Analysis of Career Salaries by Major, College Type and Region

Introduction

The Wall Street Journal collected data about a student's major, type of college attended and region of college and their career earnings at starting and mid career of their careers. They also went ahead and collected data about the salary collected at various percentiles during mid career.

In this project, I conducted an EDA to determine how each of the above mentioned factors contribute to a student's prospective career earning. College applicants consider salary as an important factor before deciding which college path to follow.

Objective

Identify how undergraduate major, college type and college region impacts an individual's career earning potential and help individuals make an informed decision.

Collecting the Data

The data required for this project has been provided from a Kaggle notebook. The data is collected and owned by the Wall Street Journal, and since it is a reputable organization, the data can be considered reliable and able to assist us in this project. The data only contains information about colleges and universities located within the United States of America making it less ideal for college applicants who plan to pursue careers in other countries. Data also does not contain college fees required for each course leaving us unable to determine which course, college and region offer the highest value for money.

Processing the Data

For processing, analyzing and visualizing the data we need to import some key Python libraries

In [1]:
```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
plt.style.use('fivethirtyeight')
```

First we need to import the data

In [2]:
```
degrees_df = pd.read_csv("/Users/arun/Documents/Portfolio Projects/College Data/degrees.csv")
```

In [3]:
```
degrees_df
```

Out[3]:

	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
3	Anthropology	$36,800.00	$61,500.00	67.1	$33,800.00	$45,500.00	$89,300.00	$138,000.00
4	Architecture	$41,600.00	$76,800.00	84.6	$50,600.00	$62,200.00	$97,000.00	$136,000.00
5	Art History	$35,800.00	$64,900.00	81.3	$28,800.00	$42,200.00	$87,400.00	$125,000.00
6	Biology	$38,800.00	$64,800.00	67.0	$36,900.00	$47,400.00	$94,500.00	$135,000.00
7	Business Management	$43,000.00	$72,100.00	67.7	$38,800.00	$51,500.00	$102,000.00	$147,000.00
8	Chemical Engineering	$63,200.00	$107,000.00	69.3	$71,900.00	$87,300.00	$143,000.00	$194,000.00
9	Chemistry	$42,600.00	$79,900.00	87.6	$45,300.00	$60,700.00	$108,000.00	$148,000.00
10	Civil Engineering	$53,900.00	$90,500.00	67.9	$63,400.00	$75,100.00	$115,000.00	$148,000.00
11	Communications	$38,100.00	$70,000.00	83.7	$37,500.00	$49,700.00	$98,800.00	$143,000.00
12	Computer Engineering	$61,400.00	$105,000.00	71.0	$66,100.00	$84,100.00	$135,000.00	$162,000.00
13	Computer Science	$55,900.00	$95,500.00	70.8	$56,000.00	$74,900.00	$122,000.00	$154,000.00
14	Construction	$53,700.00	$88,900.00	65.5	$56,300.00	$68,100.00	$118,000.00	$171,000.00
15	Criminal Justice	$35,000.00	$56,300.00	60.9	$32,200.00	$41,600.00	$80,700.00	$107,000.00
16	Drama	$35,900.00	$56,900.00	58.5	$36,700.00	$41,300.00	$79,100.00	$153,000.00
17	Economics	$50,100.00	$98,600.00	96.8	$50,600.00	$70,600.00	$145,000.00	$210,000.00
18	Education	$34,900.00	$52,000.00	49.0	$29,300.00	$37,900.00	$73,400.00	$102,000.00
19	Electrical Engineering	$60,900.00	$103,000.00	69.1	$69,300.00	$83,800.00	$130,000.00	$168,000.00
20	English	$38,000.00	$64,700.00	70.3	$33,400.00	$44,800.00	$93,200.00	$133,000.00
21	Film	$37,900.00	$68,500.00	80.7	$33,900.00	$45,500.00	$100,000.00	$136,000.00
22	Finance	$47,900.00	$88,300.00	84.3	$47,200.00	$62,100.00	$128,000.00	$195,000.00
23	Forestry	$39,100.00	$62,600.00	60.1	$41,000.00	$49,300.00	$78,200.00	$111,000.00
24	Geography	$41,200.00	$65,500.00	59.0	$40,000.00	$50,000.00	$90,800.00	$132,000.00
25	Geology	$43,500.00	$79,500.00	82.8	$45,000.00	$59,600.00	$101,000.00	$156,000.00
26	Graphic Design	$35,700.00	$59,800.00	67.5	$36,000.00	$45,500.00	$80,800.00	$112,000.00
27	Health Care Administration	$38,800.00	$60,600.00	56.2	$34,600.00	$45,600.00	$78,800.00	$101,000.00
28	History	$39,200.00	$71,000.00	81.1	$37,000.00	$49,200.00	$103,000.00	$149,000.00
29	Hospitality & Tourism	$37,800.00	$57,500.00	52.1	$35,500.00	$43,600.00	$81,900.00	$124,000.00
30	Industrial Engineering	$57,700.00	$94,700.00	64.1	$57,100.00	$72,300.00	$132,000.00	$173,000.00
31	Information Technology (IT)	$49,100.00	$74,800.00	52.3	$44,500.00	$56,700.00	$96,700.00	$129,000.00
32	Interior Design	$36,100.00	$53,200.00	47.4	$35,700.00	$42,600.00	$72,500.00	$107,000.00
33	International Relations	$40,900.00	$80,900.00	97.8	$38,200.00	$56,000.00	$111,000.00	$157,000.00
34	Journalism	$35,600.00	$66,700.00	87.4	$38,400.00	$48,300.00	$97,700.00	$145,000.00
35	Management Information Systems (MIS)	$49,200.00	$82,300.00	67.3	$45,300.00	$60,500.00	$108,000.00	$146,000.00
36	Marketing	$40,800.00	$79,600.00	95.1	$42,100.00	$55,600.00	$119,000.00	$175,000.00
37	Math	$45,400.00	$92,400.00	103.5	$45,200.00	$64,200.00	$128,000.00	$183,000.00
38	Mechanical Engineering	$57,900.00	$93,600.00	61.7	$63,700.00	$76,200.00	$120,000.00	$163,000.00
39	Music	$35,900.00	$55,000.00	53.2	$26,700.00	$40,200.00	$88,000.00	$134,000.00
40	Nursing	$54,200.00	$67,000.00	23.6	$47,600.00	$56,400.00	$80,900.00	$98,300.00
41	Nutrition	$39,900.00	$55,300.00	38.6	$33,900.00	$44,500.00	$70,500.00	$99,200.00
42	Philosophy	$39,900.00	$81,200.00	103.5	$35,500.00	$52,800.00	$127,000.00	$168,000.00
43	Physician Assistant	$74,300.00	$91,700.00	23.4	$66,400.00	$75,200.00	$108,000.00	$124,000.00
44	Physics	$50,300.00	$97,300.00	93.4	$56,000.00	$74,200.00	$132,000.00	$178,000.00
45	Political Science	$40,800.00	$78,200.00	91.7	$41,200.00	$55,300.00	$114,000.00	$168,000.00
46	Psychology	$35,900.00	$60,400.00	68.2	$31,600.00	$42,100.00	$87,500.00	$127,000.00
47	Religion	$34,100.00	$52,000.00	52.5	$29,700.00	$36,500.00	$70,900.00	$96,400.00
48	Sociology	$36,500.00	$58,200.00	59.5	$30,700.00	$40,400.00	$81,200.00	$118,000.00
49	Spanish	$34,000.00	$53,100.00	56.2	$31,000.00	$40,000.00	$76,800.00	$96,400.00

Similarly, we need to import the remaining 2 datasets

In [4]:
```
college_type_df = pd.read_csv("/Users/arun/Documents/Portfolio Projects/College Data/college_type.csv")
```

In [5]:
```
region_df = pd.read_csv("/Users/arun/Documents/Portfolio Projects/College Data/region.csv")
```

In order to perform analysis on the data, the salaries have to be converted into 'float' datatype

In [6]:
```
degrees_df[degrees_df.columns[1:3]] = degrees_df[degrees_df.columns[1:3]].replace('[$,]', '', regex=True).astype(float)
```

In [7]:
```
degrees_df[degrees_df.columns[4:]] = degrees_df[degrees_df.columns[4:]].replace('[$,]', '', regex=True).astype(float)
```

To see if the data type conversion was successful, we check using a preview of the data

In [8]:
```
degrees_df.head()
```

Out[8]:

	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

In [9]:
```
college_type_df[college_type_df.columns[2:]] = college_type_df[college_type_df.columns[2:]].replace('[$,]','', regex=True).astype(float)
```

In [10]:
```
college_type_df.head()
```

Out[10]:

	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

In [11]:
```
region_df[region_df.columns[2:]] = region_df[region_df.columns[2:]].replace('[$,]','',regex=True).astype(float)
```

In [12]:
```
region_df.head()
```

Out[12]:

	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

The College Type dataset and Region dataset both contain some missing values. These values must be removed in order to do analysis. If possible, more research would be done to obtain those values but since that is not feasible, the rows containing missing values will be deleted.

