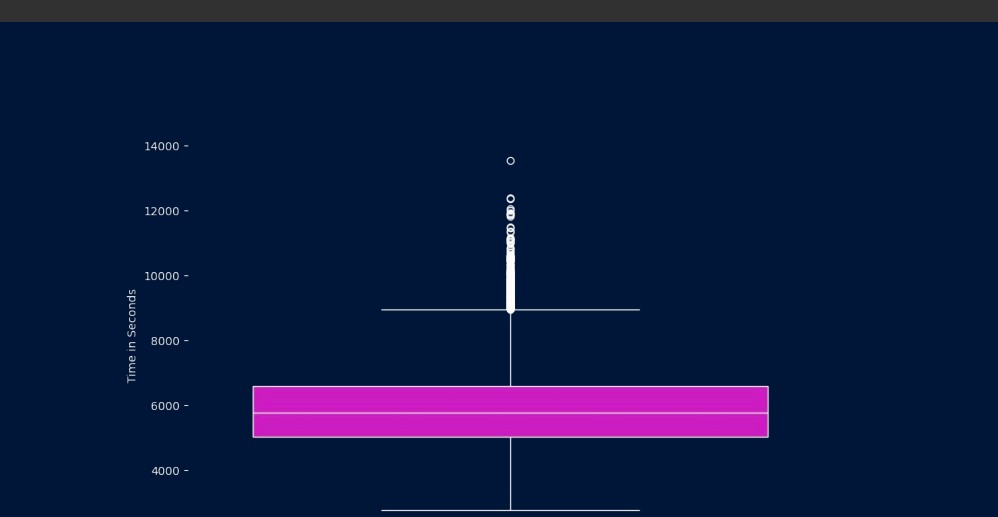
DUC TUNG DANG
(LIAM)

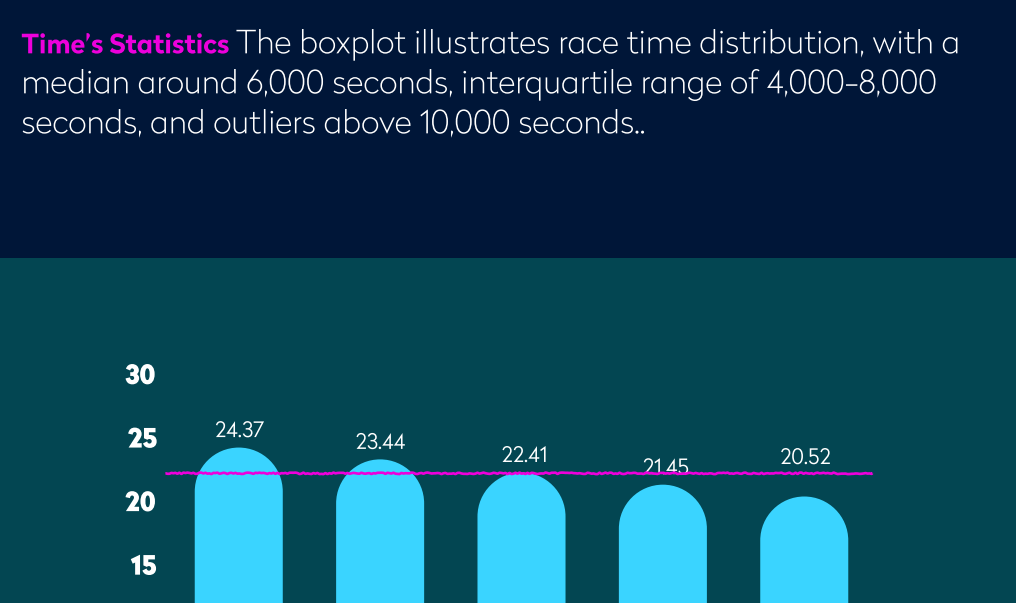
NORTH EAST
MARATHON ANALYSIS

GENERAL INFORMATION

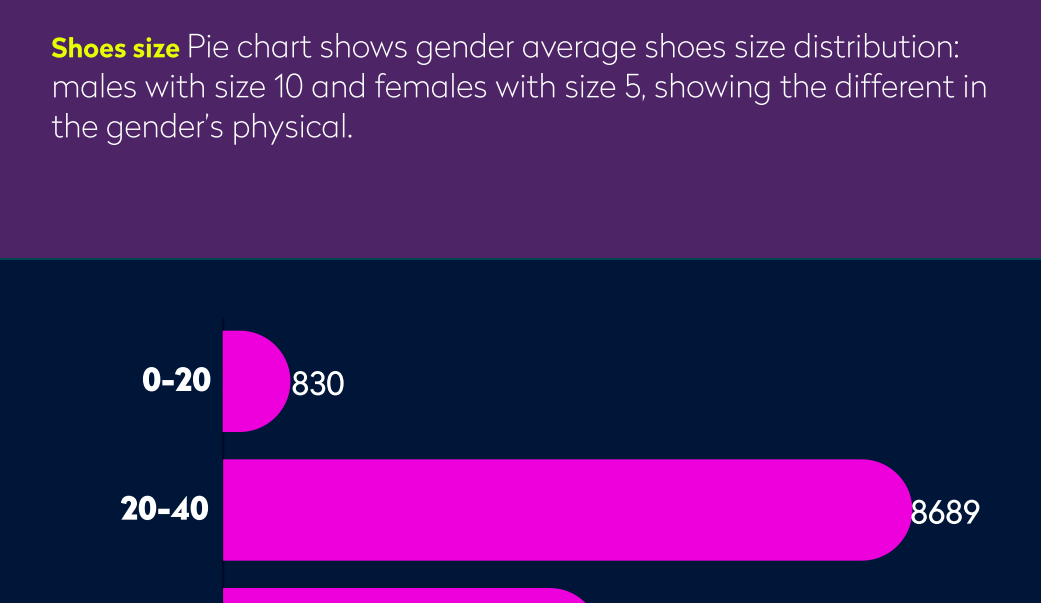


Time's Statistics The boxplot illustrates race time distribution, with a median around 6,000 seconds, interquartile range of 4,000-8,000 seconds, and outliers above 10,000 seconds.

