

Salary Study 2006

Ricky Lim

September 25, 2013

Filename: salaryStudy.Rnw

Working directory: /Users/RickyLim/Dropbox/LearnR/theArtOfR/salaryStudy

1 Data Mining

```
# download source: - http://www.flcdatacenter.com/download/Perm_FY2006_TEXT.zip zip file
# stored in input/

rawData <- unzip("input/Perm_FY2006_TEXT.zip")
all2006 <- read.csv(rawData, as.is = TRUE, header = TRUE)
head(all2006)
```

##	Case_No	Processing_Center	Final_Case_Status	Received_Date
## 1	A-05243-28497	Atlanta Processing Center	Denied	10/1/2005 0:00:00
## 2	A-05275-38245	Atlanta Processing Center	Denied	10/2/2005 0:00:00
## 3	A-05263-34450	Atlanta Processing Center	Denied	10/2/2005 0:00:00
## 4	A-05273-38122	Atlanta Processing Center	Denied	10/2/2005 0:00:00
## 5	C-05265-35535	Chicago Processing Center	Denied	10/2/2005 0:00:00
## 6	C-05275-38256	Chicago Processing Center	Denied	10/2/2005 0:00:00
##	Certified_Date	Denied_Date	Employer_Name	
## 1	10/1/2005 10:00:32		QAMAR UL ZAMAN, MD	
## 2	10/2/2005 2:48:59		HYGIA INDUSTRIES, INC.	
## 3	10/2/2005 3:49:09		TRI-SEASON LANDSCAPE & BOULDER COUNTRUCTION	
## 4	10/2/2005 9:50:15		NIPPON EXPRESS USA	
## 5	10/2/2005 12:35:50		VERNON FAIRCHILD JR.	
## 6	10/2/2005 13:51:05		VERNON FAIRCHILD JR.	
##	Employer_Address_1	Employer_Address_2	Employer_City	Employer_State
## 1	1035 RICHWOOD AVENUE		CUMBERLAND	MARYLAND
## 2	BOX 25		TALLMAN	NEW YORK
## 3	2260 SUNRISE COURT		SCOTCH PLAINS	NEW JERSEY
## 4	590 MADISON AVE.	SUITE 2401	NEW YORK	NEW YORK
## 5	4297 NORTH 1400 EAST		BUHL	IDAHO
## 6	4297 NORTH 1400 EAST		BUHL	IDAHO
##	Employer_Postal_Code	Alien_Citizenship_Country	Employer_Job_Title	
## 1	21502	PAKISTAN	Hematologist / Oncologist / Internist	
## 2	10982-0025	POLAND	COMPUTER PROGRAMMER	
## 3	7076	ECUADOR	LANDSCAPER	
## 4	10022	JAPAN	Senior Logistics Coordinator	
## 5	83316	NEW ZEALAND	Sheep Shearer-Crew Leader	
## 6	83316	NEW ZEALAND	Sheep Shearer Crew Leader	
##	Wage_Offered_From	Wage_Offered_To	Wage_Per	Prevailing_Wage_Job_Title

```
## 1      $178000.00      Year      Physician
## 2      $29.52      Year      COMPUTER PROGRAMMER
## 3      $13.35      Hour LANDSCAPING & GROUNDSKEEPING WORKERS
## 4      $62000.00      Year      Senior Logistics Coordinator
## 5      $1.70      Hour      Sheep Shearer-Crew Leader
## 6      $1.70      Hour      Sheep Shearer-Crew Leader
## Prevailing_Wage_Amount Prevailing_Wage_Level Prevailing_Wage_SOC_CODE
## 1      $163800.00      Level III      29-1062.00
## 2      $29.52      Level II      15-1021.00
## 3      $13.35      Level IV      37-3011.00
## 4      $61818.00      Level II      11-3071.01
## 5      $8.69      Level I      45-2093.00
## 6      $8.69      Level I      45-2093.00
## Prevailing_Wage_SOC_Title Prevailing_Wage_Source
## 1      Family and General Practitioners      OES
## 2      Computer Programmers      OES
## 3 Landscaping and Groundskeeping Workers      OES
## 4      Transportation Managers      OES
## 5      Farmworkers, Farm and Ranch Animals      Other
## 6      Farmworkers, Farm and Ranch Animals      Other
## Prevailing_Wage_Other_Source
## 1
## 2
## 3
## 4
## 5      Prevailing Wage Specialist
## 6      prevailig wage specialist
```

