# iNaturalist Observations Analysis

TCHILINGUIRIAN Théo

2024-06-20

## Goal of this Report

This report is an analysis of iNaturalist observations made by one individual. The objective is to uncover patterns and biases in the data that can inform future research and observation strategies. Observations are unevenly distributed depending on the month and the hour of the day, impacting the likelihood of observing certain species. This analysis aims to provide practical insights into observation methods and potential biases.

# Questions Addressed

- What is the distribution of observations over time?
- Which species are most frequently observed?
- How do observations vary by taxonomic group?

## Dataset Overview

The dataset consists of iNaturalist observations by one individual. It includes information such as the species observed, the date and time of the observation, and the taxonomic classification. The dataset used can be swapped with a collection of observations made during a mission, or by a group of individuals, to focus the analysis on the event in question.

## General Information about the Dataset

This dataset contains 9240 observations on 3627 species.

### Percentage of Missing Data per Observation Variable

| Column/Observation Variable | Percentage of Missing Data |
|---|---:|
| id | 0.00 |
| observed_on_string | 0.00 |
| observed_on | 0.00 |
| time_observed_at | 0.00 |
| time_zone | 0.00 |
| user_id | 0.00 |
| user_login | 0.00 |
| user_name | 0.00 |
| created_at | 0.00 |
| updated_at | 0.00 |
| quality_grade | 0.00 |
| license | 0.00 |
| url | 0.00 |
| image_url | 0.00 |
| sound_url | 100.00 |
| tag_list | 0.00 |
| description | 0.00 |
| num_identification_agreements | 0.00 |
| num_identification_disagreements | 0.00 |
| captive_cultivated | 0.00 |
| oauth_application_id | 97.41 |
| place_guess | 0.00 |
| latitude | 0.01 |
| longitude | 0.01 |
| positional_accuracy | 51.71 |
| private_place_guess | 0.00 |
| private_latitude | 99.66 |
| private_longitude | 99.66 |
| public_positional_accuracy | 51.52 |
| geoprivacy | 0.00 |

| Column/Observation Variable | Percentage of Missing Data |
|---|---:|
| taxon_geoprivacy | 0.00 |
| coordinates_obscured | 0.00 |
| positioning_method | 0.00 |
| positioning_device | 0.00 |
| species_guess | 0.00 |
| scientific_name | 0.00 |
| common_name | 0.00 |
| iconic_taxon_name | 0.00 |
| taxon_id | 0.00 |

## Question 1: Distribution of Observations Over Time

### Statistical Description

- Population: All possible iNaturalist observations.
- Sample: The dataset of observations by one individual.
- Variables:
  - Observed Hour: Categorical variable with 24 modalities (0 to 23 hours).
  - Observed Month: Categorical variable with 12 modalities (January to December).
- Objective: To analyze the distribution of observations over different times of the day and months of the year.

