Assignment 3 Getting and Cleaning Data

Akash Chauhan

11/26/2021

```
getwd()
```

```
## [1] "/cloud/project"
```

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#reading the data into data frame
stormdata <- read_csv("StormEvents1994.csv")
```

```
## Rows: 15627 Columns: 51
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): STATE, MONTH_NAME, EVENT_TYPE, CZ_TYPE, CZ_NAME, BEGIN_DATE_TIME, ...
## dbl (23): BEGIN_YEARMONTH, BEGIN_DAY, BEGIN_TIME, END_YEARMONTH, END_DAY, EN...
## lgl (11): EPISODE_ID, WFO, SOURCE, MAGNITUDE_TYPE, FLOOD_CAUSE, CATEGORY, TO...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nrow(stormdata)
```

```
## [1] 15627
```

```
#show fist 5 rows of the data
head(stormdata, 5)
```

```
## # A tibble: 5 x 51
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME END_YEARMONTH END_DAY END_TIME EPISODE_ID
##             <dbl>     <dbl>      <dbl>         <dbl>   <dbl>    <dbl> <lgl>
## 1          199403        27       1132        199403      27     1132 NA
```

```
## 2          199405        15       1930        199405      15     1930 NA
## 3          199406        26       2220        199406      26     2220 NA
## 4          199405        15       1347        199405      15     1347 NA
## 5          199403        27       1550        199403      27     1550 NA
## # ... with 44 more variables: EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <dbl>,
## #   YEAR <dbl>, MONTH_NAME <chr>, EVENT_TYPE <chr>, CZ_TYPE <chr>,
## #   CZ_FIPS <dbl>, CZ_NAME <chr>, WFO <lgl>, BEGIN_DATE_TIME <chr>,
## #   CZ_TIMEZONE <chr>, END_DATE_TIME <chr>, INJURIES_DIRECT <dbl>,
## #   INJURIES_INDIRECT <dbl>, DEATHS_DIRECT <dbl>, DEATHS_INDIRECT <dbl>,
## #   DAMAGE_PROPERTY <chr>, DAMAGE_CROPS <chr>, SOURCE <lgl>, MAGNITUDE <dbl>,
## #   MAGNITUDE_TYPE <lgl>, FLOOD_CAUSE <lgl>, CATEGORY <lgl>, ...
```

```
#print all column header names
colnames(x=stormdata)
```

```
##  [1] "BEGIN_YEARMONTH"    "BEGIN_DAY"          "BEGIN_TIME"
##  [4] "END_YEARMONTH"      "END_DAY"            "END_TIME"
##  [7] "EPISODE_ID"         "EVENT_ID"           "STATE"
## [10] "STATE_FIPS"         "YEAR"               "MONTH_NAME"
## [13] "EVENT_TYPE"         "CZ_TYPE"            "CZ_FIPS"
## [16] "CZ_NAME"            "WFO"                "BEGIN_DATE_TIME"
## [19] "CZ_TIMEZONE"        "END_DATE_TIME"      "INJURIES_DIRECT"
## [22] "INJURIES_INDIRECT"  "DEATHS_DIRECT"      "DEATHS_INDIRECT"
## [25] "DAMAGE_PROPERTY"    "DAMAGE_CROPS"       "SOURCE"
## [28] "MAGNITUDE"          "MAGNITUDE_TYPE"     "FLOOD_CAUSE"
## [31] "CATEGORY"           "TOR_F_SCALE"        "TOR_LENGTH"
## [34] "TOR_WIDTH"          "TOR_OTHER_WFO"      "TOR_OTHER_CZ_STATE"
## [37] "TOR_OTHER_CZ_FIPS"  "TOR_OTHER_CZ_NAME"  "BEGIN_RANGE"
## [40] "BEGIN_AZIMUTH"      "BEGIN_LOCATION"     "END_RANGE"
## [43] "END_AZIMUTH"        "END_LOCATION"       "BEGIN_LAT"
## [46] "BEGIN_LON"          "END_LAT"            "END_LON"
## [49] "EPISODE_NARRATIVE"  "EVENT_NARRATIVE"    "DATA_SOURCE"
```

