Question1_Set1_2017CSC1061.R

nitish

2020-06-05

```
# QUESTION 1

# ALL EXTERNAL PACKAGES ----
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
```

```
library(tidyr)
library(ggplot2)
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
# I & II ----
setwd("/home/nitish/Desktop/R_stuff/DataSc/PracAsgn")
disaster <- read.csv("Disaster.csv",
                     header = TRUE)
as_tibble(disaster)
```

```
## # A tibble: 2,500 x 11
##       ID ORGANIZATION YEAR_ START_DATE END_DATE  UNIT NAME  CAUSE LOCALITY
##    <int> <fct>        <int> <fct>      <fct>    <int> <fct> <fct> <fct>
##  1     0 FWS           2001 1/1/2001 ~ 1/1/200~ 81682 PUMP~ Human CAL
##  2     1 FWS           2002 5/3/2002 ~ 5/3/200~ 81682 I5    Human CAL
##  3     2 FWS           2002 6/1/2002 ~ 6/1/200~ 81682 SOUT~ Human CAL
##  4     3 FWS           2001 7/12/2001~ 7/12/20~ 81682 MARI~ Human CAL
##  5     4 FWS           1994 9/13/1994~ 9/13/19~ 81682 HILL  Human CAL
##  6     5 FWS           1994 4/22/1994~ 4/22/19~ 81682 IRRI~ Human CAL
##  7     6 FWS           1999 12/6/1999~ 12/6/19~ 81682 FIELD Human CAL
##  8    18 FWS           2003 6/3/2003 ~ 6/3/200~ 81682 CALL~ Human CAL
```

```
##  9    20 FWS           2005 8/20/2005~ 8/20/20~ 81682 OVER~ Human CAL
## 10    21 FWS           2005 12/11/200~ 12/11/2~ 81682 TRAI~ Human CAL
## # ... with 2,490 more rows, and 2 more variables:
## #   DESTRUCTION..in.Thousand.Dollars. <int>, TOTAL_ACRES <dbl>
```

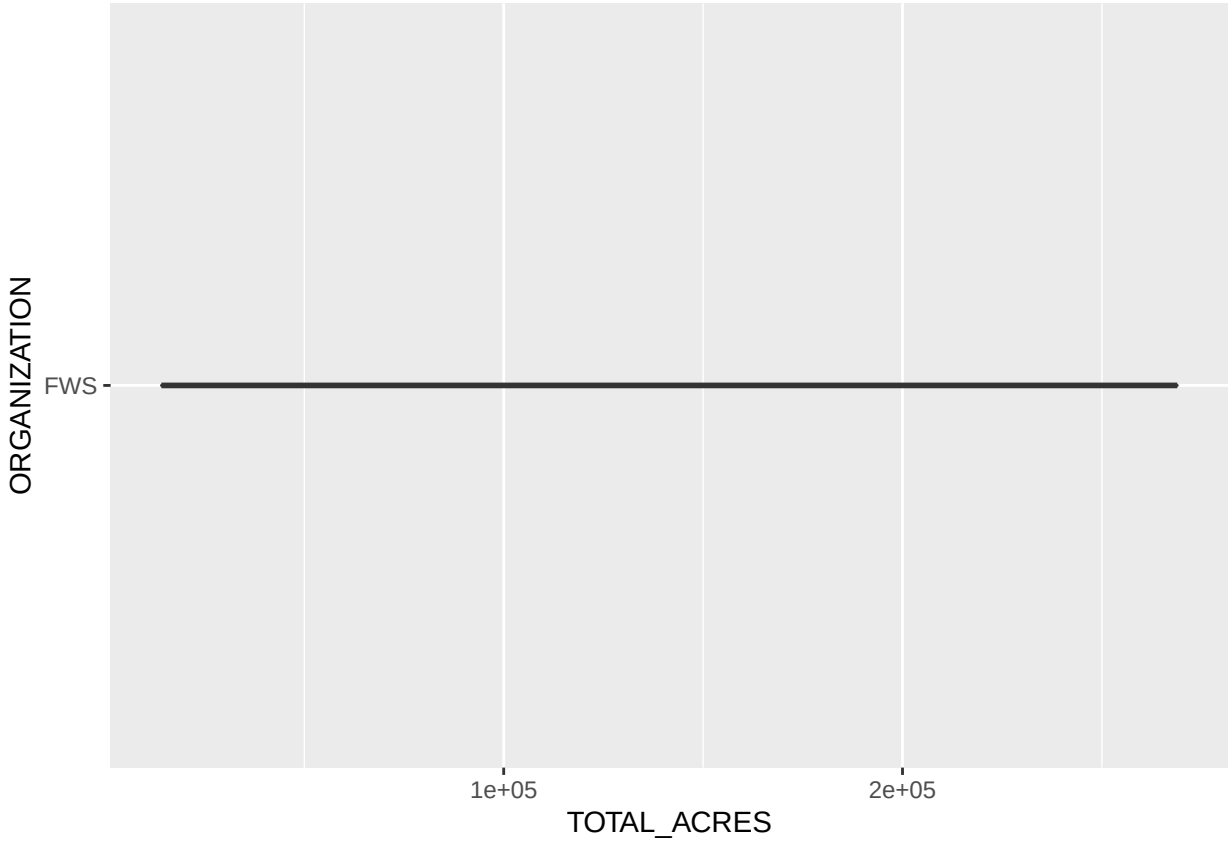
```
# III ----
nrow(disaster)
```

```
## [1] 2500
```

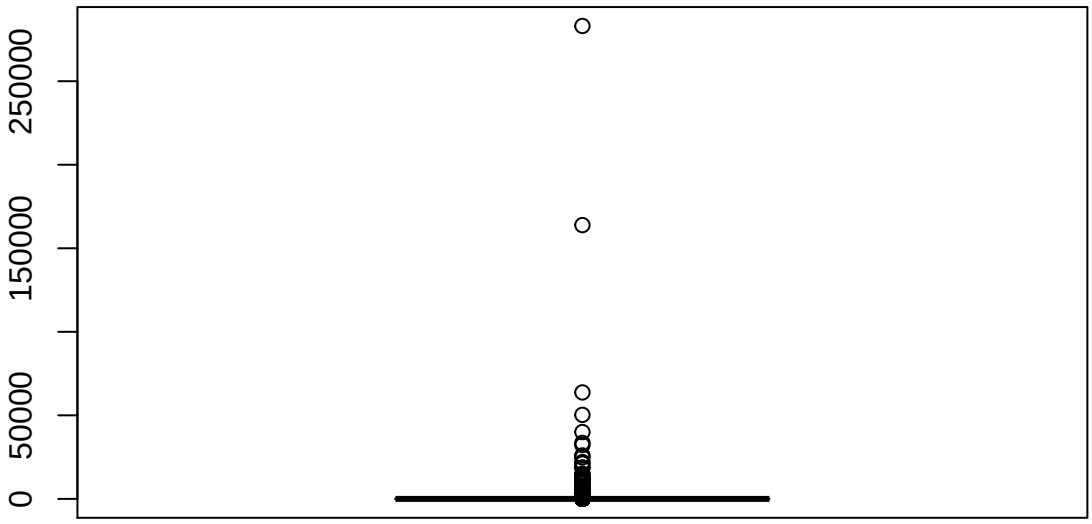
```
# IV ----
View(disaster)

# V ----
disaster %>%
  group_by(ORGANIZATION)
```

```
