**Presentation (15 points)**

**Content**

The presentation outlines the project, including the following:

* Selected topic

**Analysis of tick prevalence and exploration of model predicted tick-borne illness diagnoses in dogs**

* Reason topic was selected

**In this project, we explore data related to the geographic distribution and prevalence of tick species which carry pathogens that cause illness/disease in the US. We also look into data pertaining to veterinary symptom records and tick-illness testing in order to develop a model that could potentially predict the likelihood of a tick-illness before testing by analyzing the symptoms. The goal of this study is to ascertain whether there is a relationship between the recently recorded new migrations of ticks across the US, and a higher number of tick-borne illness cases in dogs.**

* Description of the source of data

**-Tick data sourced from the CDC; Pet data sourced from a combination of animal shelter data, various webpages detailing signs/symptoms of tick-borne illness in dogs, and some prior knowledge.**

* Questions the team hopes to answer with the data
  + **Has the prevalence of ticks increased around the US?**
  + **Can Machine-Learning be utilized to effectively predict tick-borne illness diagnoses?**
  + **Are any symptoms in particular better indicators for potential tick-borne illness?**
* Description of the data exploration phase of the project
  + **Explore data pertaining to tick prevalence in the different regions of the US and how that has changed over the years. Looking at which species carry which pathogens (which in-turn causes particular illnesses).** 
    - **Sample Hypothesis: The changes in tick species migration effects the prevalence of certain tickborne illnesses in areas where they were previously uncommon – could lead to lack of treatment options to target the different illnesses (i.e. increased mortality rate)**
  + **Explore data pertaining to tick-borne illnesses in dogs – looking particularly at the signs and symptoms. Also, looking at potential treatment options.**
* Description of the analysis phase of the project
  + **Can run statistical analysis on the symptoms and how they relate to the the illness (essentially t-testing & chi2 testing to see if certain symptoms are statistically significant in predicting tick-borne illness)**
  + **Possible analysis of tick migration (Not sure yet how..)**

**Slides**

Presentations are drafted in Google Slides.

**GitHub Repository (10 points)**

**Main Branch**

All code in the main branch is production-ready.

The main branch should include:

* All code necessary to perform exploratory analysis
* Some code necessary to complete the machine learning portion of project

**README.md**

README.md should include:

* Description of the communication protocols
* Outline of the project (this may include images, but they should be easy to follow and digest)

**Individual Branches**

Requirements for the individual branches follow:

* At least one branch for each team member
* Each team member has at least four commits for the duration of the second segment (eight total commits per person)

**Machine Learning Model (30 points)**

The team members are expected to submit the code for the machine learning model, as well as the following:

* Description of preliminary data preprocessing
* Description of preliminary feature engineering and preliminary feature selection, including the decision-making process

**(Potentially: )**

**Using Supervised model to predict Negative vs. Positive test – model will take in information for each of the symptoms in order to predict the test outcome. This will be using data set that only includes dogs that *were* tested**

**Using Unsupervised model to predict Negative vs. Positive test – this model will use the information from animals that *were not* tested to group them into “Possibly infected” and “Possibly uninfected”.**

* Description of how data was split into training and testing sets

**For Supervised -**

* Explanation of model choice, including limitations and benefits

**Limitations: Both include much categorical data – this could make the predictions weak because the discernment of the symptoms is mostly subjective**

* + **Theoretical future study to correct for this: including more discrete data – possibly in the form of bloodwork values (i.e. chemistry panels, complete blood count (CBC), etc) *[In vet practices, these tests would almost always be run along with the 4Dx test so the information should be readily available, just inaccessible within the scope of this project.]***

**Database Integration (30 points)**

The team members are expected to present a fully integrated database, including the following:

* Database stores static data for use during the project
* Database interfaces with the project in some format (e.g., scraping updates the database)
* Includes at least two tables (or collections, if using MongoDB)
* Includes at least one join using the database language (not including any joins in Pandas)
* Includes at least one connection string (using SQLAlchemy or PyMongo)

**IMPORTANT**

If you use a SQL database, you must provide your ERD with relationships.

**Dashboard (15 points)**

A blueprint for the dashboard is created and includes all of the following:

* Storyboard on a Google Slide(s)
* Description of the tool(s) that will be used to create the final dashboard
* Description of interactive element(s)