Instructions

We are asking you to conduct some introductory analyses into the dataset(s) and to report back your results. Specifically, the information we would like to see is:

1. Focusing on the data available for all ages, what does the distribution of unemployment rates look like among the different major categories? Come up with a graphical display that allows a reader to easily make sense of the information.

1. In addition to the comprehensive, all-ages dataset, the github repository also contains data regarding just recent college graduates (ages < 28). Comparing this subset of data to the whole dataset that it comes from (all-ages) can provide us with some information about recent trends. Which majors appear to have experienced a relative boom among recent graduates and which majors are dropping off in popularity? Again, explore visual ways of describing the answer as well as numerical ones.

1. (Bonus) The previous two questions deal with only a small subset of the data contained in the repository. If you have some extra time (this question isn't required for the application), we would be curious to see something else interesting you found while exploring the data. Additionally, if there are other variables or similar data sets that you could see being useful to add to this data set, feel free to mention them here! What would you use this additional data for? (Don't worry, you don't need to actually do the linking, this is more of a hypothetical question).

Question 1

In [98]:

```
#import statements
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
%matplotlib inline
```

In [3]:

```
all_ages = pd.read_csv("all-ages.csv")
```

In [4]:

```
AA_MC = all_ages.groupby("Major_category")
```

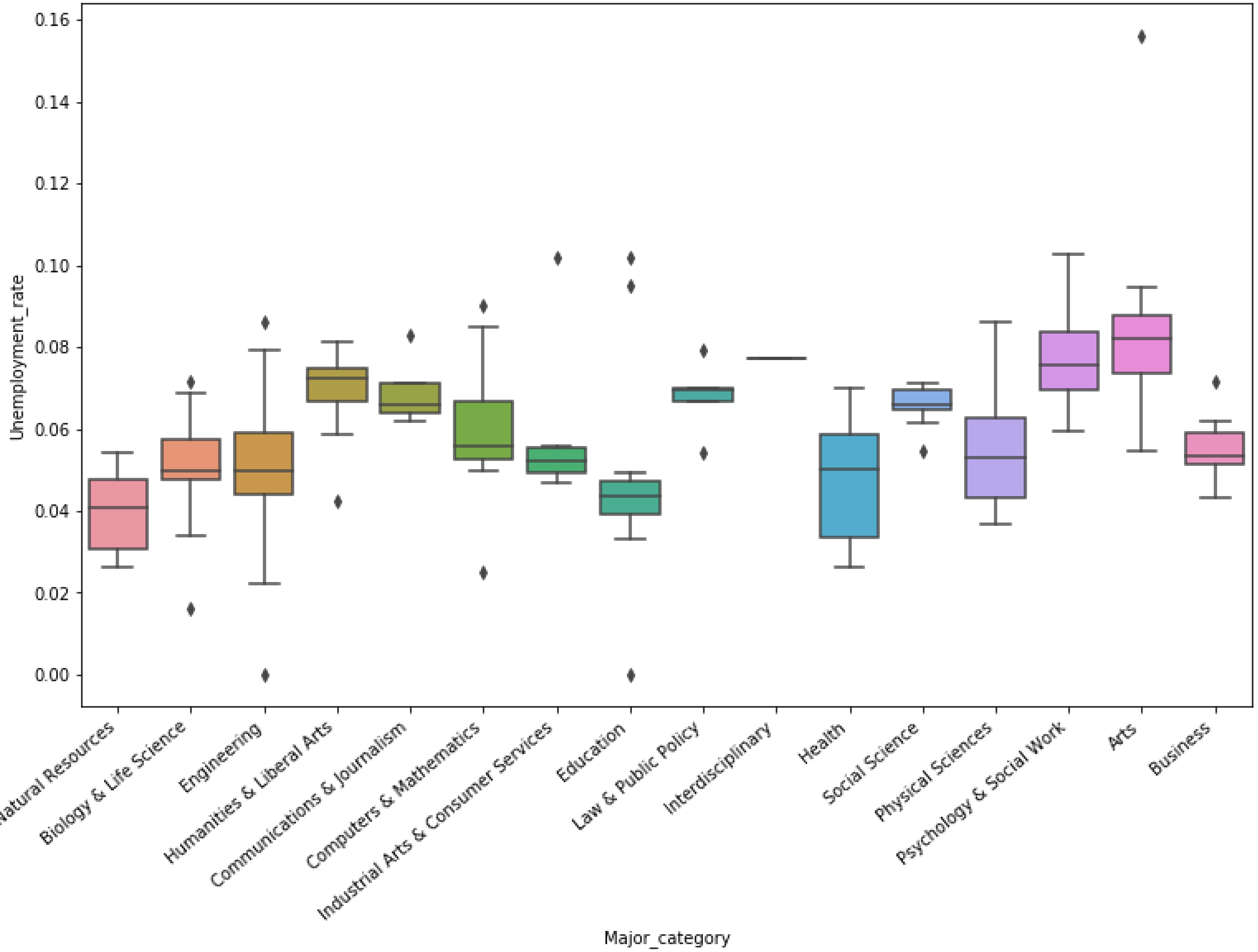
```
In [97]: AA_MC.Unemployment_rate.describe().sort_values('mean', ascending=False)
```

