Analysis of Tick-Bourne Disease Cases Reported to DPH Between 2019-2024

The structure below is one possible setup for a manuscript, or a general data analysis project (including the course project). Adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

**Authors**

* Hope Grismer (ORCID: 0000-0000-1234-5678)
* Second Author
* Third Author
* Last Author

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Southeastern Center of Excellence in Vector-Bourne Diseases (SECVBD), Gainesville, FL 32611.
3. Yet another Affiliation.

These authors contributed equally to this work.

Corresponding author: hope.grismer@uga.edu

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

Tick-borne illnesses are diseases that humans can contract from the bite of an infected tick. Ticks are small, blood-sucking arachnids that are often found in wooded, grassy, or suburban areas. When a tick feeds on an infected animal, it can acquire pathogens—such as bacteria, viruses, or parasites—that can then be transmitted to humans. Examples of tick-bourne diseases include, but are not limited to, Lyme disease, anaplasmosis, and babesiosis. These diseases are of increasing public health concern, especially as the geographic range and seasonality of ticks expand due to environmental and ecological changes.

## 2.2 Description of data and data source

This data set includes all human cases reported to the Georgia Department of Public Health (GA-DPH) between 2019 and 2024. It was graciously provided by DPH professionals in the epidemiology department. All available data for each case over this five-year period across the state of Georgia. Proper de-identification practices– the removal of first name, last name, and date of birth (DOB)– has been taken. The dataset includes variables such as disease type, patient demographics (gender, age, district, state, county, and ZIP code), medical details (incident ID, report ID, lab test dates, results, serotype, and titers), as well as disease outcomes (death status, fever, migranes, and administrative status). Reported diagnoses included in this dataset include Toxoplasmosis, Dengue, Rocky Mountain Spotted Fever, Anaplasma Phagocytophilum, Malaria, Lyme Disease, West Nile (WNV) Infection, Chikungunya, LaCrosse (LAC) Infection, Zika, Ehrlichia Chaffeensis, and Typhus (Epidemic/Louse/Flea). Some of this cases, based on adherence to specific case definitions and subsequent lab results, are categorized as suspect, probable, confirmed, or unconfirmed. By examining case reports, public health professionals can better understand the prevalence, demographics, and regional variations of these diseases. This analysis aims to further inform the public health understanding of tick-bourne disease in Georgia.

## 2.3 Questions/Hypotheses to be addressed

1. How have vector-borne diseases (VBD) changed over time, and what seasonal patterns or climatic factors influence these trends?
2. What is the distribution of vector-borne diseases across different health districts, counties, and ZIP codes?
3. How complete is the data for vector-borne diseases, and how has data completeness evolved over time?
4. What is the timeline between key dates (onset, lab test, and date of entry), excluding same-day onset and lab results, and how does this timeline vary over time?
5. What trends exist in malaria travel cases over time, and which countries are most frequently visited by affected individuals?
6. What are the trends and changes observed in relative incidence of different disease pathologies between 2019-2024?

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (1,2).

# 3. Methods

**I. Uncleaned Data**

The data contains 6255 entries representing discrete cases of tick-bourne diseases. For each case, some or all of the following data variables were collected: disease type, patient demographics (gender, age, district, state, county, and ZIP code), medical details (incident ID, report ID, lab test dates, results, serotype, and titers), and disease outcomes (death status, fever, migranes, and administrative status). For some cases, specific variables were not reported and, therefore, are reported as blank spaces in the data set.

***II. Initial Data Cleaning Stages***

Looked for Missing Data & General Trends

* Identified Disease Burden Trends & Hotspot Counties
* Generated a stacked bar graph to visualize case distribution across regions. - Checked Age Distribution
* Verified that no cases had missing age values in the cleaned dataset.
* Conducted an exploratory analysis of disease counts by age group.
* Evaluated Data Completeness Over Time
* Extracted onset date information to track missing data trends.
* Created a plot distinguishing between blank and non-blank STATUS entries over time.
* Identified periods of high missingness, which may indicate data entry errors or reporting gaps.

Analyzed Timeline Between Key Case Events

* Calculated time differences for:
  + Onset date (DOO) to first lab test date (LABDATE1)
  + First lab test (LABDATE1) to second lab test (LABDATE2)
  + Second lab test (LABDATE2) to case update date (UPDATEDATE)
  + Summed these values to determine the total case duration. - Saved the processed dataset as a CSV file for further analysis. - Created a formatted HTML table to present extracted time intervals.

Addressed Data Quality Issues

* Identified negative time values in lab test dates, indicating data entry errors
* Implemented a filtering step to remove records with invalid date sequences

***III. Analysis Process***

* Still determining this

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). **?@fig-schematic** is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

## 3.2 Data aquisition

The data used in this analysis was obtained through my internship with the Georgia Department of Health, in collaboration with the Centers for Disease Control and Prevention (CDC). Specifically, it is derived from the Nationally Notifiable Diseases Surveillance System (NNDSS) and focuses on tick-borne illnesses. This dataset provides key insights into the trends, distribution, and characteristics of reported cases, supporting public health efforts in disease monitoring and prevention.

