Goal of the Competition

The goal of this competition is to analyze horse racing tactics, drafting strategies, and path efficiency. You will develop a model using never-before-released coordinate data along with basic race information.

Your work will help racing horse owners, trainers, and veterinarians better understand how equine performance and welfare fit together. With better data analysis, equine welfare could significantly improve.

Context

Injury prevention is a critical component in modern athletics. Sports that involve animals, such as horse racing, are no different than human sport. Typically, efficiency in movement correlates to both improvements in performance and injury prevention.

A wealth of data is now collected, including measures for heart rate, EKG, longitudinal movement, dorsal/ventral movement, medial/lateral deviation, total power and total landing vibration. Your data science skills and analysis are needed to decipher what makes the most positive impact.

In this competition, you will create a model to interpret one aspect of this new data. You’ll be among the first to access X/Y coordinate mapping of horses during races. Using the data, you might analyze jockey decision making, compare race surfaces, or measure the relative importance of drafting. With considerable data, contestants can flex their creativity problem solving skills.

The New York Racing Association (NYRA) and the New York Thoroughbred Horsemen's Association (NYTHA) conduct world class thoroughbred racing at Aqueduct Racetrack, Belmont Park and Saratoga Race Course.

With your help, NYRA and NYTHA will better understand their vast data set, which could lead to new ways of racing and training in a highly traditional industry. With improved use of horse tracking data, you could help improve equine welfare, performance and rider decision making.

Our sport is currently investing significant money in collecting far more precise tracking data in the hopes of improving equine welfare. Along with stride data, we can now collect measures for heart rate, EKG, longitudinal movement, dorsal/ventral movement, medial/lateral deviation, total power and total landing vibration. However, we do not have analysts with the appropriate expertise to help decipher these data sets.

We hope this competition allows us to interact with data scientists to help find solutions to equine safety issues as well as develop a roster of academics and motivated hobbyists who lead us in analyzing the coming generations of data.

## **Overview**

Your challenge is to generate actionable, practical, and novel insights from horse tracking data that devises innovative and data-driven approaches to analyzing racing tactics, drafting strategies and path efficiency. There are several potential topics for participants to analyze.

These include, but are not limited to:

* Create a horse rating measuring expected finish position versus actual finish position. How does a horse’s expected finish position change through the running of a race? Does this metric rely solely on a horse’s own position or is it influenced by the position of competitors?
* What are optimal racing strategies? Considering different venues, surfaces and race distances. Create a jockey rating based upon path efficiency?
* Create a surface measure model which would rate the fairness of different paths on a racecourse that may be beneficial or harmful to finish position based. This may be a result of unknown barometric, weather or maintenance factors.
* Create a model measuring the existence (or not) and relevance of a drafting benefit.
* Create a model reveal optimal gait patterns. Does the model differ for such factors as age, distance, race section or surface?

Contestants should not feel limited to these suggestions.

The above list is not comprehensive, nor is it meant to be a guide for participants to cover.

Submissions that examine one idea more thoroughly are preferred versus those that examine several ideas somewhat thoroughly.

## **Scoring**

An entry to the competition consists of a Notebook submission that is evaluated on the following five components, where 0 is the low score and 100 is the high score. Submissions will be judged based on how well they address:

**Innovation (25 Points Total)**

* Is this a novel way of looking at tracking data? (10 Pts)
* Are the statistical/machine learning approaches using the most up-to-date standards? (5 Pts)
* Will the conclusions challenge the status quo of horse racing methods? (10 Pts)

**Relevance (30 Points Total)**

* Can the conclusions influence equine welfare, equine performance or rider decision making? (10 Points)
* Can the conclusions be the basis of future research on future (larger, more granular) data sets? (10 Points)
* Are the conclusions something that horse racing participants (e.g. owners, trainers, veterinarians) can understand, digest and debate? (10 Points)

**Competence (25 Points Total)**

* Given the data, are the statistical models appropriate? (5 Points)
* Are the conclusions supported by the data? (10 Points)
* Is the analysis accurate? (10 Points)

**Presentation (20 Points Total)**

* Is the writing clear and free of nomenclature? (5 Points)
* Are the charts and tables provided interesting, visually appealing, and accurate? (5 Points)
* Can the analysis thread be followed throughout the presentation? (10 Points)

## **Notebook requirements**

All notebooks submitted must be made public on or before the submission deadline to be eligible. If submitting as a team, all team members must be listed as collaborators on all notebooks submitted.

