Topic of Research:

Exploratory Analysis of Tick-Borne Illness is Dogs

Group 5• 11.03.2022

Denis Antonov ▪ Courtney Barnes ▪ Joseph Bloomfield ▪ Nichelle Francis ▪ ChiChi Ugochukwu

Overview

BRIEF: Our group explored data related to the geographic distribution and prevalence of two tick species which carry pathogens causing illness/disease to the dog population in the United States.

GOALS OF STUDY:

To develop a model that can predict the likelihood of a tick-illness prior to testing through analyzing symptoms.

To ascertain a relationship between the recently recorded new migrations of ticks across the United States, and a higher number of tick-borne illness cases in dogs.

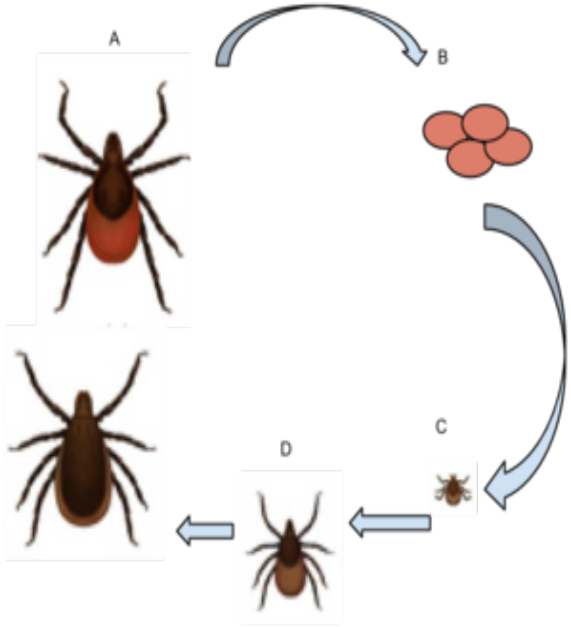
Sources

Data Sources:

Tick data sourced from the CDC.
Pet data sourced from a combination of animal shelter data, various web pages detailing signs/symptoms of tick-borne illness in dogs.

