Data Science for Everyone - Visualizing the Solar Eclipse

Spring Discovery Day STEMonstration @Lyon College

Marcus Birkenkrahe

March 13, 2024

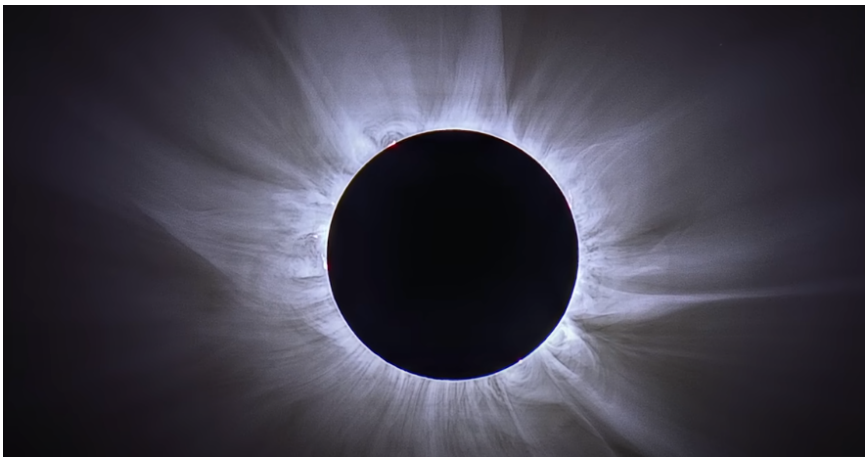
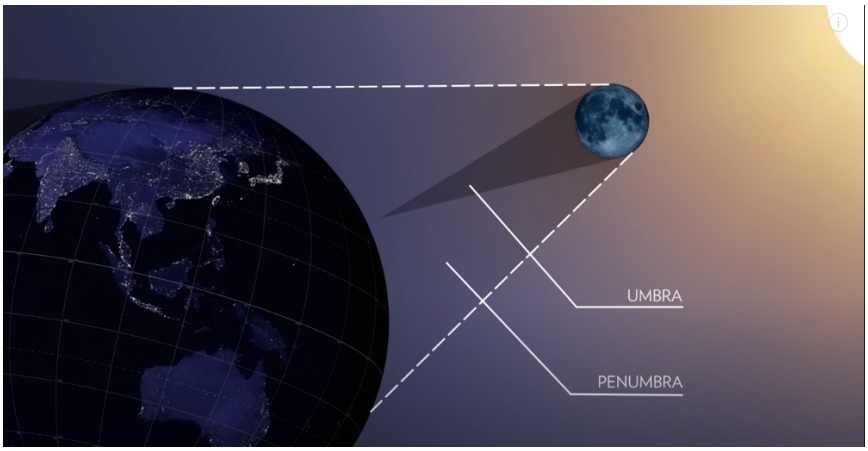


Figure 1: Image source: National Geographic (Solar Eclipse 101), 2017

How does a solar eclipse happen?



Watch: Solar Eclipse 101 | National Geographic (08/17/17)

A solar eclipse happens when the Moon passes between the Earth and the Sun, blocking all or part of the Sun's light from reaching the Earth.

This can only happen during a new moon, when Sun and Moon are in conjunction as seen from Earth.

How often does a solar eclipse happen?

Solar eclipses happen 2-5 times a year, total solar eclipses once every 18 months somewhere on Earth. Any given location on Earth might experience a solar eclipse only once every few hundred years.

Is predicting solar eclipses difficult?

