MA415_Midterm_Project_Jingxue_Feng.R

jingxuefeng

Wed Mar 22 16:36:16 2017

```r
library(foreign)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(stringr)
library(magrittr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```r
#read tables
osha <- read.dbf("osha.DBF")
info <- read.dbf("optinfo.DBF")
accid <- read.dbf("accid.DBF")
admpay <- read.dbf("admpay.DBF")
debt <- read.dbf("debt.DBF")
hazsub <- read.dbf("hazsub.DBF")
history <- read.dbf("history.DBF")
prog <- read.dbf("prog.DBF")
relact <- read.dbf("relact.DBF")
viol <- read.dbf("viol.DBF")
```

```
## check if a column has all NA in it, it's meaningless. We can drop it
indi = rep(0,ncol(osha))
for(i in 1: ncol(osha)){indi[i] = sum(!is.na(osha[,i]))}
which(indi==0)
```

```
## [1] 4 9
```

```
## [1] 4, 9

tidyosha = osha[,-c(which(indi==0))]
rm(indi, osha)

#glimpse
head(tidyosha)
```

```
##   CONTFLAG HISTFLAG OSHA1MOD PREVCTTYP PREVACTNO ACTIVITYNO REPORTID
## 1     <NA>        H 19840221      <NA>         0   10236776  0111100
## 2     <NA>        M 19910523      <NA>         0  103393633  0111100
## 3     <NA>        H 19880618      <NA>         0   18750034  0111400
## 4     <NA>        H 19880618      <NA>         0   18750042  0111400
## 5     <NA>        H 19880618      <NA>         0   18750059  0111400
## 6     <NA>        H 19880618      <NA>         0   18750067  0111400
##   JOBTITLE OPTREPTNO            ESTABNAME                        SITEADD
## 1        C 000000000        DUBE DRY WALL                   RT 1 MAIN ST
## 2        I 000000000 KNOWLTON MACHINE CO. NEW ENGLAND POWER, SALEM HARBO
## 3     <NA> 000000000       RENTAL & FROST                           <NA>
## 4     <NA> 000000000     PENN TRUCK LINES                           <NA>
## 5     <NA> 000000000       SILVERITE GUTT                           <NA>
## 6     <NA> 000000000         MARSSON CORP                           <NA>
##   SITESTATE HOSTESTKEY OWNERTYPE OWNERCODE ADVNOTICE OPENDATE CLOSEDATE
## 1        MA       <NA>      <NA>         0      <NA> 19831215         0
## 2        MA       <NA>         A         0         N 19900717  19900720
## 3        MA       <NA>      <NA>         0      <NA> 19790514  19790514
## 4        MA       <NA>      <NA>         0      <NA> 19790517  19790517
## 5        MA       <NA>      <NA>         0      <NA> 19790710  19790710
## 6        MA       <NA>      <NA>         0      <NA> 19790919  19790919
##   CAT_SH  NAICS NAICSEC NAICSINS INSPTYPE INSPSCOPE EMPCOUNT EMPCOVERED
## 1      S 000000  000000   000000        H         A        0          0
## 2      H 000000  000000   000000        B         B        0          0
## 3      S 000000  000000   000000        F         D        0          0
## 4      H 000000  000000   000000        F         D        0          0
## 5      H 000000  000000   000000        B         D        0          0
## 6      H 000000  000000   000000        B         D        0          0
##   NATEMPCNT WALKAROUND INTRVIEWD UNION CLOSECASE WHYNOINSP CLOSEDATE2
## 1         0       <NA>      <NA>     N         X      <NA>   19840206
## 2         0          X      <NA>     N         X      <NA>   19910522
## 3         0       <NA>      <NA>  <NA>         X         E   19880616
## 4         0       <NA>      <NA>  <NA>         X         E   19880616
## 5         0       <NA>      <NA>  <NA>         X         E   19880616
## 6         0       <NA>      <NA>  <NA>         X         E   19880616
##   SAFETYMANF SFTYCONST SFTYMARIT HELTHMANF HELTHCONST HELTHMARIT MIGRANT
## 1       <NA>         X      <NA>      <NA>       <NA>       <NA>    <NA>
## 2       <NA>      <NA>      <NA>         X       <NA>       <NA>    <NA>
## 3       <NA>      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>
## 4       <NA>      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>
## 5       <NA>      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>
## 6       <NA>      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>