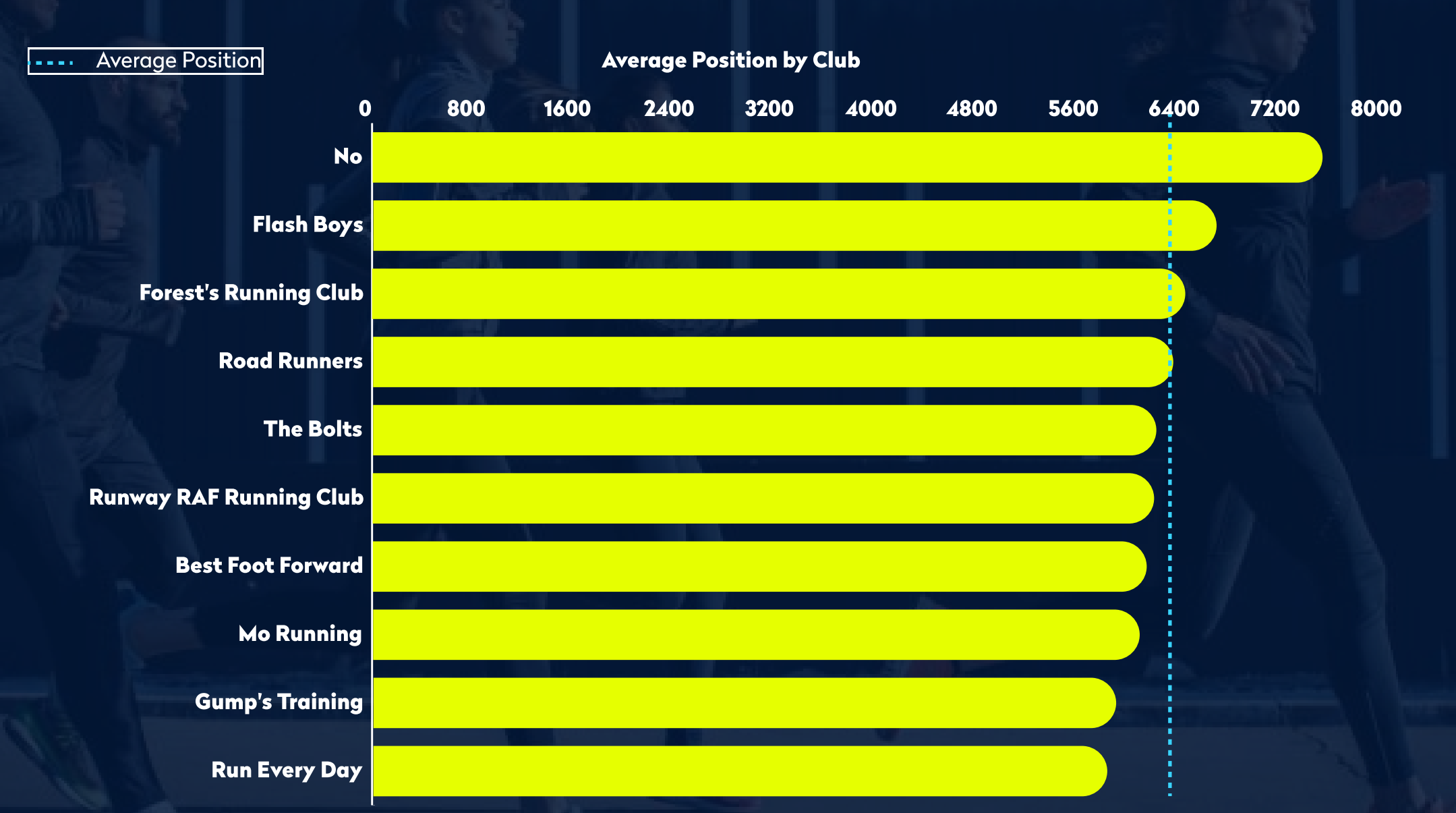
Shoes size Pie chart shows gender average shoes size distribution: males with size 10 and females with size 5, showing the different in the gender's physical.



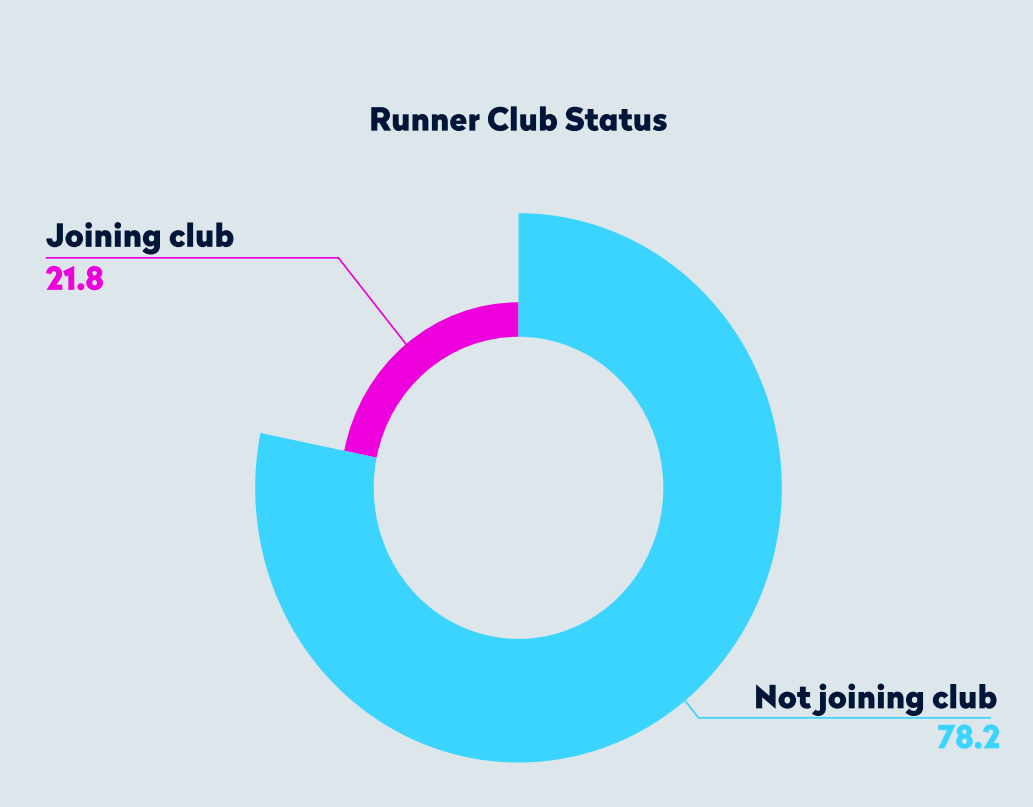
BMI Distribution by Position The bar chart shows average BMI by position bin, decreasing from 24.5 (1-3000) to 20 (12001-14567), below the overall average of 22.49.