Meet the Ticks

Ixodes Scapularis



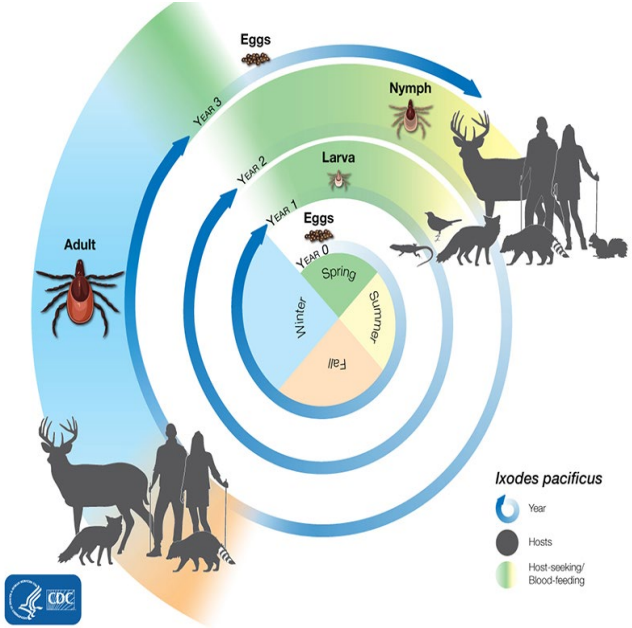
2 year lifecycle

Prevalent in the eastern , southern and midwestern states

Main vector of Lyme disease in North America

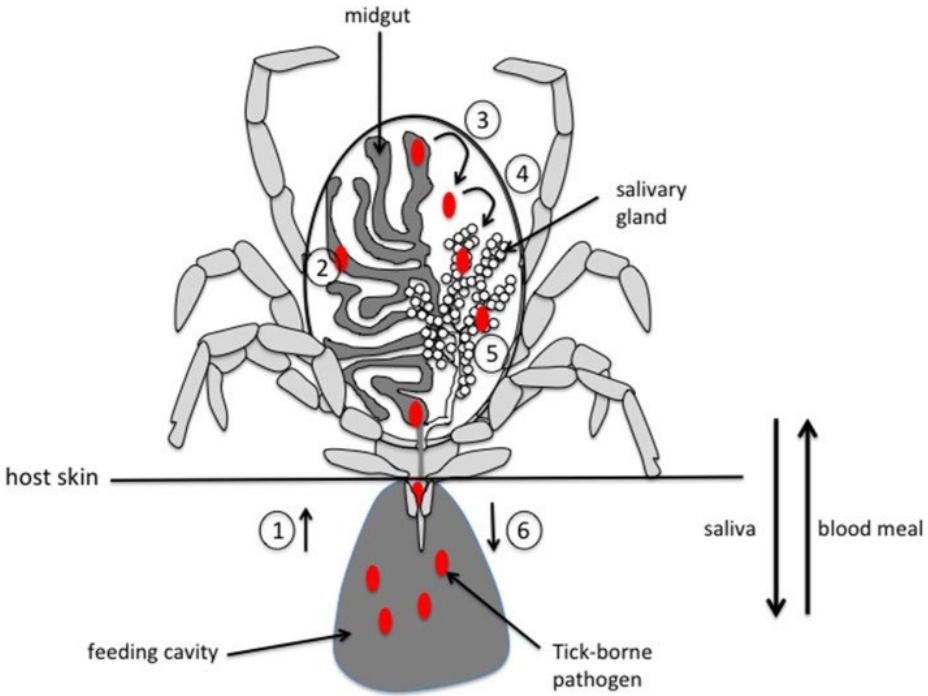
Other Diseases carried: Anaplasmosis, Babesiosis, Borrelia mayonii

Ixodes Pacificus



- 3 year lifecycle
- Prevalent on western coast of the United States
- Diseases carried: Anaplasmosis, Lyme disease

TRANSMISSION



FORM OF TRANSMISSION: The tick attaches to host via host's skin and there is an exchange of saliva and blood. The tick will insert saliva tainted with tick-borne pathogens and in return receive the necessary blood to transition into the next stage of their lifecycle. The tick has to be attached to its host for about 36-48 hours for transmission of bacteria into the host

EFFECTS: In dogs, the effects of tick-borne illness include, lethargy, lameness, fever, joint pain or swelling, and the enlargement of lymph nodes.

Questions To Explore

Has the prevalence of ticks increased around the United States?

