ARJUN MOITRA

DATA6200 Assignment 2 Report

Wildfires have become a defining environmental issue for British Columbia. In this report I use BC Wildfire Service point data (KMZ files) from 2013–2024, together with summary statistics scraped from the BCWS website, to look for patterns in how fires behave over time and space.

The analysis focuses on two questions:

1. How have wildfire **frequency**, **size**, and **spatial distribution** changed from 2013–2024?

2. How do ignition causes – especially **human** vs **lightning** – relate to those patterns?

Section 1. Data and Methods

Data sources

- **KMZ point files (2013–2024)** from the BC Wildfire Service, containing one point per fire plus a "Description" field with embedded HTML.I used a combination of the files on the archive and on the original link.

- **Official yearly summaries** scraped from the BCWS wildfire statistics page, including total fires, total hectares burned, and counts by cause.

Key wrangling steps

1. **KMZ → KML → sf points**
   All KMZ files were unzipped in R, their doc.kml files were read with sf, and all non-empty layers were combined into one large sf object.

2. **Geometry normalisation**
   Geometries were converted to simple POINTs in WGS84, dropping any Z/M dimensions and extracting longitude/latitude for plotting. A rough BC bounding box (-140°W to -113°W, 48–60°N) removed stray points.

3. **Parsing hierarchical HTML**
   The "Description" column is an HTML table. I used read_html(), xml_find_all("//tr"), and html_text() to pull out key–value pairs, then pivot_wider() to turn them into real columns. From these I constructed:

   a. FIRE_YEAR (coalesced from several possible year fields),
   b. SIZE_HA (current fire size in hectares), and
   c. CAUSE_GENERAL (raw cause descriptions).

4. **De-duplication**
   Because the KMZ bundles overlap in time, the same fire can appear multiple times. I sorted by

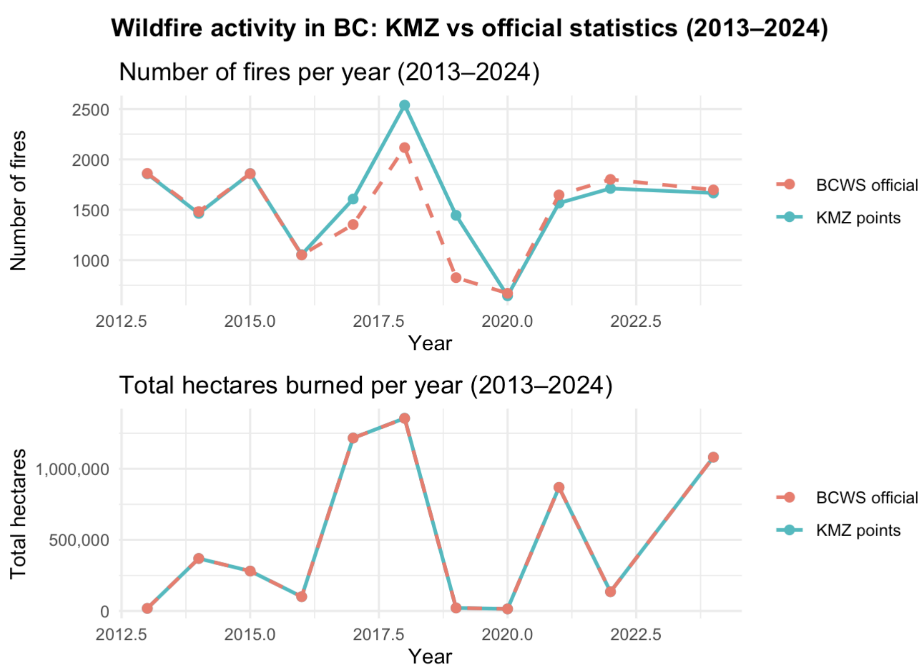
FIRE_YEAR, Name, and SIZE_HA, then kept only the largest record per fire name/year combination. This produced a unique-fire dataset for 2013–2024.

5. **Cause grouping and filters**
   Cause descriptions were lower-cased and grouped into:

   a. **Human** (anything mentioning “human”, “person”, “people”),
   b. **Lightning** (including “natural”), and
   c. **Unknown/None** (unknown, under investigation, not available, etc.).