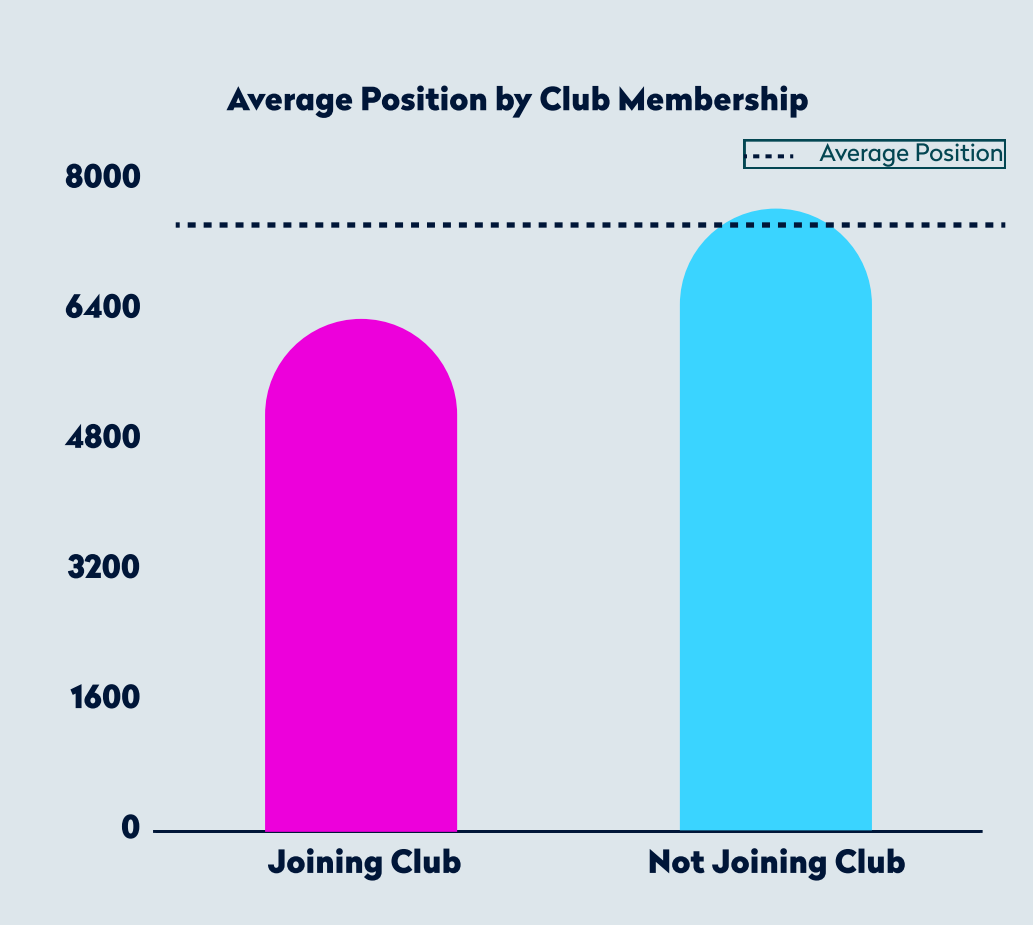


Age Group Histogram shows age distribution: 20-40 (8,689 runners), 40-60 (4,746), 0-20 (830), 60-80 (249), with highest concentration in 20-40.



Average Position by Club The graph illustrates the average position by club, highlighting variations in rankings among clubs. Overall, it indicates that club membership generally enhances runners' performance compared to those not affiliated with a club.

