EDA Analysis

An Analysis Of Data Collection Provided By Smithsonian

STP 420

Project overseen by

Neil Hatfield

Arizona State University

Spring Semester 2017

**Executive Summary**

This report summarizes the statistical analysis results associated with the data provided by the Smithsonian Vector Map project.

Assessment of the Culicidae family has become increasingly important as Big Science organizations such as the World Health Organization (WHO) and the Centers for Disease Control (CDC) begin to direct humanity’s attention to what has been called man’s greatest predator. As these organizations continue to draw from biogeographical, bio-statistical, bacteriology, virology, and computer science data, the emphasis becomes on fear. Our team decided that objectively we can analyze how this collection has changed over time with the data provided by the Smithsonian.

The purpose of this report is to compare the data provided by the Smithsonian Institute using techniques, concepts, and language taught in Neil Hatfield’s STP 420 class using JMP.

This project draws comparisons found in sample data provided by the Smithsonian Institute’s Vector Research program. The project will review data in Collecting Method from 1900 to 2015. This analysis will reveal where the samples were collected on a map of earth, and how the Collecting Methods vary compared to time, specimen count, and other variables. In addition, the analysis will show meaningful relationship between basis of record and life stage.

**Introduction**

Mosquitos have in a few generations become a public menace. From Disney PSA1s about the dangers of mosquitos to the Smithsonian Institute2 stating clearly that mosquitos are responsible for more human death than human murders, the conversation regarding the mystery of mosquitos has evolved greatly. The public knows about the dangers of several zoonotic diseases carried by the mosquito, but not many people know how the disease is carried.

Mosquitos are in the Order Diptera. Diptera are flying insect with a halter that allows them to move omni-directionally. Mosquitos are members of Culicidae which all mosquitos fall into. Culicidae are midge-like flies with proboscis, long straw like mouthparts, that the females use to drink flies. Mosquitos have 4 stages of life: eggs, larva, pupa, and adult. Most people have seen an egg, and most people have seen an adult mosquitos due to their many habitats across the world. The larvae, called by Disney a “wiggler”, is a small worm with protruding, hair-like structures at the front and sometimes the back end. They live in all types of water, but are most commonly found in standing water. After the larvae eats enough to store sufficient fat, the pupation process begins. The exoskeleton hardens and sinks until the adult is ready to emerge.

This Exploratory data analysis uses data provided by the Smithsonian that was collected from all over the globe from the past century up to 2013 and records many individual aspects of the mosquitos. From this data, the EDA report will explore the methods of collection, the amount collected, and other changes over time for the data collection provided.

**Introduction to Vocabulary**

* **Morphology**-The methodology for differentiating between species of mosquitos; different species and genera of mosquitos have a number of distinct features.
* **Habitat Importance**-Mosquitos tend to breed near water in varying quantities; this can include something as small as a puddle or as large as a lake.
* **Anopheles**-Commonly referred to as the Marsh Mosquito, this genus can transmit the parasite plasmodium as well as the type of malaria that humans contract.
* **Aedes**-This genus of mosquito can transmit zika and yellow fever, among other diseases. This genus is diurnal and can be identified by distinctive black and white spots.
* **Culex**-A specific genus of mosquito that transmits diseases such as avian malaria and west nile. Culex can be found in urban areas in the U.S.
* **Culcidae**-This family contains every genus of mosquito; the genera contain the species.
* **Proboscis**-The extended tube-like structure situated in the nasal region of the mosquito.
* **Adult, human landing/biting**-A mosquito lands on a human, bites the human, and has

their attributes measured and recorded.

* **Larval Collection**-Mosquitos are collected before they fully mature.
* **Adult, With Attraction Traps**-Adult mosquitos are trapped via luring them to the trap.
* **Adult, Non-Attraction Traps**-Adult mosquitos are trapped without a method of luring them to the trap.
* **Hand Collection (Using A Mouth Aspirator)**-A vacuum-like device is used to collect and contain the mosquitos.

