

# In Financial Terms, is going out-of-state for college worth the extra burden?

By Robert Estan

## 1. Introduction

Picking a college is an extremely important part of many young highschool graduates lives and for many out of state options seem very like the perfect oppurtunity to actualize your own potential to some and an intimading financial death sentence to others. The value of a college cannot be measured in it's totality by just financial statistics but hopefully this tutorial will provide an answer to the financial side of the question. This study is being done from the perspective of a highschool graduate in Maryland looking to major in CS.

## 2. Data Curation

We will be very optimistic in our data analysis here and assume our theoretical prospective student has been accept to all 20 of the Universities with the best CS programs in the US. We will take our definition of the best univiersites from usnews.com's ranking of them which can be found here: <https://www.usnews.com/best-graduate-schools/top-science-schools/computer-science-rankings>

### 2.1 Tuition Data

For our first piece of data we will be making a dataset containing the in state and out of state tuitions for every college within our top 20 list of colleges. Some Ivy-League Universities don't distinguish between in-state residents and out of state residents thus we will keep the per-year tuition the same for both categories for those Universities.

We will be entering our data into a custom csv file we made from the USNews ranking website as seen in the source column and read that csv into a dataframe.

```
import pandas as pd

file1 = "tuition.csv"
tuitions = pd.read_csv(file1)

tuitions
```

	university	in-state	out-of-state	source
0	Carnegie Mellon University	63829	63829	https://www.usnews.com/best-colleges/carnegie...
1	Massachusetts Institute of Technology	60156	60156	https://www.usnews.com/best-colleges/massachus...
2	Stanford University	62484	62484	https://www.usnews.com/best-colleges/stanford...
3	University of California--Berkeley	15891	48465	https://www.usnews.com/best-colleges/universit...
4	University of Illinois--Urbana-Champaign	17572	36068	https://www.usnews.com/best-colleges/universit...
5	Cornell University	66014	66014	https://www.usnews.com/best-colleges/cornell-u...
6	Georgia Institute of Technology	11764	32876	https://www.usnews.com/best-colleges/georgia-i...
7	University of Texas--Austin	11698	41070	https://www.usnews.com/best-colleges/universit...
8	University of Washington	12643	41997	https://www.usnews.com/best-colleges/universit...
9	Princeton University	59710	59710	https://www.usnews.com/best-colleges/princeton...
10	University of Michigan--Ann Arbor	17786	57273	https://www.usnews.com/best-colleges/universit...
11	Columbia University	65524	65524	https://www.usnews.com/best-colleges/columbia...
12	California Institute of Technology	63255	63255	https://www.usnews.com/best-colleges/californi...

Next steps: [View recommended plots](#)

## 2.2 Graduating Salary Data

Here we have data about the graduating salary someone will get when leaving one of our top 20 colleges. This data mentions the median graduating salary of someone immediately after college (we call that "starting"), the median salary of someone midway through their career (10 years after college) and the percentiles of salaries for graduates midway through their career. All of this data can be found through this link here: [https://www.wsj.com/public/resources/documents/info-Salaries\\_for\\_Colleges\\_by\\_Type-sort.html](https://www.wsj.com/public/resources/documents/info-Salaries_for_Colleges_by_Type-sort.html)

We will be creating a csv file to read as a dataframe in pandas. We will then be copying and pasta the data from the .html link because it simplifies the process of making the csv file rather than going through the loops of making specifically the table in the html page readable as a dataframe and then cutting only selecting the univerisities we care about from that dataframe.

```
file2 = "graduate_sals.csv"
salaries = pd.read_csv(file2)

salaries
```

	university	Starting	Mid-Career	10th-Percentile	25th-Percentile	75th-Percentile	90th-Percentile
0	Carnegie Mellon University	61800	111000	63300.0	80100.0	150000	209000.0
1	Massachusetts Institute of Technology	72200	126000	76800.0	99200.0	168000	220000.0
2	Stanford University	95200	122900	NaN	NaN	126400	NaN
3	University of California--Berkeley	59900	112000	59500.0	81000.0	149000	201000.0
4	University of Illinois--Urbana-Champaign	52900	96100	48200.0	68900.0	132000	177000.0
5	Cornell University	60300	110000	56800.0	79800.0	160000	210000.0
6	Georgia Institute of Technology	58300	106000	67200.0	85200.0	137000	183000.0
7	University of Texas--Austin	49700	93900	50100.0	67400.0	129000	188000.0
8	University of Washington	48800	85300	47000.0	59800.0	115000	149000.0
9	Princeton University	66500	131000	68900.0	100000.0	190000	261000.0

Next steps: [View recommended plots](#)

But we have a problem, we don't have data for some of the tables in this dataframe so we are going to have to do some imputation to fill in the missing values.

