H1B_Project

June 17, 2020

0.1 Write Code in Empty cells don't use existing cells otherwise output will be earesed use new cells to write code

Import Related Libraries

[1]:

download data set from this link

h1b.csv

0.1.1 read csv in Pandas DataFrame

[ ]:

[2]:

0.1.2 show columns

[ ]:

[3]:

```
[3]: Index(['Unnamed: 0', 'CASE_STATUS', 'EMPLOYER_NAME', 'SOC_NAME', 'JOB_TITLE',
       'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'YEAR', 'WORKSITE', 'lon',
       'lat'],
      dtype='object')
```

0.1.3 delete 'unnamed:0' column from data set

[ ]:

[4]:

```
[4]: Index(['CASE_STATUS', 'EMPLOYER_NAME', 'SOC_NAME', 'JOB_TITLE',
       'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'YEAR', 'WORKSITE', 'lon',
       'lat'],
      dtype='object')
```

0.2 check no of rows in data frame

[ ]:

[5]:

[5]: 3002458

0.3 Drop All rows which has any NA value and show first 5 rows after this operation

[ ]:

[7]:

[7]:

	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat
0	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	ANN ARBOR, MICHIGAN	-83.743038	42.280826
1	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	PLANO, TEXAS	-96.698886	33.019843
2	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	JERSEY CITY, NEW JERSEY	-74.077642	40.728158
3	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O…	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	DENVER, COLORADO	-104.990251	39.739236
4	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016.0	ST. LOUIS, MISSOURI	-90.199404	38.627003

0.4 reset index and check how many rows you have

[ ]:

[8]:

[8]:
```
2877765
```

0.5 Write a code to find out top 15 hiring company (Employer Name)

[ ]:

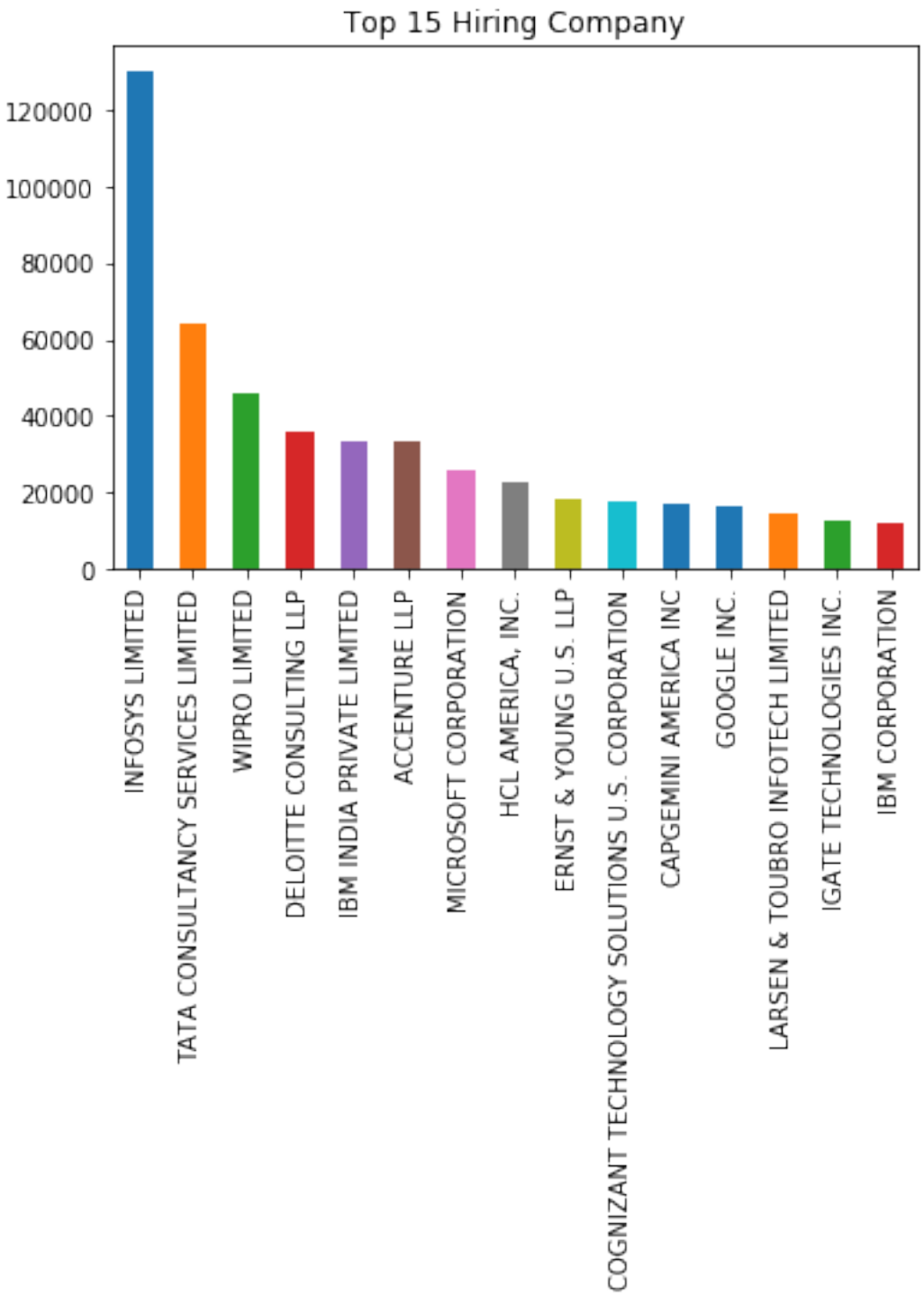
[10]:

[10]:
```
INFOSYS LIMITED                        130257
TATA CONSULTANCY SERVICES LIMITED       64273
WIPRO LIMITED                           45673
DELOITTE CONSULTING LLP                 35999
IBM INDIA PRIVATE LIMITED               33585
Name: EMPLOYER_NAME, dtype: int64
```

[ ]:

[11]:

[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23487d2fba8>
```


Top 15 Hiring Company
120000
100000
80000
60000
40000
20000
0
INFOSYS LIMITED
TATA CONSULTANCY SERVICES LIMITED
WIPRO LIMITED
DELOITTE CONSULTING LLP
IBM INDIA PRIVATE LIMITED
ACCENTURE LLP
MICROSOFT CORPORATION
HCL AMERICA, INC.
ERNST & YOUNG U.S. LLP
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION
CAPGEMINI AMERICA INC
GOOGLE INC.
LARSEN & TOUBRO INFOTECH LIMITED
IGATE TECHNOLOGIES INC.
IBM CORPORATION

0.5.1 Top 15 companies which provide highest PREVALING WAGE

[ ]:

[13]:

[13]:
```
60000.0    10185
55245.0     6745
62566.0     6480
58053.0     5683
52499.0     5492
Name: PREVAILING_WAGE, dtype: int64
```

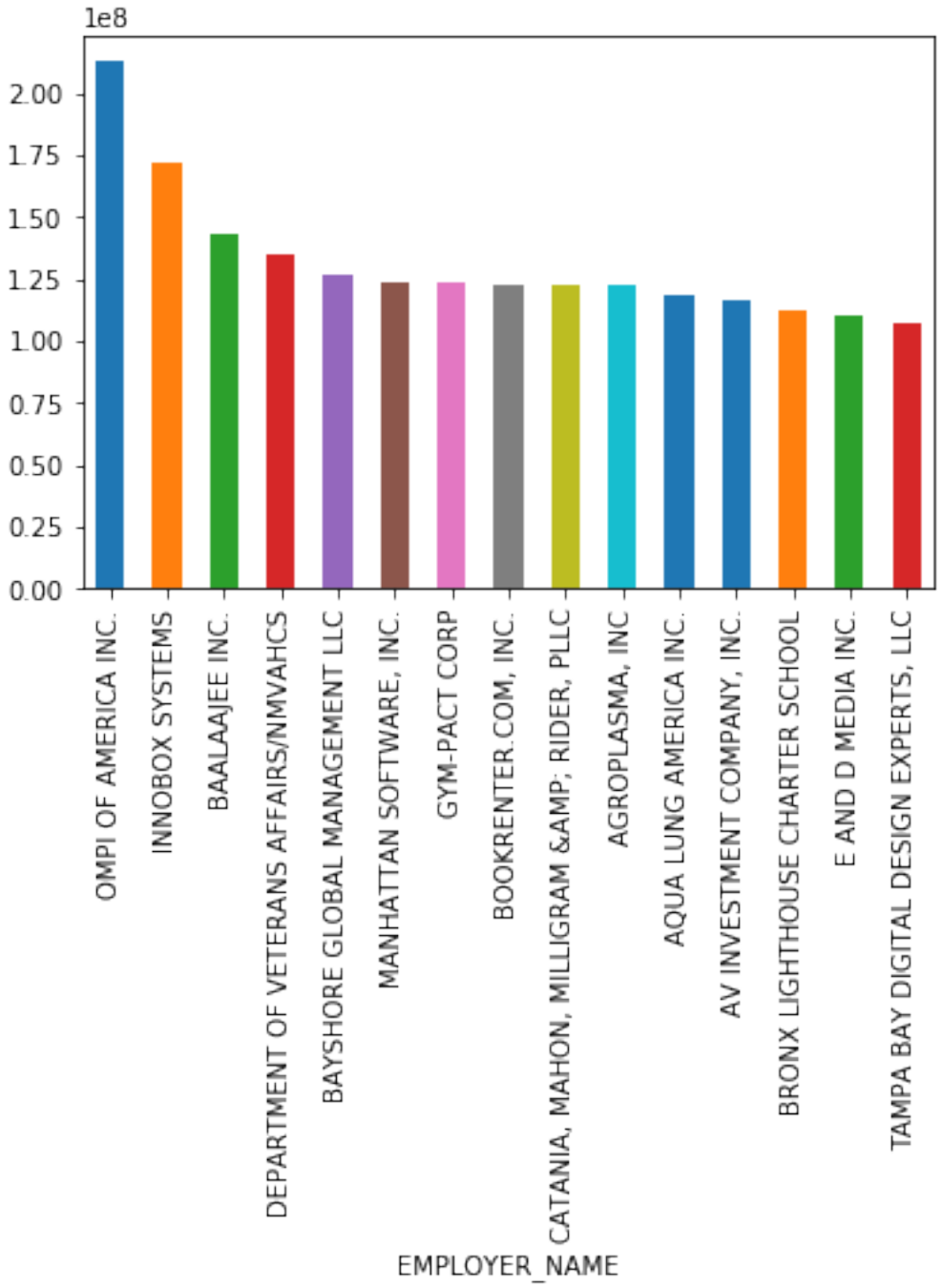
Average PREVAILING WAGE

[ ]:

[14]:

[14]:
```
145166.64888402403
```

[15]:



[16]:
```
f.columns
```

[16]:
```
Index(['CASE_STATUS', 'EMPLOYER_NAME', 'SOC_NAME', 'JOB_TITLE',
       'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'YEAR', 'WORKSITE', 'lon',
       'lat'],
      dtype='object')
```

0.5.2 Top 20 WORKSITE

[ ]:

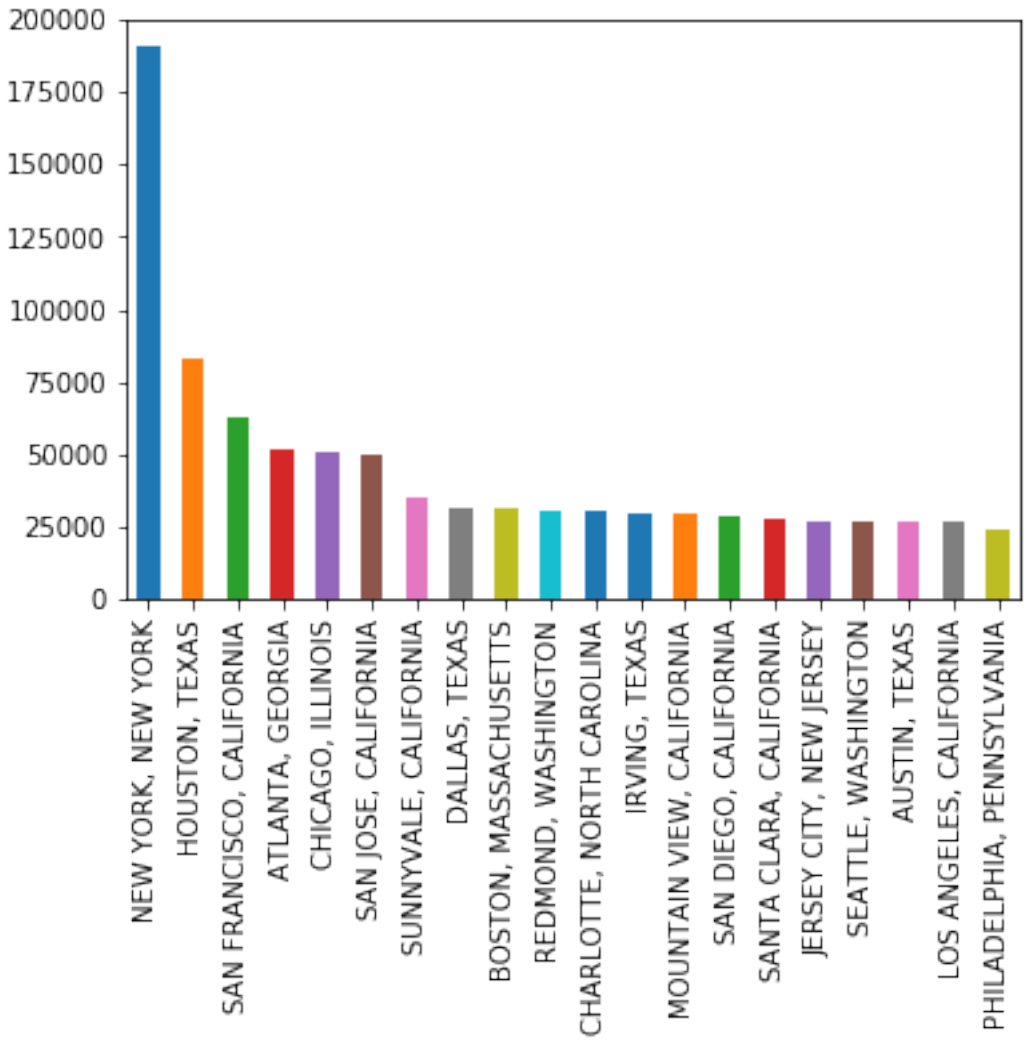
[16]:

[16]:
```
NEW YORK, NEW YORK            190863
HOUSTON, TEXAS                 83385
SAN FRANCISCO, CALIFORNIA      62457
ATLANTA, GEORGIA               52008
CHICAGO, ILLINOIS              51167
SAN JOSE, CALIFORNIA           49582
SUNNYVALE, CALIFORNIA          34968
DALLAS, TEXAS                  31509
BOSTON, MASSACHUSETTS          31336
REDMOND, WASHINGTON            30574
CHARLOTTE, NORTH CAROLINA      30176
IRVING, TEXAS                  29316
MOUNTAIN VIEW, CALIFORNIA      29245
SAN DIEGO, CALIFORNIA          28656
SANTA CLARA, CALIFORNIA        27945
JERSEY CITY, NEW JERSEY        26822
SEATTLE, WASHINGTON            26745
AUSTIN, TEXAS                  26695
LOS ANGELES, CALIFORNIA        26393
PHILADELPHIA, PENNSYLVANIA     24104
Name: WORKSITE, dtype: int64
```

[ ]:

[14]:

[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23b383f2748>
```