**Introduction to Research Questions**

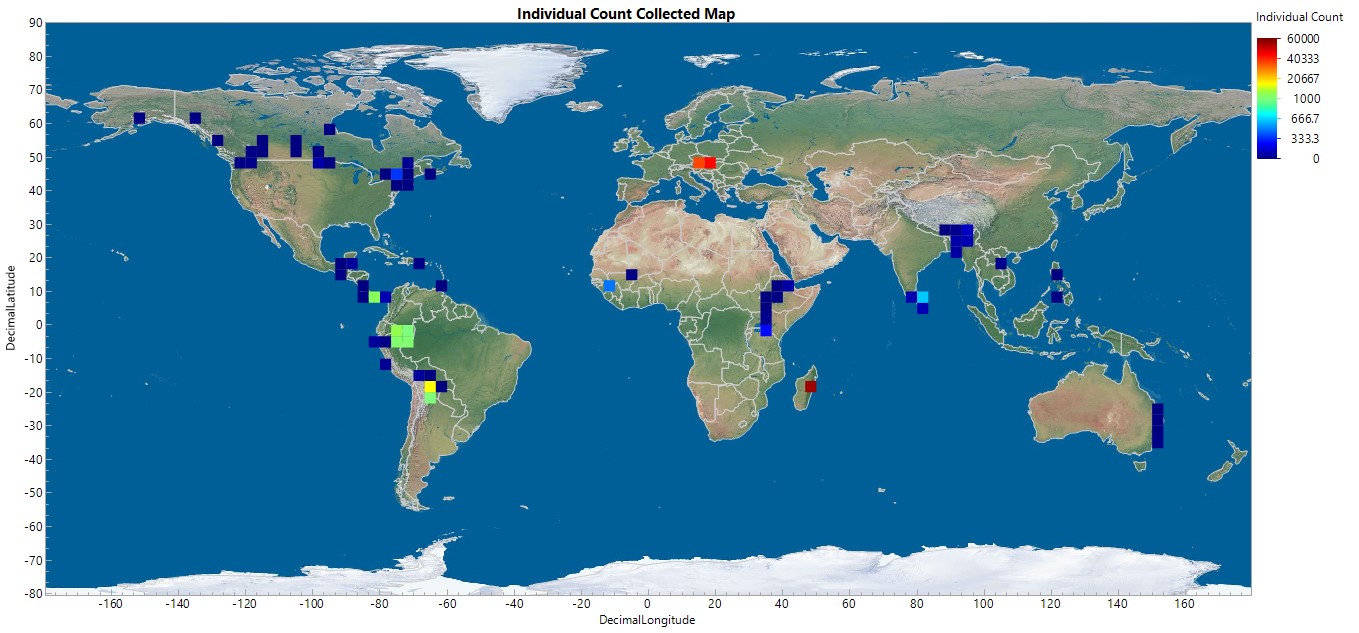
* Where are the data from, geographically speaking?
* Does the Individual Count of the collections impact our comparisons when implemented?
* Is there a difference between the Collecting Methods and the Year Collected?
* Is there an appearance of statistically significant relationship between Basis of Record and Lifestage?

**Description of the Data**

The data was provided by Smithsonian Institute’s Vector Map’s entomology department.

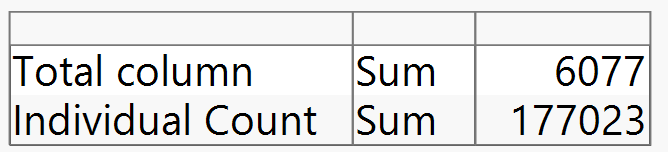
* **Description of variables**
  + **Location**-Location in this data set means the longitudinal variables and latitudinal variables combines.
  + **Year Collected**-In the case of collecting mosquitoes, this can be a lengthy process and record keeping goes back a long way. The year the mosquito was collected is an important tool for comparison.
  + **Time Collected**-The measure of the time of the specimen collected.
  + **Individual Count-** The total number of specimen in this data set. Sum of the specimens were larger collections so individual count was made.
  + **Basis of Record**-The methodology used to examine the specimen. Sometimes specimen is preserved and examined multiple times.
  + **Collecting Method**-This is how the mosquito specimen was captured.
  + **Genus (Sub-genus) Species**-Genus is a taxonomic category that is commonly used in binomial nomenclature as the first word in a scientific name
  + **Life Stage**-Mosquitoes have three main stages of growth that they go through as they mature these are the larva, pupa, and adult.
  + **Total**-This is the amount of observations or the whole number of rows in the data set.

