EDA Plan for EURO 2024 Football Event Data

1. Data Structure and Preparation

This dataset includes all events from 51 EURO 2024 games (~3,500 per match). Each event has:
- Timestamp (minute, second)
- Event type (pass, shot, duel, etc.)
- Player and team
- Spatial info (x, y)
- Descriptive text or category fields

Why: Understanding structure enables proper time series aggregation, spatial mapping, and modeling.

2. Event Distribution and Match Summary

Group and visualize events by match, team, and player.
- Count passes, shots, fouls, cards per team/match
- Histograms and bar charts for high-level view

Why: Understand tempo, discipline, style, and volume.

3. Time-Based Event Analysis

Use rolling windows (3 or 5 minutes) to track match momentum:
- Number of events, passes, shots, xG over time
- Can use `.rolling()` or group by fixed intervals

Why: Detect shifts in tempo or control (e.g., post-substitution).

4. Autocorrelation & Lag Analysis

Use ACF/PACF on time-binned series (e.g. shots/min).
- ACF detects persistence (e.g., pressure over time)
- 3-min shows micro-trends, 5-min for sustained pressure

Why: Understand repeatable behavior and response windows.

5. STL Decomposition & Seasonality

STL = Seasonal + Trend + Residual breakdown.
- Can apply to rolling xG, passes, shots per game
- Helps detect peak phases and residual volatility

Why: Uncover match rhythm and tactical timing patterns.

EDA Plan for EURO 2024 Football Event Data

6. Spatial Event Analysis (x/y)

Map passes, shots, duels using heatmaps or hexbins.

- Segment by outcome or player role
- Compute average position or zone dominance

Why: Tactical interpretation depends on space control.

7. Outlier Detection

Use IQR or z-score to identify volume spikes or extremes.

- Example: player with 40 passes in 5 minutes = not error, but pattern.
- Transform with log/sqrt if distribution skewed

Why: Investigate; don't always correct. Outliers may reflect real phenomena.

8. Missing Data

Many columns are null by design (e.g., no shot_body_part on a pass).

- Use context to decide: Null = not occurred, not missing
- Flag columns (`has_commentary = description.notna()`)

Why: Don't impute incorrectly - interpret based on domain logic.

9. Transformations & Normalization

Log, sqrt, or Box-Cox for numeric skew (e.g., distance, duration).
Normalize when comparing across players/teams with different volume.

Why: Helps models and comparisons behave consistently.

10. Categorical Variable Analysis

Use value_counts, groupby, crosstab:

- event_type vs team
- outcome vs player role
- duel_type vs time period

Why: Tactical patterns often live in these cross-relations.

11. Text Description Fields

Descriptions (e.g., 'Long ball from deep right wing') can:

- Be tokenized and analyzed for pattern detection
- TF-IDF/N-gram to uncover key recurring strategies

Why: Rich tactical context - adds nuance to categorical features.

12. Position/Role Analysis

Group players by position: Defender, Midfielder, Forward
- Compare shots, xG, passes, fouls, pressure events
- Use boxplots, ANOVA, radar charts

Why: Understand positional responsibilities and tactical usage.

13. Home/Away or Context Effects

Though EURO is neutral venue, you can:
- Track impact of match order, extra rest days, weather
- Segment stats by stage (group vs knockout)

Why: Hidden environmental/contextual bias may explain variance.

14. Stationarity Testing

Use ADF (Augmented Dickey-Fuller) to test if your time series has constant mean and variance.

Why: Required for ARIMA and other time series forecasts. Non-stationary data should be differenced.

15. Granger Causality

Does time-lagged xG predict actual goals?
- Use Granger causality test for lagged prediction
- Example: Does shot count at t-1 predict goals at t?

Why: Helps discover cause-effect relationships over time.

16. Rolling xG and Momentum

Plot xG per minute using `.rolling()`
- Combine with shots on target, possession bursts
- Overlap with match events (subs, red cards)

Why: Measures threat level over time - tactical impact of changes.

17. Event Sequences & Chains

Track sequences like:
- Pass Pass Shot
- Recovery Dribble Foul

Use `.shift()` and `.groupby()` to create chains.

Why: Reveals behavior patterns, especially for buildup or pressing.

18. Player/Team Clustering

Use PCA or t-SNE on event vectors per player/team.
Cluster based on behavior:
- Short pass %
- xG chain involvement
- Pressure actions

Why: Classify roles or discover tactical archetypes.

19. Text-Categorical Hybrid Insight

Combine structured (event_type) + unstructured (description)
- Group by keyword frequency per outcome
- Label encode categories and model outcomes

Why: Mixed-type analysis adds predictive power and depth.

20. Hidden Pattern Discovery

Examples:
- More backward passes under pressure use pass direction + pressure flag
- Substitutions increase tempo compare pre/post substitution segments
- Specific players trigger pressing detect chain events after their action

Tools:
- Groupby + lag
- Event window slicing
- t-SNE, clustering

Why: These are unseen in aggregate stats but matter tactically.