Does Your College Decision Determine Future Salary Potential?

Introduction:

The goal of this Final Technical Report is to discover if a student's college decisions predetermine their future salary potential. Our client, ACT, Inc. is a non-profit organization. Their primary focus is assessments and supporting K-12 professionals. The results of this report will help ACT, Inc. update materials for their clientele.
We will be utilizing the Wall Street Journal's dataset from Kaggle entitled "Where it Pays to Attend College" as a starting point. https://www.kaggle.com/wsj/college-salaries

Business Understanding:

ACT, Inc. is a national leader in college preparation and testing. They have 60 years of research in the field. Although they are widely known for their ACT test, they provide solutions for all ages and career stages. The results of this report will help them with their mission to offer unique solutions for all students.

Specifically, the results of this program will assist with ACT's Advisory Council of Counselors. This council addresses the ACT Profile initiative which involves personalized free college and career planning with a focus on reaching underserved demographics.
http://www.act.org/content/act/en/products-and-services/act-profile.html

Data Understanding:

We utilized the following data sources:
The Wall Street Journal's Kaggle dataset "Where it Pays to Attend College" has three sets of data:

1. 'degrees-that-pay-back.csv' - lists undergraduate majors (Accounting, Agriculture, etc.) and salary information (Degrees),
2. 'salaries-by-school-type.csv' - lists school name, school type (Engineering, State, Party, etc.), and salary information (College Type),
3. 'salaries-by-region.csv' - lists school name, school region (Northeast, Midwest, Southern, etc.) and salary information (Region).

In addition, the US Census Bureau's Kaggle dataset "Estimate of Median Household Income Group Series" was used to provide insight into cost of living measurements. This dataset has 22 US counties' individual data containing median household income values and dates (MHHI).

These datasets provide the opportunity to measure undergraduate degree choices, school types, salary outcomes and median household income data in a meaningful way.

Data Exploration/Preparation:

The datasets were manipulated directly in Kaggle using an IPython Notebook HTML.
https://www.kaggle.com/debdillerharris/mds556-project

All data exploration and modeling was conducted in the Python language utilizing the following libraries/modules:

- NumPy - Part of Python SciPy library for scientific computing
- Pandas - Python library for data analysis
- Plotly - Python library for graphing
- FigureFactory - Python module that creates additional charts not available in Plotly
- Cufflinks - Python library that binds Plotly directly to Pandas dataframes
- SciPy - Part of the core Python library inside the SciPy stack
- Statsmodels - Python module for conducting statistical tests, and statistical data exploration

```
###### This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
import plotly.plotly as py
import plotly.graph_objs as go
from plotly.tools import FigureFactory as FF
import cufflinks as cf
import scipy


import statsmodels
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

The data sets were also imported:

```
import os
print(os.listdir("../input"))

import glob

# Any results you write to the current directory are saved as output.
```

```
['college-salaries', 'estimate-of-median-household-income-group-series']
```

First, we printed out the list of 22 counties inside the Median Household Income group:

```
dict_keys(['santa-clara-county-ca', 'montgomery-county-md', 'san-francisco-county-city-ca', 's
t.-louis-county-mo', 'fulton-county-ga', 'harris-county-tx', 'fairfax-county-va', 'miami-dade-
county-fl', 'cook-county-il', 'denver-county-co', 'westchester-county-ny', 'alameda-county-c
a', 'orange-county-ca', 'sonoma-county-ca', 'dallas-county-tx', 'bergen-county-nj', 'philadelp
hia-county-city-pa', 'san-diego-county-ca', 'king-county-wa', 'milwaukee-county-wi', 'st.-loui
s-city-mo', 'los-angeles-county-ca'])
```

Fairfax County, Virginia revealed the following values:

```
hhi_data['fairfax-county-va'].tail()
```

	realtime_start	realtime_end	value	date
19	2018-12-12	2018-12-12	106690	2012-01-01
20	2018-12-12	2018-12-12	110658	2013-01-01
21	2018-12-12	2018-12-12	110507	2014-01-01
22	2018-12-12	2018-12-12	112844	2015-01-01
23	2018-12-12	2018-12-12	115518	2016-01-01

We realized that the “value” column is an aggregate of the household income which could be more than one individual. However, all counties have the same configuration so the effect would be the same on all the counties.

Next we began to structure the Median Household Income(MHHI) raw data for analysis. We constructed a Pandas DataFrame. We wanted a table with county names, the corresponding state and the income text (string) converted into a numerical form. The county column was populated with the data key names, the state was configured by stripping the last two values of the county strings and the value was converted to a numeric with one decimal point. Then we sorted alphabetically by state.

<u>Input:</u>

```
counties = list(hhi_data.keys())
hhi_2016 = pd.DataFrame(dict(
    county=counties,
    state=[c[-2:] for c in counties],
    value=[float(hhi_data[c].value.iloc[-1]) for c in counties]
)).sort_values("state")
hhi_2016
```

Output:

	county	state	value
0	santa-clara-county-ca	ca	110843.0
17	san-diego-county-ca	ca	70693.0
13	sonoma-county-ca	ca	73496.0
12	orange-county-ca	ca	81642.0
11	alameda-county-ca	ca	89472.0
21	los-angeles-county-ca	ca	61308.0
2	san-francisco-county-city-ca	ca	101873.0
9	denver-county-co	co	61038.0
7	miami-dade-county-fl	fl	45886.0
4	fulton-county-ga	ga	62824.0
8	cook-county-il	il	60025.0
1	montgomery-county-md	md	99604.0
20	st.-louis-city-mo	mo	39954.0
3	st.-louis-county-mo	mo	62756.0
15	bergen-county-nj	nj	93205.0
10	westchester-county-ny	ny	89380.0
16	philadelphia-county-city-pa	pa	41514.0
14	dallas-county-tx	tx	54429.0
5	harris-county-tx	tx	56415.0
6	fairfax-county-va	va	115518.0
18	king-county-wa	wa	85907.0
19	milwaukee-county-wi	wi	47666.0

What are the unique states within this dataset?

```
hhi_2016.state.unique()
```

```
array(['ca', 'co', 'fl', 'ga', 'il', 'md', 'mo', 'nj', 'ny', 'pa', 'tx',
       'va', 'wa', 'wi'], dtype=object)
```

Our next task was to map each state to the corresponding Regions listed within the WSJ's dataset. This will enable us to have a baseline median household income to compare against the salary data.

Input:

```
region_look_up = {
    'ca': 'California',
    'co': 'Western',
    'fl': "Southern",
    'ga': "Southern",
    'il': "Midwestern",
    'md': "Northeastern",
    'mo': "Midwestern",
    'nj': 'Northeastern',
    'ny': "Northeastern",
    'pa': "Northeastern",
    'tx': "Southern",
    'va': "Southern",
    'wa': "Western",
    'wi': "Midwestern"}
hhi_2016 = hhi_2016.assign(Region=[region_look_up[s] for s in hhi_2016.state])
hhi_2016
```

Output:

	county	state	value	Region
0	santa-clara-county-ca	ca	110843.0	California
17	san-diego-county-ca	ca	70693.0	California
13	sonoma-county-ca	ca	73496.0	California
12	orange-county-ca	ca	81642.0	California
11	alameda-county-ca	ca	89472.0	California
21	los-angeles-county-ca	ca	61308.0	California
2	san-francisco-county-city-ca	ca	101873.0	California
9	denver-county-co	co	61038.0	Western
7	miami-dade-county-fl	fl	45886.0	Southern
4	fulton-county-ga	ga	62824.0	Southern
8	cook-county-il	il	60025.0	Midwestern
1	montgomery-county-md	md	99604.0	Northeastern
20	st.-louis-city-mo	mo	39954.0	Midwestern
3	st.-louis-county-mo	mo	62756.0	Midwestern
15	bergen-county-nj	nj	93205.0	Northeastern
10	westchester-county-ny	ny	89380.0	Northeastern
16	philadelphia-county-city-pa	pa	41514.0	Northeastern
14	dallas-county-tx	tx	54429.0	Southern
5	harris-county-tx	tx	56415.0	Southern
6	fairfax-county-va	va	115518.0	Southern
18	king-county-wa	wa	85907.0	Western
19	milwaukee-county-wi	wi	47666.0	Midwestern

We wanted all the median values for each region. Using the 'groupby' function the mean was calculated for each region.

Input:

