**Background**:

I am pursuing a Master’s in Data Science as a precursor to starting a business that utilizes data science to pick horses. Accordingly, my project proposal seeks to advance this objective.

**Question I’m Trying To Answer**:

The distance a horse runs in a race is a key determinant in the horse’s ultimate finishing position. My proposed analysis/project is to look at race data at New York horse racing venues (Belmont, Aqueduct & Saratoga) to determine what insights historic races can provide toward answering the question how far a horse will run in an upcoming race. Such information would provide an important edge to a pari-mutuel investor.

Key factors that determine distance run by a horse in a race include:

* Race Distance
* Track configuration
* Number of turns in race
* Post position
* Running style of horse (front runner, mid-pack runner or closer)
* Number of Horses in the race
* Running styles of horses in the race
* Track condition

My analysis will look at some or all these factors to determine if this data set can provide the competitive edge contemplated.

**Data:**

My proposed dataset will be obtained from Trakus, a horse racing information and multi-media provider. This information is freely available at the Trakus site or at numerous horse racing venue website that provide Trakus data to their patrons.

Key benefits and weaknesses of the dataset follow:

Benefits:

* The data is uniquely able to answer my question
* The Data is Free and publicly available
* Potential to utilize gganimate

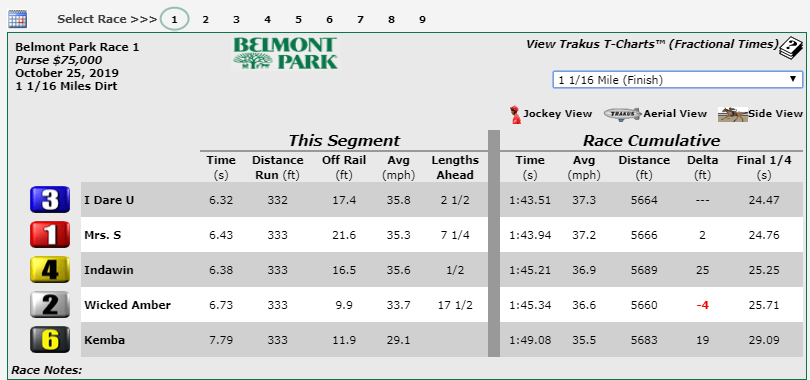
Weakness:

* Assembling the data will require manual entry or writing a screen scraper script(s) and time
* Additional data wrangling will be required
* Coming Up with inspiring visualizations could be challenging
* I believe data will require significant wrangling

The Trakus data can be located at the New York Racing Association’s (NYRA) website in the Trakus section:

<https://www.nyra.com/belmont/racing/trakus>

Sample Data



**Legend**



**Deliverables:**

The data and my analysis will ultimately determine the output of the project. Anticipated deliverables include:

1. Visualizations of Races/Distance Run
2. Box Plots Showing Distribution of Distance Run
3. Par Charts that provide a measure or measures of distance run given various inputs
4. Write-up of key findings and next steps