Wrangling & Analysis of a Mammal Collection Dataset from Western New Mexico University

Prepared By: Savannah Wallis
Date: February 6, 2026

Institution: Dartmouth Geisel School of Medicine
Collaborators: Cassie Duncan, Jillian Melbourne, Kundan Rao
Dataset Source: GBIF Mammal Collection Data (Western New Mexico University)

INTRODUCTION

Animal collections are a widely used method for studying the natural world. Aristotle is credited for being the first person who collected, dissected, and preserved animals for the purpose of scientific discovery (Miller, 2023). Now, animal collections are commonly stored in museums and continue to grow with more specimens every year. The purpose of biological collections is to preserve organisms as physical snapshots, capturing their condition at a specific point in time when they were in a particular environment. Current discoveries from collections often relate to biodiversity and climate change (Nachman et al., 2023).

This report will focus on mammal collection data. This project is in collaboration with data scientists Kundan Rao, Jillian Melbourne and Cassie Duncan. Each data scientist will examine an individual mammal collection dataset from a specific university. The overarching goal is to clean and utilize analytical methods to summarize the collection using the results from a few select research questions. Topics of interest include collection dates, collection counts, diversity and locations of the collected mammals. Therefore, four research questions have been selected to further understand the dataset:

1. How do the mammal latitudes (at collection) vary by seasons?
2. How does species diversity vary by hemisphere of collection?
3. How are collections spread across the state of New Mexico?
4. How do mammal collection counts change over time?

The methods used to answer these questions involved extensive data cleaning and coding.

MATERIALS & DATA

The dataset is accessible via GBIF and collected by Western New Mexico University. These records are considered preserved specimens within a biological collection. The dataset is downloadable and open source. I utilized the “simple” zip file in order to focus on the main content and ignore the metadata. The programming language used was R, and analyses were conducted in RStudio. Installed packages: readr, dplyr, tidyr, lubridate, ggplot2, and maps.

METHODS

I. Discovery

After the data was loaded into RStudio, an exploration session took place in which the structure and dimensions were examined. The dataset included information about each specimen, including taxonomy, collection coordinates, country, and identification details. The data was then searched for missing values or duplicate records. Several columns had NAs, however no duplicate records were found. The set was then filtered to the top ten species and countries for a broader understanding of data trends. Lastly, important numerical columns, such as years, latitudes and longitudes, were summarized.

II. Cleaning

To fully prepare and pre-process the data, a detailed cleaning procedure was conducted. The first cleaning step involved removing rows that were completely empty and added no information. The dataset contained 120 records, therefore, when a column appeared with 120 NA values, it was considered empty and dropped. This included: verbatimScientificNameAuthorship, individualCount, typeStatus, mediaType, establishmentMeans, rightsHolder, depthAccuracy, coordinatePrecision, and depth.

The next step involved dealing with the inconsistent missing values. During discovery, many blank cells were present but RStudio was not able to process them as NAs due to a string likely being present in the space. Therefore, a function was used to convert all "none", "null", "na", "n/a" strings into true NA values. This transformation would later improve visualization results.

Another cleaning task involved fixing the confusing date formats for the collection event. Many cells included a date range for the collection, such as 2008-03-09/2008-03-10. This was likely done if a specimen was found close to midnight or if the date was uncertain. I decided to split this collection date column into two columns. The first column was event_range_start_date and the second was event_range_end_date. This would make it easier for a collection event to have a date range if needed. I also ensured both new columns were in standard date format. After examining more date issues, I found that the column date_identified was not in date format and fixed it by transforming it with as.Date().

Next, after further examination of the identifier columns, it was discovered again that many strings were in cells that should have been NA, and there was no standardization of these strings either. Therefore, in the identified_by and recorded_by columns, I looked for strings like "unknown", "Collector(s): unknown", and replaced them with true NAs again.

Lastly, the timestamps for the last_interpreted column were not in a standard format and not easily processed in R. I utilized the lubridate package and its function ymd_hms to set a standard date, time, and timezone.

III. Structuring

I conducted a simple structuring procedure to ensure the data was organized and consistent. I lowercased all the column names and placed underscores for spacing. This allowed for easier readability. I also re-ordered the columns so that the ID was first, the hierarchy of taxonomy went from kingdom to species correctly, and other important columns, like coordinates and country, were closer to the front.

IV. Enrichment

I enriched the data via several methods. I added seasons, hemisphere, time since last interpreted, and the time it took to identify the species.

To add a column with seasons, I simply used ifelse statements based on the months columns. For example, the season column would display "Winter" if the month column was either c(12, 1, 2).

Additionally, if the latitude column displayed a number greater than zero, then the column would display "Northern" hemisphere. Otherwise, it would be "Southern".