##   ANTCSRVD FRSTDENY LSTREENTR LWDIRATE SHPGM DATARQD PENDUDATE FTADUDATE
## 1     <NA>        0         0        0  <NA>    <NA>  19850901         0
## 2     <NA>        0         0        0  <NA>    <NA>  19900815         0
## 3     <NA> 19790514         0        0  <NA>    <NA>         0         0
## 4     <NA> 19790517         0        0  <NA>    <NA>         0         0
## 5     <NA> 19790710         0        0  <NA>    <NA>         0         0
## 6     <NA> 19790919         0        0  <NA>    <NA>         0         0
##   DUECODE PAPREP PATRAVEL PAONSITE PATECHSUPP PARPTPREP PAOTHRCNF PALITIGN
## 1       N      0        0        0          0        40         0        0
## 2       D     40       40      100          0       180         0        0
## 3    <NA>      0        0        0          0         0         0        0
## 4    <NA>      0        0        0          0         0         0        0
## 5    <NA>      0        0        0          0         0         0        0
## 6    <NA>      0        0        0          0         0         0        0
##   PADENIAL PASUMHOURS FRSTCONTST PENREMIT FTAREMIT TOTPENLTY TOTALFTA
## 1        0         40          0      160        0       160        0
## 2        0        360          0     1820        0      1820        0
## 3        0          0          0        0        0         0        0
## 4        0          0          0        0        0         0        0
## 5        0          0          0        0        0         0        0
## 6        0          0          0        0        0         0        0
##   TOTALVIOLS TOTSERIOUS PROG_ RELACT_ OPTINFO_ DEBT_ VIOLS_ EVENT_ HAZSUB_
## 1          4          1     0       0        1     0      4      0       0
## 2          5          4     0       1        1     0      5      0       0
## 3          0          0     0       0        0     0      0      0       0
## 4          0          0     0       0        0     0      0      0       0
## 5          0          0     0       0        0     0      0      0       0
## 6          0          0     0       0        0     0      0      0       0
##   ACCID_ ADMPAY_  SIC SITEZIP SITECITY SITECNTY    DUNSNO CATSICGDE
## 1      0       1 1742   04074     1265      011 000000000      0000
## 2      0       1 3599   01970     1110      009 000000000      0000
## 3      0       0 3444   00000     0120      025 000000000      0000
## 4      0       0 4789   00000     0120      025 000000000      0000
## 5      0       0 3131   00000     0120      025 000000000      0000
## 6      0       0 2851   00000     0200      025 000000000      0000
##   CATSICINSP LSTR_DT    FRST_DT   MOD_DATE     OPENDT    CLOSEDT
## 1       0000    <NA>       <NA> 1984-02-21 1983-12-15       <NA>
## 2       0000    <NA>       <NA> 1991-05-23 1990-07-17 1990-07-20
## 3       0000    <NA> 1979-05-14 1988-06-18 1979-05-14 1979-05-14
## 4       0000    <NA> 1979-05-17 1988-06-18 1979-05-17 1979-05-17
## 5       0000    <NA> 1979-07-10 1988-06-18 1979-07-10 1979-07-10
## 6       0000    <NA> 1979-09-19 1988-06-18 1979-09-19 1979-09-19
##     CLOSEDT2    PENDUDT FTADUDT FRSTCONDT
## 1 1984-02-06 1985-09-01    <NA>      <NA>
## 2 1991-05-22 1990-08-15    <NA>      <NA>
## 3 1988-06-16       <NA>    <NA>      <NA>
## 4 1988-06-16       <NA>    <NA>      <NA>
## 5 1988-06-16       <NA>    <NA>      <NA>
## 6 1988-06-16       <NA>    <NA>      <NA>
```

```r
dim(tidyosha)
```

```
## [1] 80445    90
```

```
#[1] 80445    90

##check column head
colnames(tidyosha)
```

```
