Project Title: Big Data Derby
Group Members: Kajal Tiwary, Clare Garberg, Clara Richter, Elise Rust, Yifan Liu (Group #4)

Project Goal & Objective Summary: This project attempts to analyze horse data from the 2019 derby to inform competition strategy and improve horse health. This research is in direct response to the Big Data Derby 2022 Kaggle competition that intends to help owners, trainers, and veterinarians improve equine welfare. This analysis leverages a series of regression techniques to answer the following questions in relation to assessing competition strategy: *(1) Can we accurately predict a horse's finish position and what factors are most influential in determining this? (2) What factors are most deterministic in predicting the odds of a horse to win a race? (3) What factors are most influential in predicting the amount of money put into a race? (4) What factors determine the amount of time it takes a horse to finish a race? (5) Does the track type, course type, race type, or track condition play a role in determining horse performance or health? (6) What role does the geography and position of the horse in a race play in its outcome?*

Proposed Data Source & Methods: The data source of this project is the Big Data Derby 2022 Kaggle competition. The record dataset consists of three primary components: horse/jockey data, racetrack data, and race tracking data. It contains detailed information about the horse and jockey and specific races. The dataset also includes the geographic coordinates of the horse in a race taken every 0.25 seconds. As the project progresses, other data sources will be considered. Our analysis may include jockey and horse history statistics, scraped from EQUIBASE, where transformation and embedding techniques are used to incorporate past race performance. We may also include weather data for days races took place to identify conditions that impact horse performance. This will be done using the latitude, longitude, date, and time variables provided in the dataset and a global weather API.

We are considering experimenting with a handful of non-neural and neural network frameworks, leveraging automated hyperparameter tuning, and comparing the prediction accuracy across them to select an optimal model. The feed-forward network will act as a base case to compare against as it is a popular and common artificial neural network (ANN) for regression problems. We are also considering recurrent neural networks (RNNs) which are powerful at predicting time series data due to their ability to handle sequential tasks, as well as long short term memory networks (LSTMs) which address the vanishing gradient problem found in FFNs and RNNs. Traditional machine learning approaches, such as Random Forest and Ridge and Lasso Regression, will also be used to assess performance against neural networks and to extract important features. As the semester progresses and more types of neural networks are explored in class, this list may shift to include models that better fit regression and classification of tabular/text data tasks.

Expected Results Or Outcomes:

The end goal of this exploration is to successfully model and predict a race horse's time in the race, their position, their odds of winning (based on fan sentiment), and the amount of money bet on the horse. After the model is well-trained, we can apply it to predict the potential profit of betting on certain horses based on the predicted likelihood of winning and the amount of money invested. We aim to look at how these dependent variables are influenced by a variety of factors including the climate and weather events occurring during the time of the race. Also of interest is whether this metric relies solely on a horse's own attributes or whether it is influenced by the position of its competitors. The insights will be derived primarily from neural network models. This report will also illustrate the differences in the performance of neural networks and standard regression modeling techniques. We expect to attain accurate neural network models to predict each one of the metrics mentioned above and that geography, weight carried, weather, and race type will be significant influential variables.