```
regional_median_income = hhi_2016.groupby("Region").mean()
regional_median_income
```

Output:

	value
Region	
California	84189.571429
Midwestern	52600.250000
Northeastern	80925.750000
Southern	67014.400000
Western	73472.500000

The California region has the highest median income followed closely by the Northeastern region. The Midwestern median income is significantly lower.

Our subsequent task was to explore the WSJ's College Salaries datasets and manipulate the data as needed.

Input:

```
raw_region = pd.read_csv("../input/college-salaries/salaries-by-region.csv")
raw_college_type = pd.read_csv("../input/college-salaries/salaries-by-college-type.csv")
raw_degrees = pd.read_csv("../input/college-salaries/degrees-that-pay-back.csv")

print("Region:", raw_region.shape, raw_region.columns)
print("college_type:", raw_college_type.shape, raw_college_type.columns)
print("degrees:", raw_degrees.shape, raw_degrees.columns)
raw_degrees.head(3)
```

```
Region: (320, 8) Index(['School Name', 'Region', 'Starting Median Salary',
       'Mid-Career Median Salary', 'Mid-Career 10th Percentile Salary',
       'Mid-Career 25th Percentile Salary',
       'Mid-Career 75th Percentile Salary',
       'Mid-Career 90th Percentile Salary'],
      dtype='object')
college_type: (269, 8) Index(['School Name', 'School Type', 'Starting Median Salary',
       'Mid-Career Median Salary', 'Mid-Career 10th Percentile Salary',
       'Mid-Career 25th Percentile Salary',
       'Mid-Career 75th Percentile Salary',
       'Mid-Career 90th Percentile Salary'],
      dtype='object')
degrees: (50, 8) Index(['Undergraduate Major', 'Starting Median Salary',
       'Mid-Career Median Salary',
       'Percent change from Starting to Mid-Career Salary',
       'Mid-Career 10th Percentile Salary',
       'Mid-Career 25th Percentile Salary',
       'Mid-Career 75th Percentile Salary',
       'Mid-Career 90th Percentile Salary'],
      dtype='object')
```

Output:

	Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	Percent change from Starting to Mid-Career Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Accounting	$46,000.00	$77,100.00	67.6	$42,200.00	$56,100.00	$108,000.00	$152,000.00
1	Aerospace Engineering	$57,700.00	$101,000.00	75.0	$64,300.00	$82,100.00	$127,000.00	$161,000.00
2	Agriculture	$42,600.00	$71,900.00	68.8	$36,300.00	$52,100.00	$96,300.00	$150,000.00

We have three datasets:

- Region - 320 rows with 8 columns
- College Type - 269 rows with 8 columns
- Degrees - 50 rows with 8 columns.

Note that each dataset shares the same 6 columns:

- Starting Median Salary
- Mid-Career Median Salary
- Mid-Career 10th Percentile Salary
- Mid-Career 25th Percentile Salary
- Mid-Career 75th Percentile Salary
- Mid-Career 90th Percentile Salary

We need to conduct data cleansing on the three datasets. We need to remove the ‘$’ and ‘,’ and ‘.’ from the salary columns.

Degrees Dataset

First, we selected the columns in the Degrees dataset that had salary data. We excluded any column with the text “change” because we didn’t want to select the “Percent change from Starting to Mid-Career Salary” column.

```
cols = [c for c in raw_degrees.columns if "Salary" in c and not "change" in c]
cols
```

```
['Starting Median Salary',
 'Mid-Career Median Salary',
 'Mid-Career 10th Percentile Salary',
 'Mid-Career 25th Percentile Salary',
 'Mid-Career 75th Percentile Salary',
 'Mid-Career 90th Percentile Salary']
```

Then using Regex we stripped the ‘$’ and ‘,’ and ‘.’ from the salary columns and formatted them from strings to numerical values.

Input:

```
degrees = raw_degrees
degrees[cols] = degrees[cols].replace({'\$': '',",": ''}, regex=True).astype(float) # stripped
 the characters and
#converted to numerical value "float"
degrees.head()
```

Output:

	Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	Percent change from Starting to Mid-Career Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Accounting	46000.0	77100.0	67.6	42200.0	56100.0	108000.0	152000.0
1	Aerospace Engineering	57700.0	101000.0	75.0	64300.0	82100.0	127000.0	161000.0
2	Agriculture	42600.0	71900.0	68.8	36300.0	52100.0	96300.0	150000.0
3	Anthropology	36800.0	61500.0	67.1	33800.0	45500.0	89300.0	138000.0
4	Architecture	41600.0	76800.0	84.6	50600.0	62200.0	97000.0	136000.0

Region Dataset

Input:

```
raw_region.head()
```

Output:

	School Name	Region	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Stanford University	California	$70,400.00	$129,000.00	$68,400.00	$93,100.00	$184,000.00	$257,000.00
1	California Institute of Technology (CIT)	California	$75,500.00	$123,000.00	NaN	$104,000.00	$161,000.00	NaN
2	Harvey Mudd College	California	$71,800.00	$122,000.00	NaN	$96,000.00	$180,000.00	NaN
3	University of California, Berkeley	California	$59,900.00	$112,000.00	$59,500.00	$81,000.00	$149,000.00	$201,000.00
4	Occidental College	California	$51,900.00	$105,000.00	NaN	$54,800.00	$157,000.00	NaN

We had to repeat the same two processes:

- selected the columns in the Region dataset that had salary data
- Regex to remove the ‘$’ and ‘,’ and ‘.’ from the salary columns and formatted them from strings to numerical values.

```
cols = [c for c in raw_region.columns if "Salary" in c ]
cols
```

```
['Starting Median Salary',
 'Mid-Career Median Salary',
 'Mid-Career 10th Percentile Salary',
 'Mid-Career 25th Percentile Salary',
 'Mid-Career 75th Percentile Salary',
 'Mid-Career 90th Percentile Salary']
```

Input:

```
region = raw_region
region[cols] = region[cols].replace({'\$': '',",": ''}, regex=True).astype(float) # stripped th
e characters and
#converted to numerical value "float"
region.head()
```

Output:

	School Name	Region	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Stanford University	California	70400.0	129000.0	68400.0	93100.0	184000.0	257000.0
1	California Institute of Technology (CIT)	California	75500.0	123000.0	NaN	104000.0	161000.0	NaN
2	Harvey Mudd College	California	71800.0	122000.0	NaN	96000.0	180000.0	NaN
3	University of California, Berkeley	California	59900.0	112000.0	59500.0	81000.0	149000.0	201000.0
4	Occidental College	California	51900.0	105000.0	NaN	54800.0	157000.0	NaN

College Type Dataset

Again, we repeated the same two processes:

- selected the columns in the College Type dataset that had salary data
- Regex to remove the '$' and ',' and '.' from the salary columns and formatted them from strings to numerical values.

Input:

```
raw_college_type.head()
```

Output:

	School Name	School Type	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Massachusetts Institute of Technology (MIT)	Engineering	$72,200.00	$126,000.00	$76,800.00	$99,200.00	$168,000.00	$220,000.00
1	California Institute of Technology (CIT)	Engineering	$75,500.00	$123,000.00	NaN	$104,000.00	$161,000.00	NaN
2	Harvey Mudd College	Engineering	$71,800.00	$122,000.00	NaN	$96,000.00	$180,000.00	NaN
3	Polytechnic University of New York, Brooklyn	Engineering	$62,400.00	$114,000.00	$66,800.00	$94,300.00	$143,000.00	$190,000.00
4	Cooper Union	Engineering	$62,200.00	$114,000.00	NaN	$80,200.00	$142,000.00	NaN

```
cols = [c for c in raw_college_type.columns if "Salary" in c ]
cols
```

```
['Starting Median Salary',
 'Mid-Career Median Salary',
 'Mid-Career 10th Percentile Salary',
 'Mid-Career 25th Percentile Salary',
 'Mid-Career 75th Percentile Salary',
 'Mid-Career 90th Percentile Salary']
```

Input:

```
college_type = raw_college_type
college_type[cols] = college_type[cols].replace({'\$': '',",": ''}, regex=True).astype(float) #
 stripped the characters and
#converted to numerical value "float"
college_type.head()
```

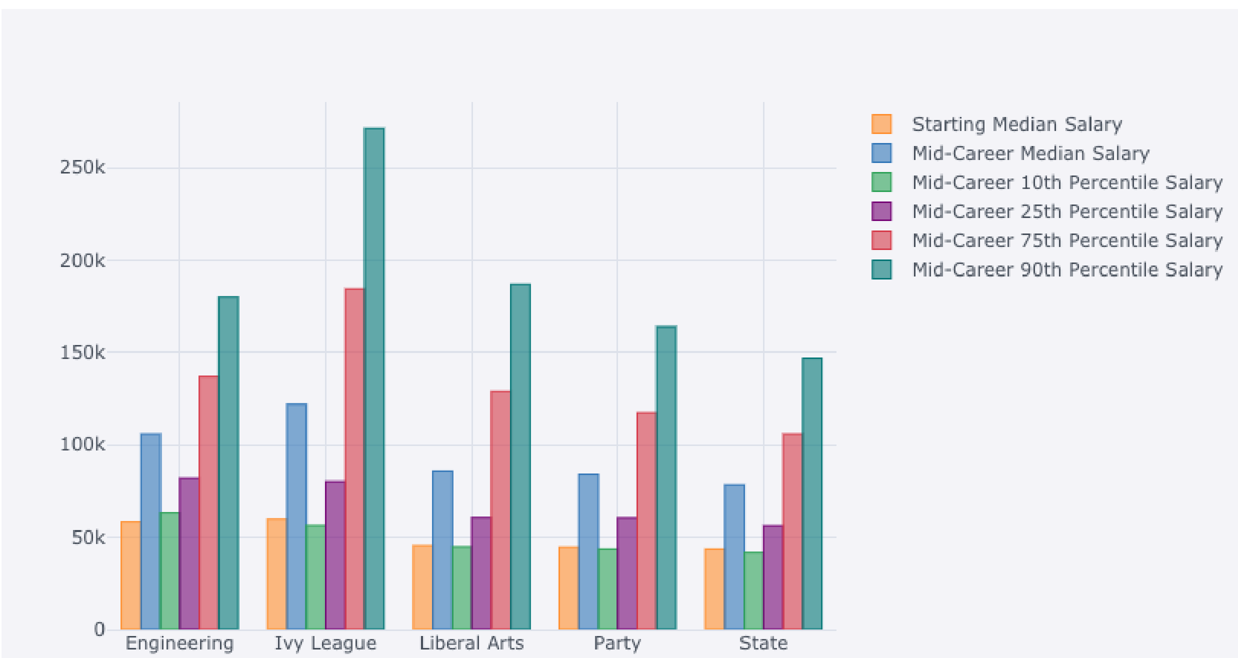
Output:

	School Name	School Type	Starting Median Salary	Mid-Career Median Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
0	Massachusetts Institute of Technology (MIT)	Engineering	72200.0	126000.0	76800.0	99200.0	168000.0	220000.0
1	California Institute of Technology (CIT)	Engineering	75500.0	123000.0	NaN	104000.0	161000.0	NaN
2	Harvey Mudd College	Engineering	71800.0	122000.0	NaN	96000.0	180000.0	NaN
3	Polytechnic University of New York, Brooklyn	Engineering	62400.0	114000.0	66800.0	94300.0	143000.0	190000.0
4	Cooper Union	Engineering	62200.0	114000.0	NaN	80200.0	142000.0	NaN

Now that the data is properly structured we can begin visualizing the relationships between the datasets. This will enable us to make some preliminary conclusions.
First, we plotted the College Type dataset grouping on the School Type column averaging the salaries within and plotting bar charts of the median value for each salary column.
Input:

```
college_type.groupby("School Type").median().iplot(kind="bar")
```

Output:



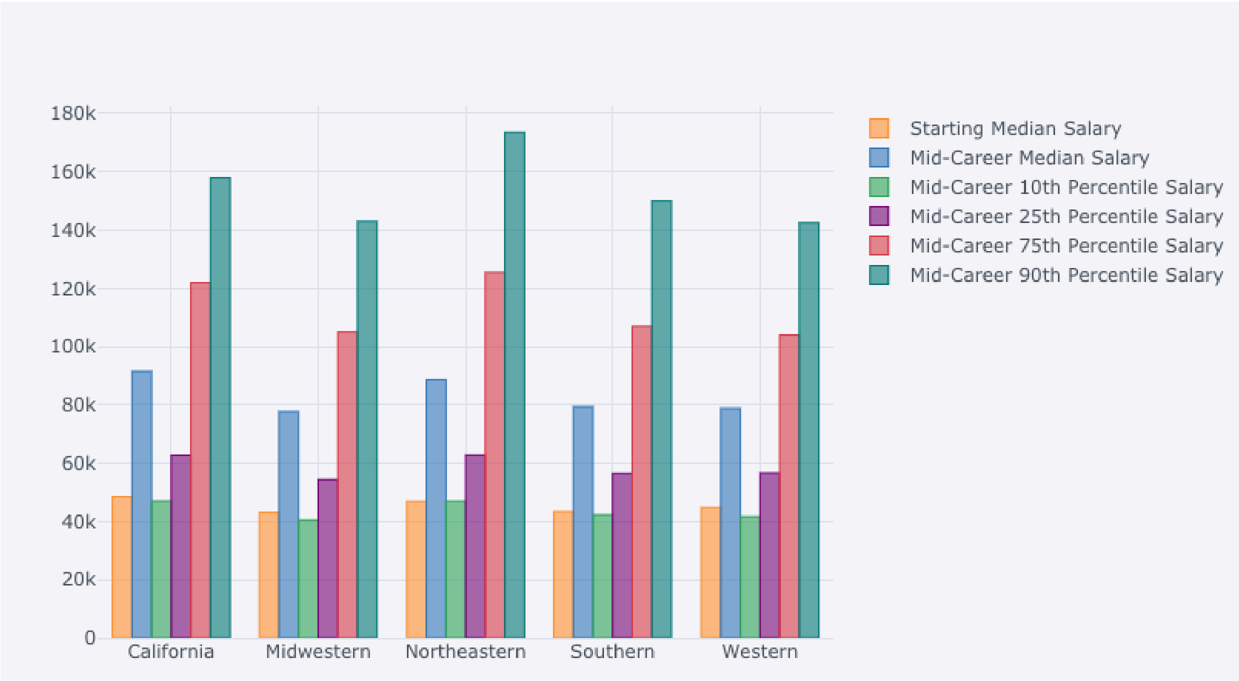
We observe that the Engineering and Ivy League salaries generally exceed Liberal Arts, Party and State.

Our next chart is on the Region dataset grouping on the Region column averaging the salaries within and plotting bar charts of the median value for each salary column.

Input:

```
region.groupby("Region").median().iplot(kind="bar")
```

Output:



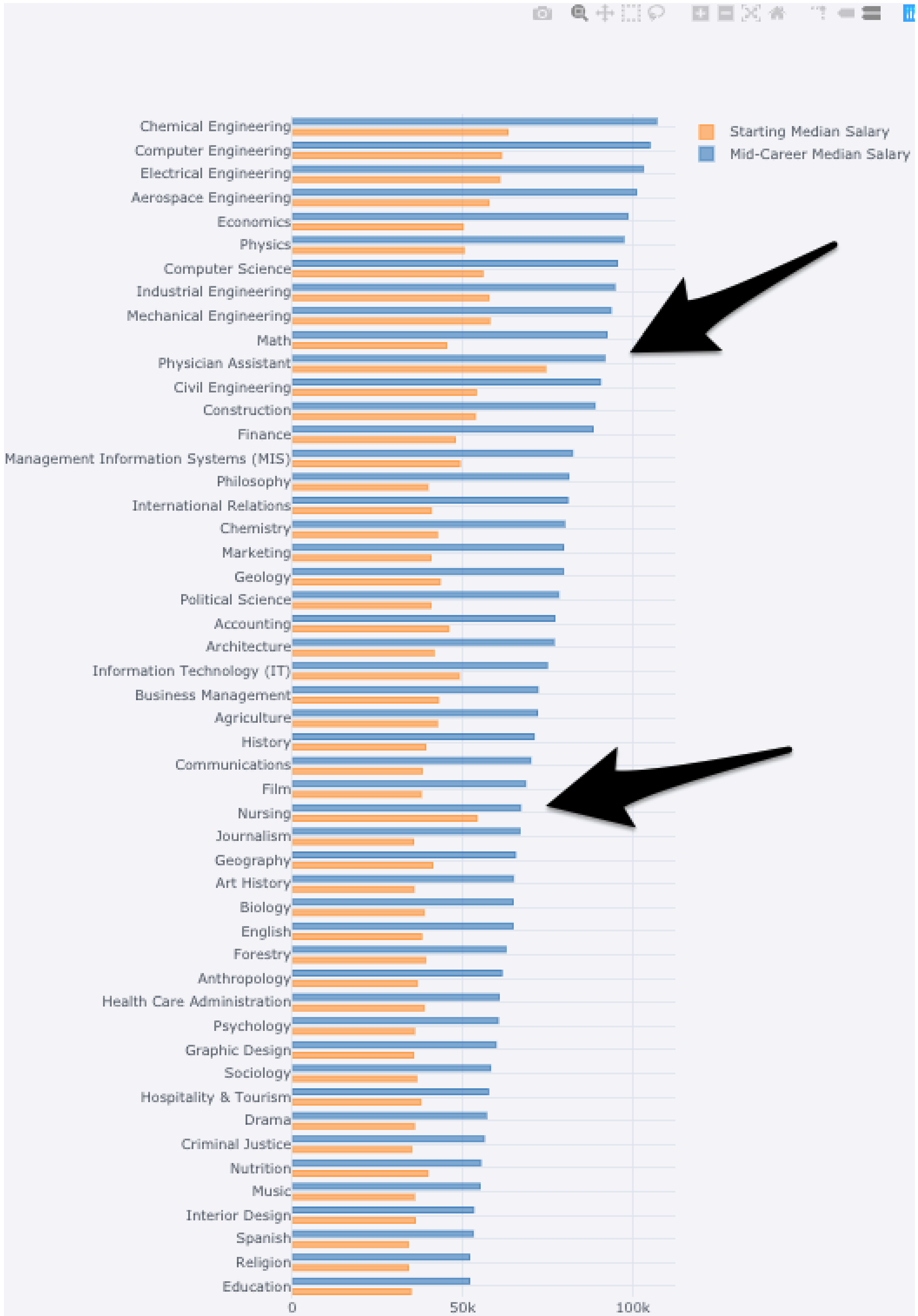
In this chart we see that the Northeastern and California regions' salaries outpace the other three regions.

Our final visualization was of the Degrees dataset. We explored the undergraduate major column and plotted the starting and mid-career median salaries.

Input:

```
columns = ["Undergraduate Major", "Starting Median Salary", "Mid-Career Median Salary"]
degrees[columns].sort_values("Mid-Career Median Salary").set_index("Undergraduate Major").iplot
(
    kind='barh', subplots=False, bargap=.1, bargroupgap=.5,
    dimensions=(800, 1200), margin=dict(l=250, r=20)
)
```

Output:



Notice the differences between the Math, Physician Assistant and Nursing degrees. The Math major virtually doubles the starting salary while the Physician Assistant Nursing degrees have comparatively small income growth.

These are the following observations from the previous three charts:

- Engineering and Ivy League salaries generally exceed Liberal Arts, Party and State.
- Northeastern and California salaries outpace the Midwestern, Western and Southern.
- Wide variation between starting and mid-career salaries in the Degrees database.

Model building and Model evaluation:

Our first model was of the **College Type dataset**. We wanted to see if School Name and School Type plays a role in determining Starting Median Salary. Based on the initial charts it appears that graduating from Ivy League and Engineering Schools will result in higher salaries. We constructed a new Pandas DataFrame using only three columns: School Name, School Type and Starting Median Salary.

Input:

```
college_type_data = pd.DataFrame(dict(
        school_name=college_type['School Name'],
        school_type=college_type['School Type'],
        starting_salary=college_type['Starting Median Salary']))

print(college_type_data.shape)
college_type_data.replace([np.inf, -np.inf], np.nan).dropna().shape
college_type_data.head()
```

```
(269, 3)
```

Output:

	school_name	school_type	starting_salary
0	Massachusetts Institute of Technology (MIT)	Engineering	72200.0
1	California Institute of Technology (CIT)	Engineering	75500.0
2	Harvey Mudd College	Engineering	71800.0
3	Polytechnic University of New York, Brooklyn	Engineering	62400.0
4	Cooper Union	Engineering	62200.0

This model college_type_lm is the least squares regression fit of starting_salary as a function of school_name and school_type. The purpose of the ANOVA test is to identify whether adding terms is valuable. The output of the ANOVA analysis shows the contribution to variance from each of the terms in the regression—both are statistically significant.

```
college_type_lm = ols('starting_salary ~ school_name+school_type', data=college_type_data).fit
() #linear model
table = sm.stats.anova_lm(college_type_lm, typ=2) # Type 2 ANOVA DataFrame

print(table)
```

```
                   sum_sq     df             F         PR(>F)
school_name  2.113485e+10  248.0  3.959293e+26  5.931430e-237
school_type  1.183310e+08    4.0  1.374385e+26  4.324498e-229
Residual     3.874382e-18   18.0           NaN            NaN
```

Both P-values are very small (p = 5.931430e-237; p = 4.3244983-229), so the F statistic quantitatively confirms our observation from data exploration that both School Type and School Name are likely to be significant in the starting salary.

Region Dataset

This model is similar to the previous but we added the MHHI to account for potential cost of living differences across regions.

Input:

```
college_region_data = pd.DataFrame(dict(
        school_name=region['School Name'],
        school_region=region['Region'],
        starting_salary=region['Starting Median Salary'],
        median_hh_income=[float(regional_median_income.loc[r]) for r in region.Region]))

print(college_region_data.shape)
print(college_region_data.replace([np.inf, -np.inf], np.nan).dropna().shape)
college_region_data.head()
```

```
(320, 4)
(320, 4)
```

Output:

	school_name	school_region	starting_salary	median_hh_income
0	Stanford University	California	70400.0	84189.571429
1	California Institute of Technology (CIT)	California	75500.0	84189.571429
2	Harvey Mudd College	California	71800.0	84189.571429
3	University of California, Berkeley	California	59900.0	84189.571429
4	Occidental College	California	51900.0	84189.571429

We repeated the process as previous model but used Starting Salary vs Median HH Income. The purpose of doing this ANOVA is to see the F statistic for comparison against the next model with added school name.

```
college_region_lm = ols('starting_salary ~ median_hh_income', data=college_region_data).fit() # linear model
table = sm.stats.anova_lm(college_region_lm, typ=2) # Type 2 ANOVA DataFrame

print(table)
```

```
                        sum_sq     df          F        PR(>F)
median_hh_income  1.217107e+09    1.0  30.355208  7.437570e-08
Residual          1.275037e+10  318.0        NaN           NaN
```

Now we have two models: college_region_lm and college_region_school. The second model has the addition of the School Name.

```
college_region_lm = ols('starting_salary ~ median_hh_income', data=college_region_data).fit() # linear model
college_region_school = ols('starting_salary ~ median_hh_income+school_name', data=college_region_data).fit()
college_region_school.compare_f_test(college_region_lm)
# (F-Statistic, p-value, increase in degrees of freedom)
```

```
(0.0, nan, 318.0)
```

According to the ANOVA test the School Name does not add any information to the model because the F-statistic is 0.

Made a new DataFrame called college_region_plusfit and added the residual column which is the leftover variance of the original model college_region_lm. If we factor in the residual from the predicted salary based on median household income we wanted to see if the School Region was relevant to predicting starting salary.

Input:

```
college_region_plusfit = college_region_data.assign(resid=college_region_lm.resid)
college_region_plusfit.head()
```

Output:

	school_name	school_region	starting_salary	median_hh_income	resid
0	Stanford University	California	70400.0	84189.571429	21766.892586
1	California Institute of Technology (CIT)	California	75500.0	84189.571429	26866.892586
2	Harvey Mudd College	California	71800.0	84189.571429	23166.892586
3	University of California, Berkeley	California	59900.0	84189.571429	11266.892586
4	Occidental College	California	51900.0	84189.571429	3266.892586

We constructed a new model least squares and then ran the ANOVA again. We wanted to compare college_region_school against college_region.