0.5.3 head of Worksite Column

[ ]:

[15]:

[15]:
```
0        ANN ARBOR, MICHIGAN
1              PLANO, TEXAS
2    JERSEY CITY, NEW JERSEY
3          DENVER, COLORADO
4        ST. LOUIS, MISSOURI
Name: WORKSITE, dtype: object
```

0.5.4 Show Column Names

[ ]:

[20]:

[20]:
```
Index(['CASE_STATUS', 'EMPLOYER_NAME', 'SOC_NAME', 'JOB_TITLE',
       'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'YEAR', 'WORKSITE', 'lon',
       'lat'],
      dtype='object')
```

0.5.5 Apply a Function on DataFrame to gather only State Name from Worksite

```
eg. worksite name current -  SAN FRANCISCO, CALIFORNIA
    worksite name after - CALIFORNIA
```

note: there sholud not be any space at the begning or end of worksite name

[21]:

[22]:

[22]:

	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POSITION	PREVAILING_WAGE	YEAR	WORKSITE	lon	lat
0	CERTIFIED-WITHDRAWN	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067.0	2016.0	MICHIGAN	-83.743038	42.280826
1	CERTIFIED-WITHDRAWN	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674.0	2016.0	TEXAS	-96.698886	33.019843
2	CERTIFIED-WITHDRAWN	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066.0	2016.0	NEW JERSEY	-74.077642	40.728158
3	CERTIFIED-WITHDRAWN	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY O…	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314.0	2016.0	COLORADO	-104.990251	
4	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	2016.0	MISSOURI	-90.199404	

```
3   39.739236
4   38.627003
```

note: if you view your analysis than you will find that 'MARIANA ISLANDS' worksite name is replaced with NA values

0.5.6 Replace all NA records in your Worksite Column with Value 'MARIANA ISLANDS'

[ ]:

[ ]:

0.5.7 Print out how many unique Worksites are there

[ ]:

[23]:

```
53
```

show column names

[ ]:

[24]:

[24]:
```
Index(['CASE_STATUS', 'EMPLOYER_NAME', 'SOC_NAME', 'JOB_TITLE',
       'FULL_TIME_POSITION', 'PREVAILING_WAGE', 'YEAR', 'WORKSITE', 'lon',
       'lat'],
      dtype='object')
```

Rename you column names as

```
{'EMPLOYER_NAME':'EMPLOYER','FULL_TIME_POSITION':'FULL_T','PREVAILING_WAGE':'PREV_WAGE','WORKSI
```

[25]:

Now Remove all Columns Except these columns

```
'CASE_STATUS','YEAR','STATE','SOC_NAME','JOB_TITLE','FULL_T','PREV_WAGE','EMPLOYER','LON',"LAT'
```

[ ]:

show colnames

[ ]:

[27]:

[27]:
```
Index(['CASE_STATUS', 'YEAR', 'STATE', 'SOC_NAME', 'JOB_TITLE', 'FULL_T',
       'PREV_WAGE', 'EMPLOYER', 'LON', 'LAT'],
      dtype='object')
```

Perform These Operations

Precise LON and LAT columns upto 2 decimal palaces

Convert YEAR Column into String

Convert PREV_WAGE column into Integer

[ ]:

[ ]:

show top 3 values to check above operations

[ ]:

[29]:

[29]:

	CASE_STATUS	YEAR	STATE	SOC_NAME	JOB_TITLE	FULL_T	PREV_WAGE	EMPLOYER	LON	LAT
0	CERTIFIED-WITHDRAWN	2016	MICHIGAN	BIOCHEMISTS AND BIOPHYSICISTS	POSTDOCTORAL RESEARCH FELLOW	N	36067	UNIVERSITY OF MICHIGAN	-83.74	42.28
1	CERTIFIED-WITHDRAWN	2016	TEXAS	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674	GOODMAN NETWORKS, INC.	-96.70	33.02
2	CERTIFIED-WITHDRAWN	2016	NEW JERSEY	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066	PORTS AMERICA GROUP, INC.	-74.08	40.73

[ ]:

0.5.8 show unique values of CASE_STATUS Column

[ ]:

[30]:

[30]:
```
array(['CERTIFIED-WITHDRAWN', 'WITHDRAWN', 'CERTIFIED', 'DENIED',
       'REJECTED', 'INVALIDATED',
```

```
'PENDING QUALITY AND COMPLIANCE REVIEW - UNASSIGNED'], dtype=object)
```

[ ]:

[ ]:

1 Calculate the petitions distributions by status

[ ]:

[31]:

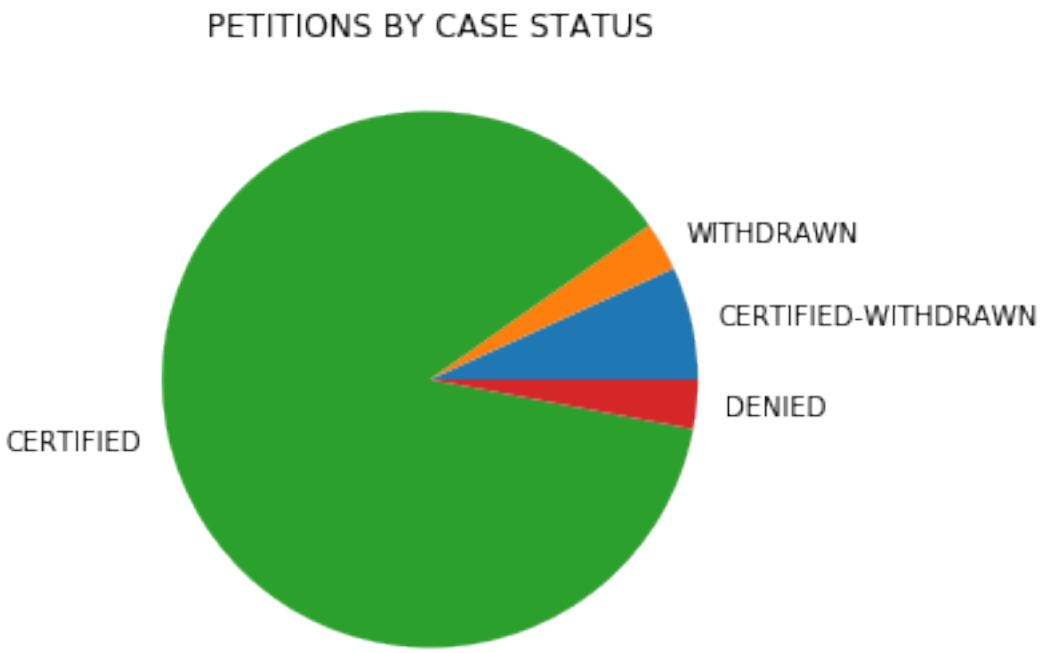
[31]: [195721, 84752, 2512114, 85161, 1, 1, 15]