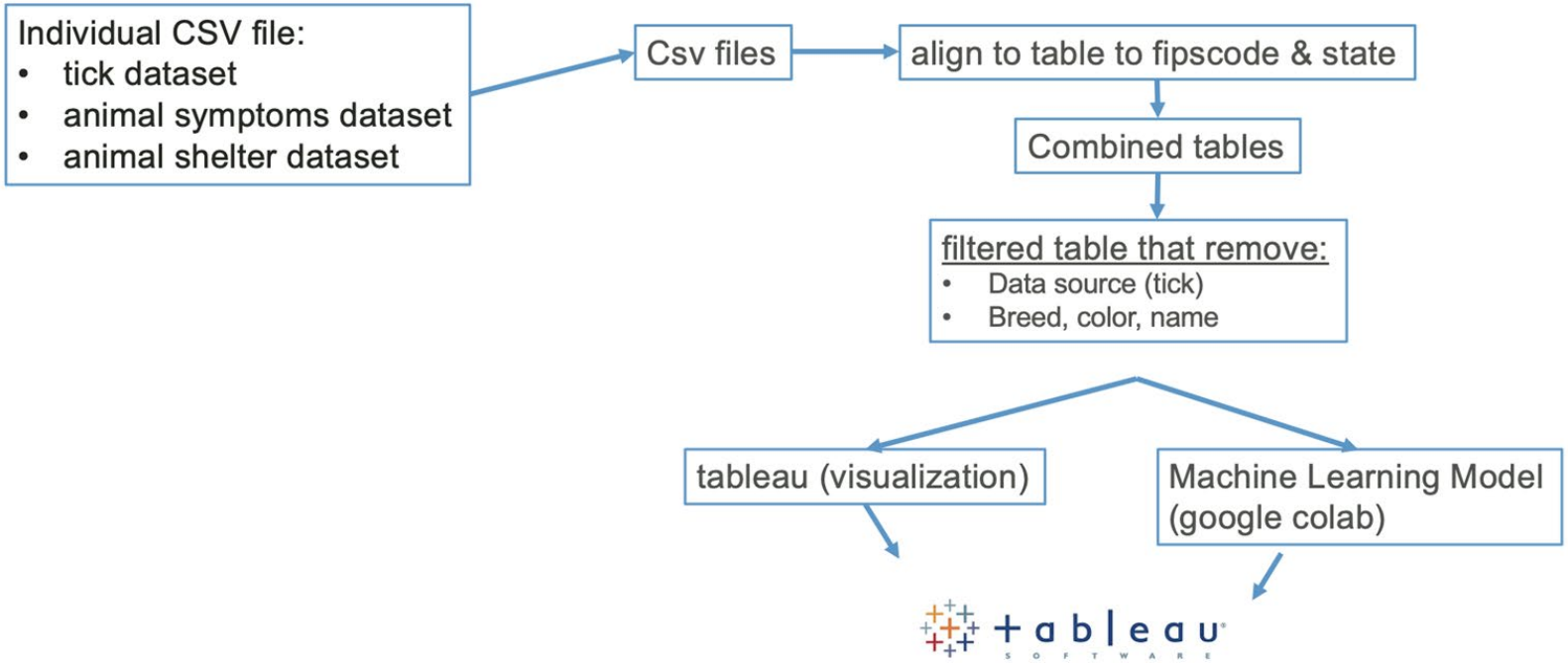
Can Machine-Learning be utilized to effectively predict tick-borne illness diagnosis?

Are certain symptoms better indicators of a potential tick-borne illness?

