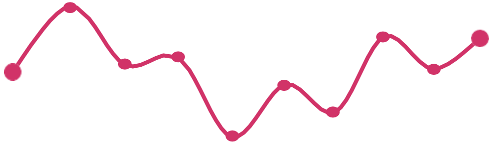
Engineering Analytics Project
H1-B Visa Petitions (2011 - 2016)

Final Report

JASWANTH BABU BANDI
U64670654
Engineering Management
jaswanthbabu@mail.usf.edu

Exploring the Data

H1-B Petitions, 2011 - 2016



Overview

The H1B visa is issued to foreign nationals who:

1. In most cases have at least a college degree.
2. Are skilled workers, have expertise in the field being hired.
3. Have on demand skills
4. Usually finished their studies in US

Requirements:

- Job Offer, Salary, Position and Bachelor’s degree

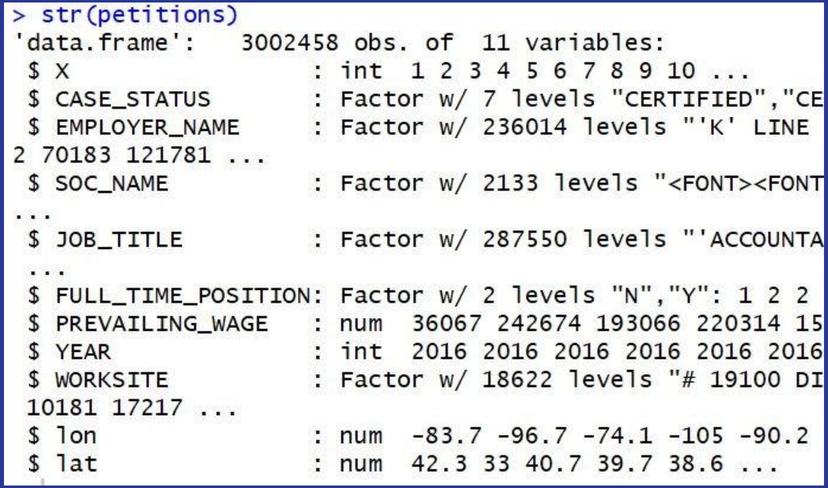
Preparing the Data

Key Variables:

a) CASE_STATUS > 6 levels (mainly CERTIFIED and DENIED)
b) EMPLOYER_NAME (for our exploration, I have considered only top employers)
c) JOB_TITLE (for our exploration, I have considered only top job titles)
d) SOC_NAME (this was eliminated by simple inspection)
e) FULL_TIME_POSITION > Y or N (binary 1 and 0)
f) PREVAILING_WAGE (key metric)

Dimensions:
> dim(petitions)
[1] 3002458   11

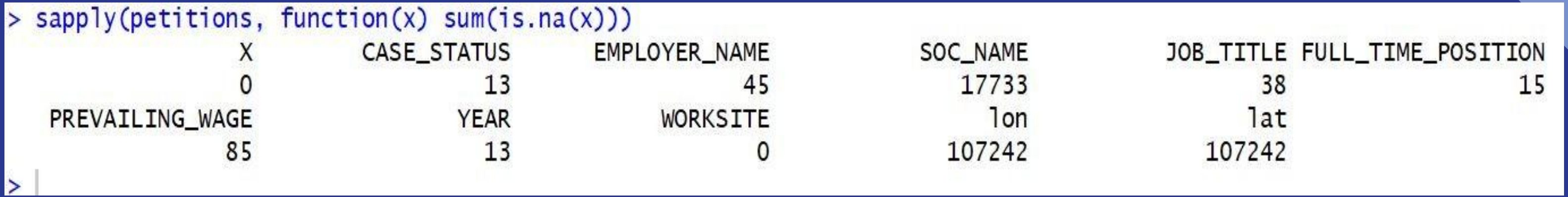
Size of the .csv file:
469.5 MB



```
> str(petitions)
'data.frame':   3002458 obs. of  11 variables:
 $ X                 : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CASE_STATUS       : Factor w/ 7 levels "CERTIFIED","CE
 $ EMPLOYER_NAME     : Factor w/ 236014 levels "'K' LINE
2 70183 121781 ...
 $ SOC_NAME          : Factor w/ 2133 levels "<FONT><FONT
...
 $ JOB_TITLE         : Factor w/ 287550 levels "'ACCOUNTA
 ...
 $ FULL_TIME_POSITION: Factor w/ 2 levels "N","Y": 1 2 2
 $ PREVAILING_WAGE   : num  36067 242674 193066 220314 15
 $ YEAR              : int  2016 2016 2016 2016 2016 2016
 $ WORKSITE          : Factor w/ 18622 levels "# 19100 DI
 10181 17217 ...
 $ lon               : num  -83.7 -96.7 -74.1 -105 -90.2
 $ lat               : num  42.3 33 40.7 39.7 38.6 ...
```

