	4933	3082
## 4	52926	52926	1376	820
## 5	17110	31030	18698	11288
## 6	17490	27360	15805	9700
## 7	20645	31060	19041	11948
## 8	18525	28245	8659	4978
## 9	17268	29028	5057	3233
## 10	37150	37150	1160	572

##	AMERICAN INDIAN / ALASKA NATIVE	ASIAN	BLACK	HISPANIC
## 1	183	601	1886	599
## 2	2	26	2345	32
## 3	17	80	372	218
## 4	10	16	210	77
## 5	46	931	3943	496
## 6	100	539	3285	402
## 7	143	140	6840	666
## 8	61	50	2030	110
## 9	23	104	1633	36
## 10	14	9	229	29

##	NATIVE HAWAIIAN / PACIFIC ISLANDER	WHITE	TOTAL	MINORITY
## 1	0	20855	3269	
## 2	0	52	2405	
## 3	1	4007	738	
## 4	1	947	359	
## 5	14	11840	5993	

```
## 6          33 10102          4684
## 7          19 9265          8294
## 8           7 5934          2258
## 9           9 2572          1941
## 10          2 738           313
##      Mid_career_pay_paidoff Early_career_pay_paidoff
## 1          60644          10544
## 2          61680          22680
## 3          48300           6200
## 4          36174          -6326
## 5          56170          17570
## 6          59040          20340
## 7          50440          13440
## 8          51755          15555
## 9          50572          15972
## 10         41750          5250
```

The tidy version of the final data, “Final\_data” is saved under the name “Tidy\_Final\_Data.xlsx” local location and will be committed from Github desktop to Github.com repository (<https://github.com/Junjie-Dylan-Yang/Data-Wrangling-Project>)

4,

#### Import fourth data: historical\_tuition

The last data source, historical\_tuition, is from “College tuition, Diversity, and Pay” in [rfordatascience/tidetuesday/2020-03-10](https://rfordatascience.tidetuesday.com/2020-03-10/), which is originally acquired from the National Center for Education Statistics. (<https://nces.ed.gov/fastfacts/display.asp?id=76>)

The fourth data, historical\_tuition, is tidy and contains the information of the trends in the cost of college education. Therefore, “historical\_tuition” is saved under the name of “Tuition\_trend.xlsx” in the same location of The tidy version of the final data.

Below is the snippet of tuition\_cost data

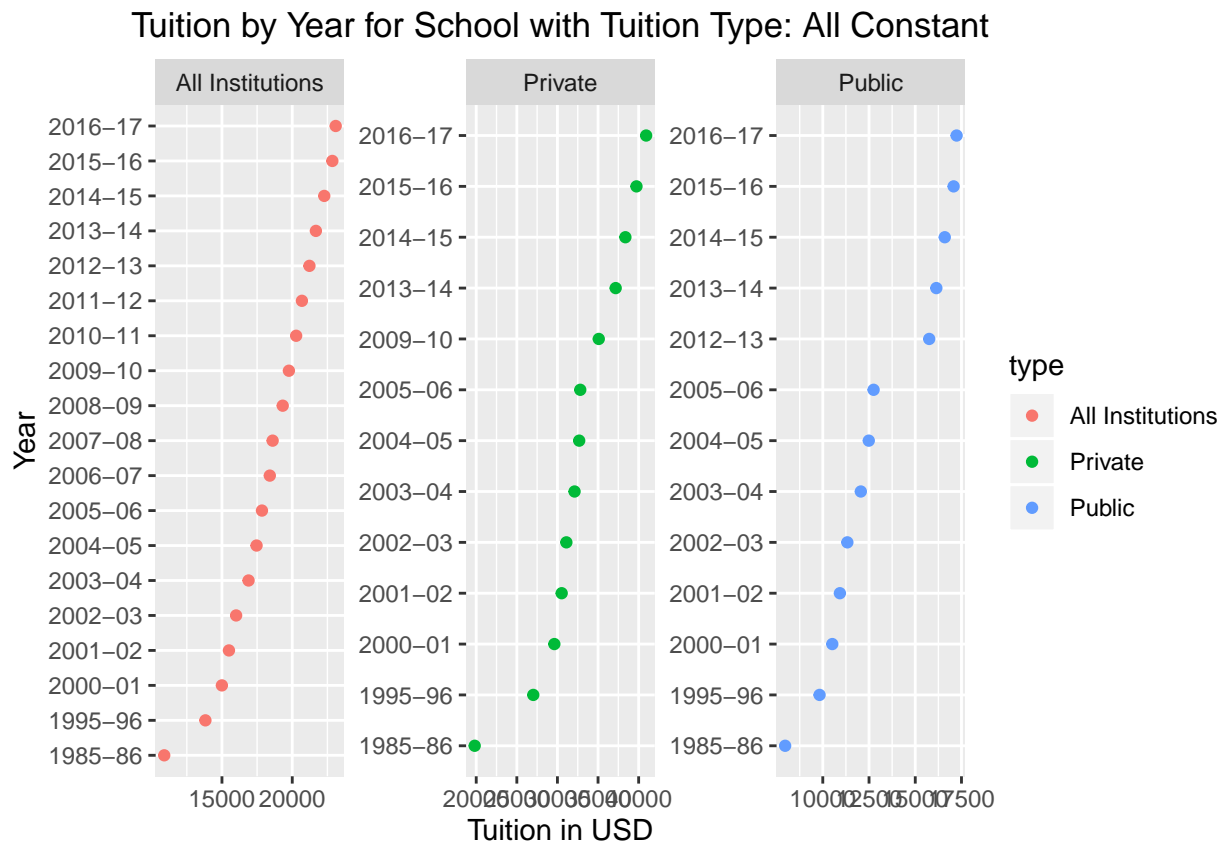
```
## # A tibble: 10 x 4
##   type          year  tuition_type  tuition_cost
##   <chr>         <chr>    <chr>         <dbl>
## 1 All Institutions 1985-86 All Constant    10893
## 2 All Institutions 1985-86 4 Year Constant  12274
## 3 All Institutions 1985-86 2 Year Constant   7508
## 4 All Institutions 1985-86 All Current    4885
## 5 All Institutions 1985-86 4 Year Current    5504
## 6 All Institutions 1985-86 2 Year Current    3367
## 7 All Institutions 1995-96 All Constant   13822
## 8 All Institutions 1995-96 4 Year Constant   16224
## 9 All Institutions 1995-96 2 Year Constant    7421
## 10 All Institutions 1995-96 All Current    8800
```

