Data Ingestion/EDA

Fatimah Niyas

2025-04-07

Data Ingestion/Cleaning and EDA

Loading in necessary libraries

```r
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(ggplot2)
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.3.3
```

```
##
## Attaching package: 'maps'
##
## The following object is masked from 'package:purrr':
##
##     map
```

```r
library(dplyr)
```

Loading in meteorite landings data

```r
setwd("~/DATA 205")
meteorite <- read.csv('Meteorite_Landings.csv')
```

Begin cleaning data

```r
meteorite1 <- meteorite |>
  filter(!is.na(mass..g.)) |> # filter out any NAs
  filter(!is.na(year)) |>
  filter(!is.na(reclat)) |>
  filter(!is.na(reclong)) |>
  rename(mass = mass..g.)|> # renaming for easier access
  filter(year > 1850 & year < 2100) |> # filter the years to avoid big outliers
  filter(reclat >= -90 & reclat <= 90, reclong >= -180 & reclong <= 180) |> # make sure the coordinates
  filter(mass > 0 & mass < 1e6) #filter out mass values
head(meteorite1)
```

```
##       name  id nametype    recclass   mass fall year    reclat    reclong
## 1   Aachen   1    Valid          L5     21 Fell 1880  50.77500    6.08333
## 2   Aarhus   2    Valid          H6    720 Fell 1951  56.18333   10.23333
## 3     Abee   6    Valid         EH4 107000 Fell 1952  54.21667 -113.00000
## 4 Acapulco  10    Valid Acapulcoite   1914 Fell 1976  16.88333  -99.90000
## 5  Achiras 370    Valid          L6    780 Fell 1902 -33.16667  -64.95000
## 6 Adhi Kot 379    Valid         EH4   4239 Fell 1919  32.10000   71.80000
##            GeoLocation
## 1    (50.775, 6.08333)
## 2 (56.18333, 10.23333)
## 3   (54.21667, -113.0)
## 4    (16.88333, -99.9)
## 5  (-33.16667, -64.95)
## 6         (32.1, 71.8)
```