2 Data Processing

```
# cleaning
all2006 <- within(all2006, {
  Wage_Offered_From <- as.numeric(gsub("\\$", "", Wage_Offered_From))
  Prevailing_Wage_Amount <- as.numeric(gsub("\\$", "", Prevailing_Wage_Amount))
  rat <- Wage_Offered_From/Prevailing_Wage_Amount
})

head(all2006)

##      Case_No      Processing_Center Final_Case_Status      Received_Date
## 1 A-05243-28497 Atlanta Processing Center      Denied 10/1/2005 0:00:00
## 2 A-05275-38245 Atlanta Processing Center      Denied 10/2/2005 0:00:00
## 3 A-05263-34450 Atlanta Processing Center      Denied 10/2/2005 0:00:00
## 4 A-05273-38122 Atlanta Processing Center      Denied 10/2/2005 0:00:00
## 5 C-05265-35535 Chicago Processing Center      Denied 10/2/2005 0:00:00
## 6 C-05275-38256 Chicago Processing Center      Denied 10/2/2005 0:00:00
##      Certified_Date      Denied_Date      Employer_Name
## 1      10/1/2005 10:00:32      QAMAR UL ZAMAN, MD
## 2      10/2/2005 2:48:59      HYGIA INDUSTRIES, INC.
## 3      10/2/2005 3:49:09 TRI-SEASON LANDSCAPE & BOULDER COUNTRUCTION
## 4      10/2/2005 9:50:15      NIPPON EXPRESS USA
```

```

## 5          10/2/2005 12:35:50          VERNON FAIRCHILD JR.
## 6          10/2/2005 13:51:05          VERNON FAIRCHILD JR.
##      Employer_Address_1 Employer_Address_2 Employer_City Employer_State
## 1 1035 RICHWOOD AVENUE          CUMBERLAND          MARYLAND
## 2          BOX 25          TALLMAN          NEW YORK
## 3 2260 SUNRISE COURT          SCOTCH PLAINS          NEW JERSEY
## 4 590 MADISON AVE.          SUITE 2401          NEW YORK          NEW YORK
## 5 4297 NORTH 1400 EAST          BUHL          IDAHO
## 6 4297 NORTH 1400 EAST          BUHL          IDAHO
##      Employer_Postal_Code Alien_Citizenship_Country          Employer_Job_Title
## 1          21502          PAKISTAN Hematologist / Oncologist / Internist
## 2          10982-0025          POLAND          COMPUTER PROGRAMMER
## 3          7076          ECUADOR          LANDSCAPER
## 4          10022          JAPAN          Senior Logistics Coordinator
## 5          83316          NEW ZEALAND          Sheep Shearer-Crew Leader
## 6          83316          NEW ZEALAND          Sheep Shearer Crew Leader
##      Wage_Offered_From Wage_Offered_To Wage_Per          Prevailing_Wage_Job_Title
## 1          178000.00          Year          Physician
## 2          29.52          Year          COMPUTER PROGRAMMER
## 3          13.35          Hour LANDSCAPING & GROUNDSKEEPING WORKERS
## 4          62000.00          Year          Senior Logistics Coordinator
## 5          1.70          Hour          Sheep Shearer-Crew Leader
## 6          1.70          Hour          Sheep Shearer-Crew Leader
##      Prevailing_Wage_Amount Prevailing_Wage_Level Prevailing_Wage_SOC_CODE
## 1          163800.00          Level III          29-1062.00
## 2          29.52          Level II          15-1021.00
## 3          13.35          Level IV          37-3011.00
## 4          61818.00          Level II          11-3071.01
## 5          8.69          Level I          45-2093.00
## 6          8.69          Level I          45-2093.00
##      Prevailing_Wage_SOC_Title Prevailing_Wage_Source
## 1          Family and General Practitioners          OES
## 2          Computer Programmers          OES
## 3 Landscaping and Groundskeeping Workers          OES
## 4          Transportation Managers          OES
## 5          Farmworkers, Farm and Ranch Animals          Other
## 6          Farmworkers, Farm and Ranch Animals          Other
##      Prevailing_Wage_Other_Source          rat
## 1          1.0867
## 2          1.0000
## 3          1.0000
## 4          1.0029
## 5          Prevailing Wage Specialist 0.1956
## 6          prevailig wage specialist 0.1956

# filtering
all2006 <- subset(all2006, Wage_Per == "Year" & Wage_Offered_From > 20000 & Prevailing_Wage_Amount >
200)