## # A tibble: 2,500 x 11
## # Groups:   ORGANIZATION [1]
##       ID ORGANIZATION YEAR_ START_DATE END_DATE  UNIT NAME  CAUSE LOCALITY
##    <int> <fct>        <int> <fct>      <fct>    <int> <fct> <fct> <fct>
##  1     0 FWS           2001 1/1/2001 ~ 1/1/200~ 81682 PUMP~ Human CAL
##  2     1 FWS           2002 5/3/2002 ~ 5/3/200~ 81682 I5    Human CAL
##  3     2 FWS           2002 6/1/2002 ~ 6/1/200~ 81682 SOUT~ Human CAL
##  4     3 FWS           2001 7/12/2001~ 7/12/20~ 81682 MARI~ Human CAL
##  5     4 FWS           1994 9/13/1994~ 9/13/19~ 81682 HILL  Human CAL
##  6     5 FWS           1994 4/22/1994~ 4/22/19~ 81682 IRRI~ Human CAL
##  7     6 FWS           1999 12/6/1999~ 12/6/19~ 81682 FIELD Human CAL
##  8    18 FWS           2003 6/3/2003 ~ 6/3/200~ 81682 CALL~ Human CAL
##  9    20 FWS           2005 8/20/2005~ 8/20/20~ 81682 OVER~ Human CAL
## 10    21 FWS           