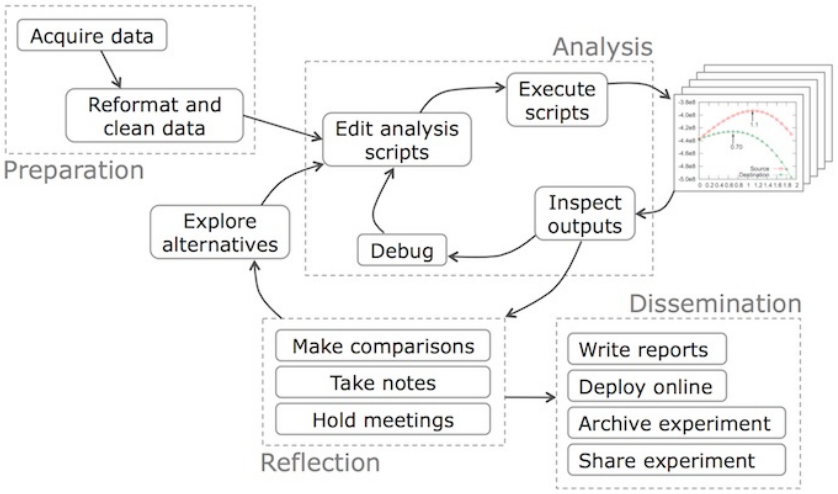
The classical basis of the prediction of the celestial positions of Sun, Moon and Earth is a three-body-problem. There is no general solution to this problem that applies to all possible initial conditions.

The two-body-problem was solved by Isaac Newton in the late 17th century. This led to the laws of motion and universal gravitation and laid the foundation for classical mechanics and enabled precise calculations of celestial motion, including the orbits of planets.

Why were the total solar eclipses 2017 and 2024 so close to one another?

The path of totality in 2017 moved from NW to SE, in 2024 from SW to NE. Timing between eclipses vary due to the complex geometry of Earth-Moon-Sun alignments and the cycle of eclipses ('Saros cycle').

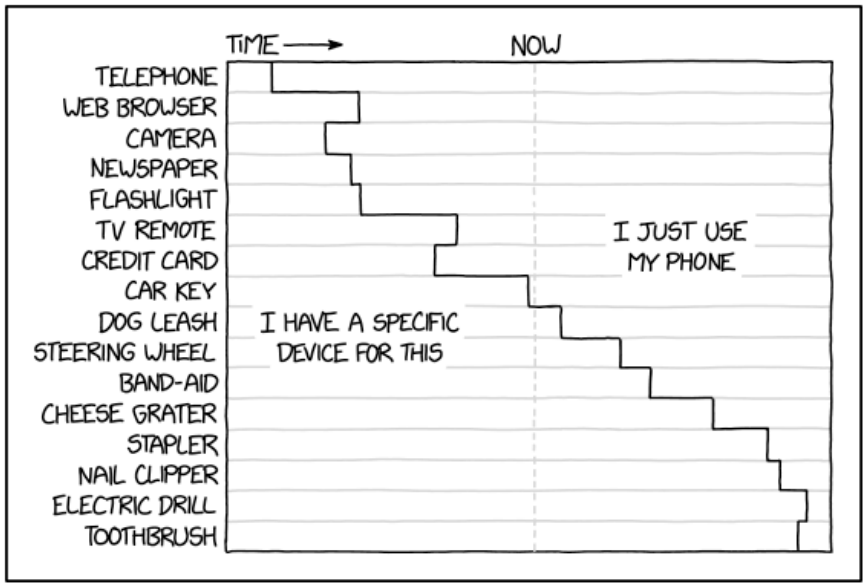
What does "data visualization" mean?



'Data visualization' is one step in the data science workflow, which includes: data gathering, data import, data cleaning, data transformation, and data modeling.

What is the difference between a "visualization" and "data visualization"?