```
#Limit the data frame to listed columns
myvars <- c("BEGIN_YEARMONTH",
            "BEGIN_DAY",
            "BEGIN_TIME",
            "BEGIN_DATE_TIME",
            "END_YEARMONTH",
            "END_DAY",
            "END_TIME",
            "END_DATE_TIME",
            "EPISODE_ID",
            "EVENT_ID",
            "STATE",
            "STATE_FIPS",
            "CZ_TYPE",
            "CZ_FIPS",
            "CZ_NAME",
            "EVENT_TYPE",
            "SOURCE",
            "BEGIN_LAT",
            "BEGIN_LON",
            "END_LAT",
            "END_LON")
```

```
#limit dataframe to above selected vars
newStormData <- stormdata[myvars]
head(newStormData, 5)
```

```
## # A tibble: 5 x 21
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME    END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <chr>                     <dbl>   <dbl>
## 1          199403        27       1132 27-MAR-94 11:32:00        199403      27
## 2          199405        15       1930 15-MAY-94 19:30:00        199405      15
## 3          199406        26       2220 26-JUN-94 22:20:00        199406      26
## 4          199405        15       1347 15-MAY-94 13:47:00        199405      15
## 5          199403        27       1550 27-MAR-94 15:50:00        199403      27
## # ... with 15 more variables: END_TIME <dbl>, END_DATE_TIME <chr>,
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <dbl>,
## #   CZ_TYPE <chr>, CZ_FIPS <dbl>, CZ_NAME <chr>, EVENT_TYPE <chr>,
## #   SOURCE <lgl>, BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>,
## #   END_LON <dbl>
```

```
colnames(x=newStormData)
```

```
##  [1] "BEGIN_YEARMONTH" "BEGIN_DAY"       "BEGIN_TIME"      "BEGIN_DATE_TIME"
##  [5] "END_YEARMONTH"   "END_DAY"         "END_TIME"        "END_DATE_TIME"
##  [9] "EPISODE_ID"      "EVENT_ID"        "STATE"           "STATE_FIPS"
## [13] "CZ_TYPE"         "CZ_FIPS"         "CZ_NAME"         "EVENT_TYPE"
## [17] "SOURCE"          "BEGIN_LAT"       "BEGIN_LON"       "END_LAT"
## [21] "END_LON"
```

```
#pad time data with 0 to become format of hhmm(hours-minutes)
install.packages("lubridate")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(lubridate)
```

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

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

```
library(dplyr)
#change to date-time class
newStormData <- newStormData %>%
mutate(END_DATE_TIME = dmy_hms(END_DATE_TIME))
newStormData <- newStormData %>%
mutate(BEGIN_DATE_TIME = dmy_hms(BEGIN_DATE_TIME))
head(newStormData, 5)
```

```
## # A tibble: 5 x 21
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME     END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
```

```
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 15 more variables: END_TIME <dbl>, END_DATE_TIME <dttm>,
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <dbl>,
## #   CZ_TYPE <chr>, CZ_FIPS <dbl>, CZ_NAME <chr>, EVENT_TYPE <chr>,
## #   SOURCE <lgl>, BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>,
## #   END_LON <dbl>
```

```
colnames(x=newStormData)
```

```
##  [1] "BEGIN_YEARMONTH" "BEGIN_DAY"       "BEGIN_TIME"      "BEGIN_DATE_TIME"
##  [5] "END_YEARMONTH"   "END_DAY"         "END_TIME"        "END_DATE_TIME"
##  [9] "EPISODE_ID"      "EVENT_ID"        "STATE"           "STATE_FIPS"
## [13] "CZ_TYPE"         "CZ_FIPS"         "CZ_NAME"         "EVENT_TYPE"
## [17] "SOURCE"          "BEGIN_LAT"       "BEGIN_LON"       "END_LAT"
## [21] "END_LON"
```

