H-1B Visa Petition

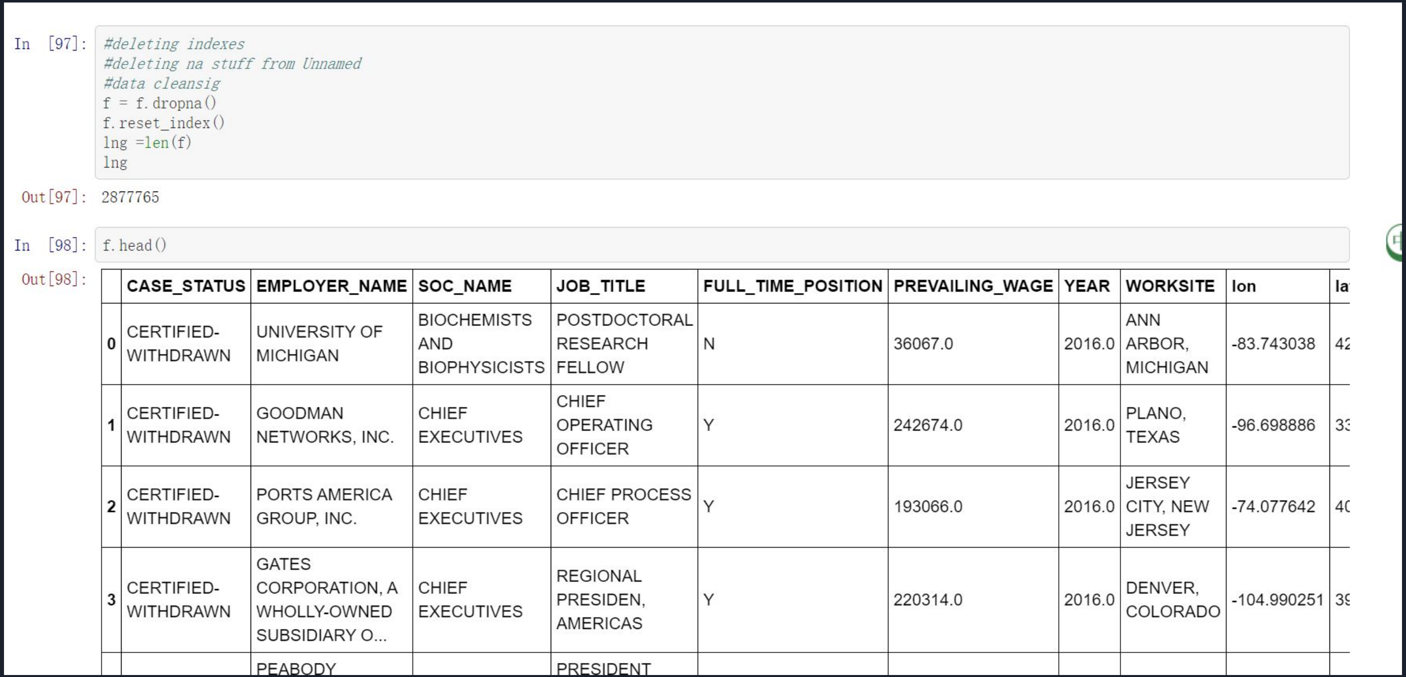
Author: Team 8
Members:
Tianyu Wang
Mingzhuo Yu
Yongji Shen
Yangyang Zhang

Problem Statement

- Perform exploratory data analysis with python to get insights from data
- Tools & Libraries applied: Pandas, Matplotlib, Seaborn, Sklearn
- Use ROC curve to visualize data with RandomForestClassifier and LogisticRegression