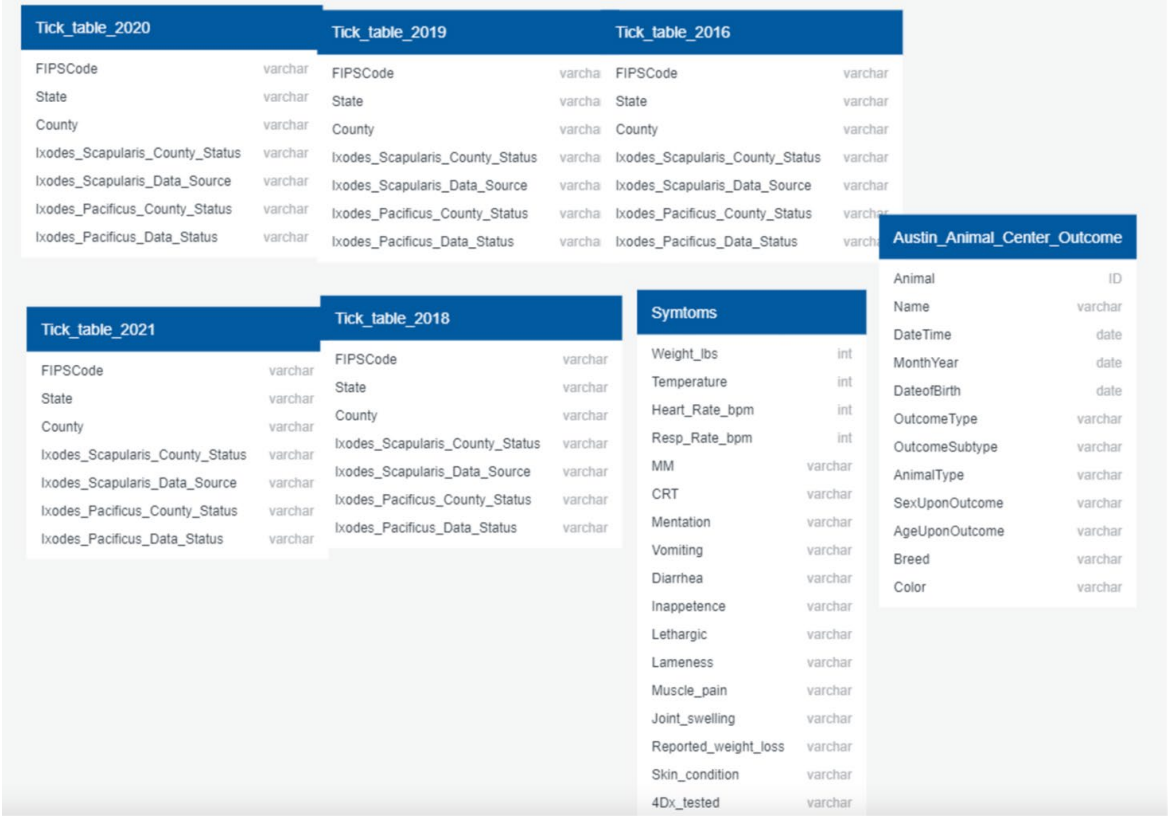
Phases of Project

Database

Database pipeline



Data Exploration Phase



Database Takeaways

Lessons Learned

- Ensuring that the data is cleaned and joined completely before beginning visualizations and machine learning.

Future analysis

- Expand Data selection

Obstacles

- Limited data
- Long query times

Machine Learning

Model Choice

Supervised Learning: Classification

Random Oversampling and SMOTE

Benefit:

Corrects for imbalance in dataset

Limitations:

Increases chance of overfitting in model

Categorical data may skew feature importance

Feature engineering and selection

- **Utilizing a Supervised Model to predict a Negative vs. Positive test.**
- **Features: The veterinary intake data**
- **Target: The test result (i.e. Negative or Positive)**

age	weight	…	vomiting	muscle_pain	Tested
…	…	…	…	…	Negative
…	…	…	…	…	Negative
…	…	…	…	…	Positive
…	…	…	…	…	Negative

Model Results

Naive Random Sampling

Balanced Accuracy Score:
0.965

	Predicted Negative	Predicted Positive
Actual Negative	3559	117
Actual Positive	22	547

	pre	rec	spe	f1	geo	iba	sup
Negative	0.993856	0.968172	0.961336	0.980846	0.964748	0.931374	3676.000000
Positive	0.823795	0.961336	0.968172	0.887267	0.964748	0.930102	569.000000
avg_pre	0.971061	0.971061	0.971061	0.971061	0.971061	0.971061	0.971061
avg_rec	0.967256	0.967256	0.967256	0.967256	0.967256	0.967256	0.967256
avg_spe	0.962252	0.962252	0.962252	0.962252	0.962252	0.962252	0.962252
avg_f1	0.968303	0.968303	0.968303	0.968303	0.968303	0.968303	0.968303
avg_geo	0.964748	0.964748	0.964748	0.964748	0.964748	0.964748	0.964748
avg_iba	0.931204	0.931204	0.931204	0.931204	0.931204	0.931204	0.931204
total_support	4245.000000	4245.000000	4245.000000	4245.000000	4245.000000	4245.000000	4245.000000

SMOTE Oversampling

Balanced Accuracy Score:
0.949

	Predicted Negative	Predicted Positive
Actual Negative	3649	27
Actual Positive	53	516

	pre	rec	spe	f1	geo	iba	sup
Negative	0.985683	0.992655	0.906854	0.989157	0.948785	0.907917	3676.000000
Positive	0.950276	0.906854	0.992655	0.928058	0.948785	0.892470	569.000000
avg_pre	0.980937	0.980937	0.980937	0.980937	0.980937	0.980937	0.980937
avg_rec	0.981154	0.981154	0.981154	0.981154	0.981154	0.981154	0.981154
avg_spe	0.918355	0.918355	0.918355	0.918355	0.918355	0.918355	0.918355
avg_f1	0.980967	0.980967	0.980967	0.980967	0.980967	0.980967	0.980967
avg_geo	0.948785	0.948785	0.948785	0.948785	0.948785	0.948785	0.948785
avg_iba	0.905847	0.905847	0.905847	0.905847	0.905847	0.905847	0.905847
total_support	4245.000000	4245.000000	4245.000000	4245.000000	4245.000000	4245.000000	4245.000000