### III. Data Analysis by Various Plot and Tables

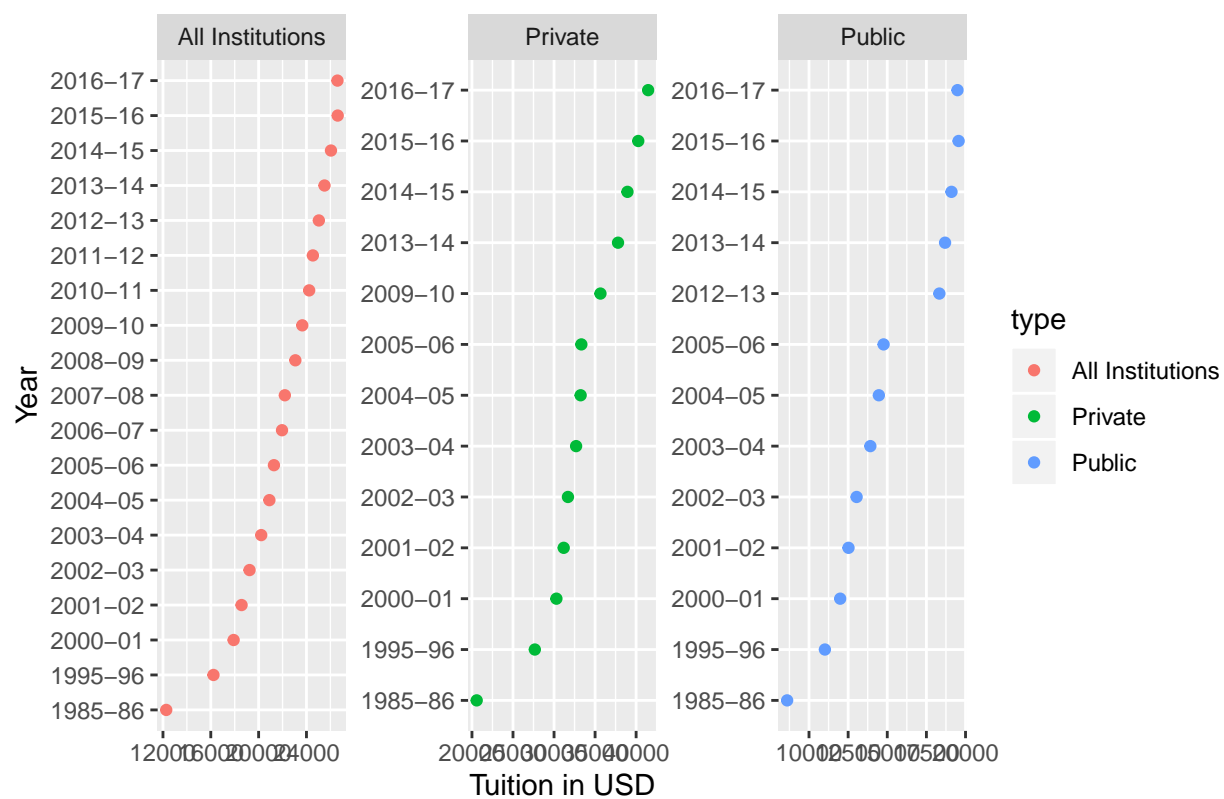
After a series of data wrangling and data cleansing conducted on several data sources from above, final data in tidy version, “Final\_data” and “historical\_tuition” data are ready to use for data analysis.

