# Cost-of-living-adjusted Salary by College Region

Brian Mahaffey

# Abstract

Cost of living in a specific area has a large impact on quality of life one is able to afford based on salary. In this analysis I normalize the cost of living in each region by CPI-U index and then project into 2018 US total CPI-U inflation adjusted dollars for a fair comparison. The dataset used is from Kaggle, courtesy of The Wall Street Journal, and contains a breakdown of salary ranges based on college geographic region in the United States. The research question I pursued is whether or not the cost-of-living-adjusted salaries among the regional colleges is consistent across regions, or if certain regions have an advantage. The data shows that though there is an increased median salary for higher expense regions, it is mostly offset by lower cost of living in other regions.

# Motivation

How much can you afford based on where you live? This is a question that goes through every adult’s mind at some point. Most big cities, particularly on either coast of the US, have substantially higher costs of living than other regions of the country. In this analysis, I hope to gain a better understanding of whether different geographic regions of colleges have a higher or lower cost-of-living adjusted salary expectation.

# Dataset(s)

- This analysis is based on the dataset "Where it Pays to Attend College" available at https://www.kaggle.com/wsj/college-salaries and provided by The Wall Street Journal.
- I used the following .csv file which contains:
  - salaries-by-region.csv
    - School Name
    - Region (California, Western, Midwestern, Southern, Northeastern)
    - Starting Median Salary
    - Mid-Career Median Salary
    - % change from Starting to Mid-Career median salary
    - Mid-Career 10th, 25th, median, 75th, and 90th percentile salary.
- I also obtained raw CPI-U data for all regions for the year 2018 from the Bureau of Labor Statistics at https://www.bls.gov/regions/home.htm.
- I was further able to obtain and use CPI-U data for California in particular, which was available from https://www.dir.ca.gov/oprl/CAPriceIndex.htm

# Data Preparation and Cleaning

- There were only a few steps of data preparation and cleaning involved in order to make these datasets useable. It is also described in detail in the attached Jupyter Notebook:
  1. Shorten & assign column names to be easier to work with
  2. Remove symbols (‘$’ & ‘,’) and convert strings to floats for all numeric columns.
  3. Append CPI data column to .csv file after gathering the regional information.
  4. Calculate new columns adjusted for CPI-U Inflation Index, and re-based in USA national average 2018 dollar value.
- I did encounter a few problems with the datasets:
  - The regional mid-career quartiles and deciles (25th, 75th, and 90th percentile salaries) had many missing values, so these were largely ignored for the purposes of this analysis.
  - CPI data is calculated on a regional basis, but California is included in the “Western” region’s calculations. It is not possible to back California out of these values, so the “Western” region CPI data ex-California is likely skewed higher. The individual states besides California do not provide CPI values.

# Research Question(s)

The research question I pursued is whether or not the cost-of-living-adjusted salaries among the regional colleges is consistent across regions, or if certain regions have an advantage.

