Assignment 3

sunil salunke

7/03/2022

```
#libraries needed for analysis
library(tidyverse)
library(ggplot2)
```

1.Reading and saving the data into R

```
#let's import the csv file using read_csv function
storm_event_details_92 <-
↪  read_csv("/Users/sunilsalunke/Documents/ana515/week6/StormEvents_details-ftp_v1.0_d1992_c20220425.cs
```

```
## Rows: 13534 Columns: 51
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (12): STATE, MONTH_NAME, EVENT_TYPE, CZ_TYPE, CZ_NAME, WFO, BEGIN_DATE_T...
## dbl (24): BEGIN_YEARMONTH, BEGIN_DAY, BEGIN_TIME, END_YEARMONTH, END_DAY, EN...
## lgl (15): EPISODE_ID, SOURCE, MAGNITUDE_TYPE, FLOOD_CAUSE, CATEGORY, TOR_OTH...
##
## 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.
```

```
#let' see the storm event details data for 1992
head(storm_event_details_92)
```

```
## # A tibble: 6 x 51
##   BEGIN_YEARMONTH BEGIN_DAY BEGIN_TIME END_YEARMONTH END_DAY END_TIME EPISODE_ID
##             <dbl>     <dbl>      <dbl>         <dbl>   <dbl>    <dbl> <lgl>
## 1          199206        24       1511        199206      24     1511 NA
## 2          199206        24       1827        199206      24     1827 NA
## 3          199206        24       1943        199206      24     1943 NA
## 4          199206        25       1950        199206      25     1950 NA
## 5          199206        26       1251        199206      26     1251 NA
## 6          199206        26       1840        199206      26     1840 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 <chr>, 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 <dbl>, SOURCE <lgl>, MAGNITUDE <dbl>,
## #   MAGNITUDE_TYPE <lgl>, FLOOD_CAUSE <lgl>, CATEGORY <lgl>, ...
```

2.Limit the dataframe to the following columns

1. the beginning and ending dates and times (make sure to keep BEGIN_DATE_TIME and END_DATE_TIME)

2. the episode ID

3. the event ID

4. the state name and FIPS

5. the "CZ" name

6. type

7. FIPS

8. the event type

9. the source

```
#Limit the data frame by selecting the required columns and save it as a new data frame
storm_event_details_92_subset <- storm_event_details_92 %>% select(BEGIN_DATE_TIME,
↪  END_DATE_TIME, EPISODE_ID, EVENT_ID, STATE, STATE_FIPS, CZ_NAME, CZ_TYPE, CZ_FIPS,
↪  EVENT_TYPE, SOURCE, BEGIN_LAT, BEGIN_LON, END_LAT, END_LON)

head(storm_event_details_92_subset)
```

10. the beginning latitude and longitude and ending latitude and longitude

```
## # A tibble: 6 x 15
##   BEGIN_DATE_TIME    END_DATE_TIME  EPISODE_ID EVENT_ID STATE STATE_FIPS CZ_NAME
##   <chr>              <chr>          <lgl>         <dbl> <chr>      <dbl> <chr>
## 1 24-JUN-92 15:11:00 24-JUN-92 15:~ NA          9985034 COLO~          8 LARIMER
## 2 24-JUN-92 18:27:00 24-JUN-92 18:~ NA          9985035 COLO~          8 EL PASO
## 3 24-JUN-92 19:43:00 24-JUN-92 19:~ NA          9985036 COLO~          8 KIT CA~
## 4 25-JUN-92 19:50:00 25-JUN-92 19:~ NA          9985037 COLO~          8 MONTRO~
## 5 26-JUN-92 12:51:00 26-JUN-92 12:~ NA          9985038 COLO~          8 EL PASO
## 6 26-JUN-92 18:40:00 26-JUN-92 18:~ NA          9985040 COLO~          8 WELD
## # ... with 8 more variables: CZ_TYPE <chr>, CZ_FIPS <dbl>, EVENT_TYPE <chr>,
## #   SOURCE <lgl>, BEGIN_LAT <dbl>, BEGIN_LON <dbl>, END_LAT <dbl>,
## #   END_LON <dbl>
```