#### 1, Tuition Trend: Going upward over time

Split the historical\_tuition into 3 subset dataset by tuition type: “All Constant”, “4 Year Constant”, and “2 Year Constant”. From below plots, it clearly shows that, college tuition increases at a rapid rate over time, on schools with all three tuition types.

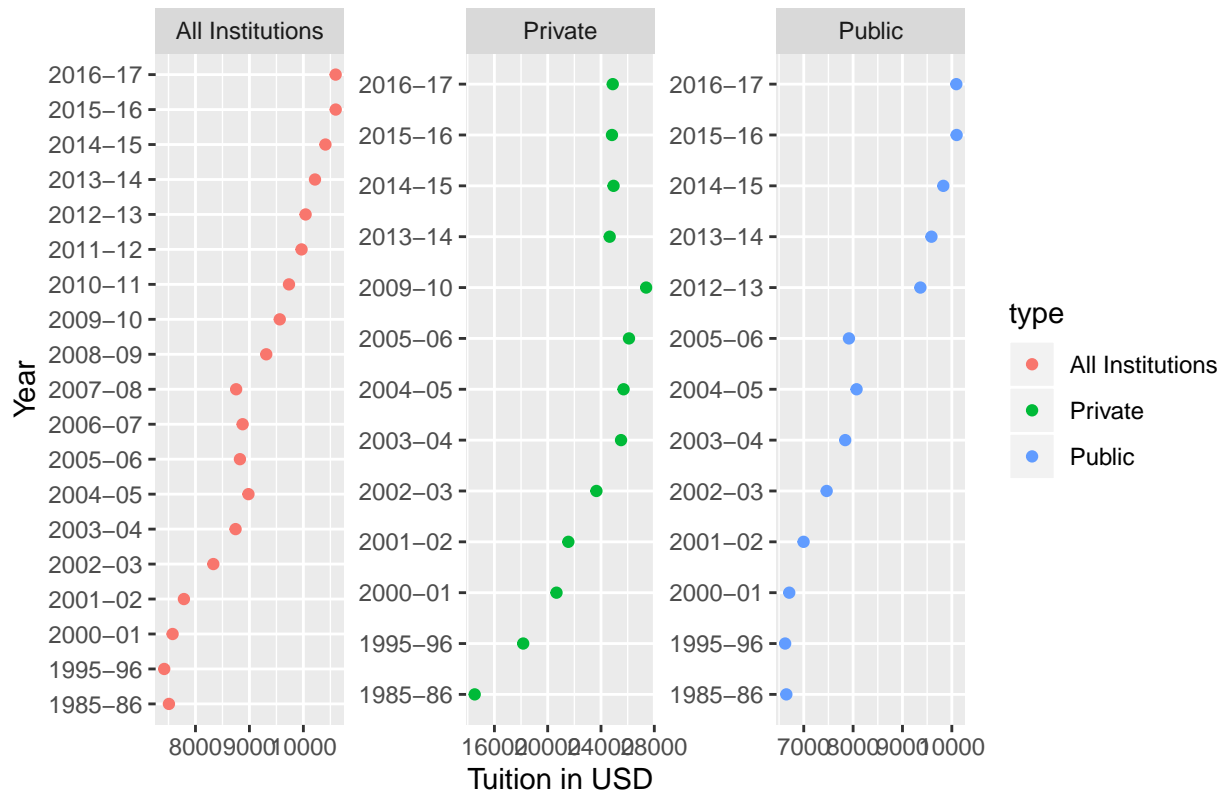


Tuition by Year for School with Tuition Type: 4 Year Constant





## Tuition by Year for School with Tuition Type: 2 Year Constant

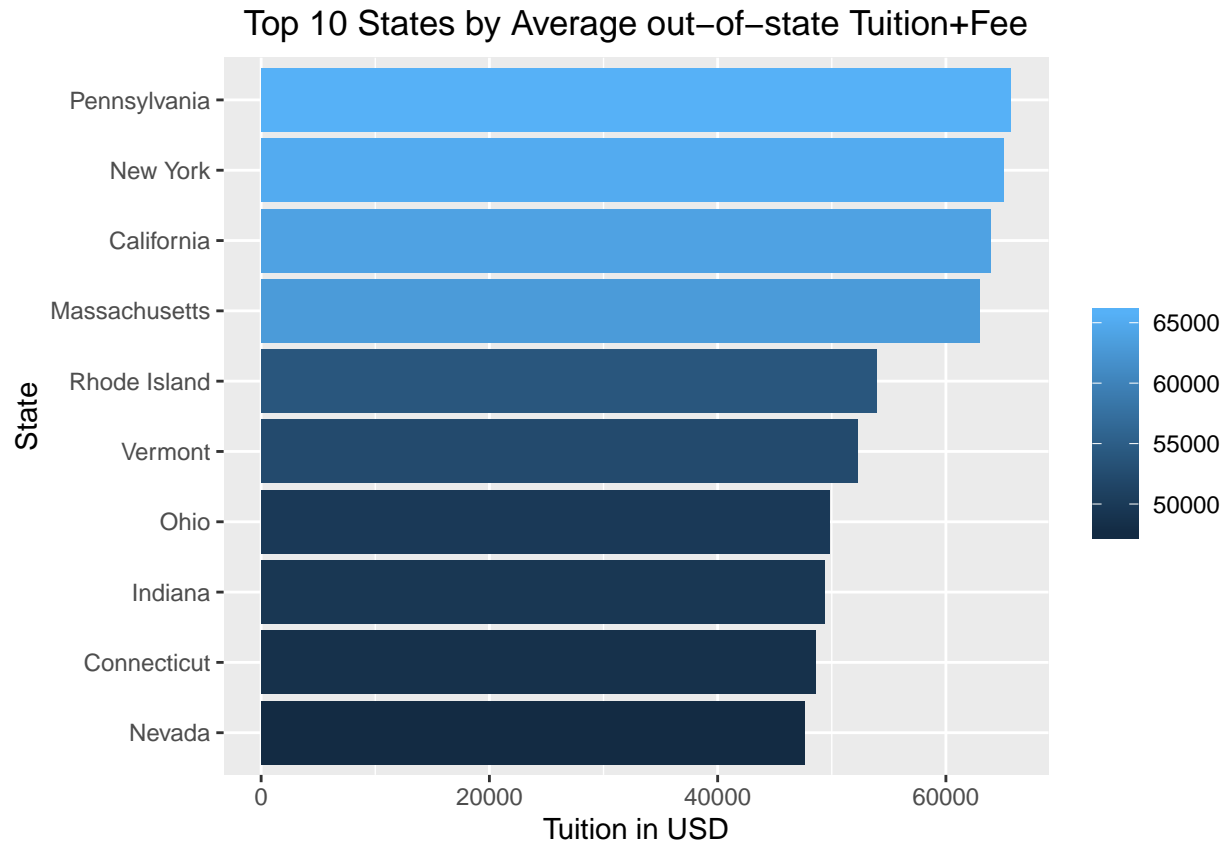