In [13]:
```
college_type_df.dropna(axis=0, how='any',thresh=None, subset=None, inplace=False)
```

Out[13]:

	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
3	Polytechnic University of New York, Brooklyn	Engineering	62400.0	114000.0	66800.0	94300.0	143000.0	190000.0
5	Worcester Polytechnic Institute (WPI)	Engineering	61000.0	114000.0	80000.0	91200.0	137000.0	180000.0
6	Carnegie Mellon University (CMU)	Engineering	61800.0	111000.0	63300.0	80100.0	150000.0	209000.0
7	Rensselaer Polytechnic Institute (RPI)	Engineering	61100.0	110000.0	71600.0	85500.0	140000.0	182000.0
...	...	...	...	...	...	...	...	...
264	Austin Peay State University	State	37700.0	59200.0	32200.0	40500.0	73900.0	96200.0
265	Pittsburg State University	State	40400.0	58200.0	25600.0	46000.0	84600.0	117000.0
266	Southern Utah University	State	41900.0	56500.0	30700.0	39700.0	78400.0	116000.0
267	Montana State University - Billings	State	37900.0	50600.0	22600.0	31800.0	78500.0	98900.0
268	Black Hills State University	State	35300.0	43900.0	27000.0	32200.0	60900.0	87600.0

231 rows × 8 columns

In [14]:
```
region_df.dropna(axis=0, how='any', thresh=None, subset=None, inplace=False)
```

Out[14]:

	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
3	University of California, Berkeley	California	59900.0	112000.0	59500.0	81000.0	149000.0	201000.0
5	Cal Poly San Luis Obispo	California	57200.0	101000.0	55000.0	74700.0	133000.0	178000.0
6	University of California at Los Angeles (UCLA)	California	52600.0	101000.0	51300.0	72500.0	139000.0	193000.0
7	University of California, San Diego (UCSD)	California	51100.0	101000.0	51700.0	75400.0	131000.0	177000.0
...	...	...	...	...	...	...	...	...
315	State University of New York (SUNY) at Potsdam	Northeastern	38000.0	70300.0	35100.0	51200.0	100000.0	179000.0
316	Niagara University	Northeastern	36900.0	69700.0	44000.0	57000.0	92000.0	128000.0
317	State University of New York (SUNY) at Fredonia	Northeastern	37800.0	66200.0	32800.0	44200.0	93300.0	181000.0
318	University of Southern Maine	Northeastern	39400.0	63600.0	40400.0	47900.0	85700.0	117000.0
319	Mercy College	Northeastern	43700.0	62600.0	35600.0	47300.0	99000.0	134000.0

273 rows × 8 columns

Analyzing the Data

Analyzing Salaries by Undergraduate Major

First, we are gonna determine which undergraduate degree offers the the highest starting median salary

In [15]:
```
high_major_start_sal = degrees_df.loc[degrees_df["Starting Median Salary"].idxmax()]
```

In [16]:
```
high_major_start_sal
```

Out[16]:
```
Undergraduate Major                                  Physician Assistant
Starting Median Salary                                           74300.0
Mid-Career Median Salary                                         91700.0
Percent change from Starting to Mid-Career Salary                   23.4
Mid-Career 10th Percentile Salary                                66400.0
Mid-Career 25th Percentile Salary                                75200.0
Mid-Career 75th Percentile Salary                               108000.0
Mid-Career 90th Percentile Salary                               124000.0
Name: 43, dtype: object
```

In [17]:
```
degrees_sorted_start = degrees_df.sort_values('Starting Median Salary', ascending=False).head(10)
```