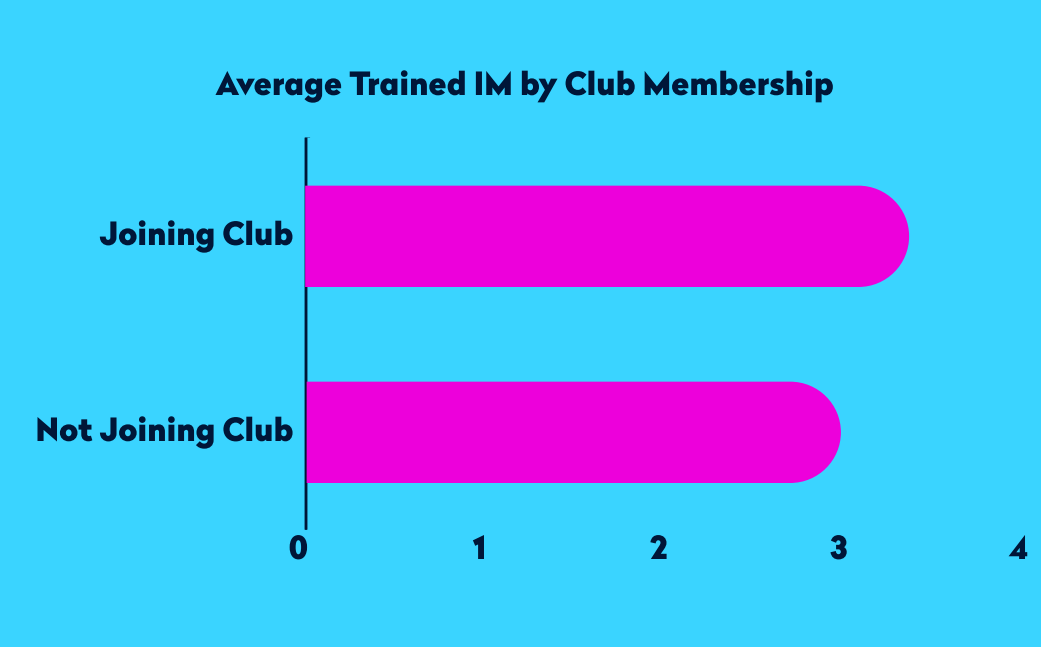




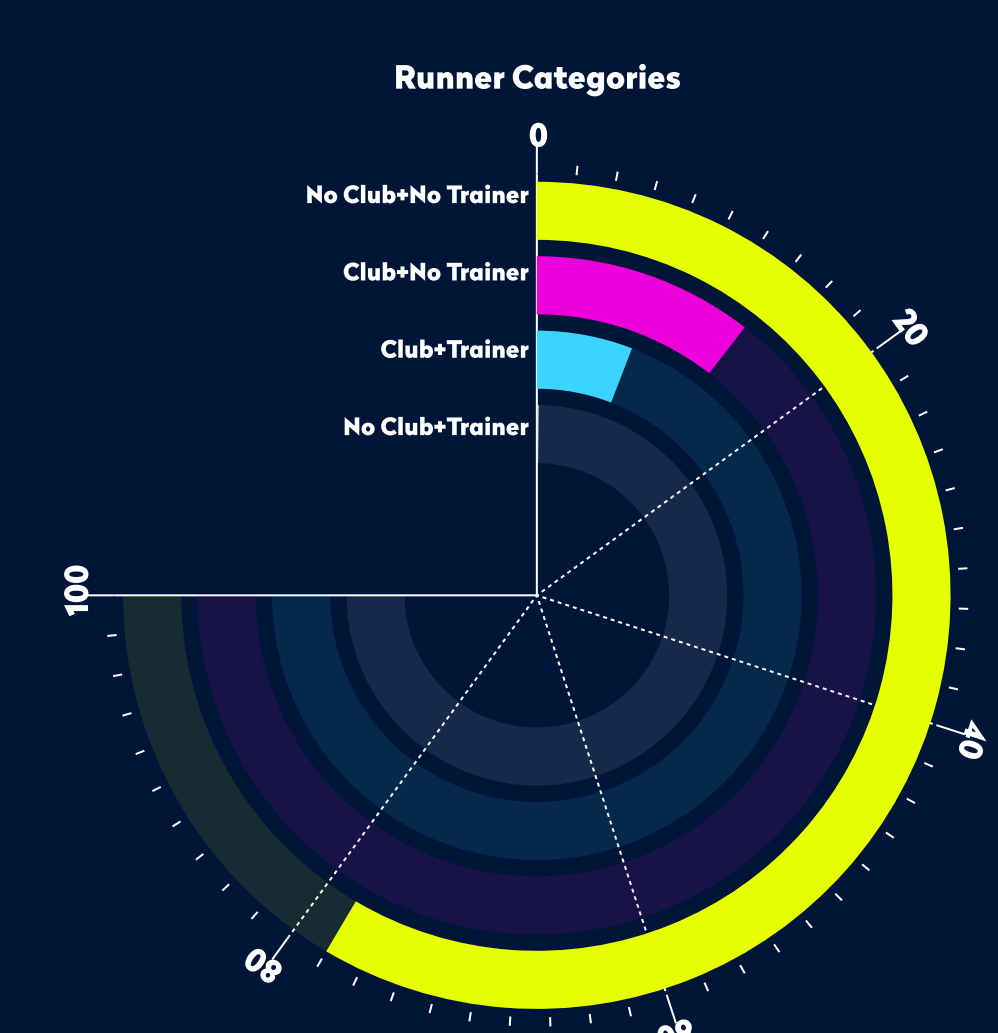
Club Membership Insights

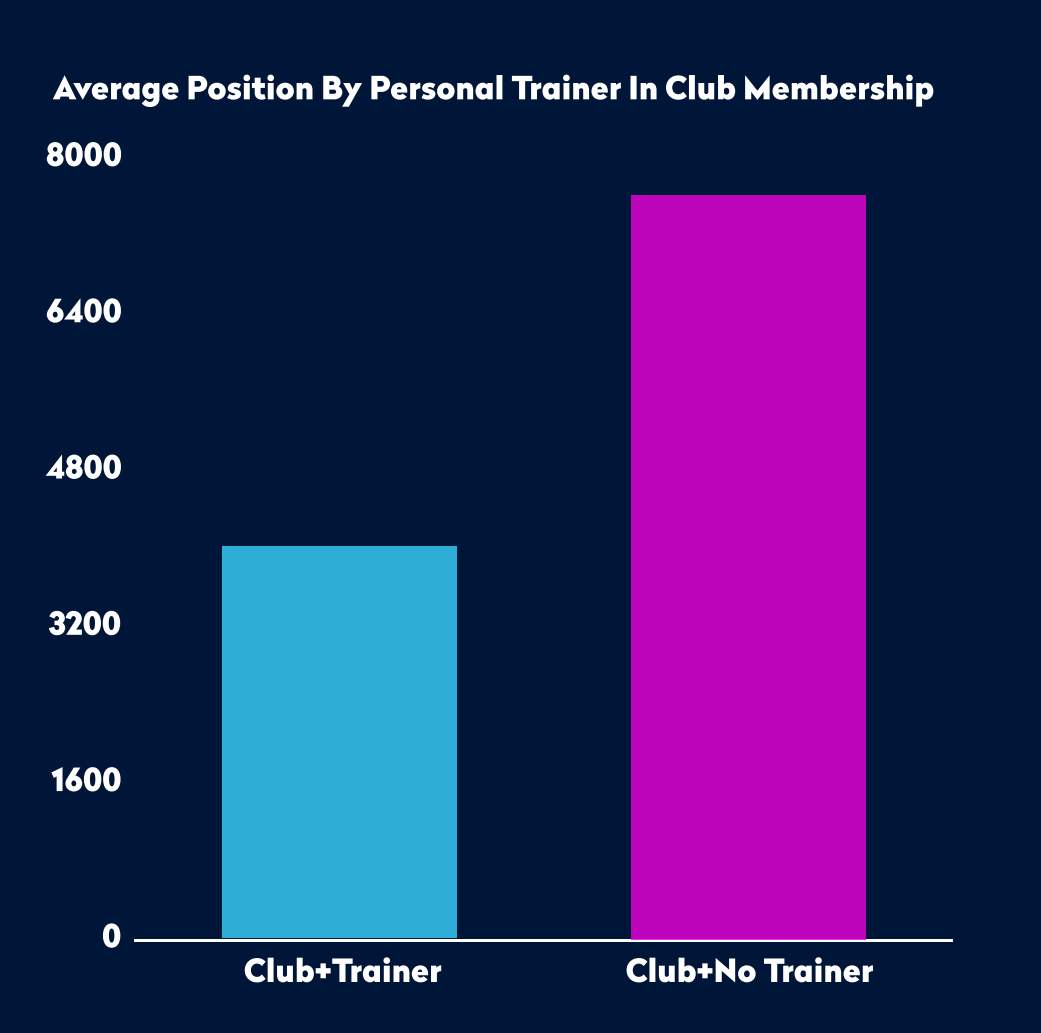
When aggregating all joined data, it is evident that the number of participants with club membership remains significantly lower than those without.

Potential barriers to joining, such as club fees, equipment costs, and training time commitments, may exist; however, the positional statistics clearly demonstrate the undeniable performance and training schedule benefits of club affiliation.