##  [1] "CONTFLAG"   "HISTFLAG"   "OSHA1MOD"   "PREVCTTYP"  "PREVACTNO"
##  [6] "ACTIVITYNO" "REPORTID"   "JOBTITLE"   "OPTREPTNO"  "ESTABNAME"
## [11] "SITEADD"    "SITESTATE"  "HOSTESTKEY" "OWNERTYPE"  "OWNERCODE"
## [16] "ADVNOTICE"  "OPENDATE"   "CLOSEDATE"  "CAT_SH"     "NAICS"
## [21] "NAICSEC"    "NAICSINS"   "INSPTYPE"   "INSPSCOPE"  "EMPCOUNT"
## [26] "EMPCOVERED" "NATEMPCNT"  "WALKAROUND" "INTRVIEWD"  "UNION"
## [31] "CLOSECASE"  "WHYNOINSP"  "CLOSEDATE2" "SAFETYMANF" "SFTYCONST"
## [36] "SFTYMARIT"  "HELTHMANF"  "HELTHCONST" "HELTHMARIT" "MIGRANT"
## [41] "ANTCSRVD"   "FRSTDENY"   "LSTREENTR"  "LWDIRATE"   "SHPGM"
## [46] "DATARQD"    "PENDUDATE"  "FTADUDATE"  "DUECODE"    "PAPREP"
## [51] "PATRAVEL"   "PAONSITE"   "PATECHSUPP" "PARPTPREP"  "PAOTHRCNF"
## [56] "PALITIGN"   "PADENIAL"   "PASUMHOURS" "FRSTCONTST" "PENREMIT"
## [61] "FTAREMIT"   "TOTPENLTY"  "TOTALFTA"   "TOTALVIOLS" "TOTSERIOUS"
## [66] "PROG_"      "RELACT_"    "OPTINFO_"   "DEBT_"      "VIOLS_"
## [71] "EVENT_"     "HAZSUB_"    "ACCID_"     "ADMPAY_"    "SIC"
## [76] "SITEZIP"    "SITECITY"   "SITECNTY"   "DUNSNO"     "CATSICGDE"
## [81] "CATSICINSP" "LSTR_DT"    "FRST_DT"    "MOD_DATE"   "OPENDT"
## [86] "CLOSEDT"    "CLOSEDT2"   "PENDUDT"    "FTADUDT"    "FRSTCONDT"
```

```
##colmns all are variable names

#check meaning of some columns in osha
#Since we are working on the most dangerous places in MA, delete state colomn
#We can also delete column sitezip
drop <- c("SITESTATE", "SITEZIP")
tidyosha = tidyosha[,!(names(tidyosha) %in% drop)]




#check level columns which that most of the observations are NA or 0
head(tidyosha)
```

```
##   CONTFLAG HISTFLAG OSHA1MOD PREVCTTYP PREVACTNO ACTIVITYNO REPORTID
## 1     <NA>        H 19840221      <NA>         0   10236776  0111100
## 2     <NA>        M 19910523      <NA>         0  103393633  0111100
## 3     <NA>        H 19880618      <NA>         0   18750034  0111400
## 4     <NA>        H 19880618      <NA>         0   18750042  0111400
## 5     <NA>        H 19880618      <NA>         0   18750059  0111400
## 6     <NA>        H 19880618      <NA>         0   18750067  0111400
##   JOBTITLE OPTREPTNO            ESTABNAME                        SITEADD
## 1        C 000000000        DUBE DRY WALL                   RT 1 MAIN ST
## 2        I 000000000 KNOWLTON MACHINE CO. NEW ENGLAND POWER, SALEM HARBO
## 3     <NA> 000000000       RENTAL & FROST                           <NA>
## 4     <NA> 000000000     PENN TRUCK LINES                           <NA>
## 5     <NA> 000000000       SILVERITE GUTT                           <NA>
## 6     <NA> 000000000         MARSSON CORP                           <NA>
##   HOSTESTKEY OWNERTYPE OWNERCODE ADVNOTICE OPENDATE CLOSEDATE CAT_SH
## 1       <NA>      <NA>         0      <NA> 19831215         0      S
## 2       <NA>         A         0         N 19900717  19900720      H
## 3       <NA>      <NA>         0      <NA> 19790514  19790514      S
## 4       <NA>      <NA>         0      <NA> 19790517  19790517      H
## 5       <NA>      <NA>         0      <NA> 19790710  19790710      H
## 6       <NA>      <NA>         0      <NA> 19790919  19790919      H
##    NAICS NAICSEC NAICSINS INSPTYPE INSPSCOPE EMPCOUNT EMPCOVERED NATEMPCNT
## 1 000000  000000   000000        H         A        0          0         0
## 2 000000  000000   000000        B         B        0          0         0
## 3 000000  000000   000000        F         D        0          0         0
## 4 000000  000000   000000        F         D        0          0         0
## 5 000000  000000   000000        B         D        0          0         0
## 6 000000  000000   000000        B         D        0          0         0
##   WALKAROUND INTRVIEWD UNION CLOSECASE WHYNOINSP CLOSEDATE2 SAFETYMANF
## 1       <NA>      <NA>     N         X      <NA>   19840206       <NA>
## 2          X      <NA>     N         X      <NA>   19910522       <NA>
## 3       <NA>      <NA>  <NA>         X         E   19880616       <NA>
## 4       <NA>      <NA>  <NA>         X         E   19880616       <NA>
## 5       <NA>      <NA>  <NA>         X         E   19880616       <NA>
## 6       <NA>      <NA>  <NA>         X         E   19880616       <NA>
##   SFTYCONST SFTYMARIT HELTHMANF HELTHCONST HELTHMARIT MIGRANT ANTCSRVD
## 1         X      <NA>      <NA>       <NA>       <NA>    <NA>     <NA>
## 2      <NA>      <NA>         X       <NA>       <NA>    <NA>     <NA>
## 3      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>     <NA>
## 4      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>     <NA>
## 5      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>     <NA>
## 6      <NA>      <NA>      <NA>       <NA>       <NA>    <NA>     <NA>
##   FRSTDENY LSTREENTR LWDIRATE SHPGM DATARQD PENDUDATE FTADUDATE DUECODE
## 1        0         0        0  <NA>    <NA>  19850901         0       N
## 2        0         0        0  <NA>    <NA>  19900815         0       D
## 3 19790514         0        0  <NA>    <NA>         0         0    <NA>
## 4 19790517         0        0  <NA>    <NA>         0         0    <NA>
## 5 19790710         0        0  <NA>    <NA>         0         0    <NA>
## 6 19790919         0        0  <NA>    <NA>         0         0    <NA>
##   PAPREP PATRAVEL PAONSITE PATECHSUPP PARPTPREP PAOTHRCNF PALITIGN
## 1      0        0        0          0        40         0        0
## 2     40       40      100          0       180         0        0
## 3      0        0        0          0         0         0        0
## 4      0        0        0          0         0         0        0
## 5      0        0        0          0         0         0        0
## 6      0        0        0          0         0         0        0
##   PADENIAL PASUMHOURS FRSTCONTST PENREMIT FTAREMIT TOTPENLTY TOTALFTA
## 1        0         40          0      160        0       160        0
## 2        0        360          0     1820        0      1820        0
## 3        0          0          0        0        0         0        0
## 4        0          0          0        0        0         0        0
## 5        0          0          0        0        0         0        0
## 6        0          0          0        0        0         0        0
##   TOTALVIOLS TOTSERIOUS PROG_ RELACT_ OPTINFO_ DEBT_ VIOLS_ EVENT_ HAZSUB_
## 1          4          1     0       0        1     0      4      0       0
## 2          5          4     0       1        1     0      5      0       0
## 3          0          0     0       0        0     0      0      0       0
## 4          0          0     0       0        0     0      0      0       0
## 5          0          0     0       0        0     0      0      0       0
## 6          0          0     0       0        0     0      0      0       0
##   ACCID_ ADMPAY_  SIC SITECITY SITECNTY    DUNSNO CATSICGDE CATSICINSP
## 1       0        1 1742      1265      011  000000000      0000        0000
## 2       0        1 3599      1110      009  000000000      0000        0000
## 3       0        0 3444      0120      025  000000000      0000        0000
## 4       0        0 4789      0120      025  000000000      0000        0000
## 5       0        0 3131      0120      025  000000000      0000        0000
## 6       0        0 2851      0200      025  000000000      0000        0000
##   LSTR_DT    FRST_DT   MOD_DATE     OPENDT    CLOSEDT   CLOSEDT2
## 1    <NA>        <NA> 1984-02-21 1983-12-15       <NA> 1984-02-06
## 2    <NA>        <NA> 1991-05-23 1990-07-17 1990-07-20 1991-05-22
## 3    <NA> 1979-05-14 1988-06-18 1979-05-14 1979-05-14 1988-06-16
## 4    <NA> 1979-05-17 1988-06-18 1979-05-17 1979-05-17 1988-06-16
## 5    <NA> 1979-07-10 1988-06-18 1979-07-10 1979-07-10 1988-06-16
## 6    <NA> 1979-09-19 1988-06-18 1979-09-19 1979-09-19 1988-06-16
##      PENDUDT FTADUDT FRSTCONDT
## 1 1985-09-01    <NA>      <NA>
## 2 1990-08-15    <NA>      <NA>
## 3       <NA>    <NA>      <NA>
## 4       <NA>    <NA>      <NA>
## 5       <NA>    <NA>      <NA>
## 6       <NA>    <NA>      <NA>
```

```
table(tidyosha$CONTFLAG)#meaningless since 80445 observation only two sample has value in this column
```

```
## 
## 1 9 
## 1 1
```

```
#1 9
#1 1

table(tidyosha$OWNERCODE) #since 80223 oberservation are 0, delete
```

```