```
college_region_plusfit_lm = ols('starting_salary ~ resid+school_region', data=college_region_pl
usfit).fit() #linear model
table = sm.stats.anova_lm(college_region_plusfit_lm, typ=2) # Type 2 ANOVA DataFrame

print(table)
#college_region_plusfit_lm.summary()
```

```
                     sum_sq     df             F  PR(>F)
school_region  1.217107e+09    4.0  3.391627e+23     0.0
resid          1.215410e+10    1.0  1.354759e+25     0.0
Residual       2.817023e-13  314.0           NaN     NaN
```

The F-Statistic is very large which we could interpret that school region plays a role in starting salary.

Input:

```
college_region_plusfit_lm.summary()
```

Output:

OLS Regression Results

Dep. Variable:	starting_salary	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	3.114e+24
Date:	Sat, 15 Dec 2018	Prob (F-statistic):	0.00
Time:	18:34:22	Log-Likelihood:	5092.5
No. Observations:	320	AIC:	-1.017e+04
Df Residuals:	314	BIC:	-1.015e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.863e+04	5.7e-09	8.54e+12	0.000	4.86e+04	4.86e+04
school_region[T.Midwestern]	-5496.8244	6.69e-09	-8.21e+11	0.000	-5496.824	-5496.824
school_region[T.Northeastern]	-567.9341	6.43e-09	-8.84e+10	0.000	-567.934	-567.934
school_region[T.Southern]	-2988.6334	6.66e-09	-4.49e+11	0.000	-2988.633	-2988.633
school_region[T.Western]	-1864.8663	7.42e-09	-2.51e+11	0.000	-1864.866	-1864.866
resid	1.0000	2.72e-13	3.68e+12	0.000	1.000	1.000

Omnibus:	70.592	Durbin-Watson:	0.543
Prob(Omnibus):	0.000	Jarque-Bera (JB):	131.539
Skew:	-1.186	Prob(JB):	2.73e-29
Kurtosis:	5.059	Cond. No.	4.98e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.98e+04. This might indicate that there are strong multicollinearity or other numerical problems.

The residuals was all of the variance that was not due to median household income across the region in the previous model. It makes sense that the R-squared is 1. All the coefficients from the different regions are negative which is confusing to interpret.
We continued to build additional models to work on a deeper understanding of these datasets.

Is salary different across the regions? We constructed a new model to attempt to answer this question.

Input:

```
college_region_anova_lm = ols('starting_salary ~ school_region', data=college_region_plusfit).f
it() #linear model
#Is salary different across the regions?
table = sm.stats.anova_lm(college_region_anova_lm, typ=2) # Type 2 ANOVA DataFrame

print(table)
college_region_anova_lm.summary()
```

```
                     sum_sq     df          F        PR(>F)
school_region  1.813378e+09    4.0  11.749409  6.593949e-09
Residual       1.215410e+10  315.0        NaN           NaN
```

Output:

OLS Regression Results

Dep. Variable:	starting_salary	R-squared:	0.130
Model:	OLS	Adj. R-squared:	0.119
Method:	Least Squares	F-statistic:	11.75
Date:	Sat, 15 Dec 2018	Prob (F-statistic):	6.59e-09
Time:	18:34:22	Log-Likelihood:	-3246.5
No. Observations:	320	AIC:	6503.
Df Residuals:	315	BIC:	6522.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.103e+04	1173.889	43.473	0.000	4.87e+04	5.33e+04
school_region[T.Midwestern]	-6806.7907	1386.167	-4.911	0.000	-9534.107	-4079.475
school_region[T.Northeastern]	-2536.1429	1328.104	-1.910	0.057	-5149.219	76.933
school_region[T.Southern]	-6510.6239	1366.172	-4.766	0.000	-9198.600	-3822.648
school_region[T.Western]	-6617.8571	1515.484	-4.367	0.000	-9599.608	-3636.106

Omnibus:	58.931	Durbin-Watson:	0.583
Prob(Omnibus):	0.000	Jarque-Bera (JB):	96.531
Skew:	1.067	Prob(JB):	1.09e-21
Kurtosis:	4.639	Cond. No.	8.68

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Interpretation:
Is salary different across the regions?

We can add the coefficient for each region to the Intercept value, which represents California, to see the predicted starting salary by region. The P-values for all coefficients are small, implying they are all statistically significant.

Prediction:
California: 51030 USD
Midwestern: 51030 - 6807 = 44223 USD
Northeastern: 51030 - 2536 = 48494 USD
Southern: 51030 - 6510 = 44520 USD
Western: 51030 - 6617 = 44413 USD

We still haven't really answered the question of how much starting salary is related to regional variations in standard of living (which we are quantifying using MHHI).
Our next step was to visualize the Region and MHHI data sets.
Input:

```
college_region_plusfit.head()
```

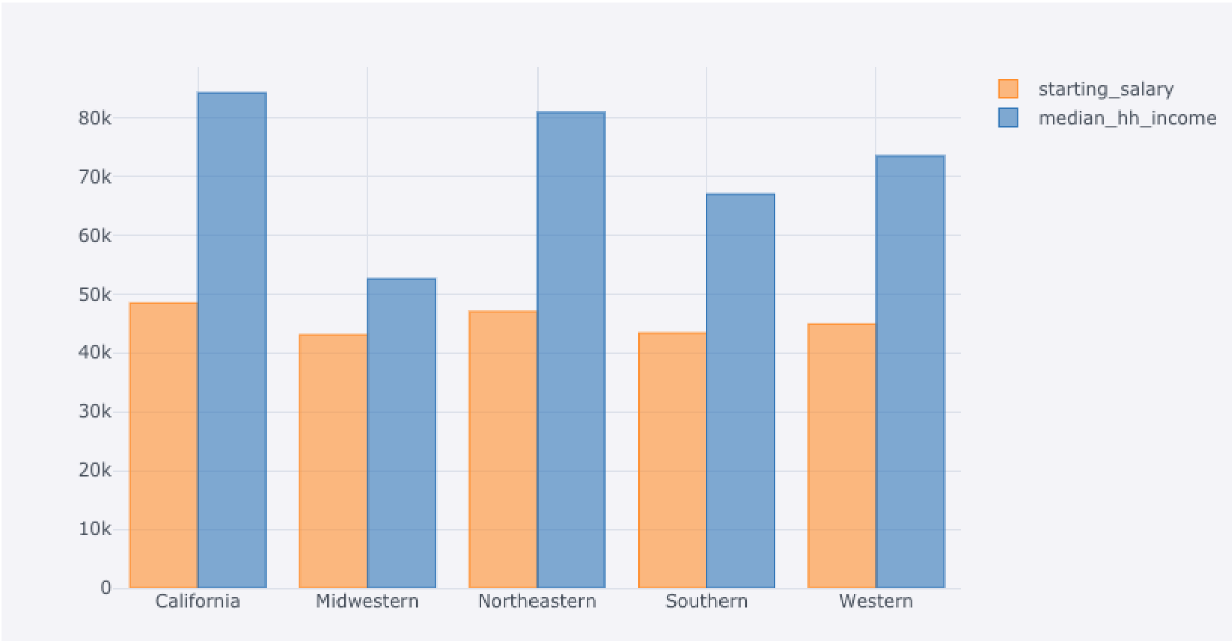
Output:

	school_name	school_region	starting_salary	median_hh_income	resid
0	Stanford University	California	70400.0	84189.571429	21766.892586
1	California Institute of Technology (CIT)	California	75500.0	84189.571429	26866.892586
2	Harvey Mudd College	California	71800.0	84189.571429	23166.892586
3	University of California, Berkeley	California	59900.0	84189.571429	11266.892586
4	Occidental College	California	51900.0	84189.571429	3266.892586

Input:

```
college_region_plusfit[['school_region','starting_salary','median_hh_income']].groupby("school_
region").median().iplot(kind="bar")
```

Output:



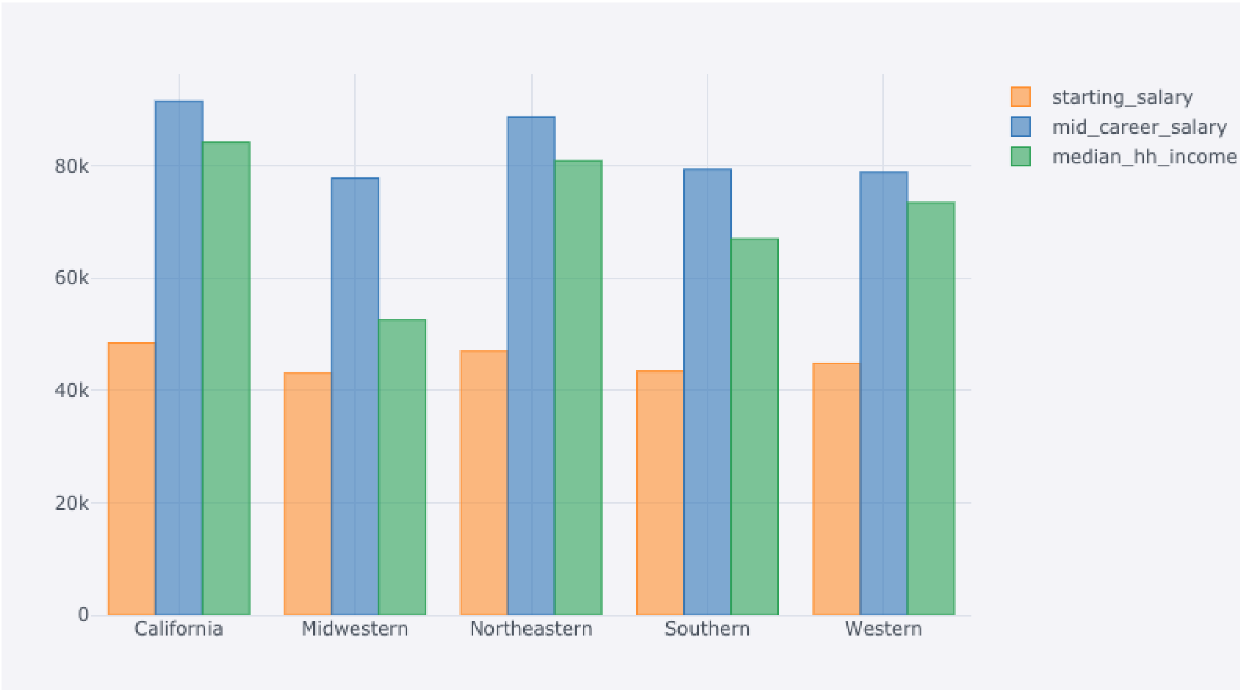
California has a higher starting salary and higher median household income. However, in the Midwestern region a college degree makes a bigger impact on your starting salary vs median household income.

We constructed an additional bar chart:

Input:

```
college_salary_data = pd.DataFrame(dict(
        school_name=region['School Name'],
        school_region=region['Region'],
        starting_salary=region['Starting Median Salary'],
        mid_career_salary=region['Mid-Career Median Salary'],
        median_hh_income=[float(regional_median_income.loc[r]) for r in region.Region]))
college_salary_data.groupby("school_region").median().iplot(kind="bar")
```

Output:



In the Midwestern region the college degree impacts your salary earnings which are substantially higher than the median hh income. Conversely, the California and Northeastern regions have little difference between mid career salary and median hh income.

We constructed a new Pandas DataFrame consisting of the college salary data, binning by school region and doing the median value on all starting salary and MHHI data. Comparing the binned data of the starting salary and the median hh income in each region, using the Kolmogorov-Smirnov test, will quantify our intuition from the bar chart: does college region affect starting salary, or is the variation all due to standard of living?

```
binned_college_data = college_salary_data.groupby("school_region").median() #bin data by school
 region
scipy.stats.ks_2samp(binned_college_data.starting_salary,binned_college_data.median_hh_income)
#KS test
```

```
Ks_2sampResult(statistic=1.0, pvalue=0.0037813540593701006)
```

The KS test had a small P-value ($p = 0.00378$) but since the statistic just compares maximum distance between values, maybe the big number is only because of the big difference in magnitude between the starting salary and the median household income. This difference in magnitude is expected because household income is typically two or more people.

This is another statistical test of our datasets. We constructed a Pandas DataFrame with starting salary, median HH income, and mid career salary. We then normalized the binned data

by their respective category medians by division. The median was chosen because we are using median values across the data sets.
Input:

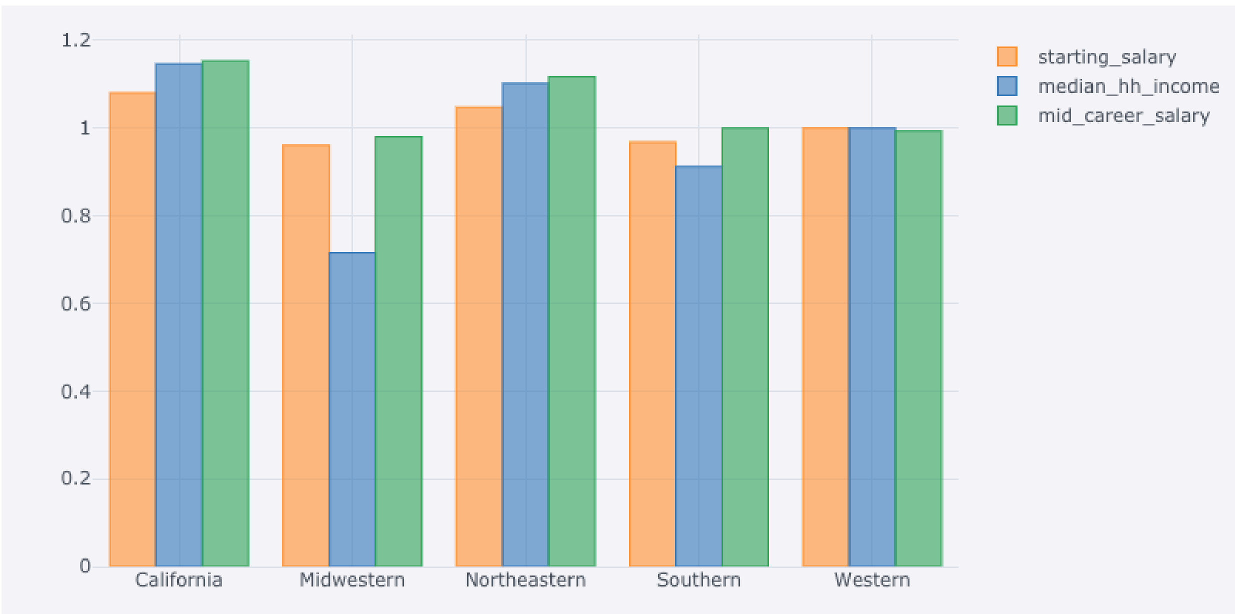
```
normalized_college_salary_data = pd.DataFrame(dict(
    starting_salary=binned_college_data.starting_salary/binned_college_data.starting_salary.med
ian(),
    median_hh_income= binned_college_data.median_hh_income/binned_college_data.median_hh_income
.median(),
    mid_career_salary=binned_college_data.mid_career_salary/binned_college_data.mid_career_sala
ry.median(),
)

)
print("Starting Salary Compared Against Median HH Income")

print(scipy.stats.ks_2samp(
   normalized_college_salary_data.starting_salary,
   normalized_college_salary_data.median_hh_income
)) #KS test



print("Starting Salary Compared Against Mid-Career Salary")

print(scipy.stats.ks_2samp(
   normalized_college_salary_data.starting_salary,
   normalized_college_salary_data.mid_career_salary
)) #KS test

print("Mid-Career Salary Compared Against Median HH Income")
print(scipy.stats.ks_2samp(
   normalized_college_salary_data.mid_career_salary,
   normalized_college_salary_data.median_hh_income
)) #KS test

#normalized_college_salary_data.groupby("school_region").median().iplot(kind="bar")
normalized_college_salary_data.iplot(kind="bar")
```

These are the results of the Kolmogorov-Smirnov tests on the normalized data.
Output:

```
Starting Salary Compared Against Median HH Income
Ks_2sampResult(statistic=0.4, pvalue=0.6974048780205908)
Starting Salary Compared Against Mid-Career Salary
Ks_2sampResult(statistic=0.4, pvalue=0.6974048780205908)
Mid-Career Salary Compared Against Median HH Income
Ks_2sampResult(statistic=0.4, pvalue=0.6974048780205908)
```

Observation:
Did our visualizations mislead us? The P-values are high in each test.
Cannot reject the null hypothesis since the K-S statistic is small and the p-value is high: the distributions of the two samples are the same—meaning regional differences in starting salary appear to be no different than the regional differences in standard of living.

This is the chart of the normalized data that was used the in Kolmogorov-Smirnov tests:



This is the table of the binned_college_data set.
<u>Input:</u>