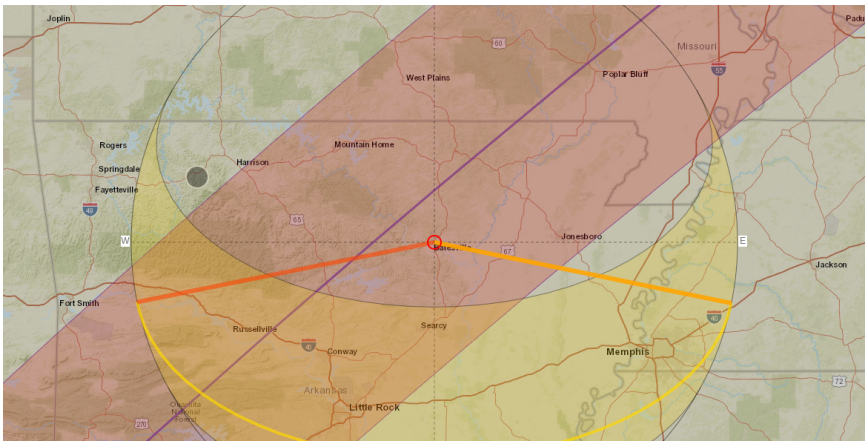
Data visualization is a scientific process: it involves using real historic (descriptive) or future (predicted) data, information about the origin of the data, a location where the data can be inspected,



This is "just" a visualization (though a fun one, and one that is rooted in common human experience - but the visualization makes no claim to be scientific or systematic).

- Science is a fickle mistress though. And when it comes to computer and data science, it's not very different from an honest craft. To fix a car (photo), you also need long training and apprenticeship.

What are interesting "solar eclipse" visualizations?



1. **The path of totality** - that is the path on Earth along which the Sun is totally blocked by the Moon. On April 8, 2024, Batesville AR lies in this path for a total time of 4 minutes 1.7 seconds. Examples:
   - NASA, Scientific Visualization Studio: Orthographic map showing many details; animated Gif and Saros cycle animation.
   - SunCalc.org: path computation with solar eclipse data for Batesville from 1500 B.C. to 3000 A.D.
   - Google maps: different lines of the path.
2. **Population impact**: how many people are in the path of totality. You need to add data sources to map approximate population density. See for example (Zeiler, 2024).
3. **Historical eclipse paths**: comparing the 2024 path with historical eclipse paths over the same region to explore frequency, duration and path. For example from eclipsophile.com.
4. **Interactive maps**: users can zoom in and out and explore specific locations. For example Google Maps.
5. **Astronomical phenomena**: visualizing the timing and positioning of other celestial bodies during the eclipse could add depth to the understanding of the event. E.g. starwalk.space, Astronomical Events 2024.
6. **Climatology and weather forecast**: weather planning for eclipse day, for example eclipsophile.com.

What is interesting the history of solar eclipse exploration?

1. **Ancient Observations:** Historical records from various civilizations, including the Babylonians, Greeks, and Chinese, provide evidence of solar eclipse observations, underscoring their importance in early astronomical studies. (E.g. as described by Herodotus during the Battle of Halys 585 BC when the sudden darkness was interpreted as a divine sign for peace).
2. **Scientific Milestones:** Solar eclipses have played pivotal roles in key scientific discoveries, including the validation of Einstein's theory of general relativity during the 1919 eclipse: Einstein had postulated that space was not the same in all directions but that gravity of large bodies could bend rays of light. Eddington measured the position of stars near the Sun's edge during an eclipse providing empirical evidence for the theory.
3. **Technological Advancements:** The study of solar eclipses has driven advancements in astronomical instruments and observational techniques, enhancing our understanding of the Sun and its influence on Earth. Example: the Antikythera from 100 BC (named after the Greek island where it was found in 1901), the earliest known analog computer designed to predict eclipses decades in advance.

4. **Cultural Impact:** Eclipses have significantly impacted human culture, inspiring myths, influencing religions, and contributing to our fascination with the cosmos. E.g. in Viking mythology, eclipses were explained as the sky wolf, Skoll, catching and devouring the Sun.

What do you need to have, know or learn to visualize the solar eclipse?

