midterm_second_version.R

ZixinDing

Wed Mar 22 23:11:03 2017

```
## load libraries
library(data.table)
library(magrittr)
require(foreign)
```

```
## Loading required package: foreign
```

```
library(ggplot2)
require(scales)
```

```
## Loading required package: scales
```

```
library(lubridate)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

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

```
library(tidyr)
```

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

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

```
library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
```

```
## ------------------------------------------------------------------------------
```

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

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
## read in and glimpse
## first look: accid
accid<-read.dbf("Accid.dbf")
acc1<-read.dbf("lookups/acc.dbf")
ncol(accid)
```

```
## [1] 16
```

```
nrow(accid)
```

```
## [1] 2147
```

```
head(accid)
```

```
##   ACTIVITYNO SITESTATE               NAME   RELINSP  SEX DEGREE NATURE
## 1   10096592        MA               <NA>  10096592 <NA>      3     21
## 2   10096592        MA               <NA>  10096592 <NA>      3     21
## 3   10096592        MA               <NA>  10096592 <NA>      3     21
## 4   10096592        MA               <NA>  10096592 <NA>      3     21
## 5   10096592        MA               <NA>  10096592 <NA>      3     21
## 6  305548745        MA M & L POWER SERVICE 305548745    M      1     21
##   BODYPART SOURCE EVENT ENVIRON HUMAN TASK HAZSUB OCC_CODE AGE
## 1       04     16    08      07    06    2   <NA>      000   0
## 2       04     16    08      07    06    2   <NA>      000   0
## 3       04     16    08      07    06    2   <NA>      000   0
## 4       04     16    08      07    06    2   <NA>      000   0
## 5       04     16    08      07    06    2   <NA>      000   0
## 6       19     15    05      18    20    1   <NA>      575  47
```

```
#first step: check NAs for each column
indi = rep(0,ncol(accid))
for(i in 1: ncol(accid)){indi[i] = sum(!is.na(accid[,i]))}
which(indi==0)
```

```
## integer(0)
```

```
# no column all NA



############################################################
#stable information of this dataset
#default setup:1
#0 means stable information for this column
#Since there is no column that is all NA, it only two cases: First some NAs and some other stuff
#Second, all other stuff(no NAs at all)
#Our goal is to find out which column has stable information(only has one level when converted to facto
index=rep(1, ncol(accid))

for(i in 1:ncol(accid))
```

```
{
  if(length(levels(factor(accid[[i]])))==1)
  {index[i]=0}
  for(j in 1:nrow(accid))
  {
    if(is.na(accid[j,i])==TRUE)
    {index[i]=1
    break}
  }
}
which(index==0)
```

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

```
colnames(accid[which(index==0)])
```

```
## [1] "SITESTATE"
```

```
tidyaccid<-accid[,-c(which(index==0))]
ncol(tidyaccid)
```

```
## [1] 15
```

```
## So we remove sitestate column since they are all MAs



############################################################
# replace the numbers in Nature, bodypart, source, event, environ, human in tidyaccid with codes in Acc
sum(acc1$CATEGORY=="PART-BODY")
```

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

```
parts<-acc1[(acc1$CATEGORY=="PART-BODY"),]
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)
```

```
colnames(parts)<-c("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 ...
##  $ VALUE   : Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 1 7 9 15 28 40 41 45 46 49 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, 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
```

```
## remove the BODYPART column since they are all
dr = function(name){return(which(colnames(tidyaccid)==name))}
tidyaccid<-tidyaccid[,-c(dr("BODYPART"))]
ncol(tidyaccid)
```

```
## [1] 15
```

```
colnames(tidyaccid)[15]<-c("BODYPART")
head(tidyaccid)
```

```
##   ACTIVITYNO                NAME   RELINSP  SEX DEGREE NATURE SOURCE EVENT
## 1   10096592                <NA>  10096592 <NA>      3     21     16    08
## 2   10096592                <NA>  10096592 <NA>      3     21     16    08
## 3   10096592                <NA>  10096592 <NA>      3     21     16    08
## 4   10096592                <NA>  10096592 <NA>      3     21     16    08
## 5   10096592                <NA>  10096592 <NA>      3     21     16    08
## 6  305548745 M & L POWER SERVICE 305548745    M      1     21     15    05
##   ENVIRON HUMAN TASK HAZSUB OCC_CODE AGE   BODYPART
## 1      07    06    2   <NA>      000   0 BODYSYSTEM
## 2      07    06    2   <NA>      000   0 BODYSYSTEM
## 3      07    06    2   <NA>      000   0 BODYSYSTEM
## 4      07    06    2   <NA>      000   0 BODYSYSTEM
## 5      07    06    2   <NA>      000   0 BODYSYSTEM
## 6      18    20    1   <NA>      575  47   MULTIPLE
```

```
## repeat the process for other columns
nature<-acc1[(acc1$CATEGORY=="NATUR-INJ"), ]
nature<-select(nature, CODE, VALUE)
head(nature)
```

```
##    CODE               VALUE
## 84   01           AMPUTATION
## 85   02             ASPHYXIA
## 86   03 BRUISE/CONTUS/ABRAS
## 87   04       BURN(CHEMICAL)
## 88   05    BURN/SCALD(HEAT)
## 89   06           CONCUSSION
```

```
colnames(nature)<-c("NATURE", "VALUE")
str(nature)
```

```
## 'data.frame':    22 obs. of  2 variables:
##  $ NATURE: Factor w/ 48 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ VALUE : Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 5 8 18 20 21 31 32 34 36 43 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, nature, 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
```

```
tidyaccid<-tidyaccid[, -c(dr("NATURE"), dr("VALUE.y"))]
ncol(tidyaccid)
```

```
## [1] 15
```

```
colnames(tidyaccid)[15]<-c("NATURE")
```

```
event<-acc1[(acc1$CATEGORY=="EVENT-TYP"), ]
event<-select(event, CODE, VALUE)
colnames(event)<-c("EVENT", "VALUE")
str(event)
```

```
## 'data.frame':    14 obs. of  2 variables:
##  $ EVENT: Factor w/ 48 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ VALUE: Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 136 25 11 48 47 135 127 76 75 2 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, 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
```

```
tidyaccid<-tidyaccid[, -c(dr("EVENT"))]
ncol(tidyaccid)
```

```
## [1] 15
```

```
colnames(tidyaccid)[15]<-c("Event")

human<-acc1[(acc1$CATEGORY=="HUMAN-FAC"), ]
human<-select(human, CODE, VALUE)
colnames(human)<-c("HUMAN", "VALUE")
str(human)
```

```
## 'data.frame':    20 obs. of  2 variables:
##  $ HUMAN: Factor w/ 48 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ VALUE: Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 99 108 107 93 37 44 94 114 128 119 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, human, by="HUMAN")
```

