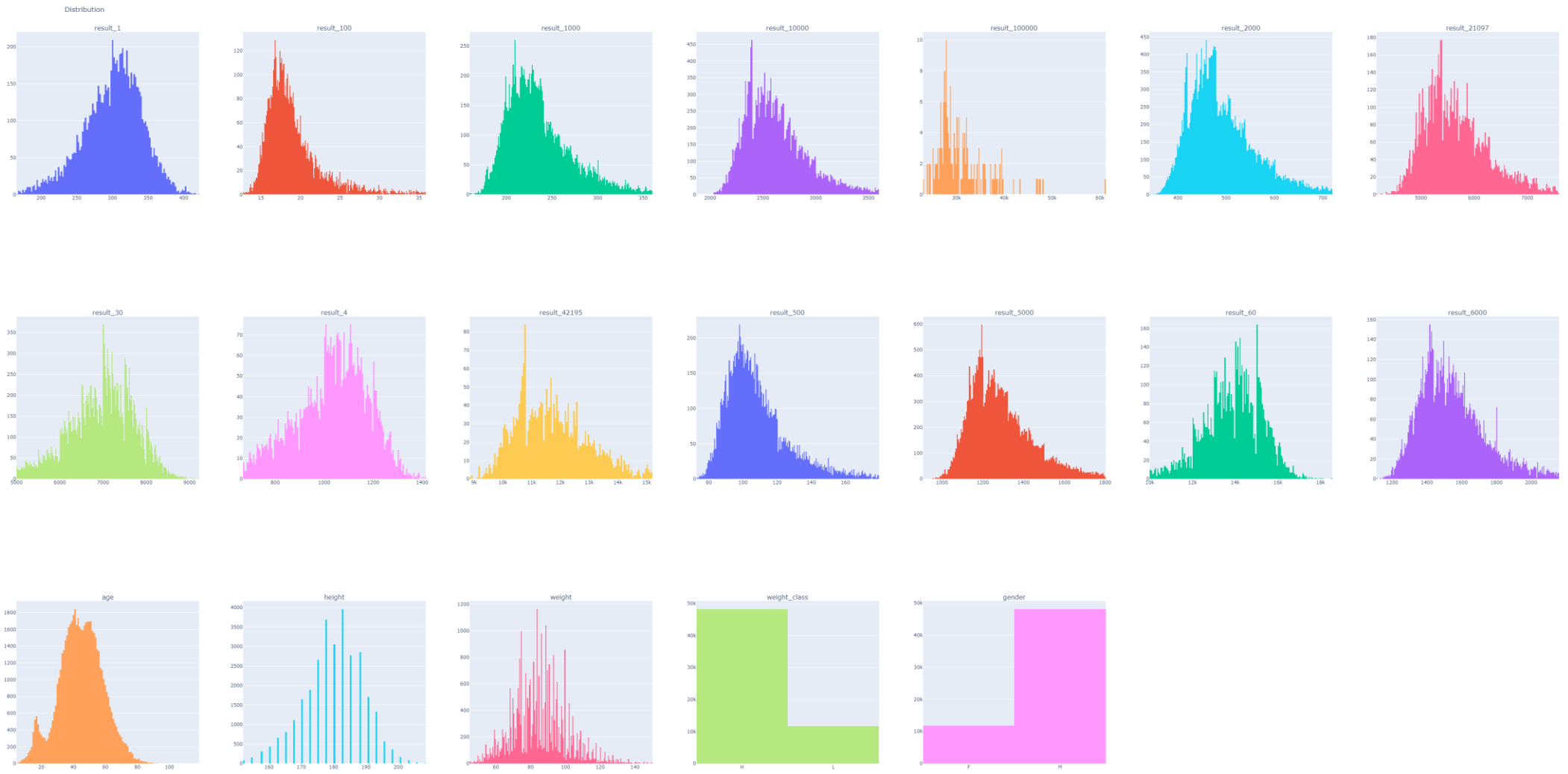
# Indoor Rowing Pace Prediction

# Indoor Rowing Pace Prediction

- Goal: Predict 2000m rowing result on a Concept2 Rowing Machine
  - Standard test and competition distance in rowing
  - Data scraped from Concept2 Logbook
- Using other test results
  - No more than two
  - Shorter than 2km
- Using other personal statistics
- Target model performance of +/- 10 seconds to be of practical use