Summary statistics of cleaned data

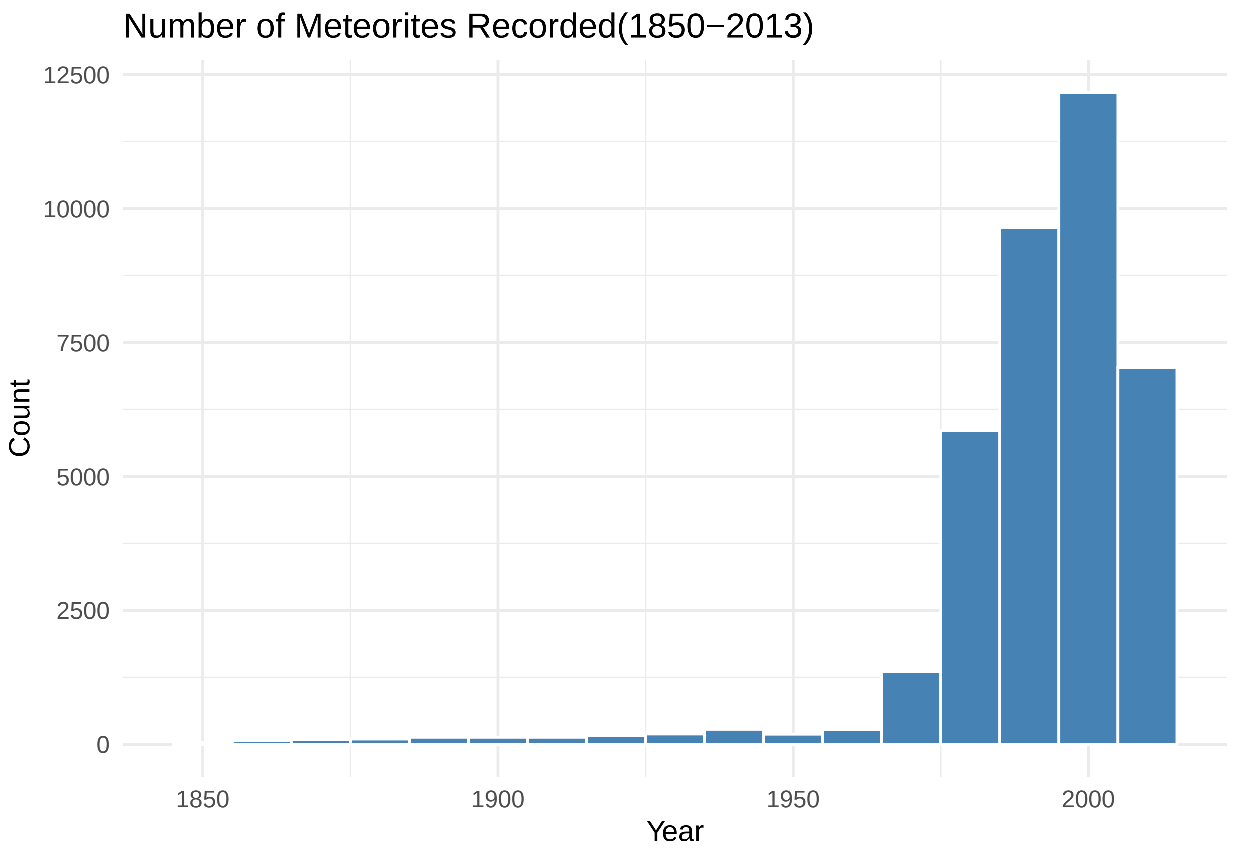
```r
summary(meteorite1)
```

```
##      name                 id          nametype           recclass
##  Length:37838       Min.   :    1   Length:37838       Length:37838
##  Class :character   1st Qu.:10857   Class :character   Class :character
##  Mode  :character   Median :21779   Mode  :character   Mode  :character
##                     Mean   :25399
##                     3rd Qu.:39946
##                     Max.   :57458
##       mass              fall                year          reclat
##  Min.   :     0.0   Length:37838       Min.   :1851   Min.   :-87.37
##  1st Qu.:     6.6   Class :character   1st Qu.:1986   1st Qu.:-76.72
##  Median :    28.6   Mode  :character   Median :1996   Median :-71.50
##  Mean   :  2452.6                      Mean   :1991   Mean   :-40.15
##  3rd Qu.:   180.0                      3rd Qu.:2003   3rd Qu.:  0.00
##  Max.   :997000.0                      Max.   :2013   Max.   : 81.17
##     reclong         GeoLocation
##  Min.   :-165.43   Length:37838
##  1st Qu.:   0.00   Class :character
##  Median :  35.67   Mode  :character
##  Mean   :  61.73
##  3rd Qu.: 157.17
##  Max.   : 178.20
```

