Norseman Race Results 2003-2023

Data Cleaning and Analysis report.

Data Sources:

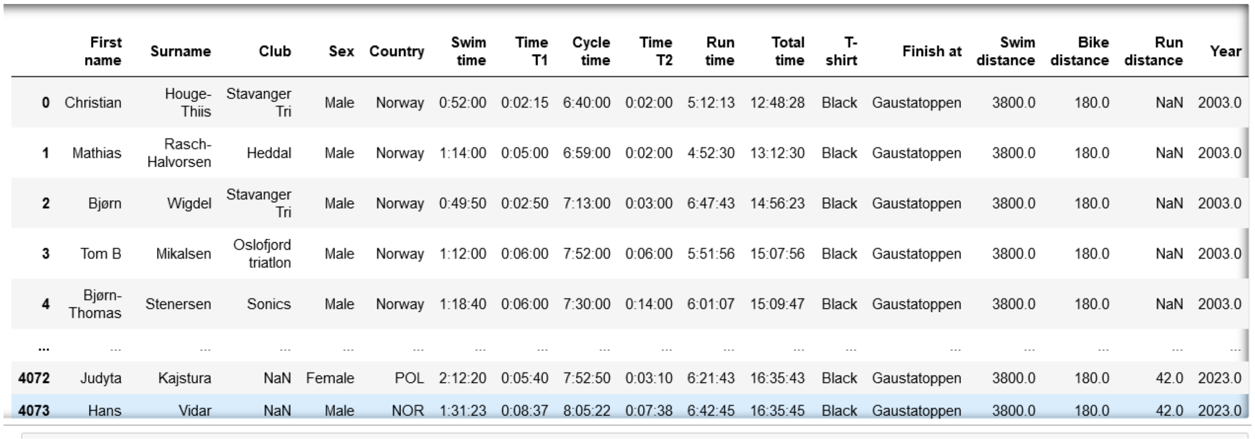
- https://nxtri.com/2002/05/19/results-all-years/
- nxtri.com for general information

Programming Language and Resources used for the Analysis:

- IDE: Jupiter Notebook
- Language: Python
- Libraries and Packages:
  Pandas
  Seaborn

Data structure:

```
        4077 entries, 0 to 4076
        Data columns (total 17 columns):
   •     #   Column         Non-Null Count  Dtype  
   •    ---  ------         --------------  -----  
   •     0   First name     4076 non-null   object 
   •     1   Surname        4076 non-null   object 
   •     2   Club           2388 non-null   object 
   •     3   Sex            4076 non-null   object 
   •     4   Country        4052 non-null   object 
   •     5   Swim time      4030 non-null   object 
   •     6   Time T1        3936 non-null   object 
   •     7   Cycle time     4032 non-null   object 
   •     8   Time T2        3939 non-null   object 
   •     9   Run time       4036 non-null   object 
   •     10  Total time     4076 non-null   object 
   •     11  T-shirt        4076 non-null   object 
   •     12  Finish at      4076 non-null   object 
   •     13  Swim distance  4076 non-null   float64
   •     14  Bike distance  4076 non-null   float64
   •     15  Run distance   258 non-null    float64
   •     16  Year           4076 non-null   float64
```



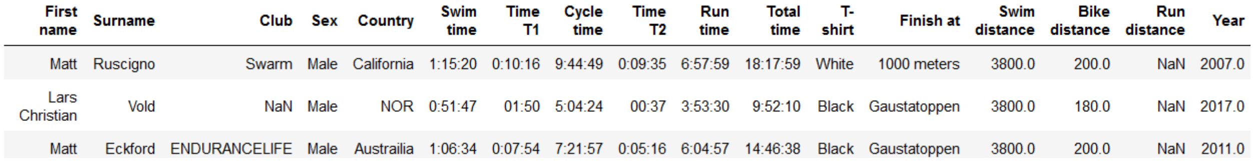
	First name	Surname	Club	Sex	Country	Swim time	Time T1	Cycle time	Time T2	Run time	Total time	T-shirt	Finish at	Swim distance	Bike distance	Run distance	Year
0	Christian	Houge-Thiis	Stavanger Tri	Male	Norway	0:52:00	0:02:15	6:40:00	0:02:00	5:12:13	12:48:28	Black	Gaustatoppen	3800.0	180.0	NaN	2003.0
1	Mathias	Rasch-Halvorsen	Heddal	Male	Norway	1:14:00	0:05:00	6:59:00	0:02:00	4:52:30	13:12:30	Black	Gaustatoppen	3800.0	180.0	NaN	2003.0
2	Bjørn	Wigdel	Stavanger Tri	Male	Norway	0:49:50	0:02:50	7:13:00	0:03:00	6:47:43	14:56:23	Black	Gaustatoppen	3800.0	180.0	NaN	2003.0
3	Tom B	Mikalsen	Oslofjord triatlon	Male	Norway	1:12:00	0:06:00	7:52:00	0:06:00	5:51:56	15:07:56	Black	Gaustatoppen	3800.0	180.0	NaN	2003.0
4	Bjørn-Thomas	Stenersen	Sonics	Male	Norway	1:18:40	0:06:00	7:30:00	0:14:00	6:01:07	15:09:47	Black	Gaustatoppen	3800.0	180.0	NaN	2003.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4072	Judyta	Kajstura	NaN	Female	POL	2:12:20	0:05:40	7:52:50	0:03:10	6:21:43	16:35:43	Black	Gaustatoppen	3800.0	180.0	42.0	2023.0
4073	Hans	Vidar	NaN	Male	NOR	1:31:23	0:08:37	8:05:22	0:07:38	6:42:45	16:35:45	Black	Gaustatoppen	3800.0	180.0	42.0	2023.0

