Homework 2 solution

```
library(readr)
library(dplyr)
library(ggplot2)
library(lubridate)

df0 = read_csv("StudyArea.csv", col_types = list(UNIT = col_character()),col_names = T)
```

```
## Warning: 420003 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## 2685 OUTDATED 1/0/T/F/TRUE/FALSE 2/16/07 0:00 'StudyArea.csv'
## 2686 OUTDATED 1/0/T/F/TRUE/FALSE 2/2/07 0:00  'StudyArea.csv'
## 2687 OUTDATED 1/0/T/F/TRUE/FALSE 1/5/07 0:00  'StudyArea.csv'
## 2688 OUTDATED 1/0/T/F/TRUE/FALSE 3/26/07 0:00 'StudyArea.csv'
## 2689 OUTDATED 1/0/T/F/TRUE/FALSE 3/23/07 0:00 'StudyArea.csv'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.
```

```
df = df0 %>%
     select(CAUSE,YEAR_,ORGANIZATI,STARTDATED,STATE,TOTALACRES) %>%
     filter(TOTALACRES >= 1000)
head(df)
```

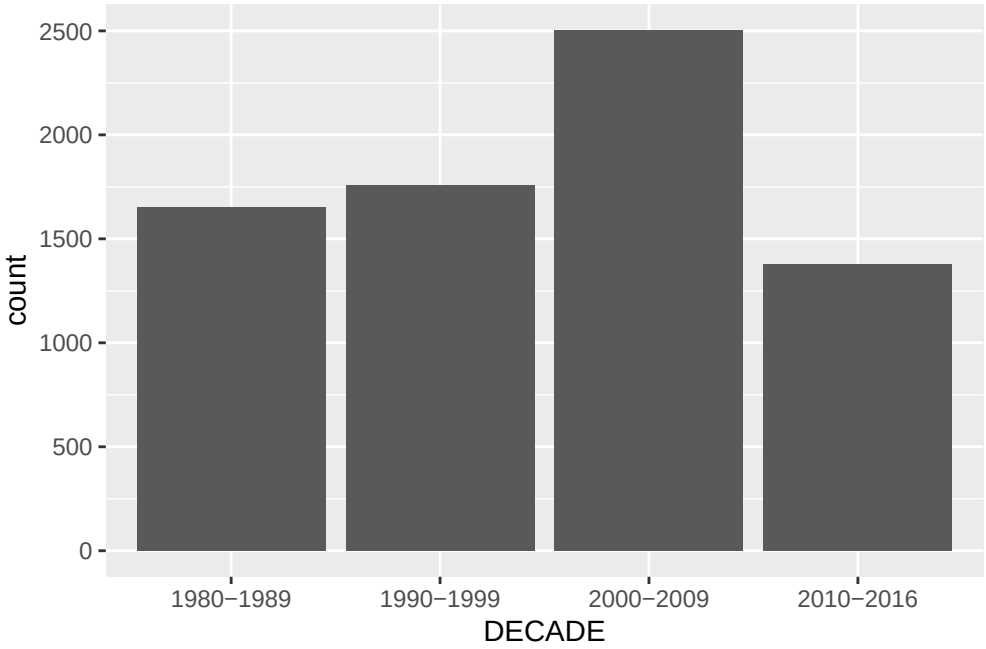
```
## # A tibble: 6 x 6
##   CAUSE YEAR_ ORGANIZATI STARTDATED   STATE   TOTALACRES
##   <chr> <dbl> <chr>      <chr>        <chr>        <dbl>
## 1 Human  1988 FWS        3/26/88 0:00 Arizona       1500
## 2 Human  1986 