```r
# StudyArea.r

library(readr)
library(dplyr)
library(lubridate)
```

```r
# read.csv
d0 = read.csv("StudyArea.csv",header=TRUE)
str(d0)
```

```
## 'data.frame':    439362 obs. of  14 variables:
##  $ FID       : int  0 1 2 3 4 5 6 18 20 21 ...
##  $ ORGANIZATI: Factor w/ 6 levels "BIA","BLM","BOR",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ UNIT      : Factor w/ 160 levels "1","13230","13290",..: 127 127 127 127 127 127 127 127 127 127
##  $ SUBUNIT   : Factor w/ 434 levels " ","102","103",..: 283 283 283 283 283 283 283 283 283 283 ...
##  $ SUBUNIT2  : Factor w/ 463 levels "Alamosa National Wildlife Refuge",..: 363 363 363 363 363 363 3
##  $ FIRENAME  : Factor w/ 149644 levels " ","  ALLEN CANYON",..: 107539 63817 125775 81086 59542 6583
##  $ CAUSE     : Factor w/ 5 levels " ","Human","Natural",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ YEAR_     : int  2001 2002 2002 2001 1994 1994 1999 2003 2005 2005 ...
##  $ STARTDATED: Factor w/ 12648 levels "","1/1/00 0:00",..: 3 7804 8137 9357 11718 6433 3836 8952 108
##  $ CONTRDATED: Factor w/ 12644 levels "","1/1/00 0:00",..: 3 7801 8134 9354 11715 6432 3850 8949 108
##  $ OUTDATED  : Factor w/ 12605 levels "","1/1/00 0:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATE     : Factor w/ 11 levels "Arizona","California",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ STATE_FIPS: int  6 6 6 6 6 6 6 6 6 6 ...
##  $ TOTALACRES: num  0.1 3 0.5 0.1 1 0.1 3 0.1 0.1 0.1 ...
```

```r
#
# all character vars converted to categorical vars (factors)

# read_csv
df0 = read_csv("StudyArea.csv",col_names=TRUE)
```

```
## Parsed with column specification:
## cols(
##   FID = col_double(),
##   ORGANIZATI = col_character(),
##   UNIT = col_double(),
##   SUBUNIT = col_character(),
##   SUBUNIT2 = col_character(),
##   FIRENAME = col_character(),
##   CAUSE = col_character(),
##   YEAR_ = col_double(),
##   STARTDATED = col_character(),
##   CONTRDATED = col_character(),
##   OUTDATED = col_logical(),
##   STATE = col_character(),
##   STATE_FIPS = col_double(),
##   TOTALACRES = col_double()
## )
```

```
## Warning: 616745 parsing failures.
##  row      col           expected       actual            file
## 2685 OUTDATED 1/0/T/F/TRUE/FALSE 2/16/07 0:00 'StudyArea.csv'
## 2686 OUTDATED 1/0/T/F/TRUE/FALSE 2/2/07 0:00  'StudyArea.csv'
## 2687 OUTDATED 1/0/T/F/TRUE/FALSE 1/5/07 0:00  'StudyArea.csv'
```