Import Data

Import Data and Delete NA Values

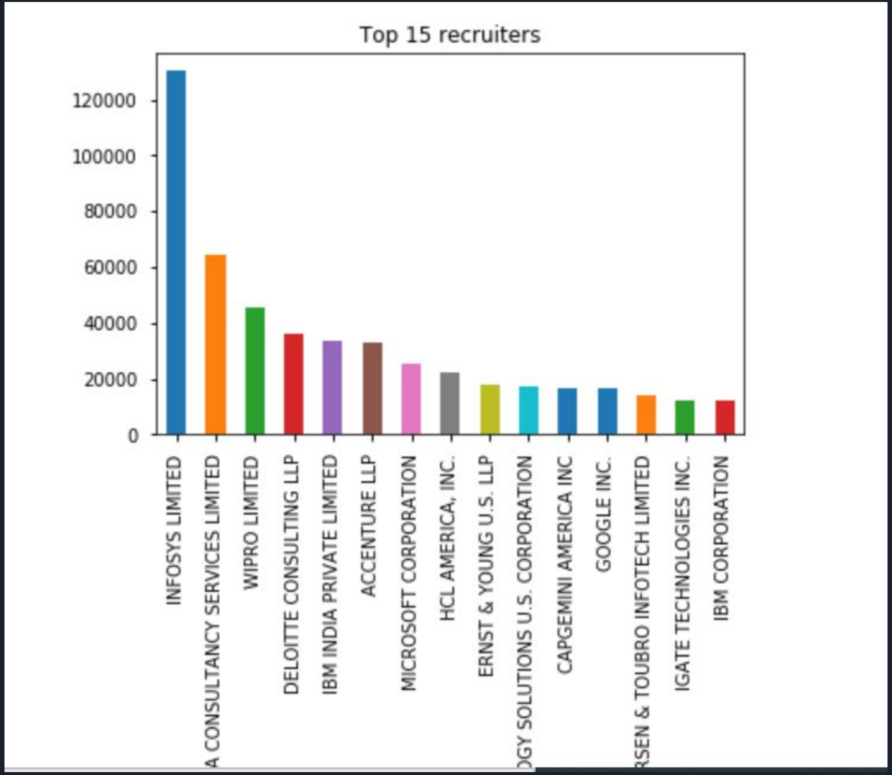


```
In  [97]: #deleting indexes
          #deleting na stuff from Unnamed
          #data cleansig
          f = f.dropna()
          f.reset_index()
          lng =len(f)
          lng

Out[97]:  2877765

In  [98]: f.head()

Out[98]:
```

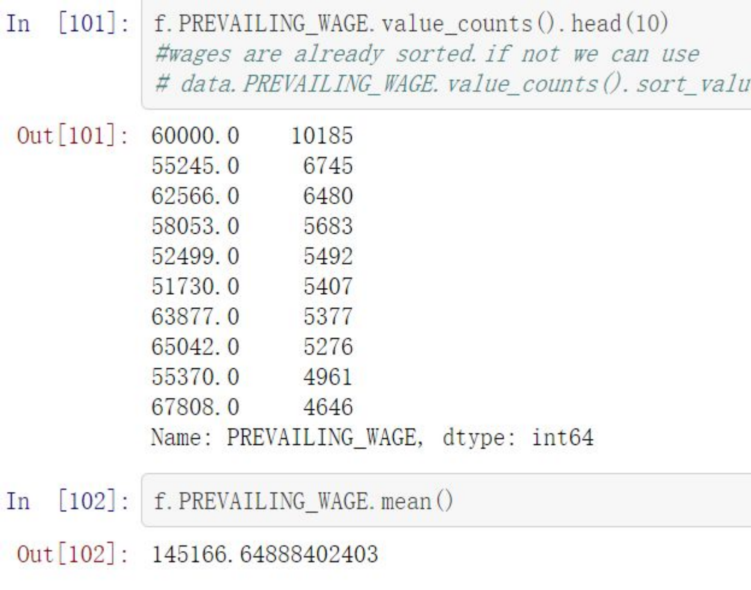
	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	la
0	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN	-83.743038	42
1	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	PLANO, TEXAS	-96.698886	33
2	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	JERSEY CITY, NEW JERSEY	-74.077642	40
3	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O...	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO	-104.990251	39
		PEABODY		PRESIDENT						

