CSCI 456 Group 2 - Project #1

Jacob Collins, Naina K.

21 February, 2025

Data Exploration and Cleaning

Select a Data Source:

Department of Agriculture- National USFS Fire Occurrence Point

Formulate Research Questions:

Develop 3-5 research questions that you aim to answer through your analysis. Ensure that these questions are specific, measurable, and relevant to the dataset you have chosen.

- How has the size and frequency of fires changed over time?
- What regions are more susceptible to forest fires, in terms of longitude and latitude?
- What is the relationship between wildfire occurrence and proximity to human settlements?

Data Variable Description:

Explain the variables in your dataset you intend to analyze.

- `STATCAUSE` - Cause: indicates cause of fire , categorical-qualitative data
- `TOTALACRES` - Total acres: represents area burned by the wildfire, quantitative- continuous data
- `LATDD83` - Latitude: north-south location, quantitative- continuous data
- `LONGDD83` - Longitude: east-west location, quantitative- continuous data
- `REVDATE` - Fire year: year wildfire occurred, quantitative- discrete data

Preparation:

Describe how you did data preparation. This may include: - Checking and handling missing values - Encoding categorical variables

Initial csv read

```
fires <- read.csv("fires.csv", fill=TRUE)
fires <- fires %>% select(TOTALACRES, LATDD83, LONGDD83, STATCAUSE, REVDATE)
nrow(fires)
```

```
## [1] 582034
```

```
summary(fires)
```

```
##    TOTALACRES          LATDD83           LONGDD83           STATCAUSE
##  Min.   :     0.0   Min.   :  -117.2   Min.   :-1038467.0   Length:582034
##  1st Qu.:     0.1   1st Qu.:    35.1   1st Qu.:    -120.6   Class :character
##  Median :     0.1   Median :    38.6   Median :    -116.2   Mode  :character
##  Mean   :   118.8   Mean   :    42.9   Mean   :    -109.8
##  3rd Qu.:     1.0   3rd Qu.:    43.7   3rd Qu.:    -109.8
##  Max.   :963309.0   Max.   :438897.0   Max.   : 1011212.8
##  NA's   :3640       NA's   :715        NA's   :715
##    REVDATE
##  Length:582034
##  Class :character
##  Mode  :character
##
##
##
##
```

There are 582,034 observations.

Cleaning `REVDATE`

```
fires$REVDATE <- as.Date(fires$REVDATE)
summary(fires$REVDATE)
```

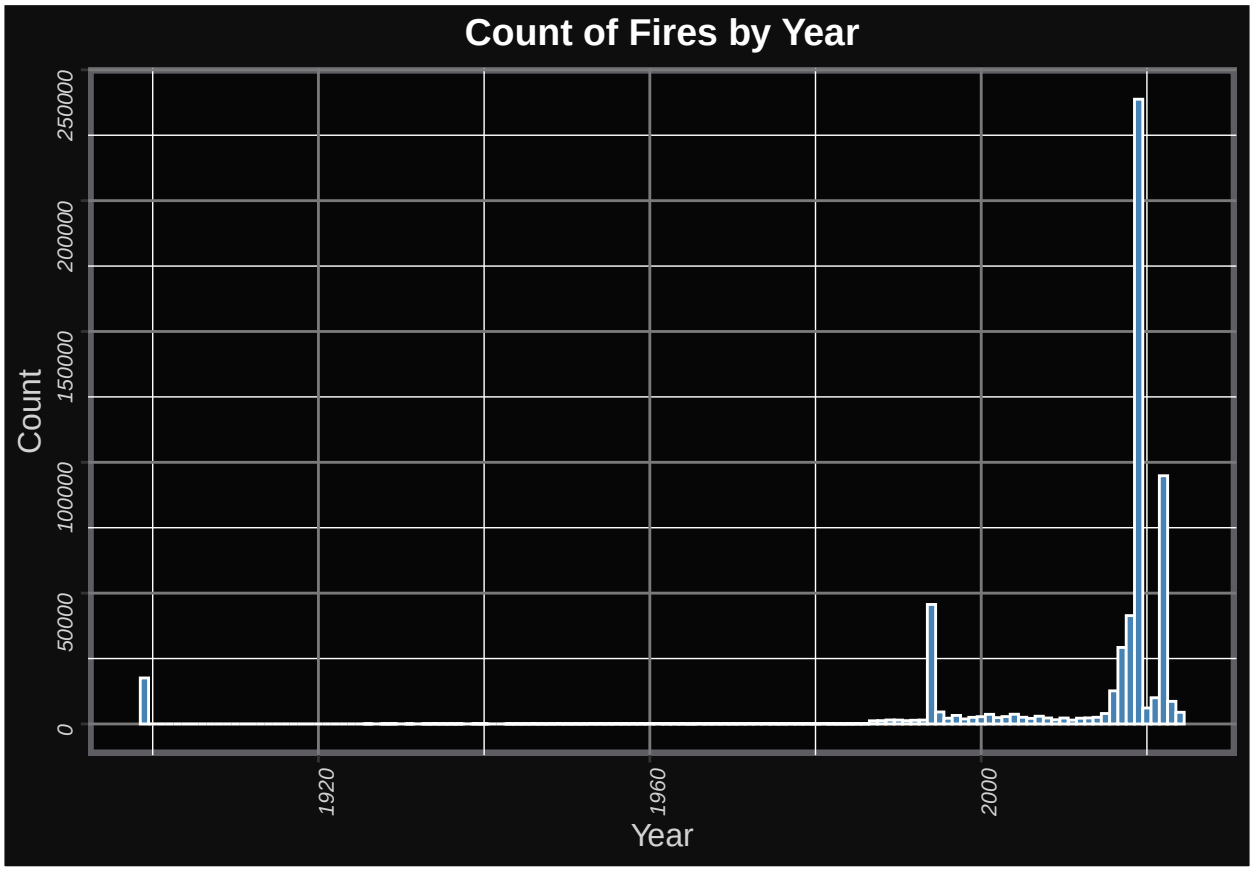
```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "0218-11-01" "2017-01-12" "2019-01-02" "2026-03-15" "2019-01-02" "9999-02-01"
##         NA's
##       "3710"
```