Data

Universal Time	Northern Limit Latitude	Northern Limit Longitude	Southern Limit Latitude	Southern Limit Longitude	Central Line Latitude	Central Line Longitude	M:S Diam. Ratio	Sun Alt	Sun Azm	Path Width km	Central Line Durat.
	° ′	° ′	° ′	° ′	° ′	° ′		°	°		
Limits	07 11.6S	158 43.9W	08 27.2S	158 20.1W	07 49.5S	158 31.9W	1.040	0	-	144	02m06.3s
16:40	-	-	07 36.2S	152 54.5W	07 38.1S	157 11.2W	1.040	1	82	146	02m08.8s
16:42	05 30.6S	149 47.6W	06 11.7S	146 38.0W	05 50.2S	148 07.8W	1.043	11	81	159	02m27.5s
16:44	04 20.5S	145 29.6W	05 08.4S	143 00.6W	04 44.0S	144 13.0W	1.044	16	81	166	02m36.8s
16:46	03 21.2S	142 27.6W	04 12.3S	140 15.6W	03 46.4S	141 20.3W	1.045	19	81	171	02m44.2s
16:48	02 27.1S	140 01.8W	03 20.2S	137 59.5W	02 53.3S	138 59.7W	1.046	22	81	174	02m50.6s
16:50	01 36.2S	137 58.5W	02 30.8S	136 02.5W	02 03.3S	136 59.7W	1.047	25	81	178	02m56.3s
16:52	00 47.7S	136 10.6W	01 43.4S	134 19.0W	01 15.4S	135 14.1W	1.048	27	81	181	03m01.6s
16:54	00 01.0S	134 34.2W	00 57.6S	132 45.9W	00 29.1S	133 39.5W	1.048	29	81	183	03m06.4s
16:56	00 44.4N	133 06.9W	00 13.0S	131 21.1W	00 15.9N	132 13.5W	1.049	31	81	186	03m10.9s
16:58	01 28.6N	131 46.8W	00 30.6N	130 03.0W	00 59.7N	130 54.5W	1.050	33	82	188	03m15.2s
17:00	02 11.8N	130 32.7W	01 13.2N	128 50.5W	01 42.7N	129 41.2W	1.050	35	82	190	03m19.3s
17:02	02 54.2N	129 23.6W	01 55.1N	127 42.8W	02 24.8N	128 32.8W	1.050	37	82	192	03m23.1s
17:04	03 35.9N	128 18.8W	02 36.4N	126 39.1W	03 06.3N	127 28.6W	1.051	38	83	193	03m26.8s
17:06	04 17.0N	127 17.7W	03 17.0N	125 39.0W	03 47.2N	126 28.0W	1.051	40	83	194	03m30.3s
17:08	04 57.5N	126 19.9W	03 57.2N	124 42.0W	04 27.5N	125 30.6W	1.052	41	84	196	03m33.7s
17:10	05 37.5N	125 24.9W	04 36.8N	123 47.8W	05 07.3N	124 36.1W	1.052	43	84	197	03m36.9s
17:12	06 17.1N	124 32.5W	05 16.0N	122 56.0W	05 46.7N	123 44.0W	1.052	44	85	198	03m40.0s
17:14	06 56.3N	123 42.4W	05 54.8N	122 06.4W	06 25.6N	122 54.1W	1.053	46	86	199	03m42.9s
17:16	07 35.0N	122 54.2W	06 33.3N	121 18.7W	07 04.3N	122 06.2W	1.053	47	86	199	03m45.8s
17:18	08 13.5N	122 07.9W	07 11.4N	120 32.8W	07 42.5N	121 20.1W	1.053	48	87	200	03m48.5s
17:20	08 51.6N	121 23.2W	07 49.2N	119 48.5W	08 20.5N	120 35.6W	1.053	49	88	201	03m51.1s

Figure 2: Solar and Lunar Eclipses (Source: Arvidsson, 2021)

- All data visualizations start with data. You can get the date, time, and location of every solar eclipses of the past 5,000 years from NASA's Goddard Space Flight Center as a CSV file (Arvidsson, 2023).