```
#convert upper case state to title case
newStormData$STATE <- str_to_title(newStormData$STATE, locale = "en")
newStormData$CZ_NAME <- str_to_title(newStormData$CZ_NAME, locale = "en")
head(newStormData, 5)
```

```
## # A tibble: 5 x 21
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME     END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 15 more variables: END_TIME <dbl>, END_DATE_TIME <dttm>,
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <dbl>,
## #   CZ_TYPE <chr>, CZ_FIPS <dbl>, CZ_NAME <chr>, EVENT_TYPE <chr>,
## #   SOURCE <lgl>, BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>,
## #   END_LON <dbl>
```

```
colnames(x=newStormData)
```

```
##  [1] "BEGIN_YEARMONTH" "BEGIN_DAY"       "BEGIN_TIME"      "BEGIN_DATE_TIME"
##  [5] "END_YEARMONTH"   "END_DAY"         "END_TIME"        "END_DATE_TIME"
##  [9] "EPISODE_ID"      "EVENT_ID"        "STATE"           "STATE_FIPS"
## [13] "CZ_TYPE"         "CZ_FIPS"         "CZ_NAME"         "EVENT_TYPE"
## [17] "SOURCE"          "BEGIN_LAT"       "BEGIN_LON"       "END_LAT"
## [21] "END_LON"
```

```
#Filter where county type is 'C' and then remove CZ_TYPE column
newSD <- filter(newStormData, CZ_TYPE == 'C')
head(newSD, 5)
```

```
## # A tibble: 5 x 21
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME     END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 15 more variables: END_TIME <dbl>, END_DATE_TIME <dttm>,
```

```
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <dbl>,
## #   CZ_TYPE <chr>, CZ_FIPS <dbl>, CZ_NAME <chr>, EVENT_TYPE <chr>,
## #   SOURCE <lgl>, BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>,
## #   END_LON <dbl>
```

```
nrow(newSD)
```

```
## [1] 15627
```

```
newSD$CZ_TYPE <- NULL
head(newSD, 5)
```

```
## # A tibble: 5 x 20
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME     END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 14 more variables: END_TIME <dbl>, END_DATE_TIME <dttm>,
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <dbl>,
## #   CZ_FIPS <dbl>, CZ_NAME <chr>, EVENT_TYPE <chr>, SOURCE <lgl>,
## #   BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>, END_LON <dbl>
```

```
colnames(x=newSD)
```

```
##  [1] "BEGIN_YEARMONTH" "BEGIN_DAY"       "BEGIN_TIME"      "BEGIN_DATE_TIME"
##  [5] "END_YEARMONTH"   "END_DAY"         "END_TIME"        "END_DATE_TIME"
##  [9] "EPISODE_ID"      "EVENT_ID"        "STATE"           "STATE_FIPS"
## [13] "CZ_FIPS"         "CZ_NAME"         "EVENT_TYPE"      "SOURCE"
## [17] "BEGIN_LAT"       "BEGIN_LON"       "END_LAT"         "END_LON"
```

```
#Pad the state and county FIPS with a "0" and unite the 2 col
newSD$CZ_FIPS <- str_pad(newSD$CZ_FIPS, width = 3, side = "left", pad="0")
newSD$STATE_FIPS <- str_pad(newSD$STATE_FIPS, width = 2, side = "left", pad="0")
head(newSD, 5)
```

```
## # A tibble: 5 x 20
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME     END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 14 more variables: END_TIME <dbl>, END_DATE_TIME <dttm>,
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, STATE_FIPS <chr>,
## #   CZ_FIPS <chr>, CZ_NAME <chr>, EVENT_TYPE <chr>, SOURCE <lgl>,
## #   BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>, END_LON <dbl>
```