Detected Problems and Resolution:

Missing Countries and Wrong Entries:

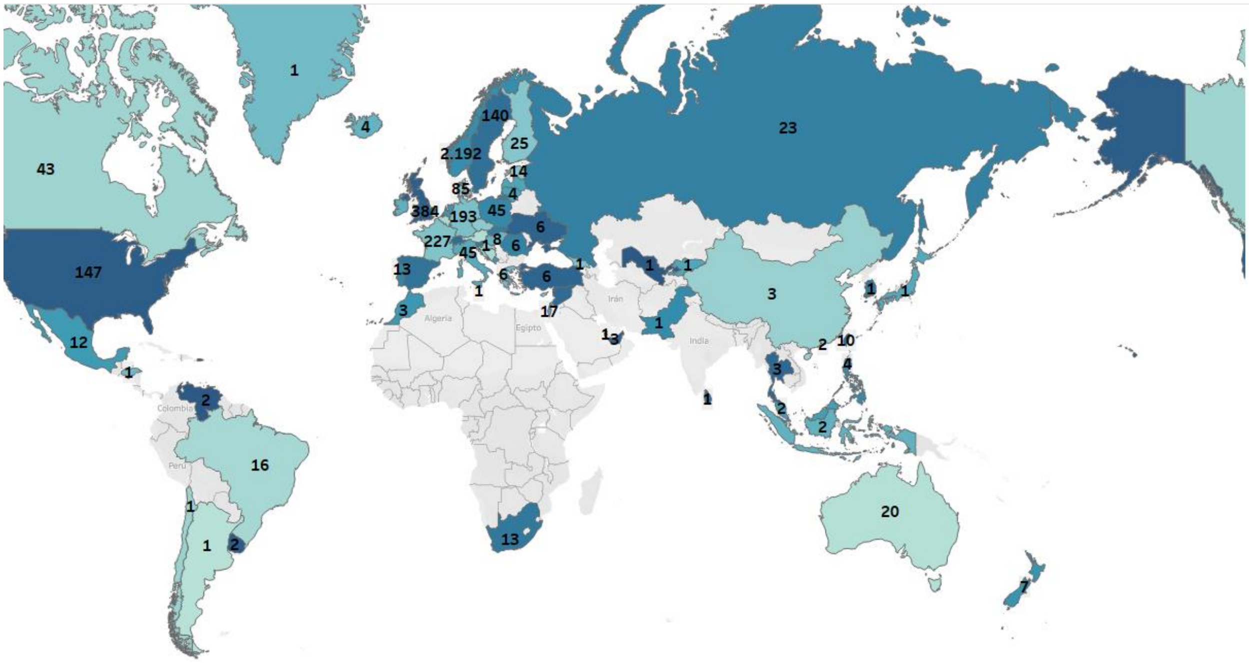
From the initial dataset the amount different countries participating in Norseman through the years was 155, when the actual amount was 72. That was due to some error in the data, mostly from the early years.

Some athletes were registered with the city and not the country, some countries were just missing capital letters or misspell and some were in 'ISO 3166-1 alfa-3' format (a 3 capital letters code for a country).