- You can also get the path data for the total solar eclipse of 2024 on April 8 from NASA (Espenak, 2014). You have to 'scrape' these data from the web page (which can be tricky).

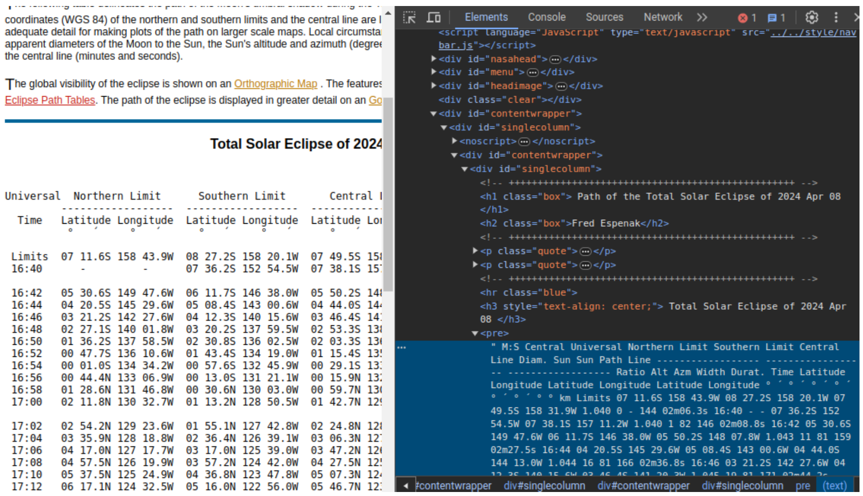


Figure 3: Total Solar Eclipse of 2024 Apr 08 (Source: Espenak, 2014)

Tools

Such as: programming languages like R or Python, data visualization software like Tableau, or symbolic languages like Wolfram Language.

Let's do some actual coding with R:

1. Importing the data into two data frames `Solar` and `Lunar`.
2. Analyzing the data - looking for structure and statistics.
3. Plotting the data.

Importing and transforming the data

After importing, we change some column names to ease analysis:

```
## Store downloaded CSV data in dataframes
solar <- read.csv("data/solar.csv", header = TRUE,stringsAsFactor=TRUE)
lunar <- read.csv("data/lunar.csv", header = TRUE,stringsAsFactor=TRUE)

## correct header names for better display
selection <- c(2,3,7,12,13,16)
names(lunar)[selection] <- c('date','time','type','lat','lon','tot')
head(lunar)[selection]
```

```
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'data/solar.csv': No such file or directory
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'data/lunar.csv': No such file or directory
Error in names(lunar)[selection] <- c("date", "time", "type", "lat", "lon",  :
  object 'lunar' not found
Error in head(lunar) : object 'lunar' not found
```

Analyzing the data - structure and statistics

- Getting a structural overview of the dataframe:

```
str(lunar)
```

```
Error in str(lunar) : object 'lunar' not found
```

```
'data.frame':   12064 obs. of  16 variables:
 $ Catalog.Number                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ date                          : Factor w/ 12064 levels "-1 January 20",..: 2697 2698
 $ time                          : Factor w/ 11198 levels "00:00:02","00:00:10",..: 660
 $ Delta.T..s.                   : int  46437 46427 46416 46404 46392 46380 46368 46357
 $ Lunation.Number               : int  -49456 -49451 -49445 -49439 -49433 -49427 -4942
 $ Saros.Number                  : int  17 -16 -11 -6 -1 4 9 14 -19 19 ...
 $ type                          : Factor w/ 8 levels "N","Nb","Ne",..: 1 1 5 5 8 8 5 5
 $ Quincena.Solar.Eclipse        : Factor w/ 11 levels "-a","-h","-p",..: 10 1 4 1 8 8
 $ Gamma                         : num  -1.098 -1.115 0.899 -0.464 0.1 ...
 $ Penumbral.Magnitude           : num  0.879 0.814 1.21 2.038 2.651 ...
 $ Umbral.Magnitude              : num  -0.192 -0.192 0.207 0.974 1.696 ...
 $ lat                           : Factor w/ 52 levels "0N","0S","10N",..: 34 13 10 7 6
 $ lon                           : Factor w/ 362 levels "0E","0W","100E",..: 192 358 33
 $ Penumbral.Eclipse.Duration..m.: num  269 233 282 343 323 ...
 $ Partial.Eclipse.Duration..m.  : Factor w/ 1808 levels "-","10.6","100.1",..: 1 1 24
 $ totality [s]                  : Factor w/ 809 levels "-","1.7","100",..: 1 1 1 1 792
```