```
binned_college_data.head()
```

<u>Output:</u>

	starting_salary	mid_career_salary	median_hh_income
school_region			
California	48450.0	91550.0	84189.571429
Midwestern	43100.0	77800.0	52600.250000
Northeastern	47000.0	88700.0	80925.750000
Southern	43400.0	79400.0	67014.400000
Western	44850.0	78850.0	73472.500000

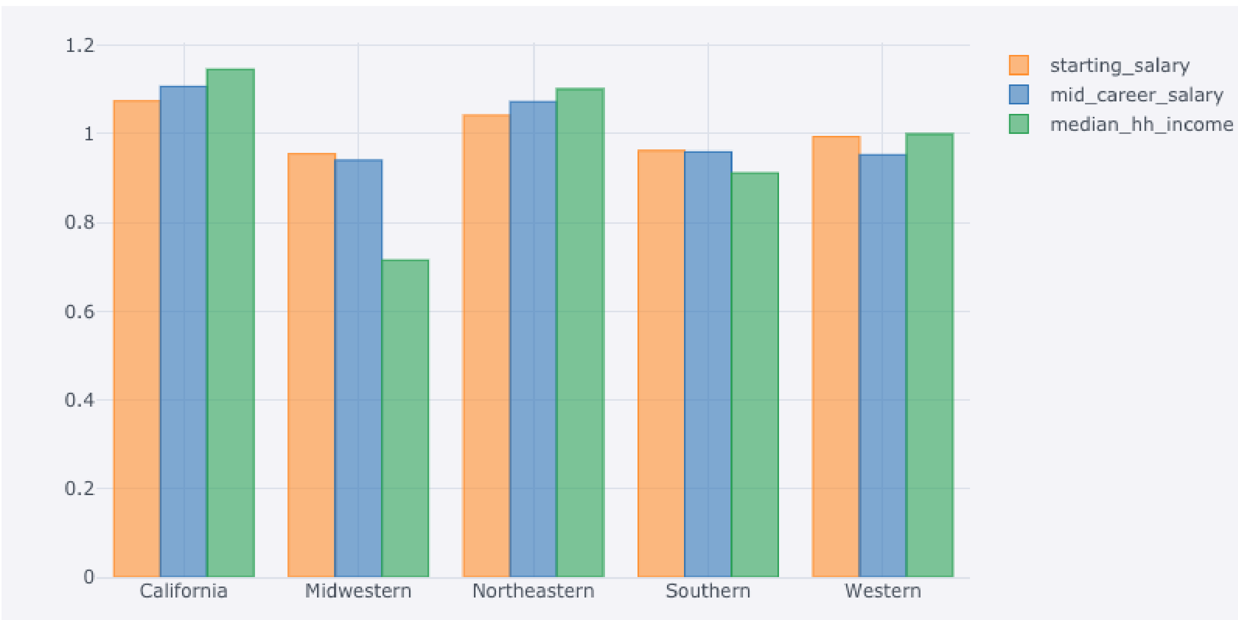
We were unsatisfied with this result because it appeared from the charts that there was a significant difference in the starting salaries by region.
We constructed another Pandas DataFrame that starts from the raw dataset and then normalizes on the median of that dataset rather than normalizing on the median value of the binned data in the previous test.
Input:

```
normalized_college_salary_data = pd.DataFrame(dict(
        school_name=region['School Name'],
        school_region=region['Region'],
        starting_salary=region['Starting Median Salary'],
        mid_career_salary=region['Mid-Career Median Salary'],
        median_hh_income=[float(regional_median_income.loc[r]) for r in region.Region]))

normalized_college_salary_data = normalized_college_salary_data.assign(
    starting_salary=normalized_college_salary_data.starting_salary/normalized_college_salary_da
ta.starting_salary.median(),
    mid_career_salary=normalized_college_salary_data.mid_career_salary/normalized_college_salar
y_data.mid_career_salary.median(),
    median_hh_income=normalized_college_salary_data.median_hh_income/normalized_college_salary_
data.median_hh_income.median()
)

print("Starting Salary Compared Against Median HH Income")

print(scipy.stats.ks_2samp(
   normalized_college_salary_data.starting_salary,
   normalized_college_salary_data.median_hh_income
)) #KS test



print("Starting Salary Compared Against Mid-Career Salary")

print(scipy.stats.ks_2samp(
   normalized_college_salary_data.starting_salary,
   normalized_college_salary_data.mid_career_salary
)) #KS test

print("Mid-Career Salary Compared Against Median HH Income")
print(scipy.stats.ks_2samp(
   normalized_college_salary_data.mid_career_salary,
   normalized_college_salary_data.median_hh_income
)) #KS test
normalized_college_salary_data.groupby("school_region").median().iplot(kind="bar")
```

Output:

```
Starting Salary Compared Against Median HH Income
Ks_2sampResult(statistic=0.284375, pvalue=6.799024513301678e-12)
Starting Salary Compared Against Mid-Career Salary
Ks_2sampResult(statistic=0.14062499999999997, pvalue=0.0031372715374436556)
Mid-Career Salary Compared Against Median HH Income
Ks_2sampResult(statistic=0.21875, pvalue=3.272547649045737e-07)
```

The results of this test show very small P-values for all three tests comparing each category against the other two categories.
This chart is of the data used in this K-S test, normalized on the raw dataset instead of the binned dataset.



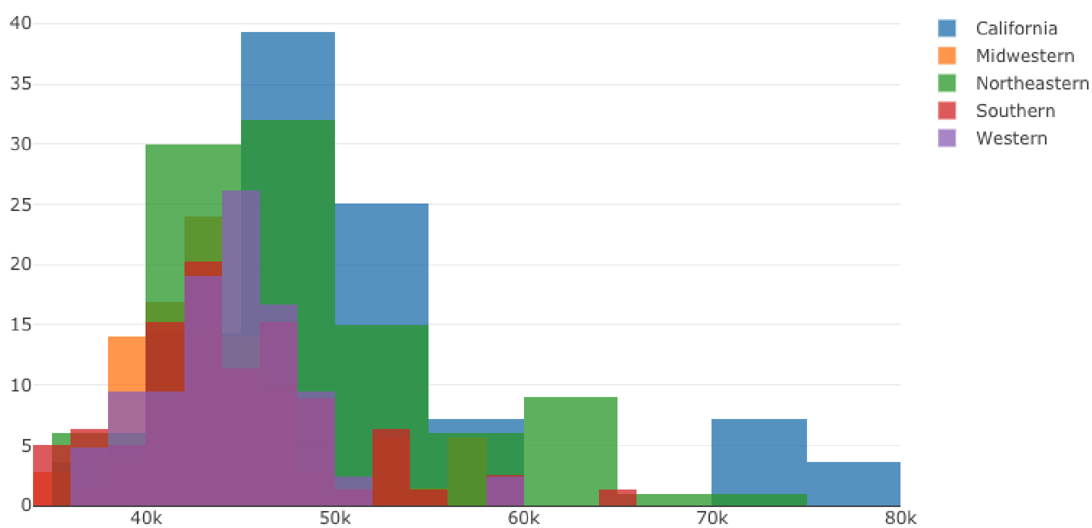
Observations:

The Midwestern region contributes the values that made the Kolmogorov-Smirnov test statistically significant. Is it due to more blue collar workers in the Midwest or because Midwesterners are leaving and going to other regions to earn higher salaries?

We completed our research with four final histograms that dramatically illustrate the variations within the datasets.

Overlaid Histogram: Starting Salaries by Region

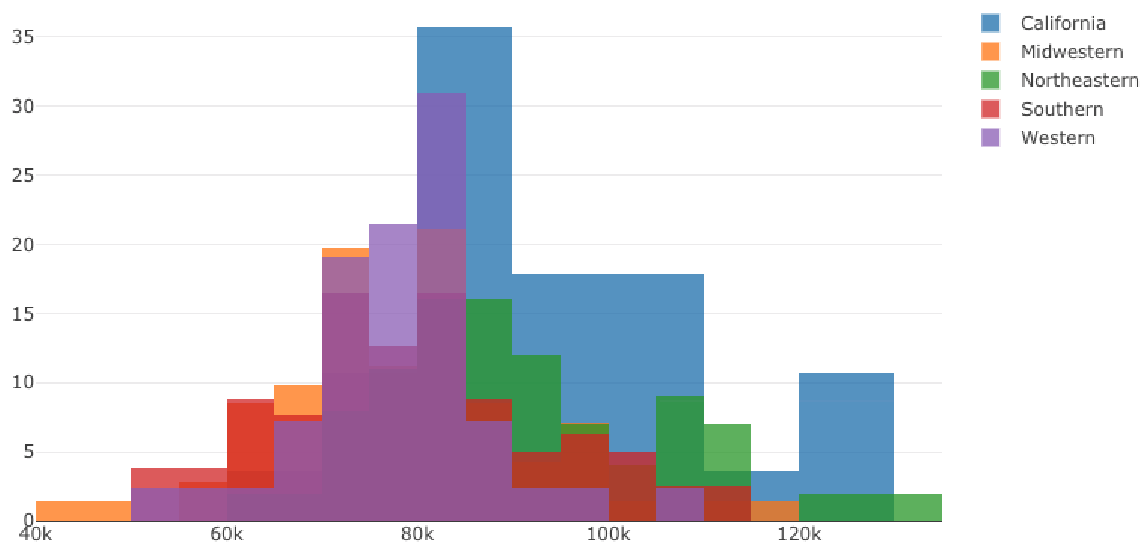
```
data = [
    go.Histogram(
        x=g,
        name=region,
        opacity=0.75,
        histnorm='percent',
    )
    for region,g in college_salary_data.groupby("school_region").starting_salary