First name	Surname	Club	Sex	Country	Swim time	Time T1	Cycle time	Time T2	Run time	Total time	T-shirt	Finish at	Swim distance	Bike distance	Run distance	Year
Matt	Ruscigno	Swarm	Male	California	1:15:20	0:10:16	9:44:49	0:09:35	6:57:59	18:17:59	White	1000 meters	3800.0	200.0	NaN	2007.0
Lars Christian	Vold	NaN	Male	NOR	0:51:47	01:50	5:04:24	00:37	3:53:30	9:52:10	Black	Gaustatoppen	3800.0	180.0	NaN	2017.0
Matt	Eckford	ENDURANCELIFE	Male	Austrailia	1:06:34	0:07:54	7:21:57	0:05:16	6:04:57	14:46:38	Black	Gaustatoppen	3800.0	200.0	NaN	2011.0

1 Country errors example image



2 Participants by Country

Missing Distances:

Some distances were missing and there were some changes in the bike and swimming distances in some years due to bad weather.

In 2005, Half Swim Distance and 200 km bike in 2007 and 2011

Errors in Partial Times:

In partial Times there was missing data, T1 and T2 in most cases where set like hours instead of minutes, some Times were in 0, and in most cases the some of partial times did not match the total

Resolution: in the case of 1 partial Time missing, I set it to be the difference between the Total Time and the sum of the other partials, in the case of more than 1 partial missing and replace it with the proportion of the time over the total time calculating the avg of times by sport.

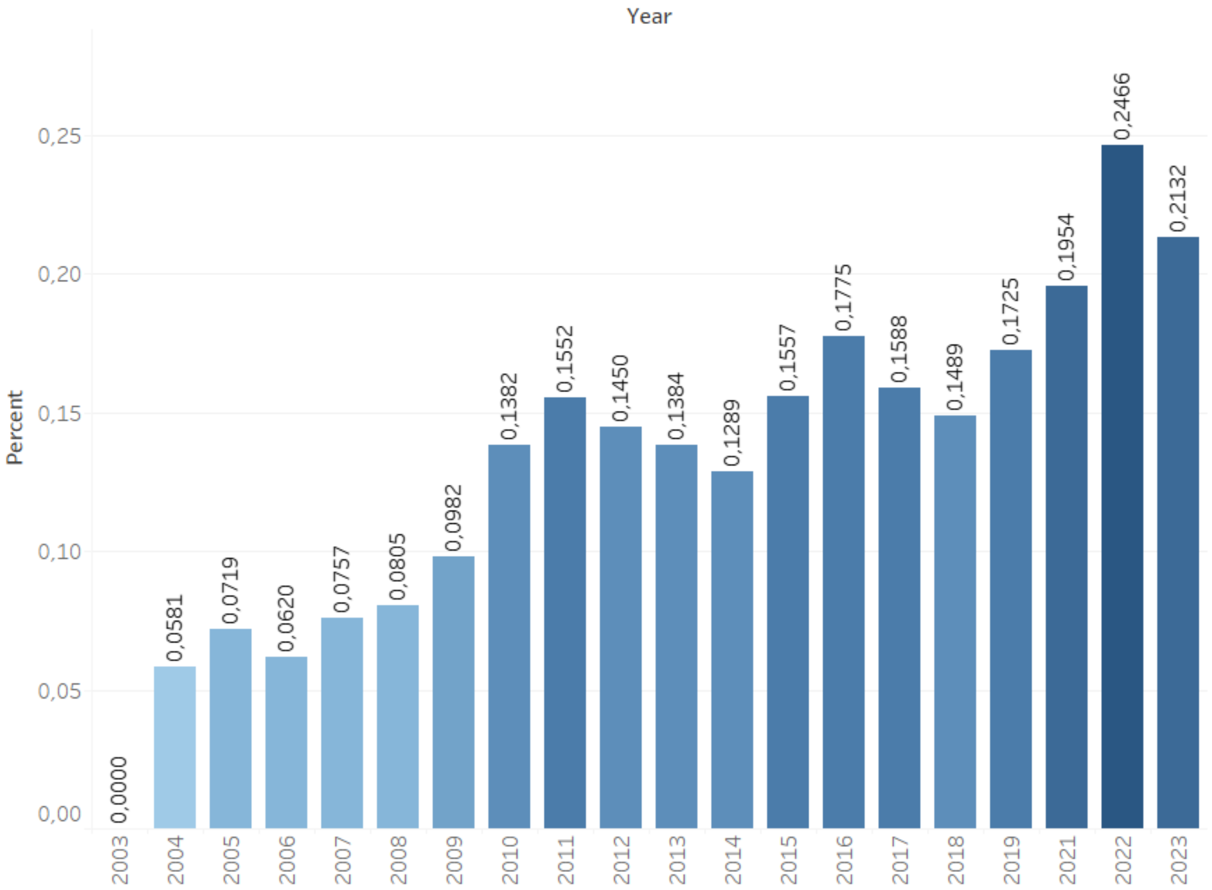
T-shirts and finish lines:

In some rows the finish line was Gaustatoppen and yet the T-shirts were White and vice versa. After some research by year I understand that this shouldn't be possible so re-checking everyone's finish line I assign the correct T-shirts to the data set.

