ChatGPT

Comprehensive Data Inventory for NFL Betting Analyzer

This document outlines **every dataset and data field** that a production-ready NFL betting analyzer should ingest, store and use. It then checks what is currently present in the project (based on `docs/current_inventory.md`, code inspection and snapshot names) and flags what is missing and therefore required.

A. League & Reference

Dataset	Required Fields (keywords only)	Present?
Teams	team_id, abbreviation, full name, conference, division, head coach/OC/DC, bye week, aliases	⚠ **Partial** – `reference_teams.csv` exists in snapshots but content not verified. Should include coaching staff and aliases.
Stadiums	stadium_id, name, team_id, city/state, latitude/longitude, surface type, roof (dome/outdoor/retractable), elevation (ft)	⚠ **Partial** – `reference_stadiums.csv` exists but no elevation or roof state confirmed.
Players	player_id, name, alternative spellings, position, team_id, status (active/inactive), height, weight, birthdate, draft info, dominant hand, college	⚠ **Partial** – `reference_players.csv` exists but does not appear to include physical/draft details.
Coaches	team_id, season, head coach, offensive coordinator, defensive coordinator, offensive/defensive scheme	**Missing** – no coaches data in snapshots.
Officials	game_id or crew_id, referee, umpire, down judge, line judge, side judge, back judge, field judge	**Missing** – no officials data present.

B. Rosters / Depth / Availability

Dataset	Required Fields	Present?
Weekly rosters	season, week, team_id, player_id, position, jersey, status, last_updated	⚠ Present as `rosters.csv` but content not verified.

Dataset	Required Fields	Present?
Depth charts	season, week, team_id, side (off/def/ST), position, slot, player_id, role, package, depth_rank	⚠ Present as `depth_charts.csv`.
Injury reports	report_date, team_id, player_id, position, practice_status (DNP/LP/FP), game_status (Q/D/O), designation (IR/PUP), return_date	`injuries.csv` exists but unknown if columns match.
Gameday inactives	season, week, team_id, player_id, reason, declaration time	Missing – no `inactives.csv`.
Transactions/ elevations	date, team_id, player_id, type (sign/waive/IR/PUP/elevation), details	Missing.

C. Schedule / Games / Context

Dataset	Required Fields	Present?
Schedule	season, week, game_id, kickoff_utc (with timezone), home_id, away_id, network, referee crew, stadium_id	⚠ Present as `schedules.csv` but timezone fields not verified.
Game metadata	game_id, roof_state, field_type, attendance, duration, closing_spread, closing_total	Missing – `games.csv` not listed in snapshots.
Rest/travel context	season, week, team_id, opponent_id, rest_days, travel_miles, tz_delta, pace_sn, pace_all, PROE, lead/trail/neutral splits	Missing – no `team_context.csv`.

D. Play-by-Play & Drives

Dataset	Required Fields	Present?
PBP	play_id, game_id, quarter, clock, offense, defense, yardline, down, distance, yards_gained, play_type, EPA, WPA, success, air_yards, YAC, pressure, blitz, personnel_off, formation, motion, shotgun, no_huddle, penalty_yards	⚠ Present as `pbp.csv` but content limited; unknown if includes advanced fields (EPA, air_yards, pressure etc.).
Drives	drive_id, game_id, offense, start_quarter, start_clock, start_yardline, end_quarter, end_clock, result, plays, yards, time_elapsed, points	Missing – no `drives.csv`.

E. Participation & Usage

Dataset	Required Fields	Present?
Snap counts	season, week, team_id, player_id, offensive snaps, defensive snaps, special teams snaps, offensive pct, defensive pct, ST pct	`snaps.csv` exists.
Routes	season, week, team_id, player_id, routes_run, route_participation	Missing – no `routes.csv`.
Usage shares	season, week, team_id, player_id, carry_share, target_share, red-zone_touch_share, goal-line_carry_share, pass_block_snaps, align_slot, align_wide, align_inline, align_backfield	Missing.

