Sydney Lockwood

The Relationship Between Sacks and Turnovers on Win Probability in the NFL

Pipeline Overview & Target Grain

One row represents a team in a single game, season, total sacks, total turnovers caused, wins, and losses. The files that feed into it are play-by-play data from NFL Savant and game win/loss data from GitHub NFLVerse. Excel is used for initial cleaning and filtering, while Python handles merging, standardizing names, and aligning game IDs.

Simple Flow

Raw files-> standardize team names ->merge with game results -> aggregate sacks/turnovers to teams -> create derived features (Win probability) -> Final clean file

ID & Mapping Strategy

Each row is uniquely identified by GameID and team. Team names were normalized to abbreviations to avoid name collisions. Mappings for all data are in the same document under different tabs for each year included in the analysis.

Standardization Rules

Sacks, turnovers, and WPA were converted to integers, and win probability was converted to decimals. The data set uses the NFL standard 3-letter list for team abbreviations.

Reshaping & Integration

Reshaping involved aggregating sacks by defensive team and aggregating turnovers by offensive team per game. When sources disagree, the game results dataset overrides for final scores and wins/losses. Play-by-play data wins for sacks and turnover data. Dropped rows include preseason and postseason games and plays with missing team codes.

Feature & Transformation Spec

Feature Name	Rule	Inputs	Before	After
SackRate	Sacks/DefensivePassSnaps	Sacks, Pass Attempts	3 sacks on 28 pass snaps	0.107
TurnoverRate	Takeaways-Giveaways	interceptions, fumbles	2 takeaways 1 giveaway	+1
Win Probability	Derived from final score	Wins, losses		Decimal (0-1)

Validation Gates (QA)

Row	Description	Row Count
Raw PBP Data	Every play, all season	232,192
Filter to regular season only	Remove preseason and postseason	232,152
Aggregate sacks & turnover to team-game level	Each team in each game	2,541
Raw game results	One row per game	1,681
Game results duplicated to a team perspective	Home & away teams Delete 2025 game data	3,089
Final merged dataset	Merged sacks/turnovers + results	2,873

Duplicate-Key Violations

Before cleaning, there were many duplicate key violations in reference to the team abbreviations across the two different datasets. After standardization of team name abbreviations, there were zero violations.

Missingness by field

About 5% of sack values were missing, mainly due to incomplete play records. These were resolved during aggregation. Turnovers were missing in 1.7% of cases and were also resolved through aggregation. Overall, missingness was low and did not meaningfully affect the dataset.

Min/max checks

Sacks per team per game ranged from 0-11, although 11 is a historically high number, this is valid and confirmed through game records. Turnovers ranged from 0-8, a valid number, not too high or out of the ordinary.

Logic Test 1 -> Turnover Directionality

A turnover credited to a defensive team must correspond to the offensive team losing possession

Resulted in 14 mismatched turnovers due to misaligned team codes. Easily fixed.

Logic Test 2 -> Sack Possession Rule

A sack can only be attributed to the defensive team when the offense attempts a pass

Resulted in 9 incorrectly attributed plays where sacks were labeled on run plays. Was able to exclude the plays where the play_type=pass

Runbook & Reproducibility

- Raw files
  - Located in the Excel workbook under each tab (2020,2021, games, etc.)
- Clean file
  - The clean dataset is in the Excel workbook under Clean Data
- Steps to re-run
  1. Load raw play-by-play data
  2. Load raw game results
  3. Standardize team names

4. Filter to regular-season games only
5. Aggregate play-level data to team-game level
6. Duplicate the game table to create a team perspective
7. Merge aggregated PBP data with game results
8. Run QA validation
9. Export the final clean dataset

- What has changed since CP3?
    - Since CP3, I have shifted from planning to execution, completing full cleaning, standardization, integration, and validation. I now have a finalized dataset ready for modeling and statistical analysis.

Data Dictionary, also attached as a file. Included for visibility.

Variable Name	Data Type	Defenition	Source (Original/Derived)	Notes (Rules?)
Game ID	String/Integer	Unique identifier for each game	Original	YYYYMMDD##
Season	Integer	NFL Season year	Original	YYYY
Team	String	Team abbreviation	Original	3-letter NFL code
Win	Integer	whether the team won the game	Derived	0=loss 1=win
Loss	Integer	Whether the team lost the game	Derived	0=loss 1=win
Sacks	Integer	Total sacks recorded by the defense	Original	
Turnovers	Integer	Total turnovers caused by the defense	Original	
sack_rate	Float	Sacks per defensive pass attempt	Derived	Calculated as sacks / defensive pass attempts
turnover_rate	Float	Net turnovers caused	Derived	Calculated as takeaways − giveaways
win_probability	Float	Predicted probability of winning	Derived	Derived from final score