Code to Extract the ABC file

The code below can be used to extrat the ABC file base on data length. The data is located here https://www.abc.ca.gov/datport/DataExport.html

Download zip file with the data

```
rm(list=ls(all=TRUE)) #start with empty workspace
startTime <- Sys.time()

temp <- tempfile()

download.file("http://www.abc.ca.gov/datport/ABC_Data_Export.zip",temp)
con <- unz(temp, "m_tape460.LST")
data <-  read.fwf(con, fill = TRUE, comment.char = "",
                  widths = c(2,8,3,8,11,11,8,3,1,2,4,2,
                             50,50,50,25,2,10,50,50,50,25,
                             2,10,16,7))
unlink(temp)

# add the variable names to the file

names(data) <- c("License Type","File Number","License Or Application",
                 "Type Status","Type Orig Issue Dates","Expiration Dates",
                 "Fee Codes","Duplicate Counts","Master Indicator",
                 "Term in Months","Geo Code","District Office",
                 "Primary Name","Premise Street Address1",
                 "Premise Street Address2","Premise City",
                 "Premise State","Premise Zip","DBA name",
                 "Mail Street Address1","Mail Street Address2",
                 "Mail City","Mail State","Mail Zip",
                 "Premise County","Premise Census Tract")

# take a look at the file

summary(data)
```

```
##   License Type   File Number     License Or Application   Type Status
##  Min.   : 1.0   Min.   :     4   APP: 10720             ACTIVE  :102929
##  1st Qu.:20.0   1st Qu.:400302   LIC:101098             PEND    :  5977
##  Median :41.0   Median :483603                          R64B    :    89
##  Mean   :34.8   Mean   :448299                          REVPEN  :   136
##  3rd Qu.:47.0   3rd Qu.:536870                          SLMSHD  :     2
##  Max.   :86.0   Max.   :568085                          SUREND  :  2495
##                                                        SUSPEN  :   190
##  Type Orig Issue Dates    Expiration Dates    Fee Codes
##              :10819     30-JUN-2016:20603   P40     :71013
##  01-JAN-1994:  557                 :10806   P0      :26626
##  22-JUN-2009:  371     31-MAY-2016: 7910   P20     : 8850
```

```
##  08-NOV-2013:  237     31-JUL-2016: 7467   PO-GL5K : 2220
##  20-MAY-2008:  171     28-FEB-2017: 7451   PO-GL20K:  721
##  18-JUL-2014:  136     31-AUG-2016: 7441   P40-GL5K:  655
##  (Other)    :99527     (Other)    :50140   (Other) : 1733
##  Duplicate Counts Master Indicator Term in Months    Geo Code
##  Min.   :  0.00   N:13788          Min.   :12     Min.   : 100
##  1st Qu.:  1.00   Y:98030          1st Qu.:12     1st Qu.:1933
##  Median :  1.00                    Median :12     Median :3100
##  Mean   :  1.41                    Mean   :12     Mean   :3058
##  3rd Qu.:  1.00                    3rd Qu.:12     3rd Qu.:3902
##  Max.   :583.00                    Max.   :12     Max.   :9999
##  NA's   :98159                                    NA's   :117
##  District Office
##  Min.   : 2.00
##  1st Qu.: 7.00
##  Median :13.00
##  Mean   :16.55
##  3rd Qu.:24.00
##  Max.   :75.00
##  NA's   :1
##                                          Primary Name
##  7 ELEVEN INC                                  :  1618
##  GARFIELD BEACH CVS LLC                        :   869
##  SAFEWAY INC                                   :   594
##  THRIFTY PAYLESS INC                           :   589
##  SMART & FINAL STORES LLC                      :   452
##  WAL MART STORES INC                           :   450
##  (Other)                                       :107246
##                                       Premise Street Address1
##                                                :   117
##  930 MCLAUGHLIN AVE                            :    88
##  20580 8TH ST E                                :    64
##  1275 COMMERCE BLVD                            :    61
##  875 HANNA DR                                  :    60
##  SF INTL AIRPORT                               :    53
##  (Other)                                       :111375
##                                       Premise Street Address2
##                                                :90957
##  STE A                                         : 1645
##  STE B                                         :  943
##  STE C                                         :  600
##  STE 100                                       :  547
##  STE D                                         :  491
##  (Other)                                       :16635
##                     Premise City   Premise State        Premise Zip
##  LOS ANGELES              : 5815   CA     :110272   92101     :   489
##  SAN FRANCISCO            : 5286   WA     :   255   94558     :   329
##  SAN DIEGO                : 4349          :   247             :   247
##  NAPA                     : 2330   OR     :   233   93940     :   240
##  SAN JOSE                 : 2118   NY     :   125   94574     :   225
##  SACRAMENTO               : 2049   TX     :    82   94133     :   220
##  (Other)                  :89871   (Other):   604   (Other)   :110068
##                                                     DBA name
##                                                :  5349
```

```