Data Analysis

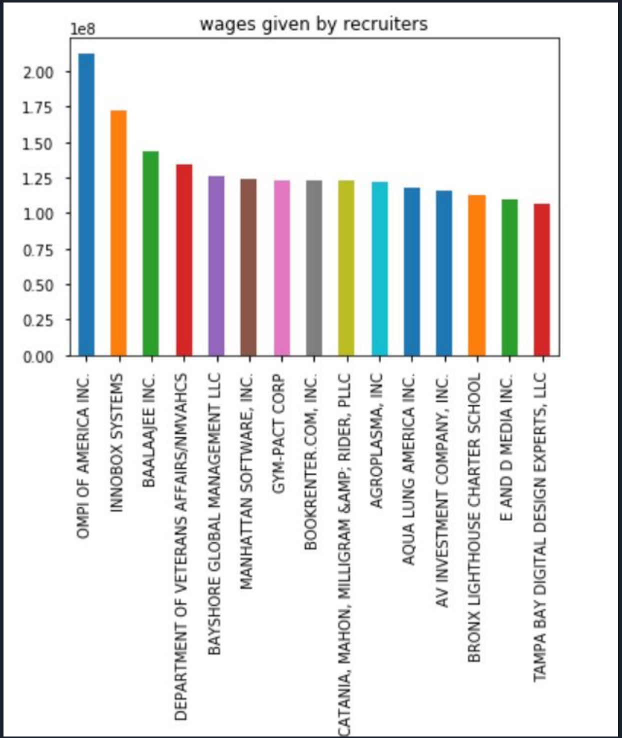
Top 15 Recruiters



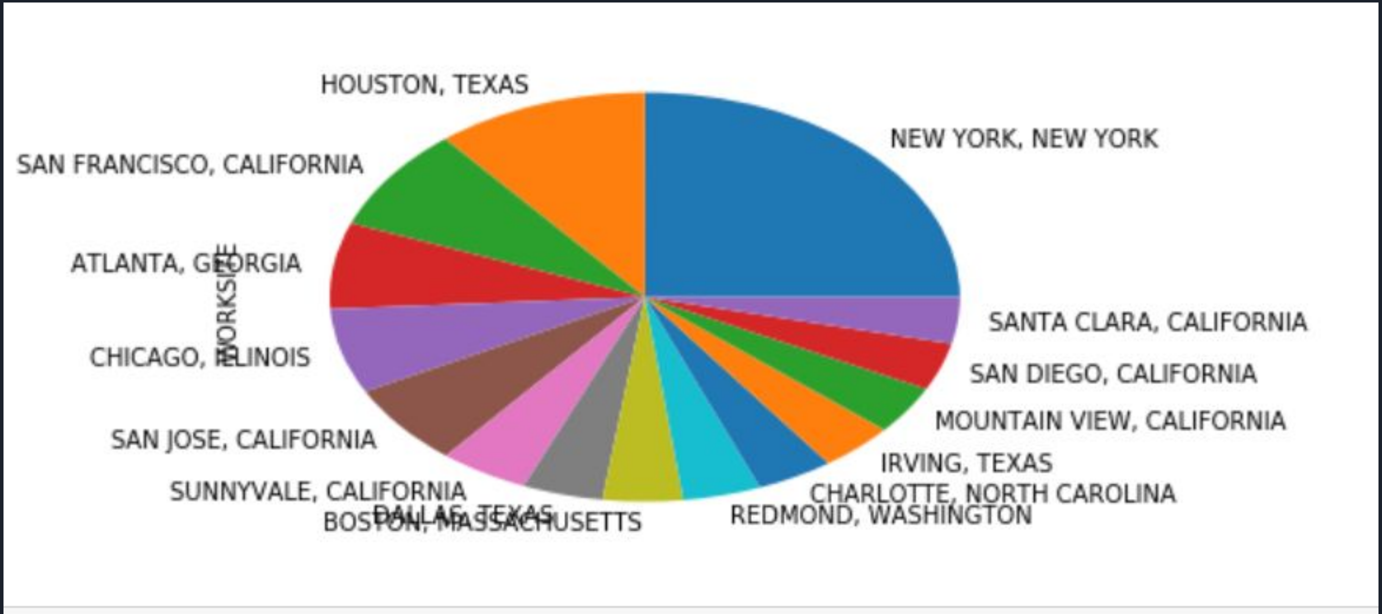
Wages and Recruiters who Give the Highest Wages



```
In [101]: f.PREVAILING_WAGE.value_counts().head(10)
          #wages are already sorted. if not we can use
          # data.PREVAILING_WAGE.value_counts().sort_valu

Out[101]: 60000.0    10185
          55245.0     6745
          62566.0     6480
          58053.0     5683
          52499.0     5492
          51730.0     5407
          63877.0     5377
          65042.0     5276
          55370.0     4961
          67808.0     4646
          Name: PREVAILING_WAGE, dtype: int64

In [102]: f.PREVAILING_WAGE.mean()

Out[102]: 145166.64888402403
```