The last columns involved minor calculations, in which the function difftime found the exact amount of time that had passed (according to the timestamp) since the specimen was last interpreted. difftime also allowed for a function in which the collection date was subtracted from the identification date. This displayed how long it took to properly interpret the species after it had been collected.

V. Data Integration

Data integration is another essential step for enriching a dataset. A key column of information missing from this dataset was the species common names, which would make the data more interpretable for all audiences. A CSV file containing thousands of mammal scientific names and common names was downloaded from the American Society of Mammalogists. A left join was used, in which the datasets were matched up via the species name, and only the common-name column was added. Although not every mammal had a match in the join, it enriched my data and filled in several animal common names.

VI. Validation

The validation step is key to ensuring every column contains logical information and numerical data. I decided to look at the minimum and maximum values for all numerical information. I was looking to see if the coordinates were as expected, the dates were possible, and elevation was rational. Next, I individually grouped character columns by counts to ensure there were no incorrect or repeating groups. I also re-did some of my discovery steps by checking for missing values or duplicates. Everything appeared correct and ready for analysis.

RESULTS & VISUALIZATIONS

The research questions from the introduction were addressed using visualizations in RStudio.

I. How do the mammal latitudes (at collection) vary by seasons?

To answer this question, ggplot2 was used to create a boxplot visualization of mammal latitude at the time of collection against the four seasons. The resulting figure displayed more variability in the mammal collection latitude during summer and winter compared to fall and spring.

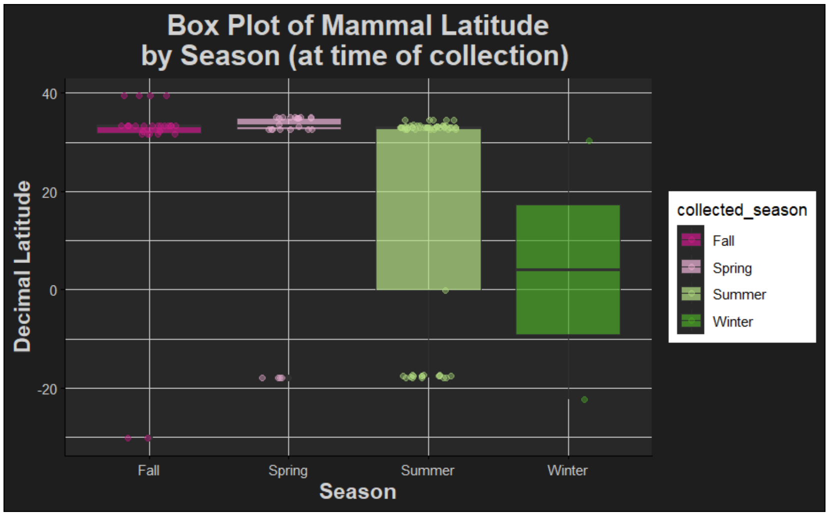

Figure 1. Boxplot of mammal collection latitudes by season

II. How does species diversity vary by hemisphere of collection?

To answer the research question, a faceted bar chart with top 10 species (by count) within each hemisphere was created. The southern hemisphere contained more bat species and much lower collection counts. The northern hemisphere contained mostly mouse species and a higher number of collections.

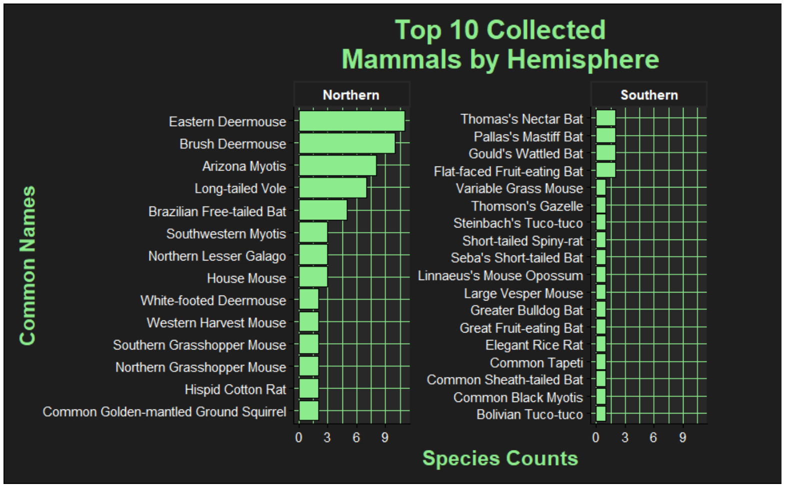

Figure 2. Bar charts of top 10 species collection counts by hemisphere

III. How are collections spread across the state of New Mexico?

To address this question, I utilized ggplot2 and the maps package to create a visualization of mammal collections within the state of New Mexico and around the university. It appeared that many collections occurred close to the university on the western side of New Mexico compared to the rest of the state.