- Getting a statistical overview of relevant features:

```
summary(lunar)
```

```
     Error in summary(lunar) : object 'lunar' not found
```

```
Catalog.Number                date            time          Delta.T..s.
Min.   :    1   -1 January 20  :    1   01:05:56:    3   Min.   :   -6
1st Qu.: 3017   -1 July 17     :    1   01:42:04:    3   1st Qu.:  962
Median : 6032   -10 December 20:    1   02:03:46:    3   Median : 5597
Mean   : 6032   -10 January 29 :    1   05:12:17:    3   Mean   :12116
3rd Qu.: 9048   -10 July 26    :    1   06:18:50:    3   3rd Qu.:20902
Max.   :12064   -100 June 1    :    1   06:34:23:    3   Max.   :46437
                (Other)        :12058   (Other) :12046
Lunation.Number   Saros.Number          type      Quincena.Solar.Eclipse
Min.   :-49456   Min.   :-20.00   P      :4207   a-     :2477
1st Qu.:-33923   1st Qu.: 40.00   N      :4020   -a     :2471
Median :-18446   Median : 80.00   T      :1405   t-     :1788
Mean   :-18531   Mean   : 80.51   T+     :1042   -t     :1787
3rd Qu.: -3068   3rd Qu.:121.00   T-     :1032   pp     :1347
Max.   : 12378   Max.   :183.00   Nx     : 141   p-     : 749
                                  (Other): 217   (Other):1445
    Gamma          Penumbral.Magnitude Umbral.Magnitude       lat
Min.   :-1.58270   Min.   :0.0004      Min.   :-1.0958   23S    : 544
1st Qu.:-0.78882   1st Qu.:0.6844      1st Qu.:-0.3340   22S    : 533
Median : 0.00175   Median :1.4175      Median : 0.4004   23N    : 514
Mean   : 0.00249   Mean   :1.4187      Mean   : 0.4002   22N    : 511
3rd Qu.: 0.79173   3rd Qu.:2.1369      3rd Qu.: 1.1179   24S    : 394
Max.   : 1.57910   Max.   :2.9089      Max.   : 1.8821   21S    : 378
                                                         (Other):9190
     lon        Penumbral.Eclipse.Duration..m. Partial.Eclipse.Duration..m.
87E    :   53   Min.   :  5.2                  -      :4378
64E    :   50   1st Qu.:223.1                  211.6  :  24
64W    :   49   Median :295.0                  213.2  :  24
129W   :   48   Mean   :270.0                  210.5  :  21
99W    :   48   3rd Qu.:327.8                  210.9  :  21
107W   :   46   Max.   :379.5                  211.2  :  21
(Other):11770                                  (Other):7575
 totality [s]
-      :8585
98.6   :  28
98.8   :  23
96     :  22
98.1   :  22
```

```
98.4   :  21
(Other):3363
```