## 3.3 Data import and cleaning

The data cleaning and preprocessing workflow begins with loading essential R packages for data handling (dplyr, tidyr, readxl), summary visualization (skimr), and file path management (here). The raw dataset, stored in an Excel file, is imported into R for structured processing. To assess data quality and structure, we utilize functions such as glimpse(), head(), and summary(), which provide an overview of column names, data types, and potential inconsistencies like missing or unexpected values. Additionally, skimr::skim() offers a detailed summary, highlighting numerical distributions and missing data patterns.

The initial cleaning step removes records where the STATUS field is either missing or not classified as “CONFIRMED,” ensuring that only officially verified cases are included. The dataset is then refined further by retaining only cases with a confirmed “positive” test result, excluding records with missing, blank, or ambiguous values (e.g., “inconclusive”). This process initially removed 15 cases that did not meet the case definition in Georgia, leaving 95 valid records. Upon further review, filtering criteria were adjusted to ensure all relevant positive cases were retained while systematically excluding non-positive values such as “Negative,” “Equivocal,” “Not detected,” and other irrelevant categories. This refinement resulted in a more accurate dataset and a clearer summary table of key case information. The cleaned dataset is saved at each stage to maintain version control and reproducibility.

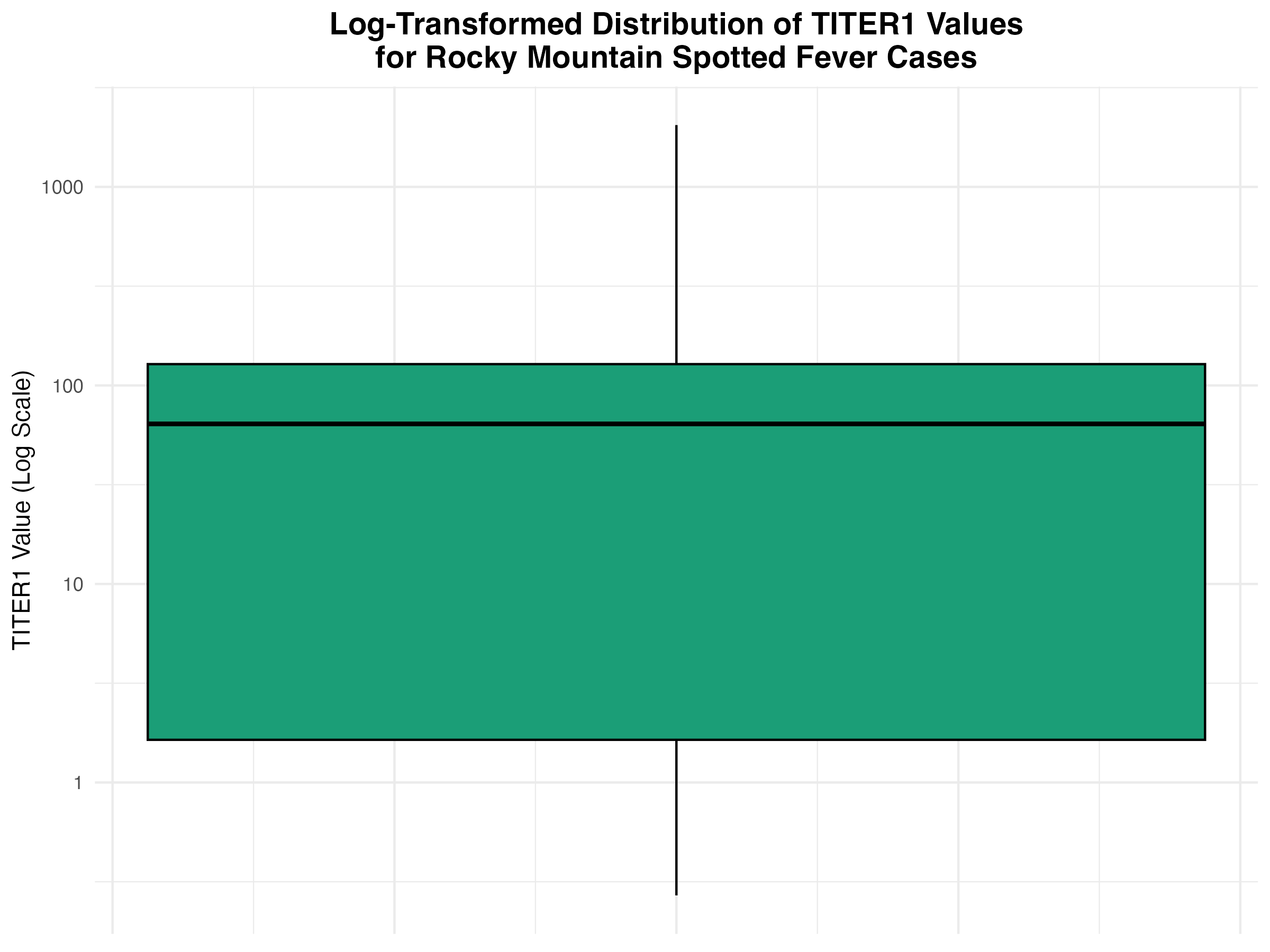
## 3.4 Key Variables in Data Wrangling - Preparing for Analysis