Columns

nyra\_start\_table.csv

* track\_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
* race\_date - date the race took place. YYYY-MM-DD.
* race\_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
* program\_number - Program number of the horse in the race passed as 3 characters. Should remain 3 characters as it isn't limited to just numbers. Is essentially the unique identifier of the horse in the race.
* weight\_carried - An integer of the weight carried by the horse in the race.
* jockey - Name of the jockey on the horse in the race. 50 character max.
* odds - Odds to win the race passed as an integer. Divide by 100 to derive the odds to 1. Example - 1280 would be 12.8-1.
* position\_at\_finish - An integer of the horse's finishing position. (added to the dataset 9/8/22)

nyra\_race\_table.csv

* track\_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
* race\_date - date the race took place. YYYY-MM-DD.
* race\_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
* distance\_id - Distance of the race in furlongs passed as an integer. Example - 600 would be 6 furlongs.
* course\_type - The course the race was run over passed as one character. M - Hurdle, D - Dirt, O - Outer turf, I - Inner turf, T - turf.
* track\_condition - The condition of the course the race was run on passed as three characters. YL - Yielding, FM - Firm, SY - Sloppy, GD - Good, FT - Fast, MY - Muddy, SF - Soft.
* run\_up\_distance - Distance in feet of the gate to the start of the race passed as an integer.
* race\_type - The classification of the race passed as as five characters. STK - Stakes, WCL - Waiver Claiming, WMC - Waiver Maiden Claiming, SST - Starter Stakes, SHP - Starter Handicap, CLM - Claiming, STR - Starter Allowance, AOC - Allowance Optionl Claimer, SOC - Starter Optional Claimer, MCL - Maiden Claiming, ALW - Allowance, MSW - Maiden Special Weight.
* purse - Purse in US dollars of the race passed as an money with two decimal places.
* post\_time - Time of day the race began passed as 5 character. Example - 01220 would be 12:20.

nyra\_tracking\_table.csv

* track\_id - 3 character id for the track the race took place at. AQU -Aqueduct, BEL - Belmont, SAR - Saratoga.
* race\_date - date the race took place. YYYY-MM-DD.
* race\_number - Number of the race. Passed as 3 characters but can be cast or converted to int for this data set.
* program\_number - Program number of the horse in the race passed as 3 characters. Should remain 3 characters as it isn't limited to just numbers. Is essentially the unique identifier of the horse in the race.
* trakus\_index - The common collection of point of the lat / long of the horse in the race passed as an integer. From what we can tell, it's collected every 0.25 seconds.
* latitude - The latitude of the horse in the race passed as a float.
* longitude - The longitude of the horse in the race passed as a float.

nyra\_2019\_complete.csv - This file is the combined 3 files into one table. The keys to join them trakus with race - track\_id, race\_date, race\_number. To join trakus with start - track\_id, race\_date, race\_number, program\_number.

* track\_id - char(3)
* race\_date - date
* race\_number - char(3)
* program\_number - char(3)
* trakus\_index - int
* latitude - float
* longitude - float
* distance\_id - int
* course\_type - char(1)
* track\_condition - char(3)
* run\_up\_distance - int
* race\_type - char(5)
* post\_time - char(5)
* weight\_carried - int
* jockey - char(50)
* odds - int
* position\_at\_finish - An integer of the horse's finishing position. (added to the dataset 9/8/22)