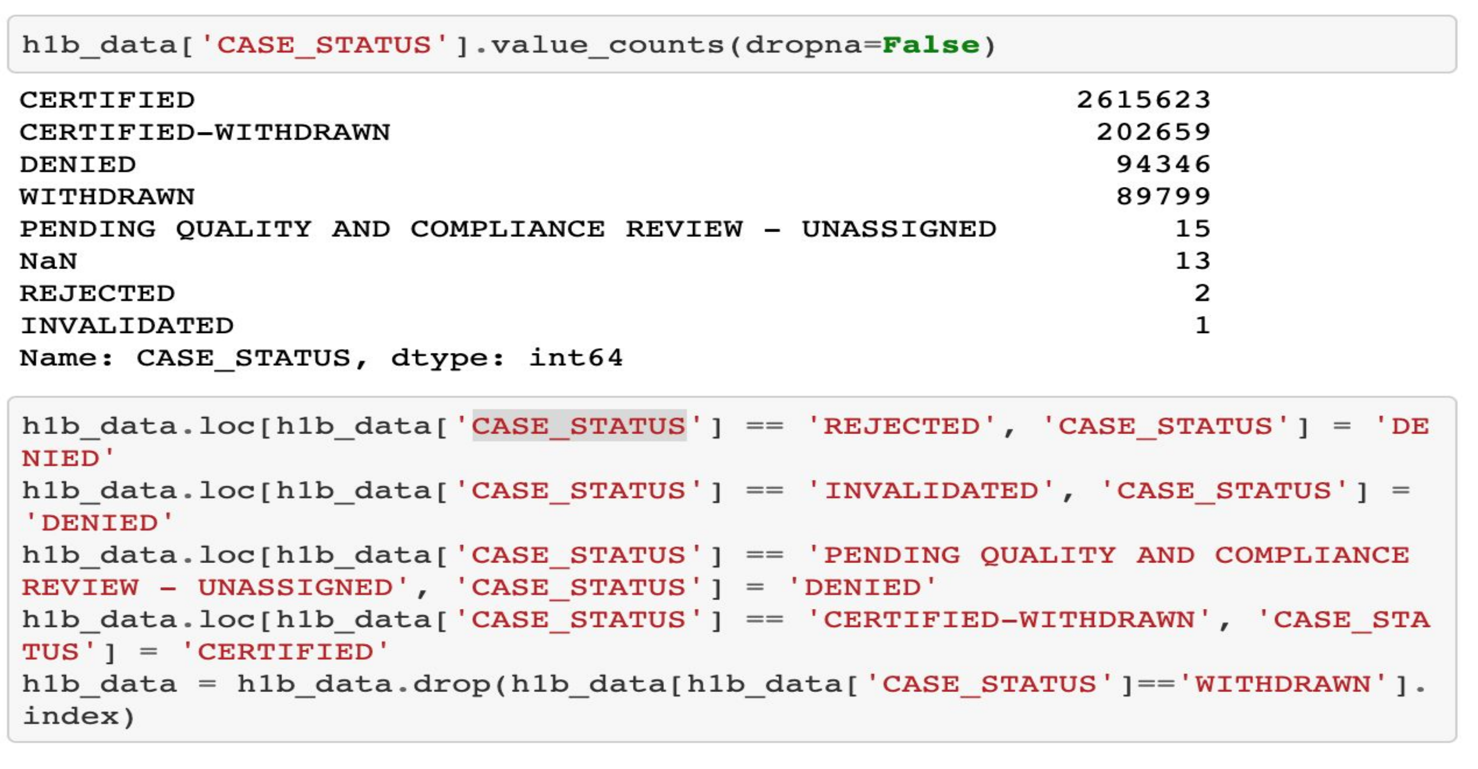
Cities with Maximum Jobs



Machine Learning

Data Pre-processing

- Transforming “CASE_STATUS” into either “Certifed” or “Deinded”