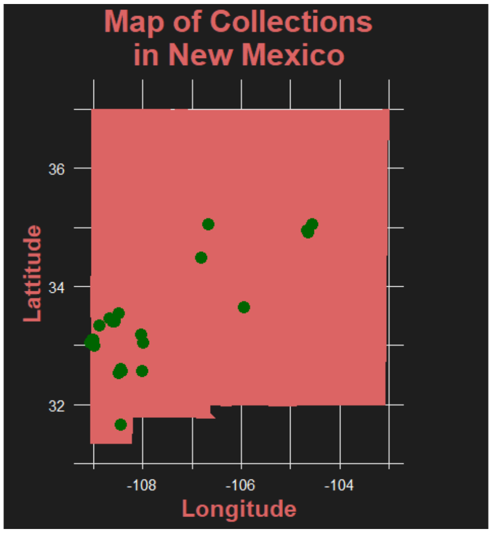

Figure 3. Map of WNMU mammal collection distribution over New Mexico

IV. How do mammal collection counts change over time?

The main goal of this question was to use a ggplot2 histogram and plot the collection counts over each year (each year as a factor). The histogram showed a slow upward trend between 1956 to 1995. However, no further records have been added to the dataset since 2008.

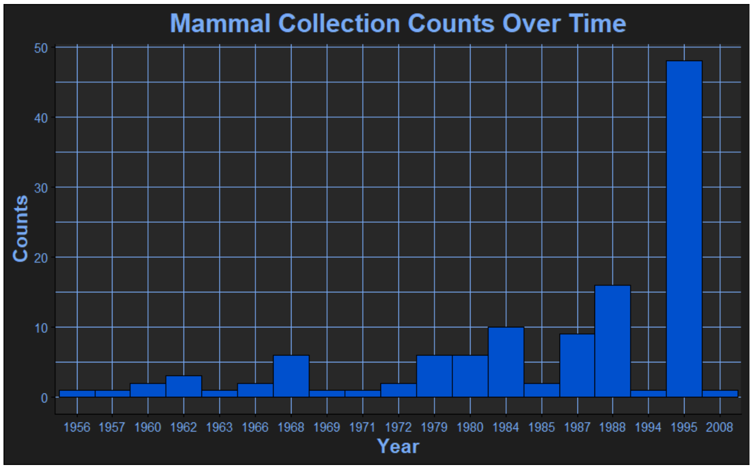

Figure 4. Histogram of WNMU mammal collection counts by year

DISCUSSION

Mammal latitude had a greater variance in winter and summer months. Therefore, collection locations tended to occur at higher latitudes during fall and spring. However, it is also likely that winter and summer breaks for WNMU students allowed for further travel. These results suggest that there may be a difference in collection latitude by season.

The species collected within each hemisphere also differed greatly when considering diversity and count data. The southern hemisphere had a high collection of bat species while the northern collected several mice species. Collection in the southern hemisphere was also completed in much lower numbers. It is important to consider that Australia, Kenya, and Bolivia were the only countries visited outside of the United States. Therefore, these results suggest that the collections from each hemisphere resulted in a uniquely diverse range of species.

The map output displayed many collection points centered around WNMU, likely meaning that collectors often didn't travel far from the university. Upon further research, a major national park of New Mexico is located in close proximity to WNMU, therefore, it is possible mammal collection was easier at this park compared to the rest of the state.

The collection over time of mammal species at WNMU showed an apparent increase over the years. Results suggest that collections were being conducted more and more frequently over time from 1956 to 2008. However, the dataset either needs to be updated, or has halted adding specimens after 2008.

CONCLUSION

Western New Mexico University's mammal collection dataset is diverse across hemispheres, has increased in collection pace over time, often includes local specimens, and seasons appear to impact the location of collection in some way.

REFERENCES

ASM Mammal Diversity Database. (n.d.). Www.mammaldiversity.org. https://www.mammaldiversity.org/

GBIF. (n.d.). www.gbif.org. https://www.gbif.org. https://www.gbif.org/occurrence/download/0015128-260129131611470

Miller, R. J. (2023). Greek Awakenings. *Oxford University Press EBooks*, 9-C1N11. https://doi.org/10.1093/oso/9780197665756.003.0002

Nachman, M. W., Beckman, E. J., Rauri CK Bowie, Cicero, C., Conroy, C. J., Dudley, R., Hayes, T. B., Koo, M. S., Lacey, E. A., Martin, C. H., McGuire, J. A., Patton, J. L., Spencer, C. L., Tarvin, R. D., Wake, M. H., Wang, I. J., Anang Achmadi, Sergio Ticul Álvarez-Castañeda, Andersen, M. J., & Arroyave, J. (2023). Specimen collection is essential for modern science. *PLOS Biology*, *21*(11), e3002318–e3002318. https://doi.org/10.1371/journal.pbio.3002318