# Available Variables



# Variable Selection

**Tests:**

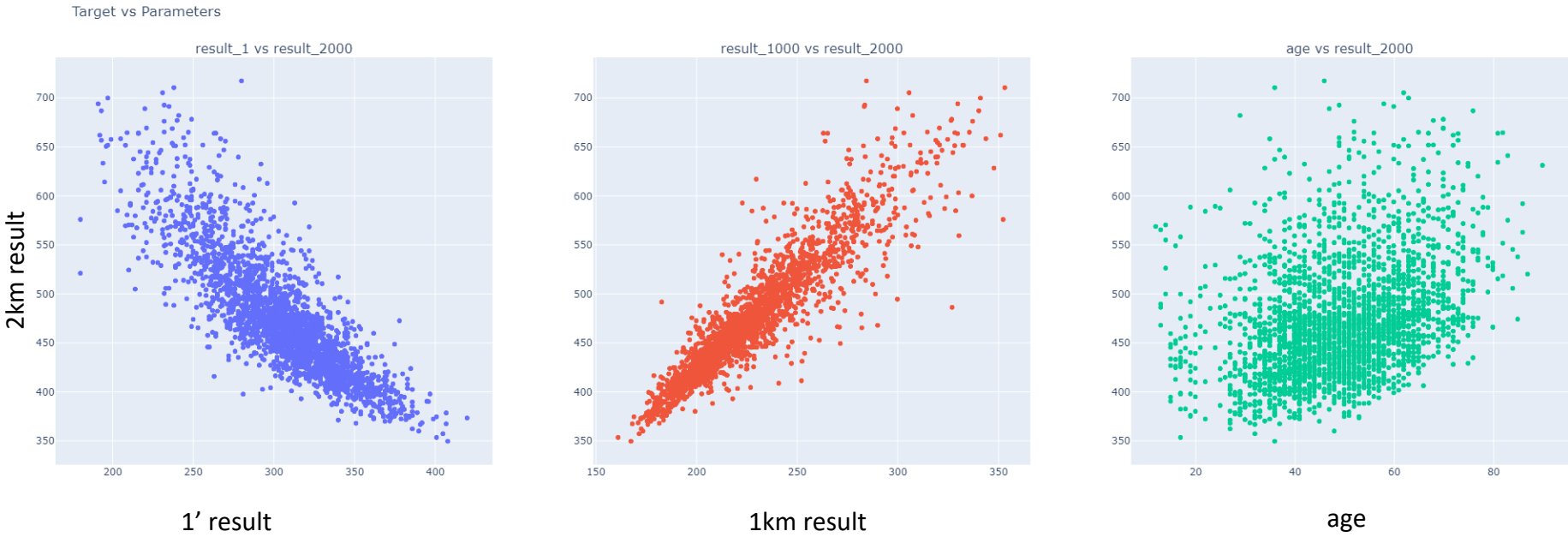
- 1km
- 1 minute

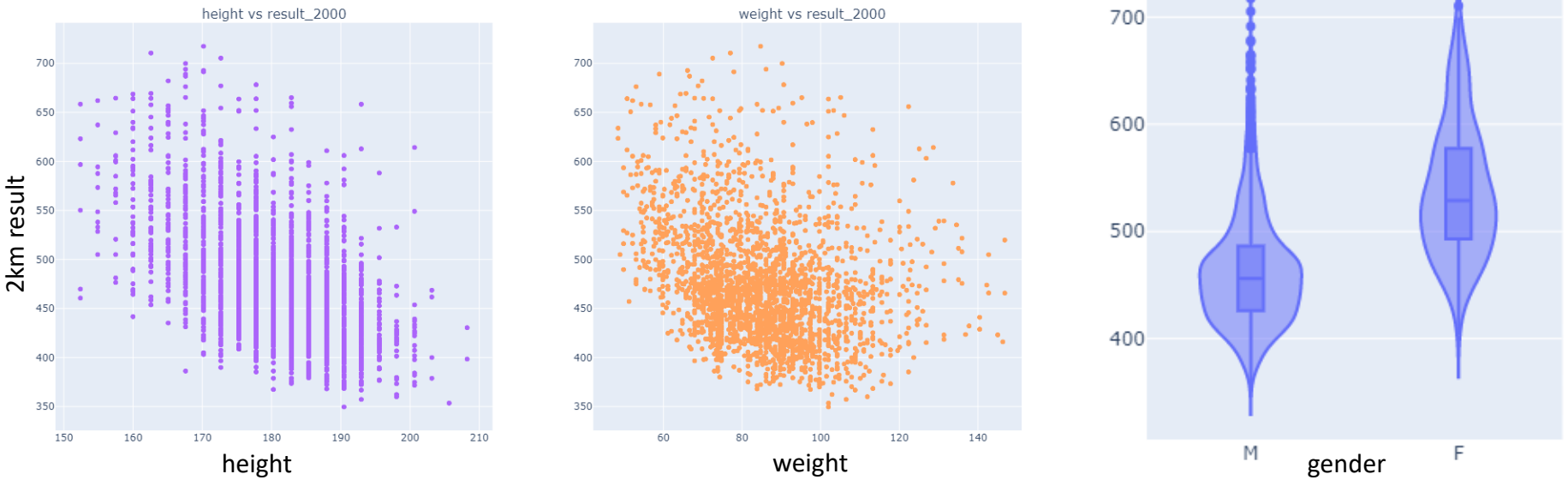
**Other:**

- Age
- Height
- Weight
- Weight Class
- Gender

**Dataset size:**

- Train: 1787
- Validate: 383
- Test: 383




# Target vs Predictors

# Linear Regression 1st Order Iteration 1

- Weight class: $p > \alpha$ (0.05)
  - Highly correlated with weight (completely explained by weight)
  - Drop from the dataset

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 265.0040 | 21.239 | 12.477 | 0.000 | 223.348 | 306.660 |
| weight_class[T.L] | 0.5819 | 1.549 | 0.376 | 0.707 | -2.456 | 3.620 |
| gender[T.M] | -3.9472 | 1.900 | -2.077 | 0.038 | -7.674 | -0.220 |
| result_1 | -0.2338 | 0.029 | -8.032 | 0.000 | -0.291 | -0.177 |
| result_1000 | 1.4589 | 0.034 | 42.476 | 0.000 | 1.392 | 1.526 |
| age | 0.2189 | 0.044 | 4.988 | 0.000 | 0.133 | 0.305 |
| height | -0.4281 | 0.085 | -5.014 | 0.000 | -0.596 | -0.261 |
| weight | 0.2139 | 0.048 | 4.441 | 0.000 | 0.119 | 0.308 |

# Linear Regression 1st Order Iteration 2

- All $p < \alpha$
- $R^2 = 0.869$
- RMSE 22.1 (4.65%)

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 266.7838 | 20.699 | 12.889 | 0.000 | 226.187 | 307.380 |
| gender[T.M] | -3.7725 | 1.842 | -2.048 | 0.041 | -7.386 | -0.159 |
| result_1 | -0.2338 | 0.029 | -8.034 | 0.000 | -0.291 | -0.177 |
| result_1000 | 1.4589 | 0.034 | 42.485 | 0.000 | 1.392 | 1.526 |
| age | 0.2181 | 0.044 | 4.977 | 0.000 | 0.132 | 0.304 |
| height | -0.4335 | 0.084 | -5.153 | 0.000 | -0.599 | -0.269 |
| weight | 0.2048 | 0.042 | 4.928 | 0.000 | 0.123 | 0.286 |

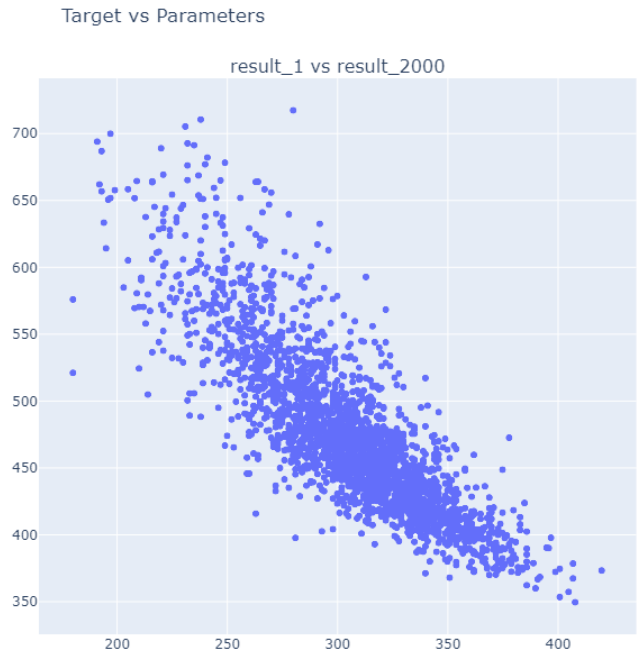
# Interpretation

| | coef |
|---|---|
| Intercept | 266.7838 |
| gender[T.M] | -3.7725 |
| result_1 | -0.2338 |
| result_1000 | 1.4589 |
| age | 0.2181 |
| height | -0.4335 |
| weight | 0.2048 |