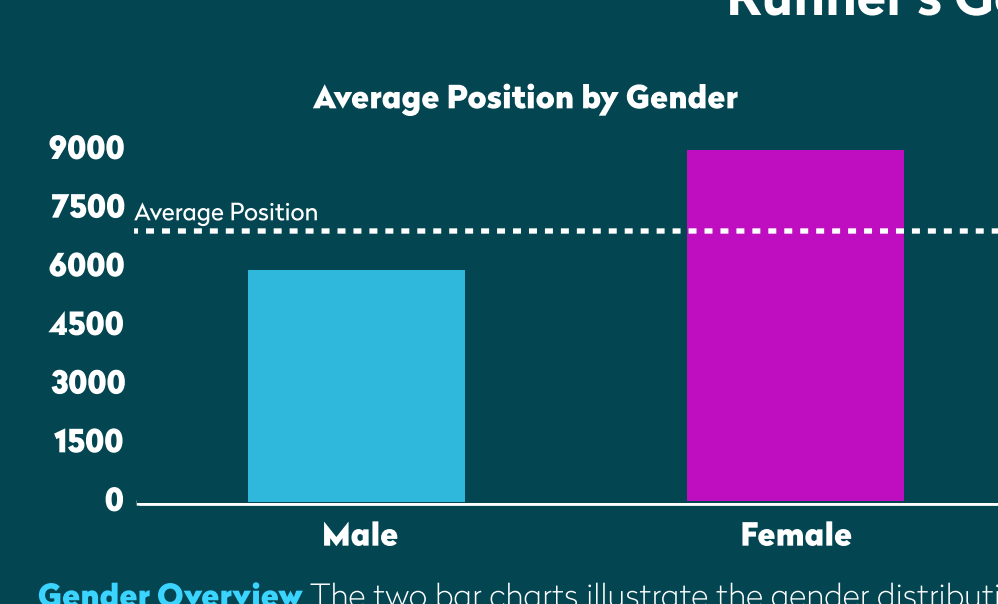
Runner Categories Insights

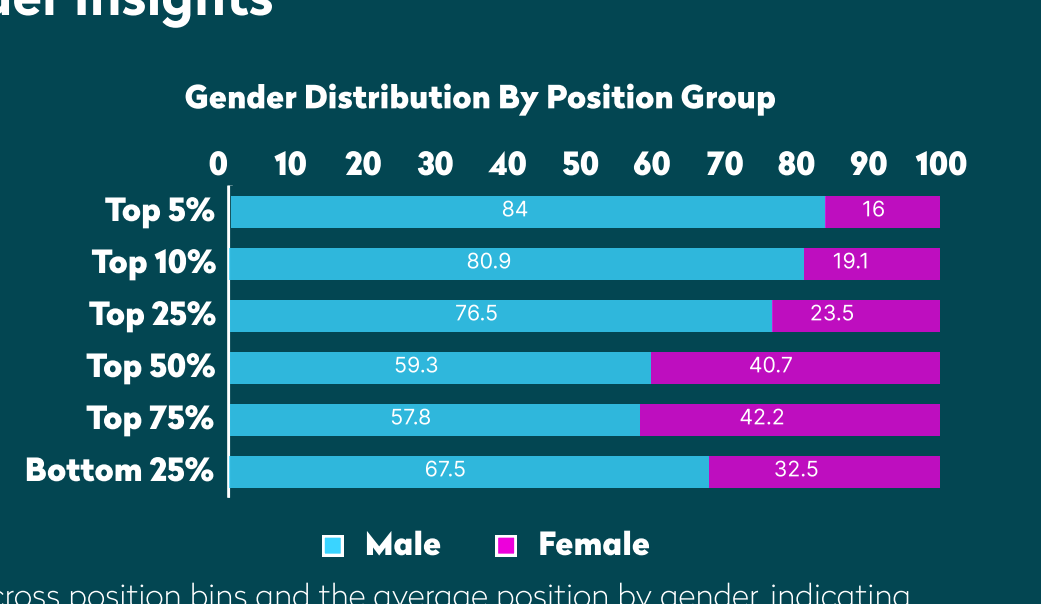




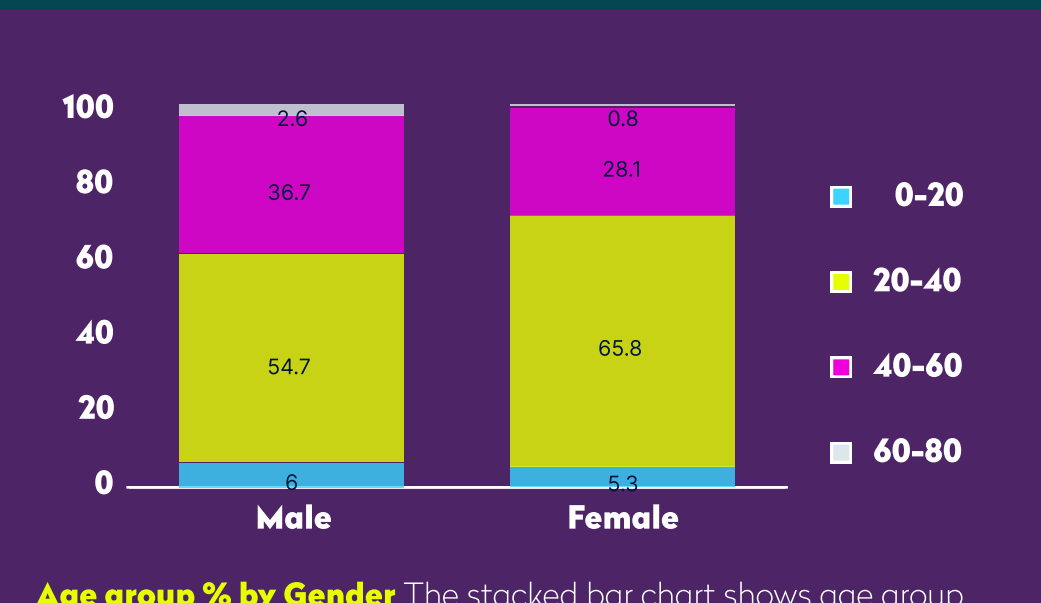
An additional variable, the presence of a personal trainer, is noteworthy. The statistics reveal that the majority of runners lack a personal trainer, with only a small proportion employing one, and most of these individuals also hold club memberships. Comparative analysis of club membership data indicates that runners with personal trainers exhibit superior performance compared to other club members.

Runner's Gender Insights

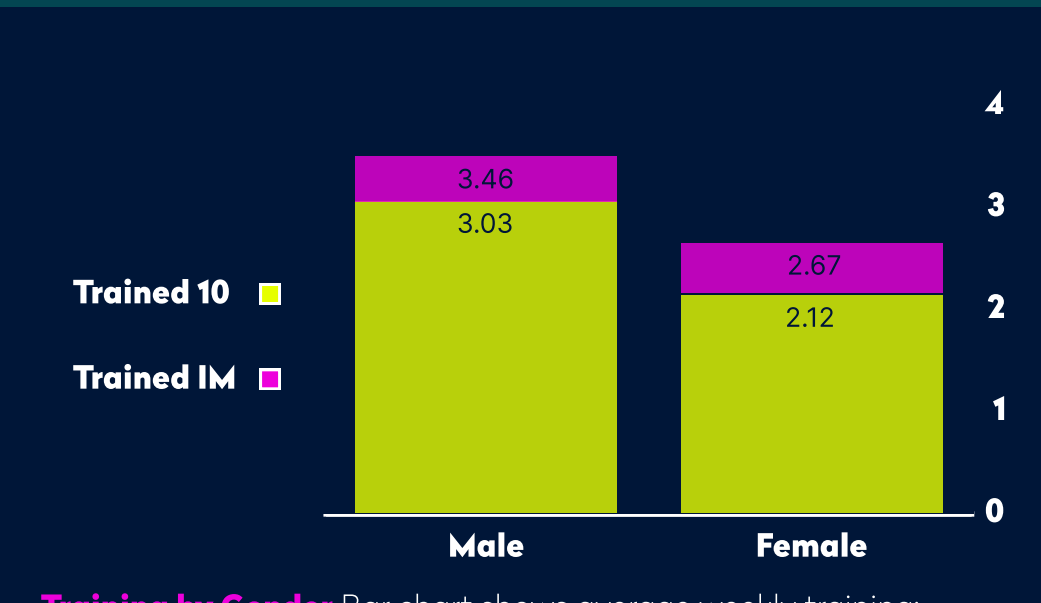




Gender Overview The two bar charts illustrate the gender distribution across position bins and the average position by gender, indicating that males achieve higher positions than females. The data reveals distinct patterns, with males predominantly securing top performance ranks, while females display a varied distribution.