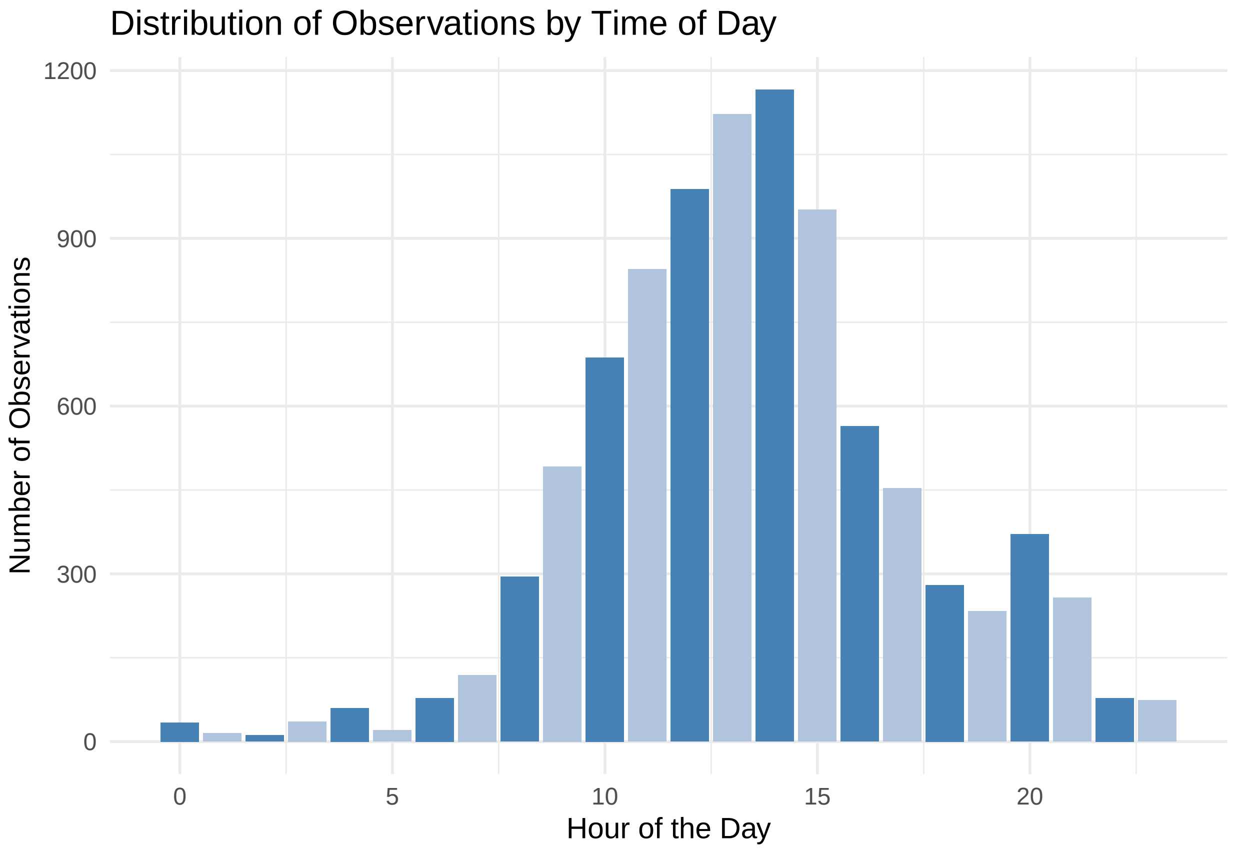
### Analysis and Visualization

To understand the temporal distribution of observations, we analyze the data by hour of the day (time of day) and by month.

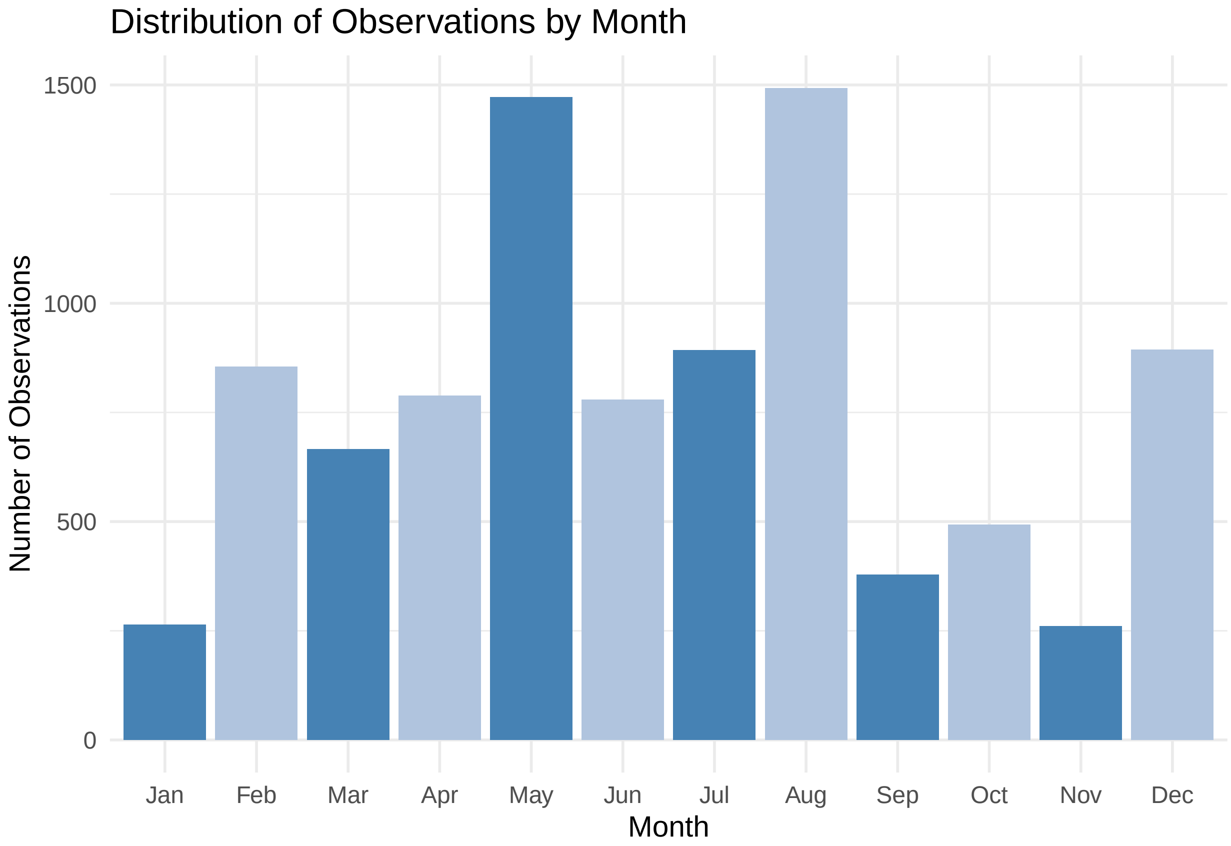
### Proportion of Observations without Temporal Data