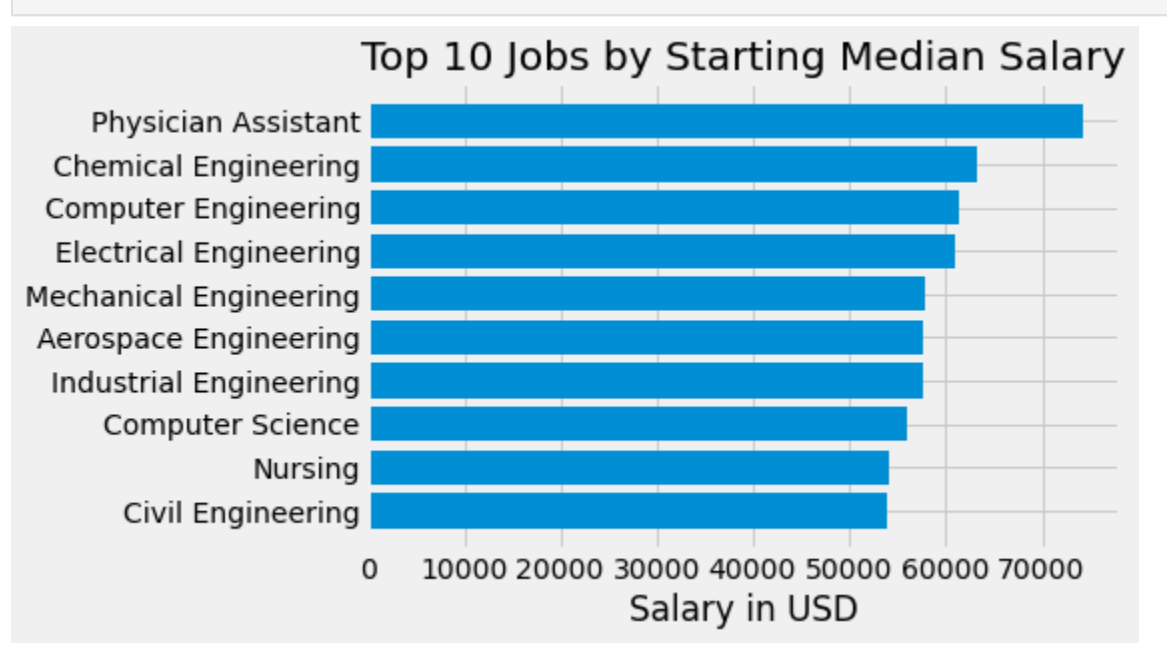
In [18]:
```
start_sal = degrees_sorted_start['Starting Median Salary'].head(10)
```

In [19]:
```
start_sal = start_sal.iloc[::-1]
```

In [20]:
```
undergrad_major_start = degrees_sorted_start['Undergraduate Major'].head(10)
```

In [21]:
```
undergrad_major_start = undergrad_major_start.iloc[::-1]
```

In [22]:
```
plt.barh(undergrad_major_start, start_sal)
plt.title('Top 10 Jobs by Starting Median Salary')
plt.xlabel('Salary in USD')
plt.show()
```



Next, we will find the highest paying career at mid-career

In [23]:
```
high_major_mid_sal = degrees_df.loc[degrees_df["Mid-Career Median Salary"].idxmax()]
```

In [24]:
```
high_major_mid_sal
```

Out[24]:
```
Undergraduate Major                                  Chemical Engineering
Starting Median Salary                                            63200.0
Mid-Career Median Salary                                         107000.0
Percent change from Starting to Mid-Career Salary                    69.3
Mid-Career 10th Percentile Salary                                 71900.0
Mid-Career 25th Percentile Salary                                 87300.0
Mid-Career 75th Percentile Salary                                143000.0
Mid-Career 90th Percentile Salary                                194000.0
Name: 8, dtype: object
```

In [25]:
```
degrees_sorted_mid = degrees_df.sort_values('Mid-Career Median Salary', ascending=False)
```

In [26]:
```
mid_sal = degrees_sorted_mid['Mid-Career Median Salary'].head(10)
```