## 2, Take a look at the final data, “Final\_data” at the level of states.

Create another dataset “state\_data” at state level from the final data, “Final\_data”. All numeric values are summarised by taking the average respect to each state. This dataset is also saved in the same location as the final data.

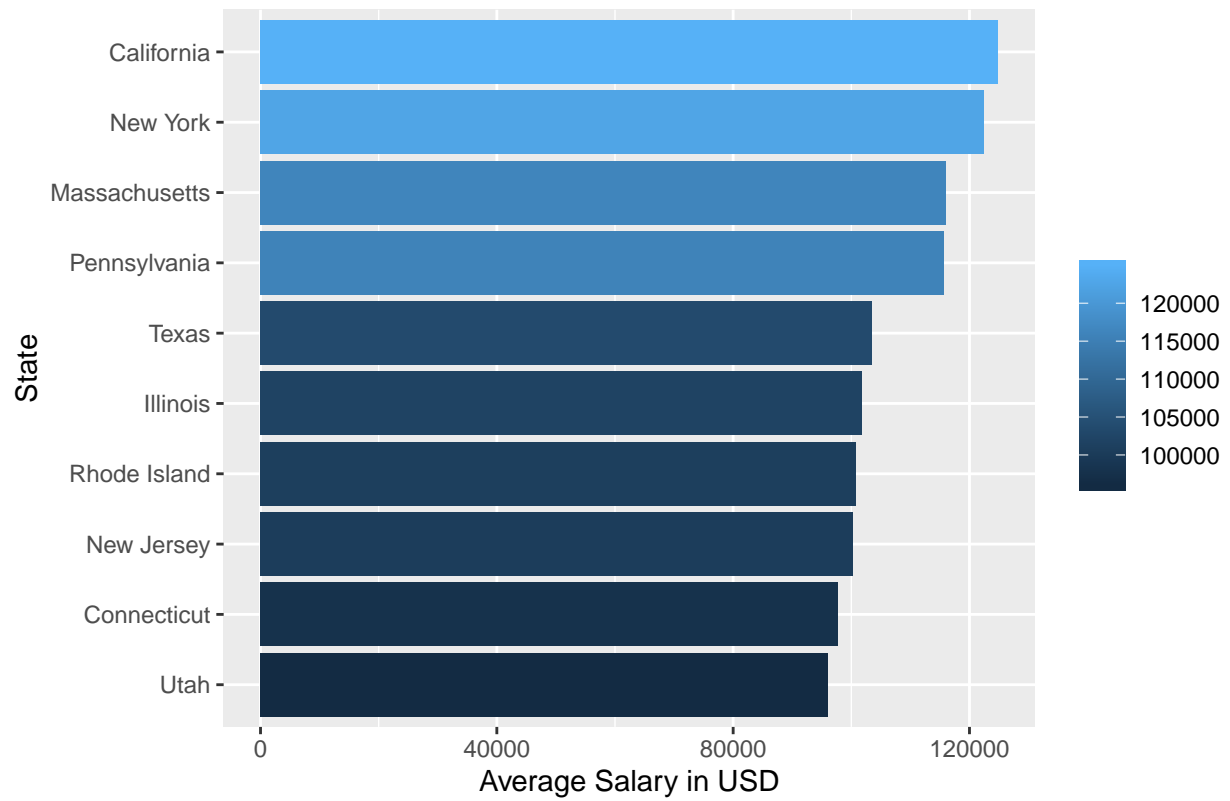
**Plots that show insightful information regarding tuition and career performance at the state level**