Out[97]:

	count	mean	std	min	25%	50%	75%	max
Major_category								
Arts	8.0	0.087601	0.030097	0.054719	0.073378	0.081994	0.087879	0.156147
Psychology & Social Work	9.0	0.077867	0.012738	0.059376	0.069667	0.075631	0.083629	0.102712
Interdisciplinary	1.0	0.077269	NaN	0.077269	0.077269	0.077269	0.077269	0.077269
Humanities & Liberal Arts	15.0	0.069429	0.009543	0.042505	0.066715	0.072374	0.074675	0.081348
Communications & Journalism	4.0	0.069125	0.009504	0.061917	0.063749	0.065788	0.071163	0.083005
Law & Public Policy	5.0	0.067854	0.009070	0.054036	0.066513	0.069655	0.069848	0.079217
Social Science	9.0	0.065686	0.005278	0.054399	0.064519	0.065804	0.069374	0.071057
Computers & Mathematics	11.0	0.059437	0.018172	0.024900	0.052366	0.055653	0.066870	0.090264
Industrial Arts & Consumer Services	7.0	0.058546	0.019373	0.046903	0.049180	0.052034	0.055363	0.101796
Physical Sciences	10.0	0.054541	0.015380	0.036726	0.043402	0.052993	0.062515	0.086022
Business	13.0	0.054496	0.007606	0.043268	0.051378	0.053415	0.058865	0.071354
Engineering	29.0	0.050630	0.015761	0.000000	0.043844	0.049846	0.058821	0.085991
Biology & Life Science	14.0	0.049936	0.013896	0.016111	0.047777	0.049899	0.057298	0.071598
Health	12.0	0.047209	0.015766	0.026292	0.033607	0.050020	0.058557	0.070010
Education	16.0	0.046762	0.023238	0.000000	0.039001	0.043830	0.047379	0.101746
Agriculture & Natural Resources	10.0	0.039569	0.010023	0.026147	0.030634	0.040897	0.047561	0.054341

In [6]:

```
a4_dims = (11.7, 8.27)
fig, ax = plt.subplots(figsize=a4_dims)
ax = sns.boxplot(x=all_ages.Major_category, y=all_ages.Unemployment_rate, orient='v')
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")
plt.tight_layout()
plt.show()
```



Interpretation

The graph above shows the distribution of unemployment rates by major category. Above the graph, we have the descriptive statistics for each major category, organized so that the highest average unemployment rate is at the top and the lowest is at the bottom. From this, we can see that the following major categories have the highest unemployment rates among all ages:

- Arts
- Pyschology & Social Work
- Interdisciplinary
- Humanities & Liberal Arts
- Communications & Journalism

Question 2

In [47]:

```
recent_grads = pd.read_csv("recent-grads.csv")
```

In [62]:

```
total_recentgrads = recent_grads.Total.sum()
total_allages = all_ages.Total.sum()
total_allages = total_allages.astype(float)
print("Total Recent Grads:", total_recentgrads, ", Total All Grads:", total_allages)
```

```
Total Recent Grads: 6771654.0 , Total All Grads: 39834398.0
```

In [68]:

```
recent_grads.sort_values('Major_code')
all_ages.sort_values('Major_code')
print("all sorted!")
```

```
all sorted!
```

In [70]:

```
rg_proportions = []
for i in recent_grads.Total:
    i = i/total_recentgrads
    rg_proportions.append(i)

aa_proportions = []
for i in all_ages.Total:
    i = i/total_allages
    aa_proportions.append(i)

all_majors = []
for i in all_ages.Major:
    all_majors.append(i)

all_categories = []
for i in all_ages.Major_category:
    all_categories.append(i)
```

In [88]:

```
data2 = {'Major': all_majors,
         'Major Categories': all_categories,
        'Recent Grads Proportions': rg_proportions,
        'All Ages Proportions': aa_proportions}
```

In [89]:

```
df = pd.DataFrame(data2)
df.head()
```

Out[89]:

	Major	Major Categories	Recent Grads Proportions	All Ages Proportions
0	GENERAL AGRICULTURE	Agriculture & Natural Resources	0.000345	0.003217
1	AGRICULTURE PRODUCTION AND MANAGEMENT	Agriculture & Natural Resources	0.000112	0.002393
2	AGRICULTURAL ECONOMICS	Agriculture & Natural Resources	0.000126	0.000852
3	ANIMAL SCIENCES	Agriculture & Natural Resources	0.000186	0.002599
4	FOOD SCIENCE	Agriculture & Natural Resources	0.004764	0.000610

In [86]:

```
df2 = pd.melt(df, id_vars="Major Categories", var_name="Dataset", value_name="Relative Popularity")
```

In [95]:

```
df["difference"] = df["Recent Grads Proportions"] - df["All Ages Proportions"]
df.sort_values("difference", ascending=False)
```

Out[95]:

	Major	Major Categories	Recent Grads Proportions	All Ages Proportions	difference
145	STUDIO ARTS	Arts	0.058145	0.002034	0.056111
76	HUMANITIES	Humanities & Liberal Arts	0.048722	0.001160	0.047562
123	PUBLIC POLICY	Law & Public Policy	0.041454	0.000371	0.041082
57	NAVAL ARCHITECTURE AND MARINE ENGINEERING	Engineering	0.034643	0.000404	0.034239
93	MILITARY TECHNOLOGIES	Industrial Arts & Consumer Services	0.031602	0.000108	0.031493
77	LIBRARY SCIENCE	Education	0.030304	0.000407	0.029898
34	SPECIAL NEEDS EDUCATION	Education	0.030922	0.003758	0.027164
137	TRANSPORTATION SCIENCES AND TECHNOLOGIES	Industrial Arts & Consumer Services	0.028748	0.003179	0.025569
35	SOCIAL SCIENCE OR HISTORY TEACHER EDUCATION	Education	0.025770	0.003189	0.022581
94	MULTI/INTERDISCIPLINARY STUDIES	Interdisciplinary	0.022568	0.001135	0.021434
113	NUCLEAR, INDUSTRIAL RADIOLOGY, AND BIOLOGICAL ...	Physical Sciences	0.021223	0.000305	0.020918
36	TEACHER EDUCATION: MULTIPLE LEVELS	Education	0.020563	0.002211	0.018352
40	GENERAL ENGINEERING	Engineering	0.029333	0.012629	0.016704
124	HUMAN SERVICES AND COMMUNITY ORGANIZATION	Psychology & Social Work	0.017047	0.002053	0.014993
139	DRAMA AND THEATER ARTS	Arts	0.018470	0.004389	0.014082
95	INTERCULTURAL AND INTERNATIONAL STUDIES	Humanities & Liberal Arts	0.015281	0.001420	0.013861
138	FINE ARTS	Arts	0.025232	0.014358	0.010873
8	ENVIRONMENTAL SCIENCE	Biology & Life Science	0.013472	0.002664	0.010808
9	FORESTRY	Agriculture & Natural Resources	0.012039	0.001743	0.010296
96	NUTRITION SCIENCES	Health	0.010724	0.001620	0.009104
41	AEROSPACE ENGINEERING	Engineering	0.010691	0.001650	0.009041
97	MATHEMATICS AND COMPUTER SCIENCE	Computers & Mathematics	0.009163	0.000180	0.008983