Begin Exploratory Data Analysis

Visualization on meteorite count over years

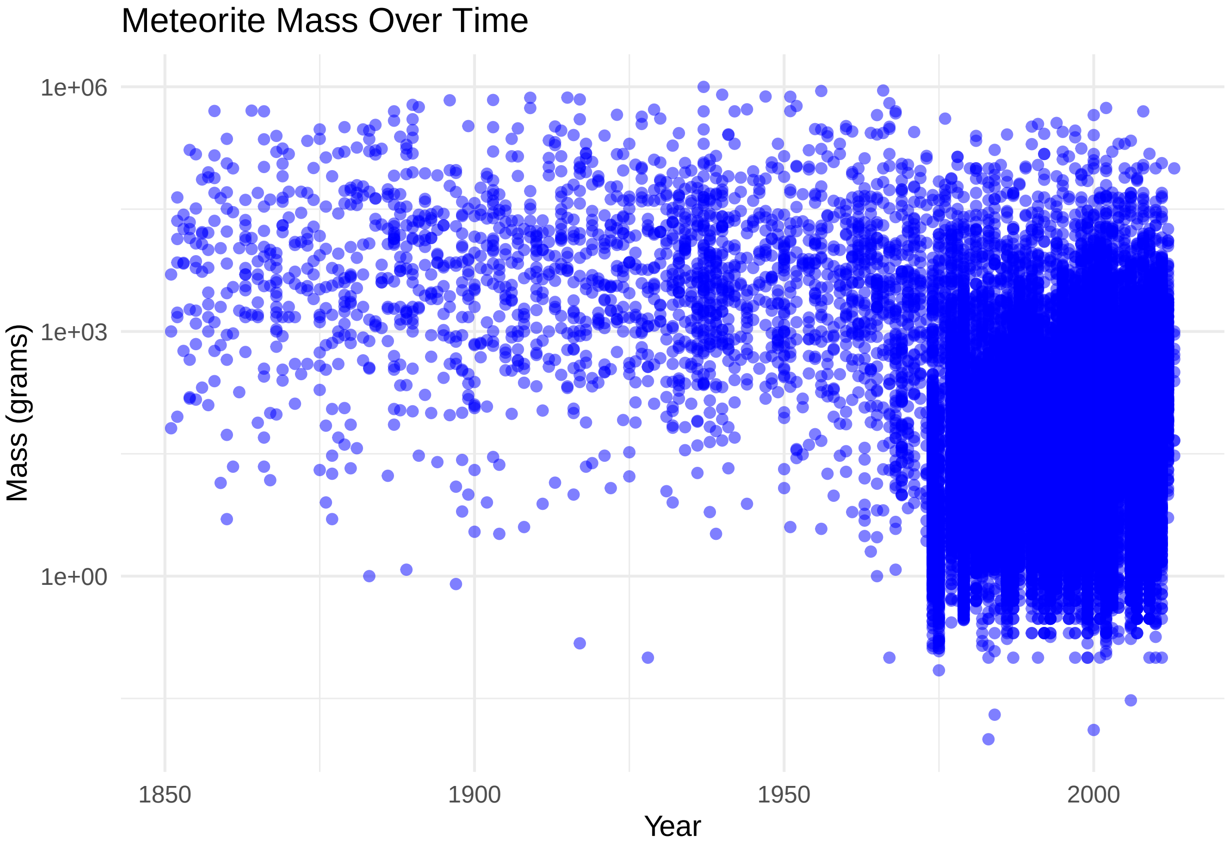
```
ggplot(meteorite1, aes(x = year)) + # visualizing meteorite count over years
  geom_histogram(binwidth = 10, fill = "steelblue", color = "white") +
  labs(title = "Number of Meteorites Recorded(1850-2013)",
       x = "Year", y = "Count") +
  theme_minimal()
```



As we can see here, the data is skewed left, clustered in more recent years likely because of advancements in technology, increased global scientific interest, and improved tracking and reporting systems. In the past, many meteorite events may have gone unnoticed or undocumented, especially in remote or less-populated areas. As scientific tools have developed, more meteorite landings have been detected, recorded, and analyzed—leading to a sharp increase in entries during the 20th and 21st centuries.

Visualization on meteorite mass over years

```
ggplot(meteorite1, aes(x = year, y = mass)) +
  geom_point(alpha = 0.5, color = "blue") +
  scale_y_log10() +
  labs(title = "Meteorite Mass Over Time",
       x = "Year",
       y = "Mass (grams)") +
  theme_minimal()
```



Here, we can see that in more recent years, the number of meteorites with smaller masses has increased also due to advancements in detection technology and improved reporting systems. Smaller meteorites that would have gone unnoticed in the past are now being recovered thanks to tools like metal detectors, satellite tracking, etc.