Calculations return that approximately 0.12% of data concerning the hour of the observation is not given.

### Distribution of Observations by Time of Day



### Distribution of Observations by Month



## Interpretation

The data shows a clear pattern in the distribution of observations by time of day and by month. Most observations are made around high noon, indicating a potential bias against nocturnal species. Additionally, certain months have higher observation frequencies, which could be due to seasonal availability or observer activity patterns.

### Limitations

The missing data for the hour of observation (approximately 0.12%) limits the completeness of the temporal analysis, however by a negligible margin. Furthermore, the dataset represents observations by only one individual, which may not generalize to a broader population.

## Question 2: Most Frequently Observed Species

### Statistical Description

- Population: All possible species that have been observed on iNaturalist.
- Sample: The species recorded in the dataset.
- Variables:
  - Species Name: Categorical variable representing the known scientific names of species observed.
  - Observation Count: Numerical variable representing the number of observations for each species.
- Objective: To identify the species that are most frequently observed in the dataset.

### Analysis and Visualization

To identify the most frequently observed species, we group the data by species and count the number of observations for each.

| Species | Number of Observations |
|---|---:|
| Linyphiidae | 145 |
| Euscorpius flavicaudis | 54 |
| Theridiidae | 51 |
| Pisaura | 50 |
| Pardosa | 48 |
| Xysticus | 43 |
| Buthus occitanus | 42 |
| Gnaphosidae | 42 |
| Heliophanus | 40 |
| Tetragnatha | 36 |
| Zelotinae | 36 |
| Nomisia | 35 |
| Synema globosum | 35 |
| Philodromus | 34 |
| Hogna radiata | 32 |
| Bdellidae | 31 |
| Scytodes thoracica | 31 |
| Enoplognatha | 30 |
| Agyneta | 27 |
| Alopecosa albofasciata | 27 |

### Interpretation

The table above lists the top 20 most observed species in the dataset. These species are observed more frequently, indicating either their abundance in the observed area or the observer's preference or ease in identifying them in their environment.

Representing this data as a table allows for an easy overview of the quantity of observations in the most frequently observed species.

### Limitations

This analysis does not account for the observer's potential biases towards certain species, nor does it consider the possibility of misidentifications.