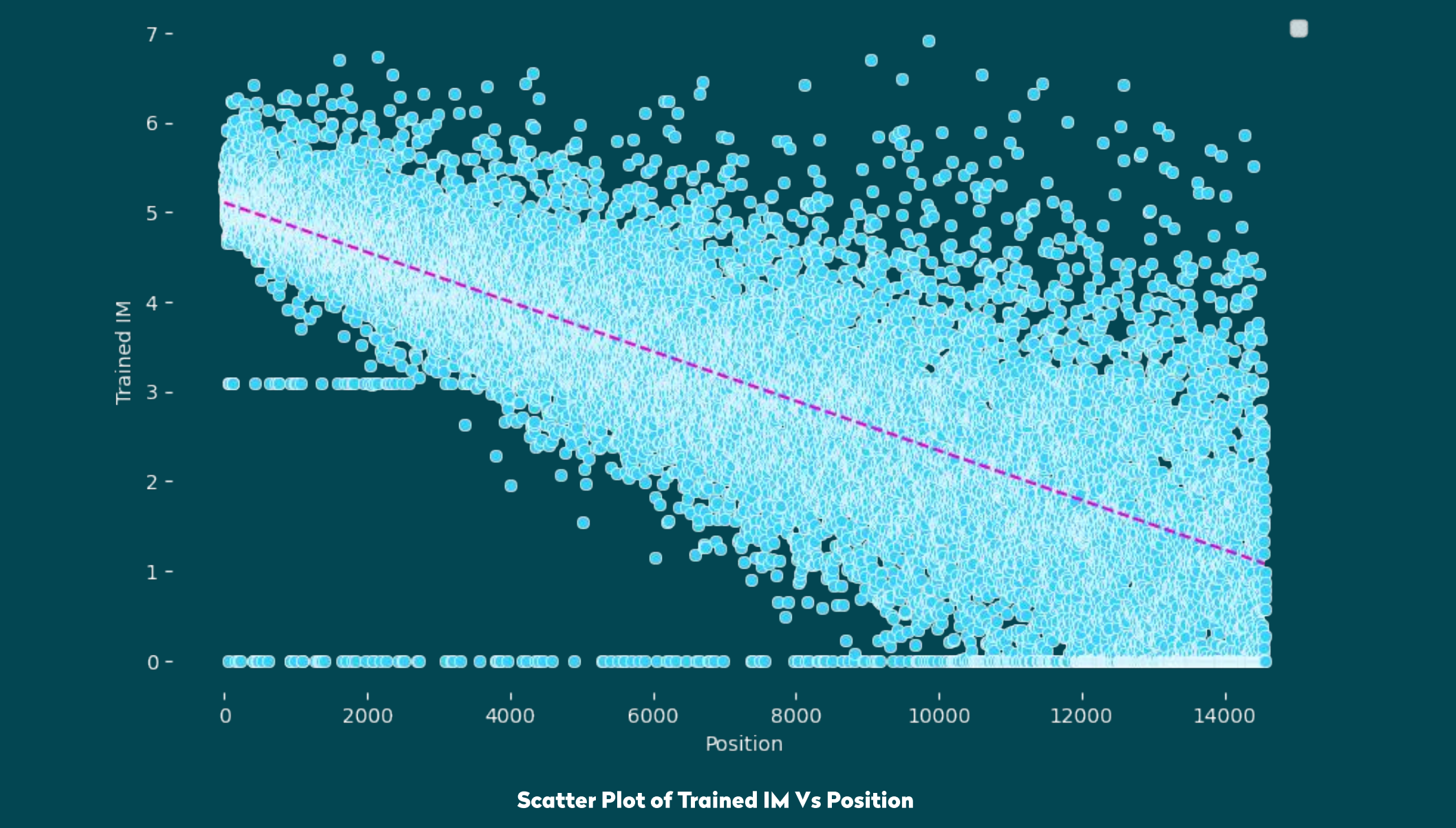


Age group % by Gender The stacked bar chart shows age group percentages by gender, with male have a higher percentage in the 20-40 age group then female, showing the more experience participants in men.



Training by Gender Bar chart shows average weekly training: females (2.12 times 10-week, 2.60 times im-week), males (3.03 times, 3.46 times). Male put more effort in training time, which will boost their performance than female.

Training Insights



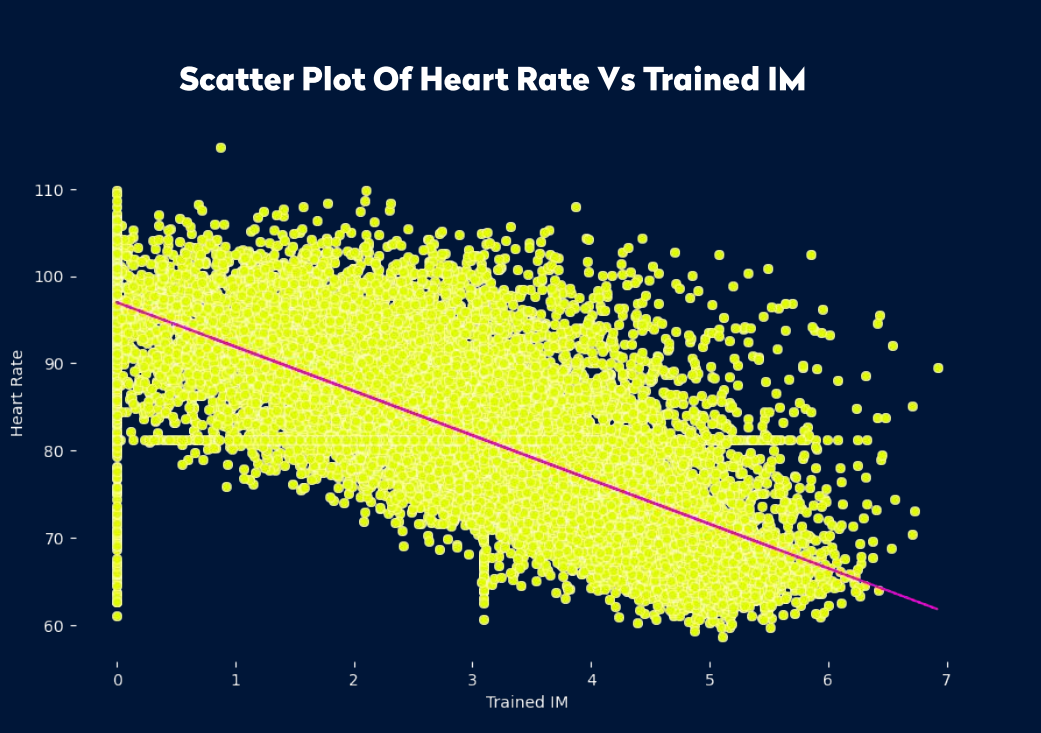
Scatter Plot of Trained IM Vs Position

Scatter Plot of Trained IM Vs Position As observed in the scatter plot, an increase in a runner's training frequency correlates with a higher race position. Upon closer examination, it becomes evident why training impacts a runner's placement.