6. When comparing causes, I restrict to **known causes only** (Human vs Lightning) to avoid being dominated by the large “Unknown/None” category in recent years.

7. **Validation against official statistics**
   I summarised the KMZ data by year (number of fires and total area burned) and joined it to the scraped BCWS table. A two-panel line plot compares KMZ vs official counts and hectares over 2013–2024. The close agreement in both panels gives confidence that the cleaned KMZ dataset is representative.



-For both fire counts and total hectares burned, the KMZ points derived values track the official series closely across 2013–2024, including the extreme peaks in 2017–2018 and 2021 and the relatively mild 2020 season. There are small differences in some years, which likely reflect the way I deduplicated fires and filtered to a BC bounding box but the overall pattern is the same. This suggests that the point data was reliable enough to use for more detailed analysis.

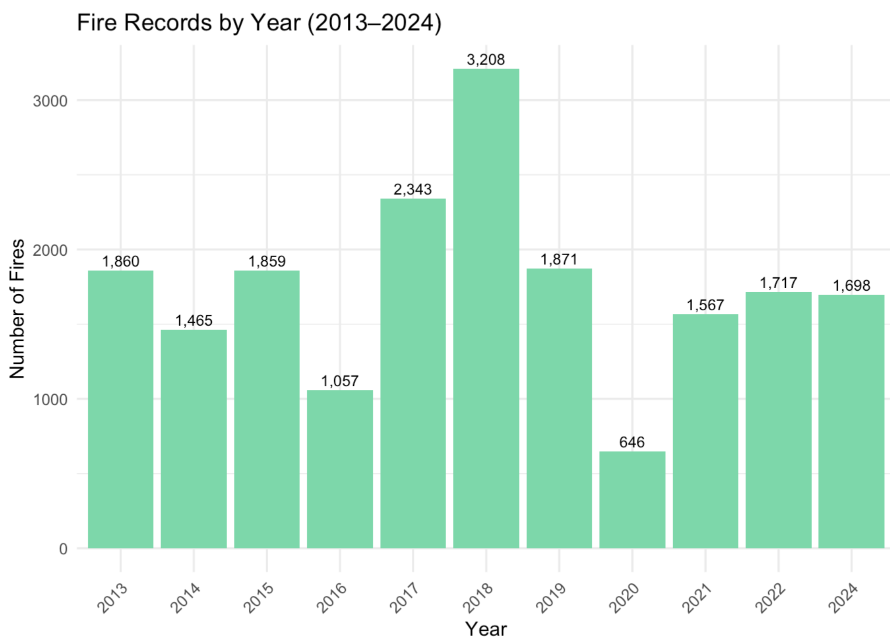
Section 2. RESULTS

1. How have fire frequency and size changed over time?

Fire frequency

The **Fire Records by Year** bar chart shows how many fires are recorded in the KMZ data each year between 2013 and 2024. Fire counts fluctuate substantially:

- Moderate activity appears in the early years (2013–2016).
- **2017 and 2018** stand out as very active seasons with the highest counts, consistent with widely reported extreme fire years in BC.
- Activity drops sharply around **2019–2020**, then rebounds in **2021–2022** and remains relatively high.