To effectively present the cleaned dataset, key variables—including DISEASE, STATUS, COUNTY, GENDER, AGE, LABTEST1, and RESULT1—are displayed in a professionally formatted table using the kableExtra package. This enhances readability and facilitates trend analysis. Several visualizations are generated using ggplot2 to explore patterns in tick-borne diseases in Georgia between 2019 and 2024: Disease Distribution: A histogram illustrates the prevalence of different tick-borne diseases, highlighting the most commonly reported cases.

**Rocky Mountain Spotted Fever (RMSF) Severity:**

A box plot of TITER1 values (log-transformed for better visualization) provides insights into the variability of test results. RMSF severity patterns are considered analogous to Alpha-Gal Syndrome, a topic of interest, though not explicitly included in this dataset as it is not a reportable disease in Georgia.

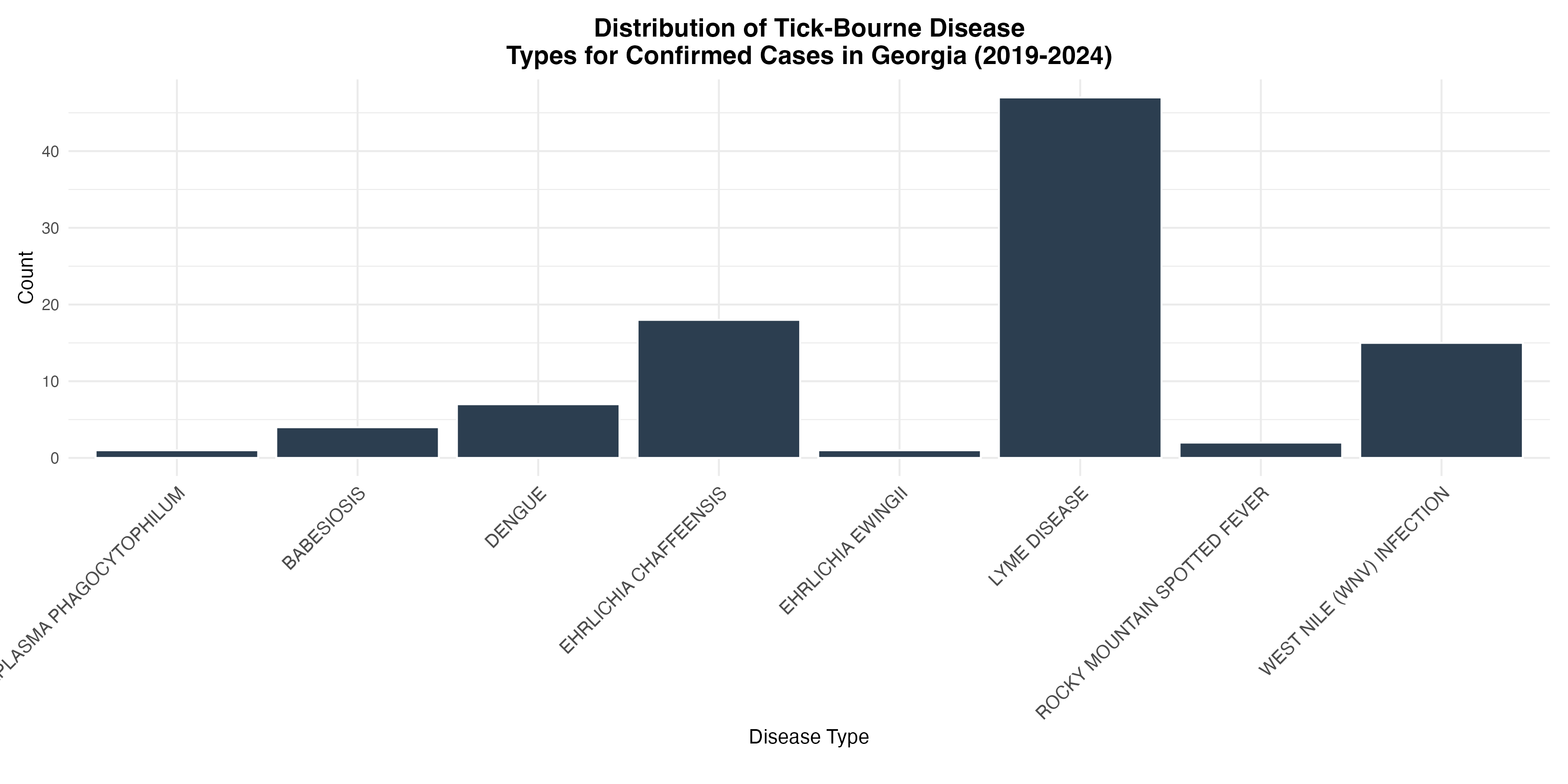
library(here)  
library(knitr)  
  