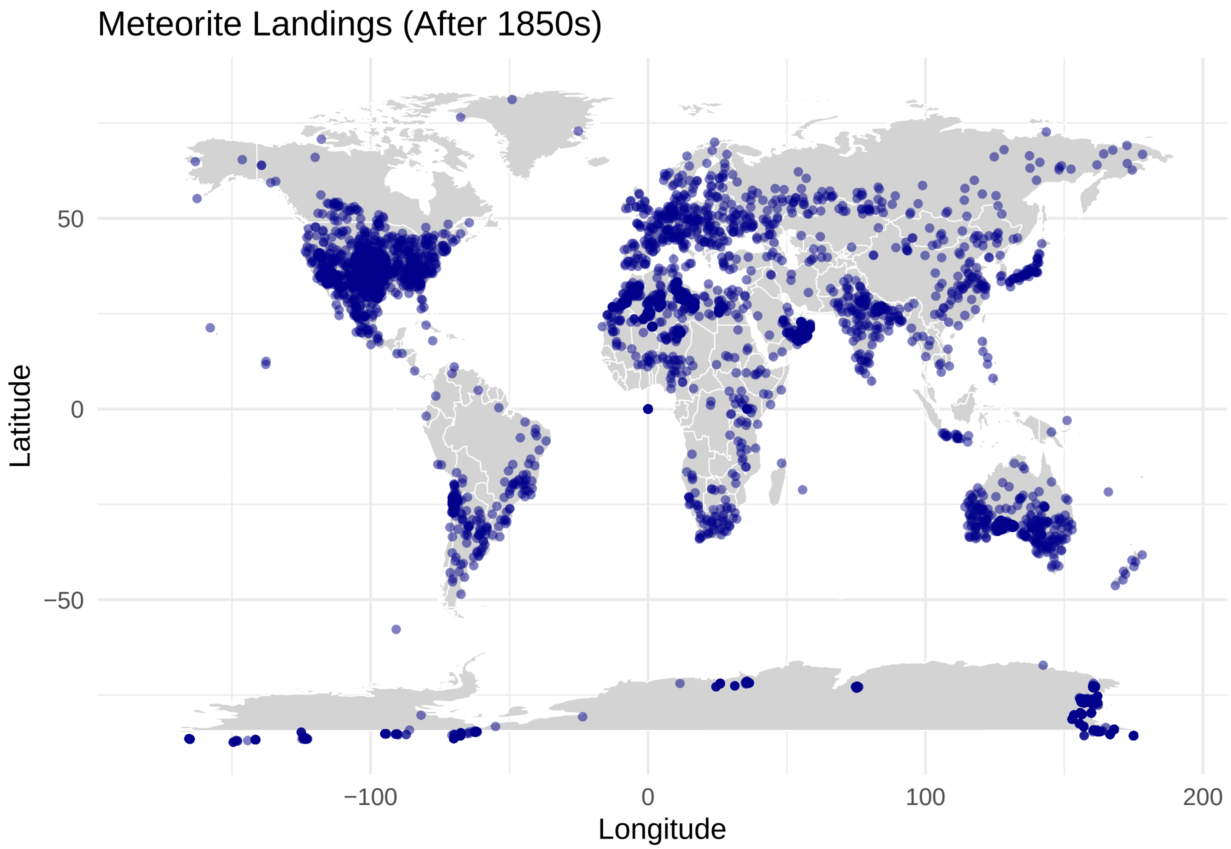
Meteorite locations on a map

```
world_map <- map_data("world")

ggplot() +
  geom_map(data = world_map, map = world_map,
           aes(x = long, y = lat, map_id = region),
           fill = "lightgray", color = "white", size = 0.2) +
  geom_point(data = meteorite1,
             aes(x = reclong, y = reclat),
             color = "darkblue", alpha = 0.5, size = 1) +
  labs(title = "Meteorite Landings (After 1850s)",
       x = "Longitude", y = "Latitude") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

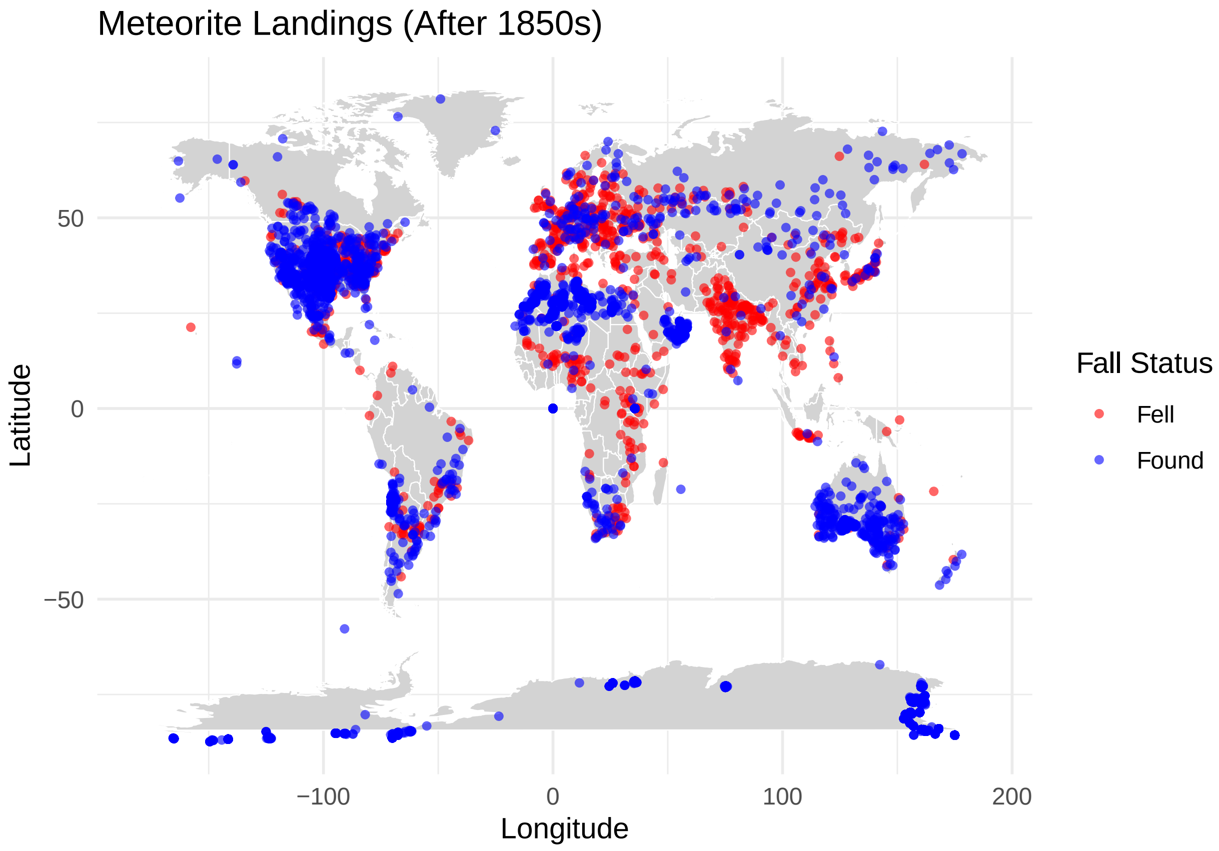
```
## Warning in geom_map(data = world_map, map = world_map, aes(x = long, y = lat, :
## Ignoring unknown aesthetics: x and y
```



Meteorite locations on a map colored by 'Fell' vs. 'Found'

```
ggplot() +
  geom_map(data = world_map, map = world_map,
           aes(x = long, y = lat, map_id = region),
           fill = "lightgray", color = "white", size = 0.2) +
  geom_point(data = meteorite1,
             aes(x = reclong, y = reclat, color = fall),
             alpha = 0.6, size = 1) +
  scale_color_manual(values = c("Fell" = "red", "Found" = "blue")) +
  labs(title = "Meteorite Landings (After 1850s)",
       x = "Longitude", y = "Latitude", color = "Fall Status") +
  theme_minimal()
```

```
## Warning in geom_map(data = world_map, map = world_map, aes(x = long, y = lat, :
## Ignoring unknown aesthetics: x and y
```


Meteorite Landings (After 1850s)
Latitude
50
0
−50
−100
0
100
200
Longitude
Fall Status
● Fell
● Found