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

```
tidyaccid<-tidyaccid[, -c(dr("HUMAN"))]
ncol(tidyaccid)
```

```
## [1] 15
```

```
colnames(tidyaccid)[15]<-c("HUMAN")

source<-acc1[(acc1$CATEGORY=="SOURC-INJ"),]
source<-select(source, CODE, VALUE)
colnames(source)<-c("SOURCE", "VALUE")
str(source)
```

```
## 'data.frame':    48 obs. of  2 variables:
##  $ SOURCE: Factor w/ 48 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ VALUE : Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 4 3 6 13 14 16 17 19 26 29 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, source, 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
```

```
tidyaccid<-tidyaccid[, -c(dr("SOURCE"))]
colnames(tidyaccid)[15]<-c("SOURCE")

environ<-acc1[(acc1$CATEGORY=="ENVIR-FAC"),]
environ<-select(environ, CODE, VALUE)
colnames(environ)<-c("ENVIRON", "VALUE")
str(environ)
```

```
## 'data.frame':    18 obs. of  2 variables:
##  $ ENVIRON: Factor w/ 48 levels "01","02","03",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ VALUE  : Factor w/ 149 levels "ABDOMEN","ABSORPTION",..: 117 24 129 133 53 112 61 97 27 52 ...
##  - attr(*, "data_types")= chr  "C" "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, 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
```

```
tidyaccid<-tidyaccid[, -c(dr("ENVIRON"))]
colnames(tidyaccid)[15]<-c("ENVIRON")

# Now we have now convert all codes in ACCID with ACC
# Then we can convert codes in HAZSUB with HZS

hzs<-read.dbf("lookups/hzs.dbf")
colnames(hzs)[1]<-c("HAZSUB")
str(hzs)
```

```
## 'data.frame':    1777 obs. of  2 variables:
##  $ HAZSUB: Factor w/ 1777 levels "0005","0010",..: 809 810 811 812 813 814 815 816 817 818 ...
##  $ TEXT  : Factor w/ 1771 levels "(DICHLOROMETHYL) BENZENE",..: 295 290 292 291 289 245 246 288 305 1
##  - attr(*, "data_types")= chr  "C" "C"
```

```
tidyaccid<-left_join(tidyaccid, hzs, 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
```

```
ncol(tidyaccid)
```

```
## [1] 16
```

```
tidyaccid<-tidyaccid[, -c(dr("HAZSUB"))]
colnames(tidyaccid)[15]<-c("HAZSUB")

# Then we convert codes in OCC_CODE with OCC
occ<-read.dbf("lookups/occ.DBF")
colnames(occ)[1]<-c("OCC_CODE")
tidyaccid<-left_join(tidyaccid, occ, 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
```

```
tidyaccid<-tidyaccid[, -c(dr("OCC_CODE"))]
# decode degree type
tidyaccid$DEGREE<-gsub(1, "fatality", tidyaccid$DEGREE)
tidyaccid$DEGREE<-gsub(2, "hospitalized", tidyaccid$DEGREE)
tidyaccid$DEGREE<-gsub(3, "nonhospitalized", tidyaccid$DEGREE)
```

```r
# decode task type
tidyaccid$TASK<-gsub(1, "regularly assigned task", tidyaccid$TASK)
tidyaccid$TASK<-gsub(2, "task other than regularly assigned", tidyaccid$TASK)
## above ACCID finish converting
tidydata<-tidyaccid[, c(dr("ACTIVITYNO"), dr("DEGREE"), dr("BODYPART"), dr("NATURE"), dr("EVENT"), dr("
# extract some column to prepare for the final data



##########################################
osha<-read.dbf("osha.DBF")
head(osha)
```

```
##   CONTFLAG HISTFLAG OSHA1MOD STFLAG PREVCTTYP PREVACTNO ACTIVITYNO
## 1     <NA>        H 19840221   <NA>      <NA>         0   10236776
## 2     <NA>        M 19910523   <NA>      <NA>         0  103393633
## 3     <NA>        H 19880618   <NA>      <NA>         0   18750034
## 4     <NA>        H 19880618   <NA>      <NA>         0   18750042
## 5     <NA>        H 19880618   <NA>      <NA>         0   18750059
## 6     <NA>        H 19880618   <NA>      <NA>         0   18750067
##   REPORTID CSHO_ID JOBTITLE OPTREPTNO            ESTABNAME
## 1  0111100    <NA>        C 000000000        DUBE DRY WALL
## 2  0111100    <NA>        I 000000000 KNOWLTON MACHINE CO.
## 3  0111400    <NA>     <NA> 000000000       RENTAL & FROST
## 4  0111400    <NA>     <NA> 000000000     PENN TRUCK LINES
## 5  0111400    <NA>     <NA> 000000000       SILVERITE GUTT
## 6  0111400    <NA>     <NA> 000000000         MARSSON CORP
##                         SITEADD SITESTATE HOSTESTKEY OWNERTYPE OWNERCODE
## 1                    RT 1 MAIN ST        MA       <NA>      <NA>         0
## 2 NEW ENGLAND POWER, SALEM HARBO        MA       <NA>         A         0
## 3                           <NA>        MA       <NA>      <NA>         0
## 4                           <NA>        MA       <NA>      <NA>         0
## 5                           <NA>        MA       <NA>      <NA>         0
## 6                           <NA>        MA       <NA>      <NA>         0
##   ADVNOTICE OPENDATE CLOSEDATE CAT_SH  NAICS NAICSEC NAICSINS INSPTYPE
## 1      <NA> 19831215         0      S 000000  000000   000000        H
## 2         N 19900717  19900720      H 000000  000000   000000        B
## 3      <NA> 19790514  19790514      S 000000  000000   000000        F
## 4      <NA> 19790517  19790517      H 000000  000000   000000        F
## 5      <NA> 19790710  19790710      H 000000  000000   000000        B
## 6      <NA> 19790919  19790919      H 000000  000000   000000        B
##   INSPSCOPE EMPCOUNT EMPCOVERED NATEMPCNT WALKAROUND INTRVIEWD UNION
## 1         A        0          0         0       <NA>      <NA>     N
## 2         B        0          0         0          X      <NA>     N
## 3         D        0          0         0       <NA>      <NA>  <NA>
## 4         D        0          0         0       <NA>      <NA>  <NA>
## 5         D        0          0         0       <NA>      <NA>  <NA>
## 6         D        0          0         0       <NA>      <NA>  <NA>
##   CLOSECASE WHYNOINSP CLOSEDATE2 SAFETYMANF SFTYCONST SFTYMARIT HELTHMANF
## 1         X      <NA>   19840206       <NA>         X      <NA>      <NA>
## 2         X      <NA>   19910522       <NA>      <NA>      <NA>         X
## 3         X         E   19880616       <NA>      <NA>      <NA>      <NA>
## 4         X         E   19880616       <NA>      <NA>      <NA>      <NA>
## 5         X         E   19880616       <NA>      <NA>      <NA>      <NA>
```

```
## 6         X         E   19880616       <NA>      <NA>      <NA>      <NA>
##   HELTHCONST HELTHMARIT MIGRANT ANTCSRVD FRSTDENY LSTREENTR LWDIRATE SHPGM
## 1       <NA>       <NA>    <NA>     <NA>        0         0        0  <NA>
## 2       <NA>       <NA>    <NA>     <NA>        0         0        0  <NA>
## 3       <NA>       <NA>    <NA>     <NA> 19790514         0        0  <NA>
## 4       <NA>       <NA>    <NA>     <NA> 19790517         0        0  <NA>
## 5       <NA>       <NA>    <NA>     <NA> 19790710         0        0  <NA>
## 6       <NA>       <NA>    <NA>     <NA> 19790919         0        0  <NA>
##   DATARQD PENDUDATE FTADUDATE DUECODE PAPREP PATRAVEL PAONSITE PATECHSUPP
## 1    <NA>  19850901         0       N      0        0        0          0
## 2    <NA>  19900815         0       D     40       40      100          0
## 3    <NA>         0         0    <NA>      0        0        0          0
## 4    <NA>         0         0    <NA>      0        0        0          0
## 5    <NA>         0         0    <NA>      0        0        0          0
## 6    <NA>         0         0    <NA>      0        0        0          0
##   PARPTPREP PAOTHRCNF PALITIGN PADENIAL PASUMHOURS FRSTCONTST PENREMIT
## 1        40         0        0        0         40          0      160
## 2       180         0        0        0        360          0     1820
## 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
##   FTAREMIT TOTPENLTY TOTALFTA TOTALVIOLS TOTSERIOUS PROG_ RELACT_ OPTINFO_
## 1        0       160        0          4          1     0       0        1
## 2        0      1820        0          5          4     0       1        1
## 3        0         0        0          0          0     0       0        0
## 4        0         0        0          0          0     0       0        0
## 5        0         0        0          0          0     0       0        0
## 6        0         0        0          0          0     0       0        0
##   DEBT_ VIOLS_ EVENT_ HAZSUB_ ACCID_ ADMPAY_  SIC SITEZIP SITECITY
## 1     0      4      0       0      0       1 1742   04074     1265
## 2     0      5      0       0      0       1 3599   01970     1110
## 3     0      0      0       0      0       0 3444   00000     0120
## 4     0      0      0       0      0       0 4789   00000     0120
## 5     0      0      0       0      0       0 3131   00000     0120
## 6     0      0      0       0      0       0 2851   00000     0200
##   SITECNTY    DUNSNO CATSICGDE CATSICINSP LSTR_DT    FRST_DT   MOD_DATE
## 1      011 000000000      0000       0000    <NA>       <NA> 1984-02-21
## 2      009 000000000      0000       0000    <NA>       <NA> 1991-05-23
## 3      025 000000000      0000       0000    <NA> 1979-05-14 1988-06-18
## 4      025 000000000      0000       0000    <NA> 1979-05-17 1988-06-18
## 5      025 000000000      0000       0000    <NA> 1979-07-10 1988-06-18
## 6      025 000000000      0000       0000    <NA> 1979-09-19 1988-06-18
##       OPENDT    CLOSEDT   CLOSEDT2    PENDUDT FTADUDT FRSTCONDT
## 1 1983-12-15       <NA> 1984-02-06 1985-09-01    <NA>      <NA>
## 2 1990-07-17 1990-07-20 1991-05-22 1990-08-15    <NA>      <NA>
## 3 1979-05-14 1979-05-14 1988-06-16       <NA>    <NA>      <NA>
## 4 1979-05-17 1979-05-17 1988-06-16       <NA>    <NA>      <NA>
## 5 1979-07-10 1979-07-10 1988-06-16       <NA>    <NA>      <NA>
## 6 1979-09-19 1979-09-19 1988-06-16       <NA>    <NA>      <NA>
```

```
fda<-read.dbf("lookups/fda.dbf")
# extract some useful column
tidyosha<-data.frame(osha$ACTIVITYNO, osha$JOBTITLE, osha$ESTABNAME, osha$OWNERCODE, osha$EMPCOUNT, osha
```

```
ncol(tidyosha)
```

```
## [1] 11
```

```
head(tidyosha)
```

```
##   osha.ACTIVITYNO osha.JOBTITLE       osha.ESTABNAME osha.OWNERCODE
## 1        10236776             C        DUBE DRY WALL              0
## 2       103393633             I KNOWLTON MACHINE CO.              0
## 3        18750034          <NA>       RENTAL & FROST              0
## 4        18750042          <NA>     PENN TRUCK LINES              0
## 5        18750059          <NA>       SILVERITE GUTT              0
## 6        18750067          <NA>         MARSSON CORP              0
##   osha.EMPCOUNT osha.NATEMPCNT osha.CLOSECASE osha.NAICS osha.SIC
## 1             0              0              X     000000     1742
## 2             0              0              X     000000     3599
## 3             0              0              X     000000     3444
## 4             0              0              X     000000     4789
## 5             0              0              X     000000     3131
## 6             0              0              X     000000     2851
##   osha.SITECITY osha.SITECNTY
## 1          1265           011
## 2          1110           009
## 3          0120           025
## 4          0120           025
## 5          0120           025
## 6          0200           025
```

```
head(fda)
```

```
##   CODE                  AGENCY
## 1   10                 C.I.A.
## 2   80        OFF OF POLICY DEV
## 3   90                 E.P.A.
## 4  200                 E.E.O.
## 5  280    NATL SECURITY COUNCIL
## 6  300 OCC SAFETY&HEALTH REVIEW
```

```
# convert codes in owmercode in tidyosha with fda
colnames(fda)[1]<-c("osha.OWNERCODE")
tidyosha<-left_join(tidyosha, fda, by="osha.OWNERCODE")
head(tidyosha)
```

```
##   osha.ACTIVITYNO osha.JOBTITLE       osha.ESTABNAME osha.OWNERCODE
## 1        10236776             C        DUBE DRY WALL              0
## 2       103393633             I KNOWLTON MACHINE CO.              0
## 3        18750034          <NA>       RENTAL & FROST              0
## 4        18750042          <NA>     PENN TRUCK LINES              0
## 5        18750059          <NA>       SILVERITE GUTT              0
## 6        18750067          <NA>         MARSSON CORP              0
##   osha.EMPCOUNT osha.NATEMPCNT osha.CLOSECASE osha.NAICS osha.SIC
## 1             0              0              X     000000     1742
## 2             0              0              X     000000     3599
## 3             0              0              X     000000     3444
## 4             0              0              X     000000     4789
## 5             0              0              X     000000     3131
```

```
## 6             0              0             X     000000     2851
##   osha.SITECITY osha.SITECNTY AGENCY
## 1          1265           011   <NA>
## 2          1110           009   <NA>
## 3          0120           025   <NA>
## 4          0120           025   <NA>
## 5          0120           025   <NA>
## 6          0200           025   <NA>
```

```
tidyosha$osha.OWNERCODE<-NULL
# convert codes in sitecity and sitecnty with scc
scc<-read.dbf("lookups/scc.dbf")
head(scc)
```

```
##   TYPE STATE COUNTY CITY            NAME
## 1    1    AK    000 0000           ALASKA
## 2    2    AK    010 0000 ALEUTIAN ISLANDS
## 3    2    AK    013 0000   ALEUTIANS EAST
## 4    2    AK    016 0000   ALEUTIANS WEST
## 5    2    AK    020 0000        ANCHORAGE
## 6    2    AK    050 0000           BETHEL
```

```
head(filter(scc, STATE=="MA"))
```

```
##   TYPE STATE COUNTY CITY          NAME
## 1    1    MA    000 0000 MASSACHUSETTS
## 2    2    MA    001 0000    BARNSTABLE
## 3    2    MA    003 0000     BERKSHIRE
## 4    2    MA    005 0000       BRISTOL
## 5    2    MA    007 0000         DUKES
## 6    2    MA    009 0000         ESSEX
```

```
sccma<-filter(scc, STATE=="MA")
sccma$TYPE<-NULL
sccma$STATE<-NULL
sccma<-rename(sccma, osha.SITECITY=CITY, osha.SITECNTY=COUNTY)
tidyosha<-left_join(tidyosha, sccma, by="osha.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
```

```
tidyosha$osha.SITECNTY.x<-NULL
tidyosha<-rename(tidyosha, osha.SITECNTY=osha.SITECNTY.y)
tidyosha$osha.SITECITY<-NULL
tidyosha$osha.SITECNTY<-NULL
tidyosha<-rename(tidyosha, PLACE=NAME)
# convert SIC code with sic.dbf
sic<-read.dbf("lookups/sic.dbf")
sic<-rename(sic, osha.SIC=SIC)
tidyosha<-left_join(tidyosha, sic, bu="osha.SIC")
```

```
## Joining, by = "osha.SIC"
```

