Optimizing Track and Field Performance through Data Science

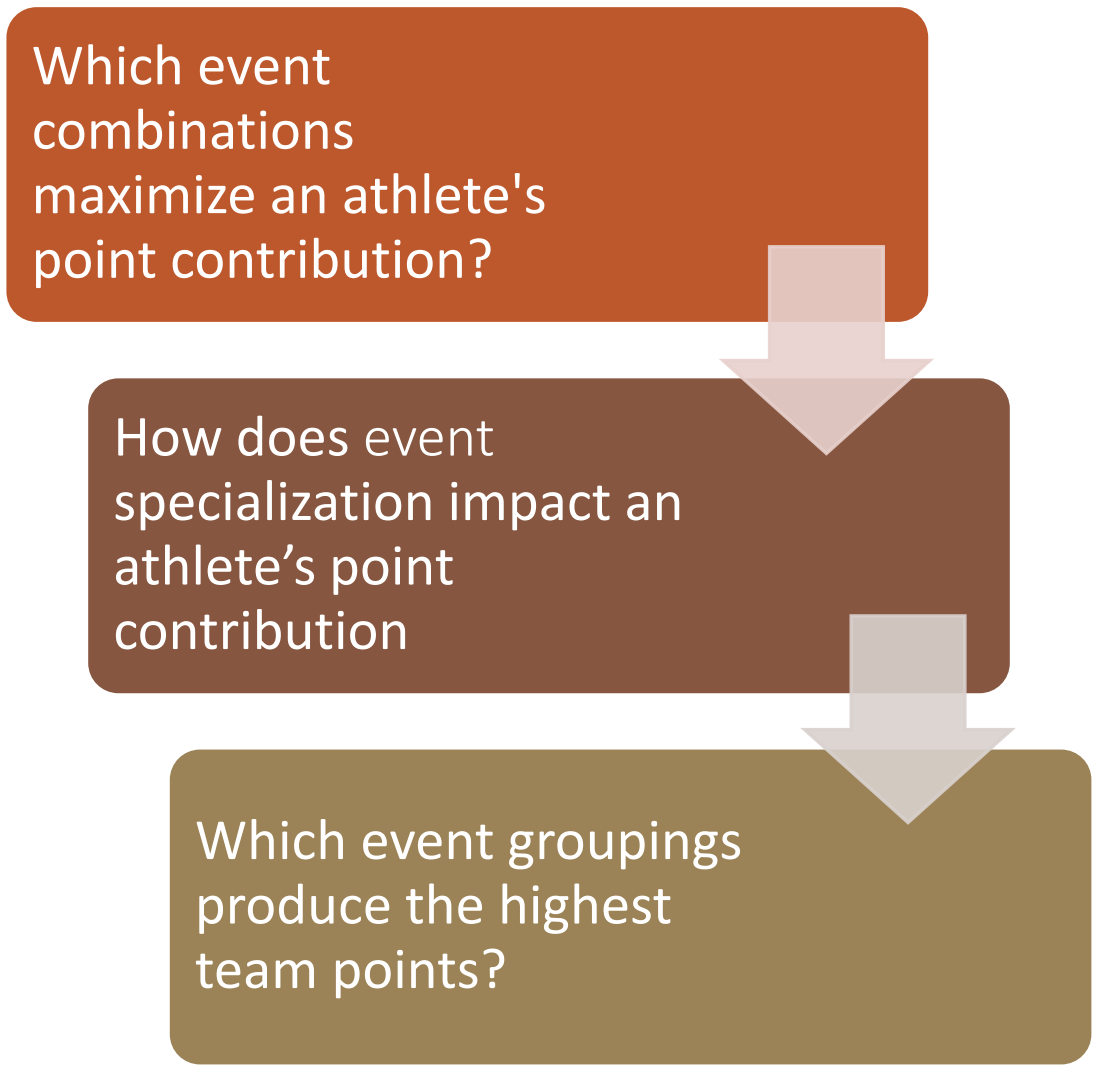## MAXIMIZING ATHLETE AND TEAM SUCCESS WITH DATA-DRIVEN INSIGHTS

Background and Motivation

Track and field provides rich quantifiable data.

Strategic event selection crucial due to 3-event limit.

Opportunity to transform data into actionable intelligence.

Which event combinations maximize an athlete's point contribution?

How does event specialization impact an athlete's point contribution

Which event groupings produce the highest team points?
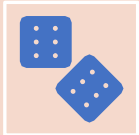
Research Questions

Data Sources and Methodology

Source: High school meet results collected through timing system.
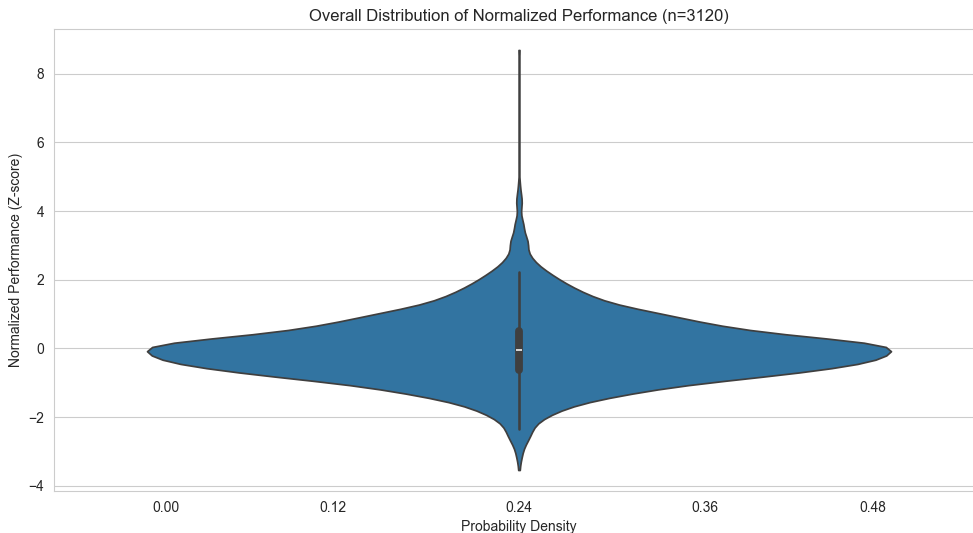
Data includes athlete info, performances, placements, and points from ten track meets.