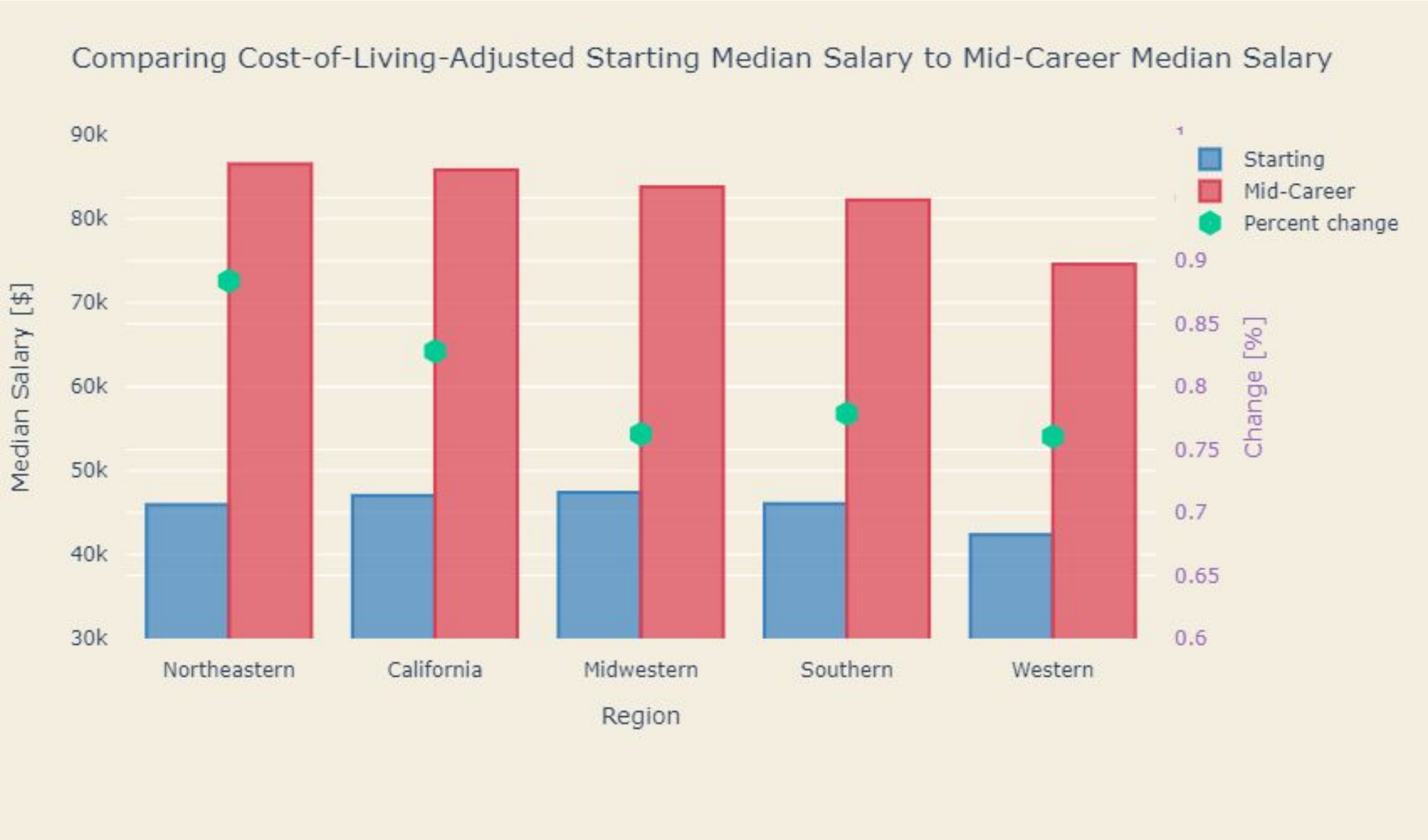
# Methods

Different regions have different costs of living, which can be approximated by the CPI-U index value.
To normalize regional differences in salary for different costs of living, I first gathered the 2018 regional CPI-U values and normalized them back into index basis year dollar values (1982-1984, basis = 100).

I then broadcast these values back into 2018 dollars using the US National Average CPI-U index so that a fair comparison could be completed which takes into account regional cost of living differences.

Once this was complete, I plotted and drew conclusions both visually and through statistical analysis of the results for regional median starting salaries, regional median mid-career salaries, and the % change between the two.

# Findings



Comparing Cost-of-Living-Adjusted Starting Median Salary to Mid-Career Median Salary
Median Salary [$]
90k
80k
70k
60k
50k
40k
30k
Change [%]
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Starting
Mid-Career
Percent change
Northeastern
California
Midwestern
Southern
Western
Region

# Findings

The Original Values:

| Region | Starting Median | Mid-Career Median | % change | Regional CPI-U Value (2018) |
|---|---|---|---|---|
| California | $ 51,032 | $ 93,132 | 82.81% | 272.51 |
| Midwestern | $ 44,225 | $ 78,180 | 76.24% | 234.29 |
| Northeastern | $ 48,496 | $ 91,352 | 88.40% | 265.139 |
| Southern | $ 44,522 | $ 79,505 | 77.86% | 242.737 |
| Western | $ 44,414 | $ 78,200 | 76.05% | 263.263 |

The CPI-Adjusted Values:

| Region | Starting Median | Mid-Career Median | % change |
|---|---|---|---|
| California | $ 47,024 | $ 85,818 | 82.81% |
| Midwestern | $ 47,400 | $ 83,792 | 76.24% |
| Northeastern | $ 45,929 | $ 86,517 | 88.40% |
| Southern | $ 46,057 | $ 82,247 | 77.86% |
| Western | $ 42,363 | $ 74,589 | 76.05% |

# Findings

## Analysis Takeaways:

1. The original raw data told us that Midwestern starting median salaries were **THE LOWEST** of the group at $44,225. After adjusting for cost of living, however, Midwestern has **THE HIGHEST** starting median salary at $47,400.
2. The same can't be said for Midwestern mid-career median salary where it ends up in the middle of the pack at $83,792.
3. For both starting median and mid-career median salary distributions, the $ value and % difference between the highest and lowest regions decreased. This demonstrates that the overall cost of living difference is working as an effective mediator of salary difference. This is as one would expect in a [mostly] capitalist system.

# Limitations

Although the overall analysis proved useful, there are many assumptions and potential sources of bias in the data (or lack of data).

1. The analysis assumes that graduates of a regional college will remain in that region. This is obviously not 100% realistic, but it is the best we can do with the available data.
2. Regional CPI and Salary values are very general over a large area (4 US regions total). For a more thorough analysis, city/MSA breakdowns would be needed where available. Unfortunately due to privacy protection practices employed by the Census Bureau, Bureau of Labor Statistics, and the IRS, this information is anonymized or omitted so a more granular analysis is not possible.
3. The provided “Western” region CPI-U value includes that of California, while the dataset counted California separately. Since California is one of the highest CPI states, the Western CPI value is probably artificially inflated when not including California in its component analysis. The cost of living for the Western region is likely overstated which skews our adjusted results (where it has the lowest starting & mid-career median salary).

# Conclusions

Assuming that CPI is a good measure of inflation and cost of living in a region, the differences in median salaries across different geographical regions are mostly offset by different costs of living. That is to say: median salary distribution is mostly equivalent when taking into account the cost of living.

There are obvious limitations to this analysis, as discussed in the previous page, but I conclude that no region has a significant advantage when viewing salaries as a function of cost of living.

# Acknowledgements

I would like to thank Kaggle and The Wall Street Journal for providing the initial datasets. I would also like to acknowledge the Census Bureau, the Bureau of Labor Statistics, and the IRS for their aggregation of the CPI data and geographical breakdowns.
Since this was initially a Kaggle competition, there were various other analyses of the data which helped me narrow down my research questions and format the data properly. I wish to thank those who came before me for laying the groundwork.
I did not receive feedback from anyone else, and I collected and compiled the CPI data myself directly from the government archives.

