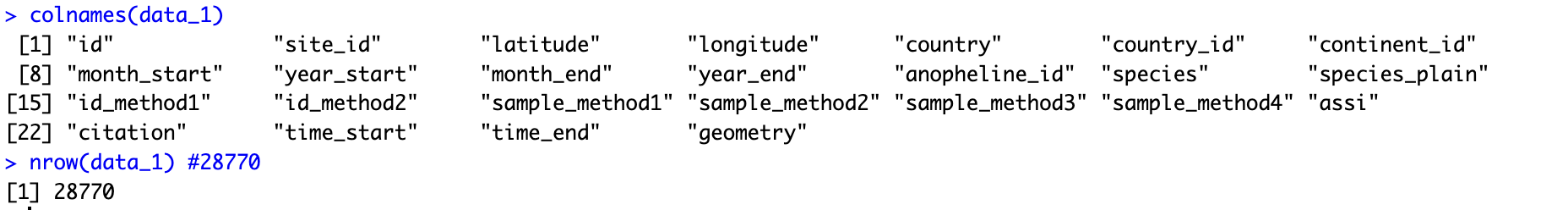
Data Structure: ETL; standardisation; cleaning; explain some species does not exist.

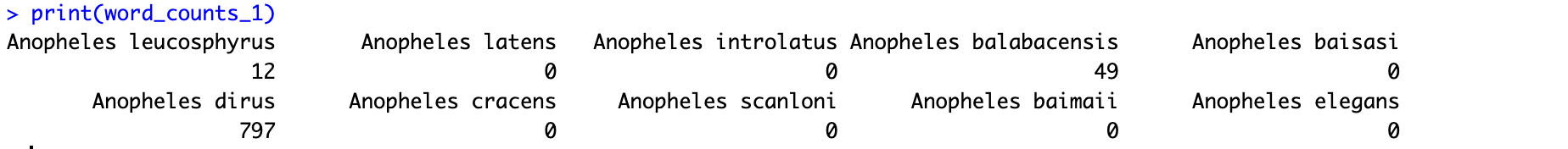
Talk the number.

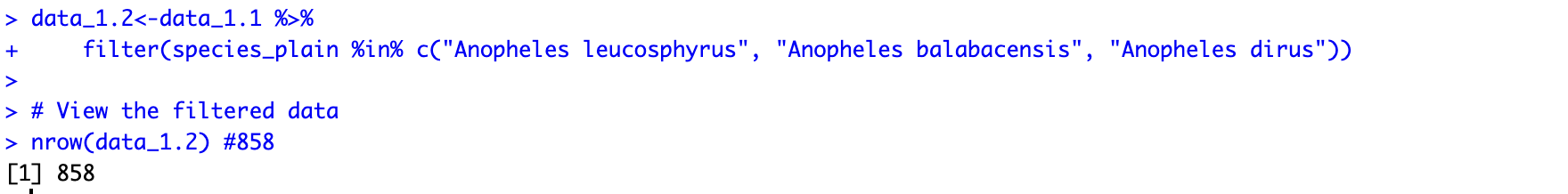
Talk the graph.

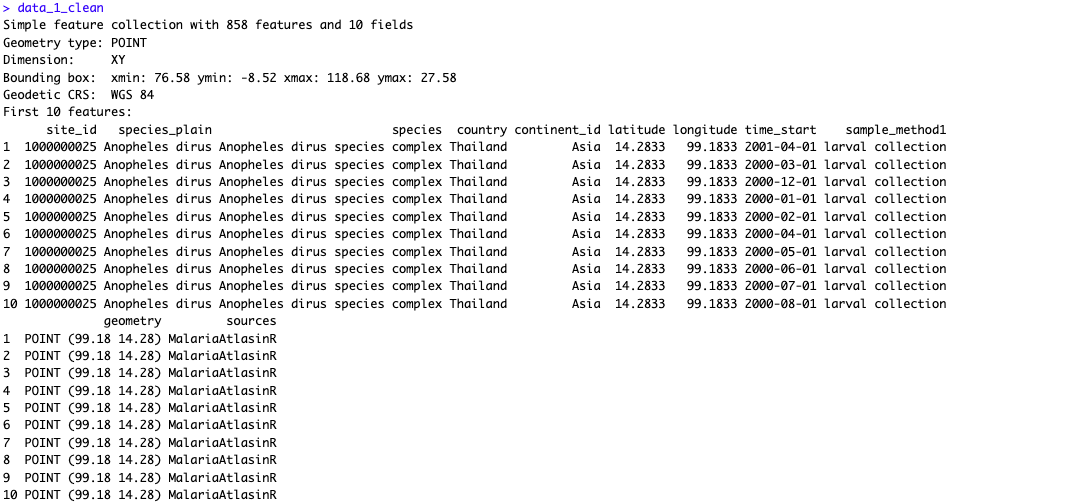
Questions & concerns

1. Malaria Atlas data









(Pre-processing)

