Clustering Race Horse Movement Profiles to Discover Trends in Injured Horses

Sara Colando[1], Jonathan Pipping[2], Kris Wilson[3]

[1]*Pomona College,* [2]*University of Florida,* [3]*North Carolina State University*

EXECUTIVE SUMMARY

BACKGROUND:

Between 2009 and 2021, over 7,200 horses died or were euthanized due to racing-related injuries (Fobar, 2023). Several horse racing associations in the United States, such as the New York Racing Association (NYRA), are invested in lowering the prevalence of race-related injuries in horses by understanding what increases injury risk (*NYRA Safety*, 2020).

GOALS:

1. Identify horses who under-raced between 2019 and 2021
2. Cluster movement profiles for horses who raced in New York in 2019
3. Discover whether certain movement clusters are more associated with injured horses

KEY FINDINGS:

Through residual analysis, PCA, and model-based clustering, we discovered that at least **251 horses** under-raced between 2019 and 2021, and the horse's speed profile is ***most*** associated with injury status, relative to cumulative lateral movement and strain profiles.

METHODS

IDENTIFYING UNDER-RACING HORSES:

Data Sources:

2019-2022 NYRA Start Lists + 2019-2020 NYRA Severe Horse Injuries →

3678 rows x 6 columns

horse_id	age_year	race_year	n_races	if_injury_reported	fatal
350	4	2019	4	TRUE	No
350	5	2019	8	TRUE	No
350	6	2020	3	TRUE	No
350	6	2021	1	TRUE	No
358	3	2019	2	TRUE	No

Data Limitations:

- Only have birth date information for 20% of horses in the NYRA tracking data
  - Likely underestimating the true number of horses who under-raced between 2019 and 2021

Expected Race Count Model:

Negative Binomial Model

- Controlled for Age of the Horse (in years)
- Controlled for the Calendar Year of the Races (as a factor)

$$log(\mathbb{E}(\text{Race Count})) = \beta_0 + \beta_1(\text{Horse Age}) + \beta_2(\text{Race Year} = 2020) + \beta_3(\text{Race Year} = 2021)$$

Residual Analysis Pipeline:

Fit a Negative Binomial Model to Combinations of Horse Age and Calendar Year in the NYRA Start Lists

↓

Calculate the Standardized Residual for Each Combination of Horse Age and Calendar Year

↓

Classify Horses as Under-Racing if Any of their Age and Calendar Year Combinations have a Standardized Residual Below -1

METHODS CONTINUED

CALCULATING STRAIN RATE:

Definition:

- The ratio of the approach velocity and the distance between any two horses in a race
  - Measured in inverse seconds

Data Source:

2019 NYRA Tracking Data →

Observations	Columns
4,965,717	41

Data Limitations:

- Tracking data is reported at 4 frames / second
  - Instantaneous approach velocities estimated by frame

CLUSTERING HORSE MOVEMENT PROFILES:

Data Source:

2019 NYRA Tracking Data →

Example row (partial)

race number	horse name	cv speed	cluster
4	Starry Rose	0.114	8

Principal Component Analysis:

- Performed on summary statistics of each trajectory: speed, acceleration, lateral movement, and strain rate
- Selected most significant principal components (explain 90% of variance)

Model-Based Clustering:

- Identified main contributor to each principal component
- Used a Gaussian mixture model to cluster horses based on significant summary statistics for each trajectory