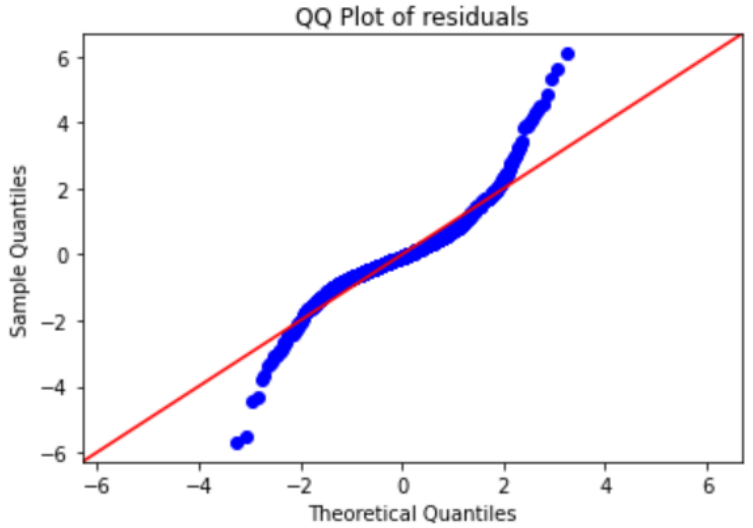
- Being male improves your 2km result.
- The better your 1-minute result, the better your 2km result.
- The better your 1km result, the better your 2km result.
- Being younger improves your 2km result.
- Being taller improves your 2km result.
- Being lighter improves your 2km result.

# Validity of Model

1. Linearity ✗
2. Statistical independence of residuals ✓
3. Homoscedasticity ✗
4. Normality ✗

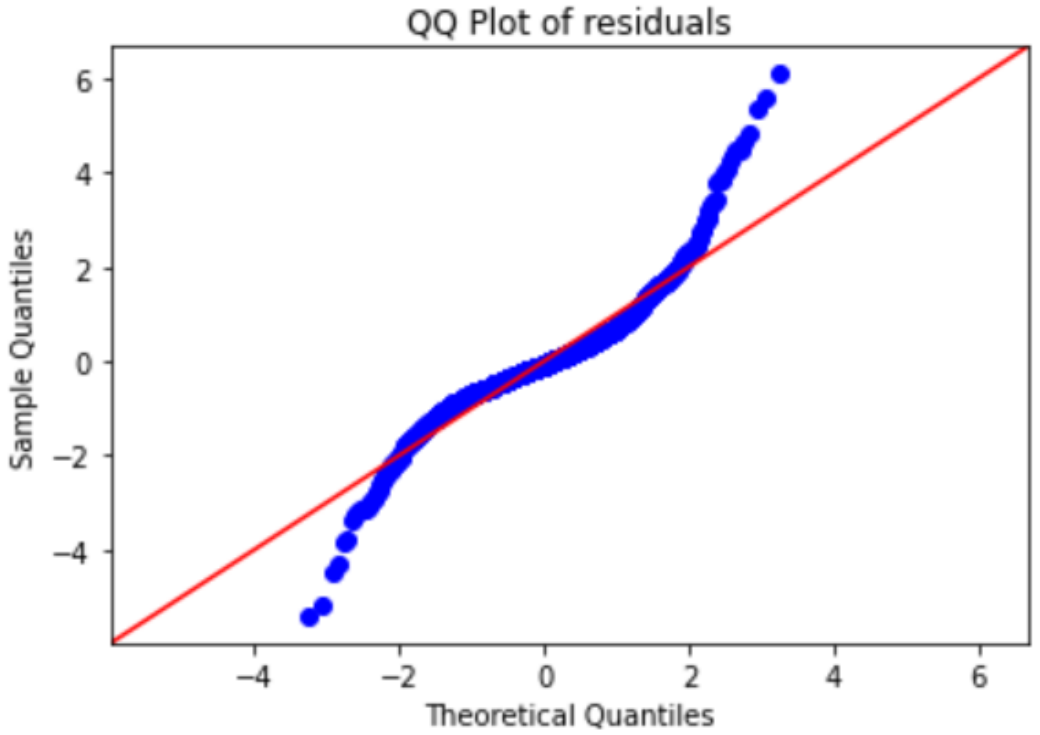




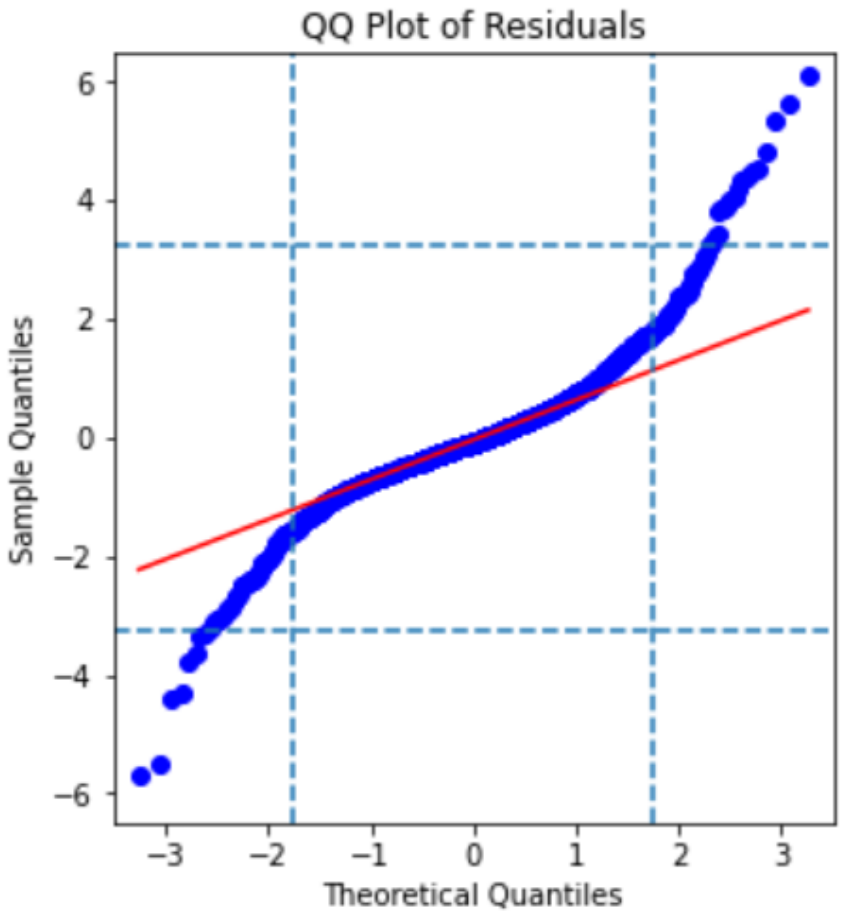
# Linear Regression
# 2nd Order
# **Iteration 4**