# Generate correct file path  
img\_path <- here::here("results", "figures", "rmsf\_titer\_boxplot\_log.png")  
  
# Display the image  
knitr::include\_graphics(img\_path)



**Most Common Tick-Borne Diseases:**

The dataset is filtered to identify the five most frequently reported tick-borne diseases. A bar chart visualizes their case counts, offering an overview of significant public health concerns.

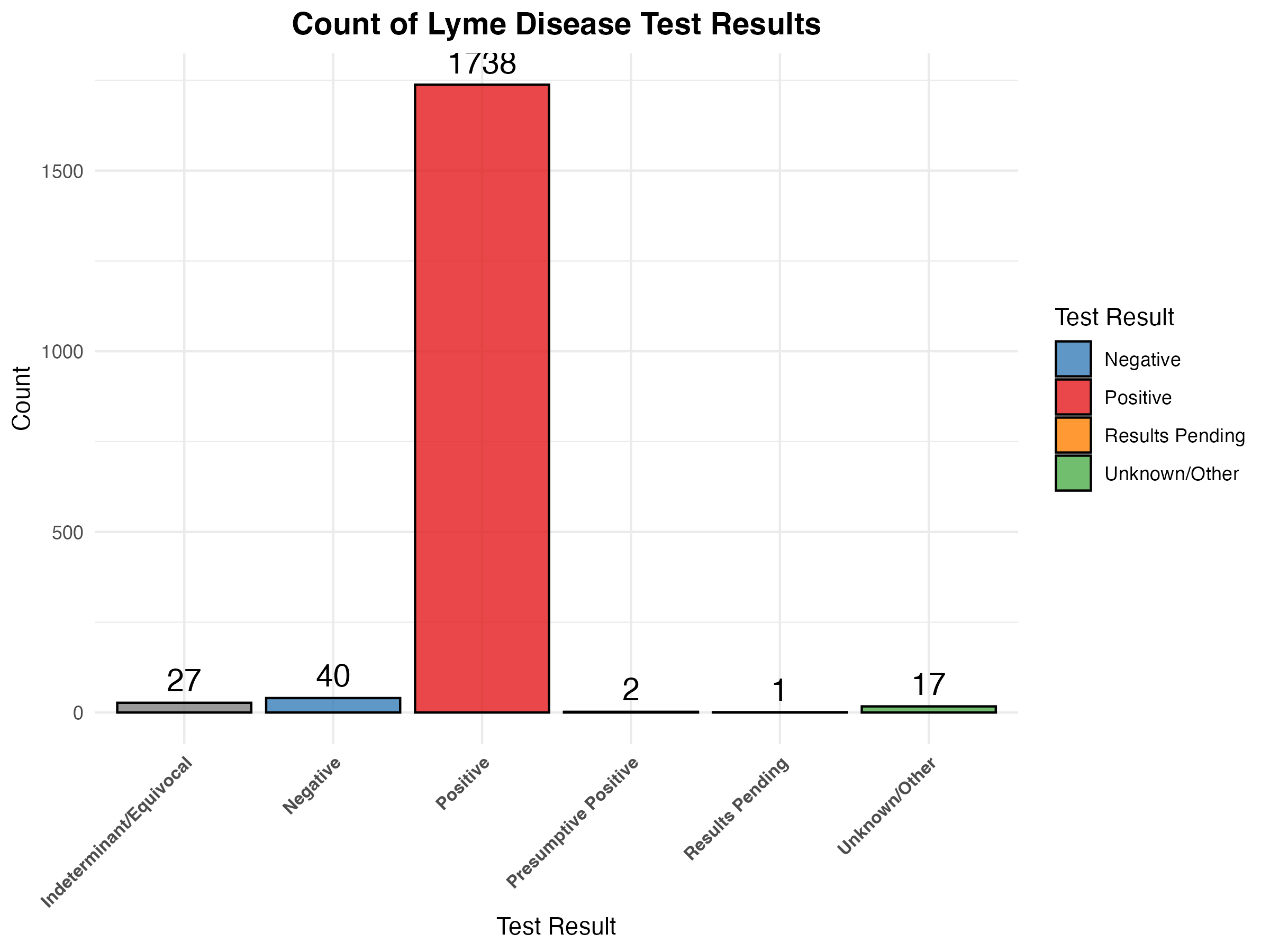
library(here)  
library(knitr)  
  