In [27]:
```
mid_sal
```

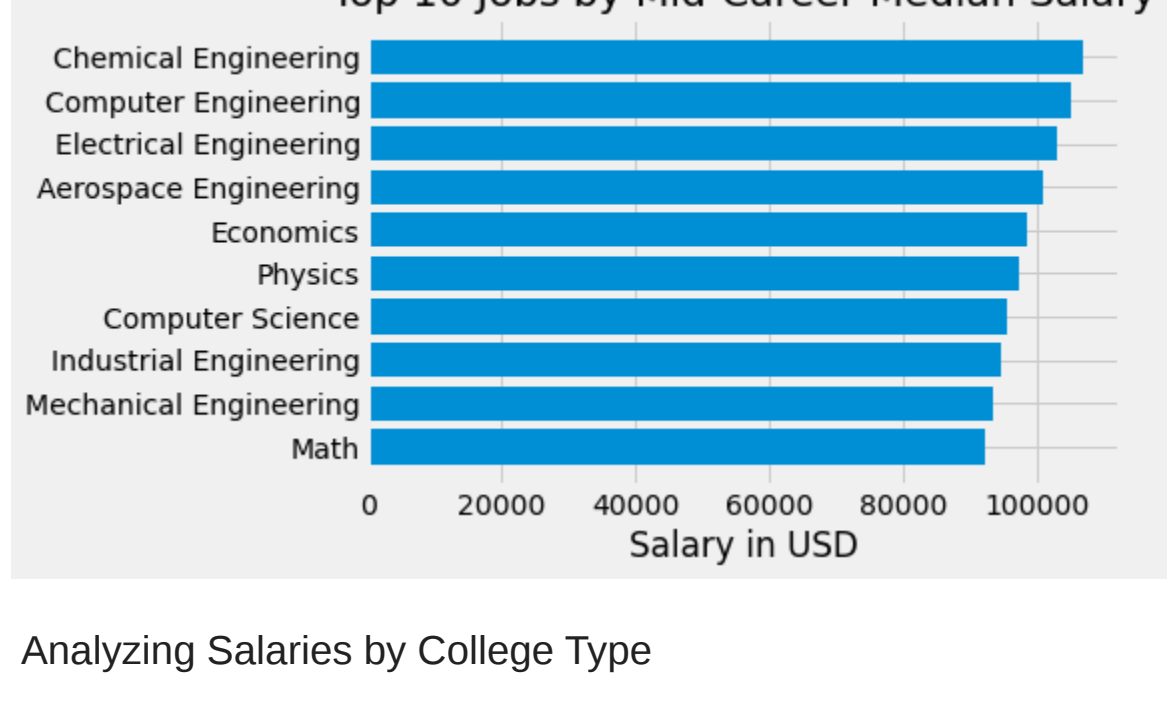
Out[27]:
```
8     107000.0
12    105000.0
19    103000.0
1     101000.0
17     98600.0
44     97300.0
13     95500.0
30     94700.0
38     93600.0
37     92400.0
Name: Mid-Career Median Salary, dtype: float64
```

In [28]:
```
mid_sal = mid_sal.iloc[::-1]
```

In [29]:
```
undergrad_major_mid = degrees_sorted_mid['Undergraduate Major'].head(10)
```

In [30]:
```
undergrad_major_mid = undergrad_major_mid.iloc[::-1]
```

In [31]:
```
plt.barh(undergrad_major_mid, mid_sal)
plt.title('Top 10 Jobs by Mid-Career Median Salary')
plt.xlabel('Salary in USD')
plt.show()
```



Analyzing Salaries by College Type

Now, we will find the which college type offers the highest starting median salary. In order to that, we will need to group all colleges of the similar type together and take the mean.

In [32]:
```
college_type = college_type_df["School Type"].unique()
```

In [33]:
```
college_type = ['Ivy Leauge', 'Engineering', 'Liberal Arts', 'Party', 'State']
```