It seems there are many years incorrectly labeled.

```
# No real data in the future
fires$REVDATE[as.numeric(format(fires$REVDATE, "%Y")) > 2025] <- NA
# No real data before 1500 is a safe assumption
fires$REVDATE[as.numeric(format(fires$REVDATE, "%Y")) < 1500] <- NA

fires$REV.YEAR <- format(fires$REVDATE, "%Y")
fires$REV.YEAR <- as.numeric(fires$REV.YEAR)
fires %>% ggplot(aes(x = REV.YEAR)) +
  geom_histogram(binwidth=1, fill = "steelblue", color="white", na.rm=TRUE) +
  default_theme +
  labs(
    x = "Year",
    y = "Count",
    title = "Count of Fires by Year",
  ) + theme(
```

```
    axis.text.x = element_text( # X-Axis Labels
      face = "italic", color = "lightgray",
      size = 8, angle = 90)
  )
```



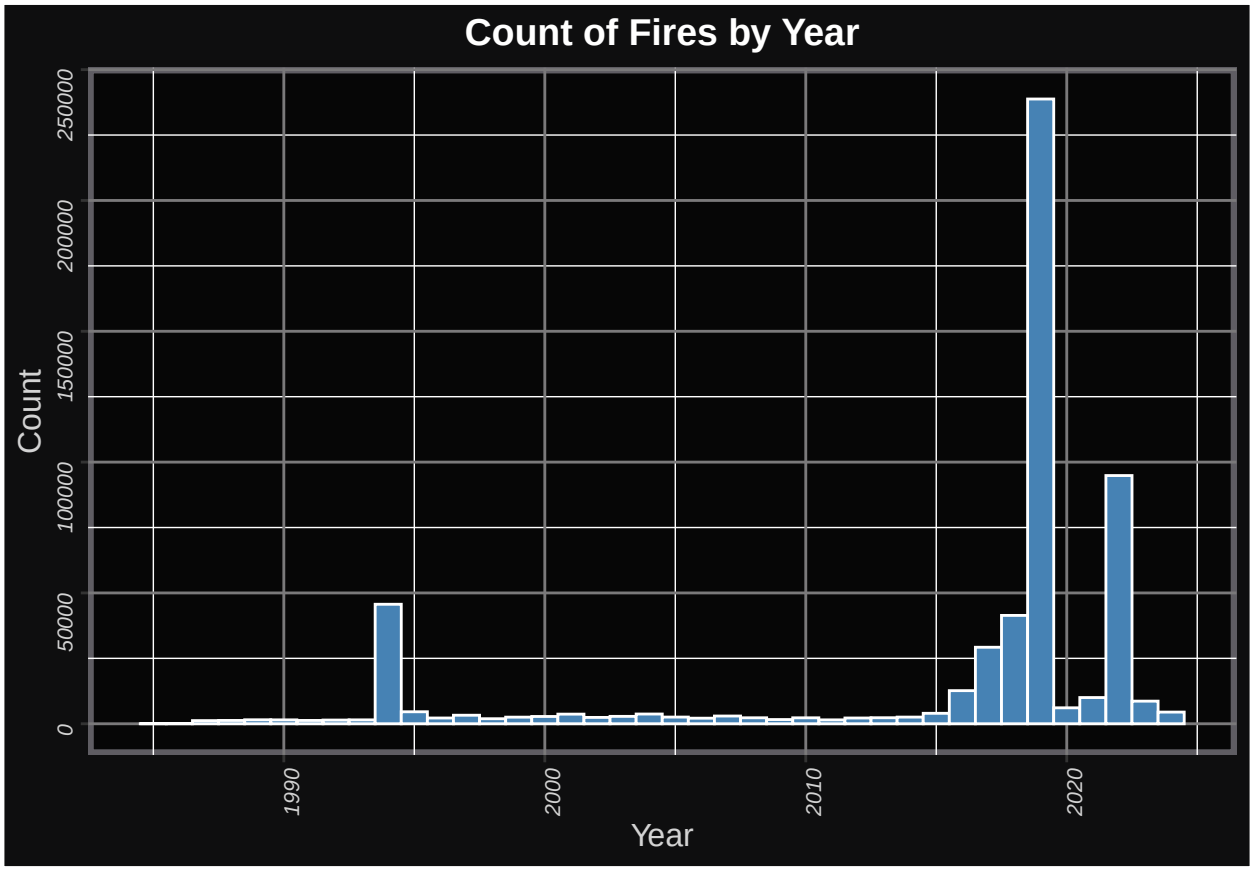
Viewing this plot shows us where the bulk of the data is, and we can prune around it.

```
# This deleted REVDATE and filled around Y, M, & D with the current md, yd, ym, respectively.
#fires <- fires %>% separate(REVDATE, into = c("REV.Y", "REV.M", "REV.D"), sep = "-")
#fires$REV.Y <- as.Date(fires$REV.Y, format = "%Y")
#fires$REV.M <- as.Date(fires$REV.M, format = "%m")
#fires$REV.D <- as.Date(fires$REV.D, format = "%d")
#summary(fires %>% select(REV.Y, REV.M, REV.D))

# No real data before 1985
fires$REVDATE[as.numeric(format(fires$REVDATE, "%Y")) < 1985] <- NA
summary(fires$REVDATE)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "1985-04-19" "2017-01-24" "2019-01-02" "2015-08-22" "2019-01-02" "2024-07-30"
##         NA's
##      "24186"
```

```r
fires$REV.YEAR <- format(fires$REVDATE, "%Y")
fires$REV.YEAR <- as.numeric(fires$REV.YEAR)
fires %>% ggplot(aes(x = REV.YEAR)) +
  geom_histogram(binwidth=1, fill = "steelblue", color="white", na.rm=TRUE) +
  default_theme +
  labs(
    x = "Year",
    y = "Count",
    title = "Count of Fires by Year",
  ) + theme(
    axis.text.x = element_text( # X-Axis Labels
      face = "italic", color = "lightgray",
      size = 8, angle = 90)
  )
```