```
## 2688 OUTDATED 1/0/T/F/TRUE/FALSE 3/26/07 0:00 'StudyArea.csv'
## 2689 OUTDATED 1/0/T/F/TRUE/FALSE 3/23/07 0:00 'StudyArea.csv'
## .... ........ .................. ............ ..............
## See problems(...) for more details.
```

```
#
# character variables not converted

# rows with problems
problems(df0)
```

```
## # A tibble: 616,745 x 5
##      row col      expected           actual       file
##    <int> <chr>    <chr>              <chr>        <chr>
##  1  2685 OUTDATED 1/0/T/F/TRUE/FALSE 2/16/07 0:00 'StudyArea.csv'
##  2  2686 OUTDATED 1/0/T/F/TRUE/FALSE 2/2/07 0:00  'StudyArea.csv'
##  3  2687 OUTDATED 1/0/T/F/TRUE/FALSE 1/5/07 0:00  'StudyArea.csv'
##  4  2688 OUTDATED 1/0/T/F/TRUE/FALSE 3/26/07 0:00 'StudyArea.csv'
##  5  2689 OUTDATED 1/0/T/F/TRUE/FALSE 3/23/07 0:00 'StudyArea.csv'
##  6  2690 OUTDATED 1/0/T/F/TRUE/FALSE 4/7/07 0:00  'StudyArea.csv'
##  7  2691 OUTDATED 1/0/T/F/TRUE/FALSE 3/31/07 0:00 'StudyArea.csv'
##  8  2692 OUTDATED 1/0/T/F/TRUE/FALSE 4/3/07 0:00  'StudyArea.csv'
##  9  2693 OUTDATED 1/0/T/F/TRUE/FALSE 2/22/07 0:00 'StudyArea.csv'
## 10  2694 OUTDATED 1/0/T/F/TRUE/FALSE 2/28/07 0:00 'StudyArea.csv'
## # ... with 616,735 more rows
```

```
d1 = problems(df0)

# cols with problems
table(d1$col)
```

```
##
## OUTDATED     UNIT
##   420003   196742
```

```
#
# all problems in cols OUTDATED, UNIT

# fix by converting to character type
df0 = read_csv("StudyArea.csv",col_types = list(UNIT = col_character(),
                                                OUTDATED = col_character()),col_names=TRUE)
#
# no more problems
#
# see OUTDATED, UNIT, character cols
df2 = select(df0,"UNIT","OUTDATED")
head(df2)
```

```
## # A tibble: 6 x 2
##   UNIT  OUTDATED
##   <chr> <chr>
## 1 81682 <NA>
## 2 81682 <NA>
## 3 81682 <NA>
## 4 81682 <NA>
## 5 81682 <NA>
```

```
## 6 81682 <NA>
```

```
sum(is.na(df2$OUTDATED))   # [1] 19359
```

```
## [1] 19359
```

```
tail(df2)
```

```
## # A tibble: 6 x 2
##   UNIT  OUTDATED
##   <chr> <chr>
## 1 PWRO  7/4/12 0:00
## 2 PWRO  9/3/12 0:00
## 3 PWRO  9/2/12 0:00
## 4 PWRO  9/8/12 0:00
## 5 PWRO  8/24/82 0:00
## 6 PWRO  7/1/97 0:00
```

```
#
# column OUTDATED includes dates with times and 19359 NAs
#
sum(is.na(df2$UNIT))
```

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

```
#
# no NAs in column UNIT
#

#
# SELECT
#
# select firename, size, year
dfFires2 = select(df0,FIRENAME,TOTALACRES,YEAR_)
head(dfFires2)
```

```
## # A tibble: 6 x 3
##   FIRENAME   TOTALACRES YEAR_
##   <chr>           <dbl> <dbl>
## 1 PUMP HOUSE        0.1  2001
## 2 I5                3    2002
## 3 SOUTHBAY          0.5  2002
## 4 MARINA            0.1  2001
## 5 HILL              1    1994
## 6 IRRIGATION        0.1  1994
```

```
#
# select cols containing word DATE or starting with TOTAL
dfFires3 = select(df0,contains("DATE"),starts_with("TOTAL"))
head(dfFires3)
```

```
## # A tibble: 6 x 4
##   STARTDATED   CONTRDATED   OUTDATED TOTALACRES
##   <chr>        <chr>        <chr>         <dbl>
## 1 1/1/01 0:00  1/1/01 0:00  <NA>            0.1
## 2 5/3/02 0:00  5/3/02 0:00  <NA>            3
## 3 6/1/02 0:00  6/1/02 0:00  <NA>            0.5
## 4 7/12/01 0:00 7/12/01 0:00 <NA>            0.1
```

```
## 5 9/13/94 0:00 9/13/94 0:00 <NA>            1
## 6 4/22/94 0:00 4/22/94 0:00 <NA>            0.1
```

```
# select columns, rename
df = select(df0,NAME=FIRENAME,CAUSE,YEAR=YEAR_,STATE,ACRES=TOTALACRES)
head(df)
```

```
## # A tibble: 6 x 5
##   NAME       CAUSE  YEAR STATE      ACRES
##   <chr>      <chr> <dbl> <chr>      <dbl>
## 1 PUMP HOUSE Human  2001 California   0.1
## 2 I5         Human  2002 California   3
## 3 SOUTHBAY   Human  2002 California   0.5
## 4 MARINA     Human  2001 California   0.1
## 5 HILL       Human  1994 California   1
## 6 IRRIGATION Human  1994 California   0.1
```