2005 12/11/200~ 12/11/2~ 81682 TRAI~ Human CAL
## # ... with 2,490 more rows, and 2 more variables:
## #   DESTRUCTION..in.Thousand.Dollars. <int>, TOTAL_ACRES <dbl>
```

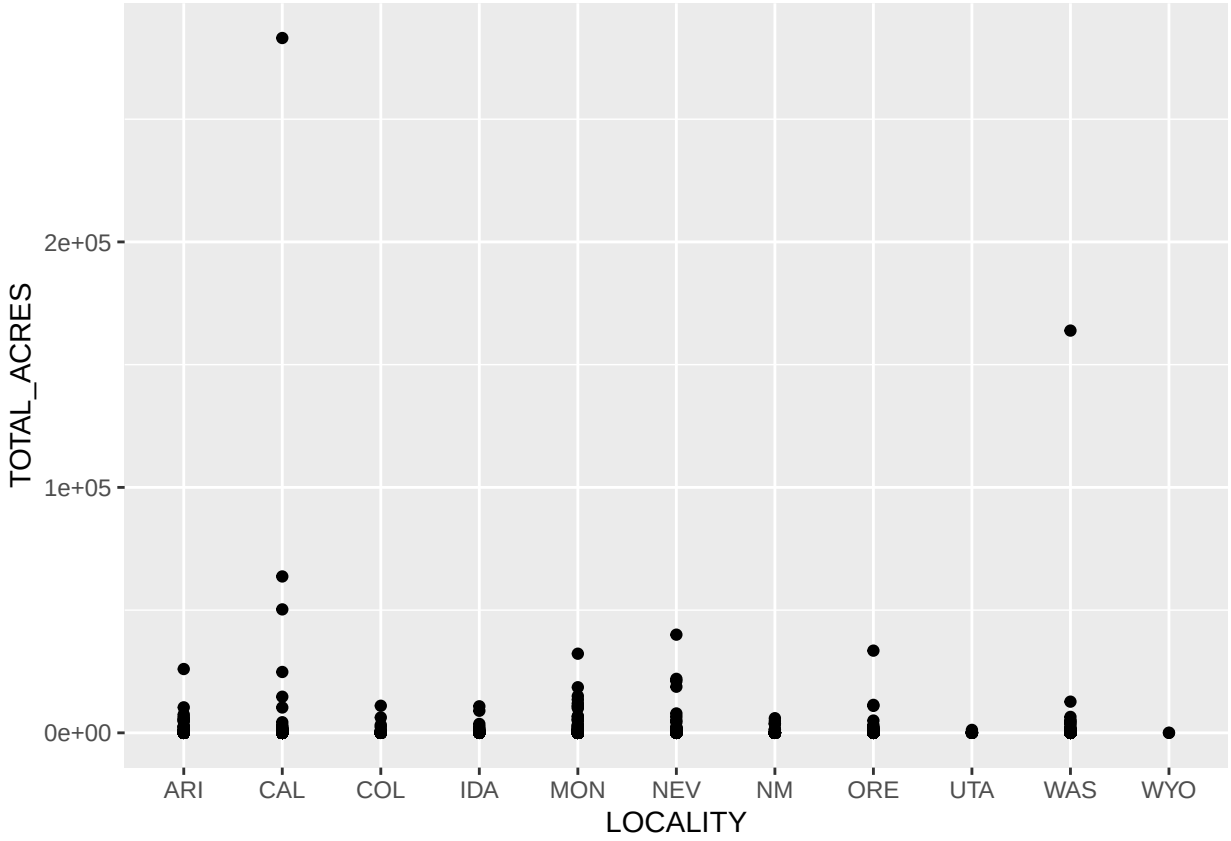
```
# VI ----
ggplot(disaster, aes(x = TOTAL_ACRES, y = ORGANIZATION)) +
  geom_boxplot()
```



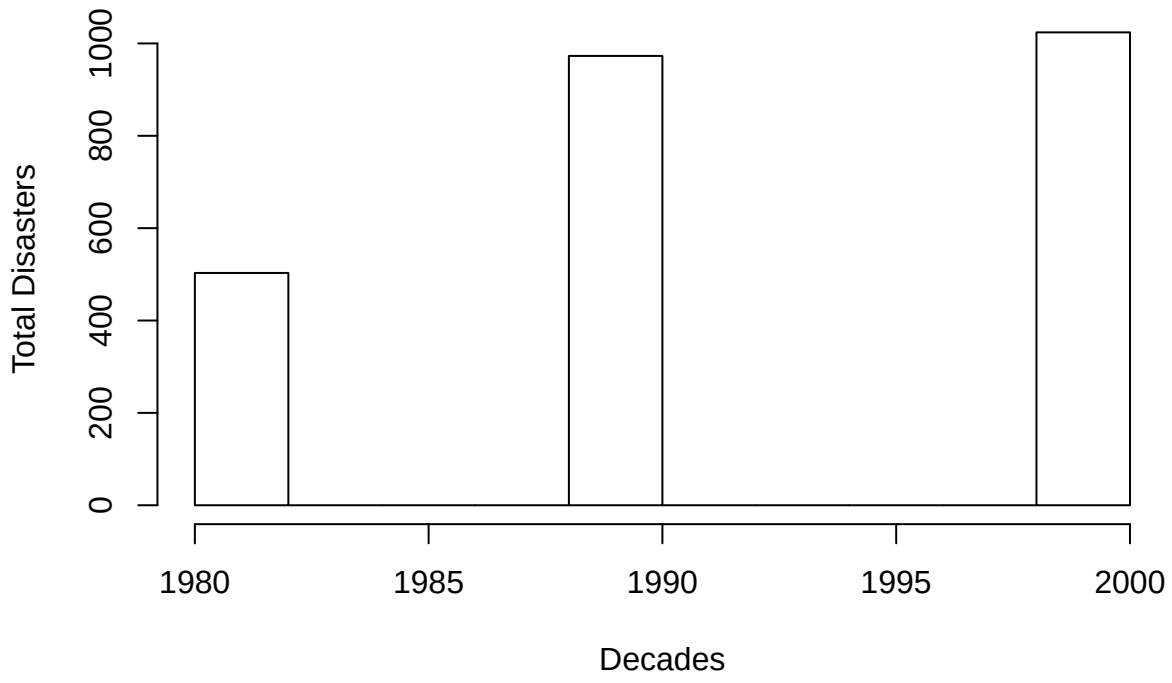

```
# VII ----
#Since TOTAL_ACRES gives us a range of numbers, we analyse on it.
outlier <- boxplot(disaster$TOTAL_ACRES)$out
```




```
# VIII ----
disaster %>%
  ggplot() +
  geom_point(mapping = aes(x = LOCALITY, y = TOTAL_ACRES)) +
  facet_null()
```




```
# IX ----
#First obtain decade(mutate), then plot.
decade <- disaster %>%
  mutate(decade = YEAR_ %/% 10 * 10) %>%
  select(decade)
hist(decade$decade,breaks = pretty(1980:2000, n=10), main = "Disasters in past decades", xlab = "Decade
```