## Question 3: Observations by Taxonomic Group

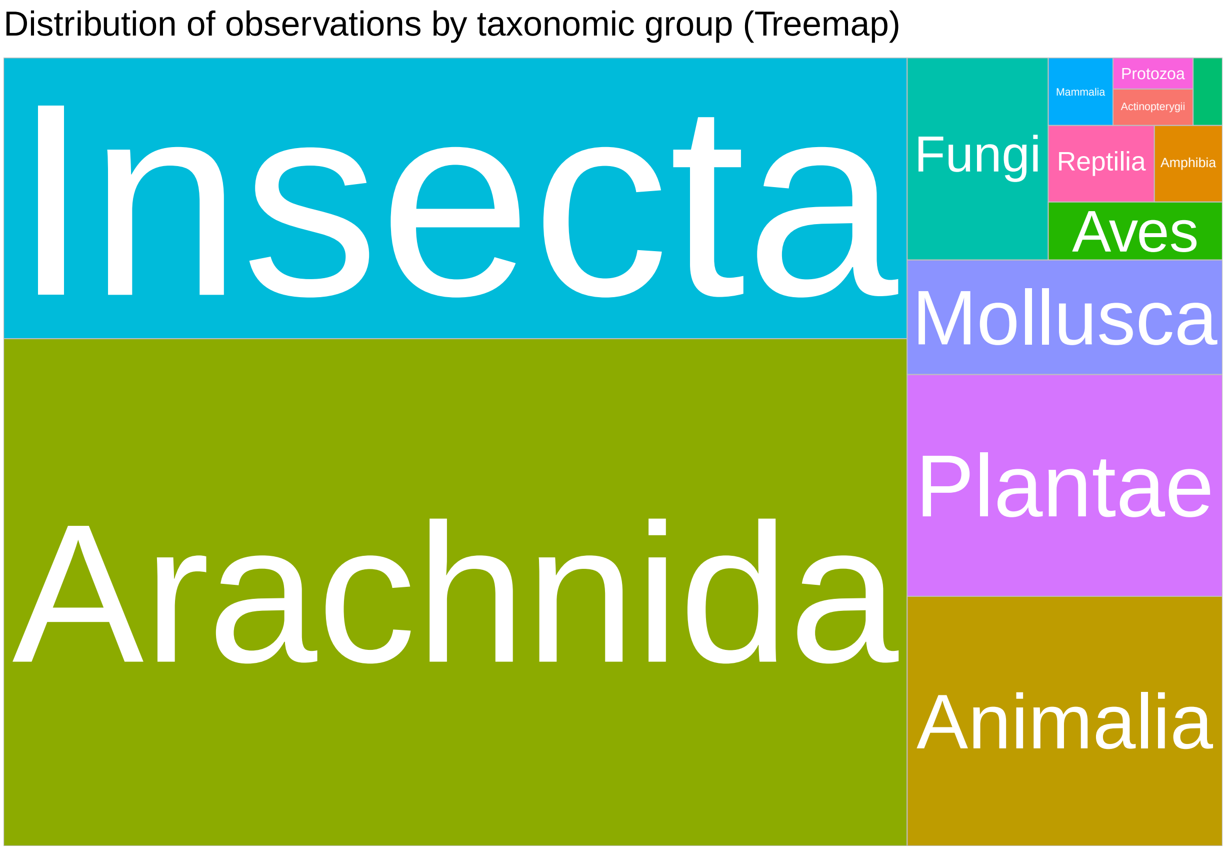
### Statistical Description

- Population: All taxonomic groups that the observed species belong to.
- Sample: The taxonomic groups recorded in the dataset.
- Variables:
  - Taxon Name: Categorical variable representing the taxonomic group of the species.
  - Observation Count: Numerical variable representing the number of observations for each taxonomic group.
- Objective: To understand the distribution of observations across different taxonomic groups.