Insights:

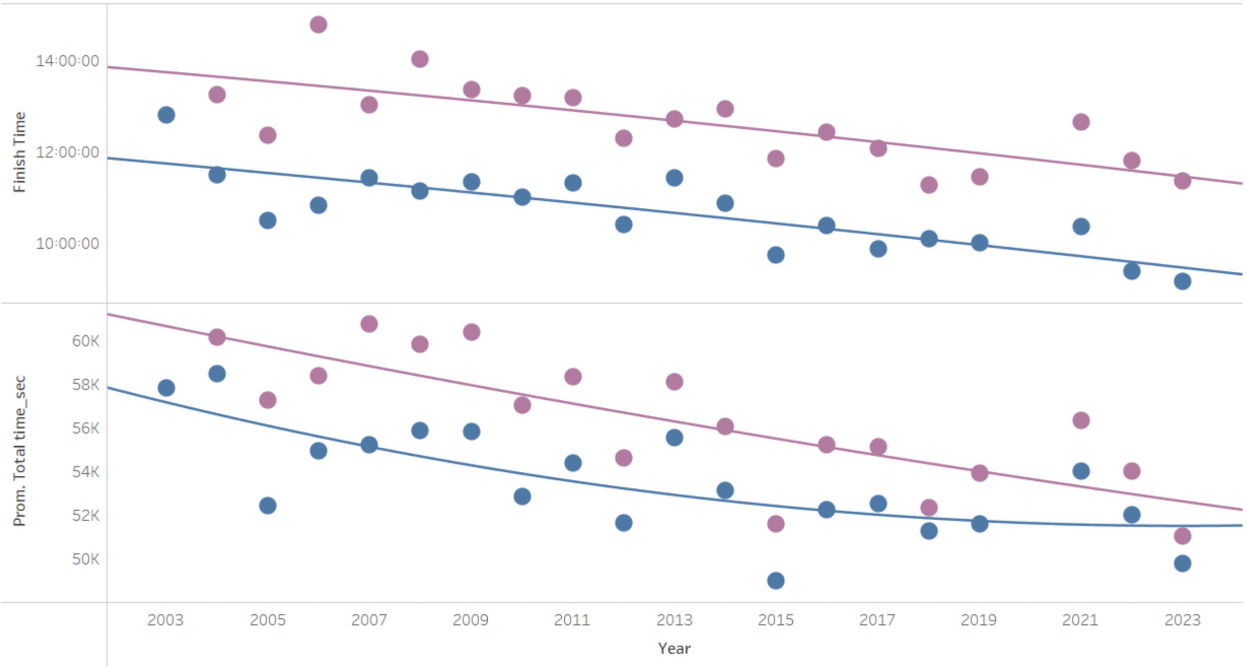
The Participation of women in the Norseman Race has increase significantly from an average of 7% of participants to 24% and 21% in the last 2 years.