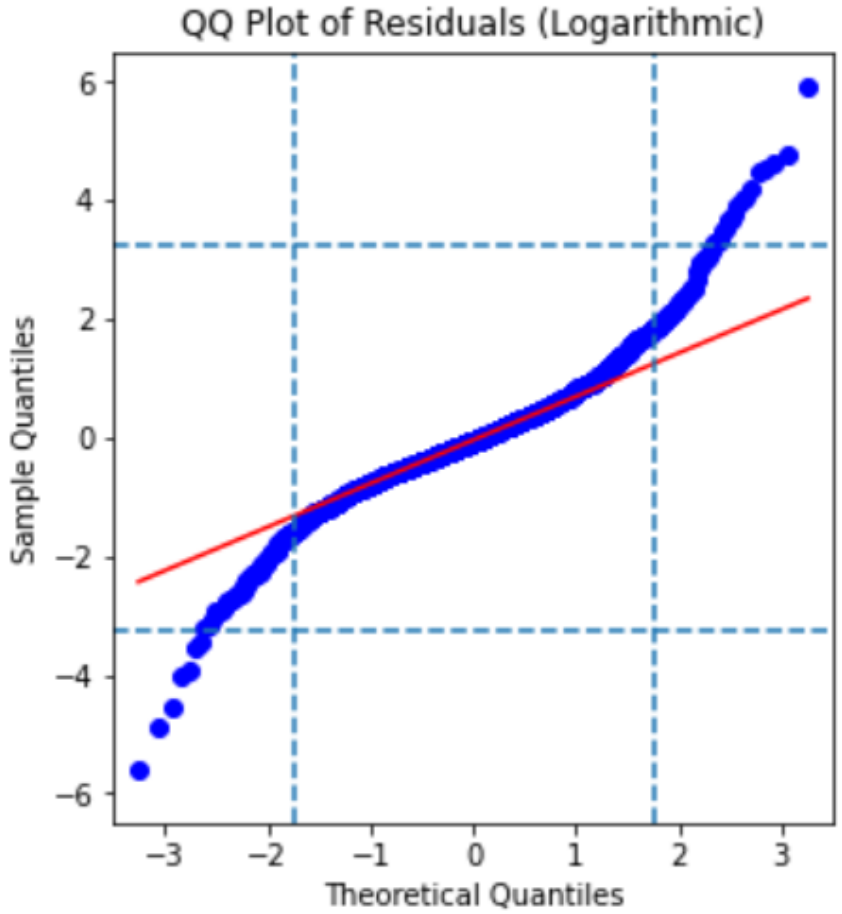
- $R^2 = 0.869$ ↔
- RMSE = 21.8 (4.59%) ↓




Linear Regression
Logarithmic Transformation
**Iteration 2**

- $R^2 = 0.872$ ↑
- RMSE = 22.2 (4.67%) ↑



# Power

Rowing machines directly measure athlete power, then convert to pace (and time & distance):