```
# Clearly, the number of disasters have increased in the past few decades.

# X ----
summary(disaster)
```

```
##        ID         ORGANIZATION     YEAR_               START_DATE
##  Min.   :    0   FWS:2500     Min.   :1980   5/25/2000 0:00:  14
##  1st Qu.: 2205                1st Qu.:1991   7/1/2006 0:00 :  12
##  Median : 6702                Median :1997   8/7/2003 0:00 :   9
##  Mean   : 5853                Mean   :1996   3/22/1997 0:00:   8
##  3rd Qu.: 8987                3rd Qu.:2002   5/18/2000 0:00:   8
##  Max.   :11597                Max.   :2006   7/16/2003 0:00:   7
##                                              (Other)       :2442
##            END_DATE          UNIT            NAME          CAUSE
##                : 216   Min.   :13290   JAMACHA:   7   Human  :1638
##  7/1/2006 0:00 :   9   1st Qu.:14560   REFUGE :   7   Natural: 862
##  3/22/1997 0:00:   7   Median :61520   165    :   6
##  7/23/2006 0:00:   6   Mean   :46833   SLOAN  :   6
##  7/28/1986 0:00:   6   3rd Qu.:81648   BADGER :   5
##  4/25/1985 0:00:   5   Max.   :84593   COYOTE :   5
##  (Other)       :2251                   (Other):2464
##     LOCALITY   DESTRUCTION..in.Thousand.Dollars.  TOTAL_ACRES
##  CAL    :660   Min.   : 4.00                     Min.   :     0.0
##  MON    :438   1st Qu.: 6.00                     1st Qu.:     0.2
##  WAS    :365   Median :30.00                     Median :     2.0
##  ORE    :270   Mean   :24.72                     Mean   :   538.1
##  ARI    :246   3rd Qu.:41.00                     3rd Qu.:    25.0
##  NEV    :215   Max.   :56.00                     Max.   :283070.0
##  (Other):306
```

```r
# XI ----
disaster %>%
  select(END_DATE,START_DATE) %>%
  filter(!START_DATE == "" , !END_DATE == "") %>%
  mutate(RECOVERY_TIME = as.Date(END_DATE,format="%m/%d/%Y") - as.Date(START_DATE,format="%m/%d/%Y")) %>
  summarise(avg_rec_time = mean(RECOVERY_TIME))
```

```
##     avg_rec_time
## 1 0.8896673 days
```

```r
# XII ----
disaster %>%
  filter(!START_DATE == "" , !END_DATE == "") %>%
  separate(START_DATE,into = c("MONTH","day","year"), sep="/", convert = TRUE) %>%
  head()
```

```
##   ID ORGANIZATION YEAR_ MONTH day      year       END_DATE  UNIT       NAME
## 1  0          FWS  2001     1   1 2001 0:00  1/1/2001 0:00 81682 PUMP HOUSE
## 2  1          FWS  2002     5   3 2002 0:00  5/3/2002 0:00 81682         I5
## 3  2          FWS  2002     6   1 2002 0:00  6/1/2002 0:00 81682   SOUTHBAY
## 4  3          FWS  2001     7  12 2001 0:00 7/12/2001 0:00 81682     MARINA
## 5  4          FWS  1994     9  13 1994 0:00 9/13/1994 0:00 81682       HILL
## 6  5          FWS  1994     4  22 1994 0:00 4/22/1994 0:00 81682 IRRIGATION
##   CAUSE LOCALITY DESTRUCTION..in.Thousand.Dollars. TOTAL_ACRES
## 1 Human      CAL                                6         0.1
## 2 Human      CAL                                6         3.0
## 3 Human      CAL                                6         0.5
## 4 Human      CAL                                6         0.1
## 5 Human      CAL                                6         1.0
## 6 Human      CAL                                6         0.1
```

```r
# XIII ----
disaster %>%
  filter(!START_DATE == "" , !END_DATE == "") %>%
  separate(START_DATE,into = c("MONTH","day","year"), sep="/", convert = TRUE) %>%
  group_by(MONTH) %>%
  summarise(mean_destr = mean(DESTRUCTION..in.Thousand.Dollars.))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 12 x 2
##    MONTH mean_destr
##    <int>      <dbl>
##  1     1       15.2
##  2     2       16.1
##  3     3       22.5
##  4     4       21.9
##  5     5       21.4
##  6     6       25.9
##  7     7       27.2
##  8     8       27.2
##  9     9       24.3
## 10    10       21.3
## 11    11       19.0
## 12    12       11.6
```

```r
# XIV ----
#read.table() function is most conveniently used in text(.txt) file format.


# XV ----
#Working of both the commands is same.
#read.csv is in the utils package and read_csv is in the readr package.
#read.csv has less arguments than read_csv.


# XVI ----
#base packages:utils,graphics,BSDA.
#third party:dplyr,ggplot2.


# XVII ----
#Z-test could be performed in two variables.
#DESTRUCTION..in.Thousand.Dollars. ,  TOTAL_ACRES.


# XVIII ----
z.test(disaster$DESTRUCTION..in.Thousand.Dollars.,sigma.x = sd(disaster$DESTRUCTION..in.Thousand.Dollars
```

```
##
##  One-sample z-Test
##
## data:  disaster$DESTRUCTION..in.Thousand.Dollars.
## z = 70.302, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  24.02811 25.40629
## sample estimates:
## mean of x
##   24.7172
```

```r
# XIX ----
#We could check the data by shapiro wilk test for normal distribution and by plotting the graph.


# XX ----
# We can change the values of the numeric columns to the common scale.
```