# Generate correct file path  
img\_path <- here::here("results", "figures", "disease\_histogram.png")  
  
# Display the image  
knitr::include\_graphics(img\_path)



**Lyme Disease Analysis:**

Given its similarity to Alpha-Gal Syndrome (which is of interest to me and the topic of my capstone), Lyme Disease cases are analyzed separately. A count histogram compares positive, negative, and other test outcomes, providing insights into testing patterns and diagnosis rates.

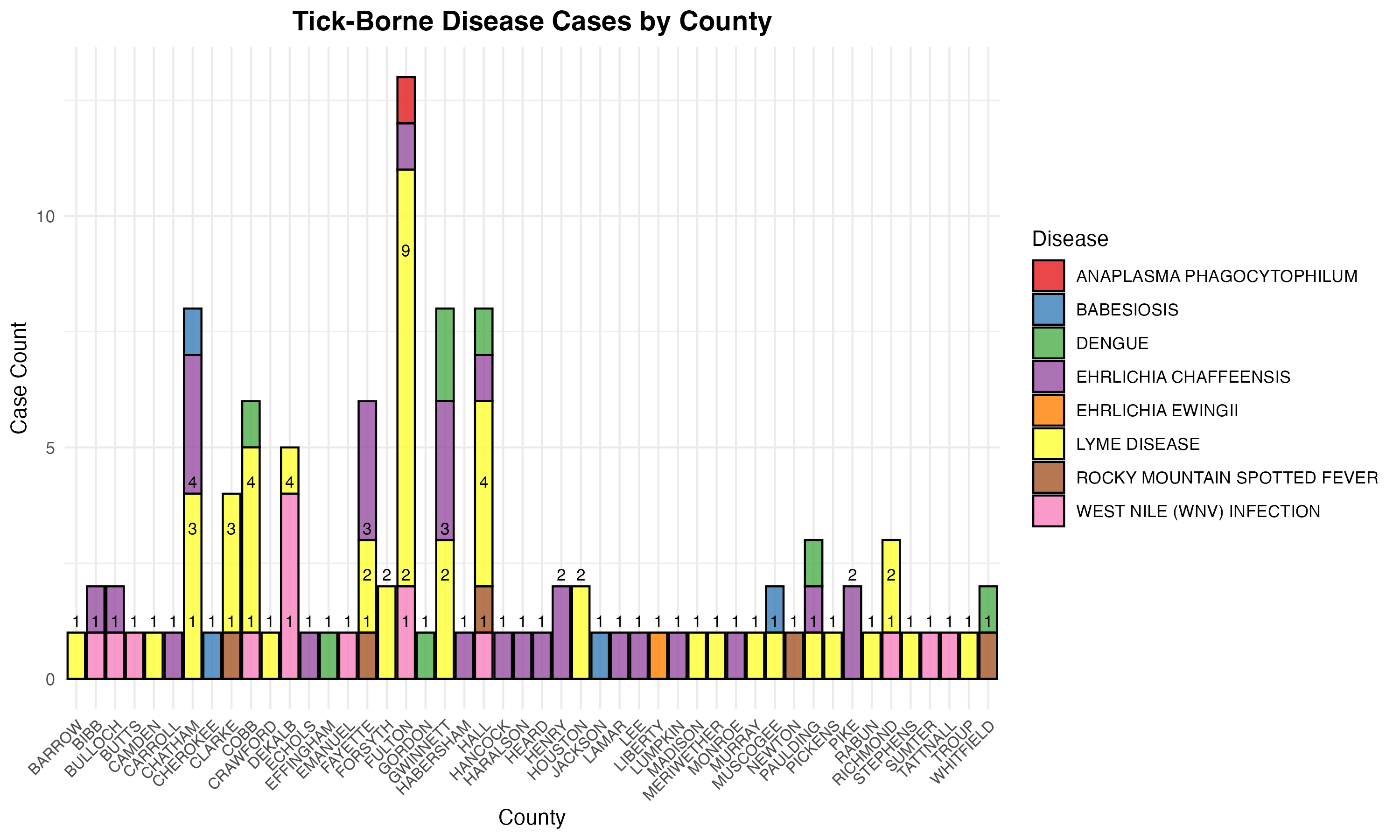
library(here)  
library(knitr)  
  
