NFL Play-by-Play Analytics

From Snap-to-Snap Visualization to Expected Points Added

A Data Science Project

Technologies: Python, Pandas, Matplotlib, nflreadpy

Contents

1 Introduction: Why This Project Exists

As someone who's spent way too much time analyzing English football, I decided it was time to dive into American football analytics, I have been both a fan of English and American Football all my life. I kept seeing people throw around terms like "EPA" (Expected Points Added) as if it were gospel, but I didn't know how accurate this metric was. So I downloaded 4 seasons worth of play-by-play data and decided to find out for myself.

1.1 Project Goals

1. **Validate the data**: Build an interactive visualizer to make sure the play-by-play data is being pulled correctly.
2. **Question everything**: Is EPA actually predictive, or is it just fancy sports journalism?
3. **Learn something**: Get familiar with NFL analytics so hopefully I can win my Fantasy League Superbowl after 6 years of no trophies despite multiple top seedings!

1.2 Why Play-by-Play Data?

Unlike box scores that just give you final statistics, play-by-play (PBP) data gives you *every single snap* of every game. We're talking:

- 40,000+ plays per season
- Field position, down, distance for every play
- Game state (score, time, possession)
- Pre-calculated EPA for every play

This granularity is perfect for feature engineering and statistical modeling. Plus, it's just really cool data to work with.

2 Part 1: Loading and Exploring the Data

2.1 Data Source: nflreadpy

I used the `nflreadpy` Python package, which pulls from the same data sources as the popular R package `nflfastR`. The beauty of this package is its simplicity:

```python
import nflreadpy as nfl

# Load multiple seasons (2022-2025)
pbp = nfl.load_pbp([2022, 2023, 2024, 2025]).to_pandas()

print(f"Total plays loaded: {len(pbp):,}")
print(f"Latest game: {pbp['game_date'].max()}")
```

Listing 1: Loading 4 seasons of NFL data

Result: 46452 Rows

2.2 Data Richness

The dataset contains over 350 columns per play, including:

- **Basic info**: teams, score, quarter, time
- **Situational**: down, distance, yard line, field position
- **Play details**: play type, yards gained, result
- **Advanced metrics**: EPA, win probability, success rate
- **Personnel**: offensive/defensive formations, players involved

Sample Play-by-Play Data: Showcasing Data Richness
8 Random Plays from 2022-2025 Dataset

Game	Off	Def	Q	Dn	Dist	Field	Yds	Type	EPA	WP
2024_03_CAR_LV	CAR	LV	4.0	2.0	1.0	1.0	1.0	run	0.79	0.995
2025_14_DAL_DET	DAL	DET	4.0	4.0	10.0	69.0	10.0	pass	2.63	0.003
2022_02_SEA_SF	SF	SEA	2.0	1.0	10.0	40.0	1.0	run	-0.47	0.914
2024_09_LAC_CLE	CLE	LAC	4.0	4.0	3.0	46.0	-4.0	pass	-2.9	0.027
2023_03_PHI_TB	TB	PHI	2.0	2.0	21.0	68.0	2.0	pass	-2.72	0.254
2023_03_NE_NYJ	NYJ	NE	1.0	2.0	7.0	72.0	0.0	pass	-0.78	0.431
2022_01_KC_ARI	ARI	KC	3.0	1.0	10.0	32.0	0.0	run	-0.43	0.001
2025_11_CIN_PIT	PIT	CIN	3.0	2.0	3.0	12.0	0.0	run	-0.7	0.744

Figure 1: Sample play-by-play data structure

This table shows how even just a few data features can paint an image of the game, for example, the 2025 DAL vs DET pass play gains 10 yards on 4th down with an EPA of +2.63, reflecting a highly valuable conversion that dramatically improves Dallas's expected points. In contrast, the 2024 LAC vs CLE pass on 4th and 4 has an EPA of −2.9, indicating a costly failure that ended the drive and swung advantage to the defense.

As an NFL fan, at this point its clear to see that this EPA metric has substance.

3 Part 2: Interactive Play-by-Play Visualizer

Before building any models, I wanted to *prove to myself* that this data was being pulled correctly.