# References

One of the best references I encountered during my research was this excellent analysis of post-college salaries: “The Economic Value of College Majors” from the Georgetown University Center on Education and the Workforce, 2015: https://1gyhoq479ufd3yna29x7ubjn-wpengine.netdna-ssl.com/wp-content/uploads/The-Economic-Value-of-College-Majors-Full-Report-web-FINAL.pdf

I also used the plotly python library for charting and spent a lot of time referencing the user guide: https://plot.ly/python/v3/user-guide/#data

# Salaries by College Major Exploration

December 22, 2019

**DSE200x Final Project: Exploring Salaries by College Major**
**By Brian Mahaffey**

This analysis is based on the dataset "Where it Pays to Attend College" available at https://www.kaggle.com/wsj/college-salaries and provided by The Wall Street Journal.

There are 3 separate csv files with information on salaries based on college, region, and academic major.

**1: Import all necessary libraries:**

[79]:
```
import numpy as np
import pandas as pd
import matplotlib
from matplotlib import pyplot as plt
%matplotlib inline
```

Import Plotly libraries for better visualizations after adding to conda:

[80]:
```
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import plotly.figure_factory as ff
```

[ ]:
```
```

**2: Import Local Datasets**

[289]:
```
major = pd.read_csv('degrees-that-pay-back.csv')
college = pd.read_csv('salaries-by-college-type.csv')
region = pd.read_csv('salaries-by-region.csv')
```

[ ]:
```
```

[290]:
```
major.head()
```

[290]:
```
     Undergraduate Major Starting Median Salary Mid-Career Median Salary  \
0             Accounting             $46,000.00               $77,100.00
1  Aerospace Engineering             $57,700.00              $101,000.00
2            Agriculture             $42,600.00               $71,900.00
3           Anthropology             $36,800.00               $61,500.00
4           Architecture             $41,600.00               $76,800.00

   Percent change from Starting to Mid-Career Salary  \
```

```
0                                               67.6
1                                               75.0
2                                               68.8
3                                               67.1
4                                               84.6

  Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                        $42,200.00                        $56,100.00
1                        $64,300.00                        $82,100.00
2                        $36,300.00                        $52,100.00
3                        $33,800.00                        $45,500.00
4                        $50,600.00                        $62,200.00

  Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
0                       $108,000.00                       $152,000.00
1                       $127,000.00                       $161,000.00
2                        $96,300.00                       $150,000.00
3                        $89,300.00                       $138,000.00
4                        $97,000.00                       $136,000.00
```

[291]:
```
major.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 8 columns):
Undergraduate Major                                  50 non-null object
Starting Median Salary                               50 non-null object
Mid-Career Median Salary                             50 non-null object
Percent change from Starting to Mid-Career Salary    50 non-null float64
Mid-Career 10th Percentile Salary                    50 non-null object
Mid-Career 25th Percentile Salary                    50 non-null object
Mid-Career 75th Percentile Salary                    50 non-null object
Mid-Career 90th Percentile Salary                    50 non-null object
dtypes: float64(1), object(7)
memory usage: 3.2+ KB
```

[292]:
```
college.head()
```

[292]:
```
                                    School Name  School Type  \
0   Massachusetts Institute of Technology (MIT)  Engineering
1      California Institute of Technology (CIT)  Engineering
2                             Harvey Mudd College  Engineering
3  Polytechnic University of New York, Brooklyn  Engineering
4                                  Cooper Union  Engineering

  Starting Median Salary Mid-Career Median Salary  \
0             $72,200.00              $126,000.00
1             $75,500.00              $123,000.00
```

```
2             $71,800.00              $122,000.00
3             $62,400.00              $114,000.00
4             $62,200.00              $114,000.00

  Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                       $76,800.00                        $99,200.00
1                              NaN                       $104,000.00
2                              NaN                        $96,000.00
3                       $66,800.00                        $94,300.00
4                              NaN                        $80,200.00

  Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
0                      $168,000.00                       $220,000.00
1                      $161,000.00                               NaN
2                      $180,000.00                               NaN
3                      $143,000.00                       $190,000.00
4                      $142,000.00                               NaN
```

[293]:
```
college.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 269 entries, 0 to 268
Data columns (total 8 columns):
School Name                          269 non-null object
School Type                          269 non-null object
Starting Median Salary               269 non-null object
Mid-Career Median Salary             269 non-null object
Mid-Career 10th Percentile Salary    231 non-null object
Mid-Career 25th Percentile Salary    269 non-null object
Mid-Career 75th Percentile Salary    269 non-null object
Mid-Career 90th Percentile Salary    231 non-null object
dtypes: object(8)
memory usage: 16.9+ KB
```