3.Arrange the data by the state name (STATE)

```
#let's arrange the data by the state name (STATE)
storm_event_details_92_subset %>% arrange(STATE)
```

```
## # A tibble: 13,534 x 15
##    BEGIN_DATE_TIME    END_DATE_TIME EPISODE_ID EVENT_ID STATE STATE_FIPS CZ_NAME
##    <chr>              <chr>         <lgl>         <dbl> <chr>      <dbl> <chr>
##  1 26-JUN-92 13:22:00 26-JUN-92 13~ NA          9982724 ALAB~          1 MACON
##  2 03-JUL-92 11:00:00 03-JUL-92 11~ NA          9982733 ALAB~          1 TUSCAL~
##  3 18-JUN-92 16:15:00 18-JUN-92 16~ NA          9982499 ALAB~          1 MARSHA~
##  4 04-NOV-92 07:00:00 04-NOV-92 07~ NA          9983918 ALAB~          1 MOBILE
##  5 04-NOV-92 07:10:00 04-NOV-92 07~ NA          9983919 ALAB~          1 MOBILE
##  6 22-NOV-92 08:40:00 22-NOV-92 08~ NA          9982810 ALAB~          1 DEKALB
##  7 22-NOV-92 08:47:00 22-NOV-92 08~ NA          9982811 ALAB~          1 JACKSON
##  8 22-NOV-92 08:50:00 22-NOV-92 08~ NA          9982812 ALAB~          1 TALLAP~
##  9 22-NOV-92 09:05:00 22-NOV-92 09~ NA          9982813 ALAB~          1 RANDOL~
## 10 22-NOV-92 09:10:00 22-NOV-92 09~ NA          9982814 ALAB~          1 CLEBUR~
## # ... with 13,524 more rows, and 8 more variables: CZ_TYPE <chr>,
## #   CZ_FIPS <dbl>, EVENT_TYPE <chr>, SOURCE <lgl>, BEGIN_LAT <dbl>,
## #   BEGIN_LON <dbl>, END_LAT <dbl>, END_LON <dbl>
```