```

3 Selection of dataframes

```

# selection based on professions
se2006 <- subset(all2006, grepl("Software Engineer", Prevailing_Wage_Job_Title))
prg2006 <- subset(all2006, grepl("Programmer", Prevailing_Wage_Job_Title))
ee2006 <- subset(all2006, grepl("Electronics Engineer", Prevailing_Wage_Job_Title))

medRat <- function(dataFrame) {
  return(median(dataFrame$rat, na.rm = TRUE))
}

makeCorp <- function(corpName) {
  return(subset(all2006, Employer_Name == corpName))
}

# selection based on corporates iteratively make an object iteratively e.g employer
# microsoft corporation as ms2006
corpList <- c("MICROSOFT CORPORATION", "ms", "INTEL CORPORATION", "intel", "SUN MICROSYSTEMS, INC.",
  "sun", "GOOGLE INC.", "google")

for (i in 1:(length(corpList)/2)) {
  corp <- corpList[2 * i - 1]
  newdtf <- paste(corpList[2 * i], "2006", sep = "")
  assign(newdtf, makeCorp(corp), pos = .GlobalEnv)
}

head(google2006)[5:10]

##      Certified_Date      Denied_Date Employer_Name      Employer_Address_1
## 6238      1/24/2006 22:19:16    GOOGLE INC. 1600 AMPHITHEATRE PARKWAY
## 6261      1/24/2006 22:26:39    GOOGLE INC. 1600 AMPHITHEATRE PARKWAY
## 6303      1/24/2006 22:26:41    GOOGLE INC. 1600 AMPHITHEATRE PARKWAY
## 6323      1/24/2006 22:26:42    GOOGLE INC. 1600 AMPHITHEATRE PARKWAY
## 6619      1/24/2006 22:26:44    GOOGLE INC. 1600 AMPHITHEATRE PARKWAY
## 6637      1/24/2006 22:26:45    GOOGLE INC. 1600 AMPHITHEATRE PARKWAY
##      Employer_Address_2 Employer_City
## 6238                      MOUNTAIN VIEW
## 6261                      MOUNTAIN VIEW
## 6303                      MOUNTAIN VIEW
## 6323                      MOUNTAIN VIEW
## 6619                      MOUNTAIN VIEW
## 6637                      MOUNTAIN VIEW

head(sun2006)[5:10]

##      Certified_Date      Denied_Date      Employer_Name Employer_Address_1
## 296      12/2/2005 13:21:20 SUN MICROSYSTEMS, INC. 7788 GATEWAY BLVD.
## 3219                      SUN MICROSYSTEMS, INC. 7788 GATEWAY BLVD.
## 7799      1/31/2006 18:06:48 SUN MICROSYSTEMS, INC. 7788 GATEWAY BLVD.
## 8822      2/3/2006 10:41:16 SUN MICROSYSTEMS, INC. 7788 GATEWAY BLVD.
## 8874      2/3/2006 15:15:54 SUN MICROSYSTEMS, INC. 7788 GATEWAY BLVD.
## 9301      2/6/2006 12:36:45 SUN MICROSYSTEMS, INC. 7788 GATEWAY BLVD.
##      Employer_Address_2 Employer_City
## 296      M/S UNWK18-303      NEWARK
## 3219      BUILDING 18      NEWARK

```

```
## 7799      BUILDING 18      NEWARK
## 8822      BUILDING 18      NEWARK
## 8874      BUILDING 18      NEWARK
## 9301      BUILDING 18      NEWARK

head(intel2006)[5:10]

##           Certified_Date      Denied_Date      Employer_Name
## 1767                12/22/2005 11:41:07 INTEL CORPORATION
## 7465                1/27/2006 15:30:23 INTEL CORPORATION
## 7478                1/30/2006 13:45:12 INTEL CORPORATION
## 12787 12/15/2005 14:54:30                INTEL CORPORATION
## 17480 1/30/2006 10:43:30                INTEL CORPORATION
## 18502                3/10/2006 15:45:16 INTEL CORPORATION
##           Employer_Address_1 Employer_Address_2 Employer_City
## 1767 2200 MISSION COLLEGE BLVD.                SANTA CLARA
## 7465 2200 MISSION COLLEGE BLVD.                SANTA CLARA
## 7478 2200 MISSION COLLEGE BLVD.                SANTA CLARA
## 12787 2200 MISSION COLLEGE BLVD.                SANTA CLARA
## 17480 2200 MISSION COLLEGE BLVD.                SANTA CLARA
## 18502 2200 MISSION COLLEGE BLVD.                SANTA CLARA

# for professional list iteratively
prof_list <- c("Software Engineer", "se", "Programmer", "prg", "Electronics Engineer", "ee")

makeProf <- function(Prof) {
  return(subset(all2006, Prevailing_Wage_Job_Title == Prof))
}

for (i in 1:(length(prof_list)/2)) {
  prof <- prof_list[2 * i - 1]
  newdtf <- paste(prof_list[2 * i], "2007", sep = "")
  assign(newdtf, makeProf(prof), pos = .GlobalEnv)
}
head(2007)

## [1] 2007
```

4 Metainfo

```
sessionInfo()

## R version 2.15.1 (2012-06-22)
## Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
```

```
## other attached packages:
## [1] codetools_0.2-8 knitr_1.1      vimcom_0.9-7    setwidth_1.0-3
##
## loaded via a namespace (and not attached):
## [1] digest_0.6.3    evaluate_0.4.3  formatR_0.7     stringr_0.6.2  tools_2.15.1
```

```
library(knitr)
knit("salaryStudy.Rnw") # compile to tex

## Error: duplicated label 'setup'

purl("salaryStudy.Rnw", documentation = 0) # extract R code only
knit2pdf("salaryStudy.Rnw")

## Error: duplicated label 'setup'
```