# Generate correct file path  
img\_path <- here::here("results", "figures", "lyme\_result\_count\_histogram2.png")  
  
# Display the image  
knitr::include\_graphics(img\_path)

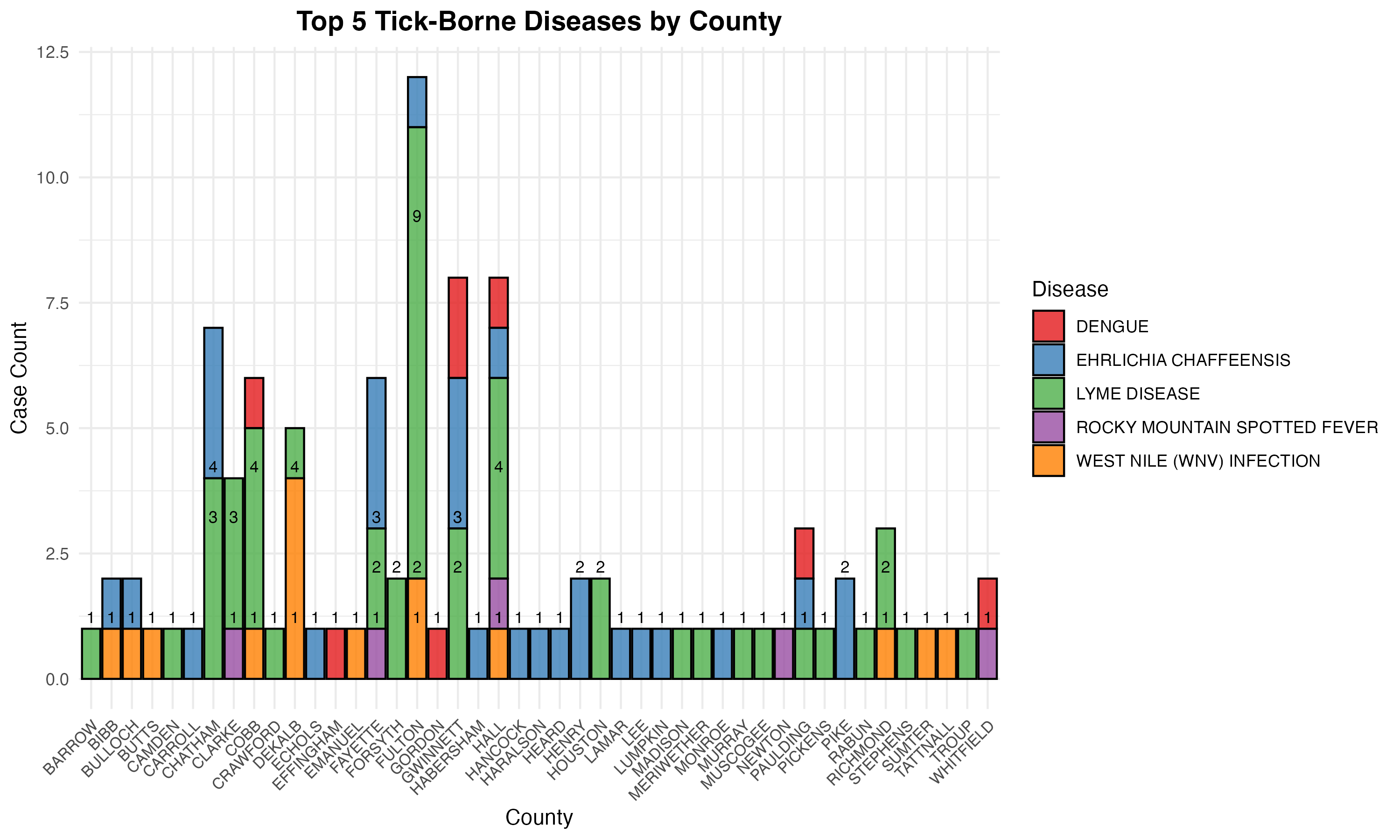


**Geographic Distribution:**

A grouped bar chart maps case counts across Georgia counties, categorized by disease type. Due to the density of information, this visualization is refined to focus on the top five tick-borne diseases and their distribution across the 12 counties with the highest confirmed case counts.

library(here)  
library(knitr)  
  
# Generate correct file paths  
img\_paths <- c(  
 here::here("results", "figures", "disease\_county\_plot.png"),  
 here::here("results", "figures", "top5\_disease\_county\_plot.png")  
)  
  