- How many total solar eclipses were recorded in `lunar.csv`, what was the longest and what was the shortest total solar eclipse? What is the first and the last recorded one?

```
## Filter for total solar eclipses, converting factors to characters as necessary
total_eclipses <- subset(lunar, grepl("T", as.character(type)))

## Count the total number of total solar eclipses
total_count <- nrow(total_eclipses)

## Convert duration to numeric while handling potential NA values
## Assuming 'tot' was imported as a factor because of stringsAsFactors=TRUE
total_eclipses$tot <- as.numeric(as.character(total_eclipses$tot))

## Find the longest and shortest total solar eclipse durations
longest_eclipse_duration <- max(total_eclipses$tot, na.rm = TRUE)
shortest_eclipse_duration <- min(total_eclipses$tot, na.rm = TRUE)

## Sort the total eclipses by date, converting factors to characters if necessary
total_eclipses_sorted <- total_eclipses[order(as.character(total_eclipses$date)),

## Get the first and last recorded total solar eclipses
first_recorded_eclipse <- total_eclipses_sorted[1, ]
last_recorded_eclipse <- total_eclipses_sorted[nrow(total_eclipses_sorted), ]

## Print the results
cat("Total Solar Eclipses:", total_count, "\n")
cat("Longest Eclipse Duration (minutes):", longest_eclipse_duration, "\n")
cat("Shortest Eclipse Duration (minutes):", shortest_eclipse_duration, "\n")

Error in subset(lunar, grepl("T", as.character(type))) :
  object 'lunar' not found
Error in nrow(total_eclipses) : object 'total_eclipses' not found
Error: object 'total_eclipses' not found
Error: object 'total_eclipses' not found
Error: object 'total_eclipses' not found
Error: object 'total_eclipses' not found
```

```
Error: object 'total_eclipses_sorted' not found
Error: object 'total_eclipses_sorted' not found
Error in cat("Total Solar Eclipses:", total_count, "\n") :
  object 'total_count' not found
Error in cat("Longest Eclipse Duration (minutes):", longest_eclipse_duration,  :
  object 'longest_eclipse_duration' not found
Error in cat("Shortest Eclipse Duration (minutes):", shortest_eclipse_duration,  :
  object 'shortest_eclipse_duration' not found
```

- According to the datasheet from Kaggle, Earth will experience 12064 lunar and 11898 solar eclipses. Data exploration must continue with an explanation of this discrepancy!

- Is the 2024, April 8 total solar eclipse contained in this dataset?

```
## Find the 2024 April 8 eclipse
april_8_2024_eclipse <- subset(lunar, as.character(date) == "2024-04-08")

## Check if the eclipse is in the data
if(nrow(april_8_2024_eclipse) > 0) {
  cat("Data for the 2024 April 8 eclipse:\n")
  print(april_8_2024_eclipse)
} else {
  cat("The 2024 April 8 eclipse is not in the dataset.\n")
}
```

```
Error in subset(lunar, as.character(date) == "2024-04-08") :
  object 'lunar' not found
Error in nrow(april_8_2024_eclipse) :
  object 'april_8_2024_eclipse' not found
```

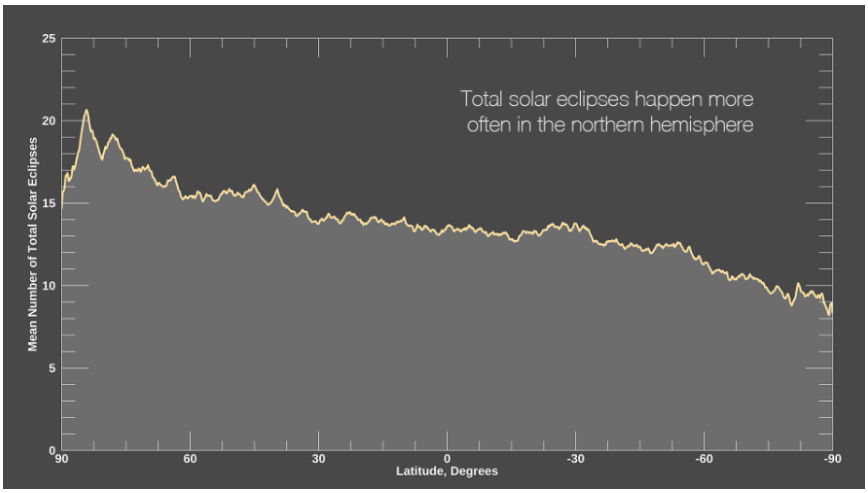
- This could be because we were looking at the lunar and not at the solar eclipse dataset, or because we got the date format wrong (this is the case: the dataframe date format is not "YYYY-MM-DD").