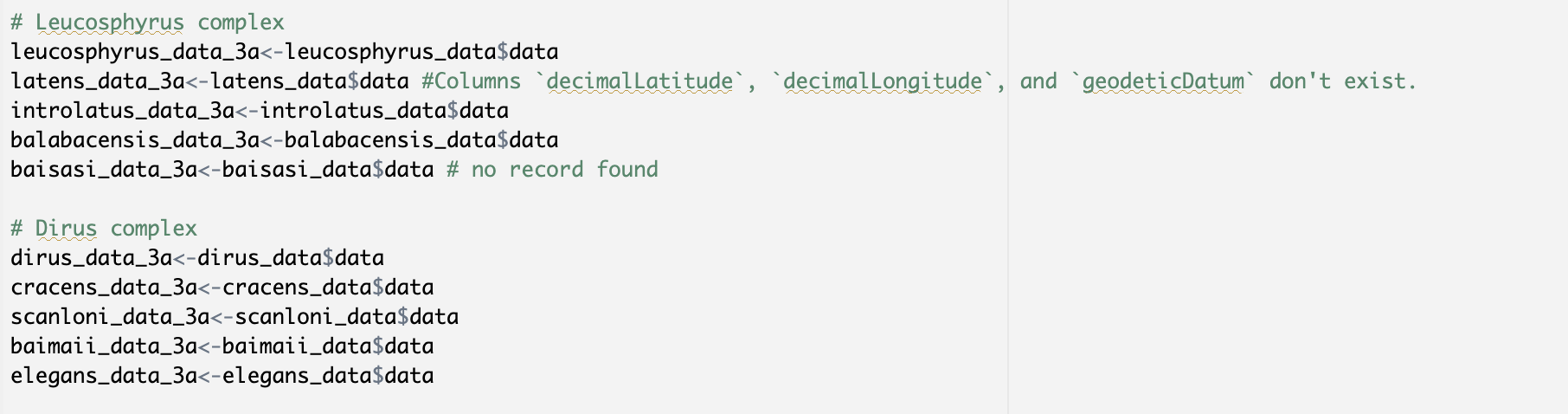
# Only three species available

# Round geometry to 2 d.p.

# 150 NAs in Event Date

## Duplicate data

1. RGBIF

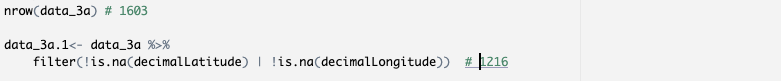


A close-up of a computer screen

Description automatically generated

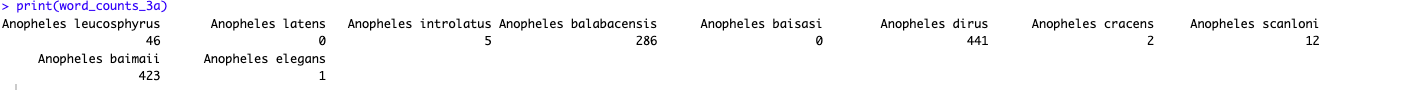


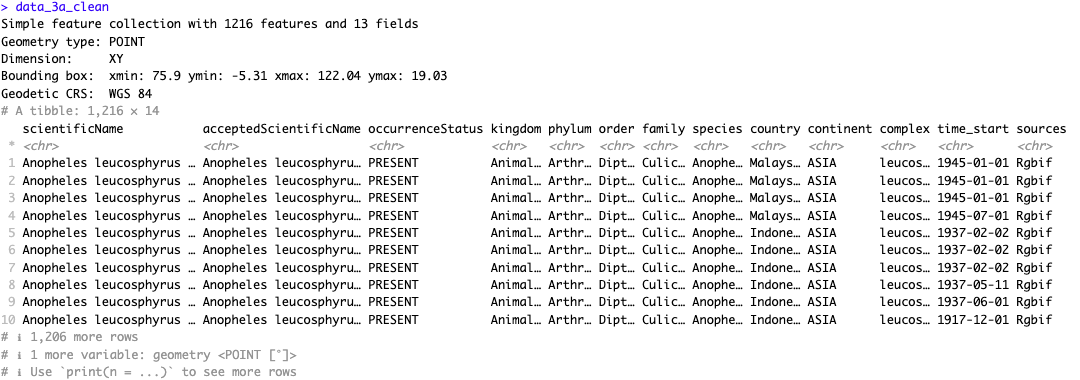




A screenshot of a computer program

Description automatically generated





(Pre-processing)  
# Anopheles latens do not have latitude& longitude.