4.Change state and county names to title case

```
#let's change state names to title case (e.g., "New Jersey" instead of "NEW JERSEY")
storm_event_details_92_subset$state_title_case <-
↪  str_to_title(storm_event_details_92_subset$STATE, locale = "en")

head(storm_event_details_92_subset$state_title_case, 100)
```

```
##   [1] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##   [5] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##   [9] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##  [13] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##  [17] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##  [21] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##  [25] "Colorado"       "Colorado"       "Colorado"       "Colorado"
##  [29] "Alabama"        "Alabama"        "Mississippi"    "Louisiana"
##  [33] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [37] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [41] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [45] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [49] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [53] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [57] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [61] "New Jersey"     "New Jersey"     "New Jersey"     "New Jersey"
##  [65] "New Jersey"     "New Jersey"     "Minnesota"      "Louisiana"
##  [69] "Nevada"         "Louisiana"      "New Jersey"     "New Jersey"
##  [73] "New York"       "Kansas"         "New Mexico"     "Missouri"
```

```
##  [77] "Pennsylvania"   "Oklahoma"       "Illinois"       "Oklahoma"
##  [81] "Florida"        "Arkansas"       "Oklahoma"       "Oklahoma"
##  [85] "Tennessee"      "Ohio"           "Ohio"           "Illinois"
##  [89] "Oklahoma"       "Pennsylvania"   "South Carolina" "New York"
##  [93] "North Carolina" "Oklahoma"       "Ohio"           "Ohio"
##  [97] "New Mexico"     "New York"       "Nebraska"       "Georgia"
```

```
#let's change county names to title case
storm_event_details_92_subset$county_title_case <-
↪  str_to_title(storm_event_details_92_subset$CZ_NAME, locale = "en")

head(storm_event_details_92_subset$county_title_case, 100)
```

```
##   [1] "Larimer"         "El Paso"         "Kit Carson"      "Montrose"
##   [5] "El Paso"         "Weld"            "Weld"            "Morgan"
##   [9] "Logan"           "Washington"      "Lincoln"         "El Paso"
##  [13] "Arapahoe"        "Arapahoe"        "Washington"      "El Paso"
##  [17] "El Paso"         "Morgan"          "Lincoln"         "Yuma"
##  [21] "Yuma"            "Arapahoe"        "Saguache"        "Phillips"
##  [25] "Adams"           "Adams"           "Adams"           "Adams"
##  [29] "Macon"           "Tuscaloosa"      "Lowndes"         "Bienville"
##  [33] "Union"           "Burlington"      "Union"           "Warren"
##  [37] "Bergen"          "Hunterdon"       "Gloucester"      "Hunterdon"
##  [41] "Cumberland"      "Atlantic"        "Cape May"        "Morris"
##  [45] "Bergen"          "Middlesex"       "Ocean"           "Monmouth"
##  [49] "Middlesex"       "Atlantic"        "Cumberland"      "Monmouth"
##  [53] "Cape May"        "Cumberland"      "Burlington"      "Burlington"
##  [57] "Middlesex"       "Middlesex"       "Somerset"        "Middlesex"
##  [61] "Union"           "Gloucester"      "Middlesex"       "Morris"
##  [65] "Burlington"      "Cumberland"      "Yellow Medicine" "Bienville"
##  [69] "Nye"             "Livingston"      "Ocean"           "Somerset"
##  [73] "Wayne"           "Leavenworth"     "Eddy"            "St. Louis"
##  [77] "Northampton"     "Logan"           "Hancock"         "Pontotoc"
##  [81] "Jackson"         "Boone"           "Pittsburg"       "Oklahoma"
##  [85] "Meigs"           "Crawford"        "Putnam"          "Morgan"
##  [89] "Noble"           "Beaver"          "Charleston"      "Broome"
##  [93] "Greene"          "Oklahoma"        "Summit"          "Cuyahoga"
##  [97] "Eddy"            "Fulton"          "Thurston"        "Cobb"
```

5.Limit to the events listed by county FIPS (CZ_TYPE of "C") and then remove the CZ_TYPE column

```
storm_event_details_92_subset %>% filter(CZ_TYPE == "C") %>% select(-CZ_TYPE)
```

```
## # A tibble: 13,534 x 16
##    BEGIN_DATE_TIME    END_DATE_TIME EPISODE_ID EVENT_ID STATE STATE_FIPS CZ_NAME
##    <chr>              <chr>         <lgl>         <dbl> <chr>      <dbl> <chr>
##  1 24-JUN-92 15:11:00 24-JUN-92 15~ NA          9985034 COLO~          8 LARIMER
##  2 24-JUN-92 18:27:00 24-JUN-92 18~ NA          9985035 COLO~          8 EL PASO
##  3 24-JUN-92 19:43:00 24-JUN-92 19~ NA          9985036 COLO~          8 KIT CA~
```

```
##  4 25-JUN-92 19:50:00 25-JUN-92 19~ NA          9985037 COLO~          8 MONTRO~
##  5 26-JUN-92 12:51:00 26-JUN-92 12~ NA          9985038 COLO~          8 EL PASO
##  6 26-JUN-92 18:40:00 26-JUN-92 18~ NA          9985040 COLO~          8 WELD
##  7 26-JUN-92 18:40:00 26-JUN-92 18~ NA          9985041 COLO~          8 WELD
##  8 26-JUN-92 18:43:00 26-JUN-92 18~ NA          9985042 COLO~          8 MORGAN
##  9 26-JUN-92 20:10:00 26-JUN-92 20~ NA          9985043 COLO~          8 LOGAN
## 10 27-JUN-92 14:28:00 27-JUN-92 14~ NA          9985044 COLO~          8 WASHIN~
## # ... with 13,524 more rows, and 9 more variables: CZ_FIPS <dbl>,
## #   EVENT_TYPE <chr>, SOURCE <lgl>, BEGIN_LAT <dbl>, BEGIN_LON <dbl>,
## #   END_LAT <dbl>, END_LON <dbl>, state_title_case <chr>,
## #   county_title_case <chr>
```