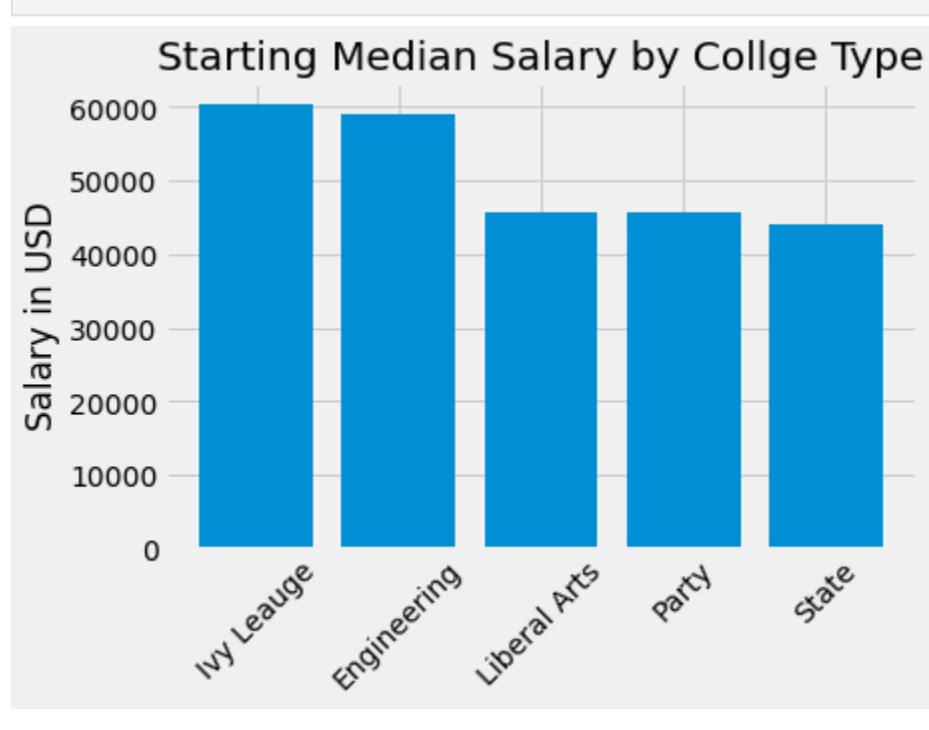
In [34]:
```
college_type_start_mean = college_type_df.groupby("School Type")["Starting Median Salary"].mean()
```

In [35]:
```
college_type_start_mean
```

Out[35]:
```
School Type
Engineering     59057.894737
Ivy League      60475.000000
Liberal Arts    45746.808511
Party           45715.000000
State           44126.285714
Name: Starting Median Salary, dtype: float64
```

In [36]:
```
college_type_start_plot = college_type_start_mean.sort_values().iloc[::-1]
```

In [37]:
```
plt.bar(college_type, college_type_start_plot)
plt.title("Starting Median Salary by Collge Type")
plt.ylabel("Salary in USD")
plt.xticks(rotation=45)
plt.show()
```



Next, we will find the highest mid-career salary by college type

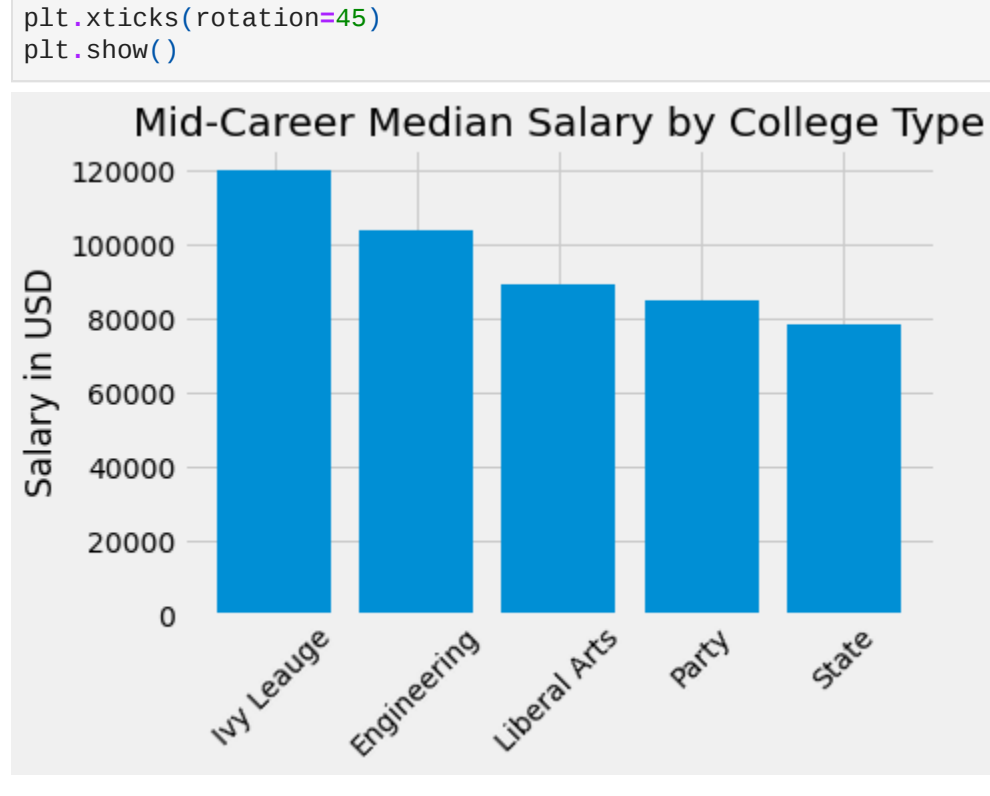
In [38]:
```
college_type_mid_mean = college_type_df.groupby("School Type")["Mid-Career Median Salary"].mean()
```