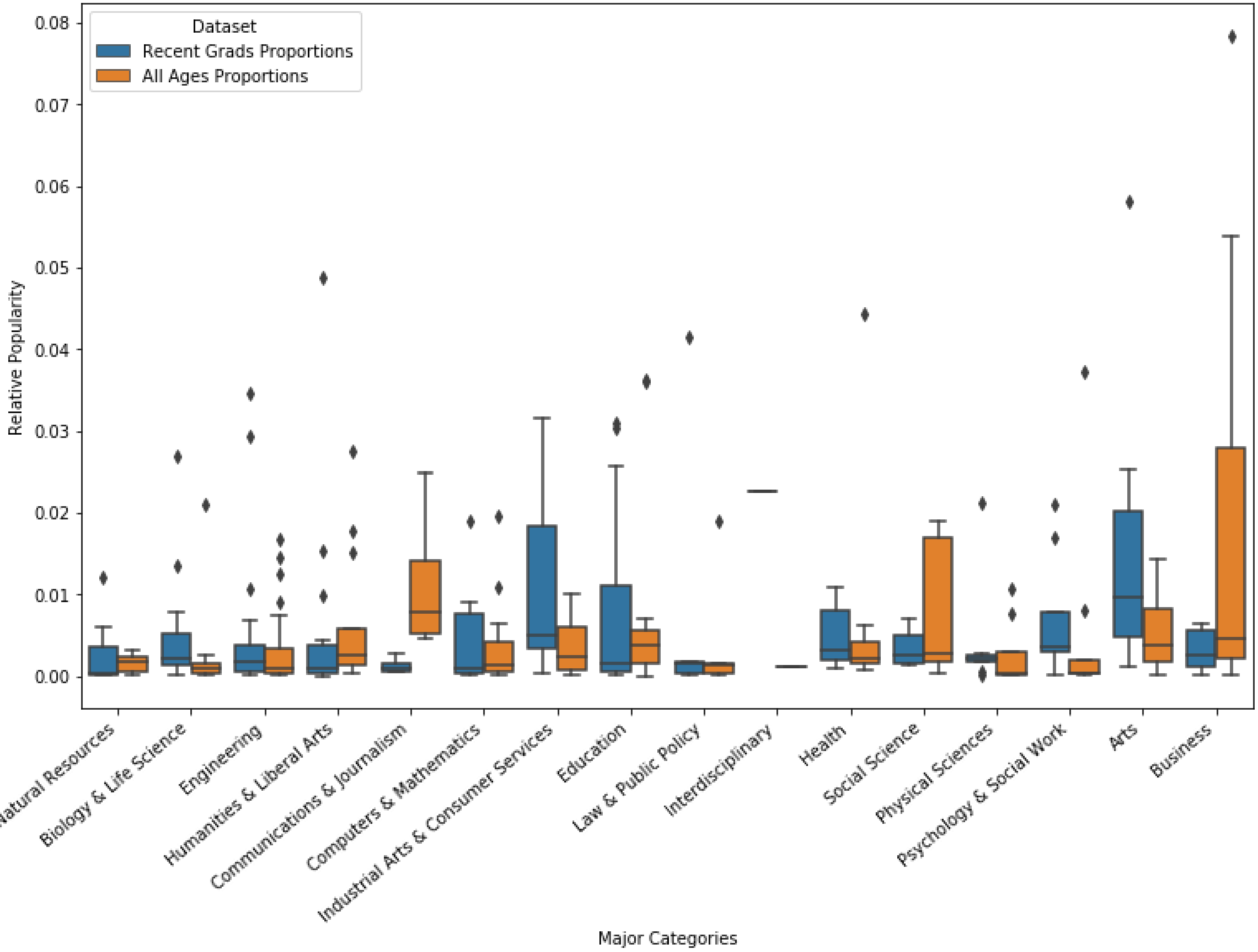
	Major	Major Categories	Recent Grads Proportions	All Ages Proportions	difference
146	MISCELLANEOUS FINE ARTS	Arts	0.008954	0.000214	0.008740
74	COMPOSITION AND RHETORIC	Humanities & Liberal Arts	0.009825	0.001486	0.008338
149	HEALTH AND MEDICAL ADMINISTRATIVE SERVICES	Health	0.010993	0.002724	0.008269
98	COGNITIVE SCIENCE AND BIOPSYCHOLOGY	Biology & Life Science	0.007851	0.000173	0.007677
17	COMMUNICATION TECHNOLOGIES	Computers & Mathematics	0.009031	0.001560	0.007471
150	MEDICAL ASSISTING SERVICES	Health	0.008565	0.001615	0.006951
115	EDUCATIONAL PSYCHOLOGY	Psychology & Social Work	0.007125	0.000352	0.006772
58	NUCLEAR ENGINEERING	Engineering	0.006855	0.000247	0.006608
...	...	...	...	...	...
106	CHEMISTRY	Physical Sciences	0.002237	0.007734	-0.005496
33	SECONDARY TEACHER EDUCATION	Education	0.000106	0.005630	-0.005524
18	COMPUTER AND INFORMATION SYSTEMS	Computers & Mathematics	0.000417	0.006371	-0.005954
30	PHYSICAL AND HEALTH EDUCATION TEACHING	Education	0.000598	0.007071	-0.006473
46	CIVIL ENGINEERING	Engineering	0.000923	0.009002	-0.008079
70	FAMILY AND CONSUMER SCIENCES	Industrial Arts & Consumer Services	0.000445	0.010093	-0.009648
14	JOURNALISM	Communications & Journalism	0.000638	0.010496	-0.009858
90	MATHEMATICS	Computers & Mathematics	0.000292	0.010865	-0.010573
112	MULTI-DISCIPLINARY OR GENERAL SCIENCE	Physical Sciences	0.000101	0.010743	-0.010642
142	COMMERCIAL ART AND GRAPHIC DESIGN	Arts	0.001908	0.012669	-0.010761
127	ECONOMICS	Social Science	0.006446	0.019019	-0.012574
54	MECHANICAL ENGINEERING	Engineering	0.000566	0.014599	-0.014033
133	SOCIOLOGY	Social Science	0.002529	0.016934	-0.014405
75	LIBERAL ARTS	Humanities & Liberal Arts	0.000360	0.015093	-0.014733
114	PSYCHOLOGY	Psychology & Social Work	0.020963	0.037256	-0.016294
48	ELECTRICAL ENGINEERING	Engineering	0.000260	0.016861	-0.016601