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

```
tidyosha$osha.SIC=NULL
# convert code in osha.JOBTITLE by following guidelines in osha.txt
```

```
# A = area director C = safety officer I = health officer L = safety trainee M = health trainee
# N = national offic e management O = area office support staff P = compliance program manager R =
# S = supervisor T = safety and health technician U = area office analyst V = discrim. invest'r
# W = regional mgt. X = regional FSO Y = regional tech. supp. Z = regional management
x<-c("A", "C", "I", "L", "M", "N", "O", "P", "S", "T", "U", "V", "W", "X", "Y", "Z")
r<-c("area director", "safety officer", "health officer", "safety trainee",
"health trainee", "national office management", "area office support staff", "compliance program manage
"supervisor", "safety and health technician", "area office analyst", "discrim. invest'r", "regional mgt
"regional FSO", "regional tech. supp.", "regional management")
jobtitle<-data.frame(osha.JOBTITLE=x, JOBTITLE=r)
tidyosha<-left_join(tidyosha, jobtitle, by="osha.JOBTITLE")
```

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

```
tidyosha$osha.JOBTITLE<-NULL
# convert code in NAICS by following guidelines in naics.dbf
naics<-read.dbf("lookups/naics.dbf")
naics<-rename(naics, osha.NAICS=NAICS)
tidyosha<-left_join(tidyosha, naics, by="osha.NAICS")
```