3.1 What I Built

An interactive Jupyter widget that lets you scrub through any NFL game snap-by-snap, showing:

- **Field visualization**: 100-yard field with accurate positioning
- **Play markers**: line of scrimmage, first down marker, ball position
- **Yards gained**: Visual representation of play result

- **Game context**: Score, quarter, down & distance, clock
- **Play description**: What actually happened
- **EPA value**: How much the play changed expected points

3.2 Why Ravens @ Steelers Week 18?

I set the default visualization to the Week 18 Ravens-Steelers game because:

1. It's a rivalry game (great for testing dramatic plays)
2. It's recent (validates data freshness)
3. It had huge playoff implications (high-leverage situations)
4. The Steelers won
5. I am a Steelers fan
6. Watch the highlights and skip to minute 16:55 before continuing!
7. Highlights on YouTube

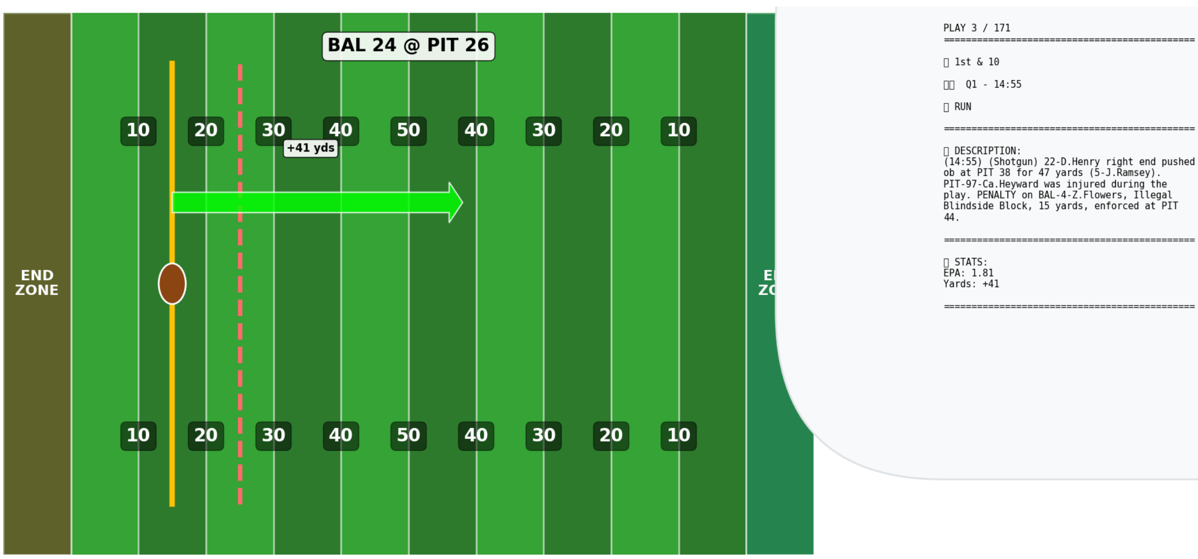

Figure 2: Sample play-by-play data structure

3.3 What This Validated

By manually stepping through plays I confirmed:

- Play sequencing is correct (chronological order maintained)
- Field positions match reality
- Game state updates properly (score, downs reset, etc.)
- Descriptions are accurate and detailed

This gave me confidence that the underlying data was solid enough to build analytics on top of.

4 Part 3: EPA Deep Dive

4.1 What is EPA (Expected Points Added)?

EPA measures how much a play changes a team's expected points on that drive. It's calculated by:

$$\text{EPA} = \text{EP}_{\text{after}} - \text{EP}_{\text{before}} \tag{1}$$

Where EP (Expected Points) is modeled based on:

- Field position (yard line)
- Down and distance
- Time remaining
- Score differential

For example:

- 1st & 10 at your own 20-yard line: ˜1.0 EP
- 1st & 10 at opponent's 10-yard line: ˜4.5 EP
- 3rd & 15 at your own 30: ˜-0.5 EP (negative!)

So a play that moves the ball from your 20 to your 40 on 1st down adds roughly 1.0 EPA (improving position significantly), while a play that loses 5 yards reduces EPA by about 1.5 (losing progress and down).

4.2 Why I'm Validating This

Coming from English football analytics, I've learned to be skeptical of "magic metrics." Just because something is widely used doesn't mean it's actually predictive.