```
colnames(x=newSD)
```

```
##  [1] "BEGIN_YEARMONTH" "BEGIN_DAY"       "BEGIN_TIME"      "BEGIN_DATE_TIME"
##  [5] "END_YEARMONTH"   "END_DAY"         "END_TIME"        "END_DATE_TIME"
##  [9] "EPISODE_ID"      "EVENT_ID"        "STATE"           "STATE_FIPS"
## [13] "CZ_FIPS"         "CZ_NAME"         "EVENT_TYPE"      "SOURCE"
## [17] "BEGIN_LAT"       "BEGIN_LON"       "END_LAT"         "END_LON"
```

```
#newSD <- unite(newSD, CZ_FIPS, STATE_FIPS, sep = "", remove = TRUE)
newSD <- newSD %>% unite("CZ_SATE_FIPS", CZ_FIPS:STATE_FIPS, sep = "", remove = TRUE)
head(newSD, 5)
```

```
## # A tibble: 5 x 19
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME BEGIN_DATE_TIME     END_YEARMONTH END_DAY
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 13 more variables: END_TIME <dbl>, END_DATE_TIME <dttm>,
## #   EPISODE_ID <lgl>, EVENT_ID <dbl>, STATE <chr>, CZ_SATE_FIPS <chr>,
## #   CZ_NAME <chr>, EVENT_TYPE <chr>, SOURCE <lgl>, BEGIN_LAT <dbl>,
## #   BEGIN_LON <dbl>, END_LAT <dbl>, END_LON <dbl>
```

```
colnames(x=newSD)
```

```
##  [1] "BEGIN_YEARMONTH" "BEGIN_DAY"       "BEGIN_TIME"      "BEGIN_DATE_TIME"
##  [5] "END_YEARMONTH"   "END_DAY"         "END_TIME"        "END_DATE_TIME"
##  [9] "EPISODE_ID"      "EVENT_ID"        "STATE"           "CZ_SATE_FIPS"
## [13] "CZ_NAME"         "EVENT_TYPE"      "SOURCE"          "BEGIN_LAT"
## [17] "BEGIN_LON"       "END_LAT"         "END_LON"
```

```
#rename all columns to lower case
newSD <- newSD %>% rename_all(tolower)
head(newSD, 5)
```

```
## # A tibble: 5 x 19
##   begin_yearmonth begin_day begin_time begin_date_time     end_yearmonth end_day
##             <dbl>     <dbl>      <dbl> <dttm>                     <dbl>   <dbl>
## 1          199403        27       1132 1994-03-27 11:32:00        199403      27
## 2          199405        15       1930 1994-05-15 19:30:00        199405      15
## 3          199406        26       2220 1994-06-26 22:20:00        199406      26
## 4          199405        15       1347 1994-05-15 13:47:00        199405      15
## 5          199403        27       1550 1994-03-27 15:50:00        199403      27
## # ... with 13 more variables: end_time <dbl>, end_date_time <dttm>,
## #   episode_id <lgl>, event_id <dbl>, state <chr>, cz_sate_fips <chr>,
## #   cz_name <chr>, event_type <chr>, source <lgl>, begin_lat <dbl>,
## #   begin_lon <dbl>, end_lat <dbl>, end_lon <dbl>
```

```
colnames(x=newSD)
```

```
##  [1] "begin_yearmonth" "begin_day"       "begin_time"      "begin_date_time"
##  [5] "end_yearmonth"   "end_day"         "end_time"        "end_date_time"
##  [9] "episode_id"      "event_id"        "state"           "cz_sate_fips"
## [13] "cz_name"         "event_type"      "source"          "begin_lat"
## [17] "begin_lon"       "end_lat"         "end_lon"
```

```
#New Data frame with 3 coloumns
us_state_info <-data.frame(state=state.name, region=state.region, area=state.area)
us_state_info
```

```
##            state        region   area
## 1        Alabama         South  51609
## 2         Alaska          West 589757
```