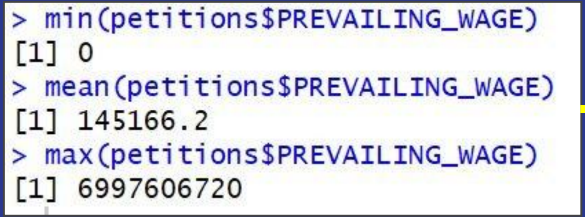
Problems

a) Missing Values

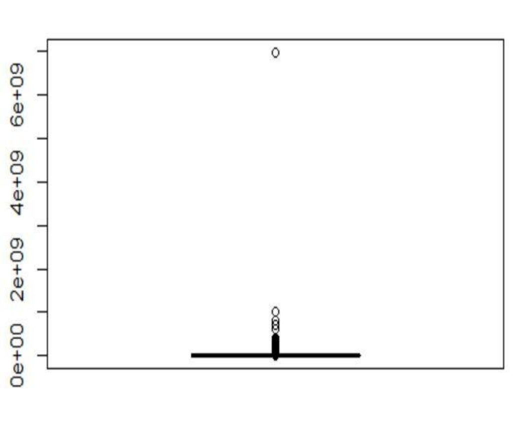


```
> sapply(petitions, function(x) sum(is.na(x)))
              X     CASE_STATUS   EMPLOYER_NAME          SOC_NAME       JOB_TITLE FULL_TIME_POSITION
              0              13              45             17733              38                 15
PREVAILING_WAGE            YEAR        WORKSITE               lon             lat
             85              13               0            107242          107242
> 
```

b) PREVAILING_WAGE



```
> min(petitions$PREVAILING_WAGE)
[1] 0
> mean(petitions$PREVAILING_WAGE)
[1] 145166.2
> max(petitions$PREVAILING_WAGE)
[1] 6997606720
```



c) CASE_STATUS

- ○ Certified = 1558571
- ○ Denied = 38858

The value for DENIED cases are low compared to CERTIFIED.

My Solution

Subsetting

Since the data was spread over 6 years, split the data into 6 parts to analyze the year-over-year effect and changes.

- 2011 => 159314 cases
- 2012 => 194803 cases
- 2013 => 225095 cases
- 2014 => 282617 cases
- 2015 => 351301 cases
- 2016 => 384299 cases

Setting Limits

From basic statistics and box plot diagram of PREVAILING_WAGE, I saw that there are too many outlying values. Since the values were too widely spread out, it was best to take an assumption.

- Minimum wage for H1B => $ 60K
- Maximum wage for H1B => $150K

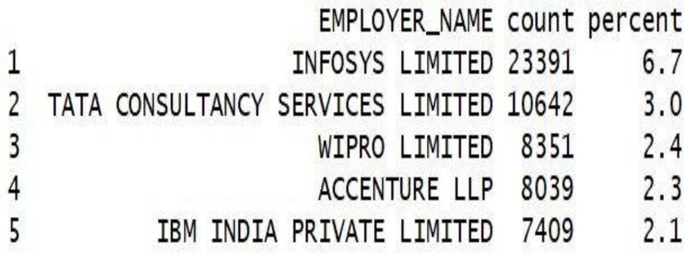
Reason: In 98% cases where the wages are above $150K, the CASE_STATUS is NOT DENIED.

