analysis of Yale summer internship data

In [1]:
```
%matplotlib
import numpy as np
import pandas as pd
```

```
Using matplotlib backend: Qt5Agg
```

In [61]:
```
file_path = "./Pandas_Summer_2019_Peer_List.xlsx"
```

In [62]:
```
df = pd.read_excel(file_path)
```

In [63]:
```
df.head()
```

Out[63]:

	Last Name	First Name	Email	Major	Class Year (After Position)
0	Sung	Christopher	christopher.sung@yale.edu	History	Undergraduate: Junior
1	Model	Max	max.model@yale.edu	Computer Science & Mathematics	Undergraduate: Junior
2	Zhou	Huahao	huahao.zhou@yale.edu	Computer Science	Undergraduate: Junior
3	Baker	Morgan	morgan.baker@yale.edu	Women'sGender&SexualityStudies	Undergraduate: Junior
4	Williams	Marina	marina.williams@yale.edu	Psychology	Undergraduate: Senior

In [64]:
```
df.shape
```

Out[64]:
```
(1708, 16)
```

In [65]:
```
df.columns
```

Out[65]:
```
Index(['Last Name', 'First Name', 'Email', 'Major',
       'Class Year (After Position)', 'Type of Position',
       'Field Research Project Title (if relevant)', 'City',
       'Country (if outside the U.S.)', 'U.S. State or Territory', 'Em
ployer',
       'Employer: Industry', 'Employer: Function (Role)',
       'Briefly describe the interview process for this position',
       'Describe (1) the projects you worked on and (2) how much inter
action you had with your supervisor during the summer',
       'Describe the work atmosphere and culture of the organizatio
n'],
      dtype='object')
```

In [66]:
```
#df.dtypes
# rename columns
df2 = df.copy()
```

In [67]:
```
df2.rename(columns={
    df2.columns[0]: "last_name",
    df2.columns[1]: "first_name",
    df2.columns[2]: "email",
    df2.columns[3]: "major",
    df2.columns[4]: "class_year",
    df2.columns[5]: "job_type",
    df2.columns[6]: "project_name",
    df2.columns[7]: "city",
    df2.columns[8]: "country",
    df2.columns[9]: "state",
    df2.columns[10]: "employer",
    df2.columns[11]: "industry",
    df2.columns[12]: "job_role",
    df2.columns[13]: "interview",
    df2.columns[14]: "work_info",
    df2.columns[15]: "work_culture"
}, inplace = True)
```

In [68]:
```
df2.fillna('', inplace=True)
```

In [69]:
```
df2.head()
```

Out[69]:

	last_name	first_name	email	major	class_yea
0	Sung	Christopher	christopher.sung@yale.edu	History	Undergraduat Junio
1	Model	Max	max.model@yale.edu	Computer Science & Mathematics	Undergraduat Junio
2	Zhou	Huahao	huahao.zhou@yale.edu	Computer Science	Undergraduat Junio
3	Baker	Morgan	morgan.baker@yale.edu	Women'sGender&SexualityStudies	Undergraduat Junio
4	Williams	Marina	marina.williams@yale.edu	Psychology	Undergraduat Senio

top 10 majors

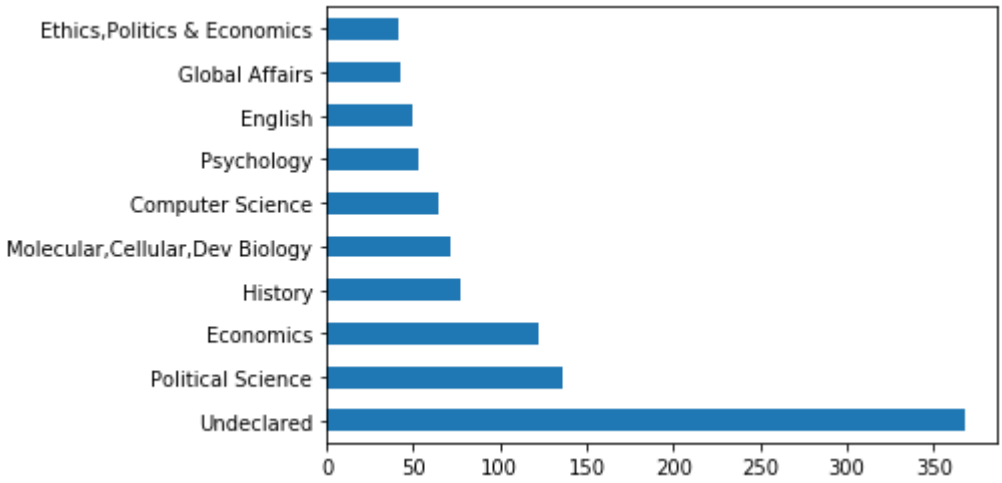
In [83]:
```
major = pd.value_counts(df2.major)
```