### Analysis and Visualization

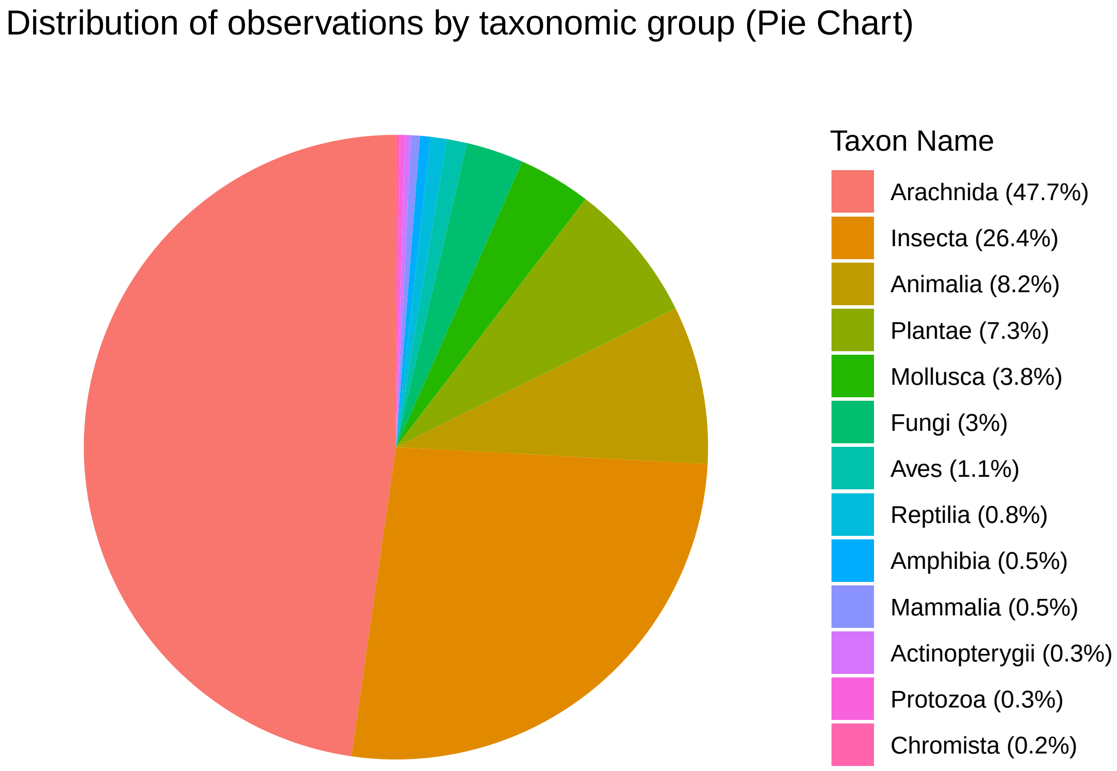
To understand the distribution of observations across different taxonomic groups, we use both a treemap and a pie chart for visualization.

### Distribution by Taxonomic Group (Treemap)



The idea to make a treemap diagram came to me while browsing existing R Shiny applications.