Analysis - 2015 v/s2016

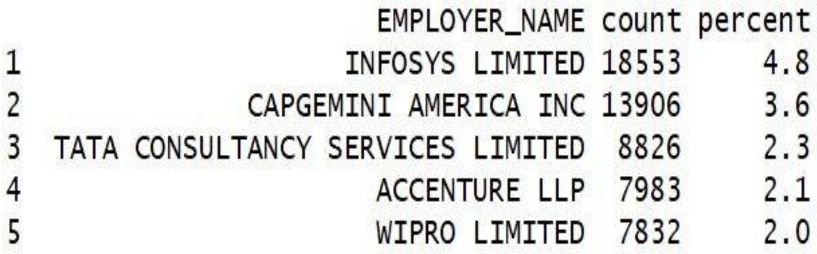
Employers & Titles in 2015

Employers & Titles in 2016

EMPLOYER

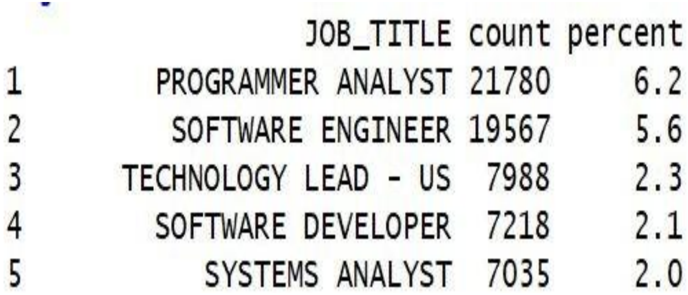


```
                    EMPLOYER_NAME count percent
1                INFOSYS LIMITED 23391     6.7
2 TATA CONSULTANCY SERVICES LIMITED 10642     3.0
3                  WIPRO LIMITED  8351     2.4
4                  ACCENTURE LLP  8039     2.3
5      IBM INDIA PRIVATE LIMITED  7409     2.1
```

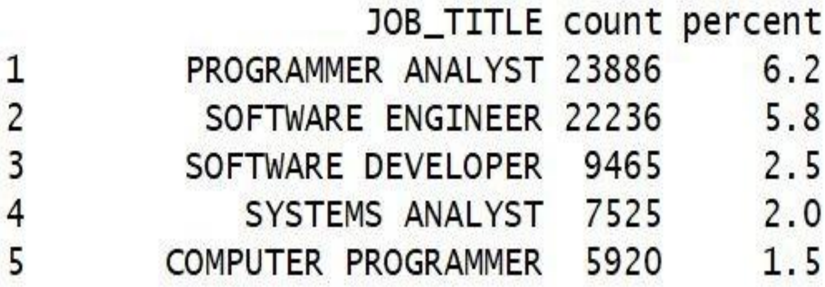


```
                      EMPLOYER_NAME count percent
1                   INFOSYS LIMITED 18553     4.8
2             CAPGEMINI AMERICA INC 13906     3.6
3 TATA CONSULTANCY SERVICES LIMITED  8826     2.3
4                     ACCENTURE LLP  7983     2.1
5                     WIPRO LIMITED  7832     2.0
```

TITLES



```
           JOB_TITLE count percent
1 PROGRAMMER ANALYST 21780     6.2
2  SOFTWARE ENGINEER 19567     5.6
3  TECHNOLOGY LEAD - US  7988     2.3
4 SOFTWARE DEVELOPER  7218     2.1
5   SYSTEMS ANALYST  7035     2.0
```



```
            JOB_TITLE count percent
1  PROGRAMMER ANALYST 23886     6.2
2   SOFTWARE ENGINEER 22236     5.8
3  SOFTWARE DEVELOPER  9465     2.5
4     SYSTEMS ANALYST  7525     2.0
5 COMPUTER PROGRAMMER  5920     1.5
```