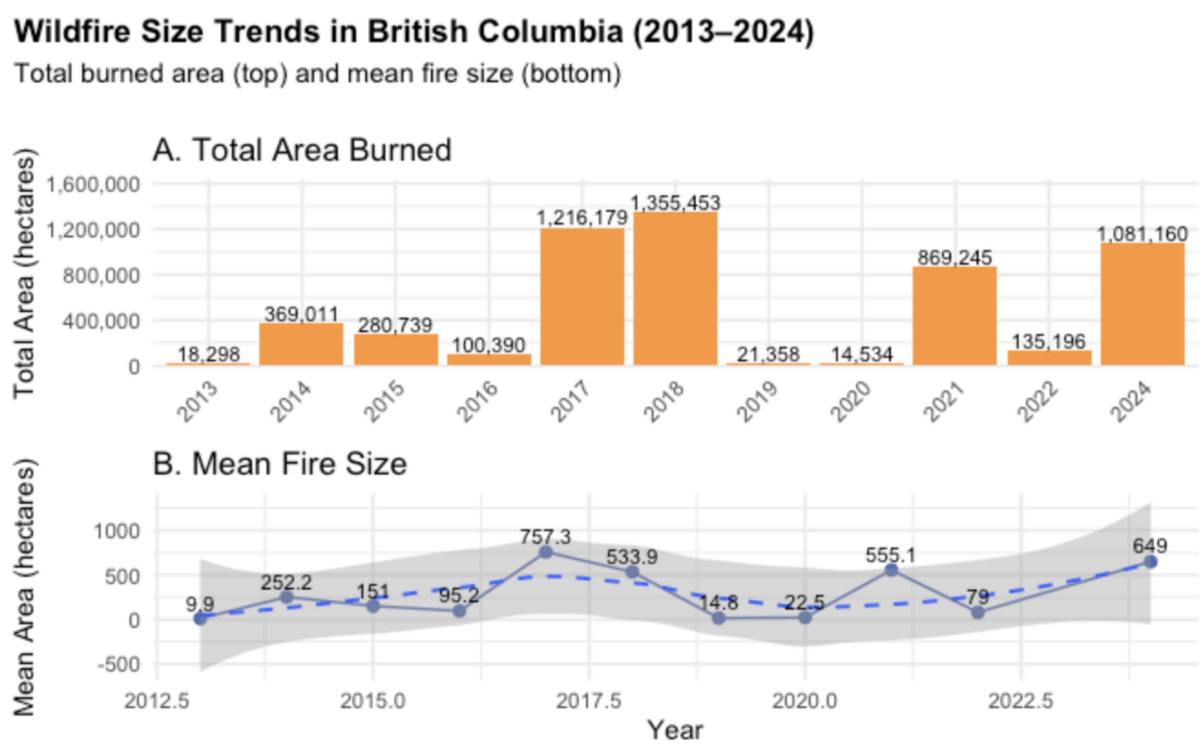
Total area burned and mean fire size

The **two-panel "Wildfire Size Trends"** figure tracks both the **total area burned per year** (top panel) and the **mean fire size** (bottom panel):

- The top panel shows that **a handful of years dominate total burned area**. In particular, 2017, 2018, 2021 and 2024 have extremely high total hectares burned, even when fire counts are not always the highest.

- The bottom panel shows mean fire size bouncing up in those same extreme years, then dropping back down in quieter seasons. The dashed smooth line highlights this pattern clearly.

Taken together, these plots indicate that **BC's burned area is driven by a few very large seasons**, not by a steady increase in "typical" fire size.