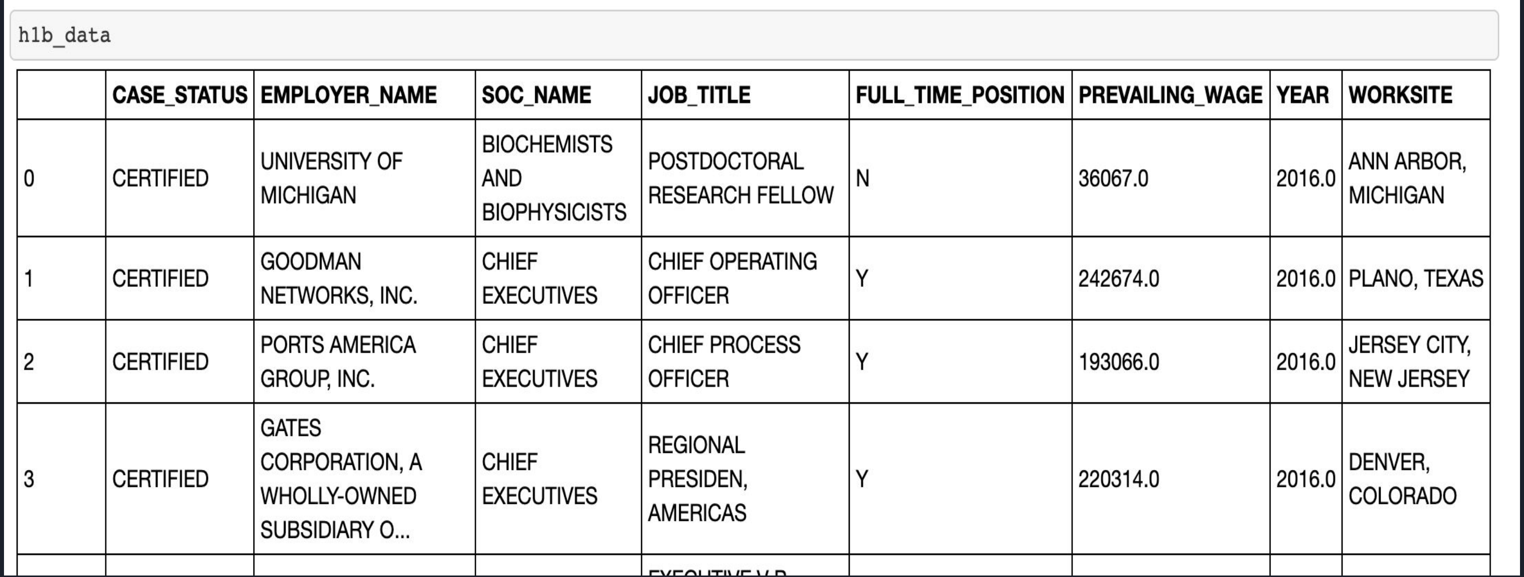
```
h1b_data['CASE_STATUS'].value_counts(dropna=False)
```

```
CERTIFIED                                             2615623
CERTIFIED-WITHDRAWN                                    202659
DENIED                                                  94346
WITHDRAWN                                               89799
PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED         15
NaN                                                        13
REJECTED                                                    2
INVALIDATED                                                 1
Name: CASE_STATUS, dtype: int64
```

```
h1b_data.loc[h1b_data['CASE_STATUS'] == 'REJECTED', 'CASE_STATUS'] = 'DE
NIED'
h1b_data.loc[h1b_data['CASE_STATUS'] == 'INVALIDATED', 'CASE_STATUS'] =
'DENIED'
h1b_data.loc[h1b_data['CASE_STATUS'] == 'PENDING QUALITY AND COMPLIANCE
REVIEW - UNASSIGNED', 'CASE_STATUS'] = 'DENIED'
h1b_data.loc[h1b_data['CASE_STATUS'] == 'CERTIFIED-WITHDRAWN', 'CASE_STA
TUS'] = 'CERTIFIED'
h1b_data = h1b_data.drop(h1b_data[h1b_data['CASE_STATUS']=='WITHDRAWN'].
index)
```

Data Pre-processing

- Transforming “EMOLOYER_NAME” into either “University” or “Company”