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

```
tidyosha$osha.NAICS<-NULL
names(tidyosha) = sub("osha.","",names(tidyosha))
head(tidyosha)
```

```
##   ACTIVITYNO            ESTABNAME EMPCOUNT NATEMPCNT CLOSECASE AGENCY
## 1   10236776        DUBE DRY WALL        0         0         X   <NA>
## 2  103393633 KNOWLTON MACHINE CO.        0         0         X   <NA>
## 3   18750034       RENTAL & FROST        0         0         X   <NA>
## 4   18750042     PENN TRUCK LINES        0         0         X   <NA>
## 5   18750059       SILVERITE GUTT        0         0         X   <NA>
## 6   18750067         MARSSON CORP        0         0         X   <NA>
##        PLACE                             INDUSTRY       JOBTITLE NAICSTEXT
## 1 SUNDERLAND PLASTERING, DRYWALL, AND INSULATION safety officer      <NA>
## 2      SALEM           INDUSTRIAL MACHINERY, NEC health officer      <NA>
## 3     BOSTON                     SHEET METAL WORK           <NA>      <NA>
## 4     BOSTON        TRANSPORTATION SERVICES, NEC           <NA>      <NA>
## 5     BOSTON                   FOOTWEAR CUT STOCK           <NA>      <NA>
## 6    CHELSEA          PAINTS AND ALLIED PRODUCTS           <NA>      <NA>
```

```
## combine all data together
head(tidydata)
```

```
##   ACTIVITYNO          DEGREE   BODYPART NATURE
## 1   10096592 nonhospitalized BODYSYSTEM  OTHER
## 2   10096592 nonhospitalized BODYSYSTEM  OTHER
## 3   10096592 nonhospitalized BODYSYSTEM  OTHER
## 4   10096592 nonhospitalized BODYSYSTEM  OTHER
## 5   10096592 nonhospitalized BODYSYSTEM  OTHER
## 6  305548745        fatality   MULTIPLE  OTHER
##                              HUMAN             SOURCE
## 1    EQUIP. INAPPROPR FOR OPERATION          FIRE/SMOKE
## 2    EQUIP. INAPPROPR FOR OPERATION          FIRE/SMOKE
```

```
## 3    EQUIP. INAPPROPR FOR OPERATION          FIRE/SMOKE
## 4    EQUIP. INAPPROPR FOR OPERATION          FIRE/SMOKE
## 5    EQUIP. INAPPROPR FOR OPERATION          FIRE/SMOKE
## 6 INSUF/LACK/PROTCV WRK CLTHG/EQUIP ELEC APPARAT/WIRING
##                         ENVIRON HAZSUB   OCCUPATION
## 1 GAS/VAPOR/MIST/FUME/SMOKE/DUST   <NA>         <NA>
## 2 GAS/VAPOR/MIST/FUME/SMOKE/DUST   <NA>         <NA>
## 3 GAS/VAPOR/MIST/FUME/SMOKE/DUST   <NA>         <NA>
## 4 GAS/VAPOR/MIST/FUME/SMOKE/DUST   <NA>         <NA>
## 5 GAS/VAPOR/MIST/FUME/SMOKE/DUST   <NA>         <NA>
## 6                          OTHER   <NA> ELECTRICIANS
```

```
head(tidyosha)
```

```
##   ACTIVITYNO            ESTABNAME EMPCOUNT NATEMPCNT CLOSECASE AGENCY
## 1   10236776        DUBE DRY WALL        0         0         X   <NA>
## 2  103393633 KNOWLTON MACHINE CO.        0         0         X   <NA>
## 3   18750034       RENTAL & FROST        0         0         X   <NA>
## 4   18750042     PENN TRUCK LINES        0         0         X   <NA>
## 5   18750059       SILVERITE GUTT        0         0         X   <NA>
## 6   18750067         MARSSON CORP        0         0         X   <NA>
##        PLACE                              INDUSTRY       JOBTITLE NAICSTEXT
## 1 SUNDERLAND PLASTERING, DRYWALL, AND INSULATION safety officer      <NA>
## 2      SALEM           INDUSTRIAL MACHINERY, NEC health officer      <NA>
## 3     BOSTON                      SHEET METAL WORK          <NA>      <NA>
## 4     BOSTON        TRANSPORTATION SERVICES, NEC          <NA>      <NA>
## 5     BOSTON                  FOOTWEAR CUT STOCK          <NA>      <NA>
## 6    CHELSEA          PAINTS AND ALLIED PRODUCTS          <NA>      <NA>
```

```
tidydata<-left_join(tidyosha, tidydata, by="ACTIVITYNO")
## check stable information
## only two columns
index=rep(0, 2)
for(i in 1:nrow(tidydata))
{
  if(tidydata[i,3]!="0")
  {
    index[1]=1
  }
  if(tidydata[i,4]!="0")
  {
    index[2]=1
  }
}
index
```

```
## [1] 0 0
```

```
## these two columns are stable information
tidydata$EMPCOUNT<-NULL
tidydata$NATEMPCNT<-NULL
## decode closecase
closecase<-data.frame(CLOSECASE=c("X", NA), CLOSE=c("Yes", "NO"))
tidydata<-left_join(tidydata, closecase, by="CLOSECASE")
tidydata$CLOSECASE<-NULL
```

```
#################################################################
# violation
viol<-read.dbf("viol.DBF")
# extract some columns
tidyviol<-data.frame(viol$ACTIVITYNO, viol$EMPHASIS, viol$GRAVITY, viol$VIOLTYPE, viol$STD_LOOKUP, viol$
names(tidyviol) = sub("viol.","",names(tidyviol))
# decode emphasis
levels(tidyviol$EMPHASIS)
```

```
## [1] "X"
```

```
tidyviol$EMPHASIS=gsub("X", "egregious", tidyviol$EMPHASIS)
# glimpse at gravity
levels(tidyviol$GRAVITY)
```

```
##  [1] "00" "01" "02" "03" "04" "05" "06" "07" "08" "09" "10"
```

```
colnames(tidyviol)[3]<-c("GRAVITYLEVEL")
head(tidyviol)
```

```
##   ACTIVITYNO EMPHASIS GRAVITYLEVEL VIOLTYPE STD_LOOKUP INSTANCES  REC
## 1   10236776     <NA>         <NA>        S   19260451         1 <NA>
## 2   10236776     <NA>         <NA>        O   19260400         1 <NA>
## 3   10236776     <NA>         <NA>        O   19260401         1 <NA>
## 4   10236776     <NA>         <NA>        O   19260401         1 <NA>
## 5  103393633     <NA>           07        S   19260058         1    C
## 6  103393633     <NA>           08        S   19260058         1    C
##   NUMEXPOSED ABATEDONE HAZCAT
## 1          0         N   <NA>
## 2          0         N   <NA>
## 3          0         N   <NA>
## 4          0         N   <NA>
## 5         10         W   <NA>
## 6         10         W   <NA>
```

```
# decode at violtype
violtype<-data.frame(VIOLTYPE=c("O", "R", "S", "U", "W"), CODE=c("other", "repeat", "serious", "unclassi
head(violtype)
```

```
##   VIOLTYPE         CODE
## 1        O        other
## 2        R       repeat
## 3        S      serious
## 4        U unclassified
## 5        W      willful
```

```
tidyviol<-left_join(tidyviol, violtype, by="VIOLTYPE")
head(tidyviol)
```

```
##   ACTIVITYNO EMPHASIS GRAVITYLEVEL VIOLTYPE STD_LOOKUP INSTANCES  REC
## 1   10236776     <NA>         <NA>        S   19260451         1 <NA>
## 2   10236776     <NA>         <NA>        O   19260400         1 <NA>
## 3   10236776     <NA>         <NA>        O   19260401         1 <NA>
## 4   10236776     <NA>         <NA>        O   19260401         1 <NA>
## 5  103393633     <NA>           07        S   19260058         1    C
```

```
## 6  103393633     <NA>           08        S   19260058         1    C
##   NUMEXPOSED ABATEDONE HAZCAT    CODE
## 1          0         N   <NA> serious
## 2          0         N   <NA>   other
## 3          0         N   <NA>   other
## 4          0         N   <NA>   other
## 5         10         W   <NA> serious
## 6         10         W   <NA> serious
```

```
tidyviol$VIOLTYPE<-NULL
tidyviol<-rename(tidyviol, VIOLTYPE=CODE)
head(tidyviol)
```

```
##   ACTIVITYNO EMPHASIS GRAVITYLEVEL STD_LOOKUP INSTANCES  REC NUMEXPOSED
## 1   10236776     <NA>         <NA>   19260451         1 <NA>          0
## 2   10236776     <NA>         <NA>   19260400         1 <NA>          0
## 3   10236776     <NA>         <NA>   19260401         1 <NA>          0
## 4   10236776     <NA>         <NA>   19260401         1 <NA>          0
## 5  103393633     <NA>           07   19260058         1    C         10
## 6  103393633     <NA>           08   19260058         1    C         10
##   ABATEDONE HAZCAT VIOLTYPE
## 1         N   <NA>  serious
## 2         N   <NA>    other
## 3         N   <NA>    other
## 4         N   <NA>    other
## 5         W   <NA>  serious
## 6         W   <NA>  serious
```

```
# decode for STD_LOOKUP
std<-read.dbf("lookups/STD.dbf")
colnames(std)[2]<-c("STD_LOOKUP")
tidyviol<-left_join(tidyviol, std, by="STD_LOOKUP")
```

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

```
tidyviol$STD_LOOKUP<-NULL
tidyviol<-rename(tidyviol, STD_LOOKUP=TEXT)
head(tidyviol)
```

```
##   ACTIVITYNO EMPHASIS GRAVITYLEVEL INSTANCES  REC NUMEXPOSED ABATEDONE
## 1   10236776     <NA>         <NA>         1 <NA>          0         N
## 2   10236776     <NA>         <NA>         1 <NA>          0         N
## 3   10236776     <NA>         <NA>         1 <NA>          0         N
## 4   10236776     <NA>         <NA>         1 <NA>          0         N
## 5  103393633     <NA>           07         1    C         10         W
## 6  103393633     <NA>           08         1    C         10         W
##   HAZCAT VIOLTYPE STATE                                    STD_LOOKUP
## 1   <NA>  serious    FE                                   SCAFFOLDING
## 2   <NA>    other    FE               ELECTRICAL, GENERAL INTRODUCTION
## 3   <NA>    other    FE                     ELECTRICAL, APPLICABILITY
## 4   <NA>    other    FE                     ELECTRICAL, APPLICABILITY
## 5   <NA>  serious    FE ASBESTOS, TREMOLITE, ANTHOPHYLLITE & ACTINOLITE
## 6   <NA>  serious    FE ASBESTOS, TREMOLITE, ANTHOPHYLLITE & ACTINOLITE
```

```
# decode for related event
rec<-data.frame(REC=c("A", "C", "I", "R", "V"), RELATEDEVENT=c("FAT/CAT (fatality/catastrophe), acciden
```

```
tidyviol<-left_join(tidyviol, rec, by="REC")
```