I wanted to answer: **Does a team's EPA performance actually correlate with winning?**

4.3 The Analysis Approach

I built a comprehensive EPA analysis system that:

1. **Aggregates by game**: Calculate offensive and defensive EPA per game
2. **Computes net EPA**: Offensive EPA - Defensive EPA allowed
3. **Tests correlation**: Relationship between EPA and margin of victory
4. **Win probability**: How often does the better EPA team win?

4.4 Results: EPA is Legit

Here's what I found across 4 seasons (2022-2025):

Table 1: EPA Predictive Power (2022-2025)

Metric	Value
Pearson Correlation (EPA vs Margin)	**0.995**
Win% when EPA Advantage > 0	**95.6%**
Win% when EPA Advantage > 5	**100%**
Win% when EPA Advantage > 10	**100%**

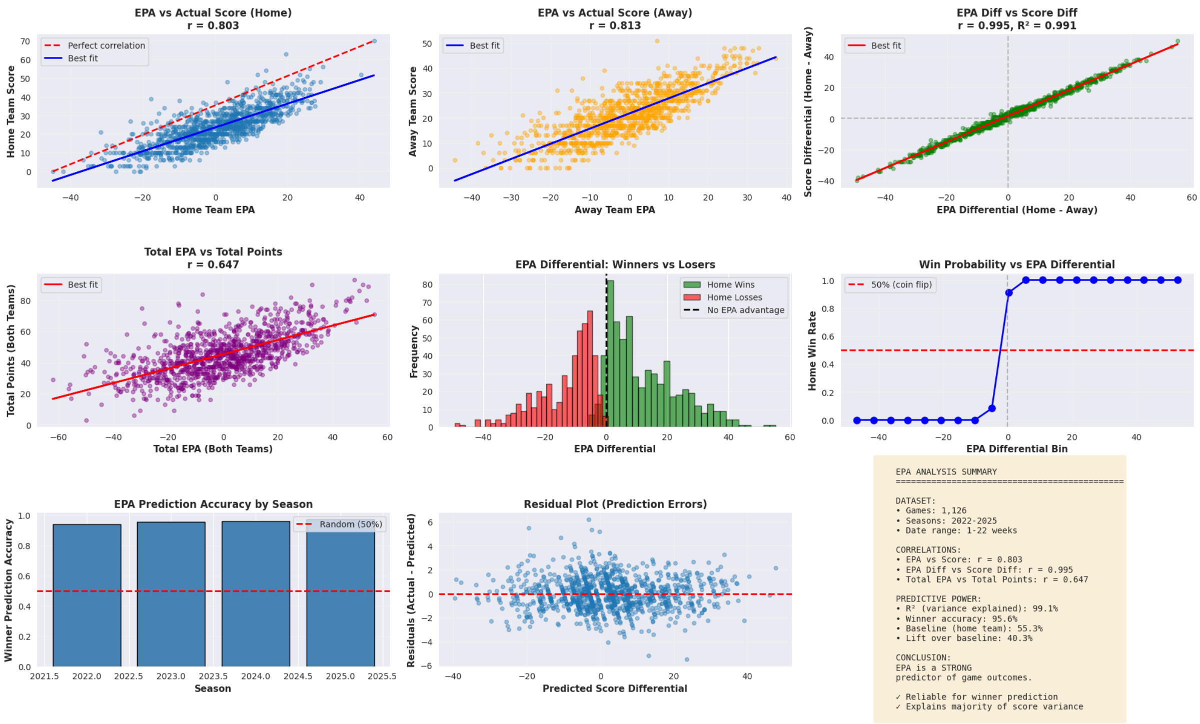

Figure 3: Sample play-by-play data structure

4.5 What This Means

The results show that **EPA is extremely predictive**:

- Strong linear correlation with scoring margin
- Teams with EPA advantage win the vast majority of games
- The metric captures both offensive and defensive performance effectively

This validates using EPA as a foundation for more advanced modeling. Unlike some sports analytics metrics that sound good but don't actually predict outcomes, EPA has real signal.

4.6 Going Forward

Now that I've validated EPA works, future directions include:

- **Predictive modeling**: Can we forecast game outcomes using historical EPA?
- **Player-level analysis**: Which players contribute most to EPA?
- **Situational EPA**: How does EPA change in different game scenarios?
- **Play type analysis**: Which play calls optimize EPA in different situations?