## 
##     0    90  1105  1120  1200  1207  1208  1350  1424  1500  1502  1504 
## 80223     1     1     1     1     4     4     1     1     1     2     1 
##  1600  1701  2000  2002  2142  2200  2520  2531  2550  3011  3100  3900 
##     3     1     1     2     4     1     3     1     4     1    26    20 
##  4500  4800  5100  6002  7107  8501  9118  9141  9301  9631 
##    42     1    62     4    17     1     1     4     3     2
```

```
table(tidyosha$OPTREPTNO) #meaningless
```

```
## 
## 000000000 
##     80445
```

```
table(tidyosha$CATSICGDE) #meaningless since 78727 are 0
```

```
## 
##  0000  0111  0171  0175  0783  1429  1522  1531  1541  1542  1611  1622 
## 78727     1     3     2     2     1    17     1    16    69     9    27 
##  1623  1629  1711  1721  1731  1741  1742  1743  1751  1752  1761  1771 
##    59     9    47     8    44    21    19     2    17     3    22    11 
##  1781  1791  1793  1794  1795  1796  1799  2011  2013  2015  2024  2026 
##     2    29     3    19     5     2    42     7     5     2     3     3 
##  2032  2033  2034  2035  2037  2038  2041  2043  2051  2052  2061  2062 
##     2     1     3     1     1    10     1     1    18     3     1     1
##   2064  2074  2077  2079  2082  2086  2087  2091  2092  2095  2096  2097
##      7     1     3     1     2     7     2     4    29     2     1     1
##   2098  2099  2221  2231  2241  2261  2273  2295  2298  2326  2353  2369
##      2     6     1     3     7     5     1     6     1     1     5     1
##   2391  2392  2394  2396  2399  2421  2426  2431  2434  2439  2441  2451
##      3     8     1     6     4     2     1    15     6     1     3     1
##   2491  2499  2511  2515  2521  2541  2542  2599  2621  2631  2652  2653
##      2    18     3     9     4    10     5     7     3     1     2     8
##   2672  2675  2677  2678  2679  2711  2732  2752  2754  2759  2789  2796
##      4     1     2     2     7     3     2     2     1     2     8     1
##   2819  2821  2833  2851  2865  2891  2893  2899  2911  2951  3011  3021
##      1     2     1     6     1     3     1     2     1     4     4     1
##   3052  3053  3069  3083  3084  3086  3089  3111  3131  3142  3143  3172
##      1     1    16     1     2     1    51     7    11     1     1     2
##   3199  3211  3221  3229  3231  3269  3272  3273  3281  3291  3299  3312
##      2     3     1     1     1     1     7    10    10     4     1     3
##   3315  3316  3317  3321  3324  3325  3339  3341  3351  3354  3357  3363
##     13     1     1     6     1     3     2     4     4     2     5     4
##   3366  3398  3399  3412  3421  3423  3429  3431  3433  3441  3442  3443
##      4     4     6     3     2     4    11     1     5    15    13    10
##   3444  3446  3448  3449  3451  3452  3463  3469  3471  3479  3491  3493
##     36     6     1     3    10     6     2    33     7    27     3     1
##   3494  3495  3496  3497  3499  3531  3532  3535  3536  3537  3541  3544
##      1     1     6     1    22     3     2     3     1     2     2     3
##   3545  3548  3549  3552  3554  3555  3559  3561  3563  3564  3565  3567
##      1     1     3     3     1     2     2     1     2     6     2     2
##   3569  3571  3582  3585  3589  3592  3596  3599  3612  3621  3632  3633
##      9     2     2     4     7     1     1     5     2     2     1     1
##   3641  3643  3644  3645  3646  3648  3669  3671  3672  3674  3676  3677
##      3     2     1     1     2     1     1     1     2     1     1     1
##   3679  3691  3711  3713  3714  3715  3728  3731  3732  3751  3799  3812
##      3     4     3     2    17     1     1     5     7     1     3     4
##   3822  3823  3825  3827  3841  3842  3873  3911  3914  3931  3944  3949
##      2     1     1     1     3     8     1     3     3     3     1    11
##   3951  3952  3961  3993  3999  4173  4212  4215  4222  4225  4226  4231
##      1     1     1    11    22     1     2     2     2    10     1     1
##   4311  4491  4493  4512  4581  4731  4785  4911  4924  4953  