Methods: EDA, clustering, correlation analysis, concurrence analysis

Overall Athlete Performance Distribution

Overall Distribution of Normalized Performance (n=3120)

• Performance distribution is sharply centered around the mean (Z=0), confirming strong competitive parity.

• A clear right skew suggests a small group of athletes significantly underperform relative to peers.

• The plot reveals a dense cluster near the mean with a long positive tail, indicating the presence of statistical outliers.

• Even modest performance improvements could move athletes ahead of many peers in this compressed performance space.

Distribution of Normalized Performance for Top 6 Events

Performance by Top 6 Events
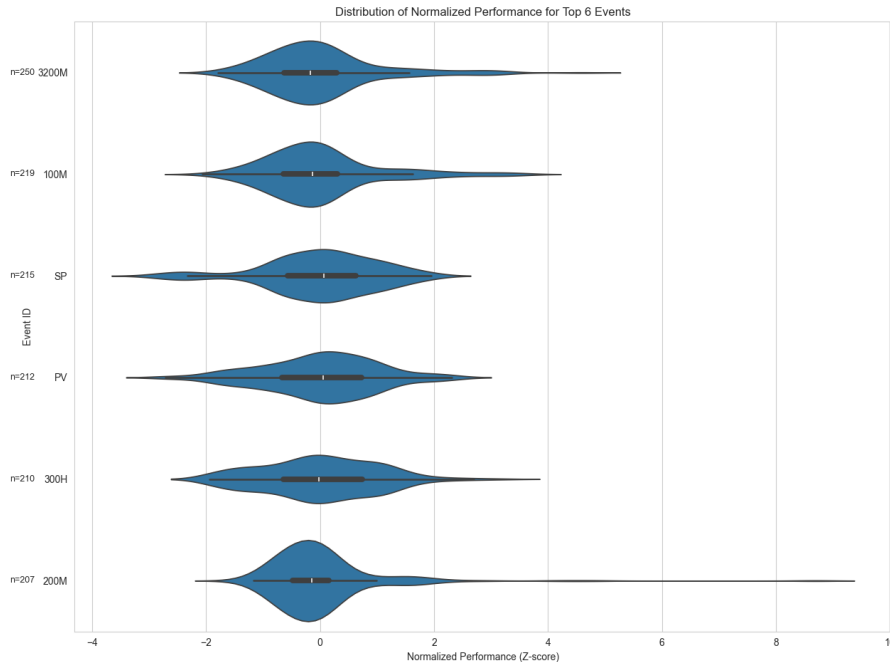
•Sprint events (100M, 200M) display tight, centralized performance clusters, reinforcing how small gains are critical in fast-twitch explosive events.

•Distance events (3200M) show broader and more asymmetrical distributions, reflecting larger variability in endurance performance.

•Hurdle and middle-distance events (300H) demonstrate moderate spread, indicating variability in both speed and technique.

•Shot Put and Pole Vault (SP, PV) shows right-skewed variation with outliers — perhaps from mismatched field strength or technical inconsistencies.

•Overall, performance consistency is higher in technical and power-based events, while endurance-based disciplines exhibit more natural variation.

Performance by Grade Level

Normalized Performance Distribution by Grade Level
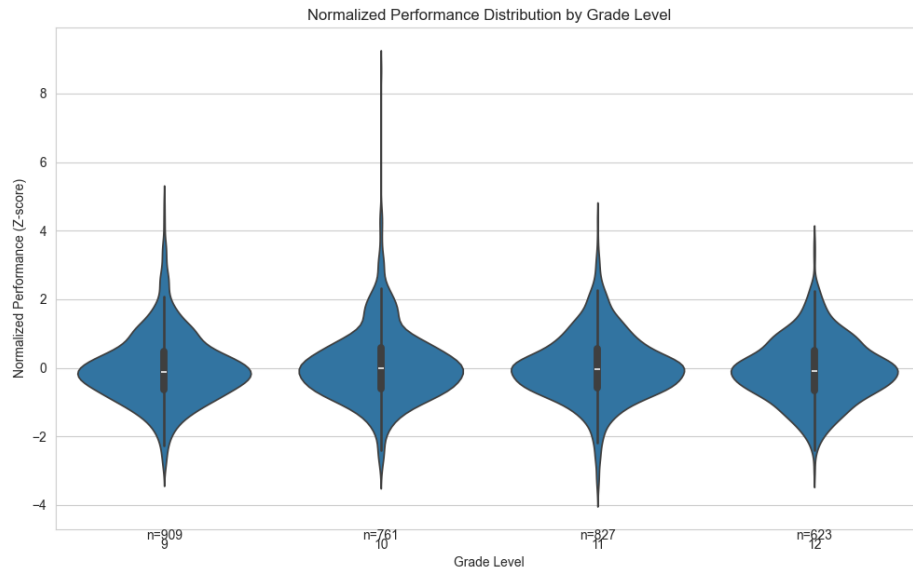
- Median performance remains tightly centered near the mean (Z=0) across all grades.

- Seniors (grade 12) still show a slight edge, but the improvement is marginal rather than substantial.

- Performance variability remains high at all levels, reinforcing that standout athletes emerge early and late alike.

- Distributions are similarly shaped, suggesting development potential exists at every stage.

- Coaching strategies should focus on consistent development rather than expecting peak performance only in upper grades.

Performance vs Points Earned

•Track events show strong negative correlations (faster = more points); field events show positive correlations (higher/farther = more points).

•Top performances cluster tightly around maximum points, emphasizing the reward for excellence.

•Strategic event selection is key: athletes near average performance levels can still contribute points in less competitive events.

•Different events offer different "point efficiency" — small improvements in sprints yield larger point gains than in distance races.