# Display the images  
knitr::include\_graphics(img\_paths)



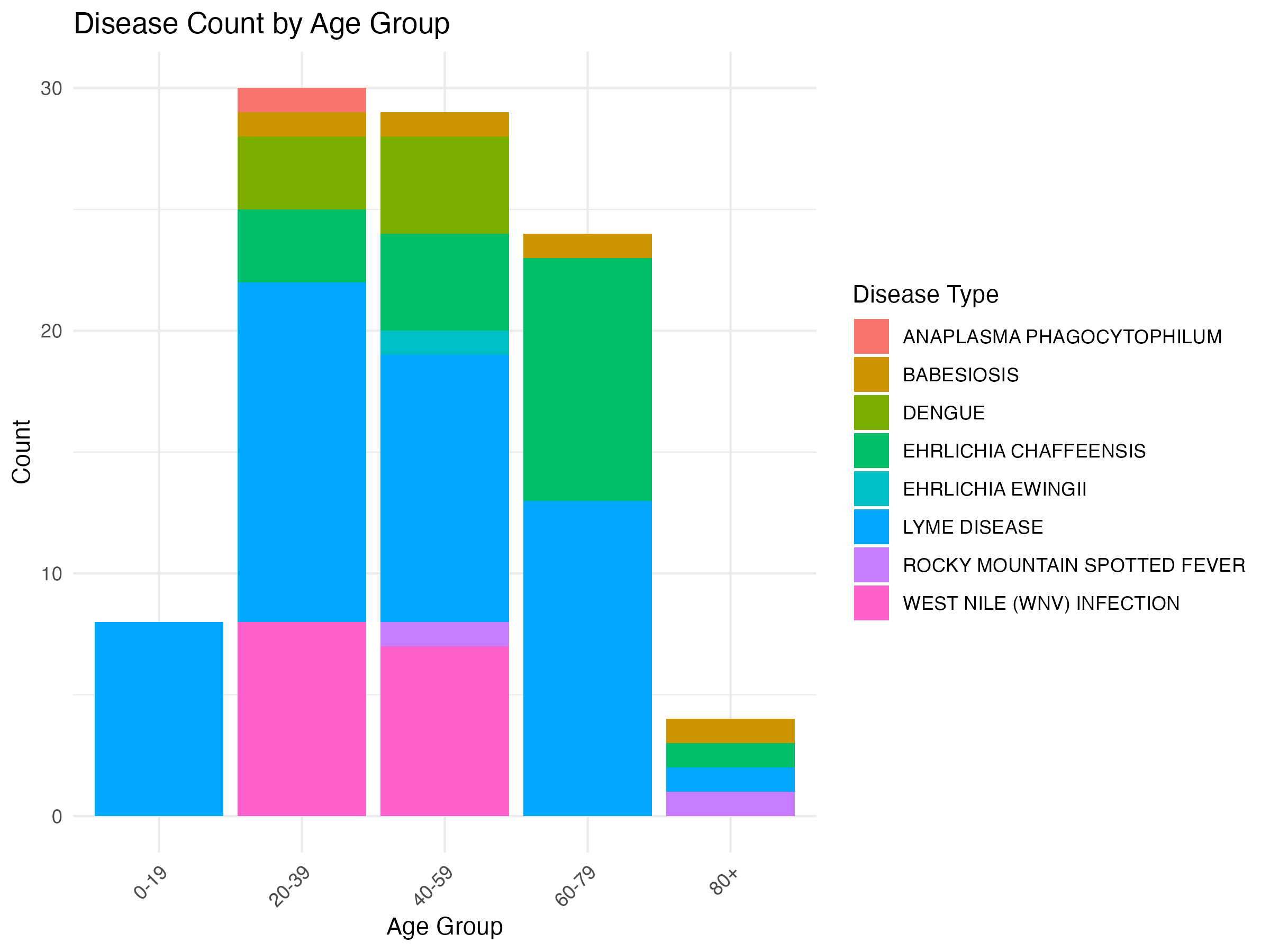


To analyze disease burden trends and identify hotspot counties, a stacked bar graph is generated to visualize case distribution across different regions.