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

```
head(tidyviol)
```

```
##   ACTIVITYNO EMPHASIS GRAVITYLEVEL INSTANCES  REC NUMEXPOSED ABATEDONE
## 1   10236776     <NA>         <NA>         1 <NA>          0         N
## 2   10236776     <NA>         <NA>         1 <NA>          0         N
## 3   10236776     <NA>         <NA>         1 <NA>          0         N
## 4   10236776     <NA>         <NA>         1 <NA>          0         N
## 5  103393633     <NA>           07         1    C         10         W
## 6  103393633     <NA>           08         1    C         10         W
##   HAZCAT VIOLTYPE STATE                                    STD_LOOKUP
## 1   <NA>  serious    FE                                    SCAFFOLDING
## 2   <NA>    other    FE                ELECTRICAL, GENERAL INTRODUCTION
## 3   <NA>    other    FE                       ELECTRICAL, APPLICABILITY
## 4   <NA>    other    FE                       ELECTRICAL, APPLICABILITY
## 5   <NA>  serious    FE ASBESTOS, TREMOLITE, ANTHOPHYLLITE & ACTINOLITE
## 6   <NA>  serious    FE ASBESTOS, TREMOLITE, ANTHOPHYLLITE & ACTINOLITE
##   RELATEDEVENT
## 1         <NA>
## 2         <NA>
## 3         <NA>
## 4         <NA>
## 5    complaint
## 6    complaint
```

```
tidyviol$REC<-NULL
tidyviol$ABATEDONE<-gsub("X", "abatement, PPE, report completed", tidyviol$ABATEDONE)%>%
gsub("E", "abatement, PPE, plan, report not completed, employer out of business")%>%
gsub("W", "abatement, PPE, plan, report not completed, worksite changed")%>%
gsub("S", "abatement, PPE, plan, report not complete, ad discretion")%>%
gsub("N", "national indicator (older files)")%>%
gsub("I", "abatement completed immediately upon receipt of citation")%>%
gsub("Q", "quick fix (fixed during the walkaround)")%>%
gsub("A", "abatement, PPE, plan, report not completed, ad Discretion")
```

```
## Warning in gsub(., "E", "abatement, PPE, plan, report not completed,
## employer out of business"): argument 'pattern' has length > 1 and only the
## first element will be used
```

```
#################################################################
# combine with tidydata above
tidydata<-left_join(tidydata, tidyviol, by="ACTIVITYNO")
head(tidydata)
```

```
##   ACTIVITYNO            ESTABNAME AGENCY      PLACE
## 1   10236776        DUBE DRY WALL   <NA> SUNDERLAND
## 2   10236776        DUBE DRY WALL   <NA> SUNDERLAND
## 3   10236776        DUBE DRY WALL   <NA> SUNDERLAND
## 4   10236776        DUBE DRY WALL   <NA> SUNDERLAND
## 5  103393633 KNOWLTON MACHINE CO.   <NA>      SALEM
## 6  103393633 KNOWLTON MACHINE CO.   <NA>      SALEM
##                              INDUSTRY       JOBTITLE NAICSTEXT DEGREE
```

```
## 1 PLASTERING, DRYWALL, AND INSULATION safety officer      <NA>   <NA>
## 2 PLASTERING, DRYWALL, AND INSULATION safety officer      <NA>   <NA>
## 3 PLASTERING, DRYWALL, AND INSULATION safety officer      <NA>   <NA>
## 4 PLASTERING, DRYWALL, AND INSULATION safety officer      <NA>   <NA>
## 5           INDUSTRIAL MACHINERY, NEC health officer      <NA>   <NA>
## 6           INDUSTRIAL MACHINERY, NEC health officer      <NA>   <NA>
##   BODYPART NATURE HUMAN SOURCE ENVIRON HAZSUB OCCUPATION CLOSE EMPHASIS
## 1     <NA>   <NA>  <NA>   <NA>    <NA>   <NA>       <NA>   Yes     <NA>
## 2     <NA>   <NA>  <NA>   <NA>    <NA>   <NA>       <NA>   Yes     <NA>
## 3     <NA>   <NA>  <NA>   <NA>    <NA>   <NA>       <NA>   Yes     <NA>
## 4     <NA>   <NA>  <NA>   <NA>    <NA>   <NA>       <NA>   Yes     <NA>
## 5     <NA>   <NA>  <NA>   <NA>    <NA>   <NA>       <NA>   Yes     <NA>
## 6     <NA>   <NA>  <NA>   <NA>    <NA>   <NA>       <NA>   Yes     <NA>
##   GRAVITYLEVEL INSTANCES NUMEXPOSED
## 1         <NA>         1          0
## 2         <NA>         1          0
## 3         <NA>         1          0
## 4         <NA>         1          0
## 5           07         1         10
## 6           08         1         10
##                                                 ABATEDONE HAZCAT
## 1 abatement, PPE, plan, report not completed, ad Discretion   <NA>
## 2 abatement, PPE, plan, report not completed, ad Discretion   <NA>
## 3 abatement, PPE, plan, report not completed, ad Discretion   <NA>
## 4 abatement, PPE, plan, report not completed, ad Discretion   <NA>
## 5 abatement, PPE, plan, report not completed, ad Discretion   <NA>
## 6 abatement, PPE, plan, report not completed, ad Discretion   <NA>
##   VIOLTYPE STATE                                     STD_LOOKUP
## 1  serious    FE                                    SCAFFOLDING
## 2    other    FE                ELECTRICAL, GENERAL INTRODUCTION
## 3    other    FE                       ELECTRICAL, APPLICABILITY
## 4    other    FE                       ELECTRICAL, APPLICABILITY
## 5  serious    FE ASBESTOS, TREMOLITE, ANTHOPHYLLITE & ACTINOLITE
## 6  serious    FE ASBESTOS, TREMOLITE, ANTHOPHYLLITE & ACTINOLITE
##   RELATEDEVENT
## 1         <NA>
## 2         <NA>
## 3         <NA>
## 4         <NA>
## 5    complaint
## 6    complaint
```

```
ncol(tidydata)
```

```
## [1] 26
```

```
nrow(tidydata)
```

```
## [1] 308651
```

```
# So I combine scc sic fda naics osha accid hazsub acc occ viol std


#################################################################
# checking duplicated rows
head(duplicated(tidydata, incomparables = FALSE))
```

```
## [1] FALSE FALSE FALSE  TRUE FALSE FALSE
```

```
# remove all duplicated rows
tidydata<-distinct(tidydata)
# save them
save(tidydata, file="tidydata.Rdata")



#################################################################
load("tidydata.Rdata")
###Play with data
library(data.table)
library(ggplot2)
#################################################################
require(knitr)
```

```
## Loading required package: knitr
```