Training Insights (Cont)

By utilizing the marathon distance and the recorded time of each runner, the speed variable for individual runners can be calculated.

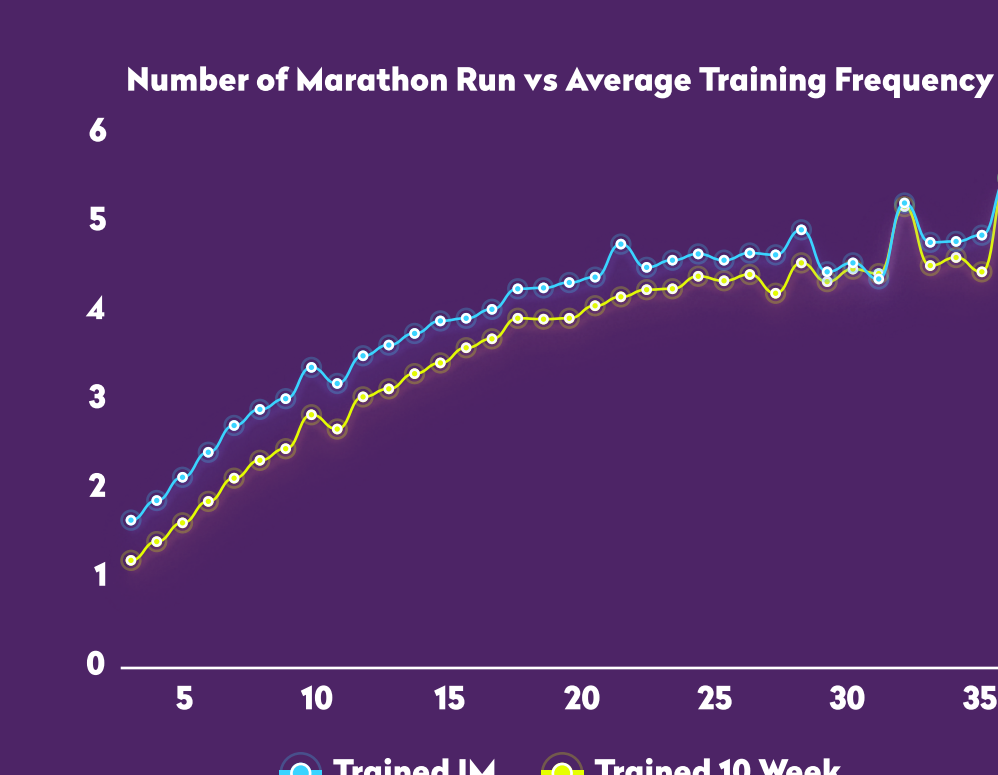
The scatter plot shows speed (2-7 m/s) vs. Trained IM (0-8 times/week), with a positive magenta trend. Dense points (2-5 m/s, 0-6 times/week for IM) suggest correlation; minimal outliers (6-7 m/s, low IM) indicate training boosts speed.



The scatter plot depicts the relationship between Trained IM (0-7) and heart rate (60-110 bpm), with a downward-sloping trendline indicating a negative correlation. A dense cluster of data points between 1-5 IM and 70-100 bpm suggests most observations lie here, showing higher training linked to lower heart rates and more endurance during the race.

Factors Influence Training

In the context of athletic performance studies, variables such as gender, the total number of marathons completed, and participants' age are critical factors significantly influencing individuals' training frequency and intensity.



The graph shows the relationship between marathons run and average training frequency (days) for Trained IM (blue) and Trained 10 Weeks (orange). Both rise from 1 to 5 days (0-35 marathons), with IM exceeding 10 Weeks, showing greater variability and stronger correlation with experience.

The bar chart illustrates average Trained IM increasing from 3.0 to 4.0 across age groups, with the 60-80 group peaking at the highest. Limited training time for 20-60 groups may result from work and family demands, while health issues may affect older adults.