We will be going with median imputation for this dataset since mode imputation is usually used when there are recurring common values in the dataset (which there aren't in our case), hot imputation would have us take values from another college's data which seems counter productive and dangerous and cold imputation would have us take the value from another dataset recording the same thing our dataset is which isn't available at the moment. We could use mean imputation but that's more a matter of personal preference and we would have to round our values down into integers if we used means.

```
import numpy as np
```

```
#here we get the medians of all the columns with missing values
```

```
med_10 = salaries['10th-Percentile'].median()
```

```
med_25 = salaries['25th-Percentile'].median()
```

```
med_90 = salaries['90th-Percentile'].median()
```



```
#here we fill in our missing values with our medians
```

```
salaries['10th-Percentile'] = salaries['10th-Percentile'].fillna(med_10)
```

```
salaries['25th-Percentile'] = salaries['25th-Percentile'].fillna(med_25)
```

```
salaries['90th-Percentile'] = salaries['90th-Percentile'].fillna(med_90)
```

```
salaries
```

	university	Starting	Mid-Career	10th-Percentile	25th-Percentile	75th-Percentile	90th-Percentile		
0	Carnegie Mellon University	61800	111000	63300.0	80100.0	150000	209000.0		
1	Massachusetts Institute of Technology	72200	126000	76800.0	99200.0	168000	220000.0		
2	Stanford University	95200	122900	51500.0	75400.0	126400	190500.0		
3	University of California--Berkeley	59900	112000	59500.0	81000.0	149000	201000.0		
4	University of Illinois--Urbana-Champaign	52900	96100	48200.0	68900.0	132000	177000.0		
5	Cornell University	60300	110000	56800.0	79800.0	160000	210000.0		
6	Georgia Institute of Technology	58300	106000	67200.0	85200.0	137000	183000.0		
7	University of Texas--Austin	49700	93900	50100.0	67400.0	129000	188000.0		
8	University of Washington	48800	85300	47000.0	59800.0	115000	149000.0		
9	Princeton University	66500	131000	68900.0	100000.0	190000	261000.0		

Next steps: [View recommended plots](#)

And voila some well populated datasets ready to be explored.

## ✓ 2.3 Merging our datasets

The astute readers among you will have noticed that our two datasets have two columns that are exactly the same. That's right the univeristy name column. To simplify our work and centralize our data we can merge our datasets using this name column as our basis for merging.

```
proj_data = pd.merge(tuitions, salaries, on="university",how="inner")
```

```
proj_data
```

	university	in-state	out-of-state	source	Starting	Mid-Career	10th-Percentile
0	Carnegie Mellon University	63829	63829	https://www.usnews.com/best-colleges/carnegie-...	61800	111000	63300.0
1	Massachusetts Institute of Technology	60156	60156	https://www.usnews.com/best-colleges/massachus...	72200	126000	76800.0
2	Stanford University	62484	62484	https://www.usnews.com/best-colleges/stanford-...	95200	122900	51500.0
3	University of California--Berkeley	15891	48465	https://www.usnews.com/best-colleges/universit...	59900	112000	59500.0
4	University of Illinois--Urbana-Champaign	17572	36068	https://www.usnews.com/best-colleges/universit...	52900	96100	48200.0
5	Cornell University	66014	66014	https://www.usnews.com/best-colleges/cornell-u...	60300	110000	56800.0
6	Georgia Institute of Technology	11764	32876	https://www.usnews.com/best-colleges/georgia-i...	58300	106000	67200.0
7	University of Texas--Austin	11698	41070	https://www.usnews.com/best-colleges/universit...	49700	93900	50100.0
8	University of Washington	12643	41997	https://www.usnews.com/best-colleges/universit...	48800	85300	47000.0
9	Princeton University	59710	59710	https://www.usnews.com/best-colleges/princeton...	66500	131000	68900.0
10	University of Michigan--Ann Arbor	17786	57273	https://www.usnews.com/best-colleges/universit...	52700	93000	50900.0
11	Columbia University	65524	65524	https://www.usnews.com/best-colleges/columbia-...	59400	107000	50300.0
12	California Institute of Technology	63255	63255	https://www.usnews.com/best-colleges/californi...	75500	123000	51500.0
13	University of California--Los Angeles	13752	46326	https://www.usnews.com/best-colleges/ucla-1315...	52600	101000	51300.0
14	University of California--San Diego	16056	48630	https://www.usnews.com/best-colleges/universit...	51000	101000	51700.0
15	University of Wisconsin--Madison	11205	40603	https://www.usnews.com/best-colleges/universit...	48900	87800	47400.0
16	Harvard University	59076	59076	https://www.usnews.com/best-colleges/harvard-u...	63400	124000	54800.0
	University of						