[294]:
```
region.head()
```

[294]:
```
                                 School Name      Region  \
0                        Stanford University  California
1  California Institute of Technology (CIT)  California
2                          Harvey Mudd College  California
3           University of California, Berkeley  California
4                           Occidental College  California

  Starting Median Salary Mid-Career Median Salary  \
0            $70,400.00              $129,000.00
1            $75,500.00              $123,000.00
2            $71,800.00              $122,000.00
3            $59,900.00              $112,000.00
```

```
4           $51,900.00             $105,000.00

  Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                        $68,400.00                        $93,100.00
1                               NaN                       $104,000.00
2                               NaN                        $96,000.00
3                        $59,500.00                        $81,000.00
4                               NaN                        $54,800.00

  Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary     CPI
0                       $184,000.00                       $257,000.00  272.51
1                       $161,000.00                               NaN  272.51
2                       $180,000.00                               NaN  272.51
3                       $149,000.00                       $201,000.00  272.51
4                       $157,000.00                               NaN  272.51
```

[295]:
```
region.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 320 entries, 0 to 319
Data columns (total 9 columns):
School Name                          320 non-null object
Region                               320 non-null object
Starting Median Salary               320 non-null object
Mid-Career Median Salary             320 non-null object
Mid-Career 10th Percentile Salary    273 non-null object
Mid-Career 25th Percentile Salary    320 non-null object
Mid-Career 75th Percentile Salary    320 non-null object
Mid-Career 90th Percentile Salary    273 non-null object
CPI                                  320 non-null float64
dtypes: float64(1), object(8)
memory usage: 22.6+ KB
```

[296]:
```
cpi.head()
```

[296]:

| | CPI | Year | Region |
|---|---|---|---|
| 0 | 272.510 | 2018 | California |
| 1 | 265.139 | 2018 | Northeastern |
| 2 | 234.290 | 2018 | Midwestern |
| 3 | 242.737 | 2018 | Southern |
| 4 | 263.263 | 2018 | Western |

**Data Cleaning & Preparation**

We can see that the data is formatted well for the most part. There are some missing entries for which we will remove the rows.

The '$' symbols look like the numbers are strings, not ints or floats…let's check

[297]:
```
type(college['Starting Median Salary'][1])
```

[297]: str

The data are strings: let's convert to floats:
But first, let's manipulate the column names to be easier to work with for all three dataframes:

[298]:
```
dataframe_list = [major, college, region]
```

[299]:
```
college_columns = {
    "School Name" : "school",
    "School Type" : "type",
    "Starting Median Salary" : "start_med",
    "Mid-Career Median Salary" : "mid_med",
    "Mid-Career 10th Percentile Salary" : "mid_10",
    "Mid-Career 25th Percentile Salary" : "mid_25",
    "Mid-Career 75th Percentile Salary" : "mid_75",
    "Mid-Career 90th Percentile Salary" : "mid_90"
}

college.rename(columns=college_columns, inplace=True)

region_columns = {
    "School Name" : "school",
    "Region" : "region",
    "Starting Median Salary" : "start_med",
    "Mid-Career Median Salary" : "mid_med",
    "Mid-Career 10th Percentile Salary" : "mid_10",
    "Mid-Career 25th Percentile Salary" : "mid_25",
    "Mid-Career 75th Percentile Salary" : "mid_75",
    "Mid-Career 90th Percentile Salary" : "mid_90",
    'CPI' : 'CPI'
}

region.rename(columns=region_columns, inplace=True)

major_columns = {
    "Undergraduate Major" : "major",
    "Starting Median Salary" : "start_med",
    "Mid-Career Median Salary" : "mid_med",
    "Percent change from Starting to Mid-Career Salary" : "increase",
    "Mid-Career 10th Percentile Salary" : "mid_10",
    "Mid-Career 25th Percentile Salary" : "mid_25",
    "Mid-Career 75th Percentile Salary" : "mid_75",
    "Mid-Career 90th Percentile Salary" : "mid_90"
}

major.rename(columns=major_columns, inplace=True)
```

[ ]:
```

```

[300]:
```
major.head()
```

[300]:
```
                 major  start_med     mid_med  increase     mid_10  \
0           Accounting $46,000.00  $77,100.00      67.6 $42,200.00
1 Aerospace Engineering $57,700.00 $101,000.00      75.0 $64,300.00
2          Agriculture $42,600.00  $71,900.00      68.8 $36,300.00
3         Anthropology $36,800.00  $61,500.00      67.1 $33,800.00
4         Architecture $41,600.00  $76,800.00      84.6 $50,600.00

       mid_25       mid_75       mid_90
0  $56,100.00  $108,000.00  $152,000.00
1  $82,100.00  $127,000.00  $161,000.00
2  $52,100.00   $96,300.00  $150,000.00
3  $45,500.00   $89,300.00  $138,000.00
4  $62,200.00   $97,000.00  $136,000.00
```

Now we'll remove the $ and , and convert to floats

[ ]:

[301]:
```
selected_columns = ["start_med", "mid_med", "mid_10", "mid_25", "mid_75",␣
↪"mid_90"]

for dataframe in dataframe_list:
    for column in selected_columns:
        dataframe[column] = dataframe[column].str.replace("$","")
        dataframe[column] = dataframe[column].str.replace(",","")
        dataframe[column] = pd.to_numeric(dataframe[column])
```

[302]:
```
major.head()
```

[302]:
```
                 major  start_med   mid_med  increase   mid_10   mid_25  \
0           Accounting    46000.0   77100.0      67.6  42200.0  56100.0
1 Aerospace Engineering   57700.0  101000.0      75.0  64300.0  82100.0
2          Agriculture    42600.0   71900.0      68.8  36300.0  52100.0
3         Anthropology    36800.0   61500.0      67.1  33800.0  45500.0
4         Architecture    41600.0   76800.0      84.6  50600.0  62200.0

    mid_75    mid_90
0 108000.0  152000.0
1 127000.0  161000.0
2  96300.0  150000.0
3  89300.0  138000.0
4  97000.0  136000.0
```

