

Data Analyst Challenge – RAC

# Objective

To create a challenge where you demonstrate your abilities to analyze data, find patterns, generate insights and help the business by detecting data needs but also benefits stakeholders might not even know about.

# The problem and expectations

You will analyze information about horse races. The premise is as follows:

*Horse races are just like any human sport. There are strategies beyond ‘just running fast’. The health of the horses is a priority. How can humans (trainers, vets) help horses to explode their full potential?*

As you detected, the premise is quite general. There’s no simple requirement as “create a nice dash” or something similar. Even though the objective is as generic as it is, as a Data Analyst you must take the data, help with strategies to **improve the performance of the horses and propose solutions** for problems trainers/vets might not even know about.

However, as a DA, you are also expected to **make questions** and solve them yourself. You’re expected to show both curiosity and proactivity by investigating data, finding insights and proposing strategies for stakeholders to get to the final objective.

Sky’s the limit. Imagine the possibilities of providing a way to understanding data (w/no technical complications for stakeholders) and revolution this world.

Some suggested topics to dig around are:

* How do positions change through the race? What factors are related to these changes?
* What are the best strategies to follow? Are these significant or purely coincidental?
* What are your recommendations for trainers? What care measures must be implemented?

Remember: Do not lock yourself with these suggestions. You’re free to provide recommendations always they are data driven. Visualizations are also encouraged to generate since you will be presenting this to non-technical stakeholders. Technical understanding and statistical validation are also encouraged.

# The data

You will receive three tables that might be joined. Tables and their fields are:

* **nyra\_start\_table.csv** - horse/jockey race data
  + 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** - racetrack race data
  + 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** - tracking data
  + 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.

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.

# Presentation

You’re expected to present this in an hour (with 15+ minutes for questioning from the audience). Your audience will be composed of Data Scientist, Operational executives and Finance Experts. Be prepared to receive questions about your methodology but also to make recommendations. Justify any recommendations you provide.

To complete this, you will have one week.

Any questions feel free to reach:

* Raúl Romero, DS & Risk Director ([raul.romero@rentacenter.com](mailto:raul.romero@rentacenter.com)).
* Sara Pérez, HR Manager ([sara.perez@acima.com](mailto:sara.perez@acima.com)).