	Major	Major Categories	Recent Grads Proportions	All Ages Proportions	difference
121	CRIMINAL JUSTICE AND FIRE PROTECTION	Law & Public Policy	0.001881	0.019007	-0.017126
171	HISTORY	Humanities & Liberal Arts	0.000683	0.017887	-0.017204
132	POLITICAL SCIENCE AND GOVERNMENT	Social Science	0.001352	0.018802	-0.017450
165	FINANCE	Business	0.001655	0.020499	-0.018844
164	MARKETING AND MARKETING RESEARCH	Business	0.005551	0.027981	-0.022431
13	COMMUNICATIONS	Communications & Journalism	0.000632	0.024795	-0.024163
73	ENGLISH LANGUAGE AND LITERATURE	Humanities & Liberal Arts	0.000018	0.027580	-0.027562
25	GENERAL EDUCATION	Education	0.007849	0.036121	-0.028272
28	ELEMENTARY EDUCATION	Education	0.001349	0.036318	-0.034969
153	NURSING	Health	0.002400	0.044431	-0.042032
159	ACCOUNTING	Business	0.002507	0.044665	-0.042158
158	GENERAL BUSINESS	Business	0.004461	0.053941	-0.049480
161	BUSINESS MANAGEMENT AND ADMINISTRATION	Business	0.000220	0.078412	-0.078193
21	INFORMATION SCIENCES	Computers & Mathematics	NaN	0.001953	NaN