- We're out of time for now, but there's still work to be done before we can begin to think about plotting.

Plotting the data

An interesting plot would be to see if total solar eclipses happen more often in the Northern or in the Southern hemisphere.

Here is a plot from NASA. This is easy to do in R or Python as well.



Relevance

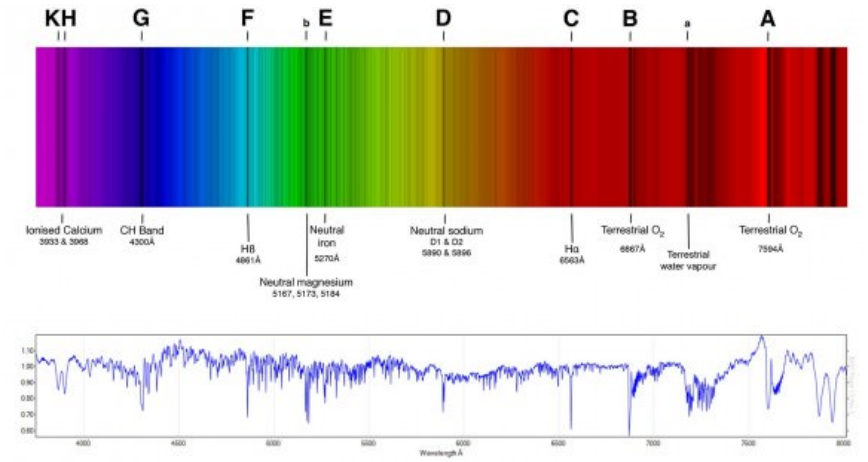


Figure 4: Fraunhofer lines (Credit: eventbrite.com)

Understanding of what you want to show and whom to show it to: a clear objective and a specific audience.

- Data never exist out of context. To invest time into gathering, importing, transforming, analysing and visualizing data, we must first convince ourselves of the relevance of our research question.
- There is no "data science for its own sake", though there can be surprise discoveries in the data (i.e. answers to questions not asked, patterns not suspected before, etc.).
- An example is the observation of helium in the Sun's atmosphere during the solar eclipse of August 18, 1868. Astronomers observed a yellow spectral line in the light from the Sun's chromosphere during the eclipse. This observation could not be explained by any known chemical elements at that time. Turns out it was "helium" (after the Greek god of the Sun, Helios), which was only found on Earth 27 years after its initial discovery in the solar spectrum.
- The image shows Fraunhofer lines - dark absorption lines that correspond to different chemical elements.

What can you study at Lyon to learn more about this?

- At Lyon, you can learn all about data in courses on:
  1. Introductory and advanced data science with R and Python
  2. Data visualization (to visualize data in maps or graphs)
  3. Machine learning (to predict events from data)
  4. Databases (to store large amounts of data)
  5. Algorithms (to search through large data sets)
  6. Geographical Information Systems (GIS)
  7. Data modeling (to derive statistical insights from data)

How can you find out more about us?

- Visit us on campus, come talk to me and audit any class!
- Participate in our summer programs (2024: creating games in JavaScript, HTML and CSS; 45 programming languages in 45 minutes).
- Follow us on X.com (@LyonCollege, @birkenkrahe) or on Youtube: @CareerPathwaysPodcast