In [84]:
```
major.head(10)
```

Out[84]:
```
Undeclared                      368
Political Science               136
Economics                       122
History                          77
Molecular,Cellular,Dev Biology   72
Computer Science                 64
Psychology                       53
English                          50
Global Affairs                   43
Ethics,Politics & Economics      41
Name: major, dtype: int64
```

In [85]:
```
major[:10].plot(kind='barh')
```

Out[85]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f60d809ffd0>
```



class distribution

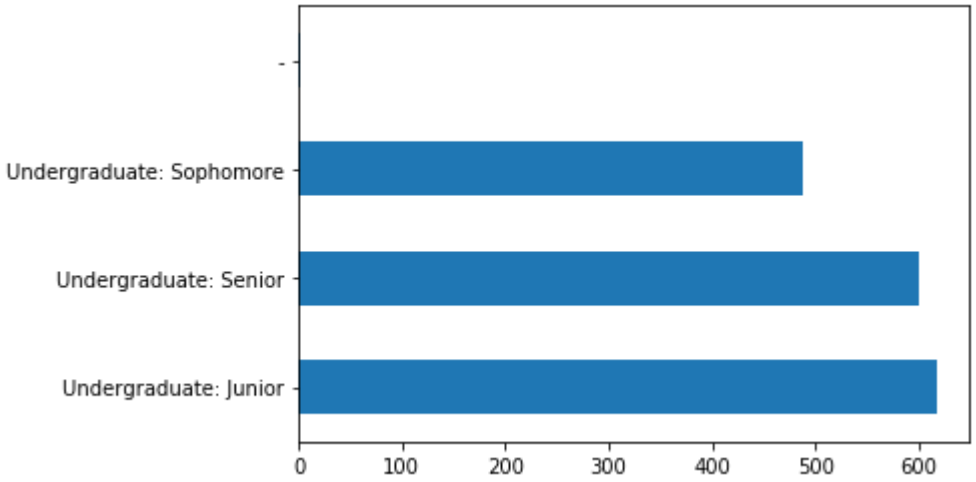
In [86]:
```
class_year = pd.value_counts(df2.class_year)

class_year.head()
```

Out[86]:
```
Undergraduate: Junior       618
Undergraduate: Senior       600
Undergraduate: Sophomore    488
-                             2
Name: class_year, dtype: int64
```

In [87]:
```
class_year.plot(kind='barh')
```

Out[87]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f60cf903630>
```