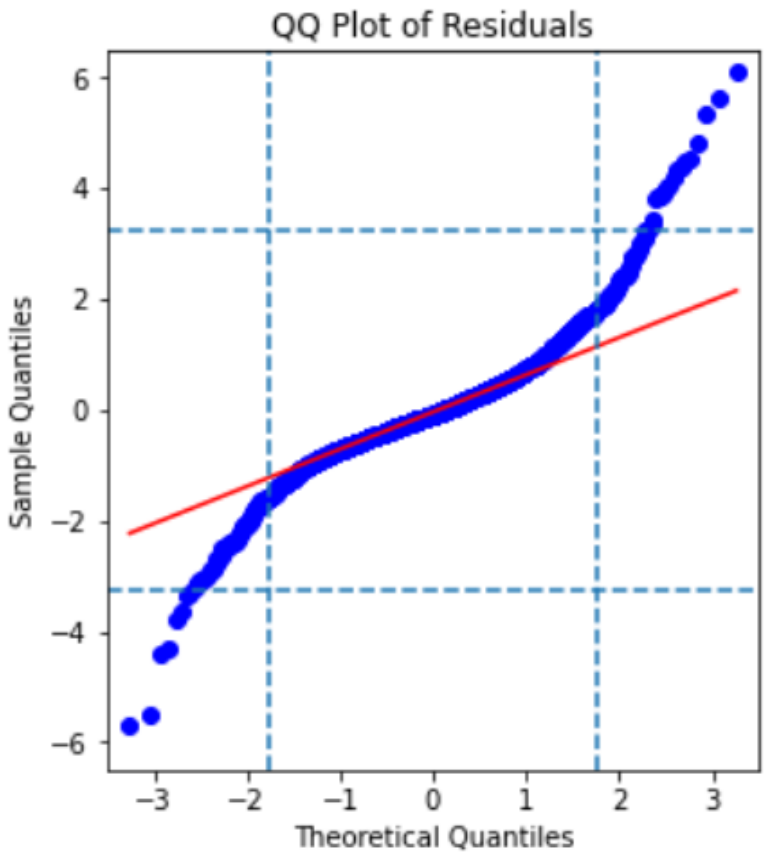
$$P = \frac{2.80}{pace^3}$$

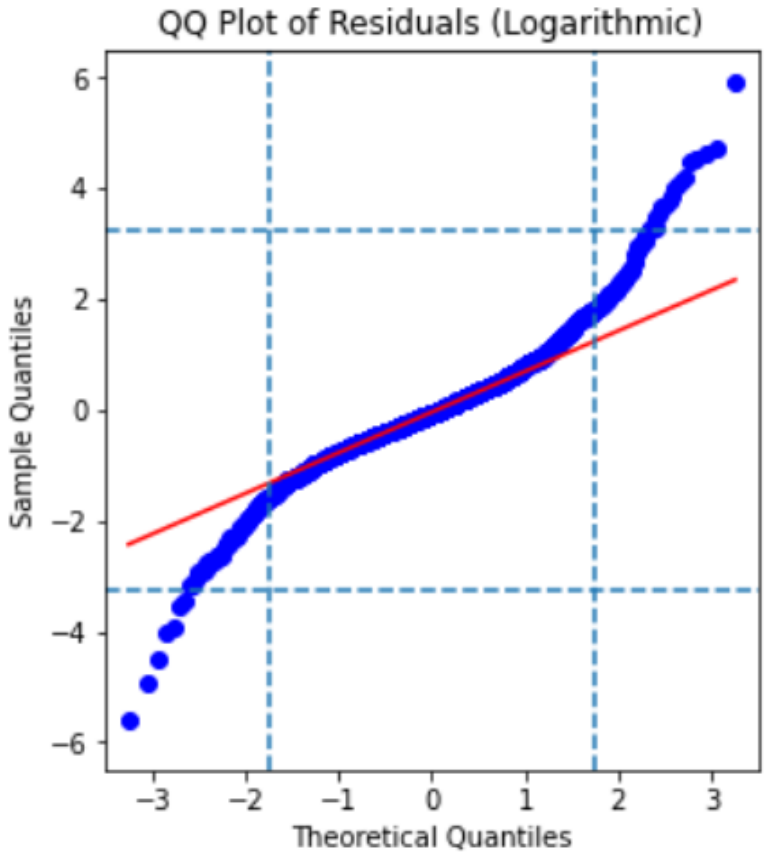
Where: $pace = t/d$

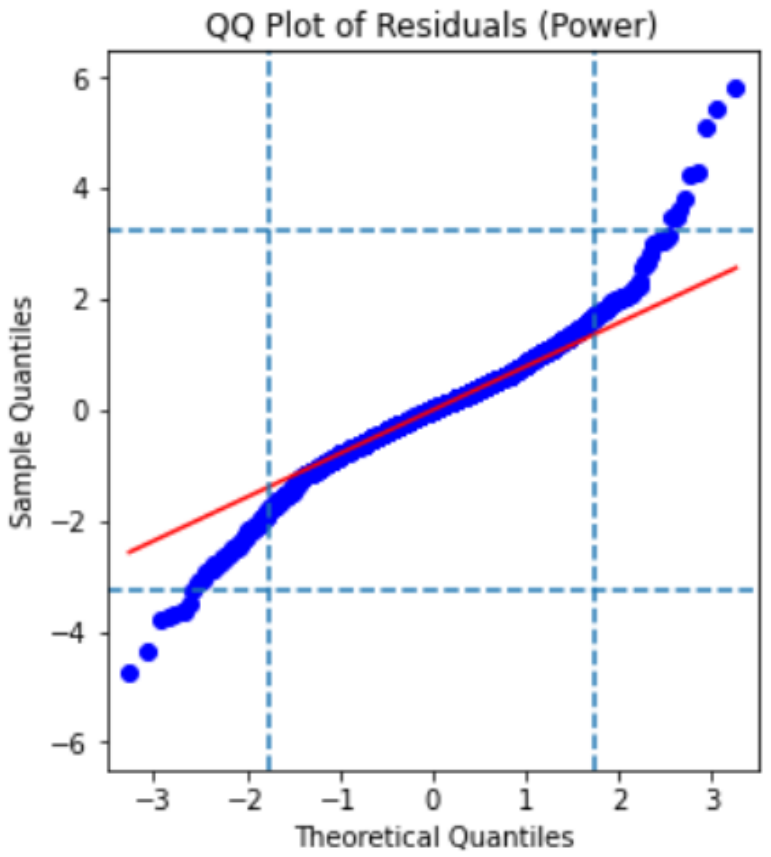
<u>Non-linear transformation</u>

# Linear Regression Power
# Iteration 2

- $R^2 = 0.890$ ↑
- RMSE = 22.2 (4.65%) ↓

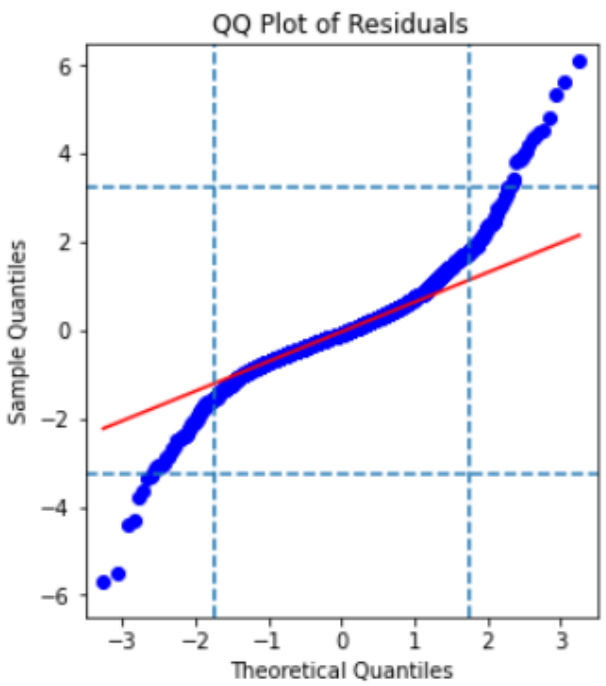


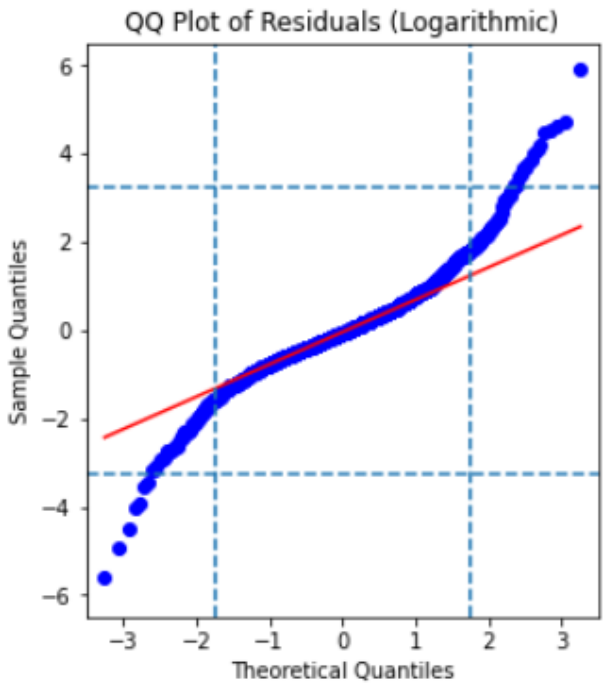



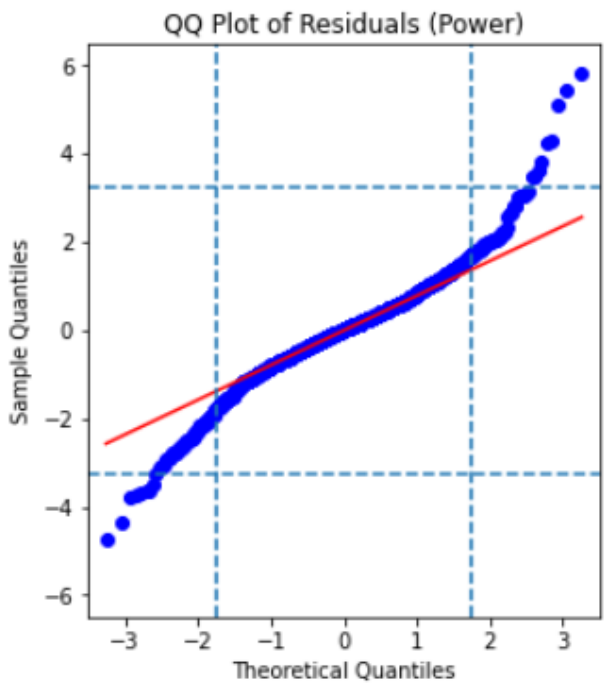



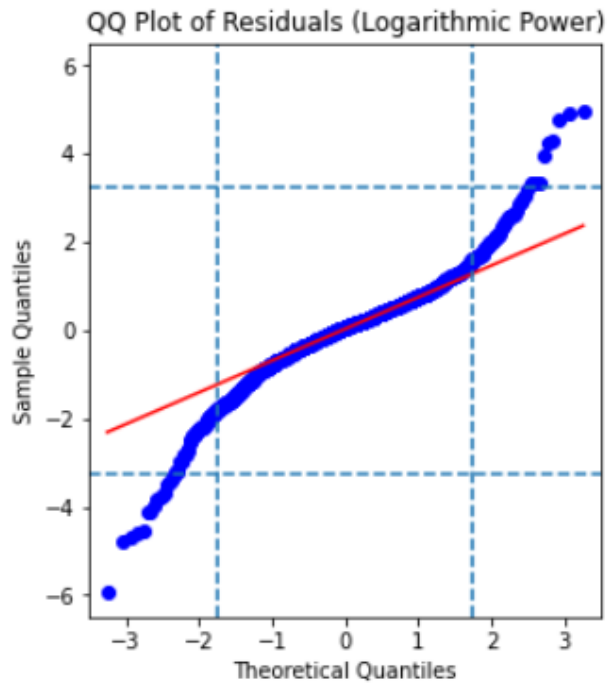
Linear Regression
Logarithmic Power
**Iteration 2**