Performance vs Points by Event (Potential Optimization)
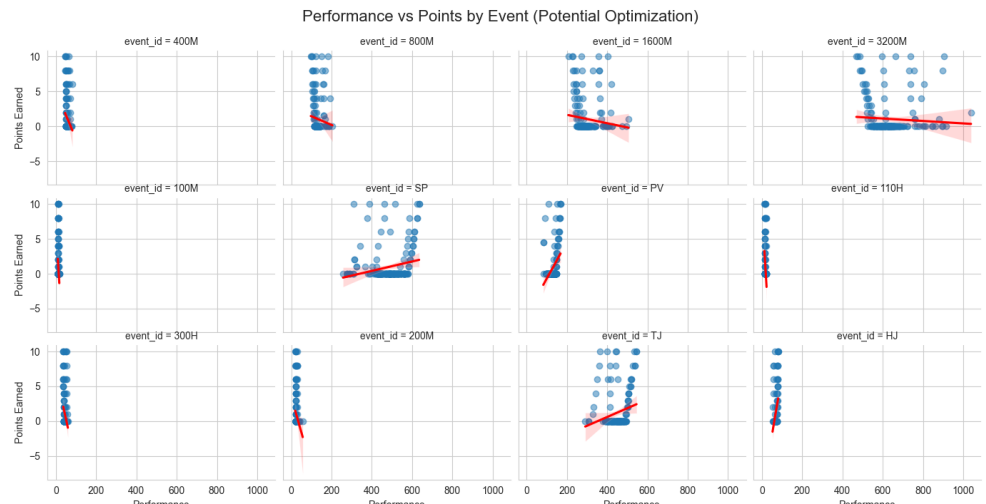
Distribution of Events per Athlete

Number of events
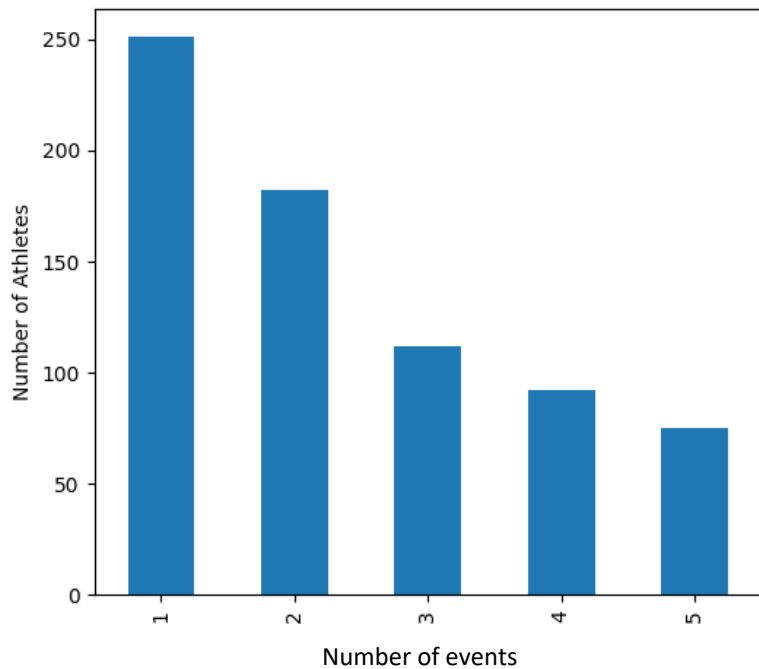
- Most athletes specialize in **only 1–2 events**.
- Athlete count sharply **drops after 3 events**.
- **Few athletes** compete in **5+ events** — rare versatility.
- **Coaching Focus:** Prioritize specialization; identify rare multi-event athletes for strategic advantage.

Total Points by Event Category

• Relays generate the most total points, offering high strategic value.

• Jumps and Distance events contribute significantly to team totals.

• Sprints are impactful but slightly behind in total contribution.

• Throws and Hurdles yield lower point totals overall.

• Coaching Strategy: Focus training and athlete development on relays and jumps to maximize team scoring potential.
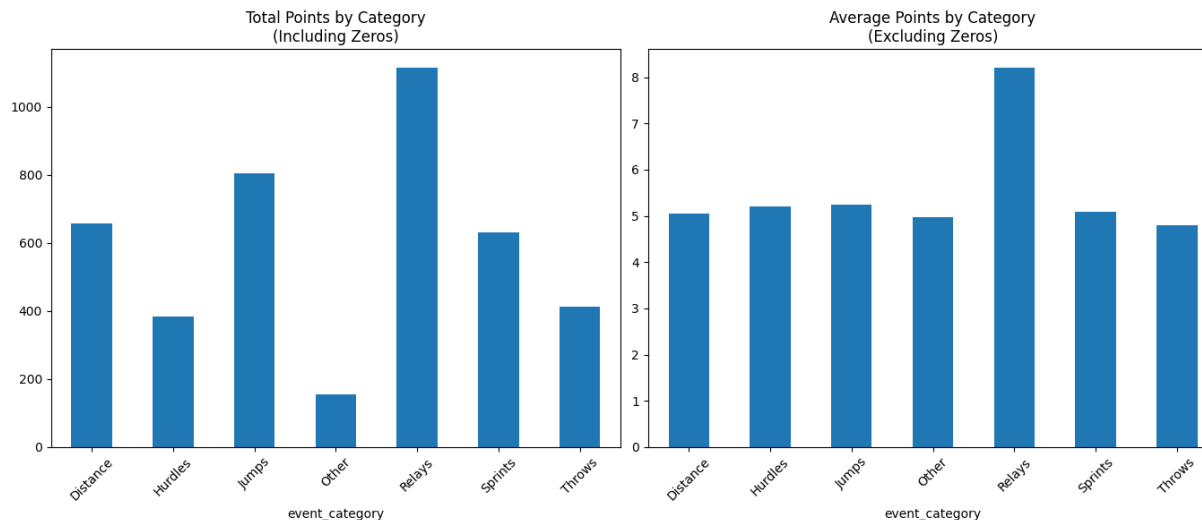
• Sprints and Distance events show consistent performance with low variability.

• Jumps and Throws have wider spreads, indicating diverse athlete skill levels.