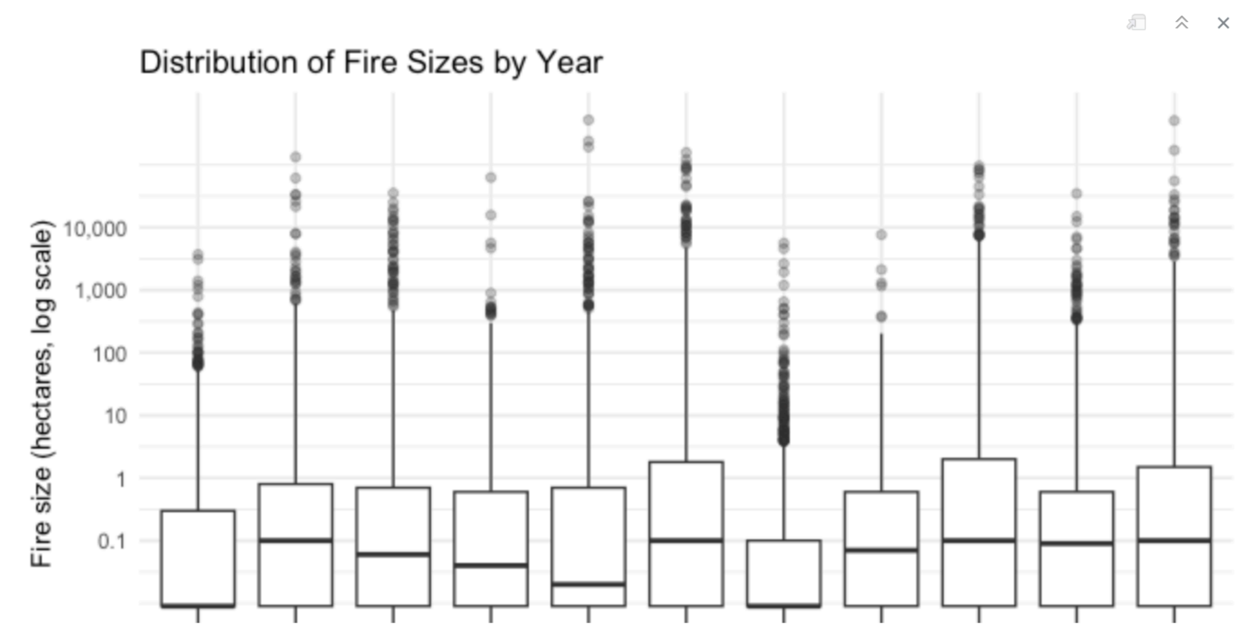


Distribution of fire sizes by year

The **boxplot of fire size by year on a log scale** gives more detail. Every year has a very low median (well under 1 ha), a long upper tail and many small outliers at the bottom:

- Most fires are **tiny**, burning less than a hectare.
- A relatively small number of fires extend far up the log scale, representing the **large and mega-fires** that drive total area burned.
- In the extreme seasons, the upper whiskers and outliers stretch higher, showing more very large fires rather than a systematic shift in the median.

Note: The log scale is crucial here: without it, the pattern would be dominated by a few monster fires.



Section 3. How are fires distributed across space?

Static period maps

The **Wildfire Locations in BC by Period** triple map shows point locations in three time windows (2013–2015, 2016–2019, 2020–2024). Each panel uses a different colour but shares the same BC bounding box.

Across all periods:

- Fires are **heavily concentrated in BC's interior**, especially the central plateau and southern interior.

- The **coastal strip** is relatively quiet, it could be due to its cooler, wetter climate.

- The overall spatial footprint is **stable**.

The main visual change over time is a **slight spread north and east** in the later period, hinting at increasing fire activity in more northern regions.