[303]:
```
type(major['start_med'][1])
```

[303]: numpy.float64

Good, now we have numeric data which we can analyze

[304]:
```
college.describe()
```

[304]:
```
             start_med        mid_med        mid_10         mid_25  \
count   269.000000     269.000000    231.000000     269.000000
mean  46068.401487   83932.342007  44250.649351   60373.234201
std    6412.616242   14336.191107   8719.612427   11381.348857
min   34800.000000   43900.000000  22600.000000   31800.000000
25%   42000.000000   74000.000000  39000.000000   53200.000000
50%   44700.000000   81600.000000  43100.000000   58400.000000
75%   48300.000000   92200.000000  47400.000000   65100.000000
max   75500.000000  134000.000000  80000.000000  104000.000000

              mid_75         mid_90
count     269.000000     231.000000
mean   116275.092937  157705.627706
std     22952.334054   34823.348157
min     60900.000000   87600.000000
25%    100000.000000  136000.000000
50%    113000.000000  153000.000000
75%    126000.000000  170500.000000
max    234000.000000  326000.000000
```

[305]:
```
major.describe()
```

[305]:
```
             start_med       mid_med    increase        mid_10        mid_25  \
count     50.000000      50.00000   50.000000     50.000000     50.000000
mean   44310.000000   74786.00000   69.274000  43408.000000  55988.000000
std     9360.866217   16088.40386   17.909908  12000.779567  13936.951911
min    34000.000000   52000.00000   23.400000  26700.000000  36500.000000
25%    37050.000000   60825.00000   59.125000  34825.000000  44975.000000
50%    40850.000000   72000.00000   67.800000  39400.000000  52450.000000
75%    49875.000000   88750.00000   82.425000  49850.000000  63700.000000
max    74300.000000  107000.00000  103.500000  71900.000000  87300.000000

              mid_75         mid_90
count      50.000000      50.000000
mean   102138.000000  142766.000000
std     20636.789914   27851.249267
min     70500.000000   96400.000000
25%     83275.000000  124250.000000
50%     99400.000000  145500.000000
75%    118750.000000  161750.000000
max    145000.000000  210000.000000
```

[306]:
```
region.describe()
```

[306]:
```
             start_med        mid_med        mid_10         mid_25  \
count   320.000000     320.000000    273.000000     320.000000
mean  46253.437500   83934.375000  45253.113553   60614.062500
std    6617.038001   15191.443091   8562.834333   11786.436432
min   34500.000000   43900.000000  25600.000000   31800.000000
```

```
25%     42000.000000   73725.000000  39500.000000   53100.000000
50%     45100.000000   82700.000000  43700.000000   59400.000000
75%     48900.000000   93250.000000  48900.000000   66025.000000
max     75500.000000  134000.000000  80000.000000  104000.000000

               mid_75         mid_90         CPI
count      320.000000     273.000000  320.000000
mean    116496.875000  160442.124542  253.162622
std      24104.265214   36785.768186   14.051447
min      60900.000000   85700.000000  234.290000
25%      99825.000000  136000.000000  242.737000
50%     113000.000000  154000.000000  263.263000
75%     129000.000000  178000.000000  265.139000
max     234000.000000  326000.000000  272.510000
```

[307]:

```
region['start_adj'] = region['start_med'] / region['CPI'] * (100*2.51107)
region['midmed_adj'] = region['mid_med'] / region['CPI'] * (100*2.51107)
region['mid10_adj'] = region['mid_10'] / region['CPI'] * (100*2.51107)
region['mid25_adj'] = region['mid_25'] / region['CPI'] * (100*2.51107)
region['mid75_adj'] = region['mid_75'] / region['CPI'] * (100*2.51107)
region['mid90_adj'] = region['mid_90'] / region['CPI'] * (100*2.51107)
region['change'] = (region['mid_med'] - region['start_med']) /␣
 ↪region['start_med']
region['change_adj'] = (region['midmed_adj'] - region['start_adj']) /␣
 ↪region['start_adj']
```

[308]:

```
region.head()
```

[308]:

```
                                     school      region  start_med   mid_med  \
0                       Stanford University  California    70400.0  129000.0
1  California Institute of Technology (CIT)  California    75500.0  123000.0
2                         Harvey Mudd College  California    71800.0  122000.0
3        University of California, Berkeley  California    59900.0  112000.0
4                        Occidental College  California    51900.0  105000.0

    mid_10    mid_25    mid_75    mid_90     CPI     start_adj     midmed_adj  \
0  68400.0   93100.0  184000.0  257000.0  272.51  64870.767311  118868.309420
1      NaN  104000.0  161000.0       NaN  272.51  69570.212102  113339.550842
2      NaN   96000.0  180000.0       NaN  272.51  66160.810979  112418.091079
3  59500.0   81000.0  149000.0  201000.0  272.51  55195.439800  103203.493450
4      NaN   54800.0  157000.0       NaN  272.51  47823.761697   96753.275109

      mid10_adj     mid25_adj      mid75_adj      mid90_adj    change  \
0  63027.847785  85787.903930  169548.596382  236815.159077  0.832386
1           NaN  95831.815346  148355.021834            NaN  0.629139
2           NaN  88460.137243  165862.757330            NaN  0.699164
3  54826.855895  74638.240799  137297.504679  185213.412352  0.869783
4           NaN  50495.995009  144669.182782            NaN  1.023121
```

```
   change_adj
0    0.832386
1    0.629139
2    0.699164
3    0.869783
4    1.023121
```