• Relays show moderate variability due to team composition.

• Outliers in all categories highlight exceptional performers worth developing.

• Variability levels differ by event type — key for tailoring coaching strategies.
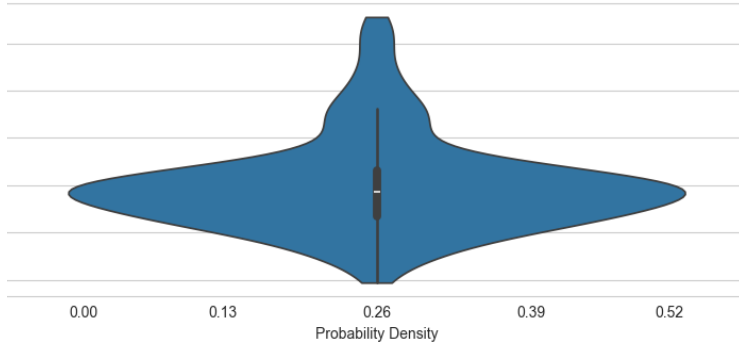
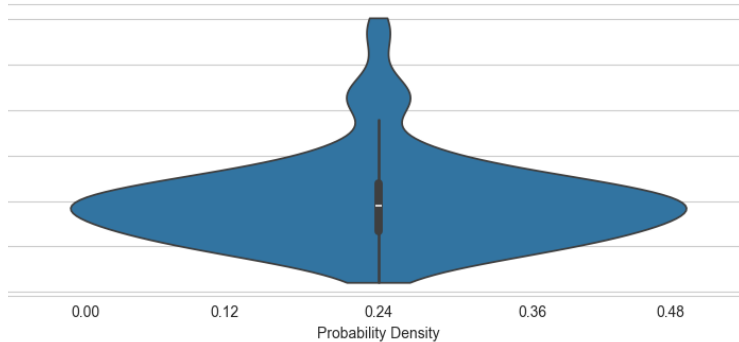# Performance by Event Category

Performance Distribution by Event Category (n=3120)
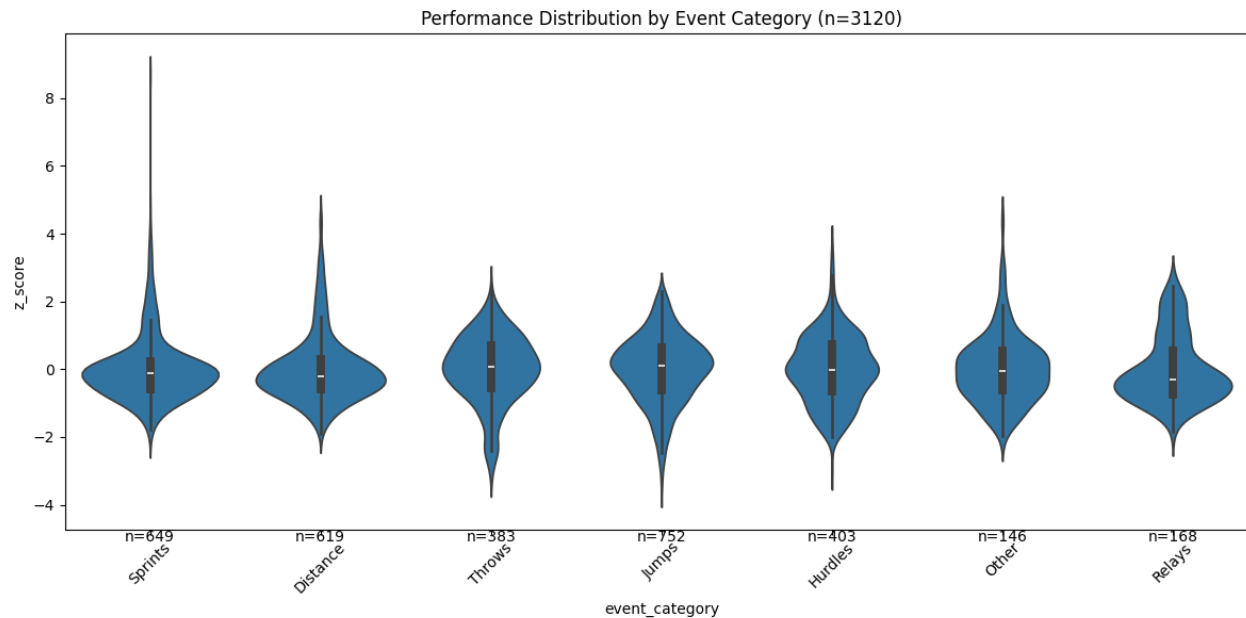
Sprints – Event Performance Distributions
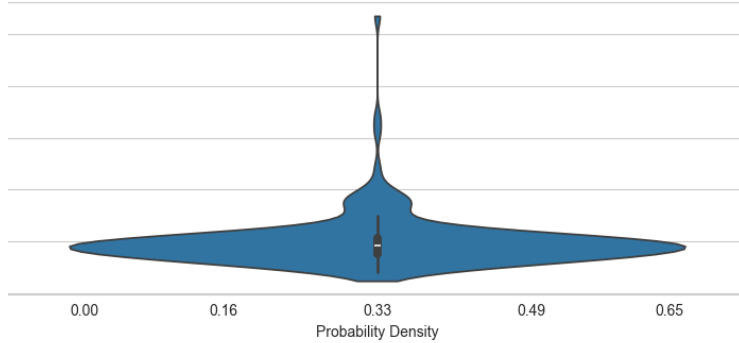
Distance – Event Performance Distributions
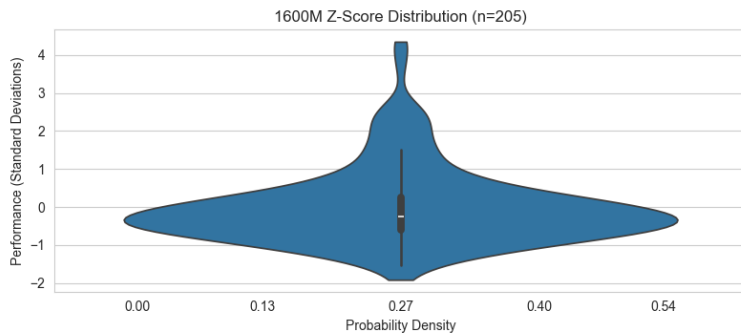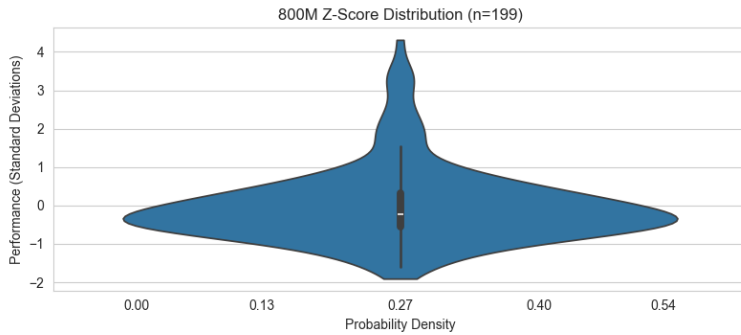
Hurdles — Event Performance Distributions