**Collection Map**



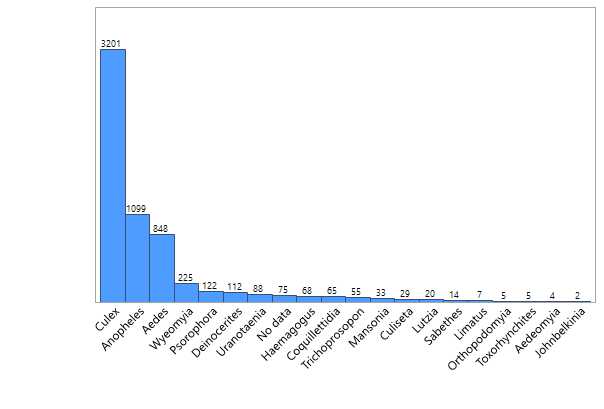
The Smithsonian Vector Map data collection that was given to us has been collected from all over the world. We can see from the map that the samples have been collected from most continents, excluding Antarctica. In addition, we can see that most of the individual counts of mosquitoes collected have been small. The larger counts are from the more recent collections.

**Tabulated Sums of Total Data Elements and Individual Count**



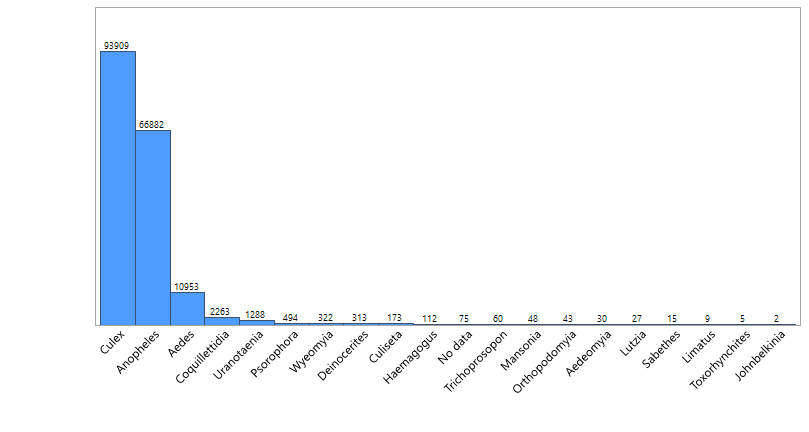
This visualization compares the total number of columns with the total number of individual mosquitos within the collections. This comparison will be critical in answering our research question “Does the individual count of the collection impact our comparisons when implemented?”

**Distribution of Genus**



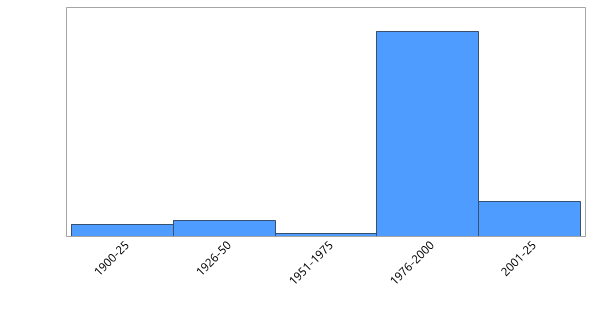
When we distribute the collection by genus we find that the collections visually show more Culex than any other genus and that Culex, Anopheles, Aedes are visually more present than any other genus. The frequency of species with “no data” on their genus is comparable to the frequency of the other minor genera. This portrays the collections without individual count.