Classification Trees

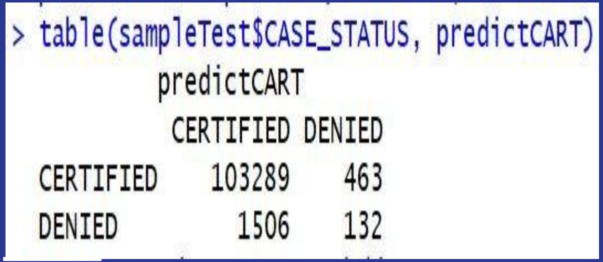
fit <- rpart(CASE_STATUS ~ EMPLOYER_NAME + FULL_TIME_POSITION + PREVAILING_WAGE , method= “class”)

Seed = 134
Data = sampleTest and sampleTest
Libraries: rpart & rpart.plot
Split is 0.7 for 2015 and 0.6 for 2016

Regressor: CASE_STATUS

The decision tree plot was complicated. Our inference was that there were too many levels for variables. This is why the image wasn’t proper.

Accuracy: 98.13%



```
> table(sampleTest$CASE_STATUS, predictCART)
           predictCART
            CERTIFIED DENIED
  CERTIFIED    103289    463
  DENIED         1506    132
```

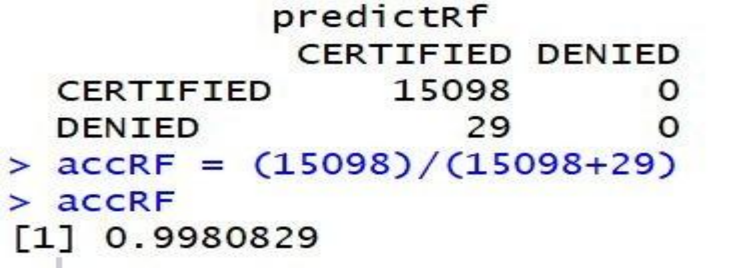
Random Forest

rf <- randomForest(CASE_STATUS ~ EMPLOYER_NAME + FULL_TIME_POSITION + PREVAILING_WAGE , data = sampleTrain, ntree = 500)

Seed = 134
Data = sampleTrain and sampleTest
Libraries: randomForest
Split is 0.7 for 2015 and 0.6 for 2016

This was run on a sample from 2015 petitions and 2016 petitions.