##  ROUND TABLE PIZZA                                :   324
##  GROCERY OUTLET                                   :   244
##  CHIPOTLE MEXICAN GRILL                           :   218
##  BEVMO                                            :   204
##  WHOLE FOODS MARKET                               :   195
##  (Other)                                          :105284
##                                          Mail Street Address1
##                                                   :51828
##  PO BOX 219088                                    : 1603
##  PO BOX 29096                                     :  885
##  1 CVS DR                                         :  848
##  2600 CAPITOL AVE                                 :  581
##  702 SW 8TH ST                                    :  483
##  (Other)                                          :55590
##                                          Mail Street Address2
##                                                   :90202
##  ATT: 7 ELEVEN LICENSING                          : 1598
##  STE 300                                          : 1251
##  MAIL DROP 23062A                                 :  854
##  STE 100                                          :  787
##  MAIL STOP #6516                                  :  589
##  (Other)                                          :16537
##                      Mail City       Mail State          Mail Zip
##                           :51851          :51851             :51851
##  LOS ANGELES              : 2921   CA     :47066   75221-9088: 1612
##  SAN FRANCISCO            : 2021   TX     : 2863   85038-9096:  885
##  DALLAS                   : 2020   AZ     : 1308   02895-6146:  870
##  SAN DIEGO                : 1510   RI     :  894   95816-5930:  584
##  SACRAMENTO               : 1426   IL     :  763   90040-1562:  452
##  (Other)                  :50069   (Other): 7073   (Other)   :55564
##           Premise County  Premise Census Tract
##  LOS ANGELES     :22497   Min.   :   1.0
##  SAN DIEGO       : 8802   1st Qu.:  86.0
##  ORANGE          : 7565   Median : 525.2
##  SAN FRANCISCO   : 5240   Mean   :1834.2
##  SANTA CLARA     : 4721   3rd Qu.:3131.0
##  ALAMEDA         : 4649   Max.   :9832.0
##  (Other)         :58344   NA's   :2309
```

```
str(data)
```

```
## 'data.frame':    111818 obs. of  26 variables:
##  $ License Type           : int  20 20 41 41 9 18 17 21 21 47 ...
##  $ File Number            : int  100196 100275 100297 100298 100306 100306 100306 100409 100472 1006
##  $ License Or Application : Factor w/ 2 levels "APP","LIC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Type Status            : Factor w/ 7 levels "ACTIVE  ","PEND    ",..: 6 1 1 1 1 1 1 1 1 1 ...
##  $ Type Orig Issue Dates  : Factor w/ 9004 levels "          ",..: 6037 5064 2562 1498 5537 7241 553
##  $ Expiration Dates       : Factor w/ 26 levels "          ",..: 6 23 16 16 6 6 6 12 2 14 ...
##  $ Fee Codes              : Factor w/ 28 levels "        ","GL20K   ",..: 28 28 12 12 12 12 12 28 12
##  $ Duplicate Counts       : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ Master Indicator       : Factor w/ 2 levels "N","Y": 2 2 2 2 1 2 2 2 2 2 ...
##  $ Term in Months         : int  12 12 12 12 12 12 12 12 12 12 ...
##  $ Geo Code               : int  3309 3405 4902 1701 1401 1401 1401 1933 1008 3711 ...
##  $ District Office        : int  8 23 27 27 6 6 6 4 21 9 ...
```

```
##  $ Primary Name           : Factor w/ 64260 levels "
##  $ Premise Street Address1: Factor w/ 76750 levels "
##  $ Premise Street Address2: Factor w/ 4299 levels "
##  $ Premise City           : Factor w/ 1948 levels "                          ",..: 803 1506 739 921 15
##  $ Premise State          : Factor w/ 50 levels "  ","AK","AL",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Premise Zip            : Factor w/ 37882 levels "          ","01007     ",..: 12570 36084 33277 33
##  $ DBA name               : Factor w/ 69399 levels "
##  $ Mail Street Address1   : Factor w/ 21330 levels "
##  $ Mail Street Address2   : Factor w/ 2170 levels "
##  $ Mail City              : Factor w/ 1841 levels "                          ",..: 1 1 1476 1476 1 1 1
##  $ Mail State             : Factor w/ 50 levels "  ","AK","AL",..: 1 1 6 6 1 1 1 1 1 6 ...
##  $ Mail Zip               : Factor w/ 13043 levels "          ","01085-4596",..: 1 1 10479 10479 1 1
##  $ Premise County         : Factor w/ 59 levels "               ",..: 34 35 50 18 15 15 15 20 11 38
##  $ Premise Census Tract   : num  455 38 1539 4 4 ...
```

```
# extract county codes

library(RCurl)
```

```
## Loading required package: bitops
```

```
library(dplyr)
```

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

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

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

```
linkToCountyCodes <- getURL("http://www2.census.gov/geo/docs/reference/codes/files/st06_ca_cou.txt")

CountyCodes <- read.csv(text = linkToCountyCodes,
                        header = FALSE,
                        col.names = c("State",
                                      "StateCode",
                                      "CountyCode",
                                      "CountyName",
                                      "ExtraCode")) %>%
  mutate(CountyNameCleaned = gsub(" County",
                                  "",
                                  CountyName),
         CountyNameCleaned =  toupper(CountyNameCleaned),
         CountyCodePadded = stringr::str_pad(CountyCode, 3,pad = "0"))

# join the sets

data <- data %>%
```

```
  mutate(PremiseCounty = stringr::str_trim( `Premise County`)) %>%
  left_join(CountyCodes,
            by = c(PremiseCounty = "CountyNameCleaned"))
```

Export the data to tab separated file for further analysis

```
write.table(data, file = "~/ABCdata.txt", sep="\t", row.names = FALSE)

# get the time
endTime <- Sys.time()
```

The analysis was completed on Mon Apr 04 12:09:17 PM 2016 in 1.1015768 minutes.