**Distribution of Genus (frequency individual count)**

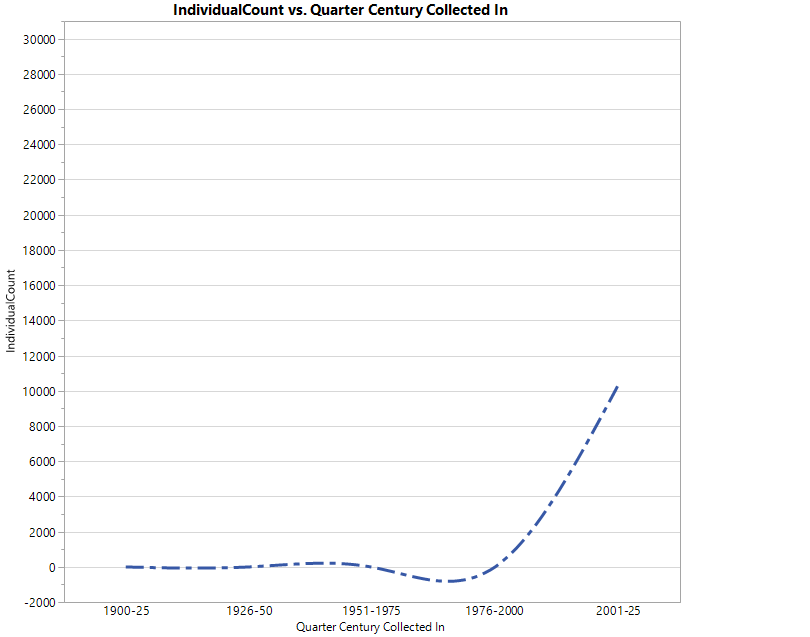


The above visualization compares the genus distribution with the individual count as the frequency. Just as the prior distribution suggested that the majority of the collections were Culex, the distribution with the individual count as frequency showcases the same top 3 genus as the previous, with an unchanged amount of “no data**.”** This provides an example of the Individual Count having a large impact when compared to the distribution without the individual count.

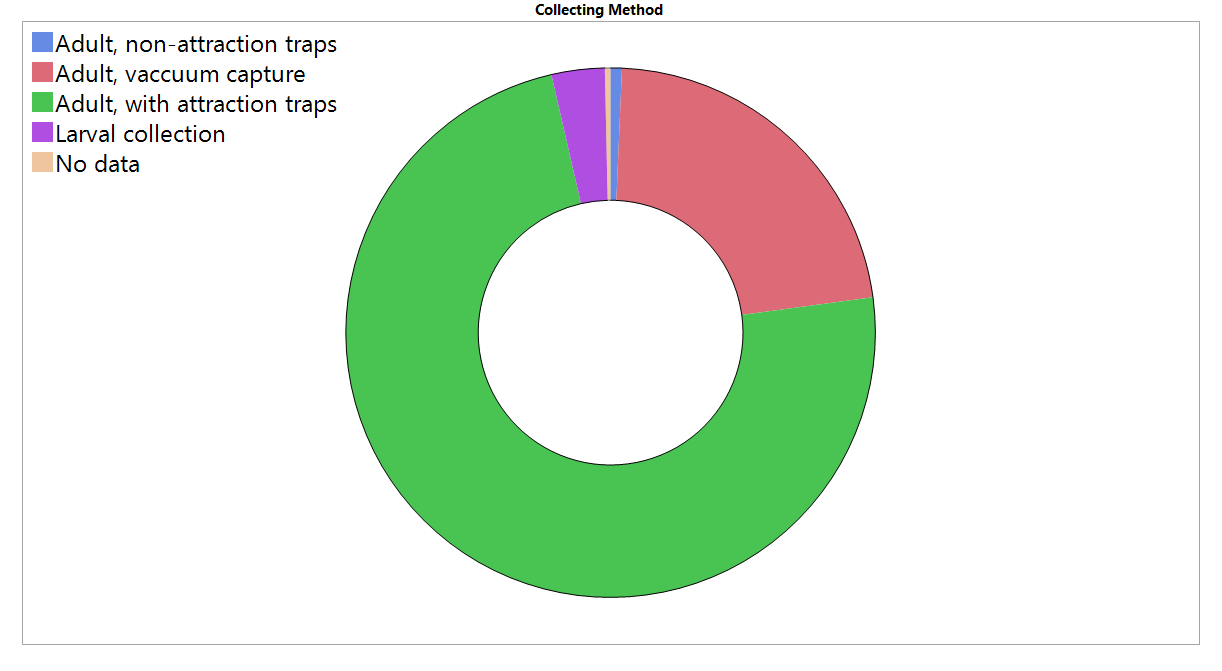
**Quarter Century Distribution of Collections**