6. Pad the state and county FIPS with a "0" at the beginning and then unite the two columns to make one fips column with the 5 or 6-digit county FIPS code

```
#let's pad state FIPS
storm_event_details_92_subset$STATE_FIPS_padded <-
↪  str_pad(storm_event_details_92_subset$STATE_FIPS, width =3, side = "left", pad = "0")

head(storm_event_details_92_subset$STATE_FIPS_padded, 100)
```

```
##   [1] "008" "008" "008" "008" "008" "008" "008" "008" "008" "008" "008" "008"
##  [13] "008" "008" "008" "008" "008" "008" "008" "008" "008" "008" "008" "008"
##  [25] "008" "008" "008" "008" "001" "001" "028" "022" "034" "034" "034" "034"
##  [37] "034" "034" "034" "034" "034" "034" "034" "034" "034" "034" "034" "034"
##  [49] "034" "034" "034" "034" "034" "034" "034" "034" "034" "034" "034" "034"
##  [61] "034" "034" "034" "034" "034" "034" "027" "022" "032" "022" "034" "034"
##  [73] "036" "020" "035" "029" "042" "040" "017" "040" "012" "005" "040" "040"
##  [85] "047" "039" "039" "017" "040" "042" "045" "036" "037" "040" "039" "039"
##  [97] "035" "036" "031" "013"
```

```
#let's pad county FIPS
storm_event_details_92_subset$CZ_FIPS_padded <-
↪  str_pad(storm_event_details_92_subset$CZ_FIPS, width =3, side = "left", pad = "0")

head(storm_event_details_92_subset$CZ_FIPS_padded, 100)
```

```
##   [1] "069" "041" "063" "085" "041" "123" "123" "087" "075" "121" "073" "041"
##  [13] "005" "005" "121" "041" "041" "087" "073" "125" "125" "005" "109" "095"
##  [25] "001" "001" "001" "001" "087" "125" "087" "013" "039" "005" "039" "041"
##  [37] "003" "019" "015" "019" "011" "001" "009" "027" "003" "023" "029" "025"
##  [49] "023" "001" "011" "025" "009" "011" "005" "005" "023" "023" "035" "023"
##  [61] "039" "015" "023" "027" "005" "011" "173" "013" "023" "063" "029" "035"
##  [73] "117" "103" "015" "189" "095" "083" "067" "123" "063" "009" "121" "109"
##  [85] "121" "033" "137" "137" "103" "007" "019" "007" "079" "109" "153" "035"
##  [97] "015" "035" "173" "067"
```

```
#let's union the padded state and county FIPS to make one fips column
storm_event_details_92_subset <- storm_event_details_92_subset %>%
  unite("fips", c(STATE_FIPS_padded,CZ_FIPS_padded), remove = FALSE, sep = "")
```

```
head(storm_event_details_92_subset$fips, 100)
```

```
##   [1] "008069" "008041" "008063" "008085" "008041" "008123" "008123" "008087"
##   [9] "008075" "008121" "008073" "008041" "008005" "008005" "008121" "008041"
##  [17] "008041" "008087" "008073" "008125" "008125" "008005" "008109" "008095"
##  [25] "008001" "008001" "008001" "008001" "001087" "001125" "028087" "022013"
##  [33] "034039" "034005" "034039" "034041" "034003" "034019" "034015" "034019"
##  [41] "034011" "034001" "034009" "034027" "034003" "034023" "034029" "034025"
##  [49] "034023" "034001" "034011" "034025" "034009" "034011" "034005" "034005"
##  [57] "034023" "034023" "034035" "034023" "034039" "034015" "034023" "034027"
##  [65] "034005" "034011" "027173" "022013" "032023" "022063" "034029" "034035"
##  [73] "036117" "020103" "035015" "029189" "042095" "040083" "017067" "040123"
##  [81] "012063" "005009" "040121" "040109" "047121" "039033" "039137" "017137"
##  [89] "040103" "042007" "045019" "036007" "037079" "040109" "039153" "039035"
##  [97] "035015" "036035" "031173" "013067"
```