5012  5013
##      2    10     4     6     2     1     1     1     1     4     1     4
##   5021  5031  5051  5084  5087  5091  5093  5094  5112  5141  5143  5146
##      1     2     3     1     1     1     6     1     1     3     2     1
##   5147  5148  5149  5181  5182  5199  5211  5231  5311  5441  5461  5712
##      3     1     7     9     5     2     5     1     3     1     2     1
##   5812  5943  5949  5999  6531  7353  7371  7372  7382  7389  7538  7542
##      4     1     1     1     1     5     1     1     1     2     1     1
##   7641  7692  7699  7999  8021  8051  8052  8059  8062  8069  8071  8331
##      1     2     1     6     1    38     2     4     1     1     1     1
##   8351  8711  8732  8741  9621
##      1     5     1     3     1
```

```r
# since we delete CASTICGDE, CATSICINSP is also meaningless
table(tidyosha$EMPCOUNT) #meaningless
```

```
## 
##     0 
## 80445
```

```
table(tidyosha$EMPCOVERED) #meaningless
```

```
##
##     0
## 80445
```

```
#delete colomn PENDUDT since it's the convert version of PENDUDATE

#remove all the meaningless columns
drop1 <- c("CONTFLAG", "OWNERCODE","OPTREPTNO","CATSICGDE","CATSICINSP","EMPCOUNT","EMPCOUNT","EMPCOVER
tidyosha = tidyosha[,!(names(tidyosha) %in% drop1)]


# some date columns have the same meaning, except for different form
# find those columns, delete the one which is not in form YYYY-MM-DD
open1 <- select(tidyosha, matches("OPEN"))
lstr1 <- select(tidyosha, matches("LSTR"))
frst1 <- select(tidyosha, matches("FRST"))
mod1 <- select(tidyosha, matches("MOD"))
close1 <- select(tidyosha, matches("CLOSE"))
pendu1 <-select(tidyosha, matches("PENDU"))
ftadu1 <- select(tidyosha, matches("FTADU"))


drop2 <- c("OPENDATE", "LSTREENTR", "FRSTDENY", "FRSTCONTST")
tidyosha = tidyosha[,!(names(tidyosha) %in% drop2)]




########  Accid
label1 <- read.dbf("lookups/acc.dbf")
if(sum(accid$SITESTATE=="MA") == dim(accid)[1]){accid %<>% select(-SITESTATE)}
dim(label1)
```

```
## [1] 153   3
```

```
sum(label1$CATEGORY=="PART-BODY")
```

```
## [1] 31
```

```
parts <- label1[(label1$CATEGORY== "PART-BODY"),]
dim(parts)
```

```
## [1] 31  3
```

```
parts <- select(parts, CODE, VALUE)
head(parts)
```

```
##   CODE      VALUE
## 1   01    ABDOMEN
## 2   02   ARM-MULT
## 3   03       BACK
## 4   04 BODYSYSTEM
## 5   05      CHEST
## 6   06     EAR(S)
```

```r
colnames(parts) <- c("BODYPART", "BODYPART_VALUE")
str(parts)
```

```
## 'data.frame':    31 obs. of  2 variables:
##  $ BODYPART      : Factor w/ 48 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ BODYPART_VALUE: Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 1 7 9 15 28 40 41 45 46 49 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```r
accid_1 <- left_join(accid, parts, by="BODYPART")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```r
#since we are looking for the most dangerous place to work, occupation is also important
#add a column in accid of occupation
lable2 <- read.dbf("lookups/occ.dbf")
if(sum(lable2$STATE=="MA") == dim(lable2)[1]) {lable2 %<>% select(-STATE)}
dim(lable2)
```

```
## [1] 503   2
```

```r
colnames(lable2) <- c("OCC_CODE", "OCCUPATION")
accid_clear <- left_join(accid_1, lable2, by = "OCC_CODE")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```r
# decode column NATURE

lable3 <- read.dbf("lookups/acc.dbf")
nature_inj <- lable3[(lable3$CATEGORY == "NATUR-INJ"),]
dim(nature_inj)
```

```
## [1] 22  3
```

```r
colnames(nature_inj) <- c("NA-INJ", "NATURE","NATURE_VALUE")
accid_clear <- left_join(accid_clear, nature_inj, by = "NATURE")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```r
#decode column SOURCE

souc_inj <- lable3[(lable3$CATEGORY == "SOURC-INJ"),]
colnames(souc_inj) <- c("SOURC-INJ","SOURCE","SOURCE_VALUE")
accid_clear <- left_join(accid_clear, souc_inj, by = "SOURCE")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```r
#decode column EVENT
event <- lable3[(lable3$CATEGORY == "EVENT-TYP"),]