```
# use quotation marks for the names if they have blank spaces

#
# COUNTING

# fires by CAUSE
table(df$CAUSE)
```

```
##
##         Human       Natural Undetermined       Unknown
##        194466        243486           169             9
```

```
#
# fires by STATE
table(df$STATE)
```

```
##
##    Arizona California   Colorado      Idaho    Montana     Nevada New Mexico
##      80625      90522      30928      36510      39209      21590      29619
##     Oregon       Utah Washington    Wyoming
##      52820      24862      20647      12030
```

```
# STATE abbreviations
state.name
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
## [21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
## [25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
## [29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
## [33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
## [37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
## [41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [49] "Wisconsin"      "Wyoming"
```

```
name = names(table(df$STATE))
name
```

```
##  [1] "Arizona"    "California" "Colorado"   "Idaho"      "Montana"
##  [6] "Nevada"     "New Mexico" "Oregon"     "Utah"       "Washington"
## [11] "Wyoming"
```

```
# match state names with abbreviations
match(name,state.name)
```

```
##  [1]  3  5  6 12 26 28 31 37 44 47 50
```

```
state.abb[match(name,state.name)]
```

```
##  [1] "AZ" "CA" "CO" "ID" "MT" "NV" "NM" "OR" "UT" "WA" "WY"
```

```
# barplot
d2 = table(df$STATE)
name = names(d2)
name2 = state.abb[match(name,state.name)]
barplot(d2,names.arg=name2,xlab="State",ylab="number of fires")

#
# FILTER
#
# large fires in CA
df25k = filter(df,ACRES > 25000,STATE == 'California')
nrow(df25k)
```

```
## [1] 126
```

```
# >1000-acre fires during 2016
df1k = filter(df,ACRES > 1000, YEAR == 2016)
nrow(df1k)
```

```
## [1] 150
```

```
# >1000-acre fires in 2010-2012
df1k = filter(df,ACRES>1000, YEAR%in% c(2010,2011,2012))
nrow(df1k)
```

```
## [1] 715
```

```
# largest fire in AZ
dfAZ = filter(df,STATE == 'Arizona')
which.max(dfAZ$ACRES)
```

```
## [1] 6354
```

```
#
# get the row
dfAZ[which.max(dfAZ$ACRES), ]
```

```
## # A tibble: 1 x 5
##   NAME   CAUSE  YEAR STATE    ACRES
##   <chr>  <chr> <dbl> <chr>    <dbl>
## 1 WALLOW Human  2011 Arizona 538049
```

```
#
# SORT
#
# sort large fires during 2016
df25k = filter(df,ACRES > 25000, YEAR == 2016)
```

```r
d2 = arrange(df25k,ACRES)
head(d2)
```

```
## # A tibble: 6 x 5
##   NAME        CAUSE    YEAR STATE       ACRES
##   <chr>       <chr>   <dbl> <chr>       <dbl>
## 1 Copper King Human    2016 Montana    28553
## 2 Cedar       Human    2016 California 29193.
## 3 Juniper     Natural  2016 Arizona    30641
## 4 Jack        Natural  2016 Arizona    33850
## 5 Cliff Creek Natural  2016 Wyoming    34313
## 6 Cherry Road Human    2016 Oregon     35194
```

```r
tail(d2)
```

```
## # A tibble: 6 x 5
##   NAME     CAUSE    YEAR STATE       ACRES
##   <chr>    <chr>   <dbl> <chr>       <dbl>
## 1 Maple    Natural  2016 Wyoming     45425
## 2 Cedar    Natural  2016 Arizona     45977
## 3 Erskine  Human    2016 California  48007
## 4 Range 12 Human    2016 Washington 171915
## 5 Junkins  Human    2016 Colorado   181320
## 6 PIONEER  Human    2016 Idaho      188404
```

```r
# sort descending
d2 = arrange(df1k,desc(ACRES))

#
# MUTATE
#
# add STARTDATE column
df1 = mutate(df,START=df0$STARTDATED)
head(df1)
```

```
## # A tibble: 6 x 6
##   NAME       CAUSE  YEAR STATE      ACRES START
##   <chr>      <chr> <dbl> <chr>      <dbl> <chr>
## 1 PUMP HOUSE Human  2001 California   0.1 1/1/01 0:00
## 2 I5         Human  2002 California   3   5/3/02 0:00
## 3 SOUTHBAY   Human  2002 California   0.5 6/1/02 0:00
## 4 MARINA     Human  2001 California   0.1 7/12/01 0:00
## 5 HILL       Human  1994 California   1   9/13/94 0:00
## 6 IRRIGATION Human  1994 California   0.1 4/22/94 0:00
```