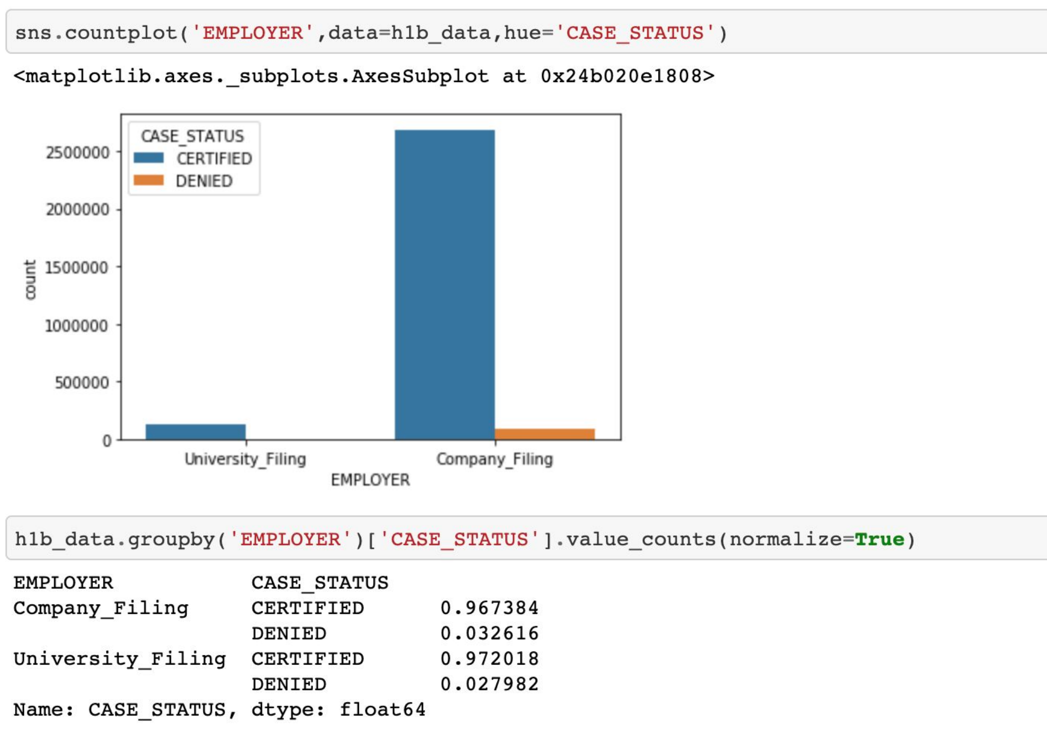


```
h1b_data
```

	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE
0	CERTIFIED	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN
1	CERTIFIED	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	PLANO, TEXAS
2	CERTIFIED	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	JERSEY CITY, NEW JERSEY
3	CERTIFIED	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O...	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO
				EXECUTIVE V.P.				

Data Pre-processing

- Transforming “EMOLOYER_NAME” into either “University” or “Company”



```
sns.countplot('EMPLOYER',data=h1b_data,hue='CASE_STATUS')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x24b020e1808>
```