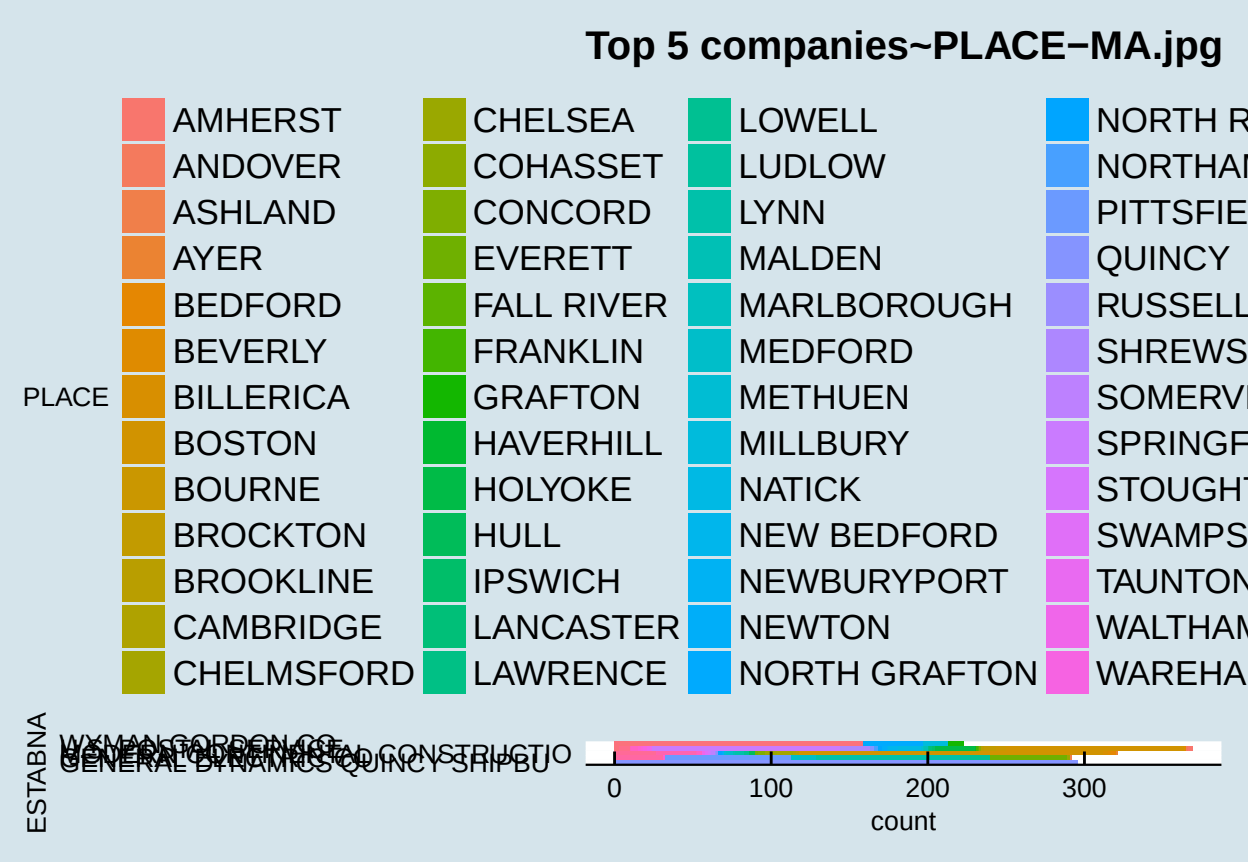
```
require(ggthemes)
```

```
## Loading required package: ggthemes
```

```
## Find the top 5 estabname and places associated with it
names(summary(tidydata$ESTABNAME))[1:5]
```

```
## [1] "U.S. POSTAL SERVICE"            "MODERN CONTINENTAL CONSTRUCTIO"
## [3] "GENERAL DYNAMICS QUINCY SHIPBU" "GENERAL ELECTRIC CO"
## [5] "WYMAN GORDON CO"
```

```
# Find the top 1 estabname in tidydata and which places consists it
sub<-filter(tidydata, ESTABNAME%in%names(summary(tidydata$ESTABNAME))[1:5])
g<-ggplot(sub, aes(ESTABNAME, fill=PLACE))+theme(axis.text.y=element_text(size=8))+coord_flip()+theme_e
g+geom_bar(position = "stack") + ggtitle("Top 5 companies~PLACE-MA.jpg")
```



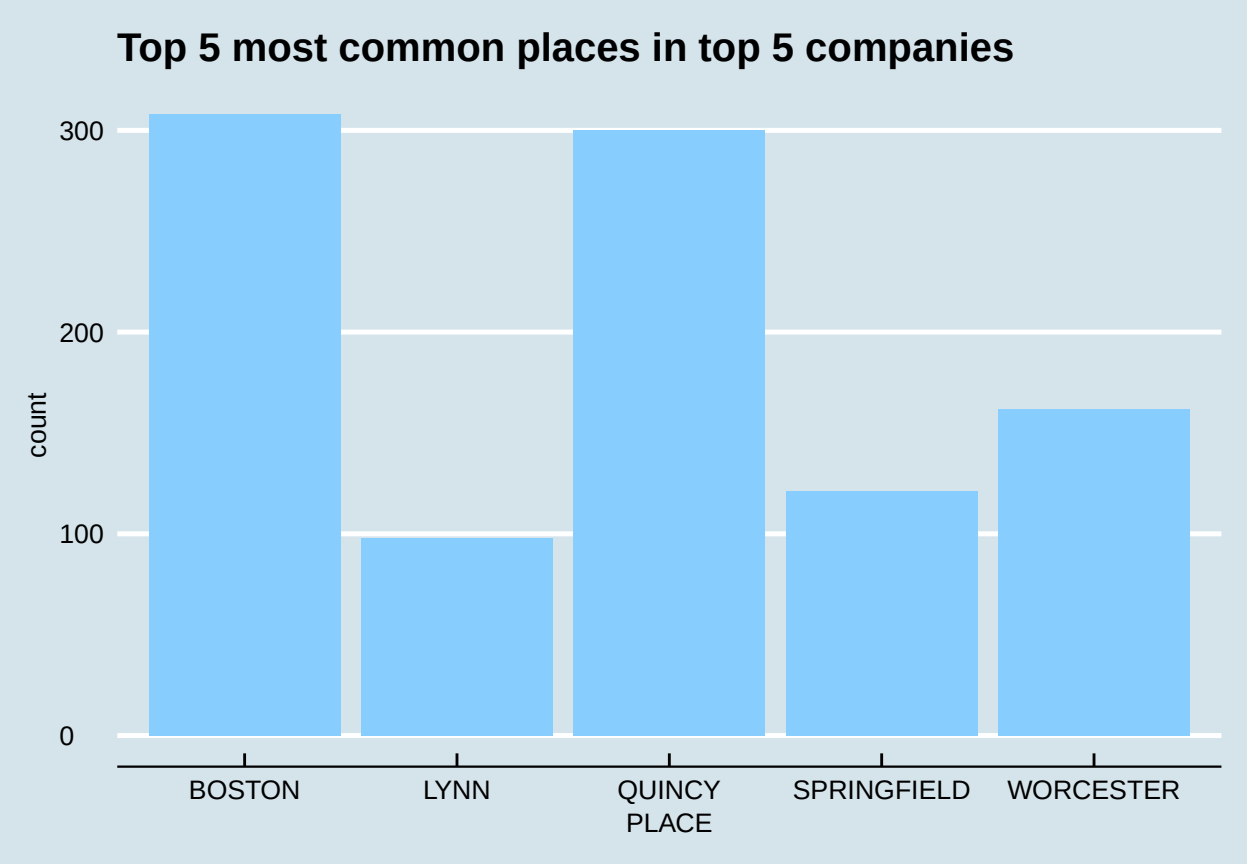

```
png(filename="Top 5 companies consists of places in MA")
plot(g+geom_bar(position = "stack") + ggtitle("Top 5 companies~PLACE-MA"))
dev.off()
```

```
## pdf
##   2
```

```
## ZOOM IN: if we only look at top 5 places in top 5 companies:
names(summary(sub$PLACE))[1:5]
```

```
## [1] "BOSTON"      "QUINCY"      "WORCESTER"   "SPRINGFIELD" "LYNN"
```

```
subsub<-filter(sub, PLACE%in%names(summary(sub$PLACE))[1:5])
s<-ggplot(subsub, aes(PLACE), fill=..count..)+theme(axis.text=element_text(size=10))+theme_economist()+s
s+geom_bar(fill="skyblue1")+ggtitle("Top 5 most common places in top 5 companies")
```