173 rows × 5 columns

In [87]:

```
a4_dims = (11.7, 8.27)
fig, ax = plt.subplots(figsize=a4_dims)
ax = sns.boxplot(x='Major Categories', y='Relative Popularity', hue='Dataset', data=df2)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")
plt.tight_layout()
plt.show()
```



Interpretation

Because there are so many majors, I thought the visual would be meaningless unless the majors were grouped into major categories. Here I made side by side bar graphs that help people see how popular major categories are among all ages and recent graduates.

However, I did look at individual majors numerically. Above is a chart of all 172 majors available in both datasets, the proportions of graduates in each major compared to the dataset the numbers come from, and the difference in proportions. Any major that has a positive number in the "difference" column became more popular recently. The top ten majors that had a notably high increase in popularity are:

- Studio Arts
- Humanities
- Public Policy
- Naval Architecture and Marine Engineering
- Military Technologies
- Library Science
- Special Needs Education
- Transportation Sciences and Technologies
- Social Science or History Education
- Interdisciplinary Studies

Above the bargraph, the dataframe is sorted from highest increase in popularity to lowest increase in popularity, so if you would like to look at other majors that increased in popularity please reference that.

Bonus Question

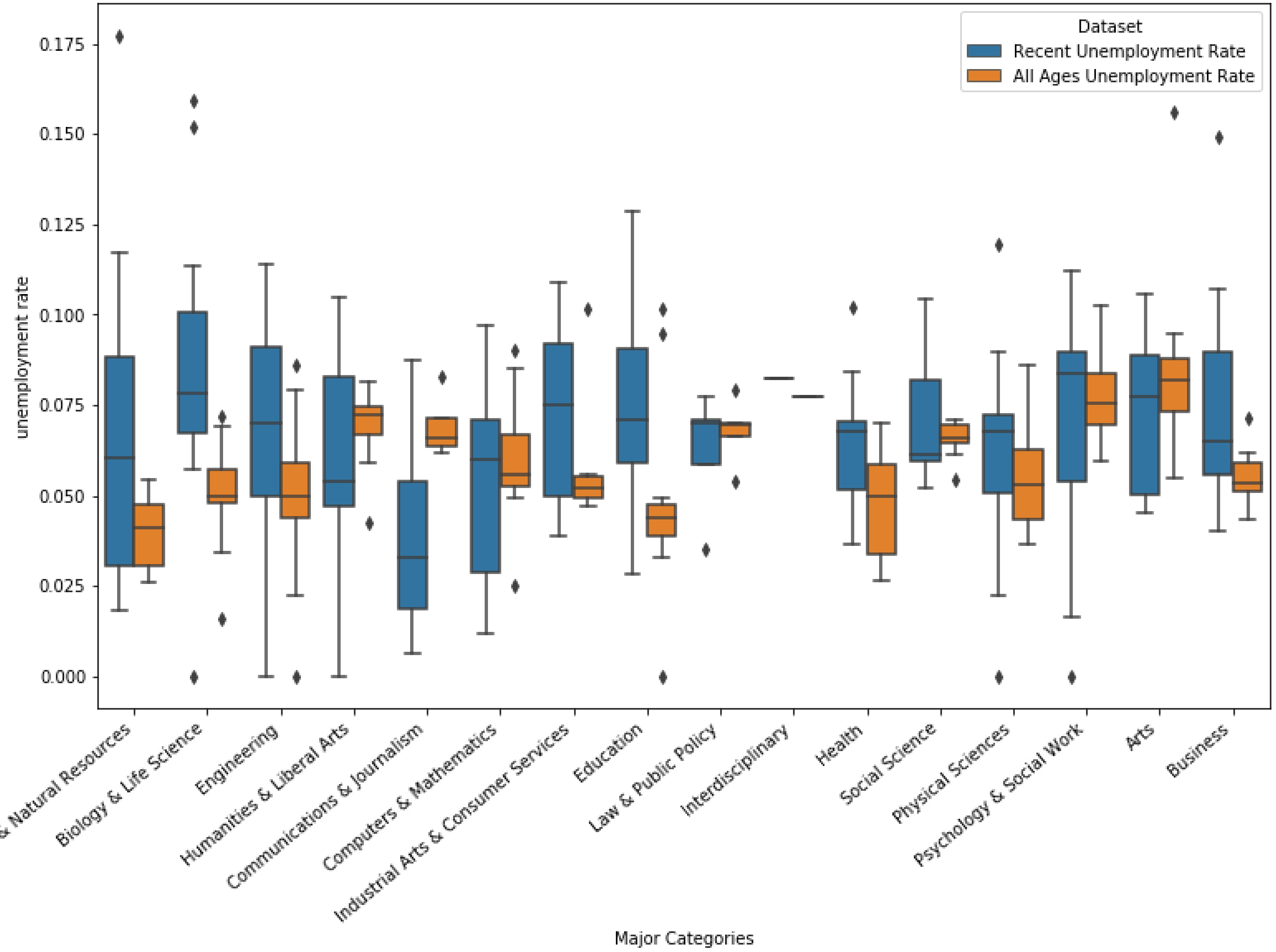