- $R^2 = 0.877$ ↓
- RMSE = 22.0 (4.61%) ↓

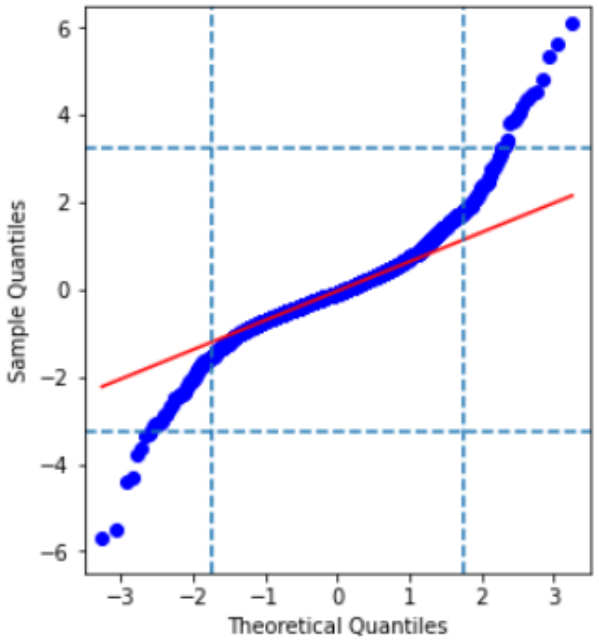


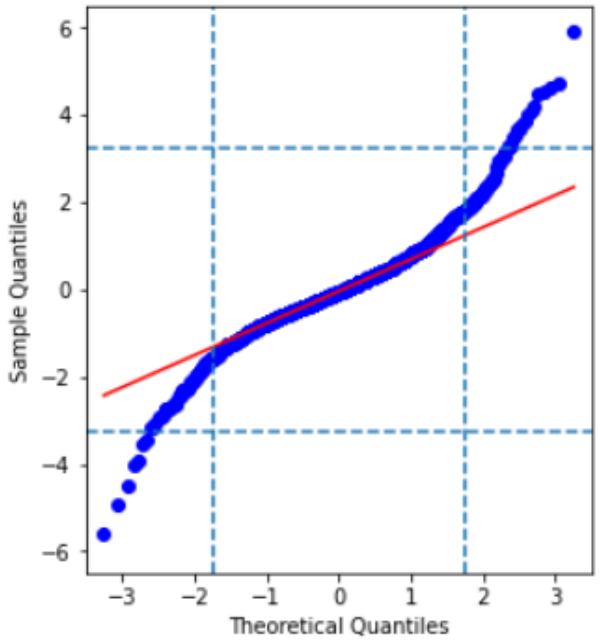
## 2nd Order Linear




$R^2 = 0.869$

RMSE = 21.8 (4.59%)

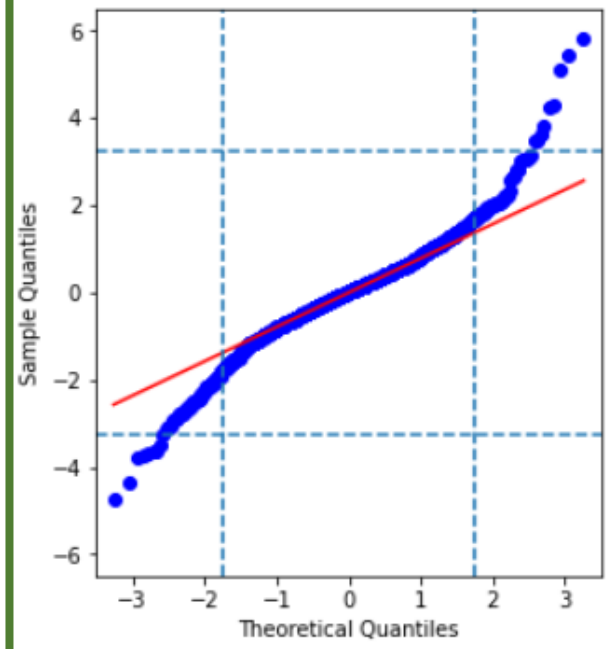
## Logarithmic




$R^2 = 0.872$

RMSE = 22.2 (4.65%)

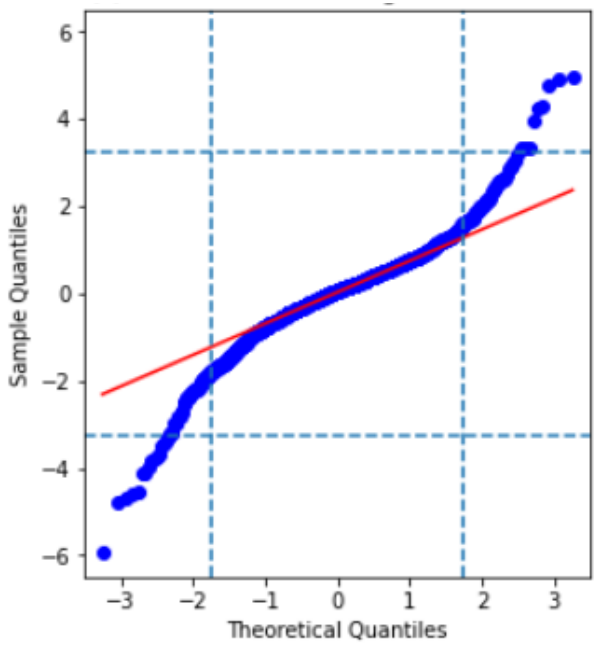
## Power




$R^2 = 0.890$

RMSE = 22.2 (4.65%)

## Logarithmic Power




$R^2 = 0.877$

RMSE = 22.0 (4.61%)

Best $R^2$

Better residuals

Real basis for transformation

Easier to interpret

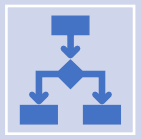
# Interpretation

| | coef |
|---|---|
| gender[F] | -26.1714 |
| gender[M] | -21.6351 |
| result_1 | 0.0656 |
| result_1000 | 0.6441 |
| age | -0.2577 |
| height | 0.5055 |
| weight | -0.2547 |

- Being male improves your 2km result.
- The better your 1-minute result, the better your 2km result.
- The better your 1km result, the better your 2km result.
- Being younger improves your 2km result.
- Being taller worsens your 2km result.
- Being lighter improves your 2km result.