```r
# filter for large fires due to human and nature
#df1 = filter(df1,ACRES>=1000 & CAUSE %in% c('Human','Natural'))
#head(df1)
#nrow(df1)
#table(df1$CAUSE)


# lubridate::yday() function to create column DOY, day of the year
df2 = mutate(df1,DOY = yday(as.Date(df1$START,format = '%m/%d/%y%H:%M')))
head(df2)
```

```
## # A tibble: 6 x 7
##   NAME       CAUSE  YEAR STATE      ACRES START          DOY
##   <chr>      <chr> <dbl> <chr>      <dbl> <chr>        <dbl>
## 1 PUMP HOUSE Human  2001 California   0.1 1/1/01 0:00      1
## 2 I5         Human  2002 California   3   5/3/02 0:00    123
## 3 SOUTHBAY   Human  2002 California   0.5 6/1/02 0:00    152
## 4 MARINA     Human  2001 California   0.1 7/12/01 0:00   193
## 5 HILL       Human  1994 California   1   9/13/94 0:00   256
## 6 IRRIGATION Human  1994 California   0.1 4/22/94 0:00   112
```

```
#
# SUMMARIZE
#
# create new column DECADE

# cut
d6=df1
aux = cut(df1$YEAR,breaks = c(0,  1980,   1990, 2000,  2010,3000),
                   labels = c("0","1980-1989","1990-1999","2000-2009","2010-2016"),
                   right=F)
d6$DECADE = aux
head(d6)
```

```
## # A tibble: 6 x 7
##   NAME       CAUSE  YEAR STATE      ACRES START        DECADE
##   <chr>      <chr> <dbl> <chr>      <dbl> <chr>        <fct>
## 1 PUMP HOUSE Human  2001 California   0.1 1/1/01 0:00  2000-2009
## 2 I5         Human  2002 California   3   5/3/02 0:00  2000-2009
## 3 SOUTHBAY   Human  2002 California   0.5 6/1/02 0:00  2000-2009
## 4 MARINA     Human  2001 California   0.1 7/12/01 0:00 2000-2009
## 5 HILL       Human  1994 California   1   9/13/94 0:00 1990-1999
## 6 IRRIGATION Human  1994 California   0.1 4/22/94 0:00 1990-1999
```

```
# mutate
d6 = mutate(df1,DECADE= ifelse(YEAR %in% 1980:1989,"1980-1989",
                        ifelse(YEAR %in% 1990:1999,"1990-1999",
                        ifelse(YEAR %in% 2000:2009,"2000-2009",
                        ifelse(YEAR %in% 2010:2016,"2010-2016","-99")))))
head(d6)
```

```
## # A tibble: 6 x 7
##   NAME       CAUSE  YEAR STATE      ACRES START        DECADE
##   <chr>      <chr> <dbl> <chr>      <dbl> <chr>        <chr>
## 1 PUMP HOUSE Human  2001 California   0.1 1/1/01 0:00  2000-2009
## 2 I5         Human  2002 California   3   5/3/02 0:00  2000-2009
## 3 SOUTHBAY   Human  2002 California   0.5 6/1/02 0:00  2000-2009
## 4 MARINA     Human  2001 California   0.1 7/12/01 0:00 2000-2009
## 5 HILL       Human  1994 California   1   9/13/94 0:00 1990-1999
## 6 IRRIGATION Human  1994 California   0.1 4/22/94 0:00 1990-1999
```

```
dim(d6)
```

```
## [1] 439362      7
```

```
#
# group dataframe by DECADE
grp = group_by(d6,DECADE)
```

```r
class(grp)
```

```
## [1] "grouped_df" "tbl_df"     "tbl"        "data.frame"
```

```r
#
# grp is called a 'grouped dataframe'
#
dim(grp)
```

```
## [1] 439362      7
```

```r
head(grp)
```

```
## # A tibble: 6 x 7
## # Groups:   DECADE [2]
##   NAME        CAUSE  YEAR STATE      ACRES START        DECADE
##   <chr>       <chr> <dbl> <chr>      <dbl> <chr>        <chr>
## 1 PUMP HOUSE Human  2001 California   0.1 1/1/01 0:00  2000-2009
## 2 I5         Human  2002 California   3   5/3/02 0:00  2000-2009
## 3 SOUTHBAY   Human  2002 California   0.5 6/1/02 0:00  2000-2009
## 4 MARINA     Human  2001 California   0.1 7/12/01 0:00 2000-2009
## 5 HILL       Human  1994 California   1   9/13/94 0:00 1990-1999
## 6 IRRIGATION Human  1994 California   0.1 4/22/94 0:00 1990-1999
```

```r
#
# avg size of wildfires by decade
sm = summarize(grp,mean(ACRES))
sm
```