Because I accidentally misterpreted the second question at first, I actually had the opportunity to look at the difference in unemployment rates between the "all ages" and "recent grads" dataset by major category. The boxplot is shown below.

From this boxplot, we can see that unfortunately, many disciplines are suffering from higher unemployment rates. However. recent graduates within the Humanities & Liberal Arts, Communications & Journalism, and Social Science categories have lower unemployment rates on average.

I noticed that there is an additional data set of women in STEM and graduate students. I would be curious to see the unemployment rates split up by gender, especially for women in STEM. Would there be a greater commitment to hiring women in STEM more recently comepared the all ages dataset? How do unemployment rates in major categories fluctuate as someone gains a master degree or PhD?

In [13]:

```
a4_dims = (11.7, 8.27)
fig, ax = plt.subplots(figsize=a4_dims)
ax = sns.boxplot(x='Major Categories', y='unemployment rate', hue='Dataset', data=df2)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")
plt.tight_layout()
plt.show()
```


Dataset
🟦 Recent Unemployment Rate
🟧 All Ages Unemployment Rate
unemployment rate
Agriculture & Natural Resources
Biology & Life Science
Engineering
Humanities & Liberal Arts
Communications & Journalism
Computers & Mathematics
Industrial Arts & Consumer Services
Education
Law & Public Policy
Interdisciplinary
Health
Social Science
Physical Sciences
Psychology & Social Work
Arts
Business
Major Categories