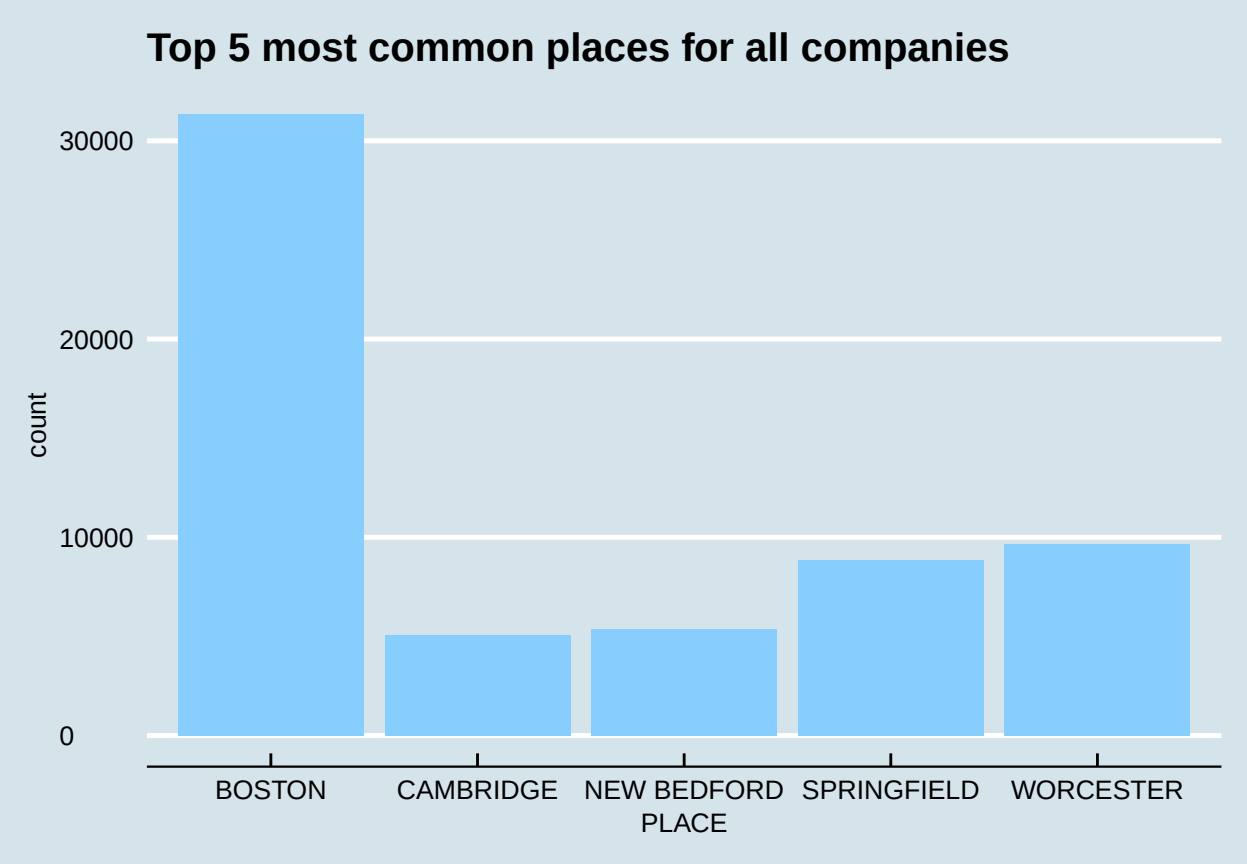

```
png(filename="Top 5 most common places in top 5 companies")
plot(s+geom_bar(fill="skyblue1")+ggtitle("Top 5 most common places in top 5 companies"))
dev.off()
```

```
## pdf
##   2
```

```
## Quincy Boston Lynn Springfield Worcester
## Let's look at a bigger picture
## How about the top 5 most common places for all companies
names(summary(tidydata$PLACE))[1:5]
```

```
## [1] "BOSTON"      "WORCESTER"   "SPRINGFIELD" "NEW BEDFORD" "CAMBRIDGE"
```

```
sub2<-filter(tidydata, PLACE%in%names(summary(tidydata$PLACE))[1:5])
a<-ggplot(sub2, aes(PLACE), fill=..count..)+theme(axis.text=element_text(size=8))+theme_economist()+scal
a+geom_bar(fill="skyblue1") + ggtitle("Top 5 most common places for all companies")
```




```
png(filename="Top 5 most common places for all companies in osha")
plot(a+geom_bar(fill="skyblue1") + ggtitle("Top 5 most common places for all companies"))
dev.off()
```

```
## pdf
##   2
```

```
# Boston Worcester Springfield New Bedford Cambridge

# Boston is the most dangerous place in MA
# So let's look at which industries(jobs) are most dangerous(easy to have accidents) for MA and Boston

## Top 10 most accid prone industries in MA
names(summary(tidydata$INDUSTRY))[1:10]
```

```
##  [1] "NONRESIDENTIAL CONSTRUCTION, NEC"
##  [2] "ELECTRICAL WORK"
##  [3] "PLUMBING, HEATING, AIR-CONDITIONING"
##  [4] "WATER, SEWER, AND UTILITY LINES"
##  [5] "PLASTICS PRODUCTS, NEC"
##  [6] "STRUCTURAL STEEL ERECTION"
##  [7] "MASONRY AND OTHER STONE WORK"
##  [8] "RESIDENTIAL CONSTRUCTION, NEC"
##  [9] "SPECIAL TRADE CONTRACTORS, NEC"
## [10] "CARPENTRY WORK"
```

```
tmptmp<-filter(tidydata, INDUSTRY==names(summary(tidydata$INDUSTRY))[1:10])
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## Warning in `==.default`(structure(c(912L, 912L, 912L, 561L, 561L, 561L, :
## longer object length is not a multiple of shorter object length
```

```
t<-ggplot(tmptmp, aes(INDUSTRY))+theme(axis.text = element_text(size=8))+coord_flip()+theme_economist()+
t+geom_bar(fill="skyblue3")+ggtitle("Top 10 most accident prone industries in MA")
```


```
png(filename="Top 10 most accident prone industries in MA")
plot(t+geom_bar(fill="skyblue3")+ggtitle("Top 10 most accident prone industries in MA"))
dev.off()
```

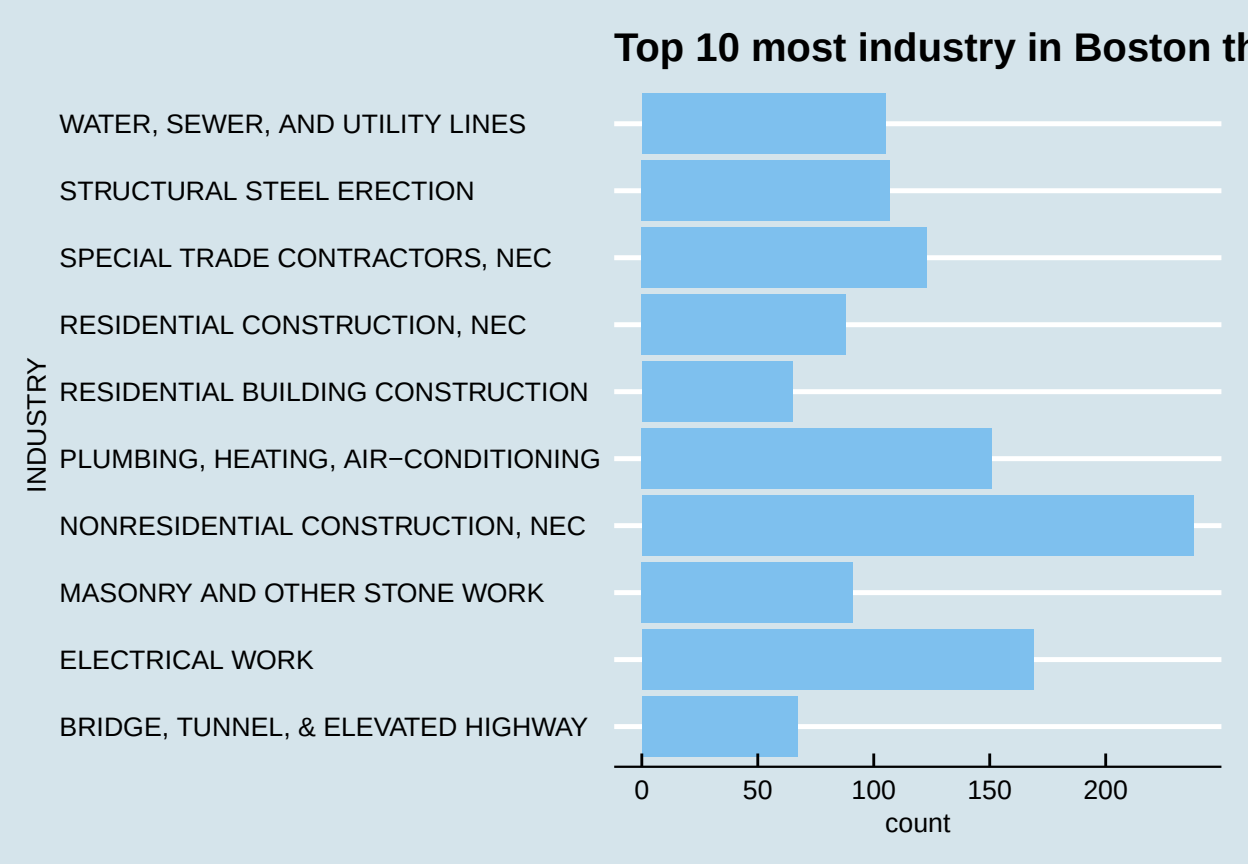
```
## pdf
##   2
```

```
## Top 10 industry in Boston that is most dangerous
tmp<-filter(tidydata, PLACE=="BOSTON")
tmp<-filter(tmp, INDUSTRY==names(summary(tmp$INDUSTRY)[1:10]))
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## Warning in `==.default`(structure(c(1079L, 1191L, 428L, 430L, 430L, 430L, :
## longer object length is not a multiple of shorter object length
```

```
m<-ggplot(tmp, aes(INDUSTRY))+coord_flip()+theme_economist()+scale_colour_economist()
m+geom_bar(fill="skyblue2")+ggtitle("Top 10 most industry in Boston that are most dangerous")
```