In [39]:
```
college_type_mid_mean
```

Out[39]:
```
School Type
Engineering     103842.105263
Ivy League      120125.000000
Liberal Arts     89378.723404
Party            84685.000000
State            78567.428571
Name: Mid-Career Median Salary, dtype: float64
```

In [40]:
```
college_type_mid_mean = college_type_mid_mean.sort_values().iloc[::-1]
```

In [41]:
```
plt.bar(college_type, college_type_mid_mean)
plt.title("Mid-Career Median Salary by College Type")
plt.ylabel("Salary in USD")
plt.xticks(rotation=45)
plt.show()
```



Analyzing Salaries by College Region

In order to get a clear image on salaries by region, we need to group colleges by the region they are present in and then find mean. First, we will find the mean starting median salary by college region

In [42]:
```
regions = region_df['Region'].unique()
```

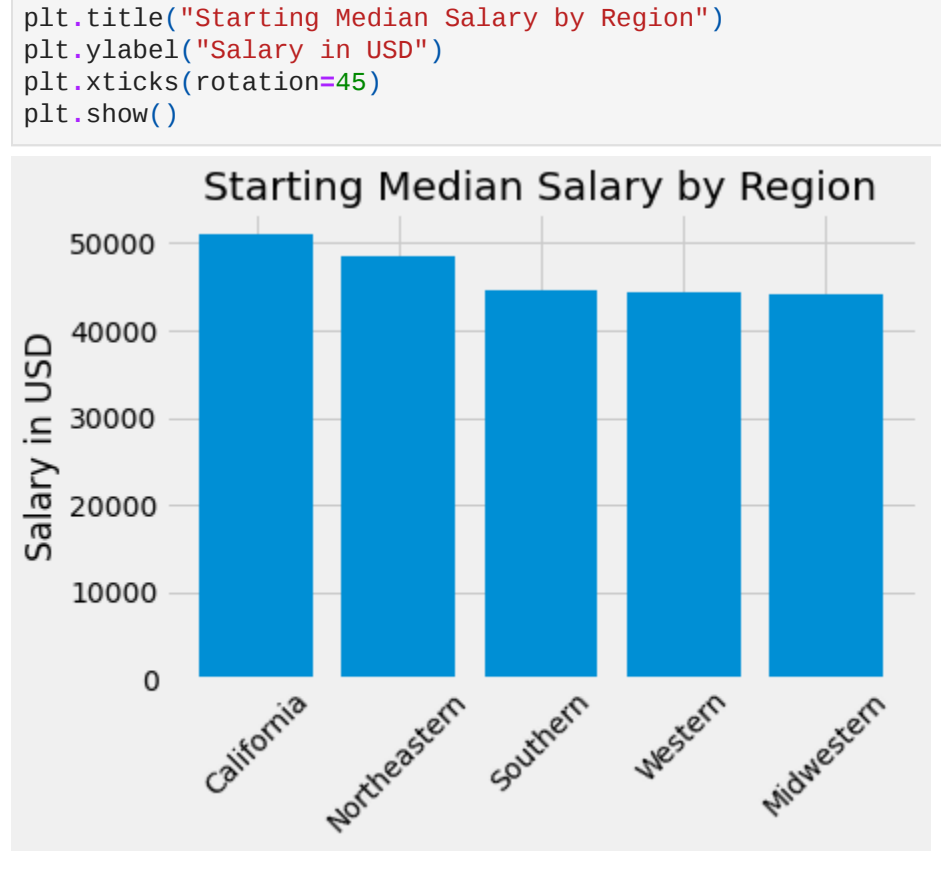
In [43]:
```
regions = ['California', 'Northeastern', 'Southern', 'Western', 'Midwestern' ]
```