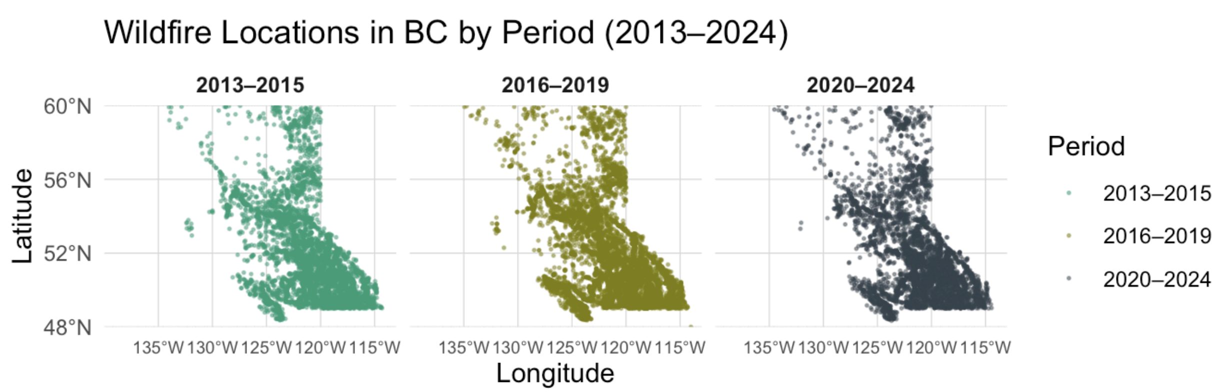
Interactive period map

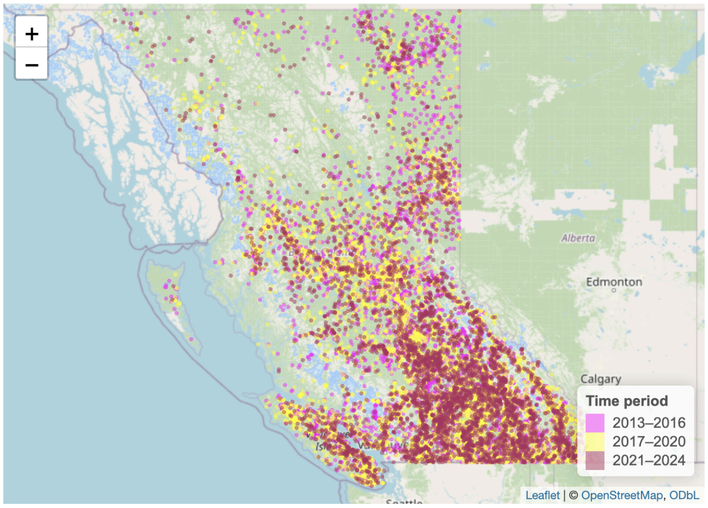
The first interactive Leaflet map (period-based colours) lets the reader:

- **Zoom in and pan** to explore individual clusters of fires.

- **Hover/click** on points to see the year, period, cause group, and size for each fire.

This interactive view reinforces the static map: regardless of period, most activity clusters in the dry interior, with more scattered points in northern and mountainous regions.





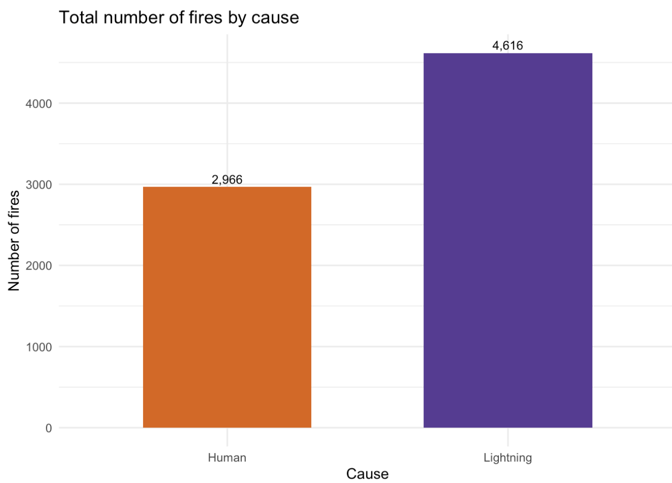
Section 4.How do human vs lightning causes relate to wildfire patterns?

4.1 Frequency by cause

The **bar chart "Total number of fires by cause (2013–2024, known causes only)"** compares the total number of fires with known ignition causes:

- Roughly **2,966 human-caused** fires vs **4,616 lightning-caused** fires.
- In other words, among fires where the cause is recorded, **lightning accounts for about 60% of ignitions** and humans for about 40%.

The takeaway is that, in this dataset, **lightning is the dominant known ignition source**, but human activity still contributes a substantial fraction.