Relays — Event Performance Distributions

Jumps – Event Performance Distributions

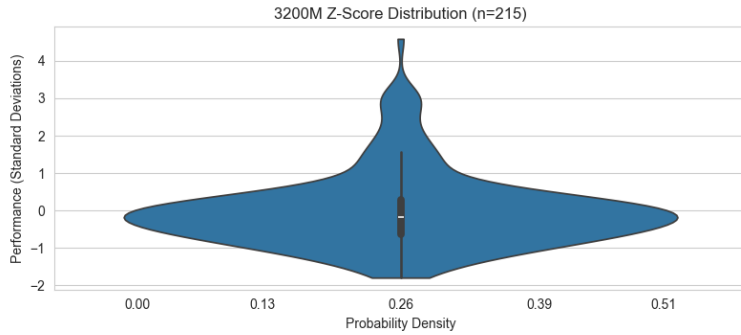SP Z-Score Distribution (n=207)

DT Z-Score Distribution (n=176)
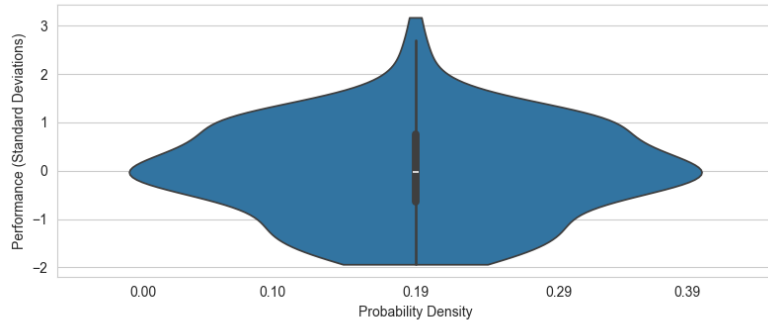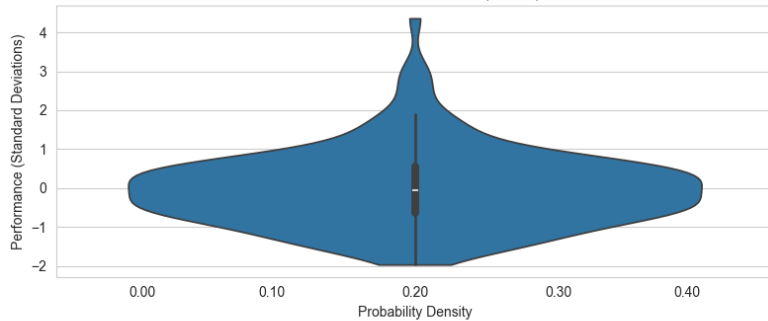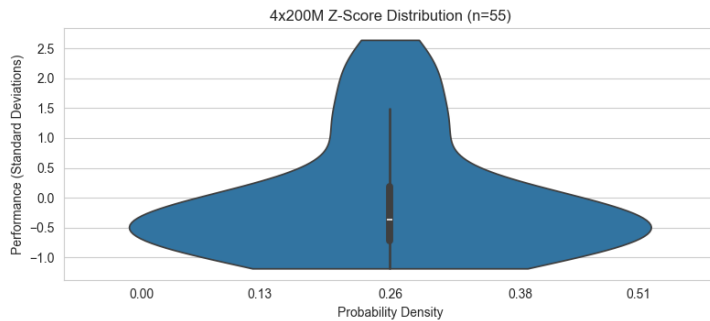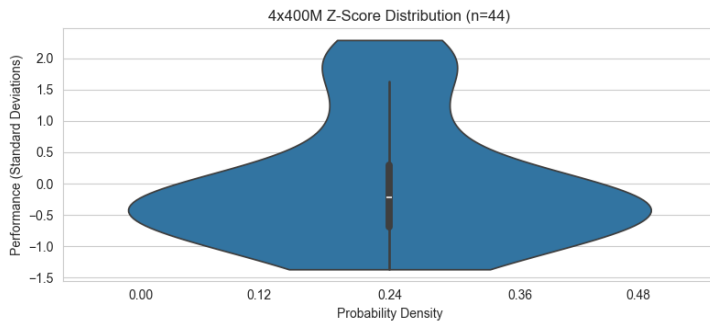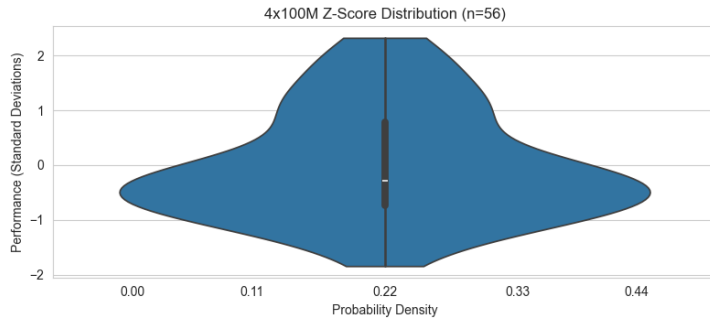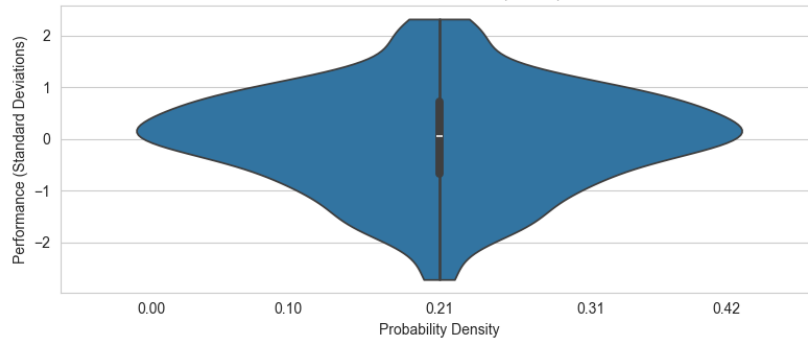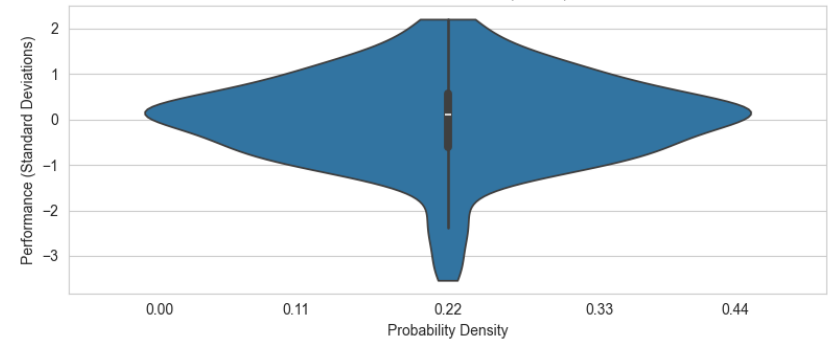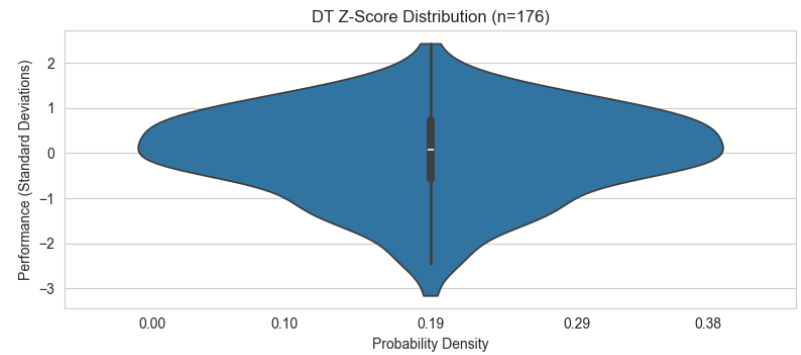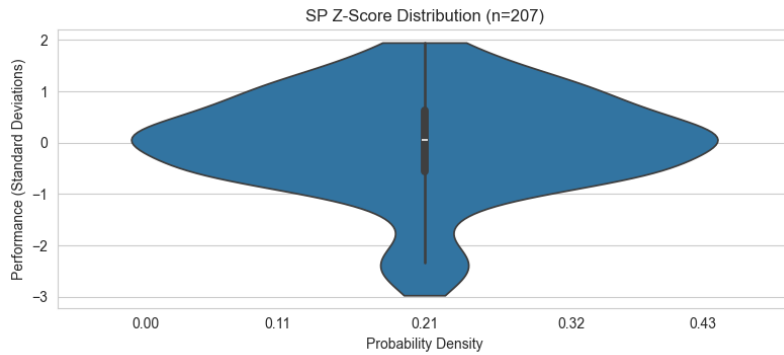
Throws – Event Performance Distributions

Top Event Co-Occurrence Matrix

• The **100M and 200M** sprint events are the most commonly paired, indicating a strong sprinting specialization among athletes.

• **Distance events** like the **800M**, **1600M**, and **3200M** show high co-occurrence, suggesting many distance runners participate across multiple long races.

• **Jumps and Throws** (e.g., PV, SP, HJ) show lower co-occurrence with track events, indicating these athletes tend to specialize within their category.

• **Strategic Insight**: Understanding common event pairings can help in designing training programs and maximizing athlete point contributions across meets.