Model Choice

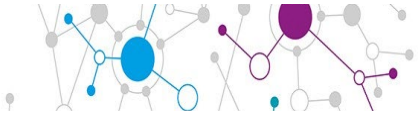
Unsupervised Learning: Clustering

KMeans

Li

Benefit:

Great for exploring trends in data



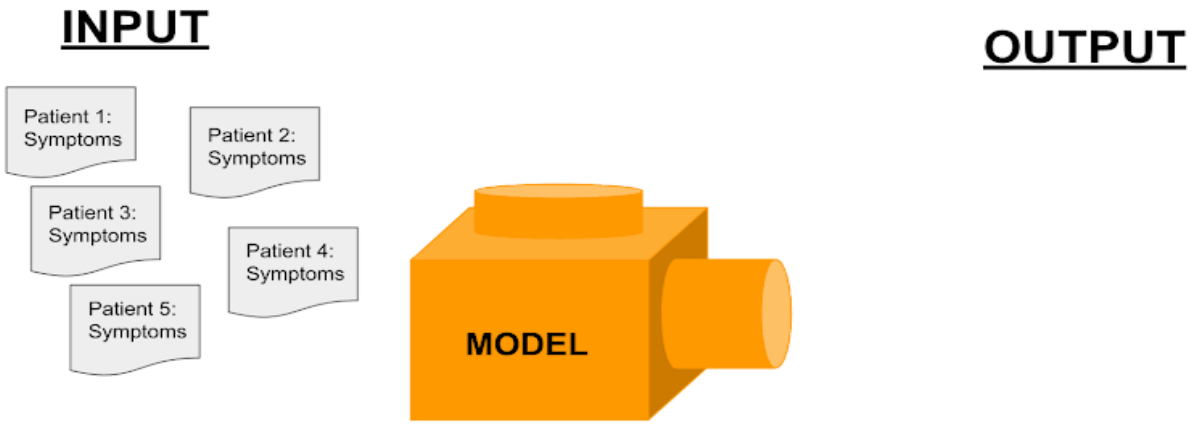
Difficulties in interpreting whether model outcomes are correct/beneficial

Sensitive to scaling

Difficult to predict the number of clusters

Feature engineering and selection

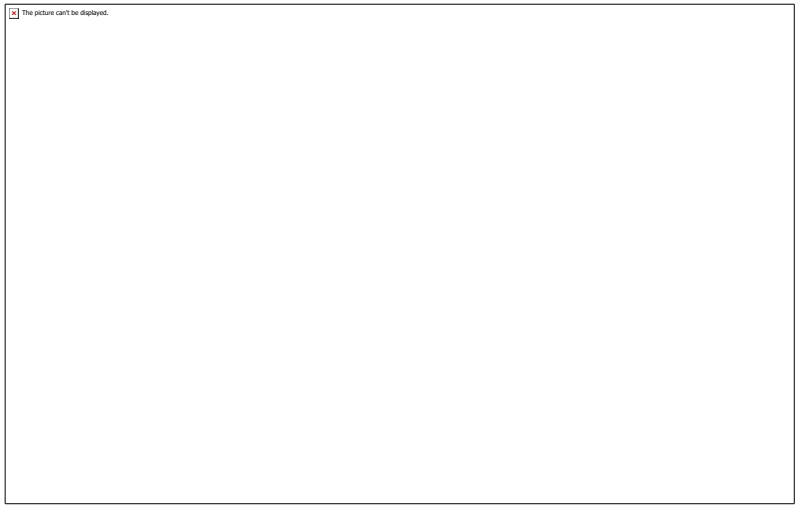
- **Utilizing an Unsupervised Model to predict Negative vs. Positive test.**
- **Features: Veterinary intake data**
- **Clusters: Two**