top 10 job types

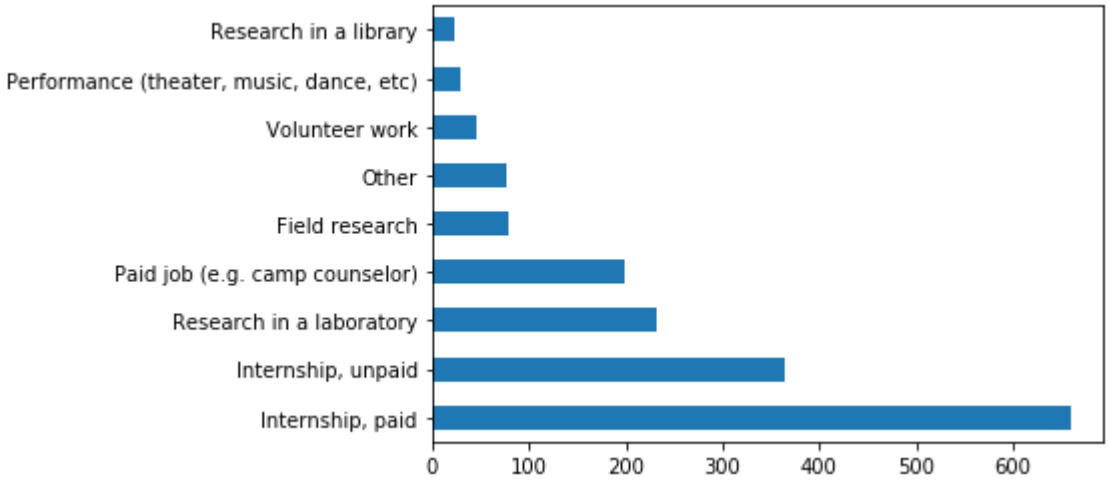
In [88]:
```
job_type = pd.value_counts(df2.job_type)

job_type.head(10)
```

Out[88]:
```
Internship, paid                          659
Internship, unpaid                        364
Research in a laboratory                  232
Paid job (e.g. camp counselor)            198
Field research                             79
Other                                      77
Volunteer work                             46
Performance (theater, music, dance, etc)   29
Research in a library                      24
Name: job_type, dtype: int64
```

In [89]:
```
job_type[:10].plot(kind='barh')
```

Out[89]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f60cf865940>
```



top 20 Employers

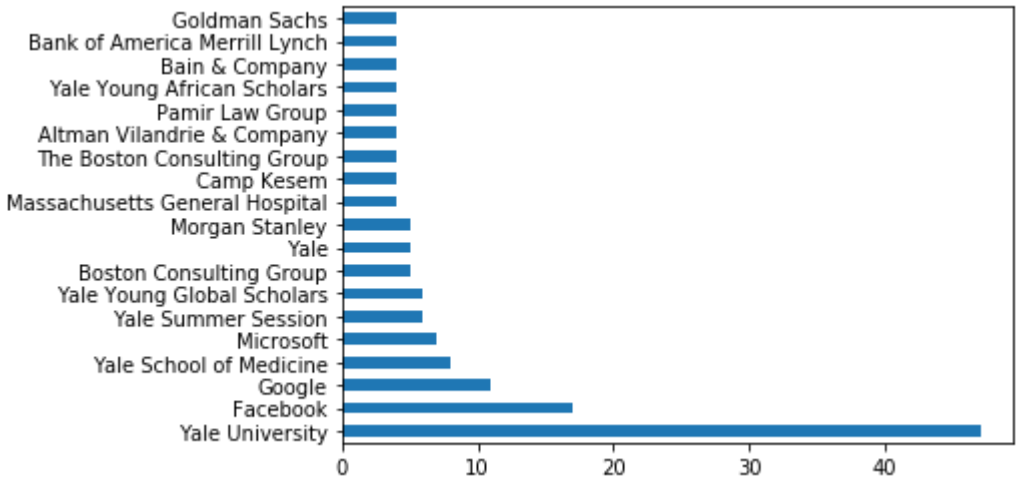
In [90]:
```
employer = pd.value_counts(df2.employer)

employer.head(20)
```

Out[90]:
```
                                209
Yale University                  47
Facebook                         17
Google                           11
Yale School of Medicine           8
Microsoft                         7
Yale Summer Session               6
Yale Young Global Scholars        6
Boston Consulting Group           5
Yale                              5
Morgan Stanley                    5
Massachusetts General Hospital    4
Camp Kesem                        4
The Boston Consulting Group       4
Altman Vilandrie & Company        4
Pamir Law Group                   4
Yale Young African Scholars       4
Bain & Company                    4
Bank of America Merrill Lynch     4
Goldman Sachs                     4
Name: employer, dtype: int64
```

In [94]:
```
employer[1:20].plot(kind='barh')
```

Out[94]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f60cf75eb70>
```



top 20 Industries

In [75]:
```
# replace "-" with ""
df2['industry'] = df2['industry'].apply(lambda x: "" if x.strip() == "-
```