FWS        5/15/86 0:00 Arizona      10390
## 3 Human  1986 FWS        6/27/86 0:00 Montana       1400
## 4 Human  2002 FWS        2/28/02 0:00 Arizona       1035
## 5 Human  2000 FWS        4/9/00 0:00  Arizona       5700
## 6 Human  2000 FWS        5/14/00 0:00 Arizona       2750
```

```
# 1)
d6 = df %>%
  mutate(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"))))) %>%
  group_by(DECADE) %>%
  summarize(count=n())
head(d6)
```

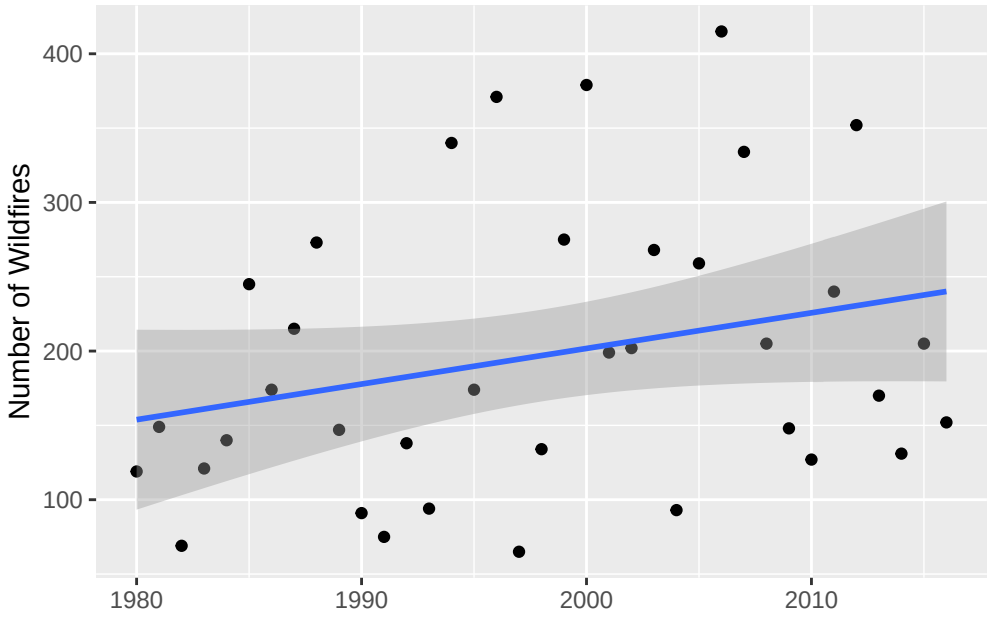
```
## # A tibble: 4 x 2
##   DECADE    count
##   <chr>     <int>
## 1 1980-1989  1652
## 2 1990-1999  1757
## 3 2000-2009  2502
## 4 2010-2016  1377
```

```
# table shows that wildfires have increased in the last few decades.
#
```

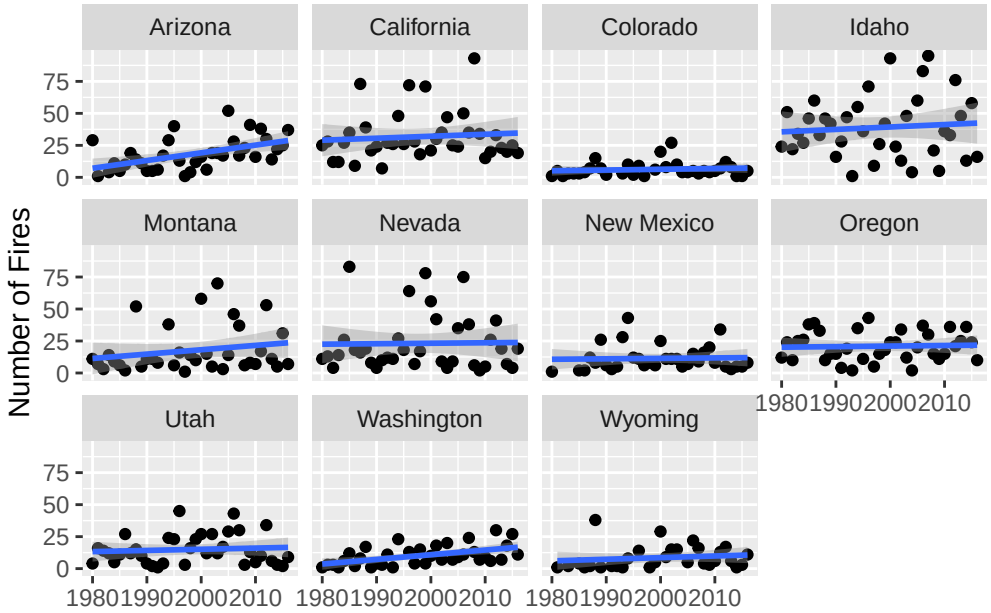
```r
# bar plot of the previous table
d6 = df %>%
     mutate(DECADE= ifelse(YEAR_ %in% 1980:1989,"1980-1989",
                        ifelse(YEAR_ %in% 1990:1999,"1990-1999",
                        ifelse(YEAR_ %in% 2000:2009,"2000-2009",