7.Change all the column names to lower case

```
#let's rename all column names to lower case
storm_event_details_92_subset <- storm_event_details_92_subset %>%  rename_all(tolower)

#let's check if column names have been changed to lower case
colnames(storm_event_details_92_subset)
```

```
##  [1] "begin_date_time"   "end_date_time"     "episode_id"
##  [4] "event_id"          "state"             "state_fips"
##  [7] "cz_name"           "cz_type"           "cz_fips"
## [10] "event_type"        "source"            "begin_lat"
## [13] "begin_lon"         "end_lat"           "end_lon"
## [16] "state_title_case"  "county_title_case" "fips"
## [19] "state_fips_padded" "cz_fips_padded"
```

8.There is data that comes with base R on U.S. states (data("state")). Use that to create a dataframe with these three columns: state name, area, and region

```
us_state_data <- data.frame(state=state.name, region=state.region, area=state.area)

head(us_state_data)
```

```
##        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
```

9.Create a dataframe with the number of events per state in the year of birth. Merge in the state information dataframe created in step 8. Remove any states that are not in the state information dataframe.

```
#let's create a dataframe with the number of events per state in the year of my birth
event_freq_per_state <- data.frame(table(storm_event_details_92_subset$state))
head(event_freq_per_state)
```

```
##          Var1 Freq
## 1     ALABAMA  347
## 2     ARIZONA   82
## 3    ARKANSAS  567
## 4  CALIFORNIA   33
## 5    COLORADO  266
## 6 CONNECTICUT   46
```

```
#Let's merge in the state information dataframe created in step 8. Let's remove any
↪  states that are not in the state information dataframe.

#Before doing merge let's make sure column names and data matches

#let's rename column name Var1 to state in event_freq_per_state dataframe
event_freq_per_state <- event_freq_per_state %>% rename(state = Var1)
head(event_freq_per_state)
```

```
##         state Freq
## 1     ALABAMA  347
## 2     ARIZONA   82
## 3    ARKANSAS  567
## 4  CALIFORNIA   33
## 5    COLORADO  266
## 6 CONNECTICUT   46
```

```
#let's change the case of the states in us_state_data from lower case to upper case
↪  before merge
us_state_data$state <- str_to_upper(us_state_data$state, locale = "en")
head(us_state_data)
```

```
##        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
```

```
#As the column names and case of the data matches for column state in both us_state_data,
↪  event_freq_per_state data frames, let's merge them
state_storms_merged_data <- merge(x=event_freq_per_state, y=us_state_data, by.x =
↪  "state", by.y = "state")
```

```
head(state_storms_merged_data,10)
```

```
##          state Freq    region   area
## 1      ALABAMA  347     South  51609
## 2      ARIZONA   82      West 113909
## 3     ARKANSAS  567     South  53104
## 4   CALIFORNIA   33      West 158693
## 5     COLORADO  266      West 104247
## 6  CONNECTICUT   46 Northeast   5009
## 7     DELAWARE   23     South   2057
## 8      FLORIDA  368     South  58560
## 9      GEORGIA  393     South  58876
## 10       IDAHO   68      West  83557
```

Create a plot of number of storm events in 1992 vs land area in square miles

```
state_storms_merged_data %>%
  ggplot(aes(x=area, y=Freq)) +
  geom_point(aes(color = region)) +
  labs(x = "Land area (square miles)",
       y = "# storm events in 1992")
```