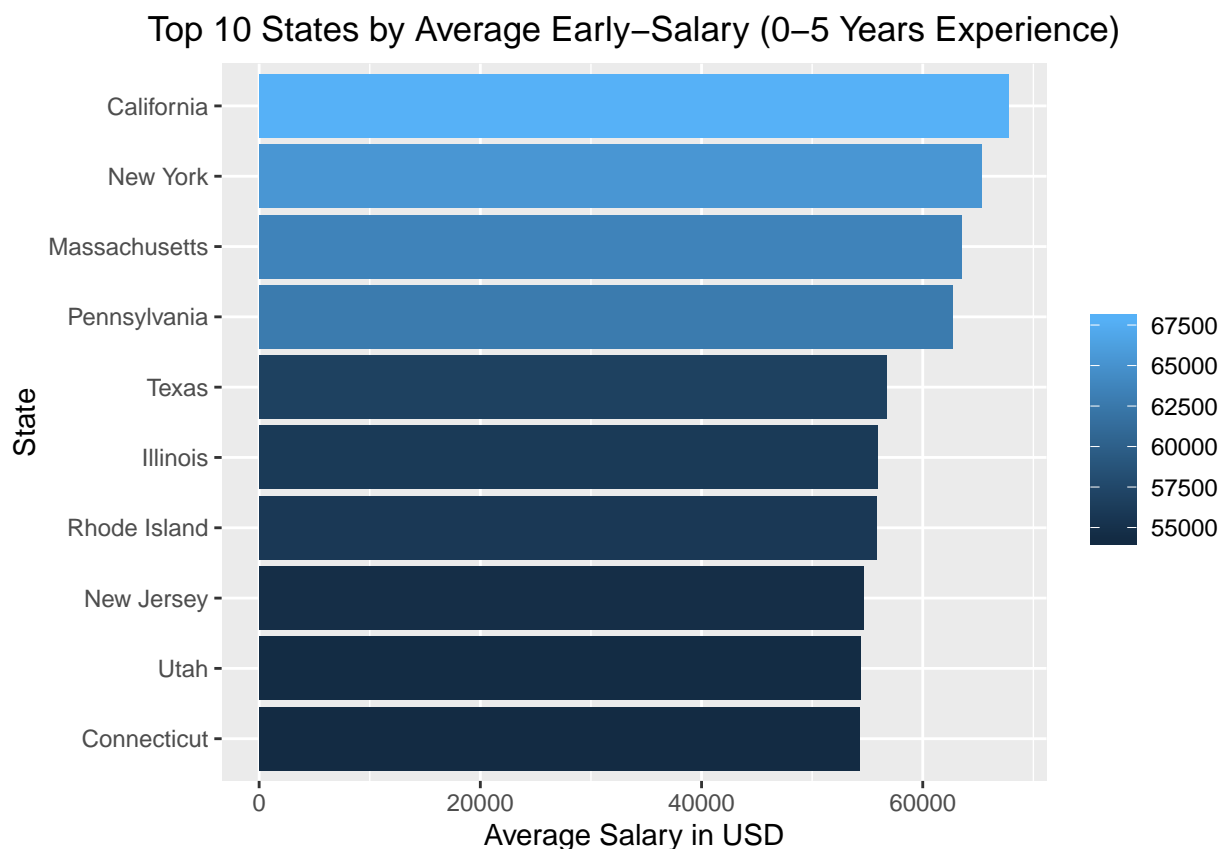
- (1) As people expected, big states like Pennsylvania, New York, California, and Massachusetts have the highest average out-of-state college cost because of high income and high levels of consumption rate.