                        ifelse(YEAR_ %in% 2010:2016,"2010-2016","-99"))))) %>%
     group_by(DECADE) %>%
     summarize(count=n()) %>%
     ggplot(aes(x=DECADE,y=count)) + geom_col()
d6
```




```r
# Overall number of fires has increased during last few decades
# (note that last bar is a partial decade)
#
# or
df1 = df %>%
  select(STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  group_by(YR) %>%
  summarize(count=n()) %>%
  ggplot(mapping = aes(x=YR, y=count)) + geom_point() + geom_smooth(method=lm, se=TRUE) +
         xlab("") + ylab("Number of Wildfires")
df1
```

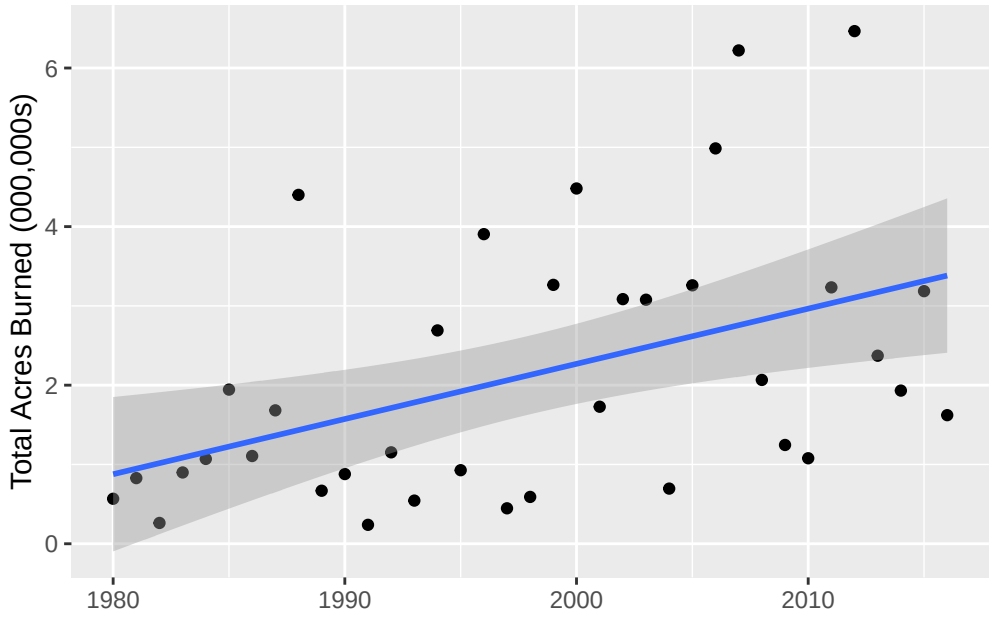



```
#
# Overall number of fires has increased during last few years

# 2)

df1b = df %>%
  select(STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  group_by(STATE, YR) %>%
  summarize(cnt = n()) %>%
  ggplot(mapping = aes(x=YR, y=cnt)) + geom_point() + facet_wrap(~STATE) +
  geom_smooth(method=lm, se=TRUE) +
  xlab("") + ylab("Number of Fires")
df1b
```

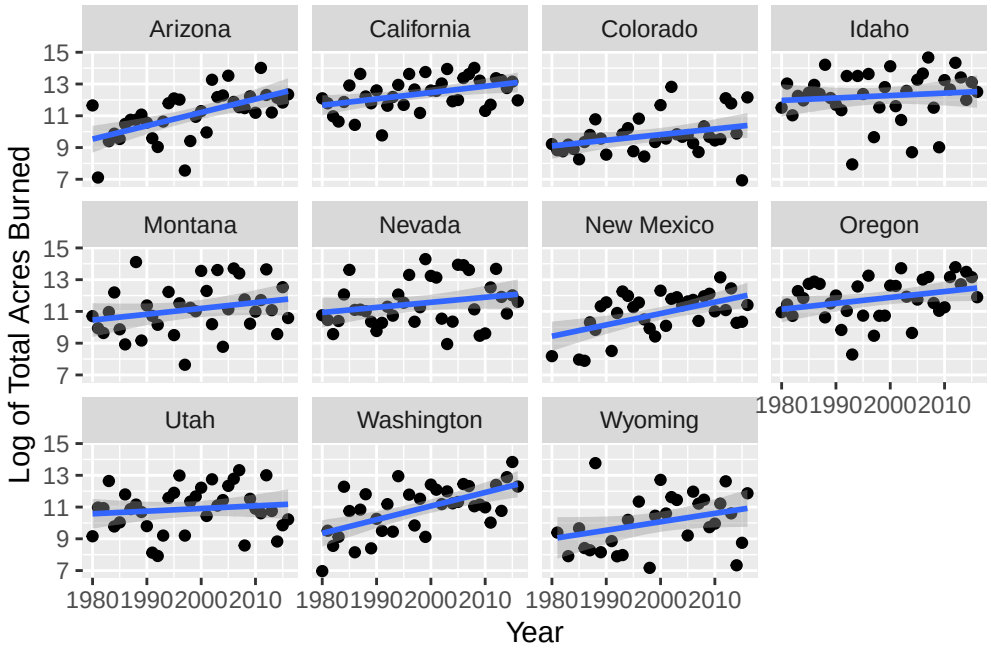



```
#
# Arizona has the largest increase in the number of wildfires
#
#
```