]

layout = go.Layout(barmode='overlay')
fig = go.Figure(data=data, layout=layout)

iplot(fig)# filename='overlaid histogram')
#iplot([go.Histogram(x=data)])
```



California does not have a normal distribution; it is right-skewed. Its range is lower ($45 - 80,000) than the Northeastern region. Interestingly, California has a 10% island of values in the $70-80,000 range. The Northeastern region does not have the same density in the higher range. It clusters heavily in the $40-55,000 range. The Midwestern starting-salaries are right-skewed with the greatest density in the $42-43,900 salary range. The Southern region has the lowest starting salaries. The Western region has a tight clustering of salaries in the $38-49,900 range.

Overlaid Histogram: Mid Career Salaries by Region

```
data = [
    go.Histogram(
        x=g,
        name=region,
        opacity=0.75,
        histnorm='percent',
    )
    for region,g in college_salary_data.groupby("school_region").mid_career_salary

]

layout = go.Layout(barmode='overlay')
fig = go.Figure(data=data, layout=layout)

iplot(fig)# filename='overlaid histogram')
#iplot([go.Histogram(x=data)])
```

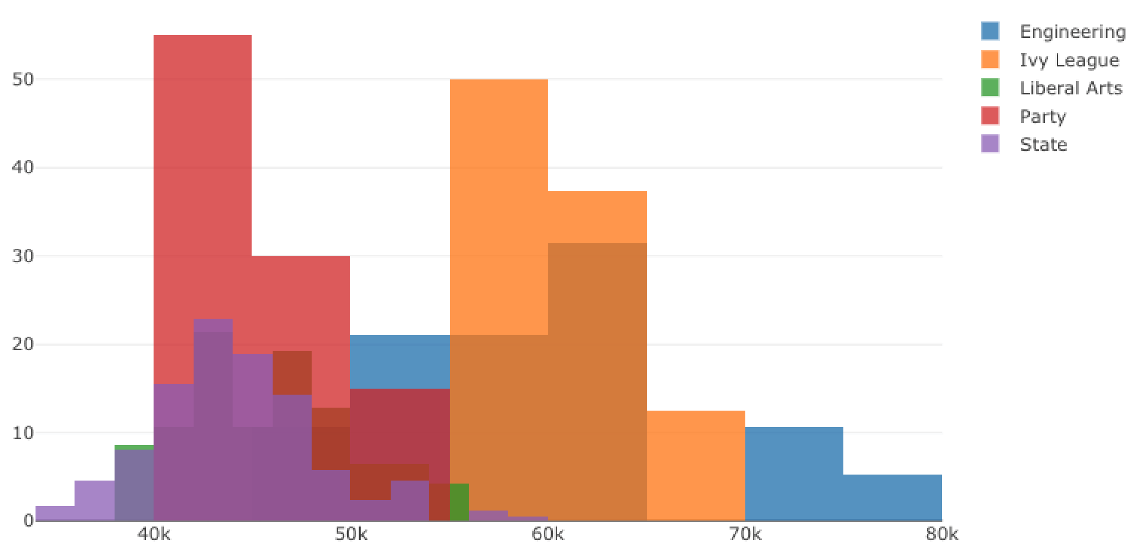


California does not have a normal distribution. The largest area is in the $90 - 110,000 salary range. The distribution for the Northeastern region is right-skewed but not as drastically as California's. It's salary range is more diffused than California's. The remaining regions have more normal distributions but their salary ranges are significantly less than the California and Northeastern regions. Their largest area of convergence is the $70-90,000 salary range. The

Western region has 70% of it’s distribution within the $70-85,000 salary range. The Southern is bimodal: $70-74,9000 and $80-84,900.

Overlaid Histogram: Starting Salary by College Type

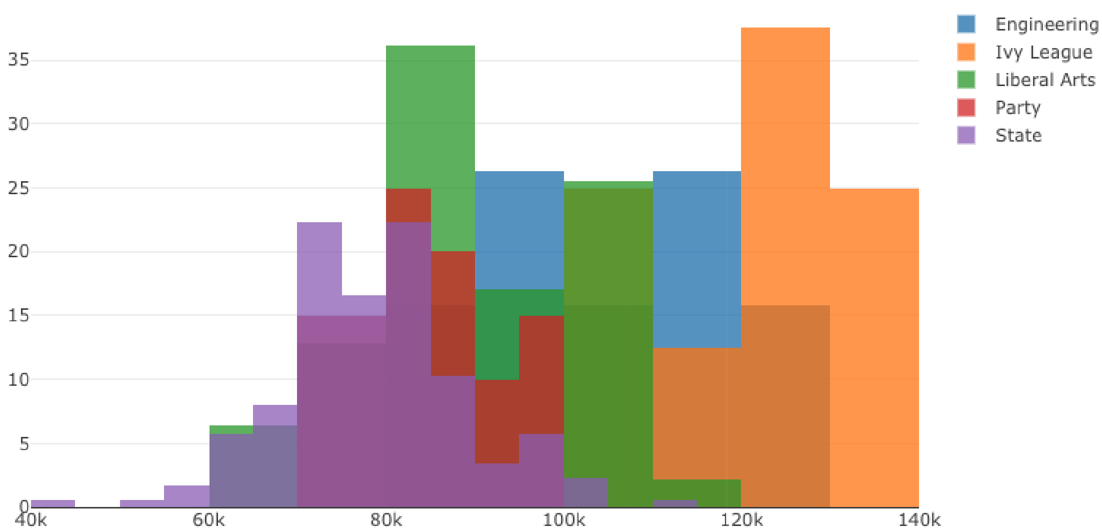
```
data = [
    go.Histogram(
        x=g,
        name=school_type,
        opacity=0.75,
        histnorm='percent',
    )
    for school_type,g in college_type.groupby("School Type")['Starting Median Salary']

]

layout = go.Layout(barmode='overlay')
fig = go.Figure(data=data, layout=layout)

iplot(fig)
```



Engineering colleges have the highest range of starting salaries: $50-80,000. The Ivy League has a very strong presence in the $55-65,000 salary range, 87.5%. The Party college graduates have 55% starting salaries in the $40-44,900 range with an upper range of $50-54,900. Surprisingly, the State and Liberal Arts colleges have a nearly identical distribution.

Overlaid Histogram: Mid-Career Salary by College Type

```
data = [
    go.Histogram(
        x=g,
        name=school_type,
        opacity=0.75,
        histnorm='percent',
    )
    for school_type,g in college_type.groupby("School Type")['Mid-Career Median Salary']

]

layout = go.Layout(barmode='overlay')
fig = go.Figure(data=data, layout=layout)

iplot(fig)
```



The Ivy League is heavily left skewed with 37.5% of the distribution in the $120-129,000 range. It is decisively the college with the greatest compensation by mid-career. The Engineering distribution is bimodal with a range of $80-129,000. The two modes are: $95-99,900 and $110-119,000. The Liberal Arts college has 36% of its value in the $80-89,900 range and it has a higher salary range than Party or State. The Party colleges outperform the State by mid-career with the bulk of State's salary in the $70-99,900 range whereas State is in the $60-89,900 range.

Conclusion:

We are not sure if the Kolmogorov-Smirnov test was the best choice for comparing the Starting Salary, Mid-Career Salary and Median Household Income values by Region. Do we need to renormalize the data?

The data strongly suggests that your college choice does make a difference in your starting and mid-career salaries. Ivy League and Engineering colleges definitively result in higher salaries over time. There are also substantial regional differences in salaries. What we aren't certain of is whether the higher salaries in the California and Northeastern regions truly result in a higher living standard or if those higher salary are mitigated by their regions' higher cost of living. Quite possibly, the ideal solution for a college educated individual would be to graduate from an Ivy League or Engineering college and live in a region with a lower cost of living.

References:

Quandl 3.4.5
Pandas 0.22.0
NumPy 1.14.6
Matplotlib 2.1.2
Statsmodels 0.8.0
Seaborn 0.7.1
Python 3.6.7
SciPy 1.2.0
Cufflinks 0.8.2.
Plotly 3.4.2

https://plot.ly/python/anova/

https://www.investopedia.com/exam-guide/cfa-level-1/quantitative-methods/hypothesis-testing.asp

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ks_2samp.html

https://www.kaggle.com/wsj/college-salaries

https://www.kaggle.com/census/estimate-of-median-household-income-group-series

http://www.act.org/content/act/en/products-and-services/act-profile.html