colnames(event) <- c("EVENT-TYP","EVENT","EVENT_VALUE")
accid_clear <- left_join(accid_clear, event, by = "EVENT")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```r
#decode ENVIRON
environ <- lable3[(lable3$CATEGORY == "ENVIR-FAC"),]
```

```
colnames(environ) <- c("ENVIR-FAC","ENVIRON","ENVIRON_VALUE")
accid_clear <- left_join(accid_clear, environ, by = "ENVIRON")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
#decode HAZSUB
haz <- read.dbf("lookups/hzs.dbf")
colnames(haz) <- c("HAZSUB","HAZSUB_VALUE")
accid_clear <- left_join(accid_clear, haz, by = "HAZSUB")
```

```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
#### delete meaningless columns
drop <- c("SOURC-INJ","SOURCE","NA-INJ", "NATURE","OCC_CODE","BODYPART","EVENT-TYP","EVENT","ENVIR-FAC"
accid_clear = accid_clear[,!(names(accid_clear) %in% drop)]
# done with accid form


#### combine accid and osha
tidyosha <- left_join(accid_clear, tidyosha, by = "ACTIVITYNO")

#delete ACCID_
tidyosha = tidyosha[,!(names(tidyosha) %in% c("ACCID_"))]

## find duplicated data
library(data.table)
```

```
## --------------------------------------------------------------------------
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
## --------------------------------------------------------------------------
##
## Attaching package: 'data.table'
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
b <- colnames(tidyosha[1:ncol(tidyosha)])

a <- data.table(tidyosha, key= b)
dupli_rows <- a[unique(a[duplicated(a)]),which = T]
length(dupli_rows)
```

```
## [1] 297
```

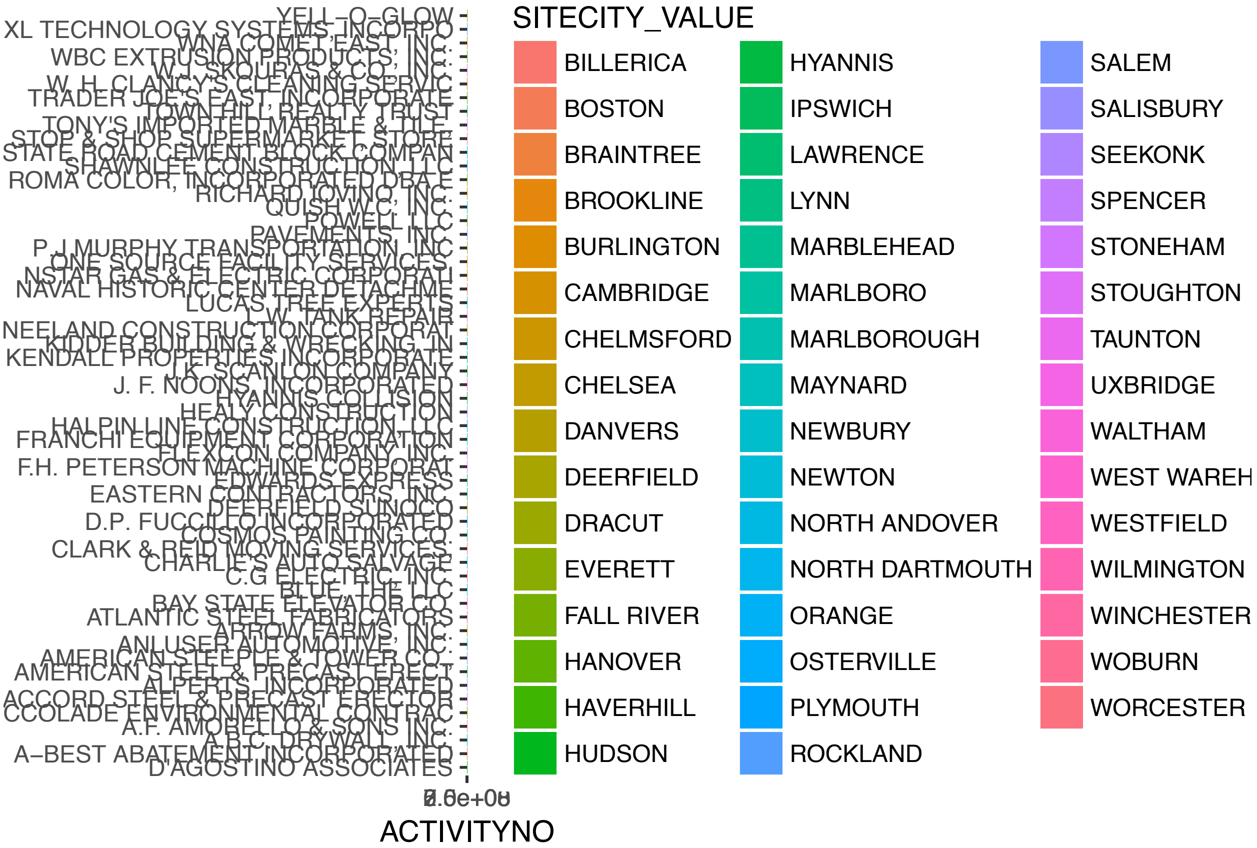
```
tidyosha = tidyosha[-dupli_rows,]
```

```
##decote SITECITY
sitecity <- read.dbf("lookups/scc.dbf")
sitecity_1 <- sitecity[(sitecity$STATE=="MA"),]
colnames(sitecity_1) <- c("TYPE", "STATE","SITECNTY","SITECITY","SITECITY_VALUE")
tidyosha <- left_join(sitecity_1, tidyosha, by = "SITECITY")
```

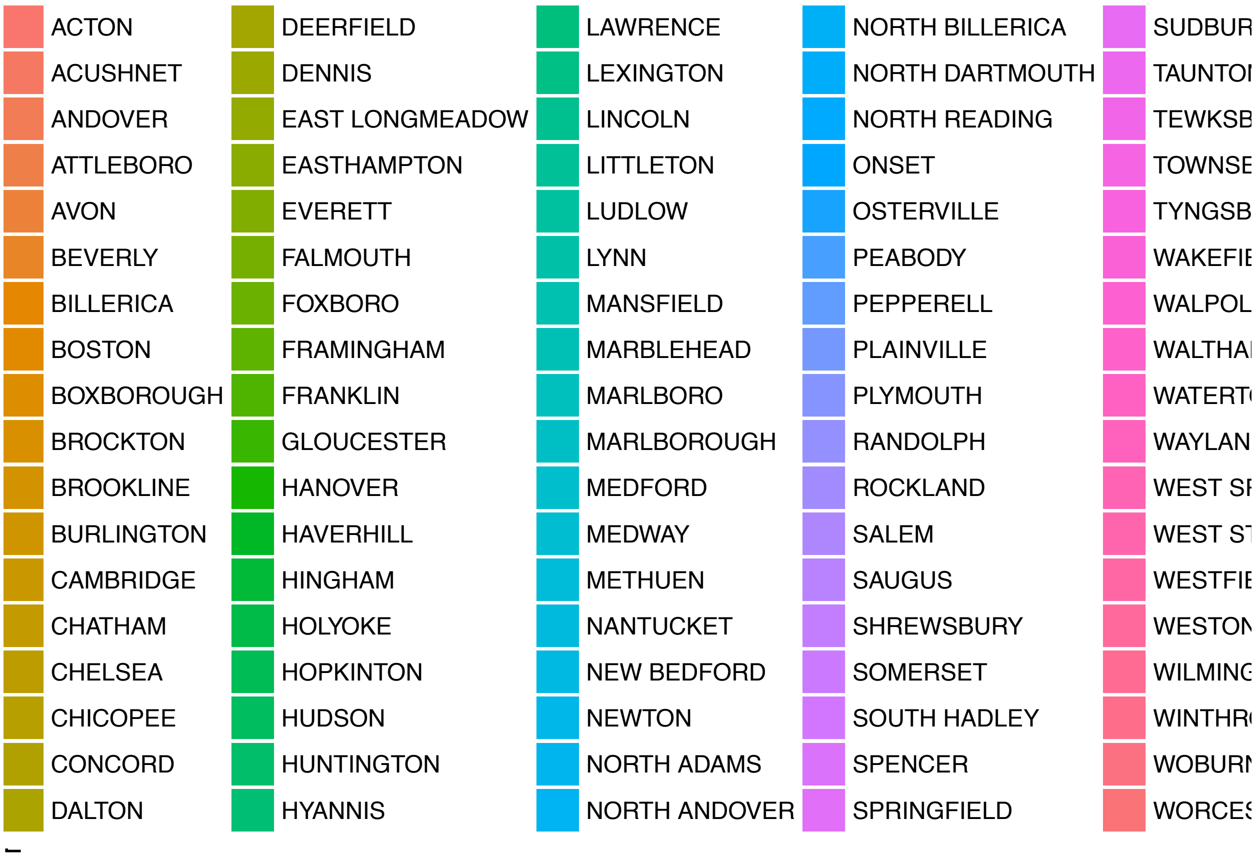
```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factors with different levels, coercing to character vector
```

```
#delete meaningless columns

drop <- c("TYPE", "STATE","SITECNTY","SITECITY")
tidyosha = tidyosha[,!(names(tidyosha) %in% drop)]







####plot

#barplot of SITECITY and ACTIVITYNO, because there are too any observations, choose the ones that are g
dangerous<- filter(tidyosha, ACTIVITYNO>300000000)

#histogram of 50 companies with the most activity numbers in alphabetical order of companies' names
# the 50th activity number is 306806944
dangerous_companies <- subset(dangerous, subset = ACTIVITYNO > 306806944)
dangerous_companies <- within(dangerous_companies,
                              ESTABNAME <- factor(ESTABNAME,
                                                  levels = names(sort(table(ESTABNAME), decreasing = TRU
c <- ggplot(dangerous_companies, aes(ESTABNAME, ACTIVITYNO))
c + geom_histogram(binwidth = 2, stat = "identity" , aes(fill = SITECITY_VALUE)) + coord_flip()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

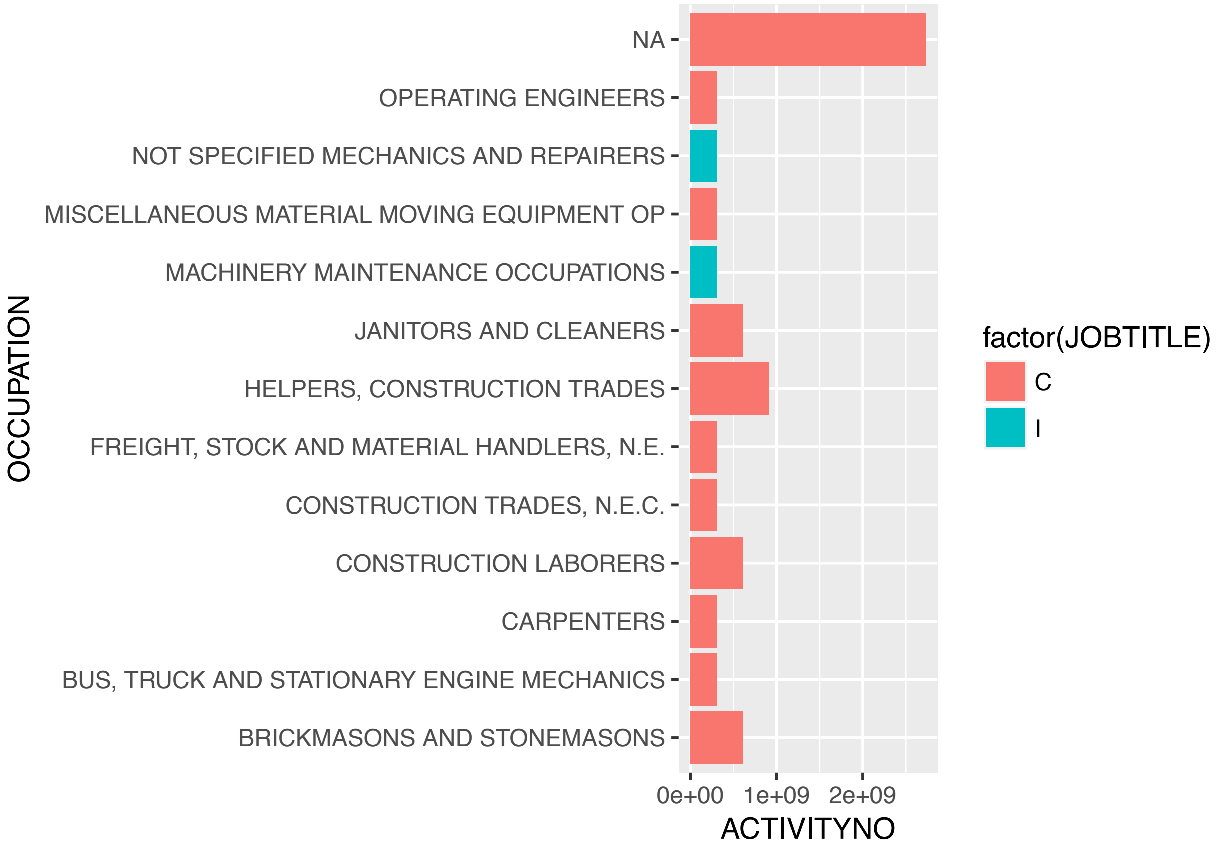


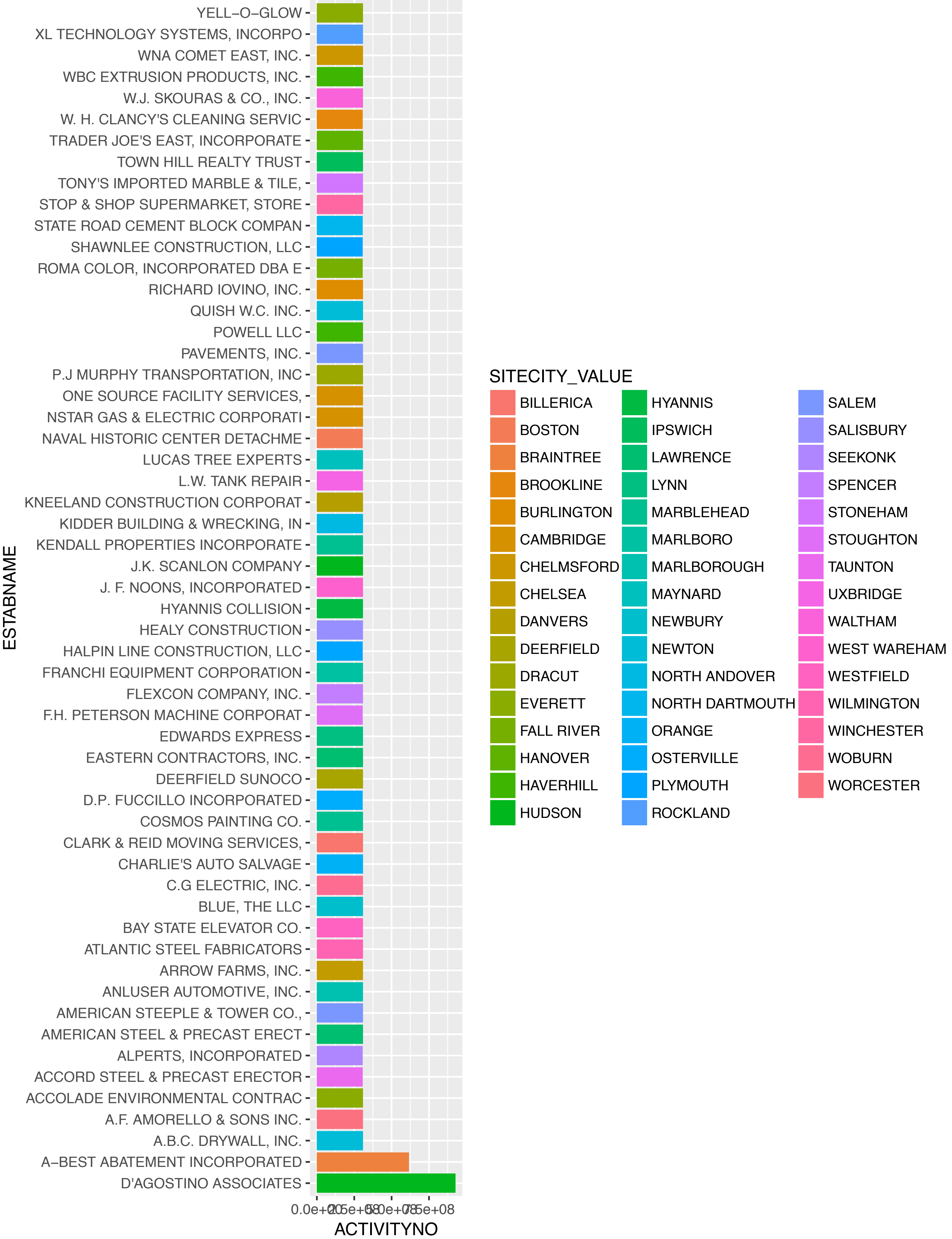
```r
#bar plot of job title and activity number in terms of sitecity
dangerous_cities = subset(dangerous, subset = OCCUPATION != "NA" )
d <- ggplot(dangerous_cities, aes(JOBTITLE, ACTIVITYNO))
d + geom_bar(stat = "identity", aes(fill = SITECITY_VALUE))
```

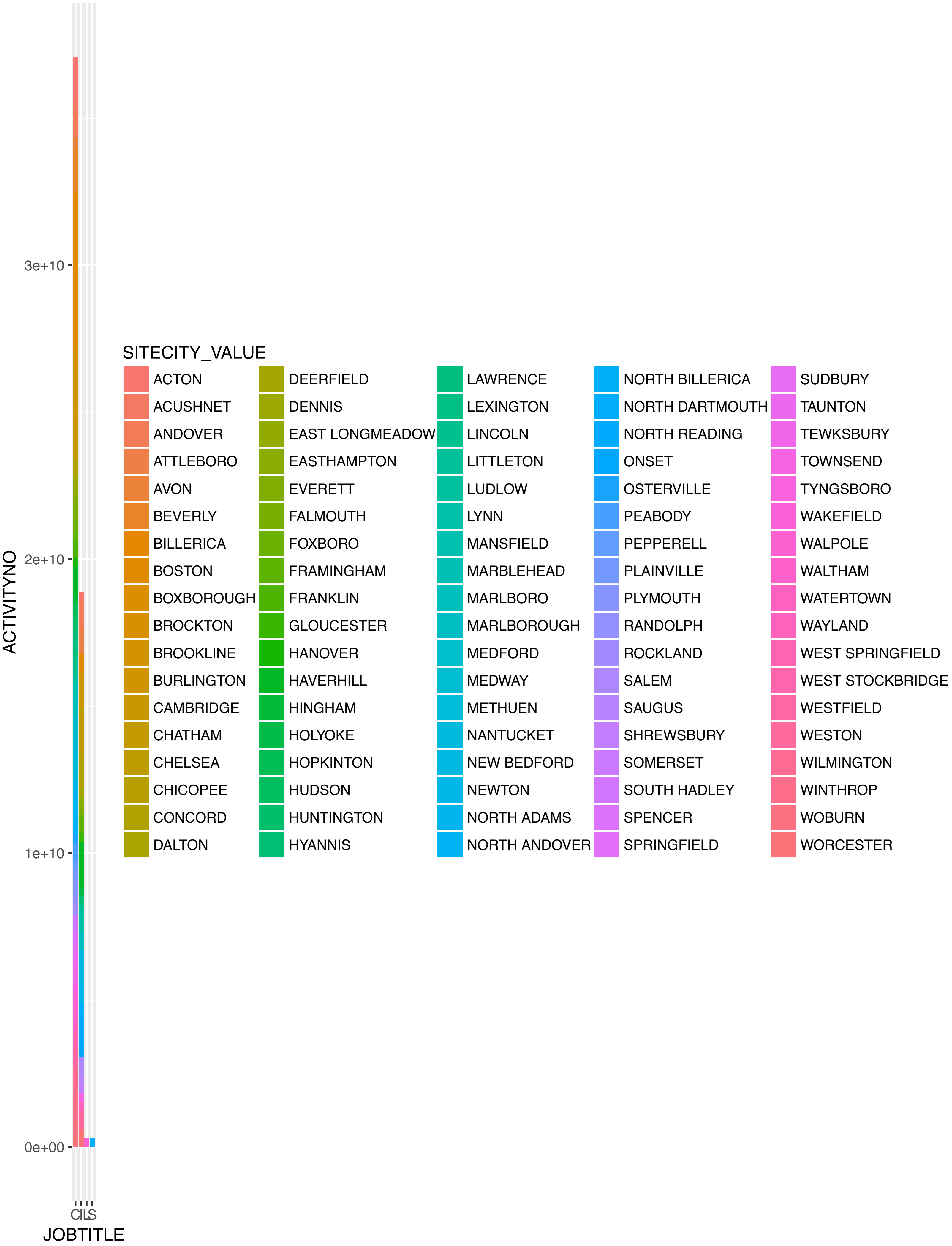


```r
#plot of activity numbers in BOSTON in terms of their job title and Latest date activity applied agains
BOSTON = subset(dangerous, subset = SITECITY_VALUE == "BOSTON")
p <- ggplot(BOSTON, aes(x = OCCUPATION, y = ACTIVITYNO))
p + geom_histogram(stat = "identity", aes(fill = factor (JOBTITLE)), size = 1)+coord_flip()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


NA
OPERATING ENGINEERS
NOT SPECIFIED MECHANICS AND REPAIRERS
MISCELLANEOUS MATERIAL MOVING EQUIPMENT OP
MACHINERY MAINTENANCE OCCUPATIONS
JANITORS AND CLEANERS
HELPERS, CONSTRUCTION TRADES
FREIGHT, STOCK AND MATERIAL HANDLERS, N.E.