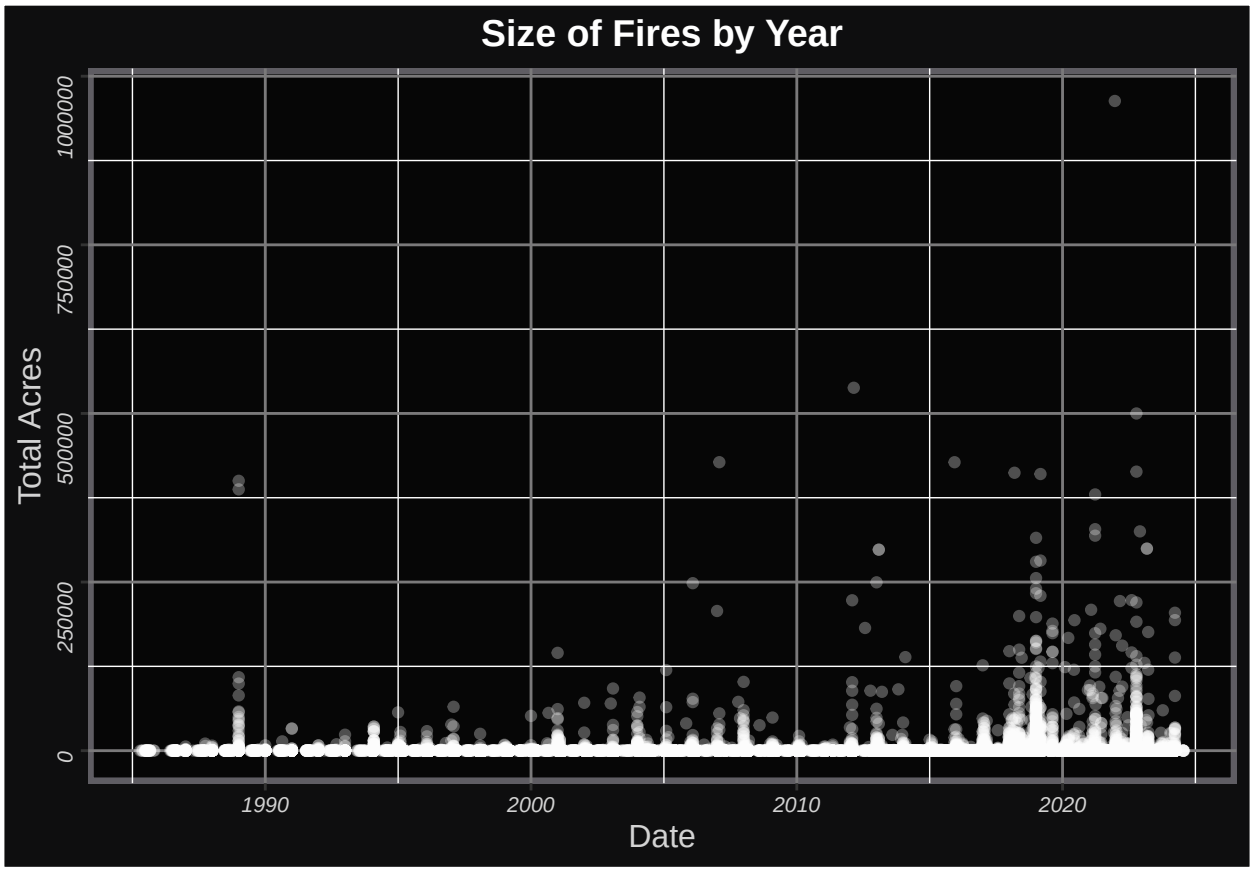


Cleaning Acres

```r
summary(fires$TOTALACRES)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##       0.0       0.1       0.1     118.8       1.0 963309.0      3640
```

```
fires %>% ggplot(mapping=aes(x=REVDATE, y=TOTALACRES)) +
  geom_point(color="white", alpha=0.3, na.rm=TRUE) +
  default_theme +
  labs(
    x = "Date",
    y = "Total Acres",
    title = "Size of Fires by Year"
  )
```



Acres looks alright.

Cleaning Longitude and Latitude

```
summary(fires %>% select(LATDD83, LONGDD83))
```

```
##     LATDD83           LONGDD83
##  Min.   :  -117.2   Min.   :-1038467.0
##  1st Qu.:    35.1   1st Qu.:    -120.6
##  Median :    38.6   Median :    -116.2
##  Mean   :    42.9   Mean   :    -109.8
##  3rd Qu.:    43.7   3rd Qu.:    -109.8
##  Max.   :438897.0   Max.   : 1011212.8
##  NA's   :715        NA's   :715
```

The valid range for latitude is -90 to 90. The valid range for longitude is -180 to 180.

```
fires$LATDD83[fires$LATDD83 > 90] <- NA
fires$LATDD83[fires$LATDD83 < -90] <- NA
fires$LONGDD83[fires$LONGDD83 > 180] <- NA
fires$LONGDD83[fires$LONGDD83 < -180] <- NA
summary(fires %>% select(LATDD83, LONGDD83))
```

```
##     LATDD83          LONGDD83
##  Min.   :-89.77   Min.   :-150.4
##  1st Qu.: 35.05   1st Qu.:-120.6
##  Median : 38.56   Median :-116.2
##  Mean   : 39.06   Mean   :-111.3
##  3rd Qu.: 43.67   3rd Qu.:-109.8
##  Max.   : 84.92   Max.   : 110.0
##  NA's   :925      NA's   :723
```