Accuracy: 99.8%

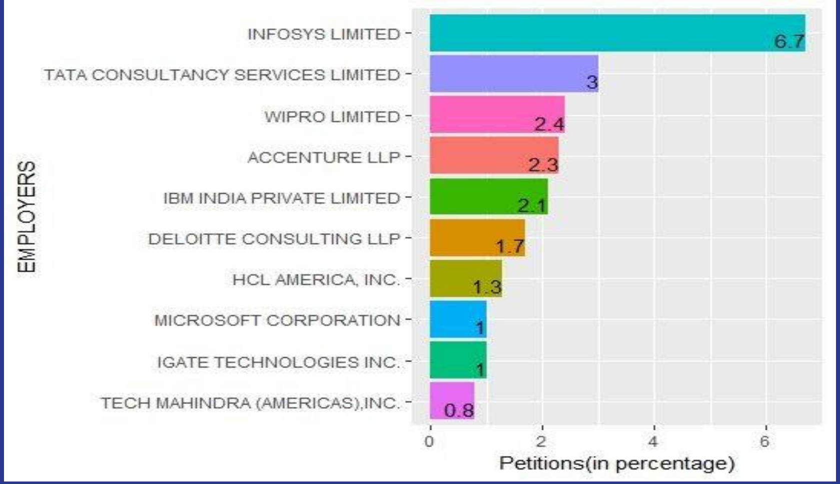


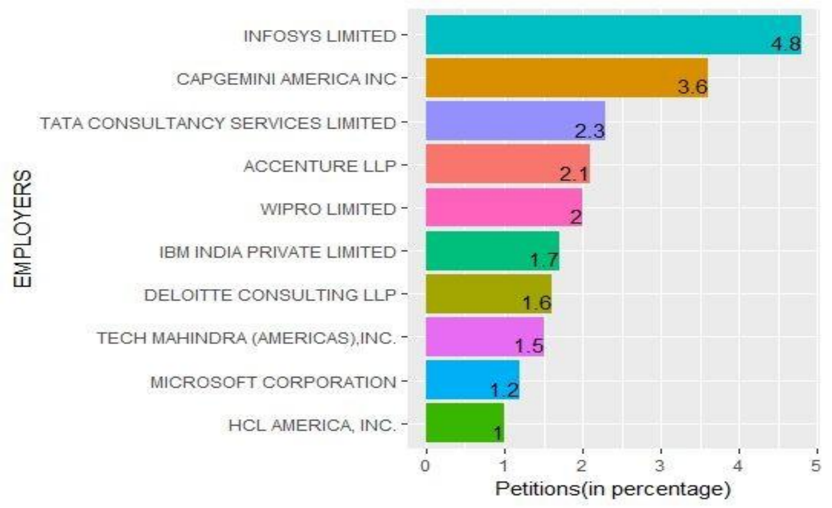
```
           predictRf
            CERTIFIED DENIED
  CERTIFIED     15098      0
  DENIED           29      0
> accRF = (15098)/(15098+29)
> accRF
[1] 0.9980829
```

Visualization

Employers in 2015 v/s 2016

- In 2016 Infosys made 6.7% of the total petitions which was 1.9% more than the previous year
- Capgemini America Inc is one company that heavily brought down their H1B visa employment
- Companies such as Microsoft, Tech Mahindra and Deloitte remained unchanged.