**Age Distribution Analysis**

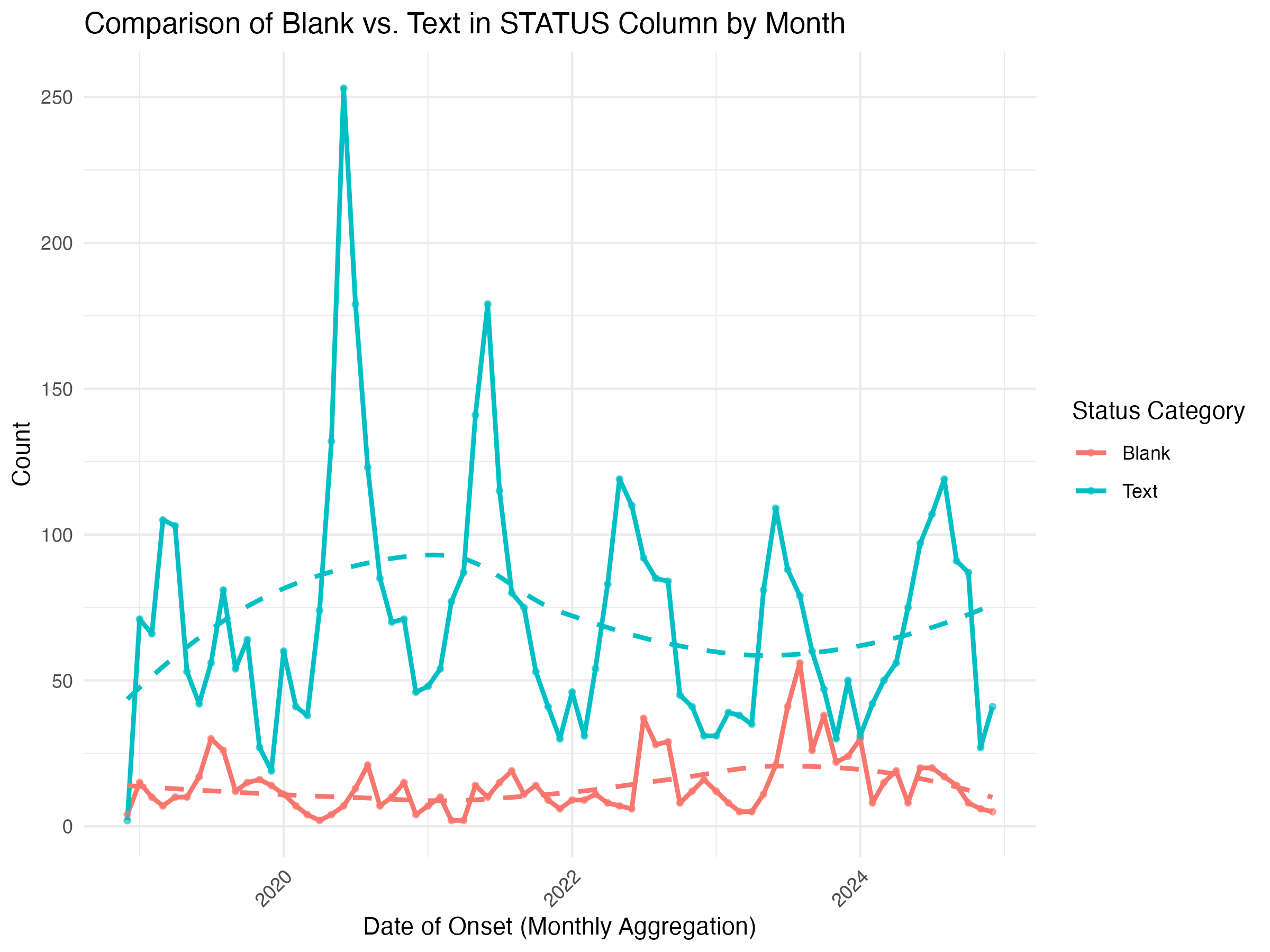
Age distribution is assessed as part of the initial data cleaning process, confirming that no cases in the cleaned dataset have missing age values. Additionally, an exploratory analysis examines disease counts by age group to identify potential patterns in disease prevalence.

library(here)  
library(knitr)  
  
# Generate correct file path  
img\_path <- here::here("results", "figures", "disease\_count\_by\_age.png")  
  
# Display the image  
knitr::include\_graphics(img\_path)



**Data Completeness Over Time** To evaluate how data completeness has evolved, the dataset is processed to extract onset date information. A visualization distinguishes between blank and non-blank entries in the STATUS column over time, helping to identify periods with high missingness. Spikes in blank values may indicate data entry errors, system issues, or reporting gaps that should be addressed before conducting further analysis.

library(here)  
library(knitr)  
  
# Generate correct file path  
img\_path <- here::here("results", "figures", "status\_completeness\_plot.png")  
  
# Display the image  
knitr::include\_graphics(img\_path)



**Timeline Analysis of Key Case Events** Given the frequent loss-to-follow-up observed in tick-borne illness reporting, an analysis is conducted to examine time gaps between critical dates, including:

**Onset date (DOO) to first lab test date (LABDATE1)**

**First lab test (LABDATE1) to second lab test (LABDATE2)**

**Second lab test (LABDATE2) to case update date (UPDATEDATE)**

These durations are summed to determine the total case timeline for each record. The processed dataset is saved as a CSV file for further analysis, and a formatted HTML table is generated to present the extracted time intervals in a structured manner. Data Quality Adjustments Upon reviewing the time calculations, inconsistencies were identified where negative time values appeared between initial lab results and follow-up tests. These errors, likely due to data entry mistakes, made it impossible to determine the correct dates. To address this, a filtering step was added to remove erroneous records where follow-up test dates occurred before initial lab dates.

## 3.5 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

## 4.2 Basic statistical analysis

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

# 6. References

1. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

2. McKay B, Ebell M, Dale AP, et al. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) *Proceedings. Biological sciences*. 2020;287(1927):20200496.