Cleaning STATCAUSE

```
unique(fires$STATCAUSE)
```

```
##  [1] "Camping"              "Lightning"            "Undetermined"
##  [4] "Smoking"              "Debris/Open Burning"  "Other Human Cause"
##  [7] "Incendiary"           "Equipment"            ""
## [10] "3"                    "1"                    "Railroad"
## [13] "Other Natural Cause"  "Utilities"            "2"
## [16] "5"                    "4"                    "9"
## [19] "6"                    "7"                    "8"
## [22] "Firearms/Weapons"     "Undertermined"        "Natural"
## [25] "Human"                "Miscellaneous"        "Debris Burning"
## [28] "Equipment Use"        "Children"             "Firearms/Weapons "
## [31] "0"                    "9 -  Miscellaneous"   "5 -  Debris Burning"
## [34] "Camping "             "1 - Lightning"        "9 - Miscellaneous"
## [37] "4 - Campfire"         "5 - Debris burning"   "Arson"
## [40] "Campfire"             "7-Arson"              "5-Debris burning"
## [43] " Undetermined"        "Powgen/trans/distrib" "Equip/vehicle use"
## [46] "Other causes"         "Investigated But Und" "Cause not Identified"
```

These can be joined by matching number or obvious typos like two spaces instead of one.

Blank results, miscellaneous, undet., etc. shall be marked NA.