```
## # A tibble: 4 x 2
##   DECADE    `mean(ACRES)`
##   <chr>             <dbl>
## 1 1980-1989          154.
## 2 1990-1999          115.
## 3 2000-2009          241.
## 4 2010-2016          302.
```

```r
# avg size of wildfires by decade
sm = summarize(grp,AVG=mean(ACRES),MAX=max(ACRES))
sm
```

```
## # A tibble: 4 x 3
##   DECADE      AVG     MAX
##   <chr>     <dbl>   <dbl>
## 1 1980-1989  154. 427680
## 2 1990-1999  115. 231389
## 3 2000-2009  241. 590620
## 4 2010-2016  302. 558198.
```

```r
#
# using tapply
tapply(d6$ACRES,d6$DECADE,mean)
```

```
## 1980-1989 1990-1999 2000-2009 2010-2016
##  153.8853  114.7012  241.2522  302.1051
```

```r
#
# using aggregate
```

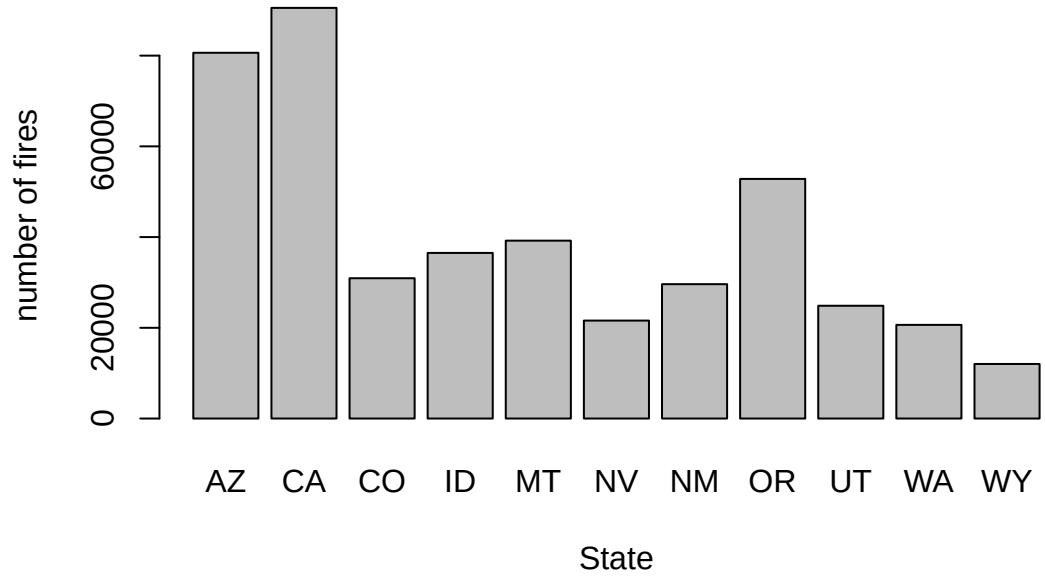
```r
aggregate(d6$ACRES,by=list(decade = d6$DECADE),FUN=mean,na.rm=TRUE)
```

```
##      decade        x
## 1 1980-1989 153.8853
## 2 1990-1999 114.7012
## 3 2000-2009 241.2522
## 4 2010-2016 302.1051
```

```r
# avg and max size of wildfies by decade
aggregate(d6$ACRES,by=list(decade = d6$DECADE),FUN=function(x) c(AVG=mean(x),MAX=max(x)))
```

```
##      decade       x.AVG       x.MAX
## 1 1980-1989    153.8853 427680.0000
## 2 1990-1999    114.7012 231389.0000
## 3 2000-2009    241.2522 590620.0000
## 4 2010-2016    302.1051 558198.3000
```

```r
#
# PLOT
#
library(ggplot2)
```



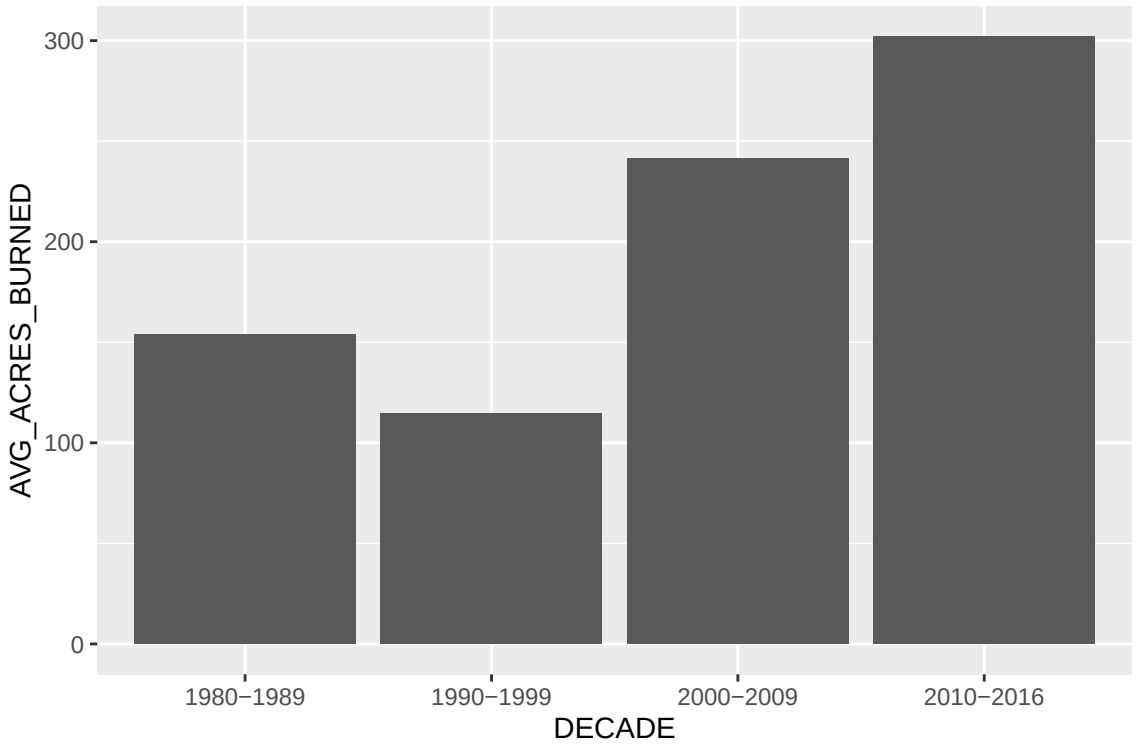

```r