### Distribution by Taxonomic Group (Pie Chart)



### Interpretation

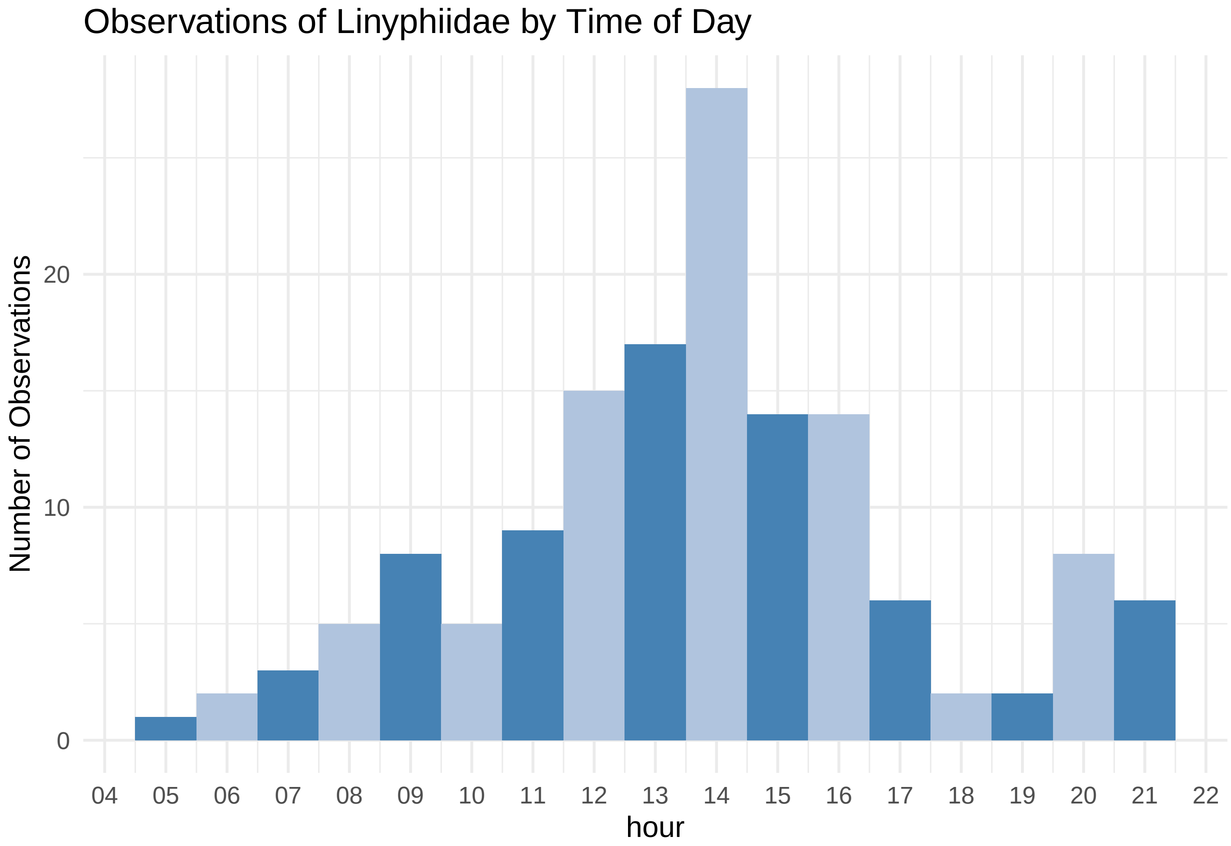
Both visualizations highlight the distribution of observations across various taxonomic groups. The treemap provides a quick visual representation of the relative abundance of each group, while the pie chart gives a more precise percentage breakdown alongside the visuals sorted by proportion.

### Limitations

Taxonomic identification is subject to the observer's expertise and potential errors. Taxons exist in different levels of precision, so these representations can only serve as imprecise overviews, unless the taxons are precisely identified before the dataset analysis.

## Additional Analysis: Observations of Linyphiidae by Time of Day

Linyphiidae, a family of spiders, is the most observed family of spiders (Arachnida) in the dataset. Analyzing their observations by time of day provides further insights into observation patterns.



Number of observations for Linyphiidae : 145.