# Testing

**Results**



Chosen model — RMSE: 73.3 (15.6%)

Comparison to Paul's Law — +5s for each doubling of distance; RMSE = 27.9

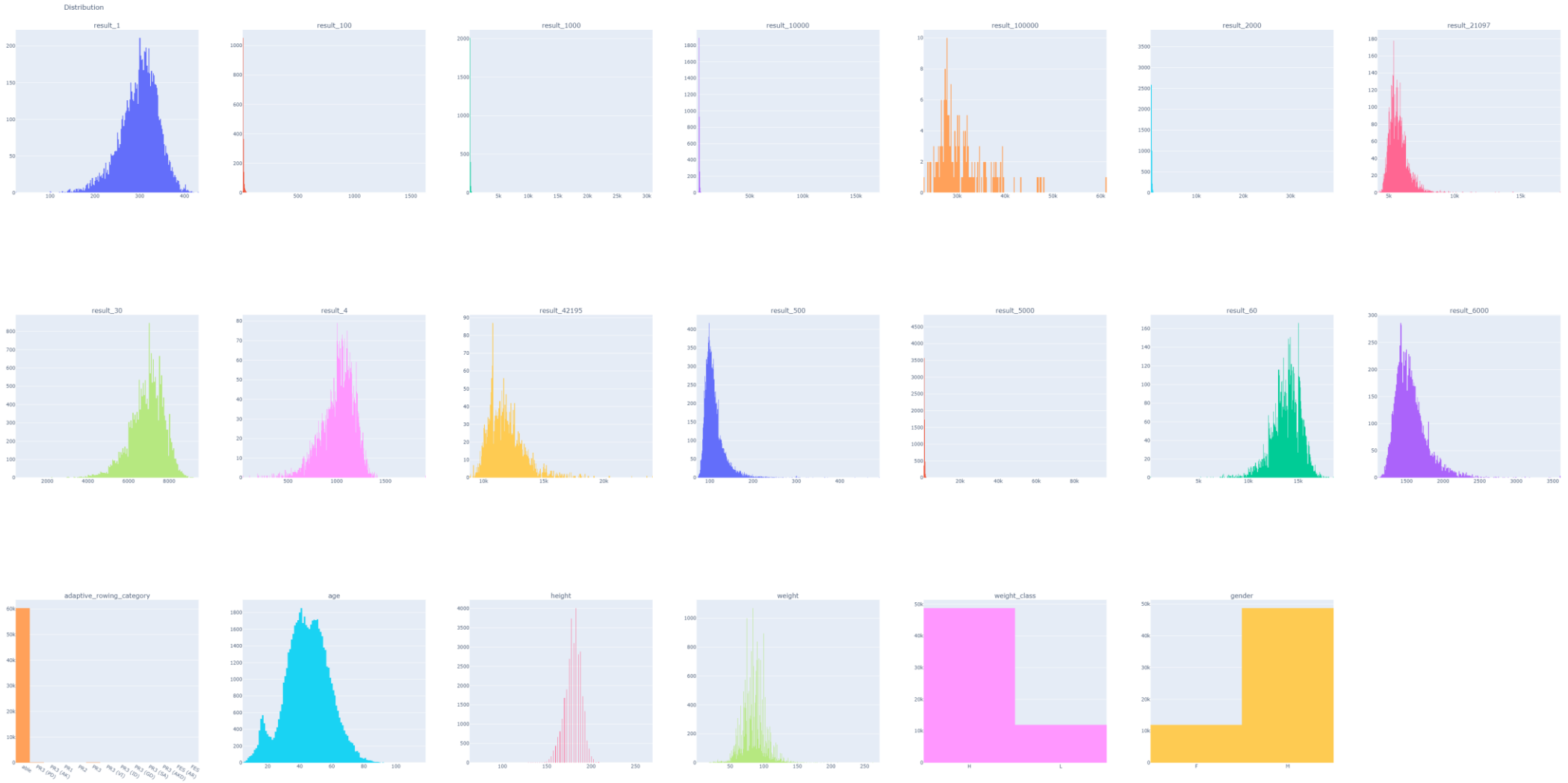
**Conclusions**

- Huge RMSE compared to validation data
- Non-normal residuals resulting in unpredictable behaviour
- Not accurate enough to be useful
- Worse than "rule of thumb"

