I’m building an **IPL Strategy Dashboard (Streamlit + GitHub portfolio project)** and I want to continue from my current cleaned/processed data pipeline.

✅ Goal (Complete + Categorized)

My end goal is to build a **complete IPL analytics + strategy dashboard** using ball-by-ball data, match metadata, teams, and player role classification — and then optionally extend it with ML features only after analytics is stable.

---

1) 📦 Data Foundation (Must be solid)

- Use the processed datasets only (raw untouched)
- Validate schema + uniqueness (ball keys)
- Confirm no unexpected nulls in key columns
- Confirm role coverage for batter and bowler
- Confirm team id/name mapping is consistent across files

---

2) 🏏 Match-Level Analytics

- Match summary (match_id, season, date, venue, teams, winner)
- Toss impact analysis (decision vs outcome)
- Win patterns (by runs / by wickets)
- Venue-level match trends
- Season trends (matches per season, result distribution)

---

3) 🎯 Innings & Over-Level Analysis

- Powerplay / middle overs / death overs scoring patterns
- Over-by-over run rate and wicket impact
- Inning momentum trends (runs + wickets progression)
- Team phase performance comparison

---

4) 👤 Batter Performance Analytics

- Total runs, balls faced, strike rate
- Boundary analysis (4s/6s if derivable)
- Consistency metrics (match-wise contribution, average runs per match)
- Performance by phase (PP / Middle / Death)

- Batter performance vs team / venue / season
- Best batters vs specific teams (opposition analysis)

---

5) 🎳 Bowler Performance Analytics

- Balls bowled, overs, wickets (if derivable)
- Economy rate (runs conceded per over)
- Dot ball percentage (if derivable)
- Wicket types distribution (if usable)
- Bowling performance by phase (PP / Middle / Death)
- Bowler performance vs venue / season
- Best bowlers vs specific teams (opposition analysis)

---

6) 🧩 Player Role Classification (Final Layer for Strategy)

We already created role logic, but the dashboard must support:

- Role distribution summary (Batter/Bowler/All-Rounder + tiers)
- Role confidence via sample size (matches involved)
- Filters by role + experience tier
- Role validation reports (spot checks for top/edge cases)
- Ability to refine role thresholds later without breaking the pipeline

---

7) ⚔️ Matchups (Batter vs Bowler)

- Batter vs bowler head-to-head runs, balls, SR
- Wicket involvement mapping (where possible)
- Identify favorable/unfavorable matchups
- Venue-specific matchup trends (optional)

---

8) 🏆 Team Analytics + Strategy View

- Team run rate trends by phase
- Team bowling strength by phase
- Team matchup strengths (vs opponents)
- Venue adaptability (home-like advantage patterns)
- Season-wise evolution of teams
- Best XI logic support (future extension)

---

9) 🏟️ Venue Insights

- Venue run scoring trends
- Venue wicket trends (if derivable)
- Bat-first vs chase patterns per venue
- High-scoring vs low-scoring venue classification
- Venue phase behaviors (PP/Middle/Death patterns)

---

10) 📊 KPI Library (Dashboard Metrics Bank)

We will build a reusable metric bank including:

- Runs, balls, SR
- Overs, economy (ER)
- Wickets / dismissal counts (if derivable)
- Phase metrics (PP/Middle/Death)
- Toss + venue effects
- Matchups
- Role distribution + confidence tiers

---

11) 🧱 Dashboard Build (Streamlit Modules)

Create the dashboard in clean modules:

- Overview/Home
- Match Explorer
- Team Dashboard
- Player Dashboard
- Role Explorer
- Matchups Explorer
- Venue Dashboard
  Each module should have:
- Filters (season, team, venue, player, phase)
- Key KPIs
- Charts + tables
- Download option for filtered results

---

12) 🤖 Machine Learning (Optional, Only After Analytics)

We will NOT start ML until analytics is stable.
When we do ML, it must be meaningful:

- Win prediction (pre-match or mid-match)
- Player impact forecasting
- Phase scoring forecast
- Matchup advantage scoring
- Team strength index model
  Also confirm:
- Whether to use CLEAN vs FULL master dataset
- Whether FULL may skew results due to missing match metadata rows

---

✅ SOP (How we will work)

1. Work in small phases, one deliverable at a time
2. First: **display + validate outputs**, then save files
3. No unnecessary notebook markdown, only code + checks
4. Save outputs into `/data/processed/` only (raw stays untouched)
5. Every phase must produce:
   - a clean dataframe output
   - sanity checks (rows, nulls, duplicates)
   - a saved CSV
   - sample preview rows
   - **Save a CSV only if the output is reusable** for downstream phases or the Streamlit dashboard (master table, dimension table, fact table, or stable aggregates).
   - If the phase is only **EDA / validation / one-time analysis**, do **not** save a CSV.

✅ Work completed so far (already done)

1) Matches cleanup

- Venue mapping + removal venues handled
- Final cleaned matches file exists

2) Players cleanup

- Used an Excel mapping sheet (`player_name_vs_full_name.xlsx`)
- Removed old `player_name` + `player_full_name`
- Created `player_updated_name`
- Saved updated raw player file

3) Ball-by-ball cleanup

- Mapped incorrect player names → `batter_updated_name` and `bowler_updated_name`
- Saved cleaned ball-by-ball file

4) Player role classification (final logic)

- Output role column: `player_role_final`
- Sample-size tiers included
  Saved file: `player_performance_with_roles.csv`

5) Team standardization

Saved:

- `teams_reference.csv`
- `ipl_matches_teamnames_standardized.csv`

6) Master dataset created

Two versions exist:

- FULL: `ipl_master_ball_by_ball_full.csv`
  - Rows: 278,205
  - ~14.82% rows missing match metadata (172 match_ids)
- CLEAN: `ipl_master_ball_by_ball_clean.csv`
  - Rows: 236,978
  - Only rows with match metadata available

✅ Final files we will use going forward

Use ONLY these in analysis (do not re-clean unless required):

1. `data/processed/ipl_master_ball_by_ball_clean.csv` ✅ Primary analytics dataset
2. `data/processed/ipl_master_ball_by_ball_full.csv` ✅ Optional for extra coverage
3. `data/processed/player_performance_with_roles.csv`
4. `data/processed/ipl_matches_teamnames_standardized.csv`
5. `data/processed/teams_reference.csv`

✅ What I want next

Start building the dashboard in **phases**, starting from the highest ROI modules first.

First Step in this new chat:

Load `ipl_master_ball_by_ball_clean.csv`, show shape/columns, validate key fields, and propose the **Phase 1 module** with the first KPIs.