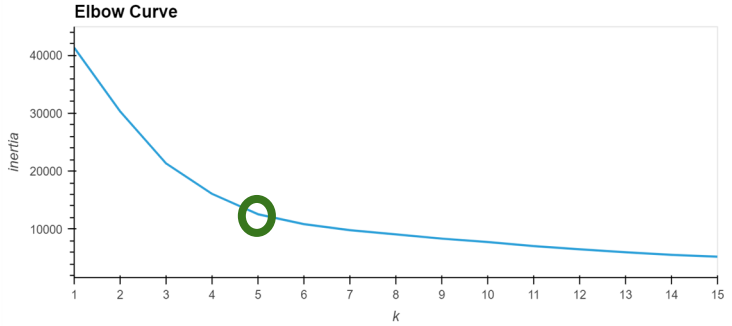


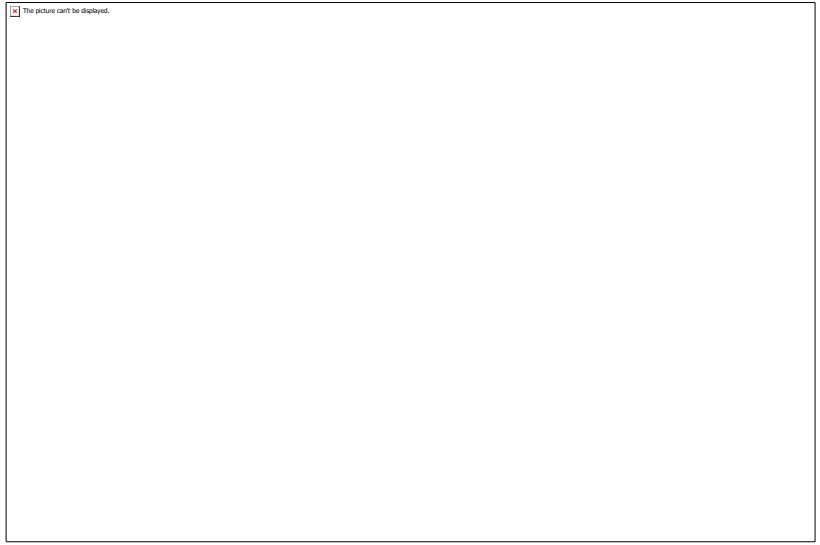
Model Results

Right: Checked elbow curve to see what data indicated as k-value. Plotted predictions with 5 clusters (k=5) on 3D plot.

Below: Predictions with 2 clusters on 3D Plot







Conclusions

Supervised Learning:

- Potentially reliable for predicting diagnoses
  - SMOTE model performed slightly better compared to Random
- High scores could be attributed to lack of variation in data

Unsupervised Learning:

- Results were not very conclusive
- Could potentially be improved by
  - Increasing the number of principal components
  - Having a more varied dataset and reducing the number of features