Co-occurrence Matrix for Top Events

Event Co-occurrence Matrix
(How often athletes compete in both events)
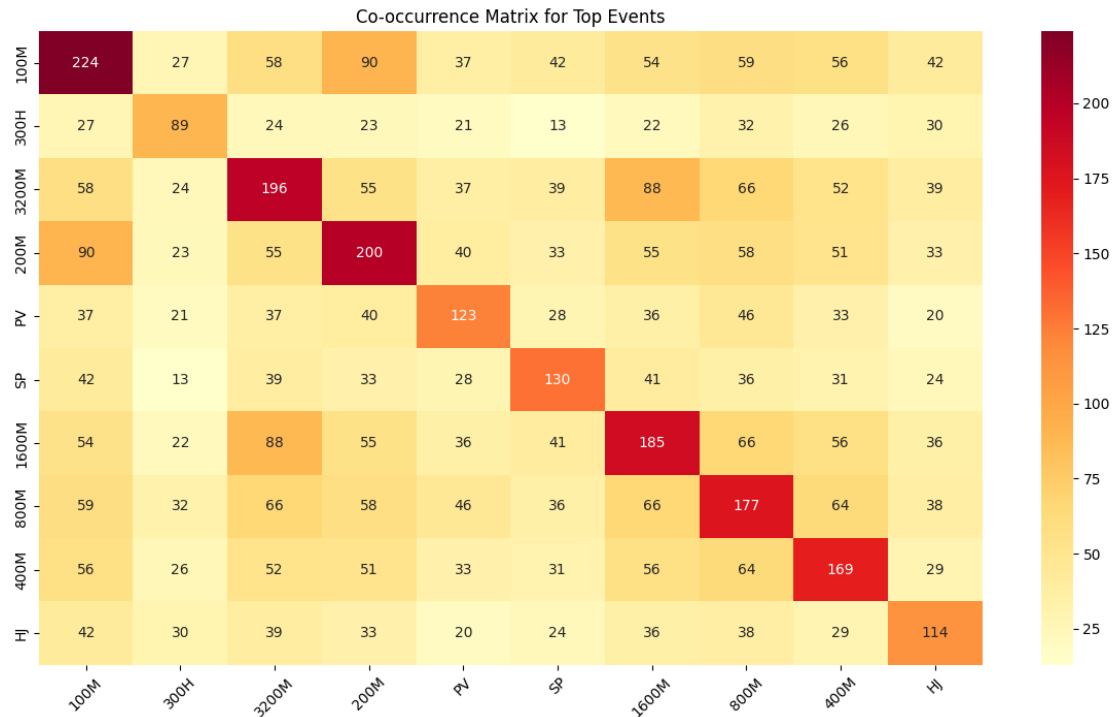
Event Co-Occurrence Matrix (Full Dataset)
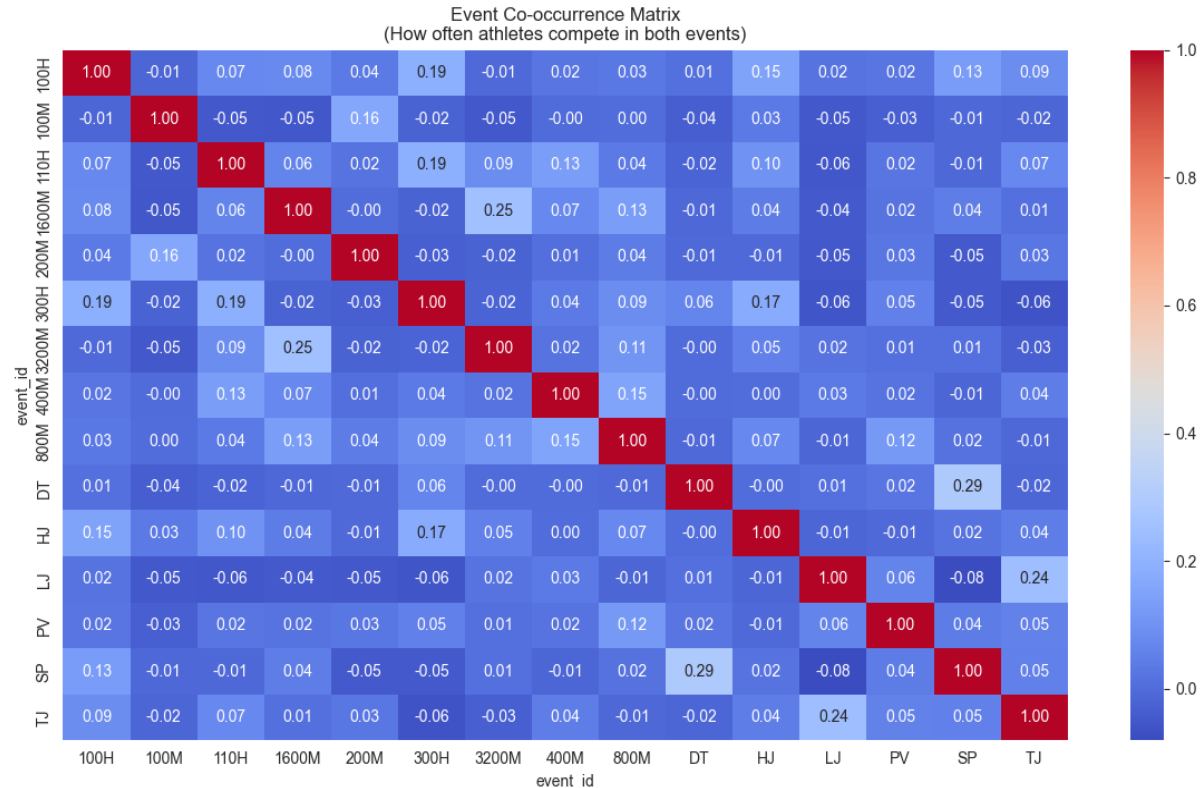
- **Specialization dominates:** Athletes overwhelmingly compete within their primary event category, with minimal crossover.

- **Sprints, distance races, and field events** (jumps/throws) display weak or even negative co-occurrence with one another, highlighting strong separation between disciplines.

- **Minor positive relationships** exist within categories—for example, **Long Jump (LJ) correlates modestly with Triple Jump (TJ)** (0.24), and **Shot Put (SP) with Discus Throw (DT)** (0.29).

- **Negative or near-zero values** between track and field events indicate that athletes rarely cross disciplines (e.g., 100M vs SP = 0.13, LJ vs 800M = 0.04).

- These patterns reinforce the idea that **event-specific training** is critical, and athletes are unlikely to contribute meaningfully across different event types.

Event Performance Correlation
(How similar are normalized performances)