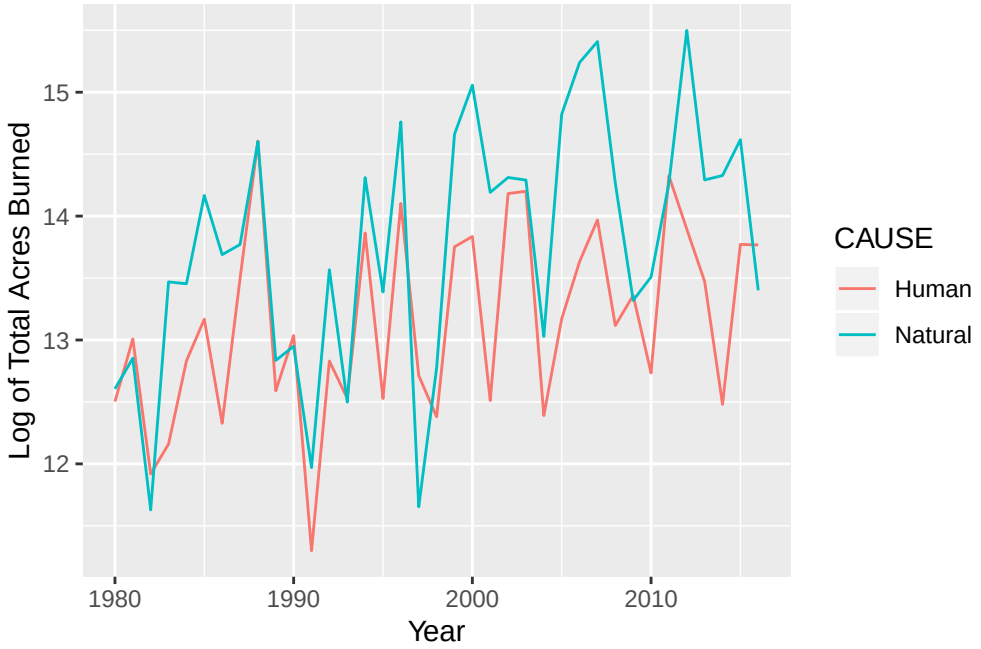
```r
# 3)
df2 = df %>%
  select(STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  group_by(YR) %>%
  summarize(totalacres = sum(ACRES)) %>%
  ggplot(mapping = aes(x=YR, y=totalacres/1000000)) +
     geom_point() + geom_smooth(method=lm, se=TRUE) +
     xlab("") + ylab("Total Acres Burned (000,000s)")
df2
```




```r
#
# The trend shows that the overall acreage burned has increased
#
# 4)

df2b = df %>%
  select(STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  group_by(STATE, YR) %>%
  summarize(totalacres = sum(ACRES)) %>%
  ggplot(mapping = aes(x=YR, y=log(totalacres))) +
  geom_point() + facet_wrap(~STATE) +
  geom_smooth(method=lm, se=TRUE) +
  xlab("Year") + ylab("Log of Total Acres Burned")
df2b
```

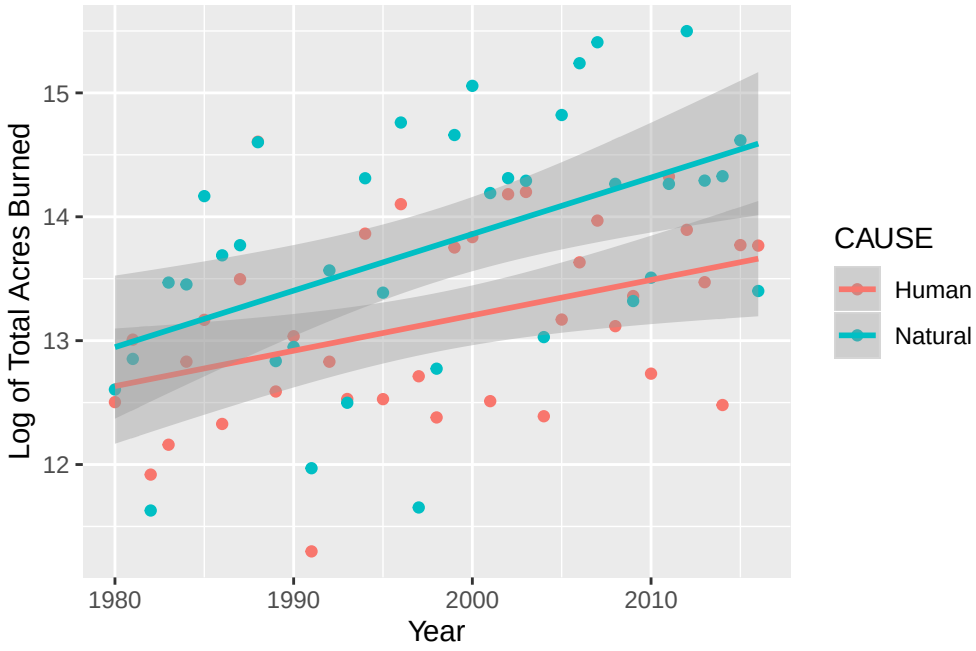



```
#
# Arizona and Washington have the largest increase in total acres burned.

# 5)
df2c = df %>%
  select(STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  filter(CAUSE %in% c('Human', 'Natural')) %>%