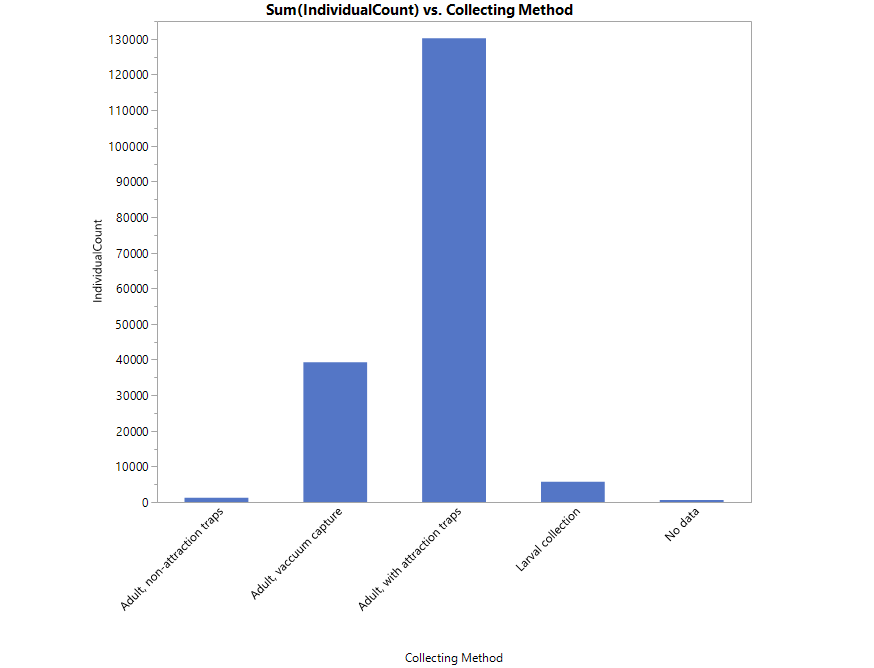
This distribution portrays the amount of collections provided in our data set broken up over quarter centuries. From the above we can also see that the collections within the data specific era has the most variety. This visualization also showcases the increase in sample collection starting after 1975. This increase could be explained by a global initiative to promote understanding and awareness of the mosquito threat.