# rename cols
names(sm) = c('DECADE','AVG_ACRES_BURNED')
```

```
## Warning: The `names` must have length 3, not 2.
## This warning is displayed once per session.
```

```r
head(sm)
```

```
## # A tibble: 4 x 3
##   DECADE    AVG_ACRES_BURNED      NA
##   <chr>                <dbl>   <dbl>
## 1 1980-1989             154. 427680
## 2 1990-1999             115. 231389
## 3 2000-2009             241. 590620
## 4 2010-2016             302. 558198.
```

```
#
# barplot
ggplot(data=sm) + geom_col(mapping=aes(x=DECADE,y=AVG_ACRES_BURNED))
```




```
#
# PIPES
#
# use pipeline to create a column for the day of the year
#
df = read_csv("StudyArea.csv",col_types = list(UNIT = col_character(),
                                               OUTDATED = col_character()),col_names=TRUE)%>%
  select(STATE,TOTALACRES,CAUSE,STARTDATED) %>%
  filter(TOTALACRES >= 1000 & CAUSE %in% c('Human','Natural')) %>%
  mutate(DOY = yday(as.Date(STARTDATED,format = '%m/%d/%y %H:%M')))
head(df)
```

```
## # A tibble: 6 x 5
##   STATE   TOTALACRES CAUSE STARTDATED     DOY
##   <chr>        <dbl> <chr> <chr>        <dbl>
## 1 Arizona       1500 Human 3/26/88 0:00    86
## 2 Arizona      10390 Human 5/15/86 0:00   135
## 3 Montana       1400 Human 6/27/86 0:00   178
## 4 Arizona       1035 Human 2/28/02 0:00    59
## 5 Arizona       5700 Human 4/9/00 0:00    100
## 6 Arizona       2750 Human 5/14/00 0:00   135
```