1. About Database

|  |
| --- |
| > h1b\_kaggle <- read.csv("C:/Users/thai/Documents/workspace-new/hu/502/project/h1b\_kaggle.csv", header=T)  > str(h1b\_kaggle)  '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  $ CASE\_STATUS : Factor w/ 7 levels "CERTIFIED","CERTIFIED-WITHDRAWN",..: 2 2 2 2 7 2 2 2 2 7 ...  $ EMPLOYER\_NAME : Factor w/ 236014 levels "'K' LINE LOGISTICS USA INC",..: 219680 86104 163491 81046 158424 33454 32942 84992 70183 121781 ...  $ SOC\_NAME : Factor w/ 2133 levels "<FONT><FONT>CARPINTEROS</FONT></FONT>",..: 177 276 276 276 276 276 276 276 276 276 ...  $ JOB\_TITLE : Factor w/ 287550 levels "'ACCOUNTANT",..: 163571 43293 43368 187285 164917 88219 43293 43336 164842 164842 ...  $ FULL\_TIME\_POSITION: Factor w/ 2 levels "N","Y": 1 2 2 2 2 2 2 2 2 2 ...  $ PREVAILING\_WAGE : num 36067 242674 193066 220314 157518 ...  $ YEAR : int 2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...  $ WORKSITE : Factor w/ 18622 levels "# 19100 DIV CD 19124 PLANO, TEXAS",..: 1006 13056 8200 4528 15700 10372 7743 14519 10181 17217 ...  $ lon : num -83.7 -96.7 -74.1 -105 -90.2 ...  $ lat : num 42.3 33 40.7 39.7 38.6 ... |

3) What are the most common Job Titles applied for by the high applicant employers?

|  |
| --- |
| library('plyr')  job\_titles <- count(h1b\_kaggle, c('JOB\_TITLE'))  job\_titles\_sorted <- job\_titles[order(-job\_titles$freq),]  job\_titles\_10 <- job\_titles\_sorted[1:10,]  barplot(job\_titles\_10$freq, ylab='freq',names.arg = job\_titles\_10$JOB\_TITLE,cex.names=0.55,las=1.5) |
| JOB\_TITLE freq  174773 PROGRAMMER ANALYST 249038  234609 SOFTWARE ENGINEER 121307  50152 COMPUTER PROGRAMMER 70570  262715 SYSTEMS ANALYST 61965  232637 SOFTWARE DEVELOPER 42907  35091 BUSINESS ANALYST 39681  51644 COMPUTER SYSTEMS ANALYST 35086  271260 TECHNOLOGY LEAD - US 28350  223988 SENIOR SOFTWARE ENGINEER 27133  270992 TECHNOLOGY ANALYST - US 26055 |
|  |

3)Top 10 employers

|  |
| --- |
| employers <- count(h1b\_kaggle, c('EMPLOYER\_NAME'))  employers\_sorted <- employers[order(-employers$freq),]  employers\_10 <- employers\_sorted[1:10,]  employers\_10 |
| EMPLOYER\_NAME freq  101654 INFOSYS LIMITED 130592  202618 TATA CONSULTANCY SERVICES LIMITED 64726  230856 WIPRO LIMITED 48117  57049 DELOITTE CONSULTING LLP 36742  98215 IBM INDIA PRIVATE LIMITED 34219  3529 ACCENTURE LLP 33447  135627 MICROSOFT CORPORATION 25576  91916 HCL AMERICA, INC. 22678  70005 ERNST & YOUNG U.S. LLP 18232  45956 COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION 17528 |
|  |

4- Top states for year 2014-2016

|  |
| --- |
| h1b <- h1b\_kaggle  h1b$state <- trimws(gsub("^.\*,", "", h1b$WORKSITE))  state\_names <- c('CALIFORNIA', 'NEW YORK', 'TEXAS', 'NEW JERSEY', 'FLORIDA', 'VIRGINIA', 'PENNSYLVANIA','OHIO','ILLINOIS','WASHINGTON')  state\_names <- state\_names[order(state\_names)]  states <- count(h1b[h1b$state %in% state\_names,], c('state','YEAR'))  states <- states[states$YEAR %in% 2011:2016,]  states\_sorted <- states[order(states$state),]  df <- data.frame(Year=integer(), 'CALIFORNIA'=integer(), 'FLORIDA'=integer(), 'ILLINOIS'=integer(), 'NEW JERSEY'=integer(), 'NEW YORK'=integer(), 'OHIO'=integer(), 'PENNSYLVANIA'=integer(),'TEXAS'=integer(),'VIRGINIA'=integer(),'WASHINGTON'=integer())  for(year in 2011:2016){  row <- states[states$YEAR==year,3]  row <- c(year, row)  newrow <- data.frame(Year=row[1],'CALIFORNIA'=row[2],'FLORIDA'=row[3],'ILLINOIS'=row[4],'NEW JERSEY'=row[5],'NEW YORK'=row[6],'OHIO'=row[7],'PENNSYLVANIA'=row[8],'TEXAS'=row[9],'VIRGINIA'=row[10],'WASHINGTON'=row[11])  df <- rbind(df, newrow)  }  matplot(df, type = c('b'), pch=1,col=2:7)  legend("topleft", legend =state\_names, col=11:2, pch=1) |
| Year CALIFORNIA FLORIDA ILLINOIS NEW.JERSEY NEW.YORK OHIO PENNSYLVANIA TEXAS VIRGINIA WASHINGTON  1 2011 65690 15227 18595 23611 41769 8582 12896 31431 11062 11359  2 2012 76402 16368 22350 27856 44512 10426 15552 37943 12423 13217  3 2013 83852 15283 24510 29794 42565 11642 16779 42216 13122 15493  4 2014 98512 17644 27407 36783 48877 13515 19150 51851 15475 17917  5 2015 115743 20401 32768 47662 55017 16066 22202 62550 19045 21665  6 2016 119741 20850 35184 48370 58670 16344 23380 68463 18901 23150 |
|  |