- (2) There is no surprise that people graduated from colleges/universities big states like Pennsylvania, New York, Massachusetts, and California would have better career performance in terms of early-salary pay(0-5 year experience) and mid-salary pay(10+ year experience) because schools in those states have the most wide range of education resources.

Top 10 States by Average Mid–Salary (10+ Years Experience)





### (3) Interesting Finding:

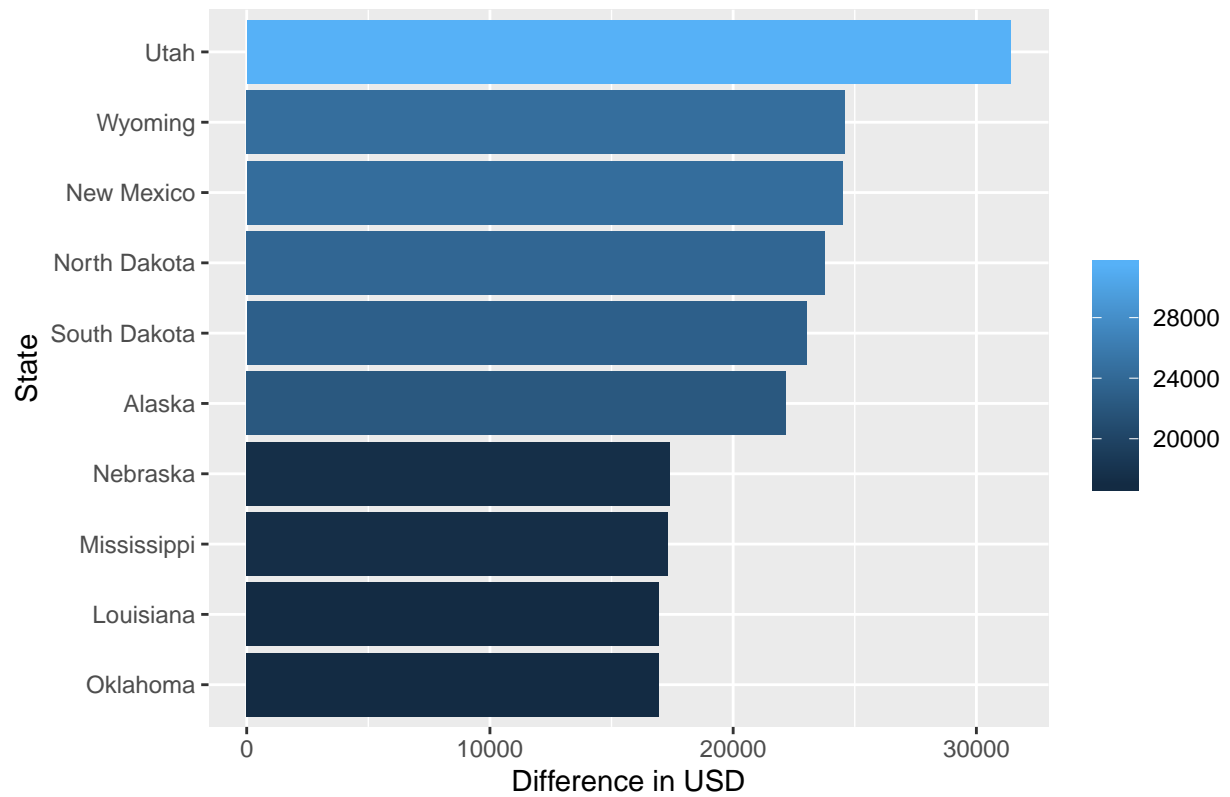
If people consider going to college is a good investment and decide to go to the colleges in big states like Pennsylvania, New York, Massachusetts, and California based on the above plots of career performance in terms of salary, they should also take a look at the bar plots below.

“Mean\_early\_paidoff” and “Mean\_mid\_paidoff” are created based on “Early\_career\_pay\_paidoff” and “Mid\_career\_pay\_paidoff” during previous data cleansing steps in Part II.