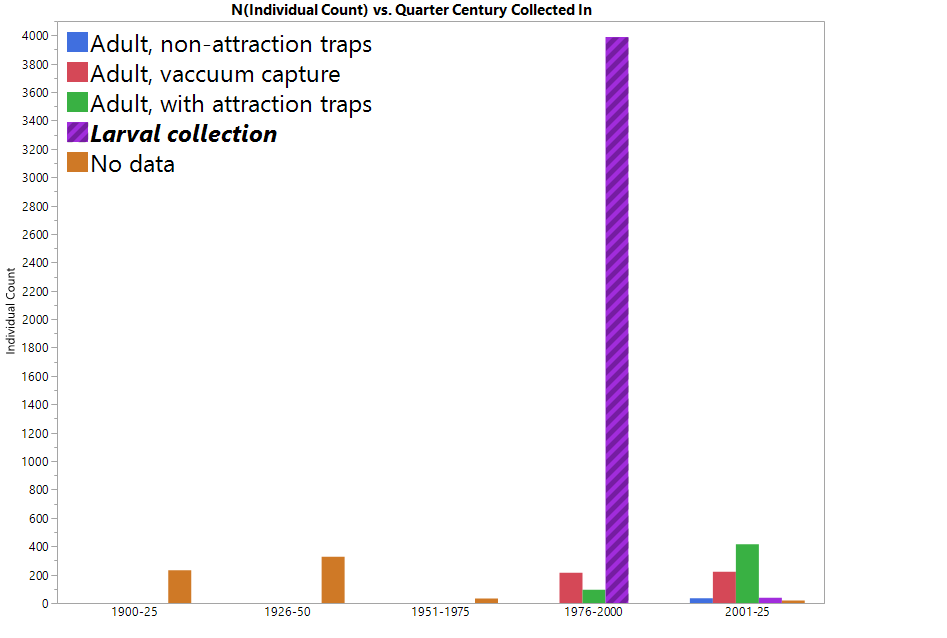


This graph displays the trend line of the data by Quarter Century. As the graph indicates, the number of mosquitoes collected remained similar from 1900 to 1950 before our data exhibit a dramatic decrease from1951-1975, and subsequently seeing a massive boom in the specimen count thereafter. This negative trend for between 1951-1975 symbolizes a lack of collections from this collection and is the reason we chose this type of graph. Post 1976, there is an increase in the amount of mosquitoes we are looking at.

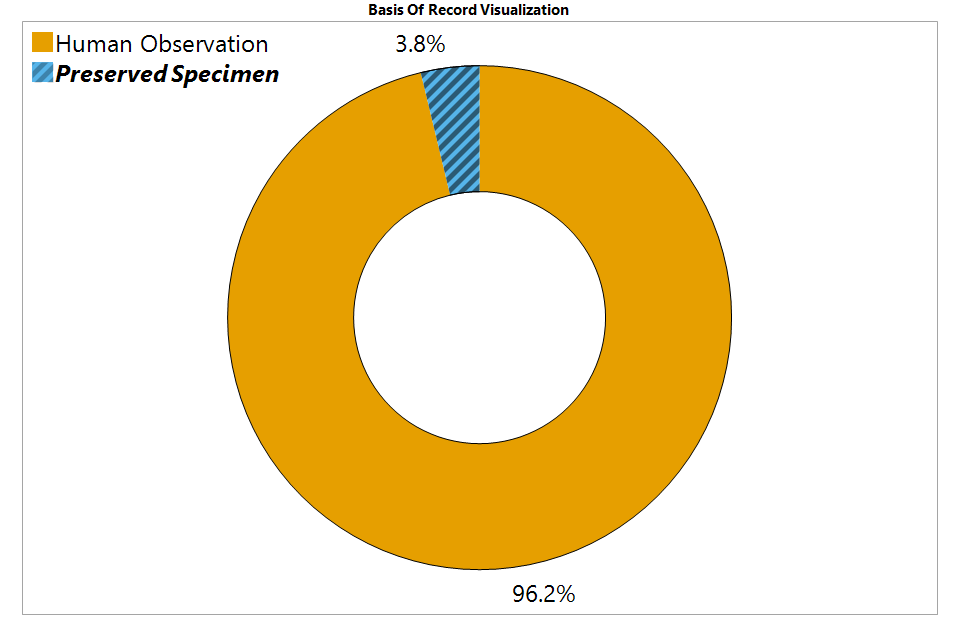


This visualization indicates that the most common method to capture mosquitoes is with attraction traps, followed by vacuum capture, with non-attraction, "no data" and larval collection occupying comparatively small portions of the ring.



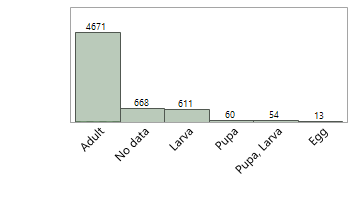
This graph compares the Collecting Method and the Individual Count. We can see from this graph that the Individual Count of Attraction Traps is visually more than double any of the others and potentially all of the other collection methods compared.

This visualization compares individual count, quarter century, and Collecting Method. A massive spike in larval collection from 1976-2000 is very obviously the greatest amount of mosquitoes caught. The earlier years lack a defined Collecting Method, and the the quarter century that we are currently in has the most variety. Also notable is that the attraction traps increased from 1976-2000 to 2001-25.