### Interpretation

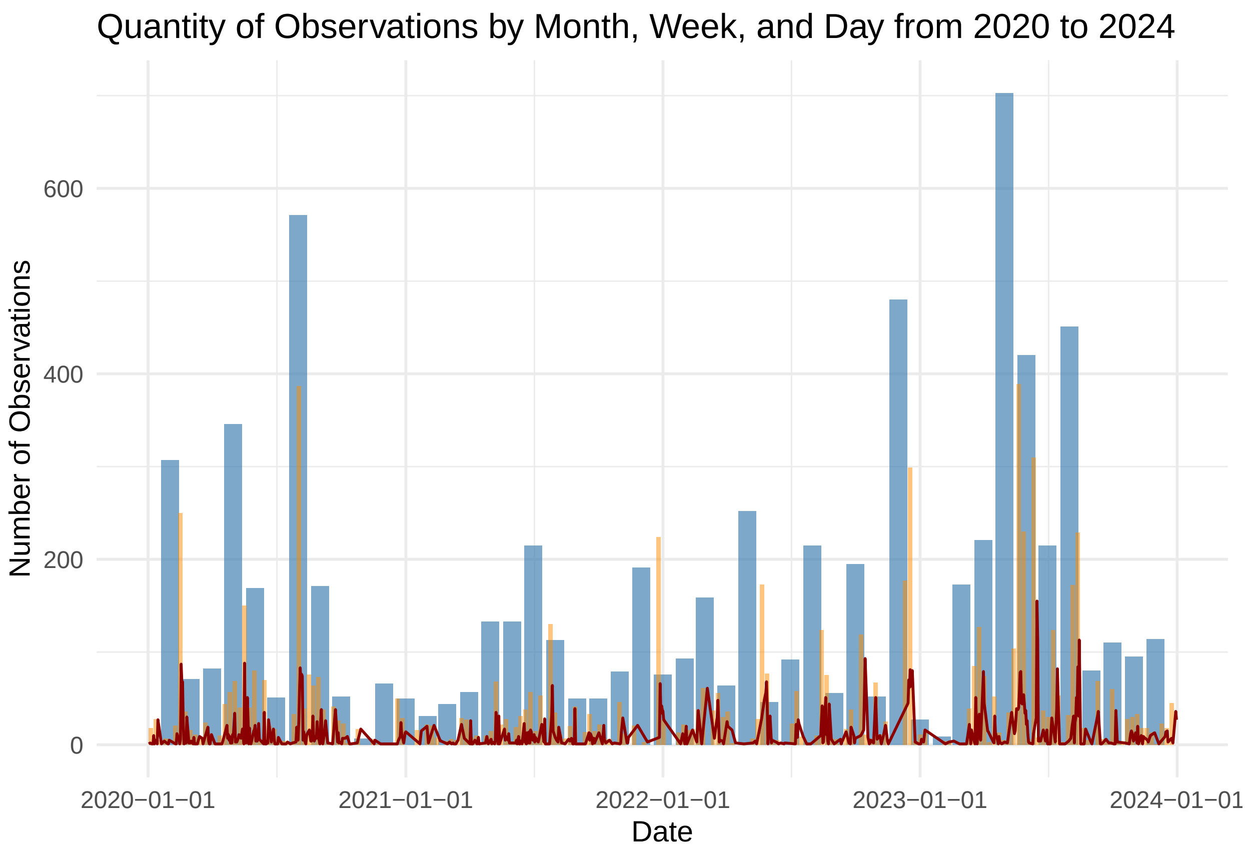
Linyphiidae is a family of spiders (order Aranea, class Arachnida). It is the most present family of spiders in the dataset.

The histogram indicates the time of day when Linyphiidae observations are most frequent. This pattern can inform the best times for observing these spiders.

This analysis can be considered an answer to the three last questions in one, focusing on one particular family of spides of the dataset.

## Additional Analysis: Quantity of Observations by Month, Week, and Day between 2020 and 2024

This diagram is experimental and should not be used for research purposes.



**Interpretation**

This presentation of the distribution of observations by month, week and day shows a handsome overview of the work done over years.

It may prove useful to compare the work between years, while still retaining useful levels of granularity.

**Limitations**

I have had issues with this graph, either from the data flooring, or from the presentation in columns, but some data does not seem to fit in the correct bars.

Due to these complications, this diagram is experimental and should not be used for research purposes.

## Conclusion

This report provides a detailed analysis of iNaturalist observations, focusing on temporal distribution, species frequency, and taxonomic group distribution. The findings highlight potential biases and patterns that can guide future research, observation and identification efforts.

### Difficulties

I encountered a few challenges while studying this dataset. I studied possible correlations between the variables in the dataset ; I either found completely uncorrelated data (null coefficient) or non-linear correlations that I didn't have the skills to study.

Moreover, efforts to make a map of the distribution of observations using the dataset coordinates (longitude, latitude, and positional accuracy) were halted by Google Maps API needing a non-free API key ; and the other possible mapping services not being available at the time of writing due to maintenance or inoperability.

### Future Works

- Future analyses to incorporate data from multiple observers to enhance generalizability.
- Efforts to reduce missing data.
- Further studies could explore additional factors influencing observation patterns, such as weather conditions and habitat types or surrounding environment.

### Interesting links and sources

The following is a collection of links to websites or databases I discovered or was introduced to, that inspired the subject for this project.

- iNaturalist: https://www.inaturalist.org/
- GBIF: https://www.gbif.org/
- Titan Database: http://titan.gbif.fr/
- World Arachnida Catalog (WAC): https://wac.nmbe.ch/
- World Spider Catalog (WSC): https://wsc.nmbe.ch/
- Aranea (spider identification): https://araneae.nmbe.ch/
- The Scorpion Files (scorpion identification): https://www.ntnu.no/ub/scorpion-files/