Statistical Analysis

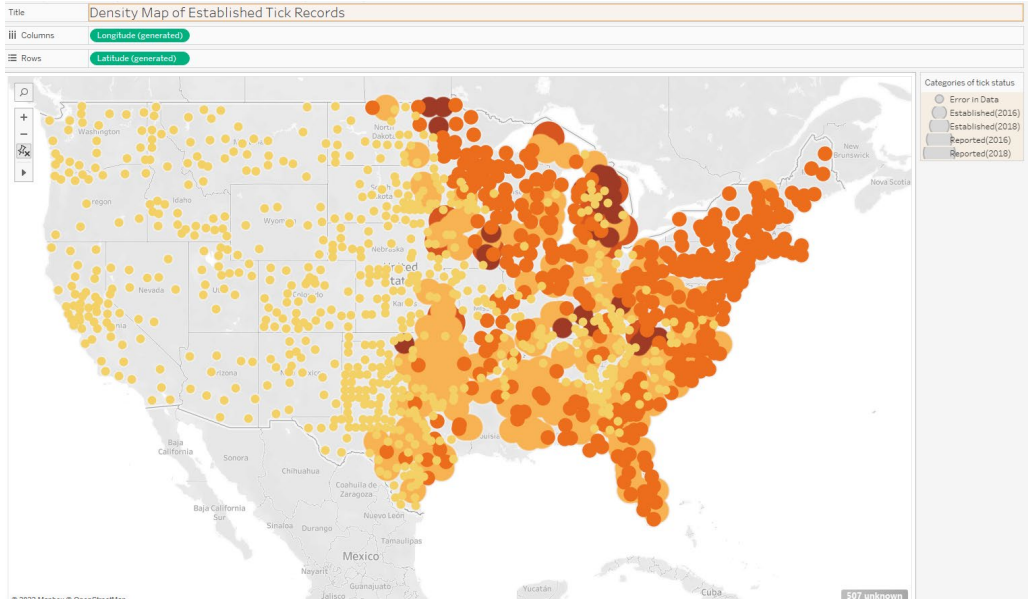
Visualizations

Tableau Integration

- For our visualization portion we used tableau desktop in order to utilize the full features of integration which allowed us to connect to our AWS instance of Postgres and cohesively collaborate and make updates instantly for the team to view
- The integration piece was surprisingly simple when using AWS and after setting up the database tables in Postgres. All team members were able to keep a local instance in Postgres and update as needed.
- To further integrate our work google drive was used to connect with tableau in order to pull data exports directly into tableau to create additional visualizations

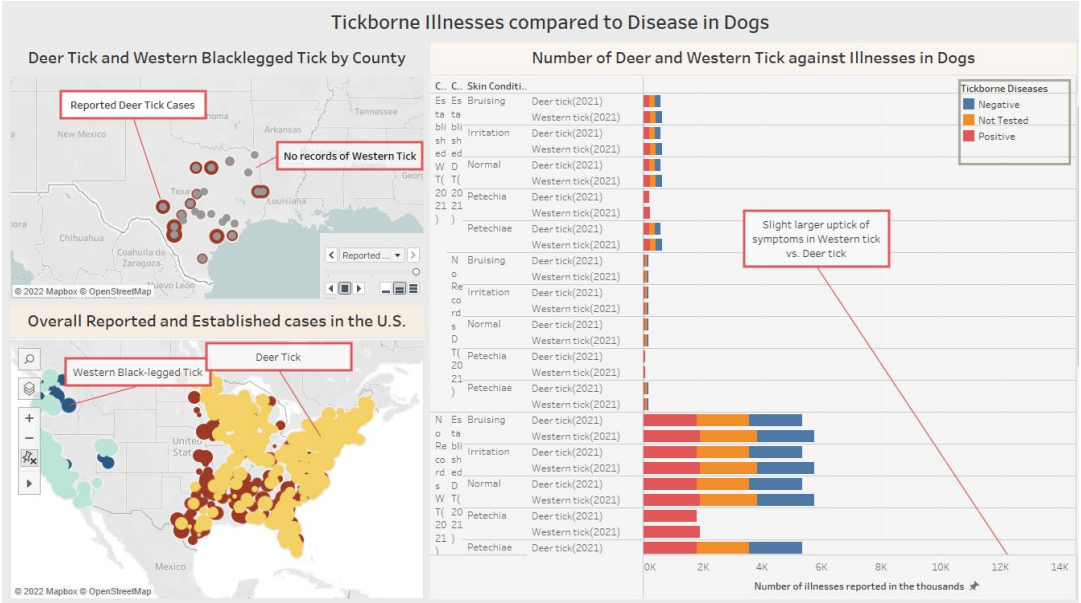
Analysis

- The initial analysis yielded results that seemed promising showing data pulled from the tick data. The below representing deer tick data and showing errors based on data not being present



Analysis

- After creating some calculated fields to combine the tick data, working through the joined data and comparing the data in table the below representation was the result.



Obstacles and lessons learned

- Data cleaning yielded some bad results in some cases
- Tableau public limited use of integration capabilities with Postgres and AWS
- Tableau desktop is not shareable on Tableau cloud without all users having an account.
- Working in tableau can be very slow depending on how much dependencies are involved

Live Dashboard

https://public.tableau.com/app/profile/joseph.bloomfield/viz/Tableaudashboard_16673515630850/TickDashboard#1

Presentation Format

Per section:

- What we did
- “Why” we did it
- “How” we did it/ results
- Obstacles

End of presentation:

- Lessons learned
- Further analysis/ alternate ideas / limitations (per section)