Women Participation

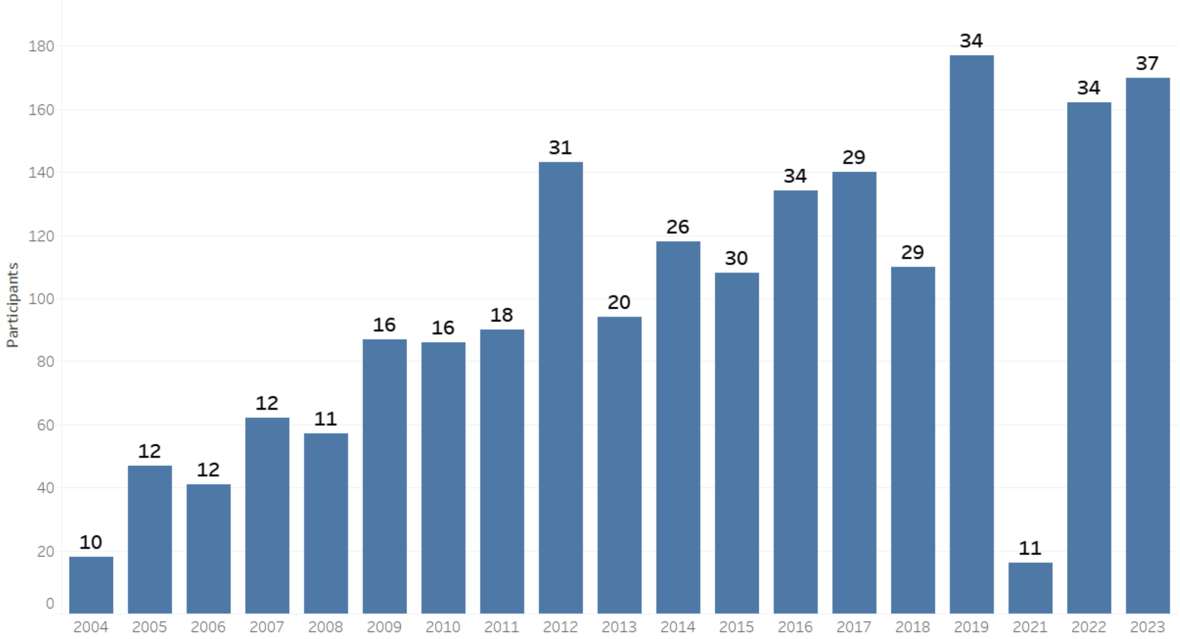


The Athletes Performance has been getting better by year in a very straight trend line, and on the graph on the bottom you can see the women average performance getting closer to man as the percent of women participation increases

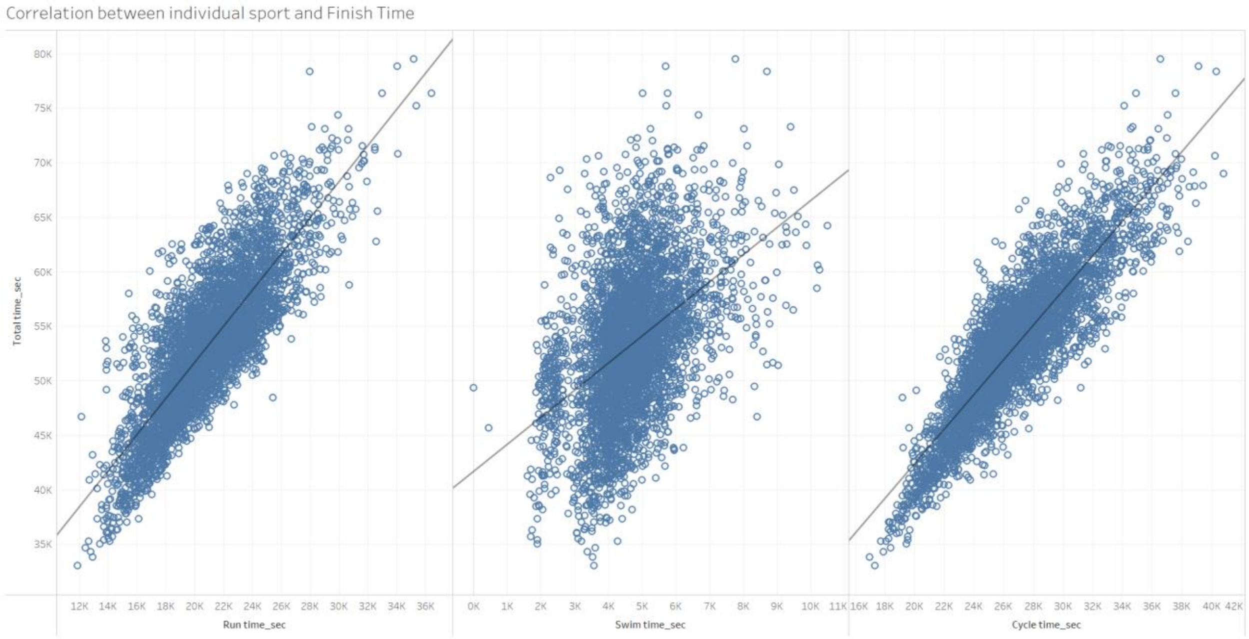
Women and Man, Records By Year



The participation of foreign countries has increased, reaching over 170 participants a year from more than 30 different countries. In the graph in visible the impact of covid in the event in the year 2021.



For last I show the correlation between individual sports and finish time, showing how best Cycle Times have the biggest impact in the final result, follow by Running and finish with Swimming.


Correlation between individual sport and Finish Time
Total time_sec
80K
75K
70K
65K
60K
55K
50K
45K
40K
35K
12K 14K 16K 18K 20K 22K 24K 26K 28K 30K 32K 34K 36K
Run time_sec
0K 1K 2K 3K 4K 5K 6K 7K 8K 9K 10K 11K
Swim time_sec
16K 18K 20K 22K 24K 26K 28K 30K 32K 34K 36K 38K 40K 42K
Cycle time_sec