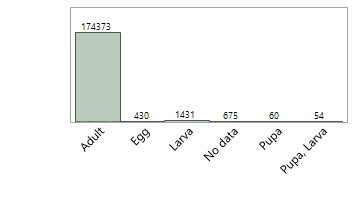
This visualization supports the notion that the overwhelming majority of data came from human observations rather than preserved specimen. This could be due to the relative cost efficiency of observation to preservation, or the integrity of live specimen in comparison to preserved. Preservation obviously comes with the caveat that the preserved being is no longer alive, meaning certain tests and statistics are impossible to determine post-preservation. In addition, this could play a role in the strong preference shown toward human observation. From the data, we are aware of many different Collecting Methods, whereas preservation methods were only discussed in this kind of detail, so the expansive list of Collecting Methods could be a factor in the tendency to observe rather than preserve.

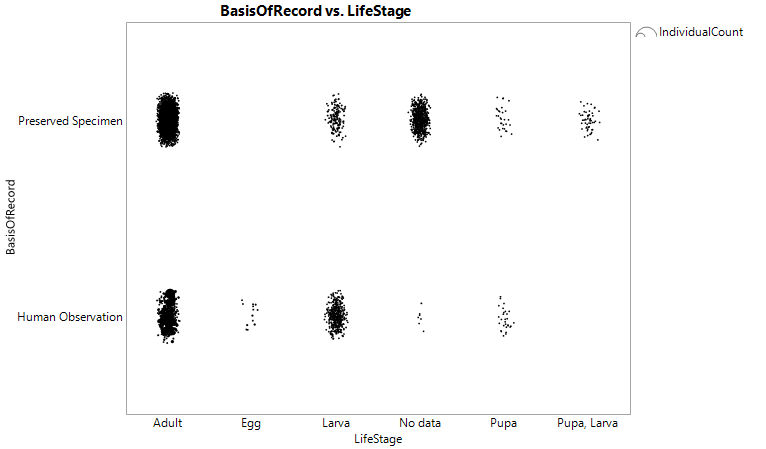
**Distribution of Life Stages of Collections**



This distribution of Life Stage of the collections within the data implies that there are more collections of adults.

**Distribution of Life stages (with Individual Collection)**



This distribution exemplifies that when we factor in the Individual Counts for a distribution of Life Stages, that the majority of the individual mosquito in the data are in the adult phase. 

This visualization compares the Life Stage of the collected mosquitoes with the Basis of Record for the collected mosquitoes. With individual count to size each data point the visualization unveils clusters of data. Using the above visualization to infer what form of Basis of Record will be used depending on the Life Stage of the mosquito becomes possible.

**Conclusion**

From the Individual Count being implemented in several circumstances we have visually shown that the Individual Count variable had an impact on the our Data.

* Where are the data from, geographically speaking? Excluding Antarctica, the data comes from every single continent on the planet.

There is a Visual difference between Years Collected and the Collecting Methods, especially when we use the Individual Count.

In a comparison between Life Stage and Basis of Record the hypotheses are:

H0: Basis of Record and Life Stage are statistically independent.

H1: There is a statistically significant relationship between Basis of Record and Life Stage.

Upon conducting a Pearson’s Chi Square test, there appears to be a statistically significant and worrisomely strong (VCrammer’s = 0.43742) dependence between Basis of Record and the Life Stage (χ25=33870.3, p<0.0001). This worrisomely strong dependence establishes that there is a statistically significant relationship between Basis of Record and Life Stage, allowing us to reject the null hypothesis.

Works Cited

1. Nuwer, R. (2014, April 30). Mosquitoes Kill More Humans Than Human Murderers Do . Retrieved March & April, 2017, from <http://www.smithsonianmag.com/smart-news/mosquitoes-kill-more-humans-human-murderers-do-180951272/>
2. Smithsonian Vector Map. (n.d.). Retrieved April & may, 2017, from http://vectormap.si.edu/

The website was contacted through email and the data was provided at request.

1. Winged Scurge 43' (n.d.). Retrieved March & April, from <https://www.youtube.com/watch?v=y68F8YwLWdg>