Next steps:

[View recommended plots](#)

And now we are completely done with our data cleaning portion and ready to move onto analysis

### 3. Exploratory Data Analysis

Time to make some plots! For this section we will be making three plots to understand certain trends in our data.

#### 3.1

Our first plot will be examining the differences between Universities' in-state and out-of-state tuitions using a regular scatter plot.

```

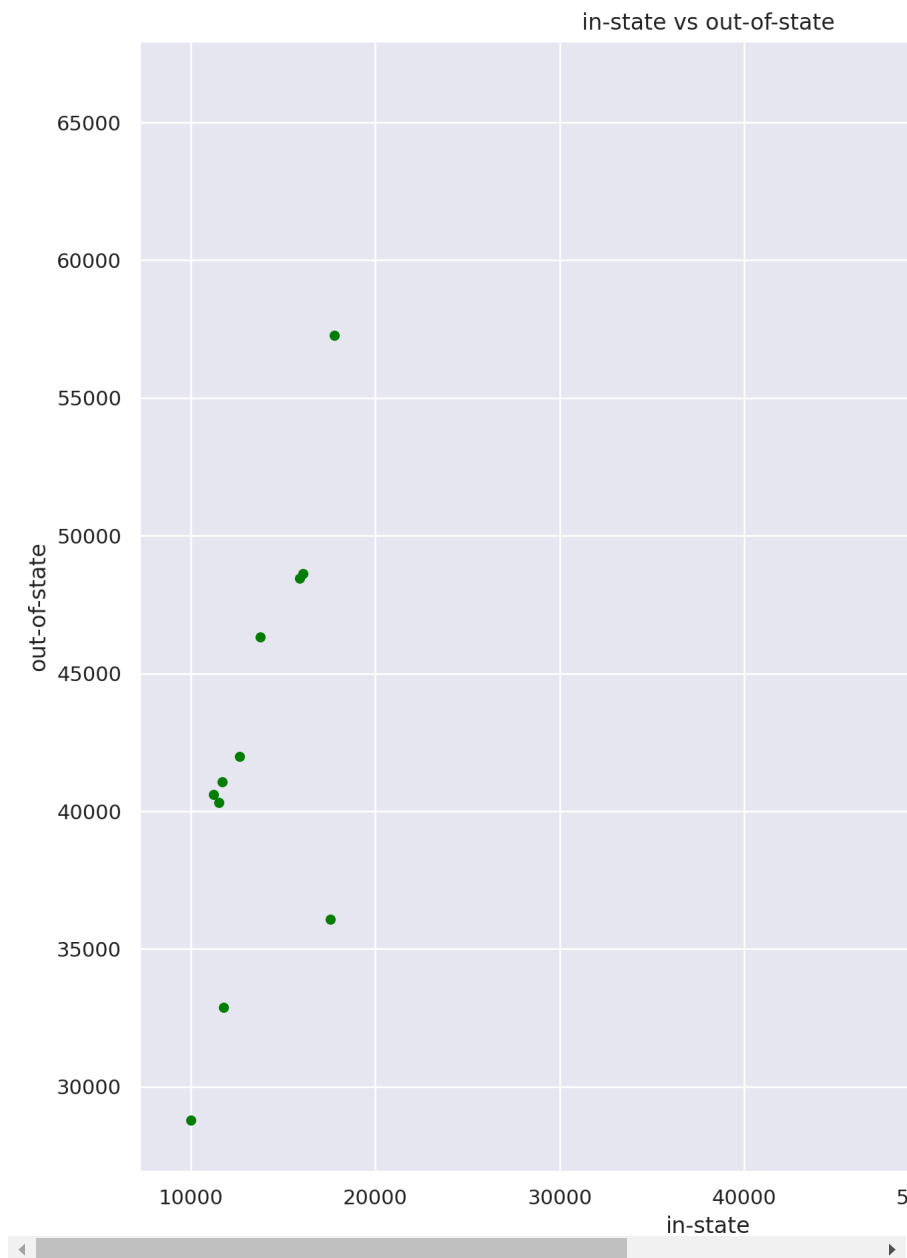
import matplotlib
import matplotlib.pyplot as plt

#setting adjustment which is mainly visual and optional for you
matplotlib.rcParams['figure.figsize'] = [11, 11]

#a function to help us make more scatter plots in the future
def scatter_func(our_frame, x_col, y_col):
    our_frame.plot(kind="scatter", x=x_col, y=y_col, title=x_col + " vs " + y_col, color="green")
    plt.show()

# plotting student populations vs faculty count
scatter_func(proj_data, "in-state", "out-of-state")