[309]:
```
major.sort_values(by = 'start_med', ascending=False, inplace=True)
major.head()
```

[309]:
```
                       major  start_med   mid_med  increase   mid_10   mid_25  \
43       Physician Assistant    74300.0   91700.0      23.4  66400.0  75200.0
8       Chemical Engineering    63200.0  107000.0      69.3  71900.0  87300.0
12      Computer Engineering    61400.0  105000.0      71.0  66100.0  84100.0
19    Electrical Engineering    60900.0  103000.0      69.1  69300.0  83800.0
38    Mechanical Engineering    57900.0   93600.0      61.7  63700.0  76200.0

      mid_75    mid_90
43  108000.0  124000.0
8   143000.0  194000.0
12  135000.0  162000.0
19  130000.0  168000.0
38  120000.0  163000.0
```

[ ]:

[ ]:

[ ]:

[311]:
```
major_sort = major.sort_values("mid_med", ascending=False).head(40)

def cut_name(x):
    if len(x) <= 18:
        return x
    else:
        return x[0:15] + "..."

trace1 = go.Bar(
    x = major_sort["major"].apply(cut_name).tolist(),
    y = major_sort["start_med"].tolist(),
    name='Starting',
    marker=dict(
        color='rgba(55, 128, 191, 0.7)',
        line=dict(
            color='rgba(55, 128, 191, 1.0)',
            width=2,
        )
    )
```

```
)
trace2 = go.Bar(
    x = major_sort["major"].apply(cut_name).tolist(),
    y = major_sort["mid_med"].tolist(),
    name='Mid-Career',
    marker=dict(
        color='rgba(219, 64, 82, 0.7)',
        line=dict(
            color='rgba(219, 64, 82, 1.0)',
            width=2,
        )
    )
)

trace3 = go.Scatter(
    x = major_sort["major"].apply(cut_name).tolist(),
    y = major_sort["increase"].tolist(),
    name='Percent change',
    mode = 'markers',
    marker=dict(
        symbol="hexagon-dot",
        size=15
    ),
    yaxis='y2'
)

data = [trace1, trace2, trace3]
layout = go.Layout(
    barmode='group',
    title = 'Comparing Starting Median Salary to Mid-Career Median Salary',
    width=850,
    height=500,
    margin=go.Margin(
        l=75,
        r=75,
        b=120,
        t=80,
        pad=10
    ),
    paper_bgcolor='rgb(244, 238, 225)',
    plot_bgcolor='rgb(244, 238, 225)',
    yaxis = dict(
        title= 'Median Salary [$]',
        anchor = 'x',
        rangemode='tozero'
    ),
    yaxis2=dict(
```

```
        title='Change [%]',
        titlefont=dict(
            color='rgb(148, 103, 189)'
        ),
        tickfont=dict(
            color='rgb(148, 103, 189)'
        ),
        overlaying='y',
        side='right',
        anchor = 'x',
        rangemode = 'tozero',
        dtick = 19.95
    ),
    #legend=dict(x=-.1, y=1.2)
    legend=dict(x=0.1, y=0.05)
)

fig = go.Figure(data=data, layout=layout)
py.iplot(fig)
```

```
C:\Users\BRVR\Anaconda3\lib\site-
packages\plotly\graph_objs\_deprecations.py:410: DeprecationWarning:

plotly.graph_objs.Margin is deprecated.
Please replace it with one of the following more specific types
  - plotly.graph_objs.layout.Margin
```

[ ]:

**Salary by region compared to average cost of living (2018 Average CPI-U)**

- These regions conveniently coordinate with US Census regions (except that California is included as "Western" and not broken out on its own). Luckily California has its own CPI Calculation which I pulled from https://www.dir.ca.gov/oprl/CAPriceIndex.htm

[356]:
```
region_mean = region.groupby('region').mean()
```

[357]:
```
region_mean
```

[357]:
```
                 start_med       mid_med        mid_10        mid_25  \
region
California    51032.142857  93132.142857  47777.272727  67153.571429
Midwestern    44225.352113  78180.281690  43076.562500  57026.760563
Northeastern  48496.000000  91352.000000  49101.219512  65479.000000
Southern      44521.518987  79505.063291  43074.647887  57506.329114
Western       44414.285714  78200.000000  42985.294118  56580.952381
```

```
                   mid_75         mid_90      CPI     start_adj  \
region
California    127350.000000  167909.090909  272.510  47024.066260
Midwestern    107594.366197  147689.062500  234.290  47399.784425
Northeastern  129576.000000  181926.829268  265.139  45929.437284
Southern      109662.025316  152769.014085  242.737  46056.699508
Western       106026.190476  143823.529412  263.263  42363.484587

                midmed_adj     mid10_adj     mid25_adj      mid75_adj  \
region
California    85817.522280  44024.834401  61879.314009  117347.900811
Midwestern    83791.950123  46168.536343  61120.059605  115317.335408
Northeastern  86517.361324  46502.626653  62013.642855  122718.425550
Southern      82246.538137  44559.937739  59489.248795  113443.365417
Western       74589.165207  41000.475760  53968.363232  101130.499204

                  mid90_adj    change  change_adj
region
California    154721.471105  0.828146    0.828146
Midwestern    158289.971476  0.762361    0.762361
Northeastern  172298.682265  0.884000    0.884000
Southern      158036.759207  0.778628    0.778628
Western       137182.570281  0.760452    0.760452
```