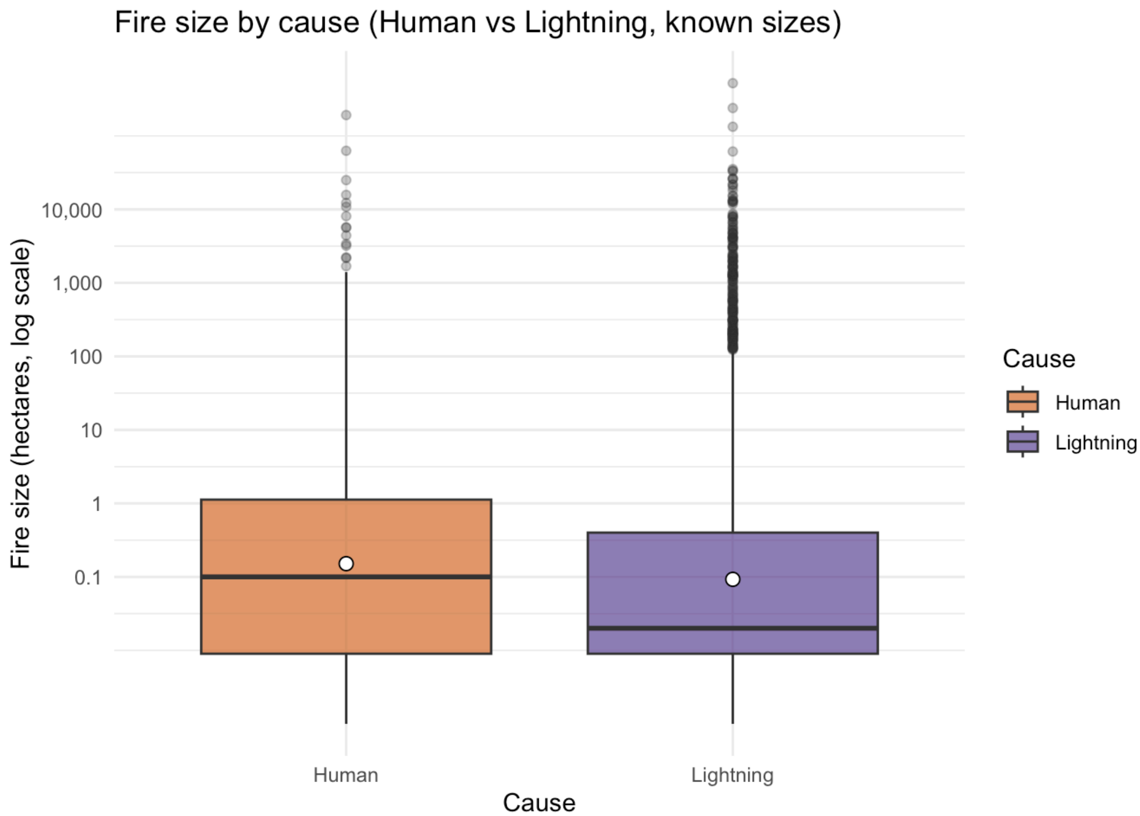


4.2 Fire size by cause

The **Fire size by cause** boxplot compares the full size distribution for human-caused and lightning-caused fires on a log scale:

For both causes, the median fire is very small (well under 1 hectare). The medians are quite similar, with lightning fires having a slightly lower median size than human fires.The real difference appears in the upper tail. Lightning-caused fires have more extreme outliers because alot of points seem to be clustered in the 1000-10000 region, the human caused fires also has outliers but it seems much less clustered.

This suggests that while most human and lightning fires are similarly small, lightning-caused fires are more likely to become very large once they escape initial stage or control.

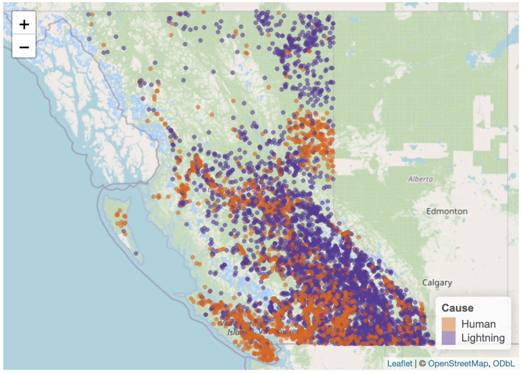


4.3 Spatial patterns of human vs lightning fires

The interactive **Human vs Lightning map** plots only fires with known causes and colours them by **Human** vs **Lightning**:

- Zooming in shows that **human-caused fires cluster more tightly around populated areas, major highways, and accessible valleys**, which is consistent with ignition sources like campfires, equipment, and debris burning.
- **Lightning fires are more evenly spread across the interior and into more remote regions**, often away from settlements.However that is not always the case.

Together, the frequency, size, and spatial patterns paint a consistent picture: **lightning causes more fires overall and is disproportionately responsible for the very largest fires, while human activity tends to create smaller but more localized ignition hot spots.**



Section 4.Limitations

Several limitations should be kept in mind when interpreting these results:

- **Missing attributes in recent years** – Some recent records (especially 2023) only have locations.
- **Large "Unknown/None" cause category** – Cause labels are often missing or vague, especially after 2017. Cause-based analyses therefore use only the subset of fires with a clear Human vs Lightning classification.
- **Point representation** – Each fire is represented by a single point, not a perimeter, so maps show **where fires started or were recorded**, not the full area burned.
- **Potential naming inconsistencies** – De-duplication is based on fire name and year, if different KMZ layers use slightly different names for the same fire, a few duplicates may remain.

Steps to Run:-

1. Open the file as a folder/project in r studio so that the data remains intact .
2. I have added alot of checks in the code to make sure the parsing , data reading etc is being done properly.