In [92]:
```
industry = pd.value_counts(df2.industry)

industry.head(20)
```

Out[92]:
```
                                                                                4
38
Academia/Education  (including University research positions)                   1
89
Finance/Insurance/Real Estate                                                   1
64
Technology                                                                      1
21
Government (including local, state, federal and military service)               1
05
Healthcare/Medical/Pharmaceutical
91
Law/Legal Services
76
Community Organizations Advocacy/Social Services
68
Consulting
65
Publishing/Media/Journalism
45
Environment
33
Entertainment/Film/Television
31
Communications/Marketing/Advertising / PR
29
Think Tank
26
Engineering
26
Arts Administration
22
Hospitality
21
Food Systems
21
International Development
18
Social Enterprise/Economic Development
17
Name: industry, dtype: int64
```

In [95]:
```
industry[1:20].plot(kind='barh')
```

Out[95]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f60cf6ca630>
```

top 20 Job Roles

In [77]:
```
# replace "-" with ""
df2['job_role'] = df2['job_role'].apply(lambda x: "" if x.strip() == "-
```

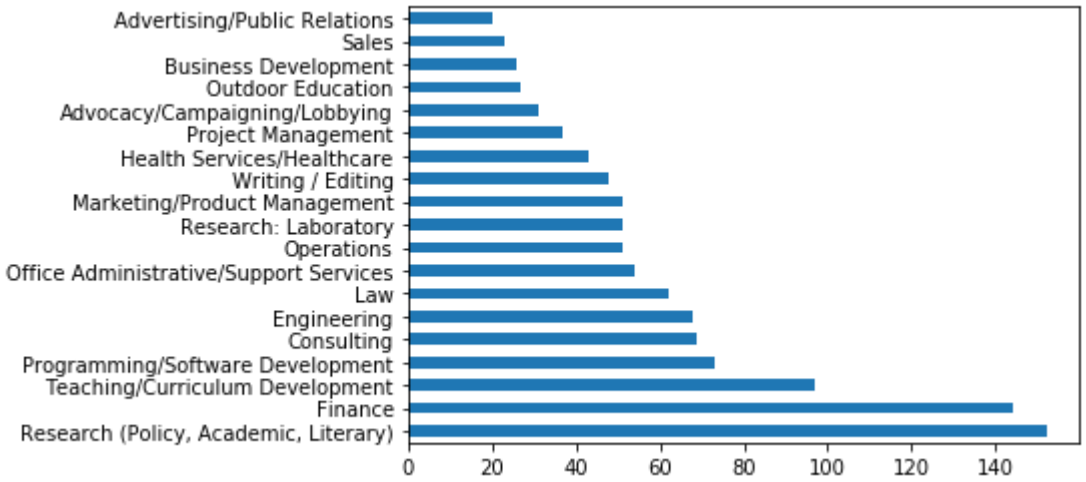
In [96]:
```
job_role = pd.value_counts(df2.job_role)

job_role.head(20)
```

Out[96]:
```
                                          439
Research (Policy, Academic, Literary)     152
Finance                                   144
Teaching/Curriculum Development            97
Programming/Software Development           73
Consulting                                 69
Engineering                                68
Law                                        62
Office Administrative/Support Services     54
Operations                                 51
Research: Laboratory                       51
Marketing/Product Management               51
Writing / Editing                          48
Health Services/Healthcare                 43
Project Management                         37
Advocacy/Campaigning/Lobbying              31
Outdoor Education                          27
Business Development                       26
Sales                                      23
Advertising/Public Relations               20
Name: job_role, dtype: int64
```

In [97]:
```
job_role[1:20].plot(kind='barh')
```

Out[97]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f60cf6006a0>
```



In [98]:
```
df2.to_excel("Summer_2019_Peer_List.xlsx")
```

In [ ]:
```

```