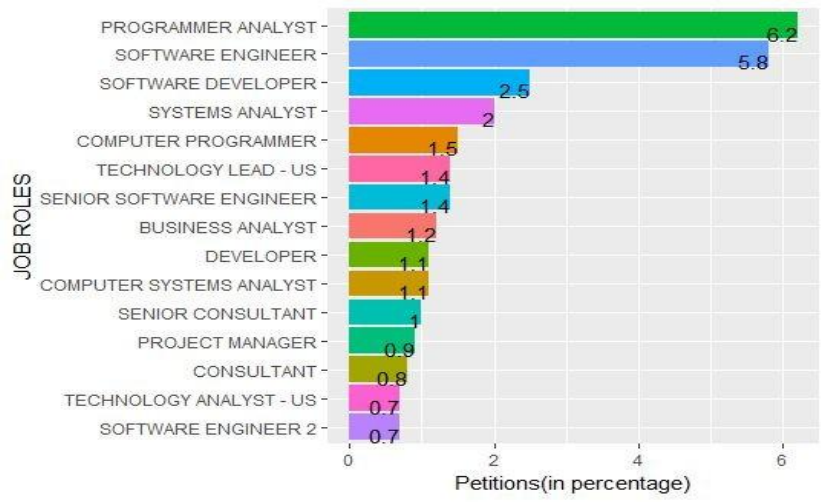


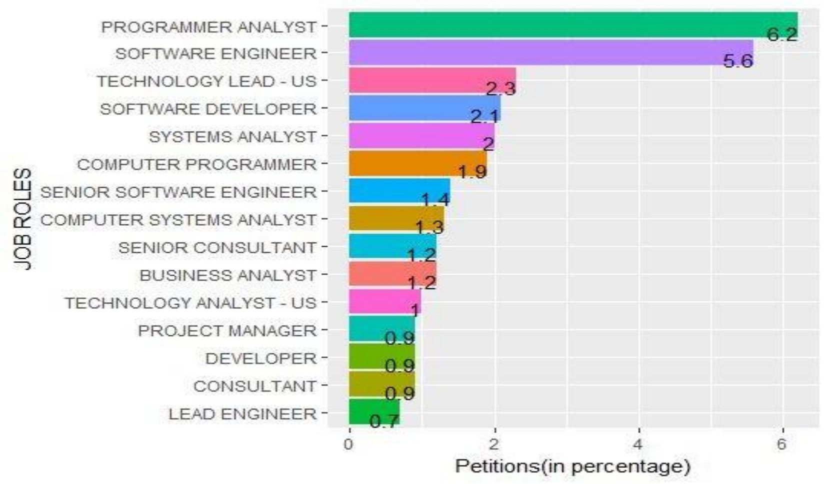


Visualization

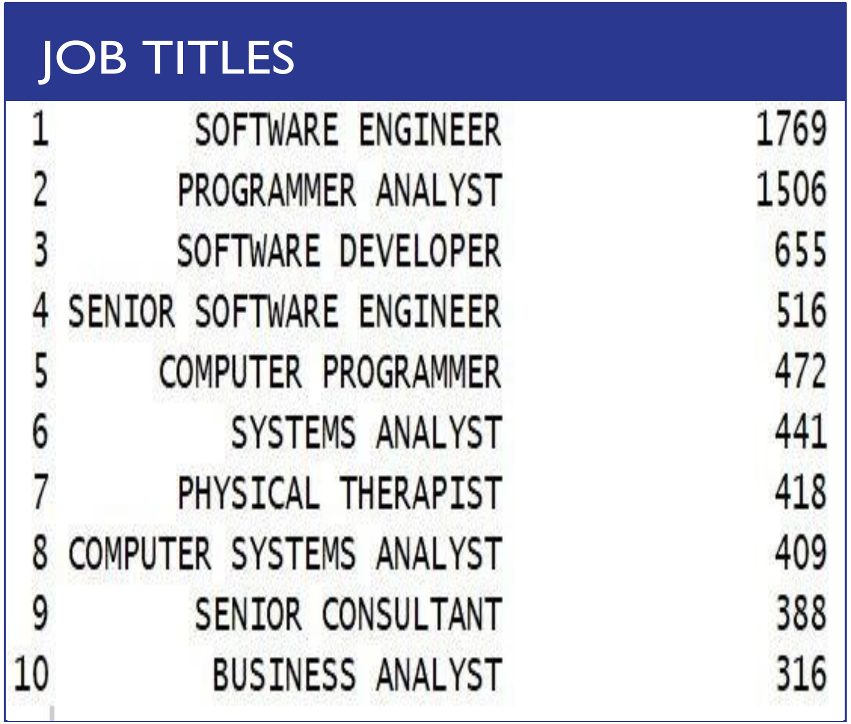
Job Titles in 2015 v/s 2016

- H1B Visa holders were primarily holding software programming positions.
- There is hardly any change in the ratio of top job titles for the petitions.
- There is a lack of managerial roles in case of H1B Visa petitioners.





CASE_STATUS as DENIED



	JOB TITLES	
1	SOFTWARE ENGINEER	1769
2	PROGRAMMER ANALYST	1506
3	SOFTWARE DEVELOPER	655
4	SENIOR SOFTWARE ENGINEER	516
5	COMPUTER PROGRAMMER	472
6	SYSTEMS ANALYST	441
7	PHYSICAL THERAPIST	418
8	COMPUTER SYSTEMS ANALYST	409
9	SENIOR CONSULTANT	388
10	BUSINESS ANALYST	316

Exploring Further

I attempted to explore the data further by splitting the WORKSITE into CITY and STATE.

This would diversify the model and allow us to study the relation of petitions with different cities and states.

We can also try to explore the change in wages for a particular position from City A to City B. (East Coast to West Coast)

Thank You :)