CONSTRUCTION TRADES, N.E.C.
CONSTRUCTION LABORERS
CARPENTERS
BUS, TRUCK AND STATIONARY ENGINE MECHANICS
BRICKMASONS AND STONEMASONS
OCCUPATION
0e+00
1e+09
2e+09
ACTIVITYNO
factor(JOBTITLE)
C
I


YELL–O–GLOW
XL TECHNOLOGY SYSTEMS, INCORPO
WNA COMET EAST, INC.
WBC EXTRUSION PRODUCTS, INC.
W.J. SKOURAS & CO., INC.
W. H. CLANCY'S CLEANING SERVIC
TRADER JOE'S EAST, INCORPORATE
TOWN HILL REALTY TRUST
TONY'S IMPORTED MARBLE & TILE,
STOP & SHOP SUPERMARKET, STORE
STATE ROAD CEMENT BLOCK COMPAN
SHAWNLEE CONSTRUCTION, LLC
ROMA COLOR, INCORPORATED DBA E
RICHARD IOVINO, INC.
QUISH W.C. INC.
POWELL LLC
PAVEMENTS, INC.
P.J MURPHY TRANSPORTATION, INC
ONE SOURCE FACILITY SERVICES,
NSTAR GAS & ELECTRIC CORPORATI
NAVAL HISTORIC CENTER DETACHME
LUCAS TREE EXPERTS
L.W. TANK REPAIR
KNEELAND CONSTRUCTION CORPORAT
KIDDER BUILDING & WRECKING, IN
KENDALL PROPERTIES INCORPORATE
J.K. SCANLON COMPANY
J. F. NOONS, INCORPORATED
HYANNIS COLLISION
HEALY CONSTRUCTION
HALPIN LINE CONSTRUCTION, LLC
FRANCHI EQUIPMENT CORPORATION
FLEXCON COMPANY, INC.
F.H. PETERSON MACHINE CORPORAT
EDWARDS EXPRESS
EASTERN CONTRACTORS, INC.
DEERFIELD SUNOCO
D.P. FUCCILLO INCORPORATED
COSMOS PAINTING CO.
CLARK & REID MOVING SERVICES,
CHARLIE'S AUTO SALVAGE
C.G ELECTRIC, INC.
BLUE, THE LLC
BAY STATE ELEVATOR CO.
ATLANTIC STEEL FABRICATORS
ARROW FARMS, INC.
ANLUSER AUTOMOTIVE, INC.
AMERICAN STEEPLE & TOWER CO.,
AMERICAN STEEL & PRECAST ERECT
ALPERTS, INCORPORATED
ACCORD STEEL & PRECAST ERECTOR
ACCOLADE ENVIRONMENTAL CONTRAC
A.F. AMORELLO & SONS INC.
A.B.C. DRYWALL, INC.
A–BEST ABATEMENT INCORPORATED
D'AGOSTINO ASSOCIATES
ESTABNAME
0.0e+00
2.5e+08
5.0e+08
7.5e+08
ACTIVITYNO
SITECITY_VALUE
BILLERICA
BOSTON
BRAINTREE
BROOKLINE
BURLINGTON
CAMBRIDGE
CHELMSFORD
CHELSEA
DANVERS
DEERFIELD
DRACUT
EVERETT
FALL RIVER
HANOVER
HAVERHILL
HUDSON
HYANNIS
IPSWICH
LAWRENCE
LYNN
MARBLEHEAD
MARLBORO
MARLBOROUGH
MAYNARD
NEWBURY
NEWTON
NORTH ANDOVER
NORTH DARTMOUTH
ORANGE
OSTERVILLE
PLYMOUTH
ROCKLAND
SALEM
SALISBURY
SEEKONK
SPENCER
STONEHAM
STOUGHTON
TAUNTON
UXBRIDGE
WALTHAM
WEST WAREHAM
WESTFIELD
WILMINGTON
WINCHESTER
WOBURN
WORCESTER


ACTIVITYNO
3e+10
2e+10
1e+10
0e+00
CILS
JOBTITLE
SITECITY_VALUE
ACTON
ACUSHNET
ANDOVER
ATTLEBORO
AVON
BEVERLY
BILLERICA
BOSTON
BOXBOROUGH
BROCKTON
BROOKLINE
BURLINGTON
CAMBRIDGE
CHATHAM
CHELSEA
CHICOPEE
CONCORD
DALTON
DEERFIELD
DENNIS
EAST LONGMEADOW
EASTHAMPTON
EVERETT
FALMOUTH
FOXBORO
FRAMINGHAM
FRANKLIN
GLOUCESTER
HANOVER
HAVERHILL
HINGHAM
HOLYOKE
HOPKINTON
HUDSON
HUNTINGTON
HYANNIS
LAWRENCE
LEXINGTON
LINCOLN
LITTLETON
LUDLOW
LYNN
MANSFIELD
MARBLEHEAD
MARLBORO
MARLBOROUGH
MEDFORD
MEDWAY
METHUEN
NANTUCKET
NEW BEDFORD
NEWTON
NORTH ADAMS
NORTH ANDOVER
NORTH BILLERICA
NORTH DARTMOUTH
NORTH READING
ONSET
OSTERVILLE
PEABODY
PEPPERELL
PLAINVILLE
PLYMOUTH
RANDOLPH
ROCKLAND
SALEM
SAUGUS
SHREWSBURY
SOMERSET
SOUTH HADLEY
SPENCER
SPRINGFIELD
SUDBURY
TAUNTON
TEWKSBURY
TOWNSEND
TYNGSBORO
WAKEFIELD
WALPOLE
WALTHAM
WATERTOWN
WAYLAND
WEST SPRINGFIELD
WEST STOCKBRIDGE
WESTFIELD
WESTON
WILMINGTON
WINTHROP
WOBURN
WORCESTER