  group_by(CAUSE, YR) %>%
  summarize(totalacres = sum(ACRES)) %>%
  ggplot(mapping = aes(x=YR, y=log(totalacres), colour=CAUSE)) +
  geom_line() +
  xlab("Year") + ylab("Log of Total Acres Burned")
df2c
```

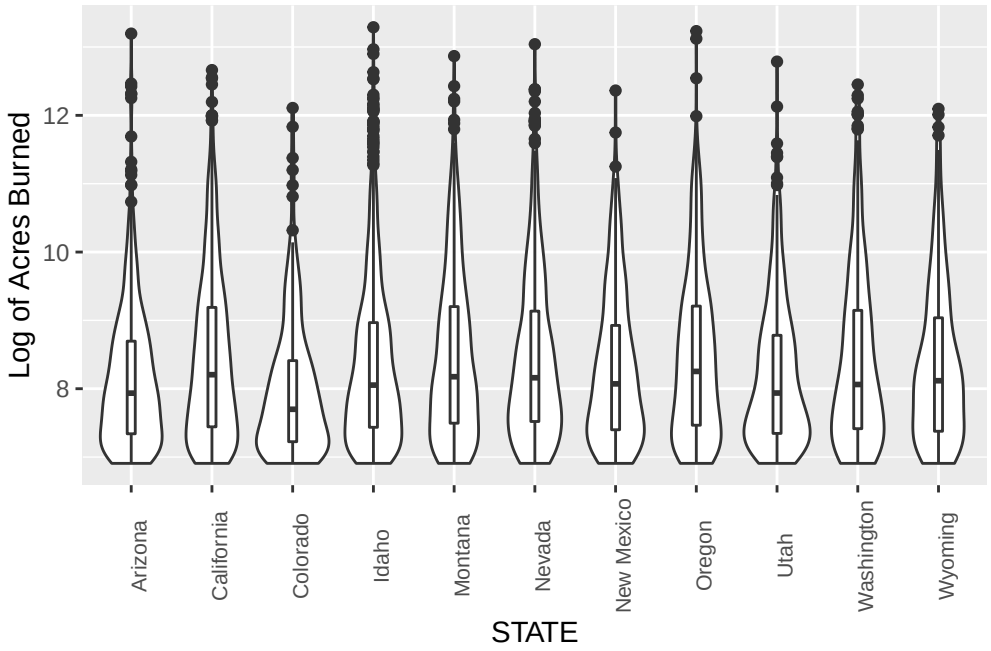



```
#
# lineplot for natural causes appears above that for human causes
```

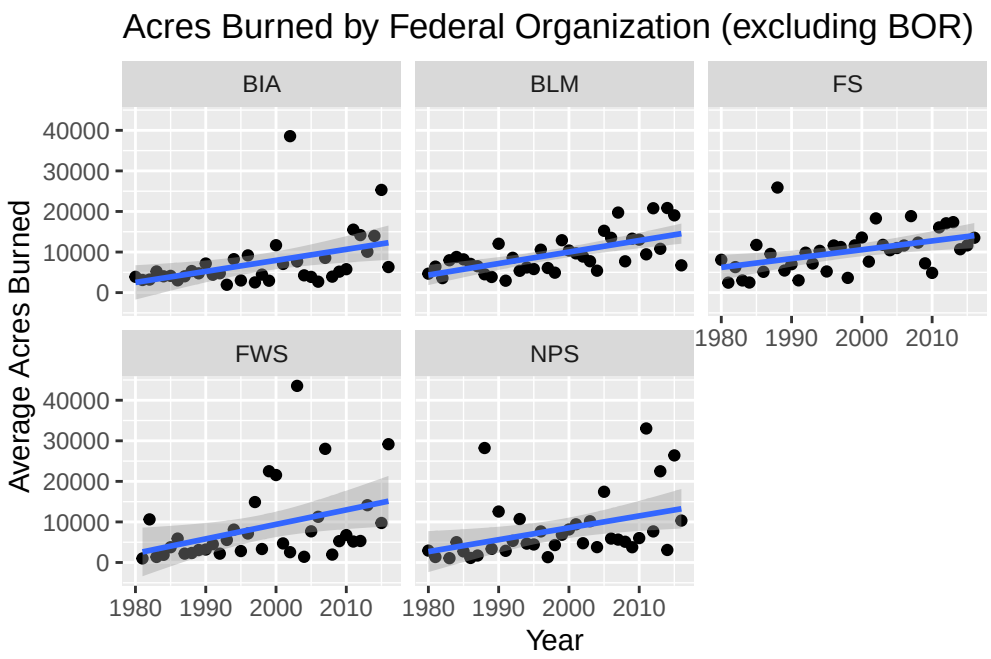
```
#
# or show a scatterplot with a trend line for each CAUSE
#
df2c = df %>%
  select(STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  filter(CAUSE %in% c('Human', 'Natural')) %>%
  group_by(CAUSE, YR) %>%
  summarize(totalacres = sum(ACRES)) %>%
  ggplot(mapping = aes(x=YR, y=log(totalacres), colour=CAUSE)) +
     geom_point() +
     geom_smooth(method=lm, se=TRUE) +
     xlab("Year") + ylab("Log of Total Acres Burned")
df2c
```



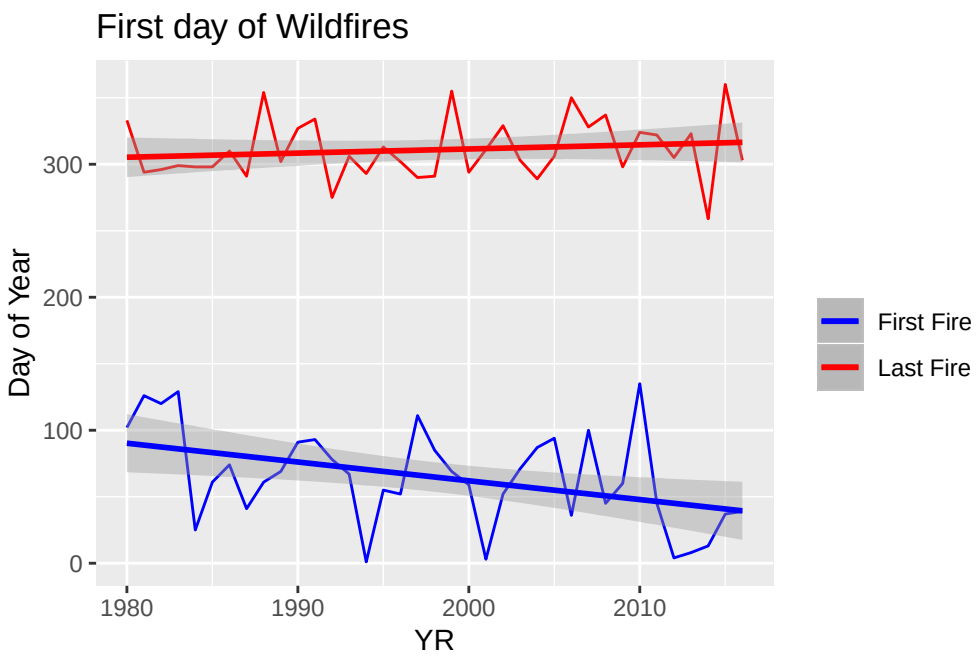

```
#
# The size of natural-cause wildfires is larger than that of human-cause fires
#
# 6) violin plots

df2d = df %>%