```
## 3         Arizona          West 113909
## 4        Arkansas         South  53104
## 5      California          West 158693
## 6        Colorado          West 104247
## 7     Connecticut     Northeast   5009
## 8        Delaware         South   2057
## 9         Florida         South  58560
## 10        Georgia         South  58876
## 11         Hawaii          West   6450
## 12          Idaho          West  83557
## 13       Illinois North Central  56400
## 14        Indiana North Central  36291
## 15           Iowa North Central  56290
## 16         Kansas North Central  82264
## 17       Kentucky         South  40395
## 18      Louisiana         South  48523
## 19          Maine     Northeast  33215
## 20       Maryland         South  10577
## 21  Massachusetts     Northeast   8257
## 22       Michigan North Central  58216
## 23      Minnesota North Central  84068
## 24    Mississippi         South  47716
## 25       Missouri North Central  69686
## 26        Montana          West 147138
## 27       Nebraska North Central  77227
## 28         Nevada          West 110540
## 29  New Hampshire     Northeast   9304
## 30     New Jersey     Northeast   7836
## 31     New Mexico          West 121666
## 32       New York     Northeast  49576
## 33 North Carolina         South  52586
## 34   North Dakota North Central  70665
## 35           Ohio North Central  41222
## 36       Oklahoma         South  69919
## 37         Oregon          West  96981
## 38   Pennsylvania     Northeast  45333
## 39   Rhode Island     Northeast   1214
## 40 South Carolina         South  31055
## 41   South Dakota North Central  77047
## 42      Tennessee         South  42244
## 43          Texas         South 267339
## 44           Utah          West  84916
## 45        Vermont     Northeast   9609
## 46       Virginia         South  40815
## 47     Washington          West  68192
## 48  West Virginia         South  24181
## 49      Wisconsin North Central  56154
## 50        Wyoming          West  97914
```

```
#9 - Create a dataframe with the number of events per state using a frequency table
eventsFreq <- data.frame(table(newStormData$STATE))
eventsFreq
```

```
##             Var1 Freq
## 1        Alabama  266
```

```
## 2         Arizona   67
## 3        Arkansas  893
## 4      California   22
## 5        Colorado  308
## 6     Connecticut   45
## 7        Delaware    8
## 8         Florida  304
## 9         Georgia  365
## 10          Idaho   18
## 11       Illinois  305
## 12        Indiana  290
## 13           Iowa  606
## 14         Kansas  899
## 15       Kentucky   89
## 16      Louisiana  456
## 17          Maine   77
## 18       Maryland   64
## 19  Massachusetts   75
## 20       Michigan  311
## 21      Minnesota  387
## 22    Mississippi  423
## 23       Missouri  569
## 24        Montana  133
## 25       Nebraska  523
## 26         Nevada   16
## 27  New Hampshire   38
## 28     New Jersey  109
## 29     New Mexico  113
## 30       New York  567
## 31 North Carolina  144
## 32   North Dakota  223
## 33           Ohio  610
## 34       Oklahoma 1670
## 35         Oregon   11
## 36   Pennsylvania  461
## 37    Puerto Rico   11
## 38   Rhode Island    7
## 39 South Carolina  328
## 40   South Dakota  249
## 41      Tennessee  232
## 42          Texas 2544
## 43           Utah   46
## 44        Vermont   59
## 45       Virginia   99
## 46     Washington    9
## 47  West Virginia  191
## 48      Wisconsin  303
## 49        Wyoming   84
```

```
eventsFreq<-rename(eventsFreq, c("state"="Var1"))
head(eventsFreq)
```

```
##         state Freq
## 1     Alabama  266
## 2     Arizona   67
```

```
## 3    Arkansas  893
## 4  California   22
## 5    Colorado  308
## 6 Connecticut   45
```

```
state_storms <- merge(x=eventsFreq,y=us_state_info,by.x="state", by.y="state")
```

```
#create plot
library(ggplot2)
storm_plot <- ggplot(state_storms,
                     aes(x=area, y=Freq))+
  geom_point(aes(color = region)) +
  labs(x = "Land area(sq. miles)",
       y = "# of storm events in 1994")
storm_plot
```