# Anopheles baisasi doesn’t exists.

# Indonesia & China contained Na for continent.

# Decimal place for lat & long.

# Removed geometry data with missing values

# Date standardisation. (Concern in Year)

# 58 NAs in Event Date

1. GBIF Website

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated



A number of numbers on a white background

Description automatically generated

A screenshot of a computer program

Description automatically generated

A close-up of a text

Description automatically generated

A screenshot of a computer

Description automatically generated

(Pre-processing)  
# Changed Country Code to full name.

# Added Continent attribute

# Date standardisation

# No missing geometry data

# Decimal place for lat & long.

# 66 NAs in event\_date

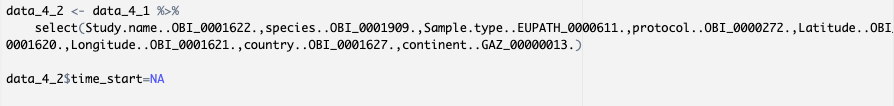
1. Vectorbase – Relational database

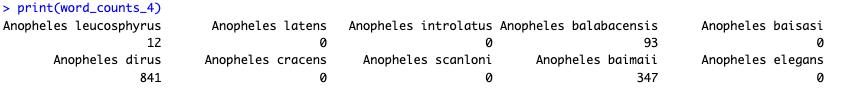
A screenshot of a computer

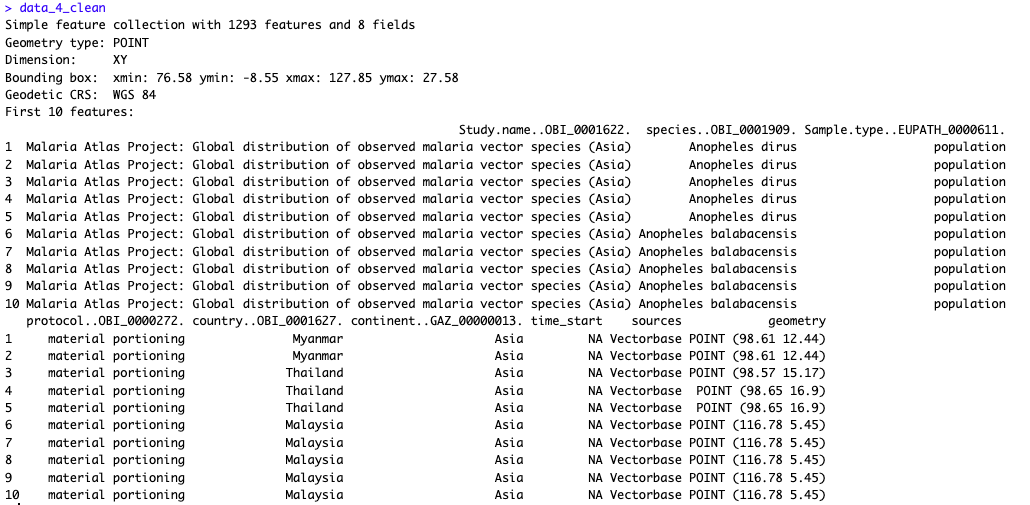
Description automatically generated

A close-up of a number

Description automatically generated







# Collection site; samples; studies

# Doesn’t contain any NA values

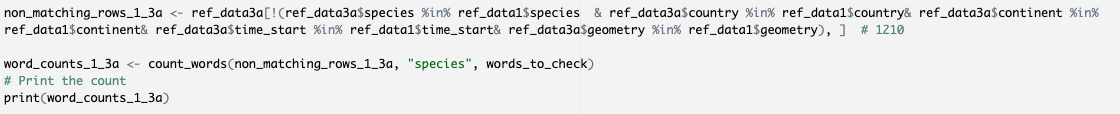
# Doesn’t contain eventDate - NAs

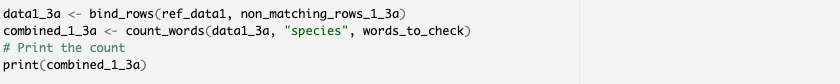
# Decimal place for lat & long

### Combine cleaned data.

A screenshot of a computer code

Description automatically generated





A screenshot of a computer

Description automatically generated



Limitation:

* Failed to include/ identify occurrences with similar attributes, but from other sources.
* More effective method to merge dataset.