HORSES WHO UNDER-RACED

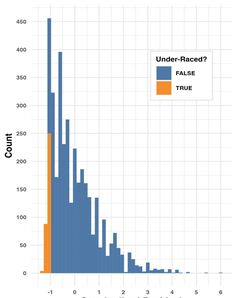


Figure 1: Standardized Residual Distribution

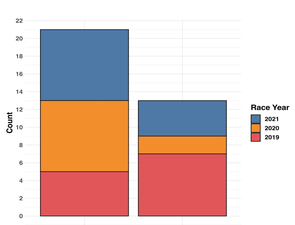


Figure 2: Racing More vs Less than Expected for Horses Severely Injured in 2019

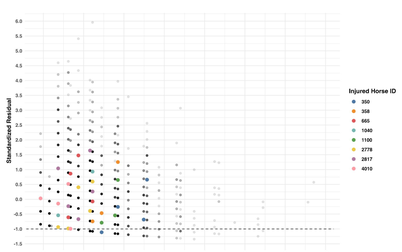
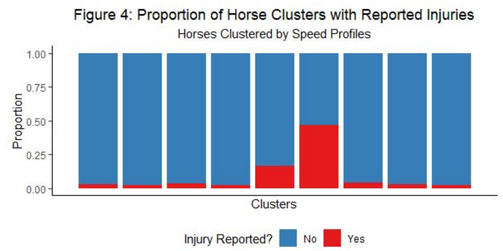


Figure 3: Standardized Residuals versus Fitted Values for the Expected Race Count Model

Key Findings

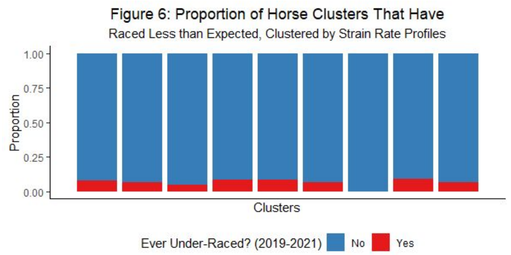
- **27%** (251/931) of horses under-raced from 2019 to 2021.
- **29.9%** (75/251) of under-racing horses under-raced more than once from 2019-2021.
- **One** horse severely injured in 2019 under-raced (and it was in 2021).
- Horses severely injured in 2019 tended to race **more than expected in 2019** and **less than expected in 2020 and 2021**.

MOVEMENT PROFILES CLUSTERING



Cluster	1	2	3	4	5	6	7
Number of Horses	539	362	463	290	147	702	254
Percentage of Horse Injuries	3.7%	3.7%	2.7%	1.0%	4.3%	3.2%	1.6%

Figure 5: Proportions of Horses who Suffered an Injury, Clustered by Lateral Movement Profiles



Key Findings

- When clustering by the speed summary statistics, ***Clusters 5 and 6*** seem to have more injuries reported. One potential explanation is that there are fewer horses in those clusters → greater relative proportion of injures.
- When clustering by lateral movement summary statistics, ***the percentage of horse injuries is mostly consistent across clusters.*** This may be due to less variation in the number of horses per cluster.
- When clustering by strain rate summary statistics, the proportion of horses who raced less than expected (under-racing) is consistent across clusters. This implies that strain rate may not be as indicative of horse availability.

CONCLUSIONS

- There is a significant proportion of horses that raced less than they were expected to, given their age. This, in addition to how they race, may provide some insight into potential patterns and factors leading to horse injures.
- Understanding the movement patterns that result in horses being susceptible to injury allows for our external partners at the NYRA to better manage their horses.

FUTURE DIRECTIONS:

- One of the data limitations is that the lateral movement variable only exists for 1600 meter races. Expanding our data set to all distances would provide more information about lateral movement and its relationship to injury risk.
- A multilevel model incorporating the distribution of horses as a random effect would capture some of the variation in the data without losing information.

ACKNOWLEDGEMENTS AND REFERENCES

We thank Dr. Ron Yurko and Quang Nguyen for all their guidance and encouragement during this project. Additionally, we are thankful to our external partners at NYRA, particularly Joe Appelbaum and Davis Klein, for offering us domain-specific advice. We are also thankful to Brendan Kumagai and his Big Data Derby 2022 team for sharing their data cleaning code and processed data set. Finally, we would like to express our gratitude to Meg Ellingwood, Shamindra Shrotriya, and the rest of SURE 2023 for the opportunity to conduct sports analytics research this summer.

REFERENCES:

Fobar, R. (2023). *Why horse racing is so dangerous*. National Geographic. https://www.nationalgeographic.com/animals/article/horse-racing-risks-deaths-sport

NYRA safety. (2020). New York Racing Association (NYRA). https://www.nyrainc.com/about/nyra-safety