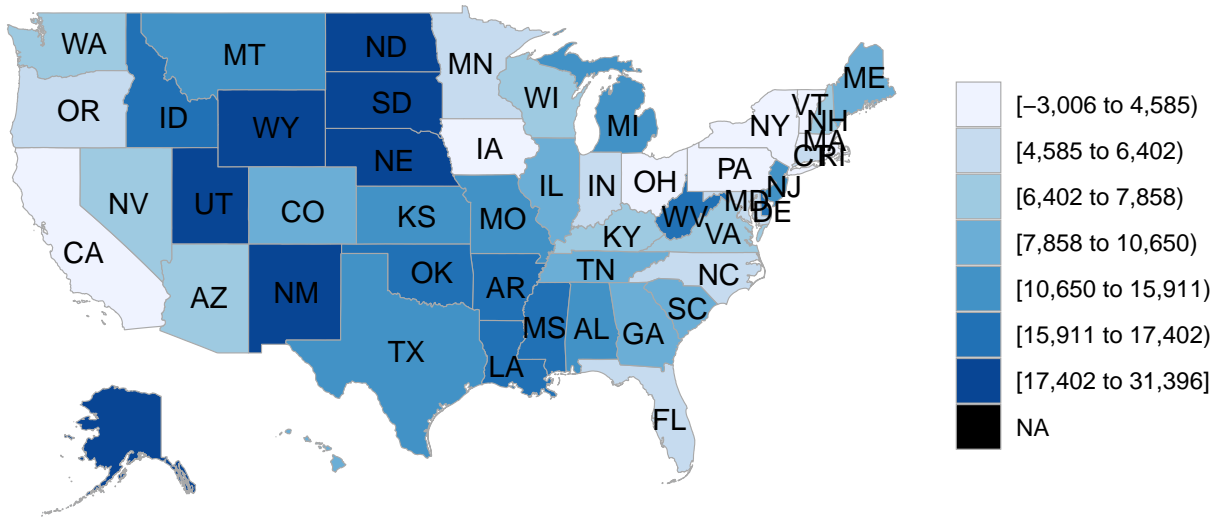
They are defined as the average different between median salary for alumni with 0-5 and 10+ years experience and out of state tuition and fee in different states.

From the below plots and maps, the schools in the states that have the best investment value in terms of “Mean\_early\_paidoff” and “Mean\_mid\_paidoff” are Utah, Wyoming, and New Mexico, etc. The schools in big states like New York and Pennsylvania are not in the top-10 list. One reason would be that those schools in the big states have the most wide range of education resources but at the same time, their college cost is way higher than the schools in other states.

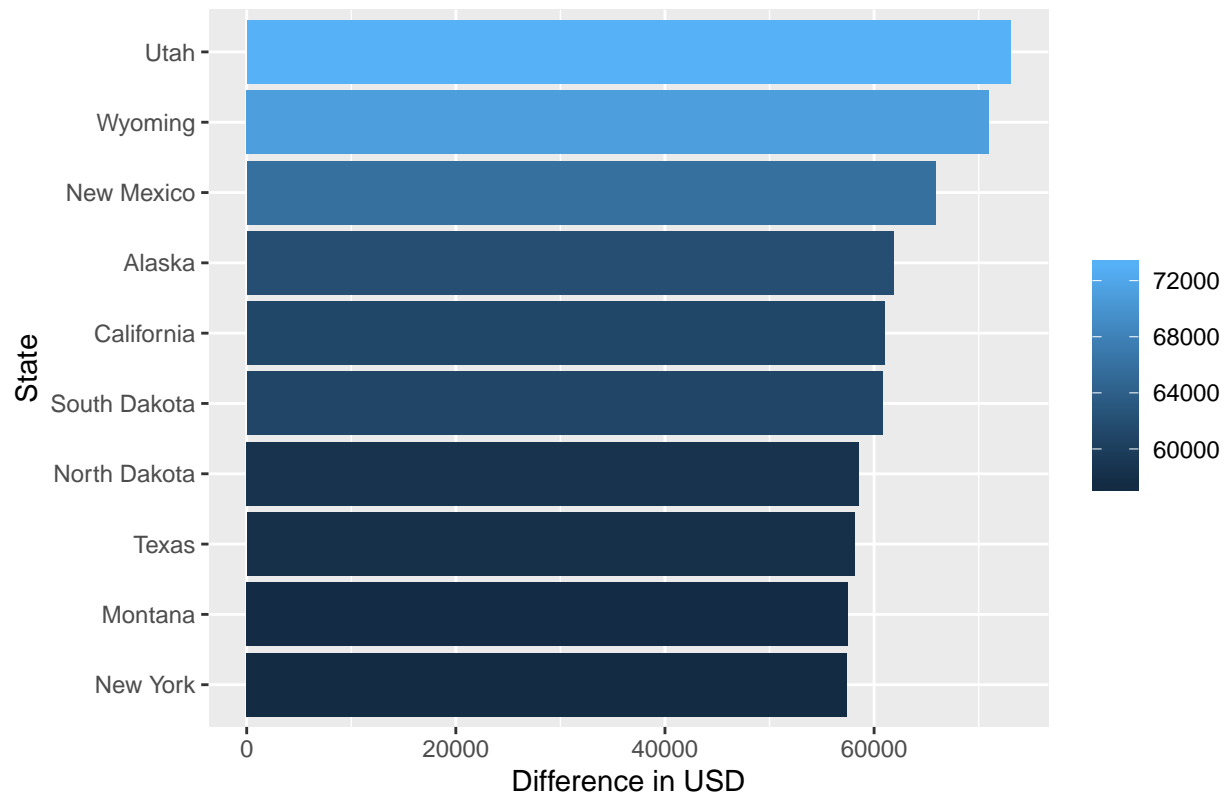
Top 10 States by Average of [Early Salary(0–5 Years Experience) – College Cost]