In [44]:
```
region_start_mean = (region_df.groupby("Region")["Starting Median Salary"].mean()).sort_values().iloc[::-1]
```

In [45]:
```
region_start_mean
```

Out[45]:
```
Region
California      51032.142857
Northeastern    48496.000000
Southern        44521.518987
Western         44414.285714
Midwestern      44225.352113
Name: Starting Median Salary, dtype: float64
```

In [46]:
```
plt.bar(regions, region_start_mean)
plt.title("Starting Median Salary by Region")
plt.ylabel("Salary in USD")
plt.xticks(rotation=45)
plt.show()
```



ADD GRAPH ABOUT START MED SAL BY REGION

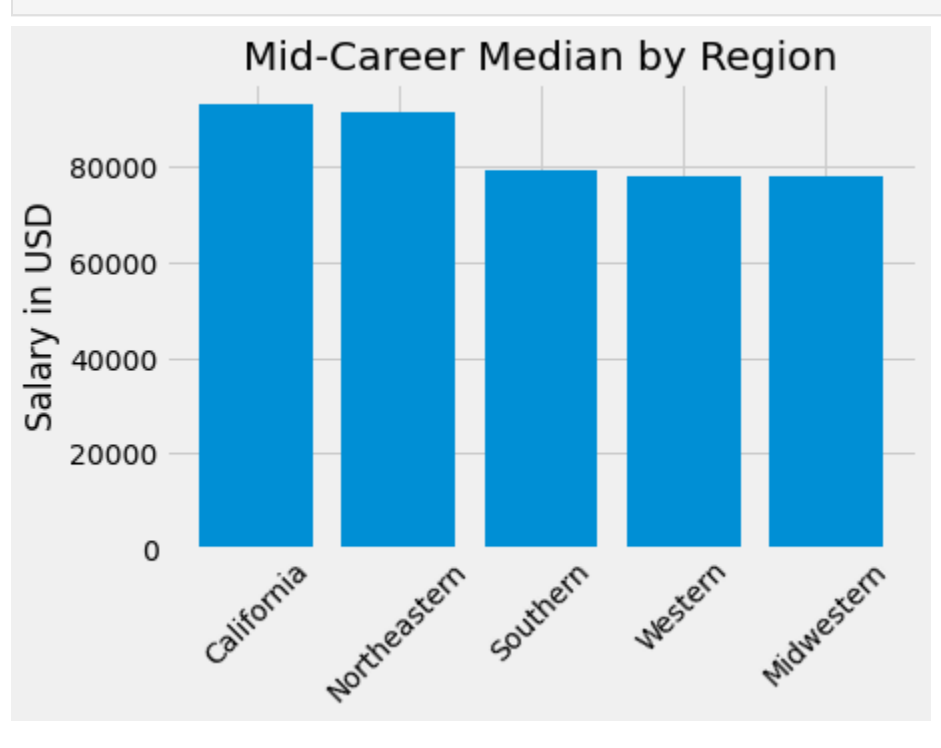
Now to find the mean mid-career salaries by region

In [47]:
```
region_mid_mean = region_df.groupby("Region")["Mid-Career Median Salary"].mean().sort_values().iloc[::-1]
```

In [48]:
```
region_mid_mean
```

Out[48]:
```
Region
California      93132.142857
Northeastern    91352.000000
Southern        79505.063291
Western         78200.000000
Midwestern      78180.281690
Name: Mid-Career Median Salary, dtype: float64
```

In [49]:
```
plt.bar(regions, region_mid_mean)
plt.title("Mid-Career Median by Region")
plt.ylabel("Salary in USD")
plt.xticks(rotation=45)
plt.show()
```



Conclusion

From the analysis of the data, it is evident Physician's Assistant offer the highest starting median salary. For individuals looking to earn more money immediately after graduation, it is recommended they pursue an Undergraduate Major to become Physician Assistant. If they are looking long term, then Chemical Engineering offers the highest median mid-career salary. If an individual is using College Type as the deciding factor, Ivy League schools offer both the highest starting median salaries and mid-career median salaries. If the individual is looking for colleges by region, California regions offering the highest starting median salaries and mid-career median salaries.