Event Performance Correlation Matrix
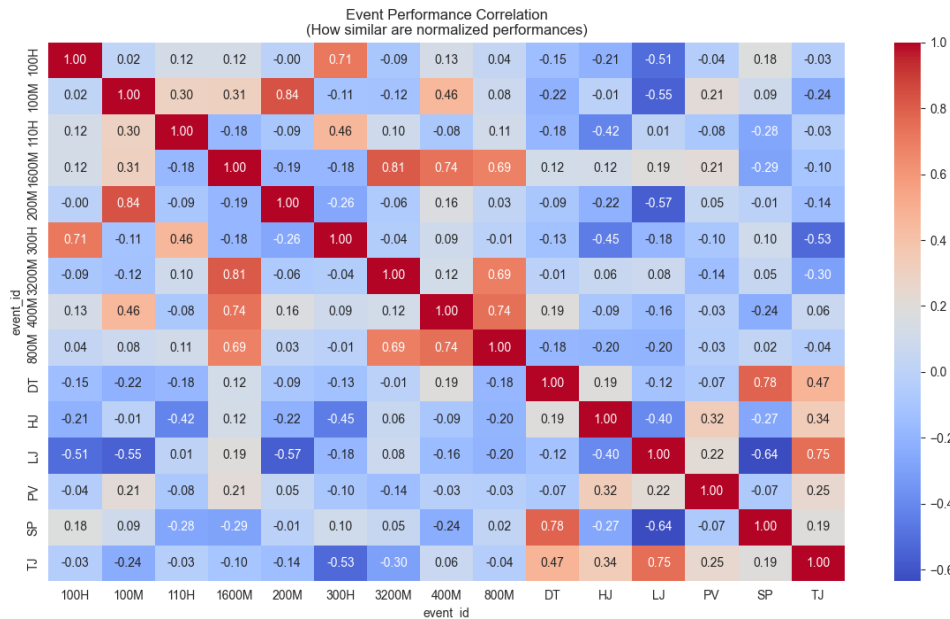
•**Natural Event Clusters Are Clear**:
Sprints like **100M, 200M, 400M, and 300H** form a tight cluster, especially with **300H and 100H** having a high correlation (r = 0.71). Distance events like **1600M, 3200M, and 800M** also correlate strongly (e.g., 1600M ↔ 3200M = 0.81, 800M ↔ 1600M = 0.74).

•**Strong Technical Event Consistency**:
Jumps and throws show high internal correlation:

•**LJ ↔ TJ** = 0.75

•**SP ↔ DT** = 0.78
These suggest shared physical or technical qualities across those events.

•**Negative or Near-Zero Cross-Category Correlations**:
There's little to no performance correlation between track and field events. In fact, some are negative:

•**LJ ↔ 100M** = -0.55

•**SP ↔ 400M** = -0.64
This reflects that excelling in power-based throws or jumps doesn't translate to success in speed/endurance events, and vice versa.

•**Strategic Insight**:
This matrix confirms that athletes typically specialize within a domain. Dual specialization across unrelated disciplines (e.g., sprint + throws) may be inefficient. Coaches can use this to guide event placement and training plans more precisely.

Athlete Clusters (PCA Projection)

Athlete Clustering Based on Event Performance Patterns

Athlete Groups
- Cluster 0: 12 athletes
- Cluster 1: 24 athletes
- Cluster 2: 636 athletes

First Principal Component (12.0% variance)
Second Principal Component (9.9% variance)
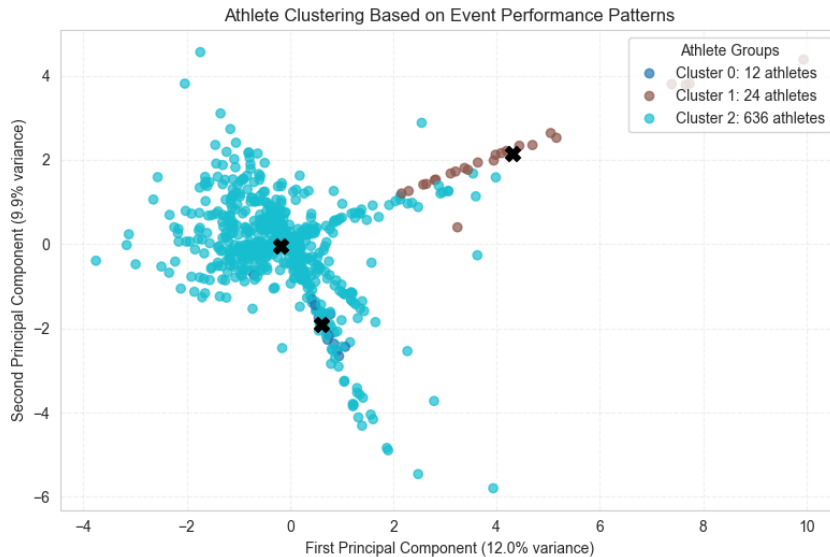
•**Data & Pre-processing**
• 3,124 performances across 672 athletes (excluding relays)
• 4 duplicate entries removed; no other nulls in raw data
• Pivoted to athlete × event Z-score matrix; filled 7,894 missing cells with 0 (average)

•**Clustering Workflow**
• Features standardized (mean 0, σ 1)
• K-means (k=3)—elbow point at k=3
• PCA projection (PC1 12.0% var; PC2 9.9% var) for visualization

•**Cluster Profiles**
• **Cluster 0 (12 athletes):** Hurdles/speed specialists (↑100H, ↑300H)
• **Cluster 1 (24 athletes):** Distance specialists (↑1600M, ↑3200M, ↑800M)
• **Cluster 2 (636 athletes):** Generalists/moderates (near-zero Z's)

•**Key Insights**
• "Arms" from main cloud reveal micro-specializations (e.g. jumps)
• Outliers may be multi-event standouts or data artifacts
• Tailor coaching: triage athletes into generalist, sprint/hurdle, or distance tracks

Conclusion

Specialization Drives Success

•Athletes naturally group into three archetypes—generalists, sprinters/hurdlers, and distance runners—with very little cross-discipline crossover.

•PCA/K-means clustering confirmed these archetypes across 3,120 performances.

Point-Efficiency Varies by Event

•Sprints reward tiny improvements (±0.1 s → 1–2 extra points); distance and field events have shallower point curves.

•Relays and jumps contribute the most total points; throws and hurdles lag behind.

Optimal Event Pairings

•100 M↔200 M and 800 M↔1600 M↔3200 M pairings maximize athlete point contribution.

•Field athletes tend to specialize narrowly—little benefit in cross-training between throws and jumps.

Data-Driven Athlete Placement

•Use clustering & co-occurrence insights to assign athletes to their natural archetype from day one.

•Focus coaching on marginal gains for highly competitive events and technical consistency in others.

Next Steps

•Build a **predictive model** to recommend event assignments based on early-season performances.

•Develop **season-long forecasting** to identify rising stars and optimize relay team selection.