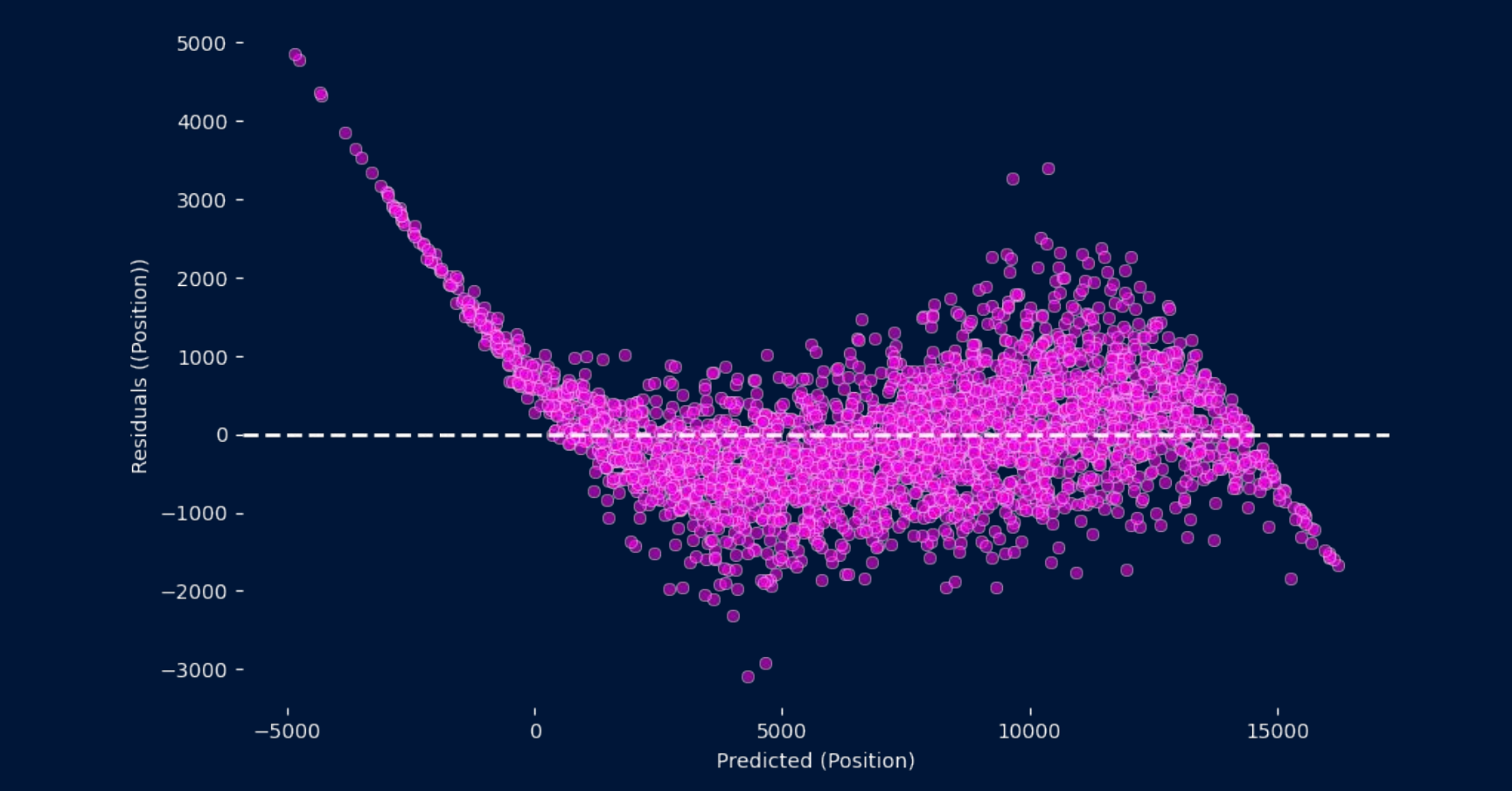
Multiple Linear Regression Model

This multiple linear regression model predict marathon time outcome (in seconds), using predictors such as speed, VO2_max, n_marathons_run, etc. Negative coefficients (e.g., speed: -1756.21, VO2_max: -680.61) indicate improved performance (faster times), while positive coefficients (e.g., heart_rate: 1253.61, has_trainer: 333.67) suggest slower times.

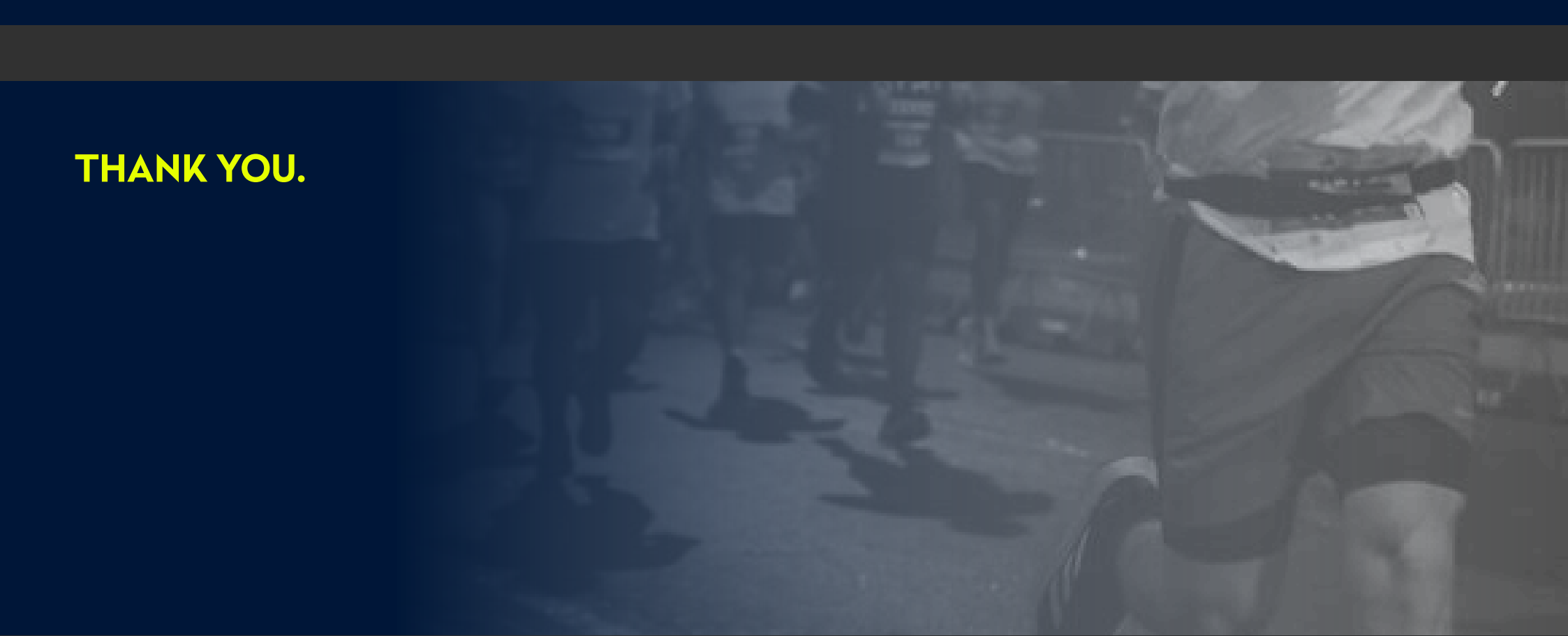
```
                            OLS Regression Results
==============================================================================
Dep. Variable:               position   R-squared:                       0.965
Model:                            OLS   Adj. R-squared:                  0.965
Method:                 Least Squares   F-statistic:                 2.880e+04
Date:                Sun, 10 Aug 2025   Prob (F-statistic):               0.00
Time:                        05:03:50   Log-Likelihood:                -93964.
No. Observations:               11612   AIC:                         1.880e+05
Df Residuals:                   11600   BIC:                         1.880e+05
Df Model:                          11
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const            7191.9529     10.914    658.989      0.000    7170.560    7213.345
gender             87.2917     15.863      5.503      0.000      56.197     118.386
trained_10_week   -77.6886     14.596     -5.323      0.000    -106.299     -49.079
trained_im       -229.1957     15.162    -15.116      0.000    -258.916    -199.475
has_trainer       333.6730     33.790      9.875      0.000     267.440     399.906
cadence          -116.1479      8.651    -13.425      0.000    -133.106     -99.190
bmi              -183.6744      9.777    -18.787      0.000    -202.839    -164.510
n_marathons_run  -238.4216     10.513    -22.678      0.000    -259.029    -217.814
VO2_max          -680.6071     16.703    -40.748      0.000    -713.347    -647.867
heart_rate       1253.6099     18.120     69.184      0.000    1218.092    1289.128
speed           -1756.2054     20.763    -84.584      0.000   -1796.904   -1715.507
join_club         117.1112     21.359      5.483      0.000      75.244     158.978
==============================================================================
Omnibus:                     1896.204   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6142.419
Skew:                           0.832   Prob(JB):                         0.00
Kurtosis:                       6.151   Cond. No.                         11.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

P-values for each independent variable:
```

const	0.000000e+00
gender	3.817845e-08
trained_10_week	1.041426e-07
trained_im	3.888021e-51
has_trainer	6.576028e-23
cadence	8.665729e-41
bmi	1.367582e-77
n_marathons_run	1.906119e-111
VO2_max	0.000000e+00
heart_rate	0.000000e+00
speed	0.000000e+00
join_club	4.268135e-08

Multiple Linear Regression Model (Cont)



Residual Plot The residual plot displays the differences between observed and predicted positional values, with the x-axis showing predicted positions (0 to 15,000) and the y-axis indicating residuals (-5,000 to 5,000). The scatter shows greater variability at higher predictions, suggesting model nonlinearity, which may require further refinement due to the some missing info at the cleaning data step.



THANK YOU.