  select(ORGANIZATI, STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE) %>%
  group_by(STATE) %>%
  ggplot(mapping = aes(x=STATE, y=log(ACRES))) +
     geom_violin() +
     geom_boxplot(width=0.1) +
     theme(axis.text.x = element_text(angle = 90,size = 8)) +
     ylab("Log of Acres Burned")
df2d
```




```
#
# Distribution of acres burned across all years look similar for all states
#
# 7) Wildfire size by Federal Organization

df5 = df %>%
  select(ORG = ORGANIZATI, STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE, STARTDATED) %>%
  filter(ORG %in% c('BIA', 'BLM', 'FS', 'FWS', 'NPS')) %>%
  group_by(ORG, YR) %>%
  summarize(meanacres = mean(ACRES)) %>%
  ggplot(mapping = aes(x=YR, y=meanacres)) +
  geom_point() + facet_wrap(~ORG) +
  geom_smooth(method=lm, se=TRUE) +
  ggtitle("Acres Burned by Federal Organization (excluding BOR)") +
  xlab("Year") + ylab("Average Acres Burned")
df5
```




```
# Yearly average wildfire sizes are somewhat different by comparing Federal
# organizations, however the trend is similar.

# 8) Has the length of the fire season increased?

library(lubridate)
df4 = df %>%
  select(ORGANIZATI, STATE, YR = YEAR_, ACRES = TOTALACRES, CAUSE, STARTDATED) %>%
# find starting date of first and last wildfire
  mutate(DOY = yday(as.Date(STARTDATED, format='%m/%d/%y %H:%M'))) %>%
  group_by(YR) %>%
  summarize(dtEarly = min(DOY, na.rm=TRUE), dtLate = max(DOY, na.rm=TRUE)) %>%
  ggplot()  +
     geom_line(mapping = aes(x=YR, y=dtEarly, color='B'))  +
     geom_line(mapping = aes(x=YR, y=dtLate, color='R')) +
     geom_smooth(method=lm, se=TRUE, aes(x=YR, y=dtEarly, color="B")) +
     geom_smooth(method=lm, se=TRUE, aes(x=YR, y=dtLate, color="R")) +
     ggtitle("First day of Wildfires") +
     ylab("Day of Year") +
     scale_colour_manual(name = "",
                         values = c("R" = "red", "B" = "blue"),
                         labels = c("First Fire", "Last Fire"))
df4
```




```
#
# the gap between the first day of first and last wildfires is increasing.
# We can conclude tht the length of the fire season has increased over time.
```