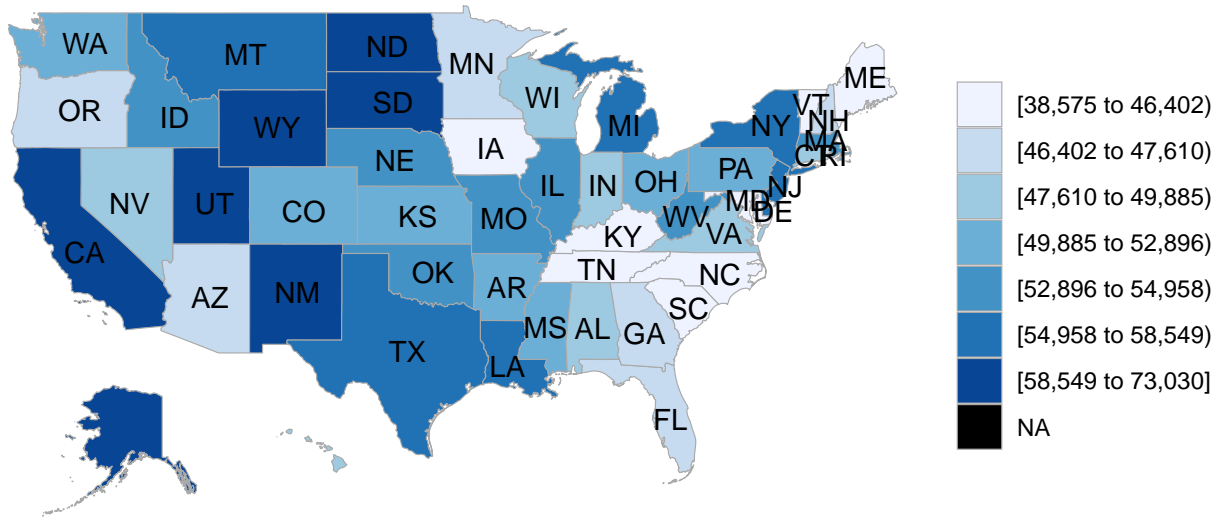
# Average of [Early Salary(0–5 Years Experience) – College Cost] in US



Top 10 States by Average of [Mid Salary(10+ Years Experience) – College Cost]



### Average of [Mid Salary(10+ Years Experience) – College Cost] in US



### 3, Take a look at the final data, “Final\_data” at the level of school type (Private vs Public)

In the average Mid-career Salary(10+ Years Experience) and the average Early-Salary(0-5 Years Experience), private schools outweighs public schools in the U.S. However, in the most important career performance metrics that I created, public schools really shows the advantage. Because of lower total tuition cost, the “Mean\_early\_paidoff”, which represents the average amount of [Early Salary(0-5 Years Experience) - College Cost] for public school is much larger than that in private school.

As a result, if people believe that college is a good investment, public schools should be highly considered.

```
## # A tibble: 2 x 11
##   type count Mean_Early_Care~ Mean_Mid_Career~ Mean_High_Meani~
##   <chr> <int>         <dbl>         <dbl>         <dbl>
## 1 Priv~   382         52397.         95199.         53
## 2 Publ~   240         48810.         87678.         54.4
## # ... with 6 more variables: Mean_STEM_Degree <dbl>,
## #   Mean_Out_Of_State_Cost <dbl>, Mean_Enrollment <dbl>, Mean_Minority <dbl>,
## #   Mean_early_paidoff <dbl>, Mean_mid_paidoff <dbl>
```



## IV. Future Development and Improvement

### 1, Create a dataset with nested states

Save for future development and improvement, such as creating linear regression model for each states to reveal significant impact that each factor might have on the relationship between college cost and career performance.

```
state_nested = Final_data %>%  
  group_by(state)%>%  
  nest()
```

### 2, More data needed

When comparing schools with different length of degrees, the comparison could be bias because there are only 2 private schools.

As for improvement, more private schools in the U.S. should be added into the dataset.

```
## # A tibble: 2 x 11  
##   degree_length count Mean_Early_Care~ Mean_Mid_Career~ Mean_High_Meani~  
##   <chr>          <int>          <dbl>          <dbl>          <dbl>  
## 1 2 Year           2          43400          76850          61.5  
## 2 4 Year          620          51038.          92347.          53.5  
## # ... with 6 more variables: Mean_STEM_Degree <dbl>,  
## #   Mean_Out_Of_State_Cost <dbl>, Mean_Enrollment <dbl>, Mean_Minority <dbl>,  
## #   Mean_early_paidoff <dbl>, Mean_mid_paidoff <dbl>
```