F. Box & Advanced Stats (per game)

Dataset	Required Fields	Present?
Passing stats	game_id, player_id, attempts, completions, yards, touchdowns, interceptions, sacks, sack_yards, yards_per_attempt, air_yards, average_depth_of_target (aDOT), fumbles	⚠ Present in root-level code but unclear if in snapshots.
Rushing stats	game_id, player_id, rush_attempts, rush_yards, rush_tds, long_run, yards_per_carry, fumbles	⚠ Present in code but not in snapshot.
Receiving stats	game_id, player_id, targets, receptions, receiving_yards, receiving_tds, air_yards, YAC, aDOT, drops, long_reception	⚠ Present in code but not in snapshot.
Defensive stats	game_id, player_id, tackles, assists, sacks, tackles for loss, QB hits, interceptions, pass breakups, defensive_tds	Missing.
Kicking stats	game_id, player_id, field_goals_made_0_39, field_goals_made_40_49, field_goals_made_50plus, field_goal_attempts, field_goals_made, extra_point_attempts, extra_points_made	Missing.

G. Team Rates & Splits

Dataset	Required Fields	Present?
Team splits	season, week, team_id, pace_sn, pace_all, PROE, red_zone_efficiency, goal_to_go_efficiency, third_down_conv_rate, fourth_down_att_rate, vs_pos_rb_yds, vs_pos_wr_yds, vs_pos_te_yds	Missing – not derived.

H. Weather

Dataset	Required Fields	Present?
Weather	game_id, stadium_id, temperature_f, humidity, wind_mph, wind_dir, precip_type, precip_prob, conditions, timestamp_utc	⚠ Present as `weather.csv` but includes no timestamp and minimal fields; not full.

I. Odds / Props

Dataset	Required Fields	Present?
Odds (point-in-time)	ts_utc, book, market, selection_id, selection_name, team_id, player_id, line, price, source	⚠ Present as `odds.csv` but based on mock odds only; real odds provider missing.
Odds history / closing lines	ts_utc, book, market, selection_id, line, price, event_id, is_closing	Missing.

J. Fantasy Scoring & Player Roles

Dataset	Required Fields	Present?
Fantasy scoring	scoring rules for PPR/half/standard; bonuses; DST scoring; fumble rules	Missing (implied but not codified).

K. Modeling Artifacts & Evaluation

Dataset	Required Fields	Present?
Model artifacts	model_id, position, target, algorithm, version, train_start, train_end, features[], r2, mae, rmse, calibration_plot	⚠ Present partially – `models/streamlined` exists in gating logic but directory contents not visible; sidecar JSON with metrics is not confirmed.
Backtest reports	hit_rate, ROI, Brier score, CRPS, n_bets, date range, plot files	Missing.
Coverage matrices	valid CSVs listing markets vs models and stats vs features	⚠ Present but malformed – needs regeneration as valid CSVs.

Summary of Imperative Missing Data

The following datasets are **missing and must be added** to achieve complete coverage:

- **Coaches** and **officials** data.

- **Gameday inactives** and **transactions** for roster churn.
- **Detailed game metadata** (roof state, field type, attendance, closing spread/total) and derived rest/travel context.
- **Drives** dataset for drive-level analysis.
- **Routes run** and **usage share** metrics (carry share, target share, red-zone and goal-line shares, alignments, pass-block snaps).
- **Full box stats** for rushing, receiving, defense, and kicking; only passing metrics appear to be partially available.
- **Team splits** (pace, PROE, red-zone efficiency, etc.).
- **Odds history and closing lines**; present odds are only from mock snapshots.
- **Fantasy scoring definitions** (e.g., PPR vs standard) in a structured file.
- **Backtest reports** with real metrics; currently absent.
- **Valid coverage matrices** (the existing CSVs are malformed).

New Documentation Recommended

- `docs/DATA_DICTIONARY.md` – includes the tables above (A–K) for quick reference.
- `docs/SNAPSHOT_SCHEMAS.md` – lists each snapshot file and required columns to enforce schema via tests.
- `docs/ODDS_PROVIDERS.md` – outlines how to configure real sportsbook APIs (e.g., The Odds API) and how to map their fields into the internal `odds.csv` schema.

Docs to Remove or Archive

- **Stale audit files** (`audit_report.md`, `audit_report.json`) from earlier phases; these misreport the state and clutter the docs.
- **Marketing-style “final verification” docs** (e.g., `FINAL_VERIFICATION_REPORT.md`) that assert success without evidence; archive them to `docs/archive/` or delete.
- Any duplicated READMEs or old instructions in subdirectories that conflict with the unified plan.