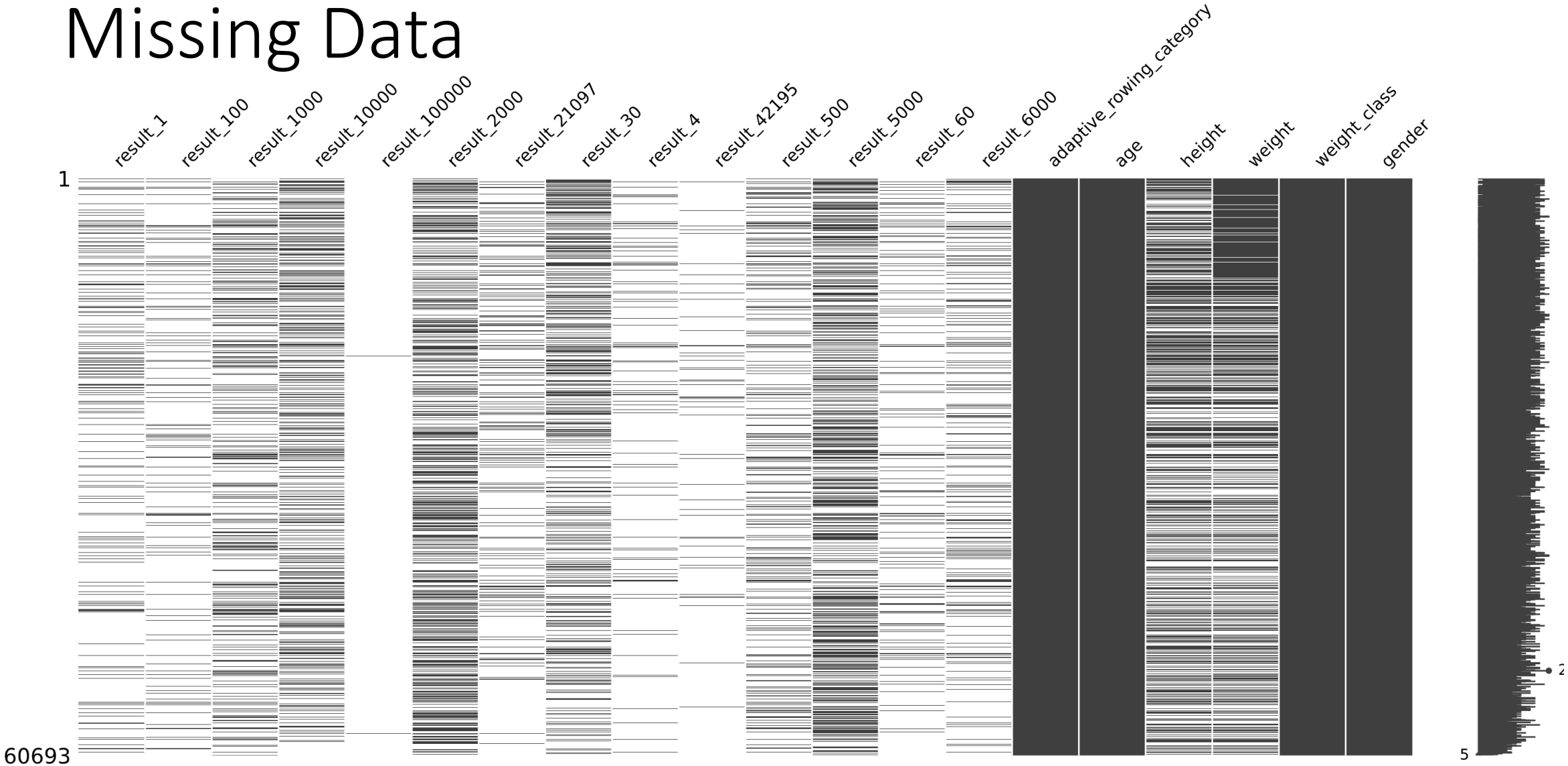
# Next Steps

- Try models that do not have normality as an assumption
- Further refine data
  - Outliers
  - Non-representative performances
- Train model over a narrower range of results

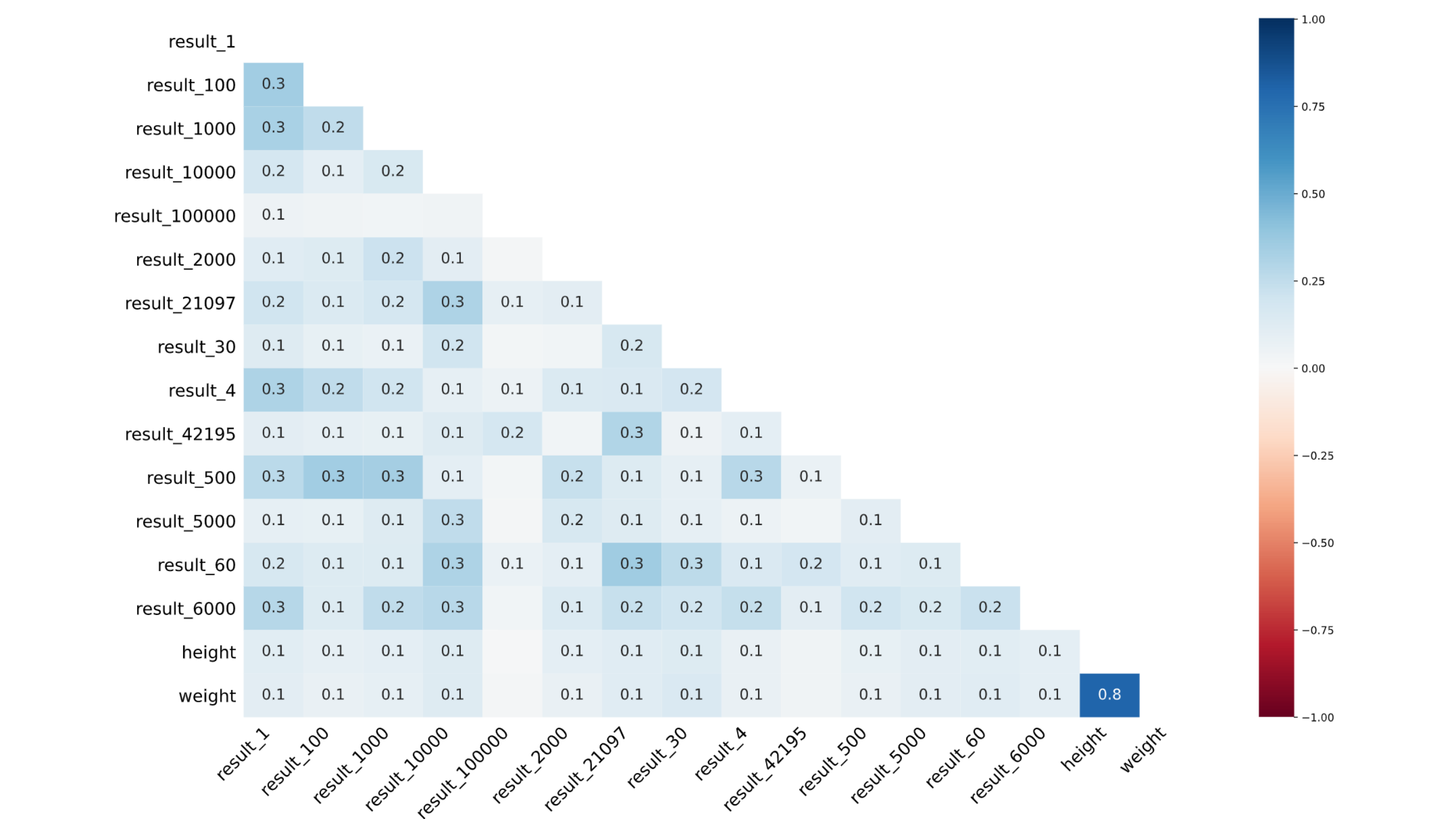
# Distribution - Before



# Missing Data




Data %


result_1
result_100 0.3
result_1000 0.3 0.2
result_10000 0.2 0.1 0.2
result_100000 0.1
result_2000 0.1 0.1 0.2 0.1
result_21097 0.2 0.1 0.2 0.3 0.1 0.1
result_30 0.1 0.1 0.1 0.2 0.2
result_4 0.3 0.2 0.2 0.1 0.1 0.1 0.1 0.2
result_42195 0.1 0.1 0.1 0.1 0.2 0.3 0.1 0.1
result_500 0.3 0.3 0.3 0.1 0.2 0.1 0.1 0.3 0.1
result_5000 0.1 0.1 0.1 0.3 0.2 0.1 0.1 0.1 0.1
result_60 0.2 0.1 0.1 0.3 0.1 0.1 0.3 0.3 0.1 0.2 0.1 0.1
result_6000 0.3 0.1 0.2 0.3 0.1 0.2 0.2 0.2 0.1 0.2 0.2 0.2
height 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
weight 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.8
result_1 result_100 result_1000 result_10000 result_100000 result_2000 result_21097 result_30 result_4 result_42195 result_500 result_5000 result_60 result_6000 height weight
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

# Missing Data

- Missing heights and weights strongly correlated with each other.
- Poor correlation between other missing data.
- Consider this as MCAR
- Can drop missing data