[32]:
```
from matplotlib.pyplot import pie,axis,show
import matplotlib as mpl
```

PETITIONS BY CASE STATUS

[ ]:

[33]:



2 Calculating the petitions distributions by year

[ ]:

[34]:

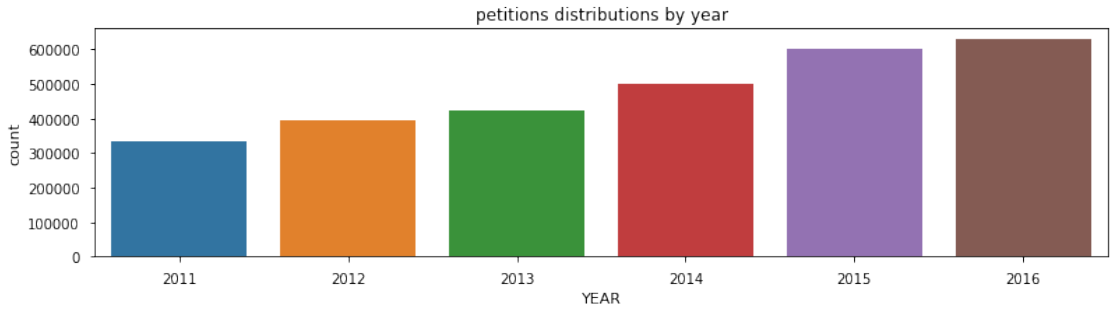
[34]: [333625, 394267, 422427, 498027, 600120, 629299]

[ ]:

[35]:

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1460:
FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

[35]: <matplotlib.axes._subplots.AxesSubplot at 0x2201782fba8>



[ ]:
```
sn.set_context("notebook",font_scale=1.0)
plt.figure(figsize=(13,3))
plt.title('petitions distributions by year')
sn.countplot(f['YEAR'])
```

[36]:
```
denied = f[f.CASE_STATUS == 'DENIED']
len(denied)
```

[36]: 85161

[37]:
```
del denied['CASE_STATUS']
```

[38]:
```
denied = denied.reset_index()
denied.head(3)
```

[38]:

	index	YEAR	STATE	SOC_NAME	\
0	39	2016	WASHINGTON	CHIEF EXECUTIVES	

```
1     47  2016  CALIFORNIA  CHIEF EXECUTIVES
2     95  2016    ILLINOIS  CHIEF EXECUTIVES

                                          JOB_TITLE FULL_T PREV_WAGE  \
0                           CHIEF EXECUTIVE OFFICER      Y    187200
1                                         PRESIDENT      Y    197683
2  PRINCIPAL (ATTORNEY) AND CHAIRMAN OF THE EXECU…      Y    226699

               EMPLOYER     LON    LAT
0      PARALLELS, INC. -122.22  47.48
1  RANCHO LA PUERTA LLC -117.16  32.72
2   BAKER & MCKENZIE PC  -87.63  41.88
```

[39]:
```
denied_year_count = [0]* 6
```

[40]:
```
for i in range(0,6):
    denied_year_count[i] = denied[denied.YEAR == years[i]]['YEAR'].count()
```

[41]:
```
denied_year_count
```

[41]: [25986, 18866, 10976, 10816, 10037, 8480]

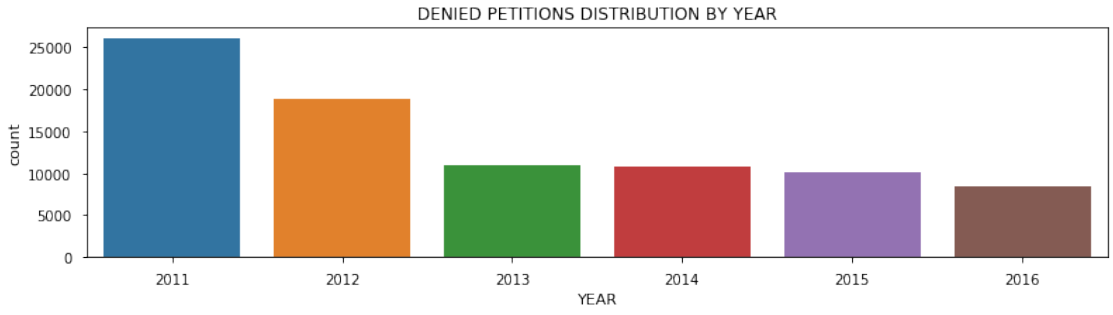
2.0.1 Denied PETITIONS DISTRIBUTION BY YEAR

[ ]:
```

```

[42]:
```

```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1460:
FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