[358]:
```
region_sort = region_mean.sort_values("midmed_adj", ascending=False).head(6)
```

[359]:
```
region_start = region_sort['start_adj']
region_mid = region_sort['midmed_adj']
```

[360]:
```
region_start
```

[360]:
```
region
Northeastern    45929.437284
California      47024.066260
Midwestern      47399.784425
Southern        46056.699508
Western         42363.484587
Name: start_adj, dtype: float64
```

[361]:
```
region_mid = pd.Series(region_mid)

#avg_highest2000s = pd.Series(avg_highest2000s)
```

[362]:
```
region_mid
```

[362]:
```
region
Northeastern    86517.361324
California      85817.522280
Midwestern      83791.950123
Southern        82246.538137
```

```
Western         74589.165207
Name: midmed_adj, dtype: float64
```

[380]:
```
trace1 = go.Bar(
    #base = regions,
    x = regions,
    y = region_start,
    name='Starting',
    marker=dict(
        color='rgba(55, 128, 191, 0.7)',
        line=dict(
            color='rgba(55, 128, 191, 1.0)',
            width=2,
        )
    )
)
trace2 = go.Bar(
    x = regions,
    y = region_mid,
    #base = region_mid,
    name='Mid-Career',
    marker=dict(
        color='rgba(219, 64, 82, 0.7)',
        line=dict(
            color='rgba(219, 64, 82, 1.0)',
            width=2,
        )
    )
)

trace3 = go.Scatter(
    x = regions,
    y = region_changeadj,
    name='Percent change',
    mode = 'markers',
    marker=dict(
        symbol="hexagon-dot",
       size=15
   ),
    yaxis='y2'
)

data = [trace1, trace2, trace3]
layout = go.Layout(
    barmode='group',
    title = 'Comparing Cost-of-Living-Adjusted Starting Median Salary to␣
 ↪Mid-Career Median Salary',
```

```
    width=850,
    height=500,
    margin=go.Margin(
        l=75,
        r=75,
        b=120,
        t=80,
        pad=10
    ),
    paper_bgcolor='rgb(244, 238, 225)',
    plot_bgcolor='rgb(244, 238, 225)',
    yaxis = dict(
        title= 'Median Salary [$]',
        anchor = 'x',
        range = (30000, 90000),
    ),
    yaxis2=dict(
        title='Change [%]',
        titlefont=dict(
            color='rgb(148, 103, 189)'
        ),
        tickfont=dict(
            color='rgb(148, 103, 189)'
        ),
        overlaying='y',
        side='right',
        anchor = 'x',
        range = (0.60, 1),

    ),
    xaxis=dict(
    title='Region'),

showlegend=True
)

fig = go.Figure(data=data, layout=layout)
py.iplot(fig)
```

[364]:
```
regions = region_sort.index  #('California', 'Northeastern', 'Midwestern',␣
↪'Southern', 'Western')
```

[366]:
```
adj_comparison = pd.DataFrame
adj_comparison = region_mean
```

[367]:
```
del adj_comparison['start_med']
del adj_comparison['mid_med']
del adj_comparison['mid_10']
```

```
del adj_comparison['mid_25']
del adj_comparison['mid_75']
del adj_comparison['mid_90']
del adj_comparison['mid10_adj']
del adj_comparison['mid25_adj']
del adj_comparison['mid75_adj']
del adj_comparison['mid90_adj']
del adj_comparison['change']
```

[368]:
```
adj_comparison
```

[368]:

| region | CPI | start_adj | midmed_adj | change_adj |
|---|---|---|---|---|
| California | 272.510 | 47024.066260 | 85817.522280 | 0.828146 |
| Midwestern | 234.290 | 47399.784425 | 83791.950123 | 0.762361 |
| Northeastern | 265.139 | 45929.437284 | 86517.361324 | 0.884000 |
| Southern | 242.737 | 46056.699508 | 82246.538137 | 0.778628 |
| Western | 263.263 | 42363.484587 | 74589.165207 | 0.760452 |

[369]:
```
adj_comparison['net_gain'] = adj_comparison['midmed_adj'] -␣
↪adj_comparison['start_adj']
```

[370]:
```
adj_comparison
```

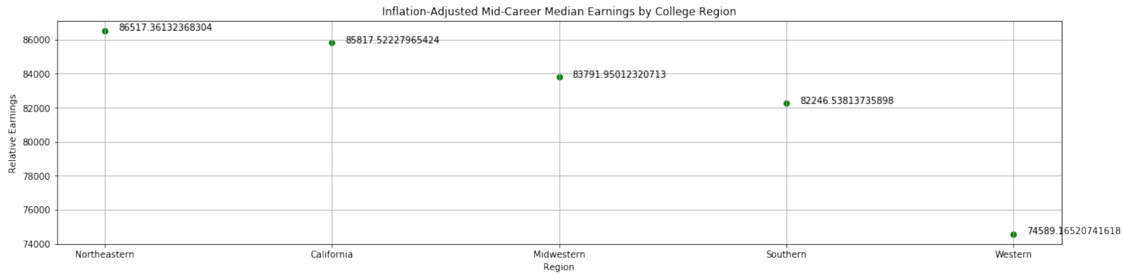
[370]:

| region | CPI | start_adj | midmed_adj | change_adj | net_gain |
|---|---|---|---|---|---|
| California | 272.510 | 47024.066260 | 85817.522280 | 0.828146 | 38793.456020 |
| Midwestern | 234.290 | 47399.784425 | 83791.950123 | 0.762361 | 36392.165698 |
| Northeastern | 265.139 | 45929.437284 | 86517.361324 | 0.884000 | 40587.924040 |
| Southern | 242.737 | 46056.699508 | 82246.538137 | 0.778628 | 36189.838630 |
| Western | 263.263 | 42363.484587 | 74589.165207 | 0.760452 | 32225.680620 |

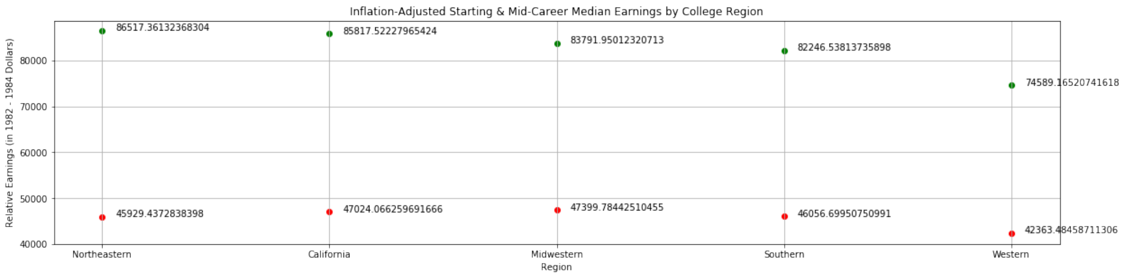
[ ]:

[208]:
```
plt.figure(figsize=(20, 10))

ax1 = plt.subplot(2,1,1)
x = [x for x in range(0, 5)]
xticks_region_list = regions
y = region_sort['midmed_adj']
plt.xticks(range(len(x)), xticks_region_list)
plt.scatter(x,y, color='g')
plt.autoscale(tight=False)
plt.title('Inflation-Adjusted Mid-Career Median Earnings by College Region')
plt.xlabel('Region')
plt.ylabel('Relative Earnings')
plt.grid(True)


for i,j in enumerate( y ):
```

```
    ax1.annotate( j, ( x[i] + 0.06, y[i] + 0.03))

plt.show()
```




[223]:
```
plt.figure(figsize=(20, 10))

ax1 = plt.subplot(2,1,1)
x = [x for x in range(0, 5)]
xticks_region_list = regions
y = region_sort['midmed_adj']
a = region_sort['start_adj']
plt.xticks(range(len(x)), xticks_region_list)
plt.scatter(x,y, color='g')
plt.autoscale(tight=False)
plt.title('Inflation-Adjusted Starting & Mid-Career Median Earnings by College␣
 ↪Region')
plt.xlabel('Region')
plt.ylabel('Relative Earnings (in 1982 - 1984 Dollars)')
plt.grid(True)

plt.scatter(x,a, color='r')


for i,j in enumerate( y ):
    ax1.annotate( j, ( x[i] + 0.06, y[i] + 0.03))

for k,l in enumerate( a ):
    ax1.annotate( l, ( x[k] + 0.06, a[k] + 0.03))


plt.show()
```


Inflation-Adjusted Starting & Mid-Career Median Earnings by College Region
Relative Earnings (in 1982 - 1984 Dollars)
Region
Northeastern
California
Midwestern
Southern
Western
86517.36132368304
85817.52227965424
83791.95012320713
82246.53813735898
74589.16520741618
45929.4372838398
47024.066259691666
47399.78442510455
46056.69950750991
42363.48458711306
40000
50000
60000
70000
80000

| Region | Starting Median | Mid-Career Median | % change | Regional CPI-U Value (2018) |
|---|---|---|---|---|
| California | $ 51,032 | $ 93,132 | 82.81% | 272.51 |
| Midwestern | $ 44,225 | $ 78,180 | 76.24% | 234.29 |
| Northeastern | $ 48,496 | $ 91,352 | 88.40% | 265.139 |
| Southern | $ 44,522 | $ 79,505 | 77.86% | 242.737 |
| Western | $ 44,414 | $ 78,200 | 76.05% | 263.263 |

| Original | Starting | Mid |
|---|---|---|
| Biggest Difference | $ 6,806.79 | $ 14,951.86 |
| as a % | 13.34% | 16.05% |

| Region | Starting Median | Mid-Career Median | % change |
|---|---|---|---|
| California | $ 47,024 | $ 85,818 | 82.81% |
| Midwestern | $ 47,400 | $ 83,792 | 76.24% |
| Northeastern | $ 45,929 | $ 86,517 | 88.40% |
| Southern | $ 46,057 | $ 82,247 | 77.86% |
| Western | $ 42,363 | $ 74,589 | 76.05% |

| Adjusted | Starting | Mid |
|---|---|---|
| Biggest Difference | $ 5,036.30 | $ 11,928.20 |
| as a % | 10.71% | 13.79% |