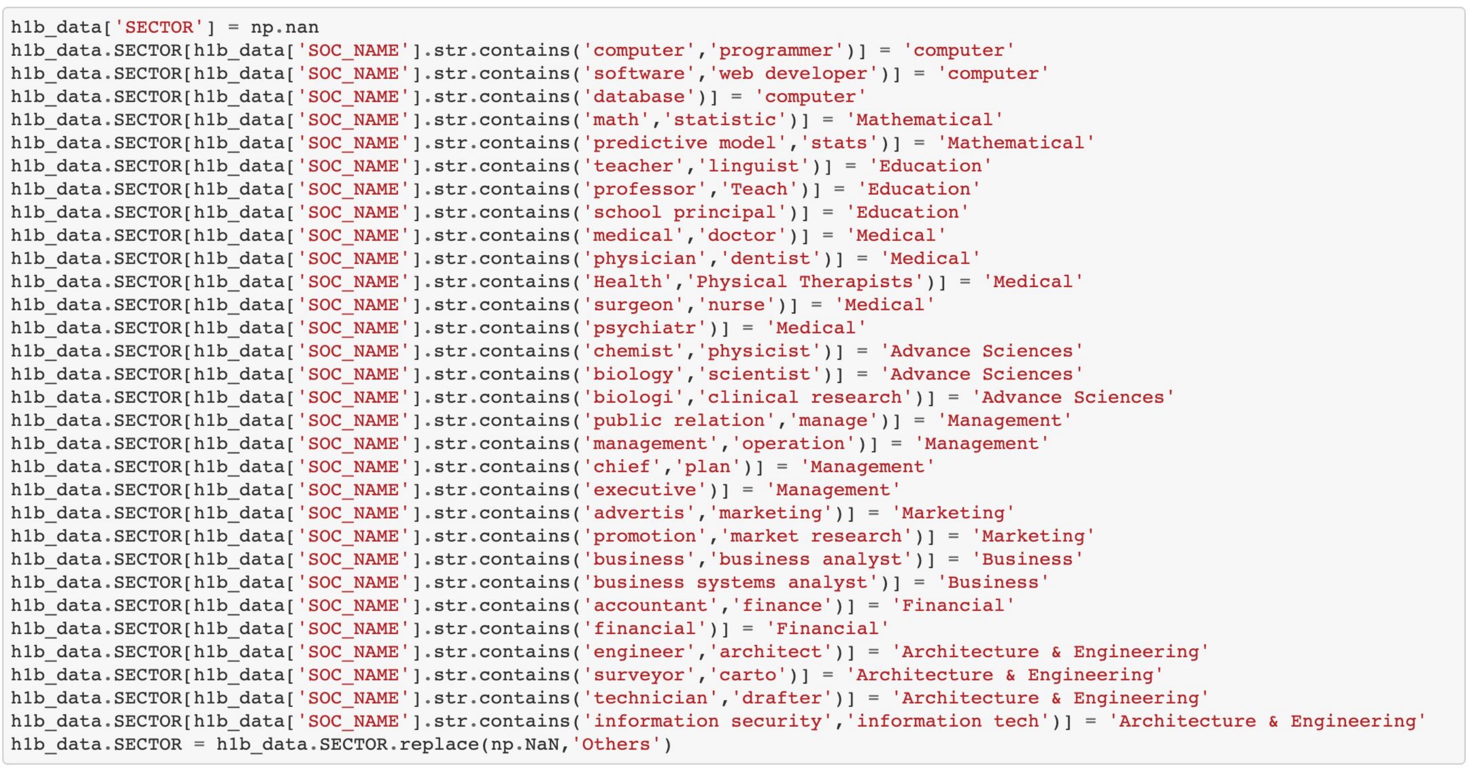
```
h1b_data.groupby('EMPLOYER')['CASE_STATUS'].value_counts(normalize=True)
```

```
EMPLOYER           CASE_STATUS
Company_Filing     CERTIFIED      0.967384
                   DENIED         0.032616
University_Filing  CERTIFIED      0.972018
                   DENIED         0.027982
Name: CASE_STATUS, dtype: float64
```

Higher chance to get H1B if a person is working for university.

Data Pre-processing

- Reorganize the occupational name



```
h1b_data['SECTOR'] = np.nan
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('computer','programmer')] = 'computer'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('software','web developer')] = 'computer'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('database')] = 'computer'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('math','statistic')] = 'Mathematical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('predictive model','stats')] = 'Mathematical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('teacher','linguist')] = 'Education'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('professor','Teach')] = 'Education'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('school principal')] = 'Education'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('medical','doctor')] = 'Medical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('physician','dentist')] = 'Medical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('Health','Physical Therapists')] = 'Medical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('surgeon','nurse')] = 'Medical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('psychiatr')] = 'Medical'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('chemist','physicist')] = 'Advance Sciences'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('biology','scientist')] = 'Advance Sciences'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('biologi','clinical research')] = 'Advance Sciences'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('public relation','manage')] = 'Management'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('management','operation')] = 'Management'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('chief','plan')] = 'Management'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('executive')] = 'Management'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('advertis','marketing')] = 'Marketing'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('promotion','market research')] = 'Marketing'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('business','business analyst')] = 'Business'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('business systems analyst')] = 'Business'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('accountant','finance')] = 'Financial'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('financial')] = 'Financial'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('engineer','architect')] = 'Architecture & Engineering'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('surveyor','carto')] = 'Architecture & Engineering'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('technician','drafter')] = 'Architecture & Engineering'
h1b_data.SECTOR[h1b_data['SOC_NAME'].str.contains('information security','information tech')] = 'Architecture & Engineering'
h1b_data.SECTOR = h1b_data.SECTOR.replace(np.NaN,'Others')
```

Model Performance and Model Comparison

- RandomForestClassifier

  Training Accuracy: 99.14%

  Testing Accuracy: 98.99%

- LogisticRegression

  Training Accuracy: 87.35%

  Test Accuracy: 87.49%

Result

- The H1b-petitions have increased from the year 2011-2016
- H1b petitions filed by the universities have more chances of getting accepted
- California, Texas, and NewYork tops the list in filling of the H1b petitions
- Majority of the people who applied for H1b were in full time roles