```



As we can see here there is a deep and large divide for college tuitions between the colleges whose tuitions cost roughly more than 60k and the colleges whose out-of-state tuitions don't even cost that much. The isolated group of colleges to the top right should be ivy-league colleges since they have large prices that don't change for state residency while the colleges closer to the left are more likely to be public colleges that have state incentive to discourage out-of-state students from applying with higher tuitions.

## ✓ 3.2

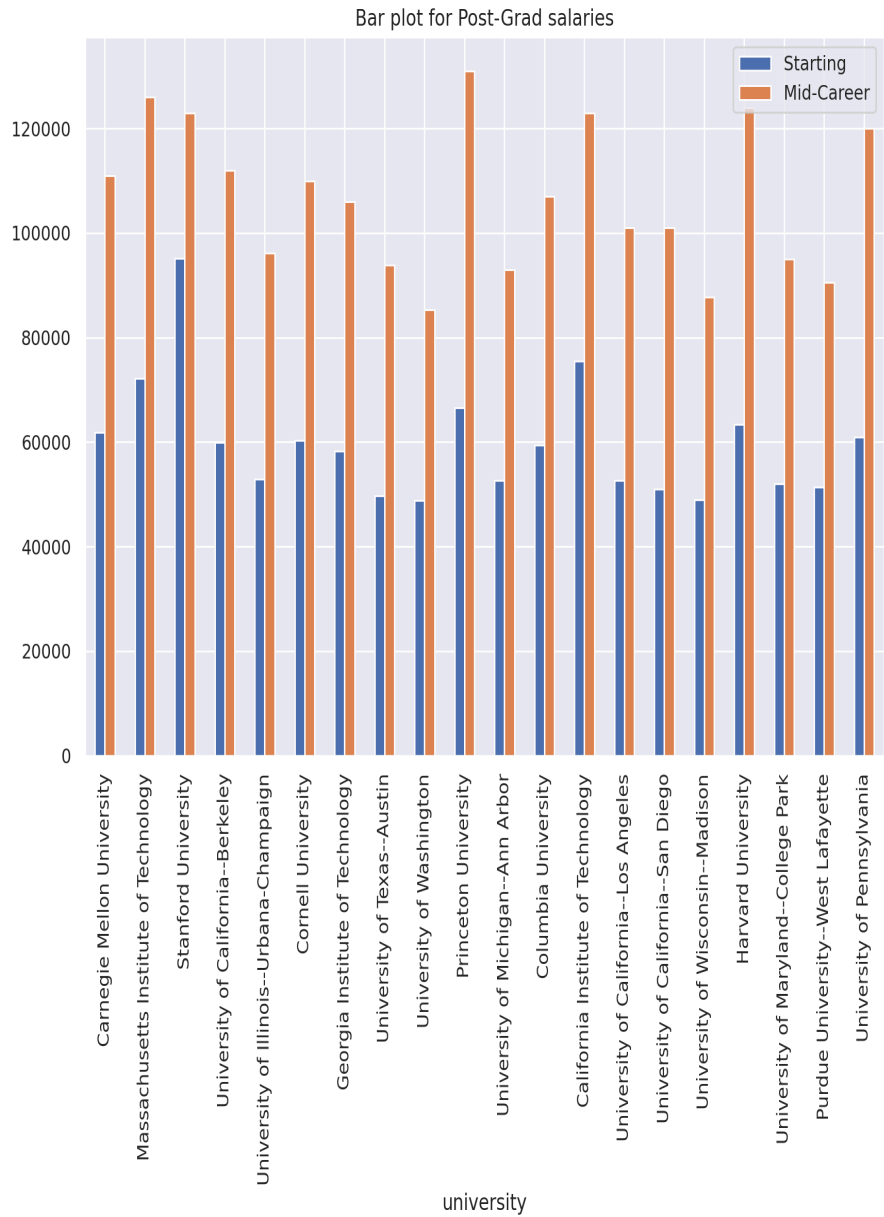
Our second type of plot will be bar plots

```
#size adjustment again
matplotlib.rcParams['figure.figsize'] = [10, 7]

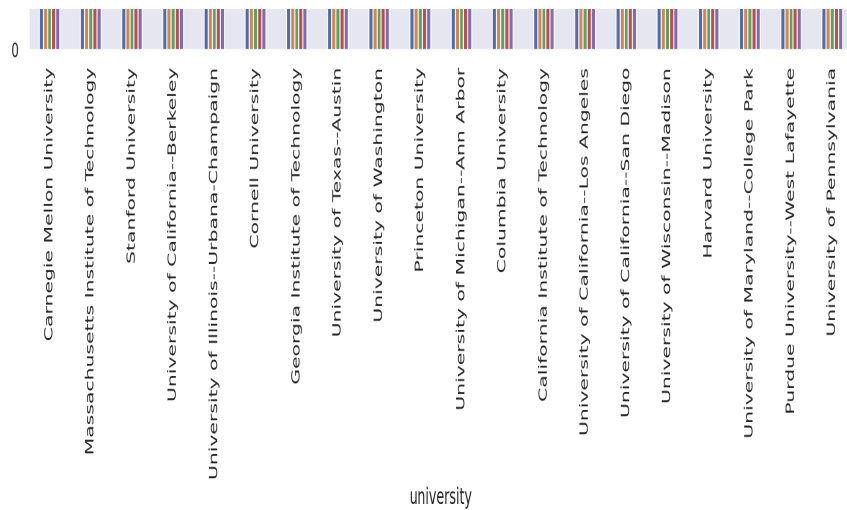
#a function to help us make more bar plots in the future
def bar_func(our_frame, x_col, nam):
    our_frame.set_index(x_col, inplace=True)
    our_frame.plot(kind="bar", title="Bar plot for " + nam + " salaries")
    plt.show()

# plotting starting salary vs mid-career salary
bar_func(proj_data[["university", "Starting", "Mid-Career"]], "university", "Post-Grad")

matplotlib.rcParams['figure.figsize'] = [14, 7]
#plotting midcareer salary variation
bar_func(proj_data[["university", "10th-Percentile", "25th-Percentile", "Mid-Career", "75th-Percentile", "90th-Percentile"]], "university")
```







From these results we can see how large of a variance there is between salaries from post-graduates midway through their career and especially in the 75th percentile and 90th percentiles there are huge jumps in salary. For the first barplot we can see how there is a pretty consistently large gap in salary from the beginning of a career right out of college and halfway through it.

### 3.3

Our final new plot method will cover histograms for the data

```
#size adjustment again
matplotlib.rcParams['figure.figsize'] = [13, 9]

#a function to help us make more histogram plots in the future
def hist_func(our_frame, x_col, nam):
    our_frame.plot(kind="hist", title="Histogram for " + nam + " salaries", xlabel=x_col)
    plt.show()

# plotting Starting Salary compared to Mid-Career Salary
hist_func(proj_data[["Starting", "Mid-Career"]], "Salaries", "Post-Grad")
hist_func(proj_data[["10th-Percentile", "25th-Percentile", "Mid-Career", "75th-Percentile", "90th-Percentile"]], "Salaries", "Post-Grad")
```