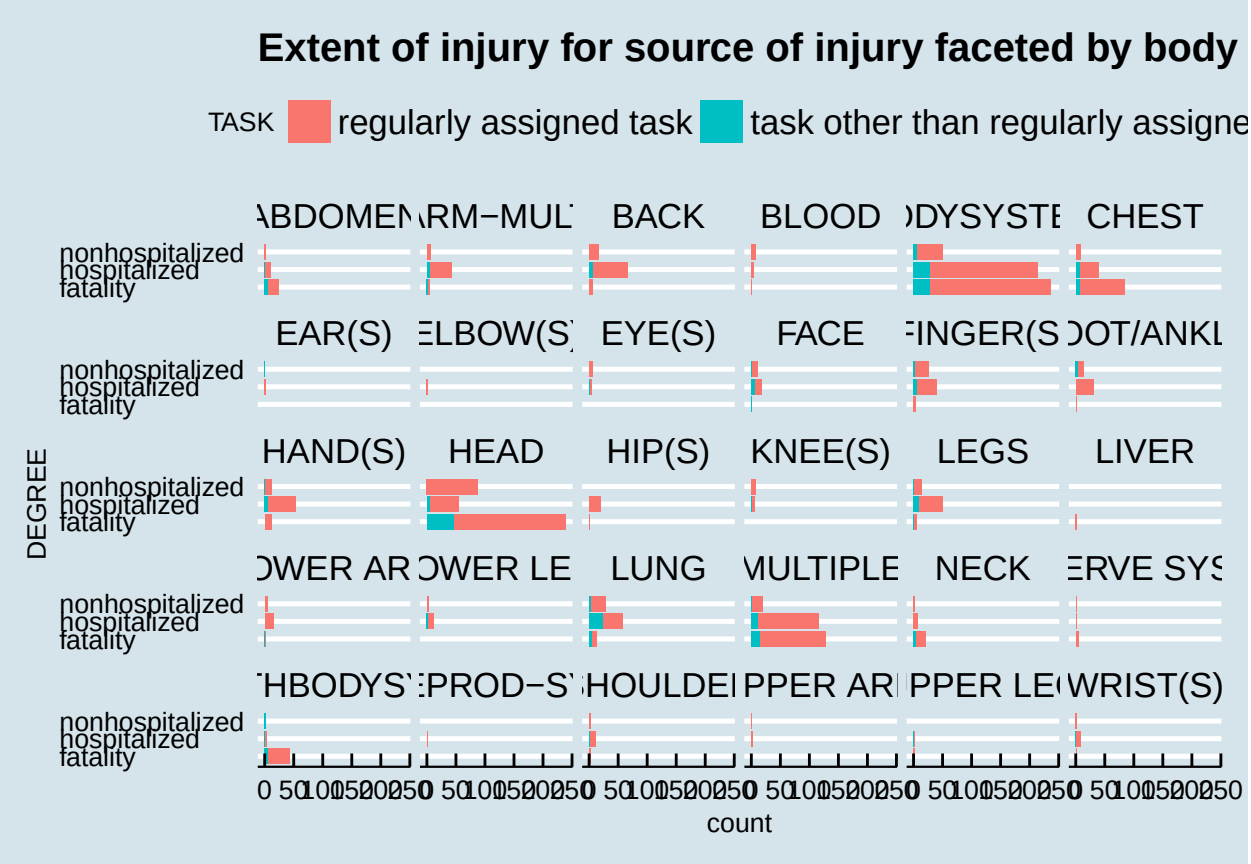



```
png(filename="Top 10 most industry in Boston that are most dangerous")
dev.off()
```

```
## pdf
##   2
```

```
# As we can see, nonresidential construction in boston and the whole MA is most dangerous

## degree(extent of injury) type for source of injury in tidyaccid faced by BODYPART
accid2<-subset(tidyaccid, TASK!=0, DEGREE!=0)
ggplot(accid2, aes(DEGREE, fill=TASK))+geom_bar()+ facet_wrap(~ BODYPART)+ theme(axis.text = element_te
```

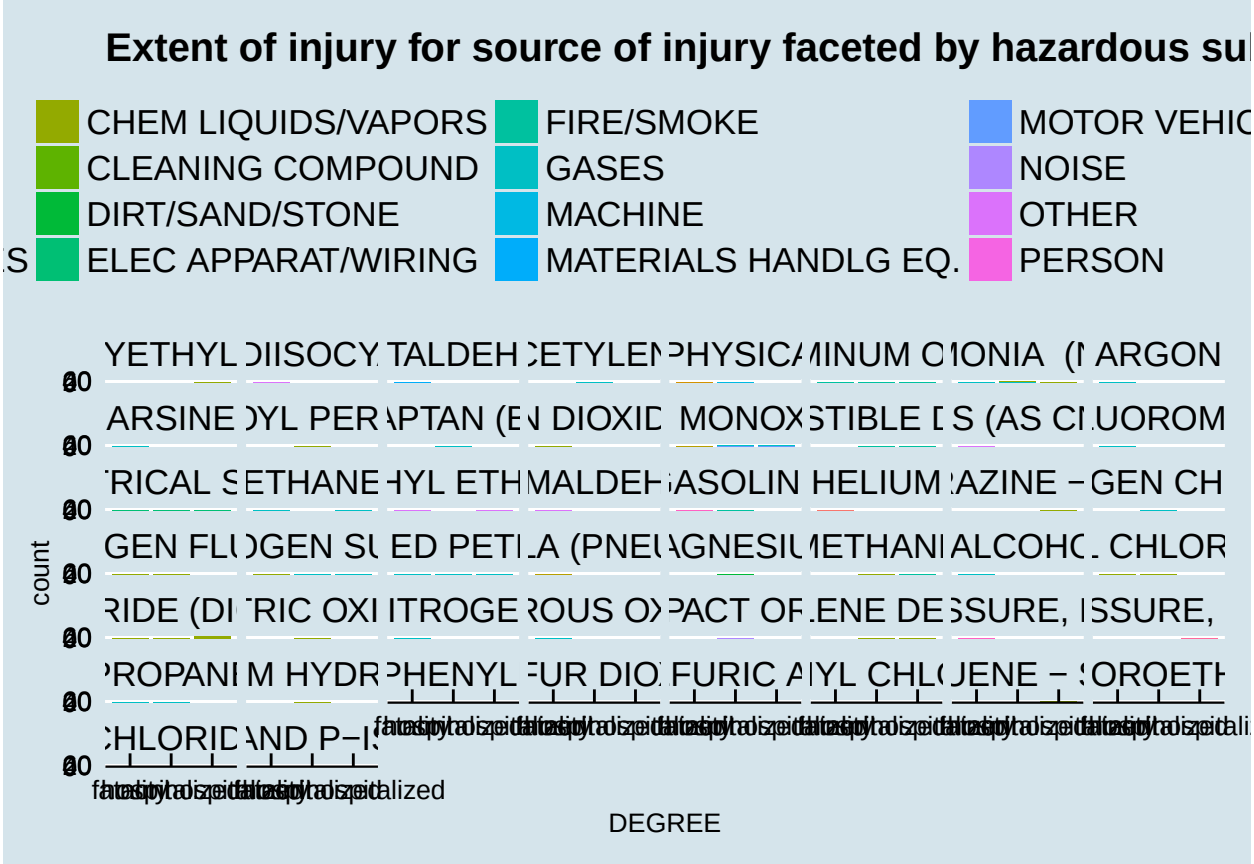



```
png(filename="Extent of injury for source of injury faceted by body part")
plot(ggplot(accid2, aes(DEGREE, fill=TASK))+geom_bar()+ facet_wrap(~ BODYPART)+ theme(axis.text = eleme
dev.off()
```

```
## pdf
##   2
```

```
# As we can see regularly assigned test causes more injury, no matter fatality, hospitalized, nonhospit
# which I didn't expect since my common sense is always that task other than regularly assigned causes
# since people are always good at what they usually do.


## Extent of injury for source of injury faceted by hazardous substance
accid3<-filter(tidyaccid, is.na(HAZSUB)==F, is.na(SOURCE)==F, is.na(DEGREE)==F)
ggplot(accid3, aes(DEGREE, fill=SOURCE))+geom_bar()+facet_wrap(~HAZSUB)+theme(axis.text = element_text(
```




```
png(filename="Extent of injury for source of injury faceted by hazardous substance")
plot(ggplot(accid3, aes(DEGREE, fill=SOURCE))+geom_bar()+facet_wrap(~HAZSUB)+theme(axis.text = element_
dev.off()
```

```
## pdf
##   2
```

```
## As we can see, METHYLENE CHLORIDE (DICHLOROMETHANE) causes most nonhospitalized injury.
```