[42]: <matplotlib.axes._subplots.AxesSubplot at 0x1fd90162748>



[ ]:
```

```

Denied % Rate By Year

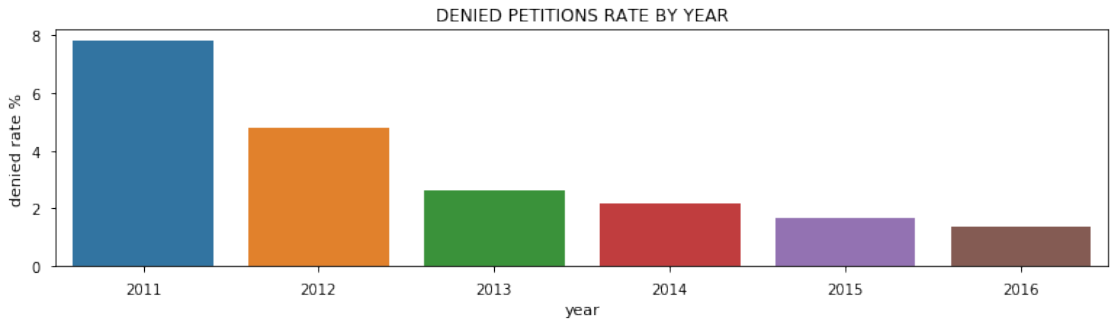
[43]:

[43]:
```
year           2011  2012  2013  2014  2015  2016
denied rate %  7.79  4.79   2.6  2.17  1.67  1.35
```

[ ]:

[44]:

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1460:
FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```



[ ]:

[ ]:

2.1 Calculate the number of petitions filed by the States

[ ]:

[45]:

[45]:
```
Index(['CASE_STATUS', 'YEAR', 'STATE', 'SOC_NAME', 'JOB_TITLE', 'FULL_T',
       'PREV_WAGE', 'EMPLOYER', 'LON', 'LAT'],
      dtype='object')
```

unique stats sorted

[ ]:

[52]:

```
['ALABAMA', 'ALASKA', 'ARIZONA', 'ARKANSAS', 'CALIFORNIA', 'COLORADO',
'CONNECTICUT', 'DELAWARE', 'DISTRICT OF COLUMBIA', 'FLORIDA', 'GEORGIA',
'HAWAII', 'IDAHO', 'ILLINOIS', 'INDIANA', 'IOWA', 'KANSAS', 'KENTUCKY',
'LOUISIANA', 'MAINE', 'MARIANA ISLANDS', 'MARYLAND', 'MASSACHUSETTS',
'MICHIGAN', 'MINNESOTA', 'MISSISSIPPI', 'MISSOURI', 'MONTANA', 'NEBRASKA',
'NEVADA', 'NEW HAMPSHIRE', 'NEW JERSEY', 'NEW MEXICO', 'NEW YORK', 'NORTH
CAROLINA', 'NORTH DAKOTA', 'OHIO', 'OKLAHOMA', 'OREGON', 'PENNSYLVANIA', 'PUERTO
RICO', 'RHODE ISLAND', 'SOUTH CAROLINA', 'SOUTH DAKOTA', 'TENNESSEE', 'TEXAS',
'UTAH', 'VERMONT', 'VIRGINIA', 'WASHINGTON', 'WEST VIRGINIA', 'WISCONSIN',
'WYOMING']
```

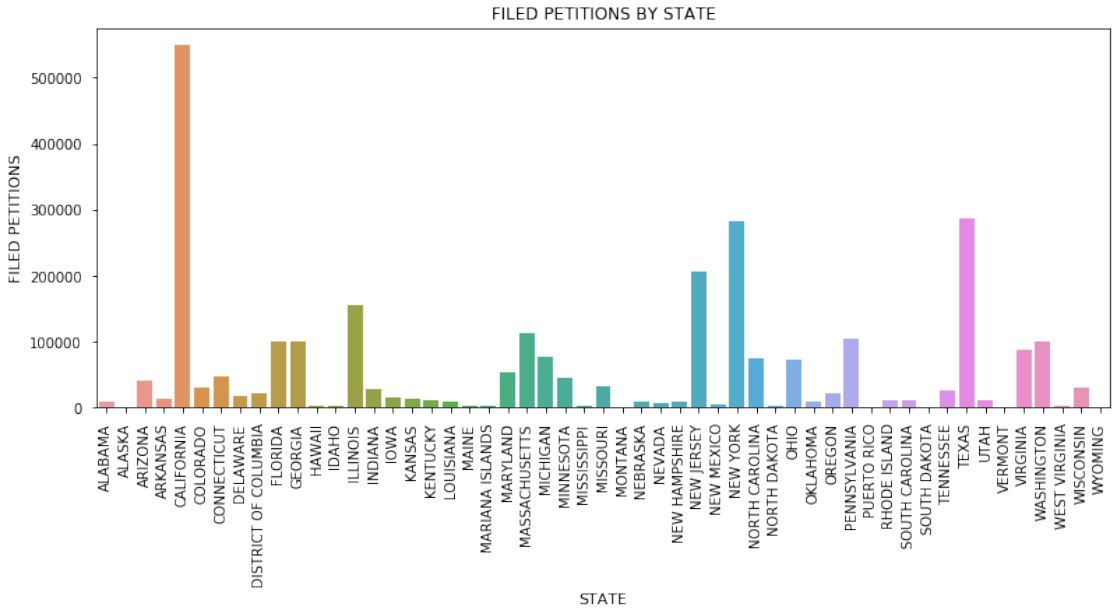
[54]:

[54]: 53

[ ]:

[60]:

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1460:
FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```



3 Number of petitions denied by the state

total denied petitions

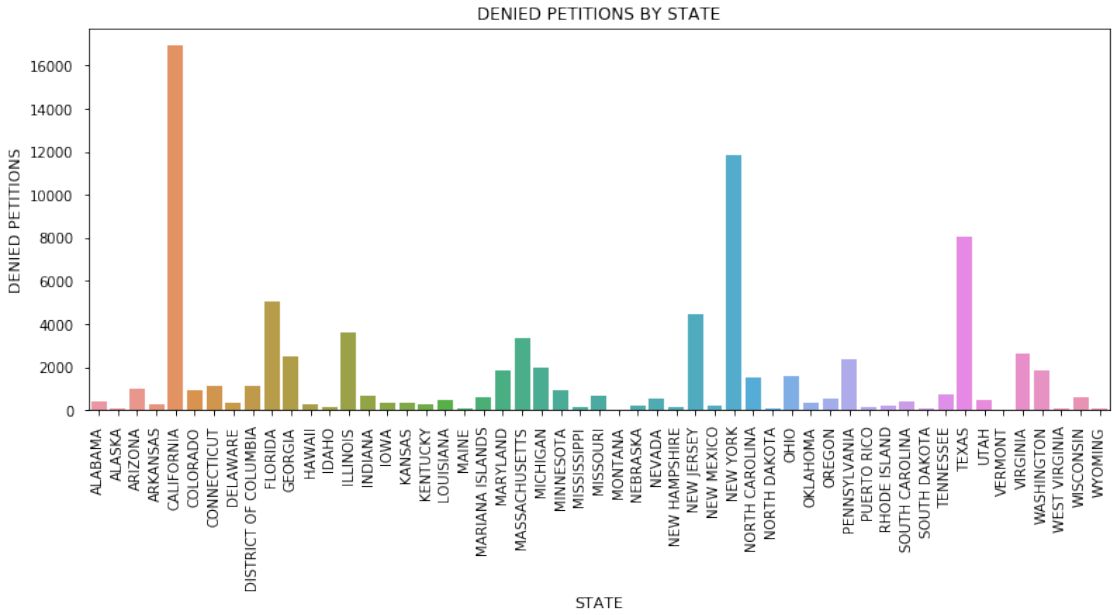
[62]:

```
85161
```

[ ]:

[66]:

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1460:
FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```

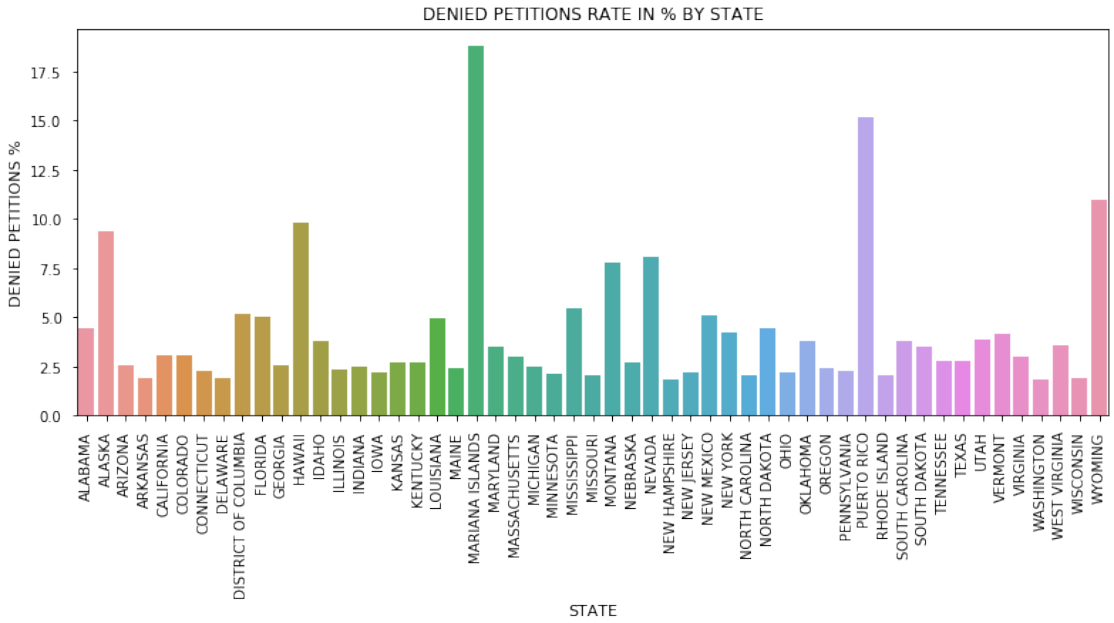


3.1 % Rate of Denied Petitions by State

[68]:

[69]:

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\categorical.py:1460:
FutureWarning: remove_na is deprecated and is a private function. Do not use.
  stat_data = remove_na(group_data)
```



Find out how many applied for Illinois State and how many how them are Denied

[ ]:

[ ]:

How Many People are Certified for Job title 'CHIEF PROCESS OFFICER' who applied for state Illinois

[ ]:

[74]:

```
2 petitions for the job of "CHIEF PROCESS OFFICER" in the state of illinois were
certified
```

[ ]:

[75]:

[75]:

```
          index CASE_STATUS  YEAR             SOC_NAME              JOB_TITLE  \
67488   1295151   CERTIFIED  2014  Management Analysts  CHIEF PROCESS OFFICER
68437   1311111   CERTIFIED  2014  Management Analysts  CHIEF PROCESS OFFICER

      FULL_T PREV_WAGE                EMPLOYER
67488      Y     67080  LITTLER MENDELSON P.C.
68437      Y     67080  LITTLER MENDELSON P.C.
```

4 Top 25 Job Titles

[ ]:

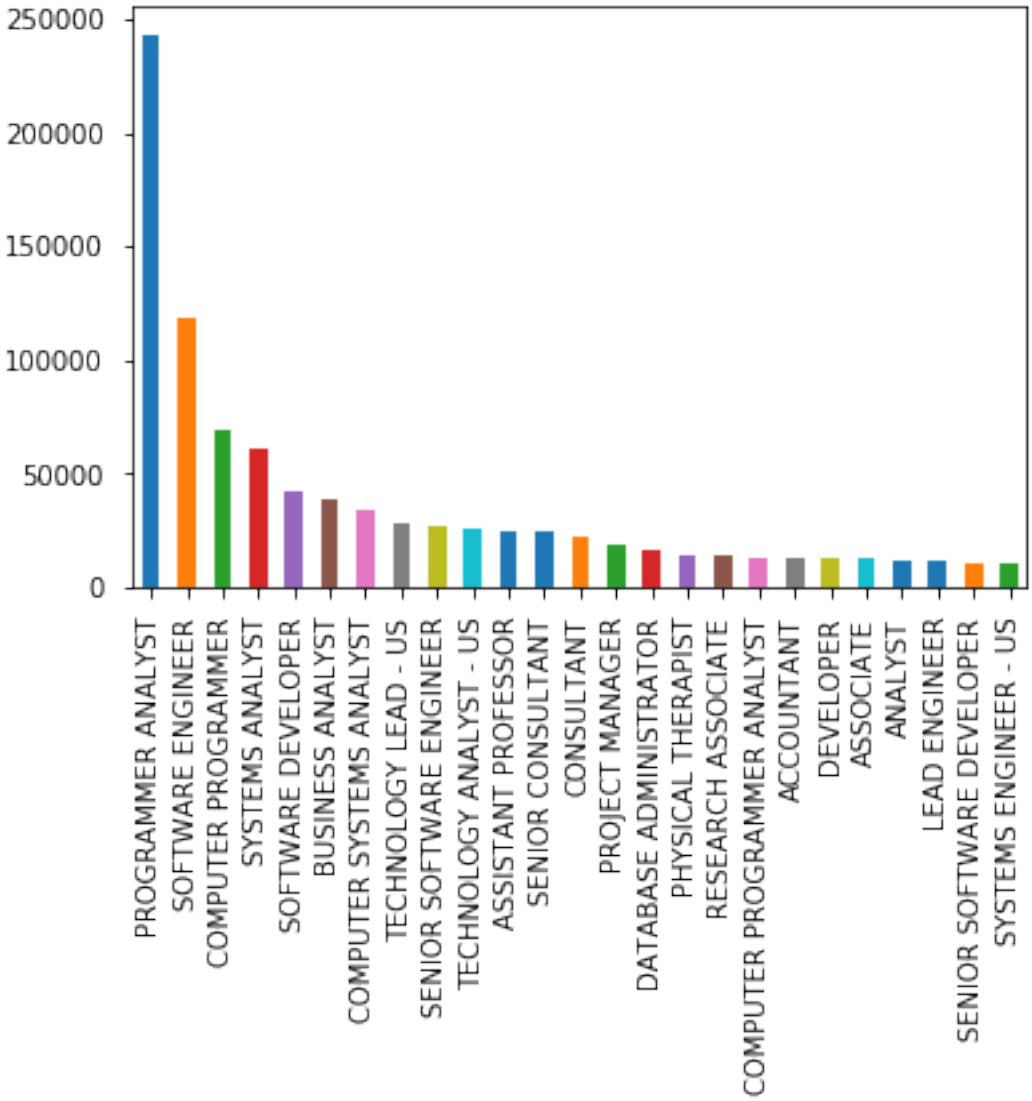
[77]:

[77]:
```
PROGRAMMER ANALYST             243357
SOFTWARE ENGINEER              118897
COMPUTER PROGRAMMER             68696
SYSTEMS ANALYST                 60754
SOFTWARE DEVELOPER              41875
BUSINESS ANALYST                38781
COMPUTER SYSTEMS ANALYST        34036
TECHNOLOGY LEAD - US            28307
SENIOR SOFTWARE ENGINEER        26617
TECHNOLOGY ANALYST - US         26010
ASSISTANT PROFESSOR             24436
SENIOR CONSULTANT               24120
CONSULTANT                      22643
PROJECT MANAGER                 19015
DATABASE ADMINISTRATOR          16108
PHYSICAL THERAPIST              14203
RESEARCH ASSOCIATE              13409
COMPUTER PROGRAMMER ANALYST     13116
ACCOUNTANT                      12934
DEVELOPER                       12737
ASSOCIATE                       12447
ANALYST                         11644
LEAD ENGINEER                   11012
SENIOR SOFTWARE DEVELOPER       10031
SYSTEMS ENGINEER - US           10020
Name: JOB_TITLE, dtype: int64
```

[ ]:

[78]:

[78]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1fd934321d0>
```



!!! Great Now Make Your Own Questions and Try to Answer Them !!